МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Нижегородский государственный университет им. Н.И. Лобачевского

М.С. Тихов
Н.В. Капкаев

ПОСТРОЕНИЕ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ

Учебно-методическое пособие

Рекомендовано методической комиссией
Института информационных технологий, математики и механики
для магистров ННГУ, обучающихся по направлению подготовки
01.04.02 «Прикладная математика и информатика»

Нижний Новгород
2022

УДК 519.2
ББК 22.1
Т-46

Т-46 Тихов М.С., Капкаев Н.В. ПОСТРОЕНИЕ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ: Учебно-методическое пособие. – Нижний Новгород: Нижегородский госуниверситет, 2022. – 133 с.

Рецензент: д.ф.-м.н., профессор **А.В. Калинин**

Учебно-методическое пособие содержит методические указания и теоретические сведения, разъясняющие тему «Построение вероятностных моделей» и практические задания по этой теме. Решение задач дает возможность полнее и глубже изучить данный раздел, а также использовать этот раздел как базу изучения темы «Построение вероятностных моделей». Пособие будет полезно при изучении дисциплины «Современные проблемы прикладной математики и информатики» студентам, обучающихся по направлению подготовки 01.04.02 «Прикладная математика и информатика».

УДК 519.2
ББК 22.1

Содержание

1. Функции случайных величин

Закономерности, которым подчиняется исследуемая случайная величина X , физически обуславливается реальным комплексом условий ее наблюдения (или эксперимента), а математически задаются соответствующим вероятностным пространством $(\Omega, \mathscr{A}, \mathbf{P})$, или, соответствующим законом распределения вероятностей. Иными словами, в теории вероятностей функция распределения $F(x)$ исследуемой случайной величины X предполагается полностью известной и по $F(x)$ мы можем предсказать, куда более вероятно попадают значения с.в. X . В математической статистике рассматриваемая задача является, в известном смысле обратной задачей теории вероятностей: по наблюдениям $x_1, x_2, \dots, x_n$ за каким либо процессом требуется указать неизвестную нам функцию распределения. При этом наблюдения $x_1, x_2, \dots, x_n$ интерпретируются как значения случайных величин $X_1, X_2, ..., X_n$. Мы будем считать, что $X_1, X_2, ..., X_n$ независимы и одинаково распределены (нор св) с неизвестной функцией распределения $F(x)$, принадлежащей известному классу $\mathscr{F}$. Степень неизвестности (информированности) о классе $\mathscr{F}$ тоже может быть различной. Так, если имеются общие сведения о классе $\mathscr{F}$ типа гладкости, т.е. множество $\mathscr{F}$ является *бесконечномерным* подпространством какого-либо функционального пространства и нет предпочтений одного распределения перед другим, то мы будем заниматься оцениванием распределений в виде гистограмм, эмпирических функций распределения, ядерных оценок плотностей и т.д.; если же в этой ситуации есть предпочтения одних распределений перед другими, например, есть гипотеза о том что искомое распределение имеет заданный вид, то мы будем заниматься проверкой гипотез о согласии. Класс $\mathscr{F}$ может иметь следующий вид: $\mathscr{F} = \{F(x;\theta), \theta \in \Theta \subset R^m\}$, функция $F(x;\theta)$ <u>известна</u>, а параметр $\theta -$ <u>неизвестен</u>, т.е. $\mathscr{F}$ является *конечномерным*, и речь идет об определении этого параметра. Если нет предпочтений одного значения параметра перед другими, то мы будем заниматься оценкой параметра, если же есть предпочтения одних значений параметра перед другими, то в этом случае задача ставится как задача проверки гипотез о параметрах. В том случае, когда мы изучаем многомерные случайные величины, то, как правило, производится анализ взаимосвязей, который включает в себя как задачи оценивания, так и задачи проверки гипотез.

Рассматриваемая задача является, в известном смысле, обратной задачей теории вероятностей. Наблюдается случайное явление, математически описываемое некоторой вероятностной мерой $\mathbf{P}$ на измеримом пространстве $(\Omega, \mathscr{A})$ всех его элементарных исходов $\omega \in \Omega$. В качестве элементарного исхода в математической статистике обычно рассматривается $\omega = (x_1, x_2, \dots, x_n)$, а в качестве $\mathscr{A}$ берется борелевская $\sigma-$алгебра. Качественное описание явления – пространство $(\Omega, \mathscr{A})$ – наблюдателю известно, а мера $\mathbf{P}$ неизвестна. Ее надо указать по наблюдениям явления.

Хорошо известно, что нетривиальные обратные задачи математической физики оказываются некорректными. Нужны дополнительные ограничения (например, требование гладкости ответа), чтобы задача получила строгое математическое содержание. В математической статистике необходимость такого ограничения более или менее узким семейством $\mathscr{P}$ была интуитивно понята на первых же этапах ее развития. Обсудим ее сейчас на формальном уровне.

В математической статистике обычно имеют дело с измеримыми пространствами $(\Omega, \mathscr{A})$ следующих трех типов: А) когда $\mathscr{A}$ – конечная алгебра (в этом случае Ω является объединением атомов A_j); В) когда Ω является счетным объединением атомов A_j ; С) когда $(\Omega, \mathscr{A})$ является лебеговым измеримым пространством. В случаях В и С алгебры $\mathscr{A}$ бесконечны, но порождаются последовательностями конечных подалгебр $\mathscr{A}_m$.

Отсутствие априорной качественной информации о $\mathbf{P}$ означает, что за семейство $\mathscr{P}$ априори допустимых ответов мы должны принять совокупность всех вероятностных распределений на $(\Omega, \mathscr{A})$.

О п р е д е л е н и е . Пусть $\mathscr{P}$ – семейство вероятностных мер (распределений) на измеримом пространстве $(\Omega, \mathscr{A})$. *Статистической структурой* называется тройка $(\Omega, \mathscr{A}, \mathscr{P})$.

Довольно часто семейство $\mathscr{P}$ имеет следующий вид $\mathscr{P} = \{ \mathbf{P}_\theta , \theta \in \Theta\}$. Понятие статистической структуры играет в математической статистике такую же роль, что и вероятностное пространство в теории вероятностей.

Пример. При статистическом эксперименте, состоящем в проведении n независимых наблюдений над случайной величиной, принимающей лишь конечное число k значений вероятности которых полностью не известны, но остаются постоянными в течение эксперимента, приходят к структуре вида:

а) $\Omega = \{1, 2, \dots, k\}^n$ с $\sigma -$алгеброй всех подмножеств Ω;

б) $\Theta -$ симплекс в R^k; $\theta_1, \theta_2, \dots, \theta_k \in \Theta$: $\theta_1 + \theta_2 + \dots + \theta_k = 1$, $\theta_1 \ge 0, \theta_2 \ge 0, \dots, \theta_k \ge 0$.

Нередко в статистических задачах делается предположение о независимости и одинаковой распределенности с.в. $X_2, X_2, \dots, X_n$. Если кроме этого о распределении $\mathbf{P}^X$ ничего не известно, то класс $\mathscr{P}$ состоит из всех функций распределения $F(x)$, которые могут быть представлены в виде: $F(x) = \prod_{i=1}^{n} G(x_i)$, где $G(x)$ может быть произвольной одномерной функцией распределения. В некоторых задачах постулируется только независимость с.в. $X_2, X_2, \dots, X_n$. Класс $\mathscr{P}$ является тогда классом всех функций распределения $F(x)$, которые могут быть представлены в виде $F(x) = \prod_{i=1}^{n} G_i(x_i)$, где $G_i(x)$ произвольные одномерные ф. р. Иногда $X_2, X_2, \dots$ предполагают стационарно связанными, с одним и тем же маргинальным распределением $\mathbf{P}^X$.

Описание выше – это, в определенной степени, классическое описание модели, когда выборка является полной. В последнее время много внимания стало уделяться как построению вероятностных моделей по неполным выборкам (цензурированные выборки, частично наблюдаемые выборки и т.д.), так и обоснованию вероятностных моделей. Это связано с построением вероятностных моделей по экономическим данным, а также с исследованием, так называемых трафиков.

Заметим, что статистический аппарат нельзя считать пригодным для одной конкретной модели, правильнее будет рассматривать его как помощь при интерпретации экспериментальных данных через посредство различных моделей. Выше в качестве таких моделей мы рассматривали параметрические модели, когда делаются предположения о классе распределений, например, нормальное, полиномиальное и т.д. Однако статистики пытаются создать и такие статистические процедуры, которые бы зависели от возможно меньшего числа предположений (так называемые, непараметрические методы), а также устойчивые статистические процедуры (оценивание параметров и проверка гипотез), когда качество процедуры сохраняется при нарушениях предположений о виде распределения.

Задачи математической и прикладной статистики. Как прикладная статистика (многомерные статистические методы), так и математическая статистика относятся к термину статистика, понимаемой как отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых количественных данных. Под математической статистикой (согласно Большому Энциклопедическому Словарю) понимается наука о методах систематизации и использования статистических данных для научных и практических выводов. В узком смысле (или в нашем понимании) под методами математической статистики принято понимать лишь те методы статистической обработки данных, разработка и использование которых апеллируют к вероятностной природе этих данных. При этом развиваемый в рамках широкого понимания прикладная статистика – это класс методов статистической переработки исходной информации, которые априори *не опираются на вероятностную природу обрабатываемых данных* (представителями методов такого типа являются, например, разнообразные методы *кластер-анализа, многомерного шкалирования, теории измерений* и др.), остается за рамками научной дисциплины «математическая статистика».

Кроме того, при конкретном применении методов статистической обработки к реальной ситуации приходится решать следующие задачи:

- глубоко вникать в содержательную сущность задачи, адекватно прилаживать исходные модельные допущения к выяснению сущности реальной задачи;
- решать задачу преобразования имеющейся исходной информации к стандартной (унифицированной) форме записи обрабатываемых статистических данных;
- разрабатывать практически реализуемые вычислительные алгоритмы и программное обеспечение с учетом специфики обрабатываемой статистической информации и возможностей имеющейся вычислительной техники.

Понятийный аппарат, методы и результаты, позволяющие проходить эту дистанцию, вместе с этапом прилаживания и доработки необходимого математического инструментария и составляют главное содержание прикладной статистики.

Таким образом, мы приходим к *определению прикладной статистики как самостоятельной научной дисциплины, разрабатывающей и систематизирующей понятия, приемы, математические методы и модели, предназначенные для организации, сбора, стандартной записи, систематизации и обработки статистических данных с целью их удобного представления, интерпретации, получения научных и практических выводов и прогноза дальнейшей ситуации*.

Подводя итог сказанному, мы будем рассматривать задачу построения вероятностной модели следующим образом: 1) исходя из конкретной ситуации, в которой мы проводим эксперимент, построим семейство распределений; 2) проверим эту модель с помощью статистических критериев; 3) вернемся снова к модели и используя накопленный опыт обоснуем её.

А.Н.Колмогоров уделял большое внимание обоснованию той или иной вероятностной модели. Об этом можно узнать из книги [6]. Здесь же мы уделяем большое внимание исследованию модели «доза-эффект». Чтобы квалифицированно обосновать ту или иную модель необходимо уметь строить вероятностные модели функций случайных величин по заданным распределениям первоначально заданным распределениям. Отметим также разницу между тем, что изучалось в классической теории вероятностей и тем, что мы будем изучать в настоящем курсе: раньше мы изучали нор св, здесь наблюдения будут неодинаково распределены и даже намешаны каким-либо распределением.

Основой анализа является ***преобразование*** (непрерывных) случайных величин (с.в.). Пусть с. в. X имеет абсолютно непрерывную функцию распределения $F_1(x)$ и плотность распределения (по лебеговой мере) $f_1(x)$, а $Y = g(X)-$ преобразование этой случайной величины. Если $g(x)-$ измеримая функция, то $Y-$тоже случайная величина. Если $Y-$непрерывная случайная величина с плотностью $f_2(y)$ и функция $g(x)-$известна, то плотности $f_1(x)$ и $f_2(y)-$ связаны. Именно, пусть функция $g(x)-$взаимно-однозначна и обратная функция $h(y)-$непрерывно дифференцируема в точке y. Тогда

$$f_2(y) = f_1(h(y))\cdot|h'(y)|. \qquad (1)$$

По-сути, соотношение (1) есть замена переменной под знаком интеграла. Если обратная функция имеет несколько непрерывно дифференцируемых ветвей $h_j(y),\ j = 1, 2, ..., m,$ то

$$f_2(y)=\sum_{j=1}^{m} f_1(h_j(y))\cdot|h_j'(y)|. \qquad (2)$$

В качестве <u>примера</u> рассмотрим распределение случайной величины $Y=X^2$, где величина X непрерывна и имеет плотность распределения $f(x)$.

Ясно, что $\boldsymbol{P}(Y\ge 0)=1$, поэтому будем рассматривать значения $y\ge 0.$ Тогда

$F_Y(y)=\boldsymbol{P}(Y<y)=\boldsymbol{P}(X^2<y)=\boldsymbol{P}(-\sqrt{y}<X<\sqrt{y})=F_X(\sqrt{y})-F_X(-\sqrt{y})$.

Поэтому $f_Y(y)=\dfrac{1}{2\sqrt{y}}(f(\sqrt{y})+f(-\sqrt{y})),\ y\ge 0.$ В частности, если $f(x)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, то $f_Y(y)=\dfrac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}},\ y>0,$ т.е. хи-квадрат распределение с одной степенью свободы (χ_1^2).

Пусть $(X_1,X_2,...,X_k)-$ случайный вектор с плотностью распределения $f_1(x_1,x_2,...,x_k)$. Рассмотрим преобразование

$$\{\ Y_j=g_j(X_1,X_2,...,X_k),\ j=1,2,...,k. \qquad (3)$$

Предположим, что система (3) имеет единственное непрерывно-дифференцируемое решение для каждой фиксированной точки $\{\ X_i=h_i(Y_1,Y_2,...,Y_k),\ i=1,2,...,k,$ причем $J=J(y_1,...,y_k)=\det(\partial x_i/\partial y_j)_{k\times k}$ есть якобиан преобразования и $J\ne 0,$ тогда плотность $f_2(y_1,y_2,...,y_k)$ есть плотность распределения вектора $(Y_1,Y_2,...,Y_k)$ будет определяться по формуле

$$f_2(y_1,y_2,...,y_k)=f_1(h_1(y_1,...,y_k),...,h_1(y_1,...,y_k))\cdot|J|. \qquad (4)$$

Замечание 1. Пусть X_1 и X_2- независимые случайные величины с плотностями по лебеговой мере соответственно $f_1(x)$, $f_2(x)$ и функциями распределения $F_1(x)$, $F_2(x)$. Рассмотрим отношение этих величин $Y_1=\dfrac{X_1}{X_2}$ (здесь $\boldsymbol{P}(X_2=0)=0$).

1) *Прямой вывод распределения частного.*

$$F_{Y_1}(y)=\boldsymbol{P}\left(\frac{X_1}{X_2}<y\right)=\boldsymbol{P}\left(\frac{X_1}{X_2}<y,X_2>0\right)+\boldsymbol{P}\left(\frac{X_1}{X_2}<y,X_2<0\right)=$$

$$=\int_0^\infty \boldsymbol{P}\left(\frac{X_1}{X_2}<y\mid X_2=x_2\right)f_2(x_2)\,dx_2+\int_{-\infty}^0 \boldsymbol{P}\left(\frac{X_1}{X_2}<y\mid X_2=x_2\right)f_2(x_2)\,dx_2=$$

$$=\int_0^\infty \boldsymbol{P}\left(\frac{X_1}{x_2}<y\mid X_2=x_2\right)f_2(x_2)\,dx_2+\int_{-\infty}^0 \boldsymbol{P}\left(\frac{X_1}{x_2}<y\mid X_2=x_2\right)f_2(x_2)\,dx_2=$$

$$=\int_0^\infty \boldsymbol{P}\left(\frac{X_1}{x_2}<y\right)f_2(x_2)\,dx_2+\int_{-\infty}^0 \boldsymbol{P}\left(\frac{X_1}{x_2}<y\right)f_2(x_2)\,dx_2=\int_0^\infty \boldsymbol{P}(X_1<yx_2)\,f_2(x_2)\,dx_2+\int_{-\infty}^0 \boldsymbol{P}(X_1>yx_2)\,f_2(x_2)\,dx_2=$$

$$=\int_0^\infty F_1(yx_2)\,f_2(x_2)\,dx_2+\int_{-\infty}^0(1-F_1(yx_2))\,f_2(x_2)\,dx_2,$$

откуда дифференцируя получаем

$$f_{Y_1}(y)=\int_{-\infty}^{\infty}|x|\ f_1(yx)\,f_2(x)\,dx \qquad (5)$$

В частности, пусть $\mathbf{P}(X_1>0)=\mathbf{P}(X_2>0)=1\Rightarrow\mathbf{P}(Y_1>0)=1,$ то $f_{Y_1}(y)=\int_{-\infty}^{\infty}|x|\ f_1(yx)\,f_2(x)\,dx$.

Например, пусть $X_1, X_2 \in N(0,1)$ и независимы, $Z = X_1 / X_2$, тогда

$$f_Z(z) = \frac{1}{\pi}\int_0^{+\infty} x e^{-\frac{x^2z^2}{2}} e^{-\frac{x^2}{2}} dx = \frac{1}{\pi(1+z^2)} \text{ – распределение Коши.}$$

Обобщим этот результат: пусть $N_v(x) = \dfrac{x^v e^{-x^2/2}\ \Gamma(v/2+1)}{\sqrt{2\pi}\cdot 2^{v/2}\ \Gamma(v+1)}\ {}_1F_1\left(\dfrac{v}{2}; v+1; \dfrac{x^2}{2}\right)$, $x \ge 0$, (плотность) $\dfrac{\Gamma(v/2+1)}{\Gamma(v+1)}\ {}_1F_1\left(\dfrac{v}{2}; v+1; \dfrac{x^2}{2}\right) = \sum\limits_{m=0}^{\infty}\dfrac{\Gamma(v/2+m)\, x^{2m}}{m!\,\Gamma(v+1+m)\cdot 2^m}$, и $X_1, X_2 \in N_v(x)$ и независимы, $Z = X_1 / X_2$, тогда $f_Z(z) = \dfrac{1}{2\pi(1+z^2)}$. Действительно, если $N_0(w) = (2\pi)^{-1/2}\exp(-w^2/2)$ и $\int\limits_0^{\infty} y\, J_v(wy)\, N_0(xy)\, dy = x^{-2} N_v(w/x)$, (см. [15], p.394, равенства (3) и (4)), а также $\int\limits_0^{\infty} y\, J_v(wy)\, N_0(y)\, dy = N_v(w)$, где $J_v(x)$ есть функция Бесселя порядка v. Эти два равенства есть преобразование Ханкеля и вместе с равенством Парсеваля они приводят к результату:

$$\int_0^{\infty} w x^{-2} N_v(w/x)\, N_v(w)\, dw = \int_0^{\infty} w\, N_0(wx)\, N_0(w)\, dw = \frac{1}{2\pi(1+x^2)} .$$

Следующий **пример** использует понятие дробной производной.

Пусть $X_2 \in N(0,1)$, а X_1 имеет распределение с плотностью $f_1(x) = \dfrac{1}{\pi\sqrt{1-x^2}}, |x| < 1$, Y и X_1 – независимы. Какое распределение имеет величина $Z = X_2 / X_1$?

Решение. Плотность распределения случайной величины Z в данном случае равна

$$f_Z(y) = \int_{-1}^{1} |x| \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2y^2}{2}} \cdot \frac{1}{\pi\sqrt{1-x^2}} dx = \sqrt{\frac{2}{\pi^3}} \int_0^1 \frac{x}{\sqrt{1-x^2}} e^{-\frac{x^2y^2}{2}} dx =$$

$$= \frac{1}{\sqrt{2\pi^3}} \int_0^1 \frac{1}{\sqrt{1-u}} e^{-\frac{y^2u}{2}} du = \frac{1}{\pi y} e^{-\frac{y^2}{2}} \operatorname{erfi}\left(\frac{y}{\sqrt{2}}\right),$$

где функция $\operatorname{erfi}(x)$ есть $\operatorname{erfi}(x) = \dfrac{2}{\sqrt{\pi}} \int\limits_0^x e^{v^2} dv$.

Отметим, что функция Доусона $F(x) = e^{-x^2}\int\limits_0^x e^{u^2} du$ есть дробная производная от e^{-x}, именно,

$$F(\sqrt{x}) \cdot \frac{2}{\sqrt{\pi}} = D^{-1/2} e^{-x}, \text{ где } D^{-p} f(t) = {}_0D_t^{-p} f(t) = \frac{1}{\Gamma(p)} \frac{d}{dt} \int_0^t \frac{f(u)}{(t-u)^{1-p}} du .$$

2) *Непрямой вывод*. Рассмотрим еще одну случайную величину $Y_2 = X_2$. Тогда $\{\ x_2 = y_2, x_1 = y_1 y_2$ и $J = \det\begin{pmatrix} y_2 & 0 \\ 0 & 1 \end{pmatrix} = y_2$. Вычисляя совместную плотность распределения пары (Y_1, Y_2) и интегрируя по второй переменной, получим формулу (5).

Пусть $\varphi_1(t)$ – характеристическая функция величины X_1 и пусть к тому же $\boldsymbol{P}(X_2 > 0) = 1$.

Тогда (5) преобразуется в формулу $f_{Y_1}(y)=\int\limits_0^{\infty} x\ f_1(yx)\, f_2(x)\,dx$, а характеристическая функция будет вычисляться как $\varphi_{Y_1}(t)=\int_{-\infty}^{\infty}\varphi_1(t/s)\, f_2(s)\,ds$.

3) *Распределение суммы независимых величин* (X_1, X_2). Как и в случае 2 введем преобразование $\{\ y_1=x_1+x_2,\ y_2=x_2,\ \Leftrightarrow\ \{\ x_1=y_1-y_2,\ x_2=y_2$. Якобиан этого преобразования равен 1, поэтому получаем для непрерывных величин соотношение:

$$f_{X_1+X_2}(y)=\int\limits_{-\infty}^{\infty} f_1(y-x)\,f_2(x)\,dx. \qquad (6)$$

Рассмотрим теперь распределение **произведения** двух случайных величин X_1 и X_2 соответственно с плотностями $f_1(x)$ и $f_2(x)$, т.е. пусть $Y=X_1\cdot X_2$. Рассуждая как выше нетрудно получить, что если $g(y)-$плотность распределения величины Y, то

$$\boxed{g(y)=\int\limits_{-\infty}^{+\infty}\frac{1}{|x|}\,f_1\left(\frac{y}{x}\right)f_2(x)\,dx.}$$

Аналогично равенству (5) в случае $Y=X_1\cdot X_2$ получим

$$\varphi_Y(t)=\int_{-\infty}^{\infty}\varphi_1(tx)\, f_2(x)\,dx$$

и если X_2 имеет равномерное распределение на интервале $(0,1)$, то $\varphi_Y(t)=\frac{1}{t}\int_0^t\varphi_1(y)\ dy$.

В случае, когда $\boldsymbol{P}(X_1>0)=1$, мы имеем $g(y)=\int\limits_0^{+\infty}\frac{1}{x}\,f_1\left(\frac{y}{x}\right)f_2(x)\,dx$. Последнее можно интерпретировать следующим образом: пусть $\frac{1}{\sigma}\,f_1\left(\frac{x}{\sigma}\right),\sigma>0-$семейство плотностей с масштабным параметром σ. Тогда $g(y)=\int\limits_0^{+\infty}\frac{1}{\sigma}\,f_1\left(\frac{y}{\sigma}\right)f_2(\sigma)\,d\sigma$ есть смесь распределений (намешанная плотностью $f_2(\sigma)$).

Отметим, что представление в виде произведения случайных величин можно использовать для **нахождения моментов** случайных величин (см. решение задачи 33), где показано, как легко можно найти математическое ожидание и дисперсию плотности

$$f_Z(z;\rho)=\frac{1}{\pi\sqrt{1-\rho^2}}\exp\left(\frac{\rho z}{1-\rho^2}\right)K_0\left(\frac{|z|}{1-\rho^2}\right),-\infty<z<\infty.$$

Прямое нахождение указанных характеристик является довольно затруднительным делом.

Одновершинные распределения

Определение. *Функция распределения U называется* ***одновершинной*** (***унимодальной***) с *модой в нуле, если на интервале* $(-\infty,0)$ *она* *<u>выпукла</u>, а на интервале* $(0,\infty)$ *<u>вогнута</u>.*

Ноль может быть точкой разрыва U, но вне нуля одновершинность предполагает у U плотность, которая монотонна на $(-\infty,0)$ и $(0,\infty)$ (интервалы постоянства исключаются).

Пусть X и $Y-$ независимые случайные величины с распределениями $F(x)$ и $G(y)$ соответственно. Предположим для простоты, что $\boldsymbol{P}(X>0)=1$, т.е. $F(0)=0$.

Рассмотрим произведение $Z=XY$. Имеем:

$$\boldsymbol{P}(Z<z\,|\,X=x)=G\left(\frac{z}{x}\right).$$

Интегрируя последнее равенство по $F(x)$, получим ф. р. $U(z)$ величины Z.

$$U(z)=\int_0^\infty G\left(\frac{z}{x}\right)dF(x).$$

В частности, когда величина X распределена равномерно на интервале $(0,1)$, получаем

$$U(z)=\int_0^\infty G\left(\frac{z}{x}\right)dx. \tag{6}$$

Теорема (критерий Хинчина). *Функция распределения* $U(z)$ *одновершинна тогда и только тогда, когда её можно представить в виде* (6). *Иными словами,* $U(z)$ *одновершинна тогда и только тогда, когда она является функцией распределения произведения* $Z = XY$ *двух независимых случайных величин, одна из которых, скажем* X *, распределена равномерно на* $(0,1)$.

Доказательство. Возьмем $h>0$ и обозначим через $U_h(z)$ функцию распределения, графиком которой является кусочно-линейная функция, совпадающая с $U(z)$ в точках $0,\pm h,...$ (т.е. $U_h(nh)=U(nh)$ и функция $U_h(z)$ линейна в интервалах между nh и $(n+1)h$). Из определения одновершинности непосредственно вытекает, что функция $U(z)$ одновершинна тогда и только тогда, когда одновершинны функции $U_h(z)$. Далее, $U_h(z)$ имеет плотность $u_h(z)$, являющиеся ступенчатыми функциями, а каждую ступенчатую функцию с разрывами в точках nh можно записать в виде

$$\sum p_n \frac{1}{|n|h} f\left(\frac{x}{nh}\right), \tag{7}$$

где $f(x)=1$ при $0<x<1$ и $f(x)=0$ в остальных точках. Функция (7) монотонна в интервалах $(-\infty,0)$ и $(0,\infty)$ тогда и только тогда, когда $p_n \ge 0$ при всех n и является плотностью, если $\sum p_n=1$. Но когда $p_n \ge 0$ и $\sum p_n=1$, (4) представляет плотность распределения произведения $Z_h = XY_h$ – двух независимых случайных величин X и Y_h, из которых X равномерно распределена на $(0,1)$, а Y_h имеет арифметическое распределение $P(Y_h = nh) = p_n$. Мы доказали, таким образом, что функция $U_h(z)$ одновершинна тогда и только тогда, когда ее можно представить в виде (6) с функцией G, замененной на функцию G_h, сосредоточенную в точках $0,\pm h,...$. Устремляя теперь h к нулю и используя теорему о монотонной сходимости, получаем нужный результат.

2. Применения

Задача 1. Пусть ξ_1 и ξ_2 – независимые случайные величины, имеющие распределение с ф.р. $F_0(x)=\frac{1}{2}+\frac{2}{\pi}\operatorname{arctg}(x), -1<x<1$. Как распределена величина $\eta=\frac{\xi_1+\xi_2}{1-\xi_1\xi_2}$?

Решение. Если φ_1,φ_2 – независимые случайные величины, равномерно распределенные на $(-\pi/4,\pi/4)$, то величины $\xi_i=\operatorname{tg}(\varphi_i), i=1,2$, независимы и имеют распределение $F_0(x)$. Но тогда $\eta=\operatorname{tg}(\varphi_1+\varphi_2)$. Носитель распределения величины $\beta=\varphi_1+\varphi_2$ есть отрезок $[-\pi/2,\pi/2]$. Функция распределения величины β равна $F(x)=2(x/\pi+1/2)^2$, если $-\pi/2<x<0$, и $F(x)=1-2(x/\pi-1/2)^2$, если $0<x<\pi/2$. Поскольку $\operatorname{arctg}(x)$ на $[-\pi/2,\pi/2]$ монотонна, то функция распределения с.в. η есть $F_\eta(x)=2((1/\pi)\operatorname{arctg}(x)+1/2)^2$, если $-\infty<x<0$, и $F(x)=1-2((1/\pi)\operatorname{arctg}(x)-1/2)^2$, если $0<x<\infty$.

Задача 2. Пусть $\boldsymbol{X} = (X_1, X_2), \boldsymbol{Y} = (Y_1, Y_2)$ – независимые и одинаково распределенные $N(\mathbf{0}, \Sigma)$, где $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ – ковариационная матрица $(-1 < \rho < 1)$. Докажите, что с.в. $Z = \beta_2 \dfrac{X_2}{Y_2} + \beta_1 \dfrac{X_1}{Y_1}, \beta_1 \geq 0, \beta_2 \geq 0, \beta_1 + \beta_2 = 1$ имеет распределение Коши C(0,1). Рассмотреть случай когда β_1 и β_2 неслучайны, а также, когда они случайны.

Решение. Векторы $\boldsymbol{X} = (X_1, X_2)^T, \boldsymbol{Y} = (Y_1, Y_2)^T$ – независимы и одинаково распределены, дисперсии величин равны 1, а коэффициент корреляции равен ρ. Рассмотрим величину $Z = \beta_2 \dfrac{X_2}{Y_2} + \beta_1 \dfrac{X_1}{Y_1}$. Обозначим $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Тогда матрица корреляций вектора $(X_1, X_2, Y_1, Y_2)^T$ примет вид:

$$R = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix} = A \otimes B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \Rightarrow \det(R) = \det(A \otimes B) = (\det(A))^2 (\det(B))^2 = (1-\rho^2)^2,$$

$$R^{-1} = (A \otimes B)^{-1} = A^{-1} \otimes B^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & -\rho \\ -\rho & 0 \end{pmatrix} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & 0 \\ -\rho & 1 & 0 & 0 \\ 0 & 0 & 1 & -\rho \\ 0 & 0 & -\rho & 1 \end{pmatrix}.$$

Здесь $A \otimes B$ есть Кронекерово произведение матриц. Совместная плотность $(X_1, X_2, Y_1, Y_2)^T$ равна

$$f(x_1, x_2, y_1, y_2) = p(x_1, x_2) \cdot p(y_1, y_2), \text{ где } \quad p(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)} \right),$$

а характеристическая функция величины Z равна

$$\varphi_Z(t) = \mathbf{E}(e^{it\left(\beta_1 \frac{X_1}{Y_1} + \beta_2 \frac{X_2}{Y_2}\right)}) =$$

$$= \iint_{R^4} e^{it\beta_1 \frac{x_1}{y_1} + it\beta_2 \frac{x_2}{y_2}} \frac{1}{4\pi^2(1-\rho^2)} \exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2 + y_1^2 - 2\rho y_1 y_2 + y_2^2}{2(1-\rho^2)} \right) dx_1 dx_2 dy_1 dy_2 =$$

$$= \frac{1}{2\pi} \iint_{R^2} e^{it\beta_1 \frac{x_1}{y_1}} \mathbf{E}(e^{it\beta_2 \frac{X_2}{Y_2}} \mid X_1 = x_1, Y_1 = y_1) e^{-\frac{x_1^2 + y_1^2}{2}} dx_1 dy_1.$$

Известно: условное распределение величины X_2 при условии, что $X_1 = x_1$ – нормальное с плотностью $\dfrac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(x_2 - \rho x_1)^2}{2(1-\rho^2)}}$, (условное распределение величины Y_2 при условии, что $Y_1 = y_1$ – нормальное с плотностью $\dfrac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(y_2 - \rho y_1)^2}{2(1-\rho^2)}}$) и при этом $\mathbf{E}((X_2 - \rho x_1)(Y_2 - \rho y_1) \mid X_1 = x_1, Y_1 = y_1) = 0$.

Утверждение. *Пусть* $\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \in N\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \alpha\sigma_1\sigma_2 \\ \alpha\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), -1 < \alpha < 1,$ *и* $\varsigma = \dfrac{\xi_1}{\xi_2}$. *Тогда плотность распределения случайной величины* ς *равна*

$$f_\varsigma(z)=\frac{\sqrt{1-\alpha^2}}{\pi\sigma_1\sigma_2\gamma^2(z)}\exp\left(-\frac{\tau^2}{2(1-\alpha^2)}\right)+\frac{\lambda(z)\eta(z)}{\sqrt{2\pi}\sigma_1\sigma_2\gamma^3(z)}\operatorname{erf}\left(\frac{\eta(z)}{\gamma(z)\sqrt{2(1-\alpha^2)}}\right),$$

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}du,\quad \gamma^2(z)=\frac{z^2}{\sigma_1^2}-\frac{2\alpha z}{\sigma_1\sigma_2}+\frac{1}{\sigma_2^2},\quad \eta(z)=\frac{a_1 z}{\sigma_1^2}-\frac{\alpha(a_1+a_2 z)}{\sigma_1\sigma_2}+\frac{a_2}{\sigma_2^2},$$

$$\lambda(z)=\exp\left(\frac{1}{2(1-\alpha^2)}\left(\frac{\eta^2(z)}{\gamma^2(z)}-\tau^2\right)\right),\quad \tau^2=\frac{a_1^2}{\sigma_1^2}-\frac{2\alpha a_1 a_2}{\sigma_1\sigma_2}+\frac{a_2^2}{\sigma_2^2}.$$

В нашем случае $a_1=\rho x_1, a_2=\rho y_1, \alpha=0, \sigma_1^2=\sigma_2^2=1-\rho^2,$ поэтому

$$\gamma^2(z)=\frac{1+z^2}{1-\rho^2},\ \eta(z)=\frac{\rho(zx_1+y_1)}{1-\rho^2},\ \tau^2=\frac{\rho^2(x_1^2+y_1^2)}{1-\rho^2},\ \lambda(z)=\exp\left(-\frac{\rho^2(zy_1-x_1)^2}{2(1-\rho^2)(1+z^2)}\right),$$

$$\nu(z)=\frac{\eta(z)}{\gamma(z)}=\frac{\rho}{\sqrt{1-\rho^2}}\cdot\frac{zx_1+y_1}{\sqrt{1+z^2}},$$

$$f_\varsigma(z)=\frac{1}{\pi(1+z^2)}\exp\left(-\frac{\rho^2(x_1^2+y_1^2)}{2(1-\rho^2)}\right)+\frac{\nu(z)\exp\left(\frac{\nu^2(z)}{2}-\frac{\rho^2(x_1^2+y_1^2)}{2(1-\rho^2)}\right)}{\sqrt{2\pi}(1+z^2)}\operatorname{erf}\left(\frac{\nu(z)}{\sqrt{2}}\right)$$

Тогда

$$\frac{1}{2\pi}f_\varsigma(z)e^{-\frac{x_1^2+y_1^2}{2}}=\frac{1}{2\pi^2(1+z^2)}\exp\left(-\frac{(x_1^2+y_1^2)}{2(1-\rho^2)}\right)+\frac{\nu(z)\exp\left(\frac{\nu^2(z)}{2}-\frac{(x_1^2+y_1^2)}{2(1-\rho^2)}\right)}{(2\pi)^{3/2}(1+z^2)}\operatorname{erf}\left(\frac{\nu(z)}{\sqrt{2}}\right),$$

$$1=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f_\varsigma(z)\frac{1}{2\pi}e^{-\frac{x_1^2+y_1^2}{2}}dz\,dx_1dx_2=I_1+I_2,$$

$$I_1=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{1}{\pi(1+z^2)}\frac{1}{2\pi}\exp\left(-\frac{(x_1^2+y_1^2)}{2(1-\rho^2)}\right)dz\,dx_1dx_2\varphi=\int_{-\infty}^{+\infty}\frac{dz}{\pi(1+z^2)}\int_0^{+\infty}\int_0^{2\pi}\frac{R}{2\pi}\exp\left(-\frac{R^2}{2(1-\rho^2)}\right)dR\,d\varphi=1-\rho^2,$$

откуда $I_2=\rho^2$. Введем дополнительный угол δ так, что $\cos\delta=\dfrac{z}{\sqrt{1+z^2}}, \sin\delta=\dfrac{1}{\sqrt{1+z^2}}$.

Сделаем замену переменных: $x_1=R\cos\varphi, y_1=R\sin\varphi,$ откуда

$$\nu(z)=\frac{\rho(R\cos\varphi\cdot\cos\delta+R\sin\varphi\cdot\sin\delta)}{\sqrt{1-\rho^2}}=\frac{\rho R\cos(\varphi_1)}{\sqrt{1-\rho^2}},\ \varphi_1=\varphi-\delta \text{ и}$$

$\dfrac{x_1}{y_1}=\operatorname{ctg}(\varphi)=\operatorname{ctg}(\varphi_1+\delta)=\dfrac{z\cos\varphi_1-\sin\varphi_1}{z\sin\varphi_1+\cos\varphi_1}$. Сделаем еще замену $w=\dfrac{z\cos\varphi-\sin\varphi}{z\sin\varphi+\cos\varphi}$, откуда

$z=\dfrac{w\cdot\cos\varphi+\sin\varphi}{\cos\varphi-w\cdot\sin\varphi}$ и $J=\left|\dfrac{dz}{dw}\right|=\dfrac{1}{(\cos\varphi-w\sin\varphi)^2}, \dfrac{1}{1+z^2}=\dfrac{(\cos\varphi-w\sin\varphi)^2}{1+w^2}, \dfrac{1}{1+z^2}\cdot J=\dfrac{1}{1+w^2}=\pi\cdot f_W(w).$

Но плотность $f_W(w)$ не зависит от φ, она есть плотность распределения Коши, а интеграл от функции $\nu(z)\exp\left(\dfrac{\nu^2(z)}{2}\right)\operatorname{erf}\left(\dfrac{\nu(z)}{\sqrt{2}}\right)$ это интеграл по отрезку длины 2π. Собирая все интегралы, получаем, что $\varphi_Z(t)=(1-\rho^2)e^{-\beta_1|t|-\beta_2|t|}+\rho^2e^{-\beta_1|t|-\beta_2|t|}=e^{-|t|}$, т.е. характеристическая функция распределения Коши $C(0,1)$. Если β_1,β_2 – случайны, то сначала зафиксируем их и поскольку вывод не зависит от их конкретных значений, то получим результат утверждения.

Оценки Питмена. Методы построения вероятностных моделей (преобразования случайных величин) мы применим к нахождению оценок Питмена.

Пусть задано семейство распределений $\mathscr{F}=\{F(x;\theta),\theta\in\Theta\subset\boldsymbol{R}\}$ и имеется выборка $X_1,\ldots,X_n$ из распределения $F(x;\theta)$.

Определение 1. *Оценкой параметра θ будем называть измеримую функцию от наблюдений $T=T_n(X_1,\ldots,X_n)$, значение которой принимается в качестве «истинного» значения неизвестного параметра.*

Рассмотрим следующее семейство распределений $\mathscr{F}=\{F(x;\theta)=F(x-\theta),\theta\in\boldsymbol{R}\}$.

Определение 2. *Статистику(оценку) $T=T_n(X_1,\ldots,X_n)$ будем называть правильной статистикой, если при всех $c\in\boldsymbol{R}$ она удовлетворяет условию*

$$T_n=T_n(x_1+c,\ldots,x_n+c)=T_n(x_1,\ldots,x_n)+c.$$

Пример 0. Статистики $\sum_{i=1}^{n}a_iX_n^{(i)},\ a_i\ge 0,\ i=1,2,\ldots,n,\ \sum_{j=1}^{n}a_j=1,$ является правильными статистиками, в частности, правильными статистиками являются $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ и $(X_n^{(1)}+X_n^{(n)})/2$.

Класс правильных оценок обозначим через $\mathscr{E}$. Если T_n' и T_n'' – две правильные оценки, то $T_n'-T_n''=\psi(x_2-x_1,\ldots,x_n-x_1)$. Нетрудно показать, что для параметра сдвига θ риск $\boldsymbol{E}_\theta(T_n-\theta)^2$ правильной оценки T_n не зависит от θ и поэтому ожидание будем рассматривать при $\theta=0$.

Определим вектор $\boldsymbol{y}=(x_2-x_1,\ldots,x_n-x_1)$ и рассмотрим оценку, которую ввел Е. Питмен (E.J. Pitman, "The estimation of the location and scale parameters of a continuous population of any given form" *Biometrika* , **30** (1939) pp. 391–421)

$$t_n=\bar{x}-\boldsymbol{E}_0(\bar{x}\mid\boldsymbol{y})=\bar{x}-\boldsymbol{E}(\bar{x}\mid\boldsymbol{y}).$$

Во-первых, можно показать, что оценка t_n является оптимальной в классе $\mathscr{E}$, а, во-вторых, если $F(x-\theta)$ абсолютно непрерывна и $f(x-\theta)$ ее плотность по мере Лебега, то

$$t_n=\frac{\int_{-\infty}^{+\infty}\xi\cdot\Pi_{i=1}^{n}f(x_i-\xi)\,d\xi}{\int_{-\infty}^{+\infty}\Pi_{i=1}^{n}f(x_i-\xi)\,d\xi}.$$

Действительно, заметим сначала, что

$$t_n=\bar{x}-\boldsymbol{E}(\bar{x}\mid\boldsymbol{y})=\bar{x}-\boldsymbol{E}(x_1+\frac{(x_2-x_1)+\ldots(x_2-x_1)}{n}\mid\boldsymbol{y})=$$
$$=\bar{x}-\boldsymbol{E}(x_1\mid\boldsymbol{y})+\frac{(x_2-x_1)+\ldots(x_2-x_1)}{n}=x_1-\boldsymbol{E}(x_1\mid\boldsymbol{y}).$$

Перейдем к переменным

$$\left\{\begin{array}{l} x_1\quad\ =u_1,\\ x_2-x_1=u_2,\\ \quad\cdot\ \cdot\ \cdot\\ x_n-x_1=u_n,\end{array}\right.\quad\Rightarrow\quad\left\{\begin{array}{l} x_1=u_1,\\ x_2=u_2+u_1,\\ \quad\cdot\ \cdot\ \cdot\\ x_n=u_n+u_1.\end{array}\right.$$

Якобиан преобразования равен 1, поэтому совместная плотность с.в. $(U_1,U_2,\ldots,U_n)$ равна

$$p(u_1,u_2,\ldots,u_n)=f(u_1)f(u_2+u_1)\ldots f(u_n+u_1).$$

Отсюда $\boldsymbol{E}(u_1\mid u_2,\ldots,u_n)=\dfrac{\int u_1f(u_1)f(u_2+u_1)\ldots f(u_n+u_1)\,du_1}{\int f(u_1)f(u_2+u_1)\ldots f(u_n+u_1)\,du_1}.$

Возвращаясь к старым переменным и обозначая $\xi = -u_1 + x_1 \ \Rightarrow\ u_1 = x_1 - \xi$, будем иметь

$$\boldsymbol{E}(x_1 \mid x_2 - x_1, \dots, x_n - x_1) = \frac{\int u_1 f(u_1) f(u_1 + x_2 - x_1) \dots f(u_1 + x_n - x_1)\, du_1}{\int f(u_1) f(u_1 + x_2 - x_1) \dots f(u_1 + x_n - x_1)\, du_1} =$$

$$= \frac{\int_{-\infty}^{\infty} (x_1 - \xi) \prod_{i=1}^{n} f(x_i - \xi)\, d\xi}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f(x_i - \xi)\, d\xi} = x_1 - \frac{\int_{-\infty}^{\infty} \xi \prod_{i=1}^{n} f(x_i - \xi)\, d\xi}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f(x_i - \xi)\, d\xi},$$

откуда получаем результат.

Пример 1. Пусть с.в. X имеет плотность $f(x-\theta) = \begin{cases} \exp(-(x-\theta)), & \theta < x < \infty, \\ 0, & \theta \ge x. \end{cases}$

Найти оценку Питмена параметра θ по выборке $x_1, \dots, x_n$.

Решение. Плотность не равна нулю, если $0 < x_i - \xi$ для всех $i = 1, 2, \dots, n$, откуда получаем, что плотность не зануляется для $\xi < x_n^{(1)}$, поэтому

$$t_n = \frac{\int_{-\infty}^{x_n^{(1)}} \xi \cdot \Pi_{i=1}^{n} \exp(-(x_i - \xi))\, d\xi}{\int_{-\infty}^{x_n^{(1)}} \Pi_{i=1}^{n} \exp(-(x_i - \xi))\, d\xi} = \frac{\int_{-\infty}^{x_n^{(1)}} \xi \cdot \exp(n\xi)\, d\xi}{\int_{-\infty}^{x_n^{(1)}} \exp(n\xi)\, d\xi} = x_n^{(1)} - \frac{1}{n}, \ \boldsymbol{D}(t_n) = \frac{1}{n^2}.$$

В стандартной схеме линейной регрессии (схема Гаусса-Маркова) наблюдения $x_1, x_2, \dots, x_n$ имеют вид $x_i = \sum_{j=1}^{m} a_{ij}\theta_j + \varepsilon_i,\ i = 1, 2, \dots, n,$ где случайные величины $\varepsilon_1, \dots, \varepsilon_n$ предполагаются н.о.р.с.в. с ожиданием 0, конечной дисперсией и плотностью $f(x)$. Векторный параметр $(\theta_1, \dots, \theta_2)^T$ предполагается неизвестным из $\mathbf{R}^m$, матрица $A = (a_{ij})_{n\times m}, \sum_{j=1}^{m} a_{ij}^2 \ne 0, i = 1, \dots, n,$ образована известными постоянными, ранг A равен $m \le n$. Здесь преобразование оценки $\tilde{\theta}(X)$ есть

$$\tilde{\theta}(X + AC) = \tilde{\theta}(X) + C,\ C \in \mathbf{R}^m.$$

Простейший пример – оценка МНК, равная $t_n = (A^T A)^{-1} A^T X$. Положим $B = I - A(A^T A)^{-1} A^T$ и $Y = BX$. Тогда оценка $\hat{\theta}_n = t_n - \boldsymbol{E}_0(t_n \mid Y)$ является оптимальной в смысле потерь $\boldsymbol{E}_\theta((\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T)$ и имеет вид

$$\hat{\theta}_n = \left(\iint_{\mathbf{R}^m} u_k \prod_{i=1}^{n} f(x_i - \sum_{j=1}^{m} a_{ij} u_j) du_1 \dots du_m \right) \left(\iint_{\mathbf{R}^m} \prod_{i=1}^{n} f(x_i - \sum_{j=1}^{m} a_{ij} u_j) du_1 \dots du_m \right)^{-1}.$$

Рассмотрим еще одно применение полученных результатов.

Рассмотрим выражение $ns^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$. Это есть квадратичная форма с матрицей

$$A_n = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ . & . & . & \dots & . \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{pmatrix},$$

у которой на диагонали стоят числа $1 - \frac{1}{n}$, а внедиагональные элементы равны $-\frac{1}{n}$.

Составим характеристическое уравнение для матрицы A_n: $P_n(\lambda) = |A_n - \lambda E_n| = 0,$ где E_n – единичная матрица. Покажем, что $P_n(\lambda) = -\lambda(1-\lambda)^{n-1}.$ Пусть $n = 2.$ Тогда

$$P_2(\lambda)=\begin{vmatrix}\frac{1}{2}-\lambda & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}-\lambda\end{vmatrix}=\left(\frac{1}{2}-\lambda\right)^2-\frac{1}{4}=\lambda(\lambda-1).$$

Имеем:

$$P_n(\lambda)=|A_n-\lambda E_n|=\begin{vmatrix}1-\lambda-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n}\\ -\frac{1}{n} & 1-\lambda-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n}\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\lambda-\frac{1}{n}\end{vmatrix}=\begin{vmatrix}1-\lambda & 0 & 0 & \cdots & 0\\ -\frac{1}{n-1} & 1-\lambda-\frac{1}{n-1} & -\frac{1}{n-1} & \cdots & -\frac{1}{n-1}\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ -\frac{1}{n-1} & -\frac{1}{n-1} & -\frac{1}{n-1} & \cdots & 1-\lambda-\frac{1}{n-1}\end{vmatrix}+$$

$$+\begin{vmatrix}-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n}\\ -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & \cdots & -\frac{1}{n(n-1)}\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & -\frac{1}{n(n-1)} & \cdots & -\frac{1}{n(n-1)}\end{vmatrix}=(1-\lambda)P_{n-1}(\lambda),$$

так как последний определитель равен 0, а первый разложен по первой строке.
Теперь возьмем вектор $\boldsymbol{e}_1^T=(1,1,...,1)$ размерности n. Здесь $A_n\cdot\boldsymbol{e}_1^T=(0,0,...,0)$, т.е. $\boldsymbol{e}_1^T$ есть собственный вектор для собственного числа $\lambda=0$. Векторы $\boldsymbol{e}_2^T=(-(n-1),1,...,1)$, $\boldsymbol{e}_3^T=(0,-(n-2),1,...,1),...,$ $\boldsymbol{e}_n^T=(0,0,...,0,-1,1)$ ортогональны и являются собственными векторами для собственного числа $\lambda=1$ кратности $n-1$.

Рассмотрим векторы
$\boldsymbol{b}_1^T=(1/\sqrt{n},1/\sqrt{n},...,1/\sqrt{n})$, $\boldsymbol{e}_2^T=(-\sqrt{(n-1)/n},1/\sqrt{(n-1)n},...,1/\sqrt{(n-1)n})$,
$\boldsymbol{e}_3^T=(0,-\sqrt{(n-2)/(n-1)},1/\sqrt{(n-2)(n-1)},...,1/\sqrt{(n-2)(n-1)})$, $\boldsymbol{e}_n^T=(0,0,...,0,-1/\sqrt{2},1/\sqrt{2})$,
которые имеют длину 1 и пусть с.в. $x_i, i=1,2,...,n,$ таковы, что $\mathbf{E}(x_i)=0$, $\mathbf{D}(x_i)=1$, $\mathbf{E}(x_ix_j)=0,\ i\neq j,\ \boldsymbol{x}^T=(x_1,x_2,...,x_n)$ и имеют совместное нормальное распределение.

Во-первых, из соотношения $\mathbf{E}(\overline{x}(x_i-\overline{x})=\frac{1}{n}\mathbf{E}((x_1+...x_n)x_i)-\mathbf{E}(\overline{x}^2)=\frac{1}{n}\mathbf{E}(x_i^2)-\frac{1}{n}=0.$

Отсюда следует, что $\overline{x}$ и $(x_i-\overline{x})-$независимы для каждого $i=1,2,...,n,$ а, значит, независимы и $\overline{x}$ и s^2. Далее, $\xi_i=\boldsymbol{b}_i^T\boldsymbol{x}, i=1,2,...,n,$ являются независимыми случайными стандартно нормально распределенными величинами. Поэтому (см. [31], с.262) $ns^2=\boldsymbol{x}^TA_n\boldsymbol{x}=\sum_{i=1}^{n}\lambda_i\xi_i^2=\sum_{i=1}^{n-1}\xi_i^2$ и поэтому имеет хи-квадрат распределение с $(n-1)-$й степенью свободы. Тем самым, квадратичную форму ns^2 мы привели к главным осям.

Рассмотрим теперь плотность распределения $\chi^2(n)$ с n степенями свободы:
$p_n(x)=\dfrac{x^{n/2-1}\exp(-x/2)}{2^{n/2}\Gamma(n/2)}$, $x>0$, и пусть теперь $\eta=\sqrt{\chi^2(n)}/n$, а ξ пусть имеет плотность $f_\xi(x)=\sqrt{\dfrac{n}{2\pi}}\exp\left(-\dfrac{nx^2}{2}\right)$. Плотность распределения величины η равна

$f_\eta(y) = \dfrac{\sqrt{2n}}{\Gamma(n/2)} \left(\dfrac{y\sqrt{n}}{\sqrt{2}} \right)^{n-1} \exp\left(-\dfrac{ny^2}{2} \right)$, $y > 0$. Найдем плотность отношения независимых величин: $z = \dfrac{\xi}{\eta}$. Имеем: $f_z(x) = \int\limits_0^{+\infty} z \sqrt{\dfrac{n}{2\pi}} \exp\left(-n\dfrac{z^2x^2}{2} \right) \dfrac{\sqrt{2n}}{\Gamma(n/2)} \left(-\dfrac{z\sqrt{n}}{2} \right)^{n-1} \exp\left(-\dfrac{nz^2}{2} \right) dz =$

$$= \frac{1}{\sqrt{n}\,\Gamma(n/2)} \int\limits_0^{+\infty} \left(\frac{z\sqrt{n}}{\sqrt{2}} \right)^{n-1} \exp\left(-\frac{nz^2(x^2+1)}{2} \right) dz.$$

Сделаем замену $u = nz^2(x^2+1)/2$. Находим

$$f_z(x) = \frac{(x^2+1)^{-(n+1)/2}}{\sqrt{n}\,\Gamma(n/2)} \int\limits_0^{+\infty} u^{(n-1)/2} e^{-u}\, du = \frac{\Gamma\left(\frac{n+1}{2} \right)}{\sqrt{n}\,\Gamma(n/2)} (x^2+1)^{-(n+1)/2}.$$

Эта плотность носит название *закона Стьюдента*.

3. Построение модели и оценка ее параметров

Пусть X – равномерно распределенная на множестве $\{1, 2, 3, \dots, \theta\}$ случайная величина, параметр θ неизвестен. Наблюдается выборка (выбор с возвращением) $x_1, x_2, \dots, x_n$ и пусть $x_n^{(n)} = \max\limits_{1 \le i \le n} x_i = k$. В качестве оценки параметра θ возьмем $t_n = \dfrac{k^{n+1} - (k-1)^{n+1}}{k^n - (k-1)^n}$. Покажем, что эта оценка является несмещенной. Найдем сначала распределение величины $Y = X_n^{(n)}$. Имеем: $\mathbf{P}(X_n^{(n)} \le k) = \mathbf{P}(\max\limits_i X_i \le k) = \mathbf{P}^n(X_1 \le k) = \left(\dfrac{k}{\theta} \right)^n$. Значит, $\mathbf{P}(X_n^{(n)} = k) = \mathbf{P}(X_n^{(n)} \le k) - \mathbf{P}(X_n^{(n)} \le k-1) = \dfrac{k^n - (k-1)^n}{\theta^n}$. Отсюда,

$$\mathbf{E}(t_n) = \sum_{k=1}^{\theta} \frac{k^{n+1} - (k-1)^{n+1}}{k^n - (k-1)^n} \cdot \frac{k^n - (k-1)^n}{\theta^n} = \frac{1}{\theta^n} \sum_{k=1}^{\theta} (k^n - (k-1)^n) = \frac{\theta^{n+1}}{\theta^n} = \theta,$$

т.е. t_n – несмещенная оценка. Ясно, что оценка $t_n^{(r)} = \dfrac{k^{n+r} - (k-1)^{n+r}}{k^n - (k-1)^n}$ будет несмещенной оценкой функции θ^r. Заметим также, что для достаточно больших θ и для достаточно больших объемов выборки $x_n^{(n)} = k -$ достаточно большое, имеем

$t_n = \dfrac{k^{n+1} - (k-1)^{n+1}}{k^n - (k-1)^n} = \dfrac{k^{n+1}(1 - (1 - 1/k)^{n+1})}{k^n(1 - (1 - 1/k)^n)} \approx k\dfrac{n+1}{n}$, поэтому для оценки параметра θ можно использовать статистику $\hat{t}_n = \dfrac{n+1}{n} X_n^{(n)}$. Пусть $n = 10$ и наблюдается максимальное значение $x_{10}^{(10)} = 910$. Тогда $t_n = \dfrac{910^{11} - 909^{11}}{910^{10} - 909^{10}} = 1000.45$, а $\hat{t}_n = \dfrac{11}{10} \cdot 910 = 1001$.

Другой простой пример состоит в следующем: пусть вы выходите на платформу метро и наблюдаете номер подошедшего поезда и этот номер равен 5 (будем считать, что номера поездов в депо идут без пропусков: 1, 2, …, θ). Тогда $n = 1$ и $t_n = \dfrac{5^2 - 4^2}{5 - 4} = 9$, т.е. число поездов мы оцениваем числом 9. Предположим, что мы наблюдаем еще один номер поезда (это может быть в другой день (тогда можно считать, что выборка повторная) и он по-прежнему равен 5. В таком случае оценка равна $t_n = \dfrac{5^3 - 4^3}{5^2 - 4^2} = \dfrac{61}{9} = 6.78$, т.е. примем это число равным 7.

Замечание. Для оценки количества поездов метро можно считать, что выборка повторная, но для оценки объема военной продукции, ясно, что выборка бесповторная. Насколько сильно будут отличаться результаты оценивания?

Для ответа на этот вопрос рассмотрим дискретное равномерное распределение, но теперь будем считать выбор *без возвращения*. Эта ситуация больше подходит для практических моделей.

Итак, $f(x;\theta) = \dfrac{1}{\theta},\ x = 1, 2, \ldots, \theta.$ Пусть $Y = X_n^{(n)},\ n \le \theta$. Тогда

$$G(y) = \mathbf{P}(Y \le y) = \mathbf{P}(X_1 \le y, X_2 \le y, \ldots, X_n \le y) = \frac{y(y-1)\cdot\ldots\cdot(y-n+1)}{\theta(\theta-1)\cdot\ldots\cdot(\theta-n+1)} = \frac{C_y^n}{C_\theta^n}.$$

Действительно, $\mathbf{P}(X_1 \le y) = \dfrac{y}{\theta}, \mathbf{P}(X_1 \le y, X_2 \le y) = \sum\limits_{k=1}^{y} \mathbf{P}(X_2 \le y, X_1 = k)\mathbf{P}(X_1 = k) = \dfrac{y-1}{\theta-1}\cdot\dfrac{y}{\theta}$,

поскольку один номер уже выбрали, и т.д. Отсюда следует, что

$$g(y) = \mathbf{P}(Y = y) = G(y) - G(y-1) = \frac{C_y^n - C_{y-1}^n}{C_\theta^n} = \frac{C_{y-1}^{n-1}}{C_\theta^n}, \text{ для } y \ge n+1,$$

и если $y = n$ (выбираются все номера 1, 2, …, n), то $g(y) = \dfrac{1}{C_\theta^n} \Rightarrow g(y) = \dfrac{C_{y-1}^{n-1}}{C_\theta^n}, n \le y \le \theta.$

Рассмотрим оценку $t_n = \dfrac{n+1}{n}Y - 1$. Тогда

$$S(n,\theta) = \mathbf{E}(t_n) = \sum_{y=n}^{\theta} \frac{(n+1)y}{n} g(y) - 1 = \sum_{y=n}^{\theta} \frac{(n+1)y}{nC_\theta^n} \frac{(y-1)!}{(n-1)!(y-n)!} - 1 = (n+1)\sum_{y=n}^{\theta} \frac{C_y^n}{C_\theta^n} - 1.$$

Обозначим $\theta = n + k$ и покажем, что $S(n,\theta) = \theta$, т.е. оценка *несмещенная*, т.е. надо показать, что

$$(n+1)\sum_{y=n}^{n+k} \frac{C_y^n}{C_\theta^n} = n + k + 1.$$

Доказывать будем по индукции. Нам надо показать, что $(n+1)\dfrac{C_n^n + C_{n+1}^n + \ldots + C_{n+k}^n}{C_{n+k}^n} = n + k + 1$.

Если $k = 0$, то $(n+1)\dfrac{C_n^n}{C_n^n} = n + 1 -$ верно. Если $k = 1$, то

$$(n+1)\frac{C_n^n + C_{n+1}^n}{C_{n+1}^n} = n + 2 \Leftrightarrow (n+1)\frac{1 + n + 1}{n+1} = n + 2 - \text{ верно.}$$

Пусть равенство (11) верно для k, т.е. $(n+1)\dfrac{C_n^n + C_{n+1}^n + \ldots + C_{n+k}^n}{C_{n+k}^n} = n + k + 1$, и покажем, что

$(n+1)\dfrac{C_n^n + C_{n+1}^n + \ldots + C_{n+k}^n + C_{n+k+1}^n}{C_{n+k+1}^n} = n + k + 2$. Но последнее равносильно

$(n+k+1)\dfrac{C_{n+k}^{n}}{C_{n+k+1}^{n}}+(n+1)\dfrac{C_{n+k}^{n}}{C_{n+k+1}^{n}}=n+k+2 \Leftrightarrow (n+k+1)\dfrac{k+1}{n+k+1}+n+1=k+1+n+1=n+k+2.$ (ил

и можно воспользоваться соотношением $C_{\theta+1}^{n+1}=\sum_{y=n}^{\theta}C_{y}^{n},\ \theta\ge n$).

Для примера Феллера $y=910, n=10$ получаем $t_n=\dfrac{11}{10}\cdot 910-1=1000$.

Найдем дисперсию нашей оценки. Из равенства $\mathbf{E}\left(\dfrac{n+1}{n}Y-1\right)=\theta$ следует, что $\mathbf{E}(Y)=\dfrac{n(\theta+1)}{n+1}$. Тогда $\mathbf{D}(t_n)=\mathbf{D}\left(\dfrac{n+1}{n}Y-1\right)=\mathbf{D}\left(\dfrac{n+1}{n}Y\right)=\dfrac{(n+1)^2}{n^2}\mathbf{E}(Y)-(\theta+1)^2$. Теперь

$$\mathbf{E}(Y^2)=\sum_{y=n}^{\theta}y^2\frac{C_{y-1}^{n-1}}{C_{\theta}^{n}}=\frac{1}{C_{\theta}^{n}}\sum_{y=n}^{\theta}y^2C_{y-1}^{n-1}=\frac{n}{C_{\theta}^{n}}\sum_{y=n}^{\theta}yC_{y}^{n}=\frac{n}{C_{\theta}^{n}}\sum_{y=n}^{\theta}(y+1-1)C_{y}^{n}=$$

$$=\frac{n(n+1)}{C_{\theta}^{n}}\sum_{y=n}^{\theta}C_{y+1}^{n+1}-\frac{n}{C_{\theta}^{n}}\sum_{y=n}^{\theta}C_{y}^{n}=\frac{n(n+1)}{C_{\theta}^{n}}C_{\theta+2}^{n+2}-\frac{n}{C_{\theta}^{n}}C_{\theta+1}^{n+1}=\frac{n(\theta+2)(\theta+1)}{n+2}-\frac{n(\theta+1)}{n}.$$

Значит, $b^2=\mathbf{D}(t_n)=\dfrac{(n+1)^2(\theta+2)(\theta+1)}{n(n+2)}-\dfrac{(n+1)(\theta+1)}{n}=\dfrac{(\theta+2)(\theta+1)}{n(n+2)}-\dfrac{(\theta+1)}{n}$. (12)

Поскольку $\theta\ge n$, то $\mathbf{D}(t_n)\ge 0$ (равенство нулю при $n=\theta$). Для повторной выборки дисперсия будет больше подсчитанной, поэтому мы ограничимся (12). При $\theta=1000, n=10$ получаем, $b^2=8258.25$, $\theta=1000, n=100$ получаем $b^2=19.8$, $\theta=1000, n=500$ получаем . $b^2=2.0$, т.е. грубо можно считать, что при $n=100$ погрешность оценивания не превосходит 13, а при $n=500-$ **четырех.**

4. Конечные и бесконечные смеси распределений. Отрицательное биномиальное распределение

Пример 2. В отдел технического контроля (ОТК) поступают партии изделий, составленные с помощью случайного извлечения из объединенной продукции двух станков (станка A и станка B). Изделия контролируются по некоторому параметру (линейному размеру) ξ мм, так что результатом контроля i-го изделия партии является число x_i мм. Изделия на станках не маркируются, так что в ОТК не известно, на каком именно станке произведено каждое из них. Производительность станка A в 1,5 раза выше производительность станка B. Задано номинальное значение контролируемого параметра a = 65 мм и известно, что точность работы станков характеризуется одинаковой величиной среднеквадратических отклонений $\sigma_A=\sqrt{\mathbf{D}(\xi_A)}$ и $\sigma_B=\sqrt{\mathbf{D}(\xi_B)}$, равной 1,0 мм. Позже выяснилось, что станок A был настроен правильно (производил изделие со средним значением $\mathbf{E}(\xi_A)=65$ мм, равным номиналу), в то время как настройка станка B была сбита в направлении завышения номинала (а именно, $\mathbf{E}(\xi_A)=67$ мм).

Известно также, что распределение размеров изделий, произведенных на каком-то определенном станке, описывается нормальным законом с параметрами $\mathbf{E}(\xi_\gamma)=a_\gamma$ и $\mathbf{D}(\xi_\gamma)=\sigma_\lambda^2$ ($\gamma=A$ или $\gamma=B$). Очевидно, анализируемая в ОТК по наблюдениям генеральная совокупность будет состоять из *смеси двух нормальных совокупностей*, одна из которых представляет продукцию станка A и описывается плотностью $f_A(x)=\varphi(x;a_A,\sigma_A^2)=\frac{1}{\sqrt{2\pi\sigma_A^2}}\exp\left(-\frac{(x-a_A)^2}{2\sigma_A^2}\right)$, а другая – продукцию станка B и описывается плотностью

$$f_B(x)=\varphi(x;a_B,\sigma_B^2)=\frac{1}{\sqrt{2\pi\sigma_A^2}}\exp\left(-\frac{(x-a_A)^2}{2\sigma_A^2}\right),.$$

Обозначим $\theta_\gamma=(a_\gamma,\sigma_\gamma^2)$, а удельный вес изделий станка γ через π_γ $(\gamma=A,B)$, можем записать уравнение функции плотности $f(x)$, описывающей закон распределения анализируемого признака ξ во *всей* (*объединенной*) генеральной совокупности, в виде:

$$f(x)=\pi_A\varphi(x,\theta_A)+\pi_B\varphi(x,\theta_B),\ \ \pi_A+\pi_B=1\ .$$

Учитывая, что в объединенной генеральной совокупности продукции станка A в 1,5 раза больше, чем продукции станка B (поскольку производительность станка A в 1,5 раза выше), а также то, что $a_A=$ 65мм, $a_B=$ 67мм, $\sigma_A^2=\sigma_B^2=1$ мм², имеем:

$$f(x)=\ 0{,}6\frac{1}{\sqrt{2p}}\,e^{-\frac{(x-65)^2}{2}}+\ 0{,}4\frac{1}{\sqrt{2p}}\,e^{-\frac{(x-67)^2}{2}}\ .$$

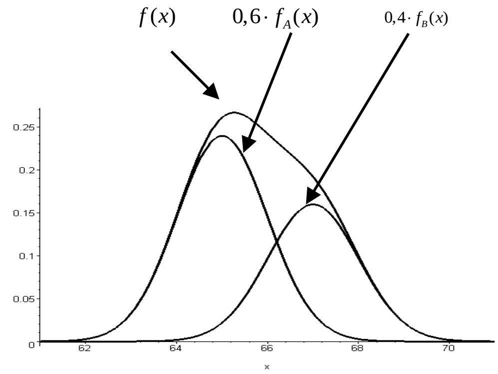


Рис. 1. Графики функций плотности отдельных компонентов и самой смеси.

В последнем соотношении величины $\pi_A=0.6$ и $\pi_B=0.4$ представляют удельные веса соответствующих компонентов смеси (их еще называют *априорными вероятностями* появления наблюдений из данного компонента смеси), а $\theta_A=(a_A,\sigma_A^2)$ и $\theta_B=(a_B,\sigma_B^2)$ – векторные параметры, от значений которых зависят законы распределения компонентов смеси.

Если сотрудники ОТК или потребители изделий-полуфабрикатов захотят по наблюдениям $x_1,x_2,\mathrm{K}\ ,x_n$ определить, на каком именно станке произведено каждое из них, то как раз и возникает одна из типичных задач классификации наблюдений в условиях отсутствия обучающих выборок (конечно, в данном примере можно представить себе специально организованное производство этих изделий, в результате которого можно получить отдельно изделия от станка A и отдельно – от станка B и использовать их в дальнейшем в качестве обучающих выборок.

Функция $f(x)$ *называется смесью вероятностных распределений (дискретных или непрерывных), если она представима в виде соответственно*

$$f(x)=\sum_{j=1}^{k}\pi_j\ f_j(x;\theta_j)\qquad \textit{или}\qquad f(x)=\int f_\omega(x;\theta_\omega)\psi(\omega)d\omega\ .$$

Нас интересует использование моделей смесей в теории и практике автоматической классификации, поэтому сузим данное выше определение смеси и будем рассматривать в дальнейшем лишь случай конечного числа k возможных значений параметра w, что соответствует ситуации, в которой величины $\psi(\omega)$ будут играть роль удельных весов (априорных вероятностей) π_j компонентов смеси ($j=1,2,\ldots,k$). Если же дополнительно постулировать однотипность компонентов-распределений $f_j(x;\theta_j)$, т.е. принадлежность всех $f_j(x;\theta_j)$ к одному общему семейству $\{f(x;\theta)\}$, то модель смеси может быть представлена в виде

$$f(x)=\sum_{j=1}^{k}\pi_j\, f_j(x;\theta_j).$$

Интерпретация в задачах автоматической классификации j-го компонента смеси (j-й генеральной совокупности) в качестве j-го искомого класса (сгустка, скопления) обусловливает естественность дополнительного условия, накладываемого на плотности (полигоны вероятностей) $f_j(x;\theta_j)$ и заключающегося в их унимодальности.

Решить задачу расщепления смеси распределений – это значит по выборке классифицируемых наблюдений $X_1, X_2, \ldots, X_n$, извлеченной из генеральной совокупности, являющейся смесью (10), построить статистические оценки для числа компонентов смеси k, их удельных весов (априорных вероятностей $\pi_1,\pi_2,...,\pi_k$ и, главное, для каждого из компонентов $f_j(x;\Theta(j))$ анализируемой смеси. В некоторых частных случаях имеющиеся априорные сведения дают исследователю точное знание числа компонентов смеси k, а иногда и априорных вероятностей $\pi_1,\pi_2,...,\pi_k$. Тогда задача расщепления смеси сводится лишь к оцениванию функций $f_j(x;\theta_j)$.

Предположим, что нам известно $k = 2$ и что функция распределения смеси имеет вид

$$F(x)=(1-\varepsilon)\Phi\left(\frac{x}{\theta}\right)+\varepsilon\Phi\left(\frac{x}{\lambda\theta}\right),\ 0\le\varepsilon\le 1, \lambda>0 - \text{ известно, } \theta>0,$$

(модель Тьюки). Оценим по выборке $x_1, x_2, \ldots, x_n$ неизвестный параметр ε. Для оценивания ε воспользуемся методом моментов. Находя 2-й и 4-й начальные моменты, получим уравнения для определения ε:

$$(1+\varepsilon(\lambda^2-1)^2=\mu(1+\varepsilon(\lambda^4-1)),\ \text{где}\quad \mu=\frac{3m_2^2}{m_4}, m_2=\frac{1}{n}\sum_{i=1}^{n}x_i^2,\ m_4=\frac{1}{n}\sum_{i=1}^{n}x_i^4.$$

Оценка для ε равна

$$\hat{\varepsilon}_n=\min\left\{\left(\frac{\mu(\sigma^2+1)-2\pm\sqrt{\mu\cdot|\mu(\sigma^2+1)-4\sigma^2|}}{2(\sigma^2-1)}\right)_+,1\right\},$$

где $a_+ = \max(0,a)$. Перед корнем берется знак плюс, если $\varepsilon<(\sigma^2+1)^{-1}$, и знак минус при выполнении противоположного неравенства. Мы будем считать, что $\sigma>1$ и ε достаточно мало. Оценка $\tilde{\varepsilon}_n$ является состоятельной, т.е. $\hat{\varepsilon}_n \xrightarrow[n\to\infty]{p} \varepsilon$, что следует из состоятельности выборочных моментов. Если λ неизвестно, то используя 1-й абсолютный и 2-й начальные моменты, как и выше, получаем оценку для $\tilde{\varepsilon}_n$:

$$\tilde{\varepsilon}_n=\min\left\{\left(\frac{\nu-\sqrt{|\nu^2-4(1-H)|}}{2(\lambda-1)}\right)_+,1\right\},\ \text{где}\quad H=\frac{\pi m_1^2}{2m_2},\ m_1=\frac{1}{n}\sum_{i=1}^{n}|x_i|,\ \nu=H(\lambda+1)-2.\ .$$

Тогда параметр λ найдем из уравнения $\tilde{\varepsilon}_n=\hat{\varepsilon}_n$.

Рассмотрим процедуру разделения смесей распределений, когда у исследователя имеется априорная информация, задаваемая в виде тех или иных ограничений. Рассмотрим случай смеси нормальных распределений с равными матрицами ковариаций, число компонентов k которой неизвестно. Кроме того, имеются обучающие выборки. Вычислительная процедура состоит из следующих шагов.

Шаг 1. Вычисляются оценки векторов средних значений $\bar{\boldsymbol{X}}_j$ $(j=1,2,\dots,l)$ и общей матрицы ковариаций $\Sigma_{\nu-l}$ по неполным обучающим выборкам. Нижний индекс указывает число степеней свободы, соответствующее оценке матрицы ковариаций. Далее для измерения расстояния между объектами $\boldsymbol{X}_i$ и $\boldsymbol{X}_r$ используется расстояние Махалонобиса

$$d^2(\boldsymbol{X}_i,\boldsymbol{X}_r)=(\boldsymbol{X}_i-\boldsymbol{X}_r)^{\mathsf{T}}\boldsymbol{S}_{\nu-l}^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_r).$$

Пусть теперь $\boldsymbol{h}$ − вектор размерности n, у которого *i*-я компонента равна номеру класса для объекта $\boldsymbol{X}_i$. Приравниваем к нулю компоненты $\boldsymbol{h}$, а объектам из обучающей выборки присваиваем соответствующие номера. Текущее значение числа классов m полагается равным . Значение счетчика числа классифицированных объектов $\nu=\sum_{j=1}^{t}\nu_j$, где ν_j − объемы обучающих выборок.

Шаг 2. Обнуляются счетчики числа классификаций объектов η и числа случаев образования новых классов ξ.

Проведем последовательный просмотр неклассифицированных объектов, т.е. объектов $\boldsymbol{X}_i$, для которых $h_i=0$. Пусть $\boldsymbol{X}_i$ – такой объект. Тогда вычисляются расстояния от $\boldsymbol{X}_i$ до центров уже образованных классов $d^2(\boldsymbol{X}_i,\bar{\boldsymbol{X}}_j)$, величины $t_j^2=\dfrac{\nu_j(\nu-m-p+1)}{p(\nu_j-1)(\nu-m)}d^2(\boldsymbol{X}_i,\bar{\boldsymbol{X}}_j)$ и значения $F(t_j^2)$ функции F-распределения с p и $\nu-m-p+1$ степенями свободы от . Вычисляется $f_j=1-F(t_j^2)$. При сделанных допущениях (нормальность, равные матрицы ковариаций) величина f_j равна вероятности реализации расстояния от $\boldsymbol{X}_i$ до $\bar{\boldsymbol{X}}_j$, большего или равного t_j^2 при условии, что $\boldsymbol{X}_i$ действительно принадлежит *j*-му классу. Пусть теперь $f_{\max}=\max\limits_{1\le j\le m} f_j$, $j_{\max}=\arg\max\limits_{1\le j\le m} f_j$, т.е. $f_{\max}=f_{j_{\max}}$.

Относительно объекта $\boldsymbol{X}_i$ принимается одно из трех решений:

1) если $f_{\max}>c_1$, то объект $\boldsymbol{X}_i$ относится к классу с номером $j_{\max}$ и проводится корреляция оценки $\bar{\boldsymbol{X}}_{j_{\max}}$ и $\boldsymbol{S}$:

$$\bar{\boldsymbol{X}}_{j_{\max}}^{нов}=(\nu\bar{\boldsymbol{X}}_{j_{\max}}+\boldsymbol{X}_i)/(\nu_{j_{\max}}+1);\quad \boldsymbol{S}^{нов}=\boldsymbol{S}+\frac{\nu+1}{\nu}(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_{j_{\max}}^{нов})(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_{j_{\max}}^{нов})^T.$$

Используя формулу Бартлетта, получаем скорректированную обратную матрицу

$$(\boldsymbol{S}^{нов})^{-1}=\boldsymbol{S}^{-1}-\frac{\dfrac{\nu+1}{\nu^2}(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_{j_{\max}}^{нов})(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_{j_{\max}}^{нов})^{\mathsf{T}}}{1+\dfrac{\nu+1}{\nu^2}(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_{j_{\max}}^{нов})(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_{j_{\max}}^{нов})^{\mathsf{T}}};\qquad \nu_{j_{\max}}=\nu_{j_{\max}}+1;\ \nu=\nu+1;\ h_i=j_{\max};$$

$\eta=\eta+1$;

2) если $f_{\max}<c_2$, то считается, что объект $\boldsymbol{X}_i$ принадлежит некоторому новому $(m+1)$-му классу; счетчик числа классов m увеличивается на 1 $(m=m+1)$ и полагается $\bar{\boldsymbol{X}}_m=\boldsymbol{X}_i$, $h_i=m$; $\xi=\xi+1$; $\nu=\nu+1$;

3) если $c_2\le f_{\max}<c_1$, то никаких действий не проводится.

Если просмотр объектов не окончен, то переходим к просмотру следующего объекта.

Шаг 3. Проверяется, все ли объекты расклассифицированы, т.е. равенство $\nu=n$. Если оно выполняется, то производится переход на шаг 5, в противном случае на шаг 4.

Шаг 4. Проверяются значения счетчиков η и ξ. Если хотя бы один из счетчиков не равен нулю, то переходят на шаг 2. Если одновременно $\eta=0$ и $\xi=0$, то, следовательно, на шаге 2 не было образовано ни одного нового класса и не было классификации объектов. Поэтому проводится уменьшение порога c_1 на величину δc_1 и увеличение порога c_2 на величину δc_2, т.е. $c_1 = c_1 - \delta c_1, c_2 = c_2 - \delta c_2$. Таким образом, увеличиваются возможности классификации объектов и образования новых классов (принятые при реализации алгоритма значения $\delta c_1 = 0.01$ и $\delta c_2 = 0.001$). Производится переход на шаг 2.

Шаг 5. Проводится реклассификация исходной совокупности объектов $\boldsymbol{X}$ так же, как на шаге, но при $c_1 = 0$ и без пересчета оценок статистических характеристик $\bar{\boldsymbol{X}}_j, \boldsymbol{S}$. Полученная классификация считается окончательной.

Пример 3. Рассмотрим распределение числа несчастных случаев на производстве, где характерно непостоянство степени риска, поскольку различные члены совокупности в различной степени подвержены несчастному случаю. Предположим первоначально пуассоновскую модель распределения числа несчастных случаев.

Таблица 1. Несчастные случаи, произошедшие с 647 женщинами, работавшими на производстве снарядов, в течение пяти недель (Гринвуд и Юл, 1920)

Число несчастных случаев	Наблюденная частота	Пуассоновское распределение $\lambda = 0.465224$	Отрицательное биномиальное распределение $\mu_2 = 0.69, c = 2.06, p = 96$
0	447	406.3	442.9
1	132	189	138.5
2	42	44	44.4
3	21	6.8	14.3
4	3	0.8	4.6
5 и больше	2	0.07	1.5
Полная частота	647	647	647
$\chi_1^2 = 70.4, \quad \chi_2^2 = 4.226, \ \chi_{0.95}^2(5) = 11.07, \chi_{0.95}^2(4) = 9.488$			

Видно, что Пуассоновское распределение дает приближение, которое хи-квадрат Пирсона отвергает. В качестве рабочей гипотезы будем считать (Гринвуд и Юл, 1920), что генеральная совокупность состоит из элементов, в различной степени подверженных несчастным случаям и имеет распределение Пуассона. Предположим еще что априорное распределение параметра λ имеет гамма-распределение с плотностью $\pi(\lambda) = \dfrac{c^p}{\Gamma(p)} e^{-c\lambda} \lambda^{p-1}, \quad 0 < \lambda < \infty, c > 0.$

Тогда $\quad \mathbf{P}(X = k) = \int\limits_0^\infty e^{-\lambda} \dfrac{\lambda^k}{k!} \dfrac{c^p}{\Gamma(p)} e^{-c\lambda} \lambda^{p-1} = \dfrac{c^p \Gamma(k+p)}{k! \Gamma(p)(c+1)^{p+k}},$

$$\mathbf{P}(X = 0) = \left(\frac{c}{c+1}\right)^p, \quad \mathbf{P}(X = k) = \left(\frac{c}{c+1}\right)^p \frac{p(p+1)\cdot \ldots \cdot (p+k-1)}{k!(c+1)^k}, k = 1, 2, \ldots .$$

При этом математическое ожидание равно $\mu_1 = \dfrac{p}{c}$, а дисперсия равна

$$\mu_2 = \mathbf{E}(X - \mu_1) = \frac{p}{c} + \frac{p}{c^2}.$$

Заметим, что при $p-$ натуральном (перейдем от p к $p+1$)

$$\mathbf{P}(X = k) = \left(\frac{c}{c+1}\right)^p \frac{(p+k)!}{k!\, p!(c+1)^k} = C_{p+k}^k q^p (1-q)^k, k = 0, 1, 2, \ldots$$

имеем отрицательное биномиальное распределение — разложение отрицательного бинома по степеням.

Пример 4. Бета-биномиальное распределение. Рассмотрим количество детей мужского пола среди семей с $n = 12$ детьми в $N = 6115$ семей, взятых из больничных записей в Саксонии XIX века и для сравнения рассмотрим две модели: 1) биномиальное распределение и 2) бета-биномиальное распределение. В первом случае мы будем предполагать, что вероятность p одна и та же для всех семей. Во втором случае она различна и будем считать, что параметр p сам случаен и имеет априорное бета-распределение $g(p) = p^{\alpha-1}(1-p)^{\beta-1} / B(\alpha,\beta),\ 0 < p < 1,$ $\alpha > 0,$ $\beta > 0,$ с неизвестными параметрами α и β. В случае 1): $b_k = \mathbf{P}(X = k; p) = C_n^k p^k (1-p)^{n-k},\ k = 0,1,\ldots,n,$ В случае 2) составное распределение дается выражением

$$f_k = \int_0^1 \mathbf{P}(X = k; p)\, g(p)\, dp = \frac{\Gamma(n+1)\Gamma(k+\alpha)\Gamma(n-k+\beta)\Gamma(\alpha+\beta)}{\Gamma(k+1)\ \Gamma(n-k+1)\Gamma(n+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)},\ \Gamma(n+1) = n!.$$

Ниже приведена статистика и предполагаемые значения по биномиальному (Bin) и бета-биномиальному (BB) распределению.

k	0	1	2	3	4	5	6	7	8	9	10	11	12
Статистика	**3**	**24**	**104**	**286**	**670**	**1033**	**1343**	**1112**	**829**	**478**	**181**	**45**	**7**
$N \cdot f_k$	2.3	22.6	104.8	310.8	655.7	1036.3	1258.1	1182.2	853.6	461.9	177.8	47.8	5.2
$N \cdot b_k$	0.9	12.1	71.8	258.5	628.1	1085.2	1367.3	1265.6	854.2	410.0	132.8	26.1	2.3

Для BB статистика $\chi^2 = 14.47,$ для Bin: $\chi^2 = 110.505.$ Критическое значение $\chi^2_{0.95}(9) = 16.919 > 14.47,$ $\chi^2_{0.95}(10) = 24.996 < 110.505,$ то BB не отвергается, а Bin – отвергается.

Распределение Линника. Эти распределения (симметричные геометрически устойчивые распределения) находят широкое применение в финансовой математике, телекоммуникационных системах, астрофизике, генетике. Такие распределения являются предельными для геометрических сумм независимых одинаково распределенных случайных величин (с.в.), распределения которых принадлежат области нормального притяжения симметричного строго устойчивого распределения. Были введены в 1953 г., имеют характеристическую функцию $\varphi_\alpha(t) = \dfrac{1}{1+|t|^\alpha},\ 0 < \alpha \le 2$. Ю.В.Линник доказал, что $\varphi_\alpha(t)$ является характеристической функцией абсолютно непрерывной случайной величины, поэтому соответствующая плотность распределения $p_\alpha(x)$ будет определяться формулой

$$p_\alpha(x) = \frac{2}{\pi}\int_0^\infty \frac{\cos(tx)}{1+t^\alpha}\, dt, \quad x > 0.$$

При $\alpha = 2$ имеем распределение Лапласа, поэтому будем считать $0 < \alpha < 2$. При других значениях α (1.5, 1.0, 0.5)плотности имеют вид, считая сверху от 2:

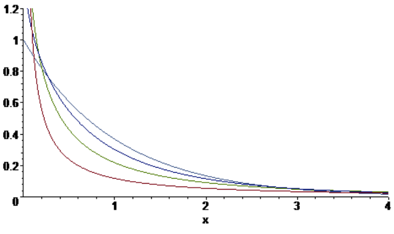


Рис.2. График плотностей распределения Линника.

При $x \to \infty,\ x^\alpha(1-F_\alpha(x)) \to \dfrac{\Gamma(\alpha)}{\pi}\cdot\sin\left(\dfrac{\pi\alpha}{2}\right),$ где $F_\alpha(x)$ – функция распределения.

Введем неотрицательную случайную величину $Y_{\alpha\beta}$, не зависящую от X_β, с плотностью распределения

$$g(s;\alpha,\beta) = \left(\frac{\beta}{\pi}\sin\frac{\pi\alpha}{2}\right)\cdot\frac{s^{\alpha-1}}{1+s^{2\alpha}+2s^\alpha\cos(\pi\alpha/\beta)},\ 0 < \alpha < \beta \le 2, 0 < s < \infty.$$

Покажем, что величина X_α является смесью распределений $Y_{\alpha\beta}$ с масштабным параметром:

$$\varphi_\alpha(t) = \int_0^\infty \varphi_\beta(t/s)\, g(s;\alpha,\beta)\, ds \quad \Leftrightarrow \quad p_\alpha(x) = \int_0^\infty p_\beta(sx)\, g(s;\alpha,\beta)\, ds. \qquad (11)$$

Доказательство (см. [13]). Заметим, что (11) эквивалентно следующему:

$$\frac{1}{1+t^\alpha} = \int_0^\infty \frac{s^\beta}{s^\beta + t^\beta}\, g(s;\alpha,\beta)\, ds,\ \ 0<\alpha<\beta\le 2,\ \ t\ge 0.$$

Чтобы доказать последнее, обозначим

$$I = \frac{\pi}{\beta \sin\frac{\pi\alpha}{\beta}} \int_0^\infty \frac{s^\beta}{s^\beta+t^\beta}\, g(s;\alpha,\beta)\, ds = \int_0^\infty \frac{s^{\beta+\alpha-1}}{(s^\beta+t^\beta)(1+s^{2\alpha}+2s^\alpha\cos\frac{\pi\alpha}{\beta})}\, g(s;\alpha,\beta)\, ds,$$

поэтому достаточно установить, что $I = \frac{\pi}{\beta\sin\frac{\pi\alpha}{\beta}} \cdot \frac{1}{1+t^\alpha}$.

Производя замену переменной $\tau = t^\beta s^{-\beta}$, мы имеем:

$$I = \frac{\pi}{\beta\sin\frac{\pi\alpha}{\beta}} \int_0^\infty \frac{\tau^{(\alpha/\beta)-1}\, d\tau}{(1+\tau)(\tau^{2\alpha/\beta}+t^{2\alpha}+2\tau^{\alpha/\beta}t^\alpha\cos\frac{\pi\alpha}{\beta})}.$$

Используем тождество:

$$\frac{1}{(1+\tau)(\tau^{2\alpha/\beta}+t^{2\alpha}+2\tau^{\alpha/\beta}t^\alpha\cos\frac{\pi\alpha}{\beta})} = \frac{1}{t^\alpha 2i\sin\frac{\pi\alpha}{\beta}}\left(\frac{1}{\tau^{\alpha/\beta}+t^\alpha e^{-i\pi\alpha/\beta}} - \frac{1}{\tau^{\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta}}\right),$$

мы получим:

$$I = \frac{1}{2i\beta\sin\frac{\pi\alpha}{\beta}}\left(\int_0^\infty \frac{\tau^{(\alpha/\beta)-1}d\tau}{(1+\tau)(\tau^{\alpha/\beta}+t^\alpha e^{-i\pi\alpha/\beta})} - \int_0^\infty \frac{\tau^{(\alpha/\beta)-1}d\tau}{(1+\tau)(\tau^{\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta})}\right).$$

Рассмотрим функцию $q(\tau) = \frac{1}{(1+\tau)(\tau^{\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta})}$

в комплексной $\tau-$плоскости разрезанной вдоль положительного луча. Определим ветвь $\tau^{\alpha/\beta}$ условием $\tau^{\alpha/\beta} = r^{\alpha/\beta}e^{i\varphi\alpha/\beta},\ \tau = re^{i\varphi},\ 0<\varphi<2\pi.$

Так как $\left|\frac{\alpha}{\beta}(\pi-\varphi)\right|$ когда $0<\varphi<2\pi$, мы имеем: $\tau^{\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta} = e^{i\varphi\alpha/\beta}(r^{\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta(\pi-\varphi)}) \ne 0.$

Следовательно, функция $q(\tau)$ аналитическая в разрезанной плоскости. Обозначим через $G_{R,\varepsilon}$ односвязную область $G_{R,\varepsilon} = \{\tau : \varepsilon < |\tau| < R\},\ 0<\varepsilon<R(>1)$, ее граница $\partial G_{R,\varepsilon}$, пройденная в направлении, которое выходит слева направо (интервал $\{\tau : \varepsilon<\tau<R\}$), проходя дважды в противоположных направлениях. По теореме вычетов Коши мы имеем

$$\oint_{\partial G_{R,\varepsilon}} q(\tau)\, d\tau = 2\pi i\,(\text{вычет}\, q(\tau)\ \text{в}\ \tau=-1) = -\frac{2\pi i}{1+t^\alpha}.$$

Устремляя в интеграле $R\to\infty,\ \varepsilon\to 0,$ мы получаем

$$\int_0^\infty \frac{\tau^{(\alpha/\beta)-1}d\tau}{(1+\tau)(\tau^{\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta})} - e^{2\pi i((\alpha/\beta)-1)}\int_0^\infty \frac{\tau^{(\alpha/\beta)-1}d\tau}{(1+\tau)(\tau^{\alpha/\beta}e^{2\pi i\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta})} = -\frac{2\pi i}{1+t^\alpha}.$$

Это равенство может быть переписано в виде

$$\int_0^\infty \frac{\tau^{(\alpha/\beta)-1}d\tau}{(1+\tau)(\tau^{\alpha/\beta}+t^\alpha e^{i\pi\alpha/\beta})} - \int_0^\infty \frac{\tau^{(\alpha/\beta)-1}d\tau}{(1+\tau)(\tau^{\alpha/\beta}+t^\alpha e^{-i\pi\alpha/\beta})} = -\frac{2\pi i}{1+t^\alpha}.$$

Подставляя в исходное равенство, получаем результат утверждения.

Пусть $\varphi(t,a)-$функция двух действительных переменных такая, что

1) при каждом фиксированном a функция $\varphi_a(t) = \varphi(t,a)$ есть характеристическая функция;

2) при каждом фиксированном t функция $\varphi_t(a)=\varphi(t,a)$ – измеримая функция. Тогда для любой функции распределения $H(x)$ функция $\gamma(t)=\int\limits_{-\infty}^{\infty}\varphi(t,a)dH(a)$ есть характеристическая функция.

Следствие (i). Пусть $\varphi_1(t),\varphi_2(t),...$ – характеристические функции и $a_1,a_2,...$ – неотрицательные числа такие, что $\sum_{n=1}^{\infty}a_n=1$. Тогда функция $\varphi(t)=\sum\limits_{n=1}^{\infty}a_n\varphi_n(t)$ есть характеристическая функция.

Чтобы понять доказательство основного результата, обратимся к доказательству следствия. Для этого рассмотрим $F(x)=\sum\limits_{n=1}^{\infty}a_nF_n(x)$, где $F_1(x),F_2(x),...$ – функции распределения, соответствующие $\varphi_1(t),\varphi_2(t),...$ соответственно. Доказательство следует из линейности интеграла.

Следствие (ii). Пусть $\varphi(t)$ – характеристическая функция. Тогда $\eta(t)=\dfrac{3+\varphi(t)}{3-\varphi(t)}-1=\dfrac{2\varphi(t)}{3-\varphi(t)}$ также является характеристической функцией. Действительно, $\dfrac{2\varphi(t)}{3-\varphi(t)}=\sum\limits_{n=1}^{\infty}\dfrac{2\varphi^n(t)}{3^n}$ является характеристической функцией.

Следствие (iii). Пусть $\varphi(t)$ – характеристическая функция и $p>0$. Тогда $\beta(t)=\dfrac{p}{t^p}\int_0^t\varphi(u)u^{p-1}du$ – также характеристическая функция.Действительно, пусть $F(x)$ – функция распределения, соответствующая характеристической функции $\varphi(t)$. Функция $h_t(x)=px^{p-1}/t^p, 0<x<t,$ является плотностью распределения. Обозначим через $\alpha_t(u)=\alpha_1(tu)$ соответствующую ей характеристическую функцию. Используя равенство Парсеваля: если $F(x)$ и $G(x)$ – функции распределения, а $f(t)$ и $g(t)$ – соответствующие им характеристические функции, то $\int\limits_{-\infty}^{\infty}e^{-itx}g(x)dF(x)=\int\limits_{-\infty}^{\infty}f(y-t)dG(y)$, поэтому

$$\beta(t)=\frac{p}{t^p}\int_0^t\varphi(u)u^{p-1}du=\int_{-\infty}^{\infty}\varphi(u)h_t(u)du=\int_{-\infty}^{\infty}\alpha_t(u)dF(u)=\int_{-\infty}^{\infty}\alpha_1(tu)dF(u),$$

и в силу нашего утверждения о смесях $\beta(t)$ есть характеристическая функция.

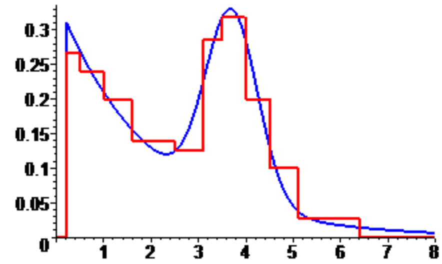
Следствие (iv). Пусть $\varphi(t)$ – характеристическая функция и $p>0$. Тогда $\gamma(t)=\exp(p\varphi(t)-1)$ также является характеристической функцией.

5. Построение вероятностных моделей по выборочным данным

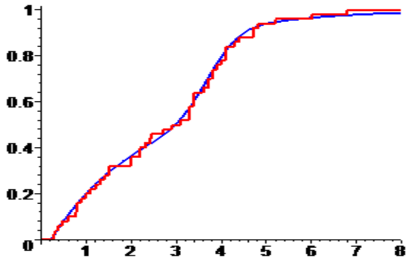
Пример 5. Ниже даны размеры (длина) 50 валунов, м, на территории Беларуси:

1.3, 1.1, 4.4, 0.32, 0.6, 0.8, 1.0, 0.9, 0.46, 0.36, 0.75, 1.5, 4.1, 1.5, 3.9, 3.8, 0.8, 0.26, 1.2, 2.2, 3.1, 3.7, 3.3, 4.7, 2.0, 3.3, 5.3, 4.1, 3.7, 2.0, 3.4, 2.3, 2.2, 1.4, 2.45, 2.4, 4.0, 3.4, 2.7, 3.4, 4.1, 4.9, 3.3, 4.7, 4.8, 4.3, 6.5, 3.6, 2.5, 3.8 (Источник: [9], , с.224-226.)

На рисунках приведены гистограмма и плотность $0,62\cdot 0,5\cdot e^{-0,5(x-0,22)}+0,38\cdot\dfrac{1}{\sqrt{2\pi}\cdot 0,55}e^{-\frac{(x-3,7)^2}{2\cdot 0,55^2}}$, а также



эмпирическая функция распределения и смесь функций распределения с весами 0,62 и 0,38.



Видно, что статистические данные представляют собой смесь двух распределений: показательного и нормального.

6. Задания

1. Рассмотрим независимые случайные величины X и Y, имеющие нормальное распределение $N(0,\sigma^2)$. Найти распределение случайной величины $\xi=\sqrt{X^2+Y^2}$.

2. Пусть случайная величина X имеет плотность распределения, равную $g(x)=\frac{1}{2\pi}\sqrt{4-x^2}\cdot I(|x|\le 2)$ (полукруговой закон Вигнера). Найти начальные моменты четного порядка (нечетные моменты равны нулю) $m_{2k}=\frac{1}{2\pi}\int\limits_{-2}^{2}x^{2k}\sqrt{4-x^2}\,dx$.

3. Пусть случайная величина (с.в.) $Y=X_1$ имеет функцию распределения

$$F_1(x)=\frac{1}{2}+\frac{1}{\pi}\arccos(2x-1),\ -1\le x\le 1,\ \text{(распределение арксинус)}, \qquad (7)$$

а величина $X=X_2$ имеет функцию распределения

$$F_2(x)=\frac{1}{2}+\frac{1}{\pi}\mathrm{arctg}(x),\ x\in\mathbf{R},\ \text{(распределение Коши)}. \qquad (8)$$

Какому преобразованию надо подвергнуть с.в. X_2, чтобы получить случайную величину X_1? (т.е. чтобы она имела распределение (7)).

4. Пусть с. в. $Z_1=X$ имеет распределение Коши C(0,1) с плотностью распределения

$$f(x)=\frac{1}{\pi(1+x^2)},\quad x\in R. \qquad (9)$$

Докажите, что для $n=2,3,\ldots$ случайные величины

$$Z_2=\frac{2X}{1-X^2},Z_3=\frac{3X-X^3}{1-3X^2},Z_4=\frac{4X-4X^3}{1-6X^2+X^4},\ldots,Z_n=\frac{C_n^1X-C_n^3X^3+C_n^5X^5-..}{1-C_n^2X^2+C_n^4X^4-C_n^6X^6+\ldots}=\frac{\sum_{k=0}^{n}(-1)^kC_n^{2k+1}X^{2k+1}}{\sum_{k=0}^{n}(-1)^kC_n^{2k}X^{2k}}$$

также имеют Коши C(0,1) распределение.

5. Случайные величины X_1 и X_2 независимы и одинаково распределены с плотностью

$$f_1(x)=f_2(x)=\frac{C}{1+x^4},\ x\in R.$$

Найти постоянную C и доказать, что величина $Y = \frac{X_1}{X_2}$ распределена по закону Коши.

6. Случайные величины X_1 и X_2 независимы и одинаково распределены с плотностью $f_1(x) = f_2(x) = \frac{1}{\pi(1+x^2)}, x \in R.$ а) Найти распределение величины $Y = \frac{X_1}{X_2}$. б) Найти распределение величины $Z = X_1 \cdot X_2$.

7. Случайные величины X_1 и X_2 независимы, а их плотности распределения равны соответственно $f_1(x) = \frac{1}{\pi\sqrt{1-x^2}}, |x| < 1,$ и $f_2(x) = x\exp(-x^2/2),\ x > 0.$

Покажите, что случайная величина $Y = X_1 \cdot X_2$ имеет нормальное распределение $N(0,1)$.

8. Пара (X_1, X_2) имеет равномерное распределение в круге радиуса 1:

$$f(x,y) = \begin{cases} \frac{1}{\pi}, & x_1^2 + x_2^2 \le 1; \\ 0, & x_1^2 + x_2^2 > 1. \end{cases}$$

Найти плотность распределения случайной величины $Y = \frac{X_1}{X_2}$.

9. Пусть ξ и η независимые случайные величины, имеющие одну и ту же плотность распределения $f(x) = \frac{2}{\pi}\frac{1}{e^x + e^{-x}}\left(f(x) = \frac{1}{2}\mathrm{sech}(\pi x/2) \right)$.

Найдите плотность распределения величины $\varsigma = \xi + \eta$ и $\tau = \frac{\xi}{\eta}$.

10. а) Пусть случайная точка (x_1, x_2) имеет нормальное распределение на плоскости с плотностью

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-a_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-a_1)(x_2-a_2)}{\sigma_1\sigma_2} + \frac{(x_2-a_2)^2}{\sigma_2^2} \right)\right).$$

Найдите плотность распределения отношения $\varsigma = \frac{x_2 - a_2}{x_1 - a_1}\cdot\frac{\sigma_1}{\sigma_2}$. б) Найдите плотность распределения $\varsigma = \frac{x_1}{x_2}$. в) Найдите плотность распределения величины $\xi = \varsigma^2 = \left(\frac{x_1}{x_2}\right)^2$.

11. Пусть $\tau -$ случайная величина, значение которой равно n, когда

$$\xi_1 \ge p, \xi_1\xi_2 \ge p, \dots, \xi_1\xi_2\cdot\ldots\cdot\xi_n \ge p, \xi_1\xi_2\cdot\ldots\cdot\xi_n\xi_{n+1} < p,$$

где $0 < p < 1 -$ задано, а $\xi_1, \xi_2, \dots, \xi_n, \dots -$ независимые с.в. равномерно распределенные на (0,1). Каково распределение с.в. τ ? (Этот прием используется на компьютере для получения значений соответствующей с.в.).

12. Проводится следующий эксперимент: наблюдается с.в. X, подчиненная пуассоновскому распределению с параметром λ. Затем производится X испытаний Бернулли с вероятностью успеха p. Каково распределение числа успехов Y в такой схеме?

13. r различимых шаров раскладываются случайным образом по n различимым ящикам. Найти вероятность того, что ровно k ячеек из n окажутся пустыми.

14. Колода состоит из sn карт (например, при $s = 4 -$ число мастей в обычной колоде). Все масти занумерованы числами 1, 2, …., n. Из колоды извлекается без возвращения выборка объема $r \ge n$. а) Вычислить вероятность q_r того, что каждое число представлено в выборке. б) Какова вероятность того, что потребуется ровно r извлечений для получения выборки, содержащей все числа. в) Чему равна эта вероятность при $s \to \infty$? Найти среднее время ожидания сбора всех купонов.

15. Пусть с.в. X_1 имеет распределение Релея $Rel(\sigma^2)$ с плотностью $f(x_1) = \frac{x_1}{\sigma^2} e^{-\frac{x_1^2}{2\sigma^2}}, x_1 > 0,$ а X_2 – распределение Релея $Rel(\lambda^2)$ и они независимы. Покажите, что плотность распределения отношения $Y = X_1 / X_2$ есть $f_Y(y) = \frac{2\sigma^2\lambda^2 y}{(\lambda^2 y^2 + \sigma^2)^2}, y > 0.$

16. Сколько в среднем надо бросать симметричную монету до появления серии $ГГ$?

17. Пусть имеется n ящиков. В эти ящики независимо один от другого бросаются r дробинок. Предположим, что вероятность попадания любой фиксированной дробинки в j – й ящик равна $1/n$ для всех $j = 1, 2, ..., n$ (иными словами, мы рассматриваем случай, когда дробинки неразличимы, ящики различимы). Обозначим через $\mu_0(r, n)$ – число пустых ящиков и пусть $\theta_j = 1$, если j – й ящик пуст ($\theta_j = 0$ – в противном случае). Тогда $\mu_0(r, n) = \theta_1 + \theta_2 + ... + \theta_n$ и $\mathbf{E}(\mu_0(r, n)) = \sum_{j=1}^{n} \mathbf{E}(\theta_j)$. В свою очередь, $\mathbf{E}(\theta_j) = \mathbf{P}(\theta_j = 1) = \left(1 - \frac{1}{n}\right)^r$, откуда $\mathbf{E}(\mu_0(r, n)) = n\left(1 - \frac{1}{n}\right)^r$.

18. Вы покупаете какой-либо продукт, например, бутылку пепси-колы, а на крышке внутри нарисован купон какого либо из n цветов (всего цветов n). Если вы соберете купоны всех цветов, то получите приз. Сколько в среднем надо купить бутылок пепси, чтобы получить приз?

19. Показать, что если с.в. имеет усеченное распределение Коши с плотностью $f_1(x) = \frac{1}{\pi(1 + x^2)\operatorname{arctg}\lambda} I(|x| < \lambda)$, то случайная величина: $Z = X$ с вероятностью $\frac{2\operatorname{arctg}\lambda}{\pi}$, $= \frac{1}{X}$ с вероятностью $1 - \frac{2\operatorname{arctg}\lambda}{\pi}$ имеет стандартное распределение Коши с плотностью $f_2(x) = \frac{1}{\pi(1 + x^2)}$.

20. Пусть случайная величина X_1 имеет распределение с плотностью

$f_1(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}, x \in R,$ а независимая от неё с.в. X_2 – с плотностью $f_2(x) = 2xe^{-x^2}, x > 0.$

Найти распределение случайной величины $Y = \frac{X_1}{X_2}$.

21. Пусть $f_1(x) \ge 0$ – монотонно убывающая, а $f_2(x) \ge 0$ – монотонно возрастающая функция и

$$f(x) = y = \begin{cases} f_1(x), & x \le x_1, \\ 0, & x_1 < x < x_2, \\ f_2(x), & x \ge x_2. \end{cases}$$

Имеем преобразование: $\eta = f(\xi)$. Найти плотность распределения случайной величины η через плотность распределения $w_\xi(x)$ случайной величины ξ. Решить эту задачу для частного случая $f_1(x) = x_1 - x,\ f_2(x) = x - x_2$. Рассмотреть также преобразование

$$f(x) = y = \begin{cases} 0, & x < x_0, \\ (x - x_0)^{\rho}, & x \ge x_0,\ \rho > 0. \end{cases}$$

Рассмотреть случай $\rho = 1$ и $\xi \in N(0, \sigma^2)$.

22. Пусть ξ_1, ξ_2 – независимые случайные величины, имеющие соответственно плотности распределения

$$f(x) = \frac{1}{p\Gamma\left(\frac{d_1}{p}\right)a_1}\left(\frac{x}{a_1}\right)^{d_1-1}\exp\left(-\left(\frac{x}{a_1}\right)^{p}\right), x > 0, \text{ и } g(y) = \frac{1}{p\Gamma\left(\frac{d_2}{p}\right)a_2}\left(\frac{y}{a_2}\right)^{d_2-1}\exp\left(-\left(\frac{y}{a_2}\right)^{p}\right), y > 0,$$

$a_1, a_2, d_1, d_2, p > 0.$ Пусть $Z = \xi_1 \cdot \xi_2$. Найдите плотность распределения величины Z.

Здесь $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}\,dt, \alpha > 0.$

23. Пусть интересующая нас система в момент времени t может находиться в одном из состояний $E_0, E_1, E_2, \ldots,$ множество которых конечно или счетно. Известно, что в момент времени $t + h$ система из состояния E_n переходит в состояние E_{n+1} с вероятностью $\lambda_n h + o(h)$ и в состояние E_{n-1} с вероятностью $\mu_n h + o(h)$. Вероятность того, что она перейдет в состояние $E_{n\pm k}, k \ge 2,$ есть $o(h)$. Отсюда следует, что вероятность остаться в том же состоянии E_n равна $1 - \lambda_n h - \mu_n h + o(h)$. Мы предполагаем, что λ_n, μ_n зависят только от n, но не от t. указанный процесс носит название ***процесса гибели и размножения***, где переход $E_n \to E_{n+1}$ означает увеличение на 1 (размножение), а переход $E_n \to E_{n-1}$ означает уменьшение на 1 (гибель).

Обозначим через $p_k(t)$ того, что изучаемая нами система в момент времени t находится в состоянии E_k.Определить вероятности $p_k(t)$, $k = 0, 1, 2, \ldots$.

24. Говорят, что случайная величина X имеет логистическое распределение, если ее функция распределения равна $F(x) = \dfrac{1}{1 + e^{-x}}, -\infty < x < \infty,$ (и плотность распределения $f(x) = \dfrac{e^{-x}}{(1 + e^{-x})^2} = F(x)(1 - F(x))$ **).** Покажите, что если:

а) U имеет равномерное распределение на $(0,1)$, то случайная величина $X = \ln\left(\dfrac{U}{1-U}\right)$ имеет логистическое распределение;

б) если X_1, X_2 – независимые случайные величины, каждая из которых имеет распределение с плотностью $f(x) = e^{-x}\exp(-e^{-x}), -\infty < x < \infty$ то $X_1 - X_2$ имеет логистическое распределение.

25. Пусть X_1, X_2, X_3 имеют распределение с плотностью $G(p_i, a), i = 1, 2, 3$, и независимы, а $U_1 = \dfrac{X_1}{X_1 + X_2}, U_2 = \dfrac{X_1 + X_2}{X_1 + X_2 + X_3}$. Покажите, что с.в. U_1 имеет бета-распределение $B(p_1, p_2)$ с плотностью распределения $g_1(u) = \dfrac{\Gamma(p_1 + p_2)}{\Gamma(p_1)\Gamma(p_2)} \cdot u^{p_1-1}(1-u)^{p_2-1}, 0 < u < 1,$ и равна 0 в противном случае, а U_2 имеет бета-распределение $B(p_1 + p_2, p_3)$.

Рассмотреть также случай $U_k = \dfrac{X_1 + X_2 + \ldots + X_{k-1}}{X_1 + X_2 + .. + X_{k-1} + X_k}$

26. Пусть X_1, X_2, X_3 имеют распределение: $X_i \in N(0,\sigma^2), i=1,2,3$, и независимы, а $Y_1 = X_1 / X_3, Y_2 = X_2 / X_3$. Покажите, что плотность распределения пары (Y_1,Y_2) есть

$$g(y_1,y_2) = \frac{1}{2\pi}\cdot\frac{1}{(1+y_1^2+y_2^2)^{3/2}}, (y_1,y_2)\in \mathbf{R}^2.$$

27. Пусть X_1, X_2, X_3 имеют распределение нормальное распределение: $X_i \in N(0,\sigma^2), i=1,2,3$, и независимы, а $V_1 = \frac{X_1}{\sqrt{X_1^2+X_2^2}}, V_2 = \frac{\sqrt{X_1^2+X_1^2}}{\sqrt{X_1^2+X_1^2+X_3^2}}$. Покажите, что плотность распределения каждой случайных величин V_1 и V_2 есть

$$h_1(v) = \begin{cases} \frac{1}{\pi}\cdot\frac{1}{\sqrt{1-v^2}}, & |v|<1, \\ 0, & |v|>1, \end{cases} \quad \text{и} \quad h_2(v) = \begin{cases} \frac{v}{\sqrt{1-v^2}}, & 0<v<1, \\ 0, & v<0, v>1. \end{cases}$$ соответственно.

28. Пусть случайная величина X имеет распределение Коши с плотностью $f(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < +\infty,$ и $a=d=\cos\alpha, b=\sin\alpha, c=-\sin\alpha$. Покажите, что тогда случайная величина $Y = \frac{aX+b}{cX+d} = \frac{X\cdot\cos\alpha+\sin\alpha}{-X\sin\alpha+\cos\alpha}$ имеет то самое распределение Коши. В частности при $\alpha = -\frac{\pi}{2}$ и $\alpha = \frac{\pi}{2}$ имеем: $Y = -\frac{1}{X}$. Но так как величины X и $-X$ одинаково распределены, то $Y = \frac{1}{X}$ имеет распределение Коши.

Точно так же величина $Y = \frac{X\cdot\cos\alpha+\sin\alpha}{X\sin\alpha-\cos\alpha}$ будет иметь распределение Коши, если X имеет распределение Коши.

29. Пусть X_1 и X_2 – независимые случайные величины, каждая из которых имеет плотность распределения $f_i(x) = \frac{\sqrt{2}}{\pi}\frac{x^2}{1+x^4}, \; -\infty < x < +\infty; \; i=1,2.$

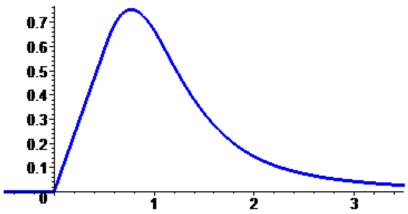
Покажите, что случайная величина $U = \frac{X_1}{X_2}$ имеет распределение Коши $C(0,1)$.

30. *Задача одномерной парковки*. Есть улица известной длины x (x не обязательно целое число), на которой вдоль прямой паркуются машины длиной 1. Однако вместо того, чтобы парковаться организованным способом они делают это наугад, выбирая случайно место парковки. Предполагая, что машины будут продолжать парковаться до тех пор пока уже не смогут больше этого сделать, каково ожидаемое количество машин сможет припарковаться ?

31. Равнобедренный треугольник на плоскости образован единичным вектором вдоль оси абсцисс и единичным вектором в случайном направлении. Найдите распределение длины третьей стороны.

32 Пусть U_1,U_2,U_3,U_4 – н.о.р.с.в., имеющие равномерное распределение на $(0,1)$. Докажите, что $\frac{U_1+U_2}{U_3+U_4}$ имеют плотность

$$f(x)=\begin{cases}\dfrac{7x}{6}, & 0<x<\dfrac{1}{2},\\ \dfrac{8}{3}-\dfrac{3x}{2}-\dfrac{2}{3x^2}+\dfrac{1}{6x^3}, & \dfrac{1}{2}\le x<1,\\ -\dfrac{2}{3}+\dfrac{x}{6}+\dfrac{8}{3x^2}-\dfrac{3}{2x^3}, & 1\le x<2,\\ \dfrac{7}{6x^3}, & 2\le x.\end{cases}$$



33 а) Пусть независимые с. в. X_1, X_2 имеют стандартное нормальное распределение $N(0,1)$. Найдите распределение величины $Y = X_1 X_2$.

б) Пусть независимые с. в. X_1, X_2, X_3, X_4 имеют стандартное нормальное распределение $N(0,1)$. Покажите, что $X_1X_2 + X_3X_4$ имеет распределение Лапласа с плотностью $\dfrac{1}{2}e^{-|x|}$.

в) Пусть с. в. X_1, X_2 имеют совместное нормальное распределение со средними 0, дисперсиями 1 и коэффициентом корреляции ρ. Покажите, что распределение случайной величины $Z = X_1X_2$ имеет плотность распределения

$$f_Z(z;\rho)=\frac{1}{\pi\sqrt{1-\rho^2}}\exp\left(\frac{\rho z}{1-\rho^2}\right)K_0\left(\frac{|z|}{1-\rho^2}\right), -\infty<z<\infty,$$

где $K_0(a\beta)=\int\limits_0^\infty \dfrac{\cos(ax)\,dx}{\sqrt{\beta^2+x^2}}$ есть функция Бесселя второго рода нулевого порядка,

$$K_\nu(z)=\frac{\sqrt{\pi}(z/2)^\nu}{\Gamma(\nu+\frac{1}{2})}\int\limits_0^\infty \exp(-z\cosh t)\cdot\cos(\nu t)\,dt, |Arg\, z|<\frac{\pi}{2},$$

$$K_\nu(z)=\sqrt{\frac{\pi}{2}}\frac{e^{-z}}{\sqrt{z}}\left(1+O\left(\frac{1}{z}\right)\right), |z|\to\infty,$$

в частности, $K_0(z)=\sqrt{\dfrac{\pi}{2z}}e^{-z}\sum\limits_{m=0}^{\infty}\dfrac{((2m-1)!!)^2}{m!(-8z)^m}$.

Найти $\lim\limits_{\rho\to 1} f_Z(z;\rho)$.

Замечание. Аналогично можно показать, что если н.о.р. с.в. $X_1, X_2, X_3, X_4 \in N(0,1)$, то случайный определитель $D=\begin{vmatrix} X_1 & X_2\\ X_3 & X_4\end{vmatrix} = X_1X_4 - X_2X_3$ имеет Лапласа распределение с плотностью $\dfrac{1}{4}\exp\left(-\dfrac{1}{2}|x|\right), -\infty<x<\infty.$

34. а) Пусть совместное распределение (X,Y) описывается плотностью

$f(x,y)=\dfrac{1}{\pi}g(x^2+y^2),\ \int\limits_0^\infty g(x)\,dx=1$. Как распределена величина $Z=\dfrac{X}{Y}$.

б) Пусть вектор (X_1, X_2, X_3) имеет плотность распределения $f(x_1,x_2,x_3)=g(\sqrt{x_1^2+x_1^2+x_1^2})$. Найти плотность распределения $R=\sqrt{X_1^2+X_2^2+X_3^2}$. Рассмотреть случай: $X_i\in N(0,1), i=1,2,3$ независимы.

35. а) Как распределена случайная величина $\eta_1 = \dfrac{\xi_1 + \xi_2\xi_3}{\sqrt{1+\xi_3^2}}$, если ξ_1, ξ_2, ξ_3 – независимые и одинаково распределенные $N(0,1)$ случайные величины.

б) Как распределена случайная величина $\eta_2 = \dfrac{\xi_1\xi_2 + \xi_3\xi_4}{\sqrt{\xi_2^2 + \xi_4^2}}$, если $\xi_1, \xi_2, \xi_3, \xi_4$ – независимые и одинаково распределенные $N(0,1)$ случайные величины.

36. Пусть $X_2 \in N(0,1)$, а X_1 имеет распределение с плотностью $f_1(x) = \dfrac{1}{\pi\sqrt{1-x^2}}, |x| < 1$, Y и X_1 – независимы. Какое распределение имеет величина $Z = X_2 / X_1$?

37. а) Случайная величина X имеет Пуассоновское распределение с параметром λ. Покажите, что $\boldsymbol{E}(\lambda g(X+1) - Xg(X)) = 0$ для любой ограниченной функции $g(x)$ на множестве целых чисел 0, 1, ...

b) пусть X – случайная величина на множестве $\{0, 1, \ldots\}$, $\lambda > 0$ – заданное число. Покажите, что X имеет распределение Пуассона с ожидаемым значением λ, если $\boldsymbol{E}(\lambda g(X+1) - Xg(X)) = 0$ для любой ограниченной функции $g(x)$ на множестве целых чисел 0, 1, ...

Указание. Рассмотрите функцию $g(x)$, определенную следующим образом: $g(i) = 1$ и $g(x) = 0$ для $x \neq i$, i – фиксированное неотрицательное целое число.

38. Непрерывная случайная величина X имеет совместную плотность распределения $f(x) = c(x + \sqrt{x})$ для $0 < x < 1$ и $f(x) = 0$ в противном случае. Найдите константу c. Найдите плотность распределения величины $\dfrac{1}{X}$.

39. Радиус круга равномерно распределен на (0,1). Найдите плотность распределения площади круга.

40. Вы выбираете наугад точку внутри прямоугольника, стороны которого имеют длины 1 и $\dfrac{3}{2}$. Случайная величина X – расстояние от точки до ближайшей стороны прямоугольника. Какова плотность распределения величины X ?

41. Непрерывные случайные величины X и Y имеют совместную плотность распределения $f(x,y) = xe^{-x(y+1)}$ для $x, y > 0$ и $f(x,y) = 0$ в противном случае. Найдите плотность распределения величины $\varsigma = XY$.

42. Непрерывные случайные величины X и Y имеют совместную плотность распределения $f(x,y) = \frac{1}{2}(x+y)e^{-(x+y)}$ для $x, y > 0$ и $f(x,y) = 0$ в противном случае. Найдите плотность распределения величины $\xi = X + Y$.

43. Случайные величины X и Y независимы и имеют равномерное распределение на (0,1). Пусть $V = X + Y, W = \dfrac{X}{Y}$. Найдите совместную плотность распределения V и W. Будут ли V и W независимы?

44. Случайные величины V и W определены как: $V = X^2 + Y^2, W = X^2 - Y^2$, где X и Y независимы и имеют стандартное нормальное распределение каждая. Найдите совместную плотность распределения V и W. Будут ли V и W независимы?

45. Непрерывные случайные величины X и Y независимы и имеют показательное распределение с параметром μ. Пусть $V = X + Y, W = \dfrac{X}{X+Y}$. Найдите совместную плотность распределения V и W. Докажите, что V и W независимы.

46. Непрерывные случайные величины X и Y имеют совместную плотность распределения $f(x,y)=cxe^{-\frac{1}{2}x(1+y^2)}$ для $x,y>0$ и $f(x,y)=0$ в противном случае. а) Найдите константу c. б) Найдите маргинальные плотности X и Y. в) Докажите, что случайные величины $Y\sqrt{X}$ и X независимы.

47. Пусть случайная величина ξ_1 имеет показательное распределение с плотностью $f_1(x)=e^{-x},\ x>0,$ а ξ_2 – арксинус-распределение с плотностью $f_2(x)=\dfrac{1}{\pi\sqrt{1-x^2}},\ |x|<1,$ ξ_1 и ξ_2 – независимы. Найти распределение случайной величины $\eta=\dfrac{\xi_1}{\xi_2}.$

48. Пусть случайные величины ξ_1 и ξ_2 независимы и одинаково распределены с плотность распределения $f(x)=\dfrac{2}{\pi}\dfrac{x^4}{(1+x^2)(1+x^4)},\ -\infty<x<+\infty,$ каждая и пусть $Y=\dfrac{\xi_1}{\xi_2}$. Покажите, что случайная величина Y имеет распределение Коши с плотностью $p(y)=\dfrac{1}{\pi}\cdot\dfrac{1}{1+y^2}$.

49. а) Пусть X_1 и X_2 – две независимых случайных величин, каждая из которых имеет показательное распределение с плотностью $f(x)=e^{-x},\ x>0;=0,\ x\le 0.$ Покажите, что если случайная величина $Y=X_1/X_2$, то Y имеет распределение Фишера $F_{2,2}$ с функцией распределения $F(y)=\dfrac{y}{1+y},\ y\ge 0,$ и плотностью распределения $p(y)=\dfrac{1}{(1+y)^2},\ y\ge 0.$

Пусть X_1 и X_2 – две независимых случайных величин, каждая из которых имеет распределение с плотностью б) $x^{-2}\exp(-x)\mathbf{1}(x>0);$ в) $(1+x^2)^{-3/2}\mathbf{1}(x>0);$ г) $x(1+x^2)^{-3/2}\mathbf{1}(x>0).$

Показать, что с.в. Y имеет распределение $F_{2,2}$.

50. Пусть $N_v(x)=\dfrac{x^v e^{-x^2/2}}{\sqrt{2\pi}\,2^{v/2}}\dfrac{\Gamma(v/2+1)}{\Gamma(v+1)}\,{}_1F_1(\frac{1}{2}v;v+1;\frac{1}{2}x^2),\ x\ge 0,$ где

$$\frac{\Gamma(v/2+1)}{\Gamma(v+1)}\,{}_1F_1\left(\frac{1}{2}v;v+1;\frac{1}{2}x^2\right)=\sum_{m=0}^{\infty}\frac{\Gamma(v/2+m)}{m!\,\Gamma(v+1+m)}\frac{x^{2m}}{2^m}.$$

Покажите, что если с.в. U,V – независимы и имеют плотности $N_v(x)$, то $\xi=U/V$ имеет плотность распределения $\dfrac{1}{\pi(1+y^2)}$.

51. Пусть X и Y две независимых случайных величины с плотностями

$$f_X(x)=\sum_{i=1}^{n}A_i\alpha_i e^{-\alpha_i x},\ x>0\ \ (\alpha_i>0),\ \text{и}\ \ f_Y(y)=\sum_{p=1}^{m}B_p\beta_p\,e^{-\beta_p y},\ y>0\ \ (\beta_p>0),$$

причем $A_i=\Pi_{j=1,\,j\ne i}^{n}\left(\frac{\alpha_j}{\alpha_j-\alpha_i}\right)$ и $B_p=\Pi_{j=1,\,j\ne p}^{n}\left(\frac{\beta_j}{\beta_j-\beta_p}\right)$. Покажите, что плотность распределения величины $Z=X/Y$ равна $f_Z(t)=\sum_{i=1}^{n}\sum_{p=1}^{m}A_iB_p\dfrac{\frac{\beta_p}{\alpha_i}}{\left(t+\frac{\beta_p}{\alpha_i}\right)^2}.$

52. Обозначим: 1) распределение Вейбулла: $G_{1,\alpha}=G_{1,\alpha}(x)=1-\exp(-x^{\alpha}),\ x\ge 0,\ \alpha>0$ с плотностью $g_{1,\alpha}(x)$; 2) распределение Парето: $G_{2,\alpha}=G_{2,\alpha}(x)=1-x^{-\alpha},\ x\ge 1,\ \alpha>0$ с плотностью $g_{2,\alpha}(x)$. Пусть случайные величины X и Y независимы, а $f(z)$ – плотность распределения величины $Z = X \cdot Y$. Покажите, что если 1) величина X распределена по закону $G_{1,\alpha}$, а величина Y распределена по закону $G_{1,\beta}$,то $f(z) = C_1 z^{\mu} \exp(-C_2 z^{\mu})(1+o(1)),\ \mu = \dfrac{\alpha\beta}{\alpha+\beta}$ при $z \to \infty$; 3) величина X распределена по закону $G_{1,\alpha}$, а величина Y распределена по закону $G_{2,\beta}$, то $f(z) = \beta \Gamma\left(1+\dfrac{\beta}{\alpha}\right) z^{-1-\beta}(1+o(1)),\ z \to \infty;$

<u>Замечание.</u> Использовать, что при больших x $\gamma(\alpha, x) = \int\limits_0^x t^{\alpha-1} e^{-t} dt = \Gamma(\alpha) - x^{\alpha-1} e^{-x} \left(1+O\left(\dfrac{1}{x}\right)\right)$

4) величина X распределена по закону $G_{2,\alpha}$, а Y – по закону $G_{2,\beta}$,то

$$f(z) = \begin{cases} \dfrac{\alpha\beta}{\alpha-\beta}(z^{-1-\alpha} - z^{-1-\beta}), \ \textit{если} \ \alpha \neq \beta, \\ \alpha^2 z^{-1-\alpha} \ln z, \ \textit{если} \ \alpha = \beta. \end{cases}$$

53. Пусть $\boldsymbol{X} = (X_1, X_2), \boldsymbol{Y} = (Y_1, Y_2)$ – независимые и одинаково распределенные $N(\mathbf{0}, \Sigma)$, где $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, -1 < \rho < 1,$ – ковариационная матрица. Докажите, что с.в. $Z = \beta_2 \dfrac{X_2}{Y_2} + \beta_1 \dfrac{X_1}{Y_1}, \beta_1 \geq 0, \beta_2 \geq 0, \beta_1 + \beta_2 = 1$ имеет распределение Коши С(0,1). Рассмотреть случай когда β_1 и β_2 неслучайны, а также, когда они случайны.

54. Пусть случайная величина X_1 имеет Максвелла распределение с плотностью $f_1(x) = \sqrt{\dfrac{2}{\pi}} a^{3/2} x^2 e^{-\frac{ax^2}{2}},\ x > 0,\ a > 0,$ а случайная величина X_2 имеет распределение Релея с плотностью $f_2(y) = \dfrac{y^2}{\sigma^2} e^{-\frac{y^2}{2\sigma^2}},\ y > 0,\ \sigma > 0,$ и функцией распределения $F(y) = 1 - e^{-\frac{y^2}{2\sigma^2}}$ и они независимы. Покажите, что случайная величина $\xi = X_1 \cdot X_2$ имеет функцию распределения $F_{\xi}(z) = 1 - e^{-\frac{z\sqrt{a}}{\sigma}} \left(1 + \dfrac{z\sqrt{a}}{\sigma}\right),\ z > 0,\ a > 0,\ \sigma > 0.$

<u>*Замечание*</u>. Если $a = \sigma = 1$, то $F_{\xi}(z) = 1 - e^{-z}(1+z)$ и плотностью распределения $f_{\xi}(z) = ze^{-z},\ z > 0,$ – гамма распределение.

55. а) Пусть случайная величина $X \in N(0,1)$. Покажите, что величина $Y = \dfrac{1}{X^2}$ имеет распределение Леви с плотностью $p_Y(x) = \dfrac{1}{\sqrt{2\pi}} x^{-3/2} \exp\left(-\dfrac{1}{2x}\right),\ x > 0.$

б) Пусть X_1, X_2 – две независимых нормальных $N(0,1)$ случайных величины. Покажите, что величина $Z = \dfrac{1}{4}\left(\dfrac{1}{X_1^2} + \dfrac{1}{X_2^2}\right)$ имеет распределение Леви.

56. Пусть случайные величины $X_1, X_2, ..., X_n$ – независимы и имеют равномерное распределение на интервале $(0,1)$. Найдите распределение величины $Z = X_1 \cdot X_2 \cdot ... \cdot X_n$.

57. С.В. X имеет гамма-распределение (обозначение $X \in G(\alpha, p), \alpha > 0, p > 0$), если ее плотность распределения равна $f(x) = \dfrac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}, 0 < x < \infty$. Покажите, что если $X_i \in G(\alpha, p_i)$ и все $X_i, i = 1,..,k,$ независимы, то $X_1 + ... + X_k \in G(\alpha, p_1 + ... p_k)$.

58. Покажите, что если $X_1 \in G(\alpha, p_1)$ и $X_2 \in G(\alpha, p_2)$ независимы, то $U_1 = X_1 + X_2$ и $U_2 = \dfrac{X_1}{X_2}$ независимы и распределены: U_1 как $G(\alpha, p_1 + p_2)$, а U_2 имеет плотность $f(u_2) = \dfrac{\Gamma(p_1 + p_2)}{\Gamma(p_1)\Gamma(p_2)} \cdot \dfrac{u_2^{p_1 - 1}}{(1 + u_2)^{p_1 + p_2}}, u_2 > 0$.

59. Пусть $X \in G(\alpha, p_1), Y \in G(\alpha, p_2)$ и независимы. Тогда случайная величина $V = \dfrac{X}{X+Y}$ имеет бета-распределение с плотностью $\dfrac{\Gamma(p_1 + p_2)}{\Gamma(p_1)\Gamma(p_2)} v^{p_1 - 1}(1 - v)^{p_2 - 1}, 0 < v < 1$.

60. (Распределение Фишера) Найти функцию распределения частного $\varsigma = \dfrac{\xi}{\eta}$, где ξ и $\eta-$ независимые случайные величины, причем ξ распределена как χ_m^2 / m, а $\eta = \chi_n^2 / n$.

61. Пусть с.в. $X_1 \in N(0,1)$, а $X_2 \in \mathscr{R}(0,1)$ и независимы. Покажите, что $Y = X_1 / X_2$ имеет плотность $f_Y(y) = y^{-2}(1 - \exp(-y^2/2)), y \in R,$ а $Y_1 = 1/Y \in N(0,1)$.

62. Покажите, что если величины, $X_1 \in N(a_1, \sigma_1^2), X_2 \in N(a_2, \sigma_2^2),$ независимы, $Y = X_1 + X_2$, то

$$f_Y(y) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-a_1)^2}{2\sigma_1^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y-x-a_2)^2}{2\sigma_2^2}\right) dx = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left(-\frac{(y-a_1-a_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right),$$

т.е. $X_1 + X_2 \in N(a_1 + a_2, \sigma_1^2 + \sigma_2^2)$.

63. Пусть X_1 и X_2 независимые случайные величины с плотностями распределения соответственно ($X_i \in B^I(\alpha_i, \beta_i)$), $f_i(x) = \dfrac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x^{\alpha_i - 1}(1-x)^{\beta_i - 1}, \ 0 < x < 1, i = 1, 2.$

Покажите, что плотность распределения величины $Z = X_1 X_2$ равна

$$f_Z(z) = \frac{\Gamma(\alpha_1 + \beta_1)\Gamma(\alpha_2 + \beta_2)}{\Gamma(\beta_1 + \beta_2)\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1 - 1}(1-z)^{\beta_1 + \beta_2 - 1} {}_2F_1(\beta_2, \alpha_1 + \beta_1 - \alpha_2; \beta_1 + \beta_2; 1 - z),$$

где $0 < z < 1$. Кроме того, если $\alpha_2 = \alpha_1 + \beta_1,$ то $Z = X_1 X_2 \in B^I(\alpha_1, \beta_1 + \beta_2)$.
Здесь

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-zt)^{-a}\, dt = 1 + \frac{ab}{c} \cdot \frac{z}{1!} + \frac{a(a+1)b(b+1)}{c(c+1)} \cdot \frac{z^2}{2!} + ..., |z| < 1.$$

64. Пусть X_1 и X_2 независимые случайные величины $X_1 \in B^I(\alpha_1, \beta_1)$ и $X_2 \in B^{II}(\alpha_2, \beta_2)$, т.е. $f_2(x) = \dfrac{\Gamma(\alpha_2 + \beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} x^{\alpha_2 - 1}(1-x)^{-\alpha_2 - \beta_2}; \ 0 < x.$ Покажите, что плотность распределения величины $Z = X_1 X_2$ равна

$$f_Z(z) = \frac{\Gamma(\beta_1 + \beta_2)\Gamma(\alpha_1 + \beta_1)\Gamma(\alpha_2 + \beta_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\beta_2)} z^{\alpha_2 - 1}(1+z)^{-(\alpha_2 + \beta_2)} {}_2F_1\left(\beta_1, \alpha_2 + \beta_2; \alpha_1 + \beta_1 + \beta_2; \frac{1}{1+z}\right), z > 0.$$

65. Пусть X_1 и X_2 независимые случайные величины $X_1 \in B^I(\alpha_1, \beta_1)$ и $X_2 \in B^I(\alpha_2, \beta_2)$.

Покажите, что плотность распределения величины $Z = X_1 / X_2$ равна

$$\frac{\Gamma(\beta_2)\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_1+\alpha_2+\beta_2)\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(\alpha_2)\Gamma(\beta_2)} z^{\alpha_1-1}\,{}_2F_1\left(\alpha_1+\alpha_2, 1-\beta_1; \alpha_1+\alpha_2+\beta_2; z\right), 0<z<1, \quad \text{и}$$

$$\frac{\Gamma(\beta_1)\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_1+\alpha_2+\beta_1)\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(\alpha_2)\Gamma(\beta_2)} z^{-\alpha_2-1}\,{}_2F_1\left(\alpha_1+\alpha_2, 1-\beta_2; \alpha_1+\alpha_2+\beta_1; \frac{1}{z}\right), z>1.$$

66. Пусть X_1 и X_2 независимые случайные величины $X_1 \in B^I(\alpha_1, \beta_1)$ и $X_2 \in B^I(\alpha_2, \beta_2)$. Покажите, что плотность распределения величины $T = X_1 / (X_1 + X_2)$ равна

$$\frac{\Gamma(\beta_2)\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_1+\alpha_2+\beta_2)\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(\alpha_2)\Gamma(\beta_2)} t^{\alpha_1-1}(1-t)^{-\alpha_1-1}\,{}_2F_1\left(\alpha_1+\alpha_2, 1-\beta_1; \alpha_1+\alpha_2+\beta_2; \frac{t}{1-t}\right), \text{ для } 0<t<1/2,$$

$$\frac{\Gamma(\beta_1)\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_1+\alpha_2+\beta_1)\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(\alpha_2)\Gamma(\beta_2)} t^{-\alpha_2-1}(1-t)^{\alpha_2-1}\,{}_2F_1\left(\alpha_1+\alpha_2, 1-\beta_2; \alpha_1+\alpha_2+\beta_1; \frac{1-t}{t}\right), \text{ для } 1/2<t<1.$$

67. Покажите, что если X и Y – независимые нормально распределенные с. в. с нулевыми средними и дисперсиями $\sigma^2 > 0$ и $\tau^2 > 0$ соответственно, тогда случайная величина $Z = \frac{XY}{\sqrt{X^2+Y^2}}$ нормальна с математическим ожиданием нуль и дисперсией $\frac{\sigma^2\tau^2}{(\sigma+\tau)^2}$.

68. а) Предположим, что X_1 и X_2 – независимые случайные величины одинаково распределенные с функцией распределения $F(x) = 1 - e^{-\lambda x}, x>0, \lambda>0$. Покажите, что X_1 и $|X_1 - X_2|$ – одинаково распределены.

б) Предположим, что X и Y – независимые случайные величины с функцией распределения $F(x) = x^\lambda, 0<x<1$. Покажите, что величины X и $\min\left(\frac{X}{Y}, \frac{Y}{X}\right)$ одинаково распределены.

69. Пусть случайная величина X имеет функцию распределения $F(x) = \frac{1}{2} + \frac{\arcsin x}{\pi}, -1<x<1$. Покажите, что X^2 и $\frac{1+X}{2}$ – одинаково распределены.

70. По заданной таблице найти условное распределение и условные математические ожидания.

$X \quad Y$	-1	0	1
-1	$1/8$	$1/12$	$1/24$
1	$5/24$	$1/6$	$1/8$

Найти закон распределения величины $Z = X + Y$.

71. Случайная величина U имеет равномерное распределение на $(0,1)$, $0<p<1$ – задано. Покажите: случайная величина $X = \left[\frac{\ln U}{\ln(1-p)}\right]$ имеет геометрическое распределение.

72. Дана плотность распределения $f(x)$ случайной величины X, определенная для $0<x<\infty$. Найти плотность распределения случайной величины $y = \ln X$.

73. Найти плотность распределения случайной величины $Z = aX^2$ если $X \in N(0, \sigma^2), a>0$.

74. Найти распределение суммы двух независимых случайных величин X_1 и X_2, если $X_i \in \mathscr{P}(\lambda_i), i=1,2$.

75. Пусть $X \in N(0,1), Y \in N(0,1)$ – независимы. Найти распределение $Z = X/Y$.

76. X и Y независимы и равномерно распределены со значениями $0, 1, 2$. Найти распределение случайной величины $Z = X - Y$.

77. Покажите, что если $Y \in \mathscr{P}(\lambda)$, то $P(Y \le k) = P(\chi^2_{2k+2} \ge 2\lambda)$.

Замечание. Можно от обеих частей взять производную!

78 . Дискретная случайная величина X имеет ряд распределения:

	–2	–1	0	1	2
p_i	0.1	0.2	0.3	0.3	0.1

Построить ряд распределения случайных величин $Y = X^2 + 1, Y = |X|$.

79. Показать, что если случайная величина X имеет непрерывную функцию распределения $F(x) = \mathbf{P}(X < x)$, то случайная величина $Y = F(X)$ имеет равномерное распределение на $(0, 1)$.

80. Найти плотность распределения объема куба, ребро которого X – случайная величина, равномерно распределенная на интервале $(0, a)$.

81. В схеме Бернулли с вероятностью успеха p обозначим через τ_k $(k = 2, 3, \dots)$ разность номеров испытаний, при которых происходят k–й и $(k-1)$–й успехи, а τ_1 – число испытаний до первого успеха. Найти совместное распределение величин τ_1 и τ_2. Являются ли эти величины независимыми?

82. Найти плотность распределения случайной величины $Y = \dfrac{X_1}{X_1 + X_2}$, если X_1 и X_2 независимы и равномерно распределены на интервале $(0,1)$.

83. Найти характеристическую функцию $\varphi(t)$ дискретной случайной величины, подчиняющейся закону распределения Паскаля $\boldsymbol{P}(X = m) = \dfrac{a^m}{(a+1)^{m+1}}$, $a > 0$, $m = 0, 1, \dots$.

По $\varphi(t)$ найти $\boldsymbol{E}(X)$ и $\boldsymbol{D}(X)$.

84. Случайная величина X имеет плотность распределения

$$f(x) = \begin{cases} \alpha\, e^{-\alpha x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Найти ее характеристическую функцию и начальные моменты.

85. Характеристическая функция случайной величины X равна $\varphi(t) = \dfrac{e^{it}(1 - e^{itn})}{n(1 - e^{it})}$.

Какой случайной величине она соответствует? Найти ее закон распределения.

86. Дана характеристическая функция $\varphi(t) = \dfrac{1}{2e^{-it} - 1}$. Показать, что она соответствует дискретной случайной величине. Найти ее закон распределения.

87. Найти характеристическую функцию отрицательного биномиального распределения:
$\mathbf{P}(X = k) = C_{m+k-1}^{m-1}\, p^m q^k$, $k = 0, 1, \dots$, $0 < p < 1$, $q = 1 - p$, m – задано.

88. Случайные величины $X_1, X_2, \dots$ – независимы и имеют показательное распределение с параметром λ: $F(x) = \boldsymbol{P}(X < x) = 1 - e^{-\lambda x}$, $x > 0, i = 1, 2, \dots, \lambda > 0$. Найти $\boldsymbol{E}(X_1 + X_2 + \dots + X_n)^k$ при любом значении k и $n = 1, 2, \dots$.

89. Случайная величина X имеет начальные моменты $\mu_k = \boldsymbol{E}(X^k) = (k+1)!\, 2^k$, $k = 1, 2, \dots$. Найдите характеристическую функцию случайной величины X. Определите распределение величины X.

90. Случайная величина X имеет плотность распределения $f(x) = \dfrac{1}{4} x e^{-\frac{x}{2}}, x > 0.$

Найти ее характеристическую функцию и начальные моменты.

91. Пусть X и Y – независимые случайные величины. Приведите пример случайной величины $Z = \phi(X, Y)$, для которой $\boldsymbol{E}(Z \mid X) = X$ и $\boldsymbol{E}(Z \mid Y) = Y$.

92. Пусть $X=\begin{pmatrix}X_1\\X_2\\X_3\end{pmatrix}\in N\left(\begin{pmatrix}2\\3\\1\end{pmatrix},\begin{pmatrix}5&2&7\\2&5&7\\7&7&14\end{pmatrix}\right)$. а) Найдите распределение с. в. $Y_1=X_1+X_2-X_3$.

б) Найдите распределение случайной величины $Y_2=X_1+X_2+X_3$. в) Найдите $\boldsymbol{E}(Y_2\,|\,X_1=5,X_2=3)$. г) Найдите $\boldsymbol{E}(Y_2\,|\,X_1=5,X_2<3)$. д) Найдите $\boldsymbol{P}(Y_2<10\,|\,X_1=5,X_2<3)$.

93. Пусть $W_n=V_1^2+V_2^2+...+V_n^2$, где $\{V_j\}_{j=1}^n$ – независимые случайные величины, $V_j\in N(\mu_j,1)$. Покажите, что: 1) случайная величина W_n имеет нецентральное хи-квадрат распределение с параметром нецентральности $2\lambda=\sum_{j=1}^n\mu_j^2$ и плотностью распределения

$$f(x)=e^{-(x+\lambda)/2}\frac{1}{2}\left(\frac{x}{\lambda}\right)^{(n-2)/4}I_{(n-2)/2}(\sqrt{\lambda x}),\ x>0,\ \ I_\nu(z)=\sum_{k=0}^{\infty}\frac{(z/2)^{2k+\nu}}{k!\Gamma(k+\nu+1)};$$

2) моменты этого распределения равны: $\boldsymbol{E}(W_n)=n+2\lambda,\ \boldsymbol{D}(W_n)=2n+4\lambda;$

3) производящая функция моментов равна $\boldsymbol{E}(e^{W_n\theta})=\dfrac{1}{(1-2\theta)^{n/2}}\exp\left(\dfrac{2\lambda\theta}{1-2\theta}\right).$

7. Ответы

1. $f_1(x)=\dfrac{x}{\sigma^2}\exp\left(-\dfrac{x^2}{2\sigma^2}\right),\ x>0.$ **2.** $m_{2k}=\dfrac{4k-2}{k+1},\ k=1,2,...$. **3.** Если X_2 имеет распределение (8), то $X_1=\dfrac{1}{2}\left(\dfrac{X_2}{\sqrt{1+X_2^2}}+1\right)$ имеет распределение (7). **5.** $C=\dfrac{\sqrt{2}}{\pi}$. **6.** $f_Y(y)=\dfrac{\ln(y^2)}{\pi^2(y^2-1)},\ y\in\mathbf{R}.$

8. $f_Y(t)=\dfrac{1}{\pi(1+t^2)}$. **9.** $f_\xi(x)=f_1(x)=\dfrac{1}{\pi}\mathrm{sech}(x),\ f_{\xi+\eta}(x)=f_2(x)=\dfrac{x}{\pi^2\mathrm{sh}(x)}$,

$$\mathrm{sech}(x)=\frac{1}{\mathrm{ch}(x)}=\frac{2}{e^x+e^{-x}},\ \mathrm{csch}(x)=\frac{1}{\mathrm{sh}(x)}=\frac{2}{e^x-e^{-x}}.$$

10. а) $f_\varsigma(z)=\dfrac{\sqrt{1-\rho^2}}{\pi(z^2-2\rho z+1)},\ \ \varsigma=\dfrac{y_2}{y_1}=\dfrac{x_2-a_2}{x_1-a_1}\cdot\dfrac{\sigma_1}{\sigma_2}.$

б) $$f_{\frac{x_1}{x_2}}(z)=\frac{\sqrt{1-\rho^2}}{\pi\sigma_1\sigma_2\gamma^2}\exp\left(-\frac{\tau^2}{2\,(1-\rho^2)}\right)+\frac{\eta\lambda}{\sqrt{2\pi}\sigma_1\sigma_2\gamma^3}\left(2\Phi\left(\frac{\eta}{\gamma\sqrt{1-\rho^2}}\right)-1\right),$$

$$\gamma^2=\frac{z^2}{\sigma_2^2}-2\rho\frac{z}{\sigma_1\sigma_2}+\frac{1}{\sigma_1^2},\ \eta=\frac{z}{\sigma_2}\left(\frac{a_2}{\sigma_2}-\rho\frac{a_1}{\sigma_1}\right)+\frac{1}{\sigma_1}\left(\frac{a_1}{\sigma_1}-\rho\frac{a_2}{\sigma_2}\right),\tau^2=\frac{a_2^2}{\sigma_2^2}-2\rho\frac{a_1a_2}{\sigma_1\sigma_2}+\frac{a_1^2}{\sigma_1^2},$$

$\lambda=\exp\left(-\dfrac{\eta^2}{2(1-\rho^2)\gamma^2}+\dfrac{\tau^2}{2(1-\rho^2)}\right)$. в) $f_{\left(\frac{x_1}{x_2}\right)^2}(x)=\dfrac{1}{2\sqrt{x}}\left(f_{\frac{x_1}{x_2}}(\sqrt{x})+f_{\frac{x_1}{x_2}}(-\sqrt{x})\right).$

11. Распределение Пуассона с параметром $\lambda=\ln\dfrac{1}{p}>0$.

18. $\mathbf{E}(T)=n\left(1+\dfrac{1}{2}+...+\dfrac{1}{n}\right).$ **20.** $f_Y(y)=\dfrac{1}{2}(1+x^2)^{-3/2}.$

22. $$f(z)=\frac{2pz^{(d_1+d_2)/2-1}}{\Gamma(d_1/p)\Gamma(d_2/p)(a_1a_2)^{(d_1+d_2)/2}}K_{(d_2-d_1)/p}\left[2\left(\frac{z}{a_1a_2}\right)^{p/2}\right],\ z>0.$$

27. $f_Y(y)=\dfrac{1}{\pi(1+y^2)}$. **30.** $\mu(x)=c_1(x+1)-1+O\left(\left(\dfrac{2e}{x}\right)^{x-3/2}\right)$, где $c_1=0,7475979203...$.

36. $\eta\in N(0,1)$. **37.** $f_Z(y)=\dfrac{1}{\pi y}e^{-\frac{y^2}{2}}\operatorname{erfi}\left(\dfrac{y}{\sqrt{2}}\right)$, где $\operatorname{erfi}(x)=\dfrac{2}{\sqrt{\pi}}\int\limits_0^x e^{v^2}dv$.

83. $\varphi(t)=\dfrac{1}{1+a(1-e^{it})}$, $\boldsymbol{E}(X)=a$, $\boldsymbol{D}(X)=a(a+1)$. **84.** $\varphi(t)=\dfrac{\alpha}{\alpha-it}$, $\mu_k=k!\alpha^{-k}$.

85. Дискретная. $\boldsymbol{P}(X=k)=\dfrac{1}{n}$, $k=1,2,...,n$. **86.** $\boldsymbol{P}(X=k)=\dfrac{1}{2^k}$, $k=1,2,...$.

87. $\varphi(t)=\left(\dfrac{p}{1-qe^{it}}\right)^m$, $q=1-p, 0<p<1$. **88.** $\mu_k=(k+1)!\lambda^{-k}$. **89.** $\varphi(t)=\dfrac{1}{(1-2it)^2}$.

90. $\mu_k=\dfrac{(k+1)!}{2^k}$. **91.** Если $\boldsymbol{E}(X)=\boldsymbol{E}(Y)=0$, то $Z=X+Y-\boldsymbol{E}(X)$.

92. а) $\boldsymbol{D}(Y_1)=0\Rightarrow X_1+X_2=X_3$, $\boldsymbol{P}(Y_1=4)=1$. б) $Y_2\in N(6,56)$. в) Найдя условное распределение $f(x_3\,|\,x_2,x_1)$, получим $\boldsymbol{E}(Y_2\,|\,X_1=5,X_2=3)=\boldsymbol{E}(X_3\,|\,X_1=5,X_2=3)+8=12$. г) и д) *Указание*. Найдите сначала условное распределение $f(y_2,x_2\,|\,x_1)$, обозначая $z_1=y_2, z_2=x_2, z_3=x_1$, найдя частный коэффициент корреляции, получим $\rho_{12\cdot 3}=1$.

8. Характеристические функции

8.1 Определение и свойства характеристических функций

Определение. *Пусть X – действительная случайная величина с функцией распределения $F(x)$. Характеристической функцией случайной величины X называется математическое ожидание случайной величины e^{itX}, где t – действительное число, а $i=\sqrt{-1}$, т.е.*

$$\varphi(t)=\varphi_X(t)=\mathbf{E}(e^{itX}).$$

Для дискретной случайной величины

$$\varphi_X(t)=\sum_k e^{itx_k}\cdot p_k \text{, где } p_k=\mathbf{P}(X=x_k).$$

Для непрерывной случайной величины

$$\varphi_X(t)=\int_{-\infty}^{+\infty} e^{itx} f(x)\,dx \text{, где } f(x) - \text{ плотность распределения с.в. } X.$$

При специальных предположениях относительно случайных величин наряду с характеристической функцией используются их аналоги:

- для целочисленной величины X – производящая функция вероятностей

$$\gamma(t)=\sum_{k=0}^{\infty} p_k t^k, \quad -1\le t\le 1;$$

- для положительной величины X – производящая функция моментов (преобразование Лапласа)

$$p(t)=\int_0^{\infty} e^{-tx}\,dF(x)$$

(если она существует для интервала t на действительной прямой, содержащего 0).

Производящая функция вероятностей полезна для определения факториальных моментов величины X:

$$\left.\frac{d^k\gamma(t)}{dt^k}\right|_{t=1}=\mathbf{E}(X(X-1)\cdot\ldots\cdot(X-k+1)).$$

ТЕОРЕМА 1 (Единственность). *Пусть $F_1(x)$ – функция распределения , а $\varphi_1(t)$ – характеристическая функция случайной величины , $F_2(x)$ – функция распределения , а $\varphi_2(t)$ – характеристическая функция случайной величины X_2. В таком случае $\varphi_1(t)=\varphi_2(t)$ тогда и только тогда, когда $F_1(x)=F_2(x)$.*

(Иными словами функция распределения однозначно определяется своей характеристической функцией.)

Дополнением к теореме единственности для характеристических функций служат явные выражения $\mathbf{P}^X(A)$ для тех или иных классов множеств через характеристические функции:

$$\mathbf{P}^X(A)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\left(\int_A e^{-itx}\,dx\right)\varphi_X(t)\,dt.$$

В случае, когда $\varphi(t)$ – абсолютно интегрируема, то случайная величина X – непрерывна, а ее плотность $f(x)$ через характеристическую функцию определяется следующим образом:

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\varphi(t)\,dt.$$

Функция распределения через характеристическую функцию определяется как следующий предел:

$$F(\beta)-F(\alpha)=\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-it\alpha}-e^{-it\beta}}{it}\varphi(t)\,dt .$$

Если $\xi -$ целочисленная с.в. и $\varphi(t)$ её характеристическая функция, то

$$\boldsymbol{P}(\xi=k)=\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{-itk}\varphi(t)\,dt .$$

ТЕОРЕМА 2. *Характеристическая функция* $\varphi(t)$
(i) *определена,* (ii) *равномерно непрерывна для любого действительного* t *,*
(iii) *ограничена. Более точно,*

$$|\ \varphi(t)\ |\le\varphi(0)=1 \quad \textit{для любого } t\in R^1 .$$

(iv) $\varphi(t)$ *неотрицательно определена, т.е. если* $a_{ij}=\varphi(t_i-t_j)$, $1\le i,\ j\le n$, $(\xi_1,\dots,\xi_n)\in C^n$,

то $\sum\limits_{j=1}^{n}\sum\limits_{k=1}^{n}\xi_j\overline{\xi}_k a_{jk}\ge 0$.

(v) *Х.ф. эрмитова, т.е.* $\varphi(-t)=\overline{\varphi(t)}$*, для любого* $t\in R^1$.

(vi) $\begin{cases}\varphi_{\lambda X}(t)=\varphi_X(\lambda t),\\ \varphi_{X+a}(t)=e^{ita}\varphi_{\lambda X}(t),\end{cases}$ *для любых* $\lambda, a, t\in R^1$.

(vii) *Если* $(X,Y)-$ *пара независимых с.в., то*

$$\varphi_{X+Y}(t)=\varphi_X(t)\cdot\varphi_Y(t)$$

(viii) *Если* $(X,Y)-$ *пара независимых с.в., то* $\boldsymbol{E}(\varphi_X(Y))=\boldsymbol{E}(\varphi_Y(X))=\varphi_{XY}(1)$.

Доказательство.

(i) + (iii) Действительно, существование следует из свойств интеграла и

$$|\ e^{itx}\ |=1, \text{ для любого } t\in R^1 .$$

(ii) Докажем равномерную непрерывность х. ф. Имеем

$$|\ \varphi(t+h)-\varphi(t)\ |=\left|\ \int_{-\infty}^{\infty}\left(\ e^{i(t+h)x}-e^{itx}\right)dF\ \right|\le\int_{-\lambda}^{\lambda}|\ e^{i(t+h)x}-e^{itx}\ |\ dF+$$

$$+\int_{|x|>\lambda}|\ e^{i(t+h)x}-e^{itx}\ |\ dF\le\ 2\mathbf{P}(A_\lambda)+\int_{-\lambda}^{\lambda}|\ e^{i(t+h)x}-e^{itx}\ |\ dF\ , \text{ где } A_\lambda=(|\ x\ |>\lambda\ ).$$

Поскольку $|\ e^{ib}-e^{ia}\ |=\left|\ i\int_a^b e^{iz}dz\ \right|\le|\ b-a\ |$, то $|\ \varphi(t+h)-\varphi(t)\ |\le 2\mathbf{P}(|X|>\lambda)+\lambda h,$

и выбирая $\lambda>0$ так, чтобы $\mathbf{P}(A_\lambda)<\dfrac{\varepsilon}{3}$ для заданного $\varepsilon>0$ и $h=\dfrac{\varepsilon}{3\lambda}$, получим

$$|\ \varphi(t+h)-\varphi(t)\ |<\varepsilon .$$

(iv) $\sum\limits_{j=1}^{n}\sum\limits_{k=1}^{n}\xi_j\overline{\xi}_k a_{jk}=\boldsymbol{E}\left(\sum\limits_{j=1}^{n}\xi_j e^{it_jX}\overline{\left(\sum\limits_{k=1}^{n}\xi_k e^{it_kX}\right)}\right)=\boldsymbol{E}\left(\left|\sum\limits_{j=1}^{n}\xi_j e^{it_jX}\right|^2\right)\ge 0$.

(v) (Доказать самим). Отсюда вытекает, что если х.ф. действительна, то она четная, а случайная величина X имеет симметричное распределение.

(vi) (Доказать самим).

(vii) В самом деле

$\varphi_{X+Y}(t)=\mathbf{E}\ (e^{it(X+Y)})=\mathbf{E}\ (e^{itX}\cdot e^{itY})=\mathbf{E}\ (e^{it\,X})\mathbf{E}\ (e^{itY})=\varphi_X(t)\cdot\varphi_Y(t),$

поскольку из независимости (X,Y) следует независимость величин e^{itX} и e^{itY}. Отметим, что это свойство не является характеристикой независимости. В самом деле, можно построить пару (X, Y), зависимую, такую, что х.ф. $\varphi_{X+Y}(t), \varphi_X(t), \varphi_Y(t)$ будут связаны соотношением (1).

(viii) (Доказать самим).

Обратная теорема в общем случае неверна.

Пример 6. Пусть (X, Y) имеет плотность распределения

$$f(x, y) = \frac{1}{4}(1 + xy(x^2 - y^2)), \quad |x| < 1, \ |y| < 1.$$

Тогда $f_1(x) = \int\limits_{-\infty}^{\infty} f(x, y)\,dx = \frac{1}{2} + \frac{1}{4}\int\limits_{-1}^{1} y\,dy \int\limits_{-1}^{1} x\,(x^2 - y^2)\,dx = \frac{1}{2}, \quad f_2(x) = \frac{1}{2}$ и

$f(x, y) \neq f_1(x) f_2(y)$, т.е. случайные величины X и Y – зависимы.

Покажем, что $\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t)$. Действительно,

$$\varphi_1(t) = \varphi_2(t) = \frac{e^{it} - e^{-it}}{2it}, \quad \varphi(t, t) = \frac{1}{4}\int\limits_{-1}^{1} dy \int\limits_{-1}^{1} dx (1 + xy(x^2 - y^2)) e^{itx} e^{ity} =$$

$$= \frac{1}{4}\int\limits_{-1}^{1} e^{ity} dy \ (\int\limits_{-1}^{1} e^{itx} dx + \int\limits_{-1}^{1} (\cos(tx) + i\sin(tx))(x^3 y - xy^3)\,dx = \frac{e^{it} - e^{-it}}{2it} \cdot \frac{e^{it} - e^{-it}}{2it},$$

так как

$$\int\limits_{-1}^{1} (\cos(tx) + i\sin(tx))(x^3 y - xy^3)\,dx = i\int\limits_{-1}^{1} \sin(tx)(x^3 y - xy^3)\,dx = = I_3 y - iI_1 y^3,$$

$$\int\limits_{-1}^{1} (\cos(ty) + i\sin(ty))(iI_3 y - iI_1 y^3)\,dx = i^2(I_3 I_1 - I_1 I_3) = 0.$$

Покажем, что $\int\limits_{-\infty}^{\infty} f^2(x)\,dx = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} |\varphi(t)|^2\,dt$.

Действительно, рассмотрим независимые случайные величины X_1, X_2 с плотностью распределения $f(x)$. Тогда характеристическая функция их разности будет равна

$\varphi_{X_1 - X_2}(t) = |\varphi_{X_1}(t)|^2 = |\varphi(t)|^2$. По формуле свертки $f_{X_1 - X_2}(x) = \int\limits_{-\infty}^{\infty} f(x + y) f(y)\,dy$.

Запишем для нее формулу обращения,

$$f_{X_1 - X_2}(x) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} e^{-itx} |\varphi(t)|^2\,dt = \int\limits_{-\infty}^{\infty} f(x + y) f(y)\,dy.$$

Подставляя $x = 0$, получаем требуемое утверждение.

С помощью характеристической функции можно оценить вероятность $\mathbf{P}(\ |X| > A)$.

Лемма 1. *Существует константа* K *такая, что*

$$\mathbf{P}(\ |X| > A) \le K \cdot A\int\limits_{0}^{1/A} (1 - \mathrm{Re}\ \varphi(t))\,dt, \quad \textit{для любого} \quad A > 0, \quad K = 1 - \sin 1 \ge 1/7.$$

Д о к а з а т е л ь с т в о . Имеем:

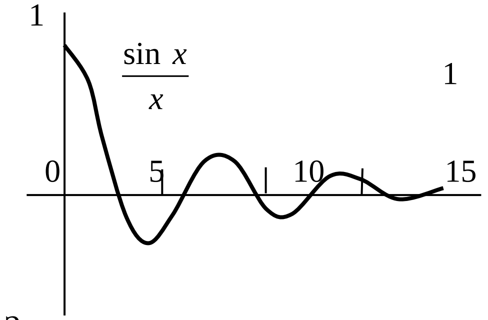
$$A\int\limits_{0}^{1/A} (1 - \mathrm{Re}\ \phi(t))\,dt = A\int\limits_{0}^{1/A} (1 - \int\limits_{-\infty}^{\infty} \cos(tx)\ dF(x))\,dt =$$

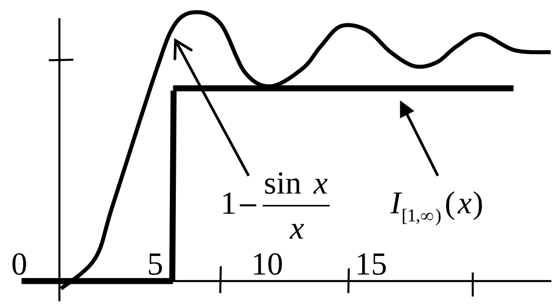
$$= A\int\limits_{0}^{1/A} \int\limits_{-\infty}^{\infty} (1 - \cos(tx)\ dF(x))\,dt = A\int\limits_{-\infty}^{\infty} \int\limits_{0}^{1/A} (1 - \cos(tx)\,dt)\ dF(x) =$$

$$= \int\limits_{-\infty}^{\infty} A\left(\frac{1}{A} - \frac{\sin(x/A)}{x}\right) dF(x) = \int\limits_{-\infty}^{\infty} \left(1 - \frac{\sin(x/A)}{x/A}\right) dF(x) \ge$$

$$\geq C \int\limits_{-\infty}^{\infty} I_{[-1,1]^c}\left(\frac{x}{A}\right) dF(x) = C\left(\int\limits_{A}^{\infty} dF(x) + \int\limits_{-\infty}^{-A} dF(x)\right) = C\ \mathbf{P}(\,|\,X\,| \geq A),$$

$$C = 1/K.$$





Здесь

$$\left(1 - \frac{\sin x}{x}\right) \geq \inf_{|x|\geq 1}\left(1 - \frac{\sin x}{x}\right) = 1 - \sin 1 \approx 0.158529 \geq \frac{1}{7} = 0.142857142.$$

Замечание. Этому результату можно придать несколько иной вид:

$$\frac{1}{u}\int_{-u}^{u}(1-\varphi_X(t))\,dt \geq \boldsymbol{P}\left(|X| \geq \frac{2}{u}\right),\ u > 0.$$

По х.ф. $\varphi(t)$ можно узнать, какому типу случайной величины она принадлежит:

а) если $\limsup\limits_{|t|\to\infty} |\varphi(t)| = 1$, то X – дискретная случайная величина;

б) если $\limsup\limits_{|t|\to\infty} |\varphi(t)| = 0$, то X – непрерывная случайная величина.

Критерий непрерывности: функция распределения $F(x)$ абсолютно непрерывна тогда и только тогда, когда $\int\limits_{-\infty}^{\infty} |\,\varphi(t)\,|dt < \infty$.

В общем случае $\lim\limits_{T\to\infty}\dfrac{1}{2T}\int\limits_{-T}^{T} |\,\varphi(t)\,|^2\,dt = \sum\limits_j p_j^2$, p_j – скачки функции распределения $F(x)$.

<u>Определение.</u> Дискретное распределение называется решетчатым, если его точки разрыва принадлежат некоторому множеству, образующему арифметическую прогрессию, то есть множеству точек $a + jd$, где a и d – постоянные числа и j принимает целые значения.

Примерами решетчатых распределений являются биномиальное и пуассоновское распределения. Если $F(x)$ – решетчатое распределение, то его характеристическая функция равна

$$\varphi(t) = e^{ita}\sum_j p_j e^{itjd}.$$

Очевидно, что $\left|\varphi\left(\dfrac{2\pi}{d}\right)\right| = |e^{2\pi ai/d}| = 1$.

Следовательно, любое решетчатое распределение F обладает следующим свойством: существует действительное число $t_o \neq 0$ такое, что $|\,\varphi(t_o)\,| = 1$. Мы сейчас покажем, что это свойство является характеристическим для решетчатых распределений.

ТЕОРЕМА 3. *Характеристическая функция $\phi(t)$ соответствует решетчатому распределению тогда и только тогда, когда существует действительное число $t_o \neq 0$ такое, что $|\,\varphi(t_o)\,| = 1$.*

Доказательство. Предположим, что существует число $t_o \neq 0$ такое, что $|\varphi(t_o)|=1$. Это означает, что $\varphi(t_o)=e^{it_o\xi}$ для некоторого действительного ξ, или $\int\limits_{-\infty}^{\infty} e^{it_o x}\, dF(x) = e^{it_o\xi}$.

Отсюда следует, что $\int\limits_{-\infty}^{\infty} (1-\cos(t_o(x-\xi)))\, dF(x)=0$.

Так как функция $1-\cos(t_o(x-\xi))$ непрерывна и неотрицательна, то последнее равенство может иметь место только тогда, когда $F(x)$ является чисто дискретным распределением, точки разрыва которого содержатся в множестве нулей функции $1-\cos(t_o(x-\xi))$. Поэтому точки разрыва $F(x)$ необходимо имеют вид $\xi+2\pi s/t_o$, (s – целое). Следовательно, $F(x)$ – решетчатое распределение. Теорема доказана.

Заметим также, что для сингулярного распределения, построенного на основе Канторова совершенного множества, характеристическая функция равна

$$\varphi(t)=e^{it/2}\lim_{n\to\infty}\prod_{j=1}^{n}\cos\left(\frac{t}{3^j}\right).$$

ТЕОРЕМА 4. *Пусть* $c_i \geq 0$, , $\varphi_{X_i}(t)$ – *характеристическая функция с.в.* X_i *с ф.р.* $F_i(x)$. *Тогда* $\sum_{i=1}^{n} c_i\varphi_{X_i}(t)$ – *х.ф. с.в.* Y *с ф.р.* $\sum_{i=1}^{n} c_i F_i(x)$.

Доказать самим.

ТЕОРЕМА 5. *Если* φ – *х.ф., то х.ф. будут* φ^n – (n *целое* ≥ 1), $\overline{\varphi}$, $|\varphi|^2$, $\mathcal{Re}\,\varphi$.

Д о к а з а т е л ь с т в о . Пусть φ – х.ф. с.в. X .

(i) φ^n – х.ф. суммы n независимых и одинаково с X распределенных с.в.;

(ii) $\overline{\varphi}$ – х.ф. с.в. $-X$;

(iii) $|\varphi|^2$ – х.ф. с.в. $(X-X')$, где X' – с.в., независимая и одинаково распределенная с X (разность $(X-X')$ называют симметризацией X);

(iv) $\mathcal{Re}\,\varphi=\dfrac{\varphi+\overline{\varphi}}{2}$ – есть характеристическая функция, поскольку является выпуклой комбинацией характеристических функций.

Пример 7. Пусть с.в. X имеет распределение Бернулли, т.е.

$$X=\begin{cases}1, & \text{с вероятностью } p,\\ 0, & \text{с вероятностью } q,\end{cases} \quad , \; p+q=1 \quad .$$

Тогда $\varphi_X(t)=p\,e^{it}+q$.

Пример 8. Пусть $X\in B(n,\,p)$, т.е. , $m=0,1,\ldots,n$.

Тогда

$$\varphi_X(t)=\sum_{m=0}^{n} e^{itm}C_n^m p^m q^{n-m}=\left(p e^{it}+q\right)^n.$$

С другой стороны, если рассмотреть случайную величину $Y=X_1+\ldots+X_n$, где $X_i, 1\leq i\leq n$, – независимые с.в., имеющие распределение Бернулли, то $\varphi_Y(t)=(pe^{it}+q)^n$. В силу теоремы единственности случайные величины X и Y одинаково распределены (биномиальное распределение).

Пример 9. Пусть X имеет распределение Пуассона ($X\in\mathcal{P}(\lambda)$), т.е.

$$\boldsymbol{P}(X=m)=e^{-\lambda}\frac{\lambda^m}{m!} \qquad , \ m=0,1,\ldots \quad .$$

Тогда $\varphi_X(t)=\sum_{m=0}^{\infty}e^{itm-\lambda}\frac{\lambda^m}{m!}=e^{-\lambda}\sum_{m=0}^{\infty}\frac{1}{m!}(\lambda e^{it})^m=\exp(\lambda(e^{it}-1))$.

Пример 10. Пусть X имеет нормальное распределение $X\in N(0,1)$ с плотностью

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \quad x\in R^1.$$

Тогда $\varphi(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\cos(tx)\exp\left(-\frac{x^2}{2}\right)dx+\frac{i}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sin(tx)\exp\left(-\frac{x^2}{2}\right)dx=$

$$=\exp\left(-\frac{t^2}{2}\right)+\frac{i}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sin(tx)\exp\left(-\frac{x^2}{2}\right)dx=\exp\left(-\frac{t^2}{2}\right).$$

Пример 11. Пусть Y имеет нормальное распределение $N(a,\sigma^2)$ с плотностью

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad x\in R^1.$$

Тогда , $\varphi_X(t)=\exp(-t^2/2)$, откуда $\varphi_Y(t)=\exp\left(iat-\frac{\sigma^2t^2}{2}\right)$.

Пример 12. Случайная величина X равномерно распределена на интервале $(-a,a)$. Характеристическая функция равна

$$\varphi(t)=\int_{-a}^{a}e^{itx}\frac{dx}{2a}=\frac{\sin(at)}{at}.$$

Через f_X обозначим плотность с.в. X .

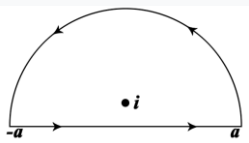
Пример 13. Покажем, что для распределения Коши х.ф. равна $\varphi(t)=\exp(-|t|)$.

Воспользуемся теоремой о вычетах. Пусть функция $f(x)$ аналитична в замкнутой односвязной области G за исключением конечного числа особых точек $a_1,a_2,\ldots,a_n$, ни одна из которых не принадлежит граничному контуру ∂G . Тогда справедлива следующая формула

$\int_{\partial G}f(z)dz=2\pi i\sum_{k=1}^{n}\operatorname*{res}_{z=a_k}f(z)$, где $\operatorname*{res}_{z=a_k}f(z)$ есть вычет функции $f(z)$ в точке $z=a_k$.

Обход контура ∂G производится против часовой стрелки. Для использования теоремы о вычислении вещественных интегралов нужно интегрируемую функцию $f(x)$ аналитически продолжить на комплексную плоскость и найти её вычеты.

Найдем интеграл $\int_{-\infty}^{+\infty}\frac{e^{itx}}{x^2+1}dx$. Функция e^{itx} есть целая функция, а функция $z^2+1=(z+i)(z-i)$, т.е. имеет нули в точках $\pm i$, поэтому для $t>0$ возьмем контур



В пределах этого контура лежит лишь одна точка $z=i$ и

$$\frac{e^{itz}}{z^2+1}=\frac{e^{itx}}{2i}\left(\frac{1}{z-i}+\frac{1}{z+i}\right)=\frac{e^{itx}}{2i(z-i)}+\frac{e^{itx}}{2i(z+i)}.$$

Вычет функции $f(z)$ в точке $z=i$ равен $\operatorname*{res}_{z=ш} f(z)=\frac{e^{-t}}{2i}$, поэтому

$$\oint_{\partial G}\frac{e^{itz}}{z^2+1}dz=\int_{-a}^{a}\frac{e^{itx}}{x^2+1}dx+\int_{C_a}\frac{e^{itz}}{z^2+1}dz,\quad \left|\int_{C_a}\frac{e^{itz}}{z^2+1}dz\right|\le\int_{C_a}\frac{1}{|z^2+1|}dz\le\int_{C_a}\frac{dz}{|z|^2-1}=\int_0^{\pi}\frac{d\varphi}{a^2-1}\xrightarrow[a\to\infty]{}0,$$

поэтому $\int_{-\infty}^{+\infty}\frac{e^{itx}}{x^2+1}dx=\pi\cdot e^{-t}\ \Rightarrow\ \varphi(t)=e^{-t}$. Можно было доказать, что характеристическая функция имеет заданный вид, используя обратное преобразование Фурье, а также используя, что (см. [18], 3.723(2), 3.723(7)))

$$\int_0^{\infty}\frac{\cos(ax)}{\beta^2+x^2}dx=\frac{\pi}{2\beta}e^{-a\beta}\,(a>0,\beta>0),\quad \text{и}\quad \int_0^{\infty}\frac{\cos(a(b-x))}{c^2+x^2}dx=\frac{\pi}{c}e^{-ac}\cos(b)\ (a>0,b\ge 0,c>0).$$

Для $t<0$ надо взять аналогичный контур, охватывающий точку $z=-i$.

Лемма 2. *Пусть с. в.* $Z\in N(0,1)$, *не зависит от* X *и* $\sigma>0$. *Тогда для любой с. в.* X,

$$f_{X+\sigma Z}(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi_X(t)e^{izy}e^{-\frac{\sigma^2z^2}{2}}dz.$$

Доказательство. По формуле свертки плотность распределения случайной величины $X+\sigma Z$ (для простоты обозначения мы ограничимся абсолютно непрерывной случайной величиной) $f_{X+\sigma Z}(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{\boldsymbol{R}}e^{-\frac{(x-z)^2}{2\sigma^2}}f_X(z)\,dz,$ поэтому

$$f_{X+\sigma Z}(0)=\frac{1}{\sqrt{2\pi}\sigma}\int_{\boldsymbol{R}}e^{-\frac{z^2}{2\sigma^2}}f_X(z)\,dz=\frac{1}{\sqrt{2\pi}\sigma}\boldsymbol{E}\left(\varphi_{\frac{Z}{\sigma}}(X)\right)=\frac{1}{\sqrt{2\pi}\sigma}\boldsymbol{E}\left(\varphi_X\left(\frac{Z}{\sigma}\right)\right)=\frac{1}{2\pi}\int_{\boldsymbol{R}}\varphi_X(z)e^{-\frac{\sigma^2z^2}{2}}dz.$$

Равенство выполнено для $y=0$. В общем случае применим его для с.в. $X-y$:

$$f_{X+\sigma Z}(y)=f_{(X-y)+\sigma Z}(0)=\frac{1}{2\pi}\int_{\boldsymbol{R}}\varphi_{X-y}(z)e^{-\frac{\sigma^2z^2}{2}}dz=\frac{1}{2\pi}\int_{\boldsymbol{R}}\varphi_X(z)e^{-izy}e^{-\frac{\sigma^2z^2}{2}}dz.$$

Лемма 3. *Пусть* $\int_{\boldsymbol{R}}|\varphi_X(t)|dt<\infty$. *Тогда с.в.* X *имеет плотность* f_X и

$$f_X(y)=\frac{1}{2\pi}\int_{\boldsymbol{R}}\varphi_X(t)e^{-ity}dt\ .$$

Доказательство. Пусть $Z\in N(0,1)$ не зависит от X. По предыдущей лемме

$f_{X+\sigma Z}(y)=\frac{1}{2\pi}\int_{\boldsymbol{R}}\varphi_X(z)e^{-izy}e^{-\frac{\sigma^2z^2}{2}}dz,$ поэтому если $g\in C(\boldsymbol{R})$, то

$$\boldsymbol{E}(g(X+\sigma Z))=\int_{\boldsymbol{R}}g(y)f_{X+\sigma Z}(y)\,dy=$$

$$=\frac{1}{2\pi}\int_{\boldsymbol{R}}g(y)\int_{\boldsymbol{R}}\varphi_X(t)e^{-ity}e^{-\frac{\sigma^2t^2}{2}}dtdy=\frac{1}{2\pi}\int_{\boldsymbol{R}}g(y)e^{-\frac{\sigma^2t^2}{2}}\int_{\boldsymbol{R}}\varphi_X(t)e^{-ity}dtdy.$$

Заметим также, что при $\sigma\to 0,\ X+\sigma Z\xrightarrow{d}X$. Отсюда

$$\boldsymbol{E}(g(X))=\lim_{\sigma\to 0}\boldsymbol{E}(g(X+\sigma Z))=\lim_{\sigma\to 0}\frac{1}{2\pi}\int_{\boldsymbol{R}}g(y)e^{-\frac{\sigma^2t^2}{2}}\int_{\boldsymbol{R}}\varphi_X(t)e^{-ity}dtdy=\frac{1}{2\pi}\int_{\boldsymbol{R}}g(y)\int_{\boldsymbol{R}}\varphi_X(t)e^{-ity}dydt,$$

что влечет равенство $f_X(y)=\frac{1}{2\pi}\int_{\boldsymbol{R}}\varphi_X(t)e^{-ity}dt$.

8.2 Характеристические функции и моменты

Рассмотрим связь между моментами и производной в нуле характеристических функций.

ТЕОРЕМА 6. *Если* $\mathbf{E}(|X|^n) < \infty$, *т.е. существует момент порядка* n, *то характеристическая функция* $\varphi(t) = \varphi_X(t)$ *непрерывно дифференцируема до порядка* n *включительно, и можно вычислять производные характеристических функций дифференцированием под знаком интеграла:*

$$\varphi^{(k)}(t) = \int_{-\infty}^{\infty} (ix)^k e^{itx} f(x)\,dx, \quad k = 1, 2, \dots, n.$$

Доказательство следует из возможности дифференцирования под знаком интеграла. Рассмотрим случай $k = 1$. Тогда

$$\lim_{h\to 0} \frac{\varphi(h) - \varphi(0)}{h} = \lim_{h\to 0} \int_{-\infty}^{\infty} \frac{e^{ixh} - 1}{h} dF(x) = \int_{-\infty}^{\infty} \lim_{h\to 0} \frac{e^{ith} - 1}{h} dF(x) = \int_{-\infty}^{\infty} ix\,dF(x),$$

так как

$$\left| \frac{e^{ixh} - 1}{h} \right| \le |x| \quad \text{и} \quad \int_{-\infty}^{\infty} |x|\,dF(x) = \boldsymbol{E}(|X|) < \infty.$$

Задание. Написать аналогичную формулу для дискретной случайной величины.

Из теоремы 1 следует, что для $k = 1, 2, \dots, n$ имеем

$$\begin{cases} |\varphi^{(k)}(t)| \le \mathbf{E}(|X|^k), \quad \forall t \in R^1, \\ \varphi^{(k)}(0) = i^k \mathbf{E}(X^k). \end{cases}$$

Иными словами, зная характеристическую функцию $\phi(t)$ можно найти начальный момент порядка k: $\mu_k = \mathbf{E}(X^k)$ по формуле $\boxed{\mu_k = (-i)^k \varphi^{(k)}(0)}$

Замечание. Теорема, обратная этой теореме 8, неверна. Однако можно показать, что если характеристическая функция $\varphi(t)$ величины X имеет производную порядка $2n$ в начале координат, тогда моменты с.в. X существуют и конечны до порядка $2n$ включительно. Отсюда, в частности, следует, что если характеристическая функция $\phi(t)$ бесконечное число раз дифференцируема в начале координат, то с.в. X имеет моменты всех порядков.

Следующий пример показывает, что этот результат улучшить нельзя.

Пример 14. Пусть $\mathbf{P}(X = j) = \mathbf{P}(X = -j) = \dfrac{1}{2C\, j^2 \ln j}$, $j = 2, 3, \dots$, $C = \sum\limits_{j=2}^{\infty} \dfrac{1}{j^2 \ln j}$.

Характеристическая функция равна $\varphi(t) = C^{-1} \sum\limits_{j=2}^{\infty} \dfrac{\cos(jt)}{j^2 \ln j}$,

а математическое ожидание $\mathbf{E}(X) = C^{-1} \sum\limits_{j=2}^{\infty} \dfrac{1}{j \ln j}$.

Покажем, что ряд $\sum\limits_{j=2}^{\infty} \dfrac{1}{j \ln j}$ расходится. Сравним его с рядом $\sum\limits_{j=2}^{\infty} (\ln\ln(j+1) - \ln\ln j)$, который расходится. Имеем

$$\ln\ln(j+1) - \ln\ln j = \frac{1}{(j+\theta_j)\ln(j+\theta_j)} \le \frac{1}{j \ln j}, \quad 0 < \theta_j < 1,$$

откуда следует требуемое. Но $\varphi'(t) = C^{-1} \sum\limits_{j=2}^{\infty} \dfrac{\sin(jt)}{j \ln j}$.

Функция $\varphi'(t)$ – непрерывная функция t, поскольку $\frac{1}{\ln j} \underset{j\to\infty}{\to} 0$. В частности, $\varphi'(0)=0$, следовательно, $\varphi(t)=1+o(|t|)$ при $t\to 0$.

ТЕОРЕМА 7. *Если* $\mathbf{E}(|X|^n)<\infty$, *т.е. существует момент порядка* $n\ge 1$, *то характеристическую функцию* $\varphi(t)$ *можно представить в следующем виде*

$$\varphi(t)=\sum_{k=0}^{n}\frac{(it)^k}{k!}\mathbf{E}(X^k)+o\,(|t|^n) \quad \forall t\in R^1.$$

В окрестности $t=0$ можно определить функцию

$$\psi_X(t)=\ln \varphi_X(t)$$

(выберем значение, которое зануляется в начале координат).

ТЕОРЕМА 8. *Если* $\mathbf{E}(|X|^n)<\infty$, *т.е. существует момент порядка* $n\ge 1$, *то* $\psi_X(t)$ *можно представить в следующем виде*

$$\psi_X(t)=\sum_{j=0}^{n}\frac{(it)^j}{j!}\kappa_j+o\,(|t|^n) \quad \forall t\in R^1.$$

Коэффициенты κ_j называют кумулянтами (полуинвариантами, семиинвариантами) порядка j (для $j\ge 2$ они являются инвариантами относительно сдвига системы координат, для моментов это свойство не выполняется). Эта терминология оправдана следующим свойством: кумулянт порядка j суммы независимых случайных величин равен сумме соответствующих кумулянтов порядка j.

Свойства:

1. $\psi_{aX+b}(t)=itb+\psi_X(at)$, т.е. при линейном преобразовании изменяется только коэффициент при t.
2. Кумулянты κ_j связаны с начальными моментами $\mu_j=\mathbf{E}(X^j)$ по следующим формулам:

$$\kappa_1=\mu_1=\mathbf{E}\,(X), \qquad \mu_1=\kappa_1,$$
$$\kappa_2=\mu_2-\mu_1^2=\mathbf{D}\,(X), \qquad \mu_2=\kappa_2+\kappa_1^2,$$
$$\kappa_3=\mu_3-3\mu_1\mu_2+2\mu_1^3, \qquad \mu_3=\kappa_3+3\kappa_1\kappa_2+\kappa_1^3,$$
$$\kappa_4=\mu_4-3\mu_2^2-4\mu_1\mu_3+12\mu_1^2\mu_2-6\mu_1^4,$$
$$\mu_4=\kappa_4+3\kappa_2^2+4\kappa_1\kappa_3+6\kappa_1^2\kappa_2+\kappa_1^4$$

Если рассматривать центральные моменты $\alpha_k=\mathbf{E}((X-\mu_1)^k)$, то эта связь выглядит следующим образом: $\kappa_2=\alpha_2$, $\kappa_3=\alpha_3$, $\kappa_4=\alpha_4-3\alpha_2^2$.

Следствие. Аддитивность дисперсий.

Пусть $X_1, X_2, \dots, X_k$ система k независимых случайных величин. Из аддитивности κ_2 выводим:

$$\sigma^2_{X_1+X_2+\ldots+X_k}=\sum_{j=1}^{k}\sigma^2_{X_j}.$$

Для нормальной случайной величины , $\kappa_1=a\kappa_1=a$, $\kappa_j=0$ для $j\ge 3$.

Пример 15. Пусть независимые случайные величины X_1 и X_2 распределены нормально $X_1\in N(a_1,\sigma_1^2)$, $X_2\in N(a_2,\sigma_2^2)$. Тогда случайная величина $X=X_1+X_2$ также распределена нормально с параметрами a_1+a_2, $\sigma_1^2+\sigma_2^2$.

Пример 16. Пусть независимые случайные величины X_1 и X_2 распределены по закону Пуассона, причем $\mathbf{P}(X_1=k)=\frac{\lambda_1^k}{k!}e^{-\lambda_1}$, $\mathbf{P}(X_2=k)=\frac{\lambda_2^k}{k!}e^{-\lambda_2}$.

Тогда случайная величина $X = X_1 + X_2$ распределена по закону Пуассона с параметром $\lambda = \lambda_1 + \lambda_2$, т.е. $\mathbf{P}(X_1 = k) = \dfrac{(\lambda_1 + \lambda_2)^k}{k!} e^{-(\lambda_1+\lambda_2)}$.

Предположим, что у случайной величины X существует начальный момент μ_k любого порядка k. Можно ли восстановить распределение, зная все моменты? В общем случае это не так. Мы приведем достаточные условия, когда это возможно сделать. Именно, любое из следующих условий является достаточным для того, чтобы последовательность абсолютных моментов μ_k однозначно определяла функцию распределения:

а) область значений с.в. X конечна;

б) $\sum_{j=1}^{\infty} \left(\mu_{2j} \right)^{-\frac{1}{2j}} = \infty$ (если область значений с.в. $(-\infty, \infty)$);

в) $\sum_{j=1}^{\infty} \left(\mu_{j} \right)^{-\frac{1}{2j}} = \infty$ (если область значений с.в. $(0, \infty)$);

г) **Критерий Карлемана**: $\overline{\lim_{n\to\infty}} \dfrac{\mu_n^{1/n}}{n} < \infty$ (равносильное условие $\overline{\lim_{n\to\infty}} \left(\dfrac{\mu_n}{n!} \right)^{1/n} < \infty$).

Рассмотрим критерий Карлемана. Так как $\mu_n = \mathbf{E}(|X|^n)$, то пусть это условие выполнено, т.е. $\overline{\lim_{n\to\infty}} \left(\dfrac{\mu_n}{n!} \right)^{1/n} \le \dfrac{1}{t_o} < \infty \Leftrightarrow \left(\dfrac{\mathbf{E}(|X|^n t_o^n)}{n!} \right)^{1/n} < 1$. Следовательно, по признаку Коши ряд $\sum_{n=1}^{\infty} \dfrac{(it)^n \mu_n}{n!}$ сходится для $t \le t_o$.

В общем случае может быть, что моменты существуют, но соответствующий радиус сходимости равен нулю, и , значит ряд $\sum_{n=1}^{\infty} \dfrac{(it)^n \mu_n}{n!}$ неоднозначно определяет функцию.

Пример распределения, для моментов которого *<u>не выполнено условие Карлемана</u>*.

Пусть плотность распределения равна (логнормальное распределение)

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, x > 0; \mu \in \boldsymbol{R}, \sigma > 0, \Rightarrow \mu_n = \boldsymbol{E}(X^n) = e^{n\mu + n^2\sigma^2/2}, \text{ и } \overline{\lim_{n\to\infty}} \frac{\mu_n^{1/n}}{n} = \infty .$$

Пусть $f_a(x) = f(x)(1 + a\sin(2\pi \ln x)), -1 \le a \le 1, x > 0$. Так как $f_a(x) \ge 0$ и $\int_0^{\infty} x^k f(x) \sin(2\pi \ln x)\, dx = 0, k = 0,1,...,$ (подстановка $\ln x = t = y + k$ приводит его к $(2\pi)^{-1/2} \exp(k^2/2) \int_{-\infty}^{\infty} \exp(-y^2/2) \sin(2\pi y)\, dy = 0 -$ интеграл от нечетной функции), то моменты этих двух распределений равны, но это разные распределения.

ТЕОРЕМА 9. *Существование $k-$й производной в нуле влечет существование момента $k-$го порядка при k четном. Это также верно и при k нечетном, если случайная величина X ограничена сверху или снизу.*

Доказательство. Докажем сначала, что если у характеристической функции существует вторая производная в нуле, то существует и момент порядка 2. Итак, пусть существует $\varphi''(0)$. Обозначим действительную часть функции $\varphi(t)$ через $u(t)$. Рассмотрим $|x| \le T$ и воспользуемся тем, что $1-\cos\dfrac{x}{T}=2\sin^2\dfrac{x}{2T}$. Если $x=T$, то $2\sin^2\dfrac{1}{2}\approx 0.459>\dfrac{1}{3}$, откуда $2\sin^2\dfrac{x}{2T}\ge\dfrac{x^2}{6T^2}$. Следовательно, $1-u\left(\dfrac{1}{T}\right)\ge\int\limits_{-T}^{T}\left(1-\cos\dfrac{x}{T}\right)dF(x)\ge\dfrac{1}{6T^2}\int\limits_{-T}^{T}x^2\,dF(x)$.

Отсюда, полагая $\eta=\dfrac{1}{T}$, получим $\dfrac{1}{6T^2}\int\limits_{-T}^{T}x^2\,dF(x)\le\dfrac{1-u(\eta)}{\eta^2}=-\dfrac{u'(\theta\eta)}{\eta}\le-\dfrac{u'(\theta\eta)}{\theta\eta}\underset{\eta\to 0}{\to}-\varphi''(0)$,

где $0\le\theta\le 1$. Это означает, что существует момент μ_2.

По индукции выводится, что при четном k утверждение имеет место и в общем случае. Действительно, утверждение верно для $k=2$ и пусть оно верно для $2k$, т.е.

$\dfrac{\varphi^{(2k)}(0)}{(i)^{2k}}=\mu_{2k}=\int\limits_{-\infty}^{\infty}x^{2k}\,dF(x)$. Здесь $\dfrac{\varphi^{(2k)}(t)}{(it)^{2k}}=\int\limits_{-\infty}^{\infty}e^{itx}x^{2k}\,dF(x)$.

Рассмотрим распределение $C\,x^{2k}\,dF(x)$. Характеристическая функция этого распределения равна $\lambda(t)=\dfrac{C\varphi^{(2k)}(t)}{(it)^{2k}}=C\int\limits_{-\infty}^{\infty}e^{itx}x^{2k}\,dF(x)$. Из предположения теоремы следует, что существует вторая производная $\lambda''(t)$, откуда получаем, что существует и $\mu_{2k+2}=\int\limits_{-\infty}^{\infty}x^{2k+2}\,dF(x)$.

Задание. Используя характеристическую функцию покажите, что если с.в. X имеет биномиальное распределение $B(N,p)$, то **третий** центральный момент $\alpha_3=\mathbf{E}(X-np)^3=npq(q-p)$,

а **четвертый** центральный момент $\alpha_4=\mathbf{E}(X-np)^4=3n^2p^2q^2+npq\,(1-6\,pq)$.

8.3 Предельные теоремы для характеристических функций. Теорема Муавра-Лапласа

Одно из важных свойств характеристических функций связано со *сходимостью по распределению случайных величин*.

Определение. Говорят, что последовательность случайных величин X_n с функциями распределения $F_n(x)$ сходится *по распределению*, когда $n\to\infty$, к случайной величине с функцией распределения $F(x)$, если для всякой точки непрерывности x функции $F(x)$ имеем:

$$\lim_{n\to\infty}F_n(x)=F(x)\qquad\text{(обозначение: } X_n\underset{n\to\infty}{\overset{d}{\to}}X\text{)}$$

Две следующие теоремы уточняют соотношения между сходимостью по распределению последовательности $\{X_n\}$ и сходимостью последовательности $\{\varphi_n\}$ соответствующих характеристических функций.

ТЕОРЕМА 10. Если $X_n\underset{n\to\infty}{\overset{d}{\to}}X$, то последовательность $\{\varphi_n\}$ характеристических функций величин X_n сходится к характеристической функции φ величины X, причем эта сходимость равномерна во всяком конечном интервале.

ТЕОРЕМА 11. (P.Levy) Если последовательность характеристических функций φ_n сходится поточечно к функции φ, действительная часть которой $\mathcal{Re}\,\varphi$ непрерывна в начале координат, тогда: 1) φ – характеристическая функция;

2) последовательность функций распределения F_n величин X_n сходится к функции распределения F, характеристическая функция которой – φ (в смысле поточечной сходимости) во всякой точке непрерывности функции F.

Пусть $X_1, X_2, \dots, X_n$ – последовательность независимых случайных величин, причем

$$X_i = \begin{cases} 1, \text{ с вероятноссгью } p, \\ 0, \text{ с вероятноссгью } q. \end{cases}$$

Положим $\quad S_n = \sum_{i=1}^{n} X_i$.

ТЕОРЕМА 12. (Муавр и Лаплас). Если $0 < p < 1$, тогда при

1) $\mathbf{P}\left(a \le \dfrac{S_n - np}{\sqrt{npq}} \le b \right) \underset{n\to\infty}{\to} \dfrac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\dfrac{x^2}{2} \right) dx = \Phi(b) - \Phi(a$;

2) $\mathbf{P}(S_n = k) \approx \dfrac{1}{\sqrt{2\pi\, npq}} \exp\left(-\dfrac{(k-np)^2}{2npq} \right) \mathbf{P}(S_n = k) \approx \dfrac{1}{\sqrt{2\pi\, npq}} \exp\left(-\dfrac{(k-np)^2}{2npq} \right)$.

Доказательство. 1) Имеем: $\quad \varphi_{S_n}(t) = (p e^{it} + q)^n$.

Тогда $\quad \varphi_{\frac{S_n - np}{\sqrt{npq}}}(t) = \left(p \exp\left(it\sqrt{\dfrac{q}{np}} \right) + q \exp\left(-it\sqrt{\dfrac{p}{nq}} \right) \right)^n = \left(1 - \dfrac{t^2}{2n} + R_n \right)^n \underset{n\to\infty}{\to} \exp\left(-\dfrac{t^2}{2} \right)$

– характеристическая функция стандартного нормального закона. Отсюда и из теоремы 12 следует результат части 1. Действительно,

$$p\,\exp\left(it\sqrt{\frac{q}{np}} \right) = p\left(1 + it\sqrt{\frac{q}{np}} - \frac{t^2}{2}\frac{q}{np} + o\left(\frac{1}{n} \right) \right) = p + it\sqrt{\frac{pq}{n}} - \frac{t^2}{2}\frac{q}{n} + o\left(\frac{1}{n} \right),$$

$$q\,\exp\left(-it\sqrt{\frac{p}{nq}} \right) = q\left(1 - it\sqrt{\frac{p}{nq}} - \frac{t^2}{2}\frac{p}{nq} + o\left(\frac{1}{n} \right) \right) = q - it\sqrt{\frac{pq}{n}} - \frac{t^2}{2}\frac{p}{n} + o\left(\frac{1}{n} \right).$$

Складывая два этих равенства, получаем $\; p\exp\left(it\sqrt{\dfrac{q}{np}} \right) + q\exp\left(-it\sqrt{\dfrac{p}{nq}} \right) = 1 - \dfrac{t^2}{2n} + o\left(\dfrac{1}{n} \right)$.

Отсюда следует результат первой части.

2) Так как S_n принимает целые значения, то

$$\mathbf{P}(S_n = k) = \mathbf{P}(k \le S_n < k+1) = \mathbf{P}\left(\frac{k-np}{\sqrt{npq}} \le \frac{S_n - np}{\sqrt{npq}} < \frac{k+1-np}{\sqrt{npq}}\right) \approx$$

$$\approx \frac{1}{\sqrt{2\pi}} \int_{\frac{k-np}{\sqrt{npq}}}^{\frac{k+1-np}{\sqrt{npq}}} \exp\left(-\frac{x^2}{2} \right) dx \approx \frac{1}{\sqrt{2\pi\, npq}} \exp\left(-\frac{(k-np)^2}{2npq} \right).$$

Замечание. Отыскание класса предельных распределений является, в некоторых условиях, задачей, называемой «центральной предельной», на которой мы подробнее остановимся в лекции 9. Отметим только, что пуассоновское распределение можно рассматривать как предельное биномиального распределения, когда вероятность p меняется вместе с n (схема серий).

Задача 3. Пусть $\varphi(t)$ – характеристическая функция с.в. X. Покажите, что $\exp(\varphi(t)-1)$ тоже характеристическая функция.

Решение. $e^{\varphi(t)-1} = \lim\limits_{n\to\infty}\left(1 + \dfrac{\varphi(t)-1}{n} \right)^n = \lim\limits_{n\to\infty}\left(\left(1 - \dfrac{1}{n} \right) + \dfrac{1}{n}\varphi(t) \right)^n$.

Но $1-$ х.ф. вырожденного распределения $\boldsymbol{P}(X=0)=1$, смесь двух х.ф. – тоже х.ф., а затем переходя к пределу – получаем функцию, непрерывную в нуле, т.е. х.ф.

ТЕОРЕМА 13. (Пуассон) *Пусть при каждом* $n \ge 1$ *независимые одинаково распределенные случайные величины* $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$ *таковы, что*

$$\mathbf{P}(\xi_{nk}=1)=p_{nk}, \quad \mathbf{P}(\xi_{nk}=0)=q_{nk}=1-p_{nk}, \quad 1\le k\le n,$$

$$\max_{1\le k\le n} p_{nk} \underset{n\to\infty}{\to} 0, \quad p_{n1}+\ldots+p_{nn} \underset{n\to\infty}{\to} \lambda>0, \quad S_n=\sum_{k=1}^{n}\xi_{nk}.$$

Тогда $$\mathbf{P}(S_n=m) \underset{n\to\infty}{\to} \frac{e^{-\lambda}\lambda^m}{m!}, \quad m=0,1,\ldots .$$

Доказательство. Имеем: $\mathbf{E}(e^{it\xi_{nk}})=p_{nk}e^{it}+q_{nk}=1+p_{nk}(e^{it}-1)$.

Поэтому $\left(\max\dfrac{1}{1-p_{nk}}=\dfrac{1}{1-\max p_{nk}}\le\dfrac{1}{1-C_1}\right)$

$$\varphi_{S_n}(t)=\prod_{k=1}^{n}(1+p_{nk}(e^{it}-1)), \quad \text{и} \qquad \ln\varphi_{S_n}(t)=\sum_{k=1}^{n}\ln(1+p_{nk}(e^{it}-1))=\sum_{k=1}^{n}p_{nk}(e^{it}-1)+B_n,$$

где $B_n=\sum_{k=1}^{n}\sum_{s=2}^{\infty}\dfrac{(-1)^{s-1}}{s}p_{nk}^s(e^{it}-1)^s$. Тогда

$$|B_n|\le\sum_{k=1}^{n}\sum_{s=2}^{\infty}\frac{1}{2}p_{nk}^s|e^{it}-1|^s\le\sum_{k=1}^{n}\sum_{s=2}^{\infty}\frac{1}{2}(2p_{nk})^s=\sum_{k=1}^{n}\frac{(2p_{nk})^2}{2(1-p_{nk})}\le2\sum_{k=1}^{n}p_{nk}^2\le\max_{1\le k\le n}p_{nk}\sum_{k=1}^{n}p_{nk}\underset{n\to\infty}{\to}0.$$

Определение. Функцию распределения $F(x)$ и соответствующую ей характеристическую функцию будем называть *безгранично делимой*, если $\varphi(t)=\left(\varphi_n(t)\right)^n$ для любого натурального n.

Задание. Покажите, что нормальное распределение является безгранично делимым распределением.

8.4 Геометрические случайные суммы

Если в случайной сумме $S_N=\sum_{i=1}^{N}X_i$ число слагаемых N имеет <u>геометрическое распределение</u> $\boldsymbol{P}(N=n)=p(1-p)^n, n=0,1,2,\ldots,$ то случайная сумма S_N называется *геометрической случайной суммой.*

Если $N, X_1, X_2, \ldots$ независимы в совокупности, то характеристическая функция геометрической случайной суммы равна

$$\varphi_{S_N}(t)=\frac{p}{1-q\varphi(t)}, q=1-p. \qquad (*)$$

где $\varphi(t)$ – характеристическая функция слагаемых. Действительно,

$$\varphi_{S_N}(t)=\boldsymbol{E}(e^{itS_N})=\sum_{n=1}^{\infty}\boldsymbol{E}(e^{itS_N}\mid N=n)pq^n=\sum_{n=0}^{\infty}\boldsymbol{E}(e^{itS_n}\mid N=n)pq^n=\sum_{n=0}^{\infty}\boldsymbol{E}(e^{itS_n})pq^n=p\sum_{n=0}^{\infty}(\varphi(t)q)^n=$$

$$=\frac{p}{1-q\varphi(t)}, \quad q=1-p, 0<q<1, |\varphi(t)|\le1.$$

Аналогично показывается, что если $\boldsymbol{P}(N=k)=C_{k+m-1}^{m-1}p^m q^k, k=0,1,2,\ldots; m\ge1-$ фиксировано, то характеристическая функция отрицательной биномиальной случайной суммы равна

$$\varphi_{S_N}(t) = \left(\frac{p}{1-q\varphi(t)}\right)^m, \ q = 1-p. \qquad (**)$$

где $\varphi(t)$ – характеристическая функция слагаемых.

ТЕОРЕМА 14 (Реньи). Пусть случайная величина N имеет геометрическое распределение с параметром p, $X_1, X_2, \ldots$ — независимые одинаково распределённые случайные величины, $X_i \geq 0$ и $0 < a = \boldsymbol{E}(X_i) < \infty$. Тогда

$$\sup_x \left| \boldsymbol{P}\left(\frac{p}{a} S_N < x\right) - G(x) \right| \underset{p \to 0}{\to} 0,$$

где $G(x)$ — функция распределения стандартного показательного распределения: $G(x) = 1 - e^{-x}, \ x \geq 0.$

Поясним, как получается распределение $G(x)$: так как величины X_i имеют математическое ожидание, то $f(t) = 1 + iat + o(|t|), |t| \to 0,$ поэтому подставляя в формулу (*) это представление, получим характеристическую функцию $f_{S_N}(t) = \dfrac{1}{1-it}$, т.е. х.ф. показательного закона. Если $b^2 = \boldsymbol{E}(X^2)$, тогда $\sup\limits_x \left| \boldsymbol{P}\left(\frac{p}{a} S_N < x\right) - G(x) \right| \leq \dfrac{pb^2}{(1-p)a^2}$.

Аналогичную теорему можно сформулировать и для отрицательно биномиальных случайных сумм, только теперь в качестве предельного распределения будет выступать гамма-распределение.

Связь между геометрическими и пуассоновскими случайными суммами.

Любая геометрическая случайная сумма является пуассоновской случайной суммой. Более точно, если $M, X_1, X_2, \ldots$ — независимые случайные величины такие, что M имеет геометрическое распределение и $X_1, X_2, \ldots$ – одинаково распределены с общей характеристической функцией $f(t)$, то

$$X_1 + X_2 + \ldots + X_M = Y_1 + Y_2 + \ldots + Y_N \quad \text{(имеют одинаковое распределение)},$$

где N — пуассоновская случайная величина с параметром $\lambda = \ln\dfrac{1}{p}$, случайные величины $N, Y_1, Y_2, \ldots$ независимы, причём $Y_1, Y_2, \ldots$ имеют одно и то же распределение с характеристической функцией $f_{Y_1}(t) = \dfrac{\ln(1-(1-p)f(t))}{\ln p}$.

Следствие (обратная связь).

Пусть S_N — пуассоновская случайная сумма с характеристической функцией $f(t) = e^{\lambda(f(t)-1)}$. Если функция $g(t) = \dfrac{1-e^{\lambda f(t)}}{1-e^{-\lambda}}$ (A) является характеристической, то S_N — геометрическая случайная сумма, $S_N = Y_1 + Y_2 + \ldots Y_M$, где случайные величины $M, Y_1, Y_2, \ldots$ независимы, причём M имеет геометрическое распределение с $p = e^{-\lambda}$, а $Y_1, Y_2, \ldots$ одинаково распределены с общей характеристической функцией $g(t)$.

Пример. Пусть с.в. $X_i, i = 1, 2, \ldots,$ независимы и каждая имеет логарифмическое распределение $\mathbf{P}(X_i = k) = \dfrac{p^k}{k \ln(1-p)}, k = 1, 2, \ldots,$ его х.ф. равна $\varphi_L(t) = \dfrac{\ln(1-pe^{it})}{\ln(1-p)}$ и $S_N = \sum_{i=1}^N X_i$, $X_j, j = 1, 2, ..$ и $N-$ независимы, с.в. N имеет распределение Пуассона с параметром λ .Тогда

$$\varphi(t)=\varphi_{S_N}(t)=\mathbf{E}(e^{itS_N})=\sum_{n=1}^{\infty}\mathbf{E}(e^{itS_N}\mid N=n)\mathbf{P}(N=n)=\sum_{n=1}^{\infty}\mathbf{E}(e^{itS_n}\mid N=n)\mathbf{P}(N=n)=$$

$$=\sum_{n=1}^{\infty}\mathbf{E}(e^{itS_n})\mathbf{P}(N=n)=\sum_{n=1}^{\infty}\left(\frac{\ln(1-pe^{it})}{\ln(1-p)}\right)^n\frac{\lambda^n}{n!}e^{-\lambda}=\exp\left(\lambda\frac{\ln(1-pe^{it})}{\ln(1-p)}-\lambda\right)=$$

$$=\exp\left(\frac{\lambda}{\ln(1-p)}\ln\left(\frac{1-pe^{it}}{1-p}\right)\right)=\left(\frac{1-p}{1-pe^{it}}\right)^{\nu},\quad \nu=-\frac{\lambda}{\ln(1-p)}.$$

8.5 Характеристическая функция случайного вектора

Характеристической функцией $n-$ мерной случайной величины $X_1, X_2, \dots, X_n$ называется математическое ожидание случайной величины $\exp\left(i\sum_{j=1}^{n}t_jX_j\right)$, где $t_1, t_2, \dots, t_n$ – вещественные, т.е. $\varphi(t_1, t_2, \dots, t_n)=\mathbf{E}\left(\exp\left(i\sum_{j=1}^{n}t_jX_j\right)\right)$.

Функция распределения однозначно определяется своей характеристической функцией.

Как и в одномерном случае х.ф. $n-$мерной случайной величины равномерно непрерывна во всем пространстве и удовлетворяет следующим соотношениям:

$$\varphi(0,\dots,0)=1,\qquad |\varphi(t_1,\dots,t_n)|\le 1,\ \varphi(-t_1,\dots,-t_n)=\overline{\varphi(t_1,\dots,t_n)}\ .$$

По характеристической функции величины ($X_1, X_2, \dots, X_n$) легко найти характеристическую функцию $k-$мерной величины $(X_{i_1},\dots,X_{i_k})$ $(k<n)$. Для этого нужно положить нулю $t_s,\ s\ne i_j$. Например, характеристическая функция X_1 есть $\varphi_{X_1}(t_1)=\varphi(t_1, 0, \dots, 0)$.

Из определения вытекает, что если ($X_1, X_2, \dots, X_n$) независимы, то

$$\varphi(t_1, t_2, \dots, t_n)=\varphi_{X_1}(t_1)\cdot\varphi_{X_2}(t_2)\cdot\ldots\cdot\varphi_{X_n}(t_n)\ .$$

Справедливо и обратное соотношение.

Если $\varphi(t_1, t_2, \dots, t_n)$ – характеристическая функция $(X_1, X_2, \dots, X_n)$, то характеристическая функция суммы равна $\varphi(t, t, \dots, t)$, т.е. когда все значения аргументов одинаковы.

Если вектор $(X_1, X_2, \dots, X_n)$ имеет $n-$мерное нормальное распределение с плотностью

$$f(x_1, x_2, \dots, x_n)=(2\pi)^{-\frac{n}{2}}|\Delta|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\sum_{j,k}\frac{(x_j-a_j)(x_k-a_k)}{\sigma_j\sigma_k\rho_{jk}}\right),$$

то его характеристическая функция равна

$$\varphi(t_1, t_2, \dots, t_n)=\exp\left(i\sum_{j=1}^{n}a_jt_j-\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n}\sigma_j\sigma_k\rho_{jk}\,t_jt_k\right).$$

Для двумерного случайного вектора (X_1, X_2) с плотностью

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-a_1)^2}{\sigma_1^2}-2\rho\cdot\frac{(x_1-a_1)}{\sigma_1}\cdot\frac{(x_2-a_2)}{\sigma_2}+\frac{(x_2-a_2)^2}{\sigma_2^2}\right)\right)$$

характеристическая функция равна

$$\varphi(t_1, t_2)=\exp\left(ia_1t_1+ia_2t_2-\frac{1}{2}(\sigma_1^2t_1^2+2\sigma_1\sigma_2\rho t_1t_2+\sigma_2^2t_2^2)\right).$$

С помощью производящих функций решим следующую задачу.

Задача 4. В трамвайном билете 6 цифр. Какова вероятность получить счастливый билет, т.е. чтобы сумма первых трех цифр равнялась сумме трех последних.

Решение. Решим общую задачу, т.е. пусть имеется 2ν цифр. Какова вероятность того, что сумма первых ν цифр равна сумме ν последних.

Рассмотрим одну цифру. Тогда $\mathbf{P}(X_1 = k) = p_k = \dfrac{1}{10}, \quad k = 0, 1, \dots, 9$.

Производящая функция вероятностей этой случайной величины равна

$$\varphi_{X_1}(u) = \sum_{k=0}^{9} p_k u^k = \frac{1}{10}\,(1 + u + \dots + u^9) = \frac{1-u^{10}}{1-u}\cdot\frac{1}{10}.$$

Далее, если случайные величины X_1 и X_2 независимы и имеют распределение вероятностей $\{p_k\}$ и $\{q_l\}$, то производящая функция суммы $X_1 + X_2$ равна

$$\varphi_{X_1+X_2}(u) = \sum_{l=0}^{\infty}\left(\sum_{k=0}^{l} p_k q_{l-k}\right)u^l = \sum_{k=0}^{\infty}\sum_{l=k}^{\infty} p_k q_{l-k} u^l = \sum_k p_k u^k \sum_{l=k}^{\infty} q_{l-k} u^{l-k} = \psi_{X_1}(u)\cdot\psi_{X_2}(u)$$

Пусть $X_1, X_2, \dots, X_\nu$ – независимы. Тогда $\varphi_{X_1+X_2+\dots+X_\nu}(u) = \dfrac{1}{10^\nu}\left(\dfrac{1-u^{10}}{1-u}\right)^\nu = \varphi_1(u)$.

Здесь коэффициент при u^l есть вероятность того, что сумма первых ν цифр равна l.

Точно так же

$$\varphi_2(u) = \frac{1}{10^\nu}\left(\frac{1-u^{-10}}{1-u^{-1}}\right)^\nu$$

есть производящая функция; коэффициент при u^{-l} есть вероятность того, что сумма последних ν цифр равна l. В таком случае, коэффициент при u^0 у функции

$$\varphi(u) = \varphi_1(u)\cdot\varphi_2(u)$$

есть вероятность того, что сумма первых ν цифр равна сумме ν последних цифр. Найдем этот коэффициент. Имеем

$$\varphi(u) = \frac{1}{10^{2\nu}}\left(\frac{1-u^{10}}{1-u}\right)^\nu\left(\frac{1-u^{-10}}{1-u^{-1}}\right)^\nu = \frac{1}{10^{2\nu}u^{9\nu}}\cdot\frac{(1-u^{10})^{2\nu}}{(1-u)^{2\nu}}.$$

Кроме того, $(1-u^{10})^{2\nu} = \sum_{j=0}^{2\nu} C_{2\nu}^j(-1)^j u^{10j}$, $(1-u)^{-2\nu} = \sum_{k=0}^{\infty}\dfrac{\Gamma(k+2\nu)}{k!\,\Gamma(2\nu)}u^k = \sum_{k=0}^{\infty} C_{2\nu+k-1}^{2\nu-1}u^k = \sum_{k=0}^{\infty} C_{2\nu+k-1}^{k}u^k$,

поэтому коэффициент при u^0 равен

$$p = 10^{-2\nu}\left(C_{11\nu-1}^{2\nu-1} - C_{2\nu}^1 C_{11\nu-11}^{2\nu-1} + C_{2\nu}^2 C_{11\nu-21}^{2\nu-1} - \dots\right).$$

Суммирование производится до тех пор, пока не станет $2\nu - 1 > 11\nu - 10k - 1 \ge 0$, т.е.

$$k \le \left[\frac{9\nu}{10}\right] = \nu - \left[\frac{\nu}{10}\right].$$

Например, при $\quad \nu = 1$: $C_{10}^1 = 10$, $p = \dfrac{10}{10^2} = 0.1$;

$\nu = 2$: $C_{21}^3 - C_4^1 C_{11}^3 = 670$, т.е. вероятность равна $p = \dfrac{670}{10^4} = 0.067$;

$\nu = 3$: $C_{32}^5 - C_6^1 C_{22}^5 + C_6^2 C_{12}^5 = 55252$, $\quad p = \dfrac{55252}{10^6} = 0.055$;

$\nu = 4$: $C_{43}^7 - C_8^1 C_{33}^7 + C_8^2 C_{23}^7 - C_8^3 C_{13}^7 = 4816030$, $\quad p = 0.048$.

В случае, когда основание счисления равно q, то последняя формула примет вид:

$$p = q^{-2\nu}(C_{(q+1)\nu-1}^{2\nu-1} - C_{2\nu}^1\cdot C_{(q+1)\nu-(q+1)}^{2\nu-1} + C_{2\nu}^2\cdot C_{(q+1)\nu-2q-1)}^{2\nu-1} - \dots).$$

Пример 16. Пусть

$$X_u = \sum_{j=-\infty}^{\infty} c_{t+j}\eta_j,\ u \ge 1, \qquad (5)$$

где $c_k = \begin{cases} \dfrac{1}{s^k}, \text{ если } k = 1,2,..., \\ 0, \text{ если } k = ...-2,-1,0, \quad (s \ge 2), \end{cases}$

$s-$ задано, $\{\eta_k\}-$ независимы и одинаково распределены:

$$\boldsymbol{P}(\eta_k = 0) = \boldsymbol{P}(\eta_k = 1) = ... = \boldsymbol{P}(\eta_k = s-1) = \frac{1}{s},$$

и, следовательно, имеют характеристическую функцию $\varphi_{\eta_k}(t) = \sum_{k=0}^{s-1} \frac{1}{s} e^{itk}$.

В таком случае характеристическая функция величины $s^{-k}\eta_k$ равна $\psi_k(t) = \frac{1}{s}\left(\sum_{j=0}^{s-1} e^{itj/s^k} \right)$.

Характеристическая функция равномерного на интервале $(0,a)$ распределения равна $\frac{e^{ita}-1}{ita}$, в частности, равномерного на $(0,1)$ распределения $r(t) = \frac{e^{it}-1}{it}$.

Из тождества $r(t) = \frac{e^{it}-1}{it} = \frac{1+e^{it/s}+e^{2it/s}+...+e^{(s-1)it/s}}{s} \cdot \frac{e^{it/s}-1}{it/s}$

следует, что $r(t) = \prod_{k=1}^{n} \frac{1+e^{its^{-k}}+...+e^{(s-1)its^{-k}}}{s} \cdot \frac{e^{its^{-n}}-1}{its^{-n}}$.

Поскольку $\lim_{n\to\infty} \frac{e^{its^{-n}}-1}{its^{-n}} = 1,$ то $r(t)$ может быть записана как бесконечное произведение

$$r(t) = \frac{e^{it}-1}{it} = \prod_{k=1}^{\infty} \frac{1+e^{its^{-k}}+...+e^{(s-1)its^{-k}}}{s}.$$

Из этого представления следует, что равенство (5) есть s-ичное разложение числа X_u.

Из (6) следует также, что $X_u = \frac{X_{u-1}+\xi_u}{s},\ u = 1,2,...$,

где $s \ge 2-$ целое число, $\{\xi_u\}-$ последовательность независимых и одинаково распределенных на множестве $\{0,1,...,s-1\}$ случайных величин, X_0 равномерно распределена на $(0,1)$. Ясно, что $\{X_u\}, u \ge 1$, есть стационарная в узком смысле последовательность равномерно распределенных на $(0,1)$ случайных величин, и , кроме того, $\{X_u\}, u \ge 1-$авторегрессия первого порядка.

Покажите, что если $G(x,y) = P(X_0 < x, X_1 < y)-$совместная функция распределения, где $0 < x < 1, 0 < y < 1,$ то $G(x,x) = x + \frac{(s-1)(x-1)}{s}$ для $\frac{s-1}{s} < x < 1.$

Решение. Действительно, пусть $\frac{s-1}{s} < x < 1$. Тогда $s-1 < sx < s$ и $s-1-j < sx-j < s-j.$

Имеем: $G(x,x) = \boldsymbol{P}(X_0 < x, X_0 + \xi_1 < x) = \frac{1}{s}\sum_{j=0}^{s-1} \boldsymbol{P}(X_0 < x, X_0 < sx-j)$.

Если $j \le s-2,$ то $1 \le (s-1)-j < sx-j$ и $\boldsymbol{P}(X_0 < sx-j) = 1.$ Поэтому для $j \le s-2$

$$\boldsymbol{P}(X_0 < x, X_0 < sx-j) = \boldsymbol{P}(X_0 < x) = x.$$

Если $j = s-1,$ то $0 < sx-(s-1) < 1$ и $x > sx-(s-1)$. Значит,

$$\boldsymbol{P}(X_0 < x, X_0 < sx-(s-1)) = \boldsymbol{P}(X_1 < sx-(s-1)) = sx-(s-1).$$

Отсюда

$$G(x,x) = \frac{1}{s}\sum_{j=0}^{s-2}\boldsymbol{P}(X_0 < x, X_0 < sx-j) + \frac{1}{s}\boldsymbol{P}(X_0 < x, X_0 < sx-(s-1)) = \frac{s-1}{s}x + \frac{sx-(s-1)}{s} =$$

$$= x + \frac{(s-1)(x-1)}{s}.$$

8.6 Эмпирическая характеристическая функция

Рассмотрим случайную выборку из распределения F: $X_1, X_2, ..., X_n$ независимых и одинаково распределенных случайных величин с общей функцией распределения $F(x)$. Обозначим ее характеристическую функцию через $\varphi(t)$. Эмпирическая функция распределения, ассоциированная с выборкой $X_1, X_2, ..., X_n$ определяется как

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i<x\}}, \text{ где } I_{\{X_i<x\}} = \begin{cases} 1, & X_i < x, \\ , & X_i \ge x. \end{cases}$$

Математическое ожидание и дисперсия эмпирической функции распределения равны

$$\boldsymbol{E}(F_n(x)) = F(x), \quad \boldsymbol{D}(F_n(x)) = \frac{1}{n}F(x)(1-F(x)).$$

Кроме того, эмпирическая функция распределения является строго состоятельной оценкой теоретической функции распределения, т.е.

$$\boldsymbol{P}\left(\lim_{n\to\infty}\sup_x |F_n(x) - F(x)| = 0\right) = 1 \qquad \text{(теорема Гливенко).}$$

Построим теперь характеристическую функцию, соответствующую эмпирической функции распределения $\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx} dF_n(x) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}$.

Каждая реализация эмпирической характеристической функции удовлетворяет следующим условиям:

(1) $\varphi_n(0) = 1$; (2) $|\varphi_n(t)| \le 1$; (3) $\varphi_n(-t) = \overline{\varphi_n(t)}$; (4) $\lim_{|t|\to\infty}\sup |\varphi_n(t)| = 1$;

(5) $\varphi_n(t)$ имеет производные всех порядков и, более того, аналитическая.

Эмпирическая характеристическая функция $\varphi_n(t)$ является оценкой теоретической характеристической функции $\varphi(t)$, т.е. $\boldsymbol{E}(\varphi_n(t)) = \varphi(t)$. Вычислим другие моменты. Мы имеем

$$\boldsymbol{E}((\varphi_n(t_1) - \varphi(t_1))(\varphi_n(t_2) - \varphi(t_2))) = \boldsymbol{E}(\varphi_n(t_1)\varphi_n(t_2)) - \varphi(t_1)\varphi(t_2) = \boldsymbol{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{it_1X_j}\frac{1}{n}\sum_{k=1}^{n} e^{it_2X_k}\right) - \varphi(t_1)\varphi(t_2) =$$

$$= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\boldsymbol{E}\left(e^{it_1X_j}e^{it_2X_k}\right) - \varphi(t_1)\varphi(t_2) = \frac{1}{n^2}\sum_{j=1}^{n}\boldsymbol{E}\left(e^{i(t_1+t_2)X_j}\right) + \frac{1}{n^2}\sum_{j\ne k}\boldsymbol{E}\left(e^{it_1X_j}\right)\boldsymbol{E}\left(e^{it_2X_k}\right) - \varphi(t_1)\varphi(t_2) =$$

$$= \frac{1}{n}(\varphi(t_1+t_2) - \varphi(t_1)\varphi(t_2)).$$

Аналогично,

$$\mathbf{COV}(\varphi_n(t_1)\varphi_n(t_2)) = \boldsymbol{E}((\varphi_n(t_1) - \varphi(t_1))\overline{(\varphi_n(t_2) - \varphi(t_2))}) = \frac{1}{n}(\varphi(t_1 - t_2) - \varphi(t_1)\overline{\varphi(t_2)}).$$

В частности, $\boldsymbol{E}(|\varphi_n(t_1) - \varphi(t)|^2) = \frac{1}{n}(1 - |\varphi(t)|^2)$.

Обозначим действительную и мнимую части функции $\varphi(t)$ через $u(t)$ и $v(t)$ и $u_n(t) = \mathrm{Re}(\varphi_n(t)), v_n(t) = \mathrm{Im}(\varphi_n(t))$.

Тогда $u_n(t)$ $u_n(t) = \frac{1}{n}\sum_{j=1}^{n}\cos(tX_j)$, $v_n(t) = \frac{1}{n}\sum_{j=1}^{n}\sin(tX_j)$.

Имеем: $\boldsymbol{E}(u_n(t)) = u(t)$, $\boldsymbol{E}(v_n(t)) = v(t)$.

Так как $\boldsymbol{E}(\cos(tX)\cos(sX)) = \frac{1}{2}(\boldsymbol{E}(\cos((t+s)X) + \boldsymbol{E}(\cos((t-s)X))$, то

$$\mathbf{COV}(u_n(t_1)u_n(t_2)) = \boldsymbol{E}(u_n(t_1)u_n(t_2)) - u(t_1)u(t_2) =$$

$$= \frac{1}{n^2}\sum_{j=1}^{n}\boldsymbol{E}(\cos(t_1X_j)\cos(t_2X_j)) + \frac{1}{n^2}\sum_{j\neq k}\boldsymbol{E}(\cos(t_1X_j))\boldsymbol{E}(\cos(t_2X_k)) - u(t_1)u(t_2) =$$

$$= \frac{1}{2n}(u(t_1+t_2) + u(t_1-t_2) - 2u(t_1)u(t_2)),$$

аналогично, $\mathbf{COV}(v_n(t_1)v_n(t_2)) = \frac{1}{2n}(u(t_1-t_2) - u(t_1+t_2) - 2v(t_1)v(t_2))$,

$$\mathbf{COV}(u_n(t_1)v_n(t_2)) = \frac{1}{2n}(v(t_1+t_2) + v(t_1-t_2) - 2u(t_1)v(t_2)).$$

Из последних соотношений следует сходимость в среднеквадратичном:

$$\lim_{n\to\infty}\boldsymbol{E}(|\varphi_n(t_1) - \varphi(t)|^2) = 0.$$

8.7 Центральные предельные теоремы

Для последовательности $\{X_n, n \geq 1\}$ независимых и одинаково распределенных случайных величин имеем: $\bar{X}_n \to a$ в слабом или сильном смысле, если $\mathbf{E}(X) = a$ существует. Однако это не дает нам никаких сведений о том, как аппроксимировать распределение $\bar{X}_n$ в больших выборках и как быстро $\bar{X}_n$ сходится к a, т.е. мы не знаем скорость сходимости разности $\bar{X}_n - \bar{a}_n$ к нулю Для изучения этого вопроса рассмотрим последовательность случайных величин $Y_n = \frac{\sqrt{n}(\bar{X}_n - a)}{\sigma}$, где $\sigma -$ некоторая константа (обычно $\sigma = \sqrt{\mathbf{D}(X)}$). При определенных условиях предельная функция распределения этой последовательности равна

$$\Phi(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx,$$

т.е. есть функция распределения нормального закона.

Для доказательства сходимости используют прием, состоящий в нахождении характеристической функции $\varphi_n(t)$ для величины Y_n и установлении того факта, что $\varphi_n(t) \underset{n\to\infty}{\to} e^{-t^2/2}$.

Мы рассматриваем более общую постановку, а именно, рассматриваем сходимость случайных величин $Y_n = \frac{\bar{X}_n - \bar{a}_n}{\sigma_n}$.

Теорема 15 (Линдеберг-Леви) (ц.п.т.). *Пусть* $X_1, X_2, \ldots -$ *последовательность независимых и одинаково распределенных случайных величин таких, что* $\mathbf{E}(X_i) = a$ *и* $\mathbf{D}(X_i) = \sigma^2$ *существуют для любого* i. *Если* $F_n(x) = \mathbf{P}\left(\frac{\sqrt{n}(\bar{X}_n - a)}{\sigma} < x\right)$, *то* $F_n(x) \underset{n\to\infty}{\to} \Phi(x)$.

Д о к а з а т е л ь с т в о. Пусть $\varphi(t) -$ характеристическая функция величины (х. ф.) $X_i - a$.

Поскольку первые два момента существуют, то

$$\varphi(t)=1-\frac{1}{2}\sigma^2t^2+o(t^2).$$

Тогда х. ф. величины $Y_n=\sum_{i=1}^{n}(X_i-a)/(\sigma\sqrt{n})$ равна

$$\varphi_n(t)=\left(\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n=\left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n\underset{n\to\infty}{\to}e^{-t^2/2},\text{ где } \sigma^2=\mathbf{D}(X),$$

откуда следует результат теоремы 9.

Из теоремы Линдеберга-Леви в частности вытекает, что любая равномерно ограниченная последовательность взаимно независимых и одинаково распределенных случайных величин удовлетворяет ц. п. т. Центральная предельная теорема говорит, в каких случаях следует ожидать появления нормальной случайной величины: если величина порождена действием случайных факторов, независимых в своем воздействии, каждая из которых оказывает приблизительно одинаковое и малое воздействие, то есть основание ожидать, что данная случайная величина будет нормально распределена. Следующая теорема доказана Линдебергом и Феллером.

Теорема 16. *Пусть* $\{X_i, i\ge 1\}$ *– последовательность независимых (необязательно одинаково распределенных) случайных величин и,* $\mathbf{D}(X_i)=\sigma_i^2\ne 0$. *Определим* $B_n=\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}$.

Тогда соотношения $\lim_{n\to\infty}\max_{1\le i\le n}\frac{\sigma_i}{B_n}=0$, $\mathbf{P}\left(Y_n<x\right)\underset{n\to\infty}{\to}\Phi(x)$

выполняются в том и только том случае, когда для любого $\varepsilon>0$

$$\lim_{n\to\infty}\frac{1}{B_n^2}\sum_{i=1}^{n}\int_{|x-a_i|>\varepsilon B_n}(x-a_i)^2\,dF_i(x)=0, \qquad (\text{L})$$

где $F_i(x)=\mathbf{P}(X_i<x)$.

Следствие (**Теорема Ляпунова**) (ц.п.т.). *Пусть* $\{X_i, i\ge 1\}$ *– последовательность независимых* (*необязательно одинаково распределенных*) *случайных величин и* $\mathbf{D}(X_i)=\sigma_i^2\ne 0$

$\mathbf{E}(\,|X_i-a_i|^3)=\beta_i^3$ *существуют при всех* $i=1,2,\ldots$. *Положим кроме того*

$C_n=\left(\sum_{i=1}^{n}\beta_i^3\right)^{1/3}$, $B_n=\left(\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}$. *Если* (Л)*:* $\lim_{n\to\infty}\frac{C_n}{B_n}=0$, *то* $\mathbf{P}\left(\frac{\sum_{i=1}^{n}(X_i-a_i)}{C_n}<x\right)\underset{n\to\infty}{\to}\Phi(x)$.

<u>Замечание.</u> Часто рассматривают $\mathbf{E}((X_i-a_i)^4)=\gamma_i^4,\ G_n^4=\sum_{i=1}^{n}\gamma_i^n,\ \frac{G_n^4}{B_n^4}\to 0\ \ (n\to\infty)$.

Необходимо отметить, что наши предельные теоремы и приближенные формулы справедливы только тогда, когда число испытаний фиксировано заранее ***независимо от исхода испытаний***.

Если дисперсии не существуют, то в качестве предельного распределения может уже не быть нормального распределения, в таком случае предельными могут быть так называемые устойчивые распределения.

Покажем, что условие (L) выражает свойство равномерной малости слагаемых. Именно, для любого $\varepsilon>0$ рассмотрим (μ) $\mathbf{P}\left(\max_{1\le k\le n}\frac{|X_k-a_k|}{B_n}>\varepsilon\right)$ (условие пренебрежимой малости).

Имеем: $\mathbf{P}\left(\max\limits_{1\le k\le n}\dfrac{|X_k-a_k|}{B_n}>\varepsilon\right)=\mathbf{P}\left(\bigcup\limits_{k=1}^{n}(\,|X_k-a_k|>\varepsilon B_n)\right)\le.$

$$\le\sum_{k=1}^{n}\mathbf{P}(\,|X_k-a_k|>\varepsilon B_n)\le\sum_{k=1}^{n}\frac{1}{B_n^2}\int\limits_{|x-a_k|>\varepsilon B_n}(x-a_k)^2\,dF_k(x)\underset{n\to\infty}{\to}0 \text{ по условию (L).}$$

Заметим, что без условия (μ) ц.п.т. называется ц.п.т. в *неклассической постановке*.

Покажем также, что если условие Ляпунова (Л) выполнено, то выполнено также и условие (L). Но это ясно из следующей цепочки неравенств:

$$\sum_{k=1}^{n}\frac{1}{B_n^2}\int\limits_{|x-a_k|>\varepsilon B_n}(x-a_k)^2\,dF_k(x)\le\sum_{k=1}^{n}\frac{1}{B_n^2(\varepsilon B_n)}\int\limits_{|x-a_k|>\varepsilon B_n}|x-a_k|^3\,dF_k(x)\le\frac{1}{\varepsilon}\cdot\frac{\sum_{k=1}^{n}\int|x-a_k|^3\,dF_k(x)}{B_n^3}\underset{n\to\infty}{\to}0.$$

Смысл условия Линдеберга поясняет также следующее обстоятельство. Имеем:

$$\sigma_j^2=D(X_j)=\int\limits_{|x|\le\varepsilon}x^2dF_j(x)+\int\limits_{|x|>\varepsilon}x^2dF_j(x)\le\varepsilon^2+\int\limits_{|x|>\varepsilon}x^2dF_j(x),$$

причем $\int\limits_{|x|>\varepsilon}x^2dF_j(x)\underset{n\to\infty}{\to}0$. Следовательно, $\overline{\lim\limits_{n\to\infty}}\max\limits_{j\le n}\sigma_j^2\le\varepsilon^2$, а поскольку $\varepsilon-$ любое, то это означает, что $\max\limits_{j\le n}\sigma_j^2\to0$ при $n\to\infty$. По неравенству Чебышева

$$P(|X_j|>\varepsilon)\le\sigma_j^2/\varepsilon^2\;\Rightarrow\;X_j\xrightarrow{p}0.$$

Более того, $\max P(|X_j|>\varepsilon)\to0$, т.е. сумма S_n состоит из большого числа малых слагаемых и ее предельное распределение невырожденно в силу «эффекта накопляемости» сумм.

Задача (характеризация Пойа). Пусть X_1,X_2-н.о.р. с.в., у которых существуют вторые моменты и пусть $\dfrac{X_1+X_2}{\sqrt{2}}\overset{d}{=}X_1$, т.е. распределение случайной величины $\dfrac{X_1+X_2}{\sqrt{2}}$ имеет такое же распределение, как и у X_1. Покажите, что X_1 имеет нормальное распределение с математическим ожиданием, равным нулю.

Решение. Ясно, что $\boldsymbol{E}(X_1)=0.$ Пусть $\sigma^2=\boldsymbol{D}(X_1).$ Тогда, если X_1,X_2,X_3,X_4-н.о.р.с.в., то

$$S_4=\frac{X_1+X_2+X_3+X_4}{2}\overset{d}{=}\frac{1}{\sqrt{2}}\left(\frac{X_1+X_2}{\sqrt{2}}+\frac{X_3+X_4}{\sqrt{2}}\right)\overset{d}{=}\frac{X_1+X_3}{\sqrt{2}}\overset{d}{=}X_1.$$

Из ц.п.т. известно, что $S_n=(X_1+X_2+\ldots+X_n)/\sqrt{n}\xrightarrow{d}N(0,\sigma^2)$ при $n\to\infty,$ поэтому рассматривая подпоследовательность $n_k=2^k,k\ge1$ и устремляя $k\to\infty$, получим $n_k\to\infty$, откуда

$$X_1\overset{d}{=}S_{n_k}/\sqrt{n_k}\xrightarrow{d}N(0,\sigma^2)\text{ при }k\to\infty.$$

Пример 17. Пусть $\{X_k,1\le k\le n\}-$ n независимых и одинаково распределенных случайных величин, у которых существуют математические ожидания и дисперсии. Для $(x_1,x_2,\ldots,x_n)\in R^n$ определим функции $y_i=o_i(x_1,x_2,\ldots,x_n)$, которые из набора $(x_1,x_2,\ldots,x_n)$ выбирают $i-$ ое по величине число. Тогда $X_n^{(i)}=o_i(X_1,X_2,\ldots,X_n)$ назовем $i-$й порядковой статистикой. Зададим число $0<\alpha<1$ и рассмотрим

$$T_n(\alpha)=\frac{X_n^{([n\alpha])}+\ldots+X_n^{(n)}}{X_n^{(1)}+\ldots+X_n^{(n)}}=\frac{X_n^{([n\alpha])}+\ldots+X_n^{(n)}}{S_n},$$

где $S_n = X_1 + ... + X_n$. По закону больших чисел $\frac{S_n}{n} \xrightarrow[n\to\infty]{p} a = \int\limits_{-\infty}^{\infty} x\,dF(x) = \int\limits_0^1 F^{-1}(x)\,dx$.

Можно показать, что $\frac{X_n^{([n\alpha])} + ... + X_n^{(n)}}{n} \xrightarrow[n\to\infty]{p} \int\limits_{\alpha}^{1} F^{-1}(x)\,dx$. Рассмотрим

$$D_F(\alpha) = \frac{\int\limits_{\alpha}^{1} F^{-1}(x)\,dx}{\int\limits_{0}^{1} F^{-1}(x)\,dx} = \frac{\int\limits_{\alpha}^{1} F^{-1}(x)\,dx}{a}.$$

Пусть $F(x) = 1 - x^{-\theta}$, $x > 1$ $(\theta > 1)$ – распределение Парето, ($y = 1 - x^{-\theta}, 1 - y = x^{-\theta}$, $x = (1-y)^{-1/\theta}$).

Тогда и $\int\limits_{\alpha}^{1} F^{-1}(x)\,dx = \int\limits_{\alpha}^{1} (1-x)^{-\frac{1}{\theta}}\,dx = \frac{(1-\alpha)^{-\frac{1}{\theta}+1}}{1-\frac{1}{\theta}}$, $a = \frac{\theta}{\theta - 1}$, а $D_F(\alpha) = (1-\alpha)^{-\frac{1}{\theta}+1}$.

Для $\theta = 1.16$, $\alpha = 0.8$ получаем $1 - \alpha = 0.2$ (20%) и $D_F(0.8) = 0.2^{-0.1386} = 0.8$.

Считается, что размеры вкладов хорошо описываются распределением Парето. Полученный результат говорит, что при таком значении параметра $\theta = 1.16$ на 20% респондентов приходится 80% суммарных вкладов. Такое соотношение 20 и 80 имеет довольно устойчивый характер.

Центральная предельная теорема утверждает, что если например, величины независимы, одинаково распределены и существуют математическое ожидание и дисперсия, и если $F_n(x)$ – функция распределения $Y_n = \frac{\sqrt{n}\,(\bar{X}_n - a)}{\sigma}$, то $\lim\limits_{n\to\infty} F(x) = \Phi(x)$ – функция распределения стандартного нормального закона.

Пусть случайные величины $X_1, \dots, X_n$ имеют третий μ_3 и четвертый μ_4 центральные моменты и $\varphi(x)$ – плотность стандартного нормального распределения. Тогда имеет место представление (разложение Эджворта):

$$F_n(x) \approx \Phi(x) - \frac{1}{3!}\frac{\mu_3}{\sigma^3}\varphi^{(2)}(x) + \frac{1}{4!}\left(\frac{\mu_4}{\sigma^4} - 3\right)\varphi^{(3)}(x) + \frac{10}{6!}\left(\frac{\mu_3}{\sigma^3}\right)^2 \varphi^{(5)}(x) =$$

$$= \Phi(x) - \frac{1}{3!}\gamma_1(x^2 - 1)\varphi(x) + \frac{1}{4!}\gamma_2(-x^3 + 3x)\varphi(x) + \frac{10}{6!}\gamma_1^2(x^4 - 4x^2 + 3)\varphi(x),$$

где γ_1 – асимметрия, γ_2 – эксцесс. Точность аппроксимации – $O\left(\frac{1}{n}\right)$.

Таким образом, при конечном n мы можем использовать аппроксимацию Эджворта. Отметим также, что если $(m_1, m_2, ..., m_k)$ имеют полиномиальное распределение $\Pi(p_1, ..., p_k)$ и

$$\eta_i = \frac{m_i - np_i}{\sqrt{np_i(1-p_i)}}, \text{ то } \sum_{j=1}^{k}(1-p_j)\eta_j^2 = \sum_{j=1}^{k}(1-p_j)\frac{(m_j - np_j)^2}{np_j(1-p_j)} = \sum_{j=1}^{k}\frac{(m_j - np_j)^2}{np_j} \xrightarrow[n\to\infty]{d} \chi^2_{k-1}.$$

Определение. *Говорят, что х.ф.* $\varphi(t)$ *является* <u>*безгранично делимой*</u> (*и, соответственно, распределения*)*, если для любого натурального* n *существует такая х.ф.* $\varphi_n(t)$*, что*

$$\varphi(t) = (\varphi_n(t))^n.$$

Заметим, что для независимых и одинаково распределенных с.в. в качестве предельных распределений могут выступать только безгранично делимые распределения.

Примеры безгранично делимых х.ф.

1) Вырожденное распределение $\varphi(t) = e^{iat}$, $\varphi_n(t) = e^{iat/n}$.

2) Распределение Пуассона $\varphi(t)=\exp(\lambda(e^{iat}-1)),\quad \varphi_n(t)=\exp\left(\frac{\lambda}{n}(e^{iat}-1)\right).$

3) Отрицательное биномиальное распределение

$$\varphi(t)=\left(p(1-qe^{it})^{-1}\right)^{m},\quad \varphi_n(t)=\left(p(1-qe^{it})^{-1}\right)^{m/n}.$$

4) Нормальное распределение $\varphi(t)=\exp\left(iat-\frac{\sigma^2t^2}{2}\right),\quad \varphi_n(t)=\exp\left(\frac{iat}{n}-\frac{\sigma^2t^2}{2n}\right).$

5) Распределение Коши $\varphi(t)=e^{iat-\theta|t|},\quad \varphi_n(t)=e^{iat/n-\theta|t|/n}.$

6) Гамма-распределение $\varphi(t)=\left(1-\frac{it}{\theta}\right)^{-\lambda},\quad \varphi_n(t)=\left(1-\frac{it}{\theta}\right)^{-\lambda/n}.$

Приведем некоторые утверждения, касающиеся б.д.р. и которые нетрудно доказать (или посмотреть, как это делается в книге [22], с.136-143. Здесь же есть примеры.).

(i) *Произведение любого конечного числа б.д.х.ф. безгранично делима.*

(ii) *Х.ф., являющаяся пределом б.д.х.ф., безгранично делима.*

(iii) *Пусть* $g(t)-$ *произвольная х.ф., а* $p-$ *произвольное положительное число. Тогда* $\varphi(t)=\exp(p(g(t)-1))-$ *безгранично делима.*

(iv) (теорема де Финетти) *Х.ф. является безгранично делимой тогда и только тогда, когда она представима в виде*

$$\varphi(t)=\lim_{m\to\infty}\exp(p_m(g_m(t)-1)),$$

где p_m- *положительные числа, а* $g_m(t)-$ *некоторые х.ф.*

<u>Свяжем</u> смесь распределений с геометрическими случайными величинами. Пусть с.в. Y обладает следующим свойством: для любого $p\in(0,1)$ найдется такая с.в. X_p, что

$$Y\overset{d}{=}X_p+\varepsilon_pY,$$

где X_p,Y,ε_p – независимы, $\boldsymbol{P}(\varepsilon_p=1)=p,\quad \boldsymbol{P}(\varepsilon_p=0)=1-p.$ Знак $\overset{d}{=}$ означает равенство распределений. Тогда

$$\boldsymbol{P}(Y<x)=F_Y(x)=\boldsymbol{P}(X_p+\varepsilon_pY<x)=pF_{X_p}(x)+(1-p)F_Z(x),\ Z=X_p+Y.$$

Если $\varphi(t)-$х.ф. с.в. Y, $g_p(t)-$х.ф. с.в. X_p, то

$$\varphi(t)=g_p(t)(p+(1-p)\varphi(t))\Rightarrow\varphi(t)=\sum_{j=1}^{\infty}g_p^j(t)p(1-p)^{j-1},\ \text{т.е.}\ \ Y=\sum_{j=1}^{\nu_p}X_p^{(j)},$$

для некоторых независимых и одинаково распределенных с.в. $X_p^{(1)},X_p^{(2)},\ldots$.

8.8 Метод Стейна

В конце 60-х годов, неудовлетворенный доказательствами центральной предельной теоремы (ц.п.т.) Чарльз Стейн (Ch. Stein) разработал новый способ доказательства ц.п.т., который стал популярен в последнее время для зависимых случайных величин, более того, он дает скорость сходимости допредельного и предельного распределений. Дальнейшие разработки провел А. Тихомиров. Сформулируем (без доказательства) ряд утверждений, на которых основывается доказательство ц.п.т. и которая использует метрику

$$d(\boldsymbol{P},\boldsymbol{Q})=\sup_{h\in\mathscr{H}}\left|\int hd\boldsymbol{P}-\int hd\boldsymbol{Q}\right|=\sup_{h\in\mathscr{H}}\left|\boldsymbol{E}(h(W))-\boldsymbol{E}(h(Y))\right|.$$

Здесь $\boldsymbol{P}$ и $\boldsymbol{Q}$ – вероятностные меры, заданные на измеримом пространстве.

Лемма 1. *Случайная величина* Z *имеет стандартное нормальное распределение если и только если*

$$\boldsymbol{E}(f'(Z)) - \boldsymbol{E}(Z \cdot f(Z)) = 0 \qquad (1)$$

для любой непрерывной или кусочно непрерывно дифференцируемой функции f*, для которой ожидания* (1) *существуют.*

Кратко мы рассмотрим *доказательство* этой леммы.

1) Если $Z \in N(0,1)$, то беря интеграл по частям, получим

$$\boldsymbol{E}(f'(Z)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f'(x) e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x f(x) e^{-x^2/2} dx = \boldsymbol{E}(Z \cdot f(Z)).$$

2) Пусть (1) выполнено для любой непрерывно дифференцируемой функции. Не нарушая общности,рассмотрим следующие функции: $f(x) = x^k, k \in N.$ Имеем: $f'(x) = kx^{k-1}, xf(x) = x^{k+1}.$ Если потребуем, чтобы математическое ожидание было равно нулю, а дисперсия 1 (т.е. второй начальный момент), то для нечетных k моменты будут равны 0, а для четных $k = 2m$, эти моменты будут равны $(2m-1)!!$ и так как условие Карлемана будет выполнено, то плотность по моментам восстанавливается однозначно.

Другое доказательство приведено в [21], с.14-15, которое основано на следующей лемме.

Лемма 1а. *Для фиксированного* $z \in \mathbf{R}$ *и* $\boldsymbol{P}(Z < z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$*, единственное решение* $f(w) \equiv f_z(w)$ *уравнения*

$$f'(w) - wf(w) = \mathbf{1}_{\{w<z\}} - \Phi(z) \qquad (\mathrm{A})$$

есть $$f_z(w) = \begin{cases} \sqrt{2\pi}\, e^{w^2/2} \Phi(w)(1-\Phi(w)), & w < z, \\ \sqrt{2\pi}\, e^{w^2/2} \Phi(z)(1-\Phi(w)), & w \geq z. \end{cases}$$

Доказательство. Умножая обе части равенства (A) на $e^{-w^2/2}$, получим

$$(e^{-w^2/2} f(w))' = e^{-w^2/2}(\mathbf{1}_{\{w<z\}} - \Phi(z)).$$

Интегрируя, получим

$$f_z(w) = e^{w^2/2} \int_{-\infty}^{w} (\mathbf{1}_{\{x<z\}} - \Phi(z)) e^{-x^2/2} dx = -e^{w^2/2} \int_{w}^{\infty} (\mathbf{1}_{\{x<z\}} - \Phi(z)) e^{-x^2/2} dx,$$

что доказывает лемму 1а.

Доказательство леммы 1. Из соотношения

$$0 = \boldsymbol{E}(f_z'(W) - Wf_z(W)) = \boldsymbol{E}(\mathbf{1}_{\{w<z\}} - \Phi(z)) = \boldsymbol{P}(W < z) - \Phi(z)$$

следует результат второй части леммы 1. Пусть $\|h\| = \sup_x |h(x)|$.

Приведем еще одно решение. Возьмем $f(x) = \cos(tx) + i\sin(tx).$ Тогда

$$f'(x) = -t\sin(tx) + it\cos(tx) = it(\cos(tx) + i\sin(tx)),$$

поэтому

$$\boldsymbol{E}(f'(Z)) = it\,\boldsymbol{E}(\cos(tZ) + i\sin(tZ)) = it \cdot \varphi_Z(t) \equiv it\varphi(t) = it\boldsymbol{E}(Z \cdot f(Z)) = it\boldsymbol{E}(Z \cdot e^{itZ})$$

и $\quad \varphi'(t) = i\boldsymbol{E}(Ze^{itZ}) \Rightarrow \varphi'(t) = -t\varphi(t).$

Так как $\varphi(0) = 1$, то решением последнего уравнения является $\varphi(e) = e^{-t^2/2}$, которая является характеристической функцией стандартного нормального распределения.

Лемма 2. *Пусть величина* h *есть ограниченная кусочно непрерывно дифференцируемая действительнозначная функция. Функция*

$$f(y) = e^{y^2/2} \int_{-\infty}^{y} (h(x) - \boldsymbol{E}(h(Y)) e^{-x^2/2} dx \qquad (8)$$

есть решение линейного первого порядка дифференциального уравнения

$$f'(x) - xf(x) = h(x) - \boldsymbol{E}(h(Y)),\ \ \boldsymbol{E}(h(Y)) < \infty \ \textit{для некоторой с.в. } Y,$$

и а) $\|f\| \le \sqrt{2\pi}\,\|h\|,$ б) $\|f'\| \le 2\,\|h\|,$ в) $\|f''\| \le 2\,\|h'\|.$

Предположим, что существует другая случайная величина $W,$ такая, что

$$\boldsymbol{E}(W \cdot f(W)) = \boldsymbol{E}(f'(Y)) \qquad (9)$$

для любой непрерывно дифференцируемой функции f, поэтому

$$|\boldsymbol{E}(h(W)) - \boldsymbol{E}(h(Y))| = |\boldsymbol{E}(f'(W)) - \boldsymbol{E}(W \cdot f(W))| = |\boldsymbol{E}(f'(W)) - \boldsymbol{E}(f'(Y))|,$$

откуда $|\boldsymbol{E}(h(W)) - \boldsymbol{E}(h(Y))| \le \|f''\| \cdot \boldsymbol{E}(|W - Y|).$

В качестве примера рассмотрим $W = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i,\ S_n = \sum_{i=1}^{n} X_i$, где $X_i -$ нор св, удовлетворяющие условию $\boldsymbol{E}(X_i) = 0,$ $\boldsymbol{D}(X_i) = 1.$ Пусть $\{X_n^*\}_{n\ge 1}$ будут нор св с условием $\boldsymbol{E}(X_n f(X_n)) = \boldsymbol{E}(f'(X_n^*))$. Ясно, что это условие будет выполнено тогда и для функции $f(x+a)$ для любого $a \in R.$ Определим $Y = \frac{(S_{n-1} + X_n^*)}{\sqrt{n}}$. Тогда

$$\boldsymbol{E}(f'(Y)) = \boldsymbol{E}(f'((S_{n-1} + X_n^*)/\sqrt{n})) = \boldsymbol{E}(\sqrt{n} X_n f((S_{n-1} + X_n)/\sqrt{n})) =$$

$$= \boldsymbol{E}\left(n \frac{X_n}{\sqrt{n}} f((S_{n-1} + X_n)/\sqrt{n}) \right) = \boldsymbol{E}\left(n \frac{X_n}{\sqrt{n}} f(W) \right) = \boldsymbol{E}(Wf(W)),$$

поскольку беря сначала условное ожидание при фиксированном значении S_{n-1} для функции $\sqrt{n} f(x+a)$, получим второе равенство, а последнее равенство выполнено поскольку $X_1, \dots, X_n$ одинаково распределены (в силу симметричности соотношения).

Пусть $0 < p < 1, q = 1 - p$ и рассмотрим случайную величину X_i, принимающую значения $-x_0 = -\sqrt{q/p}$ с вероятностью p и $x_1 = \sqrt{p/q}$ с вероятностью q. Пусть X_i^* имеет равномерное распределение $U(-x_0, x_1)$. Тогда математическое ожидание этой с.в. равно 0, а дисперсия 1 и

$$\boldsymbol{E}(X_i f(X_i)) = -px_0 f(-x_0) + qx_1 f(x_1) = \boldsymbol{E}(f'(X_i^*)) = \frac{1}{x_0 + x_1}\int_{-x_0}^{x_1} f'(x)\,dx = \frac{1}{x_0 + x_1} f(x)\Big|_{-x_0}^{x_1} = \frac{f(x_1) - f(-x_0)}{x_0 + x_1}.$$

В [21], с.27 рассмотрена непосредственно Бернулиевская $B(1, p), 0 < p < 1,$ случайная величина ξ со значениями $\{0,1\}$. Центрируем её, полагая $X = \xi - p$ и деля на $\sigma = \sqrt{p(1-p)} = \sqrt{pq}$. Тогда

$$\boldsymbol{E}(X \cdot f(X)) = \boldsymbol{E}((\xi - p) f(\xi - p)) = pqf(q) - qpf(p) = \sigma^2(f(q) - f(-p)) =$$

$$= \sigma^2 \int_{-p}^{q} f'(u)\,du = \sigma^2 \boldsymbol{E}(f'(x)),$$

для $U -$ равномерно распределенной на $[-p, 1-p].$

Теперь $|W - Y| = |X_n - X_n^*|/\sqrt{n} \le 1/\sqrt{npq}$, откуда получаем:

$$|\boldsymbol{E}(h(W)) - \boldsymbol{E}(h(Y))| \le 2\|h'\|/\sqrt{npq}.$$

Так как $\frac{1}{\sqrt{npq}} \underset{n\to\infty}{\to} 0,$ *то имеет ц.п.т. для нашей с.в. и для распределения Бернулли.*

<u>Как находить</u> такие пары, чтобы было выполнено условие (S) ? В [21] показано, что если выполнено условие $\boldsymbol{E}(Y|W) = (1-\lambda)W, 0 < \lambda < 1,$, то пара (Y, W) будет искомой парой.

Например, если (Y,W) будет иметь совместное нормальное распределение с ожиданиями 0, дисперсией 1 и коэффициентом корреляции $0<\rho<1$, то $\boldsymbol{E}(Y\,|\,W)=\rho W,\ \lambda=1-\rho.$

Другой пример состоит в следующем. Пусть $\{\xi_i, 1\le i\le n\}$ – независимые, одинаково распределенные случайные величины с ожиданием 0 и $\sum_{i=1}^n \boldsymbol{E}(\xi_i^2)=1$. Положим $W=\sum_{i=1}^n \xi_i$. Пусть τ – случайная величина, независимая от $\{\xi_i, 1\le i\le n\}$, имеющая равномерное распределение на множестве $\{1,2,...,n\}$. Тогда для $Y=W-\xi_\tau$ имеет место $\boldsymbol{E}(Y\,|\,W)=\left(1-\dfrac{1}{n}\right)W,\ \lambda=\dfrac{1}{n}.$

В [21], с.23 рассмотрен пример: $Y=W-\xi_\tau+\xi_\tau^*$, где $\{\xi_i^*, 1\le i\le n\}$ независимы от $\{\xi_i, 1\le i\le n\}$ и одинаково с ними распределены. Там же на с.27 указано, как можно построить такую случайную величину.

Именно,

если случайная величина X имеет математическое ожидание 0 и дисперсию $\sigma^2>0$, то существует случайная величина X^*, такая, что

$$\boldsymbol{E}(f'(X^*))=\sigma^2\boldsymbol{E}(X\cdot f(X))$$

для каждой абсолютно непрерывной функции f с $\boldsymbol{E}(|X\cdot f(X)|)<\infty.$

Кроме того, распределение с.в. X^* абсолютно непрерывно с плотность

$$g^*(x)=\boldsymbol{E}(X\cdot\mathbf{1}_{(X\ge x)})/\sigma^2=-\boldsymbol{E}(X\cdot\mathbf{1}_{(X<x)})/\sigma^2$$

и функцией распределения $\quad G^*(x)=\boldsymbol{E}(X(X-x)\cdot\mathbf{1}_{(X<x)})/\sigma^2.$

Рассмотрим еще один пример построения такой случайной величины. Именно,

Пусть с.в. ξ является симметричной случайной величиной с конечной ненулевой дисперсией σ^2 и пусть ξ^* имеет следующую функцию распределения: $dG^*(x)=\dfrac{x^2dF(x)}{\sigma^2}$, а $U\in\mathscr{R}(-1,1)$ и независима от ξ^*. Тогда $X^*\overset{d}{=}U\xi^*$ составляет пару Стейна с X. Действительно, возьмем непрерывно дифференцируемую f и пусть $g(x)=f'(x)$. Мы имеем:

$$\sigma^2\boldsymbol{E}(g(U\xi^*))=\sigma^2\boldsymbol{E}(f'(U\xi^*))=\frac{\sigma^2}{2}\boldsymbol{E}\left(\int_{-1}^1 f'(u\xi^*)\,du\right)=\frac{\sigma^2}{2}\boldsymbol{E}\left(\frac{f(\xi^*)-f(-\xi^*)}{\xi^*}\right)=$$

$$=\frac{1}{2}\boldsymbol{E}\left(X^2\,\frac{f(X)-f(-X)}{X}\right)=\frac{1}{2}\boldsymbol{E}\big(X(f(X)-f(-X))\big)=\frac{1}{2}\boldsymbol{E}\big(X\cdot f(X)-X\cdot f(-X))\big)=\boldsymbol{E}\big(X\cdot f(X)\big).$$

Следовательно, $\quad \sigma^2\boldsymbol{E}(g(U\xi^*))=\boldsymbol{E}\big(X\cdot f(X)\big)=\boldsymbol{E}(f'(X^*))=\sigma^2\boldsymbol{E}(g(X^*)).$

Значит, величины $U\xi^*$ и X^* одинаково распределены.

Метод Стейна позволяет также оценить скорость сходимости. Именно, пусть $\mathscr{L}(W)$ – распределение случайной величины W. Тогда, используя разложение Тейлора можно показать, что

$$\boldsymbol{d}(\mathscr{L}(W),N(0,1))\le\|f''\|\sum_{i=1}^n\left(\frac{1}{2}\boldsymbol{E}(|X_iX_{A_i}^2|)+\boldsymbol{E}(|X_iX_{A_i}X_{B_i\setminus A_i}|)+\boldsymbol{E}(|X_iX_{A_i}|)\boldsymbol{E}(|X_{B_i}|)\right),$$

где $\{1,2,...,n\}\supset A_i$ – такое множество, что X_i независимо от с.в. X_j с $j\notin A_i$, $\{1,2,...,n\}\supset B_i$ – множество тех $X_j\in A_i$, которые независимы от X_k с $k\notin B_i$.

Для $\{1,2,...,n\}\supset A$ определим $X_A=\sum_{j\in A}X_j$. Пусть $\|h'\|\le 1$, тогда $\|f''\|\le 2.$

Таким образом,

$$d(\mathscr{L}(W), N(0,1)) \le 2\sum_{i=1}^{n}\left(\frac{1}{2}\boldsymbol{E}(|X_i X_{A_i}^2|) + \boldsymbol{E}(|X_i X_{A_i} X_{B_i \setminus A_i}|) + \boldsymbol{E}(|X_i X_{A_i}|)\boldsymbol{E}(|X_{B_i}|)\right),$$

Для независимых $\{X_i\}$ имеем $A_i = B_i = \{i\}$ и $d(\mathscr{L}(W), N(0,1)) \le \dfrac{3\boldsymbol{E}(|X_1^3|)}{\sqrt{n}}$.

Так для нашего примера $d(\mathscr{L}(W), N(0,1)) \le \dfrac{3(p^2+q^2)}{\sqrt{npq}}$.

Заметим, что для каждого распределения характеристические распределения будут отличаться. Так для распределения Пуассона характеристическое уравнение таково: случайная величина W, принимающая значения 0, 1, 2, … имеет распределение Пуассона тогда и только тогда, когда для всякой ограниченной функции $f(x)$, определенной на множестве $\{0,1,2,...\}$ имеет место равенство

$$\boldsymbol{E}(\lambda f(X+1)) = \boldsymbol{E}(Xf(X)), \lambda > 0.$$

Лемма 2а. *Если* $Z \in \mathscr{P}(\lambda)$ (*т.е. имеет распределение Пуассона*), *тогда для любой ограниченной функции* $f : \mathbf{N}\bigcup\{0\} \to \mathbf{R}$, $\boldsymbol{E}(\lambda f(Z+1)) - \boldsymbol{E}(Zf(Z)) = 0, \lambda > 0.$

Доказательство. Имеем:

$$\boldsymbol{E}(\lambda f(Z+1)) = \lambda e^{-\lambda}\sum_{k=0}^{\infty} f(k+1)\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{j=1}^{\infty} f(j)\frac{\lambda^j}{(j-1)!} = e^{-\lambda}\sum_{j=1}^{\infty} j\cdot f(j)\frac{\lambda^j}{j!} =$$

$$e^{-\lambda}\sum_{j=0}^{\infty} j\cdot f(j)\frac{\lambda^j}{j!} = \boldsymbol{E}(Zf(Z)).$$

Для биномиального распределения для всякой функции $f(x)$, определенной на множестве $\{-1,0,1,2,...,n\}$ имеет место равенство

$$\boldsymbol{E}(p(n-X)f(X)) = \boldsymbol{E}(qXf(X-1)), 0 < p < 1, q = 1-p.$$

9. Проверка гипотез о распределении

9.1 Понятие статистической гипотезы. Критическая область, размер критерия

Не зная закона распределения случайной величины, статистик не сможет сделать каких-либо выводов о величине допущенной им случайной ошибки или же об относительной частоте появления выборочной характеристики в известных границах.

Обычно исследователь не располагает полными списками причин и следствий, имея дело со сложными взаимодействиями явлений. Поэтому полная индукция в практике научной работы заменяет неполная индукция, которая характеризуется умозаключением, в котором общий вывод обо всех фактах данного рода делается на основании изучения существенных признаков лишь у части таких фактов. Выводы, основанные на неполной индукции, вполне научны. Они покоятся на важнейшем положении Бэконовской логики, согласно которому связь между фактами, находящимися в причинной зависимости, открывается в многочисленных наблюдениях всегда более постоянно, чем связь между фактами, носящая случайный характер.

Понятие гипотезы, с которым имеет дело статистика более узко, чем общее понятие научной гипотезы. Научная гипотеза выдвигается обычно в результате наблюдения над новыми фактами. Чаще всего она возникает как вероятностное умозаключение по аналогии, делаемое на основании тождества между явлениями или их группами. Гипотеза может быть выдвинута как догадка, подсказанная каким-либо обнаруженным в изучаемом явлении порядком, который нельзя приписать случайности.

Возникая в результате наблюдения за фактами, гипотеза затем приобретает характер теоретического допущения. И пока она выражена в такой форме, исследователь еще лишен возможности её проверить. Для этого он должен вновь обратиться к фактам. В таком обосновании гипотетического построения фактами и в последующей его проверке фактами выражается диалектическое единство индукции и дедукции.

Статистические гипотезы касаются поведения наблюдаемых случайных величин.

Более точно, пусть имеется набор независимых случайных величин $X_1, X_2, \ldots, X_n$. Этот набор можно представить в виде точке в $n-$мерном выборочном пространстве. Случайной точке X соответствует некоторое распределение вероятностей, и для любой области W выборочного пространства $\mathbf{R}^n$ можно (по крайней мере, в принципе) вычислить вероятность $\mathbf{P}(X \in W)$ того, что выборочная точка X попадет в W. Любая гипотеза, связанная с $\mathbf{P}(X \in W)$ будет называться статистической. Иными словами, статистическая гипотеза есть предположение о виде неизвестного нам распределения.

Статистическая гипотеза возникает в первую очередь на основе всестороннего качественного анализа изучаемого явления. Так например. Если качественный анализ природы случайно величины X позволяет заключить, что X возникает под действием многих случайных факторов, равноправных, малых в своем воздействии, причем факторы эти действуют независимо друг от друга, то согласно центральной предельной теоремы можно предположить, что случайная величина X имеет нормальное распределение. По возможности выбор статистической модели должен основываться на понимании рассматриваемого физического явления и применении критериев. Лишь в крайнем случае, да и то с большой осторожностью – только лишь на критериях.

Примерами статистических гипотез могут служить следующие гипотезы:

а) случайная величина X имеет нормальное распределение с заданными средним и заданной дисперсией;

б) случайная величина X имеет нормальное распределение с заданными средним (о дисперсии ничего не говорится, т.е. считается, что она неизвестна);

в) случайная величина X имеет нормальное распределение (с какими-нибудь средним и дисперсией.

Чтобы проверить какую-либо гипотезу исходя из случайной выборки наблюдений, мы должны разделить все выборочное пространство (т.е. возможные наборы наблюдений) на две области. Если наблюдаемая выборочная точка X попадет в одну из этих областей, скажем W, то гипотеза отвергается, если же X попадает в дополнительную область $\mathbf{R}^n \setminus W$, то гипотеза принимается. Область W называется ***критической областью***, область $\mathbf{R}^n \setminus W$ – ***областью принятия гипотезы***.

Если известно распределение вероятностей наблюдений, соответствующее проверяемой гипотезе H_0, то можно определить W так, чтобы при выполнении гипотезы H_0 вероятность отвергнуть гипотезу (т.е. вероятность попадания X в W) не превосходила заранее заданной величины α, т.е.

$$\mathbf{P}(X \in W \mid H_0) \le \alpha .$$

Вероятность $\mathbf{P}(X \in W \mid H_0)$ называется ***размером критерия (критической области)***, а α ***уровнем значимости.***

Проверяемая гипотеза часто называется «нулевой гипотезой». Альтернативная гипотеза обозначается H_1 или K.

Что произойдет, когда выполнена какая-то другая гипотеза? Иными словами, невозможно ответить на вопрос, говорят ли полученные наблюдения в пользу данной гипотезы, если мы не знаем, с какими альтернативными гипотезами она сравнивается. Вполне может случиться, что выборка довольно неправдоподобна, если справедлива исходная гипотеза: однако она может быть более неправдоподобной, если выполнена другая гипотеза. Если ситуация такова, что мы должны выбрать либо одну гипотезу, либо другую, мы очевидно, выберем первую, несмотря на неправдоподобность наблюдений. Задача проверки гипотезы, в сущности, состоит в выборе между ней и другими гипотезами. Отсюда немедленно следует, что решение о принятии или непринятии исходной гипотезы существенно зависит от того, против каких альтернатив она проверяется.

9.2 Современные методы проверки гипотез о распределении

Существует несколько методов проверки гипотез. Очень простым, но довольно ненадежным методом проверки является использование для этой цели *вероятностных бумаг* (метод спрямления). В современных статистических пакетах на персональных компьютерах предусмотрена такая процедура (например, пакет STATISTICA – Статистический анализ и обработка данных в среде Windows). Вероятностная бумага – это сеть линий, специальным образом нанесенных на бумагу, так, чтобы функция распределения была бы прямой линией. Каждому распределению соответствует своя вероятностная бумага. Если, например, гипотеза состоит в том, что среднее и дисперсия имеют заданное значение, а распределение является нормальным, то берется вероятностная бумага, соответствующая нормальному распределению. На этой бумаге в виде точек наносятся результаты наблюдений. Если после этого все точки ложатся приблизительно на одну прямую, то гипотезу H_0 можно считать справедливой, в противном случае её можно забраковать и попробовать другую вероятностную бумагу. Ввиду своей невысокой надежности этот метод можно использовать как дополнительный довод (в случае, если данная бумага не отвергает гипотезу) в пользу H_0. Использование вероятностной бумаги является субъективным методом, поскольку определение того, не противоречат ли полученные данные принятой статистической модели основываются на визуальном наблюдении. Вопрос о том, какую линию считать прямой является субъективным делом, и два человека, рассматривающие один и тот же график, могут прийти к различным заключениям.

Посмотрим как устроена вероятностная бумага для случая показательного распределения

$$F(x) = 1 - e^{-x/\theta} \Rightarrow \frac{1}{1-F(x)} = e^{x/\theta} \Rightarrow -\ln(1-F(x)) = \frac{x}{\theta}.$$

Таким образом, для случайной величины, распределенной по показательному закону, график $-\ln(1-F(x))$ как функции натурального логарифма будет иметь вид прямой. Чтобы избежать вычисления логарифмов, оси вероятностной бумаги проградуированы таким образом, что по оси ординат, соответствующей $-\ln(1-F(x))$ можно наносить значения i/n, а по оси абсцисс можно наносить результаты упорядоченной выборки.

Другим методом является проверка гипотез с помощью критериев или тестов. Рассмотрим какое-либо расстояние между двумя функциями распределения $\Delta = \rho(F, F_0)$, где $F(x)$ – истинная функция распределения, а $F_o(x)$ – предполагаемая. Например, в качестве такого расстояния можно выбрать:

а) Колмогоровское расстояние $\rho_K(F, F_0) = \sup\limits_{x \in R} |F(x) - F_0(x)|$;

б) $\rho_S(F, F_0) = \int\limits_{-\infty}^{\infty} (F(x) - F_0(x))^2 \, dF_0(x)$;

в) расстояние полной вариации $\rho_V(P, Q) = \sup\limits_{A \in \mathscr{B}} |\mathbf{P}(A) - \mathbf{Q}(A)|$,

если $F(x)$ и $Q(x)$ – непрерывные функции распределения, соответствующие распределениям $\mathbf{P}$ и $\mathbf{Q}$ с плотностями $f(x)$ и $q(x)$ соответственно, тогда $2\rho_V(F, Q) = \int\limits_{-\infty}^{\infty} |f(x) - q(x)| dx$;

г) расстояние Леви-Прохорова

$\rho_{LP}(F, G) = \inf\{\varepsilon > 0 : F(x-\varepsilon) - \varepsilon \le G(x) \le F(x+\varepsilon) + \varepsilon, \forall x \in R\}$;

д) расстояние Кульбака-Лейблера $\rho_{KL}(Q,P)=\int\limits_{-\infty}^{\infty} q(x)\ln\frac{q(x)}{p(x)}dx$;

е) расстояние Хелингера $\rho_H^2(F,Q)=\int\limits_{-\infty}^{\infty}(\sqrt{f(x)}-\sqrt{q(x)})^2 dx$.

Пример. Пусть $f(x)=e^{-x}, x\geq 0,$ и $f(x)=0, x<0,$ а $q(x)=1, 0\leq x\leq 1,$ и ноль в противном случае. Тогда

$$2d_V(\exp,R[0,1])=\int_0^1(1-e^{-x})\,dx+\int_1^\infty e^{-x}dx=1+\frac{1}{e}-1+\frac{1}{e}=\frac{2}{e}\approx 0.7358, d_V=0.3679,$$

$$d_K(P,Q)=\sup_{x\geq 0}|1-e^{-x}-x\cdot I(0\leq x\leq 1)-I(x>1)|=e^{-1}\approx 0.1839,$$

$$d_H^2(P,Q)=2\left(1-\int_0^1\sqrt{e^{-x}}dx\right)=2\left(\frac{2}{\sqrt{e}}-1\right)\approx 0.4261,$$

$$d_{KL}(Q,P)=\int_0^1\ln\frac{1}{e^{-x}}dx=\frac{1}{2}=0.5, \qquad d_{LP}(P,Q)\approx 0.2785.$$

Соотношения между расстояниями

$$d_H^2(P,Q)\leq 2d_V(P,Q)\leq 2d_H(P,Q), \quad \frac{1}{2}d_V^2(P,Q)\leq d_{KL}(P,Q),$$

$$d_{LP}(P,Q)\leq d_K(P,Q)\leq d_V(P,Q).$$

По величине Δ мы можем судить, насколько близки F и F_0.

Поскольку истинное распределение нам неизвестно, то мы не можем, следовательно, подсчитать Δ. Поэтому мы найдем эмпирическую функцию распределения $F_n(x)$, которая служит приближением к истинной функции распределения когда n достаточно велико. Тогда мера отклонения $F_n(x)$ от $F_0(x)$ есть $\Delta_n=\rho(F_n,F_0)$. Хотя эту меру можно определить разными способами, но каждая мера отклонений Δ_n будет функцией от выборочных значений (Δ_n называется еще *статистикой теста*) и потому сама будет являться случайной величиной с определенным распределением. Теперь по величине статистики Δ_n нельзя уже сделать достоверных выводов, так как Δ_n может отклониться от 0 случайно, даже если $H_0: F(x)=F_0(x)$ справедлива и было бы ошибкой отвергнуть H_0. Поэтому здесь используют такой принцип (принцип фальсификации): если значение Δ_n слишком велико, больше некоторого C, т.е. $\Delta_n>C$, то маловероятно (при соответствующем подобранном C), что это случайность, правдоподобнее, что гипотеза H_0 неверна. Таким образом, критическая область определяется неравенством $\Delta_n>C_\alpha$, причем C_α подобрано таким образом, чтобы размер критической области был не выше выбранного значения уровня значимости α. Уровень значимости α выбирается таким образом, чтобы события с вероятностью α считать практически невозможными. Если в действительности $\Delta_n>C_\alpha$, то гипотезу H_0 следует отвергнуть, так как если бы она была верна, то этого практически бы не произошло. Следует отметить, что для данного метода при определении критической области требуется знания только одной статистики, а не всей совокупности наблюдений. Для того, чтобы находить C_α нужно знать функцию распределения величины Δ_n при условии, что гипотеза H_0 верна, т.е. рассчитанная по функции распределения $F_0(x)$. К этой задаче и сводится построение критерия (теста).

9.3 Критерий согласия χ^2 Пирсона. Двухэтапный критерий согласия χ^2.

Статистика Δ_n хи-квадрат критерия Пирсона задается следующим образом. Предположим, что область значений величины X некоторым образом разбита на k взаимно непересекающихся классов (на практике они обычно выбираются как последовательные интервалы в области значений X, хотя это и не обязательно). Поскольку гипотетическая функция распределения $F_0(x)$ известна (предполагается, что проверяем гипотезу $H_0: F(x) = F_0(x)$, которая является *простой*, т.е. задает единственное распределение), то мы можем вычислить вероятности попадания наблюдений в каждый из классов. Если обозначить их через p_{i0} $(i = 1, 2, \dots, k)$, а наблюденные частоты в k классах через m_i / n $\left(\sum_{i=1}^{k} m_i = n \right)$, то статистика Пирсона имеет вид:

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_{i0}} \left(\frac{m_i}{n} - p_{i0} \right)^2 = \sum_{i=1}^{k} \frac{(m_i - np_{i0})^2}{np_{i0}} = \frac{1}{n} \sum_{i=1}^{k} \frac{m_i^2}{p_{i0}} - n \ .$$

Асимптотически при $n \to \infty$ и при условии, что гипотеза H_0 верна, величина χ^2 имеет хи-квадрат распределение с $(k-1)$ степенями свободы. Для распределения $\chi^2(k-1)$ существуют таблицы, по которым можно найти критическое значение $C_\alpha = \chi^2_{1-\alpha}(k-1)$ – квантиль порядка $1-\alpha$ хи-квадрат распределения с $k-1$ степенью свободы. Поскольку статистика χ^2 имеет распределение $\chi^2(k-1)$ лишь в пределе, то обычно требуют, чтобы число попаданий в каждый класс было не меньше $5 \div 10$ (другое требование состоит в том, чтобы np_{i0} было относительно большим). Однако оказывается, что ошибка приближения, возникающего, когда χ^2 считают распределенной как $\chi^2(k-1)$ особенно мала, когда np_{i0} равны или близки между собой; тогда можно не требовать, чтобы они были большими.

Действительно, рассмотрим математическое ожидание и дисперсию статистики

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_{i0})^2}{np_{i0}}$$

при распределении $F_1(x)$, т.е. при вероятностях p_{i1}. Имеем:

$\mathbf{E}(\chi^2) = \sum_{i=1}^{k} \frac{1}{np_{i0}} \mathbf{E}(m_i - np_{i0})^2 = \frac{1}{n} \sum_{i=1}^{k} \frac{1}{p_{i0}} \mathbf{E}(m_i^2) - n$. Но $\mathbf{E}(m_i^2) = np_{i1}(1-p_{i1}) + n^2 p_{i1}^2$. Следовательно,

$$\mathbf{E}(\chi^2) = \sum_{i=1}^{k} \frac{p_{i1}(1-p_{i1})}{p_{i0}} + n \left(\sum_{i=1}^{k} \frac{p_{i1}^2}{p_{i0}} - 1 \right).$$

Если справедлива гипотеза H_0, то $p_{i1} = p_{i0}$ и $\mathbf{E}(\chi^2 | H_0) = k - 1$.

Если продифференцировать (1) по p_{i1} при условии: $\sum_{i=1}^{k} p_{i1} = 1$, то найдем, что (1) достигает минимума при $p_{i1} = p_{i0}$, откуда $\mathbf{E}(\chi^2 | H_1) > \mathbf{E}(\chi^2 | H_0) = k - 1$.

Далее, можно показать, что

$$\mathbf{D}(\chi^2 | H_0) = 2(k-1) + \frac{1}{n} \left(\sum_{i=1}^{k} \frac{1}{p_{i0}} - k^2 - 2k + 2 \right).$$

Минимум этого выражения достигается при $p_{i0} = \dfrac{1}{k}$, $i = 1, 2, \dots, k$. При этом

$$\mathbf{D}(\chi^2 \mid H_0) = 2(k-1) + \frac{1}{n}\left(k^2 - k^2 - 2k + 2\right) = 2(k-1)\left(1 - \frac{1}{n}\right).$$

В общем случае, если $\dfrac{k^2}{n} \underset{n\to\infty}{\to} 0$, то $\dfrac{1}{n}\sum_{i=1}^{k}\dfrac{1}{p_{i0}} \to 0$ и $\mathbf{D}(\chi^2 \mid H_0) \approx 2(k-1)$.

Например, если $k = C\sqrt[3]{n}$, то $\dfrac{k^2}{n} = \dfrac{C}{n^{1/3}} \to 0$.

Гипотеза H_0 должна отвергаться при больших значениях статистики χ^2. По этому вопросу в литературе не было полной уверенности, и более ранняя практика состоял в том, чтобы отвергать H_0 как при больших, так и при малых значениях χ^2. Эта практика основана на том, что причиной чрезвычайно малых значений χ^2 являются скорее всего численные ошибки в вычислениях; или в других случаях такие значения возникают из-за того, что частоты m_i / n смещены, возможно не намеренно, с тем, чтобы сблизить их с гипотетическими ожиданиями. Поэтому полезно определять не –только то, что гипотеза принимается или отвергается с данным уровнем значимости, но и указать для каждого χ^2 наименьший уровень значимости $\hat{\alpha} = \alpha(\chi^2)$, <u>критический уровень</u>, при котором гипотеза отвергается для данного результата наблюдения.

Необходимость группировки данных по классам приводит к потере некоторого количества информации, особенно если наблюдаемая величина непрерывна. Однако с этим недостатком связано и соответствующее достоинство: нам нет необходимости знать значения индивидуальных наблюдений.

В случае сложной гипотезы (т.е. когда гипотетическая функция распределения зависит от неизвестных параметров $\theta = (\theta_1, \theta_2, \dots, \theta_s)$) оказывается, что если применять оценки по методу минимума хи-квадрат, то асимптотическое распределение χ^2 будет иметь распределение хи-квадрат с $(k-s-1)$ степенями свободы, т.е. $\chi^2(k-s-1)$, $k -$ число интервалов, а $s -$ число оцениваемых параметров. Если для оценки неизвестных параметров используются оценки максимального правдоподобия, то распределение χ^2 заключено между $\chi^2(k-1)$ и $\chi^2(k-s-1)$, и при возрастании k они становятся столь близки, что различием можно пренебречь.

Дополнение. Покажем, что статистика $\chi^2 = \sum_{i=1}^{k}\dfrac{n}{p_i}\left(\dfrac{m_i}{n} - p_i\right)^2 = \sum_{i=1}^{k}\dfrac{(m_i - np_i)^2}{np_i} = \sum_{i=1}^{k}\dfrac{m_i^2}{np_i} - n$

при $n \to \infty$ и для случая проверяемой гипотезы сходится по распределению к хи-квадрат распределению с $(k-1)$ степенями свободы, где мы опустили индекс 0.

Известно, что вектор $M = (m_1, m_2, \dots, m_k)$ имеет полиномиальное распределение

$$p_{m_1, m_2, \dots, m_k} = \frac{n!}{m_1! m_2! \cdot \ldots \cdot m_k!} p_1^{m_1} p_2^{m_2} \cdot \ldots \cdot p_k^{m_k}, \text{ где } m_1 + m_2 + \dots + m_k = n,\ p_1 + p_2 + \dots + p_k = 1.$$

Характеристическая функция вектора M равна

$$\phi_M(t_1, t_2, \dots, t_k) = (p_1 e^{it_1} + p_2 e^{it_2} + \dots + p_k e^{it_k})^n.$$

Действительно (для простоты записи мы рассмотрим случай $k = 3$),

$$\mathbf{E}(e^{it_1 m_1 + it_2 m_2 + it_3 m_3}) = \sum_{m_1, m_2, m_3} e^{it_1 m_1 + it_2 m_2 + it_3 m_3} \cdot p_{m_1, m_2, m_3},\quad m_3 = n - m_1 - m_2,\ p_3 = 1 - p_1 - p_2.$$

Представим вероятность следующим образом:

$$p_{m_1,m_2,m_3} = C_n^{m_1} C_{n-m_1}^{m_2} p^{m_1} p^{m_2} (1-p_1-p_2)^{n-m_1-m_2}, \quad 0 \le m_2 \le n-m_1.$$

Тогда

$$\sum_{m_1,m_2,m_3} e^{it_1m_1+it_2m_2+it_3m_3} \cdot p_{m_1,m_2,m_3} = \sum_{m_1=0}^{n} e^{it_1m_1+t_3(n-m_1)} C_n^{m_1} p_1^{m_1} \sum_{m_2=0}^{n-m_1} e^{i(t_2-t_3)m_2} C_{n-m_1}^{m_2} p_2^{m_2} (1-p_1-p_2)^{n-m_1-m_2} =$$

$$= \sum_{m_1=0}^{n} e^{it_1m_1+it_3(n-m_1)} C_n^{m_1} p_1^{m_1} (e^{i(t_2-t_3)} p_2 + p_3)^{n-m_1} = (p_1e^{it_1} + p_2e^{it_2} + p_3e^{it_3})^n .$$

Положим $x_i = \dfrac{m_i - np_i}{\sqrt{np_i}} = \dfrac{m_i}{\sqrt{np_i}} - \sqrt{np_i}$, $i = 1, 2, \ldots, k$. Тогда $\sum_{i=1}^{k} x_i \sqrt{p_i} = 0$ и $\chi^2 = \sum_{i=1}^{k} x_i^2$.

Совместной характеристической функцией для $(x_1, x_2, \ldots, x_k)$ является

$$\varphi(t_1, t_2, \ldots, t_k) = e^{-i\sqrt{n}\sum_{i=1}^{k} t_i\sqrt{p_i}} \left(p_1 e^{\frac{it_1}{\sqrt{np_1}}} + \ldots + p_k e^{\frac{it_k}{\sqrt{np_k}}} \right)^n .$$

Тогда, разлагая в ряд, получаем

$$\ln \varphi(t_1, t_2, \ldots, t_k) = n \ln\left(1 + \frac{i}{\sqrt{n}} \sum_{i=1}^{k} t_i \sqrt{p_i} - \frac{1}{2n} \sum_{i=1}^{k} t_i^2 + O(n^{-3/2}) \right) -$$

$$- i\sqrt{n} \sum_{i=1}^{k} t_i \sqrt{p_i} = -\frac{1}{2} \sum_{i=1}^{k} t_i^2 + \frac{1}{2} \left(\sum_{i=1}^{k} t_i \sqrt{p_i} \right)^2 + O(n^{-1/2}),$$

т.е. $\lim\limits_{n\to\infty} \varphi(t_1, t_2, \ldots, t_k) = \exp\left(-\dfrac{1}{2} \left(\sum_{i=1}^{k} t_i^2 + \left(\sum_{i=1}^{k} t_i \sqrt{p_i} \right)^2 \right) \right)$.

Совершим ортогональное преобразование, полагая $u_k = \sum_{i=1}^{k} t_i \sqrt{p_i}$. При ортогональном преобразовании $\sum_{i=1}^{k} t_i^2 = \sum_{i=1}^{k} u_i^2$, поэтому $Q = \sum_{i=1}^{k} t_i^2 + \left(\sum_{i=1}^{k} t_i \sqrt{p_i} \right)^2 = \sum_{i=1}^{k-1} u_i^2$. Отсюда следует, что Q имеет ранг, равный $k-1$. Таком образом, $\lim\limits_{n\to\infty} \varphi = \exp\left(-\dfrac{1}{2} \left(\sum_{i=1}^{k-1} u_i^2 \right) \right)$ – характеристическая функция нормальных случайных величин $\xi_i, i = 1, \ldots, k-1$, каждая из которых имеет стандартное нормальное распределение $N(0,1)$. В таком случае, $\xi_1^2 + \xi_2^2 + \ldots + \xi_{k-1}^2 \in \chi_{k-1}^2$.

В таблице 3 приведено распределение (ошибочных) ответов на репертуар закрытого задания теста у 100 обследованных (закрытое задание – это тест с заранее данными ответами, а открытое, когда вы сами пишете ответ). Примером таких задач может служить тест Амхауэра интеллекта структуры теста:

1) У дерева всегда имеются: а) листья, б) плоды, в) почки, г) корни, д) тень.

2) Когда спор заканчивается взаимной уступкой, то это называют а) конвенцией, б) компромиссом, в) развязкой, г) сговором, д) столкновением.

Мы хотим выяснить, выбирают ли респонденты ответы случайно, или они выбирают их со знанием дела (неслучайно).

Таблица 3. Распределение ошибочных ответов теста.

	а	б	в	г	д
m_i	22	18	29	21	10
np_{i0}	20	20	20	20	20
$(m_i - np_{i0})^2$	4	4	81	1	100

Значение статистики χ^2 равно $\chi^2 = \dfrac{4+4+1+81+100}{20} = \dfrac{190}{20} = 9.5$. Гипотеза здесь простая (никакие параметры не оцениваются), поэтому число степеней свободы $\nu = 5-1 = 4$. Возьмем уровень значимости $\alpha = 0.01$. Тогда табличное значение равно $\chi^2_{0.99}(4) = 13.277$. Так как $9.5 < 13.277$, то это говорит о том, что гипотеза о случайности выбора ответов не отвергается.

Замечание. Рассмотрим следующую гипотез о параметрах $H_0 : p_j = p_{j0},\ j = 1,2,...,k,\ \sum_{j=1}^{k} p_j = 1.$

Покажем теперь, что критерий согласия χ^2 асимптотически при $n \to \infty$ является критерием отношения правдоподобия. Если через p_i обозначим вероятности попадания наблюдений в каждый из k интервалов, а наблюденные частоты через m_i / n, то m_i имеют полиномиальное распределение и функция правдоподобия равна $L(m_1, \dots, m_k; \pi_1, \dots, \pi_k) \sim \prod_{i=1}^{k} \pi_i^{m_i}$. При гипотезе эта функция правдоподобия равна $L(m_1, \dots, m_k; p_1, \dots, p_k) \sim \prod_{i=1}^{k} p_i^{m_i}$. Максимальное значение функция правдоподобия достигает тогда, когда вместо π_i подставим МП-оценки $\hat{\pi}_i = \dfrac{m_i}{n}$. В таком случае статистика отношения правдоподобия будет равна

$$\lambda = \frac{L(m_1, \dots, m_k; p_1, \dots, p_k)}{L(m_1, \dots, m_k; \hat{\pi}_1, \dots, \hat{\pi}_k)} = n^n \prod_{i=1}^{k} \left(\frac{p_i}{m_i} \right)^{m_i} = \prod_{i=1}^{k} \left(\frac{np_i}{m_i} \right)^{m_i}.$$

Гипотеза H_o отвергается, если значение λ достаточно мало.

Обозначим $\Delta_i = \dfrac{m_i - np_i}{np_i}$. Тогда

$$-2\ln\lambda = 2\sum_{i=1}^{k} m_i \ln(1+\Delta_i) = 2\sum_{i=1}^{k} \left\{ (m_i - np_i) + np_i \right\} \left\{ \Delta_i - \frac{1}{2}\Delta_i^2 + O\left(n^{-3/2}\right) \right\} =$$

$$= 2\sum_{i=1}^{k} \left\{ (m_i - np_i)\Delta_i + np_i\Delta_i - \frac{np_i}{2}\Delta_i^2 + O\left(n^{-1/2}\right) \right\}.$$

Поскольку $\sum_{i=1}^{k} np_i\Delta_i = 0$, то $-2\ln\lambda = \sum_{i=1}^{k} \left\{ np_i\Delta_i^2 + O\left(n^{-1/2}\right) \right\} = \chi^2 \left(1 + O\left(n^{-1/2} \right) \right)$, т.е. χ^2 и $2\sum_{i=1}^{k} m_i \ln \dfrac{m_i}{np_i}$ асимптотически при $n \to \infty$ и при гипотезе H_o распределены одинаково.

Последовательная проверка гипотез о распределении. Двухэтапный критерий χ^2.

Рассмотрим схему независимых испытаний с k исходами. Гипотеза H_0 состоит в том, что $p_j = p_{j0},\ j = 1,2,...,k,\ \sum_{j=1}^{k} p_j = 1.$ Пусть $n_1 < n_2 < ... < n_r$ – объемы последовательных, вло-

женных одна в другую выборок. Обозначим через ν_{ij} – число появлений j – го исхода в первых n_i испытаниях. Составим статистики Пирсона $X(n_i)=\sum_{j=1}^{k}\frac{(\nu_{ij}-n_i p_{j0})^2}{n_i p_{j0}}$. Последовательный r – кратный критерий χ^2 строится следующим образом. Выбираются критические значения $x_1^*, x_2^*, ..., x_r^*$; гипотеза H_0 принимается, если осуществляется одно из следующих событий $A_s, s=1,2,...,r$: $A_s=\{X(n_1)>x_1^*, X(n_2)>x_2^*,..., X(n_{s-1})>x_{s-1}^*, X(n_s)\le x_s^*\}$.
Если же $X(n_s)>x_s^*$ для всех $s=1,2,...,r$, то гипотеза H_0 отвергается.

Рассмотрим частный случай $r=2$. Здесь числа x_1^*, x_2^* найдем по уровню значимости α из равенства $\alpha=\iint\limits_{u_1>x_1^*/2,\,u_2>x_2^*/2} p_{2,0}(u_1,u_2)\,du_1du_2$, где

$$p_{2,0}(u_1,u_2)=\frac{\exp(-u_1/\beta-u_2/\beta)}{C_2^{1+\delta}\lambda^{\delta}\Gamma(1+\delta)\beta^2}\left(\frac{u_1u_2}{\beta^2}\right)I_{\delta}\left(\frac{2\lambda\sqrt{u_1u_2}}{\beta}\right), \lambda=\sqrt{n_1/n_2}, C_2=n_1/(n_2-n_1), \beta=1-\lambda^2, \delta=(k-3)/2.$$

$I_{\nu}(x)$ – функция Инфельда (модифицированная функция Бесселя первого рода).

9.4 Критерии согласия Колмогорова, Смирнова, Андерсона-Дарлинга

Рассмотрим статистику $\omega^2=\int\limits_{-\infty}^{\infty}\left(F_n(x)-F_0(x)\right)^2 dF_0(x)$,

которая была предложена и исследована **Смирновым Н.В.** после более ранних предложений Крамера и Мизеса. Здесь $F_n(x)$ – эмпирическая функция распределения, а $F_0(x)$ – функция распределения, определяемая гипотезой H_0, причем $F_0(x)$ непрерывна, а гипотеза H_0 – простая. При нахождении ω^2 удобна формула

$$\omega^2=\frac{1}{12n^2}+\frac{1}{n}\sum_{k=1}^{n}\left(F_0(x_n^{(k)})-\frac{2k-1}{2n}\right)^2,$$ где $x_n^{(k)}, k=1,2,\dots,n,$ – члены вариационного ряда, построенного по исходной выборке.

Если верна гипотеза H_0, то $\mathbf{E}(\omega^2)=\frac{1}{n}\int\limits_0^1 F_0(1-F_0)dF_0=\frac{1}{6n}$, так как

$\mathbf{E}\,(F_n(x)-F_0(x))^2=\frac{F_0(x)(1-F_0(x))}{n}$. Аналогично, $\mathbf{D}(\omega^2)=\mathbf{E}(\omega^4)-(\mathbf{E}\,(\omega^2))^2=\frac{4n-3}{180n^3}$.

Распределение ω^2 не зависит от $F_0(x)$, когда выполнена гипотеза H_0. Приведем предельные критические значения статистики ω^2 для уровня значимости $\alpha=0.1$, 0.05, 0.01, 0.005. Соответствующие $S_\alpha=0.347$, 0.461, 0.743, 1.168. Приведенные значения точны уже при $n>40$. Критическая область имеет вид:

$$n\omega^2>S_\alpha, \qquad (10)$$

т.е. гипотезу $H_0: F(x)=F_0(x)$ следует отвергнуть, если выполнено неравенство (10).

Для сложной гипотезы, если параметр допускает оценку с дисперсией большего порядка, чем n^{-1}, пользоваться критерием (2) нельзя.

Статистика Колмогорова А.Н. $\Delta_n=\sup\limits_x|F_n(x)-F_0(x)|$.

Распределение Δ_n не зависит от $F_0(x)$, когда верна гипотеза H_0. Кроме того,

$$\lim_{n\to\infty}\mathbf{P}(\Delta_n > z\,n^{-1/2}) = \sum_{k=-\infty}^{\infty}(-1)^k \exp\{-2k^2z^2\} \approx 1-2e^{-2z^2}+2e^{-8z^2}.$$

Предельные критические значения для $\sqrt{n}\,\Delta_n$ равны: (α и K_α

$\alpha = 0.05 \quad K_{0.05} = 1.3581, \qquad \alpha = 0.01 \quad K_{0.01} = 1.6276.$

Можно пользоваться также следующей аппроксимацией $\quad K_\alpha = \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}}$.

Простая гипотеза $H_0 : F(x) = F_0(x)$ для непрерывной $F_0(x)$ отвергается, если

$$\sqrt{n}\;\Delta_n > K_\alpha .$$

Для сложной гипотезы Δ_n будет зависеть от конкретного вида функции распределения, тем более распределение Δ_n будет зависеть от альтернативы.

Статистика Андерсона-Дарлинга.

$$A_n = n\int_{-\infty}^{\infty}\frac{(F_n(x)-F_0(x))^2}{F_0(x)(1-F_0(x))}dF_0(x) = -n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F_0(x_n^{(i)})+\ln(1-F_0(x_{n)}^{(n-i)}))] =$$

$$= -n-\frac{1}{n}\sum_{i=1}^{n}[(2i-1)\ln F_0(x_n^{(i)})+(2n-2i+1)\ln(1-F_0(x_n^{(i)}))].$$

Гипотеза $H_0 : F(x) = F_0(x)$ для непрерывной $F_0(x)$ отвергается, если $\quad A_n > A_\alpha$,

предельные значения A_α равны:

α	0.1	0.05	0.01	0.005	0.001
A_α	1.9355	2.4986	3.0916	4.5416	6.0266

Предельное распределение величины nA_n есть $\quad A = \sum_{j=1}^{\infty}\frac{Y_j}{j\,(j+1)}$, где н.о.р. с.в. $Y_j \in \chi_1^2$.

Пример 18. Выборка 0.3407, 0.1440, 0.6960, 0.8675, 0.5649, 0.5793, 0.1514, 0.5044, 0.9859, 0.4658 взята из таблиц случайных чисел. Проверить гипотезу $H_0 : F(x) = x,\, 0 < x < 1,$ о равномерном распределении используя 1) критерии Колмогорова, 2) Смирнова, 3) Андерсона-Дарлинга.

Решение. 1) $F(x) = x,\, 0 \le x \le 1,\, \sqrt{10} = 3.162277;\; d_{10,0.05} = 0.41(0.40925)$

Вариационный ряд: 0.1440, 0.1514, 0.3407, 0.4658, 0.5044, 0.5649, 0.5793, 0.6960, 0.8675, 0.9853

$\max|F_n(x)-F(x)|:\;$ 1) $\begin{cases} |0-0.144| = 0.144 \\ |0.1-0.144| = 0.044 \end{cases}$ 2) $\begin{cases} |0.1-0.1514| = 0.0514 \\ |0.2-0.1514| = 0.0486 \end{cases}$ 3) $\begin{cases} |0.2-0.3407| = 0.1407 \\ |0.3-0.3407| = 0.0407 \end{cases}$

4) $\begin{cases} |0.3-0.4658| = \mathbf{0.1658} \\ |0.4-0.4658| = 0.0658 \end{cases}$ 5) $\begin{cases} |0.4-0.5044| = 0.1044 \\ |0.5-0.5044| = 0.0044 \end{cases}$ 6) $\begin{cases} |0.5-0.5649| = 0.0649 \\ |0.6-0.5649| = 0.0351 \end{cases}$ 7) $\begin{cases} |0.6-0.5793| = 0.0207 \\ |0.7-0.5793| = 0.1207 \end{cases}$

8) $\begin{cases} |0.7-0.696| = 0.004 \\ |0.8-0.696| = 0.104 \end{cases}$ 9) $\begin{cases} |0.8-0.8675| = 0.0675 \\ |0.9-0.8675| = 0.0325 \end{cases}$ 10) $\begin{cases} |0.9-0.9859| = 0.0859 \\ |1.0-0.9859| = 0.0141 \end{cases}$

$\Delta = 0.1658 = D_n < 0.40925 -$ гипотеза H_0 **не отвергается**

$$\sqrt{10}D_{10} = 0.5243056 < \sqrt{n}\cdot 0.40925 = \mathbf{1.2941621}$$

Асимптотическое критическое значение, найденное по функции Колмогорова равно

$$K_{0.95} = \underline{1.3581.}$$

2) <u>С помощью критерия ω^2:</u>

$$n\omega^2 = \frac{1}{12n}+\sum_{k=1}^{n}\left(F_0(x_n^{(k)})-\frac{2k-1}{n}\right)^2 = \frac{1}{120}+(0.144-0.05)^2+(0.1514-0.15)^2+(0.3407-0.25)^2+(0.4658-0.35)^2+$$

$$+(0.5044-0.45)^2+(0.5649-0.55)^2+(0.5793-0.65)^2+(0.6975-0.75)^2+(0.8675-0.85)^2+(0.9859-0.95)^2 =$$

$$= 0.05114983.$$

$n\omega^2 = 0.514983 < \omega^2_{0.05} = 0.4614 -$ гипотеза о равномерном распределении **не отвергается.**

3) С помошью критерия AD:

nA:= – 10– 0.1[ln(0.144)+19ln(1-0.144)+3ln(0.1514)+17ln(1-0.1514)+5ln(0.3407)+
+15 ln(1-0.340)+7 ln(0.4658)+13 ln(1-0.4658)+9 ln(0.5044)+11 ln(1-0.5044)+11 ln(0.5649)+
+9 ln(1-0.5649)+13 ln(0.5793)+7 ln(1-0.5793)+15 ln(0.696)+
+5 ln(1-0.696)+17 ln(0.8675)+3 ln(1-0.8675)+19 ln(0.9853)+ln(1-0.9853)] = 0.36438.

$A = 0.36438 < 2.4986 = A_{0.05}$ – гипотеза о равномерном распределении тоже **не отвергается**.

Неравенство Коксма-Хлавка и статистика Колмогорова.

Начнем сначала со следующей формулы суммирования по частям.

Теорема А (преобразование Абеля). *Пусть даны две последовательности* $\{a_n\}$ *и* $\{b_n\}$. *Положим* $A_n = \sum_{k=0}^{n} a_k$ *при* $n \ge 0$; $A_{-1} = 0$. *Тогда, если* $0 \le p \le q$, *то*

$$\sum_{n=p}^{q} a_n b_n = \sum_{n=p}^{q-1} A_n (b_n - b_{n+1}) + A_q b_q - A_{p-1} b_p. \qquad (1)$$

Доказательство. Имеем

$$\sum_{n=p}^{q} a_n b_n = \sum_{n=p}^{q} (A_n - A_{n-1}) b_n = \sum_{n=p}^{q} A_n b_n - \sum_{n=p-1}^{q-1} A_n b_{n+1},$$

а последнее выражение справа равно правой части равенства (1).

Пусть $I = [0,1)$, $\overline{I} = [0,1]$ и $P -$ множество точек $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in \overline{I}^s$. Для произвольного подмножества $B = \prod_{i=1}^{s} [0, u_i) \in \overline{I}^s$ мы определим

$$A(B;P) = \sum_{n=1}^{N} \chi_B(\boldsymbol{x}_n) \text{ и дискрепанс } D_N^*(P) = \sup_B \left| \frac{A(B;P)}{N} - \prod_{i=1}^{s} u_i \right|.$$

Теорема В. *Если функция* f *имеет ограниченную вариацию (в смысле Харди и Краузе)* $V(f)$ *на* $\overline{I}^s$, *то для любой последовательности* $P = \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in \overline{I}^s$ *мы имеем*

$$\left| \frac{1}{n} \sum_{n=1}^{T} f(\boldsymbol{x}_n) - \int_{\overline{I}^s} f(\boldsymbol{u})\, d\boldsymbol{u} \right| \le V(f) D_N^*(P).$$

Доказательство. Рассмотрим одномерный случай $s = 1$ и для простоты рассуждений будем предполагать, что функция $f(u) -$ непрерывно дифференцируема. Кроме того, предположим, что $x_o = 0 < x_1 < x_2 < ... < x_n < 1 = x_{n+1}$. Используя формулу суммирования и интегрируя по частям, получим

$$\frac{1}{N} \sum_{n=1}^{N} f(x_n) - \int_0^1 f(u)\, du = -\sum_{n=0}^{N} \frac{n}{N} (f(x_n) - f(x_{n+1})) + \int_0^1 u\, df(u) = \sum_{n=0}^{N} \int_{x_n}^{x_{n+1}} \left(u - \frac{n}{N} \right) df(u).$$

Для $0 \le n \le N$, мы имеем $\left| u - \frac{n}{N} \right| \le D_N^*(P)$ для $x_n \le u \le x_{n+1}$.

Известно, что $V(f) = \int_0^1 |f'(u)| du$. Кроме того,

$$D_N^*(P) = \max_{0 \le n \le N} \sup_{x_n < u \le x_{n+1}} \left| \frac{A([0,u);P)}{N} - u \right| = \max_{0 \le n \le N} \sup_{x_n < u \le x_{n+1}} \left| \frac{n}{N} - u \right| = \max_{0 \le n \le N} \max \left(\left| \frac{n}{N} - x_n \right|, \left| \frac{n}{N} - x_{n+1} \right| \right),$$

а последнее есть статистика Колмогорова, если P смотреть как на вариационный ряд.

10. Проверка гипотез по многим малым выборкам

Пусть независимые случайные величины $x_1, x_2, ..., x_n$ имеют одно и то же распределение с плотностью $f(x)$. Тогда их совместное распределение равно $f(x_1)\cdot f(x_2)\cdot ...\cdot f(x_n)$.

Рассмотрим случайные величины

$$y_i = \frac{x_i - \bar{x}}{S}, \text{ где } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ r^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Величины y_i связаны соотношением

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} y_i^2 = 1 .$$

Это есть сфера, которую мы обозначим через Σ_0.

Введем в пространстве $\mathbf{R}^n$ новую систему координат: $\bar{x}, S, \theta_1, ..., \theta_{n-2}$, где $\theta_1, ..., \theta_{n-2}$ есть углы в полярной системе координат (т.е. система координат на сфере). В новой системе координат

$$f(x_1)\,f(x_2)...f(x_n)\,dx_1 dx_2 ... dx_n = \prod f(\bar{x} + Sy_k)\left|\frac{\partial(x_1,...,x_n)}{\partial(\bar{x}, S, \theta_1, ..., \theta_{n-2})}\right| d\bar{x}dSd\theta_1 ... d\theta_{n-2},$$

где $J(\bar{x}, S, \theta_1, ..., \theta_{n-2}) = \dfrac{\partial(x_1,...,x_n)}{\partial(\bar{x}, S, \theta_1, ..., \theta_{n-2})}$.

Перейдем теперь к переменным $z_1, z_2, ..., z_n$, при условии, что ось z_1 направлена по вектору $(1,1,...,1)$, т.е. совершим линейное преобразование $A = (a_{ij})_{n\times n}$, в котором первая строка равна $(1,1,...,1)$, а остальные ортогональны ей, тогда

$\sum_{j=1}^{n} a_{ij} = 0$ для $j = 2,3,..,n$. Потребуем, чтобы $z_1 = \sqrt{n}\cdot\bar{x}$ и пусть

$$\begin{cases} z_2 = S\cdot\cos\theta_1, \\ z_3 = S\cdot\sin\theta_1\cos\theta_2, \\ \quad . \ . \ . \ . \\ z_n = S\sin\theta_1 ... \sin\theta_{n-2}. \end{cases}$$

Хорошо известно, что якобиан последнего преобразования равен

$$J(S, \theta_1, ..., \theta_{n-2}) = S^{n-2} J(1, \theta_1, ..., \theta_{n-2}) = S^{n-2}(\sin\theta_1)^{n-3} ... \sin\theta_{n-2} .$$

Поскольку $\int J(1, \theta_1, ..., \theta_{n-2})\,d\theta_1 ... d\theta_{n-2}$ есть объем единичной сферы и мы имеем дело с плотностями, тогда если $\eta = (z_1, z_2, ..., z_n)$, то вероятность вектору η попасть на сферу Σ_0 равна

$$\mathbf{P}(\eta \in \Sigma_0) = \sqrt{n}\int_0^{\infty} dS \int_{-\infty}^{\infty} S^{n-2} \prod f(\bar{x} + Sy_k)\,d\bar{x} .$$

Пример 19. Пусть $f(x) = \varphi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Тогда

$$\prod f(\bar{x} + Sy_k) = \frac{1}{(\sqrt{2\pi})^n}\exp\left(-\frac{\sum(\bar{x} + Sy_k)^2}{2}\right) = \frac{1}{(\sqrt{2\pi})^n}\exp\left(-\frac{n\bar{x}^2 + S^2}{2}\right).$$

Поэтому

$$\mathbf{P}(\eta \in \Sigma_0) = \frac{\sqrt{n}}{(\sqrt{2\pi})^n} \int_0^{\infty} dS \int_{-\infty}^{\infty} S^{n-2} e^{-\frac{n\overline{x}^2+S^2}{2}} d\overline{x} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{2(\sqrt{\pi})^{n-1}},$$

т.е. в этом случае распределение на сфере Σ_0 является равномерным.

Таким образом, гипотезу о нормальности исходной выборки по многим малым выборкам можно проверять используя преобразование $y_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j}$, $j = 1, 2, ..., m$, и проверяя данные y_{ij} на согласие с равномерным распределением.

Применим полученный теоретический результат для проверки нормальности приведенных ниже выборочных данных на нормальность.

Пример 20. Ниже приведены данные о распределении вероятностей концентрации Na_2O (x_{ij}) в базальтах мира, где предполагается, что они имеют нормальное распределение при фиксированном j. Для этого вместо изучаемой величины x_{ij} (содержание Na_2O в %) вводятся величины $\tau_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$, где $\overline{x}_j$ – среднее по 4 анализам, s_j – несмещенная оценка стандарта для того же пункта, а затем случайно выбирается одно из них. Отобранные величины τ_{ij} распределены равномерно на отрезке $[-\sqrt{3}, \sqrt{3}]$, а затем рассматриваются модули этих величин имеющие равномерное на $[0, \sqrt{3}]$ распределение. Результаты $|\tau_{ij}|$, полученные для $n = 50$ пунктов приведены ниже (источник:[32], с.206):

1) 0.614, 0.7607, 1.2605, 1.308, 0.0343, 0.289, 0.973, 0.049, 1.201, 0.842, 0.925, 0.061, 1.262, 0.206, 0.528, 0.383, 1.06, 0.614, 1.23, 0.396, 1.439, 1.389, 1.052, 1.409, 0.562, 0.65, 0.676, 1.343, 0.409, 1.325, 0.993, 1.384, 0.325, 0.176, 0.214, 0.428, 0.575, 0.614, 0.538, 1.411, 0.09, 1.225, 1.25, 0.34, 0.266, 0.597, 1.23, 0.82, 0.993, 0.631.

2) 0.0198, 0.0281, 0.0352, 0.0531, 0.1016, 0.1188, 0.1238, 0.1539, 0.1668, 0.1875, 0.1966,0.2213, 0.2285, 0.2362, 0.2468, 0.3047, 0.3107, 0.3246, 0.3318, 0.3445, 0.3545, 0.3546, 0.3546, 0.3643,0.3752,0.3905, 0.4392, 0.4735, 0.4863, 0.534, 0.5618, 0.5732, 0.5734, 0.6072, 0.6123, 0.6932, 0.707, 0.71, 0.71, 0.7217, 0.7277, 0.7286, 0.755, 0.7651, 0.7752, 0.799, 0.8017, 0.8134, 0.8115, 0.831

По критерию хи-квадрат при числе интервалов $k = 4$ критическое значение $\chi^2_{0.95}(3) = 7.815$:

Гипотеза $H_0^{(1)} : F_0(x) = x, 0 < x < 1$. Значение статистики $\chi^2 = 2.32$.

Гипотеза $H_0^{(2)} : F_0(x) = x / 0.9, 0 < x < 0.9$. Значение статистики $\chi^2 = 1.71$.

Обе гипотезы <u>не отвергаются</u>, так как их значение не превосходит 7.815.

Если использовать критерий Колмогорова, где $\Delta = \sup\limits_{0<x<1} |F_n(x) - F_0(x)|$, $D = \sqrt{50} \cdot \Delta$, то

для гипотезы $H_0^{(1)}$ имеем $\Delta = 0.1653, D = 1.169$, а для гипотезы $H_0^{(2)}$ имеем $\Delta = 0.086, D = 0.6087$. При уровне значимости $\alpha = 0.05$ критическое значение $C_{0.05} = 1.3581$, поэтому каждая из гипотез не отвергается, а, значит, гипотеза о нормальном распределении исходных данных также не отвергается.

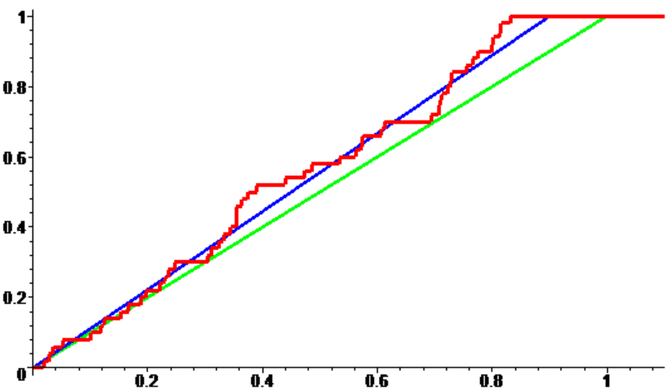


Рис. 5. Графики эмпирической и теоретической функций распределенитя.

На представленном графике зеленая линия соответствует гипотезе $H_0^{(1)}$, синяя – $H_0^{(2)}$, а красная линия – это эмпирическая функция распределения, построенная по преобразованным данным 2).

11. Вероятностные модели роста Распределение Бирнбаума-Сондерса

Предположим, что мы имеем металлическую пластину, на которой зародилась трещина и в рассматриваемый начальный момент она имеет длину x_0 и в последующие моменты времени $\Delta n, n = 1, 2, ...$ мы производим замеры $x_1, x_2, ... -$ длину растущей односторонней трещины (заметим, что если одной стороной трещина дошла до отверстия, то она будет односторонней). Нас интересует прогноз длины трещины по истечении n периодов. Естественно, на прирост длины трещины действует большое количество факторов: температура, структура металла и т.д., поэтому мы имеем дело со стохастическим прогнозом, который можно выразить примерно так: «Через $n = 50$ циклических нагрузок (а мы пока будем иметь дело с ними) с вероятностью 0.95 длина трещины будет не менее L_0 метров». Естественно речь пойдет не об одной трещине, а о массиве.

Итак, мы будет считать, что $x_1, x_2, ...$ будем считать значениями случайных величин $X_1, \ldots, X_n, ..$ и рассмотрим соседние приращения $\Delta_k = X_k - X_{k-1}$ длины трещины. Предположим, что прирост Δ_k обусловлен действием тех причин, о которых говорилось, т.е. под действием некоторого импульса ξ_k. Между Δ_k и ξ_k существует приближенная линейная связь $\Delta_k = \alpha_k \xi_k$, где α_k зависит от длины трещины X_{k-1} в момент k. Положим $\alpha_k = g(X_{k-1})$ с естественным условием неотрицательности и непрерывности функции $g(x)$. Таким образом, мы приходим к рекуррентным соотношениям, которые описывают ежегодный прирост длины,

$$X_k - X_{k-1} = \xi_k g(X_{k-1}), \; k = 1, 2, ... \qquad (1)$$

Сделаем некоторые предположения относительно распределения случайных величин $\xi_k, k = 1, 2, ...$ Будем считать, что эти случайные величины неотрицательны, независимы, одинаково распределены и обладают конечными моментами второго порядка: средним значением $a = \mathbf{E}(\xi_k)$ и дисперсией $b^2 = \mathbf{D}(\xi_k)$.

Перепишем первые n рекуррентных соотношений (1) в виде

$$\xi_k = \frac{X_k - X_{k-1}}{g(X_{k-1})}, \quad k = 1, ..., n$$

и просуммируем левые и правые части этих равенств. В результате получим

$$\sum_{k=1}^{n}\xi_k=\sum_{k=1}^{n}\frac{X_k-X_{k-1}}{g(X_{k-1})}.$$

Если каждый импульс вызывает незначительный прирост длины, то есть все $\Delta_k=X_k-X_{k-1}$ малы, то, трактуя правую часть последнего равенства как интегральную сумму, получаем приближенное равенство

$$\sum_{k=1}^{n}\xi_k=\int_{x_0}^{X}\frac{dt}{g(t)},$$

где $X=X_n-$ окончательная длина трещины.

Так как функция $g(x)$ положительна, то интеграл в правой части последнего равенства представляет собой некоторую монотонную возрастающую функцию $h(X)$. Применение центральной предельной теоремы в левой части (2) приводит к утверждению: по истечении достаточно большого срока после начала развития трещины распределение её длины X определяется соотношением $h(X)\in N(\mu,\sigma^2),$ где $\mu=na,\sigma^2=nb^2$. В силу монотонности функции $h(x)$

$$F(x)=\mathbf{P}(X<x)=\mathbf{P}(h(X)<h(x))=\Phi\left(\frac{h(x)-\mu}{\sigma}\right).$$

Осталось решить проблему с выбором функции $g(x)$. Если постулировать, что прирост её длины пропорционален уже достигнутой длине, т.е. положить $g(t)=t,$ а именно такое предположение наиболее часто используется в моделях роста, то мы придем к следующему распределению случайной величины X.

Логарифмически-нормальное распределение $LN(\mu,\sigma^2)$.

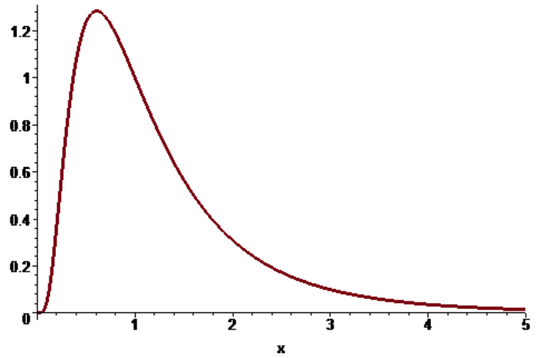


Рис. 6. Плотность логарифмически-нормального распределения.

При $g(t)=t$ интеграл в правой части (2)(с точность до постоянной равен $\ln X,$ так что $\ln X\in N(\mu,\sigma^2)$ и функция распределения X равна

$$F(x)=\Phi\left(\frac{\ln x-\mu}{\sigma}\right),\ \ x>0,$$

с функцией плотности

$$f(x)=\frac{1}{\sigma x\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right).$$

В практических приложениях целесообразнее оперировать не со случайной величиной X, а с величиной $Y=\ln X$, после чего производить расчеты, используя нормальное распределение.

Логарифмически нормальное распределение носит достаточно универсальный характер. Этому распределению подчиняется например высота дерева после n лет роста, размер трещины в испытуемом образце материала, который подвергается циклическим нагружениям (скорее следующее распределение лучше описывает модели разрушения при циклических нагрузках), роста дохода у отдельных лиц достаточно однородной человеческой популяции. Проводимые в этом направлении статистические исследования указывают на хорошее согласие с логарифмически-нормальным распределением достаточно низких доходов, в то время как при умеренных и высоких доходах более подходящим является распределение Парето.

В рамках построенной модели роста часто возникает задача определения момента достижения некоторого уровня. Это распределение можно получить путем следующих рассуждений.

Распределение Бирнбаума-Сондерса BS(λ,θ). Пусть $\tau-$ случайная величина, момент достижения заданного размера x. Тогда событие $(\tau > x)$ происходит тогда и только тогда, когда $(X_n < x) -$ к моменту времени n длина трещины еще не достигла уровня x. Поэтому

$$\mathbf{P}(\tau > n) = \mathbf{P}(X_n < x) = \mathbf{P}(h(X_n) < h(x)) = \Phi\left(\frac{h(x) - na}{b\sqrt{n}}\right).$$

Заменим теперь n на «непрерывную» переменную t и введем новые параметры λ и θ, определив их уравнениями $\lambda\sqrt{\theta} = h(x)/b,\ \lambda/\sqrt{\theta} = a/b.$ Это позволяет нам записать распределение случайного момента времени τ как

$$F(t) = \mathbf{P}(\tau < t) = 1 - \Phi\left(\lambda\left(\sqrt{\frac{\theta}{t}} - \sqrt{\frac{t}{\theta}}\right)\right),\ t > 0.$$

Это унимодальное распределение, которое называется распределением Бирнбаума-Сондерса BS (λ,θ) , плотность распределения которого похожа на плотность гамма-распределения и плотность логарифмически нормального распределения правда правая часть у BS-распределения чуть выше.

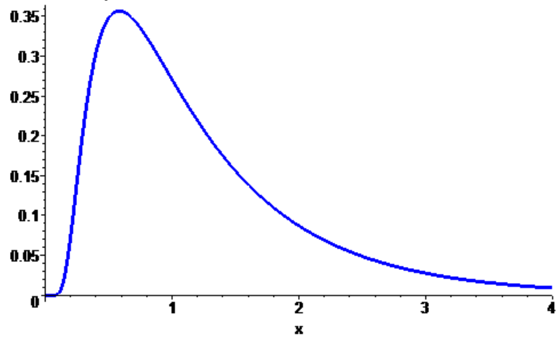


Рис. 7. Плотность распределения Бирнбаума-Сондерса (BS) с точностью до множителей.

Заметим, что случайная величина T имеет BS-распределение, если величина

$$\frac{1}{\alpha}\left(\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right)$$

имеет нормальное распределение, т.е. функция распределения равна $\Phi\left(\frac{1}{\alpha}\left(\sqrt{\frac{x}{\beta}} - \sqrt{\frac{\beta}{x}}\right)\right)$.

Для этого распределения математическое ожидание равно

$$\mu_1 = \boldsymbol{E}(X) = \beta\left(\frac{1}{2}\alpha^2 + 1\right), \tag{11}$$

дисперсия равна

$$\mu_2 = \boldsymbol{D}(X) = \alpha^2\beta^2\left(\frac{5}{4}\alpha^2 + 1\right),$$

а третий центральный момент есть

$\mu_3 = \boldsymbol{E}((X-\mu_1)^3) = \dfrac{\alpha^4\beta^3}{2}(11\alpha^2+6)$, откуда асимметрия равна $\gamma_1 = \dfrac{16\alpha^2(11\alpha^2+6)}{(5\alpha^2+4)^{3/2}}$.

Эсцесс равен $\gamma_2 = 3 + \dfrac{6\alpha^2(93\alpha^2+41)}{(5\alpha^2+4)^2}$.

Рассмотрим еще два коэффициента, связанные с распределением Бирнбаума-Сондерса

$\delta_2 = \dfrac{\mu_2}{\mu_1^2} = \dfrac{5\alpha^4+4\alpha^2}{(\alpha^2+2)^2}$ и $\delta_3 = \dfrac{\mu_3}{\mu_1\mu_2} = \dfrac{44\alpha^4+24\alpha^2}{(5\alpha^2+4)(\alpha^2+2)}$.

На основании этих показателей можно построить тест Tsukatani-Shigemitsu: если α^2, найденные с использованием последних соотношений будут близки, то это будет говорить в пользу BS-распределения.

Пример 20. Следующие данные представляют собой сроки выздоровления $N=69$ пациентов, получавших лечение от рака молочной железы в одной больнице. Время условное, данные ($x_i, i=1,2,...,69$) упорядочены.

0.3	5.0	5.6	6.2	6.3	6.6	6.8	7.5	8.4	8.4	8.7	9.0	9.8					
10.9	11.2	11.7	11.8	12.2	12.3	12.5	12.7	13.2	13.7	13.9	14.8	15.2					
16.2	17.3	17.5	17.9	19.7	20.1	20.9	21.0	21.0	21.1	23.0	23.0	23.0					
23.0	23.6	24.0	27.9	28.2	29.1	30.0	31.0	31.0	31.0	32.0	35.0	35.0					
38	40	41	41	42	44	48	51	51	52	54	56	60	74	80	89	126	154

Для этих данных оценки центральных моментов равны:

$m_1 = \bar{x} = 29.22,\quad m_2 = 727.2,\quad m_3 = 5231567.5,\quad \hat{\beta} = 20.8326,\quad \hat{\alpha} = 0.89724.$

Используя выборочный аналог соотношений (2) и (3), находим $\hat{\alpha}_1^2 = 0.8372,\ \hat{\alpha}_2^2 = 0.8242$.

Так как они близки, то это говорит в пользу BS-распределения, а потом подберем оценки для α и β минимизируя расстояние Колмогорова. Мы возьмем $\beta = 20.833,\quad \alpha = 0.897$.

Далее, преобразуем исходные данные считая, что оценки параметров – это истинные значения, по фомуле $z_i = \Phi\left(\dfrac{1}{\hat{\alpha}}\left(\sqrt{\dfrac{x_i}{\hat{\beta}}} - \sqrt{\dfrac{\hat{\beta}}{x_i}}\right)\right)$, $i=1,2,...,69$. Отобразим теперь на графике эмпирическую функцию распределения для данных z_i и прямую $y=x,\ 0 \le x \le 1$.

Значение расстояния Колмогорова равно 0.0429, а значение $\sqrt{69}\cdot 0.0429 = 0.356$. Это достаточно маленькое расхождение (критическое асимптотическое значение статистики Колмогорова $K_{0.05} = 1.358$). Если бы подсчитанное нами значение было бы больше критического значения, то гипотезу о BS-распределении мы бы отвергли, так как при истинном значении параметров, её значение было бы не меньше минимального!

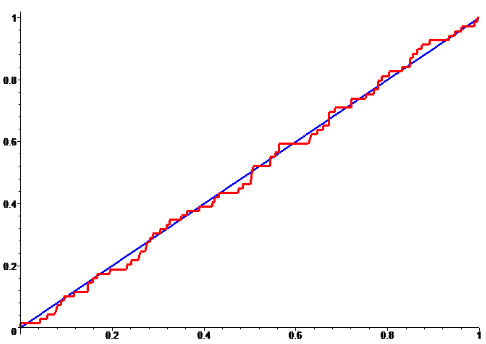


Рис.8. Графики эмпирической функции распределения и прямой.

12. Процесс броуновского движения. Распределение Вальда (IG-распределение)

Процесс броуновского движения.

Определение 1. *Случайный процесс $W_t, t \ge 0$ называется винеровским процессом (процессом броуновского движения), если он обладает следующими свойствами:*

1) *п.н. выходит из нуля, т.е.* $\mathbf{P}(W_0 = 0) = 1$*;*
2) *траектории процесса есть непрерывные функции;*
3) *случайный вектор* $(W_{t_1}, W_{t_2}, ..., W_{t_n}), 0 < t_1 < t_2 < ... < t_n < \infty$ *имеет гауссовское распределение, причем* $\mathbf{E}(W_t) = 0, \mathbf{E}(W_t W_s) = \min(t, s).$

Замечание. *Винеровский процесс является* **гауссовским** *процессом.*

Теорема 17. *Винеровский процесс является процессом с независимыми приращениями, т.е. случайные величины* $W_{t_1}, W_{t_2} - W_{t_1}, ..., W_{t_n} - W_{t_{n-1}}, 0 < t_1 < t_2 < ... < t_n < \infty$ *независимы. Кроме того,*

$$\mathbf{D}(W_t - W_s) = t - s, s \le t.$$

Доказательство. Пусть $W_t -$ винеровский процесс, $s \le t.$ Тогда

$$\mathbf{D}(W_t - W_s)^2 = \mathbf{E}(W_t - W_s)^2 = \mathbf{E}(W_t^2 - 2W_t W_s + W_s^2) = t - 2s + s = t - s.$$

Если $t_i < t_{i+1} \le t_j < t_{j+1},$ то

$$\mathbf{E}(W_{t_{i+1}} - W_{t_i})(W_{t_{j+1}} - W_{t_j})) = \mathbf{E}(W_{t_{i+1}} W_{t_{j+1}} - W_{t_{i+1}} W_{t_j} - W_{t_i} W_{t_{j+1}} W_{t_i} W_{t_j}) = t_{i+1} - t_{i+1} - t_i + t_i = 0.$$

Получаем, что приращения некоррелированы. В силу гауссовости конечномерных распределения приращения независимы.

Теорема 18. *Пусть* $s \equiv t_{0,n} < t_{1,n} < ... < t_{n-1,n} < t_{n,n} \equiv t -$ *последовательность разбиений отрезка* $[0,t]$ *с* $\lim_n \max_i |t_{i,n} - t_{i-1,n}| = 0.$ *Тогда случайная величина*

$$\xi_n = \sum_{i=0}^{n} (W_{t_{i+1},n} - W_{t_i,n})^2$$

сходится к $t - s$ *по вероятности.*

Доказательство. Имеем:

$$\mathbf{E}(\xi_n) = \mathbf{E}\left(\sum_{i=0}^{n} (W_{t_{i+1},n} - W_{t_i,n})^2 \right) = \sum_{i=0}^{n} (t_{i+1,n} - t_{i,n}) = t - s .$$

В силу независимости приращений

$$\mathbf{D}\left(\sum_{i=0}^{n} (W_{t_{i+1},n} - W_{t_i,n})^2 \right) = \sum_{i=0}^{n} \mathbf{D}\left((W_{t_{i+1},n} - W_{t_i,n})^2 \right) = \sum_{i=0}^{n} (\mathbf{E}\left((W_{t_{i+1},n} - W_{t_i,n})^4 \right) - \mathbf{E}^2 \left((W_{t_{i+1},n} - W_{t_i,n})^2 \right)).$$

Воспользовавшись тем, что $\mathbf{E}(\eta^4) = 3\sigma^4$ для $\eta \in N(0, \sigma^2)$, получаем

$$\mathbf{D}\left(\sum_{i=0}^{n} (W_{t_{i+1},n} - W_{t_i,n})^2 \right) = 2\sum_{i=0}^{n} (t_{i+1,n} - t_{i,n})^2 \le \max_i |t_{i+1,n} - t_{i,n}| \cdot (t - s) \underset{n \to \infty}{\to} 0.$$

Из сходимости в ср.кв. следует сходимость по вероятности.

Мартингал

Определение 2. *Случайный процесс в дискретном времени – это последовательнось случайных величин* $X_1, \ldots, X_n, ...$

Определение 3. ***Фильтрация*** *(поток $\sigma -$ алгебр) – это последовательнось $\sigma -$ алгебр* $\mathcal{F}_1, \ldots, \mathcal{F}_n, ...$*, такая, что* $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.

Определение 4. *Фильтрация $\mathcal{F}_n$ называется* ***естественной*** *фильтрацией для процесса* $\{X_n\}$*, если* $\mathcal{F}_n = \sigma(X_1, ..., X_n)$.

Определение 5. *Процесс* $\{X_n\}$, *для которого* $\mathbf{E}(|X_n|)<\infty, n\geq 1,$ *называется мартингалом по отношению к фильтрации* $\mathscr{F}_n$, *если выполнены 2 условия:*

1) *для любого* n *случайная величина* X_n *измерима относительно* $\mathscr{F}_n$ *;*
2) $\mathbf{E}(X_m\,|\,\mathscr{F}_n)=X_n,\ n\geq 1.$

Из последнего следует, что $\mathbf{E}(X_n)=const$. Из свойств условных математических ожиданий следует, что достаточно проверить выполнимость следующего равенства:

$$\mathbf{E}(X_{n+1}\,|\,\mathscr{F}_n)=X_n.$$

Под $\mathbf{E}(X_{n+1}\,|\,\mathscr{F}_n)=X_n$ понимается $\mathscr{F}$-измеримая с.в., такая что

$$\int_B \mathbf{E}(X_{n+1}\,|\,\mathscr{F}_n)(\omega)\,d\mathbf{P}(\omega)=\int_B X_n(\omega)\,d\mathbf{P}(\omega),\quad \forall B\in\mathscr{F}_n\ .$$

Примеры.

Пример 21. Пусть $Z_1,Z_2,Z_3...-$последовательность независимых случайных величин с нулевым математическим ожиданием. Определим $X_n=\sum_{i=1}^n Z_i$. Тогда $X_m=\sum_{i=1}^n Z_i+\sum_{i=n+1}^m Z_i,\ m\geq n+1, \mathscr{F}_n-$ естественная фильтрация. Тогда

$$\mathbf{E}(X_m\,|\,\mathscr{F}_n)=\mathbf{E}\left(\sum_{i=1}^n Z_i\,|\,\mathscr{F}_n\right)+\mathbf{E}\left(\sum_{i=n+1}^m Z_i\,|\,\mathscr{F}_n\right)=\sum_{i=1}^n Z_i+\mathbf{E}\left(\sum_{i=n+1}^m Z_i\right)=X_n,$$ т.е. мартингал.

Если $\mathbf{E}(Z_i)=a\neq 0$, то $\{X_n\}$ не является мартингалом (показать самим).

Пример 22. Вы идете на экзамен, но выучили не все билеты, а M билетов из $N\,(M<N)$. Пусть учеников тоже N. Предположим, что прошло n учеников и x_n-доля оставшихся билетов, которые вы знаете. Будем считать, что экзамен вы сдадите, если выберете билет, который знаете.

Ясно, $x_0=\dfrac{M}{N}$. Для x_{N-1}, т.е. когда вы пойдете последним, $x_{N-1}=1$, если останется билет, который вы знаете, либо $x_{N-1}=0$, если останется билет, который вы не знаете. Тогда $\mathbf{P}(x_{N-1}=1)=\dfrac{M}{N}-$известная задача из теории вероятностей.

Найдем условное математическое ожидание $\mathbf{E}(X_{n+1}\,|\,\mathscr{F}_n)=X_n$, где $\mathscr{F}_n-$ естественная фильтрация. Итак, пусть в момент n доля «хороших» билетов равна x_n. Число оставшихся «хороших» билетов равно $(N-n)x_n$. Если на $(n+1)-$м шаге будет выбран билет, который вы знаете, то число «хороших» билетов станет равно $(N-n)x_n-1$, а число оставшихся билетов тоже уменьшится на единицу и станет равно $N-1-n$, поэтому доля станет равна $x_{n+1}=\dfrac{(N-n)x_n-1}{N-1-n}$. В противном случае эта доля станет равна $x_{n+1}=\dfrac{(N-n)x_n}{N-1-n}$. Так как вероятность первого исхода равна x_n, а второго $(1-x_n)$, то условное математическое ожидание будет равно

$$\mathbf{E}(X_{n+1}\,|\,\mathscr{F}_n)=x_n\frac{(N-n)x_n-1}{N-1-n}+(1-x_n)\frac{(N-n)x_n}{N-1-n}=\frac{N-n}{N-1-n}(x_n^2+x_n-x_n^2)-\frac{x_n}{N-1-n}=$$

$$=x_n\left(\frac{N-n}{N-1-n}-\frac{1}{N-1-n}\right)=x_n,$$ т.е. $\{x_n\}-$ мартингал.

Пример 23. Пусть $Z_1, Z_2, ..., Z_n, ... -$ последовательность независимых и одинаково распределенных случайных величин с плотностью распределения $f_0(x) > 0, x \in \mathbf{R}$ и $\mathscr{F}_n -$ естественная фильтрация. Пусть кроме того $f_1(x) -$ другая плотность. Определим последовательность

$$X_n = \frac{f_1(Z_1)}{f_0(Z_1)} \cdot \frac{f_1(Z_2)}{f_0(Z_2)} \cdot ... \cdot \frac{f_1(Z_n)}{f_0(Z_n)}, \ n = 1, 2, ... \ .$$

Рассмотрим

$$\mathbf{E}(X_{n+1} | \mathscr{F}_n) = \mathbf{E}\left(\frac{f_1(Z_1)}{f_0(Z_1)} \cdot \frac{f_1(Z_2)}{f_0(Z_2)} \cdot ... \cdot \frac{f_1(Z_n)}{f_0(Z_n)} \cdot \frac{f_1(Z_{n+1})}{f_0(Z_{n+1})} \middle| \mathscr{F}_n \right) = X_n \mathbf{E}\left(\frac{f_1(Z_{n+1})}{f_0(Z_{n+1})} \middle| \mathscr{F}_n \right) = X_n \int_{-\infty}^{\infty} f_1(x)\, dx = X_n,$$

т.е. $\{X_n\} -$ мартингал.

Пример 24. Рассмотрим броуновское движение $\{B_t\}_{t \ge 0}$ и $\mathscr{F}_t -$ естественная фильтрация. Пусть $\lambda > 0$ и

$$M_t = e^{\lambda B_t - \lambda^2 t/2}, \ t \ge 0.$$

Покажем, что процесс $M = \{M_t\}_{t \ge 0}$ является мартингалом с $\mathbf{E}(M_t) = 1$. Для $t > s$ имеем:

$$\mathbf{E}(M_t | \mathscr{F}_s) = \mathbf{E}(M_t | B_s) = \mathbf{E}(e^{\lambda(B_t - B_s) + \lambda B_s - \lambda^2(t-s)/2 - \lambda^2 s/2} | B_s) = e^{\lambda B_s - \lambda^2 s/2} \cdot \mathbf{E}(e^{\lambda(B_t - B_s) - \lambda^2(t-s)/2} | B_s).$$

Заметим, что если $X \in N(0, (t-s))$, то $\mathbf{E}(e^{\lambda X}) = e^{\lambda^2(t-s)/2}$, поэтому

$$\mathbf{E}(M_t | \mathscr{F}_s) = e^{\lambda B_s - \lambda^2 s/2} = M_s,$$

т.е. мартингал. К тому же $\mathbf{E}(M_t) = \mathbf{E}(M_0) = 1.$

Распределение Вальда.

Определим случайную величину τ_a следующим образом:

$$\tau_a = \inf\{t \ge 0 : B_t \ge a\}.$$

Можно показать, что $M_{t \wedge \tau_a}$ является мартингалом, поэтому $\mathbf{E}(M_{t \wedge \tau_a}) = 1.$

Далее, $0 \le M_{t \wedge \tau_a} = e^{\lambda B_{t \wedge \tau_a} - \lambda^2 (t \wedge \tau_a)/2} \le e^{\lambda B_{t \wedge \tau_a}} \le e^{\lambda B_{\tau_a}} = e^{\lambda a}, \ t \wedge \tau_a = \min(t, \tau_a).$

поэтому по теореме о мажорируемой сходимости в интеграле Лебега

$$1 = \lim_{t \to \infty} \mathbf{E}(M_{t \wedge \tau_a}) = \lim_{t \to \infty} \left(\mathbf{E}(e^{\lambda B_{\tau_a} - \lambda^2 \tau_a/2} I(\tau_a \le t) + \mathbf{E}(e^{\lambda B_t - \lambda^2 t/2} I(\tau_a > t)) \right) =$$

$$= \mathbf{E}(e^{\lambda B_{\tau_a} - \lambda^2 \tau_a/2} I(\tau_a \le t) + \lim_{t \to \infty} \mathbf{E}(e^{\lambda B_t - \lambda^2 t/2} I(\tau_a > t)).$$

По усиленному закону больших чисел

$$e^{\lambda B_t - \lambda^2 t/2} = e^{t(\lambda B_t / t - \lambda^2/2)} \to 0, \quad t \to \infty.$$

Применяя теорему о мажорируемой сходимости устремляя λ к нулю, получаем, что $\mathbf{E}(I(\tau_a < \infty)) = 1 \Rightarrow \mathbf{P}(\tau_a < \infty) = 1.$ Поэтому заключаем, что

$$\mathbf{E}(e^{-\lambda^2 \tau_a/2}) = e^{-\lambda a}.$$

Иначе говоря, преобразование Лапласа есть

$$\mathbf{E}(e^{-t\tau_a}) = e^{-a\sqrt{2t}},$$

поскольку обозначая $\lambda^2 / 2 = t$, получаем $\lambda = \sqrt{2t}$.

Из теории преобразования Лапласа известно, что

$$\int_0^{\infty} e^{-\lambda t} \frac{a\varphi_t(a)}{t} dt = e^{-a\sqrt{2\lambda}}, \text{ где } \varphi_t(a) = \frac{1}{\sqrt{2\pi t}} e^{-a^2/(2t)}.$$

Тем самым плотность распределения $f_{\tau_a}(t)$ случайной величины $\tau_ф$ задается формулой

$$f_{\tau_a}(t)=\frac{a\varphi_t(a)}{t}=\frac{a}{\sqrt{2\pi t^3}}e^{-a^2/(2t)}.$$

Это есть плотность распределения Вальда (другое название – отрицательное гауссово распределение).

Задача. Пусть есть две конкурирующие железнодорожные компании A и B, имеющие по одному поезду, курсирующему между Чикаго и Лос-Анжелесом. Эти два поезда отправляются и прибывают одновременно и оборудованы примерно одинаково. Можно представить, что каждый пассажир независимо от другого бросает монетку, если выпадает герб, он садится в поезд A, если решка – в поезд B.

Очевидно, что мы имеем дело с испытаниями Бернулли, в которых вероятность успеха $p=\frac{1}{2}$. Всего имеется n испытаний. Если в поезде имеется $s<n$ мест, то существует положительная вероятность $p(s)$ того, что явится больше s пассажиров и мест не хватит. Найдем приближенное значение этой вероятности.

Пусть есть n пассажиров и $\xi_i,\ i=1,2,...,n,$ означает решение *i*-го пассажира сесть в первый $(\xi_i=1)$ или во второй $(\xi_i=0)$, причем ξ_i принимает эти значения с вероятностями 1/2. Тогда $S_n=\sum_{i=1}^{n}\xi_i$ есть количество пассажиров, выбравших первый поезд. Случайная величина S_n имеет биномиальное распределение с математическим ожиданием $\frac{n}{2}$ и дисперсией $\frac{n}{4}$ и по теореме Муавра-Лапласа

$$p(s)=\boldsymbol{P}(S_n>s)=\boldsymbol{P}\left(\frac{S_n-n/2}{\sqrt{n/4}}>\frac{s-n/2}{\sqrt{n/4}}\right)=1-\boldsymbol{P}\left(\frac{S_n-n/2}{\sqrt{n/4}}\le\frac{s-n/2}{\sqrt{n/4}}\right)\approx$$

$$\approx1-\Phi\left(\frac{s-n/2}{\sqrt{n/4}}\right)=1-\Phi\left(\frac{2s-n}{\sqrt{n}}\right)=1-\Phi\left(\frac{2s}{\sqrt{n}}-\sqrt{n}\right),$$

где $\Phi(x)-$ функция распределения стандартного нормального закона.

Выберем s так, чтобы вероятность $p(s)\le0.01$ например. Тогда $\frac{2s-n}{\sqrt{n}}=z_{0.99}=2.326$, поэтому при $n=1000$ получаем $s=537$.

<u>Прямой вывод распределения Вальда.</u> Зафиксируем сумму $S_n=m$, тогда число наблюдений v будет случайно, и мы имеем отрицательное биномиальное распределение:

$$\boldsymbol{P}(v=n)=Q_{m,p}(n)=C_{n-1}^{m-1}p^mq^{n-m},\quad n=m,m+1,...,\ q=1-p.$$

Заметим, что $Q_{m,p}(n)=\frac{m}{n}C_n^m p^m q^{n-m}\equiv\frac{m}{n}\cdot P_{n,p}(m).$

Кроме того, х.ф. отрицательного биномиального распределения равна

$$\varphi_m(t)=\left(\frac{pe^{it}}{1-qe^{it}}\right)^m=p^m(e^{-it}-q)^{-m}.$$

По $\varphi_m(t)\equiv\varphi(t)$ найдем первый μ_1 и второй μ_2 начальные моменты. Имеем:

$$\varphi'(t)=mip^m(e^{-it}-q)^{-m-1}\ \Rightarrow\ \mu_1=\frac{\varphi'(0)}{i}=\frac{m}{p}.$$

Аналогично, $\mu_2=\frac{\varphi''(0)}{i^2}=\frac{m(m+q)}{p^2}\ \Rightarrow\ \mu_2-\mu_1^2=\frac{m^2}{p^2}+\frac{mq}{p^2}-\frac{m^2}{p^2}=\frac{mq}{p^2}.$

Пусть $m\to\infty$, тогда $n\to\infty$ и пусть $y=\frac{np}{m},\ c=\frac{m}{q},\ \boldsymbol{E}(v)=\frac{m}{p}$. Тогда по локальной теореме Муавра-Лапласа $C_n^m p^m q^{n-m} \approx \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{npq}}\cdot e^{-\frac{(m-np)^2}{2npq}} \Rightarrow$

$$\frac{(m-np)^2}{2npq} = \frac{m^2}{2npq} - \frac{m}{q} + \frac{np}{2q} = \frac{1}{2}\cdot\left(\frac{c}{y} - 2c + cy\right) = \frac{c}{2}\cdot\left(\frac{1}{\sqrt{y}} - \sqrt{y}\right)^2,\quad \frac{1}{\sqrt{npq}} = \frac{p\sqrt{c}}{m\cdot y\sqrt{y}},$$

$$Q_{m,p}(n) \approx \frac{p}{m}\sqrt{\frac{c}{2\pi}}\, y^{-\frac{3}{2}} e^{-\frac{c}{2}\left(\sqrt{y}-\frac{1}{\sqrt{y}}\right)^2} = w_c(y)\frac{p}{m}, \quad \text{и} \quad \boldsymbol{P}\left(\frac{\nu}{\boldsymbol{E}(\nu)} < y\right) = \sum_{h=m}^{[y\boldsymbol{E}(\nu)]} Q_{m,p}(k) \approx \int_0^y w_c(z) = W_c(y),$$

т.е. распределение Вальда.
Рассмотрим следующий практический пример.

Пример 23. ([23,24]). Ниже приведены данные концентраций воздействия хлорбензола (15 мин ожидания) для рабочего-химика за время полной рабочей смены.

13.7, 10.2, 9.9, 4.3, 5.6, 45.6, 42.0, 14.1, 3.8, 9.3,
10.6, 91.3, 2.2, 3.8, 6.0, 17.8,
131.8, 31.0, 4.2, 2.6, 27.6, 1.7, 7.0, 2.1, 1.5, 7.5,
2.5, 2.4, 51.9, 12.9, 12.3 (ppm).

Предполая IG распределение, $\mu = 19.0$ (ppm), $\lambda = 7.2$ (ppm), соответственно, проверить гипотезу согласия. Вычислим сначала $m_1 = 19.0064$ и второй центральный момент $m_2 = 815.36$. Заметим, что значение $\lambda = 7.2$ получена как оценка максимального правдоподобия. Мы подберем значение λ из условия минимума расстояния Колмогорова и получим $\underline{\lambda = 6.2}$.

Совершим следующие преобразования исходных данных $x_i, i = 1,2,...,31$:

$$u_i = \Phi\left(\sqrt{\frac{\lambda}{x_i}}\left(\frac{x_i}{\mu} - 1\right)\right) + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-\sqrt{\frac{\lambda}{x_i}}\left(\frac{x_i}{\mu} + 1\right)\right).$$

Если бы $\lambda = 6.2$ и $\mu = 19.0$ были бы истинными значениями параметра, то u_i имели бы равномерное распределение. Тогда график эмпирической функции распределения, построенной по u_i и прямая $y = x,\ 0 < x < 1,$ были бы близки. Для нашей выборки имеем:

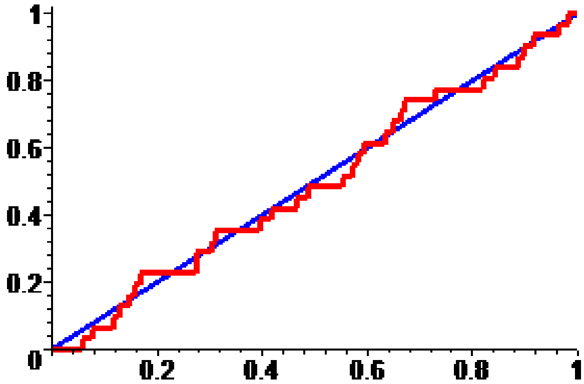


Рис. 8. Спрямленная теоретическая и эмпирическая функции распределения.

Здесь значение расстояния Колмогорова $\Delta_n = 0.043,$ а $D_n = \sqrt{n}\cdot\Delta_n = 0.356$. (здесь $n = 31$).

Так как мы подбирали параметры из условия минимума расстояний Колмогорова, то значение этого расстояния для истинных значений параметров будет не меньше найденного. В нашем случае оно меньше критического значения, которое при уровне значимости $\alpha = 0.05$ равно 1.3581. Найдя значение статистики хи-квадрат, разбив интервал (0,1) на четыре интервала, получим число попаданий в интервалы, равными соответственно 7,8,9,7, найдя оценки по методу минимума хи-квадрат, мы получим значение статистики, равным 0.354 и выбирая критическое значение при числе степеней свободы, равным 1 и уровне значимости 5%, найдем критическое значение, равным 3.841, что больше 0.354. **Вывод:** выборочные данные не противоречат гипотезе согласия: IG-распределению.

13. Процессы рождения и гибели. Пуассоновский процесс. Распределение Пуассона. Распределение Бартлетта.

Рассмотрим систему, которая в момент времени t может находиться в одном из состояний $E_n, n = 0,1,2,...$. В качестве такого примера можно рассматривать число требований (или заявок в системе обслуживания). В следующий момент времени она может перейти в другое состояние или остаться в том же самом состоянии. Обозначим вероятность этого через $P_n(t)$. При любом t имеет место равенство

$$P_0(t) + P_1(t) + P_{n\geq 2}(t) = 1 .$$

Сделаем следующие предположения:

1) вероятность того, что в интервале $(t, t+\Delta t)$ произойдет одно событие (изменение состояния на единицу) равна $\lambda\Delta t + o(\Delta t)$;
2) вероятность того, что в интервале $(t, t+\Delta t)$ не произойдет ни одного события равна $1-\lambda\Delta t + o(\Delta t)$;
3) вероятность того, что в интервале $(t, t+\Delta t)$ произойдет больше одного события равна $o(\Delta t)$ (*ординарность*);
4) процесс стационарен и обладает свойством: события на непересекающихся интервалах – независимы.

По формуле полной вероятности (при $k \geq 1$)

$$P_k(t+\Delta t) = P_0(\Delta t)\cdot P_k(t) + P_1(\Delta t)\cdot P_{k-1}(t) + \sum_{j=2}^{k} P_j(\Delta t)\cdot P_{k-j}(t).$$

Но $\quad \sum_{j=2}^{k} P_j(\Delta t)\cdot P_{k-j}(t) \leq \sum_{j=2}^{k} P_j(\Delta t) \leq \sum_{j=2}^{\infty} P_j(\Delta t) = o(\Delta t).$

Таким образом,

$$P_k(t+\Delta t) = P_0(\Delta t)\cdot P_k(t) + P_1(\Delta t)\cdot P_{k-1}(t) + o(\Delta t).$$

Из 1) и 2) получаем $\quad P_k(t+\Delta t) = (1-\lambda\Delta t)\cdot P_k(t) + \lambda\Delta t\cdot P_{k-1}(t) + o(\Delta t).$

Отсюда $\quad \dfrac{P_k(t+\Delta t) - P_k(t)}{\Delta t} = -\lambda P_k(t) + \lambda P_{k-1}(t) + o(1).$

При $\Delta t \to 0$ предел правой части существует, поэтому существует предел и левой части. В результате получаем уравнение

$$\frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t).$$

Выберем начальные условия: $P_0(0) = 1, P_k(0) = 0, \quad k \geq 1.$

Рассмотрим сначала $P_0(t)$ и промежуток времени 1. Разбивая его на n равных промежутков. В силу условия 4) $p = \left(P_0\left(\dfrac{1}{n}\right)\right)^n \Rightarrow P_0\left(\dfrac{1}{n}\right) = p^{1/n}$. Отсюда при любом целом k

$$P_0\left(\frac{k}{n}\right) = p^{k/n} .$$

Зададим t и выберем такое k (при заданном n), чтобы $\dfrac{k-1}{n} \leq t < \dfrac{k}{n}$.

Вероятность $P_0(t)$ есть убывающая функция времени, поэтому

$$p^{(k-1)/n} = P_0\left(\frac{k-1}{n}\right) \geq P_0(t) \geq P_0\left(\frac{k}{n}\right) = p^{k/n} .$$

Если $\lim\limits_{n\to\infty}\frac{k}{n}=t,$ то $P_0(t)=p^t=e^{-\lambda t},\ \lambda>0.$

Решение (3) будем искать в виде:

$P_k(t)=\exp(-\lambda t)v_k(t),$ при этом мы имеем: $P_0(t)=e^{-\lambda t}\ \Rightarrow\ v_0(t)=1.$

где $v_k(t)$ удовлетворяет условиям: $v_0(0)=1, v_k(0)=0,\ k\ge 1.$

Подстановка приводит нас к уравнению

$$\frac{dv_k(t)}{dt}=\lambda v_{k-1}(t). \qquad (7)$$

В частности, $\dfrac{dv_1(t)}{dt}=\lambda.$

Последовательно решая (7), получаем $v_k(t)=\dfrac{(\lambda t)^k}{k!}.$

Таким образом, $P_k(t)=\dfrac{(\lambda t)^k}{k!}\exp(-\lambda t),\ k\ge 0.$

Процессы рождения и гибели. Пусть

$$\frac{dP_k(t)}{dt}=\underbrace{\lambda_{k-1}P_{k-1}(t)-(\lambda_k+\mu_k)P_k(t)}_{\text{рождение}}+\underbrace{\mu_{k+1}P_{k+1}(t)}_{\text{гибель}},\ k=1,2\dots\ .$$

$$\frac{dP_0(t)}{dt}=-\lambda_0P_0(t)+\mu_1P_1(t),\ \mu_0=0.$$

Из соотношения:

$$P_k(t+\Delta t)=\lambda_{k-1}P_{k-1}(t)\Delta t+(1-\lambda_k\Delta t-\mu_k\Delta t)P_k(t)+\mu_{k+1}P_{k+1}(t)\Delta t+o(\Delta t),\ k=1,2\dots\ .$$

$$P_0(t+\Delta t)=(1-\lambda_0\Delta t)P_0(t)+\mu_1P_1(t)\Delta t+o(\Delta t),$$

$P_k(t)-$ вероятность, что система находится в момент времени t в состоянии E_k.

$E_k\to E_{k+1}$ с вероятностью $\lambda_k\Delta t$, $E_k\to E_{k-1}$ с вероятностью $\mu_k\Delta t$.

Рассмотрим стационарный случай: $\dfrac{dP_k(t)}{dt}=0, k=0,1,2,\dots$.

Тогда

1) $-\lambda_0P_0+\mu_1P_1=0\ \Rightarrow\ P_1=\dfrac{\lambda_0}{\mu_1}P_0,$

2) $\lambda_0P_0-(\lambda_1+\mu_1)P_1+\mu_2P_2=0\ \Leftrightarrow\ \lambda_0P_0-(\lambda_1+\mu_1)\dfrac{\lambda_0}{\mu_1}P_0+\mu_2P_2=0,$

$$-\frac{\lambda_0\lambda_1}{\mu_1}P_0+\mu_2P_2=0\ \Rightarrow\ P_2=\frac{\lambda_0\lambda_1}{\mu_1\mu_2}P_0,$$

. . . .

$$P_k=\frac{\lambda_0\lambda_1\cdot\ldots\cdot\lambda_{k-1}}{\mu_1\mu_2\cdot\ldots\cdot\mu_k}P_0\ .$$

Пусть а) $\lambda_k=\lambda, \mu_k=\mu, k=0,1,\dots, \rho=\dfrac{\lambda_0}{\mu_1}=\dfrac{\lambda_1}{\mu_2}=\dots=\dfrac{\lambda}{\mu}.$

Тогда

$P_1=\rho P_0, P_2=\rho^2P_0,\dots\ \Rightarrow\ P_0+P_1+P_2+\dots=1,$

$\Rightarrow\ P_0(1+\rho+\rho^2+\dots)=1\ \Rightarrow\ P_0=(1-\rho),\quad 0<\rho<1,$

$P_k=(1-\rho)\rho^k,\ k=0,1,2,\dots$.

б) $r=\dfrac{\lambda_0}{\mu_1},\ \dfrac{\lambda_1}{\mu_2}=\dfrac{\lambda_2}{\mu_3}=\dots=\rho.$ (в этом случае получим распределение Бартлетта)

Тогда

$$P_1=\frac{\lambda_0}{\mu_1}P_0=rP_0,$$

$$P_2=\frac{\lambda_0}{\mu_1}\frac{\lambda_1}{\mu_2}P_0=\rho rP_0,\quad P_3=\rho^2 rP_0,\dots$$

$$P_0+P_1+P_2+\dots=1\ \Rightarrow\ P_0(1+r+r\rho+r\rho^2+\dots)=1\ \Rightarrow P_0\left(1+\frac{r}{1-\rho}\right)=1,$$

$$P_0=\frac{1-\rho}{1-\rho+r}=1-\frac{r}{1-\rho+r}=1-\lambda,\quad \lambda=\frac{r}{1-\rho+r}.$$

$$r=\lambda-\rho\lambda+r\lambda,\ r(1-\lambda)=\lambda(1-q),\ rP_0=\lambda(1-\rho),$$

$$P_1=(1-\lambda)\cdot\frac{\lambda}{1-\lambda}(1-\rho)=\lambda(1-\rho),$$

$$P_2=rP_0\rho=\lambda(1-\rho)\rho,$$

$$P_3=\lambda(1-\rho)\rho^2,\dots .$$

Формула Полячека-Хинчина. Рассмотрим систему вида $M|G|1$, где буква M означает, что входной поток ***пуассоновский*** (с интенсивностью λ, значит, имеем пуассоновское распределение с параметром λt, если рассматриваем промежуток длины t), а G, что время обслуживания t имеет ***произвольное*** распределение (в частности, детерминированное). Будем предполагать, что математическое ожидание $\boldsymbol{E}(t)=\dfrac{1}{\mu}$, а дисперсия $\boldsymbol{D}(t)=\sigma^2(t)$. В данном пункте интерес способ вывода формулы Полячека-Хинчина (см. Саати Т. Элементы теории массового обслуживания и ее приложения, М.: Советское радио, 1965 – 510 с., с.69-71).

Введем параметр $\rho=\dfrac{\lambda}{\mu}<1$ и будем также пользоваться обобщенной формулой полной вероятности: если с.в. X и Y – зависимы и существуют приводимые ниже условные математические ожидания, то

$$\boldsymbol{E}(X)=\boldsymbol{E}(\boldsymbol{E}(X|Y))=\int_{-\infty}^{\infty}\boldsymbol{E}(X|Y=y)f_2(y)\,dy,\ \text{а}\ \boldsymbol{E}(X|Y=y)=\int_{-\infty}^{\infty}x\cdot f(x|y)\,dx,$$

$f(x|y)=\dfrac{f(x,y)}{\int_{-\infty}^{\infty}f(x,y)\,dy}$ – условная, а $f(x,y)$ – совместная п. р. величин X и Y.

Допустим, что в момент, когда обслуженный клиент покидает систему, в ней остается q клиентов, считая того, который находится на обслуживании. Время обслуживания этого клиента равно t. Пусть за время t поступит еще r клиентов. Если после ухода следующего клиента в системе остается еще $\tilde{q}$ клиентов, то величины q и $\tilde{q}$ будут связаны следующим образом:

$$\tilde{q}=\max(q-1,0)+r=q-1+\delta+r,\ \text{где}\ \delta=\begin{cases}0, & q\ge 1;\\ 1, & q=0.\end{cases}\qquad(1)$$

Заметим, что $\delta q=0$.

Допустим, что в равновесном состоянии существуют моменты первого и второго порядка $\boldsymbol{E}(q)$ и $\boldsymbol{E}(\tilde{q})$, а q рассматривается как случайная величина. Так как величина $\delta-$ бинарная, то $\delta^2=\delta$. Из стационарности системы (а мы будем это предполагать) следует, что $\boldsymbol{E}(q)=\boldsymbol{E}(\tilde{q})$ и $\boldsymbol{E}(q^2)=\boldsymbol{E}(\tilde{q}^2)$.

Беря математические ожидания, получаем

$\boldsymbol{E}(\tilde{q})=\boldsymbol{E}(q)-1+\boldsymbol{E}(\delta)+\boldsymbol{E}(r)$, откуда $\boldsymbol{E}(\delta)=1-\boldsymbol{E}(r)$.

Если длительность обслуживания равна t, то

$$\boldsymbol{E}(r\,|\,t)=\sum_{r=0}^{\infty} r\cdot\frac{(\lambda t)^r}{r!}e^{-\lambda t}=\lambda t,\quad \boldsymbol{E}(r)=\boldsymbol{E}(\boldsymbol{E}(r\,|\,t))=\frac{\lambda}{\mu}=\rho,\ \text{если}\ \boldsymbol{E}(t)=\frac{1}{\mu}.$$

Аналогично, $\boldsymbol{E}(r^2\,|\,t)=(\lambda t)^2+\lambda t,$

$$\boldsymbol{E}(r^2)=\lambda^2E(t^2)+\rho=\lambda^2\sigma^2(t)+\lambda^2\boldsymbol{E}^2(t)+\rho=\lambda^2\sigma^2(t)+\rho^2+\rho.$$

Возведем обе части (1) в квадрат. Получим:

$$\tilde{q}=q-1+\delta+r\quad\Rightarrow \tilde{q}^2=(q-1+\delta+r)^2=q^2+1+\delta^2+r^2-2q+2q\delta+2qr-2\delta-2r+2\delta r.$$

Отсюда

$$\boldsymbol{E}(\tilde{q}^2)-\boldsymbol{E}(q^2)=0=1+\boldsymbol{E}(\delta)+\boldsymbol{E}(r^2)-2\boldsymbol{E}(q)+2\boldsymbol{E}(q)\boldsymbol{E}(r)-2\boldsymbol{E}(\delta)-2\boldsymbol{E}(r)+2\boldsymbol{E}(\delta)\boldsymbol{E}(r)=$$
$$=1+1-\rho+\boldsymbol{E}(r^2)-2\boldsymbol{E}(q)+2\rho\boldsymbol{E}(q)-2+2\rho-2\rho+2(1-\rho)\rho.$$

Значит, $\boldsymbol{E}(q)=\dfrac{2(1-\rho)\rho}{2(1-\rho)}+\dfrac{\boldsymbol{E}(r^2)-\rho}{2(1-\rho)}=\rho+\dfrac{\lambda^2\sigma^2(t)+\rho^2}{2(1-\rho)}.$

Если t имеет показательное распределение с плотностью $\mu e^{-\mu x},\ x>0,$ то

$\boldsymbol{E}(t)=\dfrac{1}{\mu},\sigma^2(t)=\dfrac{1}{\mu^2},$ откуда $\boldsymbol{E}(q)=\rho+\dfrac{\rho^2}{1-\rho}=\dfrac{1}{1-\rho}$.

Для показательного распределения обслуживания

$$\boldsymbol{E}(\text{длина очереди})=\boldsymbol{E}(q)-(1-P_0)=\boldsymbol{E}(q)-\rho=\frac{\rho^2}{1-\rho}.$$

Действительно, $\boldsymbol{E}(\text{длина очереди})=0\cdot P_0+\sum\limits_{n=1}^{\infty}(n-1)\cdot(1-\rho)\rho^n=\sum\limits_{n=1}^{\infty}n\cdot(1-\rho)\rho^n-\sum\limits_{n=1}^{\infty}(1-\rho)\rho^n=$

$$=\boldsymbol{E}(q)-(1-P_0).$$

Для детерминированного обслуживания $\sigma^2(t)=0$, поэтому

$\boldsymbol{E}(\text{длина очереди})=\dfrac{\rho^2}{2(1-\rho)}$, что в среднем в 2 раза меньше, чем раньше.

14. Механизм слоеобразования. Усеченное нормальное распределение

Рассмотрим следующую задачу распределения мощности слоев в геологических отложениях, предложенную в статье Колмогорова А.Н. Решение одной задачи из теории вероятностей, связанной с вопросом о механизме слоеобразования. ДАН СССР, 1949, т.65, №6, с.793-796 (перепечатана в книге: [6], №37, с.335-339. Для этого сначала обратимся к рисунку.

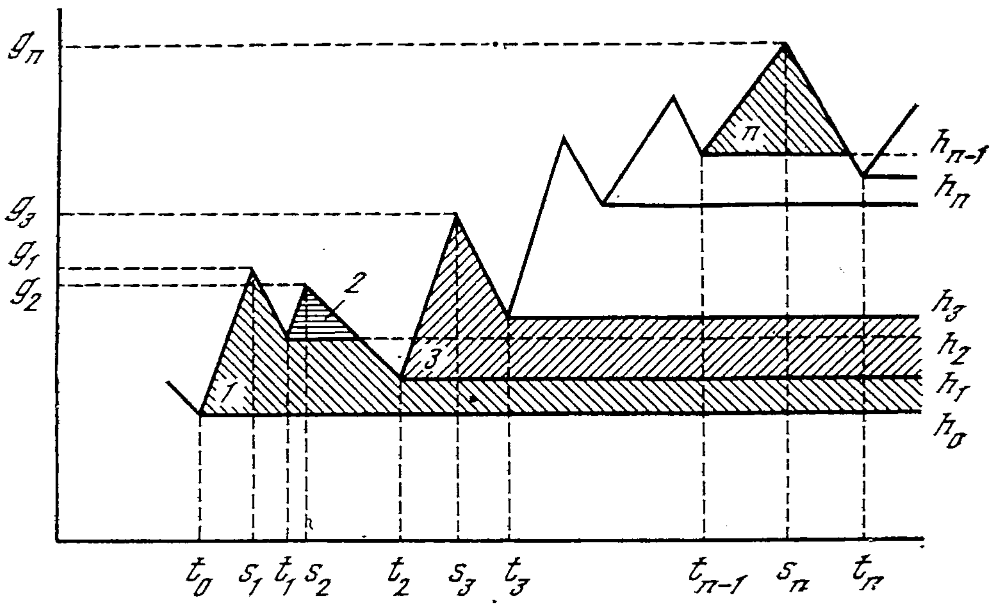


Рис. 8. Механизм слоеобразования.

По горизонтальной оси откладывается время, а по вертикальной – мощность накоплений и размывов. Слой с номером n откладывается в промежуток времени (t_{n-1}, s_n) и имеет (до последующих размывов) мощность $\xi_n = g_n - h_{n-1}$. В промежутки времени (s_n, t_n) происходят размывы на глубину $\eta_n = g_n - h_n$.

В результате такого чередования накоплений и размывов некоторые слои могут размываться несколько раз (слой 1 размывается два раза – при первом и втором размывах), а некоторые могут и совсем исчезнуть (слой 2 полностью смывается при втором размыве). Если считать, что в среднем мощности накоплений больше, чем глубины размывов, то каждый слой будет подвергаться риску быть размытым лишь в течение небольшого числа следующих за его образованием чередований накоплений и размывов. Это позволяет говорить о вероятности «окончательного» сохранения слоя и об условном распределении вероятностей «окончательных» мощностей сохранившихся слоев. Именно эти условные распределения и следует сопоставить с фактически наблюдаемыми статистическими распределениями имеющихся в каком-либо разрезе слоев мощности. По данным А.Б.Вистелиуса, эти последние часто имеют типичную форму распределений резко асимметричных и как бы «урезанных» слева на нулевой абсциссе.

Обозначим $\delta_n = \xi_n - \eta_n = h_n - h_{n-1}$ разность между первоначальной мощностью $n-$го слоя и глубиной непосредственно следующего за его возникновением размыва. Мы будем предполагать заданной бесконечную последовательность случайных величин $\delta_1, \delta_2, ..., \delta_n, ...$ и сделаем следующие допущения:

1) случайные величины δ_n взаимно независимы и имеют одну и ту же функцию распределения $G(x)$;

2) математическое ожидание $\mu = \mathbf{E}(\delta_n) = \int_{-\infty}^{\infty} x\,dG(x)$ положительно;

3) распределение величин непрерывно и $g(x)$ их плотность распределения.

Второе допущение гарантирует, что суммы $\varsigma_n^{(r)} = \delta_n + \delta_{n+1} + ... + \delta_{n+r}$ при неограниченном возрастании второго индекса будут стремиться к $+\infty$ и поэтому нижние грани $\varphi_n = \inf(\varsigma_n^{(0)}, \varsigma_n^{(1)}, ..., \varsigma_n^{(r)}, ...)$ будут конечны и будут достигаться при каком-либо (зависящем от случая) конечном номере r.

Понятно, что слой с номером n в случае $\varphi_n \le 0$ полностью размывается, а в случае от него сохраняется окончательная мощность φ_n. Поэтому задача сводится к определению вероятности $p = \mathbf{P}(\varphi_n > 0)$ и условного распределения величины φ_n при условии $\varphi_n > 0$. Из условия 1) ясно, что как вероятность p, так и указанное распределение не зависят от номера n.

Из допущения 3) можно вывести, что распределение случайных величин φ_n непрерывно предположим с плотностью $f(x)$. Тогда $p = \int_0^\infty f(x)dx$, а условное распределение величины φ_n при гипотезе $\varphi_n > 0$ задается плотностью распределения

$$s(x) = \frac{f(x)}{p} \text{ при } x > 0, \quad s(x) = 0 \text{ при } x \le 0. \qquad (1)$$

Теорема 18. *При допущениях* 1 – 3 *функция* $s(x)$ *является единственным решением интегрального уравнения*

$$s(x) = g(x) + \int_{-\infty}^{0} g(x-y)s(y)dy. \qquad (2)$$

Уравнение (2) равносильно уравнению

$$f(x) = pg(x) + \int_{-\infty}^{0} g(x-y)f(y)dy. \qquad (3)$$

Доказательство равенства (3) основано на следующем:

из определения величин φ_n следует, что

$$\varphi_n = \delta_n, \text{ если } \varphi_{n+1} > 0 \quad \text{и} \quad \varphi_n = \delta_n + \varphi_{n+1}, \text{ если } \varphi_{n+1} < 0. \qquad (4)$$

В силу допущения 1 величины φ_{n+1} и δ_n взаимно независимы, поэтому представив $f(x)$ в виде

$$f(x) = pf_1(x) + (1-p)f_2(x), \qquad (5)$$

где $f_1(x)$ и $f_2(x)$ суть условные плотности для φ_n соответственно при условиях $\varphi_{n+1} > 0$ и $\varphi_{n+1} \le 0$ и заметив, что условная плотность для φ_{n+1} при гипотезе $\varphi_{n+1} \le 0$ есть $\frac{f(x)}{1-p}$ при $x < 0$ и равна 0 при $x > 0$, получим:

$$f_1(x) = g(x), \quad f_2(x) = \frac{1}{1-p}\int_{-\infty}^{0} g(x-y)f(y)dy.$$

Отсюда получаем утверждение теоремы.

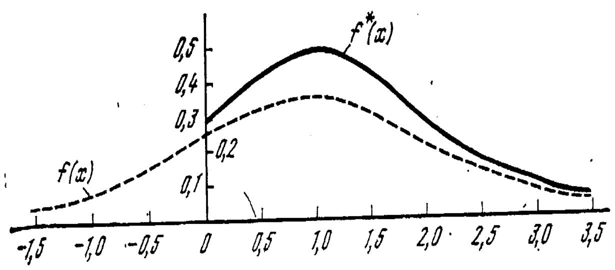


Рис. 9. Усеченное нормальное распределение.

Пример 24. В таблице 5 работы [33, p.13] приведены частоты встречаемости максимальных высот H_s и периода T_{02} волн (станция Tromsøflaket в Норвегии). Эти данные основаны на $n = 15564$ измерениях. Здесь воспроизводятся маргинальные частоты высот H_s.

0 – 0.5	0.5 – 0.9	1.0 – 1.4	1.5 – 1.9	2.0 – 2.4	2.5 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4	4.5 – 4.9	5.0 – 5.4	5.5 – 5.9
32	**1519**	**3011**	**3005**	**2624**	**1771**	**1274**	**810**	**536**	**405**	**235**	**133**

6.0 – 6.4	6.5 – 6.9	7.0 – 7.4	7.5 – 7.9	8.0 – 8.4	8.5 – 8.9	9.0 – 9.4	9.5 – 9.9	10.0 – 10.4	10.5 – 10.9	11.0 – 11.4	11.5 – 11.9
82	**60**	**42**	**25**	**15**	**10**	**9**	2	**2**	**1**	**0**	**2**

В интервалы (9.5, 11.9) попадает 7 наблюдений из 15564, объединим их в один интервал и объединим первый и второй интервал. Таким образом, имеем $N = 15564$ наблюдений.

Принимая модель Колмогорова, получаем, что распределение высот волн должно подчинятьсяусеченному нормальному распределению, а значит, распределение максимальных высот волн – распределению с ф.р. $F(x) = \exp(-\exp((x-\mu)/\sigma)), \ -\infty < x < +\infty.$

В качестве интервалов примем следующие: (0,0.4), (0.4,0.97), (0.97, 1.6), (1.6,2.14), 2.14,2.67), (2.67,3.15), (3.15,3.65), (3.65,4.1), (4.1,4.57), (4.57,5.15), (5.15,5.67), (5.67,6.2), (6.2,6.7), (6.7,7.3), (7.3,8.2), $(8.2, +\infty)$.

Значения соответствующих частот, подсчитанные по модельному распределению равны:

$n_1 = 1561.05, \ n_2 = 3023.8, \ n_3 = 3059.7, \ n_4 = 2604.8, \ n_5 = 1774.3, \ n_6 = 1288.6, \ n_7 = 777.7,$

$n_8 = 537.1, \ n_9 = 406.2, \ n_{10} = 213.5, \ n_{11} = 130, \ n_{12} = 73.6, \ n_{13} = 51.3, \ n_{14} = 37, \ n_{15} = 25.3,$

$n_{16} = 12.2, \ n_{17} = 7.4, \ n_{18} = 4.5, \ n_{19} = 7.$

Соответствующее значение статистики хи-квадрат равно $\chi^2 = 8.1$, а критическое значение есть $\chi^2_{0.95}(16) = 26.296 > 8.1$. Таким образом, критерий хи-квадрат не отвергает гипотезу согласия.

15. Максимальные порядковые статистики, неоднородная выборка

Рассмотрим пример, взятый из книги *Тихов М.С.* Асимптотические методы в теории порядковых статистик и их приложения, изд-во ННГУ, 1985, с.25.

Теорема 19. *Пусть* $X_1, X_2, \dots, X_{l(n)}, X_{l(n)+1}, \dots, X_n$ *– последовательность независимых случайных величин, причем* $X_1, X_2, \dots, X_{l(n)}$ *распределены с функцией распределения* $F(x)$, *а* $X_{l(n)+1}, \dots, X_n$ *– с функцией распределения* $F_1(x)$.

Пусть

1) $F(0) = 0, F_1(0) = 0$ *и для всякого* $\varepsilon > 0$ *имеет место* $F(\varepsilon) > 0, F_1(\varepsilon) > 0$;

2) *для любого* $\tau > 0$

$$\lim_{x \downarrow 0} \frac{F(\tau x)}{F(\tau)} = \tau^{\alpha}, \alpha > 0; \quad \lim_{x \downarrow 0} \frac{F_1(\tau x)}{F_1(\tau)} = \tau^{\alpha+\beta}, \beta > 0.$$

Возьмем $a_n = n^{-\frac{1}{\alpha+\beta}}, \ l(n) = n^{-\frac{\alpha}{\alpha+\beta}}$. *Тогда*

$$\mathbf{P}(X_n^{(1)} < x) \underset{n\to\infty}{\to} 1 - \exp(-x^{\alpha} - x^{\alpha+\beta}), \ x > 0.$$

Доказательство. Действительно,

$$\mathbf{P}(X_n^{(1)} < x) = 1 - (1 - F(a_n x))^{l(n)} (1 - F_1(a_n x))^{n - l(n)}.$$

Но

$$(1 - F(a_n x))^{l(n)} \underset{n\to\infty}{\to} \exp(-x^{\alpha}), \ (1 - F_1(a_n x))^{n-l(n)} \underset{n\to\infty}{\to} \exp(-x^{\alpha+\beta}), \quad x > 0.$$

Следовательно, предельная функция распределения с.в. $X_n^{(1)}$ равна

$$\Phi_1(x) = 1 - \exp(-x^{\alpha} - x^{\alpha+\beta}),\ x > 0.$$

Замечание. Если при x, близких к 0 $F(x) = x^{\alpha}$, а $F_1(x) = x^{\alpha+\beta}$, то

$$(1 - F(a_n x))^{l(n)} = \left(1 - \frac{x^{\alpha}}{n^{\frac{\alpha}{\alpha+\beta}}}\right)^{n^{\frac{\alpha}{\alpha+\beta}}} \underset{n\to\infty}{\to} \exp(-x^{\alpha}),$$

$$(1 - F_1(a_n x))^{n-l(n)} = \left(1 - \frac{x^{\alpha+\beta}}{n}\right)^{n - n^{\frac{\alpha}{\alpha+\beta}}} \approx \left(1 - \frac{x^{\alpha+\beta}}{n}\right)^{n} \underset{n\to\infty}{\to} \exp(-x^{\alpha+\beta}).$$

16. Кривые Пирсона

16.1. Рассмотрим хорошо известное гипергеометрическое распределение

$$\mathbf{P}(Y = r) = y_r = \frac{C_M^r C_{N-M}^{n-r}}{C_N^n}$$

и пусть

$$\Delta y_r = y_r - y_{r-1} = \frac{M!(N-M)!}{r!(M-r)!(n-r)!(N-M-n+r)!C_N^n} -$$

$$- \frac{M!(N-M)!}{(r-1)!(M-r+1)!(n-r+1)!(N-M-n+r-1)!C_N^n} =$$

$$= B \cdot \frac{(M-r+1)(n-r+1) - r(N-M-n+r)}{r \cdot (N-M-n+r)(M-r+1)(n-r+1)},$$

где

$$B = \frac{M!(N-M)!}{C_N^n \cdot (r-1)!(M-r)!(n-r)!(N-M-n+r-1)!},$$

поэтому

$$\Delta y_r = \frac{C_M^r C_{N-M}^{n-r}}{C_N^n} \cdot \frac{(M-r+1)(n-r+1) - r(N-M-n+r)}{(M-r+1)(n-r+1)}\ .$$

Числитель:

$$(M-r+1)(n-r+1) - r(N-M-n+r) = Mn + n + M + 1 - r(N+2) = a + br.$$

Знаменатель:

$$(M-r+1)(n-r+1) = r^2 - r(n+M+2) + Mn + n + 1 = er^2 + dr + c.$$

Отношение равно

$$\frac{\Delta y_r}{y_r} = \frac{a + br}{c + dr + er^2}.$$

В пределе мы получим какую-нибудь плотность распределения и мы приходим к уравнению

$$\frac{df}{f} = \frac{a + bx}{c + dx + ex^2} dx,$$

которое представим в виде

$$\frac{df}{dx} = \frac{(x-a)f}{b_0 + b_1 x + b_2 x^2}.$$

Оказывается, что большинство плотностей распределения можно представить в таком виде. Семейство плотностей, определяемых этой формулой, известно под названием «семейство распределений Пирсона».

Имеем:

$(b_0+b_1x+b_2x^2)\,df=(x-a)\,f\,dx.$

Умножим обе части (2) на x^n, получим

$x^n(b_0+b_1x+b_2x^2)\,df=x^n(x-a)\,f\,dx.$

Интегрируя по частям и предполагая, что интегралы существуют, находим

$$[x^n(b_0+b_1x+b_2x^2)\,f\,]\Big|_{-\infty}^{\infty}-\int_{-\infty}^{\infty}(nb_0x^{n-1}+(n+1)b_1x^n+(n+2)b_2x^{n+1}))\,f\,dx=$$

$$=\int_{-\infty}^{\infty}x^{n+1}f\,dx-a\int_{-\infty}^{\infty}x^n f\,dx.$$

Если существует момент порядка $n+2$, то выражение в квадратных скобках будет равно нулю и обозначая $\alpha_k=\int\limits_{-\infty}^{\infty}x^k f\,dx$, запишем:

$$-nb_0\delta_{n-1}-(n+1)b_1\delta_n-(n+2)b_2\delta_{n+1}=\delta_{n+1}-a\delta_n,$$

или

$$nb_0\delta_{n-1}+((n+1)b_1-a)\delta_n+((n+2)b_2+1)\delta_{n+1}=0.$$

Полагая последовательно $n=0,1,2,3,$ $(\delta_{-1}=0,\ \delta_0=1)$ мы находим

$$a=-\frac{\mu_3(\mu_4+3\mu_2^2)}{A}=b_1,\ b_0=-\frac{\mu_2(4\mu_2\mu_4-3\mu_3^2)}{A},\ \ b_2=-\frac{(2\mu_2\mu_4-3\mu_3^2-6\mu_2^3)}{A}.$$

Здесь μ_2,μ_3,μ_4 – центральные моменты, $A=10\mu_4\mu_2-18\mu_2^3-12\mu_3^2.$

Имеем: $\mu_2=\delta_2-\delta_1^2,\ \ \mu_3=\delta_3-3\delta_2\delta_1+2\delta_1^3,\ \ \mu_4=\delta_4-4\delta_3\delta_1+6\delta_2^2-3\delta_1^4$

Параметр a здесь является модой. Принимая моду за начало отсчета, первое уравнение можно записать в виде

$$\frac{d}{dx}(\ln f)=\frac{x-a}{B_0+B_1(x-a)+B_2(x-a)^2}\ \Leftrightarrow\ \frac{d}{dX}(\ln f)=\frac{X}{B_0+B_1X+B_2X^2}.$$

В зависимости от коэффициентов B_0,B_1,B_2 получаем 12 типов распределений.

Кривые Пирсона классифицируются в зависимости от характера корней уравнения $b_0+b_1x+b_2x^2$. Семейство кривых Пирсона составляют 12 типов и нормальное распределение. Приведем некоторые примеры и условия, при которых возникают соответствующие распределения.

Тип I. Бета-распределение

Пусть

$$B_0+B_1X+B_2X^2=B_2(x+\alpha_1)(X-\alpha_2),\ \ \alpha_1,\alpha_2>0.$$

Тогда

$$\frac{d}{dX}(\ln f)=\frac{X}{B_0+B_1X+B_2X^2}=\frac{X}{B_2(X+\alpha_1)(X-\alpha_2)}=\frac{\alpha_1}{B_2(\alpha_1+\alpha_2)(X+\alpha_1)}+\frac{\alpha_2}{B_2(\alpha_1+\alpha_2)(X-\alpha_2)},$$

отсюда

$$f=k(X+\alpha_1)^{\frac{\alpha_1}{B_2(\alpha_1+\alpha_2)}}(X-\alpha_2)^{\frac{\alpha_2}{B_2(\alpha_1+\alpha_2)}}\ \Leftrightarrow\ f=k\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_{12}}\right)^{m_2},\ \text{где}\ \ \frac{m_1}{a_1}=\frac{m_2}{a_2}.$$

Если $B_2=0,$ то получаем гамма распределение (III тип распределения)

$$f=k\left(1+\frac{x}{a}\right)^p e^{-px/a},\ -a<x<\infty.$$ И т.д.

Классификация типов ***кривых Пирсона***.

Тип I. Бета-распределение.

$$f(x)=k\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}, \quad -a_1<x<a_2, m_1>-1, m_2>-1.$$

Тип II. Частный случай бета-распределения, когда $a_1=a_2=a, m_1=m_2=m$:

$$f(x)=k\left(1-\frac{x^2}{a^2}\right)^{m}, \quad -a<x<a, m>-1.$$

Чаще всего бета-распределение записывают в виде

$$f(x)=\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}, \quad 0<x<1;\ p,q>0.$$

Для этого распределения

1) математическое ожидание $\mu=\dfrac{p}{p+q}$. 2) дисперсия $\sigma^2=\dfrac{pq}{(p+q)^2(p+q+1)}$.

3) асимметрия $\beta_1=\dfrac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$

4) эксцесс $\beta_2=6\dfrac{p^3-p^2(2q-1)+q^2(q+1)-2pq(q+2)}{pq(p+q+2)(p-q+3)}$.

Тип III. Гамма-распределение с плотностью

$$f(x)=\frac{k^{m+1}}{\Gamma(m+1)}(a+x)^m e^{-k(a+x)}, \quad x>-a,\ k>0.$$

Обычно гамма-распределение $\Gamma(\alpha,\theta)$ записывают в виде

$$f(x)=\frac{x^{\alpha-1}}{\theta^{\alpha}\Gamma(\alpha)}e^{-x/\theta}, \quad x>0;\ \theta>0,\alpha>0.$$

При $\alpha=\dfrac{n}{2}, \theta=2$ – имеем распределение хи-квадрат-распределение с n степенями свободы.

Числовые характеристики:

1) $\mu=\alpha\theta$; 2) $\sigma^2=\alpha\theta^2$; 3) $\beta_1=\dfrac{2}{\sqrt{\alpha}}$; 4) $\beta_2=\dfrac{6}{\alpha}$.

Тип IV. Условия: $D=b_0b_2-b_1^2>0, 0<\lambda=\dfrac{b_1^2}{b_0b_2}<1, b_0+2b_1x+b_2x^2=(a^2+x^2)b_2$.

Плотность распределения равна $f(x)=c(a^2+x^2)^{-m}\exp\left(-\nu\cdot\operatorname{arctg}\dfrac{x}{a}\right), x\in R, m\geq\dfrac{1}{2}$.

На рисунке приведен график функции $\dfrac{1}{1+x^2}\exp(-\operatorname{arctg}(x))$.

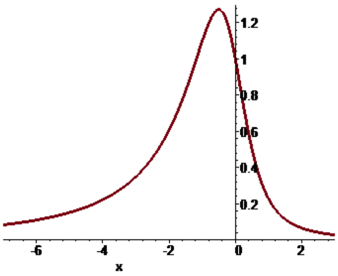


Рис. 10. Плотность распределения типа IV.

Тип V. Условия: $D=0, \lambda=1, b_0+2b_1x+b_2x^2=(a+x)^2 b_2$.
Плотность распределения равна

$$f(x)=\frac{\gamma^{m-1}}{\Gamma(m-1)\,x^m}e^{-\frac{\gamma}{x}},\ \gamma>0, m>1,\ \ x>0.$$

На рисунке представлен график функции $x^{-2}\exp(-1/x)$.

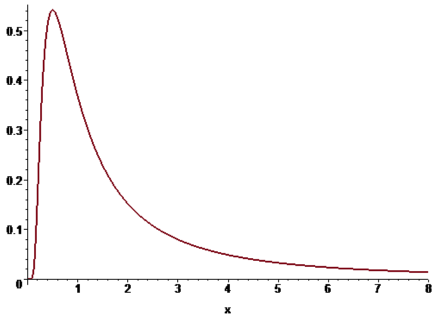


Рис. 11. Плотность распределения типа V.

Тип VI. Условия: $D<0, \lambda>1, b_0+2b_1x+b_2x^2=(a+x)(x-b)b_2$.
Плотность распределения

$$f(x)=\frac{(a+b)^{-(m+n+1)}}{B(-m-n-1, n+1)}(x+a)^m(x-b)^n,\ x>b, m>1, n>-1.$$

Тип VII. Распределение Стьюдента. Условия: $D>0, \lambda=0, b_0+2b_1x+b_2x^2=(a^2+x^2)b_2$. Плотность распределения равна

$$f(x)=\frac{a}{B(m-\frac{1}{2},\frac{1}{2})}(a^2+x^2)^{-m},\ x\in R,\ m\geq\frac{1}{2}.$$

Тип VIII. Условия: $D<0, \lambda<0, b_0+2b_1x+b_2x^2=(a+x)xb_2$.
Плотность распределения равна

$$\frac{m+1}{a^{m+1}}(x+a)^m,\ x\in[-a,0],\ -1<m<0.$$

Тип IX. Условия: $D<0, \lambda<0, b_0+2b_1x+b_2x^2=(a+x)xb_2$.

Плотность распределения равна $\dfrac{m+1}{a^{m+1}}(x+a)^m,\ x\in[-a,0],\ m<-1.$

Тип X. Показательное распределение. Условия: $D=0, \lambda=0, b_0+2b_1x+b_2x^2=b_0, a=0$. Плотность равна $f(x)=\gamma e^{-\gamma x},\ x>0;\ \gamma>0.$

Тип XI. Нормальное распределение. Условия $D=0, \lambda-$ неопределенно, $b_0+2b_1x+b_2x^2=b_0$.
Плотность распределения равна

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right),\ \ x\in R.$$

Тип XII. Условия как для типа I с условиями $\alpha=\beta, m=-n.$

16.2. Распределение Panjer.

Рассмотрим сначала (дискретное) биномиальное распределение

$$b_k=C_n^k p^k(1-p)^{n-k},\ k=0,1,\ldots,n,$$

и найдем отношение этих вероятностей:

$$\frac{b_k}{b_{k-1}}=\frac{C_n^k p^k(1-p)^{n-k}}{C_n^{k-1}p^{k-1}(1-p)^{n-k+1}}=\frac{(n-k+1)p}{k(1-p)}=-\frac{p}{1-p}+\frac{(n+1)p}{k(1-p)}=a+\frac{b}{k},\ k=1,2,\ldots,n,$$

где $a=-\frac{p}{1-p}, b=\frac{(n+1)p}{1-p}$.

Аналогично, для распределения Пуассона с $\pi_k=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,\dots,$ имеем: $\frac{\pi_k}{\pi_{k-1}}=\frac{\lambda}{k}=a+\frac{b}{k}$, т.е. можно написать, что $a=0, b=\lambda$.

Для отрицательного биномиального распределения

$$q_k=\frac{(m+k-1)}{k!(m-1)!}p^m(1-p)^k,\ k=0,1,\dots,\ \frac{q_k}{q_{k-1}}=(1-p)+\frac{(1-p)(m-1)}{k},\ m\ge 1-\text{фиксировано},\ k=0,1,\dots.$$

Существует универсальная формула для этих распределений, называемая распределением Pajer ($Y\in P(a,b)$)

$$p_k=\left(1+\frac{\lambda}{\alpha}\right)^{-\alpha}\frac{\lambda^k}{k!}\prod_{j=1}^{k-1}\frac{\alpha+j}{\alpha+\lambda},\ \text{где}\ a=\frac{\lambda}{\lambda+\alpha}, b=\frac{(\alpha-1)\lambda}{\alpha+\lambda}.$$

Отметим, что распределение Пуассона получается как предельный случай $\alpha\to\pm\infty$, оно совпадает с (обобщенным) отрицательным биномиальным распределением для положительных конечных $\in\mathbf{R}_{>0}$ и с биномиальным распределением для отрицательных целых чисел $-\alpha\in\mathbf{Z}$.

Используя соотношение $p_k/p_{k-1}=a+b/k$, легко вывести, что $\mathbf{E}(Y)=\lambda$, $\mathbf{D}(Y)=\frac{\lambda(\alpha+\lambda)}{\alpha}$.

Для гипергеометрического распределения с $g_k=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, \frac{g_k}{g_{k-1}}=\frac{(M-k+1)(n-k+1)}{k(N-M-n+k-1)}=1+\frac{b}{k}+\frac{c}{k+n-M-N}$,

где $b=-\frac{(M+1)(N+1)}{N-n+M}, c=-\frac{(n-M+1)(N-n-1)}{N-n+M}$

Для логарифмического распределения $l_k=\frac{p^k}{(-\ln q)k}, k=1,2,\dots, a=p, b=-p, q=1-p,$

и $l_k=l_{k-1}\left(a+\frac{b}{k}\right)$, правда, в отличие от предыдущего случая, распределение начинается с 1.

17. Бета распределение и закон арксинуса

Семейство бета-распределений образуется распределениями с плотностью вида

$$f(y)=\frac{1}{B(p,q)}\frac{(y-a)^{p-1}(b-y)^{q-1}}{(b-a)^{p+q-1}},\ a\le y\le b;\ B(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Преобразование

$$X=\frac{Y-a}{b-a}$$

приводит к плотности

$$f(x)=\frac{1}{B(p,q)}x^{p-1}(1-x)^{q-1},\ 0\le x\le 1.$$

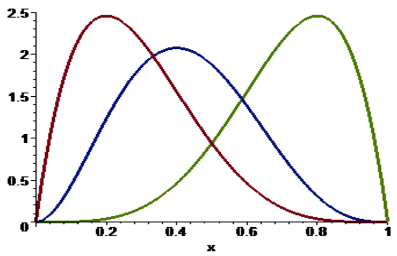


Рис. 12. Плотность бета-распределения:

1) $p=2, q=5$ 2) $p=3, q=4$ 3) $p=5, q=2$ (слева направо)

Если $q=1$, то получаем степенное распределение $f(x)=px^{p-1}, 0\le x\le 1.$

Если $p=1, q=1$, то получаем равномерное на $(0,1)$ распределение с плотностью

$$f(x)=1,\ 0\le x\le 1;\quad =0 \text{ в противном случае.}$$

Если $p=\frac{1}{2}, q=\frac{1}{2}$, то получаем

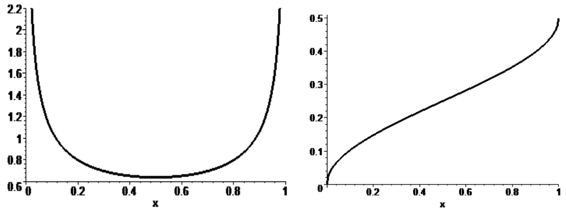


Рис. 13. Распределение арксинус с плотностью $f(x)=\dfrac{1}{\pi\sqrt{x(1-x)}}, 0\le x\le 1,$ и функцией распределения $F(x)=\dfrac{2}{\pi}\arcsin\sqrt{x},\ 0\le x\le 1.$

Задача 5. Покажите, что если $X_1\in G(\alpha,p_1)$ и $X_2\in G(\alpha,p_2)$ независимы, то $U_1=X_1+X_2$ и $U_2=\dfrac{X_1}{X_2}$ независимы и распределены: U_1 как $G(\alpha,p_1+p_2)$, а U_2 имеет плотность.

Плотность распределения U_2 равна $\dfrac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)}\cdot\dfrac{u_2^{p_1-1}}{(1+u_2)^{p_1+p_2}}$.

Задача. Пусть $X\in G(\alpha,p_1), Y\in G(\alpha,p_2)$ и независимы. Тогда случайная величина $V=\dfrac{X}{X+Y}$ имеет бета-распределение с плотностью

$$\frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)}v^{p_1-1}(1-v)^{p_2-1}, 0<v<1.$$

Рассмотрим сначала **числа Каталана**.

Порядок вычислений в арифметических выражениях задается расстановкой скобок, например,

$$(3-1)\cdot(4+(15-9)\cdot(2+6)).$$

Если стереть все элементы выражения, за исключением скобок, то оставшиеся скобки образуют *правильную скобочную структуру*:

$$(\ )((\ )\ (\ )).$$

Выпишем все правильные скобочные структуры пар скобок:

$n=1$	( )	1
$n=2$	( )( ) (( ))	2
$n=3$	( )( )( ) ( )(( )) (( ))( ) (( )( )) ((( )))	5
$n=4$		14

$n = 5$ 42

$n = 6$ 132

Всякая правильная скобочная структура удовлетворяет следующим двум условиям:

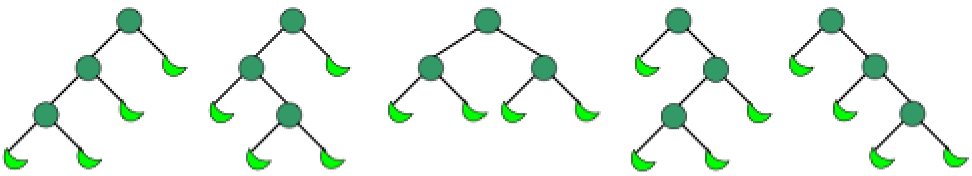
1) число левых и правых скобок в правильной скобочной структуре одинаково;
2) число левых скобок в любом начальном отрезке правильной скобочной структуры не меньше числа правых скобок.

В правильной скобочной структуре все скобки разбиваются на пары: каждой левой скобке соответствует парная ей правая скобка. Парная правая скобка выделяется следующим правилом: это первая правая скобка справа от данной левой скобки, такая, что между выбранными двумя скобками стоит правильная скобочная структура.

Определение 1. Числом Каталана c_n называется число различных правильных скобочных структур из n пар скобок.

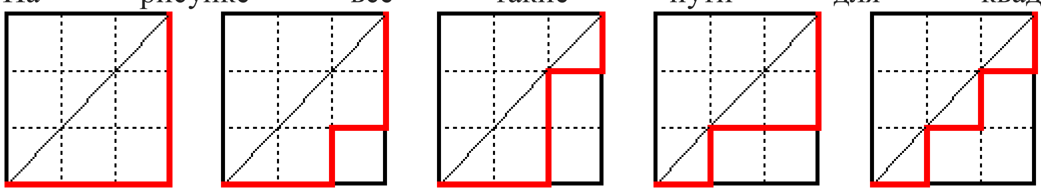
Английская Википедия утверждает, что известно, как минимум 66 различных конструкций, которые приводят к появлению чисел Каталана. Вот некоторые из них: Правильные скобочные последовательности – наборы открывающихся и закрывающихся скобок, в которых каждой открывающейся скобке соответствует закрывающаяся. Число возможных последовательностей с фиксированным числом пар скобок выражается числом Каталана. Например, 14 правильных последовательностей из четырех пар скобок: (((()))), ((()())), ((())()), ((()))(), (()(())), (()()()), (()())(), (())(()), (())()(), ()((())), ()(()()), ()(())(), ()()(()), ()()()()

Двоичные деревья – деревья, из каждого узла которых (кроме листьев) выходит ровно две ветки. Количество бинарных деревьев с заданным числом листьев – число Каталана. На рисунке представлены пять деревьев с 4 листьями в каждом.

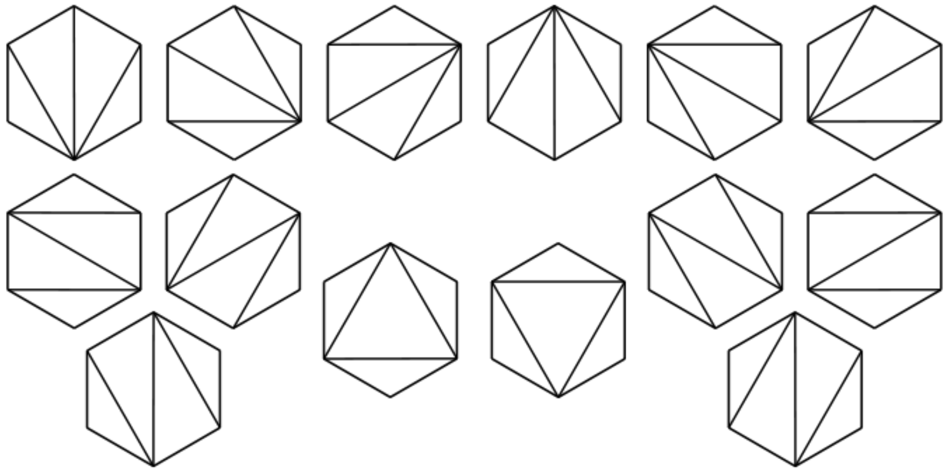


Любые деревья. Число неизоморфных деревьев с заданным числом вершин также равно числу Каталана.

- Монотонные пути в квадрате – маршруты из левого нижнего угла квадрата в правый верхний, которые идут по линиям сетки вверх или вправо и не заходят выше диагонали. На рисунке все такие пути для квадрата 3x3.



- Триангуляции многоугольника. Количество различных триангуляций выпуклого многоугольника диагоналями равно числу Каталана.



Разбиение вершин многоугольника на пары. Четное число точек на окружности можно объединить в пары непересекающимися хордами. Число способов таких объединений также равно числу Каталана.

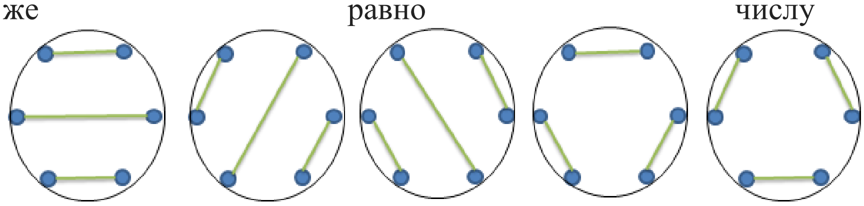


Таблица Юнга – прямоугольник, заполненный последовательными числами так, чтобы они возрастали во всех строках и столбцах. Число таблиц Юнга размером 2xn также выражается числом Каталана.

1	2	3
4	5	6

1	2	4
3	5	6

1	2	5
3	4	6

1	3	4
2	5	6

1	3	5
2	4	6

Положим $c_0 = 1$. Тогда последовательность чисел Каталана начинается так:

1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, …

Чтобы вывести производящую функцию для чисел Каталана, найдем сначала рекуррентное соотношение для этих чисел.

Рассмотрим в правильной скобочной структуре из $n+1$ пар скобок пару скобок, в которую входит самая левая скобка структуры. Тогда последовательность скобок внутри этой пары образует правильную скобочную структуру и последовательность скобок вне этой пары образует правильную скобочную структуру: (…..) …, где каждое многоточие обозначает некоторую правильную скобочную структуру. Если число пар скобок во внутренней структуре равно k, то во внешней структуре $n-k$ пар скобок. Наоборот, по каждой паре скобочных структур из k и $n-k$ пар скобок можно восстановить структуру из $n+1$ пар скобок, заключив первую структуру в скобки и приписав к результату справа вторую структуру.

Отсюда мы получаем рекуррентное соотношение для чисел Каталана:

$$c_{n+1} = \sum_{j=1}^{n} c_j c_{n-j} .$$

Например, $(c_0)^2 = c_1$, $c_5 = 1\cdot 14 + 1\cdot 5 + 2\cdot 2 + 5\cdot 1 + 14\cdot 1 = 42$.

Рассмотри производящую функцию для чисел Каталана:

$\varphi(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3 + \ldots = 1 + t + 2t^2 + 5t^3 + \ldots$.

Возводя ее в квадрат и умножив результат на t, получим

$$t\varphi^2(t) = (c_0)^2 t + (c_0 c_1 + c_1 c_0)t^2 + (c_0 c_2 + c_1 c_1 + c_2 c_0)t^3 + \ldots = c_1 t + c_2 t^2 + c_3 t^3 + \ldots = \varphi(t) - 1,$$

что дает нам квадратное уравнение на производящую функцию $t\varphi^2(t) - \varphi(t) + 1 = 0$, Откуда $$\varphi(t)=\frac{1-\sqrt{1-4t}}{2t}.$$

Второй корень отбрасывается, так как $(1+\sqrt{1-4t})/2t=1/t+...$ содержит отрицательные степени t. Найдем явный вид чисел Каталана. Согласно формуле Тейлора

$$c_n=\frac{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{2}\cdot\ldots\cdot\frac{2n-1}{2}\cdot 4^{n+1}}{2(n+1)!},$$

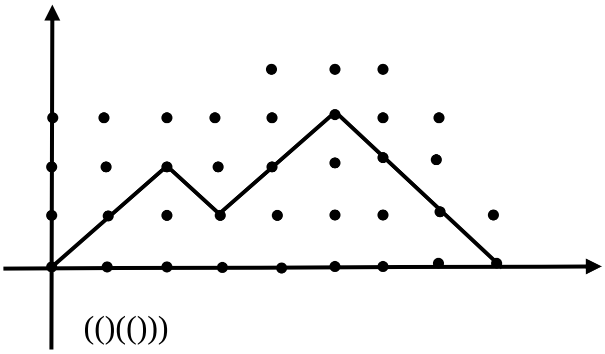
откуда, умножая числитель и знаменатель на $n!$ и сокращая на 2^{n+1}, получим

$$c_n=\frac{(2n)!}{n!(n+1)!}=\frac{1}{n+1}\binom{2n}{n} \qquad \text{и} \qquad c_{n+1}=\frac{4n+2}{n+2}c_n .$$

Еще одна важная реализация чисел Каталана связана с путями Дика на плоскости.

Рассмотрим целочисленную решетку в положительном квадранте.

Путем Дика называется непрерывная ломаная в верхней полуплоскости, составленная из векторов $(1,1)$ и $(1,-1)$, начинающаяся в начале координат и заканчивающаяся на оси абсцисс.



(()(()))

Рис. 13. Путь Дика и соответствующая ему скобочная структура.

Соответствие между путями Дика и правильными скобочными структурами: нужно поставить вектору $(1,1)$ левую скобку, а $(1,-1)$ – правую скобку. Тогда условие, что путь заканчивается на оси абсцисс и находится в верхней полуплоскости, есть в точности условие правильной скобочной структуры. Поэтому

Число путей Дика из $2n$ *звеньев равно* n*-му числу Каталана* c_n.

Случайные блуждания. Пусть $\xi_1,\xi_2,\ldots,\xi_n$ – последовательность независимых одинаково распределенных случайных величин с распределением

$$\mathbf{P}(\xi_i=1)=\frac{1}{2},\ \mathbf{P}(\xi_i=-1)=\frac{1}{2},\ S_k=\xi_1+\ldots+\xi_k, 1\le k\le n,\ S_0=0.$$

Обозначим $\sigma_{2n}=\min\{1\le k\le 2n: S_k=0\}$, полагая $\sigma_{2n}=\infty$, если $S_k\ne 0$ при всех $1\le k\le 2n$: σ_{2n} – это момент первого возвращения в нуль.

Пусть $u_{2k}=\mathbf{P}(S_{2k}=0),\quad f_{2k}=\mathbf{P}(\sigma_{2n}=2k),\ 1\le k\le n.$

Имеем: $u_0=1$ и $u_{2k}=C_{2k}^k\left(\frac{1}{2}\right)^{2k}$,

так как чтобы попасть в нуль надо k раз двигаться вверх из $2k$ попыток. Всего случаев 2^{2k}.

Имеем: $c_n=\dfrac{(2n)!}{n!(n+1)!}=\dfrac{1}{n+1}\dbinom{2n}{n}=C_{2n}^n-C_{2n}^{n-1}.$

Приведем формулу Стирлинга: $\dfrac{n!}{\sqrt{2\pi n}\cdot\left(\dfrac{n}{e}\right)^n} \sim 1$. Из нее получаем

$$u_{2k}=\frac{2k!}{(k!)^2}\cdot\frac{1}{2^{2k}}\sim\frac{(2k)^{2k}\cdot\sqrt{2\pi\cdot 2k}\cdot e^{2k}}{e^{2k}\cdot 2\pi k\cdot k^{2k}}\cdot\frac{1}{2^{2k}}=\frac{1}{\sqrt{\pi k}},\ k\to\infty,\quad f_{2k}\sim\frac{1}{\sqrt{\pi}\cdot k^{3/2}},\ \ k\to\infty.$$

Обозначим через $P_{2k,2n}$ вероятность того, что на отрезке $[0,2n]$ частица проводит $2k$ единиц времени на положительно стороне (мы говорим, что в интервале $[m-1,m]$ частица находится на положительной стороне, если по крайней мере одно из значений S_{m-1} или S_m положительно).

Лемма. *Пусть* $u_0=1$ *и* $1\le k\le n$. *Тогда*

$$P_{2k,2n}=u_{2k}\cdot u_{2(n-k)}=\frac{C_{2k}^{k}\cdot C_{2n-2k}^{n-k}}{2^{2n}}.$$

Доказательство. При $k=0$ и при $k=n$ формула верна. Пусть теперь $1\le k\le n-1$. Пусть $2r$ есть момент первого возвращения в нуль. Возможны два случая: 1) когда $S_1>0,\,k\le 2r,$ и $S_1<0,\,k\le 2r.$ Тогда

$$P_{2k,2n}=\frac{1}{2}\sum_{r=1}^{k}f_{2r}\cdot P_{2(k-r),2(n-r)}+\frac{1}{2}\sum_{r=1}^{k}f_{2r}\cdot P_{2k,2(n-r)},$$

поскольку во втором слагаемом частица на отрицательной стороне.

По предположении индукции $P_{2k,2m}=u_{2k}\cdot u_{2m-2k}$ для $m=1,\ldots,n-1$. Подставляя в последнее равенство, находим, что

$$P_{2k,2n}=\frac{1}{2}u_{2n-2k}\sum_{r=1}^{k}f_{2r}\cdot u_{2k-2r}+\frac{1}{2}u_{2k}\sum_{r=1}^{k}f_{2r}\cdot u_{2n-2k-2r}=\frac{1}{2}u_{2n-2k}u_{2k}+\frac{1}{2}u_{2k}u_{2n-2k}=u_{2k}u_{2n-2k}.$$

Значит, при $k\to\infty,\ n-k\to\infty,$

$$P_{2k,2n}=u_{2k}\cdot u_{2(n-k)}\sim\frac{1}{\pi\sqrt{k(n-k)}}=\frac{1}{\pi n\sqrt{z(1-z)}},\ z=\frac{k}{n}.$$

Рассмотрим следующий пример (см. [34])

Т а б л и ц а. **Ряд распределения возраста научных работников СССР, в %** (**1928 г.**)

Возраст	Число научных работников	Возраст	Число научных работников	Возраст	Число научных работников
20 – 25	14	45 – 50	151	70 – 75	2
25 – 30	93	50 – 55	106	75 – 80	2
30 – 35	165	55 – 60	67	80 – 85	1
35 – 40	182	60 – 65	32	Σ	1000
40 – 45	179	65 – 70	6		

Найдем выборочные моменты: $m_1=42.045,\ m_2=99.22,\ m_3=426.4,\ m_4=28555.9.$

Отсюда найдем асимметрию $\gamma_1=\dfrac{m_3}{m_2^{3/2}}=0.43$ и эксцесс $\gamma_2=\dfrac{m_4}{m_2^2}-3=-0.1.$

Приводя выборочные данные к интервалу $(0,1)$, имеем плотность распределения

$$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ \ 0<x<1.$$

Так как $\boldsymbol{E}(X)=\dfrac{\alpha}{\alpha+\beta},\ \boldsymbol{D}(X)=\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, то оценки для (α,β) найдем по методу моментов. В нашем случае они равны: $\hat{\alpha}=3.5,\ \hat{\beta}=7.3.$ Область возможных значений разобьем на 10 интервалов одинаковой длины, присоединив четыре последних интервала. Найдем оценки минимума хи-квадрат. Они равны $(\tilde{\alpha}=3, \tilde{\beta}=5.2)$. Провеhим гипотезу согласия с бета-распределением с указанными значениями $(\tilde{\alpha}, \tilde{\beta})$ и подсчитаем предполагаемые частоты попадания в интервалы. Мы имеем:

$n_1=17.1, n_2=86, n_3=155.7, n_4=191.5, n_5=187.6, n_6=154.2, n_7=107.3, n_8=62, n_9=28.2, n_{10}=10.4.$

Значение статистики хи-квадрат равно

$$\chi^2=\frac{(14-17.1)^2}{17.1}+\frac{(93-86)^2}{86}+\frac{(165-155.7)^2}{155.7}+\frac{(182-191.6)^2}{191.6}+\frac{(179-187.6)^2}{187.6}+\frac{(151-154.2)^2}{154.2}+\frac{(106-107.3)^2}{107.3}+\frac{(67-62)^2}{62}+\frac{(32-28.2)^2}{28.2}+\frac{(11-10.4)^2}{10.4}=$$
$$=3.6.$$

Так как оценивалось два параметра, а число интервалов равно 10, то степень свободы критерия хи-квадрат равна $k=10-2-1=7$. Выбрав уровень значимости равным $\alpha=0.05$, найдем критическое значение $\chi^2_{0.95}(7)=14.067>3.6$. Отсюда делаем вывод: **выборочные данные не противоречат гипотезе о бета-распределении**.

18. Зависимость доза-эффект. Оценки типа NW. Состоятельность и асимптотическая нормальность

Изложение данного вопроса начнем со следующей леммы (см. [8], т.2, VII.1, с. 255).

Рассмотрим семейство распределений $F_{n,\theta}, n=1,2,...,$ с математическим ожиданием θ и дисперсией $\sigma_n^2(\theta), \theta\in[a,b]\subset\mathbf{R}$. Пусть $g(x)-$ограниченная и непрерывная функция. Её математическое ожидание равно

$$\boldsymbol{E}_{n,\theta}(g(X))=\int_{-\infty}^{\infty}g(x)\,dF_{n,\theta}(x).$$

Лемма 4. *Если* $\sigma_n^2(\theta)\to 0$ *при любом* θ, *то*

$$\boldsymbol{E}_{n,\theta}(g(X))\to g(\theta)$$

равномерно в любом замкнутом интервале, где $\sigma_n^2(\theta)\to 0$ *равномерно и функция* g *равномерно непрерывна.*

Доказательство. Имеем:

$$|\,\boldsymbol{E}_{n,\theta}(g(X))-g(\theta)|\le\int_{-\infty}^{\infty}|g(x))-g(\theta)|\,dF_{n,\theta}(x).$$

Существуют δ, зависящее от θ, и ε, такие, что при $|x-\theta|<\delta$ подынтегральное выражение меньше ε. Вне этой окрестности подынтегральное выражение меньше некоторой постоянной M, а по неравенству Чебышева вероятность попадания в область $|x-\theta|\ge\delta$ не больше чем $\sigma_n^2(\theta)\delta^{-2}$. Таким образом, правая часть не превосходит 2ε, если n столь велико, что $\sigma_n^2(\theta)<\varepsilon\delta^2/M$. Эта граница не зависит от θ, если $\sigma_n^2(\theta)\to 0$ равномерно и функция g равномерно непрерывна.

Примеры 25. а) Если $F_{n,\theta}-$ биномиальное распределение, сосредоточенное в точках $k/n, k=0,1,...,n,$ то $\sigma_n^2(\theta)=\theta(1-\theta)n^{-1}\to 0$ и

$$\sum_{k=0}^{n}g\left(\frac{k}{n}\right)C_n^k\theta^k(1-\theta)^{n-k}\to g(\theta)\ \textit{равномерно при}\ 0\le\theta\le 1.$$

б) Если $F_{n,\theta}-$ распределение Пуассона, в точках $k/n, k=0,1,...,$ то $\sigma_n^2(\theta)=\theta n^{-1}\to 0$ и

$$e^{-n\theta}\sum_{k=0}^{n}g\left(\frac{k}{n}\right)\frac{(n\theta)^k}{k!}\to g(\theta)\ \textit{равномерно в каждом конечном интервале значений}\ \theta.$$

в) Если $F_{n,\theta}-$гамма-распределение, с ожиданием θ и дисперсией θ/n, получаем

$\frac{1}{(n-1)!}\int_{-\infty}^{\infty} g(x)\left(\frac{nx}{\theta}\right)^{n-1} e^{-nx/\theta}\,\frac{n\,dx}{\theta} \to g(\theta)$ *равномерно в каждом конечном интервале значений* θ.

Пусть $X_1, X_2, ..., X_n$ – независимые и одинаково распределенные (нор) случайные величины (св) с неизвестной непрерывной функцией распределения $F(x)$, $U_1, U_2, ..., U_n$ – нор св, независимые от X_i – ов, с неизвестным распределением G. Мы наблюдаем $\mathscr{U}^{(n)} = \{(U_i, W_i), i = 1, 2, ..., n\}$, где $W_i = I(X_i < U_i)$ есть индикатор события $(X_i < U_i)$. Рассматривается задача оценивания функции распределения $F(x)$ для фиксированного значения x на основе интервальных цензурированных данных $\mathscr{U}^{(n)}$.

Отметим, что если X и U независимы, то $\boldsymbol{E}(W \mid U = x) = F(x)$, т.е. задача оценивания $F(x)$ сводится в этом случае к задаче оценивания регрессии.

Одна из **интепретаций** данной ситуации следующая: величина U – введенная в организм доза какого-либо вещества, X – нижняя граница, с которой начинается реакция организма. Если $(U > X)$, то есть реакция (эффект) и $W = I(U > X) = 1$, в противном случае $W = 0$.

Рассмотрим последовательности оценок: NW-оценки (Надарая, 1964), (Watson, 1964)

$$F_n(x) = \frac{S_{2n}(x)}{S_{1n}(x)},\; S_{1n}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{U_i - x}{h}\right),\; S_{2n}(x) = \frac{1}{nh}\sum_{i=1}^{n} W_i K\left(\frac{U_i - x}{h}\right),$$

где $K(x)$ – ограниченная четная неотрицательная с компактным носителем функция ($K(x) = 0$ для $x \notin [-1,1]$), $h = h(n)$ – неслучайная последовательность такая, что $h \to 0$, но $nh \to \infty$ при $n \to \infty$, обычно мы берем $h = Mn^{-1/5}, M > 0.$

Если $\sum_{i=1}^{n} K((U_i - x)/h) = 0,$ то полагаем $F_n(x) = 0$.

ТЕОРЕМА 20. *Пусть* $(X_i, U_i), i = 1, 2, ..., n,$ – *независимые и одинаково распределенные двумерные случайные векторы с совместной абсолютно непрерывной функцией распределения и предположим, что функции плотности* $G'(x) = g(x) > 0$ *и* $F'(x) = f(x)$ *непрерывны и ограничены, существуют константы* $L_0 > 0$ *и* $M_1 > 0, M_2 > 0,$ *такие, что*

$$|g(u_2) - g(u_1)| \le L_0 |u_2 - u_1|,\; g(u) \le M_1,\; f(x) \le M_2.$$

Тогда

$$F_n(x) = \frac{S_{2n}(x)}{S_{1n}(x)} \xrightarrow[n\to\infty]{p} F(x).$$

Д о к а з а т е л ь с т в о. Достаточно доказать, что

$$S_{2n}(x) \xrightarrow[n\to\infty]{p} F(x)g(x),\; S_{1n}(x) \xrightarrow[n\to\infty]{p} g(x).$$

Пусть $\boldsymbol{E}(S_{1n}(x)) = g_n(x)$ Рассмотрим ожидание $g_n(x) - g(x)$. Имеем:

$$g_n(x) - g(x) = \frac{1}{h}\int_{x-h}^{x+h} K\left(\frac{u-x}{h}\right)(g(u) - g(x))\,du = \int_{-1}^{1} K(z)(\,g(x+zh) - g(x))\,dz.$$

Из условий теоремы и так как функция $|zK(z)|$ интегрируема, то

$$|g_n(x) - g(x))| \le L_0 h\int_{-1}^{1} |z| K(z)\,dz \to 0 \quad (n \to \infty).$$

Точно так же

$$\boldsymbol{E}((S_{1n}(x) - g(x))^2) = \boldsymbol{D}(S_{1n}(x)) + \boldsymbol{E}^2(S_{1n}(x) - g(x)),$$

$$\boldsymbol{E}((S_{1n}(x) - \boldsymbol{E}(S_{1n}(x))^2) = \boldsymbol{D}(S_{1n}(x)) = \frac{1}{nh^2}\boldsymbol{D}\left(K\left(\frac{U-x}{h}\right)\right) \le \frac{1}{nh^2}\boldsymbol{E}\left(K^2\left(\frac{U-x}{h}\right)\right) =$$

$$= \frac{1}{nh}\int_{x-h}^{x+h} K^2(z)\,g(x+zh)\,dz = \frac{1}{nh}\int_{x-h}^{x+h} K^2(z)(g(x+zh)-g(x))\,dz + \frac{1}{nh}g(x)\int_{-1}^{1} K^2(z)\,dz \approx$$

$$= \frac{1}{nh}g(x)\int_{-1}^{1} K^2(z)\,dz + o\left(\frac{1}{nh}\right) \approx \frac{g(x)\|K\|^2}{nh}.$$

Обозначим $\|K\|^2 = \int_{-\infty}^{\infty} K^2(z)\,dz,\ \alpha_1 = \int_{-\infty}^{\infty} |z|\,K^2(z)\,dz,\ \mu_2(K) = \int_{-\infty}^{\infty} z^2K(z)\,dz.$

Тогда

$$\int_{-\infty}^{\infty} K^2(z)\,g(x+zh)\,dz = \|K\|^2\ g(x) + \int_{-\infty}^{\infty} K^2(z)(g(x+zh)-g(x))\,dz \le \|K\|^2\ g(x) + L_0\alpha_1 h.$$

Таким образом,

$$\boldsymbol{E}((S_{1n}(x) - \boldsymbol{E}(S_{1n}(x))^2) \le \frac{\|K\|^2\ g(x)}{nh} + \frac{L_0\alpha_1}{n} \to 0,\quad n\to\infty.$$

Из закона больших чисел в форме теоремы Чебышева, примененной к сумме $S_{1n}(x)$, следует, что $S_{1n}(x)$ сходится по вероятности к $g(x)$:

$$S_{1n}(x) \underset{n\to\infty}{\overset{p}{\to}} g(x).$$

Перейдем к сумме $S_{2n}(x)$. Имеем:

$$\boldsymbol{E}(S_{2n}(x)) = \boldsymbol{E}\left(\frac{1}{nh}\sum_{i=1}^{n} W_i K\left(\frac{U_i - x}{h}\right)\right) = \frac{1}{h}\int_{-\infty}^{\infty} \boldsymbol{E}\left(I(X<U)K\left(\frac{U-x}{h}\right)|U=u\right)g(u)\,du =$$

$$= \frac{1}{h}\int_{-\infty}^{\infty} F(u)K\left(\frac{u-x}{h}\right)g(u)\,du.$$

Сделаем замену $u = x + zh$. В таком случае

$$\boldsymbol{E}(S_{2n}(x)) = \int_{-\infty}^{\infty} F(x+zh)\,g(x+zh)\,K(z)\,dz.$$

Следовательно,

$$|\,\boldsymbol{E}(S_{2n}(x)) - F(x)g(x)| \le \int_{-\infty}^{\infty} |\,F(x+zh)\,g(x+zh) - F(x)g(x)\,|\,K(z)\,dz$$

и

$$\boldsymbol{E}(S_{2n}(x)) = F(x)g(x) + o(1/n).$$

Так как $F(x)\le 1,\ g(x)\le M_1, |F(x+h)-F(x)| \le M_2\,|z|h,$ то

$|F(x+zh)\,g(x+zh) - F(x)g(x)| \le |F(x+zh)\,g(x+zh) - F(x+h)g(x)| + |F(x+zh)\,g(x) - F(x)g(x)| \le$
$\le |\,g(x+zh) - g(x) + M_1\,|F(x+zh) - F(x)| \le L_0\ |z|h + M_1M_2|z|h = C_1|z|h,$

где $C_1 = L_0 + M_1M_2.$ Отсюда

$$\boldsymbol{D}(S_{2n}(x)) = \frac{1}{nh^2}\boldsymbol{D}\left(WK\left(\frac{U-x}{h}\right)\right) = \frac{1}{nh^2}\left(\boldsymbol{E}\left(W^2K^2\left(\frac{U-x}{h}\right)\right) - \boldsymbol{E}^2\left(WK\left(\frac{U-x}{h}\right)\right)\right) \le$$

$$\le \frac{1}{nh^2}\boldsymbol{E}\left(WK^2\left(\frac{U-x}{h}\right)\right) = \frac{1}{nh}\int_{-\infty}^{\infty} F(x+zh)g(x+zh)K^2(z)\,dz \le \frac{g(x)\|K\|^2}{nh} \ge \frac{M_1\,\|K\|^2}{nh} \to 0,\quad n\to\infty.$$

Значит, $\qquad S_{2n}(x) \underset{n\to\infty}{\overset{p}{\to}} F(x)g(x).$

Из (1) и (2) следует утверждение Теоремы 1.

Обозначим

$$r_1(x)=\frac{1}{2}M^{5/2}\mu_2^2(K)g''(x),\quad r_2(x)=\frac{1}{2}M^{5/2}\mu_2^2(K)(F(x)g(x))'',$$

$\sigma_1^2(x)=\|K\|^2\ g(x),\quad \sigma_2^2(x)=\|K\|^2\ F(x)g(x),\ M-$ константа, выбор которой мы уточним позже.

ТЕОРЕМА 21. *Пусть* $(X_i,U_i),\ i=1,2,...,n,-$ *независимые и одинаково распределенные двумерные случайные векторы с совместной абсолютно непрерывной функцией распределения* $F(x)G(u),\ (x,u)\in R^2$.

Предположим, что

(A1) для любого x $G(x)$ *имеет ограниченную плотность* $g(x)>0$;

(A2) *плотности* $g(x)$ *и* $f(x)$ *имеют ограниченные производные до третьего порядка включительно, причем* $|g'''|\le L_3$ *;*

(A3) *для любого* $x\in[0,1]$ *функция* $K(x)$ *ограничена, т.е.* $\sup\limits_{x\in[-1,1]} K(x)\le K_1$.

Тогда

(i) $\quad \sqrt{nh}\,(S_{1n}(x)-g(x))\xrightarrow[n\to\infty]{d} N(r_1(x),\sigma_1^2(x));$

(ii) $\quad \sqrt{nh}\,(S_{2n}(x)-F(x)g(x))\xrightarrow[n\to\infty]{d} N(r_2(x),\sigma_2^2(x)).$

Д о к а з а т е л ь с т в о. Изучим сначала поведение $S_{1n}(x)$. На множестве $(-\infty,-1)\bigcup(1,+\infty)$ функция $K(x)$ равна нулю, поэтому

$$g_n(x)-g(x)=\int_{-1}^{1}K(z)(g(x+zh)-g(x))\,dz=\int_{-1}^{1}K(z)\left(zhg'(x)+\frac{1}{2}z^2h^2g''(x)+\frac{1}{6}z^3h^3g'''(z^*)\right)dz=$$

$$=\frac{h^2g''(x)}{2}\int_{-1}^{1}z^2K(z)\,dz+\frac{h^3}{6}\int_{-1}^{1}z^3g'''(z^*)K(z)\,dz,$$

где z^*-точка из интервала $(x,x+x+zh)$.

Из условия (A3) теоремы следует, что

$$\left|\int_{-1}^{1}z^3g'''(z^*)K(z)\,dz\right|\le L_3\int_{-1}^{1}|z|^3\ K(z)\,dz<\frac{L_3K_1}{2}<\infty$$

и $\quad g_n(x)-g(x)=\dfrac{h^2g''(x)}{2}\displaystyle\int_{-1}^{1}z^2K(z)\,dz+O(h^4).$

Теперь рассмотрим $\boldsymbol{E}((S_{1n}(x)-g(x))^2)=\boldsymbol{D}(S_{1n}(x))+(g_n(x)-g(x))^2$. При $n\to\infty$ второе слагаемое эквивалентно $\dfrac{h^4(g''(x))^2}{4}\mu_2^4(K)$. Первый член есть дисперсия величины $S_{1n}(x)$. Так как

$$\boldsymbol{D}(S_{1n}(x))=\frac{1}{nh^2}\boldsymbol{E}\left(K^2\left(\frac{U-x}{h}\right)\right)-\frac{1}{n}\left(\frac{1}{h}\boldsymbol{E}\left(K\left(\frac{U-x}{h}\right)\right)\right)^2,$$

то

$$\boldsymbol{D}(S_{1n}(x))=\frac{1}{nh}\int_{-\infty}^{+\infty}K^2(z)\,g(x+zh)\,dz+O\left(\frac{1}{n}\right)=\frac{\|K\|^2\ g(x)}{nh}+o\left(\frac{1}{nh}\right).$$

Кроме того,

$$\int\limits_{K^2(z)>\varepsilon\sqrt{nh}}K^2(z)\,g(x+zh)\,dz\le\frac{L}{h}\int\limits_{K^2((u-x)/h)>\varepsilon\sqrt{nh}}K^2\left(\frac{u-x}{h}\right)g(u)\,du\to0,\ n\to\infty.$$

Из центральной предельной теоремы Линдеберга-Феллера, примененной к последовательности сумм $S_{1n}(x)$ следует, что

$$\lim_{n\to\infty} \boldsymbol{P}\left(\frac{\sqrt{nh}\,(S_{1n}(x)-g(x))}{\sigma_1(x)} < y \right) = \Phi(y-r_1(x)) = \int_{-\infty}^{y-r_1(x)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Рассмотрим теперь сумму $S_{2n}(x)$ обозначив $\varphi_n(x)=\boldsymbol{E}(S_{2n}(x))$ и $\varphi(x)=F(x)g(x)$.

Имеем:

$$\boldsymbol{E}(S_{2n}(x)) = \int_{-\infty}^{+\infty} \varphi(x+zh)K(z)\,dz = F(x)g(x) + \frac{h^2\varphi''(x)\mu_2^2(K)}{2} + \frac{h^3}{6}\int_{-\infty}^{\infty} |z|^3 K(z)\varphi'''(z^{**})dz,$$

где $\left| \int_{-\infty}^{\infty} |z|^3 K(z)\varphi'''(z^{**})dz \right| \le L_3 \int_{-\infty}^{\infty} |z|^3 K(z)\,dz.$

Отсюда

$$\varphi_n(x) - F(x(g(x) = \frac{h^2\varphi''(x)\mu_2^2(K)}{2} + O(h^3).$$

Теперь

$$\boldsymbol{E}((S_{2n}(x)-F(x)g(x))^2) = \boldsymbol{D}(S_{2n}(x)) + (\varphi_n(x)-\varphi(x))^2 = \boldsymbol{D}(S_{2n}(x)) + \frac{h^4\mu_2^4(K)(\varphi''(x))^2}{4}(1+o(1)).$$

Так как $\boldsymbol{D}(S_{2n}(x)) = \frac{1}{nh^2}\int_{-\infty}^{\infty} K\left(\frac{u-x}{h}\right)\varphi(u)\,du - \frac{1}{n}\left(\frac{1}{h}\boldsymbol{E}\left(K\left(\frac{U-x}{h}\right)\right)\right)^2,$

то $\boldsymbol{D}(S_{2n}(x)) = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(z)\varphi(x+zh)\,dz + O\left(\frac{1}{n}\right) = \frac{\|K\|^2\varphi(x)}{nh} + o\left(\frac{1}{nh}\right).$

Кроме того,

$$\lim_{n\to\infty} \boldsymbol{P}\left(\frac{\sqrt{nh}\,(S_{2n}(x)-F(x)g(x))}{\sigma_2(x)} < y \right) = \Phi(y-r_2(x)).$$

ТЕОРЕМА 22. *Предположим* $h = Mn^{-1/5}$, $F(x)$ *и* $g(x) > 0 -$ *трижды непрерывно дифференцируемы. Тогда*

(i) $\dfrac{\sqrt{nh}\,(F_n(x)-F(x))}{b(x)\|K\|}$ *сходится по распределению в точке* x *к нормальной случайной величине со средним* $M^{5/2}\cdot a(x) \ne 0$ *и единичной дисперсией, где*

$$b^2(x) = \frac{F(x)(1-F(x)\|K\|^2}{g(x)},\ \ a(x) = \frac{f'(x)+2f(x)g'(x)}{g(x)};$$

(ii) *минимум среднеквадратичного отклонения* $\boldsymbol{E}((F_n(x)-F(x))^2)$ *достигается при*

$$M = \left(\frac{b^2(x)\|K\|^4}{a^2(x)\mu_2^4(K)} \right)^{1/5}.$$

Д о к а з а т е л ь с т в о. Обозначим $a_1 = g(x), a_2 = F(x)g(x), S_2 = S_{2n}(x), S_1 = S_{1n}(x).$ Используя разложение функции $1/(1+x)$ в окрестности нуля, получим:

$$\frac{S_2}{S_1} = \frac{S_2-a_2+a_2}{S_1-a_1+a_1} = \frac{(S_2-a_2)+a_2}{a_1\left(1+\dfrac{S_1-a_1}{a_1}\right)} = \frac{S_2-a_2}{a_1}\left(1-\frac{S_1-a_1}{a_1}+O\left(\left(\frac{S_1-a_1}{a_1}\right)^2\right)\right) + \frac{a_2}{a_1}\left(1-\frac{S_1-a_1}{a_1}+O\left(\left(\frac{S_1-a_1}{a_1}\right)^2\right)\right) =$$

$$= \frac{a_2}{a_1} + \frac{S_2-a_2}{a_1} - \frac{a_2}{a_1^2}(S_1-a_1) + O\left(\frac{(S_2-a_2)(S_1-a_1)}{a_1^2}\right) + O\left(\frac{a_2(S_1-a_1)^2}{a_1^3}\right).$$

Таким образом,

$$\frac{S_2}{S_1}-\frac{a_2}{a_1}=\frac{S_2-a_2}{a_1}-\frac{a_2}{a_1^2}(S_1-a_1)+O\left(\frac{(S_2-a_2)(S_1-a_1)}{a_1^2}\right)+O\left(\frac{a_2(S_1-a_1)^2}{a_1^3}\right).$$

Поскольку $(S_2-a_2)(S_1-a_1)$ и $(S_1-a_1)^2$ сходится к нулю быстрее, чем (S_1-a_1) и (S_2-a_2), то

$$\frac{S_2}{S_1}-\frac{a_2}{a_1}\approx\frac{(S_2-a_2)}{a_1}-\frac{a_2}{a_1^2}(S_1-a_1),\ \left(\frac{S_2}{S_1}-\frac{a_2}{a_1}\right)^2\approx\frac{(S_2-a_2)^2}{a_1^2}-\frac{a_2^2}{a_1^4}(S_1-a_1)^2-2\frac{a_2}{a_1^3}(S_2-a_2)(S_1-a_1).$$

Отсюда

$$\boldsymbol{E}\left(\frac{S_2}{S_1}-\frac{a_2}{a_1}\right)\approx\frac{\boldsymbol{E}(S_2-a_2)}{a_1}-\frac{a_2}{a_1^2}\boldsymbol{E}(S_1-a_1)=\frac{h^2a_2''\,\|K\|^2}{2a_1}-\frac{h^2a_2a_1''\,\|K\|^2}{2a_1^2}+o(h^2)=$$

$$=\frac{h^2\,\|K\|^2\,(a_2''a_1-a_1''a_2)}{2a_1^2}+o(h^2)=\frac{M^2a(x)}{n^{2/5}}+o(n^{-2/5}),$$

$$\boldsymbol{D}\left(\frac{S_2}{S_1}-\frac{a_2}{a_1}\right)\approx\frac{\boldsymbol{D}(S_2-a_2)}{a_1^2}+\frac{a_2^2}{a_1^4}\boldsymbol{D}(S_1-a_1)-2\frac{a_2}{a_1^3}\mathbf{COV}(S_1-a_1,S_2-a_2)=$$

$$=\frac{\boldsymbol{D}(S_2-a_2)}{a_1^2}+\frac{a_2^2}{a_1^4}\boldsymbol{D}(S_1-a_1)-2\frac{a_2}{a_1^3}\mathbf{COV}(S_1,S_2).$$

Учитывая, что

$$\mathbf{COV}(S_1,S_2)=\frac{1}{nh^2}\int_{-\infty}^{\infty}K^2\left(\frac{u-x}{h}\right)F(u)g(u)\,du+O\left(\frac{1}{n}\right)=\frac{\|K\|^2\,F(x)g(x)}{nh}+o\left(\frac{1}{nh}\right),$$

получаем

$$\boldsymbol{D}\left(\frac{S_2}{S_1}-\frac{a_2}{a_1}\right)=\frac{\|K\|^2\,a_2}{nha_1^2}+\frac{\|K\|^2\,a_2^2a_1}{nha_1^4}-2\frac{\|K\|^2\,a_2^2}{nha_1^3}+o\left(\frac{1}{nh}\right)=\frac{\|K\|^2\,a_2}{nha_1^2}\left(1-\frac{a_2}{a_1}\right)+o\left(\frac{1}{nh}\right)=\frac{b^2(x)}{nh}+o\left(\frac{1}{nh}\right)$$

Перейдем к части (ii) теоремы. Мы имеем

$$\boldsymbol{E}\left(\left(\frac{S_2}{S_1}-\frac{a_2}{a_1}\right)^2\right)=\frac{\boldsymbol{E}((S_2-a_2)^2)}{a_1^2}+\frac{a_2^2}{a_1^4}\boldsymbol{E}((S_1-a_1)^2)-2\frac{a_2}{a_1^3}\boldsymbol{E}((S_1-a_1)(S_2-a_2)).$$

Используя соотношения

$$\boldsymbol{E}((S_2-a_2)^2)=\frac{\mu_2^2(K)a_2}{nh}+\frac{h^4\,\|K\|^4\,(a_2'')^2}{4}+o\left(\frac{1}{n^{4/5}}\right),\ \boldsymbol{E}((S_1-a_1)^2)=\frac{\mu_2^2(K)a_1}{nh}+\frac{h^4\,\|K\|^4\,(a_1'')^2}{4}+o\left(\frac{1}{n^{4/5}}\right),$$

$$\boldsymbol{E}((S_1-a_1)(S_2-a_2))=\mathbf{COV}(S_1,S_2)+(\boldsymbol{E}(S_1)-a_1)(\boldsymbol{E}(S_2)-a_2)=\frac{\mu_2^2(K)a_2}{nh}+\frac{h^4\,\|K\|^4\,(a_2'')^2(a_1'')^2}{4}+o\left(\frac{1}{n^{4/5}}\right),$$

получаем $\boldsymbol{E}\left(\left(\frac{S_2}{S_1}-F(x)\right)^2\right)=\frac{F(x)(1-F(x))\mu_2^2(K)}{nh}+\frac{h^4\,\|K\|^4}{4}\left(\frac{(a_2'')^2}{a_1^2}+\frac{a_2^2(a_1'')^2}{a_1^4}-2\frac{a_2a_1''a_2''}{a_1^3}\right)+o\left(\frac{1}{n^{4/5}}\right)=$

$$=\frac{b^2(x)\mu_2^2(K)}{nh}+\frac{h^4\,\|K\|^4}{4}\left(\frac{a_2''-a_2a_1''}{a_1^2}\right)^2+o\left(\frac{1}{n^{4/5}}\right)=$$

$$=\frac{b^2(x)\mu_2^2(K)}{nh}+\frac{h^4\,\|K\|^4}{4}a^2(x)+o\left(\frac{1}{n^{4/5}}\right)=\frac{b^2(x)\mu_2^2(K)}{Mn^{4/5}}+\frac{M^4\,\|K\|^4}{4n^{4/5}}a^2(x)+o\left(\frac{1}{n^{4/5}}\right).$$

Рассмотрим функцию $\psi(M)=\frac{b^2(x)\mu_2^2(K)}{M}+\frac{1}{4}M^4\,\|K\|^4\,a^2(x).$

Минимум этой функции достигается при $M,$ являющимся решением уравнения

$$\psi'(M)=0\quad\text{или}\quad-\frac{b^2(x)\mu_2^2(K)}{M^2}+M^3\,\|K\|^4\,a(x)=0,\ \text{откуда}\ M=\left(\frac{b^2(x)\,\|K\|^4}{a^2(x)\mu_2^4(K)}\right)^{1/5}.$$

При таком M среднеквадратичное уклонение равно

$$\boldsymbol{E}\left(\left(\frac{S_2}{S_1}-F(x)\right)^2\right)=\frac{5(b^8(x)\mu_2^4(K)\,\|K\|^4)^{1/5}}{4n^{4/5}}+o\left(\frac{1}{n^{4/5}}\right).$$

Замечание. Покажем конспективно как получается соотношение для корреляций. Обозначим $K_i=K((x-u_i)/h)$. Тогда как и при выводе дисперсии,

$$\mathbf{cov}(S_1,S_2)=\frac{1}{n^2h^2}\mathbf{E}\left(\sum_{i=1}^{n}(K_i-\mathbf{E}(K_i))\sum_{j=1}^{n}(W_jK_j-\mathbf{E}(W_jK_j))\right)=\frac{1}{n^2h^2}\mathbf{E}\left(\sum_{i=1}^{n}(W_iK_i-\mathbf{E}(W_iK_i))(K_i-\mathbf{E}(K_i))\right)\approx$$

$$\approx\frac{1}{n^2h^2}\mathbf{E}\left(\sum_{i=1}^{n}W_iK_i^2\right)=\frac{1}{nh^2}\int_{-\infty}^{+\infty}\frac{1}{h}K\left(\frac{u-x}{h}\right)F(u)g(u)\,du\approx\frac{1}{nh}F(x)g(x)\|K\|^2.$$

Ниже приведены моделированные данные (табл.1) и реальные данные (табл.2) и графики оценок функций распределения величины X.

Таблица 20. Моделированные значения переменных: $X_i\in N(17.5;4)$

i	$U_i\in N(18.0;4)$	$W_i=I(X_i<U_i)$	i	$U_i\in N(18.0;4)$	$W_i=I(X_i<U_i)$
1	12.965	0	**26**	18.109	1
2	13.861	0	**27**	18.116	1
3	14.232	0	**28**	18.134	1
4	14.483	0	**29**	18.136	1
5	14.673	0	**30**	18.149	1
6	14.696	0	**31**	18.229	1
7	14.919	0	**32**	18.324	0
8	14.924	0	**33**	18.349	1
9	15.557	0	**34**	18.435	0
10	15.640	0	**35**	18.607	1
11	15.863	0	**36**	18.775	1
12	15.887	0	**37**	18.810	0
13	15.904	1	**38**	18.949	1
14	16.099	0	**39**	19.201	1
15	16.426	0	**40**	19.415	1
16	16.506	0	**41**	19.833	1
17	16.655	0	**42**	19.968	0
18	16.734	1	**43**	20.046	1
19	16.776	0	**44**	20.204	1
20	16.897	0	**45**	20.211	1
21	17.058	1	**46**	20.954	1
22	17.085	0	**47**	21.590	1
23	17.126	1	**48**	21.662	1
24	17.145	0	**49**	22.144	1
25	17.480	1	**50**	23.580	1

```
U:=(12.3,13.861,14.232,14.483,14.673,14.696,14.919,14.924,15.557,15.6
40,15.863,15.887,15.904,16.099,16.426,16.506,16.655,16.734,16.776,16.
897,17.058,17.085,17.126,17.145,17.480,18.109,18.116,18.134,18.136,18
.149,18.229,18.324,18.349,18.435,18.607,18.775,18.810,18.949,19.201,1
9.415,19.833,19.968,20.046,20.204,20.211,20.954,21.590,21.662,22.144,
23.580);
W:=(0,0,0,0,0,0,0,0,0,0,0,0,0.7,0,0,0,0,0.5,0,0,0.9,0,1,0,1,0.9,1,1,
0.9,1,1,0,1,0,1,1,0.4,1,1,1,1,1,0.9,1,1,1,1,1,1,1);
 n:=50;h:=2.5;
 Kern:=proc(x) if x<-1 then 0 elif (x<1) then 0.75*(1-x^2) else 0 fi
 end:
```

$s1:=\mathbf{proc}(x)$

```
local i,s1;
    s1:=0; for i to n do s1:=s1+Kern((x-U[i])/h) end do
end proc
s2:=proc(x)
local i,s2;
    s2:=0; for i to n do s2:=s2+W[i]*Kern((x-U[i])/h) end do
end proc
plot({stats[statevalf,cdf,normald[17.5,2]],s2/s1},9..27,y=0..1,thic
kness=3);
```

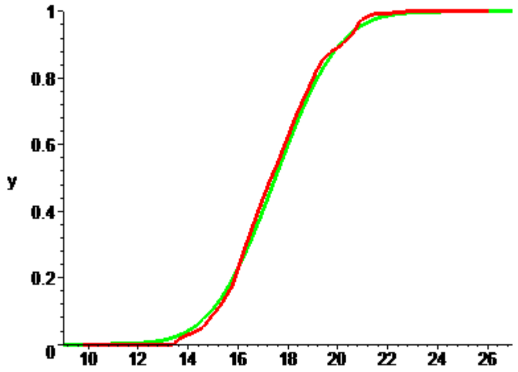


Рис.14. График функции распределения величины X (данные табл.20).

Таблица 21. Источник: Finney D. Probit-analysis, 1984, $N=\sum_{i=1}^{25}M_i=3918$.

i	Возраст U_i	Число положительных исходов (единиц) $W_i=I(X_i<U_i)$	Число случаев M_i
1	9.21	0	376
2	10.21	0	200
3	10.58	0	93
4	10.83	2	120
5	11.08	2	90
6	11.33	5	88
7	11.58	10	105
8	11.83	17	111
9	12.08	16	100
10	12.33	29	93
11	12.58	39	100
12	12.83	51	108
13	13.08	47	99
14	13.33	67	106
15	13.58	81	105
16	13.83	88	117
17	14.08	79	98
18	14.33	90	97
19	14.58	113	120
20	14.83	95	102
21	15.08	117	122
22	15.33	107	111
23	15.58	92	94
24	15.83	112	114
25	17.58	1049	1049

```
with(plots); with(stats); with(stats[statplots]);
```

U := (9.21, 10.21, 10.58, 10.83, 11.08, 11.33, 11.58, 11.83, 12.08, 12.33, 12.58, 12.83, 13.08, 13.33, 13.58, 13.83, 14.08, 14.33, 14.58, 14.83, 15.08, 15.33, 15.58, 15.83, 17.58);
M := (376., 200., 93., 120., 90., 88., 105., 111., 100., 93., 100., 108., 99., 106., 105., 117., 98., 97., 120., 102., 122., 111., 94., 114., 1049.); *W* := (0., 0., 0., 2., 2., 5., 10., 17., 16., 29., 39., 51., 47., 67., 81., 88., 79., 90., 113., 95., 117., 107., 92., 112., 1049.);
N := 376+200+93+120+90+88+105+111+100+93+100+108+99+106+105+117+98+97+120+ +102+122+111+94+114+1049; *N* := 3918; *n* := 25; *h* := .7;
Kern := **proc** (*x*) **if** *x* < – 1 **then** 0 **elif** *x* < 1 **then** $0.9375*(1-x^2)^2$ **else** 0 **end if end proc**;
s1 := **proc** (*x*) local *i*, *s1*; *s1* := 0; **for** *i* **to** *n* **do** *s1* := *s1*+*M*[*i*]**Kern*((*x*–*U*[*i*])/*h*) **od**; **end**;
s2 := **proc** (x) **local** *i*, *s2*; *s2* := 0; **for** *i* **to** *n* **do** *s2* := *s2*+*W*[*i*]**Kern*((*x*–*U*[*i*])/*h*) **od**; **end**;
$t4 := x \to \frac{s2(x)}{s1(x)}$;
plot({*t4*, *stats*[*statevalf*, *cdf*, *normald*[12.9854, 1.1]]}, 10 .. 16.5, *y* = 0 .. 1.0, *thickness* = 3);

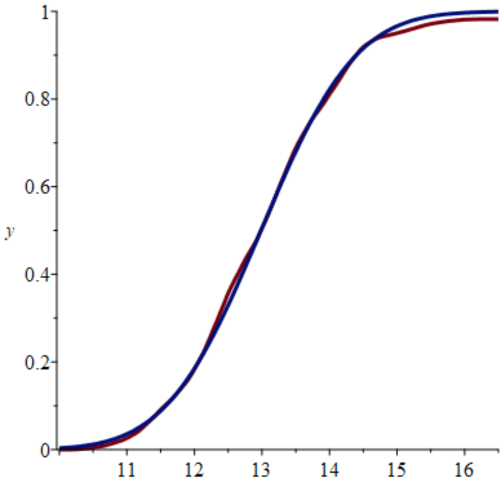


Рис. 14. График функции распределения величины X (данные табл.21).

Оценка медианы $\hat{m} = 12.9458$. Доверительные границы: $u_{0.05} = 11.1761, u_{0.95} = 14.795$.

19. Многомерные *kNN*-оценки функции распределения

Пусть $(\boldsymbol{U}_1, W_1), (\boldsymbol{U}_2, W_2), ..., (\boldsymbol{U}_n, W_n)$ – независимые, одинаково распределенные пары, где $\{\boldsymbol{U}_i\}, 1 \le i \le n -$ d-мерные случайные векторы (мы будем рассматривать случай $d \ge 2$) с ограниченной непрерывной плотностью распределения $f(\boldsymbol{x})$, $W_i = I(\boldsymbol{X}_i < \boldsymbol{U}_i)$ – индикатор события $(\boldsymbol{X}_i < \boldsymbol{U}_i)$, d-мерный вектор $\boldsymbol{X}_i$ имеет функцию распределения $Q(\boldsymbol{x})$ и непрерывную плотность распределения $q(\boldsymbol{x}) > 0$. Требуется по выборке $\mathscr{U}^{(n)} = \{(\boldsymbol{U}_i, W_i),\ \ 1 \le i \le n\}$ оценить неизвестную функцию распределения $Q(\boldsymbol{x})$.

Обычно в качестве оценки для $Q(\boldsymbol{x})$ используют непараметрические оценки. В случае $d = 1$ применяются ядерные оценки $\hat{Q}_n(x) = S_{2n}(x) / S_{1n}(x)$, где $S_{1n}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{u_i - x}{h}\right)$, $S_{2n}(x) = \frac{1}{nh}\sum_{i=1}^{n} W_i K\left(\frac{u_i - x}{h}\right)$, где и $K(x)$– ядерная функция (обычно четная финитная плотность некоторого распределения). Мы полагаем $\hat{F}_n(x) = 0$, если $S_{1n}(x) = 0$. В качестве параметра h берут значение $h = n^{-1/5}$.

Другой используемой оценкой является оценка Янга [1] вида $Q_n^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{i/n - F_n(x)}{h}\right) W_n^{[i]}$, где $F_n(x) = n^{-1}\sum_{i=1}^{n} I(U_i < x) -$ эмпирическая функция распределения величины случайной U, одинаково распределенной с $U_1, U_2, ..., U_n$, $W_n^{[i]} -$ индуцированная порядковая статистика, построенная по выборке $\mathscr{U}^{(n)}$, т.е. если $U_j = U_n^{(i)} -$ i-я порядковая статистика, то $W_n^{[i]} = W_j$.

Асимптотически эквивалентной ей является оценка k-ближайших соседей (kNN-оценка)

$$\tilde{Q}_n(x) = \tilde{S}_{2n}(x) / \tilde{S}_{1n}(x),$$

где $S_{1n}(x) = \frac{1}{n\rho}\sum_{i=1}^{n} K\left(\frac{u_i - x}{\rho}\right)$, $S_{2n}(x) = \frac{1}{n\rho}\sum_{i=1}^{n} W_j K\left(\frac{u_i - x}{\rho}\right)$.

Для заданного x исходные данные $U_1, U_2, ..., U_n$ преобразуем следующим образом. Рассмотрим величины $\xi_i = |x - U_i|$, $i = 1, 2, ..., n$, из которых получим вариационный ряд: $\xi_n^{(1)} \le \xi_n^{(2)} \le ... \le \xi_n^{(n)}$. Положим $\rho = \xi_n^{(k)}$ и для так выбранного ρ найдем значение оценки $\tilde{Q}_n(x)$. Примем $k = [n^{4/5}]$, где $[a]$ есть целая часть числа a.

Оценка $\hat{Q}_n(x)$ имеет предельную (при $n \to \infty$) дисперсию $\sigma_1^2 = Q(x)(1 - Q(x))\|K\|^2 / f(x)$, где $\| K \|^2 = \int K^2(x)\,dx$, поэтому, если плотность распределения $f(x)$ величины U равна нулю в точке x, то в этом случае лучше использовать оценку (2) или оценку (3). В настоящем сообщении мы исследуем поведение многомерных *kNN-оценок функции распределения* $Q(\boldsymbol{x})$. Многомерные *kNN-оценки плотности* были рассмотрены в работах ([24],[25]).

Непараметрические оценки широко используются в эконометрике, в биологической и медицинской статистике, при этом большая часть асимптотических результатов опирается на предположения о гладкости функции распределения. Однако зачастую на практике эти предположения могут не выполняться. Вмешательство в экономику регулирующих органов в форме, например, транфертов или различных ставок процентов и налогов ведет к нарушению гладкости функции распределения. Более реалистичные формы пособий, сложные правила налогообложения приводят к менее очевидным нарушениям гладкости распределений. В биологии риски смертности, например, после инфарктов также приводят к нарушению гладкости распределений. При условии существования функции плотности, возможно, не непрерывной или непрерывной, но с неизвестным количеством производных, этот подход приводит преобладанию асимптотического смещения в оценках или к предельному распределению оценок, которое не является нормальным. Были предложены комбинированные оценки плотности, устраняющие эти недостатки, но такой подход оказался малоэффективным в задачах оценки функции распределения в зависимости «доза-эффект». В рассматриваемой работе для оценки функции распределения мы используем *kNN*-оценки и показываем их состоятельность, асимптотическую нормальность и асимптотическую несмещенность.

Итак, рассмотрим оценку функции распределения $Q(\boldsymbol{x})$ посредством

$$\hat{Q}_n(\boldsymbol{x}) = \frac{T_{2n}(\boldsymbol{x})}{T_{1n}(\boldsymbol{x})},$$

где

$$T_{2n}(\boldsymbol{x}) = \frac{1}{n\rho^d}\sum_{j=1}^{n} W_j \mathscr{K}\left(\frac{\boldsymbol{U}_j - \boldsymbol{x}}{\rho}\right) = \frac{1}{n}\sum_{j=1}^{n} W_j \mathscr{K}_\rho(\boldsymbol{U}_j - \boldsymbol{x}),\ T_{1n}(\boldsymbol{x}) = \frac{1}{n\rho^d}\sum_{j=1}^{n} \mathscr{K}\left(\frac{\boldsymbol{U}_j - \boldsymbol{x}}{\rho}\right) = \frac{1}{n}\sum_{j=1}^{n} \mathscr{K}_\rho(\boldsymbol{U}_j - \boldsymbol{x}),$$

$\rho = \rho_n(\boldsymbol{x}) -$ евклидово расстояние между $\boldsymbol{x}$ и k-м ближайшим соседом среди всех $\boldsymbol{U}_j$, $j = 1, 2, ..., n$, $\mathscr{K}(\boldsymbol{x})$ есть нормированная ограниченная интегрируемая весовая функция

$$\int \mathscr{K}(\boldsymbol{u})\,d\boldsymbol{u} = 1,$$

$k = k(n) -$ последовательность положительных целых такая, что $k \to \infty,\ k/n \to 0$ при $n \to \infty$.

Вспомогательные результаты

Рассмотрим вначале вероятностную плотность $p(\boldsymbol{x})$ расстояния ρ между $\boldsymbol{x}$ и k-м ближайшим соседом $\boldsymbol{x}$. Пусть $S_r = \{\boldsymbol{z}: \| \boldsymbol{z} - \boldsymbol{x} \| < r\},\ G(r) = \mathbf{P}(S_r),$ и

$$G'(r) = \lim_{\delta \to 0} \frac{1}{\delta}\left[\int_{S_{r+\delta}} f(\boldsymbol{t})\,d\boldsymbol{t} - \int_{S_r} f(\boldsymbol{t})\,d\boldsymbol{t}\right] = \int_{\|\boldsymbol{x}-\boldsymbol{t}\|=r} f(\boldsymbol{t})\,d\sigma(\boldsymbol{t}),$$

где $\mathbf{P}$ есть вероятностная мера с плотностью f, а $\Sigma_0 = \sigma/\beta_d -$ равномерное распределение на поверхности сферы радиуса r, $\beta_d = d \cdot c_d -$ площадь поверхности сферы $\| \boldsymbol{x} - \boldsymbol{t} \| = r$, $c_d = \dfrac{\pi^{d/2} r^d}{\Gamma((d+2)/2)} -$ объем шара $B_r = \{\boldsymbol{t}: \| \boldsymbol{x} - \boldsymbol{t} \| < r\}$ радиуса r.

Плотность распределения величины ρ равна

$$p_n(r) = \frac{n!}{(k-1)!(n-k)!} G^{k-1}(r)(1-G(r))^{n-k} G'(r).$$

Рассмотрим теперь условную плотность распределения $(k-1)$ величин $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{k-1}$, расстояние которых до $\boldsymbol{x}$ меньше, чем расстояние до k-го ближайшего сосед, $(n-k)$ величин $\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_{n-k}$, расстояние которых до $\boldsymbol{x}$ больше, чем расстояние до k-го ближайшего соседа и величины $\boldsymbol{h}$, расстояние которого до $\boldsymbol{x}$ равно r (см. [4]). Она равна

$$p(\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}; \boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_{n-k}; \boldsymbol{h} | r) = \prod_{j=1}^{k-1}\left(\frac{f(\boldsymbol{y}_j)}{G(r)}\right)\prod_{l=1}^{n-k}\left(\frac{f(\boldsymbol{v}_l)}{1-G(r)}\right)\frac{f(\boldsymbol{h})}{G'(r)}.$$

Здесь $\{Y_j\}_{1\le j \le k-1}$, $\{V_l\}_{k+1 \le l \le n}$ условно независимы при данном $\rho = r$ соответственно, с маргинальными плотностями $\dfrac{f(\boldsymbol{y}_j)}{G(r)},\ \dfrac{f(\boldsymbol{v}_l)}{1-G(r)}, \dfrac{f(\boldsymbol{h})}{G'(r)}$,

в областях $\{\boldsymbol{t}: \| \boldsymbol{x} - \boldsymbol{t} \| < r\}$, $\{\boldsymbol{t}: \| \boldsymbol{x} - \boldsymbol{t} \| > r\}$, $\{\boldsymbol{t}: \| \boldsymbol{x} - \boldsymbol{t} \| = r\}$.

Отметим, что если f ограничена и непрерывна, то

$$G(r) = \int_{B_r} f(\boldsymbol{u})\,d\boldsymbol{u} = c_d f(\boldsymbol{x}) r^d + \int_{B_r} (f(\boldsymbol{u}) - f(\boldsymbol{x}))\,d\boldsymbol{u} = c_d f(\boldsymbol{x}) r^d + o(r^d)$$

при $r \downarrow 0$. Отсюда следует, что если $t = G(r)$, для которого $f(\boldsymbol{x}) > 0$, то

$$(G^{-1}(t))^{\lambda} = r^{\lambda} = \left(\frac{t}{c_d f(\boldsymbol{x})}\right)^{\lambda/d} + o(t^{\lambda/d}).$$

Более того, $G^{-1}(z) = \left[\dfrac{z}{c_0 f(\boldsymbol{x})}\right]^{1/d} - \left\{\dfrac{c_2 \nabla^2 f(\boldsymbol{x})}{2pc_0 f(\boldsymbol{x})}\right\} \cdot \left\{\dfrac{z}{c_0 f(\boldsymbol{x})}\right\}^{3/d} + o(z^{3/d})$ при условиях

$\int \|\boldsymbol{y}\|^2 \mathscr{K}(\boldsymbol{y})\,d\boldsymbol{y} < \infty, \int \mathscr{K}(\boldsymbol{y})\,d\boldsymbol{y} = 1, \quad \int y_i \mathscr{K}(\boldsymbol{y})\,d\boldsymbol{y} = \int y_i y_j \mathscr{K}(\boldsymbol{y})\,d\boldsymbol{y} = 0, i \ne j, \qquad \int y_i^2 \mathscr{K}(\boldsymbol{y})\,d\boldsymbol{y} > 0$ для любого i.

Основные результаты

Рассмотрим разность $\hat{Q}_n(\boldsymbol{x}) - Q(\boldsymbol{x})$. Имеем:

$$\tau_n(\boldsymbol{x}) = \frac{T_{2n}(\boldsymbol{x})}{T_{1n}(\boldsymbol{x})} - Q(\boldsymbol{x}) = \frac{T_{2n}(\boldsymbol{x})}{T_{1n}(\boldsymbol{x})} - \frac{Q(\boldsymbol{x}) f(\boldsymbol{x})}{f(\boldsymbol{x})} =$$

$$= \frac{T_{2n}(\boldsymbol{x})(f(\boldsymbol{x}) - T_{1n}(\boldsymbol{x})) + T_{1n}(\boldsymbol{x})(T_{2n}(\boldsymbol{x}) - Q(\boldsymbol{x}) f(\boldsymbol{x}))}{T_{1n}(\boldsymbol{x}) f(\boldsymbol{x})} = \frac{T_{2n}(\boldsymbol{x})(f(\boldsymbol{x}) - T_{1n}(\boldsymbol{x}))}{T_{1n}(\boldsymbol{x}) f(\boldsymbol{x})} + \frac{(T_{2n}(\boldsymbol{x}) - Q(\boldsymbol{x}) f(\boldsymbol{x}))}{f(\boldsymbol{x})}.$$

Если мы покажем, что $T_{1n}(\boldsymbol{x}) - f(\boldsymbol{x}) \xrightarrow[n\to\infty]{p} 0$ и $T_{2n}(\boldsymbol{x}) - Q(\boldsymbol{x}) f(\boldsymbol{x}) \xrightarrow[n\to\infty]{p} 0$, то из теоремы Слуцкого, ограниченности $Q(\boldsymbol{x}) f(\boldsymbol{x})$ и отделимости $f(\boldsymbol{x})$ от нуля, получим сходимость $\tau_n(\boldsymbol{x})$ по вероятности к нулю при $n\to\infty$, т.е. $\tau_n(\boldsymbol{x}) \xrightarrow[n\to\infty]{p} 0$ для каждого фиксированного $\boldsymbol{x}$. Кроме того, из этого соотношения мы получим предельное распределение величины $\tau_n(\boldsymbol{x})$.

Рассмотрим $\varphi_{1n}(\theta)$ – характеристическую функцию статистики $T_{1n}(\boldsymbol{x})$. Пусть

$$\varphi_{1n}(\theta) = \mathbf{E}\left(\exp\left(i\theta T_{1n}(\boldsymbol{x})\right)\right).$$

Используя равенство (3), выводим $\varphi_{1n}(\theta) = \int \left(\psi_{1n}(\theta,r)\right)^{k-1} \psi_{2n}(\theta,r) \left(\psi_{3n}(\theta,r)\right)^{n-k} p_n(r)dr,$ где

$$\psi_{1n}(\theta,r) = \int_{\|\boldsymbol{y}-\boldsymbol{x}\|<r} \exp\left(\frac{i\theta}{n} Q(\boldsymbol{y}) \mathscr{K}_r(\boldsymbol{y}-\boldsymbol{x})\right) \frac{f(\boldsymbol{y})}{G(r)} d\boldsymbol{y},\ \ \psi_{2n}(\theta,r) = \int_{\|\boldsymbol{t}-\boldsymbol{x}\|=r} \exp\left(\frac{i\theta}{n} Q(\boldsymbol{t}) \mathscr{K}_r(\boldsymbol{t}-\boldsymbol{x})\right) \frac{f(\boldsymbol{t})}{1-G(r)} d\boldsymbol{t}$$

$$\psi_{3n}(\theta,r) = \int_{\|\boldsymbol{v}-\boldsymbol{x}\|>r} \exp\left(\frac{i\theta}{n} Q(\boldsymbol{v}) \mathscr{K}_r(\boldsymbol{v}-\boldsymbol{x})\right) \frac{f(\boldsymbol{v})}{1-G(r)} d\boldsymbol{v}.$$ Первый член преобразуем к виду

$$\psi_{1n}(\theta,r) = \frac{1}{G(r)} \int_{\|\boldsymbol{u}\|<r} \exp\left(\frac{i\theta}{n} \mathscr{K}_r(\boldsymbol{u})\right) f(\boldsymbol{x}-\boldsymbol{u}r)\, d\boldsymbol{u}.$$

Лемма 5. *Для каждого* $\alpha \in \mathbf{R}^1$ *и* $n \ge 0$, $|e^{i\alpha} - \sum_{k=0}^{n} (i\alpha)^k / k!| \le \min\left\{ |\alpha|^{n+1}/(n+1)!, 2|\alpha|^n/n! \right\}$.

Доказательство. Если $\alpha > 0$, то интегрируя по частям интеграл $\int_0^\alpha (\alpha - t)^n e^{it} dt$ приходим к неравенству $\left| e^{i\alpha} - \sum_{k=0}^{n} \frac{(i\alpha)^k}{k!} \right| \le \frac{|\alpha|^{n+1}}{(n+1)!}$.

Используя неравенство $|e^{it} - 1| \le 2$ и заменяя в предыдущем неравенстве n на $n-1$ получим вторую часть неравенства.

Теперь, в силу условия (4) на ядро $\mathscr{K}(\boldsymbol{x})$, используя, что $\mathscr{K}(\boldsymbol{x}) \le M_1$, $f(\boldsymbol{x}) \le M_2$ применяя результат Леммы 1, заключаем, что

$$\left| \int_{\|\boldsymbol{u}\|<r} \left(\exp\left(\frac{i\theta}{n} \mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right) - 1 - \frac{1}{n}\left(i\theta \mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right) - \frac{1}{2n^2}\left(i\theta \mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right)^2 \right) f(\boldsymbol{u})\, d\boldsymbol{u} \right| \le$$

$$\le \frac{1}{6n^3} \int_{\|\boldsymbol{u}\|<r} |\theta|^3 \left(\mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right)^3 f(\boldsymbol{u})\, d\boldsymbol{u} = \frac{|\theta|^3}{6n^3 r^{2p}} \int_{\|\boldsymbol{u}\|<1} \mathscr{K}^3(\boldsymbol{u}) f(\boldsymbol{x}-\boldsymbol{u}r)\, d\boldsymbol{u} \le$$

$$\le \frac{|\theta|^3 M_1^3 M_2}{6n^3 r^{2p}} = \frac{|\theta|^3 M_1^3 M_2 c_p^2 f^2(\boldsymbol{x})}{6nk^2}.$$

Следовательно,

$$\int_{\|\boldsymbol{u}\|<r} \left(\exp\left(\frac{i\theta}{n} \mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right) - 1 - \frac{1}{n}\left(i\theta \mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right) - \frac{1}{2n^2}\left(i\theta \mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right)^2 \right) f(\boldsymbol{u})\, d\boldsymbol{u} = O(k^{-3}).$$

Далее, $\int_{\|\boldsymbol{t}\|=r}\left|\exp\left(\frac{i\theta}{n}\mathscr{K}_r(\boldsymbol{t})\right)-1\right|\frac{f(\boldsymbol{x}+\boldsymbol{t})}{G'(r)}d\boldsymbol{t}\le\frac{|\theta|}{n}\int_{\|\boldsymbol{u}\|=1}\mathscr{K}(\boldsymbol{u})\frac{f(\boldsymbol{x}+\boldsymbol{u}r)}{G'(r)}d\boldsymbol{u}=O\left(\frac{1}{n}\right),$

поэтому $\psi_{2n}(\theta,r)\to 1$ при $n\to\infty$. Точно так же,

$$\left|\int_{\|\boldsymbol{u}\|>r}\left(\exp\left(\frac{i\theta}{n}\mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right)-1-\frac{1}{n}\left(i\theta\mathscr{K}_r(\boldsymbol{x}-\boldsymbol{u})\right)\right)f(\boldsymbol{u})\,d\boldsymbol{u}\right|\le\frac{\theta^2}{2n^2}\int_{\|\boldsymbol{u}\|>r}\left(\mathscr{K}_r(\boldsymbol{u}-\boldsymbol{x})\right)^2 f(\boldsymbol{u})\,d\boldsymbol{u}=$$

$$=\frac{\theta^2}{2n^2r^p}\int_{\|\boldsymbol{u}\|>1}\left(\mathscr{K}(\boldsymbol{u})\right)^2 f(\boldsymbol{x}-\boldsymbol{u}r)\,d\boldsymbol{u}\le\frac{\theta^2M_1^2M_2}{2n^2r^d}=\frac{\theta^2M_1^2M_2c_d f(\boldsymbol{x})}{2nk}=O\left(\frac{1}{nk}\right)=o\left(\frac{1}{n}\right).$$

Для дальнейшего воспользуемся тем, что при условиях теоремы 1

$$\mathbf{E}(T_{1n}(\boldsymbol{x}))=f(\boldsymbol{x})+\frac{a_d}{(f(\boldsymbol{x}))^{2/d}}\mathscr{P}(f)(\boldsymbol{x})\left(\frac{k}{n}\right)^{2/d}+\frac{c_d f(\boldsymbol{x})}{k}\int_{\|\boldsymbol{u}\|=1}\mathscr{K}(\boldsymbol{u})\,d\Sigma_0+o\left(\left(\frac{k}{n}\right)^{2/p}+\frac{1}{k}\right),$$

где $\mathscr{P}(f)(\boldsymbol{x})=\sum_{i,j}\int u_iu_j\mathscr{K}(\boldsymbol{u})\,d\boldsymbol{u}\,D_iD_j f(\boldsymbol{x}),\quad \mathscr{H}_f(\boldsymbol{x})=(D_iD_j f(\boldsymbol{x}))-$ матрица Гессе,

$\mathbf{D}(T_{1n}(\boldsymbol{x}))=\frac{c_d f^2(\boldsymbol{x})}{k}\int\mathscr{K}^2(\boldsymbol{u})\,d\boldsymbol{u}+o\left(\frac{1}{k}\right)$, $a_d=(\Gamma((d+2)/2))^{2/d}/(2\pi)$, и, соответственно,

$$\frac{1}{n}\int_{\|\boldsymbol{v}\|>r}\mathscr{K}_r^2(\boldsymbol{v}-\boldsymbol{x})\frac{f(\boldsymbol{v})}{1-G(r)}d\boldsymbol{v}=\frac{c_d f^2(\boldsymbol{x})}{k}\int\mathscr{K}^2(\boldsymbol{u})\,d\boldsymbol{u}+o\left(\frac{1}{k}\right).$$

Поэтому равномерно по $\theta\in[-T,T]$, где $T-$ заданное действительное число, $(n\to\infty)$;

$$\int\exp\left(i\theta n^{-1}\mathscr{K}_r(\boldsymbol{u}-\boldsymbol{x})\right)f(\boldsymbol{u})\,d\boldsymbol{u}=1+i\theta f(\boldsymbol{x})k^{-1}-(1/2)\theta^2c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2k^{-1}+o\left(k^{-1}\right),\|\mathscr{K}\|^2=\int\mathscr{K}^2(\boldsymbol{u})\,d\boldsymbol{u}.$$

Разложим $\ln\varphi_{1n}(t)$ в ряд по степеням $i\alpha-\beta=i\theta\frac{f(\boldsymbol{x})}{k}-\frac{\theta^2}{2}\frac{c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2}{k}$ до

второго члена и воспользуемся тем, что $\left|\ln(1+i\alpha)-i\alpha\right|\le\frac{\alpha^2}{2}$ для любого α.

В самом деле, $\left|\ln(1+i\alpha)-i\alpha\right|=\left|i\int_0^\alpha\left(\frac{1}{1+ix}-1\right)dx\right|\le\int_0^\alpha\frac{x}{|1+ix|}dx\le\int_0^\alpha x\,dx=\frac{\alpha^2}{2},$

поскольку $|1+ix|=\sqrt{1+x^2}\ge 1$ и $\frac{1}{|1+ix|}\le 1$. Возьмем $\alpha=\theta\frac{f(\boldsymbol{x})}{k}$ и $\beta=\frac{\theta^2}{2}\frac{c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2}{k}$.

Тогда

$$\ln\left(1+i\alpha-\beta\right)=\ln\left(1-\beta\right)+\ln\left(1+\frac{i\alpha}{1-\beta}\right)=-\beta+\frac{i\alpha}{1-\beta}+O\left(k^{-2}\right)=-\beta+i\alpha+O\left(k^{-2}\right),$$

так как $\beta^2=O\left(k^{-2}\right)$, $\frac{\alpha^2}{(1-\beta)^2}\le\alpha^2=O\left(k^{-2}\right)$, при $|\beta|<\frac{1}{2}$. Отсюда

$\left|\ln\left(1+i\alpha-\beta\right)-i\alpha+\beta\right|=O\left(k^{-2}\right)$. Следовательно, при $n\to\infty$,

$$\left|\ln\left(1+i\theta f(\boldsymbol{x})k^{-1}-(1/2)\theta^2 c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2 k^{-1}\right)-\left(i\theta f(\boldsymbol{x})k^{-1}-(1/2)\theta^2 c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2 k^{-1}\right)\right|=$$

$$=O\left(k^{-2}\right)=o\left(k^{-1}\right),$$

так как для фиксированного $\boldsymbol{x}$ функция $f(\boldsymbol{x})$ ограничена, а $\|\mathscr{K}\|^2<\infty$, и $\theta\in[-T,T]$, то $o(1)$ равномерно сходится к нулю.

Поскольку $b_{1n}\,p_n(b_{1n}r+b_{0n})$, где $b_{1n},b_{0n}-$ подходящие нормирующие множители, сходится равномерно по r на любом ограниченном интервале $\left(-C,C\right)$, $C>0$ к плотности предельного распределения (см. [15]), и, учитывая, что вероятности попадания $b_{1n}^{-1}(\rho-b_{0n})$ на интервалы $\left(-\infty,-C\right]$, $\left[C,\infty\right)$ стремятся к нулю (см.[15])), а функция

$$\exp\left(i\theta f(\boldsymbol{x})k^{-1}-(1/2)\theta^2 c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2 k^{-1}\right)$$

ограничена, получаем, что $\mathbf{E}\left(\exp\left(it\sqrt{k}\left(T_{1n}(\boldsymbol{x})-f(\boldsymbol{x})\right)\right)\right)\to\exp\left(-(1/2)\theta^2 c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2\right)$ при $n\to\infty$. Теперь из сходимости характеристических функций следует, что

$$\sqrt{k}(T_{1n}(\boldsymbol{x})-f(\boldsymbol{x}))\xrightarrow[n\to\infty]{d}N(0,c_d f^2(\boldsymbol{x})\|\mathscr{K}\|^2),$$

откуда также следует, что $T_{1n}(\boldsymbol{x})-f(\boldsymbol{x})\underset{n\to\infty}{\overset{p}{\to}}0$. (5)

Покажем теперь, что $\sqrt{k}(T_{2n}(\boldsymbol{x})-Q(\boldsymbol{x})f(\boldsymbol{x}))\underset{n\to\infty}{\overset{d}{\to}}N(0,c_d Q^2(\boldsymbol{x})f^2(\boldsymbol{x})\|\mathscr{K}\|^2)$. Пусть

$$\varphi_{2n}(\theta)=\mathbf{E}(\exp(i\theta T_{2n}(\boldsymbol{x})))=\mathbf{E}\left(\exp\left(i\theta n^{-1}\sum_{j=1}^{n}I(\boldsymbol{X}_j<\boldsymbol{U}_j)\mathscr{K}_\rho(\boldsymbol{U}_j-\boldsymbol{x})\right)\right).$$

Переходя сначала к условным математическим ожиданиям при условии $\boldsymbol{U}_j$, а затем рассуждая как выше относительно характеристической функции статистики $T_{1n}(\boldsymbol{x})$, получим следующее представление $\varphi_{2n}(\theta)=\int\left(\lambda_{1n}(\theta,r)\right)^{k-1}\lambda_{2n}(\theta,r)\left(\lambda_{3n}(\theta,r)\right)^{n-k}p_n(r)\,dr$, где

$$\lambda_{1n}(\theta,r)=\int_{\|\boldsymbol{y}-\boldsymbol{x}\|<r}\exp\left(\frac{i\theta}{n}Q(\boldsymbol{y})\mathscr{K}_r(\boldsymbol{y}-\boldsymbol{x})\right)\frac{f(\boldsymbol{y})}{G(r)}d\boldsymbol{y},\quad \lambda_{2n}(\theta,r)=\int_{\|\boldsymbol{t}-\boldsymbol{x}\|=r}\exp\left(\frac{i\theta}{n}F(\boldsymbol{t})\mathscr{K}_r(\boldsymbol{t}-\boldsymbol{x})\right)\frac{f(\boldsymbol{t})}{G'(r)}d\boldsymbol{t},$$

$\lambda_{3n}(\theta,r)=\int_{\|\boldsymbol{v}-\boldsymbol{x}\|>r}\exp\left(\frac{i\theta}{n}Q(\boldsymbol{v})\mathscr{K}_r(\boldsymbol{v}-\boldsymbol{x})\right)\frac{f(\boldsymbol{v})}{1-G(r)}d\boldsymbol{v}.$ Повторяя теперь вышеприведенные рассуждения, но относительно функций $\lambda_{1n}(\theta,r)$, $\lambda_{2n}(\theta,r)$, $\lambda_{3n}(\theta,r)$ и $\varphi_{2n}(\theta)$ получаем, что

$$\sqrt{k}(T_{2n}(\boldsymbol{x})-Q(\boldsymbol{x})f(\boldsymbol{x}))\underset{n\to\infty}{\overset{d}{\to}}N(0,c_d Q^2(\boldsymbol{x})f^2(\boldsymbol{x})\|\mathscr{K}\|^2).$$

Таким образом, имеем следующий результат.

Теорема 23. *Пусть плотность* $f(\boldsymbol{x})$ *ограничена и существуют третьи непрерывные ограниченные частные производные* $f(\boldsymbol{x})$ *и* $Q(\boldsymbol{x})$, $\int \|\boldsymbol{u}\|^2 \mathcal{K}(\boldsymbol{u})\,d\boldsymbol{u} < \infty$, $\int \boldsymbol{u}\mathcal{K}(\boldsymbol{u})\,d\boldsymbol{u} = \boldsymbol{0}$.

Тогда

$$\text{(i)} \sqrt{k}\left(T_{1n}(\boldsymbol{x}) - f(\boldsymbol{x})\right) \underset{n\to\infty}{\overset{d}{\to}} N(0, c_d f^2(\boldsymbol{x})\|\mathcal{K}\|^2) \quad \text{(ii)} \sqrt{k}\left(T_{2n}(\boldsymbol{x}) - Q(\boldsymbol{x}) f(\boldsymbol{x})\right) \underset{n\to\infty}{\overset{d}{\to}}$$

$$\to N(0, c_p f^2(\boldsymbol{x}) Q^2(\boldsymbol{x})\|\mathcal{K}\|^2).$$

Следующая теорема устанавливает асимптотическую нормальность оценки $\tilde{F}_n(x)$ функции распределения $Q(\boldsymbol{x})$.

Теорема 24. *Пусть выполнены условия теоремы 1. Тогда*

$$\sqrt{k}\left(\hat{Q}_n(\boldsymbol{x}) - Q(\boldsymbol{x})\right) \underset{n\to\infty}{\overset{d}{\to}} N(0, Q(\boldsymbol{x})(1 - Q(\boldsymbol{x}))\|\mathcal{K}\|^2).$$

Доказательство. Пусть

$T_1 = T_{1n}(\boldsymbol{x})$, $T_2 = T_{2n}(\boldsymbol{x})$, $\quad Qf = Qf(\boldsymbol{x}) = Q(\boldsymbol{x}) f(\boldsymbol{x})$, $f = f(\boldsymbol{x})$. Имеем:

$$\frac{T_2}{T_1} = \frac{T_2 - Qf + Qf}{T_1 - f + f} = \frac{(T_2 - Qf) + Qf}{f\left(1 + \frac{T_1 - f}{f}\right)} = \frac{T_2 - Qf}{f}\left(1 - \frac{T_1 - f}{f} + O_p\left(\frac{(T_1 - f)^2}{f^2}\right)\right) + \frac{Qf}{f}\left(1 - \frac{T_1 - f}{f} + O_p\left(\frac{(T_1 - f)^2}{f^2}\right)\right),$$

так как $\left|\frac{1}{1+x} - 1 + x\right| = \left|\frac{x^2}{1+x}\right| \le 2x^2 = O\left(x^2\right)$, $|x| \le \frac{1}{2}$ и $T_1 - f \underset{n\to\infty}{\overset{p}{\to}} 0$, $T_2 - Qf \underset{n\to\infty}{\overset{p}{\to}} 0$.

Отсюда $\displaystyle \frac{T_2}{T_1} - \frac{Qf}{f} = \frac{T_2 - Qf}{f} - \frac{Qf}{f^2}(T_1 - f) + O_p\left(\frac{(T_2 - Qf)(T_1 - f)}{f^2}\right) + O_p\left(\frac{Qf\,(T_1 - f)^2}{f^3}\right).$

Рассуждая аналогично относительно статистик $T_{1n}(\boldsymbol{x})$ и $T_{2n}(\boldsymbol{x})$ можно показать, что с вероятностью 1

$$\overline{\lim_{n\to\infty}} \sup_{\frac{c\ln n}{n} \le h \le 1} \frac{\sqrt{nh}\left\|T_1 - \mathbf{E}(f)\right\|_\infty}{\sqrt{\max\left(\ln(1/h), \ln\ln n\right)}} = k_1(c) < \infty, \quad \overline{\lim_{n\to\infty}} \sup_{\frac{c\ln n}{n} \le h \le 1} \frac{\sqrt{nh}\left\|T_2 - \mathbf{E}(Qf)\right\|_\infty}{\sqrt{\max\left(\ln(1/h), \ln\ln n\right)}} = k_2(c) < \infty.$$

откуда для достаточно больших n, $\displaystyle \left\|\left(\frac{T_2}{T_1} - F\right) - \frac{T_2 - Qf}{f} + \frac{Qf}{f^2}(T_1 - f)\right\|_\infty \le C_1 \frac{\ln n}{k}.$

Таким образом,

$$\sqrt{k}\left\|\left(\frac{T_2}{T_1} - F\right) - \frac{T_2 - Qf}{f} + \frac{Qf}{f^2}(T_1 - f)\right\|_\infty \underset{n\to\infty}{\overset{p}{\to}} 0.$$

Далее,

$$\mathbf{D}\left(\frac{T_2}{T_1}-Q\right)=\left(\frac{1}{f^2}\mathbf{D}\left(T_2-Qf\right)+\frac{(Qf)^2}{f^4}\mathbf{D}\left(T_1-f\right)-2\frac{Qf}{f^3}\mathbf{Cov}\left(T_1-f,T_2-Qf\right)\right)\left(1+O_p\left(\frac{\ln^2 n}{k^2}\right)\right)=$$

$$=\frac{1}{f^2}\mathbf{D}\left(T_2\right)+\frac{(Qf)^2}{f^4}\mathbf{D}\left(T_1\right)-2\frac{Qf}{f^3}\mathbf{Cov}\left(T_1,T_2\right)\left(1+O_p\left(\frac{\ln^2 n}{k^2}\right)\right)\quad \text{при } n\to\infty.$$

Рассмотрим математическое ожидание $\mathbf{E}\left(T_1\cdot T_2\right)$. Для него

$$\mathbf{E}\left(T_1\cdot T_2\right)=\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathscr{K}_\rho\left(\boldsymbol{U}_i-\boldsymbol{x}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}W_i\mathscr{K}_\rho\left(\boldsymbol{U}_i-\boldsymbol{x}\right)\right)=\frac{1}{n^2}\mathbf{E}\left(\sum_{i=j=1}^{n}W_i\left(\mathscr{K}_\rho\left(\boldsymbol{U}_i-\boldsymbol{x}\right)\right)^2+\sum_{i\neq j}\mathscr{K}_\rho\left(\boldsymbol{U}_i-\boldsymbol{x}\right)W_j\mathscr{K}_\rho\left(\boldsymbol{U}_j-\boldsymbol{x}\right)\right).$$

В силу независимости и одинаковой распределенности пар $\left(\boldsymbol{U}_1,W_1\right),\ldots,\left(\boldsymbol{U}_n,W_n\right)$ заключаем,

$$\mathbf{E}\left(T_1\cdot T_2\right)=\frac{1}{n}\mathbf{E}\left(W_1\left(\mathscr{K}_\rho\left(\boldsymbol{U}_1-\boldsymbol{x}\right)\right)^2\right)+\frac{n-1}{n}\mathbf{E}\left(W_1\mathscr{K}_\rho\left(\boldsymbol{U}_1-\boldsymbol{x}\right)\right)\mathbf{E}\left(\mathscr{K}_\rho\left(\boldsymbol{U}_2-\boldsymbol{x}\right)\right)=$$

$$=\frac{1}{n}\int\mathbf{E}\left(I\left(\boldsymbol{U}_1>\boldsymbol{X}_1\right)\mathscr{K}_\rho^2\left(\boldsymbol{U}_i-\boldsymbol{x}\right)\middle|\boldsymbol{U}_1=\boldsymbol{u}\right)f(\boldsymbol{u})\,d\boldsymbol{u}+$$

$$+\frac{n-1}{n}\int\mathbf{E}\left(I\left(\boldsymbol{X}_1<\boldsymbol{U}_1\right)\mathscr{K}_\rho\left(\boldsymbol{U}_1-\boldsymbol{x}\right)\middle|\boldsymbol{U}_1=\boldsymbol{u}\right)f(\boldsymbol{u})\,d\boldsymbol{u}\times\int\mathscr{K}_\rho(\boldsymbol{u}-\boldsymbol{x})\,f(\boldsymbol{u})\,d\boldsymbol{u}.$$

Произведя замену $\boldsymbol{z}=(\boldsymbol{u}-\boldsymbol{x})r^{-1}$, получим

$$\mathbf{E}(T_1\cdot T_2)=(nr^d)^{-1}\int\mathscr{K}^2\left(\boldsymbol{u}-\boldsymbol{x}\right)Q(\boldsymbol{z}r+\boldsymbol{x})\times f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}+(r^{2d})^{-1}(1-n^{-1})\times$$

$$\times\left(\int\mathscr{K}\left(\boldsymbol{z}\right)Q(\boldsymbol{z}r+\boldsymbol{x})f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}\right)\times\left(\int\mathscr{K}\left(\boldsymbol{z}\right)f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}\right)$$

и $\quad (nr^d)^{-1}\int\mathscr{K}^2\left(\boldsymbol{u}-\boldsymbol{x}\right)Q(\boldsymbol{z}r+\boldsymbol{x})f(\boldsymbol{z}r+\boldsymbol{x})d\boldsymbol{z}=k^{-1}Q(\boldsymbol{x})f^2(\boldsymbol{x})\|\mathscr{K}\|^2+o(k^{-1}).$

Из условий на ядро $\mathscr{K}(\boldsymbol{x})$ и условий на функции $Q(\boldsymbol{x}),\ f(\boldsymbol{x})$, имеем

$\int\mathscr{K}\left(\boldsymbol{z}\right)Q(\boldsymbol{z}r+\boldsymbol{x})f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}=Q(\boldsymbol{x})f(\boldsymbol{x})+o(k^{-1}).$ Таким образом,

$$\mathbf{E}(T_1\cdot T_2)=Q(\boldsymbol{x})f^2(\boldsymbol{x})\|\mathscr{K}\|^2k^{-1}+(1-n^{-1})\left(Q(\boldsymbol{x})f(\boldsymbol{x})+f(\boldsymbol{x})\right)+o(k^{-2}),\mathbf{Cov}(T_1,T_2)=Q(\boldsymbol{x})g^2(\boldsymbol{x})\|\mathscr{K}\|^2k^{-1}+O(k^{-2}).$$

Значит, при $n\to\infty$, $\mathbf{D}(\hat{Q}_n(\boldsymbol{x})-Q(\boldsymbol{x}))=Q(\boldsymbol{x})(1-Q(\boldsymbol{x}))\|\mathscr{K}\|^2k^{-1}(1+o(1))$.

Отсюда заключаем, что $\sqrt{k}(\tilde{Q}_n(\boldsymbol{x})-Q(\boldsymbol{x}))\xrightarrow[n\to\infty]{d}\varsigma\in N(0,Q(\boldsymbol{x})\left(1-Q(\boldsymbol{x})\right)\|\mathscr{K}\|^2).$

Таким образом, $kNN-$оценка является асимптотически нормальной.

Кроме того, имеем:

$$\mathbf{E}(\hat{Q}_n(\boldsymbol{x})-Q(\boldsymbol{x}))=$$
$$a_d(k/n)^{2/d}\ /\ f(\boldsymbol{x})(Q(\boldsymbol{x})\mathscr{P}_f(\boldsymbol{x})-\mathscr{P}_{Qf}(\boldsymbol{x}))-c_d(1-Q(\boldsymbol{x}))k^{-1}\int_{\|\boldsymbol{x}\|=1}K(\boldsymbol{x})\,d\Sigma_0+o(k^{-1}).$$

20. Фурье-метод оценивания функции распределения

Пусть с.в. $Y = X + \varepsilon$, где случайная величина имеет плотность распределения $f_X(x)$, а случайная величина ε имеет функцию распределения $F_\varepsilon(x)$. Тогда случайная величина Y имеет плотность распределения $g_Y(y) = \int\limits_{-\infty}^{\infty} g_U(y-x)\,dF_\varepsilon(x)$, а для соответствующих характеристических функций имеет место равенство: $\varphi_Y(t) = \varphi_U(t)\varphi_\varepsilon(t)$, откуда $\varphi_U(t) = \dfrac{\varphi_Y(t)}{\varphi_\varepsilon(t)}$.

Пусть $\hat{\varphi}_n(t) = \hat{\varphi}_n(t) = \dfrac{1}{n}\sum\limits_{j=1}^{n} e^{itY_j}$ есть оценка (эмпирическая характеристическая функция). Поэтому $\hat{\varphi}_{n,X}(t) = \dfrac{\hat{\varphi}_n(t)}{\varphi_\varepsilon(t)}$. Например, если $\varepsilon \in N(0,\sigma_0^2)$ (σ_0^2 известна), то $\varphi_\varepsilon(t) = \exp\left(-\dfrac{t^2\sigma_0^2}{2}\right)$.

Рассмотрим оценку плотности $g(x)$ в виде

$$g_n(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y_j - y}{h}\right) = \frac{1}{h}\int\limits_{-\infty}^{\infty} K\left(\frac{z-y}{h}\right)dG_n(z) \underset{n\to\infty}{\overset{p}{\to}} g_Y(y).$$

Рассмотрим теперь $\mathscr{F}$ – Фурье-преобразование плотности $\dfrac{1}{h}K\left(\dfrac{x}{h}\right)$. Имеем:

$$\mathscr{F}\left(\frac{1}{h}K\left(\frac{x}{h}\right)\right)(t) = \int\limits_{-\infty}^{\infty} e^{itu}\frac{1}{h}K\left(\frac{u}{h}\right)du = \int\limits_{-\infty}^{\infty} e^{itzh}K(z)\,dz = \mathscr{F}(K)(th) = \varphi_K(th).$$

Далее, свертка $\dfrac{1}{h}K\left(\dfrac{x}{h}\right)$ и $g_n(y)$ с помощью характеристических функций равна

$$\psi(t) = \varphi_K(th)\cdot\frac{\hat{\varphi}_n(t)}{\varphi_\varepsilon(t)}.$$

Обратное преобразование дает

$$f_n(x) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} e^{-itx}\varphi_K(th)\cdot\frac{\hat{\varphi}_n(t)}{\varphi_\varepsilon(t)}\,dt.$$

Рассмотрим ядро

$$K_n(x) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} e^{-itx}\cdot\frac{\varphi_K(t)}{\varphi_\varepsilon(t/h)}\,dt = \mathscr{F}^{-1}\left(\frac{\varphi_K(t)}{\varphi_\varepsilon(t/h)}\right),\ \ \mathscr{F}(K_n)(t) = \frac{\varphi_K(t)}{\varphi_\varepsilon(t/h)}.$$

В таком случае, в формуле на самом деле стоит $f_n(x) = \mathscr{F}^{-1}\left(\dfrac{\varphi_K(th)}{\varphi_\varepsilon(t)}\cdot\hat{\varphi}_n(t)\right)$. Иными словами,

$$S_{1n}(x) = f_n(x) = \mathscr{F}^{-1}(\mathscr{F}(K_n))(th)\cdot\mathscr{F}(g)(t)) = \frac{1}{nh}\sum_{j=1}^{n} K_n\left(\frac{u_j - x}{h}\right) \underset{n\to\infty}{\overset{p}{\to}} \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{u-x}{h}\right)dG(u)$$

Хорошо известно, что если $\varphi(t)=\mathscr{F}(f)(t)$, то $f(x)=\mathscr{F}^{-1}(\varphi(t))(x)=\mathscr{F}(\varphi(-t))(x)$, поэтому для наших целей достаточно иметь прямое преобразование Фурье.

Аналогично, для получения значения статистики вида $\frac{1}{nh}\sum_{j=1}^{n}W_jK\left(\frac{y_j-x}{h}\right)$ достаточно совершить следующее преобразование $S_{2n}(x)=\mathscr{F}^{-1}\left(\frac{1}{nh}\sum_{j=1}^{n}w_jK_n\left(\frac{y_j-x}{h}\right)\right)$. В таком случае оценка для функции распределения $F_n(x)=\frac{S_{2n}(x)}{S_{1n}(x)}$.

21. Полукруговой закон Вигнера

В 1955 г. Вигнер в связи с рассмотрением математических моделей энергетических уровней тяжелых атомных ядер доказал теорему, которая была названа «полукруговым законом»: *если элементы* $X_{ij}, i\ge j$, *случайных симметричных матриц*

$A_N=\left(a_{ij}\right)_{i,j=1}^{N}$, $a_{ij}=\frac{X_{ij}}{\sqrt{N}}$ *независимы, симметричны,* $\mathbf{D}(X_{ij})=1, i,j=1,2,...,$ *и для всякого натурального* k $\mathbf{E}(X_{ij}^k)\le B_k<\infty$, *то* $\frac{1}{N}\sum_{j=1}^{N}I(\lambda_k<x)\xrightarrow[N\to\infty]{d}G(x)$,

где $G(x)=\frac{1}{2}+\frac{x\sqrt{4-x^2}}{4\pi}+\frac{1}{\pi}\arcsin\left(\frac{x}{2}\right)$, $G'(x)=g(x)=\frac{1}{2\pi}\sqrt{4-x^2}\,I(-2\le x\le 2)$,

$\lambda_1\le\lambda_2\le...\le\lambda_N-$ *собственные числа матрицы* A_N.

Вернемся к полукруговому закону Вигнера с плотностью $g(x)$ и рассмотрим момент m_k порядка k. В силу четности функции $g(x)$ эти моменты равны нулю и поэтому рассмотрим

$$m_{2k}=\frac{1}{2\pi}\int_{-2}^{2}x^{2k}\sqrt{4-x^2}\,dx.$$

Если сделаем замену $x=2\sin\varphi, dx=2\cos\varphi d\varphi,$ то

$$m_{2k}=\frac{1}{2\pi}2^{2k+2}\int_{-\pi/2}^{\pi/2}\sin^{2k}\varphi\cdot\cos^2\varphi d\varphi=\frac{2^{2k+1}}{\pi}\left(\int_{-\pi/2}^{\pi/2}\sin^{2k}\varphi d\varphi-\int_{-\pi/2}^{\pi/2}\sin^{2k+2}\varphi d\varphi\right).$$

Воспользуемся следующей формулой

$$\int\sin^p x\cdot\cos^q x\,dx=-\frac{\sin^{p-1}x\cdot\cos^{q+1}x}{p+q}+\frac{p-1}{p+q}\int\sin^{p-2}x\cdot\cos^q x\,dx.$$

Так как $\cos\left(\pm\frac{\pi}{2}\right)=0$, то

$$\int_{-\pi/2}^{\pi/2}\sin^{2k}\varphi d\varphi-\int_{-\pi/2}^{\pi/2}\sin^{2k+2}\varphi d\varphi=\frac{2k-1}{2k}\int_{-\pi/2}^{\pi/2}\sin^{2k-2}\varphi d\varphi-\frac{2k+1}{2k+2}\int_{-\pi/2}^{\pi/2}\sin^{2k}\varphi d\varphi=$$

$$=\frac{2k-1}{2k}\cdot\frac{2k-3}{2k-2}\cdot...\cdot\frac{3}{4}\cdot\frac{1}{2}\cdot\pi-\frac{2k+1}{2k+2}\cdot\frac{2k-1}{2k}\cdot...\cdot\frac{3}{4}\cdot\frac{1}{2}\cdot\pi=\frac{1}{2k+2}\cdot\frac{2k-1}{2k}\cdot\frac{2k-3}{2k-2}\cdot...\cdot\frac{3}{4}\cdot\frac{1}{2}\cdot\pi.$$

В частном случае $(k=1)$ имеем

$$m_2=\frac{2^3}{\pi}\cdot\frac{1}{4}\cdot\frac{1}{2}\cdot\pi=1, m_4=2.$$

Из предыдущего следует: $m_{2k+2} = \dfrac{4k+2}{k+2} m_{2k}$.

Обозначая $m_{2k} = c_k$ получаем, что $c_{k+1} = \dfrac{4k+2}{k+2} c_k$, $c_1 = 1$.

Таким образом, имеем, что $\{c_k\}$ есть последовательность чисел Каталана и поэтому производящая функция имеет указанный вид. Далее рассматривается $\dfrac{1}{N}\sum_{j=1}^{N} \lambda_j^{2k}$ и показывается, что эта сумма представляется в следующем виде:

$$\frac{1}{N}\sum_{i=1}^{N} \lambda_i^{2k} = (1+O(N^{-1}))\sum_{j=0}^{k-1}\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^{2j}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^{2(k-j-1)}\right).$$

и поскольку моменты сходятся и при $k=0$ имеем в левой части 1, то это приводит к соотношению как для чисел Каталана и поэтому в пределе получаем полукруговой закон Вигнера.

22. Современные методы построения оценок распределений и исследование их асимптотического поведения

22.1 Ассоциированные случайные последовательности

В данном параграфе мы будем рассматривать некоторое общее понятие зависимости, куда, в частном случае, будут входить и различные понятия перемешивания случайных величин, поэтому начнем с независимых величин.

Независимые случайные величины играют важную роль в математической статистике. Пусть даны две случайных величины X и Y. Они называются *независимыми*, если

$$\boldsymbol{P}((X \in A)\bigcap(Y \in B)) = \boldsymbol{P}(X \in A)\boldsymbol{P}(Y \in B), \quad \text{для любых событий } A \text{ и } B. \qquad (21)$$

Можно показать, что соотношение (21) равносильно следующему:

$$\boldsymbol{E}(f(X)g(Y)) = \boldsymbol{E}(f(X))\boldsymbol{E}(g(Y)) \quad \Leftrightarrow \quad \mathbf{Cov}(f(X), g(Y)) = 0$$

для любых борелевских ограниченных функций $f(x)$ и $g(x)$, заданных на $\mathbf{R}$. пусть конечное или бесконечное $T-$ множество индексов (можно считать, что $T = \{1,2,...,m\}$ или $T \equiv N$) и $I \subset T$. Положим $X_I = \{X_j\}_{j\in I}$.

Определение 1. Семейство X_T называется *ассоциированным* (сокращенно **A**), если для каждого конечного множества $I \subset T$ и любых функций f, g выполнено неравенство

$$\mathbf{Cov}(f(X_I), g(X_I)) \ge 0. \qquad (22)$$

В случае конечного T достаточно проверять соотношение (22) только для $I = T$.

Определение 2. Семейство X_T называется *слабо ассоциированным* или *положительно ассоциированным* (сокращенно **PA**), если для произвольных конечных <u>непересекающихся</u> подмножеств $I, J \subset T$ и любых функций f, g выполнено неравенство

$$\mathbf{Cov}(f(X_I), g(X_J)) \ge 0.$$

Определение 3. Семейство X_T называется *слабо ассоциированным* или *положительно ассоциированным* (сокращенно **NA**), если для произвольных конечных <u>непересекающихся</u> подмножеств $I, J \subset T$ и любых функций f, g выполнено неравенство

$$\mathbf{Cov}(f(X_I), g(X_J)) \le 0.$$

Теорема 25. (Хошневисан, Льюис) *Пусть функция* $h:\mathbf{R}^2 \to \mathbf{R}$ *имеет непрерывную смешанную производную второго порядка* $\partial^2 h/\partial x \partial y$. *Предположим, что случайные величины* X, Y, Z *таковы, что* $Y \overset{d}{=} Z$, *а* Z *не зависит от* X *и* Y. *Тогда*

$$\boldsymbol{E}(h(X,Y)-\boldsymbol{E}(h(X,Z)=\iint_{\mathbf{R}^2}\frac{\partial^2 h(x,y)}{\partial x\partial y}C(x,y)\,dxdy, \qquad (23)$$

где $C(x,y)=\boldsymbol{P}(X\ge x, Y\ge y)-\boldsymbol{P}(X\ge x)\boldsymbol{P}(Y\ge y)$, *при условии, что математические ожидания и интеграл в* (23) *существуют.*

Доказательство. Пусть $U-$ случайная величина такая, что вектор (U,Z) не зависит от (X,Y) и одинаково с ним распределен. Тогда

$$\iint_{\mathbf{R}^2}\frac{\partial^2 h(x,y)}{\partial x\partial y}C(x,y)\,dxdy=\iint_{\mathbf{R}^2}\frac{\partial^2 h(t,w)}{\partial t\partial w}\mathbf{Cov}(I(X\ge t),I(Y\ge w))\,dtdw=$$

$$=\frac{1}{2}\iint_{\mathbf{R}^2}\frac{\partial^2 h(t,w)}{\partial t\partial w}\boldsymbol{E}((I(X\ge t)-I(U\ge t))(I(Y\ge t)-I(Z\ge t)))\,dtdw=$$

$$=\frac{1}{2}\boldsymbol{E}\left(\iint_{\mathbf{R}^2}\frac{\partial^2 h(t,w)}{\partial t\partial w}((I(X\ge t)-I(U\ge t))(I(Y\ge t)-I(Z\ge t)))\,dtdw\right)=$$

$$=\frac{1}{2}\boldsymbol{E}\left(\int_U^X\int_Z^Y\frac{\partial^2 h(t,w)}{\partial t\partial w}dtdw\right)=\frac{1}{2}\boldsymbol{E}\left(\int_U^X\left(\frac{\partial h(t,Y)}{\partial t}-\frac{\partial h(t,Z)}{\partial t}\right)dt\right)=$$

$$=\frac{1}{2}\boldsymbol{E}(h(X,Y)-h(U,Y)-h(X,Z)+h(U,z))=\boldsymbol{E}(h(X,Y))-\boldsymbol{E}(h(X,Z)).$$

Теорема 26. (Ньюмен) *Пусть функции* $f,g:\mathbf{R}\to\mathbf{R}$ *имеют непрерывные производные. Предположим, что случайные величины* X,Y *таковы, что существуют математические ожидания* $\boldsymbol{E}(|f(X)|)$, $\boldsymbol{E}(|g(Y)|)$, $\boldsymbol{E}(|f(X)g(Y)|)$ *существуют. Тогда*

(а) *справедливо равенство*

$$\mathbf{Cov}(f(X),g(Y))=\iint_{\mathbf{R}^2}f'(x)g'(y)C(x,y)\,dxdy, \qquad (24)$$

если интеграл в правой части существует.

(б) *если случайный вектор* (X,Y) *обладает свойством* **PA** *или* **NA** *и функции* f *и* g *неубывающие, то интеграл в правой части* (6) *существует.*

Следствие. (Хёфдинг) Имеем: $\mathbf{Cov}(X,Y)=\iint_{\mathbf{R}^2}(I(X\ge x, Y\ge y)-I(X\ge x)I(Y\ge y))\,dxdy,$

Доказательство. (а) Утверждение теоремы 26 следует из теоремы 25, если взять $h(x,y)=f(x)g(y)$.

(б) Если $(X,Y)\in\mathbf{PA}$ (соответственно **NA**), то выражение под знаком интеграла (24) неотрицательно, если $f'(x)\ge 0$ и $g'(x)\ge 0$. Действительно, если $C(x,y)$ где-то < 0 (на множестве меры > 0), то рассматривая индикатор этого множества (для производных), получим противоречие с **PA**.

Замечание. Если $(X,Y)\in\mathbf{PA}$, то из соотношения (24) следует

$$|\mathbf{Cov}(f(X),g(Y))|\le\sup_x|f'(x)|\sup_y|g'(y)|\,\mathbf{Cov}(X,Y).$$

Лемма 6. *Пусть* $\{X_n, n\ge 1\}-$ *стационарная последовательность ассоциированных случайных величин. Пусть* $S_{m_n}=\sum_{j=1}^{m_n}f_n(X_j)$, *где* f_n *есть дифференцируемые функции с условием* $\sup_n\sup_x|f_n'(x)|\le c<\infty$. *Кроме того, предположим, что*

$$\sum_{j=1}^{\infty}\mathbf{Cov}(X_1,X_j)\le c<\infty.$$

Тогда

$$\boldsymbol{D}(S_{m_n})\le 2cm_n.$$

Доказательство. Заметим, что

$$\boldsymbol{D}(S_{m_n}) = \boldsymbol{D}\left(\sum_{j=1}^{m_n} f_n(X_j)\right) = m_n \boldsymbol{D}(f_n(X_1)) + 2\sum_{1\le i<j\le m_n} \mathbf{Cov}(f_n(X_i), f_n(X_j)) \le 2m_n \sum_{j=1}^{m_n} \mathbf{Cov}(f_n(X_1), f_n(X_j)),$$

и, следовательно,

$$\boldsymbol{D}(S_{m_n}) \le 2m_n c_1 \sum_{j=1}^{m_n} \mathbf{Cov}(X_1, X_j) \le 2m_n c_2.$$

22.2 Примеры

Теорема 27. *Справедливы следующие утверждения.*

(а) *Семейство из одной случайной величины ассоциировано.*

(б) *Семейство, образованное объединением независимых друг от друга совокупностей ассоциированных* (*соответственно* **PA, NA**) *случайных величин, ассоциировано* (*соответственно* **PA, NA**).

Пример 26. Пусть $\boldsymbol{P}(X=-1, Y=1) = \boldsymbol{P}(X=-1, Y=-1) = \frac{1}{4}$, $\boldsymbol{P}(X=1, Y=0) = \frac{1}{2}$.

Тогда для $f(x)=x,\ g(y)=y$ имеем: $\mathbf{Cov}(f(X), g(Y)) = \mathbf{Cov}(X,Y) = 0.$

Если $f(-1)=0,\ f(1)=1, g(-1)=g(0)=0, g(1)=1$ имеем: $\mathbf{Cov}(f(X), g(Y)) = -1/8.$

Если $f(-1)=0,\ f(1)=1, g(-1)=0, g(0)=1, g(1)=1$ имеем: $\mathbf{Cov}(f(X), g(Y)) = 1/8.$

Таким образом, (X,Y) не является ни **PA**, ни **NA**.

Пример 27. Пусть $(X,Y)-$ такой случайный вектор, что для $j=0,1,2$ имеем $\boldsymbol{P}(X=i, Y=j) = p_{ij}$, где p_{ij} заданы матрицей

$$\begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 15/64 & 0 & 1/8 \\ 0 & 9/32 & 0 \\ 1/8 & 0 & 15/64 \end{pmatrix}.$$

Тогда $(X,Y) \in \mathbf{PA}$. Действительно, для этого достаточно проверить, чтобы $\mathbf{Cov}(I(X \ge x), I(Y \ge y)) \ge 0$ при всех $x, y \in \mathbf{R}$. Последнее неравенство можно проверять лишь для $x, y \in \{1,2\}$. Например, $x=1, y=1.$ Тогда мы имеем:

$$\mathbf{Cov}(I(X \ge 1), I(Y \ge 1)) = \boldsymbol{P}(X \ge 1, Y \ge 1)) - \boldsymbol{P}(X \ge 1)\boldsymbol{P}(Y \ge 1) = \frac{33}{64} - \frac{41^2}{64^2} = \frac{431}{64^2} \text{ и т.д.}$$

Теорема 28. *Пусть* $\{X_n, n \ge 1\} -$ *стационарная последовательность ассоциированных случайных величин. Пусть* $S_{k,n} = \sum_{j=1}^{k} f_n(X_j),$ *где* f_n *есть дифференцируемые функции с условием* $\sup_n \sup_x |f_n'(x)| \le c < \infty.$ *Кроме того, предположим, что* $\boldsymbol{E}(f_n(X_1)) = 0,\ \boldsymbol{D}(f_n(X_1)) < \infty$ *и условие* (7) *выполнено. Тогда* $\boldsymbol{P}\left(\frac{S_{n,n}}{n} \underset{n\to\infty}{\to} 0\right) = 1.$

Доказательство. Утверждение теоремы есть сходимость почти наверное: $\frac{S_{n,n}}{n} \xrightarrow[n\to\infty]{п.н.} 0.$

Рассмотрим, что означает сходимость п.н.: $Z_n \xrightarrow[n\to\infty]{п.н.} 0$. Более удобно рассматривать противоположный случай: Z_n не сходится к 0. Это означает, что существует такое натуральное число m, что каждому натуральному числу n соответствует натуральное число v, для которого $|Z_{n+v}| \ge \frac{1}{m}$. Так как слово «существует» соответствует «$\bigcup$», а каждый – знаку «$\bigcap$», то

$$(Z_n \not\to 0) = \bigcup_{m=1}^{\infty}\bigcap_{n=1}^{\infty}\bigcup_{v=1}^{\infty}\left(|Z_{n+v}| \geq \frac{1}{m}\right).$$

Лемма 7 (Борель, Кантелли). *Если для каждого натурального* m,

$$\sum_{n=1}^{\infty} P\left(|Z_n| \geq \frac{1}{m}\right) < \infty, \quad \text{то } \boldsymbol{P}(Z_n \not\to 0) = 0.$$

Доказательство. Положим $A_{nm} = \bigcup_{v=1}^{\infty}\left(|X_{n+v}| \geq \frac{1}{m}\right)$, $A_m = \bigcap_{n=1}^{\infty} A_{nm}$.

Из полуаддитивности вероятности и условия леммы

$$\boldsymbol{P}(A_{nm}) = \boldsymbol{P}\left(\bigcup_{k=n+1}^{\infty}\left(|Z_k| \geq \frac{1}{m}\right)\right) \leq \sum_{k=n+1}^{\infty}\boldsymbol{P}\left(|Z_k| \geq \frac{1}{m}\right) \to 0 \text{ при } n \to \infty.$$

Заметим, что каково бы ни было n_0, $\boldsymbol{P}(A_m) = \boldsymbol{P}\left(\bigcap_{n=1}^{\infty}(A_{nm})\right) \leq \boldsymbol{P}(A_{n_0 m})$, поэтому, полагая $n_0 \to \infty$, получаем $\boldsymbol{P}(A_m) = 0$. Воспользовавшись снова полуаддитивностью, получаем

$$\boldsymbol{P}(Z_n \not\to 0) = \boldsymbol{P}\left(\bigcup_{m=1}^{\infty} A_m\right) \leq \sum_{m=1}^{\infty}\boldsymbol{P}(A_m) = 0 \Rightarrow \boldsymbol{P}(Z_n \not\to 0) = 0.$$

Используя неравенство Чебышева, заметим, что

$$\sum_{j=1}^{k^2}\boldsymbol{P}(|S_{k^2,n}| > k^2/m) \leq \frac{m^2\boldsymbol{D}(S_{k^2,n})}{k^4} \leq \frac{2cm^2k^2}{k^4} = \frac{c_1}{k^2} \Rightarrow \sum_{k=1}^{\infty}\frac{c_1}{k^2} < \infty.$$

Следовательно, по лемме Бореля-Кантелли, $\dfrac{S_{k^2,n}}{k^2} \xrightarrow[n\to\infty]{п.н.} 0.$

Каждому n можно поставить в соответствие такое натуральное число $k = k(n)$, что $k^2 \leq n < (k+1)^2$, откуда $0 \leq n - k^2 \leq 2k$ и $n \to \infty$ влечет $k \to \infty$.

Покажем, что $\dfrac{n-k^2}{n} \leq \dfrac{2}{k}$. Действительно, $n - k^2 \leq 2k \Rightarrow \dfrac{n-k^2}{n} \leq 2\dfrac{k}{n} \leq \dfrac{2}{k}$.

Положим $V_k = \max\limits_{k^2 < N \leq (k+1)^2}\left|\sum\limits_{j=k^2+1}^{N} f_n(X_j)\right|$. Тогда

$$\left|\frac{S_{n,n}}{n} - \frac{S_{k^2,n}}{k^2}\right| = \left|\left(\frac{1}{n} - \frac{1}{k^2}\right)\sum_{j=1}^{k^2} f_n(X_j) + \frac{1}{n}\sum_{j=k^2-1}^{n} f_n(X_j)\right| \leq \frac{(n-k^2)}{n}\cdot\left|\frac{S_{k^2,n}}{k^2}\right| + \left|\frac{1}{n}\sum_{j=k^2-1}^{n} f_n(X_j)\right| \leq$$

$$\leq \frac{2}{k}\cdot\left|\frac{S_{k^2,n}}{k^2}\right| + \frac{1}{n}\max_{k^2<N\leq(k+1)^2}\left|\sum_{j=k^2+1}^{N} f_n(X_j)\right| \leq \frac{2}{k}\cdot\left|\frac{S_{k^2,n}}{k^2}\right| + \frac{V_k}{k^2}.$$

Рассмотрим $\boldsymbol{E}(V_k^2)$. Имеем:

$$\boldsymbol{E}(V_k^2) = \boldsymbol{E}\left(\left(\max_{k^2<N\leq(k+1)^2}\left|\sum_{j=k^2+1}^{N} f_n(X_j)\right|\right)^2\right) = \boldsymbol{E}\left(\max_{k^2<N\leq(k+1)^2}\left(\sum_{j=k^2+1}^{N} f_n(X_j)\right)^2\right) \leq$$

$$\leq \boldsymbol{E}(f_n^2(X_{k^2+1})) + \boldsymbol{E}((f_n(X_{k^2+1}) + f_n(X_{k^2+2}))^2) + \ldots + +\boldsymbol{E}((f_n(X_{k^2+1}) + f_n(X_{k^2+2}) + \ldots + f_n(X_{(k+1)^2}))^2) \leq$$

$$\leq 2k\left(\sum_{j=k^2+1}^{(k+1)^2}\boldsymbol{E}(f_n^2(X_j)) + 2\sum_{\substack{k^2+1\leq i,j\leq(k+1)^2\\ i\neq j}}\boldsymbol{E}(f_n(X_i)f(X_j))\right) = 2k\boldsymbol{D}\left(S_{(k+1)^2,n} - S_{k^2,n}\right) \leq ck^2.$$

Кроме того, $\sum_k \boldsymbol{P}(V_k \geq k^2\varepsilon) < \infty$ для любого $\varepsilon > 0$, откуда по лемме Бореля-Кантелли,

$$\frac{V_k}{k^2} \underset{n\to\infty}{\overset{п.н.}{\to}} 0.$$

В итоге получаем, что $\frac{S_{n,n}}{n} \underset{n\to\infty}{\overset{п.н.}{\to}} 0.$

Рассмотрим следующие ядерные статистики по наблюдениям $(U_n)_{n\geq 1}$ – стационарная последовательность ассоциативных случайных величин, $g(u)$ – плотность распределения с.в. U_n. Пусть

$$g_n(x) = \frac{1}{nh_n}\sum_{j=1}^{n} K\left(\frac{x-U_j}{h_n}\right), \quad x \in J = [a,b].$$

Предположения (A)

(**A1**) $K(x)$ есть ограниченная четная плотность, заданная на интервале $[-1,1]$.

Заметим, что $K(x)$ будет удовлетворять условиям: (i) $\lim_{|u|\to\infty} |u| K(x) = 0,$

(ii) $\int_{-\infty}^{\infty} u^2 K(u)\,du < \infty$.

(**A2**) $K(x)$ дифференцируема и $\sup_x |K'(x)| \leq c < \infty.$

(**B**) Для любых l и $r \geq 0, \sum_{j:|l-j|\geq r} \mathbf{Cov}(U_j, U_l) \leq u(r),$ где $u(r) = e^{-\alpha r}$, для некоторого $\alpha > 0$.

Предположим, что $g(u)$ имеет четвертую ограниченную производную. Пусть $h_n = n^{-1/5}$. Тогда

$$\boldsymbol{E}(g_n(x)) = g(x) + \frac{h_n^2}{2} g''(x)\mu_2 + O(h_n^4), \text{ где } \mu_j = \int_{-\infty}^{\infty} x^j K(x)\,dx,\ j = 1,2,3,4, \text{ причем } \mu_1 = \mu_3 = 0.$$

Кроме того,

$$\boldsymbol{D}(g_n(x)) = \frac{1}{nh_n^2}\boldsymbol{D}\left(K\left(\frac{x-U_1}{h_n}\right)\right) + \frac{1}{n^2h_n^2}\sum_{1\leq i\neq j\leq n} \mathbf{Cov}\left(K\left(\frac{x-U_i}{h_n}\right), K\left(\frac{x-U_j}{h_n}\right)\right).$$

Аналогично случаю независимых величин,

$$\frac{1}{nh_n^2}\boldsymbol{D}\left(K\left(\frac{x-U_1}{h_n}\right)\right) = \frac{1}{nh_n}(g(x)\|K\|^2 + g''(x)h_n^2\beta_2(1+o(1))),$$

где $\|K\|^2 = \int_{-\infty}^{\infty} K^2(x)\,dx,\ \beta_2 = \int_{-\infty}^{\infty} x^2K^2(x)\,dx,$ и (здесь $\psi_n(x,y) = \frac{1}{h_n}K\left(\frac{x-y}{h_n}\right)$)

$$\frac{2}{n^2}\sum_{1\leq i<j\leq n}\left|\mathbf{Cov}\left(\frac{1}{h_n}K\left(\frac{x-U_i}{h_n}\right), \frac{1}{h_n}K\left(\frac{x-U_j}{h_n}\right)\right)\right| \leq \frac{2}{n^2}\sup_y\left\{\frac{\partial}{\partial y}\psi_n(x,y)\right\}^2 \sum_{1\leq i<j\leq n}\mathbf{Cov}(U_i,U_j) \leq$$

$$\leq \frac{c}{n^2h_n^4}\sum_{1\leq i<j\leq n}\mathbf{Cov}(U_i,U_j) \leq \frac{c}{nh_n^4}u(0) \leq \frac{c}{nh_n^4}.$$

Теорема 29. *Пусть* $\{U_n\}_{n\geq 1}$ *– стационарная последовательность ассоциированных случайных величин. Предположим, что условия* (**A**) *и* (**B**) *выполнены. Тогда для* $x \in J$,

$$g_n(x) - \boldsymbol{E}(g_n(x)) \underset{n\to\infty}{\overset{п.н.}{\to}} 0.$$

Доказательство. Положим

$$Z_{ni} = \frac{1}{h_n}\left(K\left(\frac{x-U_i}{h_n}\right) - \boldsymbol{E}\left(K\left(\frac{x-U_i}{h_n}\right)\right)\right) = \psi_n(x,U_i) - \boldsymbol{E}(\psi_n(x,U_i)), 1 \leq i \leq n.$$

Тогда $\boldsymbol{E}(Z_{ni})=0,\ \boldsymbol{D}(Z_{ni})<\infty,$ и

$$\sum_{j:|l-j|\geq r}|\mathbf{Cov}(Z_{nj},Z_{nl})|=\sum_{j:|l-j|\geq r}|\mathbf{Cov}(\psi_n(x,U_j),\psi_n(x,U_l))|\leq$$

$$\leq \sup_y\left\{\frac{\partial}{\partial y}\psi_n(x,y)\right\}^2\sum_{j:|l-j|\geq r}|\mathbf{Cov}(U_j,U_l)|\ .$$

Результат теоремы следует из условий (A2), (B) и теоремы 4.

22.3 Оценивание плотности распределения положительных случайных величин

Напомним следующую лемму (см. [8], т.2, VII.1, с.255).

Лемма 8. *Пусть* $f(t)-$ *ограниченная и непрерывная функция. Пусть* $G_{x,n}(t),\ n=1,2,...-$ *семейство распределений со средними* $\mu_n(x)\underset{n\to\infty}{\to}x$ *и дисперсиями* $h_n(x)\underset{n\to\infty}{\to}0$. *Тогда*

$$\tilde{f}_n(x)=\int_{-\infty}^{\infty}f(t)\,dG_{x,n}(t)\underset{n\to\infty}{\to}f(x).$$

Сходимость равномерна на каждом интервале, где $h_n(x)\underset{n\to\infty}{\to}0$ *и* $f-$ *равномерно непрерывна.*

Пусть теперь $F_n(x)-$ эмпирическая функция распределения и

$$\tilde{F}_n(x)=\int_{-\infty}^{\infty}F_n(t)\,dG_{x,n}(t). \qquad (8)$$

В силу теоремы Гливенко нетрудно показать, что $\sup\limits_x|\tilde{F}_n(x)-F(x)|\overset{п.н.}{\underset{n\to\infty}{\to}}0.$

Беря интеграл (8) по частям, получим

$$\tilde{F}_n(x)=1-\frac{1}{n}\sum_{j=1}^{n}G_{x,n}(X_j)\,, \qquad (9)$$

откуда беря производную, получим

$$\tilde{f}_n(x)=\frac{d\tilde{F}_n(x)}{dx}=-\frac{1}{n}\sum_{j=1}^{n}\frac{dG_{x,n}(X_j)}{dx}=-\frac{1}{n}\sum_{j=1}^{n}g_{x,n}(X_i). \qquad (10)$$

В частности, если X_j- положительные случайные величины и функция $G_{x,n}(t)=Q_{v_n}(t/x)$ убывает по $x>0$, то

$$F_n^+(x)=1-\frac{1}{n}\sum_{i=1}^{n}Q_{v_n}\left(\frac{X_i}{x}\right)\ \Rightarrow\ f_n^+(x)=\frac{1}{nx^2}\sum_{i=1}^{n}X_iq\left(\frac{X_i}{x}\right).$$

Здесь $q(x)-$ плотность, соответствующая ф.р. $Q(x)$.

Поскольку у оценки есть особенность в нуле, то будем рассматривать следующую оценку

$$f_n(x)=\frac{1}{n(x+\varepsilon_n)^2}\sum_{i=1}^{n}X_iq_{v_n}\left(\frac{X_i}{x+\varepsilon_n}\right),\ \ x\geq 0.$$

Теорема 30. *Если*

A. $v_n\to 0,\varepsilon_n\to 0$ *при* $n\to\infty$ *и*

B. $\sup_{x\geq 0}\int_0^{\infty}|\frac{d}{dx}g_{x+\varepsilon_n,n}(t)|dt=o((\frac{\ln\ln n}{\sqrt{n}})^{-1})$,

C. $\sup_{u>0,v>0}uq_v(u)<\infty,$

D. $f(x)$ *имеет ограниченную непрерывную производную на* $[0,\infty)$,

тогда

$$\sup_{x\geq 0}|f_n(x)-f(x)|\overset{п.н.}{\underset{n\to\infty}{\to}}0.$$

Пример 28. Пусть $g_{x,n}(t) = \frac{1}{x} q_v\left(\frac{t}{x}\right)$, где

$$q_v(t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} t^{\alpha-1} e^{-t/\beta}, \ t>0, \text{ где } \alpha = 1/v^2, \text{ и } \beta = 1/\alpha. \quad (\beta = v^2).$$

В этом случае $\int_0^\infty |\frac{d}{dx} g_{x+\varepsilon_n,n}(t)| dt = \frac{1}{(x+\varepsilon_n)^2 v^2} \int_0^\infty |t-(x+\varepsilon_n)| g_{x+\varepsilon_n,n}(t)\, dt = O\left(\frac{1}{(x+\varepsilon_n)v}\right)$,

откуда $\sup_{x\geq 0} \int_0^\infty |\frac{d}{dx} g_{x+\varepsilon_n,n}(t)| dt = O((v\varepsilon_n)^{-1})$. Выбирая $v\varepsilon_n = O\left(n^{-\frac{1}{2}+\delta}\right)$ для некоторого $0<\delta<1/2$, добьемся выполнения условия В теоремы.

Равенство (11) показывает, что эта оценка плотности есть н.о.р. с.в., где величины есть $Z_{in} = (X_i/(x+\varepsilon_n)^2) q_v(X_i/(x+\varepsilon_n)), i = 1,2,...,n.$

Следующая теорема дает условия на q_v и f, при которых имеет место асимптотическая нормальность этой оценки.

Теорема 31. *Предположим выполненными следующие условия:*

F. $f(x)$ *имеет ограниченную непрерывную производную на* $[0,\infty)$,

G1. $\int_0^\infty (q_v(t))^m dt = O(v^{-(m-1)})$ *при* $v \to 0$ *для* $1 \leq m \leq 3$, $I_2(q) = \lim_{v\to 0} v \int_0^\infty (q_v(t))^2 dt$ *существует,*

G2. с $q^*_{m,v}(t) = (q_v(t))^m / \int_0^\infty (q_v(x))^m dx, 1 \leq m \leq 3, \ (v \to 0)$,

(i) $\mu_{m,v} = \int_0^\infty t q^*_{m,v}(t) dt = 1 + O(v)$,

(ii) $\sigma^2_{m,v} = \int_0^\infty (t-\mu_{m,v})^2 q^*_{m,v}(t) dt = O(v^2)$,

(iii) $\sup_{0<v<\varepsilon} \int_0^\infty t^{4+\delta} q^*_{m,v}(t) dt < \infty$, *для некоторых* $\delta > 0, \varepsilon > 0$,

H. $nv \to \infty, nv\varepsilon_n \to \infty, nv^3 \to 0, nv\varepsilon_n^2 \to 0 \quad (n \to \infty)$.

Тогда при $n \to \infty$,

(a) $\sqrt{nv}(f_n(x) - f(x)) \xrightarrow{d} N(0, I_2(q)\frac{f(x)}{x})$, *для* $x > 0$,

(б) $\sqrt{nv}(f_n(0) - f(0)) \xrightarrow{d} N(0, I_2(q) f(0))$.

Проверим условия для $q_v(t)$ – гамма-распределение (пример 28).

Для $m \geq 1$, $(q_v(t))^m = (\int_0^\infty (q_v(w))^m dw) q^*_{m,v}(t)$, где $q^*_{m,v}(t)$ есть Гамма-распределение с $\alpha = mv^{-2} - m + 1, \beta = v^2 m^{-1}$ и

$$\int_0^\infty (q_v(t))^m dt = \frac{(v^{-2})^{mv^{-2}}}{(mv^{-2})^{mv^{-2}-m+1}} \cdot \frac{\Gamma(mv^{-2}-m+1)}{(\Gamma(v^{-2}))^m} \approx \frac{1}{\sqrt{m(2\pi)^{m-1}}} \cdot \frac{1}{v^{m-1}\sqrt{1-v^2}}, \text{ при } v \to 0,$$

используя Стирлинга аппроксимацию для Гамма-функции. Поэтому для любого $m \geq 1$,

(G1) $I_2(q) = \lim_{v\to 0} v \int_0^\infty (q_v(t))^2 dt = 1/\sqrt{4\pi}$,

(G2) (i) $\mu_{m,v} = \int_0^\infty t q^*_{m,v}(t)\, dt = 1 - ((m-1)/m) v^2$,

(G2) (ii) $\sigma^2_{m,v} = \int_0^\infty (t-\mu_{m,v})^2 q^*_{m,v}(t) dt = (1 - ((m-1)/m)v^2) v^2 m^{-1}$,

(G2) (iii) для любого $k \geq 1$ и любого $\varepsilon > 0$,

$$\sup_{0<v<\varepsilon}\int_0^\infty t^k q^*_{m,v}(t)\,dt = \sup_{0<v<\varepsilon}\frac{v^2\Gamma(k+mv^{-2}-m+1)}{m\Gamma(mv^{-2}-m+1)} = O\left(1+\frac{k-m+1}{m}v^2\right) < \infty.$$

22.4 Обобщенные ядерные оценки плотности Абрамсона-Новака

Рассмотрим следующую оценку плотности распределения $f(x)$:

$$f_{n,\alpha}(x) = \frac{1}{n}\sum_{i=1}^{n} f_{h\gamma}((X_i - x) f^{\alpha}(X_i)) f^{\alpha}(X_i)\mathbf{1}_i\,,$$

где $\boldsymbol{I}_i = I(|\,x - X_i\,|\, f^{\alpha}(x) < hT_+), T_+$ – константа такая, что $T_+ > T$, где плотность $f_\gamma(x)$ равна нулю вне интервала $[-T,T]$, – четная ограниченная и имеет конечное число точек разрыва (ядро).

Если $\alpha = \dfrac{1}{2}$, то эта оценка с небольшими изменениями совпадает с оценкой Абрамсона [29], а при $\alpha = 0-$это обычная ядерная оценка Розенблатта-Парзена. В общем случае такие статистики исследовал Новак С.Ю.

Казалось бы, какой смысл так устраивать оценку, если плотность $f(x)$ неизвестна? Но, если мы вместо $f(x)$ подставим подходящую обычную ядерную состоятельную оценку $\tilde{f}_n(x)$, то можно показать, что составная оценка будет состоятельной оценкой плотности.

Лемма 9. Пусть $\nu = \boldsymbol{E}(f_\gamma(\gamma))$. Тогда для любого действительного α при $h \to 0$

$$\boldsymbol{E}(f_{n,\alpha}(x)) \to f(x), \qquad nh\boldsymbol{D}(f_{n,\alpha}(x)) \to \nu f^{\alpha+1}(x).$$

Доказательство. Имеем:

$$\boldsymbol{E}(f_{n,\alpha}(x)) = \int_{-\infty}^{\infty} f^{\alpha+1}(x+hw) f_\gamma(f^{\alpha}(x+hw)) I(|\,w\,|\, f^{\alpha}(x) < T_+)\,dw,$$

откуда по теореме Лебега о мажорируемой сходимости,

$$\boldsymbol{E}(f_{n,\alpha}(x)) \to f^{\alpha+1}(x)\int_{-\infty}^{\infty} f_\gamma(wf^{\alpha}(x))\,dw = f(x) \quad \text{при} \quad h \to 0\,.$$

Аналогично заключаем, что

$$nh\boldsymbol{D}(f_{n,\alpha}(x)) = \to \int_{-\infty}^{\infty} f^{2\alpha+1}(x+hw) f_\gamma^2(wf^{\alpha}(x+hw))dw + O(h) \to \nu f^{\alpha+1}(x) \;\; (h \to 0).$$

Из этой леммы и неравенства Чебышева следует, что статистика $f_{n,\alpha}(x)$ сходится по вероятности к $f(x)$.

Список литературы

1. Гнеденко Б.В. Курс теории вероятностей. – М.: Едиториал УРСС, 2005. – 448 с.
2. Зубков А.М., Севастьянов Б.А., Чистяков В.П. Сборник задач по теории вероятностей. – М.: Наука, 1989. – 320 с.
3. Сборник задач по теории вероятностей, математической статистике и теории случайных функций под ред. Свешникова А.А. – СПб.: ЛАНЬ, 2007. – 448 с.
4. Renyi A. Probability Theory. – Budapest: Akademiai Kiado, 1970. – 666 p.
5. Dorogovtsev A.Ya. etc. Probability Theory. Collection of Problems. – AMS: Providence, R.I., 1997 – 348 p.
6. Колмогоров А.Н. Теория вероятностей и математическая статистика: избранные труды. – М.: Наука, 1986. – 534 с.
7. Рао С.Р. Линейные статистические методы и их применения. – М.: Наука, 1968. – 547 с.
8. Феллер В. Введение в теорию вероятностей и ее приложения.–М.: Мир, 1984. – 528 с.,722 с.
9. Карабанов А.К., Винокуров В.Ф., Дучиц Л.У. Культовые и исторические валуны Беларуси. – Мінск: Беларуская навука, 2011 – 235 с.
10. Мостеллер Ф. Пятьдесят занимательных вероятностных задач с решениями. – М., Наука, 1975 – 112 с.
11. Renyi A. On a one-dimensional problem concerning space filling. // Publ. Math. Inst. Hung. Acad. Sci., 1958. – №3 – p.109-127.
12. Dvoretzky A., Robbins, H. On the Parking Problem. // Publ. Math. Inst. Hung. Acad. Sci. – 1964 – vol.**9** – p. 209-224.
13. Kotz S., Ostrovskii I.V. A mixture representation of Linnik distribution. // Statistics & Probability Letters – vol.**26** – 1996 – p.61-64.
14. Tijms H. Understanding Probability. Cambridge University Press, 2012 – 562 p.
15. Watson G.N. A Treatise on the Theory of Bessel Functioms. Cambridge University Press, 1922.
16. Sneddon I.N. Fourier Transform, MacGraw-Hill, New York, 1951.
17. Федорюк М.В. Асимптотика: интегралы и ряды. – М.: Наука, 1987. – 544 с.
18. Градштейн И.С., Рыжик И.М. Таблицы интегралов, сумм, рядов и произведений. М.: ГИФМЛ, 1963. – 1100 с.
19. Хинчин А.Я. Об унимодальных распределениях. // Изв. НИИ мат. и мех. Томского ун-та – 1938 – вып. 2(2).
20. Henze B., Morgenstern H. Some elementary proofs of the normality of $XY/(X^2+Y^2)^{1/2}$ when X and Y are normal. // Comput. Math. Applic. – 1983 –vol. **15**, no. 11 – p. 943-944.
21. Puri P., Rubin H. A characterization based on the absolute difference of the two i.i.d. random variables. // Ann. Math. Statist. – 1970 – vol.**41**, no.6 – p. 2113-2122.
22. Pillai N., Meng X.L. An unexpected encounter with Cauchy and Levy. – Ann. Statist. – 2016 – vol.**44**, no.5 – p.2089-2097.
23. Shakil M., Golam Kibria B. Distributions of the product and ratio of Maxwell and Rayleigh random variables. // Stat. Papers – 2008 – vol.**49** – p.729-747.
24. Stein, C. A bound for the error in the normal approximation to the distribution of a sum ofdependent random variables. // In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* – 1972 – vol. **2** – p. 586–602.
25. Chen L., Goldstein L., Shao Q.M. Normal Approximation by Stein’s Method, NY, Springer, 2011. – 405 p.
26. Ross N. Fundamentals of Stein’s metod. – Probability Surveys, 2011, Vol. 8, pp.210-293.
27. Takagi K., Kumagai S., Matsunaga I., Kusaka Y. Application of inverse Gaussian distribution to occupational exposure data. // Ann. Occup. Hyg. – 1997 – vol.41, no.5 – p.505-514.

28. Kumagai S., Matsunaga I. Changes in the distribution of short-term exposure concentration with different averaging times. // American Industrial Hygiene Association Journal – vol.**56** – p.24-31.
29. Лукач Е. Характеристические функции. М., Наука, 1979. – 424с.
30. Булинский А.В., Шашкин А.П. Предельные теоремы для ассоциированных случайных полей и родственных систем. – М.: ФИЗМАТЛИТ, 2008. – 480 с.
31. Bagai I., Rao Pracasa B.L.S. Kernel-type density and failure rate estimation for associed sequences. // Ann. Inst. Statisat. Math. – 1995 – vol.**47**, no.2 – p. 253-266.
32. Chaubey Y.P., Sen A., Sen P.K. A new smooth density estimator for non-negative random variables. *Technical Report*, 2007, No.1/7, Department of Mathematics and Statistics, Concordia University, Montreal, Canada.
33. Abramson I.S. On bandwidth variation in kernel estimates – a square root law. // Ann. Statist. – 1982 – vol.**10**, no.4 – p.1217-1223.
34. Новак С.Ю. Обощенная ядерная оценка плотности. // Теор. вероятн. и её примен. – 1999 – т.**44**, вып.3 – с.634-645.
35. Гантмахер Ф.Р. Теория матриц. – М.: Наука, 1988. – 552 с.
36. Вистелиус А.Б. Основы математической геологии, Л.: Наука, 1980 – 389 с.
37. Athanassaoulis G., Skarsoulis E. Belibassakis K. Bivariate distributions with given marginals with an application to wave climate description. // Applied Ocean Research – 1994 – vol.**16** – pp.1-17.
38. Митропольский А.К. Техника статистических вычислений. М.: Наука, 1971 – 576 с.

Михаил Семенович **Тихов**
Никита Викторович **Капкаев**

ПОСТРОЕНИЕ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ

Учебно-методическое пособие

Компьютерная верстка – М.С. Тихов, Н.В. Капкаев

Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский Нижегородский государственный
университет им. Н.И. Лобачевского».
603950, Нижний Новгород, пр. Гагарина, 23.