# Применение статистических методов в обработке результатов измерений

Дзюба А.А., Крышень Е.Л.

24 января 2022 г.

## Аннотация

Учебное пособие *Применение статистических методов в обработке результатов измерений* предназначено для студентов, аспирантов, изучающих различные разделы физики. Оно разработано на основе курса лекций, прочитанных аспирантам НИЦ "Курчатовский институт" — ПИЯФ и содержит основные сведения по теории вероятности и математической статистики.

# Оглавление

# 1 Введение

## 1.1 Физический эксперимент

Познание свойств окружающего нас мира включает следующие этапы: первичное изучение некоторого явления при помощи наблюдения, создание теории и её проверка в рамках научного опыта при точно учитываемых условиях, которые предоставляют возможность следить за ходом исследуемого процесса. Важной составляющей является возможность повторения эксперимента при воспроизведении одних и тех же условий.

Порядок проведения физического эксперимента таков:

1. Постановка задачи и пути её экспериментального решения;
2. Планирование эксперимента;
3. Выбор структурной схемы и оценка возможных эффектов и фона;
4. Изготовление установки;
5. Подготовка плана измерений и калибровка средств измерений;
6. Разработка или выбор методов контроля аппаратуры;
7. Проведение измерения;
8. Обработка результатов.

Под *измерением* следует понимать совокупность действий для определения отношения одной (измеряемой) величины к другой однородной величине, принятой за единицу, иногда хранящуюся в средстве измерений. А под *средством измерения* — техническое средство, предназначенное для измерений, имеющее нормированные метрологические характеристики, воспроизводящее и (или) хранящее единицу физической величины, размер которой принимают неизменным (в пределах установленной погрешности) в течение известного интервала времени.

При планировании эксперимента и обработке его результатов используются методы *математической статистики*. Задачей этого раздела математики является построение *статистического вывода* (*англ.* statistical inference), — обобщение информации из выборки для получения представления о свойствах *генеральной совокупности* (*англ.* statistical population) — совокупности всех объектов, относительно которых предполагается делать выводы при изучении конкретной задачи. Под *выборкой* (*англ.* sample) понимают часть генеральной совокупности элементов, которая охватывается экспериментом или наблюдением.[1] Результатом статистического вывода является *статистиче-*

[1] для физического эксперимента понятие генеральной совокупности плохо определено.

*ское суждение*, (*англ.* statistical proposition) например: точечная оценка, доверительный интервал, отвержение гипотезы.

Математическая статистика неразрывно связана с другой дисциплиной математики — *теорией вероятности* (*англ.* probability theory).

## 1.2 Основные определения и законы теории вероятности

В основе теории вероятности лежит понятие *случайного эксперимента* (*англ.* experiment или trial), то есть математической модели, соответствующей реальному эксперименту, результат которого невозможно точно предсказать. К этой модели предъявляются следующие требования:

- она должна адекватно описывать эксперимент;
- совокупность множества наблюдаемых результатов должна быть определена;
- должна существовать принципиальная возможность осуществления эксперимента со случайным исходом сколь угодное количество раз при неизменных входных данных;
- должно быть доказано требование или изначально принята *гипотеза о стохастической устойчивости относительной частоты* для любого наблюдаемого результата, определённого в рамках математической модели: то есть при бесконечном числе повторений случайного эксперимента относительные частоты его исходов стремятся к определенным значениям.

Определим *пространство элементарных событий* (*англ.* sample space) как множество $\Omega$ всех различных *исходов* (*англ.* outcome) случайного эксперимента (в каждом эксперименте реализуется один и только один исход). Объединение (сумма) некоторых подмножеств (*событий*, *англ.* event) интерпретируется как событие, заключающееся в наступлении хотя бы одного из них. Пересечение (произведение) подмножеств (событий) интерпретируется как событие, заключающееся в наступлении всех этих событий. Непересекающиеся множества интерпретируются как несовместные события (их совместное наступление невозможно). Соответственно, пустое множество $\emptyset$ означает невозможное событие.

*Вероятностью* (*англ.* probability) называется мера[2] $\mathbf{P}$, заданная на множестве событий и обладающая свойствами: *неотрицательности*, *аддитивности* (для попарно несовместных событий вероятность наступления хотя бы одного равна сумме вероятностей этих событий), *конечности* ($\mathbf{P}(\Omega) = 1$).

[2] числовая функция

Вероятность обладает следующими свойствами:

1. $\mathbf{P}\{\emptyset\} = 0$;
2. $\forall A : 0 \leq \mathbf{P}\{A\} \leq 1$;
3. если $\bar{A}$ – событие, противоположное $A$, то $\mathbf{P}\{\bar{A}\} = 1 - \mathbf{P}\{A\}$;
4. если $A \subset B$, то есть наступление события $A$ влечёт также наступление события $B$, то $\mathbf{P}\{A\} \leq \mathbf{P}\{B\}$;
5. если $A \subset B$, то $\mathbf{P}\{B \setminus A\} = \mathbf{P}\{B\} - \mathbf{P}\{A\}$;
6. вероятность наступления хотя бы одного из произвольных (не обязательно несовместных) двух событий $\mathbf{P}\{A + B\} = \mathbf{P}\{A\} + \mathbf{P}\{B\} - \mathbf{P}\{AB\}$, ибо $A + B = A + (B \setminus (AB))$.

Определим *случайную величину* (*англ.* random variable) как переменную, значения которой представляют собой исходы какого-нибудь случайного феномена или эксперимента. *Распределение вероятностей* случайной величины (*англ.* probability distribution) — это закон, описывающий область её значений и вероятности различных исходов. Распределение вероятностей задается *функцией распределения* $F(x)$ (англ. cumulative distribution function, CDF), описывающей вероятность того, что случайная величина примет значение, меньшее или равное $x$, где $x$ — произвольное действительное число.[3] Для любой случайной величины:

1. $F(x)$ — неубывающая;
2. $\lim_{x \to -\infty} F(x) = 0$ и $\lim_{x \to +\infty} F(x) = 1$;
3. $F(x)$ непрерывна справа.

При этом любая функция, удовлетворяющая этим условиям, является функцией распределения какой-то случайной величины.

Из функции распределения можно получить *плотность вероятности* $f(x)$ (*англ.* probability density function), которая является еще одним из способов задания вероятностной меры. Для непрерывных случайных величин:

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x). \tag{1.1}$$

Пример плотности вероятности и функции распределения для нормального распределения приведен на рис. 1.2.

Переход к случаю дискретных случайных величин осуществляется через $\delta$-функцию Дирака:

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}, \qquad \int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1. \tag{1.2}$$

[3] Пока рассматриваем одномерную случайную величину

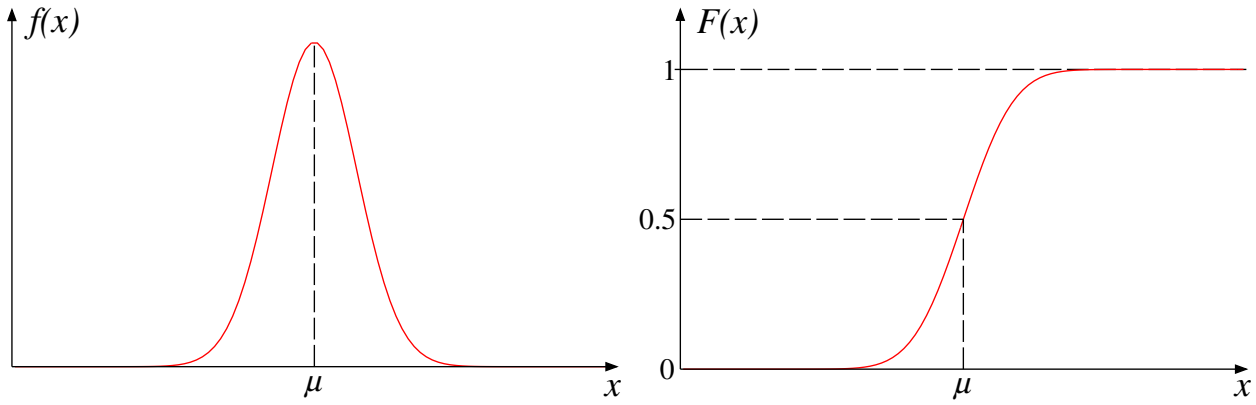


Рис. 1.1: Пример плотности вероятности $f(x)$ и функции распределения $F(x)$ для распределения Гаусса.

*Дискретная плотность вероятности* (*англ.* probability mass function) выражается через вероятности исходов ($p_i$) как:

$$f(x) = \sum_{i=1}^{n} p_i \, \delta(x - x_i),$$

то есть в этом случае $F(x)$ — ступенчатая функция. При этом

$$\mathbf{P}(x = x_i) = p_i, \qquad \sum_{i=1}^{\infty} p_i = 1.$$

Пример плотности вероятности и функции распределения для дискретной случайной величины приведен на рис. 1.2.

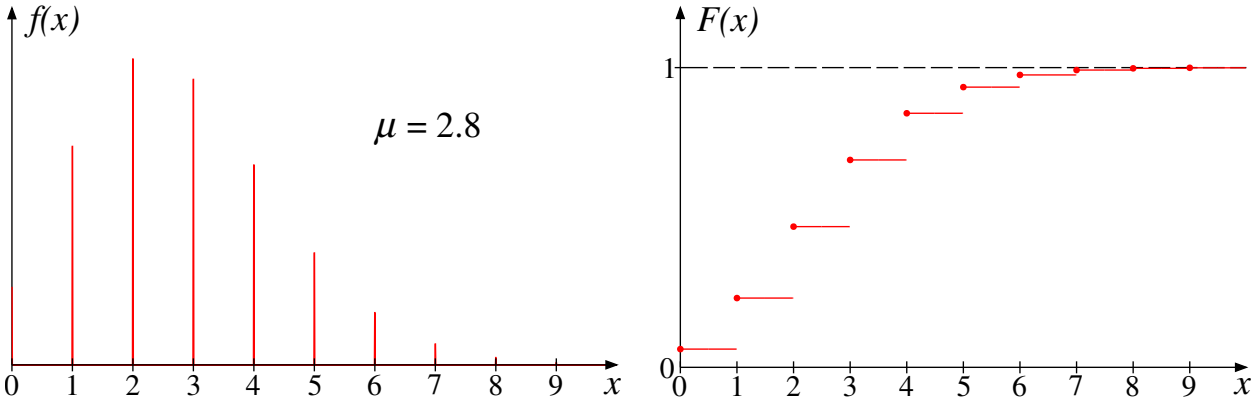


Рис. 1.2: Пример плотности вероятности $f(x)$ и функции распределения $F(x)$ для дискретной случайной величины.

*Математическим ожиданием* $\mathbf{M}[x]$ (*англ.* expected value) называют среднее значение случайной величины при стремлении количества её измерений

к бесконечности. Для непрерывной случайной величины:

$$\mathbf{M}[x] = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x, \tag{1.3}$$

а для дискретной:

$$\mathbf{M}[x] = \sum_{i=1}^{\infty} x_i\, p_i. \tag{1.4}$$

Математическое ожидание может быть определено и для преобразования случайной величины. Если $g(x)$ — интегрируемая функция и математическое ожидание случайной величины $y = g(x)$ конечно, то для непрерывных распределений оно будет вычисляться как

$$\mathbf{M}\left[g(x)\right] = \int_{-\infty}^{\infty} g(x) f(x)\, \mathrm{d}x, \tag{1.5}$$

а для дискретных

$$\mathbf{M}\left[g(x)\right] = \sum_{i=1}^{\infty} g(x_i) p_i. \tag{1.6}$$

*$k$-м начальным моментом* случайной величины $x$ (*англ.* raw moment), где $k$ — натуральное число, называется величина

$$\nu_k = \mathbf{M}\left[x^k\right] = \int_{-\infty}^{\infty} x^k\, f(x)\, \mathrm{d}x, \tag{1.7}$$

если математическое ожидание в правой части этого равенства определено. Набор моментов характеризует распределение случайной величины. Целесообразно определить функцию, зависящую от всех моментов, так называемую *производящая функция моментов* (*англ.* moment-generating function):

$$M_x(t) = \int_{-\infty}^{\infty} e^{tx}\, f(x) \mathrm{d}x, \tag{1.8}$$

где $t$ — вспомогательная переменная. При этом

$$M_x(t) = 1 + \nu_1 t + \nu_2 \frac{t^2}{2!} + \nu_3 \frac{t^3}{3!} + \ldots, \tag{1.9}$$

а моменты могут быть вычислены как

$$\nu_i = \frac{\mathrm{d}^i M_x(t)}{\mathrm{d}x^i}. \tag{1.10}$$

Важно отметить, что производящая функция моментов однозначно определяет распределение вероятностей.

*Центральными моментами k-го порядка* (*англ.* central moment) называют:

$$\mu_k = \mathbf{M}\left[(x - \mathbf{M}[x])^k\right]. \tag{1.11}$$

Второй центральный момент $\mu_2$ называется *дисперсией распределения* (*англ.* variance). Дисперсия показывает разброс распределения вокруг среднего значения. Квадратный корень из дисперсии называют *среднеквадратическим отклонением* (*англ.* standard deviation) и обозначают греческой буквой $\sigma$.

Рассмотрим сумму двух независимых случайных величин с разными математическими ожиданиями и дисперсиями $x = x_1 + x_2$. Дисперсия этой величины равна сумме дисперсий $x_1$ и $x_2$. Это вытекает из линейности математического ожидания. Действительно третий член в выражении

$$\sigma^2 = \mathbf{M}[(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + 2(x_1 - \mu_1)(x_2 - \mu_2)] \tag{1.12}$$

зануляется. Это утверждение можно обобщить на сумму нескольких случайных величин:

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2. \tag{1.13}$$

Величина $\mu_3/\sigma^3$ называется *коэффициентом асимметрии* (*англ.* skewness). Если коэффициент асимметрии положителен, то правый хвост распределения длиннее левого, если отрицателен, то наоборот. Если распределение симметрично относительно математического ожидания, то коэффициент асимметрии равен нулю.

Некоторые распределения вероятности не имеют моментов. Например, рассмотрим пушку, установленную перед бесконечной стеной, вращающуюся на лафете, и стреляющую в случайный момент времени. Распределение попаданий снарядов на стене подчиняется *распределению Коши*.

$$\mathrm{Cauchy}_{x_0,\gamma}(x) = \frac{1}{\pi}\left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right], \tag{1.14}$$

где $x_0$ — параметр сдвига, а $\gamma > 0$ — параметр масштаба. Так как интеграл Лебега не определён, ни дисперсия, ни моменты старших порядков этого распределения не определены. Интеграл 1-го момента в смысле главного значения равен:

$$\lim_{c\to\infty} \int_{-c}^{c} x \cdot \frac{1}{\pi}\left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right] \mathrm{d}x = x_0 \tag{1.15}$$

*Квантиль* (*англ.* quantile) — значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если распределение непрерывно, то $\alpha$ -квантиль однозначно задаётся уравнением

$$F(x_\alpha) = \alpha, \tag{1.16}$$

где $F(x)$ — функция распределения. *Медиана распределения* (*англ.* median) — это 0.5-квантиль. Медиана определяется для всех распределений, а в случае неоднозначности естественным образом доопределяется, в то время как математическое ожидание может быть не определено.

*Модой* (*англ.* mode) абсолютно непрерывного распределения называют любую точку локального максимума плотности этого распределения. Для дискретных распределений модой считают любое значение, вероятность которого больше, чем вероятности соседних значений.

*Информационная энтропия* (*англ.* entropy) — это мера неопределённости некоторой системы, в частности непредсказуемость появления какого-либо исхода (реализации случайного эксперимента). Для дискретной случайной величины, заданной набором вероятностей реализации ($\{p_i\}_{i=1}^n$):

$$H = -K \sum_{i=1}^{n} p_i \log_2 p_i, \tag{1.17}$$

где $K$ — положительная константа, которая нужна только для выбора единицы измерения энтропии. Изменение $K$ равносильно изменению основания логарифма. Чем больше энтропия, тем менее предсказуем исход случайного эксперимента.

Определение 1.17 для дискретных случайных событий можно формально расширить для непрерывных распределений, заданных плотностью распределения вероятностей, однако полученный функционал будет обладать несколько иными свойствами. Действительно, если определить

$$p_k = \int_{x_k}^{x_k+\Delta x} f(x)\mathrm{d}x, \tag{1.18}$$

то результат предельного перехода

$$\lim_{\Delta x \to 0} H = -\int_{-\infty}^{+\infty} f(x) \log_2 f(x)\mathrm{d}x - \lim_{\Delta x \to 0} \log_2 \Delta x, \tag{1.19}$$

стремится к бесконечности при $\Delta x \to 0$. Таким образом, непрерывные случайные величины не допускают введения конечной абсолютной меры неопределенности, однако, для них может быть введена относительная мера — *дифференциальная энтропия* (*англ.* differential entropy, continuous entropy) равная первому члену в выражении 1.19.

Можно показать, что среди всех распределений с конечной дисперсией дифференциальная энтропия максимальна в случае *нормального распределения* (*распределение Гаусса* или *Гаусса–Лапласа*, *англ.* normal distribution):

$$\mathrm{Gauss}_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1.20}$$

Среди распределений, заданных на ограниченном промежутке, максимум дифференциальной энтропии достигается для *равномерного распределения* (*англ.* uniform distribution, rectangular distribution):

$$\mathrm{Uniform}_{a,b}(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b]. \end{cases} \tag{1.21}$$

Кратко рассмотрим три теоремы о последовательности сумм случайных величин. Будем считать, что $x_n$ — независимые одинаково распределённые случайные величины с нулевым математическим ожиданием и единичной дисперсией[4], а $S_n = \sum_{i=1}^{n} x_i$.

*Закон больших чисел* (*англ.* law of large numbers) утверждает, что среднее значение конечной выборки из фиксированного распределения близко к математическому ожиданию этого распределения. При нулевом математическом ожидании это эквивалентно тому, что $\frac{S_n}{n} \longrightarrow 0$.

*Закон повторного логарифма* (*англ.* law of the iterated logarithm) определяет порядок роста делителя последовательности сумм случайных величин, при котором эта последовательность не сходится к нулю, но остается почти всюду в конечных пределах. Этот закон утверждает, что почти наверное[5]:

$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{n \ln\ln n}} = -\sqrt{2}, \qquad \limsup_{n\to\infty} \frac{S_n}{\sqrt{n \ln\ln n}} = \sqrt{2}.$$

Иными словами закон повторного логарифма утверждает, что $S_n/\sqrt{n \ln\ln n}$ будет меньше, чем любое заданное $\varepsilon > 0$ с вероятностью, стремящейся к единице, но она будет бесконечное число раз приближаться сколь угодно близко к любой точке отрезка $[-\sqrt{2}, \sqrt{2}]$ почти наверное.

*Центральная предельная теорема* (*англ.* central limit theorem) утверждает, что при $n \to \infty$ распределение суммы случайных величин $S_n/\sqrt{n}$ сходится к *стандартному нормальному распределению*[6]. В случае ненулевого математического ожидания

$$z = \sqrt{n}\,\frac{\bar{x}_n - \mu}{\sigma} \to \mathrm{Gauss}_{0,1}(x), \tag{1.22}$$

[4] для любой случайной величины с конечной дисперсией этого можно добиться при помощи линейного преобразования.

[5] это произойдет с вероятностью 1. Понятие является аналогом понятия «почти всюду» в теории меры. В то время, как во многих основных вероятностных экспериментах нет никакой разницы между «почти достоверно» и «достоверно», (то есть, событие произойдет совершенно точно), это различие важно в более сложных случаях, относящихся к случаям рассмотрения какой-либо бесконечности.

[6] нормальное распределение с $\mu = 0$ и $\sigma = 1$.

то есть, неформально говоря, сумма $n$ независимых одинаково распределённых случайных величин имеет распределение, близкое $\mathrm{Gauss}_{n\mu,n\sigma^2}(x)$, или эквивалентно, $\bar{x}_n$ имеет распределение близкое к $\mathrm{Gauss}_{\mu,\sigma^2/n}(x)$. Центральная предельная теорема доказывается, через равенство производящих функций моментов для $z$ и для нормального распределения.

Сравнивая закон повторного логарифма и центральную предельную терему, легко видеть, что в первом случае получается равномерное распределение, заданное на отрезке, а во втором, нормальное заданное на всей оси вещественных чисел. Легко показать, что отношение дифференциальных энтропий для этих распределений стремится к единице. Это отражает факт, что, суммируя случайные величины, мы получаем максимально запутанную (неопределенную) ситуацию.

В формулировках предельных законов не указывается никакого конкретного закона распределения, а требуется лишь наличие конечной дисперсии. Результат классической центральной предельной теоремы справедлив для ситуаций гораздо более общих, чем полная независимость и одинаковый закон распределения случайных событий. В частности, пусть на объект воздействует множество случайных и независимых факторов, сравнимых по своему рассеиванию (предполагается равномерно малое влияние слагаемых на рассеивание суммы). Если воздействие этих факторов складывается (имеет *аддитивный характер*), то соответствующая случайная величина имеет нормальное распределение. Если же эффекты накапливаются в объекте в зависимости от их предыдущего количества, то есть соответствующие факторы не складываются, а перемножаются (*мультипликативный характер*), то соответствующая случайная величина имеет *логнормальное распределение* (*англ.* log-normal distribution), а её логарифм имеет нормальное распределение. Плотность вероятности для логнормального распределения имеет вид:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x-\mu)^2/2\sigma^2}. \tag{1.23}$$

Логнормальное распределение удовлетворительно описывает многие эффекты, например, распределение частот частиц по их размерам при случайном дроблении, или время от заражения до острой стадии развития заболевания.

## 1.3 Интерпретация вероятности

Описание физического мира при помощи математической теории требует физической интерпретации математического понятия *вероятность*. В настоящее время в экспериментальной физике широко используются два подхода к интерпретации вероятности: *байесовская статистика* (*англ.* bayesian inference) и т.н. *частотный подход* (*англ.* frequentist inference). Соответствен-

но, перед проведением статистического анализа экспериментальных данных необходимо определиться, в рамках какого подхода он будет проведен.

Пусть два события $A$ и $B$ принадлежат одному полю вероятностного пространства. Можно ввести *условную вероятность* (*англ.* conditional probability) наступления события $A$ при условии $B$, как $\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$. Заметив, что вероятность совместного наступления событий $A$ и $B$ равна $\mathbf{P}(AB) = \mathbf{P}(A|B)\mathbf{P}(B) = \mathbf{P}(B|A)\mathbf{P}(A)$, из этого определения напрямую устанавливается связь между $\mathbf{P}(A|B)$ и $\mathbf{P}(B|A)$:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B|A)\,\mathbf{P}(A)}{\mathbf{P}(B)}. \tag{1.24}$$

Выражение 1.24 называют *теоремой (формулой) Байеса* (*англ.* Bayes' theorem) и является одной из основ *байесовской теории*, используемой как метод адаптации существующих вероятностей к вновь полученным экспериментальным данным. В основе этой теории лежит *субъективная интерпретация вероятности*, под которой понимают степень личной уверенности субъекта в возможность наступления некоторого события. При этом субъект должен быть *рациональным*, то есть продукт его веры должен подчиняться определенным правилам. Наиболее общее определение «степени уверенности» основано на пари: степень уверенности отражается величиной ставки, которую субъект готов поставить на то, что суждение истинно.

Альтернативой байесовской интерпретации понятия вероятности выступает *частотная вероятность*, определяемая как предел относительной частоты наблюдения некоторого события $A$ в серии однородных независимых испытаний.

$$\mathbf{P}(A) = \lim_{N \to \infty} \frac{n}{N}, \tag{1.25}$$

где $N$ – общее количество испытаний, $n$ – количествово наблюдений $A$. Несмотря на то, что данное определение скорее указывает на способ оценки неизвестной вероятности — путем большого количества однородных и независимых наблюдений, — тем не менее в таком определении отражено содержание понятия вероятности. В научной литературе частотную интерпретацию часто называют *эмпирической* или *объективной*[7].

## 1.4 Многомерные распределения вероятности

Для начала расширим определение функции распределения на случай двух переменных:

$$F(x, y) = \mathbf{P}\{(x' < x) \cap (y' < y)\} \tag{1.26}$$

[7] В последнее время такой подход часто именуют *фриквентистским* (*англ.* frequency — частота)

с условиями

$$F(-\infty, y) = F(x, -\infty) = 0, \quad F(\infty, \infty) = 1. \tag{1.27}$$

Это должна быть монотонно возрастающая функция по обеим переменным. Можно определить двумерную плотность вероятности через частные производные и соответствующее условие нормировки:

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}, \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \mathrm{d}x \mathrm{d}y = 1. \tag{1.28}$$

Проекции $f(x, y)$ на соответствующие оси называют *маргинальными распределениями* (*англ.* marginal distribution):

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) \mathrm{d}y, \qquad f_y(y) = \int_{-\infty}^{\infty} f(x, y) \mathrm{d}x. \tag{1.29}$$

Маргинальные распределения связаны с соответствующими условными вероятностями через теорему Байеса:

$$f_x(x \mid y) f_y(y) = f_y(y \mid x) f_x(x) = f(x, y). \tag{1.30}$$

Начальные и центральные моменты двумерного распределения определяются через математические ожидания:

$$\nu_{lm} = \mathbf{M}(x^l y^m) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^l y^m f(x, y) \mathrm{d}x \mathrm{d}y, \tag{1.31}$$

$$\mu_{lm} = \mathbf{M}((x - \nu_{10})^l (y - \nu_{01})^m). \tag{1.32}$$

Величину $\mu_{11}$ называют *ковариацией* (*англ.* covariance) и обозначают $\sigma_{xy}$. Если она отлична от нуля, то переменные *коррелированы*. Можно ввести так называемые *коэффициенты корреляции*[8] (*англ.* correlation coefficient):

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{1.33}$$

показывающие обезразмеренную[9] степень корреляции случайных величин. Предельный случай $|\rho_{xy}| = 1$ соответствует линейной корреляции между переменными.

Две случайные величины *независимы* (*англ.* independent variables), когда двумерная плотность вероятности является произведением маргинальных распределений:

$$f(x, y) = f_x(x) f_y(y), \tag{1.34}$$

[8] В определении $\sigma_x = \mu_{20}$, а $\sigma_y = \mu_{02}$.

[9] Так как справедливо *неравенство Шварца* $|\rho_{x,y}| \leq 1$.

или, что то же самое, условные вероятности тождественны маргинальным распределениям. При этом важно понимать, что равенство нулю коэффициента корреляции не означает автоматической независимости случайных величин[10]. В качестве примера некоррелированных, но зависимых случайных величин можно привести распределение

$$f(x,y) = \exp(-\sqrt{x^2+y^2})/(2\pi\sqrt{x^2+y^2}), \tag{1.35}$$

для которого $\sigma_{xy} = 0$, однако, требование 1.34 очевидно не соблюдается.

Определения 1.26 и 1.27 не сложно обобщить на $N$-мерный случай. Для него удобно ввести векторную нотацию. В такой нотации дисперсия многомерного распределения описывается так называемой *ковариационной матрицей* (*англ.* covariance matrix) с элементами:

$$C_{ij} = \mathbf{M}((x_i - \mathbf{M}(x_i))(x_j - \mathbf{M}(x_j))) = \mathbf{M}(x_i x_j) - \mathbf{M}(x_i)\mathbf{M}(x_j). \tag{1.36}$$

Базовой концепцией измерений является случай *независимых одинаково распределенных случайных величин*(*англ.* independent and identically distributed random variables), для которых

$$f(x_1, \ldots, x_N) = \prod_{i=1}^{N} f(x_i). \tag{1.37}$$

Для такого многомерного распределения вероятности ковариационная матрица диагональна, причем все её диагональные элементы равны дисперсиям случайных величин $x_i$, то есть одинаковы!

В качестве примера можно привести *двумерное нормальное распределение*, задаваемое как:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-2\rho\frac{xy}{\sigma_x\sigma_y})}. \tag{1.38}$$

Величина $\rho$ определена на отрезке $(-1, 1)$ и равна коэффициенту корреляции. Для такого распределения можно перейти к независимым случайным величинам ($x'$ и $y'$) путем поворота на угол $\phi = \mathrm{atan}[2\rho\sigma_x\sigma_y/(\sigma_x^2 - \sigma_y^2)]/2$:

$$x' = x\cos\phi + y\sin\phi, \qquad y' = -x\sin\phi + y\cos\phi. \tag{1.39}$$

Распределение для новых переменных можно получить из выражения 1.38, положив $\rho = 0$. При этом плотность распределения вероятности факторизуется.

[10] Обратное верно.

# 2 Описание данных

## 2.1 Выборочная функция распределения

Пусть $x_1, \ldots, x_n$ — выборка из распределения случайной величины $x$, задаваемой некоторой функцией распределения $F(x)$. Будем считать, что $x_i$, где $i \in \mathbf{N}$, — независимые случайные величины, определённые на некотором пространстве элементарных исходов $\Omega$. Функцию, зависящую от выборки, называют *статистикой* (*англ.* sample statistic). Для $x \in \mathbf{R}$ определим случайную величину $\hat{F}_n(x) : \Omega \to \mathbf{R}$ следующим образом:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} h(x - x_i), \tag{2.1}$$

где $h(x)$ — ступенчатая *функция Хевисайда*:

$$h(x) = \begin{cases} 0, & x < 0; \\ 1, & x \geq 0. \end{cases} \tag{2.2}$$

Случайная величина $\hat{F}_n(x)$ называется *выборочной функцией распределения* (или *эмпирической функцией распределения*, *англ.* empirical distribution function) случайной величины $x$ и является аппроксимацией для функции распределения. $\hat{F}_n(x)$ несет максимальную информацию о выборке и равномерно почти наверное сходится к $F(x)$ при $n \to \infty$ (*теорема Гливенко-Кантелли*).

Полезно рассмотреть скорость сходимости выборочной функции распределения $F_n(x)$ к её теоретическому аналогу $F(x)$. Рассмотрим величину $D_n = \sup_{x \in \mathbf{R}} |\hat{F}_n(x) - F(x)|$, равную модулю максимального отклонения выборочной функции распределения от теоретической. Тогда, согласно *теореме Колмогорова*, величина $\sqrt{n} D_n$ стремится к распределению Колмогорова:

$$\mathrm{Kolm}(x) = \begin{cases} \sum_{l=-\infty}^{\infty} (-1)^l e^{-2l^2 x^2}, & x > 0; \\ 1, & x \leq 0. \end{cases} \tag{2.3}$$

Это свойство можно использовать для оценки границ, в которые с заданной вероятностью попадает $F(x)$ (*критерий Колмогорова*):

$$\mathbf{P}\left( F_n(x) - \frac{k_\alpha}{\sqrt{n}} \leq F(x) \leq F_n(x) + \frac{k_\alpha}{\sqrt{n}}, x \in \mathbf{R} \right) \xrightarrow{n \to \infty} \mathrm{Kolm}(k_\alpha) = \alpha, \tag{2.4}$$

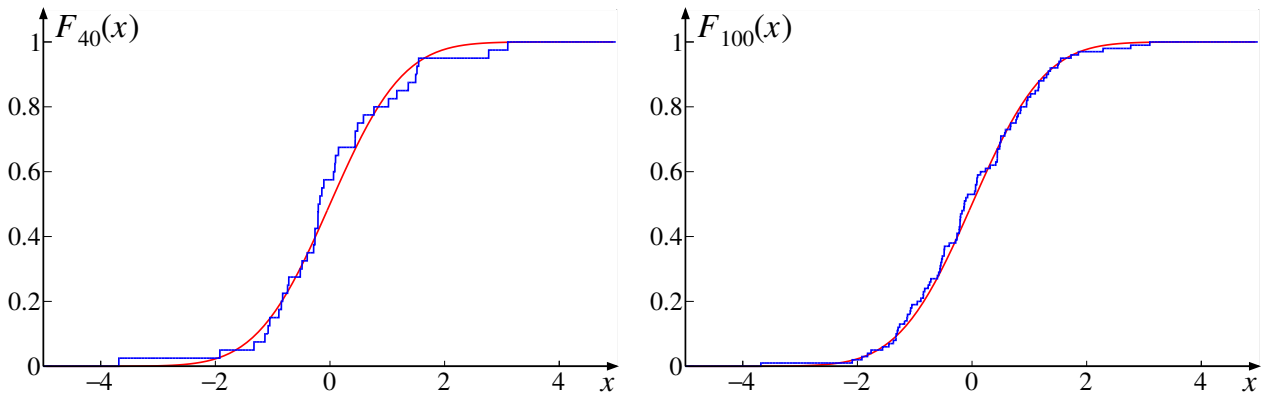


Рис. 2.1: Примеры выборочной функции распределения $F_n(x)$ для $n = 40$ (слева) и $n = 100$ (справа).

где $\alpha$ — *уровень значимости* (*англ.* significance level), т.е. наперед заданная вероятность отвергнуть правильную гипотезу[1], $k_\alpha$ — квантиль распределения Колмогорова. Критерий Колмогорова нельзя применять, если форма функции распределения $F(x)$ или её параметры определяются из той же выборки. При огромных выборках важно, чтобы измеряемая $\hat{F}_n(x)$ была гладкой.[2] При малых выборках предпочтительней использование статистики с *поправкой Большева*: $\sqrt{n}D_n + \frac{1}{6\sqrt{n}}$.

Распределение Колмогорова можно использовать для проверки *однородности выборок*, то есть того, что две исследуемые выборки подчиняются одному распределению случайной величины (*критерий Смирнова*). Для двух эмпирических функций распределения выборок объемом $n$ и $m$ строится случайная величина:

$$D_{n,m} = \sup_{x \in \mathbf{R}} |\hat{F}_n(x) - \hat{F}_m(x)|. \tag{2.5}$$

Если статистика $\sqrt{\frac{nm}{n+m}}D_{n,m}$ превышает $\alpha$-квантиль распределения Колмогорова, то гипотеза об однородности выборок отвергается.

## 2.2 Выборка и информация

Как уже отмечалось выше, максимальную информацию о выборке несет сама выборка. Допустим, известно, что выборка порождена функцией распределения, которая в свою очередь принадлежит некоторому семейству распределений, заданному плотностью вероятности $f(\theta, x)$, зависящей от параметра ($\theta \in \Theta$).[3] Очевидно, что выборка несет максимальную информацию об истинном значении этого параметра. В этом случае оперируют *функцией прав-*

[1] Далее, мы всегда будем использовать обозначение $\alpha$ в этом смысле.

[2] На больших выборках нельзя применять критерий Колмогорова, если непрерывная случайная величина измеряется дискретным прибором.

[3] Определения ниже можно распространить и на случай нескольких параметров (вектор параметров).

*доподобия* (*англ.* likelihood function) — совместное распределение выборки из параметрического распределения, рассматриваемое как функция параметра — $f_L(x|\theta)$. Для независимой выборки:

$$f_L(x|\theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \tag{2.6}$$

На практике удобно ввести логарифмическую функцию правдоподобия. В случае независимой выборки произведение переходит в сумму логарифмов, что значительно упрощает представление данных в компьютерной памяти, а также процедуру дифференцирования:

$$L(x \mid \theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta). \tag{2.7}$$

Отметим, что функция правдоподобия не несет смысла плотности вероятности, хотя бы потому, что для неё не соблюдается условие нормировки.

Функция правдоподобия связана с информацией о параметре семейства распределений, которую несет выборка. *Информацией Фишера* (*англ.* Fisher information) для данной статистической модели при $n$ испытаниях называют математическое ожидание (при данном $\theta$) квадрата частной производной логарифмической функции правдоподобия по параметру:

$$I_n(\theta) = \mathbf{M}_\theta \left( \frac{\partial L}{\partial \theta} \right)^2. \tag{2.8}$$

Если известно какому семейству распределений принадлежит функция распределения, из которой получена выборка, то в некоторых случаях можно получить соотношения (статистики), несущие столько же информации, сколько несет выборка. *Достаточной статистикой* (*англ.* sufficient statistic) называют такую функцию выборки, что при данном её значении условная вероятность выборки не зависит от значений параметра.

Например, для выборки из равномерного распределения с неизвестными параметрами $a$ и $b$ пара, состоящая из минимального и максимального значения в выборке, является достаточной статистикой. Достаточной статистикой для выборки из нормального распределения является пара: сумма значений элементов и сумма квадратов значений.

Достаточная статистика содержит столько же информации Фишера, сколько и вся выборка. Это следует из *факторизационного критерия Неймана* для достаточной статистики: если статистика $T(x)$ достаточна для $\theta$, то существуют функции $g$ и $h$ такие, что:

$$f(x, \theta) = g(T(x), \theta)h(x). \tag{2.9}$$

Действительно, из-за независимости $h$ от параметра соответствующее слагаемое при дифференцировании по $\theta$ зануляется.

## 2.3 Описательная статистика

Обработкой эмпирических данных, их систематизацией, наглядным представлением в форме графиков и таблиц, а также их количественным описанием посредством основных статистических показателей занимается *описательная статистика* (*англ.* descriptive statistics). Её можно противопоставить статистическому выводу в том смысле, что в рамках описательной статистики на основании результатов исследования частных случаев не делается никаких выводов о генеральной совокупности. Напротив, статистический вывод предполагает, что свойства и закономерности, выявленные при исследовании объектов выборки, также присущи генеральной совокупности.

Описательная статистика использует три основных метода агрегирования данных:

1. табличное представление данных,
2. расчет статистических показателей,
3. графическое представление (гистограммы, диаграммы рассеяния, графики).

Под *статистической таблицей* (*англ.* contingency table, cross tabulation) понимают систему строк и столбцов, в которой в определенной последовательности излагается статистическая информация. При обработке результатов физического эксперимента статистической таблице можно сопоставить *кортеж случайных величин*.[4,5] При этом на стадии анализа данных часто неявно предполагается, что экспериментатор имеет дело с независимыми одинаково распределенными случайными величинами, плотность распределения вероятности для которых подчиняется 1.37. Данное предположение должно всегда подвергаться проверке.

Основные статистические показатели можно разделить на две группы: *меры среднего уровня* (*англ.* measures of central tendency) и *меры рассеяния* (*англ.* measures of variability). Первые дают усредненную характеристику совокупности объектов по определенному признаку, а вторые показывают, насколько хорошо данные значения представляют данную совокупность. К мерам среднего уровня относят: среднее значение, эксцесс, асимметрию, минимум, максимум, квантили, медиану, моду и другие подобные показатели. К мерам рассеяния относят, например: дисперсию, размах вариации[6], интерквартильный размах[7], среднее абсолютное отклонение.

[4] *англ.* tuple — кортеж

[5] Например, характеристики экспериментального события, записанные системой сбора данных детектора, или производные от них характеристики.

[6] Размах вариации – разница между максимальным и минимальным показателем.

[7] Интерквартильным размахом (*англ.* interquartile range) называется разность между 0,75- и 0,25-

*Гистограмма* (*англ.* histogram) — способ представления табличных данных в виде столбчатой диаграммы. Она даёт наглядное представление функции плотности вероятности некоторой случайной величины, построенное по выборке. Чаще всего для удобства восприятия ширину прямоугольников берут одинаковой, а высоту определяют по числу элементов выборки, попадающих в соответствующий интервал. Если интервалы разные, то высота прямоугольника обычно выбирается таким образом, чтобы его площадь была пропорциональна числу элементов выборки, которые попали в этот интервал.

Часто группирование данных позволяет резко снизить влияние отдельных наблюдений, не отбрасывая их. Помимо самого распространенного разбиения данных на интервалы равной длины используют:

- разбиение на интервалы равной вероятности, также называемое равночастотным группированием. В результате такого группирования выборки осуществляется максимизация величины информационной энтропии, чем достигается наибольшая асимптотическая мощность критерия согласия $\chi^2$, либо критерия отношения правдоподобия, применяющихся при проверке гипотез о функции распределения.

- разбиение на асимптотически оптимальные интервалы. При таком разбиении минимизируются потери информации в результате группирования, то есть максимизируется информация Фишера.

Примеры гистограмм с равными и асимптотически оптимальными интервалами представлены на рис. 2.2.

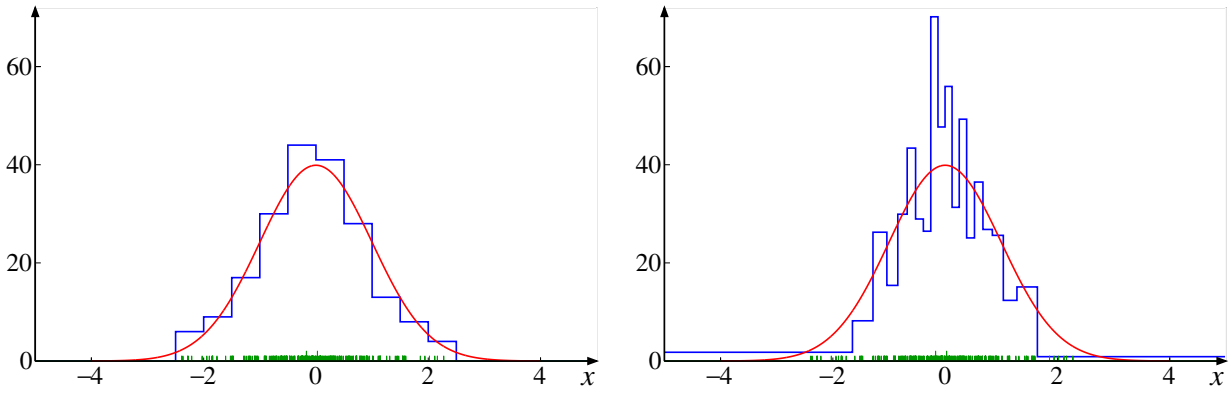


Рис. 2.2: Примеры гистограмм с равными (слева) и асимптотически оптимальными интервалами (справа). На рисунке также показаны распределения индивидуальныъх событий (зеленые линии) и теоретическая плотность вероятности.

*Диаграмма рассеяния* (также точечная диаграмма, *англ.* scatter plot) — математическая диаграмма, изображающая значения двух переменных в виде

квантилями выборочного распределения. Интерквартильный размах является характеристикой разброса распределения величины и является робастным (устойчивым) аналогом дисперсии.

точек чаще всего на декартовой плоскости. На диаграмме рассеяния каждому наблюдению (или элементарной единице набора данных) соответствует точка, координаты которой равны значениям двух каких-то параметров этого наблюдения. Такое графическое представление используют для демонстрации наличия или отсутствия корреляции между двумя переменными.

*График* (*англ.* graph) отображает зависимость одной случайной величины от другой случайной величины (или нескольких). Часто случайные величины приводятся со стандартными отклонениями. При этом по умолчанию считается, что погрешности соответствуют нормальному распределению. Если это не так, например, погрешности являются среднеквадратичным отклонением для равномерно распределенной случайной величины, то это необходимо в явном виде указывать при описании графика.

В случае широких диапазонов, в которых измеряемая величина сильно изменяется, следует помнить, что при графическом представлении результатов анализа ни центр диапазона, ни среднее значение, построенное с учетом плотности распределения, не являются лучшим положением (по $x$) для маркера, показывающего результат измерения. Для каждого из таких широких диапазонов предлагается (см. *NIM A355, 541*) использовать аналитическое или численное решение уравнения:

$$g(x_{lw}) = \frac{1}{\Delta x} \int_{x_{\min}}^{x_{\max}} g(x)\mathrm{d}x. \tag{2.10}$$

Альтернативно, при отображении теоретической кривой можно использовать представление гистограммой с таким же разбиением на интервалы, какое использовалось при анализе экспериментальных данных.

## 2.4 Статистическое оценивание

Пусть неизвестная функция распределения принадлежит некоторому семейству распределений $F(\theta, x)$, зависящему от параметра(ов) ($\theta \in \Theta$), где $\theta$ — множество на прямой (в конечномерном евклидовом пространстве). Нужно по наблюдениям $\mathbf{x} = x_1, \ldots, x_n$ оценить этот параметр (или несколько параметров). Оценка может представлять:

- либо число, предположительно близкое к оцениваемому параметру — *точечная оценка* (*англ.* point estimation),
- либо область $S \in \Theta$ для которой вероятность того, что $S$ будет содержать истинное значение параметра, не меньше чем $1 - \alpha$ (задача построения *доверительной области*; в одномерном случае — *интервальная оценка*, *англ.* interval estimation)[8].

[8] Число $1-\alpha$ называется *уровнем доверия* (*англ.* confidence level) или *доверительной вероятностью*. Оно

Оценка всегда является случайной величиной, так как представляет собой функцию от случайных величин $x_1, \ldots, x_n$.

«Хорошие» точечные оценки обладают следующими свойствами:

- *состоятельности* (*англ.* consistency) — при увеличении числа опытов оценка $\hat{\theta}$ сходится по вероятности к параметру $\theta$;
- *несмещенности* (*англ.* biasness)— ее математическое ожидание совпадает с оцениваемым параметром;
- *эффективности* (*англ.* efficiency) — несмещенная оценка с минимальной дисперсией по сравнению с другими оценками;
- *робастности* — (англ. robustness, от robust — «устойчивый») — свойство статистического метода, характеризующее независимость влияния на результат исследования различного рода выбросов, устойчивости к помехам.[9]

Мощным инструментом оценивания параметров является *метод максимального правдоподобия* (ММП, ML, MLE — *англ.* maximum likelihood estimation). Он основан на предположении, что вся информация о выборке содержится в функции правдоподобия. Точечная оценка максимального правдоподобия:

$$\hat{\theta}_{\mathrm{ML}} = \hat{\theta}_{\mathrm{ML}}(x_1, \ldots, x_n) = \operatorname*{argmax}_{\theta \in \Theta} L(x_1, \ldots, x_n \mid \theta)$$

Если функция правдоподобия дифференцируема, то необходимое условие экстремума — равенство нулю её градиента:

$$g(\theta) = \frac{\partial L(\mathbf{x}, \theta_0)}{\partial \theta} = 0$$

Достаточное условие экстремума может быть сформулировано как отрицательная определённость *гессиана* — матрицы вторых производных:

$$H = \frac{\partial^2 L(\mathbf{x}, \theta_0)}{\partial \theta \partial \theta^T}$$

Отметим, что ММП-оценки, могут быть смещёнными, но являются состоятельными. Приведем пример, показывающий смещенность оценки, полученной методом максимального правдоподобия. Пусть независимая выборка

должно быть достаточно близко к 1.

[9]Выбросы могут возникать по причине ошибки детектора, регистрирующего наблюдения. Также может присутствовать некоторое количество наблюдений, подчиняющихся другому распределению. Оценки медианы более робастны, её оценивание может быть более предпочтительным для распределений с «тяжёлыми» хвостами, если медиана совпадает с мат. ожиданием, например, для симметричных функций плотности распределения вероятности.

$x_1, \ldots, x_n$ получена из непрерывного равномерного распределения, заданного на отрезке $[0, \theta]$. Тогда функция правдоподобия имеет вид:

$$f_L(\mathbf{x} \mid \theta) = \begin{cases} \frac{1}{\theta^n}, & \theta \geq \max(x_1, \ldots, x_n) \\ 0, & \theta < \max(x_1, \ldots, x_n) \end{cases} \quad \Longrightarrow \quad \hat{\theta}_{\text{ML}} = \max(x_1, \ldots, x_n).$$

Такая оценка будет смещенной, так как математическое ожидание для функции распределения $\mathbf{P}\{\hat{\theta}_{\text{ML}} \leq x\} = \left(\frac{x}{\theta}\right)^n$ отличается от оцениваемого параметра $\theta$:

$$\text{M}(\hat{\theta}_{\text{ML}}) = \int_0^\theta x \text{d} \left(\frac{x}{\theta}\right)^n = \frac{n}{n+1}\theta. \tag{2.11}$$

Другим примером является случай выборки объемом $n$, полученной из нормального распределения $\text{Gauss}_{\mu,\sigma}(x)$. Для этого случая $\mu$ и $\sigma^2$ составляют неизвестный вектор параметров, а $\widehat{\mu}_{\text{ML}}$ и $\widehat{\sigma^2}_{\text{ML}}$ — вектор, соответствующий ММП-оценке этих параметров. Логарифмическая функция правдоподобия для этого случая имеет вид:

$$L(\mathbf{x} \mid \mu, \sigma^2) = \ln \left[ \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \frac{(x_i - \mu)^2}{2\sigma^2} \right] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Чтобы найти её максимум, приравняем к нулю частные производные:

$$\begin{cases} \dfrac{\partial}{\partial \mu} L(\mathbf{x} \mid \mu, \sigma^2) = 0 \\ \dfrac{\partial}{\partial \sigma^2} L(\mathbf{x} \mid \mu, \sigma^2) = 0 \end{cases} \Rightarrow \begin{cases} \dfrac{\sum\limits_{i=1}^n x_i - n\mu}{\sigma^2} = 0 \\ -\dfrac{n}{2\sigma^2} + \dfrac{\sum\limits_{i=1}^n (x_i - \mu)^2}{2\left(\sigma^2\right)^2} = 0 \end{cases},$$

откуда можно получить ММП-оценку параметров: $\hat{\mu}_{\text{ML}} = \bar{x}$, $\widehat{\sigma^2}_{\text{ML}} = s_n^2$, где $\bar{x}$ и $s_n^2$ — выборочное среднее и дисперсия:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \tag{2.12}$$

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \tag{2.13}$$

Полученная оценка дисперсии является смещенной, так как из свойств математического ожидания следует, что

$$\mathbf{M}(s_n^2) = \frac{n-1}{n} \sigma^2. \tag{2.14}$$

## 2.5 Оценки моментов распределений

Рассмотрим точечные оценки моментов для распределения произвольной формы. Выборочное среднее, определяемое по 2.12, является несмещенной оценкой среднего генеральной совокупности всегда, когда последнее существует.

В случае зашумленности сигнала оценка медианы распределения оказывается более устойчивой, чем выборочное среднее. Если объем упорядоченной по возрастанию выборки — нечетное число ($n = 2k + 1$), то *выборочной медианой* (*англ.* sample median) называют элемент выборки $m = x_{k+1}$, а если четное ($n = 2k$), то $m = \frac{1}{2}(x_k + x_{k+1})$. Для выборки из распределения с плотностью вероятности $f(x)$, распределение выборочной медианы асимптотически нормально со средним $m$ и дисперсией $1/(4nf(m)^2)$.

Если выборка независимая, то для 2.12 можно примерить 1.13, получив таким образом *дисперсию среднего* (*англ.* variance of the sample mean):

$$\sigma_{\bar{x}}^2 = \mathbf{M}((\bar{x} - \mu)^2) = \frac{\sigma^2}{n}. \tag{2.15}$$

Подчеркнем, что полученное выражение не зависит от формы функции распределения случайной величины, а лишь требует существования её дисперсии. Флуктуации независимых измерений компенсируют друг друга.

*Оценка дисперсии распределения* зависит от того, известен ли первый момент распределения $\mu$, или же он также оценивается по выборке. В первом случае оценка $v_\mu^2 = \frac{1}{n}\sum(x_i - \mu)^2$ будет несмещенной оценкой дисперсии, т.е. $\mathbf{M}(v_\mu^2) = \sigma^2$. Во втором случае для оценки дисперсии вместо неизвестного $\mu$ используют выборочное среднее $\bar{x}$. Рассмотрим выражение

$$v^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.16}$$

Добавив и отняв $\mu$, получим:

$$v^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \mu) - (\bar{x} - \mu)]^2 = \tag{2.17}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - 2(\bar{x} - \mu)\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) + \frac{n}{n}(\bar{x} - \mu)^2 = \tag{2.18}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - (\bar{x} - \mu)^2 \tag{2.19}$$

$$= v_\mu^2 - (\bar{x} - \mu)^2 \tag{2.20}$$

Математическое ожидание этой величины равно:

$$\mathbf{M}(v^2) = \mathbf{M}(v_\mu^2) - \mathbf{M}((\bar{x} - \mu)^2) = \sigma^2 - \frac{\sigma^2}{n} \tag{2.21}$$

То есть величина $v^2$ является смещенной оценкой дисперсии. Тогда величина

$$s^2 = \frac{n}{n-1}v^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.22}$$

может использоваться в качестве несмещенной оценки дисперсии. Множитель перед $v^2$ называют *коррекцией Бесселя*.

Важно понимать, что, взяв квадратный корень из выражения для несмещенной оценки дисперсии, мы получаем смещенную оценку среднеквадратического отклонения! Это возникает вследствие *неравенства Йенсена*, которое гласит, что для выпуклой функции[10] $\phi(x)$:

$$\phi(tx_1 + (1-t)x_2) \geq t\phi(x_1) + (1-t)\phi(x_2). \tag{2.23}$$

На языке теории вероятностей $\phi(\mathbf{M}(x)) \geq \mathbf{M}(\phi(x))$. Смещение возникает из-за того, что квадратный корень — выпуклая функция. Для нормально распределенных случайных величин смещение можно устранить, домножив корень из выражения 2.22 на корректирующий множитель[11] $\sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}$, где $\Gamma$ — *гамма-функция Эйлера*.

Примером неравенства Йенсена для вогнутых функций может быть математическое ожидание для объема $V$ сферы с радиусом $r$, распределенным согласно нормальному распределению $f(r) = \mathrm{Gauss}_{r_0,\sigma}(r)$:

$$\mathbf{M}(V) = \int_{-\infty}^{\infty} V(r)f(r)\mathrm{d}r = \frac{4}{3}\pi(r_0^3 + 3\sigma^2 r_0). \tag{2.24}$$

Так как оценка дисперсии является случайной величиной, можно задаться вопросом о её дисперсии. Для случая независимых одинаково распределенных случайных величин:

$$\mathbf{V}(s^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right), \tag{2.25}$$

где $\mu_4$ — четвертый центральный момент. На практике использование этого выражения встречается с проблемой, что вместо величины $\mu_4$ нужно подставлять оценку для неё, которая оказывается, зачастую, смещённой. Для нормально распределенных величин относительная погрешность дисперсии равна $1/\sqrt{2n}$.

Для двумерного распределения коэффициент выборочной корреляции можно оценить по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{2.26}$$

[10] Для вогнутых функций имеет место обратное неравенство.

[11] Такая коррекция имеет смысл в случае очень малого объема выборки.

Существуют методы оценки более высоких моментов распределений, но на практике они редко применяются при обработке результатов физического эксперимента.

# 3 Интервальные оценки

## 3.1 Доверительный интервал

Помимо точечной (лучшей) оценки измеряемой величины научная публикация должна содержать оценку погрешности проведенного измерения. Поэтому фактически всегда публикуется *интервальная оценка*. В математической статистике метод построения интервальной оценки зависит от интерпретации понятия вероятность. При этом различия в методах отражаются на уровне терминологии:

- в частотной интерпретации говорят о *доверительном интервале* (*англ.* confidence interval),
- в байесовском подходе о *байесовском доверительном интервале* (*англ.* credible interval).

Для некоторых (специальных) случаев границы интервальных оценок, полученные в байесовском и частотном подходах, совпадают.

Доверительный интервал $(\theta_L, \theta_U)$ для неизвестного (фиксированного) параметра $\theta$ — численный интервал, предположительно (то есть, с той или иной степенью доверительной вероятности) содержащий истинное значение параметра $\theta$. Помимо этого доверительный интервал должен удовлетворять следующим условиям:

1. для любой случайной величины доверительный интервал не должен быть пуст;
2. он не должен содержать пропусков;
3. конечные точки интервала ему принадлежат.

Интервал конструируется в смысле фиксированной частоты попадания параметра $\theta$ в построенный промежуток при бесконечном повторении эксперимента с последующей процедурой определения доверительного интервала. С точки зрения выборки из генеральной совокупности доверительный интервал несет информацию о параметре генеральной совокупности, но не о выборке из неё, и не об индивидуальном объекте. Например, пусть 95% доверительный интервал для ежедневного времени просмотра телевизора американцами составляет (2.69, 6.04) часов. При этом нельзя утверждать, что:

- 95% всех американцев смотрят телевизор в этом диапазоне;
- или что 95% из исследованной выборки делают это.

Можно лишь утверждать, что среднее значение времени просмотра лежит в этом диапазоне с соответствующей доверительной вероятностью. В частотном подходе доверительная вероятность "вступает в игру" до того момента, как начинается сбор данных! Данные, которые только еще будут собраны, позволят построить доверительный интервал, содержащий истинное значение интересующего нас параметра.

Важным свойством доверительного интервала является так называемый *охват вероятности* (*статистическое покрытие*, *англ.* statistical coverage) — то есть вероятность попадания истинного значения параметра в построенный интервал.

Еще раз подчеркнем важное положение частотного подхода, что параметр не является случайной величиной, то есть для него в рамках этого подхода нельзя ввести распределение вероятности.

Рассмотрим *метод Неймана* (*англ.* Neyman construction), который применяется для построения двустороннего частотного доверительного интервала для параметра. Пусть $f(x, \theta)$ — некоторое зависящее от параметра $\theta$ семейство плотностей распределения вероятности. Используя $f(x, \theta)$, для каждого значения параметра, мы можем построить интервал такой, что

$$\mathbf{P}(x_1 < x < x_2, \theta) = \int_{x_1(\theta,\alpha)}^{x_2(\theta,\alpha)} f(x, \theta)\mathrm{d}x \geq 1 - \alpha. \tag{3.1}$$

В случае дискретного распределения интеграл заменяется суммой. Заметим, что, хотя метод Неймана и не требует выполнения условия

$$\int_{-\infty}^{x_1(\theta,\alpha)} f(x, \theta)\mathrm{d}x = \int_{x_2(\theta,\alpha)}^{\infty} f(x, \theta)\mathrm{d}x = \frac{\alpha}{2}, \tag{3.2}$$

разумно придерживаться его выполнения. Семейство таких интервалов для различных значений параметра $\theta$ образуют доверительный пояс (*англ.* confidence belt), см рис. 3.1. Обычно левая и правая границы этого пояса являются непрерывными монотонными функциями. После выполнения процедуры измерений экспериментатор получает наблюдаемое значение $x_0$. Если отложить вертикальную линию на плоскости $x$-$\theta$, то она пересечет правую и левую границы пояса в точках $\theta_-$ и $\theta_+$, которые будут образовывать интервал по оси $\theta$. Истинное значение параметра ($\theta_t$) будет лежать между $x_1$ и $x_2$ тогда и только тогда, когда оно принадлежит $[\theta_-, \theta_+]$, и

$$\mathbf{P}(x_1(\theta) < x < x_2(\theta)) = \mathbf{P}(\theta_-(x) < \theta < \theta_+(x)) = 1 - \alpha. \tag{3.3}$$

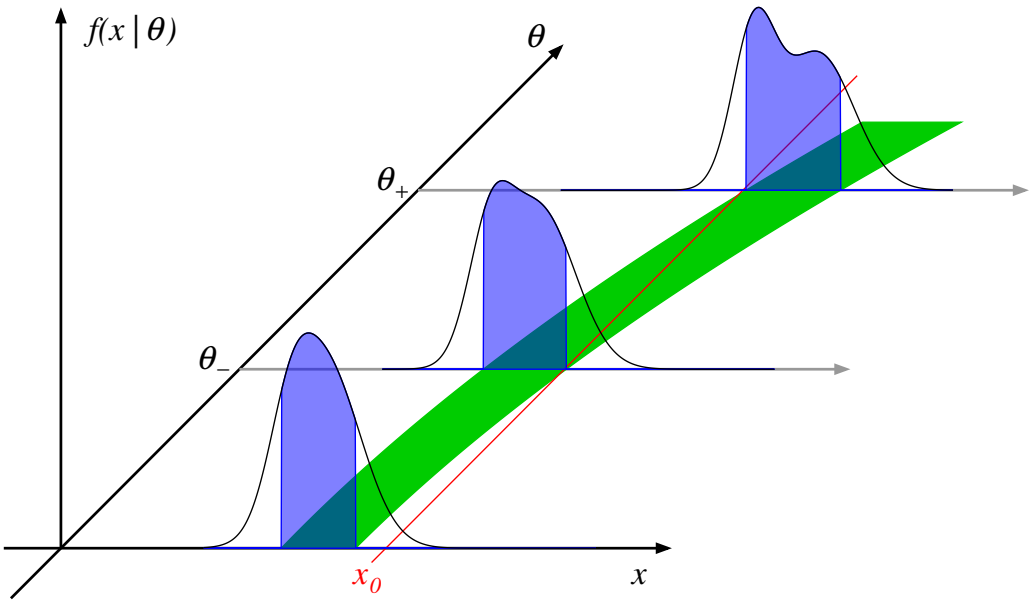


Рис. 3.1: Построение доверительного интервала методом Неймана.

Еще раз подчеркнем, что условие 3.2 не является обязательным. Оно может быть заменено на другое (см., например, метод Фельдмана-Казинса, описанный в разделе 4.1).

Доверительный интервал сам по себе является статистикой (функцией от элементов выборки), то есть случайной величиной. Рассмотрим серию псевдоэкспериментов, для которой при фиксированном исходном значении параметра $\theta$ определена серия доверительных интервалов. Заметим, что, во-первых, интервалы в ансамбле имеют различную ширину, а, во-вторых, некоторые из них не охватывают истинное значение параметра. Доля таких интервалов стремится к $\alpha$ при увеличении числа псевдоэкспериментов.

В байесовском подходе (в отличии от частотного) исследуемый параметр сам рассматривается как случайная величина. Для этой случайной величины получают распределение вероятности. Для построения байесовского доверительного интервала на основании полученных данных используют теорему Байеса, в которую подставляют функцию правдоподобия[1] и априорное распределение вероятности (*англ.* prior). В результате получается диапазон, которому (с заданной вероятностью) принадлежит исследуемый параметр.

Критики байесовского метода указывают на то, что изначально неизвестно, каким априорным распределением следует задаваться. Эта задача отдается на откуп экспериментатора. При этом полученный результат зависит от выбора априорного распределения. Также к недостаткам этого подхода относится то, что в нем часто не обеспечивается статистическое покрытие для заданной вероятности.

[1] *Принцип правдоподобия* (*англ.* likelihood principle) — предположение, согласно которому функция правдоподобия содержит в себе всю доступную (из экспериментальных данных) информацию об исследуемом параметре.

## 3.2 Оценки параметров нормального распределения

Рассмотрим построение доверительного интервала для параметров нормального распределения. Задачу можно сформулировать, например, следующим образом:

- пусть измеряется некоторая физическая величина $x_t$,
- измерения проводятся при помощи прибора с некоторым разрешением $\sigma$,
- известно, что распределение функции разрешения прибора — нормальное несмещенное $\mathrm{Gauss}_{0,\sigma}(x)$,
- Проведено $n$ измерений и получена выборка $x_1, \ldots, x_n$.

Требуется определить двухсторонний симметричный доверительный интервал для исследуемой величины, который содержит $x_t$ с заданной доверительной вероятностью равной $1 - \alpha$.

### Случай известного $\sigma$

В этом случае величина $(\bar{x} - x_t)/(\sigma/\sqrt{n})$ подчиняется стандартному нормальному распределению. Для определения доверительного интервала с доверительной вероятностью $1 - \alpha$ следует получить $\alpha/2$- и $(1 - \alpha/2)$-квантили нормального распределения — $q_{\alpha/2}$ и $q_{1-\alpha/2}$. Тогда

$$\mathbf{P}\left(\overline{x} + q_{\alpha/2}\frac{\sigma}{\sqrt{n}} < x_t < \overline{x} + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \xrightarrow{n\to\infty} 1 - \alpha, \tag{3.4}$$

что и определяет доверительный интервал. Из симметричности нормального распределения следует симметрия доверительного интервала $\bar{x} \pm q_{\alpha/2}\sigma/\sqrt{n}$.

### Случай неизвестного $\sigma$

Если параметр $\sigma$ неизвестен, то для построения статистики требуется использовать точечную несмещенную оценку дисперсии ($s^2$). Величина $(\bar{x} - x_t)/(s/\sqrt{n})$ подчиняется *распределению Стьюдента с* $n-1$ *степенями свободы* (*англ.* Student's t-distribution). Это распределение (для случая $k$ степеней свободы) задается плотностью вероятности:

$$\mathrm{Student}_k(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}}. \tag{3.5}$$

Это симметричное распределение. При $k = 1$ оно переходит в распределение Коши, а при больших $k$ стремится к нормальному. Примеры распределения Стьюдента представлены на рис. 3.2, слева.

Для выборки из нормального распределения с неизвестным средним и дисперсией следует найти квантили распределения Стьюдента с $n - 1$ степенью свободы (в силу симметрии $t_{1-\alpha/2} = -t_{\alpha/2}$) такие, что

$$1 - \alpha = \int_{t_{\alpha/2}}^{t_{1-\alpha/2}} \mathrm{Student}_{n-1}(x)\mathrm{d}x. \tag{3.6}$$

Тогда

$$\mathbf{P}\left(t_{\alpha/2} < \frac{\bar{x} - x_t}{s/\sqrt{n}} < t_{1-\alpha/2}\right) \xrightarrow{n\to\infty} 1 - \alpha, \tag{3.7}$$

что эквивалентно доверительному интервалу

$$\mathbf{P}\left(\overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} < x_t < \overline{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right) \xrightarrow{n\to\infty} 1 - \alpha. \tag{3.8}$$

### Доверительный интервал для дисперсии

Для выборки объемом $n$ из нормального распределения величина $s^2(n-1)/\sigma^2$ подчиняется *распределению хи-квадрат с* $n-1$ *степенью свободы* (*англ.* chi-squared distribution). Для случая $k$ степеней свободы это распределение задается плотностью вероятности

$$\chi^2_k(x) = \frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1}e^{-x/2}. \tag{3.9}$$

Доверительный интервал для $\sigma$ определяется соотношениями:

$$s_{\min} = \frac{s\sqrt{n-1}}{\chi^2_{n-1,1-\alpha/2}}, \qquad s_{\max} = \frac{s\sqrt{n-1}}{\chi^2_{n-1,\alpha/2}}, \tag{3.10}$$

где $\chi^2_{n-1,p}$ — это $p$-квантиль распределения хи-квадрат с $n-1$ степенью свободы. Примеры распределения хи-квадрат представлены на рис. 3.2, справа.

## 3.3 Биномиальное распределение и его свойства

Часто исследователь имеет дело с изучением структуры семейств случайных величин $x_t$, где $t \in T$ — некоторый параметр. *Реализацией* (*выборочной функцией, англ.* index set, parameter set) такого *случайного процесса* (*англ.*

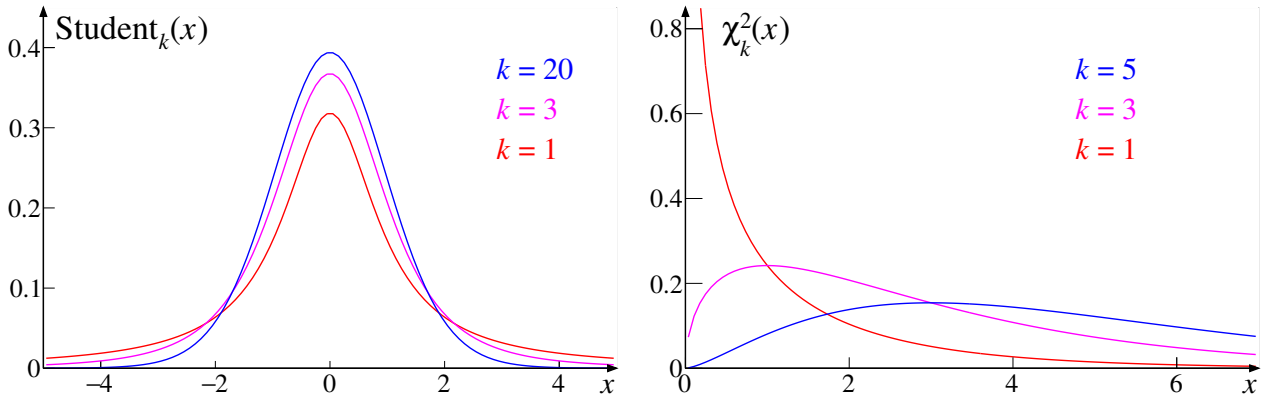


Рис. 3.2: Плотности распределения Стьюдента (слева) и хи-квадрат (справа) для разных степеней свободы $k$.

stochastic process) является функция, ставящая в соответствие каждому $t$ одно из возможных значений $x_t$. Множество параметров $T$ может быть как дискретным, так и непрерывным. Случайные процессы, у которых $T = [0, \infty)$, важны в приложениях. Для них $t$ интерпретируется как время. В общем случае случайные величины $x_t$ являются зависимыми, однако в этом пособии будут рассмотрены случайные процессы только с независимыми $x_t$.

Простейшим примером случайного процесса является *схема Бернулли* (*англ.* Bernoulli trial, binomial trial). Пусть производится $n$ независимых экспериментов, в каждом из которых с вероятностью $p$ может произойти некоторое событие $A$. Например, событие может состоять в выпадении "орла", которому будет ставится в соответствие исход 1 (*благоприятный исход*), или "решки", исход 0 (*неблагоприятный исход*), тогда вероятность $p = 1/2$. Если же благоприятный исход — выпадение "шестерки" при бросании игральной кости, а неблагоприятный исход — выпадение любой другой грани, то $p = 1/6$. В этом примере $T = [1, \ldots, n]$, а множество значений $x_t$ составляет $\{0, 1\}$. Найдем вероятность того, что в $n$ испытаниях $A$ реализуется ровно $m$ раз. Вероятность того, что $A$ реализуется в первых $m$ испытаниях, а в оставшихся $n-m$ произойдет неблагоприятный исход, составляет $p^m(1-p)^{n-m}$ (т.к. события независимы). Такой порядок реализации благоприятного исхода является лишь одним из $C_n^m$ возможных способов[2], следовательно, полная вероятность равна

$$\mathbf{P}(m) = \text{Binomial}_p(n, m) = C_n^m p^m (1-p)^{n-m}. \tag{3.11}$$

Множество вероятностей $\{\mathbf{P}(m)\}$ называется *биномиальным распределением* (*англ.* binomial distribution). Оно дискретно и

$$\sum_{m=0}^{n} \mathbf{P}(m) = (p + (1-p))^n = 1. \tag{3.12}$$

[2] $C_n^m$ — биномиальный коэффициент, $C_n^m = \frac{n!}{m!(n-m)!}$ для целых $m$ и $n$, $0 \le m \le n$.

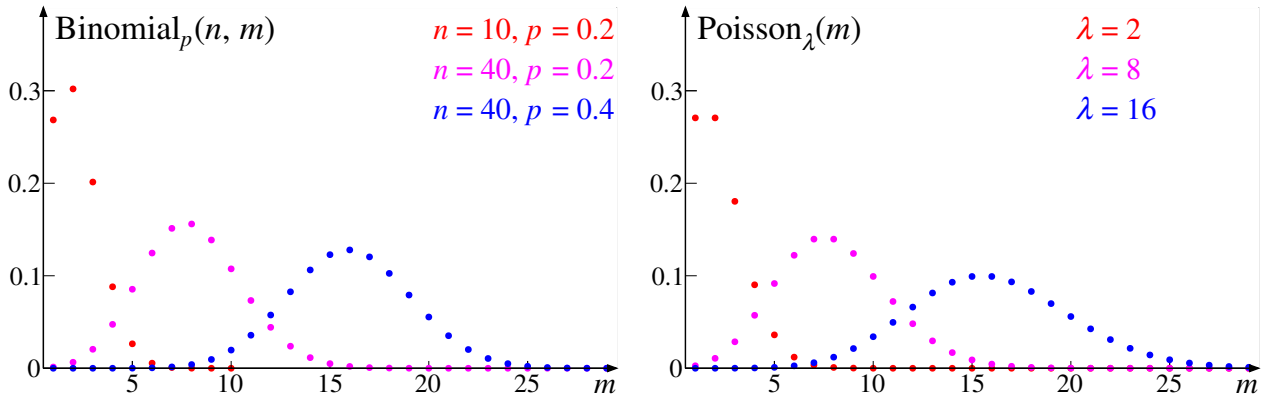


Рис. 3.3: Примеры плотности вероятности для биномиального распределения (слева) и распределения Пуассона (справа).

Для одиночного испытания ($n = 1$) биномиальное распределение называют *распределением Бернулли* (*англ.* Bernoulli distribution).

Среднее число событий $A$ в серии из $n$ испытаний составляет

$$\mu = \sum_{m=0}^{n} m \times \text{Binomial}_p(n, m) = np, \tag{3.13}$$

а дисперсия числа появлений события $A$ в $n$ испытаниях равна

$$\sigma^2 = \sum_{m=0}^{n} (m - \mu)^2 \times \text{Binomial}_p(n, m) = np(1 - p). \tag{3.14}$$

Примеры плотности вероятности для биномиального распределения с разными значениями параметров $p$ и $n$ приведены на рис. 3.3, слева.

### 3.3.1 Примеры применения биномиального распределения

#### Эффективность счетчика Гейгера-Мюллера

Пусть счетчик Гейгера-Мюллера с эффективностью 90% ($p = 0.9$) зарегистрировал $m$ частиц из $n = 1000$ частиц, попавших в него. Среднее число срабатываний будет составлять $np = 900$. Среднеквадратическое отклонение от этого числа $\sqrt{np(1-p)} = \sqrt{90} \approx 9.5$. Наблюдаемая эффективность счетчика $\varepsilon = m/n$ будет флуктуировать с $\sigma_\varepsilon = \sigma/n \approx 0.0095$.

#### Точность интегрирования методом Монте-Карло

Допустим, мы хотим вычислить значение числа $\pi$ методом Монте-Карло. Для этого случайным образом заполняем квадрат 2 на 2 (общая площадь $4\,\text{см}^2$),

центрированный относительно начала координат, $n$ точками, проекции которых на оси независимы и равномерно распределены. Число точек, попавших внутрь окружности радиусом $1\,\text{см}^2$, будет равно $m = np$, где $p = \pi/4$. Если мы хотим достичь точности расчетов 1%, то есть

$$\frac{\sigma}{np} = \frac{\sqrt{np(1-p)}}{np} = 0.01,$$

то нам потребуется $n = (1-p)/(0.01^2 p) = (4-\pi)/0.01^2\pi \approx 2732$ точек.

#### Флуктуация аксептанса для взвешенных событий

Аксептанс сложного детектора определяется методом Монте-Карло. Можно ввести плотность вероятности $f_0(\vec{x})$, где $\vec{x}$ — полный набор необходимых кинематических переменных. Чтобы избежать повторного моделирования разных физических процессов (различных сечений) определяемых распределением $f(\vec{x})$, разумно приписать каждому индивидуальному событию вес: $w_i = f(\vec{x})/f_0(\vec{x})$, где $i$ — номер события. Аксептанс для каждого события ($\varepsilon_i$) является либо нулем, либо единицей (событие зарегистрировано или нет). Таким образом, полный аксептанс для физического процесса будет вычисляться как:

$$\varepsilon_\text{T} = \frac{\sum w_i \varepsilon_i}{\sum w_i}. \tag{3.15}$$

Дисперсия каждого члена суммы в числителе есть $w_i\varepsilon_i(1-\varepsilon_i)$. Тогда дисперсия для случайной величины определяемой выражением 3.15 будет вычисляться по формуле

$$\sigma^2_\text{T} = \frac{\sum w_i^2 \varepsilon_i(1-\varepsilon_i)}{(\sum w_i)^2}. \tag{3.16}$$

### 3.3.2 Интервальные оценки для биномиального распределения

#### Аппроксимация нормальным распределением

В силу центральной предельной теоремы при больших $n$ биномиальное распределение сходится к нормальному распределению с параметрами $\mu = np$ и $\sigma = \sqrt{np(1-p)}$. Этот факт можно использовать для интервальной оценки эксперимента из $n$ испытаний, в котором было зафиксировано $m$ благоприятных исходов. Тогда доверительный интервал будет определяться как

$$p_{L,U} = \hat{p} \pm z\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \tag{3.17}$$

где $\hat{p} = m/n$ — точечная оценка вероятности, а $z$ — $(1 - \frac{\alpha}{2})$-квантиль *стандартного нормального распределения*[3]. Так как этот метод приблизительный, он подвержен абберациям, таким как выход в нефизические области. Кроме того, в случае в случае малых $m$ аппроксимация нормальным распределением имеет недостаточный охват вероятности.

Достаточное покрытие обеспечивается следующей (приблизительной) формулой:

$$p_L = \frac{2n\hat{p} + z^2 - 1 - z\sqrt{z^2 - 2 - 1/n + 4\hat{p}(n(1-\hat{p}) + 1)}}{2(n + z^2)}, \quad (3.18)$$

$$p_U = \frac{2n\hat{p} + z^2 + 1 + z\sqrt{z^2 + 2 - 1/n + 4\hat{p}(n(1-\hat{p}) - 1)}}{2(n + z^2)}, \quad (3.19)$$

которая относительно легко программируется.

### Метод Клоппера-Пирсона

Точным методом определения доверительного интервала для биномиального распределения является метод Клоппера-Пирсона (*англ.* Clopper–Pearson interval), предложенный в 1934 году. Точный доверительный интервал, содержащий все значения $p$, не отвергнутые на уровне $\alpha$, можно определить из равенств

$$\sum_{k=m,0}^{n,m} C_n^k p_{L,U}^k (1 - p_{L,U})^{n-k} = \alpha/2. \quad (3.20)$$

Можно показать, что функция распределения биномиального распределения с параметрами $(p,\, n)$ записывается как:

$$F(k;\, n, p) = \Pr(x \le k) = I_{1-p}(n - k, k + 1) = 1 - I_p(k + 1, n - k). \quad (3.21)$$

где $I_x(a,\ b)$ — регуляризованная неполная бета-функция:

$$I_x(a,\ b) = \frac{\mathrm{B}(x; a,\ b)}{\mathrm{B}(a,\ b)}, \quad (3.22)$$

определяемая через неполную бета-функцию

$$\mathrm{B}(x;\, a, b) = \int_0^x t^{a-1} (1 - t)^{b-1}\, dt, \quad (3.23)$$

[3] нормальное распределение с параметрами $\mu = 0$ и $\sigma = 1$

и полную бета-функцию $\mathrm{B}(a,b) = \mathrm{B}(1;\, a, b)$. Соответственно, $p_{L,U}$ определяются, через квантили *бета-распределения*[4] (*англ.* beta distribution):

$$p_L = B(\frac{\alpha}{2}; m, n-m+1) \tag{3.24}$$

$$p_U = B(1-\frac{\alpha}{2}; m+1, n-m) \tag{3.25}$$

Бета-распределение, в свою очередь, связано с *распределением Фишера* (*англ.* F-distribution, Fisher–Snedecor distribution), для которого существуют обширные статистические таблицы. Границы доверительного интервала определяются как:

$$p_L = \left(1 + \frac{n-m+1}{m\, F\left[\frac{\alpha}{2}; 2m, 2(n-m+1)\right]}\right)^{-1}, \tag{3.26}$$

$$p_U = \left(1 + \frac{n-m}{(m+1)\; F\left[1-\frac{\alpha}{2}; 2(m+1), 2(n-m)\right]}\right)^{-1}, \tag{3.27}$$

где $F(x; k, l)$ — квантили распределения Фишера, которое для случая степеней свободы $k$ и $l$ задается плотностью вероятности

$$\mathrm{Fisher}_{k,l}(x) = \frac{\sqrt{\frac{(kx)^k\; l^l}{(kx+l)^{k+l}}}}{x\,\mathrm{B}\left(\frac{k}{2}, \frac{l}{2}\right)}. \tag{3.28}$$

### Случай $m = 0$

Важным частным случаем является отсутствие благоприятных (или неблагоприятных) исходов. Очевидно, что в этом случае интервал односторонний, т.е. $p_L = 0$. В случае больших $n$ верхняя граница 95-процентного доверительного интервала хорошо аппроксимируется *правилом тройки* (*англ.* rule of three):

$$p_U \approx \frac{3}{n}. \tag{3.29}$$

Действительно, требуется:

$$(1-p)^n = 0.05.$$

Логарифмируя обе части равенства, получаем

$$n \ln(1-p) = \ln 0.05 \approx -3.$$

При этом $\ln(1-p) \approx -p$ для малых $p$, что уже выполняется с достаточной точностью для $n > 20$. Правило тройки следует использовать только для прикидочных вычислений.

[4] плотность вероятности бета-распределения задается формулой $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$

### 3.3.3 Байесовское оценивание параметра распределения Бернулли

Рассмотрим набор из $m = 500$ деталей, предназначенных для проверки. Пусть исход каждой такой проверки $y_i$ принимает значение 0, если деталь соответствует стандарту, и 1 в случае брака ($i = 1, \ldots, n$). Пусть $\theta$ — вероятность того, что случайно выбранная деталь имеет дефект. Каждое наблюдение можно представить распределением Бернулли:

$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}.$$

Набору независимых измерений $\vec{y} = [y_1, \ldots, y_n]$ может быть сопоставлена вероятность, пропорциональная каждой $p(y_i|\theta)$. Таким образом вероятностная модель записывается как

$$p(\vec{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}.$$

Вводя среднее $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ имеем

$$p(\vec{y}|\theta) = \theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})}.$$

Полученная вероятностная модель отражает функцию правдоподобия данных, относительно гипотезы $\theta$. Для получения байесовской апостериорной вероятности необходимо учесть априорную вероятность для этого параметра, который в байесовском подходе является случайной величиной.

$$p(\theta|\vec{y}, I) = \frac{p(\theta|I)p(\vec{y}|\theta, I)}{p(\vec{y}|I)},$$

где $I$ — факты, известные до проведения измерения, а $p(\vec{y}|I)$ вычисляется как

$$p(\vec{y}|I) = \int p(\vec{y}|I)p(\vec{y}|\theta, I)d\theta.$$

Если до проведения измерений никакой информации о проценте брака не было, то априорную плотность распределения ($p(\theta|I)$) разумно принять за равномерное от 0 до 1.

Вычислив функцию $p(\theta|\vec{y}, I)$, можно проводить (байесовское) точечное и интервальное оценивание для неё. При этом

$$\hat{\theta} = \int \theta p(\theta|\vec{y}, I)d\theta,$$

а доверительный интервал задается, например, условием

$$\int_0^{\theta_L} p(\theta|\vec{y}, I)d\theta = \int_{\theta_U}^{1} p(\theta|\vec{y}, I)d\theta = \alpha.$$

Если нужно произвести уточнение значения параметра $\theta$, то необходимо выполнить измерения и повторить процедуру, применив в качестве априорного распределения вероятности полученное на предыдущем этапе апостериорное распределение.

# 3.4 Распределение Пуассона

## 3.4.1 Процесс Пуассона

*Пуассоновский процесс* (*англ.* Poisson process) — поток однородных событий, для которого:

1. число событий в некотором временном интервале $t$ не зависит от числа событий в любых интервалах, не пересекающихся с $t$;
2. вероятность отдельного события за малый интервал времени $\Delta t$ пропорциональна длительности интервала, т.е. вероятность события на промежутке $(t, t + \Delta t)$ равна $r\Delta t + o(\Delta t)$, где $o(\Delta t) \ll \Delta t$ [5];
3. вероятность двух и более событий за $(t, t + \Delta t)$ есть $o(\Delta t)$.

Определим вероятность того, что за промежуток времени $(0, t + \Delta t)$ не произойдет ни одного события, т.е. их не будет в интервалах $(0, t)$, ни в $(t, t + \Delta t)$.

$$\mathbf{P}_0(t + \Delta t) = \mathbf{P}_0(t)(1 - r\Delta t + o(\Delta t)). \tag{3.30}$$

Получаем дифференциальное уравнение

$$\frac{\mathbf{P}_0(t + \Delta t) - \mathbf{P}_0(t)}{\Delta t} = -r\mathbf{P_0}(t) \qquad \Rightarrow \qquad \frac{d\mathbf{P}_0(t)}{dt} = -r\mathbf{P}_0(t). \tag{3.31}$$

с граничным условием $\mathbf{P}_0(0) = 1$. Его решение

$$\mathbf{P}_0(t) = e^{-rt}. \tag{3.32}$$

Отсюда сразу следует, что распределение интервалов времени между событиями пуассоновсокого процесса — экспоненциальное распределение (*англ.* exponential distribution):

$$\mathrm{Exp}_\lambda(x) = \lambda e^{-\lambda x} \tag{3.33}$$

Действительно, событие не должно произойти в промежуток $(0, t)$ (вероятность этого определяется 3.32), но обязано должно случиться в $(t, t + \Delta t)$. Вероятность последнего равна $r\Delta t$ из условий, описывающих пуассоновский поток. Таким образом, средний промежуток времени между событиями равен $\bar{t} = \mathbf{M}(\mathrm{Exp}_r(t)) = 1/r$. Также заметим, что мода экспоненциального

[5] $r$ интерпретируется как число событий в единицу времени

распределения равна нулю, т.е. наиболее вероятный временной интервал для пуассоновского процесса также равен нулю.

Важно понимать, что выбор случайной точки на временной прямой не эквивалентен выбору случайного временного интервала между событиями пуассоновского процесса. Чем длиннее интервал, тем больше вероятность попадания в него. Можно определить вес каждого интервала как $t/\bar{t} = rt$. Таким образом, распределение временных интервалов между случайным моментом времени и наступлением события оказывается $I_s(t) = r^2 t e^{-rt}$. В среднем имеем

$$\bar{t}_s = \int_0^\infty t I_s(t) dt \; / \int_0^\infty I_s(t) dt = 2/r, \tag{3.34}$$

что вдвое больше $\bar{t}$. О том, что среднее время от начала измерения до регистрации первого события в два раза больше, чем величина, обратная скорости счета, стоит помнить, когда проводится поиск очень редких событий.

Вернемся к выводу распределения для пуассоновского процесса. Вероятность того, что за время от 0 до $t + \Delta t$ произойдет ровно одно событие записывается как

$$\mathbf{P}_1(t + \Delta t) = \mathbf{P}_1(t)(1 - r\Delta t) + \mathbf{P}_0(t) r \Delta t, \tag{3.35}$$

здесь опущены члены содержащие $o(\Delta t)$. Отсюда получаем дифференциальное уравнение

$$\frac{d\mathbf{P}_1(t)}{dt} = -r\mathbf{P}_1(t) + re^{-rt}, \tag{3.36}$$

решением которого будет $\mathbf{P}_1(t) = rte^{-rt}$. Такую схему рассуждений можно продолжить для любого $m > 0$:

$$\frac{d\mathbf{P}_m(t)}{dt} = -r\mathbf{P}_m(t) + r\mathbf{P}_{m-1}(t), \tag{3.37}$$

Решение для общего случая имеет вид

$$\mathbf{P}_m(t) = \frac{(rt)^m}{m!} e^{-rt} = \text{Poisson}_{rt}(m), \tag{3.38}$$

причем оказывается, что $\sum_{m=0}^{\infty} \mathbf{P}_m(t) = 1$ так как вероятность зарегистрировать любое число событий равна 1. Распределение 3.38 называют *распределением Пуассона* (*англ.* Poisson distribution). Примеры плотности вероятности для распределения Пуассона с разными значениями параметра $\lambda$ приведен на рис. 3.3, справа.

## Закон редких событий

Распределение Пуассона является предельной формой биномиального распределения, когда

$$\lim_{n\to\infty} np = \lambda. \tag{3.39}$$

Этот факт иногда называют *законом редких событий* (*англ.* law of rare events). Действительно, биномиальное распределение можно факторизовать следующим образом:

$$\text{Binomial}_p(n,m) = \frac{n!}{n^m(n-m)!}\frac{(np)^m}{m!}\Big(1-\frac{np}{n}\Big)^n\Big(1-\frac{np}{n}\Big)^{-m}. \tag{3.40}$$

При выполнении условия 3.39 первый и четвертый сомножители стремятся к единице, третий к $e^{-\lambda}$, а второй к $\lambda^m/m!$. Таким образом пуассоновский процесс возникает в форме, где временной параметр $rt$ заменяется пространственным параметром $\lambda$.

$$\text{Poisson}_\lambda(m) = \frac{\lambda^m}{m!}e^{-\lambda}. \tag{3.41}$$

Можно проиллюстрировать эту ситуацию следующим примером. Рассмотрим систему точек в евклидовом пространстве $E$. Пусть $N_S$ обозначает число точек, содержащихся в области $S$ этого пространства. Предположим, что $N_S$ является случайной величиной. Совокупность $\{N_S\}$, где область изменения индекса $S$ состоит из всех возможных подмножеств $E$, представляет собой пуассоновский процесс, если выполняются условия:

1. количество точек в неперекрывающихся областях — независимые случайные величины;
2. для любой области $S$ конечного объема $N_S$ подчиняется распределению Пуассона со средним $\lambda V(S)$, где $V(S)$ — объем области $S$. Параметр $\lambda$ фиксирован и в некотором смысле служит мерой интенсивности распределения, которая не зависит от размера и формы.

В физике ядра и элементарных частиц сечение реакции записывается как отношение числа реакций к числу прошедших через мишень частиц, умноженному на число рассеивающих центров на единицу поверхности мишени. Очевидно, что число реакций подчиняется распределению Пуассона.

Среднее число появления событий составляет

$$\mu = \sum_{m=0}^{\infty} m \times \text{Poisson}_\lambda(m) = \lambda, \tag{3.42}$$

а дисперсия числа появлений события $A$ в $n$ испытаниях равна

$$\sigma^2 = \sum_{m=0}^{\infty} (m-\lambda)^2 \times \text{Poisson}_\lambda(m) = \lambda. \tag{3.43}$$

Производящая функция моментов для распределения Пуассона равна

$$M(t) = e^{\lambda(e^t-1)}. \tag{3.44}$$

В силу центральной предельной теоремы с увеличением $\lambda$ распределение Пуассона стремится к нормальному распределению с параметрами $\mu = \sigma^2 = \lambda$. Третий центральный момент распределения Пуассона равен $\lambda$, поэтому коэффициент асимметрии для него равен $1/\sqrt{\lambda}$ и стремится к нулю при $\lambda \to \infty$.

Легко показать, что сумма двух независимых пуассоновских случайных величин также является пуассоновской случайной величиной. *Теорема Райкова* (*англ.* Raikov's theorem) утверждает, что верно и обратное, то есть, если сумма двух независимых случайных величин подчиняется распределению Пуассона, то каждая из них также является пуассоновской случайной величиной.

### 3.4.2 Выбор оптимального времени измерения

Пусть $S$ — скорость счета для пуассоновского потока сигнальных событий, $B$ — для фоновых. Общая скорость счета, когда измеряется сигнал, составляет $S + B$. Если $T_{S+B}$ — время измерения сигнала, а $T_B$ — фона, и общее время измерения $T = T_{S+B} + T_B$ фиксировано, то существует оптимальное соотношение $T_{S+B}/T_B$.

Оценкой $S$ будет

$$\hat{S} = \frac{N_{S+B}}{T_{S+B}} - \frac{N_B}{T_B}. \tag{3.45}$$

Для погрешности измерения (с учетом того, что $\sigma^2_{N_i} = N_i$ и скорость счета равна $N_i/T_i$) получаем

$$\sigma_S = \Big(\frac{S+B}{T_{S+B}} + \frac{B}{T_B}\Big)^{1/2}, \tag{3.46}$$

откуда находим минимум при $T_{S+B}/T_B = \sqrt{(S+B)/B}$. В случае выбора оптимального соотношения $T_{S+B}/T_B$ относительная погрешность измерения $\delta_S = \sigma_S/S$ уменьшается с увеличением общего времени измерения как

$$\delta_S = \frac{1}{\sqrt{T}} \frac{\sqrt{S+B} + \sqrt{B}}{S}. \tag{3.47}$$

В предельных случаях имеем

$$S \gg B, \quad \delta_S = \sqrt{\frac{1}{TS}}, \tag{3.48}$$

и

$$S \ll B, \quad \delta_S = \sqrt{\frac{4B}{TS^2}}. \tag{3.49}$$

### 3.4.3 Специальные случаи

#### Пуассоновский поток с пересчетом

На практике часто нужно записывать не каждое событие из пуассоновского потока. Такая необходимость возникает, например, при регистрации нормировочного канала, скорость счета для которого обычно на порядки больше, чем для измеряемого. Для этого используются так называемые пересчетные устройства, записывающие одно (последнее) событие из $N$ поступивших на вход. Такой поток уже не является пуассоновским. Распределение временных интервалов между записанными событиями определяется:

$$I_N(t) = \frac{r^N t^{N-1} e^{-rt}}{(N-1)!}. \tag{3.50}$$

Распределение 3.50 называют *распределением Эрланга* (*англ.* Erlang distribution).

Средний временной интервал для такого потока в $N$ раз больше, чем для обычного

$$\bar{t}_N = \int_0^\infty t I_N(t) dt \;/\; \int_0^\infty I_N(t) dt = N/r, \tag{3.51}$$

однако мода распределения 3.50, определяемая из равенства нулю его производной, равна $(N-1)/r$. При увеличении $N$ временные интервалы становятся все более равномерными. Влияние этого эффекта важно учитывать в случае, когда так называемое *мертвое время* (*англ.* dead time) детектора сопоставимо с измеряемыми временными интервалами.

#### Модифицированное распределение Пуассона

Распределение Пуассона превосходно описывает поток событий радиоактивных распадов при условии, что время измерения много меньше периода полураспада. Это условие трудно соблюсти при измерении распадов короткоживущих изотопов. Введем среднее истинное число распадов за период времени от 0 до $T$:

$$\bar{r} = \frac{1}{T} \int_0^T r(t)\, dt. \tag{3.52}$$

Тогда вероятность наблюдать ровно $m$ событий запишется как $\mathrm{Poisson}_{\bar{r}T}(m)$. Таким образом, если измерение состоит в подсчете числа событий за период $T$, то статистика Пуассона опишет результаты множества таких измерений. Однако, если измерение состоит из множества измерений для коротких промежутков времени, распределение набора записанных событий уже не будет подчиняться распределению Пуассона.

Разобьем интервал $[0, T]$ на $n$ равных частей ($\tau = T/n$). Тогда интересующее нас распределение будет записываться как

$$\mathbf{P}(m) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{Poisson}_{r_j \tau}(m). \tag{3.53}$$

Для радиоактивного распада с параметром $\lambda$: $r(t) = r_0 e^{-\lambda t}$, можно записать

$$\mathbf{P}(m) = \mathrm{Poisson}_{r_0 \tau}(m) \times C(m), \tag{3.54}$$

где корректирующий фактор $C(m)$

$$C(m) = \frac{1}{T} \int_0^T \exp[-m\lambda t + r_0 \tau (1 - e^{\lambda t})]\, dt \tag{3.55}$$

может быть вычислен численно или выражен через специальные функции. Выражение 3.53 часто называют *модифицированным распределением Пуассона.*

### 3.4.4 Одноканальный эксперимент

Допустим, регистрируется пуассоновский поток сигнальных событий, характеризующийся неизвестной интенсивностью $s$, а также присутствует известная фоновая составляющая $b$, также распределенная по Пуассону. Экспериментально зарегистрировано $m$ событий. Требуется оценить вклад сигнала — $\hat{s}$, который по определению положителен.

#### Предельные случаи

Функция распределения для пуассоновской случайной величины задается суммой

$$F_{\mathrm{Poisson}}(m; \lambda) = \sum_{i=0}^{\lfloor m \rfloor} \mathrm{Poisson}_{\lambda}(i), \tag{3.56}$$

где $\lfloor m \rfloor$ — функция округления до ближайшего целого в меньшую сторону. Она связана с функцией распределения случайной величины хи-квадрат:

$$F_{\mathrm{Poisson}}(m; \lambda) = 1 - F_{\chi^2}(2\lambda; 2m + 2), \tag{3.57}$$

для целых $m$. Для больших $m$ можно использовать нормальное приближение

$$F_{\mathrm{Poisson}}(x; \lambda) \approx F_{\mathrm{normal}}(x; \mu = \lambda, \sigma^2 = \lambda). \tag{3.58}$$

Для начала рассмотрим случай $b = 0$. Если при измерении зафиксировано $m$ событий, то для параметра $\lambda$ будет справедлива следующая интервальная оценка:

$$\tfrac{1}{2}\chi^2_{2m,\alpha/2} \leq \lambda \leq \tfrac{1}{2}\chi^2_{2m+2,1-\alpha/2}, \tag{3.59}$$

где $\chi^2_{n,p}$ — это $p$-квантиль хи-квадрат распределения с $m$ степенями свободы. Численные значения интервалов, полученные таким способом, приведены в таблице 4.1. Следует отметить, что точно такие же численные значения могут быть получены в байесовском подходе в предположении, что параметр распределения Пуассона подчиняется равномерному распределению вероятности.

Случай, если $s$ большое, и для него применима аппроксимация нормальным распределением, также не вызывает затруднений: интервальная оценка для $s$ составляет:

$$(m - b) \pm \sqrt{m + b}. \tag{3.60}$$

Случай больших $b$ интересен с точки зрения минимального сигнала, доступного для измерения. Рассмотрим случай, когда $b$ определяется экспериментально. Как показано выше, в случае $b \gg s$ оптимально разбить измерение на два равных интервала времени. В один из этих интервалов следует оценить фон $b \pm \sqrt{b}$, а во втором провести измерение сигнала (при наличии фона) — $m \pm \sqrt{m}$. При построении доверительного интервала на вклад сигнала, следует постараться минимизировать так называемые *ошибки первого*[6] и *второго рода*[7]. Если нет дополнительных ограничений разумно требовать, чтобы вероятность ошибок первого и второго рода была одинаковой.

Для получения ответа опять воспользуемся аппроксимацией распределения Пуассона нормальным распределением. Для последнего функция распределения может быть записана в следующем виде:

$$F_{\mu,\sigma}(x) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\frac{x - \mu}{\sqrt{2\sigma^2}}, \tag{3.61}$$

где $\operatorname{erf}(z)$ — *функция ошибок* (*англ.* error function):

$$\operatorname{erf}\, z = \frac{2}{\sqrt{\pi}} \int\limits_0^z e^{-t^2}\, \mathrm{d}t. \tag{3.62}$$

Эта (и обратная ей) функция может быть рассчитана численно. Минимальный сигнал доступный для измерения вычисляется как решение системы

[6] *Ошибки первого рода* — ложноположительное срабатывание (англ. type I errors, $\alpha$-errors, false positive) — фиксируем наличие сигнала, когда его нет.

[7] *Ошибки второго рода* — ложноотрицательное срабатывание (англ. type II errors, $\beta$ errors, false negative) — фиксируется отсутствие сигнала, когда он есть.

уравнений:

$$\operatorname{erf} \frac{x-b}{\sqrt{2b}} = 1 - 2\alpha, \tag{3.63}$$

$$\operatorname{erf} \frac{x-b-s_{\min}}{\sqrt{2(b+s_{\min})}} = 2\alpha - 1, \tag{3.64}$$

в которой $\alpha$ — это численное значение, равное вероятности допустить ошибки первого и второго рода. Для $\alpha = 0.95$ решение приближенно можно записать в виде:

$$s_{\min} = 4.653\sqrt{b} + 2.706. \tag{3.65}$$

Выражение 3.65 часто называют *уравнением Кюри*. Особо отметим, что полученное выражение можно использовать только в случае, когда вклад фона оценивается по измерению такой же длительности как измерение в режиме сигнал + фон. Часто, погрешность оценки параметра фона $b$ намного меньше чем $\sqrt{b}$. Например, фоновые события дают одинаковый вклад во все каналы многоканального эксперимента, а сигнал ожидается лишь в одном канале. Усредняя фон по многим каналам, что равносильно аппроксимации фоновой гистограммы постоянной величиной, можно получить значительный выигрыш в погрешности оценки $b$.

## 3.5 Непараметрические методы

Если функциональная зависимость плотности распределения неизвестна, для оценки моментов распределения можно использовать непараметрические методы, основанные на выборочной функции распределения.

Рассмотрим так называемый *метод складного ножа* (*англ.* jackknife), используемый для оценки погрешности в статистическом выводе. Его можно применить для погрешности распределения среднего и несмещенной оценки дисперсии. Суть метода заключается в построении распределения этих статистик, содержащего $n-1$ значений, каждое из которых получено из выборки за исключением одного из её элементов. Оценка дисперсии для какого-либо параметра ($\hat{\theta}$) будет определяться выражением:

$$\hat{\sigma}^2_{\mathrm{JK}} = \frac{n-1}{n} \sum_{i=1}^{n} \left( \theta_{(i)} - \frac{1}{n} \sum_{j=1}^{n} \theta_j \right)^2, \tag{3.66}$$

где $\theta_{(i)}$ — оценка среднего, вычисленная без $i$-го элемента выборки. Этот метод дает несмещенную состоятельную оценку дисперсии асимптотически нормальных ММП-оценок.

Другим методом, который может использоваться для оценки погрешностей моментов распределения является *непараметрический бутстраппинг* (*англ.*

bootstrapping). Суть этого метода состоит в том, что при помощи математического датчика случайных чисел можно создать ансамбль псевдовыборок объемом $n$ путем случайного выбора элемента выборки $n$ раз. Некоторые элементы не попадут в псевдовыборку, а некоторые войдут в неё один или несколько раз. Для каждой такой псевдовыборки строится интересующая нас оценка и определяется дисперсия её распределения. Метод хорош тем, что позволяет получить хорошую оценку распределения статистик, рассчитать которые аналитически не представляется возможным.

# 4 Редкие события

## 4.1 Метод Фельдмана-Казинса

Для получения оценки числа сигнальных событий в случае малого числа событий и присутствия фона необходимо решать уравнение типа:

$$\mathbf{P}(m) = \sum_{i=0}^{m} \mathrm{Poisson}_i(s+b). \tag{4.1}$$

В случае флуктуации фона в сторону нижних значений, $\hat{s}_{\min}$ становится меньше нуля, то есть верхний предел оказывается в нефизической области параметров.[1] Как следствие доверительный интервал оказывается пуст, что противоречит одному из свойств, которым он должен обладать.

Интересный подход к решению проблемы предложили Фельдман и Казинс. В рамках этого подхода сначала определяется величина сигнала $s_{\mathrm{best}}$, для которого вероятность $L(m|s_{\mathrm{best}}+b)$, а далее вероятности суммируются в соответствии с убыванием отношения отношения $R = L(i, s_i+b)/L(m, s_{\mathrm{best}}+b)$, пока сумма не достигнет величины $(1-\alpha)$. Если точка $s_i$ попала в область суммирования, то она принадлежит доверительному интервалу. Подобная схема удовлетворяет определению доверительного интервала и обеспечивает гладкий переход от двухстороннего предела к одностороннему.

Оптимальность подхода базируется на *лемме Неймана-Пирсона*, которая утверждает, что наиболее *мощной статистикой*[2] (*англ.* tests with the most power) будет являться отношение функций правдоподобия. Для счетного эксперимента это:

$$Q = \frac{L(s+b)}{L(b)}. \tag{4.2}$$

К минусам данного метода следует отнести следующий парадокс. Пусть есть два эксперимента, которые не нашли сигнал (в обоих случаях $s = 0$), но при этом уровни ожидаемого фона были разные: $b_1 = 0$ и $b_2 = 5$. Тогда расчет для уровня доверия 95% дает интервалы [0, 3.09] для первого случая и [0, 1.54] для второго. То есть, худший эксперимент дает лучшую оценку!

[1] Например, при ожидаемых $s = 0$, $b = 3$ и измеренном $m = 0$ оценка, проведенная из уравнения 4.1, даст $\hat{s}_{\max} = -0.7$.

[2] Статистика, которая при принятии статистического решения обладает минимальной вероятностью допустить ошибку второго рода ($\beta$) при фиксированной вероятности допустить ошибку первого рода ($\alpha$)

Таблица 4.1: Доверительные интервалы, рассчитанные для случая отсутствия фона ($b = 0$) и $1 - \alpha = 0.95$

| n | $\chi^2$-распределение | Фельдман-Казинс |
|---|---|---|
| 0 | 0, 3.00 | 0, 3.09 |
| 1 | 0.051, 4.47 | 0.05, 5.14 |
| 2 | 0.355, 6.30 | 0.36, 6.72 |
| 3 | 0.818, 7.75 | 0.82, 8.25 |
| 4 | 1.37, 9.15 | 1.37, 9.76 |
| 5 | 1.97, 10.51 | 1.84, 11.26 |
| 6 | 2.61, 11.84 | 2.21, 12.75 |
| 7 | 3.29, 13.15 | 2.58, 13.81 |
| 8 | 3.98, 14.43 | 2.94, 15.29 |
| 9 | 4.70, 15.71 | 4.36, 16.77 |
| 10 | 5.43, 16.96 | 4.75, 17.82 |

## 4.2 $p$-значение и $\mathrm{CL}_s$-метод

При поиске сигнала важной величиной является вероятность ошибки первого рода при отвержении гипотезы об отсутствии сигнала ($H_0$). Пусть $x$ — это некоторая статистика, подчиняющаяся распределению $g$ при условии выполнения $H_0$. Тогда значение статистики, при которой $H_0$ отвергается на *уровне значимости* $\alpha$ вычисляется из

$$\alpha = \int_{x_\alpha}^{\infty} g(x \mid H_0)\mathrm{d}x. \tag{4.3}$$

Величину $(1 - \alpha)$ называют *доверительной вероятностью* (*англ.* confidence level). Экспериментатор должен пред началом измерения задаться уровнем значимости, вычислить $x_\alpha$ и, если наблюдаемое значение для статистики ($x_{\mathrm{obs}}$) превосходит $x_\alpha$, отвергнуть $H_0$. Величину

$$p = \int_{x_{\mathrm{obs}}}^{\infty} g(x \mid H_0)\mathrm{d}x. \tag{4.4}$$

называют *p-значением* (*англ.* p-value). Она имеет смысл вероятности принять статистические флуктуации за проявление сигнала.

В экспериментальной физике атомного ядра и элементарных частиц сложилась практика указывать статистическую значимость $\delta$, определенную как вероятность выпасть из центральной области ($x = \mu \pm \delta$) нормального распределения:

$$1 - \alpha = \frac{1}{2\pi\sigma} \int_{\mu-\delta}^{\mu+\delta} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x. \tag{4.5}$$

Таблица 4.2: $\delta$ для некоторых значений $\alpha$.

| $\alpha$ | $\delta$ | $\alpha$ | $\delta$ |
|---|---|---|---|
| 0.3173 | $1\sigma$ | 0.200 | $1.28\sigma$ |
| $4.55\times10^{-2}$ | $2\sigma$ | 0.100 | $1.64\sigma$ |
| $2.7\times10^{-3}$ | $3\sigma$ | 0.050 | $1.96\sigma$ |
| $6.3\times10^{-5}$ | $4\sigma$ | 0.010 | $2.58\sigma$ |
| $5.7\times10^{-7}$ | $5\sigma$ | 0.001 | $3.29\sigma$ |
| $2.0\times10^{-9}$ | $6\sigma$ | $10^{-4}$ | $3.89\sigma$ |

Некоторые значения $\delta$ (при фиксированном параметре $\alpha$) приведены в таблице 4.2.

Общеупотребима следующая терминология:

- если статистическая значимость соответствует $x_\alpha$ меньшему чем $3\sigma$ для стандартного нормального распределения, то говорят, что наблюдения не противоречат гипотезе об отсутствии сигнала;
- в случае попадания в диапазон от 3 до $5\sigma$, говорят об указании (*англ.* evidence) на возможное наличие эффекта;
- при превышении уровня $5\sigma$, говорят об открытии (*англ.* discovery) или о наблюдении (*англ.* observation) эффекта.

Как уже отмечалось выше, при поиске малых сигналов из-за флуктуаций фона метод Фельдмана-Казинса может дать более жесткое ограничение на $s$ для эксперимента с меньшей чувствительностью. Чтобы избежать этого парадокса, в современной практике анализа данных принято ограничивать пространство параметров, убирая статистические модели, которые заведомо не чувствительны к величине сигнала при заданной оценке уровня фона. Для этих целей применяют $\mathrm{CL}_s$-*метод*, базирующийся на статистике:

$$\mathrm{CL}_s = \frac{p_{s+b}}{1 - p_b}. \tag{4.6}$$

В определение $\mathrm{CL}_s$-статистики входят $p$-значения для гипотезы отсутствия сигнала ($p_b$) и гипотезы присутствия сигнала на уровне $s$ — $p_{s+b}$. Точка в пространстве параметров модели[3] исключается на уровне значимости $\alpha$, если выполнено условие:

$$\mathrm{CL}_s \leq \alpha. \tag{4.7}$$

Интервальные оценки, полученные при помощи $\mathrm{CL}_s$-метода совпадают с байесовскими оценками сделанными в предположении равномерного априорного

[3]Для одноканального эксперимента этой точке соответствует значение ожидаемого сигнала.

распределения параметров для процессов, подчиняющихся распределениям Пуассона и Гаусса.

К недостаткам метода можно отнести больший (чем требуется) статистический охват. В этом смысле данная методика дает консервативную оценку значения измеряемого параметра.

## 4.3 Типовая задача

### Постановка задачи

Проводится измерение некоторого параметра $\Lambda$, который связан с величиной сигнала ($s$), регистрируемого детектором $s = f(\Lambda)$. Сигнал представляет собой величину, распределенную в соответствии с распределением Пуассона. Одновременно с сигналом регистрируется фоновая составляющая, также подчиняющаяся распределению Пуассона. Параметр этого распределения известен с некоторой точностью $b \pm \sigma_b$. Подчеркнем, принципиальное отличие статистической флуктуации параметра $b$ при однократном измерении[4] и систематическую погрешность ($\sigma_b$) связанную с неточностью оценки среднего значения фонового параметра. В эксперименте измеряется величина $m = s + b$, то есть по его результату будет известно число $m_{\text{obs}}$. Для таких условий возможно два принципиально различных вопроса:

1. получить интервальную оценку для $\Lambda$ при известной величине $m_{\text{obs}}$;
2. оценить чувствительность эксперимента к параметру $\Lambda$ при неизвестном $m_{\text{obs}}$.

Заметим, что второй вопрос имеет смысл до того, как проведена процедура открытия данных (*англ.* analysis unblinding). Например, если сигнал сконцентрирован лишь в одном бине экспериментальной гистограммы, то это означает, что параметры $b$ и $\sigma_b$ могут быть довольно точно определены из эксперимента, путем аппроксимации формы фонового распределения.

Процедура планирования эксперимента подразумевает, что вопросы 1 и 2 будут задаваться последовательно. Необходимо заранее определить, в каких случаях будет определяться односторонний, а в каких случаях двухсторонний доверительный интервал. Обычной практикой является выбор одностороннего интервала в случае, если данные не противоречат гипотезе об отсутствии сигнала, и двухстороннего в противном случае. Также стоит заранее определиться с процедурой определения предела на значение сигнала для случая сильной флуктуации фона в сторону меньших значений.

[4] дисперсия, соответствующая этой флуктуации, равна $b$

### Случай $s = \Lambda^2$ и $b > 25$

Случай $b > 25$ позволяет воспользоваться асимптотическими свойствами распределения Пуассона, а именно его стремлением к нормальному распределению. Даже если точность оценки параметра фонового распределения точна (малые $\sigma_b$), при однократном измерении вклад фона будет флуктуировать как $\sqrt{b}$. Граница $b = 25$ определяется довольно естественным образом, ведь стандартное отклонение соответствует $5\sigma$ для нормального распределения.

Чувствительность при заданном уровне доверительной вероятности ($\mathrm{CL} = 1 - \alpha$) определяется величиной $\Lambda_{\max}$, соответствующей минимальному значению зарегистрированного сигнала $s_{\min}$.

Для случая $\sigma_b = \sqrt{b}$ и $\mathrm{CL} = 0.95$ можно воспользоваться выражением 3.65, а для других значений CL следует численно решить систему уравнений 3.63 и 3.64.

Если, же $\sqrt{b} \gg \sigma_b$, то $s_{\min}$ можно определить, воспользовавшись функцией нормального распределения (выражение 3.61), решив относительно $b_{\max}$ и $s_{\min}$ систему уравнений вида:

$$F_{b,\sigma_b}(b_{\max}) = \mathrm{CL}, \qquad F_{b+s_{\min},\sqrt{b_{\max}}}(b_{\max}) = 1 - \mathrm{CL}. \tag{4.8}$$

Функция $\mathrm{erf}(x)$ хорошо табулирована, а также имплементирована во все стандартные пакеты компьютерных программ. Например:

- Python 3: `scipy.special.erf`;
- CERN ROOT: `ROOT::Math::erf`.

Когда измерение проведено, и известна $m_{\mathrm{obs}}$, доверительный интервал для $s$ определяется выражением 3.60.

### $m_{\mathrm{obs}}$ известно, $b < 25$

Пусть до процедуры измерения было решено строить односторонний доверительный интервал, если статистическая значимость сигнала меньше $3\sigma$. Чтобы проверить, выполняется ли это условие, следует рассчитать $p$-значение (см. раздел 4.4), и сравнить его со значением, взятым из Таблицы 4.2. Для гипотезы отсутствия сигнала параметр пуассоновского распределения равен $b$. Соответствующий квантиль распределения Пуассона можно вычислить, воспользовавшись связью его функции распределения с функцией распределения случайной величины хи-квадрат (см. 3.57) или воспользоваться стандартными решениями:

- Python 3: `scipy.stats.poisson.cdf` — интеграл вычисляется по левой части распределения;

- CERN ROOT: `ROOT::Math::poisson_cdf` — интеграл вычисляется по левой части распределения;
- CERN ROOT: `ROOT::Math::poisson_cdf_c` — интеграл вычисляется по правой части распределения.

В случае построения одностороннего доверительного интервала его нижняя граница $s_{\text{low}} = 0$, а верхняя ($s_{\text{up}}$) определяется решением уравнения 4.1. Решение выражается через квантиль распределения хи-квадрат:

$$s_{\text{up}} = \frac{1}{2}\chi^2_{2m+2,1-\alpha} - b. \tag{4.9}$$

Функции для вычисления квантиля хи-квадрат распределения содержаться в стандартных библиотеках:

- Python 3: `scipy.stats.chi2.cdf` — интеграл вычисляется по левой части распределения;
- CERN ROOT: `ROOT::Math::chisquared_cdf` — интеграл вычисляется по левой части распределения;
- CERN ROOT: `ROOT::Math::chisquared_cdf_c` — интеграл вычисляется по правой части распределения.

# Литература


1. В.Э. Гмурман *Теория вероятностей и математическая статистика*, М., Высшая школа, 2008.

2. М.Дж. Кендалл, А. Стьюарт, *Теория распределений*, М., Наука, 1966.

3. М.Дж. Кендалл, А. Стьюарт, *Статистические выводы и связи*, М., Наука, 1973.

4. М.Б. Лагутин, *Наглядная математическая статистика*, М., БИНОМ, 2009.

5. А.И. Кобзарь, *Прикладная математическая статистика*, М., Физматлит, 2012.

6. Д. Худсон, *Статистика для физиков*, М., Мир, 1970.

7. ГОСТ Р 50779.10-2000, *Статистические методы. Вероятность и основы статистики. Термины и определения.*

8. G. Bohm, G. Zech, *Introduction to statistics and data analysis for physicists*, Hamburg, Deutsches Elektronen-Synchrotron, 2010.

9. Particle Data Group, *Review of Particle Physics*, раздел 40: *Statistics*, Prog. Theor. Exp. Phys., 083C01, 2020.

10. G.F. Knoll, *Radiation Detection and Measurement 3rd Edition*, New York, John Wiley & Sons, 2000.