ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Установочная лекция по статистике

Высшая школа цифровой культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Введение в статистику

1.1 Самое-самое введение и мотивировка

Здравствуйте, уважаемые слушатели. В этой лекции мы осветим те базовые понятия статистики, которые будут подробно изучаться в дальнейшем, а также покажем, какие задачи решает математическая статистика.

При решении различных прикладных задач, при наблюдении тех или иных явлений, оказывается, что многие закономерности носят вероятностный характер. Если быть более точным, вероятностная модель этих закономерностей позволяет получить разумные результаты и выводы. Как же адекватно формализовать фразу: «носят вероятностный характер», что это означает? Это означает, что изучаемая нами закономерность (явление) случайна в своих проявлениях и, тем самым, описывается случайной величиной, которая, в свою очередь, имеет вероятностное распределение. Понятие распределения было введено нами в первой части курса, касающейся теории вероятностей, а посему мы знаем, что знание распределения – это очень много, ведь распределение, грубо говоря, говорит нам, что, когда и с какой вероятностью может произойти с величиной ξ.

Кроме того, можно вычислять такие полезные штуки, как: математическое ожидание $\mathsf{E}\xi$, дисперсию $\mathsf{D}\xi = \mathsf{E}\left(\xi - \mathsf{E}\xi\right)^2$, медиану, вероятности попадания в те или иные множества, плотность f_ξ и многое-многое другое. Иными словами, зная распределение интересующей нас случайной величины, мы сразу знаем очень много, а также можем получать сразу кучищу полезной информации.

Давайте приведем пример. Предположим, что случайная величина ξ – это рост взрослого человека. Ясное дело, что он случаен и меняется от человека к человеку. В то же время оказывается, что он описывается некоторым распределением. Кстати, это распределение хорошо аппроксимируется нормальным распределением (конечно, в разумных пределах, так как нормально распределенная случайная величина может принимать и отрицательные значения, и сколь угодно большие, что для роста, конечно, невозможно). Распределение же характеризуется, например, плотностью f_ξ. Ну а зная плотность, можно вычислить и средний рост, и среднеквадратическое отклонение, и даже вероятность встретить дядю Стёпу (ну она-то равна нулю, это понятно). Но и тут нужно быть осторожным в интерпретации результатов: параметры этого распределения меняются, как минимум, от страны к стране, от национальности к национальности, ну и так далее. Кроме того, ясное дело, средний рост женщин и мужчин отличается, а значит распределения роста по полу, приближаемые нормальными N_{a,σ^2}, приближаются разными нормальными, то есть нормальными, но с разными параметрами математического ожида-

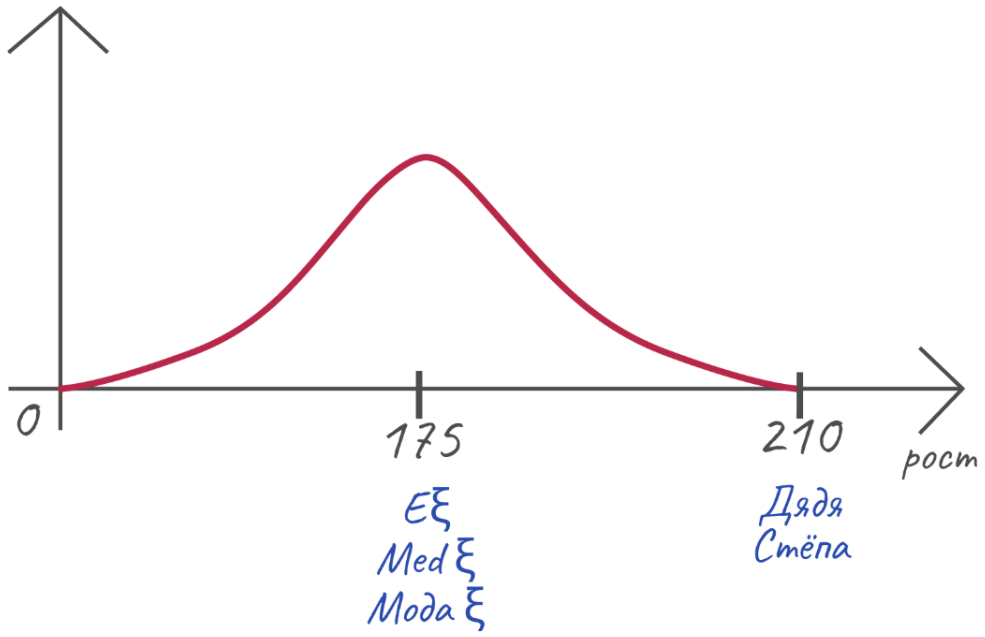

Рис. 1: Нормальное распределение

ния a и дисперсии σ^2! Поэтому важно, чтобы наши выводы, сделанные по выборке, касались именно того распределения, из которого эта выборка. Будет очень неправильно, если по выборке из роста мужчин будут сделаны выводы по пошиву одежды для женщин.

Ну а зачем нам знать эти параметры? О чем они нам говорят? В случае нормального распределения мода (локальный максимум плотности), медиана и математическое ожидание совпадают, и все могут быть интерпретированы, как среднее значение случайной величины. σ^2 – это дисперсия, а значит σ – среднеквадратическое отклонение. Иными словами, в среднем, с вероятностью примерно 0.68, значения случайной величины находятся в пределах

$$(a - \sigma,\ a + \sigma) = (\mathsf{E}\xi - \sigma_\xi,\ \mathsf{E}\xi + \sigma_\xi)\,.$$

И что же это значит? А это значит, что одежды, с размерами роста из приведенного интервала имеет смысл шить куда больше, чем с размерами вне этого интервала – она просто не будет продаваться в достаточном объеме! А что если взять не σ, а 2σ? Тогда вероятность того, что наша случайная величина попадет в интервал

$$(a - 2\sigma,\ a + 2\sigma)$$

примерно равна 0.96. Вообще, для нормального распределения есть так называемое правило 3σ. Оказывается, что вероятность случайной величине ξ, имеющей нормальное распределение N_{a,σ^2}, попасть в интервал $(a - 3\sigma, a + 3\sigma)$ равна 0.997, что помнить вовсе необязательно, но что неплохо осознавать: вероятность чрезвычайно велика!

Для других распределений ситуация, вообще говоря, не такая. Например, для случайной величины ξ, имеющей показательное распределение Exp_λ, мода – это $x = 0$, математическое ожидание – это $\mathsf{E}\xi = \frac{1}{\lambda}$ и медиана $\mathsf{med}\xi = \frac{\ln 2}{\lambda}$. Все три величины различны. А тогда что и для какой оценки

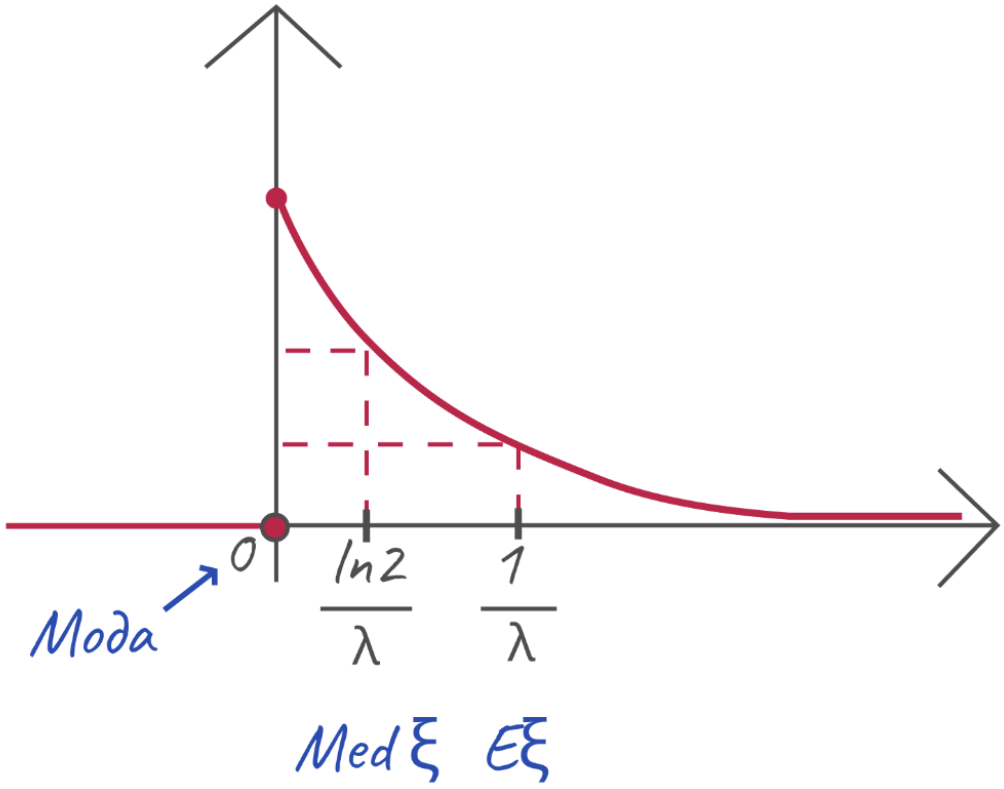

Рис. 2: Показательное распределение

выбирать? Все зависит от задачи. Мы уже вспоминали задачу про зарплаты, напомним ее.

Предположим, что имеется некоторая фирма, в которой работает 100 человек, один из которых начальник. Заработная плата начальника равна 101000 долларов в месяц, а заработная плата каждого работника равна 1000 долларов в месяц. Пусть случайная величина ξ – зарплата работника, тогда ее распределение может быть задано, как

ξ	1000	101000
P	$\frac{99}{100}$	$\frac{1}{100}$

Легко понять, что математическое ожидание случайной величины ξ равно

$$\mathsf{E}\xi = 1000 \cdot \frac{99}{100} + 101000 \cdot \frac{1}{100} = 2000,$$

то есть средняя зарплата равна 2000 долларов в месяц. В то же время, медиана равна 1000 и медианная зарплата равна 1000. Ясно, что в этом примере

гораздо более «честной» является медиана, нежели математическое ожидание, из-за такого сильного выброса в зароботной плате начальника.

В принципе, можно сказать, что нам хочется найти или оценить плотность f_ξ случайной величины ξ, или ее дискретный аналог – ряд распределения, потому что дальше все хорошо: суммы-интегралы дают какие-то числа, которые уже изучены вдоль и поперек, и даже в этом небольшом фрагменте прозвучали не раз. Но плотность нам нужна вовсе не только для этого. По плотности можно понять: а однородно ли распределение?

Не вдаваясь в детали, можно сказать, что однородное распределение – это распределение с одной модой (унимодальные). Давайте все же сформулируем определение моды в одном месте.

Определение 1.1.1 *Модой случайной величины ξ, имеющей абсолютно непрерывное распределение с плотностью f_ξ, называется точка локального максимума f_ξ.*

Если же случайная величина ξ имеет дискретное распределение, заданное таблицей

ξ	a_1	a_2	…	a_n	…
P	p_1	p_2	…	p_n	…

то модой случайной величины ξ называется произвольное значение a_i такое, что

$$p_{i-1} \le p_i, \quad p_{i+1} \le p_i,$$

если хотя бы одна из соседствующих вероятностей определена.

Рисунки иллюстрируют введенные определения. Для дискретного распреде-

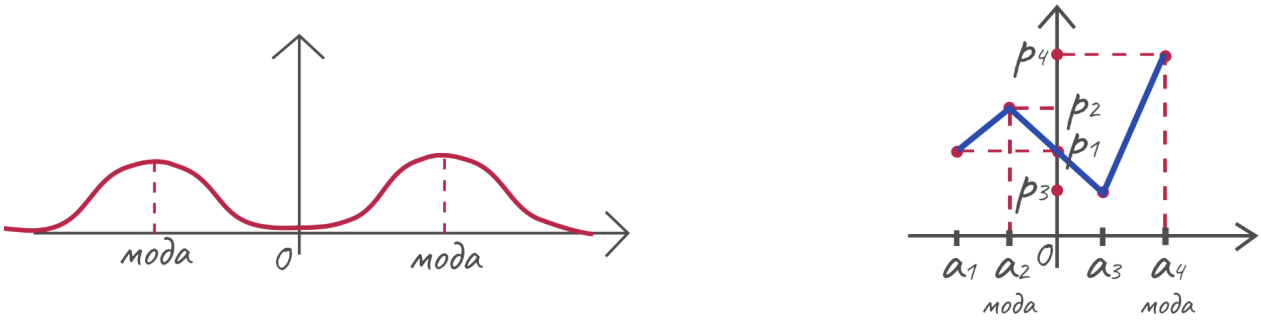

Рис. 3: Моды

ления часто вводят понятие многоугольника распределения (тогда узловые точки (a_i, p_i) видны куда лучше), он тоже изображен на рисунке. Многоугольник получается просто соединением отрезком соседних точек (a_i, p_i) и (a_{i+1}, p_{i+1}).

Легко понять (кроме как, возможно, для биномиального распределения $\mathsf{Bin}_{n,p}$), что почти что все распределения, которые мы рассматривали: вырожденное, Бернулли, биномиальное, Пуассона, показательное, нормальное (короче все, кроме равномерного и левого хвоста показательного) – унимодальные. А что вообще может означать, что распределение не однородно и имеет несколько мод? Часто это означает, что рассматривается смесь унимодальных распределений (линейная комбинация их плотностей). Не углубляясь в детали, поясним, что смесь – это когда с какой-то вероятностью p проявляется один закон распределения, а с обратной, то есть с вероятностью $(1-p)$ – другой закон распределения. У смеси плотность равна линейной комбинации плотностей этих распределений с коэффициентами p и $(1-p)$. Например, можно смешать крокодилов и черепах и рассматривать распределение смеси их весов. Если есть неоднородность (то есть смесь разного), то такое достаточно бессмысленно исследовать: зачем кому-то нужен средний вес (математическое ожидание) смеси крокодилов и черепах?

В общем, много всего интересного нам хочется узнать про распределение. И тут мы сталкиваемся с главной проблемой: а как это сделать? Нам же не дано распределение, мы наблюдаем только какие-то значения случайной величины. Иными словами, есть проявления некоторой вероятностной закономерности, по которым нужно узнать про генеральное распределение, но формулки-то нет: нет ни плотности, ни ряда распределения. Тут и возникает математическая статистика.

1.2 Немного о выборке

Итак, пусть у нас имеется n проявлений некоторой вероятностной закономерности. Иными словами, мы наблюдаем n значений некоторой случайной величины ξ, называемой еще генеральной совокупностью, имеющей какое-то (неизвестное нам) распределение. Тогда перед нами обычный числовой n-мерный вектор $X = (x_1, x_2, ..., x_n)$ – это выборка после эксперимента. Итак,

Определение 1.2.1 *Пусть ξ – рассматриваемая нами случайная величина. Выборкой (после эксперимента) $X = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ называется n независимых реализаций случайной величины ξ. Последнюю часто называют генеральной совокупностью.*

Например, пусть ξ – это рост случайно взятого человека, тогда

$$X = (175, 182, 168, 155, 192)$$

это выборка объема 5 из генеральной совокупности ξ. Хорошо, допустим теперь мы хотим вычислить среднее, а точнее прикинуть, чему равно математическое ожидание ξ. Как бы это сделать? Наверное, вы скажете, что логично

посчитать среднее арифметическое, называемое выборочным средним:

$$\overline{X} = \frac{175 + 182 + 168 + 155 + 192}{5} = 174.4.$$

И как, все готово? То есть мы получили математическое ожидание ξ? Немного подумав, наверное, можно бросить осторожную фразу вроде: «Видимо, это оценка этого математического ожидания»? И что это значит? Это значит, что, поменяв выборку, мы получим, вообще говоря, другое число. Видимо, это число имеет какое-то отношение к реальному, но какое? Каждый раз меняя «измеряемых» людей, мы получаем другое число. И как тут строить теорию? Опять что-то усреднять? Так можно циклиться до бесконечности.

Важно еще понимать и вот какой момент, называемый неоднородностью. Придя в детский сад и померив людей в нем, получим одно число (вероятно, сравнительно небольшое, даже рост воспитателей и нянечек не поможет), а измерив рост членов баскетбольной команды – разительно отличающееся. Дело в том, что выборки-то берутся из разных распределений! Поэтому мы всегда должны понимать, какое именно распределение мы изучаем по данной выборке. А то получится вроде ситуации с опросом в интернете с вопросом: «Пользуетесь ли вы интернетом»?

Ну ладно, давайте подумаем еще немного. Все-таки ξ – случайная величина, а значит она определена на каком-то вероятностном пространстве. Значит, меняя элементарные исходы пространства элементарных исходов, эта величина ξ принимает, вообще говоря, различные значения. У нас элементарные исходы – это люди, потому взяв разных людей, получим, вообще говоря, разные значения роста. Конечно, можно быть совсем занудным: нет людей одинакового роста, все зависит от точности измерений, но не будем об этом, и не будем приводить такие оговорки в дальнейшем. В итоге, разумно привести общее определение выборки.

Определение 1.2.2 *Пусть ξ – рассматриваемая нами случайная величина. Выборкой $X = (X_1, X_2, ..., X_n)$ называется n независимых случайных величин, имеющих распределение такое же, как и ξ.*

Итак, введенное определение выборки – это, по сути, определение выборки до конкретного эксперимента. В нашем примере, до того, как мы измерим рост конкретных n людей, выборка – это случайный вектор (вектор из n независимых случайных величин), а после – это набор из n чисел $X = (x_1, x_2, ..., x_n)$. Именно на основе этого понятия выборки, как случайного вектора, и строится вся теория математической статистики. В дальнейшем мы не будем конкретизировать то, что мы понимаем под выборкой (числовой вектор или вектор из случайных величин), это должно будет быть понятно из контекста. В принципе, к выборке логично относиться, как к некоторой функции от ω

до реализации на конкретном элементарном исходе ω, и к «значению» этой функции на конкретном ω после эксперимента. Ну что, давайте перейдем к задачам математической статистики.

2 Оценки характеристик распределения

2.1 Немного об оценках

Пусть θ – некоторая характеристика распределения случайной величины ξ, например, пусть θ – математическое ожидание $\mathsf{E}\xi$ случайной величины ξ (о точечном оценивании более строго мы будем говорить позднее). Мы бы хотели оценить этот параметр, используя выборку. Статистика $\widehat{\theta}$ – это некоторая функция от выборки $X = (X_1, X_2, ..., X_n)$, то есть

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n).$$

Например,

$$\widehat{\theta} = \overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

это статистика. Так как у выборки есть два определения, то и статистика может пониматься в двух смыслах. Если выборка рассмотрена на конкретном эксперименте, то $\widehat{\theta}$ – это конкретное число:

$$\widehat{\theta} = \widehat{\theta}(x_1, x_2, ..., x_n).$$

В нашем примере на конкретной выборке $X = (x_1, x_2, ..., x_n)$ в результате эксперимента получаем

$$\widehat{\theta} = \overline{X} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Впрочем, это выражение для нас уже не ново, ведь чуть раньше мы уже считали что-то вроде «среднего роста» по данной выборке. Мы не зря рассматриваем этот пример с выборочным средним, он нам потребуется и далее.

Начали мы с вопроса оценки математического ожидания и, видимо, хочется сказать, что именно выборочное среднее $\overline{X}$ является хорошей оценкой истинного математического ожидания $\mathsf{E}\xi$ генеральной совокупности ξ. А что значит, что оценка «хорошая»? И почему вообще такая оценка «естественна»? Откуда она берется?

2.2 Эмпирическое распределение

На самом деле, с каждой конкретной выборкой, полученной в результате эксперимента, то есть с выборкой $X = (x_1, x_2, ..., x_n)$, разумно связать новую случайную величину ξ^*. Какую, спросите вы? Ну, вроде бы ясно какую: такую, которая каждое значение выборки x_i принимает с вероятностью $\frac{1}{n}$ – у нас же нет любимчиков в выборке. В итоге, можно написать следующую табличку

ξ^*	x_1	x_2	...	x_n
$\widetilde{\mathsf{P}}$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

.

Кстати, как легко видеть, случайная величина, а точнее ее распределение, конечно, зависит от n, от объема выборки, но мы не будем наделять ее дополнительным индексом.

Важно отметить, что случайная величина ξ^*, конечно, задана на другом вероятностном пространстве, нежели ξ, потому вероятность на ее вероятностном пространстве будем обозначать $\widetilde{\mathsf{P}}$. Например, имея в результате эксперимента выборку

$$X = (1, 2, 1, 3, 1)$$

объема 5, соответсвующее распределение новой, разыгранной случайной величины ξ^* записывается, как

ξ^*	1	2	3
$\widetilde{\mathsf{P}}$	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

.

Как же вычислить вероятность попадания случайной величины ξ^* в множество A? Да как обычно, ровно как мы и поступаем в дискретном случае: после эксперимента, на конкретной числовой выборке, вероятность попасть в A вычисляется, как

$$\widetilde{\mathsf{P}}(\xi^* \in A) = \frac{|\{x_i : \ x_i \in A\}|}{n},$$

иными словами, вероятность события A – это отношение количества элементов выборки, лежащих в A, к объему выборки n.

Если же выборка «абстрактная», то есть эксперимент не произошел, то $\widetilde{\mathsf{P}}(\xi^* \in A)$ – это случайная величина, ведь

$$\widetilde{\mathsf{P}}(\xi^* \in A) = \frac{|\{X_i : \ X_i \in A\}|}{n}$$

и справа стоит функция от выборки.

Итак, хотелось бы, чтобы распределение представленной нами случайной величины неплохо приближало распределение настоящей генеральной

совокупности ξ с ростом n. И интуиция не подводит, а именно, справедлива теорема, которую мы готовы сформулировать. На самом деле, чтобы не мучить вас всякими σ-алгебрами и на этом этапе не затуманивать суть дела, давайте считать, что A – произвольное подмножество $\mathbb{R}$, и A – событие.

Теорема 2.2.1 (ЗБЧ для эмпирического распределения) *Для каждого множества $A \subset \mathbb{R}$ имеет место сходимость почти наверное:*

$$\frac{|\{X_i : \ X_i \in A\}|}{n} \xrightarrow[n\to+\infty]{\text{п. н.}} \mathsf{P}(\xi \in A).$$

Доказательство. Доказательство рассмотренной теоремы немедленно следует из закона больших чисел. Ясно, что вероятность ξ^* попасть в множество A – это отношение количества элементов X_i, попавших в A, к общему числу элементов n.

$$\widetilde{\mathsf{P}}(\xi^* \in A) = \frac{|\{X_i : \ X_i \in A\}|}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{I}(X_i \in A),$$

где I – индикатор события A, то есть $\mathsf{I}(X \in A) = 1$, если $X \in A$, и ноль иначе. Все эти индикаторы (А они, кстати, случайные величины (!)) одинаково распределены, независимы и имеют распределение Бернулли с параметром $p = \mathsf{P}(X_i \in A) = \mathsf{P}(\xi \in A)$. Значит, и математическое ожидание слагаемых равно $\mathsf{P}(\xi \in A)$, а тогда по ЗБЧ в форме Колмогорова

$$\widetilde{\mathsf{P}}(\xi^* \in A) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{I}(X_i \in A) \xrightarrow[n\to+\infty]{\text{п. н.}} \mathsf{P}(\xi \in A).$$

□

Итак, наше распределение поточечно почти всюду (или почти наверное) сходится к истинному. Иными словами, при увеличении n эмпирическое распределение все более и более похоже на истинное.

2.3 Оценки параметров распределения

Наверное, теперь ясна основная идея: может быть, в качестве оценки некоторой характеристики генеральной совокупности ξ, можно использовать соответствующую характеристику ξ^*, раз распределение последней так неплохо приближает истинное? Попробуем.

Тогда в качестве оценки $\widehat{\mathsf{E}\xi}$ математического ожидания ξ логично взять соответствующее математическое ожидание $\widetilde{\mathsf{E}}\xi^*$, то есть, по сути, выборочное среднее:

$$\widehat{\mathsf{E}\xi} = \widetilde{\mathsf{E}}\xi^* = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

В качестве оценки $\widehat{\mathsf{D}\xi}$ дисперсии генеральной совокупности ξ логично взять дисперсию разыгранной случайной величины:

$$\widehat{\mathsf{D}\xi} = \widetilde{\mathsf{D}}\xi^* = \widetilde{\mathsf{E}}(\xi^* - \widetilde{\mathsf{E}}\xi^*)^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Аналогичным образом можно построить оценки и других моментов, но сейчас мы этого делать не будем.

Для построения оценки медианы, нам потребуется понятие вариационного ряда. Упорядочим выборку по порядку, получим новый набор чисел (или случайных величин) $X_{(i)}$, $i \in \{1, ..., n\}$:

$$X_{(1)} \le X_{(2)} \le ... \le X_{(n)},$$

где, например, $X_{(1)} = \min(X_1, X_2, ..., X_n)$, $X_{(n)} = \max(X_1, X_2, ..., X_n)$. Полученный таким образом набор чисел и называется вариационным рядом. В вариационном ряде i-ый член называется i-ой порядковой статистикой. Рангом элемента выборки X_i называется его номер в вариационном ряде. Логично взять в качестве оценки медианы случайной величины ξ либо средний член вариационного ряда, если он определен однозначно (то есть если n нечетно) и, скажем, полусумму центральных членов вариационного ряда, если n четно. Иными словами,

$$\widehat{\mathsf{med}\ \xi} = X_{((n+1)/2)}, \text{при } n \text{ нечетном},$$

$$\widehat{\mathsf{med}\ \xi} = \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right), \text{при } n \text{ четном}.$$

Вернемся к нашему синтетическому примеру. Напомним, что выборка из генеральной совокупности ξ такова

$$X = (1, 2, 1, 3, 1),$$

а распределение случайной величины задается таблицей

ξ^*	1	2	3
$\widetilde{\mathsf{P}}$	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

.

Ясно, что оценкой математического ожидания генеральной совокупности ξ в нашем случае оказывается

$$\widehat{\mathsf{E}\xi} = \widetilde{\mathsf{E}}\xi^* = 1 \cdot \frac{3}{5} + 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} = \overline{X} = \frac{1+2+1+3+1}{5} = 1.6.$$

Что с дисперсией? Давайте оценим ее. Аналогичным образом получаем, что

$$\widehat{\mathsf{D}\xi} = \widetilde{\mathsf{D}}\xi^* = S^2 = \frac{16}{25}.$$

И, наконец, для оценивания медианы составим вариационный ряд. Он примет вид

$$1 \leq 1 \leq 1 < 2 < 3.$$

Так как объем выборки равен 5, а это число нечетное, то $\widehat{\mathsf{med}\ \xi} = X_{((5+1)/2)} = X_{(3)} = 1$.

2.4 Немного о качестве оценок

Хорошо, мы даже научились получать какие-то оценки, и поняли, из каких соображений их можно получать. А как понять, что они хорошие? Пока что мы смогли обосновать «хорошесть» только эмпирического распределения.

Можно измерить то, насколько ошибка хорошая, используя среднеквадратическую ошибку:

$$\mathsf{MSE} = \mathsf{E}\left(\widehat{\theta} - \theta\right)^2.$$

Распишем последнее выражение подробнее, раскрыв скобки и учитывая, что $\theta \in \mathbb{R}$:

$$\mathsf{E}\left(\widehat{\theta} - \theta\right)^2 = \mathsf{E}\left(\widehat{\theta}^2 - 2\widehat{\theta}\theta + \theta^2\right) = \mathsf{E}\widehat{\theta}^2 - 2\theta\mathsf{E}\widehat{\theta} + \theta^2 =$$

$$= \mathsf{E}\widehat{\theta}^2 - \left(\mathsf{E}\widehat{\theta}\right)^2 + \left(\mathsf{E}\widehat{\theta}\right)^2 - 2\theta\mathsf{E}\widehat{\theta} + \theta^2 = \mathsf{D}\widehat{\theta} + \left(\mathsf{E}\widehat{\theta} - \theta\right)^2,$$

откуда

$$\mathsf{MSE} = \mathsf{D}\widehat{\theta} + \left(\mathsf{E}\widehat{\theta} - \theta\right)^2.$$

Ясно, что чтобы оценка была разумной, хотелось бы, чтобы ошибка MSE стремилась к нулю при объеме выборки n, стремящемся к бесконечности.

Дадим несколько простых определений, более строго к которым подойдем в главе про точечные оценки.

Определение 2.4.1 *Оценка $\widehat{\theta}$ называется несмещенной, если*

$$\mathsf{E}\widehat{\theta} = \theta.$$

Ясно, что несмещенность, по сути, означает, что «в среднем» значение оценки совпадает с истинным значением параметра. Например, выборочное среднее – это несмещенная оценка математического ожидания ξ. Действительно,

$$\mathsf{E}\overline{X} = \mathsf{E}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}X_i = \frac{n\mathsf{E}X_1}{n} = \mathsf{E}X_1 = \mathsf{E}\xi,$$

так как X_i одинаково распределены. Дадим еще одно определение.

Определение 2.4.2 *Оценка $\widehat{\theta}$ называется асимптотически несмещенной, если*

$$\lim_{n\to+\infty} \mathsf{E}\widehat{\theta} = \theta.$$

Естественно, любая несмещенная оценка оказывается и асимптотически несмещенной. Кстати, мы уже знакомы со смещенной, но асимптотически несмещенной оценкой – это оценка дисперсии S^2. Вычисление $\mathsf{E}S^2$ уже не делается в две строчки, поэтому его мы оставим до соответствующей лекции, а пока что просто поверим в ее (оценки) смещенность, но, в то же время, асимптотическую несмещенность.

Ну и зачем нам это все? А выходит, и это совершенно очевидно, что если оценка $\widehat{\theta}$ оказывается несмещенной или асимптотически несмещенной, то из соотношения

$$\mathsf{MSE} = \mathsf{D}\widehat{\theta} + \left(\mathsf{E}\widehat{\theta} - \theta\right)^2.$$

следует, что стремление к нулю MSE равносильно стремлению к нулю дисперсии оценки $\mathsf{D}\widehat{\theta}$ с ростом объема выборки n.

Определение 2.4.3 *Оценка $\widehat{\theta}$ называется состоятельной в смысле среднеквадратического, если MSE стремится к нулю с ростом n.*

Давайте проверим, является ли оценка $\overline{X}$ состоятельной в смысле среднеквадратического? Достаточно вычислить ее дисперсию, предположив, что существует $\mathsf{D}\xi$:

$$\mathsf{D}\overline{X} = \mathsf{D}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}X_i = \frac{1}{n}\mathsf{D}X_1 = \frac{1}{n}\mathsf{D}\xi \xrightarrow[n\to+\infty]{} 0.$$

Итого, состоятельность в среднеквадратичном доказана.

2.5 Параметрические модели

Иногда оказывается, что про распределение случайной величины ξ что-то известно. Если это учесть, то есть шанс получить что-то более хорошее, нежели работая в режиме «слепых котят». Параметрические модели предполагают знание распределения случайной величины ξ с точностью до каких-то параметров. Например, может быть известно, что генеральная совокупность имеет нормальное распределение N_{a,σ^2} с неизвестными параметрами a и σ^2, или равномерное распределение $\mathsf{U}_{a,b}$ с неизвестными параметрами $a < b$, или геометрическое распределение G_p с параметром p, или, что очень часто, биномиальное распределение $\mathsf{Bin}_{n,p}$ с параметром p. И наша задача преобразуется в задачу оценивания конкретных параметров распределения.

Часто параметры модели можно интерпретировать. Так как мы знаем, что у нормального распределения параметр a отвечает за математическое

ожидание, а параметр σ^2 – за дисперсию, то мы уже готовы предложить оценки для этих параметров: выборочное среднее и выборочную дисперсию, а именно

$$\widehat{a} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\widehat{\sigma}^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Кроме того, мы можем даже найти распределение случайной величины $\overline{X}$.

Лемма 2.5.1 *Выборочное среднее имеет нормальное распределение* $\mathsf{N}_{a,\sigma^2/n}$.

Почему это так? Да потому, что согласно свойствам нормального распределения, если случайные величины X_1 и X_2 независимы и $X_1 \sim \mathsf{N}_{a_1,\sigma_1^2}$, а $X_2 \sim \mathsf{N}_{a_2,\sigma_2^2}$, то $X_1 + X_2 \sim \mathsf{N}_{a_1+a_2,\sigma_1^2+\sigma_2^2}$. Осталось заметить, что $X_1 + X_2 + ... + X_n \sim \mathsf{N}_{na,n\sigma^2}$ и воспользоваться свойством линейных преобразований над случайными величинами.

2.6 А как сравнивать оценки между собой?

Давайте теперь вспомним, что, вообще-то, математическое ожидание и медиана у случайной величины, имеющей нормальное распределение N_{a,σ^2}, совпадают. Как думаете, что мы хотим предложить? Конечно, еще одну оценку для математического ожидания такой генеральной совокупности – выборочную медиану. Как же теперь их сравнить? Да все так же, по MSE:

$$\mathsf{MSE} = \mathsf{D}\widehat{\theta} + \left(\mathsf{E}\widehat{\theta} - \theta\right)^2.$$

При этом, если оценки несмещенные (ну-ка быстро вспомните, что это!), то достаточно сравнить их дисперсии: логично считать, что чем дисперсия меньше, тем оценка лучше, да? Осознайте, а для этого вспомните, что такое дисперсия, что она показывает? А теперь подумайте «чисто алгебраически», ведь для несмещенных оценок $\mathsf{MSE} = \mathsf{D}\widehat{\theta}$. MSE – это ошибка (не важно какая), а мы знаем, что чем меньше ошибка, тем лучше. Значит, вроде как, чем она меньше, тем лучше.

В несмещенности выборочного среднего $\overline{X}$ мы уже убедились. Несмещенность выборочной медианы можно объяснить, например, так. Пусть $X = (X_1, X_2, ..., X_n)$ – выборка из нормального распределения N_{a,σ^2}. Тогда $Y = (Y_1, Y_2, ..., Y_n)$, где $Y_i = X_i - a$ – это выборка из нормального распределения N_{0,σ^2}. При этом, согласно свойствам линейных преобразований, $-Y \sim \mathsf{N}_{0,\sigma^2}$. Но медианы Y и $-Y$ отличаются знаком, то есть $\mathsf{med}\ Y = -\mathsf{med}(-Y)$, а

значит ее среднее $\mathsf{E}(\mathsf{med}\ Y)$, так как его распределение не поменялось, должно быть равно нулю. Ну а тогда $\mathsf{E}(\mathsf{med}\ X) = \mathsf{E}(a + \mathsf{med}\ Y) = a$, и медиана оказывается несмещенной оценкой математического ожидания.

Дисперсию выборочного среднего мы уже вычисляли, она равна $\frac{\sigma^2}{n}$. Несколько сложнее вычислить дисперсию выборочной медианы, опустим этот момент, и приведем результат:

$$\mathsf{D}\left(\widehat{\mathsf{med}\ \xi}\right) = \frac{\pi}{2} \cdot \frac{\sigma^2}{n}.$$

В итоге, выборочная медиана оказывается хуже в смысле среднеквадратической ошибки MSE. И что, занавес? Эх, было бы все так просто. С точки зрения теории – да, но реальность вносит свои коррективы?

Прикладная статистика имеет дело с реальными данными, а данные, конечно же, частенько содержат ошибки и выбросы, попавшие в выборку случайно (как крокодил, случайно затесавшийся среди черепах). А что, это проблема. Представьте, что вы измеряете средний размер черепашек (ну пусть в условных единицах), а в выборку затесался крокодил. Он неплохо так увеличит числитель нашей дроби у выборочного среднего, смотрите. Например, пусть

$$X = (5, 7, 3, 8, 4, 2, 100).$$

В конце – крокодил, а в начале – аквариумные маленькие черепашки. Смотрите, без крокодила выборка бы была объема 6 и выборочное среднее бы давало

$$\frac{5 + 7 + 3 + 8 + 4 + 2}{6} = \frac{29}{6} \approx 4.8.$$

А что с крокодилом? Мы получим

$$\frac{5 + 7 + 3 + 8 + 4 + 2 + 100}{7} \approx 18.4.$$

Как вам черепашки? Это ж пол аквариума, наверное. Конечно, если выброс один, а выборка велика, то его вклад постепенно нивелируется. А что если выбросы появляются с периодичностью? Прибор там барахлит, или крокодилы хорошо притворяются? Тогда вся наука будет долго недоумевать нашим измерениям.

Тут и всплывает понятие устойчивости оценки. Устойчивая (или робастная) оценка — оценка, которая почти не меняется от добавления в данные выброса. И с точки зрения устойчивость выборочная медиана лучше! Ну смотрите, что, в нашем примере выборочная медиана разве что-то бы почувствовала? Да нет, конечно. Без крокодила вариационный ряд бы был

$$2 < 3 < 4 < 5 < 7 < 8$$

и $\widehat{\mathsf{med}\ \xi} = \frac{4+5}{2} = 4.5$. А когда крокодил добавился, вариационный ряд стал

$$2 < 3 < 4 < 5 < 7 < 8 < 100$$

и $\widehat{\mathsf{med}\ \xi} = \frac{4+5}{2} = 5$. Видите, изменение совсем маленькое, по сравнению с выборочным средним.

Вообще, можно сказать следующее: если какая-то оценка выражается через порядковые статистики (или ранги), то она будет иметь повышенную устойчивость.

2.7 А существует ли наилучшая оценка?

Хорошо, допустим, мы выбрали из нескольких несмещенных оценок ту, у которой самая маленькая дисперсия. Будет ли она наилучшей в смысле среднеквадратического подхода? Вдруг есть оценка, дисперсия которой еще меньше, просто мы эту оценку еще не нашли? В случае параметрической модели ответ возможен, и его можно получить, используя так называемое неравенство Рао-Крамера. Неравенство Рао-Крамера применимо в случае, когда распределение удовлетворяет так называемым условиям регулярности (о них мы будем говорить в соответствующей лекции) и говорит, что в классе несмещенных оценок выполняется соотношение:

$$\mathsf{D}\widehat{\theta} \geq \frac{1}{nI(\theta)},$$

где $I(\theta)$ – так называемая информация Фишера для одного наблюдения (так можно поступать, так как мы имеем дело с повторной независимой выборкой), вычисляемая по распределению генеральной совокупности. В итоге, если нам удается найти ту оценку, для которой в неравенстве Рао-Крамера возникает равенство, то, в случае регулярного распределения, мы можем быть уверены, что ничего лучше (с точки зрения теории) получиться просто не может. Все остальные оценки будут как минимум не лучше, чем данная. Например, для нормального распределения N_{a,σ^2} при неизвестном $\theta = a$ и известной дисперсии σ^2 информация Фишера равна

$$I(\theta) = \frac{1}{\sigma^2},$$

а тогда в неравенстве Рао-Крамера оценка снизу на дисперсию оказывается такой:

$$\mathsf{D}\widehat{\theta} \geq \frac{1}{nI(\theta)} = \frac{\sigma^2}{n},$$

откуда следует, что полученная нами оценка $\overline{X}$ оказывается наилучшей среди несмещенных в смысле среднеквадратического подхода, или, как еще говорят, эффективной, так как, как мы уже неоднократно повторяли, и даже

вычислили,

$$\mathsf{D}\overline{X} = \frac{\sigma^2}{n}.$$

2.8 А как вообще получать оценки?

Отлично. Мы уже обсудили с вами то, что такое оценка и зачем она нужна, как сравнивать оценки между собой и даже как определить, является ли оценка наилучшей в некотором классе. Но до сих пор за кадром остался такой важный вопрос: а как вообще получать оценки? И, конечно, как получать «хорошие» оценки (теперь-то мы знаем, что такое хорошесть)?

Оказывается, в параметрической модели есть метод, который позволяет получить самую лучшую оценку (конечно, при объеме выборки, стремящемся к бесконечности). Метод этот называется методом максимального правдоподобия. При выполнении условий регулярности, оценки метода максимального правдоподобия оказываются асимптотически несмещенными, то есть

$$\lim_{n\to+\infty} \mathsf{E}\widehat{\theta} = \theta,$$

асимптотически эффективными (то есть неравенство Рао-Крамера в пределе становится равенством):

$$\lim_{n\to+\infty} \left(\mathsf{D}\widehat{\theta} \cdot nI(\theta)\right) = 1$$

и асимптотически нормальными (новое для нас понятие), которое означает, что

$$Y_n = \frac{\widehat{\theta} - \theta}{\sqrt{\mathsf{D}\widehat{\theta}}} \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Иными словами, так называемая центрированная $(\widehat{\theta} - \theta)$ и нормированная случайная величина сходится по распределению к случайной величине, имеющей стандартное нормальное распределение.

3 Интервальное оценивание

3.1 Немного об интервальном оценивании

То, что мы с вами изучали так подробно – это точечное оценивание. Мы оценивали какую-то числовую характеристику, или числовой параметр распределения, числом. Хорошо ли это? Смотря как посмотреть. Дело в том, что мы, конечно, не знаем истинного значения параметра (или характеристики), иначе зачем его оценивать, а потому возникает вот какой вопрос: число 5

– это хорошая оценка, скажем, математического ожидания, или нет? Теоретически выборочное среднее – это хорошая оценка, но мы-то находимся в рамках конкретного эксперимента!

Все зависит от того, насколько эта оценка реально близка к истинному математическому ожиданию. Точечное оценивание, в свою очередь, не дает нам об этом никакой информации: у нас нет информации о погрешности $|\widehat{\theta}-\theta|$. Да, мы знаем, что если оценка хорошая, то она, конечно, приближается к истинному значению параметра (вспомним MSE), но ведь это приближение идет в среднем, с ростом n, точнее на бесконечности. А мы на практике, конечно, не можем рассматривать выборки бесконечного размера, можем только конечного, и тут же возникает проблема: а какое значение n достаточно, при каком значении n оценка «близка» к истинному значению параметра?

Тут вступает в бой так называемое интервальное, или доверительное оценивание. Оно дает возможность написать интервал, в который, с заданной наперед вероятностью, попадет значение истинного параметра. Не стоит наивно ожидать, что значение вероятности берется равным единице, и дело в шляпе: в этом случае наш доверительный интервал будет очень большим, а пользы не будет никакой (подробнее об этом мы поговорим в соответствующей лекции).

Зададим большую вероятность, например, 0.95, называемую уровнем доверия (доверительным уровнем, уровнем надежности).

Определение 3.1.1 *Доверительный интервал, построенный по выборке* $X = (X_1, X_2, ..., X_n)$ *уровня доверия* 0.95 *– это интервал с концами* $\theta^{\pm}(X_1, X_2, ..., X_n)$*, для которого*

$$\mathsf{P}\left(\theta^- < \theta < \theta^+\right) \geq 0.95.$$

Концы доверительного интервала $\theta\pm$ – это случайные величины, поэтому они случайны.

Пусть опять выборка берется из нормального распределения N_{a,σ^2} с неизвестным параметром $\theta = a$ и известной дисперсией σ^2. Тогда, как мы знаем, выборочное среднее $\overline{X}$ имеет распределение $\mathsf{N}_{a,\sigma^2/n}$ и легко получить (и это мы обязательно получим), что

$$\mathsf{P}\left(\overline{X} - c_{0.95}\frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + c_{0.95}\frac{\sigma}{\sqrt{n}}\right) = 0.95,$$

где число $c_{0.95} \approx 1.96$ (что это за числа мы тоже детально обсудим в дальнейшем). Что видно из этого интервала? Во-первых, его середина – это выборочное среднее. Во-вторых, с ростом объема выборки n, длина интервала уменьшается и стремится к нулю со скоростью порядка $n^{-1/2}$. Так, например, если n увеличить в 100 раз, то длина интервала уменьшится в 10 раз.

4 Проверка гипотез

4.1 А что такое гипотезы?

Кроме оценивания характеристик распределения, часто бывает важно проверить какую-то гипотезу, то есть какое-то предположение об этом распределении.

Например, мы измеряем объем продаж какого-то товара и хотим проверить гипотезу, что продажи в этом году, в среднем, больше, чем в прошлом году. Просто сравнивать средние больше-меньше нельзя, так как могло случайно получиться больше. Гипотеза же, в свою очередь, не про выборочные средние, а про математические ожидания неизвестного генерального распределения, которые нам неизвестны.

Можно привести и другой пример. Есть монетка, с помощью которой решается «кто сегодня готовит ужин»? Нам очень важно, чтобы монетка давала справедливое решение, т.е. с вероятностью 0.5 выпадала решка и с вероятностью 0.5 – орел. Как это проверить? Да просто бросить монетку много раз и посчитать, сколько раз что выпало. Вообще говоря, мы получим выборку из распределения Бернулли с вероятностью успеха p. Гипотеза, которая нас интересует, звучит как $p = 0.5$.

Итак, гипотеза – это некоторое утверждение про распределение генеральной совокупности ξ.

4.2 Проверка гипотез

Мы будем обычно проверять гипотезу H_0, которая, как правило, говорит, что какой-то эффект не наблюдается. Например, что продажи не изменились, или монетка правильная.

Если все случайно, невозможно делать безошибочные выводы. Именно поэтому выбирается небольшая вероятность α, с которой мы согласны ошибочно сказать, что верная нулевая гипотеза не верна. Она называется уровнем значимости (или так называемой вероятностью ошибки первого рода).

Критерий δ – это правило, зависящее от выборки и α, по которому мы говорим, что нулевая гипотеза верна или не верна. Обычно содержательные выводы делаются в случае, если нулевая гипотеза отвергается. Как правило, в нулевой гипотезе содержится утверждение, что некоторого эффекта нет. Соответственно, если гипотеза отвергается, то делается вывод, что эффект есть. Но можем ли мы делать такой вывод?

Кроме нулевой гипотезы H_0 есть еще альтернативная гипотеза H_1, которая конкретизирует, что мы хотим понимать под фразой: «нулевая гипотеза неверна». Например, если нулевая гипотеза говорит о том, что новое лекарство лучше старого на 5%, то альтернативная гипотеза может утверждать,

что лекарство либо хуже, либо лучше, но не больше, чем на 4%. Вариант отклонения от нулевой гипотезы на, например, 0.1%, нас не волнует.

Если выбрана альтернативная гипотеза, то возникает так называемая ошибка второго рода, которая означает, что верна альтернативная гипотеза, но мы не отвергли нулевую гипотезу. Вероятность ошибки второго рода может быть большой (мы ее не выбираем, в отличие от вероятности ошибки 1 рода), поэтому мы не говорим фразу: «гипотезы принимаем», а говорим - не отвергаем.

Мощность критерия, равная единице минус вероятность ошибки второго рода, то есть вероятность отвергнуть гипотезу H_0, если верна H_1, характеризует умение критерия обнаруживать отличие H_1 от нулевой гипотезы. Она обладает следующими свойствами: во-первых, чем больше объем выборки n, тем мощность больше; во-вторых, чем больше уровень значимости α, тем мощность больше; в третьих, чем дальше альтернативная гипотеза от основной гипотезы, тем мощность больше. В частности, отсюда следует, что чтобы увеличить мощность, выбирается уровень значимости максимальным из допустимых. Понятие допустимости зависит от того, каковы риски в случае принятия неверного решения. Например, если мы потеряем 100 рублей в случае ошибки 1 рода, мы можем взять $\alpha = 0.1$, а если 1000000 рублей, то α должно быть существенно меньше.

Так как уровень значимости α определяется не внутри процедуры проверки гипотезы, а задается исследователем, то часто результатом критерия является пороговое значение p, которое называется p-значением. Если уровень значимости α больше p, то гипотеза H_0 отвергается, а если меньше – нет оснований отвергать гипотезу. Поэтому p-значение – это минимальный уровень значимости, при котором гипотеза отвергается (и максимальный, при котором не отвергается). Если p большое, например, 0.7, то при всех разумных уровнях значимости нулевая гипотеза не отвергается. А если, например, $p = 0.0001$, то гипотеза будет отвергаться. Для промежуточных p, например, в районе 0.01 – 0.1, результат проверки гипотезы зависит от конкретного уровня значимости.

Обычно мы не можем одновременно уменьшить α и увеличить мощность (то есть уменьшить ошибку второго рода). Давайте разберем такой шуточный пример. Пусть у нас есть выборка объема 1 из нормального распределения $\mathsf{N}_{a,1}$. Рассмотрим две гипотезы: $\mathsf{H}_0: \ a = 0$ и $\mathsf{H}_1: \ a = 1$ и критерий

$$\delta(X_1) = \begin{cases} \mathsf{H}_0, & X_1 \le d \\ \mathsf{H}_1, & X_1 > d \end{cases}.$$

Что такое ошибка первого рода? Это вероятность не отвергнуть гипотезу H_1 в случае, когда верна гипотеза H_0, то есть $\alpha = \mathsf{P}_{\mathsf{H}_0}(\delta = H_1) = \mathsf{P}_{\mathsf{H}_0}(X_1 > d)$.

Что такое ошибка второго рода? Это вероятность не отвергнуть гипотезу H_0 в случае, когда верна гипотеза H_1, то есть $\mathsf{P}_{\mathsf{H}_1}(\delta = \mathsf{H}_0) = \mathsf{P}_{\mathsf{H}_1}(X_1 \leq d)$. На рисунке четко видно, что уменьшая вероятность ошибки первого рода (α), автоматически возрастает вероятность ошибки второго рода, тем самым уменьшается мощность критерия, и наоборот. Давайте покажем то, что кри-

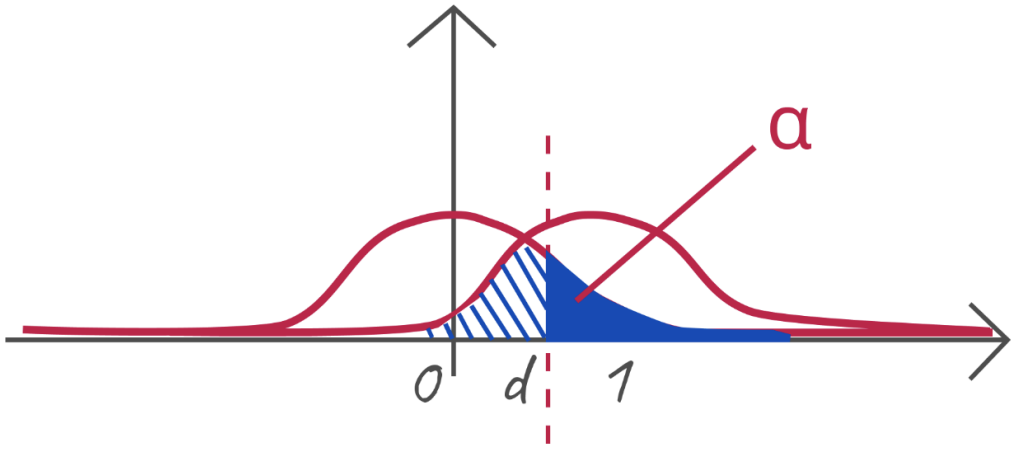

Рис. 4: Ошибки первого и второго рода

терии по-разному регулируют ошибки первого и второго рода. Напомним, что ошибку первого рода мы фиксируем, а ошибка второго рода – как получится, так получится. Тем самым, может быть не все равно, что рассмотреть в качестве нулевой гипотезы, а что в качестве альтернативы.

Например, мы собираемся лететь на самолете и есть две гипотезы – самолет исправный и самолет неисправный. При проверки гипотезы возникают две ошибки – отвергнуть гипотезу, что самолет исправный, хотя самолет исправный (соответственно, зря сдать билет и зря потерять деньги) и отвергнуть гипотезу, что самолет неисправный, хотя самолет неисправный (соответственно, полететь на самолете и попасть в катастрофу).

В каком случае ошибку важнее контролировать? Конечно, во втором. Поэтому, в идеале, надо, чтобы нулевая гипотезы была о том, что самолет неисправен. В этом случае мы полетим на неисправной самолете с выбранной нами (и, конечно, очень маленькой) вероятностью. Ну а если не сможем контролировать ошибку второго рода, то просто есть шанс, что потеряем в деньгах. Так часто бывает, что по смыслу ошибок они друг от друга отличаются по критичности последствий. Соответственно, лучше, если более критичное последствие произойдет в случае ошибки первого рода, выбираемой нами самими.

4.3 Сравнение критериев

Как ранее оценок для одного параметра могло быть много, так и критериев для проверки одной и той же гипотезы может быть много. Как их

сравнивать?

Сначала нужно зафиксировать альтернативу так, чтобы она включала туда те отличия от нулевой гипотезы, которые нам важно обнаружить.

1. Критерии можно сравнивать по мощности против выбранной альтернативы, т.е. вероятности отвергнуть гипотезу, если верна альтернатива.

2. Также, как и оценки, важным свойством критериев является их устойчивость к выбросам. Поэтому критерии, которые используют не сами значения выборки, а только их ранги, являются более устойчивыми.

Например, пусть $\mathsf{H}_0 : \ \mathsf{E}\xi = 0$, где ξ — то, на сколько возросли продажи товара. Если у нас учет хороший, то мы можем использовать так называемый t-тест. А если плохой и предполагаем, что там могут быть ошибки, то нужен более устойчивый тест. Подробнее об этих вещах мы будем говорить в соответтсвующих лекциях.