Финансовый университет
при правительстве Российской Федерации

Шамраева Виктория Викторовна

кандидат физико-математических наук,
доцент кафедры
математики и анализа данных

Теория вероятностей и математическая статистика

НАПРАВЛЕНИЕ ПОДГОТОВКИ: «Прикладная математика - ПМ»

КВАЛИФИКАЦИЯ (СТЕПЕНЬ): бакалавр

Раздел 5. Статистическая проверка гипотез

Статистическая проверка гипотез

Раздел 5. Статистическая проверка гипотез

Основные понятия

Раздел 5. Статистическая проверка гипотез

Одна из **основных задач математической статистики** состоит в том, чтобы по **реализации** $x_1, \ldots, x_n$ **случайной выборки** $X_1, \ldots, X_n$ проверить определенную гипотезу

о виде

или

параметрах генерального распределения.

Раздел 5. Статистическая проверка гипотез

Статистической гипотезой называется любое предположение о виде или параметрах генерального распределения.

Раздел 5. Статистическая проверка гипотез

Статистическая гипотеза называется **параметрической**, если она основана на предположении, что генеральное распределение известно с точностью до конечного числа параметров, и **непараметрической** в противном случае.

Раздел 5. Статистическая проверка гипотез

Примеры параметрических гипотез:

1) математическое ожидание $E(X)$ случайной величины X равно 5 в предположении, что X имеет нормальный закон распределения с известной дисперсией $D(X) = 1$;
2) дисперсии случайных величин X и Y совпадают между собой в предположении, что X и Y нормально распределены.
3) гипотеза о равенстве математического ожидания доходности акции конкретному значению 0,14 в предположении, что доходность имеет нормальный закон распределения.

4) предположение о том, что вероятность наступления риска угона для определенной категории страхователей КАСКО равна 0,03.

Раздел 5. Статистическая проверка гипотез

Примеры непараметрических гипотез:

1) случайная величина X имеет нормальный закон распределения;
2) функции распределения случайных величин X и Y совпадают;
3) предположение о том, что логарифм объема торгов акциями имеет нормальный закон распределения.

Раздел 5. Статистическая проверка гипотез

Предположим, что генеральное распределение зависит от некоторого параметра θ.

Статистическая гипотеза называется **полностью определенной** или **простой**, если предположение о ее справедливости однозначно определяет распределение случайной величины;

в противном случае гипотеза называется **неопределенной** или **сложной**.

Раздел 5. Статистическая проверка гипотез

Гипотезу, подлежащую проверке, называют **основной** или **нулевой** и обозначают H_0.

В качестве **основной гипотезы** всегда должна выступать **полностью определенной простая гипотеза**.

Если в исходной постановке задачи на роль основной гипотезы выдвигается **неопределенная (сложная) гипотеза**, то она всегда доводится до уровня определенной путем введения уточнений.

Раздел 5. Статистическая проверка гипотез

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу.

Вторую гипотезу называют **альтернативной** или **конкурирующей гипотезой** и обозначают H_1.

Раздел 5. Статистическая проверка гипотез

В каждом конкретном случае важно правильно выбирать альтернативную гипотезу, удобную для принятия содержательного решения.

На практике, как правило, в качестве **нулевой гипотезы** выбирают такую, которая соответствует ≪нормальному положению дел≫

(например, человек здоров; обвиняемый невиновен; деталь изготовлена с соблюдением стандарта; потенциальный заемщик способен погасить кредит; транзакция не является мошеннической и т. д).

Раздел 5. Статистическая проверка гипотез

Соответственно, **альтернативная гипотеза** соответствует противоположной ситуации, которая обычно рассматривается как менее вероятная, но при этом требующая обязательной реакции

(человек болен; обвиняемый виновен; деталь изготовлена с нарушением стандарта; потенциальный заемщик не способен погасить кредит; транзакция является мошенической и т. д).

Раздел 5. Статистическая проверка гипотез

Пример. После изменения конфигурации компьютерной сети были собраны случайным образом 200 замеров скорости передачи сообщений.

Основная гипотеза H_0: изменение конфигурации не имело эффекта.

Альтернативная гипотеза H_1: эффект от изменения статистически значим.

Раздел 5. Статистическая проверка гипотез

Поскольку нулевая гипотеза выдвигается на основании выборочных данных, то она может оказаться как правильной, так и неправильной – **мы не знаем!**

И поэтому она подлежит ***статистической проверке***.

Раздел 5. Статистическая проверка гипотез

Нулевая гипотеза H_0 подлежит проверке, по результатам которой ее можно принять либо отклонить. «Принять» означает «не получить убедительных аргументов для отклонения гипотезы».

Альтернативная гипотеза H_1 принимается только тогда, когда есть убедительное статистическое доказательство для отклонения основной гипотезы.

Принять основную гипотезу $\mathbf{H_0}$ ⟷ Отвергнуть $\mathbf{H_0}$ и принять $\mathbf{H_1}$

Раздел 5. Статистическая проверка гипотез

Пусть $\vec{x} = (x_1, \ldots, x_n)$ — реализация случайной выборки $\vec{X} = (X_1, \ldots, X_n)$.

Критерием или **статистическим критерием** называется **правило**, по которому, на основании данных выборки $\vec{x}$, оценивается степень достоверности гипотезы H_0.

Проверка **параметрических гипотез** проводится на основе **критериев значимости**, а **непараметрических** - **критериев согласия**.

Статистическая гипотеза называется **параметрической**, если она основана на предположении, что генеральное распределение известно с точностью до конечного числа параметров, и **непараметрической** в противном случае.

Раздел 5. Статистическая проверка гипотез

Пусть $T(\vec{X})$ — некоторая **статистика**, характеризующая отклонение эмпирических данных от тех гипотетических значений, которые соответствуют проверяемой гипотезе H_0.

И пусть, кроме того, **распределение** этой **статистики**, в случае справедливости гипотезы H_0, **известно** (точно или хотя бы приближенно).

Обозначим $T_{\text{набл}}$ — **наблюдаемое (или выборочное) значение статистики**, т. е. значение статистики $T(\vec{x})$, вычисленное для полученной реализации случайной выборки.

Раздел 5. Статистическая проверка гипотез

Зафиксируем достаточно малое число $\alpha \in (0; 1)$.

Разобьем множество всех возможных значений статистики T на две части:

критическую область критерия K и его дополнение $\overline{K} = D$.

Критическая область *K* выбирается так, чтобы выполнялось соотношение

$$P_{H_0}(\{T \in K\}) \leq \alpha.$$

(Через $P_{H_0}(A)$ обозначена вероятность события A, вычисленная в предположении, что гипотеза H_0 верна.)

Таким образом, критическая область включает в себя маловероятные значения статистики T, при условии, что верна основная гипотеза H_0.

Раздел 5. Статистическая проверка гипотез

Критерий проверки гипотезы теперь можно сформулировать следующим образом:

По имеющейся выборке $x_1, \ldots, x_n$ находится наблюдаемое значение статистики $T_{\text{набл}}$.

Если окажется, что $T_{\text{набл}} \in K$, то делается вывод о том, что в предположении справедливости гипотезы H_0 произошло маловероятное событие. Поэтому эта гипотеза должна быть **отвергнута**, как противоречащая статистическим данным $x_1, \ldots, x_n$, полученным в результате эксперимента.

В противном случае (т. е. если $T_{\text{набл}} \in D = \overline{K}$) считается, что **данные не противоречат $\boldsymbol{H_0}$**.

Критерием или **статистическим критерием** называется **правило**, по которому, на основании данных выборки $\vec{x}$, оценивается степень достоверности гипотезы H_0.

Раздел 5. Статистическая проверка гипотез

Замечание.

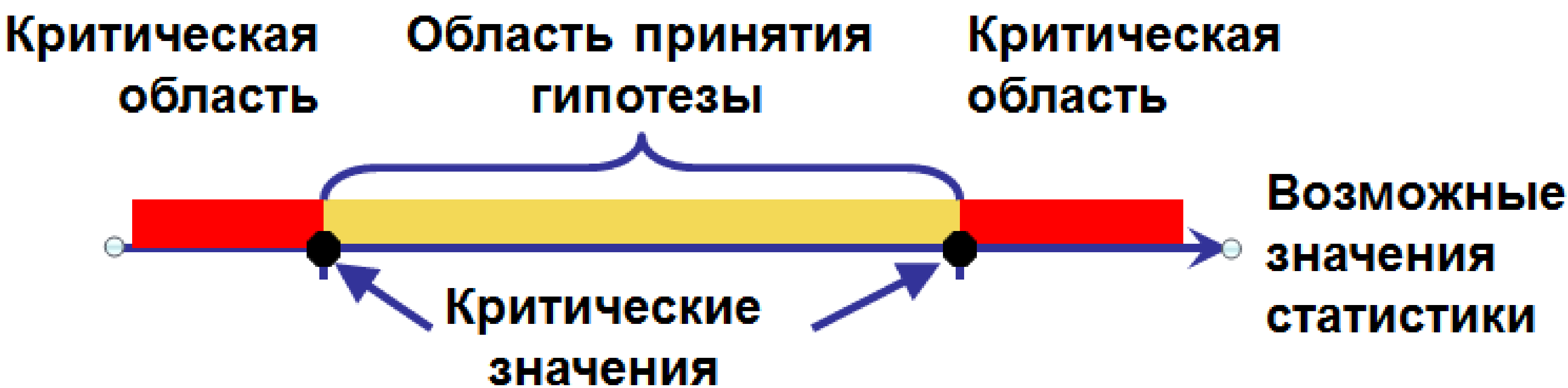
Если $T_{\text{набл}} \notin K$, то гипотеза H_0 принимается, но сам по себе тот факт, что $T_{\text{набл}} \notin K$, не является доказательством истинности H_0.

Также и тот факт, что $T_{\text{набл}} \in K$, не является доказательством истинности H_1.

Раздел 5. Статистическая проверка гипотез

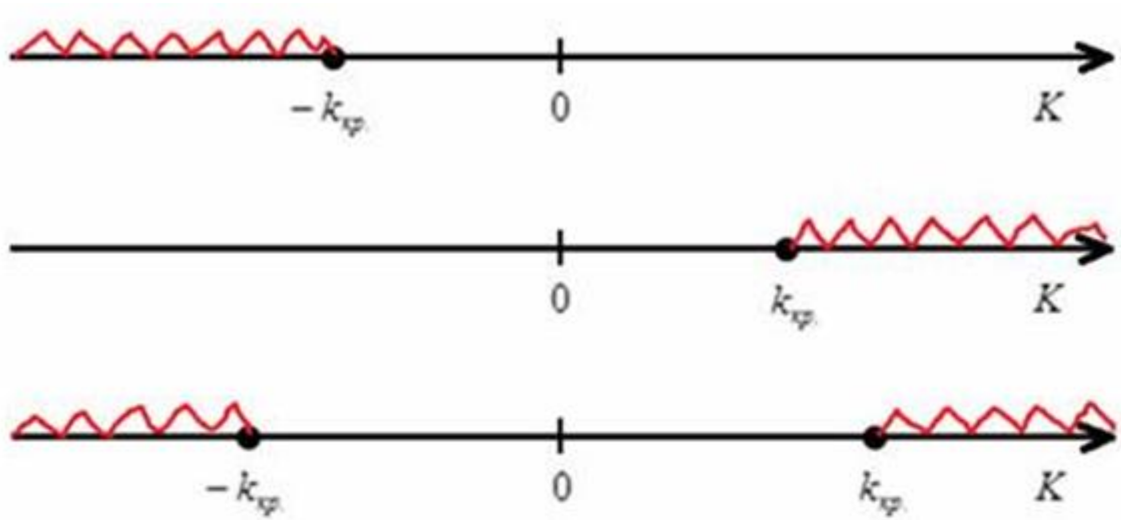
Способ выбора критической области К зависит от вида статистики T и гипотез H_0 и H_1.

Критическая область ограничена **критическими значениями**, или **границами критической области**



Раздел 5. Статистическая проверка гипотез

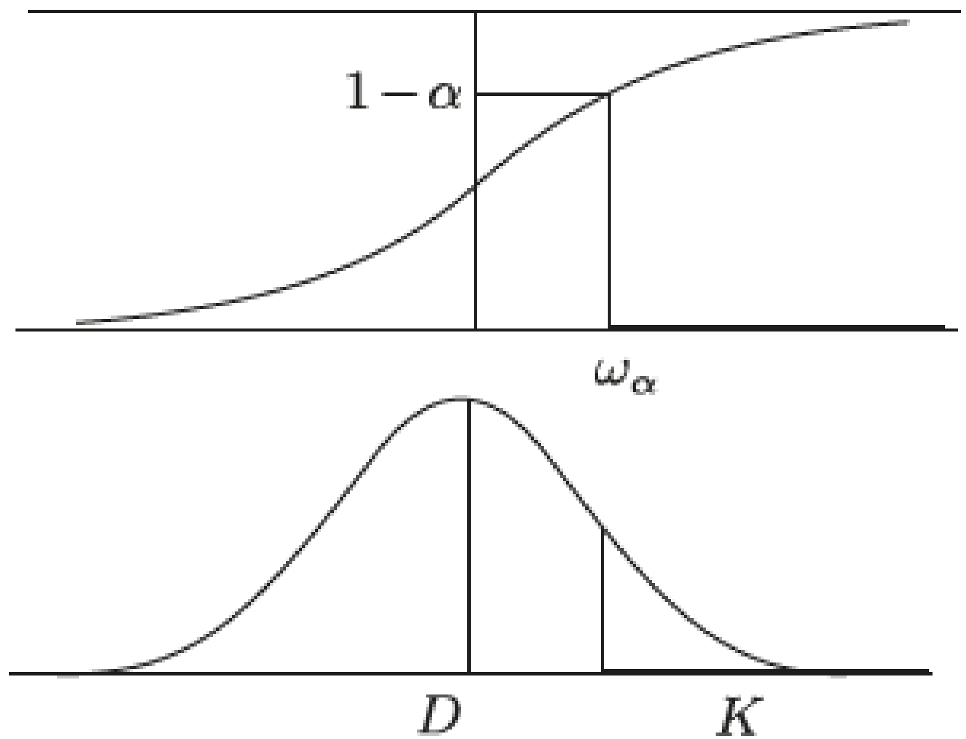
Как правило, критическая область задается одним из **следующих трех способов**.



Раздел 5. Статистическая проверка гипотез

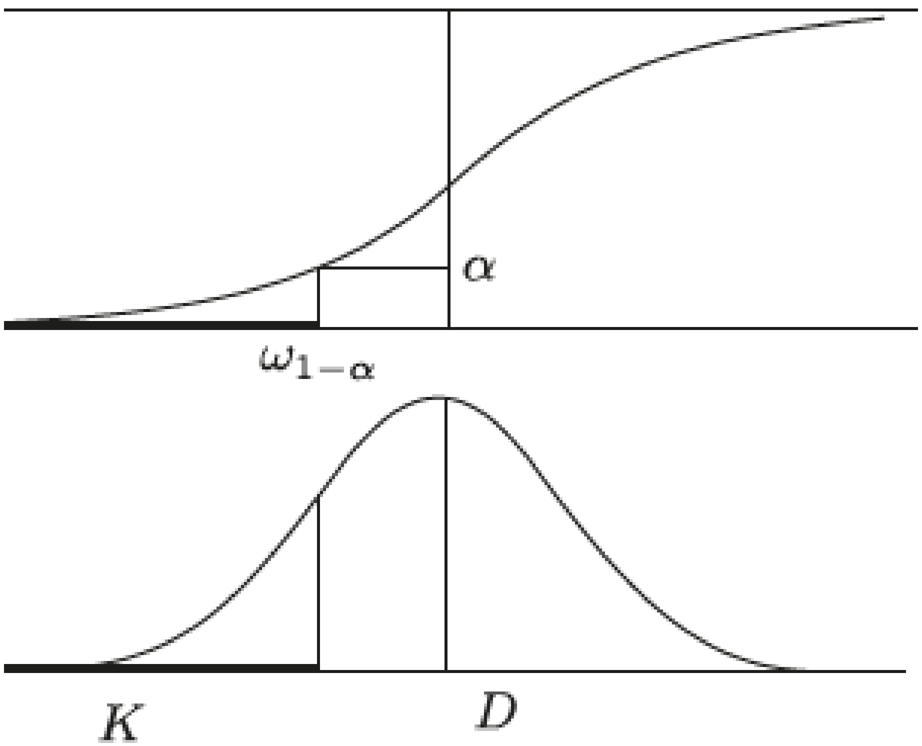
1. **Правосторонняя критическая область**:

$$K = \{\vec{x}: T(\vec{x}) > c\}.$$



2. Левосторонняя критическая область:

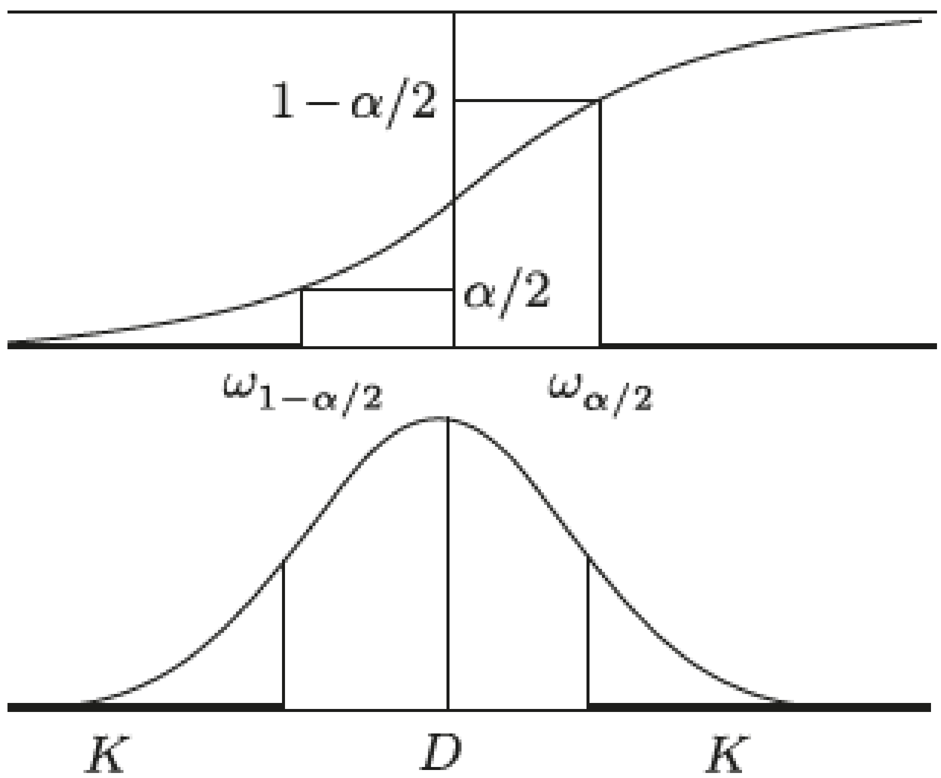
$$K = \{\vec{x} : T(\vec{x}) < c\}.$$



Раздел 5. Статистическая проверка гипотез

3. Двусторонняя критическая область:

$$K = \{\vec{x} : T(\vec{x}) < c_1\} \cup \{\vec{x} : T(\vec{x}) > c_2\}.$$



Раздел 5. Статистическая проверка гипотез

Функция $T(\vec{X})$ называется **статистикой критерия**.

Число c (или числа c_1 , c_2) называется **критическим значением статистики критерия**.

Раздел 5. Статистическая проверка гипотез

Если используется **правосторонняя критическая область**, то статистику критерия $T(\vec{X})$ и константу c выбирают из следующих соображений:

1) случайная величина $T(\vec{X})$ должна иметь некоторое известное распределение вероятностей, для которого имелась бы возможность вычислять вероятности вида $T(\vec{X}) > c$;

2) вероятность события $T(\vec{X}) > c$ (т. е. события, при осуществлении которого гипотеза H_0 отклоняется), в ситуации, когда гипотеза H_0 верна, должна не превосходить некоторого заданного и близкого к нулю уровня.

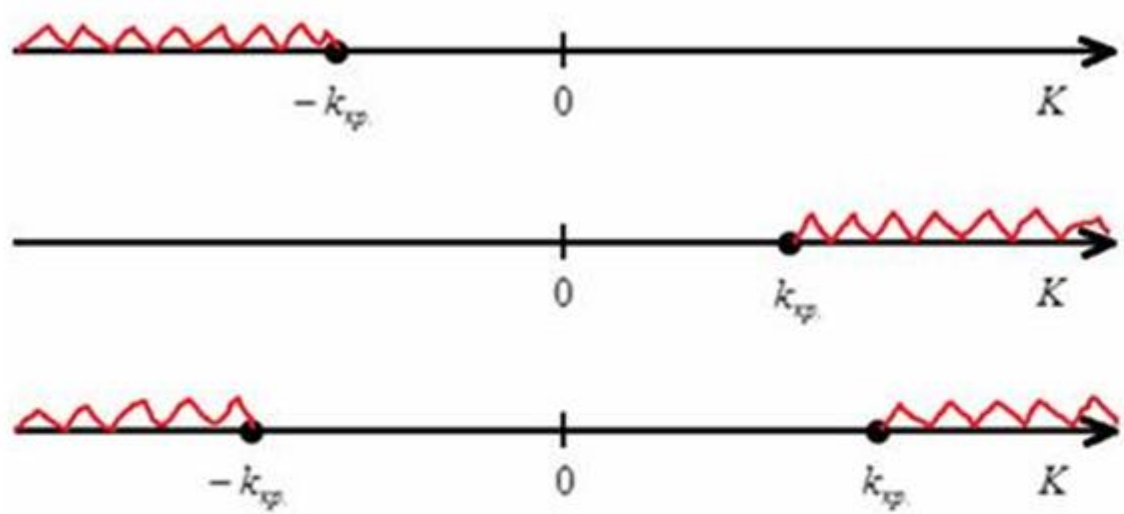
Раздел 5. Статистическая проверка гипотез

Аналогичные правила используются и для других способов задания критической области.

Раздел 5. Статистическая проверка гипотез

Если значение статистики попало в область принятия гипотезы, то гипотеза H_0 принимается

Если значение статистики попало в критическую область, то гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1



Раздел 5. Статистическая проверка гипотез

Общий принцип проверки статистических гипотез

1. Формулируем основную и альтернативную гипотезы
2. Задаем уровень значимости
3. Выбираем статистику - критерий проверки гипотезы
4. Определяем критическую область
5. Вычисляем значение статистики по выборке
6. Сравниваем значение статистики с критической областью, делаем вывод

Раздел 5. Статистическая проверка гипотез

При использовании любого критерия возможны **ошибки** следующих видов:

1) отклонить гипотезу H_0, когда в действительности она верна. Такое ошибочное решение называется **ошибкой первого рода**;

2) принять гипотезу H_0, когда в действительности она неверна. Такое решение называется **ошибкой второго рода**.

Раздел 5. Статистическая проверка гипотез

Вероятность ошибки первого рода называется **уровнем значимости** критерия и обозначается α.

В практических задачах значение уровня значимости α выбирают равной 0,1; 0,05; 0,01 и т.д.

Вероятность ошибки второго рода обозначается β, а величина $1 - \beta$ называется **мощностью** критерия.

Раздел 5. Статистическая проверка гипотез

ошибка I рода	*ошибка II рода*
Отвергается основная (нулевая) гипотеза, хотя она верна.	Отвергается конкурирующая гипотеза, хотя она верна.
Вероятность ошибки $P(H_1 \vert H_0) = \alpha$, α – ***уровень значимости критерия*** (обычно $\alpha = 0{,}05$; 0,01; 0,005; 0,001).	Вероятность ошибки $P(H_0 \vert H_1) = \beta$ (величина β, как правило, заранее неизвестна)
Вероятность принять верную (нулевую) гипотезу $P(H_0 \vert H_0) = 1 - \alpha$.	Вероятность принять верную (конкурирующую) гипотезу $P(H_1 \vert H_1) = 1 - \beta$, $(1 - \beta)$ – ***мощность критерия***.

Раздел 5. Статистическая проверка гипотез

Аналитик сам управляет **ошибкой первого рода** - задаёт вероятность её наступления.

Ошибкой второго рода он управлять не может - всегда существует вероятность того, что может быть принята неверная гипотеза.

Поэтому, чтобы избежать нежелательных последствий от принятия неверной гипотезы, основная гипотеза формулируется **таким образом**, чтобы риск от последствий принятия неверной гипотезы был минимальным.

Раздел 5. Статистическая проверка гипотез

Во многих случаях масштаб негативных последствия от **ошибок первого и второго рода** различается:

- положить в больницу здорового человека менее опасно, чем не лечить больного;
- для общества значительно опаснее ошибка, связанная с осуждением невиновного, чем с оправданием виновного; и т.п.

Поэтому при проверке гипотез бывает полезно оценивать ожидаемые доходы и убытки.

Раздел 5. Статистическая проверка гипотез

Поэтому, статистические гипотезы часто формулируются следующим образом:

"**статистическая значимость между факторами незначима**",

"**выборки незначимо отличаются по своим свойствам**",

"**фактор не имеет значимого влияния на исследуемый процесс**".

Раздел 5. Статистическая проверка гипотез

Мощность критерия показывает, насколько сильна статистика, насколько тест может обнаруживать ошибки.

Чем больше мощность, тем меньше вероятность не отвергнуть неверную гипотезу.

Увеличить мощность критерия можно двумя основными способами:

Раздел 5. Статистическая проверка гипотез

1) Увеличить размер выборки (размер определяет ошибку выборки: с увеличением числа наблюдений уменьшается стандартная ошибка → увеличивается мощность).

<u>Проблема</u>: не всегда возможно увеличить выборку (ограниченность ресурсов)

Раздел 5. Статистическая проверка гипотез

2) Увеличить область отвержения гипотезы: С увеличением области отвержения увеличивается вероятность отвергнуть верную гипотезу – ошибка первого рода (при этом увеличивается вероятность получить статистически значимый результат) → когда увеличивается ошибка первого рода, ошибка второго рода (β) уменьшается → мощность увеличивается (между β и мощностью обратная зависимость).

***Проблема**:* при таком способе увеличении мощности мы увеличиваем вероятность ошибки первого рода (нельзя одновременно уменьшить обе ошибки).

Раздел 5. Статистическая проверка гипотез

+ еще можно уменьшить дисперсию совокупности (примерно тот же эффект, что и от увеличения выборки).

Этого можно добиться, например, увеличивая точность измерений.

Раздел 5. Статистическая проверка гипотез

Например, основная гипотеза состоит в том, что предприятие получает прибыль.

Если это правильная гипотеза, то ошибка первого рода состоит в том, что данная гипотеза отвергается.

Если принимается решение о том, что прибыль предприятие не получает, то это ошибка второго рода.

Обычно ошибка первого рода влечет за собой ошибку второго рода: если отвергнута гипотеза о том, что предприятие получает прибыль, то, естественно, принимается решение о том, что оно не имеет прибыли.

Раздел 5. Статистическая проверка гипотез

Однако на практике возможны и другие ситуации. В большинстве случаев рассматриваются гипотезы о законах распределения.

Если отвергается правильный закон распределения, то совершается ошибка первого рода. Но после этого может быть принято решение уточнить данные, т.е. другая гипотеза не принимается.

Если же принимается другое распределение, то совершается ошибка второго рода.

Иногда ошибку первого рода называют «**альфа-риск**», а ошибку второго рода – «**бета-риск**».

Раздел 5. Статистическая проверка гипотез

Величина β равна той доле наблюдений, для которой значения статистики критерия проверки H_0 не попадают в критическую область критерия проверки некоторой полностью определенной гипотезы $H_0' \in H_1$, конкретизирующей сложную гипотезу H_1.

По этой причине бессмысленно говорить просто о мощности критерия, имеет смысл только термин «мощность критерия относительно гипотезы H_0».

Раздел 5. Статистическая проверка гипотез

Ошибку первого рода часто называют **ложной тревогой** —

например, анализ крови показал наличие заболевания, хотя на самом деле человек здоров.

Раздел 5. Статистическая проверка гипотез

Лемма Неймана-Пирсона

Раздел 5. Статистическая проверка гипотез

При построении критерия для проверки статистической гипотезы, как правило, исходят из необходимости максимизировать мощность критерия $w = 1 - \beta$ при фиксированном уровне значимости α.

Предположим, что генеральное распределение имеет зависящую от параметра θ положительную при всех x плотность $f(x;\theta) > 0$.

Пусть H_0 и H_1 – простые гипотезы вида

$$H_0 : \theta = \theta_0 \text{ и } H_1 : \theta = \theta_1.$$

Запишем **функции правдоподобия**, соответствующие этим гипотезам:

$$L_0(\theta_0, x_1 \dots, x_n) = f(x_1;\theta_0) \cdot \dots \cdot f(x_n;\theta_0),$$
$$L_1(\theta_1, x_1 \dots, x_n) = f(x_1;\theta_1) \cdot \dots \cdot f(x_n;\theta_1).$$

Раздел 5. Статистическая проверка гипотез

Теорема (**лемма Неймана–Пирсона**). Для любого $\alpha \in (0,1)$ существует такая константа c_α, что критерий с критической областью

$$\frac{L_1(\theta_1, x_1 \dots, x_n)}{L_0(\theta_0, x_1 \dots, x_n)} > c_\alpha,$$

является наиболее мощным критерием среди всех статистических критериев с какой-либо критической областью K, предназначенных для проверки H_0 против H_1 с уровнем значимости α.

Доказательство. (Без доказательства.)

Раздел 5. Статистическая проверка гипотез

Использование наиболее мощного критерия гарантирует при заданной вероятности α ошибки первого рода наименьшую, по сравнению с другими критериями, вероятность ошибки β второго рода, поскольку мощность критерия равна (1- β).

Раздел 5. Статистическая проверка гипотез

Пример. Пусть имеется выборка $X_1 \dots, X_n$ из распределения $N(m;\ \sigma^2)$ с известной дисперсией σ^2. Проверяется гипотеза $H_0 : m\ = m_0$ против альтернативы $H_1 : m\ = m_1$.

Предположим для определенности, что выполняется неравенство: $m_0\ < m_1$. Используем следующую **статистику**:

$$Z = \frac{\bar{X} - m_0}{\sigma/\sqrt{n}}.$$

Задается **правосторонняя критическая область**: $K = (z_\alpha;\ \infty)$.

Покажем, что так определенный критерий является наиболее мощным среди всех критериев по проверке H_0 против H_1 с уровнем значимости α.

Решение.

P-значение критерия

Раздел 5. Статистическая проверка гипотез

Описанная процедура проверки статистической гипотезы позволяет принять решение о принятии или отклонении основной гипотезы по результатам сравнения наблюдаемого значения статистики с критическим значением, которое зависит от заданного уровня значимости. Если будет задано новое значение уровня значимости, процедуру проверки нужно проводить заново.

Выбор уровня значимости довольно произволен, и для одной и той же выборке может оказаться так, что при одном значении уровня значимости проверяемая гипотеза отвергается, а при другом значении уровня значимости — не отвергается.

Раздел 5. Статистическая проверка гипотез

При проверке гипотез оказывается полезен подход, основанный на вычислении наблюдаемого уровня значимости.

Метод, основанный на использовании так называемого P-значения критерия, позволяет принять решение о принятии или отклонении основной гипотезы одновременно для всех уровней значимости.

Раздел 5. Статистическая проверка гипотез

Пусть имеется статистический критерий $T(\vec{X})$ с критической областью K. И пусть получена реализация $\vec{x}$ случайной выборки $\vec{X} = (X_1, \dots, X_n)$.

P*-значением** (P*-value**) $p(\vec{x})$ статистического критерия $T(\vec{x})$ называется наименьшая величина уровня значимости, при котором нулевая гипотеза отклоняется:

$$p(\vec{x}) = \min\{\alpha \colon T(\vec{x}) \in K\}.$$

Величина $p(\vec{x})$ задает фактический уровень значимости.

Раздел 5. Статистическая проверка гипотез

Для всех значений уровня значимости, таких, что $\boldsymbol{\alpha} \le \boldsymbol{p}(\vec{\mathbf{x}})$, гипотеза $\boldsymbol{H_0}$ принимается, при всех $\boldsymbol{\alpha} > \boldsymbol{p}(\vec{\mathbf{x}})$ гипотеза $\mathbf{H_0}$ отклоняется.

Чем меньше P-значение, тем сильнее основания отклонить нулевую гипотезу.

P **-значением** (***P*** **-value**) $p(\vec{x})$ статистического критерия $T(\vec{x})$ называется наименьшая величина уровня значимости, при котором нулевая гипотеза отклоняется:

$$p(\vec{x}) = \min\{\alpha\colon\ T(\vec{x}) \in K\}.$$

Раздел 5. Статистическая проверка гипотез

Способы задания критических областей

Пусть $T = T(\vec{X})$ — статистика критерия, ω_q — $100q$%-я точка статистки критерия, α — уровень значимости.

1. **Правосторонняя критическая область**. Гипотеза H_0 отклоняется, если выполняется неравенство $T_{\text{набл}} > \omega_\alpha$.

2. **Левосторонняя критическая область**. Гипотеза H_0 отклоняется, если выполняется неравенство $T_{\text{набл}} < \omega_{1-\alpha}$.

3. **Двусторонняя критическая область**. Гипотеза H_0 отклоняется, если выполняется условие:

$$T_{\text{набл}} \in (\infty; \omega_{1-\alpha/2}) \cup (\omega_{\alpha/2}; \infty).$$

Раздел 5. Статистическая проверка гипотез

Приведенные различные варианты задания критических областей можно описать единым образом.

Рассмотрим двустороннюю критическую область вида

$$K = \{\vec{x}\colon T(\vec{x}) < t_{1cr}\} \cup \{\vec{x}\colon T(\vec{x}) > t_{2cr}\},$$

представляющую собой объединение областей $(-\infty; t_{1cr})$ и $(t_{2cr};\ \infty)$.

Разобьем суммарный уровень значимости α, задающий вероятность попадания в критическую область, на два слагаемых:

$$\alpha = \alpha_1 + \alpha_2 = \alpha\delta + \alpha(1 - \delta).$$

В этом выражении α_1 и α_2 представляют собой вероятности попадания, соответственно, в левую и правую части критической области, а весовой коэффициент δ задает «важность» для лица, принимающего решение, первой или второй части критической области.

Раздел 5. Статистическая проверка гипотез

Описанные типовые способы задания критических областей получаются следующим образом:

– случай $\delta \to 0$ соответствует **правосторонней критической области**;
– случай $\delta \to 1$ соответствует **левосторонней критической области**;
– случай $\delta = 1/2$ соответствует **двусторонней критической области**.

Сформулируем теперь **правила вычисления Р-значения** для тех критериев, в которых нулевая гипотеза принимается согласно приведенным формулам.

$$\alpha = \alpha_1 + \alpha_2 = \alpha\delta + \alpha(1 - \delta)$$

Раздел 5. Статистическая проверка гипотез

Теорема. Пусть $T = T(\vec{X})$ — статистика критерия, $T_{\text{набл}} = T(\vec{x})$ — наблюдаемое значение статистики критерия.

Если критическая область имеет вид: $\{T_{\text{набл}} > \omega_\alpha\}$, где ω_α — 100 α -процентная точка статистики T (**правосторонняя критическая область**), то P-значение $p_1(\vec{x})$ находится по формуле: $p_1(\vec{x}) = P_{H_0}(\{T > T_{\text{набл}}\})$, где через $P_{H_0}(A)$ обозначена вероятность события A, вычисленная в предположении справедливости гипотезы H_0.

Если критическая область имеет вид: $\{T_{\text{набл}} < \omega_{1-\alpha}\}$ (**левосторонняя критическая область**), то P-значение $p_2(\vec{x})$ находится по формуле: $p_2(\vec{x}) = P_{H_0}(\{T < T_{\text{набл}}\}) = 1 - p_1(\vec{x})$.

Если критическая область имеет вид: $T_{\text{набл}} \in (-\infty; \omega_{1-\frac{\alpha}{2}}) \cup (\omega_{\frac{\alpha}{2}}; +\infty)$ (**двусторонняя критическая область**), то P-значение $p_3(\vec{x})$ находится по формуле: $p_3(\vec{x}) = 2 \cdot \min\{p_1(\vec{x}), p_2(\vec{x})\}$.

Раздел 5. Статистическая проверка гипотез

Поясним на примере проверки *односторонней гипотезы.*

Формулировка задачи.

Из *генеральной совокупности* имеющей **нормальное распределение** с неизвестным μ и известной *дисперсией* σ^2 взята *выборка* размера n. Необходимо проверить *статистическую гипотезу* о равенстве неизвестного μ заданному значению μ_0.

Иными словами:

$$H_0\colon \mu = \mu_0$$

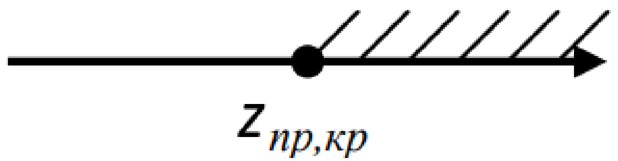
При альтернативной гипотезе:

$$H_1\colon \mu > \mu_0$$

Раздел 5. Статистическая проверка гипотез

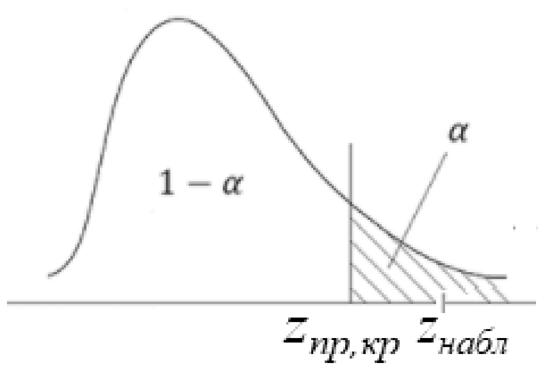
$H_0\colon \mu = \mu_0 \quad H_1\colon \mu > \mu_0$

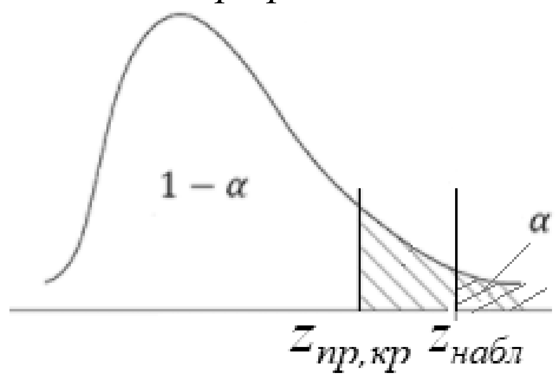
Критическая область – **правосторонняя**:



$\Rightarrow P(Z > z_{пр,кр}) = \alpha$

$P(Z > z_{пр,кр}) = \alpha \Rightarrow P(Z < z_{пр,кр}) = 1 - P(Z > z_{пр,кр}) = 1 - \alpha$





$P(Z < z_{набл}) = 1 - \alpha' = 1 - \boldsymbol{p_{value}}$

$p_1(\vec{x}) = \boldsymbol{p_{value}} = 1 - P(Z < z_{набл})$ = **Z.ТЕСТ**(<*выборка*>; μ_0 ;σ)

=1-**НОРМ.СТ.РАСП**($z_{набл}$;1)

Раздел 5. Статистическая проверка гипотез

Критическая область – **левосторонняя**:

H_0: $\mu = \mu_0$ H_1: $\mu < \mu_0$

$P(Z< z_{набл})$ = α' = $\boldsymbol{p_{value}}$

$p_2(\vec{x})$ = $\boldsymbol{p_{value}}$ =$P(Z< z_{набл})$

P-value =1- **Z.ТЕСТ**(<*выборка*>; μ_0;σ)

=**НОРМ.СТ.РАСП**($z_{набл}$;1)

Раздел 5. Статистическая проверка гипотез

Критическая область – **двусторонняя**:

H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$

$P(Z< z_{набл})$ =1- α' /2 = 1- $\boldsymbol{p_{value}}$ /2

$P(Z< -z_{набл})$ =α' /2 = $\boldsymbol{p_{value}}$ /2

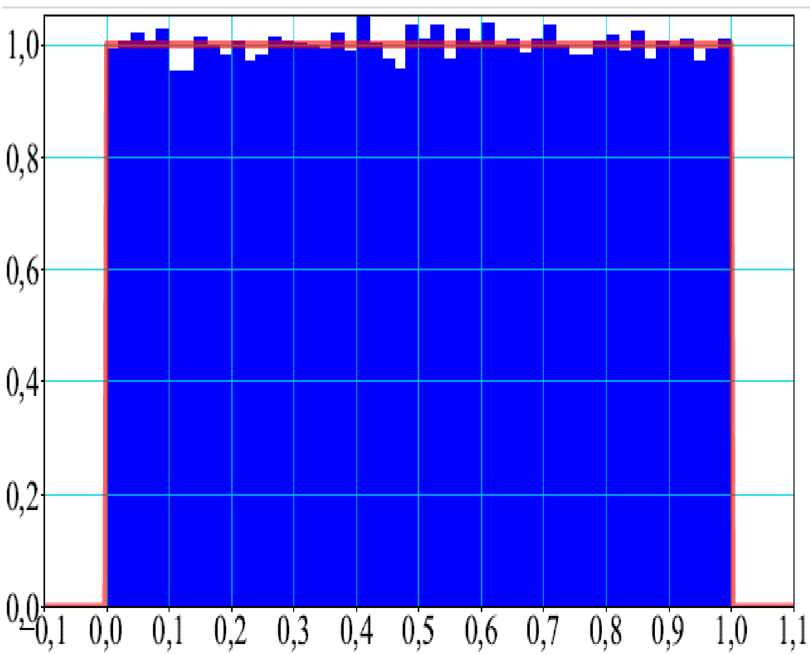
$p_з(\vec{x})$ = $\boldsymbol{p_{value}}$ =

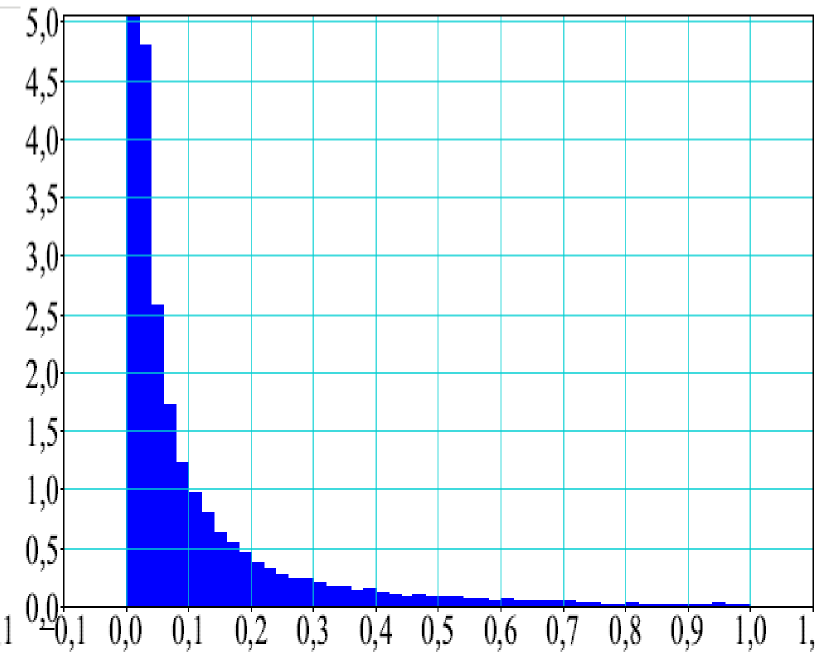
=2***МИН**(**Z.ТЕСТ**(<*выборка*>; μ_0;σ);1-**Z.ТЕСТ**(<*выборка*>; μ_0 ;σ))

=2***МИН**(1-**НОРМ.СТ.РАСП**($z_{набл}$;1); **НОРМ.СТ.РАСП**($z_{набл}$;1))

Раздел 5. Статистическая проверка гипотез

Теорема. Если H_0 – верна, то случайная величина $PV \sim Unif[0;1]$ (то есть распределение с.в. PV является равномерным на $[0;1]$).





Распределение с.в. PV, когда

1) H_0 – верна; 2) H_0 – отвергается.

Критерий значимости коэффициента корреляции

Раздел 5. Статистическая проверка гипотез

Проверяется **гипотеза о наличии значимой линейной корреляционной связи** между случайными величинами X и Y.

Для проверки гипотезы используется **выборочный коэффициент корреляции $\boldsymbol{\rho_{xy}}$**.

В случае отклонения гипотезы **о равенстве коэффициента корреляции нулю** говорят, что он **значим**.

В противном случае говорят: $\boldsymbol{\rho_{xy}}$ **не значим**.

Раздел 5. Статистическая проверка гипотез

Итак, проверяется гипотеза

H_0: $\rho_{xy} = 0$;

H_1: $\rho_{xy} \neq 0$.

Задан **уровень значимости** α.

Раздел 5. Статистическая проверка гипотез

Проверка гипотезы основана на следующем результате:

при справедливости нулевой гипотезы **статистика**

$$T = \frac{\tilde{\rho}_{xy}}{\sqrt{1 - \tilde{\rho}_{xy}^2}} \sqrt{n-2},$$

где $\tilde{\rho}_{xy}$ — выборочный коэффициент корреляции, имеет **распределение Стьюдента** с $n - 2$ степенями свободы.

Раздел 5. Статистическая проверка гипотез

Гипотеза H_0 отклоняется, если вычисленное по выборочным данным значение статистики $T_{\text{набл}}$ удовлетворяет неравенству:

$$|T_{\text{набл}}| > t_{\frac{\alpha}{2};n-2}.$$

Действительно, при условии справедливости гипотезы H_0 имеем:

$$T \sim t(n-2).$$

Следовательно, **вероятность ошибки первого рода**

$$P\left(\left\{|T| > t_{\frac{\alpha}{2};n-2}\right\}\right) =$$
$$= P\left(\left\{T < -t_{\frac{\alpha}{2};n-2}\right\}\right) + P\left(\left\{T > t_{\frac{\alpha}{2};n-2}\right\}\right) =$$
$$= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

Раздел 5. Статистическая проверка гипотез

Пример. Имеются данные о расходах на рекламу (X) и объемах продаж (Y) шести моющих средств:

Расходы на рекламу (ден. ед.)	2,3	5,7	4,8	7,3	5,9	6,2
Объем продаж (шт.)	77	105	96	118	102	95

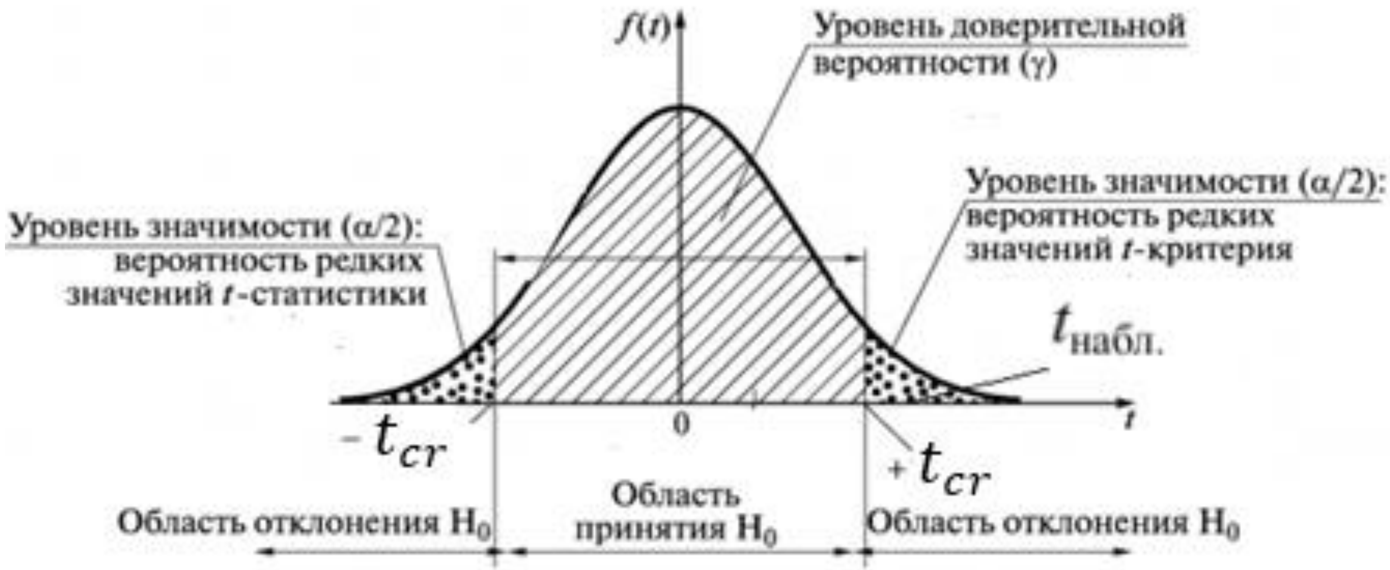
Требуется проверить, существует ли значимая линейная корреляционная связь между X и Y на уровне значимости $\alpha = 0{,}05$.

Решение.

H_0: $\rho_{xy} = 0$;

H_1: $\rho_{xy} \neq 0$

Меры связи случайных величин



Раздел 5. Статистическая проверка гипотез

Параметрические критерии

Проверка гипотезы о равенстве математического ожидания теоретическому значению.
Проверка гипотезы о равенстве двух математических ожиданий.
Проверка гипотезы о равенстве дисперсии теоретическому значению.
Проверка гипотезы о равенстве двух дисперсий. Использование аппарата проверки гипотез в экономике и управлении. Реализация критериев проверки статистических гипотез в пакете Microsoft Excel и средствами Python .

Раздел 5. Статистическая проверка гипотез

Гипотеза о генеральной средней нормального распределения

Постановка задачи: предполагается, что **генеральная средняя нормального распределения** равна некоторому значению μ_0.

Это нулевая гипотеза: H_0: $\mu = \mu_0$.

Раздел 5. Статистическая проверка гипотез

Для проверки гипотезы на уровне значимости α проводится **выборка** объема n и рассчитывается **выборочная средняя** $\overline{X}_в$. Исходя из полученного значения и специфики той или иной задачи, можно сформулировать следующие конкурирующие гипотезы:

1) H_1: $\mu < \mu_0$
2) H_1: $\mu > \mu_0$
3) H_1: $\mu \neq \mu_0$
4) H_1: $\mu = \mu_1$,

где μ_1 – конкретное альтернативное значение *генеральной средней*.

Раздел 5. Статистическая проверка гипотез

Если имеется достоверная информация о направлении отклонения от нулевой гипотезы, то использование односторонней альтернативы предпочтительнее двусторонней, поскольку это повышает мощность критерия.

Раздел 5. Статистическая проверка гипотез

При этом возможны две принципиально разные ситуации:

а) если генеральная дисперсия σ^2 известна.

б) генеральная дисперсия σ^2 НЕ известна.

Раздел 5. Статистическая проверка гипотез

а) если генеральная дисперсия σ^2 известна.

Тогда в качестве *статистического критерия* *K* рассматривают случайную величину

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

где $\bar{X}$ – случайное значение *выборочной средней*.

Почему случайное?

Потому что в разных выборках мы будем получать разные значения выборочной средней, и заранее предугадать это значение невозможно.

Раздел 5. Статистическая проверка гипотез

Далее находим ***критическую область К***.

Для конкурирующих гипотез

$H_1\colon\ \mu < \mu_0$ и $H_1\colon\ \mu = \mu_1$ (случай $\mu_1 < \mu_0$) строится ***<u>левосторонняя область</u>***,

$H_1\colon\ \mu > \mu_0$ и $H_1\colon\ \mu = \mu_1$ (случай $\mu_1 > \mu_0$) –***<u>правосторонняя</u>***,

$H_1\colon\ \mu \neq \mu_0$ – ***<u>двусторонняя</u>*** – так как конкурирующее значение генеральной средней может оказаться как больше, так и меньше μ_0-го.

Раздел 5. Статистическая проверка гипотез

Чтобы найти критическую область *K* нужно отыскать критическое значение $z_{кр}$.

Оно определяется из соотношения

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2}$$

– для одностронней области *(лево- или право-)* и

$$\Phi(z_{кр}) = \frac{1-\alpha}{2}$$

– для двусторонней области, где α – выбранный уровень значимости, а $\Phi(z)$ – старая знакомая функция Лапласа.

Раздел 5. Статистическая проверка гипотез

Теперь на основании выборочных данных рассчитываем *наблюдаемое значение критерия:*

$$z_{\text{набл}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

это можно было сделать и раньше, но такой порядок более последователен и логичен.

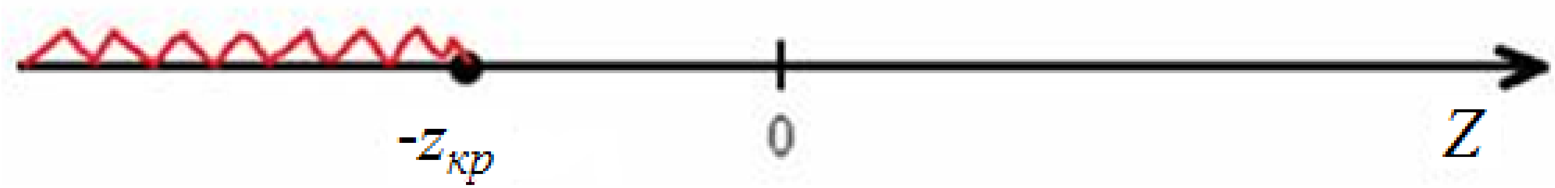
Результаты:

Раздел 5. Статистическая проверка гипотез

1) Для <u>левосторонней критической области</u>.

Если $z_{набл}>-z_{кр}=z_{1-\alpha}$, то гипотеза H_0 на уровне значимости α принимается.

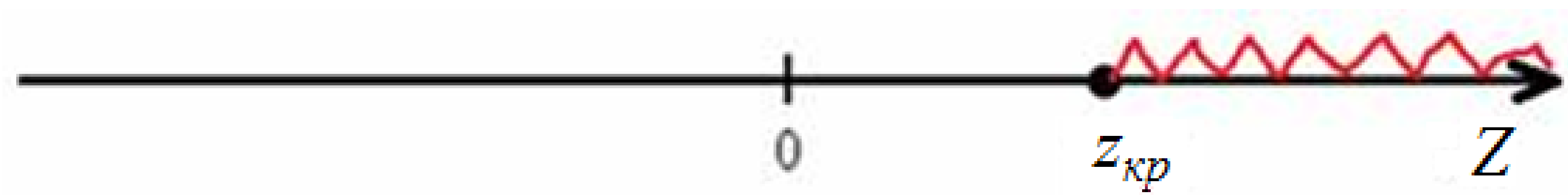
Если $z_{набл}<-z_{кр}=-z_{1-\alpha}$, то отвергается *(красный цвет)* .



2) Правосторонняя критическая область.

Если $z_{набл}<z_{кр}=z_{1-\alpha}$, то гипотеза H_0 на уровне значимости α принимается.

Если $z_{набл}>z_{кр}=z_{1-\alpha}$, то отвергается *(красный цвет)* .

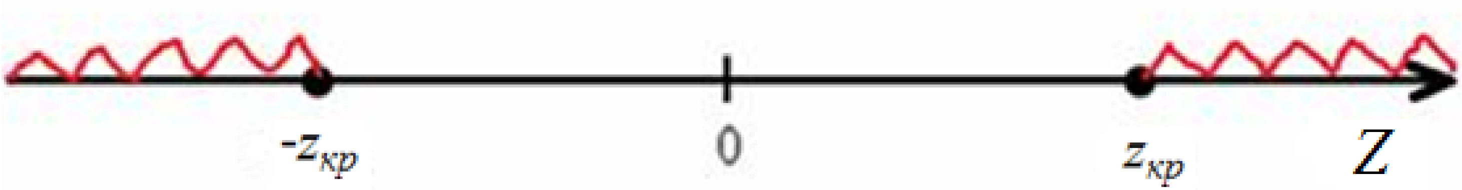


Раздел 5. Статистическая проверка гипотез

3) Двусторонняя критическая область.

Если

$-z_{кр}=-z_{1-\alpha/2}<z_{набл}<z_{кр}=z_{1-\alpha/2}$ *(незаштрихованный интервал),* то гипотеза H_0 принимается , в противном случае – отвергается:



условие принятия гипотезы часто записывают компактно – с помощью **модуля**: $|z_{наб}|<z_{кр}$.

Раздел 5. Статистическая проверка гипотез

$z_{кр} = z_{1-\alpha} = \mathbf{НОРМ.СТ.ОБР}(<1 - \alpha>)$

— квантиль уровня $1 - \alpha$ случайной величины *Z*, распределенной по стандартному нормальному закону

```
import scipy.stats as sts

zcr=sts.norm.ppf(1-alpha)
```

Раздел 5. Статистическая проверка гипотез

$z_{кр} = z_{1-\alpha/2}$= **НОРМ.СТ.ОБР**(<1 – α/2>)
— квантиль уровня 1 – α/2 случайной величины *Z*, распределенной по стандартному нормальному закону.

```
import scipy.stats as sts

zcr=sts.norm.ppf(1-alpha/2)
```

В силу симметричности распределения

$$z_{1-\alpha} = -z_{\alpha} \text{ и } z_{1-\alpha/2} = -z_{\alpha/2}$$

Раздел 5. Статистическая проверка гипотез

```
import scipy.stats as sts
```

```
alpha=0.05
```

```
zcr=sts.norm.ppf(1-alpha/2)
```

```
zcr
```

```
1.959963984540054
```

```
zcr=sts.norm.isf(alpha/2)
```

```
zcr
```

```
1.9599639845400545
```

Раздел 5. Статистическая проверка гипотез

В силу симметричности распределения

$$z_{1\text{-}\alpha} = -z_{\alpha} \text{ и } z_{1\text{-}\alpha/2} = -z_{\alpha/2}$$

```
import scipy.stats as sts
```

```
alpha=0.05
```

```
z_right_cr=sts.norm.ppf(1-alpha/2)
z_left_cr=sts.norm.ppf(alpha/2)
```

```
print('z_right_cr=',z_right_cr)
print('z_left_cr=',z_left_cr)
```

```
z_right_cr= 1.959963984540054
z_left_cr= -1.9599639845400545
```

Раздел 5. Статистическая проверка гипотез

Если, например, известно, что отклонение математического ожидания μ от гипотетического значения μ_0 может произойти только в большую сторону, то в качестве альтернативы следует взять гипотезу H_1: $\mu > \mu_0$ (**односторонняя**).

Критическая область уровня α в этом случае будет состоять не из двух бесконечных полуинтервалов $(-\infty;-z_{1-\alpha/2}]$ и $[z_{1-\alpha/2};+\infty)$, а из одного - $[z_{1-\alpha};+\infty)$.

Пример. Случайная величина X распределена по нормальному закону с математическим ожиданием m и дисперсией $\sigma^2 = 25$. Проверяется гипотеза $H_0 : m = 0$ при альтернативе $H_1 : m > 0$. Для проверки гипотезы используется статистика

$$Z = \frac{\bar{X} - m_0}{\sigma/\sqrt{n}}$$

с **правосторонней критической областью** $Z > z_\alpha$. В результате испытаний получена выборка объема $n = 100$, такая, что $\bar{x} = 1$. Требуется вычислить P-значение критерия.

Решение.

Наблюдаемое значение **статистики** критерия:

$$Z_{\text{набл}} = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} = \frac{1}{5/\sqrt{10}} = 2.$$

Решение (продолжение). В данном примере критическая область **правосторонняя** и P-значение

$$p_1(\vec{x}) = P_{H_0}(\{Z > Z_{\text{набл}}\}).$$

Случайная величина Z, при условии справедливости гипотезы H_0, имеет **стандартный нормальный закон распределения** и

$$P_{H_0}(\{Z > Z_{\text{набл}}\}) = P_{H_0}(\{Z > 2\}) = 0{,}5 - \Phi(2) = 0{,}0228.$$

# MS Excel	# Python
=**1-НОРМ.СТ.РАСП(2;1)**	import scipy.stats as sts P_H0=sts.norm.sf(2) # P_H0=1-sts.norm.cdf(2) P_H0

Следовательно, при любом уровне значимости $\alpha > 0{,}0228$ нулевая гипотеза отклоняется.

При любом $\alpha \leq 0{,}0228$ нулевая гипотеза принимается.

Пример. Из нормальной генеральной совокупности с известной дисперсией σ^2 =3,2 извлечена выборка объёма n=25 и по ней найдена выборочная средняя $\bar{x} = 19{,}3$. Требуется на уровне значимости 0,01 проверить нулевую гипотезу $H_0\colon \mu = 20$ против конкурирующей гипотезы $H_1\colon \mu = 19$.

Раздел 5. Статистическая проверка гипотез

Прежде чем приступить к решению, пару слов о **смысле такой задачи**.

Есть **генеральная совокупность** с известной **дисперсией** и есть веские основания полагать, что **генеральная средняя** равна 20 (**нулевая гипотеза**).

В результате выборочной проверки получена **выборочная средняя** 19,3, и возникает **вопрос**:

это результат случайный или же генеральная средняя и на самом деле меньше двадцати? – в частности, равна 19 (**конкурирующая гипотеза**).

Раздел 5. Статистическая проверка гипотез

Решение: по условию, известна **генеральная дисперсия** σ^2 =3,2, поэтому для проверки гипотезы использует случайную величину

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Найдём критическую область.

Раздел 5. Статистическая проверка гипотез

Для этого нужно найти критическое значение. Так как конкурирующее $H_1\colon \mu = \mu_1 = 19$ значение **меньше** чем μ_0=20, то критическая область будет левосторонней.

Критическое значение определим из соотношения:

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2} = 0,5 - \alpha = 0,5 - 0,01 = 0,49$$

для уровня значимости α=0,01 $\Rightarrow$

$$z_{кр} = z_{1-\alpha} = \textbf{НОРМ.СТ.ОБР}(1\text{-}0,01) \approx 2,326348$$

или

$$z_{кр} = \textbf{НОРМСТОБР}(0,49+0,5) \approx 2,326348$$

Решение (продолжение).

Критическое значение статистики:

$$z_{cr} = z_{1-\alpha} = 2{,}326348.$$

 # MS Excel $z_{cr} = 2,33$

	A	B	C	D
1	alpha=	0,01		
2	zcr=	2,326348	=НОРМ.СТ.ОБР(1-B1)	

 # Python

```
import scipy.stats as sts
```

```
alpha=0.01
```

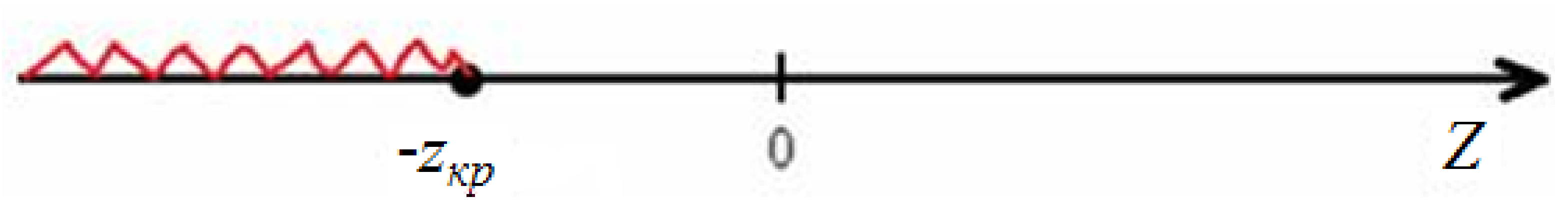
```
# zcr=sts.norm.ppf(1-alpha)
  zcr=sts.norm.isf(alpha)
```

```
zcr
```

```
2.3263478740408408
```

Раздел 5. Статистическая проверка гипотез

Таким образом, при $z_{набл}$<-$z_{кр}$ *(красная критическая область)* нулевая гипотеза отвергается, а при $z_{набл}$>-$z_{кр}$ – принимается:



В данном случае -$z_{кр}$ ≈ -2,33.

Вычислим наблюдаемое значение критерия:

$$z_{набл} = \frac{(\overline{x}_в - a_0)\sqrt{n}}{\sigma} = \frac{(19,3-20)\cdot\sqrt{25}}{\sqrt{3,2}} =$$

$$= \frac{-0,7\cdot 5}{\sqrt{3,2}} = \frac{-3,5}{\sqrt{3,2}} \approx -1,96$$

Раздел 5. Статистическая проверка гипотез

Имеем:

$z_{набл}$=-1,96>-$z_{кр}$≈-2,33

поэтому **на уровне значимости** α=0,01 **нулевую гипотезу** H_0: $\mu = 20$ **принимаем**.

Раздел 5. Статистическая проверка гипотез

Такой, вроде бы неожиданный результат, объясняется тем, что генеральное стандартное отклонение достаточно великО: $\sigma = \sqrt{3,2} \approx 1,79$, а посему нет оснований отвергать «главное» значение $\mu_0 = 20$ (несмотря на то, что выборочная средняя $\overline{x}_в = 19,3$ гораздо ближе к конкурирующему значению $\mu_1 = 19$).

Иными словами, такое значение выборочной средней, вероятнее всего, объясняется естественным разбросом **вариант** x_i.

Раздел 5. Статистическая проверка гипотез

Что означает *«на уровне значимости 0,01»*?

Это означает, что мы с 1%-ной вероятностью рисковали отвергнуть нулевую гипотезу, при условии, что она действительно справедлива.

Однако не нужно забывать, что на самом деле она может быть и неверной и существует β-вероятность того, мы приняли неправильную гипотезу.

Пример. Руководство сети магазинов модной одежды должно принять решение, имеет ли смысл продлевать договор аренда помещения для магазина в одном из торговых центров. Из опыта сети магазинов известно, что магазин прибылен, если среднее число посетителей в день превышает $\mu_0 = 400$ чел. Также известна дисперсия σ^2=400. При этом за последние n=365 дней в данном магазине среднее число посетителей в день оказалось равно $\overline{x}_в = 404$ чел.

Раздел 5. Статистическая проверка гипотез

Пример.

По результатам n=5 измерений температуры в печи найдено $\overline{x}_в = 256^0\,\mathrm{C}$. Предполагается, что ошибка измерения есть нормальная случайная величина с σ=6^0 C. Проверить на уровне значимости α=0,05 гипотезу против конкурирующей гипотезы H_1: $a>250^0\,\mathrm{C}$.

Раздел 5. Статистическая проверка гипотез

Пример. Средний вес таблетки сильнодействующего лекарства (номинал) должен быть равен 0,5 мг. Выборочная проверка n=100 выпущенных таблеток показала, что средний вес таблетки равен $\overline{x}_в = 0,508$ мг.

Многократными предварительными опытами по взвешиванию таблеток, изготавливаемых фармацевтическим заводом, установлено, что вес таблеток распределен нормально со средним квадратическим отклонением σ=0,11 мг.

Требуется на уровне значимости α=0,1 проверить гипотезу о том, что средний вес таблеток действительно равен

$$H_0\colon\ \mu = 0{,}5.$$

Раздел 5. Статистическая проверка гипотез

Пример. По выборке объёма $n = 13$, извлечённой из нормальной генеральной совокупности с известным средним квадратичным отклонением $\sigma = 3$, при уровне значимости 0,05 проверяется нулевая гипотеза $H_0: \mu = \mu_0 = 17$. Требуется:

а) найти мощность критерия для проверки гипотезы $H_1: \mu = \mu_1 = 16$.

б) найти объём выборки n, при котором мощность критерия не меньше 0,92.

Решение.

Проверка гипотезы о равенстве заданному числу математического ожидания нормально распределенной случайной величины с неизвестной дисперсией (одновыборочный *t*-критерий)

На практике, значение σ не бывает известно, поэтому рассмотренными критериями проверки гипотез можно пользоваться только при достаточно больших выборках. В общем же случае при неизвестном значении σ можно использовать тот факт, что статистика

$$T_{n-1} = \frac{(\bar{X} - \mu_0)\sqrt{n}}{s}$$

распределена по закону Стьюдента с $n-1$ степенью свободы.

Итак:

Раздел 5. Статистическая проверка гипотез

б) генеральная дисперсия σ^2 НЕ известна.
(одновыборочный t-критерий)

В этом случае остаётся ориентироваться на **исправленную выборочную испесию** s^2 и критерий

$$T_{n-1} = \frac{(\bar{X} - \mu_0)\sqrt{n}}{s}$$

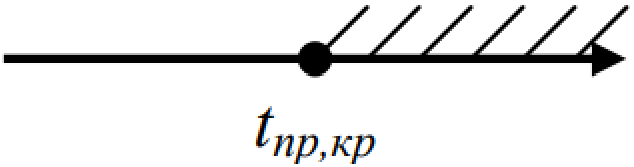
$\bar{X}$ случайное значение *выборочной средней* и s – соответствующее **исправленное стандартное отклонение**. Данная случайная величина имеет **распределение Стьюдента** с степенями свободы $k = n - 1$.

Алгоритм решения полностью сохраняется.

Раздел 5. Статистическая проверка гипотез

1. H_1: $\mu > \mu_0$

Критическая область – правосторонняя:



$\Rightarrow p(T>t_{пр,кр})=\alpha$

$p(T>t_{пр,кр})=\alpha \quad \Rightarrow \quad p(T<t_{пр,кр})=1\text{-}p(T>t_{пр,кр})=1\text{-}\alpha \quad \Rightarrow$
$p(T<t_{пр,кр})=F(t_{пр,кр})$,

где $F(x;k)$ – функция распределения Стьюдента с $k = n - 1$ степенями свободы

$\Rightarrow F(t_{пр,кр};k)=1\text{-}\alpha$

Раздел 5. Статистическая проверка гипотез

2. H_1: $\mu < \mu_0$

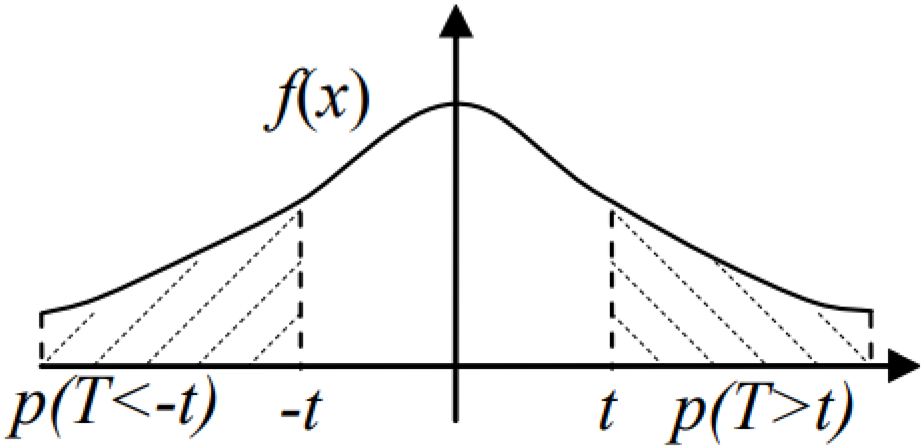
Критическая область – <u>левосторонняя</u>:

$$\Rightarrow p(T<t_{лев,кр})=\alpha$$

$$p(T<t_{лев,кр})=\alpha \Rightarrow p(T<t_{лев,кр})=F(t_{лев,кр};k)=\alpha$$

где $F(x;k)$ – функция распределения Стьюдента с $k = n - 1$ степенями свободы

Раздел 5. Статистическая проверка гипотез



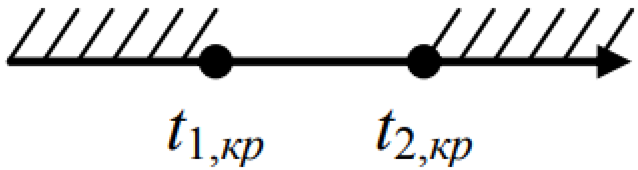
$\Rightarrow p(T>t)=p(T<-t)$

Критическая точка $t_{пр,кр}$ находится из требования: $p(T>t_{пр,кр})=\alpha \Rightarrow p(T<-t_{пр,кр})=\alpha \Rightarrow$ $-t_{пр,кр}$ является критической точкой для левосторонней области: $t_{лев,кр}=-t_{пр,кр}$

Раздел 5. Статистическая проверка гипотез

3. H_1: $\mu \neq \mu_0$

Критическая область – двусторонняя:



Пусть

$<t_{1,кр})=p(T>t_{2,кр})=\alpha/2$

В силу чётности плотности распределения Стьюдента:

$t_{1,кр}=-t_{2,кр}$

$F(t_{2,кр};k)=(1-\alpha)/2$ $\qquad$ $t_{1,кр}=-t_{2,кр}$

или

$F(t_{1,кр};k)=\alpha/2$ $\qquad$ $t_{2,кр}=-t_{1,кр}$

Раздел 5. Статистическая проверка гипотез

Соответственно, критическая область для проверки гипотезы

$$H_0\colon\ \mu = \mu_0$$

против **двусторонней** альтернативы

$$H_1\colon\ \mu \neq \mu_0$$

будет состоять из двух бесконечных полуинтервалов

$$(-\infty;-t_{n\text{-}1;1\text{-}\alpha/2}] \text{ и } [t_{n\text{-}1;1\text{-}\alpha/2};+\infty),$$

против **односторонней** альтернативы

$$H_1\colon\ \mu > \mu_0$$

- из одного полуинтервала $[t_{n\text{-}1;1\text{-}\alpha};+\infty)$ и против **односторонней** альтернативы

$$H_1\colon\ \mu < \mu_0$$

- также из одного полуинтервала $(-\infty;-t_{n\text{-}1;1\text{-}\alpha}]$, где обозначают квантили t-распределения с n-1 степенями свободы соответствующего уровня значимости.

Раздел 5. Статистическая проверка гипотез

$t_{кр}=t_{n\text{-}1;1\text{-}\alpha}$ = **СТЬЮДЕНТ.ОБР**(< 1 - α >; <*n* – 1>) — квантиль уровня 1 - α случайной величины, распределенной по Стьюдента с n-1 степенью свободы.

$t_{кр}=t_{n\text{-}1;1\text{-}\alpha/2}$= **СТЬЮДЕНТ.ОБР**(<1 - α/2>;< *n* – 1>)
— квантиль уровня 1 – α/2 случайной величины *T*, распределенной по Стьюдента с n-1 степенью свободы.

(в силу симметричности ***t*-распределения** справедливы равенства $t_{n\text{-}1;1\text{-}\alpha/2}=-t_{n\text{-}1;\alpha/2}$ и $t_{n\text{-}1;1\text{-}\alpha} =- t_{n\text{-}1;\alpha}$).

Раздел 5. Статистическая проверка гипотез

ПРИМЕР. Производитель электроламп утверждает, что средний срок их службы составляет a_0 =1000 ч. Проверим эту гипотезу на 5%-ном уровне значимости по выборке из n=25 ламп, для которой $\bar{x}_в = 875$ ч, s=50 ч.

Будем проверять на уровне значимости α=5%=0,05 гипотезу H_0 :$\mathbf{E}(X)=a_0$ о том, что математическое ожидание срока службы лампы равно a_0 =1000 ч при альтернативной гипотезе H_1:$\mathbf{E}(X)<a_0$ (состоящей в том, что на самом деле математическое ожидание срока службы лампы меньше, чем заявляет их производитель).

Сравнение параметров двух нормальных распределений

Раздел 5. Статистическая проверка гипотез

Пусть

$$\vec{X} = (X_1, \ldots, X_m) -$$

выборка из $N(\mu_x, \sigma_x^2)$, а

$$\vec{Y} = (Y_1, \ldots, Y_n) -$$

выборка из нормального распределения $N\left(\mu_y, \sigma_y^2\right)$.

Далее считаем выборки **независимыми**, что означает независимость в совокупности $m + n$ случайных величин $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.

Сравнение генеральных средних

Раздел 5. Статистическая проверка гипотез

Способ проверки гипотез о соотношениях между генеральными средними совокупностей (распределений) $N(\mu_x, \sigma_x^2)$ и $N(\mu_y, \sigma_y^2)$ определяется тем, **известны или нет дисперсии** σ_x^2 и σ_y^2, а также от вида альтернативной гипотезы.

Сравнение генеральных средних при известных дисперсиях

Раздел 5. Статистическая проверка гипотез

Предположим, что дисперсии $\boldsymbol{\sigma_x^2}$ **и** $\boldsymbol{\sigma_y^2}$ **известны**, генеральные средние $\boldsymbol{\mu_x}$ **и** $\boldsymbol{\mu_y}$ **неизвестны**.

Рассматривается **основная гипотеза**

$$H_0 : \mu_x = \mu_y.$$

Альтернативная гипотеза чаще всего имеет вид

1) $H_1 : \mu_x > \mu_y$;

2) $H_1 : \mu_x < \mu_y$;

3) $H_1 : \mu_x \neq \mu_y$.

Раздел 5. Статистическая проверка гипотез

При проверке H_0 против H_1 вида 1), 2) или 3) используется одна и та же статистика

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}.$$

$\boldsymbol{\sigma_x^2}$ **и** $\boldsymbol{\sigma_y^2}$ **<u>известны</u>,** $\boldsymbol{\mu_x}$ **и** $\boldsymbol{\mu_y}$ **<u>неизвестны</u>**.

$H_0 : \mu_x = \mu_y$

1) $H_1 : \mu_x > \mu_y$;

2) $H_1 : \mu_x < \mu_y$;

3) $H_1 : \mu_x \neq \mu_y$.

Раздел 5. Статистическая проверка гипотез

Утверждение 1. Если верна H_0, то $Z \sim N(0;1)$.

Доказательство.

$\boldsymbol{\sigma_x^2}$ и $\boldsymbol{\sigma_y^2}$ <u>известны</u>, $\boldsymbol{\mu_x}$ и $\boldsymbol{\mu_y}$ <u>неизвестны</u>.

$H_0 : \mu_x = \mu_y$

1) $H_1 : \mu_x > \mu_y$;

2) $H_1 : \mu_x < \mu_y$;

3) $H_1 : \mu_x \neq \mu_y$.

<u>Статистика</u>

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}$$

Раздел 5. Статистическая проверка гипотез

Пусть Z_α, $Z_{\alpha/2}$ – соответствующие процентные точки стандартного нормального распределения $N(0,1)$. При проверке H_0 против H_1 применяется критерий с критической областью, определяемой по таблице

H_1	K
1) $H_1 : \mu_x > \mu_y$	$Z > Z_\alpha$
2) $H_1 : \mu_x < \mu_y$;	$Z < -Z_\alpha$
3) $H_1 : \mu_x \neq \mu_y$	$\lvert Z \rvert > Z_{\alpha/2}$

$\boldsymbol{\sigma_x^2}$ и $\boldsymbol{\sigma_y^2}$ <u>известны</u>, $\boldsymbol{\mu_x}$ и $\boldsymbol{\mu_y}$ <u>неизвестны</u>.

$H_0 : \mu_x = \mu_y$

1) $H_1 : \mu_x > \mu_y$;

2) $H_1 : \mu_x < \mu_y$;

3) $H_1 : \mu_x \neq \mu_y$.

<u>Статистика</u>

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}$$

Раздел 5. Статистическая проверка гипотез

Утверждение 2. Критерий с критической областью 1) – 3) имеет уровень значимости α.

Доказательство.

Утверждение 1. Если верна H_0, то $Z \sim N(0; 1)$.

$\boldsymbol{\sigma_x^2}$ **и** $\boldsymbol{\sigma_y^2}$ **известны**,
$\boldsymbol{\mu_x}$ **и** $\boldsymbol{\mu_y}$ **неизвестны**.

$H_0 : \mu_x = \mu_y$

1) $H_1 : \mu_x > \mu_y$;

2) $H_1 : \mu_x < \mu_y$;

3) $H_1 : \mu_x \neq \mu_y$.

Статистика

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}$$

H_1	K
1) $H_1 : \mu_x > \mu_y$	$Z > Z_\alpha$
2) $H_1 : \mu_x < \mu_y$;	$Z < -Z_\alpha$
3) $H_1 : \mu_x \neq \mu_y$	$\lvert Z \rvert > Z_{\alpha/2}$

Сравнение генеральных средних при неизвестных и равных дисперсиях

Раздел 5. Статистическая проверка гипотез

Теперь предположим, что обе генеральные дисперсии $\boldsymbol{\sigma_x^2}$ и $\boldsymbol{\sigma_y^2}$ **неизвестны**, но **одинаковы** $\boldsymbol{\sigma_x^2 = \sigma_y^2 = \sigma^2}$ (генеральные средние $\boldsymbol{\mu_x}$ **и** $\boldsymbol{\mu_y}$ **неизвестны**).

С учётом этого равенства

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Поскольку значение σ неизвестно, попробуем заменить σ на корень из несмещенной оценки дисперсии.
В данном случае имеется, по крайней мере, две такие оценки:

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2\,;\; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

Раздел 5. Статистическая проверка гипотез

Оказывается, что вместо того, чтобы использовать одну из них, лучше воспользоваться линейной комбинацией

$$s^2 = \frac{m-1}{m+n-2} s_x^2 + \frac{n-1}{m+n-2} s_y^2.$$

Очевидно проверяется, что s^2 – **несмещенная оценка** σ^2.

$$s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2 \,;\; s_y^2 = \frac{1}{n-1} \sum_{i=1}^{m} (Y_i - \bar{Y})^2$$

Раздел 5. Статистическая проверка гипотез

Кроме того, можно доказать, что

$$\frac{(m-1)s_x^2}{\sigma^2} = \frac{(X_1-\bar{X})^2+\cdots+(X_m-\bar{X})^2}{\sigma^2} \sim \chi^2(m-1)$$

и

$$\frac{(n-1)s_y^2}{\sigma^2} = \frac{(Y_1-\bar{Y})^2+\cdots+(Y_n-\bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

Отсюда с учетом независимости следует, что

$$\frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2} \sim \chi^2(m+n-2).$$

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2\,;\; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{m}(Y_i-\bar{Y})^2$$

Таким образом, коэффициенты при s_x^2 и s_y^2 в

$$s^2 = \frac{m-1}{m+n-2} s_x^2 + \frac{n-1}{m+n-2} s_y^2$$

Можно интерпретировать как доли числа степеней свободы “иксов” и “игреков” в числе степеней свободы распределения случайной величины

$$\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} \sim \chi^2(m+n-2),$$

которая отличается от оценки s^2 лишь постоянным множителем $\frac{m+n-2}{\sigma^2}$.

$$s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2 \,;\; s_y^2 = \frac{1}{n-1} \sum_{i=1}^{m} (Y_i - \bar{Y})^2$$

Раздел 5. Статистическая проверка гипотез

Заменив σ на s в записи статистики

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

получим новую статистику

$$T = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2\,;\; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$$

Примем без доказательства следующую теорему.

Теорема 1. Если верна H_0, то $T \sim t(m+n-2)$, где $t(m+n-2)$ – **распределение Стьюдента** с $m+n-2$ степенями свободы**.**

Раздел 5. Статистическая проверка гипотез

В силу данной теоремы статистику T можно использовать для построения критерия по проверке H_0.

Как и раньше рассматриваются **три основных случая альтернативной гипотезы $\boldsymbol{H_1}$**, в соответствии с которыми выбирается критическая область K того или иного вида:

H_1	K
1) $H_1 : \mu_x > \mu_y$	$T > t_\alpha(m+n-2)$
2) $H_1 : \mu_x < \mu_y$;	$T < -t_\alpha(m+n-2)$
3) $H_1 : \mu_x \neq \mu_y$	$\lvert T \rvert > t_{\alpha/2}(m+n-2)$

Раздел 5. Статистическая проверка гипотез

Утверждение 3. Критерий с критической областью 1) – 3) имеет уровень значимости α.

$\boldsymbol{\sigma_x^2}$ **и** $\boldsymbol{\sigma_y^2}$ **неизвестны**, но
$\boldsymbol{\sigma_x^2 = \sigma_y^2 = \sigma^2}$
$\boldsymbol{\mu_x}$ **и** $\boldsymbol{\mu_y}$ **неизвестны**.
$H_0 : \mu_x = \mu_y$
1) $H_1 : \mu_x > \mu_y$;
2) $H_1 : \mu_x < \mu_y$;
3) $H_1 : \mu_x \neq \mu_y$.

Статистика

$$T = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

H_1	K
1) $H_1 : \mu_x > \mu_y$	$T > t_\alpha(m+n-2)$
2) $H_1 : \mu_x < \mu_y$;	$T < -t_\alpha(m+n-2)$
3) $H_1 : \mu_x \neq \mu_y$	$\lvert T \rvert > t_{\alpha/2}(m+n-2)$

Сравнение генеральных средних с неизвестными и неравными дисперсиями

Раздел 5. Статистическая проверка гипотез

Задача сравнения средних двух нормально распределенных совокупностей **при неизвестных и неравных дисперсиях** известна как **проблема Беренса-Фишера**.

Точного решения этой задачи до настоящего времени нет.

Одно из приближений даёт **критерий Кохрана–Кокса**.

Критерий Кохрана–Кокса.

Статистика критерия:

$$T = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}};$$

Критическое значение статистики:

$$t_\alpha = \frac{\frac{s_x^2}{m}t_{\alpha;m-1}+\frac{s_y^2}{n}t_{\alpha;n-1}}{\frac{s_x^2}{m}+\frac{s_y^2}{n}}.$$

Если выполняется неравенство $|T_{\text{набл.}}| > t_\alpha$, гипотеза H_0 отклоняется.

Структура критерия по проверке гипотезы о равенстве дисперсий двух нормальных распределений зависит от того, известно или нет генеральное среднее, а также от вида альтернативной гипотезы.

Раздел 5. Статистическая проверка гипотез

Гипотеза о равенстве генеральных дисперсий двух нормальных распределений (двухвыборочный *F*-критерий)

Две средние мы **уже сравнивали**, очередь за дисперсиями.

Из двух нормальных ген. совокупностей извлечены независимые выборки объёмом n и m и найдены их исправленные дисперсии: s^2_x и s^2_y соответственно.

Совершенно понятно, что эти значения случайны и отличны друг от друга.

Раздел 5. Статистическая проверка гипотез

Но возникает **вопрос**: значимо или незначимо это отличие?

Для ответа на этот вопрос на уровне значимости α проверяется гипотеза о равенстве генеральных дисперсий $H_0 : \sigma_x^2 = \sigma_y^2$.

Если она будет принята, то различие между выборочными значениями $s^2{}_x$ и $s^2{}_y$ объяснимо случайными факторами.

Рассматриваемая гипотеза часто возникает, когда требуется сравнить точность двух приборов, инструментов, станков, двух методов исследования.

Раздел 5. Статистическая проверка гипотез

Для проверки этой гипотезы используют критерий

$$F = \frac{S_Б^2}{S_М^2},$$

где $S_Б^2$ – бОльшая исправленная дисперсия, а $S_М^2$ – мЕньшая.

Данная случайная величина имеет **распределение Фишера-Снедекора** (так называемое ***F-распределение***) со степенями свободы

$k_1=n-1$, $k_2=m-1$, если $s^2_x > s^2_y$ **или**
$k_1=m-1$, $k_2=n-1$, если $s^2_y > s^2_x$.

$(s_x^2 = \frac{n}{n-1}\widehat{D}(X),\ s_y^2 = \frac{m}{m-1}\widehat{D}(Y))$

Раздел 5. Статистическая проверка гипотез

То есть, степень свободы соответствует выборке с бОльшей исправленной дисперсией.

Раздел 5. Статистическая проверка гипотез

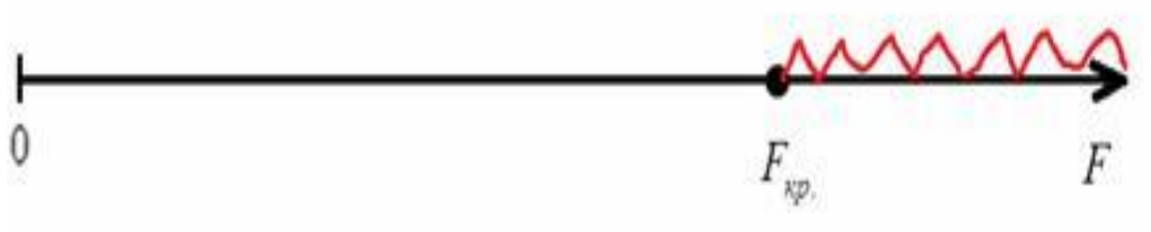
В качестве альтернативы рассматривают одну из следующих гипотез:

Раздел 5. Статистическая проверка гипотез

1) $H_1: \sigma_x^2 > \sigma_y^2$ (если $s^2{}_x > s^2{}_y$)

либо $H_1: \sigma_y^2 > \sigma_x^2$ (если $s^2{}_y > s^2{}_x$).

Для этой гипотезы строят <u>правостороннюю</u> критическую область:

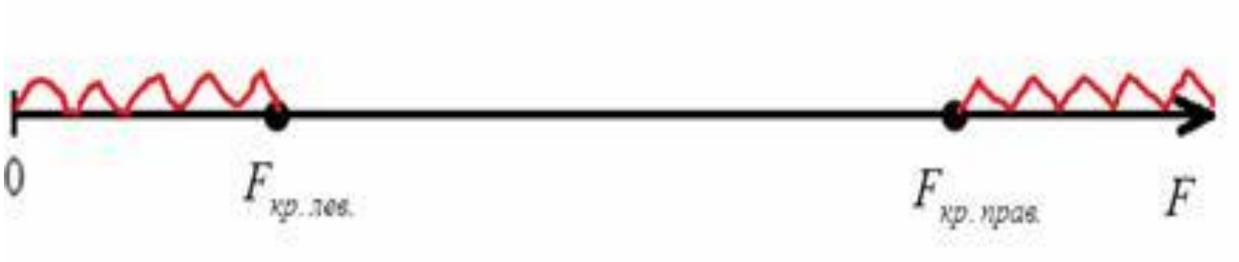


Критическое значение можно найти с помощью стандартной функции Экселя

$F_{кр}$=**F.ОБР**($<1-\alpha>$;$<k_1>$;$<k_2>$)

Раздел 5. Статистическая проверка гипотез

2) $H_1: \sigma_x^2 \neq \sigma_y^2$ – для этой гипотезы строится двусторонняя критическая область:



Однако для решения нашей задачи достаточно найти лишь правое критическое значение

$F_{кр.прав.}$=**F.ОБР**(<1-α/2>;<k_1>;<k_2>)

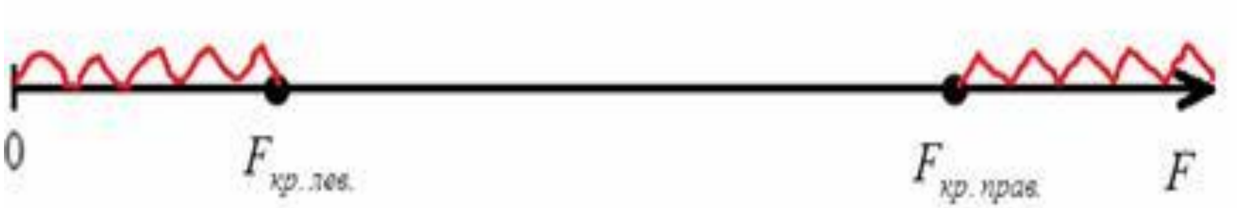
Раздел 5. Статистическая проверка гипотез

Дело в том, что

$F_{\text{кр.лев.}}=\textbf{F.ОБР}(<\alpha>;<k_1>;<k_2>)<1$

и поэтому случайное значение $F=\dfrac{S_Б^2}{S_М^2}$,

(бОльшее единицы) заведомо не может попасть в левый кусок критической области.



Раздел 5. Статистическая проверка гипотез

Далее на основании выборочных данных рассчитывается наблюдаемое значение критерия

$$F_{набл} = \frac{s_Б^2}{s_М^2},$$

и если оно попадает в критическую область ($F_{набл} > F_{кр}$ для обоих случаев), то гипотеза $H_0 : \sigma_x^2 = \sigma_y^2$ отвергается.

Если $F_{набл} < F_{кр}$, то принимается.

Пример.

Некоторая физическая величина измерена n=7 и m=5 раз двумя различными способами.

По результатам измерений найдены соответствующие погрешности s^2_x=6,3, s^2_y=10,1.

Требуется на уровне значимости 0,05 проверить, одинаковую ли точность обеспечивают эти способы измерений.

Раздел 5. Статистическая проверка гипотез

Критерии случайности, независимости, однородности

Раздел 5. Статистическая проверка гипотез

Для описания и изучения свойств генеральной совокупности по выборке из нее необходимо, чтобы результаты измерений, формирующие выборку, удовлетворяли ряду требований:

– **случайности**;

– **независимости**;

– **однородности**;

– **отсутствия аномальных наблюдений**.

Критерии отсутствия аномальных наблюдений

Раздел 5. Статистическая проверка гипотез

Под **аномальными значениями** выборки понимаются те значения, которые не отвечают потенциальным возможностям исследуемой системы и которые, оставаясь в выборке, оказывают существенное влияние на статистические характеристики.

Критерий Граббса

Раздел 5. Статистическая проверка гипотез

Наиболее часто для проверки наблюдений на выброс применяют **простые критерии Граббса**. Критерии используются для проверки на аномальность наблюдений, принадлежащих выборкам **из нормальной генеральной совокупности**.

Статистики критерия Граббса предусматривают возможность проверки на наличие в выборке либо одного аномального наблюдения (наименьшего или наибольшего), либо двух (двух наименьших в выборке или двух наибольших).

Рассмотрим **критерий Граббса проверки выборки на одно аномальное значение**.

Раздел 5. Статистическая проверка гипотез

Проверяемая **гипотеза** $\boldsymbol{H_0}$ заключается в том, что все элементы выборки $X_1, \ldots, X_n$ принадлежат одной генеральной совокупности.

Раздел 5. Статистическая проверка гипотез

1. При проверке на выброс наибольшего значения выборочного значения $X_{(n)}$, конкурирующая гипотеза H_1 заключается в том, что первые $n-1$ значений вариационного ряда $X_1, \ldots, X_{n-1}$ принадлежат одному закону, а крайний элемент $X_{(n)}$ — некоторому другому, существенно сдвинутому вправо.

Статистика критерия Граббса имеет вид

$$G_n = \frac{X_{(n)} - \bar{X}}{S},$$

где

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,\ \ S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, S = \sqrt{S^2}.$$

Раздел 5. Статистическая проверка гипотез

Критическое значение $G_{n,\alpha}$ вычисляется по формуле

$$G_{n,\alpha} = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{\alpha}{n};n-2}}{n-2+t^2_{\frac{\alpha}{n};n-2}}},$$

где $t^2_{\frac{\alpha}{n};n-2}$ процентная точка **распределения Стьюдента** порядка α/n с $n-2$ степенями свободы, α — уровень значимости.

Максимальное значение $X_{(n)}$ считается **выбросом**, если наблюдаемое значение статистики превысит критическое:

$$G_{n,\text{набл}} > G_{n,\alpha}.$$

2. При проверке на выброс наименьшего значения выборочного значения $X_{(1)}$, конкурирующая гипотеза H_1 заключается в том, что значения вариационного ряда $X_2, \ldots, X_n$ принадлежат одному закону, а крайний элемент $X_{(1)}$ — некоторому другому, существенно сдвинутому влево.

Статистика критерия Граббса имеет вид

$$G_n = \frac{\bar{X} - X_{(1)}}{S},$$

где

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,\; S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, S = \sqrt{S^2}.$$

Раздел 5. Статистическая проверка гипотез

Критическое значение $G_{n,\alpha}$ вычисляется по формуле

$$G_{1,\alpha} = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\frac{\alpha}{n};n-2}}{n-2+t^2_{\frac{\alpha}{n};n-2}}},$$

где $t^2_{\frac{\alpha}{n};n-2}$ процентная точка **распределения Стьюдента** порядка α/n с $n-2$ степенями свободы, α — уровень значимости.

Минимальное значение $X_{(1)}$ считается **выбросом**, если наблюдаемое значение статистики превысит критическое:

$$G_{1,\text{набл}} > G_{1,\alpha}.$$

Если для какого-либо крайнего значения выборки принимается гипотеза H_1, **проверку на аномальность** следует повторить для следующего элемента вариационного ряда.

При этом, значения $\bar{x}$ и s следует пересчитать с учетом удаленного крайнего значения.

Критерий случайности

Критерий инверсий

Раздел 5. Статистическая проверка гипотез

Рассматривается исходная неупорядоченная выборка.

Говорят, что пара x_i, x_j в наборе $x_1, \ldots, x_n$ образует инверсию, если $i < j$, но при этом $x_i > x_j$.

H_0: числа $x_1, \ldots, x_n$ образуют случайный набор данных;

H_1: значения $x_1, \ldots, x_n$ не образуют случайный набор данных.

В качестве **статистики** берется следующая «мера беспорядка»:

R_n — общее количество инверсий в наборе $x_1, \ldots, x_n$.

Теорема.

$$E(R_n) = \frac{n(n-1)}{4};$$

$$D(R_n) = \frac{n(n-1)(2n+5)}{72};$$

$$\frac{R_n - E(R_n)}{\sqrt{D(R_n)}} \longrightarrow N(0;1)$$

при $n \longrightarrow \infty$.

Статистика критерия имеет вид:

$$Z = \frac{R_n - E(R_n)}{\sqrt{D(R_n)}}.$$

Критическим значением критерия является процентная точка нормального распределения $z_{\alpha/2}$ (α — уровень значимости).

Если вычисленное по выборке наблюдаемое значение статистики $Z_{\text{набл}}$ удовлетворяет неравенству

$$|Z_{\text{набл}}| > z_{\alpha/2},$$

то гипотеза H_0 о том, что набор $x_1, \dots, x_n$ образует случайную выборку, отклоняется на уровне значимости α.

Данный метод применим, если объем выборки $n \geq 70$

(в этом случае статистика Z достаточно близка к нормальному закону распределения).

Критерии независимости

Критерий независимости χ^2

Раздел 5. Статистическая проверка гипотез

Имеются две дискретные случайные величины X и Y.

Требуется проверить **гипотезу об их независимости**.

Пусть различаются r значений случайной величины X (обозначим их $x_1, \dots, x_r$) и s значений случайной величины Y (обозначим их $y_1, \dots, y_s$).

Через k_{ij} обозначим общее количество таких элементов выборки, в которых X принимает значение x_i, а Y — значение y_j.

Тогда

$$\sum_{i=1}^{r} \sum_{j=1}^{s} k_{ij} = n,$$

где n — объем выборки. Введем также обозначения:

$$\nu_i = \sum_{j=1}^{s} k_{ij}\,;\ \mu_j = \sum_{i=1}^{r} k_{ij}\,.$$

Раздел 5. Статистическая проверка гипотез

В рассматриваемом случае результаты наблюдений удобно оформлять в виде таблицы, называемой **таблицей сопряженности признаков**:

X/Y	y_1	y_2	…	y_s	
x_1	k_{11}	k_{12}	…	k_{1s}	ν_1
x_2	k_{21}	k_{22}	…	k_{2s}	ν_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	k_{r1}	k_{r2}	…	k_{rs}	ν_r
	μ_1	μ_2	…	μ_s	n

Раздел 5. Статистическая проверка гипотез

Пусть далее

$$p_{ij} = P(\{X = x_i, Y = y_j\});$$
$$p_i = P(\{X = x_i\}); q_j = P(\{Y = y_j\}).$$

$H_0: p_{ij} = p_i q_j, i = 1, \ldots, r, j = 1, \ldots, s.$

$H_1: p_{ij} \neq p_i q_j$ для некоторых $i = 1, \ldots, r, j = 1, \ldots, s.$

Рассматривается следующая **статистика**:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{k_{ij}^2}{\nu_i \mu_j} - 1\right).$$

Гипотеза H_0 **отклоняется**, если вычисленное по выборочным данным значение статистики $\chi^2_{\text{набл}}$ удовлетворяет неравенству:

$$\chi^2_{\text{набл}} > \chi^2_{\alpha;(r-1)(s-1)}.$$

Действительно, при условии справедливости H_0, статистика χ^2 имеет распределение χ^2 с $(r-1)(s-1)$ степенями свободы.

Тогда вероятность ошибки первого рода

$$P\left(\left\{\chi^2 > \chi^2_{\alpha;(r-1)(s-1)}\right\}\right) = \alpha.$$

Раздел 5. Статистическая проверка гипотез

Пример. Отношение зрителей к включению одной из телепередач в программу выразилось следующими данными:

	Положительное	**Безразличное**	**Отрицательное**
Мужчины	14	24	2
Женщины	29	36	15

Можно ли считать, что отношение к включению передачи в программу не зависит от пола зрителя на уровне значимости $\alpha = 0{,}1$?

Решение.

Пусть $X_i =$ {пол зрителя}; $i = 1, 2$;

$i = 1$: мужчины, $i = 2$: женщины.

$Y_j =$ {отношение к телепередаче}; $j = 1, 2, 3$;

$j = 1$: положительное,

$j = 2$: безразличное,

$j = 3$: отрицательное.

Решение (продолжение).

$$k_{11} = 14, k_{12} = 24, k_{13} = 2, k_{21} = 29, k_{22} = 36, k_{23} = 15$$

$$\nu_1 = \sum_{j=1}^{3} k_{1j} = 40; \nu_2 = \sum_{j=1}^{3} k_{2j} = 80;$$

$$\mu_1 = \sum_{i=1}^{2} k_{i1} = 43; \mu_2 = \sum_{i=1}^{2} k_{i2} = 60; \mu_3 = \sum_{i=1}^{2} k_{i3} = 17.$$

	Положительное	**Безразличное**	**Отрицательное**	**Σ**
Мужчины	14	24	2	**40**
Женщины	29	36	15	**80**
Σ	**43**	**60**	**17**	**120**

Раздел 5. Статистическая проверка гипотез

Решение (продолжение).

Находим **наблюдаемое значение статистики** ($r = 2; s = 3$):

$$\chi^2_{\text{набл}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{k_{ij}^2}{\nu_i\mu_j} - 1\right) =$$

$$= 120\left(\frac{14^2}{40\cdot 43} + \frac{24^2}{40\cdot 60} + \frac{2^2}{40\cdot 17} + \frac{29^2}{80\cdot 43} + \frac{36^2}{80\cdot 60} + \frac{15^2}{80\cdot 17} - 1\right) =$$

$$\approx 120 \cdot 0{,}03975 \;\approx\; 4{,}77.$$

	Положительное	**Безразличное**	**Отрицательное**	**Σ**
Мужчины	14	24	2	ν_1 =**40**
Женщины	29	36	15	ν_2 =**80**
Σ	μ_1 = **43**	μ_2 = **60**	μ_3 = **17**	**120**

Решение (продолжение).

Критическое значение статистики:

$$\chi^2_{\alpha;(r-1)(s-1)} = \chi^2_{0,1;2} = 4{,}61.$$

$\chi^2_{\text{кр}}$ = **ХИ2.ОБР**($1 - \alpha;\ (r-1)(s-1)$) **# MS Excel** $\boldsymbol{\chi^2_{\text{кр}} = 4,60517}$

 # Python

```
import scipy.stats as sts
```

```
alpha= 0.1
r=2
s=3
```

```
chi_cr = sts.chi2.isf(alpha,(r-1)*(s-1))
print('Критическое значение: chi_cr=',chi_cr)
```

```
Критическое значение: chi_cr= 4.605170185988092
```

Раздел 5. Статистическая проверка гипотез

Решение (продолжение).

Так как

$$\chi^2_{\text{набл}} = 4{,}77 > \chi^2_{0,1;2} = 4{,}61,$$

то гипотеза H_0 **отклоняется**, т. е. отношение к телепередаче зависит от пола.

	Положительное	**Безразличное**	**Отрицательное**	$\boldsymbol{\Sigma}$
Мужчины	14	24	2	ν_1 =**40**
Женщины	29	36	15	ν_2 =**80**
$\boldsymbol{\Sigma}$	$\mu_1 =$ **43**	$\mu_2 =$ **60**	$\mu_3 =$ **17**	**120**

Критерий однородности

Критерий однородности χ^2

Раздел 5. Статистическая проверка гипотез

Проверяется **гипотеза** о том, что **две выборки принадлежат одной генеральной совокупности**.

Данные должны быть представлены в виде интервального статистического ряда.

Имеются

выборка объема n_1 из генеральной совокупности X_1 и выборка объема n_2 из генеральной совокупности X_2;
l — количество интервалов группировки (одинаковое для обеих выборок);
μ_i и ν_i — количество попаданий в i-й интервал группирования, соответственно, первой и второй выборок, $i = 1, 2, \ldots, l$;
уровень значимости α.

Раздел 5. Статистическая проверка гипотез

Пусть $F_j(x)$ — функция распределения случайной величины X_j, $j = 1, 2$.

Проверяется **гипотеза**

$H_0\colon F_1(x) = F_2(x), x \in \mathbb{R}$,

$H_1\colon F_1(x) \neq F_2(x)$, для некоторых $x \in \mathbb{R}$.

Статистика критерия имеет следующий вид:

$$\chi^2 = n_1\, n_2 \sum_{i=1}^{l} \frac{\left(\frac{\mu_i}{n_1} - \frac{\nu_i}{n_2}\right)^2}{\mu_i + \nu_i}.$$

Раздел 5. Статистическая проверка гипотез

В случае совпадения объемов выборок: $n_1 = n_2 = n$ **статистика** вычисляется по формуле

$$\chi^2 = \sum_{i=1}^{l} \frac{(\mu_i - \nu_i)^2}{\mu_i + \nu_i}.$$

Критическое значение статистики: $\chi^2_{\alpha;l-1}$.

Раздел 5. Статистическая проверка гипотез

Гипотеза $\boldsymbol{H_0}$ **отклоняется**, если вычисленное по выборочным данным значение статистики $\chi^2_{\text{набл}}$удовлетворяет неравенству:

$$\chi^2_{\text{набл}} > \chi^2_{\alpha;l-1}.$$

Действительно, при условии справедливости гипотезы H_0 **распределение статистики** χ^2 при n_1 и n_2, стремящимся к бесконечности, **сходится** к χ^2 -распределению с $l-1$ степенями свободы.

Поэтому при достаточно больших n_1, n_2 выполняется соотношение:

$$P(\{\chi^2 > \chi^2_{\alpha;l-1}\}) \approx \alpha.$$

Раздел 5. Статистическая проверка гипотез

Пример. В банке в течение двух дней проводилось исследование времени обслуживания клиентов. Данные представлены в таблице.

Номер интервала группирования i	**Время обслуживания (мин)**	**(1-й день)**	**(2-й день)**
1	10-12	2	2
2	12-14	4	4
3	14-16	8	9
4	16-18	12	13
5	18-20	16	16
6	20-22	10	8
7	22-24	3	3

Требуется проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что обе выборки принадлежат одной генеральной совокупности.

Раздел 5. Статистическая проверка гипотез

Решение. Имеем: $n_1 = n_2 = 55$; $l = 7$. Находим наблюдаемое значение статистики критерия:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{l} \frac{(\mu_i - \nu_i)^2}{\mu_i + \nu_i} = \frac{(2-2)^2}{4} + \frac{(4-4)^2}{8} + \frac{(8-9)^2}{17} + \frac{(12-13)^2}{25} + \frac{(16-16)^2}{32} + \frac{(10-8)^2}{18} + \frac{(3-3)^2}{6} \approx 0{,}32.$$

Номер интервала группирования i	Время обслуживания (мин)	(1-й день)	(2-й день)
1	10-12	2	2
2	12-14	4	4
3	14-16	8	9
4	16-18	12	13
5	18-20	16	16
6	20-22	10	8
7	22-24	3	3

Решение (продолжение).

Критическое значение статистики: $\chi^2_{\alpha;l-1} = \chi^2_{0,05;6} = 12{,}59.$

Python

```
import scipy.stats as sts
alpha= 0.05
l=7
```

```
chi_cr = sts.chi2.isf(alpha,l-1)
print('Критическое значение: chi_cr=',chi_cr)
```

```
Критическое значение: chi_cr= 12.59158724374398
```

MS Excel

=ХИ2.ОБР(1-0,05;6)

Номер интервала группирования i	Время обслуживания (мин)	(1-й день)	(2-й день)
1	10-12	2	2
2	12-14	4	4
3	14-16	8	9
4	16-18	12	13
5	18-20	16	16
6	20-22	10	8
7	22-24	3	3

Решение (продолжение).

Так как

$$\chi^2_{\text{набл}} = 0{,}32 < \chi^2_{0{,}05;6} = 12{,}59,$$

то оснований для отклонения гипотезы H_0 нет.

Номер интервала группирования i	**Время обслуживания (мин)**	**(1-й день)**	**(2-й день)**
1	10-12	2	2
2	12-14	4	4
3	14-16	8	9
4	16-18	12	13
5	18-20	16	16
6	20-22	10	8
7	22-24	3	3

Критерий однородности Колмогорова-Смирнова

Раздел 5. Статистическая проверка гипотез

Имеются две выборки – объема n_1 из генеральной совокупности X_1 и объема n_2 из генеральной совокупности X_2.

Предполагается, что случайные величины X_j — непрерывные с функциями распределения

$$F_j(x),\ j = 1, 2.$$

$H_0: F_1(x) = F_2(x), x \in \mathbb{R},$

$H_1: F_1(x) \neq F_2(x)$, для некоторых $x \in \mathbb{R}$.

Раздел 5. Статистическая проверка гипотез

Проверка гипотезы производится **по следующей схеме**:

1. По имеющимся выборкам находятся **эмпирические функции распределения** $F_1^*(x)$ и $F_2^*(x)$.

2. Рассматривается **статистика** следующего вида:

$$D = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max_x |F_1^*(x) - F_2^*(x)|\,.$$

3. По таблицам распределения Колмогорова определяется величина **k_α — 100α-процентная точка распределения Колмогорова уровня α**.

4. **Гипотеза H_0 отклоняется** на уровне значимости α, если вычисленное по выборочным данным значение статистики $D_{\text{набл}}$ удовлетворяет неравенству:

$$D_{\text{набл}} > k_\alpha.$$

Действительно, согласно доказанной Н.В.Смирновым теореме, при $x > 0$

$$\lim_{n\to\infty} P\{D < x\} = K(x),$$

где $K(x)$ — **функция распределения Колмогорова**.

При условии истинности гипотезы H_0 при достаточно больших n_1 и n_2 выполняется соотношение:

$$P(D > k_\alpha) \approx \alpha.$$

Замечание.

Критерий Колмогорова–Смирнова применяется при

$$n_1, n_2 \geq 50.$$

Критические точки для статистики Колмогорова D_n

Объем выборки n	Уровень значимости α			
	0,10	0,05	0,02	0,01
1	0,95	0,98	0,99	0,995
2	0,78	0,84	0,90	0,93
3	0,64	0,71	0,78	0,83
4	0,57	0,62	0,69	0,73
5	0,51	0,56	0,62	0,67
6	0,47	0,52	0,58	0,62
7	0,44	0,48	0,54	0,58
8	0,41	0,45	0,51	0,54
9	0,39	0,43	0,48	0,51
10	0,37	0,41	0,46	0,49
11	0,35	0,39	0,44	0,47
12	0,34	0,38	0,42	0,45
13	0,33	0,36	0,40	0,43
14	0,31	0,35	0,39	0,42
15	0,30	0,34	0,38	0,40
16	0,29	0,33	0,37	0,39
17	0,29	0,32	0,36	0,38
18	0,28	0,31	0,34	0,37
19	0,27	0,30	0,34	0,36
20	0,26	0,29	0,33	0,35

Критические точки распределения Колмогорова

$$Q(\lambda) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2\lambda^2}$$

α	0,10	0,05	0,02	0,01
$\lambda_{кр}$	1,23	1,36	1,52	1,63

Раздел 5. Статистическая проверка гипотез

Пример. Даны две выборки:

X_1 объема $n_1 = 30$:

1; 4; 9; 2; 5; 6; 5; 8; 4; 8; 7; 5; 6; 5; 7; 9; 9; 9; 4; 8; 2; 4; 8; 9; 6; 2; 6; 9; 8; 6

X_2 объема $n_2 = 20$:

8; 4; 5; 6; 6; 3; 2; 9; 6; 1; 2; 3; 6; 6; 2; 1; 14; 1; 1; 4.

Требуется проверить гипотезу $H_0 : F_1(x) = F_2(x)$ против альтернативной гипотезы $H_1 : F_1(x) \neq F_2(x)$ на уровне значимости $\alpha = 0{,}05$.

Решение.

Строим **вариационные** ряды:

X_1: 1; 2; 2; 2; 4; 4; 4; 4; 5; 5; 5; 5; 6; 6; 6; 6; 6; 7; 7; 8; 8; 8; 8; 8; 9; 9; 9; 9; 9; 9.

X_2: 1; 1; 1; 1; 2; 2; 2; 3; 3; 4; 4; 5; 6; 6; 6; 6; 6; 8; 9; 14.

Решение (продолжение).

Строим **статистические** ряды:

$\boldsymbol{X_1}$:

x_i	1	2	4	5	6	7	8	9
n_i	1	3	4	4	5	2	5	6

$\boldsymbol{X_2}$:

x_i	1	2	3	4	5	6	8	9	14
n_i	4	3	2	2	1	5	1	1	1

$\boldsymbol{X_1}$: 1; 2; 2; 2; 4; 4; 4; 4; 5; 5; 5; 5; 6; 6; 6; 6; 6; 7; 7; 8; 8; 8; 8; 8; 9; 9; 9; 9; 9; 9.
$\boldsymbol{X_2}$: 1; 1; 1; 1; 2; 2; 2; 3; 3; 4; 4; 5; 6; 6; 6; 6; 6; 8; 9; 14.

Решение (продолжение).

Строим **эмпирические функции распределений**:

X_1:

x_i	1	2	4	5	6	7	8	9
n_i	1	3	4	4	5	2	5	6
$\sum n_i$	1	4	8	12	17	19	24	30
$F_1^*(x_i)$	$\frac{1}{30}$	$\frac{4}{30}$	$\frac{8}{30}$	$\frac{12}{30}$	$\frac{17}{30}$	$\frac{19}{30}$	$\frac{24}{30}$	1

Δ_i	$(-\infty; 1]$	$(1; 2]$	$(2; 4]$	$(4; 5]$	$(5; 6]$	$(6; 7]$	$(7; 8]$	$(8; 9]$	$(9; \infty)$
$F_1^*(x_i)$	0	$\frac{1}{30}$	$\frac{4}{30}$	$\frac{8}{30}$	$\frac{12}{30}$	$\frac{17}{30}$	$\frac{19}{30}$	$\frac{24}{30}$	1

Решение (продолжение).

Строим **эмпирические функции распределений**:

$\boldsymbol{X_2}$:

x_i	1	2	3	4	5	6	8	9	14
n_i	4	3	2	2	1	5	1	1	1
$\sum n_i$	4	7	9	11	12	17	18	19	20
$F_2^*(x_i)$	$\frac{4}{20}$	$\frac{7}{20}$	$\frac{9}{20}$	$\frac{11}{20}$	$\frac{12}{20}$	$\frac{17}{20}$	$\frac{18}{20}$	$\frac{19}{20}$	1

Δ_i	$(-\infty; 1]$	$(1; 2]$	$(2; 3]$	$(3; 4]$	$(4; 5]$	$(5; 6]$	$(6; 8]$	$(8; 9]$	$(9; 14]$	$(14; \infty)$
$F_2^*(x_i)$	0	$\frac{4}{20}$	$\frac{7}{20}$	$\frac{9}{20}$	$\frac{11}{20}$	$\frac{12}{20}$	$\frac{17}{20}$	$\frac{18}{20}$	$\frac{19}{20}$	1

Решение (продолжение). Находим разности значений эмпирических функций распределения в каждом интервале группировки:

$(-\infty; 1]: |0-0| = 0;$

$(1; 2]: \left|\frac{4}{20} - \frac{1}{30}\right| = \frac{10}{60};$

$(2; 3]: \left|\frac{7}{20} - \frac{4}{30}\right| = \frac{13}{60};$

$(3; 4]: \left|\frac{9}{20} - \frac{4}{30}\right| = \frac{19}{60};$

$(4; 5]: \left|\frac{11}{20} - \frac{8}{30}\right| = \frac{17}{60};$

$(5; 6]: \left|\frac{12}{20} - \frac{12}{30}\right| = \frac{12}{60};$

$(6; 7]: \left|\frac{17}{20} - \frac{17}{30}\right| = \frac{17}{60};$

$(7; 8]: \left|\frac{17}{20} - \frac{19}{30}\right| = \frac{13}{60};$

$(9; 14]: \left|\frac{19}{20} - \frac{30}{30}\right| = \frac{3}{60};$

$(14; \infty): |1-1| = 0$

Видим, что $\max|F_1^*(x) - F_2^*(x)| = \frac{19}{60}$

Δ_i	$(-\infty; 1]$	$(1; 2]$	$(2; 4]$	$(4; 5]$	$(5; 6]$	$(6; 7]$	$(7; 8]$	$(8; 9]$	$(9; \infty)$
$F_1^*(x_i)$	0	$\frac{1}{30}$	$\frac{4}{30}$	$\frac{8}{30}$	$\frac{12}{30}$	$\frac{17}{30}$	$\frac{19}{30}$	$\frac{24}{30}$	1

Δ_i	$(-\infty; 1]$	$(1; 2]$	$(2; 3]$	$(3; 4]$	$(4; 5]$	$(5; 6]$	$(6; 8]$	$(8; 9]$	$(9; 14]$	$(14; \infty)$
$F_2^*(x_i)$	0	$\frac{4}{20}$	$\frac{7}{20}$	$\frac{9}{20}$	$\frac{11}{20}$	$\frac{12}{20}$	$\frac{17}{20}$	$\frac{18}{20}$	$\frac{19}{20}$	1

Решение (продолжение). Значит

$$D_{\text{набл}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max_x |F_1^*(x) - F_2^*(x)| = \sqrt{\frac{30 \cdot 20}{30+20}} \cdot \frac{19}{60} \approx 1{,}10.$$

Из таблиц распределения Колмогорова для $\alpha = 0{,}05$ находим $k_\alpha = 1{,}36$.

α	0,10	0,05	0,02	0,01
$\lambda_{кр}$	1,23	1,36	1,52	1,63

Так как $D_{\text{набл}} < k_\alpha$, то гипотеза H_0 принимается.

$$\max|F_1^*(x) - F_2^*(x)| = \frac{19}{60}$$

Проверка гипотез о законе распределения генеральной совокупности

Раздел 5. Статистическая проверка гипотез

Среди множества различных статистических критериев имеется ряд критериев, называемых **критериями согласия**.

Критерием согласия называется статистический критерий о предполагаемом законе распределения.

Критерий согласия предназначен для проверки согласованности основной гипотезы H_0 с выборочными данными, однако, в отличие от рассмотренных ранее критериев, альтернативная гипотеза явным образом не выдвигается.

При этом общая схема проверки гипотезы H_0, практически, остается неизменной.

Так, если выборка $X_1, \dots, X_n$ попадает в критическую область, гипотеза H_0 отвергается.

Если же выборка оказывается вне критической области, говорят, что нет оснований отклонить основную гипотезу.

Фактически, конечно, это означает, что гипотеза H_0 принимается, поскольку эмпирические данные согласуются с H_0.

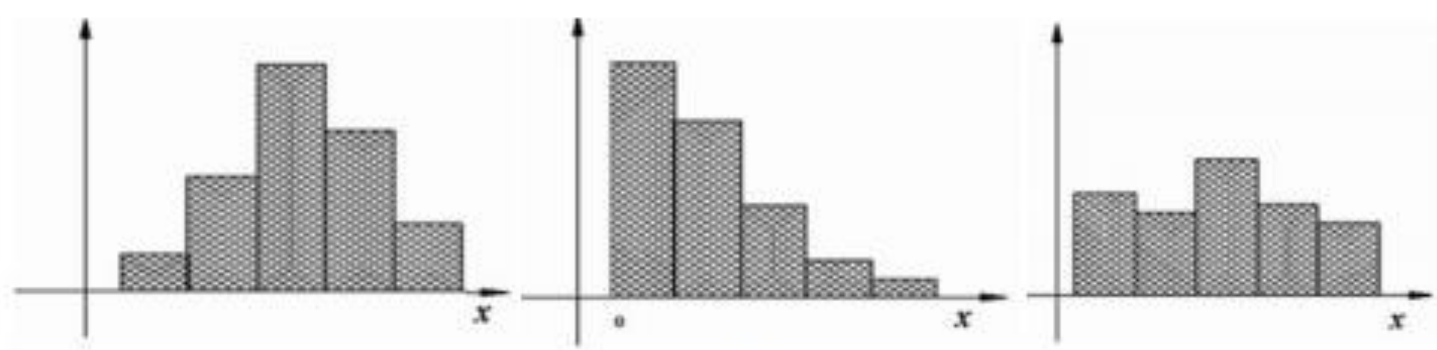
Раздел 5. Статистическая проверка гипотез

В статистической практике к основным критериям относят **критерии Пирсона, Колмогорова, Романовского, Ястремского и др**. Большинство критериев согласия базируется на использовании **отклонений эмпирических частот от теоретических**.

Очевидно, что чем меньше эти отклонения, тем лучше теоретическое распределение соответствует эмпирическому (или описывает его).

Кроме этого, сравнение гистограммы с известными кривыми функций плотностей позволяет также выдвинуть гипотезу о виде распределения генеральной совокупности.

Так, исходя из приведенных ниже гистограмм, можно предположить, что исследуемая генеральная совокупность распределена по нормальному (а), показательному (б) и равномерному (в) закону распределения.



а *б* *в*

Раздел 5. Статистическая проверка гипотез

Итак, рассмотрим **генеральную совокупность**, распределение которой неизвестно. Однако есть основание полагать, что она распределена по некоторому закону (чаще всего, **нормально**).

Это предположение может появиться как до, так и в результате статистического исследования, когда мы извлекли и изучили **выборку** объёма n.

Раздел 5. Статистическая проверка гипотез

И нам требуется **на уровне значимости** α проверить **нулевую гипотезу** H_0 – о том, что **генеральная совокупность распределена по закону** Z против конкурирующей гипотезы H_1 о том, что **она по нему НЕ распределена**.

Как проверить эту гипотезу?

Раздел 5. Статистическая проверка гипотез

Критерий согласия Пирсона – наиболее часто употребляемый критерий для проверки простой гипотезы о законе распределения.

Производится серия повторных независимых испытаний, n – число испытаний, ω_t – элементарный исход испытания с номером $t = 1, \dots, n$.

Поскольку испытания повторные, в качестве их общей вероятностной модели принимается одно и то же вероятностное пространство $(\Omega, \mathcal{F}, P)$, очевидно, что все элементарные исходы $\omega_t \in \Omega$.

Предположим, что $A_1, \ldots, A_l \in \mathcal{F}$ – попарно несовместные события, такие что

$$A_1 + \ldots + A_l \in \Omega.$$

В качестве H_0 примем гипотезу, состоящую в том, что вероятности событий A_i $(i = 1, \ldots, l)$ заданы таблицей

Событие	A_1	…	A_l
Вероятность	p_1	…	p_l

Пусть n_i – эмпирическая частота события A_i, т.е. число испытаний, в которых A_i наступило.

Эквивалентно: n_i – число тех элементарных исходов ω_t, для которых $\omega_t \in A_i$.

Исходными данными для критерия χ^2 Пирсона является таблица эмпирических частот

Событие	A_1	...	A_l
Частота	n_1	...	n_l

Раздел 5. Статистическая проверка гипотез

Если основная гипотеза верна, согласно статистическому определению вероятности

$$\hat{p}_i \approx p_i,$$

где $\hat{p}_i = n_i/n -$ относительная частота события A_i.

В качестве меры одновременной близости l пар чисел $(\hat{p}_i, p_i)$ можно принять любую сумму вида

$$c_1(\hat{p}_1 - p_1)^2 + \cdots + c_l(\hat{p}_l - p_l)^2,$$

в которой $c_i > 0$ – какие-либо положительные числа.

Раздел 5. Статистическая проверка гипотез

К.Пирсон обнаружил, что если придать большие веса маловероятным событиям, положив

$$c_i = n/p_i,$$

то при неограниченном увеличении n распределение статистики

$$\chi^2 = \sum_{i=1}^{l} \frac{n}{p_i}(\hat{p}_i - p_i)^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$$

$$[\frac{n}{p_i}(\hat{p}_i - p_i)^2 = \frac{n}{p_i}(\frac{n_i}{n} - p_i)^2 = \frac{(n_i - np_i)^2}{np_i}]$$

перестает зависеть от конкретных значений вероятностей p_i и стремится к распределению хи-квадрат с $l-1$ степенями свободы.

Заметим, что при верной H_0, случайные величины n_i распределены по **биномиальному закону** с параметрами n и p_i (то есть $n_i \sim Bin(n, p_i)$), вследствие чего $np_i = E(n_i)$ называется **ожидаемой (теоретической) частотой события** A_i .

Раздел 5. Статистическая проверка гипотез

Опуская подробности, найдем асимптотическое (при $n \longrightarrow \infty$) распределение статистики χ^2. Рассмотрим с этой целью l – мерный вектор

$$\vec{u}_0 = (\sqrt{p_1}, \dots, \sqrt{p_l}).$$

Вектор $\vec{u}_0$ имеет единичную длину, вследствие чего может быть дополнен до ортонормированного базиса.

$$\vec{u}_0, \dots, \vec{u}_{l-1}$$

в $\mathbb{R}^l$.

Определим l величин Z_k следующими формулами:

$$Z_0 = u_{0,1}X_1 + \cdots + u_{0,l}X_l,$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$Z_{l-1} = u_{l-1,1}X_1 + \cdots + u_{l-1,l}X_l,$$

где $u_{k,i} - i$ - я компонента вектора $\vec{u}_k$, а X_i определяется соотношением $X_i = \frac{n_i - np_i}{\sqrt{np_i}}$.

Раздел 5. Статистическая проверка гипотез

Из

$$\chi^2 = \sum_{i=1}^{l} \frac{n}{p_i} (\hat{p}_i - p_i)^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$$

и

$$X_i = \frac{n_i - np_i}{\sqrt{np_i}}$$

следует, что

$$\chi^2 = \sum_{i=1}^{l} X_i^2,$$

из ортонормированности базиса $\vec{u}_0, \dots, \vec{u}_{l-1}$ вытекает, что суммы квадратов величин $X_1, \dots, X_l$ и $Z_0, \dots, Z_{l-1}$ совпадают:

$$\sum_{i=1}^{l} X_i^2 = \sum_{k=0}^{l-1} Z_k^2 .$$

Нетрудно видеть, что

$$Z_0 = \sqrt{p_1}\frac{n_1 - np_1}{\sqrt{np_1}} + \cdots + \sqrt{p_l}\frac{n_l - np_l}{\sqrt{np_l}} =$$
$$= \frac{1}{\sqrt{n}}\big(n_1 + \cdots + n_l - (p_1 + \cdots + p_l)\big) = 0.$$

Значит

$$\chi^2 = \sum_{k=1}^{l-1} Z_k^2 .$$

$$\chi^2 = \sum_{i=1}^{l} X_i^2, \sum_{i=1}^{l} X_i^2 = \sum_{k=0}^{l-1} Z_k^2$$

Можно доказать, что случайный вектор $\vec{Z} = (Z_1, \dots, Z_{l-1})$ Имеет нулевое математическое ожидание и единичную ковариационную матрицу.

Если бы при этом вектор $\vec{Z}$ был еще и нормальным, то все его компоненты были бы независимыми и распределенными по стандартному нормальному закону $N(0,1)$. Отсюда, поскольку

$$\chi^2 = \sum_{k=1}^{l-1} Z_k^2$$

вытекало бы, что $\chi^2 \sim \chi^2(l-1)$.

На самом деле, конечно, вектор $\vec{Z}$ не является нормальным ни при каком n.

Однако применяя многомерный аналог центральной предельной теоремы, можно доказать, что при $n \longrightarrow \infty$ распределение вектора $\vec{Z}$ стремится к $(l-1)$ -мерному нормальному распределению.

В итоге получаем, что распределение статистики χ^2 Пирсона при достаточно большом n близко к распределению χ^2 с $(l-1)$ степенями свободы.

Можно также доказать, что если гипотеза H_0 не верна, то при $n \longrightarrow \infty$ вероятность $P(\chi^2 > c) \longrightarrow 1$ для любого c, что в конечном счете определяет достаточно высокую мощность критерия Пирсона, по крайней мере, для выборок большого объема.

$$\chi^2 = \sum_{i=1}^{l} \frac{n}{p_i} (\hat{p}_i - p_i)^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$$

Раздел 5. Статистическая проверка гипотез

В итоге приходим к заключению, что для проверки по эмпирическим данным

Событие	A_1	…	A_l
Частота	n_1	…	n_l

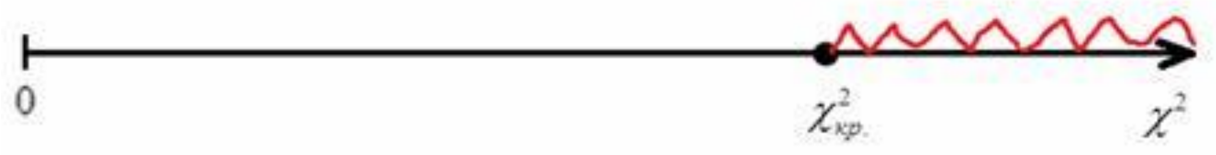
справедливости распределения

Событие	A_1	…	A_l
Вероятность	p_1	…	p_l

с асимптотическим уровнем значимости α можно использовать критерий согласия, основанный, на статистике $\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$ Пирсона и критической области $\chi^2 > \chi^2_\alpha(l-1)$.

Раздел 5. Статистическая проверка гипотез

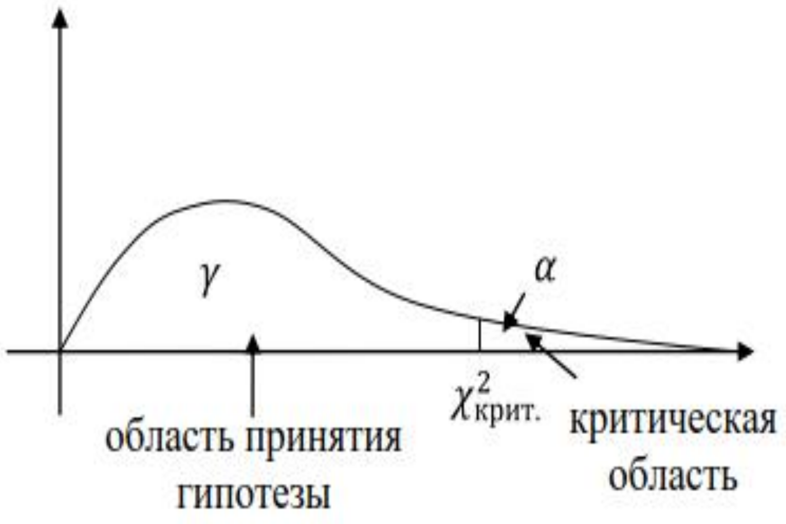
Строится правосторонняя **критическая область:**



$P(\chi^2 > \chi^2_{кр}) = \alpha \Rightarrow P(\chi^2 < \chi^2_{кр}) = 1 - P(\chi^2 > \chi^2_{кр}) = 1 - \alpha$

$P(\chi^2 < \chi^2_{кр}) = F(\chi^2_{кр})$,

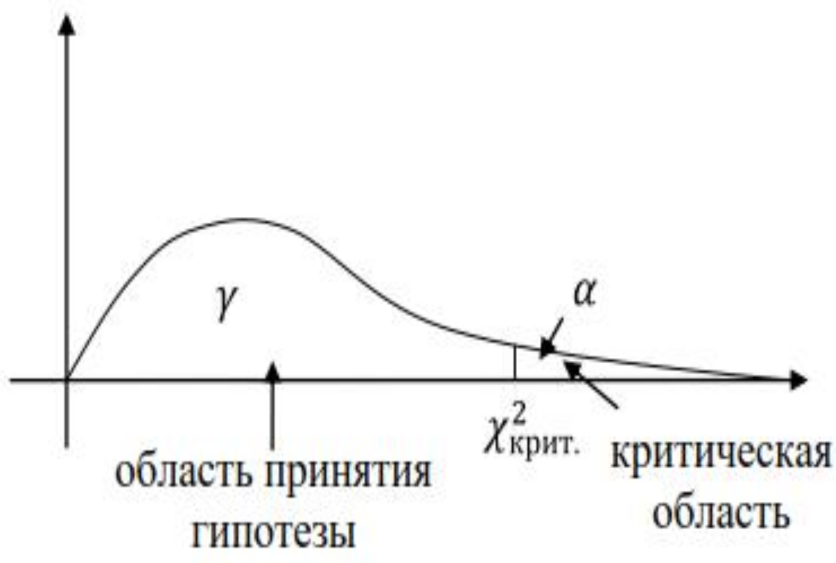
$F(x)$ - функция распределения $\chi^2 \Rightarrow F(\chi^2_{кр}) = 1 - \alpha$



Раздел 5. Статистическая проверка гипотез

На практике данный критерий Пирсона применяется, если объем выборки $n > 50$ и все ожидаемые частоты $np_i > 5$.

Несоблюдение данных условий обычно приводит к значительному отклонению фактического уровня значимости $P_{H_0}(\chi^2 > \chi^2_\alpha(l-1))$ от требуемого уровня α.



Замечание:

Малочисленные частоты $(n_i < 5)$следует объединить. В этом случае и соответствующие им теоретические частоты также надо сложить.

Если производилось объединение частот, то при определении числа степеней свободы следует принять число интервалов, оставшихся после объединения частот.

Раздел 5. Статистическая проверка гипотез

Итак, если $\chi^2_{набл}<\chi^2_{кр}$, то на уровне значимости α **нет оснований отвергать гипотезу** H_0 **о том, что генеральная совокупность распределена по закону** Z.

То есть, различие между эмпирическими и теоретическими частотами ***незначимо*** и обусловлено случайными факторами (случайностью самой выборки, способом группировки данных и т.д.).

Если $\chi^2_{набл} > \chi^2_{кр}$, то нулевую гипотезу отвергаем, иными словами эмпирические и теоретические частоты отличаются *значимо*, и это различие вряд ли случайно.

Пример. По результатам переписи населения установлен следующий возрастной состав

	До 50 лет	От 50 лет и старше
Женщины	35%	20%
Мужчины	35%	10%

Спустя несколько лет после переписи было отобрано случайным образом 1000 человек и для них подсчитано число мужчин и женщин в двух возрастных группах:

	До 50 лет	От 50 лет и старше
Женщины	343	212
Мужчины	343	102

Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что возрастной состав не изменился.

Решение.

$$n = 1000; l = 4;\ \alpha = 0{,}05; n_1 = 343; n_2 = 212; n_3 = 343; n_4 = 102;$$

Раздел 5. Статистическая проверка гипотез

Решение.

Проверим гипотезу о том, что возрастной состав не изменился при уровне значимости $\alpha = 0{,}05$.

События A_1, A_2, A_3, A_4 состоят в том, что отобранный человек оказался:

A_1– женщиной до 50 лет;

A_2– женщиной от 50 лет и старше;

A_3– мужчиной до 50 лет;

A_4– мужчиной от 50 лет и старше.

Очевидно, что события A_1, A_2, A_3, A_4 образуют полную группу, а минимальная ожидаемая частота 100 > 5.

С учетом малочисленности выборки относительно всего населения, считаем отборы людей независимыми испытаниями, поэтому проверяем гипотезу **по критерию Пирсона**.

	До 50 лет	От 50 лет и старше
Женщины	35%	20%
Мужчины	35%	10%

	До 50 лет	От 50 лет и старше
Женщины	343	212
Мужчины	343	102

Раздел 5. Статистическая проверка гипотез

Решение (продолжение). $n = 1000; l = 4; \alpha = 0{,}05;$
$n_1 = 343; n_2 = 212; n_3 = 343; n_4 = 102;$

Таким образом, $n_1 = 343, n_2 = 212$, $n_3 = 343$, $n_4 = 102$– эмпирические частоты.

Теоретические частоты:

$np_1 = 1000 \cdot 0{,}35 = 350$; $np_2 = 1000 \cdot 0{,}2 = 200$;

$np_3 = 1000 \cdot 0{,}35 = 350$; $np_4 = 1000 \cdot 0{,}1 = 100$.

Составим статистику χ^2

Статистика:

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$$

$$\chi^2 = \frac{(343-350)^2}{350} + \frac{(212-200)^2}{200} + \frac{(343-350)^2}{350} +$$

$$+ \frac{(102-100)^2}{100} = \frac{49}{350} + \frac{144}{200} + \frac{49}{350} + \frac{4}{100} = \frac{49}{175} + 0{,}72 + 0{,}04 = 1{,}04$$

	До 50 лет	От 50 лет и старше
Женщины	35%	20%
Мужчины	35%	10%

	До 50 лет	От 50 лет и старше
Женщины	343	212
Мужчины	343	102

Решение (продолжение).

Затем находим 5-процентную точку распределения хи-квадрат с 3 степенями свободы: $\chi^2_{0,05}(4-1) = 7{,}815.$

Поскольку

$$\chi^2_{\text{набл.}} = 1{,}04 \ < \chi^2_{\text{кр.}} = 7{,}815,$$

приходим к выводу, что данные выборочного исследования не обнаруживают значимого изменения возрастного состава населения.

$n = 1000; l = 4;$
$\alpha = 0{,}05;$
$n_1 = 343; n_2 = 212;$
$n_3 = 343;\ n_4 = 102;$

$\chi^2_{\text{кр.}}$**=ХИ2.ОБР(1-0,05;4-1)** **# MS Excel** $\chi^2_{\text{кр.}} = 7{,}815$

 # Python

```
import scipy.stats as sts
import numpy as np
```

```
n=1000; l=4; alpha = 0.05
ni=np.array([343,212,343,102])
pi=np.array([0.35,0.2,0.35,0.1])
npi=n*pi
chi2_набл=np.sum((ni-npi)**2/npi)
chi2_cr=sts.chi2.ppf(1-alpha,l-1)
print ("Значение chi2_набл=:", chi2_набл)
print ("Значение chi2_cr=:", chi2_cr)
```

```
Значение chi2_набл=: 1.04
Значение chi2_cr=: 7.814727903251179
```

Теория вероятностей и математическая статистика

Конец лекции