Случайная величина -

Генеральная совокупность X (сл. вел. X) – множество возможных значений случайной величины X

Закон распределения ген. сов. X – закон распределения сл. вел. X

Случайная выборка из ген. сов. X – совокупность независимых сл. вел. $X_1, \dots, X_n$, каждая из которых имеет то же распределение, что и сл. вел. X. Записывается $\overline{X_n} = (X_1, \dots, X_n)$

Выборка из ген. сов. X (реализация сл. выборки $\overline{X_n}$) – это любое возможное значение $\overline{x_n} = (x_1, \dots, x_n)$ сл. выборки $\overline{X_n}$. Интерпретируется как результаты n независимых наблюдений над сл. величиной X.

Выборочный метод – свойства сл. вел. X устанавливаются путем изучения тех же свойств на случайной выборке.

Выборочное пространство - \chi _n χ_n – множество значений сл. выборки $\overline{X_n}$.

Как выражается функция распределения сл. выборки $F_{\overline{X_n}}(t_1, \dots, t_n)$ через функцию распределения генеральной совокупности (т.е. через ф-цию распр. X).

$$F_{\vec{X}}(t_1, \dots, t_n) = \mathbf{P}\{X_1 < t_1, \dots, X_n < t_n\} = \prod_{i=1}^{n} \mathbf{P}\{X_i < t_i\} = \prod_{i=1}^{n} F(t_i), \quad (1.1)$$

где $F(t)$ — функция распределения случайной величины X (генеральной совокупности X).

стр 21. *Статистическая модель* – выборочное пространство, на котором задан класс распределений сл. выборки (если мы знаем тип функции распределения, но не знаем ее параметры).

Параметрическая модель

Статистика (выборочная характеристика) – любая функция случайной выборки $g(X_1, \dots, X_n) = g(\overline{X_n})$ – она является случайной величиной с распределением, называемым **выборочным распределением**

Выборочное распределение выборочной характеристики

Выборочное значение выборочной характеристики – значение $g(\overline{x_n})$ выборочной характеристики $g(\overline{X_n})$, определенное по реализации $\overline{x_n}$ случайной выборки $\overline{X_n}$.

стр. 24 ***Сходимость по вероятности***

Сходимость по распределению (слабая)

Задачи мат. статистики: оценка неизв. параметров, проверка стат. гипотез, установление формы и степени связи между сл. вел.

Два подхода к оценке неизвестный параметров функции распределения генеральной совокупности – точечная оценка и интервальная оценка.

Точечная оценка (или просто **оценка**) неизвестного параметра θ ф-ции распр. генеральной совокупности – это статистка (функция) $\hat{\theta}(\overline{X_n})$, выборочное значение $\hat{\theta} = \hat{\theta}(\overline{x_n})$ которой для любой реализации $\overline{x_n}$ принимают за приближенное значение неизвестного параметра θ.

Значение точечной оценки - $\hat{\theta}$.

Интервальная оценка с коэффициентом доверия γ неизвестного параметра θ ф-ции распр. генеральной совокупности – это пара статистик (функций) $\underline{\theta}(\overline{X_n})$ и $\overline{\theta}(\overline{X_n})$ таких, что с вероятностью γ выполняется неравенство $\underline{\theta}(\overline{X_n}) \le \theta \le \overline{\theta}(\overline{X_n})$ (то есть таких, что $P\{\underline{\theta}(\overline{X_n}) \le \theta \le \overline{\theta}(\overline{X_n})\} = \gamma$).

Доверительный интервал для θ с коэффициентом доверия γ - $\left(\underline{\theta}(\overline{X_n}), \overline{\theta}(\overline{X_n})\right)$.

Статистическая гипотеза – любое предположение о распределении вероятностей (о вероятностных свойствах) наблюдаемой сл.вел. (гипотеза о величине м.о., об однородности (т.е. равенстве) дисперсий, о виде распределения и т.д.).

Корреляционный и *дисперсионный* анализ – наличие связи между величинами и ее существенность. *Регрессионный* анализ – построение регрессионной модели (т.е. зависимости ср. знач. сл. величины от знач. других сл. величин).

Вариационный ряд выборки $(\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$ – упорядоченная последовательность элементов выборки $(x_{(1)}, \dots, x_{(n)})$.

Вариационный ряд случайной выборки $(\boldsymbol{X_1}, \dots, \boldsymbol{X_n})$ – последовательность случайных величин $(X_{(1)}, \dots, X_{(n)})$, где $X_{(i)}$ – сл. величина, которая при каждой реализации $\overline{x_n}$ случайной выборки $\overline{X_n}$ принимает значение, равное i-му члену вариационного ряда выборки $\overline{x_n}$.

Функции распр. крайних членов вариационного ряда ($\boldsymbol{X_{(1)}}$ и $\boldsymbol{X_{(n)}}$). Вывод

показать (см. пример 2.20), что для ***крайних членов вариационного ряда*** случайной выборки $X_{(1)}$ и $X_{(n)}$ их функции распределения имеют вид

$$\mathbf{P}\{X_{(1)} < x\} = 1 - (1 - F(x))^n$$

и

$$\mathbf{P}\{X_{(n)} < x\} = F^n(x).$$

Статистический ряд – таблица, которая в первой строчке содержит уникальные отсортированные значения элементов выборки, а во второй – количество их повторений.

Частота – количество раз, которое встречается элемент в выборке.

Относительная частота (частость) – отношение частоты значения элемента выборке к общему количеству элементов в выборке.

Интервальный статистический ряд – отрезок, содержащий все значения выборки, делят на равные части и составляют статистический ряд, в котором количество элементов подсчитывается на интервале.

Оптимальное число интервалов для гистограммы – по правилу Стёрджеса – $m = 1 + \lfloor \log_2 n \rfloor$.

Стр 32 **Выборочная функция распределения**

Эмпирическая функция распределения

Теоретическая функция распределения

Эмпирич. плотность распр.

Гистограмма, Полигон частот

Выборочные числовые моменты

Теоретические (генеральные) числовые характеристики

Выборочный начальный момент k-го порядка,

Выборочный центральный момент k-го порядка

Выборочное среднее, Выборочная дисперсия,

Выборочное ср/квадр. отклонение

Выборочный корреляционный момент,

Выборочный коэффициент корреляции

Кор. момент выборки,

Коэф. кор. выборки

из лабы 1 **Коэффициент вариации**

Стандартное отклонение

Стандартизованная асимметрия

Стандартизованный эксцесс

! Актуальные критерии нормальности распределения (асимметрия и эксцесс и что-то еще?)

из лабы 2 **Состоятельность оценки**

! Правила для определения достаточного объема выборки

Законы больших чисел

Теорема Бернулли – закон больших чисел https://studopedia.ru/12_163342_reshenie.html

При неограниченном увеличении числа однородных независимых опытов частота события будет сколь угодно мало отличаться от вероятности события в отдельном опыте.

Иначе, вероятность того, что отклонение относительной частоты m/n наступления события A от постоянной вероятности p события A очень мало при n -> +oo, стремится к 1 при любом eps > 0

$$P\left\{\left|\frac{m}{n}-p\right|<\varepsilon\right\}\underset{n\to\infty}{\to}1$$

Геометрическое распределение - распределение вероятностей случайной величины X равной количеству «неудач» до первого «успеха» в серии испытаний Бернулли и принимающей значения $n = 0,1,2, \ldots$ либо распределение вероятностей случайной величины $Y = X + 1$ равной номеру первого «успеха» и принимающей значения $n = 1,2,3, \ldots$.

Функция вероятности	$q^n p$
Функция распределения	$1 - q^{n+1}$

Экспоненциальное распределение - абсолютно непрерывное распределение, моделирующее время между двумя последовательными свершениями одного и того же события.

Обозначение	$\mathrm{Exp}(\lambda)$
Параметры	$\lambda > 0$ - интенсивность или обратный коэффициент масштаба
Носитель	$x \in [0; \infty)$
Плотность вероятности	$\lambda e^{-\lambda x}$
Функция распределения	$1 - e^{-\lambda x}$
Математическое ожидание	λ^{-1}

Распределение Бернулли – дискретное распределение вероятностей, моделирующее случайный эксперимент произвольной природы, при заранее известной вероятности успеха (p) или неудачи ($p - 1$).

Функция вероятности	$\begin{matrix} q & k=0 \\ p & k=1 \end{matrix}$
Функция распределения	$\begin{matrix} 0 & k<0 \\ q & 0\leq k<1 \\ 1 & k\geq 1 \end{matrix}$

Биномиальное распределение - распределение количества «успехов» в последовательности из n независимых случайных экспериментов, таких, что вероятность «успеха» в каждом из них постоянна и равна p.

Пусть $X_1, \ldots, X_n$- конечная последовательность независимых случайных величин, имеющих одинаковое распределение Бернулли с параметром p. Тогда сл. вел. $Y = X_1 + \cdots + X_n$ имеет биномиальное распределение с параметрами n и p. $Y \sim Bin(n, p)$.

Носитель	$k \in \{0, \ldots, n\}$
Функция вероятности	$\binom{n}{k} p^k q^{n-k}$

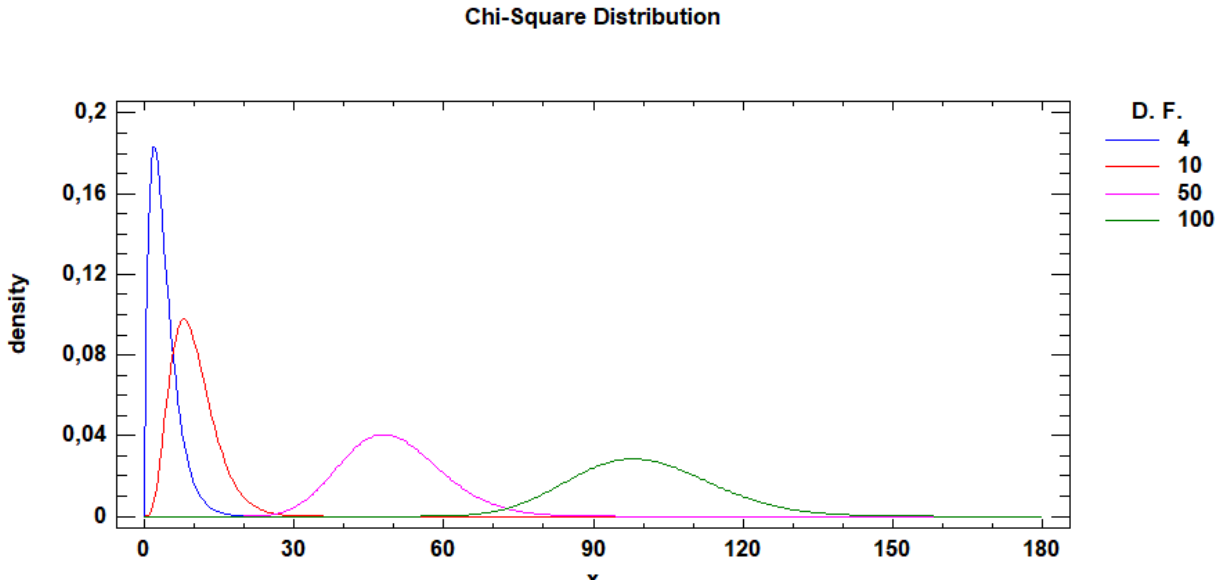
$$F_Y(y) \equiv \mathbb{P}(Y \leqslant y) = \sum_{k=0}^{\lfloor y \rfloor} \binom{n}{k} p^k q^{n-k}, \ y \in \mathbb{R},$$

Распределение Пуассона – вероятностное распределение дискретного типа, моделирует случайную величину, представляющую собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга. $Y \sim P(\lambda)$, где $\lambda > 0$ – м.о.

Функция вероятности $p(k) \equiv \mathbb{P}(Y = k) = \dfrac{\lambda^k}{k!} e^{-\lambda}$, Функция распределения $\dfrac{\Gamma(k+1, \lambda)}{k!}$.

Распределение χ^2 (Chi-Squared) с k степенями свободы — это распределение суммы квадратов k независимых стандартных нормальных случайных величин.

Пусть $Z_1, \ldots, Z_k$ – совместно независимые стандартные нормальные случайные величины, то есть $Z_i \sim N(0,1)$. Тогда случайная величина $X = Z_1^2 + \cdots + Z_k^2$ имеет распределение хи-квадрат с k степенями свободы, т.е. $X \sim f_{\chi^2(k)}(X)$.

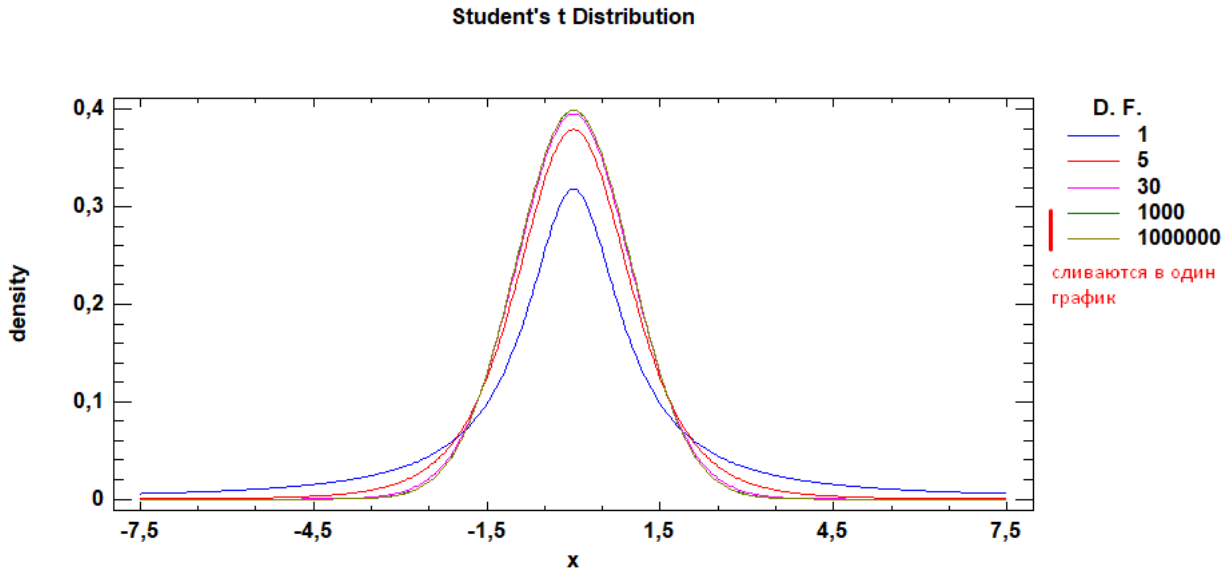


Распределение Стьюдента (Student's t) – это однопараметрическое семейство абсолютно

непрерывных распределений. Пусть $Y_0, \dots, Y_n$ – конечная последовательность независимых стандартных нормальных случайных величин, т.е. $Y_i \sim N(0,1)$. Тогда распределение сл. вел.

$$t = \frac{Y_0}{\sqrt{\frac{1}{2}\sum_{i=1}^{n} Y_i^2}}$$

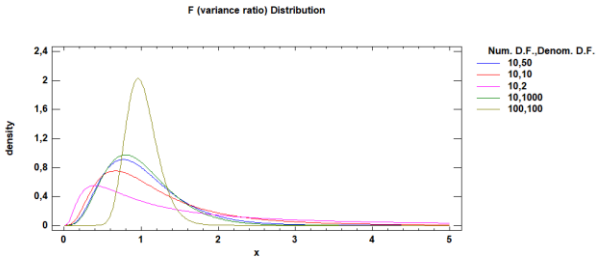
имеет распределение Стьюдента с n степенями свободы. $t \sim t(n)$.

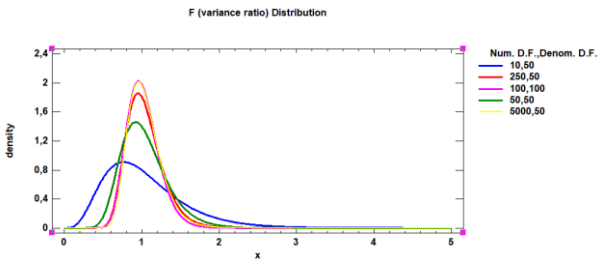


Распределение Фишера F (Variance Ratio) – это двухпараметрическое семейство абсолютно непрерывных распределений. Пусть Y_1, Y_2 — две независимые случайные величины, имеющие распределение хи-квадрат: $Y_i \sim \chi^2(d_i)$, где $d_i \in \mathbb{N}, i = 1, 2$. Тогда распределение случайной величины

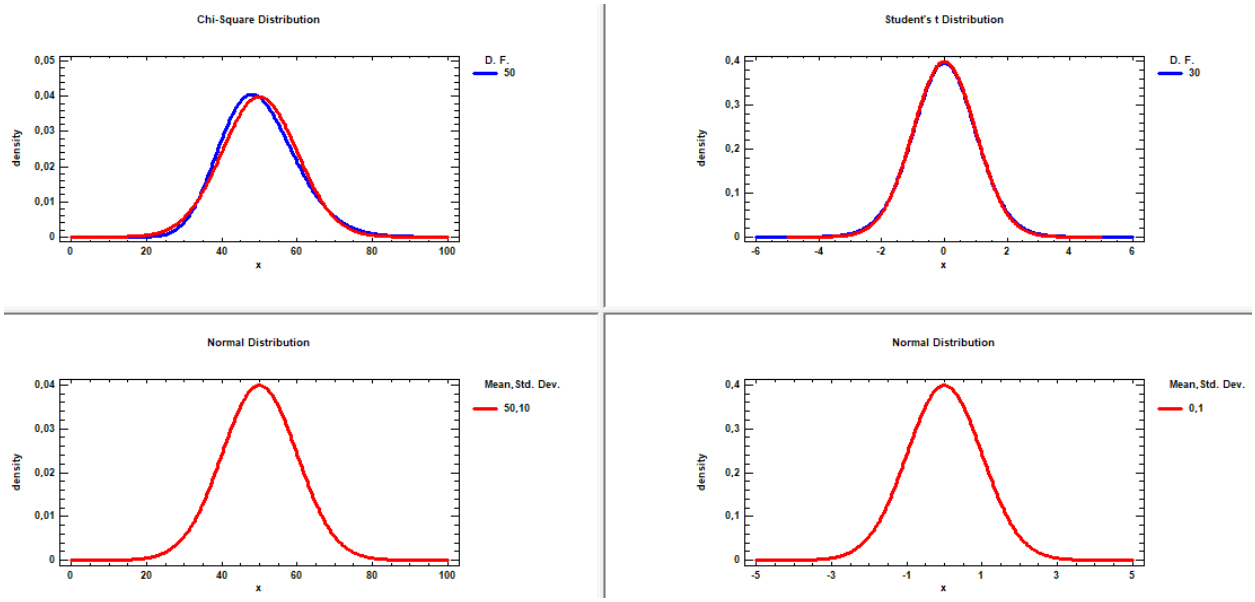
$$F = \frac{Y_1/d_1}{Y_2/d_2},$$

называется распределением Фишера со степенями свободы d_1 и d_2. Пишут $F \sim F(d_1, d_2)$.





Асимптотическая нормальность распределений Стьюдента и χ^2 – $\mathrm{t}(30) \cong \mathrm{N}(0,1)$, а $\chi^2(\mathrm{v}) \cong \mathrm{N}(\mathrm{v}, \sqrt{2 * \mathrm{v}})$ при $\mathrm{v} \geq 50$.



Bin(0.5, 100), Bin(0.01, 100), Bin(0.99, 100)

А) Для какой из выборок гистограмма «похожа» на нормальную кривую? Почему это можно было ожидать (вспомните предельные теоремы из теории вероятностей (какую???).

Б) На какое распределение должна быть «похожа» гистограмма для второго распределения? Наложите это распределение на гистограмму.

В) Почему нормальная аппроксимация дает плохой результат для третьей выборки?

~~! Тест Колмогорова-Смирнова для проверки нормальности~~

~~! Критерий χ^2~~

Локальная теорема Муавра — Лапласа

Если в схеме Бернулли n стремится к бесконечности, величина $p \in (0,1)$ постоянна, а величина $x_m = \dfrac{m-np}{\sqrt{npq}}$ ограничена равномерно по m и n (то есть $\exists a, b : -\infty < a \leqslant x_m \leqslant b < +\infty$), то

$$P_n(m) = \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{x_m^2}{2}\right)(1+\alpha_n(m))$$

где $|\alpha_n(m)| < \dfrac{c}{\sqrt{n}}, c = \text{const} > 0.$

Приближённую формулу

$$P_n(m) \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{x_m^2}{2}\right)$$

рекомендуется применять при $n > 100$ и при $m > 20$.

γ-доверительная интервальная оценка – $\left(\underline{\theta}(\overline{X_n}), \overline{\theta}(\overline{X_n})\right)$.

Нижняя и верхняя границы интервальной оценки - пара статистик (функций) $\underline{\theta}(\overline{X_n})$ и $\overline{\theta}(\overline{X_n})$.

Коэффициент доверия (доверительная вероятность, уровень доверия).

Односторонняя нижняя (и соответственно верхняя) – доверительная граница – $\underline{\theta}(\overline{X_n})$, когда $P\left\{\underline{\theta}(\overline{X_n}) \leq \theta\right\} = \gamma$.

Пример 3.1. Пусть θ — среднее значение предела прочности X некоторого материала, которое оценивают независимо друг от друга в каждой из N различных лабораторий по результатам n независимых натурных испытаний. Иначе говоря, среднее значение предела прочности в каждой лаборатории оценивают по „своим“ *экспериментальным данным*, представленным *выборкой объема* n, и в каждой лаборатории получают „свои“ значения верхней и нижней границ γ-доверительного интервала (рис. 3.1).

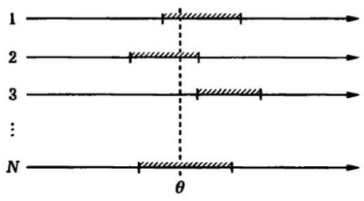


Рис. 3.1

Возможны случаи, когда γ-доверительный интервал для параметра θ не накрывает его истинного значения. Если M — число таких случаев, то при больших значениях N должно выполняться приближенное равенство $\gamma \approx (N-M)/N$. Таким образом, если опыт — получение выборки объема n в лаборатории, то уровень доверия γ — доля тех опытов (при их многократном независимом повторении), в каждом из которых γ-доверительный интервал накрывает истинное значение оцениваемого параметра.

Определение. Пусть $\alpha \in (0,1)$. Две статистики $\widehat{\theta}_1$ и $\widehat{\theta}_2$ определяют границы *доверительного интервала для параметра* θ *с коэффициентом доверия* $1-\alpha$, если при всех $\theta \in \Theta$ для выборки $\boldsymbol{X} = (X_1, \ldots, X_n)$ из закона распределения $F_\theta(x)$ справедливо неравенство

$$\mathbf{P}\left(\widehat{\theta}_1(\boldsymbol{X}) < \theta < \widehat{\theta}_2(\boldsymbol{X})\right) \geqslant 1-\alpha. \qquad (1)$$

Часто на практике полагают $\alpha = 0{,}05$. Если вероятность в левой части неравенства (1) стремится к $1-\alpha$ при $n \to \infty$, то интервал называется *асимптотическим*. Как правило, длина доверительного интервала возрастает при увеличении коэффициента доверия $1-\alpha$ и стремится к нулю с ростом размера выборки n.

§ 2. ИНТЕРВАЛЫ В НОРМАЛЬНОЙ МОДЕЛИ

Пример 2. Допустим, что элементы выборки X_i распределены по закону $\mathcal{N}(\theta, \sigma^2)$, причем параметр масштаба σ известен, а параметр сдвига θ — нет. Эту модель часто применяют к данным, полученным при независимых измерениях некоторой величины θ с помощью прибора (или метода), имеющего известную среднюю погрешность (стандартную ошибку) σ (рис. 3).

Пусть $\Phi(x) = (2\pi)^{-1/2} \int\limits_{-\infty}^{x} e^{-u^2/2} du$ — функция распределения закона $\mathcal{N}(0,1)$. Для $0 < \alpha < 1$ обозначим через x_α так называемую α-квантиль этого закона, т. е. решение уравнения $\Phi(x_\alpha) = \alpha$ (см. § 3 гл. 7). Приведем некоторые значения $x_{1-\alpha/2}$ (см. также таблицу Т2):

α	0,05	10^{-2}	10^{-3}	10^{-5}
$x_{1-\alpha/2}$	1,96	2,58	3,29	4,26

Согласно примеру 4 гл. 9, эффективной оценкой для θ служит $\overline{X}$. Известно, что $\overline{X} \sim \mathcal{N}(\theta, \sigma^2/n)$. Тогда $\sqrt{n}\,(\overline{X}-\theta)/\sigma \sim \mathcal{N}(0,1)$. Поэтому в качестве границ интервала с коэффициентом доверия $1-\alpha$ можно взять $\widehat{\theta}_1 = \overline{X} - \sigma x_{1-\alpha/2}/\sqrt{n}$ и $\widehat{\theta}_2 = \overline{X} - \sigma x_{\alpha/2}/\sqrt{n}$:

$$\mathbf{P}(\widehat{\theta}_1 < \theta < \widehat{\theta}_2) = \mathbf{P}(x_{\alpha/2} < \sqrt{n}\,(\overline{X}-\theta)/\sigma < x_{1-\alpha/2}) = 1-\alpha.$$

В силу четности плотности закона $\mathcal{N}(0,1)$ верно равенство $x_{\alpha/2} = -x_{1-\alpha/2}$. Таким образом, из приведенной выше таблицы видим, что с вероятностью 0,95 истинное значение параметра сдвига θ находится в интервале $\overline{X} \pm 1{,}96\,\sigma/\sqrt{n} \approx \overline{X} \pm 2\sigma/\sqrt{n}$ (**правило двух сигм**).

Статистический критерий – правило, позволяющее принять или отвергнуть гипотезу Н на основе реализации выборки $x_1, \dots, x_n$.

Статистика критерия – $T(x_1, \dots, x_n)$ – статистика (функция), для которой типично принимать умеренные значения в случае, когда гипотеза Н верна, и большие (малые), когда Н не выполняется.

Уровень значимости – α – вероятность, с которой мы можем позволить себе отвергнуть верную гипотезу (вероятность ошибочного отклонения правильной гипотезы).

Если значение T попало в область, имеющую при выполнении гипотезы H высокую вероятность, то можно заключить, что данные *согласуются* с гипотезой H. Отсюда происходит термин *«критерии согласия»*.

Схема

Берем $T(x_1, \dots, x_n)$ – статистика (какая-то функция),
$(x_1, \dots, x_n)$ – реализация выборки (данные эксперимента).
Делаем гипотезу Н, выбираем приемлемый уровень значимости α.

Если Н – верна, то у нас есть определенные ожидания от значения Т.
Мы находим $x_{1-\alpha}$, такое, что $\boldsymbol{P}(T(X_1, \dots, X_n) \geq x_{1-\alpha}) \leq \alpha$, то есть $x_{1-\alpha}$ – это максимальное значение для функции Т такое, что её вероятность быть больше этого значения «равна» α (точнее не больше α, то есть маленькая). Потом вычисляем реальное $T(x_1, \dots, x_n) = t_0$ и смотрим, $t_0 < x_{1-\alpha}$? (что более ожидаемо при выполнении Н) или нет?

Критическое значение – $x_{1-\alpha}$ – значение статистики критерия $T(x_1, \dots, x_n)$, при превышении которого мы должны отвергнуть гипотезу (так как по факту произошло маловероятное событие и наше предположение, наша гипотеза, скорее всего, не верна).

Фактический уровень значимости – $\alpha_0 = \boldsymbol{P}\big(T(X_1, \dots, X_n) \geq t_0 = T(x_1, \dots, x_n)\big) \leq \alpha$ – вероятность, с которой значение статистики может превысить ее фактическое значение на данной реализации случайной выборки $(x_1, \dots, x_n)$.

Статистическая гипотеза, Простая гипотеза, Сложная гипотеза

введем формально понятие статистической гипотезы.

Напомним, что под статистической моделью в § 1 гл. 6 понималось семейство функций распределения $\{F(x,\theta), \theta \in \Theta\}$, где Θ — множество возможных значений параметра. При этом данные $x_1, \dots, x_n$ рассматривались как реализация выборки $X_1, \dots, X_n$, элементы которой имеют функцию распределения $F(x,\theta_0)$ с неизвестным значением $\theta_0 \in \Theta$.

Пусть выделено некоторое подмножество $\Theta_0 \subset \Theta$. Под *статистической гипотезой* H понимается предположение о том, что $\theta_0 \in \Theta_0$. Если множество Θ_0 состоит всего из одной точки, то гипотеза H называется *простой*, иначе — *сложной*. В последнем случае задача заключается в проверке принадлежности закона распределения величин X_i целому классу функций распределения $\{F(x,\theta), \theta \in \Theta_0\}$.

Нулевая гипотеза

Альтернативная гипотеза

Двусторонняя гипотеза

Односторонняя

Критерии согласия – критерии проверки гипотез о типе распределения генеральной совокупности (о соответствии эмпирического распределения теоретическому закону распределения).

Общие критерии согласия – применимы к самой общей формулировке гипотезы, а именно к гипотезе о согласии наблюдаемых результатов с любым априорно предполагаемым распределением вероятностей.

Специальные критерии согласия – предполагают специальные нулевые гипотезы, формулирующие согласие с определенной формой распределения вероятностей.

Критерий согласия Пирсона χ^2 –

! Тест Колмогорова-Смирнова для проверки нормальности

Параметрическая гипотеза

Ошибка первого рода

Ошибка второго рода

Критическая область

Наилучшая критическая область (область принятия решений)

Параметрическая/Непараметрическая гипотеза

Статистическая значимость какого-либо значения

Непараметрический критерий

Парные/непарные наблюдения/критерии

Независимые (непарные) наблюдения

Критерий знаков

Критерий знаковых ранговых сумм

Однородность выборки

Равномерность выборки

Сравнение нескольких выборок

Таблица сопряженности – таблица частот (?)

Методика	Оценка			
	отлично	хорошо	удовлетвори-тельно	неудовлетвори-тельно
1	7	12	15	7
2	10	18	26	9
3	40	50	87	10