Доверительные интервалы. Проверка гипотез.

Леонид Иосипой

Курс «Вероятностные модели и статистика»
Центр непрерывного образования, ВШЭ

20 апреля 2021

- Повторение
- Доверительные интервалы
- Бутстрэп
- Проверка гипотез
- Критерии согласия

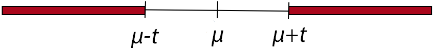
Повторение

1. Закон больших чисел (ЗБЧ).

Теорема (Закон больших чисел, ЗБЧ)

Пусть $X_1, X_2, X_3, \ldots$ *— н.о.р.с.в. Пусть также* $\mu = \mathbb{E}[X_1]$ *и* $\sigma^2 = \mathrm{Var}(X_1)$ *существуют. Определим* $S_n = X_1 + \ldots + X_n$. *Тогда для любого* $t > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > t\right) \to 0 \quad \text{при } n \to \infty.$$



Это утверждение иногда записывают так: $\dfrac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$.

Повторение

2. Центральная предельная теорема (ЦПТ).

Теорема (Центральная предельная теорема, ЦПТ)

В обозначениях, которые для ЗБЧ, для любых $a < b$,

$$\mathbb{P}\left(a < \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} < b\right) \to \mathbb{P}(a < Z < b) \quad \textit{при } n \to \infty,$$

где $Z \sim \mathcal{N}(0, 1)$.

Повторение

3. Свойства оценок.

Оценка $\widehat{\theta}(x_1, \ldots, x_n)$ параметра θ называется несмещенной, если

$$\mathbb{E}_\theta \left[\widehat{\theta}(X_1, \ldots, X_n)\right] = \theta \;\text{ для всех } \theta \in \Theta.$$

Несмещенность означает, что при многократном вычислении оценки на разных данных среднее арифметическое полученных оценок будет стремится к истинному значению параметра θ.

Повторение

3. Свойства оценок.

Оценка $\widehat{\theta}(x_1, \ldots, x_n)$ параметра θ называется состоятельной, если для всех $\theta \in \Theta$

$$\widehat{\theta}(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}_\theta} \theta \quad \text{при } n \to \infty.$$

Состоятельность оценки означает концентрацию оценки около истинного значения параметра с ростом размера выборки n (что устремив $n \to \infty$, оценка сойдется к истинному значению параметра θ).

Повторение

4. Оценивание характеристик распределения.

- Несмещенная и состоятельная оценка $\mathbb{E}[X]$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Несмещенная и состоятельная оценка $\mathrm{Var}(X)$:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- Несмещенная и состоятельная оценка $\mathbb{E}[g(X)]$:

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

Доверительные интервалы

Пусть имеется реализация выборки $x_1, \ldots, x_n$ из некоторого распределения с неизвестным (одномерным) параметром

$$\theta \in \Theta \subset \mathbb{R}.$$

До сих пор мы занимались «точечным оцениванием» неизвестного параметра — находили оценку, способную в некотором смысле заменить параметр.

Существует другой подход к оцениванию, при котором мы указываем интервал, накрывающий параметр с заданной наперед вероятностью. Такой подход называется «интервальным оцениванием».

Доверительные интервалы

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

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

Как правило, длина доверительного интервала возрастает при увеличении коэффициента доверия $1 - \alpha$ и стремится к нулю с ростом размера выборки n.

Доверительные интервалы

Задача

Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения $\mathcal{N}(\theta, \sigma^2)$ с неизвестным параметром $\theta \in \mathbb{R}$ и известным параметром $\sigma^2 > 0$.

Построить точный доверительный интервал для параметра θ уровня доверия $1 - \alpha$.

Доверительные интервалы

Решение. Будем пользоваться фактом, что нормальное распределение устойчиво по суммированию:

если

- $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$,
- $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$,
- X_1 и X_2 независимы,

то

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Доверительные интервалы

Поэтому распределение суммы элементов выборки нормально:

$$n\overline{X} = X_1 + \ldots + X_n \sim \mathcal{N}(n\theta, n\sigma^2).$$

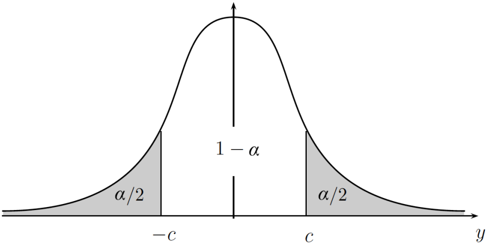
Следовательно, после стандартизации суммы мы получим стандартное нормальное распределение:

$$\frac{n\overline{X} - n\theta}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}(\overline{X} - \theta)}{\sigma} \sim \mathcal{N}(0, 1).$$

Доверительные интервалы

По заданному $\alpha > 0$ найдём число c такое, что

$$\mathbb{P}\left(-c < \frac{\sqrt{n}(\overline{X} - \theta)}{\sigma} < c\right) = 1 - \alpha.$$



Доверительные интервалы

Разрешив затем неравенство внутри вероятности относительно θ, получим точный доверительный интервал:

$$\mathbb{P}\left(\overline{X} - \frac{c\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{c\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Это можно записать и так:

$$\theta \in \left(\overline{X} - \frac{c\sigma}{\sqrt{n}}, \overline{X} + \frac{c\sigma}{\sqrt{n}}\right) \quad \text{с вероятностью } 1 - \alpha.$$

Доверительные интервалы

Пусть $F(x)$ — функция распределения некоторого закона.
Число c_α называется квантилью уровня α, если $F(c_\alpha) = \alpha$.

Если функция F строго монотонна, квантиль определяется единственным образом.

Доверительные интервалы

Итак, искомый точный доверительный для нормального распределения имеет вид:

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

где мы использовали тот факт, что $c_{\alpha/2} = -c_{1-\alpha/2}$.

- Какова середина полученного доверительного интервала?
- Какова его длина?
- Что происходит с его границами при $n \to \infty$?

Доверительные интервалы

- Зачем мы брали симметричные квантили?
- Какой будет длина, например, у такого доверительного интервала?

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/3}\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{c_{1-2\alpha/3}\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- Какой из двух доверительных интервалов одного уровня доверия и разной длины следует предпочесть?

Доверительные интервалы

Алгоритм построения точных доверительных интервалов:

1. Найти функцию $G(\mathbf{X}, \theta)$ с известным распределением, которое не зависит от неизвестного параметра θ. Необходимо, чтобы функция $G(\mathbf{X}, \theta)$ была обратима по θ.
2. Найти числа c_1 и c_2 — квантили распределения, для которых
$$\mathbb{P}(c_1 < G(\mathbf{X}, \theta) < c_2) = 1 - \alpha.$$
3. Разрешив неравенство $c_1 < G(\mathbf{X}, \theta) < c_2$ относительно θ получить точный доверительный интервал.

Доверительные интервалы в нормальной модели

Задача

Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения $\mathcal{N}(\mu, \sigma^2)$. Мы построили точный доверительный интервал для среднего $\mu \in \mathbb{R}$ нормального распределения $\mathcal{N}(\mu, \sigma^2)$ при известной дисперсии $\sigma^2 > 0$.

Построим оставшиеся точные доверительные интервалы: для σ^2 при известном и при неизвестном μ, а также для μ при неизвестной σ^2.

Доверительные интервалы в нормальной модели

Можно ли, пользуясь схемой построения доверительного интервала для μ, построить доверительный интервал для σ^2?

Попробуйте разрешить неравенство относительно σ:

$$-c < \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} < c.$$

- Чем плох интервал бесконечной длины?
- А получился ли интервал бесконечной длины?

Доверительные интервалы в нормальной модели

Для решения этих задач требуется отыскать такие функции от выборки и неизвестных параметров, распределения которых не зависят от этих параметров.

Особый интерес к нормальному распределению связан, разумеется, с центральной предельной теоремой: почти всё в этом мире нормально (или близко к нормальному).

Доверительные интервалы в нормальной модели

Пусть $X_1, \ldots, X_k$ независимы и имеют стандартное нормальное распределение $\mathcal{N}(0, 1)$.

Распределением χ^2 (хи-квадрат) с k степенями свободы называется распределение случайной величины

$$Y = X_1^2 + \ldots + X_k^2.$$

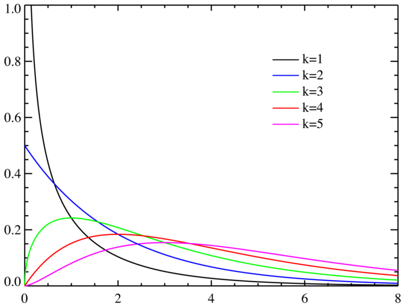
Обозначение: χ_k^2.

Доверительные интервалы в нормальной модели

Плотность распределения хи-квадрат с k степенями свободы:

$$f(u) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)}\, u^{k/2-1} e^{-u/2}, & u > 0, \\ 0, & u \le 0, \end{cases}$$

где $\Gamma(u)$ — гамма-функция Эйлера (специальная функция).



Доверительные интервалы в нормальной модели

Английский статистик Госсет, публиковавший научные труды под псевдонимом Стьюдент, ввёл следующее распределение.

Пусть $X_0, X_1, \ldots, X_k$ независимы и имеют стандартное нормальное распределение $\mathcal{N}(0, 1)$.

Распределением Стьюдента с k степенями свободы называется распределение случайной величины

$$Y = \frac{X_0}{\sqrt{\frac{X_1^2 + \ldots + X_k^2}{k}}}.$$

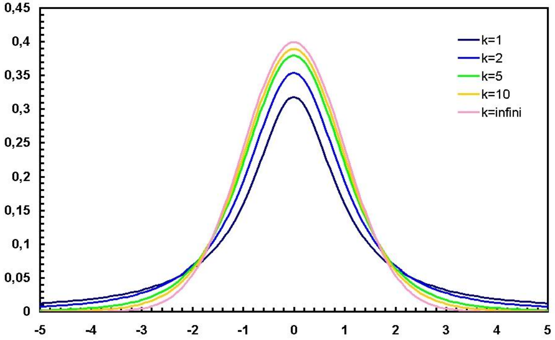
Обозначение: t_k.

Доверительные интервалы в нормальной модели

Плотность распределения Стьюдента с k степенями свободы:

$$f(u) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})} \left(1 + \frac{u^2}{k}\right)^{-\frac{k+1}{2}},$$

где $\Gamma(u)$ — гамма-функция Эйлера (специальная функция).



Доверительные интервалы в нормальной модели

Теорема (Лемма Фишера)

Пусть $X_1, \dots, X_n$ *— выборка из нормального распределения* $\mathcal{N}(\mu, \sigma^2)$*. Обозначим*

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

Тогда случайные величины $\overline{X}$ *и* S^2 *независимы и*

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Доверительные интервалы в нормальной модели

Вернемся к задаче построения доверительных интервалов в нормальной модели. Какие статистики можно использовать?

- Доверительный интервал для μ при известном σ^2:

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1).$$

- Доверительный интервал для μ при неизвестном σ^2:

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}.$$

Доверительные интервалы в нормальной модели

- Доверительный интервал для σ^2 при известном μ:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{nS_0^2}{\sigma^2}\sim\chi_n^2,\quad \text{где } S_o^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

- Доверительный интервал для σ^2 при неизвестном μ:

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

Доверительные интервалы в нормальной модели

После разрешения неравенства относительно неизвестного параметра внутри вероятности, получим следующий ответ.

- Доверительный интервал для μ при известном σ^2:

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{c_{1-\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

  где $c_{1-\alpha/2}$ — квантиль распределения $\mathcal{N}(0, 1)$.

- Доверительный интервал для μ при неизвестном σ^2:

$$\mathbb{P}\left(\overline{X} - \frac{c_{1-\alpha/2}S}{\sqrt{n}} < \mu < \overline{X} + \frac{c_{1-\alpha/2}S}{\sqrt{n}}\right) = 1 - \alpha,$$

  где $c_{1-\alpha/2}$ — квантиль распределения t_{n-1}.

Доверительные интервалы в нормальной модели

- Доверительный интервал для σ^2 при известном μ:

$$\mathbb{P}\left(\frac{nS_o^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{nS_o^2}{c_{\alpha/2}}\right) = 1-\alpha,$$

где $c_{\alpha/2}$ и $c_{1-\alpha/2}$ — квантили распределения χ^2_n.

- Доверительный интервал для σ^2 при неизвестном μ:

$$\mathbb{P}\left(\frac{(n-1)S^2}{c_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{c_{\alpha/2}}\right) = 1-\alpha,$$

где $c_{\alpha/2}$ и $c_{1-\alpha/2}$ — квантили распределения χ^2_{n-1}.

Бутстрэп

Бутстрэп — это набор практических методов, который основан на многократной генерации выборок на базе одной имеющейся выборки.

Бутстрэп используется для оценки каких-то параметров распределений, построения доверительных интервалов и т.д.

Рассмотрим параметрический и непараметрический бутстрэп.

Бутстрэп

Параметрический бутстрэп:

- Делается предположение, что данные получены из некоторого параметрического семейства F_θ.
- Новые выборки генерируются из закона $F_{\widehat{\theta}}$, где $\widehat{\theta}$ — некоторая оценка неизвестного параметра θ.
- Если семейство распределений F_θ непрерывно зависит от параметра и оценка $\widehat{\theta}$ не сильно уклонилась от истинного значения, то $F_{\widehat{\theta}}$ будет близко к закону, из которого получена выборка.
- Новые выборки используем для оценки того, что нужно.

Бутстрэп

Непараметрический бутстрэп:

- Никакого предположения относительно семейства распределений F_θ не делается.
- Новые выборки генерируются с помощью выбора с возвращением из исходной выборки.
- У этой идеи есть теоретическое подспорье: мы тем самым генерируем новую выборку из эмпирической функции распределения, которая является хорошим приближением истинной функции распределения.
- Новые выборки используем для оценки того, что нужно.

Бутстрэп

Теперь подробнее о теоретическом обосновании непараметрического бутстрэпа.

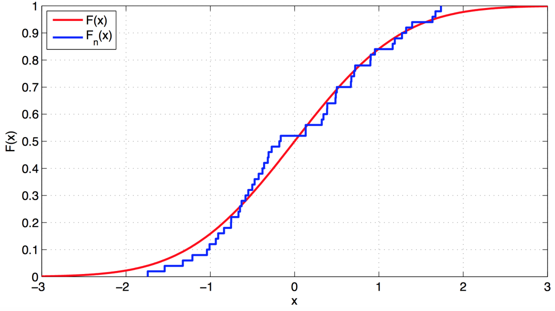
Эмпирическая функция распределения $\widehat{F}_n(u)$ определяется формулой

$$\widehat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{x_i \le u\}},$$

где $\mathbf{I}_{\{x_i \le u\}}$ — индикатор события $\{x_i \le u\}$.

Бутстрэп

График $\widehat{F}_n(x)$ представляет собой ступенчатую функцию, растущую скачками высоты $1/n$. Скачки происходят в точках с координатами $x_1, \ldots, x_n$.



Бутстрэп

Известно, что эмпирическая функция распределения является очень хорошим приближением для истинной функции распределения.

Следовательно, чтобы сгенерировать бустрэп-выборку, можно использовать закон, соответствующий эмпирической функции распределении. А это и будет выбором с возвращением.

Бутстрэп

Как строить доверительные интервалы с помощью бутстрэпа?

Существует и несколько методов построения доверительных интервалов. Наиболее простой из них — pivotal интервал.

Его идея заключается в том, чтобы посчитать некоторую характеристику, доверительный интервал для которой мы хотим построить, много раз на основе бутрэп-выборок и затем по отрезать выборочные квантили.

Бутстрэп

Пример. Допустим мы хотим построить доверительный интервал для некоторого параметра θ параметрического распределения F_θ на основе одной выборки $x_1, \dots, x_n$.

- Сгенерируем m новых бутстрэп-выборок.
  (Для параметрического бутстрэпа построим оценку $\widehat{\theta}$ параметра θ и будем генерировать из $F_{\widehat{\theta}}$, а для непараметрического бутстрэпа будем выбирать с возвращением из $x_1, \dots, x_n$.)
  На их основе мы посчитаем m новых оценок $\widehat{\theta}_1, \dots, \widehat{\theta}_m$.
- Упорядочим $\widehat{\theta}_i$ и выберем те из них, $\widehat{\theta}_-$ и $\widehat{\theta}_+$, которые стоят на местах $[(\alpha/2)m]$ и $[(1-\alpha/2)m]$ по возрастанию;
- В качестве доверительного интервала возьмем

$$\left(\widehat{\theta}_-\,,\,\widehat{\theta}_+\right).$$

Бутстрэп

Плюсы и минусы бустрэпа.

Бутстрэп прост в использовании, не требует сложных вычислений и применим даже к весьма громоздким моделям.

С другой стороны, мы не можем явным образом оценить его погрешность. В случае, если оценка $\widehat{\theta}$ значимо промахнулась мимо θ_0 или эмпирическая функция распределения $\widehat{F}_n$ сильно отличается от истинной F, мы рискуем сильно ошибиться в выводах на основе бутстрэп-выборок.

Проверка гипотез

В проверке гипотез мы делаем предположение о процессе, генерирующем данные, и наша задача состоит в том, чтобы определить, содержат ли данные достаточно информации, чтобы отвергнуть это предположение или нет.

Чтобы иметь возможность отвергнуть предположение, нам необходимо зафиксировать альтернативу — другое предположение о данных, относительно которого мы будем решать, отвергать основную гипотезу или нет.

Проверка гипотез

Пример

Предположим, что кто-то подбросил 10 раз монетку, и в 8 случаях она упала гербом вверх. Можно ли считать эту монетку симметричной?

Пусть $X_1, \ldots, X_n \sim \mathbf{B_p}$.

$H_0: \; p = \frac{1}{2}$ (основная гипотеза).
$H_1: \; p \neq \frac{1}{2}$ (альтернативная гипотеза).

Как проверить гипотезу H_0 о том, что $p = 1/2$?

Проверка гипотез

Правило, позволяющее принять или отвергнуть гипотезу H_0 на основе данных называется статистическим критерием.

Обычно критерий задается при помощи статистики критерия $T(x_1, \ldots, x_n)$ такой, что для нее типично принимать умеренные значения в случае, когда гипотеза H_0 верна, и большие (иногда малые) значения, когда H_0 не выполняется.

Проверка гипотез

Статистика критерия T должна обладать важным свойством:

- при верной H_0 статистика T должна иметь известное нам распределение G_0;
- при неверной H_0 должна иметь какое-либо распределение отличное от G_0.

Проверка гипотез

В нашем примере в качестве статистики T можно взять

$$T(x_1, \ldots, x_n) = x_1 + \ldots + x_n.$$

Тогда гипотезе $H_0: \ p = 1/2$ противоречат значения, которые близки к 0 или n.

Более того,

- при верной H_0 имеет биномиальное распределение $\mathbf{B_{n,1/2}}$;
- при верной H_1 имеет биномиальное распределение $\mathbf{B_{n,p}}$, но с $p \neq 1/2$.

Проверка гипотез

Если значение T попало в область, имеющую при выполнении гипотезы H_0 малую вероятность, то можно заключить, что данные противоречат гипотезе H_0 в пользу альтернативы H_1.

Если значение T попало в область, имеющую при выполнении гипотезы H_0 большу́ю вероятность, то можно заключить, что данные не свидетельствуют против гипотезы H_0 в пользу альтернативы H_1.

Проверка гипотез

Формализация задачи:

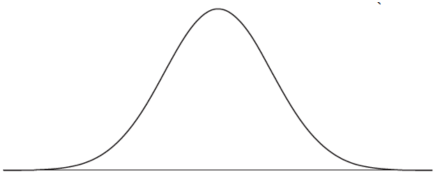
выборка: $\mathbf{X} = (X_1, \ldots, X_n),\ X_i \sim F$

нулевая гипотеза: $H_0 :\ F \in \mathcal{F}_0$

альтернатива: $H_1 :\ F \in \mathcal{F}_1,\ \mathcal{F}_1 \cap \mathcal{F}_0 = \varnothing$

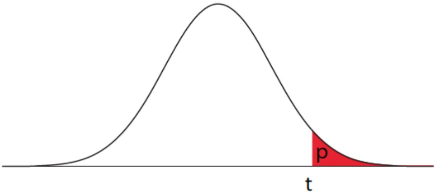
статистика: $T(x_1, \ldots, x_n),\ \ T(\mathbf{X}) \sim G_0$ при H_0

$T(\mathbf{X}) \nsim G_0$ при H_1



Проверка гипотез

реализация выборки:	$\mathbf{x} = (x_1, \dots, x_n)$
реализация статистики:	$t = T(\mathbf{x})$
достигаемый уровень значимости или p-value:	$p(\mathbf{x}) = \mathbb{P}(T(\mathbf{X}) \geq t \mid H_0)$ (если для T экстремальные значения — большие)



Проверка гипотез

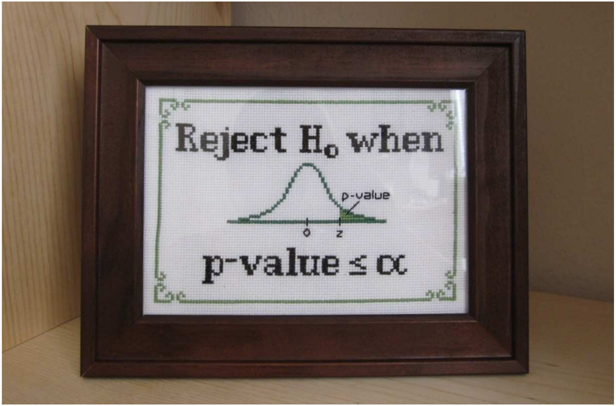
Достигаемый/Фактический уровень значимости (p-value) — это вероятность для статистики T при верной H_0 принять значение t или ещё более экстремальное.

Если для для статистики T экстремальными значениями являются большие значения, то это можно записать так:

$$p(\mathbf{x}) = \mathbb{P}(T(\mathbf{X}) \geq t \mid H_0).$$

Нулевая гипотеза H_0 отвергается при $p(\mathbf{x}) \leq \alpha$, α — уровень значимости, который мы задаем.

Проверка гипотез



Спасибо за внимание!