Прикладная статистика. Проверка гипотез. Введение

Леонид Иосипой

Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

1 декабря 2020

- Повторение
- Проверка гипотез

Повторение

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

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

Повторение

Если вероятность в левой части неравенства в пределе не превосходит $1 - \alpha$ при $n \to \infty$, то есть выполняется

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

то доверительный интервал называется асимптотическим.

Повторение

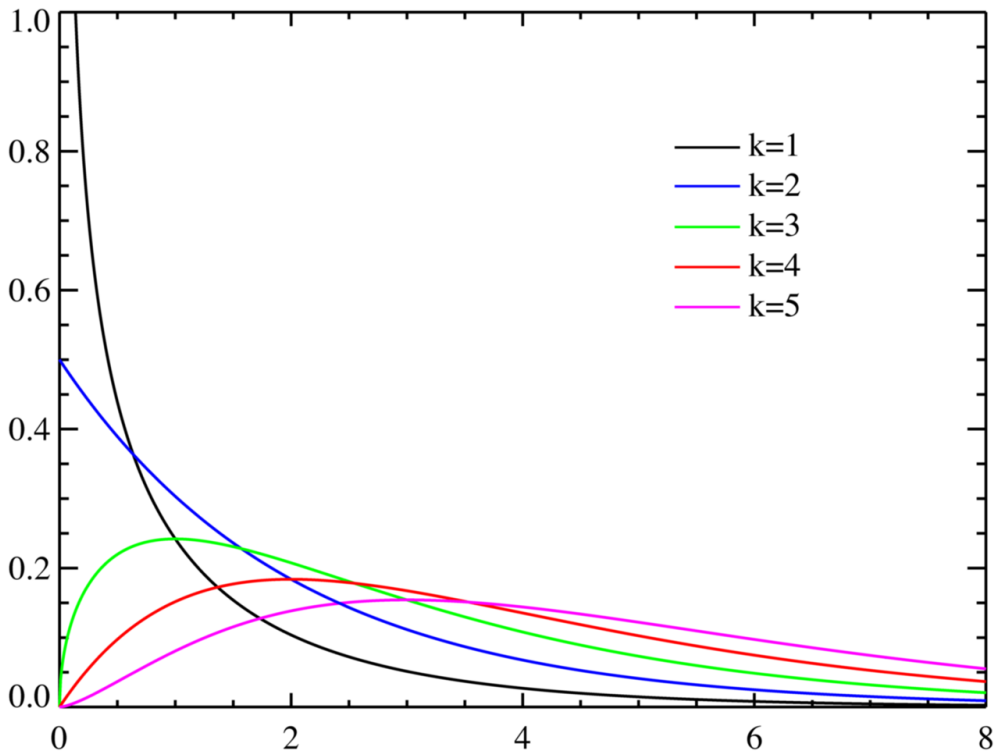
Пусть $X_1, \ldots, X_k$ независимы и имеют стандартное нормальное распределение $\mathcal{N}(0, 1)$.

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 или H_k.

Повторение



Повторение

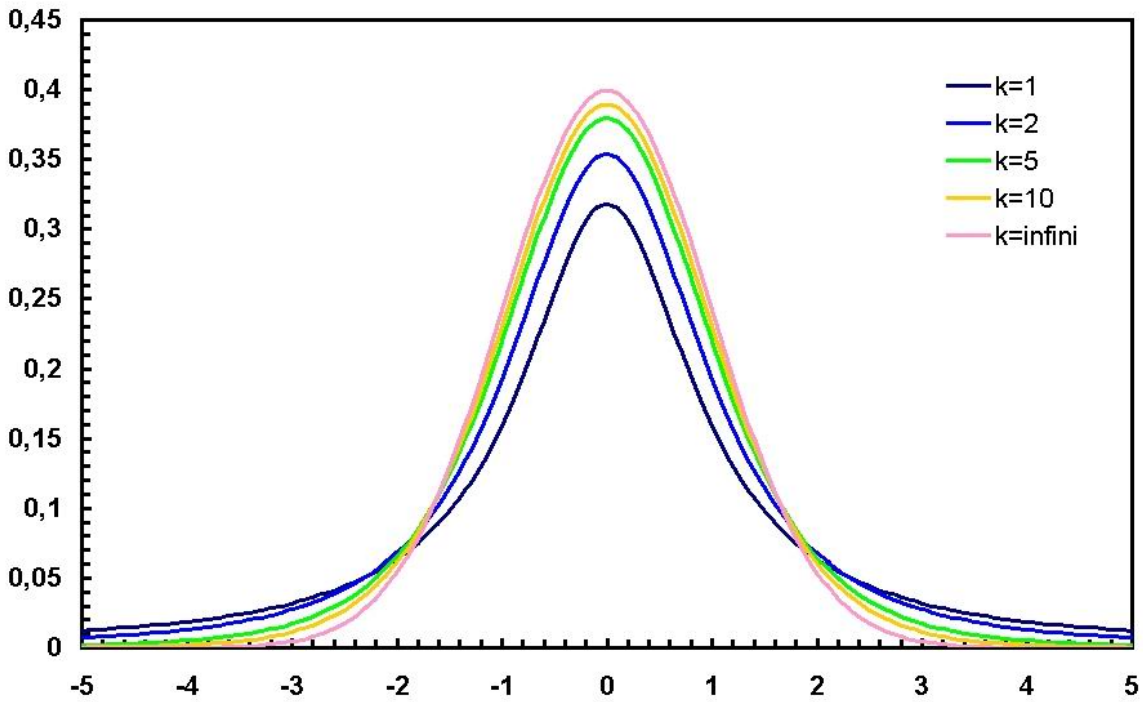
Пусть $X_0, X_1, \ldots, X_k$ независимы и имеют стандартное нормальное распределение $\mathcal{N}(0, 1)$.

Распределением Стьюдента называется распределение случайной величины

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

Обозначение: T_k.

Повторение


0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
k=1
k=2
k=5
k=10
k=infini

Повторение

Доверительные интервалы в нормальной модели.
Пусть $X_1, \ldots, X_n$ — выборка из $\mathcal{N}(\mu, \sigma^2)$.

- 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 σ^2 при известном μ:

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

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

Повторение

- 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}.

- доверительный интервал для σ^2 при неизвестном μ^2:

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

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

Повторение

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

Повторение

Параметрический бутстрэп:

- Данные пришли из некоторого параметрического семейства F_θ.
- Чтобы сгенерировать новые выборки необходимо найти оценку $\widehat{\theta}$ и генерировать из $F_{\widehat{\theta}}$.
- 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.

Повторение

Непараметрический бутстрэп:

- Не делаем предположения относительно какого-либо «семейства» распределений F_θ.
- Чтобы сгенерировать новые выборки используем выбор 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.

Повторение

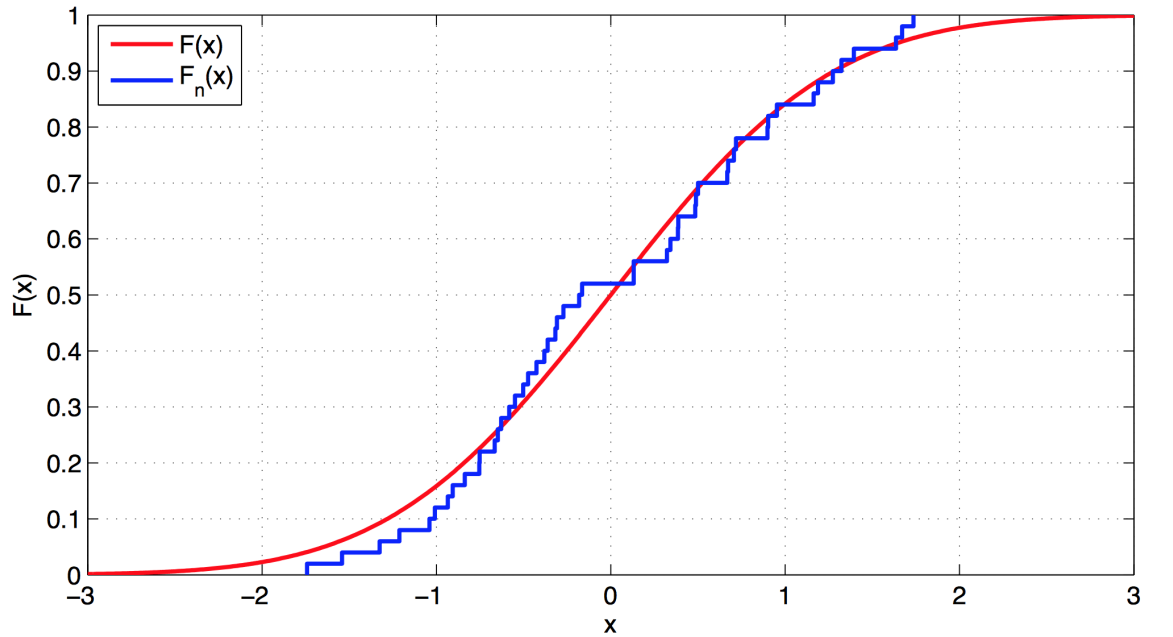
Эмпирическая функция распределения $\widehat{F}_n(u)$ определяется формулой

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

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

Повторение

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



Повторение

Доверительный интервал на основе бутстрэпа:

- Сгенерируем m выборок с помощью параметрического или непараметрического бутстрэпа.
- Посчитаем m раз величину, доверительный интервал для которой мы хотим построить. Обозначим эти «оценки» через $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$.
- 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.
- Тогда нашим интервалом будет:

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

Повторение

Задача. Пусть имеется реализация выборки $x_1, \ldots, x_n$ из равномерного распределения на $[0, \theta]$.

Допустим мы оценили θ с помощью $2\overline{x}$. А затем берем новую выборку из равномерного распределения на $[0, 2\overline{x}]$ и оцениваем с помощью ее среднего параметр θ.

Какую дисперсию будет иметь эта новая оценка?

Проверка гипотез

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

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

Проверка гипотез

Задача. 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 (малые) значения, когда H_0 не выполняется.

Проверка гипотез

Для нашего эксперимента в качестве статистики T можно взять сумму:

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

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

Проверка гипотез

Статистика критерия T должна обладать важным свойством:

- при верной H_0 статистика T должна иметь известное нам распределение F_0;
- при верной H_1 должна иметь какое-либо распределение отличное от F_0.

Проверка гипотез

В нашем примере это свойство выполняется: статистика

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

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

Если произошло обратное, то есть значение T попало в область, имеющую при выполнении гипотезы H_0 большую вероятность, то можно заключить, что данные не противоречат гипотезе H_0.

Вероятности можно посчитать, так как нам известно распределение F_0!

Проверка гипотез

Формализация задачи: в случае простых гипотез

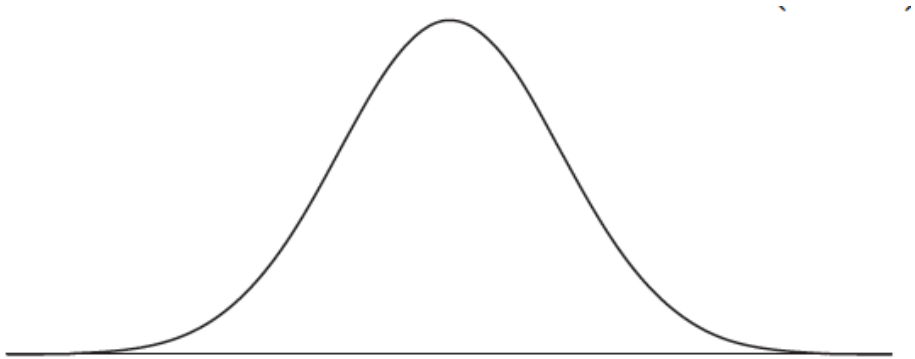
выборка: $\mathbf{X} = (X_1, \ldots, X_n)$

нулевая гипотеза: $H_0 : \; X_i \sim G_0$

альтернатива: $H_1 : \; X_i \sim G_1 \neq G_0$

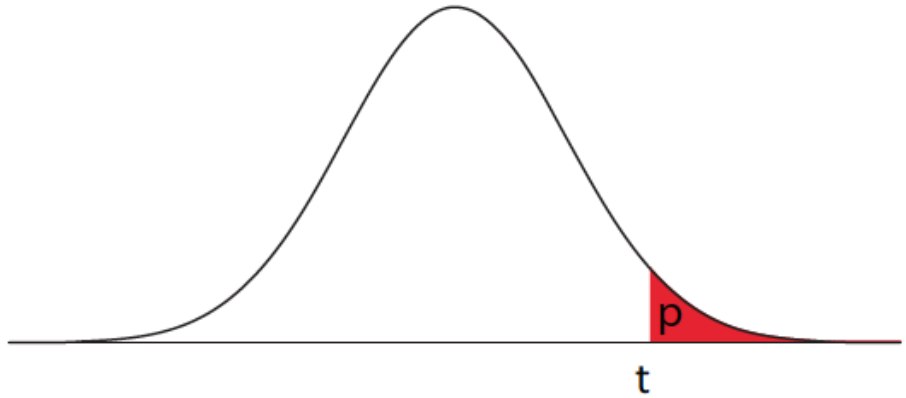
статистика: $T(x_1, \ldots, x_n), \;\; T(\mathbf{X}) \sim F_0$ при $\mathbf{X} \sim G_0$

$T(\mathbf{X}) \nsim F_0$ при $\mathbf{X} \sim G_1$



Проверка гипотез

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



Проверка гипотез

Достигаемый уровень значимости или p-value:

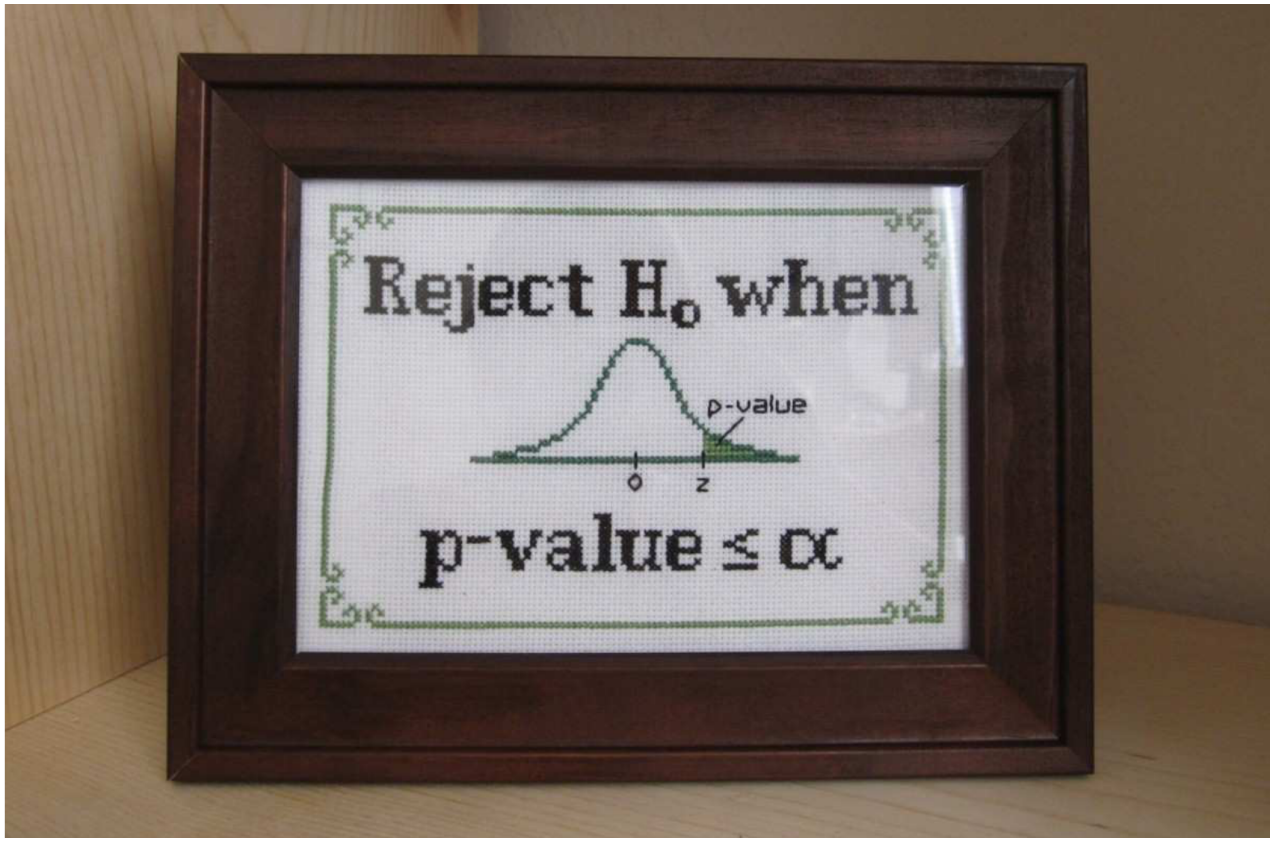
$$p(\mathbf{x}) = \mathbb{P}(\, T(\mathbf{X}) \geq t \mid H_0)$$

$p(\mathbf{x})$ — вероятность для статистики T при верной H_0 получить значение t или ещё более экстремальное.

Здесь для удобства мы считали, что экстремальными значениями для статистики T являются большие значения. Так бывает не всегда.

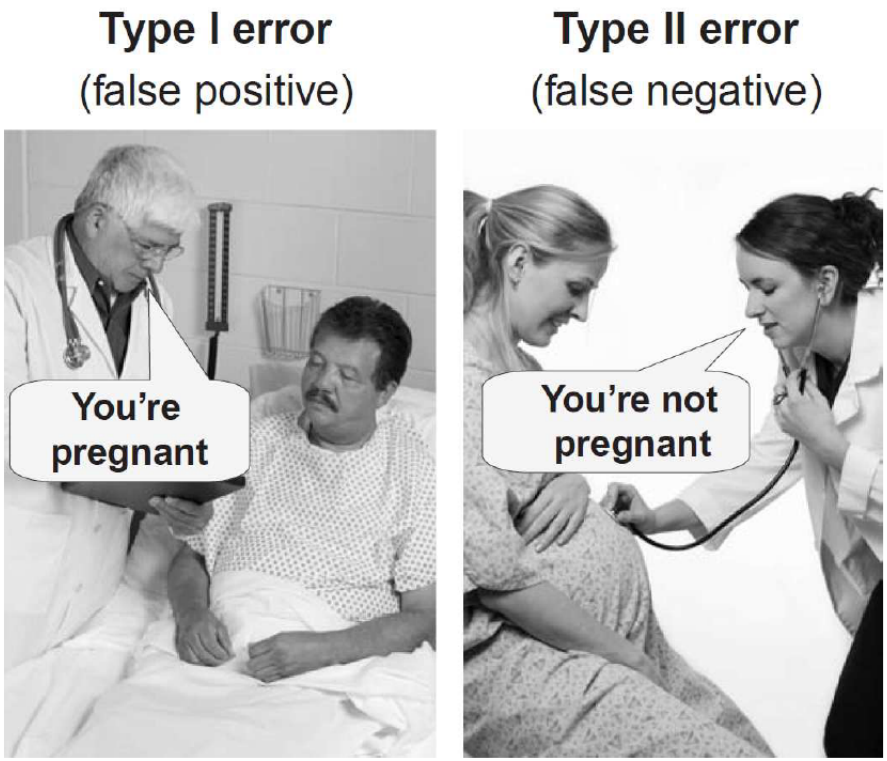
Гипотеза отвергается при $p(\mathbf{x}) \leq \alpha$, α — уровень значимости.

Проверка гипотез



Проверка гипотез

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода (False negative)
H_0 отвергается	Ошибка первого рода (False positive)	H_0 верно отвергнута



Проверка гипотез

Если величина p-value достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Если величина p-value недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы!

Проверка гипотез

Вероятность отвергнуть нулевую гипотезу зависит не только от того, насколько она отличается от истины, но и от размера выборки.

По мере увеличения n нулевая гипотеза может сначала приниматься, но потом выявятся более тонкие несоответствия выборки гипотезе H_0, и она будет отвергнута.

Пример проверка гипотез

Задача. Джеймс Бонд говорит, что предпочитает мартини взболтанным, но не смешанным. Давайте проверим, так это или нет.

Проведём слепой тест: n раз предложим ему пару напитков и выясним, какой из двух он предпочитает.

Пример проверка гипотез

Выборка: $\mathbf{X} = (X_1, \ldots, X_n)$, где $X_i \sim \mathbf{B_p}$.

Реализация выборки: $\mathbf{x} = (x_1, \ldots, x_n)$ — это бинарный вектор длины n, где

- 1 — Джеймс Бонд выбрал взболтанный мартини
- 0 — Джеймс Бонд выбрал смешанный мартини

H_0: Д.Б. не различает два вида мартини, $p = 1/2$.

H_1: Д.Б. предпочитает взболтанный мартини, $p > 1/2$.

Пример проверка гипотез

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

Реализация статистики: $t = T(\mathbf{x})$.

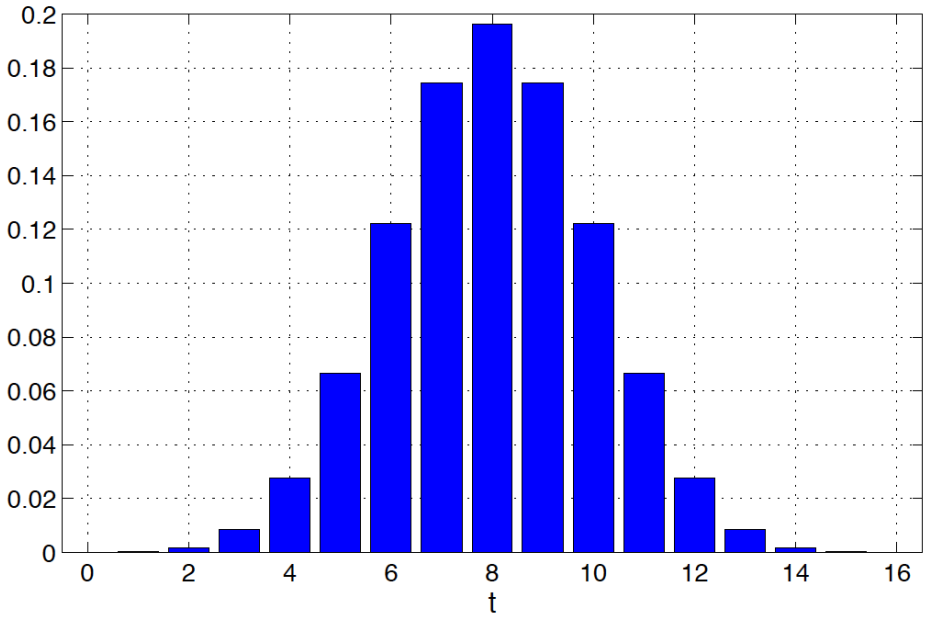
Какие значения T считаются экстремальными?

При альтернативе H_1 экстремальными являются большие значения t (они свидетельствуют против H_0 в пользу H_1).

Пример проверка гипотез

Если H_0 справедлива и Джеймс Бонд не различает два вида мартини, то T будет иметь распределение $\mathbf{B}_{n,1/2}$.

Пусть $n = 16$, тогда $\mathbf{B}_{n,1/2}$ будет иметь следующий вид:

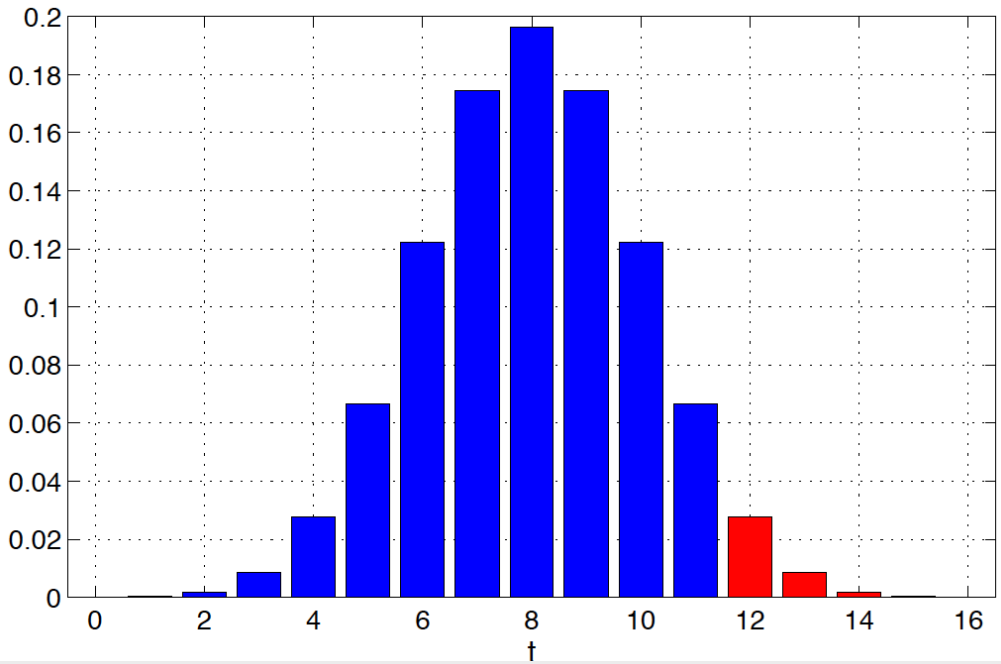


Пример проверка гипотез

Допустим, что $t = 12$, то есть в 12 случаях из 16 Джеймс Бонд выбрал взболтанный мартини.

Тогда достигаемый уровень значимости p-value равен:

$$\mathbb{P}(T(\mathbf{X}) \geq 12 | H_0) = \frac{2517}{65536} \approx 0.0384.$$



Пример проверка гипотез

Давайте поменяем альтернативу.

H_1: Джеймс Бонд предпочитает какой-то определённый вид мартини, но неизвестно какой, то есть $p \neq 1/2$.

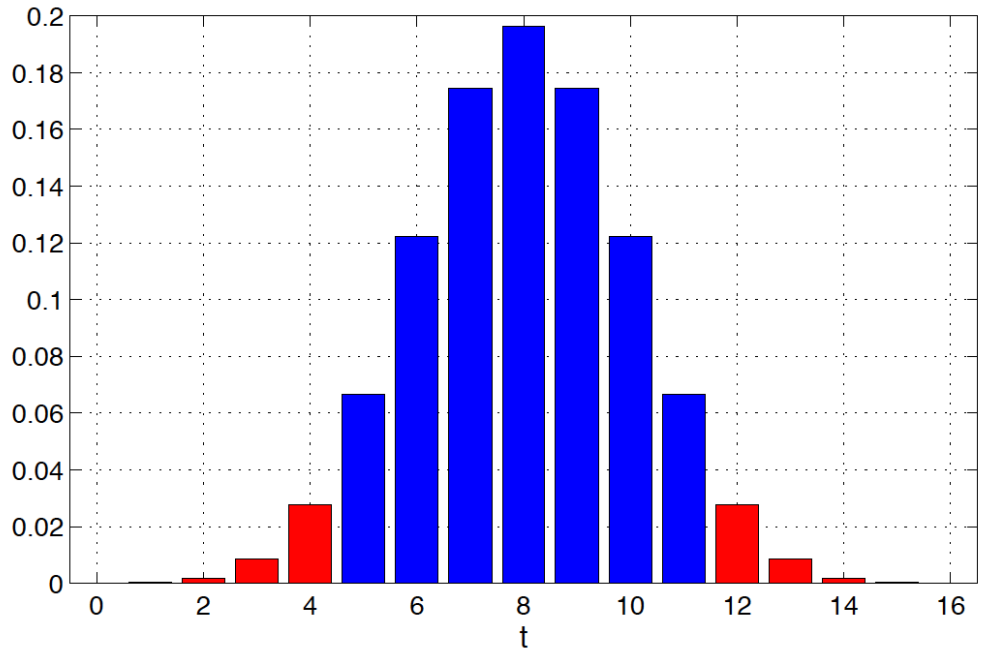
При такой альтернативе и большие, и маленькие значения t свидетельствуют против H_0 в пользу H_1.

Пример проверка гипотез

Допустим, что $t = 12$, то есть в 12 случаях из 16 Джеймс Бонд выбрал взболтанный мартини.

Тогда достигаемый уровень значимости p-value равен:

$$\mathbb{P}(T(\mathbf{X}) \geq 12 \text{ или } T(\mathbf{X}) \leq 4 | H_0) = \frac{5034}{65536} \approx 0.0768.$$



Пример проверка гипотез

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Достигаемый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

Спасибо за внимание!