Проверка гипотез. Критерии согласия и однородности.

Леонид Иосипой

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

22 апреля 2021

- Повторение
- Проверка гипотез
- Критерии согласия
- Квантильный график

Повторение

1. Распределения, связанные с нормальным.

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

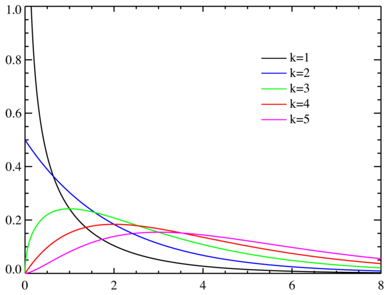
   Распределением χ^2 (хи-квадрат) с k степенями свободы называется распределение случайной величины

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

   Обозначение: χ_k^2.

Повторение

1. Распределения, связанные с нормальным.



Повторение

1. Распределения, связанные с нормальным.

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

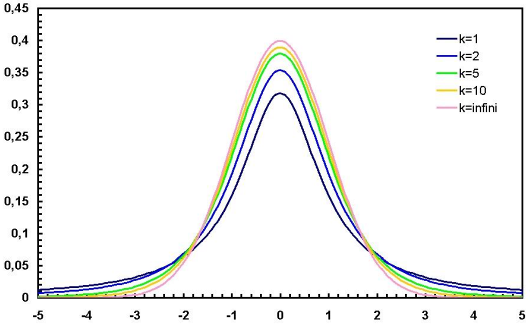
   Распределением Стьюдента с k степенями свободы называется распределение случайной величины

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

   Обозначение: t_k.

Повторение

1. Распределения, связанные с нормальным.



Повторение

2. Доверительные интервалы в нормальной модели.

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

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

- 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 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}.

Повторение

2. 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}.

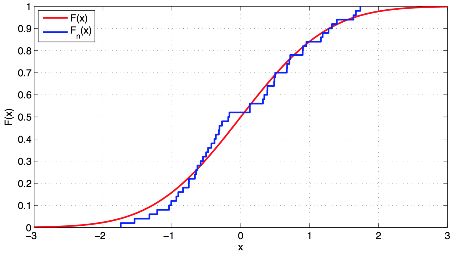
Повторение

3. Эмпирическая функция распределения.

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

4. 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.

Повторение

4. 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.

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 бо́льшую вероятность, то можно заключить, что данные не свидетельствуют против гипотезы H_0 в пользу альтернативы H_1.

Проверка гипотез

Формализация задачи:

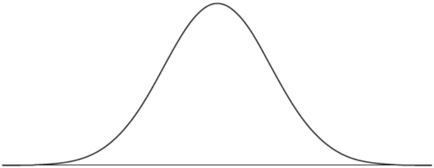
выборка: $\mathbf{X} = (X_1, \ldots, X_n)$, $X_i \sim F$

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



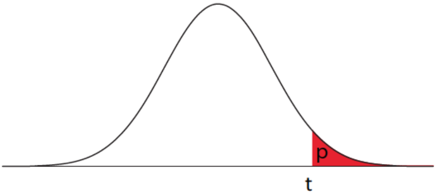
Проверка гипотез

реализация выборки: $\mathbf{x} = (x_1, \ldots, x_n)$

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

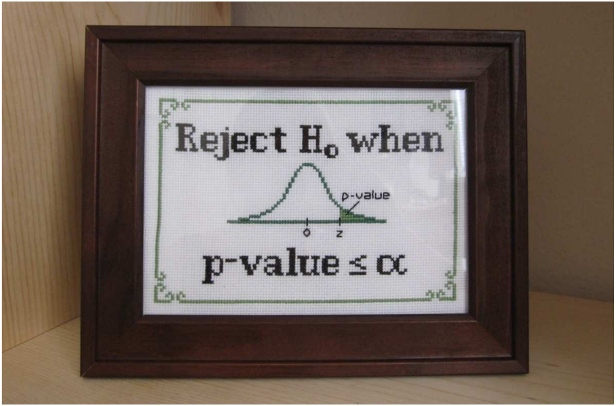
Достигаемый/Фактический уровень значимости (p-value) — это вероятность для статистики T при верной H_0 принять значение t или ещё более экстремальное.

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

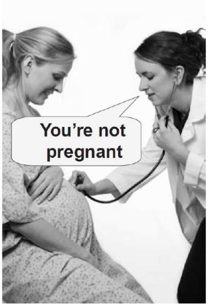


Проверка гипотез

	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

Type I error
(false positive)

Type II error
(false negative)



Проверка гипотез

Если величина p-value достаточно мала, то данные свидетельствуют против нулевой гипотезы H_0 в пользу альтернативы H_1.

Если величина p-value недостаточно мала, то данные не свидетельствуют против нулевой гипотезы H_0 в пользу альтернативы H_1.

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.

Проверка гипотез

Задача

Джеймс Бонд говорит, что предпочитает взболтанный мартини, но не смешанный. Проверим, так это или нет.

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

Проверка гипотез

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

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

- 0 — Джеймс Бонд выбрал смешанный мартини;
- 1 — Джеймс Бонд выбрал несмешанный мартини.

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

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

Проверка гипотез

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

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

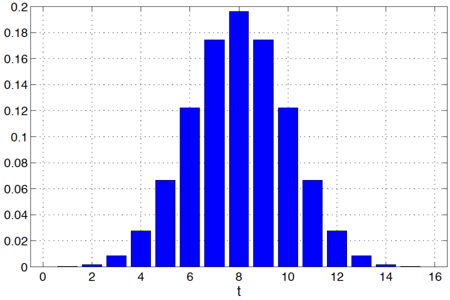
Какие значения T считаются экстремальными?

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

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:

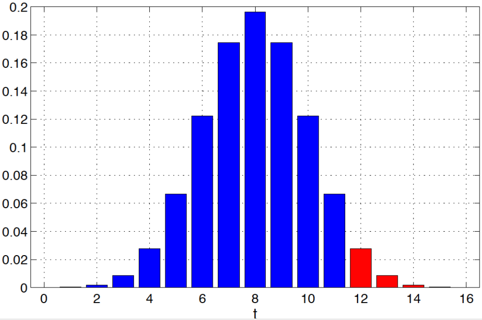


Проверка гипотез

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

Тогда достигаемый уровень значимости p-value равен:

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



Проверка гипотез

Давайте поменяем альтернативу.

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

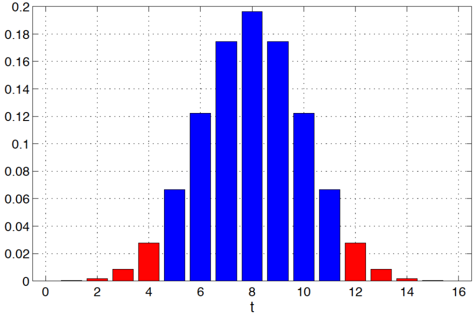
При такой альтернативе и большие, и маленькие значения t свидетельствуют против H_0 в пользу H_1.

Проверка гипотез

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

Тогда достигаемый уровень значимости p-value равен:

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



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!

Критерии согласия

Пусть у нас есть выборка $X_1, \ldots, X_n \sim F$, где F — некоторое неизвестное распределение.

Начнем изучение критериев с критериев согласия, в которых в качестве H_0 будем рассматривать $F \in \mathcal{F}_\theta$, то есть принадлежность F какому-то параметрическому семейству.

Альтернативой H_1 мы будем считать принадлежность F всем остальным распределениям.

Критерии согласия

Критерии согласия так называются, потому что они отвечают на вопрос, согласуется ли наша выборка с каким-то параметрическим семейством или нет.

В англоязычной литературе такие критерии называют Goodness of Fit Tests.

Критерии согласия

Для построения критерия согласия достаточно найти некоторое свойство, которые бы выполнялось для всех распределений из нашего класса и на его основе придумать статистику.

При этом сколько-то удовлетворительно мажорировать вероятность ошибки второго рода не удается, поскольку вне нашего параметрического семейства есть сколь угодно похожие на наши распределения.

Но по крайней мере, можно искать критерий, от которого мы ожидаем, что при верной альтернативе он будет чаще отвергать нулевую гипотезу.

Критерии согласия

Мы будем говорить, что произвольная гипотеза H является простой, если $H : F = F_0$, то есть гипотеза состоит из равенства одному распределению.

В противном случае мы будем называть гипотезу сложной.

Рассмотрим сперва проверку простой гипотезы.

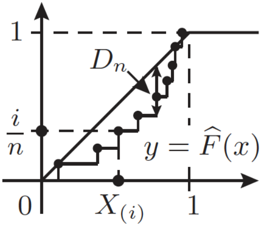
Критерии согласия

1. Критерий Колмогорова.

Критерий Колмогорова базируется на эмпирической функции распределения $\widehat{F}_n$ и ее отклонении от F_0.

Статистика критерия основана на величине

$$D_n = \sup_{u \in \mathbb{R}} \left| \widehat{F}_n(u) - F_0(u) \right|.$$



Критерии согласия

1. Критерий Колмогорова.

Для выборки достаточно большого размера, при верной H_0, значение D_n не должно существенно отклоняться от 0.

Теорема (Гливенко-Кантелли)

Пусть F_0 — функция распределения элементов выборки. Тогда статистика D_n стремится к 0 с вероятностью 1.

Критерии согласия

1. Критерий Колмогорова.

Как количественно охарактеризовать значимость отклонения D_n от нуля на конкретных данных?

Теорема (Колмогоров)

Пусть F_0 — функция распределения элементов выборки. Если F_0 непрерывна, то для любого $t > 0$, при $n \to \infty$,

$$\mathbb{P}(\sqrt{n}D_n \leq t) \to K(t) := 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2t^2}.$$

$K(t)$ называется функцией Колмогорова, а соответствующее распределение — распределением Колмогорова.

Критерии согласия

1. Критерий Колмогорова.

Быстрая сходимость к предельному закону позволяет пользоваться этим приближением уже при $n \geq 20$.

Условие непрерывности функции распределения необходимо. Например, в схеме Бернулли статистика $\sqrt{n}D_n$ имеет другой предельный закон распределения.

Повторение

1. Критерий Колмогорова

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

$X_i \sim F$, $\underline{F \text{ непрерывно}}$

нулевая гипотеза: $H_0 : \; F = F_0$

альтернатива: $H_1 : \; F \neq F_0$

статистика: $\sqrt{n} D_n = \sqrt{n} \cdot \sup_{u \in \mathbb{R}} \left| \widehat{F}_n(u) - F_0(u) \right|$

нулевое распределение: $\sqrt{n} D_n \sim K$ – распределение Колмогорова

Критерии согласия

2. Критерий Пирсона (хи-квадрат).

Критерий Пирсона (критерий хи-квадрат) основан уже на другой статистике — частотах.

Этот критерий можно использовать для проверки простой гипотезы о равенстве распределения в дискретном случае. (Существует и обобщение критерия хи-квадрат на непрерывный случай, но мы его рассматривать не будем.)

Критерии согласия

2. Критерий Пирсона (хи-квадрат).

Пусть нам дана выборка $X_1, \ldots, X_n$ из дискретного закона

X	a_1	a_2	$\ldots$	a_k
$\mathbb{P}$	p_1	p_2	$\ldots$	p_k

Статистикой критерия является величина

$$T_n = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i},$$

где ν_i — количество значений a_i в реализации $x_1, \ldots, x_n$.

Критерии согласия

2. Критерий Пирсона (хи-квадрат).

Как количественно охарактеризовать значимость отклонения T_n от нуля на конкретных данных?

Теорема (Пирсон)

Пусть реализация $x_1, \dots, x_n$ получена из закона X. Тогда, при $n \to \infty$, распределение статистики T_n сходится к закону χ^2_{k-1}.

Приближение распределения статистики T_n с помощью закона χ^2_{k-1} является достаточно точным при $n \geq 50$ и $np_i \geq 5$ для всех $i = 1, \dots, k$.

Критерии согласия

2. Критерий Пирсона (хи-квадрат).

Что делать если количество возможных значений X счетно?

В этом случае необходимо «сгруппировать» значения, которые принимаются с малыми вероятностями (причем так, чтобы получилось $np_i \geq 5$ для всех $i = 1, \ldots, k$).

Повторение

2. Критерий Пирсона (хи-квадрат)

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

$X_i \sim F$, F дискретно

нулевая гипотеза: $H_0: \; F = F_0$

альтернатива: $H_1: \; F \neq F_0$

статистика: $T_n = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$

нулевое распределение: $T_n \sim \chi^2_{k-1}$ — хи-квадрат с $k-1$ степенью свободы (k — количество возможных значений)

Критерии согласия

Перейдем теперь к сложным гипотезам.

Гораздо чаще у нас есть гипотеза о принадлежности к параметрическому семейству, например, что выборка нормальная, но с неизвестными параметрами.

Как быть в этом случае?

Критерии согласия

Можно оценить неизвестные параметры состоятельными оценками, но эта процедура может сместить распределение статистик критерия.

Например, в случае с нормальным распределением, оценить среднее и дисперсию с помощью оценок максимального правдоподобия и применить критерий Колмогорова.

Однако в этом случае предельным распределением уже будет распределение Лиллиефорса, а не Колмогорова.

Это замечание крайне важно и зачастую игнорируется малоопытными аналитиками!

Критерии согласия

Рассматривать критерии с подстановкой состоятельных оценок мы не будем, информация о них будет в дополнительном задании к лекции.

Вместо этого мы рассмотрим довольно мощные специализированные критерии для некоторых конкретных семейств распределений.

Критерии согласия

1. Проверка экспоненциальности (показательности)

Под гипотезой экспоненциальности понимается сложная гипотеза

$$H_0 : F \in \{F_\lambda\}_{\lambda>0},$$

где класс $\{F_\lambda\}_{\lambda>0}$ образуют функции распределения экспоненциального закона, то есть

$$F_\lambda(u) = (1 - e^{-\lambda u})\mathbf{I}_{\{u\geq 0\}}.$$

Критерии согласия

1. Проверка экспоненциальности (показательности)

(a) Исключение неизвестного параметра

- Положим $S_k = X_1 + \ldots + X_k$, $k = 1, \ldots, n$.
- Можно доказать, что для экспоненциального распределения случайный вектор $(S_1/S_n, \ldots, S_{n-1}/S_n)$, распределен так же, как и упорядоченный ряд из равномерного распределения на $[0, 1]$ размера $n - 1$.
- Данное преобразование сводит задачу к проверке равномерности, которую можно проверить с помощью критерия Колмогорова. Однако, за исключение «мешающего» параметра λ приходится платить уменьшением размера выборки на 1.

Критерии согласия

1. Проверка экспоненциальности (показательности)

(б) Критерий Гини (Gini)

Этот критерий базируется на статистике

$$G_n = \frac{\sum_{i=1}^{n} X_{(i)}(2i - n - 1)}{n(n-1)\overline{X}},$$

которая при нормировке $12(n-1)(G-0.5)$ сходится к нормальному распределению. Здесь и далее $X_{(i)}$ — это i-ый элемент в упорядоченной по возрастанию выборке.

Критерии согласия

1. Проверка экспоненциальности (показательности)

Для проверки экспоненциальности существует и ряд других критериев (например, Шапиро-Уилка для экспоненциального случая или Андерсона-Дарлинга).

Другие критерии могут быть основаны на других идеях.

Критерии согласия

2. Проверка нормальности

Под гипотезой нормальности понимается сложная гипотеза

$$H_0 : F \in \{F_{\mu,\sigma}\}_{\mu\in\mathbb{R},\, \sigma>0},$$

где класс $\{F_{\mu,\sigma}\}_{\mu\in\mathbb{R},\, \sigma>0}$ образуют функции распределения нормального закона. Напомним, что по определению $Y \sim \mathcal{N}(\mu,\sigma)$, если $Y = \mu + \sigma X, \quad X \sim \mathcal{N}(0,1)$. Поэтому можно записать, что

$$F_{\mu,\sigma}(u) = \Phi\left(\frac{u-\mu}{\sigma}\right),$$

где Φ — функция распределения $\mathcal{N}(0,1)$.

Критерии согласия

2. Проверка нормальности

(a) Критерий Шапиро-Уилка (Shapiro-Wilk)

Этот критерий базируется на статистике

$$SW_n = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2},$$

где a_i — некоторые константы.

Этот тест показывает очень хорошие результаты даже на небольших выборках.

Критерии согласия

2. Проверка нормальности

(б) Критерий Харке-Бера (Jarque-Bera)

Этот критерий использует статистику

$$JB_n = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right), \quad S = \frac{\mu_3}{\mu_2^{3/2}}, \quad K = \frac{\mu_4}{\mu_2^2},$$

где μ_k — k-ый центрированный выборочный момент:

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

Этот критерий тоже показывает хорошие результаты на практике.

Квантильный график(Q-Q Plot)

Согласие выборки с распределением, которое образовано сдвигом/масштабом, можно проверить визуально с помощью квантильного графика (Q-Q Plot).

К таким распределениям относятся: равномерное, экспоненциальное, нормальное и т.д.

Квантильный график(Q-Q Plot)

Рассмотрим построение квантильного графика на примере нормального распределения (именно для него он чаще всего строится).

Напомним, что в этом случае, для некоторых $\mu \in \mathbb{R}$ и $\sigma > 0$,

$$F_{\mu,\sigma}(u) = \Phi\left(\frac{u-\mu}{\sigma}\right).$$

Квантильный график(Q-Q Plot)

Идея квантильного графика заключается в следующем:

- Возьмем в качестве приближения функции распределения выборки F эмпирическую функцию распределения $\widehat{F}_n$.
- Рассмотрим следующий график $y(x) = \Phi^{-1}(\widehat{F}_n(x))$. Если $F \in F_{\mu,\sigma}(u)$, то $y(x) \approx \Phi^{-1}(F(x)) = (x - \mu)/\sigma$.
- Это означает, что данный график не должен сильно отличаться от линейного.

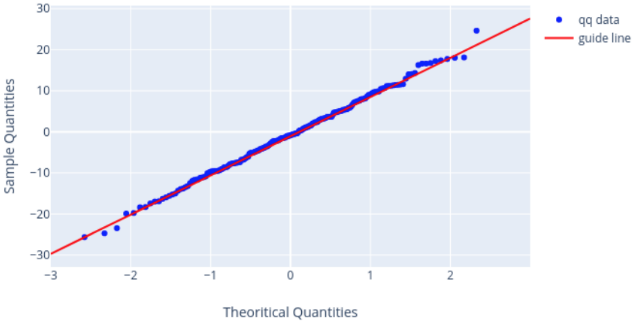
Квантильный график(Q-Q Plot)

Для реализации этого способа достаточно отметить только точки, которые соответствую «скачкам» $\widehat{F}_n$ и подогнать под это облако точек прямую.

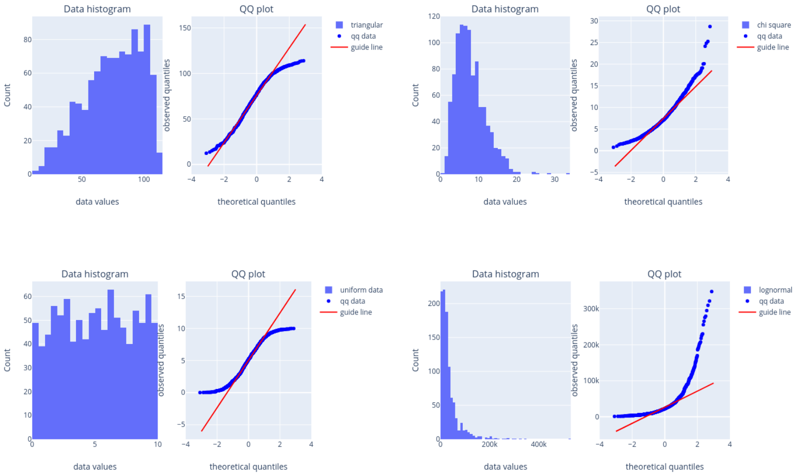
Если точки будут лежать далеко от прямой, то, скорее всего, предположение о том, что выборка взята из нормального распределения, не выполняется.

Обратите внимание, что на графике будут отложены точки $(x_{(i)}, \Phi^{-1}(i/n))$, то есть эмпирические и теоретические квантили. Поэтому график так называется.

Квантильный график(Q-Q Plot)



Квантильный график(Q-Q Plot)



Квантильный график(Q-Q Plot)

Квантильный график можно построить не только для семейства сдвига/масштаба, но и для двух выборок, чтобы визуально проверить гипотезу о том, что выборки взяты из одного и того же распределения.

Спасибо за внимание!