Доверительные интервалы. Гипотезы

Центр биоэлектрических интерфейсов, 5 декабря 2018 г.

Денис Деркач, Влад Белавин

Оглавление

Интервальные оценки
Байесовские доверительные интервалы
Доверительные интервалы
Доверительные интервалы на основе функции правдоподобия
Доверительное оценивание на основе бутстрепа
Понятие статистической гипотезы
Статистический критерий и его характеристики
Типы статистических критериев
- Критерий Вальда
- Распределение хи-квадрат (χ^2)
- Критерий перестановок
- Критерий на основе отношения правдоподобия
- Критерий Неймана-Пирсона для случая двух простых гипотез
- t-критерий

Интервальные оценки

Интервальные оценки: мотивация

- Обычно мы пытаемся измерить параметр на конечной выборке.
- Было бы интересно понять не только точечную оценку из имеющейся выборки, но и предположение о том, где лежит настоящее значение параметра (классический подход) или насколько мы уверены в полученном значении параметра (байесовкий подход).

Для этого мы вводим интервальные оценки.

Требования к интервальным оценкам

- как можно более объективно рассказать о результатах эксперимента;
- предъявить интервал, который покрывает настоящее значение параметра, с выбранной вероятностью;
- предоставить информацию, необходимую для принятия решения;
- сделать вывод о параметре, который включает в себя предыдущие знания.

NB: в случае большой выборки скорее всего достаточно будет использовать точечную оценку параметра и его стандартное отклонение.

Bayes vs. Frequentist

Как всегда, у нас возникают разные подходы к решению задачи, в зависимости от интерпретации вероятностей.

Байесовские доверительные интервалы

Байесовский доверительный интервал (credibility interval)

Определение

Байесовский p-доверительный интервал — это интервал $[L, U]$, к котором значение параметра θ принадлежит с апостериорной вероятностью p:

$$\mathbb{P}(L \leqslant \theta \leqslant U | X) = p.$$

NB: сокращение Cr.L. (часто используют C.L.).

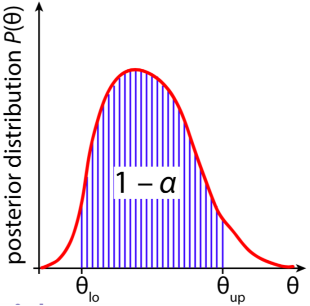
Байесовский доверительный интервал

Байесовский подход:

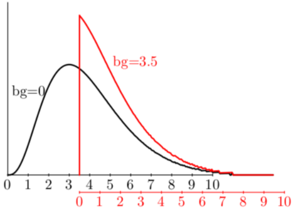
$$1 - \alpha = \int_{\theta_{\mathrm{lo}}}^{\theta_{\mathrm{up}}} p(\theta|X)d\theta$$

Подходы к выбору θ_{lo} и θ_{hi}:

- HPD (highest probability density) — брать только наиболее высокие вероятности.
- Центральный интервал - интегрировать от пика.
- Односторонний интервал — интегрировать от бесконечности.



Поведение вблизи границ



Bayesian 90% Upper Limits (Uniform Prior)

observed =	0	1	2	3
background = 0.0	2.30	3.89	5.32	6.68
0.5	2.30	3.50	4.83	6.17
1.0	2.30	3.26	4.44	5.71
2.0	2.30	3.00	3.87	4.92
3.0	2.30	2.83	3.52	4.37

Поведение вблизи границ получается в байесовском подходе очень просто — мы используем априорное распределение с информацией о физической границе.

Полученные результаты очень логичны, если использовать плоское априорное распределение с чёткой левой границей.

Мешающие параметры (Nuisance parameters)

Определение

Мешающий параметр — любой неизвестный параметр вероятностного распределения в статистической задаче, связанной с изучением других параметров данного распределения.

В байесовском подходе, включение мешающих параметров также происходит простым способом (если нам известно его распределение $P(b)$, просто принетгрируем по нему).:

$$\mathbb{P}(\theta|\text{data}) = \int_b \frac{\mathbb{P}(\text{data}|\theta, b)\mathbb{P}(\theta)}{\mathbb{P}(\text{data})}\mathbb{P}(b)db.$$

Комбинирование измерений

Ещё одним хорошим свойством байесовского подхода является простая комбинация нескольких измерений:

$$\mathbb{P}(\theta|\mathrm{data}) = \frac{\mathbb{P}_1(\mathrm{data}|\theta)\dots\mathbb{P}_N(\mathrm{data}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathrm{data})}.$$

При этом достаточно использовать только одно априорное распределение.
NB: иногда бывает полезно считать произведение в несколько заходов.

Априорная вероятность

В процессе определения границ эксперимента, мы заботились только о левой границе. А что происходит с правой? Должна ли она быть на бесконечности? в таком случае:

$$\int_b^a \mathrm{Uniform}(x)dx = 0, \forall a, b$$

То есть мы должны также ограничивать правую сторону, причём о ней у нас нет (или почти нет информации).

Кроме того, использование плоского априорного распределения довольно случайно ;-)

Априорная вероятность Джеффриса

Изначально вводились так, чтобы быть инвариантны относительно некоторых преобразований координат. Для каждого семейства кривых, вероятность Джеффриса может быть подсчитана из этого условия. Например, для распределения Пуассона Джеффрис предлагал её сделать инвариантной к маштабу $\sim 1/\mu$.

Bayesian 90% Upper Limits ($1/\mu$ Jeffreys Prior)

observed =	0	1	2	3
background = 0.0	0.00	2.30	3.89	5.32
0.5	0.00	0.00	0.00	0.00
1.0	0.00	0.00	0.00	0.00
2.0	0.00	0.00	0.00	0.00
3.0	0.00	0.00	0.00	0.00

Априорная вероятность Джеффриса

Сейчас предпочитают использовать распределения, которые минимизируют информацию Фишера. Для Пуассоновского распределения:

$$P(\mu) = \frac{1}{\sqrt{mu}}.$$

Что не даёт правильных интервалов в присутствии шума. Исправим:

$$P(\mu) = \frac{1}{\sqrt{mu + b}}.$$

То есть, наше априорное знание о сигнале зависит от знания о шуме :-(

Доверительные интервалы

Доверительные интервалы (confidence intervals)

Определение

Доверительный интервал — это интервал, построенный с помощью случайной выборки из распределения с неизвестным параметром, такой, что он содержит данный параметр с заданной вероятностью. То есть

$$\mathbb{P}(L \leqslant \theta \leqslant U) = p.$$

Заметим, что в Байесовской вероятности мы оцениваем

$$\mathbb{P}(L \leqslant \theta \leqslant U|X)$$

Покрытие (coverage)

Метод, который позволяет построить интервал $(\theta_a; \theta_b)$ такой, что $\mathbb{P}(\theta_a \leq \theta_0 \leq \theta_b) = \beta$, где θ_0 - настоящее значение параметра, обладает свойством покрытия.

Частотные интервалы будут флуктуировать вместе с новыми выборками. Потому, покрытие определяют как доля интервалов, которая содержит настоящее значение θ_0.

NB: наличие покрытия у байесовкого подхода под вопросом.

Покрытие частотных интервалов

На практике, в основном, используются методы, обладающие ассимптотическим покрытием. Если покрытие $\mathbb{P} \leq \beta$, говорят о ”недопокрытии” (undercoverage), если $\mathbb{P} \geq \beta$ (overcoverage).

В принципе, overcoverage — меньшая проблема (но с точки зрения экспериментатора это ухудшает качества эксперимента).

Нормальная теория

Пусть мы берём $X \sim N(\mu; \sigma^2)$. Для известных μ и σ^2:

$$\beta = \mathbb{P}(a \leq X \leq b) = \int_a^b N(\mu, \sigma^2) dX'.$$

При этом, если μ неизвестна, мы больше не сможем подсчитать этот интеграл, вместо этого мы можем оценить вероятность $[\mu + c, \mu + d]$:

$$\beta = \mathbb{P}(\mu + c < X < \mu + d) = \int_{\mu+c}^{\mu+d} N(\mu, \sigma^2) dX' =$$

$$= \int_{c/\sigma}^{d/\sigma} \frac{1}{\sqrt{2\pi}} \exp[\frac{1}{2} Y^2] dY.$$

То есть, мы можем переписать как $\beta = \mathbb{P}(X - d \leq \mu \leq X - c)$

Нормальная теория для интервальных оценок

Такого рода оценка сработала так как:

- была получена функция от $(X-\mu)^2$;
- мы подразумевали, что функция интегрируема и область интегрирования не имеет границ.

Если вспомнить свойства оценки правдоподобия, то асимптотически эти пункты будут выполнены. NB: для этого

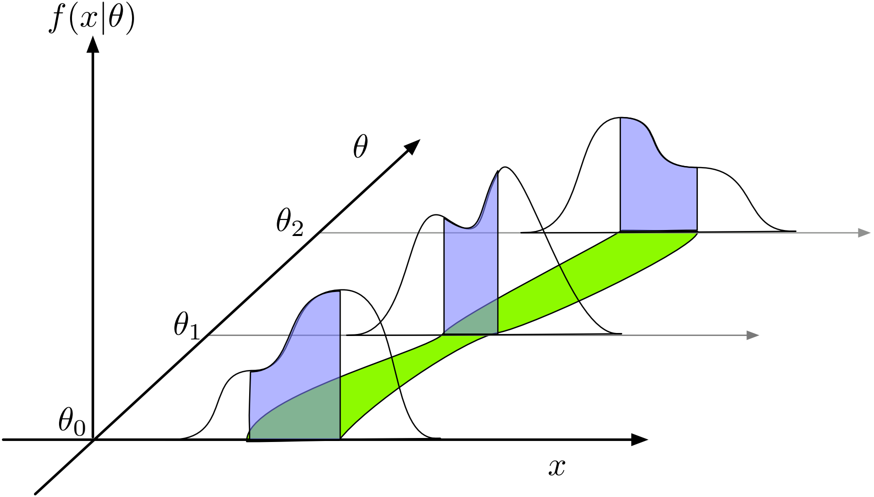
необходимо большое количество событий. NB2: все выводы очень просто распространяются на многомерные модели.

Построение Неймана

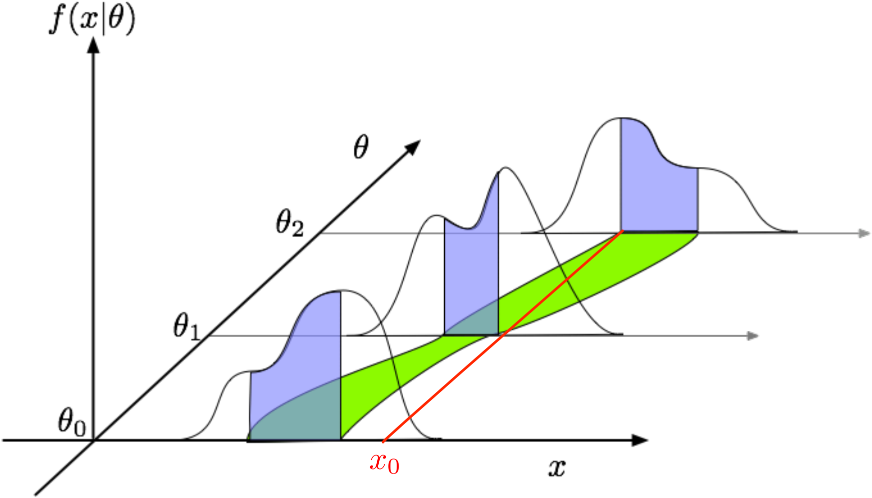
The Neyman construction for constructing frequentist confidence intervals involves the following steps:

- Given a true value of the parameter θ, determine a p.d.f. $f(x;\theta)$ for the outcome of the experiment. Often x is an estimator for the θ.
- Using some procedure, define an interval in x that has a specified probability (say, 90%) of occurring
- Do this for all possible true values of θ, and build a confidence belt of these intervals
- Compute the confidence belt given the value of x.

Построение Неймана

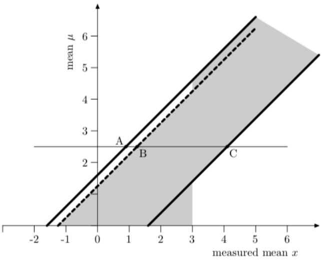


Построение Неймана



При таком построении покрытие получается всегда 100%.

Построение Неймана: проблемы



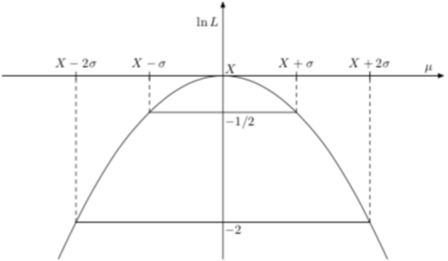
При таком построении появляются проблемы, связанные с поведением вблизи границ:

- пустые интервалы;
- ”флип-флоп” в районе перехода к отделяемому от физической границы пределу.

Эти проблемы решаются дополнительными построениями, например, Фельдман-Казинс предлагает дополнить запрещённые регионы, анализируя относительное правдоподобие.

Доверительные интервалы на основе функции правдоподобия

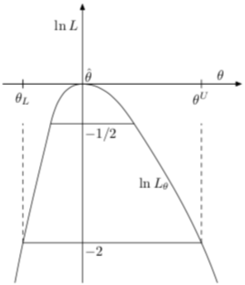
Мотивация



Log-likelihood function for Gaussian X, distributed $N(\mu, \sigma^2)$.

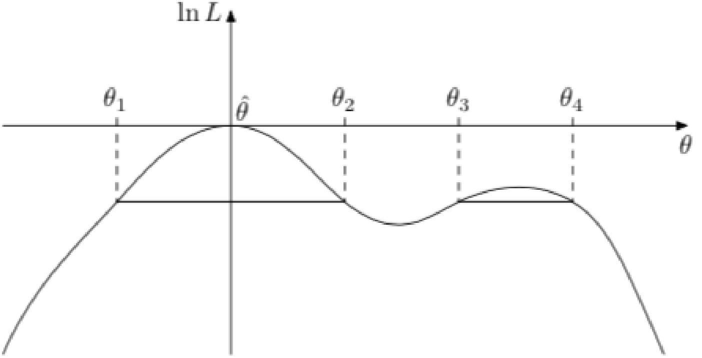
На предыдущих слайдах мы видели, что нормальная теория позволяет честно получать доверительные интервалы для величин, распределённых по Гауссу. Этот результат можно читать по-другому: если правдобие представляет собой параболу, то мы можем честно подсчитать доверительные интервалы.

Независимость правдоподбия от параметризации



В случае, если функция правдоподобия непараболическая, мы (почти) всегда можем привести её к параболическому виду некоторой трансформацией $g(\theta)$. При этом сама фукнция от параметризации независит, потому мы можем оценивать θ_L и θ_H через $\ln L = \ln L_{\max} - 1/2$ (для 68% интервала).

Сложные случаи



“Pathological” log-likelihood function.

В случае многомодальной функции правдоподобия при подобном построении есть шанс найти вторую моду.

Многомерные случаи

Самые большие проблемы начинаются в многомерном случае.

- Использовать нормальную теорию (если правдоподобие гаусово).
- Простой способ, использовать профильную функцию правдоподобия:

  $$g(x_k) = \max_{x_i, i \neq k} \ln L(X).$$

  Этот способ даст возможность анализировать простые негаусовы правдоподобия.
- Использовать объединённый метод, эквивалент бутсрепа, но без фиксированных мешающих параметров.

Систематические погрешности

В принципе, каждый источник систематической погрешности характеризуется своей случайной величиной (вернее, почти каждый). Предположим, что мы знаем плотность этой случайной величины:

- байесовский способ: без проблем, просто маргинализируем правдоподобие;
- классический способ: задача становится очень многомерной;
- смешанный способ: давайте сделаем вид, что мы байесовцы, маргинализируем, а потом используем как классический вывод.

Какой способ препочесть?

- Универсального способа не существует.
- В случае достаточно большой статитистики, нам всё равно.
- Для малой статистики или слишком близкой границы каждый метод имеет недостатки, которые надо учитывать.

Доверительное оценивание на основе бутстрепа

Нормальный интервал

- Если предположить, что данные распределены нормально, то имеет смысл рассмотреть следующий доверительный интервал:

$$(T_n - z_{\alpha/2}\widehat{se}_{boot}, T_n + z_{\alpha/2}\widehat{se}_{boot}),$$

- При этом $z_\alpha : F_{N(0,1)}(z_\alpha) = 1 - \alpha$, $\widehat{se}_{boot} = \sqrt{v_{boot}}$.

Центральный интервал

- Пусть $\theta = T(F)$ и $\widehat{\theta}_n = T(\widehat{F}_n)$.
- Пусть $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}$ — получены итерированием шагов 1 и 2 алгоритма бутстрепа.
- Пусть θ^*_β — обозначает β-квантиль для $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$.
- Тогда центральный $(1-\alpha)$-доверительный интервал :

$$C_n = (2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}).$$

Центральный интервал

Теорема

При некоторых несильных условия на $T(F)$,

$$\mathrm{P}_F(T(F) \in C_n) \to 1 - \alpha, \qquad n \to \infty,$$

$$C_n = (2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2})$$

Покрытие Бутстрепа

Coverage Accuracy	Methods		
	Pivoting	Percentile	Test Inversion
1^{st} order	Non-studentized Bootstrap	Simple Percentile Bias Corrected Percentile	
2^{nd} order	Bootstrap-t	BCa Percentile Automatic Percentile	Bootstrap LR Profile Bootstrap LR
higher order	Bootstrap Calibration or Iteration		

Понятие статистической гипотезы

Понятие статистической гипотезы

Определение

Статистическая гипотеза — определённое предположение о распределении вероятностей, лежащем в основе наблюдаемой выборки данных.

Выделяют 2 типа статистических гипотез:

- **Простая гипотеза** однозначно определяет функцию распределения на рассматриваемом множестве, например $\theta = \theta_0$ — простая гипотеза.
- **Сложная гипотеза** утверждает принадлежность распределения к некоторому множеству распределений на рассматриваемом множестве, например $\theta > \theta_0$ или $\theta < \theta_0$ — сложная гипотеза.

Понятие статистической гипотезы

Определение

Проверка статистической гипотезы — это процесс принятия решения о том, противоречит ли рассматриваемая статистическая гипотеза наблюдаемой выборке данных.

Всегда рассматривается задача проверки гипотезы H_0 против альтернативы H_1 (или же задача (H_0, H_1)).

Статистический критерий и его характеристики

Статистический критерий и его характеристики

Определение

Статистический критерий — строгое математическое правило, по которому принимается или отвергается статистическая гипотеза.

В зависимости от типа статистической гипотезы выделяют односторонние и двусторонние статистические критерии:

- Односторонний критерий

$$H_0 \;: \theta \leq \theta_0 \;\; vs \;\; H_1 \;: \theta > \theta_0.$$

- Двусторонний критерий

$$H_0 \;: \theta = \theta_0 \;\; vs \;\; H_1 \;: \theta \neq \theta_0.$$

Статистический критерий и его характеристики

Также возможно деление на параметрические и непараметрические критерии:

Определение

Параметрический критерий — критерий , предполагающий, что выборка порождена распределением из заданного параметрического семейства. В частности, существует много критериев, предназначенных для анализа выборок из нормального распределения.

Определение

Непараметрический критерий — критерий, не опирающийся на дополнительные предположения о распределении.

Статистический критерий и его характеристики

- Статистический критерий — это правило, которое для каждой реализации x выборки X должно приводить к одному из двух решений: принять гипотезу H_0 или отклонить ее (принять ее альтернативу H_1).
- В связи с этим каждому критерию соответствует некоторое разбиение выборочного пространства χ на два взаимно дополняющих множества χ_0 и χ_1.
- χ_0 состоит из тех реализаций выборки x, для которых H_0 принимается, а χ_1 из тех, для которых H_0 отвергается (принимается H_1).

Статистический критерий и его характеристики

Определение

В определениях предыдущего слайда

- χ_0 – область принятия гипотезы H_0,
- χ_1 – область ее отклонения (критическая область).

Таким образом любой критерий проверки гипотезы H_0 однозначно задается соответствующей критической областью χ_1.

Статистический критерий и его характеристики

Выбор критической области в конкретной задаче делается на основе общего принципа принятия решения.

Определение

Общий принцип принятия решений состоит в следующем:

- Если в эксперименте наблюдается маловероятное при справедливости гипотезы H_0 событие, то считается, что гипотеза H_0 не согласуется с данными и в этом случае она отклоняется.
- В противном случае считается что данные согласуются с H_0 и H_0 принимается.

Статистический критерий и его характеристики

В соответвии с общим принципом принятия решения критическую область χ_1 выбирают так, чтобы была мала вероятность $P(X \in \chi_1 | H_0)$.

Определение

Говорят, что критерий **имеет уровень значимости** α, если

$$P(x \in \chi_1 | H_0) \leq \alpha.$$

Статистический критерий и его характеристики

Следуя любому критерию мы можем принять правильное решение, либо совершить одну из двух ошибок — первого или второго рода.

Более наглядно это показано в таблице ниже:

		Верная гипотеза	
		H_0	H_1
Результат применения	H_0	OK	ошибка 2-го рода
критерия	H_1	ошибка 1-го рода	OK

Статистический критерий и его характеристики

- Пусть критерий имеет критическую область χ_1, а $\mathcal{F} = F_0 \cup F_1$ — множество всех допустимых распределений выборки X.
- При этом F_0 — множество распределений, удовлетворяющих гипотезе H_0, а F_1 — соответственно H_1.

Определение

Функционал $W(F) = W(F; \chi_1) = P(X \in \chi_1 | F), F \in \mathcal{F}$ называется **функцией мощности критерия.**

Другими словами мощность критерия показывает вероятность попадания значения выборки X в критическую область χ_1, когда F — ее истинное распределение.

Статистический критерий и его характеристики

Через функцию мощности легко выразить вероятности обоих типов ошибок, свойственных нашему критерию.

Определение

$W(F)$ — **вероятность ошибки первого рода** при $F \in F_0$.
$1 - W(F)$ — **вероятность ошибки второго рода** при $F \in F_1$.

Статистический критерий и его характеристики

Определение

Размер критерия:

$$\alpha = sup_{F \in F_0} W(F).$$

Отсюда легко видеть что если размер критерия не превосходит α, то его уровень равен α.

Статистический критерий и его характеристики

- Логично стремление построить критерий так, чтобы свести к минимуму вероятности ошибок обоих типов.
- Однако при фиксированном объеме выборки сумма вероятностей ошибок обоих типов не может быть сделана сколь угодно малой.
- Поэтому руководствуются рациональным принципом выбора критической области.

Определение

Из всех критических областей удовлетворяющих заданному уровню значимости выбирается та, для которой вероятность ошибки 2-го рода минимальна.

Статистический критерий. Пример

- Пусть $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, где σ — известно.
- Рассмотрим 2 гипотезы $H_0 : \mu \le 0 \quad H_1 : \mu > 0$.
- Рассмотрим критерий – H_0 отклоняется если $T = \bar{X} > c$.
- Тогда критическая область $\chi_1 = \{(x_1, ...x_n) : T(x_1, ..., x_n) > c\}$.
- Значит

$$W(\mu) = P_\mu(\bar{X} > c) = P_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}\right)$$

$$= P\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right),$$

где Z — случайная величина со стандартным нормальным распределением.

Статистический критерий и его характеристики

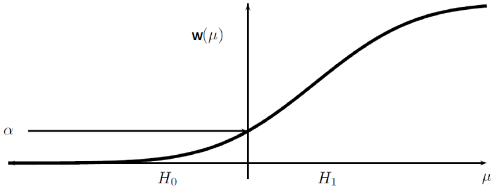


Рис.: $W(\mu)$

Пример

Отсюда легко видеть, что размер критерия равен
$W(0) = 1 - \Phi(\frac{\sqrt{n}c}{\sigma})$

Статистический критерий и его характеристики

Пример

Чтобы размер критерия равнялся α необходимо, чтобы $c = \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}$.

Определение

Наиболее мощный критерий — критерий, который имеет максимальную мощность относительно гипотезы H_1 (т.е максимальную $1 - W(F)$ при $F \in F_1$) среди всех критериев размера α.

Статистический критерий и его характеристики

- В конкретных задачах наиболее мощный критерий не всегда достижим, поэтому в реальных задачах часто приходится ограничиваться более умеренными требованиями.
- Минимальным таким требованием является требование несмещенности.

Определение

Статистический критерий называется несмещенным если при любом альтернативном распределении данных мы должны попадать в критическую область с большей вероятностью нежели при нулевой гипотезе.

Статистический критерий и его характеристики

В случае выборок большого объема важным также является условие состоятельности.

Определение

Статистический критерий называется состоятельным если в случае истинности альтернативной гипотезы при большом числе наблюдений мы будем попадать в критическую область с вероятностью близкой к 1 (т.е., отклоняя нулевую гипотезу, мы будем принимать правильное решение).

Типы статистических критериев

Типы статистических критериев

После краткого введения в теорию проверки гипотез пререйдем непосредственно к разлизным статистическим критериям.

- Критерий Вальда.
- P-значение.
- Распределение хи-квадрат и критерий Пирсона.
- Критерий перестановок.
- Критерий на основе отношения правдоподобия.
- Критерий согласия.
- Критерий Неймана-Пирсона для случая двух простых гипотез, t-критерий.

Критерий Вальда

Пусть:

- θ — скалярный параметр;
- $\hat{\theta}$ — его оценка;
- $\hat{se}$ — оценка стандартной ошибки оценки $\hat{\theta}$.

Гипотеза:

$$H_0 : \theta = \theta_0 \;\; vs. \;\; H_1 : \theta \neq \theta_0.$$

Определение (Критерий Вальда размера α)

Если $\hat{\theta}$ — асимптотически нормальная оценка параметра θ, т.е.

$$W = \frac{\hat{\theta} - \theta_0}{\hat{se}} \to \mathbb{N}(0, 1), n \to \infty,$$

то гипотеза H_0 отклоняется, если $|W| > z_{\alpha/2}$.

Теорема

Асимптотически размер критерия Вальда равен α, то есть

$$W(F) = \mathbb{P}(|W| > z_{\alpha/2}|f) \to \alpha, \quad n \to \infty, \quad f \in F_0.$$

Доказательство.

При условии, что $\theta = \theta_0$, в силу асимптотической нормальности оценки выполнено $\frac{\hat{\theta}-\theta_0}{\hat{se}} \rightsquigarrow \mathbb{N}(0,1)$. Следовательно, вероятность отклонить основную гипотезу, когда она на самом деле верна, равняется:

$$\mathbb{P}(|W| > z_{\alpha/2}|f) = \mathbb{P}(\frac{|\hat{\theta} - \theta_0|}{\hat{se}} > z_{\alpha/2}|f) \to$$
$$\to \mathbb{P}(|Z| > z_{\alpha/2}) = \alpha.$$

$\square$

Пример (Сравнение средних значений)

- Пусть $X_1, ..., X_m$ и $Y_1, ..., Y_n$ - две независимые выборки из генеральных совокупностей.
- Средние значения которых равны μ_1 и μ_2 соответсвенно.
- s_1^2 и s_2^2 — выборочные дисперсии.
- Положим $\delta = \mu_1 - \mu_2$.

$$H_0 : \delta = 0 \;\; vs. \;\; H_1 : \delta \neq 0; \quad \hat{\delta} = \bar{X} - \bar{Y}; \quad \hat{se} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}.$$

Гипотеза H_0 отвергается, если $|W| > z_{\alpha/2}$, где

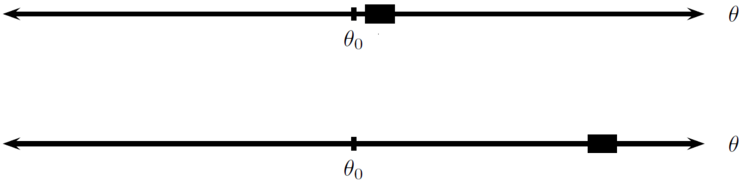
$$W = \frac{\hat{\delta} - 0}{\hat{se}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}.$$

Теорема

Критерий Вальда размера α отклоняет гипотезу $H_0 : \theta = \theta_0$ в пользу $H_1 : \theta \neq \theta_0$ если и только если $\theta_0 \notin C$, где

$$C = (\hat{\theta} - \hat{se} z_{\alpha/2}, \hat{\theta} + \hat{se} z_{\alpha/2})$$

Таким образом, тестирование гипотезы эквивалентно проверке, попало ли значение Θ_0 в доверительный интервал.



P-значение

Определение

Пусть для каждого $\alpha \in (0,1)$ имеется критерий размера α для некоторой статистики (функции от выборки) $T(X^n)$ с критической областью R_α. Тогда

$$p - value = inf\{\alpha : T(X^n) \in R_\alpha\}.$$

- Таким образом, $p - value$ - это наименьший уровень значимости, на котором еще можно отклонить H_0.
- Чем меньше $p - value$ - тем вероятнее, что H_0 надо отклонить.

Типичные значения для $p - value$:

- $< 0.01 \rightarrow H_0$ - заведомо не верна
- $0.01 - 0.05 \rightarrow H_0$ - не верна
- $0.05 - 0.10 \rightarrow H_0$ - скорее не верна
- $> 0.1 \rightarrow$ ничего определенного о гипотезе H_0 сказать нельзя.

Большое p-value не является подтверждением гипотезы H_0.
Большое $p - value$ появляется, если:

- H_0 - верна.
- H_0 - неверна, но мощность критерия недостаточна.

Теорема

Пусть критерий размера α, построенный для статистики $T(X^n)$, имеет вид: H_0 отвергается, если $T(X^n) > c_\alpha$
Гипотезе H_0 соответствует семейство распределений F_0.
Тогда

$$p - value = \sup_{f \in F_0} \mathbb{P}(T(X^n) \geq T(x^n)|f),$$

где x^n - реализация выборки X^n. Если $F_0 = f$, то

$$p - value = \mathbb{P}(T(X^n) \geq T(x^n)|f).$$

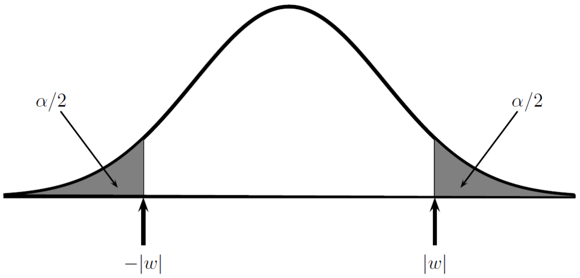
То есть $p - value$ - это вероятность (при выполнении гипотезы H_0) того, что статистика $T(X^n)$ примет значение больше либо равное тому, которое реализовалось в опыте (реализация x^n)

Теорема

Пусть $w = \frac{(\hat{\theta}-\theta_0)}{\hat{se}}$ - наблюдаемое значение статистики Вальда W. Тогда:

$$p - value = \mathbb{P}(|W| > |w||f) \simeq \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|),$$

где $Z \sim N(0,1), \quad f \in F_0$.



Пример (Равенство значений холестерина в крови)

Здесь как и в примере про сравнение средних значений:

$$W = \frac{\hat{\delta} - 0}{\hat{se}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{216.2 - 195.3}{\sqrt{5^2 + 2.4^2}} = 3.78.$$

Пусть $Z \sim N(0, 1)$, тогда

$$p - value = \mathbb{P}(|Z| > 3.78) = 2 \cdot \mathbb{P}(Z < -3.78) = 0.0002$$

.

Распределение хи-квадрат (χ^2)

Определение (распределение хи-квадрат(χ^2))

Пусть $Z_1, ..., Z_k$ - независимые стандартно нормально распределенные случайные величины. $V = \sum_{i=1}^{k} Z_i^2$, тогда $V \sim \chi_k^2$ - хи-квадрат с k степенями свободы

$$F(V) = \frac{v^{\frac{k}{2}-1} e^{-\frac{V}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})}, \quad \mathbb{E}(V) = k, \quad \mathbb{V}(V) = 2k$$

$\chi_{k,\alpha}^2 = F^{-1}(1-\alpha)$ - верхняя квантиль, F - функция распределения, т.е.

$$\mathbb{P}(\chi_k^2 > \chi_{k,\alpha}^2) = \alpha.$$

Критерий перестановок

Определение (Постановка задачи)

Критерий перестановок применяется для проверки того, отличаются ли распределения.
Пусть $X_1, ...X_m \sim F_X$ и $Y_1, ..., Y_n \sim F_Y$ - две независимые выборки. Требуется решить:

$$H_0 : F_X = F_Y \quad vs. \quad H_1 : F_X \neq F_Y$$

Критерий перестановок - ”точный” в том смысле, что он не использует предположения об асимптотической сходимости к нормальному распределению.

Критерий перестановок:

1. Обозначим через $T(x_1, ..., x_m, y_1, ..., y_n)$ некоторую тестовую статистику, например, $T(X_1, ..., X_m, Y_1, ..., Y_n) = |\bar{X}_m - \bar{Y}_n|$.
2. Положим $N = m + n$ и рассмотрим все $N!$ перестановок объединенной выборки $X_1, ..., X_m, Y_1, ..., Y_n$.
3. Для каждой из перестановок подсчитаем значение статистики T.
4. Обозначим эти значения $T_1, ..., T_{N!}$.

Теорема (Критерий перестановок)

Если H_0 верна, то при фиксированных упорядоченных значениях $\{X_1, ..., X_m, Y_1, ..., Y_n\}$ значение статистики T распределены равномерно на множестве $T_1, ..., T_{N!}$.

Теорема

Обозначим как перестановочное распределение статистики T такое, согласно которому:

$$P_0(T = T_i) = \frac{1}{N!}, \quad i = 1, ..., N!$$

Пусть t_{obs} - значение статистики, которое было получено в опыте. Тогда:

$$p - value = \mathbb{P}(T > t_{obs}|f) = \frac{1}{N!}\sum_{j=1}^{N!}\mathbb{I}(T_j > t_{obs}), \quad f \in F_0$$

Пример

Допустим, что $(X_1, X_2, Y_1) = (1, 9, 3)$. Пусть
$T(X_1, X_2, Y_1) = |\bar{X} - \bar{Y}| = 2$, тогда

Перестановка	Значение T	Вероятность
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

$$p - value = \mathbb{P}(T > 2) = 4/6$$

Критерий на основе отношения правдоподобия

Определение

Рассмотрим две конкурирующие гипотезы

$$H_0 : f \in F_0 \quad vs. \quad H_1 : f \in F_1$$

Пусть $\hat{f}$ - ОМП и $\hat{f}_0$ - ОМП при $f \in F_0$ Статистика отношения правдоподобия имеет вид:

$$\lambda = 2 \log \frac{\sup_{f \in \mathcal{F}} \mathfrak{L}(f)}{\sup_{f \in F_0} \mathfrak{L}(f)} = 2 \log \frac{\mathfrak{L}(\hat{f})}{\mathfrak{L}(\hat{f}_0)}.$$

Теорема

Допустим, что $f = (\theta_1, \ldots, \theta_q, \theta_{q+1}, ..., \theta_r)$. Пусть $F_0 = \{f : (\theta_{q+1}, \ldots, \theta_r) = (\theta_{0,q+1}, ..., \theta_{0,r})\}$.
Пусть λ - критерий на основе отношения правдоподобия. При гипотезе $H_0 : f \in F_0$

$$\lambda(x^n) \rightsquigarrow \chi^2_{q,\alpha},$$

где q - размерность F за вычетом размерности F_0. $p-value$ для критерия равно $\mathbb{P}(\chi^2_q > \lambda)$.

Пример

Пусть $f = (\theta_1, \ldots, \theta_5)$, необходимо проверить, что $\theta_4 = \theta_5 = 0$. Тогда у предельного распределения имеется 3 степени свободы.

Пример (Горох Менделя $1/3$)

Пример. Горох Менделя. Два типа: круглые желтые зерна и сморщенные зеленые зерна. Имеется 4 типа потомков: круглые желтые, сморщенные желтые, круглые зеленые и сморщенные зеленые. Количество потомков каждого типа образуют мультиномиальное распределение с вероятностью $p = (p_1, p_2, p_3, p_4)$.

Из теории следует, что

$$p = (\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}).$$

В опыте получено, что $n = 556, X = (315, 101, 108, 32)$.

Пример (Горох Менделя $2/3$)

Статистика отношений правдоподобия для

$$H_0 : p = p_0 \quad vs. \quad H_1 : p \neq p_0$$

принимает вид:

$$\lambda = 2\log\frac{\mathfrak{L}(\hat{p})}{\mathfrak{L}(\hat{p_0})} = 2\sum_{j=1}^{4} X_j \log\frac{\mathfrak{L}(\hat{p})}{\mathfrak{L}(\hat{p_0})} = 2\cdot(315\log(\frac{315/556}{9/16})+$$

$$+101\log(\frac{101/556}{3/16}) + 108\log(\frac{108/556}{3/16}) + 32\log(\frac{32/556}{1/16})) = 0.48$$

Пример (Горох Менделя $3/3$)

При гипотезе H_1 4 параметра. Так как сумма параметров должна равняться 1, то размерность пространства параметров равна 3. При гипотезе H_0 свободных параметров нет, значит количество степеней свободы равно 3 и χ_3^3 является предельным распределением.

$$p - value = \mathbb{P}(\chi_3^2 > 0.48) = 0.92$$

Замечание

Как правило, и критерий χ^2, и критерий отношения правдоподобий дают примерно одинаковые результаты при условии, что размер выборки достаточно большой.

Критерий Неймана-Пирсона для случая двух простых гипотез

Лемма (Неймана-Пирсона)

$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1$

Статистика Неймана-Пирсона:

$$T = \frac{\mathfrak{L}(\theta_1)}{\mathfrak{L}(\theta_0)} = \frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)}. \tag{1}$$

Допустим, что H_0 отвергается при $T > k$. Выберем k так, что $\mathbb{P}_{\theta_0}(T > k) = \alpha$.

Тогда, критерий Неймана-Пирсона (на основе статистики (1)) будет иметь наибольшую мощность $W(\theta_1)$ среди всех критериев размера α.

t-критерий

Определение

Случайная величина имеет распределение Стьюдента (t-распределение) с k степенями свободы, если:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1+\frac{t^2}{k})^{\frac{k+1}{2}}}$$

При $k \to \infty$ t-распределение стремится к стандартному нормальному распределению. При $k = 1$ t-распределение совпадает с распределением Коши.

t-критерий используют, когда распределение данных близко к нормальному, а размер выборки невелик.

Теорема (t-критерий)

Пусть $X_1, ..., X_n \sim N(\mu, \sigma^2)$, где параметры (μ, σ^2) неизвестны.

$$H_0 : \mu = \mu_0 \quad vs. \quad H_0 : \mu \neq \mu_0$$

Обозначим через S_n^2 выборочную дисперсию. Тогда статистика критерия:

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n}$$

Основная гипотеза отвергается, если $|T| > t_{n-1,\alpha/2}$, где $t_{n-1,\alpha/2}$— квантиль распределения Стьюдента с **n-1** степенями свободы.

При больших n выполняется $T \sim N(0, 1)$, то есть при больших n t-критерий эквивалентен критерию Вальда.