Теория вероятностей и математическая статистика

Доверительные интервалы II.

Глеб Карпов

ФКН ВШЭ

t-распределение Стьюдента

Definition

Мы говорим, что случайная величина имеет t распределение с k степенями свободы, если она построена как функция от стандартной нормальной случайной величины и $\chi^2\,(k)$ величины:

$$t(k\ df) = \frac{Z}{\sqrt{\frac{\chi^2(k\ df)}{k}}}$$

Степень свободы для t полностью определяется степенью свободы величины χ^2, которая использовалась для построения.

Построение t-распределения

Утверждение: Пусть у нас есть случайная выборка $\mathcal{X} = (X_1, X_2, \dots, X_n)$ (независимые, одинаково распределенные) с $\mu \equiv E[X_i]$, $\sigma^2 \equiv Var[X_i]$, а также $X_i \sim \mathcal{N}(\mu, \sigma^2)$, то есть исследуемая случайная величина приходит из нормального распределения. Тогда случайная величина:

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

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

имеет распределение Стьюдента с $(n-1)$ степенями свободы. (да-да, именно $(n-1)$, это не баг).

Построение t-распределения

Доказательство

Начнём с определения t-переменной:

$$t_{k\ df} = \frac{Z}{\sqrt{\frac{\chi^2(k\ df)}{k}}}.$$

Получим по цветам отдельные части этой формулы и соединим вместе :)

1. Z — это получим из стандартизации выборочного среднего:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

2. $\chi^2(k\ df)$ — это получим из распределения выборочной дисперсии:

$$\chi^2(n-1\ df) = \frac{S^2(n-1)}{\sigma^2}$$

3. k — а это число степеней свободы у распределения выборочной дисперсии: $k = n - 1$

Построение t-распределения

Доказательство

- Собираем разноцветную формулу вместе:

$$t_{(n-1)\ df} = \frac{Z}{\sqrt{\frac{\chi^2(n-1\ df)}{n-1}}} = \frac{\bar{X}-\mu}{\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}.$$

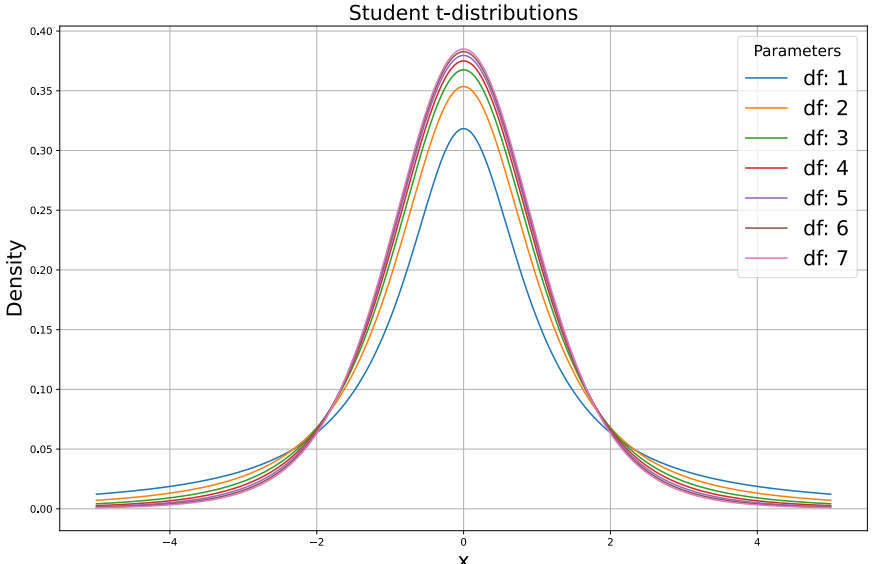
- В процессе сократились $(n-1)$ и стандартные отклонения σ, и мы получили изначальное утверждение.
- Эта форма t-распределения активно используется в статистике.

Асимптотические свойства t-распределения

При увеличении числа степеней свободы функция плотности t-распределения стремится к функции плотности стандартного нормального распределения.

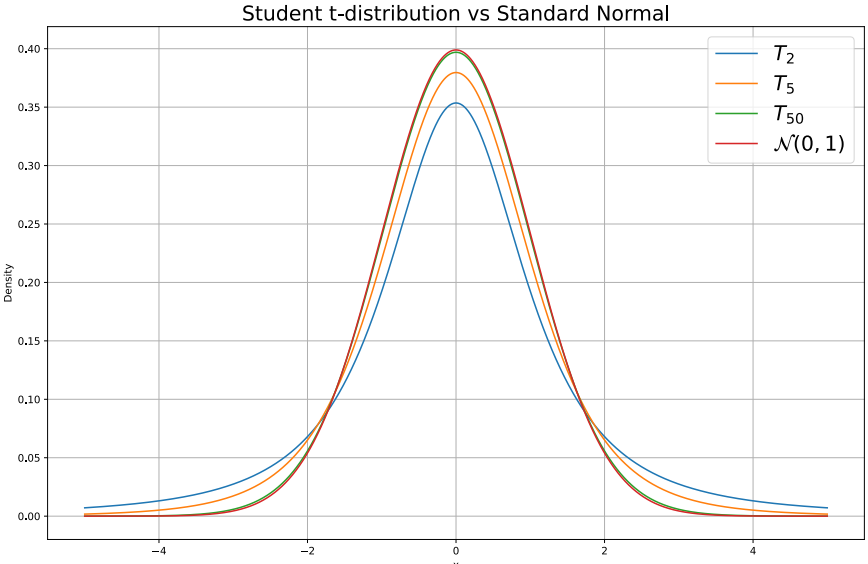
t-распределение Стьюдента

Функции плотности при разных степенях свободы



t-распределение Стьюдента

Асимптотические свойства t-распределения



Доверительные интервалы для неизвестного матожидания

Дисперсия исследуемой случайной величины неизвестна

- В реальности дисперсия интересующей нас переменной неизвестна.
- Чтобы всё же построить желаемый доверительный интервал, мы используем t-распределение Стьюдента.

$$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

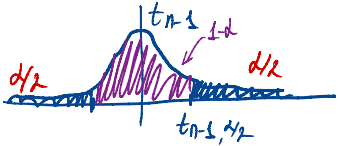
$E[S^2] = \sigma^2$

- Вооружившись этой новой идеей, мы действуем уже знакомым способом:

$$1-\alpha = P(L(X) < \mu < U(X)) = P(-U < -\mu < -L) =$$

$$P\left(\frac{\bar{X}-U}{\frac{S}{\sqrt{n}}} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < \frac{\bar{X}-L}{\frac{S}{\sqrt{n}}}\right) = P\left(-t_{n-1,\alpha/2} < t_{(n-1)} < t_{n-1,\alpha/2}\right) = 1-\alpha$$

t_{n-1} ~ T



Доверительные интервалы для неизвестного матожидания

Дисперсия исследуемой случайной величины неизвестна

- После нахождения требуемой **критической** точки $t_{n-1,\,\alpha/2}$, такой что $P\left(t_{(n-1)} > t_{n-1,\,\alpha/2}\right) = \frac{\alpha}{2}$, мы восстанавливаем верхнюю и нижнюю границы как:

$$t_{n-1,\,\alpha/2} = \frac{\bar{X} - L}{\frac{S}{\sqrt{n}}} \rightarrow L = \bar{X} - t_{n-1,\,\alpha/2}\frac{S}{\sqrt{n}}$$

$$-t_{n-1,\,\alpha/2} = \frac{\bar{X} - U}{\frac{S}{\sqrt{n}}} \rightarrow U = \bar{X} + t_{n-1,\,\alpha/2}\frac{S}{\sqrt{n}}$$

- На практике $(1-\alpha)100\%$ доверительный интервал для неизвестного матожидания $\mu \equiv E[X]$ исследуемой случайной величины записывается как:

$$\mu \in \left(\bar{x} - t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1\,\alpha/2}\frac{s}{\sqrt{n}}\right),$$

где $\bar{x}$, s — **реализации** выборочного среднего и выборочного стандартного отклонения.

Иллюстративные задачи

Пример 2: Анализ службы поддержки клиентов

Отдел обслуживания клиентов хочет проанализировать время их ответа на запросы клиентов. Они случайным образом выбрали 30 обращений в службу поддержки и зафиксировали время ответа (в минутах) для каждого обращения. Выборочное среднее время ответа составило 45.2 минут с выборочным стандартным отклонением 12.8 минут.

1. Постройте 95% доверительный интервал для истинного математического ожидания времени ответа на обращения в службу поддержки.
2. У отдела целевое время ответа составляет 40 минут. Основываясь на вашем доверительном интервале, можете ли вы сделать вывод о том, достигают ли они этой цели?

X – время ответа

$E[X]$, $Var[X]$ – ?

$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$ d.f.

$var[X]$ Неизвестна

Иллюстративные задачи

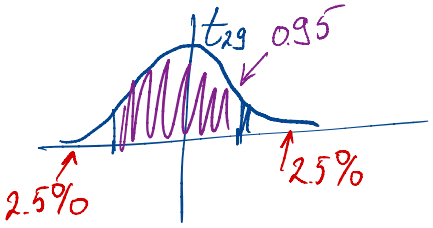
Решение

1. $n = 30$, $\bar{x} = 45.2$, $s = 12.8$, $t_{29,0.025} = 2.045$:

$$\mu \in \left(45.2 - 2.045 \cdot \frac{12.8}{\sqrt{30}}, 45.2 + 2.045 \cdot \frac{12.8}{\sqrt{30}}\right) = (40.4, 50.0)$$

2. Целевое значение 40 минут не попадает в интервал. Нельзя однозначно заключить, что цель достигнута, так как даже в лучшем случае (левая граница интервала) средняя продолжительность звонка оказывается больше, чем целевое значение.

Если бы целевое значение было 50 минут, тогда бы мы могли утверждать, что цель достигается, так как даже в худшем случае средняя продолжительность укладывалась бы в 50 минут.



Доверительные интервалы для разности истинных долей

Мотивация

- В реальном мире часто возникает необходимость **сравнивать** доли между двумя группами:
  - Какая версия сайта лучше конвертирует посетителей в покупателей?
  - Какая маркетинговая кампания эффективнее привлекает клиентов?
  - На какой производственной линии меньше брака?
  - Какое лекарство эффективнее лечит заболевание?
- Пусть p_1 — истинная доля в первой генеральной совокупности, p_2 — истинная доля во второй генеральной совокупности. Нас интересует **разность** $\theta = p_1 - p_2$.
- Для бизнеса, маркетинга, медицины и социологии критически важно знать, есть ли **статистически значимая разница** между долями в двух группах. Это позволяет принимать обоснованные решения о выборе стратегии, продукта или лечения.
- Новый вопрос в статистике: как построить доверительный интервал для разности $p_1 - p_2$?
- **Важная идея**: если доверительный интервал для $p_1 - p_2$ не содержит нуль, это означает статистически значимую разницу между долями в двух группах.

Точечная оценка для разности истинных долей

- Предположим, у нас есть случайная выборка $\mathcal{X} = \{X_1, \dots, X_n\}$ из распределения Бернулли с $P(X_i = 1) = p_1$, и случайная выборка $\mathcal{Y} = \{Y_1, \dots, Y_m\}$ из распределения Бернулли с $P(Y_i = 1) = p_2$. Тогда обе случайные выборки - процессы Бернулли длины n и m соответственно.
- Нас интересует разность истинных долей (или, то же самое, разность вероятностей успеха):

$$\theta = p_1 - p_2$$

- Если $n, m > 30$, то по ИТМЛ:

$$\hat{p}_1 \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right), \quad \hat{p}_2 \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right)$$

- Введём точечную оценку $\hat{\theta} = \hat{p}_1 - \hat{p}_2$ — разность двух выборочных долей.
- Свойства точечной оценки: $E[\hat{\theta}] = p_1 - p_2$, $Var[\hat{\theta}] = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$.
- Так как сумма двух нормальных случайных величин — нормальная случайная величина:

$$\hat{\theta} \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\right)$$

Доверительные интервалы для разности истинных долей

- Если выполнены условия ИТМЛ, то действуем знакомым способом:

$$1-\alpha = P(L < p_1 - p_2 < U) = P(-U < -(p_1 - p_2) < -L) =$$

$$P\left(\frac{\hat{\theta}-U}{\sqrt{Var(\hat{\theta})}} < \frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} < \frac{\hat{\theta}-L}{\sqrt{Var(\hat{\theta})}} \right) = P\left(-z_{\alpha/2} < Z < z_{\alpha/2} \right)$$

- Находим точку $z_{\alpha/2}$, такую, что $P\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2}$, выполняем обратное преобразование, и находим теоретические границы такого доверительного интервала:

$$L = \hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}, \quad U = \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$$

где p_1 и p_2 — истинные параметры (константы), которые заменяются на их точечные оценки на практике, так как истинные параметры неизвестны.

Доверительные интервалы для разности истинных долей

На практике $(1-\alpha)100\%$ доверительный интервал для разности долей генеральных совокупностей:

$$p_1 - p_2 \in \left(\tilde{p}_1 - \tilde{p}_2 - z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{m}},\ \tilde{p}_1 - \tilde{p}_2 + z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{m}} \right),$$

где $\tilde{p}_1$, $\tilde{p}_2$ — реализации выборочных долей.

Иллюстративные задачи

Пример 1: сравнение версий сайта

Компания проводит сравнение двух версий главной страницы сайта. Версия A показана 500 пользователям, из них 85 совершили покупку. Версия B показана 480 пользователям, из них 112 совершили покупку. Постройте 95% доверительный интервал для разности долей конверсии между версиями.

Иллюстративные задачи

Решение

$\tilde{p}_1 = \frac{85}{500} = 0.17$, $\tilde{p}_2 = \frac{112}{480} = 0.233$, $z_{0.025} = 1.96$:

$$p_1 - p_2 \in \left(0.17 - 0.233 - 1.96\sqrt{\frac{0.17 \cdot 0.83}{500} + \frac{0.233 \cdot 0.767}{480}},\ 0.17 - 0.233 + 1.96\sqrt{\frac{0.17 \cdot 0.83}{500} + \frac{0.233 \cdot 0.767}{480}}\right)$$

$$= (-0.103, -0.023)$$

Интервал не содержит нуль, можем сделать вывод, что версия сайта B имеет статистически значимо более высокую конверсию.

Доверительные интервалы для разности математических ожиданий

Мотивация

- В коммерческих и научных исследованиях бывает необходимость сравнить средние значения между двумя группами:
  - Какая производственная линия более производительна?
  - Какой метод обучения даёт лучшие результаты?
  - В каком регионе выше средний доход населения?
- Пусть μ_X — истинное математическое ожидание в первой генеральной совокупности, μ_Y — истинное математическое ожидание во второй генеральной совокупности. Нас интересует **разность** $\theta = \mu_X - \mu_Y$.
- Важно знать, есть ли **статистически значимая разница** между математискими ожиданиями в двух группах. Это позволяет принимать дальнейшие обоснованные решения.
- Новый вопрос: как построить доверительный интервал для разности $\mu_X - \mu_Y$?
- **Идея**: если доверительный интервал для $\mu_X - \mu_Y$ не содержит нуль, это означает статистически значимую разницу между средними значениями в двух группах.

Точечная оценка разности математических ожиданий

Если дисперсии известны

- Предположим, у нас есть две независимые выборки: $\mathcal{X} = \{X_1, \dots, X_n\}$, $\mathcal{Y} = \{Y_1, \dots, Y_m\}$. Характеристики называем $\mu_X \equiv E[X_i]$, $\sigma_X^2 \equiv Var[X_i]$, и соответственно $\mu_Y \equiv E[Y_i]$, $\sigma_Y^2 \equiv Var[Y_i]$
- Дисперсии σ_X^2 и σ_Y^2 предполагаем **известными**.
- Нас интересует разность истинных математических ожиданий:

$$\theta = \mu_X - \mu_Y$$

- Введём точечную оценку $\hat{\theta} = \bar{X} - \bar{Y}$ — разность двух выборочных средних.
- Свойства точечной оценки: $E[\hat{\theta}] = \mu_X - \mu_Y$, $Var[\hat{\theta}] = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$.
- При $n, m > 30$ работает ЦПТ и распределение точечной оценки для θ:

$$\hat{\theta} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

Доверительные интервалы для разности математических ожиданий

Если дисперсии известны

- Если выполнены условия (большие выборки или нормальное распределение), то действуем знакомым способом:

$$1-\alpha = P(L < \mu_X - \mu_Y < U) = P(-U < -(\mu_X - \mu_Y) < -L) =$$

$$P\left(\frac{\bar{X}-\bar{Y}-U}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} < \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} < \frac{\bar{X}-\bar{Y}-L}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}\right) = P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right)$$

- Находим точку $z_{\alpha/2}$, такую, что $P\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2}$, выполняем обратное преобразование, и находим теоретические границы такого доверительного интервала:

$$L = \bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}, \quad U = \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$$

где σ_X^2 и σ_Y^2 — известные дисперсии (константы).

Доверительные интервалы для разности математических ожиданий

Если дисперсии известны

На практике $(1-\alpha)100\%$ доверительный интервал для разности математических ожиданий:

$$\mu_1 - \mu_2 \in \left(\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}},\ \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right),$$

где $\bar{x}$, $\bar{y}$ — реализации выборочных средних, а σ_X^2 и σ_Y^2 — известные дисперсии исследуемых случайных процессов.

Иллюстративные задачи

Пример 2: Сравнение производительности двух производственных линий

Производственная компания хочет сравнить среднюю производительность двух производственных линий. Известно, что стандартное отклонение производительности для первой линии составляет $\sigma_1 = 12$ единиц продукции в час, а для второй линии — $\sigma_2 = 15$ единиц продукции в час.
Было проведено тестирование: - Первая линия: выборка из $n = 40$ часов работы, средняя производительность $\bar{x} = 145$ единиц/час - Вторая линия: выборка из $m = 35$ часов работы, средняя производительность $\bar{y} = 138$ единиц/час

1. Постройте 95% доверительный интервал для разности средних производительностей двух линий.
2. Можете ли вы сделать вывод о том, какая линия более производительна?

Иллюстративные задачи

Решение

1. $n = 40$, $m = 35$, $\bar{x} = 145$, $\bar{y} = 138$, $\sigma_1 = 12$, $\sigma_2 = 15$, $z_{0.025} = 1.96$:

$$\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} = \sqrt{\frac{12^2}{40} + \frac{15^2}{35}} = \sqrt{3.6 + 6.43} = \sqrt{10.03} \approx 3.17$$

$$\mu_1 - \mu_2 \in (145 - 138 - 1.96 \cdot 3.17,\ 145 - 138 + 1.96 \cdot 3.17) = (0.79, 13.21)$$

2. Интервал $(0.79, 13.21)$ не содержит нуль и полностью находится в положительной области. Это означает, что мы можем с определенной степенью полагать, что первая линия имеет статистически значимо более высокую производительность, чем вторая.