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:

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

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:

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

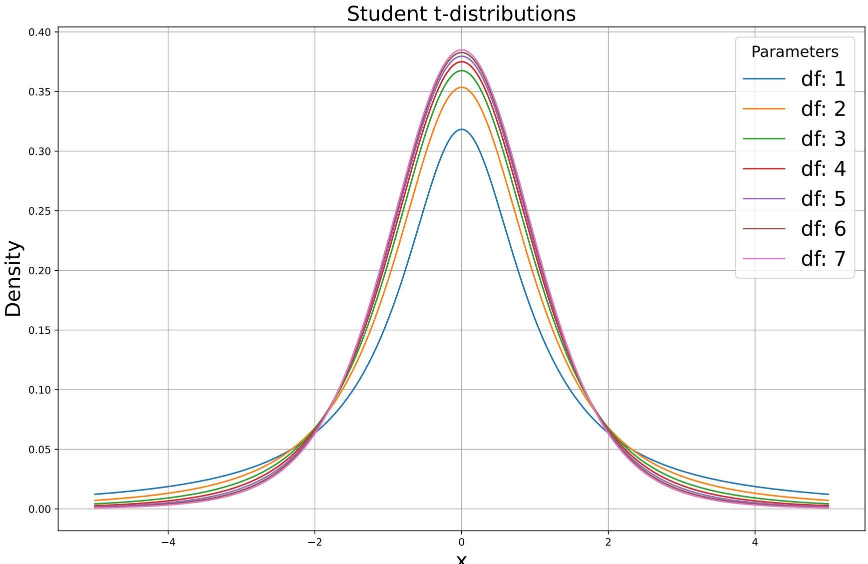
- В процессе сократились $(n-1)$ и стандартные отклонения σ, и мы получили изначальное утверждение.
- Эта форма t-распределения активно используется в статистике.

Асимптотические свойства t-распределения

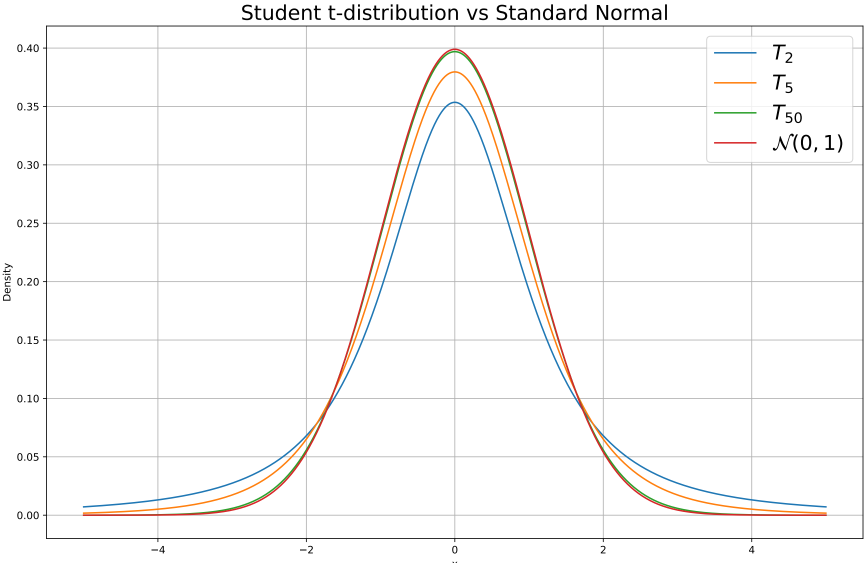
При увеличении числа степеней свободы функция плотности t-распределения стремится к функции плотности стандартного нормального распределения.

t-распределение Стьюдента

Функции плотности при разных степенях свободы



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.

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

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

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

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

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:

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

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

- 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.

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?

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.