Теория вероятностей и математическая статистика

Интервальные оценки II: интервалы для одного параметра. Распределение Стьюдента.

Глеб Карпов

ВШБ Бизнес-информатика

Напоминание: распределения Бернулли и биномиальное

- Случайный эксперимент Бернулли — эксперимент с двумя исходами, часто представляемый как случайная величина со значениями 0 и 1.
- Вероятность успешного исхода ($X = 1$) обозначается как p, поэтому функция вероятности имеет вид: $P(X = 1) = p$, $P(X = 0) = 1 - p = q$. Математическое ожидание равно $E[X] = p$.
- Если случайная величина Бернулли независимо реализуется в последовательности с фиксированной вероятностью p, это называется **процессом Бернулли**. Пример: подбросить одинаковую монету 7 раз подряд.
- Биномиальное распределение: распределение случайной величины, обозначающей количество успешных результатов в процессе Бернулли. Пример: каждая новая последовательность из 7 подбрасываний монеты, вероятно, будет иметь разное количество орлов.
- Если Y обозначает число успешных результатов, её функция вероятности записывается как:

$$P(Y = k) = C_n^k \, p^k \, q^{(n-k)}$$

- Математическое ожидание $E[Y] = np$, а дисперсия $Var[Y] = npq$.

Доля генеральной совокупности: мотивация

- - Нравится ли вам качество бренда?
  - Нравится ли вам этот новый тип кузова автомобиля?
  - Вы курите?
  - Является ли этот продукт бракованным?

  Ответы на все эти и подобные вопросы в основном бинарные или могут быть сделаны бинарными.
- Большая группа людей или объектов (генеральная совокупность) часто обладает некоторым бинарным признаком или атрибутом. Можно предположить, что существует некоторая *доля* p людей или объектов в генеральной совокупности, обладающих одним и тем же конкретным бинарным признаком.
- Для бизнеса, независимой социологии, медицины и некоторых естественных наук часто важно знать эту долю. Однако, чтобы сделать это честно, нужно исследовать *всех* людей или объектов, что обычно буквально невозможно.
- Новый вопрос в статистике: как приблизительно оценить эту p только по выборке бинарных ответов?
- **Важная идея**: величину p можно также интерпретировать как вероятность случайно взять элемент, обладающий искомым признаком, из генеральной совокупности

Точечная оценка для истинной доли

- Предположим, у нас есть выборка $\{X_1, \dots, X_n\}$ из распределения Бернулли с $P(X_i = 1) = p$. Вся выборка тогда может рассматриваться как процесс Бернулли длины n.
- Введём Y — случайную величину, показывающую количество положительных ответов в выборке, она имеет биномиальное распределение с $E[Y] = np$ и $Var[Y] = np(1-p)$.
- Тогда $\hat{p} = \frac{Y}{n}$ — случайная величина, называемая **выборочной долей** и показывающая долю положительных ответов к размеру выборки.
- Её свойства могут быть выведены из Y, а именно:

$$E[\hat{p}] = \frac{E[Y]}{n} = p, \quad Var[\hat{p}] = \frac{Var[Y]}{n^2} = \frac{p(1-p)}{n}$$

- Как следствие ИТМЛ, если $n > 30$, то:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right), \quad \text{или} \quad \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1)$$

Доверительные интервалы для истинной доли признака в генеральной совокупности

- Если выполнены условия ИТМЛ, то действуем знакомым способом:

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

$$P\left(\frac{\hat{p}-U}{\sqrt{\frac{p(1-p)}{n}}} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{\hat{p}-L}{\sqrt{\frac{p(1-p)}{n}}}\right) = P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right)$$

- Находим из таблицы или иным образом критическую точку $z_{\alpha/2}$, такую, что $P\left(Z > z_{\alpha/2}\right) = \frac{\alpha}{2}$, и выполняем обратное преобразование:

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

- Однако, эта формула над нами смеется. Мы не знаем истинное значение p, хотим его поймать в доверительный интервал, а оно присутствует в формуле. :clown_face:

Доверительные интервалы для истинной доли признака в генеральной совокупности

Поэтому на практике $(1-\alpha)100\%$ доверительный интервал для истинной доли признака p записывается как:

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

где $\tilde{p}$ — **реализация** выборочной доли, т.е. реальные, окончательные числа в нашем распоряжении, показывающие долю бинарного признака в имеющейся выборке.

Пример: выборка размера 200 содержит 15 бракованных продуктов, в выборке из 80 студентов 64 выполняют домашние задания самостоятельно, *и т.д.*

Иллюстративные задачи

Пример 1: Анализ рынка

Маркетинговая команда премиальной кофейной сети хочет понять предпочтения клиентов относительно нового сезонного напитка. Они провели опрос среди 400 случайно выбранных клиентов, и 156 из них выразили заинтересованность в новом напитке.

1. Постройте 90% доверительный интервал для истинной доли всех клиентов, которые были бы заинтересованы в новом сезонном напитке.
2. Маркетинговая команда хочет быть уверена на 95%, что их оценка доли генеральной совокупности находится в пределах ± 0.03 (3 процентных пункта) от истинной доли. Какой размер выборки им потребуется для достижения этого уровня точности?
3. Основываясь на доверительном интервале из пункта 1, порекомендовали бы вы запуск нового напитка, если компания требует, чтобы по крайней мере 35% клиентов были заинтересованы для того, чтобы запуск был прибыльным? Объясните ваши рассуждения.

Иллюстративные задачи

Решение

1. $\tilde{p} = \frac{156}{400} = 0.39$, $z_{0.05} = 1.645$:

$$p \in \left(0.39 - 1.645\sqrt{\frac{0.39 \cdot 0.61}{400}}, 0.39 + 1.645\sqrt{\frac{0.39 \cdot 0.61}{400}}\right) = (0.350, 0.430)$$

2. $z_{0.025} = 1.96$. Полуширина интервала $= E = 0.03$.
Так как мы не знаем, какая доля получится в новой выборке, используем самую консервативную оценку доли $\tilde{p} = 0.5$, при ней интервал получается наиболее широким при прочих фиксированных переменных:

$$E = z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$$

$$n = \frac{z^2_{\alpha/2}\tilde{p}(1-\tilde{p})}{E^2} = \frac{1.96^2 \cdot 0.5 \cdot 0.5}{0.03^2} \approx 1068$$

3. Нижняя граница интервала 0.35, совпадает с минимальной границей, нужной для запуска. Даже в худшем случае (левая граница интервала) доля заинтересованных клиентов как раз нужная. Можно рекомендовать продукт к запуску.

t-распределение Стьюдента

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:

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

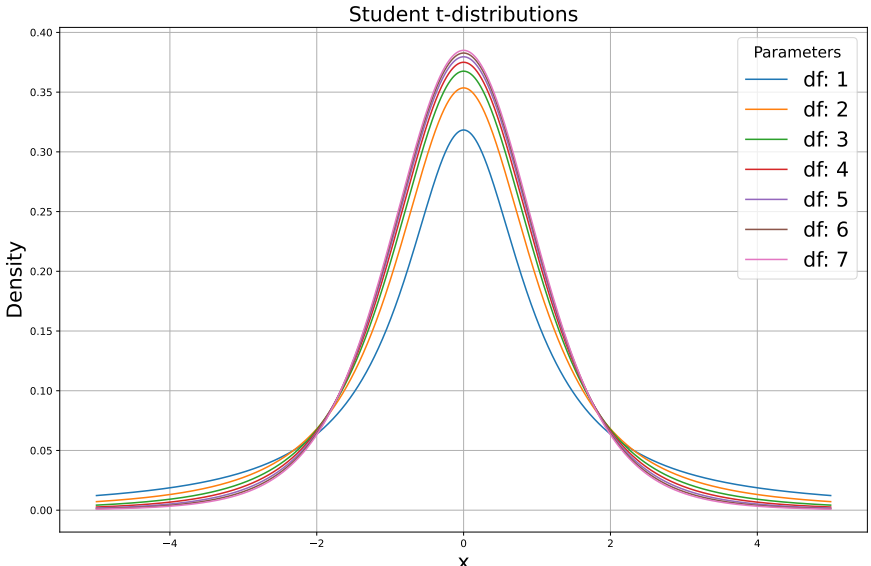
- В процессе сократились $(n-1)$ и стандартные отклонения σ, и мы получили изначальное утверждение.
- Эта форма t-распределения активно используется в статистике.

i Асимптотические свойства t-распределения

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

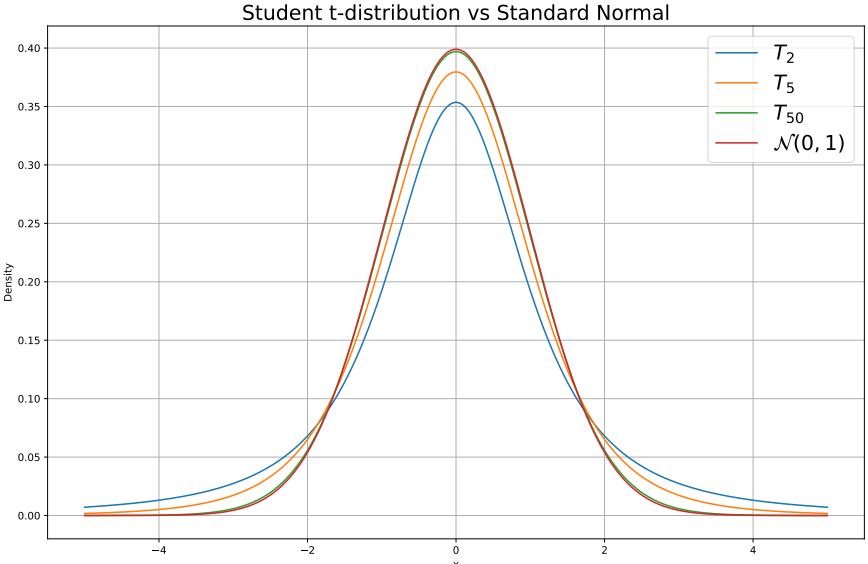
t-распределение Стьюдента

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.

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

- 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.

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?

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.