Теория вероятностей и математическая статистика

Доверительные интервалы III.

Глеб Карпов

ФКН ВШЭ

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:

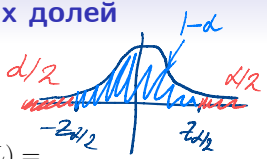
$$\hat{p}_1 \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right), \quad \hat{p}_2 \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right)$$

- 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:

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

$$\frac{\hat{\theta} - (p_1 - p_2)}{StD[\hat{\theta}]} \sim N(0,1)$$

Доверительные интервалы для разности истинных долей



- Если выполнены условия ИТМЛ, то действуем знакомым способом:

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

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

- 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

$1-\alpha = 0.95$

$\alpha = 0.05$

$z_{\alpha/2}: P(Z > z_{\alpha/2}) = 0.025 = 1 - F_Z(z_{\alpha/2})$

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.

X – Bern, покупка на версии A

Y – Bern, — // — B

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

$X \sim Bern(p_1)$

$Y \sim Bern(p_2)$

$\theta = p_1 - p_2$

Иллюстративные задачи

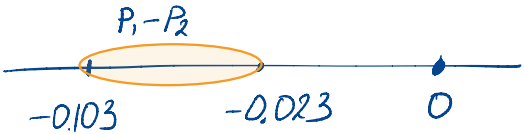
Решение

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

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

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

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:

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

$$\frac{\hat{\theta} - (\mu_X - \mu_Y)}{std[\hat{\theta}]} \sim N(0, 1)$$

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?

X – произв. линии 1, $E[X] \equiv \mu_X$, $Var[X] \equiv \sigma_X^2 \equiv \sigma_1^2$

Y – // – 2, $E[Y] \equiv \mu_Y$, $Var[Y] \equiv \sigma_Y^2 \equiv \sigma_2^2$

Иллюстративные задачи

$1-\alpha = 0.76$

$\alpha = 0.24$ $\quad \alpha/2 = 0.12 \Rightarrow$ $z_{\alpha/2}$: $P(Z > z_{\alpha/2}) = 0.12 = 1 - F_Z(z_{\alpha/2})$

$z_{\alpha/2} = 1.18$ $\quad (l, u) = (7 - 1.18 \cdot 3.17,\ 7 + 1.18 \cdot 3.17) = (3.26,\ 10.74)$

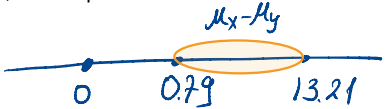
Решение

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.



Иллюстративные задачи

Пример 3: средний чек в ресторане

Менеджер ресторана хочет оценить среднюю сумму m, которую посетитель тратит на обед. По выборке из $n = 36$ посетителей получено выборочное среднее $\bar{x} = 3.60$ (в долларах). Известно, что стандартное отклонение расходов одного посетителя равно $\sigma = 0.72$.

Найдите уровень доверия, которому соответствует доверительный интервал $(3.5;\ 3.7)$ для m.

Иллюстративные задачи

Пример 4: разность долей конверсии

Сравнивают две рекламные кампании. В первой опросили $n = 30$ клиентов, из них 13 совершили покупку. Во второй — $m = 1.5n = 45$ клиентов, из них 15 совершили покупку. Построенный доверительный интервал для разности долей конверсии $p_1 - p_2$ равен $(-0.07;\ 0.27)$.

Найдите уровень доверия этого интервала.