

Теория вероятностей и математическая статистика  
Интервальные оценки II: интервалы для разности параметров.

Глеб Карпов

ВШБ Бизнес-информатика

# 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 нуля, 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  $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$ , 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 А показана 500 пользователям, из них 85 совершили покупку. Версия В показана 480 пользователям, из них 112 совершили покупку. Постройте 95% доверительный интервал для разности долей конверсии между версиями.

## Иллюстративные задачи

### Решение

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

$$\begin{aligned} p_1 - p_2 \in \left( 0.17 - 0.233 - 1.96\sqrt{\frac{0.17 \cdot 0.83}{500} + \frac{0.233 \cdot 0.767}{480}}, 0.17 - 0.233 + 1.96\sqrt{\frac{0.17 \cdot 0.83}{500} + \frac{0.233 \cdot 0.767}{480}} \right) \\ = (-0.103, -0.023) \end{aligned}$$

Интервал не содержит нуль, можем сделать вывод, что версия сайта В 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  $\mu_X$  — истинное математическое ожидание в первой генеральной совокупности,  $\mu_Y$  — истинное математическое ожидание во второй генеральной совокупности. Нас интересует **разность**  
 $\theta = \mu_X - \mu_Y$ .
- Важно знать, есть ли **статистически значимая разница** между математическими 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  $\sigma_X^2$  и  $\sigma_Y^2$  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  $\theta$ :

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

# 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  $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$ , 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  $\sigma_X^2$  и  $\sigma_Y^2$  — 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  $\sigma_X^2$  и  $\sigma_Y^2$  — 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.

## Точечная оценка разности математических ожиданий

Если дисперсии неизвестны, но предполагаются равными

- 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  $\sigma_X^2$  и  $\sigma_Y^2$  **неизвестны**, но для простоты предполагаем, что они равны:  $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ .
- Выборки должны быть либо:
  - обе большие ( $n > 30$ ,  $m > 30$ ), чтобы  $\bar{X}$  и  $\bar{Y}$  были нормально распределены по ЦПТ; либо
  - распределения генеральных совокупностей (хотя бы приблизительно) нормальные:  $X \sim \mathcal{N}(\mu_X, \sigma^2)$ ,  $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$ .
- 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:  $E[\hat{\theta}] = \mu_X - \mu_Y$ ,  $Var[\hat{\theta}] = \sigma^2 \left( \frac{1}{n} + \frac{1}{m} \right)$ .
- Распределение (если  $\sigma^2$  было бы известно):

$$\hat{\theta} \sim \mathcal{N} \left( \mu_X - \mu_Y, \sigma^2 \left( \frac{1}{n} + \frac{1}{m} \right) \right)$$

## Использование $t$ -распределения для разности матожиданий

Если дисперсии неизвестны, но предполагаются равными

- Проблема: мы не можем использовать  $\sigma^2$  при выводе границ интервала, так как дисперсия неизвестна!
- Решение: заменяем неизвестную дисперсию  $\sigma^2$  на её оценку — объединённую выборочную дисперсию  $S_p^2$ , и используем  $t$ -распределение вместо нормального.
- Вводим  $t$ -распределённую переменную:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{(n+m-2)}$$

- Объединённая дисперсия  $S_p^2$  — это взвешенное среднее выборочных дисперсий:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

- **Интуиция:** мы "объединяем" информацию о дисперсии из обеих выборок, используя веса, пропорциональные размерам выборок минус один (степени свободы). Идея в том, что чем больше размер выборки, тем точнее реализации выборочной дисперсии, и тем больше будет вес у этого слагаемого в сумме.
- Число степеней свободы:  $n + m - 2$  (сумма степеней свободы обеих выборок).

# Доверительные интервалы для разности математических ожиданий

Если дисперсии неизвестны, но предполагаются равными

- Если выполнены условия, то действуем знакомым способом:

$$1 - \alpha = P(L < \mu_X - \mu_Y < U) = P(-U < -(\mu_X - \mu_Y) < -L) =$$
$$P\left(\frac{\bar{X} - \bar{Y} - U}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} < \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} < \frac{\bar{X} - \bar{Y} - L}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}\right) = P\left(-t_{n+m-2, \alpha/2} < t_{(n+m-2)} < t_{n+m-2, \alpha/2}\right)$$

- Находим точку  $t_{n+m-2, \alpha/2}$ , такую, что:

$$P\left(t_{(n+m-2) df} > t_{(n+m-2, \alpha/2)}\right) = \frac{\alpha}{2}$$

случайная величина

константа, вещественное число

- Выполняем обратное преобразование, и находим теоретические границы такого доверительного интервала:

$$L = \bar{X} - \bar{Y} - t_{(n+m-2, \alpha/2)} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \quad U = \bar{X} - \bar{Y} + t_{(n+m-2, \alpha/2)} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

где  $S_p^2$  — объединённая дисперсия (случайная величина).



# Доверительные интервалы для разности математических ожиданий

Если дисперсии неизвестны, но предполагаются равными

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

$$\mu_X - \mu_Y \in \left( \bar{x} - \bar{y} - t_{(n+m-2, \alpha/2)} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{x} - \bar{y} + t_{(n+m-2, \alpha/2)} s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right),$$

где:

- $\bar{x}, \bar{y}, s_p$  — реализации выборочных средних и объединённого стандартного отклонения,
- $t_{(n+m-2, \alpha/2)}$  — критическая точка  $t$ -распределения с  $(n + m - 2)$  степенями свободы, такая что

$$P(t_{(n+m-2) df} > t_{(n+m-2, \alpha/2)}) = \frac{\alpha}{2}$$

случайная величина

константа, вещественное число

## Иллюстративные задачи

### Пример 3: Сравнение дневной прибыли двух магазинов

Сеть розничных магазинов хочет сравнить среднюю дневную прибыль двух своих магазинов. Были собраны данные за случайно выбранные дни:

- Магазин А: выборка из  $n = 25$  дней, средняя прибыль  $\bar{x} = 12500$  рублей, выборочное стандартное отклонение  $s_x = 1800$  рублей
  - Магазин В: выборка из  $m = 22$  дня, средняя прибыль  $\bar{y} = 11800$  рублей, выборочное стандартное отклонение  $s_y = 1950$  рублей
1. Постройте 95% доверительный интервал для разности средних дневных прибылей двух магазинов.
  2. Можете ли вы сделать вывод о том, какой магазин в среднем приносит больше прибыли?

## Иллюстративные задачи

### Решение

1.  $n = 25$ ,  $m = 22$ ,  $\bar{x} = 12500$ ,  $\bar{y} = 11800$ ,  $s_x = 1800$ ,  $s_y = 1950$ ,  $t_{45,0.025} \approx 2.014$ :

$$s_p^2 = \frac{(25 - 1) \cdot 1800^2 + (22 - 1) \cdot 1950^2}{25 + 22 - 2} = \frac{24 \cdot 3240000 + 21 \cdot 3802500}{45} \approx 3465000$$

$$s_p = \sqrt{3465000} \approx 1861$$

$$\mu_X - \mu_Y \in \left( 12500 - 11800 - 2.014 \cdot 1861 \cdot \sqrt{\frac{1}{25} + \frac{1}{22}}, 12500 - 11800 + 2.014 \cdot 1861 \cdot \sqrt{\frac{1}{25} + \frac{1}{22}} \right)$$

$$\approx (700 - 2.014 \cdot 1861 \cdot 0.292, 700 + 2.014 \cdot 1861 \cdot 0.292) = (-395, 1795)$$

2. Интервал  $(-395, 1795)$  содержит нуль, поэтому нельзя сделать статистически значимый вывод о том, какой магазин приносит больше прибыли.