Схема математической статистики

Схема математической статистики

Выборка: $x_1, \ldots, x_n$ Параметр: θ

$\hat{\theta} \longrightarrow f_{\hat{\theta}}(t)$

Точность оценки, прогнозов

доверительные интервалы

Ответы на вопросы

проверка гипотез

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Союзники

Асимптотические (при большом n)

- ЦПТ
- Дельта-метод

Точные

- Теорема Фишера
- $\chi^2_n, t_n, F_{n,k}$
- Ещё союзники!

Метод моментов

Метод моментов

$X_1, \dots, X_n$ одинаково независимо распределены $(\ iid\ )$

Момент $\mathbb{E}\left(X_i^k\right)$ зависит от неизвестного параметра θ:

$$\mathbb{E}\left(X_i^k\right) = f(\theta)$$

Оценкой метода моментов называется случайная величина:

$$\hat{\theta}_{MM} = f^{-1}(\overline{X^k})$$

То есть оценка получается решением уравнения

$$\mathbb{E}\left(X_i^k\right) \approx \frac{\sum x_i^k}{n}$$

Метод моментов

Чаще всего хватает первого момента и берут $k = 1$, то есть решают уравнение:

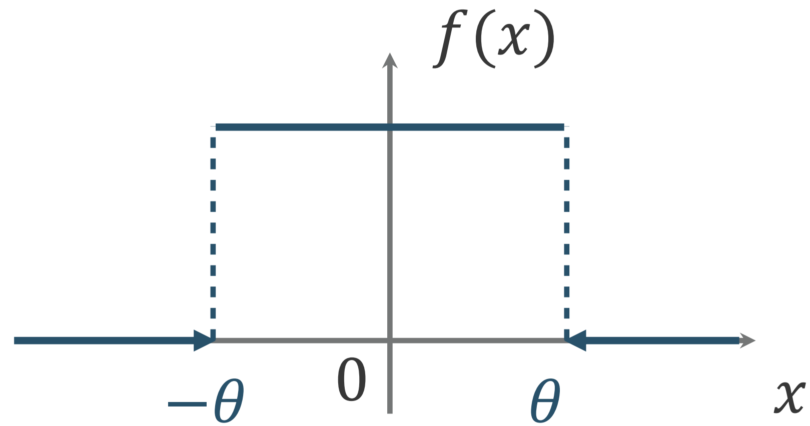
$$\mathbb{E}(X_i) \approx \frac{\sum x_i}{n}$$

Метод моментов

Если оказывается, что $\mathbb{E}(X_i) = 0$, тогда используют моменты более высоких порядков:

$$X_1, \dots, X_n \sim iid\ U[-\theta; \theta]$$

$$\mathbb{E}(X_i) = 0 \Rightarrow \bar{x} = 0$$



$$\mathbb{E}(X_i) = 0$$

Используя первый момент, нельзя получить оценку

$$\mathbb{E}\left(X_i^2\right) = \frac{\theta^2}{3} \Rightarrow \overline{x^2} = \frac{\theta^2}{3}$$

$$\Rightarrow \hat{\theta}_{MM} = \left(3\overline{x^2}\right)^{0.5}$$

Метод моментов

Если у распределения несколько параметров, используют несколько моментов:

$$X_1, \dots, X_n \sim iid\ N(\mu, \sigma^2)$$

Нужно оценить два параметра: дисперсию и математическое ожидание, используем два момента:

$$\begin{cases} \mathbb{E}(X_i) \approx \bar{x} \\ \mathbb{E}\left(X_i^2\right) \approx \overline{x^2} \end{cases} \Leftrightarrow \begin{cases} \mu = \bar{x} \\ \sigma^2 + \mu^2 = \overline{x^2} \end{cases} \Leftrightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2 \end{cases}$$

Резюме

- Метод моментов позволяет оценить параметр неизвестного распределения по выборке
- Перед тем, как использовать метод моментов, мы должны предположить, из какого распределения выборка была получена
- Обычно для оценки достаточно первого момента
- Если у распределения есть несколько параметров, используют несколько моментов

Асимптотический доверительный интервал для среднего

Схема математической статистики

Выборка: $x_1, \ldots, x_n$ Параметр: θ

$$\hat{\theta} \longrightarrow f_{\hat{\theta}}(t)$$

Точность оценки, прогнозов

доверительные интервалы

Ответы на вопросы

проверка гипотез

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Союзники

Асимптотические (при большом n)

- ЦПТ
- Дельта-метод

Точные

- Теорема Фишера
- $\chi^2_n, t_n, F_{n,k}$
- Ещё союзники!

Схема математической статистики

Выборка: $x_1, ..., x_n$ Параметр: θ

$$\hat{\theta} \longrightarrow f_{\hat{\theta}}(t)$$

Точность оценки, прогнозов

доверительные интервалы

Ответы на вопросы

проверка гипотез

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Союзники

Асимптотические (при большом n)

- ЦПТ
- Дельта-метод

Точные

- Теорема Фишера
- $\chi^2_n, t_n, F_{n,k}$
- Ещё союзники!

Мощь средних

ЗБЧ даёт нам возможность с помощью метода моментов построить оценку $\hat{\theta}_{MM}$

ЦПТ даёт нам информацию о распределении $\hat{\theta}_{MM}$,
мы можем построить доверительный интервал:

$$\mathbb{P}\left(\hat{\theta}_L \le \theta \le \hat{\theta}_R\right) = 1 - \alpha$$

$$\mathbb{P}\left(\hat{\theta}_L \le \theta \le \hat{\theta}_R\right) = 0.95$$

α – уровень значимости

Если мы 100 раз попытаемся сесть на поезд на уровне значимости 0.05, в среднем мы будем опаздывать 5 раз

Мощь средних

ЗБЧ даёт нам возможность с помощью метода моментов построить оценку $\hat{\theta}_{MM}$

ЦПТ даёт нам информацию о распределении $\hat{\theta}_{MM}$, мы можем построить доверительный интервал:

$$X_1, \ldots, X_n \sim iid$$

$$\mathbb{E}(X_i) = \mu,\ Var(X_i) = \sigma^2$$

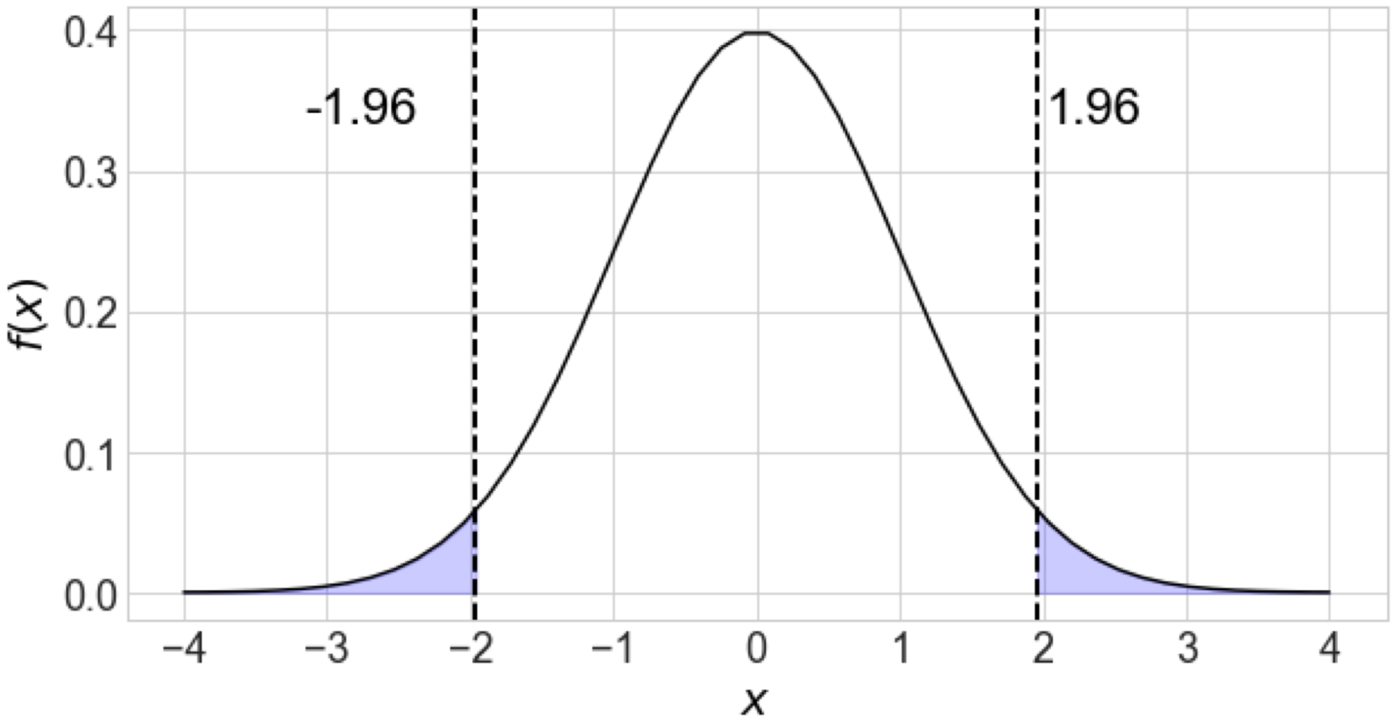
$$\hat{\mu} = \bar{x} \overset{asy}{\sim} N\left(\mu, \frac{\hat{\sigma}^2}{n}\right) \Leftrightarrow \bar{x} - \mu \overset{asy}{\sim} N\left(0, \frac{\hat{\sigma}^2}{n}\right) \Leftrightarrow \frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \overset{asy}{\sim} N(0, 1)$$

центрирование　　　　нормирование

Мощь средних

Вероятность того, что наша случайная величина окажется между -1.96 и 1.96 равна 0.95

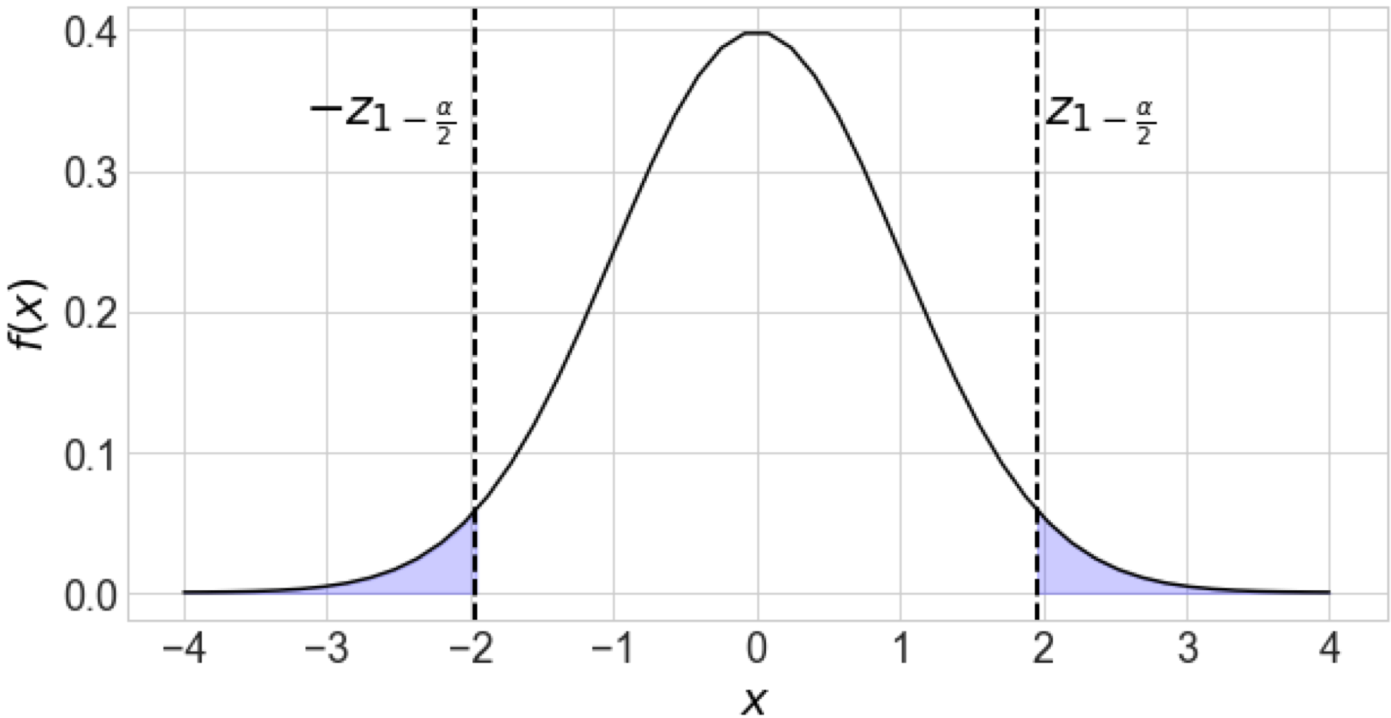
$$\frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \overset{asy}{\sim} N(0,1)$$



Мощь средних

Можно зафиксировать любую надежность $1 - \alpha$ и построить **доверительный интервал:**

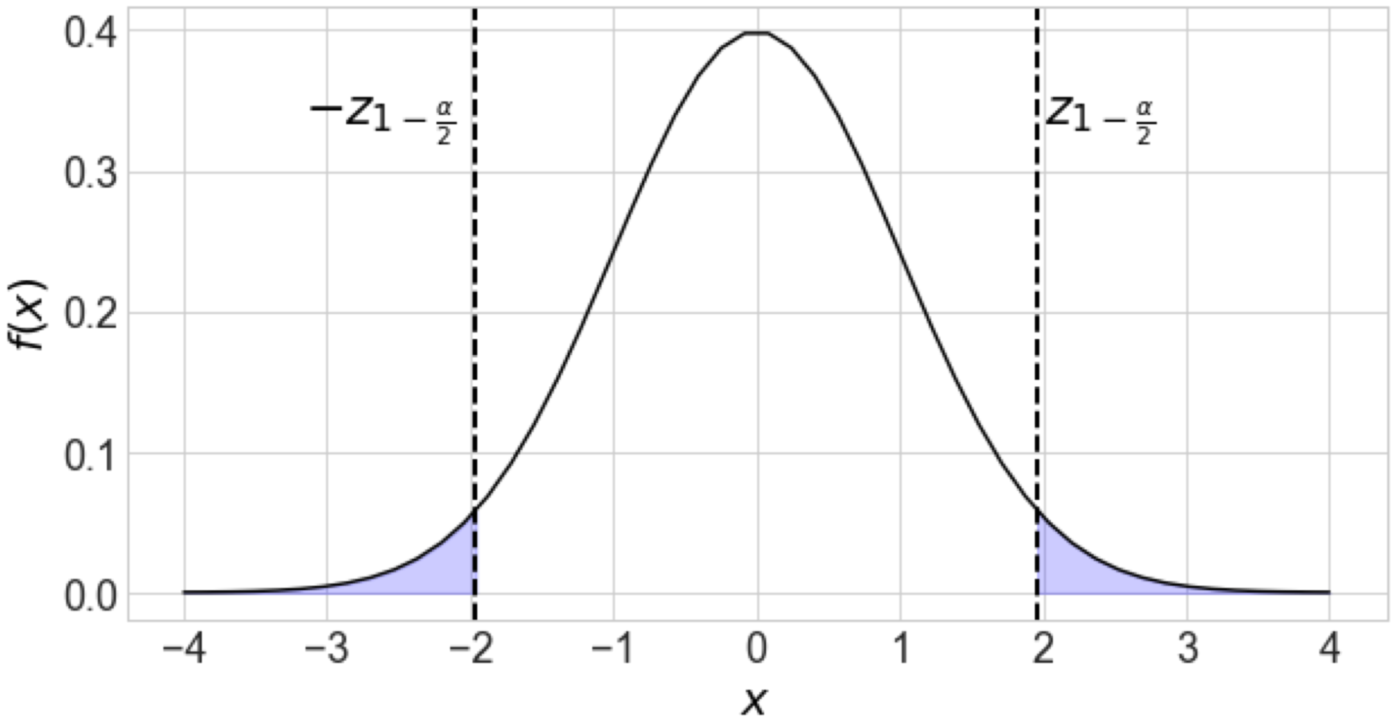
$$\frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \overset{asy}{\sim} N(0,1)$$



Мощь средних

Можно зафиксировать любую надежность $1 - \alpha$ и построить **доверительный интервал:**

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$



Мощь средних

Можно зафиксировать любую надежность $1 - \alpha$ и построить **доверительный интервал:**

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \le \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}} \le z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \le \bar{x} - \mu \le z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

$$\mathbb{P}\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

При бесконечном повторении эксперимента интервал будет накрывать истинное значение параметра μ в $100 \cdot (1 - \alpha)\%$ случаев

Мощь средних

$$\mathbb{P}\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

Иногда кратко пишут:

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

Мощь средних

Длина интервала:

$$\Delta = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

При росте n длина интервала падает

При росте дисперсии длина интервала увеличивается

При росте надёжности $1 - \alpha$ длина увеличивается

Резюме

- Центральная предельная теорема позволяет построить для среднего асимптотический доверительный интервал
- Доверительный интервал позволяет описать степень неуверенности в полученной оценке
- Такой доверительный интервал верен при большом количестве наблюдений, если в выборке нет аномалий

Асимптотический доверительный интервал для разницы средних

Разность средних

Цены на недвижимость в двух районах города:

$$X_1, \dots, X_n \sim iid \qquad Y_1, \dots, Y_m \sim iid$$

$$\bar{x} \overset{asy}{\sim} N\left(\mu_1, \frac{\sigma_1^2}{n}\right) \qquad \bar{y} \overset{asy}{\sim} N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

Разность нормальных случайных величин – нормальная случайная величина:

$$\mathbb{E}(\bar{x} - \bar{y}) = \mathbb{E}(\bar{x}) - \mathbb{E}(\bar{y}) = \mu_1 - \mu_2$$

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$$

$$\bar{x} - \bar{y} \overset{asy}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

Разность средних

Цены на недвижимость в двух районах города:

$$X_1, \dots, X_n \sim iid \qquad\qquad Y_1, \dots, Y_m \sim iid$$

$$\bar{x} - \bar{y} \overset{asy}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

Асимптотический доверительный интервал для $\mu_1 - \mu_2$:

$$(\bar{x} - \bar{y}) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}$$

Асимптотические доверительные интервалы для долей

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

$$X_1, \dots, X_n \sim iid \qquad X_i = \begin{cases} 1, \text{если любит кофе} \\ 0, \text{если не любит кофе} \end{cases}$$

X_i	0	1
$\mathbb{P}(X_i = k)$	$1 - p$	p

$$\hat{p} = \frac{X_1 + \dots + X_n}{n} = \bar{x}$$

Из-за того, что X_i принимают значение либо 0, либо 1, для оценки доли можно посчитать среднее

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

$$\hat{p} = \frac{X_1 + \dots + X_n}{n} = \bar{x}$$

X_i	0	1
$\mathbb{P}(X_i = k)$	$1 - p$	p

Найдём математическое ожидание и дисперсию оценки, а потом воспользуемся ЦПТ

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

$$\hat{p} = \frac{X_1 + \dots + X_n}{n} = \bar{x}$$

X_i	0	1
$\mathbb{P}(X_i = k)$	$1 - p$	p

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n} \cdot n \cdot \mathbb{E}(X_1) = p$$

$$Var(\hat{p}) = Var\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} \cdot n \cdot Var(X_1) = \frac{p(1-p)}{n}$$

$$\bar{x} \overset{asy}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

$$\hat{p} = \frac{X_1 + \dots + X_n}{n} = \bar{x}$$

X_i	0	1
$\mathbb{P}(X_i = k)$	$1 - p$	p

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n} \cdot n \cdot \mathbb{E}(X_1) = p$$

$$Var(\hat{p}) = Var\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} \cdot n \cdot Var(X_1) = \frac{p(1-p)}{n}$$

$$\bar{x} \overset{asy}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \hat{p} = \bar{x} \overset{asy}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Мощь долей

Получаем доверительный интервал для доли:

$$\bar{x} \overset{asy}{\sim} N\left(\mu, \frac{\hat{\sigma}^2}{n}\right) \qquad \hat{p} = \bar{x} \overset{asy}{\sim} N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Получаем доверительный интервал для разности долей:

$$\bar{x} - \bar{y} \overset{asy}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

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

Мощь долей

Получаем доверительный интервал для доли:

$$\bar{x} \overset{asy}{\sim} N\left(\mu, \frac{\hat{\sigma}^2}{n}\right) \qquad \hat{p} = \bar{x} \overset{asy}{\sim} N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Получаем доверительный интервал для разности долей:

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

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$$

Число наблюдений

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Можно определить число наблюдений, чтобы длина доверительного интервала не превышала заранее выбранный диапазон

$$\Delta = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\mathrm{n} = \frac{4 \cdot z_{1-\frac{\alpha}{2}}^{2} \cdot \hat{p}(1-\hat{p})}{\Delta^{2}}$$

Число наблюдений

$$n = \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \cdot \hat{p}(1-\hat{p})}{\Delta^2}$$

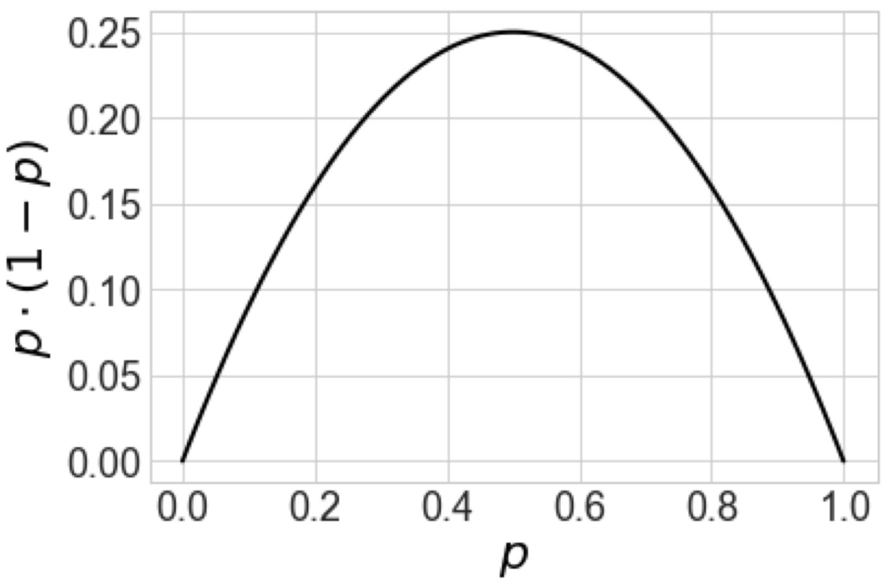
До начала испытаний мы не знаем $\hat{p}$, но мы знаем, что величина $\hat{p}(1-\hat{p})$ никогда не будет превышать 0.25

$$f(p) = p \cdot (1 - p) = p - p^2$$

$$f'(p) = 1 - 2p = 0$$

$$\Rightarrow p = 0.5$$

$$f(p) = 0.25$$



Число наблюдений

$$\mathrm{n} = \frac{4 \cdot z_{1-\frac{\alpha}{2}}^{2} \cdot \hat{p}(1-\hat{p})}{\Delta^2}$$

До начала испытаний мы не знаем $\hat{p}$, но мы знаем, что величина $\hat{p}(1-\hat{p})$ никогда не будет превышать 0.25

Эту оценку сверху мы можем использовать для поиска необходимого значения n:

$$\mathrm{n} = \frac{4 \cdot z_{1-\frac{\alpha}{2}}^{2} \cdot \hat{p}(1-\hat{p})}{\Delta^2} \leq \frac{z_{1-\frac{\alpha}{2}}^{2}}{\Delta^2}$$

Резюме

- Доля – это среднее, посчитанное по выборке из нулей и единиц
- С помощью ЦПТ можно построить доверительные интервалы для долей
- Из-за того, что вероятность принимает значения на отрезке от нуля до единицы, мы можем оценить, сколько наблюдений нам нужно собрать для определённой ширины интервала