Лекция 5

Асимптотический доверительный интервал для доли

Определение объёма выборки

Асимптотический доверительный интервал

Последовательности случайных величин $\{L_n\}$ и $\{U_n\}$ образуют асимптотический доверительный интервал для параметра θ с уровнем доверия γ, если

$$\lim_{n\to\infty} P\left(L_n < \theta < U_n\right) = \gamma, \quad \forall\theta\in\Theta.$$

зачем это нужно?

Асимптотический доверительный интервал для доли

Пусть $X_1, \ldots, X_n$ независимы,

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Хотим получить доверительный — интервал для p (доли единичек в ген. совокупности).

Уровень доверия: 1-α (α — вероятность ошибки)

Вспоминаем точечную оценку:

$$\hat{p} = \frac{X_1 + \ldots + X_n}{n}.$$

Распределение выборочной доли:

$$\boxed{\begin{gathered} \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{asy}}{\sim} N(0,1). \\[2em] \hat{p} \overset{\text{app}}{\sim} N\left(p, \frac{p(1-p)}{n}\right). \end{gathered}}$$

Асимптотический доверительный интервал для доли

Пусть $X_1,\ldots, X_n$ независимы,

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Хотим получить доверительный — интервал для p (доли единичек в ген. совокупности).
Уровень доверия: 1-α (α — вероятность ошибки)

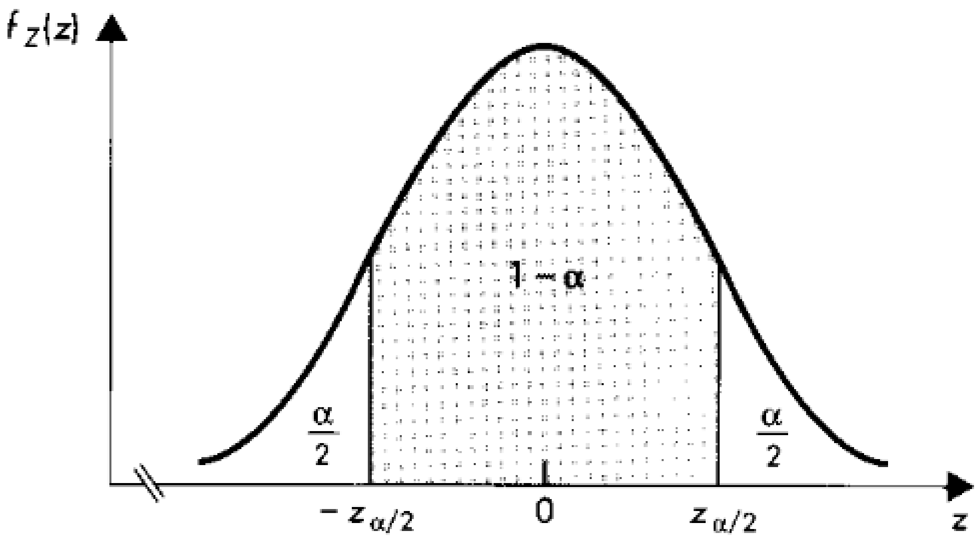
Вспоминаем точечную оценку: $\hat{p} = \dfrac{X_1+\ldots+X_n}{n}$.

Будем считать, что $Z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1)$.

Возьмём такое число $z_{\frac{\alpha}{2}}$, что $P\left(Z>z_{\frac{\alpha}{2}}\right)=\dfrac{\alpha}{2}$.

Тогда $1-\alpha = P\left(-z_{\frac{\alpha}{2}}<Z<z_{\frac{\alpha}{2}}\right) =$

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



Асимптотический доверительный интервал для доли

Пусть $X_1, \ldots, X_n$ независимы,

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

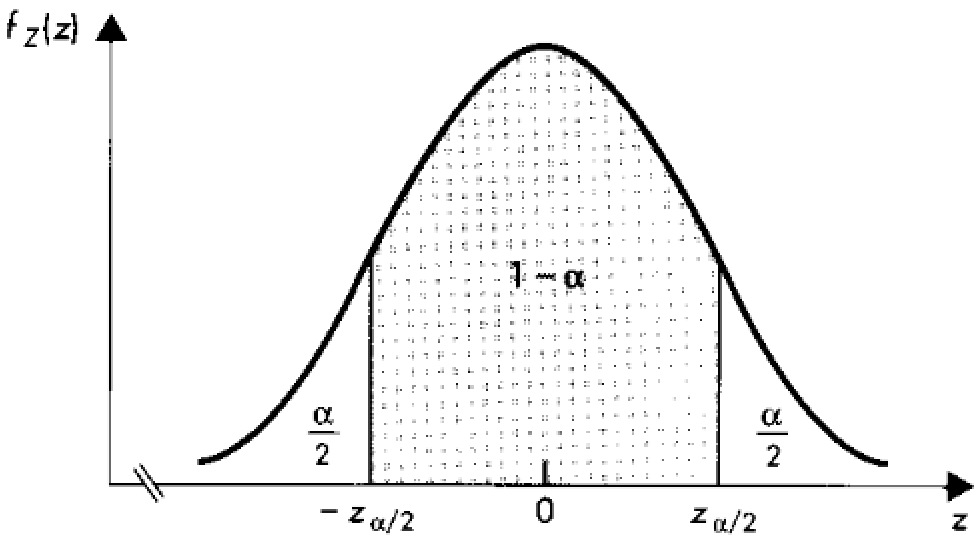
Хотим получить доверительный — интервал для p (доли единичек в ген. совокупности).
Уровень доверия: 1-α (α — вероятность ошибки)

Вспоминаем точечную оценку: $\hat{p} = \dfrac{X_1 + \ldots + X_n}{n}$.

Будем считать, что $Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1)$.

Возьмём такое число $z_{\frac{\alpha}{2}}$, что $P\left(Z > z_{\frac{\alpha}{2}}\right) = \dfrac{\alpha}{2}$.

Тогда $1 - \alpha = P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) =$



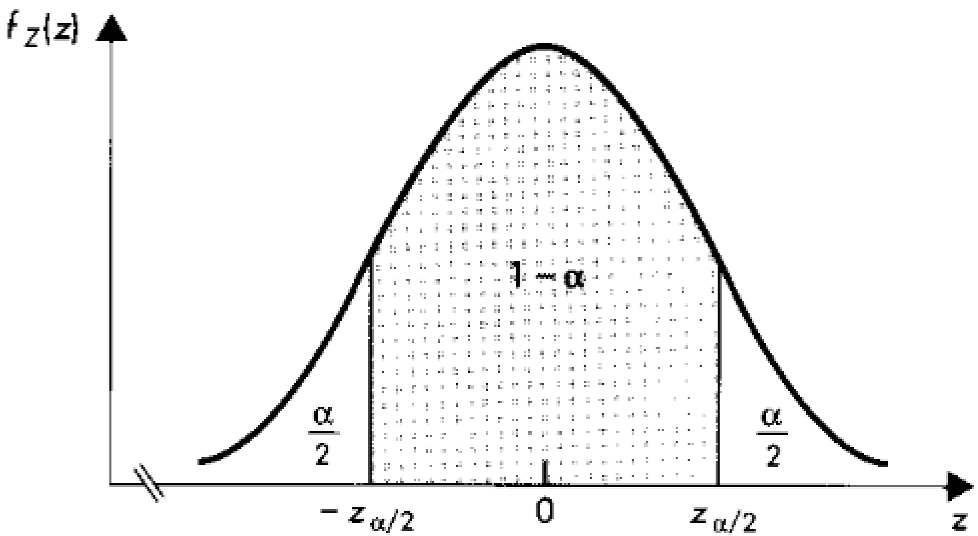
Асимптотический доверительный интервал для доли

Пусть $X_1,\ldots,X_n$ независимы,

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Хотим получить доверительный — интервал для p (доли единичек в ген. совокупности).
Уровень доверия: 1-α (α — вероятность ошибки)

Вспоминаем точечную оценку: $\hat{p} = \dfrac{X_1+\ldots+X_n}{n}$.



Будем считать, что $Z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1)$.

Возьмём такое число $z_{\frac{\alpha}{2}}$, что $P\left(Z > z_{\frac{\alpha}{2}}\right) = \dfrac{\alpha}{2}$.

Тогда $1-\alpha = P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) =$

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

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

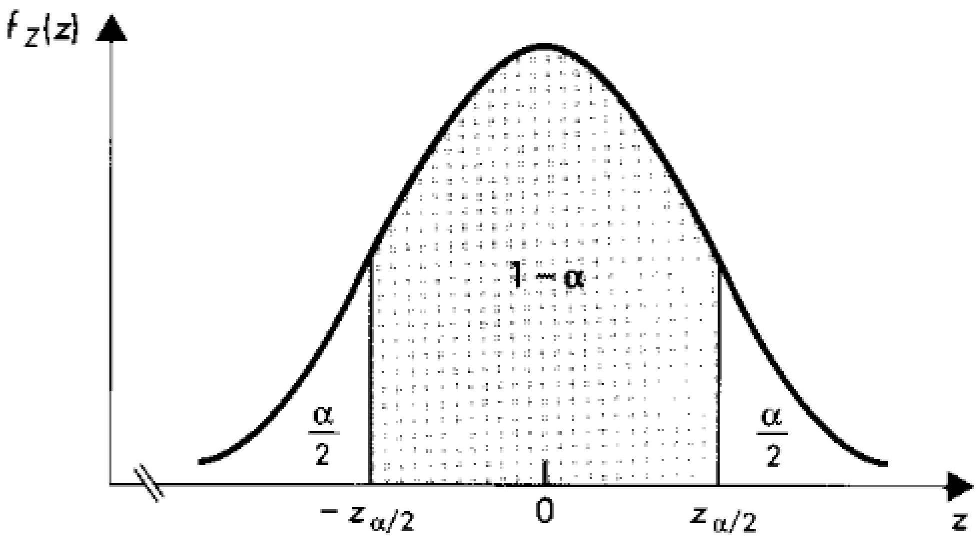
Асимптотический доверительный интервал для доли

Пусть $X_1,\ldots, X_n$ независимы,

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}.$$

Хотим получить доверительный — интервал для p (доли единичек в ген. совокупности).
Уровень доверия: 1-α (α — вероятность ошибки)

Вспоминаем точечную оценку: $\hat{p} = \dfrac{X_1+\ldots+X_n}{n}.$



Будем считать, что $Z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1).$

Возьмём такое число $z_{\frac{\alpha}{2}}$, что $P\left(Z > z_{\frac{\alpha}{2}}\right) = \dfrac{\alpha}{2}.$

Тогда $1-\alpha = P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) =$

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

$$= P\left(-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < \hat{p}-p < z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = P\left(\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right).$$

Асимптотический доверительный интервал для доли

Получили:

$$1-\alpha \approx P\left(\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right).$$

почему?

К сожалению, интервал $\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$ совершенно бесполезен.

Асимптотический доверительный интервал для доли

Получили:

$$1-\alpha \approx P\left(\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right).$$

почему?

К сожалению, интервал $\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$ совершенно бесполезен.

Можно подставить оценённую долю в границы интервала:

$$\boxed{\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}<p<\hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Это и будет асимптотический доверительный интервал для доли с уровнем доверия 1-α.

Примечания.

► Этот интервал имеет смысл использовать при больших объёмах выборки (>100).

► На самом деле, известен и точный (не асимптотический) доверительный интервал для доли.

► Более строгий вывод интервала можно найти в книге А.С. Шведова «Теория вероятностей и математическая статистика — 2».

Пример

В выборке из 240 жителей Твери 90 человек оказались голубоглазыми. Рассчитайте 99% доверительный интервал для доли голубоглазых среди всех тверяков.

Пример

В выборке из 240 жителей Твери 90 человек оказались голубоглазыми. Рассчитайте 99% доверительный интервал для доли голубоглазых среди всех тверяков.

Решение.

Из условия $n = 240, \quad \hat{p} = \frac{90}{240} = 0.375.$

Из таблицы $z_{\frac{\alpha}{2}} = 2.576.$

Левая граница доверительного интервала:

$$\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} = 0.375 - 2.576 \sqrt{\frac{0.375\left(1-0.375\right)}{240}} = 0.2945.$$

Правая граница доверительного интервала:

$$\hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} = 0.375 + 2.576 \sqrt{\frac{0.375\left(1-0.375\right)}{240}} = 0.4555.$$

Вывод. Доля голубоглазых среди тверяков лежит в пределах от 29.5% до 45.6%.

А если интервал кажется слишком широким?

Либо уменьшаем доверительную вероятность, либо увеличиваем объём выборки.

Определение объёма выборки

Итак, мы рассмотрели задачи оценивания параметров генеральной совокупности, в т.ч. точечные оценки и доверительные интервалы для:

- ▶ среднего,
- ▶ дисперсии,
- ▶ доли.

На самом деле, перед оцениванием разумно задаться вопросом: как много данных нужно собрать, чтобы оценить параметр с нужной мне точностью?

Точнее:

▶ каким должен быть объём выборки, чтобы рассчитанная по этой выборке точечная оценка $\hat{\theta}$ параметра θ отличалась от оцениваемого параметра более чем на b с вероятностью, не превышающей α :

$$P\left(|\hat{\theta}-\theta|>b\right)\leq\alpha,$$

или

$$P\left(|\hat{\theta}-\theta|\leq b\right)\geq 1-\alpha.$$

Оценивание среднего при известной дисперсии

нормальная генеральная совокупность

Пусть $X_1, \ldots, X_n$ независимы, $X_i \sim N(\mu, \sigma^2)$.

Выборочное среднее: $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$.

Задача. Найти число наблюдений n, при котором для допустимой величины ошибки b и вероятности ошибки α выполняется:

$$P\left(|\bar{X} - \mu| \le b\right) \ge 1 - \alpha.$$

Оценивание среднего при известной дисперсии

нормальная генеральная совокупность

Пусть $X_1, \ldots, X_n$ независимы, $X_i \sim N(\mu, \sigma^2)$.

Выборочное среднее: $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$.

Задача. найти число наблюдений n, при котором для допустимой величины ошибки b и вероятности ошибки α выполняется:

$$P(|\bar{X} - \mu| \le b) \ge 1 - \alpha.$$

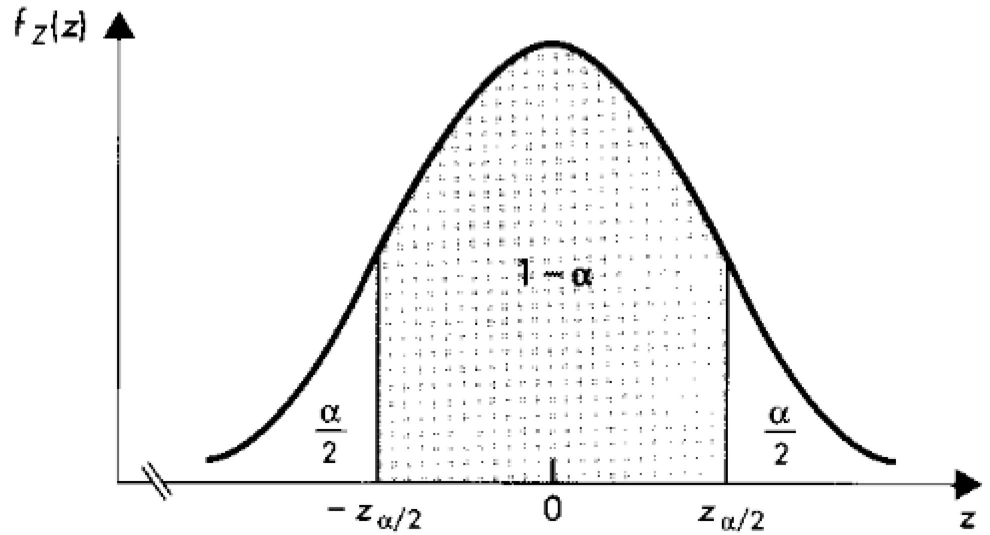
Решение. $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

Центрируем и нормируем: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Попробуем найти n, в точности обеспечивающее допустимую вероятность ошибки:

$$1 - \alpha = P(|\bar{X} - \mu| \le b) = P(-b \le \bar{X} - \mu \le b) = P\left(-\frac{b}{\sigma/\sqrt{n}} \le Z \le \frac{b}{\sigma/\sqrt{n}}\right).$$

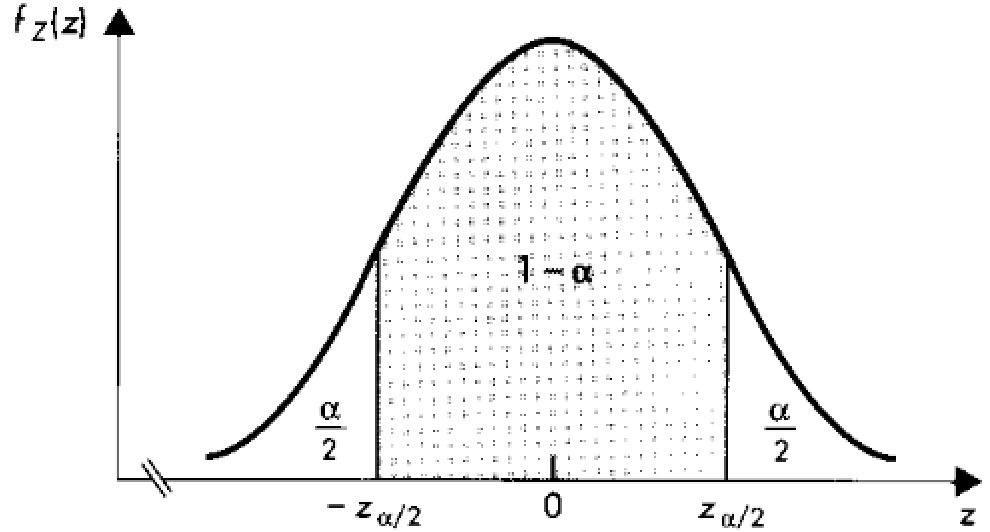
$$1-\alpha = P\left(-\frac{b}{\sigma/\sqrt{n}} \le Z \le \frac{b}{\sigma/\sqrt{n}}\right).$$



Таким образом,

$$\frac{b}{\sigma/\sqrt{n}} = z_{\frac{\alpha}{2}}.$$

$$1-\alpha = P\left(-\frac{b}{\sigma/\sqrt{n}} \le Z \le \frac{b}{\sigma/\sqrt{n}}\right).$$



Таким образом,

$$\frac{b}{\sigma/\sqrt{n}} = z_{\frac{\alpha}{2}}.$$

Решаем уравнение относительно n:

$$\frac{b\sqrt{n}}{\sigma} = z_{\frac{\alpha}{2}} \Rightarrow \sqrt{n} = \frac{z_{\frac{\alpha}{2}}\sigma}{b}.$$

$$n = \frac{z_{\frac{\alpha}{2}}^2 \sigma^2}{b^2}.$$

Это n, скорее всего, дробное => округляем вверх:

$$\boxed{n = \left\lceil \frac{z_{\frac{\alpha}{2}}^2 \sigma^2}{b^2} \right\rceil.}$$

Пример

(Ратникова, Шведов, «Сборник задач по теории вероятностей и математической статистике», 2004)

Вес коробки сахара нормально распределён со стандартным отклонением 4 грамма. Из генеральной совокупности отбираются n коробок.

а) Пусть $n = 16$. Какова вероятность, что ошибка при определении среднего будет больше, чем 2 грамма?

б) Каким должен быть объём выборки, чтобы вероятность получения ошибки в один грамм или больше не превосходила 0.0455?

Пример

(Ратникова, Шведов, «Сборник задач по теории вероятностей и математической статистике», 2004)

Вес коробки сахара нормально распределён со стандартным отклонением 4 грамма. Из генеральной совокупности отбираются n коробок.

а) Пусть n = 16. Какова вероятность, что ошибка при определении среднего будет больше, чем 2 грамма?

б) Каким должен быть объём выборки, чтобы вероятность получения ошибки в один грамм или больше не превосходила 0.0455?

Решение.

Пусть X_i — вес i-й коробки сахара. Будем считать, что $X_1, \ldots, X_n$ независимы.

Выборочное среднее $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} = \frac{4^2}{n}\right)$.

а) n = 16, так что $\bar{X} \sim N(\mu, 1)$.

Пусть $Z = \frac{\bar{X} - \mu}{1}$. Тогда $Z \sim N(0, 1)$.

$$P(|\bar{X} - \mu| > 2) = P(|Z| > 2) = 1 - P(-2 < Z < 2) = 1 - \underbrace{0.9545}_{\text{из таблицы}} = 0.0455.$$

Пример

(Ратникова, Шведов, «Сборник задач по теории вероятностей и математической статистике», 2004)

Вес коробки сахара нормально распределён со стандартным отклонением 4 грамма. Из генеральной совокупности отбираются n коробок.

а) Пусть n = 16. Какова вероятность, что ошибка при определении среднего будет больше, чем 2 грамма?

б) Каким должен быть объём выборки, чтобы вероятность получения ошибки в один грамм или больше не превосходила 0.0455?

Решение.

Пусть X_i — вес i-й коробки сахара. Будем считать, что $X_1, \ldots, X_n$ независимы.

Выборочное среднее $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} = \frac{4^2}{n}\right).$

б) Нужно найти такое n, что $P\left(|\bar{X} - \mu| \geq 1\right) \leq 0.0455.$

В общем виде: $n = \left\lceil \frac{z^2_{\frac{\alpha}{2}} \sigma^2}{b^2} \right\rceil.$

В настоящей задаче: $\sigma = 4, \; b = 1, \; \alpha = 0.0455 \;\Rightarrow\; z_{\frac{\alpha}{2}} = 2.$ *(см. пункт (а) и таблицы)*

Объём выборки: $n = \left\lceil \frac{2^2 \times 4^2}{1^2} \right\rceil = \lceil 64 \rceil = 64.$

Вывод: нужно отобрать 64 коробки. *(можно больше, но незачем)*

Оценивание доли

Пусть $X_1, ..., X_n$ независимы, $X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$.

Выборочная доля: $\hat{p} = \dfrac{X_1 + ... + X_n}{n}$.

Задача. Найти число наблюдений n, при котором для допустимой величины ошибки b и вероятности ошибки α выполняется:

$$P\left(|\hat{p} - p| \le b\right) \ge 1 - \alpha.$$

Оценивание доли

Пусть $X_1, \dots, X_n$ независимы, $X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$.

Выборочная доля: $\hat{p} = \dfrac{X_1 + \dots + X_n}{n}$.

Задача. Найти число наблюдений n, при котором для допустимой величины ошибки b и вероятности ошибки α выполняется:

$$P\left(|\hat{p} - p| \le b\right) \ge 1 - \alpha.$$

Решение. Опираемся на нормальное приближение:

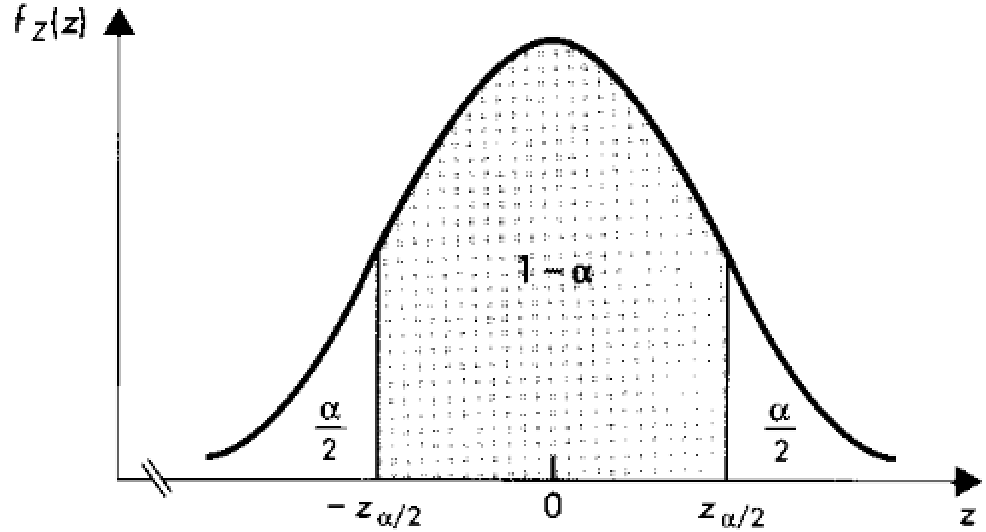
$$\hat{p} \overset{asy}{\sim} N\left(p, \frac{p(1-p)}{n}\right).$$

Центрируем и нормируем: $$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \overset{asy}{\sim} N(0,1).$$

Дальше всё как обычно:

$$1 - \alpha = P\left(|\hat{p} - p| \le b\right) = P\left(-b \le \hat{p} - p \le b\right) = P\left(-\frac{b}{\sqrt{\dfrac{p(1-p)}{n}}} \le Z \le \frac{b}{\sqrt{\dfrac{p(1-p)}{n}}}\right).$$

$$1-\alpha = P\left(-\frac{b}{\sqrt{\frac{p(1-p)}{n}}} \le Z \le \frac{b}{\sqrt{\frac{p(1-p)}{n}}}\right).$$



Таким образом,

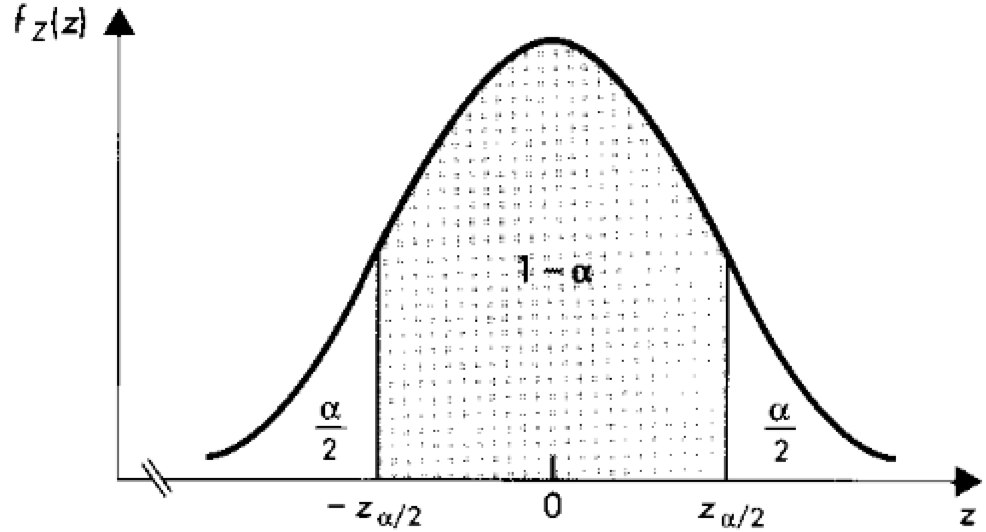
$$\frac{b}{\sqrt{\frac{p(1-p)}{n}}} \approx z_{\frac{\alpha}{2}}.$$

Решаем уравнение относительно n:

$$\frac{b\sqrt{n}}{\sqrt{p(1-p)}} = z_{\frac{\alpha}{2}} \;\Rightarrow\; \sqrt{n} = \frac{z_{\frac{\alpha}{2}}\sqrt{p(1-p)}}{b}.$$

$$n = \frac{z_{\frac{\alpha}{2}}^2\, p(1-p)}{b^2}.$$

$$1-\alpha = P\left(-\frac{b}{\sqrt{\frac{p(1-p)}{n}}} \le Z \le \frac{b}{\sqrt{\frac{p(1-p)}{n}}}\right).$$



Таким образом,

$$\frac{b}{\sqrt{\frac{p(1-p)}{n}}} \approx z_{\frac{\alpha}{2}}.$$

Решаем уравнение относительно n:

$$\frac{b\sqrt{n}}{\sqrt{p(1-p)}} = z_{\frac{\alpha}{2}} \;\Rightarrow\; \sqrt{n} = \frac{z_{\frac{\alpha}{2}}\sqrt{p(1-p)}}{b}.$$

$$n = \frac{z_{\frac{\alpha}{2}}^2\, p(1-p)}{b^2}.$$

Вот только p мы не знаем. Мы пытаемся его найти.

Опять засада.

Решение: ориентироваться на худшее значение p.

Мы не можем найти $n = \dfrac{z_{\frac{\alpha}{2}}^2 p(1-p)}{b^2}$.

Зато мы можем найти n такое, что $n \geq \dfrac{z_{\frac{\alpha}{2}}^2 p(1-p)}{b^2}$.

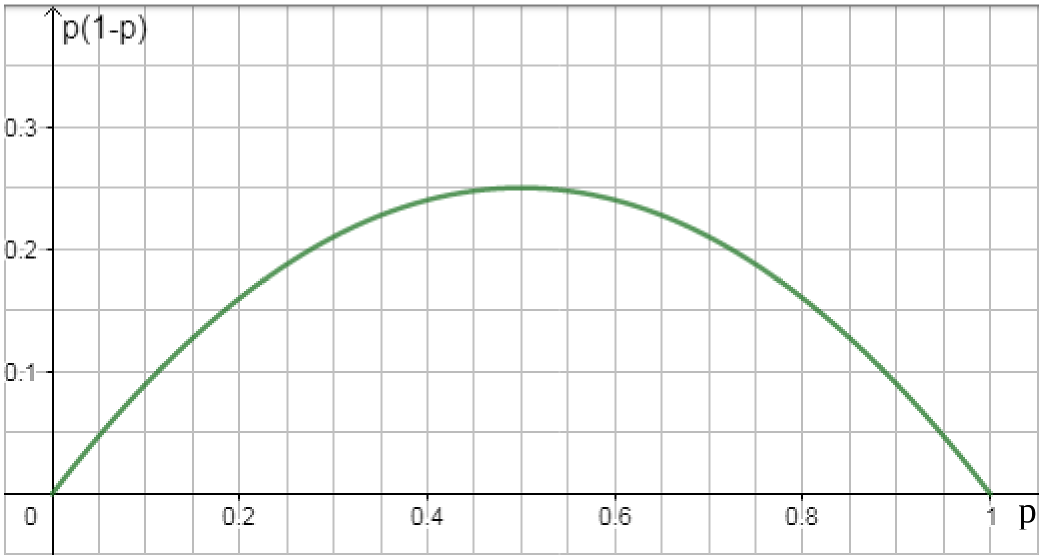
Как?

Решение: ориентироваться на худшее значение p.

Мы не можем найти $n = \dfrac{z^2_{\frac{\alpha}{2}}\, p(1-p)}{b^2}$.

Зато мы можем найти n такое, что $n \geq \dfrac{z^2_{\frac{\alpha}{2}}\, p(1-p)}{b^2}$.

$p(1-p)$ — парабола.



Её вершина — точка $\left(p=\dfrac{1}{2}, p(1-p)=\dfrac{1}{4}\right)$.

Значит, $p(1-p) \leq \dfrac{1}{4}$.

Получаем объём выборки:

$$\frac{z^2_{\frac{\alpha}{2}}\, p(1-p)}{b^2} \leq \frac{z^2_{\frac{\alpha}{2}}}{4b^2},$$

так что достаточно взять

$$\boxed{n = \left\lceil \frac{z^2_{\frac{\alpha}{2}}}{4b^2} \right\rceil.}$$

Пример

Перед выборами планируется проведение опроса населения с целью выяснить уровень поддержки кандидата А. Сколько человек нужно опросить, чтобы с вероятностью 95% доля поддерживающих кандидата А в выборке отличалась от доли в генеральной совокупности не более чем на 0.06?

Пример

Перед выборами планируется проведение опроса населения с целью выяснить уровень поддержки кандидата А. Сколько человек нужно опросить, чтобы с вероятностью хотя бы 95% доля поддерживающих кандидата А в выборке отличалась от доли в генеральной совокупности не более чем на 0.06?

Решение.

Нужно найти число наблюдений n, которое обеспечивает выполнение неравенства:

$$P\left(|\hat{p}-p|\leq 0.06\right)\geq 0.95.$$

Общий вид: $n=\left\lceil \frac{z_{\frac{\alpha}{2}}^2}{4b^2}\right\rceil.$

В данном случае: $b=0.06,\ \ \alpha=1-0.95=0.05\ \Rightarrow\ z_{\frac{\alpha}{2}}=z_{\frac{0.05}{2}}=1.96.$

Значит, $n=\left\lceil \frac{1.96^2}{4\times 0.06^2}\right\rceil=\lceil 266.78\rceil=277.$

Ответ: достаточно опросить 277 человек.

Следующая лекция:

проверка гипотез.