ОП «Политология», 2023-24
Введение в ТВиМС
Доверительные интервалы (памятка)
А. А. Макаров, А. А. Тамбовцева

Доверительный интервал для доли

Постановка задачи

Представьте, что нам необходимо оценить долю «жаворонков» – любителей рано вставать по утрам, среди всех жителей России. Всех жителей России опросить не получится, но мы можем случайным образом выбрать 1000 человек, провести опрос и выяснить, сколько среди них «жаворонков». Допустим, в этой выборке оказалось 170 респондентов, любящих рано вставать по утрам, интересующая нас выборочная доля равна 0.17. Устроит ли нас такая оценка?

С одной стороны, устроит, выборка случайная и большая по объёму. С другой стороны, какая бы выборка не была, мы не можем однозначно утверждать, что 17% россиян любят вставать по утрам, потому что, когда мы оцениваем долю по одной единственной выборке, мы получаем значение с некоторой погрешностью и при этом понимаем, что на другой выборке такого же объёма другой исследователь может получить другой результат.

Что же в таком случае сделать? Зафиксировать уровень уверенности в расчётах и вместо одного значения для доли определить интервал, в пределах которого эта доля может находиться. Другими словами, построить **доверительный интервал** для интересующей нас доли.

Составные части доверительного интервала

Чтобы не запутаться, введём следующие обозначения:

- p – доля «жаворонков» в генеральной совокупности, то есть среди всех жителей России, её мы не знаем, но хотим оценить, как раз построив доверительный интервал на основе какой-нибудь одной выборки;
- $\hat{p}$ – доля «жаворонков» в конкретной выборке, выборочная доля, которую мы получаем по итогам проведённого опроса, например, в выборке из 1000 человек.

Доля «жаворонков» $\hat{p}$ в выборке – случайная величина, мы не знаем, какую долю получим по итогам опроса. Каждый исследователь на новой выборке такого же объёма будет получать что-то своё, поскольку люди в его выборку попадут уже другие. Получается, выборочная доля изменчива, и степень этой изменчивости можно оценить. И поможет нам в этом центральная предельная теорема.

Согласно центральной предельной теореме, выборочная доля имеет следующее распределение:

$$\hat{p} \sim N(p, \sigma^2 = \frac{pq}{n}) \text{ или } \hat{p} \sim N(p, \sigma = \frac{\sqrt{pq}}{\sqrt{n}}),$$

где p – интересующая нас доля в генеральной совокупности (вероятность успеха), $q = 1 - p$ – обратная к ней доля (вероятность неудачи), n – объём выборки.

Используя эту теорему, мы можем понять, чему равно стандартное отклонение доли, то есть понять, насколько, в среднем, доля изменяется от выборки к выборке. Однако обычно значения p и q нам неизвестны, так как мы ничего не знаем про генеральную совокупность. Как быть? Приблизить их с помощью того, что нам известно, а известны нам доли, посчитанные на основе имеющейся выборки:

$$p \approx \hat{p};$$
$$q \approx \hat{q} = 1 - \hat{p}.$$

Подставляем эти значения в выражение для стандартного отклонения σ, и получаем **стандартную ошибку доли** (от *standard error*), оценку стандартного отклонения доли, полученную на основе выборки:

$$\text{se} = \frac{\sqrt{\hat{p}\hat{q}}}{\sqrt{n}}.$$

Возьмём данные из нашего примера про «жаворонков» и вычислим стандартную ошибку доли ($n = 1000$, $\hat{p} = 0.17$, $\hat{q} = 0.17$):

$$\text{se} = \frac{\sqrt{0.17 \cdot 0.83}}{\sqrt{1000}} \approx 0.01.$$

Итак, теперь мы можем сообщить более общую информацию, сказать, что доля «жаворонков» не просто равна 0.17, а равна 0.17 ± 0.01, допуская, что доля от выборки к выборке может меняться. Это значение стандартной ошибки доли мы можем использовать и для более широких интервалов, используя правило трёх сигм:

- в 68% случаев выборочная доля лежит в интервале $[\hat{p} - \text{se}; \hat{p} - \text{se}]$, т.е. в интервале [0.16; 0.18];
- в 95% случаев выборочная доля лежит в интервале $[\hat{p} - 2 \times \text{se}; \hat{p} + 2 \times \text{se}]$ т.е. в интервале [0.15; 0.19];
- в 99.8% случаев выборочная доля лежит в интервале $[\hat{p} - 3 \times \text{se}; \hat{p} + 3 \times \text{se}]$ т.е. в интервале [0.14; 0.20].

Если наша выборка была достаточно большой и при этом репрезентативной, мы можем верить, что практически всегда «жаворонков» в выборках из 1000 респондентов, будет от 14% до 20%. Если другой исследователь получит на своей аналогичной выборке 40%, это будет выглядеть странно, возможно, он сформировал нерепрезентативную выборку, например, среди прочих, опросил много пенсионеров, которые обычно встают по утрам с большей охотой.

Как мы знаем, правило трёх сигм даёт приблизительные результаты и применяется тогда, когда мы изменчивость формулируем через стандартные отклонения. А нам нужен более точный и универсальный способ получить возможный интервал значений – доверительный интервал.

Доверительный интервал для доли – это обычный числовой отрезок, на котором с определённой степенью уверенности может лежать истинное значение доли p. Мы допускаем, что при оценивании доли по выборке мы получаем результат с некоторой погрешностью ε, эта погрешность называется **предельной ошибкой выборки**, так как это – максимальная ошибка, которую мы разрешаем себе допустить при приближении p с помощью $\hat{p}$ с учётом выбранной степени уверенности:

$$|p - \hat{p}| < \varepsilon$$
$$\hat{p} - \varepsilon < p < \hat{p} + \varepsilon.$$

Эта степень уверенности в наших выводах, другими словами, **уровень доверия**, влияет на значение предельной ошибки ε. Уровень доверия (обозначается буквой «бета») выбирается исследователем самостоятельно. Обычно в исследованиях используется уровень доверия $\beta = 95\%$, иногда $\beta = 90\%$ и $\beta = 99\%$. Что означает уровень доверия 95%? Если мы будем повторять аналогичное исследование много раз, независимо друг от друга, на выборках одного и того же размера, в 95% случаев доверительные интервалы, построенные по выборкам, будут накрывать истинное значение параметра генеральной совокупности. **Предельная ошибка выборки** при работе с долями вычисляется так:

$$\varepsilon = z^* \cdot \text{se} = z^* \cdot \frac{\sqrt{\hat{p}\hat{q}}}{\sqrt{n}},$$

где z^* – значение z-статистики, соответствующее выбранному уровню доверия β.

Итого, **границы доверительного интервала для доли** определяются так:

$$\hat{p} - z^* \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z^* \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}},$$

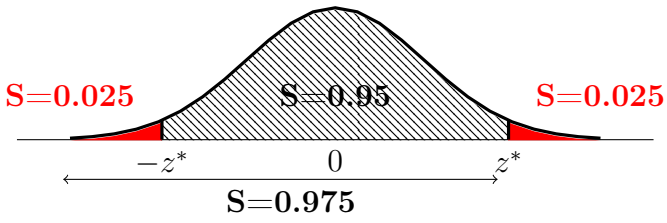
Осталось разобраться, как найти нужное z-значение.

Поиск z-значения для построения доверительного интервала

Рассмотрим три распространённых уровня доверия: $\beta = 95\%$, $\beta = 90\%$, $\beta = 99\%$.

- 95%-ный доверительный интервал ($\beta = 95\%$)

  График плотности стандартного нормального распределения, на котором заштрихована область, соответствующая 95% доверительному интервалу, выглядит следующим образом:

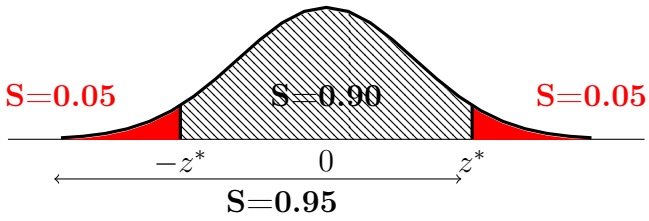


Видно, что если мы заштрихуем область, соответствующую уровню доверия (всегда симметрична относительно $z = 0$), то у нас останутся два одинаковых «хвоста» с вероятностями по 0.025. Интересующее нас значение z^* является квантилем уровня 0.975 (0.95 и левый «хвост» в 0.025).

Находим значение в таблице стандартного нормального распределения:

$$z^* = 1.96.$$

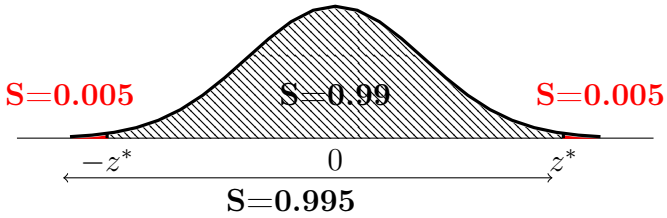
- 90%-ный доверительный интервал ($\beta = 90\%$)



Интересующее нас значение z^* является квантилем уровня 0.95 (0.90 и левый «хвост» в 0.05). Значение по таблице стандартного нормального распределения:

$$z^* = 1.65.$$

- **99%-ный доверительный интервал** ($\beta = 99\%$)



Интересующее нас значение z^* является квантилем уровня 0.995 (0.99 и левый «хвост» в 0.005). Значение по таблице стандартного нормального распределения:

$$z^* = 2.58.$$

Если перейти от графического представления к аналитическому, можем записать логику нахождения z^* в общем виде:

$$z^* = z_{\beta+\frac{1-\beta}{2}}.$$

Или (как на лекции) ввести $\alpha = 1 - \beta$ и вычислять так:

$$z^* = z_{1-\frac{\alpha}{2}}.$$

Доверительный интервал для среднего

Составные части доверительного интервала

Поставка задачи в случае среднего аналогична, мы хотим оценить, выбрав некоторый уровень доверия, в какой диапазоне могут лежать значения выборочного среднего (а значит, и среднее генеральной совокупности тоже). Например, мы не знаем, сколько времени, в среднем, уходит у студентов на выполнение домашнего задания, но мы можем опросить 30 студентов, по этой выборке посчитать среднее арифметическое и выборочное стандартное отклонение, и на его основе этих данных построить доверительный интервал.

Введём обозначения:

- a – среднее генеральной совокупности, его мы не знаем, но хотим оценить, как раз построив доверительный интервал на основе какой-нибудь одной выборки;
- $\bar{x}$ – среднее арифметическое выборки, которое мы получаем по итогам проведённого исследования.

Согласно центральной предельной теореме, распределение выборочной оценки среднего, среднего арифметического:

$$\bar{x} \sim N(a, \frac{\sigma^2}{n}),$$

где μ – среднее генеральной совокупности, σ^2 – дисперсия генеральной совокупности, n – объём выборки. Получается, дисперсия среднего $\bar{x}$ равна $\frac{\sigma^2}{n}$, значит, его стандартное отклонение равно $\frac{\sigma}{\sqrt{n}}$. Стандартное отклонение σ генеральной совокупности нам неизвестно, мы приближаем его с помощью стандартного отклонения выборки:

$$\sigma \approx s = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}.$$

Тогда **стандартная ошибка среднего**:

$$\text{se} = \frac{s}{\sqrt{n}}.$$

Предельная ошибка выборки при работе со средним:

$$\varepsilon = t^* \cdot \text{se} = t^* \cdot \frac{s}{\sqrt{n}}.$$

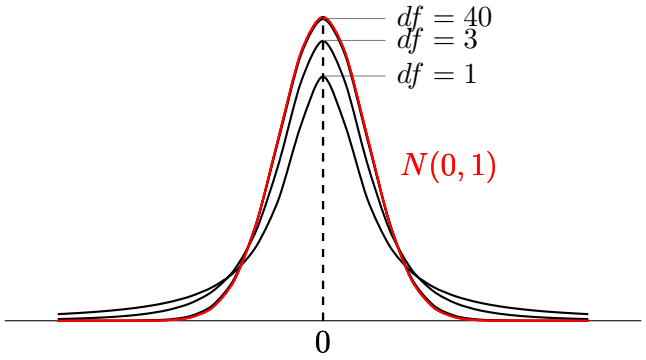
И **границы доверительного интервала для среднего** следующие:

$$\bar{x} - t^* \cdot \frac{s}{\sqrt{n}} < a < \bar{x} + t^* \cdot \frac{s}{\sqrt{n}},$$

где t^* – значение t-статистики, соответствующее выбранному уровню доверия β.

Поиск t-значения для построения доверительного интервала

Значение t^* берётся из распределения Стьюдента. График плотности распределения Стьюдента похож на график плотности стандартного нормального распределения, только более «плоский» и с более толстыми «хвостами». Главное отличие распределения Стьюдента от нормального распределения заключается в том, что у него есть специфический параметр – число степеней свободы (df). Число степеней свободы определяет форму распределения; чем больше число степеней свободы, тем ближе распределение Стьюдента к стандартному нормальному распределению:



Значение t^* для уровня доверия 95% – квантиль уровня 0.975 (та же логика, что и для z^*), число степеней свободы $\text{df} = n - 1$, где n – объём выборки. Аналогично для других уровней доверия. Например, для выборки в 25 наблюдений:

$$t^* = t(p = 0.975, \text{df} = 24) = 2.064.$$

Интерпретация доверительного интервала

95%-ный доверительный интервал для доли людей, любящих рано вставать по утрам, следующий: [0.15; 0.19].

✓ С 95%-ной уверенностью мы можем утверждать, что доля людей, любящих вставать по утрам, среди всех россиян, лежит в интервале от 0.15 до 0.19. Если мы будем проводить аналогичное исследование на выборках одного и того же размера много раз, независимо друг от друга, 95% доверительных интервалов будут включать истинное значение доли любителей рано вставать.

✓ Если мы будем проводить аналогичное исследование на выборках одного и того же размера много раз, независимо друг от друга, в 95% случаев истинное значение доли любителей рано вставать будет лежать в пределах от 0.15 до 0.19 (в предположении о том, что стандартная ошибка не изменяется от выборки к выборке).

Чтобы окончательно разобраться, можно посмотреть эту визуализацию. В ней фиксируется среднее генеральной совокупности (которое мы обычно не знаем, но оцениваем), берутся выборки одинакового размера из этой совокупности, по каждой выборке считается среднее, строится доверительный интервал и считается, сколько раз доверительные интервалы включили среднее, а сколько раз «прошли мимо» него.