*Алла Тамбовцева*

# Доверительные интервалы

Представьте, что нам необходимо оценить долю людей в России, которые любят рано вставать по утрам. Всех жителей России опросить не получится, но мы можем случайным образом выбрать 100 человек, провести опрос и выяснить, сколько среди них любителей вставать по утрам. Допустим, интересующая нас доля равна 0.17. Устроит ли нас такая оценка? С одной стороны, устроит: выборка хорошая и достаточно большая. 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 выборке, мы получаем значение с некоторой погрешностью. Что же мы можем в таком случае сделать? Зафиксировать уровень уверенности в наших расчетах и вместо одного значения для доли определить интервал, в пределах которого эта доля находится. Другими словами, мы можем построить **доверительный интервал**.

Построение доверительного интервала обычно выглядит следующим образом. У нас есть параметр $\Theta$ (например, доля людей, любящих вставать по утрам, по всей России), который мы не знаем, но хотим оценить. Оцениваем его по выборке ($\hat{\Theta}$) и делаем это с какой-то погрешностью. Мы допускаем, что при оценивании параметра по выборке максимум, что мы можем позволить – это отклониться от истинного значения параметра на некоторую величину, которая называется **предельной ошибкой выборки** ($\varepsilon$).

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

Значение предельной ошибки зависит от **уровня доверия**, который мы выбираем ($\gamma$[1]). Обычно в исследованиях используется уровень доверия 95%, иногда 90% и 99%. Что означает уровень доверия? Степень уверенности в оценках, которые мы будем получать на наших данных. Если мы будем повторять аналогичное исследование много раз, независимо друг от друга, в 95% случаев истинное значение параметра будет попадать в доверительный интервал. Например, мы захотели построить 95%-ный доверительный интервал для доли людей в России, которые любят вставать по утрам. Определили, что предельная ошибка выборки $\varepsilon = 0.07$. Это означает, что если мы будем проводить аналогичное исследование 100 раз, в 95 случаях доля любителей рано вставать, посчитанная по выборке, будет отклоняться от доли любителей рано вставать по всей России не больше, чем на 0.07 (считая, что предельная ошибка выборки не изменяется от выборки к выборке, то есть от исследования к исследованию).

Как считается предельная ошибка?

$$\varepsilon = const \cdot se,$$

где $const$ – это некоторое число, которое зависит от выбранного уровня доверия, а $se$ – это стандартная ошибка. **Стандартная ошибка** – это стандартное отклонение оценки, посчитанное по выборке. Для удобства разберем на примере оценки среднего значения.

Мы обсуждали, что оценка среднего значения генеральной совокупности – случайная величина, которая имеет примерно нормальное распределение со средним $a$ и стандартным отклонением $\frac{\sigma}{\sqrt{n}}$, где $a$ и $\sigma$ – это среднее значение и стандартное отклонение генеральной совокупности, а $n$ – число наблюдений в выборке. Обычно стандартное отклонение генеральной совокупности нам неизвестно, но мы можем его приблизить с помощью стандартного отклонения по выборке:

$$\sigma \approx s = \sqrt{\frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

[1] Это буква «гамма», просто печатная и прописная буквы выглядят по-разному.

где $\bar{x}$ – выборочное среднее, $n$ – число наблюдений в выборке. Тогда стандартная ошибка для нашего случая ($se$) – это $\frac{s}{\sqrt{n}}$, ее мы можем посчитать по выборке и подставить в формулу для доверительного интервала. Осталось понять, как определять, какое число брать в качестве *const* в формуле. Но это лучше обсуждать на конкретных примерах.

## Доверительный интервал для доли

1. Нас интересует некоторая доля в генеральной совокупности $p$, ее мы не знаем, но хотим оценить. Можно говорить, что доля $p$ – вероятность успеха.

   *Пример: доля людей, любящих вставать по утрам, среди всех жителей России.*

2. У нас есть выборка из $n$ наблюдений. Нам известна выборочная доля $\hat{p}$, оценка доли в генеральной совокупности; вероятность успеха, посчитанная по выборке.

   *Пример: доля людей, любящих вставать по утрам, среди всех людей в выборке.*

3. Нам известна выборочная доля $\hat{q}$ – вероятность неудачи, посчитанная по выборке, $\hat{q} = 1 - \hat{p}$.

   *Пример: доля людей, не любящих вставать по утрам, среди всех людей в выборке.*

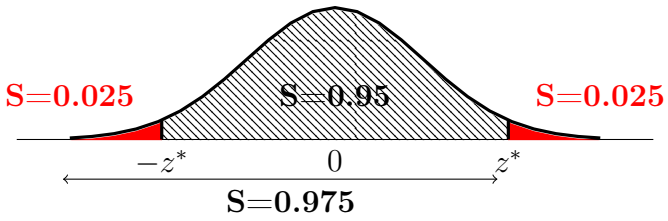
Тогда границы доверительного интервала определяются следующим образом:

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

где $z^*$ – значение $z$-статистики, соответствующее выбранному уровню доверия $\gamma$.

- **95%-ный доверительный интервал** ($\gamma = 95\%$)

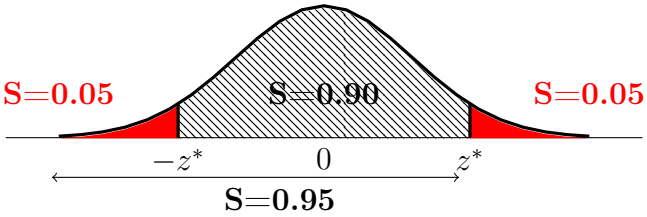
  График плотности стандартного нормального распределения, на котором заштрихована область, соответствующая 95% доверительному интервалу, выглядит следующим образом:




  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 квантилью уровня 0.975 (0.95 и левый «хвост» в 0.025).

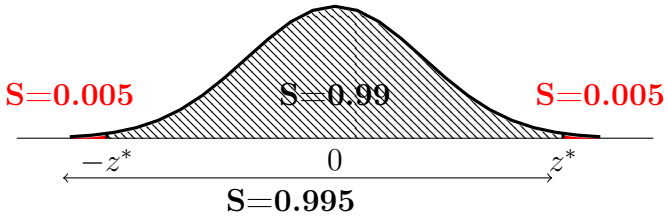
  Находим значение в таблице стандартного нормального распределения: $z^* = 1.96$.

- **90%-ный доверительный интервал** ($\gamma = 90\%$)



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** ($\gamma = 99\%$)




Интересующее нас значение $z^*$ является квантилью уровня 0.995 (0.99 и левый «хвост» в 0.005). Значение по таблице стандартного нормального распределения: $z^* = 2.58$.

## Доверительный интервал для среднего

1. Нас интересует среднее значение генеральной совокупности $\mu$, его мы не знаем, но хотим оценить.

   *Пример: средний рост студентов московских вузов.*

2. У нас есть выборка из $n$ наблюдений. Нам известно выборочное среднее $\bar{x}$ и выборочное стандартное отклонение $s$.

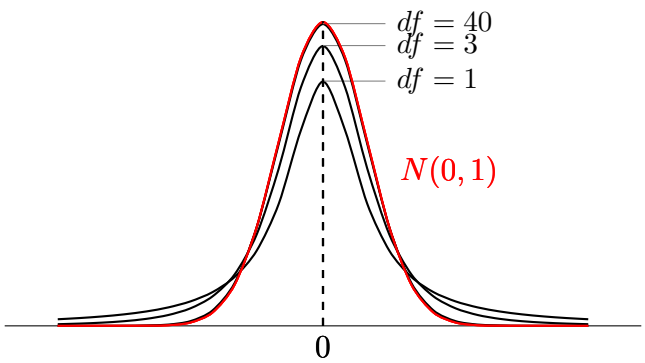
   *Пример: средний рост студентов в выборке и стандартное отклонение значений роста студентов в выборке.*

Тогда границы доверительного интервала определяются следующим образом:

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

где $t^*$ – значение $t$-статистики, соответствующее выбранному уровню доверия $\gamma$.

Значение $t$-статистики берется из распределения Стьюдента ($t$-распределения). 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 $df = n - 1$, где $n$ – число элементов в выборке. Аналогично для других уровней доверия.

Например, для выборки в 25 наблюдений, $t^* = t(p = 0.975, df = 24) = 2.064$.

**NB**: если вдруг нужного числа степеней свободы нет в таблице распределения (не все таблицы достаточно подробны), берите ближайшее число степеней свободы, которое в таблице есть и указывайте, какое значение степеней свободы в итоге используется в расчетах.

## Интерпретация доверительного интервала

*95%-ный доверительный интервал для доли людей, любящих вставать по утрам: [0.1; 0.24].*

✓ С 95%-ной уверенностью мы можем утверждать, что доля людей, любящих вставать по утрам, среди всех россиян, лежит в интервале от 0.1 до 0.24. 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 0.1 до 0.24 (в предположении о том, что предельная ошибка выборки/стандартная ошибка не изменяется от выборки к выборке).

✗ С вероятностью 0.95 истинная доля любителей вставать по утрам среди всех россиян лежит в интервале от 0.1 до 0.24.

? 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 0.1 до 0.24.

### Почему первые два утверждения корректны?

Утверждения содержат универсальную интерпретацию доверительного интервала и объяснение того, что означает выбранный уровень доверия.

### Почему третье утверждение не совсем корректно?

Уверенность – не то же самое, что вероятность. Когда вы говорите, что на 95% уверены, что сдадите экзамен, не готовясь, вы, возможно, знаете из предыдущего опыта, что в 19 случаях из 20 вы действительно сдали экзамен, не готовясь. Но это не означает, что с вероятностью 0.95 сдадите следующий экзамен, не готовясь. Вы можете его сдать (и тогда оцененная после сдачи вероятность будет 1) или завалить (оцененная после сдачи вероятность будет 0). Аналогичная ситуация с доверительными интервалами. Мы считаем доверительный интервал на основе *одной* выборки. И истинное значение доли любителей рано вставать либо войдет в этот интервал (вероятность попадания равна 1), либо не войдет (вероятность попадания равна 0).

**NB:** Существуют разные подходы к определению вероятности. В связи с этим некоторые исследователи считают подобные интерпретации доверительных интервалов (со словом «вероятность») *допустимыми*. Но во избежание недоразумений лучше говорить о степени уверенности, так как эта интерпретация точно не вызывает сомнений.

### Почему последнее утверждение может вызывать вопросы?

Когда мы говорим об уровне доверия 95% мы на самом деле имеем в виду следующее: если бы мы повторно проводили аналогичное исследование на выборках такого же размера, мы в 95% случаев получали бы оценки доли, которые отклоняются от истинного значения доли не больше, чем на величину $1.96 \cdot se$. А величина стандартной ошибки для каждой выборки своя! Один исследователь возьмет одну выборку в 100 человек, выяснит, что в ней 17% людей любят вставать по утрам ($se \approx 0.07$), другой возьмет другую выборку в 100 человек, выяснит, что в ней 25% людей любят вставать по утрам ($se \approx 0.09$) и так далее. Таким образом, каждый исследователь получит свой доверительный интервал, возможно, отличный от других, но в 95

случаях из 100 эти разные доверительные интервалы будут накрывать истинное значение доли любителей рано вставать.

**NB:** Обычно допускается, что предельная/стандартная ошибка не изменяется от выборки к выборке. В таком случае это утверждение **ничему не противоречит**.

Интерпретировать доверительный интервал так **можно**, но желательно указать соответствующее допущение о предельной/стандартной ошибке (как во втором утверждении).

Чтобы окончательно убедиться, 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.