ОП «Политология», 2023-24
Введение в ТВиМС
О доверительных интервалах и проверке гипотез (некоторые факты)
А. А. Макаров, А. А. Тамбовцева

Об обозначениях

В силу ограниченности ресурсов мы часто имеем дело не с генеральными совокупностями – всеми объектами интереса, а с выборками – той частью объектов, которые мы непосредственно обследуем. Параметры генеральной совокупности на практике нам часто неизвестны, например, мы не знаем долю интересующих нас людей среди всех жителей России, среднюю заработную плату всех работников сферы образования, дисперсию числа чашек кофе, выпиваемых всеми студентами университета во время написания кода на Python. Но эти параметры мы всегда можем оценить на выборках (собранных данных по результатам опроса/эксперимента/исследования) и получить выборочные оценки – выборочную долю, выборочное среднее, выборочную дисперсию. И, если изучаемая нами выборка достаточно большая и репрезентативная, эти оценки будут достаточно точными.

Доверительные интервалы мы строим для параметров генеральной совокупности и гипотезы формулируем тоже для них, а не для значений, полученных на конкретной выборке. Нет необходимости выбирать степень уверенности в результатах или фиксировать величину максимальной допустимой ошибки в выводах для исследуемой выборки, потому что про эту выборку мы и так всё 100%-но знаем. Все значения выборки перед нашими глазами, в каком-нибудь файле с данными, мы можем однозначно посчитать по ним среднее/дисперсию/долю и прочие характеристики. А вот чтобы перенести эти характеристики на всю генеральную совокупность (и на все возможные выборки такого размера), мы уже строим доверительные интервалы и проверяем статистические гипотезы.

Отсюда возникает необходимость в разных обозначениях, одни для выборочных оценок, другие – для генеральной совокупности:

- p – доля в генеральной совокупности, $\hat{p}$ – выборочная доля, $\hat{q} = 1 - \hat{p}$.
- μ – среднее генеральной совокупности, $\hat{\mu} = \bar{x}$ – выборочное среднее;
- σ^2 – дисперсия генеральной совокупности, s^2 – выборочная дисперсия.

О некоторых терминах и формулах

Характеристики выборки

- $\bar{x}$ – выборочное среднее, s^2 – выборочная дисперсия;
- s – выборочное стандартное отклонение, показывает, на сколько, в среднем, значения в выборке отклоняются от среднего выборки $\bar{x}$.

Пример 1. По выборке большого объёма посчитали среднее и выборочное стандартное отклонение суммы, затрачиваемой студентами на обед в столовой 14 мая.

Среднее равно 240 рублям, выборочное стандартное отклонение $s = 10$ рублей. Можем заключить, что, в среднем, студенты тратят на обед 240 ± 10 рублей. Если самой выборки у нас перед глазами нет, но выборка большая по объёму, и у нас есть основания считать, что она взята из нормального распределения, можем воспользоваться правилом трёх сигм и представить, в каком интервале могут лежать значения этой выборки:

$$240 \pm 3 \times 10 \to [210; 270]$$

Это не означает, что минимальное и максимальное значения выборки будут ровно такими, этот интервал показывает, примерно в каком диапазоне находятся траты студентов на обед (не средние траты, а сами значения расходов на обед 14 мая).

Стандартные ошибки

Стандартная ошибка доли/среднего показывает, насколько, в среднем, может изменяться выборочная доля/среднее от выборки к выборке (среди выборок одинакового размера n).

- стандартная ошибка доли: $\mathrm{se}_{\hat{p}} = \dfrac{\sqrt{\hat{p}\hat{q}}}{\sqrt{n}}$.
- стандартная ошибка среднего: $\mathrm{se}_{\hat{\mu}} = \dfrac{s}{\sqrt{n}}$.

Пример 2. По выборке объёма $n = 100$ хотим определить долю людей, любящих молочный шоколад. Выборочная доля $\hat{p} = 0.5$, стандартная ошибка доли $\mathrm{se} = 0.05$. Можем заключить, что при повторении аналогичного опроса на выборках из 100 респондентов, доля людей, любящих молочный шоколад будет равна 0.5 ± 0.05. Если наша выборка в 100 человек была репрезентативна, то мы можем воспользоваться правилом трёх сигм и получить интервал, в котором будет лежать 99.8% возможных значений доли:

$$0.5 \pm 3 \times 0.05 \to [0.35; 0.65]$$

Если при проведении аналогичного исследования доля любителей молочного шоколада окажется далеко за рамками этого интервала, одно из двух:

- выборка нерепрезентативна, её состав сильно отличается от того, что было в нашем исследовании;
- произошло какое-то качественное изменение, люди стали любить молочный шоколад больше/меньше, что объясняет неожиданно большое/маленькое значение выборочной доли.

Предельные ошибки, уровень доверия и уровень значимости

- предельная ошибка выборки ε, максимальная погрешность, которую мы разрешаем себе допустить при оценке доли/среднего по выборке на фиксированном уровне доверия β;

- уровень доверия β – степень уверенности в наших выводах, шансы построить такой доверительный интервал для доли или среднего, который будет включать истинное значение доли или среднего генеральной совокупности;
- уровень значимости α – величина, в некотором смысле обратная уровню доверия, так как $\alpha = 1 - \beta$, вероятность ошибиться, отвергнув верную нулевую гипотезу.

Пример 3. Строим 95%-ный доверительный интервал для доли жителей Москвы, которые любят молочный шоколад. Значит, мы допускаем, что с уверенностью 95% доверительный интервал, который мы построим по конкретной выборке из ограниченного числа жителей Москвы, включит долю любителей молочного шоколада по всем жителям Москвы, то есть будет «удачным».

Пример 4. Проверяем гипотезу о равенстве доли любителей молочного шоколада среди жителей Москвы значению 0.7 на уровне значимости $\alpha = 5\%$. Это означает, что при повторении аналогичного исследования на выборках такого же объёма, как и у нас, в 5% случаев нулевая гипотеза о равенстве доли 0.7 будет ошибочно отвергаться. То есть, в 95% случаев данные будут справедливо свидетельствовать в пользу её отвержения, а в 5% – нет. Это наше допущение в исследовании.

Работа с долями

Значение z^* для получения предельной ошибки ε при построении доверительного интервала для доли, а также критическое и наблюдаемое значения статистики критерия для проверки гипотезы о равенстве доли числу принадлежат стандартному нормальному распределению $Z \sim N(0, 1)$:

- $z^* = z_{1-\frac{\alpha}{2}}$, где $\alpha = 1 - \beta$;
- в случае двусторонней альтернативы $z_{\text{крит}} = z_{1-\frac{\alpha}{2}}$ и $z_{\text{крит}} = z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$;
- в случае правосторонней альтернативы $z_{\text{крит}} = z_{1-\alpha}$;
- в случае левосторонней альтернативы $z_{\text{крит}} = z_{\alpha} = -z_{1-\alpha}$.

Работа со средними

Значение t^* для получения предельной ошибки ε при построении доверительного интервала для среднего, а также критическое и наблюдаемое значения статистики критерия для проверки гипотезы о равенстве среднего числу принадлежат распределению Стьюдента с числом степеней свободы $n - 1$, где n – число наблюдений в выборке, т.е. $\mathrm{t}(\mathrm{df} = n - 1)$. При большом n, на практике при $n \geqslant 30$, можно пользоваться значениями из стандартного нормального распределения $Z \sim N(0, 1)$.

- $t^* = t_{1-\frac{\alpha}{2},\, \mathrm{df}=n-1}$, где $\alpha = 1 - \beta$;
- в случае двусторонней альтернативы $t_{\text{крит}} = t_{1-\frac{\alpha}{2},\, \mathrm{df}=n-1}$ и $t_{\text{крит}} = -t_{1-\frac{\alpha}{2},\, \mathrm{df}=n-1}$;
- в случае правосторонней альтернативы $t_{\text{крит}} = t_{1-\alpha,\, \mathrm{df}=n-1}$;
- в случае левосторонней альтернативы $t_{\text{крит}} = t_{\alpha,\, \mathrm{df}=n-1} = -t_{1-\alpha,\, \mathrm{df}=n-1}$.