ОП «Политология», 2020-21
Введение в ТВиМС
Генеральная совокупность vs выборка. (22 марта 2021 г.)
А. А. Макаров, А. А. Тамбовцева

В статистике и анализе данных есть два важных термина: генеральная совокупность и выборка. **Генеральная совокупность** включает в себя все объекты интереса. **Выборка** включает только те объекты интереса, которые мы непосредственно обследуем. Обычно для изучения всех элементов генеральной совокупности не хватает ресурсов (времени, средств, людей), поэтому мы можем работать с выборкой, взятой из генеральной совокупности, и пытаться по ней делать выводы о всей генеральной совокупности.

Вопрос 1. С какими проблемами мы можем столкнуться, используя такой подход?

Изучая выборку вместо всей генеральной совокупности, мы можем столкнуться с серьёзной проблемой, связанной с адекватностью результатов, если наша выборка плохо соотносится с генеральной совокупностью. Так как мы заинтересованы в корректных результатах, мы хотим получать такие выборки, которые хорошо бы отражали свойства генеральной совокупности. В статистике и количественных исследованиях такие выборки называются **репрезентативными**. Плюс, выборка должна быть достаточно большой. Нет единого соглашения о том, какая выборка считается достаточно большой, зависит от исследовательской области, но, когда речь заходит о статистических законах и теоремах, обычно говорят о выборке из 30 наблюдений и больше ($n \geqslant 30$).

Вопрос 2. Если мы знаем, что в определённом городе женщины составляют 60% населения, а мужчины – 40%, можем ли мы считать репрезентативной выборку из 80% мужчин и 20% женщин?

Если выборка сильно отличается от генеральной совокупности, такая выборка называется смещённой (*biased*). Существуют разные типы смещений, но большинство из них сильно зависят от процедуры сбора данных. Если мы организуем опрос среди наших близких друзей и опубликуем результаты опроса, результаты такого опроса не будут надёжными, так как наша выборка смещённая – редко близкие друзья могут представлять жителей целого города точным образом. Подобная проблема может возникнуть, если мы опросим только женщин или только людей старше 40 лет.

Генеральная совокупность может быть описана с помощью случайной величины с определёнными параметрами. Итак, когда мы берём выборку из генеральной совокупности, мы можем смотреть на неё как на выборку из случайной величины с определённым распределением. Если выражаться более формально, когда мы берём выборку из n наблюдений, мы берём n независимых реализаций случайной величины. Чтобы было понятнее, рассмотрим следующий пример. Студентка проводит эксперимент: бросает игральный кубик и, если выпадает 6 очков, она записывает 1, иначе – записывает 0. Результат эксперимента можно описать с помощью бинарного

распределения с вероятностью успеха $p = 1/6$. Что мы получили, если у нас есть такая выборка:

$$0,\ 1,\ 0,\ 0,\ 0?$$

В действительности, у нас есть записи по пяти броскам кубика. Другими словами, мы просим студентку независимо бросать кубик 5 раз (независимо от результатов в предыдущих бросках), проверять, что получилось и записывать 1 и 0.

Пример 1. Известно, что человеческий рост распределён нормально с математическим ожиданием μ и дисперсией σ^2. Это интуитивно понятно: много людей, чей рост несильно отличается от среднего и мало людей, которые слишком низкие или высокие. Мы можем взять 100 человек (только мужчин или женщин, так как средний рост и дисперсия роста отличается в зависимости от пола) и получить выборку из нормальной случайной величины $N(\mu, \sigma^2)$.

Теперь рассмотрим пару задач.

Задача 1. Рассмотрим бинарную случайную величину с параметром, с вероятностью успеха $p = 1/4$. Предложите репрезентативную (наиболее вероятную) выборку размера $n = 10$ из этой случайной величины.

Решение. Если мы хотим, чтобы выборка хорошо отражала свойства генеральной совокупности, она должна содержать (примерно) 1/4 «единиц» и 3/4 «нулей». На первый взгляд, кажется, что невозможно получить 1/4 «единиц», так как $10 \cdot 1/4$ не является целым числом, но это не должно смущать нас. Так как наша выборка очень маленькая, доля «единиц» и «нулей» в выборке может сохраняться лишь примерно, поэтому нормально, если мы получим 2 «единицы» в выборке (1/5 из 10) или 3 «единицы» в выборке (3/10 из 10). Итак, мы можем предложить следующую выборку:

$$0,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.$$

Напротив, следующая выборка не будет репрезентативной:

$$0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.$$

Задача 2. Может ли следующая выборка быть репрезентативной (правдоподобной) выборкой из стандартной случайной величины:

$$-8,\ -3,\ 4,\ -5,\ 7,\ 9,\ 2.5?$$

Решение. Стандартная нормальная величина имеет распределение $Z \sim N(0, 1)$. По правилу трёх сигм мы можем получить следующее: 99.7% значений из Z лежат в интервале $[-3; 3]$. Давайте посмотрим на нашу выборку.Только два значения попадают в этот интервал! И эти значения даже находятся на границе типичных значений Z (2.5 близко к 3, а -3 – сама нижняя граница). Итак, мы можем сказать, что такая выборка не является репрезентативной выборкой из генеральной совокупности, описываемой стандартной нормальной величиной.