ОП «Политология», 2023-24
Введение в ТВиМС
Предельные теоремы (6 марта)
А. А. Макаров, А. А. Тамбовцева

На семинаре мы зафиксировали следующие неформальные определения важных понятий в прикладном анализе данных:

- Генеральная совокупность – все объекты интереса.
- Выборка – объекты интереса, которые мы непосредственно обследуем.

На практике часто невозможно работать со всей генеральной совокупностью в силу ряда ограничений (недостаточно ресурсов для опроса большого числа людей, недоступность данных, пропущенные значения из-за отказа участвовать в исследовании или некорректных ответов), поэтому работают с выборкой – организуют выборочное обследование.

Примеры: генеральная совокупность – все жители России, выборка – случайно выбранные жители, которые участвуют в опросе и ответы которых мы фиксируем; генеральная совокупность – все районы Москвы, выборка – случайно выбранные из каждого административного округа районы, число районов из каждого округа пропорционально размеру округа.

Если описывать выборочные обследования более формально, получим следующее:

- Генеральная совокупность описывается **случайной величиной**, которая имеет некоторое распределение с определёнными параметрами.
- Выборка извлекается из случайной величины, т.е. из некоторого распределения.

Пример 1. Генеральная совокупность – все жители Москвы, нас интересует согласие с некоторым утверждением в самом общем виде, т.е. ответы «да» или «нет». Из предыдущих аналогичных исследований известно, что доля согласных примерно равна 0.2. Значит, если мы закодируем ответы «да» значением 1, а ответы «нет» – значением 0, такую генеральную совокупность можно описать бинарной случайной величиной с параметром $p = 0.2$. Выборка объёма $N = 10$, набор ответов 10 респондентов, может выглядеть так (три человека ответили «да»):

$$0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0$$

Так как в реальности мы обычно не знаем параметр p (не всегда есть аналогичные масштабные исследования, тем более, актуальные), его мы **оцениваем** на основе имеющейся выборки. В данном случае оценка вероятности p – это доля единиц в выборке, её иногда обозначают $\hat{p}$, так как символ «крышечка» означает оценку, примерное значение параметра распределения:

$$p \approx \hat{p}, \text{ если выборка достаточно большая и репрезентативная}$$

Значения p и $\hat{p}$ не обязаны совпадать (в этом примере $\hat{p} = 0.3$, а $p = 0.2$), особенно на маленьких выборках, однако при увеличении объёма выборки разница между ними уменьшается.

Пример 2. Генеральная совокупность – все студенты 1 курса ОП «Политология», нас интересует время, затраченное на выполнение первого домашнего задания (в минутах). Исходя из опыта, предполагаем, что время, затраченное на выполнение первого домашнего задания имеет нормальное распределение со средним 100 минут и стандартным отклонением 20 минут. Пример выборки объёма $N = 6$, набор ответов 6 студентов:

$$120 \quad 90 \quad 80 \quad 140 \quad 100 \quad 85$$

Как мы убедились в интерактивном режиме на семинаре, если случайным образом выбирать 6 студентов для опроса, каждый раз мы будем получать разные выборки. И среднее арифметическое по выборке тоже каждый раз будет новым, меняться от выборки к выборке. А это значит, что среднее арифметическое является случайной величиной! Какое распределение у этой случайной величины? Ответ на этот вопрос даёт центральная предельная теорема. В одном из вариантов она выглядит так.

Центральная предельная теорема. Пусть есть случайная величина с математическим ожиданием a и дисперсией σ^2. Если из этой величины независимым образом много-много раз извлекать выборки достаточно большого[1] объёма N, то среднее арифметическое таких выборок будет иметь нормальное распределение с математическим ожиданием a и дисперсией $\frac{\sigma^2}{N}$.

Задача 1. Известно, что время, затраченное на выполнение первого домашнего задания имеет нормальное распределение со средним 100 минут и стандартным отклонением 20 минут. В рамках мини-исследования мы случайным образом извлекаем выборки объёма $N = 6$, то есть имитируем опрос шести студентов [2].

(a) Запишите параметры распределения среднего арифметического выборки.
(b) Вычислите вероятность того, что среднее выборки будет менее 90 минут.
(c) Вычислите вероятность того, что среднее выборки превысит 120 минут.
(d) Вычислите вероятность того, что среднее выборки отклонится от среднего генеральной совокупности не более, чем на 5 минут.
(e) Вычислите вероятность того, что среднее выборки отклонится от среднего генеральной совокупности не более, чем на 2%.

[1] Много-много раз – в идеале все возможные выборки, хотя в реальности мы всё равно работаем с какой-то одной. Достаточно большого объёма – хотя бы с $N = 30$.

[2] Важно: центральная предельная теорема начинает действовать на выборках хотя бы из 30 наблюдений, но здесь мы продолжаем работать с выборкой из шести человек в продолжение интерактивного примера на семинаре, просто пока отметим, что наши результаты будут неточными!