Математические и статистические методы в психологии
Описательные статистики. (11 апреля 2019 г.)
А. А. Тамбовцева

Описательные статистики

Номинальные (категориальные) переменные

Какие характеристики используются для описания номинальных (категориальных или качественных) переменных?

Так как нет никакого смысла работать с номинальными переменными как с числовыми, некорректно пытаться считать среднее или дисперсии. Однако, описать номинальную переменную всё-таки можно.

- **Частоты:** частоты могут быть абсолютными (результаты подсчёта) или относительными (доли или %).
- **Мода:** самое часто встречающееся значение в выборке.

Количественные (числовые) переменные

Какие характеристики используются для описания количественных переменных?

Базовые статистики

- Минимальное значение: min;
- Максимальное значение: max;

Меры центральной тенденции

- Среднее: $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_k}{n}$, где n – размер выборки;
- Медиана (см. ниже).

Меры разброса (изменчивости)

- Размах: $\mathrm{range} = \max - \min$;
- Выборочная дисперсия: $s^2 = \dfrac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_k - \bar{x})^2}{n-1}$;
- Выборочное стандартное отклонение: $s = \sqrt{s^2}$;
- Межквартильный размах: $\mathrm{IRQ} = \mathrm{Q}_3 - \mathrm{Q}_1$;

Выборочные квантили

Квантиль уровня p – значение, которое остальные значения в выборке не превышают с вероятностью p (вероятность здесь можно рассматривать как относительную частоту).

Пример 1. Дана выборка X:

$$8\ 7\ 3\ 0\ 1\ 2\ 6\ 9\ 12\ 9$$

Чтобы найти квантили разных уровней вручную, выборку сначала надо упорядочить:

$$0\ 1\ 2\ 3\ 6\ 7\ 8\ 9\ 9\ 12$$

Теперь найдём квантиль уровня 0.2. Здесь это 1, так как 20% значений в выборке (2 из 10) не превышают 1.

Пример 2. Если мы знаем, что 32 – выборочный квантиль уровня 0.4 переменной `age` в наших данных, мы можем заключить, что 40% людей в нашем датасете не старше 32 лет.

Существуют квантили особых уровней (25%, 50%, 75%, 100%), которые называются **квартилями**. Этот термин следует из от того факта, что квартили делят выборку на четыре равные части (см. рис. 1).

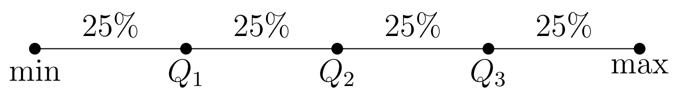

Рис. 1: Квартили

Q_1 – **нижний квартиль** (1-ый квартиль), значение, которое отделяет первые 25% наблюдений в выборке. Q_3 – **верхний квартиль** (3-й квартиль), значение, которое отделяет первые 75% of наблюдений в выборке. Q_2 обычно не называется вторым квартилем, называется **медианой**, так как делит выборку на две равные части, первые 50% и вторые 50% наблюдений.

Пример 3. Если нам известно, что для переменной `income`:

- 1-ый квартиль: 18000 руб.
- медиана: 35000 руб.
- 3-ый квартиль: 52000 руб.,

мы можем заключить, что 25% респондентов зарабатывают не более 18000 рублей в месяц, 50% респондентов зарабатывают не более 35000 рублей, и 75% респондентов зарабатывают не более 52000 рублей (или 25% людей зарабатывают 52000 рублей.

Используя квартили, мы можем посчитать **межквартильный размах**, меру изменчивости, которая более устойчива к наличию нетипичных, слишком больших или маленьких значений в выборке по сравнению с «обычным» размахом. Межквартильный размах считается следующим образом:

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1.$$

Пример 4. Рассмотрим выборку (уже упорядочена по возрастанию):

$$2 \quad 2.5 \quad 2.8 \quad 3 \quad 3.4 \quad 4.8 \quad 5.2 \quad 5.3 \quad 7.1 \quad 8.2 \quad 8.8 \quad 100$$

Если мы попытаемся делать выводы об этой выборке по обычному размаху, мы решим, что значения в выборке довольно сильно разбросаны (разнообразны) ($\mathrm{range} = \max - \min = 98$). Однако, это результат обеспечивается только за счёт того, что одно значение очень большое. Если мы посчитаем межквартильный размах, результат будет более скромным, и при этом гораздо лучше отражать реальность ($\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1 = 5.6$). Межквартильный размах не такой большой, и мы видим, что значения несильно отличаются друг от друга, если мы исключим из рассмотрения 100.