Теория вероятностей и математическая статистика

Вебинар 3

Описательная статистика. Качественные и количественные характеристики популяции. Графическое представление данных

На предыдущих занятиях мы говорили о теоретических характеристиках случайных событий и величин. Зачем нам это нужно?

Любая выборка представляет собой значения некоторой случайной величины.

На предыдущих занятиях мы говорили о теоретических характеристиках случайных событий и величин. Зачем нам это нужно?

Любая выборка представляет собой значения некоторой случайной величины.

На практике, как правило, мы имеем некоторую выборку. Мы ничего не знаем про случайную величину, из которой взята эта выборка. А хотелось бы.

Математическая статистика

На предыдущих занятиях мы говорили о теоретических характеристиках случайных событий и величин. Зачем нам это нужно?

Любая выборка представляет собой значения некоторой случайной величины.

На практике, как правило, мы имеем некоторую выборку. Мы ничего не знаем про случайную величину, из которой взята эта выборка. А хотелось бы.

В дальнейшем мы научимся по выборкам:

- строить оценки для параметров случайных величин,
- проверять гипотезы о значениях этих параметров, а также об общих характеристиках и свойствах случайных величин,
- строить доверительные интервалы для параметров случайных величин.

Точечная оценка параметров. Статистики. Выборочное среднее

Для точечного оценивания параметров случайной величины используются различные статистики. *Статистика* — это любая функция от выборки.

Пусть дана выборка $X = (x_1, x_2, \dots, x_n)$ из значений случайной величины. Одной из наиболее естественных статистик таких выборок является *среднее арифметическое* (или *выборочное среднее*). Оно обозначается как $\overline{X}$:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Выборочное среднее является *оценкой* для математического ожидания. Это означает, что, как правило, чем больше элементов в выборке, тем ближе выборочное среднее этой выборки к математическому ожиданию соответствующей случайной величины.

Выборочная дисперсия

Выборочная дисперсия, как следует из названия, оценивает дисперсию случайной величины:

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

Выборочная дисперсия

Выборочная дисперсия, как следует из названия, оценивает дисперсию случайной величины:

$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \overline{X} \right)^2$$

Несмотря на кажущуюся естественность, данная оценка является не очень хорошей в силу своей *смещённости* (об этом на следующем слайде). Поэтому в практических задачах используют *несмещённую оценку дисперсии*:

$$\sigma_{X,\, unbiased}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{X} \right)^2$$

Несмещённая и смещённая оценки

На самом деле каждый объект из выборки — это тоже случайная величина (поскольку выбирается случайным образом). В таком случае и любая статистика (т.е. функция от выборки) является случайной величиной.

Оценка некоторого параметра случайной величины называется *несмещённой*, если математическое ожидание этой оценки равняется реальному значению этого параметра.

Например, пусть выборка X берётся из значений случайной величины x. Тогда выборочное среднее является несмещённой оценкой математического ожидания:

$$M\left(\overline{X}\right) = M(x)$$

В практическом смысле это означает, что если мы рассмотрим большое количество различных выборок, то, хотя выборочное среднее каждой из них вряд ли будет равно математическому ожиданию x, в среднем мы получим именно его.

Выборочная дисперсия

Оказывается, обычная оценка дисперсии является смещённой:

$$M\left(\sigma_X^2\right) = \frac{n-1}{n} D(x)$$

Замечание. При оценке дисперсии (да и вообще) используют именно несмещённые оценки. В дальнейшем под σ_X^2 мы будем понимать именно несмещённую оценку:

$$\sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2$$

Среднее квадратическое отклонение

Вообще, дисперсия является не очень наглядной мерой разброса, поскольку имеет другой масштаб. Поэтому часто наряду с дисперсией используют *среднее квадратическое отклонение*, равное корню из дисперсии.

Оценивается среднее квадратическое отклонение аналогично дисперсии. Смещённая и несмещённая оценки:

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}, \quad \sigma_{X,\, unbiased} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

Как и в случае с дисперсией, под σ_X мы будем в будущем понимать именно *несмещённую* оценку.

Мода, медиана, квантиль

Мода и медиана

Мода — наиболее часто встречающееся в выборке значение.

Обычно мода рассматривается в том же контексте, что и выборочное среднее: она позволяет получить некоторую информацию о выборке «в среднем».

Мода и медиана

Мода — наиболее часто встречающееся в выборке значение.

Обычно мода рассматривается в том же контексте, что и выборочное среднее: она позволяет получить некоторую информацию о выборке «в среднем».

Медиана — такое значение t, что половина элементов из выборки меньше, либо равна t, и, соответственно, половина больше, либо равна t.

Медиана представляет собой *середину* выборки: если отсортировать элементы выборки по возрастанию, то медиана приходится на середину.

Медиана может приходиться как на промежуток между элементами выборки, так и на конкретный элемент.

Медиана и квантили

Медиана является частным случаем более общего понятия — *квантиля*.

Пусть $\alpha \in (0, 1)$. *Квантиль порядка α* — такое число t_α, что «α процентов» всех элементов выборки меньше t_α и, соответственно, «$(1-\alpha)$ процентов» элементов — больше t_α.

Как и в случае с медианой, квантиль может как приходиться на один из элементов выборки, так и лежать где-то между ними.

Квартили, децили, перцентили

Из определения следует, что медиана является квантилем порядка 0.5. Кроме того, часто используют:

- *первый квартиль* — квантиль порядка 0.25 (т.е. значение, которое не превышают 25% значений из выборки),
- *второй квартиль* — то же, что и медиана,
- *третий квартиль* — квантиль порядка 0.75.

Квартили, децили, перцентили

Из определения следует, что медиана является квантилем порядка 0.5. Кроме того, часто используют:

- *первый квартиль* — квантиль порядка 0.25 (т.е. значение, которое не превышают 25% значений из выборки),
- *второй квартиль* — то же, что и медиана,
- *третий квартиль* — квантиль порядка 0.75.

Также могут встречаться:

- *децили* — то же, что и квартили, но делим мы не на 4 части, а на 10. Например, медиана будет пятым децилем,
- *перцентили* — это просто другой способ задать квантиль. Здесь мы используем не долю $\alpha \in (0, 1)$, а процент. Например, третий квартиль будет 75-перцентилем.

Интерквартильный размах

Интерквартильный размах — это отрезок между первым и третьим квартилями. Это отрезок, в который попадают 50% значений выборки.

Интерквартильный размах используется для измерения разброса значений выборки вокруг среднего. Иногда его использование оказывается более предпочтительным, чем использование среднего квадратического отклонения, поскольку не учитывает выбросы в данных.

Квантиль случайной величины

Понятие квантиля также можно определить для случайной величины. Суть определения такая же, что и в случае выборки, но выглядит немного страшнее.

Квантилем порядка α случайной величины X называется такое значение t_α, что

$$P(X \leq t_\alpha) = \alpha, \ \ P(X \geq t_\alpha) = 1 - \alpha$$

Идея та же: в доле α всех случаев значение случайной величины X окажется меньше t_α и в доле $(1 - \alpha)$ случаев — больше t_α.

Квантиль случайной величины

Использование квантилей позволяет в некотором смысле «обратить» функцию распределения.

Замечание. Функцию распределения мы определим на следующем занятии, пока спойлер: для случайной величины X функция распределения выглядит следующим образом:

$$F_X(x) = P(X \leq x)$$

Прямая задача выглядит так: имеется случайная величина X и пороговое значение t. Требуется найти вероятность того, что величина X не превосходит значения t. Для этого нужна функция распределения.

Квантиль случайной величины

Использование квантилей позволяет в некотором смысле «обратить» функцию распределения.

Замечание. Функцию распределения мы определим на следующем занятии, пока спойлер: для случайной величины X функция распределения выглядит следующим образом:

$$F_X(x) = P(X \leq x)$$

Прямая задача выглядит так: имеется случайная величина X и пороговое значение t. Требуется найти вероятность того, что величина X не превосходит значения t. Для этого нужна функция распределения.

Часто в задачах математической статистики требуется решить обратную задачу: имеется случайная величина X и значение вероятности $\alpha \in (0, 1)$. Требуется найти пороговое значение t, такое, что $P(X \leq t) = \alpha$. Это и есть квантиль порядка α.

Графическое представление данных

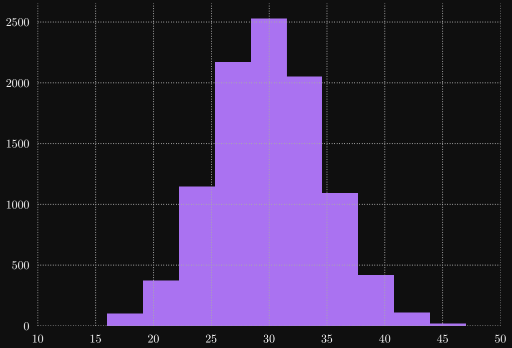
Гистограмма

Для визуализации распределения значений выборки часто используется *гистограмма*.
Как строится гистограмма?

1. По оси x откладываются все возможные значения из выборки.
2. Вся ось разбивается на какое-то заданное число одинаковых отрезков.
3. Для каждого отрезка вычисляется число значений выборки, которые лежат в этом отрезке, и это число откладывается по оси y.

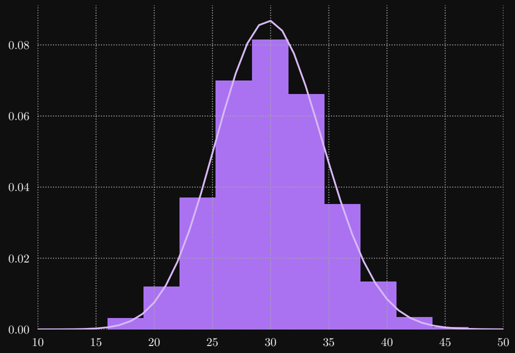
Гистограмма

Например, справа изображена гистограмма для выборки размера 10000 из биномиального распределения с параметрами $n = 100$ и $p = 0.3$. Разбиение оси x здесь производится на 10 частей.



Гистограмма

Гистограмма по форме напоминает график распределения вероятностей случайной величины. Нужно лишь нормировать откладываемые по оси y значения, чтобы сумма площадей колонок стала равной 1.



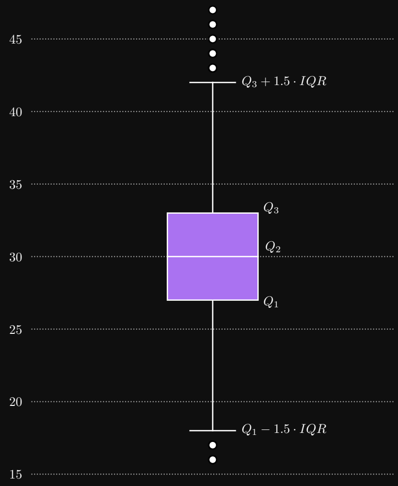
Boxplot

Другой способ визуализировать одномерные данные — *boxplot* или *ящик с усами*. В самом ящике отмечены квартили Q_1, Q_2 (медиана), Q_3. «Усы» здесь — границы отрезка

$$[Q_1 - 1.5 \cdot IQR,\ Q_3 + 1.5 \cdot IQR]\,,$$

где IQR — интерквартильное расстояние.
Всё, что выходит за границы этого отрезка, считается выбросами (отмечены кружками). Обычно это порядка 0.7% выборки.

Например, справа изображён boxplot той же выборки из биномиального распределения с параметрами $n = 100$, $p = 0.3$.



На следующем занятии

Непрерывные случайные величины. Функция распределения и плотность распределения. Равномерное и нормальное распределение. Центральная предельная теорема