

РАСПРЕДЕЛЕНИЕ

Элементы теории вероятностей

Случайность

- Результаты статистического наблюдения
 - случайные величины
 - большое количество данных
- Результаты обработки статистических данных
 - случайные величины
- Характеристики случайности
 - вероятность
 - распределение

Вероятность

- *объективное* определение:
 - результат расчетов
 - много событий (массовые явления)
 - относительная **частота** (доля) исходов

$$p(A) = \frac{n_A}{n}$$

- *субъективное* определение:
 - экспертные знания
 - единичные (редкие) события
 - уверенность в наступлении события, степень доверия к утверждению

Значение вероятности

$$0 \leq p \leq 1$$

$$p = 0$$

$$p = 1$$

- P – probability

Значение вероятности

- *Теоретическая* вероятность $p(A)$
 - «аналитические» методы
 - теория вероятностей
 - формулы (уравнения)
 - «точные» решения
- *Эмпирическая* вероятность $\lim_{n \rightarrow \infty} \hat{p}_A = p(A)$
 - «численные» методы
 - вычисления по фактическим данным
 - оценки (приближенные решения)
 - «Эмпирический» – опытный, основанный на опыте, на наблюдениях
 - [Gr. εμπειρος (опытный) = εμπ-, εν- (в, внутри, в пределах) + πειρα (опыт, умение)] == «практик»

Оценка

$\hat{p}$

- Оценка – приближенное значение, вычисленное по фактическим данным
- Оценка приближается к точному истинному значению при большом количестве исходных данных
- Условное обозначение: «с крышкой»

Распределение случайной величины (distribution)

- Закон (функция), описывающий частоту (вероятность) появления различных значений признака

$$P(X)$$

X : значения случайной величины (аргумент)

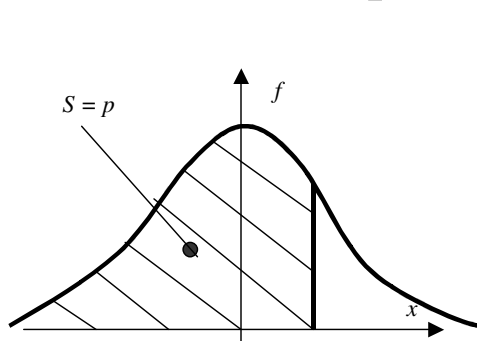
P : вероятность появления значений (функция)

- Использование распределения

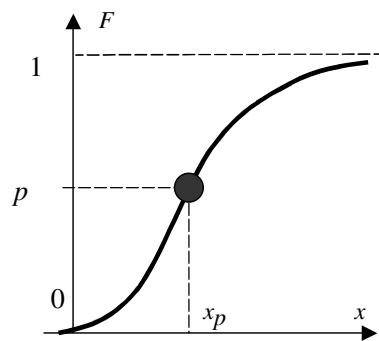
$$X \rightarrow P$$

$$P \rightarrow X$$

Закон распределения



Плотность вероятности



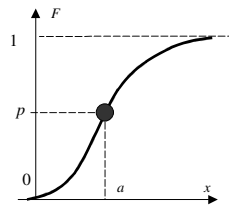
Функция распределения

Функция распределения

- вероятность того, что показатель x не превысит заданного значения a :

$$F(a) = \text{Prob}\{x \leq a\}$$

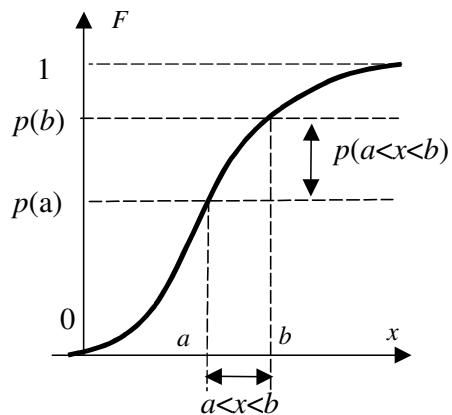
- принимает значения от 0 до 1
- неубывающая (монотонная)
 - остается постоянной или возрастает



Вероятность на интервале

- Вероятность того, что случайная величина x окажется в интервале от a до b :

$$\begin{aligned} \text{Prob}\{a \leq x < b\} &= \\ &= F(b) - F(a) = \\ &= \text{Prob}\{x \leq b\} - \text{Prob}\{x \leq a\} \end{aligned}$$



Производная и интеграл

- Плотность вероятности – производная функции распределения:

$$p(x) = \frac{dF}{dx} = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\text{Prob}\{x \leq \xi < x + \Delta x\}}{\Delta x}$$

- Функция распределения – интеграл от плотности вероятности:

$$F(x) = \int_{-\infty}^x p(x) dx$$

«Плотность» вероятности

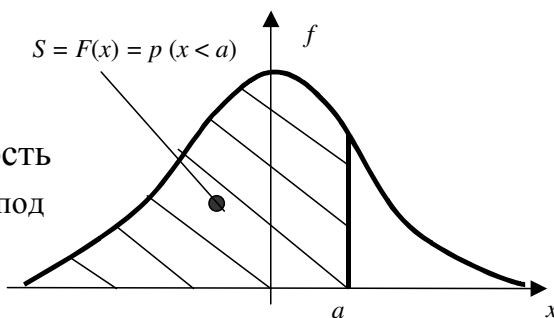
- плотность вещества – масса на объем
- плотность вероятности – вероятность на интервал

– Площадь = вероятность

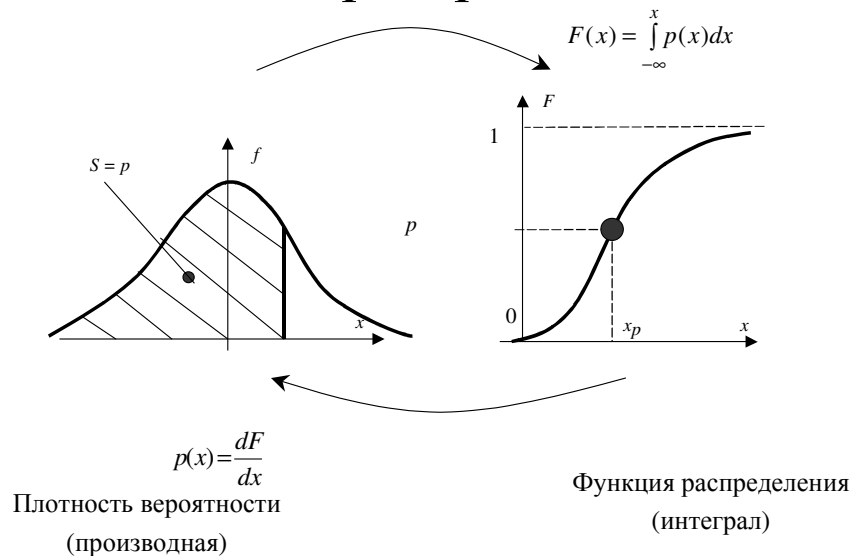
- интеграл от фигуры под графиком

– масса

- интеграл от плотности



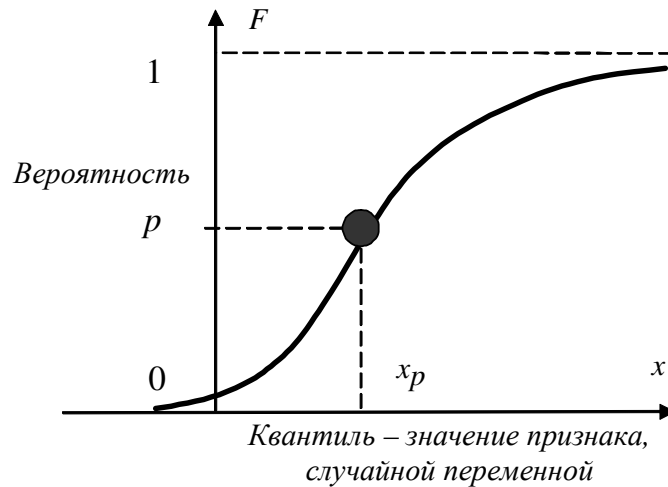
Закон распределения



Квантиль

- аргумент функции распределения (**значение** случайной величины) при заданной вероятности
- процентная точка распределения
- решение уравнения: $x_p : F(x_p) = p$
- «*статистика*» - *функция выборки*
 - Квантиль = «количество», «значение», «величина»
(в противоположность вероятности)
 - [L. *quantus* – насколько большой; *quantum* – сколько]
 - [It. *Quanto costa?* – «сколько стоит?»]
 - [E. *quantity* – количество]
 - квант, квантование

Квантиль



Характеристики распределений

- Математическое ожидание

- expected value

$$E(x) = \mu = \int_{-\infty}^{+\infty} x p(x) dx$$

- Дисперсия

- мера разброса, рассеяния вокруг среднего значения
- средний квадрат отклонения от среднего значения

$$D(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx$$

- [L. *dispersus* (разбросанный, рассыпанный)
- = *dis-* (от-, раз-) + *spargo* (сыпать)] = «разброс»

«Сигма»

- Названия показателя:
 - среднее квадратическое отклонение
 - среднее квадратичное отклонение
 - с.к.о.
 - стандартное отклонение
 - стандарт распределения
 - стандартная ошибка
 - Standard Error, Standard Deviation, STD

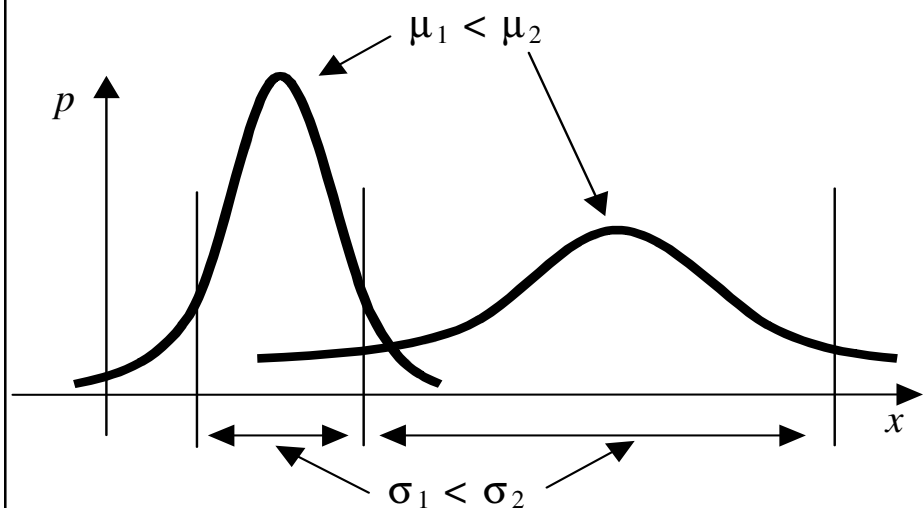
– «сигма» σ

$$\sigma = \sqrt{D}$$

- Вычисление: корень из дисперсии

$$D = \sigma^2$$

Примеры: М.О. и с.к.о.



Мода

- Наиболее часто встречающееся значение признака
- Значение признака с наибольшей частотой
- Значение с максимальной вероятностью
- Пик на графике плотности вероятности
- Решение уравнения

$$Mo = \arg \max_x p(x)$$

Мода

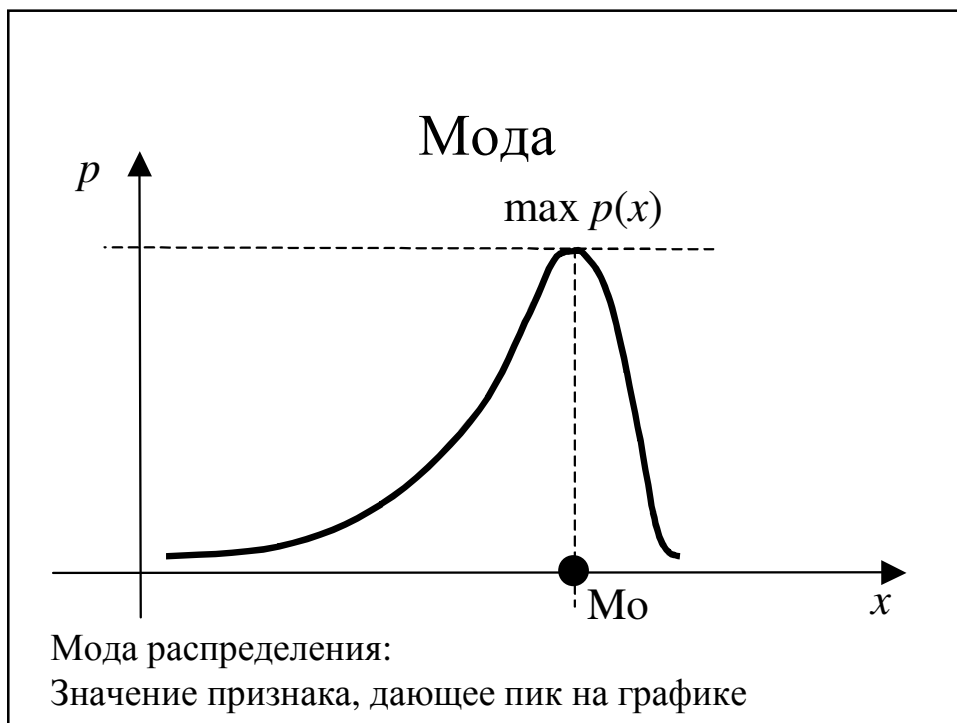
- 1. Господство вкусов в отношении одежды, предметов быта
- 2. Образцы предметов одежды, отвечающие таким вкусам
- 3. Манера поведения
- 4. (стат.) наиболее часто встречающийся тип явления
 - 5. (теор.колеб.) одно из собственных колебаний системы
 - [F. *mode* – манера, способ; мода в одежде]
 - [L. *modus* – манера, метод, настрой, мера, образ, способ, правило, предписание]
 - » Модная одежда
 - » Самый популярный, ходовой размер
 - » Высота стопки одежды на полке
 - » Родственные слова: Модель, модератор, модификация, модуль, модуляция

Задача: найти моду

x : 10, 20, 15, 20, 10, 20

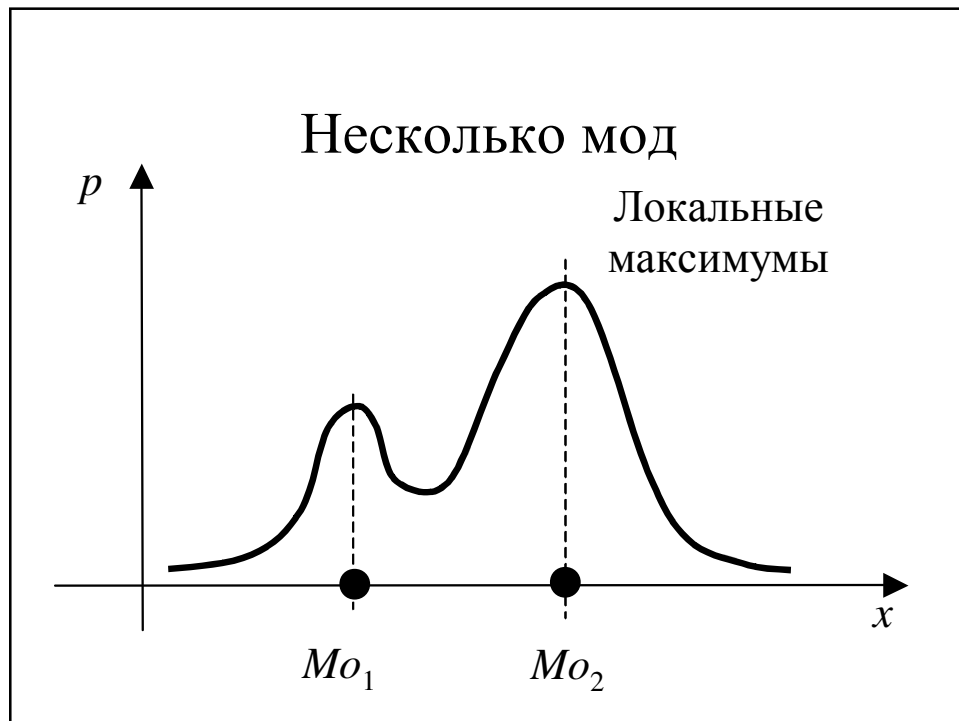
$Mo = ?$

Определить моду по выборке:
Какое число повторяется чаще всего?



Несколько мод

- Наличие нескольких мод свидетельствует о неоднородности совокупности
 - *Однородность* – однотипность, все единицы совокупности имеют общие свойства
 - *Неоднородность* – наличие принципиально разных объектов:
 - дети и взрослые
 - частные и государственные предприятия



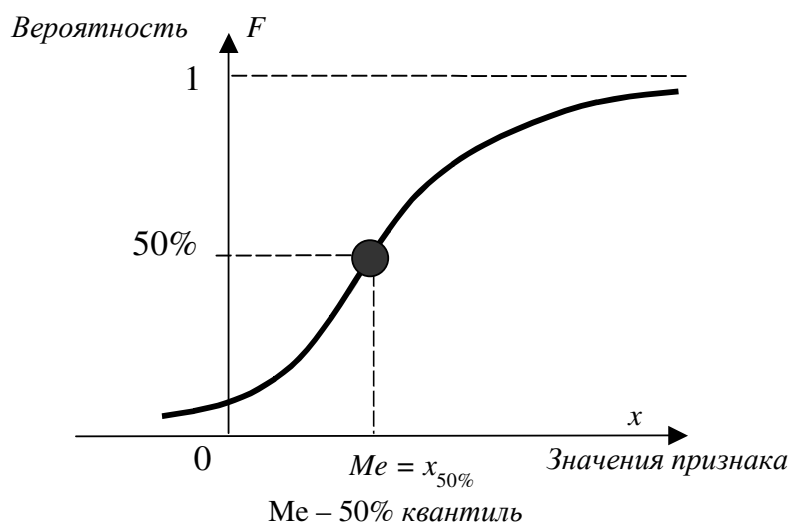
Медиана

- 50% квантиль
- Слева и справа – половина совокупности
- Центральное значение
- Упорядочить и взять «середину»

Медиана

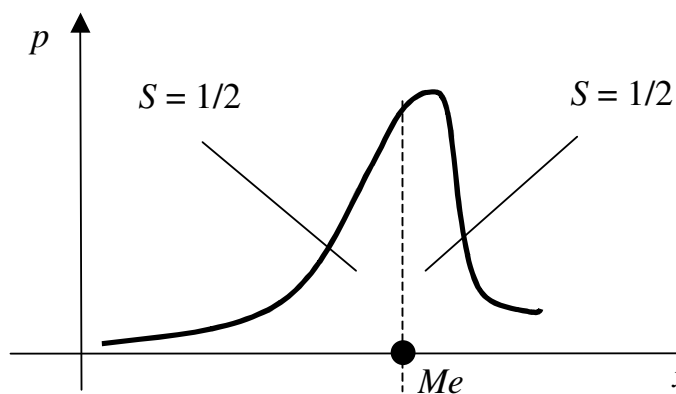
- [**L.** *medianus* – средняя, разделяющий пополам]
- 1. (геом.) линия, соединяющая вершину треугольника с серединой противоположной стороны
- 2. (стат.) срединное, центральное значение
 - «Медиа» = середина, среда
 - Родств. слова: Медиатор, медиевист, медиум, мультимедиа, масс-медиа

Медиана на графике

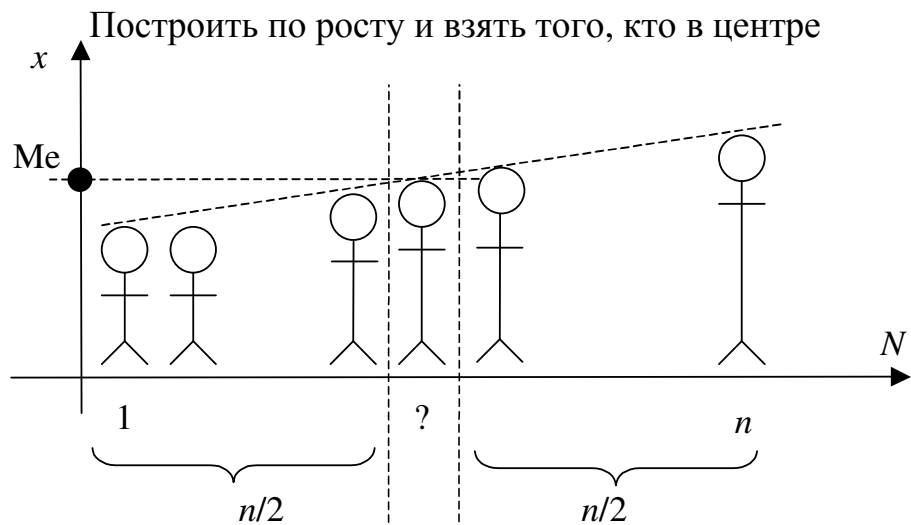


Медиана

- делит пополам площадь геометрической фигуры



Медиана по выборке



Медиана

- Если n – нечетное:

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

- Пример:

– $n = 5$, Me – 3-й элемент упорядоченного ряда

- Если n – четное:

– Me – среднее арифм. двух центральных элементов

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

- Пример:

– $n = 6$, Me – среднее арифм. 3-го и 4-го элементов упорядоченного ряда

Задача: найти медиану

x : 10, 20, 15

$Me = ?$

Упорядочить по возрастанию

Взять центральное значение

Задача: найти медиану

x : 10, 20, 15, 5

$Me = ?$

Упорядочить по возрастанию

Взять среднее арифметическое двух
центральных значений

Стандартные распределения

- *нормальное*: сумма независимых случайных величин
- *хи-квадрат*: сумма квадратов нормальных
- *Стьюдента*: отношение нормальной и корня из хи-квадрат
- *Фишера*: отношение двух хи-квадратов

Стандартные распределения

- Использование станд. распределений:
 - задают вероятность и находят квантиль: $p \rightarrow x$
 - задают квантиль и находят вероятность: $x \rightarrow p$
- Значение квантилей станд. распределений:
 - специальные функции, например гамма-функция
 - статистические таблицы
 - встроенные функции
- Стандартные распределения – инструмент:
 - приблизительное описание реальных процессов
 - неизбежная погрешность

Нормальное (гауссовское) распределение

- Сумма большого количества независимых случайных величин
 - Стандартное нормальное распределение:
 - нулевое математическое ожидание
 - единичная дисперсия
 - Плотность:

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$
 - функция Гаусса
 - Функция распределения:

$$\Phi(x) = F(x) - \frac{1}{2}$$
 - функция Лапласа
- $z = \frac{x - \mu_x}{\sigma_x}$

Домашнее задание

- Постройте график функции плотности вероятности нормального распределения с помощью MS Excel
 - Используйте функции и операции
 - [Правка – Заполнить – Прогрессия]
 - КОРЕНЬ
 - EXP
 - ^
 - ПИ
 - Диапазон значений
 - $-6 < z < +6$
- $$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Домашнее задание

- Для чего вместо интегральной функции нормального распределения вводится функция Лапласа?
- Почему минус 0,5?

$$\Phi(x) = F(x) - \frac{1}{2}$$

Распределение Пирсона Распределение хи-квадрат

Сумма квадратов n независимых случайных величин со стандартным нормальным распределением

$$\chi_n^2 = z_1^2 + z_2^2 + \dots + z_n^2$$

Число слагаемых n – число степеней свободы

Распределение Стьюдента

- Отношение стандартной нормально распределенной к корню из хи-квадрат, деленной на число степеней свободы

$$t_n = \frac{z}{\sqrt{\chi_n^2 / n}}$$

- t -распределение с n степенями свободы
- при большом числе слагаемых распределения хи-квадрат и Стьюдента приближаются к нормальному

Распределение Фишера

- отношение двух независимых случайных величин с распределением хи-квадрат, деленных на число степеней свободы

$$F_{n_1, n_2} = \frac{\chi_{n_1}^2 / n_1}{\chi_{n_2}^2 / n_2}$$

- F -распределение с числом степеней свободы n_1 и n_2

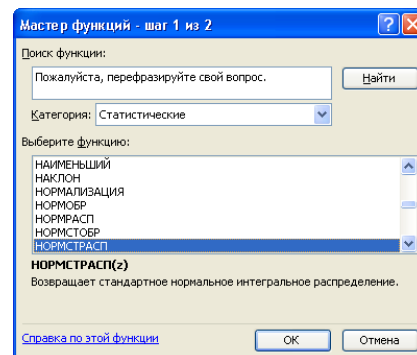
Распределения статистических показателей

- Среднее (дисперсия известна) $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- Дисперсия $D = s^2 = \frac{\sum (x - \bar{x})^2}{n} \sim \chi^2$
- Среднее (дисперсия неизвестна) $\bar{x} \sim t_n\left(\mu, \frac{s}{\sqrt{n}}\right)$
- Отношение дисперсий $\frac{D_1}{D_2} \sim F_{n_1, n_2}$

Домашнее задание

– Статистические функции MS Excel

- освоить НОРМРАСП, НОРМОБР, НОРМСТРАСП, НОРМСТОБР для вычисления значения квантиля по вероятности и значения вероятности по квантилю
- построить графики плотности вероятности и функции распределения для четырех стандартных распределений



Форма распределения

- кривая распределения
 - островершинность, плосковершинность
 - количество вершин
 - наличие симметрии
- характеристики формы распределения
 - асимметрия
 - эксцесс

Моменты

- μ_k - центральный момент k -го порядка
(статистическая характеристика)
$$\mu_k = \frac{\sum (x_i - \bar{x})^k}{n}$$
- Центрированная случайная величина $(x - \bar{x})$
- Вычисление моментов
 - начальный момент
 - исходные значения величин
 - центральный момент
 - отклонения от среднего значения
 - смешанные моменты
 - значения разных случайных величин
 - Среднее значение – начальный момент 1-го порядка
 - Дисперсия – центральный момент 2-го порядка

Асимметрия

- Показатель нарушения симметрии

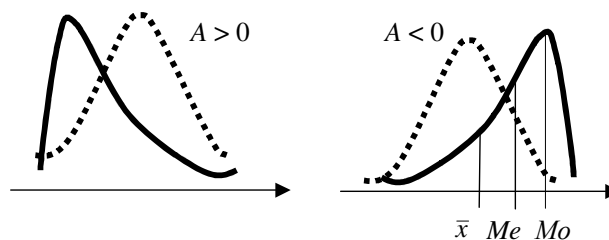
$$A = \frac{\bar{x} - Mo}{\hat{\sigma}}$$

$$A = \frac{\bar{x} - Me}{\hat{\sigma}}$$

$$A = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$$

- μ_3 - центральный момент 3-го порядка
- Для нормального распределения $A = 0$
- *SKEWNESS*
- «скос»

Примеры асимметрии



- Алфавитное правило:
 - Характеристики расположены в алфавитном порядке:
 - Mean – Median – Mode

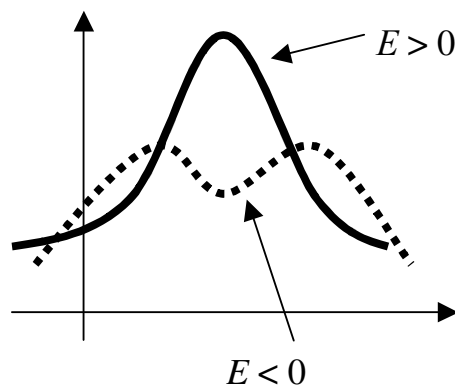
Эксцесс

- показатель островершинности распределения
- мера концентрации распределения вокруг среднего значения

$$E = \frac{\mu_4}{\mu_2^2} - 3$$

- μ_4 - центральный момент 4-го порядка
- Для нормального распределения $E = 0$
- *KURTOSIS*, «островершинность»

Примеры эксцесса



Домашнее задание

- С помощью надстройки «Анализ данных» MS Excel
 - сгенерируйте 4 столбца случайных чисел со стандартными распределениями
 - определите показатели асимметрии и эксцесса

Неоднородность совокупности

- Если обрабатывать данные по **неоднородной** совокупности, то статистические расчеты и заключения будут содержать большую погрешность.
- Статистические выводы делаются на основе стандартных распределений – исходя из предположения однородности.
- Перед углубленным статистическим анализом приходится проверять **однородность**
- Неоднородную совокупность приходится делить на однородные части и обрабатывать