Биостатистика

Кафедра медичної інформатики та комп'ютерних технологій навчання ,
Національний медичний університет імені О.О.Богомольця

18 февраля 2012 г.

Выучи теорию вероятностей за 15мин для чайников

- Вероятность, Случайная величина
- Матожидание и Дисперсия
- Плотность вероятности, Вероятностные распределения
- Центральная предельная теорема

Вероятность, рабоче-крестьянское определение...

Вероятность – это *мера* возможности появления какого-либо случайного события.

Вероятность – это *мера* возможности появления какого-либо случайного события.

$$P(A) = \frac{|A|}{|\Omega|}$$

Вероятность – это *мера* возможности появления какого-либо случайного события.

$$P(A) = \frac{|A|}{|\Omega|}$$

$|A|$ – Количество реализаций события A

$|\Omega|$ – Общее количество всех возможных случаев

Вероятность

$$P(A) = \frac{|A|}{|\Omega|}$$

$$\Omega = \{\text{орел, решка}\}$$
$$A = \text{орел}$$
$$P(A) = \frac{|A|}{|\Omega|} = \frac{1}{2}$$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
$$A = \{5, 6\}$$
$$P(A) = \frac{|A|}{|\Omega|} = \frac{2}{6} = \frac{1}{3}$$

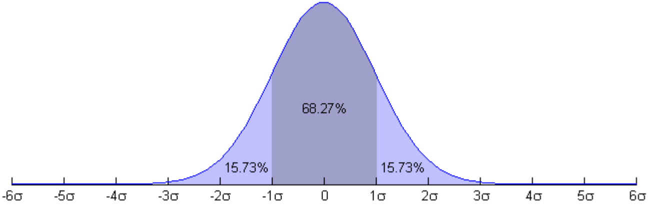
Случайная величина

- **Случайная величина** (почти) любая[1] вещественная функция от случайного события:
  $\xi : \Omega \longrightarrow \mathbb{R}$
- Пример:

$$Y(\omega) = \begin{cases} 1, & \text{if} \quad \omega = \text{heads}, \\ 0, & \text{if} \quad \omega = \text{tails}. \end{cases}$$

[1] измеримая

- **Функция распределения**: $F_\xi(x) = P\{\xi < x\}$
- **Плотность распределения**: Если существует функция $p(x)$, что: $F_\xi(y) = \int_{-\infty}^{y} p(x)dx$, то говорят что случайная величина имеет плотность $p(x)$
- $P(a < \xi < b) = \int_a^b p(x)dx$



- **Математическое ожидание** случайной величины – это средневзвешенное среднее.
- Дискретный случай: $\mathrm{M}[\xi] = \sum_i x_i p_i$
- Непрерывный случай: $\mathrm{M}[\xi] = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}\,x$
- Пример:
  ξ – число, выпавшее после подкидывание кубика d6.
  Возможные значения ξ: 1, 2, 3, 4, 5, 6, все с вероятностью $\frac{1}{6}$. Матожидание X:

$$\mathrm{M}[\xi] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$

Свойства матожидания

- $M[c] = c, \forall c \equiv \mathrm{const}$
- $M[a\xi + b\mu] = aM[\xi] + bM[\mu]$

Дисперсия

- **Дисперсия** – среднее значение квадрата отклонения случайной величины от ее среднего .
- Если ξ имеет **матожидание** $\mu = M[\xi]$, то дисперсия ξ:

$$\mathrm{D}(\xi) = \mathrm{M}[(\xi - \mu)^2]$$

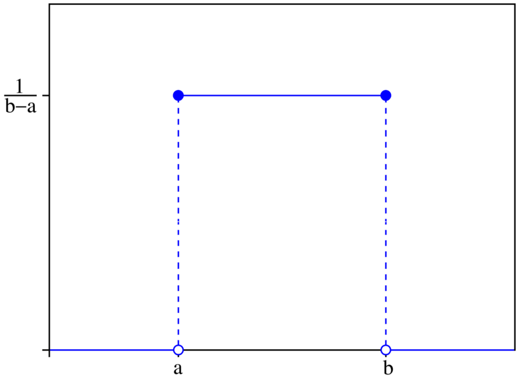
- "Физический смысл": дисперсия – это степень разброса значений
- Для дискретной случайной величины: $\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$, где $\mu = \sum_{i=1}^{n} p_i \cdot x_i$
- ξ – число, выпавшее после подкидывание кубика d6. Дисперсия X: $\sum_{i=1}^{6} \frac{1}{6}(i - 3.5)^2 = \frac{1}{6}\sum_{i=1}^{6}(i - 3.5)^2 = \frac{1}{6}\left((-2.5)^2+(-1.5)^2+(-0.5)^2+0.5^2+1.5^2+2.5^2\right) = \frac{1}{6} \cdot 17.50 = \frac{35}{12} \approx 2.92$

Свойства дисперсии

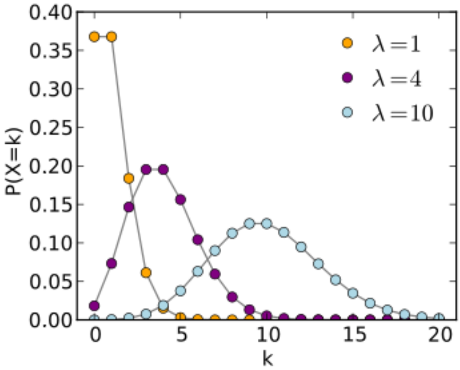
- $\mathrm{D}(\xi) = \mathrm{M}[(\xi - \mu)^2] = \mathrm{M}[\xi^2] - (\mathrm{M}[\xi])^2$
- $D[c\xi] = c^2 D[\xi]$
- $D[\xi] \geq 0,\ D[\xi] = 0 \Leftrightarrow \xi = 0$
- $D[\xi + c] = D[\xi]$

Равномерное распределение

- Плотность: $\begin{cases} \frac{1}{b-a} & \text{если } x \in [a, b] \\ 0 & \text{иначе} \end{cases}$
- $\mathrm{M}[\xi] = \frac{1}{2}(a+b)$, $\mathrm{D}[\xi] = \frac{1}{12}(b-a)^2$

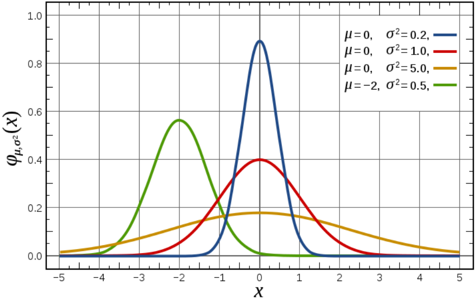


- Плотность: $p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $\mathrm{M}[X] = \lambda$, $\mathrm{D}[X] = \lambda$



Нормальное распределение

- Плотность: $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $\mathrm{M}[\xi] = \mu$, $\mathrm{D}[\xi] = \sigma$



Центральная предельная теорема

Пускай $\xi_1, \xi_2, \dots \xi_n$ –независимые одинаковораспределенные случайные величины. Тогда:

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1), \mu = \mathrm{M}\,\xi_i, \sigma = \mathrm{D}\,\xi_i, S_n = \sum_n \xi_i$$

- ~~Статистика и биология~~
- Эксперимент и наблюдение
- Основные статистические характеристики выборки
- Корреляция и Регрессия
- ~~Тестирование гипотез~~[2]

[2] Рассмотрим на следующей практике

Почему статистика так важна?

Ложь, наглая ложь и **статистика**

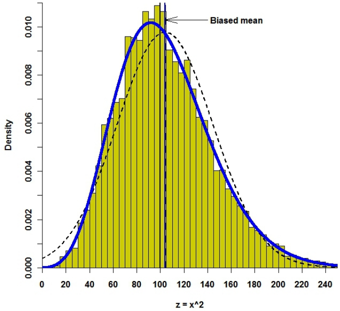
- **Выборка** – наблюдения одной и той же случайной величины, при повторных случайных экспериментах.
- Проще говоря выборка – это результаты эксперимента.
- **Вариационный ряд** - значения выборки, упорядоченные по неубыванию.

Среднее, Мода, Медиана. Дисперсия, Отклонение

- **Выборочное матожидание**: $\bar{x} = \frac{1}{N} \sum_{k=1}^{N} x_k$
- **Мода** – значение, которое встречается в выборке чаще всего
- **Медиана** – значение, которое дели вариационный ряд на 2 половины
- **Выборочная дисперсия**: $\sigma_N^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- **Среднеквадратичное(Стандартное) отклонение**: $\sigma_N = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2}$

Эмпирическая функция распределения:
$F_n(x) = \frac{1}{n}\sum_{i=0}^{n} I_{x_i \le x}$
Гистограмма:

$$p_n(x) = \sum_{k=1}^{N} \frac{1}{|\Delta_k|} v_k I_{\Delta_k}(x),\; v_k = \frac{1}{n}\sum_{n}^{j=1} I_{\Delta_k}(X_j)$$



- **Ошибка репрезентативности** – ошибка, которая вызвана тем, что мы работаем с выборкой, а не всей генеральной совокупностью:

$$m = \frac{\sigma}{\sqrt{N}}$$

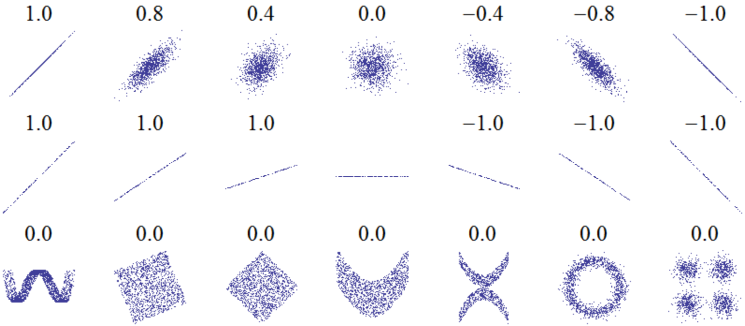
σ – среднеквадратичное отклонение, N – размер выборки

- **Корреляция** – степень зависимости между двумя наблюдаемыми случайными величинами.
- Степень *линейной* зависимости показывает **коэфициент корреляции Пирсона**:

$$\mathbb{R}_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_X \cdot \sigma_Y}$$

Корреляция	$\lvert\mathbb{R}_{X,Y}\rvert$
Отсутствует	0.0 to 0.09
Слабая	0.1 to 0.3
Умеренная	0.3 to 0.5
Сильная	0.5 to 1.0

Корреляция – диаграммы рассеивания



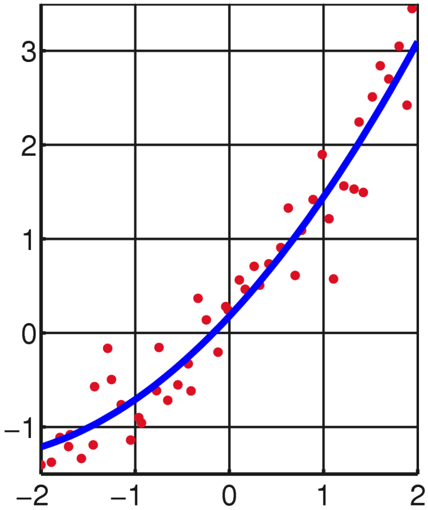
Модель регрессии

- У нас есть две парные выборки X и Y
- Необходимо построить кривую, которая отображает зависимость Y(X)
- Вид зависимости задается эмпирически(те. угадывается)
- Для построения кривой чаще всего пользуются – **методом наименьших квадратов**:
- Из всех возможных кривых мы выбираем такую, что сумма квадратов расстояний до не от всех точек минимальна.

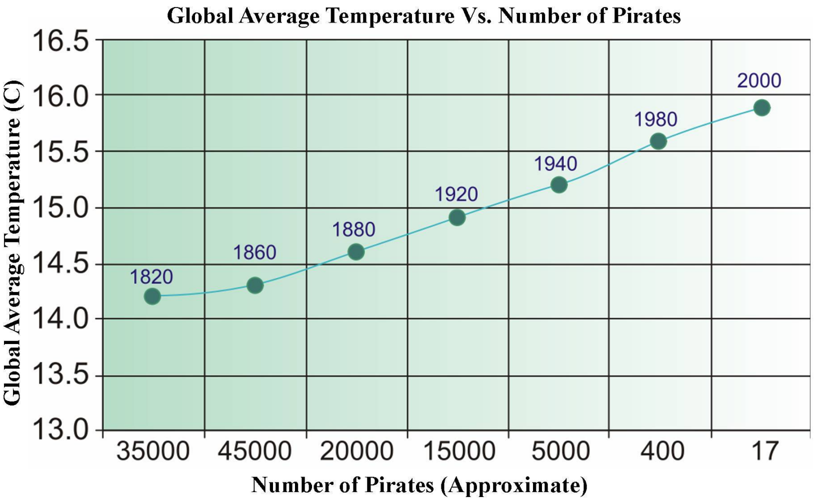
Линейная оценка МНК в модели регрессии

- $y = kx + b$
- Тут $< x >$ – среднее величины x
- $k = \frac{<x^2><y> - <x><xy>}{<x^2> - <x>^2}$
- $b = \frac{<xy> - <x><y>}{<x^2> - <x>^2}$

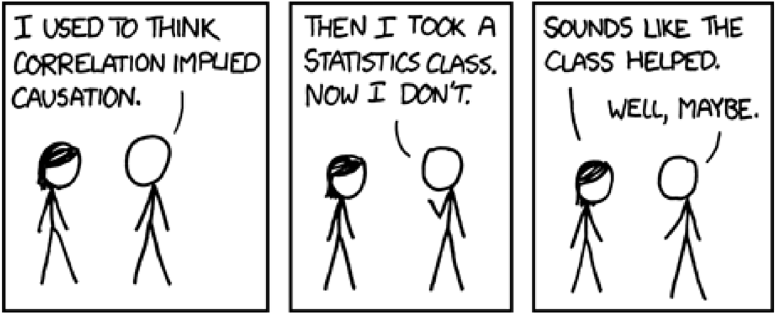
Модель регрессии



Post hoc ergo propter hoc



Post hoc ergo propter hoc



Dixi