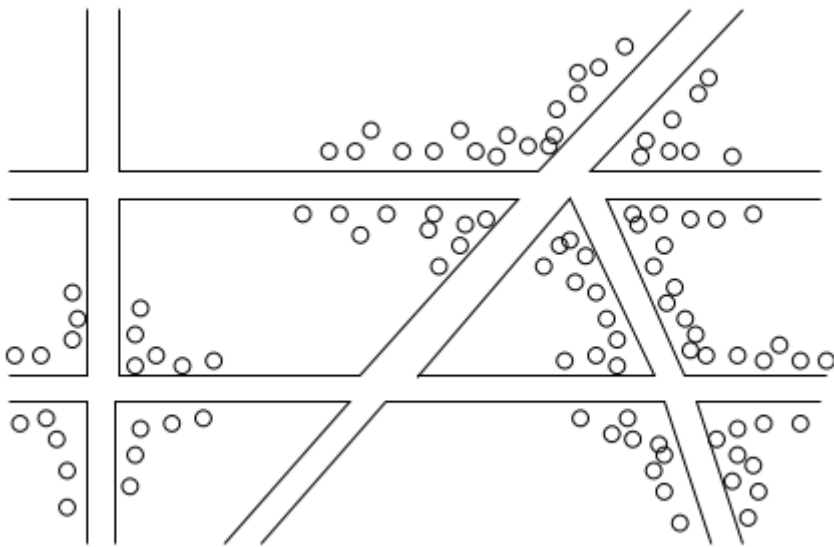


Lecture 02 Random variables, measures of central tendency and distributions 随机变量 集中趋势的度量 分布

- Topics:
 - Logistics and Recap
 - Introduction(continued)
 - Random variable and distributions
- 注意:任何的数据研究都需要提前深入了解现有的数据是什么、代表什么含义，只有在足够深入了解现有数据的基础上才可以做算法分析、数据处理等操作.
- tips:<https://lexfridman.com/>

数据挖掘是个新领域吗？



- 1854年伦敦霍乱，约翰·斯诺根据死亡地图推断出水井是导致霍乱传染的根源
- 数据挖掘 vs 机器学习
 - 数据挖掘:一个跨学科领域拥有大数据集，专注于发掘模型(数学、计算机科学、软件工程、debug、可视化、社会分析、心理学)
 - 这是如何完成的？
 - 机器学习是一种方式(使用算法发现数据集中的模型) 在连串的语句中，机器需要填充所在预警与回忆之前的语句
 - 机器学习就在我们周围
 - Netflix
 - Google
 - Amazon

- imdb
- 我们听说过很多故事，自动驾驶汽车、情绪感音环境，为什么尿片和啤酒会放在一起等等

计算机科学	机器学习
确定的规则	泛化/归纳是关键Generalization
错误是无法容忍的	错误是结果的一部分 新冠检测 自动驾驶
算法不会学习而被提供	算法学习(反向传播backprop、遗传编程genetic programming)
编程(数据)==> 输出 程序本身最重要	数据(程序)==> 模型 ==>输出 数据本身最重要

- 机器学习的问题:会归纳出之前没见过的结果
 - 但是我们可以简单的看到所有的数据吗?
- 数据挖掘与机器学习的讨论
 - ACM KDD (Knowledge Discovery and Data Mining), <http://www.kdd.org>
 - ICML (International Conference on Machine Learning)
 - ACM CIKM (International Conference on Information and Knowledge Management)
 - SDM (SIAM International Conference on Data Mining)
 - NeurIPS
- 期刊
 - IEEE Transactions on Pattern Analysis and Machine Intelligence
 - ACM Transactions on Knowledge Discovery from Data
 - IEEE Transactions on Knowledge and Data Engineering
- 相关资源
 - Kaggle (<https://www.kaggle.com>)
 - Kdnuggets (<https://www.kdnuggets.com>)
 - <https://machinelearningmastery.com>
 - <https://towardsdatascience.com>

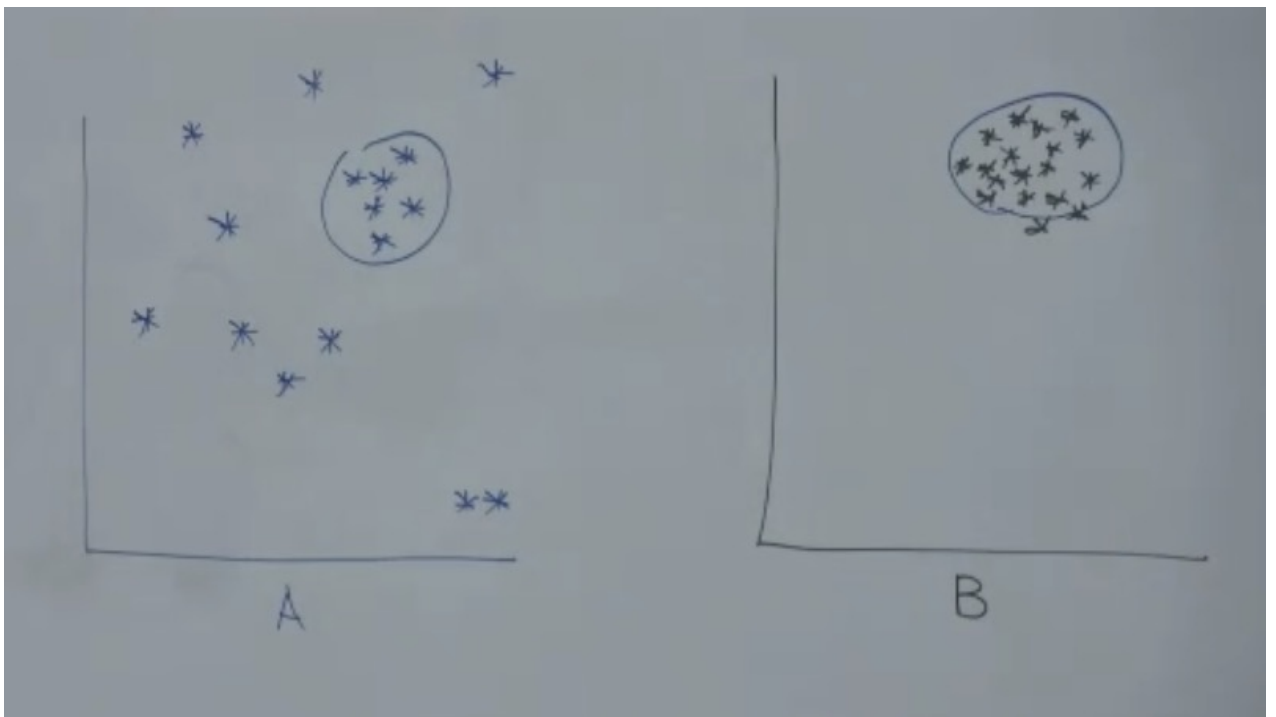
随机变量 Random variables

- 随机变量，X是一个变量，其可能的值是从一个随机环境中获取
 - 投掷硬币
 - 掷骰子
- 两种类型

- 离散
- 连续
- 总体Population vs 样本sample
 - 假定现在有个列向量, D:

$$D = \begin{pmatrix} X \\ - \\ x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \quad X \in \mathbb{R}^n$$

- 我们假设这些数据都是随机样本从X中取出, 每个 x_i 都是独立的(iid)
- 通常来说, 从X中取出的分布都是未知的, 并且矩也是未知的(moment)
 - 零阶矩, (概率)质量的总和, 也就是1
 - 一阶原点矩, 即均值mean, 衡量数据的平均水平 表位置
 - 二阶中心矩, 方差variance, 衡量数据的离散/集中程度 表胖瘦
 - 二阶原点矩, EX^2 , 衡量数据移动至平均位置所需要的平均能量
 - 三阶中心矩, 偏度, 衡量偏离中心的点的位置情况, 也就是偏离中心的点的平均水平(正负、大小) 表歪斜
 - 四阶中心矩kurtosis, 峰度, 俗称方差的方差, 衡量偏离中心的点的密集程度, 是俗话说的尖峰厚尾的理论基础 表尾巴胖瘦
 - 我们仅有样本, 希望取出的样本内分布接近总体内分布
 -



- 从上图展示的内容来看, 图A内数据相对分散, 图B数据更加聚类话, 对于机器学习和数据挖掘来说更愿意分析/处理图A数据, 图B的数据最终可以得到的结论只能是在有

点大那块区间内.图A中既存在了部分聚类也出现了部分分散的情况，就数据方来看更具说服力，我们想要离散更高的数据集.

- 假定我们给出离散随机变量X，有以下几种概率分布

- $$\begin{cases} 1 & 20\% \\ 2 & 30\% \\ 3 & 40\% \\ 4 & 10\% \end{cases}$$

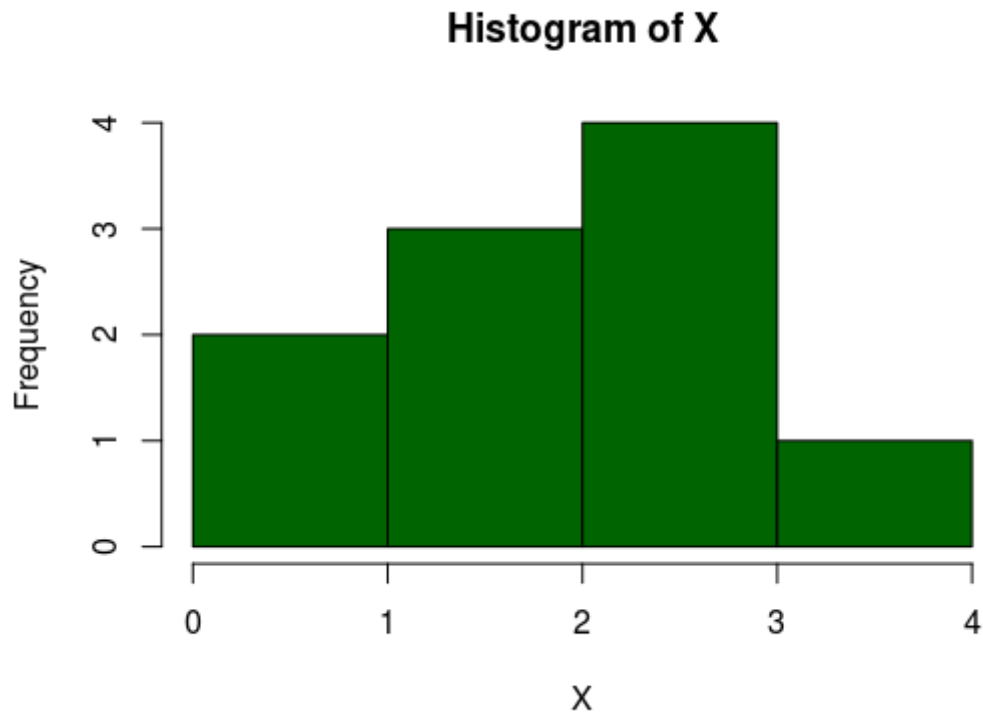
- 那么随机变量X的出现概率可以用以下公式和直方图表示

- $$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^n I(x_i = x)$$

where

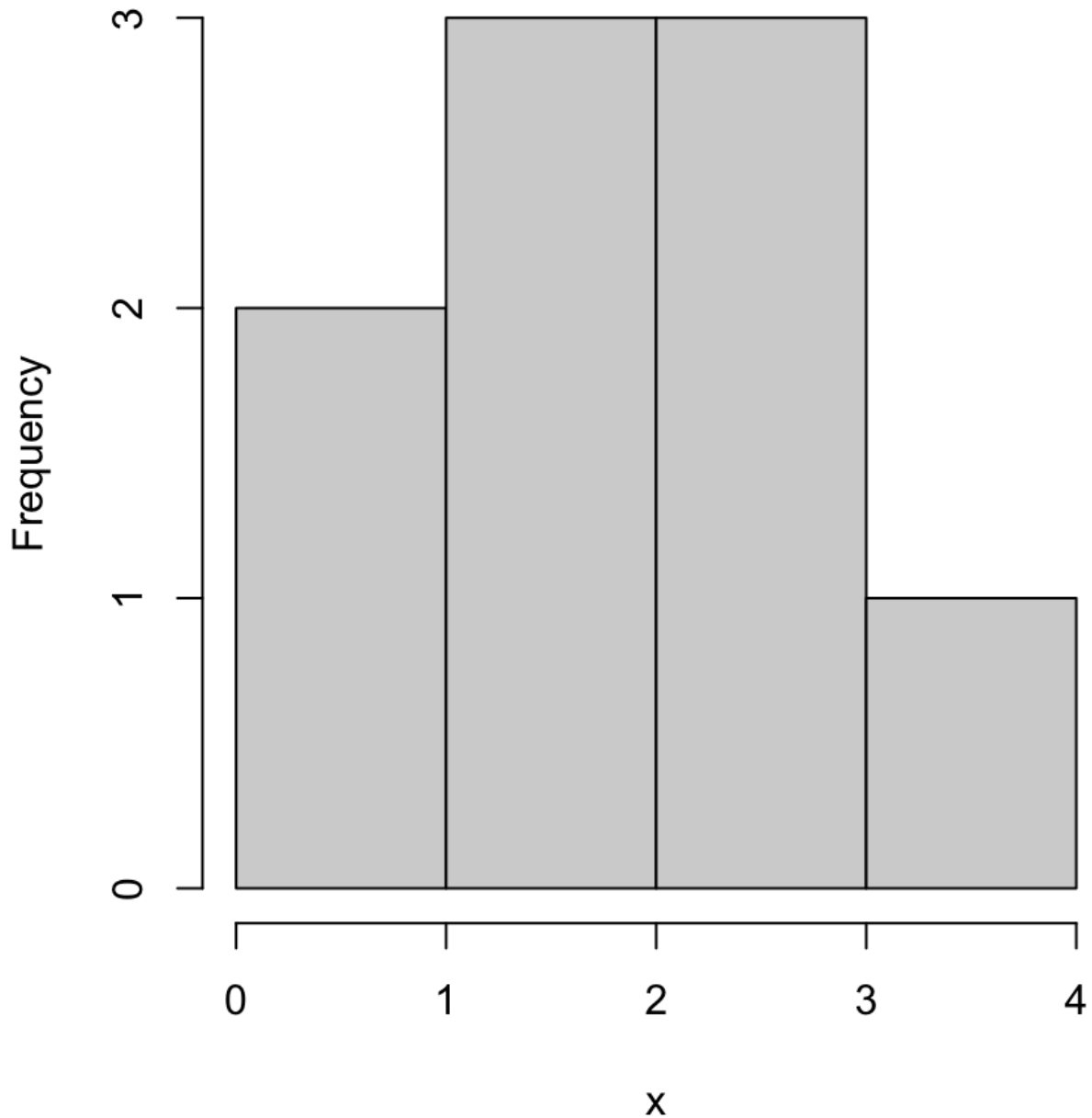
$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

-

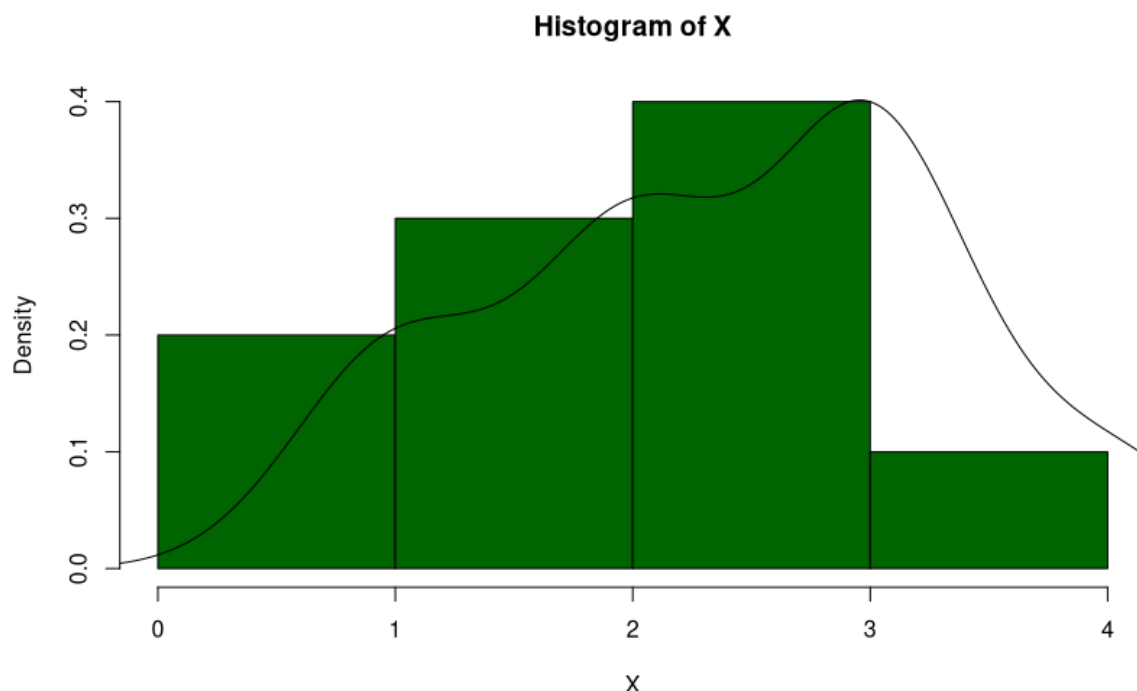


```
x <- c(1,1,2,2,2,3,3,3,4)
a <- hist(x, c(0.0,1.0,2.0,3.0,4.0))
plot(a)
```

Histogram of x



- 与PMF(probability mass function)相关的是PDF(probability density function)
 - PMF:概率质量函数是离散随机变量在各特定取值上的概率
 - PDF:概率密度函数是一个描述这个随机变量的输出值，在某个确定的取值点附近的可能性的函数
 - CDF(cumulative distribution function):累积分布函数又叫分布函数，是概率密度函数的积分，能完整描述一个实随机变量X的概率分布
 - 密度图可视化是通过绘制数据的概率分布适当的连续曲线



集中趋势的度量 Measures of central tendency

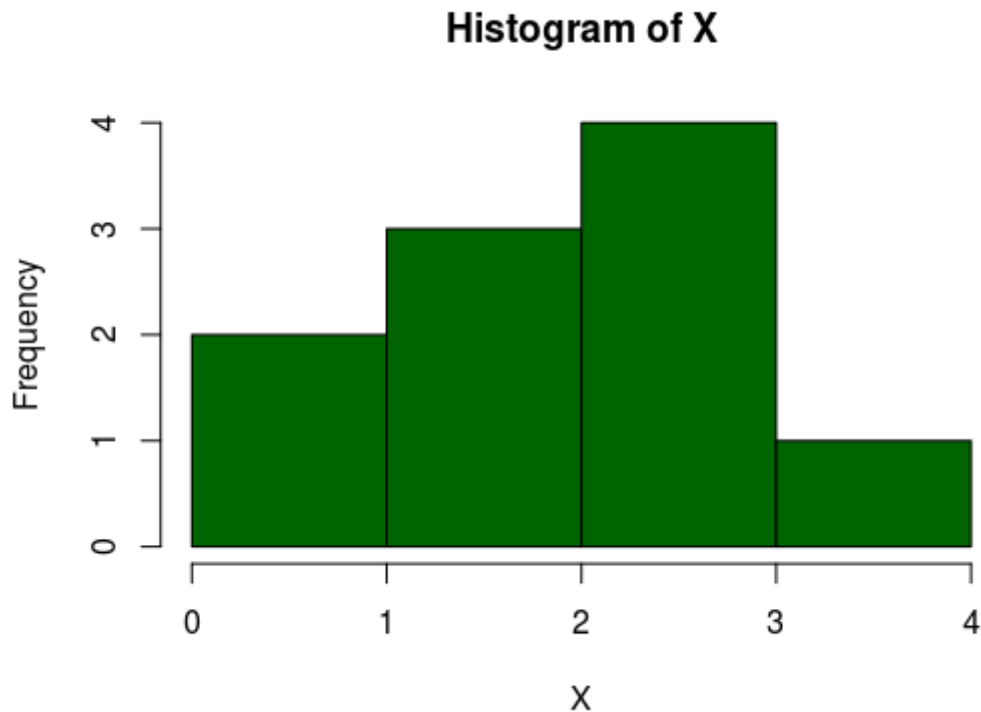
- 平均值mean(sample):

- $$\hat{\mu} = \frac{1}{n-1} \sum_{i=1}^n x_i$$

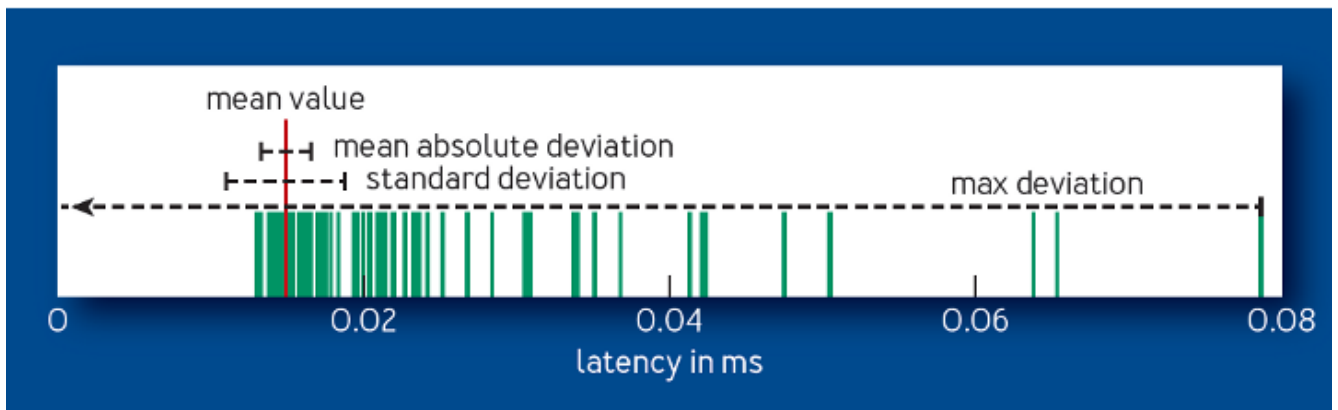
- 平均值是否稳健robust(或稳定stable)
 - 我们将稳健型定义为不受极值影响
 - 通常来说，平均值并不是稳健的
 - 一个稳健的度量是修剪平均值trimmed mean，它发生在两边的极值被丢弃后
- 对随机变量r.v random variable的期望(与平均值有关，但概念上不同)
 - 平均数（均值）是一个统计学的概念；期望是一个概率论的概念
 - 平均数是实验后根据实际结果统计得到的样本的平均值；期望是实验前根据概率分布“预测”的样本平均值
 - 之所以说是预测是因为 在实验前能得到的期望与实际实验得到的样本的平均数总会不可避免的存在偏差，毕竟随机实验的结果永远充满着不确定性
 - 如果我们能进行无穷次随机实验并计算出其样本的平均数的话，那么这个平均数其实就是期望。但是实验样本的平均数会随着实验样本的增多越来越接近期望，就像频率随着实验样本的增多会越来越接近概率一样

- $$E[X] = \sum_{i=1}^k x_i p_i = x_1 p_1 + x_2 p_2 + \dots + x_k p_k$$

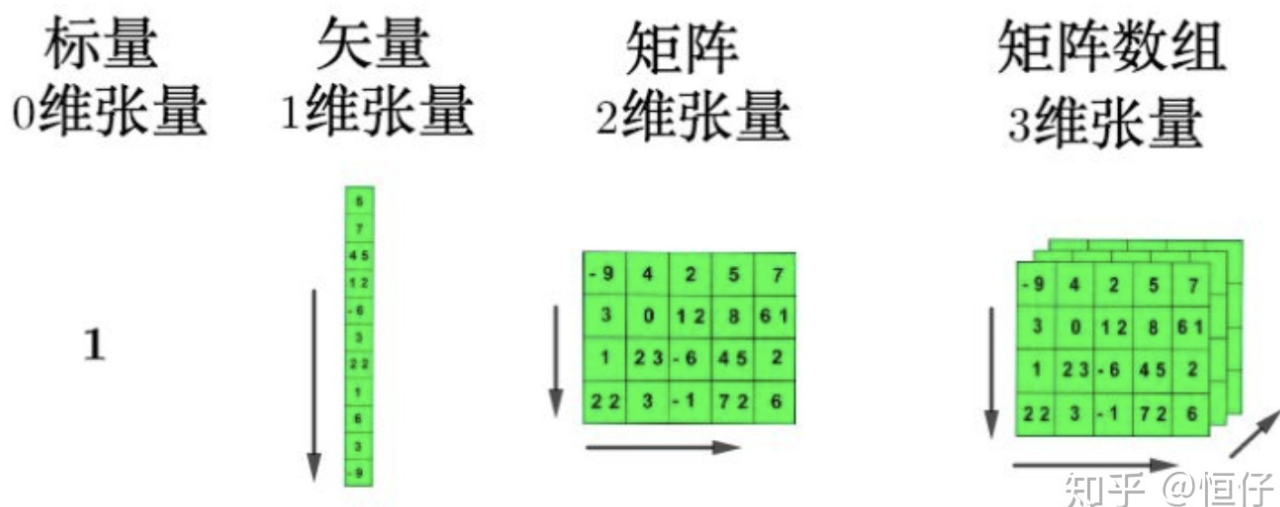
- 离散型随机变量的一切可能的取值，与对应的概率乘积之和
 - 期望的属性
 - $E[X + Y] = E[X] + E[Y]$ 期望的线性 X与Y的期望和等于两个期望分开计算后和
 - $E[aX] = aE[X]$ a是个常量
 - $E[XY] = E[X] \cdot E[Y]$ X和Y是独立的
 - $E[E[X]] = E[X]$
 - 例子:骰子有6面，掷到每一面的概率相同，都是六分之一
 - 平均值:3.5
 - 期望值:如果投掷的次数够多，期望值就越接近平均值即3.5
 - 中位数median(sample)
 - $P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq \frac{1}{2}$
 - 中位数是否robust(或稳定stable)
 - 是的
 - 不受极端值影响
 - 取的是随机变量中的实际值
 - 众数mode(sample)
 - 在随机变量中出现最多的值
 - 在剧中趋势的度量中可能用不到
 - $mode(X) = \arg \max \hat{f}(x)$
- $$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^n I(x_i = x)$$
- where
- $$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$
-



- 离散程度描述:方差variance和标准差standard deviation
 - 方差:衡量随机变量或一组数据时离散程度的度量, 概率论中方差用来度量随机变量和其数学期望(即均值)之间的偏离程度
 - 样本方差公式: $var(X) = \sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$
 - 样本标准差:样本方差的开方
 - 样本标准差公式: $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2}$
- 并不仅仅只有标准差
 - 最大偏差Maximal deviation: $maxdev(X) = \max(|x_i - \mu|) \forall x_i \in X$
 - 平均离差Mean absolute deviation: $mad(X) = \frac{1}{n} \sum_{i=1}^n |x_i - \mu| \forall x_i \in X$
 - FIGURE 11: **A REQUEST LATENCY DATASET**



- 同样我们会考虑双变量分析，试图理解双变量 x_1 与 x_2 的关系，通常我们可以将向量看作是一个2D空间
 - 标量Scalar:标量只有大小概念，没有方向的概念，通过一个具体的数值就能表达完整。
 - 重量、温度、长度、提及、时间、热量等都数据标量
 - 向量Vector:物理学上也叫矢量，指由大小和方向共同决定的量(跟「标量」相区别).如力、速度等.向量主要有2个维度：大小、方向
 - 矩阵Matrix:矩阵（Matrix）是一个按照长方阵列排列的复数或实数集合，元素是实数的矩阵称为实矩阵，元素是复数的矩阵称为复矩阵.而行数与列数都等于n的矩阵称为n阶矩阵或n阶方阵.由 $m \times n$ 个数排成的m行n列的数表称为m行n列的矩阵，简称 $m \times n$ 矩阵
 - 张量Tensor:实际上就是一个多维数组(multidimensional array)，其目的是能够创造更高维度的矩阵、向量，其具有3个属性
 - 维度rank:number of dimensions
 - 行列数shape:number of rows and columns
 - 元素类型type:data type of tensor's elements



- 矩阵的一阶二阶中(平均值、方差)与之前计算一致，除非返回的不是向量
 -

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

- 每个属性的方差分布计算， σ_1^2 代表 X_1 ， σ_2^2 代表 X_2
- 总方差 $var(D) = \sigma_1^2 + \sigma_2^2$
- 关联度量:协方差Covariance

- 协方差在概率论和统计学中用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况.协方差表示的是两个变量的总体的误差，这与只表示一个变量误差的方差不同.如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值.如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值.

- 公式:
$$\text{cov}(X_1, X_2) = \hat{\sigma}_{12} = \frac{1}{n-1} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$

- 对于有两属性的方差协方差variance-covariance可以表示成如下:

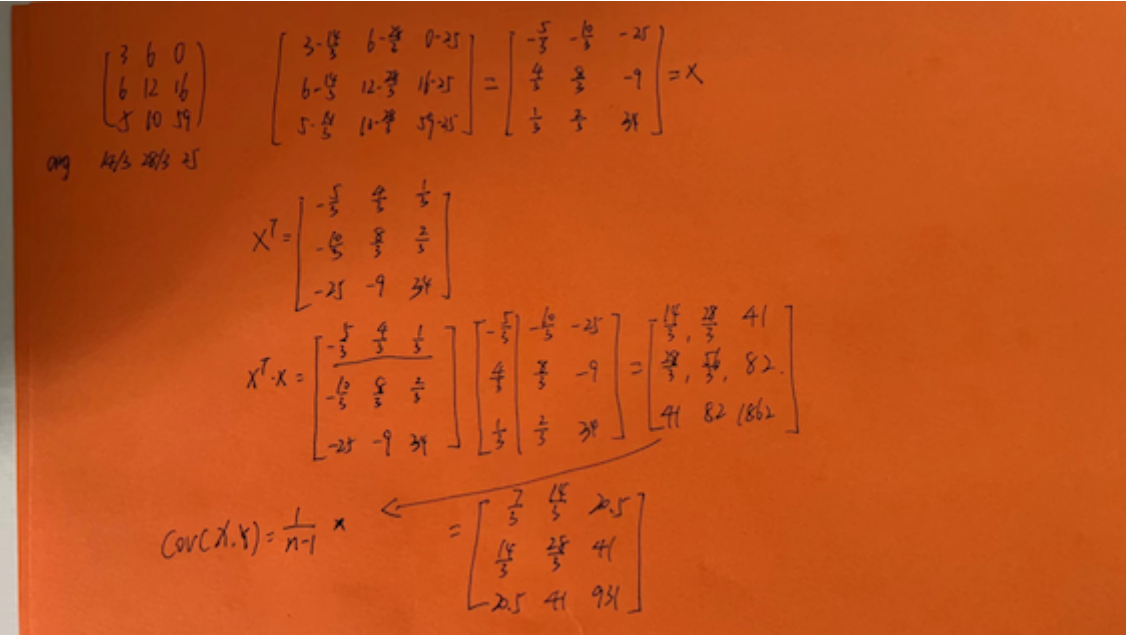
- $$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}$$

- 正方形

- 对称性

- 对角线上的元素为各个随机变量的方差，非对角线上的元素为两两随机变量之间的协方差

- 总方差为对角线的值和，通常以 \sum 表示

- 

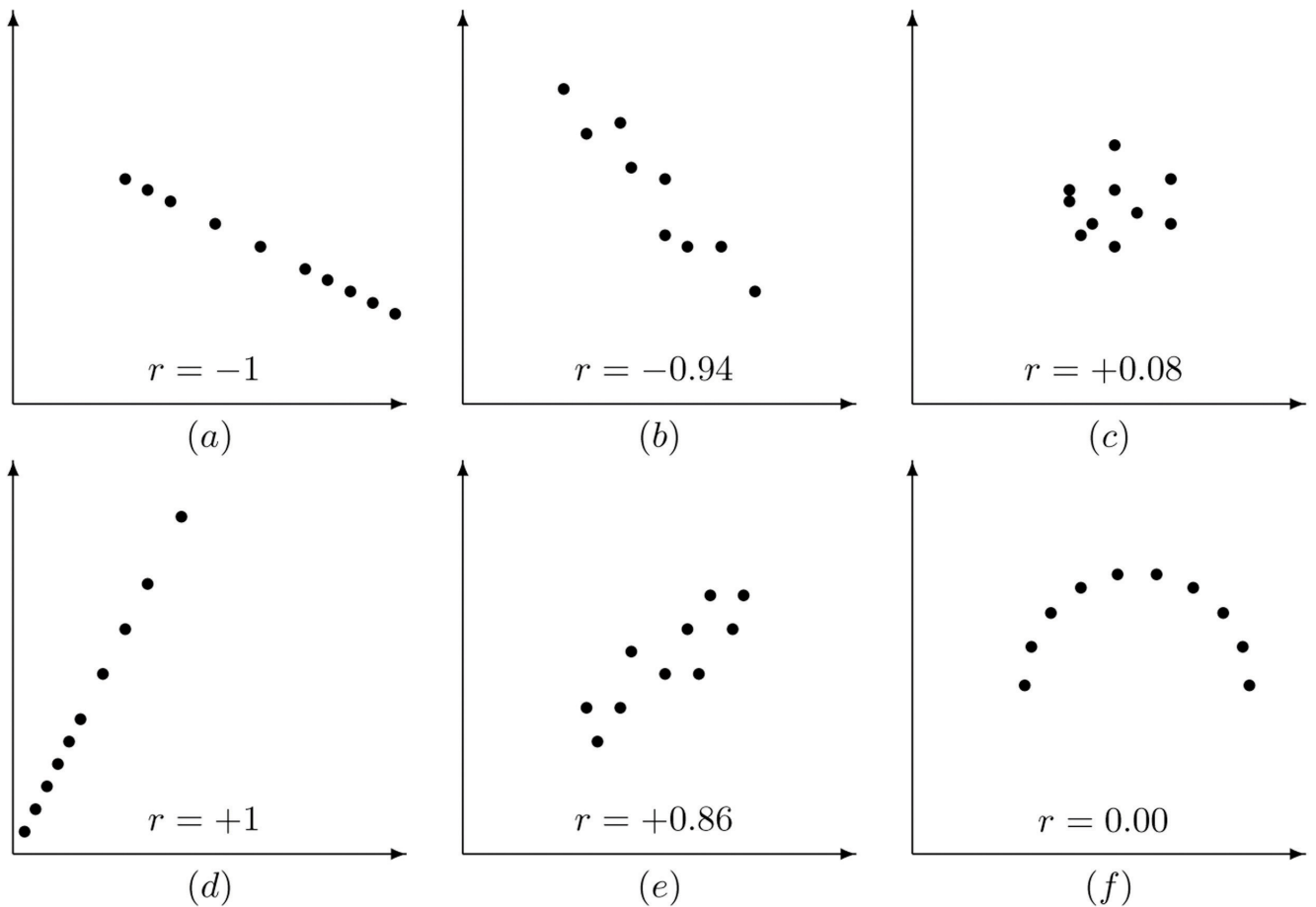
- 与协方差有关系的是相关

-

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}}$$

- 为什么既有协方差又有相关性

- 协方差的值范围是: $[-\infty, \infty]$
 - 相关性的范围是: $[-1, 1]$
 - 相关性是无量纲，它只是变量之间关系的无单位度量
 - 相关性并不代表因果关系
 -



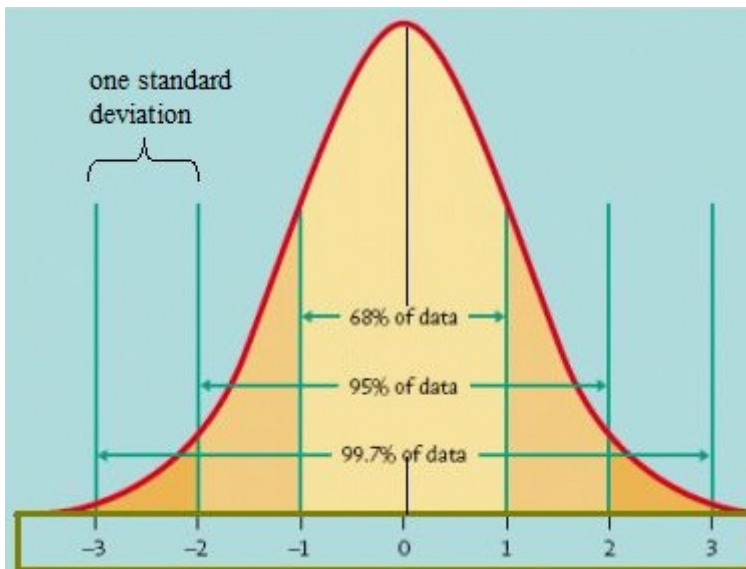
- 相关性与共线性collinearity
 - 相关性是用来度量两个变量之间的关系
 - 当两个变量相关性很高时，可以通过一个来预测另一个，换句话说一个变量是另一个变量的线性组合
 - 当大于两个预测变量是相互关联

分布 Distributions

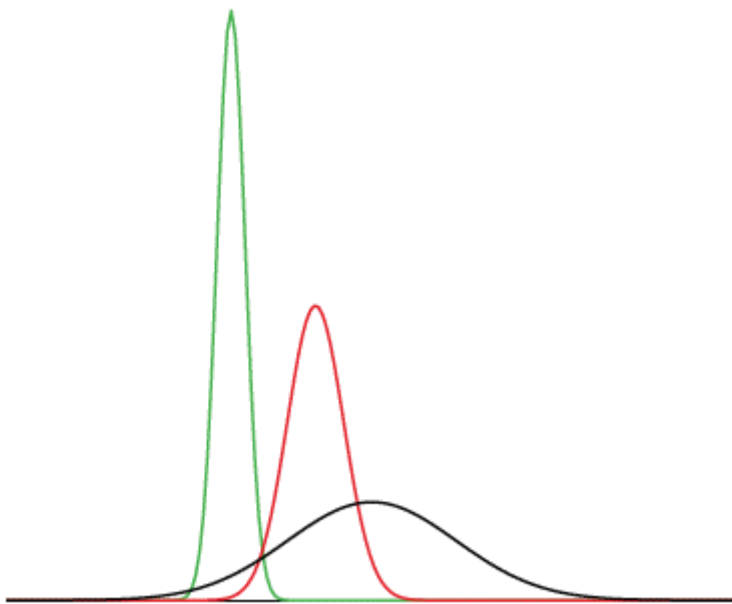
- 正态/高斯分布 Normal/Gaussian distribution
 - 参数有均值(μ)和标准差(σ)
 - 公式为:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ 是正态分布的位置参数，描述正态分布的集中趋势位置。概率规律为取与 μ 邻近的值的概率大，而取离 μ 越远的值的概率越小。正态分布以 $X=\mu$ 为对称轴，左右完全对称。正态分布的期望、均数、中位数、众数相同，均等于 μ 。
- σ 描述正态分布资料数据分布的离散程度， σ 越大，数据分布越分散， σ 越小，数据分布越集中。也称为是正态分布的形状参数， σ 越大，曲线越扁平，反之， σ 越小，曲线越瘦高。
-



○



■ 图形特征:

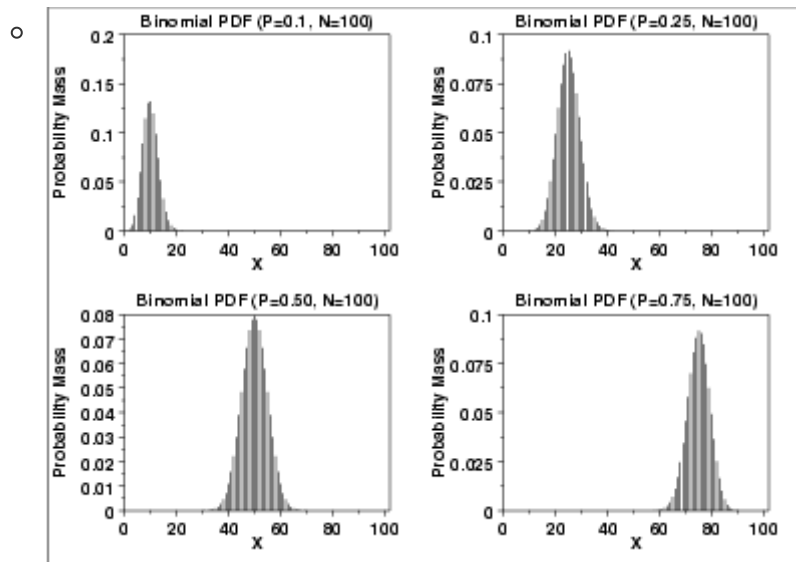
- 集中性：正态曲线的高峰位于正中央，即均数所在的位置。
- 对称性：正态曲线以均数为中心，左右对称，曲线两端永远不与横轴相交。
- 均匀变动性：正态曲线由均数所在处开始，分别向左右两侧逐渐均匀下降。

• 二项分布 Binomial distribution

- 参数有单次事件发生的概率 p 和尝试的次数 n
- 公式为:

$$\binom{n}{x} (p)^x (1-p)^{(n-x)} \quad \text{for } x = 0, 1, 2, \dots, n$$

- 平均值/期望= np
- 中位数: $\lfloor np \rfloor$ 或 $\lceil np \rceil$
- 方差: $np(1-p)$
- 特点:独立的、可重复的
- 使用场景:结果只有两种 yes no



- 泊松分布 Poisson distribution

- 参数为 λ 是单位时间(或单位面积)内随机事件的平均发生次数 事件之间独立

- 公式:

$$\frac{\lambda^k e^{-\lambda}}{k!}$$

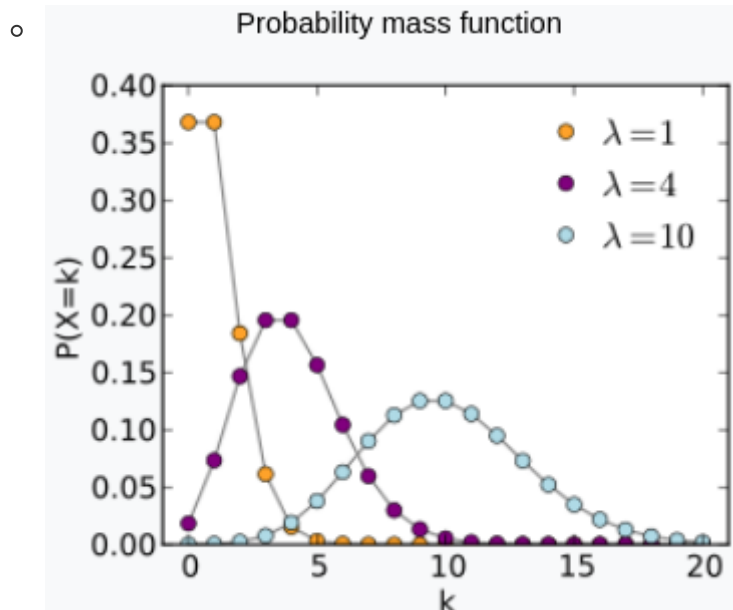
- 平均值/期望: λ

- 中位数:

$$\lambda - \ln 2 \leq \nu < \lambda + \frac{1}{3}.$$

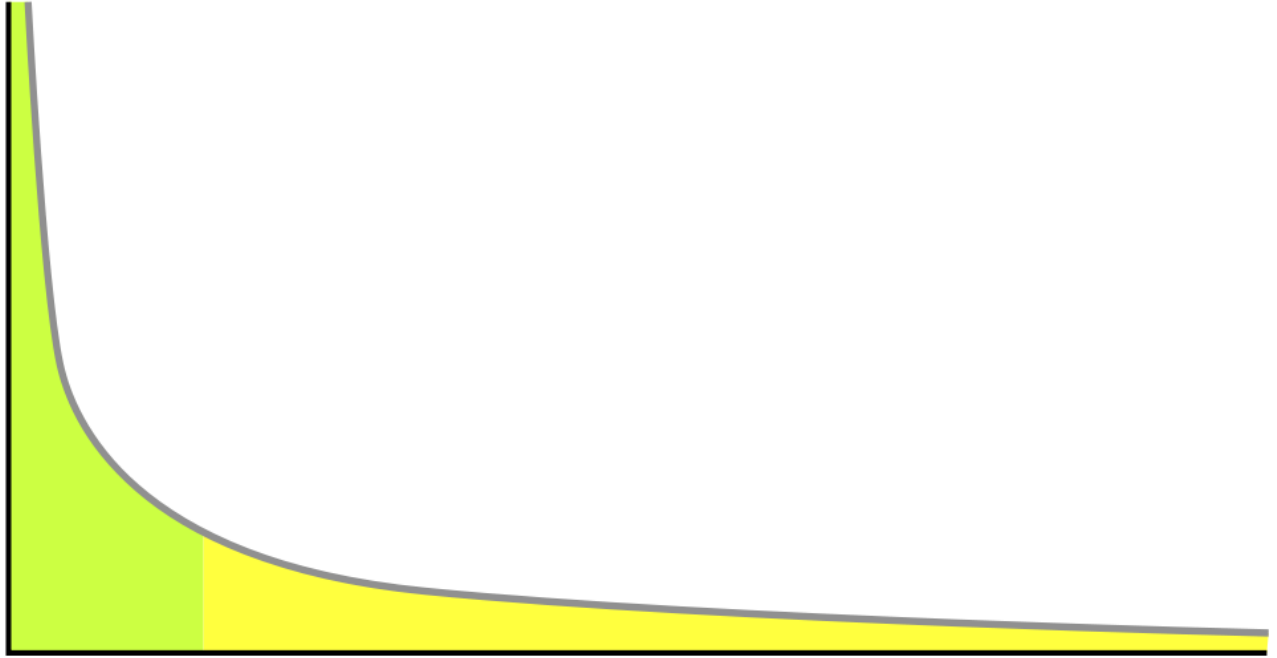
- 方差: λ

- 使用场景/例子:某电话交换台收到的呼叫、来到某公共汽车站的乘客、某放射性物质发射出的粒子、显微镜下某区域中的白血球等等,以固定的平均瞬时速率 λ (或称密度)随机且独立地出现时,那么这个事件在单位时间(面积或体积)内出现的次数或个数就近似地服从泊松分布 $P(\lambda)$ 。

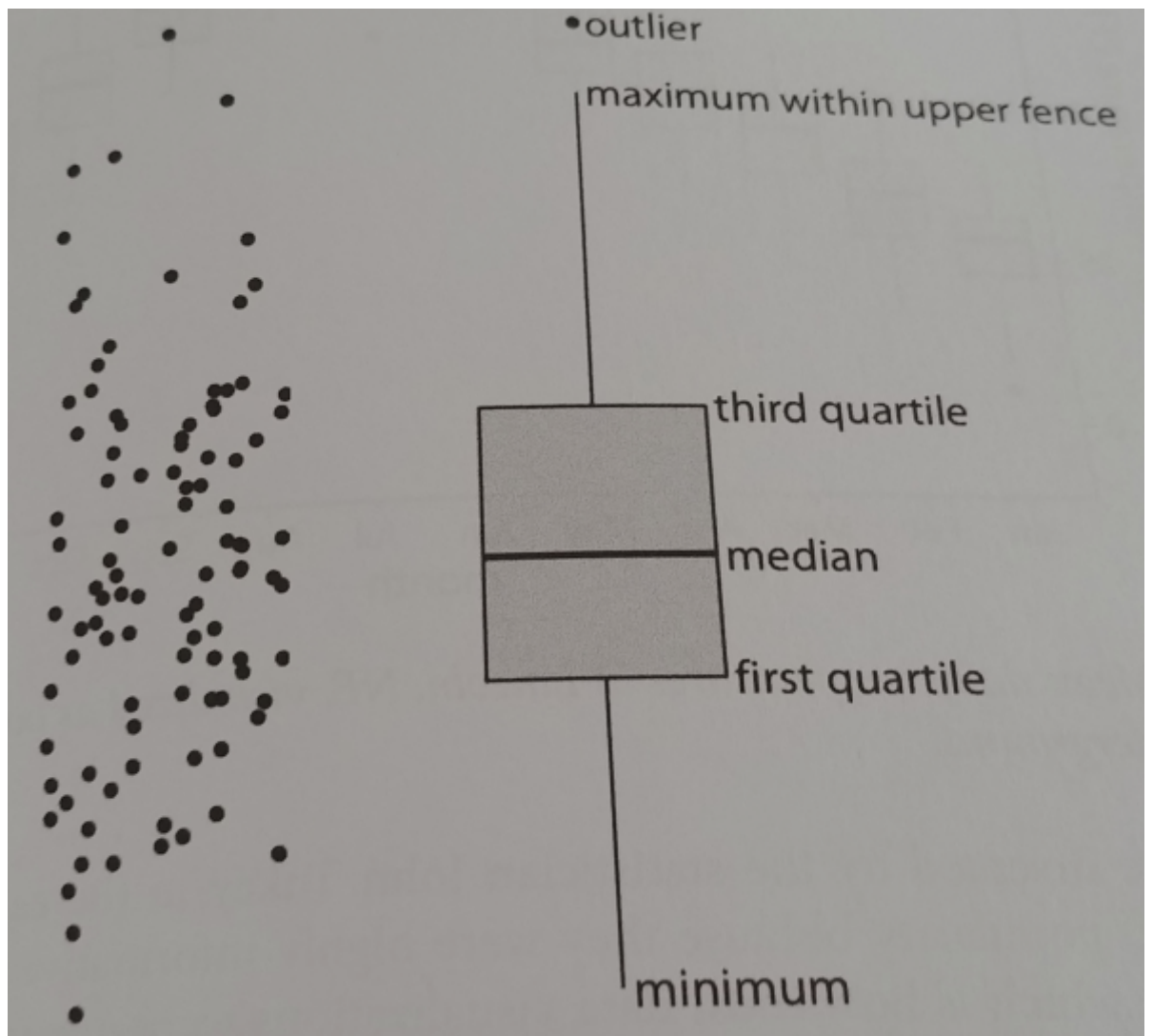


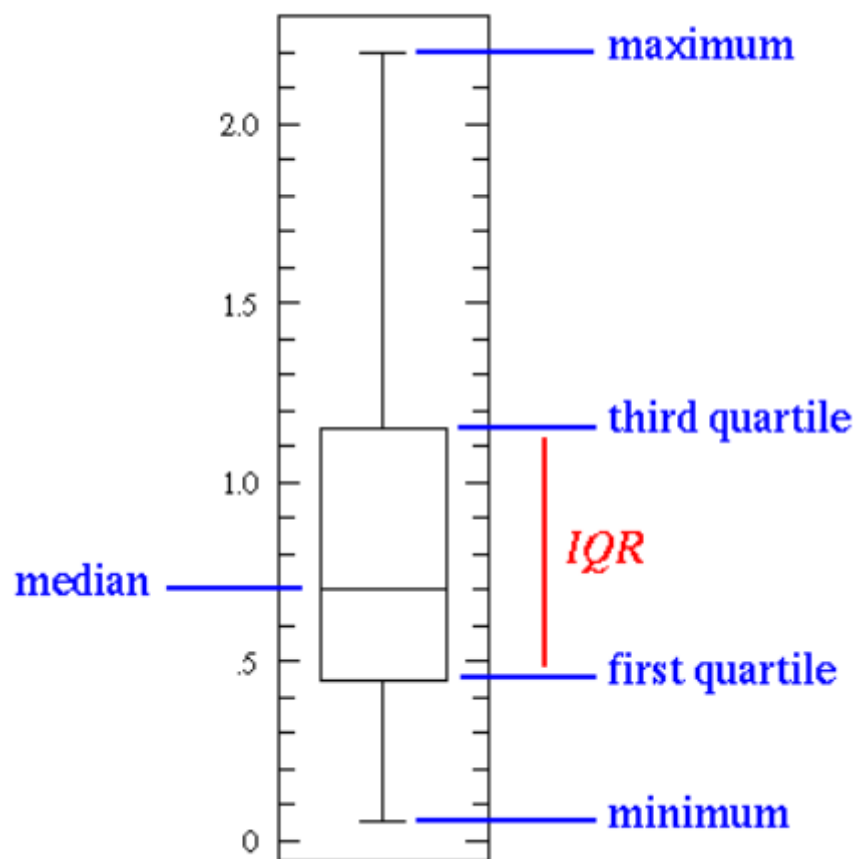
- 幂次分布律 Power-law distributions

- 一个变量在数据中起到决定性作用(随着正方形边长增加, 面积大幅度增加)
- 例子:公司内的20-80原则, 20%的人做着80%重要的工作
- 同样被称作长尾巴分布
- 不同于其它分布, 很难定义为幂次分布律, 在某一时刻下可以定义, 但是在其它时刻很难说。
-



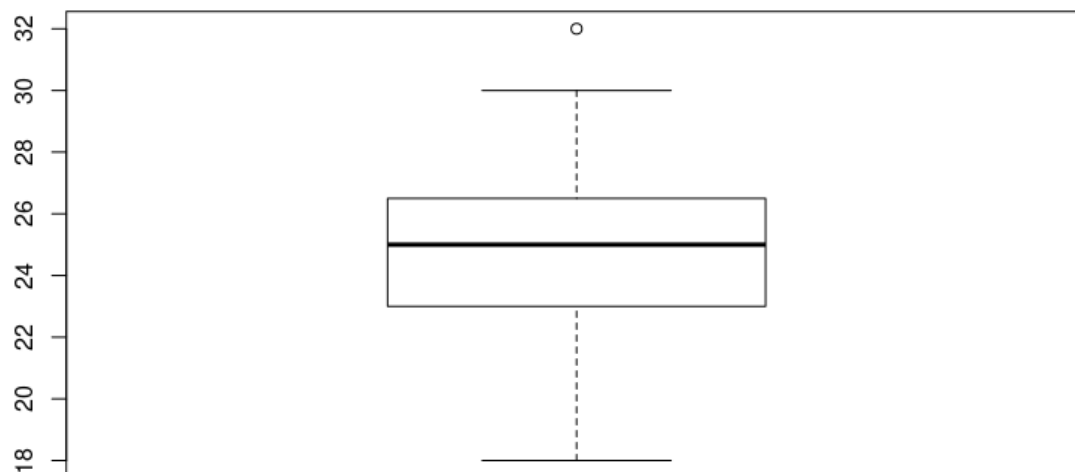
- 可视化很重要
 - 额外资源: Claus O. Wilke, Fundamentals of Data Visualization, O'Reilly Publishing. (Uses R/RStudio/R Markdown.)
 - 没时间仔细了解图中的细节, 但是可以通过
 - 箱线图
 -





■ **Console** ~/ ↻

```
> summary(a)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00  23.00   25.00   24.80  26.25   32.00
> boxplot(a)
> |
```

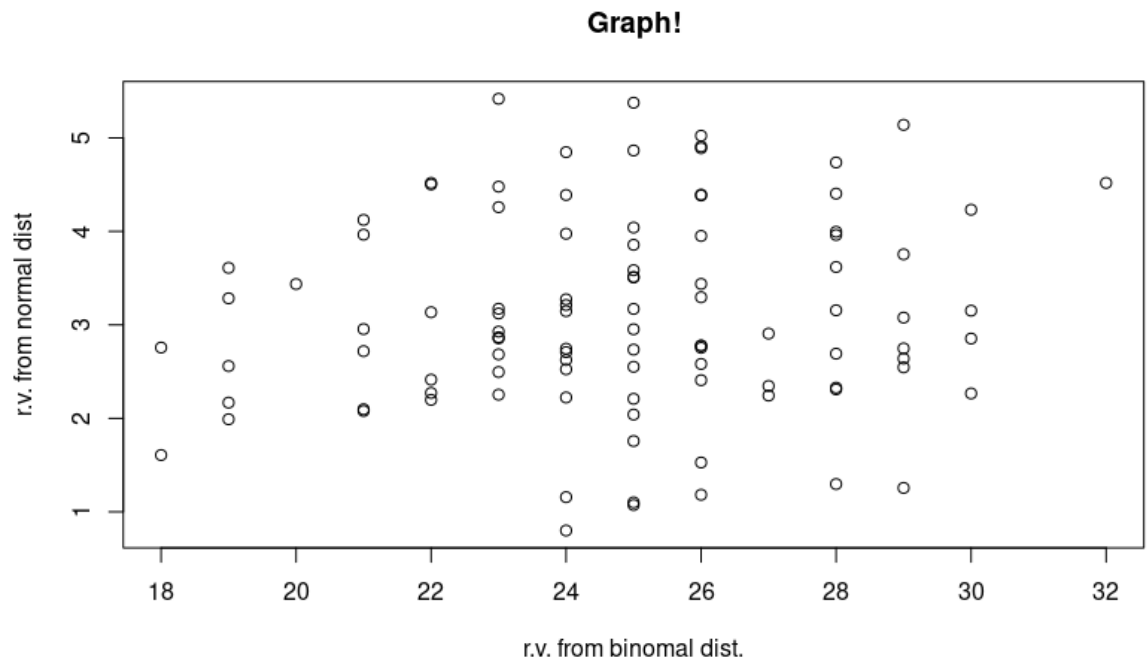


■ 正常的xy轴图


```

Console ~/
> a <- rbinom(100, 50, 0.5)
> head(a)
[1] 25 23 26 26 28 21
> a
[1] 25 23 26 26 28 21 23 22 26 25 29 24 28 25 26 24 26 27 19 27 29 22 26 28 28 30 25 28 23 19 25 26 32 21 26 24 29 25 25 27 24 24
[43] 23 24 30 21 28 24 26 29 23 26 22 19 29 23 18 28 21 22 30 28 23 23 29 22 26 24 25 25 22 24 26 21 26 20 25 26 24 18 26 23 23 25
[85] 24 30 25 25 24 24 25 28 23 25 19 25 19 29 28 21
> b <- rnorm(100, 3.0, 1.0)
> head(b)
[1] 1.072663 4.259295 4.890823 4.908722 3.619042 2.078458
> plot(a, b, xlab="r.v. from binomial dist.", ylab="r.v. from normal dist", main="Graph!")

```



- 经验累积分布函数 Empirical cumulative distribution function

- CDF(y):数据集X中小于等于y的个数占比

- 公式:

$$CDF(X,y) = \frac{|i| : i \leq y}{|X|}$$

- CDF的衍生物是PDF

- 例子:X = [2, 7, 8, 9, 10, 15, 16, 20]

- CDF(X, 15) = 6/8 = 0.75

- **Console** ~/

```

> plot(ecdf(a), verticals = T, do.points=F)
> |

```

ecdf(a)

