Impossible event $P=0$

certain event $P=1$

- contingency table
- decision tree
- Venn 图

mutually exclusive event 互斥事件：不可能同时发生

collectively exhaustive event：整个样本集至少有一个发生

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

互斥A、B $= P(A) + P(B)$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

如果A、B独立 $P(A|B) = P(A)$

$P(A \text{ and } B) = P(A)P(B)$

Bayes's theorem:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\underbrace{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)}_{=P(A)}}$$

- quantitative data → 的跳远成绩
- categorical data → 男姓

- explanatory data → 吸烟与
- Response data → 肺癌的关系

- frequency
- relative frequency
- distribution

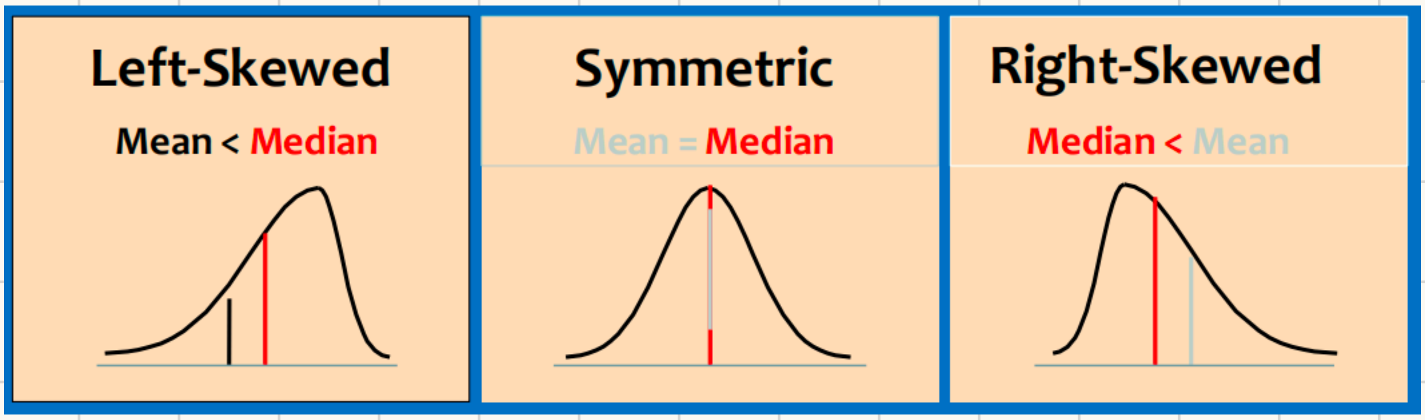
(3 5 7 8 12 13 14 18 21)

简洁描述 5 numbers

- Median 12
- Lower quartile $\frac{5+7}{2}=6$
- upper quartile $\frac{14+18}{2}=16$
- minimum 3
- maximum 21

stem and leafs plot

1	2 4 5
2	2 3 5
3	2 5



center: { mean
mode: 出现频率最高
median

spread { range ($X_{max} - X_{min}$)
IQR ($Q_3 - Q_1$)
variance
standard deviation

- **Sample Variance**

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

- **Population Variance**

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

- **Note the different between Population (N) and Sample (n-1)**

Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

$$z = \frac{X - u}{6}$$ mean 标准差

距均值的距离

±0.3是outlier

{ $z = 0$ =均值
$z > 0$ >均值
$z < 0$ <均值

emprirical rule

68% $u \pm 16$

95% $u \pm 26$

99.7% $u \pm 36$

Discrete Random Variable

$f(x) = P(X = x)$

$f(x) \geq 0 \ \forall x$ and $\sum_x f(x) = 1$

$\mu = E(X) = \sum_{i=1}^{N} x_i P(x_i)$

$\sigma^2 = \sum_{i=1}^{N} [x_i - E(X)]^2 P(x_i)$

$\sigma = \sqrt{\sigma^2}$

Binomial Distribution

二项分布 { 独立 / 只有两种

X表示在n次试验中事件发生的次数

$X \sim Binomial(n, p)$

$P(X=k) = C(n,k)\, p^k (1-p)^{n-k}$

$\mu = np$

$\sigma^2 = np(1-p)$

Poisson Distribution

泊松分布：某事件发生k次的概率（频率）

$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$

$\mu = \lambda$

$\sigma^2 = \lambda$

λ较小时 poisson右偏

λ较大时 逐渐接近正态分布 $\lambda = np$

Geometric Distribution

几何分布：第一次成功发生在第k次试验的概率

$P(X=k) = (1-p)^{k-1} p$

$\mu = \frac{1}{p}$

$\sigma^2 = \frac{1-p}{p^2}$

无记忆性

Continuous Random Variables

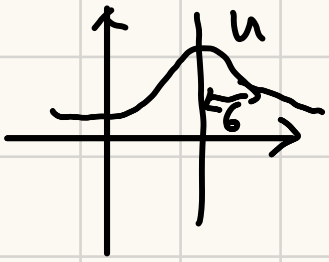
pdf: f(x)

$\int e^x = e^x$

$\int \sin = -\cos x$

$\int \cos x = \sin x$

The Normal Distribution



$$f(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \qquad X\sim N(\mu,\sigma^2)$$

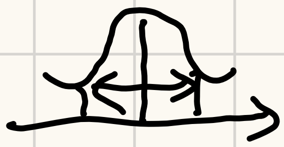
标准化 Standardized Normal

$Z \sim N(0,1)$

$Z = \frac{X-\mu}{\sigma}$

比均值高n个标准差

use standardized normal table $Z \le$ (左侧)



$$\left\{\begin{array}{ll} \mu \pm 1\sigma & 68.26\% \\ \mu \pm 2\sigma & 95.4\% \\ \mu \pm 3\sigma & 99.7\% \end{array}\right.$$

$X = \mu + Z \cdot \sigma$

Interpolation 插值法

$$\frac{Z-Z_1}{Z_2-Z_1} = \frac{P-P_1}{P_2-P_1}$$

检验是否为 Normal Distribution

① 图表

② emprirical rule $\mu \pm 1\sigma$ $\mu \pm 2\sigma$

③ $Q_3 - Q_1 \approx 1.33\sigma$

Exponential Distribution

:应用于事件发生的时间间隔

$$f(x) = \lambda e^{-\lambda x} \quad (x \geq 0)$$

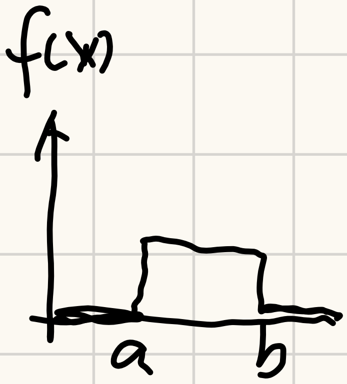
$$F(x) = 1 - e^{-\lambda x}$$

$$u = \frac{1}{\lambda}$$

$$\sigma^2 = \frac{1}{\lambda^2}$$

Uniform Distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$



$$F(x) = \begin{cases} 0 & x < a \\ \int_a^x \frac{1}{b-a}\,dt & a \leq x \leq b \\ 1 & x > b \end{cases}$$