De Morgan: $(A \cup B)^c = A^c \cap B^c$

Cartesian product: $A \times B = \{(a,b) : a \in A \wedge b \in B\}$

Inclusion-Exclusion-Principle:
$|A \cup B| = |A| + |B| - |A \cap B|$

Number of all combinations: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Disjoint: $A \cap B = \emptyset$; $P(A \cup B) = P(A) + P(B)$

total Probability:
$P(\Omega) = 1$; $\Omega = \{w_1, w_2, ...\}$ $\forall A \in \mathcal{F} : P(A) \geq 0$
$\forall i \leq n : p_i = P(w_i),\ 0 \leq p_i \leq 1$
$\sum_i p_i = 1$; $P(A) = \sum_{w_i \in A} p_i$

conditional probability:
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

independence: A, B independent:
$\Leftrightarrow P(A|B) = P(A)$
$\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$
$\Leftrightarrow P(A \cap B^c) = P(A) \cdot P(B^c)$
$\Leftrightarrow P(A^c \cap B^c) = P(A^c) \cdot P(B^c)$

Law of total probability:
$$P(A) = P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)$$
$$P(A) = \sum_{i=1}^{\infty} P(A \cap C_i) = \sum_{i=1}^{\infty} P(A|C_i) \cdot P(C_i)$$

Bayes theorem:
$$P(C_i|A) = \frac{P(A|C_i) \cdot P(C_i)}{\sum_{j=1}^{\infty} P(A|C_j) \cdot P(C_j)}$$

Rules of probability:
$P(A^c) = 1 - P(A)$
$A \subseteq B \Rightarrow P(A) \leq P(B)$

probability mass function:
$p(a) = P(X = a); \forall a : 0 \leq p(a) \leq 1$
Discrete

probability density function:
$P(c \leq X \leq d) = \int_c^d f(x)\,dx$
$f(x) \geq 0$; $\int_{-\infty}^{+\infty} f(x)\,dx = 1$
Continuous

Cumulative distribution function
$F_X(x) = P(X \leq x)$

cdf-properties:
$0 \leq F(x) \leq 1$
$x \leq y \Rightarrow F(x) \leq F(y)$
$\lim_{x \to -\infty} F(x) = 0$
$\lim_{x \to +\infty} F(x) = 1$
$\lim_{h \to 0+} F(x+h) = F(x)$

cdf- important properties:
$P(X > x) = 1 - P(X \leq x) = 1 - F(x)$
$P(a \leq X \leq b) = F(b) - F(a)$

cdf - continuous random var.
$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$
$F'(x) = f(x)$
$P(a \leq X \leq b) = \int_a^b f(x)\,dx$

Generalized inverse and quantile f
$F^{-1}(p) := \inf\{x \mid F(x) \geq p\}$ $p \in (0,1)$
$x_p = F^{-1}(p)$
$p \in (0,1) \Leftrightarrow F(x_p) = p$

Expected value discrete:
$E(X) = \sum_{i=1}^{\infty} x_i \cdot p(x_i)$

Expected value continuous:
$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx$

Properties of E(X):
$E(aX + b) = a E(X) + b$
$E(aX + bY) = a E(X) + b E(Y)$
discrete: $E(h(X)) = \sum_{i=1}^{\infty} h(x_i) \cdot p(x_i)$
continuous: $E(h(X)) = \int_{-\infty}^{+\infty} h(x) \cdot f(x)\,dx$
$\mu_k = E((X - E(X))^k)$
$\mu_2 = E((X - E(X))^2) = Var(X)$
$\sigma = \sqrt{Var(X)}$
X, Y independent: $E(X \cdot Y) = E(X) \cdot E(Y)$

Properties of Var(X):
$Var(X) = E(X^2) - (E(X))^2$
$Var(aX + b) = a^2 Var(X)$
X, Y independent:
$Var(X + Y) = Var(X) + Var(Y)$

Bernoulli distribution $X \sim b(p)$
$E(X) = p$ $\quad Var(X) = p \cdot (1-p)$
$P(X = 1) = p$ $\quad P(X = 0) = 1 - p = q$

Binomial distribution $X \sim B(n,p)$
$E(X) = n \cdot p$ $\quad Var(X) = n \cdot p \cdot (1-p)$
$p(x) = P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$

Geometric distribution:
$p(x) = P(X = x) = (1-p)^x \cdot p$
$E(X) = \frac{1-p}{p}$ $\quad Var(X) = \frac{1-p}{p^2}$

Poisson distribution $X \sim Poi(\lambda)$
$p(x) = P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$; $\lambda > 0$

Rule of thumb $n \geq 50$, $p \leq \frac{1}{10}$, $np \leq 10$
$X \sim B(n,p)$ $\quad Y \sim Poi(n \cdot p)$
$P(X = x) \approx P(Y = x)$

Uniform distribution $X \sim U(a,b)$
$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b) \\ 0, & \text{else} \end{cases}$$
$E(X) = \frac{a+b}{2}$ $\quad Var(X) = \frac{(b-a)^2}{12}$
$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \geq b \end{cases}$$

Exponential distribution $X \sim exp(\lambda)$
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
$E(X) = \frac{1}{\lambda}$ $\quad Var(X) = \frac{1}{\lambda^2}$

Normal distribution $X \sim N(\mu, \sigma^2)$
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$E(X) = \mu$ $\quad Var(X) = \sigma^2$
$\phi(z) = P(Z = z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}\,dt$
$Z \sim N(0,1)$ $\quad \phi(-z) = 1 - \phi(z)$
$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$

68-95-99.7-Rule:
$P(\mu - \sigma \leq X \leq \mu + \sigma) = P(|Z| \leq 1) \approx ,68$
$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(|Z| \leq 2) \approx ,95$
$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(|Z| \leq 3) \approx ,997$

Properties of normal distribution
$X \sim N(\mu, \sigma^2)$
$\Rightarrow Y = a + bX \sim N(a + b\mu, b^2\sigma^2)$
$\Rightarrow x_p = \mu + \sigma \cdot z_p$; $z_p = \phi^{-1}(p)$
$z_p = -z_{1-p}$
X_i independent $N(\mu, \sigma^2)$
$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$

Addition rules: $S = X_1 + ... + X_k$
$X_i \sim bern(p) \Rightarrow S \sim B(k,p)$
$X_i \sim B(n_i, p) \Rightarrow S \sim B(n_1 + ... + n_k, p)$
$X_i \sim Poi(\lambda_i) \Rightarrow S \sim Poi(\lambda_1 + ... + \lambda_k)$
$X_i \sim N(\mu_i, \sigma_i^2) \Rightarrow S \sim N(\mu_1 + ... + \mu_k, \sigma_1^2 + ... + \sigma_k^2)$

Sample mean
$X_1, ..., X_n$ independent identical distr. (i.i.d)
$E(X) = \mu$ $\quad Var(X) = \sigma^2$
$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
$E(\bar{X}) = \mu$ $\quad Var(\bar{X}) = \frac{\sigma^2}{n}$

Law of large numbers:
$E(X_i) = \mu$ $\quad Var(X_i) = \sigma^2 < \infty$
$\lim_{n \to \infty} P(|\bar{X}_n - \mu| < a) = 1$

Central Limit Theorem

$X_1, X_2 \dots$ i.i.d

$E(X_i)=\mu$ $\quad Var(X_i)=\sigma^2$

$S_n = X_1+\dots+X_n$ $\quad \bar X_n = \frac{1}{n} \cdot S_n$

$S_n \approx N(n\cdot\mu, n\cdot\sigma^2)$

$\bar X_n \approx N(\mu, \frac{\sigma^2}{n})$

Normal approximation continuity correction

$P(a\le S_n\le b)\approx \phi\left(\frac{b+\frac12-E(X_i)}{\sqrt{Var(X_i)}}\right)-\phi\left(\frac{a-\frac12-E(X_i)}{\sqrt{Var(X_i)}}\right)$

reasonable if $\min\{n\cdot p, n\cdot p(1-p)\}\ge 10$

Covariance:

$Cov(X,Y)=E((X-E(X))\cdot(Y-E(Y)))$

Properties of Covariance:

$Cov(X,Y)=E(X\cdot Y)-E(X)\cdot E(Y)$

$Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$

$Cov(X,X)=Var(X)$

X,Y independent $\Rightarrow Cov(X,Y)=0$

$Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$

$Cov(aX+b,cY+d)=ac\,Cov(X,Y)$

Correlation coefficent

$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}}$

$\rho(X,Y)=Cov\left(\frac{X-E(X)}{\sqrt{Var(X)}},\frac{Y-E(Y)}{\sqrt{Var(Y)}}\right)$

$-1\le\rho(X,Y)\le1$

$\rho(X,Y)=1 \iff Y=aX+b$ with $a>0$

$\rho(X,Y)=-1\iff Y=aX+b$ with $a<0$

$E(X\cdot Y)=\sum_{i,j} x_i\cdot y_j\cdot P(X=i, Y=j)$

Mean and empirical standard deviation

Data: $x_1,x_2,\dots x_n$

$\bar x=\frac{1}{n}\sum_{i=1}^n x_i$

$s^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar x)^2$

five statistics:

Minimum; Maximum

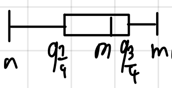
Median $(=m)$, at least 0.5 below / above

1st quartile, at least 25% $\le q_{\frac14}\le$ 75%

3rd quartile, at least 75% $\le q_{\frac34}\le$ 25%

Interquartile range: $q_{\frac34}-q_{\frac14}$

Boxplot:



Standardization: $\bar X=\frac{1}{n}\sum_{i=1}^n X_i$

$\bar X\sim N(\mu,\frac{\sigma^2}{n})$ $\quad \frac{\bar X-\mu}{\sigma/\sqrt n}\sim N(0,1)$

Hypothesis Test z-Test:

$H_0:\mu=\mu_0$ $\quad X_1,\dots,X_n$ i.i.d $\sim N(\mu,\sigma^2)$

$z=\frac{\bar x-\mu_0}{\sigma/\sqrt n}$ $\quad x_1,\dots,x_n$ realizations

$Z=\frac{\bar X-\mu_0}{\sigma/\sqrt n}\sim N(0,1)$

$p=P_{H_0}(|Z|\ge|z|)$ event as or more extreme

$p\le\alpha \iff z\in R\iff$ reject H_0

$p>\alpha\iff z\notin R\iff\neg$ reject H_0

Two-sided and one-sided testing

Two sided: $H_0:\mu=\mu_0$; $H_A:\mu\ne\mu_0$

$R=(-\infty,q_{\frac\alpha2}]\cup[q_{1-\frac\alpha2},\infty)$

$p=P_{H_0}(|Z|\ge|z|)$

Left sided: $H_0:\mu\ge\mu_0$; $H_A:\mu<\mu_0$

$R=(-\infty,q_\alpha]$

$p=P_{H_0}(Z\le z)$

right sided: $H_0:\mu\le\mu_0$; $H_A:\mu>\mu_0$

$R=[q_{1-\alpha},\infty)$

$p=P_{H_0}(Z\ge z)$

Null hypothesis	reject	$\neg$reject
true	T1 α significance	$1-\alpha$
false	power $1-\beta$	T2

$\alpha_1<\alpha_2\iff R_1\subseteq R_2\iff q_{1-\frac{\alpha_1}{2}}>q_{1-\frac{\alpha_2}{2}}$

t-Test:

estimate:

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

$T=\frac{\bar X-\mu_0}{S/\sqrt n}=\frac{\bar X-\mu_0}{SEM}\sim t(n-1)$

$t=\frac{\bar x-\mu_0}{s/\sqrt n}=\frac{\bar x-\mu_0}{SEM}$

$p=P_{H_0}(|T|\ge|t|)$

$SEM=\frac{s}{\sqrt n}$

Confidence interval (C.I.)

$H_0:\mu=\mu_0$

$I=(\bar X-q_{1-\frac\alpha2}\cdot SEM, \bar X+q_{1-\frac\alpha2}\cdot SEM)$

$P(\mu_0\in I)=1-\alpha$

$t\notin R\iff\mu_0\in I$

use when data not bell-shape

Two sample t-test: $H_0:\mu_1=\mu_2$

$t=\frac{\bar y-\bar x}{SEM_{x,y}}$ $\quad SEM_{x,y}=\sqrt{sem_y^2+sem_x^2}$

$T=\frac{\bar Y-\bar X}{\sqrt{SEM_y^2+SEM_x^2}}\sim t(\nu)$

$H_0: d=d_0$; $d_0:=\mu_2-\mu_1$

$T=\frac{(\bar Y-\bar X)-d_0}{\sqrt{SEM_y^2+SEM_x^2}}\sim t(\nu)$

$I=[(\bar Y-\bar X)-q_{1-\frac\alpha2}\cdot SEM_{x,y}, (\bar Y-\bar X)+q_{1-\frac\alpha2}\cdot SEM_{x,y}]$

Proportions relative frequency:

$Y_i=\begin{cases}1, & \text{1st category}\\0, & \text{else}\end{cases}$ $\quad H_0:p=p_0$

$h=\frac{1}{n}\sum_{i=1}^n y_i$; $H=\frac{1}{n}\sum_{i=1}^n Y_i$; $Y_i\sim ber(p)$

$SE_H:=\sqrt{\frac{H\cdot(1-H)}{n}}$; $\quad n=\frac{q_{1-\alpha}^2\cdot p\cdot(1-p)}{SEM^2}$

$Z=\frac{H-p_0}{SE_H}\sim N(0,1)$

$Z=\frac{(H_2-H_1)-0}{\sqrt{SE_{H_2}^2+SE_{H_1}^2}}\sim N(0,1)$

χ^2-statistic for goodness of fit

$\chi^2=\sum_{k=1}^d\frac{(X_k-E(X_k))^2}{E(X_k)}\sim\chi^2(d-1)$

$\chi^2\ge0$; $E(\chi^2)=d$, $Var(\chi^2)=2d$

$R=[q_{1-\alpha},\infty)$

χ^2-statistic for independence

$\chi^2=\sum_{j,k}\frac{(X_{j,k}-n\cdot p_{j,\cdot}\cdot p_{\cdot,k})^2}{n\cdot p_{j,\cdot}\cdot p_{\cdot,k}}\approx\sum_{j,k}\frac{(X_{j,k}-\frac{X_{j,\cdot}\cdot X_{\cdot,k}}{n})^2}{\frac{X_{j,\cdot}\cdot X_{\cdot,k}}{n}}$

$\chi^2=\sum_{j,k}\frac{(X_{j,k}-\frac{X_{j,\cdot}\cdot X_{\cdot,k}}{n})^2}{\frac{X_{j,\cdot}\cdot X_{\cdot,k}}{n}}\sim\chi^2((d_1-1)(d_2-1))$

$X_{z,s}\dots$ realization in row z; columns