# Probalities and Statistics

Pierre Colson

## Contents

**Markdown** version on *github*

## General Stuff

- Given $n$ distinct objects, the number of different **permutation** (without repetition) of length $r \leq n$ is :

$$n(n-1)(n-2)...(n-r+1) = \frac{n!}{(n-r)!}$$

- **Binomial coefficient** (The number of ways of distributing a set of $r$ objects from a set of $n$ distinct objects without repetition):

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- **Geometric Series** :

$$\sum_{i=0}^{n} a\theta^i = \begin{cases} a\frac{1-\theta^{n+1}}{1-\theta}, & \theta \neq 1 \\ a(n+1), & \theta = 1 \end{cases}$$

## Probability

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
- **Bayes' Theorem** : $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A \mid B_i)P(B_i)$
- $P(A_1 \cap A_2 \cap A_3) = P(A_3 \mid A_1 \cap A_2)P(A_2 \mid A_1)P(A_1)$
- $A$ and $B$ are **independant** iff $P(A \cap B) = P(A)P(B)$

## Random Variable

- A random variable that takes onmy the values 0 and 1 is called an **indicator variable** or a **Bernouilli random variable**, or a **bernouilli trial**.

- **Probability mass function** (PMF) is $f_X(x) = P(X = x) = P(A_x)$
- **Binomial** random variable is used when we are considering the number of 'sucesses' of a trial which is independently repeated a fixed number of times, and where each trial has the same probability of succecess. $X$ has PMF :

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n, \quad n \in \mathbb{N}, \quad 0 \leq p \leq 1$$

  - $E(X \mid N = n) = np$
  - $var(X \mid N = n) = np(1-p)$

- **Geometric** random variable models the waiting time until a first event, in a series of independant trials having the same sucess porbability. $X$ has PMF :

$$f_X(x) = p(1-p)^{x-1}, \quad x = 1, 2, ..., \quad 0 \leq p \leq 1$$

  We write $X \sim Geom(p)$ $E(X) = \frac{1}{p}$ and $var(X) = \frac{1-p}{p^2}$

- **Negative Binomial** random variable models the wainting time until the $n$th sucess in a series of independant trials having the same sucess probability. $X$ has PMF :

$$f_X(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}, \quad x = n, n+1, n+2, ..., \quad 0 \leq p \leq 1$$

- **Hypergeometric** random variable, independant draw, without replacement from a finite population of size $N$ We draw a sample of $m$ balls without replacement from an urn containing $w$ white balls and $b$ black balls. Let $X$ be the number of white balls drawn :

$$f_X(x) = \frac{\binom{w}{x}\binom{b}{m-x}}{\binom{w+b}{m}}, \quad x = max(0, m-b), ..., min(w, m)$$

- **Poisson** random variable appears often as a model for counts, or for number of rare events. It also provides approximation to probabilities, for example for random permutations or the binomial distribution.

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, ..., \quad \lambda > 0$$

  Is $X$ based on independant trials (0/1) with a same propability $p$, or on draws from a finite population, with replacement

  - if **Yes**, is the total number of trials $n$ fixed, so $X \in \{1, ..., n\}$ ?
    - if **Yes**: use the *binomial* distribution, $X \sim B(n, p)$ (and use the *bernouilli* distribution is $n = 1$).
      ▷ if $n \approx \infty$ or $n \gg np$, we can use the *poisson* distribution $X \sim P(np)$
    - if **No**, then $X \in \{n, n+1, ...\}$, and we use the *geometric* (if $X$ is the number of trials until one sucesses) or *negative binomial* (if $X$ is the number of trials until the last of several successes) ditribution.
  - if **No**, then if the draw is independant but without replacement from a finite population, then $X \sim$ *hypergeometric* distribution.

- **Cumulative distribution function** (CDF) of a random variable is : $F_X(x) = P(X \leq x), \quad x \in \mathbb{R}$ if $X$ is discrete we can write

$$F_X(x) = \sum_{\{x_i \in D_X : x_i \leq x\}} P(P = x_i)$$

- The **Expectation** (or **expected value** or **mean**) of $X$ is

$$E\{g(X)\} = \sum_{x \in D_X} g(x)P(X = x) = \sum_{x \in D_X} g(x)f_X(x)$$

1) $E(.)$ is linear operator, $E(aX + b) = aE(X) + b$
2) $E\{g(X) + h(X)\} = E\{g(X)\} + E\{h(X)\}$
3) if $P(X = b) = 1$, then $E(X) = b$
4) if $P(a < X \leq b) = 1$, then $a < E(X) \leq b$
5) $\{E(X)\}^2 \leq E(x^2)$
6) $E(X) = \sum_{i=1}^{\infty} E(X \mid B_i)P(B_i)$
7) $E(X) = E_Y\{E(X \mid Y = y)\}$
8) $E\{g(X,Y)\} = E_X[E\{g(X,Y) \mid X = x\}]$

- The **Variance** of $X$ is $var(X) = E[\{X - E(X)\}^2]$

1) $var(X) = E(X^2) - E(X)^2 = E\{X(X-1)\} + E(X) - E(X)^2$
2) $var(aX + b) = a^2 var(X)$
3) $var(X) = 0 \implies X$ is constant with probability 1.
4) $var\{g(X,Y)\} = E_X[var\{g(X,Y) \mid X = x\}] + var_X[E\{g(X,Y) \mid X = x\}]$

- The **Standard deviation** of $X$ is defined as $\sqrt{var(X)}$
- **Law of small numbers** Let $X_n \sim B(n, p_n)$, and suppose that $np_n \to \lambda >$ à when $n \to \infty$. Then $X_n \xrightarrow{D} X$, where $X \sim P(\lambda)$

## Continuous Random Variable

- A random variable $X$ is **continuous** if ther exist a function $f(x)$, called the **probability density function** (PDF) of $X$, such that :

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du, \quad x \in \mathbb{R}$$

The porperties of $F$ imply that $(i) f(x) \geq 0$ and $(ii) \int_{-\infty}^{\infty} f(x)dx = 1$

- **Uniform distribution**. The random variable $U$ having density :

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

and :

$$F(u) = \begin{cases} \frac{u}{b-a}, & a \leq u \leq b \\ 0, & otherwise \end{cases}$$

is called a **uniform random variable**. We write $U \sim U(a, b)$.

- **Exponential random variable**. The random variable $X$ having density :

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & otherwise \end{cases}$$

and :

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, \quad x > 0 \\ 0, \quad otherwise \end{cases}$$

is called and **exponential random variable** wiht parameter $\lambda > 0$.

  - We write $X \sim exp(\lambda)$
  - We $E(x) = \frac{1}{\lambda}$

- **Gamma distribution**. The random variable $X$ having density !

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad x > 0 \\ 0, \quad otherwise \end{cases}$$

  is called a **gamma random variable** with parameters $\alpha, \lambda > 0$. We write $X \sim Gamma(\alpha, \lambda)$

  - $E(X) = \frac{\alpha}{\beta}$
  - $var(X) = \frac{\alpha}{\beta^2}$

- **Laplace**. The random variable $X$ having density :

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x-\eta|}, \quad x \in \mathbb{R}, \quad \eta \in \mathbb{R}, \lambda > 0$$

  is called a **Laplace random variable**. $\eta$ is the *median* of teh dostribution.

- **Pareto**. The random variable $X$ witch cumulative distribution function :

$$F(x) = \begin{cases} O, \quad x < \beta \\ 1 - (\frac{\beta}{x})^\alpha, \quad x \geq \beta, \end{cases}$$

  is called a **Pareto random variable**

- We define the **expectation** of $g(X)$ to be :

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f(x) dx$$

- The **variance** of X is :

$$var(X) = \int_{-\infty}^{\infty} \{x - E(X)\}^2 f(x) dx = E(X^2) - E(X)^2$$

| . | Discrete | Continuous |
|---|---|---|
| Support $D_X$ | countable | contains an interval $(x_-, x_+) \subset \mathbb{R}$ |
| $f_X$ | mass function dimensionless | density function units $[x]^{-1}$ |
| | $O \leq f_X(x) \leq A \sum_{x \in \mathbb{R}} f_X(x) = 1$ | $0 \leq f_X(x) \int_{-\infty}^{\infty} f_X(x) dx = 1$ |
| $F_X(a) = P(X \leq a)$ | $\sum_{x \leq a} f_X(x)$ | $\int_{-\infty}^{a} f_X(x) dx$ |
| $P(X \in A)$ | $\sum_{x \in A} f_X(x)$ | $\int_A f_X(x) dx$ |
| $P(a < X \leq b)$ | $\sum_{\{x : a < x \leq b\}} f_X(x)$ | $\int_a^b f_X(x) dx$ |
| $P(X = a)$ | $f_X(a) \geq 0$ | $\int_a^b f_X(x) dx = 0$ |
| $E\{g(X)\}$ (if well defined) | $\sum_{x \in \mathbb{R}} g(x) f_X(x)$ | $\int_{-\infty}^{\infty} g(x) f_X(x) dx$ |

- Let $0 < p < 1$. We define the $p$ **quantile** of the cumulative distribution function $F(x)$ to be :

  $$x_p = inf\{x : F(x) \geq p\}$$

  For most continuous random variable, $x_p$ is unique and equals $x_p = F^{-1}(p)$, where $F^{-1}$ is the inverse function $F$. Then $x_p$ is the value for which $P(X \leq x_p) = p$.

  The $p$ quantile of $Y$ is the value $y_p$ that solves $F_Y(y_p) = p$.

- A random variable $X$ having density :

  $$f(x) = \frac{1}{(2\pi)^{1/2}\sigma} exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0$$

  is a **normal random variable** with the expectation $\mu$ and variance $\sigma^2$: we write $X \sim N(\mu, \sigma^2)$. When $\mu = 0, \sigma^2 = 1$ the corresponding variable $Z$ is **standard normal**, $Z \sim n(0,1)$ with density :

  $$\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}, \quad z \in \mathbb{R}$$

  Then :

  $$F_Z(x) = P(Z \leq x) = \Phi(x) = \int_{-\infty}^{\infty} \phi(z) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} e^{-z^2/z} dz$$

  The **normal distribution** is often called the **Gaussian distribution**.

  $Z \sim N(0,1)$ :

  1) Then density os symmetric with respect to $z = 0$ i.e. $\phi(z) = \phi(-z)$
  2) $P(Z \leq z) = \phi(z) = 1 - \phi(-z) = 1 - P(Z \geq z)$
  3) We have :

     $$\phi'(z) = -z\phi(z), \quad \phi''(z) = (z^2-1)\phi(z), \quad \phi'''(z) = -(z^3 - 3z)\phi(z), ...$$

     This implies thath $E(Z) = 0, var(Z) = 1, E(Z^3) = 0$ etc.
  4) If $X \sim N(\mu, \sigma^2)$ then $Z = (X - \mu)/\sigma \sim N(0,1)$

  MGF of Normal law : $e^{\mu t + \sigma^2 t^2/2}$

- **Moivre-Laplace** Let $X_n \sim B(n,p)$, let :

  $$\mu_n = E(X_n) = np, \quad \sigma_n^2 = var(X_n) = np(1-p)$$

  When $n \to \infty$ we can approximate : $X_n \sim N\{np, np(1-p\}$

  This gives us :

  $$P(X_n \leq r) = P(\frac{X_n - \mu_n}{\sigma_n} \leq \frac{r - \mu_n}{\sigma_n}) = \Phi(\frac{r - \mu_n}{\sigma_n})$$

  A better approcximation of $P(X_n \leq r)$ is given by replacing $r$ by $r + \frac{1}{2}$. The $\frac{1}{2}$ is called the *conitnuity correction*

Which density ?

- **Uniform** variables lie in a finite interval, and give equal probability to each part of the interval
- **Exponential** and **Gamma** variables lie in $(0, \infty)$, adn are ofter used to model waiting times and other positive quantities,
  - The gamma has two parameters and is more flexible, but the exponential is simpler and has some elegant properties
- **Pareto** variables lie in the interval $(\beta, \infty)$, so are not appropriate for arbitrary positive quantities (wich could be smaller than $\beta$) but are oftenused to model finincial losses over some treshhold $\beta$
- **Normal** variables lie in $\mathbb{R}$ and are used to model quantities that arise (or might arise) through averaging of many small effects (e.g, height and weight, which are influenced by many genetic factors), or where measurements are subject to error.
- **Laplace** variables lie in $\mathbb{R}$. the Laplace distribution can be used in place of the normal in situations where outliers might be present.

## Several random variable

- Let $(X, Y)$ be an discrete random variable: the set :

$$D = \{(x, y) \in \mathbb{R}^2 : P\{(X, Y) = (x, z)\} > 0\}$$

is countable. The **joint probability mass function** of $(X, Y)$ is

$$f_{X,Y}(x, y) = P\{(X, Y) = (x, y)\}, \quad (x, y) \in \mathbb{R}^2$$

and the **joint cumulative distribution function** of $(X, Y)$ is

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y), \quad (x, y) \in \mathbb{R}^2$$

if then random variable is *continuous* then the **joint cumulative function** of $(X, Y)$ can be written :

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{x} f_{X,Y}(u, v) du dv \quad (x, y) \in \mathbb{R}^2$$

this implies :

$$f_{X,Y}(x, y) = \frac{\delta^2}{\delta x \delta y} F_{X,Y}(x, y)$$

- The **Marginal probability mass/density function** of $X$ is :

$$f_X(x) = \begin{cases} \sum_y f_{X,Y}(x, y) & \text{discrete case} \\ \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, & \text{continuous case} \end{cases}$$

- Let $X$, $Y$ be random variables of density $f_{X,Y}(x, y)$. Then if $E\{|g(X, Y)|\} < \infty$, we can define the **expectation** of $g(X, Y)$ to be :

$$E\{g(X, Y)\} = \begin{cases} \sum_{x,y} g(x, y) f_{X,Y}(x, y), & \text{discrete case} \\ \int \int g(x, y) f_{X,Y}(x, y) dx dy, & \text{continuous case.} \end{cases}$$

- We define the **covariance** of $X$ and $Y$ as :

$$cov(X,Y) = E(XY) - E(X)E(Y)$$

  - $cov(X,X) = var(X)$
  - $cov(a,X) = 0$
  - $cov(X,Y) = cov(Y,X)$ (*symetry*)
  - $cov(a+bX+cY,Z) = bcov(X,Z) + ccov(Y,Z)$ (*bilinearity*)
  - $cov(a+bX, c+dY) = bdcov(X,Y)$
  - $var(a+bX+cY) = b^2var(X) + 2bccov(X,Y) + c^2var(Y)$
  - $cov(X,Y)^2 \leq var(X)var(Y)$ (*Cauchy-Schwarz inequality*)
- $X$, $Y$ independant $\implies$ $cov(X,Y) = 0$. Howerver, the converse is false
- The **correlation** of $X$ and $Y$ is :

$$corr(X,Y) = \frac{cov(X,Y)}{\{var(X)var(Y)\}^{1/2}}$$

  Let $\rho = corr(X,Y)$. Then :

  - $-1 \leq \rho \leq 1$
  - If $\rho = \pm 1$n then there exist $a,b,c \in \mathbb{R}$ such that : $aX + bY + c = 0$. With probability 1 ($X$ and $Y$ are then linearly indepenant)
  - If $X$ and $Y$ are independent then $corr(X,Y) = 0$
  - The effect of the transformation $(X,Y) \mapsto (a+bX, c+dY)$ is $corr(X,Y) \mapsto sign(bd)corr(X,Y)$
- We define the **moment-genrating function** (MGF) of a random variable $X$ by :

$$M_X(t) = E(e^{tX})$$

  - $M_X(t)$ is also called the **Laplace tranform** of $f_X(x)$
  - The MGF is useful as s summary of all properties of $X$, we can write :

$$M_X(t) = E(e^{tX}) = E(\sum_{r=0}^{\infty} \frac{t^r X^r}{r!}) = \sum_{r=0}^{\infty} \frac{t^r}{r!} E(X^r)$$

  - $M_X(0) = 1$
  - $M_{a+bX+cY}(t) = e^{at} M_X(bt) M_Y(ct)$
  - $E(X^r) = \frac{\partial M_X(t)}{\partial t^r}|_{t=0}$
  - $E(X) = M'_X(0)$
  - $var(X) = M''_X(0) - M'_X(0)^2$
  - We say that the random variables $\{X_n\}$ **converge in distribution** to $X$, if, for all $x \in \mathbb{R}$ where $F$ is continuous :

$$F_n(x) \to F(x), \quad n \to \infty$$

  - The cumulative case :

$$M_X(t) = E(e^{t^T X}) = E(e^{\sum_{r=1}^{p} t_r X_r}), \quad t \in \mathcal{T}$$

- The **cumulant-generating function** (CGF) of $X$ is $K_X(t) = \log M_X(t)$. The cumulants $\kappa_r$ of $X$ ar definde by :

$$K_X(t) = \sum_{r=1}^{\infty} \frac{t^r}{r!}\kappa_r, \quad \kappa_r = \frac{d^r K_X(t)}{dt^r}\Big|_{t=0}$$

  It is easy to verify that $E(X) = \kappa_1$ and $var(X) = \kappa_2$ The CGF is equivalent to MGF, and so shares its properties, but it is often easier to work with th CGF.

- The random vector $X = (X_1, \ldots, X_p)^T$ has a **multivariate normal distribution** if there exist a $p \times 1$ vector $\mu = (\mu_1, \ldots, \mu_p)^T \in \mathbb{R}^p$ and a $p \times p$ symmetric matrix $\Omega$ with elements $w_{ij}$ such that :

$$u^T X \sim \mathcal{N}(u^T\mu, u^T\Omega u), \quad u \in \mathbb{R}^p$$

  then we write $X \sim \mathcal{N}(\mu, \Omega)$.

  - We have :

$$E(X_j) = \mu_j, \quad var(X_j) = wjj, \quad cov(X_j, X_k) = w_{jk}, \quad j \neq k$$

    so $\mu$ and $\Omega$ are called the *mean vector* and *covariance matrix* of $X$.

  - The moment-genrating function of $X$ is $M_X(u) = exp(u^T\mu, +\frac{1}{2}u^T\Omega u)$, for $u \in \mathbb{R}^p$.
  - If $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, p\}$, and $\mathcal{A} \cap \mathcal{B} = \emptyset$ then

$$X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \Leftrightarrow \Omega_{\mathcal{A},\mathcal{B}} = 0$$

  - If $X_1, \ldots, X_n \overset{idd}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $X_{n\times 1} = (X_1, \ldots, X_n)^T \sim \mathcal{N}(\mu 1_n, \sigma^2 I_n)$
  - Linear combination of normal variables are normal :

$$a_{r\times 1} + B_{r\times p}X \sim \mathcal{N}_r(a + B\mu, B\Omega B^T)$$

    .

  - The random vector $X \sim \mathcal{N}(\mu, \Omega)$, has density function on $\mathbb{R}^p$ if on only if $\Omega$ is positive definite, i.e., $\Omega$ has rank $p$. If so, then density function is :

$$f(x; \mu, \Omega) = \frac{1}{(2\pi)^{p/2}|\Omega|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T\Omega^{-1}(x-\mu)\}, \quad x \in \mathbb{R}^p$$

- **Maarginal and conditional distributions**. Let $X \sim \mathcal{N}(\mu_{p\times 1, \Omega_{p\times p}})$, where $|\Omega| > 0$, and let $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, p\}$ with $|\mathcal{A}| = q < p, |\mathcal{B}| = r < p$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$ Let $\mu_{\mathcal{A}}, \Omega_{\mathcal{A}}$ and $\Omega_{\mathcal{AB}}$ be respectively the $q \times 1$ subvector of $\mu, q \times q$ and $q \times r$ submatrices of $\Omega$ conformable with $\mathcal{A}, \mathcal{A} \times \mathcal{A}$ and $\mathcal{A} \times \lfloor$. Then :

  - The marginal distribtuion of $X_{\mathcal{A}}$ is normal,

$$X_{\mathcal{A}} \sim \mathcal{N}(\mu_{\mathcal{A},\Omega_{\mathcal{A}}})$$

  - The conditional distribution of $X_{\mathcal{A}}$ given $X_{\mathcal{B}} = x_{\mathcal{B}}$ is normal,

$$X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}} \sim \mathcal{N}\{\mu_{\mathcal{A}} + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}), \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}}\}$$

    This has two important implications:

    - ∗ Implies that any subvector of $X$ also has a multivariate normal distribution.
    - ∗ Implies that two components of $X_{\mathcal{A}}$ are conditionally independent gievn $X_{\mathcal{B}}$ id and only if the corresponding off-dialog element of $\Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}}$ equals zero.

**Reminder** : Transformation pf random variables

We often want to calculate the distributions of random variables based on other random variables
* Let $Y = g(X)$, where the function $g$ is known. We want to obtain $F_Y$ and $f_Y$ from $F_X$ and $f_X$.
* Let $g : \mathbb{R} \mapsto \mathbb{R}, \mathcal{B} \subset \mathbb{R}$, and $g^{-1}(\mathcal{B}) \subset \mathbb{R}$ be the set for which $g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$. Then

$$P(Y \in \mathcal{B}) = P\{g(X) \in \mathcal{B}\} = P\{X \in g^{-1}(\mathcal{B})\}.$$

Since $X \in g^{-1}(\mathcal{B})$ iff $g(X) = Y \in g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$. * To find $F_Y(y)$, we take $\mathcal{B} = (-\infty, y]$, giving

$$F_Y(y) = P(Y \leq y) = P\{g(X) \in \mathcal{B}_y\} = P\{X \in g^{-1}(\mathcal{B}_y)\}$$

* If the function $g$ is monotonic increasing with (monotonic increasing) inverse $g^{-1}$, then

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d_X\{g^{-1}(y)\}}{dy} = f_X\{g^{-1}(y)\} \times \left|\frac{dg^{-1}(y)}{dy}\right|$$

where the $|.|$ ensures that the same formula holds with monotonic decreasing $g$.

- **Sum of independent variables** If $X$, $Y$ are independant random variables, then the PDF of their sum $S = X + Y$ is the convolution $f_X * f_Y$ of the PDFs $f_X$, $f_Y$:

$$f_S(s) = f_X * f_Y(s) = \begin{cases} \int_{-\infty}^{\infty} f_X(x) f_Y(s-x)dx, & X, Y \text{continuous} \\ \sum_x f_X(x) f_Y(s-x), & X, Y \text{discrete.} \end{cases}$$

- The **order statistics** of the rv's $X_1, \ldots, X_n$ are the ordered values

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

If the $X_1, \ldots, X_n$ are continuous, then no two of the $X_j$ can be equal, i.e.,

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

In particular, the *minimum* is $X_{(1)}$, then *maximum* is $X_{(n)}$, and the *median* is

$$X_{(m+1)} \quad (n = 2m+1, \text{odd}), \qquad \frac{1}{2}(X_{(m)} + X_{(m+1)}) \quad (n = 2m, \text{even})$$

The median is the central value of $X_1, \ldots, X_n$.

  - Let $X_1, \ldots, X_n \overset{idd}{\sim} F$, from a continuous distribution with density $f$, then:
    - ∗ $P(X_{(n)} \leq x) = F(x)^n$
    - ∗ $P(X_{(1)} \leq x) = 1 - \{1 - F(x)\}^n$
    - ∗ $$f_{X_{(r)}(x)} = \frac{n!}{(r-1)!(n-r)!} F(x)^{r-1} f(x) \{1 - F(x)\}^{n-r}, \quad r = 1, \ldots, n$$

## Approximation and Convergence

- Inequalities

  If $X$ is a random variable, $a > 0$ a constant, $h$ a non-negative function and $g$ a convex function, then

  - $P\{h(X) \geq a\} \leq E\{h(X)\}/a$, (basic inequality)
  - $P(|X| \geq a) \leq E(|X|)/a$, (Markov's inequality)
  - $P(|X| \geq a) \leq E(X^2)/a^2$, (Chebyshov's inequality)
  - $E\{g(X)\} \geq g\{E(X)\}$. (Jensen's inequality)

  On replacing $X$ by $X - E(X)$, Chebyshov's inequality gives :

$$P\{|X - E(X)| \geq a\} \leq var(X)/a^2$$

- **Hoeffdings's inequality** Let $Z_1, \ldots, Z_n$ be independent random variables such that $E(Z_i) = 0$ and $a_i \leq Z_i \leq b_i$ for constants $a_i \leq b_i$. If $\epsilon > 0$, then for all $t > 0$,

$$P(\sum_{i=1}^{n} Z_i \geq \epsilon) \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}$$

- **Convergence**

  Let $X_1, X_2, \ldots$ be random variable with cumulative distribution function $F, F_1, F_2, \ldots$ Then :

  - $X_n$ converges to $X$ *almost surely*, $X_n \xrightarrow{a.s} X$, if

$$P(\lim_{n\to\infty} X_n = X) = 1$$

  - $X_n$ converges to $X$ *in mean square*, $X_n \xrightarrow{2} X$, if

$$\lim_{n\to\infty} E\{(X_n - X)^2\} = 0, \quad \text{where} E(X_n^2), E(X^2) < \infty$$

  - $X_n$ converges to $X$ *in probability*, $X_n \xrightarrow{P} X$, if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

  - $X_n$ converges to $X$ *in distributioon*, $X_n \xrightarrow{D} X$, if

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \text{at each point } x \text{ where } F(X) \text{ is continuous.}$$

  - If $X_n \xrightarrow{a.s} X$, $X_n \xrightarrow{2} X$, $X_n \xrightarrow{P} X$, then $X_1, X_2, \ldots, X$ must all be defined with respect to only one probability space. This is not the case for $X_n \xrightarrow{D} X$, which only concerns the probabilities. This last is thus weaker than the other.
  - This mode of convergence are related to one anothr in the following way :

$$X_n \xrightarrow{a.s} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$$

$$X_n \xrightarrow{2} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$$

- **Continuity Theorem** Let $\{X_n\}$, $X$ be random variables with cumulative distribution functions {F_n}, F, whose MGFs $M_n(t)$, $M(t)$ exist for $0 \leq |t| < b$. If there exists $a0 < a < b$ such that $M_n(t) \leftarrow M(t)$ for $|t| \leq a$ when $n \to \infty$, then $X_n \xleftarrow{D} X$, thath is to say, $F_n(x) \leftarrow F(x)$ at each $x \in \mathbb{R}$ where $F$ is continuous.
  - We could replace $M_n(t)$ and $M(t)$ by the cumulant-generating functions $K_n(t) = \log M_n(t)$ and $K(t) = \log M(t)$
- Let $x_0, y_0$ be constants, $X, Y, \{X_n\}, \{Y_n\}$ random variables, and $h$ a function continuous at $x_0$. Then

$$X_n \xrightarrow{D} x_0 \Rightarrow X_n \xrightarrow{P} x_0$$

$$X_n \xrightarrow{P} x_0 \Rightarrow h(X_n) \xrightarrow{P} h(x_0)$$

$$X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{P} y_0 \Rightarrow X_n + Y_n \xrightarrow{D} X + y_0, X_nY_n \xrightarrow{D} Xy_0$$

  The third line is know as **Slutsky's lemma**. (very usefull)
- **Laws of Large Number** Let $X_1, X_2, \ldots$ be a sequence of independant identically distributed random variables with finite expectation $\mu$, and write their average as

$$\bar{X} = n^{-1}(X_1 + \cdots + X_n)$$

  Then $\bar{X} \xrightarrow{P} \mu$; i.e, for all $\epsilon > 0$,

$$P(|\bar{X} - \mu| > \epsilon) \to 0, \quad n \to \infty$$

- **Strong law of large number** Under the conditions of the last theorem, $\bar{X} \xrightarrow{a.s.} \mu$:

$$P(\lim_{n\to\infty} \bar{X} = \mu) = 1$$

- **Central limit Theorem** (CLT) Let $X_1, X_2, \ldots$ be independent random variables with expectation $\mu$ and variance $0 < \sigma^2 < \infty$ Then

$$Z_n = \frac{n^{1/2}(\bar{X} - \mu)}{\sigma} \xrightarrow{D} Z, \quad n \to \infty$$

  Where $Z \sim N(0,1)$

  Thus

$$P\Big\{\frac{n^{1/2}(\bar{X} - \mu)}{\sigma} \leq z\Big\} \doteq P(Z \leq z) = \phi(z)$$

  for large $n$.
- **Delta Model** Let $X_1, X_2, \ldots$ be independent radom variable with expectatin $\mu$ and variance $0 < \sigma^2 < \infty$ and let $g'(\mu) \neq 0$, where $g'$ is the derivative of $g$. Then

$$\frac{g(\bar{X}) - g(\mu)}{\{g'(\mu)^2\sigma^2/2\}^{1/2}} \xrightarrow{D} N(0,1), \quad n \to \infty$$

This implies that for large $n$, we have $g(\bar(X)) \dot{\sim} N\{g(\mu), g'(\mu)^2\sigma^2/n\}$ Combined with Slutsky's lemma, we have :

$$g(\bar{X}) \dot{\sim} N\{g(\mu), g'(\bar{X})^2 S^2/n\}, \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

- Might need to add Sample quantiles

## Statistical Inference

- The **Method of moments estimate** of a parameters $\theta$ is the value $\theta$ that matches the theorical and empirical moments.

  For a model with $p$ unknown parameters, we set the theorical moments of the population equla to the empirical moments of the sample $y_1, \dots, y_n$ and solve the resulting equations :

  $$E(Y^r) = \int y^r f(y;\theta)dy = \frac{1}{n}\sum_{j=1}^{n} y_j^r, \quad r = 1, \dots, p$$

- **Maximum likehood estimation** If $y_1, \dots, y_n$ is a random sample from the density $f(y, \theta)$, then the likehood for $\theta$ is :

  $$L(\theta) = f(y_1, \dots, y_n; \theta) = f(y_1;\theta) \times f(y_2;\theta) \times \cdots \times f(y_n;\theta)$$

  The data are treated as fixed, and the likehood $L(\theta)$ is regarded as a function of $\theta$

  The *maximum likelihood estimate* (MLE) $\widehat{\theta}$ of a parameter $\theta$ is the value that gives the observed data the highest likehood. Thus :

  $$L(\widehat{\theta}) \geq L(\theta) \text{ for each } \theta$$

  We simplify the calculations by maximising $l(\theta) = \log(L(\theta))$ rather than $L(\theta)$

  - Calculate the log-likehood $l(\theta)$
  - Find the value $\widehat{\theta}$ maximising $l(\theta)$, which often satisfies $dl(\widehat{\theta})/d\theta = 0$
  - Check that $\widehat{\theta}$ gives a maximum, often by cheking that $d^2l(\widehat{\theta})/d\theta^2 < 0$

- **Mean square error** of the estimator $\hat{\theta}$ of $\theta$ is :

  $$MSE(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\} = \cdots = var(\hat{\theta}) + b(\theta)^2$$

- **M-estimation** This generalises maximum likehood estimation. We maximise a function of the form :

  $$\rho(\theta; Y) = \sum_{j=1}^{n} \rho(\theta; Y_j)$$

  Where $\rho(\theta; y)$ if concave as a function of $\theta$ and for all $y$. Equivalently we minimise $-\rho(\theta; Y)$

- Let $Y = (Y_1, \dots, Y_n)$ be sampled from a distribution $F$ with parameter $\theta$. Then a **pivot** is a function $Q = q(Y, \theta)$ of the data and the parameters $\theta$, where the ditribution of $Q$ is known and does not depend on $\theta$. We say that $Q$ is **pivotal**.

- **Confidence intervals** Let $Y = (Y_1, \dots, Y_n)$ be the data from a parametric statistical model with scalar parameter $\theta$. A *confidence interval* (CI) $(C, L)$ for $\theta$ with lower bound $L$ and upper bound $U$ is a random interval that contains $\theta$ with specified probability, called the (confidence) level of the interval.
  - $L = l(Y)$ and $U = u(Y)$ are statistics that can be computed from the data $Y_1, \dots, Y_n$. They do not depend in $\theta$.
  - In a continuous setting (so $<$ gives the sample probabilities as $\leq$), and if we wirte the probabilities that $\theta$ lies below and above the interval as :

    $$P(\theta < L) = \alpha_L, \quad P(U < \theta) = \alpha_U$$

    then $(L, U)$ has confidence level

    $$P(L \leq \theta \leq U) = 1 - P(\theta < L) - P(U < \theta) = 1 - \alpha_L - \alpha_U$$

  - Often we seek an interval with equal probabilities of not containing $\theta$ at each end, with $\alpha_L = \alpha_U = \alpha/2$, giving an *qui-tailed* $(1-\alpha) \times 100$ *confidence interval*
  - We usually take standard values of $\alpha$, such that $1 - \alpha = 0.9, 0.95, 0.99, \dots$
- If $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

  $$Z = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$$

  is a pivot that provides an exact $(1 - \alpha_L - \alpha_U)$ confidence interval for $\mu$, of the form

  $$(L, U) = \Big(\bar{Y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha_L}, \bar{Y} - \frac{\sigma}{\sqrt{n}} z_{\alpha_U}\Big)$$

  where $z_p$ denotes the $p$ quantile of the standard normal distribution.
  - In application $\sigma^2$ is usually unknow. Then we have :

    $$\frac{\bar{Y} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}, \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

    are pivots that provide confidence intervals for $\mu$ and $\sigma^2$, respectivly,

    $$(L, U) = \Big(\bar{Y} - \frac{S}{\sqrt{n}} t_{n-1}(1 - \alpha_L), \bar{Y} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha_U)\Big)$$

    $$(L, U) = \Big(\frac{(n-1)S^2}{\chi^2_{n-1}(1 - \alpha_L)}, \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha_U)}\Big)$$

    where :
    * $t_v(p)$ is the $p$ quantile of the *Student $t$ distribution with $v$ degrees of fredom*
    * $\chi^2_v(p)$ is the $p$ quantile of the *chi-quare distribution with $v$ degrees of freedom*

- For symmetric densities such as the normal and the Student $t$, the quantiles satisfy :

  $$z_p = -z_{1-p}, \quad t_v(p) = -t_v(1-p)$$

  so equi-tailed $(1-\alpha) \times 100$ Cls have the forms :

  $$\bar{Y} \pm n^{-1/2}\sigma z_{1-\alpha/2}, \quad \bar{Y} \pm n^{-1/2} S t_{n-1}(1-\alpha/2)$$

8.4 pas fait