Statistics & Stochastic Processes # 229642 22987633 56A8756 634424

Probability (MAST20004-2020S1) 2-25

1. Axioms and Conditional Probability
2. Random Variables and Distributional Functions
3. Special Univariate Random Variables
4. Bivariate Random Variables
5. Sum of Independent Random Variables and Limit Theorems
6. ~~Basics of Stochastic Process~~

Statistics (MAST20005-2021SUM) 26-65

1. Random Samples and Plots
2. Point Estimation
3. Interval Estimation and Hypothesis Testing Under i.i.d. Model
4. ~~Simple Linear Regression Model~~
5. Order Statistics, Quantiles and Distribution-Free Methods
6. ~~ANOVA and Likelihood Ratio Test~~
7. ~~Bayesian Methods~~
8. Asymptotics and Optimality

Linear Statistical Models (MAST30025-2021S1) 66-112

1. Review of Linear Algebra
2. Random Vectors
3. The Full Rank Model
4. The Full Rank Model Relevance
5. The Less-Than-Full Rank Model
6. The Less-Than-Full Rank Model Relevance and ANCOVA
7. Experimental Design

Stochastic Modelling (MAST30001-2021S2) 113-135

1. Discrete-Time Markov Chains
2. Poisson Process
3. Continuous-Time Markov Chains
4. Queuing Systems
5. Renewal Theory
6. Gaussian Processes

Probability is a set function from $\mathcal{P}(\Omega)$, the powerset of sample space Ω, to $[0, 1]$.

The Kolmogorov Axioms of Probability:

1. $\mathbb{P}(A) \geq 0 \ \forall A$.
2. $\mathbb{P}(\Omega) = 1$.
3. Countable additivity: $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ where $\{A_1, A_2, \dots\}$ is any sequence of mutually exclusive events.
4. $\mathbb{P}(\emptyset) = 0$ since $\emptyset \cup \emptyset \cup \cdots = \emptyset$, but $\mathbb{P}(A) = 0 \not\Rightarrow A = \emptyset$.
5. Finite additivity.
6. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ since $A \cup A^c = \Omega$.
7. $A \subset B$ implies $\mathbb{P}(A) \leq \mathbb{P}(B)$ since $A \cup (A^c \cap B) = B$ and $A \cap (A^c \cap B) = \emptyset$.
8. $\mathbb{P}(A) \leq 1$ since $A \subset \Omega$.
9. Addition theorem: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
10. Continuity with either of
    a. $A_1 \subset A_2 \subset \cdots$ and $B = \bigcup_{i=1}^{\infty} A_i$
    b. $A_1 \supset A_2 \supset \cdots$ and $B = \bigcap_{i=1}^{\infty} A_i$

    occurs, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(B)$.

We study a random experiment in the context of a Probability Space $\langle \Omega, \mathcal{F}, \mathbb{P} \rangle$.

1. Ω is the sample space: the set of all possible outcomes of our random experiment.
2. $\mathcal{F}$ is a sigma-field: a collection (with certain properties) of subsets of Ω. We view these as events we can see or measure.
3. $\mathbb{P}$ is a probability measure: a function (with certain properties) defined on the elements of $\mathcal{F}$ with certain properties.

Conditional probability: $\mathbb{P}(A|H) = \mathbb{P}(A \cap H)/\mathbb{P}(H)$ if $\mathbb{P}(H) > 0$.

1. $\mathbb{P}(A|B) > \mathbb{P}(A)$ or $\mathbb{P}(B|A) > \mathbb{P}(B)$, then A, B are positively related.
2. $\mathbb{P}(A|B) < \mathbb{P}(A)$ or $\mathbb{P}(B|A) < \mathbb{P}(B)$, then A, B are negatively related.
3. $\mathbb{P}(A|B) = \mathbb{P}(A)$ or $\mathbb{P}(B|A) = \mathbb{P}(B)$ or $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then A, B are independent (and also their complements).

The Law of Total Probability: If A_i's are disjoint and exhaustive then, for any event H,

$$\mathbb{P}(H) = \sum_i \mathbb{P}(H \cap A_i) = \sum_i \mathbb{P}(H|A_i)\,\mathbb{P}(A_i).$$

Bayes' Formula: If A_i's are disjoint and exhaustive then, for any event H,

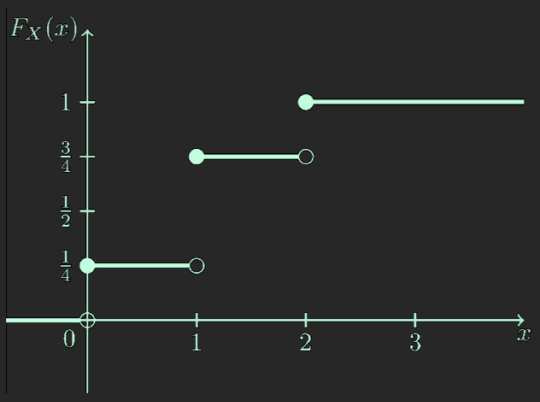
$$\mathbb{P}(A_i|H) = \frac{\mathbb{P}(A_i \cap H)}{\mathbb{P}(H)} = \frac{\mathbb{P}(H|A_i)\mathbb{P}(A_i)}{\mathbb{P}(H)} = \frac{\mathbb{P}(H|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(H|A_j)\mathbb{P}(A_j)}.$$

A function X which assigns every outcome $\omega \in \Omega$ to $X(\omega) \in \mathbb{R}$ is called a random variable. We shall denote the set of possible values (called state space) of X by $S_X \subset \mathbb{R}$.

Probability mass function: $S_X \to [0,1]$.

1. $\sum_{x_i \le x} p_X(x_i) = F_X(x)$.
2. $\sum_{x \in S_X} p_X(x) = 1$.

Cumulative distribution function: $\mathbb{R} \to [0,1]$.



1. $F_X(x) = \mathbb{P}(X \le x)$ where $x \in \mathbb{R}$.
2. $\mathbb{P}(a < x \le b) = F_X(b) - F_X(a)$.
3. Right-continuous.
4. $\mathbb{P}(X = x) = F_X(x) - \lim_{h \downarrow 0} F_X(x-h)$, i.e., the jump in F_X at x for discrete rv's.

Probability density function: $\mathbb{R} \to [0, \infty)$.

1. $\int_{-\infty}^{x} f_X(t)dt = F_X(x)$, and $\frac{d}{dx}[F_X(x)] = f_X(x)$.
2. $\int_a^b f_X(t)dt = F_X(b) - F_X(a) = \mathbb{P}(a < X \le b) = \mathbb{P}(a \le X < b) = \mathbb{P}(a \le X \le b) = \mathbb{P}(a < X < b)$ since the endpoints have zero probability.
3. $\int_{-\infty}^{\infty} f_X(t)dt = 1$.

The k'th moment (about the origin): $\mu_k = \mathbb{E}(X^k)$.

The k'th central moment (about the mean):

$$v_k = \mathbb{E}((X-\mu)^k) = \mathbb{E}\left(\sum_{j=0}^{k} \binom{k}{j} X^j (-\mu)^{k-j}\right) = \sum_{j=0}^{k} \binom{k}{j} \mu_j (-\mu)^{k-j}.$$

Mean: $\mathbb{E}(X) = \mu_X = \sum_{x \in S_X} x p_X(x) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$ for discrete rv's.

$\mathbb{E}\big(\psi(X)\big) = \sum_{x \in S_X} \psi(x) p_X(x)$ for any real-valued ψ.

Rules similarly apply for the (absolute) continuous case.

Note that $\mathbb{E}\big(\psi(X)\big) \neq \psi\big(\mathbb{E}(X)\big)$, but $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$.

Variance: $\mathbb{V}(X) = \sigma^2 = \mathbb{E}\left(\big(X - \mathbb{E}(X)\big)^2\right) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathrm{sd}^2(X)$.

$\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$.

The standardised random variable $X_s = (X - \mu_X)/\sigma_X$ has mean 0 and variance 1.

Tail probability:

1. If $\mathbb{P}(X \geq 0) = 1$ with rv continuing, then for $n > 0$:
   $\mathbb{E}(X^n) = \int_0^\infty x^n f_X(x)dx = n\int_0^\infty \left(\int_0^x y^{n-1}dy\right) f_X(x)dx =$
   $n\int_0^\infty \left(\int_y^\infty f_X(x)dx\right) y^{n-1}dy = n\int_0^\infty y^{n-1}\mathbb{P}(X > y)dy = n\int_0^\infty y^{n-1}\big(1 - F_X(y)\big)dy.$
2. If $\mathbb{P}(X \leq 0) = 1$ with rv continuous, then for $n > 0$:
   $\mathbb{E}(X^n) = \int_{-\infty}^0 x^n f_X(x)dx = -n\int_{-\infty}^0 \left(\int_x^0 y^{n-1}dy\right) f_X(x)dx =$
   $-n\int_{-\infty}^0 \left(\int_{-\infty}^y f_X(x)dx\right) y^{n-1}dy = -n\int_{-\infty}^0 \mathbb{P}(X < y)y^{n-1}dy = -n\int_{-\infty}^0 y^{n-1}F_X(y)dy.$
3. If $\mathbb{P}(X \geq 0) = 1$ with rv discrete, then for $n > 0$:
   For $x \in [i, i+1)$ with $i \in \mathbb{Z}^{0+}$, observe that $F_X(x) = F_X(i)$.
   Then, $\mathbb{E}(X^n) = n\int_0^\infty x^{n-1}\big(1 - F_X(x)\big)dx = \sum_{i=0}^\infty n\int_i^{i+1} x^{n-1}\big(1 - F_X(x)\big)dx =$
   $\sum_{i=0}^\infty n\big(1 - F_X(i)\big)\int_i^{i+1} x^{n-1}dx = \sum_{i=0}^\infty \big(1 - F_X(i)\big)\big((i+1)^n - i^n\big).$
4. …

Bernoulli Distribution: $X \sim \text{Ber}(p)$.

1. para — success rate $\coloneqq p \in [0,1]$
2. rvmeaning — $X(\text{success}) = 1, X(\text{failure}) = 0$
3. pmf — $p_X(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$
4. mean — p
5. var — $p(1-p)$
6. pgf — $1 - p + pz, \ z \in \mathbb{R}$
7. mgf — $1 - p + pe^t$

Binomial Distribution: $X \sim \text{Bi}(n,p)$.

1. para — no. of mutually independent trials $\coloneqq n \in \mathbb{N}$
success rate $\coloneqq p \in [0,1]$
2. rvmeaning — no. of successes in n mutually independent Bernoulli trials
3. pmf — $p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \dots, n$
4. mean — np # n independent Bernoulli random variables
5. var — $np(1-p)$
6. pgf — $(1 - p + pz)^n, \ z \in \mathbb{R}$
7. mgf — $(1 - p + pe^t)^n$

Geometric Distribution: $N \sim \text{G}(p) \sim \text{Nb}(1,p)$.

1. para — success rate $\coloneqq p \in (0,1]$
2. rvmeaning — no. of failure until the first success occurs for independent Bernoulli trials
3. pmf — $p_N(n) = (1-p)^n p, \ n \in \mathbb{Z}^{0+}$
4. mean — $(1-p)/p$ # $\sum_{k=0}^{\infty} kn^{k-1} = \frac{d}{dn}[\sum_{k=0}^{\infty} n^k] = (1-n)^{-2}$
5. var — $(1-p)/p^2$ # find $\mathbb{E}\big(N(N-1)\big)$
6. pgf — $p(1-(1-p)z)^{-1}, \ |z| < 1/(1-p)$
7. mgf — $p(1-(1-p)e^t)^{-1}$

Lack-of-memory Property: If $T \sim \text{G}(p)$, then for $t = 0, 1, \dots$

$$\mathbb{P}(T \geq t) = p(1-p)^t + p(1-p)^{t+1} + \cdots = p\left(\frac{1}{1-(1-p)} - \frac{1-(1-p)^t}{1-(1-p)}\right) = (1-p)^t.$$

Then for given a and $t = 0, 1, \dots$ we have $\mathbb{P}(T - a \geq t \mid T \geq a) = (1-p)^t$.

The information that there have been no successes in the past a trials has no effect on the future waiting time to a success - the process "forgets".

Use the method of tail probability to find $\mathbb{V}(N)$.

$$\mathbb{E}(N) = \sum_{i=0}^{\infty} 1 - F_N(i) = \sum_{i=0}^{\infty} \mathbb{P}(N > i) = \sum_{i=0}^{\infty} \mathbb{P}(N \geq i) - p_N(i)$$
$$= \sum_{i=0}^{\infty} (1-p)^i - p(1-p)^i = (1-p)\sum_{i=0}^{\infty} (1-p)^i = \frac{1-p}{p}$$

$$\mathbb{E}(N^2) = ((i+1)^2 - i^2)\big(1 - F_N(i)\big) = 2\sum_{i=0}^{\infty} i(1-p)^{i+1} + \sum_{i=0}^{\infty} (1-p)^{i+1}$$
$$= 2(1-p)^2 \sum_{i=0}^{\infty} i(1-p)^{i-1} + \frac{1-p}{p} = \frac{2(1-p)^2}{\big(1-(1-p)\big)^2} + \frac{1-p}{p}$$
$$= \frac{(2-p)(1-p)}{p^2}$$

$$\mathbb{V}(N) = \frac{(2-p)(1-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}.$$

Negative Binomial Distribution: $Z \sim \mathrm{Nb}(r,p)$.

$$\begin{array}{c|c} F\ F\ \ldots\ F \quad S\ S\ \ldots\ S & S \\ \leftarrow \quad z \quad \rightarrow \quad \leftarrow r-1 \rightarrow & \end{array}$$

1. para — success rate $:= p \in (0,1]$
2. rvmeaning — no. of failures until the rth success
3. pmf — $p_Z(z) = \binom{-r}{z} p^r (p-1)^z = \binom{z+r-1}{r-1} p^r (1-p)^z,\ z \in \mathbb{Z}^{0+}$
4. mean — $r(1-p)/p$ # use PGF
5. var — $r(1-p)/p^2$
6. pgf — $p^r(1-(1-p)z)^{-r},\ |z| < 1/(1-p)$
7. mgf — $p^r(1-(1-p)e^t)^{-r}$

Negative Binomial to Binomial:

$$\{\ \mathrm{Nb}(r,p) \leq n-r\ \} \sim \{\ \text{at most } n-r \text{ failures before the } r\text{th success}\ \}$$
$$\sim \{\ \text{at most } n \text{ trials to get } r \text{ successes}\ \}$$
$$\sim \{\ \text{no. of successes in first } n \text{ trials} \geq r\ \} \sim \{\ \mathrm{Bi}(n,p) \geq r\ \}$$

Extended Binomial Theorem: For any $r \in \mathbb{R}$ (more than integer) we have

$$(1+b)^r = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} b^k = \sum_{k=0}^{\infty} \frac{r(r-1)\cdots(r-k+1)}{k!} b^k = \sum_{k=0}^{\infty} \binom{r}{k} b^k$$

which converges provided $|b| < 1$.

Provided $p \in (0,1]$, it follows that $\sum_{z=0}^{\infty} \binom{-r}{z} p^r (p-1)^z = p^r\big(1+(p-1)\big)^{-r} = 1$.

Hypergeometric Distribution: $Z \sim \mathrm{Hg}(n, D, N)$.

1. para — no. of defectives $:= D$; $0 \leq \min\{n, D\} \leq N$
2. rvmeaning — (sampling without replacement) no. of defectives obtained in the sample of n
3. pmf — $p_X(x) = \binom{D}{x}\binom{N-D}{n-x} / \binom{N}{n}$, $\max\{0, n+D-N\} \leq x \leq \min\{n, D\}$
4. mean — nD/N
5. var — $(nD(N-D)/N^2) \times \big(1 - (n-1)/(N-1)\big)$

From an ordinary deck of 52 cards, cards are drawn one by one, at random and without replacement. What is the probability the the fourth heart is drawn on the tenth draw?

Let F be the event that in the first nine draws there are exactly three hearts, and E be the event that the tenth draw is a heart. Then

$$\mathbb{P}(E \cap F) = \mathbb{P}(F)\mathbb{P}(E|F) = \binom{13}{3}\binom{39}{6} / \binom{52}{9} \times \frac{10}{43} = 0.059.$$

Poisson Distribution: $N \sim \mathrm{Pn}(\lambda)$.

1. para — avg no. of events by time 1 $:= \lambda > 0$ # rate parameter
2. rvmeaning — no. of events which occur by (continuous) time 1
3. pmf — $p_N(n) = e^{-\lambda}\lambda^n / n!$, $n \in \mathbb{Z}^{0+}$
4. mean — λ
5. var — λ # find $\mathbb{E}\big(X(X-1)\big)$
6. pgf — $e^{-\lambda(1-z)}$, $z \in \mathbb{R}$
7. mgf — $e^{\lambda(e^t - 1)}$

Poisson as an approximate model: $\mathrm{Bi}(n, p) \approx \mathrm{Pn}(np)$ for $p \leq 0.05$.

Discrete Uniform Distribution: $X \sim \mathrm{U}(m, n)$.

1. para — $m \leq n$
2. pmf — $p_X(x) = (n-m+1)^{-1}$, $x = m, m+1, \ldots, n$
3. mean — $(m+n)/2$
4. var — $((n-m+1)^2 - 1)/12$

Find the variance of discrete uniform distribution $U(m,n)$.

$$\mathbb{E}(X^2) = \frac{1}{n-m+1} \times \sum_{x=m}^{n} x^2$$

$$= \frac{\sum_{x=m}^{n}((x+1)^3 - x^3) - 3\sum_{x=m}^{n} x - \sum_{x=m}^{n} 1}{3(n-m+1)} \qquad \text{\# telescoping series}$$

$$= \frac{(n+1)^3 - m^3 - \frac{3(m+n)(n-m+1)}{2} - (n-m+1)}{3(n-m+1)}$$

$$= \frac{2m^2 + 2n^2 + 2mn + n - m}{6}$$

$$\mathbb{V}(X) = \frac{2m^2 + 2n^2 + 2mn + n - m}{6} - \left(\frac{m+n}{2}\right)^2 = ((n-m+1)^2 - 1)/12\,.$$

Continuous Uniform (Rectangular) Distribution: $X \sim \mathrm{R}(a,b)$.

1. para $\quad a \le b$
2. pdf $\quad f_X(x) = (b-a)^{-1}, \;\; a \le x \le b$
3. cdf $\quad F_X(x) = (x-a)/(b-a), \;\; a \le x \le b$
4. mean $\quad (a+b)/2$
5. var $\quad (b-a)^2/12$
6. mgf $\quad M_X(t) = \begin{cases} 1, & t = 0 \\ (e^{tb} - e^{ta})/t(b-a), & t \neq 0 \end{cases}$

For a continuous distribution, $F(X) \sim \mathrm{Unif}(0,1)$ since $\mathbb{P}(F(X) \le w) = \mathbb{P}\big(X \le F^{-1}(w)\big) = F\big(F^{-1}(w)\big) = w, \;\; 0 \le w \le 1.$

Let $Y = X_1 + \cdots + X_{15}$ be the sum of i.i.d. rv's, each with PDF $f(x) = (3/2)x^2$ where $-1 < x < 1$. Approximate $\mathbb{P}(-0.3 < Y < 0.5)$ by simulation.

First, calculate the CDF,

$$F(x) = \int_{-1}^{x} \frac{3}{2} y^2 dy = \left[\frac{1}{2} y^3\right]_{-1}^{x} = \frac{x^3+1}{2}\,, \quad -1 < x < 1.$$

Then invert to get the inverse CDF: $F^{-1}(p) = (2p-1)^{1/3}$, which we use to simulate X.

```
# Function to handle powers for negative numbers properly.
exponent <- function(x, p)
    sign(x) * abs(x)^p

# Function to generate random X's.
rx <- function(n)
    exponent(2 * runif(n) - 1, 1/3)
```

```
# Function to generate random Y's.
ry <- function(n) {
    y <- 1:n
    for (i in 1:n)
        y[i] <- sum(rx(15))
    return(y)
}

# A more efficient way to do the same thing is:
ry <- function(n)
    replicate(n, sum(rx(15)))

# Simulate Y's.
ys <- ry(10000)

# Estimate the probability.
mean((-0.3 < ys) & (ys < 1.5))
```

```
[1] 0.2277
```

Exponential Distribution: $T \sim \exp(\alpha) \sim \gamma(1, \alpha)$.

1. para — avg no. of events by time 1 $:= \alpha > 0$ — # rate parameter
2. rvmeaning — waiting time until the first event occurs
3. pdf — $f_T(t) = \alpha e^{-\alpha t},\ t \geq 0$ — # from CDF
4. cdf — $F_T(t) = 1 - e^{-\alpha t},\ t \geq 0$ — # $\mathbb{P}(T > t) = \lim_{n\to\infty} \left(1 - \frac{\alpha}{n}\right)^{nt}$
5. mean — α^{-1} — # use the tail probability
6. var — α^{-2}
7. mgf — $\alpha/(\alpha - t),\ t > \alpha$

Let T, the lifetime (in years) of a radio tube, be exponentially distributed with mean $1/\lambda$. Prove that $\lfloor T \rfloor$, which is the complete number of years that the tube works, is a geometric random variable.

$$p_N(n) = \mathbb{P}(\lfloor T \rfloor = n) = \mathbb{P}(n \leq T < n + 1) = F_T(n + 1) - F_T(n) = \left(e^{-\lambda}\right)^n \left(1 - e^{-\lambda}\right).$$

Lack-of-memory Property: If $T \sim \exp(\alpha)$ then, for $t \in [0, \infty)$, $\mathbb{P}(T \geq t) = e^{-\alpha t}$.

Then for given $x, y \in [0, \infty)$, we have $\mathbb{P}(T - a \geq t | \, T \geq a) = e^{-\alpha t}$.

The time between the first and second heart attacks for a certain group of people is an exponential random variable. If 40% of those who have had a heart attack will have another one within the next five years, what is the probability that a person who had one heart attack five years ago will not have another one in the next five years?

Let T be the waiting time until the next heart attack.

$$\mathbb{P}(T \geq 10 | \, T \geq 5) = \mathbb{P}(T \geq 5) = 1 - F_T(5) = 1 - 0.4 = 0.6\,.$$

Mr. Jones is waiting to make a phone call at a train station. There are two public telephone booths next to each other, occupied by two persons, say A and B. If the duration of each telephone call is an exponential random variable with $\alpha = 1/8$, what is the probability that among Mr. Jones, A, and B, Mr. Jones will not be the last to finish his call?

Let X, Y, Z be the duration of phone call by Mr. jones, A, B (or B, A - depends on who finishes his call first) respectively.

From the memoryless property, we have $\mathbb{P}(Z - Y > X | \, Z > Y) = \mathbb{P}(Z > X)$.

Assume Mr. Jones' call takes time t to finish, then use Law of Total Probability:

$$\mathbb{P}(Z > X) = \int_{\text{all } t} \mathbb{P}(Z > t)\mathbb{P}(X = t)dt = \int_{\text{all } t} e^{-\frac{t}{8}}\left(\frac{1}{8}e^{-\frac{t}{8}}\right)dt = 0.5\,.$$

Gamma (Erlang) Distribution: $Z \sim \gamma(r, \alpha)$. # $\gamma(r, \alpha) = n\gamma(r, \alpha/n)$

1. para — no. of events occurred $:= r \in \mathbb{R}^+$; avg no. of events by time $1 := \alpha > 0$ # rate parameter
2. rvmeaning — waiting time until the rth event occurs
3. pdf — $f_Z(z) = \left(\alpha^r z^{r-1}/\Gamma(r)\right)e^{-\alpha z}, \;\; z > 0$ # from CDF
4. cdf — $F_Z(z) = 1 - \sum_{k=0}^{r-1}((\alpha z)^k/k!)e^{-\alpha z}, \;\; z > 0$ # find $\mathbb{P}(Z > z)$
5. mean — r/α # the kth moment is $\Gamma(r + k)/\Gamma(r)\alpha^k$
6. var — r/α^2
7. mgf — $\alpha^r/(\alpha - t)^r, \;\; t < \alpha$

Beta Distribution: $X \sim \text{Beta}(\alpha, \beta)$. # $\text{Beta}(1, 1) \sim \text{U}(0, 1)$

1. para — $\alpha, \beta > 0$
2. pdf — $f_X(x) = x^{\alpha-1}(1 - x)^{\beta-1}/\,\text{B}(\alpha, \beta)\,, \;\; 0 \le x \le 1$
3. mean — $\alpha/(\alpha + \beta)$ # the kth moment is $\text{B}(\alpha + k, \beta)\,/\,\text{B}(\alpha, \beta)$
4. var — $\alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1)$
5. mode — $(\alpha - 1)/(\alpha + \beta - 2)$ if $\alpha, \beta > 2$

Pareto Distribution: $X \sim \text{Pareto}(\alpha, \gamma)$.

1. para — $\alpha, \gamma > 0$
2. pdf — $f_X(x) = \gamma\alpha^\gamma/x^{\gamma+1}, \;\; x \ge \alpha$
3. cdf — $F_X(x) = 1 - \alpha^\gamma/x^\gamma, \;\; x \ge \alpha$
4. mean — $\gamma\alpha/(\gamma - 1)$ if $\gamma > 1$
5. var — $\gamma\alpha^2/(\gamma - 1)^2(\gamma - 2)$ if $\gamma > 2$

Normal (Gaussian) Distribution: $X \sim \mathrm{N}(\mu, \sigma^2) \sim \mu + \sigma Z$.

1. para $\sigma > 0$
2. pdf $f_X(x) = \left(\sigma\sqrt{2\pi}\right)^{-1} e^{-(x-\mu)^2/2\sigma^2}, \ x \in \mathbb{R}$
3. mean μ # from the standard form
4. var σ^2
5. mgf $e^{\mu t + \sigma^2 t^2/2}, \ t \in \mathbb{R}$

Standard Normal Distribution: $Z \sim \mathrm{N}(0, 1)$.

1. pdf $\varphi(z) = f_Z(z) = \left(\sqrt{2\pi}\right)^{-1} e^{-z^2/2}, \ z \in \mathbb{R}$
2. cdf $\Phi(z) = F_Z(z) = \int_{-\infty}^{z} \left(\sqrt{2\pi}\right)^{-1} e^{-t^2/2} dt, \ z \in \mathbb{R}$
3. mean 0
4. var 1
5. mgf $e^{t^2/2}, \ t \in \mathbb{R}$

As $\varphi(z)$ is an even function, we have $\Phi(-z) = 1 - \Phi(z)$.

Find the mean and variance of the normal distribution $\mathrm{N}(\mu, \sigma^2)$.

For $Z \sim \mathrm{N}(0, 1)$, we have $\mathbb{E}(Z^n) = \int_{-\infty}^{\infty} z^n \left(\sqrt{2\pi}\right)^{-1} e^{-\frac{1}{2}z^2} dz$.

Integrating by parts with $u = z^{n-1}$ and $dv = \left(\sqrt{2\pi}\right)^{-1} z e^{-\frac{1}{2}z^2} dz$, then obtain

$$\begin{aligned}\mathbb{E}(Z^n) &= \int_{\infty}^{\infty} u \, dv = [uv]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{du}{dz} v \, dz \\ &= \left[z^{n-1}\left(-\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\right)\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (n-1) z^{n-2} \left(-\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\right) dz\end{aligned}$$

Therefore, we have $\mathbb{E}(Z^n) = (n-1)\mathbb{E}(Z^{n-2})$.

Also, we know that $\mathbb{E}(Z^0) = \mathbb{E}(1) = 1$ and $\mathbb{E}(Z^1) = 0$ as the PDF is an even function. It follows that $\mathbb{E}(Z^{2k+1}) = 0$ and $\mathbb{E}(Z^{2k}) = (2k-1)(2k-3)\cdots 1 = (2k)!/2^k k!$.

For $k = 1$ we have $\mathbb{E}(Z^2) = 1$ and hence $\mathbb{V}(Z) = 1 - 0^2 = 1$. Consequently for $X \sim \mathrm{N}(\mu, \sigma^2)$, $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \sigma^2$.

Normal as an approximate model:

1. $\mathrm{Bi}(n, p) \approx \mathrm{N}\big(np, np(1-p)\big)$ for $np > 5, n(1-p) > 5$;
2. $\mathrm{Pn}(\lambda) \approx \mathrm{N}(\lambda, \lambda)$ for … ;
3. $\gamma(r, \alpha) \approx \mathrm{N}(r/\alpha, r/\alpha^2)$ for … .

If $Z_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$, then $\sum_{i=1}^{n} a_i z_i \sim \mathrm{N}(\sum_{i=1}^{n} a_i \mu_i \, , \sum_{i=1}^{n} a_i \sigma_i^2)$.

Standard Normal Probabilities

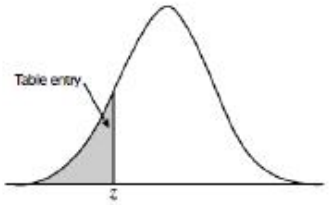


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

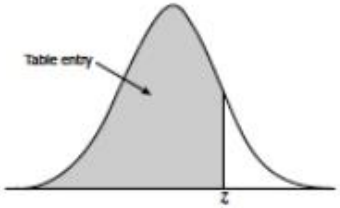


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Weibull Distribution: $X \sim \text{Weibull}(\beta, \gamma)$.

1. para $\beta, \gamma > 0$
2. pdf $f_X(x) = (\gamma x^{\gamma-1}/\beta^{\gamma})e^{-(x/\beta)^{\gamma}}, \ x \geq 0$
3. cdf $F_X(x) = 1 - e^{-(x/\beta)^{\gamma}}, \ x \geq 0$
4. mean $\beta\Gamma\big((\gamma+1)/\gamma\big)$ # by substituting $u = (x/\beta)^{\gamma}$
5. var $\beta^2\left[\Gamma\big((\gamma+2)/\gamma\big) - \Big(\Gamma\big((\gamma+1)/\gamma\big)\Big)^2\right]$

Cauchy Distribution: $X \sim \text{C}(m, a)$.

1. para location $\coloneqq m$; scale $\coloneqq a > 0$
2. pdf $f_X(x) = \pi^{-1}a/(a^2 + (x-m)^2), \ x \in \mathbb{R}$
3. cdf $F_X(x) = \pi^{-1}\arctan\big((x-m)/a\big) + 1/2, \ x \in \mathbb{R}$
4. mean, var, mgf N.D.

Transformations of random variables.

Let $Y = \psi(X)$, for $y \in \mathbb{R}$, if ψ is continuous and strictly increasing on S_X:

$$F_Y(y) = \mathbb{P}(\psi(X) \leq y) = \mathbb{P}\big(X \leq \psi^{-1}(y)\big) = F_X\big(\psi^{-1}(y)\big)\,;$$

If ψ is discrete and strictly decreasing:

$$F_Y(y) = \mathbb{P}(\psi(X) \leq y) = \mathbb{P}\big(X \geq \psi^{-1}(y)\big) = 1 - \lim_{h \downarrow 0} F_X(\psi^{-1}(y) - h)\,.$$

Lognormal Distribution: $Y \sim \text{LN}(\mu, \sigma^2)$. # derived from $Y = e^X, \ X \sim N(\mu, \sigma^2)$

1. para $\sigma > 0$
2. pdf $f_Y(y) = \big(\sqrt{2\pi}\sigma y\big)^{-1}e^{-(\ln y - \mu)^2/2\sigma^2}, \ y > 0$ # from CDF
3. mean $e^{\mu + \sigma^2/2}$
4. var $e^{2\mu+\sigma^2}\big(e^{\sigma^2} - 1\big)$
5. mgf N.D.

Find the mean and variance of the Lognormal Distribution $\text{LN}(\mu, \sigma^2)$.

Let $Z \sim \text{N}(0, 1)$, then for $r \in \mathbb{R}$,

$$\mathbb{E}(e^{rZ}) = \int_{-\infty}^{\infty} e^{rz}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = e^{\frac{r^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-r)^2}{2}}dz = e^{r^2/2}.$$

Then we can write $\mathbb{E}(Y^r) = \mathbb{E}\big(e^{(\mu+\sigma Z)r}\big) = e^{\mu r}\mathbb{E}\big(e^{(\sigma r)Z}\big) = e^{\mu r + \sigma^2 r^2/2}$.

A bivariate random variable is a function which maps Ω into $\mathbb{R}^2$.

Joint PMF:

1. $p_{(X,Y)}(x,y) = \mathbb{P}(X = x, Y = y), \quad (x,y) \in S_{(X,Y)}$
2. $\sum_x \sum_y p_{(X,Y)}(x,y) = 1.$

Marginal PMF:

1. $p_X(x) = \sum_{y \in S_Y} p_{(X,Y)}(x,y)$
2. $p_Y(y) = \sum_{X \in S_X} p_{(X,Y)}(x,y).$

Discrete Bivariate CDF:

1. $F_{(X,Y)}(x,y) = \mathbb{P}(X \le x, Y \le y) = \sum_{u \le x, v \le y} p_{(X,Y)}(u,v)$
2. $\mathbb{P}(a < x \le b, c < Y \le d) = F(b,d) - F(a,d) - F(b,c) + F(a,c)$
3. $F_{(X,Y)}(x,\infty) = F_X(x)$, and $F_{(X,Y)}(\infty,y) = F_Y(y)$.

Joint PDF and Continuous Bivariate CDF:

1. $f_{(X,Y)}(x,y) \ge 0$
2. $\frac{\partial^2}{\partial x \partial y}\left[F_{(X,Y)}(x,y)\right] = f_{(X,Y)}(x,y)$, and $\iint_{\{(u,v):\, u \le x \,\wedge\, v \le y\}} f_{(X,Y)} du dv = F_{(X,Y)}(x,y)$
3. $\mathbb{P}(a < X \le b, c < Y \le d) = \int_a^b \int_c^d f_{(X,Y)}(x,y) dy\, dx.$

Marginal PDF:

1. $f_X(x) = \int_{\text{all } y} f_{(X,Y)}(x,y) dy\,, \quad x \in \cdots$
2. $f_Y(y) = \int_{\text{all } x} f_{(X,Y)}(x,y) dx\,, \quad y \in \cdots$
3. Having the same marginal PDF $\not\Rightarrow$ having a same joint PDF.

On a line segment AB of length l, two points C and D are placed at random and independently. What is the probability that C is closer to D than to A?

Let X and Y be the distance to point A from point C and point D (uniform distributions).

$$f_{(X,Y)}(x,y) = \begin{cases} f_X(x) f_Y(y) = l^{-2}, & \text{if } 0 \le x \le l, 0 \le y \le l \\ 0, & \text{otherwise.} \end{cases}$$

Hence

$$\mathbb{P}(C \text{ is closer to } D \text{ than } A) = \mathbb{P}(|X - Y| < X) = \mathbb{P}\left(l \ge X \ge \frac{Y}{2}, l \ge Y \ge 0\right)$$
$$= \int_0^l \int_{y/2}^l l^{-2} dx\, dy = 3/4\,.$$

Conditional PMF & PDF:

1. $p_{X|Y}(x|y) = \mathbb{P}(X = x|\, Y = y) = p_{(X,Y)}(x,y)/p_Y(y)$ if $p_Y(y) > 0$
2. $f_{X|Y}(x|y) = f_{(X,Y)}(x,y)/f_Y(y)$ if $f_Y(y) > 0$
3. $f_Y(y) = \int_{\text{all } x} f_{(X,Y)}(x,y)dx = \int_{\text{all } y} f_{X|Y}(x|y)f_Y(y)dx\,,\ \ y \in \cdots.$

Standard Bivariate Normal Distribution: $(X,Y) \sim \mathrm{N}_2(\rho)$.

1. para — correlation coefficient $:= \rho \in (-1,1)$
   # $\rho > 0 \Rightarrow$ positively related; $\rho = 0 \Rightarrow$ independence
2. rvmeaning — $X, Y \sim \mathrm{N}(0,1)$
3. joint_pdf — $f_{(X,Y)}(x,y) = \dfrac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\dfrac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right),\ \ x,y \in \mathbb{R}$
4. marg_pdf — $f_X(z) = f_Y(z) = \left(\sqrt{2\pi}\right)^{-1} e^{-z^2/2}$
5. cond_pdf — $f_{X|Y}(x|y) = \dfrac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\dfrac{(x-\rho y)^2}{2(1-\rho^2)}\right) \sim \mathrm{N}(\rho y, 1-\rho^2)$

Find the marginal PDF $f_Y(y)$ of the standard bivariate normal distribution $\mathrm{N}_2(\rho)$.

$$
\begin{aligned}
f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dx &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dx \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}y^2} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}y^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} \sqrt{1-\rho^2}\, du \qquad \left[\text{substitute } u = \frac{x-\rho y}{\sqrt{1-\rho^2}}\right] \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}
\end{aligned}
$$

General Bivariate Normal Distribution: $(X,Y) \sim \mathrm{N}_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}_2\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_Y\sigma_X \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right).$$

1. rvmeaning — $X \sim \mathrm{N}(\mu_X, \sigma_X^2)\,, Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$
2. cond_pdf — $(X|\, Y = y) \sim \mathrm{N}\left(\mu_X + \sigma_X\rho\left(y - \dfrac{\mu_Y}{\sigma_Y}\right), \sigma_X^2(1-\rho^2)\right)$
3. joint_pdf — $f_{(X,Y)}(x,y) = f_{X|Y}(x|y)f_Y(y)$

Independence requires either of:

1. CDF equals the product of its marginal CDF's;
2. PMF or PDF is the product of its marginal PMF's or PDF's;
3. conditional PMF or PDF is same as its marginal PMF or PDF.

Let X and Y be two independent random variables with distribution functions F and G, respectively. Find the distribution functions of $U = \max\{X, Y\}$ and $V = \min\{X, Y\}$.

$$F_U(u) = F(u)G(u), \text{ and } F_V(v) = F(v) + G(v) - F(v)G(v).$$

Probability distribution of sum. If X and Y are independent:

1. discrete case:

$$p_{X+Y}(z) = \mathbb{P}(X + Y = z) = \sum_{i=0}^{z} \mathbb{P}(X = i)\mathbb{P}(Y = z - i) = \sum_{i=0}^{z} p_X(i)p_Y(z - i)\,;$$

2. continuous case:

$$\begin{aligned} F_{X+Y}(z) = \mathbb{P}(X + Y \le z) &= \iint_{\{(x,y):\, x+y\le z\}} f_{(X,Y)}(x, y)\, dxdy \\ &= \int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^{z-y} f_X(x)dx \right) dy = \int_{-\infty}^{\infty} f_Y(y)F_X(z - y)dy \end{aligned}$$

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} \frac{\partial}{\partial z}[f_Y(y)F_X(z - y)]dy = \int_{S_Y} f_Y(y)f_X(z - y)I(\text{dom}(z - y))dy.$$

Let X and Y be two random variables randomly selected from the interval $(0, 1)$ and $(1, 3)$. Find the probability density function of their sum.

Let $T = X + Y$, the range of T is $(1, 4)$. Now since Y is defined under $(1, 3)$, we have $1 < t - x < 3$ and so $\max\{0, t - 3\} < x < \min\{1, t - 1\}$.

The required PDF is:

$$f_T(t) = \begin{cases} \int_{t-3}^{1} \frac{1}{2} dx = 2 - \frac{1}{2}t, & 3 < t < 4 \\ \int_{0}^{1} \frac{1}{2} dx = \frac{1}{2}, & 2 \le t \le 3 \\ \int_{0}^{t-1} \frac{1}{2} dx = \frac{1}{2}t - \frac{1}{2}, & 1 < t < 2 \\ 0, & \text{otherwise} \end{cases}$$

Probability distribution of product. If $U = XY$, where X and Y are independent:

$$\begin{aligned} F_U(u) = \mathbb{P}(XY \le u) &= \int_{S_X} \mathbb{P}(xY \le u) f_X(x)dx \\ &= \int_{S_X \cap (-\infty,0)} \mathbb{P}\left(Y \ge \frac{u}{x}\right) f_X(x)dx + \int_{S_X \cap (0,\infty)} \mathbb{P}\left(Y \le \frac{u}{x}\right) f_X(x)dx \\ &= \int_{S_X \cap (-\infty,0)} \left(1 - F_Y\left(\left(\frac{u}{x}\right)^-\right)\right) f_X(x)dx + \int_{S_X \cap (0,\infty)} F_Y\left(\frac{u}{x}\right) f_X(x)dx. \end{aligned}$$

If (X, Y) are continuous then, the PDF of U can be obtained by simple differentiation:

$$f_U(u) = \int_{S_X \cap (0,\infty)} \frac{1}{x} f_Y\left(\frac{u}{x}\right) f_X(x)dx - \int_{S_X \cap (-\infty,0)} \frac{1}{x} f_Y\left(\frac{u}{x}\right) f_X(x)dx.$$

An alternative expression can be obtained by interchanging X and Y.

If the density is symmetric in x and y (i.e., value does not change by swapping x and y), then the marginal PDF's are the same, and X and Y are said to be identically distributed.

If X and Y are independent, then $\mathbb{E}(X_1X_2 \cdots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2) \cdots \mathbb{E}(X_n)$. However, the converse is not always true.

For any random variable X and Y, $\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)$.

Suppose that 80 balls are placed into 40 boxes at random and independently. What is the expected number of empty boxes?

Each time place one ball, each box has a probability of 39/40, that it won't receive the ball. Repeat 80 times. At the end each box has a probability of $(39/40)^{80} = 0.1319$ that it won't receive any ball.

Thus, the expected value of empty boxes is just $40 \times 0.1319 = 5.276$.

Variance of sum:

$$\begin{aligned} \mathbb{V}(X + Y) = \mathbb{E}\left(\left(X + Y - (\mu_X + \mu_Y)\right)^2\right) &= \mathbb{E}\left(\left((X - \mu_X) + (Y - \mu_Y)\right)^2\right) \\ &= \mathbb{E}((X - \mu_X)^2) + 2\mathbb{E}\big((X - \mu_X)(Y - \mu_Y)\big) + \mathbb{E}((Y - \mu_Y)^2) \\ &= \mathbb{V}(X) + 2\,\mathrm{Cov}(X, Y) + \mathbb{V}(Y). \end{aligned}$$

where we define covariance

$$\begin{aligned} \mathrm{Cov}(X, Y) = \sigma_{XY} = \mathbb{E}\big((X - \mu_X)(Y - \mu_Y)\big) &= \mathbb{E}(XY) - \mu_X\mathbb{E}(Y) - \mu_Y\mathbb{E}(X) + \mu_X\mu_Y \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(X|Y)Y) - \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

Clearly $\mathrm{Cov}(X, Y)$ contains the information about the relationship between X and Y.

If $\text{Cov}(X,Y) > 0$ then, X, Y are positively related; if $\text{Cov}(X,Y) < 0$ then, X, Y are negatively related; if $\text{Cov}(X,Y) = 0$ then, X, Y are uncorrelated.

If X, Y are independent, X, Y are also uncorrelated; but the converse is not always true.

Using induction, this gives $\mathbb{V}(X_1 + X_2 + \cdots + X_n) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + \cdots + \mathbb{V}(X_n)$ for mutually independent $X_1, X_2, \ldots, X_n$.

Properties of covariance:

1. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
2. $\text{Cov}(X,X) = \mathbb{V}(X)$
3. $\text{Cov}(X+Y,Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$
4. $\text{Cov}(cX,Y) = c\,\text{Cov}(X,Y)$ where c is a constant
5. $\mathbb{V}(X-Y) = \mathbb{V}(X) - 2\,\text{Cov}(X,Y) + \mathbb{V}(Y)$.

Define $X_0 = X - \mu_X$ and $Y_0 = Y - \mu_Y$, then:

1. $\mathbb{E}(X_0) = \mathbb{E}(Y_0) = 0$
2. $\mathbb{V}(X) = \mathbb{E}(X_0^2),\ \mathbb{V}(Y) = \mathbb{E}(Y_0^2)$
3. $\text{Cov}(X,Y) = \mathbb{E}(X_0 Y_0)$
4. If $Z = aX + bY$ then, $Z_0 = aX_0 + bY_0$.

Therefore, deduce that $\text{Cov}(aX+bY, cX+dY) = \mathbb{E}\big((aX_0+bY_0)(cX_0+dY_0)\big) = ac\mathbb{E}(X_0^2) + (ad+bc)\mathbb{E}(X_0Y_0) + bd\mathbb{E}(Y_0^2) = ac\mathbb{V}(X) + (ad+bc)\,\text{Cov}(X,Y) + bd\mathbb{V}(Y)$.

The correlation coefficient is defined as $\rho(X,Y) = \text{Cov}(X,Y) \,/\, \text{sd}(X)\,\text{sd}(Y)$.

There is a linear relation between X and Y, if $\rho = \pm 1$. If $\rho = 1$ then, $\mathbb{P}(X = a + bY) = 1$ where $b > 0$; if $\rho = -1$ then, $\mathbb{P}(X = c + dY) = 0$ where $d > 0$.

Prove that $|\rho| < 1$.

$$\mathbb{V}(zX+Y) = \mathbb{V}(zX) + 2\,\text{Cov}(zX,Y) + \mathbb{V}(Y) = \mathbb{V}(X)z^2 + 2\,\text{Cov}(X,Y)\,z + \mathbb{V}(Y) \geq 0$$

$$\triangle = b^2 - 4ac = (2\,\text{Cov}(X,Y))^2 - 4\mathbb{V}(X)\mathbb{V}(Y) \leq 0 \ \Rightarrow\ |\rho| \leq 1$$

If X is discrete, then $\mathbb{E}(Y|X) = \sum_i \mathbb{E}(Y|X = x_i)\,\mathbf{1}_{\{X=x_i\}}$.

$\mathbb{E}\big(\psi(X,Y)\big) = \mathbb{E}\big(\mathbb{E}(\psi(X,Y)|Y)\big)$, and $\mathbb{V}(X) = \mathbb{V}\big(\mathbb{E}(X|Y)\big) + \mathbb{E}\big(\mathbb{V}(X|Y)\big)$.

What is the expected number of random digits that should be generated to obtain three consecutive zeros?

Let X be the number of random digits to be generated until three consecutive zeroes are obtained. Let Y be the number of random digits to be generated until the first nonzero digit is obtained (including the first nonzero digit). Then

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} \mathbb{E}(X|Y)\mathbb{P}(Y=i) = \sum_{i=1}^{3} \mathbb{E}(X|Y)\mathbb{P}(Y=i) + \sum_{i=4}^{\infty} \mathbb{E}(X|Y)\mathbb{P}(Y=i)$$
$$= \sum_{i=1}^{3} \big(i + \mathbb{E}(X)\big) \times 0.1^{i-1}0.9 + \sum_{i=4}^{\infty} 3 \times 0.1^{i-1}0.9 = 0.999\mathbb{E}(X) + 0.11$$

Hence $\mathbb{E}(X) = 1100$.

Suppose the number of insurance claims is Poisson with mean 20 per year. Suppose the claim sizes are independent rv's – each with mean 200 and standard deviation 200, and suppose the number of claims and claim sizes are independent. Determine the variance of the total value of claims in a year.

Let N be a non-negative integer-valued random variable independent of X_i's, where X_i's are mutually independent with mean and all with 200 and variance 400. The total value of claims in a year is $T = \sum_{i=1}^{N} X_i$.

$$\mathbb{E}(T|\, N = n) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = 200n, \;\; \mathbb{V}(T|\, N = n) = \mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = 40000n$$

This gives $\mathbb{V}(T) = \mathbb{V}\big(\mathbb{E}(T|N)\big) + \mathbb{E}\big(\mathbb{V}(T|N)\big) = \mathbb{V}(200N) + \mathbb{E}(40000N) = \$\, 1.6 \times 10^6$.

Bienaymé Inequality: If X has mean μ and variance σ^2 then, $\mathbb{P}(|X-\mu|/\sigma \geq k) \leq k^{-2}$.

Chebyshev's Inequality: Let $\epsilon = k\sigma$, then $\mathbb{P}(|X-\mu| \geq \epsilon) \leq \sigma^2/\epsilon^2$.

The average IQ score on a certain campus is 110. If the variance of these scores is 15, what can be said about the percentage of students with an IQ below 140?

Let X be a student's IQ score,

$$\mathbb{P}(X \leq 140) = 1 - \mathbb{P}(X - 110 \geq 140 - 110) = 1 - \mathbb{P}(|X-110| \geq 30) + \mathbb{P}(X \leq 80)$$
$$\geq 1 - \frac{15}{30^2} + \mathbb{P}(X \leq 80) \geq 0.983\,.$$

Probability generating functions (<u>for $x \in \mathbb{Z}^{0+}$</u>):

1. $P_X(z) = \mathbb{E}(z^X) = \sum_{x=0}^{\infty} p_X(x) z^x$
2. PGF will always converge for $|z| \leq 1$
3. $P_X(1) = 1$
4. $p_X(k) = P_X^{(k)}(0)/k!$ # with respect to z
5. $\mathbb{E}(X) = P_X'(1)$, $\mathbb{E}\big(X(X-1)\big) = P_X''(1)$, then $\mathbb{V}(X) = P_X''(1) + P_X'(1) - P_X'(1)^2$
6. $\mathbb{P}(X \text{ is even}) = \sum_{i=0}^{\infty} p_X(2i) = \big(P_X(1) + P_X(-1)\big)/2$, and $\mathbb{P}(X \text{ is odd}) = \sum_{j=0}^{\infty} p_X(2j+1) = \big(P_X(1) - P_X(-1)\big)/2$
7. If X and Y are independent, and $W = X + Y$, then $P_W(z) = \mathbb{E}(z^{X+Y}) = \mathbb{E}(z^X)\mathbb{E}(z^Y) = \mathbb{E}(z^{X+Y})$.

Find the PGF for the Negative Binomial Distribution $X \sim \text{Nb}(r, p)$.

$$p_X(z) = \sum_{x=0}^{\infty} \binom{-r}{x} p^r (p-1)^x z^x = p^r \sum_{x=0}^{\infty} \binom{-r}{x} \big((p-1)z\big)^x = p^r \big(1 - (1-p)\big)^{-r}.$$

The ratio test gives that $P_X(z)$ converges absolutely only for z with $|z| < 1/(1-p)$ since

$$\lim_{x\to\infty} \left| \frac{\binom{-r}{x+1}\big((p-1)z\big)^{x+1}}{\binom{-r}{x}\big((p-1)z\big)^x} \right| = \left| \frac{(x+r)(p-1)z}{x+1} \right| = |(p-1)z| < 1.$$

Moment generating functions:

1. $M_X(t) = \mathbb{E}(e^{Xt}) = \mathbb{E}(\sum_{k=0}^{\infty} (Xt)^k/k!) = \mathbb{E}(\sum_{k=0}^{\infty} \mu_k t^k/k!)$
2. $M_X(t) > 0$
3. $M_X(0) = 1$
4. $\mathbb{E}(X^k) = M_X^{(k)}(0)$ for any k, then $\mathbb{V}(X) = M_X''(0) - M_X'(0)^2$ # with respect to t
5. If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$

6. If X and Y are independent, and $Z = X + Y$, then $M_Z(t) = \mathbb{E}\left(e^{t(X+Y)}\right) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY}) = M_X(t)M_Y(t)$
7. If X is a discrete random variable defined on $\mathbb{Z}^{0+}$, having PGF $P_X(z)$, then $M_X(t) = P_X(e^t)$ and $P_X(z) = M_X(\log z)$.
8. The central moment generating function is given by

$$N_X(t) = \mathbb{E}\left(e^{(X-\mu)t}\right) = \mathbb{E}\left(\sum_{k=0}^{\infty} v_k t^k/k!\right) \quad \text{for discrete case,}$$

and we see that $N_X(t) = \mathbb{E}(e^{Xt}e^{-\mu t}) = e^{-\mu t}M_X(t)$

Find the MGF for the Normal Distribution $X \sim \mathrm{N}(\mu, \sigma^2)$.

$$M_Z(t) = \int_{-\infty}^{\infty} \left(\sqrt{2\pi}\right)^{-1} e^{zt - z^2/2} dz = e^{t^2/2} \int_{-\infty}^{\infty} \left(\sqrt{2\pi}\right)^{-1} e^{-(z-t)^2/2} dz = e^{t^2/2}$$

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2} \text{ by property 5.}$$

Let $X \sim N(1, 2)$ and $Y \sim N(4, 7)$ be independent random variables. Find the probability of the event $3X + 4Y > 20$.

$$M_{3X+4Y}(t) = M_{3X}(t)M_{4Y}(t) = M_X(3t)M_Y(4t) = e^{19t+65t^2} = M_{N(19,130)}(t)$$

$$\mathbb{P}(3X + 4Y > 20) = \mathbb{P}\left(Z > \frac{20 - \sqrt{130}}{19}\right) = 1 - F_Z(0.45) = 0.3264.$$

The MGF of the sum of independent rv's equals the product of their MGF's. To turn the product form into a sum, we can take a log.

Cumulant generating functions:

1. $K_X(t) = \ln M_X(t) = \sum_{r=1}^{\infty} \kappa_r t^r/r!$,
   where $\kappa_r = K_X^{(r)}(0)$ is the rth cumulant of X # by Taylor's expansion
2. If X and Y are independent, and $Z = X + Y$, then $K_Z(t) = K_X(t) + K_Y(t)$
3. $\kappa_1 = \mathbb{E}(X)$ since $K_X'(t) = M_X'(t)/M_X(t)$, and $\kappa_2 = \mathbb{V}(X)$
4. $\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 = \mathbb{E}\left(\left(X - \mathbb{E}(X)\right)^3\right)$
   κ_3 is the skewness, which reflects level of symmetry of the distribution around its mean; to remove the scale effect, the coefficient of skewness is $\mathrm{skew}(X) = \kappa_3/\sigma^3$: negative for left-skew (has a long left tail), positive for right-skew, 0 for symmetric.
5. $\kappa_4 = \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4 = \mathbb{E}\left(\left(X - \mathbb{E}(X)\right)^4\right) - 3\sigma^4$
   κ_4 is the kurtosis. For a flatter and shorter-tailed distribution than $\mathrm{N}(\kappa_1, \kappa_2)$, the kurtosis is negative (e.g., rectangular distribution). The coefficient of kurtosis is $\mathrm{kurt}(X) = \kappa_4/\sigma^4$.

Laplace transform (for $x \geq 0$):

1. $L_X(t) = M_X(-t) = \mathbb{E}(e^{-tX})$, exist for all $t > 0$
2. $F_X(x) = \lim_{t\to\infty} \sum_{k\leq tx}((-t)^k/k!)L_X^{(k)}(t)$ can be uniquely determined for continuous rv X.

Let $X \sim \exp(1)$, compute $L_X(t)$ then recover the CDF of X.

$$L_X(t) = \int_0^\infty e^{-tx}e^{-x}dx = \frac{1}{t+1}, \quad t > 0$$

$$L_X^{(k)}(t) = (-1)^k k!\,(t+1)^{-(k+1)}, \quad t > 0$$

$$F_X(x) = \lim_{t\to\infty} \sum_{k\leq tx} \frac{(-t)^k}{k!}(-1)^k k!\,(t+1)^{-(k+1)} = \frac{1}{t+1}\left[\lim_{t\to\infty} \sum_{k=0}^{\lfloor tx\rfloor}\left(\frac{t}{t+1}\right)^k\right]$$

$$= \frac{1}{t+1}\left[\lim_{t\to\infty} \frac{1-\left(\frac{t}{t+1}\right)^{\lfloor tx\rfloor+1}}{1-\frac{t}{t+1}}\right] = 1 - \lim_{t\to\infty}\left(\frac{t}{t+1}\right)^{\lfloor tx\rfloor+1}$$

$$= 1 - \lim_{t\to\infty}\left(\left(1+\frac{1}{t}\right)^t\right)^{-\frac{\lfloor tx\rfloor+1}{t}} = 1 - e^{-x}, \quad x \geq 0.$$

The characteristic function is defined by $g_X(t) = \mathbb{E}\left(e^{itX}\right)$, where $i = \sqrt{-1}$.

Its advantage is that it is defined for all $t \in \mathbb{R}$ for all rv's as $\left|e^{itX}\right| = 1$. If X is a continuous rv with PDF f_X then, the characteristic function of X

$$g_X(t) = \int_{-\infty}^{\infty} e^{itX} f_X(x)dx$$

is the Fourier transform of the PDF f_X, which is inverse to itself. This inversion formula allows us to recover the distribution from the characteristic function.

Law of Large Numbers: Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (i.i.d.) rv's with $\mathbb{E}(X_i) = \mu$, and let $S_n = X_1 + X_2 + \cdots + X_n$. Then

$$\frac{S_n}{n} \approx \mu \quad \text{as} \quad n \to \infty,$$

with RHS interpreted as a rv which is constant with probability 1.

$$M_{\frac{S_n}{n}}(t) = \mathbb{E}\left(e^{\frac{tS_n}{n}}\right) = M_{S_n}\left(\frac{t}{n}\right) = M_{X_1+X_2+\cdots+X_n}\left(\frac{t}{n}\right) = \left(M_X\left(\frac{t}{n}\right)\right)^n$$

$$= \left(\sum_{k=0}^{\infty} \mu_k\left(\frac{t}{n}\right)^k/k!\right)^n = \left(1 + \frac{\mu t}{n} + \sum_{k=2}^{\infty}\mu_k\left(\frac{t}{n}\right)^k/k!\right)^n \approx \left(1+\frac{\mu t}{n}\right)^n$$

$$\to e^{\mu t} = M_\mu(t) \quad \text{as} \quad n \to \infty.$$

Central Limit Theorem: Let $X_1, X_2, \dots, X_n$ be i.i.d. rv's with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$, and let $S_n = X_1 + X_2 + \cdots + X_n$. Then the standardised random variable

$$Z_n = \frac{(S_n - n\mu)}{\sigma\sqrt{n}} \approx \mathrm{N}(0,1) \quad \text{as } n \to \infty.$$

In other words, for large n, $S_n \sim \mathrm{N}(n\mu, n\sigma^2)$ and $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$.

$$\begin{aligned} M_{Z_n}(t) = M_{\frac{(S_n - n\mu)}{\sigma\sqrt{n}}}(t) &= \mathbb{E}\left(e^{\frac{t(S_n - n\mu)}{\sigma\sqrt{n}}}\right) = M_{S_n - n\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) = M_{\sum_{i=1}^n X_i - \mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= \left(M_{X-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n = \left(\sum_{k=0}^{\infty} v_k \left(\frac{t}{\sigma\sqrt{n}}\right)^k / k!\right)^n \\ &= \left(1 + 0 + \frac{t^2}{2n} + \sum_{k=3}^{\infty} \mu_k \left(\frac{t}{\sigma\sqrt{n}}\right)^k / k!\right)^n \approx \left(1 + \frac{\frac{t^2}{2}}{n}\right)^n \to e^{\frac{t^2}{2}} \\ &\sim M_{\mathrm{N}(0,1)}(t) \quad \text{as } n \to \infty. \end{aligned}$$

A fair coin is tossed successively. Using the central limit theorem, find an approximation for the probability of obtaining at least 25 heads before 50 tails.

For $i = 1, 2, \dots, 50$, let X_i be the number of heads between the $(i-1)$th and the ith tails. Then $X \sim \mathrm{G}(0.5)$, CLT gives $S_n \sim (50, 100)$, and so

$$\begin{aligned} \mathbb{P}(\geq 25 \text{ heads before } 50 \text{ tails}) &= \mathbb{P}(S_n \geq 25) = \mathbb{P}\left(Z \geq \frac{25 - 50}{\sqrt{100}}\right) = 1 - F_Z(-2.5) \\ &= 0.9938. \end{aligned}$$

Poisson Limit Theorem: Let $X_1, X_2, \dots$ be independent Bernoulli random variables with $\mathbb{P}(X_i = 1) = p_i$. If $\lambda_n = p_1 + \cdots + p_n$, $W_n = X_1 + X_2 + \cdots + X_n$, then for each $x \in \mathbb{R}$,

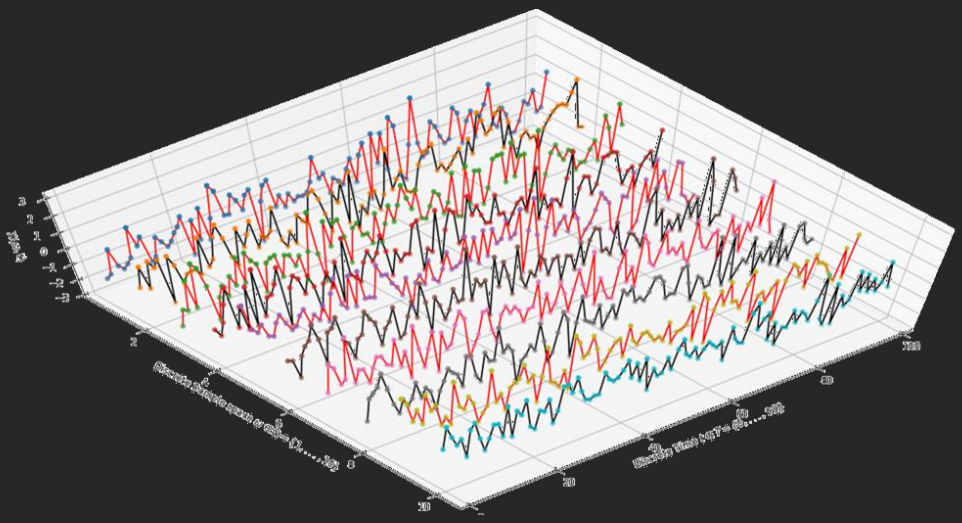
$$F_{W_n}(x) \approx F_{Y_n}(x),$$

where $Y_n \sim \mathrm{Pn}(\lambda_n)$.

A stochastic process is a sequence of random variables $X(t)$ where $t \in T$.

For each $t \in T$, $X(t)$ rakes values in a set S, called the state space of the stochastic process. The state space may be either discrete or continuous.

The set T is called the index set, with t usually denoting time. The index set may also be either discrete or continuous. It is a common practice to denote a discrete-time stochastic process by $\{X_n, n = 0, 1, \dots\}$, while using $\{X(t), t \in \cdots\}$ for a continuous-time stochastic process.



The discrete-state, continuous-time analogue of a sequence of Bernoulli trials is called the Poisson Process. For the Poisson Process, $X(t)$ is the number of "points" that occur in time $[0, t]$, the state space for the stochastic process $\{X(t)\}$ is given by $S = \mathbb{Z}_+$ and the index set is given by $T = \mathbb{R}_+$.

Discrete-time Markov Chain: A stochastic process with index set $T = \mathbb{Z}_+$ and a countable state space.

1. For all $n > 0$, all j, all $i_0, \dots, i_n$, we have $\mathbb{P}(X_{n+1} = j | X_0 = i_0 \cap \dots \cap X_n = i_n) = \mathbb{P}(X_{n+1} = j | X_n = i_n)$, where $\mathbb{P}(X_{n+1} = j | X_n = i_n)$ is called the transition probability.
2. Homogeneous if for all i, j, m, n, $\mathbb{P}(X_{n+m} = j | X_n = i) = \mathbb{P}(X_m = j | X_0 = i)$.

Transition matrix: For a homogeneous discrete-time Markov chain,

1. $P_{ij}^{(m)} = \mathbb{P}(X_{n+m} = j | X_n = i) \ \forall n$
2. $P_{ij} = P_{ij}^{(1)}$
3. $P^{(m)} = \left[P_{ij}^{(m)}\right]_{i,j \in S} = P^m$ # by induction
4. $P = P_1$
5. Every entry is non-negative, every row sums to 1.

Let $Y_1, Y_2, \ldots$ be the number of patients arriving at a hospital on days 1, 2, … and suppose the Y_k's are i.i.d. with $\mathbb{P}(Y_k = j) = 2^{-j-1}, j = 0, 1, 2, \ldots$

Let X_n be the total number of arrivals after n days, then $X_n = \sum_{1 \le i \le n} Y_i$ and $\{X_n : n = 0, 1, 2, \ldots\}$ $(X_0 = 0)$ is a Markov chain. Find the transition matrix P.

Hint. $\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(Y_{n+1} = j - i) = 2^{i-j-1}$.

Limiting Distribution:

1. $\overrightarrow{\pi_0}$ is the initial state distribution, $\overrightarrow{\pi_n}$ is the long-term distribution;
   $\overrightarrow{\pi_n} = (\mathbb{P}(X_n = 0), \mathbb{P}(X_n = 1), \ldots) = \overrightarrow{\pi_0} p^n$
2. Assume that the state space S is finite, and $P_{ij}^{(m)} \to \pi_j$ as $m \to \infty$, i.e., the limiting probability of state j will be approximately π_j; the vector $\vec{\pi} = (\pi_1, \pi_2, \ldots)$ is known as the equilibrium distribution.
3. $P_{ij}^{(m)} = \sum_k P_{ik}^{(m-1)} P_{kj} \;\Rightarrow\; \pi_j = \sum_k \pi_k P_{kj} \;\forall\, j \;\Rightarrow\; \vec{\pi} P = \vec{\pi}$;
   $\vec{\pi}$ is a left-eigenvector of P associated with the eigenvalue 1, and so $\vec{\pi}$ has a stationary interpretation. This could be solved with $\sum \pi_j = 1$.

Statistics is a study to estimate on the population distribution, based on observed data.

A random sample on rv X, is a sequence of i.i.d. rv's $X_1, X_2, \dots, X_n$.

A statistic $T = \phi(X_1, \dots, X_n)$ is a function (or a random variable) of the sample and its realization is denoted by $t = \phi(x_1, \dots, x_n)$.

Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$; Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Empirical (Sample) CDF: $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)$.

Empirical (Sample) PMF: $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x)$ for discrete underlying rv.

If the underlying rv is continuous, we prefer to obtain an approximation of the PDF.

1. Histogram, $\hat{f}_h$ (h is the bin length). First divide the entire range of values into a series of small intervals and then count how many values fall into each interval. For interval $[a, b)$, where $b - a = h$, draw a rectangle with height:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} I(a \le x_i \le b).$$

2. Smoothed PDF, $\hat{f}_h$ (h is the bandwidth parameter),

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right),$$

where $K(\cdot)$ is the kernel (a non-negative function that integrates to 1 and with mean 0) and h is a parameter that controls the level of smoothing.

Box plot is a summary of data from a single variable. The whiskers are $x_{(1)}$ and $x_{(n)}$.

Scatter plot is used for comparing data from two variables.

Quantile-quantile (QQ) plots compare the similarity of data against a theoretical distribution. Based on 'Type 6' quantile, $\hat{\pi}_p = x_{(k)}$ where $p = k/(n+1)$. Plot the points

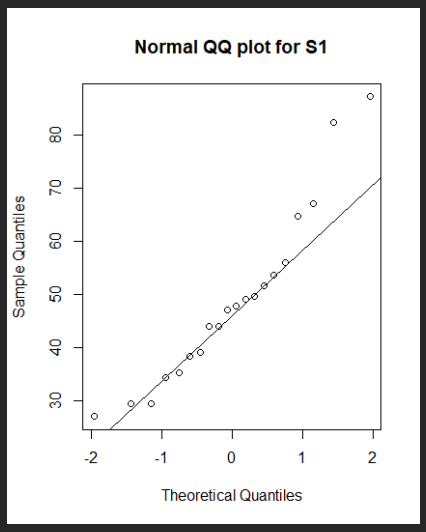
$$\left(\hat{\pi}_p, \pi_p\right) = \left(x_{(k)}, F^{-1}\left(k/(n+1)\right)\right), \quad k = 1, \dots, n.$$

If the normal model is correct, $x_{(k)} \approx \mu + \sigma\Phi^{-1}\left(k/(n+1)\right)$. The normal QQ plot:

$$\left(x_{(k)}, \Phi^{-1}\left(k/(n+1)\right)\right), \quad k = 1, \dots, n,$$

the result should be a straight line with intercept μ and slope σ. The values $\Phi^{-1}\left(k/(n+1)\right)$ are called normal scores.

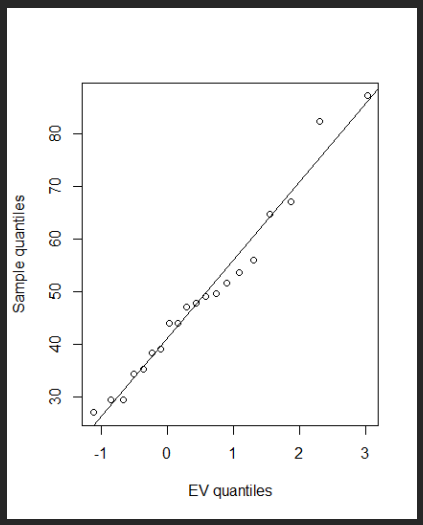
```
qqnorm(s1, main = "Normal QQ plot for S1")
qqline(s1)
```



Although the central part of the data distribution is compatible with the normality assumption, note that the right tail deviates from the straight line.

Probability theory suggests that a better model for maxima is the Extreme Value (EV) Distribution with inverse CDF $\mu + \sigma F^{-1}(p)$, where $F^{-1}(p) = -\log(-\log(p))$ is the inverse CDF of the standard EV distribution.

```
Finv <- function(p) {-log(-log(p))}  # quantile function
p <- (1:20) / 21
y <- sort(s1)  # order statistics
x <- Finv(p)   # theoretical quantiles
plot(x, y, ylab = "Sample quantiles", xlab = "EV quantiles")
fit <- lm(y ~ x)  # linear model computes the "line of best fit"
abline(fit)       # plots the "line of best fit"
```



From the last QQ plot the EV model seems to be more appropriate than the normal model, since the points in EV QQ plot are a little closer to the straight line compared to the previous normal QQ plot.

The sampling distribution of a statistic is its probability distribution across an arbitrarily large number of samples, each involving multiple observations, given an assumed population distribution and a sampling scheme (e.g., random sampling).

A point estimator is a statistic that is used to best estimate a parameter. An estimate is the observed value of the estimator for a given dataset. 'Hat' notation: If T is an estimator for θ, then we usually refer to it by $\hat{\theta}$ for convenience. Examples:

1. By CLT, $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{V}(\bar{X}) = \sigma^2/n$ for large n ;
2. $\mathbb{E}(S^2) = \sigma^2$; # for normal distribution, $\mathbb{V}(S^2) = 2\sigma^4/(n-1)$

   Since $\sigma^2 = \mathbb{E}(X^2) - \mu^2$, we see that $\mathbb{E}(X^2) = \sigma^2 + \mu^2$. A similar argument shows that $\mathbb{E}(\bar{X}^2) = \sigma^2/n + \mu^2$. Then

$$\mathbb{E}(S^2) = \frac{n}{n-1}\{\mathbb{E}(X_i^2) - \mathbb{E}(\bar{X}^2)\} = \frac{n}{n-1}\left\{(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)\right\} = \sigma^2.$$

3. For a discrete rv, let the population proportion be $p = \mathbb{P}(X = a)$, sample frequency is $\text{freq}(a) = \sum_{i=1}^{n} I(X_i = a) \sim \text{Bi}(n, p)$. For large n, the sample proportion is approximated as $\hat{p} = \text{freq}(a)/n \approx \text{N}(p, p(1-p)/n)$.

The bias of the estimator is $\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$. Unbiased if $\mathbb{E}(\hat{\theta}) = \theta$.

Is the sample standard deviation biased for the population standard deviation?

$$\mathbb{V}(S) = \mathbb{E}(S^2) - \mathbb{E}(S)^2 = \sigma^2 - \mathbb{E}(S)^2 \geq 0 \quad \Rightarrow \quad \mathbb{E}(S) \neq \sigma.$$

Compare two estimators in terms of their mean square errors (MSE):

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}\left((\hat{\theta} - \theta)^2\right) = \mathbb{E}\left((\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right) \\ &= \mathbb{E}\left((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right) + \mathbb{E}\left((\mathbb{E}(\hat{\theta}) - \theta)^2\right) + 2\mathbb{E}\left((\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)\right) \\ &= \mathbb{V}(\hat{\theta}) + \{\text{Bias}(\hat{\theta})\}^2.\end{aligned}$$

Method of moments estimator (MME) makes the population distribution resemble the empirical distribution by equating theoretical moments with sample moments. Can use the variance instead of the second moment.

Usually biased and usually not optimal.

Method of maximum likelihood estimator (MLE) finds the most likely explanation for the data, that is, find parameter values that maximize the probability of the data.

Regard the random sample $X_1, \dots, X_n$, the likelihood function with parameter $\theta_1, \dots, \theta_m$ and data $x_1, \dots, x_n$ is defined as

$$L(\theta_1, \dots, \theta_m) = \prod_{i=1}^{n} f(x_i | \theta_1, \dots, \theta_m)$$

If X is discrete, for f use the PMF.

Often (but not always) useful to take logs (called log-likelihood) and then differentiate and equate derivatives to zero to find MLE's. The final answer (the maximizing value of p) is the same, since the log of non-negative numbers is a one-to-one function whose inverse is the exponential.

Sampling from $X \sim \mathrm{N}(\theta_1, \theta_2)$.

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x_i - \theta_1)^2}{2\theta_2}\right),$$

$$\ell(\theta_1, \theta_2) = \ln L(\theta_1, \theta_2) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2.$$

Take partial derivative with respect to θ_1 and θ_2:

$$\begin{cases} \dfrac{\partial}{\partial\theta_1}[\ln L(\theta_1, \theta_2)] = \dfrac{1}{\theta_2}\displaystyle\sum_{i=1}^{n}(x_i - \theta_1) \\ \dfrac{\partial}{\partial\theta_2}[\ln L(\theta_1, \theta_2)] = -\dfrac{n}{2\theta_2} + \dfrac{1}{2\theta_2^2}\displaystyle\sum_{i=1}^{n}(x_i - \theta_1)^2 \end{cases}$$

Set both to zero and solve. This gives estimate: $\theta_1 = \bar{x}$ and $\theta_2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The MLE's are therefore: $\widehat{\theta_1} = \bar{X}$, $\widehat{\theta_2} = (n-1)S^2/n$.

Suppose we know $\hat{\theta}$ but are actually interested in $\phi = g(\theta)$ rather than θ itself. We can use $\hat{\phi} = g(\hat{\theta})$. This is known as the invariance property of the MLE. The consequence is that MLEs are usually biased since expectations are not invariant under transformations.

An interval estimate is a pair of statistics (L, U) defining a random interval that aims to convey an estimate of a parameter with uncertainty.

Standard deviation of the estimator $\mathrm{sd}(\widehat{\Theta}) = \sqrt{\mathbb{V}(\widehat{\Theta})}$ tells us a typical amount by which the estimate will vary from one sample to another, and thus (for an unbiased estimator) how close to the true parameter value it is likely to be.

We estimate $\mathrm{sd}(\widehat{\Theta})$ by substituting point estimates into the expression for the variance. We call this estimate "the standard error of the estimate for θ", write $\mathrm{se}(\hat{\theta})$.

The form of "est $\pm$ error" is an example of interval estimate.

Take a sample of size $n = 100$ and observe $\mathrm{freq}(a) = 30$, then we get the sample proportion $\hat{p} = 30/100 = 0.3$. Since $\hat{p} \sim \mathrm{N}(p, p(1-p)/n)$,

$$\mathrm{se}(\hat{p}) = \widehat{\mathrm{sd}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.3 \times (1-0.3)}{30}} = 0.046$$

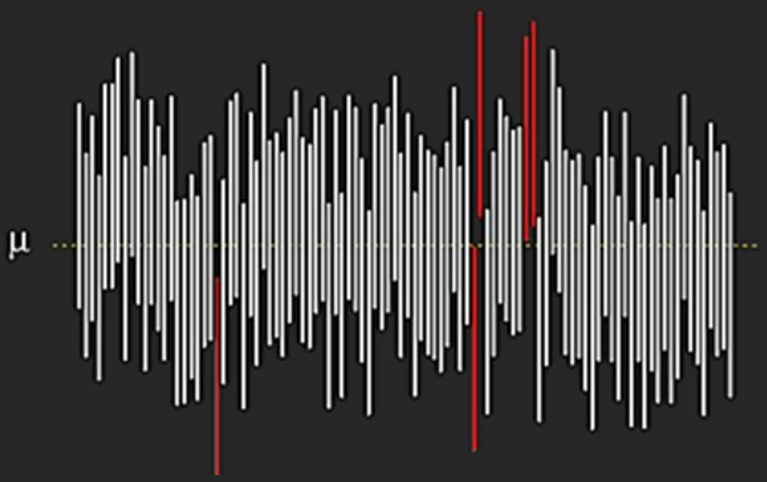
The population proportion of a is estimated to be 0.3 ± 0.046 .

A confidence interval (CI) is the most common type of interval estimate.

Under repeated sampling, the corresponding interval estimator has a probability, known as the confidence level $(1-\alpha)$, that would contain the true value of the parameter: $\mathbb{P}(L < \theta < U) = 1 - \alpha$.

Note that the random elements are the endpoints, not the parameter.

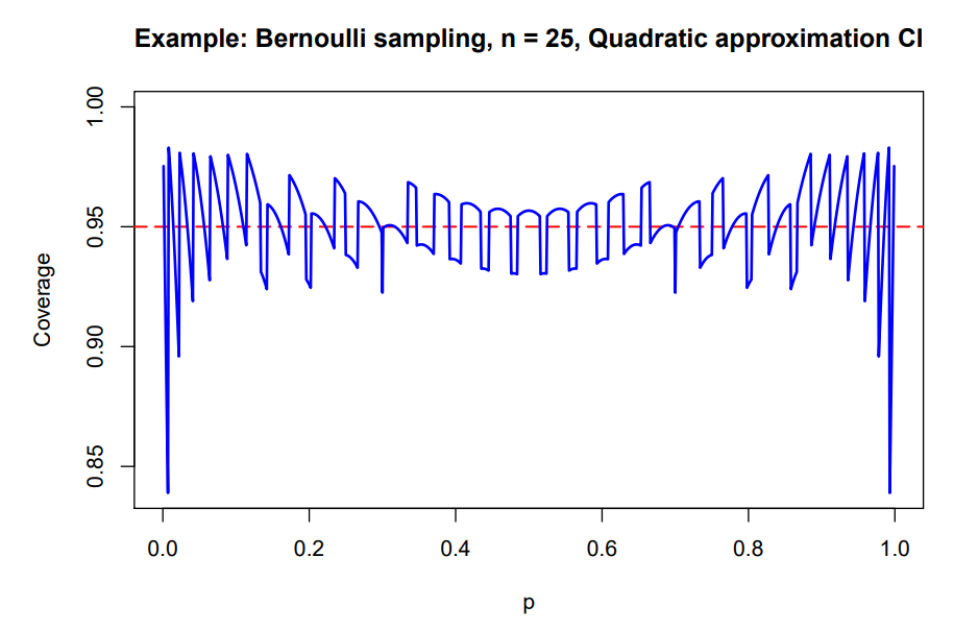
After the sample is taken, we have a realised interval. It no longer has a probabilistic interpretation; it either contains or it doesn't.



CI based on discrete statistics: If cannot guarantee an exact confidence level, instead should aim for an "at least ($\geq$)" probability.

The coverage or coverage probability of a CI is the probability it really contains the true value of the parameter. Usually this is equal to the confidence level, which is also known as the nominal coverage probability.

However, due to various approximations we use, the actual coverage achieved may vary from the confidence level, e.g., the quadratic approximation.



General CI techniques: Start with an estimator T whose sampling distribution is known. Let $a = \pi_{\alpha/2}$ and $b = \pi_{1-\alpha/2}$, write the central probability interval as

$$\mathbb{P}(a < T < b) = 1 - \alpha.$$

Take a random sample of size n from an exponential distribution with rate parameter λ. Derive an exact 95% CI for λ.

$$n\bar{X} = \sum_{i=1}^{n} X_i \sim \gamma(n, \lambda) \quad \Rightarrow \quad \lambda n\bar{X} \sim \gamma(n, 1)$$

$$\mathbb{P}\left(F^{-1}(0.025) < \lambda n\bar{X} < F^{-1}(0.975)\right) = 0.95$$

CI from CLT approximation: If n is large enough, CLT gives that $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \approx \mathrm{N}(0, 1)$. If we want a $100 \cdot (1 - \alpha)\%$ CI, write the central probability interval as

$$\mathbb{P}\left(-c < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < c\right) = 1 - a \quad \text{where} \quad c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Chi-squared Distribution: $X^2 \sim \chi_k^2 \sim Z_1^2 + \cdots + Z_k^2$. $\quad \# \chi_{2n}^2 \sim \gamma(n, 0.5),\ n \in \mathbb{Z}^{0+}$

1. para — degrees of freedom $:= k > 0$
2. shape — bounded below by zero and is right-skewed
3. pdf — $f_T(t) = t^{k/2-1}e^{-t/2} \,/\, 2^{k/2}\,\Gamma\left(\frac{k}{2}\right),\ \ t \geq 0$
4. mean — k
5. var — $2k$
6. mgf — $(1-2t)^{-k/2}$

When sampling from a normal distribution, the sample variance

$$S^2 = \frac{\sigma^2}{n-1}\left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}\right) = \frac{\sigma^2}{n-1}\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} - \frac{n(\bar{X}-\mu)^2}{\sigma^2}\right)$$
$$= \frac{\sigma^2}{n-1}\left(\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2\right) \sim \frac{\sigma^2}{n-1}(\chi_n^2 - \chi_1^2) \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2.$$

Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(0, \theta^2)$. Consider the estimators S^2 and $\hat{\theta}^2 = n^{-1}\sum_{i=1}^n X_i^2$. Show that $\hat{\theta}^2$ is a better estimator than S^2.

We have already known that S^2 is unbiased. And also

$$\mathbb{E}(\hat{\theta}^2) = \mathbb{E}\left(n^{-1}\sum_{i=1}^n X_i^2\right) = n^{-1}\sum_{i=1}^n \mathbb{E}(X_i^2) = n^{-1}n(\mathbb{E}(X_i^2) - E(X_i)^2) = \mathbb{V}(X) = \theta^2,$$

meaning that it is also unbiased. So we compare the variance of the two estimators.

To derive the variance of the estimator, first note that $\mathbb{V}(X_i^2) = \mathbb{E}(X_i^4) - \mathbb{E}(X_i^2)^2 = \mathbb{E}(X_i^4) - \theta^4$. The 4th moment for a normal distribution is $M_X^{(4)}(0) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 = 3\theta^4$. Hence, we have $\mathbb{V}(X_i^2) = 2\theta^4$.

Now derive the variance of the estimator $\hat{\theta}^2 = n^{-1}\sum_{i=1}^n X_i^2$,

$$\mathbb{V}\left(n^{-1}\sum_{i=1}^n X_i^2\right) = n^{-2}n\,\mathbb{V}(X^2) = \frac{2\theta^4}{n}.$$

Also, we know that,

$$\mathbb{V}(S^2) = \mathbb{V}\left(\frac{\sigma^2}{n-1}\chi_{n-1}^2\right) = \frac{\sigma^4}{(n-1)^2}\mathbb{V}(\chi_{n-1}^2) = \frac{\sigma^4 \cdot 2(n-1)}{(n-1)^2} = \frac{2\theta^4}{n-1},$$

which would give more uncertainty.

Student's T Distribution: $T \sim t_k \sim Z/\sqrt{X_k^2/k}.$ # $t_k^2 \sim F_{1,k}$

1. para — degrees of freedom $:= k > 0$
2. shape — symmetric, similar to normal but with wide tails, asymptotically $\to \mathrm{N}(0,1)$
3. pdf — $f_T(t) = \left(\Gamma\left(\frac{k+1}{2}\right)/\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)\right) \times \left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < t < \infty$
4. mean — 0 if $k > 1$
5. var — $k/(k-2)$ if $k > 2$

If $X_1, \dots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, both μ and σ^2 are unknown, then the statistic

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Fisher-Snedecor Distribution (F distribution): $W \sim F_{m,n} \sim (X_m^2/m)/(X_n^2/n).$

1. para — $m, n > 0$, degrees of freedom $:= k > 0$
2. pdf — $f_W(w) = \left(\beta\left(\frac{m}{2}, \frac{n}{2}\right)\right)^{-1} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1+\left(\frac{m}{n}\right)w\right)^{-(m+n)/2}$

CI from pivots: Find a pivot, i.e., a function of the data and parameters, $Q(X_1, \dots, X_n; \theta)$, whose distribution does not depend on the values of parameters.

Pivots are usually not a statistic; if a pivot is a statistic, we call it an ancillary statistic.

Pivots for normal distribution: Take random samples from each population $X_1, \dots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \dots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$.

1. Inference for a single mean, known σ

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

2. Inference for a single mean, unknown σ

   $$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

3. Comparison of two means, known σ # different in sample size?

   $$\bar{X}-\bar{Y} \sim \mathrm{N}\left(\mu_X-\mu_Y, \frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}\right) \Rightarrow Z = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim \mathrm{N}(0,1)$$

4. Inference for the mean of difference, paired samples # $(X_1, Y_1), \dots, (X_n, Y_n)$

   $X_i - Y_i \sim \mathrm{N}(\mu_D, \sigma_D^2)$ where $\mu_D = \mu_X - \mu_Y$ and $\sigma_D^2 \neq \sigma_X^2 + \sigma_Y^2$ # dependent?

   $$T = \frac{\bar{D} - (\mu_D)}{S_D/\sqrt{n}} \sim t_{n-1}$$ # a single mean, unknown σ

5. Comparison of two means, unknown σ, many samples # estimate σ by S

   $$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim \mathrm{N}(0,1)$$

6. Comparison of two means, unknown common σ

   $$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathrm{N}(0,1)\,, \qquad U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

   $$T = \frac{Z}{\sqrt{\frac{U}{n+m-2}}} \sim t_{n+m-2} \sim \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}}$$ where

   $$S_P = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$$ # the pooled estimate of the common variance

7. Comparison of two means, unknown different σ

   $$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t_r$$ where # Welch's approximation

   $$r = \left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2 \Big/ \left(\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}\right)$$

8. Inference for a single variance

   $$T = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

9. Comparison of two variances

   $$F = \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} = \frac{\left(\frac{(m-1)S_Y^2}{\sigma_Y^2}\right)/(m-1)}{\left(\frac{(n-1)S_X^2}{\sigma_X^2}\right)/(n-1)} \sim F_{m-1,n-1}$$

Pivots for proportions: Let $p = \mathbb{P}(X = a)$, the CI be $(1-\alpha)$, and $c = \pi_{1-\alpha/2}$

1. Single proportion # from CLT, sometimes called Wald approximation

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx \mathrm{N}(0,1)$$

2. Double proportion, compare proportions between two different samples

$$Z = \left(\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)\right) / \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \mathrm{N}(0,1)$$

Quadratic approximation for proportion: Using the CLT approximation for $\hat{p}$,

$$\mathbb{P}\left(\hat{p} - c\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + c\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha.$$

$$\mathbb{P}\left((p-\hat{p})^2 < c^2\frac{p(1-p)}{n}\right) \approx 1 - \alpha$$

$$\mathbb{P}\left(p^2 + \hat{p}^2 - 2p\hat{p} < \frac{c^2 p}{n} - \frac{c^2 p^2}{n}\right) \approx 1 - \alpha$$

$$\mathbb{P}\left(\left(1 + \frac{c^2}{n}\right)p^2 - 2\left(\hat{p} + \frac{c^2}{2n}\right)p + \hat{p}^2 < 0\right) \approx 1 - \alpha \qquad \text{\# quadratic in } p$$

The solution to this inequality is an interval with the following endpoints:

$$\left(\hat{p} + \frac{c^2}{2n} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{c^2}{4n^2}}\right) / \left(1 + \frac{c^2}{n}\right)$$

The determination of sample size depends on how much precision is required, often measured by the desired width ϵ of a CI.

The unemployment rate has been 8% for a while. A researcher wishes to take new sample to estimate it and wants to be 'very certain', by using a 99% CI, that the new estimate is within 0.001 of true proportion.

$$\mathbb{P}\left(\hat{p} - c\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + c\sqrt{\hat{p}(1-\hat{p})/n}\right) = 1 - \alpha \;\Rightarrow\; c\sqrt{\hat{p}(1-\hat{p})/n} = 0.001$$

$$n = \frac{c^2\hat{p}(1-\hat{p})}{0.001^2} = 10^6 \times \left(\Phi^{-1}(0.995)\right)^2 \times 0.08 \times 0.92 = 488393$$

Prediction is to estimate the value of a future observation X^*, rather than a parameter of the distribution. In the prediction interval (PI) estimator, all quantities are random variables: $\mathbb{P}(L < X^* < U) = 1 - \alpha$.

For random sample $X_1, \dots, X_n$ on $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$\bar{X} - X^* \sim \mathrm{N}\left(0, \left(\frac{1}{n} + 1\right)\sigma^2\right) \;\Rightarrow\; \mathbb{P}\left(\bar{X} - c\sigma\sqrt{\left(\frac{1}{n} + 1\right)} < X^* < \bar{X} + c\sigma\sqrt{\left(\frac{1}{n} + 1\right)}\right) = 1 - \alpha$$

As $n \to \infty$, the PI tends to be between $\mu \pm c$, which is wider than CI's.

A hypothesis is a statement about the population distribution.

A null hypothesis is a hypothesis that specifies "no effect" or "no change", usually denoted H_0. An alternative hypothesis is a hypothesis that specifies the effect of interest, usually denoted H_1. The alternative can be either one-sided, or two-sided.

A hypothesis test (also called statistical test) is a decision rule for deciding between H_0 and H_1. A test statistic, T, is a statistic on which the test is based.

The significance level (or size) of a test is

$$\alpha = \mathbb{P}(\text{Type I error/false positive}) = \mathbb{P}(\text{reject } H_0 |\, H_0 \text{ true}).$$

The power of a test is

$$\begin{aligned} K(\theta) = 1 - \beta &= 1 - \mathbb{P}(\text{Type II error/false negative}) = 1 - \mathbb{P}(\text{fail to reject } H_0 |\, H_0 \text{ false}) \\ &= \mathbb{P}(\text{reject } H_0 |\, \theta). \end{aligned}$$

The classical theory (Neyman & Pearson), using critical values:

1. Assume H_0 is true, find the realised value of test statistic z_{obs}.
2. Find the critical value c(s) associated with the significance level, α. The alternative hypothesis will specify the rejection region A on the distribution as well as its side option.
3. Reject H_0 if $z_{\text{obs}} \in A$.

The hypothesis testing based on CI: # just discussed

1. Calculate a $100(1 - \alpha)\%$ CI for the parameter, from samples.
2. Reject H_0 if p_0 is not in the interval.

Fisher's significance testing, using the p-value:

1. Find the realised value of test statistic z_{obs} like before.
2. Calculate a p-value, i.e., the probability of observing data that is as or more extreme than the observed statistic, under H_0.

3. When we have a two-sided alternative hypothesis, we simply double the relevant tail probability, even if the two tails are found asymmetric.
4. Reject H_0 if the p-value is less than the significance level.

Modern hypothesis testing:

1. Largely use the terminology and formulation of the classical theory but commonly report the results using a p-value and talk about "not rejecting" rather than "accepting" the null.

The hypothesis testing in R code:

```
## Single proportion: test p > p0
> p1 = prop.test(x = 29, n = 40, p = 0.5, conf.level = 0.95, alternative =
+         "greater", correct = TRUE)  # using quadratic approximation
> p1

    1-sample proportions test with continuity correction

data: 29 out of 40, null probability 0.5
X-squared = 7.225, df = 1, p-value = 0.003595
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5843256 1.000000
sample estimates:
    p
0.725

> sqrt(p1$statistic)  # z_obs > 1.96
2.687936

> 1 - pnorm(2.688)  # p-value
[1] 0.00359407
```

$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

```
## Single proportion, exact test, small sample size: test p > p0
> binom.test(29, 40, alternative = "greater")  # default p is 0.5

    Exact binomial test

data: 29 and 40
number of successes = 29, number of trials = 40, p-value = 0.003213
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5861226 1.0000000
sample estimates:
probability of success
                 0.725
```

```
## Two proportions: test pX > pY
> x = c(459, 425)
> n = c(500, 500)
> p2 = prop.test(x, n, alternative = "greater")
```

```
> p2

    2-sample test for equality of proportions with continuity correction

data: x out of n
X-squared = 10.62, df = 1, p-value = 0.0005594
alternative hypothesis: greater
95 percent confidence interval:
 0.03287541 1.00000000
sample estimates:
prop 1  prop 2
 0.918   0.850
```

$$Z = \frac{Y_1/n - Y_2/m}{\sqrt{p_{12}(1-p_{12})(1/n + 1/m)}}$$

```
## Normal, two means, unknown pooled variance: test muX < muY
> x = c(0.8, 1.8, 1.0, 0.1, 0.9, 1.7, 1.0, 1.4, 0.9, 1.2, 0.5)
> y = c(1, 0.8, 1.6, 2.6, 1.3, 1.1, 2.4, 1.8, 2.5, 1.4, 1.9, 2, 1.2)
> t.test(x, y, alternative = "less", var.equal = TRUE)  # using t test

    Two Sample t-te

data: x and y
t = -2.8112, df = 22, p-value = 0.005086
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.2468474
sample estimates:
mean of x   mean of y
 1.027273    1.661538
```

$$T = \frac{\bar{X} - \bar{Y}}{S_P\sqrt{1/n + 1/m}}$$

```
## Normal, two variances: test varX ≠ varY
> head(fish, 4)

  Length Group
1   5.20     X
2   4.70     X
3   5.75     X
4.  7.50     X

> var.test(Length ~ Group, data = fish)

    F test to compare two variances

data: Length by Group
F = 3.2054, num df = 29, denom df = 29, p-value = 0.002458
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.525637 6.734441
sample estimates:
ratio of variances
          3.205357
```

$$F = S_Y^2/S_X^2$$

Given that X and Y are paired. Often want to predict on Y based on observations of x, where Y_i's are independent but may not be identically distributed.

The simple linear regression of response variable Y on the predictor variable x gives the predictor function (conditional mean) as:

$$\mathbb{E}(Y|x) = \mu(x) = \alpha + \beta x \qquad \text{\# also called regression curve, model equation}$$

where the variance of the error is $\mathbb{V}(Y|x) = \sigma^2$.

The parameters α and β are called regression coefficients. A regression model is called linear if it is a linear combination of the coefficients and the predictor variables.

By parameterization, let $\alpha_0 = \alpha + \beta\bar{x}$, and thus $\mu(x) = \alpha_0 + \beta(x - \bar{x})$.

Method of ordinary least squares estimator (OLS) gives the point estimation of the mean. Choose α_0 and β to minimise the sum of squared deviation:

$$H(\alpha_0, \beta) = \sum_{i=1}^{n} \big(y_i - \alpha_0 - \beta(x_i - \bar{x})\big)^2.$$

Solve this by finding the partial derivatives and setting to zero. Then get

$$0 = \frac{\partial}{\partial \alpha_0}[H(\alpha_0, \beta)] = 2\sum_{i=1}^{n} \big(y_i - \alpha_0 - \beta(x_i - \bar{x})\big)(-1) \;\Rightarrow\; \hat{\alpha}_0 = \bar{Y};$$

$$0 = \frac{\partial}{\partial \beta}[H(\alpha_0, \beta)] = 2\sum_{i=1}^{n} \big(y_i - \alpha_0 - \beta(x_i - \bar{x})\big)\big(-(x_i - \bar{x})\big)$$

$$\Rightarrow\; \hat{\beta} = \left(\sum_{i=1}^{n} (x_i - \bar{x})Y_i\right)/K = \left(\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})\right)/K = S_{XY}/S_X^2$$

$$\text{where } K = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)S_X^2;$$

$$\hat{\alpha} = \hat{\alpha}_0 - \hat{\beta}\bar{x} = \bar{Y} - \hat{\beta}\bar{x} \;\Rightarrow\; \hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x = \bar{Y} + \hat{\beta}(x - \bar{x}).$$

These estimators of coefficients are all linear combinations of Y_i's, which allows us to easily calculate means and variances.

$$\mathbb{E}(\hat{\alpha}_0) = \mathbb{E}(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \alpha_0 + \beta(x_i - \bar{x}) = \alpha_0;$$

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}\left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} Y_i\right) = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{K}\mathbb{E}(Y_i)$$

$$= K^{-1}\sum_{i=1}^{n} (x_i - \bar{x})\big(\alpha_0 + \beta(x_i - \bar{x})\big) = K^{-1}\sum_{i=1}^{n} \beta(x_i - \bar{x})^2 = \beta;$$

$$\mathbb{E}\big(\hat{\mu}(x)\big) = \alpha_0 + \beta(x - \bar{x}) = \mu(x); \qquad \text{\# unbiased}$$

$$\mathbb{V}(\hat{\alpha}_0) = \mathbb{V}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(Y_i) = \frac{\sigma^2}{n} \to 0 \quad \text{as } n \to \infty ;$$

$$\mathbb{V}(\hat{\beta}) = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{K}\right)^2 \mathbb{V}(Y_i) = K^{-2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2 = \frac{\sigma^2}{K} = \frac{\frac{1}{n-1}\sigma^2}{S_X^2} \to 0 ;$$

$$\mathbb{V}(\hat{\alpha}) = \left(\frac{1}{n} + \frac{\bar{x}^2}{K}\right)\sigma^2 ;$$

$$\begin{aligned}\mathrm{Cov}(\hat{\alpha}_0, \hat{\beta}) &= \mathrm{Cov}\left(\sum_{i=1}^{n}\frac{1}{n}Y_i, \sum_{j=1}^{n}\frac{x_i - \bar{x}}{K}Y_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(\frac{1}{n}Y_i, \frac{x_i - \bar{x}}{K}Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{x_i - \bar{x}}{nK}\mathrm{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n}\frac{x_i - \bar{x}}{nK}\mathbb{V}(Y_i) + \sum_{i\neq j}\frac{x_i - \bar{x}}{nK}\cdot 0 = 0 \quad \Rightarrow \quad \text{Independence} ;\end{aligned}$$

$$\mathbb{V}(\hat{\mu}(x)) = \frac{\sigma^2}{n} + \frac{\sigma^2}{K}(x - \bar{x})^2 = \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{K}\right)\sigma^2.$$

ANOVA decomposition on the simple linear model:

$$\begin{aligned}\sum_{i=1}^{n}\big(Y_i - \mathbb{E}(Y_i|x_i)\big)^2 &= \sum_{i=1}^{n}\big(Y_i - \alpha_0 - \beta(x_i - \bar{x})\big)^2 \qquad \text{\# sum of sq deviation} \\ &= \sum_{i=1}^{n}\Big(\big(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x})\big) + (\hat{\alpha}_0 - \alpha_0) + \big(\hat{\beta} - \beta\big)(x_i - \bar{x})\Big)^2 \\ &= \sum_{i=1}^{n}\big(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x})\big)^2 + n(\hat{\alpha}_0 - \alpha_0)^2 + K\big(\hat{\beta} - \beta\big)^2 + 2(t_1 + t_2 + t_3)\end{aligned}$$

where

$$\begin{aligned}t_1 &= \sum_{i=1}^{n}\big(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x})\big)(\hat{\alpha}_0 - \alpha_0) \\ &= (\hat{\alpha}_0 - \alpha_0)\sum_{i=1}^{n}(Y_i - \hat{\alpha}_0) - (\hat{\alpha}_0 - \alpha_0)\hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= (\hat{\alpha}_0 - \alpha_0)\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0 ,\end{aligned}$$

$$\begin{aligned}t_2 &= \sum_{i=1}^{n}\big(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x})\big)\big(\hat{\beta} - \beta\big)(x_i - \bar{x}) \\ &= \big(\hat{\beta} - \beta\big)\Big(\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2\Big) = 0 ,\end{aligned}$$

$$t_3 = \sum_{i=1}^{n}(\hat{\alpha}_0 - \alpha_0)\big(\hat{\beta} - \beta\big)(x_i - \bar{x}) = (\hat{\alpha}_0 - \alpha_0)\big(\hat{\beta} - \beta\big)\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Let the inferred mean for each observation be called its fitted value, $\hat{Y}_i = \hat{\alpha}_0 + \hat{\beta}(x_i - \bar{x})$. The deviation from each fitted value is called a residual, $R_i = Y_i - \hat{Y}_i$.

Taking expectations of the ANOVA formula, gives:

$$n\sigma^2 = \mathbb{E}(D^2) + \sigma^2 + \sigma^2 \;\Rightarrow\; \mathbb{E}(D^2) = (n-2)\sigma^2 \;\Rightarrow\; \hat{\sigma}^2 = \frac{D^2}{n-2} \qquad \text{\# unbiased}$$

where $\hat{\sigma}^2$ is also called the residual variance, and the sum of squared residuals is

$$\begin{aligned} D^2 = \sum_{i=1}^{n} R_i^2 &= \sum_{i=1}^{n} \left(Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x})\right)^2 = \sum_{i=1}^{n} \left((Y_i - \bar{Y}) - \hat{\beta}(x_i - \bar{x})\right)^2 \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \hat{\beta} \sum_{i=1}^{n} \hat{\beta}(x_i - \bar{x})^2 - 2\hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \\ &\quad - 2\hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \end{aligned}$$

Later, we call D^2 the squared sum of error.

Further discussion requires assumptions about the population distribution. From now on, let's assume a normally distributed random sample:

$$Y_i \sim \mathrm{N}(\mu(x_i), \sigma^2) \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2) \sim \mathrm{N}(\alpha_0 + \beta(x_i - \bar{x}), \sigma^2)$$
$$\Rightarrow\; Y_i - \alpha - \beta x_i = \epsilon_i \sim \mathrm{N}(0, \sigma^2). \qquad \text{\# error}$$

With the assumed distribution, method of maximum likelihood estimation gives:

$$L(\alpha_0, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right\}$$

$$\ell(\alpha_0, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2} H(\alpha_0, \beta)$$

$$\Rightarrow \begin{cases} 0 = \dfrac{\partial}{\partial \alpha_0}[\ell(\alpha, \beta, \sigma^2)] = -\dfrac{1}{2\sigma^2} \cdot \dfrac{\partial}{\partial \alpha_0}[H(\alpha_0, \beta)] \\ 0 = \dfrac{\partial}{\partial \beta}[\ell(\alpha, \beta, \sigma^2)] = -\dfrac{1}{2\sigma^2} \cdot \dfrac{\partial}{\partial \alpha_0}[H(\alpha_0, \beta)] \end{cases} \qquad \text{\# same as OLS}$$

$$0 = \frac{\partial}{\partial \sigma^2}[\ell(\alpha, \beta, \sigma^2)] = -\frac{n}{\sigma} + \sigma^{-3} \frac{\partial}{\partial \alpha_0}[H(\alpha_0, \beta)] \;\Rightarrow\; \hat{\sigma}^2_{\mathrm{MLE}} = \frac{D^2}{n} \qquad \text{\# biased}$$

Except for σ^2, our estimators are just linear combinations of the Y_i so will also have normal distributions.

1. $\hat{\alpha}_0 = \sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i \sim \mathrm{N}\left(\sum_{i=1}^{n}\left(\frac{1}{n}\right)(\alpha_0+\beta(x_i-\bar{x})), \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2\right) \sim \mathrm{N}\left(\alpha_0, \frac{\sigma^2}{n}\right)$;
2. $\hat{\beta} = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{K}\right)Y_i \sim N\left(\beta, \frac{\sigma^2}{K}\right)$;
3. $\hat{\alpha} = \hat{\alpha}_0 - \hat{\beta}\bar{x} \sim \mathrm{N}\left(\alpha, \left(\frac{1}{n}+\frac{\bar{x}^2}{K}\right)\sigma^2\right)$;
4. $\hat{\mu}(x) = \hat{\alpha}_0 + \hat{\beta}(x-\bar{x}) \sim \mathrm{N}\left(\mu(x), \left(\frac{1}{n}+\frac{(x-\bar{x})^2}{K}\right)\sigma^2\right)$. # do not use α here

Using the ANOVA formula, divide σ^2 on both sides:

$$\sum_{i=1}^{n}\left(\frac{Y_i-\mathbb{E}(Y_i)}{\sigma}\right)^2 = \frac{D^2}{\sigma^2} + \left(\frac{\hat{\alpha}_0-\alpha_0}{\sigma/\sqrt{n}}\right)^2 + \left(\frac{\hat{\beta}-\beta}{K/\sqrt{n}}\right)^2 \Rightarrow \frac{D^2}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Therefore, we can define pivots for the various mean parameters.

$$\frac{\frac{\hat{\alpha}_0-\alpha_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{\hat{\alpha}_0-\alpha_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-2} \sim \frac{\hat{\beta}-\beta}{\frac{\hat{\sigma}}{\sqrt{K}}} \sim \frac{\hat{\mu}(x)-\mu(x)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{K}}} \quad \#\ \mathrm{se}(\hat{\mu}(x))$$

Pivots for PIs on the simple linear model: # PI predicts on Y^*, CI predicts on $\mu(x)$

$$Y^* \sim \mathrm{N}(\mu(x^*), \sigma^2), \qquad \hat{\mu}(x^*) \sim \mathrm{N}\left(\mu(x^*), \left(\frac{1}{n}+\frac{(x^*-\bar{x})^2}{K}\right)\sigma^2\right)$$

$$\Rightarrow\ Y^*-\hat{\mu}(x^*) \sim \mathrm{N}\left(0, \left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{K}\right)\sigma^2\right).$$

Define the sample covariance as:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \frac{n}{n-1}(\overline{X_iY_i}-\bar{X}\bar{Y})$$

$$\mathbb{E}(S_{XY}) = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left(\left((X_i-\mu_X)+(\mu_X-\bar{X})\right)\left((Y_i-\mu_Y)+(\mu_Y-\bar{Y})\right)\right)$$

$$= \frac{1}{n-1}(t_1-t_2-t_3+t_4) = \mathrm{Cov}(X,Y) \qquad \#\ \text{unbiased}$$

where

$$t_1 = \sum_{i=1}^{n} \mathbb{E}\big((X_i - \mu_X)(Y_i - \mu_Y)\big) = \sum_{i=1}^{n} \operatorname{Cov}(X_i, Y_i) = n \operatorname{Cov}(X, Y)\,;$$

$$t_2 = \sum_{i=1}^{n} \mathbb{E}\big((\bar{X} - \mu_X)(Y_i - \mu_Y)\big) = \sum_{i=1}^{n} \operatorname{Cov}\left(\frac{X_1 + \cdots + X_n}{n}, Y_i\right)$$
$$= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}\big(X_j, Y_i\big) = \operatorname{Cov}(X, Y)\,;$$

$$t_3 = \sum_{i=1}^{n} \mathbb{E}\big((X_i - \mu_X)(\bar{Y} - \mu_Y)\big) = \operatorname{Cov}(X, Y)\,;$$

$$t_4 = \sum_{i=1}^{n} \mathbb{E}\big((\bar{X} - \mu_X)(\bar{Y} - \mu_Y)\big) = \sum_{i=1}^{n} \operatorname{Cov}\left(\frac{X_1 + \cdots + X_n}{n}, \frac{Y_1 + \cdots + Y_n}{n}\right)$$
$$= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \operatorname{Cov}\big(X_j, Y_k\big) = \operatorname{Cov}(X, Y)\,.$$

To estimate ρ, define the sample correlation coefficient, also known as Pearson's correlation coefficient, as:

$$R = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \in [-1, 1] \qquad \text{\# biased}$$

An alternative ANOVA decomposition on the simple linear model:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\left(Y_i - \widehat{Y}_i\right) + \left(\widehat{Y}_i - \bar{Y}_i\right)\right)^2$$
$$= \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \left\{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right\}^2$$
$$= \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 + \left\{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}\right\}^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$= (1 - R^2)\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + R^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

This implies that R^2 is the proportion of the variation in Y explained by x, which is the same as the proportion of X explained by y. In this usage, R^2 is called the coefficient of determination.

CI for ρ from Fisher transformation:

$$g(r) = \operatorname{arctanh}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right);$$

$$g(R) \approx \mathrm{N}(g(\rho), (n-3)^{-1})$$

$$\mathbb{P}\left(\tanh\left(\operatorname{arctanh}(R) - \frac{c}{\sqrt{n-3}}\right) < \rho < \tanh\left(\operatorname{arctanh}(R) + \frac{c}{\sqrt{n-3}}\right)\right) = 1 - \alpha.$$

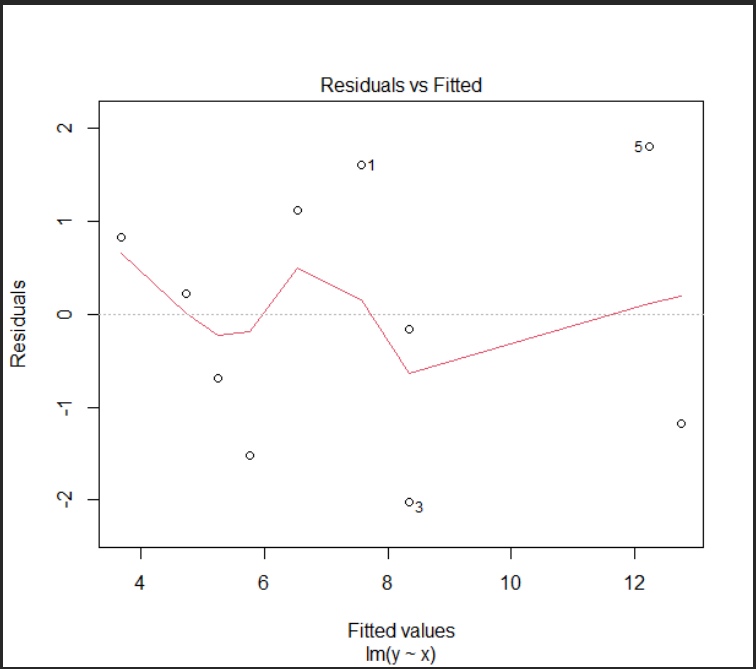
The linear regression in R code:

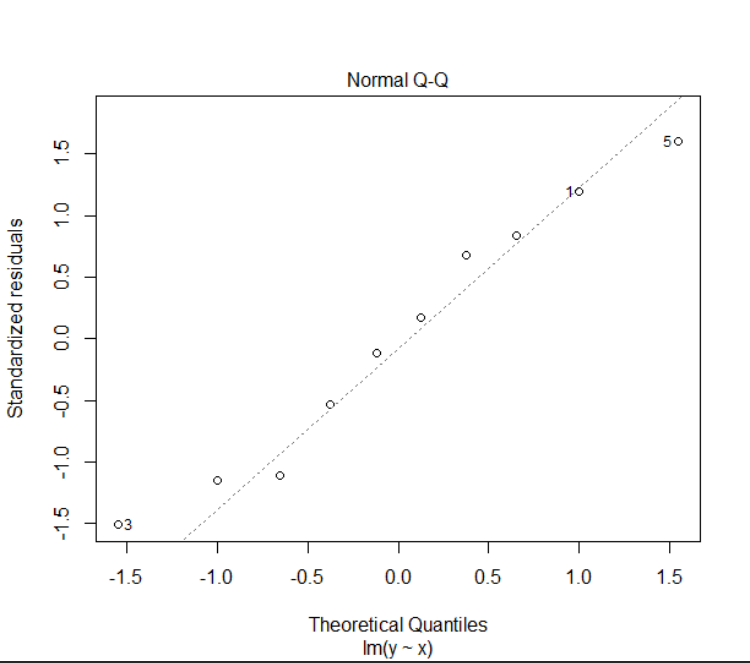
```
> x <- c(1.80, 1.40, 2.10, 0.30, 3.60, 0.70, 1.10, 2.10, 0.90, 3.80)
> y <- c(9.18, 7.66, 6.33, 4.51, 14.04, 4.94, 4.24, 8.19, 4.55, 11.57)
```

```
> cor(x, y)  # sample covariance
[1] 0.9148421

> cor.test(x, y)

    Pearson's product-moment correlation

data: x and y
t = 6.4078, df = 8, p-value = 0.0002074
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:  # CI for rho
 0.6726924 0.9799873
sample estimates:
      cor
0.9148421
```

```
## "model1" is an object that contains all the results of the regression
> model1 <- lm(y ~ x)
> summary(model1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01970 -1.05963  0.02808  1.04774  1.80580

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   2.9114      0.8479    3.434  0.008908 **   # H1 : alpha ≠ 0
x             2.5897      0.4041    6.408  0.000207 ***  # H1 : beta ≠ 0
---
Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.419 on 8 degrees of freedom     # sigma.hat
Multiple R-squared: 0.8369,    Adjusted R-squared: 0.8166  # R^2
F-statistic: 41.06 on 1 and 8 DF,  p-value: 0.0002074      # H1: beta ≠ 0
```

```
> plot(model1, 1:2)
```

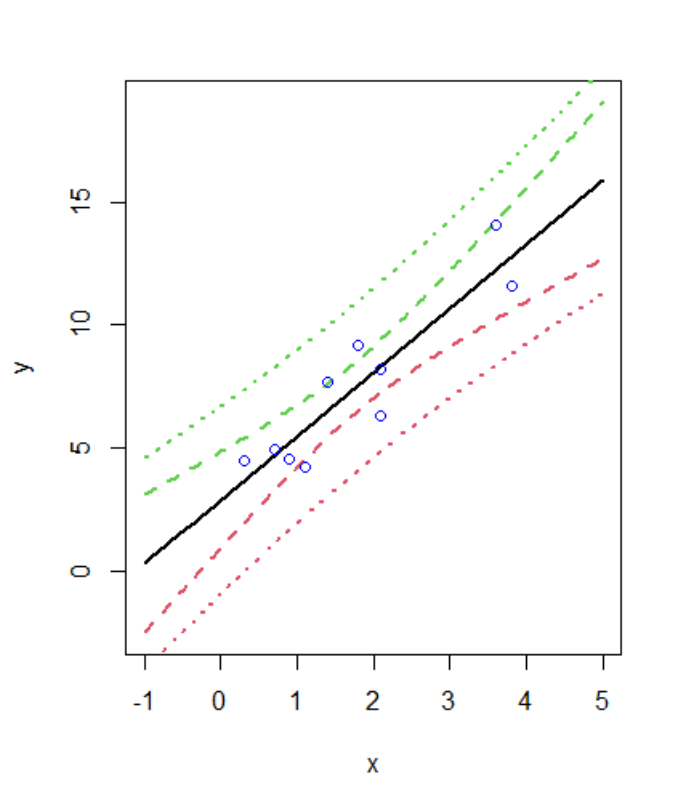





```
> confint(model1)  # CI for alpha and beta
                2.5 %   97.5 %
(Intercept) 0.9560629 4.866703
x           1.6577220 3.521623
```

```
> data2 <- data.frame(x = 3)  # calculate CI and PI for mu(3)
> predict(model1, newdata = data2, interval = "confidence", level = 0.95)
      fit      lwr      upr
1 10.6804 9.142823 12.21798
> predict(model1, newdata = data2, interval = "prediction", level = 0.95)
      fit   lwr     upr
1 10.6804 7.064 14.2968
```

```
## Plot CI bands against PI bands, PI bands are wider
> data3 <- data.frame(x = seq(-1, 5, 0.05))
> y.conf <- predict(model1, data3, interval = "confidence")
> y.pred <- predict(model1, data3, interval = "prediction")
> matplot(data3$x, y.conf, type = "l", lty = c(1,2,2), lwd = 2,
+         xlab = "x", ylab = "y")
> matplot(data3$x, y.pred, type = "l", lty = c(1,3,3), lwd = 2,
+         xlab = "x", ylab = "y", add = TRUE)
> points(x, y, col = "blue")
```



The kth order statistic $X_{(k)}$ in a sample is the sample's kth-smallest value.

Sampling from a continuous distribution, the CDF of $X_{(k)}$ is

$$G_k(x) = \mathbb{P}\left(X_{(k)} \le x\right) = \mathbb{P}(\text{at least } k\ X_i\text{'s less than } x) = \sum_{i=k}^{n} \binom{n}{i} F_X(x)^i \left(1 - F_X(x)\right)^{n-i};$$

Thus, differentiate to get the PDF as

$$\begin{aligned} g_k(x) &= \sum_{i=k}^{n} i \binom{n}{i} F_X(x)^{i-1}\left(1 - F_X(x)\right)^{n-i} f_X(x) \\ &\quad + \sum_{i=k}^{n} (n-i) \binom{n}{i} F_X(x)^i \left(1 - F_X(x)\right)^{n-i-1} \left(-f_X(x)\right) \\ &= k \binom{n}{k} F_X(x)^{k-1}\left(1 - F_X(x)\right)^{n-k} f_X(x) \end{aligned}$$

Alternatively,

$$\begin{aligned} g_k(x)dy &\approx \mathbb{P}\left(X - \frac{1}{2}dy < X_{(k)} \le x + \frac{1}{2}dy\right) \\ &= \binom{n}{k-1}\binom{n-k+1}{n-k} \mathbb{P}\left(X_i \le x - \frac{1}{2}dy\right)^{k-1} \mathbb{P}\left(X_i > x + \frac{1}{2}dy\right)^{n-k} \mathbb{P}\left(x - \frac{1}{2}dy < X_i \le x + \frac{1}{2}dy\right) \approx k \binom{n}{k} F_X(x)^{k-1}\left(1 - F_X(x)\right)^{n-k} f_X(x)dy. \end{aligned}$$

Let $W_{(k)} = F\left(X_{(k)}\right)$, which is an order statistic from a Unif$(0, 1)$ distribution.

Then the PDF of $W_{(k)}$ is derived as

$$g_{(k)}(w) = k \binom{n}{k} F_X(x)^{k-1}\left(1 - F_X(x)\right)^{n-k} f_X(x) = k \binom{n}{k} w^{k-1}(1-w)^{n-k}.$$

This is a beta distribution, $F\left(X_{(k)}\right) \sim \text{Beta}(k, n-k+1)$.

For continuous rv's, the p-quantile is a number π_p such that $\pi_p = F^{-1}(p)$.

For discrete rv's, the p-quantile is the smallest value π_p such that $\pi_p \ge F^{-1}(p)$.

The 0.5 quantile is the median, $m = \pi_{0.5}$. When $f(x)$ is symmetric, $m = \mu$.

The 0.25 and 0.75 quantiles are the first and third quartiles, $Q_1 = \pi_{0.25}$ and $Q_3 = \pi_{0.75}$. The difference of them is called the interquartile range (IQR).

The sample quantile is $\hat{\pi}_p = x_{(k)}$, where

1. 'Type 6': $p = k/(n+1)$; # $\mathbb{E}\left(F\left(X_{(k)}\right)\right) = k/(n+1)$
2. 'Type 7': $p = (k-1)/(n-1)$; # mode $\left(F\left(X_{(k)}\right)\right) = (k-1)/(n-1)$
3. 'Type 1': $k = \lceil np \rceil$. # $\hat{F}\left(x_{(\lceil np \rceil)}\right) = n^{-1}\sum_{i=1}^{n} I\left(x_i \le x_{(\lceil np \rceil)}\right) \approx p$

The sample median is

$$\widehat{m} = \begin{cases} X_{((n+1)/2)} & \text{when } n \text{ is odd} \\ \left(X_{(n/2)} + X_{(n/2+1)}\right)/2 & \text{when } n \text{ is even.} \end{cases}$$

For large sample sizes, it can be shown that

$$\hat{\pi}_p \approx \mathrm{N}\left(\pi_p, \frac{p(1-p)}{nf\left(\pi_p\right)^2}\right), \qquad \text{\# proof?}$$

where f is the PDF of the population distribution.

From this, we derive that the median, $\widehat{M} \approx \mathrm{N}(m, (4nf(m)^2)^{-1})$.

Let $W = \sum_{i=1}^{n} I\left(X_i < \pi_p\right) \sim \mathrm{Bi}(n, p)$. CI for arbitrary quantiles:

$$\mathbb{P}\left(X_{(i)} < \pi_p < X_{(j)}\right) = \mathbb{P}(i \le W \le j-1) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k} = 1 - \alpha.$$

Here we do not make distributional assumptions, but it is still possible to obtain exact or asymptotic sampling distributions for various statistics. This technique is called the distribution-free methods.

The sign test tests for the median ($H_0 : m = m_0$) for continuous distributions.

Since this test doesn't use the information about the size of the differences, it may have large type II error or small power, so most often used for ordinal data.

The test statistic is the number of positive signs amongst $X_1 - m_0, \ldots, X_n - m_0$,

$$Y = \sum_{i=1}^{n} I(\mathrm{sgn}(X_i - m_0) > 0) \sim \mathrm{Bi}(n, 0.5).$$

```
## Test m < 6.2, equivalent to test p < 0.5
> x <- c(6.80, 5.70, 6.90, 5.30, 4.10, 9.80, 1.70, 7.00, 2.10, 19.00, 18.90,
+         16.90, 10.40, 44.10, 2.90, 2.40, 4.80, 18.90, 4.80, 7.90)
> binom.test(sum(x > 6.2), length(x), alternative = "less")

    Exact binomial test

data: 11 and 20
number of successes = 11, number of trials = 20,
 p-value = 0.7483   # pbinom(11, 20, 0.5)
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.7413494
sample estimates:
probability of success
                  0.55
```

Wilcoxon singed-rank test (or called Wilcoxon one-sample test) assumes the underlying distribution is also symmetrical and continuous.

The test statistic is the sum of signed ranks

$$W = \sum_{i=1}^{n} \operatorname{sgn}(X_i - m_0) \cdot \operatorname{rank}(|X_i - m_0|)$$
$$= \sum_{i=1}^{n} i \cdot \operatorname{sgn}(X_j - m_0) \quad \text{where } \operatorname{rank}(|X_j - m_0|) = i.$$

A popular alternative is to use the sum of the positive ranks only

$$V = \sum_{i=1}^{n} I(\operatorname{sgn}(X_i - m_0) > 0) \cdot \operatorname{rank}(|X_i - m_0|) = \frac{W}{2} + \frac{n(n+1)}{4}.$$

We assumed a continuous population distribution. Thus, all observations will differ with probability 1. In practice, the data are reported to finite precision, so we could have exactly equal values. If this happens, the 'rank' assigned for the tied values should be equal to the average of the ranks they span.

Under $H_0 : m = m_0$,

$$\mathbb{E}(W) = \sum_{i=1}^{n} \mathbb{E}(i \cdot \operatorname{sgn}(X_j - m_0)) = \sum_{i=1}^{n} 0 = 0,$$

$$\mathbb{V}(W) = \mathbb{E}(W^2) = \sum_{i=1}^{n} \mathbb{E}\left(i^2 \cdot \operatorname{sgn}(X_j - m_0)^2\right) = \sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6.$$

For large n, the statistic W follows a normal distribution,

$$Z = \frac{W}{\sqrt{n(n+1)(2n+1)/6}} \approx \mathrm{N}(0,1),$$

which allows us to determine the rejection region.

```
## Test m > 3.7
> x <- c(5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3)
> wilcox.test(x, mu = 3.7, alternative = "greater",  # mu ≈ median
+         exact = TRUE)  # the exact sampling distribution of V

    Wilcoxon signed rank test

data: x
V = 40, p-value = 0.1162   # 1 - psignrank(v = 39, n = 10)
alternative hypothesis: true location is greater than 3.7

# Calculate approximate p-value, based on W
> z <- 25 / sqrt(10 * 11 * 21 / 6)
> 1 - pnorm(z)
[1] 0.1013108
```

Indeed, this test is most often used for paired samples. The assumption of symmetry is quite reasonable in this setting, since under H_0 we would typically assume X and Y have the same distribution and therefore $D = X - Y$ becomes symmetrical as well.

Wilcoxon rank-sum test (or called Wilcoxon two-sample test, Mann-Whitney U test) takes independent random samples $X_1, \dots, X_{nX}$ and $Y_1, \dots, Y_{nY}$ from two different populations with median m_X and m_Y.

To test $H_0 : m_X = m_Y$, order the **combined** sample and let W be the sum of ranks of Y_i's. This is the Wilcoxon rank-sum statistic.

$\mathbb{E}(W) = n_Y(n_X + n_Y + 1)/2$ and $\mathbb{V}(W) = n_X n_Y(n_X + n_Y + 1)/12$. W is approximately normally distributed when nX and nY are large.

A popular alternative is to use U, the number of all pairs (X_i, Y_i) such that $Y_j \le X_i$. U and W are deterministically related. $\mathbb{E}(U) = n_X n_Y/2$ and $\mathbb{V}(U) = \mathbb{V}(W)$. # proof?

```
## Test mX ≠ mY
> x <- c(117.1, 121.3, 127.8, 121.9, 117.4, 124.5, 119.5, 115.1)
> y <- c(123.5, 125.3, 126.5, 127.9, 122.1, 125.6, 129.8, 117.2)
> wilcox.test(x, y)

    Wilcoxon rank sum test

data: x and y
W = 13,  # R uses U but calls it W
 p-value = 0.04988   # 2 * pwilcox(13, 8, 8)
alternative hypothesis: true location shift is not equal to 0
```

A goodness-of-fit test tells how well a given model fits a set of data. The most commonly used is Pearson's chi-squared test. This test operates on categorical data. Can also apply it on continuous data by first partitioning the data into separate classes.

The chi-squared test statistic measures the badness of fit, so it is always right-tailed, i.e., becomes more extreme on the right.

Binomial model $Y_1 \sim \text{Bi}(n, p_1)$, the success rate is p_1:

$$Q_1 = Z^2 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)}$$
$$= \frac{(Y_1 - np_1)^2}{np_1} + \frac{\big(n - Y_1 - n(1-p_1)\big)^2}{n(1-p_1)} = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} \approx \chi_1^2$$

where $O_i = Y_i$ is the observed number, and $E_i = np_i$ is the expected number.

Multinomial model $Y_i \sim \text{Bi}(n, p_i)$, with k possible outcomes:

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi_{k-1}^2$$

```
## Test whether the data fits the multinomial model
> x <- c(26, 15, 32, 7)
> p <- c(0.25, 0.15, 0.5, 0.1)
> t1 <- chisq.test(x, p = p)
> t1

    Chi-squared test for given probabilities

data: x
X-squared = 4.275, df = 3, p-value = 0.2333

> t1$residuals   # (Oi - Ei)/sqrt(Ei)
[1] 1.3416408 0.8660254 -1.2649111 -0.3535534
> sum(t1$residuals^2)
[1] 4.275
> 1 - pchisq(4.275, 3)   # p-value
[1] 0.2332594
```

Poisson model $Y \sim \mathrm{Pn}(\lambda)$, without an exact model to compare against:

```
## Test with the data fits the Poisson model with an unspecified λ
> x <- (7, 4, 3, 6, 4, 4, 5, 3, 5, 3, 5, 5, 3, 2, 5, 4, 3, 3, 7, 6, 6, 4, 3,
+         9, 11, 6, 7, 4, 5, 4, 7, 3, 2, 8, 6, 7, 4, 1, 9, 8, 4, 8, 9, 3, 9,
+         7, 7, 9, 3, 10)

> mean(x)  # estimate λ by MLE
[1] 5.4

# Normal approximation requires Ei ≥ 5, so needs partition
> x1 <- cut(x, breaks = c(0, 3.5, 4.5, 5.5, 6.5, 7.5, 100))
> t1 <- table(x1)
> x <- as.numeric(t1)
> x
[1] 13 9 6 5 7 10

> n <- sum(x)
> p1 <- sum(dpois(0:3, 5.4));
> p2 <- dpois(4, 5.4)
> p3 <- dpois(5, 5.4)
> p4 <- dpois(6, 5.4)
> p5 <- dpois(7, 5.4)
> p6 <- 1 - (p1 + p2 + p3 + p4 + p5)
> p <- c(p1, p2, p3, p4, p5, p6)

> chisq.test(x, p = p)

    Chi-squared test for given probabilities

data: x
X-squared = 2.7334, df = 5, p-value = 0.741

# df = k − p − 1, where p is the number of estimated parameters
> df = 6 - 1 - 1
> 1 - pchisq(2.7334, 4)   # true p-value
[1] 0.6033828
```

Suppose we have multiple categorical variables, or continuous variables partitioned into classes. A contingency table records the number of observations for each possible cross-classification $(A_1, \dots, A_r \times B_1, \dots, B_c)$ of these variables.

Define that

$$p_{ij} = \mathbb{P}(A_i \cap B_j),$$

$$p_{i\bullet} = \sum_{j=1}^{c} p_{ij} = \mathbb{P}(A_i), \qquad p_{\bullet j} = \sum_{i=1}^{r} p_{ij} = \mathbb{P}(B_j).$$

Test the independence model $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}$.

Estimate that $\hat{p}_{ij} = \hat{p}_{i\bullet}\hat{p}_{\bullet j} = (Y_{i\bullet}/n)(Y_{\bullet j}/n) = Y_{i\bullet}Y_{\bullet j}/n^2$.

Pearson's chi-squared statistic for given p_{ij} is

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(Y_{ij} - Y_{i\bullet}Y_{\bullet j}/n)^2}{Y_{i\bullet}Y_{\bullet j}/n} \approx \chi^2_{(r-1)(c-1)}$$

Here, $k = rc$, the total number of cells in the table. We estimated $r - 1$ marginal probabilities for the rows and $c - 1$ for the columns. Therefore, the d.f. is $(r - 1)(c - 1)$.

```
## Test for relationship between sex and whether or not were firstborn
> x <- rbind(male = c(first = 34, later = 74),
+         female = c(first = 20, later = 22))
> x
       first later
male      34    74
female    20    22

> c1 <- chisq(x, correct = FALSE)
> c1

    Pearson's Chi-squared test

data: x
X-squared = 3.418, df = 1, p-value = 0.06449

> c1$observed
       first later
male      34    74
female    20    22

> c1$expected
       first later
male   38.88 69.12
female 15.12 26.88
```

We wish to determine if two groups of nurses distribute their time in six different categories about the same way. That is, the hypothesis under consideration is $H_0 : p_{i1} = p_{i2},\ i = 1, \dots, 6$. To test this, nurses are observed at random throughout several days,

each observation resulting in a mark in one of the six categories. The summary data are given in the following frequency table:

	Category						
	1	2	3	4	5	6	Total
Group I	95	36	71	21	45	32	300
Group II	53	26	43	18	32	28	200

Do a chi-squared test with $\alpha = 0.05$.

This a problem where we want to do a goodness-of-fit test of a particular model but where we need to first estimate some of the parameters. We can set it up in one of two ways.

The first way is to think about the null distribution and work out which parameters need to be estimated. Under H_0 we have $p_{i1} = p_{i2}$, so let's call both of them p_i (since they are equal). These define the probabilities of each category (columns) that apply to each group of nurses (rows). Note that these are conditional probabilities, $p_i = \mathbb{P}(\text{category } i \mid \text{group I}) = \mathbb{P}(\text{category } i \mid \text{group II})$. To complete the model we also need to estimate the marginal probabilities of the two groups, let's call these $g_j = \mathbb{P}(\text{group } j)$, for $j = 1, 2$. The null model, therefore, is that the probability of an observation for category i in group j is $g_j p_i$. Note that there are 6 independent parameters to estimate (5 conditional column probabilities and one row probability), so ultimately we'll end up with a test with $12 - 6 - 1 = 5$ d.f..

The other way is to note that this model is equivalent to the usual test of independence of a contingency table, we end up estimating the same parameters and apply the same test as described above.

Under either setup, the observed and expected frequencies are:

		Category					
		1	2	3	4	5	6
Group I	O	95.0	36.0	71.0	21.0	45.0	32.0
	E	88.8	37.2	68.4	23.4	46.2	36.0
Group II	O	53.0	26.0	43.0	18.0	32.0	28.0
	E	59.2	24.8	45.6	15.6	30.8	24.0

and as there are 5 df

$$\chi^2 = \frac{(95 - 88.88)^2}{88.88} + \cdots + \frac{(28 - 24)^2}{24} = 3.23 < 11.07 \qquad (0.95 \text{ quantile of } \chi^2_5)$$

so we cannot reject H_0.

Analysis of variance (ANOVA) is a technique used to compare the means of more than two populations. It assumes that each population having a normal distribution with common variance.

One-way ANOVA takes random samples from k populations (treatments).

Define that

1. Total sample size $\quad n = n_1 + n_2 + \cdots + n_k$
2. Group mean $\quad \bar{X}_{i\bullet} = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$
3. Grand mean $\quad \bar{X}_{\bullet\bullet} = n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = n^{-1} \sum_{i=1}^{k} n_i \bar{X}_{i\bullet}$
4. Error sum of squares (within group SS)

   $$SS(E) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\bullet}\right)^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$$

   $$SS(E)/\sigma^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2/\sigma^2 \sim \chi^2_{n-k} \;\Rightarrow\; \hat{\sigma}^2 = SS(E)/(n-k) \quad \text{\# unbiased}$$

5. Treatment sum of squares (between groups SS)

   $$SS(T) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2$$

6. Total sum of squares

   $$SS(TO) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\bullet} + \bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2$$
   $$= SS(T) + SS(E)$$

   $$SS(TO)/\sigma^2 = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1} \quad \text{under } H_0$$

   $$SS(T)/\sigma^2 = SS(TO)/\sigma^2 - SS(E)/\sigma^2 \sim \chi^2_{n-1} - \chi^2_{n-k} \sim \chi^2_{k-1} \quad \text{under } H_0$$

Tests the null hypothesis, which states that samples in all groups are drawn from populations with the same mean values:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu \quad \text{versus} \quad H_1 : \bar{H}_0.$$

Under H_0 we should have the test statistic

$$F = \frac{SS(T)/(k-1)}{SS(E)/(n-k)} \sim F_{k-1,n-k}\,;$$

Under H_1 the numerator will tend to be larger. Therefore, reject H_0 if F is too large.

```
> head(data1)
  Position Force
1        1    92
2        1    90
3        1    87
4        1   105
5        1    86

> table(data1$Position)
1 2 3 4 5
7 7 7 7 7

# "factor" denotes categorical variables
> model1 <- lm(Force ~ factor(Position), data = data1)
> anova(model1)
Analysis of Variance Table

Response: Force

                 Df  Sum Sq Mean Sq F value    Pr(>F)      # MS = SS / Df
factor(Position)  4 16672.1  4168.0  44.202 3.664e-12 ***  # treatment
Residuals        30  2828.9    94.3                        # error
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In one-way ANOVA, the observations were partitioned into k groups. In other words, they were defined by a single factor (categorical variables). If we have two such factors, we can extend the procedure to give two-way ANOVA.

Suppose that Factor 1 has a levels, Factor 2 has b levels, and we have exactly c observations per factor combination.

Let $X_{ijk} \sim \mathrm{N}(\mu_{ijk}, \sigma^2)$, $i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, c$ be independent.

Consider two different models: # the only random element is ϵ_{ijk}

$$X_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad \text{where} \quad \sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0.$$

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \qquad \text{\# requires } > 1 \text{ observations per cell}$$

$$\text{where} \quad \sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = 0, \quad \sum_{j=1}^{b} \gamma_{ij} = 0.$$

where μ is an overall effect, α_i and β_j are the main effects from the ith row, and the jth column, γ_{ij} is the interaction between factors, ϵ_{ijk} is the error term.

Define that

1. $\bar{X}_{ij\bullet} = c^{-1} \sum_{k=1}^{c} X_{ijk}$ # $\mathbb{E}(\bar{X}_{ij\bullet}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$
2. $\bar{X}_{i\bullet\bullet} = (bc)^{-1} \sum_{j=1}^{b} \sum_{k=1}^{c} X_{ijk}$ # $\mathbb{E}(\bar{X}_{i\bullet\bullet}) = \mu + \alpha_i$

3. $\bar{X}_{\bullet j\bullet} = (ac)^{-1}\sum_{i=1}^{a}\sum_{k=1}^{c} X_{ijk}$ # $\mathbb{E}(\bar{X}_{\bullet j\bullet}) = \mu + \beta_j$

4. $\bar{X}_{\bullet\bullet\bullet} = (abc)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk}$ # $\mathbb{E}(\bar{X}_{\bullet\bullet\bullet}) = \mu$

5. $$SS(TO) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(X_{ijk} - \bar{X}_{\bullet\bullet\bullet}\right)^2$$
$$= bc\sum_{i=1}^{a}(\bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet\bullet\bullet})^2 + ac\sum_{j=1}^{b}\left(\bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet}\right)^2$$
$$+ c\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{X}_{ij\bullet} - \bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet j\bullet} + \bar{X}_{\bullet\bullet\bullet}\right)^2$$
$$+ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(X_{ijk} - \bar{X}_{ij\bullet}\right)^2$$
$$= SS(A) + SS(B) + SS(AB) + SS(E)$$

Now interested in testing

$$H_{0A} : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0\,; \quad H_{0B} : \beta_1 = \beta_2 = \cdots = \beta_b = 0\,;$$
$$H_{0AB} : \gamma_{ij} = 0,\ \ i = 1, \ldots, a,\ \ j = 1, \ldots, b. \quad \text{\# interaction model only}$$

Believing there is no interaction, and both H_{0A} and H_{0B} are true, then we have

$$SS(A)/\sigma^2 \sim \chi^2_{a-1},\ \ SS(B)/\sigma^2 \sim \chi^2_{b-1},\ \ SS(TO)/\sigma^2 \sim \chi^2_{abc-1},\ \ SS(E)/\sigma^2 \sim \chi^2_{abc-a-b+1}.$$

To test H_{0A} and H_{0B}, use:

$$F_A = \frac{SS(A)/(a-1)}{SS(E)/\big((a-1)(b-1)\big)} \quad \text{and} \quad F_B = \frac{SS(B)/(b-1)}{SS(E)/\big((a-1)(b-1)\big)}.$$

```
> head(data2)
  Car Fuel Consumption
1   1    1          16
2   1    2          18
3   1    3          21
4   1    4          21
5   2    1          14

> model2 <- lm(Consumption ~ factor(Car) + factor(Fuel), data = data2)
> anova(model2)
Analysis of Variance Table

Response: Consumption
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(Car)   2     24 12.0000      18 0.002915 **
factor(Fuel)  3     30 10.0000      15 0.003401 **
Residuals     6      4  0.6667
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now look at the interaction, but still need to reject H_{0AB} at significance level α when:

$$F_{AB} = \frac{SS(AB)/\big((a-1)(b-1)\big)}{SS(E)/\big(ab(c-1)\big)} > c$$

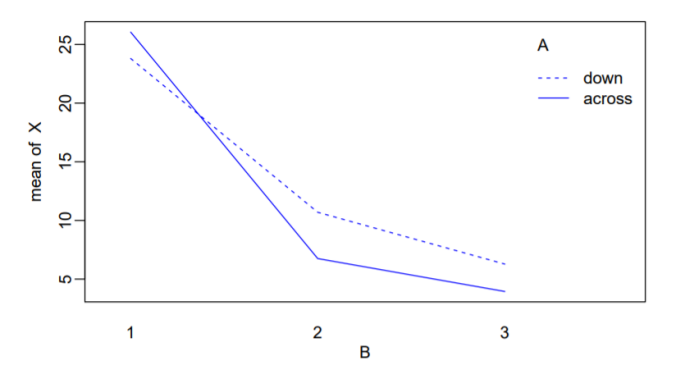
where c is the $(1-\alpha)$ quantile of $F_{(a-1)(b-1),ab(c-1)}$.

If we fail to reject H_{0AB} (i.e., there is interaction), to test H_{0A}, H_{0B}, use:

$$F_A = \frac{SS(A)/(a-1)}{SS(E)/\big(ab(c-1)\big)} \quad \text{and} \quad F_B = \frac{SS(B)/(b-1)}{SS(E)/\big(ab(c-1)\big)}.$$

```
> head(data3)
     A B    X
1 down 1 19.5
2 down 1 18.5
3 down 1 32.0
4 down 1 21.5
5 down 1 28.5
6 down 1 33.0

> table(data3[, 1:2])
        B
A        1  2  3
 down   18 18 18
 across 18 18 18

> model3 <- lm(X ~ factor(A) * factor(B), data = data3)
> anova(model3)
Analysis of Variance Table

Response: X
                     Df Sum Sq Mean Sq  F value  Pr(>F)
factor(A)             1   48.7    48.7   2.8849 0.09246 .
factor(B)             2 8022.7  4011.4 237.7776 < 2e-16 ***
factor(A):factor(B)   2  185.9    93.0   5.5103 0.00534 **
Residuals           102 1720.8    16.9
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Interaction plot, mean of X in plot denotes X̄ij.
with(data3, interaction.plot(B, A, X, col = "blue"))
```



ANOVA on the simple linear regression:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}\left((Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y})\right)^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$$
$$\Rightarrow\ SS(TO)=SS(E)+SS(R)$$

where $SS(R)$ is called the regression SS or model SS.

Usually interested in testing the significance of regression, $H_0:\beta=0$. Reject when

$$MS(E)=\frac{SS(E)}{n-2}=\hat{\sigma}^2,\ \ MS(R)=SS(R)\ \ \Rightarrow\ \ F=\frac{MS(R)}{MS(E)}\sim F_{1,n-2}>c$$

```
> anova(model4)
Analysis of Variance Table

Response: final_exam
            Df Sum Sq Mean Sq F value   Pr(>F)     # H1: beta ≠ 0
prelim_test  1 416.39  416.39  15.301 0.004471 **  # model
Residuals    8 217.71   27.21                      # error
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observe that F-test and t-test gives the same p-value, i.e., they are equivalent.

Let the test statistics be $T\sim t_{n-2}$ and $F\sim F_{1,n-2}$ then

$$\mathbb{P}\left(T\geq\frac{|\hat{\beta}|}{\hat{\sigma}/\sqrt{K}}\right)=\mathbb{P}\left(T^2\geq\frac{K\hat{\beta}^2}{\hat{\sigma}^2}\right)=\mathbb{P}\left(F\geq\frac{K\hat{\beta}^2}{\hat{\sigma}^2}\right)=\mathbb{P}\left(F\geq\frac{\sum_{i=1}^{n}\hat{\beta}^2(x_i-\bar{x})^2}{\hat{\sigma}^2}\right)$$
$$=\mathbb{P}\left(F\geq\frac{\sum_{i=1}^{n}\left(\hat{\alpha}+\hat{\beta}(x_i-\bar{x})-\bar{Y}\right)^2}{\hat{\sigma}^2}\right)=\mathbb{P}\left(F\geq\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{\hat{\sigma}^2}\right)$$
$$=\mathbb{P}\left(F\geq\frac{MS(R)}{MS(E)}\right).$$

The likelihood ratio test (LRT) is a general procedure that can find the optimal test for a given problem.

Suppose we have H_0 and H_1 and both are composite and of the form:

$$H_0:\theta\in A_0\ \ \text{versus}\ \ H_1:\theta\in A_1.$$

The likelihood ratio is

$$\lambda=L_0/L_1=\max_{\theta\in A_0}L(\theta)\,/\,\max_{\theta\in A_1}L(\theta).$$

Small λ supports more for H_1 over H_0, so reject H_0 when $\lambda<k$.

Suppose $X_i \sim \mathrm{N}(\mu, \sigma^2)$. We want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Under H_0 we have $\mu = \mu_0$, and under H_1 we need to use its MLE, $\hat{\mu} = \bar{x}$.

And we also need to estimate both variances by MLE,

$$\hat{\sigma}_0^2 = n^{-1}\sum_{k=1}^{n}(x_k - \mu_0)^2\,, \qquad \hat{\sigma}_1^2 = n^{-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

Then reject when

$$\lambda = \frac{L_0}{L_1} = \frac{(2\pi n^{-1}\sum_{i=1}^{n}(x_i-\mu_0)^2)^{-\frac{n}{2}}\exp\left\{-\dfrac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{2n^{-1}\sum_{k=1}^{n}(x_k-\mu_0)^2}\right\}}{(2\pi n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2)^{-\frac{n}{2}}\exp\left\{-\dfrac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{2n^{-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}\right\}} = \left[\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^{-\frac{n}{2}}$$

$$= \left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^{-\frac{n}{2}} = \left[1 + \frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^{-\frac{n}{2}} \leq k$$

$$\Rightarrow \quad \frac{|\bar{X}-\mu_0|}{S/\sqrt{n}} \sim t_{n-1} \leq c = \sqrt{\left(k^{-\frac{2}{n}} - 1\right)(n-1)}\,.$$

Frequentist probability is a (classical) interpretation of probability; it defines an event's probability as the limit of its relative frequency under hypothetical repetitions.

The Law of Large Numbers shows such limit exists.

The degree of plausibility, or strength of belief, of a given statement based on existing knowledge and evidence, expressed as a probability, is known as Bayesian probability.

Bayesian probability can be assigned to any statement, even when no random process is involved, and irrespective of whether the event has yet occurred or not.

In Bayesian inference, parameters and hypotheses are modelled as random variables. They quantify and express our uncertainty, both before and after seeing any data.

$$f(\theta|x_1, x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n|\theta)f(\theta)}{f(x_1, x_2, \dots, x_n)} \propto f(x_1, x_2, \dots, x_n|\theta)f(\theta)$$

$$\underline{\text{Posterior } \propto \text{ Likelihood } \times \text{ Prior}}$$

The prior as being equivalent to unobserved data, or called pseudo-observations. The prior should be diffuse enough to allow the data, if sufficient enough, to overwhelm it.

The choice of prior is often seen as a weakness, since it is difficult to specify a full probability model for complex problems. But on the other hand, frequentist inference needs further choice on, e.g., which estimate to use, as well.

Take probability intervals from the posterior, referred to as credible intervals.

If the posterior distributions are in the same probability distribution family as the prior probability distribution, the prior and posterior are then called conjugate distributions.

Random sample: $X_1, \dots, X_n \sim \mathrm{N}(\theta, \sigma^2)$, with σ^2 known.

Prior: $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$.

1. Summarising the data by $\bar{X} \sim \mathrm{N}(\theta, \sigma^2/n)$.

$$f(\theta|\bar{x}) \propto f(\bar{x}|\theta)f(\theta) = \left(\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}\right)^{-1} e^{-\frac{(\bar{x}-\theta)^2}{2\sigma^2/n}} \left(\sigma_0\sqrt{2\pi}\right)^{-1} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}}$$

$$\propto \exp\left\{-\frac{(\bar{x}-\theta)^2}{2\sigma^2/n} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\} \propto \exp\left\{\frac{\bar{x}\theta}{\sigma^2/n} - \frac{\theta^2}{2\sigma^2/n} - \frac{\theta^2}{2\sigma_0^2} + \frac{\mu_0\theta}{\sigma_0^2}\right\}$$

$$= \exp\left\{-\left(\theta^2 - \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2/n}}{\frac{1}{2\sigma_0^2} + \frac{1}{2\sigma^2/n}}\theta\right) \Big/ 2\left(\frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}}\right)\right\}$$

$$\propto \exp\left\{-\left(\theta - \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2/n}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}}\right)^2 \Big/ 2\left(\frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}}\right)\right\} = \exp\left\{\frac{-(\theta-\mu_1)^2}{2\sigma_1^2}\right\}$$

$$\Rightarrow \quad f(\theta|\bar{x}) \sim \mathrm{N}(\mu_1, \sigma_1^2) \quad \text{where} \quad \mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2/n}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}} = \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}}\mu_0 + \frac{\frac{1}{\sigma^2/n}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}}\bar{x}$$

$$\text{precision} = \frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}$$

2. Use the joint PDF to find the likelihood function

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} \left(\sigma\sqrt{2\pi}\right)^{-1} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} = \left(\sigma\sqrt{2\pi}\right)^{-n} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}}$$

$$f(\theta|x_1, \dots, x_n) \propto f(x_1, \dots, x_n|\theta) f(\theta) \propto \exp\left\{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\}$$

$$= \exp\left\{-\frac{\sum_{i=1}^{n} x_i^2 + n\theta^2 - 2\theta\sum_{i=1}^{n} x_i}{2\sigma^2} - \frac{\theta^2}{2\sigma_0^2} + \frac{\mu_0\theta}{\sigma_0^2}\right\}$$

$$\propto \exp\left\{\frac{\bar{x}\theta}{\sigma^2/n} - \frac{n\theta^2}{2\sigma^2} - \frac{\theta^2}{2\sigma_0^2} + \frac{\mu_0\theta}{\sigma_0^2}\right\} \propto \cdots \qquad \text{\# leads to the same result}$$

Can make the prior "non-informative" on θ by setting $\sigma_0 \to \infty$, called a diffuse prior. Using a diffuse prior, the credible interval is the same as the confidence interval.

This type of prior (cannot integrate to 1) is called an improper prior.

Random Sample: $X_1, \dots, X_n \sim \mathrm{Bi}(n, \theta)$.

Prior: $\theta \sim \mathrm{Beta}(\alpha, \beta)$ # α successes and β failures in $(\alpha + \beta)$ pseudo-counts

$$f(\theta|x) \propto f(x|\theta) f(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \sim \mathrm{Beta}(x+\alpha, n-x+\beta), \quad 0 \le \theta \le 1$$

$$\mathbb{E}(\theta|x) = \frac{\alpha + x}{\alpha + \beta + n} = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right) + \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{x}{n}\right)$$

When using $\mathrm{U}(0, 1) \sim \mathrm{Beta}(1, 1)$ as a "flat" prior, the posterior mode is equal to MLE.

Random Sample: $X_1, \dots, X_n \sim \mathrm{Exp}(\text{rate} = \lambda)$.

Prior: $\lambda \sim \mathrm{Gamma}(\alpha, \beta)$.

$$f(\lambda|x_1, \dots, x_n) \propto f(x_1, \dots, x_n|\lambda) f(\lambda) = \lambda^n e^{-\lambda\sum_{i=1}^{n} x_i} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \lambda^{n+\alpha-1} e^{-(\sum_{i=1}^{n} x_i + \beta)\lambda} \sim \mathrm{Gamma}(n+\alpha, \textstyle\sum_{i=1}^{n} x_i + \beta)$$

$$\mathbb{E}(\lambda|x_1, \dots, x_n) = \frac{n+\alpha}{\sum_{i=1}^{n} x_i + \beta}$$

Random Sample: $X_i \sim \theta + \text{Exp}(1)$; $f(x) = e^{-(x-\theta)}$, $x > \theta$. Use uniform prior.

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} e^{-(x_i-\theta)} I(x_i > \theta) \propto e^{n\theta} I\left(\theta < x_{(1)}\right)$$

$$\int_{-\infty}^{\infty} c e^{n\theta} I\left(\theta < x_{(1)}\right) d\theta = c \int_{-\infty}^{x_{(1)}} e^{n\theta} d\theta = c \left[\frac{e^{n\theta}}{n}\right]_{-\infty}^{x_{(1)}} = 1 \;\Rightarrow\; c = n e^{-n x_{(1)}}$$

$$F(x_1, \dots, x_n|\theta) = e^{n\left(\theta - x_{(1)}\right)}, \quad \theta < x_{(1)}$$

$\Rightarrow$ 95% credible interval is $\left(x_{(1)} + n^{-1}\log(0.05)\,, x_{(1)}\right)$.

Suppose that $X_1, \dots, X_n$ is a random sample from $f(x,\theta)$, which is continuous. Assume θ is not a boundary parameter, i.e., the support set of $f(x,\theta)$ does not depend on θ.

The score function and observed information function:

$$U(\theta) = \frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(X_i;\theta)}{\partial \theta} = \sum_{i=1}^{n} U_i\,;$$

$$V(\theta) = -\frac{\partial U}{\partial \theta} = -\frac{\partial^2 \ell}{\partial \theta^2} = \sum_{i=1}^{n} -\frac{\partial^2 \ln f(X_i;\theta)}{\partial \theta^2} = \sum_{i=1}^{n} V_i\,.$$

Determine $\mathbb{E}\big(U(\theta)\big)$ by exchanging the order of integration and differentiation:

$$\begin{aligned}\mathbb{E}\big(U(\theta)\big) &= \mathbb{E}\left(\sum_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial \ln f(X_i;\theta)}{\partial \theta}\right) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{\partial \ln f(x;\theta)}{\partial \theta} f(x;\theta)\,dx \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta}\,dx = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x;\theta)\,dx = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} 1 = 0\,.\end{aligned}$$

Determine $\mathbb{V}\big(U(\theta)\big)$ by taking derivative on one of the above results,

$$\begin{aligned}0 &= \int_{-\infty}^{\infty} \frac{\partial \ln f(x;\theta)}{\partial \theta} f(x;\theta)\,dx = \int_{-\infty}^{\infty} \left\{\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2} f(x;\theta) + \frac{\partial \ln f(x;\theta)}{\partial \theta}\frac{\partial f(x;\theta)}{\partial \theta}\right\} dx \\ &= \int_{-\infty}^{\infty} \left\{\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2} f(x;\theta) + \left\{\frac{\partial \ln f(x;\theta)}{\partial \theta}\right\}^2 f(x;\theta)\right\} dx\end{aligned}$$

$$\begin{aligned}\Rightarrow\ -\int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x;\theta)}{\partial \theta^2} f(x;\theta)\,dx &= \int_{-\infty}^{\infty} \left\{\frac{\partial \ln f(x;\theta)}{\partial \theta}\right\}^2 f(x;\theta)\,dx \ \Rightarrow\ \mathbb{E}(V_i) = \mathbb{E}(U_i^2) \\ &= \mathbb{V}(U_i) \ \Rightarrow\ \mathbb{V}\big(U(\theta)\big) = \mathbb{E}\big(V(\theta)\big) = n\mathbb{E}(V_i)\,.\end{aligned}$$

The quantity $I(\theta) = \mathbb{E}\big(V(\theta)\big)$ is called Fisher information.

The MLE satisfies:

$$\begin{aligned}0 = U\big(\hat{\theta}\big) = \frac{\partial \ln L\big(\hat{\theta}\big)}{\partial \theta} &\approx \frac{\partial \ln L(\theta)}{\partial \theta} + \big(\hat{\theta} - \theta\big)\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = U(\theta) - \big(\hat{\theta} - \theta\big)V(\theta) \\ \Rightarrow\ V(\theta)\big(\hat{\theta} - \theta\big) &\approx U(\theta). \qquad \text{\# Taylor's polynomial}\end{aligned}$$

As the sample size gets large, the Law of Large Numbers gives $V \to I(\theta)$, and Central Limit Theorem gives $U \approx \mathrm{N}\big(0, I(\theta)\big)$. Then we have the result $\hat{\theta} \approx \mathrm{N}(\theta, I(\theta)^{-1})$.

Furthermore, we use the normal distribution to construct approximate CI.

As the standard error of the MLE we can use:

$$\mathrm{se}\big(\hat{\theta}\big) \approx \frac{1}{\sqrt{I\big(\hat{\theta}\big)}} \approx \frac{1}{\sqrt{V\big(\hat{\theta}\big)}}\,.$$

Let $X_1, \dots, X_n$ be an exponential random sample with scale parameter θ.

$$V_i(\theta) = -\frac{\partial^2}{\partial\theta^2}\ln f(x_i|\theta) = -\frac{\partial^2}{\partial\theta^2}[-\ln\theta - x\theta^{-1}] = -\theta^{-2} + 2x_i\theta^{-3}$$

$$I_i(\theta) = \mathbb{E}\big(V_i(\theta)\big) = \mathbb{E}(-\theta^{-2} + 2X\theta^{-3}) = \theta^{-2} \;\Rightarrow\; I(\theta) = n\theta^{-2} \;\Rightarrow\; \hat{\theta} \approx N(\theta, \theta^2/n).$$

Suppose we observe $n = 20$ and $\bar{x} = 3.7$. An approximate 95% CI is

$$3.7 \pm 1.96\sqrt{3.7^2/20} = [2.1, 5.3].$$

Suppose the parameter does not define a boundary, we can find a lower bound on the variance of an unbiased estimator, known as the Cramér-Rao lower bound.

This bound is equal to the asymptotic variance of the MLE.

Let T be an unbiased estimator of θ.

$$\mathrm{Cov}(T, U) = \mathbb{E}(TU) - \mathbb{E}(T)\mathbb{E}(U) = \mathbb{E}(TU) = \int T\frac{\partial \ln L(\theta)}{\partial\theta}L(\theta)\,d\mathbf{x} = \int T\frac{\partial L(\theta)}{\partial\theta}d\mathbf{x}$$

$$= \frac{\partial}{\partial\theta}\int TL(\theta)\,d\mathbf{x} = \frac{\partial}{\partial\theta}\mathbb{E}(T) = \frac{\partial}{\partial\theta}\theta = 1$$

$$1 \geq \rho^2 = \frac{\mathrm{Cov}(T, U)^2}{\mathbb{V}(T)\mathbb{V}(U)} = \frac{1}{\mathbb{V}(T)\mathbb{V}(U)} \;\Rightarrow\; \mathbb{V}(T) \geq \frac{1}{\mathbb{V}(U)} = I(\theta)^{-1}.$$

We define the efficiency of the unbiased estimator T as its variance relative to the lower bound, i.e., $\mathrm{eff}(T) = I(\theta)^{-1}/\mathbb{V}(\theta)$.

The statistic $T = g(X_1, \dots, X_n)$ is sufficient for an underlying parameter θ if the conditional probability distribution of the data $(X_1, \dots, X_n)$, given the statistic $g(X_1, \dots, X_n)$, does not depend on the parameter θ.

Sometimes need more than one statistic, e.g., T_1 and T_2, in which case we say they are jointly sufficient for θ.

Once the sufficient statistics are known there is no additional information on the parameter in the sample. Samples that have the same values of the sufficient statistic yield the same estimates.

Factorization Theorem. Let $X_1, \dots, X_n$ has joint PDF/PMF $f(x_1, \dots, x_n; \theta)$.

$T = g(x_1, \dots, x_n)$ is sufficient for θ iff

$$f(x_1, \dots, x_n; \theta) = \phi\{T; \theta\}h(x_1, \dots, x_n). \qquad \text{\# knowning the pop. dist.}$$

In any problem there are, in fact, many sufficient statistics.

The MLE must be a function of the sufficient statistics, since it maximizes $L(\theta)$ by maximising $\phi\{T; \theta\}$.

Find the sufficient statistics for θ (where $\theta > 0$) when we observe X from the following PDFs:

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad 0 < x < \infty.$$

$$f(x_1, \dots, x_n; \theta) = \theta^{-n} e^{-\theta^{-1} \sum_{i=1}^n x_i} \prod_{i=1}^n I(x_i > 0) = \left\{\theta^{-n} e^{-\theta^{-1} \sum_{i=1}^n x_i}\right\} I\left(x_{(1)} > 0\right).$$

The sufficient statistic for θ is $Y = \sum_{i=1}^n x_i$.

We often use distributions which have PDFs of the form:

$$f(x;\theta) = \exp\{K(x)p(\theta) + S(x) + q(\theta)\}.$$

This is called the exponential family.

Let $X_1, \dots, X_n$ be i.i.d. from an exponential family. Then $\sum_{i=1}^n K(X_i)$ is sufficient for θ.

To prove this note that the joint PDF is

$$\exp\left\{p(\theta) \sum K(x_i) + \sum S(x_i) + nq(\theta)\right\}$$
$$= \left[\exp\left\{p(\theta) \sum K(x_i) + nq(\theta)\right\}\right] \exp\left\{\sum S(x_i)\right\}.$$

The factorization theorem then shows sufficiency.

Neyman-Pearson Lemma. The most powerful test for a given signif. level, is the LRT.

Asymptotic distribution of the likelihood ratio: Consider the test,

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

The likelihood ratio is, $\lambda = L(\theta_0)/L\left(\hat{\theta}\right)$. The function $2\ln(\lambda) \approx \chi_1^2$ asymptotically.

Matrix Partitioning. Partitioned matrices can be manipulated as if the submatrices were single elements using matrix multiplication, once their dimensions are compatible.

$$X = \left[\begin{array}{cc|c} 2 & 1 & 0 \\ 3 & 4 & 1 \end{array}\right] = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right], \qquad Y = \left[\begin{array}{cc} 1 & 0 \\ 2 & 4 \\ \hline 3 & -1 \end{array}\right] = \left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right]$$

$$XY = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right]\left[\begin{array}{c} Y_{11} \\ \hline Y_{21} \end{array}\right] = \left[X_{11}Y_{11} + X_{12}Y_{21}\right] = \left[\begin{array}{c} [\,2 \quad 1\,]\begin{bmatrix} 1 & 0 \\ 2 & 4 \end{bmatrix} + [\,0\,][\,3 \quad -1\,] \\ \hline [\,3 \quad 4\,]\begin{bmatrix} 1 & 0 \\ 2 & 4 \end{bmatrix} + [\,1\,][\,3 \quad -1\,] \end{array}\right]$$

Transposition. The transpose of a matrix results when the rows and columns are interchanged.

1. X is symmetric iff $X^T = X$.
2. $(X^T)^T = X$.
3. $(XY)^T = Y^T X^T$.

Identity. The matrix identity I is a square matrix of arbitrary size with 1's on the diagonal and 0's off the diagonal.

1. $XI_n = I_m X = X$ where I_k is the $k \times k$ identity matrix.

Inverse. If X is a square matrix such that $|X| \neq 0$, it is nonsingular (or invertible) then,

1. $XX^{-1} = X^{-1}X = I$.
2. $(X^{-1})^{-1} = X$.
3. $(XY)^{-1} = Y^{-1}X^{-1}$.
4. $(X^T)^{-1} = (X^{-1})^T$.

Linear independence. Suppose we have a set of vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}$.

We say that this set is linearly dependent iff there exists some numbers $a_1, a_2, \ldots, a_k$, which are not all zero, such that $a_1\mathbf{x_1} + a_2\mathbf{x_2} + \cdots + a_k\mathbf{x_k} = 0$.

If the only way in which this equation is satisfied is for all a's to be zero, then we say that the x's are linearly independent.

If a set of vectors is linearly dependent, then at least one of the vectors can be written as a linear combination of some or all of the rest.

Rank. The rank of an $n \times k$ matrix X, denoted by $r(X)$, is the dimension of column space of X, i.e., the greatest number of linearly independent vectors in the set $\{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_k}\}$; i.e., the number of nonzero rows in the (reduced) row-echelon form of X.

1. If $n \geq k$ and $r(X) = k$, we say that X is of full rank.
2. For any matrix X, we have $r(X) = r(X^T) = r(X^T X) = r(XX^T)$.
3. If P is $n \times n$ and nonsingular, and Q is $k \times k$ and nonsingular, then $r(X) = r(PX) = r(XQ)$.
4. If X is $k \times k$, then X is nonsingular iff X is of full rank.
5. $r(XY) \leq r(X), r(Y)$.

Trace. The trace of a square matrix X, denoted by $tr(X)$, is the sum of diagonal entries.

1. If c is a scalar, $tr(cX) = c\ tr(X)$.
2. $tr(X \pm Y) = tr(X) \pm tr(Y)$.
3. If XY and YX both exist, $tr(XY) = tr(YX)$.

Orthogonality. Two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal iff $\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i = 0$.

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ is called orthogonal iff $\langle \mathbf{v_i}, \mathbf{v_j} \rangle = 0 \ \ \forall\, i \neq j$. Every orthogonal set of nonzero vectors $\{\mathbf{v_1}, \dots, \mathbf{v_k}\}$ is linearly independent.

A set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ is called orthonormal iff it is orthogonal, and each vector has unit length.

An $n \times n$ square matrix X is orthogonal iff $X^T X = I$, then $X^{-1} = X^T$. The columns of X form an orthonormal basis (independently span) for $\mathbb{R}^n$.

Eigenthings. Suppose A is a $n \times n$ matrix and $\mathbf{x}$ is a $n \times 1$ nonzero vector which satisfies the equation $A\mathbf{x} = \lambda\mathbf{x}$. λ is an eigenvalue of A, with associated eigenvector $\mathbf{x}$.

To find the eigenvalues, solve the characteristic equation: $|A - \lambda I| = 0$. Substituting back can always find an infinite number of solutions for associated eigenvectors.

If A is real and symmetric, then its eigenvalues are all real, and its eigenvectors are orthogonal.

Suppose there are two eigenvectors $\mathbf{v_1}$ and $\mathbf{v_2}$ then,

$$\begin{gathered} A\mathbf{v_1} = \lambda_1\mathbf{v_1} \\ \mathbf{v_2}^T A\mathbf{v_1} = \lambda_1 \mathbf{v_2}^T\mathbf{v_1} \\ (\mathbf{v_2}^T A\mathbf{v_1})^T = (\lambda_1 \mathbf{v_2}^T\mathbf{v_1})^T \\ \mathbf{v_1}^T A^T\mathbf{v_2} = \lambda_1 \mathbf{v_1}^T\mathbf{v_2} \\ \mathbf{v_1}^T A\mathbf{v_2} = \lambda_1 \mathbf{v_1}^T\mathbf{v_2} \end{gathered} \qquad \begin{gathered} A\mathbf{v_2} = \lambda_2\mathbf{v_2} \\ \mathbf{v_1}^T A\mathbf{v_2} = \lambda_2 \mathbf{v_1}^T\mathbf{v_2} \end{gathered} \qquad \Rightarrow \qquad 0 = (\lambda_1 - \lambda_2)\mathbf{v_1}^T\mathbf{v_2} \ \Rightarrow \ \mathbf{v_1}^T\mathbf{v_2} = 0$$

Orthogonal diagonalization. Let A be a symmetric $k \times k$ matrix. Then an orthogonal matrix P exists such that

$$P^T AP = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix},$$

where λ_i , $i = 1, 2, \dots, k$, are the eigenvalues of A, and the columns of P are eigenvectors of A associated with the respective eigenvalues. We say that P diagonalizes A.

Let $A_1, A_2, \dots, A_m$ be a collection of symmetric $k \times k$ matrices. Then the following are equivalent:

1. There exists an orthogonal matrix P such that $P^T A_i P$ is diagonal for all $i = 1, 2, \dots, m$;
2. $A_i A_j = A_j A_i$ for every pair $i, j = 1, 2, \dots, m$.

Idempotence. A square matrix A is idempotent iff $A^2 = A$.

1. The only nonsingular idempotent matrices are the identity matrices.
2. $I - A$ is idempotent iff A is idempotent.
3. A (symmetric) matrix is idempotent iff all it eigenvalues are either 0 or 1.

   Let A be an idempotent matrix with eigenvalue λ and associated eigenvector $\mathbf{x}$.

   $(\Rightarrow)$ $\lambda^2\mathbf{x} = \lambda A\mathbf{x} = A\lambda\mathbf{x} = A^2\mathbf{x} = A\mathbf{x} = \lambda\mathbf{x} \;\Rightarrow\; (\lambda^2 - \lambda)\mathbf{x} = 0.$

   $(\Leftarrow)$ …
4. Iff A is symmetric and idempotent matrix, $r(A) = r(\Lambda) = tr(\Lambda) = tr(A)$.

   Using diagonalization, since P is orthogonal, both P and P^T are nonsingular. Then $r(A) = r(P^T A) = r(P^T AP)$. $P^T AP$ is diagonal, $r(P^T AP)$ is the number of nonzero eigenvalues of A. But A is idempotent, so its eigenvalues are either 0 or 1.

   Therefore, $r(A) = tr\big((P^T A)P\big) = tr\big(P(P^T A)\big) = tr(A)$.
5. Let $A_1, A_2, \dots, A_m$ be a collection of symmetric $k \times k$ matrices. Then any two of the following conditions implies the third:
   - All $A_i, i = 1, 2, \dots, m$ are idempotent;
   - $\sum_{i=1}^m A_i$ is idempotent;
   - $A_i A_j = 0$ for all $i \neq j$.
6. If the conditions in (5) are true, then $r(\sum_{i=1}^m A_i) = \sum_{i=1}^m r(A_i)$.

   As a sum of symmetric matrices, $\sum_{i=1}^m A_i$ is also symmetric. Thus $r(\sum_{i=1}^m A_i) = tr(\sum_{i=1}^m A_i) = \sum_{i=1}^m tr(A_i) = \sum_{i=1}^m r(A_i)$.

Quadratic forms. Let A be a $n \times n$ matrix and $\mathbf{y}$ a $n \times 1$ vector containing variables. The scalar quantity

$$q = \mathbf{y}^T A \mathbf{y} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} y_i y_j$$

is called a quadratic form in $\mathbf{y}$, and A is called the matrix of the quadratic form.

If $\mathbf{y}^T A \mathbf{y} > 0 \ \ \forall \, \mathbf{y} \neq \mathbf{0}$, then A and the quadratic form are positive definite.

If $\mathbf{y}^T A \mathbf{y} \geq 0 \ \ \forall \, \mathbf{y}$, then A and the quadratic form are positive semidefinite.

A symmetric matrix A is positive definite iff all of its eigenvalues are strictly positive. A symmetric matrix A is positive semidefinite iff all of its eigenvalues are non-negative.

Let $\lambda_1, \dots, \lambda_n \geq 0$ be the eigenvalues of A. For any $\mathbf{x}$ we have, for $\mathbf{z} = P^T \mathbf{x} = (z_1, \dots, z_n)^T$,

$$(\Leftarrow) \qquad \mathbf{x}^T A \mathbf{x} = \mathbf{x}^T P \Lambda P^T \mathbf{x} = \mathbf{z}^T \Lambda \mathbf{z} = \sum_{i=1}^{n} z_i^2 \lambda_i \geq 0.$$

Let A be positive semidefinite. Let $\mathbf{x_i}$ be its normalised i'th eigenvector, then

$$(\Rightarrow) \qquad 0 \leq \mathbf{x_i}^T A \mathbf{x_i} = \mathbf{x_i}^T (A \mathbf{x_i}) = \mathbf{x_i}^T \lambda_i \mathbf{x_i} = \lambda_i \mathbf{x_i}^T \mathbf{x_i} = \lambda_i.$$

$X^T X$ is positive semidefinite. An inverse of positive definite is also positive definite.

Differentiation of quadratic forms. Suppose we have a vector of variables $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, and some scalar function of them: $z = f(\mathbf{y})$.

We define the partial derivative of z with respect to $\mathbf{y}$ as: $\partial z / \partial \mathbf{y} = \begin{bmatrix} \partial z / \partial y_1 \\ \partial z / \partial y_2 \\ \dots \\ \partial z / \partial y_n \end{bmatrix}$.

1. If $z = \mathbf{a}^T \mathbf{y}$ where $\mathbf{a}$ is a vector of constants, then $\partial z / \partial \mathbf{y} = \mathbf{a}$.
2. If $z = \mathbf{y}^T \mathbf{y}$, then $\partial z / \partial \mathbf{y} = 2\mathbf{y}$.
3. If $z = \mathbf{y}^T A \mathbf{y}$, then $\partial z / \partial \mathbf{y} = A\mathbf{y} + A^T \mathbf{y}$. In particular, if A is symmetric, then $\partial z / \partial \mathbf{y} = 2A\mathbf{y}$.

   The derivative with respect to the k'th variable is then by product rule:

$$\begin{aligned} \frac{\partial}{\partial y_k} [\mathbf{y}^T A \mathbf{y}] &= \frac{\partial}{\partial y_k} \left[\sum_{i=1}^{n} y_i \sum_{j=1}^{n} a_{ij} y_j \right] \\ &= \sum_{i=1}^{n} \left(\frac{\partial}{\partial y_k} [y_i] \sum_{j=1}^{n} a_{ij} y_j + y_i \frac{\partial}{\partial y_k} \left[\sum_{j=1}^{n} a_{ij} y_j \right] \right) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial y_k} [y_i] \sum_{j=1}^{n} a_{ij} y_j + \sum_{i=1}^{n} y_i \sum_{j=1}^{n} a_{ij} \frac{\partial}{\partial y_k} [y_j] \\ &= \sum_{j=1}^{k} a_{kj} y_j + \sum_{i=1}^{n} a_{ik} y_i. \end{aligned}$$

If then you arrange these derivatives into a column vector, you get:

$$\begin{bmatrix} \sum_{j=1}^{k} a_{1j}y_j + \sum_{i=1}^{n} a_{i1}y_i \\ \vdots \\ \sum_{j=1}^{k} a_{1j}y_j + \sum_{i=1}^{n} a_{i1}y_i \end{bmatrix} = A\mathbf{y} + A^T\mathbf{y}.$$

Matrix square root of a matrix A is a matrix B such that $B^2 = A$. In general, the square root is not unique.

If A is symmetric and positive semidefinite, there is a unique symmetric positive semidefinite square root, called the principal root, denoted $A^{1/2}$.

Suppose that P diagonalises A, i.e., $P^T AP = \Lambda$. Then $A = P\Lambda P^T = P\Lambda^{1/2}\Lambda^{1/2}P^T = P\Lambda^{1/2}I\Lambda^{1/2}P^T = P\Lambda^{1/2}P^T P\Lambda^{1/2}P^T$, so $A^{1/2} = P\Lambda^{1/2}P^T$.

Define the expectation of a random vector $\mathbf{y}$ to be:

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}] = \begin{bmatrix} \mathbb{E}[y_1] \\ \mathbb{E}[y_2] \\ \vdots \\ \mathbb{E}[y_n] \end{bmatrix} \text{ if } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

1. If $\mathbf{a}$ is a vector of constants, then $\mathbb{E}[\mathbf{a}] = \mathbf{a}$, and $\mathbb{E}[\mathbf{a}^T\mathbf{y}] = \mathbf{a}^T\mathbb{E}[\mathbf{y}]$.
2. If A is a matrix of constants, then $\mathbb{E}[A\mathbf{y}] = A\mathbb{E}[\mathbf{y}]$.

Define the variance (or covariance matrix) of a random vector $\mathbf{y}$ to be:

$$\mathbb{V}[\mathbf{y}] = \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T].$$

Suppose that $\mathbf{y}$ is an arbitrary random vector with $\mathbb{V}[\mathbf{y}] = V$. Then:

1. V is always symmetric.
2. If $\mathbf{a}$ is a vector of constants, then $\mathbb{V}[\mathbf{a}^T\mathbf{y}] = \mathbf{a}^T V\mathbf{a}$.
3. If A is a matrix of constants, then $\mathbb{V}[A\mathbf{y}] = AVA^T$.

$$\mathbb{V}[A\mathbf{y}] = \mathbb{V}[\,[\,\mathbf{a_1}\ \cdots\ \mathbf{a_n}\,]\,\mathbf{y}\,] = \mathbb{V}\left[\begin{bmatrix} \mathbf{a_1^T} \\ \vdots \\ \mathbf{a_n^T} \end{bmatrix}^T \mathbf{y}\right] = \begin{bmatrix} \mathbf{a_1^T} \\ \vdots \\ \mathbf{a_n^T} \end{bmatrix}^T V \begin{bmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{a_n} \end{bmatrix} = AVA^T.$$

4. V is positive semidefinite, since $\mathbf{u}^T V\mathbf{u} = \mathbb{V}[\mathbf{u}^T\mathbf{y}] \geq 0$.

Let $\mathbf{z}$ be a $k \times 1$ vector of independent standard normal random variables, A an $n \times k$ matrix, and $\mathbf{b}$ an $n \times 1$ vector. We define that

$$\mathbf{x} = A\mathbf{z} + \mathbf{b} \sim \text{MVN}(\boldsymbol{\mu}, \Sigma)$$

has (an n-dimensional) multivariate normal distribution, with mean $\boldsymbol{\mu} = \mathbb{E}[A\mathbf{z}] + \mathbb{E}[\mathbf{b}] = A\mathbf{0} + \mathbf{b} = \mathbf{b}$ and covariance matrix $\Sigma = \mathbb{V}[A\mathbf{z} + \mathbf{b}] = A\mathbb{V}[\mathbf{z}]A^T = AIA^T = AA^T$.

In particular, for any $\boldsymbol{\mu}$ and any symmetric positive semidefinite matrix Σ, let $\mathbf{z}$ be a vector of independent standard normals. We can construct

$$\boldsymbol{\mu} + \Sigma^{1/2}\mathbf{z} \sim \text{MVN}(\boldsymbol{\mu}, \Sigma), \qquad \mathbf{z} \sim \text{MVN}(\mathbf{0}, I).$$

Then any linear combination of multivariate normals results in another multivariate normal: if $\mathbf{x} \sim \text{MVN}(\boldsymbol{\mu}, \Sigma)$ is $k \times 1$, A is $n \times k$, and $\mathbf{b}$ is $n \times 1$, then

$$\mathbf{y} = A\mathbf{x} + \mathbf{b} \sim \text{MVN}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T). \qquad \text{\# the general rule}$$

Put $\mathbf{x} = \boldsymbol{\mu} + \Sigma^{1/2}\mathbf{z}$, then $\mathbf{y} = A\left(\boldsymbol{\mu} + \Sigma^{1/2}\mathbf{z}\right) + \mathbf{b} = A\Sigma^{1/2}\mathbf{z} + (A\boldsymbol{\mu} + \mathbf{b})$.

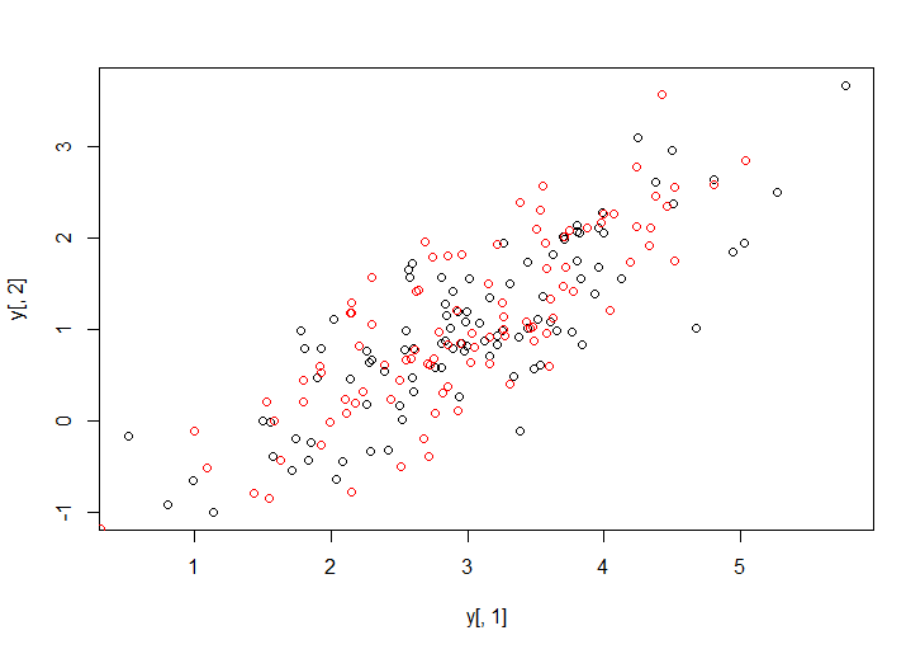
By definition, $\mathbf{y} \sim \text{MVN}\left(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma^{1/2}\left(A\Sigma^{1/2}\right)^T\right) \sim \text{MVN}\left(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma^{1/2}\left(\Sigma^{1/2}\right)^T A^T\right)$.

Since $\left(\Sigma^{1/2}\right)^T = \left(P\Lambda^{1/2}P^T\right)^T = (P^T)^T\left(\Lambda^{1/2}\right)^T P^T = P\Lambda^{1/2}P^T = \Sigma^{1/2}$, then $\mathbf{y} \sim \text{MVN}\left(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma^{1/2}\Sigma^{1/2}A^T\right) \sim \text{MVN}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

If $\mathbf{x} \sim \text{MVN}(\boldsymbol{\mu}, \Sigma)$ and Σ is $k \times k$ positive definite, then $\mathbf{x}$ has the density:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

```
## Generate a sample of size 100 with distribution MVN(a,V)
> a <- as.vector(c(3,1))
> V <- matrix(c(1,.8,.8,1), 2, 2, byrow = FALSE)

> library(MASS)
> y <- mvrnorm(100, mu = a, Sigma = V)
> plot(y[,1], y[,2])

# manually
> P <- eigen(V)$vectors
> sqrtV <- P %*% diag(sqrt(eigen(V)$values)) %*% t(P)
> z <- matrix(rnorm(200), 2, 100)
> y_new <- sqrtV %*% z + rep(a, 100)
> points(y_new[1,], y_new[2,], col = "red")
```



If $\mathbf{z} = (z_1, z_2)^T$ is multivariate normal, then z_1 and z_2 are independent iff they are uncorrelated.

In general, suppose that z_1 and z_2 are normal random variables, $\mathbf{z} = (z_1, z_2)^T$ does not have to be multivariate normal, if z_1 and z_2 are uncorrelated but not independent.

For example, $u \sim \text{U}(-1, 1)$, $z_1 \sim \text{N}(0, 1)$, and $z_2 = z_1 \,\text{sgn}(u) \sim \text{N}(0, 1)$.

Let $\mathbf{y}$ be a random vector with $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\mathbb{V}(\mathbf{y}) = V$, and let A be a matrix of constants. Then the expectation of the random quadratic form is

$$\begin{aligned}\mathbb{E}[\mathbf{y}^T A\mathbf{y}] &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}y_i y_j\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\mathbb{E}[y_i y_j] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(\sigma_{ij} + \mu_i\mu_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\sigma_{ji} + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\mu_i\mu_j = tr(AV) + \boldsymbol{\mu}^T A\boldsymbol{\mu}.\end{aligned}$$

```
> library(MASS)
> mu <- c(1,3)
> V <- matrix(c(2,1,1,5), 2, 2)
> y <- t(mvrnorm(100, mu = mu, Sigma = V))
> A <- matrix(c(4,1,1,2), 2, 2)
> sum(diag(A %*% V)) + drop(t(mu) %*% A %*% mu)
[1] 48
> quadform <- function(y, A) { t(y) %*% A %*% y }
> mean(apply(y, 2, quadform, A = A))
[1] 46.26975
```

Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, I_k)$ be a $k \times 1$ random vector. Then

$$x = \mathbf{y}^T\mathbf{y} = \sum_{i=1}^{k} y_i^2 \sim \chi^2_{k,\lambda}$$

has a noncentral χ^2 distribution with k d.f. and noncentrality parameter $\lambda = \boldsymbol{\mu}^T\boldsymbol{\mu}/2 > 0$. Note that the distribution of x depends on $\boldsymbol{\mu}$ only through λ.

The density of the noncentral χ^2 distribution is:

$$f(x; k, \lambda) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} g(x; k + 2i)$$

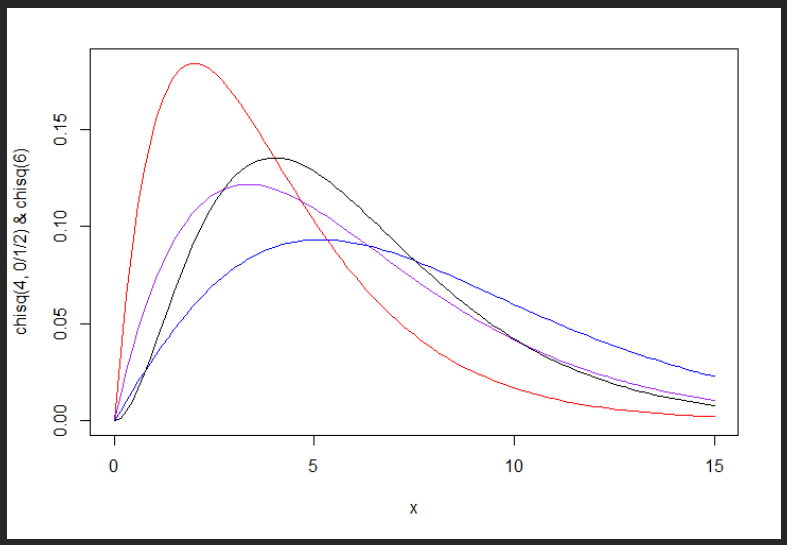
where g is the density of the (ordinary) χ^2 distribution:

$$g(x; k) = \frac{1}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)} x^{k/2-1}e^{-x/2} = f(x; k, 0), \quad x > 0.$$

The expectation of x is $\mathbb{E}[x] = \mathbb{E}[\mathbf{y}^T\mathbf{y}] = tr(I_k I_k) + \boldsymbol{\mu}^T I_k\boldsymbol{\mu} = k + 2\lambda$.

$\mathbb{V}[x]$ can also be shown to be $2k + 8\lambda$. # proof?

```
> curve(dchisq(x,4,0), col = "red")    # R uses λ = μ^T μ
> curve(dchisq(x,4,2), add = TRUE, col = "purple")
> curve(dchisq(x,4,4), add = TRUE, col = "blue")
> curve(dchisq(x,6,0), add = TRUE, col = "black")
```



Let $X^2_{k_1,\lambda_1}, \ldots, X^2_{k_n,\lambda_n}$ be a collection of n independent noncentral χ^2 random variables. Then $\sum_{i=1}^n X^2_{k_i,\lambda_i}$ has a noncentral χ^2 distribution with $\sum_{i=1}^n k_i$ d.f. and noncentrality parameter $\sum_{i=1}^n \lambda_i$. # proof?

Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, I_n)$ be a $n \times 1$ random vector and let A be a $n \times n$ symmetric matrix. Then $\mathbf{y}^T A\mathbf{y}$ has a noncentral χ^2 distribution with k d.f. and noncentrality parameter $\lambda = \boldsymbol{\mu}^T A\boldsymbol{\mu}/2$ iff A is idempotent and has rank k. # ⇒ ?

(⇐) Since A is idempotent and symmetric, we know that $P^T AP = \Lambda$, where all its eigenvalues are either 0 or 1, also $tr(\Lambda) = r(A) = k$.

Now arrange the columns of P so that all the 1 eigenvalues are first. Then we can partition the diagonalisation of A as:

$$\Lambda = P^T AP = \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right].$$

Now define the random vector $\mathbf{z} = P^T\mathbf{y} \sim \text{MVN}(P^T\boldsymbol{\mu}, I_n)$. Partition the matrices as

$$\mathbf{z} = \left[\begin{array}{c} \mathbf{z_1} \\ \hline \mathbf{z_2} \end{array}\right], \qquad P = [\, P_1 \mid P_2 \,]$$

where $\mathbf{z_1}$ is $k \times 1$ and P_1 is $n \times k$. Then $\mathbf{z_1} = P_1^T\mathbf{y} \sim \text{MVN}(P_1^T\boldsymbol{\mu}, I_k)$, and

$$\mathbf{y}^T A\mathbf{y} = (P\mathbf{z})^T A(P\mathbf{z}) = \mathbf{z}^T P^T APz = [\, \mathbf{z_1}^T \mid \mathbf{z_2}^T \,] \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} \mathbf{z_1} \\ \hline \mathbf{z_2} \end{array}\right] = \mathbf{z_1}^T\mathbf{z_1}.$$

Therefore, $\mathbf{y}^T A\mathbf{y} = \mathbf{z_1}^T\mathbf{z_1}$ has a noncentral χ^2 distribution with k d.f. and noncentrality parameter

$$\lambda = \frac{1}{2}\boldsymbol{\mu}^T P_1 P_1^T \boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}^T A\boldsymbol{\mu},$$

since

$$A = P\Lambda P^T = [\, P_1 \mid P_2 \,] \left[\begin{array}{c|c} I_k & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} P_1^T \\ \hline P_2^T \end{array}\right] = P_1 P_1^T.$$

(Corollary.) Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, \sigma^2 I_n)$ be a $n \times 1$ random vector and let A be a $n \times n$ symmetric matrix. Then $\mathbf{y}^T A\mathbf{y}/\sigma^2$ has a noncentral χ^2 distribution with k d.f. and noncentrality parameter $\lambda = \boldsymbol{\mu}^T A\boldsymbol{\mu}/2\sigma^2$ iff A is idempotent and has rank k.

Hint. Define the random vector $\mathbf{z} = \sigma^{-1}\mathbf{y} \sim \text{MVN}(\sigma^{-1}\boldsymbol{\mu}, I_n)$.

Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector, and let A be a $n \times n$ symmetric matrix. Then $\mathbf{y}^T A\mathbf{y}$ has a noncentral χ^2 distribution with k d.f. and noncentrality parameter $\lambda = \boldsymbol{\mu}^T A\boldsymbol{\mu}/2$ iff AV is idempotent and has rank k.

Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector with nonsingular variance V, and let A and B be symmetric $n \times n$ matrices. Then the two quadratic forms $\mathbf{y}^T A\mathbf{y}$ and $\mathbf{y}^T B\mathbf{y}$ are independent iff $AVB = 0$.

($\Leftarrow$) Suppose that $AVB = 0$. Since V is symmetric and positive definite (can be shown from nonsingularity and positive semidefiniteness), we have $V = C^2$ for some unique positive definite C.

Let $R = CAC$, $S = CBC$, then $RS = CAVBC = 0$. Because A, B, C are symmetric, R, S are also symmetric, thus $SR = S^T R^T = (RS)^T = 0 = RS$. Then, *we can find an orthogonal matrix P which diagonalises R and S simultaneously.

Since C is nonsingular, $r(R) = r(CAC) = r(A)$. Thus

$$P^T RP = \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & 0 \end{array}\right]$$

where D_1 has dimension $r(A) \times r(A)$. Because $RS = 0$, it can be shown that

$$P^T SP = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & D_2 \end{array}\right]$$

where the partition has the same dimension.

Now define $\mathbf{z} = P^T C^{-1}\mathbf{y}$. Then $\mathbf{z}$ is multivariate normal with

$$\mathbb{E}[\mathbf{z}] = P^T C^{-1}\boldsymbol{\mu}, \qquad \mathbb{V}[\mathbf{z}] = P^T C^{-1} V (P^T C^{-1})^T = P^T (C^{-1} V C^T) P = P^T P = I.$$

From this, observe that the elements of $\mathbf{z}$ are independent.

Partition $\mathbf{z}$ into $\mathbf{z} = \left[\begin{array}{c} \mathbf{z_1} \\ \hline \mathbf{z_2} \end{array}\right]$ where $\mathbf{z_1}$ has dimension $r(A) \times 1$.

By rewriting our equation, we see that $\mathbf{y} = CP\mathbf{z}$, $A = C^{-1} R C^{-1}$, $B = C^{-1} S C^{-1}$. So

$$\mathbf{y}^T A\mathbf{y} = \mathbf{z}^T P^T C C^{-1} R C^{-1} CP\mathbf{z} = \mathbf{z}^T P^T RP\mathbf{z} = \left[\, \mathbf{z_1}^T \,\middle|\, \mathbf{z_2}^T \,\right] \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} \mathbf{z_1} \\ \hline \mathbf{z_2} \end{array}\right] = \mathbf{z_1}^T D_1 \mathbf{z_1}$$

and similarly, $\mathbf{y}^T B\mathbf{y} = \mathbf{z_2}^T D_2 \mathbf{z_2}$.

But $\mathbf{z_1}$ and $\mathbf{z_2}$ are mutually independent of each other, since all elements of $\mathbf{z}$ are independent. Therefore, $\mathbf{y}^T A\mathbf{y}$ and $\mathbf{y}^T B\mathbf{y}$ are independent.

Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector, and let A be a $n \times n$ symmetric matrix and B a $m \times n$ matrix. Then $\mathbf{y}^T A\mathbf{y}$ and $B\mathbf{y}$ are independent iff $BVA = 0$.

Take a normal sample of size n. Prove that $\bar{y}$ and s^2 are independent.

Hint. $\bar{y} = n^{-1}\mathbf{1}^T\mathbf{y}$, $s^2 = (n-1)^{-1}\mathbf{y}^T(I - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbf{y}$.

(Combined.) Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, I)$ be a random vector, and let $A_1, \ldots, A_m$ be a set of symmetric matrices. If any two of the following statements are true:

- All A_i are idempotent;
- $\sum_{i=1}^m A_i$ is idempotent;
- $A_i A_j = 0$ for all $i \neq j$;

then so is the third, and

- For all i, $\mathbf{y}^T A_i \mathbf{y}$ has a noncentral χ^2 distribution with $r(A_i)$ d.f. and noncentrality parameter $\lambda_i = \boldsymbol{\mu}^T A_i \boldsymbol{\mu}/2$;
- $\mathbf{y}^T A_i \mathbf{y}$ and $\mathbf{y}^T A_j \mathbf{y}$ are independent for $i \neq j$; and
- $\sum_{i=1}^m r(A_i) = r(\sum_{i=1}^m A_i)$. # this can replace one of the two conditions above

When $\sum_i A_i = I$ (idempotent), the previous result is related to the following theorem.

Cochran-Fisher Theorem. Let $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, \sigma I^2)$ be a $n \times 1$ random vector. Decompose the sum of squares of $\mathbf{y}/\sigma$ into the quadratic forms

$$\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{y} = \sum_{i=1}^m \frac{1}{\sigma^2}\mathbf{y}^T A_i \mathbf{y}.$$

Then the quadratic forms are independent and have noncentral χ^2 distributions with parameters $r(A_i)$ and $\boldsymbol{\mu}^T A_i \boldsymbol{\mu}/2\sigma^2$, respectively, iff

$$\sum_{i=1}^m r(A_i) = r\left(\sum_i A_i\right) = r(I) = n.$$

The linear model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

for all $i = 1, 2, \ldots, n$, or

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} \quad = \quad X \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\epsilon}$$

where $\mathbf{y}$ and $\boldsymbol{\epsilon}$ (assume mean $\mathbf{0}$ and variance $\sigma^2 I$) are random vectors, and $\boldsymbol{\beta}$ is the vector of parameters. Although it is common for X to be a measurement, we assume that there is no uncertainty/error in these measurements.

Therefore, this model gives that $\mathbb{E}[\mathbf{y}] = X\boldsymbol{\beta}$.

A model is said to be of full rank when the design matrix X has full rank, i.e., $r(X) = p = k + 1$. Then, it means $X^T X$ is nonsingular and $(X^T X)^{-1}$ exists.

For this section and the next, we assume that the design matrix X is of full rank.

Suppose that $\mathbf{b} = (b_0, \ldots, b_k)^T$ is an estimate of $\boldsymbol{\beta}$, then $\hat{\mathbf{y}} = \widehat{\mathbb{E}[\mathbf{y}]} = X\mathbf{b}$.

The i'th residual is defined as $e_i = y_i - \widehat{\mathbb{E}[y_i]}$.

If our estimates are good, the residuals should be very close to the error, and has mean close to $\mathbf{0}$:

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbb{E}[\mathbf{y}]} \approx \mathbf{y} - \mathbb{E}[\mathbf{y}] = \boldsymbol{\epsilon}.$$

We choose our estimates to minimise the sum of the squares of the residuals. This is the method of least squares estimation:

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\mathbf{b} - \mathbf{b}^T X^T\mathbf{y} + \mathbf{b}^T X^T X\mathbf{b}$$
$$= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T X\mathbf{b} + \mathbf{b}^T X^T X\mathbf{b} = \mathbf{y}^T\mathbf{y} - 2(X^T\mathbf{y})^T\mathbf{b} + \mathbf{b}^T(X^T X)\mathbf{b}$$

$$\Rightarrow \quad \frac{\partial}{\partial \mathbf{b}}[SS_{Res}] = \mathbf{0} - 2X^T\mathbf{y} + (X^T X)\mathbf{b} + (X^T X)^T\mathbf{b} = \mathbf{0} \quad \Rightarrow \quad \mathbf{b} = (X^T X)^{-1}X^T\mathbf{y}.$$

```
> y <- c(50,40,52,47,65)
> X <- matrix(c(rep(1,5),1,5,5,10,20,1,1,2,2,3), 5, 3)
> b <- solve(t(X)%*%X, t(X)%*%y)
> str(b)
 num [1:3, 1] 33.06 -0.19 10.72
```

The least squares estimator is unbiased since $\mathbb{E}[\mathbf{b}] = (X^T X)^{-1}X^T\mathbb{E}[\mathbf{y}] = (X^T X)^{-1}X^T X\boldsymbol{\beta} = \boldsymbol{\beta}$. Furthermore, $\mathbb{V}[\mathbf{b}] = (X^T X)^{-1}X^T\mathbb{V}[\mathbf{y}]((X^T X)^{-1}X^T)^T = (X^T X)^{-1}\sigma^2$.

Linear estimators are those which take the form $L\mathbf{y}$, where L is a matrix of constants.

If $\mathbb{E}[\mathbf{b}] = \boldsymbol{\beta}$ and variances of $b_0, \dots, b_k$ are minimised over all estimators, then $\mathbf{b}$ is called a best linear unbiased estimator (or BLUE) of $\boldsymbol{\beta}$.

Gauss-Markov Theorem. In the full rank general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the least squares estimator $\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y}$ is the unique BLUE for $\boldsymbol{\beta}$.

Suppose we have another unbiased linear estimator for $\boldsymbol{\beta}$, written as

$$\mathbf{b}^* = ((X^TX)^{-1}X^T + B)\mathbf{y}$$

where B is a $p \times n$ matrix. Then $\mathbb{E}[\mathbf{b}^*] = ((X^TX)^{-1}X^T + B)X\boldsymbol{\beta} = (I + BX)\boldsymbol{\beta}$.

Since $\mathbf{b}^*$ is an unbiased estimator for $\boldsymbol{\beta}$, implies that $(I + BX)\boldsymbol{\beta} = \boldsymbol{\beta}$, then $BX = 0$.

Now look at the variance of $\mathbf{b}^*$:

$$\begin{aligned}\mathbb{V}[\mathbf{b}^*] &= ((X^TX)^{-1}X^T + B)\sigma^2 I((X^TX)^{-1}X^T + B)^{\boldsymbol{T}} \\ &= \sigma^2((X^TX)^{-1} + 2BX(X^TX)^{-1} + BB^T) = (X^TX)^{-1}\sigma^2 + BB^T\sigma^2 \\ &= \mathbb{V}[\mathbf{b}] + BB^T\sigma^2.\end{aligned}$$

$$\mathbb{V}[b_i^*] = \left[\mathbb{V}[\mathbf{b}^*]\right]_{ii} = \mathbb{V}[b_i] + \sigma^2 \sum_{j=1}^{n} B_{Ij}^2.$$

The minimum is obtained iff $B_{ij} = 0$ for all i, j, in which case $\mathbf{b}^* = \mathbf{b}$.

Take the full rank general linear model $\mathbf{y} = X\boldsymbol{\beta} + \epsilon$ and let $\mathbf{t}$ be a $p \times 1$ vector of constants. Then the BLUE for $\mathbf{t}^T\boldsymbol{\beta}$ is $\mathbf{t}^T\mathbf{b}$, where $\mathbf{b}$ is the least squares estimator for $\boldsymbol{\beta}$.

Define $H = X(X^TX)^{-1}X^T$, then $\hat{\mathbf{y}} = X\mathbf{b} = H\mathbf{y}$. H is symmetric and idempotent.

The sample variance

$$s^2 = SS_{Res}/(n-p) = (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b})/(n-p) \qquad \text{\# versus } S^2?$$

is an unbiased estimator for σ^2.

$$\begin{aligned}(n-p)\mathbb{E}[s^2] = \mathbb{E}[SS_{Res}] &= \mathbb{E}[(\mathbf{y} - H\mathbf{y})^T(\mathbf{y} - H\mathbf{y})] = \mathbb{E}[\mathbf{y}^T(I-H)^T(I-H)\mathbf{y}] \\ &= \mathbb{E}[\mathbf{y}^T(I-H)\mathbf{y}] = tr\big((I-H)\sigma^2 I_n\big) + \boldsymbol{\beta}^T X^T(I-H)X\boldsymbol{\beta} \\ &= \sigma^2\big(tr(I-H)\big) + 0 = \sigma^2\big(r(I_n) - r(H)\big) = \sigma^2(n-p) \;\Rightarrow\; \mathbb{E}[s^2] = \sigma^2.\end{aligned}$$

```
> e <- y - X%*%b
> str(e)
 num [1:5, 1] 6.41 -2.83 -1.55 -5.6 3.58
> (SSRes <- sum(e^2))
[1] 95.67587
> (s2 <- SSRes/(5-3))
[1] 47.83794
> diag(solve(t(X)%*%X))*s2   # estimated variances of parameter estimators
[1] 110.388463 1.233391 94.618683
```

The leverage of point i is H_{ii}. The size of H_{ii} reflects how much $\hat{y}_i$ is based on y_i, as opposed to other y_j. The points with large leverage have an unusually large effect on the estimated parameters.

The variance of residuals is

$$\mathbb{V}[\mathbf{e}] = \mathbb{V}[\mathbf{y} - H\mathbf{y}] = \mathbb{V}[(I - H)\mathbf{y}] = (I - H)\sigma^2 I(I - H)^T = \sigma^2(I - H),$$

where $I - H$ is symmetric and idempotent.

Since we don't know σ^2, we use s^2 instead. This creates the standardised residuals:

$$z_i = e_i/\sqrt{s^2(1 - H_{ii})}.$$

How is the variance of residual for design variables (row) "away from the centre"?

If there is a large leverage combined with an extreme residual, then the corresponding point may distort the fit.

To check this, we calculate the Cook's distance of each point. This measures the change in the estimated parameters $\mathbf{b}$ if we remove the point:

$$D_i = \frac{\left(\mathbf{b}_{(-i)} - \mathbf{b}\right)^T X^T X\left(\mathbf{b}_{(-i)} - \mathbf{b}\right)}{ps^2} = p^{-1} z_i^2 \left(\frac{H_{ii}}{1 - H_{ii}}\right),$$

where $\mathbf{b}_{(-i)}$ is the estimated parameters if point i is removed.

It is generally considered large if it is greater than 1, and small if it is less than 0.5.

From now on, we assume the normality of errors, i.e., $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2 I)$.

To find MLEs, choose parameter values to maximise the likelihood of the observed values of the response:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} \left(\sigma\sqrt{2\pi}\right)^{-1} e^{-\epsilon_i^2/2\sigma^2} = (2\pi\sigma^2)^{-n/2} e^{-\sum_{i=1}^{n} \epsilon_i^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})/2\sigma^2}. \end{aligned}$$

Then the MLE for $\boldsymbol{\beta}$ is also the least squares estimator. The MLE for σ^2 is given by $\hat{\sigma}^2 = SS_{Res}/n$. This is a biased estimator, but has the same asymptotic properties as the sample variance.

Furthermore, the least squares estimators

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y} \quad \text{and} \quad s^2 = SS_{Res}/(n - p)$$

are jointly sufficient for $\boldsymbol{\beta}$ and σ^2, and they have the lowest variance among all unbiased estimators. This is a stronger condition than BLUE because it includes non-linear estimators. We call this UMVUE (uniformly minimum variance unbiased estimator).

Many of previous random vectors are just linear combinations of $\boldsymbol{\epsilon}$, so they also have multivariate normal distributions.

1. $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 I)$.
2. $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y} \sim \text{MVN}(\boldsymbol{\beta}, (X^T X)^{-1}\sigma^2)$.
3. $\hat{\mathbf{y}} = X\mathbf{b} \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 H)$.
4. $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \sim \text{MVN}\big(\mathbf{0}, \sigma^2(I - H)\big)$.
5. Differently, $(n-p)s^2/\sigma^2 = SS_{Res}/\sigma^2 \sim \chi^2_{n-p}$. # gives CI for σ^2

   Have shown earlier that $SS_{Res} = \mathbf{y}^T(I - H)\mathbf{y}$, where $I - H$ is symmetric and idempotent, rank = trace $= n - p$. And by assumption, $\mathbf{y} \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 I)$.

   Then $\mathbf{y}^T(I - H)\mathbf{y}/\sigma^2$ has a noncentral χ^2 distribution with $n - p$ d.f. and $\lambda = (X\boldsymbol{\beta})^T(I - H)X\boldsymbol{\beta}/2\sigma^2 = (X\boldsymbol{\beta})^T(X\boldsymbol{\beta} - HX\boldsymbol{\beta})/2\sigma^2 = 0$.
6. $\mathbf{b}$ and s^2 are independent, since $BVA = (X^T X)^{-1} X^T \sigma^2 I(I - H)/\sigma^2 = 0$.

We can now create CIs for the parameters. Consider the covariance matrix of $\mathbf{b}$:

$$(X^T X)^{-1}\sigma^2 = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{11} & c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{bmatrix} \sigma^2.$$

The least squares estimator of β_i is b_i. The variance of b_i is the i'th diagonal element of the covariance matrix, denoted $c_{ii}\sigma^2$.

Then the pivot is

$$\left(\frac{b_i - \beta_i}{\sigma\sqrt{c_{ii}}}\right) \Big/ \left(\sqrt{\frac{SS_{Res}/\sigma^2}{n-p}}\right) = \frac{b_i - \beta_i}{s\sqrt{c_{ii}}} \sim t_{n-p}.$$

We also want to predict the value of the "expected" response for a given set of inputs.

Let $\mathbf{x}^* = (1, x_1^*, \ldots, x_k^*)^T$, we can show that $(\mathbf{x}^*)^T\mathbf{b} \sim \text{MVN}(\mathbb{E}[y^*], (\mathbf{x}^*)^T(X^T X)^{-1}\mathbf{x}^*\sigma^2)$ is the BLUE for $\mathbb{E}[y^*] = (\mathbf{x}^*)^T\boldsymbol{\beta}$.

Then the pivot is

$$\left(\frac{(\mathbf{x}^*)^T\mathbf{b} - \mathbb{E}[y^*]}{\sqrt{(\mathbf{x}^*)^T(X^T X)^{-1}\mathbf{x}^*\sigma^2}}\right) \Big/ \left(\sqrt{\frac{SS_{Res}/\sigma^2}{n-p}}\right) = \frac{(\mathbf{x}^*)^T\mathbf{b} - \mathbb{E}[y^*]}{s\sqrt{(\mathbf{x}^*)^T(X^T X)^{-1}\mathbf{x}^*}} \sim t_{n-p}.$$

Given a set of inputs, a 95% CI for the response gives an interval that contains the "expected" response 95% of the time. In contrast, given a set of inputs, a 95% PI

produces an interval in which we are 95% sure that any given response with those inputs lies in.

Suppose we have inputs $\mathbf{x}^*$ just like before, the response should be

$$y^* = (\mathbf{x}^*)^T\boldsymbol{\beta} + \epsilon^* \quad \Rightarrow \quad y^* - (\mathbf{x}^*)^T\mathbf{b} = (\mathbf{x}^*)^T\boldsymbol{\beta} + \epsilon^* - (\mathbf{x}^*)^T\mathbf{b},$$

where ϵ^* is an error associated with the future observation y^*, and $\mathbf{b}$ depends only on the current observations $\mathbf{y}$.

This gives

$$\mathbb{V}[(\mathbf{x}^*)^T\mathbf{b} - y^*] = \mathbb{V}[(\mathbf{x}^*)^T\mathbf{b}] + \mathbb{V}[\epsilon^*] = (1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*)\sigma^2,$$

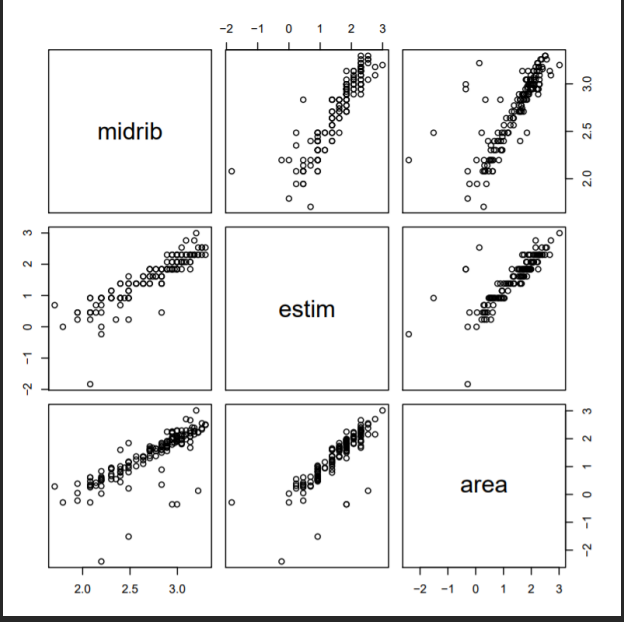
and since the estimator is unbiased, the expectation is $\mathbf{0}$.

Following exactly the previous arguments, we derive that

$$\frac{(\mathbf{x}^*)^T\mathbf{b} - y^*}{s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*}} \sim t_{n-p}. \qquad \text{\# wider than CI's}$$

R example: clover leaves.

```
> logclover <- log(read.csv("clover.csv"))    # can take log
> pairs(logclover)
```




```
> library(Matrix)
> y <- logclover$area
> X <- cbind(1, logclover$midrib, logclover$estim)
> dim(X)    # n p
[1] 145 3
> rankMatrix(X)[1]    # full rank
[1] 3
```

```
> model <- lm(area ~ midrib + estim, data = logclover)
> summary(model)
..
Residuals:
     Min       1Q   Median       3Q      Max
-2.31730 -0.07022 0.08005 0.18787 1.14160

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1741     0.4604   -2.55   0.0118 *
midrib        0.5240     0.2248    2.33   0.0212 *
estim         0.7338     0.1157    6.34 2.87e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4659 on 142 degrees of freedom   # s
Multiple R-squared: 0.7078,    Adjusted R-squared: 0.7036
F-statistic: 172 on 2 and 142 DF,  p-value: < 2.2e-16   # beta = (*,0,..,0)
```

```
> model$coefficients
(Intercept)    midrib     estim
 -1.1741275 0.5239692 0.7337812
> str(model$residuals)
Named num [1:145] 0.0575 -0.0524 -0.4043 -0.1323 0.3702 ...
> rstandard(model)[1]   # z_1
..
> str(model$fitted.values)
Named num [1:145] 0.2277 -0.2353 0.1811 0.1811 0.0151 ...
> model$rank
[1] 3
> model$df.residual
[1] 142
> deviance(model)/model$df.residual   # s^2 = SSRes/(n-p)
[1] 0.2170816
> influence(model)$hat[1]   # H_11
..
> cooks.distance(model)[1]   # D_1
..
```

```
> C <- solve(t(X)%*%X)
> confint.b0 <- b[1] + c(-1,1)*qt(0.975,df=n-p)*sqrt(s2*C[1,1])
> confint(model2, level = 0.95)   # CI for parameters
                2.5 %     97.5 %
(Intercept) -1.7871886 -0.9757665
..
```

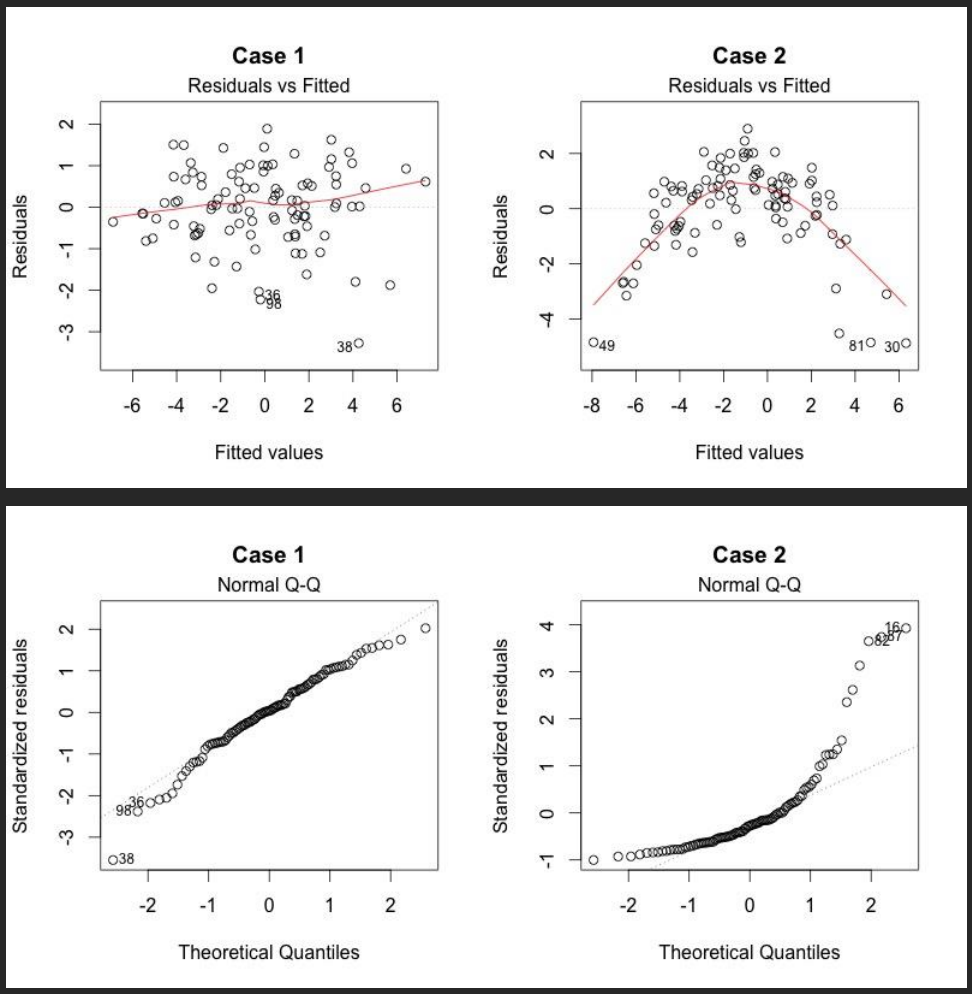
```
> (n-p)*s2/qchisq(c(0.975,0.025), n-p)   # CI for error variance
[1] 0.02774825 0.04471258
```

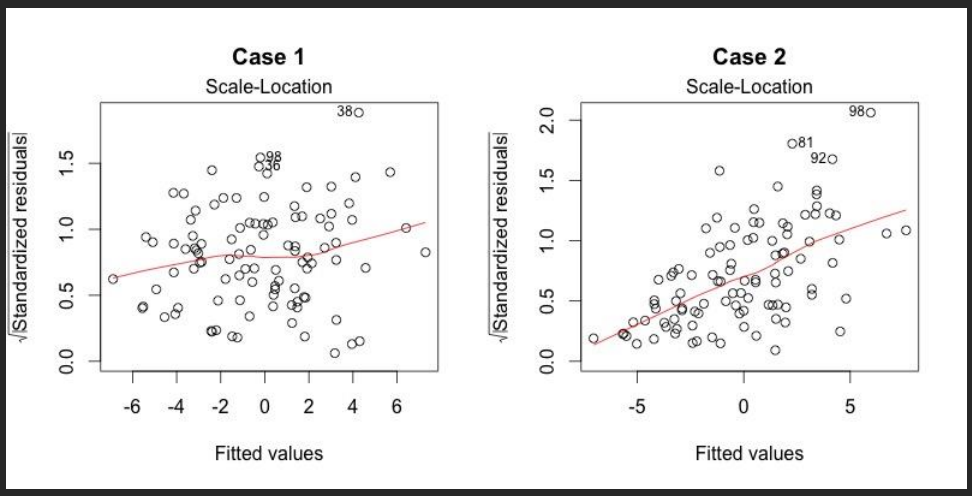
```
> newclover <- data.frame(midrib = log(10), estim = log(10))
> predict(model2, newclover, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 1.709429 1.538316 1.880541
> predict(model2, newclover, interval = "prediction", level = 0.95)
       fit      lwr     upr
1 1.709429 1.303147 2.11571
```

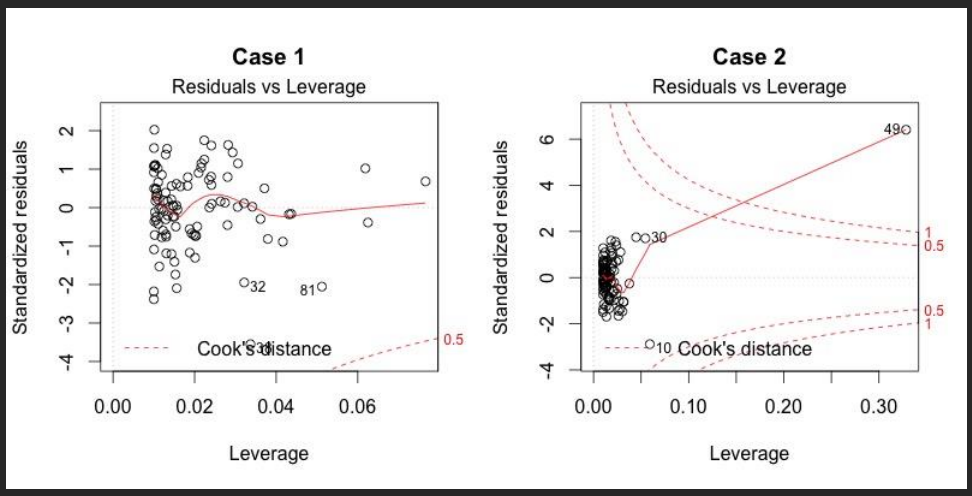
R produces many useful diagnostic plots for checking the fit of the model and deviations from assumptions, then we can remove any offending points.

```
> plot(model, which = c(1:3,5))
> goodlogclover <- logclover[-c(..),]
```

1. Residuals vs. fitted values:
   - non-zero residual sample mean ⇒ bias, or non-zero error means;
   - line not horizontal ⇒ non-linearity.
2. Normal QQ plot of the standardised residuals:
   - outliers to be removed;
   - line poor fit ⇒ non-normal errors.
3. Square roots of absolute values of standardised residuals against fitted values:
   - a pattern of spread around a (non-?)horizontal line ⇒ unequal variances (heteroskedasticity).
4. Leverage vs. standardised residuals: # which = 5
   - points with high leverage ⇒ potentially dangerous;
   - points with high Cook's distance ⇒ model poor fit.







Finding CI's for each parameter β_i individually does NOT guarantee that all of them will be satisfied at once, since they can be dependent. We need to find a joint confidence region for a number of parameters, i.e., $\boldsymbol{\beta}$, at the same time.

The least squares estimator gives $\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y} \sim \text{MVN}(\boldsymbol{\beta}, (X^TX)^{-1}\sigma^2)$, and so $(\mathbf{b}-\boldsymbol{\beta})^TX^TX(\mathbf{b}-\boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$. We also know that $SS_{Res}/\sigma^2 \sim \chi^2_{n-p}$. Since $\mathbf{b}$ and s^2 are independent, the two χ^2 variables are independent, which means that

$$\left(\frac{(\mathbf{b}-\boldsymbol{\beta})^TX^TX(\mathbf{b}-\boldsymbol{\beta})}{p\sigma^2}\right) \Big/ \left(\frac{SS_{Res}}{(n-p)\sigma^2}\right) = \frac{(\mathbf{b}-\boldsymbol{\beta})^TX^TX(\mathbf{b}-\boldsymbol{\beta})}{ps^2} \sim F_{p,n-p}.$$

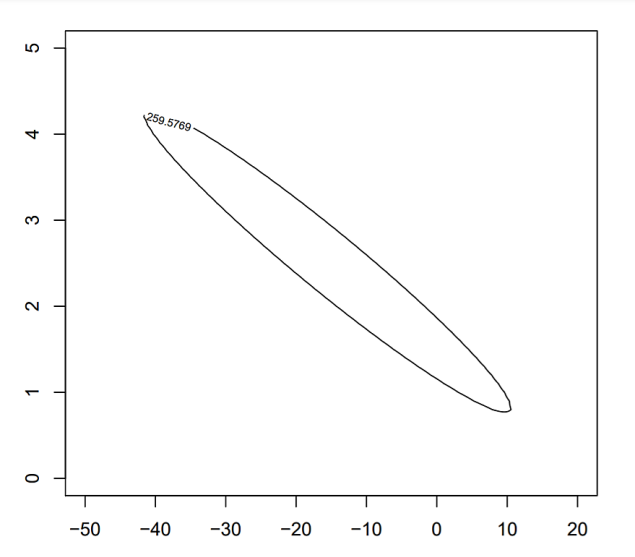
Because this statistic is based on $\mathbf{b}-\boldsymbol{\beta}$, which we want to be small, we use the right-hand tail of the F distribution to create a confidence region:

$$(\mathbf{b}-\boldsymbol{\beta})^TX^TX(\mathbf{b}-\boldsymbol{\beta}) \le ps^2f_{1-\alpha}.$$

This region has the form of an ellipse or ellipsoid.

```
> ..
> b1 <- seq(-50, 20, .2)
> b2 <- seq(0, 5, .1)
```

```
> f <- function(beta1, beta2) {
+     f.out <- rep(0, length(beta1))
+     for (i in 1:length(beta1)) {
+         beta <- as.vector(c(beta1[i],beta2[i]))
+         f.out[i] <- t(b - beta) %*% t(X) %*% X %*% (b - beta)
+     }
+     return(f.out)
+ }
> z <- outer(b1, b2, f)
> contour(b1, b2, z, levels = p*s2*qf(0.95,p,n-p))
```



So far, three assumptions have been made about our linear model:

1. the normality of errors.
2. $\mathbf{0}$ error mean; otherwise, we should find another model.
3. $\sigma^2 I$ error variance; otherwise, …

Suppose that $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, V)$. The MLE now maximises:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{k/2}|V|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}-X\boldsymbol{\beta})^T V^{-1}(\mathbf{y}-X\boldsymbol{\beta})},$$

that is, minimises $(\mathbf{y} - X\boldsymbol{\beta})^T V^{-1}(\mathbf{y} - X\boldsymbol{\beta})$. This gives the generalised least square estimators

$$\mathbf{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} \mathbf{y}.$$

This estimator is unbiased, and $\mathbb{V}[\mathbf{b}] = (X^T V^{-1} X)^{-1}$.

Moreover, it can be shown that the Gauss-Markov Theorem still holds, i.e., the generalised least squares estimator is BLUE.

In the situation that errors are uncorrelated but do not have a common variance:

$$\mathbb{V}[\boldsymbol{\epsilon}] = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2).$$

To estimate the parameter with MLE, we minimise

$$(\mathbf{y} - X\boldsymbol{\beta})^T V^{-1}(\mathbf{y} - X\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\epsilon_i}{\sigma_i}\right)^2.$$

That is, we weight each residual by the inverse of the corresponding standard deviation. So a point with high variance influences **b** less than a point with low variance.

Non-linearity. Signs which may indicate that a transformation is required:

1. All the values are positive.
2. The distribution of the data is skewed.
3. There is an obvious non-linear relationship with another variable.
4. The variances show a relationship with one of the variables.

Logarithmic transformations are very common because they convert multiplicative effects into additive ones. Some useful transformations are:

1. $x,\ \ln y$	exponential
2. $\ln x,\ \ln y$	power law
3. $\sqrt{y}$	areas, or occurrences inside areas
4. $\sqrt[3]{y}$	volumes
5. $1/y$	rates
6. $\ln y/(1-y)$	proportions

Do NOT forget to transform the model back later.

From ANOVA, the Residual (Error) SS is

$$SS_{Res} = (\mathbf{y} - H\mathbf{y})^T(\mathbf{y} - H\mathbf{y}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T H\mathbf{y} + \mathbf{y}^T H^2\mathbf{y} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T H\mathbf{y}$$
$$= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\hat{\mathbf{y}} = \mathbf{y}^T\mathbf{y} - \hat{\mathbf{y}}^T\hat{\mathbf{y}} = SS_{Total} - SS_{Reg}\ ,$$

where SS_{Total} is the Total SS, and SS_{Reg} is the Regression SS.

In the full rank general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we have (proved earlier) $SS_{Res}/\sigma^2 \sim \chi^2_{n-p}$, and SS_{Reg}/σ^2 has a noncentral χ^2 distribution with p d.f. and noncentrality parameter $\lambda = \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}/2\sigma^2$.

It can be proved that SS_{Res} and SS_{Reg} are independent, knowing that both of them are quadratic forms in $\mathbf{y}$. Alternatively, we write $SS_{Res} = (n-p)s^2$ and $SS_{Reg} = \mathbf{b}^T X^T X\mathbf{b}$, observe that $\mathbf{b}$ and s^2 are independent.

The general linear hypothesis is

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\delta}^* \quad \text{vs.} \quad H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\delta}^*,$$

where C is an $r \times p$ matrix of rank $r \leq p$ and $\boldsymbol{\delta}^*$ is an $r \times 1$ vector of constants.

This hypothesis makes it possible to test for many things, including relationships among the parameters, or testing the individual parameters against a constant.

Now we are going to develop a test statistic. Since $\mathbf{b} \sim \text{MVN}(\boldsymbol{\beta}, (X^T X)^{-1}\sigma^2)$, we have $C\mathbf{b} - \boldsymbol{\delta}^* \sim \text{MVN}(C\boldsymbol{\beta} - \boldsymbol{\delta}^*, C(X^T X)^{-1}C^T\sigma^2)$. Then the quadratic form

$$(C\mathbf{b} - \boldsymbol{\delta}^*)^T(C(X^T X)^{-1}C^T)^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*)/\sigma^2 \qquad \#\ C(X^T X)^{-1}C^T \text{ nonsingular?}$$

has a noncentral χ^2 distribution with r d.f. and noncentrality parameter

$$\lambda = (C\boldsymbol{\beta} - \boldsymbol{\delta}^*)^T(C(X^T X)^{-1}C^T)^{-1}(C\boldsymbol{\beta} - \boldsymbol{\delta}^*)/2\sigma^2.$$

If the null hypothesis is true, then $C\boldsymbol{\beta} = \boldsymbol{\delta}^*$ and the quadratic form has an ordinary χ^2 distribution. Since the quadratic form depends only on $\mathbf{b}$, it is independent from s^2. Therefore, the statistic is

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T(C(X^T X)^{-1}C^T)^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*)/r}{SS_{Res}/(n-p)} \sim F_{r,n-p}.$$

To justify a one-tailed test, the expected value of the numerator can be calculated to be

$$\mathbb{E}[(C\mathbf{b} - \boldsymbol{\delta}^*)^T(C(X^T X)^{-1}C^T)^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*)/r] = tr(AV) + \boldsymbol{\mu}^T A\boldsymbol{\mu}$$
$$= \sigma^2 + \frac{1}{r}(C\boldsymbol{\beta} - \boldsymbol{\delta}^*)^T(C(X^T X)^{-1}C^T)^{-1}(C\boldsymbol{\beta} - \boldsymbol{\delta}^*).$$

where $C(X^T X)^{-1}C^T$ is positive definite. If the null hypothesis is true, then the expectation of numerator is σ^2, and the F statistic should be close to 1. Therefore, we reject H_0 when the statistic is large.

Recall the joint confidence region for the parameters of a full rank linear model?

```
## Test H0: beta0 = -1, beta1 = beta2, on prior clover example
> C <- matrix(c(1,0,0,1,0,-1), 2, 3)
> dst <- c(-1,0)   # delta star

> library(car)
> linearHypothesis(model, C, dst)
..
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    # (SoS/Df)/(RSS/Res.Df)
1    138 6.0792
2    136 4.7221  2    1.3571 19.543 3.464e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Particularly, if the null hypothesis $\boldsymbol{\beta} = \mathbf{0}$ is true, then $C = I$, $\boldsymbol{\delta}^* = \mathbf{0}$. Our test statistic is:

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T (C(X^T X)^{-1} C^T)^{-1} (C\mathbf{b} - \boldsymbol{\delta}^*)/r}{SS_{Res}/(n-p)} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p)} = \frac{MS_{Reg}}{MS_{Res}} \sim F_{p,n-p}.$$

If a particular β_i is zero, then it is best to remove it from the model. Otherwise, it will only fit noise, and reduce the ability of the model to predict. Thus, we need to find a way of testing whether parts of the parameter vector are $\mathbf{0}$ or not.

Split the parameter vector

$$\boldsymbol{\beta} = [\, \beta_0 \;\; \cdots \;\; \beta_{r-1} \mid \beta_r \;\; \cdots \;\; \beta_k \,]^T = [\, \boldsymbol{\gamma}_1^T \mid \boldsymbol{\gamma}_2^T \,]^T,$$

and test the hypotheses

$$H_0 : \boldsymbol{\gamma}_1 = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\gamma}_1 \neq \mathbf{0}.$$

By relabelling the indies, we can test the relevance of any subset of the parameters.

An important thing to note is that we are testing $\boldsymbol{\gamma}_1 = \mathbf{0}$ in the presence of the other parameters, not by itself alone. In other words, we are comparing two models: the full model in H_1, against the reduced model in H_0:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{vs.} \quad \mathbf{y} = X_2\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}_2.$$

where X_2 contains the last $p - r$ columns of $X = [\, X_1 \mid X_2 \,]$.

We can do this in the framework of the general linear hypothesis. Let $C = [\, I_r \mid 0 \,]$ and $\boldsymbol{\delta}^* = \mathbf{0}$. Then $C\boldsymbol{\beta} = \boldsymbol{\delta}^*$ iff $\boldsymbol{\gamma}_1 = \mathbf{0}$. Define the regression SS for $\boldsymbol{\gamma}_1$ in the presence of $\boldsymbol{\gamma}_2$ as

$$\begin{aligned} R(\boldsymbol{\gamma}_1 | \boldsymbol{\gamma}_2) &= (C\mathbf{b} - \boldsymbol{\delta}^*)^T (C(X^T X)^{-1} C^T)^{-1} (C\mathbf{b} - \boldsymbol{\delta}^*) = \widehat{\boldsymbol{\gamma}_1}^T A_{11}^{-1} \widehat{\boldsymbol{\gamma}_1} \\ &= SS_{Reg}(\boldsymbol{\beta}) - SS_{Reg}(\boldsymbol{\gamma}_2), \end{aligned}$$

where $\widehat{\boldsymbol{\gamma}_1}$ is the least squares estimator, A_{11} the $r \times r$ upper-left submatrix of $(X^T X)^{-1}$.

Therefore, we should reject the null when

$$\frac{R(\boldsymbol{\gamma}_1 | \boldsymbol{\gamma}_2)/r}{SS_{Res}/(n-p)} \sim F_{r,n-p} > c.$$

Test $H_0 : \beta_1 = \cdots = \beta_k = 0$ versus $H_1 = \overline{H_0}$. If the null is true, we have

$$SS_{Reg}(\boldsymbol{\gamma_2}) = \mathbf{y}^T\mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{y} = \left(\sum_{i=1}^{n} y_i\right)^2 /n \sim \chi_1^2 ,$$

$$R(\boldsymbol{\gamma_1}|\boldsymbol{\gamma_2}) = SS_{Reg}(\boldsymbol{\beta}) - SS_{Reg}(\boldsymbol{\gamma_2}) \sim \chi_{p-1}^2 ,$$

$$SS_{cor} = SS_{Res}(\boldsymbol{\gamma_2}) + R(\boldsymbol{\gamma_1}|\boldsymbol{\gamma_2}) = SS_{Total} - SS_{Reg}(\boldsymbol{\gamma_2}) = \mathbf{y}^T\mathbf{y} - \left(\sum_{i=1}^{n} y_i\right)^2 /n \sim \chi_{n-1}^2 .$$

where SS_{cor} refers to the corrected SS.

```
> str(X)
 num [1:11, 1:4] 1 1 1 1 1 1 1 1 1 1 ...
> y
[1] 22.6 15.0 78.1 29.0 80.5 24.5 20.5 147.6 4.2 48.2 20.5

> model <- lm(y ~ X[,-1])
> null1 <- lm(y ~ 0)          # beta = (0,0,..,0)
> null2 <- lm(y ~ 1)          # beta = (*,0,..,0), corrected

> anova(null2, model)
Analysis of Variance Table
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     10 17866.5
2      7   688.6  3     17178 58.205 2.578e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Suppose that we have a number of explanatory variables in a model, but it's not obvious if all of them are relevant. We could fit a model using all of them, but this runs the risk of overfitting: using irrelevant variables to explain noise by coincidence. Ideally, we prefer to select a parsimonious model, i.e., using a minimal number of explanatory variables, so if we were to test if any parameter is 0, in the presence of the other model parameters, we should always reject the null.

Conceivably, we could test all the possible parameters sets to find the largest $\boldsymbol{\gamma_1}$ such that the hypothesis $\boldsymbol{\gamma_1} = \mathbf{0}$ is not rejected. However, problems can happen when two variables are strongly correlated, and given one of them the other isn't needed, but you need to have at least one of them. For example, we might reject $\beta_1 = \beta_2 = 0$, but not reject $\beta_1 = 0$ given β_2 and β_3, and also not reject $\beta_2 = 0$ given β_1 and β_3.

Now we consider partial tests, i.e., p tests of the form $H_0 : \beta_i = 0$, given all the other parameters are in the model. The problem of partial tests is that they are not independent, i.e., "acceptance" or rejection of H_0 just means that the parameter is useful or not in the full model, instead of in the best model.

To avoid this, we can consider a nested sequence of models, supported by the theorem:

$$\frac{\mathbf{y}^T\mathbf{y}}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} + \frac{R(\beta_0)}{\sigma^2} + \frac{R(\beta_1|\beta_0)}{\sigma^2} + \frac{R(\beta_2|\beta_0,\beta_1)}{\sigma^2} + \cdots + \frac{R(\beta_k|\beta_0,\ldots,\beta_{k-1})}{\sigma^2}$$

where the quadratic forms on the right are all independent with noncentral χ^2 distributions. SS_{Res} has $n-p$ d.f. and the rest have 1 d.f. each. Then

1. Forward testing: Start with a simple model and sequentially add parameters until adding parameters does not significantly improve the fit.
   (i) Start with an empty model.
   (ii) Calculate the F-values for the tests $H_0 : \beta_i = 0$, for all parameters not in the model, in the presence of parameters already in the model.
   (iii) If none of the tests are significant (no null hypothesis rejected), then stop.
   (iv) Otherwise add the most significant parameter (i.e., parameter with the largest F-value).
   (v) Return to step (ii).

```
> basemodel <- lm(y ~ 1, data = heat)
> add1(basemodel, scope= ~. + x1 + x2 + x3 + x4, test = "F")
..
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>              2715.76 71.444
x1      1   1450.08 1265.69 63.519 12.6025 0.0045520 **
x2      1   1809.43  906.34 59.178 21.9606 0.0006648 ***
x3      1    776.36 1939.40 69.067  4.4034 0.0597623 .
x4      1   1831.90  883.87 58.852 22.7985 0.0005762 ***
..
> model2 <- lm(y ~ x4, data = heat)
> add1(model2, scope= ~. + x1 + x2 + x3, test = "F")
..
       Df Sum of Sq    RSS    AIC  F value    Pr(>F)
<none>              883.87 58.852
x1      1    809.10  74.76 28.742 108.2239 1.105e-06 ***
x2      1     14.99 868.88 60.629   0.1725    0.6867
x3      1    708.13 175.74 39.853  40.2946 8.375e-05 ***
..
> model3 <- lm(y ~ x1 + x4, data = heat)
> add1(model3, scope= ~. + x2 + x3, test = "F")
..
       Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>              74.762 28.742
x2      1    26.789 47.973 24.974  5.0259 0.05169 .
x3      1    23.926 50.836 25.728  4.2358 0.06969 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Backward elimination: Start with a full model and sequentially remove parameters until removing parameters significantly worsens the fit.
   (i) Start with the full model.

(ii) Calculate the F-values for the tests $H_0 : \beta_i = 0$, for all parameters in the model, in the presence of the other parameters in the model.

(iii) If all of the tests are significant, then stop.

(iv) Otherwise remove the least significant parameter.

(v) Return to step (ii).

```
> fullmodel <- lm(y ~., data = heat)
> drop1(fullmodel, scope= ~., test = "F")
..
       Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>              47.864 26.944
x1      1   25.9509 73.815 30.576  4.3375 0.07082 .
x2      1    2.9725 50.836 25.728  0.4968 0.50090
x3      1    0.1091 47.973 24.974  0.0182 0.89592
x4      1    0.2470 48.111 25.011  0.0413 0.84407
..
> model2 <- lm(y ~ x1 + x2 + x4, data = heat)
> drop1(model2, scope= ~., test = "F")
..
       Df Sum of Sq    RSS    AIC  F value    Pr(>F)
<none>               47.97 24.974
x1      1    820.91 868.88 60.629 154.0076 5.781e-07 ***
x2      1     26.79  74.76 28.742   5.0259   0.05169 .
x4      1      9.93  57.90 25.420   1.8633   0.20540
..
> model3 <- lm(y ~ x1 + x2, data = heat)
> drop1(model3, scope= ~., test = "F")
..
       Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>                57.90 25.420
x1      1    848.43  906.34 59.178  146.52 2.692e-07 ***
x2      1   1207.78 1265.69 63.519  208.58 5.029e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F tests require the variable set of one model to be fully contained in the variable set of the other model, and an arbitrary choice of a significance level is needed. Goodness-of-fit measures tell how good a model is, independently of other models:

1. SS_{Res}, but always decreases as a new variable added, suffer from overfitting.
2. s^2, discharges overfitting sightly since $n - p$ also decreases.
3. $R^2 = 1 - SS_{Res}/(SS_{Total} - (\sum_{i=1}^{n} y_i)^2/n) \in [0, 1]$, the proportion of corrected total sums of squares explained by the model; suffer from overfitting.
4. adj $R^2 = 1 - \big((n-1)/(n-1-k)\big)(1 - R^2)$ performs better than s^2, assuming that β_0 is always in the model.

5. Akaike's information criterion (AIC), likelihood is the maximised likelihood; a smaller value of AIC indicates a better model:

$$AIC = -2\ln(\text{likelihood}) + 2p = n\ln(2\pi) + n\ln(\sigma^2) + \frac{(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})}{\sigma^2} + 2p$$
$$\approx n\ln\left(\frac{SS_{Res}}{n}\right) + 2p + n\ln(2\pi) + \frac{SS_{Res}}{\left(\frac{SS_{Res}}{n}\right)}$$
$$= n\ln\left(\frac{SS_{Res}}{n}\right) + 2p + \text{const.}$$

6. Bayesian information criterion (BIC) penalises extra parameters more harshly than AIC, so promotes a model with fewer variables; smaller = better:

$$BIC = -2\ln(\text{likelihood}) + p\ln n = n\ln\left(\frac{SS_{Res}}{n}\right) + p\ln n + \text{const.}$$

7. Mallow's C_p statistic compares the residual SS of an intermediate model against the residual SS for a full model; smaller = better:

$$C_p = SS_{Res}(\text{model})/s^2(\text{full model}) + 2p - n.$$

Note that any goodness-of-fit statistic should only to be used to compare various models for **the same data**. There is no absolute measure of how good a model is, for any of them.

3. Stepwise selection: Using a goodness-of-fit measure, e.g. AIC, an already added variable can be removed, and vice versa; however, the final model depends on the starting model, so it does not necessarily find a global optimum.
   (i) Start with any model.
   (ii) Compute the AICs of all models which either have one extra variable or one less variable than the current model.
   (iii) If the AICs of all such models are more than the current, stop.
   (iv) Otherwise change to the model with the lowest AIC.
   (v) Return to step (ii).

```
> model2 <- step(basemodel, scope= ~. + x1 + x2 + x3 + x4, steps = 1)
..
       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444
..
> model3 <- step(model2, scope= ~. + x1 + x2 + x3, steps = 1)
..
       Df Sum of Sq     RSS    AIC
+ x1    1    809.10   74.76 28.742
+ x3    1    708.13  175.74 39.853
```

```
<none>               883.87 58.852
+ x2    1     14.99  868.88 60.629
- x4    1   1831.90 2715.76 71.444
..
> model4 <- step(model3, scope= ~. + x2 + x3, steps = 1)
..
       Df Sum of Sq     RSS    AIC
+ x2    1     26.79   47.97 24.974
+ x3    1     23.93   50.84 25.728
<none>                74.76 28.742
- x1    1    809.10  883.87 58.852
- x4    1   1190.92 1265.69 63.519
..
> step(model4, scope= ~. + x3)
..
       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- x4    1      9.93  57.90 25.420
+ x3    1      0.11  47.86 26.944
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629

Call:
lm(formula = y ~ x4 + x1 + x2, data = heat)

Coefficients:
(Intercept)      x4     x1     x2
    71.6483 -0.2365 1.4519 0.4161
```

We can also use a t test for a partial test of one parameter. Let us compare this with our partial F test. The statistics we use are:

$$\frac{b_i}{s\sqrt{c_{ii}}} \sim t_{n-p} \quad \text{and} \quad \frac{R(\beta_i|\beta_0,\dots,\beta_{i-1},\beta_{i+1},\dots,\beta_k)}{SS_{Res}/(n-p)} \sim F_{1,n-p}$$

where c_{ii} is the (i,i)'th entry of $(X^TX)^{-1}$.

We saw previously that the numerator is

$$R(\beta_i|\beta_0,\dots,\beta_{i-1},\beta_{i+1},\dots,\beta_k) = \widehat{\boldsymbol{\gamma_1}}^T A_{11}^{-1} \widehat{\boldsymbol{\gamma_1}}$$

where $\widehat{\boldsymbol{\gamma_1}} = b_i$, and A_{ii} is the top-left submatrix of $(X^TX)^{-1}$ after the columns have been re-arranged so that the i'th column comes first, i.e., c_{ii}, and so

$$\frac{R(\beta_i|\beta_0,\dots,\beta_{i-1},\beta_{i+1},\dots,\beta_k)}{SS_{Res}/(n-p)} = \frac{b_i(c_{ii})^{-1}b_i}{s^2} = \frac{b_i^2}{s^2 c_{ii}}.$$

This is exactly the square of the t statistic.

Furthermore, $T_{n-p}^2 \sim \left(Z/\sqrt{X_{n-p}/(n-p)}\right)^2 \sim (X_1/1)/\left(X_{n-p}/(n-p)\right) \sim F_{1,n-p}$.

The Ridge regression minimise the residual SS, but include a term which penalises the size of the parameters. We choose **b** to minimise $\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{k} b_j^2$. The penalised least squares estimators can be calculated to be $\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$.

The (positive) λ term controls the amount of "shrinkage" of the parameters. This needs to be tuned by minimising a modified version of AIC:

$$AIC = -n \ln\left(\frac{SS_{Res}}{n}\right) + 2df,$$

where $df = tr(H) = tr(X(X^T X + \lambda I)^{-1} X^T)$ is the effective d.f..

Another approach is the LASSO, which minimises $\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{k} |b_j|$.

The ridge regression will never shrink parameters to 0. The LASSO actually shrinks small parameters to 0.

A common method to choose shrinking parameter λ is cross-validation, which estimates the predictive power of the model by removing parts of the dataset and using them as test sets.

For this section and the next, the design matrix X can be less-than-full rank.

In the one-way classification model, samples come from k distinct sub-populations with different characteristics, in which the i'th population has n_i samples.

Let y_{ij} be the j'th sample taken from the i'th population. Then the model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{21} \\ y_{22} \\ \vdots \\ y_{k,n_k} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{k,n_k} \end{bmatrix},$$

for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, n_i$.

In this model, $r(X) = k$, i.e., less-than-full rank. We can re-parametrise the model as $y_{ij} = \mu_i + \epsilon_{ij}$, and easily convert to a full rank model.

However, this is NOT always possible. Consider the two-way classification model without interaction (also known as the additive model), having a levels of factor 1, b levels of factor 2, and n_{ij} observations for each combination (i, j):

$$y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk}, \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}.$$

$$X = \left[\begin{array}{c|cccc|cccc} 1 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\ \hline 1 & 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 1 \\ \hline \vdots & & & & & & & & \\ \hline 1 & 0 & 0 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \end{array}\right], \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_a \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_b \end{bmatrix}.$$

Let A be a $n \times p$ matrix. The $p \times n$ matrix A^c is called a (nonunique) conditional inverse for A iff $AA^cA = A$. If A is nonsingular, then $A^{-1} = A^c$.

Conditional inverses can be constructed as follows:

1. Find a submatrix M of A which is nonsingular and of dimension $r(A) \times r(A)$.
2. Replace M in A with $(M^{-1})^T$ and other entries with zeros.
3. Transpose the resulting matrix.

Let A be a $n \times p$ matrix of rank r, where $n \geq p \geq r$. Then

1. $r(A) = r(AA^c) = r(A^cA) = r(A^c)$, since $r(A) \geq r(AA^c) \geq r(AA^cA) = r(A)$.
2. $(A^c)^T = (A^T)^c$.
3. A^cA and AA^c are idempotent.
4. $A = A(A^TA)^cA^TA$ and $A^T = A^TA(A^TA)^cA^T$, since $A^TA = A^TA(A^TA)^cA^TA$.
5. $A(A^TA)^cA^T$ is unique, symmetric and idempotent. An expression involving a conditional inverse is unique if it is invariant to the choice of conditional inverse.
6. $r(A(A^TA)^cA^T) = r$, since $r(A(A^TA)^cA^T) \geq r(A(A^TA)^cA^TA) = r(A) \geq r(A(A^TA)^cA^T)$.

```
> library(MASS)
> A <- matrix(c(2,-6,3,1,6,4,3,0,7), 3, 3)   # det = 0

# Moore-Penrose Pseudoinverse, is unique
> (Ac1 <- ginv(A))
            [,1]         [,2]       [,3]
[1,] 0.025713835 -0.084240416 0.03659883
[2,] 0.009149708  0.080454330 0.04369774
[3,] 0.034863543 -0.003786086 0.08029658

# manually
> Ac2 <- matrix(0,3,3)
> Ac2[1:2,1:2] <- t(solve(A[1:2,1:2]))
> Ac2 <- t(Ac2)
> Ac2
          [,1]        [,2] [,3]
[1,] 0.3333333 -0.05555556    0
[2,] 0.3333333  0.11111111    0
[3,] 0.0000000  0.00000000    0
```

The system $A\mathbf{x} = \mathbf{g}$ is consistent (i.e., has at least one solution) iff the rank of $[\, A \mid \mathbf{g} \,]$ is equal to the rank of A.

($\Leftarrow$) $r([\, A \mid \mathbf{g} \,]) = r(A)$, $\mathbf{g}$ must be a linear combination of the columns of A. Therefore, there exists constants $x_1, x_2, \ldots, x_p$ such that $x_1\mathbf{a_1} + x_2\mathbf{a_2} + \cdots + x_p\mathbf{a_p} = \mathbf{g}$, where $\mathbf{a_i}$ is the i'th column of A.

In the general linear model $y = X\beta + \epsilon$, the normal equations $X^TX\mathbf{b} = X^T\mathbf{y}$ is consistent.

$r([\, X^TX \mid X^T\mathbf{y} \,]) = r(X^T[\, X \mid \mathbf{y} \,]) \leq r(X^T) = r(X^TX) \leq r([\, X^TX \mid X^T\mathbf{y} \,])$.

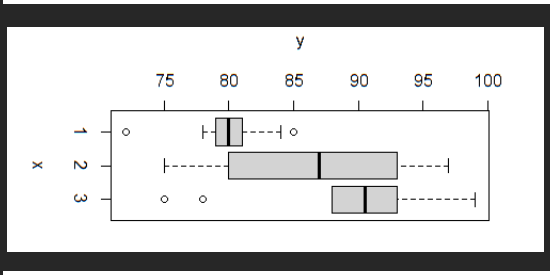
Let $A\mathbf{x} = \mathbf{g}$ be a consistent system. Then $A^c\mathbf{g}$ is a solution to the system, where A^c is any conditional inverse for A.

Since $A\mathbf{x}^* = \mathbf{g}$ for some $\mathbf{x}^*$, $A\mathbf{x}^* = AA^cA\mathbf{x}^* = AA^c\mathbf{g}$.

Furthermore, any $\mathbf{x} = A^c\mathbf{g} + (I - A^cA)\mathbf{z}$ solves the system, where $\mathbf{z}$ is an arbitrary $p \times 1$ vector. Particularly if $\mathbf{x}' = \text{ginv}(A)\,\mathbf{g}$, then $\mathbf{x} = \mathbf{x} + (I - A^cA)\mathbf{x}' \;\; \forall\, \mathbf{x}$.

Therefore, $\mathbf{b} = (X^T X)^c X^T \mathbf{y}$ is a (nonunique) solution to the normal equations. But note that $H = X(X^T X)^c X^T$ is unique, symmetric and idempotent.

```
> maths <- read.csv("maths.csv")
> maths$class.f <- factor(maths$class.f)
> str(maths)
'data.frame':   30 obs. of  4 variables:
 $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ maths.y: int  81 84 81 79 78 79 81 85 72 79 ...
 $ iq     : int  99 103 108 109 96 104 96 105 94 91 ...
 $ class.f: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
> plot(maths$class.f, maths$maths.y)
```




```
> y <- maths$maths.y
> n <- dim(maths)[1]
> k <- length(levels(maths$class.f))
> X <- matrix(c(rep(1,n),rep(0,n*k)), n, k+1)
> X[maths$class.f==1, 2] <- 1
> X[maths$class.f==2, 3] <- 1
> X[maths$class.f==3, 4] <- 1
```

```
## Re-parameterisation
> Xre <- X[,-1]
> (b1 <- solve(t(Xre) %*% Xre, t(Xre) %*% y))
     [,1]
[1,] 79.9
[2,] 86.5
[3,] 89.4
> modelre <- lm(y ~ 0 + X[,2] + X[,3] + X[,4])

## By conditional inverse
> library(MASS)
> (b2 <- ginv(t(X) %*% X) %*% t(X) %*% y)
      [,1]
[1,] 63.95
[2,] 15.95
[3,] 22.55
[4,] 25.45
```

In a general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, a quantity $\mathbf{t}^T\boldsymbol{\beta}$ is said to be estimable if there exists a vector $\mathbf{c}$ such that $\mathbb{E}[\mathbf{c}^T\mathbf{y}] = \mathbf{t}^T\boldsymbol{\beta}$, i.e., there is a linear unbiased estimator for it.

(Extended.) … iff the linear system $X^T X\mathbf{z} = \mathbf{t}$ is consistent.

($\Leftarrow$) Let $\mathbf{z}^*$ be a solution to $X^T X\mathbf{z} = \mathbf{t}$ and put $\mathbf{c} = X\mathbf{z}^*$. Then $\mathbb{E}[\mathbf{c}^T\mathbf{y}] = \mathbb{E}[\mathbf{z}^{*T} X^T\mathbf{y}] = \mathbf{z}^{*T} X^T\mathbb{E}[\mathbf{y}] = \mathbf{z}^{*T} X^T X\boldsymbol{\beta} = \mathbf{t}^T\boldsymbol{\beta}$.

(Extended..) ... iff $\mathbf{t}^T(X^TX)^cX^TX = \mathbf{t}^T$, for **any and all** (?) conditional inverse of X^TX.

($\Leftarrow$) Transpose on both sides of $\mathbf{t}^T(X^TX)^cX^TX = \mathbf{t}^T$, then $X^TX(X^TX)^c\mathbf{t} = \mathbf{t}$. Observe that $(X^TX)^c\mathbf{t}$ is a solution to the linear system $X^XX\mathbf{z} = \mathbf{t}$.

($\Rightarrow$) Suppose that $\mathbf{t}^T\boldsymbol{\beta}$ is estimable, then there exists a solution to the system $X^TX\mathbf{z} = \mathbf{t}$. We know that a solution is $\mathbf{z} = (X^TX)^c\mathbf{t}$. Note that the conditional inverse is arbitrary.

Gauss-Markov Theorem. In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, suppose $\mathbf{t}^T\boldsymbol{\beta}$ is estimable. Then the BLUE for $\mathbf{t}^T\boldsymbol{\beta}$ is $\mathbf{z}^TX^T\mathbf{y}$, where $\mathbf{z}$ is a solution to the system $X^TX\mathbf{z} = \mathbf{t}$. Furthermore, this estimate is the same for any solution of the system, and can be written $\mathbf{t}^T\mathbf{b}$, where $\mathbf{b}$ is any solution to the normal equations.

(Extended.) ..., elements of $X\boldsymbol{\beta}$ are estimable. For example, in a one-way classification model with any number of levels, $\mu + \tau_i$ is always estimable.

(Extended..) Let $\mathbf{t}_1^T\boldsymbol{\beta}, \mathbf{t}_2^T\boldsymbol{\beta}, \mathbf{t}_3^T\boldsymbol{\beta}$ be estimable functions, and let $z = a_1\mathbf{t}_1^T\boldsymbol{\beta} + a_2\mathbf{t}_2^T\boldsymbol{\beta} + \cdots + a_k\mathbf{t}_k^T\boldsymbol{\beta}$. Then z is estimable, and the BLUE for z is $a_1\mathbf{t}_1^T\mathbf{b} + a_2\mathbf{t}_2^T\mathbf{b} + \cdots + a_k\mathbf{t}_k^T\mathbf{b}$.

In a one-way classification model, any treatment contrast (the same categorical variable) $a_1\tau_1 + a_2\tau_2 + \cdots + a_k\tau_k$ is estimable, as long as $\sum_{i=1}^k a_i = 0$. Then

$$z = \sum_{i=1}^k a_k\mu + a_1\tau_1 + a_2\tau_2 + \cdots + a_k\tau_k = a_1(\mu + \tau_1) + a_2(\mu + \tau_2) + \cdots + a_k(\mu + \tau_k)$$

is a linear combination of the estimable function $\mu + \tau_i$, and is therefore estimable.

For some $i \neq j$, $\tau_i - \tau_j = (\mu + \tau_i) - (\mu + \tau_j) = \bar{y}_i - \bar{y}_j$.

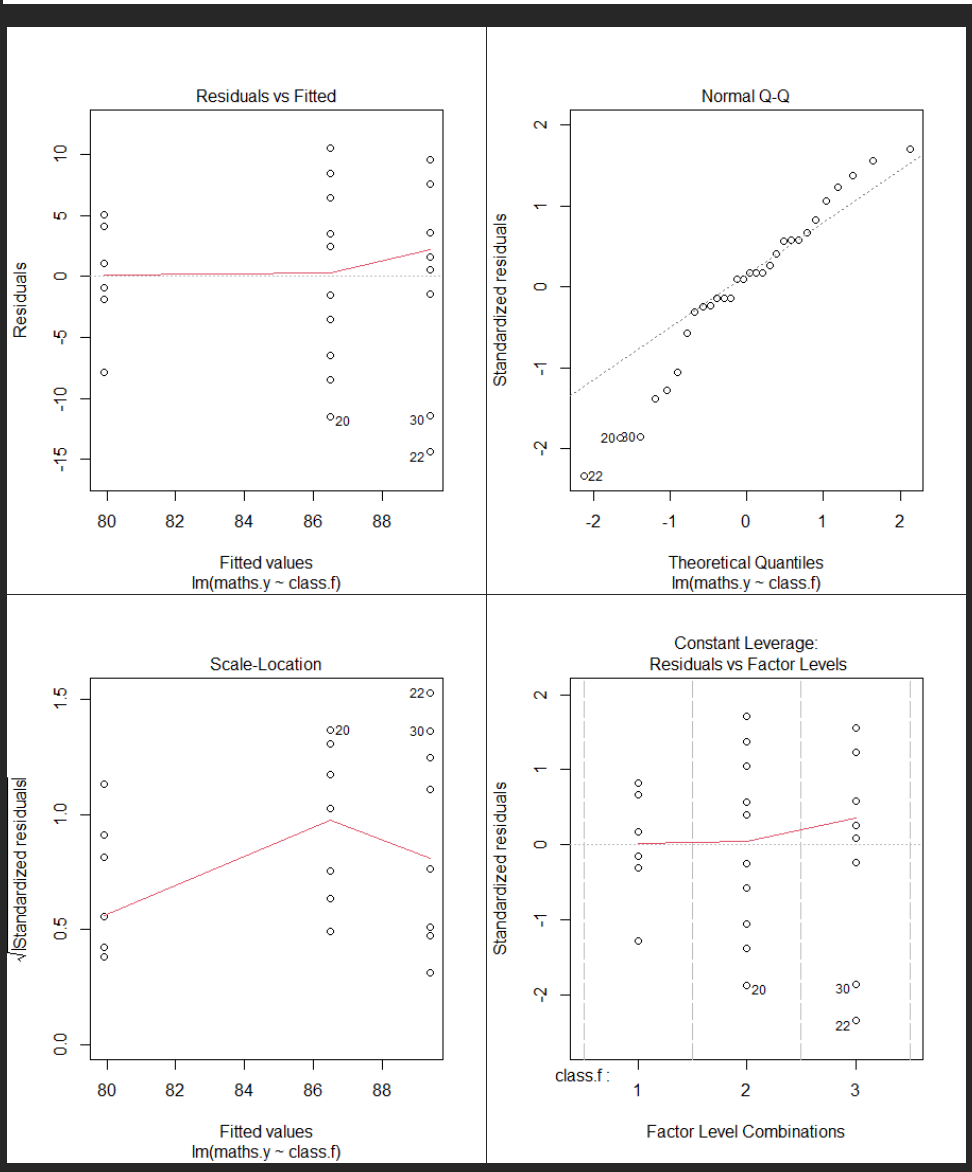
The two main contrast sets are contr.treatment and contr.sum. For the one-way classification model:

Label	contr.treatment		contr.sum	
Intercept	μ_1	$\mu + \tau_1$	$\bar{\mu}$	$\mu + k^{-1}\sum\tau_i$
Factor 1	\		$\mu_1 - \bar{\mu}$	$\tau_1 - k^{-1}\sum\tau_i$
Factor 2	$\mu_2 - \mu_1$	$\tau_2 - \tau_1$	$\mu_2 - \bar{\mu}$	$\tau_2 - k^{-1}\sum\tau_i$
⋮	⋮	⋮	⋮	⋮
Factor $k-1$	$\mu_{k-1} - \mu_1$	$\tau_{k-1} - \tau_1$	$\mu_{k-1} - \bar{\mu}$	$\tau_{k-1} - k^{-1}\sum\tau_i$
Factor k	$\mu_k - \mu_1$	$\tau_k - \tau_1$	\ $= -(\mu_1 - \bar{\mu}) - \cdots - (\mu_{k-1} - \bar{\mu})$	

```
## R re-parameterises the model using contrasts
> contrasts(maths$class.f) <- contr.treatment(k)   # R's default
> model1 <- lm(maths.y ~ class.f, data = maths)
> summary(model1)
..
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   79.900      2.053  38.922  < 2e-16 ***
class.f2       6.600      2.903   2.273  0.03117 *
class.f3       9.500      2.903   3.272  0.00292 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
> contrasts(maths$class.f) <- contr.sum(k)
> model2 <- lm(maths.y ~ class.f, data = maths)
> summary(model2)
..
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   79.900      2.053  38.922  < 2e-16 ***
class.f2       6.600      2.903   2.273  0.03117 *
class.f3       9.500      2.903   3.272  0.00292 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> plot(model1, which = c(1:3,5))
```



For the less-than-full rank model we can still define the RSS as $(\mathbf{y}-X\mathbf{b})^T(\mathbf{y}-X\mathbf{b})$, invariant to the choice of $\mathbf{b}$. Although $\mathbf{b}$ can vary, $X\mathbf{b}$ will not, because $X\boldsymbol{\beta}$ is estimable.

$\mathbb{E}[SS_{Res}] = r(I-H)\sigma^2 = (n-r)\sigma^2$, where $r = r(X)$. And $SS_{Res}/\sigma^2 = (n-r)s^2/\sigma^2$ has a χ^2 distribution with $\boldsymbol{n-r}$ d.f..

If $\mathbf{t}^T\boldsymbol{\beta}$ is **estimable**, then $\mathbf{t}^T\mathbf{b}$ is independent of s^2. # proof?

The steps to derive a CI are similar to that for the full rank case, but with two small differences. Firstly, we can only find CIs for quantities that are **estimable**. Secondly, we replace $(X^TX)^{-1}$ by $(X^TX)^c$.

The CI for the estimable quantity $\mathbf{t}^T\boldsymbol{\beta}$, using a t distribution with $n-r$ d.f., is $\mathbf{t}^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{\mathbf{t}^T(X^TX)^c\mathbf{t}}$.

To find the difference between class 3 and the average of the other two classes, using contr.treatment, we need

$$\tau_3 - \frac{1}{2}\tau_2 - \frac{1}{2}\tau_1 = (\tau_3 - \tau_1) - \frac{1}{2}(\tau_2 - \tau_1) = \text{class. f3} - \frac{1}{2}\text{class. f2}$$

```
> library(gmodels)
> ci <- estimable(model, c(0,-0.5,1), conf.int = 0.95)
> c(ci$Lower, ci$Upper)
[1] 1.041325 11.358675
```

Consider a linear model with only categorical predictor, written in matrix form as $\mathbf{y} = X_1\boldsymbol{\beta_1} + \boldsymbol{\epsilon_1}$. Suppose we add some continuous predictors, resulting in an expanded model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Now consider a quantity $\mathbf{t}^T\boldsymbol{\beta}$, where $\mathbf{t}^T = [\,\mathbf{t_1}^T \mid \mathbf{t_2}^T\,]$ is partitioned according to the categorical and continuous predictors. Show that if $\mathbf{t_1}^T\boldsymbol{\beta_1}$ is estimable in the first model, then $\mathbf{t}^T\boldsymbol{\beta}$ is estimable in the second model.

If you write $X = [\,X_1 \mid X_2\,]$, you may assume that $r(X) = r(X_1) + r(X_2)$.

Because $\mathbf{t_1}^T\boldsymbol{\beta_1}$ is estimable, the linear system $X_1^TX_1\mathbf{z_1} = \mathbf{t_1}$ can have a solution for $\mathbf{z_1}$. Because X_2 of full rank, $X_2^TX_2$ is nonsingular, and so the linear system $X_2^TX_2\mathbf{z_2} = \mathbf{t_2}$ always has $\mathbf{z_2} = (X_2^TX_2)^{-1}\mathbf{t_2}$ as its solution.

Firstly, easily observe that $r(X^TX) \le r([\,X^TX \mid \mathbf{t}\,])$ since if we add a new column vector $\mathbf{t}$ to X^TX, the dimension of the column space will not reduce, so is the rank.

Secondly,

$$\begin{aligned} r([\,X^TX \mid \mathbf{t}\,]) &= r\left(\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & \mathbf{t_1} \\ X_2^TX_1 & X_2X_2^T & \mathbf{t_2} \end{array}\right]\right) = r\left(\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & X_1^TX_1\mathbf{z_1} \\ X_2^TX_1 & X_2X_2^T & X_2^TX_2\mathbf{z_2} \end{array}\right]\right) \\ &= r\left(\begin{bmatrix} X_1^T & 0 \\ 0 & X_2^T \end{bmatrix}\left[\begin{array}{cc|c} X_1 & X_2 & X_1\mathbf{z_1} \\ X_1 & X_2 & X_2\mathbf{z_2} \end{array}\right]\right) \le r\left(\begin{bmatrix} X_1^T & 0 \\ 0 & X_2^T \end{bmatrix}\right) = r(X_1^T) + r(X_2^T) \\ &= r(X_1) + r(X_2) = r(X) = r(X^TX). \end{aligned}$$

Combine the two inequalities, we get $r(X^TX) = r([\, X^TX \mid \mathbf{t}\,])$.

That is, the linear system $X^TX\mathbf{z} = \mathbf{t}$ is consistent, hence $\mathbf{t}^T\boldsymbol{\beta}$ is estimable.

A linear hypothesis H_0 is testable if there exists a set of estimable functions $\mathbf{c}_1^T\boldsymbol{\beta}, \mathbf{c}_2^T\boldsymbol{\beta}, \dots, \mathbf{c}_m^T\boldsymbol{\beta}$ such that H_0 is true iff $\mathbf{c}_1^T\boldsymbol{\beta} = \mathbf{c}_2^T\boldsymbol{\beta} = \mathbf{c}_m^T\boldsymbol{\beta} = 0$, and $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_m$ are linearly independent.

Now recall that a linear function $\mathbf{c}^T\boldsymbol{\beta}$ is estimable iff $\mathbf{c}^T(X^TX)^cX^TX = \mathbf{c}^T$. Therefore $H_0 : C\boldsymbol{\beta} = \mathbf{0}$ is testable iff C is of full rank and $C(X^TX)^cX^TX = C$, where C is $m \times p$ of rank m.

Note that since $r = r((X^TX)^cX^TX) = r(X)$, the maximum number of linearly independent estimable functions in a less-than-full rank model is r, so $m \le r \le p$.

Therefore if $C\boldsymbol{\beta} = \mathbf{0}$ is **testable**, then the F statistic is similar to that for the full rank case

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T(C(X^TX)^cC^T)^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*)/m}{SS_{Res}/(n-r)} \sim F_{m,n-r}.$$

In a one-way classification model, if we look closer at the F statistic, we can see that if we are given s^2, the only other information that we need to test our hypothesis is the means, and numbers, of samples from the various populations.

$$X^TX = \begin{bmatrix} n & n_1 & n_2 & \cdots & n_k \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_k & 0 & 0 & \cdots & n_k \end{bmatrix}, \qquad (X^TX)^c = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & n_1^{-1} & 0 & \cdots & 0 \\ 0 & 0 & n_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & n_k^{-1} \end{bmatrix},$$

$$X^T\mathbf{y} = \begin{bmatrix} \sum_{ij} y_{ij} \\ \sum_j y_{1j} \\ \sum_j y_{2j} \\ \vdots \\ \sum_j y_{kj} \end{bmatrix}, \qquad \mathbf{b} = (X^TX)^cX^T\mathbf{y} = \begin{bmatrix} 0 \\ \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_k \end{bmatrix}.$$

Furthermore, any hypothesis that we can test in a one-factor model can be tested for each factor in an additive two-factor mode.

Interaction happens when one factor affects the effect of another factor. And the model becomes

$$y_{ijk} = \mu + \tau_i + \beta_j + \xi_{ij} + \epsilon_{ijk}, \quad i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n_{ij}.$$

We are often interested in testing whether there is interaction or not. However, the hypothesis $H_0 : \xi_{ij} = 0 \ \forall\, i, j$ is NOT even testable, nor is $H_0 : \xi_{ij}$ the same $\forall\, i, j$.

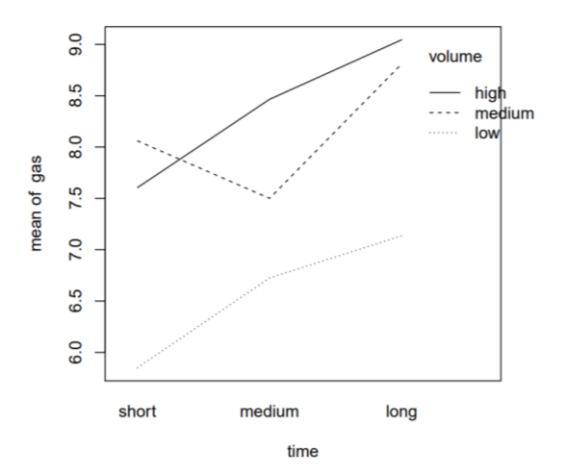
Instead, we say there is no interaction iff $(\xi_{ij} - \xi_{ij'}) - (\xi_{i'j} - \xi_{i'j'}) = 0 \ \forall\, i \neq i', j \neq j'$. These quantities are all estimable, since this is equivalent to $(\mu_{ij} - \mu_{ij'}) = (\mu_{i'j} - \mu_{i'j'})$, and the group means are all elements of $X\boldsymbol{\beta}$.

It turns out that $(a-1)(b-1)$ tests are sufficient.

Consider these things when testing for interaction:

1. If we have one sample per combination of factors, it is impossible to test for interaction. This is because $r = r(X) = n$ and therefore $n - r = 0$ minimally. The residual d.f. is 0, so we have no way to estimate the variance.
2. Even if we test for interaction and find that there is none, we cannot be sure that there is no interaction, we just haven't found any. However, for practical purposes, this may take away too many d.f. from the Residual SS.
3. It is possible to have interaction between three or more factors.

```
> engine <- read.csv("engine.csv")
> engine$time <- factor(engine$time)
> engine$volume <- factor(engine$volume)
> str(engine)
'data.frame':   18 obs. of  3 variables:
 $ gas   : num  6.27 8.08 7.34 5.43 8.04 7.87 6.94 7.48 8.61 6.51 ...
 $ volume: Factor w/ 3 levels "low","medium",..: 1 2 3 1 2 3 1 2 3 1 ...
 $ time  : Factor w/ 3 levels "short","medium",..: 1 1 1 1 1 1 2 ...
> with(engine, interaction.plot(time, volume, gas))
```




```
> library(MASS, Matrix)
> X <- matrix(c(rep(1,n),rep(0,n*15)), n, 7+9)
> X[cbind(1:n,as.numeric(engine$volume)+1)] <- 1
> X[cbind(1:n,as.numeric(engine$time)+4)] <- 1
> X[cbind(1:n,as.numeric(engine$time)*3+as.numeric(engine$volume)+4)] <- 1
> (r <- rankMatrix(X)[1])
[1] 9
> XtXc <- ginv(t(X)%*%X)
> b <- XtXc%*%t(X)%*%y
> s2 <- sum((y-X%*%b)^2)/(n-r)

## Test for interaction munually
> C <- matrix(0, 4, 16)
> C[1,c(8,9,11,12)] <- c(1,-1,-1,1)
> C[2,c(9,10,12,13)] <- c(1,-1,-1,1)
> C[3,c(11,12,14,15)] <- c(1,-1,-1,1)
> C[4,c(12,13,15,16)] <- c(1,-1,-1,1)
```

```
> C[,-(1:7)]
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]    1   -1    0   -1    1    0    0    0    0
[2,]    0    1   -1    0   -1    1    0    0    0
[3,]    0    0    0    1   -1    0   -1    1    0
[4,]    0    0    0    0    1   -1    0   -1    1
> (Fstat <- (t(C%*%b)%*%solve(C%*%XtXc%*%t(C))%*%
+                  C%*%b/4)/s2)
         [,1]
[1,] 4.47684
> pf(Fstat,4,n-r,lower=F)
           [,1]
[1,] 0.02891813

## Test for interaction using anova()
> amodel <- lm(gas ~ volume + time, data = engine)
> imodel <- lm(gas ~ volume * time, data = engine)
> imodel <- lm(gas ~ (volume + time) ^ 2, data = engine)
> anova(amodel, imodel)
..
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     13 2.08158
2      9 0.69625  4    1.3853 4.4768 0.02892 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Interaction significant, so include the interaction term
> summary(imodel)
..
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.8500     0.1967  29.745 2.68e-10 ***
volumemedium               2.2100     0.2781   7.946 2.34e-05 ***
volumehigh                 1.7550     0.2781   6.310 0.000139 ***
timemedium                 0.8750     0.2781   3.146 0.011815 *
timelong                   1.2850     0.2781   4.620 0.001254 **
volumemedium:timemedium   -1.4350     0.3933  -3.648 0.005333 **
volumehigh:timemedium     -0.0150     0.3933  -0.038 0.970413
volumemedium:timelong     -0.5350     0.3933  -1.360 0.206882
volumehigh:timelong        0.1550     0.3933   0.394 0.702715
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R's contr.treatment estimates:

Intercept	$\mu_{11} = \mu + \tau_1 + \beta_1 + \xi_{11}$
f2	$\mu_{21} - \mu_{11} = \tau_2 - \tau_1 + \xi_{21} - \xi_{11}$
f3	$\mu_{31} - \mu_{11} = \tau_3 - \tau_1 + \xi_{31} - \xi_{11}$
g2	$\mu_{12} - \mu_{11} = \beta_2 - \beta_1 + \xi_{12} - \xi_{11}$
g3	$\mu_{13} - \mu_{11} = \beta_3 - \beta_1 + \xi_{13} - \xi_{11}$
f2:g2	$\mu_{22} - \mu_{21} - \mu_{12} + \mu_{11} = \xi_{22} - \xi_{21} - \xi_{12} + \xi_{11}$
f3:g2	$\mu_{32} - \mu_{31} - \mu_{12} + \mu_{11} = \xi_{32} - \xi_{31} - \xi_{12} + \xi_{11}$
f2:g3	$\mu_{23} - \mu_{21} - \mu_{13} + \mu_{11} = \xi_{23} - \xi_{21} - \xi_{13} + \xi_{11}$
f3:g3	$\mu_{33} - \mu_{31} - \mu_{13} + \mu_{11} = \xi_{33} - \xi_{31} - \xi_{13} + \xi_{11}$

```
> b[1]+b[2]+b[5]+b[8]
[1] 5.85
> b[c(3,4)] - b[2] + b[c(9,10)] - b[8]
[1] 2.210 1.755
> b[c(6,7)] - b[5] + b[c(11,14)] - b[8]
[1] 0.875 1.285
> b[c(12,13,15,16)] - b[c(9,10,9,10)] - b[c(11,11,14,14)] + b[c(8,8,8,8)]
[1] -1.435 -0.015 -0.535 0.155
```

We can also do analysis of covariance (ANCOVA) using the linear model framework.

In this case, we have one (or more) categorical predictors and one (or more) continuous predictors. For example:

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \xi_i x_{ij} + \epsilon_{ij}.$$

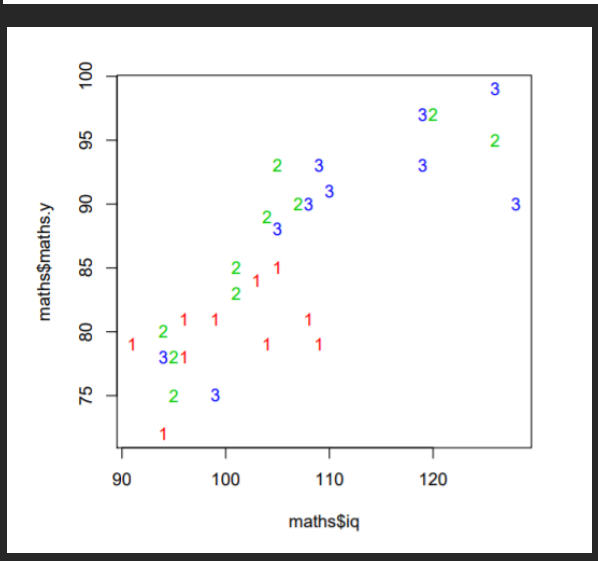
We can think of this simple model as fitting several regression lines, one to each population (assuming equal variances across populations).

Interaction in this case means that the slopes of the regression lines (effect of **continuous** predictor) are different for each population.

Suppose we fit the lines $y = \alpha_i + \beta_i x$ to each subpopulation. For the interaction model, R's contr.treatment estimates:

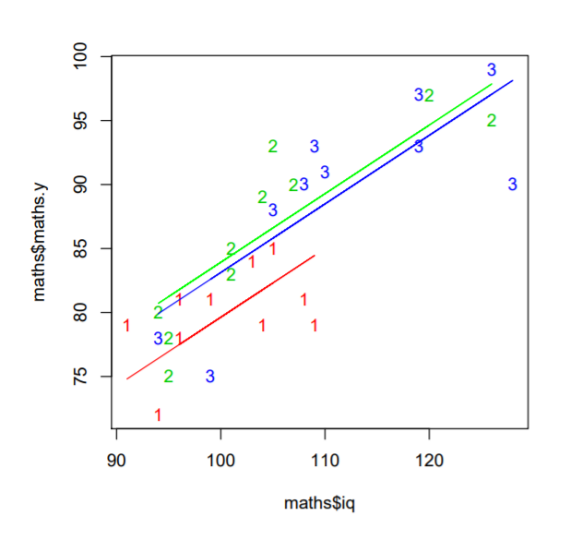
Intercept	$\alpha_1 = \mu + \tau_1$
f2	$\alpha_2 - \alpha_1 = \tau_2 - \tau_1$
f3	$\alpha_3 - \alpha_1 = \tau_3 - \tau_1$
x	$\beta_1 = \beta + \xi_1$
f2:x	$\beta_2 - \beta_1 = \xi_2 - \xi_1$
f3:x	$\beta_3 - \beta_1 = \xi_3 - \xi_1$

```
> plot(maths$iq, maths$maths.y, pch=array(maths$class.f),
+         col=maths$class+1)
```



```
## Test for interaction
> amodel <- lm(maths.y ~ class.f + iq, data = maths)
> imodel <- lm(maths.y ~ class.f * iq, data = maths)
> anova(amodel, imodel)
..
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     26 423.42
2     24 392.36  2    31.062 0.95 0.4008
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Interaction not significant, so fit an additive model
> summary(amodel)
..
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.02809    8.23338   3.161  0.00396 **
class.f2     4.29503    1.83799   2.337  0.02743 *
class.f3     3.49636    2.01959   1.731  0.09526 .
iq           0.53604    0.08093   6.623 5.03e-07 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Why X is called the design matrix? Because X has been designed.

An observational study cannot determine causality, only correlation. Instead, you need a designed experiment/trial. Even then, we can never be sure.

Hill's criteria (a designed experiment can help with the *'s):

1. Strength of association.
2. Consistency of association: Since the 1600's there has been global warming and a decrease in pirates, therefore pirates help stop global warming?
3. Consistent with existing knowledge.
4. * Monotonic response (increasing A makes B more likely).
5. * Temporal relationship: The more firemen, the bigger the fire, therefore firemen cause fires?
6. * Plausibility of alternatives, i.e., a nuisance/confounding factor: Sleeping with shoes on causes headaches? (Drinking the night before.)
7. Predictive value of link.

Experimental design aims to avoid the effects of confounding factors, known and unknown, by:

1. Control and comparison: Keep everything the same except the variables you know about. Often a group with no treatment is used as a basis for comparison, called a control group.
2. Blocking: Partition the population into homogeneous groups in terms of the confounding variables.
3. Randomisation: Randomly assign units to different treatments, to avoid lurking factors (confounding factors we are unaware of).
4. Blind: In a blind experiment, the patients do not know if they are being treated or not. In a double-blind experiment, neither the patients nor those administering treatments know which treatment is being given to whom. We do this to avoid response bias and the placebo effect.

Suppose we have a factor of interest - the treatment - and zero or more confounding/ blocking factors, which are dealt with by blocking.

Complete randomised design (CRD): No blocking factor (1-way classification).

```
> n <- c(5,6,4)
> nsum <- sum(n)
> x <- sample(nsum, nsum)
> j1 <- x[1:n[1]]
> j2 <- x[n[1]+(1:n[2])]
> j3 <- x[(n[1]+n[2]+1):nsum]
```

In a CRD with n test units, an optimal allocation of test units to factor levels tries to minimise the variance of statistics:

$$\sum_{i=1}^{k} \mathbb{V}[\hat{\mu}_i] = \sigma^2 \sum_{i=1}^{k} n_i^{-1}$$

Using Lagrangian multipliers to deal with the constraint in sample size. Take

$$f(n_1, \dots, n_k, \lambda) = \sigma^2 \sum_{i=1}^{k} n_i^{-1} + \lambda \left(\sum_{i=1}^{k} n_i - n \right).$$

We minimise this function with respect to all variables; the equation $\partial f / \partial \lambda = 0$ ensures that the total sample size is constrained to n.

$$\frac{\partial f}{\partial n_i} = -\frac{\sigma^2}{n_i^2} + \lambda = 0 \;\; \Rightarrow \;\; n_i^2 = \frac{\sigma^2}{\lambda}.$$

This does not depend on i, so to satisfy these equations we must choose all n_i equal.

How to minimise the variance of other statistics, e.g., treatment contrasts (against?)?

Complete block design (CBD): One blocking factor (2-way classification).

Partition the experimental units into b blocks, each having size of [an integer multiple of] k (the number of treatment), which are homogeneous in the blocking factor. In different blocks, an equal number of units receive each treatment, assigned randomly.

$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij},\ i = 1, \dots, b,\ j = 1, \dots, k,$ where β_i are the block effects and τ_j are the treatment effects.

$$\begin{bmatrix} y_2 \\ y_1 \\ y_3 \\ \hline y_4 \\ y_5 \\ y_6 \\ \hline y_7 \\ y_8 \\ y_9 \\ \hline y_{10} \\ y_{11} \\ y_{12} \end{bmatrix} = \left[\begin{array}{ccccc|ccc} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \hline \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \epsilon_2 \\ \epsilon_1 \\ \epsilon_3 \\ \hline \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \hline \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \hline \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{bmatrix}$$

In the general linear model, write

$$\mathbf{y} = [\, X_1 \mid X_2 \,] \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \boldsymbol{\epsilon}.$$

Then $\begin{bmatrix}\mathbf{b_1}\\ \mathbf{b_2}\end{bmatrix}$ is a solution to the normal equations only if $\mathbf{b_2}$ is a solution to the reduced normal equations: $X_2^T[I-H_1]X_2\mathbf{b_2} = X_2^T[I-H_1]\mathbf{y}$, where $H_1 = X_1(X_1^TX_1)^cX_1^T$.

$$X^TX\mathbf{b} = X^T\mathbf{y}$$

$$\begin{bmatrix}X_1^T\\ X_2^T\end{bmatrix}[\,X_1 \mid X_2\,]\begin{bmatrix}\mathbf{b_1}\\ \mathbf{b_2}\end{bmatrix} = \begin{bmatrix}X_1^TX_1\mathbf{b_1}+X_1^TX_2\mathbf{b_2}\\ X_2^TX_1\mathbf{b_1}+X_2^TX_2\mathbf{b_2}\end{bmatrix} = \begin{bmatrix}X_1^T\mathbf{y}\\ X_2^T\mathbf{y}\end{bmatrix} \Leftrightarrow \begin{cases}X_1^TX_1\mathbf{b_1}+X_1^TX_2\mathbf{b_2} = X_1^T\mathbf{y}\\ X_2^TX_1\mathbf{b_1}+X_2^TX_2\mathbf{b_2} = X_2^T\mathbf{y}\end{cases}$$

$$X_2^TX_1(X_1^TX_1)^c(X_1^TX_1\mathbf{b_1}+X_1^TX_2\mathbf{b_2}) = X_2^TX_1(X_1^TX_1)^cX_1^T\mathbf{y} \quad [1]$$

$$X_2^TX_1\mathbf{b_1}+X_2^TH_1X_2\mathbf{b_2} = X_2^TH_1\mathbf{y}$$

$$X_2^T[I-H_1]X_2\mathbf{b_2} = X_2^T[I-H_1]\mathbf{y} \quad [2]$$

$$([I-H_1]X_2)^T[I-H_1]X_2\mathbf{b_2} = ([I-H_1]X_2)^T\mathbf{y}$$

The reduced normal equations are the same as the normal equation for the linear model $\mathbf{y} = [I-H_1]X_2\,\boldsymbol{\beta_2}+\boldsymbol{\epsilon}$. We write $X_{2|1} = [I-H_1]X_2$, called the reduced design matrix.

If we have k treatments and b blocks, each of size k. Let J_k be the $k\times k$ matrix of 1's, and $\mathbf{1_k}$ the $k\times 1$ vector of 1's. Then we have

$$X_1 = \begin{bmatrix}\mathbf{1_k} & \mathbf{1_k} & \cdots & 0\\ \vdots & & \ddots & \\ \mathbf{1_k} & 0 & & \mathbf{1_k}\end{bmatrix}, \qquad X_1^TX_1 = \begin{bmatrix}n & k & \cdots & k\\ k & k & & 0\\ \vdots & & \ddots & \\ k & 0 & & k\end{bmatrix},$$

$$H_1 = \begin{bmatrix}\mathbf{1_k} & \mathbf{1_k} & \cdots & 0\\ \vdots & & \ddots & \\ \mathbf{1_k} & 0 & & \mathbf{1_k}\end{bmatrix}\begin{bmatrix}0 & 0 & \cdots & 0\\ \vdots & k^{-1} & & 0\\ 0 & & \ddots & \\ 0 & 0 & & k^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{1_k^T} & \cdots & \mathbf{1_k^T}\\ \mathbf{1_k^T} & & 0\\ & \ddots & \\ 0 & & \mathbf{1_k^T}\end{bmatrix}$$

$$= \frac{1}{k}\begin{bmatrix}\mathbf{1_k} & \mathbf{1_k} & \cdots & 0\\ \vdots & & \ddots & \\ \mathbf{1_k} & 0 & & \mathbf{1_k}\end{bmatrix}\begin{bmatrix}0 & \cdots & 0\\ \mathbf{1_k^T} & & 0\\ & \ddots & \\ 0 & & \mathbf{1_k^T}\end{bmatrix} = \frac{1}{k}\begin{bmatrix}J_k & 0 & \cdots & 0\\ 0 & J_k & & 0\\ \vdots & & \ddots & \\ 0 & 0 & & J_k\end{bmatrix},$$

$$X_{2|1} = [I-H_1]X_2 = \left(I-\frac{1}{k}\begin{bmatrix}J_k & 0 & \cdots & 0\\ 0 & J_k & & 0\\ \vdots & & \ddots & \\ 0 & 0 & & J_k\end{bmatrix}\right)\begin{bmatrix}I_k\\ \vdots\\ I_k\end{bmatrix} = \begin{bmatrix}I_k\\ \vdots\\ I_k\end{bmatrix}-\frac{1}{k}\begin{bmatrix}J_k\\ \vdots\\ J_k\end{bmatrix},$$

$$X_{2|1}^TX_{2|1} = [\,I_k\ \dots\ I_k\,]\begin{bmatrix}I_k\\ \vdots\\ I_k\end{bmatrix}-\frac{1}{k}[\,I_k\ \dots\ I_k\,]\begin{bmatrix}J_k\\ \vdots\\ J_k\end{bmatrix}-\frac{1}{k}[\,J_k\ \dots\ J_k\,]\begin{bmatrix}I_k\\ \vdots\\ I_k\end{bmatrix}$$
$$+\frac{1}{k^2}[\,J_k\ \dots\ J_k\,]\begin{bmatrix}J_k\\ \vdots\\ J_k\end{bmatrix} = bI_k-\frac{b}{k}J_k-\frac{b}{k}J_k+\frac{b}{k^2}kJ_k$$
$$= b\left[I_k-\frac{1}{k}J_k\right],$$

$$\left(X_{2|1}^T X_{2|1}\right)\left(\frac{1}{b}I_k\right)\left(X_{2|1}^T X_{2|1}\right) = X_{2|1}^T X_{2|1} \;\Rightarrow\; \left(X_{2|1}^T X_{2|1}\right)^c = \frac{1}{b}I_k\,,$$

$$\mathbf{b_2} \;=\; \left(X_{2|1}^T X_{2|1}\right)^c X_{2|1}^T \mathbf{y} \;=\; \frac{1}{b}I_k\left[\,[\,I_k \;\ldots\; I_k\,] - \frac{1}{k}[\,J_k \;\ldots\; J_k\,]\right]\begin{bmatrix} y_{11} \\ \vdots \\ y_{1k} \\ \vdots \\ y_{b1} \\ \vdots \\ y_{bk} \end{bmatrix} = \begin{bmatrix} \bar{y}_{\bullet 1} - \bar{y}_{\bullet\bullet} \\ \vdots \\ \bar{y}_{\bullet k} - \bar{y}_{\bullet\bullet} \end{bmatrix}.$$

In a CBD, if $\mathbf{t}^T\boldsymbol{\tau}$ is estimable, then

$$\mathbf{t}^T = \mathbf{t}^T\left(X_{2|1}^T X_{2|1}\right)^c\left(X_{2|1}^T X_{2|1}\right) = \mathbf{t}^T\frac{1}{b}I_k\; b\left[I_k - \frac{1}{k}J_k\right] \;\Rightarrow\; \mathbf{t}^T\mathbf{1_k} = \mathbf{t}^T\left[I_k - \frac{1}{k}J_k\right]\mathbf{1_k} = 0.$$

That is, $\mathbf{t}$ must be a contrast. Thus, if $\mathbf{t}^T\boldsymbol{\tau}$ is estimable, then its estimate is

$$\mathbf{t}^T\begin{bmatrix} \bar{y}_{\bullet 1} - \bar{y}_{\bullet\bullet} \\ \vdots \\ \bar{y}_{\bullet k} - \bar{y}_{\bullet\bullet} \end{bmatrix} = \mathbf{t}^T\begin{bmatrix} \bar{y}_{\bullet 1} \\ \vdots \\ \bar{y}_{\bullet k} \end{bmatrix}.$$

This is what we would get if we ignored the blocks and treated the experiment as a CRD.

However, for CBD, $y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$,

$$\mathbb{V}\left[\widehat{\mathbf{t}^T\boldsymbol{\tau}}\right] = \mathbb{V}[\mathbf{t}^T\mathbf{b_2}] = \mathbb{V}\left[\mathbf{t}^T\left(X_{2|1}^T X_{2|1}\right)^c X_{2|1}^T\mathbf{y}\right] = \mathbf{t}^T\left(X_{2|1}^T X_{2|1}\right)^c X_{2|1}^T\sigma^2 I X_{2|1}\left(X_{2|1}^T X_{2|1}\right)^c\mathbf{t}$$
$$= \mathbf{t}^T\left(X_{2|1}^T X_{2|1}\right)^c\mathbf{t}\,\sigma^2 = \mathbf{t}^T\left(\frac{1}{b}I_k\right)\mathbf{t}\,\sigma^2 = \frac{1}{b}\sum_{i=1}^{k} t_i^2\sigma^2\,;$$

For CRD, $y_{ij} = \mu + \tau_i + \epsilon'_{ij}$,

$$\mathbb{V}\left[\widehat{\mathbf{t}^T\boldsymbol{\tau}}\right] = \frac{1}{b}\sum_{i=1}^{k} t_i^2\sigma^{2\prime} = \frac{1}{b}\sum_{i=1}^{k} t_i^2\mathbb{V}[\beta_u + \epsilon_{iu}] = \frac{1}{b}\sum_{i=1}^{k} t_i^2\left(\sigma^2 + \frac{1}{b}\sum_{j=1}^{b}\left(\beta_j - \bar{\beta}_\bullet\right)^2\right).$$

Thus, if there is an effect due to the blocks, then ignoring it means your estimators are less accurate than they could be.

However, note that in CRD, s^2 has $(bk - k)$ d.f., while in CBD, s^2 has $\big(bk - (b + k - 1)\big)$ d.f.. If there is really no blocking effect, then the CRD is better than the CBD.

In theory, a CBD can treat the two confounding factors A and B as a single confounding factor C, where the levels of C are given by all possible combinations of levels from A and B. Latin squares are used as a more efficient way than a CBD. However, they require **the treatment and both blocking factors to have the same number of levels**, t say.

A Latin square is a $t \times t$ matrix which contains the numbers 1 to t such that every number occurs once in each row and column, chosen randomly.

```
> library(magic)
> rlatin(5)
     [,1] [,2] [,3] [,4] [,5]
[1,]    3    5    2    4    1
[2,]    5    3    1    2    4
[3,]    1    2    4    5    3
[4,]    2    4    3    1    5
[5,]    4    1    5    3    2
```

The model for a Latin square design is a 3-way classification model: $y_{ijk} = \mu + \beta_i + \gamma_j + \tau_k + \epsilon_{ijk}$, where $1 \le i, j \le t$ and $k = k(i,j)$ is determined by i and j.

$$\mathbf{y} = X_1 \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} + X_2\boldsymbol{\tau} + \boldsymbol{\epsilon}, \quad \text{where } X_1 = \begin{bmatrix} \mathbf{1_t} & \mathbf{1_t} & 0 & \cdots & 0 & I_t \\ \mathbf{1_t} & 0 & \mathbf{1_t} & \cdots & 0 & I_t \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1_t} & 0 & 0 & \cdots & \mathbf{1_t} & I_t \end{bmatrix}, \; X_2 = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_t \end{bmatrix},$$

and each P_i is a <u>permutation matrix</u> (zero except for a single 1 for each row and column), such that $\sum_{i=1}^{t} P_t = J_t$.

$$X_1^T X_2 = \begin{bmatrix} \mathbf{1_t}^T & \cdots & \mathbf{1_t}^T \\ \mathbf{1_t}^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{1_t}^T \\ I_t & \cdots & I_t \end{bmatrix} \begin{bmatrix} P_1 \\ \vdots \\ P_t \end{bmatrix} = \begin{bmatrix} t\mathbf{1_t}^T \\ J_t \\ J_t \end{bmatrix}, \quad X_1^T \frac{1}{t} J_{t^2 \times t} = \begin{bmatrix} t\mathbf{1_t}^T \\ J_t \\ J_t \end{bmatrix}$$

$$\Rightarrow \; H_1 X_2 = X_1 (X_1^T X_1)^c X_1^T \frac{1}{t} J_{t^2 \times t} = H_1 \frac{1}{t} J_{t^2 \times t} = \frac{1}{t} J_{t^2 \times t},$$

$$X_{2|1} = \begin{bmatrix} P_1 \\ \vdots \\ P_t \end{bmatrix} - \frac{1}{t} \begin{bmatrix} J_t \\ \vdots \\ J_t \end{bmatrix}, \quad X_{2|1}^T X_{2|1} = t\left[I_t - \frac{1}{t} J_t \right], \quad \left(X_{2|1}^T X_{2|1} \right)^c = \frac{1}{t} I_t,$$

$$\mathbf{b_2} = \left(X_{2|1}^T X_{2|1} \right)^c X_{2|1}^T \mathbf{y} = \frac{1}{t} \left[\begin{bmatrix} P_1^T & \dots & P_t^T \end{bmatrix} - \frac{1}{t} \begin{bmatrix} J_t & \dots & J_t \end{bmatrix} \right] \mathbf{y} = \begin{bmatrix} \bar{y}_{\bullet\bullet 1} - \bar{y}_{\bullet\bullet\bullet} \\ \vdots \\ \bar{y}_{\bullet\bullet t} - \bar{y}_{\bullet\bullet\bullet} \end{bmatrix}.$$

This result is again the same as the CRD.

<u>Balanced incomplete block design (BIBD)</u>: Suppose we have t treatment levels, and b blocks of size $k < t$, e.g., twins (2). One blocking factor but CBD impossible.

A design is called BIBD if

1. Each treatment occurs at most once in a block;
2. Each treatment occurs exactly $r = bk/t$ times (first-order balance);
3. Each pair of treatments occurs in the same number of blocks, say λ (second-order balance). There are $t(t-1)/2$ different pairs of treatments and $bk(k-1)/2$ available slots, so we must have $\lambda = r(k-1)/(t-1)$.

Given t and k, we can always find a BIBD with $b = \binom{t}{k}$ blocks, by taking all possible subsets of size k. In this case,

$$r = \binom{t}{k}\frac{k}{t} = \binom{t-1}{k-1} \quad \text{and} \quad \lambda = \binom{t-1}{k-1}\frac{k-1}{t-1} = \binom{t-2}{k-2}.$$

The model for a BIBD is: $y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$ for $1 \le i \le b,\ j \in S(i),$ where $S(i)$ are the treatments in block i.

$$\mathbf{y} = X_1\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\beta}\end{bmatrix} + X_2\boldsymbol{\tau} + \boldsymbol{\epsilon}, \quad \text{where } X_1 = \begin{bmatrix}\mathbf{1_k} & \mathbf{1_k} & 0 & \cdots & 0\\ \mathbf{1_k} & 0 & \mathbf{1_k} & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{1_k} & 0 & 0 & \cdots & \mathbf{1_k}\end{bmatrix},\ X_2 = \begin{bmatrix}T_1\\ T_2\\ \vdots\\ T_n\end{bmatrix}.$$

The reduced normal equations for this model can be written as

$$\frac{\lambda t}{k}\left[I_t - \frac{1}{t}J_t\right]\mathbf{b_2} = \mathbf{t} - X_2^T X_1 \mathbf{b} =: \mathbf{q}\,,$$

where $\mathbf{t} = [\, y_{\bullet 1} \ \dots \ y_{\bullet t}\,]^T$ are the treatment totals, and $\mathbf{b} = [\, y_{1\bullet} \ \dots \ y_{\mathrm{b}\bullet}\,]^T$ are the block totals.

$$\left(X_{2|1}^T X_{2|1}\right)^c = \frac{k}{\lambda t} I_t, \quad \mathbf{b_2} = \frac{k}{\lambda t}\mathbf{q}.$$

Here $\mathbf{q}$ can be thought of as the "adjusted" treatment totals after the effect of the blocks is taken into account. For a CBD, the presence of the blocks does not affect $\mathbf{b_2}$, but this is not the case for a BIBD.

A collection of random variables $(X_t)_{t\in I}$ on a common probability space $(\Gamma, \mathcal{F}, \mathbb{P})$ is called a stochastic process. The index variable t is called "time".

If both ω and t are fixed, then $X_t(\omega)$ is a real number. For a fixed t, the function $X_t : \Omega \to \mathbb{R}$ is a random variable. For a fixed ω, the function $X_\bullet(\omega) : I \to \mathbb{R}$ is a trajectory (step function) of t. If we allow ω to vary, we get a collection of trajectories.

Assume that the transition probability $p_{i,j}(t) \coloneqq \mathbb{P}(X_{t+1} = j \mid X_t = i)$ do not depend on t, in which case the DTMC is called time-homogeneous, and we write $p_{i,j} \coloneqq p_{i,j}(t)$.

The initial distribution is denoted by $\boldsymbol{\pi}^{(0)} \coloneqq \left(\pi_i^{(0)}\right)_{i\in S}$.

If the state space S has $m \in \mathbb{N}$ elements, then by relabelling values if necessary, we may assume that $S = \{1, 2, \dots, m\}$. Then define the (one step)-transition matrix whose ij-th entry is $p_{i,j}$, and each row sums to 1.

Strong Markov property: Let $(X_t)_{t\in\mathbb{Z}}$ be a DTMC, and T be a stopping time for the chain. Then $\mathbb{P}(X_{T+1} = j \mid T = t, X_0 = x_0, \dots, X_T = i) = \mathbb{P}(X_{T+1} = j \mid T < \infty, X_T = i) = p_{i,j}$. This says that looking at the next step of a Markov chain at a stopping time is the same as starting the process from the random state X_T.

A state j is an absorbing state if $p_{j,j} = 1$.

State j is accessible from state i, denoted by $i \to j$, if there exists an $n \geq 0$ such that $p_{i,j}^{(n)} > 0$. That is, either $j = i$ or we can get from i to j in a finite number of steps.

If $i \to j$ and $j \to i$, then states i and j communicate, denoted by $i \leftrightarrow j$. The communication relation is known as an equivalence relation, which has the properties of reflexivity, symmetry and transitivity.

Partition the state space S of a DTMC into several communicating classes such that $i, j \in S_m$ iff $i \leftrightarrow j$. If a DTMC has only one communicating class, then it is called irreducible.

Let

1. $h_{i,j} \coloneqq \mathbb{P}(j \in \{X_0, X_1, \dots\} \mid X_0 = i)$, be the hitting probability that we ever reach state j, starting from state i.
   - $h_{i,i} = 1$.
   - $h_{i,j} > 0$ iff $i \to j$.
2. $f_i \coloneqq \mathbb{P}(i \in \{X_1, X_2, \dots\} \mid X_0 = i)$, be the return probability.
3. $T(j) \coloneqq \inf\{n \geq 0 : X_n = j\}$, be the first hitting (stopping) time of j.
4. $m_{i,j} \coloneqq \mathbb{E}[T(j) \mid X_0 = i]$, be the expected time to reach state j starting from state i.
   - $m_{j,j} = 0$.
   - If $h_{i,j} < 1$ then $\mathbb{P}(T(j) = \infty \mid X_0 = i) > 0$ so $m_{i,j} = \infty$.
   - If $h_{i,j} = 1$ then $m_{i,j}$ may or may not be finite.

5. $\mu_i \coloneqq \mathbb{E}[T^+(i)|\, X_0 = i]$, is the expected time to return to state i.
6. $\triangle_j\,(i)$, be the time between the $(j+1)$'st and j'th visit to state i. $N(i)$, be the number of visits to state i.
   - If $f_i = 1$, then each $\triangle_j\,(i)$ is finite since the chain is certain to return to i in finite time, and $N(i) = \infty$.
   - If $f_i < 1$, then with probability $1 - f_i > 0$ it will never return to i. From the Markov probability we see that $N(i)$ has a geometric distribution. Specifically, for $n \geq 1$, $\mathbb{P}(N(i) = n|\, X_0 = i) = f_i^{n-1}(1 - f_i)$.
7. $\gamma_i \coloneqq \mathbb{E}[N(i)|X_0 = i] = (1 - f_i)^{-1}$.

For an irreducible DTMC, the following are equivalent:

1. The chain is recurrent, i.e., $h_{i,j} = 1$ for any and all $i, j \in S$. Otherwise, it is transient.
2. $f_i = 1$ for every $i \in S$.
3. $f_i = 1$ for some $i \in S$.
4. $\gamma_i = \infty$ for every $i \in S$.
5. $\gamma_i = \infty$ for some $i \in S$.

(3) $\Leftrightarrow$ (5): $\gamma_i = (1-1)^{-1} = \infty$.

(1) $\Rightarrow$ (2): $f_i = \sum_{j \in S} p_{i,j} h_{j,i} = \sum_{j \in S} p_{i,j} = 1$.

(3) $\Rightarrow$ (1): Let $j \in S$. Since $i \to j$ we must have $h_{j,i} = 1$ (otherwise the chain could escape form i by visiting j). And every time the chain reaches state i it has a fixed probability of hitting j before returning to i, so $h_{i,j} = 1$. Now for $j, k \in S$ since we are guaranteed to hit i starting from j and guaranteed to hit k started from i, we are guaranteed to hit k starting from j, i.e., $h_{j,k} = 1$. …

Irreducible finite-state chains are recurrent.

Let $S = \{1, \dots, k\}$. Now $\sum_{j=1}^{k} N(j) = \infty$, so for any i with $\mathbb{P}(X_0 = i) > 0$, we have $\mathbb{E}\left[\sum_{j=1}^{k} N(j) \,\middle|\, X_0 = i\right] = \infty$. Thus, there must exist $j \in S$ such that $\mathbb{E}[N(j)|X_0 = i] = \infty$. For this j, $\gamma_j \geq \mathbb{E}[N(j)|\, X_0 = i] = \infty$.

Let $(Y_i)_{i \in \mathbb{N}}$ be i.i.d. random variables with $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = -1) = 1 - p$. Let $S_n = \sum_{i=1}^{n} Y_i$ for each $n \in \mathbb{N}$. Then $(S_n)_{n \in \mathbb{Z}_+}$ is a DTMC called a simple random walk.

1. Note that the expected number of visits to 0 is

   $\gamma_0 \coloneqq \mathbb{E}[N(0)|\, X_0 = 0] = \sum_{m=0}^{\infty} \mathbb{E}\left[\mathbf{1}_{\{S_m = 0\}}\right] = \sum_{m=0}^{\infty} \mathbb{P}(S_m = 0) = \sum_{m=0}^{\infty} p_{0,0}^{(m)}$.

   Note that $p_{j,j}^{(m)} = 0$ if m is odd. If $m = 2n$, then $p_{j,j}^{(m)} = p_{j,j}^{(2n)} = \binom{2n}{n} p^n (1-p)^n$.

Stirling's approximation: $n! \approx \sqrt{2\pi n} n^n e^{-n}$, gives us the fact that $p_{j,j}^{(2n)} \approx \big(4p(1-p)\big)^n/\sqrt{n\pi}$, and so series $\sum_{m=0}^{\infty} p_{0,0}^{(m)}$ diverges if $p = 1/2$, and converges if $p \neq 1/2$ (transient).

2. If $i > 0$, then $h_{i,0} = ph_{i+1,0} + (1-p)h_{i-1,0}$. Let $x = h_{1,0}$. We have $x = ph_{2,0} + 1 - p = px^2 + 1 - p$. Solving the quadratic gives solutions $x = 1$, and $x = (1-p)/p$.

   If $p \leq 1/2$, then $h_{1,0} = 1$ is the only possible value for $h_{1,0}$, and in this case $h_{i,0} = 1$ for every $i > 0$. If $p > 1/2$, then the walk is transient to the right, so $h_{i,0}$ cannot be 1 for $i > 0$. Thus $h_{1,0} = (1-p)/p$ and $h_{i,0} = \big((1-p)/p\big)^i$.

   The vector of hitting probabilities $\big(h_{i,B}\big)_{i\in S}$ is the unique minimal nonnegative solution to the equations

$$h_{i,B} = \begin{cases} 1, & \text{if } i \in B \\ \sum_{j\in S} p_{i,j} h_{j,B}\,, & \text{otherwise.} \end{cases}$$

   Let $h_{i,B}^{(n)} \coloneqq \mathbb{P}(T(B) \leq n |\, X_0 = i)$, and $(x_i)_{i\in S}$ be a nonnegative solution. We'll show that $h_{i,B}^{(n)} \leq x_i$ for each $n \in \mathbb{Z}_+$, by induction. This is sufficient since $h_{i,B} = \lim_{n\to\infty} h_{i,B}^{(n)}$.

   For $n = 0$, note that $h_{i,B}^{(0)} = 1$ if $i \in B$ and $h_{i,B}^{(0)} = 0$ if $i \notin B$. Since $(x_i)_{i\in S}$ are nonnegative and equal 1 for $i \in B$, we have $h_{i,B}^{(0)} \leq x_i$ for all $i \in S$. Proceeding by induction, suppose that $h_{i,B}^{(n)} \leq x_i$ for all $i \in S$. Then $h_{i,B}^{(n+1)} = \sum_{j\in S} p_{i,j} h_{j,B}^{(n)} \leq \sum_{j\in S} p_{i,j} x_j = x_i$.

   The vector $\big(m_{i,A}\big)_{i\in S}$ of mean hitting times is the minimal nonnegative solution to

$$m_{i,A} = \begin{cases} 0, & \text{if } i \in A \\ 1 + \sum_{j\in S} p_{i,j} m_{j,A}\,, & \text{otherwise.} \end{cases}$$

3. Recall that, to solve

$$a(d^2y/dt^2) + b(dy/dt) + cy = 0,$$

   we try a solution of the form $y = y(t) = e^{\lambda t}$ to obtain the characteristic equation $a\lambda^2 + b\lambda + c = 0$. If the equation has distinct roots, λ_1 and λ_2, the general solution has the form $y = Ae^{\lambda_1 t} + Be^{\lambda_2 t}$. If the roots are coincident, the general solution has the form $y = Ae^{\lambda_1 t} + Bte^{\lambda_1 t}$. In both cases, the values of the constants A and B are determined by the initial conditions.

   The method for solving second-order linear difference equation with constant coefficients is similar. To solve

$$av_{i+1} + bv_i + cv_{i-1} = 0,$$

   we try a solution of the form $v_i = z^i$ to obtain the characteristic equation $az^2 + bz + c = 0$. The characteristic equation of $h_{i,0} = ph_{i+1,0} + (1-p)h_{i-1,0}$ is $pz^2 - z + (1-p) = 0$, which has the general solution for $i \geq 1$ is of the form:

$$h_{i,0} = \begin{cases} A + B\big((1-p)/p\big)^i & \text{if } (1-p)/p > 1 \\ A + Bi & \text{if } (1-p)/p = 1. \end{cases}$$

In either case, these can only be probabilities if $B = 0$ and then notice $A = h_{1,0} = ph_{2,0} + (1-p) = pA + (1-p)$, so $A = 1$.

A state $i \in S$ has period $d(i) \geq 1$ if $\left\{n \geq 1 : p_{i,i}^{(n)} > 0\right\}$ is nonempty and has greatest common divisor $d(i)$. If state i has period 1, then we say that it is aperiodic (otherwise it is periodic).

If $i \leftrightarrow j$, then $d(i) = d(j)$.

Assume that state i has period $d(i)$ and $i \leftrightarrow j$. Then, as before, there must exist s and t such that $p_{i,j}^{(s)} > 0$ and $p_{j,i}^{(t)} > 0$. We know straight away that $d(i)$ divides $s + t$.

Now take a path from j to itself in r steps, and we have an $s + r + t$ step path from i to itself. So $d(i)$ divides $s + r + t$ which means that $d(i)$ divides r. So $d(i)$ divides $d(j)$. (?)

Now we can switch i and j in the argument to conclude that $d(j)$ divide $d(i)$ which means that $d(i) = d(j)$, and all states in the same communicating class have a common period.

An irreducible DTMC is periodic with period d if every states have period $d > 1$.

A recurrent DTMC is positive recurrent if $\mu_i < \infty$ for every $i \in S$. Otherwise, it is null recurrent. The simple random walk with $p = 1/2$ is null recurrent.

Any irreducible DTMC with finite state space is positive recurrent.

There should exist $n_0 \in \mathbb{N}$ and $\epsilon > 0$, such that $\mathbb{P}(T(i) \leq n_0 | X_0 = k) > \epsilon$ for every $k \in S$. Starting from any state j that is reached from i, observe whether the chain has gone back to i within n_0 steps. This has probability at least ϵ. If it has not reached i then observe it for the next n_0 steps.

The number of blocks of n_0 steps that we have to observe is dominated by a Geometric(ϵ) random variable. Therefore, $m_{j,i} \leq n_0/\epsilon$ for every j. Thus, $\mu_i \leq 1 + n_0/\epsilon < \infty$.

$(*)$ Let $(X_i)_{i\in\mathbb{Z}}$ be an irreducible DTMC with state space S, and $N_n(j) = \sum_{i=0}^{n-1} \mathbf{1}_{\{X_i = j\}}$ denote the number of visits to state j before time n. Then

$$\mathbb{P}(N_n(i)/n \to \mu_i^{-1}) = 1,$$

where we interpret the limit μ_i^{-1} as $0 = 1/\infty$ if the chain is not positive recurrent (null recurrent or transient).

A vector $\boldsymbol{\pi} = (\pi_i)_{i\in S}$ with nonnegative entries is a stationary measure for a stochastic matrix P if $\pi_i = \sum_{j\in S} \pi_j p_{j,i}$ for each $i \in S$. These equations are called the full balance equations. If $\sum_{i\in S} \pi_i = 1$ then $\boldsymbol{\pi}$ is called a stationary distribution for P.

$(**)$ An irreducible DTMC with countable state space S has a stationary measure. It has a unique stationary distribution iff the chain is positive recurrent, and in this case $\pi_i = \mu_i^{-1}$ for each $i \in S$.

Combining with $(*)$, we see that for a irreducible and positive recurrent DTMC, the long-run proportion of time spent in state i is the stationary probability π_i.

A distribution $(a_i)_{i\in S}$ is called the limiting probability for a DTMC $(X_t)_{t\in\mathbb{Z}_+}$ if $\lim_{n\to\infty} \mathbb{P}(X_n = i) = a_i$ for each $i \in S$. A limiting distribution needs not exist, but if it exists, it's unique and depends on both the initial distribution and P. The simple random walk has no limiting distribution, with any p value.

For a DTMC, if a limiting distribution exists then it is a stationary distribution.

$(***)$ Let $(X_n)_{n\in\mathbb{Z}_+}$ be an irreducible, aperiodic DTMC with countable state space S. Then for all $i, j \in S$, $p_{i,j}^{(n)} = \mathbb{P}(X_n = j | X_0 = i) \to \mu_j^{-1}$ as $n \to \infty$.

A DTMC is ergodic if the limiting distribution exists and does not depend on the starting distribution, and so is equal to the stationary distribution. An irreducible DTMC is ergodic iff it is aperiodic and positive recurrent.

An $m \times m$ stochastic matrix P is called doubly-stochastic if each column sums to 1.

A finite-state DTMC has a doubly-stochastic transition matrix iff the uniform distribution $\boldsymbol{\pi} = (m^{-1}, \dots, m^{-1})$ is a stationary distribution.

An irreducible DTMC is called reversible if there exists a probability distribution $\boldsymbol{\pi} = (\pi_i)_{i\in S}$ such that $\pi_i p_{i,j} = \pi_j p_{j,i}$ for each $i, j \in S$, i.e., the amount that flows in equals the amount that flows out. These equations are called the detailed balance equations.

Any solution to these equations is a stationary distribution since if $\boldsymbol{\pi}$ satisfies the detailed balance equations then $\sum_{j\in S} \pi_j p_{j,i} = \sum_{j\in S} \pi_i p_{i,j} = \pi_i \left(\sum_{j\in S} p_{i,j}\right) = \pi_i$.

Kolmogorov's reversibility criterion. An irreducible DTMC with state space S is reversible iff it has a stationary distribution and P satisifies

$$p_{j_n j_1} \prod_{i=1}^{n-1} p_{j_i j_{i+1}} = p_{j_1 j_n} \prod_{i=1}^{n-1} p_{j_{i+1} j_i}$$

for **every** n and **every** $\{j_1, j_2, \dots, j_n\}$.

Let P be an irreducible (so that it has a stationary measure) stochastic matrix with $p_{i,j} = 0$ unless $j \in \{i-1, i, i+1\}$. Such chain is called a birth-and-death chain.

For any sequence of n transitions taking us from state i back to itself: if $n = 2$, then $p_{j_2 j_1} p_{j_1 j_2} = p_{j_1 j_2} p_{j_2 j_1}$; else if $n > 2$, we have $p_{j_n j_1} = p_{j_1 j_n} = 0$. Therefore, the Kolmogorov criterion is always satisfied for such a stochastic matrix, provided it has a stationary distribution.

Let $(X_n)_{n \in \mathbb{Z}_+}$ be a DTMC with $S = \mathbb{Z}_+$ and transition probabilities $p_{i,i+1} = p_i \in (0,1)$ for each $i \in S$, and $p_{0,0} = 1 - p_{0,1}$ and $p_{i,i-1} = 1 - p_i$ for $i \geq 1$.

This is a birth-and-death chain, so if it has a stationary distribution then it satisfies the detailed balance equations:

$$\pi_i p_i = \pi_{i+1}(1 - p_{1+1}), \quad \text{i.e.} \quad \pi_{i+1} = \big(p_i/(1 - p_{1+1})\big)\pi_i = \rho_i \pi_i \,.$$

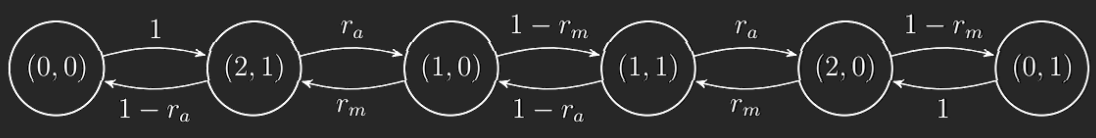
It follows that $\pi_i = x \prod_{j=0}^{i-1} \rho_j$ gives a solution to the detailed balance equations. Thus if $\sum_{i=0}^{\infty} \big(\prod_{j=0}^{i-1} \rho_j\big) < \infty$, i.e., $p < 1/2$, then there is a stationary distribution (otherwise not).

The chain is irreducible and aperiodic (due to the boundary), so also ergodic, and the limiting distribution is equal to the stationary distribution:

$$1 = \sum_{i=0}^{\infty} \pi_i = \sum_{i=0}^{\infty} \pi_0 \rho^i = \pi_0/(1-p) \;\; \Rightarrow \;\; \pi_0 = 1 - \rho.$$

Each morning at 8am, it is raining with probability $r_m \in (0,1)$, and each afternoon at 4pm it is raining with probability $r_a \in (0,1)$ (both are independent of all previous weather conditions). Suppose that your MAST30001 lecturer has 2 umbrellas, and that he departs home for work at 8am each day and departs from work to home at 4pm each day (with a very short commute). Whenever it is raining at departure time s/he takes an umbrella on the trip if there was one available at the departure point. Find the long run proportion of trips for which it is raining on his departure but he has no umbrella available.

Let N_n denote the number of umbrellas at the lecturer's current location, and L_n denote the current location. Then (N_n, L_n) is a discrete-time (time homogeneous) Markov chain with transition diagram:



A nonnegative integer-valued process $(N_t)_{t\geq 0}$ is a Poisson process with a rate λ if

1. It has independent increments on disjoint intervals: for $k \geq 2$ and $0 \leq s_1 < t_1 \leq s_2 < \cdots < t_k$, $N_{t_1} - N_{s_1}, \ldots, N_{t_k} - N_{s_k}$ are independent variables; and
2. For each $t > s \geq 0$, $N_t - N_s \sim \mathrm{Pn}\big(\lambda(t-s)\big)$.

Let $T_0 = 0$ and $T_j = \min\{t : N_t = j\}$, and define the time difference $\tau_j = T_j - T_{j-1}$.

$(N_t)_{t\geq 0}$ is a Poisson process with rate λ iff $(\tau_j)_{j\in\mathbb{N}} \sim \exp(\lambda)$ and are independent.

The key to the proof is to observe that the event $\{T_j \leq t\}$ is the same as $\{N_t \geq j\}$. That is the waiting time until the jth jump is less than or equal to t iff there are j or more jumps up to (and including) time t.

Assume that $(N_t)_{t\geq 0}$ is a Poisson process. Then $\mathbb{P}(T_1 \leq t) = \mathbb{P}(N_t \geq 1) = 1 - \mathbb{P}(N_t = 0) = 1 - e^{-\lambda t}$, so $T_1 \sim \exp(\lambda)$.

Furthermore, we have

$$F_{T_j}(t) = \mathbb{P}\big(T_j \leq t\big) = \mathbb{P}(N_t \geq j) = \sum_{k=j}^{\infty} e^{-\lambda t}(\lambda t)^k/k! = F_{\gamma(j,\lambda)}(t).$$

So the waiting time until the jth event is the sum of j independent exponentially-distributed inter-event times with parameter λ.

The argument also holds in reverse.

Assuming that $\tau_1 \sim \exp(\lambda)$, we know that $\mathbb{P}(T_1 \leq t) = 1 - e^{-\lambda t}$, which tells us that $\mathbb{P}(N_t = 0) = e^{-\lambda t}$. Alternatively, for $j > 1$, if $\{\tau_1, \ldots, \tau_j\}$ are i.i.d., then $T_j \sim \gamma(j, \lambda)$. So $\mathbb{P}(N_t \geq j) = \mathbb{P}\big(T_j \leq t\big) = 1 - \sum_{k=0}^{j-1} e^{-\lambda t}(\lambda t)^k/k!$, which tells us that $N_t \sim \mathrm{Pn}(\lambda t)$.

Also, following from the memoryless property of the exponential distribution and the independence of the τ_j's, $N_{t_i} - N_{s_i} \sim \mathrm{Pn}\big(\lambda(t_i - s_i)\big)$ and are independent over sets $[s_i, t_i)$ of disjoint intervals.

If $Y_1, \ldots, Y_k$ are i.i.d. with distribution function F, then the distribution function of $Y_{(i)}$ is

$$F_{Y_{(i)}}(x) = \sum_{l=i}^{k} \binom{k}{l} F(x)^l \big(1 - F(x)\big)^{k-l}.$$

If they are also absolutely continuous, with density f, then the density of the order statistics $Y_{(i)}$ is

$$\begin{aligned} f_{Y_{(i)}}(x) &= \binom{k}{i-1}\binom{k-i+1}{1}\binom{k-i}{k-i} F(x)^{i-1} f(x) \big(1 - F(x)\big)^{k-i} \\ &= \binom{k}{i} i\, F(x)^{i-1} f(x) \big(1 - F(x)\big)^{k-i}. \end{aligned}$$

Similarly, the joint densities for $1 \le r \le k$ and $x_1 < \cdots < x_r$ are

$$f_{Y_{(i_1)},\ldots,Y_{(i_r)}}(x_1, \ldots, x_r) = \binom{k}{i_1 - 1, 1, i_2 - i_1 - 1, 1, \cdots, 1, k - i_r} \times \prod_{j=1}^{r} f(x_j) \times \prod_{j=0}^{r} \left(F(x_{j+1}) - F(x_j)\right)^{i_{j+1} - i_j - 1},$$

where $\binom{l}{a_1, \ldots, a_j}$ is the number of ways to choose subsets of sizes $a_1, \ldots, a_j$ from a set of size l, and for the sake of brevity, we set $x_0 = -\infty$ and $x_{r+1} = \infty$ so $F(x_0) = 0$ and $F(x_{r+1}) = 1$.

In particular for $r = k,\ x_1 < \cdots < x_r$,

$$f_{Y_{(1)},\ldots,Y_{(k)}}(x_1, \ldots, x_k) = k! \prod_{j=1}^{k} f(x_j).$$

The conditional distribution of $(T_1, \ldots, T_k)$ given that $N_t = k$ is the same as the distribution of order statistics of a sample of k independent and identically-distributed random variables uniformly distributed on $[0, t]$. That is,

$$(T_1, \ldots, T_k) \mid \{N_t = k\} \sim \left(U_{(1)}, \ldots, U_{(k)}\right)$$

where $U_1, \ldots, U_k \sim U(0, t)$ and are independent.

According to our derivation for order statistics, $\left(U_{(1)}, \ldots, U_{(k)}\right)$ has density $k!\, t^{-k}$ for $0 = x_0 < x_1 < \cdots < x_k < t$. So we show the LHS has the same density:

$$\begin{aligned} &\mathbb{P}(T_1 \in dx_1, \ldots, T_k \in dx_k \mid N_t = k) \\ &= \frac{\mathbb{P}(\tau_1 \in dx_1, \tau_2 \in d(x_2 - x_1), \ldots, \tau_k \in d(x_k - x_{k-1}), \tau_{k+1} > t - x_k)}{\mathbb{P}(N_t = k)} \\ &= \frac{\left(\prod_{i=1}^{k} \lambda e^{-\lambda(x_i - x_{i-1})}\right) e^{-\lambda(t - x_k)}}{(\lambda t)^k e^{-\lambda t}/k!}\, dx_1 \ldots dx_k = k!\, t^{-k} dx_1 \ldots dx_k\,. \end{aligned}$$

The theorem implies that if $\tau_1, \ldots, \tau_n$ are i.i.d. exponential variables, then

$$\left(\frac{\tau_1}{\sum_{j=1}^{n+1} \tau_j}, \frac{\tau_1 + \tau_2}{\sum_{j=1}^{n+1} \tau_j}, \ldots, \frac{\sum_{j=1}^{n+1} \tau_j}{\sum_{j=1}^{n+1} \tau_j}\right)$$

have the same distribution as $U(0, 1)$ order statistics.

Superposition Theorem. Let $(N_t)_{t \ge 0}$ and $(M_t)_{t \ge 0}$ be two independent Poisson processes with rates λ_1 and λ_2 respectively and $L_t = N_t + M_t$. Then $(L_t)_{t \ge 0}$ is a Poisson process with rate $\lambda_1 + \lambda_2$.

Thinning Theorem. Suppose in a Poisson process $(N_t)_{t\geq 0}$ each jump is marked independently with probability p. Let M_t count the number of marks that arrive on $[0,t]$. The process $(M_t)_{t\geq 0}$ and $(N_t - M_t)_{t\geq 0}$ are **independent** Poisson processes with rates λp and $\lambda(1-p)$ respectively.

$$\begin{aligned}\mathbb{P}(M_t = j, N_t - M_t = k) &= \mathbb{P}(M_t = j, N_t = k+j) = \mathbb{P}(M_t = j \mid N_t = k+j)\mathbb{P}(N_t = k+j)\\ &= \binom{k+j}{j} p^j (1-p)^k \frac{e^{-\lambda t}(\lambda t)^{k+j}}{(k+j)!} = \frac{e^{-p\lambda t}(p\lambda t)^j}{j!}\frac{e^{-(1-p)\lambda t}\big((1-p)\lambda t\big)^k}{k!}\\ &= \mathbb{P}(M_t = j)\mathbb{P}(N_t - M_t = k).\end{aligned}$$

Note that $\mathbb{V}[N_t + M_t] = \mathbb{V}[N_t - M_t + 2M_t] = \mathbb{V}[N_t - M_t] + 4\mathbb{V}[M_t] = \lambda(1-p) + 4\lambda$.

Suppose that $(N_t)_{t\geq 0}$ is a Poisson process and $(X_i)_{i\in\mathbb{N}}$ are independent and identically-distributed random variables, which are also independent of $(N_t)_{t\geq 0}$. For $t \geq 0$, define $Y_t = \sum_{j\leq N_t} X_j$. Then $(Y_t)_{t\geq 0}$ is called a compound Poisson process.

Can find mean and variance of Y_t using decomposition laws as $\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X|Y]\big]$, and $\mathbb{V}[X] = \mathbb{V}\big[\mathbb{E}[X|Y]\big] + \mathbb{E}\big[\mathbb{V}[X|Y]\big]$.

A stochastic process $(X_t)_{t\geq 0}$ in continuous time, taking values in a countable state space $S \in \mathbb{R}$ is said to be a CTMC if, for all $k \geq 1$, $0 \leq t_1 < t_2 < \cdots < t_{k+1}$ and $i_1, i_2, \dots, i_{k+1} \in S$,

$$\mathbb{P}\left(X_{t_{k+1}} = i_{k+1} \middle| X_{t_1} = i_1, \dots, X_{t_k} = i_k\right) = \mathbb{P}\left(X_{t_{k+1}} = i_{k+1} \middle| X_{t_k} = i_k\right),$$

whenever the LHS is well-defined.

If $\mathbb{P}(X_{s+t} = k | X_s = j)$ for $j, k \in S$ do not depend on s then we say that the CTMC is time-homogeneous, and we can write $p_{j,k}^{(t)}$.

For DTMC, if $X_n = i$ then we waited for a Geometric$(1 - p_{i,i})$ amount of time before jumping to a new state. At the time we jump to a new state, the probability of jumping to j is $b_{i,j} = p_{i,j}/(1 - p_{i,i})$.

For CTMC, if $X_t = i$, we wait an $\exp(\lambda_i)$ time and then jump to a new state. The probability of jumping to j is $b_{i,j}$. Set the rate of transition $q_{i,j} = \lambda_i b_{i,j}$, then we can draw a transition diagram for it.

The jump chain of a CTMC $(X_t)_{t\geq 0}$ is the DTMC $\left(X_n^J\right)_{n\in\mathbb{Z}_+}$ defined by $X_n^J = X_{T_n}$ where $T_0 = 0$, and $T_i = \inf\{t > T_{i-1} : X_t \neq X_{T_{i-1}}\}$ are the jump chains of the CTMC.

DTMC properties:

1. Whether a state i is absorbing, i.e., $\lambda_i = 0$ and $b_{i,i} = 1$.
2. Whether $i \to j$, i.e., $p_{i,j}^{(t)} > 0$ for some t.
3. Communicating classes, i.e., $i \leftrightarrow j$.
4. Irreducibility, i.e., $i \leftrightarrow j$ for every $i, j \in S$.
5. Hitting probabilities, i.e., $h_{i,A} = \mathbb{P}(X_t \in A \text{ for some } t \geq 0 | X_0 = i)$.
6. Recurrence ($h_{i,j} = 1$ for every $i, j \in S$ for irreducible chains) and transience.
7. Expected hitting times. For a CTMC with state space S, and $A \subset S$, the vector of mean hitting times $\left(m_{i,A}\right)_{i\in S}$ is the minimal nonnegative solution to:

$$m_{i,A} = \begin{cases} 0, & \text{if } i \in A, \\ \lambda_i^{-1} + \sum_{j\in S} b_{i,j} m_{j,A}, & \text{if } i \notin A. \end{cases}$$

8. Positive recurrence, i.e., $m_{i,j} < \infty$ for every $i, j \in S$.
9. The long-run behaviour of the chain.

Items (2)-(6) above only depend on $\left(b_{i,j}\right)_{i,j\in S}$ and hence the CTMC has the given property iff its jump chain has that property.

On the other hand, items (7)-(9) also depend on how long we wait at every state.

An explosive CTMC is a process that jumps infinitely many times in a finite amount of time. For example, if $S = \mathbb{Z}_+$ and $\lambda_i = \lambda^i$ for some $\lambda > 1$ and $b_{i,i+1} = 1$ for each i then the CTMC is explosive. We henceforth assume that our CTMC is non-explosive.

A distribution $\pi = (\pi_i)_{i\in S}$ is called a stationary distribution for the family $\left(P^{(t)}\right)_{t\geq 0}$ if $\pi P^{(t)} = \pi$ for each $t \geq 0$.

Let $q_{i,i} \coloneqq -\sum_{j\neq i} q_{i,j} = -\lambda_i$ and let Q denote the rate matrix (called the infinitesimal generator) whose i,j'th entry is $q_{i,j}$. For non-explosive CTMCs, a distribution $\pi = (\pi_i)_{i\in S}$ is a stationary distribution for Q iff $\pi Q = 0$. This is equivalent to $\pi_i \lambda_i = \sum_{j\neq i} \pi_j q_{j,i}$ for $i \in S$, which are referred to as the full balance equations.

It's possible for explosive CTMC to have a solution to $\pi Q = 0$, with $\sum_j \pi_j = 1$ that is not the stationary distribution.

An irreducible and positive recurrent CTMC is ergodic, i.e., the limiting distribution exists and does not depend on the initial distribution, and the limiting proportion of time spent in state i is the stationary probability π_i.

For $i \in S$, let $T_1^{(i)} = \inf\{\, t > 0 : X_t \neq i\}$ and $T^{i,i} = \inf\left\{\, t > T_1^{(1)} : X_t = i\right\}$. Then

$$\pi_i = \mathbb{E}\left[\, T_1^{(i)} \,\middle|\, X_0 = i\right] / \mathbb{E}\left[\, T^{i,i} \,\middle|\, X_0 = i\right].$$

Periodicity is not an issue for a CTMC.

An irreducible CTMC is called reversible if there exists a probability distribution $\boldsymbol{\pi} = (\pi_i)_{i\in S}$ such that $\pi_i q_{i,j} = \pi_j q_{j,i}$ for each $i, j \in S$. Then we say that Q is reversible, and these equations are called the detailed balance equations.

Let $(X_t)_{t\geq 0}$ be a birth-and-death process with birth rates $(v_n)_{n\in S}$, death rates $(\mu_n)_{n\in S}$.

The generator has the form:

$$Q = \begin{bmatrix} -v_0 & v_0 & 0 & 0 & \cdots \\ \mu_1 & -(\mu_1 + v_1) & v_1 & 0 & \cdots \\ 0 & \mu_2 & -(\mu_2 + v_2) & v_2 & \cdots \\ 0 & 0 & \mu_3 & -(\mu_3 + v_3) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

To derive the stationary distribution (if it exists), we solve:

$$v_k \pi_k = \mu_{k+1} \pi_{k+1} \quad \text{where } k \in \mathbb{Z}_+ \quad \Rightarrow \quad \pi_k = \pi_0 \prod_{l=1}^{k} (v_{l-1}/\mu_l).$$

So a stationary distribution exists iff $\sum_{k=0}^{\infty} \prod_{l=1}^{k} (v_{l-1}/\mu_l) < \infty$, in which case

$$\pi_0 = \left(\sum_{k=0}^{\infty} \prod_{l=1}^{k} (v_{l-1}/\mu_l)\right)^{-1}.$$

Let $dP^{(t)}/dt = P^{(t)}Q$ and $dP^{(t)}/dt = QP^{(t)}$ be the forward and backward equations.

For (non-explosive) CTMCs, the matrix Q determines the transition probability completely by solving the backward or forward equations to get

$$P^{(t)} = \exp(tQ) \coloneqq \sum_{k=0}^{\infty} \frac{1}{k!} t^k Q^k,$$

subject to $P^{(0)} = I$.

The Poisson process is a CTMC with generator

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & \cdots \\ 0 & 0 & 0 & -\lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

One can show that $(Q^n)_{i,j} = 0$ if $j \notin \{i, i+1, \dots, i+n\}$, and otherwise

$$(Q^n)_{i,j} = \lambda^n \binom{n}{j-i} (-1)^{n-(j-i)}.$$

Then for $j \geq i$,

$$p_{i,j}^{(t)} = \sum_{n=0}^{\infty} \frac{1}{n!} t^n (Q^n)_{i,j} = \frac{(t\lambda)^{j-i}}{(j-i)!} e^{-t\lambda} \;\Rightarrow\; p_{0,k}^{(t)} = e^{-t\lambda}(t\lambda)^k/k!.$$

Alternatively, we have

$$\frac{d}{dt} p_{0,0}^{(t)} = \left[P^{(t)}Q\right]_{0,0} = \sum_{i=0}^{\infty} P_{0,i}^{(t)} Q_{i,0} = -\lambda p_{0,0}^{(t)} \;\Rightarrow\; p_{0,0}^{(t)} = ce^{-\lambda t},$$

so with the condition that $p_{0,0}^{(0)} = 1$ we get $p_{0,0}^{(t)} = e^{-\lambda t}$. Similarly,

$$\frac{d}{dt} p_{0,k}^{(t)} = \lambda\left(p_{0,k-1}^{(t)} - p_{0,k}^{(t)}\right).$$

Then by induction, we get exactly the same result as above.

A Markov chain $(X_t)_{t\geq 0}$ with state space $S = \{1, 2, 3\}$ has generator

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 2 & -4 & 2 \\ 0 & 1 & -1 \end{bmatrix}.$$

Using symmetry and that $\sum_i p_{2,i}(t) = 1$, we have $p_{2,1}(t) = p_{2,3}(t) = \left(1 - p_{2,2}(t)\right)/2$, so the above equation reduces to

$$p'_{2,2}(t) = p_{2,1}(t) - 4p_{2,2}(t) + p_{2,3}(t) = \left(1 - p_{2,2}(t)\right) - 4p_{2,2}(t) = 1 - 5p_{2,2}(t).$$

We can easily solve this with the boundary condition $p_{2,2}(0) = 1$, to find

$$p_{2,2}(t) = \frac{1}{5} + \frac{4}{5}e^{-5t}.$$

Also, we can calculate $p_{1,2}(t)$ from this result. In order to be at state 2 at time t, starting from state 1, we must jump away from 1 before time t. But then the chain always jumps to state 2, so we condition on the time of the jump and use the Markov property to find:

$$p_{1,2}(t) = \int_0^t e^{-s} p_{2,2}(t-s)ds = \int_0^t e^{-s}\left(\frac{1}{5} + \frac{4}{5}e^{-5(t-s)}\right)ds = \frac{1}{5}(1 - e^{-5t}).$$

Queueing theory is the mathematical study of the operation of stochastic systems describing processing of flows of jobs. Queues occur when current demand for service exceeds the capacity of the service facility.

- There's a total of m spaces for both receiving service and waiting for it. If there's an idle server, service commences for an arriving customer immediately.
- The service time $S_i^{(j)}$ of the i'th customer at the j'th server is a random variable, which is i.i.d. for each fixed j.
- Arrival times $T_1, T_2, T_3, \ldots$ The (i.i.d.) inter-arrival times are $\tau_1 = T_1 - T_0, \ldots$
- Counting process N_t gives the number of arrivals in $[0, t]$, $t \geq 0$.
- X_t is the number of customers in the system (including those being serverd) at time t.
- If all servers are busy, then the arriving customers join a queue if there is enough space, otherwise, the customer is rejected.
- We only consider FIFO in this course.

Kendall's notation takes the form $A/B/m/n$ where

- A describes the arrival process. $A = M$ means (Markovian) inter-arrival times are exponentially-distributed.
- B describes the service process.
- n gives the number of servers.
- m gives the capacity of the system. When $m = \infty$, this is usually omitted.

$M/M/1$ queues have arrival stream as Poisson process with intensity λ, and $n = 1$ server that service time $S \sim \exp(\mu)$. This is a birth-and-death process with nonzero transition rates $q_{i,i+1} = \lambda$ and $q_{i+1,i} = \mu$ for all $i \in \mathbb{Z}_{0+}$.

Using our results from CTMCs, we see that a stationary distribution for $(X_t)_{t\geq 0}$ exists iff the chain is positive recurrent. This is equivalent to the traffic intensity $\rho \coloneqq \lambda/\mu < 1$, in which case,

$$\pi_n = (1-\rho)\rho^n \quad \text{for } n \in \mathbb{Z}_{0+}.$$

So the stationary distribution for **the number of customers in the system** is Geometric$(1-\rho)$.

In the stationary regime, a tagged arriving customer will find a random number N of customers ahead where $N \sim (\pi_k)_{k\in\mathbb{Z}_{0+}}$. (PASTA principal)

If $N = 0$, then the customer will go straight into service. Otherwise, the remaining (independent) service time S_i for the customer being served $\sim \exp(\mu)$. So the waiting time for our tagged customer is $W = \sum_{j=1}^{N} S_j$.

The distribution of a nonnegative random variable Y is characterized by its Laplace transform $M_Y(-s) = \mathbb{E}[e^{-sY}]$ for $s > 0$. We can write

$$\begin{aligned}\mathbb{E}_\pi[e^{-sW}] &= \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[e^{-s\sum_{j=1}^N S_j}\middle|N\right]\right] = \mathbb{E}_\pi[(\mathbb{E}_\pi[e^{-sS_1}])^N] = \mathbb{E}_\pi\left[\left(\int_{x\geq 0} e^{-sx} f_X(x)dx\right)^N\right] \\ &= \mathbb{E}_\pi\left[\left(\mu\int_{x\geq 0} e^{-(s+\mu)x}dx\right)^N\right] = \mathbb{E}_\pi\left[\left(\mu\left[-\frac{e^{-(s+\mu)x}}{s+u}\right]_{x=0}^{\infty}\right)^N\right] \\ &= \mathbb{E}_\pi\left[\left(\frac{u}{s+u}\right)^N\right] = \mathbb{E}_\pi\left[e^{N\log(u/(s+u))}\right] = M_N\big(\log(u/(s+u))\big) \\ &= \frac{1-\rho}{1-\rho u/(s+u)} = (1-\rho) + \rho\frac{\mu-\lambda}{s+\mu-\lambda},\end{aligned}$$

and we see that the distribution of W is a mixture of a 0-random variable and an $\exp(\mu-\lambda)$ random variable. To be precise,

$$\mathbb{P}(W=0) = 1-\rho, \qquad \mathbb{P}(W>x) = \rho e^{-x(\mu-\lambda)} \quad \text{for } x>0.$$

It follows that **the expected waiting time** is $\mathbb{E}[W] = \rho/(\mu-\lambda)$.

Once we have the expected waiting time, we can calculate **the expected total time** d in the system via the formula:

$$d = \mathbb{E}[W] + \mu^{-1} = \frac{1}{\mu-\lambda}.$$

Recall that l Is the expected number of customers in the system, while l_q is the expected number of customers waiting for service (both at stationarity).

Little's Law says that $l = \lambda d,\ l_q = \lambda\mathbb{E}[W]$. Note that $l_q \neq l-1$ since the system might be empty.

$M/M/a$ queues have $a \geq 1$ servers.

The transition rates are $q_{i,i+1} = \lambda$ for $i \geq 0$, and $q_{i,i-1} = \mu \times \min\{a, i\}$ for $i \geq 1$.

Let $\kappa_j = v_0 \dots v_{j-1}/(\mu_1 \dots \mu_j) = \begin{cases}(\lambda/\mu)^j/j! & \text{if } j<a,\\ (\lambda/\mu)^j/(a!\,a^{j-a}) & \text{if } j\geq a.\end{cases}$ Then, $\sum_{j=0}^\infty \kappa_j < \infty$ occurs if $\lambda < a\mu$ (and so ergodic), where

$$\sum_{j=0}^\infty \kappa_j = \sum_{k=0}^{a-1}\frac{\lambda^k}{k!\,\mu^k} + \frac{\lambda^a}{a!\,\mu^a}\frac{a\mu}{a\mu-\lambda}.$$

In this case, the stationary distribution is given by

$$\pi_k = \begin{cases}\pi_0(\lambda/\mu)^k/k! & \text{if } k<a\\ \pi_0(\lambda/\mu)^k/(a!\,a^{k-a}) & \text{if } k\geq a,\end{cases} \quad \text{where } \pi_0 = \left(\sum_{j=0}^\infty \kappa_j\right)^{-1}.$$

The proportion of time δ_q that all the servers are busy is the same as the probability that an arriving customer will have to wait.

We have

$$\delta_q = \sum_{k=a}^{\infty} \pi_k = \pi_0 \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}.$$

The expected queue length is

$$\begin{aligned} l_q = \mathbb{E}[\max\{X_t - a, 0\}] &= \sum_{n=a}^{\infty} (n-a) \frac{\pi_0 \left(\frac{\lambda}{\mu}\right)^n}{a!\,a^{n-a}} = \sum_{m=0}^{\infty} m \frac{\pi_0 \left(\frac{\lambda}{\mu}\right)^{m+a}}{a!\,a^m} \\ &= \pi_0 \frac{\lambda^a}{a!\,\mu^a} \sum_{m=0}^{\infty} m \left(\frac{\lambda}{a\mu}\right)^m = \pi_0 \frac{\lambda^a}{a!\,\mu^a} \frac{\frac{\lambda}{a\mu}}{\left(1 - \frac{\lambda}{a\mu}\right)^2} = \pi_0 \frac{\lambda^a}{a!\,\mu^a} \frac{\lambda a\mu}{(a\mu - \lambda)^2} \\ &= \frac{\lambda}{a\mu - \lambda} \delta_q. \end{aligned}$$

In stationarity, the expected number of busy servers is

$$\begin{aligned} b_s = \mathbb{E}[\min\{X_t, a\}] &= \sum_{k=0}^{a-1} k\pi_0 \left(\frac{\lambda}{\mu}\right)^k / k! + a\delta_q = \frac{\sum_{k=1}^{a-1} \frac{\lambda^k}{(k-1)!\,\mu^k} + a \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}}{\sum_{k=0}^{a-1} \frac{\lambda^k}{k!\,\mu^k} + \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}} \\ &= \frac{\sum_{k=0}^{a-2} \frac{\lambda^{k+1}}{k!\,\mu^{k+1}} + a \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}}{\sum_{k=0}^{a-1} \frac{\lambda^k}{k!\,\mu^k} + \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}} \\ &= \frac{\sum_{k=0}^{a-1} \frac{\lambda^{k+1}}{k!\,\mu^{k+1}} - \frac{\lambda^a}{(a-1)!\,\mu^a} + a \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}}{\sum_{k=0}^{a-1} \frac{\lambda^k}{k!\,\mu^k} + \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}} \\ &= \frac{\sum_{k=0}^{a-1} \frac{\lambda^{k+1}}{k!\,\mu^{k+1}} + \frac{\lambda^a}{a!\,\mu^a} \left(\frac{a^2\mu}{a\mu - \lambda} - a\right)}{\sum_{k=0}^{a-1} \frac{\lambda^k}{k!\,\mu^k} + \frac{\lambda^a}{a!\,\mu^a} \frac{a\mu}{a\mu - \lambda}} = \lambda/\mu. \end{aligned}$$

Note that provided that $\lambda < a\mu$, this does not depend on a. Then the expected number of customers in the system is

$$l = b_s + l_q = \frac{\lambda}{\mu} + \frac{\lambda}{a\mu - \lambda} \delta_q.$$

Using Little's Law, we can compute the expected waiting time and delay to be $\mathbb{E}[W] = l_q/\lambda = \delta_q/(a\mu - \lambda)$, and $d = \mu^{-1} + \delta_q/(a\mu - \lambda)$.

At a small airport, arriving passengers arrive as a Poisson process of rate 1 per minute and each such passenger is assigned uniformly at random (independent of the length of the queues) to one of k servers (that each serve the customers in their queue in the order in which they arrived). Service times at each queue are Exponential random variables with mean 2.5 minutes (independent of everything else). Let $N_t^{(i)}$ denote the number of customers in queue $i = 1, \dots, k$ at time t, and $N_t = \sum_{i=1}^{k} N_t^{(i)}$.

Hint. This system is not $M/M/a$, consider it as combination of multiple $M/M/1$'s.

In a certain computer queuing system, jobs arrive according to a Poisson process with rate 4. There are two servers that process jobs: Server A works at exponential rate 3 and Server B at exponential rate 2. Since Server A is faster than Server B, the system works as follows. When an arriving job finds the system empty, Server A always processes the job. If only one server is free, then an arriving job goes immediately into service with the free server. When both servers are busy, jobs queue in an infinite buffer.

Model the number of jobs in the system (including those being worked on) as a continuous time Markov chain $(X_t)_{t\geq 0}$ with appropriate state space, and specify its generator. Find the stationary distribution of the Markov chain.

We view the system as a CTMC with states $\{0, (1,0), (0,1), 2, 3, \dots\}$, where $(1,0)$ means the Server A is busy and Server B is not, and $(0,1)$ means Server B is busy and Server A is not; note in both these cases the number of jobs in the system is 1. The generator is

$$\begin{bmatrix} -4 & 4 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 3 & -7 & 0 & 4 & 0 & 0 & 0 & \cdots \\ 2 & 0 & -6 & 4 & 0 & 0 & 0 & \cdots \\ 0 & 2 & 3 & -9 & 4 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 5 & -9 & 4 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 5 & -9 & 4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where it continues for $j \geq 3$ as a birth-death process. (The chain is irreducible and ergodic since the maximum service rate is greater than the arrival rate.)

To determine the stationary distribution of this system, we solve $\pi A = 0$. For $j \geq 3$, the equations are $9\pi_j = 4\pi_{j-1} + 5\pi_{j+1}$, which can be solved in the usual way, and so

$$\pi_j = a + b(4/5)^j.$$

The a coefficient must be zero. Now substitute for the states 0, (1, 0), (0, 1), and get

$$\pi_0 = 13/113, \quad \pi_{(1,0)} = 12/113, \quad \pi_{(0,1)} = 8/113,$$

$$\pi_j = \frac{16}{113}\left(\frac{4}{5}\right)^{j-2} \quad \text{for } j \geq 2.$$

A renewal process $\{N_t : t \geq 0\}$ is a counting process for which the times $\tau_j \geq 0$ between successive events, called renewals, are i.i.d. nonnegative-real-valued random variables with an arbitrary distribution function F. A renewal process that is not a Poisson process is not Markovian.

Counting process description vs. waiting time description:

- $\{N_t \geq n\} = \{T_n \leq t\}$
- $\{N_t < n\} = \{T_n > t\}$
- $\{N_t = n\} = \{T_n \leq t < T_{n+1}\}$
- $T_{N_t} \leq t < T_{N_t+1}$.

For any fixed $t < \infty$, $\mathbb{P}(N_t = \infty) = 0$, i.e., there cannot be an explosion.

To see this, write

$$\mathbb{P}(N_t = \infty) = \lim_{n\to\infty} \mathbb{P}(N_t \geq n) = \lim_{n\to\infty} \mathbb{P}(T_n \leq t) = \lim_{n\to\infty} \mathbb{P}\left(\sum_{i=1}^{n} \tau_i \leq t\right)$$
$$= \lim_{n\to\infty} \mathbb{P}\left(e^{-\sum_{i=1}^{n}\tau_i} \geq e^{-t}\right).$$

Using Markov's inequality:

$$\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a \quad \text{for } X \geq 0 \text{ and } a > 0,$$

we have $\mathbb{P}\left(e^{-\sum_{i=1}^{n}\tau_i} \geq e^{-t}\right) \leq e^{t}\mathbb{E}\left[e^{-\sum_{i=1}^{n}\tau_i}\right] = e^{t}(\mathbb{E}[e^{-\tau_1}])^n \to 0$ as $n \to \infty$, since $[e^{-\tau_1}] < 1$.

If $\mathbb{E}[\tau_j] = \mu, \mathbb{V}[\tau_j] = \sigma^2 < \infty$, then

$$\mathbb{P}(N_t \geq n) = \mathbb{P}(T_n \leq t) = \mathbb{P}\left(\sum_{i=1}^{n} \tau_i \leq t\right) \approx \mathbb{P}\left(Z \leq \frac{t - n\mu}{\sqrt{n\sigma^2}}\right) = \mathbb{P}\left(Z \geq \frac{n\mu - t}{\sqrt{n\sigma^2}}\right).$$

Now, we choose $n = \mathrm{n}(x, t)$ such that

$$\frac{n\mu - t}{\sqrt{n\sigma^2}} \approx x \;\Rightarrow\; n(x,t) \approx \frac{t}{\mu} + x\sqrt{\frac{t\sigma^2}{\mu^3}}.$$

Thus as $t \to \infty$,

$$\mathbb{P}(Z \geq x) = \mathbb{P}(N_t \geq \mathrm{n}(x,t)) = \mathbb{P}\left(N_t \geq \frac{t}{\mu} + x\sqrt{\frac{t\sigma^2}{\mu^3}}\right) \;\Rightarrow\; \frac{N_t - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \sim \mathrm{N}(0,1).$$

The residual lifetime at time t is the amount of time until the next renewal time, i.e., $R_t = T_{N_t+1} - t > 0$. The age of the renewal process at time t is the amount of time since the most recent renewal, i.e., $A_t = t - T_{N_t}$.

If F is non-lattice (i.e., does not concentrate its mass at multiples of a fixed amount) with finite mean μ and $x \geq 0$, then

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{P}(R_t \leq x) &= 1 - \lim_{t\to\infty} \mathbb{P}(R_t > x) = 1 - \int_0^{T_n} \mathbf{1}_{\{R_s>x\}} ds \,/T_n \\
&= 1 - \frac{\sum_{i=1}^n (\tau_i - x)\mathbf{1}_{\{R_s>x\}}}{\sum_{i=1}^n \tau_i} = 1 - \frac{n^{-1}\sum_{i=1}^n (\tau_i - x)\mathbf{1}_{\{R_s>x\}}}{n^{-1}\sum_{i=1}^n \tau_i} \\
&= 1 - \frac{\mathbb{E}\left[(\tau_i - x)\mathbf{1}_{\{R_s>x\}}\right]}{\mathbb{E}[\tau_1]} \\
&= 1 - \frac{1}{\mu}\int_0^\infty \mathbb{P}\left((\tau_i - x)\mathbf{1}_{\{R_s>x\}} > y\right) dy \qquad \text{[Tail Probability]} \\
&= 1 - \frac{1}{\mu}\int_x^\infty \mathbb{P}(\tau_i > u) du = \frac{1}{\mu}\int_0^x \big(1 - F(y)\big)\, dy.
\end{aligned}$$

For $x, y \geq 0$, then events $\{R_t > x, A_t > y\}$ and $\{R_{t-y} > x + y\}$ are equal so

$$\lim_{t\to\infty} \mathbb{P}(R_t > x, A_t > y) = \lim_{t\to\infty} \mathbb{P}\big(R_{t-y} > x + y\big) = \frac{1}{\mu}\int_{x+y}^\infty \big(1 - F(z)\big) dz\,. \quad \text{\# pdf}$$

Setting $x = 0$ we get

$$\lim_{t\to\infty} \mathbb{P}(A_t \leq y) = \frac{1}{\mu}\int_0^y \big(1 - F(z)\big)\, dz = \lim_{t\to\infty} \mathbb{P}(R_t \leq y).$$

For large t, we can also find the joint pdf of (R_t, A_t):

$$\begin{aligned}
\mathbb{P}(R_t \leq x, A_t \leq y) &= \mathbb{P}(A_t \leq y) - \mathbb{P}(R_t \leq x) + \mathbb{P}(R_t > x, A_t > y) \\
\Rightarrow \quad \frac{\partial^2 \mathbb{P}(R_t \leq x, A_t \leq y)}{\partial x \partial y} &= \frac{\partial^2 \mathbb{P}(R_t > x, A_t > y)}{\partial x \partial y} \\
&= \frac{\partial^2}{\partial x \partial y}\left[\frac{1}{\mu}\int_{x+y}^\infty \big(1 - F(z)\big) dz\right].
\end{aligned}$$

A random vector $\vec{X} = (X_1, \dots, X_d)$ is multivariate normal iff every linear combination $\vec{x} \cdot \vec{X}$ is univariate normal. The distribution of a multivariate normal vector can be specified by the mean vector and the covariance matrix.

A Gaussian process $(X_t)_{t \in I}$ is a random process for which the finite-dimensional distributions are all multivariate normal, i.e., $(X_{t_1}, \dots, X_{t_r})$ is multivariate normal for every $r \in \mathbb{N}$, $t_1 < t_2 < \cdots < t_r$ all in I. Typically, I is $[0, \infty)$ or $[0, 1]$.

It follows that the distribution of continuous Gaussian process $(X_t)_{t \in I}$ is determined by two functions: the mean function $\mu(t) = \mathbb{E}[X_t]$ for $t \in I$, and the covariance function $\sum(s, t) = \mathrm{Cov}(X_s, X_t)$ for $s \leq t$ both in I.

(Standard) Brownian motion $(W_t)_{t \geq 0}$ is a continuous Gaussian process with $\mu(t) = 0$ and $\sum(s, t) = s$ for $s \leq t$. (Standard) Brownian bridge $(B_t)_{t \in [0,1]}$ is a continuous Gaussian process with $\mu(t) = 0$ and $\sum(s, t) = s(1 - t)$ for $s \leq t$.

Sketch construction of Brownian motion.

Let $(Z_q)_{q \in \mathbb{Q} \cap [0,1]}$ be i.i.d. standard normal random variables.

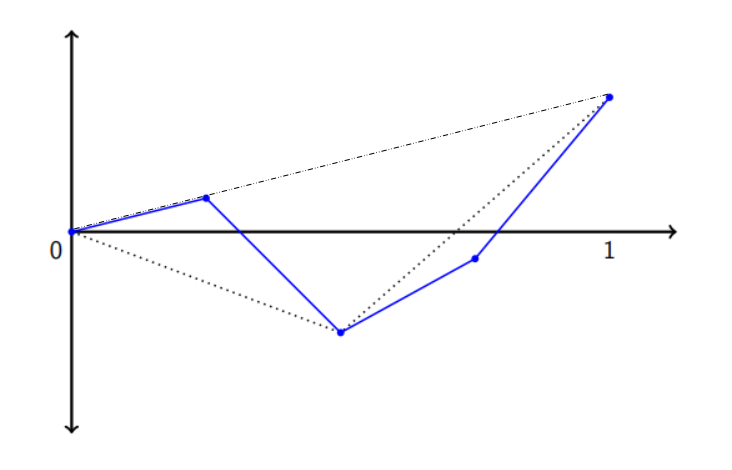
Define a sequence of random functions $\left(W_t^{(n)}\right)_{t \in [0,1]}$ for $n \in \mathbb{N}$ by $W_t^{(1)} = tZ_1$, i.e., set $W_0^{(1)} = 0$ and $W_1^{(1)} = Z_1$, and then linearly interpolate.

Set $W^{(2)}$ to be the same at 0 and 1 but set

$$W_{1/2}^{(2)} = W_{1/2}^{(1)} + \frac{1}{\sqrt{2^2}} Z_{1/2},$$

and then linearly interpolate in between.

More generally define $W^{(n+1)}$ to be equal to $W^{(n)}$ at points $2i/2^n$, and define $W^{(n+1)}$ at the points q of the form $(2i + 1)/2^n$ by adding some extra randomness $Z_q/\sqrt{2^n}$ to $W_q^{(n)}$ and then linearly interpolating.



This sequence of (random) continuous functions converges as $n \to \infty$ (uniformly) to a random continuous function $(W_t)_{t\in[0,1]}$. This random function has the correct mean and covariance functions since it does at every dyadic rational point.

BM has independent increments: if $0 \le s_1 < t_1 \le s_2 < t_2, \dots, \le s_n < t_n$, then $\left(W_{t_i} - W_{s_i}\right)_{i\le n}$ are independent random variables.

$$\begin{aligned}
&\mathrm{Cov}\left(W_{t_i} - W_{s_i}, W_{t_j} - W_{s_j}\right) \\
&\quad = \mathbb{E}\left[\left(W_{t_i} - W_{s_i}\right)\left(W_{t_j} - W_{s_j}\right)\right] - \left(\mathbb{E}[W_{t_i}] - \mathbb{E}[W_{s_i}]\right)\left(\mathbb{E}\left[W_{t_j}\right] - \mathbb{E}\left[W_{s_j}\right]\right) \\
&\quad = \mathbb{E}\left[W_{t_i}W_{t_j}\right] - \mathbb{E}\left[W_{s_i}W_{t_j}\right] - \mathbb{E}\left[W_{t_i}W_{s_j}\right] + \mathbb{E}\left[W_{s_i}W_{s_j}\right] \\
&\quad = \mathrm{Cov}\left(W_{t_i}, W_{t_j}\right) - \mathrm{Cov}\left(W_{s_i}, W_{t_j}\right) - \mathrm{Cov}\left(W_{t_i}, W_{s_j}\right) + \mathrm{Cov}\left(W_{s_i}, W_{s_j}\right) \\
&\quad = t_j - t_j - s_j + s_j = 0.
\end{aligned}$$

Definition of BM is equivalent to saying that $(W_t)_{t\ge0}$ is a continuous process with:

1. $W_0 = 0$ and,
2. with independent increments (if they're disjoint)
3. and $W_t - W_s \sim \mathrm{N}(0, t-s)$ for every $t > s \ge 0$.

BM is both a Markov process (with state space $\mathbb{R}$), and a Martingale. BM is recurrent. BM is not differentiable at any point.

Suppose that $(W_t)_{t\ge0}$ is a BM:

1. If $s > 0$ and $X_t = W_{t+s} - W_s$, then $(X_t)_{t\ge0}$ is a BM and independent of $(W_u)_{u\le s}$.
2. If $X_0 = 0$ and $X_t = tW_{1/t}$ for $t > 0$, then $(X_t)_{t\ge0}$ is a BM.
3. If $c > 0$ and $X_t = W_{ct}/\sqrt{c}$, then $(X_t)_{t\ge0}$ is a BM.
4. If $X_t = W_t - tW_1$, then $(X_t)_{t\in[0,1]}$ is a Brownian bridge. Brownian bridge does not have independent increments, e.g., $B_1 - B_{1/2} = -\left(B_{1/2} - B_0\right)$.

Functional Central Limit Theorem.

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. random variables with mean 0 and variance 1, and let $Z_t^{(n)} = \sum_{i=1}^{\lfloor nt\rfloor} X_i/\sqrt{n}$. Then $\left(Z_t^{(n)}\right)_{t\ge0} \sim (W_t)_{t\ge0}$.

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. with cdf F, and let $F^{(n)}(x) = \sum_{i=1}^{n} \mathbf{1}_{\{X_i\le x\}}/n$. Then $\left(\sqrt{n}\left(F^{(n)}(x) - F(x)\right)\right)_{x\in\mathbb{R}} \sim \left(B_{F(x)}\right)_{x\in\mathbb{R}}$. If $X_i \sim \mathrm{U}(0,1)$, then $B_{F(t)} = B_t$.

An insurance company receives 10 thousand dollars per day in payments. For $i = 1, 2, \ldots$, let X_i be the amount, in thousands of dollars, that the insurance company pays out in claims i days from now. Assume that the X_i are i.i.d. standard gamma with parameter 10, having density $x^9 e^{-x}/9!$ for $x > 0$.

The company currently has 100 thousand dollars in its bank account for paying out claims. Assuming payments and claims are made at the same time each day, use the approximation of random walk by Brownian motion to estimate the probability that the insurance company has a positive amount in its bank account every day for the next 1000 days. You may want to use the fact that $M_t := \min_{0 \le s \le t}\{B_s\} \approx -|B_t|$.

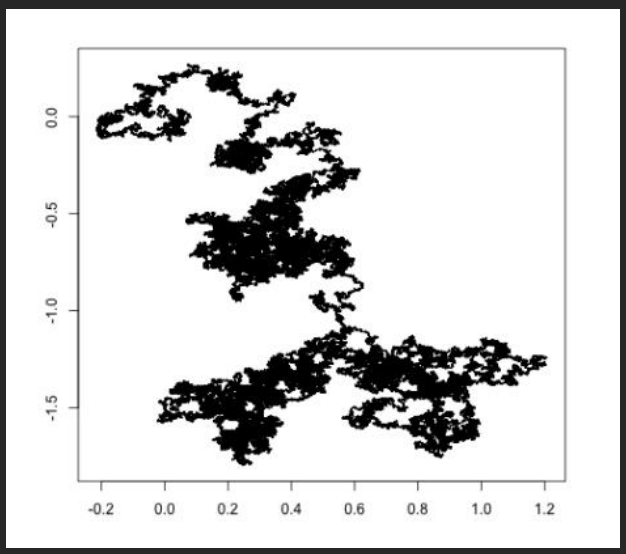
The amount of profit/loss each day is i.i.d. distributed $X_i - 10$, which has mean 0 and variance 10. Thus if $S_n = \sum_{i=1}^{n}(X_i - 10)$, then we define the process

$$B_t^{(n)} = \left(\frac{S_{\lfloor nt \rfloor}}{\sqrt{10}\sqrt{n}}\right)_{0 \le t \le 1} \approx W_t \sim \mathrm{N}(0, t)$$

is well approximately by a Brownian motion as $n \to \infty$. In particular,

$$\begin{aligned}\mathbb{P}(\min_{0 \le k \le 1000}\{S_k\} > -100) &= \mathbb{P}\left(\min_{0 \le t \le 1}\left\{\frac{S_{\lfloor 1000t \rfloor}}{\sqrt{10}\sqrt{1000}}\right\} > -1\right) \\ &= \mathbb{P}(\min_{0 \le t \le 1} B_t > -1) = \mathbb{P}(|B_1| < 1) = 2\Phi^{-1}(1) - 1 = 0.683.\end{aligned}$$

Let $\left(W_t^{[i]}\right)_{t \ge 0}$ be independent BM for $i \in \mathbb{N}$. Then $\left(\left(W_t^{[1]}, W_t^{[2]}\right)\right)_{t \ge 0}$ is a 2-dimensional BM, $\left(\left(W_t^{[1]}, W_t^{[2]}, W_t^{[3]}\right)\right)_{t \ge 0}$ is a 3-dimensional BM, etc.



(1-dimensional) BM visits every point in $\mathbb{R}$ infinitely often.

For 2-dimensional BM, for every $k \in \mathbb{Z}_+$ there are (random) points in $\mathbb{R}^2$ visited exactly k times. Every neighbourhood of every point is visited infinitely often.

For 3-dimensioanl BM, there are (random) points visited exactly twice, and no point in $\mathbb{R}^3$ is visited 3 or more times, $|B_t| \to \infty$ as $t \to \infty$.

For 4-dimensional BM, no point in $\mathbb{R}^4$ is hit more than once.

Useful Formulae:

1. $\Gamma(r) = \int_0^\infty e^{-x} x^{r-1} dx = (r-1)! \quad \forall\, r > 0$

2. $\mathrm{B}(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$

3. $\binom{-r}{z} = \binom{z+r-1}{r-1}(-1)^z, \quad z \in \mathbb{Z}^{0+}$

4. $\lim_{n\to\infty} \left(1 + \dfrac{x}{n}\right)^n = e^x$

5. $\sinh(x) = \dfrac{1}{2}(e^x - e^{-x}) = -i\sin(ix) \qquad \cosh(x) = \dfrac{1}{2}(e^x + e^{-x}) = \cos(ix)$

   $\operatorname{arcsinh}(x) = \log\left(x + \sqrt{x^2+1}\right), \quad x \in \mathbb{R}$

   $\operatorname{arccosh}(x) = \log\left(x + \sqrt{x^2-1}\right), \quad x \geq 1$

   $\operatorname{arctanh}(x) = \dfrac{1}{2}\log\left(\dfrac{1+x}{1-x}\right), \quad |x| < 1$

6. Taylor Polynomial: $f(x) = \lim_{n\to\infty} \sum_{k=0}^{n} \dfrac{f^{(k)}(a)(x-a)^k}{k!}$

   $\exp(z) = \sum_{k=0}^{\infty} \dfrac{z^k}{k!} \qquad \log(z+1) = \sum_{k=0}^{\infty} (-1)^{k+1} \dfrac{z^k}{k}$

   $\cos(z) = \sum_{k=0}^{\infty} \dfrac{(-1)^k z^{2k}}{(2k)!} \qquad \sin(z) = \sum_{k=0}^{\infty} \dfrac{(-1)^k z^{2k+1}}{(2k+1)!}$

   $\cosh(z) = \sum_{k=0}^{\infty} \dfrac{z^{2k}}{(2k)!} \qquad \sinh(z) = \sum_{k=0}^{\infty} \dfrac{z^{2k+1}}{(2k+1)!}$

7. Integration by Parts: $\int_a^b f'(t)g(t)dt = [f(t)g(t)]_a^b - \int_a^b f(t)g'(t)dt$

8. $\int \dfrac{1}{\sqrt{a^2-x^2}} dx = \arcsin\left(\dfrac{x}{a}\right) + C \qquad \int \dfrac{1}{\sqrt{x^2+a^2}} dx = \operatorname{arcsinh}\left(\dfrac{x}{a}\right) + C$

   $\int \dfrac{-1}{\sqrt{a^2-x^2}} dx = \arccos\left(\dfrac{x}{a}\right) + C \qquad \int \dfrac{1}{\sqrt{x^2-a^2}} dx = \operatorname{arccosh}\left(\dfrac{x}{a}\right) + C$

   $\int \dfrac{1}{a^2+x^2} dx = \dfrac{1}{a}\arctan\left(\dfrac{x}{a}\right) + C \qquad \int \dfrac{1}{a^2-x^2} dx = \dfrac{1}{a}\operatorname{arctanh}\left(\dfrac{x}{a}\right) + C$

9. Geometric Progression: $\sum_{k=1}^{n} ar^{k-1} = \frac{a(1-r^n)}{1-r}$

   Geometric Series: $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ if $|r| < 1$

10. $(AB)^{-1} = B^{-1}A^{-1}$

11. $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = (ad-bc)^{-1}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$