A Brief Summary of Statistics Course

统计学课程知识总结

Vincent

2021 年 4 月 21 日

目录

Chapter. I 概率论部分

Instructor: Wanlu Deng

Chapter Overview

- Basic axioms

Cover: Basic axioms, random events, σ-field; random variable/vector and their properties, some special distributions; E & σ^2 & cov and their properties; probability-generating/moment-generating/characteristic function; weak/strong law of large number, central limit thm.; intro. to multivariate normal distribution.

Section 1.1 Some Important Distributions

X	$p_X(k)//f_X(x)$	E	σ^2	PGF	MGF
$B(p)$		p	pq		$q+pe^s$
$B(n,p)$	$C_n^k p^k(1-p)^{n-k}$	np	npq	$(q+ps)^n$	$(q+pe^s)^n$
$G(p)$	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{ps}{1-qs}$	$\frac{pe^s}{1-qe^s}$
$H(n,M,N)$	$\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$	$n\frac{M}{N}$	$\frac{nM(N-n)(N-M)}{N^2(n-1)}$		
$P(\lambda)$	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ	$e^{\lambda(s-1)}$	$e^{\lambda(e^s-1)}$
$U(a,b)$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$		$\frac{e^{sb}-e^{sa}}{(b-a)s}$
$N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2		$e^{\frac{\sigma^2 s^2}{2}+\mu s}$
$\epsilon(\lambda)$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$		$\frac{\lambda}{\lambda-s}$
$\Gamma(\alpha,\lambda)$	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$		
$B(\alpha,\beta)$	$\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$		
χ_n^2	$\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}$	n	$2n$		
t_ν	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$	0	$\frac{\nu}{\nu-2}$		
$F(m,n)$	$\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\frac{m^{\frac{m}{2}}n^{\frac{n}{2}}x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}$	$\frac{n}{n-2}$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$		

Definition of PGF, MGF, CF see section 1.5.

More Properties of χ^2, t, F see section 1.8.2.

Section 1.2 Probability and Probability Model

What is **Probability**?

A 'belief' in the chance of an event occurring?

1.2.1 Sample and σ-Field

Def. sample space Ω: The set of all possible outcomes of one particular experiment.

Def. $\mathscr{F}$ a σ-field(or a σ-algebra) as a collection of some subsets of Ω **if**

- $\Omega \in \mathscr{F}$
- if $A \in \mathscr{F}$,then $A^C \in \mathscr{F}$
- if $A_n \in \mathscr{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$

And $(\Omega, \mathscr{F})$ is a measurable space.

1.2.2 Axioms of Probability

P is probability measure (or probability function) defined on $(\Omega, \mathscr{F})$, satisfying

- Nonnegativity

$$P(A) \geq 0 \qquad \forall A \in \Omega \tag{1.1}$$

- Normalization

$$P(\Omega) = 1 \tag{1.2}$$

- Countable Additivity

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots \quad (A_i \parallel A_j \quad \forall i \neq j) \tag{1.3}$$

Then $(\Omega, \mathscr{F}, P)$ is probability space.

Properties of Probability:

- Monotonicity

$$P(A) \leq P(B) \quad \text{for } A \subset B \tag{1.4}$$

- Finite Subadditivity (Boole Inequality)

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \tag{1.5}$$

- Inclusion-Exclusion Formula

$$\begin{aligned} P(\Cup_{i=1}^{n} A_i) = \sum_{1\leq i\leq n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) \\ + \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k) - \cdots \\ + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n) \end{aligned}$$

- Borel-Cantelli Lemma

$$\sum_{n=1}^{\infty} P(A_n) < \infty \Rightarrow P(\lim_{n\to\infty} \sup A_n) = 0$$

$$\sum_{n=1}^{\infty} P(A_n) = \infty \Rightarrow P(\lim_{n\to\infty} \sup A_n) = 1 \quad \text{if } A_i \text{ independent}$$

1.2.3 Conditional Probability

Def. **Conditional Probability** of B given A:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{1.6}$$

(Actually a change of σ-field from Ω to B)

Application of conditional probability:

- Multiplication Formula
$$P(\Cap_{i=1}^{n} A_i) = P(A_1) \prod_{i=2}^{n} P(A_i|A_1 \cap A_2 \cap \cdots \cap A_{i-1}) \tag{1.7}$$
- Total Probability Thm.
$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i) \tag{1.8}$$
where $\{A_i\}$ is a partition of Ω.
- Bayes's Rule
$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n} P(A_j)P(B|A_j)} \tag{1.9}$$
where $\{A_i\}$ is a partition of Ω.
- Statistically Independence
$$P(A \cap B) = P(A)P(B), \text{ for } A \parallel B \tag{1.10}$$

Section 1.3 Properties of Random Variable and Vector

1.3.1 Random Variable

Def. Random Variable: a **function** X defined on sample space Ω, mapping from Ω to some $\mathscr{X} \in \mathbb{R}$.

Then def. Cumulative Distribution Function (CDF).

$$F_X(x) = P(X \leq x) \tag{1.11}$$

For Discrete case, consider CDF as right-continuity.

- PMF:
$$p_X(x) = F_X(x^+) - F_X(x^-) \tag{1.12}$$
PDF:
$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} \tag{1.13}$$
- Indicator function:
$$I_{x\in A}(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{1.14}$$
- Convolution
  - $W = X + Y$
$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w - x) \mathrm{d}x \tag{1.15}$$

- $V = X - Y$

$$f_V(v) = \int_{-\infty}^{\infty} f_X(x) f_Y(x - v) \mathrm{d}x \tag{1.16}$$

- $Z = XY$

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y(\frac{z}{x}) \mathrm{d}x \tag{1.17}$$

- Order Statistics

  Def $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ as order statistics of $\vec{X}$

$$g_{X_{(i)}} = n! \prod f(x_i) \qquad \text{for } x_1 < x_2 \cdots < x_n \tag{1.18}$$

  PDF of $X_{(k)}$

$$g_k(x_k) = nC_{n-1}^{k-1}[F(x_k)]^{k-1}[1 - F(x_k)]^{n-k} f(x_k) \tag{1.19}$$

- p-fractile

$$\xi_p = F^{-1}(p) = \inf\{x | F(x) \geq p\} \tag{1.20}$$

1.3.2 Random Vector

A general case of random variable.

n-dimension Random Vector $\vec{X} = (X_1, X_2, \ldots, X_n)$ defined on $(\Omega, \mathscr{F}, P)$.

CDF $F(x_1, \ldots, x_n)$ defined on $\mathbb{R}^n$:

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \tag{1.21}$$

Joint PDF of random vector:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n} \tag{1.22}$$

k-dimensional Marginal Distribution: For $1 \leq k < n$ and index set $S_k = \{i_1 \ldots, i_k\}$, distribution of $\vec{X} = (X_{i_1}, X_{i_2}, \ldots, X_{i_k})$

$$F_{S_k}(x_{i_1}, X_{i_2} \leq x_{i_2} \ldots, x_{i_k}) = P(X_{i_1} \leq x_{i_1}, \ldots, X_{i_k} \leq x_{i_k}; X_{i_{k+1}}, \ldots, X_{i_n} \leq \infty) \tag{1.23}$$

Marginal distribution:

$$g_{S_k}(x_{i_1}, \ldots, x_{i_k}) = \int_{\mathbb{R}^{n-k}} f(x_1, \ldots, x_n) \mathrm{d}x_{i_{k+1}} \ldots \mathrm{d}x_{j_n} = \frac{\partial^{n-k} F(x_1, \ldots, x_n)}{\partial x_{i_{k+1}} \ldots \partial x_{i_n}} \tag{1.24}$$

△ Function of r.v.

For $\vec{X} = (X_1, X_2, \cdots, X_n)$ with PDF $f(\vec{X})$ and define

$$\vec{Y} = (Y_1, Y_2, \cdots, Y_n) = (y_1(\vec{X}), y_2(\vec{X}), \cdots, y_n(\vec{X})) \tag{1.25}$$

with inverse mapping

$$\vec{X} = (X_1, X_2, \cdots, X_n) = (x_1(\vec{Y}), x_2(\vec{Y}), \cdots, x_n(\vec{Y})) \tag{1.26}$$

then

$$g(\vec{Y}) = f(x_1(\vec{Y}), x_2(\vec{Y}), \cdots, x_n(\vec{Y})) \left| \frac{\partial \vec{X}}{\partial \vec{Y}} \right| I_{D_Y} \tag{1.27}$$

(Intuitively: $g(\vec{Y})\mathrm{d}\vec{Y} = \mathrm{d}P = f(\vec{X})\mathrm{d}\vec{X}$)

Section 1.4 Properties of E , σ^2 and cov

Expectation and Variance of common distributions see sec.1.1.

1.4.1 Expection

Expectation of r.v. $g(X)$ def.:

$$E[g(X)] = \begin{cases} \int_\Omega g(x) f_X(x) \mathrm{d}x = \int_\Omega g(x) \mathrm{d}F(x) \\ \sum_\Omega g(X) f_X(x) \end{cases} \tag{1.28}$$

Properties of expectation $E(\cdot)$:

- Linearity of Expectation

$$E(aX + bY) = aE(X) + bE(Y) \tag{1.29}$$

- Conditional Expectation

$$E(X|A) = \frac{E(XI_A)}{P(A)} \tag{1.30}$$

  Note: if take A as Y is also a r.v. then

$$m(Y) = E(X|Y) = \int x f_{X|Y}(x) \mathrm{d}x \tag{1.31}$$

  is actually a function of Y

- Law of Total Expectation

$$E\{E[g(X)|Y]\} = E[g(X)] \tag{1.32}$$

- r.v.& Event

$$P(A|X) = E(I_A|X) \Rightarrow E[P(A|X)] = E(I_A) = P(A) \tag{1.33}$$

-

$$E[h(Y)g(X)|Y] = h(Y)E[g(X)|Y] \tag{1.34}$$

1.4.2 Variance

Variance of r.v. X:

$$var(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2 \tag{1.35}$$

(sometimes denoted as σ_X^2.)

Properties:

- Linear combination of Variance

$$var(aX + b) = a^2 var(X) \tag{1.36}$$

- Conditional Variance

$$var(X|Y) = E[X - E(X|Y)]^2|Y \tag{1.37}$$

- Law of Total Variance

$$var(X) = E[var(X|Y)] + var[E(X|Y)] \tag{1.38}$$

Standard Deviation def. as :

$$\sigma_X = \sqrt{var(X)} \tag{1.39}$$

Then can construct **Standardization** of r.v.

$$Y = \frac{X - E(X)}{\sqrt{var(X)}} \tag{1.40}$$

1.4.3 Covariance and Correlation

Covariance of r.v. X and Y:

$$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y) \tag{1.41}$$

And (Pearson's) Correlation Coefficient

$$\rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} \tag{1.42}$$

Remark: correlation $\not\Rightarrow$ cause and effect.

Properties:

- Bilinear of Covariance

$$cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)$$
$$cov(X, Y + Z) = cov(X, Y) + cov(X, Z)$$

- Variance and Covariance

$$label EqaVarOfSumOfRV var(X + Y) = var(X) + var(Y) + 2cov(X, Y) \tag{1.43}$$

- Covariance Matrix

  Def $\Sigma = E[(X - \mu)(X - \mu)^T] = \{\sigma_{ij}\}$ (where X should be considered as a column vector)

$$\Sigma = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \dots & cov(X_1, X_n) \\ cov(X_2, X_1) & var(X_2) & \dots & cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & \dots & var(X_n) \end{pmatrix} \tag{1.44}$$

Attachment: Independence:

$$X_i || X_j \Rightarrow \begin{cases} f(x_1, x_2, \cdots, x_n) = \prod f(x_i) \\ F(x_1, x_2, \cdots, x_n) = \prod F(x_i) \\ E(\prod X_i) = \prod E(X_i) \\ var(\sum X_i) = \sum var(X_i) \end{cases} \tag{1.45}$$

Section 1.5 PGF, MGF and C.F

Generating Function: Representation of P in function space. $P \Leftrightarrow$ Generating Function.

1.5.1 Probability Generating Function

PGF: used for non-negative, integer X

$$g(s) = E(s^X) = \sum_{j=0}^{\infty} s^j P(X=j), s \in [-1,1] \tag{1.46}$$

Properties

- $P(X=k) = \dfrac{g^{(k)}(0)}{k!}$
- $E(X) = g^{(1)}(1)$
- $var(X) = g^{(2)}(1) + g^{(1)}(1) - [g^{(1)}(1)]^2$
- For $X_1, X_2, \cdots, X_n$ independent with $g_i(s) = E(s^{X_i})$, $Y = \sum_{i=1}^{n} X_i$, then

$$g_Y(s) = \prod_{i=1}^{n} g_i(s), s \in [-1,1] \tag{1.47}$$

- For X_i i.i.d with $\psi(s) = E(s^{X_i})$, Y with $G(s) = E(s^Y)$, $W = X_1 + X_2 + \cdots + X_Y$,then

$$g_W(s) = G[\psi(s)] \tag{1.48}$$

- 2-Dimensional PGF of (X,Y)

$$g(s,t) = E(s^X t^Y) = \sum_{i=o}^{\infty}\sum_{j=0}^{\infty} P_{(X,Y)}(X=i, Y=j) s^i t^j,\ s,t \in [-1,1] \tag{1.49}$$

1.5.2 Moment Generating Function

MGF:

$$M_X(s) = E(e^{sX}) = \begin{cases} \sum_j e^{sx} P(X=x_j) \\ \int_{-\infty}^{\infty} e^{sx} f_X(x) \mathrm{d}x \end{cases} \tag{1.50}$$

Properties

- MGF of $Y = aX + b$: $M_Y(s) = e^{sb} M(sa)$
- $E(X^k) = M^{(k)}(0)$
- $P(X=0) = \lim_{s \to -\infty} M(s)$
- For $X_1, X_2, \cdots, X_n$ independent with $M_{X_i}(s) = E(e^{sX_i})$, $Y = \sum_{i=1}^{n} X_i$, then

$$M_Y(s) = \prod_{i=1}^{n} M_{X_i}(s) \tag{1.51}$$

1.5.3 Characteristic Function

C.F is actually the Fourier Transform of f.

$$\phi(t) = E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\mathrm{d}x \tag{1.52}$$

Properties

- if $E(|X|^k) < \infty$,then
$$\phi^{(k)}(t) = i^k E(X^k e^{itX}) \qquad \phi^{(k)}(0) = i^k E(X^k) \tag{1.53}$$
- For $X_1, X_2, \cdots, X_n$ independent with $\phi_{X_i}(t) = E(e^{itX_i})$, $Y = \sum\limits_{i=1}^{n} X_i$, then
$$\phi_Y(t) = \prod_{i=1}^{n} \phi_{X_i}(t) \tag{1.54}$$
- Inverse (Fourier) Transform
$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t)\mathrm{d}t \tag{1.55}$$

Section 1.6 Convergence and Limit Distribution

1.6.1 Convergence Mode

$$\begin{cases} \text{Convergence in Distribution} & X_n \xrightarrow{\mathscr{L}} X : \lim\limits_{n\to\infty} F_n(x) = F(x) \\ \text{Convergence in Probability} & X_n \xrightarrow{p} X : \lim\limits_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0\,, \forall \varepsilon > 0 \\ \text{Almost Sure Convergence} & X_n \xrightarrow{\text{a.s.}} X : P(\lim\limits_{n\to\infty} X_n = X) = 1 \\ L_p \text{ Convergence} & X_n \xrightarrow{L_p} X : \lim\limits_{n\to\infty} E(|X_n - X|^p) = 0 \end{cases} \tag{1.56}$$

Relations between convergence:

$$\begin{tikzcd}
\boxed{X_n \xrightarrow{\text{a.s.}} X} \arrow[dr] & & \\
& \boxed{X_n \xrightarrow{p} X} \arrow[r] & \boxed{X_n \xrightarrow{\mathscr{L}} X} \\
\boxed{X_n \xrightarrow{L_p} X} \arrow[ur] & &
\end{tikzcd}$$

Useful Thm.:

- Continuous Mapping Thm.: For continuous function $g(\cdot)$

  1. $X_n \xrightarrow{\text{a.s.}} X \Rightarrow g(X_n) \xrightarrow{\text{a.s.}} g(X)$
  2. $X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$

3. $X_n \xrightarrow{\mathscr{L}} X \Rightarrow g(X_n) \xrightarrow{\mathscr{L}} g(X)$

- Slutsky's Thm.: For $X_n \xrightarrow{\mathscr{L}} X, Y_n \xrightarrow{p} c$

1. $X_n + Y_n \xrightarrow{\mathscr{L}} X + c$
2. $X_n Y_n \xrightarrow{\mathscr{L}} cX$
3. $X_n / Y_n \xrightarrow{\mathscr{L}} X/c$

- Continuity Thm.

$$\lim_{n\to\infty} \phi_n(t) = \varphi(t) \Leftrightarrow X_n \xrightarrow{\mathscr{L}} X \tag{1.57}$$

1.6.2 Law of Large Number & Central Limit Theorem

- WLLN

$$\frac{1}{n}\sum X_i \xrightarrow{p} E(X_1) \tag{1.58}$$

- SLLN

$$\frac{1}{n}\sum X_i \xrightarrow{\text{a.s.}} C \tag{1.59}$$

- CLT

$$\frac{1}{\sigma\sqrt{n}}\sum (X_k - \mu) \xrightarrow{\mathscr{L}} N(0,1) \tag{1.60}$$

- de Moivre-Laplace Thm.

$$P(k \le S_n \le m) \approx \Phi(\frac{m + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{k - 0.5 - np}{\sqrt{npq}}) \tag{1.61}$$

- Stirling Eqa

$$\frac{\lambda^k}{k!}e^{-\lambda} \approx \frac{1}{\sqrt{\lambda}\sqrt{2\pi}} e^{-\frac{(k-\lambda)^2}{2\lambda}} \xrightarrow[\lambda=n]{k=n} n! \approx \sqrt{2\pi n}(\frac{n}{e})^n \tag{1.62}$$

Section 1.7 Inequalities

- Cauchy-Schwarz Inequality

$$|E(XY)| \le \sqrt{E(X^2)E(Y^2)} \tag{1.63}$$

- Bonferroni Inequality

$$P(\bigcup_{i=1}^{n} A_i) \ge \sum_{1\le i\le n} P(A_i) + \sum_{1\le i<j\le n} P(A_i \cap A_j) \tag{1.64}$$

- Markov Inequality

$$P(|X| \ge \epsilon) \le \frac{E(|X|^\alpha)}{\epsilon^\alpha} \tag{1.65}$$

- Chebyshev Inequality

$$P(|X - E(X)| \ge \epsilon) \le \frac{var(X)}{\epsilon^2} \tag{1.66}$$

- Jensen Inequality: For convex function $g(x)$:

$$E[g(X)] \ge g(E(X)) \tag{1.67}$$

Section 1.8 Multivariate Normal Distribution

For $X_1, X_2, \cdots, X_n$ independent and $X_k \sim N(\mu_k, \sigma_k^2),\ k = 1, \cdots, n$, $T = \sum_{k=1}^{n} c_k X_k$, ($c_k$ const), then

$$T \sim N(\sum_{k=1}^{n} c_k \mu_k, \sum_{k=1}^{n} c_k^2 \sigma_k^2) \tag{1.68}$$

Deduction in some special cases:

- Given $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$, i.e. X_k i.i.d., then

$$T \sim N(\mu \sum_{k=1}^{n} c_k, \sigma^2 \sum_{k=1}^{n} c_k^2) \tag{1.69}$$

- Further take $c_1 = c_2 = \cdots = c_n = \frac{1}{n}$, i.e. $T = \sum_{k=1}^{n} X_k / n = \bar{X}$, then

$$T = \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{1.70}$$

1.8.1 Linear Transform

First consider $\epsilon_1, \epsilon_2, \cdots, \epsilon_m$ i.i.d. $\sim N(0,1)$, $n \times 1$ const column vector $\vec{\mu}$, $n \times m$ const matrix $\mathbf{B} = \{b_{ij}\}$, def.$X_i = \sum_{j=1}^{m} b_{ij}\epsilon_j$, i.e.

$$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix} + \vec{\mu} \tag{1.71}$$

We have: $\vec{X} \sim N(\vec{\mu}, \Sigma)$, where Σ, as defined in eqa.1.44 is

$$\Sigma = E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T] = \mathbf{B}\mathbf{B}^T = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \dots & cov(X_1, X_n) \\ cov(X_2, X_1) & var(X_2) & \dots & cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & \dots & var(X_n) \end{pmatrix} = \{\sigma_{ij}\} \tag{1.72}$$

Furthur Consider $\vec{Y} = (Y_1, \cdots, Y_n)^T$, $n \times n$ const square matrix $\mathbf{A} = \{a_{ij}\}$ and def. $\vec{Y} = \mathbf{A}\vec{X}$ i.e.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{1.73}$$

Then $\vec{Y} \sim N(\mathbf{A}\vec{\mu}, \mathbf{A}\Sigma\mathbf{A}^T)$

Special case: $X_1, \cdots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$, $\vec{X} = (X_1, \cdots, X_n)^T$,

$$
\begin{aligned}
E(Y_i) =& \mu \sum_{k=1}^{n} a_{ik} \\
var(Y_i) =& \sigma^2 \sum_{k=1}^{n} a_{ik}^2 \\
cov(Y_i, Y_j) =& \sigma^2 \sum_{k=1}^{n} a_{ik} a_{jk}
\end{aligned}
$$

Specially when $\mathbf{A} = \{a_{ij}\}$ orthonormal, we have $Y_1, \cdots, Y_n$ independent

$$Y_i \sim N(\mu \sum_{k=1}^{n} a_{ik}, \sigma^2) \tag{1.74}$$

1.8.2 Distributions of Function of Normal Variable: χ^2, t & F

Consider $X_1, X_2, \ldots, X_n$ i.i.d. $\sim N(0,1)$; $Y, Y_1, Y_2, \ldots, Y_m$ i.i.d. $\sim N(0,1)$

- χ^2 Distribution: Def. χ^2 distribution with degree of freedom n:

$$\xi = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2 \tag{1.75}$$

PDF of χ_n^2:

$$g_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2} e^{-x/2} I_{x>0} \tag{1.76}$$

Properties

  – E and var of $\xi \sim \chi_n^2$

$$E(\xi) = n \qquad var(\xi) = 2n \tag{1.77}$$

  – For independent $\xi_i \sim \chi_{n_i}^2$, $i = 1, 2, \ldots, k$:

$$\xi_0 = \sum_{i=1}^{k} \xi_i \sim \chi^2_{n_1+\ldots+n_k} \tag{1.78}$$

  – Denoted as $\Gamma(\alpha, \lambda)$:

$$\xi = \sum_{i=1}^{n} X_i \sim \Gamma(\frac{n}{2}, \frac{1}{2}) = \chi_n^2 \tag{1.79}$$

- t Distribution: Def. t distribution with degree of freedom n:

$$T = \frac{Y}{\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}}} = \frac{Y}{\sqrt{\frac{\xi}{n}}} \sim t_n \tag{1.80}$$

(Usually take ν instead of n)

PDF of t_ν:

$$t_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{1.81}$$

Denote: Upper α-fractile of t_ν, satisfies $P(T \geq c) = \alpha$:

$$c = t_{\nu,\alpha} \tag{1.82}$$

(Similar for χ_n^2 and $F_{m,n}$ etc.)

- F Distribution: Def. F distribution with degree of freedom m and n:

$$F = \frac{\sum_{i=1}^m Y_i}{\sum_{i=1}^n X_i} \sim F_{m,n} \tag{1.83}$$

PDF of $F_{m,n}$:

$$f_{m,n}(x) = \frac{\Gamma(\frac{m+n}{2}) m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} x^{\frac{m}{2}-1}(mx+n)^{-\frac{m+n}{2}} I_{x>0} \tag{1.84}$$

Properties

  - If $Z \sim F_{m,n}$, then $\dfrac{1}{Z} \sim F_{n,m}$.
  - If $T \sim t_n$, then $T^2 \sim F_{1,n}$
  - $F_{m,n,1-\alpha} = \dfrac{1}{F_{n,m,\alpha}}$

☐ Some useful Lemma (uesd in statistic inference, see section 2.3.3):

- For $X_1, X_2, \dots, X_n$ independent with $X_i \sim N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \sim \chi_n^2 \tag{1.85}$$

- For $X_1, X_2, \dots, X_n$ i.i.d.$\sim N(\mu, \sigma^2)$, then

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1} \tag{1.86}$$

For $X_1, X_2, \dots, X_m$ i.i.d.$\sim N(\mu_1, \sigma^2)$, $Y_1, Y_2, \dots, Y_n$ i.i.d.$\sim N(\mu_2, \sigma^2)$, denote sample pooled variance $S_\omega^2 = \dfrac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$, then

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_\omega} \cdot \sqrt{\frac{mn}{m+n}} \sim t_{m+n-2} \tag{1.87}$$

- For $X_1, X_2, \dots, X_m$ i.i.d.$\sim N(\mu, \sigma^2)$, $Y_1, Y_2, \dots, Y_n$ i.i.d.$\sim N(\mu_2, \sigma^2)$, then

$$T = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1} \tag{1.88}$$

- For $X_1, X_2, \dots, X_n$ i.i.d. $\sim \epsilon(\lambda)$, then

$$2\lambda n\bar{X} = 2\lambda \sum_{i=1}^n X_i \sim \chi_{2n}^2 \tag{1.89}$$

Remark: for $X_i \sim \epsilon(\lambda) = \Gamma(1, \lambda) \Rightarrow 2\lambda \sum_{i=1}^n X_i \sim \Gamma(n, 1/2) = \chi_{2n}^2$.

Chapter. II 统计推断部分

Instructor: Jiangdian Wang

Statistical Inference: use sample to estimate population.

Two main tasks of Statistical Inference:

- Parameter Estimation
  - Point Estimation: 2.2
  - Interval Estimation: 2.3
- Hypothesis Testing: 2.4

Section 2.1 Statistical Model and Statistics

Random sample comes from population X. In parametric model case, we have population distribution family:

$$\mathscr{F} = \{f(x;\vec{\theta})|\vec{\theta} \in \Theta\} \tag{2.1}$$

where parameter $\vec{\theta}$ reflect some quantities of population (e.g. mean, variance, etc.), each $\vec{\theta}$ corresponds to a distribution of population X.

Sample space: Def. as $\mathscr{X} = \{\{x_1, x_2, \ldots, x_n\}, \forall x_i\}$, then $\{X_i\} \in \mathscr{X}$ is random sample from population $X \sim f(x;\vec{\theta})$.

2.1.1 Statistics

Statistic(s): function of random sample $\vec{T}(X_1, X_2, \ldots, X_n)$, **but not a function of parameter**.

Some useful statistics, e.g.

- Sample mean (Consider X_i i.i.d.)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.2}$$

- Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.3}$$

- Sample moments
  - Origin moment

$$a_{n,k} = \frac{1}{n}\sum_{i=1}^{k} X_i^k \qquad k = 1, 2, 3, \ldots \tag{2.4}$$

  - Center moment

$$m_{n,k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \qquad k = 2, 3, 4, \ldots \tag{2.5}$$

- Order statistics

$$(X_{(1)}, X_{(2)}, \ldots, X_{(n)}), \text{ for } X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \tag{2.6}$$

- Sample p-fractile

$$m_p = X_{(m)}, \quad m = [(n+1)p] \tag{2.7}$$

- Sample coefficient of variation

$$\hat{\nu} = \frac{S}{\bar{X}} \tag{2.8}$$

- Skewness and Kurtosis

$$\hat{g}_1 = \frac{m_{n,3}}{m_{n,2}^{3/2}} \qquad \hat{g}_2 = \frac{m_{n,4}}{m_{n,2}^2} - 3 \tag{2.9}$$

□ Properties

Statistic T is a function of random sample $\{X_i\}$, thus has distribution (say $g_T(t)$) called **Sampling Distribution**.

For X_i i.i.d. from $X \sim f(x)$ with population mean μ and variance σ^2

- Calculation of sample variance S^2

$$(n-1)S^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{2.10}$$

- E and var of $\bar{X}$ and S^2

$$E(\bar{X}) = \mu \qquad var(\bar{X}) = \frac{\sigma^2}{n} \qquad E(S^2) = \sigma^2 \tag{2.11}$$

Further if X_i i.i.d. from $X \sim N(\mu, \sigma^2)$ where μ and σ^2 unknown.

- Independence of $\bar{X}$ and S^2

$$\bar{X} \text{ and } S^2 \text{ are independent} \tag{2.12}$$

  - Distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{2.13}$$

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

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2.14}$$

2.1.2 Exponential Family

Def. $\mathscr{F} = \{f(x;\vec{\theta}|\vec{\theta} \in \Theta)\}$ is **Exponential Family** if $f(x;\vec{\theta})$ has the form as

$$f(x;\vec{\theta}) = C(\vec{\theta})h(x)\exp\left[\sum_{i=1}^{k} Q_i(\vec{\theta})T_i(x)\right] \quad \vec{\theta} \in \Theta \tag{2.15}$$

Canonical Form: Take $Q_i(\vec{\theta}) = \varphi_i$, then $\vec{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_k)$ =$(Q_1(\vec{\theta}), Q_2(\vec{\theta}), \ldots, Q_k(\vec{\theta}))$ is a transform from Θ to Θ^*, s.t. $\mathscr{F}$ has canonical form, i.e.

$$f(x;\vec{\varphi}) = C^*(\vec{\varphi})h(x)\exp\left[\sum_{i=1}^{k} \varphi_i T_i(x)\right] \quad \vec{\varphi} \in \Theta^* \tag{2.16}$$

Θ^* is canonical parameter space.

□ Why we need exponential family? Have some nice properties.

2.1.3 Sufficient and Complete Statistics

Note: For simplification, the following parts denote $\vec{\theta}, \vec{T}, \dots$ as $\theta, T, \dots$ etc.

▶ A **Sufficient Statistic** $T(\vec{X})$ for θ contains all the information of sample when infer θ, i.e.

$$f(\vec{X}; T(\vec{X})) = f(\vec{X}; T(\vec{X}), \theta) \tag{2.17}$$

Properties

- **Factorization Thm.** $T(\vec{X})$ is sufficient **if and only if** $f_{\vec{X}}(\vec{x};\theta) = f(\vec{x};\theta)$ can be written as

$$f(\vec{x};\theta) = g[t(\vec{x});\theta]h(\vec{x}) \tag{2.18}$$

- If $T(\vec{X})$ sufficient, then $T'(\vec{X}) = g[T(\vec{X})]$ also.(require g single-valued and invertible)
- If $T(\vec{X})$ sufficient, then (T, T_1) also.
- Minimal sufficient statistic $T_\theta(\vec{X})$ satisfies

$$\forall \text{sufficient statistic } S,\ \exists\, q_S(\cdot),\ \text{s.t.} T_\theta = q_S(S) \tag{2.19}$$

  A minimal sufficient statistic not always exists.

  Sufficient & Complete $\Rightarrow$ Minimal sufficient.

- Usually dimension of $\vec{T}_\theta$ and θ equals.

Sufficient statistic is **not** unique.

▶ A **Complete Statistic** $T(\vec{X})$ for θ satisfies

$$\forall\theta \in \Theta\,;\ \forall\varphi \text{ satisfies } E[\varphi(T(\vec{X}))] = 0, \text{ we have } P[\varphi(T) = 0;\theta] = 1 \tag{2.20}$$

Explanation: $T \sim g_T(t)$. Rewrite as

$$\int \varphi(t) g_T(t)\,\mathrm{d}t = 0\ \forall\,\theta \Rightarrow \varphi(T) = 0 \text{ a.s.} \tag{2.21}$$

i.e. span$\{g_T(t);\forall\theta\}$ is a complete space. Or to say that $\nexists$ none-zero $\varphi(t)$ so that $E(\varphi(T)) = 0$ (unbiased estimation)

$$\varphi(T) \neq 0\ \forall\theta \Rightarrow E[\varphi(T(\vec{X}))] \neq 0 \tag{2.22}$$

So make sure the uniqueness of unbiased estimation of $\hat{\theta}$ using T.

Properties

- If $T(\vec{X})$ complete, then $T'(\vec{X}) = g[T(\vec{X})]$ also.(require g measurable)
- A complete statistic not always exists.

- An **Ancillary Statistic** $S(\vec{X})$ is a statistic whose distribution does not depend on θ

  Basu Thm.: $\vec{X} = (X_1, X_2, \ldots, X_n)$ is sample from $\mathscr{F} = \{f(x;\theta), \theta \in \Theta\}$. $T(\vec{X})$ is a complete and minimal sufficient statistic, $S(\vec{X})$ is ancillary statistic, then $S(\vec{X}) \parallel T(\vec{X})$.

- Exponential family: For $\vec{X} = (X_1, X_2, \ldots, X_n)$ from exponential family with canonical form, i.e.

$$f(\vec{x};\theta) = C(\theta)h(\vec{x}) \exp\left[\sum_{i=1}^{k} \theta_i T_i(\vec{x})\right], \quad \theta \in \Theta \tag{2.23}$$

  Then if $\Theta \in \mathbb{R}^k$ interior point exists, then $T(\vec{X}) = (T_1(\vec{X}), T_2(\vec{X}), \ldots, T_k(\vec{X}))$ is sufficient & complete statistic.

Section 2.2 Point Estimation

For parametric distribution family $\mathscr{F} = \{f(x,\theta), \theta \in \Theta\}$, random sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ from $\mathscr{F}$. $g(\theta)$ is a function defined on Θ.

Mission: use sample $\{X_i\}$ to estimate $g(\theta)$, called **Parameter Estimation**.

$$\text{Parameter Estimation} \begin{cases} \text{Point Estimation} & \surd \\ \text{Interval Estimation} & \end{cases} \tag{2.24}$$

Point estimation: when estimating θ or $g(\theta)$, denote the estimator (defined on sample space $\mathscr{X}$) as

$$\hat{\theta}(\vec{X}) \qquad \hat{g}(\vec{X}) \tag{2.25}$$

Estimator is a statistic, with sampling distribution.

2.2.1 Optimal Criterion

Some nice properties of estimators (that we expect)

- Unbiasedness

$$E(\hat{\theta}) = \theta \quad \text{or} \quad E(\hat{g}(\vec{X})) = g(\theta) \tag{2.26}$$

  Otherwise, say $\hat{\theta}$ or $\hat{g}$ biased. Def. **Bias**: $E(\hat{\theta}) - \theta$

  Asymptotically unbiasedness

$$\lim_{n\to\infty} E(\hat{g}_n(\vec{X})) = g(\theta) \tag{2.27}$$

- Efficiency: say $\hat{g}_1(\vec{X})$ is more efficient than $\hat{g}_2(\vec{X})$, if

$$var(\hat{g}_1) \leq var(\hat{g}_2) \quad \forall \theta \in \Theta \tag{2.28}$$

- Mean Squared Error (MSE)

$$\text{MSE} = E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) + [Bias(\hat{\theta})]^2 \tag{2.29}$$

  For unbiased estimator, i.e. $Bias(\hat{\theta}) = 0$, we have

$$\text{MSE} = E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) \tag{2.30}$$

- (Weak) Consistency

$$\lim_{n\to\infty} P(|\hat{g}_n(\vec{X}) - g(\theta)| \geq \varepsilon) = 0 \quad \forall \varepsilon > 0 \tag{2.31}$$

- Asymptotic Normality

2.2.2 Method of Moments

Review: Population moments & Sample moments

$$\alpha_k = E(X^k) \qquad \mu_k = E[(X - E(X))^k]$$

$$a_{n,k} = \frac{1}{n}\sum_{i=1}^n X_i^k \qquad m_{n,k} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^k$$

Property: $a_{n,k}$ is the unbiased estimator of α_k.(while $m_{n,k}$ unually biased for μ_k)

For sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ from $\mathscr{F} = \{f(x;\theta, \theta \in \Theta)\}$, unknown parameter (or its function) $g(\theta)$ can be written as

$$g(\theta) = G(\alpha_1, \alpha_2, \ldots, \alpha_k; \mu_2, \mu_3, \ldots, \mu_l) \tag{2.32}$$

Then its **Moment Estimate** $\hat{g}(\vec{X})$ is

$$\hat{g}(\vec{X}) = G(a_{n,1}, a_{n,2}, \ldots, a_{n,k}; m_{n,2}, m_{n,3}, \ldots, m_{n,l}) \tag{2.33}$$

Example: coefficient of variance & skewness

$$\hat{\nu} = \frac{S}{\bar{X}} \quad \hat{\beta}_1 = \frac{m_{n,3}}{m_{n,2^{3/2}}} = \sqrt{n}\frac{\sum_{i=1}^n (X_i - \bar{X})^3}{[\sum_{i=1}^n (X_i - \bar{X})^2]^{\frac{3}{2}}} \tag{2.34}$$

☐ Note:

- G may not have explicit expression.
- Moment estimate may not be unique.
- If $G = \sum_{i=1}^k c_i\alpha_i$ (linear combination of α, without μ), then $\hat{g}(\vec{X}) = \sum_{i=1}^k c_i a_{n,i}$ unbiased.

  Usually $\hat{g}(\vec{X})$ is asymptotically unbiased.
- For small sample, not so accurate.
- May not contain all the information about θ, i.e. may not be sufficient statistic.
- Do not require a statistic model.

2.2.3 Maximum Likelihood Estimation

For sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ with distribution $f(\vec{x};\theta)$ from $\mathscr{F} = \{f(x;\theta), \theta \in \Theta\}$, def. **Likelihood Function** $L(\theta;\vec{x})$, defined on Θ (as a function of θ)

$$L(\theta;\vec{x}) = f(\vec{x};\theta) \qquad \theta \in \Theta,\, \vec{x} \in \mathscr{X} \tag{2.35}$$

Also def. log-likelihood function $l(\theta;\vec{x}) = \ln L(\theta;\vec{x})$.

If estimator $\hat{\theta} = \hat{\theta}(\vec{X})$ satisfies

$$L(\hat{\theta};\vec{x}) = \sup_{\theta\in\Theta} L(\theta;\vec{x}), \quad \vec{x} \in \mathscr{X} \tag{2.36}$$

Or equivalently take $l(\theta;\vec{x})$ instead of $L(\theta;\vec{x})$.

Then $\hat{\theta} = \hat{\theta}(\vec{X})$ is a **Maximum Likelihood Estimate**(MLE) of $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$

How to identify MLE?

- Differentiation: Fermat Lemma

$$\left.\frac{\partial L}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0 \qquad \left.\frac{\partial^2 L}{\partial \theta_i \partial \theta_j}\right|_{\theta=\hat{\theta}} \text{negative definite} \qquad \forall i,j = 1,2,\ldots,k \tag{2.37}$$

- Graphing method.
- Numerically compute maximum.

☐ Some properties of MLE

- (Depend on the case, not always) unbiased.
- Invariance of MLE: If $\hat{\theta}$ is MLE of θ, invertible function $g(\theta)$, then $g(\hat{\theta})$ is MLE of $g(\theta)$.
- MLE and Sufficiency: $T = T(X_1, X_2, \ldots, X_n)$ is a sufficient statistic of θ, if MLE of θ exists, say $\hat{\theta}$, then $\hat{\theta}$ is a function of T, i.e.

$$\hat{\theta} = \hat{\theta}(\vec{X}) = \hat{\theta}^*(T(\vec{X})) \tag{2.38}$$

- Asymptotic Normality:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma_\theta^2), \quad \sigma_\theta^2 = \frac{1}{E_\theta[\frac{\partial}{\partial\theta} \ln f(\vec{X};\theta)]^2} \tag{2.39}$$

  i.e.

$$\hat{\theta}_n \xrightarrow{d} N(\theta, \frac{\sigma_\theta^2}{n}) \tag{2.40}$$

☐ Comparison: MoM and MLE

- MoM do not require statistic model; MLE need to know PDF.
- MoM is more robust than MLE.

MLE in Exponential Family:

For sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ from canonical exponential family $\mathscr{F} = \{f(x;\theta), \theta \in \Theta\}$

$$f(x;\theta) = C(\theta)h(x)\exp\left[\sum_{i=1}^{k}\theta_i T_i(x)\right] \quad \theta = (\theta_1, \ldots, \theta_k) \in \Theta \tag{2.41}$$

Likelihood function $L(\theta, \vec{x}) = \prod_{j=i}^{n} f(x_j;\theta)$ and log-likelihood function $l(\theta, \vec{x})$

$$L(\theta, \vec{x}) = C^n(\theta)\prod_{j=1}^{n} h(x_j)\exp\left[\sum_{i=1}^{k}\theta_i\sum_{j=1}^{n}T_i(x_j)\right]$$

$$l(\theta, \vec{x}) = n\ln C(\theta) + \sum_{j=1}^{n}\ln h(x_j) + \sum_{i=1}^{k}\theta_i\sum_{j=1}^{n}T_i(x_j)$$

Solution of MLE: (Require $\hat{\theta} \in \Theta$)

$$\frac{n}{C(\theta)}\frac{\partial C(\theta)}{\partial\theta_i}\bigg|_{\theta=\hat{\theta}} = -\sum_{j=1}^{n}T_i(x_j), \quad i = 1, 2, \ldots, k \tag{2.42}$$

2.2.4 Uniformly Minimum Variance Unbiased Estimator

MSE: For $\hat{g}(\vec{X})$ is estimate of $g(\theta)$,then MSE

$$\text{MSE}(\hat{g}(\vec{X})) = E[(\hat{g}(\vec{X}) - g(\theta))^2] = var(\hat{g}) + [Bias(\hat{g})]^2 \tag{2.43}$$

Note: Unbiased estimator (i.e. $Bias(\hat{g}) = 0$) not unique; not always exist.

Now only consider unbiased estimators of $g(\theta)$ exists, say $\hat{g}(\vec{X})$, then

$$\text{MSE}(\hat{g}(\vec{X})) = var(\hat{g}(\vec{X})) \tag{2.44}$$

If $\forall$ unbiased estimate $\hat{g}\prime(\vec{X})$, $\hat{g}$ satisfies

$$var[\hat{g}(\vec{X})] \le var[\hat{g}\prime(\vec{X})] \tag{2.45}$$

□ Then $\hat{g}(\vec{X})$ is **Uniformly Minimum Variance Unbiased Estimator(UMVUE)** of $g(\theta)$

How to determine UMVUE? (Not an easy task)

- Zero Unbiased Estimate Method
- Sufficient and Complete Statistic Method
- Cramer-Rao Inequality

1. **Zero Unbiased Estimate Method**

Let $\hat{g}(\vec{X})$ be an unbiased estimate with $var(\hat{g}) < \infty$. If $\forall\ E(\hat{l}(\vec{X})) = 0$, $\hat{g}$ holds that

$$cov(\hat{g}, \hat{l}) = E(\hat{g} \cdot \hat{l}) = 0, \quad \forall\theta \in \Theta \tag{2.46}$$

Then $\hat{g}$ is a UMVUE of $g(\theta)$ (sufficient & necessary).

2. **Sufficient and Complete Statistic Method**

   For $T(\vec{X})$ sufficient statistic, $\hat{g}(\vec{X})$ unbiased estimate of $g(\theta)$, then

$$h(T) = E(\hat{g}(\vec{X})|T) \tag{2.47}$$

   is an unbiased estimate of $g(\theta)$ and $var(h(T)) \leq var(\hat{g})$.

   Remark:

   - A method to improve estimator.
   - A UMVUE has to be a function of sufficient statistic.

   Lehmann-Scheffé Thm.: For $\vec{X} = (X_1, X_2, \ldots, X_n)$ from population $X \sim \mathscr{F} = \{f(x, \theta, \theta \vec{\in} \Theta)\}$. $T(\vec{X})$ sufficient and complete, and $\hat{g}(T(\vec{X}))$ be an unbiased estimator, then $\hat{g}(T(\vec{X}))$ is the unique UMVUE.

   Can be used to construct UMVUE: given $T(\vec{X})$ sufficient and complete and some unbiased estimator $\hat{g}\prime(\theta)$ then

$$\hat{g}(T) = E(\hat{g}\prime|T) \tag{2.48}$$

   is the unique UMVUE.

3. **Cramer-Rao Inequality**

   Core idea: determine a lower bound of $var(\hat{g})$.

   Consider $\theta = \theta$ (One dimension parameter); For $\{X_i\}$ i.i.d. $f(x, \theta)$: def.

   - **Score function**: Reflects the steepness/slope of likelihood function f.

$$S(\vec{x}; \theta) = \frac{\partial \ln f(\vec{x}; \theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i; \theta)}{\partial \theta} \tag{2.49}$$

$$E[S(\vec{X}; \theta)] = 0 \tag{2.50}$$

   - **Fisher Information**: Variance of $S(\vec{x}; \theta)$, reflects the accuracy to conduct estimation, i.e. reflects information of statistic model that sample brings.

$$I(\theta) = E\left[\left(\frac{\partial \ln f(\vec{x}; \theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln f(\vec{x}; \theta)}{\partial \theta^2}\right] \tag{2.51}$$

   Consider $\mathscr{F}$ satisfies some regularity conditions(in most cases, regularity conditions do hold), then the lower bound of $var(\hat{g})$ satisfies **Cramer-Rao Inequality**:

$$var(\hat{g}(\vec{X})) \geq \frac{[g'(\theta)]^2}{nI(\theta)} \tag{2.52}$$

   Special case: $g(\theta) = \theta$ then

$$var(\hat{\theta}) \geq \frac{1}{nI(\theta)} \tag{2.53}$$

   note:

   - C-R Inequality determine a lower bound, not the infimum(i.e. UMVUE$\not\Rightarrow$ $var(\hat{g}(\vec{X})) = \dfrac{[g'(\theta)]^2}{nI(\theta)}$).

- Take '=': Only some cases in Exponential family.
- **Efficiency**: How good the estimator is.

$$e_{\hat{g}(\vec{X})}(\theta) = \frac{[g'(\theta)]^2/(nI(\theta))}{var(\hat{g}(\vec{X}))} \tag{2.54}$$

4. **Multi-Dimensional Cramer-Rao Inequality**

ReDef. Fisher Information:

$$\mathbf{I}(\theta) = \{I_{ij}(\theta)\} = \{E\left[\left(\frac{\partial \ln f(\vec{x};\theta)}{\partial\theta_i}\right)\left(\frac{\partial \ln f(\vec{x};\theta)}{\partial\theta_j}\right)\right]\} \tag{2.55}$$

Then covariance matrix $\Sigma(\theta)$ satisfies **Cramer-Rao Inequality**

$$\Sigma(\theta) \geq (n\mathbf{I}(\theta))^{-1} \tag{2.56}$$

Note: '$\geq$' holds for all diagonal elements, i.e.

$$var(\hat{\theta}_i) \geq \frac{I^*_{ii}(\theta)}{n}, \quad \forall i = 1, 2, \ldots, k \tag{2.57}$$

2.2.5 MoM and MLE in Linear Regression

Note: More detailed knowledge see sec.3 Linear Regression Analysis.

☐ Linear Regression Model(1-dimension case):

$$y_i = \beta_0 + \beta_1 x_0 + \epsilon_i \tag{2.58}$$

where β_0, β_1 are regression coefficient, and ϵ_i are unknown random **error**.

Basic Assumptions (Guass-Markov Assumption):

$$\text{Zero-Mean: } \epsilon_i \text{ are i.i.d.}$$
$$\text{Homogeneity of Variance: } E(\epsilon_i|x_i) = 0$$
$$\text{Independent: } var(\epsilon_i) = \sigma^2$$

Mission: use data $\{(x_i, y_i)\}$ to estimate β_0, β_1(i.e. regression line), and error ϵ_i.

1. OLS (Ordinary Least Squares): Take β_0, β_1 so that MSE min, i.e. SSE min

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.59}$$

(Express in Matrix Notation (eqa.2.74), so that it can be generalized to multidimensional case) SSE can be expressed as the **Excliean Distance** between $\{y_i\}$ and $\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}$, i.e.

$$\arg\min d(y, X\hat{\beta}) \tag{2.60}$$

i.e. $\hat{\beta}$ is the Projection of y onto hyperplane X, then

$$(X\hat{\beta})^T(y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty \tag{2.61}$$

Solution for 2-D case:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} - \hat{\beta}_1 \bar{x} \\ \dfrac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \end{bmatrix} \tag{2.62}$$

So get regression line:$y = \hat{\beta}_0 + \hat{\beta}_1 x$

Def. Residuals

$$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{2.63}$$

Residuals can be used to estimate ϵ_i: $E[(\epsilon_i)^2] = \sigma^2$

$$\hat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{2.64}$$

2. MoM: Consider r.v. $\epsilon \sim f(\varepsilon; x, y, \beta_0, \beta_1)$, sample $\{\epsilon_i | \epsilon_i = y_i - \beta_0 - \beta_1 x_i\}$, then obviously

$$\bar{\epsilon} = \bar{y} - \beta_0 - \beta_1 \bar{x} \tag{2.65}$$

Take moment estimate of ϵ, we have

$$E(\epsilon_i) = 0 \qquad E(\epsilon_i x_i) = 0 \text{ (note that} E(\epsilon | x) = 0) \tag{2.66}$$

$$\text{i.e.} \begin{cases} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{1}{n} \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases} \tag{2.67}$$

Solution:

$$\begin{cases} \hat{\beta}_0 & = \bar{y} - \beta_1 \bar{x} \\ \hat{\beta}_1 & = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{cases} \tag{2.68}$$

(Same as OLS)

Moment estimate of σ^2

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{2.69}$$

3. MLE: Assume $\epsilon_i \sim N(0, \sigma^2)$, then $y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Get likelihood function:

$$L(\beta_0, \beta_1, \sigma^2; x_1, \ldots, x_n, y_1, \ldots, y_n) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)}{2\sigma^2} \right] \tag{2.70}$$

Log-likelihood:

$$l(\beta_0, \beta_1, \sigma^2; x_1, \ldots, x_n, y_1, \ldots, y_n) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.71}$$

MLE, use Fermat Lemma:

$$\begin{cases} \frac{\partial l}{\partial \beta_0} = 0 & \Rightarrow -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial l}{\partial \beta_1} = 0 & \Rightarrow -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial l}{\partial \sigma^2} = 0 & \Rightarrow -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases} \tag{2.72}$$

Solution:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \end{aligned}$$

□ Linear Regression Model(Multi-dimension case):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{2.73}$$

Denote: $\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$, $\vec{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$, then for each i: $y_i = x_i^T\beta + \epsilon_i$

Further denote: Matrix form:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} = X\vec{\beta} + \vec{\epsilon} \tag{2.74}$$

Basic Assumptions: Gauss-Markov Assumptions

- OLS unbiased

$$E(\epsilon_i|x_i) = 0 \qquad E(y_i|x_i) = x_i^T\beta \tag{2.75}$$

- Homogeneity of ϵ_i

$$var(\epsilon_i) = \sigma^2 \tag{2.76}$$

- Independent of ϵ

- (For MLE) ϵ_i i.i.d. $\sim N(0, \sigma^2)$

Residuals:

$$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - x_i^T\beta \tag{2.77}$$

Def. Error Sum of Squares (SSE)

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \tag{2.78}$$

Estimator exists and unique:($\hat{\sigma}^2$ is after bias correction)

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^Ty \\ \hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\hat{\beta})^2 \\ \hat{\sigma}^2 &= \frac{1}{n-p-1}\sum_{i=1}^n (y_i - x_i^T\hat{\beta})^2\end{aligned} \tag{2.79}$$

2.2.6 Kernel Density Estimation

Given random sample $\{X_i\}$. Def. Empirical CDF.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,x]}(X_i) \tag{2.80}$$

Problem: Overfitting when getting $\hat{f}$. Solution: Using **Kernel Estimate**, replace $I_{(-\infty,x]}(\cdot)$ with Kernel function $K(\cdot)$, then

$$\hat{f}_n(x) = \frac{F_n(x+h_n) - F - n(x-h_n)}{2h_n} = \frac{1}{nh_n}\sum_{i=1}^n K(\frac{x-X_i}{h_n}) \tag{2.81}$$

where h_n is **bandwidth**. Take proper kernel function K to get estimate of f.

Can be considered as a convolution of sample $\{X_i\}$ and kernel function K.

Useful Kernel Functions:

- $K(x) = \frac{1}{2}I_{[-\frac{1}{2},\frac{1}{2}]}$
- $K(x) = (1-|x|)I_{[-1,1]}$
- $K(x) = \frac{1}{2\pi}e^{-\frac{x^2}{2}}$
- $K(x) = \frac{1}{\pi(1+x^2)}$
- $K(x) = \frac{1}{2\pi}\text{sinc}^2(\frac{x}{2})$

Section 2.3 Interval Estimation

$$\text{Parameter Estimation}\begin{cases}\text{Point Estimation} \\ \text{Interval Estimation} \quad \surd\end{cases} \tag{2.82}$$

Interval Estimation: to estimate $g(\theta)$, give **two** estimators $\hat{g}_1(\vec{X})$, $\hat{g}_2(\vec{X})$ defined on $\mathscr{X}$ as the two ends of interval (i.e. give an interval $[\hat{g}_1(\vec{X}),\ \hat{g}_2(\vec{X})]$), then random interval $[\hat{g}_1(\vec{X}),\ \hat{g}_2(\vec{X})]$ is an **Interval Estimation** of $g(\theta)$.

2.3.1 Confidence Interval

How to judge an interval estimation?

- Reliability

$$P(g(\theta) \in [\hat{g}_1, \hat{g}_2]) \tag{2.83}$$

- Precision

$$E(\hat{g}_2 - \hat{g}_1) \tag{2.84}$$

Trade off: (in most cases)

Given a level of reliability, find an interval with the highest precision above the level

□ For a given $0 < \alpha < 1$, if

$$P(\hat{g}_1 \leq g(\theta) \leq \hat{g}_2) \geq 1 - \alpha \tag{2.85}$$

then $[\hat{g}_1, \hat{g}_2]$ is a **Confidence Interval** for $g(\theta)$, with **Confidence Level** $1 - \alpha$.

Confidence Coefficient:

$$\inf_{\forall \theta \in \Theta} P(\theta \in \text{CI}) \tag{2.86}$$

Other cases:

- **Confidence Limit**:Upper/Lower Confidence Limit

$$P(g \leq \hat{g}_U) \geq 1 - \alpha$$
$$P(\hat{g}_L \leq \theta) \geq 1 - \alpha$$

- **Confidence Region**: For high dimensional parameters $\vec{g} = (g_1, g_2, \ldots, g_k)$

$$P(\vec{g} \in S(\vec{X})) \geq 1 - \alpha \quad \forall \theta \in \Theta \tag{2.87}$$

Mission: Determine $\hat{g}_1, \hat{g}_2$.

2.3.2 Pivot Variable Method

Idea: Based on point estimation, construct a new variable and thus find the interval estimation.

Def. **Pivot Variable** T, satisfies:

- Expression of T contains θ (thus T is not a statistic).
- Distribution of T independent of θ.

In different cases, construct different pivot variable, usually base on sufficient statistics and transform.

Knowing a proper pivot variable $T = T(\hat{\varphi}, g(\theta)) \sim f$, ($f$ is some distribution independent of θ), $\hat{\varphi}$ is a sufficient statistic), then we can take T satisfies:

$$P(f_{1-\frac{\alpha}{2}} \leq T \leq f_{\frac{\alpha}{2}}) = 1 - \alpha \tag{2.88}$$

Construct the inverse mapping of $T = T(\hat{\varphi}, g(\theta)) \rightleftarrows g(\theta) = T^{-1}(T, \hat{\varphi})$, we get

$$P[T^{-1}(f_{1-\frac{\alpha}{2}}, \hat{\varphi}) \leq \hat{g} \leq T^{-1}(f_{\frac{\alpha}{2}}, \hat{\varphi})] = 1 - \alpha \tag{2.89}$$

Thus get a confidence interval for θ with confidence coefficient $1 - \alpha$.

2.3.3 Confidence Interval for Common Distributions

Some important properties of χ^2, t and F see section 1.8.2.

1. Single normal population: $\vec{X} = \{X_1, X_2, \dots, X_n\} \in \mathscr{X}$ i.i.d from Normal Distribution population $N(\mu, \sigma^2)$. Denote sample mean and sample variance:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad S_\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \text{(for } \mu \text{ known)} \tag{2.90}$$

Estimating μ & σ^2: construction of pivot variable under different circumstances:

Estimation	Pivot Variable	Confidence Interval
σ^2 known, estimate μ	$T = \dfrac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$	$\left[\bar{X} - \dfrac{\sigma}{\sqrt{n}} N_{\frac{\alpha}{2}}, \bar{X} + \dfrac{\sigma}{\sqrt{n}} N_{\frac{\alpha}{2}}\right]$
σ^2 unknown, estimate μ	$T = \dfrac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$	$\left[\bar{X} - \dfrac{S}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}, \bar{X} + \dfrac{S}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}\right]$
μ known, estimate σ^2	$T = \dfrac{nS_\mu^2}{\sigma^2} \sim \chi^2_n$	$\left[\dfrac{nS_\mu^2}{\chi^2_{n,\frac{\alpha}{2}}}, \dfrac{nS_\mu^2}{\chi^2_{n,1-\frac{\alpha}{2}}}\right]$
μ unknown, estimate σ^2	$T = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$	$\left[\dfrac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \dfrac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right]$

2. Double normal population: $\vec{X} = \{X_1, X_2, \dots, X_m\}$ i.i.d. from $N(\mu_1, \sigma_1^2)$; $\vec{Y} = \{Y_1, Y_2, \dots, Y_n\}$ i.i.d. from $N(\mu_2, \sigma_2^2)$

Denote sample mean, sample variance and pooled sample variance:

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{n} X_i \qquad S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \qquad S_{\mu_1}^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu_1)^2, (\mu_1 \text{ known})$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad S_{\mu_2}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_2)^2, (\mu_2 \text{ known})$$

$$S_\omega^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

Estimating $\mu_1 - \mu_2$:

When $\sigma_1^2 \neq \sigma_2^2$ unknown, estimate $\mu_1 - \mu_2$: Behrens-Fisher Problem, remain unsolved, but can deal with simplified cases.

Estimation	Pivot Variable	Confidence Interval
σ_1^2 & σ_2^2 known, estimate $\mu_1-\mu_2$	$T=\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}\sim N(0,1)$	$\left[\bar{X}-\bar{Y}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}},\ \bar{X}-\bar{Y}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}\right]$
$\sigma_1^2=\sigma_2^2$ unknown, estimate $\mu_1-\mu_2$	$T=\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_\omega\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}\sim t_{m+n-2}$	$\left[\bar{X}-\bar{Y}-S_\omega t_{m+n-2,\frac{\alpha}{2}}\sqrt{\dfrac{1}{m}+\dfrac{1}{n}},\ \bar{X}-\bar{Y}+S_\omega t_{m+n-2,\frac{\alpha}{2}}\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}\right]$
Welch's t-Interval (when m, n large enough)	$T=\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{S_X^2}{m}+\dfrac{S_Y^2}{n}}}\xrightarrow{\mathscr{L}}N(0,1)$	$\left[\bar{X}-\bar{Y}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{S_1^2}{m}+\dfrac{S_2^2}{n}},\ \bar{X}-\bar{Y}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{S_1^2}{m}+\dfrac{S_2^2}{n}}\right]$

Estimating $\dfrac{\sigma_1^2}{\sigma_2^2}$:

Estimation	Pivot Variable	Confidence Interval
μ_1,μ_2 known, estimate $\dfrac{\sigma_1^2}{\sigma_2^2}$	$T=\dfrac{S_{\mu_2}^2}{S_{\mu_1}^2}\dfrac{\sigma_1^2}{\sigma_2^2}\sim F_{n,m}$	$\left[\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}\dfrac{1}{F_{m,n,\frac{\alpha}{2}}},\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}\dfrac{1}{F_{m,n,1-\frac{\alpha}{2}}}\right]$ or $\left[\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}F_{m,n,\frac{\alpha}{2}},\dfrac{S_{\mu_1}^2}{S_{\mu_2}^2}F_{n,m,\frac{\alpha}{2}}\right]$
μ_1,μ_2 unknown, estimate $\dfrac{\sigma_1^2}{\sigma_2^2}$	$T=\dfrac{S_Y^2}{S_X^2}\dfrac{\sigma_1^2}{\sigma_2^2}\sim F_{n-1,m-1}$	$\left[\dfrac{S_X^2}{S_Y^2}\dfrac{1}{F_{m-1,n-1,\frac{\alpha}{2}}},\dfrac{S_X^2}{S_Y^2}\dfrac{1}{F_{m-1,n-1,1-\frac{\alpha}{2}}}\right]$ or $\left[\dfrac{S_X^2}{S_Y^2}\dfrac{1}{F_{m-1,n-1,\frac{\alpha}{2}}},\dfrac{S_X^2}{S_Y^2}F_{n-1,m-1,\frac{\alpha}{2}}\right]$

3. Non-normal population:

Estimation	Pivot Variable	Confidence Interval
Uniform Distribution: $\vec{X}$ i.i.d. from $U(0,\theta)$	$T=\dfrac{X_{(n)}}{\theta}\sim U(0,1)$	$\left[X_{(n)},\dfrac{X_{(n)}}{\sqrt[n]{\alpha}}\right]$
Exponential Distribution: $\vec{X}$ i.i.d. from $\epsilon(\lambda)$	$T=2n\lambda\bar{X}\sim\chi_{2n}^2$	$\left[\dfrac{\chi^2_{2n,1-\frac{\alpha}{2}}}{2n\bar{X}},\dfrac{\chi^2_{2n,\frac{\alpha}{2}}}{2n\bar{X}}\right]$
Bernoulli Distribution: $\vec{X}$ i.i.d. from $B(1,\theta)$	$T=\dfrac{\sqrt{n}(\bar{X}-\theta)}{\sqrt{\bar{X}(1-\bar{X})}}\xrightarrow{\mathscr{L}}N(0,1)$	$\left[\bar{X}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}(1-\bar{X})}{n}},\bar{X}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}(1-\bar{X})}{n}}\right]$
Poisson Distribution: $\vec{X}$ i.i.d. from $P(\lambda)$	$T=\dfrac{\sqrt{n}(\bar{X}-\lambda)}{\sqrt{\bar{X}}}\xrightarrow{\mathscr{L}}N(0,1)$	$\left[\bar{X}-N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}}{n}},\bar{X}+N_{\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}}{n}}\right]$

4. General Case: Use asymptotic normality of MLE to construct CLT for large sample. MLE of θ satisfies:

$$\sqrt{n}(\hat{\theta}^* - \theta) \xrightarrow{\mathscr{L}} N(0, \frac{1}{I(\theta)}) \tag{2.91}$$

where $\hat{\theta}^*$ is MLE of θ. Replace $\frac{1}{I(\theta)}$ by $\sigma^2(\hat{\theta}^*)$, then

$$T = \frac{\sqrt{n}(\hat{\theta}^* - \theta)}{\sigma(\hat{\theta}^*)} \xrightarrow{\mathscr{L}} N(0,1) \tag{2.92}$$

confidence interval:

$$\left[\hat{\theta}^* - \frac{N_{\frac{\alpha}{2}}}{\sqrt{n}}\sigma(\hat{\theta}^*), \hat{\theta}^* + \frac{N_{\frac{\alpha}{2}}}{\sqrt{n}}\sigma(\hat{\theta}^*)\right] \tag{2.93}$$

2.3.4 Fisher Fiducial Argument*

Idea: When sample is known, we can get '**Fiducial Probability**' of θ, thus can find an interval estimation based on fiducial distribution.(Similar to the idea of MLE)

Remark: Fiducial probability (denoted as $\tilde{P}(\theta)$) is 'probability of parameter', in the case that sample is known. **Fiducial probability is different from Probability**.

Thus get

$$\tilde{P}(\hat{g}_1 \leq g(\theta) \leq \hat{g}_2) = 1 - \alpha \tag{2.94}$$

Section 2.4 Hypothesis Testing

Hypothesis is a statement about the characteristic of population, e.g. distribution form, parameters, etc.

Mission: Use sample to test the hypothesis, i.e. judge whether population has some characteristic.

2.4.1 Basic Concepts

Parametric hypothesis testing.

For random sample $\vec{X} = (X_1, X_2, \ldots, X_n) \in \mathscr{X}$ i.i.d. from $\mathscr{F} = \{f(x;\theta); \theta \in \Theta\}$

- Null Hypothesis H_0 & Alternative Hypothesis H_1: Wonder whether a statement is true. Def. **Null Hypothesis**: $H_0 : \theta \in \Theta_0 \subset \Theta$, **a statement that we try to reject based on sample**; $H_1 : \theta \in \Theta_1 = \Theta - \Theta_0$ is **Alternative Hypothesis**.

  ☐ Note: **Cannot** exchange H_0 and H_1, because when the evidence is ambiguity, we have to accept H_0, regardless of what H_0 is. So it is **very important** to pick the proper H_0.

  Thus Hypothesis Testing:

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1 \tag{2.95}$$

- Rejection Region R & Acceptance Region R^C: Judge whether to reject H_0 from sample, Def. **Rejection Region**:

$$R \subset \mathscr{X}: \text{reject } H_0 \text{ if } \vec{X} \in R \tag{2.96}$$

  Acceptance Region: accept H_0 if $\vec{X} \in R^C$

- Test Function: Describe how to make a decision.
  - Continuous Case:

$$\varphi(\vec{X}) = \begin{cases} 1, & \vec{X} \in R \\ 0, & \vec{X} \in R^C \end{cases} \tag{2.97}$$

    i.e. $R = \{\vec{X} : \varphi(\vec{X}) = 1\}$. Where R to be determined.
  - Discrete Case: Randomized Test Function

$$\varphi(\vec{X}) = \begin{cases} 1, & \vec{X} \in R - \partial R \\ r, & \vec{X} \in \partial R \\ 0, & \vec{X} \in R^C \end{cases} \tag{2.98}$$

    Where R and r to be determined.

- Type I Error & Type II Error: Sample is random, possible to make a wrong judge.
  - Type I Error (弃真): H_0 is true but sample falls in R, thus H_0 is rejected.

$$P(\text{type I error}) = P(\vec{X} \in R|H_0) = \alpha(\theta) \tag{2.99}$$

  - Type II Error (取伪): H_0 is wrong but sample falls in R^C, thus H_0 is accepted.

$$P(\text{type II error}) = P(\vec{X} \notin R|H_1) = \beta(\theta) \tag{2.100}$$

	Judgement		
		Accept H_0	Reject H_0
Real Case	H_0	$\surd$	Type I Error
	H_1	Type II Error	$\surd$

表 1: 'Confusion Matrix'

Impossible to make probability of Type I & II Error small simultaneously, how to pick a proper test $\varphi(\vec{x})$?

☐ **Neyman-Pearson Principle**: First control $\alpha \leq \alpha_0$, then take min β.

How to determine α_0? Depend on specific problem.[1]

- p-value: probability to get larger bias than observed $\vec{x}_0$ under H_0.

  e.g. For reject region $R = \{\vec{X}|T(\vec{X}) \geq C\}$, p-value:

$$p(\vec{x}) = P[T(\vec{X}) \geq t(\vec{x}_0)|H_0] \tag{2.101}$$

  Remark: Under H_0, the probability to get a **worse** result than $\vec{x}_0$.

  Rule: Reject H_0 if $p(\vec{x}_0) \leq \alpha_0$

[1] In most cases, take $\alpha_0 = 0.05$.

- Power Function: (when H_0 is given), probability to reject H_0 by sampling.

$$\pi(\theta) = \begin{cases} P(\text{type I error}), & \theta \in \Theta_0 \\ 1 - P(\text{type II error}), & \theta \in \Theta_1 \end{cases} = \begin{cases} \alpha(\theta), & \theta \in \Theta_0 \\ 1 - \beta(\theta), & \theta \in \Theta_1 \end{cases} \tag{2.102}$$

Express as test function:

$$\pi(\theta) = E[\varphi(\vec{X})|\theta] \tag{2.103}$$

A nice test: $\pi(\theta)$ small under H_0, large under H_1 (and grows very fast at the boundary of H_0 and H_1).

☐ **General Steps of Hypothesis Testing:**

1. Propose H_0 & H_1.
2. Determine R (usually in the form of a statistic, e.g. $R = \{\vec{X} : T(\vec{X}) \geq c\}$).
3. Select a proper α (to determine c).
4. Sampling, get sample (as well as $t(\vec{x})$), then
   - compare with R and determine whether to reject/accept H_0, or
   - calculate p-value and determine whether to reject/acceptH_0

2.4.2 Hypothesis Testing of Common Distributions

For some common distribution populations, determine rejection region R under certain H_0 with confidence coefficient α.

Definition of necessary statistics see section 2.3.3.

1. Single normal population:

Condition	H_0	H_1	Testing Statistic T	Rejection Region R
σ^2 known, test μ	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim N(0,1)$	$\|T\| > N_{\frac{\alpha}{2}}$
	$\mu \leq \mu_0$	$\mu > \mu_0$		$T > N_\alpha$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$T < -N_\alpha$
σ^2 unknown, test μ	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \sim t_{n-1}$	$\|T\| > t_{n-1,\frac{\alpha}{2}}$
	$\mu \leq \mu_0$	$\mu > \mu_0$		$T > t_{n-1,\alpha}$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$T < -t_{n-1,\alpha}$
μ known, test σ^2	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$T = \frac{nS_\mu^2}{\sigma_0^2} \sim \chi_n^2$	$T < \chi^2_{n,1-\frac{\alpha}{2}} \cup T > \chi^2_{n,\frac{\alpha}{2}}$
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$T > \chi^2_{n,\alpha}$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$T < \chi^2_{n,1-\alpha}$
μ unknown, test σ^2	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$T = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$T < \chi^2_{n-1,1-\frac{\alpha}{2}} \cup T > \chi^2_{n-1,\frac{\alpha}{2}}$
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$T > \chi^2_{n-1,\alpha}$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$T < \chi^2_{n-1,1-\alpha}$

2. Double normal population:

Condition	H_0	H_1	Testing Statistic T	Rejection Region R
σ_1^2, σ_2^2 known, test $\mu_1 - \mu_2$	$\mu_1-\mu_2=\mu_0$ $\mu_1-\mu_2\le\mu_0$ $\mu_1-\mu_2\ge\mu_0$	$\mu_1-\mu_2\ne\mu_0$ $\mu_1-\mu_2>\mu_0$ $\mu_1-\mu_2<\mu_0$	$T=\dfrac{\bar{X}-\bar{Y}-\mu_0}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}\sim N(0,1)$	$\lvert T\rvert>N_{\frac{\alpha}{2}}$ $T>N_\alpha$ $T<-N_\alpha$
σ_1^2, σ_2^2 unknown, test $\mu_1 - \mu_2$	$\mu_1-\mu_2=\mu_0$ $\mu_1-\mu_2\le\mu_0$ $\mu_1-\mu_2\ge\mu_0$	$\mu_1-\mu_2\ne\mu_0$ $\mu_1-\mu_2>\mu_0$ $\mu_1-\mu_2<\mu_0$	$T=\dfrac{\bar{X}-\bar{Y}-\mu_0}{S_\omega}\sqrt{\dfrac{mn}{m+n}}$ $\sim t_{m+n-2}$	$\lvert T\rvert>t_{m+n-2,\frac{\alpha}{2}}$ $T>t_{m+n-2,\alpha}$ $T<-t_{m+n-2,\alpha}$
μ_1, μ_2 known, test $\dfrac{\sigma_1^2}{\sigma_2^2}$	$\sigma_1^2=\sigma_2^2$ $\sigma_1^2\ge\sigma_2^2$ $\sigma_1^2\le\sigma_2^2$	$\sigma_1^2\ne\sigma_2^2$ $\sigma_1^2<\sigma_2^2$ $\sigma_1^2>\sigma_2^2$	$T=\dfrac{S_{\mu_2}^2}{S_{\mu_1}^2}\sim F_{n,m}$	$T<F_{n,m,1-\frac{\alpha}{2}}$ $\cup T>F_{n,m,\frac{\alpha}{2}}$ $T>F_{n,m,\alpha}$ $T<F_{n,m,1-\alpha}$
μ_1, μ_2 unknown, test $\dfrac{\sigma_1^2}{\sigma_2^2}$	$\sigma_1^2=\sigma_2^2$ $\sigma_1^2\ge\sigma_2^2$ $\sigma_1^2\le\sigma_2^2$	$\sigma_1^2\ne\sigma_2^2$ $\sigma_1^2<\sigma_2^2$ $\sigma_1^2>\sigma_2^2$	$T=\dfrac{S_2^2}{S_2^2}\sim F_{n-1,m-1}$	$T<F_{n-1,m-1,1-\frac{\alpha}{2}}$ $\cup T>F_{n-1,m-1,\frac{\alpha}{2}}$ $T>F_{n-1,m-1,\alpha}$ $T<F_{n-1,m-1,1-\alpha}$

3. None normal population:

Condition	H_0	H_1	Testing Statistic T	Rejection Region R
$\vec{X}$ from $B(1,p)$, test p	$p=p_0$	$p\ne p_0$	$T=\dfrac{\sqrt{n}(\bar{X}-p_0)}{\sqrt{p_0(1-p_0)}}\xrightarrow{\mathscr{L}}N(0,1)$	$\lvert T\rvert>N_{\frac{\alpha}{2}}$
$\vec{X}$ from $P(\lambda)$, test λ	$\lambda=\lambda_0$	$\lambda\ne\lambda_0$	$T=\dfrac{\sqrt{n}(\bar{X}-\lambda_0)}{\sqrt{\lambda_0}}\xrightarrow{\mathscr{L}}N(0,1)$	$\lvert T\rvert>N_{\frac{\alpha}{2}}$

2.4.3 Likelihood Ratio Test

Idea: To test $H_0:\theta\in\Theta_0 \longleftrightarrow H_1:\theta\in\Theta_1$ known $\vec{x}$, examine the likelihood function $L(\theta;\vec{x})$ and **compare** $L_{\theta\in\Theta_0}$ and $L_{\theta\in\Theta}$ to see the likelihood that H_0 is true.

Def. **Likelihood Ratio** (LR):

$$\Lambda(\vec{x})=\frac{\sup\limits_{\theta\in\Theta_0}L(\theta;\vec{x})}{\sup\limits_{\theta\in\Theta}L(\theta;\vec{x})} \tag{2.104}$$

Reject H_0 if $\Lambda(\vec{x})<\Lambda_0$. Or equivalently: Reject H_0 if $-2\ln\Lambda(\vec{x})>C(=-2\ln\Lambda_0)$.

where Λ_0 (or equivalently $C=-2\ln\Lambda_0$) satisfies:

$$E_{\Theta_0}[\varphi(\vec{X})]\le\alpha,\quad \forall\theta\in\Theta_0 \tag{2.105}$$

LR and sufficient statistic: $\Lambda(\vec{x})$ can be expressed as $\Lambda(\vec{x})=\Lambda^*(T(\vec{x}))$, where $T(\vec{X})$ is sufficient statistic.

☐ LRT for one-sample t-test: For $X_1, X_2, \dots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$, test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0 \quad \text{when } \sigma^2 \text{ unknown}$$

Can prove:

$$\Lambda^{2/n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2}$$

Denote $T = \dfrac{\sqrt{n}(\bar{x} - \mu_0)}{S}$, then LRT is

$$\Lambda = \left(1 + \frac{T^2}{n-1}\right)^{-n/2}$$

The Multivariate case see sec. 4.4, where T^2 itself is the Hotelling's T^2 statistic.

☐ Limiting Distribution of LR: Wilks' Thm.

If $\dim \Theta = k > \dim \operatorname{span}\{\Theta_0\} = s$[2], then under $H_0 : \theta \in \Theta_0$:

$$\Lambda_{\theta \in \Theta_0}(\vec{x}) = -2 \ln \lambda(\vec{x}) \xrightarrow{\mathscr{L}} \chi^2_{k-s} \tag{2.106}$$

2.4.4 Uniformly Most Powerful Test

Idea: Neyman-Pearson Principle: control α, find min β. i.e. control α, find max $\pi(\theta)$

Def. **Uniformly Most Powerful Test** (UMP) φ_{UMP} with level of significance α satisfies

$$\pi_{\text{UMP}}(\theta) \geq \pi(\theta), \ \forall \theta \in \Theta_1 \tag{2.107}$$

Neyman-Pearson Lemma: For $\vec{X} = (X_1, X_2, \dots, X_n)$ i.i.d. from $f(\vec{x}; \theta)$.

Test hypothesis $H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta = \theta_1$. Def. test function φ as:

$$\varphi(\vec{x}) = \begin{cases} 1, & \dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} > C \\ r, & \dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} = C \\ 0, & \dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} < C \end{cases} \tag{2.108}$$

Then there exists C and r such that

- $E[\varphi(\vec{x})|\theta_0] = P(\dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} > C) + rP(\dfrac{f(\vec{x};\theta_1)}{f(\vec{x};\theta_0)} = C) = \alpha$
- This φ is UMP of level of significance α

Actually kind of 1-dimensional case of LRT.

Note: UMT exist for **simple** H_0, H_1, otherwise may not exist.

UMP and sufficient statistics: Test function $\varphi(\vec{X}$ given by eqa.2.108 is function of sufficient statistics $T(\vec{X})$, i.e. $\varphi(\vec{X}) = \varphi^*(T(\vec{X}))$.

[2]Here 'dimension' refers to 'degree of freedom'.

UMP and Exponential Family: For sample $\vec{X} = (X_1, X_2, \ldots, X_n)$ from exponential family:

$$f(\vec{x};\theta) = C(\theta)h(\vec{x})\exp\{Q(\theta)T(\vec{x})\} \tag{2.109}$$

Test single hypothesis $H_0: \theta = \theta_0 \longleftrightarrow H_1: \theta = \theta_1$, (where $\theta_0 < \theta_1$). If

- θ_0 is inner point of Θ
- $Q(\theta)$ monotone increase with θ

Then UMP exists, in the form of:

$$\varphi(\vec{x}) = \begin{cases} 1, & T(\vec{x}) > C \\ r, & T(\vec{x}) = C \\ 0, & T(\vec{x}) < C \end{cases} \tag{2.110}$$

where C and r satisfies $E[\varphi(\vec{x})|\theta_0] = \alpha$.

Note: or take $Q(\theta)$ mono decreased, then in eqa.2.110, take opposite inequality operators.

□ **General Steps of UMP**:

1. Find a point $\theta_0 \in \Theta_0$ and a point $\theta_1 \in \Theta_1$. (Note: **one** point)
2. Construct test function in the form of eqa.2.108, use $E[\varphi(\vec{x})|\theta_0] = \alpha$ to determine C and r.
3. Get R and $\varphi(\vec{x})$.
4. If φ does **not** depend on θ_1, then H_1 can be generalized to $H_1: \theta \in \Theta_1$.
5. If φ satisfies $E_{\theta\in\Theta_0}(\varphi) \leq \alpha$, then H_0 an be generalized to $H_0: \theta \in \Theta_0$.

2.4.5 Duality of Hypothesis Testing and Interval Estimation

- Thm.: $\forall\theta_0 \in \Theta$ there exists hypothesis testing $H_0: \theta = \theta_0 \longleftrightarrow H_1: \theta \neq \theta_0$ of level α with rejection region R_{θ_0}. Then

$$C(\vec{X}) = \{\theta : \vec{X} \in R_\theta^C\} \tag{2.111}$$

  is a $1-\alpha$ confidence region for θ

- Thm.: $C(\vec{X})$ is a $1-\alpha$ confidence region for θ. Then $\forall\theta_0 \in C(\vec{X})$, the rejection region of hypothesis testing $H_0: \theta = \theta_0 \longleftrightarrow H_1: \theta \neq \theta_0$ of level α satisfies

$$R_{\theta_0}^C = \{\vec{X} : \theta_0 \in C(\vec{X})\} \tag{2.112}$$

□ Idea:

$$H_0: \theta = \theta_0 \longleftrightarrow H_1: \theta \neq \theta_0$$

$$\updownarrow \tag{2.113}$$

$$P(R^C(\vec{X})|H_0) = P(R^C(\vec{X})|\theta_0) = 1-\alpha$$

$$\updownarrow \tag{2.114}$$

$$\text{Confidence Interval: } \theta_0 \in R^C(\vec{X})$$

Similar for Confidence Limit and One-Sided Testing.

2.4.6 Introduction to Non-Parametric Hypothesis Testing

Motivation: Usually distribution form unknown, cannot use parametric hypothesis testing.

Useful Method:

- Sign Test: Used for paired comparison $\vec{X} = (X_1, X_2, \dots, X_n, \vec{Y} = (Y_1, Y_2, \dots, Y_n)$.

  Take $Z_i = Y_i - X_i$ i.i.d., denote $E(Z) = \mu$. Test $H_0 : \mu = 0 \longleftrightarrow H_1 : \mu \neq 0$.

  Denote $n_+ = \#(\text{positive } Z_i)$ and $n_- = \#(\text{negative } Z_i)$, $n_0 = n_+ + n_-$. Then $n_+ \sim B(n_0, \theta)$, test $H_0 : \theta = \frac{1}{2} \longleftrightarrow H_1 : \theta \neq \frac{1}{2}$

  Then use Binomial Testing or large sample CLT Normal Testing.

  Remark:

  - Also can test $H_0 : \theta \leq \frac{1}{2} \longleftrightarrow H_1 : \theta > \frac{1}{2}$
  - Drawback: ignores magnitudes.

- Wilcoxon Signed Rank Sum Test: Improvement of Sign Test. Base on order statistics.

  Order Statistics of Z_i: $Z_{(1)} < Z_{(2)} < \dots < Z_{(n)}$, where each $Z_{(j)}$ corresponds to some Z_i, denote as $Z_i = Z_{(R_i)}$, then R_i is the rank of Z_i.[3]

  Def. $\vec{R} = (R_1, R_2, \dots, R_n)$ is **Rank Statistics** of $(Z_1, Z_2, \dots, Z_n)$

  Def. **Sum of Wilcoxon Signed Rank**:

  $$W^+ = \sum_{i=1}^{n_0} R_i I_{Z_i > 0} \tag{2.115}$$

  Distribution of W^+ is complex. E and var of W^+ under H_0:

  $$E(W^+) = \frac{n_0(n_0+1)}{4} \qquad var(W^+) = \frac{n_0(n_0+1)(2n_0+1)}{24} \tag{2.116}$$

  Usually consider large sample CLT, construct normal approximation:

  $$T = \frac{W^+ - E(W^+)}{\sqrt{var(W^+)}} \xrightarrow{\mathscr{L}} N(0,1) \tag{2.117}$$

  Rejection Region: $R = \{|T| > N_{\frac{\alpha}{2}}\}$

- Wilcoxon Two-Sample Rank Sum Test: Used for two independent sample comparison.

  Assume $\vec{X} = (X_1, \dots, X_m)$ i.i.d. $\sim f(x)$; $\vec{Y} = (Y_1, \dots, Y_n)$ i.i.d. $\sim f(x - \theta)$, test $H_0 : \theta = 0 \longleftrightarrow H_1 : \theta \neq 0$.

  Rank X_i and Y_i as:

  $$Z_1 \leq Z_2 \leq \dots \leq Z_{m+n} \tag{2.118}$$

[3] If some $X_i, X_j, \dots$ equal, then take same rank $R = \text{mean}\{R_i, R_j, \dots\}$.

in which denote rank of Y_i as R_i, and def. **Wilcoxon two-sample rank sum**:

$$W = \sum_{i=1}^{n} R_i \tag{2.119}$$

E and var of W under H_0:

$$E(W) = \frac{n(m+n+1)}{2} \qquad var(W) = \frac{mn(n+m+1)}{12} \tag{2.120}$$

Use large sample approximation, construct CLT:

$$T = \frac{W - E(W)}{\sqrt{var(W)}} \xrightarrow{\mathscr{L}} N(0,1) \tag{2.121}$$

- Goodness-of-Fit Test: For $\vec{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. from some certain population X. Test $H_0 : X \sim F(x)$.

  where F is theoretical distribution, can be either parametric or non-parametric.

  Idea: Define some *quantity* $D = D(X_1, \ldots, X_n; F)$ to measure the difference between F and sample. And def. *Goodness-of-fit* when observed value of D (say d_0) is given:

$$p(d_0) = P(D \geq d_0 | H_0) \tag{2.122}$$

  Goodness-of-Fit Test: Reject H_0 if $p(d_0) < \alpha$.

  Pearson χ^2 Test: Usually used for discrete case.

  Test $H_0 : P(X_i = a_i) = p_i$, $i = 1, 2, \ldots, r$. Denote $\#(X_j = a_i) = \nu_i$, take D as:

$$K_n = K_n(X_1, \ldots, X_n; F) = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \tag{2.123}$$

  Pearson Thm.: For K_n defined as eqa.2.123, then under H_0:

$$K_n \xrightarrow{\mathscr{L}} \chi^2_{r-1-s} \tag{2.124}$$

  Here s is number of unknown parameter, $r - 1 - s$ is the degree of freedom.

  Note:

  – a_i must **not** depend on sample.

  – For continuous case, construct division:

$$\mathbb{R} \to (-\infty, a_1, a_2, \ldots, a_{r-1}, \infty = a_r) \tag{2.125}$$

  and test $H_0 : P(X \in I_j) = p_j$

  Criterion: Pick proper interval so that np_i and ν_i both ≥ 5.

- Contingency Table Independence & Homogeneity Test

- Independence Test:

  Test a two-parameter sample and to see whether these two parameters(features) are independent. Denote $Z = (X, Y)$ are some 'level' of sample, n_{ij} is number of sample with level (i, j)

  Contingency Table:

X \ Y	1	…	j	…	s	$\sum$
1	n_{11}	…	n_{1j}	…	n_{1s}	$n_{1\cdot}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
i	n_{i1}	…	n_{ij}	…	n_{is}	$n_{i\cdot}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
r	n_{r1}	…	n_{rj}	…	n_{rs}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	…	$n_{\cdot j}$	…	$n_{\cdot s}$	n

  Test $H_0 : X \,\&\, Y$ are independent. i.e. $H_0 : P(X = i, Y = j) = P(X = i)P(Y = j) = p_{i\cdot}p_{\cdot j}$.

  Construct χ^2 test statistic:

$$K_n = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{[n_{ij} - n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})]^2}{n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) \tag{2.126}$$

  Then under H_0, $K_n \xrightarrow{\mathscr{L}} \chi^2_{rs-1-(r+s-2)} = \chi^2_{(r-1)(s-1)}$

  Reject H_0 if $p(k_0) = P(K_n \geq k_0) < \alpha$

- Homogeneity Test:

  Test R groups of sample with category rank, to see whether these groups has similar rank distribution.

Group \ Category	Category 1	…	Category j	…	Category C	$\sum$
Group 1	n_{11}	…	n_{1j}	…	n_{1C}	$n_{1\cdot}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
Group i	n_{i1}	…	n_{ij}	…	n_{iC}	$n_{i\cdot}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
Group R	n_{R1}	…	n_{Rj}	…	n_{RC}	$n_{R\cdot}$
$\sum$	$n_{\cdot 1}$	…	$n_{\cdot j}$	…	$n_{\cdot C}$	n

  Denote $P(\text{Category } j|\text{Group } i) = p_{ij}$. Test $H_0 : p_{ij} = p_j, \forall 1 \leq i \leq R$.

  Construct χ^2 test statistic:

$$D = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{[n_{ij} - n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})]^2}{n(\frac{n_{i\cdot}}{n})(\frac{n_{\cdot j}}{n})} = n\left(\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) \tag{2.127}$$

  Then under H_0, $D \xrightarrow{\mathscr{L}} \chi^2_{R(C-1)-(C-1)} = \chi^2_{(R-1)(C-1)}$

- Test of Normality: normality is a good & useful assumption.

  For $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)$,

  Test H_0 : exists μ & σ^2 such that Y_i i.i.d. $\sim N(\mu, \sigma^2)$.

  – Kolmogorov-Smirnov Test: Assume $\vec{X}$ form population CDF $F(x)$, test $H_0 : F(x) = F_0(x)$(where can take $F_0 = \Phi$ or some other known CDF).

    use $F_n(x)$ (as defined in eqa.2.80) as approx. to $F(x)$, test

    $$D_n = \sum_{-\infty<x<+\infty} |F_n(x) - F_0(x)| \tag{2.128}$$

    Reject H_0 if $D_n > c$

    or use goodness-of-fit: denote observed value of D_n as d_n. Reject H_0 if

    $$p(d_n) = P(D_n > d_n | H_0) < \alpha \tag{2.129}$$

  – Shapiro-Wilk Test:

    Test H_0 : exists μ & σ^2 such that X_i i.i.d. $\sim N(\mu, \sigma^2)$.

    Denote $Y_{(i)} = \dfrac{X_{(i)} - \mu}{\sigma}$, $m_i = E(Y_{(i)})$

    Under H_0, $(X_{(i)}, m_i)$ falls close to straight line. Test Statistic: Correlation

    $$R^2 = \frac{\left(\sum_{i=1}^n (X_{(i)} - \bar{X})(m_i - \bar{m})\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (m_i - \bar{m})^2} = corr(X_{(i)}, m_i) \tag{2.130}$$

    Reject H_0 if $R^2 < c$

    Shapiro-Wilk correction:

    $$W = \frac{\left(\sum_{i=1}^{[n/2]} a_i (X_{(n+1-i)} - X_{(i)})\right)^2}{\sum_{i=1}^n (X_{(i)} - \bar{X})^2} \tag{2.131}$$

☐ Summary: Useful Non-Parameter Hypothesis Testing.

- Non-Parameter Hypothesis Testing
  - One Population Sample
    - χ^2 Test
    - Binomial Test
    - One-Sample K-S Test
    - Wilcoxon Sign Test
    - Runs Test
  - Two Population Sample
    - Independent Sample
      - Mann-Whitney Test
      - K-S Test
      - Wald-Wolfowitz Test
      - Moses Test of Extreme Reactions
    - Relative Sample
      - Sign Test
      - McNemar Test
      - Wilcoxon Rank Sum Test
      - Marginal Homogeneity Test
  - Multi-Population Sample
    - Independent Sample
      - Median Test
      - K-W One-Way ANOVA Test
      - Jonckheere-Terpstra Test
    - Relative Sample
      - Friedman Rank Sum Test
      - Kendall’s Coefficient of Concordance Test
      - Cochran Q Test

Chapter. III 线性回归分析部分

Instructor: Zaiying Zhou

☐ Steps in Regression Analysis

1. Statement of the problem;
2. Selection of potentially relevant **variables**;
3. Data collection;
4. Exploratory Data Analysis (**EDA**)
5. **Model** specification;
6. Choice of fitting method;
7. Model fitting;
8. Model validation and criticism;
9. Using the chosen model(s) for the solution of the posed problem;
10. **Explain** the result.

R. Code for EDA

```
libaray('GGally')
head(df)
ggpairs(df)
str(df)
summary(df)
```

☐ Used Packages in R.

```
library('ggplot2')
libaray('GGally')
library('car')
library('moments')
library('lmtest')
library('nortest')
library('MASS')
library('tseries')

source('package.r')
```

Section 3.1 Linear Regression Model

3.1.1 Data and Model for Simple Linear Regression

We will observe pairs of variables, called 'cases'(样本点). A sample is $(X_1, Y_1), \ldots, (X_n, Y_n)$

▷ R. **Code**

Example data import:

```
df <- read.table('dataset/CH01PR27.txt',header=FALSE,
    sep=',',col.names = c('y','x'))
```

Linear Model: [4] [5]

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3.1}$$

with Guass-Markov Assumption:

$$\begin{aligned} \text{Zero-Mean: } & E(\epsilon_i|X_i) = 0 \\ \text{Homogeneity of Variance: } & var(\epsilon_i) = \sigma^2 \\ \text{Independent: } & \epsilon_i \text{ i.i.d. } \sim \varepsilon \end{aligned} \tag{3.2}$$

Normal Error Assumption: Further in most cases, we consider $\varepsilon \sim N(0, \sigma^2)$ —-because of its well-property distribution, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ i.i.d. $N(0, \sigma^2)$.[6]

What does Linear Regression do? Under Linear Model, try to estimate

- β_0 (intercept) ;
- β_1 (slope) ;
- σ^2 (variance of error).

(Thus Linear Regression is also a Statistics Inference process: deduce properties of model from data)

3.1.2 The Ordinary Least Square Estimation

Aim: use (x_i, y_i) to estimate $\beta_0, \beta_1, \sigma^2$. The idea is to define a 'loss function' to reflect the 'distance' from sample point to estimation point.

Estimate Principle: [7]

---

[4] Here in linear regression, we consider X_i only as real number, **without** randomness. So here Y_i can be considered as an r.v. with X_i as parameter, i.e. $Y_i|_{X_i=x_i}$

[5] Note: Why we need ε as 'random error term'?

- It represents the intrinsic random property of the model.
- Based on ε, we can take r.v. into our statistic model.

[6] i.e. Y_i are independent

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \quad i = 1, 2, \ldots, n \tag{3.3}$$

[7] Detailed Definition and Derivation see sec.2.2.5.

- Ordinary Least Squares:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^{n} (y - \beta_0 - \beta_1 x_i)^2 \tag{3.4}$$

- MLE or MoM Estimation.

And get $\hat{\beta}_1, \hat{\beta}_0$ as well as $\hat{\sigma^2}$(see eqa(3.9):[8]

$$\begin{aligned}
\hat{\beta}_1 =& \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta}_0 =& \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\sigma^2} =& \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned} \tag{3.6}$$

Def. **Residual**: distance from sample point to estimate point, to reflect how the sample points fit the model.

$$e_i = y_i - \hat{y}_i = \text{observed value of } \varepsilon_i \tag{3.7}$$

Note: under least square estimation, we have[9]

$$\sum e_i = 0 \qquad \sum x_i e_i = 0 \tag{3.8}$$

Then use e_i to estimate σ^2 (because it is ε_0 that are i.i.d., not Y_i), where $(n-p-1)$ is Degree of Freedom (df or dof)[10]

$$\begin{aligned}
\hat{\sigma^2_n} &= \frac{1}{n} \sum e_i^2 \quad \text{(use MLE or MoM)} \\
\hat{\sigma^2} &= \frac{1}{n-p-1} \sum e_i^2 = \frac{1}{n-2} \sum e_i^2 \quad \text{(use OLS, unbiased)}
\end{aligned} \tag{3.9}$$

Degree of Freedom of a Quadric Form:

- Intuitively: the number of independent variable;
- Rigorously: for quadric SS $= x'Ax$:

$$dof_{SS} = \text{rank}(A) \tag{3.10}$$

▷ R. **Code**

```r
lmfit <- lm(formula,df)
summary(lmfit,cor=TRUE)
ggcoef(lmfit)
```

[8] A memory trick: use $\frac{Y}{\sqrt{s_Y}} = r_{XY} \frac{X}{\sqrt{s_X}}$ to get formular of $Y \sim X$:

$$\hat{\beta}_1 = r_{XY} \frac{\sqrt{s_Y}}{\sqrt{s_X}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{3.5}$$

[9] Intuitively, they each means '$E(\varepsilon) = 0$' and '$X \parallel \varepsilon$'.

[10] Generally, MLE and LSE are different.

Comment from R.A.Fisher: $\sum e_i^2$ should be divided by 'number of e_i^2 that contribute to variance'. Here $(n-p-1)$ corresponds to 'degree of freedom' $= (n-2)$, $p = 1$ corresponds to 'one' variable (see sec.2.2.5, eqa(2.79)), and correponds to the two equations of e_i, eqa(3.8) `lmfit` includes parameters `lmfit$coefficient` and `lmfit$residuals`

Example `lm()` output:

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-16.1368  -6.1968  -0.5969   6.7607  23.4731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 156.3466     5.5123   28.36   <2e-16 ***
x            -1.1900     0.0902  -13.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.173 on 58 degrees of freedom
Multiple R-squared:  0.7501,    Adjusted R-squared:  0.7458
F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

3.1.3 Statistical Inference to β_0,β_1,e_i

□ Sampling Distribution of $\hat{\beta}_1, \hat{\beta}_0$

Consider $\hat{\beta}_1, \hat{\beta}_0$ as statistics of sample, then we can examine the sampling distribution of $\hat{\beta}_1, \hat{\beta}_0$. Their randomness comes from

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3.11}$$

(The following part treats $\hat{\beta}_1, \hat{\beta}_0$ as r.v., and note that X_i are **not** r.v.. And for convenience and conciseness, denote $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$)

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{XX}} \varepsilon_i$$
$$\hat{\beta}_0 = \beta_0 + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) \varepsilon_i$$

Denote corresponding variance as $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_0}$, using eqa(1.69) to get:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{XX}} \qquad \sigma^2_{\hat{\beta}_0} = \sigma^2 (\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}) \tag{3.12}$$

And under normal error assumption, distribution of $\hat{\beta}_1, \hat{\beta}_0$ are

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1}) = N(\beta_1, \frac{\sigma^2}{S_{XX}})$$
$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}) = N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}))$$

Based on sampling distribution of $\hat{\beta}_1, \hat{\beta}_0$, we can conduct statistical inference, including CI and HT.[11]

Note: In linear regression model, we usually focus more on β_1. And note that when 0 is **not** within the fitting range,β_0 is not so important.[12]

□ Sampling Distribution of e_i Consider e_i as r.v. satisfies

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \tag{3.13}$$

and get the expression of $\hat{e}_i$

$$\hat{e}_i = \varepsilon_i - \sum_{k=1}^{n} \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \right) \varepsilon_k \tag{3.14}$$

$$E(e_i) = 0 \qquad \sigma^2_{e_i} = \sigma^2 \left(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}} \right) \tag{3.15}$$

Under normal assumption:

$$e_i \sim N(0, \sigma^2 \left(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}} \right)) \tag{3.16}$$

Further we can get $\hat{\sigma}^2 = E(\frac{1}{n-2} \sum_{i=1}^n e_i^2)$ where $e_i^2 \sim \sigma^2 \left(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}} \right) \chi^2$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sigma^2 \sum_{i=1}^{n} (1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}) = \sigma^2 \tag{3.17}$$

More definition of refined residuals see sec.3.3.1 in page 3.3.1.

□ Why we choose OLS to get regression coefficients?

Gauss-Markov Thm.: the OLS estimator has the lowest sampling variance within the class of linear unbiased estimators, i.e. OLS is the Best Linear Unbiased Estimator(BLUE).[13]

3.1.4 Prediction to Y_h

For a new X_h at which we wish to **predict** the corresponding Y_h (based on other known point (X_i, Y_i)), denote the estimator as $\hat{\mu}_h$:

$$\hat{\mu}_h = \hat{\beta}_1 X_h + \hat{\beta}_0 = \beta_1 X_h + \beta_0 + \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(X_i - \bar{X})(X_h - \bar{X})}{S_{XX}} \right) \varepsilon_i \tag{3.18}$$

[11] Detail see sec.2.4, estimating/testing $\hat{\beta}_1, \hat{\beta}_0$ usually corresponds to 'estimate μ, with σ^2 unknown'.

[12] Two reason:

- The etimation error of Y from $\hat{\beta}_1$ increases with $X_h - \bar{X}$;
- $\beta_1 == 0$ is important: decides whether linear model can be used.

[13] This Thm. does **not** require normal error assumption.

Thus we can get[14]

$$E(\hat{\mu}_h) = \beta_1 X_h + \beta_0 \qquad \sigma^2_{\hat{\mu}_h} = \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\right)\sigma^2 \tag{3.19}$$

Under Normal assumption:

$$\hat{\mu}_h \sim N(\beta_1 X_h + \beta_0, \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}\right)\sigma^2) \tag{3.20}$$

Base on distribution we can give CI and HT.

Note: We can either consider

- Y_h **itself as an r.v.** : Confidence Interval of Y_h;

  And we can just use $\sigma^2_{\hat{\mu}_h}$ to construct CI;

  ▷ R. **Code**

```r
predict(lmfit,data.frame(x=c(df$x,40)),
    interval="confidence",level=0.95)
```

- **predicted** Y_h **from other sample points**: Prediction Interval of Y_h

  Need to have the randomness of $\hat{\beta}_0, \hat{\beta}_1$ considered(if they are unknown).

  Def. Prediction Error: Y_h itself is an Y of the linear model, i.e. $Y_i = \beta_0 + \beta_1 X_h + \varepsilon_h$, we can and define **Prediction Error**:

$$d_h = Y_h - \hat{\mu}_h \tag{3.21}$$

$$E(d_h) = 0 \qquad \sigma^2_{d_h} = var(Y_h - \hat{\mu}_h) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})}{S_{XX}}\right] > \sigma^2_{\hat{\mu}_h} \tag{3.22}$$

  ▷ R. **Code**

```r
predict(lmfit,data.frame(x=c(df$x,40)),
    interval="prediction",level=0.95)
```

☐ Simultaneous Confidence Band(SCB)

Confidence Band is **not** the CI at each point, but really a **band** for the **entire** regression line.[15]

Aim: Find lower and upper function $L(x)$ and $U(x)$ such that

$$P[L(x) < (\beta_0 + \beta_1 x) < U(x), \forall x \in I_x] = 1 - \alpha \tag{3.23}$$

and get **Confidence Band**:

$$\{(x, y) | L(x) < y < U(x) | \forall x \in I_x\} \tag{3.24}$$

Where $(L(x), U(x))$ can be derived as

$$(L(x), U(x)) = \hat{\mu}_x \pm s_{\hat{\mu}_x} W_{2,n-2,1-\alpha} \tag{3.25}$$

[14] So $\sigma^2(\hat{\mu}_h)$ increases with $X_h - \bar{X}$. Intuitively it make sense, because $(\bar{X}, \bar{Y})$ must falls on regression line.

[15] Why they are different? We require the confidence band have a **simultaneous** converage probability. For the same band $(L(x), U(x))$, $P(\text{the whole line}) < P(\text{each point})$, so Confidence Band is wider than $\bigcup$CIs to hold the same $1 - \alpha$.

Also, we will see that for linear model, the boundary of SCB forms hyperbola, which make sense considering its asymptotic line.

Where W correponds to W distribution: $W_{m,n} = \sqrt{2F_{m,n}}$

Small sample case: Bonferroni correction.

▷ R. Code

```
library(ggplot2)
ggplot(df,aes(x,y))+geom_point()+geom_smooth(method='lm',formula=y~x)
```

Section 3.2 Analysis of Variance

ANalysis **O**f **VA**riance (ANOVA): One-sample t test $\rightsquigarrow$ Two sample t test $\rightsquigarrow$ ANOVA

☐ **Key Point Of ANOVA**: Take Partition of Total Sum of Square To Examine **Variation**.

Because Y_i are not i.i.d. (different mean), so has different parts of variation from Regression Model/Error Term.

3.2.1 Monovariate ANOVA

Measure of Variation: Sum of Square (SS) & Mean Sum of Square (MS).

MS: Divide each SS by corresponding dof. Definition of dof see eqa(3.10).

$$\text{MS} = \frac{\text{SS}}{dof} \tag{3.26}$$

- SST: Total Sum of Squares

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad dof_{\text{SST}} = n - 1 \tag{3.27}$$

- SSRegression: Variation due to Regression Model (which is explained by regression line);[16]

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad dof_{\text{SSR}} = 1 \tag{3.28}$$

- SSError: Variation attribtes to ε (which is reflected by residuals).

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) \qquad dof_{\text{SSE}} = n - 2 \tag{3.29}$$

△ **IMPORTANT:** In some books

- SSRegression → SSExplained of SSModel;
- SSError → SSResidual.

And Cause **Confusion**! In this summary we take the former.

Idea: take partition of SST. i.e.

$$Y_i - \bar{Y} = (Y_i - \hat{Y}) + (\hat{Y} - \bar{Y}) = e_i \tag{3.30}$$

And we can prove that

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \text{SSR} + \text{SSE} \tag{3.31}$$

[16] $\text{SSR} = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$, so $dof_R = 1$

That is: we **partition** SST into two parts, so that we can examine them seperately.

Properties:

$$E(\text{MSE}) = \sigma^2 \qquad E(\text{MSR}) = \sigma^2 + \beta_1^2 S_{XX} \tag{3.32}$$

3.2.2 Multivariate ANOVA

Sampling Notation see eqa(3.81), still consider $p+1$ -dim $(\mathbf{1}_n, X_i)$ v.s. 1-dim Y, and $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$

- SST:

$$\text{SST} = (Y - \bar{Y}\mathbf{1}_n)'(Y - \bar{Y}\mathbf{1}_n) \qquad dof_{\text{SST}} = n-1 \tag{3.33}$$

- SSR:

$$\text{SSR} = (\hat{Y} - \bar{Y}\mathbf{1}_n)'(\hat{Y} - \bar{Y}\mathbf{1}_n) \qquad dof_{\text{SSR}} = p \tag{3.34}$$

  Denoted in hat matrix H and $\mathcal{J}$ in eqa(4.9)

$$\text{SSM} = Y'(H - \frac{1}{n}\mathcal{J})Y \tag{3.35}$$

- SSE:

$$\text{SSE} = (Y - \hat{Y})'(Y - \hat{Y}) \qquad dof_{\text{SSE}} = n-p-1 \tag{3.36}$$

  Denoted in residual e and hat matrix H:

$$\text{SSE} = e'e = Y'(I - H)Y \tag{3.37}$$

3.2.3 ANOVA Table

Source	dof	SS	MS	F-Statistic
SSRegression	p	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	SSR$/dof_R$	MSR/MSE
SSError	$n-p-1$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	SSE$/dof_E$	
SSTotal	$n-1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	SST$/dof_T$	

▷ `R.` **Code**

```
anova(lmfit)
```

3.2.4 Hypotheses Testing to Slope

Main focus: whether the linear relation exist:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \longleftrightarrow H_1 : \exists \beta_i \neq 0,\ i = 1, 2, \ldots, p \tag{3.38}$$

- ANOVA F-Test:

  We can examine
  $$F = \frac{\text{MSR}}{\text{MSE}} \sim F_{p,n-p-1}$$

- General Linear Test (GLT)

  First we introduce the examine models:

  – Full model:
  $$Y_i = X_i'\beta + \varepsilon_i = \beta_0 + \sum_{j=1}^{n} X_{ij}\beta_j + \varepsilon_i$$
  And define SSE_F with $dof_\text{F} = n - p - 1$ under Full Model.

  – Reduced model:
  $$Y_i = \beta_0 + \varepsilon_i$$
  And define SSE_R with $dof_\text{R} = n - 1$ under Reduced Model.

  and examine
  $$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{F})/(dof_\text{R} - dof_\text{F})}{\text{SSE}_\text{F}/dof_F} \sim F_{p,n-p-1} \tag{3.39}$$

  ▷ R. **Code**

```
nullmodel <- lm(y ~ 1,df)
anova(nullmodel,lmfit)
```

- Pearson Correlation Coefficient r and Coefficient of Multiple Determination R^2:

  Pearson's r:
  $$r = c\hat{o}v(Y, \hat{Y}) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

  CMD R^2:
  $$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

  Adjusted R^2:
  $$R_\text{a}^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{n-1}{n-p}\frac{\text{SSE}}{\text{SST}}$$

  – Relation between r and R^2: Under Simple Linear Model, we have
  $$R^2 = r^2$$

  – Relation between R^2 and F-Statistic:
  $$F = \frac{R^2}{1 - R^2}\frac{n-p}{n-1}$$

Section 3.3 Model Assumption, Diagnostics and Remedies

To apply OLS, we need the basic Gauss-Markov Assumption eqa(3.2); or we further need better properties of the model, so need Normal Assumption.

Assumptions:

$$\begin{aligned} \text{Zero-Mean: } & E(\epsilon_i|X_i) = 0 \\ \text{Homogeneity of Variance: } & var(\epsilon_i) = \sigma^2 \\ \text{Independent: } & \epsilon_i \text{ i.i.d. } \sim \varepsilon \\ \text{Normal: } & Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \end{aligned} \tag{3.40}$$

Or sum up as

$$\vec{\varepsilon} \sim N_n(\vec{0}, \sigma^2 I_n) \tag{3.41}$$

Thus we need to conduct Diagnostics and Remedies to

- examine whether these assumptions are satisfies;
- perform correction to regression method.

3.3.1 Diagnostics

Preliminary Diagnostics: ▷ R. **Code**

```
lmfit <- lm(y~x,lmfit)
par(mfrow = c(2, 2))
plot(lmfit)
```

□ Diagnostics to X

Considering the dependence of Y_i on X_i, to get a more reliable $\hat{\beta}_1$, we cannot just focus on the (marginal) distribution of Y_i, we would also need a better 'distribution' of X_i

- 4 statistics(parameters);[17]
  - Mean: Location;
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.42}$$
  - Standard Deviation: Variability;
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.43}$$
  - Skewness: Lack of Symmertry;
$$\hat{g}_1 = \frac{m_{n,3}}{m_{n,2}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})\right)^{3/2}} \tag{3.44}$$

    Adjusted Skewness (Least MSE):
$$\frac{\sqrt{n(n-1)}}{n-2}\hat{g}_1 \tag{3.45}$$

[17] See sec.2.1.1

* $\hat{g}_1 > 0$: Right skewness, longer right tail;
* $\hat{g}_1 < 0$: Left skewness, longer left tail.

Fisher-Pearson coefficient of skewness.

– Kurtosis: Heavy/Light Tailed.

$$\hat{g}_2 = \frac{m_{n,4}}{m_{n,2}^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} - 3 \tag{3.46}$$

$\hat{g}_2 = 0 \Rightarrow$ similar to normal.

* $\hat{g}_2 > 0$: Leptokurtic, heavy tail, slender;
* $\hat{g}_2 < 0$: Platykurtic, light tail, broad.

Note: In expression of $\hat{g}_1$ and $\hat{g}_2$, we already divide the variance. So Skewness and Kurtosis only reflect the difference from normal, but **not** related to variance.

Best tool to determine Kurtosis: QQ-Plot.

▷ `R.` **Code**

```
summary(df$x)
```

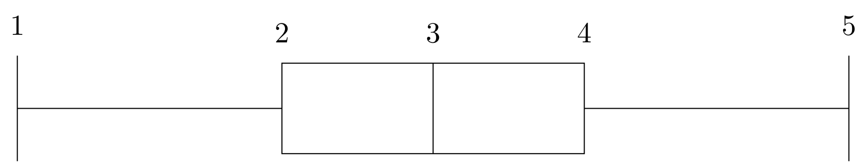
Other moments use package `moments`

• Useful Plots:

– BoxPlot: to examine the similarity of distribution.

Notation:

1. min point above 25% quantile-1.5IQR;
2. 25% quantile;
3. median;
4. 75% quantile;
5. max point below 75% quantile+1.5IQR.



– Histogram Plots: Frequency distribution (can deal with many-peak)

– Quantile-Quantile Plots: Examine the similarity between distribution.

For two CDF $q = F(x)$ and $q = G(x)$(where q for quantile), with $x = F^{-1}(q)$, $x = G^{-1}(q)$. And Plot $F^{-1}(q)$-$G^{-1}(q)$.

Usually test normality, take $G = \Phi$

▷ `R.` **Code**

```
boxplot(df$x)

hist(df$x)

hist(df$x,freq=FALSE)
lines(density(df$x))

stem(df$x)

qqnorm(df$x)
qqline(df$x,col='red)
```

- Normality;
- Bias:
  - Selection Bias: Not completely random sampling;
  - Information Bias: Difference between 'designed' and 'get', e.g. no response;
  - Confounding: Exist another important variable, while the model actually focuses on a less important variable, or even reverse the causality.

☐ Diagnostics to Residual

- Linearity: use Residual Plot: Reflect the linearity and variance assumption. ▷ R. **Code**

```
lmfit <- lm(y~x,df)
scatter(df$x,lmfit$residuals)
abline(h=0)
```

- The Assumption of Equal Variances:
  - Bartlett's test:

    Idea: divide the sample into groups g, and get each MSE

$$\mathrm{MSE}_g = \frac{1}{n_g}\sum_{i=1}^{n_g}(Y_{gi} - \hat{Y}_g)^2 \tag{3.47}$$

    and take statistic

$$S = -\frac{(N-g)\ln\left[\sum_g \frac{n_g}{N-n_g}\mathrm{MSE}_g\right] - \sum_g (n_g - 1)\ln\frac{n_g}{N-n_g}\mathrm{MSE}_g}{1 + \frac{1}{3(G-1)}\sum_g\left(\frac{1}{n_g-1} - \frac{1}{N-G}\right)} \sim \chi^2 \tag{3.48}$$

    to conduct test.

    Note: **sensitive** to normal assumption, not robust. Used when normal assumption is satisfied.

- Levene's test: Divide the sample into G groups. Denote **mean** of residual within each group as $\tilde{e}_g$, and in each group compute

$$d_{ig} = |e_{ig} - \tilde{e}_g| \Rightarrow \bar{d}_g = \frac{1}{n_g}\sum_{j=1}^{n_g} d_{ig} \tag{3.49}$$

Then conduct ANOVA to d_{ig}.

If $G = 2$: 2-sample t-test,

$$T = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{\mathscr{L}} t_{n-2} \qquad s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2} \tag{3.50}$$

- Brown-Forsythe's Test (Modified Levene's test): For skewed sample, take the **mean** as **median**, more robust.
- Breusch-Pagan Test:

Assume variance of ε_i dependent on X_i as m^{th} polynomial:

$$\sigma_i^2 = \alpha_0 + \sum_{k=1}^{m} \alpha_k X_i^k \tag{3.51}$$

and test

$$H_0 : \alpha_k = 0\,\forall k = 1, 2, \ldots, m \longleftrightarrow H_1 \tag{3.52}$$

Method: First conduct OLS to get regression line $\hat{l}_1$ and residuals e_i and SSE, and conduct regression of e_i^2 over X_i to get another regression line $\hat{l}_2$ and corresponding SSR*.

Then statistic

$$S = \frac{\text{SSR}^*/2}{(\text{SSE}/n)^2} \xrightarrow{\mathscr{L}} \chi^2_m \tag{3.53}$$

▷ `R.` **Code**

Example for $G = 2$:

```r
group <-factor(rep(c(1,2),length.out=length(df$x),
    each=(ceiling(length(df$x)/2))))

bartlett.test(lmfit$residuals~group,group)

library(car)
leveneTest(lmfit$residuals~group,group,center=mean)
leveneTest(lmfit$residuals~group,group,center=median)

library(lmtest)
bptest(lmfit)
```

- The Assumption of Normality:

In most case we use S-W Test($n < 2000$) and K-S Test($n > 2000$):

  - QQ-plot of ordered residuals;

⋆ Shapiro-Wilk Test (Most Powerful)[18]:

$$R^2 = \frac{\left(\sum_{i=1}^n (X_{(i)} - \bar{X})(m_i - \bar{m})\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (m_i - \bar{m})^2} = corr(X_{(i)}, m_i) \tag{3.54}$$

– Kolmogorov-Smirnov Test:

$$D_n = \sum_x |F_n(x) - \Phi(x)| \tag{3.55}$$

– Cramér-von Mises Test:

$$T = n \int_{-\infty}^{\infty} (F_n(x) - \Phi(x))^2 \, \mathrm{d}\Phi(x) \tag{3.56}$$

– Anderson-Darling Test:

$$A^2 - n \int_{-\infty}^{\infty} (F_n(x) - \Phi(x))^2 \frac{1}{\Phi(x)(1 - \Phi(x))} \, \mathrm{d}\Phi(x) \tag{3.57}$$

▷ **R. Code**

```
qqnorm(lmfit$residuals)
qqline(lmfit$residuals)

qqp <- qqnorm(lmfit$residuals)
cor(qqp$x,qqp$y)

shapiro.test(lmfit$residuals)

ks.test(jitter(lmfit$residuals),pnorm,mean(lmfit$residuals),
    sd(lmfit$residuals))

library(nortest)
cvm.test(lmfit$residuals)

ad.test(lmfit$residuals)
```

- The Assumption of Independence:

  – Durbin-Watson Test:

$$d = \frac{\sum_{j=2}^n (e_j - e_{j-1})^2}{\sum_{j=1}^n e_j^2} \tag{3.58}$$

  $d \in (1.5, 2.5)$ is fine.

  – Ljung-Box Test:

$$Q = n(n+2) \sum_{k=1}^n \frac{\hat{\rho}_k^2}{n-k} \tag{3.59}$$

▷ **R. Code**

[18] Detail of S-W Test and K-S Test see Test of Normality in sec.2.4.6

```
dwtest(lmfit)
```

☐ Diagnostics to Influentials

An intuitive explanation to extreme values:

- Outliers: Extreme case for Y;
- High Leverage: Extreme case for X;
- Influentials: Cases that influence the regression line.

Influentials = Outliers ∩ High Leverage

In OLS part, we got the $\hat{\beta}$ as $\hat{\beta} = (X'X)^{-1}X'Y$ and got $\hat{Y}$ as

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'y = \hat{H}Y \tag{3.60}$$

where $\hat{H}$ is the **Hat Matrix**[19]

Denote in matrix derivation as $H = \dfrac{\partial \hat{Y}}{\partial Y}$. The diagonal elements of $\hat{H}$ is self-sensitivity:

$$h_{ii} = \frac{\partial \hat{Y}_i}{\partial Y_i} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}} \tag{3.61}$$

Note: the distribution of e_i in eqa.(3.16) thus can be written in h_{ii} as

$$e_i \sim (0, \sigma^2(1 - h_{ii})) \tag{3.62}$$

Some refined residuals to help conduct Diagnostics:

- Standardized Residual:

$$\frac{e_i}{\sigma_{e_i}} = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}} \tag{3.63}$$

- (Internal) Studentized Residual: replace σ with $s = \hat{\sigma}$

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{3.64}$$

- (External) Studentized Residual: To avoid self-influence, take **deleted** residual:

  Delete the i^{th} case and conduct regression to the $n - 1$ sample cases, denote the regression parameter as

$$\hat{\beta}_{1(\wedge i)} \qquad \hat{\beta}_{0(\wedge i)} \tag{3.65}$$

  and deleted residual defined as

$$d_i = Y_i - Y_{i(\wedge i)} = \frac{e_i}{1 - h_{ii}} \tag{3.66}$$

  external studentized residual:

$$t_i = \frac{d_i}{\sigma_{(\wedge i)\sqrt{1 - h_{ii}}}} \tag{3.67}$$

[19] It can also be considered as the projection matrix onto span$\{X\}$.

Cook's Distance:

$$D_i = \frac{\sum_{k=1}^{n}(Y_k - \hat{Y}_{k(\wedge i)})^2}{p\hat{\sigma}^2} = \frac{e_i^2}{p\hat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \tag{3.68}$$

Comment:

$$D_i = \frac{e_i^2}{p\hat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}} \times r_i^2 \tag{3.69}$$

where $\frac{1}{p}\frac{h_{ii}}{1-h_{ii}}$ correponds to hige leverage, and r_i^2 correponds to outliers, multiply to get influentials.

3.3.2 Remedies

☐ General Linear Model

$$E(Y) = g(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots) \tag{3.70}$$

☐ Remedies: Conduct Transformation

- Stablize Variance;
- Improve Normality;
- Simplify the Model.

Transformation Methods:

- Variance Stabilizing Transformations: For $E(Y_X) = \mu_X$, $var(Y_X) = h(\mu_X)$, take transformation $f(Y)$ such that $var(f(Y)) =$ const, satisfies

$$f(\mu) = \int \frac{c\,\mathrm{d}\mu}{\sqrt{h(\mu)}} \tag{3.71}$$

  Examples:

$$\begin{aligned} h(\mu) =& \mu^2 \Rightarrow f(\mu) = \ln \mu \\ h(\mu) =& \mu^{2\nu} \Rightarrow f(\mu) = \mu^{1-\nu} \end{aligned}$$

- Box-Cox Transformation: Take

$$Y^* = \frac{Y^\lambda - 1}{\lambda} \tag{3.72}$$

  Examples:

$$\begin{aligned} \lambda = 1 &\Rightarrow Y^* \sim Y \\ \lambda = 0.5 &\Rightarrow Y^* \sim \sqrt{Y} \\ \lambda = 0 &\Rightarrow Y^* \sim \ln Y \\ \lambda = -1 &\Rightarrow Y^* \sim 1/Y \end{aligned}$$

  And conduct regression to model

$$Y^* = \beta_0 + \beta_1 X + \varepsilon_i \tag{3.73}$$

Likelihood Function

$$L(\beta,\sigma^2;\lambda)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(Y_i^*-\beta_0-\beta_1X_i)^2\right)J(\frac{\partial Y^*}{\partial Y}) \tag{3.74}$$

where the Jacobi Matrix denoted in Geometric Mean $\mathrm{GM}(Y)=\prod_{i=1}^n Y_i^{1/n}$

$$J(\frac{\partial Y^*}{\partial Y})=\prod_{i=1}^n Y_i^{\lambda-1}=\mathrm{GM}(Y)^{n(\lambda-1)} \tag{3.75}$$

MLE Estiamtor:

$$\begin{aligned}\hat\beta^*&=(X'X)^{-1}X'Y^*\\ \hat\sigma_n^2&=\frac{1}{n}\mathrm{SSE}^*\\ \mathrm{SSE}^*&=\sum_{i=1}^n(Y_i^*-\hat Y_i^*)^2\end{aligned}$$

And when β, σ^2 take MLE estimator, $L(\beta,\sigma^2;\lambda)$ can be regarded a function of λ:

$$\ln L(\beta,\sigma^2;\lambda)=l(\lambda)=-\frac{n}{2}\ln\frac{\hat\sigma_n^2}{\mathrm{GM}(Y)^{2(\lambda-1)}}+\mathrm{const} \tag{3.76}$$

For simplification, denote $Z=Y*/J^{1/n}$ and get

$$l(\lambda)=-n\ln\sigma_{n_Z}^2+\mathrm{const} \tag{3.77}$$

where

$$Z_i^*=\begin{cases}\dfrac{Y_i^\lambda-1}{\lambda}\dfrac{1}{\prod_{k=1}^n Y_k^{\frac{\lambda-1}{n}}}, & \lambda\neq 0\\ \ln Y_i\prod_{k=1}^n Y_k^{\frac{1}{n}}, & \lambda=0\end{cases} \tag{3.78}$$

Plot $l(\lambda)$-λ to determine a proper λ and transform $Y^*=\dfrac{Y^\lambda-1}{\lambda}$:

- Selected λ should be closed to $\lambda_{\arg\max l}$, at least within CI[20]

$$\{\lambda|l(\lambda)\geq l(\lambda_{\arg\max l})-\frac{1}{2}\chi^2_{1,1-\alpha}\} \tag{3.79}$$

- Should pick a λ which is **Interpretable**. e.g. If $\lambda=1$ is within range, then take $\lambda=1$ (does not transform).

Section 3.4 Multiple Linear Regression

☐ Sample Geometry Notation

In sample matrix notation:

$$Y=X\beta+\varepsilon\leftrightharpoons Y_i=X\beta_i+\varepsilon_i,\ \forall i=1,2,\ldots,q \tag{3.80}$$

[20]Here CI can be derived using Wilk's Thm.

where

$$\underset{n\times q}{Y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1q} \\ y_{21} & y_{22} & \dots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nq} \end{bmatrix} = \begin{bmatrix} y_1, y_2, \dots, y_q \end{bmatrix} \qquad y_i = \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{ni} \end{bmatrix} \tag{3.81a}$$

$$\underset{n\times (p+1)}{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \qquad x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \tag{3.81b}$$

$$\underset{(p+1)\times q}{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0q} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1q} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pq} \end{bmatrix} = \begin{bmatrix} \beta_1, \beta_2, \dots, \beta_q \end{bmatrix} \qquad \beta_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i1} \\ \vdots \\ \beta_{ip} \end{bmatrix} \tag{3.81c}$$

$$\underset{n\times q}{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1q} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{nq} \end{bmatrix} = \begin{bmatrix} \varepsilon_1, \varepsilon_2, \dots, \varepsilon_q \end{bmatrix} \qquad \varepsilon_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{ni} \end{bmatrix} \tag{3.81d}$$

Under matrix notation, model and assumptions eqa(3.2) can be expressed in condensed notation:

$$Y_i = X\beta_i + \varepsilon_i \sim N_n(X\beta_i, \sigma_i^2 I_n), \quad i = 1, 2, \dots, q \tag{3.82}$$

To conduct OLS

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min}(Y - X\beta)^T(Y - X\beta) \tag{3.83}$$

Here we introduce two approaches:

- Analytical: Take matrix differciation (See sec.4.1.2)

$$\begin{aligned} 0 &= \frac{\partial (Y - X\beta)^T(Y - X\beta)}{\partial\beta} = \frac{\partial}{\partial\beta}(Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta) \\ &= -X^TY - X^TY + (X^TX + XX^T)\beta = -2X^T(Y - X\beta) \end{aligned}$$

  Thus we get OLS:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.84}$$

- Geometric/Algebraical: Use hyper-projection.

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min}\, d(Y, X\beta) \tag{3.85}$$

  i.e. $\hat{\beta}$ is the (hyper-)projection of Y onto X (within Euclidean Space), naturally we have

$$(X\beta)^T(Y - X\beta) = 0 \Rightarrow \hat{\beta} = (X'X)^{-1}X'Y \tag{3.86}$$

☐ Matrix Notation of OLS Estimator:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.87}$$

(For simplificaiton, the following part consider multivariate $\underset{n\times(p+1)}{X}$ with one $\underset{n\times 1}{Y}$)

Properties & Extrapolation

- Sampling Districution of $\hat{\beta}$: (Here consider normal case $Y \sim N(X\beta, \sigma^2 I_n)$, and use eqa(4.36))

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N_n(\beta, \sigma^2(X'X)^{-1}) \tag{3.88}$$

  Comment: $cov(\beta_i, \beta_j)$ are generally not 0, $\Rightarrow \beta_i, \beta_j$ dependent.

- Predicted Response & Hat Matrix H:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y \equiv HY = P_X Y \tag{3.89}$$

  where **Hat Matrix**/Influence matrix/Projection matrix $H = P_X = X(X'X)^{-1}X'$, with properties

  – Symmertric: $H^T = H$;

  – Idempotence: $H^2 = H$

  – H and self-influene factor h_{ii}: Note the linearity of $\hat{Y}$ on Y

$$\hat{Y} = HY \Rightarrow H = \frac{\partial \hat{Y}}{\partial Y} \tag{3.90}$$

  The diagonal elements of H is

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} = X_i(X'X)^{-1}X_i' \tag{3.91}$$

  Comment on h_{ii}: $var(e_i) = \sigma^2(1 - h_{ii})$, for $h_{ii} \to 1$, i.e. the regression line always pass y_i, thus it's 'influential'.

  – H and Residual e

- Residual:

$$e = Y - \hat{Y} = (I - H)Y \sim N_n\left(0, \sigma^2(I - H)\right) \tag{3.92}$$

  Covariance Matrix of Residual:

$$cov(e) = \sigma^2(I - H) = \sigma^2 \begin{bmatrix} 1 - h_{11} & -h_{12} & \dots & -h_{1n} \\ -h_{21} & 1 - h_{22} & \dots & -h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{n1} & -h_{n2} & \dots & 1 - h_{nn} \end{bmatrix} \tag{3.93}$$

- Estimator and Distribution of σ^2:

  First use eqa(4.37) to get [21]

$$E(\text{SSE}) = E(e'e) = E(Y'(I - H)Y) = (X\beta)'(I - H)X\beta + tr((I - H)\sigma^2 I_n) = \sigma^2(n - p - 1) \tag{3.95}$$

[21] Also we need the property of idmpotnet matrix

$$\lambda_i = 0 \text{ or } 1 \Rightarrow tr(H) = \text{rank}(H) = \sum_{i=1}^{n} \lambda_i = \#(\lambda = 1) \tag{3.94}$$

dof of Residual e (use definition eqa(3.10)):

$$dof_e = dof_{(I-H)Y} = \text{rank}(I-H) = n-p-1 \tag{3.96}$$

Thus the unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \text{MSE} = \frac{e'e}{n-p-1} = \frac{Y'(I-H)Y}{n-p-1} \tag{3.97}$$

Distribution (under normal assumption):

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{3.98}$$

- Gauss-Markov Thm.: OLS Estimator of β is the BLUE Estimator.
- Leverage and Mahalanobis Distance:

  Mahalanobis Distance between X and Y as defined in eqa(4.19)

$$d_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1}(\vec{x}-\vec{\mu})} \tag{3.99}$$

  And we can proof d_M of a case item $X_{i\cdot} = (1, X_{i1}, X_{i2}, \dots, X_{ip})$ is

$$d_M^2(X_{i\cdot}) = (n-1)(h_{ii} - \frac{1}{n}) \tag{3.100}$$

Test of Normality: Jarque-Bera Test , using skewness $\hat{g}_1$ and kurtosis $\hat{g}_2$

$$\text{JB} = \frac{n}{6}(\hat{g}_1^2 + \frac{1}{4}\hat{g}_2^2) \xrightarrow{\mathscr{L}} \chi^2_2 \tag{3.101}$$

▷ `R.` **Code**

```
library(tseries)
jarque.bera.test(df$y)
```

Chapter. IV 多元统计分析部分

Instructor: Dong Li & Tianying Wang

Section 4.1 Multivariate Data

In this section, we consider a **Multivariate Statistic Model**. Sample comes from p dimension multivariate population $f(x_1, x_2, \ldots, x_p)$.

Notation : In this section, we still denote random variable in upper case and observed value in lower case, specially express random vector in bold font. **But** in this section we usually omit the vector symbol $\vec{\cdot}$. e.g. random vector with n **variable** is denoted as $\mathbf{X} = (X_{\cdot 1}, X_{\cdot 2}, \ldots, X_{\cdot p})$; sample of size n from the multivariate population is a $n \times p$ matrix $\{x_{ij}\}$, each sample item (a row in sample matrix) is denoted as x_i' or x_i^T.[22]

4.1.1 Matrix Representation

- Random Variable Representation
- Sample Representation
- Statistics Representation
- Sample Statistics Properties

☐ Random Variable Representation:

- Random Matrix: Definition and basic properties of r.v. see section 1.3. Now extend the definition to matrix $X = \{X_{ij}\}$.

$$X = \{X_{ij}\} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{n2} & \ldots & X_{np} \end{bmatrix} \tag{4.1}$$

  And we can further define $E(X) = \{E(X_{ij})\}$. For any const matrix A, B we have

$$E(AXB) = AE(X)B \tag{4.2}$$

- Random Vector: For a $p \times 1$ random vector $\vec{X} = (X_1, X_2, \ldots, X_p)^T$, denote (Marginal) expectation and variance, and covariance, correlation coefficient between X_i, X_j as follows:

$$\begin{aligned} \mu_i &= E(X_i) \\ \sigma_{ii} &= \sigma_i^2 = E(X_i - \mu_i)^2 \\ \sigma_{ij} &= E[(X_i - \mu_i)(X_j - \mu_j)] \\ \rho_{ij} &= \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} \end{aligned}$$

[22]Here sample item (or sample case) $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$ is a column vector.

and we have covariance matrix (as defined in section 1.4.3, eqa.1.44)

$$\Sigma = E[(X-\mu)(X-\mu)^T] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} \tag{4.3}$$

and Standard Deviation Matrix

$$V^{1/2} = diag\{\sqrt{\sigma_{ii}}\} \tag{4.4}$$

Based on $\vec{X} = (X_1, X_2, \dots, X_p)$, consider the linear combination:$Y = c'X = c_1X_1 + c_2X_2 + \dots c_pX_p$

$$E(y) = c'\mu \qquad var(Y) = c'\Sigma c$$

and $Z_i = \sum_{j=1}^{p} c_{ij}X_j$ (i.e. $Z = CX$):

$$\mu_Z = E(Z) = C\mu_X \qquad \Sigma_Z = C\Sigma_X C^T \tag{4.5}$$

and Correlation Matrix

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{p2} & \dots & \rho_{pp} \end{bmatrix} = V^{-1/2}\Sigma V^{-1/2} \tag{4.6}$$

☐ Sample Representation:

Sample of n items from population characterized by p variables

Item \ Variable	Variable 1	Variable 2	$\dots$	Variable j	$\dots$	Variable p
Item 1	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
Item 1	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Item j	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Item n	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}

Or represented in condense notation:

$$X = \{x_{ij}\} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \dots & y_p \end{bmatrix} \tag{4.7}$$

☐ Statistics Representation

- Unit 1 vector:

$$\mathbf{1}_k = (\underbrace{1, 1, \ldots, 1}_{k\ 1\text{ in total}})^T \tag{4.8}$$

Unit 1 matrix:

$$\mathcal{J}_n = \{1\}_{n\times n} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}_{n\times n} \tag{4.9}$$

- Sample mean:

$$\bar{x}_i = \frac{x_{1i} + x_{2i} + \ldots + x_{ni}}{n} = \frac{y_i'\mathbf{1}_n}{n} \tag{4.10}$$

- Deviation of measurement of the i^{th} variable:

$$d_i = y_i - \bar{x}_i\mathbf{1}_n = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix} \tag{4.11}$$

- Covariance Matrix:

  – Variance of y_i:

$$s_i^2 = s_{ii} = \frac{1}{n}d_i'd_i = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2, \quad i = 1, 2, \ldots p \tag{4.12}$$

  – Covariance between y_i and y_j:

$$s_{ij} = \frac{1}{n}d_i'd_j = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j), \quad i, j = 1, 2, \ldots p \tag{4.13}$$

  – Correlation Coefficient:

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} = \frac{\sum\limits_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum\limits_{k=1}^{n}(x_{ki} - \bar{x}_i)^2}\sqrt{\sum\limits_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}}, \quad i, j = 1, 2, \ldots p \tag{4.14}$$

In condense notation, define Covariance Matrix from sample of size n:

$$S_n^2 = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{p2} & \ldots & s_{pp} \end{bmatrix} \tag{4.15}$$

and sample Correlation Coefficient Matrix:

$$R_n = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1p} \\ r_{21} & r_{22} & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{p2} & \ldots & r_{pp} \end{bmatrix} \tag{4.16}$$

- Generalized sample variance: $|S| = \lambda_1\lambda_2\ldots\lambda_p$, where λ_i are eigenvalues.
- 'Statistical Distance' between vectors: to measure the difference between two vectors $x = (x_1, x_2, \ldots, x_p)$ and $y = (y_1, y_2, \ldots, y_p)$.
  - Euclidean Distance:

$$d_E(x, y) = \sqrt{(x-y)^T(x-y)} \tag{4.17}$$

  - **Mahalanobis Distance**: Scale invariant distance, and include information about relativity:

$$d_M(x, y) = \sqrt{(x-y)'S^{-1}(x-y)} \tag{4.18}$$

    Note: P, Q are from the same distribution with covariance matrix S_p. When $S = I$, return to Euclidean distance.

    Remark: Mahalanobis distance is actually the normalized Euclidean distance in principal component space. So we can actually define the Mahalanobis distance for one sample case $\vec{x} = (x_1, x_2, \ldots, x_p)$ from distribution of $(\vec{\mu}, \Sigma)$

$$d_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})} \tag{4.19}$$

    Note: the hyper-sruface $d_M(\vec{x})$ forms a ellipsoid.

☐ Sample Statistics Properties

Consider take an n cases sample from r.v. population $\vec{X} = (X_1, X_2, \ldots, X_p)$, population mean μ and covariance matrix Σ.

- $E(\bar{\bar{X}}) = \mu$;
- $cov(\bar{X}) = \frac{1}{n}\Sigma$;
- $E(S_n) = \frac{n-1}{n}\Sigma$

4.1.2 Review: Some Matrix Notation & Lemma

- Orthonormality: For square matrix P satisfies:

$$x_i^T x_j = \delta_{ij} \tag{4.20}$$

  where x_i, x_j are columns of P.

- Eigenvalue and Eigenvector: For square matrix A, its eigenvalues λ_i and corresponding eigenvectors e_i satisfies:

$$Ae_i = \lambda_i e_i,\ \forall i = 1, 2, \ldots p \tag{4.21}$$

  Denote $P = [e_1, e_2, \ldots, e_p]$, which is an orthonormal matrix. And denote $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$.

$$A = \sum_{i=1}^{p} \lambda_i e_i e_i^T = P\Lambda P^T = P\Lambda P^{-1} \tag{4.22}$$

  is called the Spectral Decomposition of A

- Square root matrix: Def. as

$$A^{1/2} = \sum_{i=1}^{p} \sqrt{\lambda_i} e_i e_i^T = P\Lambda^{1/2}P^T \tag{4.23}$$

  Properties:

  - $A^{1/2}A^{1/2}A$;
  - $A^{-1/2} = (A^{1/2})^{-1} = PL^{-1/2}P^T$;
  - $tr(A) = \sum_{i=1}^{n} \lambda_n$;
  - $|A| = \prod_{i=1}^{n} \lambda_n$.

- (Symmetric) Positive Definite Matrix: Say A a Positive Definite Matrix if

$$x^T A x > 0, \forall x \in \mathbb{R}^p \tag{4.24}$$

  where $x^T A x$ is called a Quadric Form.

  Properties:

  - Use the Spectral Decomposition of A, we can write the Quadric Form as

$$x^T A x = x^T P \Lambda P^T x = y^T \Lambda y = \sum_{i=1}^{p} \lambda_i y_i^2 = \sum_{i=1}^{p} (\sqrt{\lambda_i} y_i)^2 \tag{4.25}$$

  - Eigenvalues $\lambda_i > 0$, $\forall i = 1, 2, \ldots, p$
  - A can be written as product of symmetric matrix: $A = Q^T Q$ (Q is symmetric);

- Trace of Matrix: For $p \times p$ square matrix A

$$tr(A) = \sum_{i=1}^{p} a_{ii} \tag{4.26}$$

  Properties:

  - $tr(AB) = tr(BA)$;
  - $x'Ax = tr(x'Ax) = tr(Axx')$

- Matrix Partition: partition matrix $\underset{p\times p}{A}$ as

$$A = \begin{bmatrix} \underset{q_1\times q_1}{A_{11}} & \underset{q_1\times q_2}{A_{12}} \\ \underset{q_2\times q_1}{A_{21}} & \underset{q_2\times q_2}{A_{22}} \end{bmatrix}$$

  where $p = q_1 + q_2$

  Property:

$$|A| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}| = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}|$$

  | A |=| A22 || A11 - A12A-122A21 |=| A11 || A22 - A21A-111A12 |

- Calculus Notations: We want to take derivative of $y = (y_1, y_2, \dots, y_q)^T$ over $x = (x_1, x_2, \dots, x_p)^T$

  We use 'Denominator-layout', which is

$$\frac{\partial y}{\partial x} = \frac{\partial y^T}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \dots & \frac{\partial y_q}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_p} & \frac{\partial y_2}{\partial x_p} & \dots & \frac{\partial y_q}{\partial x_p} \end{bmatrix} \tag{4.27}$$

  Properties (under denominator-layout):

  - $\frac{\partial}{\partial x} Ax = A^T$;
  - $\frac{\partial}{\partial x} x^T A = A$;
  - $\frac{\partial}{\partial x} x^T x = 2x$;
  - $\frac{\partial}{\partial x} x^T Ax = Ax + A^T x$;
  - $\frac{\partial}{\partial x} \log(x^T Ax) = \frac{2Ax}{x^T Ax}$;
  - $\frac{\partial |A|}{\partial A} = |A| A^{-1}$;
  - $\frac{\partial tr(AB)}{\partial A} = B^T$;
  - $\frac{\partial tr(A^{-1}B)}{\partial A} = -A^{-1} B^T A^{-1}$

- Kronecker Product: For matrix $\underset{m\times n}{A} = \{a_{ij}\}$, $\underset{p\times q}{B} = \{b_{ij}\}$. Their Kronecker product

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix} \tag{4.28}$$

4.1.3 Useful Inequalities

- Cauchy-Schwartz Inequality:

  Let b, d are any $p \times 1$ vectors.

$$(b'd)^2 \leq (b'b)(d'd) \tag{4.29}$$

- Extended Cauchy-Schwartz Inequality:

  Let B be a positive definite matrix.

$$(b'd)^2 \le (b'Bb)(d'B^{-1}d) \tag{4.30}$$

- Maximazation Lemma:

  d be a given vector, for any non-zero vector x,

$$\frac{(x'd)^2}{x'Bx} \le d'B^{-1}d \tag{4.31}$$

  Take Maximum when $x = cB^{-1}d$.

Section 4.2 Statistical Inference to Multivariate Population

Statistics model: a n cases sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, where each $\mathbf{X}_i$ i.i.d. from a multivariate population (usually consider a multi-normal). i.e.

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{n2} & \ldots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix} \tag{4.32}$$

Section 4.3 Multivariate Normal Distribution

Univariate Noraml Distribution: $N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2} \tag{4.33}$$

Multivariate Normal Distribution: $X \sim N_p(\vec{\mu}, \Sigma)$[23]

$$f_{\mathbf{X}}(\vec{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{(\vec{x}-\vec{\mu})'\Sigma^{-1}(\vec{x}-\vec{\mu})}{2}\right) \tag{4.34}$$

Note: Here in the exp, the $(\vec{x}-\vec{\mu})'\Sigma^{-1}(\vec{x}-\vec{\mu})$ is the Mahalanobis Distance d_M defined in eqa.4.19

Remark: A n-dimension multivariate normal has $\dfrac{p(p+1)}{2}$ free parameters. Thus for a very high dimension, contains too many free parameters to be determined!

Properties: Consider $X \sim N_p(\mu, \Sigma)$

- Linear Transform:

  - For a $p \times 1$ vector a:

$$X \sim N_p(\mu, \Sigma) \Leftrightarrow a'X \sim N(a'\mu, a'\Sigma a),\ \forall a \in \mathbb{R}^p \tag{4.35}$$

    (Proof: use characteristic function.)

[23] Detailed derivation see section 1.8

– For a $q \times p$ const matrix A:

$$AX + a \sim N_q(A\mu + a, A\Sigma A') \tag{4.36}$$

– For a $p \times p$ square matrix A:

$$E(X'AX) = \mu' A\mu + tr(A\Sigma) \tag{4.37}$$

- Conditional Distribution: Take partition of $\underset{p\times 1}{X} \sim N(\underset{p\times 1}{\mu}, \underset{p\times p}{\Sigma})$ into $\underset{q_1\times 1}{X_1}$ and $\underset{q_2\times 1}{X_2}$, where $q_1 + q_2 = p$. Write in matrix form:

$$\underset{p\times 1}{X} = \begin{bmatrix} \underset{q_1\times 1}{X_1} \\ \underset{q_2\times 2}{X_2} \end{bmatrix} \qquad \underset{p\times 1}{\mu} = \begin{bmatrix} \underset{q_1\times 1}{\mu_1} \\ \underset{q_2\times 2}{\mu_2} \end{bmatrix} \qquad \underset{p\times p}{\Sigma} = \begin{bmatrix} \underset{q_1\times q_1}{\Sigma_{11}} & \underset{q_1\times q_2}{\Sigma_{12}} \\ \underset{q_2\times q_1}{\Sigma_{21}} & \underset{q_2\times q_2}{\Sigma_{22}} \end{bmatrix} \tag{4.38}$$

i.e.

$$\underset{p\times 1}{X} = \begin{bmatrix} \underset{q_1\times 1}{X_1} \\ \underset{q_2\times 2}{X_2} \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \underset{q_1\times 1}{\mu_1} \\ \underset{q_2\times 2}{\mu_2} \end{bmatrix}, \begin{bmatrix} \underset{q_1\times q_1}{\Sigma_{11}} & \underset{q_1\times q_2}{\Sigma_{12}} \\ \underset{q_2\times q_1}{\Sigma_{21}} & \underset{q_2\times q_2}{\Sigma_{22}} \end{bmatrix} \right) \tag{4.39}$$

Independence: $X_1 \parallel X_2 \Leftrightarrow \Sigma_{21} = \Sigma_{12}^T = 0$

And the conditional dictribution $X_1|X_2 = x_2$ is given by [24]

$$X_1|_{X_2=x_2} \sim N_p(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{4.40}$$

- Multivariate Normal & χ^2

  Let $X \sim N_p(\mu, \Sigma)$, then

$$(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi_p^2 \tag{4.41}$$

- Linear Combination: Let $X_1, X_2 \ldots, X_n$ with $X_i \sim N_p(\mu_i, \Sigma)$ (different μ_i, same Σ). And denote $V_1 = \sum_{i=1}^n c_i X_i$, then

$$V_1 \sim N_p(\sum_{i=1}^n c_i\mu_i, \sum_{i=1}^n c_i^2 \Sigma) \tag{4.42}$$

4.3.1 MLE of Multivariate Normal

Under the notation in eqa(4.32), i.e. each sample case $\mathbf{X}_i$ i.i.d. $\sim N_p(\mu, \Sigma)$, we can get the joint PDF of $\mathbf{X}$:

$$f_{\mathbf{X_1},\ldots,\mathbf{X_n};\mu,\Sigma}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)'\Sigma^{-1}(x_i - \mu)}{2} \right) \tag{4.43}$$

and at the same time get likelihood function[25]:

[24] In eqa(4.36), take

$$\underset{p\times p}{A} = \begin{bmatrix} \underset{q\times q}{I} & -\underset{q\times(p-q)}{\Sigma_{12}\Sigma_{22}^{-1}} \\ \underset{(p-q)\times q}{0} & \underset{(p-q)\times(p-q)}{I} \end{bmatrix}$$

[25] Here we need to use the property of trace

$$x'Ax = tr(x'Ax) = tr(Ax'x) \tag{4.44}$$

$$L(\mu,\Sigma;x_1,\dots,x_n)=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left[-\frac{1}{2}tr\left(\Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'+n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right)\right] \tag{4.45}$$

And we can get the MLE of μ and Σ as follows[26]:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$$
$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'=\frac{n-1}{n}S^2$$

And we can furthur construct MLE of function of μ, Σ (use invariance property of MLE), for example

$$|\hat{\Sigma}|=|\hat{\Sigma}|$$

Note: $(\hat{\mu},\hat{\Sigma})$ is sufficient statistic of multi-normal population.

4.3.2 Sampling distribution of $\bar{X}$ and S^2

$\hat{\mu}=\bar{X}$ and $\hat{\Sigma}=\dfrac{n-1}{n}S^2$ are statistics, with sampling distribution.

□ Sampling distribution of $\bar{X}$

Similar to monovariate case:

$$\bar{X}\sim N_p(\mu,\frac{1}{n}\Sigma)$$

□ Sampling distribution of S^2

- Monovariate case: Consider $(X_1,X_2,\dots,X_n)$ i.i.d. $\sim N(\mu,\sigma^2)$

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

- Multivariate case: Consider $(\mathbf{X}_1,\mathbf{X}_2,\dots,\mathbf{X}_n)$ i.i.d. $\sim N_p(\mu,\Sigma)$

  Then
  $$(n-1)S^2\sim W_p(n-1,\Sigma)$$

  Where $W_p(n-1,\Sigma)$ is Wishart Distribution, details as follows:

  For r.v. $Z_1,Z_2,\dots,Z_m$ i.i.d. $\sim N_p(0,\Sigma)$, def p dimensional **Wishart Distribution** with dof m as $W_p(m,\Sigma)$.[27]

  $$W_p=\sum_{i=1}^{n}Z_iZ_i' \tag{4.46}$$

  PDF of $W_p(m,\Sigma)$:

  $$f_W(w;p,m,\Sigma)=\frac{|w|^{\frac{m-p-1}{2}}\exp\left(-\frac{1}{2}tr(\Sigma^{-1}w)\right)}{2^{\frac{mp}{2}}|\Sigma|^{-1/2}\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma(\frac{m-i+1}{2})} \tag{4.47}$$

---

[26] Detailed proof see '*Applied Multivariate Statistical Analysis*' P130

[27] $W_p(m,\Sigma)$ is a distribution defined on $p\times p$ matrix space.

C.F.

$$\phi(T) = |I_p - 2i\Sigma T|^{-\frac{m}{2}} \tag{4.48}$$

Properties:

- For independent $A_1 \sim W_p(m_1, \Sigma)$ and $A_2 \sim W_p(m_2, \Sigma)$, then

$$A_1 + A_2 \sim W_p(m_1 + m_2, \Sigma)$$

- For $A \sim W_p(m, \Sigma)$, then

$$CAC' \sim W_p(m, C\Sigma C')$$

- Wishart distribution is the matrix generization of χ^2_n. When $p = 1, \Sigma = \sigma^2 = 1$, $W_p(m, \Sigma)$ naturally reduce to χ^2_m.

$$\chi^2_n = W_1(n, 1)$$

▷ `R.` **Code**

Distribution functions are in package `MCMCpack`, or use `rWishart()` function.

☐ Large sample $\bar{X}$ and S^2

- $\sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathscr{L}} N_p(0, \Sigma)$;
- $n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \xrightarrow{\mathscr{L}} \chi^2_p$

Section 4.4 Multivariate Statistical Inference

4.4.1 Hypothesis Testing for Normal Population

- **One-Population Hypothesis Testing**:

Conduct hypothesis testing to μ:

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0$$

☐ Hotelling's T^2 test

- One-Dimensional case: t-test

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

i.e.

$$T^2 = [\sqrt{n}(\bar{X} - \mu_0)](S^2)^{-1}[\sqrt{n}(\bar{X} - \mu_0)] \sim t^2_{n-1} = F_{1,n-1}$$

- Multi-Dimensional case: Hotelling's T^2

$$T^2 = [\sqrt{n}(\bar{X} - \mu_0)'](S^2)^{-1}[\sqrt{n}(\bar{X} - \mu_0)] \sim N_p(0, \Sigma)' \frac{W_p(n-1, \Sigma)}{n-1} N_p(0, \Sigma) = \frac{p}{n-p}(n-1) F_{p,n-p}$$

And we can get the distribution of **Hotelling's** T^2:

$$\frac{n-p}{p} \frac{T^2}{n-1} \sim F_{p,n-p}$$

Rejection Rule:

$$T^2 > \frac{p(n-1)}{n-p} F_{p,n-p,\alpha}$$

Property:

Invariant for X transform: For $Y = CX + d$, then

$$T_Y^2 = n(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) = T_X^2$$

☐ LRT of $\hat{\mu}$

Monovariate case see sec.2.4.3.

LRT uses the statistic:

$$\Lambda = \frac{\max_{H_0} L(\mu_0, \Sigma)}{\max_{H_0 \cup H_1} L(\mu, \Sigma)} = (1 + \frac{T^2}{n-1})^{-n/2}$$

where $T^2 = n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0)$

- **Two-Population Hypothesis Testing**:

  Conduct hypothesis testing to $\delta = \mu_1 - \mu_2$:

  $$H_0 : \delta = \delta_0 \longleftrightarrow H_1 : \delta \neq \delta_0$$

  Notation: The two sample of size n_1, n_2, each denoted as

  $$X_{1,ij} \qquad X_{2,ij}$$

  with mean μ_1, μ_2 and covariance matrix Σ_1, Σ_2

  – Paired Samples: $n_1 = n_2$

    For two paires samples $\{X_{1,ij}\}$, $X_{2,ij}$, take subtraction as

    $$D_{ij} = X_{1,ij} - X_{2,ij}$$

    denote $\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j$, $S_D^2 = \frac{1}{n-1} \sum_{j=1}^n (D_j - \bar{D}'(D_j - \bar{D}))$
    and conduct test to

    $$H_0 : \bar{D} = \delta_0 \longleftrightarrow H_1 : \bar{D} \neq \delta_0$$

    And the folloeing steps are as in One-population testing, test

    $$T^2 = n(\bar{D} - \delta)' (S_D^2)^{-1} (\bar{D} - \delta) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

  – Under Equal Unknown Variance: $\Sigma_1 = \Sigma_2$

$$\bar{X}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} X_{1,j} \qquad \bar{X}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{1,j} \tag{4.49}$$

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_{1,j} - \bar{X}_1)(X_{1,j} - \bar{X}_1)' \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2,j} - \bar{X}_2)(X_{2,j} - \bar{X}_2)' \tag{4.50}$$

And denote pooled variance

$$S^2_{\text{pooled}} = \frac{1}{n_1+n_2-2}\left((n_1-1)S_1^2+(n_2-1)S_2^2\right) \sim \frac{W_p(n_1+n_2-2,\Sigma)}{n_1+n_2-2}$$

Under H_0, we have

$$T^2 = \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}}(\bar{X}_1-\bar{X}_2-\delta_0)'(S^2_{\text{pooled}})^{-1}(\bar{X}_1-\bar{X}_2-\delta_0) \sim \frac{p(n_1+n_2-2)}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}$$

4.4.2 Confidence Region

Estimate the confidence region for μ of $X\sim N_p(\mu,\Sigma)$, Monovariate case see sec.2.3.3

- Confidence Region:

  Also use Hotelling's T^2

  $$\frac{n-p}{p}\frac{T^2}{n-1} \sim F_{p,n-p}$$

  And take $100(1-\alpha)\%$ confidence region of μ as

  $$R(x) = \{x|T^2\le c^2\} \qquad c^2 = \frac{p}{n-p}(n-1)F_{p,n-p,\frac{\alpha}{2}}$$

  The shape of $R(x)$ is an ellipsoid.

- Individual Converage Interval

  Use the decomposition of S^2 as a positive finite matrix $S^2 = A^TA$, where A is some $p\times p$ matrix, then

  $$T^2 = [\sqrt{n}(\bar{X}-\mu_0)'](S^2)^{-1}[\sqrt{n}(\bar{X}-\mu_0)] = [A^{-1\prime}\sqrt{n}(\bar{X}-\mu_0)]'[A^{-1\prime}\sqrt{n}(\bar{X}-\mu_0)]$$

  Thus denote $Z = A^{-1\prime}(X-\mu_0)\sim N_p(0,A^{-1\prime}\Sigma A^{-1})$, the T^2 estimator of Z would be

  $$T_Z^2 = [\sqrt{n}\bar{Z}]'(S_Z^2)^{-1}[\sqrt{n}\bar{Z}] = n\bar{Z}'\bar{Z} = \frac{1}{n}\sum_{i=1}^n \bar{Z}_i^2 \sim F_{p,n-p}$$

  As a simplified case, we can take the **Individual Converage Interval** of Z_i, which is

  $$\frac{\sqrt{n}Z_i}{s_{Z_i}} \sim t_{n-1}$$

  And we can take the Confidence Region[28] as

  $$R(z) = \bigotimes_{i=1}^n(\bar{Z}_i \pm s_{Z_i}t_{n-1,\frac{\beta}{2}})$$

  where β take

  $$1-p\beta = 1-\alpha$$

  Note: Consider that

  $$P(\text{all } Z_i \text{ in CI}_i) \ge 1-m\beta = 1-\alpha$$

  So the real CR for μ should be larger.

  The shape of $R(x)$ is an oblique cubold.

[28] The confidence region of Z can be transformed to that of X using $\hat{Z} = A^{-1\prime}(\hat{X}-\bar{X})$.

4.4.3 Large Sample Multivariate Inference

Basic point:

$$\bar{X} \xrightarrow{\mathscr{L}} \mu \qquad S^2 \xrightarrow{\mathscr{L}} \Sigma$$

- One-sample Mean:

$$n(\bar{X}-\mu)(S^2)^{-1}(\bar{X}-\mu) \xrightarrow{\mathscr{L}} \chi_p^2$$

- Unequal Variance Two-sample Mean:

$$\bar{X}_1 - \bar{X}_2 \xrightarrow{\mathscr{L}} N(\mu_1-\mu_2, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2) \qquad \frac{1}{n_1}S_1^2 + \frac{1}{n_2}S_2^2 \xrightarrow{\mathscr{L}} \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$$

Test:

$$T^2 = \left[(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)\right]'(\frac{1}{n_1}S_1^2+\frac{1}{n_2}S_2^2)^{-1}\left[(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)\right] \xrightarrow{\mathscr{L}} \chi_p^2$$

参考文献

[1] 概率导论 (第二版・修订版). Dimitri P. Bertsekas, John N. Tsitsiklis. 人民邮电出版社.

[2] 数理统计（第二版）. 韦来生. 科学出版社.

[3] Statistical Inference(2nd Edition). George Casella, Roger L. Berger. Duxbury Press.

[4] Applied Linear Statistical Models(5th Edition). Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li. McGraw-Hill Compaines, Inc.

[5] 线性模型引论. 王松桂 et. al. 科学出版社.

[6] Linear Models with R(2nd Edition). Julian J. Faraway. CRC Press.

[7] 实用多元统计分析（第六版）. Richard A. Johnson, Dean W. Wichern. 清华大学出版社.

索引