Statistical Modelling - Concise Notes

MATH50011

Arnav Singh

Colour Code - **Definitions** are **green** in these notes, **Consequences** are **red** and **Causes** are **blue**

Content from MATH40005 assumed to be known.

Mathematics
Imperial College London
United Kingdom
February 24, 2022

Contents

1 Statistical Models

1.2 Parametric Statistical Models

Definition 1.1 ***Statistical Model***

Statistical model; collection of probability distribution $\{P_\theta : \theta \in \Theta\}$ on a given sample space.
Set Θ - (**Parameter Space**) - set of all possible parametric values, $\Theta \subset \mathbb{R}^p$

Definition 1.2 ***Identifiablee***

Statistical model is **identifiable** if map $\theta \mapsto P_\theta$, one-to-one, $P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2 \quad \forall \theta_1, \theta_2 \in \Theta$

1.3 Using Models

Requirements for a model

1. Agree with observed data "reasonable" well
2. reasonably simple (no excess parameters)
3. easy to interpret (parameter have practical meaning)

2 Point Estimation

Definition 2.1 ***Statistic***

Statistic - function of observable random variable.

Definition 2.2 ***Estimate/Estimators***

t a statistic
$t(y_1, \ldots, y_n)$ called **estimate** of θ
$T(Y_1, \ldots, Y_n)$ an **estimator** of Θ

2.1 Properties of estimators

2.1.1 Bias

Definition 2.3 ***Bias***

T estimator for $\theta \in \Theta \subset \mathbb{R}$

$$bias_\theta(T) = E_\theta(T) - \theta$$

unbiased if $bias_\theta(T) = 0, \quad \forall \theta \in \Theta$

If $\Theta \subset \mathbb{R}^k$ often interested in $g(\theta)$, $g : \theta \to \mathbb{R}$

$$\text{extend } bias_\theta(T) = E_\theta(T) - g(\theta)$$

2.1.2 Standard error

Definition 2.4

T estimator for $\theta \in \Theta \subset \mathbb{R}$

$$SE_\theta(T) = \sqrt{Var_\theta(T)}$$

Standard error, is standard deviation of sampling distribution of T

2.1.3 Mean Square Error

Definition 2.5

T estimator for $\theta \in \Theta \subset \mathbb{R}$
Mean square error of T

$$\begin{aligned} MSE_\theta(T) &= E_\theta(T - \theta)^2 \\ &= Var_\theta(T) + [bias_\theta(T)]^2 \end{aligned}$$

3 The Cramér-Rao Lower Bound

Theorem 3.1 *(Cramér-Rao Lower Bound)*

$T = T(X)$ unbiased estimator for $\theta \in \Theta \subset \mathbb{R}$ for $X = (X_1, \ldots, X_n)$ with just pdf $f_\theta(x)$ under mild regularity conditions:

$$Var_\theta(T) \geq \frac{1}{I(\theta}$$

For I_θ the **Fisher information of sample**

$$\begin{aligned} I(\theta) &= E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f_\theta(x)\right\}^2\right] \\ &= -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\right] \\ I_n(\theta) &= -nE_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\right] \end{aligned}$$

Proposition.

For a random sample: Fisher info proportional to sample size

Jensen's inequality

For X a random variable with φ a convex function

$$\varphi(E[X]) \leq E\left[\varphi(X)\right]$$

Call $E\left[\varphi(X)\right] - \varphi\left(E[X]\right)$ the **Jensen gap**

4 Asymptotic Properties

Definition 4.1

Sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ called **(weakly) consistent** if $\forall\theta \in \Theta$

$$T_n \xrightarrow{P_\theta} g(\theta) \quad (n \to \infty)$$

Definition 4.2

Convergence in probability: $T_n \xrightarrow{P_\theta} g(\theta)$

$$\forall\epsilon > 0 : \lim_{n\to\infty} P_\theta(|T_n - g(\theta)| < \epsilon) = 1$$

Lemma - (Portmanteau Lemma)

X, X_n real valued random value.
Following are equivalent:

1. $X_n \to X$ as $n \to \infty$
2. $E[f(X_n)] \to E[f(X)] \quad n \to \infty$ for all bounded + continuous functions $f : \mathbb{R} \to \mathbb{R}$

Definition 4.3

Sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ **asymptotically unbiased** if $\forall\theta \in \Theta$

$$E_\theta \to g(\theta) \quad n \to \infty$$

Lemma.

(T_n) asymptotically unbiased for $g(\theta)$ and $\forall\theta \in \Theta$

$$Var_\theta(T_n) \to 0 \quad n \to \infty$$

$\implies (T_n)$ consistent for $g(\theta)$

Definition 4.4

Sequence (T_n) of estimators for $\theta \in \mathbb{R}$ **asymptotically normal** if

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$$

for some $\sigma^2)(\theta)$

Theorem 4.1 *(Central Limit Theorem)*

$Y_1, \dots, Y_n$ be iid random variable with $E(Y_i) = \mu,\ Var(Y_i) = \sigma^2$

$$\implies \text{ sequence } \sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Remark.
Under mild regularity conditions for asymptotically normal estimators T_n

$$SE_\theta(T_n) \approx \frac{\sigma(T_n)}{\sqrt{n}}$$

Lemma. *(Slutsky)*

X_n, X, Y_n random variables

If $X_n \underset{d}{\to} X$ and $Y_n \underset{p}{\to} c$ for constant c

1. $X_n + Y_n \underset{d}{\to} X + c$
2. $Y_n X_n \underset{d}{\to} cX$
3. $Y_n^{-1} X_n \underset{d}{\to} c^{-1} X \quad$ provided $c \neq 0$

Theorem 4.2 *(Delta Method)*

Suppose T_n asymptotically normal estimator of θ with

$$\sqrt{n}(T_n - \theta) \underset{d}{\to} N(0, \sigma^2(\theta))$$

$g : \Theta \to \mathbb{R})$ differentiable function with $g'(\theta) \neq 0$. Then

$$\sqrt{n}[g(T_n) - g(\theta)] \underset{d}{\to} N(0, g'(\theta)^2 \sigma^2(\theta))$$

Theorem 4.3 *(Continuous Mapping Theorem)*

$k, m \in \mathbb{N}, X, X_n, \quad \mathbb{R}^k-$valued random variable.
$g : \mathbb{R}^k \to \mathbb{R}^m$ continuous function at every point of C s.t $P(X \in C) = 1$

- If $X_n \underset{d}{\to} X \implies g(X_n) \underset{d}{\to} g(x)$ as $n \to \infty$
- If $X_n \underset{p}{\to} X \implies g(X_n) \underset{p}{\to} g(X)$ as $n \to \infty$
- If $X_n \underset{a.s}{\longrightarrow} X \implies g(X_n) \underset{a.s}{\longrightarrow} g(X)$ as $n \to \infty$

5 Maximum Likelihood Estimation

Definition 5.1 *(Likelihood function)*

Suppose observer Y with realisation y
Likelihood function

$$L(\theta) = L(\theta : y) = \begin{cases} P(Y = y : \theta) & \text{discrete data} \\ f_Y(y : \theta) & \text{absolutely continuous data} \end{cases}$$

Likelihood function is the joint pdf/pmf or observed data as a function of unknown parameter.

Random sample $Y = (Y_1, \dots, Y_n)$ $\quad Y_i$ iid.
If Y_i has pdf $f(\cdot; \theta)$

$$\implies L(\theta) = \prod_{i=1}^{n} f(y_i : \theta)$$

Definition 5.2 *(Maximum Likelihood Estimator)*

MLE of θ is estimator $\hat{\theta}$ s.t

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$$

5.1 Properties of Maximum Likelihood estimators

5.1.1 MLEs functionally invariant

g bijective function
$\hat{\theta}$ MLE of $\theta \implies \hat{\phi} = g(\hat{\theta})$ a MLE of $\phi = g(\theta)$

5.1.2 Large Sample property

Theorem 5.1

$X_1, X_2, \dots$ iid observations with pdf/pmf f_θ
$\theta \in \Theta, \;\Theta$ an open interval
$\theta_0 \in \Theta$ - true parameter.

Under regularity conditions ($\{x : f_\theta(x) > 0\}$ indpendent of θ). We have

1. $\exists$ consistent sequence $(\hat{\theta})_{n \in \mathbb{N}}$ of MLE
2. $(\hat{\theta})_{n \in \mathbb{N}}$ consistent sequence of MLEs $\implies \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_f(\theta_0))^{-1})$ $\quad$ *(Asymptotic normality of MLE)*
   Where $I_f\theta$ Fisher information of sample size $= 1$

Remark: if MLE unique $(\forall n) \implies$ sequence of MLEs consistent

Remark
Limiting distribution depends on $I_f(\theta_0)$, which is often unknown in practical situations. $\implies$ need to estimate $I_f(\theta_0)$

iid sample; $I_f(\theta_0)$ estimated by

- $I_f(\hat{\theta})$
- $\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \log(f(x_i : \theta))|_{\theta = \hat{\theta}} \right)^2$
- $-\frac{1}{n} \sum_{i=1}^{n} (\frac{\partial}{\partial \theta})^2 \log(f(x_i : \theta))|_{\theta = \hat{\theta}}$

Often consistent $\implies$ converge to $I_f(\theta_0)$ in probability
Remark

Standard error of asymptotically normal MLE $\hat{\theta}_n$
Approximated by $SE(\hat{\theta}_n) = \sqrt{\hat{I}_n^{-1}} / \sqrt{n}$ $\;\hat{I}_n$ estimator from above.

Remark - Multivariate version.
$\Theta \subset \mathbb{R}^k$ open set, $\hat{\theta}_n$ MLE based on n observation.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_f(\theta_{\mathbf{0}})^{-1})$$

$\theta_{\mathbf{0}}$ the true parameter, $I_f(\theta)$ **Fisher information matrix**

$$\begin{aligned} I_f(\theta) &:= E_\theta\left[(\nabla \log f(X;\theta))^T (\nabla \log f(X;\theta))\right] \\ &:= -E_\theta\left[\nabla^T \nabla \log f(X:\theta)\right] \end{aligned}$$

Definition 5.3

Converges in distribution for random vector
$\mathbf{X}, \mathbf{X_1}, \mathbf{X_2}$ random vectors of dimension k

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \quad (n \to \infty)$$

If $P(\mathbf{X}_n \leq z) \xrightarrow[n\to\infty]{} P(\mathbf{X} \leq z) \quad \forall z \in \mathbb{R}^k : z \mapsto P(X \leq Z)$ continuous

6 Confidence Regions

Definition 6.1 *(Confidence interval)*

$1-\alpha$ **confidence interval** for θ, a random interval I containing 'true' paramter with probability $\geq 1-\alpha$

$$P_{\theta \in I} \geq 1 - \alpha \quad \forall \theta \in \Theta$$

6.1 Construction of confidence intervals

Definition 6.2

Pivotal Quantity for θ a function $t(Y,\theta)$ of data and θ
s.t distribution of $t(Y,\theta)$ known (no dependency on unknown parameters)
Know distribution of $t(Y,\theta) \implies$ can find constant a_1, a_2 s.t $P(a_1 \leq t(Y_1,\theta) \leq a_2) \geq 1-\alpha$
$\implies P(h_1(Y) \leq \theta \leq h_2(Y)) \geq 1 - \alpha$

Call $[h_1(Y), h_2(Y)]$ a **random interval**
with observed interval $[h_1(y), h_2(y)]$ a $1-\alpha$ **confidence interval for** θ

6.2 Asymptotic confidence intervals

We often know

$$\begin{aligned} &\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)) \\ \implies &\underbrace{\sqrt{n}(\frac{T_n - \theta}{\sigma(\theta)})}_{\text{use as pivotal quantity}} \xrightarrow{d} N(0,1) \end{aligned}$$

Definition 6.3

Sequence of random intervals I_n
an **asymptotic $1-\alpha$ Confidence Interval** if

$$\lim_{n\to\infty} P_\theta(\theta \in_n) \geq 1-\alpha \quad \theta$$

Simplification

Given consistent estimator $\hat{\sigma}_n$ for $\sigma(\theta)$ $\hat{\sigma}_n \xrightarrow{P_\theta} \sigma(\theta)$ $\forall\theta$

$$\sqrt{n}(\frac{T_n-\theta}{\sigma(\theta)}) \xrightarrow{d} N(0,1)$$

$$T_n \pm c_{\alpha/2} \times \underbrace{\frac{\hat{\sigma}_n}{\sqrt{n}}}_{\text{estimates } SE(T_n)}$$

$$T_n \pm c_{\alpha/2} SE(T_n)$$

Simplification.
$\hat{\sigma}^2 = \frac{Y}{n}(1-\frac{Y}{n})$ $\quad\hat{\sigma}^2 \xrightarrow{P} \theta(1-\theta)$

$$\underbrace{\sqrt{n}\frac{Y/n-\theta}{\sqrt{\frac{Y}{n}(1-\frac{Y}{n})}}}_{\text{pivotal quantity}} \implies \frac{y}{n} \pm \frac{c_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{y}{n}(1-\frac{y}{n}}$$

6.3 Simultaneous Confidence Interval/Confidence regions.

Definition 6.4

$\theta = (\theta_1, \dots, \theta_k)^T \in \Theta \in \mathbb{R}^k$
With random intervals $(L_i(\mathbf{Y}), U_i(\mathbf{Y}))$ s.t

$$\forall\theta : P_\theta(L_i(\mathbf{Y} < \theta_i < U_i(\mathbf{Y}), i \in \{1, \dots, k\}) \geq 1-\alpha$$

$(L_i(\mathbf{y}, U_i(\mathbf{y}))$ $i \in \{1, \dots, k\}$ a **$1-\alpha$ simultaneous confidence interval** for $\theta_1, \dots, \theta_k$
Remark - (Bonferroni correction)
take $[L_i, U_i]$ a $1-\alpha$ confidence interval for $\theta_i,\ i \in \{1, \dots, k\}$

7 Hypothesis Testing

7.1 Prelim

Definition 7.1 *(Hypotheses)*

We have 2 complementary hypothesis

- H_0 : Null hypothesis - consider to be the status quo
- H_1: Alternative hypothesis

Definition 7.2 *(Hypthesis Test)*

Hypothesis test a rule that specifies for which valus of a sample $x_1, \dots, x_n$ a decision is to be made

- accept H_0 as true
- reject H_0 and accept H_1

Rejection region/Critical region - subset of sample space for which H_0 rejected

Definition 7.3 *(Types of error)*

	H_0 True	H_0 False
Don't reject H_0	✓	**Type II Error**
Reject H_0	**Type I Error**	✓

7.2 Power of a Test

Definition 7.4 *(Power function)*

Θ parameter space with $\quad \Theta_0 \subset \Theta,\ \Theta_1 = \Theta \backslash \Theta_0$
Consider:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

Given a test for this hypothsis, we have a **Power function**

$$\beta : \theta \to [0,1]$$
$$\beta(\theta) = P_\theta(\text{reject} H_0)$$

$\theta \in \Theta_0 \implies$ want $\beta(\theta)$ small
$\theta \in \Theta_1 \implies$ want $\beta(\theta)$ large

7.3 p-Value

Definition 7.5 *(p-value)*

$$p = \sup_{\theta \in \Theta_0} P_\theta(\text{observing something 'at least as extreme' as the observation})$$

reject $H_0 \iff p \leq \alpha$
For test based on statistic T with rejection for large value of T we have

$$p = \sup_{\theta \in \Theta_0} P_\theta(T \geq t)$$

for t our observed value

7.4 Connection between tests & confidence intervals

7.4.1 Constructing a test from confidence region

Y a random observation.
$A(Y)$ a $1-\alpha$ confidence region for θ

$$P(\theta \in A(Y)) \geq 1 - \alpha \quad \forall \theta \in \Theta$$

Define test for $\begin{matrix} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \notin \Theta_0 \end{matrix}$ for $\Theta_0 \subset \Theta$ a fixed subset with level α s.t

$$\text{Reject } H_0 \text{ if } \Theta_0 \cap A(Y) = \emptyset$$

$$\begin{aligned} P_\theta(\text{Type I error}) = P_\theta(\text{reject}) &= P_\theta(\Theta_0 \cap A(Y) = \emptyset) \\ &\leq P_\theta(\theta \notin A(Y)) \leq \alpha \end{aligned}$$

7.4.2 Constructing confidence region from tests

Suppose $\forall \theta_0 \in \Theta$ we have a level α test ϕ_{θ_0} for

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

A decision rule ϕ_{θ_0} to reject/not reject $H_0^{\theta_0}$ satisfying:

$$P_{\theta_0}(\phi_{\theta_0} \text{ reject } H_0^{\theta_0}) \leq \alpha$$

Consider random set:

$$A := \left\{ \theta_0 \in \Theta : \phi_{\theta_0} \text{ doesn't reject } H_0^{\theta_0} \right\}$$

We see A a $1-\alpha$ confidence region for θ
$\forall \theta \in \Theta\ P_\theta(\theta \in A) = P_\theta(\phi_\theta \text{ not rejects }) = 1 - P_\theta(\phi_\theta \text{ rejects }) \geq 1 - \alpha$

8 Likelihood Ratio Tests

(Numbers don't line up with official notes!!!)

Definition 8.1 *(Likelihood ratio statistic)*

$$t(\mathbf{y}) = \frac{sup_{\theta \in \Theta} L(\theta; \mathbf{y})}{sup_{\theta \in \Theta_0} L(\theta; \mathbf{y})} = \frac{\text{max likelihood under } H_0 + H_1}{\text{max likelihood under } H_0}$$

Theorem 8.1

$X_1, \ldots, X_n \sim N(0,1),\ X_i$ independent

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

Theorem 8.2

Under regularity conditions

$$2 \log t(\mathbf{Y}) \xrightarrow{D} \chi_r^2 \quad (n \to \infty)$$

under H_0 where r the number of independent restrictions on θ needed to define H_0