

Statistics 2 - Reviewed Notes

Dom Hutchinson

January 12, 2020

Contents

1	General	2
1.1	Definitions	2
1.2	Theorems	3
2	Estimation	3
2.1	Likelihood	3
2.2	Estimators	6
2.3	Confidence Sets	6
2.4	Convergence	9
2.5	Performance of Estimators	10
2.6	Asymptotic Distribution of Estimators	14
2.6.1	Confidence Intervals	16
2.7	Efficiency of Estimators	16
3	Testing	17
3.1	Hypothesis Testing	17
3.1.1	Neyman-Pearson Approach	20
3.1.2	Generalised Hypothesis Testing	22
3.2	Categorical Distributions & Pearson's χ^2 -Test	23
4	Bayesian Inference	25
4.1	Credible Intervals	26
4.2	Bayesian Hypothesis Testing	27
0	Appendix	28
0.1	Notation	28
0.2	Definitions	28
0.3	Theorems	30
0.4	Probability Distributions	31
0.5	Identities	34
0.5.1	Likelihood	34

1 General

1.1 Definitions

Definition 1.1 - Probability Space, $(\Omega, \mathcal{F}, \mathbb{P})$

A *Probability Space* is a mathematical construct for modelling the real world. A *Probability Space* has three elements

- i) Ω , Sample space;
- ii) \mathcal{F} , Set of events; and,
- iii) \mathbb{P} , Probability Measure

and must fulfil the following criteria

- i) $\Omega \in \mathcal{F}$;
- ii) $\forall A \in \mathcal{D} \implies A^c \in \mathcal{F}$;
- iii) $\forall A_0, \dots, A_n \in \mathcal{F} \implies \left(\bigcup_{i=0}^n A_i \right) \in \mathcal{F}$;
- iv) $\mathbb{P}(\Omega) = 1$; and,
- v) $\mathbb{P}\left(\bigcup_{i=0}^n A_i\right) = \sum_{i=0}^n \mathbb{P}(A_i)$ for any n disjoint A_0, \dots, A_n .

Definition 1.2 - Random Variable

A *Random Variable* is a function which maps an event in the sample space to a value. X is a random variable if it satisfies the signature

$$X : \Omega \rightarrow \mathbb{R}$$

Definition 1.3 - Parametric Models

Parametric Models are the class of statistical distributions whose probability mass/density function take parameters. These parameters represent values of interest in the population, such as mean or variance. We generally do not know these values so we estimate them from samples.

Definition 1.4 - Quantity of Interest

When analysing distributions it often helps to define *Quantities of Interest* about the distributions (*e.g.* Mean). These are defined as functions in terms of the parameters $\tau(\theta)$. We estimate *Quantities of Interest* by passing estimated values of the parameters $\hat{\tau} = \tau(\hat{\theta})$.

Definition 1.5 - Frequentist Approach

The *Frequentist Approach* to probability is an interpretation of probability where *Probability* refers to the limiting relative frequencies of events. *Probabilities* are objective properties of the world.

$$\mathbb{P}(X = x) = \lim_{n \rightarrow \infty} \frac{k}{n}$$

where k is the number of times x is observed in n samples.

N.B. Most of this module follows this approach.

Definition 1.6 - Bayesian Approach

The *Frequentist Approach* to probability is an interpretation of probability where *Probability* is a reasonable expectation given our beliefs about the system so we can model features beyond the data. We encode our beliefs using the components of *Bayes' Theorem*.

1.2 Theorems

Theorem 1.1 - *Samples from a Normal Distribution are χ^2 Distributed*

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$. Then

$$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

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

Theorem 1.2 - *Distance between Sample Mean & Population Mean is t_r Distributed*

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$. Then

$$\frac{\sqrt{n}}{s}(\bar{X} - \mu) \sim t_{n-1}$$

N.B. We don't need to know σ^2 to estimate the distance between \bar{X} and μ .

Theorem 1.3 - *Multidimension Transform of a Random Variable*

Consider an n -dimensional *continuous* random variable $\mathbf{X} \sim f_{\mathbf{X}}(\cdot)$ which we wish to transform.

Define a continuously differentiable bijective function $\mathbf{g} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $\mathbf{h} := \mathbf{g}^{-1}$.

Then if $\mathbf{Y} := \mathbf{g}(\mathbf{X}) \sim f_{\mathbf{Y}}(\cdot)$ we have

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{h}(\mathbf{y}))J_{\mathbf{h}}(\mathbf{y})$$

where $J_{\mathbf{h}}(\mathbf{y}) := \left| \det \left(\frac{\partial \mathbf{h}}{\partial \mathbf{y}} \right) \right| = \left| \det \begin{pmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{pmatrix} \right|$.

Theorem 1.4 - *Weak Law of Large Numbers*

Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of independent & identically distributed random variables.

If $\mathbb{E}(X_i) = \mu < \infty$ then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mu$$

Theorem 1.5 - *Central Limit Theorem*

Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of independent & indetically distributed with $\mathbb{E}(X_i) = \mu < \infty$ and

$\text{Var}(X_i) = \sigma^2 < \infty$. Then

$$\sqrt{\frac{n}{\sigma^2}}(Z_n - \mu) \rightarrow_{\mathcal{D}} Z \sim \text{Normal}(0, 1)$$

2 Estimation

2.1 Likelihood

Definition 2.1 - *Likelihood Function*

The *Likelihood Function* is a family of functions which measure the likely of a certain realisation of a random variable is given the parameters of a model have a certain value.

$$L(\boldsymbol{\theta}; \mathbf{x}) := C f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) \text{ for } C > 0$$

where $\mathbf{X} \sim f_n(\cdot; \boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^*$ unknown and \mathbf{x} is a realisation of \mathbf{X} .

N.B. *Likelihood Functions* have signature $L(\cdot; \mathbf{x}) : \boldsymbol{\theta} \rightarrow [0, \infty)$.

N.B. This is also known as the *Observed Likelihood Function*.

Definition 2.2 - Log-Likelihood Function

The *Log-Likelihood Function* is the family of functions which are equivalent to the natural log of the *Likelihood Function*.

$$\ell(\theta; \mathbf{x}) := \ln f_n(\mathbf{x}; \theta) + C \text{ for } \underbrace{C}_{\equiv \ln C} \in \mathbb{R}$$

N.B. This is increasing with $L(\cdot; \mathbf{x})$.

Remark 2.1 - Likelihood for Independent & Identically Distributed Random Variables

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$ and \mathbf{x} be a realisation of \mathbf{X} . Then

$$\begin{aligned} L_n(\theta; \mathbf{x}) &:= \prod_{i=1}^n L(\theta; x_i) \\ \ell_n(\theta; \mathbf{x}) &:= \sum_{i=1}^n \ell(\theta; x_i) \end{aligned}$$

Theorem 2.1 - The Likelihood Function is Invariant under Bijective Transformations which are independent of Model Parameters

Consider $\mathbf{X} \sim f_{\mathbf{X}}(\cdot; \theta)$ and $\mathbf{g} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a bijective function which is independent of θ . Define $\mathbf{Y} := \mathbf{g}(\mathbf{X}) \sim f_{\mathbf{Y}}(\cdot; \theta)$. Then

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) \propto f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}); \theta)$$

Hence

$$L_{\mathbf{Y}}(\theta; \mathbf{g}(\mathbf{x})) \propto L_{\mathbf{X}}(\theta; \mathbf{x})$$

Proof 2.1 - Theorem 2.1

Consider $\mathbf{X} \sim f_{\mathbf{X}}(\cdot; \theta)$ and $\mathbf{g} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a bijective function which is independent of θ . Define $\mathbf{h} := \mathbf{g}^{-1}$ and $\mathbf{Y} := \mathbf{g}(\mathbf{X})$.

We consider the cases where \mathbf{X} is discrete & continuous independently

Discrete Case Let \mathbf{X} be a discrete random variable. Then

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \theta) &= \mathbb{P}(\mathbf{Y} = \mathbf{y}; \theta) \\ &= \mathbb{P}(\mathbf{g}^{-1}(\mathbf{Y}) = \mathbf{g}^{-1}(\mathbf{y}); \theta) \\ &= \mathbb{P}(\mathbf{h}(\mathbf{Y}) = \mathbf{h}(\mathbf{y}); \theta) \\ &= \mathbb{P} * \mathbf{X} = \mathbf{h}(\mathbf{y}); \theta \\ &= f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}); \theta) \end{aligned}$$

Continuous Case Let \mathbf{X} be a continuous random variable.

By **Theorem 1.3**

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}); \theta) J_{\mathbf{g}^{-1}}(\mathbf{y})$$

Since $J_{\mathbf{g}^{-1}}$ is independent of θ this case is solved.

In both cases $L_{\mathbf{Y}}(\theta; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \theta) \propto f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y}); \theta) = L_{\mathbf{X}}(\theta; \mathbf{x})$. □

Definition 2.3 - Maximum Likelihood Estimate

Let $\mathbf{X} \sim f_n(\cdot; \theta)$ and \mathbf{x} be a realisation of \mathbf{X} .

The *Maximum Likelihood Estimate* of \mathbf{X} is the value $\hat{\theta} \in \Theta$ which produces the greatest value of the *Likelihood Function* of \mathbf{X} .

$$\hat{\theta}_{\text{MLE}}(\mathbf{x}) := \operatorname{argmax}_{\theta} L(\theta; \mathbf{x}) = \operatorname{argmax}_{\theta} \ell(\theta; \mathbf{x})$$

N.B. The *Maximum Likelihood Estimate* is not necessarily unique.

Theorem 2.2 - Maximum Likelihood Estimate of Reparameterisation

Define random variable $\tau = g(X)$ where $g : \mathbb{R} \rightarrow \mathbb{R}$. Then

$$\hat{\tau}_{\text{MLE}} = \tau(\hat{\theta}_{\text{MLE}})$$

Proof 2.2 - Theorem 2.2

This is a proof by contradiction.

Suppose $\exists \tau^* \in G$ st $\tilde{f}(x; \tau^*) > \tilde{f}(x; \tau)$.

We know that $\forall \theta \in \Theta$, $f(x; \theta) = \tilde{f}(x; g(\theta))$ and $\forall \tau \in G$, $f(x; g^{-1}(\tau)) = \tilde{f}(x; \tau)$.

We deduce that

$$\begin{aligned} f(x; g^{-1}(\tau^*)) &= \tilde{f}(x; \tau^*) \\ &> \tilde{f}(x; \hat{\tau}) \text{ by assumption} \\ &= f(x; g^{-1}(\hat{\tau})) \\ &= f(x; \hat{\theta}) \end{aligned}$$

This contradicts the assumption that $\hat{\theta}$ is an maximum likelihood estimate of θ . □

Remark 2.2 - Finding Maximum Likelihood Estimates - Multivariate

Let $X \sim f_X(\cdot; \theta)$ be continuous random variable where $f_X(\cdot)$ is differentiable and θ is an n -dimensional parameter.

Let \mathbf{x} be a realisation of \mathbf{X} .

To find a *Maximum Likelihood Estimate* for θ

- i) Find the gradient of $\ell(\theta; \mathbf{x})$ wrt θ .

$$\nabla \ell(\theta; \mathbf{x}) := \begin{pmatrix} \frac{\partial}{\partial \theta_1} \ell(\theta; \mathbf{x}) & \dots & \frac{\partial}{\partial \theta_n} \ell(\theta; \mathbf{x}) \end{pmatrix}$$

- ii) Equate $\nabla \ell(\theta; \mathbf{x})$ to the zero-vector and solve for each θ to find extrema of ℓ .

$$\nabla \ell(\theta; \mathbf{x}) = \mathbf{0}$$

- iii) Calculate the *Hessian* of $\ell(\theta; \mathbf{x})$

$$\nabla^2 \ell(\theta; \mathbf{x}) = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} \ell(\theta; \mathbf{x}) & \dots & \frac{\partial^2}{\partial \theta_1 \partial \theta_n} \ell(\theta; \mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_n \partial \theta_1} \ell(\theta; \mathbf{x}) & \dots & \frac{\partial^2}{\partial \theta_n^2} \ell(\theta; \mathbf{x}) \end{pmatrix}$$

- iv) Test each extremum $\hat{\theta}$ to see if it is a maximum

If $\det(H(\hat{\theta})) > 0$ and $\frac{\partial}{\partial \theta_1^2} \ell(\hat{\theta}; \mathbf{x}) < 0$ then $\hat{\theta}$ is a local maximum.

i.e. Check $H(\hat{\theta})$ is *negative definite*.

Definition 2.4 - Likelihood Ratio

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$ for $\theta^* \in \Theta$ and $\{\hat{\theta}_i\}_{i \in \mathbb{N}}$ be a sequence of consistent *Maximum Likelihood Estimators* of $\theta^* \in \Theta$.

We define the *Likelihood Ratio* as

$$\Lambda_n(\mathbf{x}) := \frac{L(\theta^*; \mathbf{x})}{L(\hat{\theta}_n; \mathbf{x})} \in [0, 1] \text{ for } \mathbf{x} \in \mathcal{X}^n$$

Theorem 2.3 - Asymptotic Distribution of Likelihood Ratio

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$ for $\theta^* \in \Theta$ and $\{\hat{\theta}_i\}_{i \in \mathbb{N}}$ be a sequence of consistent *Maximum Likelihood Estimators* of $\theta^* \in \Theta$.

Suppose the conditions of **Theorem 2.13** hold (i.e. X_n is asymptotically normal). Then

$$-2 \ln \Lambda_n(\mathbf{X}_n) \rightarrow_{\mathcal{D}(\cdot; \theta^*)} \chi_1^2$$

2.2 Estimators**Definition 2.5 - Estimation**

Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ with $\theta^* \in \Theta$ and \mathbf{x} be a realisation of \mathbf{X} .

As *Estimation* of model parameter θ^* is a statistic, $\hat{\theta}(\mathbf{x}) = T(\mathbf{x})$, which is intended to approximate the true value of θ^* .

N.B. Interchangeable with *Estimate*.

Definition 2.6 - Estimator

Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ with $\theta^* \in \Theta$ and \mathbf{x} be a realisation of \mathbf{X} .

An *Estimator* of model parameter θ^* is the random variable $\hat{\theta} := \hat{\theta}(\mathbf{X})$ where $\hat{\theta}(\mathbf{x})$ is an *estimation* of θ^* .

Definition 2.7 - Bias

The *Bias* of an *Estimator*, $\hat{\theta}$, is its expected error.

i.e. By how much an estimator consistently deviates from the true value of the parameter).

Let θ^* be the true value of parameter θ . Then

$$\begin{aligned} \text{Bias}(\hat{\theta}; \theta^*) &:= \mathbb{E}(\hat{\theta} - \theta^*; \theta^*) \\ &= \mathbb{E}(\hat{\theta}; \theta^*) - \theta^* \end{aligned}$$

N.B. An *Estimator* is *Unbiased* if $\forall \theta \in \Theta \text{ Bias}(\theta^*; \theta) = 0 \iff \mathbb{E}(\hat{\theta}; \theta) = \theta$.

Definition 2.8 - Mean Square Error

The *Mean Square Error* of an *Estimator*, $\hat{\theta}$, measures the average of its square error.

Let θ^* be the true value of parameter θ . Then

$$\begin{aligned} \text{MSE}(\hat{\theta}; \theta^*) &:= \mathbb{E} \left[(\hat{\theta}(\mathbf{X}) - \theta^*)^2; \theta^* \right] \\ &= \text{Var}(\hat{\theta}; \theta^*) + \text{Bias}(\hat{\theta}; \theta^*)^2 \end{aligned}$$

Definition 2.9 - Distribution of an Estimator

Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ with $\theta^* \in \Theta \subseteq \mathbb{R}$.

Let $\hat{\theta}(\mathbf{X})$ be a real-valued *Estimator* of θ^* . Then

$$\begin{aligned} F_{\hat{\theta}(\mathbf{X})}(t; \theta^*) &:= \mathbb{P}(\hat{\theta}(\mathbf{X}) \leq t; \theta^*) \\ &= \int_{\mathcal{X}^n} \mathbb{1}\{\hat{\theta}(\mathbf{x}) \leq t\} f_n(\mathbf{x}; \theta^*) d\mathbf{x} \end{aligned}$$

N.B. The distribution of an *Estimator* depends on the true value of the parameter it is estimating.

N.B. As sample size increases the distribution of an estimator should converge to a more standard distribution.

2.3 Confidence Sets**Definition 2.10 - Random Interval**

A *Random Interval* is an interval of values which depends on a random variable and thus does not have fixed values.

$$\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$$

Definition 2.11 - Observed Confidence Interval

Let \mathbf{X} be a random variable, $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$ and \mathbf{x} be a realisation of \mathbf{X} .
 $\mathcal{I}(\mathbf{x}) = [L(\mathbf{x}), U(\mathbf{x})]$ is an *Observed Confidence Interval*.

Definition 2.12 - Coverage of an Interval

Let $\mathbf{X} \sim f_n(\cdot; \theta)$ for $\theta \in \Theta = \mathbb{R}$.

Define $L : \mathcal{X}^n \rightarrow \Theta$ & $U : \mathcal{X}^n \rightarrow \Theta$ st $\forall \mathbf{x} \in \mathcal{X}^n, L(\mathbf{x}) < U(\mathbf{x})$.

The *Coverage* of the *Random Interval* $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$ at θ is defined to be

$$C_{\mathcal{I}}(\theta) := \mathbb{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})]; \theta)$$

N.B. *Coverage* is the probability that a realisation of a random variable lies in a given random interval for a given parameter value.

Definition 2.13 - Confidence Interval

Let $\alpha \in [0, 1]$ and $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$ be a random interval.

We say that $\mathcal{I}(\mathbf{X})$ is a $1 - \alpha$ *Confidence Interval* if

$$\forall \theta \in \Theta, C_{\mathcal{I}}(\theta) \geq 1 - \alpha$$

N.B. $\mathcal{I}(\mathbf{X})$ is an Exact *Confidence Interval* if $\forall \theta \in \Theta, C_{\mathcal{I}}(\theta) = 1 - \alpha$.

Proposition 2.1 - Transformed Confidence Interval

Let $\mathbf{X} \sim f(\cdot; \theta^*)$ for $\theta^* \in \Theta$ and $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$ be a confidence interval for θ^* .

Let $\tau := g(\theta)$ be a bijective, continuously differential function. If

- $g(\cdot)$ is **increasing** then $[L(\mathbf{x}), U(\mathbf{x})] = [g(L(\mathbf{x})), g(U(\mathbf{x}))]$.
- $g(\cdot)$ is **decreasing** then $[L(\mathbf{x}), U(\mathbf{x})] = [g(U(\mathbf{x})), g(L(\mathbf{x}))]$.

Proposition 2.2 - Confidence Interval for Reparameterisations

Let $\mathbf{X}_n \sim f(\cdot; \theta^*)$ for $\theta^* \in \Theta \subseteq \mathbb{R}$ and $\tau_n := g(\theta)$ be a bijective & continuously differentiable function.

When \mathbf{X}_n is a regular statistical model we have

$$\sqrt{n\tilde{I}(\tau^*)}(\hat{\tau}_n - \tau^*) \rightarrow_{\mathcal{D}(\cdot; \tau^*)} Z \sim \text{Normal}(0, 1)$$

which leads to the *Confidence Interval*

$$\tilde{\mathcal{I}}(\mathbf{X}) := [\tilde{L}(\mathbf{X}), \tilde{U}(\mathbf{X})] \text{ where } \tilde{L}(\mathbf{X}) = \hat{\tau}_n - z_{\alpha/2} \sqrt{\frac{g'(\theta^*)^2}{nI(\theta^*)}} \text{ and } \tilde{U}(\mathbf{X}) = \hat{\tau}_n + z_{\alpha/2} \sqrt{\frac{g'(\theta^*)^2}{nI(\theta^*)}}$$

N.B. This confidence interval is **not** necessarily the same as transforming $[L(\mathbf{x}), U(\mathbf{x})]$ directly.

Proposition 2.3 - Confidence Intervals with unknown variance, σ^2

When variance, σ^2 , is unknown we can define a consistent sequence of estimators $\{\hat{\sigma}_n^2\}_{n \in \mathbb{N}}$

$$\hat{\sigma}_n^2 := \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \hat{\mu}_n)^2$$

Definition 2.14 - Wald's Approach

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$ for $\theta^* \in \Theta \subset \mathbb{R}$.

Using *Wald's Approach* we can define a confidence interval for θ^* using the asymptotic distribution of the *Maximum Likelihood Estimator* for θ^* .

$$\mathcal{I}(\tau^*) := [L(\mathbf{X}), U(\mathbf{X})] \text{ where } L(\mathbf{x}) := \hat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{nI(\theta^*)}} \text{ and } U(\mathbf{x}) = \hat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{nI(\theta^*)}}$$

When we have a realisation we use the *Observed Fisher Information*

$$L(\mathbf{x}) := \hat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{nJ(\theta^*)}} \text{ and } U(\mathbf{x}) = \hat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{nJ(\theta^*)}}$$

N.B. This definition ensures that as $\mathbb{P}(\theta \in \mathcal{I}(\mathbf{X})) \xrightarrow[n \rightarrow \infty]{} 1 - \alpha$.

Remark 2.3 - Limitations of Wald's Approach

Let $\mathcal{I}(\theta^*)$ be a *Confidence Interval* defined using *Wald's Approach*.

There are certain limitations of *Wald's Approach*

- i) It is possible $\exists \theta \notin \mathcal{I}(\theta^*)$ st $\exists \theta' \in \mathcal{I}(\theta^*)$ where $L(\theta; \mathbf{x}) > L(\theta'; \mathbf{x})$.
- ii) It is possible $\exists \theta \in \mathcal{I}(\theta^*)$ where $L(\theta; \mathbf{x}) = 0$.
- iii) *Wald Confidence Intervals* are not invariant under reparameterisation.

Definition 2.15 - Confidence Set

Let $\mathbf{X}_n \stackrel{\text{iid}}{\sim} f_n(\cdot; \theta^*)$ for $\theta^* \in \Theta$ and $\hat{\theta}_n$ be an estimator of θ .

Confidence Sets for θ^* are the possible values of θ whoses likelihood is close to that of the *Maximum Likelihood Estimate* of θ .

Confidence Sets are not necessarily contiguous.

$$C(\mathbf{X}_n) := \left\{ \theta \in \Theta : \ell(\hat{\theta}_n; \mathbf{X}_n) - \ell(\theta; \mathbf{X}_n) \leq \frac{1}{2} \chi_{1,\alpha}^2 \right\} \subseteq \Theta$$

Confidence Interval sets are asymptotically $1 - \alpha$ for θ^* since

$$\mathbb{P}(\theta^* \in C(\mathbf{X}_n); \theta^*) \xrightarrow[n \rightarrow \infty]{} 1 - \alpha$$

N.B. This definition and result are applications of **Definition 2.4** & **Theorem 2.3**.

N.B. *Confidence Sets* are hard to define explicitly without a computer.

N.B. This is known as *Wilk's Approach*.

Theorem 2.4 - Confidence Set of Reparameterisation

Let $\mathbf{X} \sim f(\cdot; \theta^*)$ for $\theta^* \in \Theta$ and $\tau := g(\theta)$ where $g : \Theta \rightarrow G$ is a bijection.

Let $C(\mathbf{x})$ be a confidence set for θ^* and $\tilde{C}(\mathbf{x})$ be a confidence set for τ^* . Then

$$\forall \mathbf{x} \in \mathcal{X}^n, \theta^* \in \Theta \text{ we have } \theta \in C(\mathbf{x}) \iff g(\theta) \in \tilde{C}(\mathbf{x})$$

$$\text{N.B. } \tilde{C}(\mathbf{x}) := \left\{ \theta \in \Theta : \tilde{\ell}_n(\hat{\theta}_n; \mathbf{x}) - \tilde{\ell}(\theta; \mathbf{x}) \leq \frac{1}{2} \chi_{1,\alpha}^2 \right\}.$$

Proof 2.3 - Theorem 2.4

Let $\mathbf{x} \in \mathcal{X}^n$ be arbitrary.

Everything rests on the observation that

$$\forall \theta \in \Theta, \ell(\theta; \mathbf{x}) = \ln f(\mathbf{x}; \theta) = \ln f(\mathbf{x}; g(\theta)) = \tilde{\ell}(g(\theta); \mathbf{x})$$

and similary

$$\forall \tau \in G, \tilde{\ell}(\tau; \mathbf{x}) = \ln \tilde{f}(\mathbf{x}; \tau) = \ln f(\mathbf{x}; g^{-1}(\tau)) = \ell(g^{-1}(\tau); \mathbf{x})$$

Note that $g(\hat{\theta}_n)$ is the *Maximum Likelihood Estimate* of τ .

Assume $\theta \in C(\mathbf{x})$. Then

$$-2 \left[\ell(\theta; \mathbf{x}) - \ell(\hat{\theta}_n; \mathbf{x}) \right] \leq \chi_{1,\alpha}^2$$

Thus

$$-2 \left[\tilde{\ell}(g(\theta); \mathbf{x}) - \tilde{\ell}(g(\hat{\theta}_n); \mathbf{x}) \right] \leq \chi_{1,\alpha}^2$$

Thus $g(\theta) \in \tilde{C}(\mathbf{x})$.

So $\theta \in C(\mathbf{x}) \implies g(\theta) \in \tilde{C}(\mathbf{x})$.

Similarly, assume that $g(\theta) \in \tilde{C}(\mathbf{x})$. Thus

$$-2 \left[\ell(\theta; \mathbf{x}) - \ell(\hat{\theta}_n; \mathbf{x}) \right] \leq \chi_{1,\alpha}^2$$

Thus $\theta \in C(\mathbf{x})$.

So $\theta \in C(\mathbf{x}) \iff g(\theta) \in \tilde{C}(\mathbf{x})$.

For the last part, this correspondence implies that

$$\{\mathbf{x} \in \chi^n; \theta^* \in C(\mathbf{x})\} = \{\mathbf{x} \in \chi^n : g(\theta^*) \in \tilde{C}(\mathbf{x})\}$$

Thus, we can conclude from the equivalence of the events

$$\{\theta^* \in C(\mathbf{X}) = \{g(\theta^*) \in \tilde{C}(\mathbf{X})\}$$

Remark 2.4 - Confidence Set Rule of Thumb

Under the conditions of **Theorem 2.3** there is a rule of thumb that

$$\mathbb{P}(\theta^* \in C(\mathbf{x})) \approx 0.95 \text{ where } C \approx \left\{ \theta \in \Theta : \ell(\hat{\theta}_n; \mathbf{x}) - \ell(\theta; \mathbf{x}) \leq 2 \right\}$$

2.4 Convergence

Definition 2.16 - Convergence

Let $\{z_n\}_{n \in \mathbb{N}}$ be a deterministic sequence of real values and $z \in \mathbb{R}$.

We say $\{z_n\}$ *converges* to limit z if

$$\forall \varepsilon > 0 \exists n_0 \in \mathbb{N} \text{ st } \forall n \geq n_0 \quad |z_n - z| \leq \varepsilon$$

N.B. This is the same for vectors.

Definition 2.17 - Convergence in Probability

Let $\{Z_n\}_{n \in \mathbb{N}}$ be a sequence of random variables and Z be a random variable in the same probability space.

We say that $\{Z_n\}_{n \in \mathbb{N}}$ *Converges in Probability* to Z if

$$\forall \varepsilon > 0 \quad \lim_{n \rightarrow \infty} \mathbb{P}(|Z_n - Z| > \varepsilon) = 0$$

N.B. This is denoted as $Z_n \rightarrow_{\mathbb{P}} Z$.

Definition 2.18 - Convergence in Distribution

Let $\{Z_n\}_{n \in \mathbb{N}}$ be a sequence of random variables and Z be a random variable, not necessarily in the same probability space.

We say $\{Z_n\}_{n \in \mathbb{N}}$ *Converges in Distribution* to Z if

$$\forall z \in Z \text{ where } F_Z(z) \text{ is continuous } \lim_{n \rightarrow \infty} F_{Z_n}(z) = F_Z(z)$$

i.e. F_{X_n} converges in value to F_X as $n \rightarrow \infty$.

N.B. This is denoted as $Z_n \rightarrow_{\mathcal{D}} Z$.

Definition 2.19 - *Convergence in Quadratic Mean*

Let $\{Z_n\}_{n \in \mathbb{N}}$ be a sequence of random variables and Z be a random variable, not necessarily in the same probability space.

We say $\{Z_n\}_{n \in \mathbb{N}}$ *Converges in Quadratic Mean* to Z if

$$\lim_{n \rightarrow \infty} \mathbb{E}[(Z_n - Z)^2] = 0$$

N.B. This is denoted as $Z_n \rightarrow_{\text{qm}} Z$.

Theorem 2.5 - $Z_n \rightarrow_{\mathbb{P}} Z \implies Z_n \rightarrow_{\mathcal{D}} Z$

Theorem 2.6 - $Z_n \rightarrow_{\text{qm}} Z \implies Z_n \rightarrow_{\mathbb{P}} Z$

Theorem 2.7 - $Z_n \rightarrow_{\mathbb{P}} a \iff Z_n \rightarrow_{\mathcal{D}} a$ for $a \in \mathbb{R}$

Theorem 2.8 - *Continuous Mapping Theorem*

Let $\{Z_n\}_{n \in \mathbb{N}}$ be a sequence of random variables and Z be a random variable.

Let $g : Z \rightarrow G$ be a function which maps from the space of random variable Z to a space G .

Then

- i) If $Z_n \rightarrow_{\mathbb{P}} Z$ then $g(Z_n) \rightarrow_{\mathbb{P}} g(Z)$.
- ii) If $Z_n \rightarrow_{\mathcal{D}} Z$ then $g(Z_n) \rightarrow_{\mathcal{D}} g(Z)$.

Theorem 2.9 - *Slutsky's Theorem*

Let $\{Y_n\}_{n \in \mathbb{N}}$ & $\{Z_n\}_{n \in \mathbb{N}}$ be sequences of random variables, Y be a random variable & $c \in \mathbb{R} \setminus \{0\}$. If $Y_n \rightarrow_{\mathcal{D}} Y$ and $Z_n \rightarrow_{\mathcal{D}} c$. Then

- i) $Y_n + Z_n \rightarrow_{\mathcal{D}} Y + c$.
- ii) $Y_n Z_n \rightarrow_{\mathcal{D}} Y c$.
- iii) $\frac{Y_n}{Z_n} \rightarrow_{\mathcal{D}} \frac{Y}{c}$.

Definition 2.20 - *Consistent Sequence of Estimators*

Let $\mathbf{X}_n \sim f_n(\cdot; \theta)$ be a random vector and $\{\hat{\theta}_n(\cdot) : \mathcal{X}^n \rightarrow \Theta\}_{n \in \mathbb{N}}$ be a sequence of estimators for θ .

We say $\{\hat{\theta}_n\}$ is *Consistent* if

$$\forall \theta \in \Theta \quad \hat{\theta}_n(\mathbf{X}_n) \rightarrow_{\mathbb{P}(\cdot; \theta)} \theta$$

Theorem 2.10 - $\hat{\theta}_n \rightarrow_{\text{qm}} \theta \implies \{\hat{\theta}_n\}$ is consistent

2.5 Performance of Estimators

Remark 2.5 - *Measuring Performance of an Estimator*

We measure the performance of an estimator $\hat{\theta}$ in terms of variance since its mean should be θ^* and is thus a bad measure.

Lower variance indicates better performance.

Definition 2.21 - *Fisher Information Regularity Conditions*

Define $\Theta \subset \mathbb{R}$ and $f(x; \theta)$ be a probability mass/density function.

If a model fulfils the following criteria then it is sufficiently *regular* for *Fisher Information* to be drawn from it

- i) $\forall x \in \mathcal{X}$ **both** $L'(\theta; x) = \frac{d}{d\theta} f(x; \theta)$ and $L''(\theta; x) = \frac{d^2}{d\theta^2} f(x; \theta)$ *exist*.
- ii) $\forall \theta \in \Theta$ the set $S := \{x \in \mathcal{X} : f(x; \theta) > 0\}$ is *independent* of $\theta \in \Theta$.
- iii) The identity below *exists*

$$\int_S \frac{d}{d\theta} f(x; \theta) dx = \frac{d}{d\theta} \int_S f(x; \theta) dx = 0$$

N.B. Statistical Models which fulfil all these criteria are described as *Regular*.

Definition 2.22 - *Score Function - Single Random Variable*

Let $X \sim f(\cdot; \theta)$ for some $\theta \in \Theta$ and x be a realisation of X .

The *Score Function* measures the sensitivity of the likelihood function wrt the parameter it is estimating.

$$\ell'(\theta; x) := \frac{d}{d\theta} \ell(\theta; x) = \frac{\frac{d}{d\theta} f(x; \theta)}{f(x; \theta)}$$

Definition 2.23 - *Score Function - Independent & Identically Distributed Random Variables*

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$ with $\theta \in \Theta$ and \mathbf{x} be a realisation of \mathbf{X} .

$$\ell'_n(\theta; \mathbf{x}) := \sum_{i=1}^n \frac{d}{d\theta} \ell(\theta; x_i)$$

Proof 2.4 - *By Regularity Conditions* $\mathbb{E}(\ell'(\theta; X); \theta) = 0 \forall \theta \in \Theta$

$$\begin{aligned} \mathbb{E}(\ell'(\theta; X); \theta) &= \int_S \frac{\frac{d}{d\theta} f(x; \theta)}{f(x; \theta)} f(x; \theta) dx \\ &= \int_S \frac{d}{d\theta} f(x; \theta) dx \\ &= \frac{d}{d\theta} \int_S f(x; \theta) dx \\ &= \frac{d}{d\theta} (1) \\ &= 0 \forall \theta \in \Theta \end{aligned}$$

Definition 2.24 - *Fisher Information - Single Random Variable*

Let $X \sim f(\cdot; \theta)$ be an sufficiently regular (see **Definition 2.14**) observable random variable with θ unknown.

Fisher Information measures the amount of information X carries about θ .

$$\begin{aligned} I(\theta) &:= \mathbb{E}(\ell'(\theta; X)^2; \theta) \\ &= \text{Var}(\ell'(\theta; X); \theta) \text{ by } \mathbf{Proof 2.3} \end{aligned}$$

N.B. This is the expectation of the score, squared \equiv The second moment of the score.

Definition 2.25 - *Fisher Information - Independent & Identically Distributed Random Variables*

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$ with $\theta \in \Theta$ and \mathbf{x} be a realisation of \mathbf{X} .

$$\begin{aligned} I_n(\theta) &:= \mathbb{E}(\ell'_n(\theta; \mathbf{X})^2; \theta) \\ &= \text{Var}(\ell'_n(\theta; \mathbf{X}); \theta) \\ &= nI(\theta) \end{aligned}$$

Definition 2.26 - Observed Fisher Information

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$ be a random n -dimensional vector.

The *Observed Fisher Information* at θ is

$$J_n(\theta) = -\frac{1}{n} \ell''(\theta; \mathbf{X}) = -\frac{1}{n} \sum_{i=1}^n \ell''(\theta; X_i)$$

N.B. $\mathbb{E}(J_n(\theta^*; \theta^*)) = I(\theta^*)$. This is a deterministic value, not an expectation like *Fisher Information*.

Theorem 2.11 - Fisher Information of Reparameterisation

Let $X \sim f(\cdot; \theta)$ for $\theta \in \Theta \subseteq \mathbb{R}$ and $\tau := g(\theta)$ be a bijective & continuously differentiable function.

Consider the reparameterisation $\tilde{f}(x; \tau) := f(x; g(\theta)) = f(x; g^{-1}(\tau))$.

The *Fisher Information* for this reparameterisation, \tilde{f} is given by

$$\tilde{I}(\tau) = \frac{I(\theta)}{g'(\theta)^2}$$

Proof 2.5 - Theorem 2.9

Since $\tilde{f}(x; \tau) = f(x; g^{-1}(\tau))$ the log-likelihood for τ is

$$\tilde{\ell}(\tau; x) = \ln \tilde{f}(x; \tau) = \ln f(x; g^{-1}(\tau))$$

The score is therefore

$$\begin{aligned} \tilde{\ell}'(\tau; x) &= \frac{d}{d\tau} \ln f(x; g^{-1}(\tau)) \\ &= \frac{d}{d\theta} \ln f(x; g^{-1}(\tau)) \times \frac{d}{d\tau} g^{-1}(\tau) \\ &= \ell'(g^{-1}(\tau); x) \times \frac{1}{g'(g^{-1}(\tau))} \\ &= \frac{\ell'(\theta; x)}{g'(\theta)} \end{aligned}$$

No we use the definition of *Fisher Information*

$$\begin{aligned} \tilde{I}(\tau) &= \mathbb{E}(\tilde{\ell}'(\tau; X)^2; \tau) \\ &= \mathbb{E}\left(\frac{\ell'(\theta; X)^2}{g'(\theta)^2}; \theta\right) \\ &= \frac{1}{g'(\theta)^2} \mathbb{E}(\ell'(\theta; X)^2; \theta) \\ &= \frac{I(\theta)}{g'(\theta)^2} \end{aligned}$$

Theorem 2.12 - Alternative Expression of Fisher Information

Let $X \sim f(\cdot; \theta)$ be a sufficiently regular random variable. Then

$$\text{if } \forall \theta \in \Theta \int_{\mathcal{X}} \frac{d^2}{d\theta^2} f(x; \theta) dx = \frac{d}{d\theta} \int_{\mathcal{X}} \frac{d}{d\theta} f(x; \theta) dx \text{ then } I(\theta) = -\mathbb{E}\left(\frac{d^2}{d\theta^2} \ell(\theta; X); \theta\right)$$

Proof 2.6 - Theorem 2.9

By the *Quotient Rule*

$$\begin{aligned} \frac{d^2}{d\theta^2} \ell(\theta; x) &= \frac{d}{d\theta} \frac{\frac{d}{d\theta} f(x; \theta)}{f(x; \theta)} \\ &= \frac{\frac{d^2}{d\theta^2} f(x; \theta)}{f(x; \theta)} - \left(\frac{\frac{d}{d\theta} f(x; \theta)}{f(x; \theta)}\right)^2 \end{aligned}$$

Consequently

$$\begin{aligned}
 \mathbb{E} \left(\frac{d^2}{d\theta^2} \ell(\theta; X); \theta \right) &= \int_S \frac{\frac{d^2}{d\theta^2} f(x; \theta)}{f(x; \theta)} f(x; \theta) dx - \int_S \left(\frac{\frac{d}{d\theta} f(x; \theta)}{f(x; \theta)} \right)^2 f(x; \theta) dx \\
 &= \int_S \frac{d^2}{d\theta^2} f(x; \theta) dx - \int_S \ell'(\theta; x)^2 f(x; \theta) dx \\
 &= 0 - \mathbb{E}(\ell'(\theta; X)^2; \theta) \\
 &= -I(\theta) \\
 \Rightarrow I(\theta) &= -\mathbb{E} \left(\frac{d^2}{d\theta^2} \ell(\theta; X); \theta \right)
 \end{aligned}$$

□

Theorem 2.13 - Distribution of Maximum Likelihood Estimators for Regular Models

Let $\mathbf{X}_n \stackrel{\text{iid}}{\sim} f_n(\cdot; \theta^*)$ be a sufficiently regular statistically model and $\{\hat{\theta}_n\}_{n \in \mathbb{N}}$ be a consistent sequence of *Maximum Likelihood Estimators* for θ^* . Then

$$\sqrt{nI(\theta^*)}(\hat{\theta}_n - \theta^*) \rightarrow_{\mathcal{D}(\cdot; \theta^*)} Z \sim \text{Normal}(0, 1)$$

Here $I(\theta^*)$ is unknown so we replace it with

i) $I(\hat{\theta}_n)$ when

- (a) $I(\theta)$ is continuous in a neighbourhood of θ^* ;
- (b) And, the interval $[L(\mathbf{X}), U(\mathbf{X})]$ with $L(\mathbf{x}) := \hat{\theta}_n - z_{\alpha/2} \sqrt{nI(\hat{\theta}_n)}$ and $U(\mathbf{x}) := \hat{\theta}_n + z_{\alpha/2} \sqrt{nI(\hat{\theta}_n)}$ is an asymptotically exact $1 - \alpha$ confidence interval for θ^* .

ii) $J_n(\hat{\theta}_n) := -\frac{1}{n} \sum_{i=1}^n \ell''(\hat{\theta}_n; X_i)$ when

- (a) $\hat{\theta}_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} \theta^*$;
- (b) $I(\theta) = -\mathbb{E}(\ell''(\theta; X); \theta) \forall \theta \in \Theta$;
- (c) $\exists C : \mathcal{X} \rightarrow [0, \infty)$ st $\mathbb{E}(C(X_1); \theta^*) < \infty$, $\Xi \subset \Theta$ is an open set containing θ^* and $\Delta(\cdot) : \Xi \rightarrow [0, \infty)$ is continuous at 0 st $\Delta(0) = 0$, and st $\forall \theta, \theta^*, x \in \Xi^2 \times \mathcal{X}$

$$|\ell''(\theta; x) - \ell''(\theta'; x)| \leq C(x)\Delta(\theta - \theta')$$

- (d) And, the interval $[L(\mathbf{X}), U(\mathbf{X})]$ with $L(\mathbf{x}) := \hat{\theta}_n - z_{\alpha/2} \sqrt{nJ_n(\hat{\theta}_n)}$ and $U(\mathbf{x}) := \hat{\theta}_n + z_{\alpha/2} \sqrt{nJ_n(\hat{\theta}_n)}$ is an asymptotically exact $1 - \alpha$ confidence interval for θ^*

Theorem 2.14 - Cramer-Rao Inequality

Let *Cramer-Rao Inequality* provides us with a *lower bound* for the performance of all estimators.

Let $\mathbf{X}_n \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$ be a sufficiently regular random vector and $\hat{\theta}_n(\cdot)$ be an estimator of θ with expectation $m_1(\theta) := \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n); \theta)$.

$$\text{if } \forall \theta \in \Theta, \underbrace{\frac{d}{d\theta} \int \hat{\theta}_n(\mathbf{x}) f_n(\mathbf{x}; \theta) d\mathbf{x}}_{\mathbb{E}(\hat{\theta}_n)} = \int \hat{\theta}_n(\mathbf{x}) \frac{d}{d\theta} f_n(\mathbf{x}; \theta) d\mathbf{x}$$

Then

$$\forall \theta \in \Theta, \text{Var}(\hat{\theta}_n(\mathbf{X}); \theta) \geq \frac{m_1'(\theta)^2}{nI(\theta)}$$

Proof 2.7 - Cramer-Rao Inequality

We notice that

$$\begin{aligned} m'(\theta) &= \frac{d}{d\theta} \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n); \theta) \\ &= \frac{d}{d\theta} \int_{S^n} \hat{\theta}_n(\mathbf{x}_n) f_n(\mathbf{x}_n; \theta) d\mathbf{x}_n \end{aligned}$$

The clever part of this proof is to observe that

$$\begin{aligned} \text{Var}(\hat{\theta}_n(\mathbf{X}_n); \theta) nI(\theta) &= \text{Var}(\hat{\theta}_n(\mathbf{X}_n); \theta) \text{Var}(\ell'_n(\theta; \mathbf{X}_n); \theta) \\ &\geq \text{Cov}(\hat{\theta}_n(\mathbf{X}_n), \ell'_n(\theta; \mathbf{X}_n); \theta)^2 \text{ by Covariance Inequality} \end{aligned}$$

Thus

$$\begin{aligned} \text{Cov}(\hat{\theta}_n(\mathbf{X}_n), \ell'_n(\theta; \mathbf{X}_n); \theta)^2 &= \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n) \ell'_n(\theta; \mathbf{X}_n); \theta) - \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n); \theta) \mathbb{E}(\ell'_n(\theta; \mathbf{X}_n); \theta) \\ &= \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n) \ell'_n(\theta; \mathbf{X}_n); \theta) - \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n); \theta) \times 0 \\ &= \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n) \ell'_n(\theta; \mathbf{X}_n); \theta) \\ &= \int_{S^n} \hat{\theta}_n(\mathbf{x}_n) \ell'_n(\theta; \mathbf{x}_n) f_n(\mathbf{x}_n; \theta) d\mathbf{x}_n \\ &= \int_{S^n} \hat{\theta}_n(\mathbf{x}_n) \frac{\frac{d}{d\theta} f_n(\mathbf{x}_n; \theta)}{f_n(\mathbf{x}_n; \theta)} f_n(\mathbf{x}_n; \theta) d\mathbf{x}_n \\ &= \int_{S^n} \hat{\theta}_n(\mathbf{x}_n) \frac{d}{d\theta} f_n(\mathbf{x}_n; \theta) \\ &= \frac{d}{d\theta} \int_{S^n} \hat{\theta}_n(\mathbf{x}_n) f_n(\mathbf{x}_n; \theta) d\mathbf{x}_n \text{ by regularity assumption} \\ &= m'(\theta) \\ \implies \text{Var}(\hat{\theta}_n(\mathbf{X}_n); \theta) nI(\theta) &\geq m'(\theta)^2 \end{aligned}$$

Remark 2.6 - Cramer-Rao Inequality with an Unbiased Estimator

Let $\hat{\theta}_n$ be an unbiased estimator of θ (i.e. $m_1(\theta) = \theta$). Then

$$\text{Var}(\hat{\theta}_n(\mathbf{X}_n); \theta) = \text{MSE}(\hat{\theta}_n(\mathbf{X}_n); \theta) \geq \frac{1}{nI(\theta)}$$

2.6 Asymptotic Distribution of Estimators**Theorem 2.15 - Asymptotic Distribution of Maximum Likelihood Estimators**

Suppose that $\mathbf{X}_n \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$ for some $\theta^* \in \Theta$ and assume that

- i) The sequence of maximum likelihood estimators $\{\hat{\theta}_n(\mathbf{X}_n)\}$ is consistent;
- ii) The Fisher Information Regularity Conditions (**Definition 6.2**) hold and $I(\theta^*) = -\mathbb{E}[\ell''(\theta; X); \theta] > 0$.
- iii) $\exists C : \mathcal{X} \rightarrow [0, \infty)$ such that $\mathbb{E}[C(X_1); \theta^*] < \infty$ and $\Delta : \Xi \rightarrow [0, \infty)$, where $\Xi \subset \Theta$ st $\theta^* \in \Xi$, that is continuous at 0 st $\Delta(0) = 0$, such that

$$\forall (\theta, \theta', x) \in \mathcal{X}^2 \times \mathcal{X}, \quad |\ell''(\theta; x) - \ell''(\theta'; x)| \leq C(x) \Delta(\theta - \theta')$$

Then $\forall \theta^* \in \Theta$

$$\sqrt{nI(\theta^*)}(\hat{\theta}_n(\mathbf{X}_n) - \theta^*) \rightarrow_{\mathcal{D}(\cdot; \theta^*)} Z \sim \text{Normal}(0, 1)$$

Proof 2.8 - Theorem 2.11

By **Theorem 2.11** $\ell'_n(\hat{\theta}_n; \mathbf{X}) = \ell'_n(\theta^*; \mathbf{X}) + (\hat{\theta}_n - \theta^*)[\ell''_n(\theta^*; \mathbf{X}) + R_n]$ where $\frac{1}{n}R_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} 0$.

Since $\hat{\theta}_n$ is the maximum likelihood estimator & the Fisher Information Regularity Conditions hold, the score at $\ell'(\hat{\theta}_n; X) = 0$.

Hence, $0 = \ell''(\hat{\theta}_n; X) = \ell'_n(\theta; X) + (\hat{\theta}_n - \theta^*)\{\ell''(\theta; X) + R_n\}$.

Rearranging & rescaling by \sqrt{n} gives

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \frac{\frac{1}{\sqrt{n}} \ell'(\theta^*; X)}{-\frac{1}{\sqrt{n}} \{\ell''(\theta^*; X) + R_n\}} =: \frac{U_n}{V_n - \frac{R_n}{n}}$$

Recall that $\ell'_n(\theta^*; X) = \sum_{i=1}^n \ell'(\theta; X_i)$ and $\ell''_n(\theta^*; X) = \sum_{i=1}^n \ell''(\theta^*; X_i)$.

Since $\mathbb{E}(\ell'(\theta^*; X_i); \theta^*) = 0$ and $\text{Var}(\ell'(\theta^*; X_i); \theta^*) = I(\theta^*)$

$\implies U_n \rightarrow_{\mathcal{D}(\cdot; \theta^*)} U \sim \text{Normal}(0, I(\theta^*))$ by the *Central Limit Theorem*.

We observed that $V_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} I(\theta^*)$ by the *Weak Law of Large Numbers* since $\mathbb{E}(-\ell''(\theta^*; X_i); \theta^*) = I(\theta^*)$.

It follows that $V_n - \frac{1}{n}R_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} I(\theta^*)$ by *Slutsky's Theorem*.

Using *Slutsky's Theorem* again

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \frac{U_n}{V_n - \frac{1}{n}R_n} \rightarrow_{\mathcal{D}(\cdot; \theta^*)} \frac{\sqrt{I(\theta^*)}}{I(\theta^*)} Z \text{ where } Z \sim \text{Normal}(0, 1)$$

We can rewrite this as

$$\sqrt{nI(\theta^*)}(\hat{\theta}_n - \theta^*) \rightarrow_{\mathcal{D}(\cdot; \theta^*)} Z \sim \text{Normal}(0, 1)$$

Theorem 2.16 - *Convergence of Score of Maximum Likelihood Estimators*

Under the conditions in **Theorem 2.11**, with $\hat{\theta}_n$ a Maximum Likelihood Estimator

$$\ell'_n(\hat{\theta}_n; \mathbf{X}) = \ell'_n(\theta^*; \mathbf{X}) + (\hat{\theta}_n - \theta^*)[\ell''_n(\theta^*; \mathbf{X}) + R_n]$$

where $\frac{1}{n}R_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} 0$.

Proof 2.9 - *Theorem 2.12*

This is a non-examinable, sketch proof of Theorem 8.2.

By the regularity conditions and the mean value theorem

$$\frac{\ell'_n(\theta; \mathbf{x}) - \ell'_n(\theta^*; \mathbf{x})}{\theta - \theta^*} = \ell''_n(\tilde{\theta}; \mathbf{x})$$

for some $\tilde{\theta} \in (\theta, \theta^*)$. Hence, we deduce that

$$\begin{aligned} \ell'_n(\theta; \mathbf{x}) - \ell'_n(\theta^*; \mathbf{x}) &= (\theta - \theta^*)\ell''_n(\tilde{\theta}; \mathbf{x}) \\ &= (\theta - \theta^*)\{\ell''_n(\theta^*; \mathbf{x}) + [\ell''_n(\tilde{\theta}; \mathbf{x}) - \ell''_n(\theta^*; \mathbf{x})]\} \\ &= (\theta - \theta^*)\{\ell''_n(\theta; \mathbf{x}) + R_n(\theta, \theta^*, \mathbf{x})\} \end{aligned}$$

Now we replace θ with the maximum likelihood estimator $\hat{\theta}_n := \hat{\theta}_n(\mathbf{X})$. We find

$$\ell'_n(\hat{\theta}_n; \mathbf{X}) = \ell'_n(\theta^*; \mathbf{X}) + (\hat{\theta}_n - \theta^*)\{\ell''_n(\theta^*; \mathbf{X}) + R_n(\hat{\theta}_n, \theta^*, \mathbf{x})\}$$

and we need to analyse R_n .

Since $\hat{\theta}_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} \theta^*$ we can take n large enough that $\mathbb{P}(\hat{\theta}_n \in \Xi; \theta^*)$ with arbitrarily high probability.

On the event $\{\hat{\theta} \in \Xi\}$ and we have $\{\tilde{\theta}_n \in \Xi\}$ since $\tilde{\theta}_n \in (\hat{\theta}_n, \theta^*)$ and

$$\begin{aligned} \left| \frac{1}{n}R_n \right| &= \frac{1}{n} |\ell''_n(\tilde{\theta}_n; \mathbf{X}) - \ell''_n(\theta^*; \mathbf{X})| \\ &= \frac{1}{n} \left| \sum_{i=1}^n \ell''(\tilde{\theta}_n; X_i) - \ell''(\theta^*; X_i) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \left| \ell''(\tilde{\theta}_n; X_i) - \ell''(\theta^*; X_i) \right| \\ &\leq \Delta(\tilde{\theta}_n - \theta^*) \left\{ \frac{1}{n} \sum_{i=1}^n C(X_i) \right\} \end{aligned}$$

from the smoothness condition on ℓ'' .

From the *Weak Law of Large Numbers*

$$\frac{1}{n} \sum_{i=1}^n C(X_i) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} \mathbb{E}(C(X_1); \theta^*) < \infty$$

and from the consistency of $\{\hat{\theta}_n\}$ and $\{\tilde{\theta}_n\}$ and continuity of $\Delta(\cdot)$ we have by the *Continuous Mapping Theorem*

$$\Delta(\tilde{\theta}_n - \theta^*) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} 0$$

Hence, $\frac{1}{n}R_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} 0$ □

2.6.1 Confidence Intervals

Theorem 2.17 - Convergence in Distribution of Confidence Intervals

Let $\mathbf{X} \sim f(\cdot; \theta^*)$ with $\theta \in \Theta$ and define $\{\hat{\theta}_n\}_{n \in \mathbb{N}}$ be a consistent sequence of estimators of θ^* . Suppose that $\{\hat{\theta}_n\}$ is asymptotically normal in the sense that

$$\exists \sigma^2 > 0 \text{ st } \frac{\hat{\theta}_n(\mathbf{X}) - \theta^*}{\sqrt{\sigma^2/n}} \rightarrow_{\mathcal{D}(\cdot; \theta^*)} Z \sim \text{Normal}(0, 1)$$

Then

$\forall \alpha \in (0, 1)$, $\mathcal{I}_n(\mathbf{X}) = [L_n(\mathbf{X}), U_n(\mathbf{X})]$ is an asymptotically exact $1 - \alpha$ confidence interval where $L_n(\mathbf{x}) := \hat{\theta}_n(\mathbf{x}) - z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}}$ and $U_n(\mathbf{x}) := \hat{\theta}_n(\mathbf{x}) + z_{\alpha/2} \sqrt{\frac{\sigma^2}{n}}$.

Proof 2.10 - Theorem 2.13

Let $\{W_n\}_{n \in \mathbb{N}}$ be defined by $W_n := \frac{\hat{\theta}_n(X) - \theta^*}{\sqrt{\sigma^2/n}}$.

Since $W_n \rightarrow_{\mathcal{D}(\cdot; \theta^*)} Z \sim \text{Normal}(0, 1)$ we have

$$\begin{aligned} \mathbb{P}(-z_{\alpha/2} \leq W_n \leq z_{\alpha/2}) &= F_{W_n}(z_{\alpha/2}) - F_{W_n}(-z_{\alpha/2}) \\ &\xrightarrow{n \rightarrow \infty} \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\ &= 1 - \alpha \end{aligned}$$

Similar to before we have the equivalence of events

$$\{-z_{\alpha/2} \leq W_n \leq z_{\alpha/2}\} = \left\{ \hat{\theta}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \theta^* \leq \hat{\theta}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

So $\lim_{n \rightarrow \infty} \mathbb{P} \left(\hat{\theta}_n(X) - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \theta^* \leq \hat{\theta}_n(X) + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \theta^* \right) = 1 - \alpha$. □

2.7 Efficiency of Estimators

Definition 2.27 - Efficient Estimator

Let $\hat{\theta}$ be an estimator of parameter θ .

$\hat{\theta}$ is said to be an *Efficient Estimator* if its variance is equal to the *Cramer-Rao Lower Bound* $\forall \theta^*$.

$$\forall \theta^*, \text{Var}(\hat{\theta}; \theta^*) = \frac{m'(\theta^*)^2}{nI(\theta)}$$

Definition 2.28 - Asymptotically Efficient Sequence of Estimators

Let $\mathbf{X} \sim f(\cdot; \theta)$ for $\theta \in \Theta$ and $\{\hat{\theta}_n(\mathbf{X})\}_{n \in \mathbb{N}}$ be a sequence of estimators.

The sequence $\{\hat{\theta}_n\}$ is *Asymptotically Efficient* if either

i) its *Mean-Squared Error* converges in value to the *Cramer-Rao Lower Bound*

$$\forall \theta \in \Theta, n\text{MSE}(\hat{\theta}_n(\mathbf{X}_n); \theta) \xrightarrow{n \rightarrow \infty} \frac{1}{I(\theta)}$$

ii) Or, $\hat{\theta}_n$ *Converges in Distribution* to a standard Normal

$$\forall \theta \in \Theta, \sqrt{nI(\theta)}(\hat{\theta} - \theta) \rightarrow_{\mathcal{D}(\cdot; \theta)} Z \sim \text{Normal}(0, 1)$$

Remark 2.7 - Under the conditions of Theorem 2.11 Maximum Likelihood Estimators are Asymptotically Efficient

3 Testing

3.1 Hypothesis Testing

Definition 3.1 - Hypothesis

A *Hypothesis* is a statement about the value of one or more parameters in a parameteric model.

$$H : \theta \in \Theta_0 \text{ where } \Theta_0 \subseteq \Theta$$

Definition 3.2 - Simple Hypothesis

A *Simple Hypothesis* is a *Hypothesis* which states that θ has an exact value.

i.e. $H : \theta \in \Theta_0$ where $|\Theta_0| = 1$.

Definition 3.3 - Composite Hypothesis

A *Composite Hypothesis* is a *Hypothesis* which states that θ takes one of a range of values.

i.e. $H : \theta \in \Theta_0$ where $|\Theta_0| > 1$.

Definition 3.4 - Hypothesis Testing

Hypothesis Testing is the process using observed data to determine which of two *hypotheses* is more consistent with the data.

For the *hypotheses* we define a *Null Hypothesis*, $H_0 : \theta \in \Theta_0$, which is our default position & an *Alternative Hypothesis*, $H_1 : \theta \in \Theta_1$ where $\Theta_1 := \Theta \setminus \Theta_0$.

Proposition 3.1 - Hypothesis Testing Process

Let \mathbf{x} be a realisation of \mathbf{X} .

- i) Define a *Model* $f(\cdot; \theta)$, for $\theta \in \Theta$, st $\mathbf{X} \sim f(\cdot; \theta)$.
- ii) Define a *Null Hypothesis*, H_0 , and an *Alternative Hypothesis*, H_1 .
- iii) Define a *Test Statistic*, $T(\cdot)$.
- iv) Choose a *Significan Level*, α , and calculate the resulting *Critical Value*, c .
- v) Calculate the observed value of the *Test Statistic*, $t = T(\mathbf{x})$.
- vi) If $t \geq c$ then reject H_0 in favour of H_1 , otherwise accept H_0 .

Definition 3.5 - One-Sided Hypothesis Test

Consider the two hypotheses $H_0 : \theta \in \Theta_0$ & $H_1 : \theta \in \Theta_1$.

A *Hypothesis Test* on these two hypotheses is said to be a *One-Sided Hypothesis Test* if both Θ_0 & Θ_1 are continuous regions of the parameter space.

i.e. $\exists \theta_0 \in \Theta$ st $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$ (visa-versa) are equivalent definitions to above.

Definition 3.6 - Two-Sided Hypothesis Test

Consider the two hypotheses $H_0 : \theta \in \Theta_0$ & $H_1 : \theta \in \Theta_1$.

A *Hypothesis Test* on these two hypotheses is said to be a *Two-Sided Hypothesis Test* if at least one of Θ_0 & Θ_1 is not a continuous region of the parameter space.

i.e. $\exists \theta_0, \theta_1 \in \Theta$ st $H_0 : \theta \in [\theta_0, \theta_1]$ and $H_1 : \theta \notin [\theta_0, \theta_1]$ (visa-versa) are equivalent definitions to above.

Definition 3.7 - Type I & Type II Error

Consider the table below

Truth \ Action	Retain H_0	Reject H_0
H_0 is True	Correct	Type I Error
H_1 is True	Type II Error	Correct

Type I Error occurs when the *Null Hypothesis* is rejected, when in fact it is true.

Type II Error occurs when the *Null Hypothesis* is accepted, when in fact it is false.

Definition 3.8 - Significance Level

Significance Level, α , is the rate at which we allow *Type I Errors* to occur

$$\alpha := \mathbb{P}(\text{Type I Error}) \in [0, 1]$$

i.e. What is an acceptable proportion of times to reject H_0 when it is in fact true.

N.B. Typically $\alpha \leq 0.05$.

Remark 3.1 - Significance Level is directly related to the phrase "Statistical Significance"

Definition 3.9 - Test Statistic

A *Test Statistic* is a random variable, T , whose value depends on the observed data set and is used to determine the outcome of a hypothesis test. *Test Statistics* are defined in such a way that they measure how likely a given observation is given a particular hypothesis. Thus if an observation is deemed sufficiently unlikely by a *Test Statistic* then we reject that hypothesis, in favour of the alternative.

N.B. $T : \mathcal{X}^n \rightarrow \mathbb{R}$ where n is the number of observed values.

Proposition 3.2 - Common Test Statistics

Test Statistic	Use
$T(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$	Testing mean

Definition 3.10 - Equivalent Statistics

Let $T(\cdot)$ & $T'(\cdot)$ be *Test Statistics* and \mathbf{X} be a *Random Variable*.

We say $T(\cdot)$ and $T'(\cdot)$ are *Equivalent Statistics* if

$$\forall c \in \mathbb{R} \exists c' \in \mathbb{R} \text{ st } \{\mathbf{x} \in \mathcal{X}^n : T(\mathbf{x}) \geq c\} \equiv \{\mathbf{x} \in \mathcal{X}^n : T'(\mathbf{x}) \geq c'\}$$

Proposition 3.3 - Verifying Equivalent Statistics

Let $T(\cdot)$ & $T'(\cdot)$ be *Test Statistics*.

To verify that $T(\cdot)$ and $T'(\cdot)$ are *Equivalent Statistics* it is sufficient to factorise $T(\cdot)$ as

$$T(\mathbf{x}) = Mf(T'(\mathbf{x}))$$

for some M, f where M is independent of \mathbf{x} and $f(\cdot)$ is an increasing, bijective function.

Proof 3.1 - Proposition 3.3

$$\begin{aligned} T(\mathbf{x}) \geq c &\iff Mf(T'(\mathbf{x})) \geq c \\ &\iff f(T'(\mathbf{x})) \geq \frac{c}{M} \\ &\iff T'(\mathbf{x}) \geq \underbrace{f^{-1}\left(\frac{c}{M}\right)}_{c'} \end{aligned}$$

Definition 3.11 - Critical Value

A *Critical Value*, $c \in \mathbb{R}$, is an explicit value which if the observed value of the test statistic, $T(\mathbf{x})$, exceeds then we reject the *Null-Hypothesis*.

i.e. If $T(\mathbf{x}; H_0) \geq c$ then we reject H_0 .

N.B. The *Critical Value* depends on the *Test Statistic* & the *Significance Level* used in a given

test.

Definition 3.12 - Critical Region

The *Critical Region*, R , is the set of observations which would lead to us rejecting the *Null-Hypothesis*.

Let $T(\cdot)$ be a *Test Statistic* & c be a *Critical Value* then

$$R := \{\mathbf{x} \in \mathcal{X}^n : T(\mathbf{x}) \geq c\}$$

N.B. $\mathcal{X}^n = R \cup R^c$.

Definition 3.13 - Power Function

The *Power Function*, $\pi(\cdot)$ measures the probability of rejecting the *Null-Hypothesis* given that the true value of the parameter is θ .

Let $\mathbf{X} \sim f(\cdot; \theta^*)$, $T(\cdot)$ be a test statistic, c be a *Critical Value* & R be the *Critical Region*. Then

$$\pi(\theta; T, c) := \mathbb{P}(\mathbf{X} \in R; \theta^* = \theta) = \mathbb{P}(T(\mathbf{X}) \geq c; \theta^* = \theta)$$

N.B. $\pi(\cdot; T, c) : \Theta \rightarrow [0, 1]$.

Remark 3.2 - $\pi(\cdot; T, c) \equiv 1 - \mathbb{P}(\text{Type II Error})$

Definition 3.14 - Uniformly Most Powerful Test

Define two *Composite Hypotheses*, $H_0 : \theta \in \Theta_0$ & $H_1 : \theta \notin \Theta_0$ for $|\Theta_0| > 1$ and a *Test*, (T, c) , for these hypotheses.

We say that this *Test*, (T, c) , is a *Uniformly Most Powerful Test* for these hypotheses if

$$\forall (T', c'), \pi(\theta; T, c) \geq \pi(\theta; T', c') \text{ for } \theta \in \Theta_1 := \Theta \setminus \Theta_0$$

N.B. We refer to T in this case as the *Uniformly Most Powerful Test Statistic*.

Remark 3.3 - A *Uniformly Most Powerful Test* is not Guaranteed to exist

Proposition 3.4 - Procedure for Hypothesis Testing with Composite Hypotheses

- i) Calculate the *Likelihood Ratio Test Statistic*, $T_{NP}(\cdot)$.
- ii) Find the simplest *Equivalent* test statistic, $T(\cdot)$, to the *Likelihood Ratio Test Statistic*.
- iii) Compute the *p-Value* using the distribution of $T(\cdot)$ under the *Null-Hypothesis*
- iv) Determine whether you accept the *Null-Hypothesis* given the computed *p-Value*.

Definition 3.15 - p-Value

The *p-Value* of a *Test Statistic* is the probability of observing a test statistic, $T(\mathbf{X})$, at least as extreme as a realisation of the test statistic, $T(\mathbf{x})$, under the *Null Hypothesis*.

Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ be a *Random Vector* for $\theta^* \in \Theta$, \mathbf{x} be a realisation of \mathbf{X} , $T(\cdot)$ be a *Test Statistic* and define a *Null Hypothesis*, $H_0 : \theta \in \Theta_0$.

$$p(\mathbf{x}) := \sup_{\theta_0 \in \Theta_0} \mathbb{P}(T(\mathbf{X}) \geq T(\mathbf{x}); \theta_0)$$

N.B. $p(\mathbf{x})$ is the smallest *Significance Level* at which we would reject the *Null Hypothesis*.

Remark 3.4 - *p-Value* is a measure of the evidence against the *Null-Hypothesis*

Definition 3.16 - Size of a Test

The *Size of a Test* is the maximum power of a test under the *Null-Hypothesis*.

Let $T(\cdot)$ be a *Test Statistic* & c be a *Critical Value*

$$\alpha := \sup_{\theta \in \Theta_0} \pi(\theta; T, c)$$

i.e. The greatest possible probability of making a *Type I Error*

3.1.1 Neyman-Pearson Approach**Remark 3.5 - Motivation**

TODO

Definition 3.17 - Likelihood Ratio Test Statistic

Let \mathbf{x} be a realisation of $\mathbf{X} \sim f_n(\cdot; \theta)$.

Consider two *Simple Hypotheses* $H_0 : \theta = \theta_0$ & $H_1 : \theta = \theta_1$.

The *Likelihood Ratio Test Statistic* is

$$T_{\text{NP}}(\mathbf{x}) := \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} = \frac{f_n(\mathbf{x}; \theta_1)}{f_n(\mathbf{x}; \theta_0)}$$

N.B. AKA *Neyman-Pearson Test Statistic*.

Theorem 3.1 - The Neyman-Pearson Lemma

Let \mathbf{x} be a realisation of $\mathbf{X} \sim f_n(\cdot; \theta)$.

Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ using the *Neyman-Pearson Test Statistic*, T_{NP} .

Let c_{NP} be the *Critical Value* st $(T_{\text{NP}}, c_{\text{NP}})$ has *Size* α .

$$\text{i.e. } c_{\text{NP}} \text{ st } \mathbb{P}(T_{\text{NP}} \geq c_{\text{NP}}; \theta_0) = \alpha$$

Then $(T_{\text{NP}}, c_{\text{NP}})$ is *Equivalent* to the *Uniformly Most Powerful α -Level Test*.

Proof 3.2 - Theorem 3.1

Consider for an arbitrary level α test (T, c) , the linear combination of *Type I Errors* and *Type II Errors*.

$$\phi(T, c) := c_{\text{NP}} \alpha(T, c) + \beta(T, c)$$

where $\alpha(T, c) = \mathbb{P}(T(\mathbf{X}) \geq c; \theta_0) = \mathbb{P}(\text{Type I Error})$ and

$\beta(T, c) = \mathbb{P}(T(\mathbf{X}) < c; \theta_1) = 1 - \mathbb{P}(T(\mathbf{X}) \geq c; \theta_1) = \mathbb{P}(\text{Type II Error})$.

Then

$$\begin{aligned} \phi(T, c) &= c_{\text{NP}} \alpha(T, c) + \beta(T, c) \\ &= c_{\text{NP}} \mathbb{P}(T(\mathbf{X}) \geq c; \theta_0) + [1 - \mathbb{P}(T(\mathbf{X}) \geq c; \theta_1)] \\ &= \left[c_{\text{NP}} \int \mathbb{1}\{T(\mathbf{x}) \geq c\} f_n(\mathbf{x}; \theta_0) d\mathbf{x} \right] + \left[1 - \int \mathbb{1}\{T(\mathbf{x}) \geq c\} f_n(\mathbf{x}; \theta_1) d\mathbf{x} \right] \\ &= 1 + \int \mathbb{1}\{T(\mathbf{x}) \geq c\} [c_{\text{NP}} f_n(\mathbf{x}; \theta_0) - f_n(\mathbf{x}; \theta_1)] d\mathbf{x} \\ &= 1 + \int \mathbb{1}\{T(\mathbf{x}) \geq c\} \left[c_{\text{NP}} - \frac{f_n(\mathbf{x}; \theta_1)}{f_n(\mathbf{x}; \theta_0)} \right] f_n(\mathbf{x}; \theta_0) d\mathbf{x} \\ &= 1 + \int \mathbb{1}\{T(\mathbf{x}) \geq c\} (c_{\text{NP}} - T_{\text{NP}}(\mathbf{x})) f_n(\mathbf{x}; \theta_0) d\mathbf{x} \end{aligned}$$

Now consider the difference

$$\phi(T, c) - \phi(T_{\text{NP}}, c_{\text{NP}}) = \int (\mathbb{1}\{T(\mathbf{x}) \geq c\} - \mathbb{1}\{T_{\text{NP}}(\mathbf{x}) \geq c_{\text{NP}}\}) (c_{\text{NP}} - T_{\text{NP}}(\mathbf{x})) f_n(\mathbf{x}; \theta_0) d\mathbf{x}$$

We observe that

$$\mathbb{1}\{T_{NP}(\mathbf{x}) \geq c_{NP}\} = 1 \iff c_{NP} - T_{NP}(\mathbf{x}) \leq 0$$

and

$$\mathbb{1}\{T_{NP}(\mathbf{x}) \geq c_{NP}\} = 0 \iff c_{NP} - T_{NP}(\mathbf{x}) > 0$$

Thus

$$\forall \mathbf{x} \in \mathcal{X}^n, \quad [\mathbb{1}\{T(\mathbf{x}) \geq c\} - \mathbb{1}\{T_{NP}(\mathbf{x}) \geq c_{NP}\}](c_{NP} - T_{NP}(\mathbf{x})) \geq 0$$

and hence as the integral of a non-negative function

$$\phi(T, c) - \phi(T_{NP}, c_{NP}) \geq 0$$

We have established

$$\begin{aligned} 0 &\leq \phi(T, c) - \phi(T_{NP}, c_{NP}) \\ &= c_{NP}\alpha(T, c) + \beta(T, c) - c_{NP}\alpha(T_{NP}, c_{NP}) - \beta(T_{NP}, c_{NP}) \\ &= \underbrace{c_{NP}[\alpha(T, c) - \alpha(T_{NP}, c_{NP})]}_{\geq 0} + \underbrace{\beta(T, c) - \beta(T_{NP}, c_{NP})}_{\geq 0} \end{aligned}$$

Since (T, c) specifies an α -level test, we know $\alpha(T, c) \geq c$ while (T_{NP}, c_{NP}) specifies an α -size test so $\alpha(T_{NP}, c_{NP}) = \alpha$.

It follows that

$$\alpha(T, c) - \alpha(T_{NP}, c_{NP})$$

so we have

$$\beta(T, c) - \beta(T_{NP}, c_{NP}) \geq 0$$

which means (T_{NP}, c_{NP}) 's *Type II Error* rate is no higher than (T, c) .

Since (T, c) is an arbitrary α level test, we conclude that (T_{NP}, c_{NP}) is the most powerful test with level α . \square

Remark 3.6 - Neyman-Pearson Lemma with Non-Continuous Random Variables

If $T(\mathbf{X})$ is not a continuous random variable, then it is possible that no c_{NP} exists.

In this situation we perform an appropriate *randomised* test, and this will also be the most powerful α -size test.

N.B. This is out of scope of this course.

Proposition 3.5 - Neyman-Pearson Testing Procedure

From **Theorem 3.1** we can deduce the *Neyman-Pearson Testing Procedure* for testing two *Simple Hypotheses*, H_0 against H_1 .

- i) Use the *Likelihood Ratio Test Statistic* as the *Test Statistic*

$$T_{NP}(\mathbf{x}) := \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} = \frac{f_n(\mathbf{x}; \theta_1)}{f_n(\mathbf{x}; \theta_0)}$$

- ii) Find a critical value, c , st we achieve the desired significance level, α .

$$\alpha = \pi(\theta_0; T, c) = \mathbb{P}(T_{NP}(\mathbf{x}) \geq c; \theta_0)$$

- iii) Compute the *Power* of the *Alternative Hypothesis*

$$\pi(\theta_1; T, c) = \mathbb{P}(T_{NP}(\mathbf{X}) \geq c; \theta_1)$$

- iv) Compute the observed test statistic, $t_{\text{obs}} := T(\mathbf{x})$ and report whether $T(\mathbf{x}) \geq c$.

- v) Report the power of the *Alternative Hypothesis*, $\pi(\theta_1; T_{NP}, c)$

Remark 3.7 - *Limitations of Neyman-Pearson Approach to Hypothesis Testing*

- i) Reporting *rejection/acceptance* of the *Null-Hypothesis* does not show the strength of the evidence against the *Null-Hypothesis*.
- ii) We may wish to set the *Significance Level*, $\alpha := \pi(\theta_0)$, & *Type II Error Rate*, $\beta := 1 - \pi(\theta_1)$ together, or optimise both to be as minimal as possible.

3.1.2 Generalised Hypothesis Testing

Definition 3.18 - *Generalised Likelihood Ratio Test*

Let $\mathbf{X} \sim f_n(\cdot; \theta)$ be a *Random Vector* and consider *Composite Hypotheses* $H_0 : \theta \in \Theta_0$ & $H_1 : \theta \in \Theta_1$.

We define the *Generalised Likelihood Ratio Test* to be

$$\Lambda(\mathbf{x}) := \frac{\sup_{\theta \in \Theta_0} f_n(\mathbf{x}; \theta)}{\sup_{\theta \in \Theta} f_n(\mathbf{x}; \theta)} = \min \left\{ \underbrace{1}_{\hat{\theta} \in \Theta_0}, \underbrace{\frac{\sup_{\theta \in \Theta_0} f_n(\mathbf{x}; \theta)}{\sup_{\theta \in \Theta_1} f_n(\mathbf{x}; \theta)}}_{\hat{\theta} \notin \Theta_0} \right\}$$

N.B. This compares the best fit for the data under the *Null Hypothesis* to the best fit from the whole parameter space.

Definition 3.19 - *Nested Parameter Space*

Assume the *Parameter Space* is $\Theta \subseteq \mathbb{R}^d$ for some $d \geq 1$.

Define a continuously differentiable bijection, $\phi(\cdot) := (\phi_1(\cdot), \phi_2(\cdot)) : \Theta \rightarrow \Phi_1 \times \Phi_2$ where $\Phi_1 \subseteq \mathbb{R}^r$ & $\Phi_2 \subseteq \mathbb{R}^{d-r}$ for some $r \in \mathbb{N}$.

$\Theta_0 \subseteq \Theta$ is said to be *Nested* in Θ if

$$\Theta_0 := \{\theta \in \Theta : \phi(\theta) = c\} \text{ for some } c \in \Phi_1 \subseteq \mathbb{R}^r$$

N.B. $\dim(\Theta_0) = d - r$.

Theorem 3.2 -

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$ be a *Random Vector* for some $\theta \in \Theta_0$ where Θ_0 is *Nested* in Θ .

Then

$$T_n(\mathbf{X}) := -2 \ln \Lambda_n(\mathbf{X}) \rightarrow_{\mathcal{D}(\cdot; \theta)} W \sim \chi_r^2$$

where $r = \dim(\Theta) - \dim(\Theta_0)$.

N.B. The proof of this relies on a Taylor Expansion of the Likelihood function.

Remark 3.8 - *The fact that $-2 \ln \Lambda_n(\mathbf{X}) \rightarrow_{\mathcal{D}(\cdot; \theta)} W \sim \chi_r^2$, is a generalisation of the result which motivates Wilks Confidence Sets*

Proposition 3.6 - *Computing an Approximate p-Value for Composite Hypotheses*

- i) Compute *Observed Test Statistic*, $T_n(\mathbf{x}) := -2 \ln \Lambda_n(\mathbf{x})$.
- ii) Determine $r = \dim(\Theta) - \dim(\Theta_0)$.
- iii) Compute the approximate *p-Value*

$$p(\mathbf{x}) = \mathbb{P}(\chi_r^2 \geq -2 \ln \Lambda_n(\mathbf{x}))$$

3.2 Categorical Distributions & Pearson's χ^2 -Test

Definition 3.20 - Categorical Distributions

Consider a scenario where a random variable Y takes one of m possible values, $\{1, \dots, m\}$ (i.e. Categories) and $p_i := \mathbb{P}(Y = i)$. Then Y is said to have a *Categorical Distribution*

$$Y \sim \text{Categorical}(\mathbf{p})$$

where \mathbf{p} is a vector of probabilities (i.e. $\sum p_i = 1$ & $p_i \geq 0 \forall i$).

Definition 3.21 - Counts in Categorical Distribution

Let $\mathbf{Y} \stackrel{\text{iid}}{\sim} \text{Categorical}(\mathbf{p})$ be n random variables.

Definition 3.22 - Multinomial Distribution

Let $\mathbf{Y} \stackrel{\text{iid}}{\sim} \text{Categorical}(\mathbf{p})$ be n random variables and $\mathbf{X} := \{N_1, \dots, N_m\}$, where $N_k := \sum_{i=1}^n \mathbb{1}\{Y_i = k\}$, represent the counts from \mathbf{Y} .

Then \mathbf{X} is said to have a *Multinomial Distribution*

$$\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p})$$

with

$$\begin{aligned} f_n(\mathbf{x}; \mathbf{p}) &= \mathbb{1}\left\{\sum_{i=1}^m x_i = n\right\} \left[\frac{n!}{\prod_{i=1}^m x_i!}\right] \prod_{i=1}^m p_i^{x_i} \\ \mathbb{E}(N_i) &= np_i \\ \text{Var}(N_i) &= np_i(1 - p_i) \end{aligned}$$

Theorem 3.3 - Maximum Likelihood Estimate - Multinomial Distribution

Let $\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p}^*)$ & \mathbf{x} be a realisation of \mathbf{X} . Then

$$\hat{\mathbf{p}}_{\text{MLE}}(\mathbf{x}) = (\hat{p}_1(\mathbf{x}), \dots, \hat{p}_m(\mathbf{x})) = \left(\frac{x_1}{n}, \dots, \frac{x_m}{n}\right)$$

Proof 3.3 - Theorem 3.3

Note that

$$\sum_{i=1}^m p_i = 1 \implies p_m = 1 - \sum_{i=1}^{m-1} p_i$$

Hence there are only $m - 1$ independent variables and

$$\begin{aligned} L(\mathbf{p}, \mathbf{x}) &= L(p_1, \dots, p_{m-1}; \mathbf{x}) \\ &\propto \prod_{j=1}^m p_j^{x_j} \\ &= \left(\prod_{j=1}^{m-1} p_j^{x_j}\right) \left(1 - \sum_{i=1}^{m-1} p_i\right)^{x_m} \end{aligned}$$

So

$$\ell(p_1, \dots, p_{m-1}; \mathbf{x}) = C + \left(\sum_{i=1}^{m-1} x_i \ln p_i\right) + x_m \ln \left(1 - \sum_{i=1}^{m-1} p_i\right)$$

Now for $k = 1, \dots, m - 1$.

$$\begin{aligned} \text{Setting } \frac{\partial}{\partial p_k} \ell(p_1, \dots, p_{m-1}; \mathbf{x}) &= \frac{x_k}{p_k} - \frac{x_m}{1 - \sum_{i=1}^{m-1} p_i} \\ &= 0 \\ \implies \frac{x_k}{p_k} &= \frac{x_m}{p_m} \quad \forall k \in [1, m] \end{aligned}$$

So $\frac{x_1}{p_1} = \dots = \frac{x_m}{p_m} = c$ and $\sum_{i=1}^m p_i = 1$.

$$\implies \sum_{i=1}^m \frac{x_i}{c} = 1 \implies \sum_{i=1}^m x_i = c \implies n = c$$

Hence $\frac{x_k}{p_k} = n \implies \hat{p}_j = \frac{x_k}{n} \forall k \in [1, m]$.

In order to confirm that this is a maximum we will show that $\ell(\mathbf{p}; \mathbf{x})$ is concave.

i.e. for $\lambda \in [0, 1]$ $\ell(\lambda \mathbf{p} + (1 - \lambda) \mathbf{p}'; \mathbf{x}) \geq \lambda \ell(\mathbf{p}; \mathbf{x}) + (1 - \lambda) \ell(\mathbf{p}'; \mathbf{x})$.

$$\begin{aligned} \ell(\lambda \mathbf{p} + (1 - \lambda) \mathbf{p}'; \mathbf{x}) &= \sum_{i=1}^m x_i \ln(\lambda p_i + (1 - \lambda) p'_i) \\ &\geq \sum_{i=1}^m x_i [\lambda \ln p_i + (1 - \lambda) \ln p'_i] \text{ since } \ln x \text{ is concave} \\ &= [\lambda \sum_{i=1}^m x_i \ln p_i] + x_i (1 - \lambda) \ln p'_i \\ &= \lambda \ell(\mathbf{p}; \mathbf{x}) + (1 - \lambda) \ell(\mathbf{p}'; \mathbf{x}) \end{aligned}$$

Thus concave.

It follows that

$$\Lambda_n(\mathbf{x}) = \frac{f_n(\mathbf{x}; \mathbf{p}_0)}{\sup_{\mathbf{p} \in \mathcal{S}_m} f_n(\mathbf{x}; \mathbf{p})} = \prod_{i=1}^m \frac{p_{0,i}^{x_i}}{\hat{p}_i^{x_i}} = \prod_{i=1}^m \frac{p_{0,i}^{x_i}}{(x_i/n)^{x_i}}$$

so that

$$T_n(\mathbf{x}) = -2 \ln \Lambda_n(\mathbf{x}) = -2 \sum_{i=1}^m x_i \{\ln p_{0,i} - \ln(x_i/n)\}$$

is the *Generalised Likelihood Ratio* test statistic. From the general theorem

$$T_n(\mathbf{x}) \rightarrow_{\mathcal{D}(\cdot; \mathbf{p}_0)} \chi_{m-1}^2$$

since $\dim(\mathcal{S}_m) = m - 1$.

Many people rewrite this statistic as

$$\begin{aligned} T_n(\mathbf{x}) &= 2 \sum_{j=1}^m o_j \ln \left(\frac{o_j}{e_j} \right) \\ &= 2 \sum_{j=1}^m n_j \ln \left(\frac{x_j/n}{p_{0,j}} \right) \\ &= -2 \sum_{j=1}^m n_j \ln \left(\frac{x_j}{np_{0,j}} \right) \end{aligned}$$

where $o_j = n_j$ is the observed number in category j and $e_j = np_{0,j}$ is the expected number in category j . \square

Definition 3.23 - Pearson's χ^2 Test Statistic

Let $\mathbf{X} \sim \text{Categorical}(\mathbf{p})$ where $\mathbf{p} := (p_0, \dots, p_m)$ and \mathbf{x} is a realisation of \mathbf{X} .

We define *Pearson's χ^2 Test Statistic* as

$$T_{\text{Pearson}}(\mathbf{x}) := \sum_{j=1}^m \frac{(x_j - np_j)^2}{np_j} = \sum_{j=1}^m \frac{(o_j - e_j)^2}{e_j} \rightarrow_{\mathcal{D}(\cdot; \mathbf{p})} \chi_{m-1}^2$$

where o_j is the number of observations of category j and e_j is the expected number of observations of category j .

N.B. TODO - something about degrees of freedom.

4 Bayesian Inference

Theorem 4.1 - Bayes' Theorem

Consider $X \sim f(\cdot; \theta)$.

$$\underbrace{p(\theta|X)}_{\text{Posterior}} = \frac{\overbrace{p(X|\theta)}^{\text{Likelihood}} \overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(X)}_{\text{Evidence}}}$$

Definition 4.1 - Prior Distribution, $p(\theta)$

Consider random variable $X \sim f(\cdot; \theta)$.

A *Prior Distribution* encodes our beliefs about the model parameters, $p(\theta)$, before any data is observed. Typically *Priors* have less affect as the number of observed data points increases.

Definition 4.2 - Posterior, $\mathbb{P}(\theta|X)$

Consider random variable $X \sim f(\cdot; \theta)$ and let \mathbf{x} be a set of realisations of X .

A *Posterior Distribution* is used to learn possible values for the parameters of a model, given a set of observations from the model.

Definition 4.3 - Conjugacy

A *Prior* is said to be *Conjugate* if its distribution is in the same family as the *Poseterior*. When a *Prior* is not *Conjugate* one typically requires a computer to conduct *Bayesian Inference*.

Proposition 4.1 - Modelling Parameters

Consider $X \sim f(\cdot; \theta)$.

Here we can consider θ to be a realisation of some random variable ϑ and theorise the distribution of ϑ in our *Prior*.

N.B. Often β -Distributions are used, $\vartheta \sim \text{Beta}(\alpha, \beta)$. To set the values of α & β we set the *mean* & *variance* and then solve the resulting simultaneous equations.

Proposition 4.2 - Generalisation of Posterior Distributions

Consider *Random Variable* $\mathbf{X} \sim f_n(\cdot; \theta)$ and \mathbf{X} be a realisations of X .

Given a *Prior Distribution*, $p(\theta)$, we can generalise the *Posterior Distribution*

$$p(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}; \theta)p(\theta)}{\int_{\Theta} f_n(\mathbf{x}; \xi)p(\xi)d\xi} = \frac{L(\theta; \mathbf{x})p(\theta)}{\int_{\Theta} L(\xi; \mathbf{x})p(\xi)d\xi} \propto L(\theta; \mathbf{x})p(\theta)$$

N.B. This is equivalent to a *Maximum Likelihood Estimate* from the *Frequentist Approach*, but is a distribution rather than a point estimate.

Proposition 4.3 - Making Estimations

Given a *Poseterior Distribution*, $p(\theta|\mathbf{x})$, there are an infinite number of point estimates which could be used. Ones worth considering using are

- i) The *Mean* of the *Poseterior Distribution* (Common when the *Posterior* is *Unimodal*).

$$\hat{\theta} = \int_{\Theta} \theta p(\theta|\mathbf{x})d\theta$$

- ii) The *Maximum a Posteriori*. This might be misleading in certain situations.

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} p(\theta|\mathbf{x})$$

iii) The *Median of the Posterior Distribution* (Or other quantiles)

$$p(\theta \geq \hat{\theta}) = 0.5$$

iv) The *Variance of the Posterior Distribution* which depends on the model used.

$$\hat{\theta} = \text{Var}(X)$$

Definition 4.4 - Posterior Expected Loss

Consider a *Random Vector* $\mathbf{X} \sim f_n(\cdot; \theta)$, \mathbf{x} be a realisation of \mathbf{X} & $\hat{\theta}$ be an estimate of θ . Then the *Posterior Expected Loss* of $\hat{\theta}$ is defined to be

$$R(\hat{\theta}|\mathbf{x}) = \int_{\Theta} L(\theta, \hat{\theta}) p(\theta|\mathbf{x}) d\theta$$

where $L(\theta, \hat{\theta})$ is a non-negative *Loss Function*.

N.B. AKA *Posterior Risk*.

Proposition 4.4 - Loss Functions

A *Loss Function* is a measure of how much an estimate of a parameter deviates from the true value.

Some popular *Loss Functions* are

Name	Form	Bayes Estimate
Squared Error Loss	$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$	$\hat{\theta}_{\text{Bayes}} = \int_{\Theta} \theta p(\theta \mathbf{x}) d\theta$ (Posterior Mean)
Absolute Value	$L(\theta, \hat{\theta}) = \theta - \hat{\theta} $	$\hat{\theta}_{\text{Bayes}} = \theta$ where $p(\theta \mathbf{x}) = 0.5$ (Posterior Median)
	$L(\theta, \hat{\theta}) = \mathbb{1}(\theta \neq \hat{\theta})$	$\hat{\theta}_{\text{Bayes}} = \text{argmax}_{\theta \in \Theta} p(\theta \mathbf{x})$ (Posterior Mode)

Definition 4.5 - Bayes Estimate

Consider a *Random Vector* $\mathbf{X} \sim f_n(\cdot; \theta)$ & \mathbf{x} be a realisation of \mathbf{X} .

A *Bayes Estimate* of θ is

$$\hat{\theta}_{\text{Bayes}} = \text{argmin}_{\theta \in \Theta} R(\hat{\theta}|\mathbf{x})$$

i.e. The value which minimise the *Posterior Expected Loss*

4.1 Credible Intervals

Definition 4.6 - Symmetric Credible Interval

Consider a *Random Variable* $\mathbf{X} \sim f_n(\cdot; \theta)$ and a realisation \mathbf{x} of \mathbf{X} .

Let $\alpha \in [0, 1]$.

An interval (θ_1, θ_2) , for $\theta_1, \theta_2 \in \Theta$, is called a *Symmetric $(1 - \alpha)$ Credible Interval* if

$$\mathbb{P}(\theta \in [\theta_1, \theta_2]|\mathbf{x}) = \int_{\theta_1}^{\theta_2} p(\theta|\mathbf{x}) d\theta = 1 - \alpha$$

N.B. This is hard to generalise to the multidimensional scenario.

Definition 4.7 - High Posterior Density Set

Consider a *Random Variable* $\mathbf{X} \sim f_n(\cdot; \theta)$ and a realisation \mathbf{x} of \mathbf{X} .

Let $\alpha \in [0, 1]$.

The $(1 - \alpha)$ *High Posterior Density* is the *Level Set* defined as

$$\mathcal{HPD}_v := \{\theta \in \Theta : p(\theta|\mathbf{x}) \geq v\}$$

where v is chosen st

$$\mathbb{P}(\theta \in \mathcal{HPD}_v|\mathbf{x}) = \int_{\mathcal{HPD}_v} p(\theta|\mathbf{x}) d\theta = 1 - \alpha$$

N.B. These sets are difficult to compute without a computer.

Theorem 4.2 - The $\mathcal{HPD}(1 - \alpha)$ Credible Set is the smallest Subset of Θ containing exactly $1 - \alpha$ of the total density/probability.

4.2 Bayesian Hypothesis Testing

Remark 4.1 - Difference to Frequentist Approach

Consider a Random Variable $\mathbf{X} \sim f_n(\cdot; \theta)$ where θ is a realisation of ϑ and \mathbf{x} is a realisation of \mathbf{X} . Consider testing the Composite Hypotheses: $H_0 : \theta \in \Theta_1$, $H_1 : \theta \in \Theta_2$.

In the Bayesian Approach we actually calculate

$$\begin{aligned} \mathbb{P}(\vartheta \in \Theta_0 | \mathbf{X} = \mathbf{x}) &= \int_{\Theta_0} p(\theta | \mathbf{x}) d\theta \\ \text{and } \mathbb{P}(\vartheta \in \Theta_1 | \mathbf{X} = \mathbf{x}) &= \int_{\Theta_1} p(\theta | \mathbf{x}) d\theta \end{aligned}$$

Example 4.1 - Bayesian Hypothesis Testing

Consider the testing the Simple Hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

In a Bayesian framework we can take into account the cost of making an error, and base our decision on the minimisation of this cost.

We construct a loss table

Truth \ Decision	θ_0	θ_1
θ_0	L_{00}	L_{10}
θ_1	L_{10}	L_{11}

L_{ij} is the loss for choosing H_i when actually $\theta = \theta_j$.

We assume $L_{ik} > L_{jj}$ for $i \neq j$ (i.e. it is always more costly to make a wrong decision).

We aim to choose the hypothesis with the smallest Posterior Expected Cost.

When we choose H_0 we risk the following loss

$$R_0 = L_{00} \times p(\theta_0 | \mathbf{x}) + L_{01} \times p(\theta_1 | \mathbf{x})$$

and when we choose H_1 we risk the loss

$$R_1 = L_{10} \times p(\theta_0 | \mathbf{x}) + L_{11} \times p(\theta_1 | \mathbf{x})$$

Hence we choose H_1 if $R_1 < R_0$ (and visa-versa). i.e.

$$\begin{aligned} & L_{10} \times p(\theta_0 | \mathbf{x}) + L_{11} \times p(\theta_1 | \mathbf{x}) < L_{00} \times p(\theta_0 | \mathbf{x}) + L_{01} \times p(\theta_1 | \mathbf{x}) \\ \Rightarrow & p(\theta_0 | \mathbf{x}) [L_{10} - L_{00}] < p(\theta_1 | \mathbf{x}) [L_{01} - L_{11}] \\ \Rightarrow & \frac{p(\theta_1 | \mathbf{x})}{p(\theta_0 | \mathbf{x})} > \frac{L_{10} - L_{00}}{L_{01} - L_{11}} \\ \equiv & \frac{p(\mathbf{x} | \theta_1) p(\theta_1)}{p(\mathbf{x} | \theta_0) p(\theta_0)} > \frac{L_{10} - L_{00}}{L_{01} - L_{11}} \\ \Rightarrow & \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} > \frac{p(\theta_0) [L_{10} - L_{00}]}{p(\theta_1) [L_{11} - L_{10}]} \end{aligned}$$

Note that is exactly the same form as the *Neyman-Pearson Test*, expt that the "critical value" is chosen according to our prior and our assessment of the risk of taking the wrong decision.

In particular

- i) The greater the cost of a Type I Error, L_{10} , the higher the threshold.
- ii) The greater the cost of a Type II Error, L_{01} , the lower the threshold.
- iii) The greater the prior probability of H_0 , $p(\theta_0)$, the higher the threshold.
- iv) The greater the prior probability of H_1 , $p(\theta_1)$, the lower the threshold.

0 Appendix

0.1 Notation

Notation 0.1 - Convergence

$\{z_n\}_{n \in \mathbb{N}} \rightarrow z$ denotes that the sequence of deterministic values $\{z_n\}_{n \in \mathbb{N}}$ converges in value to $z \in \mathbb{R}$.

$\{Z_n\}_{n \in \mathbb{N}} \rightarrow_{\mathbb{P}} Z$ denotes that the sequence of random variables $\{Z_n\}_{n \in \mathbb{N}}$ converges in probability to random variable Z .

$\{Z_n\}_{n \in \mathbb{N}} \rightarrow_{\mathbb{P}(\cdot; \theta)} Z$ denotes that the sequence of random variables $\{Z_n\}_{n \in \mathbb{N}}$ converges in probability to random variable Z , dependent upon parameter θ .

$\{Z_n\}_{n \in \mathbb{N}} \rightarrow_{\mathcal{D}} Z$ denotes that the sequence of random variables $\{Z_n\}_{n \in \mathbb{N}}$ converges in distribution to random variable Z .

Notation 0.2 - Gamma Function

$$\Gamma(x) := \int_0^{\infty} t^{x-1} e^{-t} dt$$

0.2 Definitions

Definition 0.1 - Correlation

Let X & Y be random variables.

Correlation is a measure of dependence between two random variables

$$\text{Corr}(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \in [-1, 1]$$

Definition 0.2 - Covariance

Covariance is a measure of the joint variability of two random variables.

Consider random variable X & Y

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

If X & Y are independent then $\text{Cov}(X, Y) = 0$.

By definition of *Covariance* $\text{Cov}(X, X) = \text{Var}(X)$.

Definition 0.3 - Estimation

Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ with $\theta^* \in \Theta$ and \mathbf{x} be a realisation of \mathbf{X} .

As *Estimation* of model parameter θ^* is a statistic, $\hat{\theta}(\mathbf{x}) = T(\mathbf{x})$, which is intended to approximate the true value of θ^* .

N.B. Interchangeable with *Estimate*.

Definition 0.4 - Estimator

Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ with $\theta^* \in \Theta$ and \mathbf{x} be a realisation of \mathbf{X} .

An *Estimator* of model parameter θ^* is the random variable $\hat{\theta} := \hat{\theta}(\mathbf{X})$ where $\hat{\theta}(\mathbf{x})$ is an *estimation* of θ^* .

Definition 0.5 - Expectation

Expectation is the mean value for a random variable.

Consider *continuous* random variable X with pdf f_X and *discrete* random variable Y with pmf f_Y . Then

$$\mathbb{E}(X) := \int_{\mathbb{R}} x f_X(x) dx \quad \text{and} \quad \mathbb{E}(Y) := \sum_{y \in \mathcal{Y}} y p_Y(y)$$

For a function $g : \mathbb{R} \rightarrow \mathbb{R}$ we have

$$\mathbb{E}(g(X)) := \int_{\mathbb{R}} g(x)f_X(x)dx \quad \text{and} \quad \mathbb{E}(g(Y)) := \sum_{y \in \mathcal{Y}} g(y)p_Y(y)$$

For linear transformations of a random variable Z we find

$$\mathbb{E}(aZ + b) = a\mathbb{E}(Z) + b \quad \text{for } a, b \in \mathbb{R}$$

Definition 0.6 - Five-Number Summary

The *Five-Number Summary* of a sample contains the sample's: median; lower hinge; upper hinge; minimum value; & maximum value.

Definition 0.7 - Hinges

Hinges describe the spread of data in a sample, while trying to ignore extreme data. The *Lower Hinge*, H_1 , is the median of the set containing the median & values with rank less than the sample median. The *Upper Hinge*, H_3 , is the median of the set containing the median & values with rank greater than the sample median.

Definition 0.8 - Median

The *Median* is the central value of a data set.

Consider a data set x_0, \dots, x_n

- If $\exists m \in \mathbb{N}$ st $n = 2m + 1$ (i.e. n is odd) then the median is $x_{(m+1)}$.
- Else $\exists m \in \mathbb{N}$ st $n = 2m$ (i.e. n is even) then the median is $x_{(m+1)}$.

Definition 0.9 - Moments

The *Moments* of a random variable X are the expected values of powers of X .

$$n^{\text{th}} \text{ moment of } X := \mathbb{E}(X^n)$$

N.B. $\mathbb{E}(X^n) \neq \mathbb{E}(X)^n$.

Definition 0.10 - Order Statistic

An *Order Statistic* is a data set where the data has been placed in increasing order of value, not time. We use $x_{(i)}$ to denote the i^{th} lowest value in (x_0, \dots, x_n) .

Definition 0.11 - Quartiles

Quartiles describe the spread of data in a sample. The *Lower Quartile*, Q_1 , is the median of the set of values with rank less than the sample median. The *Upper Quartile*, Q_3 , is the median of the set of values with rank greater than the sample median.

N.B. These sets do not contain the median.

Definition 0.12 - Sample Mean

The *Sample Mean* is the mean value of all data points within a sample. Consider a sample $\{x_1, \dots, x_n\}$

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

Definition 0.13 - Sample Variance

Sample Variance is a measure of spread of data in a sample around the sample mean. For a sample $\{x_1, \dots, x_n\}$

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

Definition 0.14 - Statistic

Let \mathbf{x} be some data.

A *Statistic* is any function of the data, $T(\mathbf{x})$.

N.B. *Statistics* are independent of unknown model parameters.

Definition 0.15 - Trimmed Sample Mean

The *Trimmed Sample Mean* is the average value of a subset of data points within a sample. The subset is defined to ignore the $\frac{\Delta}{2}\%$ largest & smallest values of the sample. For a $\Delta\%$ trimmed mean we define

$$\bar{x}_{\Delta} := \frac{1}{n - 2k} \sum_{i=k+1}^{n-k-1} x_i \text{ with } k = \left\lfloor \frac{n\Delta}{100} \right\rfloor$$

Definition 0.16 - Variance

Variance measures how far a set of random numbers are spread from their average value.

Consider random variable X

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

For linear transformation of a random variable X we find

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

For a linear transformation of two random variables X & Y we have

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y) \quad \text{for } a, b \in \mathbb{R}$$

Definition 0.17 - Skew

Skew describes the spread of values in a sample which are less than the median, relative to the spread of values greater than the median. A sample is *Left-Skewed* if $|H_3 - H_2| < |H_1 - H_2|$. A sample is *Right-Skewed* if $|H_3 - H_2| > |H_1 - H_2|$.

0.3 Theorems**Theorem 0.1 - Cauchy-Schwarz Inequality**

Let X & Y be real-valued random variables in the same probability space. Then

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Theorem 0.2 - Chebyshev's Inequality

Let X be a random variable.

Define $\mu := \mathbb{E}(X)$ and $\sigma^2 := \text{Var}(X)$. Then

$$\forall a > 0 \quad \mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

Theorem 0.3 - Covariance Inequality

Let X & Y be real-valued random variables in the same probability space. Then

$$\text{Cov}(X, Y)^2 \leq \text{Var}(X)\text{Var}(Y)$$

Theorem 0.4 - Joint Probability Density of Simple Random Sample

Let \mathbf{X}_1, \dots, X_n be a set of independent random variables with pdfs f_{X_1}, \dots, f_{X_n} , respectively,

and x_1, \dots, x_n be a realisation of X_1, \dots, X_n .

The probability of obtaining x_1, \dots, x_n is

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i; \theta)$$

Theorem 0.5 - Markov's Inequality

Let $X \sim f_X(\cdot)$ be a non-negative continuous random variable. Then

$$\forall a > 0 \quad \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

0.4 Probability Distributions

Definition 0.18 - β -Distribution

Let $X \sim \text{Beta}(\alpha, \beta)$.

A *continuous* random variable with shape parameters $\alpha, \beta > 0$. Then

$$\begin{aligned} f_X(x) &\propto x^{\alpha-1}(1-x)^{\beta-1} \mathbb{1}\{x \in [0, 1]\} \\ \mathbb{E}(X) &= \frac{\alpha}{\alpha + \beta} \\ \text{Var}(X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\ \mathcal{M}_X(t) &= 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r} \right) \frac{t^k}{k!} \end{aligned}$$

Definition 0.19 - Bernoulli Distribution

Let $X \sim \text{Bernoulli}(p)$.

A *discrete* random variable which takes 1 with probability p & 0 with probability $(1 - p)$. Then

$$\begin{aligned} p_X(k) &= \begin{cases} 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \\ 0 & \text{otherwise} \end{cases} \\ P_X(k) &= \begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } k \in [0, 1) \\ 1 & \text{otherwise} \end{cases} \\ \mathbb{E}(X) &= p \\ \text{Var}(X) &= p(1 - p) \\ \mathcal{M}_X(t) &= (1 - p) + pe^t \end{aligned}$$

N.B. Often we define $q := 1 - p$ for simplicity.

Definition 0.20 - Binomial Distribution

Let $X \sim \text{Binomial}(n, p)$.

A *discrete* random variable modelled by a *Binomial Distribution* on n independent events and rate of success p .

$$\begin{aligned} p_X(k) &= \binom{n}{k} p^k (1 - p)^{n-k} \\ P_X(k) &= \sum_{i=1}^k \binom{n}{i} p^i (1 - p)^{n-i} \\ \mathbb{E}(X) &= np \\ \text{Var}(X) &= np(1 - p) \\ \mathcal{M}_X(t) &= [(1 - p) + pe^t]^n \end{aligned}$$

N.B. If $Y := \sum_{i=1}^n X_i$ where $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$ then $Y \sim \text{Binomial}(n, p)$.

Definition 0.21 - Categorical Distribution

Let $X \sim \text{Categorical}(\mathbf{p})$.

A *discrete* random variable where probability vector \mathbf{p} for a set of events $\{1, \dots, m\}$.

$$f_X(i) = p_i$$

Definition 0.22 - χ^2 Distribution

Let $X \sim \chi_r^2$.

A *continuous* random variable modelled by the χ^2 Distribution with r degrees of freedom. Then

$$\begin{aligned} f_X(x) &= \frac{1}{2^{r/2} \Gamma(r/2)} x^{\frac{r}{2}-1} e^{-\frac{x}{2}} \\ F_X(x) &= \frac{1}{\Gamma(k/2)} \gamma\left(\frac{r}{2}, \frac{x}{2}\right) \\ \mathbb{E}(X) &= r \\ \text{Var}(X) &= 2r \\ \mathcal{M}_X(t) &= \mathbb{1}\{t < \frac{1}{2}\} (1 - 2t)^{-\frac{r}{2}} \end{aligned}$$

N.B. If $Y := \sum_{i=1}^k Z_i^2$ with $\mathbf{Z} \stackrel{\text{iid}}{\sim} \text{Normal}(0, 1)$ then $Y \sim \chi_k^2$.

Definition 0.23 - Exponential Distribution

Let $X \sim \text{Exponential}(\lambda)$.

A *continuous* random variable modelled by a *Exponential Distribution* with rate-parameter λ . Then

$$\begin{aligned} f_X(x) &= \mathbb{1}\{t \geq 0\} \cdot \lambda e^{-\lambda x} \\ F_X(x) &= \mathbb{1}\{t \geq 0\} \cdot (1 - e^{-\lambda x}) \\ \mathbb{E}(X) &= \frac{1}{\lambda} \\ \text{Var}(X) &= \frac{1}{\lambda^2} \\ \mathcal{M}_X(t) &= \mathbb{1}\{t < \lambda\} \frac{\lambda}{\lambda - t} \end{aligned}$$

N.B. Exponential Distribution is used to model the wait time between decays of a radioactive source.

Definition 0.24 - Gamma Distribution

Let $X \sim \Gamma(\alpha, \beta)$.

A *continuous* random variable modelled by a *Gamma Distribution* with shape parameter $\alpha > 0$ & rate parameter β . Then

$$\begin{aligned} f_X(x) &= \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x} \\ F_X(x) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha)} (\alpha, \beta x) \\ \mathbb{E}(X) &= \frac{\alpha}{\beta} \\ \text{Var}(X) &= \frac{\alpha}{\beta^2} \\ \mathcal{M}_X(t) &= \mathbb{1}\{t < \beta\} \left(1 - \frac{t}{\beta}\right)^{-\alpha} \end{aligned}$$

N.B. There is an equivalent definition of a *Gamma Distribution* in terms of a shape & scale parameter. The scale parameter is 1 over the rate parameter in this definition.

Definition 0.25 - Multinomial Distribution

Let $\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p})$.

A *discrete* random variable which models n events with probability vector \mathbf{p} for events $\{1, \dots, m\}$.

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \mathbb{1} \left\{ \sum_{i=1}^m x_i \equiv n \right\} \frac{n!}{x_1! \cdots x_n!} \prod_{i=1}^n p_i^{x_i} \\ \mathbb{E}(X_i) &= np_i \\ \text{Var}(X_i) &= np_i(1 - p_i) \\ \text{Cov}(X_i, x_j) &= -np_i p_j \text{ for } i \neq j \\ \mathcal{M}_{X_i}(\theta_i) &= \left(\sum_{i=1}^m p_i e^{\theta_i} \right)^n \end{aligned}$$

N.B. In a realisation \mathbf{x} of \mathbf{X} , x_i is the number of times event i has occurred.

Definition 0.26 - Normal Distribution

Let $X \sim \text{Normal}(\mu, \sigma^2)$.

A *continuous* random variable with mean μ & variance σ^2 .

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ F_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\ \mathbb{E}(X) &= \mu \\ \text{Var}(X) &= \sigma^2 \\ \mathcal{M}_X(\theta) &= e^{\mu\theta + \sigma^2\theta^2(1/2)} \end{aligned}$$

Definition 0.27 - Pareto Distribution

Let $X \sim \text{Pareto}(x_0, \theta)$.

A *continuous* random variable modelled by a *Pareto Distribution* with minimum value x_0 & shape parameter $\alpha > 0$. Then

$$\begin{aligned} f_X(x) &= \frac{\alpha x_0^\alpha}{x^{\alpha+1}} \\ F_X(x) &= 1 - \left(\frac{x_0}{x} \right)^\alpha \\ \mathbb{E}(X) &= \begin{cases} \infty & \alpha \leq 1 \\ \frac{\alpha x_0}{\alpha - 1} & \alpha > 1 \end{cases} \\ \text{Var}(X) &= \begin{cases} \infty & \alpha \leq 2 \\ \frac{x_0^2 \alpha}{(\alpha - 1)^2 (\alpha - 2)} & \alpha > 2 \end{cases} \\ \mathcal{M}_X(t) &= \mathbb{1}\{t < 0\} \alpha (-x_0 t)^{\alpha-1} \Gamma(-\alpha, -x_0 t) \end{aligned}$$

Definition 0.28 - Poisson Distribution

Let $X \sim \text{Poisson}(\lambda)$.

A *discrete* random variable modelled by a *Poisson Distribution* with rate parameter λ . Then

$$\begin{aligned} p_X(k) &= \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{for } k \in \mathbb{N}_0 \\ P_X(k) &= e^{-\lambda} \sum_{i=1}^k \frac{\lambda^i}{i!} \\ \mathbb{E}(X) &= \lambda \\ \text{Var}(X) &= \lambda \\ \mathcal{M}_X(t) &= e^{\lambda(e^t - 1)} \end{aligned}$$

N.B. Poisson Distribution is used to model the number of radioactive decays in a time period.

Definition 0.29 - *t*-Distribution

Let $X \sim t_r$.

A *continuous* random variable with r degrees of freedom. Then

$$\begin{aligned} f_X(k) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ \mathbb{E}(X) &= \begin{cases} 0 & \text{if } \nu > 1 \\ \text{undefined} & \text{otherwise} \end{cases} \\ \text{Var}(X) &= \begin{cases} \frac{\nu}{\nu-2} & \text{if } \nu > 2 \\ \infty & 1 < \nu \leq 2 \\ \text{undefined} & \text{otherwise} \end{cases} \\ \mathcal{M}_X(t) &= \text{undefined} \end{aligned}$$

N.B. Let $Y \sim \text{Normal}(0, 1)$ & $Z \sim \chi_r^2$ be independent random variables then $X := \frac{Y}{\sqrt{Z/r}} \sim t_r$.

Definition 0.30 - Uniform Distribution - Uniform

Let $X \sim \text{Uniform}(a, b)$.

A *continuous* random variable with lower bound a & upper bound b . Then

$$\begin{aligned} f_X(x) &= \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \\ F_X(x) &= \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & \text{otherwise} \end{cases} \\ \mathbb{E}(X) &= \frac{1}{2}(a+b) \\ \text{Var}(X) &= \frac{1}{12}(b-a)^2 \\ \mathcal{M}_X(t) &= \begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & t \neq 0 \\ 1 & t = 0 \end{cases} \end{aligned}$$

0.5 Identities

0.5.1 Likelihood

Proposition 0.1 - Binomial

Let $X \sim \text{Binomial}(n, p)$ with n & p unknown and x be a realisation of X . Then

$$\begin{aligned} L(n, p; x) &\propto \binom{n}{x} p^x (1-p)^{n-x} \\ \ell(n, p; \mathbf{x}) &= \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p) + C \\ \hat{n}_{\text{MLE}} &= \frac{x}{\hat{p}} \\ \hat{p}_{\text{MLE}} &= \frac{x}{\hat{n}} \end{aligned}$$

Proposition 0.2 - Normal

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$ with μ & σ^2 unknown and \mathbf{x} be a realisation of \mathbf{X} . Then

$$\begin{aligned} L(\mu, \sigma^2; \mathbf{x}) &\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ \ell(\mu, \sigma^2; \mathbf{x}) &= -n \ln \sigma^2 - \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 + C \\ \hat{\mu}_{\text{MLE}} &= \bar{\mathbf{x}} \\ \hat{\sigma}_{\text{MLE}}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \end{aligned}$$

Proposition 0.3 - Poisson

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda)$ with λ unknown and \mathbf{x} be a realisation of \mathbf{X} . Then

$$\begin{aligned} L(\lambda; \mathbf{x}) &\propto e^{-\lambda n} \lambda^{n\bar{x}} \\ \ell(\lambda; \mathbf{x}) &= -\lambda n + n\bar{x} \ln \lambda + C \\ \hat{\lambda}_{\text{MLE}} &= \bar{x} \end{aligned}$$

Proposition 0.4 - Uniform

Let $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Uniform}(a, b)$ with a & b unknown and \mathbf{x} be a realisation of \mathbf{X} . Then

$$\begin{aligned} L(a, b; \mathbf{x}) &\propto \begin{cases} \frac{1}{(b-a)^n} & a \leq x_i \leq b \ \forall x_i \in \mathbf{x} \\ 0 & \text{otherwise} \end{cases} \\ \ell(a, b; \mathbf{x}) &= \begin{cases} -\ln(b-a) & a \leq x_i \leq b \ \forall x_i \in \mathbf{x} \\ 0 & \text{otherwise} \end{cases} \\ \hat{a}_{\text{MLE}} &= \min\{x_i : x_i \in \mathbf{x}\} \\ \hat{b}_{\text{MLE}} &= \max\{x_i : x_i \in \mathbf{x}\} \end{aligned}$$