

# Statistics 2 - Notes

Dom Hutchinson

November 26, 2019

## Contents

|          |                                                                                           |           |
|----------|-------------------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Estimation</b>                                                                         | <b>2</b>  |
| 1.1      | Introduction . . . . .                                                                    | 2         |
| 1.2      | The Likelihood Function . . . . .                                                         | 3         |
| 1.3      | Maximum Likelihood Estimates . . . . .                                                    | 4         |
| 1.4      | Determining MLEs - The Tractable Case . . . . .                                           | 5         |
| 1.5      | Statistics and Estimators . . . . .                                                       | 7         |
| 1.6      | Probabilistic Convergence . . . . .                                                       | 9         |
| 1.7      | Probabilistic Convergence & Estimators . . . . .                                          | 11        |
| 1.8      | The Fisher Information . . . . .                                                          | 12        |
| 1.9      | Efficiency and The Cramer-Rao Bound . . . . .                                             | 15        |
| 1.10     | Asymptotic Distribution of the Maximum Likelihood Estimator . . . . .                     | 17        |
| 1.11     | Confidence Sets Around the Maximum Likelihood Estimator . . . . .                         | 19        |
| 1.12     | Asymptotic Approximation of Confidence Intervals . . . . .                                | 21        |
| 1.13     | Estimating the Information for Maximum Likelihood Estimates . . . . .                     | 22        |
| 1.14     | Transformations and Confidence Intervals . . . . .                                        | 25        |
| 1.15     | Likelihood Ratio Confidence Sets - Wilk's Approach . . . . .                              | 27        |
| 1.16     | Transformation Invariant Confidence Sets . . . . .                                        | 29        |
| <b>2</b> | <b>Testing</b>                                                                            | <b>30</b> |
| 2.1      | Introduction to Hypothesis Tests . . . . .                                                | 30        |
| 2.2      | Hypothesis Testing - Significance and Power . . . . .                                     | 31        |
| 2.2.1    | Power . . . . .                                                                           | 32        |
| 2.3      | Designing Tests - Neyman-Pearson Approach . . . . .                                       | 34        |
| 2.4      | Testing - p-Values, Equivalent Test Statistics and Computing the Power Function . . . . . | 36        |
| 2.5      | Uniformly Most Powerful Tests . . . . .                                                   | 38        |
| 2.6      | Generalised Likelihood Ratio Test . . . . .                                               | 40        |
| 2.7      | Categorical Distributions and Pearson's Chi-Squared Test . . . . .                        | 42        |
| <b>0</b> | <b>Appendix</b>                                                                           | <b>43</b> |
| 0.1      | Notation . . . . .                                                                        | 43        |
| 0.2      | R . . . . .                                                                               | 43        |
| 0.3      | Probability Distributions . . . . .                                                       | 43        |

# 1 Estimation

## 1.1 Introduction

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

A mathematical construct for modelling the real world. A *Probabilty Space* has three elements

- i)  $\Omega$  - Sample space.
- ii)  $\mathcal{F}$  - Set of events.
- iii)  $\mathbb{P}$  - Probability measure.

and most fulfil the following conditions

- i)  $\Omega \in \mathcal{F}$ ;
- ii)  $\forall A \in \mathcal{F} \implies A^c \in \mathcal{F}$ ;
- iii)  $\forall A_0, \dots, A_n \in \mathcal{F} \implies \left( \bigcup_i A_i \right) \in \mathcal{F}$ ;
- iv)  $\mathbb{P}(\Omega) = 1$ ; and,
- v)  $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$  for disjoint  $A_1, A_2, \dots$  (Countable Additivity).

**Definition 1.2** - *Random Variable*

A function which maps an event in the sample space to a value *e.g.*  $X : \Omega \rightarrow \mathbb{R}$ .

**Remark 1.1** - *Probability Density Function for iid Random Variable Vector*

For  $\mathbf{X} \sim f_n(\cdot; \theta)$  where each component of  $\mathbf{X}$  is independent and identically distribution the probability density function of  $\mathbf{X}$  is

$$f_n(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

**Definition 1.3** - *Expectation*

The mean value for a random variable. For rv  $X$

$$\mathbb{E}(X) := \sum_{x \in \mathcal{X}} x f_X(x) \quad \& \quad \mathbb{E}(X) := \int_{\mathbb{R}} x f_X(x) dx$$

**Theorem 1.1** - *Expection of a Function*

For a function  $g : \mathbb{R} \rightarrow \mathbb{R}$  and rv  $X$  with pmf  $f_X$

$$\mathbb{E}(g(X)) := \sum_{g(x) \in \mathcal{X}} x f_X(x) \quad \& \quad \mathbb{E}(g(X)) := \int_{\mathbb{R}} g(x) f_X(x) dx$$

**Theorem 1.2** - *Expectation of a Linear Operator*

For rv  $X$  with pmf  $f_X$  &  $a, b \in \mathbb{R}$

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

**Definition 1.4** - *Variance*

For rv  $X$

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

**Theorem 1.3 - Variance of a Linear Operator**

For rv  $X$  and  $a, b \in \mathbb{R}$

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

**Definition 1.5 - Moment of a Random Variable**

For rv  $X$  the  $n^{\text{th}}$  moment of  $X$  is defined as  $\mathbb{E}(X^n)$ .

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

**Definition 1.6 - Covariance**

For rv  $X$  &  $Y$

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

**Theorem 1.4 - Properties of Covariance**

Let  $X$  &  $Y$  be independent random variables

i)  $\text{Cov}(X, X) = \text{Var}(X)$ ;

ii)  $\text{Cov}(X, Y) = 0$

**Theorem 1.5 - Variance of two Random Variables with linear operators**

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

**Theorem 1.6 - Independent Random Variables**

Random variables  $X_1, \dots, X_n$  are independent iff

$$\mathbb{P}(X_1 \leq a_1, \dots, X_n \leq a_n) = \prod_{i=1}^n \mathbb{P}(X_i \leq a_i) \quad \forall a_1, \dots, a_n \in \mathbb{R}$$

**1.2 The Likelihood Function****Definition 2.1 - Likelihood Function**

Define  $\mathbf{X} \sim f_n(\cdot; \theta^*)$  for some unknown  $\theta^* \in \Theta$  and let  $\mathbf{x}$  be an observation of  $\mathbf{X}$ .

A *Likelihood Function* is any function,  $L(\cdot; \mathbf{x}) : \Theta \rightarrow [0, \infty)$ , which is proportional to the PMF/PDF of the observed realisation  $\mathbf{x}$ .

$$L(\theta; \mathbf{x}) := C f_b(\mathbf{x}; \theta) \quad \forall C > 0$$

*N.B.* Sometimes this is called the *Observed Likelihood Function* since it is dependent on observed data.

**Definition 2.2 - Log-Likelihood Function**

Let  $\mathbf{X} \sim f_n(\cdot; \theta^*)$  for some unknown  $\theta^* \in \Theta$  and  $\mathbf{x}$  be an observation of  $\mathbf{X}$ .

The *Log-Likelihood Function* is the natural log of a *Likelihood Function*

$$\ell(\theta; \mathbf{x}) := \ln f_n(\mathbf{x}; \theta) + C, \quad C \in \mathbb{R}$$

**Theorem 2.1 - Multidimensional Transforms**

Let  $\mathbf{X}$  be a continuous random vector in  $\mathbb{R}^n$  with PDF  $f_{\mathbf{X}}$ ;  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a continuous differentiable bijection; and,  $h := g^{-1}$ .

Then  $\mathbf{Y} = g(\mathbf{X})$  is a continuous random vector and its PDF is

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

where

$$J_h := \left| \det \left( \frac{\partial h}{\partial \mathbf{y}} \right) \right|$$

**Proposition 2.1** - *Invariance of Likelihood Function by bijective transformation of the observations independent of  $\theta$*

Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a bijective transformation which is independent of  $\theta$ ; and  $\mathbf{Y} := g(\mathbf{X})$ .

Then  $\mathbf{Y}$  is a random variable with PDF/PMF

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

Hence, if  $\mathbf{y} = g(\mathbf{x})$  then  $L_{\mathbf{Y}}(\theta; \mathbf{y}) \propto L_{\mathbf{X}}(\theta; \mathbf{x})$

**Proof 2.1** - *Proposition 2.1*

Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a bijective transformation which is independent of  $\theta$ ;  $h := g^{-1}$ ;  $\mathbf{X}, \mathbf{Y}$  be a rvs st  $\mathbf{Y} := g(\mathbf{X})$ .

i) *Discrete Case* - Consider the case when  $\mathbf{X}$  is a discrete rv. Then

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

ii) *Continuous Case* - Consider the case when  $\mathbf{X}$  is a continuous rv.

Then, by **Theorem 2.1**

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

Since  $J_{g^{-1}}$  does not depend on  $\theta$  this case is solved.

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

### 1.3 Maximum Likelihood Estimates

**Definition 3.1** - *Maximum Likelihood Estimate*

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

The *Maximum Likelihood Estimate* is the value  $\hat{\theta} \in \Theta$  st

$$\forall \theta \in \Theta \quad f_n(\mathbf{x}; \hat{\theta}) \geq f_n(\mathbf{x}, \theta)$$

Equivalently

$$\forall \theta \in \Theta \quad L(\hat{\theta}; \mathbf{x}) \geq L(\theta; \mathbf{x}) \quad \text{or} \quad \ell(\hat{\theta}; \mathbf{x}) \geq \ell(\theta; \mathbf{x})$$

i.e.  $\hat{\theta}(\mathbf{x}) := \operatorname{argmax}_{\theta} (L(\theta; \mathbf{x}))$ .

**Remark 3.1** - *The Maximum Likelihood Estimate may not be unique*

**Example 3.1** - *MLE for Uniform Distribution*

Consider  $\mathbf{X} \stackrel{\text{iid}}{\sim} U[0, \theta]$  for  $\theta > 0$ .

Then

$$\begin{aligned} L(\theta; \mathbf{x}) &\propto f_n(\mathbf{x}; \theta) \\ &= \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}\{x_i \in [0, \theta]\} \\ &= \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}\{x_i \in [0, \theta]\} \\ \implies \hat{\theta} &= \max\{x_i : x_i \in \mathbf{x}\} \end{aligned}$$

**Remark 3.2 - MLE of Reparameterisation**

Define  $\tau(\theta) : \mathbb{R} \rightarrow \mathbb{R}$ . Then

$$\hat{\tau} = \tau(\hat{\theta})$$

*N.B.* We often write  $\tilde{f}$  to represent the pmf when  $\tau$  is taken as a parameter rather than  $\theta$ . *i.e.*  $f(x; \theta) = \tilde{f}(x; \tau(\theta))$ .

**Theorem 3.1 - Invariance of MLE under bijective Reparameterisation**

Let  $g : \Theta \rightarrow G$  be a bijective transformation of the statistical parameter  $\theta$ .

Let  $\mathbf{X} \sim f(\cdot; \theta) = \tilde{f}(\cdot; g(\theta))$  for some  $\theta$ , and let  $\mathbf{x}$  be a realisation of  $\mathbf{X}$ .

If  $\hat{\theta}$  is an MLE of  $\theta$  then  $\hat{\tau} = g(\hat{\theta})$  is an MLE of  $\tau$ .

**Proof 3.1 - Theorem 3.1**

*This is a proof by contradiction.*

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

□

**Remark 3.3 - Not all Reparameterisations are Bijective**

When reparameterisations  $g : \mathbb{R} \rightarrow \mathbb{R}$  is not bijective it is helpful to consider the *induced likelihood*

$$L^*(\tau; \mathbf{x}) := \max_{\theta \in G_\tau} L(\theta; \mathbf{x}) \text{ where } G_\tau := \{\theta : g(\theta) = \tau\}$$

Since this reduces the domain to only where  $g$  is bijective.

**1.4 Determining MLEs - The Tractable Case****Proposition 4.1 - Differentiable Likelihood in the continuous case - Multivariate**

When  $L(\theta; \mathbf{x})$  is differentiable one can find MLEs by considering its extrema. This is done equating & solving the cases when the gradient is zero, *i.e.*  $\nabla L(\theta; \mathbf{x}) = 0$ , and then checking whether this is a maximum or minimum point.

A point is a local minimum if the Hessian at the point is *Negative Definite* *i.e.*  $\mathbf{x}^T \mathbf{A} \mathbf{x} < 0 \forall \mathbf{x} \neq \mathbf{0}$ .

**Example 4.1 - MLE of Normal Distribution**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned}
 L(\mu, \sigma^2; \mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\
 \Rightarrow \ell(\mu, \sigma^2; \mathbf{x}) &= C - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\
 \Rightarrow \nabla \ell(\mu, \sigma^2; \mathbf{x}) &= \left( \frac{-1}{\sigma^2} \sum_{i=1}^n (x_i - \mu), \quad -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 \right) \\
 \text{Setting } \frac{-1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) &= 0 \\
 \Rightarrow \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \\
 \text{Setting } -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 &= 0 \\
 \Rightarrow \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i \hat{\mu})^2
 \end{aligned}$$

We now want to check whether  $(\hat{\mu}, \hat{\sigma}^2)$  is a minimum.

$$\begin{aligned}
 \nabla^2 \ell(\mu, \sigma^2; \mathbf{x}) &= \begin{pmatrix} \frac{\partial^2 \ell(\mu, \sigma^2; \mathbf{x})}{\partial \mu^2} & \frac{\partial^2 \ell(\mu, \sigma^2; \mathbf{x})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ell(\mu, \sigma^2; \mathbf{x})}{\partial \mu \partial \sigma^2} & \frac{\partial^2 \ell(\mu, \sigma^2; \mathbf{x})}{\partial (\sigma^2)^2} \end{pmatrix} \\
 &= \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix}
 \end{aligned}$$

Since  $\begin{pmatrix} z_1 & z_2 \end{pmatrix} \begin{pmatrix} -a & 0 \\ 0 & -b \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = -az_1^2 - bz_2^2 < 0 \forall a, b > 0$  and we have  $\frac{n}{\hat{\sigma}^2}, \frac{n}{2\hat{\sigma}^4} > 0$  then we can conclude that  $\nabla^2 \ell$  is negative definite.

Thus  $\hat{\mu} = \bar{x}$  &  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i \hat{\mu})^2$  is an MLE for the normal distribution.

#### Example 4.2 - MLE for Capture-Recapture Model

Suppose you are wanting to calculate the unknown size of a population,  $n$ . The Capture-Recapture Model is one technique that can be used. You tag  $t \leq n$  members of the population; wait for a while; then recapture  $c \leq n$  members of which  $x \leq \min\{t, c\} \leq n$  are tagged.

With  $t, c, x$  known produce a MLE for  $n$ .

We first work out the associated probability distribution for  $X$ , the population size. We have

- i)  $\binom{t}{x}$  ways of choosing  $x$  members among the tagged ones;
- ii)  $\binom{n-t}{c-x}$  ways of choosing the remaining members among the non-tagged ones;
- iii)  $\binom{n}{c}$  ways of choosing  $c$  members in a population of  $n$  individuals.

Thus

$$f_X(x; n) = \frac{\binom{t}{x} \binom{n-t}{c-x}}{\binom{n}{c}}$$

This means that  $X \sim \text{Hypergeometric}(t, n, c)$  with  $t$  &  $c$  known.

Now we calculate the MLE for  $X$

$$\begin{aligned}
 L(n; x) &= f_X(x; n) \\
 &= \frac{\binom{t}{x} \binom{n-t}{c-x}}{\binom{n}{c}} \\
 &= \frac{t!}{x!(t-x)!} \frac{(n-t)!}{(c-x)!(n-t-c+x)!} \\
 &= \frac{n!}{c!(n-c)!}
 \end{aligned}$$

Now we consider  $L(n; x) = 0$  when  $x > \min\{t, c\}$ . We want to identify values of  $n$  for which  $L(n; x) \geq L(n-1; x)$ .

Consider  $n-1 \geq \min\{t, c\} \implies L(n-1; x) > 0$

$$\begin{aligned}
 \text{Let } r(n) &:= \frac{L(n; x)}{L(n-1; x)} \\
 &= \frac{n-t}{n-t-c+x} \frac{n-c}{n} \\
 \Rightarrow 1 &\leq r(n) \\
 \Leftrightarrow 1 &\leq \frac{n-t}{n-t-c+x} \frac{n-c}{n} \\
 \Leftrightarrow n(n-t-c+x) &\leq (n-t)(n-c) \\
 \Leftrightarrow n^2 - nt - cn + xn &\leq n^2 - nt - cn + ct \\
 \Leftrightarrow xn &\leq ct \\
 \Leftrightarrow x &\leq \frac{ct}{n}
 \end{aligned}$$

So  $L(n; x)$  is increasing for  $n \leq \lfloor \frac{ct}{x} \rfloor$  & decreasing for  $n > \lfloor \frac{ct}{x} \rfloor$ .

Consequently  $\hat{n}_{\text{MLE}}(x) = \lfloor \frac{ct}{x} \rfloor$

## 1.5 Statistics and Estimators

### Definition 5.1 - Statistic

Given some data  $\mathbf{x}$  a statistic is a function of the data  $T(\mathbf{x})$ .

*N.B.* A statistic cannot depend on an unknown statistical parameter.

### Definition 5.2 - Estimate

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

An *Estimate*  $\theta^*$  is a statistic  $\hat{\theta}(\mathbf{x}) = T(\mathbf{x})$  which is intended to approximate the real value of  $\theta^*$ .

*N.B.* An *Estimate* is a real value & thus is hard to evaluate.

### Definition 5.3 - 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  $\theta^*$  is  $\hat{\theta}$  where  $\hat{\theta}(\mathbf{x})$  is an *estimate*.

*N.B.* We call  $T(\mathbf{X})$  an estimator. This is a random variable.

### Definition 5.4 - Distribution of an Estimator

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

If  $\hat{\theta}(\mathbf{X})$  is a real-valued random variable, we can write its CDF as

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

### Remark 5.1 - Estimator depends upon true value

The distribution of  $\hat{\theta}(\mathbf{X})$  depends on the distribution of  $\mathbf{X}$  which in turn depends upon the

distribution of  $\theta^*$ .

Thus the distribution of an estimator depends on the true parameter of the variable it is estimating.

**Remark 5.2 - Estimator Distribution & Sample Size**

As sample size increases the distribution of an estimator may converge to a more standard distribution (e.g. Normal, Poisson).

**Definition 5.5 - Bias**

*Bias* is a measure of how much an estimator deviates from the true value, on average.

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

**Definition 5.6 - Unbiased Estimator**

An *Estimator*,  $\hat{\theta}$ , is said to be *Unbiased* if  $\forall \theta \in \Theta$ ,  $\text{Bias}(\hat{\theta}; \theta) = 0$ .  
Equivalently  $\mathbb{E}(\hat{\theta}; \theta) = \theta$ .

**Definition 5.7 - Mean Square Error**

The *Mean Square Error* of an estimator is the mean of the squared error associated with rv  $\hat{\theta}$ .

$$MSE(\hat{\theta}; \theta^*) := \mathbb{E} \left[ (\hat{\theta}(\mathbf{X}) - \theta^*)^2; \theta^2 \right]$$

**Proposition 5.1 - Simplification of MSE Formula**

The MSE is a combination of variance & bias.

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

**Example 5.1 - Sample mean as an Estimator**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda^*)$ .

Suppose we are using the sample mean,  $\hat{\lambda}(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n x_i$ , as an estimate of  $\lambda^*$ . We first want to show this estimator is *Unbiased*

$$\begin{aligned}\mathbb{E}(\hat{\lambda}; \lambda) &= \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n X_i; \lambda \right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i; \lambda) \\ &= \frac{1}{n} n \lambda \\ &= \lambda\end{aligned}$$

Thus  $\hat{\lambda}$  is unbiased.

Now we consider the MSE of  $\hat{\lambda}$

$$\begin{aligned}MSE(\hat{\lambda}; \lambda) &= \text{Var}(\hat{\lambda}; \lambda) \\ &= \text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i; \lambda \right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i; \lambda) \\ &= \frac{1}{n^2} n \lambda \\ &= \frac{\lambda}{n}\end{aligned}$$

This shows that as the sample size increases the MSE of  $\hat{\lambda}$  converges to 0.



## 1.6 Probabilistic Convergence

### Remark 6.1 - Motivation

Here we consider the properties of a maximum likelihood estimators as the sample size increases.

### Theorem 6.1 - Markov's Inequality

For a *non-negative* random variable  $X$  and a constant  $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

### Proof 6.1 - Markov's Inequality

Consider continuous  $X$ . We have

$$\begin{aligned} a\mathbb{P}(X \geq a) &= a \int_a^\infty f_X(x) dx \\ &\leq \int_a^\infty x f_X(x) dx \\ &\leq \int_0^\infty x f_X(x) dx \\ &= \mathbb{E}(X) \\ \implies a\mathbb{P}(X \geq a) &= \mathbb{E}(X) \\ \implies \mathbb{P}(X \geq a) &\leq \frac{\mathbb{E}(X)}{a} \end{aligned}$$

□

### Theorem 6.2 - Chebyshev's Inequality

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

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

### Proof 6.2 - Chebyshev's Inequality

We have

$$\begin{aligned} \mathbb{P}(|X - \mu| \geq a) &= \mathbb{P}(|X - \mu|^2 \geq a^2) \\ &\leq \frac{\mathbb{E}((X - \mu)^2)}{a^2} \text{ By Markov's Inequality} \\ &= \frac{\sigma^2}{a^2} \end{aligned}$$

□

### Definition 6.1 - Convergence in Probability

We say the sequence of random variables  $\{Z_n\}_{n \in \mathbb{N}}$  converges in probability to the random variable  $Z$  if

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

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

*N.B.* The random variables  $\{Z_n\}_{n \in \mathbb{N}}$  &  $Z$  must be in the same probability space.

### Theorem 6.3 - Weak Law of Large Numbers

If  $\{X_n\}_{n \in \mathbb{N}}$  are independent & identically distributed and  $\mathbb{E}(X_1) = \mu < \infty$  then

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

*N.B.* This is an example of Convergence in Probability.

**Definition 6.2 - Convergence in Distribution**

We say the sequence of random variables  $\{Z_n\}_{n \in \mathbb{N}}$  converges in distribution to random variable  $Z$  if

$$\forall z \in \mathbb{R} \text{ where } \mathbb{P}(Z \leq z) \text{ is continuous, } \lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z)$$

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

*N.B.* The random variables  $\{Z_n\}_{n \in \mathbb{N}}$  &  $Z$  need not be in the same probability space.

**Remark 6.2 - Equivalent Statements to Convergence in Distribution**

Saying that  $Z_n \rightarrow_{\mathcal{D}} Z$  is equivalent to saying that

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

**Theorem 6.4 - Central Limit Theorem**

If  $\{X_n\}_{n \in \mathbb{N}}$  are independent & identically distributed,  $\mathbb{E}(X_1) = \mu < \infty$  and  $\text{Var}(X_1) = \sigma^2 < \infty$  then

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

**Theorem 6.5 - Convergence in Probability & Distribution**

Convergence in probability  $\implies$  Convergence in distribution, **but** the opposite is not necessarily true.

**Theorem 6.6 - Convergence in Probability & Distribution to a Constant**

Convergence in distribution to a constant **and** convergence in probability to a constant are equivalent.

**Example 6.1 -**

Let  $X \sim \text{Bernoulli}(\frac{1}{2})$  and  $\{X_n\}_{n \in \mathbb{N}}$  be a sequence of random variables where  $X_i := (1 - X) + \frac{1}{n}$ . We have

$$F_X(x) = \begin{cases} 0 & , x < 0 \\ \frac{1}{2} & , x \in [0, 1) \\ 1 & , x \geq 1 \end{cases} \quad F_{X_n}(x) = \begin{cases} 0 & , x < \frac{1}{n} \\ \frac{1}{2} & , x \in [\frac{1}{n}, 1 + \frac{1}{n}) \\ 1 & , x \geq 1 + \frac{1}{n} \end{cases}$$

Clearly  $F_{X_n}(x) \rightarrow F_X(x)$  at all points at which  $F_X$  is continuous (i.e.  $x \in \mathbb{R} \setminus \{0, 1\}$ ).

Thus  $X_n \rightarrow_{\mathcal{D}} X$ .

**Theorem 6.7 - Continuous Mapping Theorem**

Let  $g : Z \rightarrow G$  be a *continuous* function. 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 6.8 - 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 &  $c \in \mathbb{R} \setminus \{0\}$  be a constant.

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)  $Y_n Z_n \rightarrow_{\mathcal{D}} Yc$ ; and,
- iii)  $\frac{Y_n}{Z_n} \rightarrow_{\mathcal{D}} \frac{Y}{c}$ .

**Definition 6.3 - Convergence in Quadratic Mean**

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

We say that  $\{Z_n\}_{n \in \mathbb{N}}$  Converges in Quadratic Mean to the random variable  $Z$  if

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

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

**Theorem 6.9** - If  $Z_n \rightarrow_{qm} Z$  then  $Z_n \rightarrow_{\mathbb{P}} Z$

**Proof 6.3 - Theorem 5.9**

Fix any  $\varepsilon > 0$ . We have

$$\begin{aligned} \mathbb{P}(|Z_n - Z| > \varepsilon) &= \mathbb{P}(|Z_n - Z|^2 > \varepsilon^2) \\ &\leq \frac{1}{\varepsilon^2} \mathbb{E}[(Z_n - Z)^2] \text{ by Markov's Inequality} \\ &\rightarrow 0 \text{ since } Z_n \rightarrow_{qm} Z. \end{aligned}$$

Hence  $Z_n \rightarrow_{\mathbb{P}} Z$ . □

**1.7 Probabilistic Convergence & Estimators****Definition 7.1 - Consistency of a Sequence of Estimators**

A sequence of estimators,  $\{\hat{\theta}_n(\cdot) : \chi^n \rightarrow \Theta\}$ , are said to be *Consistent* if

$$\forall \theta \in \Theta \text{ with } \mathbf{X}_n \sim f_n(\cdot; \theta), \hat{\theta}_n(\mathbf{X}_n) \rightarrow_{\mathbb{P}(\cdot; \theta)} \theta$$

**Remark 7.1 - Consistency of a Sequence of Estimators**

- i) In numerous situations one will talk about the consistency of *the* estimator, *e.g.* for the MLE, but also for the mean, etc. This implicitly refers to the corresponding sequence of MLEs, sequence of means, etc.
- ii) Note the  $\mathbb{P}(\cdot; \theta)$  in the limit above, and in particular the dependence on  $\theta$ . This is often omitted in practice, you should however not forget what the symbols actually mean.
- iii) Quadratic mean / Mean Square convergence  $\implies$  consistency.  
That is, if the MSE of the estimator converges to 0, the estimator is consistent.

**Example 7.1 - Consistency of Flipping Coins**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta^*)$  for some  $\theta^* \in [0, 1]$ .

The maximum likelihood estimate and method of moments for  $\hat{\theta}_n$  are the sample mean.

$$\hat{\theta}_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i$$

By the *Weak Law of Large Numbers* we have that *consistency* of  $\{\hat{\theta}_n\}$ , since  $\mathbb{E}(X_1) = \theta^*$ .

**Example 7.2 - Crude Confidence Interval when Flipping Coins**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta^*)$  for some  $\theta^* \in [0, 1]$  and define  $\hat{\theta}_n := \hat{\theta}_n(X_1, \dots, X_n)$ .

We shall produce a *confidence interval* for  $\theta^*$ .

$$\mathbb{E}(\hat{\theta}_n; \theta^*) = \theta^* \quad \text{and} \quad \text{Var}(\hat{\theta}_n; \theta^*) = \frac{\theta^*(1 - \theta^*)}{n}$$

$$\begin{aligned}
\mathbb{P}\left(|\hat{\theta}_n - \theta^*| \geq \varepsilon; \theta^*\right) &\leq \frac{\theta^*(1-\theta^*)}{n\varepsilon^2} \quad \text{by Chebyshev's Inequality} \\
\text{We don't know } \theta^*, \text{ but can deduce that } \theta^*(1-\theta^*) &\leq \frac{1}{4} \\
\implies \mathbb{P}\left(|\hat{\theta}_n - \theta^*| \geq \varepsilon; \theta^*\right) &\leq \frac{1}{4n\varepsilon^2} \\
&\text{Define } \alpha := \frac{1}{4n\varepsilon^2} \\
\implies \mathbb{P}\left(|\hat{\theta}_n - \theta^*| \geq \frac{1}{2\sqrt{n\alpha}}; \theta^*\right) &\leq \alpha \\
\implies \mathbb{P}\left(\hat{\theta}_n - \frac{1}{2\sqrt{n\alpha}} < \theta^* < \hat{\theta}_n + \frac{1}{2\sqrt{n\alpha}}; \theta^*\right) &\geq 1 - \alpha
\end{aligned}$$

This means the random interval  $(\hat{\theta}_n - \frac{1}{2\sqrt{n\alpha}}, \hat{\theta}_n + \frac{1}{2\sqrt{n\alpha}}; \theta^*)$  contains  $\theta^*$  with probability  $1 - \alpha$ .

We can note that the interval decreases as  $n$  increases, and increases as  $\alpha$  decreases. *N.B.*  $\hat{\theta}_n$  is a random variable, while  $\theta^*$  is not.

### Example 7.3 - Asymptotically Exact Confidence Interval when Flipping Coins

This is an improvement on the bound produced in **Example 5.3**.

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta^*)$  for some  $\theta^* \in [0, 1]$ ,  $W \sim \text{Normal}(0, 1)$  and define  $\hat{\theta}_n := \hat{\theta}_n(X_1, \dots, X_n)$ .

We shall show that

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta^*)}{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}} \rightarrow_D W$$

We know that  $\text{Var}(X_1) = \theta^*(1 - \theta^*)$ .

By the *Weak Law of Large Numbers*  $\hat{\theta}_n \rightarrow_{\mathbb{P}} \theta^*$ .

By the *Central Limit Theorem*

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta^*)}{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}} \rightarrow_D W$$

$$\text{Define } Y_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta^*)}{\sqrt{\theta^*(1 - \theta^*)}} \text{ and } Z_n = \frac{\sqrt{\theta^*(1 - \theta^*)}}{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}}.$$

By the *Continuous Mapping Theorem* tells us that  $Z_n \rightarrow_D 1$  and  $Z_n \rightarrow_{\mathbb{P}} 1$ .

Hence, by *Slutsky's Theorem*

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta^*)}{\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}} = Y_n Z_n \rightarrow_D W$$

This gives us random interval

$$\left( \hat{\theta}_n - z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}}, \hat{\theta}_n + z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}} \right)$$

This interval captures  $\theta^*$  asymptotically (in  $n$ ) with probability  $1 - \alpha$ .

*N.B.*  $z_{\alpha} = \Phi^{-1}(1 - \alpha)$  where  $\Phi$  is the cumulative density function of a  $\text{Normal}(0, 1)$ .

## 1.8 The Fisher Information

### Remark 8.1 - Motivation

In the next part of the content we shall show that given  $\mathbf{X}_n \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$  then for sufficiently regular models

- i) There exists a lower bound on the achievable performance of any estimate of  $\theta^*$ .
- ii) A scaled & centered sequence of maximum likelihood estimators  $\{\hat{\theta}_n(\mathbf{X}_n)\}$  become asymptotically normal as  $n \rightarrow \infty$ .

**Remark 8.2 - Measuring Performance of Estimator**

We measure the performance of an estimator  $\hat{\theta}$  in terms of variance, since its mean should be  $\theta^*$ . Lower variance indicates better performance.

**Definition 8.1 - The Score Function**

Let  $\ell(\theta; x) := \ln f(x; \theta)$ .

The *Score Function* is a measure of the sensitivity of the likelihood function wrt  $\theta$

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

**Remark 8.3 -  $\theta^*$  is a turning point of  $\ell(\theta; x)$** 

Note that under the *Fisher Information Regularity Conditions* we have that  $\forall \theta \in \Theta$

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

This shows that we expect the derivative to equal 0 at  $\theta^*$ . Further, this means  $\theta^*$  is a turning point of the log-likelihood function (hopefully a maximum).

**Example 8.1 - Application of Remark 6.3**

Let  $X \sim \text{Poisson}(\theta)$ . Then  $f_X(x; \theta) = \frac{\theta^x}{x!} e^{-\theta} \mathbf{1}\{x \in \mathbb{N}\}$ .

$$\begin{aligned} \implies \ell(\theta; x) &= -\theta + x \ln \theta - \ln x! \\ \implies \ell'(\theta; x) &= -1 + \frac{x}{\theta} \\ \implies \mathbb{E}(\ell'(\theta; X); \theta) &= -1 + \frac{\theta}{\theta} \\ &= 0 \end{aligned}$$

**Definition 8.2 - Fisher Information Regularity Conditions**

Let  $\Theta$  be an open interval in  $\mathbb{R}$  and  $f(x; \theta)$  be a pmf/pdf.

Below are conditions which a model is required to meet in order to be considered sufficiently regular such that *Fisher Information* can be drawn from it.

- i) 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 for any  $x \in \mathcal{X}$ .
- ii)  $\forall \theta \in \Theta$  the set  $S := \{x \in \mathcal{X} : f(x; \theta) > 0\}$  does not depend on  $\theta \in \Theta$ .
- iii) The identity below exists

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

**Definition 8.3 - Fisher Information**

*Fisher Information* is a technique for measuring the amount of information that an observable random variable  $X$  carries about an unknown parameter  $\theta$  upon which the probability of  $X$  depends.

Let  $X \sim f(\cdots; \theta)$ . Then the *Fisher Information* for any  $\theta \in \Theta$  is

$$I(\theta) := \mathbb{E}(\ell'(\theta; X)^2; \theta) \geq 0$$

*N.B.* This is the *Expectation of the score, squared*  $\equiv$  *Second moment of the score*.

**Remark 8.4 - Fisher Information**

- i) *Fisher Information* is a function of the parameter,  $\theta$ , not the data,  $X$ .
- ii)  $I(\theta)$  can be thought of as being the average *information* brought by a single observation  $X$  about  $\theta$ , assuming  $X \sim f(\cdot; \theta)$ .
- iii) Since  $\forall \theta \in \Theta$ ,  $\mathbb{E}(\ell'(\theta; X); \theta) = 0$  then

$$I(\theta) = \text{Var}(\ell'(\theta; X); \theta)$$

The variance of the score.

**Example 8.2 - Fisher Information of Poisson**

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

From **Example 6.1** we know that  $\ell'(\theta; x) = -1 + \frac{x}{\theta}$ . Then

$$\begin{aligned} I(\theta) &= \text{Var}(\ell'(\theta; X); \theta) \\ &= \text{Var}\left(-1 + \frac{X}{\theta}; \theta\right) \\ &= \text{Var}\left(\frac{X}{\theta}; \theta\right) \\ &= \frac{1}{\theta^2} \text{Var}(X; \theta) \\ &= \frac{1}{\theta^2} \cdot \theta \text{ since } X \sim \text{Poisson}(\theta) \\ &= \frac{1}{\theta} \end{aligned}$$

**Theorem 8.1 - Alternative Expression of Fisher Information**

Let  $f(x; \theta)$  be a pmf/pdf which satisfies the conditions of **Definition 6.2**. If

$$\forall \theta \in \Theta \quad \int_{\mathcal{X}} \frac{d^2}{d\theta^2} f(x; \theta) dx = \frac{d}{d\theta} \int_{\mathcal{X}} \frac{d}{d\theta} f(x; \theta) dx$$

Then

$$I(\theta) = -\mathbb{E}\left(\frac{d^2}{d\theta^2} \ell(\theta; X); \theta\right)$$

N.B.  $\frac{d}{d\theta} \int_{\mathcal{X}} \frac{d}{d\theta} f(x; \theta) dx = 0$  by the regularity conditions.

**Proof 8.1 - Theorem 6.1**

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

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

□

## 1.9 Efficiency and The Cramer-Rao Bound

### Definition 9.1 - IID Score Function

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$  for some  $\theta \in \Theta$ . Then the *Score Function* is

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

N.B.  $\frac{d}{d\theta} \ell_n(\theta; \mathbf{x}) = \frac{d}{d\theta} \sum \ell(\theta; x_i) = \sum \ell'(\theta; x_i)$ .

### Definition 9.2 - IID Fisher Information

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$  for some  $\theta \in \Theta$ . Then the *Fisher Information* is

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

### Theorem 9.1 - Relationship between IID Fisher Information & Fisher Information

Consider the situation where  $\forall \theta \in \Theta$ ,  $f_n(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$ . Then

$$\forall \theta \in \Theta, I_n(\theta) = nI(\theta)$$

### Proof 9.1 - Theorem 7.1

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$ . Then

$$\begin{aligned} I_n(\theta) &= \text{Var}(\ell'_n(\theta; \mathbf{X}); \theta) \\ &= \text{Var}\left(\sum_{i=1}^n \ell'(\theta; X_i); \theta\right) \\ &= n\text{Var}\left(\sum_{i=1}^n \ell'(\theta; X_1); \theta\right) \\ \implies I_n(\theta) &= nI(\theta) \end{aligned}$$

□

### Theorem 9.2 - Cauchy-Schwarz Inequality for Expectation

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

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

### Proof 9.2 - Theorem 7.2

If  $\mathbb{E}(Y^2) = 0$  then  $\mathbb{P}(Y = 0) = 1$  so  $\mathbb{E}(XY) = 0$  and the statement holds.

Thus, assume  $\mathbb{E}(Y^2) > 0$  and define  $\lambda := \frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}$ . Then

$$\begin{aligned} 0 &\leq \mathbb{E}(X - \lambda Y)^2 \\ &= \mathbb{E}(X^2) - 2\lambda\mathbb{E}(XY) + \lambda^2\mathbb{E}(Y^2) \\ &= \mathbb{E}(X^2) - 2\frac{\mathbb{E}(XY)^2}{\mathbb{E}(Y^2)} + \frac{\mathbb{E}(XY)^2}{\mathbb{E}(Y^2)} \\ &= \mathbb{E}(X^2) - \frac{\mathbb{E}(XY)^2}{\mathbb{E}(Y^2)} \\ \implies \mathbb{E}(XY)^2 &\leq \mathbb{E}(X^2)\mathbb{E}(Y^2) \end{aligned}$$

□

### Theorem 9.3 - Covariance Inequality

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

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

**Proof 9.3 - Theorem 7.3**

Let  $W = X - \mathbb{E}(X)$  and  $Z = Y - \mathbb{E}(Y)$  giving  $\mathbb{E}(WZ) = \text{Cov}(X, Y)$ ,  $\mathbb{E}(W^2) = \text{Var}(X)$  and  $\mathbb{E}(Z^2) = \text{Var}(Y)$ .

By applying the *Cauchy-Schwarz inequality* we get

$$\text{Cov}(X, Y)^2 = \mathbb{E}(WZ)^2 \leq \mathbb{E}(W^2)\mathbb{E}(Z^2) = \text{Var}(X)\text{Var}(Y) \iff \text{Cov}(X, Y)^2 \leq \text{Var}(X)\text{Var}(Y)$$

**Remark 9.1 - Correlation value**

The result in **Theorem 7.3** is the reason why correlation is valued in  $[-1, 1]$ .

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

**Theorem 9.4 - Cramer-Rao Inequality - Scalar Parameter**

Let  $\mathbf{X}_n \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$  and assume the *Fisher Information Regularity Conditions* hold.

Let  $\hat{\theta}_n(\cdot)$  be an estimator of  $\theta$  with expectation  $m(\theta) := \mathbb{E}(\hat{\theta}_n(\mathbf{X}_n); \theta)$  which satisfies

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

Then

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

**Proof 9.4 - Theorem 7.4**

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

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

**Proposition 9.1 - Useful result from Cramer-Rao Inequality**

If  $\hat{\theta}_n(\mathbf{X}_n)$  is an unbiased estimator (i.e.  $m(\theta) = \theta$ ) then

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



This shows there is a lower bound on the possible performance of an estimator.

**Definition 9.3 - Efficient Estimator**

An *Estimator* is said to be *Efficient* when its variance is equal to the *Cramer-Rao lower bound*  $\forall \theta^*$ .

**Example 9.1 - Efficient Coin Flipping**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$  with  $\theta \in [0, 1]$ , this corresponds to flipping a coin  $n$  times and considering each flip the random variable  $X : \{H, T\} \rightarrow \{0, 1\}$  such that  $X(H) = 1$  and  $X(T) = 0$  with probability distribution such that  $\mathbb{P}(X = 1; \theta) = \theta$  and  $\mathbb{P}(X = 0; \theta) = 1 - \theta$ . We consider the intuitive estimator of  $\theta$

$$\hat{\theta}_n := \hat{\theta}_n(\mathbf{X}_n) := \frac{1}{n} \sum_{i=1}^n X_i$$

The estimator is unbiased  $\forall n \in \mathbb{N}$  and its variance is

$$\text{Var}(\hat{\theta}_n; \theta) = \frac{\text{Var}(X_1; \theta)}{n} = \frac{\mathbb{E}(X_1^2; \theta) - \mathbb{E}(X_1; \theta)^2}{n} = \frac{\theta - \theta^2}{n} = \frac{\theta(1 - \theta)}{n}$$

Now we consider the *Cramer-Rao bound*

$$\begin{aligned} \text{We find } L(\theta; x) &= \theta^x (1 - \theta)^{1-x} \\ \implies \ell(\theta; x) &= x \ln \theta + (1 - x) \ln(1 - \theta) \\ \implies \ell'(\theta; x) &= \frac{x}{\theta} - \frac{1-x}{1-\theta} \\ \implies \ell''(\theta; x) &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \end{aligned}$$

Thus we can use  $I(\theta) = -\mathbb{E}(\ell''(\theta; X); \theta)$

$$\begin{aligned} \implies I(\theta) &= -\mathbb{E}\left(-\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2}; \theta\right) \\ &= \mathbb{E}\left(\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2}; \theta\right) \\ &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\ &= \frac{1}{\theta} + \frac{1}{1-\theta} \\ &= \frac{1}{\theta(1-\theta)} \\ I_n(\theta) &= nI(\theta) \text{ Since } X_1, X_2, \dots \text{ are iid} \end{aligned}$$

The *Cramer-Rao bound* for the variance is

$$\frac{1}{nI(\theta)} = \frac{\theta(1 - \theta)}{n}$$

Thus our estimator is efficient.

## 1.10 Asymptotic Distribution of the Maximum Likelihood Estimator

**Theorem 10.1 -**

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)  $\exists C(\cdot) : \mathcal{X} \rightarrow [0, \infty)$  such that  $\mathbb{E}(C(X_1); \theta^*) < \infty$ ,  $\Xi \subset \Theta$  an open set containing  $\theta^*$  and  $\Delta(\cdot) : \Xi \rightarrow [0, \infty)$  continuous at 0 st  $\Delta(0) = 0$ , st  $\forall \theta, \theta', x \in \Xi \times \mathcal{X}$ .

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

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

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

**Theorem 10.2 -**

Under the conditions of **Theorem 8.1**, with  $\hat{\theta}_n := \hat{\theta}_n(\mathbf{X})$  the 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 10.1 - Theorem 8.1**

By **Theorem 8.2**  $\ell'_n(\hat{\theta}_n; \mathbf{X}) = \ell'_n(\theta^*; \mathbf{X}) + (\hat{\theta}_n - \theta^*)\{\ell''_n(\theta^*; \mathbf{X}) + R_n\}$  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  $\text{Var}(\ell'(\theta^*; X_i); \theta^*) = I(\theta^*)$

$\Rightarrow U_n \rightarrow_{\mathcal{D}(\cdot; \theta^*)} U \sim \text{Normal}(0, I(\theta^*))$  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

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

**Proof 10.2 - Theorem 8.2**

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

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

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

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

from the smoothness condition on  $\ell''$ .

From the *Weak Law of Large Numbers*

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

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

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

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

**Definition 10.1 - Asyptically Efficient**

A sequence of estimators  $\{\hat{\theta}_n(\mathbf{X})\}$  is *Asymptotically Efficient* if either its mean-squared error converges to the *Cramer-Rao Lower Bound*

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

or  $\hat{\theta}_n$  is *Asumptotically Normally Distributed* in the sense of **Theorem 8.1**

$$\forall \theta \in \Theta, \sqrt{nI(\theta)}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}(\cdot; \theta)} Z$$

*N.B.* The variance of  $\frac{Z}{\sqrt{(nI(\theta^*))}}$  is exactly  $\frac{1}{nI(\theta)}$ .

**Theorem 10.3 -**

Under the conditions of **Theorem 8.1** the maximum likelihood estimator is *asymptotically efficient*.

**Definition 10.2 - Regular Statistical Model**

Any *Statistical Model* which satisfies the condition of **Theorem 8.1** is a *Regular Statistical Model*.

**Remark 10.1 - Why use MLE over others**

Due to the *Asymptotic Efficiency* of maximum likelihood estimators it is beter to use them in *Regular Statistical Models*.

## 1.11 Confidence Sets Around the Maximum Likelihood Estimator

**Definition 11.1 - Coverage of an Interval**

Let  $\mathbf{X} \sim f_n(\cdot; \theta)$ ,  $\theta \in \Theta = \mathbb{R}$ ,  $L(\cdot) : \mathcal{X}^n \rightarrow \Theta$  and  $U(\cdot) : \mathcal{X}^n \rightarrow \Theta$  where  $\forall \mathbf{x} \in \mathcal{X}^n$ ,  $L(\mathbf{x}) < U(\mathbf{x})$ . Then,  $\forall \theta \in \Theta$  the coverage  $C_{\mathcal{I}}(\theta)$  of the random interval  $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$  at  $\theta$  is

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

**Remark 11.1 - Coverage of an Interval in Words**

$C_{\mathcal{I}}(\theta)$  is the probability that the deterministic quantity  $\theta$  falls into the random interval  $\mathcal{I}(\mathbf{X})$  under the probability distribution  $\mathbb{P}(\cdot; \theta)$  where  $\mathbf{X} \sim f_n(\cdot; \theta)$ .

**Remark 11.2 - Multi-Dimensional Coverage**

We can extend *Coverage of an Interval* to the multi-dimensional case by considering confidence sets and then considering the probability  $\mathbb{P}(\theta \in \mathcal{I}(\mathbf{X}); \theta)$ .

**Definition 11.2 - Confidence Interval**

$\forall \alpha \in [0, 1]$  we say that an interval  $\mathcal{I}(\mathbf{X}) := [L(\mathbf{X}), U(\mathbf{X})]$  is a  $1 - \alpha$  confidence interval if  $\forall \theta \in \Theta$  its coverage is at least  $1 - \alpha$  or more formally  $\inf_{\theta \in \Theta} C_{\mathcal{I}}(\theta) \geq 1 - \alpha$ .

**Remark 11.3 - Exact Confidence Interval**

If  $C_{\mathcal{I}}(\theta) = 1 - \alpha \forall \theta \in \Theta$  then  $\mathcal{I}$  is an exact  $1 - \alpha$  confidence interval.

**Definition 11.3 - Observed Confidence Interval**

For an interval  $\mathcal{I}(\cdot) = [L(\cdot), U(\cdot)]$  with  $L : \mathcal{X}^n \rightarrow \Theta$  and  $U : \mathcal{X}^n \rightarrow \Theta$ , and a realisation  $\mathbf{x}$ , the corresponding *Observed Confidence Interval* is  $\mathcal{I}(\mathbf{x})$ .

*N.B.* Nothing interesting can be said about the probability that  $\theta \in \mathcal{I}(\mathbf{x})$  since  $\theta$  and  $\mathcal{I}(\mathbf{x})$  are deterministic.

**Notation 11.1 - Quantile of Normal(0, 1)**

For any  $\beta \in (0, 1)$  let  $z_{\beta} \in \mathbb{R}$  be such that for  $Z \sim \text{Normal}(0, 1)$ ,  $1 - \Phi(z_{\beta}) = \mathbb{P}(Z > z_{\beta}) = \beta$ .

**Example 11.1 - Confidence interval for the mean of a Normal Distribution**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$  for  $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^{\geq 0}$  and where  $\sigma^2$  is known.

Consider the estimator  $\hat{\mu}_n = \hat{\mu}_n(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i$  of  $\mu$ . Then we know that the following non-asymptotic result holds.

We have  $\frac{1}{n} \sum_{i=1}^n X_i \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$ . Thus

$$\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\sigma^2/n}} \sim \text{Normal}(0, 1)$$

Then

$$\begin{aligned} \forall \alpha \in (0, 1) \quad , \quad & \mathbb{P} \left( z_{1-\alpha/2} \leq \frac{\hat{\mu}_n(\mathbf{X}) - \mu}{\sqrt{\sigma^2/n}} \leq z_{\alpha/2}; \mu \right) \\ &= \mathbb{P} \left( \frac{\hat{\mu}_n(\mathbf{X}) - \mu}{\sqrt{\sigma^2/n}} \leq z_{\alpha/2} \right) - \mathbb{P} \left( \frac{\hat{\mu}_n(\mathbf{X}) - \mu}{\sqrt{\sigma^2/n}} \leq z_{1-\alpha/2} \right) \\ &= \left( 1 - \frac{\alpha}{2} \right) - \left( 1 - \left( 1 - \frac{\alpha}{2} \right) \right) \\ &= 1 - \alpha \end{aligned}$$

By symmetry we notice that  $z_{1-\frac{\alpha}{2}} = -z_{\alpha/2}$ .

By rearranging we have the equivalence of events

$$\left\{ -z_{\alpha/2} \leq \frac{\hat{\mu}_n(\mathbf{X}) - \mu}{\sqrt{\sigma^2/n}} \leq z_{\alpha/2} \right\} = \left\{ \hat{\mu}_n(\mathbf{X}) - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu}_n(\mathbf{X}) + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

To rearrange we separate into two events & treat them separately

$$\begin{aligned} \left\{ \frac{\hat{\mu}_n(\mathbf{X}) - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right\} &= \left\{ \frac{\hat{\mu}_n(\mathbf{X})}{\sigma/\sqrt{n}} - z_{\alpha/2} \leq \frac{\mu}{\sigma/\sqrt{n}} \right\} \\ &= \left\{ \mu \geq \hat{\mu}_n(\mathbf{X}) - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\} \end{aligned}$$

Similarly

$$\begin{aligned} \left\{ -z_{\alpha/2} \leq \frac{\hat{\mu}_n(X) - \mu}{\sqrt{\sigma^2/n}} \right\} &= \left\{ \frac{\mu}{\sigma/\sqrt{n}} \leq \frac{\hat{\mu}_n(X)}{\sigma/\sqrt{n}} + z_{\alpha/2} \right\} \\ &= \left\{ \mu \leq \hat{\mu}_n(X) + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} \end{aligned}$$

So the interval  $\mathcal{I}(X) = [L(X), U(X)]$  where  $L(\mathbf{x}) = \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  and  $U(\mathbf{x}) = \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  is an  $1 - \alpha$  exact confidence interval.

**Remark 11.4 - Confidence Intervals with unknown  $\sigma^2$**

When  $\sigma^2$  is unknown we can define  $\{\hat{\sigma}_n^2\}_{n \in \mathbb{N}}$  to be a consistent sequence of estimators of  $\sigma^2$  (e.g. the sample variance)

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

## 1.12 Asymptotic Approximation of Confidence Intervals

**Theorem 12.1 -**

Assume  $\mathbf{X} \sim f(\cdot; \theta^*)$ . Let  $\{\hat{\theta}_n\}_{n \in \mathbb{N}}$  be a consistent sequence of estimators of  $\theta^*$  and assume 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  $L_n(\mathbf{x}) := \hat{\theta}_n(\mathbf{x}) - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  and  $U_n(\mathbf{x}) := \hat{\theta}_n(\mathbf{x}) + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ .

**Proof 12.1 - Theorem 10.1**

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

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

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

**Remark 12.1 - Theorem 10.1**

The confidence interval is only asymptotically exact. For finite  $n$ , the overage of the confidence interval will be different from  $1 - \alpha$  but the difference will converge to 0 as  $n$  increases. In practice  $\sigma^2$  may be unknown, in these cases substitute for a consistent sequence of estimators of  $\sigma^2$ .

**Theorem 12.2 -**

Assum  $\mathbf{X} \sim f(\cdot; \theta^*)$  let  $\{\hat{\theta}_n\}_{n \in \mathbb{N}}$  be a consistent sequence of estimators of  $\theta^*$  and assume that  $\{\hat{\theta}_n\}$  is asymptotically normal in the sense that

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

Assume also that  $\{\hat{\sigma}_n^2\}_{n \in \mathbb{N}}$  is a consistent sequence of estimators of  $\sigma^2$ . 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  $L_n(\mathbf{x}) := \hat{\theta}_n(\mathbf{x}) - z_{\alpha/2} \sqrt{\hat{\sigma}_n^2(\mathbf{x})/n}$  and  $U_n(\mathbf{x}) := \hat{\theta}_n(\mathbf{x}) + z_{\alpha/2} \sqrt{\hat{\sigma}_n^2(\mathbf{x})/n}$ .

**Proof 12.2 - Theorem 10.2**

Define  $W_n := \frac{\hat{\theta}_n - \theta^*}{\sqrt{\hat{\sigma}_n^2(X)/n}} = \frac{\hat{\theta}_n(X) - \theta^*}{\sqrt{\hat{\sigma}_n^2(X)/n}} - \sqrt{\frac{\sigma^2}{\hat{\sigma}_n^2(X)}}$ .

By consistency of  $\{\hat{\sigma}_n^2\}_{n \in \mathbb{N}}$  and the *Continuous Mapping Theorem*

$$\sqrt{\frac{\sigma^2}{\hat{\sigma}_n^2(X)}} \xrightarrow{\mathbb{P}(\cdot; \theta^*)} 1$$

By *Slutsky's Theorem*

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

The rest of the proof is the same as for **Theorem 10.1**.

**Remark 12.2 - Theorem 10.2**

For a given  $n$  the quality of the normal approximation will be affected by this additional approximation. One may find that for less accurate estimators of  $\sigma^2$ , the  $n$  required for the confidence interval to have almost the right coverage will be higher.

### 1.13 Estimating the Information for Maximum Likelihood Estimates

**Remark 13.1 - Applying Theorem 10.2 to sequences of MLEs for regular statistical models**

When dealing with *Maximum Likelihood Estimators* for regular statistical models we have that  $\sigma^2 = 1/I(\theta^*)$  thus

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

However the *Fisher Information* is unknown so we consider two cases

- i) When the expectation,  $I(\theta^*) = -\mathbb{E}(\ell''(\theta^*; X_1); \theta^*)$ , can be calculated. In this case we replace  $\theta^*$  with  $\hat{\theta}_n$  in the equation.
- ii) When the expectation **cannot** be calculated we invoke the *Weak Law of Large Numbers* and consider the sequence of estimators,  $J_n(\hat{\theta}_n) := -\frac{1}{n} \sum_{i=1}^n \ell''(\hat{\theta}_n; X_i)$ .

**Theorem 13.1 - Case i)**

Assume  $\{\hat{\theta}_n\}$  is a sequence of *Maximum Likelihood Estimators* st  $\hat{\theta}_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} \theta^*$  and  $I$  is a continuous function of  $\theta$ . Then  $I(\hat{\theta}_n) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} I(\theta^*)$ .

*N.B.* The proof of this follows directly from the *Continuous Mapping Function*.

**Remark 13.2 - Theorem 11.1**

It is only necessary for  $I$  to be continuous in the neighbourhood of  $\theta^*$ . This is due to an extension of the *Continuous Mapping Theorem* that states

$$\begin{aligned} \text{If } X_n \rightarrow_{\mathbb{P}} X \text{ and } g \text{ is a function with discontinuity set } D \text{ then} \\ \mathbb{P}(X \in D) = 0 \implies (X_n) \rightarrow_{\mathbb{P}} g(X). \end{aligned}$$

**Theorem 13.2 - Case ii)**

Assume that  $\{\hat{\theta}_n\}$  is a sequence of *Maximum Likelihood Estimators* st

- i)  $\hat{\theta}_n \rightarrow_{\mathbb{P}(\cdot; \theta^*)} \theta^*$ ;
- ii)  $I(\theta) = -\mathbb{E}(\ell''(\theta; X); \theta) \forall \theta \in \Theta$

- iii)  $\exists C : \mathcal{X} \rightarrow [0, \infty)$  st  $\mathbb{E}(C(X_1); \theta^*) < \infty$ ,  $\Xi \subset \Theta$  is an open set containing  $\theta^*$  and  $\Delta(\cdot) : \Xi \rightarrow [0, \infty)$  is continuous at 0 st  $\Delta(0) = 0$ , and st  $\forall \theta, \theta^*, x \in \Xi^2 \times \mathcal{X} \quad |\ell''(\theta; x) - \ell''(\theta'; x)| \leq C(x)\Delta(\theta - \theta')$

Then

$$J_n(\hat{\theta}_n) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} I(\theta^*)$$

**Proof 13.1 - Theorem 11.2**

Consider the following decomposition

$$\begin{aligned} J_n(\hat{\theta}) - I(\theta^*) &= -\frac{1}{n} \sum_{i=1}^n \ell''(\hat{\theta}_n; X_i) - I(\theta^*) \\ &= T_1 + T_2 \\ \text{Where } T_1 &= -\frac{1}{n} \sum_{i=1}^n \ell''(\hat{\theta}_n; X_i) + \frac{1}{n} \sum_{i=1}^n \ell''(\theta^*; X_i) \\ \text{and } T_2 &= -\left\{ \frac{1}{n} \sum_{i=1}^n \ell''(\theta^*; X_i) \right\} - I(\theta^*) \end{aligned}$$

Now the first term can be upper bounded as follows (for sufficiently large  $n$ , with arbitrary large probability the second inequality holds)

$$\begin{aligned} |T_1| &= \left| -\frac{1}{n} \sum_{i=1}^n \ell''(\hat{\theta}_n; X_i) + \frac{1}{n} \sum_{i=1}^n \ell''(\theta^*; X_i) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \left| \ell''(\hat{\theta}_n; X_i) - \ell''(\theta^*; X_i) \right| \\ &\leq \Delta(\theta_n - \theta^*) \frac{1}{n} \sum_{i=1}^n C(X_i) \end{aligned}$$

By the *Weak Law of Large Numbers*

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

by the assumed consistency of  $\{\hat{\theta}_n\}_{n \in \mathbb{N}}$  and continuity of  $\Delta$  we have that

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

Consequently  $T_1 \xrightarrow[n \rightarrow \infty]{\mathbb{P}(\cdot; \theta^*)} 0$ .

By the *Weak Law of Large Numbers* we have

$$\begin{aligned} &-\frac{1}{n} \sum_{i=1}^n \ell''(\theta^*; X_i) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} I(\theta^*) \\ \implies T_2 &= -\frac{1}{n} \sum_{i=1}^n \ell''(\theta^*; X_i) - I(\theta^*) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} 0 \end{aligned}$$

Since  $T_1 \xrightarrow[n \rightarrow \infty]{\mathbb{P}(\cdot; \theta^*)} 0$  and  $T_2 \xrightarrow[n \rightarrow \infty]{\mathbb{P}(\cdot; \theta^*)} 0$  we deduce from the earlier decomposition that

$$J_n(\hat{\theta}_n) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} I(\theta^*)$$

□

**Remark 13.3 - Summary**

Whenever **Theorem 8.1** holds for a sequence of *Maximum Likelihood Estimators*

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

we can replace  $I(\theta^*)$  with one of two options

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

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

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

- (a) The assumptions of **Theorem 11.2** hold; and,
- (b) 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  $\theta^*$ .

### Example 13.1 - Coin Flipping

Here the new results for this chapter are applied in order to simplify methods used in previous examples when finding confidence intervals & upper bounds on  $\theta^*$ .

The sequence of estimators  $\hat{\theta}_n := \frac{1}{n} \sum_{i=1}^n X_i$  is consistent by the *Weak Law of Large Numbers* and the conditions for asymptotic normality hold  $\forall \theta \in \Theta$ . Hence

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

We can compute the *Fisher Information*  $\forall \theta \in \Theta$ . We have

$$\begin{aligned} \ell'(\theta(x)) &= \frac{x}{\theta} - \frac{1-x}{1-\theta} \\ \text{and } \ell''(\theta; x) &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \\ \implies I(\theta) &= \frac{1}{\theta} + \frac{1}{1-\theta} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

In practice  $\theta^*$  is unknown so we replace  $I(\theta^*)$  with  $I(\hat{\theta}_n)$  to give the asymptotically exact confidence interval,  $[L(\mathbf{X}), U(\mathbf{X})]$  where

$$L(\mathbf{X}) = \hat{\theta}_n - z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \text{ and } U(\mathbf{X}) = \hat{\theta}_n + z_{\alpha/2} \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$$

If we did not know how to compute  $I(\theta)$  we could instead compute

$$\begin{aligned} J_n(\hat{\theta}_n) &= -\frac{1}{n} \sum_{i=1}^n \ell''(\hat{\theta}_n; X_i) \\ &= -\frac{1}{n} \sum_{i=1}^n \left\{ -\frac{X_i}{\hat{\theta}_n^2} - \frac{1-X_i}{(1-\hat{\theta}_n)^2} \right\} \\ &= \frac{1}{\hat{\theta}_n^2} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) + \frac{1}{(1-\hat{\theta}_n)^2} \left( 1 - \frac{1}{n} \sum_{i=1}^n X_i \right) \\ &= \frac{\hat{\theta}_n}{\hat{\theta}_n^2} + \frac{1-\hat{\theta}_n}{(1-\hat{\theta}_n)^2} \\ &= \frac{1}{\hat{\theta}_n(1-\hat{\theta}_n)} \end{aligned}$$

In this case  $J_n(\hat{\theta}_n) = I(\hat{\theta}_n)$ , this is not always true.

### Definition 13.1 - Observed Fisher Information

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

The *Observed Fisher Information* at  $\theta$  is

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



*N.B.*  $\mathbb{E}(J_n(\theta^*); \theta^*) = I(\theta^*)$  and that it differs from the *Fisher Information* (under the *Fisher Information Regularity Conditions* by not being an expectation.

### 1.14 Transformations and Confidence Intervals

#### Definition 14.1 - Wald Approach

The confidence intervals seen so far fit the *Wald Approach*.

If  $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$  where  $\theta^* \in \Theta \subset \mathbb{R}$  then one can define a confidence interval for  $\theta^*$  using the asymptotic distribution of the *Maximum Likelihood Estimator*

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

which ensures that as  $n \rightarrow \infty$ ,  $\mathbb{P}(\theta^* \in [L(\mathbf{X}), U(\mathbf{X})]) \rightarrow 1 - \alpha$ .

#### Proposition 14.1 - Transformed Confidence Interval - Increasing

Let  $\tau := g(\theta)$  be a bijective, continuously differentiable & increasing function.

This gives a direct transformation of  $[L(\mathbf{x}), U(\mathbf{x})]$  to  $[g(L(\mathbf{x})), g(U(\mathbf{x}))]$ .

$$\text{i.e. } \{\mathbf{x} \in \mathcal{X}^n : L(\mathbf{x}) \leq \theta^* \leq U(\mathbf{x})\} = \{\mathbf{x} \in \mathcal{X}^n : g(L(\mathbf{x})) \leq \tau^* \leq g(U(\mathbf{x}))\}$$

Consequently

$$\begin{aligned} \mathbb{P}(\theta^* \in [L(\mathbf{X}), U(\mathbf{X})]; \theta^*) &= \mathbb{P}(\tau^* \in [g(L(\mathbf{X})), g(U(\mathbf{X}))]) \\ &\rightarrow 1 - \alpha \text{ as } n \rightarrow \infty \end{aligned}$$

*i.e.*  $[g(L(\mathbf{X})), g(U(\mathbf{X}))]$  is an asymptotically exact  $1 - \alpha$  for  $\tau^*$ .

#### Proposition 14.2 - Transformed Confidence Interval - Decreasing

Let  $\tau := g(\theta)$  be a bijective, continuously differentiable & decreasing function.

This gives a direct transformation of  $[L(\mathbf{X}), U(\mathbf{X})]$  to  $[g(U(\mathbf{X})), g(L(\mathbf{X}))]$  which is an asymptotically exact  $1 - \alpha$  confidence interval for  $\tau^*$ .

#### Remark 14.1 - Deriving Reparameterised Confidence Intervals

We can obtain a reparameterised *Confidence Interval* by working with the reparameterised likelihood,  $\tilde{f}(\mathbf{x}; \tau) := f(\mathbf{x}; g^{-1}(\tau))$ . Now we can find  $\tilde{L}(\mathbf{x})$  and  $\tilde{U}(\mathbf{x})$  directly.

#### Theorem 14.1 -

Assume  $X \in f(\cdot; \theta)$  for  $\theta \in \Theta \subseteq \mathbb{R}$  and let  $\tau := g(\theta)$  where  $g$  is bijective & continuously differentiable.

The *Fisher Information* for the parameterisation  $\tilde{f}(x; \tau) := f(x; g^{-1}(\tau))$  is

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

#### Proof 14.1 - Theorem 12.1

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

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

**Remark 14.2 -**

As a consequence, for regular statistical models

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

is equivalent to

$$\sqrt{\frac{nI(\theta^*)}{g'(\theta^*)^2}}(\hat{\tau}_n - \tau^*) \rightarrow_{\mathcal{D}(\cdot; \theta^*)} Z \sim \text{Normal}(0, 1)$$

which leads to

$$\begin{aligned}
 \tilde{L}(\mathbf{x}) &= \hat{\tau}_n - z_{\alpha/2} \sqrt{\frac{g'(\theta^*)^2}{nI(\theta^*)}} \\
 \tilde{U}(\mathbf{x}) &= \hat{\tau}_n + z_{\alpha/2} \sqrt{\frac{g'(\theta^*)^2}{nI(\theta^*)}}
 \end{aligned}$$

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

**Example 14.1 -**

Consider  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, 1)$ .

We know that the *Maximum Likelihood Estimator* of  $\mu$  is  $\bar{X} \sim \text{Normal}(\mu, \frac{1}{n})$ .

A  $1 - \alpha$  *Confidence Interval* for  $\mu$  is

$$\left[ \bar{X} - \frac{z_{\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}}{\sqrt{n}} \right]$$

Consider the parameterisation  $\tau = \frac{1}{\mu}$ . This corresponds to  $g(x) = \frac{1}{x}$  which is bijective & continuously differentiable except at 0, and is decreasing.

Hence, a  $1 - \alpha$  exact *Confidence Interval* for  $\tau$  is

$$\left[ \frac{1}{\bar{X} + z_{\alpha/2}/\sqrt{n}}, \frac{1}{\bar{X} - z_{\alpha/2}/\sqrt{n}} \right]$$

Consider the two ways to find an asymptotically  $1 - \alpha$  *Exact Confidence Interval* for  $\tau$ . After direct calculations we find that

$$\tilde{\ell}''(\tau; x) = -\frac{3}{\tau^4} + \frac{2x}{\tau^3}$$

So

$$\tilde{I}(\tau) := i\mathbb{E}(\tilde{\ell}''(\tau; X); \tau) = \frac{3}{\tau^3} - \frac{2}{\tau^4} = \frac{1}{\tau^4}$$

Noting that the *Maximum Likelihood Estimator* for  $\tau$  is  $1/\bar{X}$  we find that

$$\sqrt{\frac{n}{\tau^4}} \left( \frac{1}{\bar{X}} - \tau \right) \rightarrow_{\mathcal{D}(\cdot; \tau)} Z \sim \text{Normal}(0, 1)$$

so an asymptotically exact  $1 - \alpha$  *Confidence Interval* is

$$\left[ \frac{1}{\bar{X}} - z_{\alpha/2} \frac{\tau^2}{\sqrt{n}}, \frac{1}{\bar{X}} + z_{\alpha/2} \frac{\tau^2}{\sqrt{n}} \right]$$

Alternatively, instead of working out  $\tilde{I}(\tau)$  as above, we could use **Theorem 12.1** to find that

$$\tilde{I}(\tau) = \frac{I(\theta)}{g'(\theta)^2}, \quad \theta = g^{-1}(\tau) = \frac{1}{\tau}$$

Since  $I(\theta) = 1$  and  $g(\theta) = 1/\theta \implies g'(\theta) = -1/\theta^2 = -\tau^2$ , we have

$$\tilde{I}(\tau) = \frac{1}{(-1/\theta^2)^2} = \frac{1}{(-\tau^2)^2} = \frac{1}{\tau^4}$$

**Remark 14.3 - Example 12.1**

- i) The transformed *Confidence Interval* is exact, which the second *Confidence Interval* is not since  $\sqrt{n/\tau^4} \left( \frac{1}{\bar{X}} - \tau \right)$  is not exactly normally distributed, but only asymptotically so.
- ii) The transformed *Confidence Interval* is not generally centred at  $\hat{\tau}$ .
- iii) This serves as an example that convergence in distribution says nothing about convergence of moments. In particular, you can derive that  $\frac{1}{\bar{X}}$  does not have a mean for any  $\mu \in \mathbb{R}$ .

## 1.15 Likelihood Ratio Confidence Sets - Wilk's Approach

**Remark 15.1 - Motivation**

Consider a *Wald Confidence Interval*  $\mathcal{I}(\theta^*)$ .

It is possible for some  $\theta \notin \mathcal{I}(\theta^*)$  to have a greater likelihood interval than some  $\theta' \in \mathcal{I}(\theta^*)$ . It is possible  $\exists \theta \in \mathcal{I}(\theta^*)$  st  $L(\theta; \mathbf{x}) = 0$ .

*Wald Confidence Intervals* are not invariant under reparameterisation.

These features of *Wald Confidence Intervals* motivate why we may wish to consider a different type of *Confidence Interval*.

**Definition 15.1 - Likelihood Ratio**

Define  $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$  for some  $\theta^* \in \Theta$ , let  $\{\hat{\theta}_i\}$  be a sequence of consistent *Maximum Likelihood Estimators* of  $\theta^* \in \Theta$ .

Define  $\forall \mathbf{x} \in \mathcal{X}^n$  the *Likelihood Ratio*

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

**Theorem 15.1 -**

Define  $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta^*)$  for some  $\theta^* \in \Theta$ , let  $\{\hat{\theta}_i\}$  be a sequence of consistent *Maximum Likelihood Estimators* of  $\theta^* \in \Theta$  and assume that the conditions of **Theorem 8.1** hold (implying asymptotic normality). Then

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

**Remark 15.2 -**

We observe that

$$-2 \ln \Lambda_n(\mathbf{x}) = -2 \left( \ell(\theta^*; \mathbf{x}) - \ell(\hat{\theta}_n; \mathbf{x}) \right) = 2 \left( \ell(\hat{\theta}_n; \mathbf{x}) - \ell(\theta^*; \mathbf{x}) \right)$$

*i.e.* This is twice the difference of the log-likelihoods for  $\hat{\theta}_n$  and  $\theta^*$ .

**Definition 15.2 - Confidence Sets**

Define  $\chi_{1,\alpha}^2$  to be the number st  $\mathbb{P}(W \leq \chi_{1,\alpha}^2) = 1 - \alpha$  for  $W \sim \chi_1^2$ . The *Confidence Sets*

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

are asymptotically exact  $1 - \alpha$  *Confidence Sets* for  $\theta^*$  since

$$\mathbb{P}(\theta^* \in C(\mathbf{X}_n; \theta^*)) = \mathbb{P}(-2 \ln \Lambda_n(\mathbf{X}_n) \leq \chi_{1,\alpha}^2; \theta^*) \xrightarrow{n \rightarrow \infty} 1 - \alpha$$

**Remark 15.3 - Interpreting Confidence Sets**

$C(\mathbf{x}_n)$  contains the values  $\theta$  st  $\ell(\theta; \mathbf{x}_n)$  is not too much less than  $\ell(\hat{\theta}_n; \mathbf{x}_n)$ . Hence, these confidence intervals contain those values of  $\theta$  with the greatest likelihood values.

**Remark 15.4 -**

The observed confidence set  $C(\mathbf{x})$  is defined implicitly, and finding an explicit representation of such sets might not be easy. This difficulty explains why *Wald's Approach* has been historically popular, despite its shortcomings. However, with the help of a computer, it is often easy to determine  $C(\mathbf{x})$  numerically.

**Proof 15.1 - Theorem 13.1**

Consider the second order *Taylor Expansion* of  $\ell_n(\theta; x) = \ln f_n(x; \theta)$

$$\ell_n(\theta; x) = \ell_n(\theta_0; x) + (\theta - \theta_0)\ell'_n(\theta_0; x) + \frac{(\theta - \theta_0)^2}{2}\ell''_n(\bar{\theta}; x) \text{ for some } \bar{\theta} \in [\theta, \theta_0]$$

Rearranging we find

$$\ell_n(\theta; x) - \ell_n(\theta_0; x) = (\theta - \theta_0)\ell'_n(\theta_0; x) + \frac{(\theta - \theta_0)^2}{2}\ell''_n(\bar{\theta}; x)$$

Take  $\theta = \theta^*$  and  $\theta_0 = \hat{\theta}_n$ .

Since  $\ell'_n(\hat{\theta}_n; x) - \ell_n(\hat{\theta}_n; x)$  then

$$\begin{aligned} \ln \Lambda_n(x) &= \ell_n(\theta^*; x) - \ell_n(\hat{\theta}_n; x) \\ &= \frac{(\theta^* - \hat{\theta}_n)^2}{2}\ell''_n(\bar{\theta}_n; x) \text{ for some } \bar{\theta}_n \in [\theta^*, \hat{\theta}_n] \\ \implies -2 \ln \Lambda(x) &= -(\theta^* - \hat{\theta}_n)^2 \ell''_n(\bar{\theta}_n; x) \\ &= -\left[\sqrt{nI(\theta^*)}\right]^2 (\theta^* - \hat{\theta}_n)^2 \frac{1}{nI(\theta^*)} \ell''_n(\bar{\theta}_n; x) \end{aligned}$$

Consider the random variable  $-2 \ln \Lambda(X)$ . Then we have

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

By the *Continuous Mapping Theorem*

$$\left[\sqrt{nI(\theta^*)}\right]^2 (\hat{\theta}_n - \theta^*)^2 \rightarrow_{\mathcal{D}(\cdot; \theta^*)} Z^2$$

Since  $\bar{\theta}_n \in [\theta^*, \hat{\theta}_n]$

$$-\frac{1}{n}\ell''_n(\bar{\theta}_n; X) \rightarrow_{\mathbb{P}(\cdot; \theta^*)} I(\theta^*)$$

By *Slutsky's Theorem*

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

□

**Remark 15.5 - A Rule of Thumb**

Under the assumptions of **Theorem 13.1**, the set

$$\left\{ \theta \in \Theta : \ell(\theta; \mathbf{x}) \geq \ell(\hat{\theta}_n; \mathbf{x}) - 2 \right\}$$

is an asymptotically approximate 95% confidence set for  $\theta^*$ .

**Proof 15.2 - Remark 13.1**

We have  $\chi_{0.05}^2 = 3.84$ .

The result follows from the approximation  $1.92 \approx 2$  □

**1.16 Transformation Invariant Confidence Sets****Remark 16.1 - Motivation**

Here we investigate whether the likelihood ratio approach to determining confidence sets is invariant to transformations, in contrast to *Wald's Approach*.

Consider the reparameterisation of the likelihood in terms of  $\tau := g(\theta)$  where  $g : \Theta \rightarrow G$  is bijective. We have

$$\tilde{f}(\mathbf{x}; \tau) := f(\mathbf{x}; g^{-1}(\tau)) = f(\mathbf{x}; \theta)$$

We can now define

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

We want to know whether  $\theta \in C(\mathbf{x}) \iff g(\theta) \in \tilde{C}(\mathbf{x}) \forall \mathbf{x} \in \chi^n$ .  
i.e.  $C(\mathbf{x})$  &  $\tilde{C}(\mathbf{x})$  define the same sets up to reparameterisation.

**Theorem 16.1 -**

Let  $\mathbf{X} \sim f(\cdot; \theta^*)$ ,  $C$  and  $\tilde{C}$  defined as in **Remark 14.1**.

Assume that  $g : \Theta \rightarrow G$  is bijective. Then

$$\forall \mathbf{x} \in \chi^n \text{ and } \theta^* \in \Theta, \theta \in C(\mathbf{x}) \iff g(\theta) \in \tilde{C}(\mathbf{x})$$

Thus

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

**Proof 16.1 - Theorem 14.1**

Let  $\mathbf{x} \in \chi^n$  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 similarly

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

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

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

For the last part, this correspondence implies that

$$\{\mathbf{x} \in \chi^n; \theta^* \in C(\mathbf{x})\} = \{\mathbf{x} \in \chi^2 : g(\theta^*) \in \tilde{C}(\mathbf{x})\}$$

Thus, we can conclude from the equivalence of the events

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

## 2 Testing

### 2.1 Introduction to Hypothesis Tests

#### Remark 1.1 - Motivation

Hypothesis testing allows us to make decisions about a parameter, rather than just estimating a range of values.

#### Definition 1.1 - Hypothesis Testing

*Hypothesis Testing* is a process for deciding which of two competing hypotheses,  $H_0$  or  $H_1$ , is more consistent with an observation  $\mathbf{x} = (x_1, \dots, x_n)$  of  $\mathbf{X} = (X_1, \dots, X_n) \sim f(\cdot; \theta)$ .

#### Remark 1.2 - Difference to Statistics 1

In Statistics 1 we always had the null hypothesis be  $H_0 = \mu$ . Now we consider a more general case where

- i)  $\mathbf{X} \sim f(\cdot; \theta)$  where  $\theta \in \Theta$  is unknown.
- ii) We have an observation  $\mathbf{x}$  of  $\mathbf{X}$ ;
- iii) We have formulated a null hypothesis concerning possible values of  $\theta$  (e.g.  $H_0 : \theta \in \Theta_0$ )
- iv) We have an alternative hypothesis,  $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$ .

#### Definition 1.2 - Simple Hypothesis

A *Simple Hypothesis* is a hypothesis  $H_i$  of the form  $H_i : \theta = \theta_i$  where  $\theta_i$  is a specified value, equivalently  $H_i : \theta \in \Theta_i = \{\theta_i\}$ .

#### Definition 1.3 - Composite Hypothesis

A *Composite Hypothesis* is a hypothesis  $H_i$  of the form  $H_i : \theta \in \Theta_i$  where  $\Theta_i$  is not a singleton. (i.e.  $|\Theta_i| > 1$ ).

#### Definition 1.4 - One-Sided Test

Let  $\theta$  be a scalar &  $\theta_0 \in \Theta$  be a specified value.

A *One-Sided Test* is a hypothesis test of the form

$$H_0 : \theta \leq \theta_0 \text{ and } H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \text{ and } H_1 : \theta < \theta_0$$

**Definition 1.5 - Two-Sided Test**

Let  $\theta$  be a scalar &  $\theta_0 \in \Theta$  be a specified value.

A *Two-Sided Test* is a hypothesis test of the form

$$H_0 : \theta = \theta_0 \text{ and } H_1 : \theta \neq \theta_0$$

**Definition 1.6 - Test Statistic**

A *Test Statistic* is an operation on an observation which we use to determine the outcome of a hypothesis test. Using the distribution of specified *Test Statistic* we can determine the likelihood of see a certain observation under the null-hypothesis & thus the likelihood of the null-hypothesis being true.

*N.B.* A test statistic has the signature  $T : \chi^n \rightarrow \mathbb{R}$ .

**Definition 1.7 - Critical Value**

The *Critical Value*,  $c \in \mathbb{R}$ , is an explicit value which if the value of a test statistic  $T$  exceeds it (*i.e.*  $T(\mathbf{x}) \geq c$ ) we reject the null-hypothesis.

**Definition 1.8 - Critical Region**

The *Critical Region* is the sets of observations which cause us to reject the null hypothesis.

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

where  $T$  is a *Test Statistic* &  $c$  is a *Critical Value*.

*N.B.*  $\chi^n = R \cup R^c$ .

**2.2 Hypothesis Testing - Significance and Power****Definition 2.1 - Type I & Type II Error**

*Type I Error* occurs when  $H_0$  is rejected, when in fact it is true.

*Type II Error* occurs where  $H_0$  is accepted, when in fact it is false.

Consider the table below

|               | Retain $H_0$         | Reject $H_0$        |
|---------------|----------------------|---------------------|
| $H_0$ is True | Correct              | <i>Type I Error</i> |
| $H_1$ is True | <i>Type II Error</i> | Correct             |

**Definition 2.2 - Significance Level**

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

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

Typically this is the level of improbability at which we reject the null hypothesis.

*N.B.* Common *Significance Levels* are  $\alpha = 0.05, 0.01$ .

**Example 2.1 - Testing the mean of a normal sample**

Suppose that  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$  and we want to test

$$H_0 : \mu \leq 0 \text{ and } H_1 : \mu > 0$$

We consider the test statistic  $T(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$  with critical region

$$R := \{\mathbf{x} \in \chi^n : \bar{x} \geq c\} \text{ for } c \in \mathbb{R}$$

We want to find  $c \in \mathbb{R}$  st  $\mathbb{P}(X \in R; \mu \in \Theta_0) \leq \alpha \implies \mathbb{P}(\bar{x} \geq c; \mu \in \Theta_0) \leq \alpha$ .

We know that  $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$ .

Hence  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$ .

We have

$$\mathbb{P}(\bar{X} \geq c; \mu) = \mathbb{P}\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \geq \frac{(c - \mu)\sqrt{n}}{\sigma}; \mu\right) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right)$$

We want to ensure that

$$\begin{aligned} \mathbb{P}(\bar{X} \geq c; \mu \in \Theta_0) &\leq \alpha \\ \iff 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right) &\leq \alpha \\ \iff \frac{\sqrt{n}(c - \mu)}{\sigma} &\geq \Phi^{-1}(1 - \alpha) \\ \iff c &\geq \mu + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha) \end{aligned}$$

Now observe that, for a fixed  $c$  and considering  $\mu \leq 0$  and  $\mu \in \Theta_0$

$$\mathbb{P}(\bar{X} \geq c; \mu \in \Theta_0) \leq \mathbb{P}(\bar{X} \geq c; \mu = 0)$$

Thus we can ensure that

$$\sup_{\mu \in \Theta_0} \mathbb{P}(\bar{X} \geq c; \mu) = \alpha$$

by taking  $c = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$ .

### Remark 2.1 - Change in Critical Value

*Critical Value*,  $c$ , decreases as number of sample,  $n$ , increases.

*Critical Value*,  $c$ , increases as variance,  $\sigma$ , increases.

### Remark 2.2 -

*Significance Level*,  $\alpha$ , is directly related to the phrase "statistical significance". *Statistical Significance* relates only to the *Type I Error* rate.

### 2.2.1 Power

#### Definition 2.3 - Power Function

Let  $\mathbf{X} \sim f(\cdot; \theta^*)$ ,  $T(\cdot)$  be a test statistic &  $c$  be the critical value of  $T$ .

The power function,  $\pi(\cdot; T, c) : \Theta \rightarrow [0, 1]$ , is the probability of rejecting  $H_0$  when the true value of the parameter is  $\theta \in \Theta$ .

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

### Remark 2.3 -

For a given  $\theta \in \Theta_1$ , the probability of a *Type II Error* occurring is  $1 - \pi(\theta; T, c)$ .

### Remark 2.4 -

- i) The power is non-increasing in  $c$ , regardless of whether  $\theta \in \Theta_0$  or  $\theta \in \Theta_1$ .
- ii) To make the probability of *Type I Error* tend to 0 we should make  $c$  very large so we rarely reject  $H_0$ .
- iii) If  $c$  is really large, we will rarely reject  $H_0$  even if  $\theta \in \Theta_1$ . Thus the *Power* is low and the probability of *Type II Error* is high.



**Notation 2.1 -**

When it is clear from context what test,  $T(\cdot)$ , and critical value,  $c$ , we are referring to then we may write  $\pi(\theta)$  in place of  $\pi(\theta; T, c)$ .

**Example 2.2 - Testing the Mean of a Normal Sample - Continued**

Suppose that  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$  and we want to test

$$H_0 : \mu \leq 0 \text{ and } H_1 : \mu > 0$$

We consider the test statistic  $T(\mathbf{x}) = \bar{x}$  with critical region  $R = \{\mathbf{x} \in \mathcal{X}^n : \bar{x} \geq c\}$  for some  $c \in \mathbb{R}$ .

The *Power Function* of this test is

$$\pi(\mu; T, c) = \mathbb{P}(\bar{X} \geq c; \mu)$$

We have already derived that  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$ . Hence

$$\begin{aligned} \mathbb{P}(\bar{X} \geq c; \mu) &= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{c - \mu}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}\left(Z \geq \frac{c - \mu}{\sigma/\sqrt{n}}; \mu\right) \\ &= 1 - \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(\frac{\mu - c}{\sigma/\sqrt{n}}\right) \end{aligned}$$

**Definition 2.4 - Size of a Test**

The size of a test is the greatest possible probability of making a *Type I Error*

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

*N.B.* It is the maximum power under the null-hypothesis.

**Remark 2.5 -**

Generally we choose a critical value  $c$  so that the test has size  $\alpha$ .

**Definition 2.5 - Significance Level of a Test**

A test has level  $\alpha$  if its size is less than or equal to  $\alpha$ . The corresponding test is called a *Level  $\alpha$  Test*.

**Definition 2.6 -**

When  $\Theta_0 = \{\theta_0\}$  (i.e. simple) then  $\alpha = \pi(\theta_0; T, c)$  is the significance level.

**Definition 2.7 -**

When  $\Theta_1 = \{\theta_1\}$  (i.e. simple) then  $1 - \pi(\theta_1; T, c)$  is the probability of *Type II Error*.

**Example 2.3 - Testing the mean of a normal sample - Continued**

Suppose that  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$  and that we want to test

$$H_0 : \mu \leq 0 \text{ and } H_1 : \mu > 0$$

We consider the test statistic  $T(\mathbf{x}) = \bar{x}$  with critical region  $R$ .

A test of size  $\alpha$  is obtained by choosing

$$c = \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha) = \frac{\sigma}{\sqrt{n}} z_\alpha$$

So we consider the fact that  $c = \frac{\sigma}{\sqrt{n}}z_\alpha$  and we obtain

$$\mathbb{P}\left(\bar{X} \geq \frac{\sigma}{\sqrt{n}}z_\alpha; \mu\right) = 1 - \Phi\left(z_\alpha - \frac{\mu\sqrt{n}}{\sigma}\right)$$

This gives the power  $\forall \mu \in \mathbb{R}$  and we are interested in particular in it for  $\mu > 0$ .

## 2.3 Designing Tests - Neyman-Pearson Approach

**Remark 3.1** - *Plan for Testing at Significance Level,  $\alpha$*

- i) Define a model  $f(\cdot; \theta)$  for  $\theta \in \Theta$
- ii) Define a null hypothesis  $H_0 : \theta \in \Theta_0$  and an alternative hypothesis  $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$
- iii) Define a test statistic  $T(\mathbf{x})$ .
- iv) Choose a critical value,  $c$ , st  $\sup_{\theta \in \Theta_0} \mathbb{P}(T(\mathbf{X}) \geq c; \theta) \leq \alpha$ .

*N.B.* The value of  $c$  is determined the value of  $\alpha$  (which we set).

**Theorem 3.1** - *Neyman-Pearson Lemma*

Suppose we test  $H_0 : \theta = \theta_0$  against  $H_1 : \theta = \theta_1$  and use the *Likelihood Ratio Test Statistic*

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

Let the *Critical Value*,  $c_{NP} \geq 0$ , be st the test has size  $\alpha$

$$\mathbb{P}(T_{NP} \geq c_{NP}; \theta_0) = \alpha$$

Then, this test is the most powerful level  $\alpha$  test.

*i.e.* Among all tests with level  $\alpha$ , this test maximises the power function.

**Proof 3.1** - *Theorem 2.1*

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

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

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

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

Now consider the difference

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

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  $\alpha$  level test, we know  $\alpha(T, c) \geq c$  while  $(T_{NP}, c_{NP})$  specifies a size  $\alpha$  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  $\alpha$  level test, we conclude that  $(T_{NP}, c_{NP})$  is the most powerful test with level  $\alpha$ .  $\square$

### Remark 3.2 - Neyman-Pearson with Non-Continuous Random Variable

If  $T(\mathbf{X})$  is not a continuous random variable, then it is possible that no such  $c_{NP}$  exists. In this situation we perform an appropriate randomised test, and this will also be the most powerful size  $\alpha$  test.

*N.B.* The details of this are not covered in this course.

### Definition 3.1 - Neyman-Pearson Procedure

For **Theorem 2.1** we can deduce the *Neyman-Pearson Procedure* for testing two simple hypotheses

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

$$T(\mathbf{x}) = \frac{f_n(\mathbf{x}; \theta_1)}{f_n(\mathbf{x}; \theta_0)} = \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})}$$

- ii) Choose a critical value  $c$  in order to target a particular significance level,  $\alpha$ , st

$$\alpha = \pi(\theta_0) = \mathbb{P}(T(\mathbf{X}) \geq c; \theta_0)$$

- iii) Compute the *Power*

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

- iv) Compute  $T(\mathbf{x})$  and report whether  $T(\mathbf{x}) \geq c$  as well as the power  $\pi(\theta_1, T, c)$  or the *Type II Error* rate  $1 - \pi(\theta_1, T, c)$

**Remark 3.3 - Limitations of Neyman-Pearson Approach**

- i) Often just rejecting  $H_0$  or retaining  $H_0$  is not satisfactory, we may want more information.
- ii) It is not obvious how to calibrate a likelihood ratio test (*i.e.* TO find the critical value or compute the power function).

**2.4 Testing - p-Values, Equivalent Test Statistics and Computing the Power Function****Remark 4.1 - Motivation for p-Value**

Many studies prefer not to select in advance just one significance level  $\alpha$ , or they may wish to report something more informative than a binary decision. In such cases, they can report the  $p$ -value associated with the observed test statistic.

**Definition 4.1 - p-Value**

Let  $\mathbf{X} \sim f_n(\cdot; \theta^*)$  for some  $\theta^* \in \Theta$ .

The  $p$ -Value for a test with test statistic  $T(\mathbf{x})$  is the probability of seeing a test statistic  $T(\mathbf{X})$  at least as extreme as  $T(\mathbf{x})$ .

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

Equivalently,  $p(\mathbf{x})$  is the smallest significance level at which we would reject  $H_0$ .

**Remark 4.2 - p-Value Intuition**

Intuitively,  $p$ -value is a measure of the evidence against  $H_0$ . The smaller it is, the less likely it is that  $\mathbf{x}$  is a realisation of  $\mathbf{X} \sim f(\cdot; \theta_0)$ , resulting in strong evidence against  $H_0$ .

*N.B.* A large  $p$ -value is not evidence in favour of  $H_0$ , nor is it necessarily evidence in favour of  $H_1$  as  $H_1$  is not involved at all when computing the  $p$ -value.

**Remark 4.3 - Standard Caution**

$p(\mathbf{x})$  is *not* the probability that  $H_0$  is true. It is the probability to observe the data we observed if  $\theta_0$  is true.

**Remark 4.4 - Distribution of p-Value**

When using a simple null hypothesis  $\Theta_0 = \{\theta_0\}$  and assuming  $T(\mathbf{X})$  is a continuous random variable when  $\mathbf{X} \sim f(\cdot; \theta_0)$ , the distribution of  $p(\mathbf{X})$  is in fact uniform under the null hypothesis.

**Example 4.1 - Normal**

The model is  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, 1)$  and we want to test  $H_0 : \mu = \mu_0 < 0$  against  $H_1 : \mu = \mu_1 > 0$ . The  $p$ -value for  $T(\mathbf{x}) = \bar{x} = \frac{1}{n} \sum x_i$  is

$$p(\mathbf{x}) := \sup_{\mu \in \Theta_0} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq T(\mathbf{x}) = \bar{x}; \mu\right) = \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq T(\mathbf{x}) = \bar{x}; \mu\right)$$

A very large positive value of the empirical mean leads to a small  $p$ -value and is an indication of how unlikely it is to have observed  $\mathbf{x}$  if it was a realisation of  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu_0, 1)$ .

A large  $p$ -value is not an argument in favour of  $H_0$ , in fact it could suggest that  $T(\mathbf{x})$  is an unlikely realisation under  $H_0$ .

We have already calculated this kind of expression under the null hypothesis  $\bar{X} \sim \text{Normal}(\mu_0, \frac{1}{n})$  so

$$\begin{aligned} \mathbb{P}(\bar{X} \geq c; \mu) &= \mathbb{P}(\sqrt{n}(\bar{X} - \mu_0) \geq \sqrt{n}(c - \mu_0); \mu_0) \\ &= \mathbb{P}(Z \geq \sqrt{n}(c - \mu_0)) \\ &= 1 - \Phi(\sqrt{n}(c - \mu_0)) \end{aligned}$$

It follows that

$$p(\mathbf{x}) = 1 - \Phi(\sqrt{n}(\bar{x} - \mu_0))$$

**Definition 4.2 - Equivalent Statistics**

A statistic  $T'(\mathbf{x})$  is equivalent to  $T(\mathbf{x})$  if  $\forall$  critical values  $c \in \mathbb{R}$  of  $T(\cdot)$  we can find  $c' \in \mathbb{R}$  we can find  $c' \in \mathbb{R}$  st  $\forall \mathbf{x} \in \mathcal{X}^n$

$$T(\mathbf{x}) \geq c \iff T'(\mathbf{x}) \geq c'$$

Equivalently,  $\forall c \in \mathbb{R}$  there exist  $c' \in \mathbb{R}$  such that the corresponding critical regions of  $T(\cdot)$  and  $T'(\cdot)$  respectively coincide

$$\{\mathbf{x} \in \mathcal{X}^n : T(\mathbf{x}) \geq c\} = \{\mathbf{x} \in \mathcal{X}^n : T'(\mathbf{x}) \geq c'\}$$

**Proposition 4.1 - Proving Equivalence**

To verify that  $T'(\mathbf{x})$  is an *Equivalent Statistic* to  $T(\mathbf{x})$  it is sufficient to factorise  $T(\mathbf{x})$  as

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

where  $M$  is independent of  $\mathbf{x}$  and  $f$  is increasing & bijective.

**Proof 4.1 - Proposition 4.1**

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

**Example 4.2 - Geometric Example**

Let that  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Geometric}(p)$  so that  $f(x; p) = (1-p)^{x-1}p\mathbb{1}\{x \in \mathbb{N} \setminus \{0\}\}$ .

Suppose that we want to test  $H_0 : p = p_0$  against  $H_1 : p = p_1$  with  $p_0 > p_1$ .

$$T_{NP}(\mathbf{x}) = \frac{f_n(\mathbf{x}; p_1)}{f_n(\mathbf{x}; p_0)} = \frac{\prod_{i=1}^n f(x_i; p_1)}{\prod_{i=1}^n f(x_i; p_0)} = \prod_{i=1}^n \frac{f(\mathbf{x}_i; p_1)}{f(\mathbf{x}_i; p_0)}$$

So for  $x \in X$

$$\frac{f(x; p_1)}{f(x; p_0)} = \frac{(1-p_1)^{x-1}p_1}{(1-p_0)^{x-1}p_0} = \left(\frac{1-p_1}{1-p_0}\right)^x \left(\frac{1-p_1}{1-p_0}\right)^{-1} \left(\frac{p_1}{p_0}\right)$$

So

$$T_{NP}(\mathbf{x}) = \left(\frac{1-p_1}{1-p_0}\right)^{\sum x_i} \left(\frac{1-p_1}{1-p_0}\right)^{-n} \left(\frac{p_1}{p_0}\right)^n = \left(\frac{1-p_1}{1-p_0}\right)^{n\bar{x}} \underbrace{\left(\frac{1-p_1}{1-p_0}\right)^{-n} \left(\frac{p_1}{p_0}\right)^n}_M$$

Note that

$$p_0 > p_1 \implies 1-p_0 < 1-p_1 \implies \frac{1-p_1}{1-p_0} > 1$$

So  $\left(\frac{1-p_1}{1-p_0}\right)^{n\bar{x}}$  is increasing with  $\bar{x}$ .

It follows that  $T'(\mathbf{x}) = \bar{x}$  is an equivalent test statistic to  $T_{NP}$ .

If  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Geometric}(p)$  then  $n\bar{x} \sim \text{Negative-Binomial}(n, p)$ .

Hence we can compute  $c_{NP}$  or compute the power function.

**Example 4.3 - Normal Example**

The model is  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, 1)$  and we want to test  $H_0 : \mu = 0$  against  $H_1 : \mu = 1$ .  
The *Neyman-Pearson Test Statistic* is

$$\begin{aligned} T_{NP}(\mathbf{x}) &= \frac{f_n(\mathbf{x}; \mu = 1)}{f_n(\mathbf{x}; \mu = 0)} \\ &= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum (x_i - 1)^2}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum (x_i - 0)^2}} \\ &= e^{-\frac{1}{2} (\sum x_i^2 - 2 \sum x_i + n - \sum x_i^2)} \\ &= e^{-\frac{1}{2} (-2n\bar{x} + n)} \\ &= \underbrace{e^{-\frac{n}{2}}}_M e^{n\bar{x}} \end{aligned}$$

Since  $T_{NP}$  is increasing in terms of  $\bar{x}$ ,  $\bar{x}$  is an equivalent test statistic to  $T_{NP}$ .

To relate  $T_{NP} \geq c_{NP}$  with  $T(\mathbf{x}) = \bar{x} \geq c$ .

We have

$$\begin{aligned} T_{NP}(\mathbf{x}) \geq c_{NP} &\iff e^{n\bar{x} - \frac{n}{2}} \geq c_{NP} \\ &\iff n\bar{x} - \frac{n}{2} \geq \ln c_{NP} \\ &\iff \bar{x} - \frac{1}{2} \geq \frac{1}{n} \ln c_{NP} \\ &\iff \bar{x} \geq \underbrace{\frac{1}{2} + \frac{1}{n} \ln c_{NP}}_c \end{aligned}$$

So  $c_{NP} = e^{n(c - \frac{1}{2})}$ .

Now we can compute the power function.

We know that  $\bar{X} \sim \text{Normal}(\mu, \frac{1}{n})$  for  $\alpha \in (0, 1)$ .

We find  $c$  by solving

$$\begin{aligned} \pi(\mu_0; T, c) &= \alpha \\ \implies \mathbb{P}(\bar{X} \geq c; \mu_0) &= \alpha \\ \implies \mathbb{P}\left(Z \geq \frac{c - 0}{1/\sqrt{n}}\right) &= \alpha \\ \implies 1 - \Phi(c\sqrt{n}) &= \alpha \\ \implies \Phi(c\sqrt{n}) &= 1 - \alpha \\ \implies c\sqrt{n} &= \Phi^{-1}(1 - \alpha) \\ \implies c &= \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{n}} \\ &= \frac{z_\alpha}{\sqrt{n}} \end{aligned}$$

Hence  $c_{NP} = e^{n(\frac{z_\alpha}{\sqrt{n}} - \frac{1}{2})}$  We can also compute *Type II Error* probability

$$\begin{aligned} 1 - \pi(1) &= \mathbb{P}(\bar{X} < c; \mu = 1) \\ &= \mathbb{P}\left(Z < \frac{c - 1}{1/\sqrt{n}}\right) \\ &= \Phi(\sqrt{n}c - \sqrt{n}) \\ &= \Phi(z_\alpha - \sqrt{n}) \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

**2.5 Uniformly Most Powerful Tests****Definition 5.1 - Uniformly Most Powerful Test**

Consider a test involving composite hypothesis  $H_0 : \theta \leq \theta_0$  against  $H_1 : \theta > \theta_0$ .

A *Uniformly Most Powerful Test* is a test  $(T, c)$  which has the largest power  $\pi(\theta; T, c)$  among all possible tests, uniformly in  $\theta \in \Theta_1$ . That is a  $(T, c)$  st  $\forall \theta \in \Theta_1$  and any test statistic  $(T', c')$

$$\pi(\theta; T, c) \geq \pi(\theta; T', c')$$

**Remark 5.1 -**

The *Type II Error Rate* depends on a specific value of  $\theta \in \Theta_1$ . Typically, the *Type II Error Rate* is close to  $1 - \alpha$  for values of  $\theta \in \Theta_1$  "very close to being in"  $\Theta_0$ . i.e.  $\pi(\theta; T, c) \approx \alpha$  for  $\theta = \theta_0 + \varepsilon$  for  $\varepsilon$  very small.

**Theorem 5.1 -**

Let  $\Theta_1 = \{\theta : \theta > \theta_0\}$  for some  $\theta_0 \in \Theta$ .

Assume that for the simple hypotheses

$$H'_0 : \theta = \theta_1 \quad \text{against} \quad H'_1 : \theta = \theta_2$$

The *Neyman-Pearson Test Statistic*

$$T_{NP}(\mathbf{x}) = \frac{f_n(\mathbf{x}; \theta_2)}{f_n(\mathbf{x}; \theta_1)}$$

is equivalent to the same test statistic  $T(\mathbf{x})$  for any  $\theta_1 < \theta_2$  and  $T(\mathbf{x})$  does not depend on  $\theta_1$  or  $\theta_2$ .

Then  $T(\mathbf{x})$  is the uniformly most powerful test statistic for

$$H_0 : \theta \leq \theta_0 \quad \text{against} \quad H_1 : \theta > \theta_0$$

and the associated  $p$ -value is

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

**Example 5.1 - Poisson**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda)$  for some  $\lambda > 0$  and we want to test  $H_0 : \lambda \leq \lambda_0$  against  $H_1 : \lambda > \lambda_0$ .

Compute the  $p$ -value associated with this test.

Consider  $H_0 : \lambda = \lambda_0$  &  $H_1 : \lambda = \lambda_1$  where  $\lambda_1 > \lambda_0$ .

$$T_{NP}(\mathbf{x}) = \prod_{i=1}^n \left( \frac{e^{-\lambda_1} \lambda_1^{x_i} (x_i!)^{-1}}{e^{-\lambda_0} \lambda_0^{x_i} (x_i!)^{-1}} \right) = e^{-n(\lambda_1 - \lambda_0)} \left( \frac{\lambda_1}{\lambda_0} \right)^{\sum x_i}$$

Since  $\lambda_1 > \lambda_0$  we have  $T_{NP}(\mathbf{x})$  is an increasing function in terms of  $\sum x_i$ .

So  $S_n := \sum x_i$  is an equivalent test statistic and does not depend on  $\lambda_0$  or  $\lambda_1$ .

Hence,  $S_n$  is a *Uniformly Most Powerful Test*. To find the  $p$ -value

$$p(\mathbf{x}) = \mathbb{P}(S_n(\mathbf{X}) \geq S_n(\mathbf{x}); \lambda_0)$$

We use  $\lambda_0$  in this scenario since it is the value which is most likely to produce extreme values, in general, for  $T(\mathbf{X})$ .

We know that  $S_n(\mathbf{X}) = \sum X_i \sim \text{Poisson}(n\lambda)$ . So

$$p(\mathbf{x}) = \sum_{k \geq S_n(\mathbf{x})} e^{-\lambda_0 n} (n\lambda_0)^k \frac{1}{k!}$$

*Alternatively*

If  $n$  is large  $S_n(\mathbf{X}) \simeq \text{Normal}(n\lambda, n\lambda)$ .

So

$$p(\mathbf{x}) \simeq \mathbb{P}\left(Z \geq \frac{S_n(\mathbf{x}) - n\lambda_0}{\sqrt{n\lambda_0}}\right) = 1 - \Phi\left(\frac{S_n(\mathbf{x}) - n\lambda_0}{\sqrt{n\lambda_0}}\right)$$

where  $Z \sim \text{Normal}(0, 1)$ .

**Example 5.2 - Geometric**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Geometric}(p)$  we have already shown that for two simple hypotheses  $\bar{X}$  is equivalent

to the likelihood ratio test statistic when  $p_0 > p_1$ .

It follows that  $T(\mathbf{x}) = S_n(\mathbf{x})$  is equivalent.

We have noticed that  $S_n(\mathbf{X}) \sim \text{NegBinomial}(n, p)$ .

We shall compute the  $p$ -value associated to hypotheses  $H_0 : p \leq p_0$  against  $H_1 : p > p_0$ .

From **Example 2.4.2** we have that for  $H'_0 : p = p_0$  against  $H'_1 : p = p_1$

$$T_{NP}(\mathbf{x}) = \left(\frac{p_1}{p_0}\right)^n \left(\frac{1-p_1}{1-p_0}\right)^{-n} \left(\frac{1-p_1}{1-p_0}\right)^{n\bar{x}}$$

For  $p_1 > p_0$  we see that  $T_{NP}$  is a decreasing function (since the last two terms are  $< 0$ ) in terms of  $\sum X_i =: S_n$ .

Hence it is increasing in terms of  $T(\mathbf{x}) := -S_n(\mathbf{x})$ .

Since  $S_n$  is independent of  $p_0$  &  $p_1$  we have that  $-S_n$  is a *Uniformly Most Powerful Test*.

$$p\text{-value} := p(\mathbf{x}) - \mathbb{P}(-S_n(\mathbf{X}) \geq -S_n(\mathbf{x}); p_0) = \mathbb{P}(S_n(\mathbf{X}) \leq S_n(\mathbf{x}); p_0)$$

where  $S_n(\mathbf{X}) \sim \text{NegativeBinomial}(n, p_0)$ .

**Remark 5.2** - *Uniformly Most Powerful Tests need not exist*

In general, *Uniformly Most Powerful Tests* need not exist.

It might be the case that  $(T_1, c_1)$  is best for, say,  $\theta_{1,1} \in \Theta_1$ .

i.e.  $\forall (T', c')$

$$\pi(\theta_{1,1}; T_1, c_1) \geq \pi(\theta_{1,1}; T', c')$$

but  $\exists (T_2, c_2)$  st

$$\pi(\theta_{1,2}; T_2, c_2) > \pi(\theta_{1,2}; T_1, c_1) \quad \theta_{1,2} \in \Theta_1$$

i.e.  $(T_2, c_2)$  is better than  $(T_1, c_1)$ .

## 2.6 Generalised Likelihood Ratio Test

**Remark 6.1** - *Generalised Tests*

In the most general case we would like to test  $H_0 : \theta \in \Theta_0$  against  $H_1 : \theta \in \Theta_1$ .

There is no guarantee of the existence of an optimal test statistic.

**Proposition 6.1** - *Generalised Likelihood Ratio Test*

We can generalise the likelihood ratio test for simple hypotheses from

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

to

$$T_{\text{suggested}}(\mathbf{x}) := \frac{\sup_{\theta \in \Theta_1} (f_n(\mathbf{x}; \theta))}{\sup_{\theta \in \Theta_0} (f_n(\mathbf{x}; \theta))}$$

N.B. The generalised simple hypotheses are  $\Theta_i = \{\theta_i\}$  for  $\theta_i \in \Theta$ .

**Definition 6.1** - *Likelihood Ratio*

We define a *Likelihood Ratio*

$$\begin{aligned} \Lambda_n(\mathbf{x}) &:= \frac{\sup_{\theta \in \Theta_0} f_n(\mathbf{x}; \theta)}{\sup_{\theta \in \Theta} f_n(\mathbf{x}; \theta)} \\ &= \frac{\sup_{\theta \in \Theta_0} f_n(\mathbf{x}; \theta)}{f_n(\mathbf{x}; \underbrace{\hat{\theta}_n}_{\text{MLE}})} \\ &= \min \left\{ \underbrace{1}_{\hat{\theta}_n \in \Theta_0}, \underbrace{\frac{\sup_{\theta \in \Theta_0} f_n(\mathbf{x}; \theta)}{\sup_{\theta \in \Theta_1} f_n(\mathbf{x}; \theta)}}_{\hat{\theta}_n \notin \Theta_0} \right\} \end{aligned}$$



**Remark 6.2 - Likelihood Ratio**

- i) The denominator corresponds to plugging in the *Maximum Likelihood Estimate* in the likelihood (assuming it exists and is unique).
- ii) The last equality follows from the fact that  $\sup_{\theta \in \Theta} f_n(\mathbf{x}; \theta) \geq \sup_{\theta \in \Theta_0} f_n(\mathbf{x}; \theta)$ .  
if the inequality is strict then

$$\sup_{\theta \in \Theta} f_n(\mathbf{x}; \theta) = \sup_{\theta \in \Theta_1} f_n(\mathbf{x}; \theta) > \sup_{\theta \in \Theta_0} f_n(\mathbf{x}; \theta)$$

and if it is an equality then  $\Lambda_n(\mathbf{x}) = 1$ .

**Definition 6.2 - Nested Parameter Space**

Let  $\Theta \subseteq \mathbb{R}^d$  and define  $\phi = (\phi_1, \phi_2) : \Theta \rightarrow \Phi_1 \times \Phi_2$  to a *continuously differentiable bijection* with  $\Phi_1 \subseteq \mathbb{R}^r$  and  $\Phi_2 \subseteq \mathbb{R}^{d-r}$ .

$\Theta_0$  is said to be *Nested* within  $\Theta$  if for some  $c \in \Phi_1 \subseteq \mathbb{R}^r$

$$\Theta_0 = \{\theta \in \Theta : \phi_1(\theta) = c\}$$

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

**Example 6.1 - Nested Parameter Space**

Suppose  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$  then  $\Theta = \{\mu, \sigma^2\}$ .

Suppose  $\phi_1(\mu, \sigma^2) = \mu$  &  $\phi_2(\mu, \sigma^2) = \sigma^2$  then  $\Phi_1 = \mathbb{R}$  &  $\Phi_2 = \mathbb{R}^+ \subseteq \mathbb{R}$ .

Then  $\Theta_0 := \{(\mu, \sigma^2) : \phi_1(\mu, \sigma^2) = 0\}$  is *Nested* in  $\Theta$ .

**Theorem 6.1 - Distribution of Test Statistics in Nested Parameter Spaces**

Let  $\Theta \subset \mathbb{R}^d$  and  $\mathbf{X} \stackrel{\text{iid}}{\sim} f(\cdot; \theta)$  for  $\theta \in \Theta_0$  (*i.e.*  $H_0$  is true) where  $\Theta_0$  is nested in  $\Theta$ . Then

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

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

*N.B.* The proof is not covered in this course but relies on the Taylor expansion of the likelihood.

**Remark 6.3** - If value of the null hypothesis is fixed,  $H_0 : \theta = 0$ , then  $\dim(\Theta_0) = 0$

**Example 6.2 -**

Let  $\mathbf{X} \stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda^*)$  for some  $\lambda^* > 0$ .

Consider the test

$$H_0 : \lambda = \lambda_0 \quad \text{against} \quad H_1 : \lambda \neq \lambda_0$$

Then

$$\begin{aligned} T_n(\mathbf{x}) &= -2 \ln \left( \frac{f_n(\mathbf{x}; \lambda_0)}{f_n(\mathbf{x}; \hat{\lambda}_n)} \right) \quad \text{where } \hat{\lambda}_n \text{ is MLE of } \lambda \\ &= -2 \ln[\ell_n(\lambda; \mathbf{x}) - \ell_n(\hat{\lambda}_n; \mathbf{x})] \end{aligned}$$

We know

$$\begin{aligned} f_n(\mathbf{x}; \lambda) &= \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \quad \text{by independence} \\ &\propto e^{-n\lambda} \lambda^{n\bar{x}} \\ \implies \ell_n(\lambda; \mathbf{x}) &= c - n\lambda + n\bar{x} \ln \lambda \\ \text{Taking } \ell'_n(\lambda; \mathbf{x}) &= 0 \\ \implies -n + \frac{\bar{x}n}{\lambda} &= 0 \\ \implies \frac{\bar{x}n}{\lambda} &= \bar{x} \end{aligned}$$

We confirm  $\hat{\lambda}$  is a *Maximum Likelihood Estimate* by showing  $\ell_n''(\bar{x}; x) < 0$ .

So

$$\begin{aligned} T_n(\mathbf{x}) &= -2 \left[ -n(\lambda_0 - \hat{\lambda}_n) + n\bar{x} \ln \frac{\lambda_0}{\hat{\lambda}_n} \right] \\ &= -2 \left[ -n(\lambda_0 - \bar{x}) + n\bar{x} \ln \frac{\lambda_0}{\bar{x}} \right] \end{aligned}$$

We know that

$$T_n(\mathbf{X}) \sim \chi_{1-0}^2 = \chi_1^2 \text{ under } H_0$$

since  $\dim(\Theta) = 1$  and  $\dim(\Theta_0) = 0$ .

### Definition 6.3 - Two Sided Test

A *Two Sided Hypothesis Test* is a hypothesis test where the alternative hypothesis covers two separate regions of the parameter space. The general form is

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

### Remark 6.4 - Connection to Confidence Intervals

Recall that a test of size approximately  $\alpha$  is obtained by retain  $H_0$  if

$$T_n(\mathbf{x}) = -2 \left[ \ell_n(\theta_0; \mathbf{x}) - \ell_n(\hat{\theta}_n; \mathbf{x}) \right] < \chi_{r,\alpha}^2$$

The values of  $\theta_0$  which lead to the retention of  $H_0$  are

$$\begin{aligned} C(\mathbf{x}) &= \left\{ \theta : -2[\ell_n(\theta; \mathbf{x}) - \ell_n(\hat{\theta}_n; \mathbf{x})] < \chi_{r,\alpha}^2 \right\} \\ &= \left\{ \theta : \ell_n(\theta; \mathbf{x}) > \ell_n(\hat{\theta}_n; \mathbf{x}) - \frac{1}{2}\chi_{r,\alpha}^2 \right\} \end{aligned}$$

This is an asymptotically exact  $1 - \alpha$  confidence set for  $\theta$ .

## 2.7 Categorical Distributions and Pearson's Chi-Squared Test

### Definition 7.1 - Categorical Distribution

Let  $Y$  be a random variable which takes one value from a finite set  $\{1, \dots, m\}$ .

Let  $\mathbf{p} := (p_1, \dots, p_m)$  be a vector where  $p_i = \mathbb{P}(Y = i)$  then

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

*N.B.*  $\sum_{i=1}^m p_i = 1$ ,  $p_i \in [0, 1] \forall i \in [1, m]$  and  $\mathbf{p}$  is a vector of  $m - 1$  free parameters.

### Remark 7.1 - Bernoulli( $p$ ) $\sim$ Categorical( $1 - p, p$ )

### Definition 7.2 - Counts

Let  $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \text{Categorical}(\mathbf{p})$ .

We define the random variable  $N_k$  to model the number of times  $k$  occurs in a realisation of  $Y_1, \dots, Y_n$ .

$$N_k := \sum_{i=1}^n \mathbb{1}\{Y_i = k : i \in [1, n]\}$$

This definition gives rise to the random vector

$$\mathbf{X} := (N_1, \dots, N_m) \sim \text{Multinomial}(n, \mathbf{p})$$

with

$$\mathbb{P}(N_1 = n_1, \dots, N_m = n_m; \mathbf{p}) = \mathbb{1} \left\{ \sum_{i=1}^m n_i = n \right\} \left\{ \frac{n!}{\prod_{i=1}^m n_i} \right\} \prod_{i=1}^m p_i^{n_i}$$

## 0 Appendix

### Definition 0.1 - Gradient

$$\nabla f(\boldsymbol{\theta}; \mathbf{x}) := \left( \frac{\partial f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_1}, \dots, \frac{\partial f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_n} \right)$$

### Definition 0.2 - Hessian

$$\nabla^2 f(\boldsymbol{\theta}; \mathbf{x}) := \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_1^2} & \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_2 \partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_1 \partial \theta_n} & \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_2 \partial \theta_n} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta}; \mathbf{x})}{\partial \theta_n^2} \end{pmatrix}$$

### 0.1 Notation

| Notation                                            | Denotes                                                                                               |
|-----------------------------------------------------|-------------------------------------------------------------------------------------------------------|
| $Z_n \rightarrow_{\mathbb{P}} Z$                    | $\{Z_n\}_{n \in \mathbb{N}}$ converges in <i>Probability</i> to random variable $Z$ .                 |
| $Z_n \rightarrow_{\mathcal{D}} Z$                   | $\{Z_n\}_{n \in \mathbb{N}}$ converges in <i>Distribution</i> to random variable $Z$ .                |
| $Z_n \rightarrow_{qm} Z$                            | $\{Z_n\}_{n \in \mathbb{N}}$ converges in <i>Quadratic Mean</i> to random variable $Z$ .              |
| $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ | Scalar or vector parameter characterising a probability distribution                                  |
| $\hat{\theta}$                                      | Estimation for the value of the parameter $\theta$                                                    |
| $\theta^*$                                          | True value of the parameter $\theta$                                                                  |
| $\mathbb{P}$                                        | Probability measure $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$                                     |
| $\Omega$                                            | Sample space                                                                                          |
| $X$                                                 | Scalar random variable                                                                                |
| $\mathcal{F}$                                       | Sigma field (Set of events)                                                                           |
| $\chi$                                              | Support of rv $X$ . A set $\chi$ is definitely in it <i>i.e.</i> $\mathbb{P}(X \in \chi; \theta) = 1$ |
| $\mathbf{X}$                                        | Vector consisting of scalar random variables                                                          |

### 0.2 R

| Command                      | Result                                                       |
|------------------------------|--------------------------------------------------------------|
| <code>hist(a)</code>         | Plots a histogram of the values in array $a$                 |
| <code>mean(a)</code>         | Returns the mean value of array $a$                          |
| <code>rbinom(s, n, p)</code> | Samples $n$ of $Bi(n, p)$ random variables                   |
| <code>rep(v, n)</code>       | Produces an array of size $n$ where each entry has value $v$ |
| <code>x ← v</code>           | Maps value $v$ to variable $x$                               |

### 0.3 Probability Distributions

#### Definition 3.1 - Binomial Distribution

Let  $X$  be a discrete random variable modelled by a *Binomial Distribution* with  $n$  events and rate of success  $p$ .

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbb{E}(X) = np \quad \& \quad \text{Var}(X) = np(1-p)$$

#### Definition 3.2 - Gamma Distribution

Let  $T$  be a continuous random variable modelled by a *Gamma Distribution* with shape parameter

$\alpha$  & scale parameter  $\lambda$ . Then

$$f_T(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \quad \text{for } x > 0$$

$$\mathbb{E}(T) = \frac{\alpha}{\lambda} \quad \& \quad \text{Var}(T) = \frac{\alpha}{\lambda^2}$$

*N.B.*  $\alpha, \lambda > 0$ .

**Definition 3.3 - Exponential Distribution**

Let  $T$  be a continuous random variable modelled by a *Exponential Distribution* with parameter  $\lambda$ . Then

$$f_T(t) = \mathbf{1}\{t \geq 0\} \cdot \lambda e^{-\lambda t}$$

$$F_T(t) = \mathbf{1}\{t \geq 0\} \cdot (1 - e^{-\lambda t})$$

$$\mathbb{E}(X) = \frac{1}{\lambda} \quad \& \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

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

**Definition 3.4 - Normal Distribution**

Let  $X$  be a continuous random variable modelled by a *Normal Distribution* with mean  $\mu$  & variance  $\sigma^2$ .

Then

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

$$M_X(\theta) = e^{\mu\theta + \sigma^2\theta^2(1/2)}$$

$$\mathbb{E}(X) = \mu \quad \& \quad \text{Var}(X) = \sigma^2$$

**Definition 3.5 - Poisson Distribution**

Let  $X$  be a discrete random variable modelled by a *Poisson Distribution* with parameter  $\lambda$ . Then

$$p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{For } k \in \mathbb{N}_0$$

$$\mathbb{E}(X) = \lambda \quad \& \quad \text{Var}(X) = \lambda$$

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