Statistics

John M. Noble,
Mathematical Statistics,
Institute of Applied Mathematics and Mechanics,
Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw,
ul. Banacha 2,
02-097 Warszawa, Poland

Contents

1 Statistical Models **9**
1.1 Introduction . . . 9
1.2 Parametrisations and Parameters . . . 10
1.3 Statistics as Functions on the Sample Space . . . 12
1.4 The Empirical Distribution Function . . . 14

2 Sufficiency **29**
2.1 Sufficienct Statistics . . . 29
2.2 The Factorisation Theorem . . . 30
2.3 Minimal Sufficiency . . . 32

3 Exponential Families **41**
3.1 Properties of Exponential Families . . . 43
3.2 Rank of an Exponential Family . . . 45

4 Parameter Estimation **61**
4.1 Parameter Estimation: Basic Heuristics . . . 61
4.2 Unbiased Estimation . . . 61
4.3 Least Squares . . . 62
4.4 Method of Moments (MOM) . . . 63
4.5 Maximum Likelihood Method . . . 64
4.5.1 Maximum Likelihood as Minimum Contrast . . . 65

5 Maximum Likelihood **77**
5.1 Likelihood Equations . . . 77
5.2 Maximum Likelihood for Exponential Families . . . 79
5.3 The EM (Expectation / Maximisation) Algorithm . . . 81

6 The Information Inequality **93**
6.1 The Information Inequality: One Parameter . . . 93
6.2 The Optimal Linear Predictor . . . 100
6.2.1 Introduction and Motivation . . . 100

6.2.2 The Linear Predictor . . . 100

6.3 Information Inequality with Several Parameters . . . 101

6.4 Information and Log Partition Function for Exponential Families . . . 102

7 Confidence Intervals **115**

7.1 Introduction, Definitions and First Example . . . 115

7.2 Pivots . . . 116

7.3 An Inverse Method . . . 117

7.4 Pivotal Quantities for the Normal Distribution . . . 119

7.4.1 The Chi Squared Distribution Intervals for sigma . . . 120

7.4.2 Normal random sample: Interval for mu, unknown sigma . . . 122

8 Hypothesis Testing (I) **135**

8.1 Introduction . . . 135

8.2 Methods of Finding Tests . . . 135

8.2.1 Likelihood Ratio Test . . . 135

8.3 Evaluating Tests . . . 138

8.4 Most Powerful Tests . . . 139

9 Hypothesis Testing (II) **149**

9.1 Monotone Likelihood Ratio . . . 149

9.2 Union-Intersection and Intersection-Union Tests . . . 151

9.3 p-Values . . . 153

9.4 Interval Estimator by Inverting a Test Statistic . . . 155

10 Gaussian Linear Models (I) **169**

10.1 Introduction . . . 169

10.2 The Canonical Form of the Gaussian Linear Model . . . 171

10.3 Parameter Estimation for Gaussian Linear Models . . . 173

10.4 Confidence Intervals and Prediction . . . 176

11 Gaussian Linear Models (II) **189**

11.1 Tests and Confidence Intervals . . . 189

11.1.1 Comparison of models, unknown sigma . . . 192

11.2 Analysis of Variance . . . 195

12 Complete Sufficiency and UMVU Estimators **211**

12.1 Completeness and Parameter Estimation . . . 214

13 Asymptotic Results **225**

13.1 Consistency . . . 225

13.1.1 The Weak Law of Large Numbers . . . 225

13.1.2 Consistency of Minimum Contrast Estimators 229
13.2 The Delta Method 230
13.3 Asymptotic Results for Maximum Likelihood 233
13.4 Asymptotic Distribution of Maximum Likelihood Estimators 236

Tutorial 1

Definitions and Notation

- Let $X_1, \dots, X_n$ be independent identically distributed. Such a collection is a *random sample*. The *order statistics* are defined as: $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$; $X_{k:n}$ is the kth lowest of the collection. Another common notation is the following: the order statistics of a sample size n are often written as $X_{(1)}, \dots, X_{(n)}$

- The notation $X \sim Be(p)$ will be used to denote a Bernoulli trial with success probability p; $\mathbb{P}(X=1) = p$, $\mathbb{P}(X=0) = 1-p$.

- The notation $Y \sim Bi(n,p)$ will be used to denote the Binomial distribution;

$$\mathbb{P}(Y=k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

- Let X be a random variable. If Y is a continuous random variable with density f_Y, then

$$\mathbb{P}(X \in A) = \int \mathbb{P}(X \in A | Y = y) f_Y(y) dy.$$

Exercises

1. A generalisation of i.i.d. random variables is *exchangeable* random variables, an idea due to de Finetti (1972). The random variables $X_1, \dots, X_n$ are *exchangeable* if any permutation of any subset of them of size $k \leq n$ has the same distribution. This exercise gives an example of random variables that are exchangeable but not i.i.d. Let $X_i | P$ be i.i.d. Be(P) variables, where $P \sim U(0,1)$. That is, P is uniformly distributed on $[0,1]$. Conditioned on $P = p$, the variables $X_1, \dots, X_n$ are i.i.d. $Be(p)$ variables.

   (a) Show that the marginal distribution of any k of the Xs is the same as

$$\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \int_0^1 p^t (1-p)^{k-t} dp = \frac{t!(k-t)!}{(k+1)!},$$

   where $t = \sum_{i=1}^k x_i$.

   (b) Show that, marginally,

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \neq \prod_{j=1}^n \mathbb{P}(X_j = x_j).$$

   The distribution is exchangeable, but not i.i.d.

2. Let $U_i : 1, 2, \dots$ be i.i.d. $U(0,1)$ variables. (uniformly distributed on the interval $[0,1]$).

   (a) Consider a random sample of size n, $U_1, \dots, U_n$. Compute $p_{U_{1:n}}$, the density for the first order statistic.

(b) Suppose that X has the distribution

$$\mathbb{P}(X = x) = \frac{c}{x!} \qquad x = 1, 2, 3, \ldots$$

where $c = 1/(e-1)$. Find the distribution of

$$Z = \min\{U_1, \ldots, U_X\}.$$

Hint Condition on X.

3. Let $X_1, \ldots, X_n$ be a random sample from a population with p.d.f. (probability density function)

$$p_X(x) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x)$$

Let $X_{1:n} < \ldots < X_{1:n}$ be the *order statistics.*

(a) Show that the joint density of $(X_{1:n}, \ldots, X_{n:n})$ is:

$$p_{X_{1:n},\ldots,X_{n:n}}(x_1, \ldots, x_n) = \frac{n!}{\theta^n}\mathbf{1}(0 \leq x_1 \leq \ldots \leq x_n \leq \theta).$$

(b) Show that $\frac{X_{1:n}}{X_{n:n}}$ and $X_{n:n}$ are independent random variables.

4. Let $X_1, \ldots, X_n$ be a random sample from a population with p.d.f.

$$p_X(x) = \frac{a}{\theta^a}x^{a-1}\mathbf{1}_{[0,\theta]}(x).$$

Let $X_{1:n} < \ldots < X_{n:n}$ be the order statistics. Show that $\frac{X_{1:n}}{X_{2:n}}, \frac{X_{2:n}}{X_{3:n}}, \ldots, \frac{X_{n-1:n}}{X_{n:n}}$ and $X_{n:n}$ are mutually independent random variables. Find the distribution of each of them.

5. Let X be a random variable with continuous distribution function F. Find the distribution function of $F(X)$.

6. Let $U \sim \text{Unif}(0,1)$ (that is, U has uniform distribution over the interval $[0,1]$) and let F be a continuous, monotonically increasing cumulative distribution function. Show that the c.d.f. of $Y := F^{-1}(U)$ is F. Is the result true if F is the c.d.f. of a discrete random variable, which takes values in the set $(x_j)_{j\geq 1}$, $x_1 < x_2 < \ldots$?

7. Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with continuous c.d.f. $F(x) = \mathbb{P}(X \leq x)$ (that is $X_1, \ldots, X_n$ are independent identically distributed). Let

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{(-\infty,x]}(X_j).$$

$\widehat{F}_n$ is the *empirical distribution function.* Here $\mathbf{1}_A$ denotes the indicator function of a set A. That is,

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

(a) Show that $n\widehat{F}_n(x) \sim Bi(n, F(x))$.

(b) Compute $\mathrm{Cov}(\widehat{F}_n(x), \widehat{F}_n(y))$ (Cov denotes covariance).

(c) Let $D_n = \sup_{-\infty < x < +\infty} |\widehat{F}_n(x) - F(x)|$. Prove that the distribution of D_n is the same for all underlying continuous distribution functions F.

8. Let X and Y be independent random variables, with $\mathrm{Gamma}(\alpha, \lambda)$ and $\mathrm{Gamma}(\beta, \lambda)$ distributions respectively where $\alpha, \beta, \lambda > 0$. Show that $X + Y \sim \mathrm{Gamma}(\alpha + \beta, \lambda)$. The $\mathrm{Gamma}(\alpha, \lambda)$ distribution has density:

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbf{1}_{(0,+\infty)}(x).$$

9. Let $Z \sim N(0, 1)$ (standard normal). Find the distribution of Z^2. This is the $\chi^2(1)$ distribution.

10. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Find the distribution of:

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - \mu)^2.$$

11. Let $W_1, \ldots, W_n$ be i.i.d. $\chi^2(1)$ random variables, then $W_1 + \ldots + W_n \sim \chi^2(n)$. Let Z and Y be independent random variables with standard normal and $\chi^2(n)$ distributions respectively. Find the density function of:

$$T = \frac{Z}{\sqrt{Y/n}}.$$

Short Answers

1. (a) Since they're conditionally independent given $P = p$, it follows that, for a sequence $(x_1, \ldots, x_k)$ with t success and $k - t$ failure,

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) &= \int_0^1 \mathbb{P}(X_1 = x_1, \ldots, X_k = x_k | P = p) dp \\ &= \int_0^1 \prod_{j=1}^{k} \mathbb{P}(X_j = x_j | P = p) dp \\ &= \int_0^1 p^t (1-p)^{k-t} dp\end{aligned}$$

This is a standard integral, the beta integral. One way to compute it is by induction;

$$I(k,t) = \int_0^1 p^t (1-p)^{k-t} dp = \frac{k-t}{1+t} \int_0^1 p^{t+1} (1-p)^{k-t-1} dp = \frac{k-t}{1+t} I(k, t+1)$$

and

$$I(k,k) = \int_0^1 p^k dp = \frac{1}{k+1}$$

giving

$$I(k,t) = \frac{t!(k-t)!}{(k+1)!}.$$

(b)

$$\mathbb{P}(X_j = x_j) = \int_0^1 p^{x_j} (1-p)^{1-x_j} dp = \frac{1}{2} \qquad x_j = 0, 1$$

$$\prod_{j=1}^{n} \mathbb{P}(X_j = x_j) = \frac{1}{2^k} \qquad \{x_1, \ldots, x_k\} \in \{0,1\}^k$$

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{t!(k-t)!}{(k+1)!}$$

so they are not equal, hence the random variables are not i.i.d..

2. (a)

$$\mathbb{P}(U_{1:n} > x) = \mathbb{P}(U_1 > x, \ldots, U_n > x) = \mathbb{P}(U > x)^n = (1-x)^n \qquad 0 \le x \le 1$$

$$p_{U_{1:n}}(x) = n(1-x)^{n-1} \qquad 0 \le x \le 1$$

(b)

$$\mathbb{P}(Z > y) = \sum_{x=1}^{\infty} \mathbb{P}(U_{x:n} > y) \frac{c}{x!} = \sum_{x=1}^{\infty} \frac{c(1-y)^x}{x!} = c\left(e^{1-y} - 1\right) = \frac{e^{1-y} - 1}{e - 1} \qquad 0 \le y \le 1$$

$$F_Z(y) = 1 - \frac{e^{1-y} - 1}{e - 1} = \frac{e(1 - e^{-y})}{e - 1} \qquad 0 \le y \le 1.$$

3. (a) For any $A \subset \{0 < x_1 < \ldots < x_n \leq \theta\}$, let σ denote a permutation of $(1, \ldots, n)$ and let $A_\sigma = \{(x_1, \ldots, x_n) : (x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \in A\}$. By construction, the regions A_σ are disjoint.

$$\mathbb{P}((X_{1:n}, \ldots, X_{n:n}) \in A) = \sum_\sigma \mathbb{P}((X_1, \ldots, X_n) \in A_\sigma) = n!\mathbb{P}((X_1, \ldots, X_n) \in A)$$

so that, for all $A \subset \{0 < x_1 < \ldots < x_n \leq \theta\}$,

$$\mathbb{P}((X_{1:n}, \ldots, X_{n:n}) \in A) = \frac{n!}{\theta^n} \int \mathbf{1}_{\{0 \leq x_1 < \ldots < x_n \leq \theta\} \cap A} dx_1 \ldots dx_n$$

From this, it follows directly that the density is:

$$p_{X_{1:n}, \ldots, X_{1:n}}(x_1, \ldots, x_n) = \frac{n!}{\theta^n} \mathbf{1}(0 < x_1 < \ldots < x_n < \theta).$$

(b) Density of $(X_{1:n}, X_{n:n})$ computed as follows:

$$p_{X_{1:n}, X_{n:n}}(x_1, x_n) = \frac{n!}{\theta^n} \int_{x_1 < \ldots < x_n} \int dx_2 \ldots dx_{n-1} = \frac{n!}{\theta^n} \frac{1}{(n-2)!} (x_n - x_1)^{n-2}$$

(the integral is the area of the $n-2$ dimensional simplex) so

$$p_{X_{1:n}, X_{n:n}}(x_1, x_n) = \frac{n(n-1)}{\theta^n} (x_n - x_1)^{n-2} \qquad 0 < x_1 < x_n < \theta$$

Let $(Y, Z) = (\frac{X_{1:n}}{X_{n:n}}, X_{n:n})$ so that

$$(X_{1:n}, X_{n:n}) = (YZ, Z) \qquad Z = X_{n:n}.$$

Then

$$J_{(x_1, x_n) \to (y,z)} = \begin{pmatrix} z & 0 \\ y & 1 \end{pmatrix}$$

so the determinant is $|J_{(x_1, x_n) \to (y,z)}| = z$. It follows that

$$\begin{aligned} f_{(Y,Z)}(y, z) &= \frac{n(n-1)}{\theta^n} z(z - zy)^{n-2} \\ &= \frac{n(n-1)}{\theta^n} z^{n-1} (1-y)^{n-2} \qquad 0 < yz < z < \theta \Rightarrow 0 < y < 1, 0 < z < \theta \end{aligned}$$

Density is in product form, therefore $Y \perp Z$.

$$p_Z(z) = \frac{n}{\theta^n} z^{n-1} \qquad 0 \leq z \leq \theta$$

$$p_Y(y) = (n-1)(1-y)^{n-2} \qquad 0 \leq y \leq 1$$

4.

$$p_{X_{1:n}, \ldots, X_{n:n}}(x_1, \ldots, x_n) = \frac{n! a^n}{\theta^{na}} (x_1 \ldots x_n)^{a-1} \mathbf{1}(0 \leq x_1 \leq \ldots \leq x_n \leq \theta).$$

$$(Y_1, \ldots Y_{n-1}, Y_n) = \left(\frac{X_{1:n}}{X_{2:n}}, \ldots, \frac{X_{n:n}}{X_{n-1:n}}, X_{n:n} \right)$$

$$(x_1,\ldots,x_n)=(\prod_{j=1}^n y_j,\prod_{j=2}^n y_j,\ldots,y_n)$$

(here $x_k=\prod_{j=k}^n y_j$).

$$J_{(x_1,\ldots,x_n)\to(y_1,\ldots,y_n)}=\left(\begin{array}{ccc}\frac{\partial x_1}{\partial y_1} & \ldots & \frac{\partial x_1}{\partial y_n}\\ \vdots & & \vdots\\ \frac{\partial x_n}{\partial y_1} & \ldots & \frac{\partial x_n}{\partial y_n}\end{array}\right)=\left(\begin{array}{cccc}\prod_{j=2}^n y_j & \prod_{j\neq 2} y_j & \ldots & \prod_{j=1}^{n-1} y_j\\ 0 & \prod_{j=3}^n y_j & \ldots & \prod_{j=2}^{n-1} y_j\\ \vdots & & \vdots & \\ 0 & 0 & \ldots & 1\end{array}\right)$$

This matrix is upper triangular, with determinant $\prod_{j=2}^n y_j^{j-1}$.

Density for $Y_1,\ldots,Y_{n-1},Y_n$ is therefore

$$p_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n)=\left(\prod_{j=2}^n y_j^{j-1}\right)\frac{n!a^n}{\theta^{na}}\left(\prod_{j=1}^n y_j^{j(a-1)}\right)\qquad 0\leq\prod_{j=1}^n y_j\leq\prod_{j=2}^n y_j\leq y_{n-1}y_n\leq y_n\leq\theta$$

which reduces to

$$p_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n)=\prod_{j=1}^n f_{Y_j}(y_j)$$

where

$$\begin{aligned}p_{Y_1}(y_1)&=ay_1^{a-1}\qquad 0\leq y_1\leq 1\\ p_{Y_j}(y_j)&=jay_j^{ja-1}\qquad 0\leq y_j\leq 1\qquad 2\leq j\leq n-1.\\ p_{Y_n}(y_n)&=\frac{na}{\theta^{na}}y_n^{na-1}\qquad 0\leq y_n\leq\theta.\end{aligned}$$

5. The function F is continuous and non-decreasing. Define:

   $$F^{-1}(u)=\sup\{x:F(x)<u\}.$$

   Then $F(F^{-1}(u))=u$, so that:

   $$\mathbb{P}(F(X)\leq x)=\mathbb{P}(X\leq F^{-1}(x))=F(F^{-1}(x))=x\quad x\in[0,1]$$

   hence $F(X)\sim\text{Unif}(0,1)$.

6. $$\mathbb{P}(Y\leq x)=\mathbb{P}(F^{-1}(U)\leq x)=\mathbb{P}(U\leq F(x))=x.$$

   Yes: define

   $$F^{-1}(u)=\sup\{x:F(x)<u\}$$

   if $(x_j)_{j\geq 1}$ denotes the points where the probability mass function has positive mass, where $\mathbb{P}(X=x_j)=p_j$ where $0<x_1<x_2<\ldots$ and $\sum_j p_j=1$, then

   $$\left\{\begin{array}{ll}\mathbb{P}(F^{-1}(U)\leq x_1)=p_1 & \\ \mathbb{P}(F^{-1}(U)\leq x_{j+1})-\mathbb{P}(F^{-1}(U)\leq x_j)=p_{j+1} & j\geq 1.\end{array}\right.$$

7. (a) $n\widehat{F}_n(x) = \sum_{j=1}^n \mathbf{1}_{\{X_j \leq x\}}$. $n\widehat{F}_n(x)$ is therefore the number of 'success' in n Bernoulli trials each with success parameter $F(x)$, so $n\widehat{F}_n(x) \sim Bi(n, F(x))$. It follows that

$$\mathbb{E}\left[\widehat{F}_n(x)\right] = F(x) \qquad \mathrm{Var}\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n}$$

(b)

$$\begin{aligned}
\mathrm{Cov}(\widehat{F}_n(x), \widehat{F}_n(y)) &= \mathrm{Cov}(\frac{1}{n}\sum_{j=1}^n \mathbf{1}(X_j \leq x), \frac{1}{n}\sum_{k=1}^n \mathbf{1}(X_k \leq y)) \\
&= \frac{1}{n^2}\sum_{j,k} \mathrm{Cov}(\mathbf{1}(X_j \leq x), \mathbf{1}(X_k \leq y)) = \frac{1}{n^2}\sum_{j=1}^n \mathrm{Cov}(\mathbf{1}(X_j \leq x), \mathbf{1}(X_j \leq y)) \\
&= \frac{1}{n}\left(\mathbb{E}\left[\mathbf{1}(X_j \leq x)\mathbf{1}(X_j \leq y)\right] - F(x)F(y)\right) \\
&= \frac{1}{n}F(x)(1-F(y)) \qquad x < y
\end{aligned}$$

(c) First consider F *strictly* increasing. Then

$$D_n = \sup_{-\infty < y < +\infty} |\widehat{F}_n(y) - F(y)| = \sup_{0<x<1} |\widehat{F}_n(F^{-1}(x)) - F(F^{-1}(x))| = \sup_{0<x<1} |\widehat{F}_n(F^{-1}(x)) - x|.$$

$$\widehat{F}_n(F^{-1}(x)) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{X_j \leq F^{-1}(x)\}} = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{F(X_j) \leq x\}}$$

and the result follows since $F(X_1), \ldots, F(X_n)$ are i.i.d. $U(0,1)$ variables.

General case: use $F^{-1*}(x) = \sup\{y | F(x) < y\}$. Then $F^{-1*}(x) \leq z \Leftrightarrow x \leq F(z)$ and same proof follows.

Therefore, for any continous F, the distribution of D_n is the same as that of

$$\sup_{0 \leq y \leq 1} |\widehat{F}_n(y) - y|$$

where

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{U_j \leq y\}}$$

for a random sample $U_1, \ldots, U_n$ of $U(0,1)$ variables.

8. (8,9 and 10 form a sequence of questions, which follow on from each other). Let $W = X + Y$ then

$$F_W(w) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^w x^{\alpha-1}e^{-\lambda x}\left(\int_0^{w-x} y^{\beta-1}e^{-\lambda y}dy\right)dx$$

so that

$$\begin{aligned}
p_W(w) &= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda w} \int_0^w x^{\alpha-1}(w-x)^{\beta-1}dx \\
&= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda w} w^{\alpha+\beta-1} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \propto w^{\alpha+\beta-1}e^{-\lambda w}.
\end{aligned}$$

Since this is a density function, it follows (using the formula for a Gamma density function given in the question) that

$$p_W(w) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)} w^{\alpha+\beta-1} e^{-\lambda w} \mathbf{1}_{(0,+\infty)}(w).$$

9. For $x > 0$,

$$\mathbb{P}(Z^2 \leq x) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

so

$$p_{Z^2}(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2} \mathbf{1}_{(0,+\infty)}(x) \propto x^{-1/2} e^{-x/2} \mathbf{1}_{(0,+\infty)}(x); \qquad \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right).$$

Note: the Gamma$(\frac{1}{2}, \frac{1}{2})$ is also known as a $\chi^2(1)$ distribution.

10. Let $Z_k = \frac{X_k - \mu}{\sigma}$ then $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ and $\frac{1}{\sigma^2} \sum_{k=1}^n (X_k - \mu)^2 = \sum_{j=1}^n Z_j^2$. Using the results of 8 and 9, therefore, this has distribution Gamma $\left(\frac{n}{2}, \frac{1}{2}\right)$.

This is also known as a $\chi^2(n)$ distribution.

11.

$$\mathbb{P}(T \leq x) = \mathbb{P}\left(Z \leq x\sqrt{\frac{Y}{n}}\right) = \frac{1}{\sqrt{2\pi}} \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_0^\infty \int_{-\infty}^{x\sqrt{\frac{y}{n}}} e^{-z^2/2} y^{(n/2)-1} e^{-y/2} dz dy.$$

The density is:

$$\begin{aligned} p_T(x) &= \frac{1}{2^{(n+1)/2}\pi^{1/2}n^{1/2}\Gamma(n/2)} \int_0^\infty y^{(n-1)/2} \exp\left\{-\frac{y}{2}\left(1 + \frac{x^2}{n}\right)\right\} dy \\ &= \frac{(1+\frac{x^2}{n})^{-(n+1)/2}}{2^{(n+1)/2}\pi^{1/2}n^{1/2}\Gamma(n/2)} \int_0^\infty v^{(n+1)/2-1} e^{-v/2} dv \\ &= \frac{\Gamma(\frac{n+1}{2})}{n^{1/2}\pi^{1/2}\Gamma(\frac{n}{2})} \frac{1}{(1+\frac{x^2}{n})^{(n+1)/2}} \end{aligned}$$

Chapter 1

Statistical Models

1.1 Introduction

Studies and experiments, scientific or industrial, produce data, from which information is to be drawn. The particular angle of mathematical statistics is to view data as the outcome of a random experiment, which is described, at least approximately, by a *statistical model.*

Sampling Consider a population containing N individuals. Obtaining information on each individual may be a task that is either impossible or too expensive. A *random sample* of size m from the population is simply a subset of size m chosen at random. There are $\binom{N}{m}$ possible subsets; with random sampling from a finite population, each is chosen with probability $\frac{1}{\binom{N}{m}}$.

In many situations, the total population size N is considered to be so large that it may be considered as $+\infty$. When this happens, the observations $x_1, \ldots, x_m$ on m randomly chosen individuals are considered to be realisations of $X_1, \ldots, X_m$, m random variables.

Hypergeometric to Binomial Consider a population of size N, consisting of two types; N_1 of type 1 and N_2 of type 2, where $N = N_1 + N_2$. You select a subset of size m at random. Let X denote the number of type 1 elements in the sample. Then X has the so-called *hypergeometric distribution*

$$\mathbb{P}(X = k) = \frac{\binom{N_1}{k}\binom{N_2}{m-k}}{\binom{N}{m}} \qquad k = 0, 1, \ldots, m.$$

Now let $p = \frac{N_1}{N}$. Suppose that p is fixed; the proportion of type 1 elements in the population remains p, while $N \to +\infty$. Then it is a straightforward computation to show that

$$\mathbb{P}(X = k) \overset{N\to+\infty}{\longrightarrow} \binom{m}{k} p^k (1-p)^{m-k} \qquad k = 0, 1, \ldots, m.$$

If we consider $X = Y_1 + \ldots + Y_m$, where $Y_j = 1$ if an element of type 1 is chosen and 0 if an element of type 2 is chosen, then as $N \to +\infty$, $Y_1, \ldots, Y_m$ are m independent Bernoulli trials.

Notation Throughout, the aim is to use the convention: capital letters X denote random variables, while lower case letters x denote an observation on the random variable.

Example 1.1 (Two-Sample Model)**.**

Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ respectively be the responses of m subjects with a certain disease to a drug A and n subjects with the same condition being given drug B.

The natural assumptions are that $x_1, \ldots, x_m$ is an *observed random sample* from a distribution F, while $y_1, \ldots, y_n$ is an observed random sample from a distribution G.

Definition 1.1 (Random Sample)**.** *A* random sample $X_1, \ldots, X_n$ *is a collection of independent identically distributed (i.i.d.) random variables. An* observed random sample $x_1, \ldots, x_n$ *is a realisation, or outcome, of* $X_1, \ldots, X_n$*.*

That is, if $X = (X_1, \ldots, X_n)^t$ is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then $\underline{x} = (x_1, \ldots, x_n)^t = X(\omega)$ for some $\omega \in \Omega$.

From the data, one may try to establish whether there are differences in the performances of the two drugs; that is, whether or not $F = G$.

1.2 Parametrisations and Parameters

In its widest generality, a *statistical model* is simply a collection of probability distributions.

Definition 1.2 (Statistical Model)**.** *A* statistical model *is a family of probability distributions* $\mathcal{P}$*.*

The situation may be simplified if attention can be restricted to a *parametric family*; a collection of probability distributions described by a few parameters.

Definition 1.3 (Parametric Family / Regular Model)**.** *A* parametric family $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ *of probability distributions is a collection of distributions described by a parameter, or parameter vector* θ*, for all* $\theta \in \Theta$*. The space* Θ *of parameters is known as the* parameter space*. A parametric model is* regular *if either*

- *All the* $\mathbb{P}_\theta$ *are continuous with densities* $p(x; \theta)$ *and the set* $\mathcal{X} = \{x | p(x, \theta) > 0\}$ *is the same for each* $\theta \in \Theta$

or

- *All the* $\mathbb{P}_\theta$ *are discrete with probability mass functions* $p(x, \theta)$ *and the set* $\mathcal{X} = \{x | p(x, \theta) > 0\}$ *is the same for each* $\theta \in \Theta$*.*

Note The important point in this definition is that the *support* (the set where the density function for continuous, or mass function for discrete variables) is the same for each parameter value.

Furthermore, if $(X_1, \ldots, X_d)$ is a random vector, with state space $\mathcal{X} \subseteq \mathbb{R}^d$, where $X_1, \ldots, X_d$ are continuous variables, then the density $p_{X_1,\ldots,X_d}$ is uniquely defined, for example through

$$p_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

Example 1.2 (Example of a regular parametric family: Bernoulli(p))**.**

The family $\mathcal{P} = \{\mathbb{P}_p : p \in (0,1)\}$, where $\mathbb{P}_p(X = 1) = p$ and $\mathbb{P}_p(X = 0) = 1 - p$ is an example of a paremtric family. The parameter space $\Theta = (0,1)$ and support of the probability mass function is $\{0,1\}$ for each $p \in \Theta$.

However, if we take the parameter space $\Theta = [0,1]$ then the family $\{\mathbb{P}_p : p \in \Theta\}$ is not a regular parametric family, since for $p = 0$, $\mathbb{P}_0(X = 1) = 0$; the support is only $\{0\}$ and, for $p = 1$, $\mathbb{P}_1(X = 0) = 0$; the support is $\{1\}$. □

Example 1.3 (Example of a parametric family: Exp(λ))**.**

One example of a parametric family is the *exponential family* $\mathcal{P} = \{\mathbb{P}_\lambda : \lambda \in (0, +\infty)\}$ defined by the density function

$$p(x; \lambda) = \lambda e^{-\lambda x} \mathbf{1}_{[0,+\infty)}(x).$$

The space $\Theta = (0, +\infty)$ is the *parameter space*. This family is used if one has good reason to believe that the observations are from exponential variables (for example, if the random variable models lifetimes that satisfy the 'memoryless property'), but where the average value $\frac{1}{\lambda}$ is a priori unknown. □

There may be several ways of parametrising a distribution. In some cases, when multiple parameters are involved, the distribution may be *over parametrised*; each member of a subset of the parameters may give a the same distribution. It is useful to describe a parametric family in such a way that the parametrisation is *identifiable*.

Definition 1.4 (Identifiable Parametrisation)**.** *Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. The parametrisation θ is said to be* identifiable *if $\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$.*

Statistical models may be *non-parametric*, *semi-parametric* or *parametric*.

1. **Non-parametric** A model is non-parametric if no parametric assumptions are made. For example, with the two-sample model (Example 1.1), a *non-parametric* model is the assumption that $\{X_1, \ldots, X_m\}$ are a random sample from distribution F and $\{Y_1, \ldots, Y_n\}$ are a random sample from a distribution G, with no further assumptions about F and G.

2. **Semi-parametric** A *semi-parametric* model is a model that uses parameters, but these parameters do not give a full description. For example, with the two sample model (Example 1.1), the

variables could represent measurements of an enzyme level and there could be an assumption that drug B reduces the enzyme level by an amount θ more than drug A. Then if $\{X_1, \ldots, X_n\}$ is a random sample from a distribution with c.d.f. $F(x) = \mathbb{P}(X \leq x)$, it follows that Y is a random sample with c.d.f. $G(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(X \leq x + \theta) = F(x + \theta)$. If no further assumptions are made on F, then the model is *semi-parametric.*

3. **Parametric model** A *parametric model* is a model where the probability distributions are described fully by parametric families. For example, with the two sample model (Example 1.1), an example of a parametric model is: $\{X_1, \ldots, X_n\}$ is a $N(\mu_1, \sigma_1^2)$ random sample, while $\{Y_1, \ldots, Y_n\}$ is a $N(\mu_2, \sigma_2^2)$ random sample.

1.3 Statistics as Functions on the Sample Space

Definition 1.5 (Statistic)**.** *A* statistic T *is a map from the state space* $\mathcal{X}$ *of the random variable to a space of values* $\mathcal{T}$. *The mapping does* not *depend on the parametrisation.*

The space $\mathcal{T}$ is usually a subset of $\mathbb{R}^p$ for some p, but may be integer valued and is sometimes simply an indicator; $T : \mathcal{X} \to \{0, 1\}$.

Example 1.4 (Sample Average / Sample Mean)**.**

Let $X_1, \ldots, X_n$ be a random sample. The *sample average* is defined as:

$$T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{j=1}^{n} X_j = \overline{X}_n.$$

If the random sample is from a distribution with expectation μ and variance $\sigma^2 < +\infty$, then

$$\mathbb{E}\left[\overline{X}\right] = \mu$$

and

$$\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Recall that for a linear combination $a_1 X_1 + \ldots a_n X_n + b$,

$$\mathbb{E}[a_1 X_1 + \ldots a_n X_n + b] = a_1 \mathbb{E}[X_1] + \ldots + a_n \mathbb{E}[X_n] + b$$

and

$$\text{Var}(\sum_{j=1}^{n} a_j X_j + b) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j \text{Cov}(X_i, X_j).$$

Now set $a_1 = \ldots = a_n = \frac{1}{n}$ and $b = 0$. Use $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$.

It follows from Chebychev's inequality that:

$$\mathbb{P}\left(\left|\overline{X}-\mu\right|>\epsilon\right)\leq\frac{\sigma^2}{n\epsilon^2}\overset{n\to+\infty}{\longrightarrow}0.$$

The quantity $\overline{X}$, known as the *sample average*, is an *estimator* of μ, the *population average*. We have shown two of the properties that it satisfies:

1. **Unbiased** $\mathbb{E}\left[\overline{X}\right]=\mu$. That is, the average value of the estimator is equal to the quantity that it is estimating. An estimator $\widehat{\Theta}$ of a parameter θ which has the property that $\mathbb{E}_\theta\left[\widehat{\Theta}\right]=\theta$ is said to be *unbiased.*

2. **Consistency** The sequence $\overline{X}_n$ converges to μ in probability. A sequence of estimators $\widehat{\Theta}_n$ of a parameter (or parameter vector θ) that converges to θ in probability, i.e.

$$\mathbb{P}\left(\left|\widehat{\Theta}_n-\theta\right|>\epsilon\right)\overset{n\to+\infty}{\longrightarrow}0\qquad\forall\epsilon>0$$

is said to be *consistent.*

□

Example 1.5 (Sample Variance)**.**

Let $S:\mathbb{R}^n\to\mathbb{R}_+$ denote the function

$$S(x_1,\ldots,x_n)=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\overline{x})^2}.$$

Let $\{X_1,\ldots,X_n\}$ be a random sample. The *sample variance* is defined as:

$$S^2:=S^2(X_1,\ldots,X_n)=\frac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})^2.$$

If the parent distribution has expectation μ and variance $\sigma^2<+\infty$, then

$$\mathbb{E}[S^2]=\frac{1}{n-1}\sum_{j=1}^{n}\mathrm{Var}(X_j-\overline{X})=\frac{1}{n-1}\sum_{j=1}^{n}(\mathrm{Var}(X_j)+\mathrm{Var}(\overline{X})-2\mathrm{Cov}(X_j,\overline{X}))$$

where

$$\mathrm{Cov}(X_j,\overline{X})=\frac{1}{n}\sum_{k=1}^{n}\mathrm{Cov}(X_j,X_k)=\frac{1}{n}\mathrm{Var}(X_j)=\frac{\sigma^2}{n}$$

so that

$$\mathbb{E}[S^2]=\frac{n}{n-1}\left(\sigma^2+\frac{\sigma^2}{n}-\frac{2\sigma^2}{n}\right)=\sigma^2.$$

We have therefore shown that S^2 is an unbiased estimator of σ^2. To prove *consistency* of the sequence, a further assumption that $\mathbb{E}[X^4]<+\infty$ is needed; consistency under this condition is left as an exercise. □

Definition 1.6 (Order Statistics)**.** *Let $X_1, \ldots, X_n$ be a random sample, where each X_j has state space $\mathcal{X} \subseteq \mathbb{R}$. The* order statistic *is defined as*

$$T(X_1, \ldots, X_n) = (X_{1:n}, \ldots, X_{1:n})$$

where $(X_{1:n}, \ldots, X_{n:n}) = (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ for a permutation such that $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$. This may also be written as: $X_{1:n}, \ldots, X_{n:n}$.

1.4 The Empirical Distribution Function

Another example of a statistic is the *empirical distribution function.* Let $\widehat{F} : \mathbb{R}^{n+1} \to [0,1]$ be defined as:

$$\widehat{F}_{x_1,\ldots,x_n}(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{(-\infty,x]}(x_j).$$

Let $X_1, \ldots, X_n$ be a random sample. Then $\widehat{F}_{X_1,\ldots,X_n}(x)$, abbreviated to $\widehat{F}_n(x)$, is the *empirical distribution function* evaluated at x. This function estimates the distribution function $F_X(x) = \mathbb{P}(X \leq x)$ where F_X is the c.d.f. of the parent variable of the random sample.

$$\mathbb{E}\left[\widehat{F}_n(x)\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\mathbf{1}_{(-\infty,x]}(X_j)\right] = \mathbb{P}(X \leq x).$$

From the construction, it is clear that

$$n\widehat{F}_n(x) \sim \text{Binomial}(n, \mathbb{P}(X \leq x))$$

and therefore

$$\text{Var}\left(\widehat{F}_n(x)\right) = \frac{1}{n}F_X(x)(1 - F_X(x)).$$

□

Given a random sample $X_1, \ldots, X_n$ with parent distribution F, the *Kolmogorov-Smirnov statistic* D_n is defined by:

$$D_n := \sup_{-\infty < x < +\infty} \left|\widehat{F}_n(x) - F(x)\right|. \tag{1.1}$$

The following result is proved in tutorial Exercises 1.4 and 1.4.

Theorem 1.7 (Glivenko - Cantelli)**.**

$$D_n \xrightarrow{P} 0.$$

Proof See Tutorial 1 Exercises 1.4 and 1.4. □

Theorem 1.8 below is due to Kolmogorov [5](1933) and Theorem 1.9 due to Smirnov [9] (1939).

Theorem 1.8. *Suppose that F is continuous and define D_n by Equation* (1.1)*. Then for all $z \geq 0$,*

$$\mathbb{P}\left(n^{1/2}D_n \leq z\right) \rightarrow L(z)$$

where $L(z)$ is the cumulative distribution function given by

$$L(z) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-j^2z^2} = \frac{(2\pi)^{1/2}}{z}\sum_{j=1}^{\infty}e^{-(2j-1)^2\pi^2/8z^2}. \tag{1.2}$$

For $z \leq 0$, $L(z) = 0$.

Proof Omitted. □(OMITTED FROM THE LECTURE AND THE COURSE) There are several proofs of this. The proof given here follows Feller [3](1948). It may be of interest to the enthusiast. The style of proof was described to me as 'proof by intimidation'.

Let $x_1, \ldots, x_{n-1}$ denote the numbers such that

$$F(x_k) = \frac{k}{n} \qquad k = 1, 2, \ldots, n-1$$

This definition is unique except when $F(x) = \frac{k}{n}$ within an entire interval, in which case let x_k denote the left endpoint of the interval.

Now let $X_1, \ldots, X_n$ be i.i.d. with distribution F. Consider $(Y_1, \ldots, Y_n)$ defined by:

$$\begin{cases} Y_1 = \sum_{j=1}^{n}\mathbf{1}_{(-\infty,x_1)}(X_j) \\ Y_k = \sum_{j=1}^{n}\mathbf{1}_{[x_{k-1},x_k)}(X_j) \qquad k = 2, \ldots, n-1 \\ Y_n = \sum_{j=1}^{n}\mathbf{1}_{[x_{n-1},\infty)}(X_j). \end{cases}$$

Then

$$(Y_1, \ldots, Y_n) \sim \text{Mult}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

Let $c \in \mathbb{Z}$. Suppose that for some particular x

$$\widehat{F}_n(x) - F(x) > \frac{c}{n}. \tag{1.3}$$

The point x is contained in a *maximal* interval in which (1.3) holds. At the right endpoint ξ of the interval,

$$\widehat{F}_n(\xi) - F(\xi) = \frac{c}{n}. \tag{1.4}$$

The point $\widehat{F}_n(\xi)$ is necessarily a number of the form $\frac{r}{n}$ for some positive integer r. Since c is an integer, it follows that $F(\xi) = \frac{k}{n}$ and hence $\xi = x_k$ for some k. From (1.4) it follows that

$$X_{(k+c)} < x_k \qquad X_{(k+c+1)} \geq x_k \tag{1.5}$$

In other words, exactly $c+k$ of the n variables $(X_j)_{j=1}^n$ are smaller than x_k. Let

$$A_k(c) = \{X_{(k+c)} < x_k\} \cap \{X_{(k+c+1)} \geq x_k\}.$$

The inequality (1.3) holds for some x if, and only if, at least one of the events $A_1(c), \ldots, A_n(c)$ occurs. The argument applies equally to $c < 0$ and shows that the event $D_n > \frac{c}{n}$ occurs if and only if at least one among the events

$$A_1(c), A_1(-c), A_2(c), A_2(-c), \ldots, A_n(c), A_n(-c) \tag{1.6}$$

occurs.

In terms of $Y_1, \ldots, Y_n$, $A_k(c) = \{Y_1 + \ldots + Y_k = c + k\}$.
Let U_r and V_r be the events that in the sequence (1.6) the first event to occur are $A_r(c)$ or $A_r(-c)$ respectively. More formally (using $\overline{A}$ to denote the complement of an event A),

$$\left\{ \begin{array}{l} U_r = \overline{A}_1(c)\overline{A}_1(-c)\ldots\overline{A}_{r-1}(c)\overline{A}_{r-1}(c)A_r(c) \\ V_r = \overline{A}_1(c)\overline{A}_1(-c)\ldots\overline{A}_{r-1}(c)\overline{A}_{r-1}(c)\overline{A}_r(c)A_r(-c) \end{array} \right. \tag{1.7}$$

These events are mutually exclusive and therefore

$$\mathbb{P}\left(D_n > \frac{c}{n}\right) = \sum_{r=1}^n \mathbb{P}(U_r) + \sum_{r=1}^n \mathbb{P}(V_r). \tag{1.8}$$

From the definitions, the following two fundamental relations follow:

$$\left\{ \begin{array}{l} \mathbb{P}(A_k(c)) = \sum_{r=1}^k \mathbb{P}(U_r)\mathbb{P}(A_k(c)|A_r(c)) + \sum_{r=1}^k \mathbb{P}(V_k)\mathbb{P}(A_k(c)|A_r(-c)) \\ \mathbb{P}(A_k(-c)) = \sum_{r=1}^k \mathbb{P}(U_r)\mathbb{P}(A_k(-c)|A_r(c)) + \sum_{r=1}^k \mathbb{P}(V_r)\mathbb{P}(A_k(-c)|A_r(-c)). \end{array} \right. \tag{1.9}$$

The fact that $\mathbb{P}(A_k(c)|U_r) = \mathbb{P}(A_k(c)|A_r(c))$ and $\mathbb{P}(A_k(c)|V_r) = \mathbb{P}(A_k(c)|A_r(-c))$ follows from considering the events in terms of the multinomial distribution of $(Y_1, \ldots, Y_n)$. This is left as an exercise.

This is a system of $2n$ linear equations for the $2n$ unknowns $\mathbb{P}(U_r)$ and $\mathbb{P}(V_r)$. It may be solved by the method of *generating functions*.

By definition of x_k, it follows that $\mathbb{P}(X_j < x_k) = \frac{k}{n}$ for each j. The probability of the event $A_k(c)$ (that exactly the same inequality holds for exactly $k+c$ different $X_1, \ldots, X_n$) is therefore

$$\mathbb{P}(A_k(c)) = \binom{n}{k+c}\left(\frac{k}{n}\right)^{k+c}\left(1 - \frac{k}{n}\right)^{n-(k+c)}. \tag{1.10}$$

Similarly,

$$\begin{aligned} \mathbb{P}(A_k(c)|A_r(c)) &= \frac{\mathbb{P}(A_k(c) \cap A_r(c))}{\mathbb{P}(A_r(c))} = \frac{\mathbb{P}(Y_1 + \ldots + Y_r = c + r, Y_{r+1} + \ldots + Y_k = k - r)}{\mathbb{P}(Y_1 + \ldots + Y_r = c + r)} \\ &= \frac{\frac{n!}{(c+r)!(k-r)!(n-(c+k))!}\left(\frac{r}{n}\right)^{c+r}\left(\frac{k-r}{n}\right)^{k-r}\left(\frac{n-k}{n}\right)^{n-(c+k)}}{\frac{n!}{(c+r)!(n-(c+r))!}\left(\frac{r}{n}\right)^{c+r}\left(\frac{n-r}{n}\right)^{n-(c+r)}} \\ &= \binom{n-r-c}{k-r}\left(\frac{k-r}{n-r}\right)^{k-r}\left(1 - \frac{k-r}{n-r}\right)^{n-k-c} \end{aligned} \tag{1.11}$$

and

$$\mathbb{P}(A_k(c)|A_r(-c)) = \binom{n-r+c}{k-r+2c}\left(\frac{k-r}{n-r}\right)^{k-r+2c}\left(1-\frac{k-r}{n-r}\right)^{n-k-c} \tag{1.12}$$

The last three equations also hold for $c < 0$. They can be written in a more convenient form in terms of the quantities:

$$p_k(c) = e^{-k}\frac{k^{k+c}}{(k+c)!} \tag{1.13}$$

Note that

$$p_n(0) = e^{-n}\frac{n^n}{n!} \tag{1.14}$$

By Stirling's formula,

$$p_n(0) \simeq e^{-n}n^n(2\pi)^{-1/2}n^{-n-\frac{1}{2}}e^n \Rightarrow (2\pi n)^{1/2}p_n(0) \overset{n\to+\infty}{\longrightarrow} 1.$$

From the formula for $p_n(0)$ (1.14),

$$\mathbb{P}(A_k(c)) = \frac{n!}{(n-k-c)!(k+c)!}\frac{k^{k+c}(n-k)^{n-k-c}}{n^n} = \frac{p_k(c)p_{n-k}(-c)}{p_n(0)} \tag{1.15}$$

$$\mathbb{P}(A_k(c)|A_r(c)) = \frac{(n-r-c)!}{(k-r)!(n-k-c)!}\frac{(k-r)^{k-r}(n-k)^{n-k-c}}{(n-r)^{n-r-c}} = \frac{p_{k-r}(0)p_{n-k}(-c)}{p_{n-r}(-c)} \tag{1.16}$$

$$\mathbb{P}(A_k(c)|A_r(-c)) = \frac{(n-r+c)!}{(k-r+2c)!(n-k-c)!}\frac{(k-r)^{k-r+2c}(n-k)^{n-k-c}}{(n-r)^{n-r+c}} = \frac{p_{k-r}(2c)p_{n-k}(-c)}{p_{n-r}(c)} \tag{1.17}$$

Putting these into equation (1.9) gives:

$$\frac{p_k(c)p_{n-k}(-c)}{p_n(0)} = \sum_{r=1}^{k}\mathbb{P}(U_r)\frac{p_{k-r}(0)p_{n-k}(-c)}{p_{n-r}(-c)} + \sum_{r=1}^{k}\mathbb{P}(V_r)\frac{p_{k-r}(2c)p_{n-k}(-c)}{p_{n-r}(c)}$$

and similarly for the other equation in (1.9). The second factor in the numerator of each term cancels. A further simplification is achieved by introducing the new unknowns

$$u_r = \mathbb{P}(U_r)\frac{p_n(0)}{p_{n-r}(-c)} \qquad v_r = \mathbb{P}(V_r)\frac{p_n(0)}{p_{n-r}(c)}. \tag{1.18}$$

The fundamental relations (1.9) may then be written as:

$$\begin{cases} p_k(c) = \sum_{r=1}^{k} u_r p_{k-r}(0) + \sum_{r=1}^{k} v_r p_{k-r}(2c) \\ p_k(-c) = \sum_{r=1}^{k} u_r p_{k-r}(-2c) + \sum_{r=1}^{k} v_r p_{k-r}(0). \end{cases} \tag{1.19}$$

This is a system of convolution type and can therefore be solved by means of generating functions. Set

$$u(\lambda) = \sum_{k=1}^{\infty} u_k\lambda^k \qquad v(\lambda) = \sum_{k=1}^{\infty} v_k\lambda^k \tag{1.20}$$

and

$$p(\lambda; c) = \frac{1}{n^{1/2}}\sum_{k=1}^{\infty} p_k(c)\lambda^k. \tag{1.21}$$

Then it follows that

$$p(\lambda;c) = \frac{1}{n^{1/2}}\sum_{k=1}^{\infty}\sum_{r=1}^{k} u_r\lambda^r p_{k-r}(0)\lambda^{k-r} + \frac{1}{n^{1/2}}\sum_{k=1}^{\infty}\sum_{r=1}^{k} v_r\lambda^r p_{k-r}(0)\lambda^{k-r}$$

so that

$$\begin{cases} p(\lambda;c) = u(\lambda)p(\lambda;0) + v(\lambda)p(\lambda;2c) \\ p(\lambda;-c) = u(\lambda)p(\lambda;-2c) + v(\lambda)p(\lambda;0). \end{cases} \tag{1.22}$$

From this, $u(\lambda)$ and $v(\lambda)$ may be obtained. Equation (1.18) then determines $\mathbb{P}(U_r)$ and $\mathbb{P}(V_r)$. We are only interested in the two sums in (1.8). Set

$$\xi_k = \frac{1}{p_n(0)}\sum_{r=1}^{k} p_{k-r}(-c)u_r \qquad \eta_k = \frac{1}{p_n(0)}\sum_{r=1}^{k} p_{k-r}(c)v_r \tag{1.23}$$

From (1.18),

$$\sum_{r=1}^{n}\mathbb{P}(U_r) = \xi_n \qquad \sum_{r=1}^{n}\mathbb{P}(V_r) = \eta_n \tag{1.24}$$

and therefore, by (1.8),

$$\mathbb{P}\left(D_n > \frac{c}{n}\right) = \xi_n + \eta_n. \tag{1.25}$$

From (1.23),

$$\begin{cases} \xi(\lambda) := \sum_{k=1}^{\infty}\xi_k\lambda^k = \frac{u(\lambda)p(\lambda;-c)n^{1/2}}{p_n(0)} \\ \eta(\lambda) := \sum_{k=1}^{\infty}\eta_k\lambda^k = \frac{v(\lambda)p(\lambda;c)n^{1/2}}{p_n(0)}. \end{cases} \tag{1.26}$$

Now let $n \to +\infty$ and $c \to +\infty$ in accordance with (1.27):

$$c = zn^{1/2} \qquad n \to +\infty \tag{1.27}$$

Then, by Stirling's formula,

$$\begin{aligned} n^{1/2}p_{nt}(zn^{1/2}) &\simeq n^{1/2}e^{-nt}\frac{(nt)^{(nt)+(zn^{1/2})}}{(2\pi)^{1/2}((nt)+(zn^{1/2}))^{(nt)+(zn^{1/2})+\frac{1}{2}}}e^{(nt)+(zn^{1/2})} \\ &= e^{zn^{1/2}}(2\pi t)^{-1/2}\left(1+\frac{z}{tn^{1/2}}\right)^{-1/2}\left(1+\frac{z}{n^{1/2}t}\right)^{-nt-zn^{1/2}} \\ &\overset{n\to+\infty}{\longrightarrow} (2\pi t)^{-1/2}\exp\left\{-\frac{z^2}{2t}\right\} \end{aligned} \tag{1.28}$$

It follows that

$$\begin{aligned} p(e^{-s/n};zn^{1/2}) &= \frac{1}{n^{1/2}}\sum_{k=1}^{\infty}e^{-sk/n}p_k(zn^{1/2}) \\ &\overset{n\to+\infty}{\longrightarrow} \frac{1}{(2\pi)^{1/2}}\int_0^{\infty}\frac{1}{t^{1/2}}\exp\left\{-ts-\frac{z^2}{2t}\right\}dt \\ &= \frac{1}{(2s)^{1/2}}\exp\left\{-(2sz^2)^{1/2}\right\} \end{aligned} \tag{1.29}$$

The limiting form is the same for $p(\lambda;c)$ and $p(\lambda;-c)$. Let $u^* = \lim_{n\to+\infty}u(e^{-s/n})$ and $v^* = \lim_{n\to+\infty}v(e^{-s/n})$. It follows directly from (1.22) that

$$\begin{cases} \frac{1}{(2s)^{1/2}}e^{-(2sz^2)^{1/2}} = u^* \frac{1}{(2s)^{1/2}} + v^* \frac{1}{(2s)^{1/2}}e^{-(8sz^2)^{1/2}} \\ \frac{1}{(2s)^{1/2}}e^{-(2sz^2)^{1/2}} = u^* \frac{1}{(2s)^{1/2}}e^{-(8sz^2)^{1/2}} + v^* \frac{1}{(2s)^{1/2}} \end{cases}$$

from which

$$u^* = v^* = \frac{\exp\left\{-(2sz^2)^{1/2}\right\}}{1+\exp\left\{-(8sz^2)^{1/2}\right\}}. \tag{1.30}$$

Using this, together with the fact that $(2\pi n)^{1/2}p_n(0) \to 1$, it follows from (1.26) that

$$\lim_{n\to+\infty} \frac{1}{n}\xi(e^{-s/n}) = \lim_{n\to+\infty} \frac{1}{n}\eta(e^{-s/n}) = \left(\frac{2\pi}{2s}\right)^{1/2} \frac{\exp\left\{-(8sz^2)^{1/2}\right\}}{1+\exp\left\{-(8sz^2)^{1/2}\right\}} =: \phi(s). \tag{1.31}$$

Expanding $\phi(s)$ a geometric series gives:

$$\phi(s) = \left(\frac{2\pi}{2s}\right)^{1/2} \sum_{j=1}^{\infty}(-1)^{j-1}\exp\left\{-(8sj^2z^2)^{1/2}\right\}. \tag{1.32}$$

From the integral in (1.29), it follows that $\phi(s)$ is the Laplace transform of

$$f(t) = \sum_{j=1}^{\infty}(-1)^{j-1}\frac{1}{t^{1/2}}\exp\left\{-\frac{2j^2z^2}{t}\right\}. \tag{1.33}$$

It follows from (1.31) that

$$\phi(s) = \lim_{n\to+\infty}\sum_{k=1}^{\infty}\xi_{(k/n)n}e^{-sk/n}.$$

Using $\frac{k}{n} \to t$ that

$$\lim_{n\to+\infty}\xi_n = \lim_{n\to+\infty}\eta_n = f(1) \tag{1.34}$$

which, by (1.25), completes the proof. □

Theorem 1.9. *Let $(X_1,\dots,X_m)$ and $(Y_1,\dots,Y_n)$ be two random samples of mutually independent random variables having the same common distribution function. Let $\widehat{F}_m$ and $\widehat{G}_n$ be the corresponding empirical distribution functions and define the random variable $D_{m,n}$ by:*

$$D_{m,n} = \sup_{-\infty<x<+\infty}|\widehat{F}_m(x) - \widehat{G}_n(x)|.$$

Set $N = \frac{mn}{m+n}$. Suppose that $m \to +\infty$ and $n \to +\infty$ in such a way that $\frac{m}{n} \to a$ where a is a constant. Then for all $z \geq 0$,

$$\mathbb{P}(N^{1/2}D_{m,n} \leq z) \to L(z)$$

where L is defined by Equation (1.2).

Proof Omitted. □

Lecture 1: Summary

- Definition of random sample (finite and infinite populations)
- Definition of statistical model, regular statistical model, identifiable parametrisation.
- Definition of a Statistic. Basic examples: sample average, sample variance, order statistics. Definition of unbiased, definition of consistency for estimators.
- Empirical distribution function, Glivenko-Cantelli lemma, distribution of Kolmogorov-Smirnov statistic.

Tutorial 2

Identities for Estimating Moments

1. Let $X_1, \ldots, X_n$ be a random sample, with sample average $\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$ and sample variance $S^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2$. Show that
$$S^2 = \frac{1}{2n(n-1)}\sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2$$
You may use:
$$\sum_{j=1}^n y_j = \frac{1}{2n}\sum_{j,k=1}^n (y_j + y_k)$$
and $x^2 + y^2 = (x-y)^2 + 2xy$.

2. Assume that $\mathbb{E}[X_i^4] < +\infty$ and set $\theta_1 = \mathbb{E}[X_i]$, $\theta_j = \mathbb{E}[(X_i - \theta_1)^j]$ for $j = 2, 3, 4$. Let $Y_j = X_j - \theta_1$, $\overline{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$ and $\overline{Y^2} = \frac{1}{n}\sum_{j=1}^n Y_j^2$.
   (a) Compute $\mathbb{E}[\overline{Y}^4]$ and $\mathbb{E}[\overline{Y^2}^2]$ and $\mathbb{E}[\overline{Y^2}\,\overline{Y}^2]$ in terms of θ_1, θ_2, θ_3 and θ_4.
   (b) Show that
   $$\mathrm{Var}(S^2) = \frac{1}{n}\left(\theta_4 - \frac{n-3}{n-1}\theta_2^2\right).$$
   (c) Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population.
      i. Find expressions for $\theta_1, \theta_2, \theta_3, \theta_4$ in terms of μ and σ^2.
      ii. Hence compute $\mathrm{Var}(S^2)$ for a $N(\mu, \sigma^2)$ random sample.

3. Establish the following recursion relations for means and variances. Let $\overline{X}_n$ and S_n^2 be the mean and variance respectively of $X_1, \ldots, X_n$. Suppose another observation X_{n+1} becomes available. Show that
   (a)
   $$\overline{X}_{n+1} = \frac{X_{n+1} + n\overline{X}_n}{n+1}$$
   (b)
   $$nS_{n+1}^2 = (n-1)S_n^2 + \left(\frac{n}{n+1}\right)(X_{n+1} - \overline{X}_n)^2.$$

Parametric Families: Identifiability Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a family of probability distributions. The parametrisation θ is said to be *identifiable* if $\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$. For example, let $\theta = (\mu, \sigma^2)$ and $\mathbb{P}_\theta$ denote the $N(\mu, \sigma^2)$ distribution. The parameterisation is *identifiable* since
$$(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2) \Rightarrow \exists A \in \mathcal{B}(\mathbb{R}) : \mathbb{P}_{\theta_1}(A) \neq \mathbb{P}_{\theta_2}(A)$$
where $\mathcal{B}(\mathbb{R})$ denotes the Borel subsets of $\mathbb{R}$.

On the other hand, the parametrisation $\theta = (\mu, \nu, \sigma^2)$ where $\mathbb{P}_\theta$ is $N(\mu - \nu, \sigma^2)$ is not identifiable, since $\theta_1 = (\mu, \nu, \theta)$ and $\theta_2 = (\mu + a, \nu + a, \theta)$ give the same distribution.

4. (a) Let $X_{ij} : i = 1, \dots, p; j = 1, \dots, b$ be independent with $X_{ij} \sim N(\mu_{ij}, \sigma^2)$. Let $\mu_{ij} = \nu + \alpha_i + \beta_j$. Let $\theta = (\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_b, \nu, \sigma^2)$ and $\mathbb{P}_\theta$ the distribution of $X_{11}, \dots, X_{pb}$. Is the parametrisation identifiable? Prove or disprove.

   (b) Now suppose that $(\alpha_1, \dots, \alpha_p)$ and $(\beta_1, \dots, \beta_b)$ are restricted to the sets $\sum_{i=1}^{p} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$. Is the parametrisation identifiable? Prove or disprove.

5. A measuring instrument is being used to obtain n independent determinations of a physical constant μ. Suppose that the measuring instrument is known to be biased by a positive constant θ units, where θ is unknown and that the errors are otherwise identically distributed normal random variables with known variance σ^2. Is the parametrisation identifiable? Prove or disprove.

6. The number of eggs laid by an insect follows a Poisson distribution with unknown mean μ. Once laid, each egg has an unknown chance p of hatching, independently of the others. An entomologist studies a set of n such insects, observing only the number of eggs hatching for each nest. Is the parametrisation identifiable?

Hazard and Survival

7. Let $T_1, \dots, T_m$ and $T'_1, \dots, T'_n$ be random samples with parent variables T and T' respectively, which are the survival times of two groups of patients receiving treatments A and B respectively. The *group survival* for the two groups is defined as $X = \min_{j=1,\dots,m} T_j$ and $Y = \min_{j=1,\dots,n} T'_j$ respectively. Let $S_X(t) = \mathbb{P}(X > t)$ and $S_Y(t) = \mathbb{P}(Y > t)$ denote the *group survival functions.* Assume that the groups are independent of each other and that T and T' have the same distribution.

   (a) Show that $S_Y(t) = S_X^{n/m}(t)$.

   (b) Extending from rationals to $\delta \in (0, +\infty)$ gives the *Lehmann model*: $S_Y(t) = S_X^{\delta}(t)$. Equivalently, $S_Y(t) = S_0^{n\delta}(t)$ and $S_X(t) = S_0^{m\delta}(t)$ for some survival function S_0. Suppose that X is a non negative continuous random variable with survival function $S_X(t) = S_0^{m\delta}(t)$. Compute the distribution function of $X' := -\log S_0(X)$.

   (c) Suppose that T and Y are two non-negative continuous random variables with survival functions $S_T(t)$ and $S_Y(t)$ respectively and densities $f_T(t)$ and $f_Y(t)$ respectively. Their *hazard functions* are defined as $\alpha_T(t) = \frac{f_T(t)}{S_T(t)}$ and $\alpha_Y(t) = \frac{f_Y(t)}{S_Y(t)}$ respectively. Show that $\alpha_Y = c\alpha_T$ if and only if $S_Y = S_T^c$. Such a model is known as the *Cox proportional hazard model.*

Order Statistics and Glivenko-Cantelli Lemma

8. Let $X_1, \dots, X_n$ be i.i.d. random variables, with c.d.f. F and density f. The ordered vector $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ which is an ordering of $X_1, \dots, X_n$ from lowest to highest is the vector of *order statistics.*

   (a) Find the c.d.f. and density of $X_{k:n}$.

(b) Hence, if $X_1, \dots, X_n$ be a random sample from a $U(0,1)$ distribution (uniform on the interval $(0,1)$), show that the density function for the jth order statistic $X_{j:n}$ is

$$f_{X_{j:n}}(x) = j\binom{n}{j}x^{j-1}(1-x)^{n-j} \qquad x \in [0,1]$$

(c) Hence prove (again for a $U(0,1)$ random sample) that for positive integer p,

$$\mathbb{E}\left[X_{j:n}^p\right] = j\binom{n}{j}\frac{\Gamma(j+p)\Gamma(n-j+1)}{\Gamma(n+p+1)}.$$

You may assume the Beta integral:

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

9. Let F be a continuous cumulative distribution function, $X_1, \dots, X_n$ a random sample generated from F and $\widehat{F}_n$ the empirical distribution function. Let $D_n = \sup_{-\infty<x<+\infty} |F(x) - \widehat{F}_n(x)|$. Prove that for any $\epsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}\left(\sup_{-\infty<x<+\infty} |F(x) - \widehat{F}_n(x)| > \epsilon\right) = 0.$$

You may use the result from the previous tutorial that the distribution of D_n does not depend on the underlying F (and hence assume that the random sample is $U(0,1)$).

Short Answers

1.

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})^2 \\
&= \frac{1}{2n(n-1)}\sum_{j,k=1}^{n}\{((X_j-\overline{X})^2+(X_k-\overline{X})^2)\} \\
&= \frac{1}{2n(n-1)}\sum_{j,k=1}^{n}\{(X_j-X_k)^2+2(X_j-\overline{X})(X_k-\overline{X})\} \\
&= \frac{1}{2n(n-1)}\sum_{j,k=1}^{n}(X_j-X_k)^2
\end{aligned}$$

because $\sum_j(X_j-\overline{X})=0$.

2. (a)

$$\mathbb{E}[\overline{Y}^4]=\frac{1}{n^4}\sum_{j_1,j_2,j_3,j_4=1}^{n}\mathbb{E}[Y_{j_1}Y_{j_2}Y_{j_3}Y_{j_4}]=\frac{1}{n^3}\theta_4+3\left(\frac{n-1}{n^3}\right)\theta_2^2$$

$$\mathbb{E}[\overline{Y^2}^2]=\frac{1}{n^2}\sum_{j_1,j_2=1}^{n}\mathbb{E}[Y_{j_1}^2Y_{j_2}^2]=\frac{1}{n}\theta_4+\frac{n-1}{n}\theta_2^2.$$

$$\mathbb{E}[\overline{Y^2}\,\overline{Y}^2]=\frac{1}{n^3}\sum_{j_1,j_2,j_3}\mathbb{E}[Y_{j_1}^2Y_{j_2}Y_{j_3}]=\frac{1}{n^2}\theta_4+\frac{n-1}{n^2}\theta_2^2$$

(b) $\mathbb{E}\left[\overline{Y}^2\right]=\frac{\theta_2}{n}$ and $\mathbb{E}\left[\overline{Y^2}\right]=\theta_2$. For $j\neq k$, $\mathbb{E}[(Y_j-Y_k)^2]=2\theta_2$, so

$$\begin{aligned}
\mathrm{Var}(S^2) &= \frac{1}{4n^2(n-1)^2}\mathrm{Var}\left(\sum_{j,k}(Y_j-Y_k)^2\right) \\
&= \frac{n^2}{(n-1)^2}\mathrm{Var}\left(\overline{Y^2}-\overline{Y}^2\right) \\
&= \frac{n^2}{(n-1)^2}(\mathbb{E}\left[\overline{Y^2}^2+\overline{Y}^4-2\overline{Y^2}\,\overline{Y}^2\right]-\mathbb{E}[\overline{Y^2}]^2-\mathbb{E}[\overline{Y}^2]^2+2\mathbb{E}[\overline{Y^2}\mathbb{E}[\overline{Y}^2]) \\
&= \frac{n^2}{(n-1)^2}\left(\left(\frac{1}{n}\theta_4+\frac{n-1}{n}\theta_2^2\right)+\left(\frac{1}{n^3}\theta_4+3\left(\frac{n-1}{n^3}\right)\theta_2^2\right)\right. \\
&\quad \left.-2\left(\frac{1}{n^2}\theta_4+\frac{n-1}{n^2}\theta_2^2\right)-2\left(1-\frac{1}{n}\right)^2\theta_2^2+\left(1-\frac{1}{n}\right)^2\theta_2^2\right) \\
&= \frac{1}{n}\left(\theta_4-\frac{n-3}{n-1}\,\theta_2^2\right)
\end{aligned}$$

(c) i. $\theta_1=\mu$, $\theta_2=\sigma^2$, $\theta_3=0$, $\theta_4=3\sigma^4$. The only one that may cause problems is the last one:

$$\theta_4 = \int y^4 \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} dy = 2\int_0^\infty y^4 \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} dy$$

substitute (for example) $x = \frac{y^2}{2\sigma^2}$ $dx = \frac{ydy}{\sigma^2}$

$$\theta_4 = \frac{4\sigma^4}{\sqrt{\pi}} \int_0^\infty z^{3/2} e^{-z} dz = \frac{4\sigma^4 \Gamma(5/2)}{\sqrt{\pi}} = 3\sigma^4$$

ii.

$$\mathrm{Var}(S^2) = \frac{1}{n}\left(3 - \frac{n-3}{n-1}\right)\sigma^4 = \frac{2}{n-1}\sigma^4.$$

3. (a)

$$\overline{X}_{n+1} = \frac{1}{n+1}\sum_{j=1}^{n+1} X_j = \frac{1}{n+1}\sum_{j=1}^{n} X_j + \frac{1}{n+1}X_{n+1} = \frac{n}{n+1}\overline{X}_n + \frac{1}{n+1}X_{n+1}$$

(b)

$$\begin{aligned} nS_{n+1}^2 &= \sum_{j=1}^{n+1}(X_j - \overline{X}_{n+1})^2 = \sum_{j=1}^{n}(X_j - \overline{X}_n)^2 + n(\overline{X}_n - \overline{X}_{n+1})^2 + (X_{n+1} - \overline{X}_{n+1})^2 \\ &= (n-1)S_n^2 + n\left(\frac{1}{n+1}\overline{X}_n - \frac{1}{n+1}X_{n+1}\right)^2 + (\frac{n}{n+1}X_{n+1} - \frac{n}{n+1}\overline{X}_n)^2 \\ &= (n-1)S_n^2 + \frac{n(1+n)}{(n+1)^2}(\overline{X}_n - X_{n+1})^2 = (n-1)S_n^2 + \frac{n}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2. \end{aligned}$$

4. (a) Not identifiable: for example,

$$\mathbb{P}_{\nu,\sigma^2,\alpha_1,\ldots,\alpha_p,\beta_1,\ldots,\beta_b} = \mathbb{P}_{0,\sigma^2,\alpha_1+a\nu,\ldots,\alpha_p+a\nu,\beta_1+(1-a)\nu,\ldots,\beta_b+(1-a)\nu}$$

for any $a \in \mathbb{R}$.

(b) Yes - it is identifiable. Joint density is

$$\begin{aligned} &\frac{1}{(2\pi)^{pb/2}\sigma^{pb}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{ij}(x_{ij} - \nu - \alpha_i - \beta_j)^2\right\} \\ &\quad = \frac{1}{(2\pi)^{pb/2}\sigma^{pb}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{ij} x_{ij}^2 - \sum_{ij} x_{ij}(\nu + \alpha_i + \beta_j) + \sum_{ij}(\nu + \alpha_i + \beta_j)^2\right)\right\} \end{aligned}$$

If it is not identifiable, then different $(\nu, \underline{\alpha}, \underline{\beta})$ yield the same $\nu + \alpha_i + \beta_j$ for each (i, j). If

$$\nu_1 + \alpha_{1i} + \beta_{1j} = \nu_2 + \alpha_{2i} + \beta_{2j} \qquad \forall(i, j)$$

plus zero sum conditions, then $\nu_1 = \nu_2$. Again, sum over j gives $\alpha_{1i} = \alpha_{2i}$ for each i and summing over i gives $\beta_{1j} = \beta_{2j}$. Hence it is identifiable.

5. Not identifiable; $\mathbb{P}_{\nu_1,\theta_1,\sigma^2} = \mathbb{P}_{\nu_2,\theta_2,\sigma^2}$ for all $(\mu_1,\theta_1),(\mu_2,\theta_2)$ such that $\mu_1+\theta_1 = \mu_2+\theta_2$.

6. The parametrisation is (μ, p). Let X denote number of eggs laid, Y the number that hatch. Then

$$\mathbb{P}(Y = y|X = x) = \binom{x}{y} p^y(1-p)^{x-y} \qquad \mathbb{P}(X = x) = \frac{\mu^x}{x!}e^{-\mu}$$
$$\mathbb{P}(Y = y, X = x) = \frac{x!}{y!(x-y)!}p^y(1-p)^{x-y}\frac{\mu^x}{x!}e^{-\mu} \qquad x \geq y$$

so that

$$\mathbb{P}(Y = y) = e^{-\mu}\frac{\mu^y p^y}{y!}\sum_{x=y}^{\infty}\frac{(1-p)^{x-y}\mu^{x-y}}{(x-y)!} = \frac{(\mu p)^y}{y!}e^{-\mu p}.$$

No not identifiable.

7. (a)

$$S_Y(t) = \mathbb{P}(\min(T_1',\ldots,T_n') > t) = \mathbb{P}(T > t)^n \qquad S_X(t) = \mathbb{P}(T > t)^m$$

from which the result follows directly.

(b)

$$\begin{aligned} F_{X'}(x) &= \mathbb{P}(X' \leq x) = \mathbb{P}(-\log S_0(X) \leq t) \\ &= \mathbb{P}(S_0(X) \geq e^{-t}) = \mathbb{P}(S_X(X) \geq e^{-m\delta t}) \\ &= \mathbb{P}(F_X(X) \leq 1 - e^{-m\delta t}) = 1 - e^{-m\delta t}. \end{aligned}$$

(c)

$$\alpha_T(t) = -\frac{d}{dt}\log S_T(t) \qquad \alpha_Y(t) = -\frac{d}{dt}\log S_Y(t).$$
$$\alpha_Y = c\alpha_T \Leftrightarrow -\frac{d}{dt}\log S_T(t) = -c\frac{d}{dt}\log S_Y(t) \Leftrightarrow -\frac{d}{dt}\log S_T(t) = -\frac{d}{dt}\log S_Y^c(t)$$

Now using $S_T(0) = S_Y(0) = 1$ gives:

$$S_T(t) = S_Y^c(t) \qquad \forall t \geq 1.$$

8. (a)

$$\mathbb{P}(X_{k:n} \leq x < X_{k+1:n}) = F_{X_{k:n}}(x) - F_{X_{k+1:n}}(x)$$

and

$$\begin{aligned} F_{X_{k:n}}(x) - F_{X_{k+1:n}}(x) &= \binom{n}{k}\mathbb{P}(X_1 \leq x,\ldots,X_k \leq x, X_{k+1} > x,\ldots,X_n > x) \\ &= \binom{n}{k}F(x)^k(1-F(x))^{n-k}. \end{aligned}$$

To compute $F_{X_{k:n}}(x)$, we need $F_{X_{n:n}}(x)$, but this is easy:

$$F_{X_{n:n}}(x) = F(x)^n.$$

Therefore:

$$F_{X_{k:n}}(x) = \sum_{j=k}^{n} \binom{n}{j} F(x)^k (1-F(x))^{n-k}.$$

To compute the density, take a derivative:

$$\begin{aligned}
f_{X_{k:n}}(x) &= \sum_{j=k}^{n} \binom{n}{j} \left(jF(x)^{j-1}(1-F(x))^{n-j} - (n-j)F(x)^j(1-F(x))^{n-j-1} \right) f(x) \\
&= nf(x) \sum_{j=k}^{n} \left\{ \binom{n-1}{j-1} F(x)^{j-1}(1-F(x))^{n-j} - \binom{n-1}{j} F(x)^j (1-F(x))^{n-j-1} \right\} \\
&= n \binom{n-1}{k-1} F(x)^{k-1}(1-F(x))^{n-k} f(x)
\end{aligned}$$

so that:

$$f_{X_{k:n}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x).$$

(b) For $U(0,1)$, $F(x) = x$ for $0 \le x \le 1$ and $f(x) = \mathbf{1}_{[0,1]}(x)$ so that:

$$f_{X_{k:n}}(x) = n \binom{n-1}{k-1} x^{k-1}(1-x)^{n-k} \mathbf{1}_{[0,1]}(x)$$

as required.

(c)

$$\mathbb{E}[X_{j:n}^p] = j \binom{n}{j} \int_0^1 x^p x^{j-1}(1-x)^{n-j} dx = j \binom{n}{j} \frac{\Gamma(j+p)\Gamma(n-j+1)}{\Gamma(n+p+1)}.$$

Using $\Gamma(n+1) = n!$, it follows that

$$\mathbb{E}[X_{j:n}^p] = j \frac{n!}{j!(n-j)!} \frac{(j+p-1)!(n-j)!}{(n+p)!} = \frac{\prod_{k=0}^{p-1}(j+k)}{\prod_{k=1}^{p}(n+k)}.$$

9. First, for fixed ϵ, we consider the following grid: $x_1 = \inf\{z : F(z) \ge \epsilon\}$, $x_j = \inf\{z > x_{j-1} : F(z) - F(x_{j-1}) \ge \epsilon$, define M as the smallest integer such that $1 \ge F(x_M) > 1 - \epsilon$. Since F is continuous, $F(x_j) - F(x_{j-1}) = \epsilon$ for $j = 2, \ldots, M$.

Now, if $|\widehat{F}_n(x_j) - F(x_j)| \le \epsilon$ and $|\widehat{F}_n(x_{j+1}) - F(x_{j+1})| \le \epsilon$, then it is straightforward that $\sup_{x \in [x_j, x_{j+1}]} |\widehat{F}_n(x) - F(x)| \le 2\epsilon$. Therefore

$$\begin{aligned}
\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| > \epsilon \right) &\le \mathbb{P}\left(\max_{j \in \{1,\ldots,M\}} |\widehat{F}_n(x_j) - F(x_j)| > \frac{\epsilon}{2} \right) \\
&\le \sum_{j=1}^{M} \mathbb{P}\left(|\widehat{F}_n(x_j) - F(x_j)| > \epsilon \right) \\
&\le M \times \frac{4}{\epsilon^2} \times \sup_x \frac{F(x)(1-F(x))}{n} \le \frac{1}{n\epsilon^3} \overset{n\to+\infty}{\longrightarrow} 0
\end{aligned}$$

using the fact that $\mathbb{E}[\widehat{F}_n(x)] = F(x)$ and $\mathrm{Var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} \le \frac{1}{4n}$.

Chapter 2

Sufficiency

2.1 Sufficienct Statistics

Consider a random sample $\{X_1, \dots, X_n\}$, whose parent distribution $\mathbb{P}_\theta$ belongs to a parametric family $\mathcal{P}$. Suppose that the value of the parameter (or parameter vector) $\theta \in \Theta$ is unknown and is to be estimated from the sample. Let $X = \{X_1, \dots, X_n\}$. A *statistic* $T(X)$ represents a reduction of the data X. For example, $T(X) = (X_{(1)}, \dots, X_{(n)})$ loses information about the labels of the X_i.

The idea of *sufficiency* is to reduce the data with statistics that involve no loss of information about the parameter vector $\theta \in \Theta$ in the context of a model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$.

Definition 2.1 (Sufficient Statistic). *A statistic $T(X)$ is* sufficient *for $\mathbb{P} \in \mathcal{P}$ (or rather the parameter vector θ) if*

$$\mathbb{P}_\theta(X \in .|T(X) = t)$$

does not involve θ for any $t \in \mathcal{T}$.

Example 2.1.

Arrival of customers at a service station follows a Poisson process with arrival rate θ. Let X_1 be the arrival time of the first customer and X_2 the interarrival time (time between arrivals) from the first to the second customer. Recall that $X_1 \perp X_2$ and both are $\text{Exp}(\theta)$ variables.

Result $T = X_1 + X_2$ is sufficient for θ.

Proof Firstly, $\frac{X_1}{X_1+X_2} \perp X_1 + X_2$ and $\frac{X_1}{X_1+X_2} \sim U(0,1)$. This is seen by a change of variables: let $Y = \frac{X_1}{X_1+X_2}$ and $Z = X_1 + X_2$. Then, for $0 \leq a \leq 1$, $b \geq 0$,

$$\begin{aligned}\mathbb{P}(Y \leq a, Z \leq b) &= \mathbb{P}\left(\frac{(1-a)}{a}X_1 \leq X_2, X_1 + X_2 \leq b\right) = \theta^2 \int_0^{ab} \int_{((1-a)/a)x_1}^{b-x_1} e^{-\theta(x_1+x_2)} dx_2 dx_1 \\ &= a\left(1 - e^{-\theta b} - b\theta e^{-\theta b}\right)\end{aligned}$$

This has product form, hence

$$\mathbb{P}(Y \le a) = \begin{cases} 0 & a < 0 \\ a & 0 \le a \le 1 \\ 1 & a > 1 \end{cases}$$

and

$$\mathbb{P}(Z \le b) = \begin{cases} 0 & b < 0 \\ 1 - (1 - \theta b)e^{-\theta b} & b \ge 0. \end{cases}$$

hence they are independent and $Y \sim U(0,1)$.

Conditioned on $X_1 + X_2 = t$, the distribution of

$$X_1 = \frac{X_1}{(X_1 + X_2)} \times (X_1 + X_2)$$

is the same as that of $\frac{X_1 t}{(X_1+X_2)}$. Hence,

$$X_1 | \{X_1 + X_2 = t\} \sim U(0,t).$$

It follows that

$$(X_1, X_2) | \{X_1 + X_2 = t\} \sim (U, t - U)$$

where $U \sim U(0,t)$. This does not depend on θ and hence the statistic $X_1 + X_2$ is sufficient for the parameter θ.

2.2 The Factorisation Theorem

The following important theorem, known as the *factorisation theorem* was proved in various forms by Fisher, Neyman, Halmos and Savage.

Theorem 2.2 (Factorisation Theorem). *In a regular statistical model, a statistic $T(X)$ with range $\mathcal{T}$ is sufficient for the parameter vector θ if and only if there exists a function $g(t,\theta)$ defined for all $t \in \mathcal{T}$ and a function h defined on $\mathcal{X}$ such that*

$$p(x,\theta) = g(T(x),\theta)h(x). \tag{2.1}$$

for all $x \in \mathcal{X}$ and $\theta \in \Theta$.

Proof for the discrete case Let $\mathcal{X} = \{x_1, x_2, \ldots\}$. Set $t_i = T(x_i)$. To prove that (2.1) implies sufficiency, suppose that the formula holds. Then, for $x_j : T(x_j) = t_i$,

$$\mathbb{P}_\theta(X = x_j | T = t_i) = \frac{p(x_j,\theta)}{\mathbb{P}_\theta(T = t_i)} = \frac{h(x_j)g(t_i,\theta)}{g(t_i,\theta)\sum_{y:T(y)=t_i} h(y)} = \frac{h(x_j)}{\sum_{y:T(y)=t_i} h(y)}$$

which is independent of θ and hence sufficiency is proved.

Now suppose that the statistic is sufficient. Then, for all x_j such that $T(x_j) = t_i$, there is a function α which does not depend on θ such that

$$\alpha(x_j) = \mathbb{P}_\theta(X = x_j | T = t_i) = \frac{\mathbb{P}_\theta(X = x_j)}{\mathbb{P}_\theta(T = t_i)}$$

Hence

$$\mathbb{P}_\theta(X = x_j) = \alpha(x)\mathbb{P}_\theta(T = t_i).$$

Set $h(x) = \alpha(x)$ and $g(t, \theta) = \mathbb{P}_\theta(T = t)$. □

The proof for the continuous case is similar, replacing sums with appropriate integrals.

Example 2.2 (Example 2.1 continued)**.**

Suppose that $X_1, \ldots, X_n$ are the interarrival times for n customers. They are i.i.d. Exp(θ) variables and hence

$$p(x_1, \ldots, x_n; \theta) = \begin{cases} \theta^n \exp\left\{-\theta \sum_{j=1}^n x_j\right\} & x_1 \geq 0, \ldots, x_n \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

From the factorisation theorem, it follows that $T(X_1, \ldots, X_n) = \sum_{j=1}^n X_j$ is sufficient. Set

$$g(t, \theta) = \begin{cases} \theta^n e^{-\theta t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

$$h(x_1, \ldots, x_n) = \begin{cases} 1 & x_1 \geq 0, \ldots, x_n \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Example 2.3 (Estimating the size of a population)**.**

Consider a population with θ members, labelled $1, \ldots, \theta$. Take a random sample $X_1, \ldots, X_n$ with replacement. The probability distribution of $X = (X_1, \ldots, X_n)$ is:

$$p(x_1, \ldots, x_n; \theta) = \begin{cases} \frac{1}{\theta^n} & 1 \leq \min_j x_j \leq \max_j x_j \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Let $x_{(1)} \leq \ldots \leq x_{(n)}$ denote the order statistics, so that $x_{(n)} = \max_j x_j$. Then

$$p(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n} \mathbf{1}_{[1,\theta]}(x_{(n)}).$$

It follows from the factorisation theorem that $X_{(n)}$ is a sufficient statistic for θ. □

Example 2.4 (Regression)**.**

Suppose that $Y_1, \ldots, Y_n$ are independent, $Y_j \sim N(\mu_j, \sigma^2)$, and

$$\mu_j = \beta_0 + \beta_1 z_j \qquad j = 1, \ldots, n.$$

Assume that the given constants $\{z_j\}$ are not all identical and let $\theta = (\beta_1, \beta_2, \sigma^2)^t$ denote the parameter vector. Then

$$\begin{aligned} p(y, \theta) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 z_j)^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (\beta_0 - \beta_1 z_j)^2 \right\} \\ &\qquad \times \exp\left\{ \frac{1}{2\sigma^2} \sum_{j=1}^{n} \left(-\sum_j y_j^2 + 2\beta_0 \sum_j y_j + 2\beta_1 \sum_j z_j y_j \right) \right\}. \end{aligned}$$

It follows that the parametrisation is *identifiable* and that $T(Y) = (\sum_j Y_j, \sum_j Y_j^2, \sum_j z_j Y_j)$ is sufficient for θ. □

2.3 Minimal Sufficiency

In general, for any model there are many sufficient statistics. A sufficient statistic $T(X)$ is *minimal sufficient* if it provides a greater reduction of the data than any other sufficient statistic $S(X)$. Formally, this is expressed as follows:

Definition 2.3 (Minimal Sufficiency)**.** *A statistic $T(X)$ is said to be* minimally sufficient *if for any other sufficient statistic $S(X)$, there is a transformation r such that*

$$T(X) = r(S(X)).$$

Example 2.5 (Bernoulli trials)**.**

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli(θ) distribution. That is, the distribution for each variable is

$$p(1) = \theta, \qquad p(0) = 1 - \theta.$$

Then

$$p_{X_1,\ldots,X_n;\theta}(x_1, \ldots, x_n) = \theta^{\sum_{j=1}^n x_j} (1 - \theta)^{n - \sum_{j=1}^n x_j}$$

and hence $T := \sum_{j=1}^n X_j$ is a sufficient statistic for θ. Now consider any other sufficient statistic $S(X)$. Then

$$p(x; \theta) = g(S(x), \theta) h(x)$$

so that

$$\theta^{T(x)} (1 - \theta)^{n - T(x)} = g(S(x), \theta) h(x) \qquad \forall \theta \in [0, 1].$$

For any two fixed θ_1 and θ_2, it follows that

$$\left(\frac{\theta_1}{\theta_2}\right)^T\left(\frac{1-\theta_1}{1-\theta_2}\right)^{n-T}=\frac{g(S,\theta_1)}{g(S,\theta_2)}.$$

It follows (by taking logarithms) that

$$T(x)=\frac{1}{\log\frac{\theta_1(1-\theta_2)}{\theta_2(1-\theta_1)}}\log\left\{\frac{g(S(x),\theta_1)}{g(S(x),\theta_2)}\left(\frac{1-\theta_2}{1-\theta_1}\right)^n\right\}.$$

Now (for example) set $\theta_1=\frac{2}{3}$ and $\theta_2=\frac{1}{3}$ to get $T(x)=r(S(x))$. Hence $T(x)$ is minimal sufficient. □

Definition 2.4 (The Likelihood Function)**.** *The likelihood function, for an observed data vector x, is defined as:*

$$L(\theta;x)=p(x,\theta)\qquad\theta\in\Theta$$

where $p(x,\theta)$ denotes the probability mass function for discrete variables and the probability density function for continuous variables.

The likelihood function may be used to provide minimal sufficient statistics. This is the content of the *Dynkin, Lehmann, Scheffé* theorem.

Theorem 2.5 (Dynkin, Lehmann, Scheffé)**.** *Let $L(\theta,x)$ denote the likelihood function, where $x\in\mathcal{X}$ and $\{\mathbb{P}_\theta:\theta\in\Theta\}$ is a family of probability measures defined over $(\mathcal{X},\mathcal{F})$ satisfying $\mathbb{P}_{\theta_1}(A)=0\Leftrightarrow\mathbb{P}_{\theta_w}(A)=0$ for all $\theta_1,\theta_2\in\Theta$ and all $A\in\mathcal{F}$. Suppose there exists a function $T:\mathcal{X}\rightarrow\mathbb{R}$ such that for all $t\in\mathbb{R}$ and all $x,y\in\mathcal{X}:T(x)=T(y)=t$, the ratio $\frac{L(\theta,x)}{L(\theta,y)}$ does not depend on θ, and the ratio does depend on θ for x and y such that $T(x)\neq T(y)$. Then $T(X)$ is a minimal sufficient statistic for θ.*

Proof The result is proved in the special case where $L(\theta,x)>0$ for all $x\in\mathcal{X}$, $\theta\in\Theta$.

First, it is proved that $T(X)$ is sufficient. Let

$$\mathcal{T}=\{t:\exists x\in\mathcal{X}|t=T(x)\}.$$

Let

$$A_t=\{x|T(x)=t\}.$$

For each A_t, choose one element $x_t\in A_t$. It follows that

$$h(x):=\frac{L(\theta,x)}{L(\theta,x_{T(x)})}$$

does not depend on θ. Let $g(t,\theta)=L(\theta,x_t)$. Then

$$p(x,\theta)=h(x)g(T(x),\theta).$$

and hence by the factorisation theorem, $T(X)$ is sufficient for θ.

To prove minimality, let T' be any other sufficient statistic. By the factorisation theorem, there are functions h' and g' such that

$$p(x, \theta) = g'(T'(x), \theta)h'(x).$$

Let $x, y \in \mathcal{X}$ be such that $T'(x) = T'(y)$. Then

$$\frac{p(\theta, x)}{p(\theta, y)} = \frac{h'(x)}{h'(y)}.$$

Since this ratio does not depend on θ, it follows from the assumptions of the theorem that $T(x) = T(y)$ and hence $T = r(T')$ for some function r. Hence T is minimal. □

Lecture 2: Summary

Sufficiency

- Sufficiency, definition of a sufficient statistic, Factorisation theorem.
- Minimal sufficiency.
- Dynkin Lehmann Scheffé theorem

Tutorial 3

1. Let $X_1, \ldots, X_n$ be a random sample from a Poiss(θ) population where $\theta > 0$.

   (a) Show directly that $\sum_{j=1}^{n} X_j$ is sufficient for θ.

   (b) Establish the same result using the Factorisation Theorem.

2. Suppose that $X_1, \ldots, X_n$ is a random sample from a population with the following density:

$$p(x, \theta) = \begin{cases} \theta a x^{a-1} \exp\{-\theta x^a\} & x > 0, \quad \theta > 0, \quad a > 0 \\ 0 & \text{otherwise} \end{cases}$$

   where a is fixed. This is known as the *Weibull* density. Find a real-valued sufficient statistic for θ.

3. Let X be a random variable with state space $\mathcal{X} = \{v_1, \ldots, v_k\}$ and probability distribution $\mathbb{P}_\theta(X = v_i) = \theta_i$ for $i \in \{1, \ldots, k\}$ (so that $\sum_{i=1}^{k} \theta_i = 1$) and suppose that $\theta_i \in (0, 1)$ for each $i = 1, \ldots, k$. Let $X_1, \ldots, X_n$ be a random sample from X. Let

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{\{v_j\}}(X_i).$$

   (the number of trials such that $X_i = v_j$).

   (a) What is the distribution of $(N_1, \ldots, N_k)$?

   (b) Show that $N = (N_1, \ldots, N_k)$ is sufficient for $\theta = (\theta_1, \ldots, \theta_k)$.

4. Let $X_1, \ldots, X_n$ be a random sample from a population with density $p(x, \theta)$ where:

$$p(x, \theta) = \frac{1}{\sigma} \exp\left\{ - \left(\frac{x - \mu}{\sigma} \right) \right\} \mathbf{1}_{[\mu, +\infty)}(x).$$

   Here $\theta = (\mu, \sigma)$, $\Theta = (-\infty, +\infty) \times (0, +\infty)$.

   (a) Show that $\min(X_1, \ldots, X_n)$ is sufficient for μ when σ is fixed.
   Note: you cannot use the factorisation theorem for this part since the support of the density depends on μ.

   (b) Find a one-dimensional sufficient statistic for σ when μ is fixed.

   (c) Find a two-dimensional sufficient statistic for $\theta = (\mu, \sigma)$.

5. Let $X_1, \ldots, X_n$ be a random sample from a distribution F. Treating F as a parameter, show that the order statistic $(X_{(1)}, \ldots, X_{(n)})$ is sufficient for F.

6. Let $X_1, \ldots, X_n$ be a random sample from $f(t - \theta), \theta \in \mathbb{R}$. Show that the order statistic is *minimal sufficient* for f when f is the Cauchy density

$$f(t) = \frac{1}{\pi(1 + t^2)} \qquad t \in \mathbb{R}.$$

7. Let $X_1, \ldots, X_m; Y_1, \ldots, Y_n$ be independent and distributed according to $N(\mu, \sigma^2)$ and $N(\eta, \tau^2)$ respectively. Find minimal sufficient statistics in the following three cases, where $(\mu, \eta, \sigma, \tau) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$:

   (a) μ, η, σ, τ arbitrary.

   (b) $\sigma = \tau$, μ, η, σ arbitrary.

   (c) $\mu = \eta$, μ, σ, τ arbitrary.

8. Let $Y = (Y_1, \ldots, Y_n)^t$ be a multivariate Gaussian random vector with distribution

$$Y \sim N(X\beta, \sigma^2 I)$$

   where X is an $n \times p$ design matrix (values are given) and $\beta = (\beta_1, \ldots, \beta_p)^t$ is a parameter vector. Compute a $p+1$ dimensional sufficient statistic for (β, σ^2).

9. Let $Y_1, \ldots, Y_n$ be independent Bernoulli trials, where

$$\mathbb{P}(Y_j = 1) = \frac{1}{1 + \exp\left\{-\sum_{k=1}^{p} X_{jk}\beta_k\right\}} \qquad j = 1, \ldots, n.$$

   Compute a p dimensional sufficient statistic for β.

Short Answers

1. (a) Let $T = \sum_{j=1}^n X_j$. Note: $T \sim \text{Poiss}(n\theta)$. For $x_1 + \ldots + x_n = t$

$$\begin{aligned}\mathbb{P}_\theta((X_1,\ldots,X_n) = (x_1,\ldots,x_n)|T=t) &= \frac{\mathbb{P}_\theta((X_1,\ldots,X_n) = (x_1,\ldots,x_n))}{\mathbb{P}(T=t)} \\ &= \frac{\theta^{\sum_{j=1}^n x_j}\prod_{j=1}^n \frac{1}{x_j!}e^{-n\theta}}{\theta^t \frac{1}{t!}e^{-n\theta}} = \frac{(\sum_j x_j)!}{\prod_{j=1}^n x_j!}\end{aligned}$$

which does not depend on θ.

(b)

$$\mathbb{P}_\theta((X_1,\ldots,X_n) = (x_1,\ldots,x_n)) = \frac{\theta^{\sum_{j=1}^n x_j}}{\prod_{j=1}^n x_j!}\exp\{-n\theta\}$$

This factorises as $g(\sum_{j=1}^n x_j, \theta)h(\underline{x})$ where $g(t,\theta) = \theta^t e^{-n\theta}$ and $h(\underline{x}) = \frac{1}{\prod_{j=1}^n x_j!}$.

2.

$$f(x_1,\ldots,x_n;\theta) = \begin{cases} \theta^n a^n \left(\prod_{j=1}^n x_j\right)^{a-1} \exp\left\{-\theta\sum_{j=1}^n x_j^a\right\} & x_1 > 0,\ldots,x_n > 0 \\ 0 & \text{other}\end{cases}$$

Set $t(x_1,\ldots,x_n) = \sum_{j=1}^n x_j^a$ then

$$f(x_1,\ldots,x_n;\theta) = g(t(x_1,\ldots,x_n),\theta)h(x_1,\ldots,x_n)$$

where

$$g(t,\theta) = \theta^n e^{-\theta t}, \qquad h(x_1,\ldots,x_n) = a^n \mathbf{1}_{\{x_1>0,\ldots,x_n>0\}}\left(\prod_{j=1}^n x_j\right)^{a-1}$$

Hence $t(X_1,\ldots,X_n)$ is sufficient for θ.

3. (a)

$$(N_1,\ldots,N_k) \sim \text{mult}(n;\theta_1,\ldots,\theta_k).$$

(b)

$$\mathbb{P}_\theta\left((X_1,\ldots,X_n) = (v_{a_1},\ldots,v_{a_n})\right) = \theta_1^{n_1}\ldots\theta_k^{n_k}$$

where $n_j = \sum_{i=1}^n \mathbf{1}(a_i = j)$. This is in the required form from the factorisation theorem.

4. (a) The joint density is:

$$p(x_1,\ldots,x_n,\theta) = \frac{1}{\sigma^n}\exp\left\{-\sum_{j=1}^n \frac{x_j-\mu}{\sigma}\right\}\mathbf{1}_{\{\min_j x_j \geq \mu\}}.$$

The factorisation theorem cannot be used, since the support of the density depends on μ. We therefore show that the conditional density $p(x_1,\ldots,x_n|\min_j X_j = y)$ does not depend on μ.

Since $X-\mu \sim \text{Exp}(\frac{1}{\sigma})$, therefore $\min_{j\in\{1,\ldots,n\}} X_j - \mu \sim \text{Exp}(\frac{n}{\sigma})$ so that the density of $Y := \min_j X_j$ is:

$$p_Y(y) = \frac{n}{\sigma}\exp\left\{-\frac{n(y-\mu)}{\sigma}\right\}\mathbf{1}_{\{y\geq\mu\}}$$

and therefore

$$p(x_1,\ldots,x_n|\min_j X_j = y) = \frac{p(x_1,\ldots,x_n)}{p_Y(y)} = \frac{1}{n\sigma^{n-1}}\exp\left\{-\frac{1}{\sigma}\sum_{j=1}^n(x_j - \min_i x_i)\right\}$$

which does not depend on μ, hence $\min_i X_i$ is sufficient for μ when σ is fixed.

(b) From the factorisation theorem, it follows that $\sum_{j=1}^n X_j$ is sufficient for σ when μ is fixed.

(c) From the factorisation theorem, conditioned on $\min_j X_j = y$, $(\sum_{j=1}^n X_j, \min\{X_1,\ldots,X_n\})$ is sufficient for σ.

Hence $p(x_1,\ldots,x_n|\sum_j x_j = z, \min_j = y)$ depends neither on μ nor on σ, hence from the definition of sufficiency $(\min_j X_j, \sum_j X_j)$ is sufficient for $\theta = (\mu,\sigma)$.

5. Once the order statistics $x_{(1)},\ldots,x_{(n)}$ are given, the problem is then the random assignment (without replacement) of $x_1,\ldots,x_n$ to $x_{(1)},\ldots,x_{(n)}$. There are $n!$ permutations, each with equal probability. Suppose that there are m groups, group j contains n_j so that $n_1+\ldots+n_m = n$, and the order statistics are equal within each group. Then

$$\mathbb{P}((X_1,\ldots,X_n) = (x_1,\ldots,x_n)|x_{(1)},\ldots,x_{(n)}) = \frac{\prod_{j=1}^m n_j!}{n!}$$

which does not depend on F.

6. We use the Dynkin Lehman Scheffe lemma; a statistic T is minimal sufficient if $\frac{L(\theta;x)}{L(\theta;y)}$ does not depend on θ for $T(x) = T(y)$ and does depend on θ for $T(x)\neq T(y)$.

$$L(\theta;x_1,\ldots,x_n) = \frac{1}{\pi^n}\prod_{j=1}^n\frac{1}{(1+(x_j-\theta)^2)}$$

$$\frac{L(\theta;\underline{x})}{L(\theta;\underline{y})} = \prod_{j=1}^n\frac{(1+(y_j-\theta)^2)}{(1+(x_j-\theta)^2)}$$

Firstly, $L(x_1,\ldots,x_n;\theta) = L(x_{(1)},\ldots,x_{(n)};\theta)$ so that if $\underline{y}$ is a permutation of $\underline{x}$, then $\frac{L(\theta;\underline{x})}{L(\theta;\underline{y})} = 1$. To see that this is minimal, the function does not depend on θ only if the roots of the numerators and denominators are the same (considering as functions of θ), These are: $\theta = y_j \pm i$ for $j = 1,\ldots,n$ (for the numerator) and $\theta = x_j \pm i$ for $j = 1,\ldots,n$ for the denominator (where $i=\sqrt{-1}$). These are the same if and only if $(y_{(1)},\ldots,y_{(n)}) = (x_{(1)},\ldots,x_{(n)})$.

7. We use the Dynkin Lehman Scheffe lemma; a statistic T is minimal sufficient if $\frac{L(\theta;x)}{L(\theta;y)}$ does not depend on θ for $T(x)=T(y)$ and does depend on θ for $T(x)\neq T(y)$. This is equivalent to these properties holding for the log likelihood; $\log L(\theta;x)-\log L(\theta;y)$.

   The log likelihood function is:

$$\log L(\mu,\eta;\sigma,\tau;\underline{x},\underline{y}) = -\frac{(n+m)}{2}\log(2\pi) - m\log\sigma - n\log\tau$$
$$-\frac{1}{2\sigma^2}\left(\sum_{j=1}^m x_j^2 - \mu\sum_{j=1}^m x_j + m\mu^2\right) - \frac{1}{2\tau^2}\left(\sum_{j=1}^n y_j^2 - \eta\sum_{j=1}^n y_j + n\eta^2\right).$$

   Write out

$$\begin{aligned}\log L(\theta;\underline{x}_1,\underline{y}_1)-\log L(\theta;\underline{x}_2,\underline{y}_2) &= -\frac{1}{2\sigma^2}\left(\sum_{j=1}^m x_{1j}^2 - \sum_{j=1}^m x_{2j}^2\right) + \frac{\mu}{2\sigma^2}\left(\sum_{j=1}^m x_{1j} - \sum_{j=1}^m x_{2j}\right)\\ &\quad -\frac{1}{2\tau^2}\left(\sum_{j=1}^n y_{1j}^2 - \sum_{j=1}^n y_{2j}^2\right) + \frac{\eta}{2\tau^2}\left(\sum_{j=1}^n y_{1j} - \sum_{j=1}^n y_{2j}\right)\end{aligned}$$

   and obtain:

   (a) $\sum_{j=1}^m X_j$, $\sum_{j=1}^m X_j^2$, $\sum_{j=1}^n Y_j$, $\sum_{j=1}^n Y_j^2$.

   (b) $\sum_{j=1}^m X_j$, $\sum_{j=1}^n Y_j$, $\sum_{j=1}^m X_j^2 + \sum_{j=1}^n Y_j^2$.

   (c) $\sum_{j=1}^m X_j$, $\sum_{j=1}^n Y_j$, $\sum_{j=1}^n X_j^2$, $\sum_{j=1}^n Y_j^2$. (same as part (a)).

8. Density is:

$$f(y_1,\ldots,y_n) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}(y^t y - 2y^t X\beta + \beta^t X^t X\beta)\right\}$$

   so, by the factorisation theorem, a $p+1$ dimensional sufficient statistic is $y^t X, y^t y$.

9. For an outcome $(Y_1,\ldots,Y_n)=(y_1,\ldots,y_n)$ where $(y_1,\ldots,y_n)$ is a vector of 0's and 1's, we have

$$\mathbb{P}_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \frac{\prod_{j=1}^n \exp\{(1-y_j)\sum_{k=1}^p X_{jk}\beta_k\}}{\prod_{j=1}^n(1+\exp\left\{-\sum_{k=1}^p X_{jk}\beta_k\right\})}$$

   The sufficient statistic is therefore $(\sum_{j=1}^n y_j X_{jk} : k=1,\ldots,p)$.

Chapter 3

Exponential Families

The class of families known as *exponential families* was first introduced into statistics in the 1930's independently by Koopman [6] (1936), Pitman and Wishart [7](1936) and Darmois [2] (1935). Many of the standard families of probability distributions are *exponential families.*

Definition 3.1. *A family of distributions* $\{\mathbb{P}_\theta : \theta \in \Theta\}$ *with density or probability mass function* p *defined on* $\mathcal{X} \subseteq \mathbb{R}^q$*, where* $\Theta \subset \mathbb{R}^k$ *is said to be a* k*-parameter exponential family if there exist* k *real valued functions* $\eta_1, \ldots, \eta_k$*, a function* $B : \Theta \rightarrow \mathbb{R}$ *and real valued functions* $T_1, \ldots, T_k, h : \mathcal{X} \rightarrow \mathbb{R}$ *such that*

$$p(x, \theta) = h(x) \exp\left\{\sum_{j=1}^{k} \eta_j(\theta)T_j(x) - B(\theta)\right\} \qquad x \in \mathcal{X} \subseteq \mathbb{R}^q$$

It is clear that the statistic $(T_1, \ldots, T_k)$ is sufficient for θ and is referred to as the *natural sufficient statistic* for the family.

For a single parameter, this reduces to:

$$p(x, \theta) = h(x) \exp\{\eta(\theta)T(x) - B(\theta)\}$$

Here are some examples:

Example 3.1 (Poisson Distribution)**.**

Let $\{\mathbb{P}_\theta : \theta \in (0, +\infty)\}$ be the family of Poisson distributions, where $\mathbb{E}_\theta[X] = \theta$. That is,

$$p(x, \theta) = \frac{\theta^x}{x!}e^{-\theta} = \frac{1}{x!}\exp\{x \log\theta - \theta\}.$$

Here $q = 1$, $\eta(\theta) = \log\theta$, $B(\theta) = \theta$, $T(x) = x$ and $h(x) = \frac{1}{x!}$.

Example 3.2 (Normal Distribution)**.**

Let $\{\mathbb{P}_\theta : \theta \in (-\infty, \infty) \times (0, +\infty)\}$ be the family $N(\mu, \sigma^2)$ where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Then

$$p(x,\theta)=\frac{1}{\sqrt{2\pi\theta_2}}\exp\left\{-\frac{(x-\theta_1)^2}{2\theta_2}\right\}=\exp\left\{-\frac{x^2}{2\theta_2}+x\frac{\theta_1}{\theta_2}-\left(\frac{\theta_1^2}{2\theta_2}+\frac{1}{2}\log(2\pi\theta_2)\right)\right\}$$

Here

$$q=1,\qquad \eta_1(\theta)=-\frac{1}{2\theta_2},\qquad \eta_2(\theta)=\frac{\theta_1}{\theta_2},\qquad B(\theta)=\frac{\theta_1^2}{2\theta_2}+\frac{1}{2}\log(2\pi\theta_2),\qquad h(x)=1.$$

For random sampling from an exponential family, the distribution of the random sample is again an exponential family. For example, if $X_1,\ldots,X_n$ is a random sample from a $N(\theta_1,\theta_2)$ distribution, then its joint density is given by:

$$p(x_1,\ldots,x_n;\theta)=\exp\left\{-\frac{1}{2\theta_2}\sum_{j=1}^n x_j^2+\frac{\theta_1}{\theta_2}\sum_{j=1}^n x_j-n\left(\frac{\theta_1^2}{2\theta_2}+\frac{1}{2}\log(2\pi\theta_2)\right)\right\}.$$

It follows that the natural sufficient statistic for a random sample $(X_1,\ldots,X_n)$ from $N(\mu,\sigma^2)$ is $(\sum_{j=1}^n X_j,\sum_{j=1}^n X_j^2)$.

Definition 3.2 (Canonical Parametrisation)**.** *The* canonical parametrisation *is the parametrisation where the parameters $\eta=(\eta_1,\ldots,\eta_k)^t$ is used instead of θ. The density or probability mass function is written:*

$$p(x,\eta)=h(x)\exp\{(T(x),\eta)-A(\eta)\}\qquad x\in\mathcal{X}\subseteq\mathbb{R}^q. \tag{3.1}$$

The exponential family is said to be *generated by* (T,h). The function $A(\eta)$ is known as the *log-partition function*; it is the logarithm of the normalisation factor, by which the expression $\exp\{(T(x),\eta)\}\,h(x)$ has to be divided to make it a probability density / mass function.

Example 3.3 (Normal (continued))**.**

For the $N(\theta_1,\theta_2)$ density, $\eta_1=-\frac{1}{2\theta_2}$ and $\eta_2=\frac{\theta_1}{\theta_2}$, while $T(x)=(T_1(x),T_2(x))=(x^2,x)$.

$$A(\eta)=-\frac{\eta_2^2}{4\eta_1}+\frac{1}{2}\log(-\frac{\pi}{\eta_1})$$

□

Example 3.4 (Multinomial Trials)**.**

Consider the outcomes of n independent trials $(X_1,\ldots,X_n)$ where $\mathbb{P}_\theta(X_j=a)=\theta_a$ for $a=1,\ldots,k$ and $\sum_{a=1}^k\theta_a=1$. Let $T_a(x_1,\ldots,x_n)=\sum_{j=1}^n\mathbf{1}_{\{a\}}(X_j)$. Then

$$p(x,\theta)=\prod_{a=1}^k\theta_a^{T_a(x)}=\exp\left\{\sum_{a=1}^k T_a(x)\log\theta_a\right\}\qquad \theta\in\left\{[0,1]^k:\sum_{a=1}^k\theta_a=1\right\}.$$

Now use a parametrisation $(\lambda_1,\ldots,\lambda_k)$ where $(\lambda_1,\ldots,\lambda_k)$ is any solution to:

$$\theta_a = \frac{e^{\lambda_a}}{\sum_{b=1}^k e^{\lambda_b}} \qquad a = 1, \ldots, k,$$

then $\log \theta_a = \lambda_a - \log \sum_{b=1}^k e^{\lambda_b}$. This has the advantage that $(\lambda_1, \ldots, \lambda_k)$ are all free variables, while the restriction $\sum_{a=1}^k \theta_a = 1$ and $\theta \in [0,1]^k$ holds. Then, in terms of the parameters $(\lambda_1, \ldots, \lambda_k)$, the probability mass function is:

$$q_0(x, \lambda) = \exp\left\{ \sum_{j=1}^k \lambda_j T_j(x) - n \log \sum_{j=1}^k e^{\lambda_j} \right\}. \tag{3.2}$$

This is the canonical parametrisation, but the parametrisation is not *identifiable* since for any $\alpha \in \mathbb{R}$, $(\lambda_1+\alpha, \ldots, \lambda_k+\alpha)$ (shift each parameter value by α) gives the same distribution. This can be remedied quite simply by noting that $T_1(x) + \ldots + T_k(x) = n$ and setting $\eta_j = \lambda_j - \lambda_k$ for $j = 1, \ldots, k-1$. With this parametrisation,

$$\begin{aligned} q(x, \eta) &= \exp\left\{ \sum_{j=1}^{k-1} (\eta_j + \lambda_k) T_j(x) + \lambda_k T_k(x) - n \log \left(e^{\lambda_k} + \sum_{j=1}^{k-1} e^{\eta_j + \lambda_k} \right) \right\} \\ &= \exp\left\{ \sum_{j=1}^{k-1} \eta_j T_j(x) - n \log \left(1 + \sum_{j=1}^{k-1} e^{\eta_j} \right) \right\}. \end{aligned}$$

This is a canonical representation where the parameters are *identifiable.* Identifiability is not dealt with now; it is dealt with later, by Theorem 3.7. Firstly, show that the covariance matrix of $(T_1, \ldots, T_{k-1})$ is a strictly positive definite $k-1 \times k-1$ matrix. Recall that $\text{Var}_\theta(T_i) = n\theta_i^2$ and $\text{Cov}_\theta(T_i, T_j) = -n\theta_i\theta_j$ and that $\sum_{j=1}^{k-1} \theta_j = 1 - \theta_k < 1$. From this, it is straightforward to establish that the covariance matrix is strictly positive definite. This is equivalent (by Theorem 3.7) to identifiability of the parametrisation.

□

3.1 Properties of Exponential Families

For a random k-vector T, the moment generating function $M_T : \mathbb{R}^k \to \mathbb{R}_+$ is defined as:

$$M_T(s) = \mathbb{E}\left[\exp\left\{(s, T)\right\}\right]. \tag{3.3}$$

The notation Σ_T will be used to denote the $k \times k$ covariance matrix; $\Sigma_{T;i,j} = \text{Cov}(T_i, T_j)$.

Definition 3.3 (Natural Parameter Space)**.** *The* natural parameter space $\mathcal{E}$ *is the set of* $\eta \in \mathbb{R}^k$ *such that* $A(\eta) < +\infty$*.*

Theorem 3.4. *Let* $\mathcal{P}$ *be a canonical* k*-parameter exponential family generated by* (T, h) *with corresponding natural parameter space* $\mathcal{E}$ *and log-partition function* $A(\eta)$*. Then* *(a) $\mathcal{E}$ is convex.*

(b) $A : \mathcal{E} \to \mathbb{R}$ is convex.

(c) If $\mathcal{E}$ has non empty interior $\mathcal{E}^0$ in $\mathbb{R}^k$ and $\eta_0 \in \mathcal{E}^0$, then the moment generating function M of $T(X)$ is given by:

$$M_T(s) = \exp\left\{A(\eta_0 + s) - A(\eta_0)\right\}. \tag{3.4}$$

valid for all s such that $\eta_0 + s \in \mathcal{E}$. Since η_0 is an interior point, this set includes a ball of radius ϵ for some $\epsilon > 0$ centred at 0.

The expression for the moment generating function gives a convenient way of computing the summary statistics (mean and covariance matrix):

Corollary 3.5. *Under the conditions of Theorem 3.4,*

$$\mathbb{E}_{\eta_0}\left[T(X)\right] = \dot{A}(\eta_0)$$

$$\Sigma_T^{(\eta_0)} = \ddot{A}(\eta_0)$$

where

$$\dot{A}(\eta_0) = \left(\frac{\partial A}{\partial \eta_1}(\eta_0), \ldots, \frac{\partial A}{\partial \eta_k}(\eta_0)\right)^t$$

and $\ddot{A}(\eta_0)$ is the $k \times k$ matrix with entries $\ddot{A}_{ij}(\eta_0) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta_0)$.

Proof of Corollary 3.5 This follows directly from expression (3.4) □

Proof of Theorem 3.4 Firstly, part (b) is proved. Suppose that $\eta_1, \eta_2 \in \mathcal{E}$ and $0 \leq \alpha \leq 1$. A simple application of Hölder's inequality gives:

$$\begin{aligned} 1 &= \int h(x) \exp\left\{(T(x), \alpha\eta_1 + (1-\alpha)\eta_2)\right\} dx \exp\left\{-A(\alpha\eta_1 + (1-\alpha)\eta_2)\right\} \\ &\leq \left(\int h(x) \exp\left\{(T(x), \eta_1)\right\} dx\right)^{\alpha} \left(\int h(x) \exp\left\{(T(x), \eta_2)\right\} dx\right)^{1-\alpha} \exp\left\{-A(\alpha\eta_1 + (1-\alpha)\eta_2)\right\} \\ &= \exp\left\{\alpha A(\eta_1) + (1-\alpha)A(\eta_2) - A(\alpha\eta_1 + (1-\alpha)\eta_2)\right\}. \end{aligned}$$

It follows that for all $\eta_1, \eta_2 \in \mathcal{E}$ and all $\alpha \in (0,1)$.

$$\alpha A(\eta_1) + (1-\alpha)A(\eta_2) - A(\alpha\eta_1 + (1-\alpha)\eta_2) \geq 0.$$

and part (b) is proved.

From part (b), it follows that if $\eta_1, \eta_2 \in \mathcal{E}$, then $\alpha\eta_1 + (1-\alpha)\eta_2 \in \mathcal{E}$ and part (a) is proved.

For part (c), the proof is given in the continuous case; the discrete case is similar.

$$\begin{aligned} M_T(s) &= \mathbb{E}_\eta\left[e^{(s,T(X))}\right] = \int e^{(s,T(x))}h(x)e^{(\eta,T(x))-A(\eta)}dx \\ &= \left(\int e^{(s+\eta,T(x))}h(x)dx\right)e^{-A(\eta)} = e^{A(s+\eta)-A(\eta)} \end{aligned}$$

and the result is proved. □

Example 3.5 (Example 3.4 (continued)).

To compute the expectation vector and covariance matrix, the overparametrised version (3.2) is more convenient. Here

$$A(\lambda) = n\log\sum_{j=1}^k e^{\lambda_j}$$

$$\dot{A}(\lambda) = \frac{n}{\sum_{j=1}^k e^{\lambda_j}}\left(e^{\lambda_1},\ldots,e^{\lambda_k}\right) = (n\theta_1,\ldots,n\theta_k) = \left(\mathbb{E}_\theta\left[T_1(X)\right],\ldots,\mathbb{E}_\theta\left[T_k(X)\right]\right).$$

$$\ddot{A}_{ij}(\lambda) = \begin{cases} \frac{ne^{\lambda_i}}{\sum_{a=1}^k e^{\lambda_a}} - \frac{ne^{2\lambda_i}}{\left(\sum_{a=1}^k e^{\lambda_a}\right)^2} & i=j \\ -\frac{ne^{\lambda_i}e^{\lambda_j}}{\left(\sum_{a=1}^k e^{\lambda_a}\right)^2} & i\neq j \end{cases}$$

giving (in the parametrisation θ where derivatives are taken with respect to the natural parametrisation η)

$$\ddot{A}_{ij}(\theta) = \begin{cases} n\theta_i(1-\theta_i) & i=j \\ -n\theta_i\theta_j & i\neq j \end{cases}$$

and hence $\ddot{A}(\theta) = \Sigma_T^{(\theta)}$. □

3.2 Rank of an Exponential Family

Definition 3.6 (Rank). *An exponential family is of rank k if and only if the generating statistic T is k dimensional and $(1, T_1(X), \ldots, T_k(X))$ are linearly independent with positive probability. That is,*

$$\mathbb{P}_\eta\left(\sum_{j=1}^k a_jT_j(X) = a_{k+1}\right) < 1 \qquad \forall(a_1,\ldots,a_{k+1}) \neq 0.$$

Theorem 3.7. *Let $\mathcal{P} = \{q(x,\eta) : \eta \in \mathcal{E}\}$ be a canonical exponential family generated by (T,h), where T is k-dimensional, with natural parameter space $\mathcal{E}$ and suppose that $\mathcal{E}$ is open. Suppose that $\Sigma_T^{(\eta)}$ is well defined for each $\eta\in\mathcal{E}$. Then the following conditions are equivalent:*

1. *$\mathcal{P}$ is of rank k.*

2. *η is identifiable.*

3. *$\Sigma_T^{(\eta)}$ is positive definite for all $\eta \in \mathcal{E}$.*

4. *$\eta \to \dot{A}(\eta)$ is $1-1$ on $\mathcal{E}$*

5. *A is strictly convex on $\mathcal{E}$.*

Proof 1. ≡ 3. is straightforward: let $a = (a_1, \ldots, a_k)^t$, then

$$a^t \Sigma_T^{(\eta)} a = \mathrm{Var}_\eta \left(\sum_{j=1}^k a_j T_j \right) = 0 \Leftrightarrow \exists a_{k+1} : \mathbb{P}_\eta \left(\sum_{j=1}^k a_j T_j = a_{k+1} \right) = 1.$$

1. ≡ 2. η not identifiable $\Leftrightarrow$ there exist $\eta_1 \neq \eta_2 \in \mathcal{E}$ such that

$$\begin{aligned} \exp\left\{(\eta_1, T(x)) - A(\eta_1)\right\} h(x) &= \exp\left\{(\eta_2, T(x)) - A(\eta_2)\right\} h(x) \\ &\Leftrightarrow (\eta_1 - \eta_2, T(x)) = A(\eta_2) - A(\eta_1) \quad \forall x \end{aligned}$$

so for $\eta = \eta_1$ or η_2, take $a_j = (\eta_1 - \eta_2)_j$ and $a_{k+1} = A(\eta_2) - A(\eta_1)$ and

$$\mathbb{P}_\eta \left(\sum_{j=1}^k a_j T_j(X) = a_{k+1} \right) = 1$$

The result 1. ≡ 2. follows.

2. ≡ 5. In the proof of Theorem 3.4, the Hölder inequality gives a strict inequality if and only if $\mathbb{P}_{\eta_1} \neq \mathbb{P}_{\eta_2}$ and hence 2. $\Leftrightarrow$ 5.
3. ≡ 5. This follows directly from the definition of convexity, since $\ddot{A}(\eta) = \Sigma_T^{(\eta)}$.
3. ⇒ 4., firstly note that $\ddot{A}$ is non-negative definite. Choose two points $\eta_1 \neq \eta_2$ such that $\eta_1 \neq \eta_2$ and $\dot{A}(\eta_1) = \dot{A}(\eta_2)$. Let $\eta(s) = \eta_1 + s(\eta_2 - \eta_1)$ and let $\gamma = \eta_2 - \eta_1$. Then

$$0 = \gamma^t \dot{A}(\eta_2) - \gamma^t \dot{A}(\eta_1) = \int_0^1 \gamma^t \ddot{A}(\eta(s)) \gamma ds$$

which implies that $\ddot{A}(\eta(s)) = 0$ for all $s \in [0,1]$, which contradicts $\ddot{A}(\eta)$ positive definite for all $\eta \in \mathcal{E}$. It follows that $3 \Rightarrow 4$.
4. ⇒ 3. For any point η_0 and any γ satisfying $\|\gamma\| = 1$, choose a sequence $(\eta_{1,n})$ and $(\eta_{2,n})$ such that for each n, $\eta_0 = \frac{\eta_{1,n} + \eta_{2,n}}{2}$ and such that $\gamma := \frac{\eta_{2,n} - \eta_{1,n}}{\|\eta_{2,n} - \eta_{1,n}\|}$. Then, letting $\eta_n(s) = \eta_{1,n} + s(\eta_{2,n} - \eta_{1,n})$:

$$\frac{\dot{A}(\eta_{1,n}) - \dot{A}(\eta_{2,n})}{\|\eta_{2,n} - \eta_{1,n}\|} = \int_0^1 \ddot{A}(\eta(s)) \gamma ds \overset{n \to +\infty}{\longrightarrow} \ddot{A}(\eta_0)\gamma.$$

Since the mapping is $1-1$, this is non-zero, hence

$$0 \neq \gamma^t \ddot{A}(\eta_0) \gamma$$

thus establishing that $\ddot{A}(\eta)$ is positive definite for all $\eta \in \mathcal{E}$. It follows that 4. ⇒ 3.

The equivalences have now been established. □

Statement 4. of the previous theorem leads to the following very important corollary:

Corollary 3.8. *Suppose that the conditions of Theorem 3.7 hold. Then*

1. $\mathcal{P}$ *may be uniquely parametrised by* $\mu(\eta) = \mathbb{E}_\eta[T(X)]$, *where* μ *ranges over* $\dot{A}(\mathcal{E})$.
2. $\log p(x,\eta)$ *is a strictly concave function of* η *on* $\mathcal{E}$.

Proof This is simply a restatement of 4. and 5. of Theorem 3.7. □

The parametrisation in terms of μ is known as the *mean* parametrisation.

Example 3.6 (The p-variate Gaussian Family).

A random p-vector Y with positive definite covariance function Σ has a Gaussian $N_p(\mu,\Sigma)$ distribution if and only if its density is:

$$p(y,\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(y-\mu)^t\Sigma^{-1}(y-\mu)\right\} \qquad y \in \mathbb{R}^p.$$

It follows that

$$\log p(y,\mu,\Sigma) = -\frac{1}{2}y^t\Sigma^{-1}y + \mu^t\Sigma^{-1}y - \frac{1}{2}\left(\log|\Sigma| + \mu^t\Sigma^{-1}\mu\right) - \frac{p}{2}\log\pi.$$

Note that

$$-\frac{1}{2}y^t\Sigma^{-1}y = -\left(\sum_{1\le i<j\le p}\Sigma^{-1}_{ij}y_iy_j + \frac{1}{2}\sum_{i=1}^p\Sigma^{-1}_{ii}y_i^2\right)$$

and

$$\mu^t\Sigma^{-1}y = \sum_{i=1}^p\sum_{j=1}^p\Sigma^{-1}_{ij}\mu_jy_i.$$

It follows that this is a $k = \frac{1}{2}p(p+3)$ parameter exponential family with statistics

$$(Y_1,\ldots,Y_p,\{Y_iY_j\}_{1\le i\le j\le p})$$

and parameters

$$h(y) \equiv 1, \qquad \theta = (\mu,\Sigma), \qquad B(\theta) = \frac{1}{2}\left(\log|\Sigma| + \mu^t\Sigma^{-1}\mu\right).$$

Here $\eta_i = \sum_j \mu_j\Sigma^{-1}_{ji}$ is the natural parameter associated with y_i, while $\eta_{ii} = -\frac{1}{2}\Sigma^{-1}_{ii}$ is the natural parameter associated with y_i^2 and $\eta_{ij} = -\Sigma^{-1}_{ij}$ the natural parameter associated with y_iy_j for $i \neq j$.

Letting M denote the mean parameters, they are: $M_i = \mathbb{E}[Y_i] = \mu_i$ for $i = 1,\ldots,p$ and $M_{ij} = \mathbb{E}[Y_iY_j] = \Sigma_{ij} + \mu_i\mu_j$ for $1 \le i \le j \le p$. □

Summary of Lecture 3

Exponential Families

- Introduction to Exponential Families: Definition
- Canonical Parametrisation
- Examples
- Moment generating function.
- Rank of an exponential family.
- Theorem giving equivalence between rank, identifiability, positive definiteness of covariance for sufficient statistics, parameter to mean coordinate map $1-1$, moment map strictly convex.

Tutorial 4

1. Express the following families as exponential families, identifying the terms in the expression:

   (a) The beta family:

$$p(x; \beta_1, \beta_2) = \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} x^{\beta_1 - 1}(1-x)^{\beta_2 - 1} \qquad 0 \leq x \leq 1 \qquad \beta_1 > 0, \quad \beta_2 > 0.$$

   (b) The gamma family:

$$p(x; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha - 1} e^{-\lambda x} \qquad x \geq 0, \quad \lambda > 0, \alpha > 0$$

2. Which of the following are exponential families? Prove or disprove.

   (a) The $U(0, \theta)$ family for $\theta > 0$. That is, $X \sim U(0, \theta)$ if it has density

$$p(x; \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{other} \end{cases}$$

   (b) The family of densities:

$$p(x; \theta) = \mathbf{1}_{[0,\theta]}(x) \exp\left\{-2 \log \theta + \log(2x)\right\}$$

   where $\theta > 0$.

   (c) The family of discrete probability mass functions

$$p(x; \theta) = \frac{1}{9} \qquad x \in \{0.1 + \theta, 0.2 + \theta, \ldots, 0.9 + \theta\} \qquad \theta \in \mathbb{R}$$

   (d) The $N(\theta, \theta^2)$ family, $\theta > 0$

   (e)

$$p(x; \theta) = \frac{2(x + \theta)}{1 + 2\theta} \qquad 0 < x < 1, \quad \theta > 0$$

   (f) $p(x, \theta)$ is the conditional probability mass function for a binomial(n, θ) variable, conditioned on $X > 0$. (recall binomial has probability mass function $\binom{n}{k}\theta^k(1-\theta)^{n-k}$).

3. The *inverse Gaussian* density $IG(\mu, \lambda)$, is:

$$f(x; \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{1/2} \frac{1}{x^{3/2}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\} \mathbf{1}_{\{x>0\}} \qquad \mu > 0, \lambda > 0.$$

   (a) Show that this is an exponential family generated by $T(X) = -\frac{1}{2}(X, \frac{1}{X})$ and $h(x) = \frac{1}{(2\pi)^{1/2} x^{3/2}}$.

(b) Show that the canonical parameters (η_1, η_2) are

$$\eta_1 = \frac{\lambda}{\mu^2}, \qquad \eta_2 = \lambda$$

and that the log partition function is:

$$A(\eta_1, \eta_2) = -\left(\frac{1}{2}\log(\eta_2) + \sqrt{\eta_1\eta_2}\right), \qquad \mathcal{E} = \mathbb{R}_+^2$$

(c) Find the moment generating function of T and show that

$$\mathbb{E}[X] = \mu, \quad \mathrm{Var}(X) = \frac{\mu^3}{\lambda}, \quad \mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mu} + \frac{1}{\lambda}, \quad \mathrm{Var}\left(\frac{1}{X}\right) = \frac{1}{\lambda\mu} + \frac{2}{\lambda^2}.$$

4. Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a canonical exponential family generated by (T, h) and $\mathcal{E}^0 \neq \phi$. Show that T is minimal sufficient.

5. Let $p(x, \eta)$ be a one parameter canonical exponential family generated by $T(x) = x$ and $h(x)$: $x \in \mathcal{X} \subset \mathbb{R}$. Let $\psi(x)$ be a non-constant, non-decreasing function. Show that $\mathbb{E}_\eta[\psi(X)]$ is strictly increasing in η.

   Hint: Let Cov denote covariance. Show that

   $$\mathrm{Cov}(X, Y) = \frac{1}{2}\mathbb{E}\left[(X - X')(Y - Y')\right]$$

   where (X, Y) and (X', Y') are independent, identically distributed. Compute $\frac{\partial}{\partial\eta}\mathbb{E}_\eta[\psi(X)]$ in terms of $\mathrm{Cov}_\eta(\psi(X), X)$.

6. **Logistic Regression** In the following, $Y_1, \ldots, Y_n$ are the outcomes of random experiments $i = 1, \ldots, n$. For experiment i, you fix the values of covariates $z_{i1}, \ldots, z_{id}$. For example, suppose you are trying to find a cure for Coronavirus. For trial i, you choose the quantities of d different chemicals; these quantities are $z_{i1}, \ldots, z_{id}$. There are unknown parameters $\beta_1, \ldots, \beta_d$. You run experiment i on n_i individuals (who are independent of each other) and Y_i represents the number who are successfully cured and $n_i - Y_i$ the number for whom the treatement is not successful.

   If the model is correct, then you would like estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ of the unknown parameters and, from this, estimate the success rate of the treatment for a given covariate vector $(z_1, \ldots, z_d)$.

   $(z_{1,.}, Y_1), \ldots, (z_{n,.}, Y_n)$ are observed, where $z_{1,.}, \ldots, z_{n,.}$ are d-row vectors and $Y_1, \ldots, Y_n$ are independent and $Y_j \sim \mathrm{Binomial}(n_j, \lambda_j)$. The success probability λ_j depends on the vector $z_{j,.}$. The function

   $$l(u) = \log\frac{u}{1-u}$$

   is called the *logit* function. In logistic regression, it is assumed that

   $$l(\lambda_i) = z_{i,.}\beta$$

where $\beta = (\beta_1, \ldots, \beta_d)^t$ is the parameter vector.

Show that $Y = (Y_1, \ldots, Y_n)$ is an exponential family of rank d if and only if $z_{.,1} \ldots, z_{.,d}$ are linearly independent.

Note The family is of rank k if and only if

$$\mathbb{P}\left(\sum_{j=1}^{k} c_j T_j(X) = c_0\right) < 1$$

for all $(c_0, c_1, \ldots, c_k) \not\equiv 0$.

7. Let $(X_1, X_2, \ldots X_n)$ be a *stationary Markov chain* with two states, 0 and 1. That is,

$$\mathbb{P}(X_i = x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}) = \mathbb{P}(X_i = x_i | X_{i-1} = x_{i-1}) = p_{x_{i-1}, x_i}$$

where $\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$ is the matrix of *transition probabilities.* Suppose, furthermore, that

- $p_{00} = p_{11} = p$, so that $p_{10} = p_{01} = 1 - p$,
- $\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = \frac{1}{2}$.

(a) Show that if $0 < p < 1$ is unknown, this is a full-rank one-parameter exponential family with $T = N_{00} + N_{11}$, where N_{ij} denotes the number of transitions from i to j. For example, the sequence 01011 has $N_{01} = 2$, $N_{11} = 1$, $N_{00} = 0$ and $N_{10} = 1$.

(b) Show that $\mathbb{E}[T] = (n-1)p$.

8. Let $X = (Z, Y)$ where $Y = Z + \theta W$, $\theta > 0$, Z and W are independent $N(0, 1)$ variables. Let $X_1, \ldots, X_n$ be i.i.d. as X. Write the density of $X_1, \ldots, X_n$ as a canonical exponential family and identify T, h, η, A and $\mathcal{E}$. Find the expected value and variance of the sufficient statistic.

9. The *entropy* $h(p)$ of a random variable X with density p is defined by:

$$h(p) = \mathbb{E}\left[-\log p(X)\right] = -\int_S p(x) \log p(x) dx.$$

where $S = \{x : p(x) > 0\}$.

(a) Show that the canonical k parameter exponential family density

$$p(x, \eta) = \exp\left\{\sum_{j=1}^{k} \eta_j r_j(x) - A(\eta)\right\} \qquad x \in S$$

maximises $h(p)$ subject to the constraints

$$p(x) \geq 0, \quad \int_S p(x) dx = 1, \quad \int_S p(x) r_j(x) dx = \alpha_j, \quad 1 \leq j \leq k$$

for given $\alpha_1, \ldots, \alpha_k$ for which a solution exists, where $\eta_1, \ldots, \eta_k$ are chosen so that p satisfies the constraints.

Hint This is very easy using Lagrange multipliers; maximise the integrand

(b) Find the maximum entropy densities when $r_j(x) = x^j$ in the following cases:

i. $S = (0, +\infty), \quad \alpha_1 > 0$

ii. $S = \mathbb{R}$, $k = 2$, $\alpha_1 \in \mathbb{R}, \quad \alpha_2 \in \mathbb{R}_+$

iii. $S = \mathbb{R}$, $k = 3$, $\alpha_1 \in \mathbb{R}$, $\alpha_2 > 0$, $\alpha_3 \in \mathbb{R}$.

10. Suppose that $p(x, \theta)$ is a positive density on the real line, which is continuous in x for each θ and such that if X_1, X_2 is a sample of size 2 from $p(., \theta)$ then $X_1 + X_2$ is sufficient for θ. Show that $p(., \theta)$ corresponds to a one-parameter exponential family of distributions with $T(x) = x$.

Answers

1. (a)

$$p(x;\beta_1,\beta_2) = \exp\left\{(\beta_1-1)\log x + (\beta_2-1)\log(1-x) - \log\frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2)}\right\}$$

$$h(x)=1, \quad T_1(x)=\log x, \quad \eta_1(\beta)=\beta_1-1, \quad T_2(x)=\log(1-x), \quad \eta_2(\beta)=\beta_2-1,$$

$$B(\beta_1,\beta_2)=\log\frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2)}$$

(b)

$$p(x;\alpha,\lambda)=\exp\left\{(\alpha-1)\log(x)-\lambda x-(\log\Gamma(\alpha)-\alpha\log(\lambda)\right\}$$

$$h(x)\equiv 1, \quad T_1(x)=\log(x), \quad \eta_1(\theta)=(\alpha-1), \quad T_2(x)=x, \quad \eta_2(\theta)=-\lambda$$

$$B(\theta)=\log\Gamma(\alpha)-\alpha\log(\lambda)$$

2. (a) no: $p(x;\theta)=\theta^{-1}\mathbf{1}_{[0,\theta]}(x)$. For an exponential family:

$$p(x;\theta)=h(x)\exp\{\eta(\theta)T(x)-B(\theta)\}$$

so that

$$h(x)=\exp\{-\eta(\theta)T(x)+B(\theta)-\log\theta\}\mathbf{1}_{[0,\theta]}(x)$$

so that $h(x)=0$ for all $x>\theta$. Since h does *not* depend on θ, $x\in[0,1]$ and $\Theta=(0,+\infty)$, hence $h(x)=0$ for all $x>\theta$ for all $\theta>0$, hence $h(x)=0$ for all $x>0$, so that $p(x;\theta)\equiv 0$, which is a contradiction.

(b) no: same as for (a): assume it is exponential family then:

$$h(x)e^{(T(x),\eta(\theta))-B(\theta)}=\mathbf{1}_{[0,\theta]}(x)\exp\{-2\log\theta+\log(2x)\}$$

so that

$$h(x)=\mathbf{1}_{[0,\theta]}(x)\exp\{B(\theta)-2\log\theta+\log(2x)-(T(x),\eta(\theta))\}$$

Here $\Theta=(0,+\infty)$ and $h(x)=0$ for all $x>\theta$. This holds for all $\theta>0$ hence $h(x)\equiv 0$ so that $p(x;\theta)\equiv 0$ which is a contradiction.

(c) no;

$$p(x;\theta)=\frac{1}{9}\sum_{j=1}^{9}\mathbf{1}_{0.1j}(x-\theta)=p(x-\theta;0)$$

so that

$$h(x)\exp\{\eta(\theta)T(x)-B(\theta)\}=h(x-\theta)\exp\{\eta(0)T(x-\theta)-B(0)\}.$$

If we take $\theta = 0$, we see that $h(x)$ has support (i.e. is non-zero for) $x \in \{0.1, 0.2, \ldots, 0.9\}$. That is, $h(x) = 0$ for any x which does not belong to this set of values. Now, let us consider *arbitrary* θ, we see that $h(x)$ has support $\{0.1 + \theta, \ldots, 0.9 + \theta\}$; $h(x) = 0$ for any x which does not belong to this set of values. Since $\Theta = \mathbb{R}$, therefore so that $h \equiv 0$, which gives a contradiction.

(d)

$$p(x;\theta) = \exp\left\{-\frac{x^2}{2\theta^2} + \frac{x}{\theta} - \frac{1}{2} - \frac{1}{2}\log(2\pi) - \log\theta\right\}$$

This does (technically) satisfy the definition of an exponential family, so the answer is YES. Note, however, that Θ is one-dimensional, yet we need a two-dimensional sufficient statistic $(T_1(x), T_2(x)) = (-x^2, x)$ and a two functions $\eta_1(\theta) = \frac{1}{2\theta^2}$ and $\eta_2(\theta) = \theta$. This is known as a *curved* exponential family.

(e)

$$p(x;\theta) = 2\exp\{\log(x+\theta) - \log(1+2\theta)\}\mathbf{1}_{[0,1]}(x)$$

no: for an exponential family (taking logarithms of the density):

$$\log p(x;\theta) = \log h(x) + \eta(\theta)T(x) - B(\theta)$$

To show that, in this case it doesn't hold, we can take $\frac{\partial^2}{\partial\theta\partial x}$ to get rid of the $\log h(x)$ and $B(\theta)$ term to get (using $\log p(x;\theta) = \log 2 + \log(x+\theta) - \log(1+2\theta)$):

$$\frac{\partial^2}{\partial x\partial\theta}\log p(x;\theta) = \eta'(\theta)T'(x) = -(x+\theta)^{-2} \qquad x \in (0,1).$$

Since the right hand side is well defined, hence $\frac{\partial^2}{\partial x\partial\theta}\log p(x;\theta)$ is well defined. From the form of the exponential family, it is therefore clear that the derivatives on the right hand side are well defined.

$$(x+\theta)^2 = \frac{1}{\eta'(\theta)}\frac{1}{T'(x)}$$

From this (setting $\theta = 1$ - or, since the derivative is defined Lebesgue almost surely, choose a particular θ for which it is well defined)

$$T'(x) = \frac{1}{\eta'(1)(1+x)^2}$$
$$\eta'(\theta) = \frac{1}{T'(\frac{1}{2})(\theta+\frac{1}{2})^2}$$

Hence

$$(x+\theta)^2 = \eta'(1)T'(\frac{1}{2})(1+x)^2(\theta+\frac{1}{2})^2.$$

This must hold for all (x,θ). Clearly it does not, so we obtain a clear contradiction.

(f)

$$\mathbb{P}_\theta(X > 0) = 1 - (1-\theta)^n$$

$$p(x,\theta) = \frac{1}{1-(1-\theta)^n}\binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}\exp\left\{x\log\frac{\theta}{1-\theta} + n\log(1-\theta) - \log(1-(1-\theta)^n)\right\}$$

yes

3. (a) Comes from expanding

$$-\frac{\lambda(x-\mu)^2}{2\mu^2 x} = -\frac{\lambda x}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2x}$$

which gives sufficient statistic $-\frac{1}{2}(x, \frac{1}{x})$ and canonical coordinates $(\eta_1, \eta_2) = (\frac{\lambda}{\mu^2}, \lambda)$. The $h(x) = \frac{1}{(2\pi)^{1/2}x^{3/2}}$ comes directly from the first part of the expression for the density and the log partition function is:

$$B(\mu,\lambda) = -\frac{\lambda}{\mu} - \frac{1}{2}\log\lambda.$$

(b) For $T(x) = -\frac{1}{2}(x, \frac{1}{x})$, the above expansion also gives $\eta_1 = \frac{\lambda}{\mu^2}$, $\eta_2 = \lambda$ and

$$A(\eta_1,\eta_2) = -\frac{\lambda}{\mu} - \frac{1}{2}\log\lambda = -\sqrt{\eta_1\eta_2} - \frac{1}{2}\log\eta_2.$$

(c) Using $M_T(s) = \exp\{A(\eta+s) - A(\eta)\}$ we have:

$$M_{T;\eta}(s_1,s_2) = \left(\frac{\eta_2}{\eta_2+s}\right)^{1/2}\exp\left\{\sqrt{\eta_1\eta_2} - \sqrt{(\eta_1+s_1)(\eta_2+s_2)}\right\}$$

To compute expectations and variances, use $\dot{A}(\eta) = \mathbb{E}_\eta[T]$ and $\ddot{A}(\eta) = \Sigma_T$.

$$\dot{A}(\eta_1,\eta_2) = -\begin{pmatrix} \frac{1}{2}\frac{\eta_2^{1/2}}{\eta_1^{1/2}} \\ \frac{1}{2}\frac{\eta_1^{1/2}}{\eta_2^{1/2}} + \frac{1}{2\eta_2} \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}\left[\frac{1}{X}\right] \end{pmatrix}$$

$$\mathbb{E}[X] = \mu, \qquad \mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mu} + \frac{1}{\lambda}$$

$$\ddot{A}(\eta_1,\eta_2) = \frac{1}{4}\begin{pmatrix} \eta_1^{-3/2}\eta_2^{1/2} & -\eta_1^{-1/2}\eta_2^{-1/2} \\ -\eta_1^{-1/2}\eta_2^{-1/2} & \eta_1^{1/2}\eta_2^{-3/2} + \frac{2}{\eta_2} \end{pmatrix}$$

$$\mathrm{Var}(X) = \frac{\mu^3}{4\lambda} \qquad \mathrm{Var}\left(\frac{1}{X}\right) = \frac{1}{\mu\lambda} + \frac{2}{\lambda^2}.$$

4.

$$p(x,\theta) = h(x)\exp\left\{\sum_{j=1}^k T_j(x)\theta_j - A(\theta)\right\}$$

$$\log L(\theta,x) - \log L(\theta,y) = (\log h(x) - \log h(y)) + \sum_{j=1}^k (T_j(x) - T_j(y))\theta_j$$

clearly does not depend on θ if and only if $T(x) = T(y)$.

5. For a one-parameter exponential family,

$$p(x;\eta) = h(x)\exp\{\eta T(x) - A(\eta)\}$$

so that

$$\begin{aligned}\frac{\partial}{\partial\eta}\mathbb{E}_\eta[\psi(X)] &= \frac{\partial}{\partial\eta}\int h(x)e^{\eta T(x)-A(\eta)}\psi(x)dx = \int h(x)\left(\frac{d}{d\eta}e^{\eta T(x)-A(\eta)}\right)\psi(x)dx \\ &= \int h(x)e^{\eta T(x)-A(\eta)}(T(x)-\dot{A}(\eta))\psi(x)dx \\ &= \int p(x;\eta)T(x)\psi(x)dx - \dot{A}(\eta)\int p(x;\eta)\psi(x)dx\end{aligned}$$

and therefore, using $\dot{A}(\eta) = \mathbb{E}_\eta[T(X)]$:

$$\frac{\partial}{\partial\eta}\mathbb{E}_\eta[\psi(X)] = \mathbb{E}_\eta[X\psi(X)] - \mathbb{E}_\eta[\psi(X)]\mathbb{E}_\eta[X] = \mathrm{Cov}(X,\psi(X)).$$

Under the conditions placed on ψ, $(x-y)(\psi(x)-\psi(y))$ is non negative and positive with positive probability. The result follows.

6. Using the notations of the question, and setting $z_{j.} = (z_{j1},\ldots,z_{jd})^t$,

$$p(y_1,\ldots,y_n,\underline{\beta}) = \left(\prod_{j=1}^n \binom{n_j}{y_j}\right)\exp\left\{\sum_{i=1}^d \beta_i\left(\sum_{j=1}^n y_j z_{ji}\right) - \sum_{j=1}^n n_j\log(1-\lambda_j)\right\}$$

The family is of rank k if and only if

$$\mathbb{P}(\sum_{j=1}^k c_j T_j(X) = c_0) < 1$$

for all $(c_0, c_1, \ldots, c_k)$. Here

$$\mathbb{P}\left(\sum_{i=1}^d c_i T_i(Y) = c_0\right) = \mathbb{P}\left(\sum_{j=1}^n\left(\sum_{i=1}^d c_i z_{ji}\right)Y_j = c_0\right)$$

If the (column) vectors $z_{.1},\ldots,z_{.d}$ are not linearly independent, then (by definition) $c_1,\ldots,c_d$ may be found so that $\sum_{i=1}^d c_i z_{.,i} = 0$.

If they are linearly independent, then clearly the family is of rank d.

7. (a)

$$\mathbb{P}((X_1,\ldots,X_n) = (x_1,\ldots,x_n)) = \frac{1}{2}p_{00}^{n_{00}}p_{01}^{n_{01}}p_{10}^{n_{10}}p_{11}^{n_{11}}$$

where $n_{00}+n_{01}+n_{10}+n_{11}=n-1$, the total number of transitions. It follows that

$$\begin{aligned}\mathbb{P}((X_1,\ldots,X_n)=(x_1,\ldots,x_n)) &= \frac{1}{2}\exp\left\{(n_{00}+n_{11})\log p+(n_{01}+n_{10})\log(1-p)\right\}\\ &= \frac{1}{2}\exp\left\{(n_{00}+n_{11})\log\left(\frac{p}{1-p}\right)+(n-1)\log(1-p)\right\}\end{aligned}$$

The result now follows from the formula for an exponential family; $h(x)=\frac{1}{2}$, $T(x)=n_{00}+n_{11}$, $\eta(p)=\log\left(\frac{p}{1-p}\right)$, $B(p)=-(n-1)\log(1-p)$.

(b) Let $Y_i=1$ if transition i is either $0\mapsto 0$ or $1\mapsto 1$ and let $Y_i=0$ otherwise. Then

$$T=Y_1+\ldots Y_{n-1}.$$

Since $\mathbb{E}[Y_j]=p$, the result follows.

8. $X=\binom{Z}{Y}\sim N\left(\left(\begin{array}{c}0\\0\end{array}\right),\left(\begin{array}{cc}1&1\\1&1+\theta^2\end{array}\right)\right)$. Covariance matrix is $\Sigma=\left(\begin{array}{cc}1&1\\1&1+\theta^2\end{array}\right)$ so $|\Sigma|=\theta^2$ and $\Sigma^{-1}=\frac{1}{\theta^2}\left(\begin{array}{cc}1+\theta^2&-1\\-1&1\end{array}\right)$. It follows that

$$f_{(Z,Y)}(z,y)=\frac{1}{2\pi|\theta|}\exp\left\{-\frac{1}{2\theta^2}(z^2+(1+\theta^2)y^2-2zy)\right\}$$

giving:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\frac{1}{(2\pi)^n|\theta|^n}\exp\left\{-\frac{1}{2\theta^2}\sum_{j=1}^n(z_j-y_j)^2-\frac{1}{2}\sum_{j=1}^n y_j^2\right\}$$

so that

$$T(x_1,\ldots,x_n)=\sum_{j=1}^n(z_j-y_j)^2,\qquad h(x_1,\ldots,x_n)=\frac{1}{(2\pi)^n}e^{-\sum_{j=1}^n y_j^2},\qquad \eta=-\frac{1}{2\theta^2}$$

$$A(\eta)=-\frac{n}{2}\log\frac{1}{\theta^2}=-\frac{n}{2}\log(-2\eta),\qquad \mathcal{E}=(0,+\infty)$$

Hence

$$\mathbb{E}_\eta[T]=\frac{dA}{d\eta}=-\frac{n}{2\eta}=n\theta^2$$

$$\mathrm{Var}_\eta(T)=\frac{d^2A}{d\eta^2}=\frac{n}{2\eta^2}=2n\theta^4$$

9. (a) Lagrange method of multipliers: if we maximise the integrand pointwise, then this maximises the integral. Maximise

$$-p(x)\log p(x)-\lambda_0 p(x)-\sum_{j=1}^k p(x)r_j(x)\lambda_j$$

then choose $\lambda_0, \lambda_1, \ldots, \lambda_k$ to satisfy constraints. Taking derivative w.r.t. $p(x)$, maximum satisfies:

$$-\log p(x) - 1 - \lambda_0 - \sum_{j=1}^{k} r_j(x)\lambda_j = 0$$

so that p is of the form:

$$p(x) = \exp\left\{-(1+\lambda_0) - \sum_{j=1}^{k} \lambda_j r_j(x)\right\}$$

Choose $\lambda_0, \lambda_1, \ldots, \lambda_k$ so that the constraints are satisfied. For an exponential family, this is clearly the case if $\lambda_j = -\eta_j$ for $j = 1, \ldots, k$ and $A(\eta) = 1 + \lambda_0$.

(b) i. $p(x) = \frac{1}{\alpha_1}\exp\{-x/\alpha_1\} \qquad x \in (0, +\infty)$

ii.

$$p(x) = \exp\left\{\eta_1 x + \eta_2 x^2 - A(\eta)\right\}$$

No solution for $\alpha_2 < \alpha_1^2$; this would require random variables which satisfy: $\mathbb{E}[X^2] < \mathbb{E}[X]^2$. It follows that α_2 satisfies $\alpha_2 > \alpha_1^2$. Set $\sigma^2 = \alpha_2 - \alpha_1^2$, then

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\alpha_1)^2}{2\sigma^2}\right\} \qquad -\infty < x < +\infty$$

iii.

$$p(x) = \exp\left\{\eta_1 x + \eta_2 x^2 + \eta_3 x^3 - A(\eta)\right\} \qquad -\infty < x < +\infty$$

Clearly it doesn't exist!

10. It follows from the factorisation theorem that

$$p(x_1, \theta)p(x_2, \theta) = h(x_1, x_2)g(x_1 + x_2, \theta).$$

Fix a point θ_0 and let $r(x, \theta) = \log p(x, \theta) - \log p(x, \theta_0)$. Let $q(z, \theta) = \log g(z, \theta) - \log g(z, \theta_0)$. Then

$$r(x_1, \theta) + r(x_2, \theta) = q(x_1 + x_2, \theta)$$

so that $r(., \theta)$ and $q(., \theta)$ are linear in x;

$$r(x, \theta) = a(\theta) + b(\theta)x.$$

It follows that

$$p(x, \theta) = p(x, \theta_0)\exp\left\{a(\theta) + b(\theta)x\right\}$$

Let $h(x) = p(x, \theta_0)$, then this density is an exponential family with $T(x) = x$.

Establishing linearity in x The density is continuous and positive, hence so are r and q. Since $q(x_1 + x_2) = r(x_1) + r(x_2)$, it follows that $q(x) = r(x) + r(0)$ so that $q(0) = 2r(0)$ and $q(x_1 + x_2) = q(x_1) + q(x_2) - q(0)$. Now set $f(x) = q(x) - q(0)$ so that

$$f(x_1 + x_2) = f(x_1) + f(x_2).$$

It follows that for any $x_1, \ldots, x_n$,

$$f(x_1 + \ldots + x_n) = f(x_1) + \ldots + f(x_n).$$

In particular,

$$f(1) = nf\left(\frac{1}{n}\right) \Rightarrow f\left(\frac{1}{n}\right) = \frac{1}{n}f(1)$$

and

$$f\left(\frac{k}{n}\right) = \frac{k}{n}f(1).$$

It follows that for x rational, $f(x) = xf(1)$ and hence, by continuity, it follows that $f(x) = xf(1)$ for all x. It follows that $q(x) = a + bx$ for constants a and b and hence that $r(x) = \frac{a}{2} + bx$.

Chapter 4

Parameter Estimation

4.1 Parameter Estimation: Basic Heuristics

Consider a parametric family $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Let $X = (X_1, \ldots, X_n)$ denote a random sample from distribution $\mathbb{P}_\theta$, where θ is unknown. The aim of estimation is to find a function $\widehat{\theta}$ such that $\widehat{\theta}$ is 'close' (in an appropriate sense) to the parameter θ.

Minimum Contrast Estimates Consider a function

$$\rho : \mathcal{X} \times \Theta \to \mathbb{R}$$

and define

$$D(\theta_0, \theta) := \mathbb{E}_{\theta_0}\left[\rho(X, \theta)\right].$$

Suppose that ρ is such that $D(\theta_0, \theta)$ is uniquely minimised for $\theta = \theta_0$. The *minimum contrast* estimate for θ from ρ is $\widehat{\theta}(X)$, the value of θ that minimises $\rho(X, \theta)$. The function ρ is a *contrast function.*

Estimating Equation Estimate Suppose that $\Theta \subset \mathbb{R}^d$. Let $\Psi : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^d$. Set

$$V(\theta_0, \theta) = \mathbb{E}_{\theta_0}\left[\Psi(X, \theta)\right]$$

and suppose that $V(\theta_0, \theta) = 0$ has unique solution $\theta = \theta_0$ for all $\theta_0 \in \Theta$. Then $\widehat{\theta}$ solving

$$\Psi(X, \widehat{\theta}) = 0$$

is an *estimating equation estimate.*

4.2 Unbiased Estimation

Definition 4.1 (Bias)**.** *Let $\mathcal{P} := \{\mathbb{P} : \theta \in \Theta\}$ be a parametric family. Let $\delta(X)$ be the estimator of a parameter $q(\theta)$, where $q : \Theta \to \mathbb{R}$. The* bias *of an estimator is defined as:*

$$Bias_\theta(\delta) = \mathbb{E}_\theta\left[\delta(X)\right] - q(\theta).$$

An estimator that satisfies $Bias_\theta(\delta) = 0$ for all $\theta \in \Theta$ is called unbaised.

Example 4.1 (Unbiased Estimates)**.**

Let $\mathcal{X} = (x_1, \ldots, x_N)$, a sample space with N elements taken from $\mathbb{R}$. Let $(X_1, \ldots, X_n)$ be a random sample, drawn at random from $\mathcal{X}$, without replacement. Then

$$\mathbb{P}_{\mathcal{X}}(\{X_1, \ldots, X_n\} = \{a_1, \ldots, a_n\}) = \begin{cases} \frac{1}{\binom{N}{n}} & \{a_1, \ldots, a_n\} \subset \mathcal{X} \\ 0 & \text{otherwise} \end{cases}$$

where the notation $\{\}$ denotes a set (i.e. the collection of elements, without respect to ordering) and $\{a_1, \ldots, a_n\}$ (of course) contains n distinct elements.

Suppose we want to estimate $\overline{x} = \frac{1}{N}\sum_{j=1}^N x_j$. Then $\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$ is an unbiased estimator (exercise) with variance

$$\text{Var}(\overline{X}) = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)s^2$$

where $s^2 = \frac{1}{N}\sum_{j=1}^N (x_j - \overline{x})^2$ (exercise).
Hint: Note that for $i \neq j$, $\text{Cov}(X_i, X_j) = \text{Cov}(X_1, X_2)$. Therefore

$$\text{Var}(\overline{X}) = \frac{1}{n}\text{Var}(X_1) + \frac{n-1}{n}\text{Cov}(X_1, X_2).$$

Note also that

$$0 = \text{Var}(X_1 + \ldots + X_N) = N\text{Var}(X_1) + N(N-1)\text{Cov}(X_1, X_2).$$

Clearly $\text{Var}(X_1) = \frac{1}{N}\sum_{j=1}^N (x_j - \overline{x})^2$. □

4.3 Least Squares

Suppose that $Y_1, \ldots, Y_n$ are random variables satisfying $\mathbb{E}[Y_j] = \mu(z_{j.}) = g(\beta, z_{j.})$ where $\beta \in \mathbb{R}^d$. Let $X = ((Y_1, z_{1.}), \ldots, (Y_n, z_{n,.}))$. A natural function to consider is:

$$\rho(X, \beta) = |Y - \mu|^2 = \sum_{i=1}^n (Y_i - g(\beta, z_{i.}))^2.$$

Now suppose that $Y_j - \mu(z_{j,.})$ are i.i.d. $N(0, \sigma^2)$. Then

$$\begin{aligned} D(\beta_0, \beta) &= \mathbb{E}_{\beta_0}\left[\rho(X, \beta)\right] \\ &= \mathbb{E}_{\beta_0}\left[\sum_{j=1}^n (Y_i - g(\beta_0, z_{i.}) + g(\beta_0, z_{i.}) - g(\beta.z_{i.}))^2\right] \\ &= n\sigma^2 + \sum_{j=1}^n \left(g(\beta_0, z_{j,.}) - g(\beta, z_{j.})\right)^2. \end{aligned}$$

This is minimised at $\beta = \beta_0$ and uniquely so if and only if the parametrisation is identifiable. An estimate that minimises $\rho(X, \beta)$ exists if $g(\beta, z)$ is continuous in β and if

$$\lim_{|\beta| \to +\infty} |g(\beta, z)| = +\infty.$$

This estimate is known as the *least squares estimate.*

If, further, $g(\beta, z)$ is differentiable in β, then $\widehat{\beta}$ satisfies:

$$\nabla_\beta \rho(X, \widehat{\beta}) = 0.$$

This is equivalent to the system of estimating equations:

$$\sum_{i=1}^n \frac{\partial g}{\partial \beta_j}(\widehat{\beta}, z_{i.}) Y_i = \sum_{i=1}^n \frac{\partial g}{\partial \beta_j}(\widehat{\beta}, z_{i.}) g(\widehat{\beta}, z_i) \qquad 1 \leq j \leq d$$

Consider the important linear case:

$$g(\beta, z_i) = \sum_{j=1}^d z_{ij} \beta_j \qquad z_i = (z_{i1}, \ldots, z_{id}).$$

The system becomes:

$$\sum_{i=1}^n z_{ij} Y_i = \sum_{k=1}^d \left(\sum_{i=1}^n z_{ij} z_{ik} \right) \widehat{\beta}_k.$$

These are known as the *normal equations*, which are usually written in matrix form:

$$Z^t Y = Z^t Z \beta$$

where Z is the *design matrix.*

4.4 Method of Moments (MOM)

Suppose $X_1, \ldots, X_n$ are i.i.d. from a family $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ and that the parametrisation is identifiable. Suppose that $\mu_1(\theta), \ldots, \mu_d(\theta)$ are the first d moments of the population from which we are sampling and that they are well defined. That is $\mathbb{E}[|X|^j] < +\infty$ for each $j = 1, \ldots, d$ and

$$\mu_j(\theta) = \mu_j = \mathbb{E}[X^j].$$

Define the jth sampling moment as:

$$\widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j \qquad 1 \leq j \leq d.$$

To apply the method of moments to the problem of estimating θ, it is necessary that θ can be expressed as a continuous function of the first d moments. Suppose

$$\theta \to (\mu_1(\theta), \dots, \mu_d(\theta))$$

is $1-1$ from $\mathbb{R}^d$ to $\mathbb{R}^d$. The method of moment estimate $\widehat{\theta}$ is the solution to the problem:

$$\widehat{\mu}_j = \mu_j(\widehat{\theta}).$$

The motivation is from the law of large numbers; if μ_j is well defined, then $\widehat{\mu}_j \to_p \mu_j$.

More generally, to estimate a $\mathbb{R}^k$ valued function $q(\theta)$ of θ, a MOM estimator may be obtained by expressing $q(\theta)$ as a function of any of the first d moments $\mu_1, \dots, \mu_d$ of X, say $q(\theta) = h(\mu_1, \dots, \mu_d)$, $d \geq k$ and then using $h(\widehat{\mu}_1, \dots, \widehat{\mu}_d)$ as the estimate of $q(\theta)$.

Method of Moments as Minimum Contrast The *method of moments* method can be seen as an *estimating equation estimate.*

Suppose there are d parameters; $\theta = (\theta_1, \dots, \theta_d)$. Let $X = (X_1, \dots, X_n)$ where $X_1, \dots, X_n$ are i.i.d. observations. Let $m_p(\theta) = \mathbb{E}_\theta[X_1^p]$

$$\Psi(X.\theta) = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^n X_j - m_1(\theta) \\ \vdots \\ \frac{1}{n}\sum_{j=1}^n X_j^d - m_d(\theta) \end{pmatrix}.$$

Then $\widehat{\theta}_{MM}$ satisfies $\Psi(X, \widehat{\theta}_{MM}) = 0$ and $V(\theta_0, \theta_0) := \mathbb{E}_{\theta_0}[\Psi(X, \theta)]|_{\theta=\theta_0} = 0$.

Additional conditions are required to show that $\theta = \theta_0$ is the *unique* solution to $V(\theta_0, \theta) = 0$. □

Example 4.2 (MOM method for gamma distribution)**.**

Suppose $X \sim \text{Gamma}(\alpha, \lambda)$. That is,

$$p(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\lambda x\} \qquad x > 0, \quad \alpha > 0, \quad \lambda > 0.$$

Here $\theta = (\alpha, \lambda)$, $\mu_1 = \mathbb{E}[X] = \frac{\alpha}{\lambda}$, $\mu_2 = \mathbb{E}[X^2] = \frac{\alpha(1+\alpha)}{\lambda^2}$. Solving for θ gives:

$$\alpha = \left(\frac{\mu_1}{\sigma}\right)^2 \qquad \lambda = \frac{\mu_1}{\sigma^2}$$

where $\sigma^2 = \mu_2 - \mu_1^2$. Then, using $\widehat{\sigma}^2 = \widehat{\mu}_2 - \widehat{\mu}_1^2$, the moment method estimates are:

$$\widehat{\alpha} = \left(\frac{\overline{X}}{\widehat{\sigma}}\right)^2 \qquad \widehat{\lambda} = \frac{\overline{X}}{\widehat{\sigma}^2}.$$

4.5 Maximum Likelihood Method

The method of maximum likelihood was first proposed by Carl Friedrich Gauss in 1821, but was then forgotten. The approach is usually attributed to the English statistician R.A. Fisher, who proposed the idea in 1922 (in [4]). The approach only makes sense with *regular parametric models* (Definition 1.3).

Recall the definition of the likelihood function; $L(\theta, x) = p(x, \theta)$. The *maximum likelihood estimate* of a parameter θ, for given data x, is the value $\widehat{\theta}(x)$ of θ which maximises $L(\theta, x)$. For discrete variables, the MLE is the value of θ that gives the highest probability value to the observed data.

Example 4.3 (Estimating the size of a population)**.**

Suppose $X_1, \ldots, X_n$ are i.i.d. $U(\{1, 2, \ldots, \theta\})$ (uniform distribution on $\{1, 2, \ldots, \theta\}$ where θ is an unknown positive integer.

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n}\mathbf{1}(\max_j x_j \leq \theta).$$

This achieves its maximum at $\widehat{\theta}(x) = \max_j x_j$.

The *estimator* is $\widehat{\theta}(X_1, \ldots, X_n) = \max_j X_j$, while the *estimate* is obtained by plugging in the observed values; $\widehat{\theta}(x_1, \ldots, x_n) = \max_j x_j$.

Note that $\mathbb{E}[X] = \frac{1+\theta}{2}$ so that $\widehat{\theta}_{MOM} = 2\overline{x} - 1$. Clearly, the method of moments estimator is not satisfactory when $2\overline{x} - 1 < \max_j x_j$, since none of the observed values can be greater than θ.

4.5.1 Maximum Likelihood as Minimum Contrast

Set

$$\log L(\theta; x) = \log L(\theta; x) = \log p(x, \theta).$$

This is known as the *log likelihood.* The problem of maximising the likelihood is equivalent to that of maximising the log likelihood. When the family under consideration is an exponential family, the log likelihood takes a particularly convenient form. It is also useful for random sampling. For $p(x_1, \ldots, x_n; \theta) = \prod_{j=1}^n p(x_j; \theta)$,

$$\log L(\theta; x_1, \ldots, x_n) = \sum_{j=1}^n \log L(\theta; x_j).$$

There is also an important connection with information theory. Set $\rho(x, \theta) = -\log L(\theta, x)$ as a contrast, then $\widehat{\theta}_{MLE}(x)$ is the minimum contrast estimate. In this case,

$$D(\theta_0, \theta) = -\mathbb{E}_{\theta_0}\left[\log p(X, \theta)\right].$$

Note that

$$D(\theta_0, \theta) - D(\theta_0, \theta_0) = -\mathbb{E}_{\theta_0}\left[\log \frac{p(X, \theta)}{p(X, \theta_0)}\right] \geq -\log \int p(x, \theta_0)\frac{p(x, \theta)}{p(x, \theta_0)}dx = -\log 1 = 0$$

where the inequality is by Jensen's inequality and is strict if $\mathbb{P}_{\theta_0} \neq \mathbb{P}_\theta$. Here $D(\theta_0, \theta_0)$ is the *entropy* of X. This has been written for a family of distributions with density over $\mathbb{R}^d$; replace integrals with sums if the family is of probability mass functions over a discrete state space $\mathcal{X}$.

Definition 4.2 (Entropy, Kullback-Leibler Divergence)**.** *Let p be a probability mass function or density function. The* entropy *of p is defined as*

$$\mathcal{E}(p) = -\mathbb{E}\left[\log p(X)\right]$$

where X is a random variable with distribution p.

Let p_0 and p_1 be two probability mass functions or density functions defined on the same state space $\mathcal{X}$. The Kullback Leibler (or information) divergence $D_{KL}(p_0 \| p_1)$ is defined as:

$$D_{KL}(p_0 \| p_1) = -\mathbb{E}_0\left[\log \frac{p_1(X)}{p_0(X)}\right]$$

where $\mathbb{E}_0$ denotes expectation with respect to p_0. By convention, $\frac{0}{0} = 0$ and $0 \log 0 = 0$.

The following lemma is a straightforward consequence of Jensen's inequality:

Lemma 4.3 (Shannon [8](1948))**.** *The Kullback Leibler divergence $D_{KL}(p_0 \| p_1)$ satisfies $D_{KL}(p_0 \| p_1) \geq 0$ with equality if and only if $\{x : p_0(x) = p_1(x)\}$ has probability 1 under both $\mathbb{P}_0$ and $\mathbb{P}_1$.*

Proof The result is a straightforward consequence of Jensen's inequality and the basic argument has been given above. □

Lemma 4.3 shows that when $X_1, \ldots, X_n$ are i.i.d., the function

$$\rho(X, \theta) := -\frac{1}{n} \sum_{j=1}^{n} \log p(X_i, \theta)$$

satisfies the condition of being a contrast function and the MLE is the minimum contrast estimate.

Furthermore, let $\widehat{P}_n$ denote the empirical distribution formed from $\underline{X} = (X_1, \ldots, X_n)$, then

$$\widehat{P}_n(dx) = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j}(dx)$$

(place $\frac{1}{n}$ of a Dirac mass at each observation) and (clearly):

$$\rho(\underline{X}, \theta) = -\mathbb{E}_{\widehat{P}_n}\left[\log p(\underline{X}, \theta)\right].$$

Note that

$$D_{KL}(\widehat{P}_n \| \mathbb{P}_\theta) = \mathbb{E}_{\widehat{P}_n}\left[\log \frac{\widehat{P}_n(X)}{p(X, \theta)}\right] = \mathbb{E}_{\widehat{P}_n}\left[\log \widehat{P}_n(X)\right] + \rho(\underline{X}, \theta)$$

and, since $\mathbb{E}_{\widehat{P}_n}\left[\log \widehat{P}_n(X)\right]$ is a function of the data (and not a function of θ), therefore the MLE is the value of θ which minimizes the Kullback-Leibler divergence between $\widehat{P}_n$ and $\mathbb{P}_\theta$. □

Tutorial 5

1. Let $X_1, \ldots, X_n$ be i.i.d. $U(0, \theta)$ random variables where θ is an unknown parameter.

   (a) Let $T_1 = \frac{n+1}{n} \max_j X_j$. Compute $\mathbb{E}[T_1]$ and $\text{Var}(T_1)$.

   (b) Let $T_2 = \frac{2}{n} \sum_{j=1}^{n} X_j$. Compute $\mathbb{E}[T_2]$ and $\text{Var}(T_2)$.

   You should find that both T_1 and T_2 are unbiased estimators of θ (that is $\mathbb{E}[T_1] = \mathbb{E}[T_2] = \theta$, but that the variance of T_1 is substantially lower.

   (c) Show that $\max_j X_j$ is the maximum likelihood estimator of θ.

   (d) Show that $\frac{2}{n} \sum_{j=1}^{n} X_j$ is the Method of Moments estimator of θ (based on the first moment - the expectation).

2. Let X be a random variable with state space $\mathcal{X} = \{v_1, \ldots, v_k\}$, where $p_j = \mathbb{P}(X = v_j)$. Let $(X_1, \ldots, X_n)$ be a random sample from X. The *frequency plug-in principle* is simply the estimation procedure where $(p_1, \ldots, p_k)$ is estimated by $(\widehat{p}_1, \ldots, \widehat{p}_n) = \left(\frac{N_1}{n}, \ldots, \frac{N_k}{n}\right)$, where $N_j = \sum_{i=1}^{n} \mathbf{1}(X_i = v_j)$. The *extension* principle simply extends this, to estimating a continuous function $q(p_1, \ldots, p_k)$ by $q(\widehat{p}_1, \ldots, \widehat{p}_k)$, which is the *frequency substitution estimate.*

   Consider a population made up of three different types of individuals occurring in the Hardy-Weinberg proportions $\mathbb{P}_\theta(X = v_1) = \theta^2$, $\mathbb{P}_\theta(X = v_2) = 2\theta(1 - \theta)$ and $\mathbb{P}_\theta(X = v_3) = (1 - \theta)^2$ respectively.

   (a) Show that $T := \frac{N_1}{n} + \frac{N_2}{2n}$ is a frequency substitution estimate of θ.

   (b) Using part (a), find a frequency substitution estimate of the odds ratio $\frac{\theta}{1-\theta}$.

   (c) Suppose $v_1 = -1$, $v_2 = 0$ and $v_3 = 1$. By considering the first moment of X, show that T is a method of moment estimate of θ.

3. Let $X_1, \ldots, X_n$ be i.i.d., with Beta(β_1, β_2) distribution. Find the method of moments estimate of $\beta = (\beta_1, \beta_2)$ based on the first two moments.

4. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli trials, each with success probability θ. Let $\psi : \mathbb{R}^n \times (0, 1) \to \mathbb{R}$ be the function defined by:

$$\psi(X_1, \ldots, X_n, \theta) = \frac{1}{\theta} \sum_{j=1}^{n} X_j - \frac{1}{1 - \theta} \left(n - \sum_{j=1}^{n} X_j \right).$$

   Compute

$$V(\theta_0, \theta) = \mathbb{E}_{\theta_0} \left[\psi(X_1, \ldots, X_n, \theta) \right]$$

   and show that θ_0 is the unique solution of $V(\theta_0, \theta) = 0$. Compute the estimating equation estimate of θ.

5. **General method of moment estimates** Suppose $X_1, \ldots, X_n$ are i.i.d., as $X \sim \mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^d$ and θ identifiable. Let $g_1, \ldots, g_d$ be linearly independent functions and set

$$\mu_j(\theta) := \mathbb{E}_\theta \left[g_j(X)\right] \qquad \widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} g_j(X_i)$$

The moment method estimates are $\widehat{\theta}$ such that $\mu_j(\widehat{\theta}) = \mu_j$ for each $j = 1, \ldots, d$. Furthermore, for the parameters of a canonical exponential family, the moment method estimator of the parameter vector η is the moment method estimator based on the sufficient statistic T. Recall that for an exponential family in its canonical coordinates

$$\dot{A}(\eta) = \mathbb{E}_\eta \left[T(X)\right]$$

where $\dot{A}$ denotes the vector of partial derivatives.

Suppose that $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is a k-parameter exponential family generated by (h, T) where $T = (T_1, \ldots, T_k)$. Using $g_j = T_j$ in the above, find the method of moments estimates for the parameters in:

(a) The Rayleigh distribution:

$$p(x, \theta) = \left(\frac{x}{\theta^2}\right) \exp\left\{-\frac{x^2}{2\theta^2}\right\} \qquad x > 0, \quad \theta > 0,$$

(b) the Gamma distribution gamma(p, θ) where p is fixed.

6. When the data is not i.i.d., it may still be possible to express parameters as functions of moments and then use estimates based on replacing population moments with ‘sample’ moments.

Let $X_1, \ldots, X_n$ satisfy:

$$\begin{cases} X_i = \mu + e_i & i = 1, \ldots, n \\ e_i = \beta e_{i-1} + \epsilon_i & i = 1, \ldots, n, \qquad e_0 = 0. \end{cases}$$

where $\epsilon_1, \epsilon_2, \epsilon_3, \ldots$ are independent, identically distributed, $\mathbb{E}\left[\epsilon_j\right] = 0$ and $\mathrm{Var}(\epsilon_j) = \sigma^2$.

(a) Use $\mathbb{E}[X_i]$ to give a method of moments estimate of μ.

(b) Suppose $\mu = \mu_0$ and $\beta = b$ are fixed. Use $\mathbb{E}[U_i^2]$ where

$$U_i = \frac{X_i - \mu_0}{\left(\sum_{j=0}^{i-1} b^{2j}\right)^{1/2}}$$

to give a method of moments estimate of σ^2.

(c) If μ and σ^2 are fixed, can you give a method of moments estimate of β?

7. $X_1, \ldots, X_n$ random sample from distribution with density

$$f(x) = \theta x^{-\theta-1} \qquad x \geq 1$$

$\theta > 0$ unknown parameter

   (a) Compute $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$, the method of moments estimator.

   (b) Show that, for $\theta > 2$, $\hat{\theta}_n \rightarrow_p \theta$.

8. Let $\theta = (\theta_1, \theta_2)$ be a bivariate parameter. Suppose that X has state space $\mathcal{X}$, $T_1 : \mathcal{X} \rightarrow \mathbb{R}$ and $T_2 : \mathcal{X} \rightarrow \mathbb{R}$ are functions such that $T_1(X)$ is sufficient for θ_1 whenever θ_2 is fixed and known, whereas $T_2(X)$ is sufficient for θ_2 whenever θ_1 is fixed and known. Assume that $S = \{x | p(x, \theta) > 0\}$ does not depend on θ (p a density if X is a continuous variable, a probability function if it is discrete).

   (a) Show that if T_1 and T_2 do not depend on θ_2 and θ_1 respectively, then $(T_1(X), T_2(X))$ is sufficient for θ.

   (b) Give an example where $(T_1(X), T_2(X))$ is sufficient for θ, where $T_1(X)$ is sufficient for θ_1 whenever θ_2 is fixed and known, $T_2(X)$ is *not* sufficient for θ_2 whenever θ_1 is fixed and known.

9. **Censored geometric waiting times** Let

$$\mathbb{P}_\theta(X = k) = \theta^{k-1}(1 - \theta) \qquad k = 1, 2, \ldots$$

where $0 < \theta < 1$, where X is the time to failure. Suppose that we only record a time to failure if it occurs on or before time r, and record that it survives for longer than time r otherwise. Suppose we observe n individuals, m of which survive longer than time r and the failure times of the others are $Y_1, \ldots, Y_{n-m}$, where $1 \leq Y_j \leq r$ for $j = 1, \ldots, n-m$. Let $S = \sum_{j=1}^{n-m} Y_j$. Compute the maximum likelihood estimator of θ based on all the information.

10. **Maximum Likelihood (Normal)** Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where both μ and σ are unknown. Compute the maximum likelihood estimators for the pair (μ, σ^2). Are the estimators of μ and σ^2 unbiased?

Short Answers

1. (a)

$$\mathbb{P}(\max_j X_j \leq x) = \mathbb{P}(X_1 \leq x)^n = \begin{cases} 0 & x < 0 \\ \frac{x^n}{\theta^n} & 0 \leq x \leq \theta \\ 1 & x > \theta \end{cases}$$

so, setting $Y = \max_j X_j$, the density is

$$p_Y(x;\theta) = \frac{nx^{n-1}}{\theta^n}\mathbf{1}_{[0,\theta]}(x)$$

so that

$$\mathbb{E}[Y] = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$

Since $T_1 = \frac{n+1}{n}Y$, therefore $\mathbb{E}_\theta[T_1] = \theta$.

$$\mathbb{E}_\theta[Y^2] = \frac{n}{\theta^n}\int_0^\theta x^{n+1}dx = \frac{n}{n+2}\theta^2$$

Hence

$$\text{Var}_\theta(Y) = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n}{(n+2)(n+1)^2}\theta^2$$

so that

$$\text{Var}_\theta(T_1) = \left(\frac{n+1}{n}\right)^2 \text{Var}(Y) = \frac{1}{n(n+2)}\theta^2.$$

(b)

$$\mathbb{E}_\theta[T_2] = \frac{2}{n}\sum_{j=1}^n \mathbb{E}_\theta[X_j] = \frac{2}{n}\frac{\theta}{2} = \theta.$$

If $X \sim U(0,\theta)$ then $\text{Var}_\theta(X) = \frac{\theta^2}{12}$ so that

$$\text{Var}_\theta(T_2) = \frac{4}{n^2} \times n \times \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

(c) That $\max_j X_j$ is the maximum likelihood estimator of θ is seen as follows:

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n}\mathbf{1}_{[\max_j x_j, +\infty)}(\theta)$$

which (as a function of θ) is maximised by $\widehat{\theta}_{ML} = \max_j x_j$.

(d) That T_2 is the method of moments estimator of θ is seen as follows: $\mathbb{E}_\theta[X_1] = \frac{\theta}{2}$ so that $\theta = 2\mathbb{E}_\theta[X_1]$ and hence $\widehat{\theta}_{MM} = 2\overline{X} = \frac{2}{n}\sum_{j=1}^n X_j$.

2. (a) $p_1 = \theta^2$, $p_2 = 2\theta(1-\theta) = 2\theta - 2\theta^2 = 2\theta - 2p_1$ so that

$$\theta = p_1 + \frac{p_2}{2}$$

hence a frequency substitution estimate is:

$$\widehat{\theta} = \frac{N_1}{n} + \frac{N_2}{2n}.$$

(b) There are several answers. One answer: from before, estimate is:

$$\frac{\widehat{\theta}}{1-\widehat{\theta}} = \frac{2N_1 + N_2}{2n - 2N_1 - N_2}.$$

Another answer:

$$\frac{\theta}{1-\theta} = \frac{\theta^2}{\theta(1-\theta)} = \frac{p_1}{p_2/2} \qquad \frac{\widehat{\;\theta\;}}{1-\theta} = \frac{2N_1}{N_2}.$$

(c) $\mathbb{E}_\theta[X] = -\theta^2 + (1-\theta)^2 = 1 - 2\theta$

$$\widehat{\mu} = -\frac{N_1}{n} + \frac{N_3}{n} = 1 - \frac{N_2}{n} - \frac{2N_1}{n}.$$

Hence moment method estimate (based on first moment) is:

$$1 - 2\widehat{\theta} = 1 - \frac{N_2}{n} - \frac{2N_1}{n} \Rightarrow \widehat{\theta} = \frac{N_1}{n} + \frac{N_2}{2n}$$

as required.

3.

$$f(x) = \frac{\Gamma(\beta_1+\beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} x^{\beta_1}(1-x)^{\beta_2} \qquad x \in (0,1)$$

$$\mathbb{E}[X] = \frac{\Gamma(\beta_1+\beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \frac{\Gamma(\beta_1+1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2+1)} = \frac{\beta_1}{\beta_1+\beta_2}$$

$$\mathbb{E}[X^2] = \frac{\Gamma(\beta_1+\beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \int_0^1 x^{\beta_1+1}(1-x)^{\beta_2-1}dx = \frac{\Gamma(\beta_1+\beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)} \frac{\Gamma(\beta_1+2)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2+2)}$$

$$\mathbb{E}[X^2] = \frac{(\beta_1+1)\beta_1}{(\beta_1+\beta_2+1)(\beta_1+\beta_2)}$$

Let $m_1 = \mathbb{E}[X]$ and $m_2 = \mathbb{E}[X^2]$. Then from the first equation

$$\beta_2 = \beta_1\left(\frac{1}{m_1} - 1\right)$$

and from the second equation

$$m_2 = \frac{\beta_1+1}{\beta_1+\beta_2+1} m_1 \Rightarrow m_2 = \frac{m_1^2(\beta_1+1)}{\beta_1+m_1} \Rightarrow \beta_1 = \frac{m_1(1-m_2)}{m_2 - m_1^2}, \quad \beta_2 = \frac{(1-m_1)(1-m_2)}{m_2-m_1^2}$$

$$\widehat{\beta}_1 = \frac{\overline{x}(1-\overline{x^2})}{\overline{x^2} - \overline{x}^2} \qquad \widehat{\beta}_2 = \frac{(1-\overline{x})(1-\overline{x^2})}{\overline{x^2}-\overline{x}^2}.$$

4.
$$\mathbb{E}_{\theta_0}\left[\psi(X_1,\ldots,X_n,\theta)\right] = n\left(\frac{\theta_0}{\theta} - \frac{1-\theta_0}{1-\theta}\right) = n\frac{\theta_0-\theta}{\theta(1-\theta)}$$

which is 0 if and only if $\theta = \theta_0$ as required.

Estimating equation estimate satisfies

$$\psi(X_1,\ldots,X_n,\widehat{\theta}) = 0 = \frac{1}{\widehat{\theta}}\sum_{j=1}^{n} X_j - \frac{1}{1-\widehat{\theta}}\left(n - \sum_{j=1}^{n} X_j\right) \Rightarrow \widehat{\theta} = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

5. (a) Set $\eta = \frac{1}{2\theta^2}$ then

$$p(x,\theta) = x\exp\left\{\eta(-x^2) - (-\log(2\eta))\right\}$$
$$T(x) = -x^2$$
$$A(\eta) = -\log(2\eta) \Rightarrow \frac{d}{d\eta}A(\eta) = -\frac{1}{\eta} = -2\theta^2 = -\mathbb{E}_\theta\left[X^2\right]$$

so

$$\widehat{\theta} = \left(\frac{1}{2n}\sum_{j=1}^{n} X_j^2\right)^{1/2}$$

(b)

$$p(x,\theta) = \frac{\theta^p}{\Gamma(p)}x^{p-1}e^{-\theta x} \qquad x > 0$$

set $\eta = \theta$ then

$$p(x,\eta) = \frac{x^{p-1}}{\Gamma(p)}e^{\eta(-x)-(-p\log(\eta))}$$
$$T(x) = x \qquad A(\eta) = -p\log(\eta)$$
$$\frac{d}{d\eta}A(\eta) = -\frac{p}{\eta}$$

so that

$$\mathbb{E}_\theta[T(X)] = -\mathbb{E}_\theta[X] = -\frac{p}{\theta}$$

so

$$\widehat{\theta} = \frac{p}{\overline{X}}$$

6. (a) $\mu = \mathbb{E}\left[X_i\right]$ so $\widehat{\mu} = \overline{X}$.

(b) Here $X_i = e_i$ and

$$\mathrm{Var}(e_i) = b^2\mathrm{Var}(e_{i-1}) + \sigma^2 \Rightarrow \mathrm{Var}(e_i) = \sigma^2\left(\sum_{i=0}^{j-1} b^{2i}\right)$$

It follows that $\mathbb{E}\left[U_i^2\right] = \sigma^2$ for each $i = 1,\ldots,n$, hence

$$\widehat{\sigma}^2 := \frac{1}{n}\sum_{j=1}^{n} U_j^2$$

is a method of moments estimator of σ^2.

(c) For $i = 1, 2, \ldots, n$ (using $X_0 = \mu$ and $\mathbb{E}[X_i] = \mu$),

$$(X_i - \mu) = \beta(X_{i-1} - \mu) + \epsilon_i$$

so that

$$\text{Cov}(X_i, X_{i-1}) = \beta \text{Var}(X_{i-1})$$

hence

$$\sum_{i=1}^{n-1} \text{Cov}(X_{i+1}, X_i) = \beta \sum_{i=1}^{n-1} \text{Var}(X_i)$$

This leads to

$$\widehat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)(X_{i+1} - \mu)}{\frac{1}{n}\sum_{i=1}^{n-1}(X_i - \mu)^2}$$

This method of moments estimator is consistent provided $|\beta| < 1$, so that

$$\text{Var}(X_i) \to \frac{\sigma^2}{1 - \beta^2} \qquad \text{Cov}(X_i, X_{i+1}) \to \beta\frac{\sigma^2}{1 - \beta^2}.$$

7. (a)

$$\mathbb{E}[X] = \theta \int_1^\infty x.x^{-\theta-1}dx = \theta\left[\frac{x^{1-\theta}}{1-\theta}\right]_1^\infty = \frac{\theta}{\theta - 1}$$

$$\frac{\widehat{\theta}_n}{\widehat{\theta}_n - 1} = \overline{X}$$

$$\widehat{\theta}_n = \frac{\overline{X}}{\overline{X} - 1}.$$

(b) To prove convergence,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\overline{X}}{\overline{X}-1} - \theta\right| > \epsilon\right) &\leq \mathbb{P}\left(\left|\overline{X} - \frac{\theta}{\theta - 1}\right| > \frac{\epsilon}{(\theta - 1 + \epsilon)(\theta - 1)}\right) \\ &\leq \left(\frac{(\theta - 1 + \epsilon)(\theta - 1)}{\epsilon}\right)^2 \text{Var}(\overline{X}) \\ &= \left(\frac{(\theta - 1 + \epsilon)(\theta - 1)}{\epsilon}\right)^2 \frac{1}{n}\text{Var}(X_1) \overset{n \to +\infty}{\longrightarrow} 0\end{aligned}$$

provided $\text{Var}(X_1) < +\infty$, which requires $\theta > 2$. Techniques using the characteristic function enable this to be extended to $\theta > 1$; convergence can be proved using a characteristic function technique if $\mathbb{E}[|X_1|] < +\infty$.

8. (a) $T(X)$ sufficient if and only if there are functions g and h such that

$$p(x, \theta) = g(T(x), \theta)h(x).$$

The conditions give functions g_1, g_2, h_1 and h_2 such that

$$p(x,\theta) = g_1(T_1(x),\theta_1)h_1(x,\theta_2) = g_2(T_2(x),\theta_2)h_2(x,\theta_1).$$

It follows that

$$\frac{g_1(T_1(x),\theta_1)}{h_2(x,\theta_1)} = \frac{g_2(T_2(x),\theta_2)}{h_1(x,\theta_2)} = f(x)$$

and hence that

$$p(x,\theta) = g_1(T_1(x),\theta_1)g_2(T_2(x),\theta_2)f(x)$$

and hence $(T_1(X), T_2(X))$ is sufficient for (θ_1,θ_2).

(b) Sample $(X_1,\ldots,X_n)$ from $N(\mu,\sigma^2)$, statistics $T_1(X) = \sum_{j=1}^n X_j$ and $T_2(X) = \sum_{j=1}^n X_j^2$. Here $T_1(X)$ is sufficient for μ whether or not σ^2 is known. The statistic $\sum_{j=1}^n X_j^2$ is not sufficient for σ^2; $\sum_{j=1}^n X_j$ is also needed whether or not μ is known. This is seen from the factorisation: the formula is

$$p(x,\sigma^2) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^n x_j^2 + \frac{\mu}{\sigma^2}\sum_{j=1}^n x_j - \frac{n\mu^2}{\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)\right\}$$

and, even if μ is fixed, both $\sum_{j=1}^n x_j^2$ and $\sum_{j=1}^n x_j$ give information about σ^2.

9. Firstly, we compute the probability that failure occurs after time r. It is:

$$\mathbb{P}_\theta(X \geq r+1) = (1-\theta)\sum_{k=r+1}^{\infty}\theta^{k-1} = (1-\theta)\theta^r\sum_{k=0}^{\infty}\theta^k = \theta^r$$

Now suppose the events $Y_1 \in A_1, \ldots, Y_n \in A_n$ are observed. The likelihood is:

$$L(\theta; Y_1 \in A_1, \ldots, Y_n \in A_n) = \prod_{j=1}^n L(\theta; Y_j \in A_j)$$

since these events are independent. If $A_j = \{r+1, r+2, \ldots\}$ we have $L(\theta, Y_j \in A_j) = \theta^r$. If $A_j = \{y_j\}$, we have $L(\theta; Y_j \in A_j) = \theta^{y_j - 1}(1-\theta)$ and multiplying these together gives:

$$L(\theta) = \theta^{S-(n-m)}(1-\theta)^{n-m}\theta^{rm}.$$

To get the maximum, it is probably easier to use logarithms:

$$\log L(\theta) = (S - n + (r+1)m)\log\theta + (n-m)\log(1-\theta)$$

Taking derivative to get critical points:

$$\frac{d}{d\theta}\log L(\theta) = \frac{S-n+(r+1)m}{\theta} - \frac{n-m}{1-\theta}$$

$$\widehat{\theta} = \frac{S + (r+1)m - n}{S + rm}$$

Maximum - clear -

$$\frac{d^2}{d\theta^2} = -\frac{S - n + (r+1)m}{\theta^2} - \frac{n-m}{(1-\theta)^2} < 0$$

hence strictly concave, $\log L(\theta) > +\infty$ and bounded above for $0 < \theta < 1$ and $\log L(0) = \log L(1) = -\infty$ so exists a maximum, which is unique and given at point where $\frac{d}{d\theta} \log L = 0$.

10. Let $v = \sigma^2$

$$\log L(\mu, v) = -\frac{n}{\log}(2\pi) - \frac{n}{2} \log v - \frac{1}{2v} \sum_{j=1}^{n} (x_j - \mu)^2$$

so the likelihood equations are:

$$\begin{cases} -\frac{n}{2v} + \frac{1}{2v^2} \sum_{j=1}^{n} (x_j - \mu)^2 = 0 \\ \sum_{j=1}^{n} (x_j - \mu) = 0 \end{cases}$$

It follows that $\widehat{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j = \overline{X}$. For the other equation,

$$v = \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu)^2$$

so that $\widehat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X})^2$. This is an exponential family and conditions that MLE exists and is the unique solution of the likelihood equations are clearly satisfied. From earlier:

$$\mathbb{E}\left[\widehat{\mu}\right] = \mu, \qquad \mathbb{E}\left(\widehat{\sigma}^2\right) = \frac{n-1}{n} \sigma^2$$

so the variance estimator is biased.

Chapter 5

Maximum Likelihood

Maximum likelihood estimation has already been introduced; we continue the discussion and establish properties of the maximum likelihood estimator.

5.1 Likelihood Equations

If Θ is an open set and $\log L(\theta, x)$ is differentiable in θ and $\widehat{\theta}$ exists, then $\widehat{\theta}$ satisfies the estimating equation:

$$\nabla_\theta \log L(\theta, x) = 0$$

This is known as the *likelihood equation.* If the $X_1, \ldots, X_n$ are independent variables with log likelihood functions $\log L_i(\theta, x)$, then the log likelihood equation simplifies to:

$$\sum_{i=1}^{n} \nabla_\theta \log L(\widehat{\theta}, X_i) = 0.$$

Example 5.1 (Hardy - Weinberg Populations)**.**

Consider a population with three types of individuals labelled 1, 2 and 3, occurring in the Hardy - Weinberg proportions

$$p(1, \theta) = \theta^2 \qquad p(2, \theta) = 2\theta(1-\theta) \qquad p(3, \theta) = (1-\theta)^2$$

where $0 < \theta < 1$. Suppose we observe a random sample of five individuals, where $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 2, x_5 = 3$. Then

$$L(\theta, x) = p(1, \theta)p(2, \theta)^2 p(3, \theta)^2 = 4\theta^4(1-\theta)^6.$$

so that the log likelihood function is:

$$\log L(\theta, x) = \log L(\theta, x) = \log 4 + 4 \log \theta + 6 \log(1-\theta).$$

The likelihood equation is:

$$\frac{\partial}{\partial\theta}\log L(\theta, x) = 0$$

which is:

$$\frac{\partial}{\partial\theta}(\log 4 + 4\log\theta + 6\log(1-\theta)) = \frac{4}{\theta} - \frac{6}{1-\theta} = 0.$$

This has unique solution: $\widehat{\theta} = \frac{2}{5}$. Because

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta, x) = -\frac{4}{\theta^2} - \frac{6}{(1-\theta)^2} < 0,$$

for all $\theta \in (0,1)$ and $\log L(0,x) = \log L(1,x) = -\infty$ (in the sense of $\theta \to 0$ and $\theta \to 1$), it follows that $\widehat{\theta}$ maximises $\log L(\theta, x)$ and hence $L(\theta, x)$.

In general, let n_1, n_2, n_3 denote respectively the number of 1's, 2's and 3's in $x_1, \ldots, x_n$. Then if both $2n_1 + n_2$ and $n_2 + 2n_3$ are positive, the maximum likelihood estimate exists and is given by

$$\widehat{\theta}(x) = \frac{2n_1 + n_2}{2n}.$$

If $2n_1 + n_2 = 0$, then the likelihood is $(1-\theta)^n$ which is maximised by $\widehat{\theta} = 0$. The MLE does not exist because $\Theta = (0,1)$. Similarly, the MLE does not exist if $n_2 + 2n_3 = 0$. □

To apply the likelihood equation successfully, we need to know when a solution is a MLE. A sufficient condition is that $\log L$ is *concave* in θ. If θ is a single parameter and $\log L$ is twice differentiable, this is equivalent to

$$\frac{\partial^2}{\partial\theta^2}\log L(\theta, x) \le 0 \qquad \forall\theta \in \Theta.$$

This technique may have difficulties carrying over to multiparameter settings, where the matrix of second derivatives has to be considered and shown to be non-positive definite. An alternative approach in some cases is to consider Kullback Leibler divergence.

Example 5.2 (Multinomial Sampling)**.**

This example illustrates how Kullback-Leibler divergence may be used in the multinomial setting.

Let $(X_1, \ldots, X_n)$ be a random sample from a distribution where $\mathbb{P}(X_1 = j) = \theta_j$ for $j = 1, \ldots, k$ (so that $\sum_{j=1}^k \theta_j = 1$). Suppose that $\theta \in \Theta = \{\theta \in (0,1)^k : \sum_{j=1}^k \theta_j = 1\}$. Suppose that in the n trials, $N_j = \sum_{i=1}^n \mathbf{1}_{\{j\}}(X_i)$. Then

$$L(\theta; X) = \prod_{j=1}^k \theta_j^{N_j}$$

and

$$\log L(\theta; X) = \sum_{j=1}^{k-1} N_j \log\theta_j + N_k \log\left(1 - \sum_{j=1}^{k-1}\theta_j\right).$$

There are $k-1$ free parameters. The likelihood equation is:

$$0 = \frac{N_j}{\theta_j} - \frac{N_k}{1-\sum_{i=1}^{k-1}\theta_i} = \frac{N_j}{\theta_j} - \frac{N_k}{\theta_k}$$

so that

$$\frac{N_1}{\widehat{\theta}_1} = \ldots = \frac{N_k}{\widehat{\theta}_k} = \alpha$$

for some constant α. It follows that $\widehat{\theta}_j = \alpha N_j$ and since $\sum_{j=1}^k \widehat{\theta}_j = 1$, it follows that

$$\widehat{\theta}_j = \frac{N_j}{N} \qquad j = 1, \ldots, k.$$

The following argument shows that this maximises the likelihood function:

$$\begin{aligned} \frac{1}{N}\log L(\theta; X) &= \frac{1}{N}\sum_{j=1}^k N_j \log\theta_j = \sum_{j=1}^k \widehat{\theta}_j \log\theta_j \\ &= \sum_{j=1}^k \widehat{\theta}_j \log\widehat{\theta}_j - \sum_{j=1}^k \widehat{\theta}_j \log\frac{\widehat{\theta}_j}{\theta_j} = -\mathcal{E}(\widehat{\theta}) - D_{KL}(\widehat{\theta}\|\theta). \end{aligned}$$

where $\mathcal{E}(p)$ is the Shannon entropy. Recall that D_{KL}, the Kullback-Leibler divergence, is non negative and 0 if and only if $\widehat{\theta} = \theta$. It follows that $\widehat{\theta}$ maximises the log-likelihood and hence the likelihood. □

5.2 Maximum Likelihood for Exponential Families

For exponential families, convexity of the log-partition function simplifies the task of showing that a solution to the likelihood equations maximises the likelihood function. The following lemma is trivially clear:

Lemma 5.1. *Let $\Theta \subset \mathbb{R}^p$ be an open set. Let $\partial\Theta$ denote its boundary. Let d denote Euclidean distance. Let $l : \Theta \rightarrow \mathbb{R}$ be a continuous function. Suppose also that for any sequence $(\theta_n)_{n\geq 1}$ satisfying $d(\theta_n, \partial\Theta) \overset{n\rightarrow+\infty}{\longrightarrow} 0$, $\lim_{n\rightarrow+\infty} l(\theta_n) = -\infty$ and that for any sequence $(\theta_n)_{n\geq 0}$ such that $\theta_n \in \Theta$ and $|\theta_n| \overset{n\rightarrow+\infty}{\longrightarrow} +\infty$, $l(\theta_n) \overset{n\rightarrow+\infty}{\longrightarrow} -\infty$.*

Then there exists a value $\widehat{\theta} \in \Theta$ (not necessarily unique) such that

$$l(\widehat{\theta}) = \max\{l(\theta) : \theta \in \Theta\}.$$

When strict concavity is added, the maximum is unique.

Proposition 5.2. *Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^p$ is open. Let $\partial\Theta$ denote the boundary of Θ. Let d denote Euclidean distance. Suppose that $\log L(\theta, x)$, the log likelihood function, is strictly concave and, for any sequence $(\theta_n)_{n\geq 0}$ satisfying $d(\theta_n, \partial\Theta) \overset{n\rightarrow+\infty}{\longrightarrow} 0$, satisfies $\lim_{n\rightarrow+\infty} \log L(\theta_n, x) = -\infty$. Then the maximum likelihood estimate $\widehat{\theta}(x)$ exists and is unique.*

Proof By concavity, it follows that $\log L(\theta, x)$ is continuous in θ. It follows from the previous lemma that $\widehat{\theta}(x)$ exists. By strict concavity it follows that $\widehat{\theta}(x)$ is unique. □

The following theorem is a consequence of the properties of exponential families.

Important Note: This result states that, under the hypotheses, there is a $1-1$ mapping between the *method of moment* estimator of $\mathbb{E}_\eta[T(X)]$ (expected value of the sufficient statistic) and the *maximum likelihood* estimator of η (the canonical parameter vector).

Theorem 5.3. *Let $\mathcal{P}$ be a canonical exponential family generated by (T, h), where $T = (T_1, \dots, T_k)$, of rank k, where the natural parameter space $\mathcal{E} = \{\eta : A(\eta) < +\infty\}$ is open. Let x be the observed data vector and set $t_0 = T(x)$.*

1. *Suppose there is a $\delta > 0$ and an $\epsilon > 0$ such that $t_0 \in \mathbb{R}^k$ satisfies :*

$$\inf_{(c_1,\dots,c_k):\sum_j c_j^2 = 1} \mathbb{P}\left((c, T(X) - t_0) > \delta\right) > \epsilon \tag{5.1}$$

then the MLE $\widehat{\eta}$ exists, is unique and is the solution to the equation

$$\dot{A}(\eta) = \mathbb{E}_\eta\left[T(X)\right] = t_0. \tag{5.2}$$

2. *If t_0 does not satisfy Equation (5.1), then the MLE does not exist and equation (5.2) has no solution.*

Definition 5.4 (Convex Support). *Let $\mathbb{P}$ be a probability measure defined on $\mathcal{X} \subset \mathbb{R}^p$. The* convex support *of a probability measure $\mathbb{P}$ is the smallest* convex *set $C \supseteq \mathcal{X}$ such that $\mathbb{P}(C) = 1$.*

Proof of Theorem 5.3 Firstly existence, and then uniqueness, of the MLE is proved. Let η_0 denote a reference parameter $\eta_0 \in \mathcal{E}$. It follows that

$$p(x, \eta) = p(x, \eta_0) \exp\left\{(\eta - \eta_0, T(x) - t_0) - (A(\eta) - (\eta - \eta_0, t_0) - A(\eta_0))\right\}.$$

Let $\theta = \eta - \eta_0$ and let $S = T - t_0$, so that

$$p(x, \eta) = p(x, \eta_0) \exp\left\{(\theta, S(x)) - (A(\eta_0 + \theta) - A(\eta_0) - (\theta, t_0))\right\}.$$

Now consider a maximising sequence $\{\theta_m\}$. Suppose that $\eta_m := \eta_0 + \theta_m$ has no limit point in $\mathcal{E}$. Let $\theta_m = \lambda_m u_m$, where $\lambda_m = \|\theta_m\|$ and $u_m = \frac{\theta_m}{\|\theta_m\|}$. If $\{\eta_0 + \theta_m\}$ has no limit point in $\mathcal{E}$, then there is a subsequence $\{\eta_0 + \theta_{m_k}\}$ that has one of the following properties:

1. $\lambda_{m_k} \to +\infty$, $u_{m_k} \to u$. If this holds, then

$$\begin{aligned}\liminf_{k\to+\infty} \int e^{(\theta_{m_k}, S(x))} p(x, \eta_0) dx &= \liminf_{k\to+\infty} \mathbb{E}_{\eta_0}\left[e^{\lambda_{m_k}(u_{m_k}, S(X))}\right] \\ \geq \liminf_{k\to+\infty} e^{\lambda_{m_k}\delta} \mathbb{P}_{\eta_0}\left((u_{m_k}, S(X)) > \delta\right) &\geq \liminf_{k\to+\infty} e^{\lambda_{m_k}\delta} \mathbb{P}_{\eta_0}\left((u, S(X)) > \delta\right) = +\infty.\end{aligned}$$

From this, it follows that $e^{-(A(\eta_0+\theta_{m_k})-A(\eta_0)-(\theta_{m_k},t_0)} \to 0$, so that $A(\eta_{m_k}) \to +\infty$. It follows that $\log L(\eta_{m_k}, x) \overset{k\to+\infty}{\longrightarrow} -\infty$.

2. $\lambda_{m_k} \to \lambda$ and $u_{m_k} \to u$. Then $\lambda u \notin \mathcal{E}$ so that $A(\eta_0 + \lambda u) = +\infty$ by definition of the natural parameter space and hence $\log L(\eta_{m_k}, x) \to -\infty$.

It follows that neither of these cases are maximising sequences. It follows that a maximising sequence has a converging subsequence which converges to a limit $\widehat{\eta}$ which maximises the likelihood. Now consider two different maximisers, η_1 and η_2. Then, since the log partition function is strictly convex, for $\alpha \in (0,1)$,

$$A(\alpha\eta_1 + (1-\alpha)\eta_2) - (\alpha\eta_1 + (1-\alpha)\eta_2, T(x)) < \alpha(A(\eta_1) - (\eta_1, T(x))) + (1-\alpha)(A(\eta_2) - (\eta_2, T(x))$$

so that $\eta := \alpha\eta_1 + (1-\alpha)\eta_2$ satisfies: $p(x,\eta) > p(x,\eta_1)$ so that η_1 and η_2 do not maximise. It follows that the maximiser is unique and solves the likelihood equation $\nabla_\eta \log L(\eta) = 0$ giving

$$\nabla_\eta \left((\eta, T(x)) - A(\eta)\right) = 0 \Rightarrow t_0 = T(x) = \dot{A}(\eta).$$

Furthermore, by Corollary 3.5, it follows that $\dot{A}(\eta) = \mathbb{E}_\eta\left[T(X)\right]$. It therefore follows that $\widehat{\eta}_{mle}$ satisfies Equation (5.2). □

5.3 The EM (Expectation / Maximisation) Algorithm

In many examples, it is hard to compute the maximum likelihood estimate explicitly and numerical approximations are necessary. One popular algorithm is the EM algorithm, which is suitable for a large class of examples. The algorithm works as follows:

Suppose we do not observe X, but rather $S(X)$ where $S(X)$ is a statistic. We want to find $\widehat{\theta}_{MLE}$ based on this information.

$$J(\theta|\theta_0) = \mathbb{E}_{\theta_0}\left[\log\frac{p(X,\theta)}{p(X,\theta_0)}\middle| S(X) = s\right].$$

Note that $J(\theta|\theta_0) = D_{KL}(\theta_0\|\theta)|_{S(X)=s}$; that is, the Kullback Leibler divergence, conditioned on the event $S(X) = s$.

- **Initialise** with $\theta_{\text{old}} = \theta_0$.

- **E step** Compute $J(\theta|\theta_{\text{old}})$ for as many values of θ as needed. If this step is difficult, then the EM algorithm is probably not suitable.

- **M step** Maximise $J(\theta|\theta_{\text{old}})$ as a function of θ. Again, if this step is difficult, then EM is probably not appropriate.

- Set $\theta_{\text{new}} = \text{argmax}_\theta J(\theta|\theta_{\text{old}})$, let θ_{old} take the value of θ_{new} and repeat the process.

The rationale behind the algorithm is the following: let $q(s, \theta)$ denote the mass / density for S when the parameter value is θ. Proof of the following formula is left as an exercise:

$$\frac{q(s,\theta)}{q(s,\theta_0)} = \mathbb{E}_{\theta_0}\left[\frac{p(X,\theta)}{p(X,\theta_0)}\right| S(X) = s\left.\vphantom{\frac{p(X,\theta)}{p(X,\theta_0)}}\right]$$

From this, taking logs and differentiating gives

$$\left.\frac{\partial}{\partial\theta}\log q(s,\theta)\right|_{\theta=\theta_0} = \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log p(X,\theta)\right]\Bigg|_{\theta=\theta_0}$$

for all θ_0 under suitable regularity assumptions.

Formally,

$$\frac{\partial}{\partial\theta}J(\theta|\theta_0) = \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log p(X,\theta)\right| S(X) = s\left.\vphantom{\frac{\partial}{\partial\theta}}\right]$$

from which

$$\left.\frac{\partial}{\partial\theta}J(\theta|\theta_0)\right|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\log q(s,\theta_0).$$

It follows that a fixed point $\tilde{\theta}$ of the algorithm satisfies the likelihood equation:

$$\frac{\partial}{\partial\theta}\log q(s,\tilde{\theta}) = 0.$$

The algorithm behaves well due to the following lemma:

Lemma 5.5. *Let θ_{new} and θ_{old} be defined as in the statement of the algorithm and let $S(X) = s$. Then*

$$q(s,\theta_{new}) \geq q(s,\theta_{old}).$$

Equality holds if and only if the conditional distribution of X given $S(X) = s$ is the same for the θ_{new} as for θ_{old} and θ_{old} maximises $J(\theta|\theta_{old})$.

Proof The proof is given in the discrete case. The result holds whenever the quantities in $J(\theta|\theta_0)$ can be defined in a reasonable fashion. In the discrete case,

$$p(x,\theta) = q(s,\theta)r(x|s,\theta)$$

where r is the conditional probability mass function of X given $S(X) = s$. Then

$$J(\theta|\theta_0) = \log\frac{q(s,\theta)}{q(s,\theta_0)} + \mathbb{E}_{\theta_0}\left[\log\frac{r(X|s,\theta)}{r(X|s,\theta_0)}\right| S(X) = s\left.\vphantom{\frac{r}{r}}\right].$$

If $\theta_0 = \theta_{\text{old}}$ and $\theta = \theta_{\text{new}}$, then

$$\log\frac{q(s,\theta_{\text{new}})}{q(s,\theta_{\text{old}})} = J(\theta_{\text{new}}|\theta_{\text{old}}) - \mathbb{E}_{\theta_{\text{old}}}\left[\log\frac{r(X|s,\theta_{\text{new}})}{r(X|s,\theta_{\text{old}})}\right| S(X) = s\left.\vphantom{\frac{r}{r}}\right]$$

By definition of θ_{new},

$$J(\theta_{\text{new}}|\theta_{\text{old}}) \geq J(\theta_{\text{old}}, \theta_{\text{old}}) = 0.$$

On the other hand, it follows from Shannon's inequality (Lemma 4.3) that

$$-\mathbb{E}_{\theta_{\text{old}}}\left[\log \frac{r(X|s,\theta_{\text{new}})}{r(X|s,\theta_{\text{old}})}\middle| S(X) = s\right] \geq 0.$$

□

This leads to the following theorem for exponential families.

Theorem 5.6. *Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a canonical exponential family generated by (h, T) of rank k where the natural parameter space is open. Let $S(X)$ be any statistic. Then*

1. *The EM algorithm consists of the alternation:*

$$\begin{cases} \dot{A}(\theta_{new}) = \mathbb{E}_{\theta_{old}}[T(X)|S(X) = s] \\ \theta_{old} = \theta_{new}. \end{cases}$$

(If a solution to the first of these exists, then it is necessarily unique).

2. *If the sequence of iterates $\widehat{\theta}_m$ so obtained is bounded and the equation*

$$\dot{A}(\theta) = \mathbb{E}_\theta [T(X)|S(X) = s]$$

has a unique solution then it converges to a limit $\widehat{\theta^*}$ which is necessarily a local maximum of $q(s, \theta)$.

Proof In this case,

$$\begin{aligned} J(\theta|\theta_0) &= \mathbb{E}_{\theta_0}[(\theta - \theta_0, T(X)) - (A(\theta) - A(\theta_0))|S(X) = s] \\ &= (\theta - \theta_0, \mathbb{E}_{\theta_0}[T(X)|S(X) = y]) - (A(\theta) - A(\theta_0)). \end{aligned}$$

From this, part (a) follows. Part (b) is more difficult; a proof due to Wu [11] (1983) is sketched in one of the tutorial exercises. □

Tutorial 6

1. **Maximum Likelihood: Hypergeometric** Suppose X has probability function

$$\mathbb{P}(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad k=0,1,\dots,n \qquad 0\le n\le M\le N$$

where N, M, n are non negative integers. Show that the maximum likelihood estimate of M for N and n fixed is given by

$$\widehat{M}(X)=\left\lfloor \frac{X}{n}(N+1)\right\rfloor$$

if $\frac{X}{n}(N+1)$ is not an integer and

$$\widehat{M}(X)=\frac{X}{n}(N+1) \qquad \text{or} \qquad \frac{X}{n}(N+1)-1$$

otherwise, where $\lfloor x\rfloor$ denotes the integer part of x.

Hint: Consider the ratio $\frac{L(M+1,x)}{L(M,x)}$ as a function of M.

2. **Maximum Likelihood** Suppose $X_1,\dots,X_n$ is a sample from a population with density

$$f(x;\mu,\sigma^2)=\frac{9}{10\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)+\frac{1}{10}\phi(x-\mu)$$

where ϕ defined as $\phi=\frac{1}{\sqrt{2\pi}}\exp\{-x^2/2\}$ $-\infty<x<+\infty$ is the standard normal density function and the parameter space is $(\mu,\sigma)\in\Theta=\mathbb{R}\times(0,+\infty)$. Show that the maximum likelihood estimator for the pair (μ,σ) does not return a good answer if $\sigma>0$.

3. Let $X_1,\dots,X_n$ be i.i.d., with parent distribution $U((\theta-\frac{1}{2},\theta+\frac{1}{2}))$ where θ is an unknown parameter. That is, the distribution with density

$$p(x;\theta)=\mathbf{1}_{(\theta-\frac{1}{2},\theta+\frac{1}{2})}(x).$$

Find the maximum likelihood estimator of θ.

4. (a) Let Y be any random variable and let $R(c)=\mathbb{E}[|Y-c|]$ be the *mean absolute prediction error*. Show that either $R(c)\equiv+\infty$ or else $R(c)$ is minimised by any number c_0 such that

$$\mathbb{P}(Y\ge c_0)\ge\frac{1}{2} \qquad \text{and} \qquad \mathbb{P}(Y\le c_0)\ge\frac{1}{2}.$$

A number c satisfying this property is known as the *median.*

Hint: First show that if $c<c_0$ then

$$\mathbb{E}[|Y-c_0|]=\mathbb{E}[|Y-c|]-(c_0-c)(\mathbb{P}(Y\ge c_0)-\mathbb{P}(Y<c_0))-2\mathbb{E}[(Y-c)\mathbf{1}_{(c,c_0)}(Y)]$$

and consider the consequences if c_0 is the median. Consider a symmetric argument for $c>c_0$.

(b) Suppose that $Y_1, \ldots, Y_n$ are independent with Y_i having the Laplace density

$$\frac{1}{2\sigma} \exp\left\{-\frac{|y_i - \mu_i|}{\sigma}\right\} \qquad \sigma > 0,$$

where $\mu_i = \sum_{j=1}^{p} z_{ij}\beta_j$. The z_{ij} are fixed and known, the β_j are unknown parameters.

i. Show that the MLE of $(\beta_1, \ldots, \beta_p, \sigma)$ is obtained by finding $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ that minimises the *least absolute deviation contrast function* $\sum_{j=1}^{n} |y_j - \mu_j|$ and then setting $\widehat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{\mu}_i|$ where $\widehat{\mu}_i = \sum_{j=1}^{p} z_{ij}\widehat{\beta}_j$.

ii. Suppose $\mu_i = \mu$ for each i. Show that the *sample median* $\widehat{y}$ is the minimiser of $\sum_{i=1}^{n} |y_i - \mu|$.

5. Let $X \sim Poiss(n(\mu_1 + \mu_2))$, $Y \sim Poiss(m\mu_1)$ and $Z \sim Poiss(m\mu_2)$ be independent variables, where n and m are fixed and known. Find the MLE of (μ_1, μ_2) based on (X, Y, Z).

6. Let $X_1, \ldots, X_n$ be i.i.d. with density $\frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right)$, $\sigma > 0$ and $\mu \in \mathbb{R}$. Let $w = -\log f_0$ and assume that w'' exists and satisfies $w'' > 0$; $w(\pm\infty) = +\infty$.

(a) Show that if $n \geq 2$, the likelihood equations are:

$$\begin{cases} \sum_{i=1}^{n} w'\left(\frac{X_i - \mu}{\sigma}\right) = 0 \\ \sum_{i=1}^{n} \left\{\frac{(X_i - \mu)}{\sigma} w'\left(\frac{X_i - \mu}{\sigma}\right) - 1\right\} = 0 \end{cases}$$

and that they have a unique solution $(\widehat{\mu}, \widehat{\sigma})$.

Hint Show that the function $D(a, b) = \sum_{i=1}^{n} w(aX_i - b) - n\log a$ is strictly convex in the variables (a, b) and $\lim_{(a,b)\to(a_0,b_0)} D(a, b) = +\infty$ if either $a_0 = 0$ or $+\infty$, or $b_0 = \pm\infty$. You may use the following:

- If a strictly convex function has a minimum, then it is unique.
- For a function D of two variables, if $\frac{\partial^2 D}{\partial a^2} > 0$, $\frac{\partial^2 D}{\partial b^2} > 0$ and $\frac{\partial^2 D}{\partial a^2}\frac{\partial^2 D}{\partial b^2} > \left(\frac{\partial^2 D}{\partial a \partial b}\right)^2$ then D is strictly convex.

(b) Suggest an algorithm, using Newton-Raphson techniques applied to the problem of locating the minimum of $D(a, b)$ such that, with initial conditions $\widehat{\mu}^{(0)} = 0$, $\widehat{\sigma}^{(0)} = 1$, $\widehat{\mu}^{(i)} \to \widehat{\mu}$ and $\widehat{\sigma}^{(i)} \to \widehat{\sigma}$.

(c) Show that for the logistic distribution (c.d.f. $F_0(x) = \frac{1}{1+e^{-x}}$ for $-\infty < x < +\infty$), w is strictly convex. Give the likelihood equations in this case for μ and σ.

7. Let $X_1, \ldots, X_n$ be i.i.d. random p-vectors, with density

$$f(x; \theta) = c(\alpha) \exp\{-\|x - \theta\|^\alpha\} \qquad \theta \in \mathbb{R}^p \qquad \alpha \geq 1$$

where $\frac{1}{c(\alpha)} = \int_{\mathbb{R}^p} \exp\{-\|x\|^\alpha\}\, dx$, $\|.\|$ denotes the Euclidean norm.

(a) Show that if $\alpha > 1$, then the MLE $\widehat{\theta}$ exists and is unique.

(b) Show that if $\alpha = 1$ and $p = 1$, then the MLE $\widehat{\theta}$ exists, but is not unique if n is even.

8. Let $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$ be independent. Find the maximum likelihood estimates of θ_1 and θ_2 when it is known that $\theta_1 \leq \theta_2$.

9. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ populations respectively. Show that the MLE of $\theta = (\mu_1, \mu_2, \sigma^2)$ is

$$\widehat{\theta} = (\overline{X}, \overline{Y}, \widehat{\sigma^2})$$

where

$$\widehat{\sigma^2} = \frac{1}{m+n} \left(\sum_{i=1}^{m} (X_i - \overline{X})^2 + \sum_{j=1}^{n} (Y_j - \overline{Y})^2 \right)$$

10. Suppose that $T(X)$ is sufficient for θ and that $\widehat{\theta}(X)$ is a maximum likelihood estimator of θ. Show that if $\widehat{\theta}$ is unique, then it depends on X only through $T(X)$. (Use the factorisation theorem)

11. (a) Let $X \sim \mathbb{P}_\theta$, $\theta \in \Theta$ and let $\widehat{\theta}$ denote the MLE of θ. Suppose that h is a one-to-one function from Θ onto $h(\Theta)$. Define $\eta = h(\theta)$ and let $p(x, \eta)$ denote the density or probability mass function in terms of η (i.e. reparametrise the model using η). Show that the MLE of η is $h(\widehat{\theta})$. In other words, the MLE is unaffected by reparametrisation; they are equivalent under one-to-one transformations.

    (b) Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$, $p \geq 1$ be a family of models for X, with state space $\mathcal{X} \subset \mathbb{R}^d$. Let q be a map from Θ onto Ω, where $\Omega \subset \mathbb{R}^k$, $1 \leq k \leq p$. Show that if $\widehat{\theta}$ is a MLE of θ, then $q(\widehat{\theta})$ is a MLE of $\omega = q(\theta)$.
    Hint Let $\Theta(\omega) = \{\theta \in \Theta : q(\theta) = \omega\}$, then $\{\Theta(\omega) : \omega \in \Omega\}$ is a partition of Θ and $\widehat{\theta}$ belongs to only one member of this partition, say $\Theta(\widehat{\omega})$. Because q is onto Ω, it follows that for each $\omega \in \Omega$ there is $\theta \in \Omega$ such that $\omega = q(\theta)$. Thus the MLE of ω is by definition

$$\widehat{\omega}_{MLE} = \arg\sup_{\omega \in \Omega} \sup \{L(\theta; X) : \theta \in \Theta(\omega)\}$$

    where arg sup means the value of ω which maximises Now show that $\widehat{\omega}_{MLE} = q(\widehat{\theta})$.

12. **E-M Algorithm**

    (a) Suppose that we only observe $S(X)$, where S is a function of X (the observation). Suppose that X is discrete and let $q(s, \theta)$ denote the mass function for S when θ is the parameter. Show that

$$\frac{q(s, \theta)}{q(s, \theta_0)} = \mathbb{E}_{\theta_0} \left[\frac{p(X, \theta)}{p(X, \theta_0)} \middle| S(X) = s \right]$$

    (b) Establish part 2. of Theorem 5.6. Do this by showing that $\{(\theta_m, \theta_{m+1})\}$ has a subsequence converging to $\{(\theta^*, \theta^*)\}$ and that therefore θ^* is a global minimiser.

Short Answers

1.
$$\frac{L(M+1,x)}{L(M,x)} = \frac{(M+1)(N-M-n+x)}{(M+1-x)(N-M)}$$
$$\frac{L(M+1,x)}{L(M,x)} > 1 \Leftrightarrow M < \frac{x(N+1)}{n} - 1$$

It follows that
$$L(M+1,x) \le L(M,x) \Leftrightarrow M \ge \frac{x(N+1)}{n} - 1$$
so that $L(M,x)$ is the maximum value if and only if
$$M = \begin{cases} \left\lfloor \frac{x(N+1)}{n} \right\rfloor & \frac{x(N+1)}{n} \notin \mathbb{Z}_+ \\ \frac{x(N+1)}{n} - 1 & \frac{x(N+1)}{n} \in \mathbb{Z}_+ \\ \frac{x(N+1)}{n} & \frac{x(N+1)}{n} \in \mathbb{Z}_+ \quad \text{and} \quad \frac{L(M+1,x)}{L(M,x)} = 1 \end{cases}$$

2.
$$L(\mu,\sigma; x_1,\ldots,x_n) = \prod_{j=1}^{n} \left(\frac{9}{10\sigma}\phi\left(\frac{x_j-\mu}{\sigma}\right) + \frac{1}{10}\phi(x_j-\mu) \right)$$

Clearly, taking $\mu = x_j$ for any $j \in \{1,\ldots,n\}$:
$$\lim_{\sigma\to 0} L(x_1,\ldots,x_n; x_j,\sigma) = +\infty$$

Hence $(\widehat{\mu},\widehat{\sigma}) = (x_j, 0)$ for any $j \in \{1,\ldots,n\}$ returns a value of $+\infty$ for the likelihood. For any $\sigma > 0$ and any $\mu \in \mathbb{R}$, $L(\mu,\sigma) < +\infty$, hence $\widehat{\sigma}_{ML} = 0$ irrespective of the true value of σ.

3.
$$L(\theta; x_1,\ldots,x_n) = \prod_{j=1}^{n} \mathbf{1}_{(x_j-\frac{1}{2}, x_j+\frac{1}{2})}(\theta) = \mathbf{1}_{(\max_j x_j - \frac{1}{2}, \min_j + \frac{1}{2})}(\theta)$$

so $\widehat{\theta}_{ML}$ is not unique; any value $\widehat{\theta}_{ML} \in (\max_j x_j - \frac{1}{2}, \min_j x_j + \frac{1}{2})$ maximises the likelihood.

4. (a) For $c < c_0$,
$$\begin{aligned} \mathbb{E}[|Y-c_0|] &= \mathbb{E}[(Y-c_0)\mathbf{1}_{\{Y>c_0\}}] + \mathbb{E}[(c_0-Y)\mathbf{1}_{\{Y<c_0\}}] \\ &= \mathbb{E}[(Y-c)\mathbf{1}_{\{Y>c\}}] - \mathbb{E}[(Y-c)\mathbf{1}_{\{c<Y\le c_0\}}] - (c_0-c)\mathbb{P}(Y>c_0) \\ &\quad + \mathbb{E}[(c-Y)\mathbf{1}_{\{Y<c\}}] + \mathbb{E}[(c-Y)\mathbf{1}_{\{c<Y<c_0\}}] + (c_0-c)\mathbb{P}(Y<c_0) \\ &= \mathbb{E}[|Y-c|] - (c_0-c)\left(\mathbb{P}(Y\ge c_0) - \mathbb{P}(Y<c_0)\right) - 2\mathbb{E}[(Y-c)\mathbf{1}_{(c,c_0)}(Y)] \end{aligned}$$

Now choose c_0 such that $\mathbb{P}(Y \ge c_0) \ge \frac{1}{2}$, $\mathbb{P}(Y \le c_0) = \frac{1}{2}$, then $R(c_0) \le R(c)$. The inequality is strict unless both $\mathbb{P}(Y = c_0) = 0$ and $\mathbb{P}(Y \in (c,c_0)) = 0$. It follows that $R(c_0) < R(c)$ unless c also satisfies $\mathbb{P}(Y \ge c) \ge \frac{1}{2}$ and $\mathbb{P}(Y \le c) \ge \frac{1}{2}$. Similar arguments for $c > c_0$. It follows that a value c minimises if and only if it is a median.

(b) i. Log likelihood

$$\log L(\beta, \sigma; y_1, \ldots, y_n) = -n \log 2 - n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} |y_i - \mu_i|$$

Let f denote the maximum of $\sum_{i=1}^{n} |y_i - \mu_i|$. Then, for β that gives f,

$$\frac{\partial}{\partial \sigma} \log L = -\frac{n}{\sigma} + \frac{1}{\sigma^2} f \Rightarrow \widehat{\sigma} = \frac{1}{f}$$

The result follows.

ii. Consider the empirical distribution defined by $Y_{(1)}, \ldots, Y_{(n)}$ and apply the result of the first part.

5.

$$\mathbb{P}(X = x, Y = y, Z = z) = \frac{(n(\mu_1 + \mu_2))^x (m\mu_1)^y (m\mu_2)^z}{x!y!z!} e^{-(n(\mu_1+\mu_2))-m\mu_1-m\mu_2)}$$

$$\begin{aligned}\log L(\mu_1, \mu_2; x, y, z) = & \left(x \log \frac{n}{x!y!z!} + x \log(\mu_1 + \mu_2) + y \log m + z \log m \right) \\ & + y \log \mu_1 + z \log \mu_2 - \mu_1(n+m) - \mu_2(n+m)\end{aligned}$$

A critical point, if it is in $(0, \infty) \times (0, \infty)$ (the interior) satisfies

$$\begin{cases} \frac{\partial}{\partial \mu_1} \log L(\mu_1, \mu_2) = \frac{x}{\mu_1+\mu_2} + \frac{y}{\mu_1} - (n+m) = 0 \\ \frac{\partial}{\partial \mu_2} \log L(\mu_1, \mu_2) = \frac{x}{\mu_1+\mu_2} + \frac{z}{\mu_2} - (n+m) = 0 \end{cases}$$

If $x > 0$, $y > 0$ and $z > 0$, there is exactly one solution to these equations. From the equations, $\frac{z}{\mu_2} = \frac{y}{\mu_1}$ so that

$$\frac{x}{\mu_1 + \frac{z\mu_1}{y}} + \frac{y}{\mu_1} = n + m \Rightarrow \frac{xy}{(y+z)\mu_1} + \frac{y}{\mu_1} = n + m \Rightarrow \mu_1 = \frac{y(x+y+z)}{(y+z)(n+m)}$$

giving

$$\mu_2 = \frac{z(x+y+z)}{(n+m)(y+z)} \qquad \mu_1 = \frac{y(x+y+z)}{(n+m)(y+z)}.$$

It turns out that this (μ_1, μ_2) gives a global maximum in $\mathbb{R}_+ \times \mathbb{R}_+$ in all cases. To see this, we consider the boundaries of $\mathbb{R}_+ \times \mathbb{R}_+$ which are: $\mu_1 + \mu_2 \to +\infty$, $\mu_1 \to 0$ for $\mu_2 < +\infty$ and $\mu_2 \to 0$ for $\mu_1 < +\infty$. The different cases are as follows:

1) If $x > 0$, $y > 0$, $z > 0$, then $\log L(\mu_1, \mu_2; x, y, z)$ is strictly concave in (μ_1, μ_2), bounded above, and $\log L(\mu_1, \mu_2; x, y, z) \overset{\mu_1+\mu_2 \to +\infty}{\longrightarrow} -\infty$, $\log L(\mu_1, \mu_2) \to -\infty$ if $\mu_1 \to 0$ (μ_2 fixed) or $\mu_2 \to 0$ (μ_1 fixed).

Therefore, from strict concavity and differentiability, that the maximum is unique and is in the interior of the domain and satisfies $\frac{\partial \log L}{\partial \mu_1} = \frac{\partial \log L}{\partial \mu_2} = 0$.

2) If $x > 0$, $y > 0$, $z = 0$, then $\log L(\mu_1, \mu_2)$ is strictly concave, but there is no solution to the equations $\frac{\partial}{\partial \mu_1} \log L = \frac{\partial}{\partial \mu_2} \log L = 0$ in $(0, +\infty) \times (0, +\infty)$ and hence the maximum is on the boundary. $\mathcal{L}(\mu_1, \mu_2) \overset{\mu_1+\mu_2 \to +\infty}{\longrightarrow} -\infty$, $\log L(\mu_1, \mu_2) \overset{\mu_1 \to 0}{\longrightarrow} -\infty$ for μ_2 fixed.

Therefore, the part of the boundary where the maximum is achieved is $\mu_2 = 0$. The problem now reduces to finding μ_1 that maximises $\log L(\mu_1, 0; x, y, 0)$ which is $\mu_1 = \frac{x+y}{n+m}$.

3) Similarly for $x > 0$, $y = 0$, $z > 0$.

4) $x > 0$, $y = z = 0$ - the only thing that can be estimated is $\mu_1 + \mu_2$, the estimate is

$$\widehat{\mu_1 + \mu_2} = \frac{x}{m+n}$$

5) $x = 0$: This splits into two separate estimation problems, μ_1 which maximises $y \log \mu_1 - \mu_1(n+m)$ and μ_2 which maximises $z \log \mu_2 - \mu_2(n+m)$ which gives

$$\mu_1 = \frac{y}{n+m} \qquad \mu_2 = \frac{z}{n+m}.$$

6. (a) $\log L(\mu, \sigma) = \log \prod_{j=1}^n \left(\frac{1}{\sigma} f_0 \left(\frac{x_j - \mu}{\sigma} \right) \right) = -n \log \sigma - \sum_{j=1}^n w \left(\frac{x_j - \mu}{\sigma} \right)$. Likelihood equations are $\nabla \log L(\mu, \sigma) = 0$ giving

$$\begin{cases} \sum_{j=1}^n w' \left(\frac{x_j - \mu}{\sigma} \right) = 0 \\ \sum_{i=1}^n \left\{ \frac{(x_i - \mu)}{\sigma} w' \left(\frac{X_i - \mu}{\sigma} \right) - 1 \right\} = 0 \end{cases}$$

as required. For uniqueness, consider the function $D(a, b)$. Then

$$\begin{aligned} \nabla D(a,b) = 0 \quad &\Leftrightarrow \quad \left(\sum_{i=1}^n X_i w'(aX_i - b) - \frac{n}{a}, -\sum_{i=1}^n w'(aX_i - b) \right) = 0 \\ &\Leftrightarrow \quad \left(\sum_{i=1}^n \left\{ \frac{(X_i - \mu)}{\sigma} w' \left(\frac{X_i - \mu}{\sigma} \right) - 1 \right\}, \sum_{i=1}^n w' \left(\frac{X_i - \mu}{\sigma} \right) \right) = 0 \end{aligned}$$

using $a = \frac{1}{\sigma}$ and $b = \frac{\mu}{\sigma}$. Now,

$$\frac{\partial^2 D}{\partial a^2} = \sum_{i=1}^n X_i^2 w''(aX_i - b) + \frac{n}{a^2}, \quad \frac{\partial^2 D}{\partial b^2} = \sum_{i=1}^n w''(aX_i - b), \quad \frac{\partial^2 D}{\partial a \partial b} = -\sum_{j=1}^n X_j w''(aX_j - b)$$

$$\begin{aligned}\left(\frac{\partial^2 D}{\partial a \partial b}\right)^2 &= \left(\sum_{j=1}^n X_j w''(aX_j - b)\right)^2 \\ &\leq \sum_{j=1}^n X_j^2 w''(aX_j - b) \sum_{j=1}^n w''(aX_j - b) < \left(\frac{\partial^2 D}{\partial a^2}\right)\left(\frac{\partial^2 D}{\partial b^2}\right)\end{aligned}$$

using $|\sum c_i d_i| \leq (\sum c_i^2)^{1/2}(\sum d_i^2)^{1/2}$; $c_i = X_i\sqrt{w''(aX_i - b)}$ and $d_i = \sqrt{w''(aX_i - b)}$ from which convexity follows. We're using $w'' > 0$.

Finally, we have to show that $\lim_{(a,b)\to(a_0,b_0)} D(a,b) = +\infty$ for (a_0, b_0) as described. The only part which requires attention is: $a_0 = +\infty$. But $w'' > 0$ implies that w' is *increasing*. Since $w(\pm\infty) = +\infty$, this implies that there exists an x_0 such that $w(x_0) = \min_x w(x)$, that $\lim_{x\to+\infty}(-w'(x)) = c_1 > 0$ and $\lim_{x\to+\infty} w'(x) = c_2 > 0$ where c_1 and/or c_2 may be $+\infty$. From this, it is clear that unless $X_1 = \ldots = X_n = 0$, $\lim_{a\to+\infty} D(a,b) = +\infty$, since $\frac{d}{da}\log a = \frac{1}{a} \overset{a\to+\infty}{\longrightarrow} 0$.

(b) Minimise $D(a,b)$. The matrix of second derivatives is positive definite and well defined. Call it M and let $U = \nabla D$. Then

$$\begin{pmatrix} a^{(i+1)} \\ b^{(i+1)} \end{pmatrix} = \begin{pmatrix} a^{(i)} \\ b^{(i)} \end{pmatrix} - M^{-1}(a^{(i)}, b^{(i)})U(a^{(i)}, b^{(i)}).$$

(c) $f_0(x) = (1 + e^{-x})^{-2}e^{-x}$ so that

$$w(x) = -\log f_0(x) = 2\log(1 + e^{-x}) + x, \qquad w''(x) = (1 + e^{-x})^{-2}e^{-2x} + (1 + e^{-x})^{-1}e^{-x}$$

so it is strictly convex.

$$w'(x) = -\frac{1}{e^x + 1}$$

Likelihood equations are:

$$\left\{\begin{array}{l} \sum_{i=1}^n \frac{1}{(e^{(x_i-\mu)/\sigma}+1)} = 0 \\ \sum_{i=1}^n \left\{\frac{(x_i-\mu)/\sigma}{(e^{(x_i-\mu)/\sigma}+1)} - 1\right\} = 0 \end{array}\right.$$

7. (a) For $\alpha > 1$, $\|y\|^\alpha$ is strictly convex in y. This can be seen as follows: for $\alpha > 1$, the function $g : \mathbb{R} \to \mathbb{R}_+$ defined by $g(x) = |x|^\alpha$ is strictly convex. It follows that for $t \in (0,1)$, if $\|x\| \neq \|y\|$, then

$$\begin{aligned}\|tx + (1-t)y\|^\alpha &= \left(t^2\|x\|^2 + 2t(1-t)\langle x, y\rangle + (1-t)^2\|y\|^2\right)^{\alpha/2} \\ &\leq \left(t^2\|x\|^2 + 2t(1-t)\|x\|\|y\| + (1-t)^2\|y\|^2)^{\alpha/2}\right) \\ &= (t\|x\| + (1-t)\|y\|)^\alpha < t\|x\|^\alpha + (1-t)\|y\|^\alpha.\end{aligned}$$

and if $\|x\| = \|y\|$ but $x \neq y$, then $|\langle x, y\rangle| < \|x\|\|y\|$ where the inequality is strict, so that again

$$\|tx + (1-t)y\|^\alpha < t\|x\|^\alpha + (1-t)\|y\|^\alpha.$$

$$\log L(x_1, \ldots, x_n; \theta) = n \log c(\alpha) - \sum_{j=1}^n \|x_j - \theta\|^\alpha$$

and the sum of strictly convex functions is again strictly convex. It follows that the likelihood function has a unique maximiser $\widehat{\theta}_{ML}$.

(b)

$$f(x; \theta) = c \exp\{-|x - \theta|\}$$

$$\log L(x_1, \ldots, x_n; \theta) = n \log c - \sum_{j=1}^n |x_j - \theta|$$

Problem is therefore to find θ that minimises $\sum_{j=1}^n |x_j - \theta|$. It follows from earlier exercise that $\widehat{\theta}$ provides a minimiser where $\widehat{\theta}$ is any sample median. If n is even and $x_{(n/2)} < x_{(n/2)+1}$ then the median is not unique.

8. Minimise

$$(\theta_1 - x_1)^2 + (\theta_2 - x_2)^2$$

subject to the constraint that $\theta_1 \leq \theta_2$. If $x_1 \leq x_2$, then $(\widehat{\theta}_1, \widehat{\theta}_2) = (x_1, x_2)$. If $x_1 > x_2$, then $\widehat{\theta}_1 = \widehat{\theta}_2$ (on the boundary) so that it is the minimiser of

$$2\theta^2 - 2(x_1 + x_2)\theta + (x_1^2 + x_2^2)$$

which is: $\widehat{\theta}_1 = \widehat{\theta}_2 = \frac{x_1 + x_2}{2}$.

9. Minimise:

$$\frac{1}{2\sigma^2}\left(\sum_{j=1}^m (x_j - \mu_1)^2 + \sum_{j=1}^n (y_j - \mu_2)^2\right) + \frac{(m+n)}{2} \log \sigma^2$$

μ_1 and μ_2 are easy; $\widehat{\mu}_1 = \overline{x}$ and $\widehat{\mu}_2 = \overline{y}$. For σ^2, $\widehat{\sigma^2}$ is the point which satisfies:

$$-\frac{1}{2(\sigma^2)^2}\left(\sum_{j=1}^m (x_j - \overline{x})^2 + \sum_{j=1}^n (y_j - \overline{y})^2\right) + \frac{m+n}{2\sigma^2}$$

giving the MLE of

$$\widehat{\sigma^2} = \frac{1}{m+n}\left(\sum_{j=1}^m (X_j - \overline{X})^2 + \sum_{j=1}^n (Y_j - \overline{Y})^2\right)$$

10. Factorisation theorem gives:

$$p(x, \theta) = h(x) g(T(x), \theta);$$

maximising in terms of θ is equivalent to maximising $g(T(x), \theta)$, hence the result follows.

11. (a) $\widehat{\eta}_{ML}$ maximises $p(x,\eta) = p(x, h(\theta))$. The value of θ which maximises this is $\widehat{\theta}_{ML}$, hence if $\widehat{\theta}_{ML}$ is a value of θ which maximises $p(x,\theta)$ then $\eta_{ML} = h(\widehat{\theta}_{ML})$ is a value of η which maximises $p(x,\eta)$. Similarly, if θ does not maximise $p(x,\theta)$, then $\eta = h(\theta)$ does not maximise the reparametrised family $p(x,\eta)$.

(b) Using the hint, $\widehat{\omega}_{MLE}$ maximises $\sup_{\theta\in\Omega(\omega)} L(\theta; X)$. ω_{ML} is therefore the value of ω that satisfies $\widehat{\theta}_{ML} \in \Omega(\widehat{\omega}_{ML})$ and is therefore (by definition) $\widehat{\omega}_{ML} = q(\widehat{\theta}_{ML})$.

12. (Omitted)

Chapter 6

The Information Inequality

6.1 The Information Inequality: One Parameter

Among several possible unbiased estimators, one problem is to find the estimator with the smallest variance. The following gives a theoretical lower bound on the variance of estimators.

Throughout the discussion, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ is a *regular* parametric family and that $\Theta \subset \mathbb{R}$ is open. Furthermore, $p(x, \theta)$ is a density for each $\theta \in \Theta$. The results and discussion for discrete variables are essentially similar and therefore left as an exercise. The discussion also requires the following regularity assumptions:

1. $A = \{x : p(x, \theta) > 0\}$ does not depend on θ and for all $x \in A$ and $\theta \in \Theta$, $\frac{\partial}{\partial \theta} \log p(x, \theta)$ exists and is finite.
2. If T is any statistic such that $\mathbb{E}_\theta[|T|] < +\infty$ for all $\theta \in \Theta$ then the operations of integration and differentiation can be exchanged:

$$\frac{\partial}{\partial \theta} \int T(x) p(x, \theta) dx = \int T(x) \frac{\partial}{\partial \theta} p(x, \theta).$$

Assumption 2. is practically useless as written. If Assumption 1. holds, then it is straightforward to show that Assumption 2. holds provided that for all T such that $\mathbb{E}_\theta[|T|] < +\infty$ for all $\theta \in \Theta$, the integrals

$$\int T(x) \left(\frac{\partial}{\partial \theta} p(x, \theta) \right) dx \qquad \text{and} \qquad \int \left| T(x) \left(\frac{\partial}{\partial \theta} p(x, \theta) \right) \right| dx \tag{6.1}$$

are continuous functions of θ.

Proposition 6.1. *Let $p(x, \theta) = h(x) \exp\{\eta(\theta) T(x) - B(\theta)\}$ be an exponential family, where $\eta(\theta)$ has non-vanishing continuous derivative on Θ. Then Assumptions 1. and 2. hold.*

Proof This is standard and uses the dominated convergence theorem. Let

$$p_\delta(x, \theta) = \frac{1}{\delta} \left(p(x, \theta + \delta) - p(x, \theta) \right)$$

Then for any finite $\delta > 0$,

$$\frac{\int T(x)p(x,\theta+\delta)dx - \int T(x)p(x,\theta)dx}{\delta} = \int T(x)p_\delta(x,\theta)dx.$$

The left hand side converges to $\frac{d}{d\theta}\int T(x)p(x,\theta)dx$ while $p_\delta(x,\theta) \overset{\delta\to 0}{\longrightarrow} 0$ pointwise. We need to use the dominated convergence theorem to conclude the result. This requires a dominating function. Note:

$$p_\delta(x,\theta) = p(x,\theta)\frac{1}{\delta}\left(e^{((\eta(\theta+\delta)-\eta(\theta))T(x)-(B(\theta+\delta)-B(\theta))} - 1\right)$$

and, under hypothesis of non-vanishing continuous derivative of $\eta(\theta)$, such a function may be constructed. □

If Assumption 1. holds, it is possible to define an important characteristic of the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$, the *Fisher Information*

Definition 6.2 (Fisher Information)**.** *The Fisher information, denoted $I(\theta)$, is defined for regular parametric families, where Θ is an open subset of $\mathbb{R}$, as:*

$$I(\theta) := \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right)^2\right] = \int\left(\frac{\partial}{\partial\theta}\log p(x,\theta)\right)^2 p(x,\theta)dx.$$

The following lemma is trivial:

Lemma 6.3. *Suppose that Assumptions 1. and 2. hold and that*

$$\mathbb{E}_\theta\left[\left|\frac{\partial}{\partial\theta}\log p(X,\theta)\right|\right] < +\infty.$$

Then

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p(X,\theta)\right] = 0$$

from which

$$I(\theta) = Var\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right).$$

Proof

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p(X,\theta)\right] = \int\frac{\frac{\partial}{\partial\theta}p(x,\theta)}{p(x,\theta)}p(x,\theta)dx = \frac{\partial}{\partial\theta}\int p(x,\theta)dx = \frac{\partial}{\partial\theta}1 = 0.$$

Now use the fact that for a random variable Y satisfying $\mathbb{E}[Y^2] < +\infty$ and $\mathbb{E}[Y] = 0$,

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2].$$

□

Theorem 6.4. *Let $T(X)$ be any statistic such that $Var_\theta(T(X)) < +\infty$ for all $\theta \in \Theta$. Let $\psi(\theta) := \mathbb{E}_\theta[T(X)]$. Suppose that Assumptions 1. and 2. hold and that $0 < I(\theta) < +\infty$ for all $\theta \in \Theta$. Then for all $\theta \in \Theta$, $\psi(\theta)$ is differentiable and*

$$Var_\theta(T(X)) \geq \frac{(\psi'(\theta))^2}{I(\theta)}.$$

Proof

$$\psi(\theta) = \mathbb{E}_\theta[T(X)] = \int T(x)p(x;\theta)dx.$$

From the assumptions, the operations of $\int .dx$ and $\frac{\partial}{\partial\theta}$ can be exchanged so that:

$$\psi'(\theta) = \int T(x)\frac{\partial}{\partial\theta}p(x,\theta)dx = \int T(x)\left(\frac{\partial}{\partial\theta}\log p(x,\theta)\right)p(x,\theta)dx.$$

This uses the simple fact that for a differentiable function f, $\frac{d}{dy}\log f(y) = \frac{\frac{df}{dy}}{f}$, so that

$$\frac{\partial}{\partial\theta}p(x;\theta) = \left(\frac{\partial}{\partial\theta}\log p(x;\theta)\right)p(x;\theta).$$

Since $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p(X,\theta)\right] = 0$, therefore:

$$\psi'(\theta) = \text{Cov}_\theta\left(\frac{\partial}{\partial\theta}\log p(X,\theta), T(X)\right). \tag{6.2}$$

This follows, because for two random variables X and Y such that $\mathbb{E}[X^2] < +\infty$ and $\mathbb{E}[Y^2] < +\infty$, if $\mathbb{E}[X] = 0$ then

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

Hence, directly from the Cauchy Schwartz inequality:

$$|\text{Cov}(X,Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}$$

it follows that:

$$|\psi'(\theta)| \leq \sqrt{\text{Var}_\theta(T(X))\text{Var}_\theta\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right)} = \sqrt{\text{Var}_\theta(T(X))I(\theta)} \tag{6.3}$$

and hence

$$\text{Var}_\theta(T(X)) \geq \frac{\psi'(\theta)^2}{I(\theta)}$$

as required. □

The following corollary is immediate.

Corollary 6.5. *Suppose that the conditions of Theorem 6.4 hold. Suppose, further, that $T(X)$ is an unbiased estimator of θ; that is, $\mathbb{E}_\theta[T(X)] = \theta$ for all $\theta \in \Theta$. Then*

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

Proof Directly from $\psi(\theta) = \theta$, hence $\psi'(\theta) = 1$. □

The quantity $\frac{1}{I(\theta)}$ is often referred to as the *Craméŕ - Rao lower bound* for the variance of an unbiased estimator of θ.

Proposition 6.6. *Suppose that $X = (X_1, \ldots, X_n)$ is a random sample from a population with density $p(x, \theta)$, $\theta \in \Theta$ and that the conditions of Theorem 6.4 hold. Let*

$$I_1(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log p(X_1, \theta) \right)^2 \right].$$

Then

$$I(\theta) = nI_1(\theta) \qquad \textit{and} \qquad Var_\theta(T(X)) \geq \frac{(\psi'(\theta))}{nI_1(\theta)}.$$

Proof This follows directly:

$$I(\theta) = \text{Var}\left(\frac{\partial}{\partial \theta} \log p(X, \theta) \right) = \text{Var}\left(\sum_{j=1}^n \frac{\partial}{\partial \theta} \log p(X_i, \theta) \right) = \sum_{j=1}^n \text{Var}\left(\frac{\partial}{\partial \theta} \log p(X_i, \theta) \right) = nI_1(\theta).$$

□

$I_1(\theta)$ is the information contained in a single observation. The information in a random sample of size n is $I(\theta) = nI_1(\theta)$.

There is another identity that is sometimes useful.

Lemma 6.7. *Suppose that, in addition to Assumptions 1. and 2., $p(., \theta)$ is twice differentiable and interchange between integration and differentiation is permitted. Then*

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log p(X, \theta) \right] \tag{6.4}$$

Proof This follows simply by differentiating;

$$\frac{\partial^2}{\partial \theta^2} \log p(x, \theta) = \frac{\partial}{\partial \theta} \left(\frac{\frac{\partial}{\partial \theta} p(x, \theta)}{p(x, \theta)} \right) = \frac{1}{p(x, \theta)} \frac{\partial^2}{\partial \theta^2} p(x, \theta) - \left(\frac{\partial}{\partial \theta} \log p(x, \theta) \right)^2.$$

It follows, by integrating with respect to $p(x, \theta)$, that

$$-\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta} \log p(X, \theta) \right] = -\frac{\partial^2}{\partial \theta^2} \int p(x, \theta) dx + \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log p(X, \theta) \right)^2 \right]$$

from which the result follows. □

The following indicates that if there exists an unbiased estimator which achieves the lower bound, then the family is necessarily an exponential family.

Theorem 6.8. *Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a parameteric family which satisties Assumptions 1. and 2. and suppose that there exists an unbiased estimator T^* of $\psi(\theta)$ which achieves the lower bound of Theorem 6.4. Then $\mathcal{P}$ is a one-parameter exponential family with density function of the form*

$$p(x, \theta) = h(x) \exp \{\eta(\theta) T^*(x) - B(\theta)\}. \tag{6.5}$$

Conversely, if $\{\mathbb{P}_\theta\}$ is a one-parameter exponential family of the form of Equation (6.5) *with natural sufficient statistic $T(X)$ and $\eta(\theta)$ has a continuous non-vanishing derivative on Θ then $T(X)$ achieves the Cramér Rao lower bound and is a UMVU estimate of $\mathbb{E}_\theta [T(X)]$.*

Proof The proof presented here is essentially that of Wijsman [10] (1973). By Equation (6.2) and the conditions necessary to achieve equality in (6.3), it follows that T^* achieves the Cramér Rao lower bound for all θ if and only if there exist functions $a_1(\theta)$ and $a_2(\theta)$ such that

$$\frac{\partial}{\partial \theta} \log p(X, \theta) = a_1(\theta) T^*(X) + a_2(\theta)$$

with $\mathbb{P}_\theta$ probability 1 for each θ. Let

$$A^* = \left\{ x : \frac{\partial}{\partial \theta} \log p(x, \theta) = a_1(\theta) T^*(x) + a_2(\theta) \qquad \forall \theta \in \Theta \right\}.$$

It is required to prove that $\mathbb{P}_\theta(X \in A^*) = 1$ for all θ. Once this has been shown, integrating with respect to θ gives Equation (6.5).

The proof that $\mathbb{P}_\theta(X \in A^*)$ for all $\theta \in \Theta$ is a straightforward piece of analysis. Firstly, let

$$A_\theta = \left\{ x : \frac{\partial}{\partial \theta} \log p(x, \theta) = a_1(\theta) T^*(x) + a_2(\theta) \right\}.$$

then, from the definition, $\mathbb{P}_\theta(X \in A_\theta) = 1$. From Assumption 1., for a given $\theta \in \Theta$, it follows that $\mathbb{P}_{\theta'}(X \in A_\theta) = 1$ for all $\theta' \in \Theta$. Now let Θ^* denote a countable dense set of Θ and let $A^{**} = \cap_{\theta \in \Theta^*} A_\theta$. Then $\mathbb{P}_{\theta'}(A^{**}) = 1$ for all $\theta' \in \Theta$.

Choose an $x_1 \neq x_2$, $x_1, x_2 \in A^{**}$ such that $T^*(x_1) \neq T^*(x_2)$. Then a_1 and a_2 are linear combinations of $\frac{\partial}{\partial \theta} \log p(x_j, \theta)$ for $j = 1, 2$ and therefore continuous in θ, since (by hypothesis) $\frac{\partial}{\partial \theta} \log p(x, \theta)$ is continuous in θ for each x. It therefore follows that $\mathbb{P}_\theta(A^*) = 1$ for all $\theta \in \Theta$ and the result follows.

Conversely, for an exponential family,

$$\frac{\partial}{\partial \theta} \log p(x, \theta) = \eta'(\theta) \left(T(x) - \dot{A}(\eta(\theta)) \right).$$

It follows that

$$I(\theta) = \left(\eta'(\theta)\right)^2 \mathrm{Var}_\theta \left(T(X)\right) = \left(\eta'(\theta)\right)^2 \ddot{A}(\eta(\theta))$$

where $\dot{A}$ denotes single derivative with respect to η and $\ddot{A}$ denotes double derivative with respect to η.

This follows from:

$$\begin{cases} \dot{A}(\eta) = \mathbb{E}_\eta[T(X)] \\ \ddot{A}(\eta) = \mathbb{E}_\eta[T(X)^2] - 2\dot{A}(\eta)\mathbb{E}[T(X)] + \dot{A}(\eta)^2 = \mathbb{E}_\eta[T(X)^2] - \mathbb{E}_\eta[T(X)]^2 = \mathrm{Var}_\eta(T(X)). \end{cases}$$

Let $\psi(\theta) = \mathbb{E}_\theta[T(X)] = \dot{A}(\eta(\theta))$. It follows that $\psi'(\theta) = \ddot{A}(\eta(\theta))\eta'(\theta)$. The information bound is therefore

$$\mathrm{Var}_\theta(T(X)) \geq \frac{\left(\ddot{A}(\eta)\eta'(\theta)\right)^2}{(\eta'(\theta))^2 \ddot{A}(\eta(\theta))} = \ddot{A}(\eta(\theta)),$$

Since $\mathrm{Var}_\theta(T(X)) = \ddot{A}(\eta(\theta))$, it follows that $T(X)$ achieves the information bound and is a UMVU estimator of $\mathbb{E}_\theta\left[T(X)\right]$. □

Example 6.1 (Poisson Random Sample)**.**

Let $X_1, \ldots, X_n$ be a random sample from a Poiss(θ) population. Then

$$\frac{\partial}{\partial\theta}\log p(x,\theta) = \frac{1}{\theta}\sum_{j=1}^n x_j - n.$$

The information in the sample for parameter θ is:

$$I(\theta) = \mathrm{Var}_\theta\left(\frac{1}{\theta}\sum_{j=1}^n X_j\right) = \frac{1}{\theta^2}n\theta = \frac{n}{\theta}.$$

The MLE is $\widehat{\theta} = \overline{X}$. It is unbiased; $\mathbb{E}_\theta\left[\overline{X}\right] = \theta$ and it satisfies

$$\mathrm{Var}(\overline{X}) = \frac{\theta}{n}.$$

It achieves the Cramér Rao lower bound and hence $\overline{X}$ is a UMVU estimator of θ.

For a Poisson random sample,

$$\frac{\partial^2}{\partial^2\theta}\log p(x,\theta) = -\frac{1}{\theta^2}\sum_{j=1}^n x_j$$

and hence

$$-\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(x,\theta)\right] = \frac{1}{\theta^2}\mathbb{E}\left[\sum_{j=1}^n X_j\right] = \frac{n}{\theta} = I(\theta).$$

□

Example 6.2 (Hardy-Weinberg Model)**.**

Consider a random sample from a distribution where $p(1) = \theta^2$, $p(2) = 2\theta(1-\theta)$ and $p(3) = (1-\theta)^2$. Let $X = (X_1, \dots, X_n)$ be a random sample. Then

$$\begin{aligned} p(\underline{x}, \theta) &= \theta^{2n_1}(2\theta(1-\theta))^{n_2}(1-\theta)^{2n_3} \\ &= 2^{n_2}\exp\{(2n_1+n_2)\log\theta + (n_2+2n_3)\log(1-\theta)\} \\ &= 2^{n_2}\exp\left\{(2n_1+n_2)\log\left(\frac{\theta}{1-\theta}\right) + 2n\log(1-\theta)\right\}. \end{aligned}$$

This is an exponential family. Note that, for any c we can take $T(X) = c(2N_1+N_2)$ and $\eta(\theta) = \frac{1}{c}\log\left(\frac{\theta}{1+\theta}\right)$.

Now,

$$\mathbb{E}_\theta[2N_1+N_2] = n(2\theta^2 + 2\theta(1-\theta)) = n\theta.$$

Take $c = \frac{1}{2n}$ so that $T(X) = \frac{2N_1+N_2}{2n}$ and $\mathbb{E}_\theta[T(X)] = \theta$.

Therefore, by Theorem 6.8, $T(X) = \frac{2N_1+N_2}{2n}$ is a UMVU estimator of θ.
For an exponential family, $T(X)$ is the ML estimator of $\mathbb{E}_\eta[T(X)]$. Therefore, the ML estimator of θ is:

$$\widehat{\theta} = \frac{2N_1+N_2}{2n}$$

and:

$$\mathrm{Var}_\theta(\widehat{\theta}) = \frac{1}{I(\theta)} = \ddot{A}(\eta(\theta)).$$

Here $\eta = 2n\log\frac{\theta}{1-\theta}$ so

$$A(\eta) = -2n\log(1-\theta) = 2n\log(1+e^{\eta/2n})$$

$$\ddot{A}(\eta) = \frac{1}{2n}\theta(1-\theta)$$

so that

$$I(\theta) = \frac{2n}{\theta(1-\theta)}, \qquad \mathrm{Var}_\theta(\widehat{\theta}_{ML}) = \frac{\theta(1-\theta)}{2n}.$$

6.2 The Optimal Linear Predictor

6.2.1 Introduction and Motivation

The *optimal linear predictor* is very useful in its own right. It is introduced here so that it can be used to compute an information lower bound where $\theta = (\theta_1, \ldots, \theta_d)$ is a parameter vector. If $T(X)$ is an unbiased estimator of $\psi(\theta)$, then the idea is to find the linear function of $\nabla_\theta \log p(X, \theta)$,

$$\widehat{T} := b + \sum_{j=1}^{d} a_j \frac{\partial}{\partial \theta_j} \log p(X, \theta) \qquad \text{satisfying} \qquad \mathbb{E}_\theta[\widehat{T}] = \psi(\theta),$$

which, under this constraint, minimises $\text{Var}(T(X) - \widehat{T})$. Then standard properties of linear predictors, namely that $\text{Var}(\widehat{T}) \leq \text{Var}(T(X))$, can be used to obtain the information lower bound.

The optimal linear predictor is well used in a variety of settings; for example, it is a key component in analysis of stationary time series, where future behaviour is predicted based on past information of the time series, together with the mean and autocovariance functions.

6.2.2 The Linear Predictor

Consider a random variable Y and a random vector $X = (X_1, \ldots, X_n)$, where X is observable. Let Σ denote the covariance matrix of X and suppose that Σ is positive definite. Suppose we want to find the linear combination

$$\widehat{Y} = \sum_{j=1}^{n} a_j X_j + b \qquad (a_1, \ldots, a_n, b) \in \mathbb{R}^{n+1} \tag{6.6}$$

such that

$$\mathbb{E}\left[\widehat{Y}\right] = \mathbb{E}\left[Y\right] \tag{6.7}$$

which minimises $\text{Var}(Y - \widehat{Y})$ subject to (6.7). The matrix Σ has entries $\Sigma_{ij} = \text{Cov}(X_i, X_j)$. Let C denote the vector with entries $C_i = \text{Cov}(Y, X_i)$. Let $a = (a_1, \ldots, a_n)^t$. Then

$$\text{Var}(\widehat{Y} - Y) = a^t \Sigma a + \text{Var}(Y) - 2a^t C.$$

From this it follows that $a = \Sigma^{-1} C$ and therefore that

$$\widehat{Y} = \mathbb{E}[Y] + (X - \mathbb{E}[X])^t \Sigma^{-1} C. \tag{6.8}$$

The crucial lemma about linear predictors, which enables us to obtain the information lower bound is that the variance of the predictor is bounded from above by the variance of the variable that it is trying to predict.

Lemma 6.9.

$$Var(\widehat{Y}) \leq Var(Y).$$

Proof

$$0 \leq \text{Var}(Y - \widehat{Y}) = \text{Var}(Y) + \text{Var}(\widehat{Y}) - 2\text{Cov}(Y, \widehat{Y}).$$

while

$$\text{Cov}(Y, \widehat{Y}) = C^t \Sigma^{-1} C$$

and

$$\text{Var}(\widehat{Y}) = C^t \Sigma^{-1} \Sigma \Sigma^{-1} C = C^t \Sigma^{-1} C = \text{Cov}(Y, \widehat{Y}).$$

It follows that

$$0 \leq \text{Var}(Y - \widehat{Y}) = \text{Var}(Y) - \text{Var}(\widehat{Y})$$

and the result follows. □

6.3 Information Inequality with Several Parameters

Now suppose that $\theta = (\theta_1, \ldots, \theta_d)$. In this section, the following conditions are assumed:

- Θ is an open subset of $\mathbb{R}^d$.
- $\mathcal{P} = \{p(x, \theta) : \theta \in \Theta\}$ is a regular parametric model.
- Assumptions 1. and 2. are satisfied for differentiation with respect to θ_j, for any $j = 1, \ldots, d$.

Definition 6.10 (Fisher Information Matrix)**.** *The* Fisher Information matrix *is defined as the $d \times d$ matrix $I(\theta)$ where*

$$I_{ij}(\theta) = \mathbb{E}\left[\frac{\partial}{\partial \theta_i} \log p(X, \theta) \frac{\partial}{\partial \theta_j} \log p(X, \theta)\right].$$

The following proposition is a straightfoward generalisation of the single parameter discussion.

Proposition 6.11. *Under the conditions stated above,*

1. $$\mathbb{E}_\theta\left[\frac{\partial}{\partial \theta_j} \log p(X, \theta)\right] = 0.$$

2. $$I_{jk}(\theta) = Cov_\theta\left(\frac{\partial}{\partial \theta_j} \log p(X, \theta), \frac{\partial}{\partial \theta_k} \log p(X, \theta)\right).$$

3. *If $X_1, \ldots, X_n$ are i.i.d. then $X = (X_1, \ldots, X_n)^t$ has information matrix $nI_1(\theta)$, where I_1 is the information matrix of a single observation.*

4. *If $p(., \theta)$ is twice differentiable and double integration and differentiation under the integral sign can be interchanged,*

$$I_{jk}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log p(X, \theta)\right] \qquad (j, k) \in (1, \ldots, d)^2.$$

Proof Straightforward. □

Theorem 6.12. *Let T be an estimator for $q(\theta_1, \ldots, \theta_d)$ where the parameters $\theta_1, \ldots, \theta_d$ are unknown. Let $\psi(\theta) = \mathbb{E}_\theta [T(X)]$. Let $\dot{\psi}(\theta) = \nabla_\theta \psi(\theta)$. Suppose that the information matrix $I(\theta)$ is non singular. Then for all θ, $\dot{\psi}(\theta)$ exists and*

$$Var_\theta(T(X)) \geq \dot{\psi}(\theta)^t I^{-1}(\theta) \dot{\psi}(\theta).$$

Proof Let $Z = \nabla_\theta \log p(X, \theta)$, C_θ the vector with entries $C_{\theta,i} = \text{Cov}_\theta(Z_i, T(X))$. It follows that P, defined by

$$P = \psi(\theta) + (Z - \mathbb{E}_\theta[Z])^t I(\theta)^{-1} C$$

is the linear predictor of $T(X)$ with least variance among unbiased linear predictors. It follows from Lemma 6.9 that $\text{Var}_\theta(T(X)) \geq \text{Var}_\theta(P)$ for each $\theta \in \Theta$ and hence that

$$\text{Var}_\theta(T(X)) \geq C^t I(\theta)^{-1} C.$$

The argument is concluded by noting that

$$\dot{\psi}_j(\theta) = \int T(x) \left(\frac{\partial}{\partial \theta_j} p(x, \theta) \right) dx = \int T(x) \left(\frac{\partial}{\partial \theta_j} \log p(x, \theta) \right) p(x, \theta) dx = C_j.$$

□

6.4 Information and Log Partition Function for Exponential Families

The following lemma characterises the Fisher Information matrix in terms of the second derivatives of the log partition function when dealing with an exponential family in its canonical parametrisation.

Lemma 6.13. *Let $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in \mathcal{E}\}$ be a canonical exponential family generated by (h, T). Let $A(\eta)$ be the log partition function. Suppose that A is twice differentiable in $\mathcal{E}$ (the natural parameter space) and let $\ddot{A}(\eta)$ be the matrix of second order partial derivatives. Then*

$$\ddot{A}(\eta) = \Sigma^{(T)} = I(\eta)$$

where $\Sigma^{(T)}$ denotes the covariance matrix of the natural sufficient statistics and $I(\eta)$ denotes the Fisher Information Matrix.

Proof

$$p(x, \eta) = h(x) \exp \left\{ \sum_{j=1}^k \eta_j T(x) - A(\eta) \right\}$$

so that

$$\log p(x, \eta) = \log h(x) + \sum_{j=1}^k \eta_j T(x) - A(\eta)$$

giving

$$-\frac{\partial^2}{\partial\eta_i\eta_j}\log p(x,\eta) = \frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\eta)$$

so that

$$I_{i,j}(\eta) = -\mathbb{E}_\eta\left[\frac{\partial^2}{\partial\eta_i\eta_j}\log p(X,\eta)\right] = \frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\eta).$$

It follows that $I(\eta) = A(\eta)$.

The other identity is an earlier exercise: let $M_\eta(s) = \mathbb{E}_\eta\left[e^{\langle s,T(X)\rangle}\right]$, then

$$M_\eta(s) = e^{A(s+\eta)-A(s)}$$

giving

$$\mathbb{E}_\eta\left[T_i(X)\right] = \left.\frac{\partial A}{\partial s_i}(s+\eta)\right|_{s=0}$$

and

$$\mathbb{E}_\eta\left[T_i(X)T_j(X)\right] = \left.\frac{\partial^2}{\partial s_i\partial s_j}M_\eta(s)\right|_{s=0} = \left.\left(\frac{\partial^2}{\partial s_i\partial s_j}A(\eta+s) + \frac{\partial A}{\partial s_i}(\eta+s)\frac{\partial A}{\partial s_j}(\eta+s)\right)\right|_{s=0}$$

so that

$$\ddot{A}(\eta) = \Sigma^{(T)}.$$

□

Summary

- Uniformly Minimum Variance Unbiased Estimation
- Fisher Information.
- Cramér Rao lower bound (Information inequality)
- Theorem: under regularity assumptions on log likelihood, the Cramér Rao lower bound is achieved if and only if the parametric family is an exponential family.

Tutorial 7

1. Let $X_1, \ldots, X_n$ be a random sample from distribution $p_\theta(0) = 1 - \theta$, $p_\theta(1) = \theta$, where the parameter $\theta \in (0,1)$. Show that $\overline{X}$ is a UMVU (uniformly minimum variance unbiased) estimator of θ.

2. Let $X \sim \text{Binomial}(n, \theta)$. In other words

$$\mathbb{P}(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad x = 0, 1, \ldots, n$$

   Consider the estimators $\widehat{\theta} = \frac{X}{n}$ and $\widetilde{\theta} = \frac{X+1}{n+2}$. Compute the bias of these estimators, their variance and their mean squared errors. Are they consistent?

3. Let $(X_1, \ldots, X_n)$ be a sample from a Poisson distribution with unknown parameter λ. To estimate λ, we estimate $g(\lambda) := \mathbb{P}(X_1 = 0) = e^{-\lambda}$ and we consider two estimators $\widehat{g}_1$, $\widehat{g}_2$ of g, where

$$\widehat{g}_1 = e^{-\overline{X}}, \qquad \widehat{g}_2 = \left(1 - \frac{1}{n}\right)^{n\overline{X}},$$

   where $\overline{X}$ is the sample average. Compute the bias of the estimators $\widehat{g}_1$ and $\widehat{g}_2$.

4. Let $(X_1, \ldots, X_n)$ be a random sample from a Bernoulli(p) distribution (that is, $\mathbb{P}(X_j = 1) = p$, $\mathbb{P}(X_j = 0) = 1 - p$). Show that there do not exist unbiased estimators of the quantities

$$g_1(p) = \frac{p}{1-p}, \qquad g_2(p) = \frac{1}{p}.$$

5. Let $(X_1, X_2, \ldots, X_n)$ be a random sample with unknown expected value μ and known variance σ^2.

   (a) Show that the statistic

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{n} a_i X_i \qquad \sum_{i=1}^{n} a_i = 1$$

   is an unbiased estimator of μ.

   (b) Compute the variance of T and show that, for unbiased estimators of this form, it is minimised for $a_i = \frac{1}{n}$, $i = 1, \ldots, n$.

6. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p) random variables. Find the Cramer-Rao lower bound for the variance of an unbiased estimator of $g(p) = p(1-p)$.

7. Let $(X_1, \ldots, X_n)$ be a random sample from an exponential distribution Exp(λ). That is, the density function is:

$$f(x; \lambda) = \lambda e^{-\lambda x} \mathbf{1}_{\{x \geq 0\}}.$$

   Show that the statistic $T(X_1, \ldots, X_n) = n X_{1:n}$ is an unbiased estimator of $\frac{1}{\lambda}$, but that it is not consistent.

8. Suppose θ is an unknown parameter to be estimated, and let $f(X)$ be the estimator. Let $l(\theta, a)$ be a *loss function*, where the *loss* incurred when estimating θ by $f(X)$ is given by $l(\theta, f(X))$. The *risk function* is defined as:

$$R(\theta, f) = \mathbb{E}_\theta \left[l(\theta, f(X)) \right].$$

Suppose that the loss function $l(\theta, a)$ is strictly convex in the variable a. Suppose $g(X)$ is an unbiased estimator of $q(\theta)$ and that $T(X)$ is a sufficient statistic. Let $g^*(X) = \mathbb{E}_\theta \left[g(X) | T(X) \right]$. Show that $R(\theta, g^*) \leq R(\theta, g)$.

Hint Jensen's inequality: if ϕ is a convex function and X a random variable then

$$\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X]).$$

9. Let X have density or probability mass function $p(x, \theta)$, where $\theta \in \Theta \subset \mathbb{R}$. Suppose that the following assumptions hold:

- $\{x : p(x, \theta) > 0\}$ is the same for all $\theta \in \Theta$ and
- For any statistic T satisfying $\mathbb{E}[|T(X)|] < +\infty$ for all $\theta \in \Theta$,

$$\int T(x) \frac{\partial}{\partial \theta} p(x, \theta) dx = \frac{\partial}{\partial \theta} \int T(x) p(x, \theta) dx.$$

Suppose that h is monotone increasing and differentiable from Θ to $h(\Theta)$. Let $\eta = h(\theta)$ and $q(x, \eta) = p(x, h^{-1}(\eta))$.

(a) Let $I_p(\theta)$ and $I_q(\eta)$ denote the Fisher information in the two parametrisations. Show that

$$I_q(\eta) = \frac{1}{(h'(h^{-1}(\eta)))^2} I_p(h^{-1}(\eta)).$$

(b) Let $B_p(\theta)$ and $B_q(\eta)$ denote the information inequality lower bound for the two parametrisations. That is,

$$B_p(\theta) = \frac{(\psi_1'(\theta))^2}{I(\theta)}, \qquad B_q(\eta) = \frac{(\psi_2'(\eta))^2}{I(\eta)}$$

where $\psi_1(\theta)$ is the quantity to be estimated and $\psi_2(\eta) = \psi_1(h^{-1}(\eta))$; i.e. the same quantity under the η-parametrisation.

Show that $B_q(\eta) = B_p(h^{-1}(\eta))$. That is, the Fisher information lower bound is the same.

10. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where μ is known.

(a) Show that

$$\widehat{\sigma}^2 := \frac{1}{n} \sum_{j=1}^{n} (X_j - \mu)^2$$

is a UMVU estimator of σ^2.

(b) For parameter estimation, a *decision rule* d simply means assigning a decision $d(X)$ for the unknown parameter. A decision rule d is *inadmissible* if there is another decision rule d^* such that $R(\theta, d^*) \leq R(\theta, d)$ for all θ and $R(\theta, d^*) < R(\theta, d)$ for some θ. A decision rule is *admissible* if it is not inadmissible. Show that $\widehat{\sigma}^2$ is inadmissible under squared loss error $R(\theta, d) = \mathbb{E}_\theta\left[|\theta - d(X)|^2\right]$.

Hint Consider bias estimators of the form $a_n\widehat{\sigma}^2$ where $\widehat{\sigma}^2$ is defined above.

11. Let $Y_1, \ldots, Y_n$ be independent Poisson random variables, where

$$\mathbb{E}[Y_j] = \mu_j = \exp\left\{\alpha + \beta z_j\right\}.$$

(For example, z_j could be the level of a drug given to the jth patient with an infectious disease, and Y_j could denote the number of infectious microbes found in a unit of blood taken from patient j 24 hours after the drug was administered.

(a) Write the model for $Y_1, \ldots, Y_n$ in two-parameter canonical exponential form and give the sufficient statistic.

(b) Let $\theta = (\alpha, \beta)$. Compute $I(\theta)$, the information matrix for the model and find the lower bound on the variances of unbiased estimators $\widehat{\alpha}$ and $\widehat{\beta}$ of α and β.

(c) Let $z_i = \log\left(\frac{i}{n+1}\right)$, $i = 1, \ldots, n$. Compute $\lim_{n\to+\infty} \frac{1}{n} I(\theta)$ and give the limit of n times the lower bound on the variances of $\widehat{\alpha}$ and $\widehat{\beta}$.

Hint Use integral approximations for the sums.

Short Answers

1. Firstly,

$$\mathbb{E}_\theta\left[\overline{X}\right] = \mathbb{E}_\theta\left[X_1\right] = 0 \times (1-\theta) + 1 \times \theta = \theta$$

so the estimator is unbiased.

Secondly: there are several ways to show it is UMVU.

$$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) = \theta^{\sum_j x_j}(1-\theta)^{n-\sum_j x_j} = \exp\left\{n\overline{x}\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}$$

This is an exponential family; take $\overline{X}$ as the sufficient statistic and $\eta := n\log\left(\frac{\theta}{1-\theta}\right)$ as the canonical parameter. Hence, by the result in lectures, $\overline{X}$ is the UMVU estimator of its expected value, which is $\mathbb{E}_\theta[\overline{X}] = \theta$.

Alternatively, we can compute the Cramér-Rao lower bound directly. The estimator has variance

$$\mathbf{V}_\theta\left(\overline{X}\right) = \frac{1}{n}\mathbf{V}_\theta\left(X_1\right) = \frac{\theta(1-\theta)}{n}.$$

Now to show that this achieves the C-R lower bound, compute $I(\theta) = nI_1(\theta)$, the information in the sample.

$$\frac{d}{d\theta}\log p_\theta(0) = -\frac{1}{1-\theta} \qquad \frac{d}{d\theta}\log p_\theta(1) = \frac{1}{\theta}$$

$$I_1(\theta) = \mathbb{E}_\theta\left[\left(\frac{d}{d\theta}\log p_\theta(X)\right)^2\right] = (1-\theta)\frac{1}{(1-\theta)^2} + \theta\frac{1}{\theta^2} = \frac{1}{\theta(1-\theta)}$$

hence the information from a sample size n is:

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

and the C-R lower bound is

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

hence estimator is UMVU.

2.

$$\mathbb{E}[\widehat{\theta}] = \frac{1}{n}\mathbb{E}[X] = \frac{1}{n}n\theta = \theta$$

so this estimator is unbiased.

$$\mathbf{V}(\widehat{\theta}) = \frac{1}{n^2}\mathbf{V}(X) = \frac{1}{n^2}n\theta(1-\theta) = \frac{\theta(1-\theta)}{n} < \frac{1}{4n}$$

so

$$\mathbb{P}(|\widehat{\theta}-\theta| > \epsilon) \le \frac{\theta(1-\theta)}{\epsilon^2 n} \le \frac{1}{4\epsilon^2 n} \overset{n\to+\infty}{\longrightarrow} 0$$

hence uniformly consistent.

For $\widetilde{\theta}$

$$\mathbb{E}[\widetilde{\theta}] = \frac{n\theta+1}{n+2} = \frac{n}{n+2}\theta + \frac{1}{n+2}$$

so

$$\text{Bias}(\widetilde{\theta}) = \frac{n}{n+2}\theta + \frac{1}{n+2} - \theta = \frac{1-2\theta}{n+2}$$

$$\mathbf{V}(\widetilde{\theta}) = \frac{n\theta(1-\theta)}{(n+2)^2}$$

so that

$$\mathbb{E}\left[|\widetilde{\theta}-\theta|^2\right] = \frac{(1-2\theta)^2}{(n+2)^2} + \frac{n\theta(1-\theta)}{(n+2)^2} = \frac{1+(n-4)\theta(1-\theta)}{(n+2)^2}$$

Yes - the estimator is uniformly consistent; for $n \geq 4$,

$$\mathbb{P}(|\widetilde{\theta}-\theta|^2 > \epsilon) \leq \frac{1+(n-4)\theta(1-\theta)}{\epsilon^2(n+2)^2} \leq \frac{1+(n-4)/4}{\epsilon^2(n+2)^2} \overset{n\to+\infty}{\longrightarrow} 0.$$

3. $Y := n\overline{X} = \sum_{j=1}^n X_j \sim \text{Poiss}(n\lambda)$ so that

$$\mathbb{E}[\widehat{g}_1] = \mathbb{E}\left[e^{-Y/n}\right] = \sum_{x=0}^{\infty} \frac{(\lambda n)^x}{x!} e^{-n\lambda-(x/n)} = \sum_{x=0}^{\infty} \frac{(\lambda n e^{-1/n})^x}{x!} e^{-n\lambda} = e^{-n\lambda(1-e^{-1/n})}$$

so

$$\text{Bias}(\widehat{g}_1) = e^{-\lambda}\left(e^{-\lambda(n(1-e^{-1/n})-1)} - 1\right).$$

$$\mathbb{E}[\widehat{g}_2] = \sum_{x=0}^{\infty} \frac{\left((n\lambda)(1-\frac{1}{n})\right)^x}{x!} e^{-\lambda n} = e^{n\lambda-\lambda-n\lambda} = e^{-\lambda}$$

so that $\widehat{g}_2$ is unbiased.

4. An unbiased estimator of $g_1(p)$ is a function of the n binary variables $T(X_1, \ldots, X_n)$ satisfying

$$\mathbb{E}[T(X_1,\ldots,X_n)] = \sum_{\{0,1\}^n} T(x_1,\ldots,x_n)p^k(1-p)^{n-k} = \frac{p}{1-p}$$

where k denotes the number of 1s in the sequence $(x_1, \ldots, x_n)$.

so that

$$0 = \sum_{\{0,1\}^n} T(x_1,\ldots,x_n)p^{k-1}(1-p)^{n-k+1} - 1.$$

This holds for all p, which is a contradiction, since the equation is a polynomial of degree $n+1$ and hence has at most $n+1$ distinct roots. Similarly for g_2.

5. (a)

$$\mathbb{E}[T] = \mu\sum_{i=1}^n a_i = \mu.$$

(b)

$$\mathbf{V}(T) = \sigma^2 \sum_{i=1}^{n} a_i^2$$

$n-1$ free variables; $a_n = 1 - \sum_{j=1}^{n-1} a_j$ so that

$$\frac{\partial}{\partial a_i} \mathbf{V}(T) = 2\sigma^2 \left(a_i - a_n \right) = 0$$

so that $a_1 = \ldots = a_n$. With constraint that $\sum_{j=1}^{n} a_j = 1$, it follows that $a_j = \frac{1}{n}$ for each $j = 1, \ldots, n$.

6.

$$\mathbb{P}(x) = p^x (1-p)^{1-x} \Rightarrow \log \mathbb{P}(x) = x \log p + (1-x) \log(1-p) \qquad x \in \{0,1\}$$

$$\frac{d}{dp} \log \mathbb{P}(x) = \frac{d}{dp} \log \mathbb{P}(x) = \left(\frac{x}{p} - \frac{(1-x)}{1-p} \right) = \left(\frac{x}{p(1-p)} - \frac{1}{1-p} \right)$$

$$I(p) = \mathbf{V}_p \left(\frac{d}{dp} \log \mathbb{P}_p(X) \right) = \frac{1}{p^2(1-p)^2} \mathbf{V}_p(X) = \frac{1}{p(1-p)}.$$

For n observations, $I_n(g) = \frac{n}{p(1-p)}$. The Cramér lower bound is therefore:

$$\mathbf{V}(\widehat{g}) \geq \frac{(g'(p))^2}{I_n(g)} = \frac{(1-2p)^2 p(1-p)}{n}.$$

7. $\min_j X_j \sim \text{Exp}(n\lambda)$ hence

$$\mathbb{E}\left[n \min_j X_j \right] = n \frac{1}{n\lambda} = \frac{1}{\lambda}.$$

$$\mathbf{V}\left(n \min_j X_j \right) = n^2 \frac{1}{n^2 \lambda^2} = \frac{1}{\lambda^2},$$

hence $\mathbb{P}\left(|n \min_j X_j - \lambda| > \epsilon \right) \not\to 0$ as $n \to +\infty$.

8. This follows immediately from the definition and Jensen:

$$\begin{aligned} R(\theta, g^*) &= \mathbb{E}_\theta \left[l(\theta, g^*(X)) \right] = \mathbb{E}_\theta \left[l(\theta, \mathbb{E}_\theta [g(X) | T(X)]) \right] \\ &\leq \mathbb{E}_\theta \left[\mathbb{E}_\theta \left[l(\theta, g(X)) | T(X) \right] \right] = \mathbb{E}_\theta \left[l(\theta, g(X)) \right] = R(\theta, g). \end{aligned}$$

Note that we do not need T to be a *sufficient* statistic.

9. (a) This follows simply from the definition: let $\eta = h(\theta)$, then

$$\frac{d}{d\eta} \log q(x, \eta) = \frac{1}{dh(\theta)/d\theta} \frac{d}{d\theta} \log p(x, \theta)$$

$$I_q(\eta) = \mathbb{E}_\eta \left[\left(\frac{d}{d\eta} \log q(x, \eta) \right)^2 \right] = \frac{1}{(h'(\theta))^2} \mathbb{E}_\theta \left[\left(\frac{d}{d\theta} \log p(x, \theta) \right)^2 \right] = \frac{1}{h'(h^{-1}(\eta))} I_p(h^{-1}(\eta)).$$

(b) Consider a quantity $\psi(\theta) = \mathbb{E}_\theta[T(X)]$, so that $T(X)$ is an unbiased estimator of $\psi(\theta)$. The lower bound using parameter θ is:

$$\frac{(\psi'(\theta))^2}{I_p(\theta)},$$

while with parameter η:

$$\frac{d}{d\eta}\psi(h^{-1}(\eta)) = \psi'(h^{-1}(\eta))\frac{dh^{-1}(\eta)}{d\eta},$$

so that the lower bound is:

$$\frac{\left(\frac{d}{d\eta}\psi(h^{-1}(\eta))\right)^2}{I_q(\eta)} = \frac{\left(\psi'(h^{-1}(\eta)\right)^2}{I_p(h^{-1}(\eta))}.$$

10. (a) Unbiased is clear:

$$\mathbb{E}_{\sigma^2}\left[\widehat{\sigma}^2\right] = \frac{1}{n}\sum_{j=1}^n \mathbf{V}_{\sigma^2}(X_j) = \sigma^2$$

The variance is:

$$\mathbf{V}_{\sigma^2}(\widehat{\sigma}^2) = \frac{1}{n^2}\sum_{j=1}^n \mathbf{V}((X_j-\mu)^2) = \frac{\sigma^4}{n}\mathbf{V}\left(\left(\frac{X_1-\mu}{\sigma}\right)^2\right) = \frac{2\sigma^4}{n}$$

while the Fisher information is $I(\sigma^2) = nI_1(\sigma^2)$;

$$\frac{d}{d(\sigma^2)}\log p(x,\sigma^2) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}$$

Let $I_n(\sigma^2)$ denote the information in a sample of size n, then $I_n(\sigma^2) = nI_1(\sigma^2)$ and:

$$I_1(\sigma^2) = \mathbf{V}_{\sigma^2}\left(\left(\frac{d}{d(\sigma^2)}\log p(x,\sigma^2)\right)^2\right) = \frac{1}{4\sigma^4}\mathbf{V}_{\sigma^2}\left(\left(\frac{X-\mu}{\sigma}\right)^2\right)$$

and now use $V = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi_1^2$ so that $\mathbf{V}(V) = 2$. Then

$$I_1(\sigma^2) = \frac{1}{2\sigma^4} \Rightarrow I_n(\sigma^2) = \frac{n}{2\sigma^4}.$$

Hence the Cramér-Rao lower bound is $\frac{2\sigma^4}{n}$, which is $\mathbf{V}_\sigma^2(\widehat{\sigma}^2)$.
so that $I(\sigma^2) = \frac{n}{2\sigma^4}$ giving a lower bound of $\frac{2\sigma^4}{n}$, hence $\widehat{\sigma}^2$ is an UMVU estimator.

(b) For an unbiased estimator, this risk function is simply the variance since $\widehat{\sigma}^2$ is UMVU, it follows that any estimator with smaller risk must be biased. Try estimators of the form $a_n\widehat{\sigma}^2$. Then

$$\begin{aligned}
\mathbb{E}_{\sigma^2}\left[\left|a_n\widehat{\sigma}^2 - \sigma^2\right|^2\right] &= \mathbb{E}_{\sigma^2}\left[\left|a_n(\widehat{\sigma}^2-\sigma^2) + (a_n-1)\sigma^2\right|^2\right] \\
&= a_n^2\mathbf{V}_{\sigma^2}(\widehat{\sigma}^2) + \left((a_n-1)\sigma^2\right)^2 \\
&= a_n^2\frac{2\sigma^4}{n} + (a_n-1)^2\sigma^4
\end{aligned}$$

Minimising gives: $a_n = \frac{n}{n+2}$. This gives estimator $\widetilde{\sigma}^2 = \frac{1}{n+2}\sum_{j=1}^n (X_j - \mu)^2$ and $R(\sigma^2, \widetilde{\sigma}^2) = \frac{2}{n+2}\sigma^4$ which is smaller than the UMVU estimator.

11. (a)

$$p(y_1, \ldots, y_n; \alpha, \beta) = \frac{1}{\prod_{j=1}^n y_j!} \exp\left\{\alpha \sum_{j=1}^n y_j + \beta \sum_{j=1}^n y_j z_j - \sum_{j=1}^n \exp\{\alpha + \beta z_j\}\right\}$$

Sufficient statistic: $T(Y) = (T_1(Y), T_2(Y))$ where

$$T_1(Y) = \sum_{j=1}^n Y_j \qquad \text{and} \qquad T_2(Y) = \sum_{j=1}^n z_j Y_j.$$

(b)

$$\frac{\partial}{\partial \alpha} \log p = \sum_{j=1}^n y_j - \sum_{j=1}^n e^{\alpha + \beta z_j} \Rightarrow -\frac{\partial^2}{\partial \alpha^2} \log p = \sum_{j=1}^n e^{\alpha + \beta z_j}$$

$$\frac{\partial}{\partial \beta} \log p = \sum_{j=1}^n y_j z_j - \sum_{j=1}^n z_j e^{\alpha + \beta z_j} \Rightarrow -\frac{\partial^2}{\partial \beta^2} \log p = e^{\alpha} \sum_{j=1}^n z_j^2 e^{\beta z_j}$$

$$-\frac{\partial^2}{\partial \alpha \partial \beta} \log p = e^{\alpha} \sum_{j=1}^n z_j e^{\beta z_j}$$

so

$$I_{\alpha,\alpha}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \alpha^2} \log p(X, \alpha, \beta)\right] = \sum_{j=1}^n \mu_j = e^{\alpha} \sum_{j=1}^n e^{\beta z_j}$$

$$I_{\beta,\beta}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \beta^2} \log p(X; \alpha, \beta)\right] = e^{\alpha} \sum_{j=1}^n z_j^2 e^{\beta z_j}$$

$$I_{\alpha,\beta}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \alpha \partial \beta} \log p(X; \alpha, \beta)\right] = e^{\alpha} \sum_{j=1}^n z_j e^{\beta z_j}.$$

We have to invert the information matrix:

$$I^{-1}(\theta) = \frac{e^{-\alpha}}{\left(\sum e^{\beta z_j}\right)\left(\sum z_j^2 e^{\beta z_j}\right) - \left(\sum z_j e^{\beta z_j}\right)^2} \begin{pmatrix} \sum z_j^2 e^{\beta z_j} & -\sum z_j e^{\beta z_j} \\ -\sum z_j e^{\beta z_j} & \sum e^{\beta z_j} \end{pmatrix}$$

$$\mathbf{V}\left(\widehat{\alpha}\right) \geq e^{\alpha} \left\{ (\sum e^{\beta z_j}) - \frac{(\sum z_j e^{\beta z_j})^2}{(\sum z_j^2 e^{\beta z_j})} \right\}$$

$$\mathbf{V}\left(\widehat{\beta}\right) \geq e^{\alpha} \left\{ (\sum z_j^2 e^{\beta z_j}) - \frac{(\sum z_j e^{\beta z_j})^2}{(\sum e^{\beta z_j})} \right\}$$

Note If the information matrix is singular, then the results are correct but useless; they give variances greater than or equal to 0.

(c) Integrals are straightforward :

$$\int_0^1 x^\beta dx = \frac{1}{1+\beta}$$

$$\int_0^1 x^\beta \log x dx = \int_0^1 e^{\beta \log x} \log x dx = \frac{d}{d\beta} \int_0^1 x^\beta dx = -\frac{1}{(1+\beta)^2}$$

$$\int_0^1 x^\beta (\log x)^2 dx = \frac{d^2}{d\beta^2} \int_0^1 x^\beta (\log x)^2 dx = \frac{2}{(1+\beta)^3}$$

In the limit,

$$\frac{1}{n} I(\theta) \to e^\alpha \left(\begin{array}{cc} \frac{1}{1+\beta} & -\frac{1}{(1+\beta)^2} \\ -\frac{1}{(1+\beta)^2} & \frac{2}{(1+\beta)^3} \end{array} \right)$$

$$n I^{-1}(\theta) \to e^{-\alpha} \left(\begin{array}{cc} 2(1+\beta) & (1+\beta)^2 \\ (1+\beta)^2 & (1+\beta)^3 \end{array} \right)$$

Lower bounds: $2(1+\beta)e^{-\alpha}$ and $(1+\beta)^3 e^{-\alpha}$ respectively for α and β.

Chapter 7

Confidence Intervals

7.1 Introduction, Definitions and First Example

Consider a parametric family of probability distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Suppose $\Theta \subseteq \mathbb{R}$. Based on observed data, the aim is to find an interval in which θ can be stated, with confidence, to lie.

Definition 7.1 (Interval Estimator)**.** *Let X be a random variable (or vector) with state space $\mathcal{X}$. An* interval estimator *of a real valued parameter θ is constructed from any two functions $\theta_- : \mathcal{X} \to \mathbb{R}$, $\theta_+ : \mathcal{X} \to \mathbb{R}$ that satisfy $\theta_-(x) \le \theta_+(x)$ for all $x \in \mathcal{X}$. The interval $[\theta_-(X), \theta_+(X)]$ is the* interval estimator.

If $X = x$ is observed, the inference $\theta \in [\theta_-(x), \theta_+(x)]$ is made.

Definition 7.2 (Coverage Probability)**.** *For an interval estimator $[\theta_-(X), \theta_+(X)]$, the* coverage probability *is defined as*

$$\mathbb{P}_\theta\left(\theta_-(X) \le \theta \le \theta_+(X)\right).$$

Note that, with the definitions given so far, the coverage probability can depend on θ. We are going to develop interval estimators whose coverage probability does not depend on θ. We would like to find functions θ_- and θ_+ such that, for a specified value $\alpha < 1$,

$$\mathbb{P}_\theta(\theta_-(X) \le \theta \le \theta_+(X)) = 1 - \alpha \qquad \forall \theta \in \Theta.$$

The following example introduces some general principles for computing interval estimators.

Example 7.1 (Uniform $U(0, \theta)$ distribution)**.**

Let $X_1, \ldots, X_n$ be a random sample from a $U(0, \theta)$ distribution, with unknown parameter θ. That is, the density is

$$p(x, \theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = \max_{j \in \{1,\ldots,n\}} X_j$, $\theta_-(Y) = aY$, $\theta_+(Y) = bY$. Compute the coverage probability for the interval $[\theta_-(Y), \theta_+(Y)]$. For a fixed $\alpha < 1$, compute a and b such that $\mathbb{P}(\theta < \theta_-(Y)) = \frac{\alpha}{2}$ and $\mathbb{P}(\theta > \theta_+(Y)) = \frac{\alpha}{2}$.

Solution Firstly, the density function for Y is

$$p_Y(x,\theta) = \begin{cases} \frac{n}{\theta^n} y^{n-1} & 0 \leq y \leq \theta \\ 0 & \text{other} \end{cases}$$

$$\mathbb{P}_\theta(aY \leq \theta \leq bY) = \mathbb{P}\left(\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right) = \mathbb{P}_\theta\left(\frac{1}{b} \leq T \leq \frac{1}{a}\right)$$

The variable $T = h(Y,\theta) = \frac{Y}{\theta}$ has density

$$p_T(t) = \begin{cases} nt^{n-1} & 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

It follows that, for $1 \leq a \leq b \leq +\infty$,

$$\mathbb{P}_\theta(\theta_-(Y) \leq \theta \leq \theta_+(Y)) = \mathbb{P}_\theta\left(\frac{1}{b} \leq T \leq \frac{1}{a}\right) = \int_{1/b}^{1/a} nt^{n-1}dt = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n.$$

For the other parts,

$$\mathbb{P}_\theta(\theta < \theta_-(Y)) = \int_{1/a}^{1} nt^{n-1}dt = 1 - \left(\frac{1}{a}\right)^n = \frac{\alpha}{2} \Rightarrow a = \frac{1}{(1-(\alpha/2))^{1/n}}$$

$$\mathbb{P}_\theta(\theta > \theta_+(Y)) = \int_{0}^{1/b} nt^{n-1}dt = \left(\frac{1}{b}\right)^n = \frac{\alpha}{2} \Rightarrow b = \frac{1}{(\alpha/2)^{1/n}}.$$

Definition 7.3 (Symmetric Confidence Interval)**.** *A* symmetric confidence interval *with* confidence level $1-\alpha$ *is defined as an interval* $[\theta_-(X), \theta_+(X)]$ *such that* $\mathbb{P}_\theta(\theta < \theta_-(X)) = \mathbb{P}_\theta(\theta > \theta_+(X)) = \frac{\alpha}{2}$. *The quantity* α *is known as the* significance level *and indicates the probability of wrongly excluding* θ *from the interval.*

In the example above, $\left[\frac{1}{(1-(\alpha/2))^{1/n}}Y, \frac{1}{(\alpha/2)^{1/n}}Y\right]$ is a symmetric confidence interval for the parameter θ.

There are several methods of constructing confidence intervals. The method used above was the so-called *pivot method.* A *pivot* is a function of the random variable, which depends on the parameter, but whose distribution does not depend on the parameter. The distribution of $h(Y,\theta)$ did not depend on θ. We now discuss the pivot method; other methods are based on hypothesis testing and will be considered later.

7.2 Pivots

In the example, the random variable $h(X_1,\ldots,X_n;\theta) = \frac{1}{\theta}\max_{j\in\{1,\ldots,n\}} X_j$ has the useful property that its distribution does not depend on the parameter. Such variables can be helpful in constructing confidence intervals.

Definition 7.4 (Pivot)**.** *A random variable* $h(X,\theta)$ *is a* pivotal quantity *or* pivot *if the distribution of* $h(X,\theta)$ *is independent of all parameters.*

Example 7.2 (Gamma Pivot).

Suppose that $X_1, \ldots, X_n$ are i.i.d. $\text{Exp}(\lambda)$ variables. Then $T = \sum_{j=1}^n X_j$ is a sufficient statistic for the parameter λ and $T \sim \text{gamma}(n, \lambda)$. The density function is

$$p_T(x, \lambda) = \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x} \qquad x \geq 0$$

Let $H = h(T, \lambda) = 2\lambda T$, then $H \sim \text{gamma}(n, \frac{1}{2}) = \chi^2_{2n}$; the density is

$$p_H(x) = \frac{1}{2^n \Gamma(n)} x^{n-1} e^{-x/2} \qquad x \geq 0$$

□

Using a Pivot When a pivot is available, the *pivot method* for constructing confidence intervals may be used. For a specified α, find suitable numbers a and b, which do not depend on θ, such that

$$\mathbb{P}_\theta(a \leq h(X, \theta) \leq b) \geq 1 - \alpha$$

The $1 - \alpha$ confidence region is then the set

$$C(X) = \{\theta : a \leq h(X, \theta) \leq b\}$$

Example 7.3 (Random sample from an $\text{Exp}(\lambda)$ distribution, λ unknown).

From a random sample $X_1, \ldots, X_n$ taken from an $\text{Exp}(\lambda)$ distribution, required is a $1 - \alpha$ symmetric confidence interval for λ.

Solution $h(X_1, \ldots, X_n, \lambda) = 2\lambda \sum_{j=1}^n X_j = 2\lambda T \sim \chi^2_{2n}$. Required two constants a and b such that

$$\mathbb{P}(a \leq H) = \frac{\alpha}{2} \qquad \mathbb{P}(b \geq H) = \frac{\alpha}{2} \qquad H \sim \chi^2_{2n}$$

Define $k_{2n,\alpha}$ such that $\mathbb{P}(H \geq k_{2n,\alpha}) = \alpha$, then $a = k_{2n,1-(\alpha/2)}$, $b = k_{2n,(\alpha/2)}$, so that the $1 - \alpha$ symmetric interval is $\lambda \in [\lambda_-(\underline{X}), \lambda_+(\underline{X})]$, where

$$\lambda_-(\underline{X}) = \frac{k_{2n;1-(\alpha/2)}}{2\sum_{j=1}^n X_j}, \qquad \lambda_+(\underline{X}) = \frac{k_{2n;(\alpha/2)}}{2\sum_{j=1}^n X_j}.$$

7.3 An Inverse Method

Suppose Y is a random variable with distribution from a parametric family $\{\mathbb{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$. Suppose that an observation y is made. We would like to be able to establish a confidence interval by finding parameter values θ_1 and θ_2 such that, for specified significance α_1 and α_2, θ_1 and θ_2 solve: $\mathbb{P}_{\theta_1}(Y \leq y) = \alpha_1$ and $\mathbb{P}_{\theta_2}(Y \geq y) = \alpha_2$.

If it turns out that for each t $F_\theta(t) := \mathbb{P}_\theta(Y \leq t)$ is *monotone decreasing* in θ so that large values of Y are more likely for small values of θ, we would like to conclude that a confidence interval with significance $\alpha_1 + \alpha_2$, based on data y is $[\theta_1, \theta_2]$. We would like a similar result, with the roles of θ_1 and θ_2 reversed for *monotone increasing* for each t. The following theorem gives conditions where this holds.

Theorem 7.5. *Let T be a random variable with c.d.f. from the family $F_\theta(t) = \mathbb{P}_\theta(T \leq t)$, $\theta \in \mathbb{R}$. Let $\widetilde{F}_\theta(t) = \mathbb{P}_\theta(T \geq t)$. Suppose that either*

1. *For each t, $F_\theta(t)$ is decreasing as a function of θ. Define $\theta_L(t)$ and $\theta_U(t)$ as:*

$$\widetilde{F}_{\theta_L(t)}(t) = \alpha_1 \qquad F_{\theta_U(t)}(t) = \alpha_2$$

2. *For each t, $F_\theta(t)$ is increasing as a function of θ for each t. Define $\theta_L(t)$ and $\theta_U(t)$ by:*

$$F_{\theta_L(t)}(t) = \alpha_1, \qquad \widetilde{F}_{\theta_U(t)}(t) = \alpha_2.$$

Then $\mathbb{P}_\theta(\theta < \theta_L(T)) \leq \alpha_1$ and $\mathbb{P}_\theta(\theta > \theta_U(T)) \leq \alpha_2$. The inequalities are equalities when $F_\theta^{-1}(F_\theta(x)) = x$ for each θ and each x.

Proof Only part 1. is proved, since part 2. is similar. Firstly, let $\widetilde{F}_\theta(t) = \mathbb{P}_\theta(T \geq t)$. Then $F_\theta(T)$ and $\widetilde{F}_\theta(t)$ satisfy:

$$\begin{cases} \mathbb{P}_\theta\left(F_\theta(T) < x\right) \leq x \\ \mathbb{P}_\theta\left(\widetilde{F}_\theta(T) < x\right) \leq x. \end{cases}$$

(with equality if F_θ is the c.d.f. of a continuous variable with strictly positive density).

From the definition,

$$\{\theta : \widetilde{F}_\theta(T) < \alpha_1\} = \{\theta < \theta_L(T)\}$$

and

$$\{\theta : F_\theta(T) < \alpha_2\} = \{\theta > \theta_U(T)\}.$$

Since $\widetilde{F}_\theta(t)$ is increasing in θ for each t and $F_\theta(t)$ is decreasing in θ for each t,

$$\mathbb{P}_\theta\left(\theta < \theta_L(T)\right) = \mathbb{P}_\theta\left(\widetilde{F}_\theta(T) < F_{\theta_{L(T)}}(T)\right) = \mathbb{P}_\theta\left(\widetilde{F}_\theta(T) < \alpha_1\right) \leq \alpha_1$$

For the other,

$$\mathbb{P}_\theta(\theta > \theta_U(T)) \quad = \quad \mathbb{P}_\theta(F_\theta(T) < F_{\theta_{U(T)}}) = \mathbb{P}_\theta(F_\theta(T) < \alpha_2) \leq \alpha_2.$$

Example 7.4 (Poisson Interval Estimator).

Let $X_1, \ldots, X_n$ be a random sample from a Poiss(μ) population and let $Y = \sum_{j=1}^n X_j$. Then Y is sufficient for μ and $Y \sim$ Poiss($n\mu$). To find a symmetric confidence interval with confidence level $1-\alpha$, the functions $\mu_-(y)$ and $\mu_+(y)$ should solve the two equations:

$$\mathbb{P}_\mu(\mu < \mu_-(Y)) = \frac{\alpha}{2} \qquad \mathbb{P}_\mu(\mu > \mu_+(Y)) = \frac{\alpha}{2}.$$

Note that for $Y \sim Poiss(\mu)$, $\mathbb{P}_\mu(Y \leq y)$ is decreasing in μ for each y and hence the hypotheses of the previous theorem are satisfied. Let $\mathbb{P}_\nu$ denote the Poiss($n\nu$) distribution, then from the above theorem, a $1-\alpha$ confidence interval is given by $[\mu_1(Y), \mu_+(Y)]$ where μ_- and μ_+ solve:

$$\mu_-(y) = \left\{\nu : \mathbb{P}_\nu(Y \leq y) = \frac{\alpha}{2}\right\} \qquad \mu_+(y) = \left\{\nu : \mathbb{P}_\nu(Y \geq y) = \frac{\alpha}{2}\right\}.$$

Then, using the result above, we may conclude that for $\mu < \mu_-(y)$, $\mathbb{P}_\mu(Y \leq y) \leq \frac{\alpha}{2}$ and for $\mu > \mu_+(y)$, $\mathbb{P}_\mu(Y \geq y) \leq \frac{\alpha}{2}$.
These two equations may be solved numerically. Written out explicitly, the equations are:

$$\mu_-(y) = \nu : \sum_{j=0}^{y} \frac{(n\nu)^j}{j!} e^{-n\nu} = \frac{\alpha}{2} \qquad \mu_+(y) = \nu : 1 - \sum_{j=0}^{y-1} \frac{(n\nu)^j}{j!} e^{-n\nu} = \frac{\alpha}{2}.$$

We can obtain a more useful representation in the following way.

Recall that the Poisson and gamma families are linked via the Poisson process; $\mathbb{P}(Y \leq y)$ is the probability of y or fewer events in unit time of a Poisson process with intensity parameter $n\mu$; the $y+1$th event has not yet happened and hence, letting T denote the time of event $y+1$,

$$\mathbb{P}(Y \leq y) = \mathbb{P}(T \geq 1) \qquad T \sim \text{gamma}(y+1, n\mu).$$

If $T \sim \text{gamma}(y+1, n\mu)$ then $2n\mu T \sim \chi^2_{2(y+1)}$, so that, if $W \sim \chi^2_{2(y+1)}$, then

$$\mathbb{P}(Y \leq y) = \mathbb{P}(T > 1) = \mathbb{P}(W \geq 2n\mu) \Rightarrow \mu_-(Y) = \frac{1}{2n} k_{2(Y+1), 1-(\alpha/2)}$$

Similarly, for the upper bound, let $V \sim \chi^2_{2y}$ then, for $T \sim \text{gamma}(y, n\mu)$,

$$\frac{\alpha}{2} = \sum_{k=y}^{\infty} e^{-n\mu} \frac{(n\mu)^k}{k!} = \mathbb{P}(Y \geq y) = \mathbb{P}(T < 1) = \mathbb{P}(V < 2n\mu) \Rightarrow \mu_+(Y) = \frac{1}{2n} k_{2Y, (\alpha/2)}.$$

□

7.4 Pivotal Quantities for the Normal Distribution

If $X \sim N(\mu, \sigma^2)$, then $Z = h(X; \mu, \sigma) = \frac{X-\mu}{\sigma} \sim N(0,1)$ is a pivotal quantity. Based on a $N(\mu, \sigma^2)$ random sample $X_1, \ldots, X_n$,

$$h(X_1, \ldots, X_n; \mu, \sigma) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

This quantity is used to compute confidence intervals for μ if σ is known. Let

$$z_\alpha : \qquad \mathbb{P}(Z \geq z_\alpha) = \alpha \qquad Z \sim N(0,1)$$

Then, for example, an upper bound, with confidence level $1-\alpha$ for μ may be obtained by using

$$\mathbb{P}\left(\frac{\mu - \overline{X}}{\sigma/\sqrt{n}} \leq z_\alpha\right) = 1 - \alpha$$

giving a confidence interval $\mu \in (-\infty, \overline{X} + \frac{\sigma}{\sqrt{n}} z_\alpha]$.

7.4.1 The Chi Squared Distribution Intervals for sigma

The definition of the χ^2 distribution is as follows:

Definition 7.6 (Chi squared Distribution)**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0,1)$ random variables and let $V = Z_1^2 + \ldots + Z_n^2$, then $V \sim \chi_n^2$ where χ_n^2 denotes chi squared distribution with n degrees of freedom.*

Now suppose that $X_1, \ldots, X_n$ is a $N(\mu, \sigma^2)$ random sample, where σ^2 is unknown and μ is known. Then

$$V := g(\mu, \sigma; X_1, \ldots, X_n) := \frac{1}{\sigma^2} \sum_{j=1}^{n} (X_j - \mu)^2 \sim \chi_n^2.$$

This is a pivot variable and can be used for constructing confidence intervals for σ.

Now suppose that μ is *unknown* and that a confidence interval for σ is required. Let

$$W = w(\sigma : X_1, \ldots, X_n) := \frac{1}{\sigma^2} \sum_{j=1}^{n} (X_j - \overline{X})^2$$

Lemma 7.7. *W, thus defined, has χ_{n-1}^2 distribution.*

Proof Let $Z_j = \frac{X_j - \mu}{\sigma}$, then $Z_1, \ldots, Z_n$ is an i.i.d. $N(0,1)$ random sample, so the problem is equivalent to showing that $\sum_{j=1}^n (Z_j - \overline{Z})^2 \sim \chi_{n-1}^2$ where $Z_1, \ldots, Z_n$ is a $N(0,1)$ random sample.
The vector $(Z_1 - \overline{Z}, \ldots, Z_n - \overline{Z})^t$ may be written:

$$\begin{pmatrix} Z_1 - \overline{Z} \\ \vdots \\ Z_n - \overline{Z} \end{pmatrix} = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \ldots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \ldots & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \ldots & -\frac{1}{n} & 1 - \frac{1}{n} \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t)\underline{Z},$$

where I_n is the $n \times n$ identity matrix, $\mathbf{1} = (1, \ldots, 1)^t$ is a column vector, length n, with each entry 1 and $\underline{Z} \sim N(0, I_n)$. The matrix $I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t$ is the matrix with diagonal entries $1 - \frac{1}{n}$ and off-diagonal entries $-\frac{1}{n}$.

It is straightforward to show that

$$(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t)^2 = I_n - \frac{2}{n}\mathbf{1}\mathbf{1}^t + \frac{1}{n^2}\mathbf{1}(\mathbf{1}^t\mathbf{1})\mathbf{1}^t = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t.$$

It follows that $M_n := I_n - \frac{1}{\mathbf{1}}\mathbf{1}^t$ is *idempotent*. It has decomposition $M_n = PDP^t$ where P is orthonormal and D is a diagonal matrix with entries 0 and 1. Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of M_n. Since M_n is symmetric, it follows that

$$\sum_{j=1}^n \lambda_j = \text{trace}(M_n) = n-1.$$

It follows that D has $n-1$ entries which are 1 and one entry which is 0. Let $Y = P^t Z$, then $Y \sim N(0, P^t IP) = N(0, I)$ and

$$\sum_{j=1}^n (Z_j - \overline{Z})^2 = \underline{Z}^t M_n^t M_n \underline{Z} = \underline{Z}^t M_n \underline{Z} = \underline{Z}^t PDP^t \underline{Z} = Y^t DY = \sum_{j=1}^{n-1} Y_j^2.$$

Since $Y_1, \ldots, Y_{n-1}$ are i.i.d. $N(0,1)$, it follows from the definition of the χ^2 distribution that

$$\sum_{j=1}^n (Z_j - \overline{Z})^2 \sim \chi^2_{n-1}.$$

Therefore, for $X_1, \ldots, X_n$ a $N(\mu, \sigma^2)$ random sample, the quantity

$$W := w(\sigma; X_1, \ldots, X_n) := \frac{\sum_{j=1}^n (X_j - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

as required. □

$W = \frac{\sum_{j=1}^n (X_j - \overline{X})^2}{\sigma^2}$ is a pivot variable which can be used for constructing inteval estimators for σ^2.

Confidence interval for σ when μ is unknown Let $X_1, \ldots, X_n$ be a $N(\mu, \sigma^2)$ random sample. Suppose that μ is unknown and that a confidence interval for σ is required. The quantity

$$w(X_1, \ldots, X_n; \sigma) := \frac{\sum_{j=1}^n (X_j - \overline{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

is a pivotal quantity for σ. To find a $1-\alpha$ upper bound for σ, find a such that

$$\mathbb{P}\left(a \leq \frac{(n-1)S^2}{\sigma^2}\right) = \mathbb{P}\left(\frac{(n-1)S^2}{\sigma^2} \geq a\right) = 1-\alpha \Rightarrow a = k_{n-1,1-\alpha}$$

which gives the $1-\alpha$ upper bound

$$\sigma \leq \sqrt{\frac{(n-1)S^2}{k_{n-1,1-\alpha}}}.$$

□

7.4.2 Normal random sample: Interval for mu, unknown sigma

Now consider the problem of finding interval estimators for μ when σ is unknown. The t distribution is needed; let

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

This is an unbiased estimator of σ^2. The quantity

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}$$

where t_{n-1} denotes a t distribution with $n-1$ degrees of freedom.

Definition 7.8 (t distribution)**.** *Let $Z \sim N(0,1)$ and $V \sim \chi^2_m$ be independent random variables. The random variable T is said to have t_m distribution (t distribution with m degrees of freedom) if it has the same distribution as $\frac{Z}{\sqrt{V/m}}$.*

Exercise If $T \sim t_m$, show that its density function is

$$p_T(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\Gamma\left(\frac{m}{2}\right)}\left(1+\frac{x^2}{m}\right)^{-\left(\frac{m+1}{2}\right)} \qquad -\infty < x < +\infty$$

□

Proposition 7.9. *Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population, let $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ denote the sample average and let*

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

Then $\overline{X}$ and S^2 are independent random variables and

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Proof Firstly, to show that $\overline{X}$ and S are independent, recall that if U and V are random vectors such that

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{UU} & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_{VV} \end{pmatrix}\right)$$

and $\Sigma_{U,V} = 0$, then $U \perp V$. This may be seen by considering the moment generating function of $\binom{U}{V}$; if $\Sigma_{UV} = 0$, then

$$M_{U,V}(p,q) = \exp\left\{(p,\mu_1) + \frac{1}{2}p^t\Sigma_{UU}p\right\}\exp\left\{(q,\mu_2) + \frac{1}{2}q^t\Sigma_{VV}q\right\} = M_U(p)M_V(q).$$

Consider $U = \overline{X}$ (vector with a single component) and $V = (X_1 - \overline{X}, \ldots, X_n - \overline{X})^t$, then

$$\mathbf{C}(\overline{X}, X_j - \overline{X}) = \mathbf{C}(\overline{X}, X_j) - \mathbf{V}(\overline{X}) = \frac{1}{n}\sigma^2 - \frac{1}{n}\sigma^2 = 0.$$

It follows that $\overline{X} \perp (X_1 - \overline{X}, \ldots, X_n - \overline{X})$ and hence that

$$\overline{X} \perp \sum_{j=1}^{n} (X_j - \overline{X})^2.$$

The fact that

$$\sum_{j=1}^{n} \frac{(X_j - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

has already been established. From this, it follows directly that:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{j=1}^{n}(X_j - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

and hence that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \frac{1}{\sqrt{S^2/\sigma^2}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

as required. □

Confidence Intervals for μ when σ is unknown To construct a confidence interval for μ from a $N(\mu, \sigma^2)$ sample $X_1, \ldots, X_n$ when σ is not known, the pivot

$$T = h(X_1, \ldots, X_n; \mu) = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

may be used. Let

$$t_{m,\alpha} : \quad \mathbb{P}(T \geq t_{m,\alpha}) = \alpha \qquad T \sim t_m$$

and note that, by symmetry,

$$\mathbb{P}(T \leq -t_{m,\alpha}) = \alpha \Rightarrow t_{m,1-\alpha} = -t_{m,\alpha}.$$

then a $1 - \alpha$ symmetric confidence interval is given by

$$\mu \in \left[\overline{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right]$$

□

Summary

- Interval estimator, Coverage Probability.
- Pivot Variables (Gamma pivot, chi squared)
- theorem: c.d.f. monotone in θ, construction of confidence intervals.
- Example: Poisson Interval Estimator
- Pivotal Quantities for normal distribution: $N(0,1)$, t distribution, χ^2 distribution.

Tutorial 8

1. Let X be a single observation from a distribution with density

$$p_X(x) = \theta x^{\theta-1} \qquad 0 \leq x \leq 1.$$

   (a) Let $Y = -\frac{1}{\log X}$. Evaluate the confidence level of the interval estimator $\left[\frac{Y}{2}, Y\right]$.

   (b) Find a pivotal quantity and use it to set up an interval estimator.

2. Let $X_1, \ldots, X_n$ be a sample from a $N(\mu, \sigma^2)$ population, both μ and σ^2 unknown.

   (a) Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0,1)$ variables. The distribution of $W = Z_1^2 + \ldots + Z_n^2$ is χ^2_n. Let $\underline{Z} = (Z_1, \ldots, Z_n)^t$ where t denotes transpose. Let $Y_j = Z_j - \overline{Z}$ and let $\underline{Y} = (Y_1, \ldots, Y_n)^t$. Let $\underline{Y} = M_n \underline{Z}$ for a symmetric matrix M_n. What is M_n? Prove that $M_n^2 = M_n$. From this, what do you conclude about the eigenvalues of M_n? Use the fact that the sum of the eigenvalues is equal to the sum of the trace for symmetric matrices.

   Now consider the expression $M_n = PDP^t$ where D is diagonal and P is orthonormal. What is the distribution of $P^t \underline{Z}$? Hence what is the distribution of $\underline{Z}^t M_n^t M_n \underline{Z}$?
   Hence conclude that

$$\frac{\sum_{j=1}(X_j - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

   (b) Show that

$$\overline{X} \perp X_1 - \overline{X}, \ldots, X_n - \overline{X}.$$

3. Let $Z \sim N(0,1)$ and let $V \sim \chi^2_m$. Let $Z \perp V$. Let $T = \frac{Z}{\sqrt{V/m}}$. Compute the density function for T.

4. Let $X_1, \ldots, X_n$ be a random sample from a $N(1, \sigma^2)$ population ($\mu = 1$ is known). Construct a symmetric $1 - \alpha$ interval estimator for σ, with as many degrees of freedom as possible.

5. Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Using a pivot based on $\sum_{j=1}^n (X_j - \overline{X})^2$, construct a symmetric confidence interval with confidence level $1 - \alpha$ for $\log \sigma^2$.

6. Suppose that $Y_1, \ldots, Y_n$ are independent and that

$$Y_i \sim N(x_i \beta, \sigma^2),$$

   where $x_1, \ldots, x_n$ are given, σ is known and β is an unknown parameter.

   (a) Compute the least squares estimator $\widehat{\beta}_{LS}$ of β.

   (b) Compute a confidence interval for β of the form $[\widehat{\beta}_{LS} - c, \widehat{\beta}_{LS} + c]$ with confidence level $1 - \alpha = 0.95$.

7. Let
$$X_i = \frac{\theta}{2} t_i^2 + \epsilon_i \qquad i = 1, \ldots, n$$
where ϵ_i are i.i.d. $N(0, \sigma^2)$ variables, where σ is known.

   (a) Compute the MLE of θ.

   (b) Using a pivot based on the MLE of θ, find a symmetric confidence interval for θ with confidence level $1 - \alpha$.

   (c) Suppose the values for t_i may be chosen freely subject to the constraint that $0 \leq t_i \leq 1$ for each $i = 1, \ldots, n$. What values of t_i should be chosen to make the symmetric confidence interval as short as possible?

8. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma_1^2)$. Suppose a lower confidence bound $\overline{X} - c$, intended to be of confidence level $1 - \alpha$ is computed under the assumption that $X_j \sim N(\mu, \sigma_0^2)$. What is the actual confidence level?

9. (a) Let $X_1, \ldots, X_n$ be a $N(\mu, \sigma^2)$ sample, where μ and σ^2 are both unknown. Show that the symmetric $1 - \alpha$ confidence interval is given by
$$\left[\overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1;\alpha/2}\right]$$
where $S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2$ and $t_{n,\alpha}$ denotes the number such that $\mathbb{P}(T > t_{n,\alpha}) = \alpha$ where $T \sim t_n$.

   (b) Suppose we want to select a sample size N such that the interval in part (a) has length at most $l = 2d$ for some preassigned length. Stein's two stage procedure (1945) is the following: Begin by taking a fixed number $n_0 \geq 2$ of observations, calculate $\overline{X}_0 = \frac{1}{n_0} \sum_{j=1}^{n_0} X_j$ and $S_0^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (X_j - \overline{X}_0)^2$. Then take $N - n_0$ further observations where N is the smallest integer greater than or equal to n_0 and greater than or equal to $(S_0 t_{n_0-1;(\alpha/2)}/d)^2$.
   Show that
$$\frac{\sqrt{N}(\overline{X} - \mu)}{S_0} \sim t_{n_0 - 1}$$
where $S_0^2 = \frac{1}{n_0-1} \sum_{j=1}^{n_0} (X_j - \overline{X}_0)^2$ and $\overline{X} = \frac{1}{N} \sum_{j=1}^{N} X_j$. It follows that $\left[\overline{X} \pm \frac{S_0}{\sqrt{N}} t_{n_0-1;\alpha/2}\right]$ is a confidence interval with confidence level $1 - \alpha$ for μ of length at most $2d$.

   Hint recall that $\overline{X} \perp S^2$ where $\overline{X}$ is the estimator of μ and S^2 is the estimator of σ^2 based on a sample size n from a $N(\mu, \sigma^2)$ distribution. Consider the definition of a t distribution.

10. Let $X_1, \ldots, X_n$ be a random sample from a Rayleigh distribution
$$p(x, \sigma) = \begin{cases} \frac{x}{\sigma^2} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} & x \geq 0 \\ 0 & x < 0 \end{cases}$$
where $\sigma > 0$ is an unknown parameter.

(a) Compute the maximum likelihood estimator of σ^2.

(b) Compute a symmetric confidence interval, confidence level $1-\alpha$ of the form

$$\left[c_n \widehat{\sigma^2}_{ML}, d_n \widehat{\sigma^2}_{ML}\right]$$

for σ^2. Express your answer in terms of quantiles of an appropriate χ^2 distribution.

11. Let $X_1, \ldots, X_n$ be a $N(\mu, \sigma^2)$ random sample where σ is known. Show that the interval of shortest length of confidence $1-\alpha$ of the form

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha_1}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha_2}\right] \qquad \alpha_1 + \alpha_2 = \alpha$$

is obtained by taking $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.

Answers

1. (a) Coverage probability is:

$$\begin{aligned}\mathbb{P}_\theta\left(\frac{Y}{2} \leq \theta \leq Y\right) &= \mathbb{P}\left(\theta \leq Y \leq 2\theta\right) = \mathbb{P}(e^{-1/\theta} \leq X \leq e^{-1/(2\theta)}) \\ &= \theta \int_{e^{-1/\theta}}^{e^{-1/(2\theta)}} x^{\theta-1}dx = e^{-1/2} - e^{-1}.\end{aligned}$$

(b) X has c.d.f.

$$\mathbb{P}_\theta(X \leq x) = \begin{cases} 0 & x < 0 \\ x^\theta & 0 \leq x \leq 1 \\ 1 & x \geq 1. \end{cases}$$

Let $Z = -\theta \log X$ then:

$$\mathbb{P}_\theta(Z \leq z) = \mathbb{P}_\theta(-\theta \log X \leq z) = \mathbb{P}(X \geq \exp\{-\frac{z}{\theta}\}) = \begin{cases} 1 - e^{-z} & z \geq 0 \\ 0 & z < 0. \end{cases}$$

Hence $Z = -\theta \log X \sim \text{Exp}(1)$ is a natural pivot variable (its distribution does not depend on θ). Then

$$\mathbb{P}(Z \leq z_1) = \alpha_1 \Rightarrow 1 - e^{-z_1} = \alpha_1 \Rightarrow z_1 = \log\frac{1}{1-\alpha_1}$$
$$\mathbb{P}(Z \geq z_2) = \alpha_2 \Rightarrow e^{-z_2} = \alpha_2 \Rightarrow z_2 = \log\frac{1}{\alpha_2}$$

so that

$$1 - (\alpha_1 + \alpha_2) = \mathbb{P}_\theta\left(\log\frac{1}{1-\alpha_1} \leq -\theta \log X \leq \log\frac{1}{\alpha_2}\right)$$

giving

$$\mathbb{P}_\theta\left(\frac{\log\frac{1}{1-\alpha_1}}{-\log X} \leq \theta \leq \frac{\log\frac{1}{\alpha_2}}{-\log X}\right) = 1 - (\alpha_1 + \alpha_2)$$

An interval estimator with coverage probability $1 - (\alpha_1 + \alpha_2)$ is:

$$\left[\frac{\log(1-\alpha_1)}{\log X}, \frac{\log \alpha_2}{\log X}\right].$$

For a symmetric interval with coverage probabiltiy $1 - \alpha$, take $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.

Alternatively, from the c.d.f. for X, another pivot is: $W = X^\theta$ which has $U(0,1)$ distribution (which does not depend on θ).

Using $X^\theta \sim U(0,1)$, for any $0 \leq a < b \leq 1$,

$$\mathbb{P}(a \leq X^\theta \leq b) = b - a$$

$$b - a = \mathbb{P}(\log a \leq \theta \log X \leq \log b) = \mathbb{P}\left(\frac{\log b}{\log X} \leq \theta \leq \frac{\log a}{\log X}\right)$$

an interval estimator with confidence level $b - a$ is

$$\left[\frac{\log b}{\log X}, \frac{\log a}{\log X}\right].$$

If we want a *symmetric* confidence interval with confidence level $1 - \alpha$, then $a = \frac{\alpha}{2}$ and $b = 1 - \frac{\alpha}{2}$; interval is:

$$\left[\frac{\log(1 - \alpha/2)}{\log X}, \frac{\log(\alpha/2)}{\log X}\right].$$

2. (a) Let I_n denote the $n \times n$ identity matrix and let $\mathbf{1}_n$ denote a column vector, length n, with each entry 1. Then $\mathbf{1}_n\mathbf{1}_n^t$ is the $n \times n$ matrix with each ehtry 1. Then

$$M_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t.$$

It follows that

$$M_n^2 = (I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)^2 = I_n - \frac{2}{n}\mathbf{1}_n\mathbf{1}_n^1 t + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^t\mathbf{1}_n\mathbf{1}_n^t = M_n$$

using the fact that $\mathbf{1}_n^t\mathbf{1}_n = n$.

M_n is symmetric, hence has decomposition $M_n = PDP^t$ where D is diagonal and P is orthonormal. So $M_n^2 = M_n$ implies

$$M_n^2 = PDP^tPDP^t = PD^2P^t = PDP^t = M_n \Rightarrow D^2 = D$$

hence all the eigenvalues are 0 or 1. Using the fact that the sum of eigenvalues is equal to the trace, the sum of eigenvalues is $n - 1$. Hence 1 has multiplicity $n - 1$ and 0 has multiplicity 1.

Use: If $\underline{X} \sim N(\underline{\mu}, \Sigma)$ then $A\underline{X} + \underline{b} \sim N(A\underline{\mu} + \underline{b}, A\Sigma A^t)$. Then

$$\underline{Y} := P^t\underline{Z} \sim N(P^t\underline{0}, \sigma^2 P^t I P) = N(0, \sigma^2 I)$$

since P is orthonormal (so that $P^tP = I$).

Take the columns of P such that $D = \text{diag}(1, \ldots, 1, 0)$ (the 0 in the nth position). It follows that

$$\underline{Z}^t M_n^t M_n \underline{Z} = \underline{Y}^t D \underline{Y} = \sum_{j=1}^{n-1} Y_j^2 \sim \chi^2_{n-1}.$$

(b) Let $\Sigma^{(X)} = \sigma^2 I$ denote the covariance matrix of $\underline{X}$. Clearly, $\underline{Y} := (X_1 - \overline{X}, \ldots, X_n - \overline{X}, \overline{X})^t = A\underline{X}$ is a normal random vector, since it is a linear transformation of $\underline{X}$. Let $\Sigma^{(Y)}$ denote the covariance matrix of Y. Here $A_{ij} = -\frac{1}{n}$ for $j \neq i$, $i = 1, \ldots, n$, $A_{ii} = 1 - \frac{1}{n}$, $i = 1, \ldots, n$, $A_{n+1,j} = \frac{1}{n}$, $j = 1, \ldots, n$. Use:

$$\Sigma^{(Y)} = A\Sigma^{(X)}A^t = \sigma^2 AA^t$$

For $j \neq n+1$, $(AA^t)_{n+1,j} = \sum_{k=1}^n A_{n+1,k}A_{j,k} = \frac{1}{n}\sum_{k=1}^n A_{j,k} = 0$ hence $\mathbf{C}(\overline{X}, X_j - \overline{X}) = 0$.

3. Density of χ^2_m is:

$$p_V(v) = \frac{1}{2^{m/2}\Gamma(m/2)} v^{(m/2)-1}e^{-v/2}\mathbf{1}_{[0,+\infty)}(v)$$

so

$$\mathbb{P}(T \leq t) = \mathbb{P}(Z \leq t\sqrt{V/m}) = \frac{1}{2^{m/2}\Gamma(m/2)}\int_0^\infty dv \left(\int_{-\infty}^{t\sqrt{v/m}} \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\right) v^{(m/2)-1}e^{-v/2}$$

Density of T is

$$p_T(t) = \frac{d}{dt}\mathbb{P}(T \leq t) = \frac{1}{m^{1/2}\pi^{1/2}2^{(m+1)/2}\Gamma(m/2)}\int_0^\infty dv \left(v^{(m-1)/2}e^{-(1+(t^2/m))(v/2)}\right)$$

Now make the substitution $z = (1+\frac{t^2}{m})(v/2)$ to get:

$$p_T(t) = \frac{1}{m^{1/2}\pi^{1/2}\Gamma(m/2)}\frac{1}{(1+t^2/m)^{(m+1)/2}}\int_0^\infty z^{(m-1)/2}e^{-z}dz = \frac{\Gamma(\frac{m+1}{2})}{m^{1/2}\pi^{1/2}\Gamma(\frac{m}{2})(1+t^2/m)^{(m+1)/2}}$$

4. Here $\frac{\sum_{j=1}^n (X_j-1)^2}{\sigma^2} \sim \chi^2_n$. The interval estimator is

$$\left[\sqrt{\frac{\sum_{j=1}^n (X_j-1)^2}{k_{n,(\alpha/2)}}}, \sqrt{\frac{\sum_{j=1}^n (X_j-1)^2}{k_{n,1-(\alpha/2)}}}\right]$$

where $k_{n,\alpha}$ is the value such that $\mathbb{P}(V > k_{n,\alpha}) = \alpha$, $V \sim \chi^2_n$.

5.

$$\frac{\sum_{j=1}^n (X_j - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

and this is the pivot variable. Using tail probabilities $\mathbb{P}(V \geq k_{n,\alpha}) = \alpha$ for $V \sim \chi^2_n$,

$$\mathbb{P}\left(k_{n-1,1-(\alpha/2)} \leq \frac{\sum_{j=1}^n (X_j - \overline{X})^2}{\sigma^2} \leq k_{n-1,(\alpha/2)}\right) = 1-\alpha$$

$$\mathbb{P}\left(\log\sum_{j=1}^n (X_j - \overline{X})^2 - \log k_{n-1,(\alpha/2)} \leq \log\sigma^2 \leq \log\sum_{j=1}^n (X_j - \overline{X})^2 - \log k_{n-1,1-(\alpha/2)}\right) = \alpha$$

A symmetric interval for σ with coverage probability $1-\alpha$ is therefore:

$$\sigma \in \frac{1}{2}\left[\log\sum_{j=1}^n (X_j - \overline{X})^2 - \log k_{n-1,(\alpha/2)}, \log\sum_{j=1}^n (X_j - \overline{X})^2 - \log k_{n-1,1-(\alpha/2)}\right].$$

6. (a) Least squares estimator minimises $\frac{1}{\sigma^2}\sum_{j=1}^n (Y_i - x_i\beta)^2$ since $\mathbb{E}[Y_i] = x_i\beta$ and $\operatorname{Var}(Y_i) = \sigma^2$ for each i. It follows that

$$\widehat{\beta}_{LS} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

(b) Since $\mathbb{E}\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{\sum_{i=1}^n x_i^2 \beta}{\sum_{i=1}^n x_i^2} = \beta$ and

$$\operatorname{Var}\left(\widehat{\beta}_{LS}\right) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

$$\widehat{\beta}_{LS} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$$

giving a $1-\alpha$ confidence interval of

$$\beta \in \left[\widehat{\beta}_{LS} \pm \frac{\sigma}{\sqrt{\sum_{i=1}^n x_i^2}} z_{\alpha/2}\right]$$

where z_α denotes the value such that $\mathbb{P}(Z \geq z_\alpha) = \alpha$ for $Z \sim N(0,1)$. Here $\frac{\alpha}{2} = 0.025$ and $z_{0.025} = 1.96$ so interval is:

$$\beta \in \left[\widehat{\beta}_{LS} \pm 1.96 \frac{\sigma}{\sqrt{\sum_{i=1}^n x_i^2}} z_{\alpha/2}\right]$$

7. (a) $X_i \sim N(\frac{\theta}{2}t_i^2, \sigma^2)$.

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^n \left(x_j - \frac{\theta}{2}t_j^2\right)^2\right\}$$

Maximum

$$\sum_{j=1}^n \frac{t_j^2}{2}\left(x_j - \frac{\theta}{2}t_j^2\right) = 0 \Rightarrow \widehat{\theta}_{ML} = \frac{2\sum_{j=1}^n t_j^2 X_j}{\sum_{j=1}^n t_j^4}$$

(b)

$$\hat{\theta}_{ML} \sim N\left(\theta, \frac{4\sigma^2}{\sum_{j=1}^n t_j^4}\right).$$

$$Z = \frac{2\sum_{j=1}^n t_j^2 X_j - \theta\sum_{j=1}^n t_j^4}{2\sigma\sqrt{\sum_{j=1}^n t_j^4}} \sim N(0,1)$$

$$I = \left[\frac{2\sum_{j=1}^n t_j^2 X_j}{\sum_{j=1}^n t_j^4} \pm \frac{2\sigma}{\sqrt{\sum_{j=1}^n t_j^4}} z_{\alpha/2}\right]$$

(c) $t_i = 1$ $i = 1, \ldots, n$.

8. $c = \frac{\sigma}{\sqrt{n}} z_\alpha$. Let $1-\gamma$ denote the actual confidence level, σ_0 the assumed value of σ and σ_1 the true value. Then

$$\gamma = \mathbb{P}\left(\mu < \overline{X} - \frac{\sigma_0}{\sqrt{n}} z_\alpha\right) = \mathbb{P}\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma_1} > \frac{\sigma_0 z_\alpha}{\sigma_1}\right) = 1 - \Phi\left(\frac{\sigma_0 z_\alpha}{\sigma_1}\right)$$

9. (a) Follows directly from $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

   (b) N depends only on $(X_1, \ldots, X_{n_0})$ and does so only through S_0. S_0 is independent of $\overline{X}_0 := \frac{1}{n_0}\sum_{j=1}^{n_0} X_j$ and also of $(X_{n_0+j})_{j\geq 1}$.

$$\overline{X} = \frac{n_0 \overline{X}_0 + \sum_{j=1}^{N-n_0} X_{n_0+j}}{N}$$

For $n \geq n_0 + 1$, let

$$Z_n = \frac{\sqrt{n}((\frac{1}{n}\sum_{j=1}^n X_j) - \mu)}{\sigma}.$$

Then for any borel set $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}\left(\frac{\sqrt{N}(\overline{X}-\mu)}{\sigma} \in A \mid N = n, S_0\right) = \mathbb{P}\left(Z_n \in A | N = n, S_0\right) = \mathbb{P}(Z_n \in A).$$

The conditioning may removed, because Z_n depends on $X_1, \ldots, X_{n_0}$ only through $\overline{X}_0$, hence is independent of S_0 hence is independent of N. From this, it follows that

$$\frac{\sqrt{N}(\overline{X}-\mu)}{\sigma} \sim N(0,1), \qquad \frac{\sqrt{N}(\overline{X}-\mu)}{\sigma} \perp S_0.$$

Since $\frac{(n_0-1)S_0^2}{\sigma^2} \sim \chi^2_{n_0-1}$ and independent of $\frac{\sqrt{N}(\overline{X}-\mu)}{\sigma}$, it follows that

$$\frac{\sqrt{N}(\overline{X}-\mu)}{S_0} \sim t_{n_0-1}.$$

10. (a)

$$\log L(\sigma; x_1, \ldots, x_n) = -2n \log \sigma + \sum_{j=1}^n \log x_j - \frac{1}{2\sigma^2}\sum_{j=1}^n x_j^2$$

Take derivative with respect to σ and set to 0 for ML (convexity gives that the result is the MLE)

$$-\frac{2n}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^n x_j^2 = 0 \Rightarrow \widehat{\sigma^2}_{ML} = \frac{1}{2n}\sum_{j=1}^n X_j^2$$

If there is at least one $x_i > 0$, then log likelihood function is strictly concave, $\to -\infty$ for $\sigma \to 0$ and $\sigma \to +\infty$ hence maximum. If all $x_i = 0$, then likelihood maximised at $\sigma = 0$ as required.

(b) Distribution of $\widehat{\sigma^2}_{ML}$: let $Y = \frac{X^2}{\sigma^2}$ then

$$\begin{aligned}\mathbb{P}\left(Y > y\right) &= \mathbb{P}\left(\frac{X^2}{\sigma^2} > y\right) = \mathbb{P}\left(X > \sigma\sqrt{y}\right) \\ &= \int_{\sigma\sqrt{y}}^{\infty} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} dx = \int_y^\infty \frac{1}{2} e^{-z/2} dz = \exp\{-\frac{1}{2}y\} \qquad Y \sim \text{Exp}(\frac{1}{2})\end{aligned}$$

so $\frac{X^2}{\sigma^2} \sim \text{Exp}(\frac{1}{2})$ and $\frac{1}{\sigma^2}\sum_{j=1}^n X_j^2 \sim \Gamma(n, \frac{1}{2}) = \chi^2_{2n}$. Let $W \sim \chi^2_{2n}$, then $\widehat{\sigma^2}_{ML} \overset{(d)}{=} \frac{\sigma^2}{2n} W$ so that

$$\frac{\alpha}{2} = \mathbb{P}(\sigma^2 < c_n \sigma^2 \frac{W}{2n}) = \mathbb{P}(W > \frac{2n}{c_n}) \Rightarrow \frac{2n}{c_n} = k_{2n,(\alpha/2)} \Rightarrow c_n = \frac{2n}{k_{2n,(\alpha/2)}}.$$

$$d_n = \frac{2n}{k_{2n,1-(\alpha/2)}}.$$

$k_{2n,\alpha}$ denotes the value such that $\mathbb{P}(W > k_{2n,\alpha}) = \alpha$ for $W \sim \chi^2_{2n}$.

11. length is $\frac{\sigma}{\sqrt{n}}(z_{\alpha_2} + z_{\alpha_1})$ so the aim is to find $\alpha_1 \in [0, \alpha]$ that maximises

$$z_{\alpha_1} + z_{\alpha - \alpha_1}.$$

$$\alpha = \int_{z_\alpha}^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \Rightarrow 1 = -\frac{dz_\alpha}{d\alpha}\frac{1}{\sqrt{2\pi}} e^{-z_\alpha^2/2} \Rightarrow \frac{dz_\alpha}{d\alpha} = \sqrt{2\pi} e^{z_\alpha^2/2}$$

$$\frac{d}{d\alpha_1}(z_{\alpha-\alpha_1} + z_{\alpha_1}) = \sqrt{2\pi}\left(e^{z_{\alpha_1}^2/2} - e^{z_{\alpha-\alpha_1}^2/2}\right) = 0 \Rightarrow z_{\alpha_1} = z_{\alpha-\alpha_1} \Rightarrow \alpha_1 = \alpha - \alpha_1 = \frac{\alpha}{2}.$$

Chapter 8

Hypothesis Testing (I)

8.1 Introduction

In statistics, a *hypothesis* is a statement about a parameter. Consider a statistical model $\{\mathbb{P}_\theta | \theta \in \Theta\}$. The goal of hypothesis testing is to decide, based on a sample from a population, which of two complementary hypotheses is true. These are called the *null hypothesis* and *alternative hypothesis* and are denoted H_0 and H_1 respectively. The null hypothesis is $H_0 : \theta \in \Theta_0$ and the alternative $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 \subset \Theta$ is a strict subset of the parameter space.

Definition 8.1. *A* hypothesis testing procedure *or* hypothesis test *is a rule that specifies*

1. *For which sample values the hypothesis H_0 is not rejected.*
2. *For which sample values H_0 is rejected and H_1 is accepted as true.*

The subset of $\mathcal{X}$ (the sample space) for which H_0 is rejected is the rejection region *or* critical region.

Typically, for a random sample $X_1, \dots, X_n$, a hypothesis test is specified in terms of a *test statistic* $W(X_1, \dots, X_n) = W(\underline{X})$, a function of the sample.

8.2 Methods of Finding Tests

8.2.1 Likelihood Ratio Test

Definition 8.2 (Likelihood Ratio Test Statistic)**.** *Let $L(\theta; x)$ denote the likelihood function for parameter θ. The* likelihood ratio test statistic *for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is*

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}.$$

A likelihood ratio test *(LRT) is any test that has a rejection region of the form*

$$\mathcal{R}_{crit} = \{x | \lambda(x) < c\} \qquad 0 \le c \le 1$$

Example 8.1 (Normal LRT)**.**

Let $X_1, \dots, X_n$ be a random sample from a $N(\theta, 1)$ population. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. The LRT is

$$\lambda(\underline{x}) = \frac{L(\theta_0|\underline{x})}{L(\overline{x}|\underline{x})}$$

since the denominator is maximised (by definition) at $\widehat{\theta}_{ML} = \overline{X}$. The LRT statistic is

$$\lambda(\underline{x}) = \exp\left\{-\frac{1}{2}\left(-\sum_{i=1}^{n}(x_i - \theta_0) + \sum_{i=1}^{n}(x_i - \overline{x})^2\right)\right\} = \exp\left\{-\frac{n}{2}(\overline{x} - \theta_0)^2\right\}.$$

An LRT is a test that rejects H_0 for small values of $\lambda(\underline{x})$. The rejection region $\{\underline{x}|\lambda(\underline{x}) \leq c\}$ can be written as:

$$\left\{\underline{x} : |\overline{x} - \theta_0| \geq \sqrt{-\frac{2}{n}\log c}\right\}.$$

Clearly, the rejection region obtained from the LRT statistic has a complete description in terms of the simpler statistic $|\overline{X} - \theta_0|$.

Example 8.2 (Exponential LRT)**.**

Let $X_1, \dots, X_n$ be a random sample from an exponential population with p.d.f.

$$p(x;\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

where $\Theta = \mathbb{R}$. The likelihood function is

$$L(\theta;\underline{x}) = \begin{cases} e^{-\sum_{j=1}^{n} x_j + n\theta} & \theta \leq x_{(1)} \\ 0 & \theta > x_{(1)}. \end{cases}$$

Here $x_{(1)} = \min_j x_j$ is the lowest order statistic.

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Clearly $L(\theta;\underline{x})$ is an increasing function of θ on $-\infty < \theta < x_{(1)}$. Thus, the LRT statistic is

$$\lambda(\underline{x}) = \begin{cases} 1 & x_{(1)} \leq \theta_0 \\ e^{-n(x_{(1)}-\theta_0)} & x_{(1)} > \theta_0 \end{cases}$$

An LRT rejects H_0 if $\lambda(\underline{X}) \leq c$ for some specified c. The rejection region is therefore

$$\left\{\underline{x} : x_{(1)} \geq \theta_0 - \frac{\log c}{n}\right\}.$$

It depends on the sample only through the sufficient statistic $X_{(1)}$.

Theorem 8.3. *If $T(\underline{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\underline{x})$ are the LRT statistics based on T and $\underline{X}$ respectively, then $\lambda^*(T(\underline{x})) = \lambda(\underline{x})$ for every $\underline{x} \in \mathcal{X}$.*

Proof From the factorisation theorem, the probability mass (or density) function of $\underline{X}$ may be written as

$$p(\underline{x}, \theta) = g(T(\underline{x}), \theta)h(\underline{x})$$

where $g(t, \theta)$ is the pmf / pdf of T and $h(\underline{x})$ does not depend on θ. It follows that

$$\lambda(\underline{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\underline{x})}{\sup_{\theta \in \Theta} L(\theta|\underline{x})} = \frac{\sup_{\theta \in \theta_0} g(T(\underline{x})|\theta)}{\sup_{\theta \in \Theta} g(T(\underline{x})|\theta)} = \frac{\sup_{\theta \in \Theta_0} L^*(\theta; T(\underline{x}))}{\sup_{\theta \in \Theta} L^*(\theta; T(\underline{x}))} = \lambda^*(T(\underline{x})).$$

□

Likelihood ratio tests are also useful in situations where there are *nuisance parameters*, that is, parameters that are present in the model, but not of direct interest. Their presence can lead to different tests.

Example 8.3 (Normal LRT with unknown variance).

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population, where it is required to test $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ for a specified μ_0 and where σ^2 is unknown. The LRT statistic is

$$\lambda(\underline{x}) = \frac{\sup_{(\mu, \sigma^2):\mu \leq \mu_0} L(\mu, \sigma^2; \underline{x})}{L(\widehat{\mu}_{ML}, \widehat{\sigma^2}_{ML}; \underline{x})}$$

A test based on $\lambda(\underline{x})$ is equivalent to a test based on the t-statistic

$$T(\underline{X}) = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S} \sim t_{n-1} \qquad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

This is seen as follows: if $\mu_0 > \overline{x}$, then $\lambda(\underline{x}) = 1$, since the maximising pair for the numerator $(\tilde{\mu}, \tilde{\sigma})$ is the same as that for the denominator. If $\mu_0 < \overline{x}$, then

$$L(\mu, \sigma; \underline{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2\right\}$$

Now note that $\sum_{j=1}^{n}(x_j - \mu)^2 = \sum_{j=1}^{n} x_j^2 - 2n\overline{x}\mu + n\mu^2$, and that $\mu^2 - 2\mu\overline{x}$ is decreasing in μ for $\mu < \overline{x}$ so that the maximiser for the numerator is μ_0. Then the $\tilde{\sigma}$ that maximises the numerator satisfies:

$$\tilde{\sigma}^2 := \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu_0)^2.$$

It follows that for $\mu_0 < \overline{x}$,

$$\lambda(\underline{x}) = \exp\left\{\frac{n}{2}\log\frac{\widehat{\sigma}^2_{ML}}{\tilde{\sigma}^2}\right\}$$

so that

$$\lambda(\underline{x})^{2/n} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \mu_0)^2} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2 + n(\overline{x} - \mu_0)^2} = \frac{1}{1 + \frac{n(\overline{x} - \mu_0)^2}{(n-1)s^2}}$$

where $s^2 = \frac{1}{n-1}\sum_{j=1}^n (x_j - \overline{x})^2$.
It follows that

$$\lambda(\underline{x}) \leq c \Leftrightarrow T(\underline{x})^2 := \frac{n(\overline{x}-\mu_0)^2}{s^2} > (n-1)\left(c^{-2/n}-1\right) = k$$

where the relation between k and c is given by the formula. Since the test is only rejected for $\mu_0 < \overline{x}$, it follows that:

$$\mathcal{R}_{\text{crit}} = \{\underline{x} : T(\underline{x}) < k\}.$$

8.3 Evaluating Tests

There are two possible errors when testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$.

- **Type I error**: This is when $\theta \in \Theta_0$, but the hypothesis test incorrectly rejects H_0.
- **Type II error**: This is when $\theta \in \Theta_0^c$, but the hypothesis test fails to reject H_0.

Suppose $\mathcal{R}$ denotes the rejection region for a test. Then, for $\theta \in \Theta_0$,

$$\mathbb{P}(\text{type I error}) = \mathbb{P}_\theta(\underline{X} \in \mathcal{R}) \qquad \theta \in \Theta_0$$

and for $\theta \in \Theta_0^c$,

$$\mathbb{P}(\text{type II error}) = \mathbb{P}_\theta(\underline{X} \in \mathcal{R}^c) \qquad \theta \in \Theta_0^c.$$

The rejection region $\mathcal{R}$ is chosen so that the probability of type I error is not greater than a value specified in advance, α, known as the *significance level.* For a specified significance, the *power* of the test is defined as the probability that it will reject H_0 when H_0 is false.

Definition 8.4 (Power Function)**.** *The* power function *of a hypothesis test with rejection region $\mathcal{R}$ is the function $\beta : \Theta \to [0, 1]$ defined by*

$$\beta(\theta) = \mathbb{P}_\theta(\underline{X} \in \mathcal{R}).$$

Definition 8.5 (Size α, level α test)**.** *For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a* size α test *if*

$$\sup_{\theta\in\Theta_0} \beta(\theta) = \alpha$$

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a level α test *if*

$$\sup_{\theta\in\Theta_0} \beta(\theta) \leq \alpha.$$

Example 8.4 (Normal Power Function)**.**

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, where σ^2 is known. An LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects H_0 if

$$\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > c \qquad c \in \mathbb{R}_+$$

The power function of the test is

$$\beta(\theta) = \mathbb{P}_\theta\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) = \mathbb{P}_\theta\left(\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(c - \frac{\theta - \theta_0}{\sigma/\sqrt{n}}\right)$$

where $\Phi(x) = \mathbb{P}(Z \leq x)$, $Z \sim N(0, 1)$.

Suppose a significance level $\alpha = 0.1$ is required and, in addition, a power of 0.8 is required if $\theta \geq \theta_0 + \sigma$ (θ is more than one standard deviation from θ_0). Which values of c and n should be chosen to meet this?

The requirements will be met if

$$\beta(\theta_0) = 0.1 \qquad \beta(\theta_0 + \sigma) = 0.8.$$

$$0.1 = \beta(\theta_0) = 1 - \Phi(c) \Rightarrow c = z_{0.1} \simeq 1.28$$

$$\begin{aligned} 0.8 &= \beta(\theta_0 + \sigma) = 1 - \Phi\left(c - \sqrt{n}\right) \\ &\Rightarrow \Phi\left(\sqrt{n} - c\right) = 0.8 \Rightarrow \sqrt{n} - c = z_{0.2} \simeq 0.84 \Rightarrow n = (1.28 + 0.84)^2 = 4.49 \end{aligned}$$

Since n is an integer, choose $n = 5$. □

8.4 Most Powerful Tests

Definition 8.6. *Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathcal{C}$ with power function $\beta(\theta)$ is a* uniformly most powerful (UMP) class $\mathcal{C}$ test *if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every β' that is a power function of a test in class $\mathcal{C}$.*

The requirements of this definition are so strong that UMP tests do not exist in many realistic problems. When they do, they are very useful. The following theorem describes which tests are UMP level α tests in the situation where where the null and alternative hypotheses both consist of only one probability distribution (H_0 and H_1 are *simple* hypotheses).

Theorem 8.7 (Neyman-Pearson Lemma)**.** *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the p.d.f or p.m.f. corresponding to θ_i is $p(., \theta_i)$, $i = 0, 1$, using a test with rejection region $\mathcal{R}$ that satisfies*

$$\underline{x} \in \mathcal{R} \quad \textit{if} \quad p(\underline{x}, \theta_1) > kp(\underline{x}, \theta_0) \quad \textit{and} \quad \underline{x} \in \mathcal{R}^c \quad \textit{if} \quad p(\underline{x}, \theta_1) < kp(\underline{x}, \theta_0) \tag{8.1}$$

for some $k \geq 0$ (both inequalities are strict) and

$$\alpha = \mathbb{P}_{\theta_0}(\underline{X} \in \mathcal{R}).$$

Then

1. *(Sufficiency) Any test that satisfies these criteria is a MP (Most Powerful) level α test.*
2. *(Necessity) If there exists a test satisfying these two criteria with $k > 0$, then every MP level α test is a size α test and every UMP level α test satisfies Equation* (8.1) *except perhaps on a set A satisfying $\mathbb{P}_{\theta_0}(\underline{X} \in A) = \mathbb{P}_{\theta_1}(\underline{X} \in A) = 0$.*

The theorem has the following corollary:

Corollary 8.8. *Suppose that $T(\underline{X})$ is a sufficient statistic for θ and $g(t, \theta_i)$ the p.d.f. or p.m.f. of T corresponding to θ_i, $i = 0, 1$. Then any test based on T with rejection region $\mathcal{S}$ is a MP level α test if it satisfies*

$$t \in \mathcal{S} \qquad \textit{if} \qquad g(t,\theta_1) > kg(t,\theta_0) \qquad \textit{and} \qquad t \in \mathcal{S}^c \qquad \textit{if} \qquad g(t,\theta_1) < kg(t,\theta_0). \tag{8.2}$$

for some $k \geq 0$ where

$$\alpha = \mathbb{P}_{\theta_0}(T \in \mathcal{S}).$$

Proof (Exercise - use the theorem and the factorisation theorem) □

Before giving the proof of the Neyman-Pearson lemma, the following examples may be instructive.

Example 8.5 (UMP Binomial test)**.** *Let $X \sim$ Binomial$(2, \theta)$. Find the Most Powerful test for $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta = \frac{3}{4}$ with a given level α.*

Solution The probability ratios are:

$$\frac{p_X(0|\theta = \frac{3}{4})}{p_X(0|\theta = \frac{1}{2})} = \frac{1}{4}, \qquad \frac{p_X(1|\theta = \frac{3}{4})}{p_X(1|\theta = \frac{1}{2})} = \frac{3}{4} \qquad \frac{p_X(2|\theta = \frac{3}{4})}{p_X(2|\theta = \frac{1}{2})} = \frac{9}{4}.$$

If we choose $\frac{3}{4} < k < \frac{9}{4}$, the Neyman-Pearson lemma says that $\mathcal{R} = \{2\}$ (reject H_0 for $x = 2$) is the MP level $\alpha = p_X(2|\theta = \frac{1}{2}) = \frac{1}{4}$ test.

If we choose $\frac{1}{4} < k < \frac{3}{4}$, then the Neyman-Pearson lemma says that $\mathcal{R} = \{1, 2\}$ (reject H_0 for $x = 1$ or 2) is the MP level $\alpha = \mathbb{P}(X \in \{1, 2\}|\theta = \frac{1}{2}) = \frac{3}{4}$ test (probability of 0.75 of wrongly rejecting H_0).

If we choose $k < \frac{1}{4}$ or $k > \frac{9}{4}$, this gives the MP level $\alpha = 1$ ($\mathcal{R} = \phi$) or level $\alpha = 0$ ($\mathcal{R} = \{0, 1, 2\}$) test respectively.

If $k = \frac{3}{4}$, then Equation (8.1) says we must reject H_0 for the sample point $x = 2$ and accept for $x = 0$, but leaves $x = 1$ undetermined. If we do not reject for $x = 1$, then we get the MP level $\alpha = \frac{1}{4}$ test as above. If we reject for $x = 1$, we get the MP level $\alpha = \frac{3}{4}$ test above. □

Example 8.6 (MP Normal test)**.**

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, σ^2 known. The sample mean $\overline{X}$ is a sufficient statisticfor θ. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$. Let g denote the density function of $\overline{X}$. Then

$$g(\overline{x}, \theta_1) > kg(\overline{x}, \theta_0) \Leftrightarrow \overline{x} < \frac{\frac{1}{n}(2\sigma^2 \log k) - \theta_0^2 + \theta_1^2}{2(\theta_1 - \theta_0)}$$

The test with rejection region $\mathcal{R} = \{\overline{x} < c\}$ is the MP level α test where $\alpha = \mathbb{P}_{\theta_0}(\overline{X} < c)$. If a particular α is specified, then the MP test rejects H_0 if

$$\overline{X} < c = \theta_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$$

Proof of Neyman - Pearson Lemma For notational convenience, the proof is given for continuous variables; the proof for discrete variables is similar.

First, since Θ_0 contains only one point, any test satisfying $\mathbb{P}_{\theta_0}(X \in \mathcal{R}) = \alpha$ is a size α test and hence a level α test. Let $\mathcal{R}$ be the region defined by Equation (8.1) and $\mathbf{1}_{\mathcal{R}}$ the indicator function for this region. Let $\mathcal{R}'$ be any other region such that $\mathbb{P}_{\theta_0}(X \in \mathcal{R}') \leq \alpha$ (critical region of a level α test). Let $\beta(\theta)$ and $\beta'(\theta)$ be the power functions corresponding to rejection regions of $\mathcal{R}$ and $\mathcal{R}'$ respectively. Because $0 \leq \mathbf{1}_{\mathcal{R}'}(x) \leq 1$, it follows that

$$(\mathbf{1}_{\mathcal{R}}(x) - \mathbf{1}_{\mathcal{R}'}(x))(p(x, \theta_1) - kp(x, \theta_0)) \geq 0 \qquad \forall x \in \mathcal{X}$$

since $\mathbf{1}_{\mathcal{R}}(x) - \mathbf{1}_{\mathcal{R}'}(x) \geq 0$ when $p(x, \theta_1) > kp(x, \theta_0)$ and $\mathbf{1}_{\mathcal{R}}(x) - \mathbf{1}_{\mathcal{R}'}(x) \leq 0$ when $p(x, \theta_1) < kp(x, \theta_0)$.

It follows that

$$0 \leq \int (\mathbf{1}_{\mathcal{R}}(x) - \mathbf{1}_{\mathcal{R}'}(x))\,(p(x, \theta_1) - kp(x, \theta_0))\,dx = \big(\beta(\theta_1) - \beta'(\theta_1)\big) - k\,\big(\beta(\theta_0) - \beta'(\theta_0)\big)\,.$$

The *sufficiency* statement follows by noting that since ϕ' is a level α test and ϕ a size α test, $\beta(\theta_0) - \beta'(\theta_0) \geq 0$. It follows that

$$0 \leq (\beta(\theta_1) - \beta'(\theta_1)) - k(\beta(\theta_0) - \beta'(\theta_0)) \leq \beta(\theta_1) - \beta'(\theta_1).$$

It follows that the test with critical region $\mathcal{R}$ is the MP test.

To prove the *necessity* statement, let $\mathcal{R}'$ be the rejection region for any MP level α test. The test satisfying Equation (8.1) is also a MP level α test and therefore $\beta(\theta_1) = \beta'(\theta_1)$. It follows that

$$0 \leq -k(\beta(\theta_0) - \beta'(\theta_0))$$

and since $\mathcal{R}$ is the critical region for a size α test, it follows that $\beta'(\theta_0) = \alpha$; hence $\mathcal{R}'$ is the critical region for a size α test. It now follows that

$$\int \left(\mathbf{1}_{\mathcal{R}}(x) - \mathbf{1}_{\mathcal{R}'}(x)\right)\left(p(x, \theta_1) - kp(x, \theta_0)\right) dx = 0$$

and since the integrand is non negative, it is therefore zero everywhere. It follows that $\mathbf{1}_{\mathcal{R}'}$ is the indicator of the region from Equation (8.1) except possibly on a set A of $p(., \theta_i)$ measure 0 for $i = 0, 1$. The theorem is proved. □

Tutorial 9

1. Consider a situation where the parameter space has two elements, $\Theta = \{\theta_0, \theta_1\}$ Suppose we want to test $H_0 : \theta = \theta_0$ versus the alternative, $H_1 : \theta = \theta_1$. One way of doing this is to consider the test statistic

$$\nu(x) = \frac{L(\theta_1; x)}{L(\theta_0; x)},$$

the ratio of the likelihood functions. This is a different formulation, but gives the same test as the Likelihood Ratio statistic. We reject $H_0 : \theta = \theta_0$ in favour of $H_1 : \theta = \theta_1$ if $\nu(x)$ is large.

We have a single observation on a random variable X with distribution F, where F is either $U(0, 1)$ or $\text{Exp}(1)$. Construct the test described above, with significance level $\alpha = 0.05$ to test $H_0 : X \sim U(0, 1)$ versus the alternative $H_1 : X \sim \text{Exp}(1)$. Compute the rejection region for the test and compute its power when H_1 is true.

2. We have a single observation on the random variable X with density function

$$p(x, \theta) = \begin{cases} \theta e^{-x} + 2(1 - \theta)e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\theta \in [0, 1]$ is an unknown parameter.

   (a) Construct a test between the null hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta > 0$ with significance level $\alpha = 0.05$. (Use LRT method).

   (b) Compute the power function of this test.

3. Let $(U_j)_{j \geq 1}$ be a sequence of i.i.d. $U(0, 1)$ random variables. Let X be a random variable. It is required to test

$$H_0 : X = \min\{U_1, \ldots, U_k\} \qquad \text{versus} \qquad H_1 : X = \min\{U_1, \ldots, U_l\} \qquad l < k.$$

   (a) Construct a test with significance level α based on the statistic $\nu(x) := \frac{L(H_1;x)}{L(H_0;x)}$ where $L(H_1; x)$ and $L(H_0; x)$ denote the likelihoods based on H_1 and H_0 respectively (each hypothesis corresponds to a single parameter value).

   (b) What is the largest value of the ratio $\frac{l}{k}$ so that a test with significance $\alpha = 0.05$ has power at least 0.95?

4. Consider a population with three types of individual, labelled 1, 2 and 3, which occur in the Hardy - Weinberg proportions

$$p_\theta(1) = \theta^2 \qquad p_\theta(2) = 2\theta(1 - \theta) \qquad p_\theta(3) = (1 - \theta)^2.$$

For a sample $X_1, \ldots, X_n$ from this population, let $N_1 = \sum_{j=1}^n \mathbf{1}_1(X_j)$, $N_2 = \sum_{j=1}^n \mathbf{1}_2(X_j)$, $N_3 = \sum_{j=1}^n \mathbf{1}_3(X_j)$ denote the number of appearances of 1, 2, 3 respectively in the sample. Let $0 < \theta_0 < \theta_1 < 1$.

(a) Show that $\nu(\underline{x}; \theta_0, \theta_1) = \frac{L(\theta_1; \underline{x})}{L(\theta_0; \underline{x})}$ is an increasing function of $2N_1 + N_2$. (n is fixed).

(b) Show that if $c > 0$ and $\alpha \in (0, 1)$ satisfy

$$\mathbb{P}_{\theta_0}(2N_1 + N_2 > c) = \alpha$$

then a test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ with a given significance level α that rejects H_0 if and only if $2N_1 + N_2 > c$ corresponds to the test where $H_0 : \theta = \theta_0$ is rejected for large values of $\nu(\underline{x}; \theta_0, \theta_1)$, defined in the previous part.

5. Let $X_1, \ldots, X_n$ be i.i.d. $U(0, \theta)$ variables and let $M_n = \max\{X_1, \ldots, X_n\}$. Consider a test of $H_0 : \theta \leq \theta_0$ versus the alternative $H_1 : \theta > \theta_0$ where H_0 is rejected if and only if $M_n > c$ for some value $c > 0$.

   (a) Compute the power function of this test and show that it is monotone increasing in θ.

   (b) For $\theta_0 = \frac{1}{2}$, compute the value of c which would give the test a size exactly 0.05.

   (c) Compute the value of n so that the test of size 0.05 for $\theta_0 = \frac{1}{2}$ has power 0.98 for $\theta = \frac{3}{4}$.

6. Consider a simple hypothesis test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Suppose that the test statistic T has a continuous distribution and the null hypothesis is rejected for $t \geq c$ where t is the observed value of T for some c and that, as a function of c, the size of the test is:

$$\alpha(c) = \mathbb{P}_{\theta_0}(T \geq c).$$

   Prove that, for $\theta = \theta_0$, $\alpha(T) \sim U(0, 1)$.

7. Let $T_1, \ldots, T_r$ be independent test statistics for the same simple $H_0 : \theta = \theta_0$ and that for each j, T_j has a continuous distribution. Let $\alpha_j(c) = \mathbb{P}_{\theta_0}(T_j \geq c)$. Show that, under H_0, $\tilde{T} = -2\sum_{j=1}^r \log \alpha_j(T_j) \sim \chi^2_{2r}$.

8. Let $F_0(y) = \mathbb{P}(Y < y)$ where Y is a non negative random variable representing a survival time. Assume that F_0 has a density f_0. Let $X_1, \ldots, X_n$ be i.i.d. each with an alternative distribution, representing survival time under an alternative treatment. The new distribution is considered to take the form

$$G(y, \Delta) = 1 - (1 - F_0(y))^\Delta \qquad y > 0 \quad \Delta > 0.$$

   To test whether the new treatment is beneficial, test $H_0 : \Delta \leq 1$ versus $H_1 : \Delta > 1$. Compute the Likelihood Ratio Test and compute the critical region for a test with significance level α in terms of n and an appropriate χ^2 distribution. (This is known as the *Lehmann alternative*).

Answers

1. $f_0(x) = 1$ for $0 \leq x \leq 1$. $f_1(x) = \exp\{-x\}$ for $x \geq 0$. For the Neyman Pearson test, the ratio is:

$$\nu(x) = \frac{f_1(x)}{f_0(x)} = \begin{cases} e^{-x} & 0 \leq x \leq 1 \\ +\infty & x > 1 \\ \text{undefined} & x < 0 \end{cases}$$

By the Neyman Pearson lemma, a test is a UMP test if and only if there is a k such that

$$x \in \mathcal{R} \qquad \text{if} \qquad \nu(x) > k \qquad \text{and} \qquad x \in \mathcal{R}^c \qquad \text{if} \qquad \nu(x) < k$$

For a 5% significance level,

$$\begin{aligned} 0.05 &= \mathbb{P}(\nu(X) > k | X \sim U(0,1)) \\ &= \mathbb{P}(\{X < -\log k\} \cup \{X > 1\} | X \sim U(0,1)) = -\log k \Rightarrow k = e^{-0.05} \end{aligned}$$

Rejection region $\mathcal{R} = [0, 0.05] \cup [1, +\infty]$. The power of the test when $X \sim Exp(1)$ is

$$\mathbb{P}(\{X < 0.05\} \cup \{X > 1\} | X \sim Exp(1)) = (1 - e^{-0.05}) + e^{-1}.$$

2. (a) LRT First find $\widehat{\theta}_{ML}$;

$$L(\theta; x) = \theta(e^{-x} - 2e^{-2x}) + 2e^{-2x}$$

$$\widehat{\theta}_{ML} = \begin{cases} 0 & x < \log 2 \\ 1 & x > \log 2 \\ \in [0,1] & x = \log 2 \end{cases}$$

$$\lambda(x) = \frac{L(0,x)}{L(\widehat{\theta}_{ML}, x)} = \begin{cases} 1 & 0 \leq x \leq \log 2 \\ \frac{2e^{-2x}}{e^{-x}} = 2e^{-x} & x > \log 2. \end{cases}$$

Reject H_0 if

$$\lambda(x) < c \Rightarrow 2e^{-x} < c \Rightarrow x > -\log \frac{c}{2} = k.$$

k determined by:

$$0.05 = \mathbb{P}_{\theta=0}(X > k) = \int_k^\infty 2e^{-2x} dx = e^{-2k} \Rightarrow k = \frac{1}{2} \log 20.$$

$$\mathcal{R} = (\frac{1}{2} \log 20, +\infty).$$

(b)

$$\beta(\theta) = \mathbb{P}_\theta(X > \frac{1}{2} \log 20) = \frac{\theta}{\sqrt{20}} + (1-\theta)\frac{1}{20}.$$

3. (a) Under H_0, X has density $f_0(x) = k(1-x)^{k-1} \quad 0 \leq x \leq 1$ and under H_1, X has density $f_1(x) = l(1-x)^{l-1} \quad 0 \leq x \leq 1$.

The N-P ratio is:

$$\nu(x) = \frac{f_1(x)}{f_0(x)} = \frac{l}{k}(1-x)^{l-k}.$$

For N-P test, reject H_0 for large values of $\lambda(x)$;

$$x \in \mathcal{R} \quad \text{if} \quad \frac{l}{k}(1-x)^{l-k} > c \qquad x \in \mathcal{R}^c \quad \text{if} \quad \frac{l}{k}(1-x)^{l-k} < c.$$

Simplifying gives:

$$\mathcal{R} = \{x | \frac{l}{k}(1-x)^{l-k} > c\} \Rightarrow \mathcal{R} = \{x | x > 1 - \left(\frac{ck}{l}\right)^{1/(l-k)} = K\}$$

where we set $K := 1 - \left(\frac{ck}{l}\right)^{1/(l-k)}$. Then, for a size α test, K satisfies:

$$\alpha = \mathbb{P}_0(X > K) = \int_K^1 k(1-x)^{k-1} dx = (1-K)^k \Rightarrow K = 1 - \alpha^{1/k}.$$

so that H_0 is rejected for $X \in \mathcal{R}$ where

$$\mathcal{R} = [1 - \alpha^{1/k}, 1].$$

(b) We require a power of 0.95 when $\alpha = 0.05$ and H_1 is true. Then:

$$0.95 = \int_{1-0.05^{1/k}}^1 l(1-x)^{l-1} dx = 0.05^{l/k} \Rightarrow \frac{l}{k} = \frac{\log 0.95}{\log 0.05} = \frac{\log(20/19)}{\log 20}.$$

4. (a)

$$\begin{aligned} \nu(\underline{x}; \theta_0, \theta_1) &= \left(\frac{\theta_1}{\theta_0}\right)^{2N_1} \left(\frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}\right)^{N_2} \left(\frac{(1-\theta_1)}{(1-\theta_0)}\right)^{2n-2(N_1+N_2)} \\ &= \left(\frac{\theta_1}{\theta_0}\right)^{2N_1+N_2} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{2n-(2N_1+N_2)} \end{aligned}$$

and since $\frac{\theta_1}{\theta_0} > 1$ and $\frac{1-\theta_1}{1-\theta_0} < 1$, this is increasing in $2N_1 + N_2$ for fixed n.

(b) Let $\mathcal{R}$ denote rejection region from Neyman Pearson lemma, a test is UMP if and only if it satisfies

$$\underline{x} \in \mathcal{R} \quad \text{if} \quad \nu(\underline{x}; \theta_0, \theta_1) > k \qquad \underline{x} \in \mathcal{R}^c \quad \text{if} \quad \nu(\underline{x}; \theta_0, \theta_1) < k$$

for some k.

$$\begin{aligned}\nu(\underline{x};\theta_0,\theta_1) > k \quad &\Rightarrow \quad (2N_1+N_2)\left(\log\frac{\theta_1}{\theta_0} - \log\frac{1-\theta_1}{1-\theta_0}\right) + 2n\log\frac{1-\theta_1}{1-\theta_0} > \log k \\ &\Rightarrow \quad 2N_1+N_2 > \frac{\log k - 2n\log\frac{1-\theta_1}{1-\theta_0}}{\log\frac{\theta_1}{\theta_0} - \log\frac{1-\theta_1}{1-\theta_0}} = K.\end{aligned}$$

To get the UMP test of significance level α, K has to be chosen such that

$$\mathbb{P}(2N_1+N_2 > K) = \alpha$$

so that $K = c$.

5. (a) $P(\theta) = \mathbb{P}_\theta(M_n > c) = 1 - \left(\frac{c}{\theta}\right)^n \mathbf{1}_{\{c<\theta\}}$ (monotone non-decreasing)

   (b) $0.05 = P(\frac{1}{2}) = 1 - (2c)^n\mathbf{1}_{\{c<\theta\}} \Rightarrow c = \frac{1}{2}(0.95)^{1/n}$.

   (c) $0.98 = P(\frac{3}{4}) = 1 - 0.95\left(\frac{2}{3}\right)^n$ so that $n = \frac{\log(95/2)}{\log(3/2)}$. First integer greater than or equal to this gives $n = 10$.

6. Let $\gamma(c) = 1 - \alpha(c)$ then, since T is continuous, γ is the c.d.f. of T and hence $\gamma(T) \sim U(0,1)$. It follows that $\alpha(T) = 1 - \gamma(T) \sim U(0,1)$.

7. This follows directly from the previous exercise: for each j $\alpha_j(T_j) \sim U(0,1)$ from which it follow directly that $-\log\alpha_j(T_j) \sim \text{Exp}(1)$ and hence $-2\sum_{j=1}^r \log\alpha_j(T_j) \sim \Gamma(r,\frac{1}{2}) = \chi^2_{2r}$.

8. $g(y,\Delta) = \Delta(1-F_0(y))^{\Delta-1}f_0(y)$. It follows that

$$\lambda(\underline{x}) = \frac{\sup_{0\leq\Delta\leq 1}\Delta^n\left(\prod_{j=1}^n(1-F_0(x_j))\right)^{\Delta-1}}{\sup_{0\leq\Delta<+\infty}\Delta^n\left(\prod_{j=1}^n(1-F_0(x_j))\right)^{\Delta-1}} = \begin{cases} 1 & \Delta^*\leq 1 \\ \frac{1}{\Delta^{*n}\left(e^{-n+(n/\Delta^*)}\right)} & \Delta^* > 1\end{cases}$$

where

$$\Delta^* = -\frac{1}{\frac{1}{n}\sum_{j=1}^n\log(1-F_0(x_j))}.$$

This comes from solving

$$\left.\frac{d}{d\Delta}n\log\Delta + (\Delta-1)\log\prod_{j=1}(1-F_0(x_j))\right|_{\Delta=\Delta^*} = 0$$

giving

$$\frac{n}{\Delta^*} + \log\prod_{j=1}(1-F_0(x_j)) = 0 \Rightarrow \Delta^* = -\frac{1}{\frac{1}{n}\sum_{j=1}^n\log(1-F_0(x_j))}$$

The LRT rejects H_0 if the ratio is small; for some k to be determined

$$\mathcal{R} = \left\{ \underline{x} | \lambda(\underline{x}) < \frac{1}{k^n} \right\} = \left\{ \underline{x} | \Delta^* e^{(1/\Delta^*)-1} > k \right\}.$$

For $x > 1$, $xe^{(1/x)-1}$ is *increasing* (take derivative of log) hence critical region is

$$\mathcal{R} = \left\{ \underline{x} | -\sum_{j=1}^{n} \log(1 - F_0(x_j)) < c \right\}$$

for some constant c.

Recall that if X has c.d.f. F for F continuous, then

$$\mathbb{P}(F(X) \leq x) = \mathbb{P}(X \leq F^{-1}(x)) = F(F^{-1}(x)) = x$$

so $F(X) \sim U(0,1)$ and hence $1 - F(X) \sim U(0,1)$.

For a prescribed significance level α, under the assumption of H_0, X_j has distribution F_0 and hence

$$\alpha \quad = \quad \mathbb{P}(\underline{X} \in \mathcal{R}) = \mathbb{P}\left(-\sum_{j=1}^{n} \log U_j < c \right)$$

where $U_1, \ldots, U_n$ are i.i.d. $U(0,1)$ If $U_j \sim U(0,1)$, then $-\log U_j \sim Exp(1)$ and hence

$$-\sum_{j=1}^{n} \log U_j \sim \text{gamma}(n, 1)$$

so that $W := -2\sum_{j=1}^{n} \log U_j \sim \chi^2_{2n}$, so

$$\alpha = \mathbb{P}(W < 2c) \Rightarrow c = \frac{1}{2} k_{2n,1-\alpha}$$

where $k_{2n,\gamma}$ is the value such that $\mathbb{P}(W > k_{2n,\gamma}) = \gamma$.

Chapter 9

Hypothesis Testing (II)

9.1 Monotone Likelihood Ratio

The Neyman-Pearson lemma only enables us to conclude that when the parameter space has exactly two elements, the LRT test is the most powerful when choosing between the two. We would like to extend this, to obtain Uniformly Most Powerful tests in more general situations. The results are limited, because with a range of parameters satisfying H_1 (the alternative hypothesis), it is not clear that there is a test with a power function which is optimal over the whole range of parameters.

The *monotone likelihood ratio* property in the situation where $\Theta \subseteq \mathbb{R}$ enables us to establish uniformly most powerful tests for the situation of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. We do this by considering binary NP situations of the NP and showing that the optimal NP test is the *same* for *each* $\theta' \in \Theta \backslash \Theta_0$.

Definition 9.1. *A family of p.d.f.s (probability density functions) or p.m.f.s (probability mass functions) $g(t, \theta)$ for a univariate random variable T with real valued parameter θ has* monotone likelihood ratio *(MLR) if for every $\theta_2 > \theta_2$, $\frac{g(t,\theta_2)}{g(t,\theta_1)}$ is a monotone (non-increasing or non-decreasing) function of t on $\{t : g(t, \theta_1) + g(t, \theta_2) > 0\}$, where $\frac{c}{0}$ is defined as $+\infty$ if $c \neq 0$.*

Note that any density of the form:

$$g(t, \theta) = h(t)c(\theta) \exp\{w(\theta)t\}$$

satisfies MLR if $w(\theta)$ is a monotone function.

Theorem 9.2 (Karlin - Rubin)**.** *Consider a test $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of p.d.f.s (or p.m.f.s) for T satisfies MLR, where $\frac{g(t,\theta_2)}{g(t,\theta_1)}$ is non-decreasing in t for $\theta_2 > \theta_1$. Then for any t_0, the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test where $\alpha = \mathbb{P}_{\theta_0}(T > t_0)$.*

Proof Let $\beta(\theta) = \mathbb{P}_\theta(T > t_0)$ be the power function of the test. Fix $\theta' > \theta_0$ and consider the test $H_0' : \theta = \theta_0$ versus $H_1' : \theta = \theta'$.

Let

$$k' = \inf_{t \in \mathcal{T}} \frac{g(t,\theta')}{g(t,\theta_0)} \qquad \mathcal{T} = \{t : t > t_0 \qquad \text{and} \qquad g(t,\theta') + g(t,\theta_0) > 0\}.$$

Then, by the MLR property,

$$\{T > t_0\} \qquad \Leftrightarrow \qquad \frac{g(t,\theta')}{g(t,\theta_0)} > k'.$$

It now follows from Corollary 8.8 that the power function β^* for any other level α test satisfies $\beta(\theta') \geq \beta^*(\theta')$. Since $\beta^*(\theta_0) \leq \sup_{\theta \in \Theta_0} \beta^*(\theta) \leq \alpha$, it follows that $\beta(\theta') \geq \beta^*(\theta')$ for any level α test of H_0. The test is therefore a level α test. Since θ' is arbitrary, it follows that the test is a UMP level α test. □

Example 9.1 (Normal Random Sample, Variance Known, Testing Mean)**.**

If $X_1, \ldots, X_n$ is a $N(\mu, \sigma^2)$ random sample, where σ^2 is known, then $T := \overline{X}$ is a sufficient statistic for μ. Its density is $N(\mu, \frac{\sigma^2}{n})$ and the likelihood ratio is

$$\frac{g(t,\mu_2)}{g(t,\mu_1)} = \exp\left\{\frac{n}{2\sigma^2}\left(\mu_1 + \mu_2 - 2t\right)\left(\mu_1 - \mu_2\right)\right\}$$

and it is clear that it satisfies the MLR property. Consider the test

$$H_0 : \mu \geq \mu_0 \qquad \text{versus} \qquad H_1 : \mu < \mu_0,$$

with

$$\mathcal{R}_{\text{crit}} = \left(-\infty, \mu_0 - \frac{\sigma z_\alpha}{\sqrt{n}}\right).$$

From the preceding result, this is a UMP level α test. The power function is:

$$\beta(\mu) = \mathbb{P}_\mu\left(\overline{X} < \mu - \frac{\sigma z_\alpha}{\sqrt{n}}\right) = \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right)$$

which is clearly decreasing as a function of μ, so that

$$\beta(\mu_0) = \sup_{\mu \geq \mu_0} \beta(\mu) = \alpha.$$

□

The following simple example illustrates that the UMP test does not, in general, exist. Two tests, which satisfy the *necessary* conditions for UMP are given, but they do not satisfy the *sufficient* conditions.

Example 9.2 (Two sided normal, variance known)**.**

Let $X \sim N(\mu, \sigma^2)$, where σ^2 is known. Consider the test

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0.$$

For a specified value of α, a level α test is any test that satisfies

$$\mathbb{P}_{\mu_0}\left(\text{reject} \quad H_0\right) \leq \alpha.$$

Consider the test: reject H_0 if $x < \mu_0 - \sigma z_\alpha$ where x is the observed value of X. This test has greatest power in the region $\{\mu : \mu < \mu_0\}$. By the necessity part of the Neyman-Pearson lemma, any other level α test, for any $\mu < \mu_0$, a level α test of the same power has the same rejection region up to a set of both $\mathbb{P}_\mu$ and $\mathbb{P}_{\mu_0}$ measure zero. Let β denote the power function of this test.

Now consider the test: reject H_0 if $x > \mu_0 + \sigma z_\alpha$. This is also a level α test. Let β^* denote its power function, then it is clear that

$$\beta^*(\mu) > \beta(\mu) \qquad \forall \mu > \mu_0.$$

These contradictory tests both satisfy necessary (but not sufficient) conditions to be UMP tests, therefore there does not exist a UMP test. □

To obtain sensible results, an additional requirement has to be added to the class of tests under consideration; tests should be *unbiased.*

Definition 9.3 (Unbiased Test)**.** *A statistical test of size (level) α, $0 < \alpha < 1$ for testing $H_0 : \theta \in \Theta_0 \subset \Theta$ against an alternative $H_1 : \theta \in \Theta \backslash \Theta_0$ whose power function satisfies*

$$\begin{cases} \beta(\theta) \leq \alpha & \theta \in \Theta_0 \\ \beta(\theta) \geq \alpha & \theta \in \Theta \backslash \Theta_0 \end{cases}$$

is said to be unbiased

Exercise In the example of $X \sim N(\mu, \sigma^2)$ where σ^2 is known, the test with critical region

$$\mathcal{R}_{\text{crit}} = (-\infty, \mu_0 - \sigma z_{\alpha/2}] \cup [\mu_0 + \sigma z_{\alpha/2}, +\infty)$$

is UMP among the class of unbiased tests. □

9.2 Union-Intersection and Intersection-Union Tests

A *union-intersection* test is one where the null hypothesis is of the form:

$$H_0 : \theta \in \widetilde{\Theta} := \bigcap_{\gamma \in \Gamma} \Theta_\gamma \qquad \text{versus} \qquad H_1 : \theta \notin \widetilde{\Theta}$$

where Γ is an indexing set and $\Theta_\gamma \subset \Theta$ for each $\gamma \in \Gamma$. Suppose that a test is available for each of the problems: $H_0^\gamma : \theta \in \Theta_\gamma$ versus $H_1^\gamma : \theta \in \Theta \backslash \Theta_\gamma$ and suppose the rejection region for this test is: $\mathcal{R}_\gamma$. The rejection region for the test as a whole is $\cup_{\gamma \in \Gamma} \mathcal{R}_\gamma$. That is, the null hypothesis is rejected if it is rejected for any one of the tests.

Similarly, an intersection-union test (IUT) has null hypothesis of the form

$$H_0 : \theta \in \widetilde{\Theta} = \bigcup_{\gamma \in \Gamma} \Theta_\gamma \qquad \text{versus} \qquad H_1 : \theta \notin \widetilde{\Theta}$$

and, if the rejection region for test γ is $\mathcal{R}_\gamma$, then the rejection region for the test as a whole is $\mathcal{R} := \cap_{\gamma \in \Gamma} \mathcal{R}_\gamma$.
Let $\lambda_\gamma(x)$ be the LRT statistic for testing $H_0^\gamma : \theta \in \Theta_\gamma$ versus $H_1^\gamma : \theta \notin \Theta_\gamma$ and consider the UIT $H_0 : \theta \in \widetilde{\Theta} := \cap_{\gamma \in \Gamma} \Theta_\gamma$ versus $H_1 : \theta \notin \widetilde{\Theta}$. Then the following results gives the relationships between the overall LRT and the UIT based on $\lambda_\gamma(x)$.

Theorem 9.4. *Consider testing $H_0 : \theta \in \widetilde{\Theta} := \cap_{\gamma \in \Gamma} \Theta_\gamma$ versus $H_1 : \theta \in \Theta \backslash \widetilde{\Theta}$ and let $\lambda(x)$ denote the LRT for this test. Let $\lambda_\gamma(x)$ be the LRT statistic for $H_0 : \theta \in \Theta_\gamma$ versus $H_1 : \theta \in \Theta \backslash \Theta_\gamma$. Let*

$$T(x) = \inf_{\gamma \in \Gamma} \lambda_\gamma(x).$$

Consider two tests: the first with rejection region $\{x : T(x) < c\}$ and the second with rejection region $\{x : \lambda(x) < c\}$, for a given c. Then

1. *$T(x) \geq \lambda(x)$ for each x,*

2. *If β_T and β_λ are the likelihood functions for the UIT and LRT tests respectively, then $\beta_T(\theta) \leq \beta_\lambda(\theta)$ for each $\theta \in \Theta$.*

3. *If the LRT is a level α test, then the UIT is a level α test.*

Proof From the definition of an LRT statistic of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta \backslash \Theta_0$, $\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta, x)}{\sup_{\theta \in \Theta} L(\theta, x)}$, it is clear that $\lambda_\gamma(x) \geq \lambda(x)$ for each $\gamma \in \Gamma$, since $\widehat{\Theta} \subset \Theta_\gamma$ for each $\gamma \in \Gamma$. It follows that $T(x) := \inf_{\gamma \in \Gamma} \lambda_\gamma(x) \geq \lambda(x)$, thus proving the first part. The second follows trivially from the definition, since $\{x : T(x) < c\} \subseteq \{x : \lambda(x) < c\}$. The third also follows directly from the definition. □

The following result for IUT tests is trivially obvious:

Theorem 9.5. *Consider the test of $H_0 : \theta \in \widetilde{\Theta}$ versus $H_1 : \theta \in \Theta \backslash \widetilde{\Theta}$ where $\widetilde{\Theta} = \cup_{\gamma \in \Gamma} \Theta_\gamma$. Let α_γ be the size of the test of $H_0^\gamma : \theta \in \Theta_\gamma$ versus $H_1^\gamma : \theta \in \Theta \backslash \Theta_\gamma$. Let R_γ be the rejection region for this test and let $R = \cap_{\gamma \in \Gamma} R_\gamma$ be the rejection region for H_0 versus H_1. Then the IUT test with rejection region R is a level $\alpha = \sup_{\gamma \in \Gamma} \alpha_\gamma$ test.*

Proof For any $\theta \in \widehat{\Theta}$,

$$\mathbb{P}_\theta(X \in R) \leq \mathbb{P}_\theta(X \in R_\gamma) \leq \alpha_\gamma < \alpha.$$

□

The following theorem gives conditions under which the size of the IUT is exactly α.

Theorem 9.6. *Consider the test $H_0 : \theta \in \cup_{j=1}^k \Theta_j$, where k is a finite positive integer and let R_j denote the rejection region of a level α test for $H_0^j : \theta \in \Theta_j$. Suppose that there is an $i \in \{1, \ldots, k\}$ for which there is a sequence $\theta_l \in \Theta_i : l = 1, 2, \ldots$ such that*

1. *$\lim_{l \to +\infty} \mathbb{P}_{\theta_l}(X \in R_i) = \alpha$ and*

2. *for each $j \in \{1, \ldots, k\} \backslash \{i\}$, $\lim_{l \to +\infty} \mathbb{P}_{\theta_l}(X \in R_j) = 1$.*

Then the IUT test with rejection region $R = \cap_{j=1}^k R_j$ is a size α test.

Proof By the previous result (which is obvious), R is a level α test:

$$\sup_{\theta \in \widetilde{\Theta}} \mathbb{P}_\theta(X \in R) \leq \alpha.$$

Furthermore,

$$\begin{aligned}\sup_{\theta\in\widehat{\Theta}}\mathbb{P}_\theta(X\in R) &\geq \lim_{l\to+\infty}\mathbb{P}_{\theta_l}(X\in R)=\lim_{l\to+\infty}\mathbb{P}_{\theta_l}(X\in\cap_{j=1}^k R_j)\\ &= 1-\lim_{l\to+\infty}\mathbb{P}_{\theta_l}(X\in\cup_{j=1}^k R_j^c)\geq 1-\lim_{l\to+\infty}\sum_{j=1}^k\mathbb{P}_{\theta_l}(X\in R_j^c)\\ &= 1-(1-\alpha)=\alpha\end{aligned}$$

and hence equality. □

9.3 p-Values

A p-value is a statistic $p(X)$, which takes values in the unit interval $[0,1]$, such that if H_0 is true, then $p(X)\geq U$ where $U\sim\text{Unif}(0,1)$. If x is the observed value and $p(x)$ is small, then the null hypothesis looks unlikely. If α is the significance level and $p(x)<\alpha$, then the null hypothesis is rejected; if $p(x)>\alpha$ then the null hypothesis is not rejected. More formally, the definition is:

Definition 9.7 (p-value)**.** *A p-value $p(X)$ for a test $H_0:\theta\in\Theta_0$ is a test statistic $p:\mathcal{X}\to[0,1]$ where, for each $\theta\in\Theta_0$ and $\alpha\in[0,1]$,*

$$\mathbb{P}_\theta\left(p(X)\leq\alpha\right)\leq\alpha.$$

A level α test may be constructed on a valid p-value; the test for $H_0:\theta\in\Theta_0$ with rejection region $\mathcal{R}_{\text{crit}}=\{x:p(x)\leq\alpha\}$ is a level α test.

The following theorem gives the most common ways of defining p values.

Theorem 9.8. *Let $W(X)$ be a test statistic for $H_0:\theta\in\Theta_0$ versus $H_1:\theta\in\Theta\backslash\Theta_0$ and suppose that the rejection region for H_0 takes the form:*

$$\mathcal{R}_{crit}=\{x:W(x)>c\}$$

for some $c\in\mathbb{R}$. Define:

$$p(x)=\sup_{\theta\in\Theta_0}\mathbb{P}(W(X)\geq W(x)).$$

Then $p(X)$ is a valid p-value.

Proof Set $p_\theta(x):=\mathbb{P}_\theta(W(X)\geq W(x))$ so that $p(x)=\sup_{\theta\in\Theta_0}p_\theta(x)$. Let $F_\theta(w)=\mathbb{P}_\theta(W(X)<w)$, so that

$$F_\theta(W(x))=\mathbb{P}_\theta(W(X)<W(x))=1-p_\theta(x).$$

Since F_θ is monotone in w, it follows that

$$\mathbb{P}_\theta(F_\theta(W))<y)=\mathbb{P}_\theta(W<F_\theta^{-1}(y))\geq y,$$

so that for $\theta\in\Theta_0$,

$$\begin{aligned}\mathbb{P}_\theta(\sup_{\theta\in\Theta_0} p_\theta(X)\leq\alpha) &\leq \mathbb{P}_\theta(p_\theta(X)\leq\alpha)\\ &= \mathbb{P}_\theta\left(1-F_\theta(W(X))\leq\alpha\right)\\ &= \mathbb{P}_\theta\left(F_\theta(W(X))\geq 1-\alpha\right)=1-\mathbb{P}_\theta\left(F_\theta(W(X))<1-\alpha\right)\leq\alpha.\end{aligned}$$

□

Example 9.3 (Normal: one sided test of mean)**.** *This example gives a p-value statistic for a test of $H_0:\mu\leq\mu_0$ against $H_1:\mu>\mu_0$ for a $N(\mu,\sigma^2)$ random sample.*

Let $X=(X_1,\ldots,X_n)$ be a $N(\mu,\sigma^2)$ random sample. Consider the test of $H_0:\mu\leq\mu_0$ versus $H_1:\mu>\mu_0$. As shown earlier, the LRT rejects H_0 for large values of

$$W(X)=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\qquad S^2=\frac{1}{n-1}\sum_{j=1}^n(X_j-\overline{X})^2.$$

Let

$$p(x)=\sup_{(\mu,\sigma):\mu\leq\mu_0}\mathbb{P}_{(\mu,\sigma)}(W(X)\leq W(x)).$$

A consequence of the previous result is that $p(X)$ is a valid p-value, as the following computation shows.

For $\mu\leq\mu_0$,

$$\begin{aligned}\mathbb{P}_{(\mu,\sigma)}\left(W(X)\geq W(x)\right) &= \mathbb{P}_{(\mu,\sigma)}\left(\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\geq W(x)\right)=\mathbb{P}_{(\mu,\sigma)}\left(\frac{\overline{X}-\mu}{S/\sqrt{n}}\geq W(x)+\frac{\mu_0-\mu}{S/\sqrt{n}}\right)\\ &\leq \mathbb{P}(T\geq W(x))=\mathbb{P}_{(\mu_0,\sigma)}(W(X)\geq W(x))\end{aligned}$$

where $T\sim t_{n-1}$. Hence

$$p(x)=\mathbb{P}(T\geq W(x))=\mathbb{P}\left(T\geq\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right)=1-F_{T,n-1}\left(\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right)$$

where $x=(x_1,\ldots,x_n)$ denotes the observed random sample, $\overline{x}=\frac{1}{n}\sum_{j=1}^n x_j$ and
$s^2=\frac{1}{n-1}\sum_{j=1}^n(x_j-\overline{x})^2$ and $F_{T,n-1}$ denotes the cumulative distribution function for a T distribution with $n-1$ degrees of freedom. Clearly $p(X)\sim U(0,1)$ if $X=(X_1,\ldots,X_n)$ is an i.i.d. $N(\mu_0,\sigma^2)$ random sample; a small observed value $p(x)$ indicates that H_0 is unlikely. □

p-values by conditioning on a sufficient statistic Another way to define a p-value is to condition on a sufficient statistic $S(X)$. Let W denote a test statistic for a test with critical region $\mathcal{R}_{\text{crit}}=\{x: W(x)\geq c\}$ for some c. For $x\in\mathcal{X}$, let

$$p(x|s)=\mathbb{P}(W(X)\geq W(x)|S(X)=s).$$

By definition of 'sufficient statistic', the distribution does not depend on the parameters. This may be useful for discrete variables; here for any $\theta \in \Theta$,

$$\mathbb{P}_\theta(p(X) \leq \alpha) = \sum_s \mathbb{P}(p(X) \leq \alpha | S(X) = s)\mathbb{P}_\theta(S = s) \leq \alpha.$$

Therefore, if $p(x|s)$ is a valid p value for each s, then the overall statistic is a valid p value.

Example 9.4 (Fisher's Exact Test)**.**

Let S_1 and S_2 be two independent random variables, $S_1 \sim \text{binomial}(n_1, p_1)$ and $S_2 \sim \text{binomial}(n_2, p_2)$. Consider the test $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$. The joint probability mass function under $H_0 : p_1 = p_2 = p$ is:

$$p_{S_1,S_2}(s_1, s_2; p) = \binom{n_1}{s_1}\binom{n_2}{s_2} p^{s_1+s_2}(1-p)^{(n_1+n_2)-(s_1+s_2)},$$

so that $S = S_1 + S_2$ is a sufficient statistic under H_0. Conditioned on S, S_1 may be used as the test statistic, since S_2 gives no further information. The conditional distribution of S_1 given S is hypergeometric

$$\mathbb{P}(S_1 = k | S = s) = \frac{\binom{n_1}{k}\binom{n_2}{s-k}}{\binom{n_1+n_2}{s}}.$$

Let s_1 denote the observed value of S_1. This test rejects H_0 for $s_1 \in [x, n_1]$ for suitable x. The conditional p value is therefore:

$$p((s_1, s_2)) = \sum_{x}^{n_1} \mathbb{P}(S_1 = j | S = s).$$

The test defined by this p-value is known as *Fisher's Exact Test.* □

9.4 Interval Estimator by Inverting a Test Statistic

There is a correspondence between hypothesis testing and interval estimation. Every interval estimator corresponds to a test statistic and vice versa. This is the subject of the following theorem:

Theorem 9.9. *For each $\theta_0 \in \Theta$, let $\mathcal{R}(\theta_0)$ denote the critical region of a level α test of $H_0 : \theta = \theta_0$. For each $x \in \mathcal{X}$, let*

$$c(x) = \{\theta_0 : x \notin \mathcal{R}(\theta_0)\}.$$

Then the set $C(X)$ is a $1 - \alpha$ confidence set.
Conversely, let $C(X)$ be a $1 - \alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$\mathcal{R}(\theta_0) = \{x : \theta_0 \notin C(x)\}.$$

Then $\mathcal{R}(\theta_0)$ is the critical region for a level α test of $H_0 : \theta = \theta_0$.

Proof For the first part, for each $\theta \in \Theta$,

$$\alpha \geq \mathbb{P}_\theta(X \in \mathcal{R}(\theta)) = 1 - \mathbb{P}_\theta(\theta \in C(X)) \Rightarrow \mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha.$$

and hence $C(X)$ is a $1 - \alpha$ confidence set.

For the second part, for all $\theta \in \Theta$,

$$\mathbb{P}_\theta(X \in \mathcal{R}(\theta)) = \mathbb{P}_\theta(\theta \notin C(X)) \leq \alpha$$

so that this is a level α test. □

Example 9.5 (Inverting a LRT)**.**

Let $X_1, \ldots, X_n$ be an $\text{Exp}(\lambda)$ random sample. Construct a confidence interval for λ by inverting the LRT.

Solution The sample space is $\mathcal{X} = \mathbb{R}_+^n$. Let $\underline{x} = (x_1, \ldots, x_n) \in \mathcal{X}$ denote an outcome and $\overline{x} = \frac{1}{n}\sum_{j=1}^n x_j$.For the test $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$, the LRT is:

$$\lambda(\underline{x}) = \frac{\lambda_0^n e^{-\lambda_0 \sum_{j=1}^n x_j}}{\sup_\lambda \lambda^n e^{-\lambda \sum_{j=1}^n x_j}} = (\overline{x}\lambda)^n e^{n - n\lambda\overline{x}}.$$

For a fixed λ_0, the critical region is

$$\mathcal{R}(\lambda_0) = \left\{\underline{x} : \lambda_0 \overline{x} e^{-\lambda_0 \overline{x}} < k^*\right\}$$

where k^* is a constant chosen to satisfying

$$\mathbb{P}_{\lambda_0}\left(\underline{X} \in \mathcal{R}(\lambda_0)\right) = \alpha.$$

The inversion method gives a confidence set of

$$C(\underline{x}) = \left\{\lambda : \lambda\overline{x}e^{-\lambda\overline{x}} \geq k^*\right\}.$$

The interval depends on $\underline{x}$ only through $\overline{x}$;

$$C(\underline{x}) = \left[\frac{a}{\overline{x}}, \frac{b}{\overline{x}}\right]$$

where a

$$ae^{-a} = be^{-b} = k^*.$$

By taking logarithms, it is clear that the equation $ye^{-y} = k^*$ has no solutions for $k^* > e^{-1}$, has one solution ($y = 1$) for $k^* = e^{-1}$ and two solutions for $k^* \in (0, e^{-1})$.
Recall that $2\lambda \sum_{j=1}^n X_j \sim \chi^2_{2n}$. Let F be the c.d.f. of the χ^2_{2n} distribution, then

$$1 - \alpha = \mathbb{P}_\lambda \left(2na \leq 2\lambda \sum_{j=1}^{n} X_j \leq 2nb \right) = F(2nb) - F(2na)$$

and solutions may be obtaine numerically by finding a and b which satisfy:

$$\begin{cases} F(2nb) - F(2na) = 1 - \alpha \\ ae^{-a} = be^{-b}. \end{cases}$$

This does not give a symmetric confidence interval; the exact symmetric confidence interval for λ is:

$$\left[\frac{\frac{1}{2n} k_{2n,1-(\alpha/2)}}{\overline{X}}, \frac{\frac{1}{2n} k_{2n,\alpha/2}}{\overline{X}} \right]$$

where $k_{2n,\beta}$ is the value such that $1 - F(k_{2n,\beta}) = \beta$.

The LRT interval has the advantage over the symmetric interval that the parameter values in the interval are those which give the best likelihood ratios. Except for a few particular examples, it cannot be computed explicitly and needs numerical approximations. □

Summary

- Definition of hypothesis test.
- Likelihood ratio test.
- Errors and power function, size and level of a test.

Tutorial 10

1. We have a single observation on a random variable X from a distribution with density

$$p(x;\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

where $\theta > 0$ is unknown. We test $H_0 : \theta = 0$ against the alternative $H_1 : \theta > 0$ and we reject the null hypothesis if the observed value $x \in [c, +\infty) = \mathcal{R}_{\text{crit}}$ for an appropriate $c > 0$.

(a) Compute c if the test has significance level $\alpha = 0.05$.

(b) Determine whether or not this test is uniformly most powerful.

2. Let $X_1, \ldots, X_n$ be i.i.d. with distribution $F(x)$ where

$$F(x) = \begin{cases} 1 - e^{-x^\theta} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad \theta > 0.$$

Find the most powerful test for $H_0 : \theta = 1$ versus $H_1 : \theta = \theta_1$ for a particular $\theta_1 > 1$. For $\alpha = 0.05$, show that this does not give a UMP test for $H_0 : \theta = 1$ versus $H_1 : \theta > 1$.

3. Let

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right) \qquad i = 1, \ldots, n.$$

and suppose that $X_1, \ldots, X_n$ are independent. Consider the hypothesis test: $H_0 : \mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ versus the alternative $H_1 : \mu_1 \neq \mu_2$ or $\sigma_1 \neq \sigma_2$. Compute the likelihood ratio test statistic.

4. The $F_{n,m}$ distribution is defined as follows: if $V \sim \chi^2_m$, $W \sim \chi^2_n$ and $V \perp W$, then $F := \frac{W/n}{V/m}$ has $F_{n,m}$ distribution. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent exponential Exp(θ) and Exp(λ) samples respectively and let $\Delta = \frac{\theta}{\lambda}$.

(a) Let $f(\alpha)$ denote the value such that $\mathbb{P}(F > f(\alpha)) = \alpha$ where $F \sim F_{2n_1,2n_2}$. Show that $\left[\frac{\overline{Y}}{\overline{X}} f\left(1 - \frac{\alpha}{2}\right), \frac{\overline{Y}}{\overline{X}} f\left(\frac{\alpha}{2}\right)\right]$ is a confidence interval for Δ with confidence coefficient $1 - \alpha$.

(b) Show that the test with acceptance region (the region where H_0 is not rejected) given by $[f(1 - \alpha/2), f(\alpha/2)]$ for the test $H_0 : \Delta = 1$ versus $H_1 : \Delta \neq 1$ using test statistic $\widehat{\Delta} = \frac{\overline{X}}{\overline{Y}}$ has size α.

5. Let $X_1, \ldots, X_n$ denote the times (in days) to failure of n similar pieces of equipment which is considered to be an Exp(λ) random sample. Consider the hypothesis $H_0 : \frac{1}{\lambda} = \mu \leq \mu_0$ (the average lifetime is no greater than μ_0).

(a) Show that the test with critical region $\overline{X} \in \left[\mu_0 \frac{k_{2n,\alpha}}{2n}, +\infty\right)$ where $k_{m,\alpha}$ is the value such that $\mathbb{P}(W > k_{m,\alpha}) = \alpha$ for $W \sim \chi^2_m$, is a size α test.

(b) Give an expression for the power function in terms of the χ^2_{2n} distribution.

(c) Use the central limit theorem to show that $\Phi\left(-\frac{\mu_0 z_\alpha}{\mu}+\frac{\sqrt{n}(\mu-\mu_0)}{\mu}\right)$ is an approximation to the power function of the test in part (a). Here z_α is the value such that $\mathbb{P}(Z > z_\alpha) = \alpha$ for $Z \sim N(0,1)$ and $\Phi(z) = \mathbb{P}(Z \leq x)$.

6. Let $X_1, \ldots, X_n$ be a random sample from Poiss(θ), where θ is unknown.

(a) Construct a UMP level α test for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

(b) Show that the power function of the test is increasing in θ.

(c) What distribution tables would you need to calculate the power function of the UMP test?

(d) Give an approximate expression, derived using the central limit theorem, for the critical value (above which you reject H_0) if n is large and θ not too close to 0 or $+\infty$.

7. (a) Given a random sample $X_1, \ldots, X_n$ from a distribution with c.d.f. F, let

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{(-\infty,x]}(X_j)$$

denote the empirical distribution. Consider the test of $H_0 : F = F_0$ versus the alternative $H_1 : F \neq F_0$ for a given F_0. Let $D_n = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_0(x)|$ and consider the test: reject H_0 if and only if $D_n \geq k_\alpha$ for k_α such that $\mathbb{P}_{F_0}(D_n \geq k_\alpha) = \alpha$ under H_0. (Recall that asymptotically D_n has the Kolmogorov distribution).

Show that the *power* of this test (for a true distribution F), $\beta(F)$, is bounded below by

$$\beta(F) \geq \sup_x \mathbb{P}_F\left(|\widehat{F}_n(x) - F_0(x)| \geq k_\alpha\right).$$

(b) For $n = 80$, obtain an approximation $\widetilde{k_{0.1}}$ for $k_{0.1}$ and see how well it approximates the true value:

$$k_{0.1} \simeq \frac{1.2}{\sqrt{n}} \simeq 0.134$$

(c) Again, take $\alpha = 0.10$ and $n = 80$. Let F_0 be the $N(0,1)$ c.d.f. and

$$F(x) = \frac{1}{1+\exp\left\{-\frac{x}{\tau}\right\}} \qquad -\infty < x < +\infty \qquad \tau = \frac{\sqrt{3}}{\pi}.$$

With this choice of τ, this is the *logistic* distribution with mean zero and variance 1. Evaluate the lower bound $\mathbb{P}_F(|\widehat{F}_n(x) - F_0(x)| \geq k_\alpha)$ for $\alpha = 0.10$, $n = 80$ and $x = 1.5$ using the normal approximation to the binomial distribution of $n\widehat{F}(x)$ and the approximate critical value of the previous part. (the value of 1.2 may be obtained from tables of the Kolmogorov Smirnov distribution). If F_0 is the c.d.f. for $N(0,1)$ then $F_0(1.5) \simeq 0.93$.

(d) Show that if $F \neq F_0$ and F and F_0 are continuous, then the power of this test tends to 1 as $n \to +\infty$. You may use the fact that $\sqrt{n}D_n$ under the null hypothesis converges to the Kolmogorov Smirnov distribution. In particular, $\sqrt{n}k_{0.1} \overset{n\to+\infty}{\longrightarrow} c_{0.1} = 1.2$.

8. Again, let $X_1, \ldots, X_n$ be a random sample from a distribution with continuous c.d.f. F and let $\widehat{F}_n$ denote the empirical distribution. Let $\psi : (0,1) \to (0,+\infty)$ and $\alpha > 0$. Define the statistics:

$$S_{\psi,\alpha} = \sup_x \psi(F_0(x))|\widehat{F}(x) - F_0(x)|^\alpha$$

$$T_{\psi,\alpha} = \sup_x \psi(\widehat{F}_n(x))|\widehat{F}(x) - F_0(x)|^\alpha$$

$$U_{\psi,\alpha} = \int \psi(F_0(x))|\widehat{F}_n(x) - F_0(x)|^\alpha F_0(dx)$$

$$V_{\psi,\alpha} = \int \psi(\widehat{F}_n(x))|\widehat{F}_n(x) - F_0(x)|^\alpha \widehat{F}_n(dx)$$

For each of these statistics show that the distribution under $H_0 : F = F_0$, does not depend on F_0 (continuous).

Short Answers

1. (a)

$$0.05 = \mathbb{P}_0(X \geq c) = e^{-c} \Rightarrow c = \log 20$$

(b) This follows from the *Karlin-Rubin* theorem (Theorem 9.2). Clearly X is sufficient for θ and the likelihood ratio satisfies for $\theta_1 < \theta_2$:

$$\lambda(\theta_1, \theta_2; x) := \frac{e^{-(x-\theta_2)}}{e^{-(x-\theta_1)}} \frac{\mathbf{1}_{[\theta_2,+\infty)}(x)}{\mathbf{1}_{[\theta_1,+\infty)}(x)} = \begin{cases} \text{undefined} & x, \theta_1 \\ 0 & \theta_1 \leq x < \theta_2 \\ e^{\theta_1 - \theta_2} & x \geq \theta_2. \end{cases}$$

This is monotone in x for x in the support of at least one of $p(. : \theta_1)$ or $p(.; \theta_2)$.Hence, by Karlin-Rubin theorem, $\mathcal{R}_{\text{crit}} = \{x : x > c\}$ is UMP.

2. Here the density is $p(x; \theta) = \theta x^{\theta-1} e^{-x^\theta} \mathbf{1}_{[0,+\infty)}(x)$. Therefore, for an NP test of $H_0 : \theta = 1$ against an alternative, we need the ratio:

$$\lambda(\theta; \underline{x}) := \theta^n \left(\prod_{j=1}^n x_j \right)^{\theta-1} \exp\left\{ -\sum_{j=1}^n (x_j^\theta - x_j) \right\}.$$

where $\underline{x} = (x_1, \ldots, x_n)$ denotes the vector of observations.

The NP rejection region for H_0: $\theta = 1$ against $H_1 : \theta = \theta_1$ of size α, where $\theta_1 > 1$ is:

$$\mathcal{R}_{\text{crit};\theta_1,\alpha} = \{\underline{x} : \lambda(\theta_1, \underline{x}) > k_{\theta_1,\alpha}\}$$

where $k_{\theta_1,\alpha}$ satisfies:

$$\mathbb{P}(\lambda(\theta_1, \underline{X}) > k_{\theta_1,\alpha}) = \alpha, \qquad \underline{X} \quad \text{i.i.d.} \quad \text{Exp}(1).$$

This is the most powerful level α test.

A test (reject if $\underline{x} \in \mathcal{R}$ for some critical region $\mathcal{R}$) is UMP for all $\theta > 1$ if and only if, for the given level α it satisfies $\mathbb{P}(\underline{X} \in \mathcal{R}) \leq \alpha$ and

$$\beta(\theta) := \mathbb{P}_\theta(\underline{X} \in \mathcal{R}) \geq \mathbb{P}_\theta(\underline{X} \in \widetilde{\mathcal{R}})$$

for any other $\widetilde{\mathcal{R}}$ such that

$$\mathbb{P}_1(\underline{X} \in \widetilde{\mathcal{R}}) \leq \alpha.$$

The NP lemma gives an ‘if and only if’ condition. That is, a test is UMP if the NP tests give the same critical region for all $\theta > 1$. The critical region may be expressed as:

$$\mathcal{R}_\theta = \left\{ \underline{x} : (\theta - 1)\sum_{j=1}^{n} \log x_j - \sum_{j=1}^{n}(x_j^\theta - x_j) > c_\theta \right\}$$

where c_θ is chosen such that for $X_1, \ldots, X_n$ i.i.d. Exp(1) variables,

$$\mathbb{P}(\underline{X} \in \mathcal{R}_\theta) = 0.05.$$

For the test to be UMP, we need: for all $\underline{x} \in \mathbb{R}_+^n$,

$$(\theta_1 - 1)\sum_{j=1}^{n} \log x_j - \sum_{j=1}^{n} x_j^{\theta_1} + \sum_{j=1}^{n} x_j \geq c_{\theta_1} \Leftrightarrow (\theta_2 - 1)\sum_{j=1}^{n} \log x_j - \sum_{j=1}^{n} x_j^{\theta_2} + \sum_{j=1}^{n} x_j \geq c_{\theta_2}.$$

For $n \geq 2$, the shapes of the regions depend on the parameter θ. Indeed, we can see that $\lim_{\theta \to +\infty} F(x;\theta) = \mathbf{1}_{[1,+\infty)}(x)$; for any $\epsilon > 0$,

$$\lim_{\theta \to +\infty} \mathbb{P}_\theta(1 - \epsilon < \min(X_1, \ldots, X_n) \leq \max(X_1, \ldots, X_n) < 1 + \epsilon) = 1,$$

so that $\mathcal{R}_{\text{limit}} = \bigcap_{\theta > 1} \mathcal{R}_\theta = \{(1, \ldots, 1)\}$

and

$$\mathbb{P}_\theta((X_1, \ldots, X_n) \in \mathcal{R}_{\text{limit}}) = 0.$$

3. The samples $X_{11}, \ldots, X_{n1}$ and $X_{12}, \ldots, X_{n2}$ are independent. The log likelihood is:

$$\log L(\mu_1, \mu_2, \sigma_1, \sigma_2) = \text{const} - n \log \sigma_1 - n \log \sigma_2 - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n}(x_{i1} - \mu_1)^2 - \frac{1}{2\sigma_2}\sum_{i=1}(x_{i2} - \mu_2)^2$$

For constrained problem (under H_0) the max. likelihood of $\mu = \mu_1 = \mu_2$ is:

$$\widehat{\mu} = \frac{1}{2n}(\sum_{i=1}^{n} x_{i1} + \sum_{i=1}^{n} x_{i2}) = \overline{x}_{..}$$

and the max. likelihood for $\sigma^2 = \sigma_1^2 = \sigma_2^2$ is:

$$\widehat{\sigma}^2 = \frac{1}{2n}\left(\sum_{i=1}^{n}(x_{i1} - \overline{x}_{..})^2 + \sum_{i=1}^{n}(x_{i2} - \overline{x}_{..})^2\right).$$

For the unconstrained problem,

$$\widehat{\mu}_1 = \overline{x}_{.1} \qquad \widehat{\mu}_2 = \overline{x}_{.2}$$

while

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{i1} - \overline{X}_{.1})^2, \qquad \widehat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{i2} - \overline{X}_{.2})^2.$$

Then, for the constrained problem,

$$L(\widehat{\mu}, \widehat{\sigma}) = \frac{1}{(2\pi)^n \widehat{\sigma}^{2n}} \exp\{-n\}$$

and, for the unconstrained problem,

$$L(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1, \widehat{\sigma}_2) = \frac{1}{(2\pi)^n \widehat{\sigma}_1^n \widehat{\sigma}_2^n} \exp\{-n\},$$

so that the LRT is:

$$\lambda(x) = \frac{\widehat{\sigma}_1^n \widehat{\sigma}_2^n}{\widetilde{\sigma}^{2n}}.$$

4. (a) $\sum_{j=1}^{n_1} X_j = n_1\overline{X} \sim \Gamma(n_1, \theta)$ so that $2n_1\theta\overline{X} \sim \Gamma(n_1, \frac{1}{2}) = \chi^2_{2n_1}$. Similarly, $2\theta n_2\overline{Y} \sim \chi^2_{2n_2}$. It follows that $\frac{\theta\overline{X}}{\lambda\overline{Y}} = \Delta\frac{\overline{X}}{\overline{Y}} \sim F_{2n_1,2n_2}$. Hence the $1-\alpha$ symmetric confidence interval is $f\left(1-\frac{\alpha}{2}\right) \le \Delta\frac{\overline{X}}{\overline{Y}} \le f(\frac{\alpha}{2})$ giving a confidence interval of

$$\Delta \in \left[\frac{\overline{Y}}{\overline{X}} f(1-\frac{\alpha}{2}), \frac{\overline{Y}}{\overline{X}} f(\frac{\alpha}{2})\right]$$

as required.

(b) Under $H_0 : \Delta = 1$, $F := \frac{\overline{X}}{\overline{Y}} \sim F_{2n_1,2n_2}$. Reject for observed value of F greater than $f(\frac{\alpha}{2})$ or less than $f(1-\frac{\alpha}{2})$. Hence acceptance region is $[f(1-\frac{\alpha}{2}), f(\frac{\alpha}{2})]$.

5. (a) $T := 2\lambda\sum_{j=1}^{n} X_j \sim \chi^2_{2n}$. For $\lambda = \frac{1}{\mu_0}$,

$$\mathbb{P}_{\mu_0}(\overline{X} \ge \mu_0\frac{k_{2n,\alpha}}{2n}) = \mathbb{P}(T > k_{2n,\alpha}) = \alpha.$$

Clearly the power function is increasing in μ, so this is a size α test.

(b)

$$\beta(\mu) = \mathbb{P}_\mu\left(\overline{X} \ge \mu_0\frac{k_{2n,\alpha}}{2n}\right) = \mathbb{P}(T \ge \frac{\mu_0}{\mu}k_{2n,\alpha}) = 1 - F\left(\frac{\mu_0}{\mu}k_{2n,\alpha}\right)$$

where F is the c.d.f. for the χ^2_{2n}.

(c) From CLT, approximately $\overline{X} \sim N(\mu, \frac{\mu^2}{n})$ so that

$$\beta(\mu) \simeq 1 - \Phi\left(\frac{\mu_0\frac{k_{2n,\alpha}}{2n} - \mu}{\mu/\sqrt{n}}\right) = \Phi\left(\frac{\sqrt{n}(\mu-\mu_0)}{\mu} + \frac{\sqrt{n}\mu_0}{\mu}(1 - \frac{k_{2n,\alpha}}{2n})\right)$$

Now, by CLT, if $V \sim \chi^2_{2n}$ then V is approximately $N(2n, 4n)$, so

$$\alpha = \mathbb{P}(W > k_{2n,\alpha}) \simeq \mathbb{P}(Z > \frac{k_{2n,\alpha} - 2n}{2\sqrt{n}})$$

so $z_\alpha \simeq \frac{k_{2n,\alpha}-2n}{2\sqrt{n}}$ and hence

$$\beta(\mu) \simeq \Phi\left(\frac{\sqrt{n}(\mu-\mu_0)}{\mu} - \frac{\mu_0}{\mu} z_\alpha\right)$$

as required.

6. (a) Construct UMP by Karlin-Rubin Theorem. Let $T = \sum_i X_i$, then T is sufficient for θ. Also, $T \sim \text{Poiss}(n\theta)$. To show that test with critical region $T \in [t_0, +\infty)$ is UMP, it is sufficient by K-R to show MLR. For $\theta_1 < \theta_2$,

$$\lambda(\theta_1, \theta_2; t) = \frac{L(\theta_1, t)}{L(\theta_0, t)} = \left(\frac{\theta_1}{\theta_0}\right)^t e^{-n(\theta_1-\theta_0)}$$

This is clearly monotone in t, hence likelihood ratio satisfies MLR property, hence UMP level α test is: reject H_0 for $\sum_j X_j > k$ for an integer k, chosen as the smallest value such that

$$\mathbb{P}_{\theta_0}(\sum_j X_j > k) \leq \alpha.$$

(b) For $S \sim \text{Poiss}(n\theta)$ when the parameter value is θ,

$$\beta(\theta) = \mathbb{P}_\theta\left(S > k\right)$$

Use S is the number of events by time 1 in a Poisson process with parameter $n\theta$. Then $\{S > k\} = \{T < 1\}$ where T is the time until the kth event. $T \sim \Gamma(k, n\theta)$ so that $2n\theta T \sim \Gamma(k, \frac{1}{2}) \sim \chi^2_{2k}$. It is clear that $\mathbb{P}_\theta(T < 1)$ is increasing in θ.

(c) From part (b), the chi squared distribution.

(d) By CLT: $\sum_j X_j \sim N(n\theta, n\theta)$ approximately; reject H_0 if $\frac{\sum_j X_j - n\theta_0}{\sqrt{n\theta_0}} = \frac{k-n\theta_0}{\sqrt{n\theta_0}} > z_\alpha$; k is the lowest integer greater than $n\theta_0 + \sqrt{n}\sqrt{\theta_0} z_\alpha$.

7. (a) First part is trivial: since $D_n \geq |\widehat{F}_n(x) - F_0(x)|$ for each x, it follows that

$$\mathbb{P}_F(|\widehat{F}_n(x) - F_0(x)| \geq k_\alpha) \leq \mathbb{P}_F(D_n \geq k_\alpha)$$

for eacn $x \in \mathbb{R}$ and hence

$$\sup_{x\in\mathbb{R}} \mathbb{P}_F(|\widehat{F}_n(x) - F_0(x)| \geq k_\alpha) \leq \mathbb{P}_F(D_n \geq k_\alpha)$$

(b) Approximation for k_α:

$$\alpha > \mathbb{P}_F(|\widehat{F}_n(x) - F(x)| > k_\alpha)$$

so

$$\alpha > \mathbb{P}_F(|n\widehat{F}_n(x) - nF(x)| > nk_\alpha)$$

Consider $X_1, \ldots, X_n$ i.i.d. $U(0,1)$ variables, then $F(x) = x$ and $n\widehat{F}_n(x) \sim Bi(n,x) \sim N(nx, nx(1-x))$ (central limit approximation). Let $Y \sim N(nx, nx(1-x))$ then

$$\alpha = \sup_{x \in \mathbb{R}} \mathbb{P}(|Z| > \frac{\sqrt{n}\widetilde{k}_\alpha}{\sqrt{x(1-x)}})$$

Recall that $\alpha = 0.1$. The supremum occurs at $x = \frac{1}{2}$. With $n = 80$,

$$0.1 = \mathbb{P}(|Z| > 17.89\widetilde{k}_{0.1}) \Rightarrow 17.89\widetilde{k}_{0.1} = 1.65 \Rightarrow \widetilde{k}_{0.1} \simeq 0.09$$

(value for the Kolmogorov Smirnov statistic: $k_\alpha = 0.136$).

(c) Note that $80 \times 0.136 \simeq 11$. For F_0 the $N(0,1)$ c.d.f., $80 \times F_0(1.5) = 74.4$, so look for:

$$\mathbb{P}_F(|n\widehat{F}_n(1.5) - 74.4| \geq 11)$$

where $n\widehat{F}_n(1) = \text{Binomial}(80, \mathbb{P}_F(X \leq 1.5))$. Using $\tau = \frac{\sqrt{3}}{\pi}$, it follows that

$$\mathbb{P}_F(X \leq 1.5) = \frac{1}{1 + e^{-1.5\pi/\sqrt{3}}} \simeq 0.94.$$

The answer is therefore

$$\mathbb{P}(Y \leq 56) + \mathbb{P}(Y \geq 79) \qquad Y \sim \text{Binomial}(80, 0.94) \simeq N(75.2, 4.5).$$

which is:

$$(1 - \Phi(9.1)) + (1 - \Phi(1.65)) \simeq 0.05$$

In other words, the test is catastrophically awful. The null hypothesis is wrong, nevertheless, we only reject it with probability 0.05.

(d) Choose a point x such that $0 < F(x) < 1$, $0 < F_0(x) < 1$ and $F(x) \neq F_0(x)$ and let $c_{0.1} \simeq 1.22$ the value such that

$$\lim_{n \to +\infty} \mathbb{P}_{F_0}(\sqrt{n}D_n \geq c_{0.1}) = 0.1.$$

Using $Y := n\widehat{F}_n(x) \sim \text{Binomial}(nF(x), nF(x)(1-F(x))$, it follows from the central limit theorem that the power $\beta_n(F)$ (based on sample size n) satisfies:

$$\begin{aligned}\lim_{n \to +\infty} \beta_n(F) \quad &\geq \quad \Phi\left(\frac{\sqrt{n}(F_0(x) - F(x))}{\sqrt{F(x)(1-F(x))}} - \frac{c_\alpha}{\sqrt{F(x)(1-F(x))}}\right) \\ &\quad + \Phi\left(\frac{\sqrt{n}(F(x) - F_0(x))}{\sqrt{F(x)(1-F(x))}} - \frac{c_\alpha}{\sqrt{F(x)(1-F(x))}}\right) \to 1\end{aligned}$$

from which the result follows; if $F_0(x) > F(x)$ then the first term converges to 1 and the second to 0; if $F(x) > F_0(x)$ then the first term converges to 0 and the second to 1.

8. The first and second are similar to arguments given before (earlier tutorial exercises). The third and fourth are similar; here is the argument for the fourth.

$$\mathbb{P}(V_{\psi,\alpha} > v) = \mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\psi\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{(-\infty,X_k]}(X_j)\right)\left|\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{(-\infty,X_k]}(X_j) - F_0(X_k)\right|^{\alpha} > v\right)$$

Now let $Y_k = F_0(X_k)$ so that $Y_1, \ldots, Y_n$ are i.i.d. $U(0,1)$, then

$$\mathbb{P}(V_{\psi,\alpha} > v) = \mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\psi\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{(-\infty,Y_k]}(Y_j)\right)\left|\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{(-\infty,Y_k]}(Y_j) - Y_k\right|^{\alpha} > v\right)$$

which does not depend on F_0.

Chapter 10

Gaussian Linear Models (I)

10.1 Introduction

In the classical Gaussian linear model, an observation Y_i among n independent observations $(Y_1, \ldots, Y_n)$ depends on known constants $x_{i1}, \ldots, x_{ip}$ via the relationship:

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \qquad i = 1, \ldots, n$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma^2)$. That is,

$$Y_i = x_{i.}\beta + \epsilon_i \qquad i = 1, \ldots, n$$

where $\beta = (\beta_1, \ldots, \beta_p)^t$ denotes the parameter vector. In vector / matrix notation, this is written:

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2 I_n)$$

where I_n is the $n \times n$ identity matrix and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$.

The variable Y_i is called the *response* variable, while X is the *design matrix* and x_{ij} are the *design values.*

Note *Linear* here means linear in the *parameters* β, with additive noise. For example, if Y is a response and x is an explanatory variable, then

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

is a Gaussian linear model. The right hand side is linear in $\beta = (\beta_0, \beta_1, \beta_2)^t$.

Example 10.1 (One Way Layout)**.**

Suppose we are comparing the performance of $p \geq 2$ treatments on a population and we administer only one treatment to each subject. A sample of size n_k get treatment k, $n_1 + \ldots + n_p = n$. The *one way layout* is the model:

$$Y_{kl} = \beta_k + \epsilon_{kl} \qquad 1 \le l \le n_k \qquad 1 \le k \le p$$

where Y_{kl} is the response of the lth subject in group k to treatment k. Here $\epsilon_{kl} \sim N(0, \sigma^2)$ are independent for different kl.

The variables may be re-labelled as a vector $Y = (Y_1, \ldots, Y_n)^t$ where $Y_{1l} = Y_l$ for $l = 1, \ldots, n_1$ and for subsequent labels $Y_{jl} = Y_{n_1 + \ldots + n_{j-1} + l}$. Then the design matrix has elements:

$$x_{ij} = \begin{cases} 1 & \sum_{k=1}^{j-1} n_k + 1 \le i \le \sum_{k=1}^{j} n_k \\ 0 & \text{otherwise} \end{cases}$$

Then

$$X = \begin{pmatrix} \mathbf{1}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{1}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{1}_p \end{pmatrix}$$

where $\mathbf{1}_j$ is the n_j-column vector with each entry 1.

The one way layout model is often reparametrised by introducing

$$\alpha = \frac{1}{p} \sum_{k=1}^{p} \beta_k \qquad \delta_k = \beta_k - \alpha \qquad k = 1, \ldots, p$$

where α denotes an average response and β_k denotes the difference between the average for the kth treatment and the overall average. The new parameters are: $\beta^* = (\alpha, \delta_1, \ldots, \delta_k)$ and the linear model is

$$Y = X^* \beta^* + \epsilon \qquad \epsilon \sim N(0, \sigma^2 I_n)$$

where $X^* = (\mathbf{1}|X)$, $\mathbf{1}$ is the vector with n ones.

The parametrisation β^* is not identifiable for $\beta^* \in \mathbb{R}^{p+1}$, but it is identifiable when restricted to the subspace $\{\beta^* \in \mathbb{R}^{p+1} | \sum_{k=1}^{p} \delta_k = 0\}$.

Even if β is not identifiable, the vector of means $\mu = (\mu_1, \ldots, \mu_n)^t$ of Y always is. It is given by

$$\mu = X\beta = \sum_{j=1}^{p} \beta_j x_{.j}.$$

Let r be the rank of X. We assume that $n \ge r$. The parameter space for β is $\mathbb{R}^p$ and the parameter space for μ is

$$M = \{\mu = X\beta \qquad \beta \in \mathbb{R}^p\}$$

This is the linear space spanned by the columns $x_{.j}$. The dimension of M is the rank of X. The parametrisation (β, σ^2) is identifiable if and only if $r = p$ and $n \geq p + 1$.

Many of the results are easier to establish if the linear model is expressed in its *canonical form.*

10.2 The Canonical Form of the Gaussian Linear Model

Suppose $\beta \in \mathbb{R}^p$, X is an $n \times p$ matrix and $\dim(M) = r$ where $r \leq p$. Let $v_1, \ldots, v_n$ be an orthonormal basis of $\mathbb{R}^n$ such that the first r components, $v_1, \ldots, v_r$ span M. Let

$$V = \left(\begin{array}{c} V^{(1)} \\ \hline V^{(2)} \end{array}\right) \qquad V^{(1)} = \left(\begin{array}{c} v_1^t \\ \hline \vdots \\ \hline v_r^t \end{array}\right), \qquad V^{(2)} = \left(\begin{array}{c} v_{r+1}^t \\ \hline \vdots \\ \hline v_n^t \end{array}\right)$$

$$U_i = v_i^t Y \qquad \eta_i = \mathbb{E}[U_i] = v_i^t \mu \qquad i = 1, \ldots, n$$

In other words,

$$U = VY, \qquad \eta = V\mu \qquad \text{where} \qquad V^{(2)}\mu = 0, \qquad V^{(1)}\mu = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_r \end{pmatrix}.$$

Theorem 10.1. *The U_i are independent and $U_i \sim N(\eta_i, \sigma^2)$ for $i = 1, \ldots, n$ where*

$$\eta_i = 0 \qquad i = r + 1, \ldots, n.$$

Proof Let V be the orthonormal matrix with rows $v_1^t, \ldots, v_n^t$. Then

$$U = VY, \qquad \eta = V\mu.$$

Furthermore, the covariance matrix of U is:

$$\Sigma^{(U)} = V\Sigma^{(Y)}V^t = \sigma^2 V I_n V^t = \sigma^2 I_n.$$

It follows that

$$U \sim N(\eta, \sigma^2 I).$$

□

The log likelihood function for U is

$$\log L(\eta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{r}(u_j - \eta_j)^2 - \frac{1}{2\sigma^2}\sum_{j=r+1}^{n} u_j^2. \qquad (10.1)$$

Theorem 10.2. *In the canonical Gaussian linear model with σ^2 unknown and $n \geq r+1$,*

1. *$T = (U_1, \ldots, U_r, \sum_{i=r+1}^n U_i^2)$ is sufficient for $(\eta_1, \ldots, \eta_r, \sigma^2)$.*
2. *The MLE of $(\eta_1, \ldots, \eta_r)$ is $(U_1, \ldots, U_r)$ and the MLE of σ^2 is $\frac{1}{n}\sum_{i=r+1}^n U_i^2$.*
3. *For $i = 1, \ldots, r$, U_i is the UMVU estimator of η_i and for any constants $c_1, \ldots, c_r$ with $\alpha = \sum_{j=1}^r c_j \eta_j$, $\widehat{\alpha} = \sum_{j=1}^r c_j U_j$ is the UMVU estimator of α.*
4. *Let $U = \binom{U^{(1)}}{U^{(2)}}$ where $U^{(1)} = (U_1, \ldots, U_r)^t$ and $U^{(2)} = (U_{r+1}, \ldots, U_n)^t$. The MLE of μ is*

$$\widehat{\mu} = \sum_{i=1}^r v_i U_i = V^{(1)t} U^{(1)}$$

and $\widehat{\mu}_i$ is the UMVU estimator for μ_i, for each $i = 1, \ldots, n$. Furthermore, for $i = 1, \ldots, r$, $U_i = v_i^t \widehat{\mu}$. That is:

$$U^{(1)} = \begin{pmatrix} \widehat{\eta}_1 \\ \vdots \\ \widehat{\eta}_r \end{pmatrix} = V^{(1)} \widehat{\mu}.$$

5. *$S^2 = \frac{1}{n-r}\sum_{i=r+1}^n U_i^2$ is the UMVU estimator of σ^2.*

Proof Part 1. It follows directly from the log likelihood equation (10.1) and the factorisation theorem that $(U_1, \ldots, U_r, \sum_{i=1}^n U_i^2)$ is sufficient. This is equivalent to $T = (U_1, \ldots, U_r, \sum_{i=r+1}^n U_i^2)$.

Part 2.

$$\frac{\partial}{\partial \eta_i} \log L(\eta) = \frac{1}{\sigma^2}(u_j - \eta_j)$$

giving $\widehat{\eta}_j = U_j$ for $j = 1, \ldots, r$. That this is a maximum follows from taking the second derivative, noting that it is negative, and noting that $\log L(\eta) \to -\infty$ as $\eta_j \to -\infty$ for each $j \in \{1, \ldots, r\}$.

$$\frac{\partial}{\partial \sigma^2} \log L(\eta, \sigma) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\sum_{j=1}^r (u_j - \eta_j)^2 + \sum_{j=r+1}^n u_j^2\right)$$

so that $(\widehat{\eta}, \widehat{\sigma}^2)$ satisfies:

$$\frac{n}{2\widehat{\sigma}^2} = \frac{1}{2\widehat{\sigma}^4}\sum_{j=r+1}^n U_j^2 \quad \Rightarrow \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=r+1}^n U_j^2.$$

Part 5. The unbiased property is straightforward; for $j \geq r+1$, $\mathbb{E}[U_j] = 0$ so that $\mathbb{E}[U_j^2] = \mathbf{V}(U_j) = \sigma^2$ from which it follows that

$$\mathbb{E}[S^2] = \frac{1}{n-r} \times (n-r) \times \sigma^2 = \sigma^2.$$

Uniform minimum variance has not been covered yet; the proof will be complete after the treatment of *complete statistics.* The sufficient statistics for the Gaussian are *complete* and the result follows from the Lehmann-Scheffé theorem for complete statistics 12.7 by noting that $S^2 = \mathbb{E}[S^2|T] = S^2$.

Part 3: Unbiased is straightforward; again, uniform minimum variance is a straighforward consequence of the Lehmann-Scheffé theorem for complete statistics, which is dealt with later; Theorem 12.7.

Part 4. Firstly, $\eta = V\mu$ so that $\widehat{\eta} = V\widehat{\mu}$ and hence $\widehat{\mu} = V^t\widehat{\eta}$. Now,

$$\widehat{\eta} = \begin{pmatrix} U^{(1)} \\ 0 \end{pmatrix}$$

so that

$$\widehat{\mu} = V^t \begin{pmatrix} U^{(1)} \\ 0 \end{pmatrix} = V^{(1)t}U^{(1)}.$$

Since $V^{(1)}V^{(1)t} = I$,

$$V^{(1)}\widehat{\mu} = V^{(1)}V^{(1)t}U^{(1)} = U^{(1)}.$$

□

10.3 Parameter Estimation for Gaussian Linear Models

For a vector $\underline{t} \in \mathbb{R}^n$, let $|\underline{t}| = \sqrt{\sum_{j=1}^n t_j^2}$. Let $\mathcal{S}$ denote a subset of $\mathbb{R}^n$. The *projection* $t_0 = \pi(t|\mathcal{S})$ of a point $t \in \mathbb{R}^n$ on $\mathcal{S}$ is the point:

$$t_0 = \arg\min\{|t - t_0| : t_0 \in \mathcal{S}\}.$$

Recall that the log likelihood function for $(\underline{\beta}, \sigma^2)$ is:

$$\log L(\underline{\beta}, \sigma^2; \underline{y}) = -n\log\sigma - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}|\underline{y} - X\underline{\beta}|^2.$$

It follows that the MLE $\widehat{\beta}$ is given by:

$$\widehat{\beta} = \arg\min\{|y - X\beta| : \beta \in \mathbb{R}^q\}.$$

Note that $\widehat{\beta}$, the MLE is also the least squares estimator (LSE).

Theorem 10.3. *Let $\mathcal{S} = \{X\beta : \beta \in \mathbb{R}^p\}$. For the Gaussian linear model,*

1. *Let $\mu = X\beta = \mathbb{E}[Y]$. Then $\widehat{\mu}$, the MLE of μ, is the unique projection of Y on $\mathcal{S}$ and is given by*

$$\widehat{\mu} = X\widehat{\beta}.$$

2. *$\widehat{\mu}$ is orthogonal to $Y - \widehat{\mu}$.*

3. *Let r denote the dimension of $\mathcal{S}$, then*

$$S^2 = \frac{1}{n-r}|Y - \widehat{\mu}|^2$$

is the UMVU estimator of σ^2.

4. *If $r = q$, then β is* identifiable, *$\beta = (X^tX)^{-1}X^t\mu$. Also, $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$ (the maximum likelihood and ordinary least squares estimators of β are equal), $\widehat{\beta}_{ML}$ is unique and given by*

$$\widehat{\beta} = (X^tX)^{-1}X^t\widehat{\mu} = (X^tX)^{-1}X^tY.$$

5. *If $r = q$, then $\widehat{\beta}$ is the UMVU estimator of β and $\widehat{\mu}$ is the UMVU estimator of μ.*

Proof

1. This follows because $\mathcal{S} = \{X\beta : \beta \in \mathbb{R}^p\}$.

2.

$$Y = VU, \qquad \widehat{\mu} = V^{(1)}U, \qquad Y - \widehat{\mu} = V^{(2)}U$$

so that

$$(\widehat{\mu}, Y - \widehat{\mu}) = UV^{(1)t}V^{(2)}U = 0.$$

3.

$$|Y - \widehat{\mu}|^2 = U^tV^{(2)t}V^{(2)}U = \sum_{j=r+1}^{n} U_j^2$$

from which (from before)

$$S^2 = \frac{1}{n-r}|Y - \widehat{\mu}|^2$$

UMVU follows as before.

4. $\mu = X\beta$ so that $X^t\mu = X^tX\beta$ and $X^t\widehat{\mu} = X^tX\widehat{\beta}$. Since X has full rank, X^tX is non singular and hence

$$\beta = (X^tX)^{-1}X^t\mu \qquad \widehat{\beta} = (X^tX)^{-1}X^t\widehat{\mu}.$$

The expression to be minimised is: $|Y - X\beta|^2 = Y^tY - 2Y^tX\beta + \beta^tX^tX\beta$. The likelihood equations, also known as *normal equations* are therefore:

$$X^tX\widehat{\beta} = X^tY$$

When X^tX is invertible, it follows that

$$\widehat{\beta} = (X^tX)^{-1}X^tY$$

as required.

5. Any linear combination of $Y_1, \ldots, Y_n$ is also a linear combination of $U_1, \ldots, U_n$ and any linear combination of $U_1, \ldots, U_n$ is a UMVU estimator of its expected value by Theorem 12.7. This is dealt with later in the treatment of *complete statistics*.

□

The estimator $\widehat{\mu} = X\widehat{\beta}$ of μ is the *fitted value* of μ, while $\widehat{\epsilon} := Y - \widehat{\mu}$ is the *residual vector*, which estimates ϵ the noise. If X^tX is invertible, then

$$\mathbb{E}[\widehat{\beta}] = (X^tX)^{-1}X^t\mathbb{E}[Y] = (X^tX)^{-1}X^tX\beta = \beta$$

and

$$\Sigma^{(\widehat{\beta})} = \sigma^2(X^tX)^{-1}X^tI_nX(X^tX)^{-1} = \sigma^2(X^tX)^{-1}.$$

The *fitted values* satisfy

$$\widehat{\mu} = X(X^tX)^{-1}X^tY.$$

It is common to write $\widehat{Y}$ for $\widehat{\mu}$; this is the vector of fitted values of Y. Let

$$H = X(X^tX)^{-1}X^t$$

then it is clear that $H^t = H$ (the matrix H is symmetric) and

$$H^2 = X(X^tX)^{-1}(X^tX)(X^tX)^{-1}X^t = X(X^tX)^{-1}X^t = H$$

so that H is *idempotent*. The covariance matrix for $\widehat{Y}$, $\Sigma^{(\widehat{Y})}$ is:

$$\Sigma^{(\widehat{Y})} = \sigma^2HI_nH^t = \sigma^2H.$$

The residual vector may be written as

$$\widehat{\epsilon} = Y - \widehat{Y} = (I_n - H)Y$$

and since $I_n - H$ is symmetric and idempotent,

$$\Sigma^{(\widehat{\epsilon})} = \sigma^2(I_n - H).$$

The following corollary gathers together these results.

Corollary 10.4. *Suppose X^tX is invertible. Then*

1. $\widehat{Y} \perp \widehat{\epsilon}$.

2. $\widehat{Y} \sim N(\mu, \sigma^2H)$, $\widehat{\epsilon} \sim N(0, \sigma^2(I_n - H))$

3. $\widehat{\beta} \sim N(\beta, \sigma^2(X^tX)^{-1})$.

Proof Firstly, $(\widehat{Y}, \widehat{\epsilon})$ is a linear transform of a Gaussian random vector $\underline{Y}$ and is therefore Gaussian. The covariance between $\widehat{Y}$ and $\widehat{\epsilon}$ satisfies:

$$C(\widehat{\epsilon}, \widehat{Y}) = \sigma^2(I_n - H)I_n H^t = \sigma^2(H - H^2) = 0$$

and hence $\widehat{Y} \perp \widehat{\epsilon}$. The remaining statements follow from the fact that $\widehat{Y}$, $\widehat{\epsilon}$ and $\widehat{\beta}$ are Gaussian, since they are linear transforms of $\underline{Y}$, their expectations and covariance structure have been computed above. □

10.4 Confidence Intervals and Prediction

Consider the regression problem:

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon} \qquad \underline{\epsilon} \sim N(\underline{0}, \sigma^2 I_n)$$

where $\underline{Y}$ represents n observations according to the model:

$$Y = \beta_0 + \sum_{j=1}^{q} x_j \beta_j + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

different observations are independent. There are q explanatory or regressor variables. Suppose that $X^t X$ is invertible. Since $\widehat{\beta} \sim N(\beta, \sigma^2(X^tX)^{-1})$, confidence intervals for linear combinations of β follow easily: let v be a $q+1$ vector, then

$$v^t\widehat{\beta} \sim N(v^t\beta, \sigma^2 v^t(X^tX)^{-1}v^t).$$

The unbiased estimator of σ^2 is $S^2 = \frac{1}{n-(q+1)}Q_{\text{res}}$, which has $n-(q+1)$ degrees of freedom (since $Q_{\text{res}} = \sum_{j=q+2}^{n} U_j^2$ in the transformed canonical variables) and hence the $1-\alpha$ symmetric confidence interval for $v^t\beta$ is:

$$v^t\beta \in (v^t\widehat{\beta} \pm s\sqrt{v^t(X^tX)^{-1}v}t_{n-(q+1);\alpha/2}).$$

Hence, for explanatory variables with values $(x_1, \ldots, x_q)$, the confidence interval for $\mathbb{E}[Y](x_1, \ldots, x_q)$ is given by the above expression with $v = (1, x_1, \ldots, x_q)^t$.

The *prediction* problem is that of finding a confidence interval for the next observation. As well as the uncertainty in the estimate of the expected value, there is also an additional independent error with variance σ^2 (estimated by S^2) and hence

$$Y_{\text{pred}}(x_1, \ldots, x_q) \in (v^t\widehat{\beta} \pm s\sqrt{1 + v^t(X^tX)^{-1}v}t_{n-(q+1),\alpha/2}).$$

Other intervals may be constructed from the distributional theory that has been developed.

Tutorial 11

1. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ random samples respectively.

   (a) Find the MLE of $\theta := (\mu_1, \mu_2, \sigma^2)$. Let c_n be the value such that $S^2 = c_n \widehat{\sigma}^2$ is an unbiased estimator of σ^2. What is c_n? What is S^2?

   (b) Consider testing $H_0 : \mu_1 \leq \mu_2$ versus $H_1 : \mu_1 > \mu_2$. Assume that $\alpha < \frac{1}{2}$. Show that the likelihood ratio test is equivalent to the test with critical (rejection) region

   $$\overline{x} - \overline{y} \geq s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{n_1+n_2-2,\alpha}.$$

   Here $t_{p,\alpha}$ is the value such that $\mathbb{P}(T > t_{p,\alpha}) = \alpha$ for $T \sim t_p$.

   (c) Compute a normal approximation to the power function and use it to find the sample size n needed for the level 0.01 test to have power 0.95 when $n_1 = n_2 = \frac{n}{2}$ and $\frac{\mu_1 - \mu_2}{\sigma} = \frac{1}{2}$.

2. Consider the linear Gaussian model $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$ where $\underline{\epsilon} \sim N(\underline{0}, \sigma^2 I_n)$, put into canonical coordinates via an orthonormal transform $\underline{U} = A\underline{Y}$ where $U_i \sim N(\eta_i, \sigma^2)$ for $i = 1, \ldots, r$ and $U_i \sim N(0, \sigma^2)$ for $i = r+1, \ldots, n$ with unknown parameters $\underline{\eta} = (\eta_1, \ldots, \eta_r)^t$ and σ^2, and log likelihood function:

   $$\log L(\underline{\eta}, \sigma^2; \underline{u}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{r}(u_i - \eta_i)^2 - \frac{1}{2\sigma^2}\sum_{i=r+1}^{n} u_i^2 - \frac{n}{2}\log(2\pi\sigma^2).$$

   Show that the MLE for $(\underline{\eta}, \sigma^2)$ does not exist if $n = r$ and that it is given by
   $(U_1, \ldots, U_r, \frac{1}{n}\sum_{i=r+1}^{n} U_i^2)$ if $n \geq r+1$. Show, in particular, that $\widehat{\sigma}^2_{ML} = \frac{1}{n}\left|\underline{Y} - \widehat{\mu}\right|^2$

3. Consider *simple* linear regression; there is one explanatory variable and

   $$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \quad \text{i.i.d.} \quad i = 1, \ldots, n$$

   where $x_1, \ldots, x_n$ are not all equal. Express this as a Gaussian linear model

   $$\underline{Y} = X\underline{\beta} + \underline{\epsilon}$$

   identifying X and $\underline{\beta}$.

   (a) Show that

   $$(X^t X)^{-1} = \frac{1}{\sum_{j=1}^{n}(x_j - \overline{x})^2}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}$$

   where $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ and $\overline{x^2} = \frac{1}{n}\sum_{j=1}^{n} x_j^2$.

   (b) Let $\binom{\widehat{\beta}_0}{\widehat{\beta}_1}$ denote the maximum likelihood estimator of $\beta = \binom{\beta_0}{\beta_1}$. What is the distribution of $\binom{\widehat{\beta}_0}{\widehat{\beta}_1}$?

(c) Let

$$S^2 = \frac{1}{n-2}\sum_{j=1}^{n}(Y_j - \widehat{\beta}_0 - \widehat{\beta}_1 x_j)^2.$$

What is the distribution of $\frac{(n-2)S^2}{\sigma^2}$?

(d) Suppose

$$Y(z) = \beta_0 + \beta_1 z + \epsilon \qquad \epsilon \sim N(0,\sigma^2).$$

Let s denote the observed value of S. Using $t_{p,\alpha}$ to denote the value such that $\mathbb{P}(T > t_{p,\alpha}) = \alpha$ for $T \sim t_p$, show that a symmetric confidence interval for $\mathbb{E}[Y(z)]$ is given by:

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 z \pm s\sqrt{\frac{1}{n} + \frac{(\overline{x}-z)^2}{\sum_{j=1}^n (x_j - \overline{x})^2}} t_{n-2,\alpha/2}\right).$$

(e) Let $Y_* = \beta_0 + \beta_1 z + \epsilon_*$ where $\epsilon_* \sim N(0,\sigma^2)$ is independent of $\epsilon_1,\ldots,\epsilon_n$ (Y_* is a new observation with explanatory variable set at z). Let $\overline{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$ and let $\widehat{Y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 z$ (the predictor of Y_* based on $Y_1,\ldots,Y_n$). Show that, if $\beta_1 = 0$, then

$$\mathbb{E}[(Y^* - \widehat{Y}^*)^2] \geq \mathbb{E}[(Y^* - \overline{Y})^2]$$

4. Consider the one way layout problem

$$Y_{ij} = \beta_i + \epsilon_{ij} \quad i = 1,\ldots,p \quad j = 1,\ldots,n_i$$

where ϵ_{ij} are i.i.d. $N(0,\sigma^2)$ and $n = n_1 + \ldots + n_p$.

(a) Show that

$$S^2 = \frac{\sum_{i=1}^p \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2}{\sum_{i=1}^p (n_i - 1)}$$

is an unbiased estimator of σ^2 and that

$$\frac{\sum_{i=1}^p \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2}{\sigma^2} \sim \chi^2_{n-p}.$$

(b) Show that a level $1-\alpha$ confidence intervals for $\beta_j - \beta_i$ is:

$$\beta_j - \beta_i \in \left(\overline{Y}_{j.} - \overline{Y}_{i.} \pm S t_{n-p;\alpha/2}\sqrt{\frac{n_i + n_j}{n_i n_j}}\right)$$

where S^2 is the unbiased estimator of σ, $\overline{Y}_{k.} = \frac{1}{n_k}\sum_{i=1}^{n_k} Y_{ki}$ and $t_{p,\alpha}$ denotes the value such that $\mathbb{P}(T > t_{p,\alpha}) = \alpha$ if $T \sim t_p$. Show that the level $1-\alpha$ confidence interval for σ^2 is given by:

$$\frac{(n-p)s^2}{k_{n-p;(\alpha/2)}} \leq \sigma^2 \leq \frac{(n-p)s^2}{k_{n-p;1-(\alpha/2)}}$$

where $k_{q,\beta}$ is the value such that $\mathbb{P}(V \geq k_{q,\beta}) = \beta$ if $V \sim \chi^2_q$.

(c) Find confidence intervals for $\psi = \frac{1}{2}(\beta_2+\beta_3)-\beta_1$ and $\sigma^2_\psi := \mathbf{V}(\widehat{\psi})$ where $\widehat{\psi} = \frac{1}{2}(\widehat{\beta}_2+\widehat{\beta}_3)-\widehat{\beta}_1$.

5. Show that if C is an $n \times r$ matrix of full rank r, $r \leq n$, then the $r \times r$ matrix C^tC is of rank r and hence non singular.

   Hint: Because C^t is of rank r, it follows that for any r-vector x, $x^tC^t = 0$ implies $x = 0$. Use this to show that if x is a non zero r-vector, then $x^tCC^tx > 0$.

6. Consider the one-way layout model: k groups of observations, all random variables independent. For group j, $Y_{1,j}, \ldots, Y_{n_j,j} \sim N(\mu_j, \sigma^2)$. Let $n = n_1 + \ldots + n_k$ denote the total number of observations.

   (a) Compute the likelihood ratio test statistic for $H_0 : \mu_1 = \ldots = \mu_k$ versus $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$.

   (b) Let $Q_{\text{res}} = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_{.j})^2$ where $\overline{Y}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$, the sample average from group j. Let $Q_M = \sum_{j=1}^k n_j(\overline{Y}_{.j} - \overline{Y}_{..})^2$ where $\overline{Y}_{..} = \frac{1}{n}\sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}$ (the overall average). Here Q_{res} denotes the *residual* sum of squares, while Q_M denotes the *model* sum of squares. Show that the likelihood ratio test is equavalent to reject H_0 for $F := \frac{Q_M/(k-1)}{Q\text{res}/(n-k)} > c$ for some $c > 0$.

   (c) Show that the statistic F has $F_{k-1,n-k}$ distribution.

Answers

1. (a) Computing maximum likelihood estimators for normal distribution parameters should be straightforward. The log-likelihood function is:

$$\log L(\mu_1,\mu_2,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{j=1}^{n_1}(x_j-\mu_1)^2 + \sum_{j=1}^{n_2}(y_j-\mu_2)^2\right).$$

This is maximised for:
$\widehat{\mu}_1 = \overline{X}$, $\widehat{\mu}_2 = \overline{Y}$,

$$\widehat{\sigma}^2_{ML} = \frac{1}{n_1+n_2}\left(\sum_{j=1}^{n_1}(X_j-\overline{X})^2 + \sum_{j=1}^{n_2}(Y_j-\overline{Y})^2\right)$$

This estimator is biased; recall that

$$\frac{\sum_{j=1}^{n_1}(X_j-\overline{X})^2}{\sigma^2} \sim \chi^2_{n_1-1} \qquad \frac{\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

and that, for $V \sim \chi^2_m$, $\mathbb{E}[V] = m$. Therefore:

$$\mathbb{E}[\widehat{\sigma}^2_{ML}] = \frac{n_1+n_2-2}{n_1+n_2}\sigma^2 \Rightarrow c_n = \frac{n_1+n_2}{n_1+n_2-2}$$

The unbiased estimator required in the question is therefore:

$$S^2 = \frac{1}{n_1+n_2-2}\left(\sum_{j=1}^{n_1}(X_j-\overline{X})^2 + \sum_{j=1}^{n_2}(Y_j-\overline{Y})^2\right)$$

(b) Recall $H_0 : \mu_1 \le \mu_2$ versus $H_1 : \mu_1 > \mu_2$. The log likelihood ratio test statistic is:

$$\lambda(x,y) = \frac{\sup_{\mu_1,\mu_2,\sigma\in H_0} L(\mu_1,\mu_2,\sigma;x,y)}{\sup_{\mu_1,\mu_2,\sigma} L(\mu_1,\mu_2,\sigma;x,y)} = \frac{L(\widehat{\mu}_{0,1},\widehat{\mu}_{0,2},\widehat{\sigma}_0)}{L(\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma})}$$

where $(\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma})$ are the MLE estimators for the full space

$$\Theta = \{(\mu_1,\mu_2,\sigma^2) : (\mu_1,\mu_2,\sigma^2) \in \mathbb{R}^2\times\mathbb{R}_+\} = \mathbb{R}^2\times\mathbb{R}_+.$$

These were computed in the previous part of the exercise. The values $(\widehat{\mu}_{0,1},\widehat{\mu}_{0,2},\widehat{\sigma}_0)$ are the values which maximise the likelihood over the null hypothesis space

$$\Theta_0 = \{(\mu_1,\mu_2,\sigma^2) \in \mathbb{R}^2\times\mathbb{R}_+ : \mu_1 \le \mu_2\}.$$

If $\overline{X} \le \overline{Y}$, then (clearly) $(\widehat{\mu}_{01},\widehat{\mu}_{02},\widehat{\sigma}^2_0) = (\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}^2)$ and hence $\lambda(x,y) = 1$ for $\overline{x} < \overline{y}$.

Now consider the other case, where $\overline{x} > \overline{y}$. The maximiser clearly does not lie in the interior of the space; in this case there are no solutions to the likelihood equations $\nabla_\theta \log L(\theta) = 0$ in the space Θ_0. Therefore the maximiser lies on the boundary.

Clearly, as $\mu_1 \to -\infty$ or $\mu_2 \to +\infty$, $\log L(\mu_1, \mu_2, \sigma) \to -\infty$, so the maximiser does not lie on the part of the boundary where parameter values are $\pm\infty$. Therefore, the maximiser lies on the boundary $\mu_1 = \mu_2$. Therefore, for $\overline{x} > \overline{y}$, $\widehat{\mu}_{0,1} = \widehat{\mu}_{02} = \widehat{\mu}_0$ where $(\widehat{\mu}_0, \widehat{\sigma}_0^2)$ are the values which maximise

$$\log L(\mu, \sigma) = -\frac{(n_1+n_2)}{2}\log(2\pi) - \frac{(n_1+n_2)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{j=1}^{n_1}(x_j-\mu)^2 + \sum_{j=1}^{n_2}(y_j-\mu)^2\right).$$

From this,

$$\begin{array}{l} \widehat{\mu}_0 = \frac{1}{n_1+n_2}\left(\sum_{j=1}^{n_1} X_j + \sum_{j=1}^{n_2} Y_j\right) \\ \widehat{\sigma}_0^2 = \frac{1}{n_1+n_2}\left(\sum_{j=1}^{n_1}(X_j-\widehat{\mu}_0)^2 + \sum_{j=1}^{n_2}(Y_j-\widehat{\mu}_0)^2\right) \end{array}$$

To compute the likelihood ratio:

$$\begin{array}{rcl} L(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}^2) & = & \frac{1}{(2\pi)^{(n_1+n_2)/2}\widehat{\sigma}^{n_1+n_2}}\exp\left\{-\frac{1}{2\widehat{\sigma}^2}\left(\sum_{j=1}^{n_1}(x_j-\widehat{\mu}_1)^2 + \sum_{j=1}^{n_2}(y_j-\widehat{\mu}_2)^2\right)\right\} \\ & = & \frac{1}{(2\pi)^{(n_1+n_2)/2}\widehat{\sigma}^{n_1+n_2}}\exp\left\{-\frac{(n_1+n_2)}{2}\right\} \end{array}$$

The last simplification comes from the formula for $\widehat{\sigma}^2$. Similarly, for the case $\overline{x} > \overline{y}$,

$$\begin{array}{rcl} L(\widehat{\mu}_{0,1}, \widehat{\mu}_{0,2}, \widehat{\sigma}_0^2) & = & \frac{1}{(2\pi)^{(n_1+n_2)/2}\widehat{\sigma}_0^{n_1+n_2}}\exp\left\{-\frac{1}{2\widehat{\sigma}_0^2}\left(\sum_{j=1}^{n_1}(x_j-\widehat{\mu}_0)^2 + \sum_{j=1}^{n_2}(y_j-\widehat{\mu}_0)^2\right)\right\} \\ & = & \frac{1}{(2\pi)^{(n_1+n_2)/2}\widehat{\sigma}_0^{n_1+n_2}}\exp\left\{-\frac{n_1+n_2}{2}\right\} \end{array}$$

using $\widehat{\mu}_{01} = \widehat{\mu}_{02} = \widehat{\mu}_0$.

The LRT is therefore:

$$\lambda(x,y) = \begin{cases} 1 & \overline{x} \leq \overline{y} \\ \left(\frac{\widehat{\sigma}}{\widetilde{\sigma}}\right)^{n_1+n_2} & \overline{x} > \overline{y} \end{cases}$$

Test: reject H_0 for $\lambda(x,y) < c$ where $c < 1$, (so a necessary condition for rejection is: $\overline{x} > \overline{y}$). To get it into the format required in the question, use:

$$\begin{aligned}(n_1+n_2)\widehat{\sigma}_0^2 &= \sum_{j=1}^{n_1}(X_j-\widehat{\mu}_0)^2+\sum_{j=1}^{n_2}(Y_j-\widehat{\mu}_0)^2\\ &= \sum_{j=1}^{n_1}(X_j-\overline{X})^2+n_1(\overline{X}-\widehat{\mu}_0)^2+\sum_{j=1}^{n_2}(Y_j-\widehat{\mu}_0)^2+n_2(\overline{Y}-\widehat{\mu}_0)^2\\ &= (n_1+n_2)\widehat{\sigma}^2+\frac{n_1n_2}{n_1+n_2}(\overline{X}-\overline{Y})^2\end{aligned}$$

so that

$$\widehat{\sigma}_0^2=\widehat{\sigma}^2+\frac{n_1n_2}{(n_1+n_2)^2}(\overline{X}-\overline{Y})^2.$$

Therefore:

$$\lambda(x,y)<c\Leftrightarrow\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2}>\frac{1}{c^{2/(n_1+n_2)}}\Leftrightarrow\left(1+\frac{n_1n_2}{(n_1+n_2)^2}\frac{(\overline{X}-\overline{Y})^2}{\widehat{\sigma}^2}\right)>\frac{1}{c^{2/(n_1+n_2)}}$$

Since $\widehat{\sigma}^2=\frac{n_1+n_2-2}{n_1+n_2}S^2$ also need $\overline{X}-\overline{Y}>0$ to reject H_0, this gives a test of reject H_0 if and only if

$$\frac{\overline{X}-\overline{Y}}{S}>k$$

for a suitable value of k, which depends on the significance level α. Since

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n_1+n_2-2}$$

it follows that $\mathbb{P}_{\mu_1,\mu_2}\left(\frac{\overline{X}-\overline{Y}}{S}>k\right)$ is increasing as $\mu_1-\mu_2$ increases and the result follows.

(c) The test is: Reject H_0 for $\frac{(\overline{x}-\overline{y})}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}>t_{n_1+n_2-2;\alpha}$, where $t_{n_1+n_2;\alpha}$ is the value such that $\mathbb{P}(T>t_{n_1+n_2-2;\alpha})=\alpha$.

Let $\theta=\mu_2-\mu_1$, then $\overline{X}-\overline{Y}\sim N\left(\theta,\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$ and hence

$$Z:=\frac{(\overline{X}-\overline{Y})-\theta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim N(0,1).$$

The *power* of the test is

$$\begin{aligned}\beta(\theta) &:= \mathbb{P}\left(\frac{(\overline{X}-\overline{Y})}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}>t_{n_1+n_2-2;\alpha}\middle|\,\mu_2-\mu_1=\theta\right)\\ &= \mathbb{P}\left(\frac{(\overline{X}-\overline{Y})-\theta}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}>t_{n_1+n_2-2;\alpha}-\frac{\theta}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\middle|\,\mu_2-\mu_1=\theta\right)\end{aligned}$$

For large n_1, n_2, $S \simeq \sigma$ (law of large numbers) and $t_{n_1+n_2;\alpha} \simeq z_\alpha$ where $\mathbb{P}(Z > z_\alpha) = \alpha$ for $Z \sim N(0,1)$, so

$$\beta(\theta) \simeq \mathbb{P}(Z \geq z_\alpha - \frac{\theta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}})$$

For the numbers given, $\alpha = 0.01$ and

$$0.95 = \beta(\frac{\sigma}{2}) \simeq 1 - \Phi(z_{0.01} - \frac{\sqrt{n}}{4})$$

Using $z_{0.01} = 2.33$ and $z_{0.05} = 1.64$, we have:

$$-1.64 = 2.33 - \frac{\sqrt{n}}{4} \Rightarrow n = 253$$

2. Likelihood equations obtained by: $\frac{\partial}{\partial \eta_i} \log L = 0$, $i = 1, \ldots, r$ and $\frac{\partial}{\partial \sigma} \log L = 0$. These give directly that the ML estimate has to satisfy:

$$\begin{cases} \widehat{\eta}_i = U_i & i = 1, \ldots, r \\ \frac{1}{\widehat{\sigma}^2} \sum_{j=r+1}^n U_j^2 = n \end{cases}$$

For $r = n$, $\widehat{\eta}_i = U_i$ so that the log likelihood evaluated at $\widehat{\eta}$ is:

$$\log L(\widehat{\eta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2)$$

which is maximised for $\sigma = 0$, which is not in the (open) parameter space $(0, +\infty)$, hence $\widehat{\sigma}_{ML}$ does not exist. Hence no solution for $n = r$.

For $n \geq r + 1$,

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{j=r+1}^n U_j^2.$$

Let $U = \binom{U^{(1)}}{U^{(2)}}$ where $U^{(1)} = (U_1, \ldots, U_r)^t$ and $U^{(2)} = (U_{r+1}, \ldots, U_n)^t$. Let $A = \binom{A^{(1)}}{A^{(2)}}$ where $A^{(1)}$ is $r \times n$ and $A^{(2)}$ is $n - r \times n$. Note that $\widehat{\mu} = A^{(1)t}U^{(1)}$ so that $Y - \widehat{\mu} = A^{(2)t}U^{(2)}$. It follows that

$$\sum_{j=r+1}^n U_j^2 = U^{(2)t}U^{(2)} = U^{(2)t}A^{(2)}A^{(2)t}U^{(2)} = |Y - \widehat{\mu}|^2.$$

3. The purpose of this question is to see all the abstract results for $Y = X\beta + \epsilon$ in the concrete setting of a single explanatory variable. Here the formulae are more transparent and we can see (for example) what happens when there is ill-conditioning in the X matrix.

   (a) The matrix X is:

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

and the parameter vector is:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

To get $(X'X)^{-1}$ (so that - for example - we can compute the covariance of the parameter vector estimator):

$$(X^tX) = \left(\begin{array}{cc} n & \sum_{j=1}^n x_j \\ \sum_{j=1}^n x_j & \sum_{j=1}^n x_j^2 \end{array} \right) = n \left(\begin{array}{cc} n & \overline{x} \\ \overline{x} & \overline{x^2} \end{array} \right)$$

Using the usual formula for inverting a 2 × 2 matrix together with the obvious identity:

$$\det(X'X) = n(\overline{x^2} - \overline{x}^2) = \sum_{j=1}^n (x_j - \overline{x})^2$$

gives:

$$(X^tX)^{-1} = \frac{1}{\sum_{j=1}^n (x_j - \overline{x})^2} \left(\begin{array}{cc} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{array} \right)$$

(b) The MLE is equal to the least squares estimator. From lectures,

$$\widehat{\beta} = (X^tX)^{-1}X^tY$$

Plugging in $(X^tX)^{-1}$ which has been computed gives:

$$\begin{aligned} \left(\begin{array}{c} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{array} \right) &= (X^tX)^{-1} \left(\begin{array}{c} n\overline{Y} \\ n\overline{xY} \end{array} \right) \\ &= \frac{1}{\frac{1}{n}\sum_{j=1}^n (x_j - \overline{x})^2} \left(\begin{array}{c} \overline{x^2}\,\overline{Y} - \overline{x}\,\overline{xY} \\ \overline{xY} - \overline{x}\,\overline{Y} \end{array} \right) \\ &= \left(\begin{array}{c} \overline{Y} - \overline{x}\frac{\sum_{j=1}^n (x_j - \overline{x})(Y_j - \overline{y})}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ \frac{\sum_{j=1}^n (x_j - \overline{x})(Y_j - \overline{y})}{\sum_{j=1}^n (x_j - \overline{x})^2} \end{array} \right). \end{aligned}$$

This gives the best fitting straight line in the least squares sense. Note that

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}.$$

(c) For the standard deviation estimate,

$$\frac{(n-2)S^2}{\sigma^2} = \frac{|Y - \widehat{\mu}|^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Note: $n - 2$ degrees of freedom is obtained from the previous exercise.

We may also see it directly: the argument goes as follows: $\widehat{Y} = X(X^tX)^{-1}X^tY$ so that the residuals are:

$$Y-\widehat{Y}=(I-X(X^tX)^{-1}X^t)Y=(I-H)\epsilon$$

where $H=X(X^tX)^{-1}X^t$ and $\epsilon\sim N(0,\sigma^2I)$. This is because $Y=X\beta+\epsilon$ and $HX=X$. Note that $H^2=H$ (straightforward computation). It therefore follows that all the eigenvalues are either 0 or 1. Therefore, since X is rank 2 it follows that H is of rank 2; 2 e-values are 1, the remaining are 0 and it is straightforward that that $I-H$ is rank $n-2$; the eigenvalues of matrix $I-H$ are $n-2$ 1's and 2 0'2. Let $D=\text{diag}(1,\dots,1,0,0)$ and let $I-H=PDP^t$ where P is orthonormal. Then

$$\sum(Y_i-\widehat{\beta}_0-x_i\widehat{\beta}_1)^2=(Y-\widehat{Y})^t(Y-\widehat{Y})=\epsilon^tPDP^t\epsilon=\sum_{j=1}^{n-2}\eta_j^2$$

where $\eta=P^t\epsilon$. Since P is orthonormal, it follows that $\eta\sim N(0,\sigma^2I)$.

Therefore, it follows that:

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

$$\underline{\widehat{\beta}}\sim N(\underline{\beta},(X^tX)^{-1}\sigma^2)$$

(d) Let $\underline{v}=(1,z)^t$ then

$$\mathbb{E}[Y(z)]=\underline{v}^t\underline{\beta}$$

$$\frac{\underline{v}^t\underline{\widehat{\beta}}-\underline{v}^t\underline{\beta}}{\sigma\sqrt{\underline{v}^t(X^tX)^{-1}\underline{v}}}\sim N(0,1)$$

$$\frac{\underline{v}^t\underline{\widehat{\beta}}-\underline{v}^t\underline{\beta}}{S\sqrt{\underline{v}^t(X^tX)^{-1}\underline{v}}}\sim t_{n-2}$$

with $1-\alpha$ confidence,

$$\underline{v}^t\underline{\beta}\in\left(\underline{v}^t\underline{\widehat{\beta}}\pm s\sqrt{\underline{v}^t(X^tX)^{-1}\underline{v}}t_{n-2;\alpha/2}\right)$$

and

$$\underline{v}^t(X^tX)^{-1}\underline{v}=\frac{\overline{x^2}-2z\overline{x}+z^2}{\sum_{j=1}^n(x_j-\overline{x})^2}=\frac{\frac{1}{n}\sum_{j=1}^n(x_j-\overline{x})^2+(\overline{x}-z)^2}{\sum_{j=1}^n(x_j-\overline{x})^2}$$

and the result follows.

(e) From the previous part,

$$\mathbb{E}[(Y^*-\widehat{Y}^*)^2]=\mathbf{V}(Y^*-\widehat{Y}^*)=\mathbf{V}(Y^*)+\mathbf{V}(\widehat{Y}^*)=\sigma^2\left(1+\frac{1}{n}+\frac{(\overline{x}-z)^2}{\sum_{j=1}^n(x_j-\overline{x})^2}\right)$$

while, under the assumption $\beta_1 = 0$,

$$\mathbb{E}[(Y^* - \overline{Y})^2] = \mathbf{V}(Y^*) + \mathbf{V}(\overline{Y}) = \sigma^2 \left(1 + \frac{1}{n}\right)$$

and the result is clear.

4. (a)

$$\overline{Y}_{j.} - \overline{Y}_{i.} \sim N(\beta_j - \beta_i, \sigma^2(\frac{1}{n_j} + \frac{1}{n_i}))$$

$$S^2 = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i.})^2 \qquad n-p \qquad d.f.$$

is the unbiased estimator of σ^2. Then

$$\frac{(\overline{Y}_{j.} - \overline{Y}_{i.}) - (\beta_j - \beta_i)}{S\sqrt{\frac{n_i+n_j}{n_i n_j}}} \sim t_{n-p}$$

and the confidence interval follows. The confidence interval for σ follows from:

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

hence the $1-\alpha$ confidence bound is given by:

$$k_{n-p;1-(\alpha/2)} \leq \frac{(n-p)s^2}{\sigma^2} \leq k_{n-p;(\alpha/2)}$$

from which the result follows.

(b)

$$\widehat{\psi} \sim N\left(\psi, \sigma^2 \left(\frac{1}{4n_2} + \frac{1}{4n_3} + \frac{1}{n_1}\right)\right)$$

the estimator of σ^2 is $S^2 = Q_{\text{res}} n - p$ given above with $n-p$ degrees of freedom and hence

$$\frac{1}{2}(\beta_2 + \beta_3) - \beta_1 \in \left(\frac{1}{2}\left(\overline{Y}_{2.} + \overline{Y}_{3.}\right) - \overline{Y}_{1.} \pm s t_{n-p,\alpha/2} \sqrt{\frac{n_1 n_3 + n_1 n_2 - 4n_2 n_3}{4n_1 n_2 n_3}}\right)$$

Similarly,

$$\mathbf{V}(\widehat{\psi}) = \frac{n_1 n_3 + n_1 n_2 + 4n_2 n_3}{4n_1 n_2 n_3}\sigma^2$$

hence the confidence interval is:

$$\frac{n_1 n_3 + n_1 n_2 + 4n_2 n_3}{4n_1 n_2 n_3} \frac{(n-p)s^2}{k_{n-p;(\alpha/2)}} \leq \mathbf{V}(\widehat{\psi}) \leq \frac{n_1 n_3 + n_1 n_2 + 4n_2 n_3}{4n_1 n_2 n_3} \frac{(n-p)s^2}{k_{n-p;1-(\alpha/2)}}$$

5. $x^t C^t C x = 0$ implies that $x^t C^t = 0$ which implies that $x = 0$ so that if $x \neq 0$ then $x^t C^t C x \neq 0$ hence $C^t C$ is (strictly) positive definite.

6. (a) Let $n = n_1 + \ldots + n_k$ denote the total number of experimental units. For $H_0 : \mu_1 = \ldots = \mu_k = \mu$, we have the maximiser $\widetilde{\mu} = \frac{1}{n}\sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}$ and

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^k \sum_{i=1}^{n_j}(Y_{ij} - \widetilde{\mu})^2$$

and the maximum likelihood under the constraint H_0 is: $\frac{1}{(2\pi)^{n/2}\widetilde{\sigma}^n}e^{-n/2}$.

For the unconstrained problem, the likelihood is maximised at $\widehat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$ and

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^k \sum_{i=1}^{n_j}(Y_{ij} - \widehat{\mu}_j)^2.$$

The maximum likelihood for the unconstrained problem is: $\frac{1}{(2\pi)^{n/2}\widehat{\sigma}^n}e^{-n/2}$ and hence the likelihood ratio statistic is:

$$\lambda(y) = \left(\frac{\widehat{\sigma}}{\widetilde{\sigma}}\right)^n.$$

(b) Pythagorean identity: note that $\overline{Y}_{.j} = \widehat{\mu}_j$ and $\overline{Y}_{..} = \widetilde{\mu}$ from previous part.

$$\sum_{j=1}^k \sum_{i=1}^{n_j}(Y_{ij} - \widetilde{\mu})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{.j} + \overline{Y}_{.j} - \overline{Y}_{..})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{.j})^2 + \sum_{j=1}^k n_j(\overline{Y}_{.j} - \overline{Y}_{..})^2$$

so:

$$n\widetilde{\sigma}^2 = Q_{\text{res}} + Q_M \qquad n\widehat{\sigma}^2 = Q_{\text{res}}.$$

Therefore, the likelihood ratio test is:

$$\lambda(y) < c \Leftrightarrow \frac{Q_{res}}{Q_M + Q_{res}} < c^{2/n} \Leftrightarrow \frac{Q_M/(k-1)}{Q_{\text{res}}/(n-k)} > \left(\frac{n-k}{k-1}\right)\left(\frac{1-c^{2/n}}{c^{2/n}}\right) = k$$

establishing the result.

(c) It follows from the canonical representation (lectures) that $Q_M \perp Q_{\text{res}}$. Under $H_0 : \mu_1 = \ldots = \mu_k$, it follows that $\frac{Q_M}{\sigma^2} \sim \chi^2_{k-1}$ since the parameter space for $\mu_1, \ldots, \mu_k$ is k-dimensional and the parameter space for the mean under the null hypothesis is 1-dimensional, and $\frac{Q_{\text{res}}}{\sigma^2} \sim \chi^2_{n-k}$. The result follows from Proposition 11.4.

Chapter 11

Gaussian Linear Models (II)

11.1 Tests and Confidence Intervals

Consider a situation where two models are to be tested; model I with a parameter vector $\beta^{(1)}$ of length p and model II with a parameter vector $\beta = \begin{pmatrix} \beta^{(1)} \\ \hline \beta^{(2)} \end{pmatrix}$ where $\beta^{(2)}$ is a parameter vector of length $r-p$, where $r > p$. Suppose that model I may be written as:

$$Y = X^{(1)}\beta^{(1)} + \epsilon$$

where $X^{(1)}$ is a $n \times p$ matrix and Model II may be written as

$$Y = X\beta + \epsilon$$

where $X = (X^{(1)}|X^{(2)})$, $X^{(2)}$ is $n \times r - p$.

One very important issue in the study of linear models is often that of comparing such nested models and considering whether the mean vector $\mu = \mathbb{E}[Y]$ belongs to a linear subspace $\mathcal{S}_0 = \{X^{(1)}\beta^{(1)} : \beta^{(1)} \in \mathbb{R}^p\}$ of $\mathcal{S} = \{X\beta : \beta \in \mathbb{R}^r\}$.

Example 11.1 (A Multiple Linear Regression Problem).

Consider the regression problem:

$$Y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \epsilon_j \qquad j = 1, \ldots, n$$

Note that

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}$$

Here $p = 3$; there are 3 parameters and 2 explanatory (or regressor) variables. With multiple linear regression, there are $q = p - 1$ regressor variables with associated parameters $(\beta_1, \ldots, \beta_q)$. When finding the best fitting linear model, there is always, in addition, the intercept β_0.

Consider, for example, a medical experiment, where x_{i1} denotes the age of patient i and x_{i2} denote the dose level of a drug administered to patient i. The response Y_i denotes the reduction in blood pressure for patient i.

In order to estimate all three parameters β_0, β_1 and β_2 effectively, it is important that x_{i1} and x_{i2} are chosen so that the matrix X is of rank 3. Consider testing $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. Under H_0, $\{\mu : \mu = \beta_0 + \beta_1 x_{.1}\}$ is a two dimensional linear subspace of the three dimensional linear space defined by the full model. □

One method of testing is the Likelihood Ratio Test. Let $\mathcal{S}$ denote the space for μ in the full model and $\mathcal{S}_0$ the space for μ in the reduced model. Then the likelihood ratio is:

$$\lambda(y) = \frac{\sup_{\mu \in \mathcal{S}_0} L(\mu, y)}{\sup_{\mu \in \mathcal{S}} L(\mu, y)}$$

and for a vector of observations y, reject the null hypothesis $H_0 : \mu \in \mathcal{S}_0$ for small values of $\lambda(y)$.

σ assumed to be the same for both models Note that

$$\lambda(y) = \exp\left\{ -\frac{1}{2\sigma^2} \left\{ |y - \widehat{\mu}_0|^2 - |y - \widehat{\mu}|^2 \right\} \right\}$$

where $\widehat{\mu}_0$ and $\widehat{\mu}$ are the maximum likelihood estimates for the reduced and full models respectively.

As before, let V be an $n \times n$ orthonormal matrix which may be written as $V = \begin{pmatrix} V^{(1)} \\ \hline V^{(2)} \\ \hline V^{(3)} \end{pmatrix}$; $V^{(1)}$ is a $p \times n$ matrix whose rows span $\mathcal{S}_0$, $V^{(2)}$ is a $r - p \times n$ matrix such that the rows of $\begin{pmatrix} V^{(1)} \\ \hline V^{(2)} \end{pmatrix}$ span $\mathcal{S}$. Set

$$U = VY \qquad \eta = V\mu,$$

then $Y = V^t U$ and $\mu = V^t \eta$.

Let $U = \begin{pmatrix} U^{(1)} \\ \hline U^{(2)} \\ \hline U^{(3)} \end{pmatrix}$ where $U^{(1)}$ is of length p, $U^{(2)}$ is of length $r - p$ and $U^{(3)}$ is of length $n - r$. From Theorem 10.2,

$$\widehat{\mu}_0 = V^{(1)t} U^{(1)} \qquad \widehat{\mu} = (V^{(1)t} | V^{(2)t}) \begin{pmatrix} U^{(1)} \\ \hline U^{(2)} \end{pmatrix}.$$

From this,

$$\begin{aligned} |Y-\widehat{\mu}_0|^2-|Y-\widehat{\mu}|^2 &= |V^tU-V^{(1)t}U^{(1)}|^2-|V^tU-(V^{(1)t}|V^{(2)t})\left(\frac{U^{(1)}}{U^{(2)}}\right)|^2 \\ &= (U^{(2)t}|U^{(3)t})\left(\frac{V^{(2)}}{V^{(3)}}\right)(V^{(2)t}|V^{(3)t})\left(\frac{U^{(2)}}{U^{(3)}}\right)-U^{(3)t}V^{(3)}V^{(3)t}U^{(3)} \\ &= U^{(2)t}V^{(2)}V^{(2)t}U^{(2)}=U^{(2)t}I_{r-p}U^{(2)}=\sum_{i=p+1}^{r}U_j^2. \end{aligned}$$

Therefore:

$$-2\log\lambda(Y)=\frac{1}{\sigma^2}\sum_{j=p+1}^{r}U_j^2.$$

Under the null hypothesis that $\mu\in\mathcal{S}_0$, $U_i\sim N(0,\sigma^2)$ for $i=p+1,\ldots,n$ so that

$$W:=-2\log\lambda(Y)\sim\chi^2_{r-p}.$$

The null hypothesis is *rejected* for large values of W; for a test of significance α, reject H_0 for $W>k_{r-p,\alpha}$ where $k_{r-p;\alpha}$ denotes the value such that $\mathbb{P}(W>k_{r-p;\alpha})=\alpha$.
To compute the power of the test, the *non-central* χ^2 *distribution* is needed.

Note that $U_i\sim N(\eta_i,\sigma^2)$. Let $Z_i=\frac{U_i}{\sigma}$ and $\theta_i=\frac{\eta_i}{\sigma}$, then $Z_i\sim N(\theta_i,1)$. If H_0 is not true, then some of the values $\eta_{p+1},\ldots,\eta_r$ are non-zero.

Lemma 11.1. *Let $Z_i\sim N(\theta_i,1)$ for $i=1,\ldots,n$ be independent. Let $V=Z_1^2+\ldots+Z_n^2$. Then the distribution of V has moment generating function:*

$$M_V(\nu)=\frac{1}{(1-2\nu)^{n/2}}\exp\left\{\frac{\nu}{1-2\nu}\sum_{j=1}^{n}\theta_j^2\right\}.$$

Proof Basic calculus. □

Definition 11.2 (Non central χ^2 distribution)**.** *The distribution of V in the previous lemma is the* non central χ^2 distribution with noncentrality parameter $\lambda=\sum_{j=1}^n\theta_j^2$.

It follows that for the Gaussian linear model, $-2\log\lambda(Y)\sim\chi^2_{r-p}(\lambda)$ where

$$\lambda=\frac{1}{\sigma^2}\sum_{i=p+1}^{r}\eta_i^2=\frac{1}{\sigma^2}|\mu-\mu_0|^2$$

where μ_0 is the projection of μ onto $\mathcal{S}_0$.

11.1.1 Comparison of models, unknown sigma

Consider the same problem, choosing between

$$\begin{array}{ll} I & Y = X^{(1)}\beta^{(1)} + \epsilon_1 \\ II & Y = (X^{(1)}|X^{(2)}) \left(\begin{array}{c} \beta^{(1)} \\ \beta^{(2)} \end{array} \right) + \epsilon_2 \end{array}$$

where $\beta^{(1)}$ is a p-vector of parameters and $\beta = \left(\frac{\beta^{(1)}}{\beta^{(2)}} \right)$ and $\beta^{(2)}$ is a $r-p$ vector. Here no assumptions are placed on σ. The likelihood ratio test statistic is:

$$\lambda(Y) = \frac{\sup_{\mu \in \mathcal{S}_0, \sigma^2 \in \mathbb{R}_+} L(\mu, \sigma; Y)}{\sup_{\mu \in \mathcal{S}, \sigma^2 \in \mathbb{R}_+} L(\mu, \sigma; Y)}.$$

The esimates for the mean vectors are as before;

$$\widehat{\mu}_0 = V^{(1)t}U^{(1)} \qquad \widehat{\mu} = (V^{(1)t}|V^{(2)t}) \left(\frac{U^{(1)}}{U^{(2)}} \right).$$

where $\widehat{\mu}_0$ and $\widehat{\mu}$ are the estimators for Models I and II respectively.

To compute the likelihood ratio test statistic, the ML estimators for σ^2 for Models I and II are required. Let $\widehat{\sigma}_0^2$ and $\widehat{\sigma}^2$ be the ML estimators for Models I and II respectively, then

$$\widehat{\sigma}^2 = \frac{1}{n}|Y - \widehat{\mu}|^2 \qquad \widehat{\sigma}_0^2 = \frac{1}{n}|Y - \widehat{\mu}_0|^2.$$

Note:

$$L(\widehat{\mu}_0, \widehat{\sigma}_0; y) = \frac{1}{(2\pi)^{n/2}\widehat{\sigma}_0^n} \exp\left\{ -\frac{1}{2\widehat{\sigma}_0^2}|y - \widehat{\mu}_0|^2 \right\} = \frac{1}{(2\pi)^{n/2}\widehat{\sigma}_0^n} \exp\left\{ -\frac{n}{2} \right\}$$

and

$$L(\widehat{\mu}, \widehat{\sigma}; y) = \frac{1}{(2\pi)^{n/2}\widehat{\sigma}^n} \exp\left\{ -\frac{1}{2\widehat{\sigma}^2}|y - \widehat{\mu}|^2 \right\} = \frac{1}{(2\pi)^{n/2}\widehat{\sigma}^n} \exp\left\{ -\frac{n}{2} \right\}$$

so that the likelihood ratio statistic is:

$$\lambda(y) = \frac{L(\widehat{\mu}_0, \widehat{\sigma}_0; y)}{L(\widehat{\mu}, \widehat{\sigma}; y)} = \left(\frac{|y - \widehat{\mu}|^2}{|y - \widehat{\mu}_0|^2} \right)^{n/2}.$$

Recall that

$$\frac{1}{\sigma^2}|\widehat{\mu} - \widehat{\mu}_0|^2 = \frac{1}{\sigma^2} \sum_{j=p+1}^{r} U_j^2.$$

Using $|Y - \widehat{\mu}|^2 = \sum_{j=r+1}^{n} U_j^2$, we have the Pythagorean identity:

$$|Y-\widehat{\mu}_0|^2 = \sum_{j=p+1}^{n} U_j^2 = \sum_{j=p+1}^{r} U_j^2 + \sum_{j=r+1}^{n} U_j^2 = |\widehat{\mu}-\widehat{\mu}_0|^2 + |Y-\widehat{\mu}|^2.$$

Therefore:

$$\lambda(Y) = \left(\frac{1}{1+\left(\frac{|\widehat{\mu}-\widehat{\mu}_0|}{|Y-\widehat{\mu}|}\right)^2}\right)^{n/2}$$

and hence, for $0 < k < 1$,

$$\lambda(Y) \leq k \Leftrightarrow \frac{|\widehat{\mu}-\widehat{\mu}_0|^2}{|Y-\widehat{\mu}|} \geq c$$

where $c = \frac{1}{k^{2/n}} - 1$, so rejecting H_0 : Model I correct for $\frac{|\widehat{\mu}-\widehat{\mu}_0|^2}{|Y-\widehat{\mu}|} \geq c$ is equivalent to the Likelihood Ratio Test.

If Model 1 is correct, then $U_{p+1}, \ldots, U_r$ are i.i.d. $N(0, \sigma^2)$ so that

$$\frac{1}{\sigma^2}|\widehat{\mu}-\widehat{\mu}_0|^2 \sim \chi^2_{r-p}.$$

Recall also that $\frac{1}{\sigma^2}|Y-\widehat{\mu}|^2 \sim \chi^2_{n-r}$ and

$$\frac{1}{\sigma^2}|\widehat{\mu}-\widehat{\mu}_0|^2 \perp \frac{1}{\sigma^2}|Y-\widehat{\mu}|^2.$$

Recall the definition of an F distribution. If Model I is correct,

$$F := \frac{|\widehat{\mu}-\widehat{\mu}_0|^2/(r-p)}{|Y-\widehat{\mu}|^2/(n-r)} \sim F_{r-p,n-r}.$$

For a test with significance α, H_0, that Model I is correct, is *rejected* for $F > f_{r-p,n-r;\alpha}$ where $f_{r-p,n-r;\alpha}$ is the value such that

$$\mathbb{P}(F > f_{r-p,n-r;\alpha}) = \alpha \qquad F \sim F_{r-p,n-r}.$$

The Power of the Test To compute the *power* of the test, it is necessary to know the distribution of the test statistic when $\mu \neq \mu_0$. This requires the *non-central F distribution.*

We have already shown that

$$\frac{1}{\sigma^2}\left|\widehat{\mu}-\widehat{\mu}_0\right|^2 \sim \chi^2_{r-p}(\theta^2) \qquad \text{where} \qquad \theta^2 = \frac{1}{\sigma^2}|\mu-\mu_0|^2.$$

The canonical representation is:

$$\frac{1}{\sigma^2}\left|\widehat{\mu}-\widehat{\mu}_0\right|^2 = \frac{1}{\sigma^2}\sum_{i=q+1}^{r} U_i^2$$

Furthermore,

$$\frac{1}{\sigma^2}\,|Y-\widehat{\mu}|^2=\frac{1}{\sigma^2}\sum_{i=r+1}^{n}U_i^2\sim\chi^2_{n-r}$$

and $\frac{1}{\sigma^2}\,|Y-\widehat{\mu}|^2 \perp \frac{1}{\sigma^2}\,|\widehat{\mu}-\widehat{\mu}_0|^2$.

Definition 11.3. *Let $V\sim\chi^2_m(\lambda)$ and $W\sim\chi^2_n(0)$, where $V\perp W$. The non-central F distribution with noncentrality parameter λ and m and n degrees of freedom is defined as:*

$$F:=\frac{V/m}{W/n}\sim F_{m,n}(\lambda).$$

In general,

$$F:=\frac{|\widehat{\mu}-\widehat{\mu}_0|^2/(r-p)}{|Y-\widehat{\mu}|^2/(n-r)}\sim F_{r-p,n-r}(\theta^2). \tag{11.1}$$

If $\frac{|\widehat{\mu}-\widehat{\mu}_0|^2}{\sigma^2}=\theta^2$, then the test of size α has power

$$\beta(\theta)=\mathbb{P}(F>f_{r-p,n-r;\alpha})$$

where $F\sim F_{r-p,n-r}(\theta^2)$ and $f_{r-p,n-r;\alpha}$ is the value such that $\mathbb{P}(F_0>f_{r-p,n-r;\alpha})=\alpha$ for $F_0\sim F_{r-p,n-r}(\theta^2)$.

These results are summarised in the following proposition.

Proposition 11.4. *Consider the hypothesis test of $H_0:\mu\in\mathcal{S}_0$ versus $H_1:\mu\notin\mathcal{S}_0$. If H_0 is true, then T defined by*

$$T:=\frac{|\widehat{\mu}-\widehat{\mu}_0|/(r-p)}{|Y-\widehat{\mu}|^2/(n-r)}$$

satisfies:

$$T\sim F_{r-p,n-r}(0)$$

That is, the central F distribution. Let f_α denote the value such that $\mathbb{P}(T>f_\alpha)=\alpha$, then for a size α test, the null hypothesis is rejected for $t>f_\alpha$, where t is the observed value of T.

For $\frac{|\mu-\mu_0|^2}{\sigma^2}=\theta^2$, the power of the test is:

$$\beta(\theta)=\mathbb{P}(F>f_\alpha)$$

where $F\sim F_{r-p,n-r}(\theta^2)$.

Proof The analysis is given in the preceding discussion. □

If the reduced model is true, then $U_i\sim N(0,\sigma^2)$ for $i=p+1,\ldots,r$. It therefore follows that $\frac{1}{r-p}|\widehat{\mu}-\widehat{\mu}_0|^2$ is an unbiased estimator of σ^2 under H_0. Whether H_0 or H_1 holds, $\frac{1}{n-r}|Y-\hat{\mu}|^2$ is an unbiased estimator of σ^2. The statistic T from equation (11.1) is, under the null hypothesis, simply the ratio of these two unbiased estimators of σ^2. If H_0 does not hold, then $\frac{1}{r-p}|\widehat{\mu}-\widehat{\mu}_0|^2$ also contains a systematic component, given by the non centrality parameter θ^2.

Example 11.2 (Regression)**.**

Consider comparison of the two models for an observable Y:

$$\begin{cases} I : Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon \\ II : Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \beta_{p+1} x_{p+1} + \ldots + \beta_{p+q} x_{p+q} + \epsilon \end{cases}$$

where $\epsilon \sim N(0, \sigma^2)$ (σ^2 unknown) and errors ϵ are independent for different observations. Let $\beta = \binom{\beta^{(1)}}{\beta^{(2)}}$ where $\beta^{(1)} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ and $\beta^{(2)} = (\beta_{p+1}, \ldots, \beta_{p+q})^t$. Consider a vector $\underline{Y}$ of n observations. The full model (model II) may be written as:

$$\underline{Y} = X_1 \beta^{(1)} + X_2 \beta^{(2)} + \underline{\epsilon}$$

where $\underline{\epsilon} \sim N(\underline{0}, \sigma^2 I_n)$, the jth row of X_1 is: $(1, x_{j1}, \ldots, x_{j,p})$ and the jth row of X_2: $(x_{j,p+1}, \ldots, x_{j,p+q})$.

Assume that the matrix $X = (X_1 | X_2)$ is of full rank and consider the test: $H_0 : \beta^{(2)} = 0$ versus $H_1 : \beta^{(2)} \neq 0$. The parameter estimators for the full and reduced models respectively are:

$$\widehat{\beta} = (X^t X)^{-1} X^t \underline{Y} \qquad \widehat{\beta}_0^{(1)} = (X_1^t X_1)^{-1} X_1^t \underline{Y}$$

The residual sums of squares for the full and reduced models respectively are:

$$Q_{\text{res},II} = |\underline{Y} - \widehat{\mu}|^2 = |(I - X(X^t X)^{-1} X^t)\underline{Y}|^2 = \underline{Y}^t (I - X(X^t X)^{-1} X^t)\underline{Y}$$

$$Q_{\text{res},I} = |\underline{Y} - \widehat{\mu}_0|^2 = \underline{Y}^t (I - X_1 (X_1^t X_1)^{-1} X_1^t)\underline{Y}$$

The statistic is:

$$F = \frac{(Q_{\text{res},I} - Q_{\text{res},2})/q}{Q_{\text{res},2}/(n - (p + q + 1))}$$

which has $F_{q,n-(p+q+1)}(\theta^2)$ distribution, with noncentrality parameter

$$\theta^2 = \frac{1}{\sigma^2} \beta^{(2)t} (X_2^t X_2 - X_2^t X_1 (X_1^t X_1)^{-1} X_1^t X_2) \beta^{(2)}.$$

This is left as an exercise.

In the special case that $X_1^t X_2 = 0$, θ^2 simplifies to $\frac{1}{\sigma^2} \beta^{(2)t} (X_2^t X_2) \beta^{(2)}$, which only depends on the second set of variables and coefficients. □

11.2 Analysis of Variance

Recall the Pythagorean identity: let $X = (X^{(0)} | X^{(1)})$ where $X^{(0)}$ is $n \times p$ and $X^{(1)}$ is $n \times r - p$ where $r - p \geq 1$. Let $\beta = \binom{\beta^{(0)}}{\beta^{(1)}}$ where $\beta^{(0)}$ is a p vector and $\beta^{(1)}$ is an $r - p$ vector. Let $\mathcal{S}_0 = \{X^{(0)} \beta^{(0)} : \beta^{(0)} \in$

$\mathbb{R}^p\}$ and $\mathcal{S}_1 = \{X\beta : \beta \in \mathbb{R}^r\}$ where $r > p$. Let $\widehat{\mu}_0 = \arg\min_{\mu \in \mathcal{S}_0} |Y - \mu|$ and $\widehat{\mu} = \arg\min_{\mu \in \mathcal{S}} |Y - \mu|$. Then

$$|\underline{Y} - \underline{\widehat{\mu}}_0|^2 = |\underline{\widehat{\mu}} - \underline{\widehat{\mu}}_0|^2 + |\underline{Y} - \underline{\widehat{\mu}}|^2.$$

Now consider the special case where $X^{(0)} = \mathbf{1}_n$, the n-vector with each entry 1. Then $\mathcal{S}_0 = \mathbb{R}$ and $\underline{\widehat{\mu}}_0 = \overline{Y}\mathbf{1}_n$ and

$$|\underline{Y} - \overline{Y}\mathbf{1}_n|^2 = \sum_{j=1}^{n} (Y_j - \overline{Y})^2 = Q_T.$$

This is referred to as the *total sum of squares.* If the model

$$Y_i = \mu + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \qquad i.i.d.$$

then $\overline{Y}$ is the unbiased estimator of μ and $\frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$ is the unbiased estimator of σ^2; $\frac{\underline{Y}^t M_n \underline{Y}}{\sigma^2} \sim \chi^2_{n-1}$, where $M_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$. Here I_n denotes the $n \times n$ identity matrix and $\mathbf{1}_n$ denotes a column vector, length n, with each entry 1.

Now consider the model

$$\underline{Y} = X\beta + \underline{\epsilon} \qquad \underline{\epsilon} \sim N(\underline{0}, \sigma^2 I_n),$$

where $X = (X^{(0)}|X^{(1)})$, $X^{(0)} = \mathbf{1}_n$, $\beta = \binom{\beta_0}{\beta^{(1)}}$ where $\beta_0 \in \mathbb{R}$ and $\beta^{(1)} \in \mathbb{R}^q$ where $q = r - 1$. There are q explanatory variables, whose values for observation Y_i are given by $(X^{(1)}_{i1}, \ldots, X^{(1)}_{iq})$ and the parameters $\beta^{(1)}$ indicate the dependence on these variables. Assume that the columns of $X^{(1)}$ are linearly independent of $\mathbf{1}_n$. Analysis of Variance summarises the sums of squares, indicating whether or not the explanatory variables are significant. Let $s = \dim(\mathcal{S}) - \dim(\mathcal{S}_0)$. This is equal to q if X is of full rank. Set

$$Q_M = |\widehat{\mu} - \widehat{\mu}_0|^2 \Rightarrow \frac{Q_M}{\sigma^2} \sim \chi^2_s\left(\frac{|\mu - \mu_0|^2}{\sigma^2}\right).$$

Here M stands for ‘model’. Finally,

$$Q_{\text{res}} = |Y - \widehat{\mu}|^2 \qquad \Rightarrow \qquad \frac{Q_{\text{res}}}{n - (s+1)} \sim \chi^2_{n-(s+1)}(0).$$

To test $H_0 : \beta^{(1)} = 0$ versus $H_1 : \beta^{(1)} \neq 0$, the results may be summarised in an *analysis of variance* table:

	sum of squares	d.f.	mean square	$F -$ value
model	Q_{Mod}	s	$M_{\text{Mod}} = \frac{Q_{\text{Mod}}}{s}$	$\frac{M_{\text{Mod}}}{M_{\text{res}}}$
residual	Q_{res}	$n - (s+1)$	$M_{\text{res}} = \frac{Q_{\text{res}}}{n-(s+1)}$	
total	Q_T	$n - 1$		

Example 11.3 (One Way Layout)**.**

Consider the model

$$Y_{ij} = \beta_i + \epsilon_{ij} \qquad i = 1, \ldots, p \qquad j = 1, \ldots, n_i, \qquad \epsilon_{ij} \sim N(0, \sigma^2) \qquad \text{i.i.d.}$$

Let $n = n_1 + \ldots + n_p$ denote the total number of observations. Let $\overline{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ and let $\overline{Y}_{..} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$. Then $\widehat{\beta}_i = \overline{Y}_{i.}$ for $i = 1, \ldots, p$. Consider the test: $H_0 : \beta_1 = \ldots = \beta_p$ versus H_1 : (not H_0). Under H_0, all observations have the same mean, so

$$\widehat{\mu}_0 = (\overline{Y}_{..}, \ldots, \overline{Y}_{..})^t.$$

It follows that:

$$|\widehat{\mu} - \widehat{\mu}_0|^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i} |\overline{Y}_{i.} - \overline{Y}_{..}|^2 = \sum_{i=1}^{p} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2.$$

At the same time,

$$|Y - \widehat{\mu}|^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$$

so that

$$T = \frac{n-p}{p-1}\frac{\sum_{i=1}^{p} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2}.$$

T has a $F_{p-1,n-p}(\lambda)$ distribution where the noncentrality parameter is:

$$\lambda = \frac{1}{\sigma^2}\sum_{i=1}^{p}(\beta_i - \overline{\beta})^2 \qquad \overline{\beta} = \frac{1}{n}\sum_{i=1}^{p} n_i\beta_i.$$

The sum of squares identities may be written:

$$Q_B = \sum_{i=1}^{p} n_k(\overline{Y}_{i.} - \overline{Y}_{..})^2 \qquad \text{between treatments}$$

$$Q_W = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 \qquad \text{within treatments}$$

$$Q_T = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 \qquad \text{total}$$

The *Pythagorean identity*:

$$|Y - \widehat{\mu}_0|^2 = |\widehat{\mu} - \widehat{\mu}_0|^2 + |Y - \widehat{\mu}|^2$$

may be written:

$$Q_T = Q_B + Q_W$$

Under H_0, $\frac{Q_B}{\sigma^2} \sim \chi^2_{p-1}$ and $\frac{Q_W}{\sigma^2} \sim \chi^2_{n-p}$. These random variables are independent. This information is summarised in what is known as an *analysis of variance* (ANOVA) table:

	sum of squares	d.f.	mean square	$F-$ value
between samples	Q_B	$p-1$	$M_B = \frac{Q_B}{p-1}$	$\frac{M_B}{M_W}$
within samples	Q_W	$n-p$	$M_W = \frac{Q_W}{n-p}$	
total	Q_T	$n-1$		

Tutorial 12

1. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. All parameters are assumed unknown. Let

$$R = \frac{\sum_{j=1}^{n_2}(Y_j - \overline{Y})^2}{\sum_{j=1}^{n_1}(X_j - \overline{X})^2}$$

and $F = \frac{(n_1-1)}{(n_2-1)}R$.

(a) Show that $\frac{\sigma_1^2}{\sigma_2^2}F$ has an F_{n_2-1,n_1-1} distribution.

(b) Compute the LR test of $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$ and show that it satisfies: reject H_0 for $F > x_1$ or $F < x_2$ where (x_1, x_2) satisfy:

$$\begin{cases} F_{n_2-1,n_1-1}(x_2) - F_{n_2-1,n_1-1}(x_1) = \alpha \\ \frac{x_1}{(1+\frac{n_2-1}{n_1-1}x_1)^{1+(n_1/n_2)}} = \frac{x_2}{(1+\frac{n_2-1}{n_1-1}x_2)^{1+(n_1/n_2)}}. \end{cases}$$

Here $F_{v,w}(x) = \mathbb{P}(X \leq x)$ for $X \sim F_{v,w}$.

(c) Can you show that the LR test with significance α is asymptotically equivalent to: reject H_0 for $F > F_{n_2-1,n_1-1;\alpha/2}$ or $F > \frac{1}{F_{n_1-1,n_2-1;\alpha/2}}$?

2. Consider the regression problem

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon}$$

where $\underline{Y}$ is an n vector, X is an $n \times (p+q+1)$ matrix, $\underline{\beta} = \binom{\beta^{(1)}}{\beta^{(2)}}$, $\beta^{(1)}$ is a $p+1$ vector and $\beta^{(2)}$ is a q vector. Let X_1 be the matrix with the first $p+1$ columns of X and X_2 the matrix with the remaining q columns. Consider the hypothesis test $H_0 : \beta^{(2)} = 0$ versus $H_1 : \beta^{(2)} \neq 0$. Suppose that X has full rank.

(a) Let $\underline{\widehat{\mu}}$ denote the ML estimator of $X\underline{\beta}$ for the full model and let $\underline{\widehat{\mu}}_0$ the estimator of $X_1\underline{\beta}^{(1)}$ under the null hypothesis. Show that

$$\mathbb{E}[\widehat{\mu} - \widehat{\mu}_0] = (I - X_1(X_1^t X_1)^{-1}X_1^t)X_2\beta^{(2)}.$$

(b) Let

$$F = \frac{(Q_{\text{res},I} - Q_{\text{res},II})/q}{Q_{\text{res},II}/(n-(p+q+1))}.$$

Show that this has $F_{q,n-(p+q-1)}(\theta^2)$ distribution, where the non-centrality parameter θ^2 is:

$$\theta^2 = \frac{1}{\sigma^2}\beta^{(2)t}(X_2^t X_2 - X_2^t X_1(X_1^t X_1)^{-1}X_1^t X_2)\beta^{(2)}.$$

3. Consider the one-way layout model

$$Y_{ij} = \alpha + \beta_i + \epsilon_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n_i$$

where ϵ_{ij} are i.i.d. $N(0, \sigma^2)$ and $\sum_{i=1}^{p} n_i \beta_i = 0$. Let $n = n_1 + \ldots + n_p$.

   (a) Find the MLE $(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)^t$ of the parameter vector $(\alpha, \beta_1, \ldots, \beta_p)$.

   (b) Compute the covariance matrix for $(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)^t$.

   (c) Give symmetric confidence intervals for α and β_k.

4. Consider again the one-way layout model of the previous exercise. Consider the two models:

$$\begin{cases} I & Y_{ij} = \alpha + \epsilon_{ij} \\ II & Y_{ij} = \alpha + \beta_i + \epsilon_{ij} \end{cases}$$

where Model II is the full model and Model I is the reduced model. Let $Q_{\text{res},I}$ and $Q_{\text{res},II}$ be the residual sums of squares of the two models. Show that

$$\frac{(Q_{\text{res},I} - Q_{\text{res},II})/(p-1)}{Q_{\text{res},II}/(n-p)} \sim F_{p-1,n-p}(\delta^2)$$

where the non-centrality parameter is:

$$\delta^2 = \frac{1}{\sigma^2} \sum_{k=1}^{p} n_k \beta_k^2.$$

5. Let $X = (X_1 | X_2)$ where X_1 is $n \times p$, X_2 is $n \times q$, X is $n \times p + q$ and $X^t X$ is invertible. Show that

$$X(X^t X)^{-1} X^t X_1 = X_1.$$

6. Consider the linear model $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$. Let $\widehat{Y} = X\widehat{\beta}$ denote the fitted values, where $\widehat{\beta}$ is the least squares estimator of β. Assume that $X_{.1} = \mathbf{1}_n$ (the n-vector with each entry 1). Show that

   (a) $\text{Var}(Y) = \text{Var}(\widehat{Y}) + \text{Var}(Y - \widehat{Y})$.

   (b) $\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$.

7. Consider the linear model $Y = X\beta + \epsilon$ where the first column of X is a column of 1s. (This corresponds to multiple linear regression). Suppose that $\epsilon \sim N(0, \sigma^2 I)$. Define the *hat matrix* H as $H = X(X^t X)^{-1} X^t$. β is a p-vector of parameters. Show that:

   (a) $\frac{1}{n} \leq H_{ii} \leq 1$ for all $i = 1, \ldots, p$,

   (b) $\text{tr}(H) = p$,

(c) $H_{ii} = \mathrm{Cor}(Y_i, \widehat{Y}_i)^2$.

You may use the fact that if $X = (X^{(1)}|X^{(2)})$, $H^{(1)} = X^{(1)}(X^{(1)\prime}X^{(1)})^{-1}X^{(1)\prime}$ and $H = X(X'X)^{-1}X'$ then $HH^{(1)} = H^{(1)}H = H^{(1)}$.

8. Consider again the regression model

$$Y = X\beta + \epsilon$$

where all elements of the first column of X are 1 and $\epsilon \sim N(0, \sigma^2 I)$. Define

$$R^2 = 1 - \frac{Q\mathrm{res}}{Q_T}$$

where $\widehat{Y}_j$ are the fitted values, $\overline{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$, $Q\mathrm{res} = \sum_{j=1}^n (Y_j - \widehat{Y}_j)^2$ (the residual sum of squares) and $Q_T = \sum_{j=1}^n (Y_j - \overline{Y})^2$ (the total sum of squares).

(a) Show that

$$R^2 = \left(\frac{\sum_{i=1}^n (Y_i - \overline{Y})(\widehat{Y}_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2 \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}} \right)^2 .$$

(b) Show that the test with critical region $R^2 > c$ is equivalent to the LRT test for testing the null model (where only β_0 is non-zero) against the full model (where all coefficients are non-zero).

(c) Show that R^2 is distributed according to a Beta $\left(\frac{p-1}{2}, \frac{n-p}{2}\right)$ distribution.

9. Let $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I_n)$, X is $n \times p$ of full rank, $p < n$ and let $\widehat{Y} = X(X^tX)^{-1}X^tY$, the projection onto $\mathcal{S} = \{\mu : \mu = X\beta \qquad \beta \in \mathbb{R}^p\}$. Let $H = X(X^tX)^{-1}X^t$, the projection matrix. Let Y^* be independent and indentically distributed with Y. Show that:

$$\mathbb{E}\left[\left|Y^* - \widehat{Y}\right|^2\right] = \mathbb{E}\left[\left|Y - \widehat{Y}\right|^2\right] + 2\sigma^2 \mathrm{tr}(H).$$

Answers

1. (a)

$$W := \frac{\sum_{j=1}^{n_2}(Y_j - \overline{Y})^2}{\sigma_2^2} \sim \chi^2_{n_2-1}, \qquad V := \frac{\sum_{j=1}^{n_1}(X_j - \overline{X})^2}{\sigma_1^2} \sim \chi^2_{n_1-1}, \qquad V \perp W.$$

From the definition of an F distribution,

$$G := \frac{W/(n_2-1)}{V/(n_1-1)} \sim F_{n_2-1,n_1-1}.$$

Therefore

$$G = \frac{\sigma_1^2}{\sigma_2^2}\frac{(n_1-1)}{n_2-2)}\frac{\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}{\sum_{j=1}^{n_1}(X_j-\overline{X})^2} = \frac{\sigma_1^2}{\sigma_2^2}F \sim F_{n_2-1,n_1-1}.$$

as required.

(b) The likelihood is:

$$L(\mu_1,\mu_2,\sigma_1,\sigma_2) = \frac{1}{(2\pi)^{(n_1+n_2)/2}\sigma_1^{n_1}\sigma_2^{n_2}}\exp\{-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n_1}(x_j-\mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}(y_j-\mu_2)^2\}$$

The likelihood ratio statistic is:

$$\lambda(x,y) = \frac{\sup_{\mu_1,\mu_2,\sigma} L(\mu_1,\mu_2,\sigma,\sigma)}{\sup_{\mu_1,\mu_2,\sigma_1,\sigma_2} L(\mu_1,\mu_2,\sigma_1,\sigma_2)}$$

For the numerator (restriction to H_0 true), the likelihood, subject to the constraint that $\sigma_1 = \sigma_2 = \sigma$ is maximised for $(\mu_1,\mu_2,\sigma^2) = (\widehat{\mu}_{10},\widehat{\mu}_{20},\widehat{\sigma}_0^2)$ where

$$(\widehat{\mu}_{10},\widehat{\mu}_{20},\widehat{\sigma}_0^2) = (\overline{x},\overline{y},\frac{1}{n_1+n_2}(\sum_{j=1}^{n_1}(x_j-\overline{x})^2 + \sum_{j=1}^{n_2}(y_j-\overline{y})^2))$$

For the denominator (no restrictions on parameter space) the likelihood is maximised for $(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) = (\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}_1^2,\widehat{\sigma}_2^2)$, where

$$(\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}_1^2,\widehat{\sigma}_2^2) = (\overline{x},\overline{y},\frac{1}{n_1}\sum_{j=1}^{n_1}(x_j-\overline{x})^2,\frac{1}{n_2}\sum_{j=1}^{n_2}(y_j-\overline{y})^2).$$

Note:

$$\widehat{\sigma}_0^2 = \frac{n_1}{n_1+n_2}\widehat{\sigma}_1^2 + \frac{n_2}{n_1+n_2}\widehat{\sigma_2}^2$$

then, using the usual trick of

$$\begin{aligned}&\left(\sum_{j=1}^{n_1}(x_j-\overline{x})^2+\sum_{j=1}^{n_2}(y_j-\overline{y})^2\right)=(n_1+n_2)\widehat{\sigma}_0^2,\\&\sum_{j=1}^{n_1}(x_j-\overline{x})^2=n_1\widehat{\sigma}_1^2,\\&\sum_{j=1}^{n_2}(y_j-\overline{y})^2=n_2\widehat{\sigma}_2^2\end{aligned}$$

gives:

$$\begin{aligned}\lambda(x,y) &= \frac{\frac{1}{(2\pi)^{(n_1+n_2)/2}\widehat{\sigma}_0^{n_1+n_2}}\exp\left\{-\frac{1}{2\widehat{\sigma}_0^2}\left(\sum_{j=1}^{n_1}(x_j-\overline{x})^2+\sum_{j=1}^{n_2}(y_j-\overline{y})^2\right)\right\}}{\frac{1}{(2\pi)^{(n_1+n_2)/2}\widehat{\sigma}_1^{n_1}\widehat{\sigma}_2^{n_2}}\exp\left\{-\frac{1}{2\widehat{\sigma}_1^2}\sum_{j=1}^{n_1}(x_j-\overline{x})^2-\frac{1}{\widehat{2\sigma_2^2}}\sum_{j=1}^{n_2}(y_j-\overline{y})^2\right\}}\\
&= \frac{\widehat{\sigma}_1^{n_1}\widehat{\sigma_2}^{n_2}}{\widehat{\sigma}_0^{n_1+n_2}}=\frac{(n_1+n_2)^{n_1+n_2}}{n_1^{n_1/2}n_2^{n_2/2}}\left(\frac{n_1\widehat{\sigma}_1^2}{n_1\widehat{\sigma}_1^2+n_2\widehat{\sigma}_2^2}\right)^{n_1/2}\left(\frac{n_2\widehat{\sigma}_2^2}{n_1\widehat{\sigma}_1^2+n_2\widehat{\sigma}_2^2}\right)^{n_2/2}\\
&= \frac{(n_1+n_2)^{(n_1+n_2)/2}}{n_1^{n_1/2}n_2^{n_2/2}}\left(\frac{1}{1+\frac{\sum_j(y_j-\overline{y})^2}{\sum_j(x_j-\overline{x})^2}}\right)^{(n_1+n_2)/2}\left(\frac{\sum_j(y_j-\overline{y})^2}{\sum_j(x_j-\overline{x})^2}\right)^{n_2/2}\\
&= \frac{(n_1+n_2)^{(n_1+n_2)/2}}{n_1^{n_1/2}n_2^{n_2/2}}\frac{R^{n_2/2}}{(1+R)^{(n_1+n_2)/2}}\end{aligned}$$

Then

$$\lambda(x,y)=\frac{(n_1+n_2)^{(n_1+n_2)/2}}{n_1^{n_1/2}n_2^{n_2/2}}\frac{(\frac{n_2-1}{n_1-1}F)^{n_2/2}}{(1+\frac{n_2-1}{n_1-1}F)^{(n_1+n_2)/2}}.$$

Reject H_0 if and only if $\lambda(x,y)<c$. We would like to show that this implies: reject H_0 for $F<k_1$ and $F>k_2$ for some k_1 and k_2, which we will then compute (or at least find an expression for).

Note that $\lambda=\lambda(F)$ (it is a function of F). As a function of F,

$$\frac{d}{dF}\log\lambda(F)=0\Leftrightarrow F=\frac{1-\frac{1}{n_1}}{1-\frac{1}{n_2}}.$$

Therefore $\lambda(0)=\lambda(+\infty)=0$, $\lambda(F)$ increases from 0 to a unique maximum at $F=\frac{1-\frac{1}{n_1}}{1-\frac{1}{n_2}}$ and then decreases to 0. The rejection region therefore has the form $F<k_1$, $F>k_2$ as required. Since $F\sim F_{n_2-1,n_1-1}$, k_1 and k_2 satisfy the following two equations: with confidence level $1-\alpha$,

$$\begin{cases}F_{n_2-1,n_1-1}(k_2)-F_{n_2-1,n_1-1}(k_1)=1-\alpha\\ \frac{k_1}{(1+\frac{n_2-1}{n_1-1}k_1)^{1+(n_1/n_2)}}=\frac{k_2}{(1+\frac{n_2-1}{n_1-1}k_2)^{1+(n_1/n_2)}}\end{cases}$$

For the first of these, $F_{n_2-1,n_1-1}(x)=\mathbb{P}(F\le x)$ for $F\sim F_{n_2-1,n_1-1}$. For the second of these, since we reject for $\lambda\le k$ for some value k, we have $\lambda(k_1)=\lambda(k_2)=k$.

2. (a) For the reduced model, $\beta^{(1)}$ is estimated by

$$\beta^{(1)*} = (X_1^t X_1)^{-1} X_1^t Y$$

so that (using $\mathbb{E}[Y] = X\beta$)

$$\mathbb{E}[\beta^{(1)*}] = (X_1^t X_1)^{-1} X_1^t (X_1|X_2) \begin{pmatrix} \beta^{(1)} \\ \beta^{(2)} \end{pmatrix} = \beta^{(1)} + (X_1^t X_1)^{-1} X_1^t X_2 \beta^{(2)}$$

and hence, using $\mathbb{E}[\widehat{\mu}_0] = X_1 \mathbb{E}[\beta^{(1)*}]$ and $\mathbb{E}[\widehat{\mu}] = X\beta$, we have:

$$\begin{aligned} \mathbb{E}\left[\widehat{\mu} - \widehat{\mu}_0\right] &= (X_1|X_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - X_1 \beta^{(1)} - X_1 (X_1^t X_1)^{-1} X_1^t X_2 \beta^{(2)} \\ &= \left(I - X_1 (X_1^t X_1)^{-1} X_1^t\right) X_2 \beta^{(2)}. \end{aligned}$$

(b) The non-centrality parameter is:

$$\theta^2 = \frac{1}{\sigma^2} \left(\mathbb{E}\left[\widehat{\mu} - \widehat{\mu}_0\right]^t \mathbb{E}\left[\widehat{\mu} - \widehat{\mu}_0\right] \right)$$

and, using the previous part,

$$\begin{aligned} \mathbb{E}\left[\widehat{\mu} - \widehat{\mu}_0\right]^t \mathbb{E}\left[\widehat{\mu} - \widehat{\mu}_0\right] &= \beta^{(2)t} X_2^t (I - X_1 (X_1^t X_1)^{-1} X_1^t)(I - X_1 (X_1^t X_1)^{-1} X_1^t) X_2 \beta^{(2)} \\ &= \beta^{(2)t} (X_2^t X_2 - X_2^t (X_1^t X_1)^{-1} X_1^t X_2) \beta^{(2)} \end{aligned}$$

From lectures (moving to canonical co-ordinates): $(Q_{\text{res},I} - Q_{\text{res},II}) \sim \chi_q^2(\theta^2)$ and $Q_{\text{res},II} \sim \chi^2_{n-(p+q+1)}$. These are independent. The result follows from the definition of the non-central F distribution.

A quick reminder of lectures: consider a linear model $Y = (X_1|X_2) \left(\begin{array}{c} \beta^{(1)} \\ \hline \beta^{(2)} \end{array} \right) + \epsilon$ where $\beta^{(1)}$ is a $p+1$ vector, $\beta^{(2)}$ is a q vector and let

$$\mathcal{S}_1 = \{X_1 \beta : \beta \in \mathbb{R}^{p+1}\} \qquad \mathcal{S} = \{(X_1|X_2)\gamma : \gamma \in \mathbb{R}^{p+q+1}\}$$

then we can find an $n \times n$ orthonormal matrix $V = \left(\begin{array}{c} V^{(1)} \\ \hline V^{(2)} \\ \hline V^{(3)} \end{array} \right)$ where $V^{(1)}$ spans the space $\mathcal{S}_1$ and $\left(\begin{array}{c} V^{(1)} \\ \hline V^{(2)} \end{array} \right)$ spans the space $\mathcal{S}$. Let $U = VY$. Then

$$U \sim N(VX\beta, \sigma^2 VIV') = N(VX\beta, \sigma^2 I).$$

This is a vector of n independent random variables, where $U_i \sim N(\eta_i, \sigma^2)$ and $\eta_{p+q+2} = \ldots = \eta_n = 0$ for some $\eta_1, \ldots, \eta_{p+q+1}$.

Let

$$U^{(1)} = \begin{pmatrix} U_1 \\ \vdots \\ U_{p+1} \end{pmatrix} \qquad U^{(2)} = \begin{pmatrix} U_{p+2} \\ \vdots \\ U_{p+q+1} \end{pmatrix}$$

We have

$$\widehat{\mu} = (V^{(1)\prime}|V^{(2)\prime}) \left(\frac{U^{(1)}}{U^{(2)}} \right) \qquad \widehat{\mu}_0 = V^{(1)}U^{(1)}$$

so that

$$\begin{aligned} Q_{\text{res},I} - Q_{\text{res},II} &= |Y - \widehat{\mu}_0|^2 - |Y - \widehat{\mu}|^2 \\ &= |V'U - V^{(1)\prime}U^{(1)}|^2 - |V'U - (V^{(1)\prime}|V^{(2)\prime}) \left(\frac{U^{(1)}}{U^{(2)}} \right) |^2 \\ &= |(V^{(2)\prime}|V^{(3)\prime}) \left(\frac{U^{(2)}}{U^{(3)}} \right) |^2 - |V^{(3)\prime}U^{(3)}|^2 \\ &= U^{(2)\prime}V^{(2)}V^{(2)\prime}U^{(2)} = \sum_{j=p+2}^{p+q+1} U_j^2. \end{aligned}$$

while

$$\mathbb{E}[\widehat{\mu} - \widehat{\mu}_0] = V^{(2)\prime} \begin{pmatrix} \eta_{p+2} \\ \vdots \\ \eta_{p+q+1} \end{pmatrix}$$

so that

$$|\mathbb{E}[\widehat{\mu} - \widehat{\mu}_0]|^2 = \sum_{j=p+2}^{p+q+1} \eta_j^2$$

$\frac{Q_{\text{res},II}}{\sigma^2} = \sum_{j=p+q+2}^{n} U_j^2 \sim \chi^2_{n-(p+q+1)}$. Furthermore,

$$\frac{Q_{\text{res},I} - Q_{\text{res},II}}{\sigma^2} = \sum_{j=p+2}^{p+q+1} \left(\frac{U_j}{\sigma} \right)^2 \sim \chi^2_q \left(\sum_{j=p+2}^{q} \left(\frac{\eta_j}{\sigma} \right)^2 \right)$$

and $\frac{Q_{\text{res},I} - Q_{\text{res},II}}{\sigma^2} \perp \frac{Q_{\text{res},II}}{\sigma^2}$ and the result follows by the definition of the non-central F distribution.

3. (a) The MLE for $(\mu_1, \dots, \mu_p)$ is $(\overline{Y}_{1.}, \dots, \overline{Y}_{p.})$.

$$\overline{Y}_{i.} = \widehat{\alpha} + \widehat{\beta}_i$$

$$\sum_i n_i \overline{Y}_{i.} = n\widehat{\alpha} \Rightarrow \widehat{\alpha} = \overline{Y}_{..}$$

$$\widehat{\beta}_i = \overline{Y}_{i.} - \overline{Y}_{..}$$

(b)

$$\mathrm{Var}(\widehat{\alpha}) = \frac{\sigma^2}{n}$$

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_i) &= \mathrm{Var}(\overline{Y}_{i.} - \overline{Y}_{..}) \\
&= \mathrm{Var}((1 - \frac{n_i}{n})\overline{Y}_{i.} - \sum_{j \neq i} \frac{n_j}{n} \overline{Y}_{j.}) \\
&= \left(1 - \frac{n_i}{n}\right)^2 \frac{\sigma^2}{n_i} + \sum_{j \neq i} \frac{n_j^2}{n^2} \frac{\sigma^2}{n_j} \\
&= (1 - \frac{n_i}{n})^2 \frac{\sigma^2}{n_i} + \left(1 - \frac{n_i}{n}\right) \frac{\sigma^2}{n} \\
&= \left(\frac{1}{n_i} - \frac{2}{n} + \frac{n_i}{n^2} + \frac{1}{n} - \frac{n_i}{n^2}\right) \sigma^2 = \left(\frac{1}{n_i} - \frac{1}{n}\right) \sigma^2
\end{aligned}$$

$$\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}_i) = \mathrm{Cov}(\overline{Y}_{..}, \overline{Y}_{i.}) - \mathrm{Var}(\overline{Y}_{..}) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$$

$$i \neq j : \qquad \mathrm{Cov}(\widehat{\beta}_i, \widehat{\beta}_j) = -\mathrm{Cov}(\overline{Y}_{i.}, \overline{Y}_{..}) - \mathrm{Cov}(\overline{Y}_{j.}, \overline{Y}_{..}) + \mathrm{Var}(Y_{..}) = -\frac{\sigma^2}{n}$$

(c)

$$\alpha \in \left(\widehat{\alpha} \pm \frac{s}{\sqrt{n}} t_{n-p,a/2}\right)$$

(d)

$$\beta_j \in \left(\widehat{\beta}_j \pm s\sqrt{\frac{1}{n_j} - \frac{1}{n}} t_{n-p,a/2}\right)$$

where a is the significance and

$$s = \sqrt{\frac{Q\mathrm{res}}{n-p}} \qquad Q\mathrm{res} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\mathrm{Y}_{ij} - \widehat{\alpha} - \widehat{\beta}_j)^2$$

4.

$$\delta^2 = \frac{1}{\sigma^2} |\mu - \mu_0|^2 = \frac{1}{\sigma^2} \sum_{i=1}^{p} n_i \beta_i^2$$

5.

$$X(X^t X)^{-1} X^t X = X \Rightarrow X(X^t X)^{-1} X^t (X_1 | X_2) = (X_1 | X_2) \Rightarrow X(X^t X)^{-1} X_1 = X_1.$$

6. (a) Using $\widehat{\epsilon} = Y - \widehat{Y}$ and Σ_V to denote the covariance matrix of a random vector V,

$$\Sigma_Y = \Sigma_{\widehat{Y}} + \Sigma_{\widehat{\epsilon}} + \sigma^2 X(X^tX)^{-1}X^t(I - X(X^tX)^{-1}X^t) = \Sigma_{\widehat{Y}} + \Sigma_{\widehat{\epsilon}}$$

as required.

(b) We'll consider this in two ways. Firstly, directly and secondly, by putting into canonical variables. Directly:

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^n (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y})$$

$$\sum_{i=1}^n (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) = Y^t(I - X(X^tX)^{-1}X^t)(X(X^tX)^{-1}X^t - X_1(X_1^tX_1)^{-1}X_1^t)Y$$

where $X_1 = (1, \ldots, 1)^t$. From above (previous exercise, taking $X = (X_1|X_2)$), it follows that:

$$\sum_{i=1}^n (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) = 0$$

and the result follows.

Canonical variables: Assume X is $n \times r$, of rank r $U = A^tY$ where A is an orthonormal $n \times n$ matrix. We let $A_{.1} = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^t$ so that $U_1 = \sqrt{n}\overline{Y}$. We let $A_{.2}, \ldots, A_{.r}$ be the $r-1$ unit vectors, orthogonal to each other and to $A_{.1}$, so that $A_{.1}, \ldots, A_{.r}$ are an orthonormal basis for the space $\mathcal{S} = \{X\beta : \beta \in \mathbb{R}^r\}$. Let $B_1 = (A_{.1}|0|\ldots|0)$ (the first column $A_{.1}$ the other columns 0), $B_1 = (0|A_{.2}|\ldots|A_{.r}|0|\ldots|0)$ (the $n \times n$ matrix with the first column 0s and the subsequent $r-1$ columns $A_{.1}, \ldots, A_{.r}$ and the remaining columns 0. Let $B_3 = A - B_2 - B_1$. Then

$$\begin{aligned}\sum (Y_i - \overline{Y})^2 &= (Y - \overline{Y}\mathbf{1}_n)^t(Y - \overline{Y}\mathbf{1}_n)\\ &= U^t(A - B_1)^t(A - B_1)U = U^tB_2^tB_2U + U^tB_3^tB_3U\\ &= (\widehat{Y} - \overline{Y}\mathbf{1}_n)^t(\widehat{Y} - \overline{Y}\mathbf{1}_n) + (Y - \widehat{Y})^t(Y - \widehat{Y})\end{aligned}$$

as required.

7. (a) Firstly, to show that $H_{ii} \leq 1$ for each i: $I - H$ is non-negative definite, since $\widehat{\epsilon}^t\widehat{\epsilon} = Y^t(I - H)Y$, hence $H_{ii} \leq 1$. (otherwise there would exist a diagonal element of $I - H$ which was negative, say i. Take vector $v = (0, \ldots, 0, 1, 0, \ldots, 0)^t$ to get $v^t(I - H)v = 1 - H_{ii} < 1$). Secondly, to show that $H_{ii} \geq \frac{1}{n}$ for each i: let $X^{(1)} = (1, \ldots, 1)^t$ and $X = (X^{(1)}|X^{(2)})$ then

$$G := X(X^tX)^{-1}X^t - X^{(1)}(X^{(1)t}X^{(1)})^{-1}X^{(1)t}$$

is positive definite, since

$$G'G = G^2 = (H - H^{(1)})^2 = H^2 - HH^{(1)} - H^{(1)}H + H^{(1)2} = H - H^{(1)} = G$$

using the hint. Now,

$$H^{(1)} = X^{(1)}(X^{(1)\prime}X^{(1)})^{-1}X^{(1)\prime} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$$

where $\mathbf{1}$ is the n vector with each entry 1. Therefore $H_{ii} = \frac{1}{n} + G_{ii}$ for all $1 \leq i \leq p$. By the argument for the previous part, $G_{ii} \geq 0$, from which the result follows.

(b) Since H is idempotent, it has eigenvalues 1 or 0. Since it is of rank p, it has p eigenvalues 1 and $n-p$ eigenvalues 0. The trace is the sum of the eigenvalues, hence the result follows.

(c)

$$\mathrm{Cov}(Y, \widehat{Y}) = \mathrm{Var}(\widehat{Y})$$

$$\mathrm{Cor}(Y_i, \widehat{Y}_i) = \frac{\mathrm{Var}(\widehat{Y}_i)}{\sqrt{\mathrm{Var}(Y_i)\mathrm{Var}(\widehat{Y}_i)}} = \frac{H_{ii}}{\sqrt{H_{ii}}} = \sqrt{H_{ii}}.$$

8. (a) Follows directly from previous exercise;

$$\sum_{i=1}^{n}(Y_i - \overline{Y})(\widehat{Y}_i - \overline{Y}) = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2.$$

(b)

$$\begin{aligned}\lambda(y) &= \frac{\sup_\sigma \sup_{\beta_0} \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \beta_0)^2\right\}}{\sup_\sigma \sup_\beta \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta)\right\}} \\ &= \frac{\widetilde{\sigma}^{-n}}{\widehat{\sigma}^{-n}}\end{aligned}$$

where

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{y})^2 \qquad \widehat{\sigma}^2 = \frac{1}{n}(y - X\widehat{\beta})^t(y - X\widehat{\beta})$$

$$\widehat{\beta} = (X^tX)^{-1}X^tY \qquad \widehat{Y} = X(X^tX)^{-1}X^tY$$

It now follows that

$$\lambda(y) = \left(\frac{Q_{\mathrm{res}}}{Q_T}\right)^{n/2}$$

where $Q_{\mathrm{res}} = \sum_{j=1}^{n}(Y_j - \widehat{Y})^2$ and $Q_T = \sum_{j=1}^{n}(Y_j - \overline{Y})^2$. The likelihood ratio test is: reject H_0 for $\lambda(y) < k$ for some k which is equivalent to: reject H_0 for $R^2 > c$ for some c.

(c) Use the result: if $X \sim \mathrm{Gamma}(a, \lambda)$ and $Y \sim \mathrm{Gamma}(b, \lambda)$ then

$$\frac{X}{X+Y} \sim \mathrm{Beta}(a, b).$$

This is basic calculus: let $V = \frac{X}{X+Y}$, then for $t \in (0, 1)$,

$$\mathbb{P}(V \leq t) = \mathbb{P}(X \leq \frac{t}{1-t}Y) = \frac{\lambda^{a+b}}{\Gamma(a)\Gamma(b)}\int_0^\infty dy y^{b-1}e^{-\lambda y}\int_{ty/(1-t)}^\infty dx x^{a-1}e^{-\lambda x}$$

Take derivative with respect to t to get the density:

$$f_V(t) = \frac{\lambda^{a+b}}{\Gamma(a)\Gamma(b)} \frac{t^{a-1}}{(1-t)^{a+1}} \int_0^\infty y^{a+b-1} e^{-\lambda y/(1-t)} dy$$

Now use: $z = \frac{\lambda y}{1-t}$ to get:

$$f_V(t) = \frac{1}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1} \int_0^\infty z^{a+b-1} e^{-z} dz = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1} \qquad t \in (0,1)$$

as required.

Let $Q_M = \sum_{j=1}^n (\widehat{Y}_j - \overline{Y})^2$, then $Q_M \perp Q_{\text{res}}$ and

$$\frac{Q_M}{\sigma^2} \sim \chi^2_{p-1} = \Gamma(\frac{p-1}{2}, \frac{1}{2}) \qquad \frac{Q_{\text{res}}}{\sigma^2} \sim \chi^2_{n-p} = \Gamma(\frac{n-p}{2}, \frac{1}{2})$$

so that

$$R^2 = \frac{Q_T - Q_{\text{res}}}{Q_T} = \frac{Q_M}{Q_M + Q_{\text{res}}} \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right).$$

9. Firstly, $HX\beta = X(X^tX)^{-1}X^tX\beta = X\beta$ so that

$$\widehat{Y} - X\beta = HY - X\beta = H(Y - X\beta) = H\epsilon.$$

Therefore $\mathbb{E}[\widehat{Y}] = X\beta$ and

$$\begin{aligned} \mathbb{E}[|Y^* - \widehat{Y}|^2] &= \mathbb{E}[|Y^* - X\beta + X\beta - \widehat{Y}|^2] \\ &= \mathbb{E}[|Y^* - X\beta|^2] + \mathbb{E}[|\widehat{Y} - X\beta|^2] \\ &= n\sigma^2 + \text{tr}\text{Var}(\widehat{Y}) \\ &= n\sigma^2 + \sigma^2 \text{tr}(H) \end{aligned}$$

using $\text{Var}(\widehat{Y}) = \text{Var}(HY) = \sigma^2 HIH^t = \sigma^2 H.$

For the right hand side, using $Y - \widehat{Y} = (I - H)Y$, $\mathbb{E}[Y] = \mathbb{E}[\widehat{Y}]$ and $(I-H)^2 = I - H$ gives:

$$\mathbb{E}[|(I-H)Y|^2] = \text{tr}\text{Var}((I-H)Y) = \sigma^2 \text{tr}(I-H) = n\sigma^2 - \sigma^2\text{tr}(H)$$

so that

$$\mathbb{E}[|Y^* - \widehat{Y}|^2] = \mathbb{E}[|Y - \widehat{Y}|^2] + 2\sigma^2 \text{tr}(H).$$

Chapter 12

Complete Sufficiency and UMVU Estimators

Let x denote data from a distribution from a parametric family. Suppose T is a sufficient statistic. The original data x may be expressed as $(T(x), S(x))$ where T is a sufficient statistic and $S(x)$ is a statistic required to determine x uniquely once $T(x)$ is known. For example, if $T(x) = \overline{x}$, then $S(x) = (x_1 - \overline{x}, \ldots, x_n - \overline{x})$ gives the required information. If $T(x_1, \ldots, x_n) = (x_{1:n}, \ldots, x_{n:n})$ (the *order* statistics) then $S(x) = (r_1, \ldots, r_n)$ (the ranks of the data) gives the required information. In both cases, the distribution of the statistic S does not depend on the parameter value. S is known as an *ancillary statistic*

Definition 12.1 (Ancillary Statistic). *Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a statistical model. An* ancillary statistic *is a statistic whose distribution does not depend upon the parameter θ.*

A *minimal sufficient* statistic is not necessarily independent of ancillary statistics; there are situations where ancillary statistics can give important information for estimating a parameter θ.

Example 12.1.

Let X_1, X_2 be independent identically distributed with probability function

$$p_\theta(\theta) = p_\theta(\theta + 1) = p_\theta(\theta + 2) = \frac{1}{3}.$$

$\theta \in \Theta = \mathbb{Z}$. Let $X_{1:2} \leq X_{2:2}$ be the order statistic and let

$$R = X_{2:2} - X_{1:2}, \qquad M = \frac{X_{1:2} + X_{2:2}}{2}.$$

The joint probability distribution of (X_1, X_2) is:

$$p(x_1, x_2; \theta) = \frac{1}{9} \qquad (x_1, x_2) \in (\theta, \theta + 1, \theta + 2) \times (\theta, \theta + 1, \theta + 2).$$

The statistic (R, M) is clearly *minimal sufficient* for θ, since (R, M) gives the same information as $(X_{1:2}, X_{2:2})$ and $\frac{L(\theta; M(x_1,x_2), R(x_1,x_2))}{L(\theta; M(y_1,y_2), R(y_1,y_2))}$ does not depend on θ if and only if $(x_{1:2}, x_{2:2}) = (y_{1:2}, y_{2:2})$.

The distribution of R does not depend on θ. We have:

$$p_R(0) = \frac{1}{3} \qquad p_R(1) = \frac{4}{9} \qquad p_R(2) = \frac{2}{9},$$

irrespective of θ, hence R is *ancillary.*

The distribution of M is:

$$\begin{array}{l} \mathbb{P}_\theta(M = \theta) = \mathbb{P}_\theta(M = \theta + 2) = \frac{1}{9} \\ \mathbb{P}_\theta(M = \theta + \frac{1}{2}) = \mathbb{P}(M = \theta + \frac{3}{2}) = \frac{2}{9} \\ \mathbb{P}_\theta(M = \theta + 1) = \frac{1}{3}. \end{array}$$

With only information $M = m$, the values $\theta = m, m - \frac{1}{2}, m - \frac{3}{2}$ and $m - 2$ are all possible. The additional information that $R = 2$ forces

$$\begin{cases} X_{2:2} - X_{1:2} = 2 \\ X_{1:2} + X_{2:2} = 2m \end{cases}$$

which implies that $X_{1:2} = m - 1$ and $X_{2:2} = m + 1$. The ancillary statistic R by itself gives no information about θ, but an ancillary statistic may give information about the precision of an estimate.

Note that here (M, R) is the minimal sufficient statistic; it is not independent of R (the ancillary statistic).

To ensure that a sufficient statistic is independent of all ancillary statistics, an additional concept is required, that of *completeness.*

Definition 12.2 (Complete Statistic)**.** *Let $T(X)$ be a statistic with p.d.f or p.m.f. $p(t;\theta)$. The family of probability distributions is said to be* complete *if*

$$\mathbb{E}_\theta[g(T)] = 0 \quad \forall\theta \in \Theta \Rightarrow \mathbb{P}_\theta(g(T) = 0) = 1 \quad \forall\theta \in \Theta.$$

Equivalently, the statistic $T(X)$ is said to be a complete statistic*.*

Note that completeness is a property of the *family* of distributions and not of one particular distribution.

Example 12.2 (Complete Sufficient Statistic for Binomial)**.**

Let $T \sim \text{Binomial}(n, p)$, for $0 < p < 1$. n is fixed. Let g be a function such that

$$\mathbb{E}_p[g(T)] = 0,$$

then

$$0 = \mathbb{E}_p[g(T)] = \sum_{t=0}^{n} g(t)\binom{n}{t}p^t(1-p)^{n-t} = (1-p)^n \sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t$$

for all $0 : 0 < p < 1$. It follows that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t \qquad \forall r \in (0, +\infty).$$

It follows that $g(t)\binom{n}{t} = 0$ for each $t = 0, 1, \ldots, n$ and hence $g(t) = 0$ for $t = 0, 1, \ldots, n$. $\mathbb{P}_p(g(T) = 0) = 1$ for all p and hence T is a complete statistic for p. □

The following theorem, known as Basu's theorem, states that sufficiency, together with completeness, implies that a statistic is independent of every ancillary statistic.

Theorem 12.3 (Basu's Theorem)**.** *If $T(X)$ is complete and sufficient, then it is independent of every ancillary statistic.*

Proof The proof given here is only for discrete distributions. Let $S(X)$ be any ancillary statistic, then $\mathbb{P}_\theta(S(X) = x)$ does not depend on θ (definition of ancillarity). Furthermore, $\mathbb{P}_\theta(S(X) = s|T(X) = t)$ does not depend on θ since $T(X)$ is sufficient (by the definition of sufficiency). To show that $S(X)$ and $T(X)$ are independent, it is sufficient to show that

$$\mathbb{P}(S(X) = s|T(X) = t) = \mathbb{P}(S(X) = s) \qquad \forall t \in \mathcal{T}$$

where $\mathcal{T} = T(\mathcal{X})$. Now,

$$\mathbb{P}(S(X) = s) = \sum_{t \in \mathcal{T}} \mathbb{P}(S(X) = s|T(X) = t)\mathbb{P}_\theta(T(X) = t).$$

Let

$$g(t) = \mathbb{P}(S(X) = s|T(X) = t) - \mathbb{P}(S(X) = s).$$

Then

$$\mathbb{E}_\theta[g(T)] = \sum_{t \in \mathcal{T}} g(t)\mathbb{P}_\theta(T(X) = t) = \mathbb{P}(S(X) = s) - \mathbb{P}(S(X) = s) = 0 \qquad \forall \theta \in \Theta.$$

Since $T(X)$ is a *complete* statistic, this implies that $g(t) = 0 \qquad \forall t \in \mathcal{T}$. The result follows. □

Example 12.3.

Let $X_1, \ldots, X_n$ be i.i.d. Exp(θ) observations. Let

$$s(X) = \frac{X_n}{X_1 + \ldots + X_n}.$$

$T(X) = \sum_{j=1}^n X_j$ is sufficient for θ. Furthermore $\theta X_1, \ldots, \theta X_n$ are i.i.d. Exp(1) and hence the distribution of $s(X) = \frac{\theta X_n}{\theta X_1 + \ldots + \theta X_n}$ does not depend on θ; $s(X)$ is ancillary.

To show that $T(X)$ is complete: consider any function g such that $\mathbb{E}_\theta[g(T)] = 0$ for all $\theta \in \Theta$. Since $T \sim \Gamma(n, \theta)$, it follows that

$$0 = \frac{\theta^n}{\Gamma(n)} \int_0^\infty x^{n-1} e^{-\theta x} g(x) dx \qquad \forall \theta > 0.$$

From this it follows, by Laplace transform theory, that $x^{n-1} g(x) = 0$ for all $x > 0$, hence $g(x) = 0$ for all $x > 0$ and hence that T is complete.

It now follows from Basu's theorem that $T \perp S$. □

12.1 Completeness and Parameter Estimation

Theorem 12.4 (Rao-Blackwell)**.** *Let W be a statistic which is an unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Let $\phi(T) := \mathbb{E}[W|T]$. Then $\mathbb{E}_\theta[\phi(T)] = \tau(\theta)$ and $\mathbf{V}_\theta(\phi(T)) \leq \mathbf{V}_\theta(W)$ for all θ. That is, the variance of $\phi(T)$ is uniformly lower than that of W; both are unbiased estimators of $\tau(\theta)$.*

Proof This is straightforward: firstly, $\phi(T)$ is unbiased.

$$\tau(\theta) = \mathbb{E}_\theta[W] = \mathbb{E}_\theta[\mathbb{E}[W|T]] = \mathbb{E}_\theta[\phi(T)].$$

To show the improvement on variance,

$$\begin{aligned} \mathbf{V}_\theta(W) &= \mathbf{V}_\theta(\mathbb{E}[W|T]) + \mathbb{E}_\theta[\mathbf{V}(W|T)] \\ &\geq \mathbf{V}_\theta(\phi(T)). \end{aligned}$$

Since W is a statistic (a function only of the sample) and T is a sufficient statistic (i.e. $\mathbb{P}(.|T)$ does not depend on θ, it follows that the distribution of $W|T$ does not depend on θ, hence $\phi(T)$ is a statistic. □

Theorem 12.5. *If there exists a UMVU estimator of $\tau(\theta)$, then it is unique.*

Proof Let W be a UMVU estimator and suppose there is another UMVU estimator W'. Let $W^* = \frac{1}{2}(W + W')$. Then $\tau(\theta = \mathbb{E}_\theta[W^*]$ so that W^* is also an unbiased estimator of $\tau(\theta)$. Then

$$\mathbf{V}_\theta(W^*) = \frac{1}{4}\mathbf{V}_\theta(W) + \frac{1}{4}\mathbf{V}_\theta(W') + \frac{1}{2}\mathbf{C}_\theta(W, W').$$

By Cauchy-Schwartz, $\mathbf{C}_\theta(W, W') \leq \sqrt{\mathbf{V}_\theta(W)\mathbf{V}(W')}$ with equality if and only if $W' = a(\theta)W + b(\theta)$ for constants $a(\theta)$ and $b(\theta)$. Since W and W' are UMVU, $\mathbf{V}_\theta(W) = \mathbf{V}_\theta(W') \leq \mathbf{V}_\theta(W^*)$. It follows that $\mathbf{V}_\theta(W) = \mathbf{V}_\theta(W') \leq \mathbf{V}_\theta(W^*)$. Hence

$$\mathbf{V}_\theta(W) = \mathbf{C}_\theta(W, W') = \mathbf{C}_\theta(W, a(\theta)W + b) = a(\theta)\mathbf{V}_\theta(W)$$

from which it follows that $a(\theta) = 1$. Since both W and W' are unbiased, it follows that $b(\theta) = 0$, hence $W = W'$. □

Theorem 12.6. *Let W be an unbiased estimator of $\tau(\theta)$. Then W is UMVU if and only if W is uncorrelated with all unbiased estimators of 0.*

Proof Let W be a UMVU estimator of $\tau(\theta)$. Let U be an unbiased estimator of 0 and let $V = W + aU$. Then V is an unbiased estimator of $\tau(\theta)$ and

$$\mathbf{V}_\theta(V) = \mathbf{V}_\theta(W) + 2a\mathbf{C}_\theta(W,U) + a^2\mathbf{V}_\theta(U).$$

Suppose that for some $\theta_0 \in \Theta$, $\mathbf{C}_{\theta_0}(W,U) < 0$. Then choose $a \in (0, -\frac{2\mathbf{C}_{\theta_0}(W,U)}{\mathbf{V}_{\theta_0}(U)})$. For such a choice of a, $2a\mathbf{C}_{\theta_0}(W,U) + a^2\mathbf{V}_{\theta_0}(U) < 0$, hence $\mathbf{V}_{\theta_0}(V) < \mathbf{V}_{\theta_0}(W)$ contradicting the fact that W is UMVU. A similar argument achieves a contradiction if $\mathbf{C}_{\theta_0}(W,U) > 0$ for some $\theta_0 \in \Theta$. □

Theorem 12.7. *(Lehmann-Scheffé) Let T be complete and sufficient for θ. Let S be unbiased for $\tau(\theta)$. Then $S^* = \mathbb{E}[S|T]$ is the unique UMVUE of $\tau(\theta)$.*

Proof Firstly, $\mathbb{E}_\theta[S^*] = \mathbb{E}_\theta[\mathbb{E}[S|T]] = \mathbb{E}_\theta[S] = \tau(\theta)$, so it is unbiased. Furthermore, $\mathbf{V}_\theta(S^*) \leq \mathbf{V}_\theta(S)$ by Rao-Blackwell. Furthermore, consider any other unbiased estimator S_1 and let $S_1^* = \mathbb{E}[S_1|T]$, then

$$\mathbb{E}_\theta[S^* - S_1^*] = \mathbb{E}_\theta[\mathbb{E}[S|T] - \mathbb{E}[S_1|T]] = \tau(\theta) - \tau(\theta) = 0$$

for all $\theta \in \Theta$. It follows from the definition of completeness that $\mathbb{P}_\theta(S^* - S_1^* = 0) = 1$ for all $\theta \in \Theta$. □
The sufficient statistics for an exponential family are also *complete.*

Theorem 12.8 (Complete Statistics in an Exponential Family)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables from an exponential family with p.m.f. or p.d.f. of the form:*

$$p(x;\theta) = h(x)\exp\left\{\sum_{j=1}^k \eta_j(\theta)T_j(x) - B(\theta)\right\}$$

where $\theta = (\theta_1, \ldots, \theta_k)$. The statistic

$$T(X) = \sum_{k=1}^n T(X_k)$$

where $T = (T_1, \ldots, T_k)$ is complete *if $\{(\eta_1(\theta), \ldots, \eta_k(\theta)) : \theta \in \Theta\}$ contains an open set in $\mathbb{R}^k$.*

Proof (Sketch) Firstly, consider an exponential family in canonical form. If $\mathbb{E}_\eta[g(T)] = 0$ for all $\eta \in \mathcal{E}$, then (by definition)

$$0 \equiv \int_{\mathbb{R}^k} e^{\sum_{j=1}^k \eta_j t_j} g(t)J(t)dt$$

where $J(t)$ is the Jacobean determinant for the co-ordinate transformation from x to t. This is a Laplace transform; $\widehat{gJ}(\eta) = 0$ for all $\eta \in \mathcal{E}$. If $\mathcal{E}$ contains an open set in $\mathbb{R}^k$, this is sufficient to conclude that $g(t)J(t) = 0$ Lebesgue almost everywhere and hence that $\mathbb{P}_\eta(g(T) = 0) = 1$ for all $\eta \in \mathcal{E}$.

The result as stated now follows. □

Tutorial 13

1. Let $X_1, \ldots, X_n$ be a random sample from distribution:

$$g(x,\theta) = \begin{cases} \theta x^{\theta-1} & 0 \le x \le 1 \qquad \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find the MLE of $\frac{1}{\theta}$. Is it unbiased? Is it UMVU?

   (b) Show that $\overline{X}$ is an unbiased estimator of $\frac{\theta}{1+\theta}$. Is it UMVU?

2. Let X and Y be two discrete random variables with well defined expected values and variances. Prove that:

   (a) $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$

   (b) $\text{Var}(X) = \text{Var}(E[X|Y]) + \mathbb{E}[\text{Var}(X|Y)]$.

3. Let $X_1, \ldots, X_{n+1}$ be independent Bernoulli(p) variables and let

$$h(p) = \mathbb{P}\left(\sum_{i=1}^{n} X_i > X_{n+1} \middle| \, p \right).$$

   (a) Show that

$$T(X_1, \ldots, X_{n+1}) = \begin{cases} 1 & \sum_{j=1}^{n} X_j > X_{n+1} \\ 0 & \text{otherwise} \end{cases}$$

   is an unbiased estimator of $h(p)$.

   (b) Find the UMVUE of $h(p)$.

4. Let X be an observation from the probability with mass function:

$$p(-1,\theta) = \frac{\theta}{2}, \quad p(0,\theta) = 1-\theta, \quad p(1,\theta) = \frac{\theta}{2} \qquad \theta \in [0,1].$$

   (a) Find the maximum likelihood estimator of θ and show that it is unbiased.

   (b) Let

$$T(X) = \begin{cases} 2 & x = 1 \\ 0 & x = -1 \quad \text{or} \quad 0 \end{cases}$$

   Show that T is an unbiased estimator of θ.

   (c) Show that $\widehat{\theta}$ (maximum likelihood estimator) is minimal sufficient for θ and that $\mathbb{E}[T|\widehat{\theta}] = \widehat{\theta}$. Show that $\text{Var}(\widehat{\theta}) < \text{Var}(T)$.

5. Consider a Gaussian linear model $Y = X\beta + \epsilon$, where Y is an n-vector, X is $n \times r$ of rank r ($r < n$) and $\epsilon \sim N(0, \sigma^2 I)$ and β is an r-vector of unknown parameters. σ^2 is unknown. Recall (from lectures) that the OLS estimator of $\widehat{\beta}$ is:

$$\widehat{\beta} = (X^t X)^{-1} X^t Y.$$

Show that $\widehat{\beta}_i$ is UMVU for each $i = 1, \ldots, r$ and that $S^2 = \frac{1}{n-r}\sum_{j=1}^n (Y_j - \widehat{Y}_j)^2$ is an UMVU estimator of σ^2, where $\widehat{Y} = X(X^tX)^{-1}X^tY$.

6. Let X be the number of dots showing when a fair die is rolled; i.e.

$$p_X(x) = \frac{1}{6} \qquad x = 1, 2, 3, 4, 5, 6.$$

Let Y be the number of heads obtained when X fair coins are tossed. Find

(a) The mean and variance of Y.

(b) The MSPE (mean squared prediction error) of the optimal linear predictor of Y based on X. The optimal linear predictor is the function $\widehat{Y} = aX + b$, where a and b are chosen such that $\mathbb{E}[\widehat{Y}] = \mathbb{E}[Y]$ and, subject to this constraint, to minimise $\mathrm{Var}(Y - \widehat{Y})$.

(c) The optimal linear predictor of Y given $X = x$ for $x = 1, 2, 3, 4, 5, 6$.

7. A person walks into a clinic at time t and is diagnosed with a certain disease. At the same time (t), a diagnostic indicator Z_0 of the severity of the disease (e.g. a blood cell count or a virus count) is obtained. Let S be the unknown date in the past when the subject was infected. We are interested in the time $Y_0 = t - S$ from infection until detection. Assume that the conditional density of Z_0 (the present condition) given $Y_0 = y$ is $N(\mu + \beta y_0, \sigma^2)$. Let

$$Z = \frac{Z_0 - \mu}{\sigma}, \qquad Y = \frac{\beta}{\sigma}Y_0.$$

(a) Show that the conditional density $p(z|y)$ of Z given $Y = y$ is $N(y, 1)$.

(b) Suppose that Y has exponential density $\pi(y) = \lambda \exp\{-\lambda y\}\mathbf{1}_{\{y>0\}}$ where $\lambda > 0$. Show that the conditional distribution of Y given $Z = z$ has density

$$\pi(y|z) = \frac{1}{(2\pi)^{1/2}c}\exp\left\{-\frac{1}{2}(y - (z - \lambda))^2\right\} \qquad y > 0$$

where c is a suitable constant (depending on z and λ). Compute c in terms of the c.d.f. Φ for a $N(0, 1)$ random variable.

(c) Find the conditional density $\pi_0(y_0|z_0)$ of Y_0 given $Z_0 = z_0$.

(d) Suppose it is known that $Z_0 = z_0$. Find an expression (in terms of the c.d.f for $N(0, 1)$ and its inverse) for $g(z_0)$, the best predictor of Y_0 given $Z_0 = z_0$ using mean *absolute* prediction error $\mathbb{E}[|Y_0 - g(Z_0)|]$.

(e) Let ϕ denote the density function for a $N(0, 1)$ random variable. Show that the best Mean Squared Prediction Error (MSPE) predictor of Y given $Z = z$ is:

$$\mathbb{E}[Y|Z = z] = \frac{1}{c}\phi(\lambda - z) - (\lambda - z).$$

Answers

1. (a) For $(x_1, \ldots, x_n) \in [0,1]^n$,

$$\log L(\theta; x_1, \ldots, x_n) = n \log \theta + (\theta - 1) \sum_{j=1}^{n} \log x_j$$

$$\frac{\partial}{\partial \theta} \log L(\theta) = \frac{n}{\theta} + \sum_{j=1}^{n} \log x_j$$

$$\frac{d^2}{d\theta^2} \log L(\theta) = -\frac{n}{\theta^2}$$

while $\log L(\theta) \overset{\theta \to 0, \theta \to +\infty}{\longrightarrow} -\infty$ hence unique maximiser which is $\widehat{\theta} = \frac{-1}{\sum_{j=1}^{n} \log x_j}$. Therefore:

$$\frac{1}{\widehat{\theta}_{ML}} = -\frac{1}{n} \sum_{j=1}^{n} \log X_j$$

$$\mathbb{E}_\theta \left[\frac{1}{\widehat{\theta}_{ML}} \right] = -\theta \int_0^1 x^{\theta - 1} \log x dx = \theta \int_0^\infty e^{-\theta y} y dy = \frac{1}{\theta}$$

so $\frac{1}{\widehat{\theta}_{ML}}$ is an unbiased estimator of $\frac{1}{\theta}$.

To show that it is UMVU: this is an exponential family;

$$L(x_1, \ldots, x_n; \theta) = \frac{1}{\prod_{j=1}^{n} x_j} \prod_{j=1}^{n} \mathbf{1}_{[0,1]}(x_j) \exp \left\{ \theta \sum_{j=1}^{n} \log x_j + n \log \theta \right\}.$$

The sufficient statistic $T(x_1, \ldots, x_n) = \sum_{j=1}^{n} \log x_j$ is therefore *complete.* The UMVU estimator is therefore:

$$\mathbb{E}\left[\frac{1}{\widehat{\theta}_{ML}} | T(X) \right] = \frac{1}{\widehat{\theta}_{ML}}.$$

(b)

$$\mathbb{E}[\overline{X}] = \mathbb{E}[X] = \theta \int_0^1 x^\theta dx = \frac{\theta}{1 + \theta}$$

so $\overline{X}$ is an unbiased estimator of $\frac{\theta}{1+\theta}$.

To show that it is not UMVU, $\mathbb{E}[\overline{X} | \sum_i \log X_i]$ is the unique UMVU estimator and this is not equal to $\overline{X}$ since with probability 1, $\overline{X}$ is not a function of $\sum_i \log X_i$.

2. (a)

$$\begin{aligned} \mathbb{E}[\mathbb{E}[X|Y]] &= \sum_y p_Y(y(\sum_x x p_{X|Y}(x|y) = \sum_{x,y} x \frac{p_{X,Y}(x,y)}{p_Y(y)} \\ &= \sum_x x(\sum_y p_{X,Y}(x,y) = \sum_x x p_X(x) = \mathbb{E}[X]. \end{aligned}$$

(b)

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[\mathbb{E}[X|Y]]^2 \\ &= \mathbb{E}[\mathrm{Var}(X|Y)] + \mathbb{E}[(\mathbb{E}[X|Y])^2] - \mathbb{E}[\mathbb{E}[X|Y]]^2 \\ &= \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}(\mathbb{E}[X|Y]).\end{aligned}$$

3. (a) Trivially clear from the definition: T is a binary variable taking values in $\{0,1\}$, therefore:

$$\mathbb{E}_p[T] = \mathbb{P}_p(T = 1) = h(p).$$

(b) $\sum_{j=1}^{n+1} X_j$ is a complete sufficient statistic for p, hence unique UMVUE is

$$S := \mathbb{E}[T|\sum_{j=1}^{n+1} X_j].$$

Now, for each $y \in \{0, 1, \dots, n+1\}$:

$$\mathbb{E}[T|\sum_{j=1}^{n+1} X_j = y] = \mathbb{P}(T = 1|\sum_{j=1}^{n+1} X_j = y) = \frac{\mathbb{P}(\sum_{j=1}^{n} X_j > X_{n+1}, \sum_{j=1}^{n+1} X_j = y)}{\mathbb{P}(\sum_{j=1}^{n+1} X_j = y)}.$$

The denominator is $\binom{n+1}{y} p^y (1-p)^{n+1-y}$; the numerator is:

$$\begin{cases} 0 & y = 0 \\ \mathbb{P}(\sum_{j=1}^{n} X_j = 1, X_{n+1} = 0) = np(1-p)^n & y = 1 \\ \mathbb{P}(\sum_{j=1}^{n} X_j = 2, X_{n+1} = 0) = \frac{n(n-1)}{2} p^2 (1-p)^{n-1} & y = 2 \\ \mathbb{P}(\sum_{j=1}^{n+1} X_j = y) = \binom{n+1}{y} p^y (1-p)^{n+1-y} & y \geq 3 \end{cases}$$

Note: for $y \geq 3$, it always holds that $\sum_{j=1}^{n} X_j > X_{n=1}$. Putting this together gives:

$$\begin{cases} 0 & y = 0 \\ \frac{n}{n+1} & y = 1 \\ \frac{n-1}{n+1} & y = 2 \\ 1 & y \geq 3 \end{cases}$$

4. (a)

$$L(\theta; x) = \frac{\theta}{2} \mathbf{1}_{\{-1,1\}}(x) + (1-\theta)\mathbf{1}_{\{0\}}(x)$$

Clearly this is maximised for: $\widehat{\theta}(1) = \widehat{\theta}(-1) = 1 \qquad \widehat{\theta}(0) = 0.$

To compute its expected value:

$$\mathbb{E}_\theta[\widehat{\theta}] = \frac{\theta}{2} + \frac{\theta}{2} = \theta.$$

(b) $\mathbb{E}[T(X)] = 2 \times \frac{\theta}{2} = \theta$ so it is unbiased.

(c) Note that $\widehat{\theta}(X) = |X|$. To show sufficiency:

$$\mathbb{P}_\theta(X = x||X| = y) = \frac{\mathbb{P}_\theta(X = x, |X| = y)}{\mathbb{P}_\theta(|X| = y)} = \begin{cases} 1 & x = 0, y = 0 \\ \frac{1}{2} & x = \pm 1, y = 1 \\ 0 & \text{other} \end{cases}$$

which does not depend on θ.

To prove *minimal* sufficiency: Any *reduction* is a function $S : S(|X|) = \text{constant}$ so that $\mathbb{P}_\theta(X \in .|S) = \mathbb{P}_\theta(X \in .)$ which does depend on θ. hence S is not sufficient. Therefore $\widehat{\theta}$ is *minimal* sufficient.

Clearly:

$$\mathbb{E}[T(X)||X|] = \begin{cases} 0 & X = 0 \\ 1 & X = \pm 1 \end{cases}$$

Finally: $\widehat{\theta} \sim Be(\theta)$ so that

$$\text{Var}(\widehat{\theta}) = \theta(1 - \theta).$$

while $T = 2Z$ for $Z \sim Be(\frac{\theta}{2})$ so that

$$\text{Var}(T) = 4\frac{\theta}{2}(1 - \frac{\theta}{2}) = 2\theta(1 - \frac{\theta}{2})$$

which is clearly greater.

5. Unbiased follows directly from lectures:

$$\widehat{\beta} = (X^t X)^{-1} X^t Y$$

so that

$$\mathbb{E}[\widehat{\beta}] = (X^t X)^{-1} X^t \mathbb{E}[Y] = (X^t X)^{-1} X^t X \beta = \beta.$$

For the sample standard deviation, let $H = X(X^t X)^{-1} X^t$ then H is idempotent, of rank r and hence $H = PDP^t$ where P is orthonormal and $D = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$ where 1 appears with multiplicity r. Hence

$$Y - \widehat{Y} = (I - H)Y = (I - H)X\beta + (I - H)\epsilon = (I - H)\epsilon.$$

Let $\eta = P^t \epsilon$ then $\eta \sim N(0, \sigma^2 I)$. Also,

$$(Y - \widehat{Y})^t (Y - \widehat{Y}) = \eta^t (I - D)\eta = \sum_{r+1}^{n} \eta_j^2$$

so that

$$\frac{(n-r)S^2}{\sigma^2} = \sum_{r+1}^{n}\left(\frac{\eta_j}{\sigma}\right)^2 \sim \chi^2_{n-r}.$$

From this, $\mathbb{E}[S^2] = \sigma^2$ so that the estimator is unbiased.

Now to show that the estimators are UMVU:

$$p(y_1,\ldots,y_n) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)^t(y-X\beta)\right\}$$

and the argument inside $\exp\{-\frac{1}{2}(.)\}$ is:

$$\frac{1}{\sigma^2}(y^ty - y^tX\beta - \beta^tX^ty + \beta^tX^tX\beta).$$

The sufficient statistic is therefore:

$$T(y) = (y^ty, \sum_{j=1}^{n} X_{ji}y_j : i = 1,\ldots,r).$$

$\widehat{\beta}_i = \sum_{jk}(X^tX)^{-1}_{ij}X_{kj}y_k$ is clearly a linear function of the sufficient statistics. For the standard deviation:

$$(Y - \widehat{Y}^t)(Y - \widehat{Y}) = Y^tY - \widehat{Y}^t\widehat{Y}$$

This holds since

$$Y^t\widehat{Y} = Y^tHY = Y^tH^tHY = \widehat{Y}^t\widehat{Y}$$

Now $\widehat{Y} = X(X^tX)^{-1}X^tY$ which is a (linear) function of the sufficient statistics and hence

$$\mathbb{E}[S^2|T(Y)] = S^2.$$

The result follows by the Lehman-Scheffé theorem.

6. (a) $\mathbb{E}[Y] = \frac{7}{4}$,

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X]) = \frac{1}{4}\mathbb{E}[X] + \frac{1}{4}\mathrm{Var}(X) = \frac{3}{8} + \frac{12.5+4.5+0.5}{24} = \frac{26.5}{24} = 1\frac{5}{48}.$$

(b)

$$\widehat{Y} = aX + b$$

minimise

$$\mathrm{Var}(Y - aX - b) = \mathrm{Var}(Y) + a^2\mathrm{Var}(X) - 2a\mathbf{C}(Y,X)$$

gives:

$$a = \frac{\mathbf{C}(Y, X)}{\mathrm{Var}(X)}$$

We can show $\mathrm{Cov}(X, Y) = \frac{1}{2}\mathrm{Var}(X)$ as follows:

$$\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y|X]] = \frac{1}{2}\mathbb{E}[X^2]$$

$$\mathbb{E}[Y] = \frac{1}{2}\mathbb{E}[X]$$

hence

$$\mathbf{C}(Y, X) = \frac{1}{2}\mathrm{Var}(X) \Rightarrow a = \frac{1}{2}$$

$$\mathrm{Var}(Y - \widehat{Y}) = \mathrm{Var}(Y) - \frac{1}{4}\mathrm{Var}(X) = \frac{1}{4}\mathbb{E}[X] = \frac{7}{8}.$$

(c)

$$\mathbb{E}[\widehat{Y}] = \mathbb{E}[Y] = \frac{1}{2}\mathbb{E}[X].$$

Now using $\widehat{Y} = aX + b$ with $a = \frac{1}{2}$ gives $b = 0$ so that

$$\widehat{Y} = \frac{1}{2}X.$$

7. (a) $Z \sim N(y, 1)$ follows directly from rescaling.

(b)

$$\begin{aligned} \pi(y|z) = \frac{\pi(y)p(z|y)}{p(z)} &\propto \lambda e^{-\lambda y}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-y)^2}\mathbf{1}_{\{y>0\}} \\ &= \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{y^2}{2} + y(z-\lambda) - \frac{z^2}{2}\right\}\mathbf{1}_{\{y\geq 0\}} \end{aligned}$$

so

$$\pi(y|z) = K\exp\left\{-\frac{1}{2}(y - (z-\lambda))^2\right\}\mathbf{1}_{\{y\geq 0\}}$$

$$1 = \sqrt{2\pi}K\int_{-(z-\lambda)}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \sqrt{2\pi}K\Phi(z-\lambda)$$

where Φ is the $N(0, 1)$ c.d.f., hence

$$\pi(y|z) = \frac{1}{\sqrt{2\pi}\Phi(z-\lambda)}\exp\left\{-\frac{1}{2}(y-(z-\lambda))^2\right\}\mathbf{1}_{\{y\geq 0\}}$$

(c)

$$\pi_0(y_0|z_0) = \frac{\beta}{(2\pi)^{1/2}\sigma\Phi(\frac{z_0-\mu}{\sigma} - \lambda)}\exp\left\{-\frac{1}{2\sigma^2}(\beta^2 y_0 - (z_0 - \mu - \lambda))^2\right\}\mathbf{1}_{\{y_0>0\}}$$

(d) First find $h(z)$, the best linear predictor of Y given $Z = a$. Then $g(z_0) = \frac{\beta}{\sigma} h(\frac{z_0 - \mu}{\sigma})$. $h(z)$ is the value of h that minimises $\int_0^\infty |y - h| \pi(y|z) dy$ so that h satisfies:

$$\int_0^h \frac{1}{(2\pi)^{1/2}} \exp\{-\frac{1}{2}(y - (z - \lambda))^2\} dy = \int_h^\infty \frac{1}{(2\pi)^{1/2}} \exp\{-\frac{1}{2}(y - (z - \lambda))^2\} dy$$

giving

$$\Phi(h - (z - \lambda)) - \Phi(-(z - \lambda)) = 1 - \Phi(h - (z - \lambda)),$$

$$\Phi(h - (z - \lambda)) = \frac{1}{2}(1 + \Phi(-(z - \lambda)))$$

$$h(z) = (z - \lambda) + \Phi^{-1}\left(\frac{1}{2}(1 + \Phi(\lambda - z))\right).$$

(e) We want to find h which minimises

$$\int_0^\infty (y - h)^2 \pi(y|z) dz$$

which is given by $h(z) = \mathbb{E}[Y|Z = z]$. This is:

$$\begin{aligned} \mathbb{E}[Y|Z = z] &= (z - \lambda) + \frac{1}{(2\pi)^{1/2} c} \int_0^\infty (y - (z - \lambda)) e^{-\frac{1}{2}(y - (z - \lambda))^2} dy \\ &\quad = (z - \lambda) + \frac{1}{(2\pi)^{1/2} c} \int_0^{e^{-\frac{1}{2}(z - \lambda)^2}} dx \\ &= (z - \lambda) + \frac{1}{(2\pi)^{1/2} c} e^{-\frac{1}{2}(z - \lambda)^2}. \end{aligned}$$

Chapter 13

Asymptotic Results

This chapter is devoted to *asymptotic results*; firstly, *consistency* is discussed, the problem of when a sequence of estimators $(\widehat{q}_n)_{n\geq 1}$ converges in probability to $q(\theta)$, the quantity to be estimated. Then, techniques based on the Central Limit theorem are discussed to give conditions under which the maximum likelihood estimators of the canonical parameters an exponential family are asymptotically normal.

13.1 Consistency

Let θ be an unknown parameter from a parameter space Θ. Let $(\widehat{q}_n)_{n\geq 1}$ be a sequence of estimators of $q(\theta)$, where $q : \Theta \to \mathbb{R}^p$.

Definition 13.1. *The sequence $(\widehat{q}_n)_{n\geq 1}$ is* consistent *if for all $\theta \in \Theta$ and $\epsilon > 0$,*

$$\mathbb{P}_\theta\left(|\widehat{q}_n - q(\theta)| \geq \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{13.1}$$

where $|.|$ denotes the Euclidean norm. It is said to be uniformly consistent over $K \subseteq \Theta$ *(or simply uniformly consistent if $K = \Theta$) if*

$$\sup_{\theta\in K} \mathbb{P}_\theta\left(|\widehat{q}_n - q(\theta)| \geq \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{13.2}$$

13.1.1 The Weak Law of Large Numbers

The simplest example of *consistency* is convergence of the sample average to the population average.

Theorem 13.2. *Let $X_1, \ldots, X_n$ be i.i.d. with distribution $\mathbb{P}$. Suppose that $\mathbb{E}\left[|X_i|\right] < +\infty$, then $\overline{X} \to_{\mathbb{P}} \mathbb{E}\left[X_1\right] =: \mu$.*

Sketch of Proof Only a sketch of the proof is given, since the result is treated fully in Probability 2. Let $\phi_X(t) = \mathbb{E}\left[e^{itX}\right]$ denote the characteristic function of the random variable X. Since $|\phi_X(t)| \leq 1$ for all $t \in \mathbb{R}$, Taylor's expansion theorem may be applied to give:

$$\phi_X(t) = \mathbb{E}\left[1 + itZ + o(t)\right] = 1 + it\mu + o(t).$$

It follows that, for $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$,

$$\phi_{\overline{X}}(t) = \prod_{j=1}^{n} \phi_{X_j/n}(t) = \phi_X\left(\frac{t}{n}\right)^n = \left(1 + i\frac{t}{n}\mu + o(\frac{t}{n})\right)^n \overset{n\rightarrow+\infty}{\longrightarrow} e^{it\mu}.$$

This is the characteristic function of the constant random variable μ and hence by the Lévy continuity theorem (omitted) $\overline{X} \rightarrow_{\mathbb{P}} \mu$. □

Without further assumptions, *uniform* consistency cannot be proved only on the assumption that $\mathbb{E}_\theta\left[|X_1|\right] < +\infty$ for each $\theta \in \Theta$; stronger conditions are required. If, in addition, $\mathbf{V}_\theta(X_1) < M < +\infty$ where M does not depend on θ, Chebyshev's inequality may be used to prove uniform consistency;

$$\mathbb{P}_\theta\left(\left|\overline{X} - \mu(\theta)\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbf{V}_\theta(\overline{X}) \leq \frac{M}{n\epsilon^2}.$$

The main tool to prove consistency will be the law of large numbers. This is natural for a method of moments estimator, where the parameter estimators will be functions of moment estimators.

We start with a result about functions of estimators of *multinomial* sampling probabilities. The estimators $\widehat{p_i} := \frac{N_i}{n}$ of p_i are the sample averages and hence are consistent by the law of large numbers. Here n denotes the total number of trials and N_i the total number that gave the outcome labelled i. These are the estimators obtained by maximum likelihood or moment method or simply frequency estimators, which all turn out to be the same for multinomial sampling.

These estimators are *uniformly* consistent, since $\mathbb{V}_p(\widehat{p_i}) = \frac{p_i(1-p_i)}{n} \leq \frac{1}{4n}$; this bound does not depend on p. To prove *uniform* consistency for a function of these estimators, the function has to be *uniformly* continuous. This is obtained 'for free' if the parameter space is compact. In the following theorem for multinomial sampling, the usual parameter space is extended by taking the closure. This gives a compact space. A continuous function over a compact space is uniformly continuous.

Theorem 13.3. *Let $\mathcal{P} = \{(p_1, \ldots, p_k) : 0 \leq p_j \leq 1, \quad 1 \leq j \leq k, \quad \sum_{j=1}^k p_j = 1\}$. Let $\mathbb{P}_p$ denote the probability distribution $\underline{p} = (p_1, \ldots, p_k)$ over $\mathcal{X} = (x_1, \ldots, x_k)$. Let $X_1, \ldots, X_n$ denote a random sample from $\mathbb{P}_p$. Let $N_j = \sum_{i=1}^n \mathbf{1}_{x_j}(X_i)$ and $\widehat{p}_{n,j} = \frac{N_j}{n}$ for $j = 1, \ldots, k$. Let $q : \mathcal{P} \rightarrow \mathbb{R}^p$ be continuous. Let $\underline{\widehat{p}}_n = (\widehat{p}_{n,1}, \ldots, \widehat{p}_{n,k})$. Then $\widehat{q}_n := q(\underline{\widehat{p}}_n)$ is a uniformly consistant estimator of $q(\underline{p})$.*

Proof Let $\underline{\widehat{p}}_n = (\widehat{p}_{1,n}, \ldots, \widehat{p}_{k,n})$. Note that $\mathbb{E}\left[\widehat{p}_{nj}\right] = p_j$ for each j and $\mathbf{V}_p\left(\widehat{p}_{nj}\right) = \frac{p_j(1-p_j)}{n} \leq \frac{1}{4n}$. By Chebyshev's inequality, it follows that for all $p = (p_1, \ldots, p_k) \in \mathcal{P}$ and $\delta > 0$,

$$\begin{aligned} \mathbb{P}_p\left(\left|\underline{\widehat{p}}_n - \underline{p}\right| \geq \delta\right) &= \mathbb{P}_p\left(\sum_{j=1}^k \left(\widehat{p}_{nj} - p_j\right)^2 \leq \delta^2\right) \leq \mathbb{P}_p\left(\cup_{j=1}^k \left\{k\left(\widehat{p}_{nj} - p_j\right)^2 \geq \delta^2\right\}\right) \\ &\leq \sum_{j=1}^k \mathbb{P}_p\left(\left|\widehat{p}_{nj} - p_j\right| \geq \frac{\delta}{\sqrt{k}}\right) \leq \frac{k^2}{4n\delta^2}. \end{aligned}$$

Because q is continuous and $\mathcal{P}$ is compact, it follows that q is *uniformly* continuous on $\mathcal{P}$. It follows that for every $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that for all $\underline{p}, \underline{p}' \in \mathcal{P}$,

$$|\underline{p}' - \underline{p}| \leq \delta(\epsilon) \qquad \Rightarrow \qquad |q(\underline{p}') - q(\underline{p})| \leq \epsilon.$$

It follows that

$$\mathbb{P}_p\left(|\widehat{q}_n - q| \geq \epsilon\right) \leq \mathbb{P}_p\left(|\widehat{\underline{p}}_n - \underline{p}| \geq \delta(\epsilon)\right) \leq \frac{k^2}{4n\delta(\epsilon)^2}$$

and the result follows. □

The aim of the discussion that now follows is to establish Theorem 13.6; that is, that as $n \rightarrow +\infty$, the probability of existence of the ML estimator for an exponential family tends to 1 and the sequence of ML estimators from random samples of size n is consistent. This requires Proposition 13.4 and Lemma 13.5, which is a corollary to Theorem 5.3 (which gives conditions under which $\widehat{\eta}_{ML}$ exists for an exponential family in its canonical co-ordinates; $\widehat{\eta}_{ML} = \dot{A}^{-1}(t)$ where t is the observed value of the sufficient statistic).

Proposition 13.4. *Let $X_1, \ldots, X_n$ be i.i.d., each with state space $\mathcal{X}$ and let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a regular family of probability distributions over $\mathcal{X}$. Let $\underline{g} = (g_1, \ldots, g_d)$ map $\mathcal{X}$ onto $\mathcal{Y} \subset \mathbb{R}^d$. Suppose $\mathbb{E}_\theta[|g_j(X_1)|] < +\infty$ for $1 \leq j \leq d$ for all $\theta \in \Theta$. Let $m_j(\theta) = \mathbb{E}_\theta[g_j(X_1)]$ and let $q(\theta) = h(m(\theta))$ where $h : \mathcal{Y} \rightarrow \mathbb{R}^p$ is a continuous function. Then*

$$\widehat{q} = h(\overline{g}) = h\left(\frac{1}{n}\sum_{i=1}^n g(X_i)\right)$$

is a consistent estimate of $q(\theta)$.

Proof It follows from the weak law of large numbers that

$$\frac{1}{n}\sum_{i=1}^n g(X_i) \rightarrow_{\mathbb{P}} \mathbb{E}_{\mathbb{P}}\left[g(X_1)\right].$$

It is straightforward to establish that if $\underline{Y}_n \rightarrow_{\mathbb{P}} \underline{Y}$ and h is continuous then $h(\underline{Y}_n) \rightarrow_{\mathbb{P}} h(\underline{Y})$. □

The following result is a corollary to Theorem 5.3.

Lemma 13.5. *Suppose that $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in \mathcal{E}\}$, where $\mathcal{E}$, the natural parameter space is open; $\mathcal{P}$ is the canonical exponential family generated by (h, T) where $T = (T_1, \ldots, T_k)$, of rank k. Let C_T denote the convex support of the distribution of T under $\mathbb{P}_\eta$ for all $\eta \in \mathcal{E}$. Let $t_0 = \mathbb{E}_\eta[T(X)]$. Then $\widehat{\eta}$, the MLE exists and is unique if and only if $t_0 \in C_T^0$, the interior of C_T.*

Proof of Lemma 13.5 Recall Theorem 5.3, that if there is a $\delta > 0$ and an $\epsilon > 0$ such that t_0 satisfies:

$$\inf_{(c_1,\ldots,c_k):\sum_j c_j^2=1} \mathbb{P}\left((c, T(X) - t_0) > \delta\right) > \epsilon$$

then the MLE $\widehat{\eta}$ exists, is unique and is the solution to the equation

$$\dot{A}(\eta) = \mathbb{E}_\eta\left[T(X)\right] = t_0.$$

The point t_0 satisfies $t_0 \in C^0$ the interior of a convex set C if and only if for every $d \neq 0$, both $\{t : (d, t) > (d, t_0)\} \neq \emptyset$ and $\{t : (d, t) < (d, t_0)\} \neq \emptyset$. The equivalence of Equation (5.1) and Lemma 13.5 follows. □

The main result, which is a consequence of Proposition 13.4, together with Lemma 13.5 may now be stated and proved.

Theorem 13.6. *Let $\mathcal{P}$ be a canonical exponential family of rank d generated by $T = (T_1, \ldots, T_d)$. Let η denote the natural parameter, $\mathcal{E}$ the natural parameter space and A the log partition function. Suppose that $\mathcal{E}$ is open. Let $X_1, \ldots, X_n$ be a random sample from $\mathbb{P}_\eta \in \mathcal{P}$. Let $\widehat{\eta}$ denote the MLE. Then*

1. $\mathbb{P}_\eta\left(\widehat{\eta}_{MLE} \quad \textit{exists}\right) \overset{n\to+\infty}{\longrightarrow} 1.$

2. $(\widehat{\eta}_n)_{n\geq 1}$ *is consistent.*

Proof It follows from Lemma 13.5 that $\widehat{\eta}(X_1, \ldots, X_n)$ exists if and only if $\overline{T}_n := \frac{1}{n}\sum_{j=1}^n T(X_j)$ belongs to the interior C_T^0 of the convex support of the distribution of $\overline{T}_n$. If η_0 is the parameter value, then $\mathbb{E}_{\eta_0}\left[T(X_1)\right]$ belongs to the interior of the convex support by Theorem 5.3 since η_0 solves the equation $\dot{A}(\eta_0) = t_0 = \mathbb{E}_{\eta_0}[T(X_1)]$. By definition of the convex support, there exists a ball

$$S_\delta := \{t : |t - \mathbb{E}_{\eta_0}\left[T(X_1)\right]| < \delta\} \subset C_T^0.$$

By WLLN,

$$\frac{1}{n}\sum_{i=1}^n T(X_i) \longrightarrow_{\mathbb{P}_{\eta_0}} \mathbb{E}_{\eta_0}\left[T(X_1)\right],$$

from which

$$\mathbb{P}_{\eta_0}\left(\frac{1}{n}\sum_{i=1}^n T(X_i) \in C_T^0\right) \geq \mathbb{P}_{\eta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n T(X_i) - \mathbb{E}_{\eta_0}\left[T(X_1)\right]\right| < \delta\right) \overset{n\to+\infty}{\longrightarrow} 1.$$

Since $\widehat{\eta}$ is the solution to $\dot{A}(\eta) = \frac{1}{n}\sum_{i=1}^n T(X_i)$, it follows that the probability that $\widehat{\eta}$ exists tends to 1 and hence 1. follows.

For part 2., by Theorem 3.7, the map $\eta \to \dot{A}(\eta)$ is 1 - 1 and continuous on $\mathcal{E}$. It follows that the inverse $\dot{A}^{-1} : \dot{A}(\mathcal{E}) \to \mathcal{E}$ is continuous on S_δ and the result now follows by Proposition 13.4. □

13.1.2 Consistency of Minimum Contrast Estimators

Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a regular family, let $\rho(x, \theta)$ be a contrast function and let $X = (X_1, \ldots, X_n)$ be a random sample from $\mathbb{P}_\theta$. Let $\widehat{\theta}$ be a minimum contrast estimate that minimises

$$\rho_n(X, \theta) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \theta).$$

Recall that if ρ is a contrast function, then (by definition) $D(\theta_0, \theta) := \mathbb{E}_{\theta_0}[\rho(X_1, \theta)]$ is uniquely minimised at $\theta = \theta_0$ for all $\theta_0 \in \Theta$.

Note: Maximum Likelihood The maximum likelihood estimator is obtained by minimising the contrast function $\rho_n(X, \theta) = -\frac{1}{n}\sum_{j=1}^{n} \log L(\theta; X_j)$ where L denotes the likelihood function for a single observation. Here

$$D(\theta_0, \theta) = -\mathbb{E}_{\theta_0}[\log L(\theta; X)].$$

Theorem 13.7. *Suppose*

$$\sup_{\theta \in \Theta} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (\rho(X_i, \theta) - D(\theta_0, \theta)) \right| \right\} \xrightarrow{\mathbb{P}_{\theta_0}} 0 \qquad (13.3)$$

and

$$\inf_{\theta : |\theta - \theta_0| > \epsilon} (D(\theta_0, \theta) - D(\theta_0, \theta_0)) > 0 \qquad \forall \epsilon > 0 \qquad (13.4)$$

then $\widehat{\theta}$ is consistent.

Proof Firstly, consider the set on which $|\widehat{\theta} - \theta_0| > \epsilon$. For a minimum contrast estimate $\widehat{\theta}$ such that $|\widehat{\theta} - \theta_0| > \epsilon$, it is necessary that $\widehat{\rho}_n(X, \widehat{\theta}) < \rho_n(X, \theta_0)$. Clearly, if $|\widehat{\theta} - \theta_0| > \epsilon$, then

$$\widehat{\theta} \in \{\theta : |\theta - \theta_0| > \epsilon|\}.$$

Also,

$$\widehat{\theta} \in \{\theta : \rho_n(X, \theta) \leq \rho_n(X, \theta_0)\}.$$

It follows that

$$\mathbb{P}_{\theta_0}\left(|\widehat{\theta} - \theta_0| \geq \epsilon\right) \leq \mathbb{P}_{\theta_0}\left(\inf_{|\theta - \theta_0| \geq \epsilon} \frac{1}{n} \sum_{j=1}^{n} (\rho(X_j, \theta) - \rho(X_j, \theta_0)) \leq 0 \right). \qquad (13.5)$$

Let

$$A = \inf_{\theta : |\theta - \theta_0| \geq \epsilon} \frac{1}{n} \sum_{i=1}^{n} (\rho(X_i, \theta) - \rho(X_i, \theta_0))$$

and

$$B = \inf_{\theta : |\theta - \theta_0| > \epsilon} (D(\theta_0, \theta) - D(\theta_0, \theta_0)).$$

From the hypotheses, $\mathbb{P}_{\theta_0} \left(\sup_\theta |\rho_n(\underline{X}, \theta) - D(\theta_0, \theta)| > \delta \right) \overset{n \to +\infty}{\longrightarrow} 0$ and hence, using

$$D(\theta_0, \theta) - D(\theta_0, \theta_0) \geq 0,$$

that for all $\delta > 0$,

$$\begin{aligned} \mathbb{P}_{\theta_0}(A - B \leq -\delta) \quad & \leq \quad \mathbb{P}_{\theta_0} \left(\inf_{\theta : |\theta - \theta_0| \geq \epsilon} (\rho_n(X, \theta) - D(\theta_0, \theta)) - (\rho_n(X, \theta_0) - D(\theta_0, \theta_0)) \leq -\delta \right) \\ & \overset{n \to +\infty}{\longrightarrow} 0. \end{aligned} \qquad (13.6)$$

Now choose $\delta = \inf_{\theta : |\theta - \theta_0| > \epsilon} (D(\theta_0, \theta) - D(\theta_0, \theta_0))$. Then $\delta > 0$ and from Equation (13.6), it follows directly that the right hand side of (13.5) tends to zero. □

The following simple and important corollary gives a condition under which the MLE is consistent.

Corollary 13.8. *Let $\Theta = \{\theta_1, \ldots, \theta_d\}$ denote a finite parameter space. Suppose that*

$$\max_{j,k} \mathbb{E}_{\theta_j} \left[|\log p(X_1, \theta_k)| \right] < +\infty$$

and suppose that the parametrisation is identifiable. Let $\widehat{\theta}$ denote the MLE of θ. Then $\mathbb{P}_{\theta_j} \left(\widehat{\theta} \neq \theta_j \right) \to 0$ for all $j \in \{1, \ldots, d\}$.

Proof Since the parameter space is discrete and finite, it follows that there is an $\epsilon > 0$ such that

$$\mathbb{P}_{\theta_j} \left(\widehat{\theta} \neq \theta_k \right) = \mathbb{P}_{\theta_j} \left(\left| \widehat{\theta} - \theta_k \right| \geq \epsilon \right) \qquad \forall (j, k).$$

Recall that the MLE estimator is the minimum contrast estimator with contrast function

$$\frac{1}{n} \sum_{i=1}^{n} \log p(X_j, \theta).$$

By Shannon's lemma 4.3, $D(\theta_0, \theta)$ is minimised for $\theta = \theta_0$ for all $\theta_0 \in \Theta$. It follows that only equations (13.3) and (13.4) need to be checked.

Equation (13.3) follows from the WLLN. Equation (13.4) follows from Shannon's lemma. □

13.2 The Delta Method

Let $X_1, \ldots, X_n$ be a random sample from a parent distribution satisfying $\mathbb{E}[X_1] = \mu$ and $\mathbf{V}(X_1) = \sigma^2 < +\infty$. The central limit theorem states that

$$\sqrt{n}(\overline{X} - \mu) \overset{n \to +\infty}{\longrightarrow}_{\mathcal{L}} N(0, \sigma^2).$$

Here $\mathcal{L}$ means *law*, which is another term used for *distribution*; the terms *convergence in law* and *convergence in distribution* mean exactly the same thing.

Definition 13.9 (Convergence in law / distribution)**.** *A sequence $X_1, X_2, \ldots$ of real valued random variables is said to* converge in distribution*, or* converge in law *to a random variable X if*

$$\lim_{n \to +\infty} F_n(x) = F(x)$$

for every $x \in \mathbb{R}$ at which F is continuous. Here F_n, defined by $F_n(x) := \mathbb{P}(X_n \leq x)$ and F, defined by $F(x) := \mathbb{P}(X \leq x)$ are the cumulative distribution functions of the random variables X_n and X respectively.

For random k-vectors $(X_1, X_2, \ldots)$ where for each n, $X_n \in \mathbb{R}^k$, convergence in distribution is defined similarly. The sequence $(X_1, X_2, \ldots)$ converges in distribution to a random k- vector X if

$$\lim_{n \to +\infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$$

for every $A \in \mathcal{B}(\mathbb{R}^k)$ which is a continuity set of X. Here $\mathcal{B}(\mathbb{R}^k)$ denotes the Borel σ algebra over $\mathbb{R}^k$ and a continuity set *A of X is a set A such that $\mathbb{P}(X \in \partial A) = 0$ (∂A denotes the boundary of A).*

The *delta method* is simply the name given to the application of Taylor's expansion theorem to obtain the distribution of functions of the sample average.

Theorem 13.10 (The Delta Method)**.** *Let $X_1, \ldots, X_n$ be a random sample, where X_1 has state space $\mathbb{R}$, $\mathbb{E}[X_1] = \mu$, $\mathbf{V}(X_1) = \sigma^2 < +\infty$ and $h : \mathbb{R} \to \mathbb{R}$ a differentiable function. Then*

$$\sqrt{n}(h(\overline{X}) - h(\mu)) \overset{n \to +\infty}{\longrightarrow}_{\mathcal{L}} N\left(0, \left(h'(\mu)\right)^2 \sigma^2\right) \tag{13.7}$$

The result follows from the following lemma.

Lemma 13.11. *Let $\{U_n\}$ be a sequence of real valued random variables and $\{a_n\}$ a sequence of constants that satisfies $a_n \to +\infty$ as $n \to +\infty$. Suppose that*

1. *$a_n(U_n - u) \overset{n \to +\infty}{\longrightarrow}_{\mathcal{L}} V$ for some constant $u \in \mathbb{R}$ where V is a well defined random variable,*
2. *$g : \mathbb{R} \to \mathbb{R}$ is differentiable at u with derivative $g'(u)$. Then*

$$a_n \left(g(U_n) - g(u)\right) \overset{n \to +\infty}{\longrightarrow} g'(u)V \tag{13.8}$$

Proof of Lemma 13.11 We use heavily the following result:

- Convergence in probability implies convergence in law.
- Convergence in law implies convergence in probability *when the limiting random variable is a constant.*

Since $a_n \to +\infty$ as $n \to +\infty$, it follows that $U_n - u \to 0$ in distribution and therefore convergence in probability also holds; for every $\delta > 0$,

$$\mathbb{P}(|U_n - u| \leq \delta) \to 1.$$

From the definition of a derivative, it follows that for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|v-u| \leq \delta \Rightarrow |g(v)-g(u)-(v-u)g'(u)| \leq \epsilon |v-u|.$$

From this, it follows that

$$\mathbb{P}\left(\left|g(U_n)-g(u)-g'(u)(U_n-u)\right| \leq \epsilon \left|U_n-u\right|\right) \geq \mathbb{P}(|U_n-u| \leq \delta) \overset{n\to+\infty}{\longrightarrow} 1.$$

Since

$$\{|g(U_n)-g(u)-g'(u)(U_n-u)| \leq \epsilon|U_n-u|\} = \{|a_n(g(U_n)-g(u))-a_n g'(u)(U_n-u)| \leq \epsilon|a_n(U_n-u)|\},$$

therefore:

$$\mathbb{P}\left(\left|a_n\left(g(U_n)-g(u)\right)-g'(u)(a_n(U_n-u)\right| \leq \epsilon\left|a_n(U_n-u)\right|\right) \overset{n\to+\infty}{\longrightarrow} 1.$$

Now, $a_n(U_n-u) \rightarrow_{\mathcal{L}} V$. Let $F_V(x) = \mathbb{P}(V \leq x)$. Let $A_n(x) = \{a_n\left(g(U_n)-g(u)\right) \leq x\}$ and let $B_n(x) = \{g'(u)(a_n(U_n-u)) \leq x\}$, then it is straightforward to show that for each $x \in \mathbb{R}$ $\mathbb{P}(A_n \backslash B_n) + \mathbb{P}(B_n \backslash A_n) \overset{n\to+\infty}{\longrightarrow} 0$ and hence for each x, $|\mathbb{P}(A_n(x))-\mathbb{P}(B_n(x))| \overset{n\to+\infty}{\longrightarrow} 0$. Now, $\mathbb{P}(B_n(x)) \overset{n\to+\infty}{\longrightarrow} F_V(x)$. The result follows. □

Proof of Theorem 13.10 This follows from the lemma by setting $U_n = \overline{X}$, $a_n = n^{1/2}$, $u = \mu$ and $V \sim N(0, \sigma^2)$. □

The delta method can be extended to situations where $h : \mathbb{R} \to \mathbb{R}$ is a twice differentiable function with $h'(\mu) = 0$, but $h''(\mu) \neq 0$.

Theorem 13.12 (Second order delta method)**.** *Let (Y_n) be a sequence of random variables that satisfy $\sqrt{n}(Y_n-\mu) \rightarrow_{\mathcal{L}} N(0,\sigma^2)$. Let h be a function that is twice differentiable and satisfies $h'(\mu) = 0$, $h''(\mu) \neq 0$. Then*

$$n(h(Y_n)-h(\mu)) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} \frac{\sigma^2}{2} h''(\mu) V$$

where $V \sim \chi_1^2$.

Proof Similar to the first order delta method; consider the second derivative and recall that if $V \sim \chi_1^2$, then $V =_{\mathcal{L}} Z^2$ where $Z \sim N(0,1)$. □

The delta method extends to the multivariate setting. Firstly, Lemma 13.11 extends directly:

Lemma 13.13. *Let $\{\underline{U}_n\}$ be random d-vectors and let $\{a_n\}$ be a sequence of constants satisfying $a_n \to +\infty$ as $n \to +\infty$ and suppose that*

1. *$a_n(\underline{U}_n - \underline{u}) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} \underline{V}$ where $\underline{V}$ is a random d-vector.*

2. *$g : \mathbb{R}^d \to \mathbb{R}^p$ has a differential $g^{(1)}_{p\times d}(\underline{u})$ at $\underline{u}$. Then*

$$a_n\left(g(\underline{U}_n)-g(\underline{u})\right) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} g^{(1)}(\underline{u})\underline{V}.$$

Proof Similar to Lemma 13.11. □

From this, the multivariate version of the delta method can be stated and proved.

Theorem 13.14 (Multivariate delta method)**.** *Let $\underline{Y}_1, \ldots, \underline{Y}_n$ be i.i.d. random d-vectors with well defined expected value $\underline{\mu}$ and covariance matrix Σ. Let $h : \mathcal{O} \to \mathbb{R}^p$ where $\mathcal{O}$ is an open subset of $\mathbb{R}^d$. Suppose that h has a well defined differential $h^{(1)}(\underline{\mu})$, where*

$$h_{ij}^{(1)}(\underline{\mu}) = \frac{\partial h_i}{\partial x_j}(\underline{\mu}).$$

Then

$$h(\overline{\underline{Y}}) = h(\underline{\mu}) + h^{(1)}(\underline{\mu})\left(\overline{\underline{Y}} - \underline{\mu}\right) + o_{\mathbb{P}}(n^{-1/2}) \tag{13.9}$$

where $o_{\mathbb{P}}(n^{-1/2})$ denotes a quantity V that satisfies:

$$\mathbb{P}(n^{1/2}|V| > \epsilon) \overset{n\to+\infty}{\longrightarrow} 0 \qquad \forall \epsilon > 0$$

and

$$\sqrt{n}\left(h(\overline{\underline{Y}}) - h(\underline{\mu})\right) \overset{n\to+\infty}{\longrightarrow} N(\underline{0}, h^{(1)}(\underline{\mu})\Sigma h^{(1)t}(\underline{\mu})). \tag{13.10}$$

The proof follows in the same way as before; let $a_n = \sqrt{n}$, $\underline{U}_n = \overline{\underline{Y}}$, $\underline{u} = \underline{\mu}$ and $\underline{V} \sim N(\underline{0}, \Sigma)$. Then

$$h^{(1)}(\underline{\mu})\underline{V} \sim N(\underline{0}, h^{(1)}(\underline{\mu})\Sigma h^{(1)t}(\underline{\mu}))$$

as required. □

13.3 Asymptotic Results for Maximum Likelihood

The following result gives the asymptotic distribution for the maximum likelihood estimator of the canonical parameters.

Theorem 13.15. *Let $\mathcal{P}$ be a canonical exponential family of rank d generated by T and suppose that $\mathcal{E}$ (the natural parameter space) is open. Let $X_1, \ldots, X_n$ be a random sample from $\mathbb{P}_\eta \in \mathcal{P}$. Let $\widehat{\eta}$ be the MLE if it exists and equal to a constant vector c otherwise. Then*

1.

$$\widehat{\eta} = \eta + \ddot{A}^{-1}(\eta)\left(\frac{1}{n}\sum_{i=1}^{n} T(X_i) - \dot{A}(\eta)\right) + o_{\mathbb{P}_\eta}(n^{-1/2})$$

2.

$$\sqrt{n}(\widehat{\eta} - \eta) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} N(\underline{0}, I^{-1}(\eta))$$

where $o_{\mathbb{P}_\eta}(n^{-1/2})$ denotes a quantity V_n such $\lim_{n\to+\infty} \mathbb{P}_\eta(n^{1/2}|V_n| > \epsilon) = 0$ for all $\epsilon > 0$.

Remark The asymptotic variance matrix $I^{-1}(\eta)$ of $\sqrt{n}\,(\widehat{\eta}-\eta)$ is the matrix that gives the Cramér Rao lower bound on variances of unbiased estimators of linear combinations of $(\eta_1,\ldots,\eta_d)$. This is the *asymptotic efficiency* property of the ML estimator for exponential families.

Proof This is an immediate consequence of the multivariate delta method. To simplify the presentation, the proof is first given for a one-parameter exponential family;

$$p(x;\eta) = h(x)\exp\{\eta T(x) - A(\eta)\} \qquad x \in \mathbb{R}, \quad \eta \in \mathcal{E} \subseteq \mathbb{R}.$$

In this case, $\dot{A}(\eta) = \frac{dA}{d\eta}(\eta)$. Let H denote the inverse of $\dot{A}$; namely, the function such that

$$H\left(\frac{dA}{d\eta}(\eta)\right) = \eta.$$

Let $\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T(X_j)$, then

$$\widehat{\eta}_{ML} = H(\overline{T})$$

Then $\mathbb{P}_\eta(\overline{T} \in \dot{A}(\mathcal{E})) \to 1$ and hence $\mathbb{P}_\eta(\widehat{\eta} = H(\overline{T})) \to 1$. Now, let $y = \frac{dA}{d\eta}(\eta)$, then $H(y) = \eta$ so that

$$1 = \frac{d}{d\eta}H(y) = \frac{dH}{dy}\frac{dy}{d\eta}$$

and since $\frac{dy}{d\eta} = \frac{d^2A}{d\eta}(\eta)$, therefore:

$$\frac{d^2A}{d\eta^2}(\eta) = \frac{1}{H'(y)} \Rightarrow H'(y) = \frac{1}{\frac{d^2A}{d\eta^2}(\eta)}.$$

Recall that $\mathbb{E}_\eta[T(X)] = \frac{dA}{d\eta}(\eta)$. Also, $\mathrm{Var}_\eta(T(X)) = \frac{d^2A}{d\eta^2}(\eta)$. By the delta method, it therefore follows that:

$$\sqrt{n}\left(H(\overline{T}) - H\left(\frac{dA}{d\eta}(\eta)\right)\right) \overset{n\to+\infty}{\longrightarrow} N\left(0, \frac{1}{\left(\frac{d^2A}{d\eta^2}\right)^2}\frac{d^2A}{d\eta^2}(\eta)\right).$$

Now use that: $I(\eta) = \frac{d^2A}{d\eta^2}(\eta)$ to get:

$$\sqrt{n}\,(\widehat{\eta}_{ML} - \eta) \overset{n\to+\infty}{\underset{\mathcal{L}}{\longrightarrow}} N\left(0, \frac{1}{I(\eta)}\right).$$

The second statement is proved.

Now consider the result in the general (multivariate) setting. Firstly, as before, let

$$\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T(X_j),$$

then (as before) $\mathbb{P}_\eta\left(\overline{T} \in \dot{A}(\mathcal{E})\right) \to 1$ and hence $\mathbb{P}_\eta\left(\widehat{\eta} = \dot{A}^{-1}(\overline{T})\right) \to 1$. Because the exponential family is of full rank, therefore $\dot{A} : \mathbb{R}^d \to \mathbb{R}^d$ is $1-1$. Now, using the notation of the multivariate version of the Delta method, set $h = \dot{A}^{-1}$ and $\underline{\mu} = \dot{A}(\eta)$ in Theorem 13.14.

The following result is from Analysis 2: Let $h : \mathbb{R}^d \to \mathbb{R}^d$ be $1-1$ and continuously differentiable on an open neighbourhood $\mathcal{O}$ of $\underline{x}$. Suppose that $Dh(x) := \left(\frac{\partial h_i}{\partial x_j}\right)_{d\times d}$ be non singular. Then $h^{-1} : h(\mathcal{O}) \to \mathcal{O}$ is differentiable at $y = h(x)$ and

$$Dh^{-1}(y) = (Dh(x))^{-1}.$$

By definition, $D\dot{A} = \ddot{A}$. In Theorem 13.14, $h^{(1)}(\underline{\mu}) = \ddot{A}^{-1}(\eta)$. The first statement of the theorem now follows from (13.9). The second part follows by noting that the covariance matrix of $T(X_1)$ is $\Sigma = \ddot{A}(\eta)$, from which

$$h^{(1)}(\underline{\mu})\Sigma h^{(1)t}(\underline{\mu}) = \ddot{A}^{-1}\ddot{A}\ddot{A}^{-1}(\eta) = \ddot{A}^{-1}(\eta).$$

statement 2. now follows from (13.10). □

Example 13.1 (Normal Random Sample).

Let $X_1, \ldots, X_n$ be a $N(\mu, \sigma^2)$ random sample. Note that

$$p(x_1, \ldots, x_n; \mu, \sigma) = \frac{1}{(2\pi)^{n/2}} \exp\left\{\frac{\mu}{\sigma^2}\sum_{j=1}^n x_j - \frac{1}{2\sigma^2}\sum_{j=1}^n x_j^2 - n\frac{\mu^2}{2\sigma^2} - n\log\sigma\right\}$$

Let $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$, then the model can be re-written in canonical form as

$$p(x_1, \ldots, x_n; \underline{\eta}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{n\left(\eta_1\overline{x} + \eta_2\overline{x^2} + \frac{\eta_1^2}{4\eta_2} - \log\sqrt{-\frac{1}{2\eta_2}}\right)\right\}$$

Then $T = (T_1, T_2)$ where $T_1 = \overline{X}$ and $T_2 = \overline{X^2}$ is a sufficient statistic for the parameters. Since $\mathbb{E}_\eta[T_1] = \mu$ and $\mathbb{E}_\eta[T_2] = \sigma^2 + \mu^2$, it follows from the central limit theorem that:

$$\sqrt{n}\begin{pmatrix} T_1 - \mu \\ T_2 - (\mu^2 + \sigma^2)\end{pmatrix} \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ddot{A}(\eta)\right)$$

where $A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}(\log 2 + \log(-\eta_2))$ so that

$$\ddot{A}(\eta) = \frac{1}{2\eta_2^2}\begin{pmatrix} -\eta_2 & \eta_1 \\ \eta_1 & 1 - \frac{\eta_1^2}{4\eta_2}\end{pmatrix}$$

The maximum likelihood estimators for the normal are:

$$\widehat{\mu} = \overline{X} = T_1$$

and

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{X})^2 = T_2 - (T_1)^2.$$

It follows that $\widehat{\eta}_1 = \frac{\overline{X}}{\widehat{\sigma}^2}$ and $\widehat{\eta}_2 = -\frac{1}{2\widehat{\sigma}^2}$. By the preceding theorem,

$$\sqrt{n}\begin{pmatrix} \widehat{\eta}_1 - \eta_1 \\ \widehat{\eta}_2 - \eta_2 \end{pmatrix} \overset{n\to+\infty}{\longrightarrow_{\mathcal{L}}} N\left(\underline{0}, I^{-1}(\eta)\right).$$

□

13.4 Asymptotic Distribution of Maximum Likelihood Estimators

The result for the asymptotic distribution of MLE estimators for the canonical parameters of exponential families may be extended to a large class of minimum contrast estimators. Let $\theta \in \Theta \subset \mathbb{R}^d$, where $\theta = (\theta_1, \ldots, \theta_d)$. Let $X = (X_1, \ldots, X_n)$ and let $\rho_n(X, \theta)$ be a minimum contrast function based on the random sample of the form

$$\rho_n(X, \theta) = -\frac{1}{n}\sum_{j=1}^{n} \rho(X_j, \theta) \qquad (13.11)$$

where ρ is a contrast function for a single observation. For example, take $\rho(x, \theta) = \log p(x, \theta)$ for maximum likelihood estimation. Assume the following:

1. ρ_n is differentiable in θ_j for each $j = 1, \ldots, d$. Let $\widehat{\theta}_n$ denote the minimum contrast estimate; that is, $\widehat{\theta}_n$ satisfies

$$\frac{\partial \rho_n}{\partial \theta_j}(X, \widehat{\theta}_n) = 0 \qquad j = 1, \ldots, d. \qquad (13.12)$$

   In the case of ρ_n given by Equation (13.11), this is the maximum likelihood estimate.

2.
$$\mathbb{E}_\theta\left[\frac{\partial \rho_n}{\partial \theta_j}(X, \theta)\right] = 0. \qquad (13.13)$$

$$\mathbb{E}_\theta\left[\left|\nabla_\theta \rho_n(X, \theta)\right|^2\right] < +\infty \qquad (13.14)$$

   where $|.|$ denotes the Euclidean norm.

3. ρ_n is twice differentiable in θ and satisfies

$$\sum_{j,k} \mathbb{E}_\theta\left[\left|\frac{\partial^2}{\partial \theta_j \partial \theta_k}\rho_n(X, \theta)\right|\right] < +\infty \qquad \forall \theta \in \Theta$$

   The matrix with entries $\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j}\rho_n(X, \theta)\right]$ is non-singular for each $\theta \in \Theta$.

4. $\widehat{\theta}_n \longrightarrow_{\mathbb{P}_\theta} \theta$ for each $\theta \in \Theta$.

For the fourth of these, in the case of exponential families of full rank, where $\theta = \theta(\eta)$ for a continuous $1-1$ mapping θ, $\widehat{\theta}_n \rightarrow_{\mathbb{P}_\theta} \theta$ by virtue of Theorem 13.15.

Theorem 13.16. *Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a regular parametric family. Let $X = (X_1, \ldots, X_n)$ be a random sample. Suppose that conditions 1., 2., 3. and 4. hold. Let J be the matrix with entries*

$$J_{jk}(\theta) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k}\rho(X,\theta)\right]$$

and let K be the matrix with entries

$$K_{jk}(\theta) = \mathbb{E}_\theta\left[\frac{\partial\rho}{\partial\theta_j}(X,\theta)\frac{\partial\rho}{\partial\theta_k}(X,\theta)\right].$$

Then the minimum contrast estimate satisfies:

$$\widehat{\theta}_n = \theta + J^{-1}(\theta)\nabla\rho_n(X,\theta) + o_{\mathbb{P}_\theta}(n^{-1/2}),$$

so that

$$\sqrt{n}\left(\widehat{\theta}_n - \widehat{\theta}\right) \overset{n\rightarrow+\infty}{\longrightarrow}_{\mathcal{L}} N(0, J^{-1}(\theta)K(\theta)J^{-1}(\theta)).$$

In the case of the maximum likelihood estimate, $I = J = K$ so that

$$\sqrt{n}\left(\widehat{\theta}_n - \widehat{\theta}\right) \overset{n\rightarrow+\infty}{\longrightarrow}_{\mathcal{L}} N(0, I^{-1}(\theta)).$$

Proof By Taylor's expansion theorem:

$$\frac{\partial}{\partial\theta_k}\rho_n(X,\theta) = \frac{\partial}{\partial\theta_k}\rho_n(X,\widehat{\theta}_n) + \sum_j \frac{\partial^2}{\partial\theta_j\partial\theta_k}\rho_n(X,\theta_n^*)(\theta_j - \widehat{\theta}_{n,j}) = \sum_j \frac{\partial^2}{\partial\theta_j\partial\theta_k}\rho_n(X,\theta_n^*)(\theta_j - \widehat{\theta}_{n,j})$$

where $|\theta_{n,j}^* - \theta_j| \leq |\widehat{\theta}_{n,j} - \theta_j|$ for each j.
It follows from assumption 4. that $\theta_n^* \rightarrow \theta$ and hence from the WLLN that

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\rho_n(X,\theta_n^*) \longrightarrow_{\mathbb{P}_\theta} \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\rho(X,\theta)\right] = J_{j,k}(\theta)$$

Since $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_k}\rho_n(X,\theta)\right] = 0$, it follows that $K(\theta)$ is the covariance matrix of $\nabla\rho(X,\theta)$ and hence, from the central limit theorem, that

$$\sqrt{n}\nabla\rho_n(X,\theta) \longrightarrow_{\mathbb{P}_\theta} N(0, K(\theta)).$$

If $V \sim N(0, K(\theta))$, then $J^{-1}(\theta)V \sim N(0, J^{-1}(\theta)K(\theta)J^{-1}(\theta))$, from which it follows that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \overset{n\rightarrow+\infty}{\longrightarrow} N(0, J^{-1}(\theta)K(\theta)J^{-1}(\theta)).$$

The result for maximum likelihood follows directly, since for maximum likelihood $I = J = K$. □

Summary

- Definition of consistency
- Multinomial setting: consistent estimators.
- General setting: conditions for consistent estimation via moment method.
- Consistency of MLE for natural parameters of exponential family via moment method.
- Consistency of minimum contrast estimators.
- The Delta Method: univariate and multivariate.

Tutorial 14

1. Let $X_1, \ldots, X_n$ be i.i.d. $U(0, \theta)$; that is, the density is therefore:

$$p(x;\theta) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x)$$

Let $l(\theta; x) = -\log p(x;\theta)$.

(a) Show that $\frac{d}{d\theta}l(\theta, x) = \frac{1}{\theta}$ for $\theta > x$ and is undefined for $\theta \leq x$. If $X \sim U(0,\theta)$, conclude that $\frac{d}{d\theta}l(\theta, X)$ is defined with $\mathbb{P}_\theta$ probability 1, but that

$$\mathbb{E}_\theta\left[\frac{d}{d\theta}l(\theta; X)\right] = \frac{1}{\theta} \neq 0.$$

(b) Recall that $\widehat{\theta}_{ML} = \max\{X_1, \ldots, X_n\}$. Show that: $n(\theta - \widehat{\theta}) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}_\theta} \text{Exp}(1/\theta)$ ($\mathcal{L}_\theta$ denotes the law when the parameter value is θ).

2. Suppose $\lambda : \mathbb{R} \to \mathbb{R}$ satisfies $\lambda(0) = 0$, is bounded and has bounded second derivative λ''. Show that if $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}[X_1] = \mu$ and $\mathbf{V}(X_1) = \sigma^2 < +\infty$, then

$$\left|\sqrt{n}\mathbb{E}\left[\lambda(|\overline{X} - \mu|)\right] - \lambda'(0)\sigma\sqrt{\frac{2}{\pi}}\right| \overset{n\to+\infty}{\longrightarrow} 0$$

3. Let $V_n \sim \chi^2_n$. Show that $(\sqrt{V_n} - \sqrt{n}) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} N(0, \frac{1}{2})$ ($\mathcal{L}$ denotes law).

4. Suppose that $X_1, \ldots, X_n$ are i.i.d. variables each with probability function

$$p_X(0) = \theta^2 \qquad p_X(1) = 2\theta(1-\theta) \qquad p_X(2) = (1-\theta)^2$$

(a) Find a and b (in terms of n and θ) such that $Z_n = \frac{\overline{X}-a}{b} \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} N(0,1)$.

(b) Find c and d (in terms of n and θ) such that $Y_n = \frac{\sqrt{\overline{X}}-c}{d} \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} N(0,1)$.

5. Let $X_1, \ldots, X_n$ be a sample from a population with mean μ and variance $\sigma^2 < +\infty$. Let h be a function and let $h^{(j)}$ denote its jth derivative. Suppose that h has a second derivative continuous at μ and that $h^{(1)}(\mu) = 0$.

(a) Show that $\sqrt{n}(h(\overline{X}) - h(\mu)) \overset{n\to+\infty}{\longrightarrow} 0$, while $n\left(h(\overline{X}) - h(\mu)\right) \overset{n\to+\infty}{\longrightarrow} \frac{1}{2}h^{(2)}(\mu)\sigma^2 V$ where $V \sim \chi^2_1$.

(b) Use part (a) to show that when $\mu = \frac{1}{2}$, then

$$n\left(\overline{X}(1-\overline{X}) - \mu(1-\mu)\right) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} -\sigma^2 V \qquad V \sim \chi^2_1$$

6. Show that if $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ and $S^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2$, then

$$\sqrt{n}\begin{pmatrix} \overline{X} - \mu \\ S^2 - \sigma^2 \end{pmatrix} \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right)$$

7. Let $X_{ij} : i = 1, \dots, p, j = 1, \dots, k$ be independent with $X_{ij} \sim N(\mu_i, \sigma^2)$.

(a) Show that the MLEs of μ_i and σ^2 are:

$$\overline{\mu}_i = \frac{1}{k}\sum_{j=1}^{k} X_{ij} \qquad \widehat{\sigma^2} = \frac{1}{kp}\sum_{i=1}^{p}\sum_{j=1}^{k}(X_{ij} - \widehat{\mu}_i)^2$$

(b) Show that if k is fixed and $p \to +\infty$, then

$$\widehat{\sigma}^2 \overset{p\to+\infty}{\longrightarrow}_{\mathcal{L}} \left(1 - \frac{1}{k}\right)\sigma^2.$$

That is, the MLE $\widehat{\sigma}^2$ is not consistent.

Answers

1. (a) Let $l(\theta)$ denote the log likelihood. Then:

$$l(\theta) = \begin{cases} -\log\theta & 0 \leq x \leq \theta \\ -\infty & \text{other} \end{cases}$$

so

$$\frac{d}{d\theta} l(\theta) = \begin{cases} -\frac{1}{\theta} & 0 \leq x \leq \theta \\ \text{undefined} & \text{otherwise} \end{cases}$$

so it is defined with probability 1 and

$$\mathbb{E}_\theta \left[\frac{d}{d\theta} l(\theta; X) \right] = -\frac{1}{\theta} \neq 0.$$

(b) Let $Y = \max\{X_1, \ldots, X_n\}$, then

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X_1 \leq y)^n = \begin{cases} 0 & y \leq 0 \\ \left(\frac{y}{\theta}\right)^n & 0 < y \leq \theta \\ 1 & y > \theta \end{cases}$$

Let $W = n(\theta - Y)$, then

$$\mathbb{P}(W \leq t) = \mathbb{P}(n(\theta - Y) \leq t) = \mathbb{P}(Y \geq \theta - \frac{t}{n}) = 1 - \left(1 - \frac{t}{n\theta}\right)^n \overset{n\to+\infty}{\longrightarrow} 1 - e^{-t/\theta}$$

so that $n(\theta - \widehat{\theta}_{ML}) \overset{n\to+\infty}{\longrightarrow}_{\mathcal{L}} \text{Exp}(1/\theta)$.

2. Taylor's expansion theorem gives:

$$\sqrt{n}\lambda(|\overline{X} - \mu|) = \sqrt{n}\lambda'(0)|\overline{X} - \mu| + \sqrt{n}\,\frac{\lambda''(Y)}{2}|\overline{X} - \mu|^2$$

where $0 \leq Y \leq |\overline{X} - \mu|$. Using $\sup_x |\lambda''(x)| \leq K$, it follows that:

$$|\mathbb{E}[\frac{\sqrt{n}\lambda''(Y)}{2}|\overline{X} - \mu|^2]| \leq \frac{\sigma^2}{2\sqrt{n}}K$$

since

$$\mathbb{E}\left[\left|\overline{X} - \mu\right|^2\right] = \mathbf{V}(\overline{X}) = \frac{\sigma^2}{n}.$$

By the central limit theorem,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \overset{n\to+\infty}{\longrightarrow} N(0,1).$$

Let $Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$. Then from the definition of 'convergence in law', for any *bounded* function f,

$$\mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)] \qquad Z \sim N(0,1)$$

For any non negative $N < +\infty$,

$$\mathbb{E}[|Z_n|\mathbf{1}_{\{|Z_n|\geq N\}}] \leq \mathbb{E}[|Z_n|^2]^{1/2}\mathbb{P}(|Z_n| \geq N) = \mathbb{P}(|Z_n| \geq N) \rightarrow \mathbb{P}(|Z| \geq N),$$

so that

$$|\mathbb{E}[|Z_n|] - \mathbb{E}[|Z|]| \leq |\mathbb{E}[|Z_n| \wedge N] - \mathbb{E}[|Z| \wedge N]| + (\mathbb{P}(|Z_n| \geq N) + \mathbb{P}(|Z| \geq N)).$$

while $\mathbb{P}(|Z_n| \geq N) \overset{n\rightarrow+\infty}{\longrightarrow} \mathbb{P}(|Z| \geq N)$. It follows that (taking limit in n first and then limit in N),

$$\lim_{n\rightarrow+\infty} |\mathbb{E}[|Z|] - \mathbb{E}[|Z_n|]| \leq 2\mathbb{P}(|Z| \geq N) \overset{N\rightarrow+\infty}{\longrightarrow} 0.$$

Therefore

$$\sqrt{n}\mathbb{E}\left[\frac{|\overline{X} - \mu|}{\sigma}\right] \overset{n\rightarrow+\infty}{\longrightarrow} \sqrt{\frac{2}{\pi}}\int_0^\infty xe^{-x^2/2}dx = \sqrt{\frac{2}{\pi}},$$

from which the result follows.

3. This is a straightforward application of the Delta method. $V_n = Z_1^2 + \ldots + Z_n^2$ where $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ variables. Let $h(x) = x^{1/2}$, then $h'(x) = \frac{1}{2}x^{-1/2}$, so $(h'(1))^2 = \frac{1}{4}$.

   Let $Y_j = Z_j^2$ and $\overline{Y} = \frac{1}{n}\sum_{j=1}^n Z_j^2$. Since $\mathbb{E}[Z_1^2] = 1$ and $\mathbf{V}(Z_1^2) = 2$, it follows from the Delta method that

$$\sqrt{V_n} - \sqrt{n} = \sqrt{n}\left(\overline{Y} - 1\right) \overset{n\rightarrow+\infty}{\longrightarrow} N(0, 2 \times \frac{1}{4}) = N(0, \frac{1}{2}).$$

4.

$$\mu = \mathbb{E}[X_1] = 2\theta(1-\theta) + 2(1-\theta)^2 = 2\theta - 2\theta^2 + 2 - 4\theta + 2\theta^2 = 2(1-\theta)$$

$$\mathbb{E}[X_1^2] = 2\theta(1-\theta) + 4(1-\theta)^2 = 2(1-\theta)(\theta + 2 - 2\theta) = 2(1-\theta)(2-\theta)$$

so that

$$\sigma^2 = \mathbf{V}(X_1) = 2(1-\theta)(2-\theta) - 4(1-\theta)^2 = 2(1-\theta)(2-\theta-2+2\theta) = 2\theta(1-\theta).$$

(a) This is just the central limit theorem:

$$\frac{\overline{X} - 2(1-\theta)}{\sqrt{2\theta(1-\theta)/n}} \overset{n\rightarrow+\infty}{\longrightarrow} N(0,1)$$

$$a = 2(1-\theta) \qquad b = \sqrt{\frac{2\theta(1-\theta)}{n}}.$$

(b) Delta method: $h(x) = x^{1/2}$, $h'(x) = \frac{1}{2x^{1/2}}$, so that $h'(2(1-\theta)) = \frac{1}{2^{3/2}(1-\theta)^{1/2}}$.

$$\sqrt{n}(\sqrt{\overline{X}} - \sqrt{2(1-\theta)}) \overset{n\rightarrow+\infty}{\longrightarrow} N(0, \frac{2\theta(1-\theta)}{2^3(1-\theta)})$$

so

$$\frac{\sqrt{\overline{X}} - \sqrt{2(1-\theta)}}{\sqrt{\theta/4n}} \overset{n\rightarrow+\infty}{\longrightarrow} N(0,1).$$

$$c = \sqrt{2(1-\theta)} \qquad d = \frac{1}{2}\sqrt{\frac{\theta}{n}}.$$

5. (a) By Taylor's expansion,

$$h(\mu + (\overline{X}-\mu)) = h(\mu) + (\overline{X}-\mu)h^{(1)}(\mu) + \frac{(\overline{X}-\mu)^2}{2}h^{(2)}(\mu+z) \qquad |z| \leq |\overline{X}-\mu|$$

so

$$n(h(\overline{X}) - h(\mu)) = \sigma^2\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\right)^2 h^{(2)}(\mu+z) \qquad |z| \leq |\overline{X}-\mu|.$$

giving

$$n(h(\overline{X}) - h(\mu)) \overset{n\rightarrow+\infty}{\longrightarrow_{\mathcal{L}}} \sigma^2 h^{(2)}(\mu)V \qquad V \sim \chi^2_1$$

It follows directly that $\sqrt{n}(h(\overline{X}) - h(\mu)) \overset{n\rightarrow+\infty}{\longrightarrow_{\mathcal{L}}} 0$.

(b) $h(x) = x(1-x)$, $h^{(1)} = 1 - 2x = 0$ if $x = \frac{1}{2}$.

$$h^{(2)}(x) = -2 \Rightarrow h^{(2)}(\frac{1}{2}) = -2$$

$$n(\overline{X}(1-\overline{X}) - \frac{1}{4}) \overset{n\rightarrow+\infty}{\longrightarrow} -\sigma^2 V \qquad V \sim \chi^2_1$$

6. Firstly, asymptotic normality. This may be seen by expressing

$$S^2 = \frac{1}{n-1}\sum_{j=1}(X_j-\mu)^2 - \frac{n}{n-1}\left(\overline{X}-\mu\right)^2$$

$$\mathbf{V}((\overline{X}-\mu)^2) = \frac{\sigma^2}{n^2}\mathbf{V}\left(\frac{n(\overline{X}-\mu)^2}{\sigma^2}\right) = \frac{2\sigma^2}{n^2}$$

since $\frac{n(\overline{X}-\mu)^2}{\sigma^2} \sim \chi^2_1$ which has variance 2. Let $Y_i = (X_i-\mu)^2 - \sigma^2$. Then

$$\sqrt{n}(S^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} Y_j + \epsilon_n$$

where $\epsilon_n \rightarrow_{\mathbb{P}} 0$. It follows directly from the central limit theorem (no delta method required here) that the random vector $\sqrt{n}\begin{pmatrix}\overline{X}\\ S^2-\sigma^2\end{pmatrix}$ is asymptotically normal. It only remains to compute the covariance matrix. Firstly, $\mathbf{V}\left(\sqrt{n}(\overline{X}-\mu)\right) = \sigma^2$, as required. Secondly, $\overline{X}$ and S^2 are independent, giving the 0 covariance terms. This may be seen as follows: consider the random vector $(\overline{X}, X_1 - \overline{X}, \ldots, X_n - \overline{X})$. This is a normal random vector. Then

$$\mathbf{C}(\overline{X}, X_j - \overline{X}) = \mathbf{C}(X_j, \overline{X}) - \mathbf{V}(\overline{X}) = \frac{1}{n}\mathbf{V}(X_j) - \frac{\sigma^2}{n} = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.$$

It follows that

$$\overline{X} \perp \{(X_1 - \overline{X}), \dots, (X_n - \overline{X})\}$$

and hence that

$$\overline{X} \perp \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

Thirdly,

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

hence

$$\mathbf{V}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$$

so that asymptotically, $\frac{\frac{(n-1)S^2}{\sigma^2}-(n-1)}{\sqrt{2(n-1)}} = \frac{\sqrt{n-1}(S^2-\sigma^2)}{\sqrt{2}\sigma^2} \overset{n\to+\infty}{\longrightarrow_{\mathcal{L}}} N(0,1)$ giving

$$\sqrt{n}(S^2 - \sigma^2) \overset{n\to+\infty}{\longrightarrow_{\mathcal{L}}} N(0, 2\sigma^4)$$

and the result follows.

7. (a)

$$L(\mu_1, \dots, \mu_p, \sigma; (x_{ij})) = \frac{1}{(2\pi)^{kp/2}\sigma^{kp}} \exp\left\{-\frac{1}{\sigma^2}\sum_{ij}(x_{ij} - \mu_i)^2\right\}$$

$\widehat{\mu_i}$ from minimising $\sum_{i,j}(x_{ij} - \mu_i)^2$ which gives

$$\widehat{\mu_i} = \frac{1}{k}\sum_{j=1}^{k} X_{ij}$$

$\widehat{\sigma^2}$ from maximising

$$-\frac{kp}{2}\log(\sigma^2) - \frac{1}{(\sigma^2)}\sum_{i=1}^{p}\left(\sum_{j=1}^{k}(x_{ij} - \mu_i)^2\right)$$

giving

$$\widehat{\sigma^2} = \frac{1}{kp}\sum_{i=1}^{p}\sum_{j=1}^{k}(X_{ij} - \widehat{\mu_i})^2$$

(b) For each i,

$$\frac{\sum_{j=1}^{k}(X_{ij} - \widehat{\mu_i})^2}{\sigma^2} \sim \chi^2_{k-1}$$

and these are independent, so

$$\frac{\sum_{i=1}^{p}\sum_{j=1}^{k}(X_{ij} - \widehat{\mu_i})^2}{\sigma^2} = \frac{kp\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{p(k-1)}$$

$$\mathbf{V}\left(\frac{\widehat{\sigma^2}}{\sigma^2}\right) = \frac{2p(k-1)}{k^2p^2} = \frac{2}{p}(\frac{1}{k} - \frac{1}{k^2}) \overset{p\to+\infty}{\longrightarrow_{\mathcal{L}}} 0$$

Furthermore,

$$\mathbb{E}[\widehat{\sigma^2}] = \frac{p(k-1)}{kp}\sigma^2 = \left(1 - \frac{1}{k}\right)\sigma^2$$

so, by Chebyshev,

$$\widehat{\sigma^2} \overset{p\to+\infty}{\longrightarrow}_{\mathcal{L}} \left(1 - \frac{1}{k}\right)\sigma^2.$$

Literature Cited

[1] Andersen, E. [1970] *Sufficiency and Exponential Families for Discrete Sample Spaces* Journal of the American Statistical Association, Vol. 65, No. 331 pp 1248–1255.

[2] Darmois, G. [1935] *Sur les lois de probabilites a estimation exhaustive* C.R. Acad. Sci. Parisvol. 200 pp.1265–1266.

[3] Feller, W. [1948] *On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions* Ann. Math. Statist. vol. 19 no. 2 pp. 177 - 189

[4] Fisher, R.A. [1922]*On the Mathematical Foundations of Theoretical Statistics* Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 222, pp. 309-368

[5] Kolmogoroff, A. [1933] *Sulla determinazione empirica di una legge di distribuzione* Inst. Ital. Attuari, Giorn., vol. 4 pp. 1 - 11

[6] Koopman, B [1936] *On distribution admitting a sufficient statistic* Transactions of the American Mathematical Society, Vol. 39, No. 3 pp. 399–409.

[7] Pitman, E.; Wishart, J. [1936] *Sufficient statistics and intrinsic accuracy* Mathematical Proceedings of the Cambridge Philosophical Society vol. 32 no. 4 pp. 567–579

[8] Shannon, C.E. [1948]*A Mathematical Theory of Communication* Bell System Tech. Journal vol. 27 no. 3 pp. 379–423

[9] Smirnov, N. [1939]*On the estimation of the discrepancy between empirical curves of distribution for two independent samples* Bulletin Mathématique de l'Université de Moscou, vol. 2, fasc. 2.

[10] Wijsman, R.A. [1973]*On the Attainment of the Cramér Rao Lower Bound* Ann. Math. Statis., vol. 1, pp. 538 - 542

[11] Wu, C.F.J. [1983] *On the Convergence Properties of the EM Algorithm* Ann. Statist., vol. 11 pp. 95 - 103.