STAT3923: Advanced Statistical Inference

Charles Christopher Hyland

Semester 2 2019

Abstract

Thank you for stopping by to read this. These are notes collated from lectures and tutorials as I took this course.

Contents

STAT3923: Advanced Statistical Inference

1. Probability Theory

1.1 Probability Theory

1.1.1 Probability Theory Introduction

Definition 1.1 *(Random Variable). Let Ω be a sample space. A random variable $X : \Omega \to \mathbb{R}$ is a real-valued function defined over elements of Ω.*

Definition 1.2 *(Cumulative Distribution Function). For any random variable, its distribution is characterised by the cumulative distribution function*

$$F(x) = P(X \leq x)$$

for $-\infty < x < \infty$.

Lemma 1.3 *The following are properties of the CDF F(x)*

1. *$F(a) \leq F(b)$ for $a < b$;*
2. *$\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$;*
3. *F(x) is right continuous.*

Definition 1.4 *(Probability Mass Function). Let X be a discrete random variable taking on values $x_1 < x_2 < ...$ The PMF for the random variable X is defined as*

$$f(x_i) = P(X = x_i)$$

where $\sum_{x_i} f(x_i) = 1$.

Theorem 1.5 *Let X be a random variable. Then, $f(x_i) = F(x_i) - F(x_{i-1})$ and $F(x) = \sum_{x_i \leq x} f(x_i)$.*

Definition 1.6 *(Probability Density Function). Let X be a continuous random variable. The PDF is defined as*

$$f(x) = \frac{dF(x)}{dx}$$

where $\int_{-\infty}^{\infty} f(x)dx = 1$. Furthermore, we have that

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Definition 1.7 *(L^1-space). We denote the set of all first integrable random variables as*

$$L^1 = \{X : \Omega \to \mathbb{R} : ||X||_1 < \infty\}.$$

Definition 1.8 *(Expectation). Let $X \in L^1$. Then we define the expectation of a random variable as*

$$E(X) = \begin{cases} \sum_X xf(x) & \textit{(Discrete)} \\ \int_{-\infty}^{\infty} xf(x)dx & \textit{(Continuous)} \end{cases}$$

Lemma 1.9 *Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$. Then the random variable $Y = g(X)$ is a random variable with PMF/PDF f_Y.*

Definition 1.10 *(r-th moment) Let $X \in L^r$. Then we define the r-th moment as*

$$E(X^r) = \begin{cases} \sum_X x^r f(x) & \textit{(Discrete)} \\ \int_{-\infty}^{\infty} x^r f(x)dx & \textit{(Continuous)} \end{cases}$$

Definition 1.11 *(Variance). Let $X \in L^2$. Then we define the variance as*

$$Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2.$$

Definition 1.12 *(General Expectation) Let $X \in L^r$. Then we define the r-th moment as*

$$E(g(X)) = \begin{cases} \sum_X g(x)f(x) & \textit{(Discrete)} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \textit{(Continuous)} \end{cases}$$

Proposition 1.13 *Let X and Y be a random variables and a,b be constants. We have the following properties*

1. $E(aX + b) = aE(X) + b$;
2. $E(X + Y) = E(X) + E(Y)$;
3. $Var(aX + b) = a^2 Var(X)$;
4. $E(XY) = E(X)E(Y)$ *(if X and Y are independent)*;
5. $Var(X + Y) = Var(X) + Var(Y)$ *(if X and Y are independent).*

STAT3923: Advanced Statistical Inference

2. Random variables and distributions

2.1.2 Discrete Random Variables

A discrete random variable is a random variable whose range is finite or countably infinite.

Definition 2.14 *(Bernoulli Distribution). A random variable X has a Bernoulli distribution and it is referred to as a Bernoulli random variable if and only if its probability distribution is given by*

$$f(x;\theta) = \theta^x(1-\theta)^{1-x} \quad x \in \{0,1\}.$$

Definition 2.15 *(Binomial). A random variable X has a Binomial distribution and it is referred to as a Binomial random variable if and only if its probability distribution is given by*

$$f(x;n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad x = 0,1,...,n.$$

Theorem 2.16 *Let X be a Binomial random variable. Then*

$$f(x;n,\theta) = f(n-x;n,1-\theta).$$

Theorem 2.17 *The mean and variance of the Binomial distribution are*

$$E(X) = n\theta$$

$$Var(X) = n\theta(1-\theta).$$

Definition 2.18 *(Negative Binomial Distribution). A random variable X has a negative binomial distribution and it is referred to as a negative binomial random variable if and only if*

$$f(x;k,\theta) = \binom{x-1}{k-1}\theta^k(1-\theta)^{x-k} \quad x = k, k+1, k+2, ...$$

Theorem 2.19 *The mean and the variance of the negative binomial distribution are*

$$\mu = \frac{k}{\theta}$$

$$\sigma^2 = \frac{k}{\theta}(\frac{1}{\theta} - 1).$$

Definition 2.20 *(Geometric Distribution). A random variable X has a Geometric distribution and it is referred to as a Geometric random variable if and only if its probability distribution is given by*

$$f(x;\theta) = \theta(1-\theta)^{x-1} \quad x = 1,2,3,...$$

Theorem 2.21 *The mean and variance of the Geometric random variable are*

$$\mu = \frac{1}{p}$$

$$\sigma^2 = \frac{1-p}{p^2}.$$

Definition 2.22 *(Hypergeometric Distribution). A random variable X has a Hypergeometric distribution and it is referred to as a hypergeometric random variable if and only if its probability distribution is given by*

$$f(x; n, N, M) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad x = 0, 1, 2, ...$$

and $x \leq M$ and $n - x \leq N - M$. Here, M are the number of successes and N - M as failures.

Definition 2.23 *(Poisson). A random variable X has a Poisson distribution and it is referred to as a Poisson random variable if and only if its probability distribution is given by*

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, ...$$

for $k \in \mathbb{N}^+$.

Remark 2.24 *The Poisson random variable expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known constant rate (λ) and are independent of the time of the last event.*

Theorem 2.25 *The mean and variance of the Poisson random variable are*

$$\mu = \lambda$$

$$\sigma^2 = \lambda.$$

Theorem 2.26 *(Poisson Limit Theorem). Let p_n be a sequence of real numbers in [0,1] such that the sequence $np_n \to \lambda < \infty$. Then*

$$\lim_{n\to\infty; p_n \to 0} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}.$$

2.1.3 Continuous Random Variables

Definition 2.27 *(Uniform Distribution). A random variable X has a uniform distribution and it is referred to as a continuous uniform random variable if and only if its probability density is given by*

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha < x < \beta \\ 0 & otherwise. \end{cases}$$

Theorem 2.28 *The mean and variance of the uniform distribution are given by*

$$\mu = \frac{\alpha+\beta}{2}$$

$$\sigma^2 = \frac{1}{2}(\beta-\alpha)^2.$$

Definition 2.29 *(Gamma Function). The gamma function is defined for any complex number with a positive real part. It is defined as*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dy$$

for $\alpha > 0$.

Theorem 2.30 *The gamma function satisfies the recursion formula*

$$\Gamma(\alpha+1) = (\alpha)\Gamma(\alpha).$$

Definition 2.31 *(Gamma Distribution). A random variable X has a Gamma distribution and it is referred to as a Gamma random variable if and only if its density is given by*

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha}x^{\alpha-1}e^{-x/\beta} & x > 0 \\ 0 & otherwise \end{cases}$$

where $\alpha > 0$ *and* $\beta > 0$.

Theorem 2.32 *The mean and variance of the gamma distribution are given by*

$$\mu = \alpha\beta$$

$$\sigma^2 = \alpha\beta^2$$

The exponential and chi-square distribution are special cases of the gamma distribution.

Definition 2.33 *(Exponential Distribution). A random variable X has an exponential distribution and it is referred to as an exponential random variable if and only if its probability density is given by*

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}e^{-x/\theta} & x > 0 \\ 0 & elsewhere \end{cases}$$

for $\theta > 0$.

Remark 2.34 *The exponential distribution is the Gamma distribution for* $\alpha = 1$.

Theorem 2.35 *The mean and variance of the exponential distribution are given by*

$$\mu = \theta$$

$$\sigma^2 = \theta^2.$$

Definition 2.36 *(Chi-Square Distribution). A random variable X has a chi-square distribution and it is referred to as a chi-square random variable if and only if its probability density is given by*

$$f(x,\nu) = \begin{cases} \frac{1}{2^{\nu/2}\gamma(\nu/2)}x^{\frac{\nu-2}{2}}e^{-\frac{x}{2}} & x > 0 \\ 0 & elsewhere. \end{cases}$$

Remark 2.37 *The chi-square distribution is the Gamma distribution for* $\alpha = \nu/2$ *and* $\beta = 2$.

Theorem 2.38 *The mean and variance of the chi-square distribution are given by*

$$\mu = \nu$$

$$\sigma^2 = 2\nu.$$

Theorem 2.39 *If $Z_i \sim N(0,1)$ are i.i.d, then $X = \sum_{i=1}^{\nu} Z_i^2$, then $X \sim \chi^2_\nu$.*

Definition 2.40 *(Beta Distribution). A random variable X has a Beta distribution and it is referred to as a Beta random variable if and only if its probability density is given by*

$$f_X(x;\alpha,\beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & x \in (0,1) \\ 0 \quad elsewhere \end{cases}$$

where $\alpha > 0$ and $\beta > 0$.

Theorem 2.41 *The mean and variance of the beta distribution are given by*

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Definition 2.42 *(Normal Distribution). A random variable X has a normal distribution and it is referred to as a normal random variable if and only if its probability density is given by*

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad -\infty < x < \infty$$

where $\sigma > 0$.

Theorem 2.43 *(Linear Transformation of the Normal). Let $Z \sim N(0,1)$. Define $X = \mu + \sigma Z$. Then, $X \sim N(\mu, \sigma^2)$.*

Definition 2.44 *(Standard Normal Distribution). The normal distribution with $\mu = 0$ and $\sigma = 1$ is referred to as the standard normal distribution.*

Theorem 2.45 *(Binomial Approximation To Normal). If X is a random variable having a binomial distribution with the parameters n and θ, then the MGF of*

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

approaches to that of the standard normal distribution when $n \to \infty$.

Lemma 2.46 *Let X be a continuous nonnegative random variable. Then we have that*

$$E(X) = \int_0^\infty P(X > x)dx.$$

STAT3923: Advanced Statistical Inference

3. Moment Generating Functions

3.2 Moment Generating Functions

3.2.1 Moment Generating Functions Introduction

We are interested in MGFs for three reasons. A MGF as a real function that uniquely determines its associated probability distribution, and its derivatives at zero are equal to the moments of the random variable. Finally, a MGF is useful for finding the distribution of sums of functions.

Definition 1: Moment Generating Function

Let X be a random variable, then $Y = e^{tX} \geq 0$, so $E(Y)$ is well defined. The MGF of the random variable X is defined as

$$M(t) = E(e^{tX}) = \sum_x e^{tx} f_X(x)$$

for all t for which the right hand side is finite.

Remark 3.47 *We know that for $t = 0$, the MGF exists as $M(0) = 1 < \infty$. If X is a discrete RV, then the MGF is $M(t) = \sum_i e^{tX_i} P_X(x_i)$.*

Theorem 1: Computing moments with MGFs

If there exists a $\delta > 0$ such that $M(t) < \infty$ for all $t \in (-\delta, \delta)$ then for all $n \in \mathbb{N}$, we have

$$M^n(0) = \mathbb{E}(X^n)$$

and exist. The MGF is infinitely differentiable at 0.

Proof: Assuming there is an interval in which we can interchange differentiation and expectation, we have

$$\frac{d}{dt} E[e^{tX}] = E\Big[\frac{d}{dt} e^{tX}\Big] = E\Big[Xe^{tX}\Big].$$

Then, if we let t = 0 in the above, we get

$$E\Big[X\Big] = M'(0).$$

■

Theorem 2: Equality of distributions

Let F and G be CDFs and suppose that there exists $\delta > 0$ such that for all $t \in (-\delta, \delta)$, the MGFs $M_F(t) = M_G(t) < \infty$. Then $F = G$. It follows that all the moments of F and G exist and are equal.

Remark 3.48 *The converse to the above theorem is false. All the moment of F and G can exist, and be equal, yet $F \neq G$.*

Proposition 1: Linear transformation of MGFs

Let X be a random variable possessing a MGF $M_X(t)$. Define the linear transformation

$$Y = a + bX$$

where $a, b \in \mathbb{R}$ are two constants and $b \neq 0$. Then the random variable Y posseses a MGF $M_Y(t)$ and

$$M_Y(t) = exp(at)M_X(bt).$$

Remark 3.49 *Not every random variable posseses a moment generating function. However, every random variable posseses a characteristic function.*

Definition 3.50 *(Characteristic Function). Let X be a random variable. Let $i = \sqrt{-1}$ be the imaginary unit. The function $\phi : \mathbb{R} \to \mathbb{C}$ is defined by*

$$\phi_X(t) = \mathbb{E}[exp(itX)]$$

is called the characteristic function of X.

Proposition 2: Independence of MGFs

Let $X_1, ..., X_n$ be n mutually independent random variables. Let Z be their sum:

$$Z = \sum_{i=1}^{n} X_i.$$

Then, the MGF of Z is the product of the MGFs of $X_1, ..., X_n$:

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

Theorem 3.51 *(Continuity Theorem). Let F_n be CDFs with MGFs M_n, and let F be a CDF with MGF M, and suppose that there exists $\delta > 0$ such that $M_n(t) \to_n M(t)$ for all $t \in (-\delta, \delta)$. Then $F_n(x) \to F(x)$ for all x where F is continuous at x.*

STAT3923: Advanced Statistical Inference

4. Convergence Concepts

4.2.2 Convergence Concepts

We are interested in defining what it means for random variables to converge.

Definition 2: Convergence in probability

Let $\{X_n\}$ and X be jointly distributed random variables. We say that $X_n \xrightarrow{p} X$ in probability if for all $\epsilon > 0$, we have that

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

. We say that $X_n \xrightarrow{p} X$.

Definition 3: Convergence Almost Surely

Let $\{X_n\}$ and X be jointly distributed random variables. We say that $X_n \xrightarrow{a.s.} X$ strong or almost surely if we have that

$$P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1.$$

Remark 4.52 *Recall that random variables are real-valued functions defined on the sample space Ω. Let $s \in \Omega$ be sample points. A sequence of functions $X_n(s)$ converges to X(s) for all $s \in \Omega$ except for $s \in \mathcal{N}$ where $\mathcal{N} \subset \Omega$ and $P(\mathcal{N}) = 0$. That is, we have pointwise convergence of a sequence of functions except convergence need not occur on a set with probability 0.*

Theorem 4.53 *Convergence almost surely $\xrightarrow{a.s.}$ implies convergence in probability $\xrightarrow{p}$.*

Definition 4: Convergence in distribution

Let $\{X_n\}$ and X be jointly distributed random variables. We say that $X_n \xrightarrow{d} X$ in distribution if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all x where F_X is continuous at x. We say that $X_n \xrightarrow{d} X$.

Remark 4.54 *Note that convergence in distribution is phrased in terms of the CDFs. Hence, it is the CDFs that converges, not the random variables when we speak of convergence in distributions.*

Theorem 4.55 *Convergence in probability $\xrightarrow{p}$ implies convergence in distribution $\xrightarrow{d}$.*

The CLT states that the **sample mean** has a distribution which is approximately normal with mean μ and variance σ^2. That is, probability statements about the sample mean can be approximated using a normal distribution. **Not the random variable itself.**

Theorem 3: Central Limit Theorem

Suppose X_i are i.i.d random variables with $\sigma^2 = Var(X_i) < \infty$ and $\mu = E(X_i)$. Then with $S_n = \sum_{i=1}^{n} X_i$ for all $x \in \mathbb{R}$

$$P(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x) \xrightarrow{d} \phi(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt.$$

Let $Y_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$, then the CLT states that

$$F_{Y_n}(x) \xrightarrow{d} F_Z(x)$$

for all $x \in \mathbb{R}$ where $Z \sim N(0,1)$. In other words, $Y_n \xrightarrow{d} Z$ in distribution.

Theorem 4.56 *(Markov's Inequality). Let X be a non-negative random variable and suppose $\mathbb{E}[X]$ exists. For any $t > 0$*

$$P(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$

Theorem 4: Chebychev's Inequality

Assume the random variable X has a finite second moment σ. Then, for any real number $k > 0$,

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

Remark 4.57 *This a useful theorem for proving convergence in probability to a constant.*

Theorem 5: Weak Law of Large Numbers

Let $X_1, ..., X_n$ be a sequence of i.i.d random variables. Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Define the random variable $S_n = \frac{X_1 + ... + X_n}{n}$, then

$$S_n \xrightarrow{p} \mu$$

. That is, for any $\epsilon > 0$

$$\lim_{n \to \infty} P(|S_n - \mu| \geq \epsilon) = 0.$$

That is, $S_n \xrightarrow{p} \mu$.

That is, the mean of a large sample is close to the mean of the distribution. Hence, the distribution of the sample mean becomes more concentrated around μ as n gets large.

Theorem 4.58 *(Strong Law of Large Numbers). Let $X_1, ..., X_n$ be a sequence of i.i.d random variables. Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Define the random variable $S_n = \frac{X_1 + ... + X_n}{n}$, then*

$$S_n \xrightarrow{a.s.} \mu$$

. *That is, for any* $\epsilon > 0$

$$P(\lim_{n\to\infty} |S_n - \mu| \leq \epsilon) = 1.$$

That is, $S_n \xrightarrow{a.s.} \mu$.

STAT3923: Advanced Statistical Inference

5. Further Limit Laws

5.2.3 Further Limit Laws

Theorem 5.59 *Let $\{X_n\}$, $\{Y_n\}$ be sequences of random variables. If $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{p} d$, then*

$$X_n + Y_n \xrightarrow{p} c + d$$

where c and d are constants.

Theorem 5.60 *Let $\{X_n\}$ be a sequence of random variables. Suppose for a function $g(.)$, we have that $\lim_{x\to c} g(x) = \ell$ exists and is finite for a constant ℓ. If $X_n \xrightarrow{p} c$, then*

$$g(X_n) \xrightarrow{p} \ell.$$

Corollary 5.61 *If $g(.)$ is **continuous** at c, then*

$$g(X_n) \xrightarrow{p} g(c).$$

*If $h(.)$ is **differentiable** at c, then*

$$\frac{h(X_n) - h(c)}{X_n - c} \xrightarrow{p} h'(c).$$

STAT3923: Advanced Statistical Inference

6. Asymptotics

6.2.4 Asymptotics

Lemma 6.62 *Let X have a CDF $F(.)$ and let x be a continuity point of the CDF $F(.)$. Suppose $X_n \xrightarrow{d} X$. Then*

$$P(X_n = x) \to 0.$$

Proposition 6.63 *The sequence of random variables $X_1, X_2, ...,$ converges in probability to a constant c if and only if the sequence converges in distribution to c.*
That is, the statement

$$P(|X_n - c| > \epsilon) \xrightarrow{p} 0 \quad \text{for every } \epsilon > 0$$

is equivalent to

$$P(X_n \leq x) \xrightarrow{d} \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x > c. \end{cases}$$

Theorem 6: Continuous Mapping Theorem

Let X_n and X be a random variable. Let g(.) be a continuous function.

1. If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.
2. If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

Remark 6.64 *$g(.)$ can actually be continuous **almost surely**, i.e. $P(x \in D) = 0$ where D is the set of discontinuity points of g.*

Theorem 7: Slutsky's Theorem

Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ where c is a constant. Then

1. $X_n + Y_n \xrightarrow{d} X + c$;
2. $X_n Y_n \xrightarrow{d} cX$;
3. If $c \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.

Remark 6.65 *This theorem is equivalent for when $Y_n \xrightarrow{d} c$.*

We recall some concepts from calculus to derive the delta method, a tool that helps us approximate the mean and variance of estimators.

Definition 6.66 *(Taylor Polynomial). If a function $g(x)$ has derivatives of order r, that is, $g^{(r)}(x) = \frac{d^r}{dx^r}g(x)$ exists, then for any constant a, the Taylor polynomial of order r about a is*

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i.$$

Theorem 6.67 *If $g^r(a) = \frac{d^r}{dx^r}g(x)|_{x=a}$ exists, then*

$$\lim_{x\to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

That is, the remainder from the approximation, $g(x) - T_r(x)$, always tend to 0 faster than the highest-order explicit term.

For our purposes, we are interested in the first-order Taylor series expansion of an estimator T(.) with the differentiable function g(T) about the parameter point θ:

$$g(t) \approx g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(t_i - \theta_i).$$

We can now look at a theorem to help us determine the limiting variance of an estimator.

Theorem 8: Delta Method

Let X_n be a sequence of random variables such that $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and g(.) is differentiable and nonzero at θ. Then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2[g'(\theta)]^2).$$

Remark 6.68 *Note that the Delta method requires that X_n has a limiting normal distribution in order for us to apply the Delta method to find the limiting distribution of $g(X_n)$.*

Lemma 6.69 *Suppose $\sqrt{n}(X_n - c) \xrightarrow{d} F(.)$ for a proper CDF F(.). Then $X_n \xrightarrow{p} c$.*

Definition 6.70 *(Variance stabilising transformation). Suppose that the limiting variance is a function of an unknown parameter. A function $g(.)$ is a variance stabilising transformation if the limiting variance is no longer a function of the unknown parameter.*

STAT3923: Advanced Statistical Inference

7. Joint, Marginal, and Conditional Distributions

7.3 Multivariate Distributions

7.3.1 Joint, Marginal, and Conditional Distributions

Definition 7.71 *(Joint Probability Density Function). A bivariate function with values $f(x, y)$ defined over the xy-plane is called a joint probability density function of the continuous random variables X and Y if and only if*

$$P(X, Y) \in A = \int \int_A f(x, y)dxdy$$

for any region A in the xy-plane.

Theorem 7.72 *A bivariate function can serve as the joint probability distribution function of a pair of random variabels X and Y if and only if f(x,y) satisfies that*

1. $f(x, y) \geq 0$ *for each pair of values (x,y) within its domain;*
2. $\int_X \int_Y f(x, y) = 1$ *for each pair of values (x,y) within its domain.*

Definition 7.73 *(Joint CDF). The joint CDF of X and Y is given by*

$$f_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)dtds$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

Theorem 9: Tonelli's Theorem

The iterated/repeated integral of a non-negative function is the same as the double integral

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)dtds = \int \int_{B_{xy}} f(s, t)dsdt = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t)dsdt.$$

Fubini's extension states that if for **any** integrable f, that if one of the above three integrals is finite, when f is replaced by $|f|$, then the equalities still hold.

Theorem 7.74 *Assume the CDF function $F \in C^2$, then*

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

Definition 7.75 *(Marginal Density) If X and Y are continuous random variables and f(x,y) is the value of their joint probability density at (x,y), the function given by*

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad -\infty < x < \infty$$

is called the marginal density of X. Correspondingly, the function given by

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx \quad -\infty < y < \infty$$

is called the marginal density of Y.

Remark 7.76 *For a multivariate joint probability distribution, we can also speak of the* ***joint marginal distribution****.*

Remark 7.77 *We can derive the marginal distribution from the joint distribution but* ***not the converse****.*

Definition 7.78 *(Conditional Distribution). If f(x,y) is the value of the joint probability distribution of the random variables X and Y at (x,y) and h(y) is the value of the marginal distribution of Y at y, the function given by*

$$f(x|y) = \frac{f(x,y)}{h(y)} \quad h(y) \neq 0$$

for $-\infty < x < \infty$ is called the conditional density of X given $Y = y$. Correspondingly, if g(x) is the value of the marginal density of X at x, the function given by

$$w(y|x) = \frac{f(x,y)}{g(x)} \quad g(x) \neq 0$$

for $-\infty < y < \infty$ is called the conditional density of Y given $X = x$.

Definition 7.79 *(Independence of random variables). If $f(x_1, x_2, ..., x_n)$ is the value of the joint probability distribution of the random variables $X_1, X_2, ..., X_n$ at $(x_1, x_2, ..., x_n)$ and $f_i(x_i)$ is the value of the marginal distribution of X_i at x_i for i = 1,2,...,n, then the n random variables are independent if and only if*

$$f(x_1, ..., x_n) = f_1(x_1)f_2(x_2)...f_n(x_n)$$

for all $(x_1, x_2, ..., x_n)$ within their range.

Definition 7.80 *(Conditional Expectation). We define the conditional expectation*

$$E(X|Y = y) = \begin{cases} \sum x f_{X|Y}(x|y) & X \text{ is discrete} \\ \int x f_{X|Y}(x|y)dx & X \text{ is continuous.} \end{cases}$$

Definition 7.81 *(Conditional Variance). We define the conditional variance as*

$$V(Y|X = x) = \int (y - E(Y|X = x))^2 f(y|x)dy.$$

Theorem 7.82 *(Law of total expectation). We define the law of total expectation as*

$$E(Y) = E[E(Y|X)].$$

Theorem 7.83 *(Law of total variance). We define the law of total variance as*

$$V(Y) = E[V(Y|X)] + V(E[Y|X]).$$

7.3.2 Multivariate Distribution

Definition 5: Bivariate Normal Distribution

A pair of random variables X and Y have a bivariate normal distribution and they are referred to as jointly normally distributed random variables if and only if their joint probability density is given by

$$f(x,y) = \frac{e^{-\frac{1}{2(1-\rho)^2}[(\frac{x-\mu_1}{\sigma_1})^2] - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2}) + (\frac{y-\mu_2}{\sigma_2})^2}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

for $x \in (-\infty, \infty)$ and $y \in (-\infty, \infty)$, where $\sigma_1 > 0$, $\sigma_2 > 0$, and $-1 < \rho < 1$.

Theorem 7.84 *If X and Y have a bivariate normal distribution, the conditional density of Y given $X = x$ is a normal distribution with the mean*

$$\mu_{Y|x} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

and the variance

$$\sigma^2_{Y|x} = \sigma_2^2(1 - \rho^2)$$

and the conditional density of X given $Y = y$ is a normal distribution with the mean

$$\mu_{X|y} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2)$$

and the variance

$$\sigma^2_{X|y} = \sigma_1^2(1 - \rho^2).$$

Theorem 7.85 *If two random variables have a bivariate normal distribution, they are independent if and only if $\rho = 0$.*

Theorem 7.86 *If (X, Y) is a bivariate normal random variable, then $aX + bY$ is a normal random variable for constants a and b.*

STAT3923: Advanced Statistical Inference

8. Sampling Distributions

8.3.3 Sampling Distributions

Definition 8.87 *(Random Sample). If $X_1, ..., X_n$ are i.i.d random variables, we say that they constitute a random sample from the infinite population given by their common distribution. We can write their joint distribution as*

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i).$$

Definition 8.88 *(Statistic). A statistic T(.) is a random variable that is a function of a set of random variables $X_1, ..., X_n$ that constitute a random sample.*

Proposition 8.89 *A statistic is a random variable and hence has a sampling distribution.*

Definition 8.90 *(Sample Mean and Sample Variance). If $X_1, ..., X_n$ are a random sample, then the sample mean is given by*

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

and the sample variance is given by

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

The values of sampling statistics can be expected to vary from sample to sample, hence we find the distribution of such statistics.

Definition 8.91 *(Sampling Distribution). The distribution of the sampling statistics is known as the* ***sampling distribution.***

Theorem 8.92 *Let $X_1, ..., X_n$ be a random sample with mean μ and variance σ^2. Then,*

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Definition 8.93 *(Standard Error of the mean). Let $\bar{X}$ be the sample mean. Then, let $\sigma^2_{\bar{X}} = Var(\bar{X})$. We define $\sigma_{\bar{X}}$ as the standard error of the mean.*

Proposition 8.94 *Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Then the sample mean $\bar{X}$ and sample variance S^2 are independent.*

Proposition 3: Sampling distribution of variance

The transformed sample variance has the distribution

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

Furthermore,

$$E(\frac{(n-1)}{\sigma^2}S^2) = n-1$$
$$E(S^2) = \sigma^2$$

that is, S^2 is an unbiased estimator of σ^2.

Definition 8.95 *(T-distribution). Let $Z \sim N(0,1)$ and $Y \sim \chi^2_v$. Let Z and Y be independent. Then, we say that a random variable T*

$$T = \frac{Z}{\sqrt{\frac{Y}{2}}}$$

has a t distribution $T \sim t_v$.

Definition 8.96 *(F distribution). Let $U \sim \chi^2_{v_1}$ and $V \sim \chi^2_{v_2}$ where U and V are independent. Then we can construct the F distribution by defining*

$$F = \frac{\frac{U}{v_1}}{\frac{V}{v_2}}$$

giving us $F \sim F_{v_1,v_2}$.

STAT3923: Advanced Statistical Inference

9. Order Statistics

9.3.4 Order Statistics

Definition 6: Order Statistics

Let $X_1, ..., X_n$ be an i.i.d sample from a population with the same CDF F. Define Y_j to be the jth smallest value of $X_1, ..., X_n$. The ordered values $Y_1 < Y_2 ... < Y_n$ are called the order statistics.

Remark 9.97 *The order statistics are a permutation of the original dataset. Furthermore, the distribution of the order statistics depends on the sample size n.*

Proposition 9.98 *The CDF of the k-th order statistic is given by*

$$F_{Y_k}(x) = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}.$$

The PDF of the k-th order statistic is given by

$$f_{Y_k}(x) = k\binom{n}{k} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x).$$

Theorem 9.99 *Let Y_n be the maximum statistic. The CDF of the maximum Y_n is given by*

$$F_{Y_n}(x) = [F(x)]^n.$$

The PDF is given by

$$f_{Y_n}(x) = n[F(x)]^{n-1} f(x).$$

Theorem 9.100 *Let Y_1 be the minimum statistic. The CDF of the maximum Y_1 is given by*

$$F_{Y_1}(x) = 1 - [1 - F(x)]^n.$$

The PDF is given by

$$f_{Y_1}(x) = n[1 - F(x)]^{n-1} f(x).$$

Definition 9.101 *(Joint PDF of 2 Order Statistics). Let $\{Y_i\}$ be the order statistics. Then, for any $i < j$, the joint PDF of 2 order statistics is*

$$f_{Y_i,Y_j}(y_i, y_j) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} F(y_i)^{i-1} [F(y_j) - F(y_i)]^{j-i-1} (1 - F(y_j))^{n-j} f(y_i) f(y_j).$$

Definition 9.102 *(Joint PDF of all Order Statistics). Let $\{Y_i\}$ be the order statistics. Then, the joint PDF of all the order statistics is given by*

$$f_{Y_{(1)},Y_{(2)},...,Y_{(n)}}(y_1, y_2, ..., y_n) = \begin{cases} n!\, f(y_1)...f(y_n) & -\infty < y_1 < .... < y_n < \infty \\ 0 \quad \textit{otherwise.} \end{cases}$$

STAT3923: Advanced Statistical Inference

10. Transformation of random variables

10.4 Transformation of random variables

10.4.1 Transformation of random variables

Suppose we are given a set of random variables $X_1, ..., X_n$ and we are interested in the probability distribution or density of $Y = g(X_1, ..., X_n)$. The 3 techniques we can use are

1. Distribution Function Technique;
2. Transformation Technique;
3. Moment-Generating Function Technique.

Theorem 10.103 *(Distribution Function Technique). We obtain the CDF of Y $F_Y(y) = P(Y \leq y)$ and then differentiate with respect to y to find the probability density $f_Y(y) = \frac{dF(y)}{dy}$.*

Remark 10.104 *Typically this is used for scalar-valued function and for continuous distributions.*

Definition 10.105 *(Convolution). Let X and Y be independent random variables. Define the random variable $T = X + Y$. Then the convolution is defined as*

1. $P(T = t) = \sum_X P(X = x, Y = t - x) = \sum_X P_X(x)P_Y(t - x)$ *(Discrete)*
2. $f_T(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x)dx$ *(Continuous).*

STAT3923: Advanced Statistical Inference

11. Transformation of random variables

11.4.2 Univariate Transformation of random variables

Theorem 11.106 *Suppose $f'(x) > 0$ in (a,b). If f is strictly increasing in (a,b) and let g be its inverse function. Then g is differentiable and*

$$g'(f(x)) = \frac{1}{f'(x)}.$$

Theorem 10: Transformation techniques

Let X be a continuous random variable having PDF $f_X(.)$. Suppose that g(x) is differentiable and strictly monotonic for all x such that $f_X(x) \neq 0$. Then the random variable $Y = g(X)$ has the PDF

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)| & \text{if y = g(x)} \\ 0 \quad \text{otherwise} \end{cases}$$

Theorem 11.107 *Suppose (X_1, X_2) has the joint PDF $f(x_1, x_2)$ and define $Y = U(X_1, X_2)$. W.L.O.G, if we fix X_2, then $U(., X_2)$ satisfies the conditions required by the one-variable transformation technique. Then we can write the joint PDF of (Y, X_2) as*

$$g(y, x_2) = f(x_1, x_2)|\frac{\partial x_1}{\partial y}|.$$

Then we can marginalise out X_2 to arrive at

$$f_Y(y) = \int_{-\infty}^{\infty} g(y, x_2)dx_2.$$

STAT3923: Advanced Statistical Inference

12. Multivariate Transformation of random variables

12.4.3 Multivariate Transformation of random variables

Definition 12.108 *(Jacobian). Let (X_1, X_2) have the joint PDF $f(x_1, x_2)$. Let $g(y_1, y_2) = f(u_1(x_1, x_2), u_2(x_1, x_2))$. The Jacobian is then*

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix}.$$

Theorem 11: Multivariate Transformation

Define $T : D \subset \mathbb{R}^2 \to R \subset \mathbb{R}^2$ where D and R are open sets. Suppose that T is bijective and differentiable on D with a non vanishing Jacobian $J_T \neq 0$. Suppose (X, Y) is jointly continuous with density f_{XY} which vanishes outside of D and let $(U, V) = T(X, Y)$. Then (U, V) is jointly continuous and for all $(u, v) \in R$, we have the density

$$f_{UV}(u, v) = f_{XY}(T^{-1}(u, v))|J_{T^{-1}}(u, v)|.$$

Proof: Take $\psi = f_{XY}$, then for any open $B \subset R$, we have that

$$\begin{aligned} P((U, V) \in B) &= P((X, Y) \in A = T^{-1}(B)) \\ &= \int\int_A f_{XY}(x, y)dxdy \\ &= \int\int_{B=T(A)} f_{XY}(T^{-1}(u, v))|J_{T^{-1}}(u, v)|dudv. \end{aligned}$$

Since this holds for all $B \subset R$, (U,V) has a density which is given by

$$f_{UV}(u, v) = f_{XY}(T^{-1}(u, v))|J_{T^{-1}}(u, v)|.$$

■

Remark 12.109 *The Jacobian takes into account that the density increases (decreases) after the linear transformation and hence scales the density down (up) to compensate.*

Claim 12.110 *Suppose $X_1, X_2, ..., X_n \sim F_X$ and $\boldsymbol{Y} = T(\boldsymbol{X})$ where T is bijective from $D \subset \mathbb{R}^n \to R \subset \mathbb{R}^n$. Denote $\boldsymbol{Y} \sim G_{\boldsymbol{Y}}$, for any $\boldsymbol{Z} \sim G_{\boldsymbol{Y}}$,*

$$T^{-1}\boldsymbol{Z} \sim F_{\boldsymbol{X}}.$$

STAT3923: Advanced Statistical Inference

13. Multivariate Jacobian Technique

13.4.4 Multivariate Jacobian

Theorem 13.111 *Let X and Y be 2 vectors of random variables, related by an invertible transformation*

$$Y_1 = y_1(X), \quad Y_2 = y_2(X), \quad ... \quad Y_n = y_n(X).$$

Then the joint PDF of X is related to the joint PDF of Y via

$$f_Y(y) = f_X(x)|detJ|$$

where the inverse transforms are

$$X_1 = x_1(Y), \quad X_2 = x_2(Y), \quad ... \quad X_n = x_n(Y)$$

and the Jacobian matrix is

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & ... & \frac{\partial x_1}{\partial y_n} \\ ... & ... & ... \\ \frac{\partial x_n}{\partial y_1} & ... & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

STAT3923: Advanced Statistical Inference

14. Sufficient Statistics

14.5 Exponential Families

14.5.1 Information Theory (Background)

Information theory will make the notions of sufficiency alot more intuitive. We work with discrete random variables for ease.

Definition 14.112 *(Entropy). The entropy H(X) of a discrete random variable X is defined by*

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x) log_2 p(x).$$

The unit of measurements for entropy is called ***bits****.*

Remark 14.113 *Entropy measures the uncertainty of a random variable. That is, what is the amount of information required on average to describe the random variable. Note that entropy only depends on the probabilities of the PMF rather than the actual values that the random variable takes on.*

Lemma 14.114 *We have that*

$$H(X) \geq 0.$$

Recall that if g(X) is a function of a random variable, then

$$E(g(X)) = \sum_{x\in\mathcal{X}} g(x) p_X(x).$$

Hence, we can let $g(X) = \frac{1}{p(X)}$ and then

$$H(X) = E_p log(\frac{1}{p(X)}).$$

Definition 14.115 *(Joint Entropy). The joint entropy H(X, Y) of a pair of discrete random variables (X, Y) with a joint distribution p(x,y) is defined as*

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) log p(x,y)$$

Definition 14.116 *(Conditional Entropy). If $(X,Y) \sim P(x,y)$, the conditional entropy $H(Y|X)$ is defined as*

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) = -E log p(Y|X).$$

Theorem 14.117 *(Chain rule).*

$$H(X,Y) = H(X) + H(Y|X).$$

Corollary 14.118

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z).$$

Definition 14.119 *(Relative Entropy/Kullback-Leibler Distance). The relative entropy between two probability mass functions p(x) and q(x) is defined as*

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x) log \frac{p(x)}{q(x)} = E_p log \frac{p(X)}{q(X)}.$$

Remark 14.120 *Relative entropy measures the distance between two distributions.*

Definition 14.121 *(Mutual Information). Consider two random variables X and Y with a joint PMF $p(x,y)$ and marginal PMFs p(x) and p(y). The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution p(x)p(y);*

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) log \frac{p(x,y)}{p(x)p(y)}.$$

Remark 14.122 *Mutual information is a measure of the amount of information that one random variables contains about another random variable. It is the reduction in the uncertainty of one random variable due to knowledge of another.*

Lemma 14.123 *We can express mutual information as*

$$I(X;Y) = H(X) - H(X|Y).$$

Mutual information is therefore the reduction in the uncertainty of X due to knowledge of Y.

Definition 14.124 *The conditional mutual information of random variables X and Y given Z is defined by*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z).$$

Theorem 14.125 *(Information Inequality). Let p(x), q(x), and $x \in \mathcal{X}$ be two PMFs. Then the relative entropy*

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Corollary 14.126 *For any two random variables X and Y*

$$I(X;Y) \geq 0$$

with equality if and only if X and Y are independent.

Theorem 14.127 *Conditioning reduces entropy*

$$H(X|Y) \leq H(X)$$

with equality if and only if X and Y are independent.

Definition 14.128 *(Markov Chain). Random variables X,Y,Z are said to form a Markov chain in that order ($X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specially, the joint PMF can be written as*

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

Theorem 14.129 *(Data-processing inequality). If $X \to Y \to Z$, then*

$$I(X; Z) \leq I(X; Y).$$

Definition 14.130 *Suppose we have a family of PMFs $\{f_\theta(x)\}$ indexed by θ and let X be any sample from a distribution in this family. Let T(X) be any statistic. Then $\theta \to X \to T(X)$ is a Markov Chain and by the data-processing inequality*

$$I(\theta; T(X)) \leq I(\theta; X)$$

14.5.2 Sufficient Statistics

In this section, we work with the intuitive idea that discarding irrelevant data can never hurt performance. In fact, irrelevant data may actually impair performance. The advantage of this is that it makes inference alot easier to do.

Definition 14.131 *(Statistic). A statistic $T : X \to \mathbb{R}^n$ is a function of the data.*

Suppose we had a random sample $\tilde{X} = (X_1, ..., X_n)$ whose distribution depends on θ. We want to estimate the parameter θ using $\tilde{X}$ using a function $T(\tilde{X})$ without losing information about θ.

Definition 7: Likelihood Function

We have a random sample of size n from a distribution with PDF $f(x; \theta)$. Then the likelihood function is

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

where $\tilde{x} = (x_1, ..., x_n)$ are observed values.

Hence, a function T(X) is a **sufficient statistic** if

$$I(\theta; T(X)) = I(\theta; X)$$

as no information is lost. Hence sufficient statistics preserve mutual information.

Definition 8: Sufficient Statistic

A statistic $\tilde{T} = \tilde{T}(\tilde{X})$ is sufficient for a family of distributions if and only if the conditional distribution of $\tilde{X}$ given $\tilde{T}(\tilde{X})$ is independent of the parameters.
In general, if $X_1, ..., X_n$ are random samples from the discrete distribution with PMF $f(x;\theta)$, the conditional probability of $\tilde{X} = \tilde{x}$ given $\tilde{T} = \tilde{t}$ is

$$f(\tilde{X};\theta|\tilde{T} = \tilde{t}) = \frac{\prod_{i=1}^{n} f(x_i;\theta)}{f_T(\tilde{t};\theta)}$$

where we require the ratio of the likelihood and marginal distribution of T to be independent of the parameter θ.

If we can express the likelihood with just the parameter θ and statistic T(x), then T(x) is a sufficient statistic. In other words, a statistic T is sufficient if for all t, the conditional distribution $X|T(x) = t$ does not depend on the parameter θ.

However, it may be difficult to compute the conditional distribution which leads us to the next theorem.

Theorem 12: Neyman Factorisation Theorem

Let $f(x;\theta)$ be the PDF of a random sample $\tilde{X} = X_1, ..., X_n$. Let $\tilde{T} = \tilde{T}(\tilde{X})$ be a statistic. Then, $\tilde{T}(\tilde{X})$ is a sufficient statistic for θ if and only if the **likelihood** $\ell(\tilde{x};\theta)$ can be written in the form

$$\ell(\tilde{x};\theta) = g(\tilde{T}(\tilde{x});\theta)h(\tilde{x})$$

where $h(\tilde{x})$ is independent of θ.

Remark 14.132 *Note that the first factor $g(\tilde{T}(\tilde{x})$ means that it may depend on θ and possibly depend on x, but only through T(x).*

So why do we care about sufficient statistics? If we had a sufficient statistic, when computing the likelihood function, instead of evaluating n PDFs, we can evaluate a single PDF with the sufficient statistic placed inside.

Theorem 13: Sufficiency Principle

If T($\mathbf{X}$) is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value T($\mathbf{X}$). That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

We now present two notable examples of sufficient statistics.

Lemma 14.133 *(Max of uniform). Let $X_1, ..., X_n$ be i.i.d according to the uniform distribution $U(0,\theta)$. Then, $T(x) = \max(X_1, ..., X_n)$ is a sufficient statistic.*

Lemma 14.134 *(Order Statistics). Let $X_1, ..., X_n$ be i.i.d with any model. Then, the order statistics $T = X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ are sufficient statistics.*

Finally, note however that reduction via sufficiency can also increase the computational complexity of inference, in some instances even turning a computationally tractable inference problem into an intractable one.

STAT3923: Advanced Statistical Inference

15. Exponential Families

15.5.3 Exponential Families

Exponential families are of particular interest to us because many common distributions are exponential families. Examples include the normal, binomial, and poisson. Furthermore, the exponential families are closely linked to the notion of sufficiency.

Definition 9: Exponential family

A one parameter exponential family is a set of probability distributions that can be written in the form

$$f(x;\theta) = e^{\eta(\theta)T(x)-\phi(\theta)}h(x)I_A(x)$$

for $x \in \mathbb{R}^n$ and $\theta \in \Theta \subseteq \mathbb{R}$. We have that $\eta(.), T(.), \phi(.), h(.)$ are real-valued functions. $I_A(.)$ indicates the support of the distribution and does not depend on θ.

Remark 15.135 *η, T, ϕ, h are not unique. $h(x) \geq 0$ is known as a base (carrier) measure. $\phi(\theta)$ is a normalising constant.*

Lemma 15.136 *The parameterisation of exponential families is **not** unique.*

Proposition 4: Sufficiency of Exponential family

The exponential family always has a sufficient statistic.

Lemma 15.137 *The uniform distribution is not part of the exponential family as its support depends on the parameters (a, b).*

Proposition 15.138 *(Exponential Distribution). The density of the exponential distribution has the density*

$$f(x;\theta) = \theta e^{-\theta x} I_{x\geq 0} = e^{-\theta x + log(\theta)} I_{x\geq 0}$$

which yields a 1-dimensional exponential family with

1. $\eta(\theta) = -\theta$
2. $T_i(x) = x$
3. $\phi(\theta) = -log(\theta)$
4. $h(x) = I_{x\geq 0}$.

STAT3923: Advanced Statistical Inference

16. Canonical Paramter Exponential Families

16.5.4 Canonical Paramter Exponential Families

Definition 10: Canonical Form of exponential family

The canonical form of one-parameter exponential family is

$$f(x;\eta) = e^{\eta T(x) - \psi_0(\eta)} h(x) I_A(x) \quad x \in \mathbb{R}^n$$

where

$$\eta \in \mathcal{F} = \{\eta : e^{\psi_0(\eta)} = \int_A e^{\eta T(x)} h(x) dx < \infty\}.$$

Here, η is called the **natural parameter**; T(x) is a **natural sufficient statistic** for η.
$\mathcal{F}$ is the **natural parameter space** which describes the set of values of η for which the PDF can be defined.

This parameterises the density in terms of the **natural parameter** η rather than θ.

Definition 16.139 *(Regular). The canonical exponential family is called **regular** of $\mathcal{F}$ is an open set in $\mathbb{R}$.*

Theorem 14: Moments of sufficient statistics

For any η in the interior of $\mathcal{F}$, we have that

1. $E(T(X)) = \psi_0^{'}(\eta)$;
2. $Var(T(x)) = \psi_0^{''}(\eta)$.

Theorem 16.140 *(Moments of sufficient statistics). Suppose $T(\tilde{X})$ is a natural sufficient statistic for η based on a random sample $\tilde{X} = (X_1, X_2, ..., X_n)$ from $f(x;\eta)$, then*

1. $E(T(\tilde{X})) = n\psi_0^{'}(\eta)$;
2. $Var(T(\tilde{x})) = n\psi_0^{''}(\eta)$.

Proposition 16.141 *(Independent exponentials). If $X_1, ..., X_n$ are i.i.d with pdf $e^{\eta T(x) - \psi_0(\eta)} h(x) I_A(x) h(x)$, then, their joint pdf is*

$$f(x_1, ..., x_n; \theta) = e^{\eta \sum_{j=1}^n T(x_j) - n\psi_0(\eta)}(x) \prod_{j=1}^n h(x_j) I_A$$

Proposition 16.142 *The exponential family is infinitely differentiable with respect to η and the derivatives can be obtained by differentiating under the integral sign.*

16.5.5 Two Parameter Exponential Families

Definition 11: Two Parameter exponential families

Let $\tilde{\theta} = (\theta_1, \theta_2)$. A family of distributions is said to be 2-parameter exponential family if there exists real-valued functions $\eta_1(.), \eta_2(.), T_1(.), T_2(.), \psi(.), h(.)$ such that the PDF

$$f(x;\theta) = e^{\sum_{i=1}^{2} \eta_i(\tilde{\theta}) T_i(x) - \psi(\tilde{\theta})} h(x) I_A(x)$$

for $x \in \mathbb{R}^n$.

The beta distribution is an example of a two parameter exponential family.

16.5.6 Uniform and Exponential Spacing

Definition 16.143 *(Uniform Spacing). Let $U_1, ..., U_n$ be i.i.d uniform [0,1] with order statistics $U_{(1)} \leq U_{(2)} \leq ... \leq U_{(n)}$. The statistics S_i defined by*

$$S_i = U_{(i)} - U_{(i-1)} \quad (1 \leq i \leq n+1)$$

where $U_{(0)} = 0, U_{(0)} = 1$ are called the uniform spacings for this sample.

Theorem 16.144 *The uniform spacing $(S_1, ..., S_n)$ are uniformly distributed over the simplex*

$$A_n = \{(x_1, ..., x_n) : x_i \geq 0, \sum_{i=1}^{n} x_i \leq 1\}.$$

Definition 16.145 *(Exponential Spacing). Let $E_1, ..., E_n$ be i.i.d exponential random variables with order statistics $E_{(1)} \leq E_{(2)} \leq ... \leq E_{(n)}$. The statistics defined by*

$$(n - i + 1)(E_{(i)} - E_{(i-1)}) \quad 1 \leq i \leq n$$

are known as normalized exponential spacings.

Theorem 16.146 *Let $(n - i + 1)(E_{(i)} - E_{(i-1)}) \quad 1 \leq i \leq n$ be normalized exponential spacings. These are i.i.d. exponential random variables. Furthermore,*

$$\frac{E_1}{n}, \frac{E_1}{n} + \frac{E_2}{n-1}, ..., \frac{E_1}{n} + ... + \frac{E_1}{1}$$

are distributed as $E_{(1)}, ...E_{(n)}$.

STAT3923: Advanced Statistical Inference

17. Maximum Likelihood Estimators

17.6 Minimum Variance Unbiased Estimation

17.6.1 The Likelihood Principle

We are interested in finding the parameters θ as knowledge of θ will allow us to generate data through the pdf. We look at techniques at estimating the parameter θ and techniques for evaluating our estimations.

Definition 17.147 *(Likelihood Function). Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf of the sample $\boldsymbol{X} = (X_1, ..., X_n)$. Then, given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function of θ defined by*

$$\ell(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

is called the likelihood function.

Remark 17.148 *Recall that if $\boldsymbol{X}$ is a discrete random vector, the likelihood function will be $\ell(\theta|\boldsymbol{x}) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$ and hence for two different parameter points θ_1, θ_2, then we can interpret the likelihood function as the probability of a parameter θ_i given the sample we have $\boldsymbol{x}$*

$$P_{\theta_1}(\boldsymbol{X} = \boldsymbol{x}) = \ell(\theta_1|\boldsymbol{x}_1) > \ell(\theta_2|\boldsymbol{x}_2) = P_{\theta_2}(\boldsymbol{X} = \boldsymbol{x}).$$

Theorem 17.149 *Let X_i be i.i.d random variables with common pdf$f_\theta(.)$. Then,*

$$log\Big(\prod_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} log\Big(X_i\Big).$$

Here, the pdf fixes θ and varies $\mathbf{x}$ whereas the likelihood function fixes $\mathbf{x}$ and varies θ.

Theorem 17.150 *(Likelihood Principle). If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $\ell(\theta|\boldsymbol{x})$ is proportional to $\ell(\theta|\boldsymbol{y})$, that is, there exists a constant $C(\boldsymbol{x}, \boldsymbol{y})$ such that*

$$\ell(\theta|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y})\ell(\theta|\boldsymbol{y})$$

for all θ, then the conclusions drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ should be identical.

17.6.2 Maximum Likelihood Estimators

Definition 17.151 *(Likelihood Function). If $X_1, ..., X_n$ are i.i.d. sample from a population with pdf $f(x|\theta_1, ..., \theta_k)$, the likelihood function is defined by*

$$\ell(\theta|\boldsymbol{x}) = \ell(\theta_1, ..., \theta_k|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_k).$$

Definition 12: Maximum Likelihood Estimator

For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $\ell(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

Remark 17.152 *The range of the MLE coincides with the range of the parameter.*

The MLE is the parameter point of the MLE for which the observed sample is most likely.

Remark 17.153 *Two inherent drawbacks of MLE is that finding the global maximum can be difficult and that the estimate may be sensitive to small changes in the data. This second scenario occurs in the case of flat likelihoods.*

Proposition 5: Necessary condition for MLE

The first derivative of a function being 0 is a **necessary** condition for a maximum

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, \quad i = 1, ..., k.$$

Remark 17.154 *The zeros of the first derivative are only located in the extrema in the interior of the domain of the function. Henec, we need to check the boundaries separately for extrema.*

Remark 17.155 *When maximising a likelihood with restrictions on the parameter, we need to check for different cases of the optimal values.*

Theorem 17.156 *(Invariance Property of MLE). Suppose that $\hat{\theta}$ is the MLE of a parameter θ. Let $\tau(\theta)$ be a one to one mapping. Then, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.*

Theorem 17.157 *(MLE of multivariate likelihood). Suppose our likelihood function is $H(\theta_1, \theta_2)$. To check that $H(\theta_1, \theta_2)$ has a local maximum at $(\hat{\theta}_1, \hat{\theta}_2)$, we check the 3 conditions that*

1. *First order partial derivatives are 0*

$$\frac{\partial}{\partial \theta_1} H(\theta_1, \theta_2) = \frac{\partial}{\partial \theta_2} H(\theta_1, \theta_2) = 0$$

2. *At least one second-order partial derivative is negative;*

3. *The Jacobian J of the second-order partial derivatives is positive*

$$|J| > 0.$$

17.6.3 Mean Squared Error

We look at **finite-sample** measures of the quality of an estimator.

Definition 13: Mean Squared Error

The mean squared error (MSE) of an estimator W of a parameter θ is the function of θ defined by

$$\mathbb{E}[(W-\theta)^2].$$

Definition 14: Bias

The bias of a point estimator W of a parameter θ is the difference between the expected value of W and θ, that is,

$$\text{Bias}_\theta = \mathbb{E}[W-\theta].$$

An estimator whose bias is equal to 0 is called **unbiased** and satisfies $\mathbb{E}_\theta[W] = \theta$ for all θ.

Theorem 15: MSE Decomposition

The MSE can be decomposed as the sum of the variance of the estimator plus the square of the bias:

$$E_\theta[(W-\theta)^2] = Var_\theta(W) + (E_\theta[W-\theta])^2 = Var_\theta(W) + (\text{Bias}_\theta W)^2.$$

Corollary 17.158 *The MSE of an **unbiased** estimator is equal to its variance.*

STAT3923: Advanced Statistical Inference

18. Differentiation and Integration

18.6.4 Differentiation and Integration

We now look into when can we interchange differentiation, integration, and summation.

Theorem 18.159 *(Leibnitz's Rule). If $f(x,\theta), a(\theta), b(\theta)$ are differentiable with respect to θ, then*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - f(a(\theta),\theta)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta}f(x,\theta)dx.$$

Corollary 18.160 *If $a(\theta), b(\theta)$ are constants, then*

$$\frac{d}{d\theta}\int_{a}^{b} f(x,\theta)dx = \int_{a}^{b} \frac{\partial}{\partial\theta}f(x,\theta)dx.$$

Remark 18.161 *Note the LHS of the corollary depends on one parameter whereas the RHS depends on two parameters.*

Theorem 18.162 *(Dominated Convergence Theorem). Suppose the function $h(x,y)$ is continuous at y_0 for each x, and there exists a function $g(x)$ satisfying*

1. *$|h(x,y)| \leq g(x)$ for all x and y,*
2. *$\int_{-\infty}^{\infty} g(x)dx < \infty$.*

Then,

$$\lim_{y\to y_0}\int_{-\infty}^{\infty} h(x,y)dx = \int_{-\infty}^{\infty}\lim_{y\to y_0} h(x,y)dx.$$

Theorem 18.163 *(Interchange integration and limits). Suppose $f(x,\theta)$ is differentiable for every $\theta = \theta_0$. That is*

$$\lim_{\delta\to 0}\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta} = \frac{\partial}{\partial\theta}f(x,\theta)$$

such that

1. *$|\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta}| \leq g(x,\theta_0)$ for all x and $|\delta| \leq \delta_0$;*
2. *$\int_{-\infty}^{\infty} g(x,\theta)dx < \infty$.*

Then,

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta)dx.$$

Theorem 18.164 *We can interchange integration and differentiation for the exponential family.*

Lemma 18.165 *A derivative can always be taken inside a* ***finite*** *sum.*

Theorem 18.166 *(Interchange differentiation and summation). Suppose that the series* $\sum_{x=0}^{\infty} h(\theta, x)$ *converges for all* $\theta \in (a, b)$ *and*

1. $\frac{\partial}{\partial\theta} h(\theta, x)$ *is continuous in* θ *for each* x*,*
2. $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(\theta, x)$ *converges uniformly on every closed bounded subinterval of* (a, b)*.*

Then,

$$\frac{d}{d\theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(\theta, x).$$

Theorem 18.167 *(Interchange summation and integration). Suppose the series* $\sum_{x=0}^{\infty} h(\theta, x)$ *converges uniformly on [a,b] and that, for each* x*,* $h(\theta, x)$ *is a continuous function of* θ*. Then*

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x) d\theta = \sum_{x=0}^{\infty} \int_a^b h(\theta, x) d\theta.$$

STAT3923: Advanced Statistical Inference

19. Cramer Rao Lower Bound

19.6.5 Cramer Rao Lower Bound

When looking at the best performance of estimators, we require 3 restrictive conditions

1. We only look at unbiased estimators $E_\theta[\hat{\theta}] = \theta$.
2. We measure performance by the variance of the estimator.
3. We restrict attention to a class of **regular** problems.

Definition 15: Regularity Conditions

For the next section, we state that the regularity conditions are

1. We can interchange the order of differentiation and integration/summation;
2. The PMF $f_\theta(\tilde{x}) \neq 0$ for all $\tilde{x}$ in the support.

Remark 19.168 *Recall interchanging the order of differentiation and integration requires the dominated convergence theorem.*

Definition 16: Uniform Minimum Variance Unbiased Estimators UMVUE

An estimator $\hat{\theta}$ is called the uniform minimum variance unbiased estimator of θ if $E_\theta[\hat{\theta}] = \theta$ and for any other unbiased estimator W, we have that

$$Var_\theta \hat{\theta} \leq Var_\theta W$$

for all θ.

Remark 19.169 *UMVUE does not need to actually only apply to unbiased estimators. Suppose the estimator W^* had the bias $E_\theta[W^*] = \tau(\theta)$. Then, we say that W^* is UMVUE for $\tau(\theta)$ if $Var(W^*) \leq Var(W)$ for all estimators W such that $E_\theta[W] = \tau(\theta)$.*

Definition 17: Score Function

Let $\tilde{X}$ be a random vector with PMF $f_\theta(\tilde{X})$. Assuming the regularity conditions hold, the score function is defined by

$$s(\theta) = \frac{\partial}{\partial \theta} log f_\theta(\tilde{X}).$$

Remark 19.170 *The score indicates the steepness of the log-likelihood function and thereby the sensitivity to infinitesimal changes to the parameter values.*

We recall the following definition in order to help define the Cramer-Rao lower bound.

Definition 19.171 *(Covariance/Correlation). The covariance of two random variables Y and Z is*

$$Cov(Y,Z) = E[Y,Z] - E[Y]E[Z].$$

Theorem 19.172 *(Correlation Inequality). The correlation inequality is*

$$Corr(Y,Z)^2 = \frac{Cov(Y,Z)^2}{Var(Y)Var(Z)} \leq 1.$$

Definition 18: Cramer-Rao Lower Bound

Let $\hat{\theta}(.)$ be an **unbiased estimator** with finite variance. Furthermore, assume that regularity conditions holds. Then, the lower bound of the variance of any unbiased estimator is given by

$$Var_\theta[\hat{\theta}(\tilde{X})] \geq \frac{1}{Var_\theta[\frac{\partial}{\partial\theta} log f_\theta(\tilde{X})]}$$

Proof: Let $\hat{\theta}(.)$ be an unbiased estimator of θ. Hence, we have that

$$E_\theta[\hat{\theta}(X)] = \theta = \sum_{x\in X} ... \sum \hat{\theta}(x) f_\theta(x) = \theta \quad \forall\theta \in \Theta.$$

We then take derivative with respect to θ on both sides, interchange the derivative and summation and multiply and divide by $f_\theta(x)$. We then get

$$\sum_{x\in X} ... \sum \Big[\frac{\partial}{\partial\theta} log(f_\theta(x))\Big] \hat{\theta}(x) f_\theta(x) = 1.$$

We note that this is the definition of the expectation with respect to the pdf $f_\theta(.)$.

$$E_\theta\Big[\hat{\theta}(X), \frac{\partial}{\partial\theta} log(f_\theta(X))\Big].$$

Now, recall that $Cov(X,Y) = E[X,Y]$ if either E[X] = 0 or E[Y] = 0. As $E_\theta[\frac{\partial}{\partial\theta} log(f_\theta(X))] = 0$, we have

$$E_\theta\Big[\hat{\theta}(X), \frac{\partial}{\partial\theta} log(f_\theta(X))\Big] = Cov\Big(\hat{\theta}(X), \frac{\partial}{\partial\theta} log(f_\theta(X))\Big).$$

Now, we apply the correlation inequality to get

$$Cov\Big(\hat{\theta}(X), \frac{\partial}{\partial\theta} log(f_\theta(X))\Big) \leq Var\Big(\hat{\theta}(X)\Big) Var\Big(\frac{\partial}{\partial\theta} log(f_\theta(X))\Big).$$

As the covariance of an unbiased estimator with the score function is less than or equal to 1, we therefore have

$$1 \leq Var\Big(\hat{\theta}(X)\Big) Var\Big(\frac{\partial}{\partial\theta} log(f_\theta(X))\Big)$$

and hence the CRLB follows. ■

Remark 19.173 *An issue with the CRLB is that there may be no estimators that attain the CRLB.*

Definition 19: Fisher Information

The term in the denominator of the Cramer-Rao lower bound

$$Var_\theta[\frac{\partial}{\partial\theta}logf_\theta(\tilde{X})]$$

is known as the **Fisher information** of the sample. The more information we have of the sample, then the smaller the variance of our estimator.

Remark 19.174 *If we have an unbiased estimator $\hat{\theta}(.)$ and we show that its variance is equivalent to the Cramer-Rao lower bound, then we know that $\hat{\theta}(.)$ is an optimal estimator.*

Remark 19.175 *Alternatively, we can formulate the Fisher information as*

$$I_n(\theta) := E_\theta[-\frac{d^2}{d\theta^2}\ell(\theta; X_1, ..., X_n)].$$

*Here, the second derivative measures the curvature of the likelihood function. Taking the expectation of this tells us how curved the likelihood function is on average. The more curved the likelihood function is, the more **information** it contains and hence the more precise the MLE will be.*

Theorem 19.176 *Let $X_1, ..., X_n$ be iid and define the Fisher information as $I_n := E_\theta[-\frac{d^2}{d\theta^2}\ell(\theta; X_1, ..., X_n)]$. Then,*

$$I_n(\theta) = nI_\theta$$

*where $I(\theta)$ is the Fisher information for a **single observation**. That is, $I(\theta) = I_1(\theta)$.*

We are interested in the case of when is the CR-lower bound an equality. In particular, when is the correlation inequality an equality. This occurs when 2 random variables are linearly related.

Lemma 19.177 *Let Y and X be random variables. If $Y = a + bX$, then*

$$Cov(X, Y)^2 = Var(X)Var(Y).$$

We now can specify a way to find an estimator that attains the CRLB. That is, if an unbiased estimator is linearly dependent to the score function, then the estimator attains the Cramer-Rao lower bound!

Theorem 16: Attainment of CRLB

Let $X_1, ..., X_n$ be iid from $f_\theta(x|\theta)$ and $f_\theta(x|\theta)$ satisfies the conditions of the Cramer-Rao theorem. If the unbiased estimator $\hat{\theta}(\tilde{x})$ and $\frac{\partial}{\partial\theta}logf_\theta(\tilde{x})$ are linearly related, then the estimator $\hat{\theta}(\tilde{x})$ attains the CRLB.
Furthermore, as the expectation of the score function of the unbiased estimator must be zero, we have that

$$\frac{\partial}{\partial\theta}logf_\theta(\tilde{x}) = C_\theta[\hat{\theta}(\tilde{x}) - \theta].$$

Remark 19.178 *If the score function has the relationship above, then the estimator $\hat{\theta}$ is a MVUE. Furthermore, the constaint C_θ is the inverse of the Fisher information.*

Theorem 19.179 *If the theorem of attainment holds, we must have an exponential family and $\hat{\theta}(\tilde{x})$ must be a multiple of the sufficient statistic for θ.*

Theorem 17: Rao-Blackwell

Let W be an unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for $\tau(\theta)$. Define $\phi(T) = E(W|T)$. Then

$$\begin{cases} E_\theta \phi(T) = \tau(\theta) \\ Var_\theta \phi(T) \leq Var_\theta W \quad \forall \theta \end{cases}$$

that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Remark 19.180 *Conditioning any unbiased estimator on a sufficient statistic will result in a uniform imporvement.*

STAT3923: Advanced Statistical Inference

20. Asymptotically Minimum Variance Unbiased Estimators

20.6.6 Asymptotically Minimum Variance Unbiased Estimators

At the end of the last section, we identified that if the score function has the following relationship

$$\frac{\partial}{\partial\theta} log f_\theta(\tilde{x}) = C_\theta[\hat{\theta}(\tilde{x}) - \theta].$$

then $\hat{\theta}$ is a minimum variance unbiased estimator. However, this does not always hold. In this section, we look at cases where this almost holds when looking at asymptotics.

Example 20.181 *Let $X_1, ..., X_n$ be iid geometric(p). We have that*

$$\frac{\partial}{\partial\theta} log f_\theta(\tilde{X}) = \frac{-n}{1-\theta}\Big(\overline{X} - \frac{1}{\theta}\Big).$$

Here, we don't quite have the correct form as we have $\frac{1}{\theta}$ rather than θ. Hence, we can't identify an optimal estimator.

We use the notions of asymptotic normality and that the asymptotic variance of the estimator is the CRLB to now identify these estimators.

Definition 20.182 *(Central Limit Theorem). Suppose a sequence of random variables $\{X_n\}$ is such that*

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

or equivalently

$$Z_n = \frac{X_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

i.e. $P(Z_n \leq z) \to \phi(z)$. Then we say that X_n is asymptotically normal $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ and we write this as $X_n \sim A\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Remark 20.183 *We can have n when writing $A\mathcal{N}(\mu, \frac{\sigma^2}{n})$ but not when writing convergence in distributions $\xrightarrow{d}$.*

Remark 20.184 *We refer to σ^2/n as the asymptotic variance of X_n.*

Theorem 18: Delta Method

Suppose $X_n \sim A\mathcal{N}(\mu, \frac{\sigma^2}{n}$ and g(.) is differentiable at μ. Then

$$g(X_n) \sim A\mathcal{N}(g(\mu), \frac{g'(\mu)^2\sigma^2}{n}).$$

Proof:(Sketch). First, it can be shown that if X_n is asymptotically normal (AN), then X_n converges to the asymptotic mean, that is, $X_n \xrightarrow{p} \mu$. Now, we look at

$$\sqrt{n}[g(X_n) - g(\mu)] = [\frac{g(X_n) - g(\mu)}{X_n - \mu}]\sqrt{n}(X_n - \mu).$$

Now, we have that

$$\begin{cases} [\frac{g(X_n)-g(\mu)}{X_n-\mu}] \xrightarrow{p} g'(\mu) & \text{(definition of derivative)} \\ \\ \sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) & \text{(WLLN).} \end{cases}$$

Hence, we have that

$$[\frac{g(X_n) - g(\mu)}{X_n - \mu}]\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2\sigma^2).$$

■

Remark 20.185 *If X is normal, then $g(X)$ is not necessarily normal. This theorem is a local theorem requiring linearity. This is because we are using the fact that the derivative is a linear transformation about the point μ.*

We may not be able to find an unbiased estimator that meets the Cramer-Rao lower bound. However, we may have that it is asymptotically normal, where the mean is the true parameter θ_0 and the variance is the Cramer-Rao lower bound. This is the next best thing.

Definition 20: Asymptotically Minimum Variance Unbiased

We say that an estimator $\hat{\theta}(\tilde{x})$ is **asymptotically minimum variance unbiased** if it is $A\mathcal{N}(\theta, \frac{v}{n})$ where $\frac{v}{n}$ is the Cramer-Rao lower bound.

Remark 20.186 *Our estimator is an AMVU estimator if the asymptotic variance is the CRLB.*

Unbiased estimators which are AMVU are sometimes said to be asymptotically efficient.

We now describe the procedure on how to show that an estimator $\hat{\theta}$ is AMVU. Suppose you had $X_1, ..., X_n$ iid with common pdf $f_\theta()$. After doing some work, you attempt to arrive at

$$C_\theta\Big[\hat{\theta}(X) - \theta\Big].$$

If you can arrive at this form, then θ is in fact a MVU estimator. However, if not, we can then try to show that it is an AMVU estimator. First, define $\eta = \eta(\theta)$ such that

$$C_\eta\Big[\hat{\eta}(X) - \eta\Big]$$

is of the correct form. This shows that $\eta(X)$ is an MVU estimator of η. Now, we also have that $\eta \sim AN(\eta, C_\eta)$. We then define $g(\hat{\eta})$ where we solve for θ. Compute $g(\hat{\eta})'$ then using the delta method, we get

$$\hat{\theta} \sim AN(g(\hat{\eta}) = \theta, g'(\hat{\eta})C_\eta = C_\theta).$$

Hence, this shows that $\hat{\theta}$ is an AMVU estimator as it now achieves the CRLB (C_θ).

20.6.7 MLE is AMVU

All the techniques we have looked at apply well to exponential families when looking at their maximum likelihood estimators where they were the solutions to the score equations

$$\frac{\partial}{\partial\theta} log f_\theta(\tilde{X}) = 0.$$

Furthermore, these methods were also method of moment estimators based on the sufficient statistic $t(\tilde{X})$, i.e. solutions to the moment equation

$$E_\theta[t(\tilde{X})] = t(\tilde{(X)}).$$

Lemma 20.187 *For exponential families, maximum likelihood estimation is equivalent to method of moments estimation.*

We now explore further properties of maximum likelihood estimates beyond exponential families.

We now have the following set up. Suppose $X_1, ..., X_n$ are iid continuous random variables with common pdf $f_\theta(x)$ for a family of pdfs $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ for some $\Theta \subset \mathbb{R}$. Suppose $\mathcal{F}$ is suitably regular and we can differentiate twice under the integral sign

$$\int_{-\infty}^{\infty} f_\theta(x)dx = 1$$

to get

$$\int_{-\infty}^{\infty} \frac{\partial^2}{\partial\theta^2} f_\theta(x)dx = 0.$$

Suppose we can also multiply and divide by $f_\theta(x)$. We then get the following theorem.

Theorem 20.188 *Assuming the set up above holds, we have the following information regarding the score function*

$$\begin{cases} E_\theta[\frac{\partial}{\partial\theta} log f_\theta(x)] = 0 \\ \\ Var_\theta[\frac{\partial}{\partial\theta} log f_\theta(x)] = -E_\theta[\frac{\partial^2}{\partial\theta^2} log f_\theta(x)] \end{cases}$$

which describes the mean and variance of the score function.

Now, recall that the maximum likelihood estimate (MLE) $\hat{\theta} = \hat{\theta}(X)$ is the solution to the score equation

$$\ell'(\theta; \tilde{X}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} log f_\theta(x_i) = 0.$$

Then, assuming $\hat{\theta}$ is close to the true parameter value θ_0, we note that

$$-\frac{\ell'(\theta_0; \tilde{x})}{\hat{\theta} - \theta_0} = \frac{\ell'(\hat{\theta}; \tilde{x}) - \ell'(\theta_0; \tilde{x})}{\hat{\theta} - \theta_0} \approx \ell^{''}(\theta_0; \tilde{x}).$$

Theorem 20.189 *We can approximate the MLE $\hat{\theta}$ by the following relationship*

$$\hat{\theta} \approx \theta_0 - \frac{\ell'(\theta_0; \tilde{x})}{\ell''(\theta_0; \tilde{x})}.$$

Resultantly, we have that

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\sqrt{n}\frac{\ell'(\theta_0; \tilde{x})}{\ell''(\theta_0; \tilde{x})}$$

where $\ell'(\theta_0; \tilde{x}) \xrightarrow{d} \mathcal{N}(0, I_\theta)$ and $\ell''(\theta_0; \tilde{x}) \xrightarrow{p} -I_\theta$ where $I_\theta = Var_\theta(\frac{\partial}{\partial\theta} log f_\theta(x_1))$.

Theorem 19: MLE attains CRLB asymptotically

The MLE $\hat{\theta}$ is asymptotically normal with mean equal to the true value θ_0 and variance equal to the Cramer-Rao Lower bound

$$\hat{\theta} \sim A\mathcal{N}(\theta, \frac{1}{nI_\theta}).$$

In a more rigorous manner,

$$I_n(\theta_0)^{-0.5}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$$

as $n \to \infty$.

To interpret this, we consider the **sampling distribution** of the MLE. That is, suppose we have had sampled several datasets $X_1, ..., X_j$ where the jth dataset gives the jth MLE $\hat{\theta}_j$. The distribution of the MLE is then the distribution of these realised $\hat{\theta}_j$ values, that is, the histogram of this is the **sampling distribution**. This sampling distribution is what has a normal distribution.

Remark 20.190 *This holds because the variance of the score function is the negative of the expected value of the second derivative of the score function*

$$Var_\theta[\frac{\partial}{\partial\theta} log f_\theta(x)] = -E_\theta[\frac{\partial^2}{\partial\theta^2} log f_\theta(x)]$$

Hence, maximum likelihood estimates are optimal under regularity conditions and hence their widespread use.

STAT3923: Advanced Statistical Inference

21. Further properties of score function

21.6.8 Further properties of score function

Definition 21: Efficient

Unbiased estimators which attain the CRLB are said to be **efficient**.

Definition 21.191 *(Asymptotically Efficient). Asymptotically normal estimators which are AMVU are said to be* ***asymptotically efficient****.*

Definition 22: Asymptotic Relative Efficiency

Let $\hat{\theta}$ be our candidate estimator. Then, the ratio

$$\frac{CRLB}{\text{Asymp } Var_\theta(\hat{\theta}(X)} \leq 1$$

is said to to be the asymptotic relative efficiency (ARE) of $\hat{\theta}(\tilde{x})$.

Remark 21.192 *To interpret ARE, if ARE = 85% for our candidate estimator $\hat{\theta}$, then the optimal AMVU estimator needs only 85% of the sample size to get the same precision as $\hat{\theta}(\tilde{X})$. The CRLB is the smallest value of variance and hence the higher the ARE is for $\hat{\theta}$, the closer it is in performance compared to the AMVU estimator.*

Definition 21.193 *(Fisher's Information per observation). Let $X_1, ..., X_n$ be iid with common pdf $f_\theta(.)$ that satisfies the regularity conditions. Then, the CRLB is $\frac{1}{nI_\theta}$ where*

$$I_\theta = Var_\theta[\frac{\partial}{\partial\theta} log f_\theta(X_1)]$$

is called the Fisher's information per observation. That is, it is the variance of the score function for ***one observation****.*

Remark 21.194 *The bigger Fisher's information per obervation, the more information is in the data and the smaller the variance of the estimator.*

Theorem 21.195 *Suppose $X = (X_1, ..., X_n)$ are iid RVs with common density $f_\theta(.)$ given by*

$$f_\theta(x) = g(x - \theta)$$

for a known PDF g(.) with a continuous derivative. Thus, θ is a location parameter.
Then, the score function is

$$\sum_{i=1}^{n} \psi(X_i - \theta) = \sum_{i=1}^{n} \frac{-g'(X_i - \theta)}{g(X_i - \theta)}.$$

Furthermore, for any unbiased estimator $\hat{\theta}(X)$ of θ, we have

$$Var\Big[\hat{\theta}(X)\Big] \geq \frac{1}{nI}$$

where $I = \int \frac{[g'(x)]^2}{g(x)} dx$.

21.6.9 Completeness

We learn about some other things for fun.

Definition 21.196 *(Completeness). A statistic $T = T(X)$ is complete if for every measurable function g*

$$E_\theta g(T) = 0$$

for all θ, then

$$P_\theta(g(T) = 0) = 1$$

for all θ.

That is, for every θ, the expectation of $g(T)$.

We now state why completeness is useful.

Theorem 21.197 *(Lehmann-Scheffé Theorem). If a statistic T is unbiased, complete, and sufficient for some parameter θ, then T is MVUE.*

Definition 21.198 *(Pivot). Let $X = (X_1, ..., X_n)$ from a distribution that depends on parameter θ. Let $g(X, \theta)$ be a random variable whose distribution is* ***the same*** *for all θ. Then g is called a pivotal quantity.*

Pivotal quantities do not depend on θ in their distribution. However, the actual quantities themselves may depend on θ.
Recall that a statistic is independent of the parameters θ.

Definition 21.199 *(Ancillary statistic). A statistic that is pivotal. That is, a statistic whose distribution does not depend on the parameters θ.*

Theorem 21.200 *(Basu's Theorem). A statistic that is both boundedly complete and sufficient is independent of any ancillary statistic.*

This theorem is useful to prove independence of two statistics by showing that one statistic is complete sufficient and the other is ancillary. For example, we can use Basu's theorem to show that the sample mean and sample variance are independent.

Definition 21.201 *(Minimal sufficient). A statistic T(X) is minimal sufficient if and only if T(X) is sufficient and for any other sufficient statistic T(X), then there exists a function f such that T(X) = f(S(X)).*

STAT3923: Advanced Statistical Inference

22. Hypothesis Testing

22.7 Hypothesis Testing

22.7.1 Hypothesis Testing

In statistical inference, we observe the realisation of random observations. However, due to the randomness, there are a range of possibilities in which the data could have arisen from. However, there are some possibilities which are too "different" with our realised data and hence we can disregard them. Therefore, we can only ever disprove something with data.

In hypothesis testing, we are interested in seeing for what realised values of $\mathbf{X}$, should we reject the null hypothesis. We reduce the data to a single statistic, known as a test statistic and use it to measure the strength of evidence against the hypothesis H_0. The p-value is a measure that is used. We are interested in identifying optimal level-α tests where we maximise the power of a test. The smaller the p-value, the stronger the evidence against the hypothesis we have.

Definition 22.202 *(Hypothesis). The hypothesis $H_0 \subset \mathcal{M}$ where $\mathcal{M}$ is a larger statistical model. We call the complement of H_0 the alternative hypothesis H_1 whilst H_0 is called the null hypothesis $\mathcal{M} = H_0 \cup H_1$ and $H_0 \cap H_1 = \emptyset$.*

Definition 22.203 *(Simple and Composite Hypothesis). A hypothesis containing only 1 distribution is called simple. Otherwise, it is called composite.*

Remark 22.204 *We look at 3 cases in this course. Simple vs simple, simple vs composite, and composite vs composite. The first one is easy to find optimal tests whereas the second and third only has optimal tests in certain circumstances.*

Definition 23: Power

The power of a test is

$$P(\text{reject } H_0 | H_1 \text{ is true}).$$

In other words, the power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\mathbb{X} \in R)$.

Remark 22.205 *Let the null hypothesis be $H_0 : \theta \in \Theta_0$. The ideal power function is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$.*

Definition 22.206 *(Type I/II Error). A type 1 error is when we incorrectly reject a true null hypothesis whereas a type 2 error is when we fail to reject a false null hypothesis.*

Definition 24: Size

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.

Definition 25: Level$-\alpha$ test

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a level α test if $size = \sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

Remark 22.207 *For a simple null hypothesis, the size and level of a test coincide.*

STAT3923: Advanced Statistical Inference

23. Simple vs Simple Hypothesis

23.7.2 Simple vs Simple Hypothesis

Here, we only have two specific distributions that we are looking at $x \sim f_0(.)$ and $x \sim f_1(.)$. Suppose that when H_0 is true, the random variable

$$y = \frac{f_1(x)}{f_0(x)}$$

the likelihood ratio has a continuous distribution. We then fix the level (probability of making a type 1 error) to $0 < \alpha < 1$. Suppose there is a unique value y_α such that

$$P_{f_0}(Y \geq y_\alpha) = \alpha$$

that is, the probabliity when the true distribution if f_0 that $Y \geq y_0$ is α. y_α is the upper α quantile of the random variable Y when the null hypothesis is true. We define the critical region to be

$$C = \{x : \frac{f_1(x)}{f_0(x)} \geq y_\alpha\}$$

then $P_{f_0}(c) = P_{f_0}(x \in C) = \alpha$. C is the set of observed values of X for which we reject the null hypothesis H_0.

Definition 23.208 *(Critical Region). The set*

$$C = \{x : \frac{f_1(x)}{f_0(x)} \geq y_\alpha\}$$

is defined as the critical region.

Theorem 20: Continuous Neyman-Pearson Lemma

Let $H_0 : X \sim f_0(.)$ and $H_1 : X \sim f_1(.)$ where H_0, H_1 are both simple. Then the **likelihood ratio**

$$\frac{f_1(X)}{f_0(X)}$$

is the **most powerful** test. That is, the power of the test based on the likelihood ratio is higher than the power of any other test.

If we let the critical region $C = \{x : \frac{f_1(x)}{f_0(x)} \geq y_\alpha\}$ be based on the likelihood ratio and D be any other critical region. What the continuous Neyman-Pearson lemma says is that the power for C is greater than the power for D, that is $P_{f_1}(C) = P_{f_1}(\tilde{X} \in C) \geq P_{f_1}(\tilde{X} \in D) = P_{f_1}(D)$.

If the likelihood ratio $Y = \frac{f_1(X)}{f_0(X)}$ has a discrete distribution, then there may be no exact value y_α such that $P_0(Y \geq y_\alpha) = \alpha$ for any given α. We introduce the concept of a randomised test to help us.

Definition 23.209 *(Test Function). Let $\delta(.)$ be a function taking values in [0,1]. That is, let $U \sim U[0,1]$ be independent of X. Then, we reject if $U \leq \delta(X)$.*

Remark 23.210 *The above construction of the test function means that we reject with probability $\delta(X)$.*

Theorem 21: Discrete Neyman-Pearson Lemma

Let X be a **disrete random variable**. Hence, the likelihood ratio of the alternative over the null is a discrete random variable. The most powerful test at level α of $H_0 : P_0(.)$ vs $H_1 : P_1(.)$ is given by the test function

$$\delta(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > y \\ \gamma & \frac{f_1(x)}{f_0(x)} = y \\ 0 & \frac{f_1(x)}{f_0(x)} < y \end{cases}$$

where y and α are chosen so that $\mathbb{E}_\theta[\delta(X)] = \alpha$.

Remark 23.211 *We will need to determine the critical value y and randomisation probability γ such that $\mathbb{E}_\theta[\delta(X)] = \alpha$.*

Hence, if $Y = \frac{f_1(.)}{f_0(.)}$ has a discrete distribution, then the value y satisfies $P_0(Y \geq y) \geq \alpha \geq P_0(Y > y)$ and γ is such that we usually have that

$$P_0(Y = y).\gamma + P_0(Y > y) = \alpha$$

therefore

$$\gamma = \frac{\alpha - P(Y > y)}{P(Y = y)}.$$

STAT3923: Advanced Statistical Inference

24. Simple vs Composite Hypothesis: UMP Tests

24.7.3 Simple vs Composite Hypothesis: One-sided UMP Tests

We now have the following set up. We have a family, depending only on 1 parameter $\{f_\theta(.) : \theta \in \Theta\}$ for some $\Theta \subset \mathbb{R}$ and we model the data x as observed values of $X \sim f_\theta(.)$ where $\theta \in \Theta$ is unknown. We want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta \backslash \theta_0$. We want to find a uniformly most powerful test.

Definition 26: Uniformly Most Powerful Test

Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta_0 \in \Theta$ versus $H_1 : \theta \in \Theta \backslash \theta_0$. A test $\delta_0(.)$ in class $\mathcal{C}$, is a uniformly most powerful (UMP) test at level α if

1. $E_{\theta_0}[\delta(X)] \leq \alpha$ (so the test is of level α)
2. $E_\theta[\delta_0(X)] \geq E_\theta[\delta_1(X)]$ (the test has the biggest power)

for any other test $\delta_1(.)$, which also has a level $E_{\theta_0}[\delta_1(X)] \leq \alpha$, for all $\theta \in \Theta \backslash \theta_0$ (i.e. where H_1 is true).

Remark 24.212 *With regards to notation, E_{θ_0} denotes the expectation under the null hypothesis being true whilst E_θ denotes the expectation under the alternative hypothesis.*

Hence, we want to find the test that has the most power out of all tests that have level α.

We can also frame this as an optimisation problem. Our goal is to find a UMP level-α test δ which

$$\text{Maximise power function } E_\theta[\delta(X)]$$

$$\text{subject to } E_{\theta_0} \leq \alpha.$$

For composite alternative hypotheses, the Neyman-Pearson lemma, and the likelihood ratio test, generalises if the data model $f_\theta(.)$ satisfies the monotone-likelihood ratio property. First, we define what is meant by the monotone likelihood ratio property.

Definition 27: Monotone Likelihood Ratio

A family of pdfs/pmfs $\{f_\theta(.) : \theta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}$ is said to have monotone likelihood ratio (MLR) in a test statistic T if for any arbitrary parameter values $\theta_0 < \theta_1$ in Θ, we had that

1. $f_{\theta_0}(.)$ and $f_{\theta_1}(.)$ are different distinct distributions;
2. The ratio $\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$ is a non-decreasing function of the test statistic T(X).

Corollary 24.213 *If the functions are first-differentiable, we can also say that the family has the monotone likelihood ratio if*

$$\frac{\partial}{\partial X}\left[\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}\right] \geq 0.$$

Remark 24.214 *Note that the ratio is a function of X and not the parameter θ.*

We can now state that when the density $\{f_\theta(.) : \theta \in \Theta\}$ has the MLR property, then a uniformly most powerful test exists.

Theorem 22: One-Sided Composite Tests with MLR

Suppose a family $\{f_\theta(.) : \theta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}$ has monotone likelihood ratio in the statistic $T(X)$. Then, for any $\theta_0 \in \Theta$, a uniformly most powerful level-α test exists for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ given by the test function

$$\delta(X) = \begin{cases} 1 & T(\tilde{X}) > c \\ \gamma & T(\tilde{X}) = c \\ 0 & T(\tilde{X}) < c \end{cases}$$

where C, γ are chosen to satisfy $E_{\theta_0}[\delta(\tilde{X})] = \alpha$.

Remark 24.215 *The UMP level-α test for $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ is obtained by swapping inequalities given in the theorem above.*

Remark 24.216 *Note that we need MLR to be an INCREASING function of T(X) to invoke the above theorem. So be careful.*

Theorem 24.217 *If the family $\{f_\theta(.) : \theta \in \Theta\}$ has a monotone likelihood ratio, the power functions of the UMP 1-sided tests are strictly monotone. That is, it increases up until 1 and then becomes constant.*

We now can revisit something we have seen from first year.

Definition 24.218 *(P-value). Given our test δ defined for densities with the MLR property, we reject the null hypothesis H_0 if $T(x) > c$. Suppose that the test statistic $T(x) = t$. We call the probability*

$$P(T > t)$$

*the **p-value**.*

Remark 24.219 *The smaller the p-value, the stronger the evidence against the null hypothesis.*

The reason for the introduction of p-values is that simply stating rejecting H_0 is not very informative. With a p-value, we can then know that if we reject a test at the smallest level α, we know that we will reject for $\alpha' > \alpha$.

Remark 24.220 *A high p-value is* ***not*** *in favor of H_0. It suggests that either H_0 is true or H_0 is false but our test has low power. Furthermore, the p-value is* ***not*** *the probability that H_0 is true conditioned on the data.*

We now state an important theorem to help us with testing.

Theorem 23: Monotone likelihood ratio for exponential families

All 1-parameter exponential families have monotone likelihood ratio, which are functions of their **sufficient statistics** T(X).

Hence, we have an important corollary to the fact that MLR exists for exponential families.

Proposition 6: Existence of 1-sided UMP tests for exponential families

All 1-parameter exponential families have a 1-sided UMP test.

24.7.4 Simple vs Composite Hypothesis: Two-sided UMPU Tests

We now look at two-sided tests. Suppose we are now testing a two-sided composite hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ for θ_0 in the interior of Θ.

Recall that we stated that the power function for UMP 1-sided tests are strictly monotone. The significance of this is that for a two-sided test, performing the one-sided test will be extremely biased towards one side. That is, the power function will be terrible for one side of the test. In particular, the power of the test will be below the level of the test

$$E_\theta[\delta(X)] < E_{\theta_0}[\delta(X)]$$

for some θ in the alternative.

Proposition 24.221 *(Power function is monotone for UMP 1-sided). The power function for a UMP 1-sided test is strictly monotone.*

Hence, this is terrible for our 2 sided test as the 1-sided UMP test will do well for one side of θ but have terrible power for the other side. In fact, it will have a power even less than the level α, which we do not want. This is known as a **biased test**. Hence, we need to restrict our analysis to tests that do not have this property.

Definition 28: Unbiased test

A test $\delta(.)$ is unbiased if its power function

$$E_\theta[\delta(X)] \geq E_{\theta_0}[\delta(X)] = \alpha$$

for all θ under H_1. That is, the power for rejecting a false hypothesis is higher than the level of the test.

Definition 29: UMPU Test

An unbiased uniformly most powerful test is a test that is uniformly most powerful and unbiased.

Theorem 24: Existence of 2-sided UMPU tests for exponential families

For a 1-parameter exponential family, an UMPU test always exist as it has a monotone likelihood function.

We can now state the theorem that guarantees us a UMPU level-α test for 1-parameter exponential families.

Theorem 25: Karlin-Rubin Theorem

For a 1-parameter exponential family with sufficient statistic T(X), a UMPU level-α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ exists and is given by

$$\delta(X) = \begin{cases} 1 & T(x) > C_2 \\ \gamma_2 & T(x) = C_2 \\ 0 & C_1 < T(x) < C_2 \\ \gamma_1 & T(x) = C_1 \\ 1 & T(x) < C_2 \end{cases}$$

where $\gamma_1, \gamma_2, C_1, C_2$ are chosen such that

$$\begin{cases} E_{\theta_0}[\delta(X)] = \alpha \\ \\ E_{\theta_0}[T(x)\delta(x)] = \alpha E_{\theta_0}[T(x)] \end{cases}$$

i.e. the level is α and the test function $\delta(.)$ is uncorrelated with the sufficient statistic $T(x)$.

Remark 24.222 *Notice that we now have UMPU rather than UMP tests. This is because we now* ***restrict*** *our attention to tests that are unbiased. Then, this allows us to now find uniformly most powerful tests, that is, a test that is a test with a uniformly higher power function compared to other tests.*

Theorem 24.223 *Suppose we have a 1-parameter family indexed by θ in some interval and suppose θ_0 is an interior point of that interval. If a UMP test of level $\leq \alpha$ exists for $H_0 : \theta = \theta_0$ against the two-sided alternative $H_1 : \theta \neq \theta_0$, then the test is automatically of exact level α and unbiased.*

STAT3923: Advanced Statistical Inference

25. Simple vs Composite: General Methods

25.7.5 Simple vs Composite: General Methods

We are now interested in generalising to simple vs composite testing for **when we don't have a monotone likelihood ratio**. Suppose we have an i.i.d sample x with a common pdf $f_\theta(.)$ for a 1-parameter family $\{f_\theta(.) : \theta \in \Theta\}$ for some $\Theta \subseteq \mathbb{R}$. Recall the Neyman Pearson likelihood ratio (NPLR) test statistic.

Definition 25.224 *(Neyman Pearson likelihood ratio test statistic). Let f_{θ_1} and f_{θ_0} be the pdf under θ_1 and θ_0 respectively. Then, the Neyman Pearson likelihood ratio test statistic is*

$$\frac{\prod_{i=1}^n f_{\theta_1}(x_i)}{\prod_{i=1}^n f_{\theta_0}(x_i)}.$$

Now if we have a composite null $H_1 : \theta \in \Theta \backslash \theta_0$, we can try estimate a θ_1-value and plug it in to get an approximation to the NLPR statistic.

Definition 30: Generalised Likelihood Ratio Test

The Generalised Likelihood Ratio Test (GLRT) for testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \in \Theta \backslash \theta_0$ uses the statistic

$$\frac{\prod_{i=1}^n f_{\hat{\theta}}(x_i)}{\prod_{i=1}^n f_0(x_i)}$$

where $\hat{\theta} = \max_{\theta \in \Theta} \prod_{i=1}^n f_\theta(x_i)$ is the MLE for θ over Θ.

We can also use the results we derived of the asymptotic properties of the MLE. Recall that we linearly approximated the score function and took a 2-term Taylor series about the true value θ_0. This gives us the next theorem.

Theorem 26: Limiting distribution of GLRT

Let $\hat{\theta}$ be the MLE. Then, the limiting distribution of the $\ell(\hat{\theta}; \tilde{X}) - \ell(\theta_0; \tilde{X})$ is $\frac{1}{2}\chi_1^2$ when H_0 is true.

Alternative formulation of the likelihood ratio statistic is

$$\lambda = 2log\Big(\frac{\sup_{\theta \in \Theta} \ell(\theta)}{\ell(\theta_0)}\Big)$$

where $\ell(\theta)$ is the likelihood.

STAT3923: Advanced Statistical Inference

26. Composite vs Composite: 1-Parameter Families

26.7.6 Composite vs Composite: 1-Parameter Families

We are now interested in the set up $x \sim f_\theta(x)$ for a 1-parameter family $\mathcal{F} = \{f_\theta(.) : \theta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}$. We wish to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta \backslash \Theta_0$ for some $\Theta_0 \subseteq \Theta$. For certain kinds of composite H_0's, optimal tests exists.

Proposition 7: Composite null with MLR property

If the family $\mathcal{F}$ has a monotone likelihood ratio in a statistic T(x), the **UMP** test of $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ is of the same form as for a simple null hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. That is, the test function is given by

$$\delta(X) = \begin{cases} 1 & T(X) > C \\ \gamma & T(X) = C \\ 0 & T(X) < C \end{cases}$$

where C, γ are chosen to satisfy $E_{\theta_0}[\delta(X)] = \alpha$.

Hence, our composite null hypothesis test now becomes a simple null hypothesis test against a composite null. Recall that UMP test exists for simple vs 1-sided composite hypothesis if our family of interest has a monotone likelihood ratio in the statistic T(X).
We see again that the exponential family has nice properties.

Proposition 8: Composite null for exponential family

If the family $\mathcal{F}$ is a 1-parameter exponential family, a **UMP** test for $H_0 : \{\theta \leq \theta_1\} \cup \{\theta > \theta_2\}$ against $H_1 : \theta_1 < \theta < \theta_2$ exists and is of the form

$$\delta(X) = \begin{cases} 1 & C_1 < T(x) < C_2 \\ \gamma_i & T(x) = C_i, i = 1, 2 \\ 0 & T(x) < C_1 \text{ or } T(x) > C_2 \end{cases}$$

where C_i, γ_i are chosen to satisfy

$$E_{\theta_1}[\delta(x)] = E_{\theta_2}[\delta(x)] = \alpha$$

i.e. on the endpoints of the interval, the level is exactly α.

We can state a stronger proposition for 1-parameter exponential families for when our null hypothesis is inside an interval with our compsite being outside of that interval.

Proposition 26.225 *(Interval composite null for exponential family). If $\mathcal{F}$ is a 1-parameter exponential family with a sufficient statistic T(x), then the **UMPU** test of $H_0 : \theta_1 \leq \theta \leq \theta_2$ against $H_1 : \{\theta < \theta_1\} \cup \{\theta > \theta_2\}$ is of the form*

$$\delta(X) = \begin{cases} 1 & \{T(x) < C_1\} \cup \{T(x) > C_2\} \\ \gamma_i & T(x) = C_i, i = 1, 2 \\ 0 & C_1 < T(x) < C_2 \end{cases}$$

where the C_i, γ_i are chosen so that

$$E_{\theta_1}[\delta(x)] = E_{\theta_2}[\delta(x)] = \alpha$$

i.e. on the endpoints of the interval, the level is exactly α.

Remark 26.226 *The limiting version of this test as $\theta_1 \to \theta_0 \leftarrow \theta_2$ is the **UMPU** test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.*

26.7.7 Composite vs Composite: Multi-Parameter Families

We are now interested in testing composite nulls against composite alternatives for multi-parameter families. Previously, what we have established worked for one parameter exponential families. We now wish to generalise this and give an approximation to the NPLR statistic. That is $H_0 : \theta \in \Theta_0 \subset \Theta$ against $H_1 : \theta \in \Theta \backslash \Theta_0$.

Definition 31: Generalised Likelihood Ratio Test for composite null

The Generalised Likelihood Ratio Test (GLRT) for testing $H_0 : \theta \in \Theta_0 \subset \Theta$ against $H_1 : \theta \in \Theta \backslash \Theta_0$ uses the statistic

$$\ell(\hat{\theta}; x) - \ell(\hat{\theta}_0; x)$$

where $\hat{\theta}$ is the unrestricted maximum likelihood estimator whilst $\hat{\theta}_0 = \max_{\theta \in \Theta_0} \ell(\theta; x)$ is the null-restricted m.l.e.

We now describe the limiting distribution of the GLRT statistic.

Theorem 27: Wilk's Theorem

Suppose we had a vector of parameters $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta \backslash \Theta_0$. Under the conditions of

1. Smoothness/differentiability
2. Support being independent of θ
3. Θ_0 consists of only interior points of Θ
4. Identifiability: $\theta_1 \neq \theta_2$ means that $f_{\theta_1}(.) \neq f_{\theta_2}(.)$.

Suppose that H_0 imposes k constraints on the parameter vectors. Then

$$2\{\ell(\hat{\theta}); x) - \ell(\hat{\theta}_0; x)\} \xrightarrow{d} \chi^2_k$$

where $\hat{\theta}_0$ is the m.l.e under Θ_0, ℓ is the log likelihood and k is the difference in dimension between Θ and Θ_0.

Remark 26.227 *This allows us to convert observations into p-values.*

STAT3923: Advanced Statistical Inference

27. GLRT Examples

27.7.8 GLRT Examples

Theorem 27.228 *The 1-way ANOVA F-test is an example of the GLRT.*

Theorem 27.229 *The maximisation of the one sided likelihood GLRT is equivalent to a 1-sided t-test.*

27.7.9 Simulation based p-values

What should we do if we can't utilise large sample theory to run inference tests. We can use simulation techniques.

Definition 32: Monte-Carlo p-value

Suppose we are testing a simple null hypothesis $H_0 : \theta = \theta_0$. We can then simulate data x from the null hypothesis distribution $f_{\theta_0}(.)$ and generate an arbitrary number of realisations. We then construct a statistic T(x) from our realised sample. We repeat this process an arbitrary number of times to generate a sampling distribution of the statistic T(x). This gives us a Monte-Carlo p-value.

Algorithm 1: Size Study

Input: Data Generating Process such that H_0 is true
Output: Size level
$x^{(i)} \leftarrow$ 10000 draws from DGP with true H_0
$t^{(i)} \leftarrow$ test statistic
$size \leftarrow$ count fraction of rejections of H_0 over $\{t^{(i)}\}$
return *size*

Definition 33: Parametric Bootstrap

Suppose we are testing a composite null hypothesis $H_0 : \theta \in \Theta_0$. We estimate the null θ_0 using $\hat{\theta}_0$ under H_0. We then sample from the distribution $f_{\hat{\theta}_0}(.)$ to generate arbitrary sample sizes and construct statistics from this. Hence, we can generate a sampling distribution for our statistic T(x).

Algorithm 2: Parametric Bootstrap for one sided test

Input: Data Generating Process F_θ with unknown θ
Output: P-value
$\hat{\theta} \leftarrow$ from data x which estimates θ
$\gamma(x) \leftarrow$ test statistic with $\hat{\theta}$ and x
for $s \leftarrow 0$ **to** S **do**
  $x^{(s)} \leftarrow$ sample from DGP $F_{\hat{\theta}}$ with $\hat{\theta}$ as parameter
  $\gamma(x^{(s)}) \leftarrow$ test statistic on $x^{(s)}$
p-value $\leftarrow \frac{\text{number of draws with } \gamma(x^{(s)}) \leq \gamma(x)}{S}$
if *p-value* $< \alpha$ **then**
  return *Reject* H_0
else
  return *Don't reject* H_0

STAT3923: Advanced Statistical Inference

28. Simple Prediction Problems

28.8 Statistical Decision Theory

28.8.1 Simple Prediction Problems

We are interested in the set up where Y is a random variable with a **known distribution**. We define D to be an arbitrary set called the **decision space**.

Definition 28.230 *(Loss). For each possible value y of Y and decision $d \in D$, we suffer a loss $L(d|y)$.*

Definition 28.231 *(Risk). The expectation of the loss is the risk*

$$R(d) = E[L(d|y)]$$

where we aim to minimise the risk.

Definition 28.232 *(Admissible). Let $\tilde{d}(.)$ be a procedure. We say that $\tilde{d}(.)$ is admissible if*

$$R(\theta, \tilde{d}) \leq R(\theta, d) \quad \forall \theta \in \Theta$$

and

$$R(\theta, \tilde{d}) < R(\theta, d) \quad \forall \theta \in [a, b] \subset \Theta$$

for all other procedures d(.).

Theorem 28.233 *In the simple prediction problem, if we define the loss to be the **squared error loss** $L(d|y) = c(d-y)^2$ for $c > 0$, the optimal decision d to minimise this is the mean $d = E(y)$.*

Proof:(Sketch). Look at $R(d|y) = E[L(d|y)]$ and then find first order conditions to find optimal value for d. ■

Theorem 28.234 *In the simple prediction problem, if we define the loss to be the **absolute error loss** $L(d|y) = c|d-y|$ for $c > 0$, the optimal decision d to minimise this is the median $d = F^{-1}(\frac{1}{2}) = median(Y)$.*

Theorem 28.235 *In the simple prediction problem, if we define the loss to be the **0-1 error loss***

$$L(d|y) = 1\{|d-y| > c\} = \begin{cases} 1 & \text{if } |d-y| > c \\ 0 & \text{if } |d-y| \leq c \end{cases}$$

where the non-coverage of y be the interval $d \pm c$. Furthermore, we also assume that $f(.)$ is unimodal. The optimal decision d is chosen so that the interval $d \pm c$ is a level set of f(.). If the pdf f(.) is symmetric about m, the optimal d is m.

Theorem 28.236 *For the simple prediction problem where Y has a strictly increasing, continuous CDF $F(.)$ and $\mu = E(Y)$ exists and is finite and the decision space $D = \mathbb{R}$. We assume the loss is the asymmetric piecewise-linear loss function given by*

$$L(d|y) = \begin{cases} p(y-d) & d < y \\ \\ (1-p)(d-y) & d > y \end{cases}$$

for some $p \in (0,1)$. Then, the decision d that minimises the risk is

$$d = F^{-1}(p)$$

that is, the p-th quantile of $F(.)$.

STAT3923: Advanced Statistical Inference

29. Discrete Selection Problem

29.8.2 Discrete Selection Problem

Suppose $\mathbb{R}$ is partitioned into sets $S_1, ..., S_k$. The decision space is $D = \{1, 2, ..., k\}$. Our goal is to guess which set does y belong to. Hence, the loss function is

$$L(d|y) = \sum_{j=1}^{k} L_{d_j} 1\{y \in S_j\}$$

where we can construct the the **loss matrix** $\mathbb{L} = \{L_{d_j}; j = 1, ..., k\}$ where $L_{dd} = 0$ and $L_{dy} \geq 0$.

We can define a column vector $\tilde{p}$ where $p_j = P(Y \in S_j)$ for $j = 1, ..., k$. Then, the risk of a decision d is

$$R(d) = \sum_{j=1}^{k} L_{d_j} P(Y \in S_j) = \mathbb{L}_{\tilde{p}_j}.$$

In other words, the risk of decision d is the loss associated with the probability of y being in d.

29.8.3 Special case of Discrete Selection

Suppose that the loss matrix defined earlier $\mathbb{L}$ only depends on

1. The observed value y of Y
2. Whether the decision d is right or wrong.

Recall that for our loss matrix $\mathbb{L}$, the columns are the possible values of Y and the rows are the different decisions we label Y to be. So, $\mathbb{L}_{dd} = 0$ and $\mathbb{L}_{d_j} = L_j$ where $d \neq j$. That is, the diagonals of our loss matrix is 0 and the off-diagonals are the same in each column.

Then, the risk of decision d reduces to

$$R(d) = E[L(d|y)] = \sum_{j=1}^{k} L_{d_j} P(Y \in S_j) = \sum_{j=1 \wedge j \neq d}^{k} L_j P(Y \in S_j)$$

$$= \sum_{j=1}^{k} L_j P(Y \in S_j) - L_d P(Y \in S_d).$$

Hence, minimising $R(d)$ is equivalent to maximising the product $L_d P(Y \in S_d)$. Hence, we label Y to be in set S_d if it is highly likely or we pay a big price of $Y \in S_d$ if we don't pick it as L_d is the loss we pay if Y lands in S_d but we did not pick S_d.

STAT3923: Advanced Statistical Inference

30. Statistical Decision Theory

30.8.4 Statistical Decision Theory

Definition 30.237 *(Statistical Decision Framework). We specify the setup for the framework.*

1. *A family* $\mathcal{F} = \{f_\theta(.) : \theta \in \Theta\}$ *of distributions for a random vector* $\tilde{x}$ *taking values in a space* X.
2. *A decision space* $\mathcal{D} = \{d\}$.
3. *Non-negative valued loss function such that when a decision* $d \in D$ *is made, the true distribution* $f_\theta(.)$ *is made and the true distribution is* $f_\theta(.)$ *a loss of* $L(d|\theta)$ *is suffered.*

The problem is then to choose a D-valued decision function d(.) defined on the sample space $d : X \to D$.

Definition 34: Risk Function

Each decision function $d : X \to D$ has an associated risk function

$$R(\theta|d(.)) = E_\theta[L(d(x)|\theta)]$$

which measures the long-run average loss suffered using the decision function $d(.)$ and when $x \sim f_\theta(.)$.

Remark 30.238 *Comparing decision functions reduces down to comparing their respective risk functions.*

The issue is that we cannot compare risk functions pointwise between decisions functions as we can simply define biased decision functions which work extremely well for certain values of θ. Hence, pointwise comparison of risk functions does not work. We instead use two alternative measures of risk. That is, Bayes Risk and maximum risk.

Definition 35: Bayes Risk

Let w(.) be a non-negative weight function. We define the Bayes risk as

$$B_w(d) = \int_\Theta w(\theta) R(\theta|d) d\theta = \int_\Theta w(\theta) E_\theta[L(d(x)|\theta)] d\theta \leq +\infty.$$

Definition 36: Bayes Decision Rule

If the decision $\tilde{d}$ is such that

$$B_w(\tilde{d}) \leq B_w(d)$$

for any other decision function $d(.)$, then $\tilde{d}$ is said to be a **Bayes decision rule** with respect to the weight function w(.).

Definition 37: Maximum over a subset risk

For a given subset $\Theta_0 \subseteq \Theta$, a decision rule $\hat{d}$ is said to be minimax (over Θ_0) if

$$\max_{\theta \in \Theta_0} R(\theta|\hat{d}) \leq \max_{\theta \in \Theta_0} R(\theta|d)$$

for any other decision function $d(.)$. Hence, we minimise the maximum risk.

30.8.5 Finding Bayes Decision Rules

It is quite easy to find Bayes decision rules. We can find Bayes decision rules by reducing our problem into a simple prediction problem. Recall that the Bayes risk of a decision rule $d(.)$ is

$$\begin{aligned} B_w(d) &= \int_\Theta w(\theta) R(\theta|d) d\theta \\ &= \int_\Theta w(\theta) \Big[E_\theta [L(d(\tilde{x})|\theta)] \Big] d\theta \\ &= \int_\Theta w(\theta) \Big[\int ... \int_X L[d(\tilde{x}|\theta)] f_\theta(\tilde{x}) d\tilde{x} \Big] d\theta \\ &= \int ... \int_X \Big[\int_\Theta L[d(\tilde{x}|\theta)] w(\theta) f_\theta(\tilde{x}) d\theta \Big] d\tilde{x} \end{aligned}$$

where the last equality arises from Tonelli's theorem.

Theorem 30.239 *(Tonelli's theorem). Assume that f is a non-negative measurable function. Then*

$$\int_X \Big(\int_Y f(x,y) dy \Big) dx = \int_Y \Big(\int_X f(x,y) dx \Big) dy = \int_{X \times Y} f(x,y) d(x,y).$$

Assuming $w(\theta) f_\theta(\tilde{x})$ is ingerable, we define

$$m(\tilde{x}) = \int w(\theta) f_\theta(\tilde{x}) d\theta.$$

Hence, we have the Bayes risk as

$$B_w(d) = \int ... \int_X m(\tilde{x}) \Big[\int_\Theta L[d(\tilde{x}|\theta)] \frac{w(\theta) f_\theta(\tilde{x})}{m(\tilde{x})} d\theta \Big] d\tilde{x}$$

Definition 30.240 *Using the definition of conditional probabliity, we define*

$$p(\theta|\tilde{x}) = \frac{w(\theta) f_\theta(\tilde{x})}{m(\tilde{x})}$$

as the probabliity density of θ.

Definition 30.241 *(Bayes Risk). We define the Bayes risk as*

$$B_w(d) = \int ... \int_X m(\tilde{x}) \Big[\int_\Theta L[d(\tilde{x})|\theta] p(\theta|\tilde{x}) d\theta \Big] d\tilde{x}.$$

Our aim is to choose the decision $d(\tilde{x})$ that minimises the innter integral $\int_\Theta L[d(\tilde{x})|\theta]p(\theta|\tilde{x})d\theta$ as that will minimise the Bayes risk. Note that this is **exactly** the same form of the simple prediction problem based on a single draw from $p(\theta|\tilde{x})$ with loss $L[d|\theta]$.

Theorem 30.242 *If $\tilde{d}(\tilde{x})$ was a decision rule such that*

$$\int_\Theta L[\tilde{d}(\tilde{x})|\theta]p(\theta|\tilde{x})d\theta \leq \int_\Theta L[d(\tilde{x})|\theta]p(\theta|\tilde{x})d\theta$$

for any other decision rule d(.), then we have that

$$B_w(\tilde{d}) \leq B_w(d).$$

Definition 30.243 *(Posterior Density). The density $p(\theta|\tilde{x})$ is known as the **posterior density.***

Definition 30.244 *(Prior Density). If the weight function w(.) is a density, then it is known as a **prior density**.*

Hence, finding a Bayes rule is solving a simple prediction problem from a single draw of the posterior density.

STAT3923: Advanced Statistical Inference

31. Bayesian Interpretation

31.8.6 Bayesian Interpretation

We now consider setting our weight function w(.) to be a distribution. That is, $\int_\Theta w(\theta)d\theta = 1$.

Definition 31.245 *(Prior Distribution). Let w(.) be a probability density function. Then, we say that w(.) is a prior distribution.*

Definition 31.246 *(Bayes Theorem). Let θ have a continuous PDF f(.). Then Bayes theorem states that*

$$f(\theta|x) = \frac{f(x|\theta)w(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

where the denominator does not depend on θ.
If we had n i.i.d samples x, then we have that

$$f(\theta|\tilde{x}) = \frac{\prod_{i=1}^n f(x_i|\theta)w(\theta)}{\int f(x|\theta)f(\theta)d\theta} = \frac{L_n(\theta)w(\theta)}{c} \approx L_n(\theta)w(\theta)$$

where c is a normalising constant and $L_n(\theta)$ is the likelihood with sample size n.

Theorem 31.247 *The posterior is approximately the prior times the likelihood*

$$f(\theta|\tilde{x}) \approx L_n(\theta)w(\theta).$$

Theorem 28: Bayes Decision Rules

Let the decision space $D = \mathbb{R}$ and suppose $\tilde{x} = (x_1, ..., x_n)$ are iid random variables $f_\theta(.)$ for $\theta \in \mathbb{R} = \Theta$ with a loss function $L(d|\theta)$. Unless otherwise, let $d = \mathbb{R}$.

1. If $L(d|\theta) = (d-\theta)^2$, the Bayes decision rule is the **mean of the posterior distribution**.
2. If $L(d|\theta) = |d-\theta|$, the Bayes decision rule is the **median of the posterior distribution**.
3. If $L(d|\theta) = 1\{|d-\theta| > c\}$, the Bayes decision rule is the **midpoint of the level set of width 2c**.
4. Let $d = \{0, 1\}$. Then, define the loss function to be

$$L(d|\theta) = \begin{cases} L_0 & d = 1, \theta \leq 0 \\ L_1 & d = 0, \theta > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let p_1 be the probability placed on $(0, \infty) = \Theta_1$ by the posterior distribution and $p_0 = 1 - p_1$. Then, the Bayes decision rule is to choose 0 if $p_0 L_0 > p_1 L_1$ and 1 if $p_1 L_1 > L_0 p_0$.

The following is useful for helping us determine the risk with absolute error loss.

Lemma 31.248 *Suppose $Z \sim \mathcal{N}(0,1)$. Then, for any constant c,*

$$E_\theta\{|c+Z|\} = c\Big[1 - 2\Phi(-c)\Big] + \frac{2e^{-\frac{1}{2}c^2}}{\sqrt{2\pi}}.$$

Lemma 31.249 *Suppose $Z \sim \mathcal{N}(0,1)$. Suppose $c_n \to 0$ as $n \to \infty$. Then*

$$\lim_{n\to\infty} E_\theta\{|c_n + Z\} = \sqrt{\frac{2}{\pi}}.$$

Lemma 31.250 *Let $f(\theta|\tilde{x})$ be the posterior of θ. We can compute the point estimate by computing the mean of the posterior*

$$\frac{1}{c}\int L_n(\theta)w(\theta)\theta.$$

Definition 31.251 *(Posterior interval). Suppose we want to find a, b such that*

$$\int_{-\infty}^{a} f(\theta|\tilde{x})d\theta = \int_{b}^{\infty} f(\theta|\tilde{x})d\theta = \frac{\alpha}{2}.$$

Let $C = (a,b)$. Then

$$P(\theta \in C|\tilde{x}) = \int_a^b f(\theta|\tilde{x})d\theta = 1-\alpha.$$

C is called the $1-\alpha$ posterior interval.

Definition 31.252 *(Flat prior). Let the weight function $w(.) \approx \gamma$ where γ is a constant. Then $w(.)$ is known as a flat or uninformative prior.*

Definition 31.253 *(Improper prior). Let $w(.)$ be a weight function such that $\int w(\theta)d\theta = \infty$. $w(.)$ is not a probability density function and hence is referred to as an improper prior.*

Lemma 31.254 *Flat prior are not transformation invariant. That is, a flat prior on a parameter θ does not imply a flat prior on the transformed version of the parameter $\tau(\theta)$.*

Definition 31.255 *(Jeffrey's prior). A prior that is transformation invariant is known as Jeffrey's prior and is defined by*

$$w(\theta) \approx I(\theta)^{\frac{1}{2}}$$

where $I(\theta)$ is the Fisher information function. The Jeffrey's prior is transformation invariant.

STAT3923: Advanced Statistical Inference

32. Bayesian vs Frequentist

32.8.7 Bayesian vs Frequentist

We now describe the differences between Bayesian and frequentist approach to statistical modelling.

The frequentist approach to statistical model supposes that the data was generated from a fixed distribution from a known family. That is, the family $\{f_\theta(.) : \theta \in \Theta\}$ is given and the distribution of the data in $f_\theta(.)$ for some unknown but fixed θ. Inference then consists of hypothesis tests, point and interval estimates.

The Bayesian approach is to specify a known prior distribution $w(.)$ on Θ and assume the data was obtained by first drawing a value θ from Θ according to $w(.)$ and then conditional on θ, the data has distribution $f_\theta(.)$. Inference is done on the posterior distribution $p(\theta|\tilde{x})$ of θ given $\tilde{x}$.

We assume the frequentist point of view in this course. We assume there is a fixed non-random but unknown true parameter value.

Bayesian procedures have very desirable frequentist properties. We do not require that our weight functions are integrable (proper priors). Even if w(.) is not integrable, the resulting posterior may still be integrable.

We now look at a family of weight functions with nice properties. We can select the weight function w(.) in such a way that the corresponding posterior is of the same distribution.

Definition 38: Conjugate Family

Let $\mathcal{F}$ denote the class of PDFs $f(x|\theta)$. A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in X$.

Parameter θ	Conjugate Prior $w(\theta)$
Normal Mean	Normal
Binomial Success probability	Beta
Poisson	Gamma
Gamma Rate	Gamma
Gamma Scale	Inverse Gamma
Normal variance	Inverse Gamma
$U(0, \theta)$	Pareto
Pareto Shape	Gamma

We can use Bayes procedures as theoretical tools for finding minimax procedures. The main takeaway is that Bayes estimators with a constant risk function are minimax.

Theorem 29: Minimax

Suppose that for k = 1,2,..., $d_k(.)$ is the Bayes procedure with respect to a proper prior $w_k(.)$ on Θ and the loss function $L(.|\theta)$. If the procedure $\tilde{d}(.)$ is such that

$$\max_{\theta\in\Theta} \mathbb{E}_\theta[L(\tilde{d}(\tilde{x})|\theta)] \leq \lim_{k\to\infty} B_{w_k}(d_k(.))$$

then $\tilde{d}(.)$ is minimax over Θ for $L(.|\theta)$. Furthermore, w_k is called a least favorable prior.

32.8.8 Minimax Procedures

Recall that a decision function $\tilde{d}(.)$ such that for $\Theta_0 \subseteq \Theta$

$$\sup_{\theta\in\Theta_0} E_\theta[L(\tilde{d}(\tilde{x})|\theta)] \leq \sup_{\theta\in\Theta_0} E_\theta[L(d(\tilde{x})|\theta)]$$

for any other decision function $d(.)$, then $\tilde{d}(.)$ is said to be minimax over Θ_0 for the loss function $L(.|\theta)$. However, these procedures are harder to find compared to Bayes procedures. However, they have the advantage of not requiring the choice of a weight function.

Proposition 9: Average less than max

For any estimator d(.) and proper prior w(.),

$$B_w(d(.)) \leq \sup_{\theta\in\Theta} E_\theta[L(d(\tilde{x})|\theta)].$$

That is, the Bayes risk is always less than the maximum risk.

From the above lemma, if we can show that the **maximum risk is less than the Bayes risk, we then can conclude that the Bayes risk is equal to the maximum risk for an estimator and hence it is a minimax estimator**. We have 2 theorems that use Bayes procedures as theoretical tools for finding minimax procedures.

Theorem 30: Minimax Estimator

Suppose that for $k = 1, 2, ...,$ $d_k(.)$ is the Bayes procedures with respect to a **proper prior** $w_k(.)$ on Θ and the loss function $L(.|\theta)$. If the procedure $\tilde{d}(.)$ is such that

$$\max_{\theta\in\Theta} E_\theta[L(\tilde{d}(\tilde{x})|\theta)] \leq \lim_{k\to\infty} \int E_\theta[L(d_k(\tilde{x})|\theta)]w_k(\theta)d\theta$$

$$= \lim_{k\to\infty} B_{w_k}(d_k(.))$$

then $\tilde{d}(.)$ is a minimax estimator over Θ for $L(.|\theta)$.

Theorem 31: Hodges and Lehman

Suppose d(.) is a Bayes procedure with respect to $L(.|\theta)$ and a **proper prior** $w(.)$ on Θ. Let Θ_0 denote the support of $w(.)$. Suppose the following conditions hold

1. $\mathbb{E}_\theta[L(d(\tilde{x})|\theta)] = c$ for $\theta \in \Theta_0$
2. $\mathbb{E}_\theta[L(d(\tilde{x})|\theta)] \leq c$ for $\theta \in \Theta$

then $d(.)$ is minimax over Θ for $L(.|\theta)$.

Remark 32.256 *The reason we have w_k now is that we may have our prior $w(.)$ to be an improper prior or a flat prior, hence $w_k(.)$ are a sequence of proper priors that converges to this prior. This allows us to make sure of the 2 theorems above.*

Remark 32.257 *An estimator with constant risk is a minimax estimator.*

Proposition 10: Showing a Bayes procedure is minimax

Given a Bayes procedure with a proper prior, if we can show that it has a constant risk, then it is minimax.

STAT3923: Advanced Statistical Inference

33. Decision Theory Recap

33.8.9 Decision Theory Recap

In regression analysis, we have the covariates $X_1, ..., X_p$ and the best least squares prediction is given by the conditional mean $d(x_1, ..., x_p) = E[Y|X_1 = x_1, ..., X_p = x_p]$.

The Bayesian approach is to include a prior distribution $w(\theta)$ and make inferences from the posterior distribution

$$f(\theta|\tilde{x}) = \frac{w(\theta)f(\tilde{x}|\theta)}{m(\tilde{x})} \approx w(\theta)f(\tilde{x}|\theta)$$

where $f(\tilde{x}|\theta)$ is the likelihood function and $m(\tilde{x}) = \int_\Theta w(\theta)f(\tilde{x}|\theta)d\theta$ is the marginal likelihood.

$w(\theta)$ represents the beliefs about θ before observing $\tilde{X}$. $f(\theta|\tilde{x})$ represents beliefs about θ after observing $\tilde{X}$.

Definition 39: Posterior Expected Loss

We define the posterior expected loss as

$$\int_\Theta L(d(\tilde{x})|\theta)f(\theta|\tilde{x})d\theta.$$

Remark 33.258 *Here, we take a single draw from the posterior and predict what it is.*

Lemma 33.259 *The Bayes decision rule minimises the posterior expected loss.*

The Bayes decision rule to minimise the posterior expected loss is the same form as a simple prediction problem of a single draw from the posterior distribution $\theta|X$.

Definition 40: Limiting Risk

We define the limiting risk as the "long term risk" of an estimator

$$\lim_{n\to\infty} nR(\theta|d).$$

Theorem 32: MLE under regularity

For models that satisfy regularity conditions, MLE and Bayes estimators with reasonable priors $w(.)$ have the same large sample performance.

STAT3923: Advanced Statistical Inference

34. Non-regular Bayes estimation

34.8.10 Non-regular Bayes estimation

As stated in the last section, if a model is regular, then the MLE and Bayes estimators with reasonable priors will have similar limiting rescaled risk. We will now look at what happens when a model is **not regular**.

Definition 34.260 *(Pareto Distribution). The CDF of the Pareto (γ, m) distribution with shape parameter $\gamma > 0$ and scale $m > 0$ is*

$$F(y;\gamma,m) = \begin{cases} 0 & y < m \\ 1 - (\frac{m}{y})^\gamma & y \geq m. \end{cases}$$

The corresponding PDF is

$$f(y;\gamma,m) = \begin{cases} 0 & y < m \\ \frac{\gamma m^\gamma}{y^{\gamma+1}} & y \geq m. \end{cases}$$

Theorem 34.261 *The Pareto distribution is heavy-tailed, and the k-th moment exists only if γ is sufficiently large*

$$\mathbb{E}[Y^k] = \begin{cases} m^k \frac{\gamma}{\gamma - k} & k < \gamma \\ \infty & k \geq \gamma. \end{cases}$$

Theorem 34.262 *The mean of the Pareto distribution is*

$$\mathbb{E}[Y] = \frac{m\gamma}{\gamma - 1} \quad \gamma > 1.$$

We can now look at our irregular model. Suppose $X_1, ..., X_n$ are iid $U(0, \theta)$ random variables where $\theta \in (0, \infty)$ is unknown. For the squared error loss, we wish to determine the limiting risk $\lim_{n\to\infty} n^2 R(\theta|d)$ for the MLE $d(\tilde{X}) = \hat{\theta}_{MLE}$ and for the Bayes estimator with respect to the flat prior $w(\theta) = 1$, which we denote by $\hat{\theta}_{flat}$.

Lemma 34.263 *The MLE is the sample maximum*

$$\hat{\theta}_{MLE} = X_{(n)}.$$

Lemma 34.264 *The Bayes estimator with a flat prior and uniform likelihood function, has a posterior distribution proportional to the Pareto distribution*

$$f(\theta|\tilde{X}) = \frac{(n-1)X_{(n)}^{n-1}}{\theta^n} 1\{\theta \geq X_{(n)}\}.$$

Lemma 34.265 *Under squared error loss, the Bayes estimator d_{flat} is the posterior mean, which in this case is*

$$\hat{\theta}_{flat} = \frac{(n-1)X_{(n)}}{n-2} \quad n > 2.$$

Lemma 34.266 *The risk is the MSE under squared error loss.*

Lemma 34.267 *The limiting rescaled risk of the MLE* $\hat{\theta}_{MLE}$ *is*

$$\lim_{n\to\infty} n^2 R(\hat{\theta}_{MLE}|\theta) = \theta^2 + \theta^2.$$

Lemma 34.268 *The limiting rescaled risk of the Bayes estimator with a flat prior is*

$$\lim_{n\to\infty} n^2 R(\hat{\theta}_{flat}|\theta) = 0 + \theta^2 = \theta^2.$$

Remark 34.269 *Hence, we see that* $\hat{\theta}_{flat}$ *dominates* $\hat{\theta}_{MLE}$ *for large n.*

As $U(0,\theta)$ is not a regular model as the support depends on θ and θ is on the boundary of Θ. For such models, the MLE does not do well.

STAT3923: Advanced Statistical Inference

35. Asymptotically Minimax Estimator

35.9 Asymptotically Minimax Procedures

35.9.1 Asymptotically Minimax Estimator

Minimax estimators that are global (over the entire parameter space) are quite rare.

Theorem 35.270 *For a large sample size, the MLE will have a risk smaller than the risk of the minimax estimator, except for a small subset $\theta \in \Theta$.*

It is often impossible to find exact minimax estimators. Hence, we relax conditions and now focus on deciding whether to maximimise **max subset risk** or **limiting maximum risk**.

Definition 35.271 *(Max subset risk). The maximum subset risk for a decision d over an interval $[a, b] \subset \Theta$ is*

$$\max_{a \leq \theta \leq b} R(\theta|d).$$

Definition 41: Limiting (rescaled) maximum risk

The limiting (rescaled) maximum risk for a sequence of decisions $\{d_n\}_{n \geq 1}$ over an interval $[a, b] \subset \Theta$ is

$$\lim_{n \to \infty} \max_{a \leq \theta \leq b} R(\theta|d_n).$$

Definition 42: Asymptotically Minimax Estimators

An estimator $d(\tilde{x})$ which minimises the limiting (rescaled) maximum risk $\lim_{n \to \infty} \max_{a \leq \theta \leq b} R_n(\theta|d)$ over all choices $d(.)$ is called an **asymptotically minimax estimator**.

For a sequence of decisions d_n, to show that it is asymptotically minimax, we have two steps.

1) First, we determine a lower bound to

$$\lim_{n \to \infty} \max_{a \leq \theta \leq b} R_n(\theta|d)$$

for **any** procedure $R_n(\theta|d)$. Note that this is remarkable as we do not require the estimator to be unbiased or regular. The drawback though is that this gives us only a limiting result.

2) We then show that the procedure attains this lower bound.

STAT3923: Advanced Statistical Inference

36. Hodges' estimator and Superefficiency

36.9.2 Hodges' estimator and Superefficiency

This section, we look at showing that asymptotic mean-variance unbiased estimators (AMVU), which are asymptotically normal with asymptotic mean and asymptotic variance $\frac{1}{n}$, are not actually asymptotically efficient. That is, it is possible to construct estimators that do better than an AMVU estimator in terms of the pointwise limiting (rescaled) risk. In particular, these estimators only outperform AMVU estimators at isolated points.

Definition 43: Superefficient

A superefficient estimator is an estimator that attains a asymptotic variance that is smaller than regular efficient estimators (AMVU estimators).

Suppose we have $X_1, ..., X_n$ iid $N(\theta, 1)$ for some unknown θ.

Definition 44: Hodges' Estimator

Suppose $\hat{\theta}_n$ is a consistent estimator for a parameter θ that converges to an asymptotic distribution L_θ, where L_θ is a normal distribution with mean zero and variance depending on θ at the $\sqrt{n}$ rate

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} L_\theta.$$

Then, the Hodges' estimator $\hat{\theta}_n^H$ is defined as

$$\hat{\theta}_n^H = \begin{cases} \hat{\theta}_n & \text{if } |\hat{\theta}_n| \geq n^{-1/4} \\ 0 & \text{if } |\hat{\theta}_n| < n^{-1/4}. \end{cases}$$

Theorem 36.272 *The Hodges' estimator $\hat{\theta}_n^H$ is equal to $\hat{\theta}_n$ everywhere except on the interval $[-n^{1/4}, n^{1/4}]$, where it is equal to 0. The Hodges' estimator is superefficient as it surpasses the asymptotic behavior of the efficient estimator $\hat{\theta}_n$ on at one point $\theta = 0$.*

Lemma 36.273 *The Hodges' estimator $\hat{\theta}_n^H$ is consistent for θ and its asymptotic distribution is*

$$\begin{cases} \sqrt{n}(\hat{\theta}_n^H - \theta) \xrightarrow{d} L_\theta & \theta \neq 0 \\ n^\alpha(\hat{\theta}_n^H - \theta) \xrightarrow{d} 0 & \theta = 0, \forall \alpha \in \mathbb{R}. \end{cases}$$

That is, the estimator has the same asymptotic distribution as $\hat{\theta}_n$ for all $\theta \neq 0$ whereas for $\theta = 0$, the rate of convergence becomes arbitrarily fast.

Theorem 36.274 *Hodges' estimator improves upon a regular estimator at a single point. In general, any superefficient estimator may surpass a regular estimator at most on a set of Lebesgue measure zero.*

We now analyse the performance of Hodges' estimator against the AMVU estimator $\overline{X}$ using the criteria of limiting rescaled risk.

Theorem 36.275 *Let $d_1(\tilde{X}) = \overline{X}$. The risk function is $R(\theta|d_1) = Var_\theta(\overline{X}) = \frac{1}{n}$. Then, the rescaled risk is constant, not depending on n or θ. The limiting rescaled risk of d_1 is*

$$\lim_{n\to\infty} nR(\theta|d_1) = 1.$$

Theorem 36.276 *Let $d_2(\tilde{X})$ be Hodges' estimator $\hat{\theta}_n^H$. The limiting rescaled risk is*

$$\lim_{n\to\infty} nR(\theta|d_2) = \begin{cases} 1 & \theta \neq 0 \\ \\ 0 & \theta = 0. \end{cases}$$

Here, Hodges' estimator seems to perform uniformly better compared to our AMVU estimator, it does just as well at every point of θ but does better at the point $\theta = 0$. However, if we now look at the limiting maximum rescaled risk over an interval, we get a different story.

Theorem 36.277 *Let $d_1(\tilde{X}) = \overline{X}$. The limiting maximum rescaled risk is*

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} nR(\theta|d_1) = 1$$

as $\max_{a\leq\theta\leq b} nR(\theta|d_1)$ does not depend on n.

Theorem 36.278 *Let $d_2(\tilde{X})$ be Hodges' estimator $\hat{\theta}_n^H$. The limiting maximum rescaled risk is*

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} nR(\theta|d_2) = \begin{cases} \infty & a \leq 0 \leq b \\ \\ 1 & \textit{otherwise.} \end{cases}$$

Hence, for points very near to but not exactly zero, Hodges' estimator performs poorly.

STAT3923: Advanced Statistical Inference

37. Interchanging limit and maximum

37.9.3 Interchanging limit and maximum

The takeaway from this section is that limiting rescaled risks may not always be the best criteria to use to evaluate estimators. The **limiting maximum scaled risk** may be better in some circumstances. Furthemore, it is not always possible to swap the order of operation of taking limits and taking maximum.

We recall some facts from analysis.

Definition 37.279 *(Supremum norm). Let $f : D \to \mathbb{K}^N$ be a function. We define its supremum norm by*

$$||f||_{\infty,D} = \sup_{x \in D} ||f(x)||.$$

Corollary 37.280 *f is a bounded function if and only if $||f||_{\infty,D} < \infty$.*

Lemma 37.281 *Let $f : D \to \mathbb{K}^N$ be a continuous function. If D is a compact subset of $\mathbb{K}^d$, then f is bounded. That is, $||f||_{\infty,D} < \infty$.*

Definition 37.282 *(Uniform convergence). We say that $f_n \to f$ uniformly on D if for every $\epsilon > 0$, there exists a $n_\epsilon \in \mathbb{N}$ such that*

$$||f_n(x) - f(x)|| < \epsilon$$

for all $n > n_\epsilon$ and all $x \in D$. We say that $f_n(x) \to f(x)$ uniformly with respect to $x \in D$.

Proposition 11: Uniform Convergence of Functions

Let $f_n : D \to \mathbb{K}^N$ be a sequence of functions. Then, we have that $f_n \to f$ uniformly on the domain D if and only if

$$||f_n - f||_{\infty,D} \to 0$$

as $n \to \infty$.

37.9.4 Poisson Interval Estimation

We describe the setup for the poisson interval. Suppose $\tilde{x} = (x_1, ..., x_n)$ consists of iid Poisson(θ) r.v's with

$$P_\theta(X_1 = x) = \frac{e^{-\theta}\theta^x}{x!}$$

for $x = 0, 1, ...$ and some unknown parameter $\theta \in \Theta = (0, \infty)$.
Consider the decision problem of predicting interval estimate of θ of width $\frac{2C}{\sqrt{n}}$ where the decision space $D = \Theta = (0, \infty)$ and the loss function is $1\{|d - \theta| > \frac{C}{\sqrt{n}}\}$ for some known $C > 0$.
We define the decision $d(\tilde{x}) = \overline{X}$.

Proposition 37.283 *The risk for the Poisson interval estimate is*

$$R(\theta|\overline{X}) = P_\theta(\theta < \overline{X} - \frac{C}{\sqrt{n}}) + P_\theta(\theta > \overline{X} - \frac{C}{\sqrt{n}}).$$

If we let $T = n\overline{X}$ and recall that T is asymptotically normal, that is

$$Z_n = \frac{T - n\theta}{\sqrt{n\theta}} \xrightarrow{d} \mathcal{N}(0,1).$$

Lemma 37.284 *For any $z_n \to z$, we have that*

$$P_\theta(Z_n \leq z_n) \to \Phi(z)$$

where $\Phi(.)$ is the $\mathcal{N}(0,1)$ CDF.

Theorem 37.285 *For a fixed θ, we have that the limiting risk for the Poisson interval estimation is*

$$\lim_{n\to\infty} R(\theta|\overline{X}) = 2[1 - \Phi(\frac{C}{\sqrt{\theta}})].$$

We are interested in computing $\lim_{n\to\infty} \max_{a\leq\theta\leq b} R(\theta|\overline{X})$. We nominate that we could possibly compute this by interchanging the operations and analysing $\max_{a\leq\theta\leq b} \lim_{n\to\infty} R(\theta|\overline{X})$.

Lemma 37.286 *We have that*

$$\max_{a\leq\theta\leq b} \lim_{n\to\infty} R(\theta|\overline{X}) = 2[1 - \Phi(\frac{C}{\sqrt{b}})]$$

where b is the maximum of the interval we are maximising θ over.

Lemma 37.287 *We can upper bound the limiting maximum risk by the following*

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} R(\theta|\overline{X}) \leq 2[1 - \Phi(\frac{C}{\sqrt{b}})]$$

where b is the end point of the interval to maximise θ.

Corollary 37.288 *If we can show that $2[1 - \Phi(\frac{C}{\sqrt{b}})]$ is also a lower bound, then we have that*

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} R(\theta|\overline{X}) = 2[1 - \Phi(\frac{C}{\sqrt{b}})] = \max_{a\leq\theta\leq b} \lim_{n\to\infty} R(\theta|\overline{X}).$$

That is, we can interchange the operation of taking limits and taking maximum.

STAT3923: Advanced Statistical Inference

38. Asymptotic Minimax Lower Bound

38.9.5 Asymptotic Minimax Lower Bound

We now use the pointwise limiting rescaled risk of certain Bayes procedures to provide a lower bound to the limiting maximum rescaled risk of **any estimator**. That is, the bound applies to any estimator whereas previously, the Cramer Rao Lower bound only applied to unbiased and regular models.

Theorem 33: Asymptotic Minimax Lower Bound theorem

Suppose that for a sequence $\{L_n(.|\theta)\}$ of loss functions and any $\theta_0 < \theta_1$, the corresponding sequence of Bayes procedures $\{d_n(.)\}$ based on the uniform prior $U[\theta_0, \theta_1]$ over the interval $[\theta_0, \theta_1]$ is such that for each $\theta_0 < \theta < \theta_1$, we have that the limiting risk

$$\lim_{n\to\infty} R_n(\theta|d_n) = \lim_{n\to\infty} E_\theta\Big[L_n(d_n(x)|\theta)\Big] = S(\theta)$$

for some continuous function $S(.)$.
Then for **any other sequence of procedures** $\{\tilde{d}(.)\}$ and any $a < b$,

$$\max_{a\le\theta\le b} S(\theta) = \max_{a\le\theta\le b} \lim_{n\to\infty} R_n(\theta|d_n) \le \lim_{n\to\infty} max_{a\le\theta\le b} E_\theta\Big[L_n(\tilde{d}_n(x)|\theta)\Big].$$

Remark 38.289 *$L_n(.)$ and hence R_n absorbs the rescaling n term. Furthermore, by analysing the pointwise limiting risk of certain Bayes procedures, this gives us a lower bound to the limiting maximum risk of any procedure which is quite remarkable.*

Remark 38.290 *For any fixed a and b, it is needed that $[\theta_0, \theta_1] \subseteq [a, b]$. However, the asymptotically minimax property is to hold for any $a < b$ in the parameter space, therefore we also require that the Bayes procedure based on the $U[\theta_0, \theta_1]$ prior to have the desired property for any $\theta_0 < \theta_1$ in the parameter space.*

Remark 38.291 *There is nothing special to a uniform prior being used. Any other prior with bounded support would also work. However, it is easier to work with a uniform prior.*

The important takeaway from the above theorem is that if we have a Bayes procedures based on a uniform prior, then the maximum limiting risk is the lower bound for the limiting maximum risk of any other procedure.

STAT3923: Advanced Statistical Inference

39. Proof of Asymptotic Minimax Lower Bound theorem

39.9.6 Proof of Asymptotic Minimax Lower Bound theorem

We are interested in proving the asymptotic minimax lower bound theorem. Let us recall is first.

Theorem 39.292 *Suppose that for a statistical decision problem based on a sequence $\{L_n(d|\theta)\}$ of loss functions, for any $\theta_0 < \theta_1$, the sequence $\{\tilde{d}(.)\}$ of Bayes procedures based on the uniform $U[\theta_0, \theta_1]$ weight function $w(\theta) = \frac{1\{\theta_0 \leq \theta \leq \theta_1\}}{\theta_1 - \theta_0}$ satisfies, for all $\theta_0 < \theta < \theta_1$,*

$$\lim_{n\to\infty} E_\theta\Big[L_n(\tilde{d}_n(X)|\theta)\Big] = S(\theta)$$

for a continuous function S(.). Then, for any other sequence of procedures $\{d_n(.)\}$ and any $a < b$, we have that

$$\max_{a\leq\theta\leq b} S(\theta) \leq \lim_{n\to\infty} \max_{a\leq\theta\leq b} E_\theta\Big[L_n(d_n(X)|\theta)\Big].$$

Lemma 39.293 *For a monotone function $m(.)$, the limit*

$$\lim_{x\to\infty} m(x)$$

always exists.

Lemma 39.294 *For an arbitrary function $f(.)$, the new functions*

$$\overline{f(x)} = \sup_{y\geq x} f(y)$$

and

$$\underline{f(x)} = \inf_{y\geq x} f(y)$$

are monotone.

Corollary 39.295 *As a result, we have that*

$$\lim_{x\to\infty} \overline{f(x)} = \lim_{x\to\infty} \sup_{y\geq x} f(y) = \lim_{x\to\infty} \sup f(x)$$

and

$$\lim_{x\to\infty} \underline{f(x)} = \lim_{x\to\infty} \inf_{y\geq x} f(y) = \lim_{x\to\infty} \inf f(x)$$

always exists.

Lemma 39.296 *We have that $\lim_{x\to\infty} \inf f(x) = \lim_{x\to\infty} \sup f(x)$ if and only if $\lim_{x\to\infty} f(x)$ exists.*

Theorem 39.297 *(Fatou's Lemma). For a sequence of non-negative functions $\{f_n(.)\}$, we have that*

$$\lim_{n\to\infty} inf \int f_n(x)dx \geq \int \liminf f_n(x)dx.$$

Lemma 39.298 *For any sequence of procedure $\{d_n(.)\}$ and any $a < \theta_0 < \theta_1 < b$, we have that*

$$\lim_{n\to\infty} \inf \max_{a\le\theta\le b} E_\theta\Big[L_n(d_n(X)|\theta)\Big] \ge \frac{1}{\theta_1-\theta_0}\int_{\theta_0}^{\theta_1} S(\theta)d\theta$$

where the weight function is the uniform density over the interval $[\theta_0, \theta_1]$.

Lemma 39.299 *Under the conditions, we have that for any $a < b$*

$$\lim_{n\to\infty} \inf \max_{a\le\theta\le b} E_\theta\Big[L_n(d_n(X)|\theta)\Big] \ge \max_{a\le\theta\le b} S(\theta).$$

STAT3923: Advanced Statistical Inference

40. Interval Estimation of a normal mean parameter

40.10 Examples of Asymptotically Minimax Procedures

40.10.1 Introduction to interval estimation

Suppose our loss function is the interval non-coverage loss. The interval is the level set of the posterior of the desired width $2C_n$, that is, the set of θ values where the posterior density $p(\theta|X)$ is above a certain level ℓ

$$\{\theta : p(\theta|X) \geq \ell\}.$$

Choosing different levels of ℓ gives intervals of different widths. The trick is then to choose the level which gives an interval of width $2C_n$. We have two scenarios.

1. If the posterior density is **unimodal**, first increasing then decreasing, then it is $d \pm C_n$ where we solve for d in the equation
$$p(d - C_n|X) = p(d + C_n|X)$$
2. If the posterior density is **strictly decreasing** over some range [a, b), then so long as $a + 2C_n < b$, the level set is then simply
$$[a, a + 2C_n].$$

Hence, we are looking the interval of width $2C_n$ with the highest probability under the posterior distribution.

40.10.2 Examples with non-coverage loss: Poisson Interval Estimation Revisited

Recalled we have shown that for the procedure $\overline{X}$,

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} R(\theta|\overline{X}) \leq \max_{a\leq\theta\leq b} \lim_{n\to\infty} R(\theta|\overline{X}) = 2[1 - \Phi(\frac{C}{\sqrt{b}})]$$

and that we now wanted to show that the maximum of the limiting risk is also a lower bound. We can use the theorem of the asymptotic minimax lower bound theorem to help us show this.

First, recall that

$$\lim_{n\to\infty} R(\theta|\tilde{d}_n) = 2[1 - \Phi(\frac{C}{\sqrt{\theta}})] = S(\theta).$$

We use the uniform prior $w(\theta) = \frac{1\{\theta_0\leq\theta\leq\theta_1\}}{\theta_1-\theta_0}$. We have that the product of the prior and likelihood gives us

$$w(\theta)f_\theta(\tilde{x}) = Const\frac{\theta^{(T+1)}e^{-\eta\theta}1\{\theta_0 \leq \theta \leq \theta_1\}}{\int_{\theta_0}^{\theta_1} \theta^{(T+1)-1}e^{-n\theta}d\theta}$$

where $T = \sum_{i=1}^{n} X_i$. The posterior density is a **truncated gamma density**. That is, it is the Gamma distribution with shape T+1 and rate n where we restrict it to the interval $[\theta_0, \theta_1]$. As the loss function is the coverage loss, the Bayes procedure is the level set of the posterior of width $\frac{2C}{\sqrt{n}}$. Hence, $\tilde{d}_n(\tilde{x}) = \overline{X}$.

We can now use the asymptotic minimax lower bound theorem.

Theorem 40.300 *For any other procedure* $\{d_n\}$ *and any* $a < b$

$$\lim_{n\to\infty} \max_{a\le\theta\le b} R(\theta|d_n(.)) \ge \max_{a\le\theta\le b} S(\theta)$$

$$= \max_{a\le\theta\le b} 2[1 - \Phi(\frac{C}{\sqrt{\theta}})]$$

$$= 2[1 - \Phi(\frac{C}{\sqrt{b}})].$$

Hence, using the above, let $\{d_n\} = \overline{X}$ and we get that

$$2[1 - \Phi(\frac{C}{\sqrt{b}})] \le \lim_{n\to\infty} \max_{a\le\theta\le b} R(\theta|\overline{X}).$$

As a result, we finally get the result that

$$\lim_{n\to\infty} \max_{a\le\theta\le b} R(\theta|\overline{X}) = 2[1 - \Phi(\frac{C}{\sqrt{b}})] = \max_{a\le\theta\le b} \lim_{n\to\infty} R(\theta|\overline{X})$$

that is, we can interchange the operation of taking limits and maximum. Therefore, the procedure $\overline{X} \pm \frac{C}{\sqrt{n}}$ is asymptotically minimax over any interval for Poisson interval estimation.

STAT3923: Advanced Statistical Inference

41. Interval Estimation of a normal mean parameter

41.10.3 Examples with non-coverage loss: Interval Estimation of a normal mean parameter

The limiting risk of Bayes procedures $\tilde{d}(X)$ using a uniform prior can be derived by first deriving the limiting risk of $d_{flat}(X)$, the Bayes procedure using the flat prior $w(\theta) = 1$ and then showing that with probability tending to 1, $\tilde{d}(X) = d_{flat}(X)$.

Suppose $\tilde{X}$ consists of iid $N(\theta, 1)$ r.v's for some unknown $\theta \in \Theta = \mathbb{R}$. Consider the decision problem with $D = \Theta = \mathbb{R}$ and non-coverage loss function $L(d|\theta) = 1\{|d - \theta| > \frac{C}{\sqrt{n}}\}$ for some given $C > 0$.

We first derive the limiting risk of the Bayes estimator $\tilde{d}(\tilde{X})$ based on a uniform prior $U[\theta_0, \theta_1]$.

Lemma 41.301 *The posterior density of the Bayes estimator $\tilde{d}$ using a uniform prior $U[\theta_0, \theta_1]$ is*

$$w(\theta) f_\theta(\tilde{X}) = Const. \frac{1\{\theta_0 \leq \theta \leq \theta_1\} e^{-\frac{n}{2}(\theta - \overline{X})^2}}{\int_{\theta_0}^{\theta_1} e^{-\frac{n}{2}(\theta - \overline{X})^2}}.$$

In particular, the posterior density is a truncated normal density where we restrict the normal $N(\overline{X}, \frac{1}{n})$ density to the interval $[\theta_0, \theta_1]$.

Lemma 41.302 *The Bayes estimator $\tilde{d}(\tilde{X})$ is the midpoint of the level set of the truncated normal density of width $\frac{2C}{\sqrt{n}}$. That is, $\tilde{d}(\tilde{X}) = \overline{X}$ for when*

$$\theta_0 + \frac{C}{\sqrt{n}} < \overline{X} < \frac{C}{\sqrt{n}} - \theta_1.$$

Definition 41.303 *(Event of non-coverage). Let us denote B_n to be the event of non-coverage of the interval*

$$B_n = \{\theta < \overline{X} - \frac{C}{\sqrt{n}}\} \cup \{\theta > \overline{X} + \frac{C}{\sqrt{n}}\}.$$

Theorem 41.304 *We have that the limiting risk of the Bayes estimator $\tilde{d}$ using a uniform prior $U[\theta_0, \theta_1]$ is*

$$\lim_{n \to \infty} R(\theta|\tilde{d}) = \lim_{n \to \infty} P_\theta(B_n) = 2\Big[1 - \Phi(C)\Big] = S(\theta).$$

Corollary 41.305 *For any procedure $\{d_n(.)\}$, and any $a < b$, we have that*

$$\lim_{n \to \infty} \max_{a \leq \theta \leq b} R(\theta|d_n) \geq \max_{a \leq \theta \leq b} S(\theta) = 2\Big[1 - \Phi(c)\Big].$$

Now, to show that $\overline{X} = \hat{\theta}_n$ is asymptotically minimax, we need to show that

$$\lim_{n \to \infty} \max_{a \leq \theta \leq b} R(\theta|\hat{\theta}_n) \leq \max_{a \leq \theta \leq b} S(\theta) = 2\Big[1 - \Phi(c)\Big].$$

However, we have just shown that $\overline{X}$ has a limiting maximum rescaled risk that is exactly equal to the lower bound. Hence, $\overline{X}$ is asymptotically minimax.

We now want to show that a Bayes estimator with a conjugate normal prior $w(\theta) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\theta-\mu_0)^2}$ ($\mathcal{N}(\mu_0, \sigma_0^2)$ density) is asymptotically minimax.

Lemma 41.306 *The posterior distribution of the Bayes estimator with conjugate normal prior $\mathcal{N}(\mu_0, \sigma_0^2)$ is also a normal distribution*

$$\mathcal{N}\Big((\frac{1}{1+n\sigma_0^2})\mu_0 + (\frac{n\sigma_0^2}{1+n\sigma_0^2})\overline{X}, \frac{\sigma_0^2}{1+n\sigma_0^2}\Big).$$

Lemma 41.307 *Under the interval coverage loss, the Bayes estimator is the center of symmetry of the posterior distribution, hence*

$$\hat{\theta}_{conj} \pm \frac{C}{\sqrt{n}}$$

where $\hat{\theta}_{conj} = (\frac{1}{1+n\sigma_0^2})\mu_0 + (\frac{n\sigma_0^2}{1+n\sigma_0^2})\overline{X}$.

Lemma 41.308 *The exact risk for the Bayes estimator $\hat{\theta}_{conj}$ is*

$$R(\theta|\hat{\theta}_{conj}) = P_\theta\Big(\theta < \hat{\theta}_{conj} - \frac{C}{\sqrt{n}}\Big) + P_\theta\Big(\theta > \hat{\theta}_{conj} - \frac{C}{\sqrt{n}}\Big)$$

We analyse the probabilities of our interval being too high or too low in separate cases.

Lemma 41.309 *We have that the probability of our interval overestimating to be*

$$P_\theta\Big(\theta < \hat{\theta}_{conj} - \frac{C}{\sqrt{n}}\Big) = 1 - \Phi\Big(C(1+\frac{1}{n\sigma_0^2}) + \frac{\theta-\mu_0}{\sigma_0^2\sqrt{n}}\Big).$$

Lemma 41.310 *We have that the probability of our interval underestimating to be*

$$P_\theta\Big(\theta < \hat{\theta}_{conj} - \frac{C}{\sqrt{n}}\Big) = \Phi\Big(-C(1+\frac{1}{n\sigma_0^2}) + \frac{\theta-\mu_0}{\sigma_0^2\sqrt{n}}\Big).$$

Lemma 41.311 *Hence, for any $a < b$, we have that*

$$\max_{a\le\theta\le b} R(\theta|\hat{\theta}_{conj}) \le 1 - \Phi\Big(C(1+\frac{1}{n\sigma_0^2}) + \frac{\theta-\mu_0}{\sigma_0^2\sqrt{n}}\Big) + \Phi\Big(-C(1+\frac{1}{n\sigma_0^2}) + \frac{\theta-\mu_0}{\sigma_0^2\sqrt{n}}\Big).$$

Corollary 41.312 *We have that the limiting maximum risk for $\hat{\theta}_{conj}$ is*

$$\lim_{n\to\infty} \max_{a\le\theta\le b} R(\theta|\hat{\theta}_{conj}) \le 2\Big[1 - \Phi(C)\Big].$$

Hence, $\hat{\theta}_{conj}$ is asymptotically minimax.

STAT3923: Advanced Statistical Inference

42. Interval Estimation of a uniform scale parameter

42.10.4 Examples with non-coverage loss: Interval Estimation of a uniform scale parameter

Suppose $\tilde{X} = (X_1, ..., X_n)$ consists of iid $U[0, \theta]$ random variables for some unknown $\theta \in \Theta = (0, \infty)$. The maximum likelihood estimator is $X_{(n)}$, the sample maximum.

Remark 42.313 *In models that do not satisfy regularity conditions, the bias matters now when analysing the limiting MSE.*

Lemma 42.314 *The posterior distribution of the Bayes procedure d_{flat} using the flat prior $w(\theta) = 1$ is the Pareto distribution with shape n-1 and scale $X_{(n)}$*

$$p(\theta|\tilde{X}) = \frac{(n-1)X_{(n)}^{n-1}}{\theta^n} 1\{\theta \geq X_{(n)}\}.$$

Lemma 42.315 *Under non-coverage loss, the interval estimate with d_{flat} is the level set of width $\frac{2C}{n}$, which in this case is*

$$\Big[X_{(n)}, X_{(n)} + \frac{2C}{n}\Big].$$

Lemma 42.316 *The risk function for d_{flat} is*

$$\begin{aligned} R(\theta|d_{flat}) &= P_\theta(\theta < X_{(n)}) + P_\theta(\theta > X_{(n)} + \frac{2C}{n}) \\ &= \Big(1 - \frac{2C}{\theta n}\Big)^n. \end{aligned}$$

Corollary 42.317 *Hence, the maximum of d_{flat} over [a,b] is*

$$\max_{a \leq \theta \leq b} R(\theta|d_{flat}) = \Big(1 - \frac{2C}{bn}\Big)^n.$$

Lemma 42.318 *The limiting risk of d_{flat} is*

$$\lim_{n\to\infty} R(\theta|d_{flat}) = \lim_{n\to\infty} \Big(1 - \frac{2C}{bn}\Big)^n = e^{-\frac{2C}{\theta}}.$$

Lemma 42.319 *The limiting maximum risk of d_{flat} is*

$$\lim_{n\to\infty} \max_{a \leq \theta \leq b} R(\theta|d_{flat}) = e^{-\frac{2C}{b}}.$$

We now show that d_{flat} is asymptotically minimax. We derive the limiting risk of $\tilde{d}$ using the $U[\theta_0, \theta_1]$ prior.

Lemma 42.320 *The posterior distribution of $\tilde{d}$ using the $U[\theta_0, \theta_1]$ prior is*

$$const. \frac{1}{\theta^n} \frac{1\{max(X_{(n)}, \theta_0) \leq \theta \leq \theta_1\}}{\int_{max(X_{(n)}, \theta_0)}^{\theta_1} \frac{1}{\theta^n} d\theta}$$

which is the truncated version of the Pareto distribution we get using the flat prior.

Lemma 42.321 *As long as $\theta_0 \leq X_{(n)}$ and $X_{(n)} + \frac{2C}{n} \leq \theta$, the level is*

$$\Big[X_{(n)}, X_{(n)} + \frac{2C}{n}\Big].$$

Lemma 42.322 *We can show that*

$$P_\theta\Big[\theta_0 \leq X_{(n)} \leq \theta_1 - \frac{2C}{n}\Big] \to 1$$

for all $\theta_0 < \theta < \theta_1$.

Lemma 42.323 *Let us denote $\tilde{d}_n(.)$ to be the Bayes procedure which uses the uniform prior $U[\theta_0, \theta_1]$. Then, we have that*

$$P_\theta\Big[d_{flat}(\tilde{X}) = \tilde{d}_n(\tilde{X})\Big] \to 1$$

for all $\theta_0 < \theta < \theta_1$.

Lemma 42.324 *The limiting risk of the Bayes procedure with uniform prior, via the flat prior, is*

$$\lim_{n\to\infty} R(\theta|\tilde{d}_n) = \lim_{n\to\infty} R(\theta|d_{flat}) = e^{-\frac{2C}{\theta}} = S(\theta).$$

Hence, for any sequence of procedures $\{d_n(.)\}$,

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} R(\theta|d_n) \geq \max_{a\leq\theta\leq b} S(\theta) = e^{-\frac{2C}{b}}.$$

Therefore, d_{flat} is asymptotically minimax.

STAT3923: Advanced Statistical Inference

43. Estimating binomial proportion with known sample size

43.10.5 Examples with squared error loss: Estimating binomial proportion with known sample size

In the case of squared error loss, it turns out that in many cases, the Bayes procedure (i.e. the posterior mean) using a uniform prior has the same limiting rescaled risk as the Bayes procedure using a flat prior.

Suppose $\tilde{X} = (X_1, ..., X_n)$ consists of iid binomial $(1, \theta)$ random variables for $\theta \in \Theta = (0, 1)$. Consider the decision problem with decision space $D = \Theta = (0, 1)$ and the loss $L(d|\theta) = (d - \theta)^2$. We want to show that the Bayes procedure with the conjugate prior $w(\theta) = \frac{\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}}{beta(\alpha_0, \beta_0)}$ (beta(α_0, β_0) density) is asymptotically minimax. We assume that for any $\theta \leq \theta_0 < \theta_1 \leq 1$, the Bayes procedure using the $U[\theta_0, \theta_1]$ prior, $\tilde{d}(\tilde{X})$ is such that

$$\lim_{n\to\infty} nR(\theta|\tilde{d}) = \lim_{n\to\infty} nR(\theta|d_{flat})$$

where d_{flat} is the Bayes procedure using the flat prior $w(\theta) = 1$.

First, we find the lower bound to the limiting maximum risk. We look at the limiting rescaled risk of $\tilde{d}$ but this may be difficult, hence, we use the assumption given to us and look at the limiting rescaled risk of d_{flat} which uses a flat prior.

Lemma 43.325 *The posterior distribution for d_{flat} using a flat prior $w(\theta) = 1$ is*

$$p(\theta|\tilde{X}) = \frac{\theta^{(T+1)-1}(1-\theta)^{(n-T+1)-1}}{beta(T+1, n-T+1)}$$

which is the $beta(T + 1, n - T + 1)$ density.

Corollary 43.326 *Under a squared-error loss, the Bayes procedure for d_{flat} is the posterior mean of the $beta(T + 1, n - T + 1)$ density, which is*

$$d_{flat}(\tilde{X}) = \frac{T+1}{n+2}$$

which may be interpreted as the sample proportion we would obtain if we added 1 success and 1 failure to the data.

Lemma 43.327 *The risk under a squared error loss is the mean squared error, which can be decomposed into the variance and bias squared. Hence, the risk of d_{flat} is*

$$R(\theta|d_{flat}) = Var_\theta\Big[d_{flat}\Big] + Bias_\theta\Big[d_{flat}\Big]^2 = \frac{n\theta(1-\theta)}{(n+2)^2} + \Big(\frac{1-2\theta}{n+2}\Big)^2$$

$$= \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2}.$$

Lemma 43.328 *We have that the limiting rescaled risk for d_{flat} is*

$$\lim_{n\to\infty} nR(\theta|d_{flat}) = \lim_{n\to\infty}]bigg[\frac{n\theta(1-\theta)+(1-2\theta)^2}{(n+2)^2}\Big]$$

$$\to \theta(1-\theta) = S(\theta).$$

Hence, using our assumption, we have that

$$\lim_{n\to\infty} nR(\theta|\tilde{d}) = \lim_{n\to\infty} nR(\theta|d_{flat}) = \theta(1-\theta).$$

Corollary 43.329 *For any other procedure $\{d_n(.)\}$, we have that*

$$\lim_{n\to\infty} \max_{a\le\theta\le b} R(\theta|d_n) \ge \max_{a\le\theta\le b} S(\theta) = \max_{a\le\theta\le b} \theta(1-\theta).$$

We now look for an upper bound.

Lemma 43.330 *Using the beta(α_0,β_0) density as a weight function, we have that the posterior density is*

$$f_\theta(\tilde{X})w(\theta) = Const.\frac{\theta^{T+\alpha_0-1}(1-\theta)^{n-T+\beta_0-1}}{beta(T+\alpha_0, n-T+\beta_0)}$$

where the posterior density is the beta$(T+\alpha_0, n-T+\beta_0)$ density.

Lemma 43.331 *The Bayes procedure under the squared error loss is the posterior mean, which is,*

$$d_{conj}(\tilde{X}) = \frac{T+\alpha_0}{n+\alpha_0+\beta_0}.$$

Lemma 43.332 *The risk under the squared error loss is the MSE, hence, the risk of d_{conj} is*

$$R(\theta|d_{conj}) = Var_\theta\Big[d_{conj}(X)\Big] + Bias_\theta\Big[d_{conj}\Big]^2$$

$$= \frac{n\theta(1-\theta)+[(1-\theta)\alpha_0-\theta\beta_0]^2}{(n+\alpha_0+\beta_0)^2}.$$

Lemma 43.333 *The limiting maximum rescaled risk of d_{conj} can be bounded by*

$$\lim_{n\to\infty} \max_{a\le\theta\le b} R(\theta|d_{conj}) \le \max_{a\le\theta\le b} \theta(1-\theta) = \max_{a\le\theta\le b} S(\theta).$$

Corollary 43.334 *The bias is asymptotically negligible compared to the variance in the limit.*

Theorem 34: Exponential family variance vs risk

Exponential families have the property that the variance dominates the risk in the limit.

Hence, we have that $d_{conj}(\tilde{X})$ is asymptotically minimax.

43.10.6 Showing convergence of interval coverage

Theorem 43.335 *Suppose that $X_1, ..., X_n$ are iid $N(\theta, 1)$ are random variables and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. If $\theta_0 < \theta < \theta_1$ and $0 < C < \infty$, then*

$$P_\theta\Big\{\theta_0 + \frac{C}{\sqrt{n}} < \overline{X} < \theta_1 - \frac{C}{\sqrt{n}}\Big\} \to \infty$$

as $n \to \infty$.

Remark 43.336 *Hence, the probability that the estimator $\overline{X}$ lies in the interval is 1 as the sample size gets large.*

STAT3923: Advanced Statistical Inference

44. Estimating normal variance with known mean

44.10.7 Examples with squared error loss: Estimating normal variance with known mean

Suppose $X = (X_1, ..., X_n)$ consists of iid $N(0, \theta)$ R.Vs for some unknown $\theta \in \Theta = (0, \infty)$. We let the decision space $D = \Theta = (0, \infty)$ and loss $L(d|\theta) = (d - \theta)^2$. We want to show that **both** the maximum-likelihood estimator and the Bayes procedure with the inverse Gamma conjugate prior $w(\theta) = \frac{\lambda_0^{\alpha_0} e^{-\lambda/\theta}}{\theta^{\alpha_0+1}\Gamma(\alpha_0)}$ for known $\alpha_0, \lambda_0 > 0$ are asymptotically minimax.

We are allowed to assume that for any $0 < \theta_0 < \theta_1 < \infty$, the Bayes procedure $\tilde{d}(.)$ using the $U[\theta_0, \theta_1]$ prior, has, for all $\theta_0 < \theta < \theta_1$,

$$\lim_{n\to\infty} nR(\theta|\tilde{d}) = \lim_{n\to\infty} nR(\theta|d_{flat})$$

where d_{flat} is the Bayes procedure using the flat prior $w(\theta) = 1$.

We show two steps. First, we find the lower bound to any limiting maximum rescaled risk for **any** estimator. Then, we find the upper bounds to the limiting maximum risks for our candidate estimators. If these bounds coincide, then we have shown that our candidate estimators are asymptotically minimax.

Instead of trying to derive the limiting maximum rescaled risk with our Bayes procedure with inverse Gamma conjugate prior $\tilde{d}$, we can use the assumption given to us that it is equivalent to the limiting maximum rescaled risk of d_{flat} which uses a flat prior.

Lemma 44.337 *The posterior density of d_{flat} is*

$$w(\theta) f_\theta(\tilde{X}) = \frac{1}{(2\pi\theta)^{n/2}} e^{-\frac{1}{2\theta}T}$$

where $T = \sum_{i=1}^n X_i^2$ is the sufficient statistic.

Corollary 44.338 *The posterior density of d_{flat} is the Inverse Gamma($\frac{n}{2} - 1, \frac{T}{2}$) as we had that the conjugate prior of $\tilde{d}$ was an Inverse gamma.*

Lemma 44.339 *The Bayes procedure under squared error loss is the posterior mean, hence*

$$d_{flat}(\tilde{X}) = \frac{T/2}{(n/2 - 1) - 1} = \frac{T}{n - 4}.$$

Lemma 44.340 *Under the squared error loss, the risk of d_{flat} is the MSE of d_{flat}. Therefore,*

$$R(\theta|d_{flat}) = Var(d_{flat}) + Bias(d_{flat})^2 = \frac{2n\theta^2}{(n-4)^2} + \frac{16\theta^2}{(n-4)^2}.$$

Lemma 44.341 *The limiting maximum rescaled risk is of d_{flat} is*

$$\lim_{n\to\infty} nR(\theta|d_{flat}) = \lim_{n\to\infty}\{nVar_\theta(d_{flat})\} + \lim_{n\to\infty}\{nBias_\theta(d_{flat})^2\}$$

which is equal to

$$= 2\theta^2\Big[\lim_{n\to\infty}\Big(\frac{n}{n-4}\Big)^2 + \lim_{n\to\infty}\frac{8n}{(n-4)^2}\Big]$$
$$\xrightarrow{n\to\infty} 2\theta^2.$$

Corollary 44.342 *Under the squared error loss, the contribution of the bias to the limiting rescaled MSE/risk is negligible compared to the variance.*

Hence, using our assumption, we have that

$$\lim_{n\to\infty} nR(\theta|\tilde{d}) = \lim_{n\to\infty} nR(\theta|d_{flat}) = 2\theta^2 = S(\theta).$$

So, for any other sequence of estimators $\{d_n(.)\}$, we have, for any $0 < a < b < \infty$

$$\lim_{n\to\infty}\max_{a\le\theta\le b} nR(\theta|d_n) \ge max_{a\le\theta\le b}S(\theta) = 2b^2.$$

Lemma 44.343 *The MLE is*

$$\hat{\theta}_{ML} = \frac{T}{n}$$

Lemma 44.344 *The limiting maximum rescaled risk of the MLE is*

$$\lim_{n\to\infty}\max_{a\le\theta\le b} nR(\theta|\hat{\theta}_{ML}) = 2b^2.$$

This coincides with the lower bound and hence $\hat{\theta}_{ML}$ is asymptotically minimax.

Lemma 44.345 *For the Bayes procedure using the inverse gamma conjugate prior, the Bayes procedure is the posterior mean of the posterior inverse gamma distribution*

$$d_{conj}(\tilde{X}) = \frac{T + 2\lambda_0}{n + 2\alpha_0 - 2}.$$

Lemma 44.346 *The rescaled risk of d_{conj} is the rescaled MSE*

$$nR(\theta|d_{conj}) = 2\theta^2\Big(\frac{n}{n+2\alpha_0-2}\Big)^2 + \frac{n\Big(2\lambda_0 - 2\theta\alpha_0 + 2\theta\Big)}{(n+2\alpha_0-2)^2}.$$

Hence, the limiting maximum rescaled risk is

$$\lim_{n\to\infty}\max_{a\le\theta\le b} nR(\theta|d_{conj}) = \max_{a\le\theta\le b} 2\theta^2 = 2b^2.$$

Hence, d_{conj} is also asymptotically minimax.

STAT3923: Advanced Statistical Inference

45. Estimating normal variance with known mean

45.10.8 Examples with absolute error loss: Estimating normal mean with known variance

Suppose $\tilde{X} = (X_1, ..., X_n)$ consists of iid $N(\theta, 1)$ random variables for some unknown $\theta \in \Theta = \mathbb{R}$. Consider the decision problem with the decision space $D = \Theta$ and loss function $L(d|\theta) = |d - \theta|$.

Lemma 45.347 *The sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is **both** the MLE and the Bayes procedure using the flat prior $w(\theta) = 1$, where in the latter case, the posterior is the $\mathcal{N}(\overline{X}, \frac{1}{n})$ density.*

Corollary 45.348 *Under the absolute error loss, the Bayes procedure is the posterior median which is $\overline{X}$.*

Lemma 45.349 *The limiting maximum rescaled risk for $\overline{X}$ is*

$$\lim_{n\to\infty} \max_{a\le\theta\le b} \sqrt{n} R(\theta|\overline{X}) = \sqrt{\frac{2}{\pi}}.$$

Lemma 45.350 *The Bayes procedure $\tilde{d}(\tilde{x})$ is the Bayes procedure based on a $U[\theta_0, \theta_1]$ prior. This estimator has the form*

$$\tilde{d}(\tilde{x}) = \overline{X} + \frac{1}{\sqrt{n}}\Phi^{-1}\Big(\frac{1}{2}\Big[\Phi(\sqrt{n}(\theta_1 - \overline{X})) + \Phi(\sqrt{n}(\theta_0 - \overline{X}))\Big]\Big).$$

Lemma 45.351 *The rescaled risk for the Bayes procedure $\tilde{d}$ is*

$$\sqrt{n} R(\theta|\tilde{d}) \to \sqrt{\frac{2}{\pi}}.$$

Corollary 45.352 *Hence, for any other estimator $d_n(\tilde{X})$, we have that*

$$\lim_{n\to\infty} \max_{a\le\theta\le b} \sqrt{n} R(\theta|d_n) \ge \sqrt{\frac{2}{\pi}}.$$

STAT3923: Advanced Statistical Inference

46. Estimating uniform scale

46.10.9 Examples with absolute error loss: Estimating uniform scale

Suppose $\tilde{X} = (X_1, ..., X_n)$ which consists of iid $U[0, \theta]$ random variables for some unknown $\theta \in \Theta = (0, \infty)$. Consider the decision space $D = \Theta$ and loss $L(d|\theta) = |d - \theta|$.

Lemma 46.353 *The risk of the MLE $\hat{\theta}_{ML} = X_{(n)}$ is*

$$R(\theta|\hat{\theta}_{ML}) = \frac{\theta}{n+1}.$$

Corollary 46.354 *The limiting maximum rescaled risk of the MLE is therefore*

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} nR(\theta|\hat{\theta}_{ML}) = b.$$

We now consider the **median-unbiased** version of the MLE.

Lemma 46.355 *The MLE $X_{(n)}$ has the CDF*

$$F_n(x;\theta) = P_\theta(X_{(n)} \leq x) = \begin{cases} 0 & x < 0 \\ \left(\frac{x}{\theta}\right)^n & 0 \leq x \leq \theta \\ 1 & x > \theta. \end{cases}$$

Corollary 46.356 *The median of this distribution is the solution m to*

$$\frac{1}{2} = \left(\frac{m}{\theta}\right)^n$$

$$m = \left(\frac{1}{2}\right)^{\frac{1}{n}}\theta.$$

Definition 46.357 *(Median-unbiasd MLE). The median unbiased MLE is*

$$d_{med}(\tilde{X}) = 2^{\frac{1}{n}} X_{(n)}.$$

Lemma 46.358 *The risk of the median unbiasd MLE is*

$$R(\theta|d_{med}) = \frac{\theta n}{n+1}\left(2^{\frac{1}{n}} - 1\right).$$

Corollary 46.359 *The limiting maximum risk of the median unbiased MLE is*

$$\lim_{n\to\infty} \max_{a\leq\theta\leq b} nR(\theta|d_{med}) = blog_e 2.$$

Lemma 46.360 *The Bayes procedure using a flat weight function is the median of the posterior distribution of the Pareto(n-1,$X_{(n)}$) density. Hence, the Bayes procedure is*

$$d_{flat}(\tilde{X}) = 2^{\frac{1}{n-1}} X_{(n)}.$$

Remark 46.361 *The Bayes procedure under the flat prior is very similar to the median-unbiased MLE!*

Theorem 46.362 *The Bayes procedure $\tilde{d}(\tilde{X})$ using the $U[\theta_0, \theta_1]$ prior has the same limiting risk as the Bayes procedure with the flat prior and the median-unbiased MLE.*

$$\lim_{n\to\infty} nR(\theta|\tilde{d}) = \lim_{n\to\infty} nR(\theta|d_{flat}) = \lim_{n\to\infty} nR(\theta|d_{med}) = \theta log 2$$

for all $\theta_0 < \theta < \theta_1$.

Corollary 46.363 *Hence, $d_{med}(\tilde{X})$ and $d_{flat}(\tilde{X})$ are asymptotically minimax but $\hat{\theta}_{ML}(\tilde{X}$ is not.*

STAT3923: Advanced Statistical Inference

47. L2 convergence of estimators

47.10.10 L2 convergence of estimators

We have used numerous times that when computing the limiting risk, under squared error loss, of the Bayes procedure with a uniform prior, we can instead compute the limiting risk of the Bayes procedure with a flat prior and this gives us the same result. The advantage to this is that computing the limiting risk of the Bayes procedure with a flat prior is easier. We now seek to prove why we can do this. Intuitively, we can do this because the two procedures are "close" enough and hence have similar limiting risk.

First, we seek to define what it means for two estimators to be close.

Definition 47.364 *(L2 metric). We define the L2 or mean-squared metric as*

$$d(X,Y) = E[(X-Y)]^2$$

where we assume that $X \in L^2$ and $Y \in L^2$.

Definition 47.365 *(Mean-square convergent). Let $\{X_n\}_{n\geq 1}$ be a sequence in L^2 defined on a sample space Ω. We say that $\{X_n\}_{n\geq 1}$ is mean-square convergent (or convergent in mean square) if and only if there exists a random variable $X \in L^2$ such that*

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0.$$

X is called the mean-square limit of the sequence and denote this by

$$X_n \xrightarrow{m.s} X$$

Theorem 47.366 *Mean-square convergence implies convergence in probability.*

Theorem 47.367 *Suppose $X_n \xrightarrow{P} 0$ and $|X_n| \leq M < \infty$ for all $n \geq n_0$ for some positive integer n_0 and constant M. Then,*

$$E(X_n^2) \to 0.$$

We can now define what it means for two Bayes procedures to be close by using the L2 metric.

Theorem 35: L2 Convergence of Estimators

Suppose that for 2 estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ and some rate (sequence) $\{r_n\}$, we have

1. $r_n E_\theta\left[(\hat{\theta}_1 - \theta)^2\right] \to S(\theta) < \infty$
2. $r_n E_\theta\left[(\hat{\theta}_1 - \hat{\theta}_2)^2\right] \to 0.$

Then, we have that

$$r_n E_\theta\left[(\hat{\theta}_2 - \theta)^2\right] \to S(\theta) < \infty$$

Remark 47.368 *That is, if we have one estimator that has the same finite limiting risk, which is $S(\theta)$, and we have that the two estimators are close in mean-squared, then we can say that the other estimator has the same limiting risk as our original estimator.*

Lemma 47.369 *The MLE is not asymptotically minimax for non-regular model as the bias has the same order as the variance.*

The significance of this is that for non-regular models, the Bayes estimators automatically adjust the order of their bias in the presence of irregularity and hence perform better than the MLE for such cases.

STAT3923: Advanced Statistical Inference

48. Convergence of Bayes procedure and sample mean

48.10.11 Convergence of Bayes procedure and sample mean for normal

Suppose $X = (X_1, ..., X_n)$ consists of iid $N(\theta, 1)$ random variables for some unknown $\theta \in \Theta = \mathbb{R}$. We consider the decision space $D = \Theta$ and loss $L(d|\theta) = (d - \theta)^2$. We are interested in showing that the limiting rescaled risk of the sample mean $\overline{X}$ is identical to the limiting rescaled risk of the Bayes procedure with a uniform $U[\theta_0, \theta_1]$ prior.

Lemma 48.370 *(Mills Ratio). Let Φ be the CDF of a standard normal random variable. Then, we have*

$$\frac{1}{\sqrt{2\pi}}[1 - \Phi(\sqrt{2})] < 6.$$

Theorem 48.371 *If $\tilde{d}$ is the Bayes procedure with uniform $[\theta_0, \theta_1]$ prior, then*

$$\lim_{n\to\infty} nR(\theta|\tilde{d}) = \lim_{n\to\infty} nR(\theta|\overline{X}).$$

STAT3923: Advanced Statistical Inference

49. Overview of Asymptotically Minimax Procedures

49.10.12 Overview of Asymptotically Minimax Procedures

We summarise what we have done with regards to asymptotically minimax procedures. First, under squared-error loss, asymptotically minimax procedures generalises the notion of AMVU estimators in the case of non-regular models.

Lemma 49.372 *Under regularity conditions, maximum likelihood estimators are different to Bayes estimators. However, they tend to differ in their bias under squared error loss. However, as the bias is asymptotically negligible, they are both asymptotically minimax.*

Lemma 49.373 *We state 4 reasons for why we would use the limiting maximum risk over intervals as our criteria for statistical optimality.*

1. *It is difficult to establish non-asymptotic results and only available under certain conditions.*
2. *The limiting maximum risk gets around the issues of superefficiency.*
3. *The asymptotic minimax lower bound applies to* ***any procedure****.*
4. *The theory behind asymptotic minimax lower bound highlights intersting properties of Bayes estimators. That is, by just analysing the risk of Bayes procedures, this determines the best possible performance of any procedure in terms of their limiting maximum risk.*

STAT3923: Advanced Statistical Inference

50. Notes on Bayesian Statistics

50.11 Bayesian Statistics

50.11.1 Notes

First, we motivate the use of priors. Before that, we motivate the very use of parameters.

Definition 50.374 *(Infinite Exchangeability). We say that $(x_1, ..., x_n)$ is an infinitely exchangeable sequence of random variables if, for any n, the joint probability $p(x_1, ..., x_n)$ is invariant to permutation of the indices. That is, for any permutation π,*

$$p(x_1, ..., x_n) = p(x_{\pi_1}, ..., x_{\pi_n}).$$

Note that independent and identically distributed is a subset of infinite exchangeability. The following theorem indicates why infinite exchangeability is important.

Theorem 50.375 *(De Finetti Theorem). A sequence of random variables $(x_1, x_2, ...)$ is infinitely exchangeable if and only if for all n,*

$$p(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta)p(\theta)d\theta$$

for some distribution of θ given by $p(\theta)d\theta$.

The forward direction of the above theorem is what is so powerful. It says that

1. We have exchangeable data;
2. There must exist a parameter θ;
3. There must exist a likelihood $p(x|\theta)$;
4. There must exist a distribution P on θ;
5. The above quantities **must** exist so as to render the data $(x_1, ..., x_n)$ conditionally independent.

Hence, this is why we should use parameters and why we should priors on parameters.

Proposition 50.376 *(3 principles of Bayesian approach). We state the 3 principles behind Bayesian modelling.*

1. *Conditionality principle. If an experiment concerning inference about θ is chosen from a collection of possible experiments independently, then any experiment not chosen is irrelevant to the inference.*
2. *Likelihood principle. The relevant information in any inference about θ after x is observed is contained **entirely** in the likelihood function. That is, the likelihood function is $p(x|\theta)$ for a fixed x is a function of θ.*

3. *Sufficiency principle. If two different observations x,y are such that $T(x) = T(y)$ for sufficient statistic T, then inference based on x and y should be the same.*

The issue with taking expectations over all datasets X is that it does not adhere to the conditionality principle and hence the reason for Bayesians not liking the notion of fixing θ and taking expectation over X.
In Bayesian approach to decision theory, we construct a loss function $L(\theta, \delta(X))$ for a decision $\delta(X)$. From this, we can define the posterior risk, which conditions on x and integrates over Θ with a prior π.

Definition 50.377 *(Posterior risk). The posterior risk is defined as*

$$p(\pi, \delta(x)) = \int L(\theta, \delta(x))p(\theta|x)d\theta.$$

The Bayes action $\delta^*(x)$ for any fixed x is the decision $\delta(x)$ that minimises the posterior risk.

Lemma 50.378 *Under squared error loss, the posterior mean is the Bayes action.*

However, note that in frequentists, we can define the frequentist risk

$$R(\theta, \delta) = E_\theta L(\theta, \delta(X))$$

where we take expectation over X with parameter θ fixed. However, we can combine the two ideas.

Definition 45: Bayes rule

A Bayes rule is a function δ_π that minimises

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)d\theta$$

where $R(\theta, \delta)$ is the frequentist risk. This averages the frequentist risk over a prior distribution of θ.
The **Bayes risk** is $r(\pi) = r(\pi, \delta_\pi)$ is the Bayes rule with the Bayes rule plugged in.

Definition 50.379 *(Conjugate prior). A family of priors such that, upon being multiplied by the likelihood, yields a posterior in the same family.*

Note that we distinguish between objective priors (priors chosen based on the likelihood) and subjective priors (based on domain knowledge).

50.11.2 Exponential families and conjugate priors

Nearly everything we have seen are exponential families. Notable exceptions to this are the Cauchy distribution and t-distribution.
Recall that a conjugate prior is the case when the posterior is of the same distribution as the prior.

1. Beta is conjugate prior for the Bernoulli.
2. Gamma is the conjugate prior to the exponential.

Remark 50.380 *Note that there isn't ONE conjugate prior for a distribution. There is* ***a*** *conjugate prior for a distribution.*

Theorem 50.381 *Any exponential family has a conjugate prior.*