STAT8027 Final Review Outline

C.C

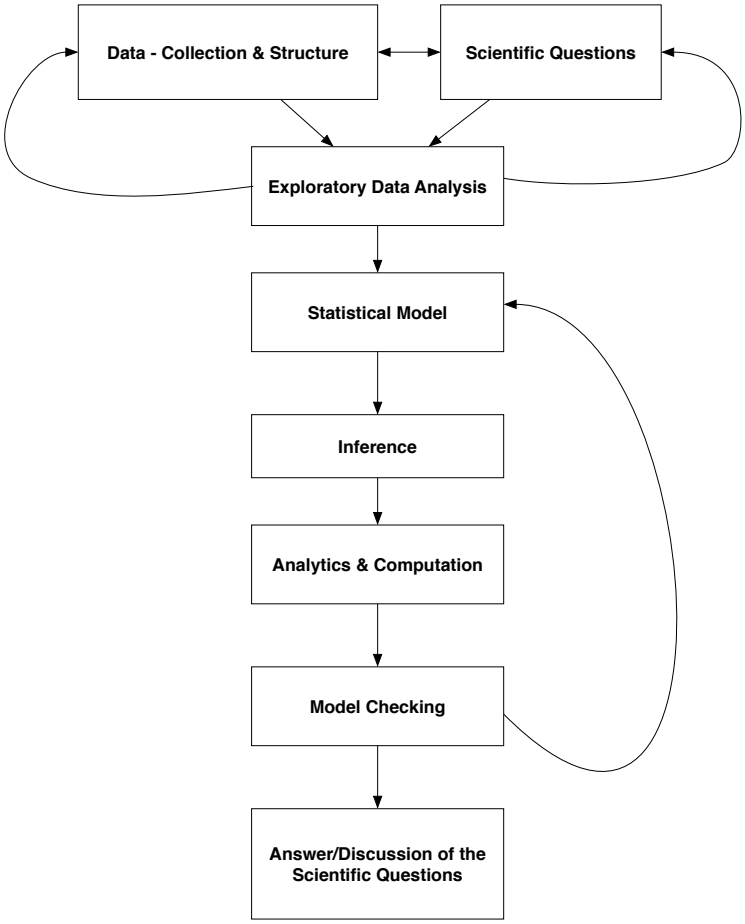
Catalogue

Reference

1. STAT8027 Lecture slides
   Lecture: Dr. Anton Westveld
   ANU - RSFAS
2. *Statistical Inference* (second edition)
   Paul Garthwaite, Ian Joliffe, and Byron Jones
   Oxford Science Publication

Introduction

Workflow



Areas:

1. Point estimation
2. Interval estimation
3. Hypothesis testing

Estimation Methods:

1. Least-squares Estimation
2. Method of moments
3. Maximum Likelihood Estimation (MLE)
4. Bayesian
5. Non-parametric
6. …

1. Generation Random Variables

1.1 Monte Carlo integration

- Many quantities of statistical analyses can be expressed as the expectation of a function of a random variable $E[h(X)]$.
- Let $f(X|\theta)$ denote the density of X
- Let μ denote the expectation of $h(X)$.
- Then when an iid sample $X_1, \ldots, X_n$ is obtained from $f(X;\theta)$, we can approximate μ by a sample average:

$$\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \int h(x)f(x)dx = \mu$$

$$\hat{\sigma}^2_{MC} = \frac{1}{n-1}\sum_{i=1}^{n}(h(X_i) - \hat{\mu}_{MC})^2 \to \sigma^2$$

1.2 Generating Random Samples

The probability inverse transform (Tutorial 0~1)

L12a

2. Theorems between distributions (Proof matters!)

Theorem R1: If Z is a standard normal random variable, the $U = Z^2$ is a χ^2 distribution with 1 degree of freedom.

Theorem R2: If $U_1, \ldots, U_n$ are independent and $X_i \sim \chi^2_{r_i}$ then

$$U_1 + U_2 \sim \chi^2_{n+m} \qquad \sum_{i=1}^{n} U_i \sim \chi^2_n$$

Theorem Extend

3. If

- $Z \sim \text{normal}(0,1)$
- $U \sim \chi^2_n$
- Z and U are independent, then:

$T = Z/\sqrt{U/n}$ is a t distrbution with n degrees of freedom

$$f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

4. If

- $U \sim \chi^2_m$
- $V \sim \chi^2_n$
- U and V are independent then: $W = \dfrac{U/m}{V/n} \sim \text{F}(m,n)$

5. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\mu, \sigma^2)$, then

1. $\bar{X} \sim \text{normal}(\mu, \sigma^2/n)$
2. $\bar{X}$ and S^2 are independent
3. $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$

6.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Properties of Estimators

Properties and MVUE

1. Unbiased

Definition 2.1: $\hat{\theta} = T(X_1, \ldots, X_n)$ is an unbiased estimator for θ if $E[T(\boldsymbol{X})] = \theta$. The bias of an estimator is defined as:

$$\text{bias}(\hat{\theta}) = E[T(\boldsymbol{X})] - \theta$$

1.1 $$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

$$\begin{aligned} E\left[(\hat{\theta} - \theta)^2\right] &= E\left[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2\right] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + 2(E(\hat{\theta}) - \theta)E[(\hat{\theta} - E(\hat{\theta}))] + E[(E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + 0 + E[(E(\hat{\theta}) - \theta)^2] \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2] + (E(\hat{\theta}) - \theta)^2 \\ &= V(\hat{\theta}) + \text{Bias}(\hat{\theta})^2 \end{aligned}$$

2. Weakly consistent

Definition 2.1: An estimator $\hat{\theta}$ is **weakly consistent** if

$$P(|\hat{\theta} - \theta| > \epsilon) \;\to\; 0 \quad \text{as} \quad n \;\to\; \infty$$

for any $\epsilon > 0$.

3. Efficiency

An unbiased estimator is said to be ***efficient*** if it has the minimum possible variance; the ***efficiency*** of and unbiased estimator is the ratio of the minimum possible variance to the variance of the estimator.

3.1 Cramer-Rao Inequality [lower bound]

Section 2.4 (Cramer-Rao Inequality [lower bound]): Let $X_1, \ldots, X_n$ be a random sample from a distribution family with density function $f_X(x; \theta)$ where θ is a scalar parameter. Also, let $T = t(X_1, \ldots, X_n)$ be an unbiased estimator for $\tau(\theta)$. Then under certain regularity (smoothness) conditions:

$$Var(T) \geq \frac{\{\tau'(\theta)\}^2}{ni(\theta)} = \{\tau'(\theta)\}^2 I(\theta)^{-1}$$

- $\tau'(\theta) = \frac{d}{d\theta}\tau(\theta)$
- Where $I(\theta) = ni(\theta)$ is called the Expected Fisher Information.
- $I(\theta) = E\left[\left(\frac{\partial l(\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right]$

C-R Inequality Extended: Let $X_1, \ldots, X_n$ be a sample [note we don't have to have iid] with pdf $f(\boldsymbol{x}|\theta)$ and let $T(\boldsymbol{X})$ be an estimator [doesn't have to be unbiased] then based on regularity conditions we have:

$$V\left[T(\boldsymbol{X})\right] \geq \frac{\left[\frac{\partial}{\partial\theta}E[T(\boldsymbol{X})]\right]^2}{E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x};\theta)\right)^2\right]} = \frac{\left[\frac{\partial}{\partial\theta}E[T(\boldsymbol{X})]\right]^2}{I(\theta)}$$

- If $E[T(\boldsymbol{X})] = \tau(\theta)$, so $T(\boldsymbol{X})$ is an unbiased estimator for $\tau(\theta)$,
- If we have iid samples:

$$V\left[T(\boldsymbol{X})\right] \geq \frac{[\tau'(\theta)]^2}{ni(\theta)}$$

Regularity conditions

- $\frac{\partial}{\partial\theta}ln\{f(x|\theta)\}$ exists for all x and θ;
- interchange of integration and differentiation is permissible;
- The expectation $i(\theta) = E\left[\left[\frac{\partial}{\partial\theta}ln\{f(x;\theta)\}\right]^2\right]$, where X is a generic random variable having distribution with density $f(x|\theta)$, is finite for all $\theta \in \Theta$.

Fisher Information

- The Fisher information, or expected Fisher information, or the information number is:

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(\boldsymbol{x};\theta)\right)^2\right] = -E\left[\left(\frac{\partial^2}{\partial\theta^2}\ log\ f(\boldsymbol{x};\theta)\right)\right]$$

- For one data point we have:

$$i(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ log\ f(x;\theta)\right)^2\right]$$

- For iid data:

$$ni(\theta) = I(\theta)$$

(Principles of Data Reduction, L03a / p2)

4. Sufficiency

Sufficiency Principle: If $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through $T(X_1, \ldots, X_n)$.

Definition 2.5: A statistic $T(X_1, \ldots, X_n)$ is **sufficient** for θ if the conditional distribution of the sample $X_1, \ldots, X_n$ given $T(X_1, \ldots, X_n)$ does not depend on θ.

Theorem 2.1 (the factorization theorem/criterion): Suppose $X_1, \ldots, X_n$, form a random sample from $f(x; \theta)$. Then $T(\boldsymbol{X})$ is a sufficiant statistic for θ **if and only if** there exists two non-negative functions K_1 and K_2, such that the likelihood $L(\theta, \boldsymbol{x})$ can be written:

$$L(\theta; \boldsymbol{x}) = K_1 \left[t(\boldsymbol{x}); \theta \right] \; K_2 \left[\boldsymbol{x} \right]$$

4.1 Minimal Sufficient

Definition 2.6: A sufficient statistic $T(\boldsymbol{X})$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(\boldsymbol{X})$, $T(\boldsymbol{X})$ is a function of $T'(\boldsymbol{X})$.

- Not easy to use the definition to find a minimal sufficient statistic!

Lemma 2.3: Let $f(\boldsymbol{x}; \theta)$ be the pdf or pmf of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$ the ratio

$$L(\theta; \boldsymbol{x}) / L(\theta; \boldsymbol{y})$$

is constant as function of θ [note: this can be a vector] if and only if

$$T(\boldsymbol{x}) = T(\boldsymbol{y}).$$

Then $T(\boldsymbol{X})$ is a minimal sufficient statistic.

Example:

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} n(\mu, \sigma^2)$, with both μ, σ^2 unknown.
- Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two sample points.
- Let $(\bar{x}, s_{\boldsymbol{x}}^2)$ and $(\bar{y}, s_{\boldsymbol{y}}^2)$ be the sample means and sample variances for the samples $\boldsymbol{x}$ and $\boldsymbol{y}$.

$$\begin{aligned} \frac{f(\boldsymbol{x}|\mu, \sigma^2)}{f(\boldsymbol{y}|\mu, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} exp(-[n(\bar{x} - \mu)^2 - (n-1)s_{\boldsymbol{x}}^2]/(2\sigma^2))}{(2\pi\sigma^2)^{-n/2} exp(-[n(\bar{y} - \mu)^2 - (n-1)s_{\boldsymbol{y}}^2]/(2\sigma^2))} \\ &= exp([-n(\bar{x}^2 - \bar{y}^2) + 2n\mu(\bar{x} - \bar{y}) - (n-1)(s_{\boldsymbol{x}}^2 - s_{\boldsymbol{y}}^2)]/(2\sigma^2)) \end{aligned}$$

- This ratio will not depend on μ and σ^2 if and only if $\bar{x} = \bar{y}$ and $s_{\boldsymbol{x}}^2 = s_{\boldsymbol{y}}^2$.
- $(\bar{X}, S^2)$ are minimally sufficient for μ, σ^2.

4.2 ***Rao-Blackwell Theorem*** (How sufficiency helps finding MVUEs)

- Let W be any unbiased estimator of $\tau(\theta)$.
- Let T be a sufficient statistic for θ.
- Define $\phi(T) = E[W|T]$.
- Then

$$E[\phi(T)] = \tau(\theta)$$

$$V[\phi(T)] \leq V[W]$$

- So if we have unbiased estimator and condition it on a sufficient statistic, our new statistic $\phi(T)$ has the same or smaller variance!!

Corollary to Theorem

If an MVUE for θ exists, then there must be a function of the minimum sufficient statistic for θ which is an MVUE.

5. Complete

Definition 2.9: Let $f_T(t;\theta)$ be a family of of pdfs or pmfs for a statistic $T(\boldsymbol{x})$. The family of probability distributions is called complete if

$$E[h(T)] = \int h(t) f_T(t) dt = 0$$

for all θ implies that P(h(T) = 0) =1

for all θ.

Scheffe Theorem

Lemma 2.6: Let $X_1, \ldots, X_n$ be a random sample from a distribution with density function $f(x;\theta)$. If $T = T(\boldsymbol{X})$ is a complete and sufficient statistic, and $\phi(T)$ is an unbiased estimator of $\tau(\theta)$, then $\phi(T)$ is the unique MVUE of $\tau(\theta)$.

How to find MVUEs?

- How to find MVUEs? It seems we have an approach:
  1. Find or construct a sufficient and complete statistic T.
  2. Find an unbiased estimator W for $\tau(\theta)$.
  3. Compute $\phi(T) = E[W|T]$, then $\phi(T)$ is the UMVUE.
- Or:
  1. Find or construct a sufficient and complete statistic T.
  2. Find a function $h(T)$, where $E[h(T)] = \tau(\theta)$ (i.e. it is unbiased).
  3. Then $h(T)$ is the MVUE.

Exponential Families

Definition 2.7: We say that a random variable belongs to the k-parameter exponential family of distributions if its pdf can be written in the following form:

$$f(x;\boldsymbol{\theta}) = exp\left(\sum_{j=1}^{k} A_j(\boldsymbol{\theta})B_j(x) + C(x) + D(\boldsymbol{\theta}))\right)$$

$$f(x;\boldsymbol{\theta}) = C^*(x)D^*(\boldsymbol{\theta})exp\left(\sum_{j=1}^{k} A_j(\boldsymbol{\theta})B_j(x)\right)$$

- If we define:

$$\phi = (\phi_1, \ldots, \phi_k) = A(\boldsymbol{\theta}) = \{A_1(\theta), \ldots, A_k(\theta)\}$$

then ϕ is referred to as the canonical parameter for the exponential family and the density function can be written in the form:

$$f(x;\theta) = exp\left\{\sum_{j=1}^{k} \phi_i B_i(x) + C(x) + D(\phi)\right\}$$

- Note:

$$\theta = A^{-1}(\phi)$$
$$D(\phi) = D\{A^{-1}(\phi)\}$$

Exponential families and MVUE

Lemma 2.4: If the usual regularity condition hold, then a vector of k sufficient statistics $\boldsymbol{T}$ exists for a vector or parameters $\boldsymbol{\theta}$ if and only if the distribution of $\boldsymbol{X}$ belong to the k-parameter exponential family.

Proof:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = exp\left\{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})\left(\sum_{i=1}^{n} B_j(x_i)\right) + nD(\boldsymbol{\theta}) + \sum_{i=1}^{n} C(x_i)\right\}$$

- Let $\boldsymbol{t} = (\sum_{i=1}^{n} B_1(x_i), \ldots, \sum_{i=1}^{n} B_k(x_i))$.

- $K_1 = exp\left\{\sum_{j=1}^{k} A_j(\boldsymbol{\theta})t_j + nD(\boldsymbol{\theta})\right\}$
- $K_2 = exp\left\{\sum_{i=1}^{n} C(x_i)\right\}$

Lemma 2.5: Under the same conditions as Lemma 2.4, $\boldsymbol{T}$ is also miniminal sufficient.

Lemma 2.8: Under the same conditions as Lemma 2.4, $\boldsymbol{T}$ is also complete.

- For exponential families, we can easily find complete and sufficient statistics.
- All that is needed to find $h(T)$ such that $E[h(T)] = \tau(\theta)$, then we have the unique MVUE!

Methods of Estimation

Method of Moments (MOM)

1. MOM

- Typically the population moments are implicitely defined by parameters $\theta = (\theta_1, \ldots, \theta_K)$.
- Equate the sample and population moments:

$$\begin{aligned} \mu_1(\theta_1, \ldots, \theta_K) &= \hat{\mu}_1(x_1, \ldots, x_n) \\ &\vdots \quad\quad \vdots \\ \mu_K(\theta_1, \ldots, \theta_K) &= \hat{\mu}_K(x_1, \ldots, x_n) \end{aligned}$$

- The estimator $T(\boldsymbol{X}) = \tilde{\theta}$ is the value for θ which solves the system of K equations.

2. Central Moments

$$\mu'_k = E_\theta(\{X - E_\theta(X)\}^k)$$

And sample moments:

$$\hat{\mu}'_k = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

for $k = 2, \ldots, K$. Set $\mu_1 = \hat{\mu}_1 = \bar{x}$.

3.Generalized MOM

$$\begin{aligned} E_\theta(g_1(X)) &= \frac{1}{n} \sum_{i=1}^{n} g_1(x_i) \\ &\vdots \\ E_\theta(g_k(X)) &= \frac{1}{n} \sum_{i=1}^{n} g_k(x_i) \end{aligned}$$

Note: If we set $g_i(x) = x^i$ then we recover the standard method of moments.

Maximum Likelihood Estimation

*"This is the **best known, most widely used, most intuitive, most important**, . . . of estimation procedures."*

1. Defintion:

Definition 3.1: The MLE is the value $\hat{\boldsymbol{\theta}}$ which maximizes $L(\boldsymbol{\theta}; \boldsymbol{x})$.

- If the likelihood is differentiable in (θ_i), possible candidates for the MLE are the values $(\theta_1, \ldots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta; \boldsymbol{x}) = 0, \quad i = 1, \ldots, k$$

- Possible: local vs. global maximum, extrema may occur on the boundary and thus the first derivative may not be 0, ...
- All the good points and bad points of optimizing a function are here!

The log likelihood function is $\ell(\theta) = log[L(\theta)]$

Score Function is the gradient of $\ell(\theta)$ $U(\theta) = \frac{\partial \ell}{\partial \theta} = \ell'(\theta)$

2. General approach for checking 2-parameter MOM

1. First-order partial derivatives (score equations) at $\hat{\theta}_1, \hat{\theta}_2$ are zero:

$$\frac{\partial}{\partial \theta_1} H(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0$$
$$\frac{\partial}{\partial \theta_2} H(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0$$

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

$$\frac{\partial}{\partial \theta_1^2} H(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0$$
$$\frac{\partial}{\partial \theta_2^2} H(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0$$

3. The determinant of the matrix of second order partial derivatives (the Hessian matrix) is positive.

$$\begin{vmatrix} \frac{\partial}{\partial \theta_1^2} & \frac{\partial}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial}{\partial \theta_2 \partial \theta_1} & \frac{\partial}{\partial \theta_2^2} \end{vmatrix}_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} > 0$$

3. **Newton-Raphson (N-R) Method**

- N-R is an extremely fast root finding approach, but is sensitive to starting values.
- Again let's consider a log likelihood function $\ell(\theta|\boldsymbol{x})$.
- Let $U(\theta) = \ell'(\theta|\boldsymbol{x})$ denote first derivative of $\ell(\theta)$.
- Let $H(\theta) = \ell''(\theta|\boldsymbol{x})$ denote the derivative of $\ell(\theta)$.
- Let θ_0 be an initial estimate of θ.
- Let $\hat{\theta}$ be the MLE.
- Let's do a Taylor series expansion of $U(\theta)$ around θ_0

$$U(\theta) = U(\theta_0) + (\theta - \theta_0)H(\theta_0) + \cdots$$

- At $\theta = \hat{\theta}$ we know $U(\hat{\theta}) = 0$, so we have

$$0 = U(\theta_0) + (\hat{\theta} - \theta_0)H(\theta_0) + \cdots$$

- Let's say the one-step approximation is reasonable enough.

$$\hat{\theta} = \theta_0 - H^{-1}(\theta_0)U(\theta_0)$$

- This suggests that $\hat{\theta}$ is well approximated by θ_1:

$$\theta_1 = \theta_0 - H^{-1}(\theta_0)U(\theta_0)$$

- We can then get an improved estimate:

$$\theta_2 = \theta_1 - U(\theta_1)H^{-1}(\theta_1)$$

- We can continue with $\theta_3, \theta_4, \ldots$ until convergence is achieved.

$$|\theta_k - \theta_{k-1}| < \epsilon = 1e-07$$

3.1 Multivariate N-R Method

- We can naturally extend the N-R approach to a multivariate setting.
- Let $U(\boldsymbol{\theta})$ denote the vector of first partial derivatives of $\ell(\boldsymbol{\theta})$.
- Let $H(\boldsymbol{\theta})$ denote the matrix of second partial derivatives of $\ell(\boldsymbol{\theta})$.

3.2 Fisher Scoring

- The method of "scoring" is a simple modification of the Newton-Raphson method.
- The Hessian $H(\boldsymbol{\theta})$ is replaced by its expectation.

$$E[H(\boldsymbol{\theta})] = -I(\boldsymbol{\theta})$$

where $I(\boldsymbol{\theta})$ is Fisher's information matrix.

$$I(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \theta_i}\right)\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \theta_j}\right)\right] = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \theta_i \partial \theta_j}\right]$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + I^{-1}(\boldsymbol{\theta}_t)U(\boldsymbol{\theta}_t)$$

- A great advantage is that $E[H(\boldsymbol{\theta})]$ is guaranteed to be positive definite, thus eliminating some possible convergence issue with the Newton-Raphson approach.

4. Expectation - Maximization (EM) Algorithm

- Presentation adapted from CB & *Computational Statistics*.
- The EM algorithm is a general algorithm to find MLEs when some of the data are missing (or the problem can be set in a manner that there are missing data).
- Suppose we observe all of the data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, then all we do to find the MLE is maximize: $\ell(\boldsymbol{\theta}; \boldsymbol{y})$
- Suppose we don't observe all the $\boldsymbol{y}$s then based on the notation by Donald Rubin we have $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss})$.

$$\begin{aligned} f(\boldsymbol{y}; \boldsymbol{\theta}) &= f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}; \boldsymbol{\theta}) \\ &= k(\boldsymbol{y}_{miss} | \boldsymbol{y}_{obs}, \boldsymbol{\theta}) g(\boldsymbol{y}_{obs}; \boldsymbol{\theta}) \end{aligned}$$

- This leads to: $g(\boldsymbol{y}_{obs}; \boldsymbol{\theta}) = \frac{f(\boldsymbol{y}; \boldsymbol{\theta})}{k(\boldsymbol{y}_{miss} | \boldsymbol{y}_{obs}, \boldsymbol{\theta})}$

$$\begin{aligned} log\left[g(\boldsymbol{y}_{obs}; \boldsymbol{\theta})\right] &= log\left[f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}; \boldsymbol{\theta})\right] - log\left[k(\boldsymbol{y}_{miss} | \boldsymbol{y}_{obs}, \boldsymbol{\theta})\right] \\ \ell_{obs}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}) &= \ell_{comp}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}) - log\left[k(\boldsymbol{y}_{miss} | \boldsymbol{y}_{obs}, \boldsymbol{\theta})\right] \end{aligned}$$

- As $\boldsymbol{y}_{miss}$ is missing, we replace the right side of the equation with its expectation:

$$\begin{aligned} \ell_{obs}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}) &= E\left\{\ell_{comp}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}) \middle| \boldsymbol{\theta}', \boldsymbol{y}_{obs}\right\} \\ &\quad - E\left\{log\left[k(\boldsymbol{y}_{miss} | \boldsymbol{y}_{obs}, \boldsymbol{\theta})\right] \middle| \boldsymbol{\theta}', \boldsymbol{y}_{obs}\right\} \end{aligned}$$

4.1 E-M Steps

- The EM algorithm seeks to maximize $\ell(\boldsymbol{\theta}; \boldsymbol{y}_{obs})$ with respect to $\boldsymbol{\theta}$ through the following process:

don't want missing data so integrate it out

1. **E step**: Calculate the expectation of the complete likelihood conditional on the observed data and the current value of $\boldsymbol{\theta}$:

$$\begin{aligned} Q\left(\boldsymbol{\theta} | \boldsymbol{\theta}^{(r)}\right) &= E\left\{\ell_{comp}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}) \middle| \boldsymbol{\theta}^{(r)}, \boldsymbol{y}_{obs}\right\} \\ &= \int \left[\ell_{comp}(\boldsymbol{\theta}; \boldsymbol{y}_{obs}, \boldsymbol{y}_{miss})\right] k(\boldsymbol{y}_{miss} | \boldsymbol{y}_{obs}, \boldsymbol{\theta}) d\boldsymbol{y}_{miss} \end{aligned}$$

conditional

2. **M step**: Maximize $Q\left(\boldsymbol{\theta} | \boldsymbol{\theta}^{(r)}\right)$ with respect to $\boldsymbol{\theta}$. Set $\boldsymbol{\theta}^{(r+1)}$ equal to the maximizer of Q.
3. Return to the E step unless a stopping criterion has been reached.

5. Invariance Property of MLEs

Lemma 3.2: Suppose that $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ represent two alternative parameterizations for some probability distirbution and that $\boldsymbol{\eta}$ is a (1-1) function of $\boldsymbol{\theta}$, so that we can write $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta}), \boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\eta})$ for appropriate functions $\boldsymbol{g}(\cdot), \boldsymbol{h}(\cdot)$.

- If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ then $\hat{\boldsymbol{\eta}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}})$ is the MLE for $\boldsymbol{\eta}$
- If the mapping is (1-1) we simply note:
$$\eta = \tau(\theta) \rightarrow \tau^{-1}(\eta) = \theta$$
- Define our likelihood based on the reparameterization ($\theta = \tau^{-1}(\eta)$):
$$L^*(\eta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \tau^{-1}(\eta)) = L(\tau^{-1}(\eta); \boldsymbol{x}) = L(\theta; \boldsymbol{x})$$
- We find the supremumum of likelihood
$$\sup_{\eta} L^*(\eta; \boldsymbol{x}) = \sup_{\eta} L(\tau^{-1}(\eta); \boldsymbol{x}) = \sup_{\theta} L(\theta; \boldsymbol{x})$$
to see that the maximum of $L^*(\eta; \boldsymbol{x})$ is when $\eta = \tau(\theta) = \tau(\hat{\theta})$.

5.1 not one-to-one case

- However, many functions of interest are not one-to-one: $\theta \rightarrow \theta^2$.
- We proceed by defining the induced likelihood function of L^* for $\tau(\theta)$
$$L^*(\eta; \boldsymbol{x}) = \sup_{\theta: \tau(\theta) = \eta} L(\theta; \boldsymbol{x})$$
- The value $\hat{\eta}$ that maximizes $L^*(\eta; \boldsymbol{x})$ will be called the MLE of η.

6. Asymptotic of MLE

Score Function

Lemma: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta)$ and let $\ddot{\theta}$ be the MLE of θ. Under regularity conditions of $f(x; \theta)$ and thus $L(\theta; \boldsymbol{x})$ (under appropriate smoothness conditions), we can state:

$$W = \frac{1}{\sqrt{n}} \ell'(\theta; \boldsymbol{x}) \xrightarrow{D} \text{normal}(0, i(\theta))$$

MLE

Lemma 3.3: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta)$. Let $\hat{\theta}$ be the MLE of θ. Under regularity conditions of $f(x; \theta)$ and thus $L(\theta; \boldsymbol{x})$ (under appropriate smoothness conditions), we have:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \text{normal}(0, i(\theta)^{-1})$$

6.1 Delta Method

Theorem: Let Y_n be a sequence of random variables such that:

$$\sqrt{n}(Y_n - \theta) \xrightarrow{D} \text{normal}(0, \sigma^2)$$

For a given function g and a specific value θ, suppose that $g'(\theta)$ exists and is not 0, then

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{D} \text{normal}(0, \sigma^2[g'(\theta)]^2)$$

Extension

Lemma: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$. Let $\hat{\theta}$ be the MLE of θ and let $tau(\theta)$ be a continuous function of θ. Under regularity conditions (i.e. under appropriate smoothness conditions) of $f(x;\theta)$ and thus $L(\theta; \boldsymbol{x})$, we have:

$$\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \xrightarrow{D} \text{normal}(0, \nu(\theta))$$

Where $\nu(\theta) = \frac{[\tau'(\theta)]^2}{i(\theta)}$ is the Cramer-Rao lower bound for a single data point.

$$\tau(\hat{\theta}) \dot{\sim} \text{normal}\left(\tau(\theta), \frac{[\tau'(\theta)]^2}{\mathbf{I}(\theta)}\right)$$

6.2 *Asymptotic of MLE*

So asymptotically, MLEs are:

1. unbiased;
2. achieve the Cramer-Rao lower bound (efficient);
3. asymptotically normally distributed.

- We can also note that MLEs are consistent estimators.
- Because these estimators achieve (1-3) they are *asymptotically efficient! **best asymptotically normal (BAN) estimators**

Hypothesis Testing

1. Definition

Definitions: Suppose that $X_1, \ldots, X_n$ represent a simple random sample from a parametric family with density function $f(x;\theta)$ for some parameter $\theta \in \Theta$.

- A statistical hypothesis is simply a subset of the parameter space, Θ.
- Any statistical hypothesis of interest, often termed the null hypothesis, is associated with a competing alternative hypothesis.
- A null hypothesis and its alternative form a partition of the parameter space Θ consisting of the sets Θ_0 and

$$\Theta_1 = \Theta_0^c \cap \Theta$$

Definition: A hypothesis testing procedure or hypothesis test is a rule that specifies:

1. For which sample values the decision is made to accept H_0.
2. For which sample values H_0 is rejected and H_1 is accepted as true.

2. Rejection Region (C)

The subset of the space for which H_0 will be rejected is called the rejection region or critical region (C). The complement of the rejection region is called the acceptance region.

we can define a statistical test in terms of a rejection region (C) which is just a set for some statistic $T(X_1, \ldots, X_n)$:

$$C = \{X \in \mathcal{X} : T(X) < k\}$$

3. Type I and Type II errors

- Type I Error: Reject H_0 given that it is true. Thus the observations fall in the rejection region C when in fact that null hypothesis, H_0, is true.
- Type II Error: Do not Reject H_0 when it is false: Thus the observed data values fall outside the rejection region when in fact the null hypothesis is false.

	Decision	
Truth	Accept H_0	Accept H_1
H_0	Correct Decision	Type I Error
H_1	Type II Error	Correct Decision

There is a strong relationship between Type I and Type II errors. Note that for a given value of θ, only one type of error can occur (since for any given θ, H_0 either is or is not true).

3.1 Significance Level & Power

Definition 4.2: The probability of a Type I error, α in a test of hpotheses is called the **size** or **signifigance level** of the test. The complement of the probability of a Type II error

$$\eta(\theta) = 1 - \beta, \quad \text{is the } \textbf{power} \text{ of the test.}$$

- **Power = 1 - P(Type II Error)**
  $= 1 - P(\boldsymbol{X} \in C^c | H_1 \text{ is true}) = P(\boldsymbol{X} \in C | H_1 \text{ is true})$
- Given that H_1 is true, what is the probability I reject H_0

- It is standard to focus on test which have sizes 0.05 or 0.01.
- If we focus on tests with rejection regions of the form $C = \{X < k_\alpha\}$, we can choose k_α such that:

Maximize the power when $\theta \in \Theta_0$

$$\begin{aligned} max_{\theta \in \Theta_0} \eta(\theta) &= max_{\theta \geq 1000} 1 - exp(-k_\alpha/\theta) \\ &= 1 - exp(-k_\alpha/1000) = \alpha \end{aligned}$$

3.2 UMP Test

Definition 4.3 *A test which minimizes β for fixed α is called a most powerful or best test of size α.*

Definition 4.5 *Suppose we have a test of size α; then it is uniformly most powerful (UMP) of size α if its power function $\eta(\theta)$ is such that $\eta(\theta) \geq \eta^*(\theta)$ for all $\theta \in \Omega - \omega$, where $\eta^*(\theta)$ is the power function of any other size-α test.*

4. Essential Nature of a Hypothesis Test (Experimental Design, Hoff 2009)

Given H_0, H_1 and data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$:

1. From the data, compute a relevant test statistic $T(\boldsymbol{x})$: The test statistic $T(\boldsymbol{x})$ should be chosen so that it can differentiate between H_0 and H_1 in ways that are scientifically relevant. Typically, $T(\boldsymbol{x})$ is chosen so that

$$T(\boldsymbol{x}) \text{ is probably } \begin{cases} \text{small under } H_0 \\ \text{large under } H_1 \end{cases}$$

2. Obtain a null distribution: A probability distribution over the possible outcomes of $T(\boldsymbol{X})$ under H_0. Here, $X = \{X_1, \ldots, X_n\}$ are potential experimental results that could have happened under H_0.

3. Compute the p-value: The probability under H_0 of observing a test statistic $T(\boldsymbol{X})$ as or more extreme than the observed statistic $t(\boldsymbol{x})$.

$$\text{p-value} = P(T(\boldsymbol{X}) \geq t(\boldsymbol{x}) | H_0)$$

If the p-value is small $\Rightarrow$ evidence against H_0
If the p-value is large $\Rightarrow$ not evidence against H_0

5. Neyman-Pearson Set-up (simple case)

- Suppose that $X_1, \ldots, X_n$ are a sample from a population characterized by a probability model with density function $f(x;\theta)$ for $\theta \in \Theta$ where $\Theta = \{\theta_0, \theta_1\}$.

$$H_0: \quad \theta = \theta_0$$
$$H_1: \quad \theta = \theta_1$$

- Consider the likelihood-ratio:

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_0; \boldsymbol{x})}{L(\theta_1; \boldsymbol{x})}$$

- The test we shall define has a critical region of the form

$$C = \{\lambda(\boldsymbol{x}) \leq k\}$$

- The ratio of the likelihood for any given sample at each of the two possible parameter values is precisely a relative measure of how plausible the two hypotheses are.
- In other words, when $\lambda(\boldsymbol{x})$ is very small, this is strong evidence that the observations arose from the alternative hypothesis rather than the null hypothesis.
- It should seem intuitively reasonable that the likelihood ratio is a good method of distinguishing between samples which support the null hypothesis versus samples which support the alternative hypothesis.
- From what we have done, we know for a given α we could compare the power $\beta(\theta)$.
- We would like to find a uniformily most powerful test . $\eta(\theta) \geq \eta(\theta^*)$
- It turns out that N-P tests lead to UMP tests.
- Suppose that H_0 and H_1 are simple hypotheses and that the test that reject Note: $0 \leq \lambda \leq 1$, and λ will be close to 1 if H_0 is true. nificance level Where $0 \leq k \leq 1$.
- **Lemma 4.2:** Then any other test for which the significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test.

Examples compare H1 = to H1> and H1<, they have the same result as the result of H1 = (applying NP), which is UMP.

5. *Maximum Likelihood Ratio Test*

$$H_0: \theta \in \omega \text{ versus } H_1: \theta \in \Omega - \omega$$

$$\lambda(\boldsymbol{x}) = \frac{\max\limits_{\Theta \in \omega} L(\theta; \boldsymbol{x})}{\max\limits_{\Theta \in \Omega} L(\theta; \boldsymbol{x})}$$

- $\max\limits_{\Theta \in \omega} L(\theta; \boldsymbol{x})$ is a restricted maximization.
- $\max\limits_{\Theta \in \Omega} L(\theta; \boldsymbol{x})$ is a unrestricted maximization.

We construct a test of the form:

$$C = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq k\}$$

Note: $0 \leq \lambda \leq 1$, and λ will be close to 1 if H_0 is true.
Where $0 \leq k \leq 1$.

5.1 Asymptotic of MLRT

Theorem For testing $H_0 : \theta \in \omega$ versus $H_1 : \theta \in \Omega - \omega$

suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ and $\hat{\theta}$ is the MLE of θ and $f(x;\theta)$ satisfies the regularity conditions (smoothness).

Then under H_0, as $n \to \infty$,

$$-2log[\lambda(\boldsymbol{x})] \xrightarrow{D} \chi^2_1$$

Theorem Extend $$-2log(\lambda) \xrightarrow{D} \chi^2_\nu$$

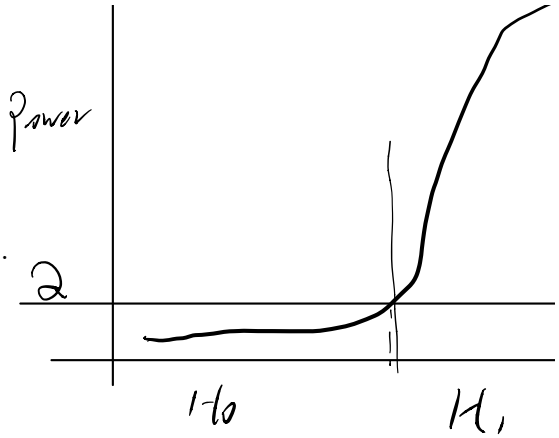
where $\nu = \#\text{number of constraints set in } H_0$.

- Another way to think about it is: Let p be the number of parameters estimated (are free) under H_1. And let p_0 be the the number of parameters estimated (are free) under H_0.
- Then $\nu = p - p_0$.

5.2 Properties of MLRTs

The MLRT

1. is asymptotically most powerful unbiased;
2. is asymptotically similar;
3. is asymptotically efficient.



6. *Unbiased* Test

Suppose that we wish to test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1.$$

A test of size α is said to be **unbiased** if

$$\eta(\boldsymbol{\theta}) \geq \alpha \text{ for all } \boldsymbol{\theta} \in \Theta_1.$$

7. *Similar* Test

Suppose that we wish to test:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1.$$

A test of size α is said to be **similar** if

$$\eta(\boldsymbol{\theta}) = \alpha \text{ for all } \boldsymbol{\theta} \in \Theta_0.$$

8. *Efficiency*

Suppose that we have ~~to~~ possible tests of H_0 vs. H_1, where both tests are simple.

If n_1 and n_2 are the minimum possible sample sizes for tests 1 and 2 for which we can achieve a size α and power $\geq \eta$, then the **relative efficiency** of test 1 compared to test 2 is:

$$n_2/n_1.$$

9. Two other tests

9.1 *The Score Test*

$$\boldsymbol{u}(\boldsymbol{\theta}) = \left(\frac{\partial \ell}{\partial \theta_1}, \frac{\partial \ell}{\partial \theta_2}, \cdots, \frac{\partial \ell}{\partial \theta_k} \right)^T$$

- Suppose that we wish to test:

$$H_0 : \ \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \Omega - \{\boldsymbol{\theta}_0\}.$$

- Test statistic:

$$\boldsymbol{u}(\boldsymbol{\theta})^T \ \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1} \boldsymbol{u}(\boldsymbol{\theta}) \ \dot{\sim} \ \chi^2_{df=k}$$

- Note: We don't have to determine the MLEs!

9.2 The Wald Test

- Suppose that we wish to test:

$$H_0 : \ \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \Omega - \{\boldsymbol{\theta}_0\}.$$

- Test statistic:

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \ \boldsymbol{I}_{\hat{\boldsymbol{\theta}}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \ \dot{\sim} \ \chi^2_{df=k}$$

- The MLRT, the Score Test, and the Wald Test are asymptotically equivalent!

Interval Estimation / Confidence set

1. Defintion

Suppose that $\{f(x;\theta);\theta \in \Omega\}$ define a family of distributions
If $S_{\boldsymbol{X}}$ is a subset of Ω, depending on $\boldsymbol{X}$, such that

$$P(\boldsymbol{X} : S_{\boldsymbol{X}} \supset \boldsymbol{\theta}) = 1 - \alpha$$

then $S_{\boldsymbol{X}}$ is a **confidence set** for $\boldsymbol{\theta}$ with **confidence coeficient** $1 - \alpha$.

Construction Methods

- parametric "exact" intervals
- parametric asymptotic intervals

Some general approaches

- Inverting a test statistic
- Pivotal Quantities
- Pivoting the CDF

2. Hypothesis Testing & Confidence Sets/Intervals

- There is a strong relationship between hypothesis testing and interval estimation. In general, every confidence set corresponds to a test and vice verse.

Lemma 5.1: Suppose that $\overline{C}(\boldsymbol{\theta}_0)$ is the acceptance region for a test of size α:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \in \boldsymbol{\Omega} - \boldsymbol{\theta}_0.$$

Then a **conficence set** for $\boldsymbol{\theta}$ with **confidence coefficient** $(1 - \alpha)$, is given by

$$S_{\boldsymbol{X}} = \{\boldsymbol{\theta}_0 : \ \boldsymbol{X} \in \overline{C}(\boldsymbol{\theta}_0)\}$$

3. Pivotal quantity

Definition 5.2: A random variable $g(\boldsymbol{X}, \boldsymbol{\theta})$ is a **pivotal quantity (or a pivot)** if the distribution of $g(\boldsymbol{X}, \boldsymbol{\theta})$ is independent of all parameters.

$X_i \sim N(\theta, 1)$ & $X - \theta \sim N(0,1)$ also a pivot.

$\bar{X} \sim N(\theta, \frac{1}{n}) \Rightarrow \frac{\bar{X} - \theta}{1/\sqrt{n}} \sim N(0,1)$ without θ, is pivot

16 / 43

we can construct a pivot $g(x, \theta)$ so that the distribution of g is not dependent on θ, thus we can use dis of g to get the interval of θ

4. MLEs & Asymptotic

$$\hat{\theta} \dot{\sim} \text{normal}(\theta, I(\theta)^{-1})$$

MLE →

$$\frac{\hat{\theta} - \theta}{1/\sqrt{I(\theta)}} \dot{\sim} \text{normal}(0, 1)$$

We have a pivotal quantity. Based on the same approach as before we can construct an asymptotic $100(1-\alpha)\%$ confidence interval as:

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

If we are interested in a function of θ, say $\tau(\theta)$, then we have:

$$\tau(\hat{\theta}) \dot{\sim} \text{normal}\left(\tau(\theta), \frac{[\tau'(\theta)]^2}{I(\theta)}\right)$$

$$\frac{\tau(\hat{\theta}) - \tau(\theta)}{\sqrt{\frac{[\tau'(\theta)]^2}{I(\theta)}}} \dot{\sim} \text{normal}(0, 1)$$

We can construct an asymptotic $100(1-\alpha)\%$ confidence interval as:

$$\left[\tau(\hat{\theta}) - z_{\alpha/2}\frac{\tau'(\hat{\theta})}{\sqrt{I(\hat{\theta})}},\ \tau(\hat{\theta}) + z_{\alpha/2}\frac{\tau'(\hat{\theta})}{\sqrt{I(\hat{\theta})}}\right]$$

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

$\pm\sqrt{CRLB(\tau(\theta))}$

4.1 Invertible Interval

interval construction, as we have done it, is not functionally equivalent!!

- Can we come up with an approach which does possess the equivariance property?
  - Yes, as long as the functional transformation in question is invertible.
  - Let's consider an asymptotic likelihood-based confidence interval procedure which is parameterization equivariant.
  - Specifically, this means that if we find a confidence region, C, for θ based on this new procedure and transform all of its values [which we sometimes denote as $\tau(C) = \{\tau(\theta) : \theta \in C\}$] then we will arrive at the same confidence region as if we had applied our new procedure to the parameter τ directly.

5. Asymptotic MLRT Interval Estimation (L09b/p26)

Did we have to use the asymptotic result of the LRT for our interval.
No, but it is more straightforward.

6. CDF Method

- Let T be a statistic with a continuous cdf $F_T(t;\theta)$.
- Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$.
- Suppose that for each $t \in T$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as:

1. If $F_T(t;\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by:
$$F_T(t;\theta_U(t)) = \alpha_1 \qquad F_T(t;\theta_L(t)) = 1-\alpha_2$$
2. If $F_T(t;\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by:
$$F_T(t;\theta_L(t)) = \alpha_1 \qquad F_T(t;\theta_U(t)) = 1-\alpha_2$$

Then the interval $[\theta_L(t), \theta_U(t)]$ is a $1-\alpha$ confidence interval for θ.

- We can prove that $F_T(t;\theta)$ is monotone in θ. See C&B.

Example: Consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} Unif(0,\theta)$.

- So we have the following CDF for X:

$$F_X(x;\theta) = \frac{x}{\theta}\mathbb{I}_{(0\le x\le\theta)}$$

- We know the MLE for θ is $T = max(X_1, \ldots, X_n)$

$$\begin{aligned} F_T(t;\theta) = Pr(T \le t) &= Pr\{max(X_1,\ldots,X_n) \le t\} \\ &= Pr\{X_1 \le t, \ldots, X_n \le t\} \\ &= Pr\{X_1 \le t\} \times \ldots \times Pr\{X_n \le t\} \\ &= \{F_X(t;\theta)\}^n \\ &= \frac{t^n}{\theta^n}\mathbb{I}_{(0\le t\le\theta)} \end{aligned}$$

- Note: $F_T(t;\theta)$ is a deceasing function for θ. Let $\alpha_1 = \alpha_2 = \alpha/2$. We have:

$$\begin{aligned} F_T(t;\theta_U(t)) &= \alpha/2 \\ \left(\frac{t}{\theta_U}\right)^n &= \alpha/2 \\ \theta_U &= t(\alpha/2)^{-(1/n)} \end{aligned}$$

$$\begin{aligned} F_T(t;\theta_L(t)) &= 1-\alpha/2 \\ \left(\frac{t}{\theta_L}\right)^n &= 1-\alpha/2 \\ \theta_L &= t(1-\alpha/2)^{-(1/n)} \end{aligned}$$

Bayesian Inference

1. Introduction

- Bayesian inference is the application of Bayes' rule to the problem of 'guessing' an unknown quantity(ies).
- The key distinction between Bayesian and sampling theory statistics is the issue of what is to be regarded as random and what is to be regarded as fixed.
  - To a Bayesian, parameters are random and data, once observed, are fixed.
  - To a sampling theorist (classical statistician), data are random even after being observed (how inference is done), but parameters are fixed.

2. Advantages

- Bayesian methods provide the user with the ability to formally incorporate prior information (Carlin and Louis, 2009).
- If $p(y|\theta)$ and $p(\theta)$ represent a rational person's beliefs then Bayes' rule is an optimal method for updating those beliefs.
- If $p(y|\theta)$ and $p(\theta)$ approximately represents beliefs then $p(\theta|y)$ is also approximate and may be useful (Hoff 2009).
- Inferences are conditional on the actual data (Carlin and Louis, 2009).
- Bayesian answers are more easily interpretable by non-statisticians (Carlin and Louis, 2009).
- All Bayesian analyses follow *directly* from the posterior; no separate theories of estimation, testing, multiple comparisons, etc. are needed (Carlin and Louis, 2009).
- In many complicated statistical problems there are no obvious non-Bayesian methods of estimation or inference (Hoff 2009).

3. ***Bayesian Inference***

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int_\Theta f(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \\ &= \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})} \\ &\propto f(\boldsymbol{x}|\theta)\pi(\theta)\end{aligned}$$

- $\pi(\theta|\boldsymbol{x})$ is the posterior distribution for θ.
- $f(\boldsymbol{x}|\theta)$ is the joint sampling distribution for $\boldsymbol{X}$ or the likelihood for θ.
- $\pi(\theta)$ is the prior distribution for θ.
- $m(\boldsymbol{x})$ is the marginal distribution for $\boldsymbol{X}$.

It is important to remember though that in a Bayesian framework θ is random. In the frequentist framework it is fixed!

- Bayesian approach:
  - We start with a prior belief about a situation (represented through a probability distribution parameterized by θ).
  - We observe data $\boldsymbol{x}$.
  - Given the data $\boldsymbol{x}$, we update our beliefs about θ via Bayes' rule and obtain the posterior distribution.

As $\pi(\theta|\boldsymbol{x})$ is a probability distribution function, we can consider a number of summaries for θ after we observe the data:

- Posterior mode: the value of θ which maximizes the posterior distribution
- Posterior median
- Posterior mean $\hat{\theta}_B = E\{[\theta|\boldsymbol{x}]\} = \int_\Theta \theta\pi(\theta|\boldsymbol{x})d\theta$
- We can also consider functions of θ, $\tau(\theta)$:

$$\widehat{\tau(\theta)}_B = E\{[\tau(\theta)|\boldsymbol{x}]\} = \int_\Theta \tau(\theta)\pi(\theta|\boldsymbol{x})d\theta$$

3.1 Conjugate *Family*

Definition (Section 6.4.2): Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$. A class $\mathcal{P}$ of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class $\mathcal{P}$ for all $f \in \mathcal{F}$, all priors in $\mathcal{P}$, and all $x \in \mathcal{X}$.

- In other words, if the prior distribution is in the same family of distributions as the posterior, then it is a conjugate prior distribution.
- For $f(\boldsymbol{x}|\theta)$ examine the kernel with regard to θ and see if your recognize the distribution. This will be the conjugate distribution.

Example: Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

- Determine a conjugate prior distribution and then determine the posterior distribution for λ.

$$\begin{aligned} f(\boldsymbol{x}|\lambda) &= \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod x_i!} \\ &\Rightarrow \lambda^{\sum x_i} e^{-n\lambda} \end{aligned}$$

- This is a kernel for a gamma distribution. Thus we have the following conjugate prior:

$$\lambda \sim \text{gamma}(a, b)$$

$$\begin{aligned} p(\lambda|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\lambda)p(\lambda) \\ &\propto \left[\lambda^{\sum x_i} e^{-n\lambda}\right]\left[\lambda^{a-1} e^{-\lambda/b}\right] \\ &= \lambda^{\sum x_i + a - 1} e^{-\lambda(n+1/b)} \\ &= \lambda^{a^*-1} e^{-\lambda/b^*} \end{aligned}$$

$$\begin{aligned} p(\lambda|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\lambda)p(\lambda) \\ &\propto \left[\lambda^{\sum x_i} e^{-n\lambda}\right]\left[\lambda^{a-1} e^{-\lambda/b}\right] \\ &= \lambda^{\sum x_i + a - 1} e^{-\lambda(n+1/b)} \\ &= \lambda^{a^*-1} e^{-\lambda/b^*} \end{aligned}$$

4. Bayesian Testing

consider the following ratio of the posterior model probabilities

$$\frac{\pi(M_2|\boldsymbol{x})}{\pi(M_1|\boldsymbol{x})} = \frac{f(\boldsymbol{x}|M_2)}{f(\boldsymbol{x}|M_1)} \times \frac{\pi(M_2)}{\pi(M_1)}$$
$$= BF(M_2; M_1) \times \frac{\pi(M_2)}{\pi(M_1)}$$

Where $BF(M_2; M_1)$ is called the Bayes factor.

- Typically $\pi(M_2) = \pi(M_1)$, so the ratio of the posterior probabilities is the Bayes factor.
- The Bayes factor looks like a likelihood ratio. However, the difference is that θ has been integrated out in both the numerator and denominator, so we have the marginal distribution of the data given the model.
- If $f(\boldsymbol{x}|M_2) > f(\boldsymbol{x}|M_1)$ or $\frac{f(\boldsymbol{x}|M_2)}{f(\boldsymbol{x}|M_1)} > 1$ then we have support for M_2 against M_1.
- Jeffreys, H. (1961 appendix B) suggested the following:

$BF(M_2; M_1) = B_{21}$	Evidence against model 1 (H_0)
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
> 100	Decisive

5. Bayesian Interval Estimation

In Bayesian estimating, we have the whole distribution for $\pi(\theta|\boldsymbol{x}) = \frac{L(\theta|\boldsymbol{x})\pi(\theta)}{m(\boldsymbol{x})}$

- To obtain an interval we simply consider: $P\pi(\theta|x)(C) = \int_C \pi(\theta|\boldsymbol{x})d\theta = 1 - \alpha$

5.1 Choices of Reject Region (C)

- Equal tailed: $\int_{-\infty}^{\theta_L} \pi(\theta|\boldsymbol{x})d\theta = \alpha/2, \quad \int_{\theta_U}^{\infty} \pi(\theta|\boldsymbol{x})d\theta = \alpha/2$
- Smallest length: We can choose C to minimize $\theta_U - \theta_L$.
- Highest posterior density region (HPD): We define C to be that set with posterior probability $1 - \alpha$ which satisfies the criterion:

$$\theta_1 \in C \quad \text{and} \quad \pi(\theta_2|\boldsymbol{x}) > \pi(\theta_1|\boldsymbol{x}) \Rightarrow \theta_2 \in C$$

C contains the values of θ which have the highest posterior density values, so that we can determine HPD regions as the set:

$$C = \{\theta \in \Theta : \pi(\theta|\boldsymbol{x}) > c_\alpha\}$$

If the posterior is unimodal then this will be the smallest length interval!

6. ***Properties*** of Bayesian Inference

6.1 Sufficient

Definition 7.1: A statistic $T(\boldsymbol{X})$ is **sufficient** for $\boldsymbol{\theta}$ *if and only if* the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ is the same as the posterior distribution of $\boldsymbol{\theta}$ given $T(\boldsymbol{X})$.

Proof: Note that Definitions 2.5 and 7.1 are the same!

Suppose that $T(\boldsymbol{X})$ satisfies Definition 2.5. Then:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = g(\boldsymbol{x}|t,\boldsymbol{\theta})h(t|\boldsymbol{\theta}) = g(\boldsymbol{x}|t)h(t|\boldsymbol{\theta})$$

- The posterior is

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &\propto f(\boldsymbol{x};\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto h(t|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta}|t) \end{aligned}$$

- Now assume that $T(\boldsymbol{X})$ satisfies Definition 7.1.

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\theta}) &= \frac{p(\boldsymbol{\theta}|\boldsymbol{x})h(\boldsymbol{x})}{p(\boldsymbol{\theta})} \\ &= \frac{p(\boldsymbol{\theta}|t)h(\boldsymbol{x})}{p(\boldsymbol{\theta})} \\ &= K_1[t|\boldsymbol{\theta}]\ K_2[\boldsymbol{x}] \end{aligned}$$

- From the **factorization theorem**, it follows that $T(\boldsymbol{X})$ is a sufficient statistic.

6.2 Asymptotic (to Normal)

$$p(\theta|\boldsymbol{y}) \approx \text{normal}\left(\hat{\theta}, \left[I(\hat{\theta})\right]^{-1}\right)$$

(rough idea -> L10b/p33)

Decision Theory

1. Basic Elements

1. A number of 'actions' are possible; we must decide which to take.
2. A number of 'states of nature' are possible; we don't know in general which will occur.
3. The relative desirability of the various actions for each state of nature can be quantified.
4. Prior information may be available regarding the relative probabilities of the various states of nature.
5. Data may be available which will add to our knowledge of the relative probabilities of the states of nature.

2. General Process

- Let $\boldsymbol{\theta}$ denote the true state of nature.
- Suppose we are able to observe some data . . . a draw from the random variable $\boldsymbol{X}$, whose distribution depends on $\boldsymbol{\theta}$. Sometimes no data are available.
- A decision procedure δ, specifies which action to take for each value of $\boldsymbol{X}$.
- If we observe $\boldsymbol{X} = \boldsymbol{x}$, the we adopt the procedure $\delta(\boldsymbol{x})$.
- Whether $\delta(\boldsymbol{x})$ was a good choice depends on the loss function, which measures the loss from $\delta(\boldsymbol{x})$ when $\boldsymbol{\theta}$ holds.

$$L_S(\boldsymbol{\theta}, \delta(\boldsymbol{x}))$$

Note: The negative of a loss function is a utility function.

- Frequentists want to minimize expected loss.

$$min_\delta \int L_S(\boldsymbol{\theta}, \delta(\boldsymbol{x})) f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x}$$

- Bayesians want to minimize posterior expected loss.

$$min_\delta \int L_S(\boldsymbol{\theta}, \delta(\boldsymbol{x})) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}$$

3. Risk Function

$$R(\boldsymbol{\theta}, \delta(\boldsymbol{x})) = \int L_S(\boldsymbol{\theta}, \delta(\boldsymbol{x})) \;\; L(\boldsymbol{\theta}; \boldsymbol{x}) d\boldsymbol{x}.$$

This is the expected loss with respect to the joint distribution (i.e. likelihood).

4. Inadmissible

A procedure δ_1 is **inadmissible** if there exists another procedure δ_2 such that

$$R(\boldsymbol{\theta}, \delta_1) \;\geq\; R(\boldsymbol{\theta}, \delta_2) \;\; \forall \boldsymbol{\theta},$$

with strict inequality for some $\boldsymbol{\theta}$.

"δ1 is never better than δ2 and sometimes worse"

5. ***Minimax*** Procedure $\max_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \delta)$ is minimized.

6. ***Bayes*** Procedure $\int R(\boldsymbol{\theta}, \delta) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is minimized.

Expected risk with regard to the prior distribution minimized.

Extended

$$\begin{aligned}
\int R(\boldsymbol{\theta}, \delta) p(\boldsymbol{\theta}) d\boldsymbol{\theta} &= \int_{\Theta} \left[\int_{\mathcal{X}} L_S(\boldsymbol{\theta}, \delta)\ L(\boldsymbol{\theta}; \boldsymbol{x})\ d\boldsymbol{x} \right]\ p(\boldsymbol{\theta})\ d\boldsymbol{\theta} \\
&= \int_{\Theta} \int_{\mathcal{X}} L_S(\boldsymbol{\theta}, \delta)\ \frac{L(\boldsymbol{\theta}; \boldsymbol{x})\ p(\boldsymbol{\theta})}{p(\boldsymbol{x})} p(\boldsymbol{x})\ d\boldsymbol{x}\ d\boldsymbol{\theta} \\
&= \int_{\Theta} \int_{\mathcal{X}} L_S(\boldsymbol{\theta}, \delta)\ p(\boldsymbol{\theta}|\boldsymbol{x})\ p(\boldsymbol{x})\ d\boldsymbol{x}\ d\boldsymbol{\theta} \\
&= \int_{\mathcal{X}} p(\boldsymbol{x}) \left[\int_{\Theta} L_S(\boldsymbol{\theta}, \delta)\ p(\boldsymbol{\theta}|\boldsymbol{x})\ d\boldsymbol{\theta} \right]\ d\boldsymbol{x}
\end{aligned}$$

For any value of $\boldsymbol{x}$ we should minimize $\int_{\Theta} L_S(\boldsymbol{\theta}, \delta)\ p(\boldsymbol{\theta}|\boldsymbol{x})\ d\boldsymbol{\theta}$, which is **posterior expected loss**.

- Under certain conditions:
  For example, if $\boldsymbol{\theta}$ is discrete and and only takes a number of finite values then (2) holds. Additionally if $P(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}$ then (1) holds.
  1. A Bayes procedure in necessarily admissible.
  2. Every admissible procedure is a Bayes procedure for some prior distribution.
- The link between Bayes and minimax procedures:
  1. A Bayes procedure with constant risk for $\boldsymbol{\theta}$ is minimax.
  2. A minimax procedure is generally a Bayes procedure for some prior distribution. In particular, the so called **least favourable prior distibution**.

7. Point Estimation

1. The possible actions are possible estimators.
2. Possible states of nature correspond to the true value of θ.
3. Loss is some function of the estimator $\hat{\theta}$ and θ: $L_S(\theta, \delta = \hat{\theta})$.
4. Prior information is quantified by means of a prior distribution $p(\theta)$.
5. The available data, are our draws from $f(x|\theta)$.

- Some simple loss functions:

1. Zero-one loss: $L_S(\theta, \delta = \hat{\theta}) = \begin{cases} 0, & |\hat{\theta} - \theta| < b, \\ a, & |\hat{\theta} - \theta| \geq b \end{cases}$ where $a, b > 0$.
2. Absolute error loss: $L_S(\theta, \delta = \hat{\theta}) = a|\hat{\theta} - \theta|$
3. Squared Error Loss (quadratic loss): $L_S(\theta, \delta = \hat{\theta}) = a(\hat{\theta} - \theta)^2$

   To minimize, we set $\hat{\theta} = \bar{\theta}$.
   - Absolute error loss: $E[|\theta - \hat{\theta}|] = \int |\theta - \hat{\theta}| p(\theta|\boldsymbol{x}) d\boldsymbol{\theta}$.
   - This is minimized when $\hat{\theta} =$ median.

(James-Stein Estimator L11a/p32)

Non-Parametric Methods

1. Empirical Distribution Function

- The empirical distribution function $\hat{F}$ is the CDF of a new discrete random variable, say X^*.
- It can be shown that $\hat{F}$ is a sufficient statistic for F (based on a random sample), so
- $\hat{F}$ and X^* mimics the relationship between F and the X.
- This leads to studying $(\hat{F}, X^*)$ to learn about (F, X).

$$\begin{aligned}\hat{F}(x) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(x_i \le x)} \\ E\{\hat{F}(x)\} &= \frac{1}{n}\sum^{n} Pr(X_i \le x) = \frac{1}{n}\sum_{i=1}^{n} F(x) = F(x) \\ Var\{\hat{F}\} &= \frac{1}{n}F(x)(1-F(x))\end{aligned}$$

we are interested in $\theta(F) = E_F(X)$:

$$\hat{\theta} = \theta(\hat{F}) = E_{\hat{F}}(X) = \sum_{x\in\mathcal{X}} x p_{\hat{F}}(x) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

2. Bootstrap

- We create B "bootstrap" (re-sampled) data sets:

$$\{X^*_{1,1},\ldots,X^*_{n,1}\},\ldots,\{X^*_{1,b},\ldots,X^*_{n,b}\},\ldots,\{X^*_{1,B},\ldots,X^*_{n,B}\}$$

- With each sample we estimate $\hat{\theta}^b = \theta(\hat{F}^*_b)$,
- Since we know the "true" distribution $\hat{F}$, we can determine exactly the bias and variance of $\theta(\hat{F}^*)$.

$$\hat{B}_B = E_{\hat{F}}\{\theta(\hat{F}^*)\} - \theta(\hat{F}) \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b - \hat{\theta}$$

$$\hat{Var}_B\{\theta(\hat{F}^*)\} \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b\right)^2$$

2.1 Disadvantages

- As implemented, the bootstrap method yields a different answer every time (of course, the differences will be very small if B is large).
- Another drawback is that if θ is complicated to calculate (perhaps because it is implicitly defined as the solution to an equation, just as the MLE was) then computing its value for each of B re-sampled data sets is computationally quite expensive and time consuming.
- If we truly believe the parametric structure we have set up, then the parametric estimators have nice optimal properties.

Advantages

- The bootstrap is a very flexible and widely applicable approach which deserves more attention than it currently gets among statistical practitioners
- The bootstrap can even be extended to circumstances beyond the *iid* setting on which we have focused here.
- But a word of warning on complicated (non iid settings):
  - We cannot always guarantee that using the bootstrap paradigm (replacing F by $\hat{F}$ and $\hat{F}$ by $\hat{F}^*$) to estimate bias and variance will yield valid estimates.

2.2 Bootstrap Interval Estimation

2.21 Asymptotically

The estimated standard deviation was:

$$\hat{\sigma}_B(\hat{\theta}) = \sqrt{\frac{1}{B-1}(\hat{\theta}^*_b - \bar{\hat{\theta}}^*)^2}$$

We saw MLEs are asymptotically normal, and in fact many estimators are, we could just use that idea:

$$\left[\hat{\theta} - z_{\alpha/2}\hat{\sigma}_B(\hat{\theta}) \;\;,\;\; \hat{\theta} + z_{\alpha/2}\hat{\sigma}_B(\hat{\theta})\right]$$

2.2.2 Empirical

We can use the empirical quantiles of the "bootstrap distribution". So we have:

$$P_{\hat{F}}(\hat{\theta}^* \leq \hat{\theta}^*_L) = \alpha/2, \qquad P_{\hat{F}}(\hat{\theta}^* \leq \hat{\theta}^*_U) = 1 - \alpha/2$$

Using B re-samples, calculate $\hat{Q}_b$ for each re-sample and approximate q_l and q_U with (where $\alpha_1 + \alpha_2 = \alpha$) using the empirical distribution of $\hat{Q}$:

$$\hat{q}_l = \hat{Q}_{\alpha_1} \quad \hat{q}_U = \hat{Q}_{1-\alpha_2}$$

$$[\theta(\hat{F}) - \hat{q}_U\hat{\sigma}(F) \;,\; \theta(\hat{F}) + \hat{q}_L\hat{\sigma}(F)]$$

2.2.3 Properties of Intervals

- Shortest intervals for a given confidence or credibility (eg. 95%).
- Range respecting
- Parameterization equivariance (We would like our interval construction procedures to transform appropriately if we change our focus from $\tau = \tau(\theta)$ to $\gamma = \gamma(\tau) = \gamma\{\tau(\theta)\}$)

Non-Parametric Testing

- We will consider two types of tests:
  - permutation/randomization tests
  - bootstrap tests

(L11b)