Lectures 1-5

1 Introduction and Definitions

- The basic inference problem: Population, Sample, Probability model, Parameters.
- Goal is to infer aspects of population from information in sample.
- Types of inference: Estimation, Hypothesis testing
- Sample space Ω. $\mathbf{X} = (X_1, \cdots, X_n)$ be a random vector defined on the sample space. The outcome of the experiment is a realization $\mathbf{x} = (x_1, \cdots, x_n)$ of the random vector $\mathbf{X}$. We call $\mathbf{x}$ the data.
- Typical model:$\mathbf{X}$ has distribution $f(x_1, \cdots, x_n|\theta)$. This distribution is known except for the parameter θ. Given the data $\mathbf{x}$, the goal is to infer the unknown parameter θ.
- $\mathcal{F}$ represents the set of all possible probability distributions for $\mathbf{X}$. We'll call $\mathcal{F}$ the model(or probability model) for the experiment.
- Often the elements of $\mathcal{F}$ are indexed by one or more parameters. We'll often denote a vector of parameters by θ and let Θ be the collection of all possible values of θ. Θ is called the parameter space.
- If $\mathcal{F}$ can be expressed as a collection of distributions indexed by finite dimensional vectors $\Theta = (\theta_1, \cdots, \theta_k)$, where Θ is a subset of $\mathbb{R}^k$, then $\mathcal{F}$ will be called a parametric family. If $\mathcal{F}$ cannot be so expressed, it will be called nonparametric.
- Suppose $\theta = (\theta_1, \theta_2)$. If θ_1 is the only parameter of interest, then θ_2 is called a nuisance parameter.
- A model is said to be identifiable if $F_{\theta_1} = F_{\theta_2}$ whenever $\theta_1 = \theta_2$.
- Let T be a real-valued or vector-valued function whose domain contains the range of $\mathbf{X}$. If T does not depend on the unknown parameter θ, then $T = T(\mathbf{X})$ is called a statistic. The probability distribution of T is called its sampling distribution.

Example 1 *Have a population of N items, possibly a shipment of manufactured goods. An unknown number M of the N items are defective. A random sample of size n is drawn without replacement and inspected. Let X be the number of defectives in the sample.*

Example 2 *There are unknown number N number of fish in a pond. You catch M of them, tag them and let them go. Allow them to mingle for a while. Then you catch n fish and note the number of tagged ones among them. Let X be the number of tagged fish in the recaptured sample.*

Example 3 *Experimenter makes n independent determinations of the value of a physical constant μ and measurements are subject to error. $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$.*

Example 4 *Let $\mathcal{F}$= family of all continuous distributions that are symmetric about 0. Then $\mathcal{F}$ is a nonparametric family.*

2 Sufficiency for data reduction

[CB6.2, BD 1.5]

2.1 Sufficiency

Definition 1 *A statistic $T(\mathbf{X})$ is a sufficient statistic if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$ does not depend on θ, regardless of what t is.*

Example 5 *Suppose $X_1, \cdots, X_n$ are i.i.d. Poisson with mean θ. Then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a sufficient statistic.*

- Basic idea of sufficiency: Given data $\mathbf{X} = (X_1, \cdots, X_n)$, can we find a statistic $T(X)$ of smaller dimension than n that contains as much information about θ as $\mathbf{X}$ does? If a statistic exists, we can reduce (perhaps greatly) the amount of data without throwing away information. The search for good estimation and testing procedures can be narrowed.
- We can think of a partition of the sample space where each set A_t in the partition is such that $T(\mathbf{x}) = t$ for each $\mathbf{x} \in A_t$. All $\mathbf{x} \in A_t$ are equivalent in that each one contains the same information about θ as the others.
- If T_1 and T_2 are any two statistics such that $T_1(x) = T_1(y)$ if and only if $T_2(x) = T_2(y)$, then T_1 and T_2 are said to be equivalent.
- Sufficiency Principle: Consider sample $\mathbf{X}$ from model $\mathcal{F}$, and let $T(\mathbf{X})$ be a sufficient statistic. Suppose experimenter 1 observes $\mathbf{X} = x$ while experimenter 2 observes$\mathbf{X} = y$. If $T(x) = T(y)$, then experimenters 1 and 2 should make the same inference about θ.

Theorem 1 *(Fisher-Neyman Factorization Theorem): Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of the data $\mathbf{X}$. A statistic $T(\mathbf{X})$ is sufficient if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$(where h does not depend on θ) such that $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$ for all $\mathbf{x}$ and all parameter values θ.*

Example 6 *Estimating the Size of a Population: Consider a population with N members labeled consecutively from I to N. The population is sampled with replacement and n members of the population are observed and their labels $X_l, \cdots, X_n$ are recorded. Then $X_{(n)}$ is indeed sufficient.*

Example 3 (revisited): $X_1, \cdots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)$.
$T(X) = (\sum X_i, \sum X_i^2)$ is jointly sufficient for θ.
Qn: Is the dimension of a sufficient statistic the always same to the dimension of the parameters?
HW: Eg 1.5.5 of BD: Linear Regression
Let $f_X(x|\theta)$ be the joint pdf or pmf of $\mathbf{X}$ and $q(t|\theta)$ be the pdf or pmf of $T(\mathbf{X})$. Then T is a sufficient statistic for θ, iff, for every x, the ratio $f_X(x|\theta)/q(T(x)|\theta)$ is constant as a function of θ.

Example 7 *Suppose we observe* $\mathbf{X} = (X_1, \cdots, X_n)$, *where*

$$X_i = \rho X_{i-1} + Z_i, \quad i = 2, 3, \cdots, n.$$

The quantity ρ *is an unknown parameter such that* $|\rho| < 1$. $Z_2, \cdots, Z_n$ *are i.i.d.* $\mathcal{N}(0, \sigma^2)$, *where* σ^2 *is another unknown parameter.* $X_1 \sim \mathcal{N}(0, \sigma^2/(1-\rho^2))$ *and* $X_1, Z_2, \cdots, Z_n$ *are mutually independent.*
The parameter space is $\Theta = (\rho, \sigma^2) : |\rho| < 1, \sigma^2 > 0$.
This model is called an autoregressive model and is used in time series analysis.

$$f(\mathbf{x}|\rho, \sigma) = (2\pi\sigma^2)^{-n/2}\sqrt{1-\rho^2}e^{\left\{-\frac{1}{2\sigma^2}(x_1^2(1-\rho^2)+\sum_{i=2}^n (x_i - \rho x_{i-1})^2)\right\}}.$$

$T_1(\mathbf{X}) = \sum_{i=2}^{n-1} X_i^2, T_2(\mathbf{X}) = \sum_{i=2}^n X_i X_{i-1}$ *and* $T_3(\mathbf{X}) = X_1^2 + X_n^2$ *are jointly sufficient statistics.*

Proposition 1 *Let* $T(\mathbf{X}) = (T_1(\mathbf{X}), \cdots, T_k(\mathbf{X}))$ *be a sufficient statistic and* r *be a 1-1 function, not depending on* θ *and with domain equal to the range of* $T(\mathbf{X})$. *Then* $r(T(\mathbf{X}))$ *is a sufficient statistic.*

2.2 Minimal Sufficiency

Definition 2 *: A statistic* $T(\mathbf{X})$ *is a minimal sufficient statistic if it is a function of every other sufficient statistic.*

Theorem 2 *Let* $f(\mathbf{x}|\theta)$ *be the pdf or pmf of* $\mathbf{X}$. *Suppose there exists a statistic* $T(\mathbf{X})$ *such that, for any two points* $\mathbf{x}$ *and* $\mathbf{y}$, *the ratio* $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ *is constant as a function of* θ *iff* $T(\mathbf{x}) = T(\mathbf{y})$. *Then* $T(\mathbf{X})$ *is a minimal sufficient statistic.*

Example 8 $X_1, \cdots, X_n$ *iid Unif*$(\theta, \theta + 1)$. *Then* $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ *is minimal sufficient.*

Proposition 2 *If* $T(X)$ *is a minimal sufficient statistic for* θ, *then its one-to-one function is also a minimal sufficient statistic for* θ.

Proposition 3 *There is always a one-to-one function between any two minimal sufficient statistics.*

Example 3 (revisited): $T_1(\mathbf{X})$=$(\bar{X}, S^2)$ is minimal sufficient.
HW: For $X_1, \cdots, X_n$ iid from cauchy distn, show that the minimal sufficient statistics is the order statistics. Does the order-statistics provide any data reduction?

2.3 Ancillarity

Definition 3 *A statistic $S(\mathbf{X})$ is an ancillary statistic if its distribution does not depend on θ.*

Example 8 continued: The range is ancillary.

Definition 4 *Let $f(x)$ be any pdf. Then for any $\mu \in \mathbb{R}$ and any $\sigma > 0$ the family of pdfs $g(x) = f((x-\mu)/\sigma)/\sigma$, indexed by the parameter (μ, σ) is called the location-scale family with standard pdf $f(x)$, and μ is called the location parameter and σ is called the scale parameter for the family.*

HW: In the above definition g is indeed a pdf.
HW: X is a random variable with pdf f if and only if there exists a random variable Z with pdf g and $X = \sigma Z + \mu$.
HW: Let $X_1, \cdots, X_n$ be iid from a location family. Show that the range is an ancillary statistic. Can you think of another ancillary statistic?
HW: Let $X_1, \cdots, X_n$ be iid from a scale family. Show that the following statistic $T(\mathbf{X})$ is ancillary. $T(\mathbf{X}) = (X_1/X_n, \cdots, X_{n-1}/X_n)$.

2.4 Completeness

Definition 5 *Let $f_T(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called complete if $E[g(T)|\theta] = 0$ for all θ implies $Pr[g(T) = 0|\theta] = 1$ for all θ. equivalently T is a complete statistic.*

Example 5 revisited: In the Poisson eg, restrict $\Theta = \{1, 2\}$. Then $g(0) = 2, g(2) = 2, g(1) = -2$ and 0 otherwise is a function that has expectation zero for all θ. Thus the family is not complete. When $\Theta = \mathbb{R}^+$, then the family is complete.

Proposition 4 *For a statistic T(X), if a non-constant function of T, say r(T) is ancillary, then T(X) cannot be complete.*

Proposition 5 *If T(X) is a complete statistic, then a function of T, say $T^* = r(T)$ is also complete.*

Proposition 6 *If a complete sufficient statistic exists, then a minimal sufficient statistic is complete.*

Theorem 3 *(Basu 1955) If T(X) is complete and minimal sufficient statistic, then T(X) is independent of every ancillary statistic.*

HW: For exponential distribution, find $E(X_1/(X_1 + \cdots + X_n))$

Lectures 6-9

3 Exponential Family

[CB3.4, BD1.6]
Binomial and normal distributions have the property that the dimension of a sufficient statistic is independent of the sample size. We would like to identify and define a broad class of models that have this and other desirable properties.

Definition 1 *Let $\{f(x;\theta) : \theta \in \Theta\}$ be a family of pdf's (or pmf's). We assume that the set $\{\mathbf{x} : f(\mathbf{x};\theta) > 0\}$ is independent of θ, where $\mathbf{x} = (x_1, \cdots, x_n)$. We say that the family $\{f(x;\theta) : \theta \in \Theta\}$ is a k–parameter exponential family if there exist real–valued functions $Q_1(\theta), \cdots, Q_k(\theta)$ and $D(\theta)$ on Θ and $T_1(\mathbf{X}), \cdots, T_k(\mathbf{X})$ and $S(\mathbf{X})$ on $\mathbb{R}^n$ such that*

$$f(x;\theta) = exp(\sum_{i=1}^{k} Q_i(\theta)T_i(\mathbf{x}) + D(\theta) + S(\mathbf{x})).$$

We can express the k–parameter exponential family in canonical form for a natural $kx1$ parameter vector $\eta = (\eta_1, \cdots, \eta_k)'$ as

$$f(\mathbf{x};\eta) = h(\mathbf{x})c(\eta)exp(\sum_{i=1}^{k} \eta_i T_i(\mathbf{x})),$$

We define the natural parameter space as the set of points $\eta \in W \subset \mathbb{R}^k$ for which the integral $\int_{\mathbb{R}^n} exp(\sum_{i=1}^{k} \eta_i T_i(\mathbf{x}))h(\mathbf{x})d\mathbf{x}$ is finite.
We shall refer to T as a natural sufficient statistic.
Ex: Verify that Binomial and Normal belong to exponential family.
Uniform distribution U([0, θ]), $\theta \in \mathbb{R}^+$ does not belong to the exponential family, since its support depends on θ
If the probability distribution of X_1 belongs to an exponential family, the probability distribution of $(X_1, \cdots, X_n)$ also belongs to the same exponential family, where X_i are iid with distribution same as X_1.

Theorem 1 *Suppose $X_1, \cdots, X_n$ is a random sample from pdf or pmf $f_X(x|\theta)$ where $f_X(x|\theta) = h(x)d(\theta)exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$ is a member of an exponential family. Define a statistic $T(X)$ by $T(X) = (\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j))$. The distribution of $T(X)$ is an exponential family of the form $f_T(u_1, \cdots, u_k|\theta) = H(u_1, \cdots, u_k)[d(\theta)]^n \exp(\sum_{i=1}^{k} w_i(\theta)u_i)$*

Theorem 2 (3.4.2 of CB) *If X is a random variable with pdf/pmf as in definition 1 then, for every j,*

$$\mathrm{E}(\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(\mathbf{X})) = -\frac{\partial}{\partial \theta_j} D(\theta)$$

$$\mathrm{Var}(\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(\mathbf{X})) = -\frac{\partial^2}{\partial {\theta_j}^2} D(\theta) - \mathrm{E}(\sum_{i=1}^{k} \frac{\partial^2 w_i(\theta)}{\partial {\theta_j}^2} t_i(\mathbf{X}))$$

Ex: Use this to derive the mean and variance of the binomial and normal distributions.

Theorem 3 *If the distribution of X belongs to a canonical exponential family and η is an interior point of W, the mgf of T exists and is given by*

$$M(s) = c(\eta)/c(s+\eta)]$$

for s in some neighbourhood of 0.

Ex: Use this to derive the mean and variance of the natural sufficient statistic of Raleigh distribution

$$p(x,\theta) = (x/\theta^2) exp(-x^2/2\theta^2), x > 0, \theta > 0.$$

In an exponential family, if the dimension of Θ is k (there is an open set subset of $\mathbb{R}^k$ that is contained in Θ), then the family is a full exponential family.
Otherwise the family is a curved exponential family.
An example of a full exponential family is $\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma > 0$.

Example 1 *An example of a curved exponential family is $\mathcal{N}(\mu, \mu^2), \mu \in \mathbb{R}$.*

Curved exponential families arise naturally in applications of CLT as approximation to binomial $\sigma^2 = p(1-p)/n$ or Poisson $\sigma^2 = \lambda/n$.

Theorem 4 *In the exponential family given by definition 1 and the set Θ contains an open subset of $\mathbb{R}^k$ then $(T_1(\mathbf{X}), \cdots, T_k(\mathbf{X}))$ is complete.*

Ex: In the curved exponential family of example 1, $k = 2$ and the set $Theta$ does not contain an open subset of $\mathbb{R}^2$. So we cannot apply the above theorem. Is it still true that $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is complete?
Ex: Show that the Cauchy family is not an exponential family.
Ex: Multinomial is a $(k-1)$ parameter exponential family.
Ex: Linear Regression model is 3 parameter exponential family.
Ex: Logistic regression model is 2-parameter exponential family.

Definition 2 *An exponential family is of rank k iff the natural sufficient statistic T is k-dimensional and $(1, T_1(X), \cdots, T_k(X))$ are linearly independent with positive probability. Formally, $P[\sum_{j=1}^k a_j T_j(X) = a_{k+1}] < 1$ unless all a_j are 0.*

Ex: multinomial is rank $k-1$.
Ex: Logistic with n=1 is rank 1 and θ_1 and θ_2 are not identifiable. For $n \geq 2$, the rank is 2.
The following theorem establishes the relation between rank and identifiability.

Theorem 5 *Suppose $\mathcal{P} = q(x, \eta); \eta \in W$ is a canonical exponential family generated by (T_{kxl}, h) with natural parameter space W such that W is open. Let $A(\eta) = -log(c(\eta))$. Then the following are equivalent.*

1. *$\mathcal{P}$ is of rank k.*
2. *η is a parameter (identifiable).*
3. *Var(T) is positive definite.*
4. *$\eta \to \dot{A}(\eta)$ is 1-1 on E*
5. *A is strictly convex in E.*

Ex: Multivariate normal. Show that this family is full rank and E is open.

Lectures 10-17

4 Methods of Point Estimation

[CB7.2, BD2]
Point estimate: Any function of the data. That is, any statistic.
Estimate vs estimator.

4.1 Method of Moments

Definition 1 *Let $X_1, \cdots, X_n$ be iid with pdf (or pmf) $f_\theta, \theta \in \Theta$. We assume that first k moments $m_1, \cdots, m_k$ of f_θ exist. If θ can be written as $\theta = h(m_1, ..., m_k)$, the method of moments estimate of θ is*

$$\hat{\theta}_{MOM} = T(X_1, ..., X_n) = h(\sum_{i=1}^{n} X_i, \cdots, \sum_{i=1}^{n} X_i^k)$$

Note:

- The Definition above can also be used to estimate joint moments. For example, we use $\sum_{i=1}^{n} X_i Y_i$ to estimate $\mathrm{E}(XY)$.
- If θ is not a linear function of the population moments, $\hat{\theta}_{MOM}$ will, in general, not be unbiased. However, it will be consistent and (usually) asymptotically Normal.
- Method of moments estimates do not exist if the related moments do not exist.
- Method of moments estimates may not be unique. If there exist multiple choices for $\hat{\theta}_{MOM}$, one usually takes the estimate involving the lowest–order sample moment.

eg. Normal
eg. Binomial with both n and p unknown.

Example 1 *$X_1, \cdots, X_n$ iid Gamma(p, λ). The first two moments of the gamma distribution are $E(X) = p/\lambda$ and $E(X^2) = p(p+1)/\lambda^2$. Use this to obtain the MOM estimator.*

Example 2 *(Different MoM estimators)Example: $X_1, \cdots, X_n$ iid Poisson(λ). The first moment is λ . Thus, the method of moments estimator based on the first moment is $\bar{X}$. We could also consider using the second moment to form a method of moments estimator.. The method of moments estimator based on the second moment solves $\bar{X^2} = \lambda + \lambda^2$ Solving this equation (by taking the positive root), we find that $\hat{\lambda} = -1/2 + (1/4 + \bar{X^2})^{1/2}$. The two method of moments*

estimators are different. For example, for the data
rpois(10,1) 2 3 0 1 2 1 3 1 2 1,
the method of moments estimator based on the first moment is 1.1 and the method of moments estimator based on the second moment is 1.096872. We choose rthtet one based on the lower moment.

Example 3 *(Hardy-Weinberg proportions) Consider (first generation of) a population in which the alleles A and a are encountered with probabilities θ and 1-θ respectively, $\theta \in (0,1)$. If the alleles are chosen at random and independently for each individual in the next generation, then the probability of having the AA genotype is θ^2, the aa genotype is $(1-\theta)^2$ and Aa genotype $2\theta(1-\theta)$. Suppose we sample n individuals from the population, observe their genotypes and would like to estimate the probability (proportion) of A allele in the population. The corresponding statistical model is an i.i.d. sample $X_1, \cdots, X_n$, where X_i takes values in AA, Aa,AA with probabilities $\theta^2, 2\theta(1-\theta) and (1-\theta)^2$ respectively. Note that $E_\theta N_{AA} = \theta^2$ and $E_\theta N_{aa} = (1-\theta)^2$. Also, $E(N_{AA} + 1/2N_{aA} = \theta$. Each of these can be used to find a method of moments estimator for θ.*

Example 4 *(The method of moments does not use all the information that is available.) $X_1, \cdots, X_n$ iid Uniform(0,θ). The method of moments estimator based on the first moment is $\hat{\theta} = 2\bar{X}$. If $2\bar{X} < max(X_i)$, we know that $\theta > \hat{\theta}$.*

Definition 2 *Suppose we are given a function $\Psi : \Omega \times \mathbb{R}^d \to \mathbb{R}^d$ and define*

$$V(\theta_0, \theta) = E_{\theta_0} \Psi(X, \theta)$$

Suppose $V(\theta_0, \theta) = 0$ has θ_0 as its unique solution for all $\theta_0 \in \Theta$. Then we say $\hat{\theta}$ solving

$$\Xi(X, \hat{\theta}) = 0$$

is an estimating equation estimate.

eg: Take $\Xi = (\hat{\mu}_1 - \mu_1, \cdots, \hat{\mu}_d - \mu_d)$ to get the method of moments estimator.
eg: Least squares as estimating equation.

Definition 3 *Consider a parameter that can be written as a function of F, i.e., $\theta = T(F)$. The plug-in estimator of θ is $T(\hat{F_n}$ where F_n is the empirical cdf.*

For parametric models plug-in estimators are not generally optimal. But they are good starting points for numerical algorithms.
Sample median is a plug-in estimator of population median $theta = F^{-1}(1/2)$, but not a MoM estimator.

4.2 Maximum likelihood estimation

Definition 4 *Let $(X_1, \cdots, X_n)$ be a random vector with pdf (or pmf) $f(x_1, \cdots, x_n; \theta), \theta \in \Theta$. We call the function $L(\theta; x_1, \cdots, x_n) = f(x_1, \cdots, x_n; \theta)$ of θ the likelihood function.*

Definition 5 *A maximum likelihood estimate (MLE) is an estimate $\hat{\theta}_{ML}$ such that*

$$L(\hat{\theta}_{ML}; x_1, \cdots, x_n) = sup_{\theta \in \Theta} L(\theta; x_1, \cdots, x_n).$$

Note: It is often convenient to work with log L when determining the maximum likelihood estimate. Since the log is monotone, the maximum is the same.
Use the derivative to find potential MLE. Use the double derivative to confirm local maximum. Check boundary and confirm global maximum.
If the function is NOT differentiable with respect to θ. Use numerical methods. Or perform directly maximization, using inequalities, or properties of the function.
For multivariate θ, second derivative test for maxima entails checking that the Hessian matrix (matrix of second derivatives) is negative definite. Sign of determinant of any principal minor is $(-1)^r$, where r is the order.
Eg: X_1, X_2, X_3, X_4 i.i.d. Bernoulli(p), $0 < p < 1$. Plot the likelihood function for x = (1, 1, 1, 1), x=(0,0,0,0) and x=(1,1,0,0).
Eg (bivariate parameter) Normal.
Eg (non-unique) Unif($\theta - 1/2, \theta + 1/2$)
Eg Unif(0, θ).
Eg Ber(p), with $\Theta = (1/2, 3/4)$. Here MLE is worse in the sense of MSE to $\hat{p} = 1/2$. Eg: MLE of θ in the Hardy Weinberg set-up.
Eg: MLE(discrete parameter space) Hypergeometric. Total 12, marked θ, pick 5. If $X = 3$ then $\hat{\theta} = 7$.
In this case, MoM estimator does not exist. 12*3/5=7.2 is not in parameter space.
The likelihood function is not a probability mass function or a probability density function: in general, it is not true that L integrates to 1 with respect to θ .
The MLE is the parameter point for which the observed sample is most likely.

Theorem 1 *Let T be a sufficient statistic for $f_\theta, \theta \in \Theta$. If MLE of θ exists, it is a function of T.*

Proof: Since T is sufficient, we can write

$$f(x, \theta) = h(x) g(T(x), \theta)$$

due to the Factorization Criterion. Maximizing the likelihood function with respect to θ takes $h(x)$ as a constant and therefore is equivalent to maximizing $g(T(x), \theta)$ with respect to θ. But $g(T(x), \theta)$ involves x only through T.

- MLE may not exist.
- MLE may not be unique.
- Computation may be difficult.

Theorem 2 *(Invariance of MLE) Let $\{f_\theta : \theta \in \Theta\}$ be a family of pdf's (or pmf's) with $\Theta \subseteq R^k, k \geq 1$. Let $h : \Theta \to \Delta$ be a mapping of Θ onto $\Delta \subseteq, R^p 1 \leq p \leq k$. If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is an MLE of $h(\theta)$.*

Proof: For each $\delta \in \Delta$, we define $\Theta_\delta = \{\theta : \theta \in \Theta, h(\theta) = \delta\}$
and $M(\delta; x) = sup_{\theta \in \Theta_\delta} L(\theta; x)$, the likelihood function induced by h.
Let $\hat{\theta}$ be an MLE let and $\hat{\delta} = h(\hat{\theta})$.
It holds $M(\hat{\delta}; x) = sup_{\theta \in \Theta_{\hat{\delta}}} L(\theta; x) \geq L(\hat{\theta}, x)$ since $\hat{\theta} \in \Theta_{\hat{\delta}}$
But also $M(\hat{\delta}; x) \leq sup_{\delta \in \Delta} M(\delta; x) = sup_{\delta \in \Delta}(sup_{\theta \in \Theta_\delta} L(\theta; x)) = sup_{\theta \in \Theta} L(\theta; x) = L(\hat{\theta}; x)$.
Therefore, $M(\hat{\delta}; x) = L(\hat{\theta}; x) = sup_{\delta \in \Delta} M(\delta; x)$.
Thus, $\hat{\delta} = h(\hat{\theta})$ is an MLE.
eg Let $X_1, ..., X_n$ be iid Ber(p). Let $h(p) = p(1-p)$. Since the MLE of p is $X = \sum(X_i)$, the MLE of h(p) is $X(1-X)$.

4.3 Bayesian methods

Model: $X_, \cdots, X_n \sim f(\mathbf{X}|\theta)$.
In the frequentist approach, θ is a fixed unknown constant.
In the Bayesian approach, we put a prior probability distribution on θ, say $\pi(\theta)$.
The model is then the conditional distribution of the data given a value of θ.
The joint distribution is, therefor the product of the prior and the model.
We use Bayes Rule to obtain the conditional distribution of θ given the data.
This is called the posterior distribution and is given below:

$$f(\theta|\mathbf{X}) = \frac{\text{joint}}{\text{marginalof}\mathbf{X}} = \frac{\pi(\theta) f(\mathbf{X}|\theta)}{\int_{\eta \in \Theta} \pi(\eta) f(\mathbf{X}|\eta) d\eta}.$$

The Bayes estimator is the conditional expectation of θ given the data, that is, the expectation of the posterior distribution and is given by:

$$E(\theta|\mathbf{X}) = \int_{\eta \in \Theta} \eta f(\eta|\mathbf{X}) d\eta = \frac{\int_{\eta \in \Theta} \eta \pi(\eta) f(\mathbf{X}|\eta) d\eta}{\int_{\eta \in \Theta} \pi(\eta) f(\mathbf{X}|\eta) d\eta}.$$

Eg: $X_i \sim iid\mathcal{N}(\theta, 1), \theta \sim \mathcal{N}(0, \sigma^2)$.
Eg Bernoulli with beta(r,r) prior

Definition 6 *A family of prior probability distributions π is said to be conjugate to a family of likelihood functions L(x; θ) if the resulting posterior distributions are in the same family as prior; the prior is called a conjugate prior for the likelihood.*

eg Poisson with Gamma prior as conjugate.

5 Numerical methods for finding MLE's

5.1 Bisection

The bisection method is a method for finding the root of a one-dimensional function that is continuous on $\mathbb{R}$, for which f is monotone increasing or decreasing. It can be used when the likelihood equation is (or can be reduced to)

a one-parameter equation. The bisection method works by repeatedly dividing an interval in half and then selecting the subinterval in which the root exists.

5.2 Coordinate ascent

The coordinate ascent method is an approach to finding the maximum likelihood estimate in a multidimensional family. The coordinate ascent method works by using the bisection method iteratively. Suppose we have a k-dimensional parameter $(\theta_1, \cdots, \theta_k)$. The coordinate ascent method is: Choose an initial estimate $(\hat{\theta}_1, \cdots, \hat{\theta}_k)$.

1. Set $(\hat{\theta}_1, \cdots, \hat{\theta}_k)_{old}$=$(\hat{\theta}_1, \cdots, \hat{\theta}_k)$
2. Maximize $l(\theta_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$ over θ_1 using the bisection method. Reset θ_1 to the value that maximizes the likelihood as $\hat{\theta}_1$.
3. Maximize l over θ_2 using the bisection method. Reset $\hat{\theta}_2$.
4. continue to θ_K
5. Stop if the distance between $(\hat{\theta}_1, \cdots, \hat{\theta}_k)_{old}$ and $(\hat{\theta}_1, \cdots, \hat{\theta}_k)$ is less than some tolerance ϵ. Otherwise return to step 1.

The coordinate ascent method converges to the maximum likelihood estimate when the log likelihood function is strictly concave on the parameter space. See Figure 2.4.1 in Bickel and Doksum.

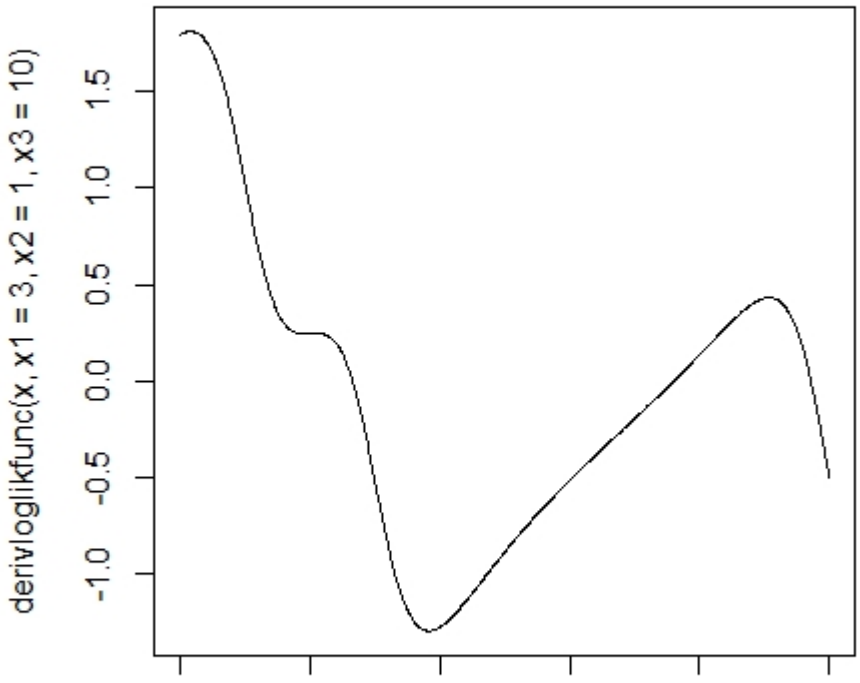
Example (Beta Distribution) This is a two parameter full rank exponential family and hence the log likelihood is strictly concave. We found the method of moments estimates and use them as initial estimates. $\hat{r} = \bar{x}(\bar{x} - \bar{x^2})/(\bar{x^2} - \bar{x}^2)$ $\hat{s} = (1 - \bar{x})(\bar{x} - \bar{x^2})/(\bar{x^2} - \bar{x}^2)$

```
R code for finding the MLE:
# Code for beta distribution MLE
# xvec stores the data
# rhatcurr, shatcurr store current estimates of r and s
# Generate data from Beta(r=2,s=3) distribution)
xvec=rbeta(20,2,3);
#xvec = (0.3184108, 0.3875947, 0.7411803, 0.4044642,  0.7240628, 0.7247060, 0.1091041, 0.138
# Set low and high starting values for the bisection searches
rhatlow=.001;
rhathigh=20;
shatlow=.001;
shathigh=20;
# Use method of moments for starting values
rhatcurr=mean(xvec)*(mean(xvec)-mean(xvec^2))/(mean(xvec^2)-mean(xvec)^2);
shatcurr=((1-mean(xvec))*(mean(xvec)-mean(xvec^2)))/(mean(xvec^2)-mean(xvec)^2);
#rhatcurr=2.239774
#shatcurr=2.893378
```

```
rhatiters=rhatcurr;
shatiters=shatcurr;
derivrfunc=function(r,s,xvec){
  n=length(xvec);
  sum(log(xvec))-n*digamma(r)+n*digamma(r+s);
}
derivsfunc=function(s,r,xvec){
  n=length(xvec);
  sum(log(1-xvec))-n*digamma(s)+n*digamma(r+s);
}
dist=1;
cc=1;
toler=.0001;
while(dist>toler){
  rhatnew=uniroot(derivrfunc,c(rhatlow,rhathigh),s=shatcurr,xvec=xvec)$root;
  shatnew=uniroot(derivsfunc,c(shatlow,shathigh),r=rhatnew,xvec=xvec)$root;
  dist=sqrt((rhatnew-rhatcurr)^2+(shatnew-shatcurr)^2);
  rhatcurr=rhatnew;
  shatcurr=shatnew;
  rhatiters=c(rhatiters,rhatcurr);
  shatiters=c(shatiters,shatcurr);
cc=cc+1}
rhatmle=rhatcurr;
shatmle=shatcurr;
#rhatmle=2.401314
#shatmle=3.117656
#cc=21
```

Example of nonconcave likelihood: Cauchy model. Log likelihood is not concave and has two local maxima between 0 and 10. There is also a local minimum. The local maximum (i.e., the solution to the likelihood equation) that the bisection method finds depends on the interval searched over.

```
R program to use bisection method
derivloglikfunc=function(theta,x1,x2,x3){
  dloglikx1=2*(x1-theta)/(1+(x1-theta)^2);
  dloglikx2=2*(x2-theta)/(1+(x2-theta)^2);
  dloglikx3=2*(x3-theta)/(1+(x3-theta)^2);
  dloglikx1+dloglikx2+dloglikx3;
}
plot(x,derivloglikfunc(x,x1=3,x2=1,x3=10),type="l")
uniroot(derivloglikfunc,interval=c(0,5),x1=3,x2=1,x3=10);
#$root=2.653812
uniroot(derivloglikfunc,interval=c(0,10),x1=3,x2=1,x3=10);
#$root=9.721143
```



5.3 Newton's Method

Newton's method is a numerical method for approximating solutions to equations. The method produces a sequence of values that, under ideal conditions, converges to the MLE . To motivate the method, we expand the derivative of the log likelihood around $\hat{\theta}_{MLE}$: $0 = l'(\hat{\theta}_{MLE}) \approx l'(\theta^{(j)}) + (\hat{\theta}_{MLE} - \theta^{(j)})l''(\theta^{(j)})$ Solving for $\hat{\theta}_{MLE}$ gives $\hat{\theta}_{MLE} = \theta^{(j)} - l'(\theta^{(j)})/l''(\theta^{(j)})$ This suggests the following iterative scheme: $\theta^{(j+1)} = \theta^{(j)} - l'(\theta^{(j)})/l''(\theta^{(j)})$ Newton's method can be extended to more than one dimension (usually called Newton-Raphson) $\theta^{(j+1)} = \theta^{(j)} - \ddot{l}^{-1}(\theta^{(j)})/\dot{l}(\theta^{(j)})$ where $\dot{l}$ denotes the gradient vector of the likelihood and $\ddot{l}$ denotes the Hessian.

Comments on methods for finding the MLE:

1. The bisection method is guaranteed to converge if there is a unique root in the interval being searched over but is slower than Newton's method.
2. Newton's method does not work if $l''(\theta^{(j)}) \approx 0$
3. Newton's method does not always converge.
4. For the coordinate ascent method and Newton's method, a good choice of starting values is often the method of moments estimator or plug-in estimator.

5. When there are multiple roots to the likelihood equation, the solution found by the bisection method, the coordinate ascent method and Newton's method depends on the starting value. These algorithms might converge to a local maximum (or a saddlepoint) rather than a global maximum.

5.4 EM Algorithm

Complete data, incomplete data.
E step: Expectation of complete data log likelihood given incomplete data.
M step: Maximize
Iterate.
This is a famous example from Rao (1973)[Linear Statistical Inference and Its Applications]. We consider the genetic linkage of 197 animals, in which the phenotypes are distributed into 4 categories: Y = (y1, y2, y3, y4) = (125, 18, 20, 34) with cell probabilities $(1/2 + \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$.
Though it is by no means impossible to maximize this multinomial likelihood directly, we illustrate how the EM algorithm brings a substantial simplification, by using the augmentation method. Specifically, we augment the observed data Y by dividing the first cell into two, with respective cell probabilities 1/2 and $\theta/4$. This gives an augmented data set X = (x1, x2, x3, x4, x5), where x1 + x2 = y1, and x3 = y2, x4 = y3, x5 = y4.
E-step: $E(l) = (E(X2) + x5)log(\theta) + (x3 + x4)log(1-\theta)$.
$X_2 \sim Bin(y1, \theta/(\theta+2))$
M-step: $\theta_{n+1} = (159\theta_n + 68)/(197\theta_n + 144)$
The alternation between estimation and maximization is clearly seen in this iteration formula. Starting with $\theta_0 = 0.5$ we obtain the sequence as follows 0.6082, 0.6243, 0.6265, 0.6268, 0.6268.

Lectures 18-21

4 Criteria for estimators

[CB7.3, BD3.4]

Definition 1 *The bias of an estimate $T(X)$ of a parameter $q(\theta)$ in a model (non-empty set of pdf/pmf) mathcalP $= P_\theta : \theta \in \Theta$ as $Bias_\theta(T) = E_\theta(T(X)) - q(\theta)$. An estimate such that $Bias_\theta(T) = 0$ is called unbiased. Any function $q(\theta)$ for which an unbiased estimate T exists is called an estimable function.*

This notion has intuitive appeal, ruling out, for instance, estimates that ignore the data, such as $T(X) = q(\theta_0)$, which can't be beat for $\theta = \theta_0$ but can obviously be arbitrarily terrible.
Eg: $\bar{X}$ and s^2 in normal distribution are unbiased for μ and σ^2. However, note that S is not an unbiased estimate of σ. Eg: (Unbiased estimates may be absurd) Let $X \sim Poisson(\lambda)$ and let $q(\lambda) = e^{-2\lambda}$. Consider $T(X) = (-1)^X$ as an estimate. It is unbiased but since T alternates between -1 and 1 while $q(\lambda) > 0$, it is not a good estimate.
Eg: (Unbiased Estimates in Survey Sampling) Suppose we wish to sample from a finite population, for instance, a census unit, to determine the average value of a variable (say) monthly family income during a time between two censuses and suppose that we have available a list of families in the unit with family incomes at the last census. Write $x_l, \cdots, x_N$ for the unknown current family incomes and correspondingly $u_1, \cdots, u_N$ for the known last census incomes. We ignore difficulties such as families moving. We let $X_l, \cdots, X_n$ denote the incomes of a sample of n families drawn at random without replacement. The parameter of interest is $\frac{1}{N}\sum_{i=1}^N x_i$. The model is

$$P(X_1 = a_1, \cdots X_n = a_n) = \binom{N}{n}^{-1} \quad \text{if} \quad \{a_1, \cdots, a_n\} \subseteq \{x_1, \cdots, x_n\}$$

Ex: $\bar{X}$ is unbiased and has variance $\frac{\sigma^2}{n}(1 - \frac{(n-1)}{(N-1)})$ where $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^2$. This method of sampling does not use the information contained in $u_1, \cdots, u_N$. One way to do this, reflecting the probable correlation between $(u_1, \cdots, u_N)$ and $(x_l, \cdots, x_N)$, is to estimate by a regression estimate

$$\bar{X}_R = \bar{X} - b(\bar{(U)} - \bar{u})$$

Ex: For each b this is unbiased.
Ex: If the correlation between U_i and X_i is positive (population) and $b < 2Cov(\bar{U}, \bar{X})/Var(\bar{U})$, this is better than $\bar{X}$.
Ex: The optimal choice of b is $Cov(\bar{U}, \bar{X})/Var(\bar{U})$. This value is unknown and can be estimated by

$$b_{opt} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(U_i - \bar{U})}{\frac{1}{N}\sum_{i=1}^N (u_i - \bar{u})^2}$$

. Ex: This estimator is biased.

4.1 Uniform Minimum Variance Unbiased (UMVU)

Note: If there exist 2 unbiased estimates T_1 and T_2 of θ, then any estimate of the form $\alpha T_1 + (1-\alpha)T_2$ for $0 \leq \alpha \leq 1$ will also be an unbiased estimate of θ. Which one should we choose?
For unbiased estimates mean square error and variance coincide.

Definition 2 *An unbiased estimate $T*(X)$ of $q(\theta)$ that has minimum MSE among all unbiased estimates for all θ is called UMVU (uniformly minimum variance unbiased). If this happens for a single parameter value θ_0 then it is locally minimum variance unbiased.*

Theorem 1 *Let U be the class of all unbiased estimates T of $\theta \in \Theta$ with $E_\theta(T^2) < \infty \forall \theta$, and suppose that U is non–empty. Let U_0 be the set of all unbiased estimates of 0, i.e., $U_0 = \{\nu : E_\theta(\nu) = 0, E_\theta(\nu^2) < \infty \forall \theta \in \Theta\}$. Then $T_0 \in U$ is UMVUE iff $E_\theta(\nu T_0) = 0 \forall \theta \in \Theta \forall \nu \in U_0$.*

Eg: Let X be unif$(\theta, \theta+1)$. Then $T = X - 1/2$ is unbiased for θ. An unbiased estimator $\nu(X)$ of zero has to satisfy $\int_\theta^{\theta+1} \nu(x)dx = 0$ for all θ. One such function is $\nu(x) = sin(2\pi x)$.

$$Cov(X - 1/2, sin(2\pi X)) = -cos(2\pi\theta)/2\pi.$$

This is non-zero. So T is not UMVU.
Eg: $X_1, \cdots, X_n$ iid unif(0,θ). Here $Y = (n+1)T/n$ is unbiased with T as $X_{(n)}$. Note that T is a sufficient statistic. We need to check if this is uncorrelated with all unbiased estimators of zero. Suppose W is an unbiased estimator of zero and cov(W,Y)>0. Then cov(E(W—Y),Y)=E(YE(W—Y))-E(W)E(Y)=EE(WY—Y)-E(W)E(Y)=cov(W,Y)¿0. So wlog, W can be considered a function of Y, equivalently a function of T. But T is complete sufficient implying W=0. Since Y is uncorrelated with W, Y is UMVE.

Theorem 2 *Let U be the non–empty class of unbiased estimates of $\theta \in \Theta$ as defined in Theorem 1. Then there exists at most one UMVUE $T \in U$ for θ.*

Theorem 3 *(Rao–Blackwell) Let W be any unbiased estimator of $\tau(\theta)$ and T be a sufficient statistic for θ. Define $\phi(T) = E(W \mid T)$. Then $\phi(T)$ is an estimator with $E(\phi(T)) = \tau(\theta)$ and $var(\phi(T)) \leq var(W)$.*

Pf: CB pf 342
This process of conditioning an unbiased estimator on a sufficient statistic is called Rao Blackwellization and leads to another unbiased estimator with uniformly lower variance. In other words, it is enough to consider the class of

unbiased estimators that are functions of sufficient statistics as any other unbiased estimator will have higher variance than one of them (the corresponding conditional correlation).

Eg: Suppose that $X_1, \cdots, X_n$ comes from density $\lambda exp(-\lambda x)$. Suppose that we want an estimate of $\theta = exp(-10\lambda)$. This corresponds to the probability P[$X_i > 10$]. The maximum likelihood estimate of λ is $1/\bar{X}$, so we could certainly claim $T = exp(-10/\bar{X})$ is the MLE of θ. This is certainly not unbiased.
Use statistic $u(X) = I(X_1 > 10)$. This statistic takes only the values 0 and 1, and it only depends on the first observation, so it's certainly a bad estimate. It is, however, unbiased.
The Rao-Blackwell theorem says that we can get a better unbiased estimate by using $u^*(X) = E[u(X)|V]$ where $V = \sum X_i$ is a sufficient statistic.
The conditional distribution of X_1/V given V is beta(1,n).
$u^*(X) = P[X_1 > 10|V] = P(beta(1, n) > 10/V) = (1 - 10/V)^n$

Eg (conditioning on an insufficient statistic): X_1, X_2 iid N$(\theta, 1)$. Then $\bar{X}$ is unbiased for θ. Let $\phi(\bar(X)) = E(\bar{X}|X_1)$. Then this is unbiased and has lower variance. But it is not an estimator (depends on θ).

4.2 Mean squared Error

Definition 3 *The Mean Squared Error (MSE) of an estimator W of a parameter θ is the function of θ defined by $E_\theta(W - \theta)^2$.*

Alternatively, Mean absolute error or expectation of any other increasing function of $|W - \theta|$ can be used as a measure of performance of an estimator. The advantage of MSE id easy tractability and the interpretation $MSE = Var + Bias^2$. (prove). For an unbiased estimator MSE=var. But a biased estimator might have lower MSE and will be preferred in most cases.

In the iid normal case, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Here $E(S^2) = \sigma^2, var[(n-1)S^2/\sigma^2] = 2(n-1), var(S^2) = 2\sigma^4/(n-1) = mse$. Now let us consider $\hat{\sigma^2}_{MLE} = (n-1)S^2/n$. Bias=$\sigma^2/n$. Var=$2(n-1)\sigma^4/n^2$. MSE=$(2n-1)\sigma^4/n^2$. This is smaller than MSE of the unbiased estimator S^2. Thus by trading off variance for bias, MSE is improved.

Eg Let $X_1, \cdots X_n$ be iid Ber(p). The MLE of p is $\bar{X}$ with MSE=Var=p(1-p)/n.
Consider the Bayes estimator with $Beta(\alpha, \beta)$ prior. The estimator equals $\hat{p}_B(\sum X_i + \alpha)/(n + \alpha + \beta)$. Taking $\alpha = \beta = \sqrt{n}/2$ makes MSE$(\hat{p}_B)$ constant as a function of p. With this prior, for small n, $\bar{X}$ has lower MSE than $\hat{p}_B$ unless p is close to zero or one. For large n, $\hat{p}_B$ has lower MSE than $\bar{X}$ unless p is close to half.

4.3 Information Inequality

Assumptions I. The set $A = x : p(x, \theta) > 0$ does not depend on θ. For all $x \in A, \theta \in \Theta, \partial/partial\theta logp(x, \theta)$ exists and is finite.
II. If T is any statistic such that E($|T|$) ¡ ∞ for all $\theta \in \Theta$, then the operations of integration and differentiation can be interchanged in $\partial/partial\theta \int T(x)p(x, \theta)dx$.

Theorem 4 *If $p(x, \theta) = h(x)exp\{\eta(\theta)T(x) - B(\theta)\}$ is an exponential family and $\eta(\theta)$ has a nonvanishing continuous derivative on Θ, then I and II hold.*

The Fisher Information is defined as $I(\theta) = E(logp(X, \theta))^2$.

Theorem 5 *Suppose that I and II hold and that $E|logp(X, \theta)| < \infty$. Then $E(logp(X, \theta)) = 0$ and $I(\theta) = Var(logp(X, \theta))$.*

Theorem 6 *(Information Inequality/ Cramer Rao Lower Bound) Let T(X) be any statistic such that Var(T(X)) ¡ ∞ for all θ. Denote E(T(X)) by $\psi(\theta)$. Suppose that I and II hold and $0 < I(\theta) < \infty$. Then for all $\theta, \psi(\theta)$ is differentiable and*

$$Var(T(X)) \geq \frac{(\psi'(\theta))^2}{I(\theta)}$$

5 Large Sample (Asymptotic) Properties of Estimators

[CB10.1, BD5.2-5.3]
Asymptotics in statistics is usually thought of as the study of the limiting behavior of statistics or, more specifically, of distributions of statistics, based on observing n i.i.d. observations $X_1, \cdots, X_n$ as $n \to \infty$. Asymptotics, in this context, always refers to a sequence of statistics $\{T_n(X_1, \cdots, X_n)\}_{n\geq 1}$, for instance the sequence of means $\{\bar{X}_n\}_{n\geq 1}$, or the sequence of medians, or it refers to the sequence of their distributions $\{\mathrm{L}_F(T_n(X_1, \cdots, X_n))\}_{n\geq 1}$. Asymptotic statements are always statements about the sequence.

The strong law of large numbers (Kolmogorov) tells us that if $X_1, X_2, \cdots, X_n$ are independent and identically distributed, the existence of a finite constant c for which $\bar{X} \stackrel{a.s.}{\to} c$ holds iff $E(X_1)$ is finite and equals c. [Serfling pg 27]

We interpret this as saying that, for n sufficiently large, X_n is approximately equal to its expectation. The trouble is that for any specified degree of approximation, say, $\epsilon = .01$, this does not tell us how large n has to be for the approximation not holding to this degree, that is $|\bar{X}(\omega) - c| > \epsilon$. Is $n > 100$ enough or does it have to be $n > 100,000$?

Central Limit Theorem(Lindeberg-Levy) If $X_1, X_2, \cdots, X_n$ are independent and identically distributed (distribution F) with finite mean ($E_F(X_1) = \mu$) and variance ($Var_F(X_1) = \sigma^2$), then $\sqrt{n}(\bar{X}_n - \mu) \Rightarrow \mathcal{N}(0, \sigma^2)$.

Relaxation of assumptions lead to other CLT's like Lindeberg Levy where independence is assumed but different means and variances are allowed satisfying suitable conditions.

As an approximation, this reads $P(\bar{X} \leq x) \approx \Phi(\sqrt{n}(x - \mu)/\sigma)$. Again we are faced with the questions of how good the approximation is for given n, x and F. What we in principle prefer are bounds, which are available in the classical situations of WLLN and CLT above.

By Chebychev's inequality, if $E_F(X_l^2) < \infty$, then $P_F[|\bar{X}_n - \mu| \geq \epsilon] \leq \sigma^2/n\epsilon^2$.As a bound this is typically far too conservative. If $|X_1| \leq 1$, the much more delicate Hoeffding bound gives $P_F[|\bar{X}_n - \mu| \geq \epsilon] \leq 2exp(-n\epsilon^2/2)$. Because $|X_1 \leq 1$ implies that $\sigma^2 \leq 1$ when σ^2 is unknown the RHS of Chebychev becomes $1/n\epsilon^2$. For $\epsilon = .1, n = 400$ Chebychev is 0.25 whereas Hoeffding is 0.14. Of course $|X_1 \leq 1$ can be replaced with $|X_1 \leq M$.

The celebrated Berry-Esseen bound states that if $E_F|X_1|^3 <^{\infty}$, $sup_x|P(\sqrt{n}(\bar{X}_n - \mu)/\sigma \leq x) - \Phi(x)| \leq CE_F|X_1|^3/\sigma^3\sqrt{n}$ where C is a universal constant known to be $< 33/4$.

5.1 Consistency

Consistency refers to convergence in probability (weak) or almost surely (strong). In the iid case, by LLN, $\bar{X}_n$ is consistent.

Ex: $X_1, \cdots X_n$ iid Bernoulli(p). $\hat{p}_n = \bar{X}_n$ is consistent for p by LLN. Consider $\theta = p(1 - p)$ which is the variance of X_1 and its method of moments

estimator $\hat{p}_n(1-\hat{p}_n)$. This is consistent.

Result: If X_n converges to X in probability and g is a continuous function, then $g(X_n)$ converges to $g(X)$ in probability.

Theorem 1 *(Consistency of MoM estimators) $X_1, \cdots, X_n$ iid. $X_i \in \mathcal{X}$. Let $g = (g_1, \cdots, g_d)$ map $\mathcal{X}$ onto $\mathcal{Y} \subseteq \mathbb{R}^d$. Suppose $E_\theta g_j(X_1) < \infty, 1 < j < d, \forall \theta$. Let $m_j(\theta) = E_\theta g_j(X_l), 1 < j < d$ and $q(\theta) = h(m(\theta))$, where $h : \mathcal{Y} \to \mathbb{R}^p$. Then, if h is continuous $\hat{q} = h(\bar{g})$ is consistent for $q(\theta)$.*

Theorem 2 *(Consistency of MLE in exponential family) Suppose $\mathcal{P}$ is a canonical exponential family of rank d generated by T. Suppose $\mathcal{E}$ the support of the canonical parameter η, is open. Then, if $X_l, \cdots, X_n$ are a sample from $P_\eta \in \mathcal{P}$*

1. *P(The MLE $\hat{\eta}$ exists) $\to$ 1*
2. *$\hat{\eta}$ is consistent.*

Pf: Pg 304 of BD. Not to be done in class.

We begin the discussion of the consistency of the MLE by defining the so-called Kullback-Liebler information.

Definition 1 *If $f_{\theta_0}(x)$ and $f_{\theta_1}(x)$ are two densities, the Kullback-Leibler information number equals $K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}$. If $P_{\theta_0}\left(f_{\theta_0}(X) > 0 \quad \text{and} \quad f_{\theta_1}(X) = 0\right) > 0$, then $K(f_{\theta_0}, f_{\theta_1})$ is defined to be 1.*

We may show that the Kullback-Leibler information must be nonnegative using Jensen's inequality.

Theorem 3 (Jensen's inequality) *If $g(t)$ is a convex function, then for any random variable $X, g(EX) \leq Eg(X)$. Furthermore, if $g(t)$ is strictly convex, then $Eg(X) = g(EX)$ only if $P(X = c) = 1$ for some constant c.*

Considering the Kullback-Leibler information once again, we first note that

$$E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = E_{\theta_1}\left(I_{f_{\theta_0}(X)>0}\right) \leq 1.$$

Therefore, by the strict convexity of the function $-\log x$,

$$K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} - \log \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq -\log E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq 0, \tag{1}$$

with equality if and only if $P_{\theta_0} f_{\theta_0}(X) = f_{\theta_1}(X) = 1$. Inequality (1) is sometimes called the Shannon- Kolmogorov information inequality.

If $X_1, \cdots, X_n$ are iid with density $f_{\theta_0}(x)$, then $l(\theta) = \sum_{i=1}^n \log f_{\theta_0}(x_i)$. Thus, the Shannon-Kolmogorov information inequality may be used to prove the consistency of the maximum likelihood estimator in the case of a finite parameter space.

Theorem 4 (Consistency of MLE) *Suppose Ω is finite and that $X_1, \cdots, X_n$ are iid with density $f_{\theta_0}(x)$. Furthermore, suppose that the model is identifiable, which is to say that different values of θ lead to different distributions. Then if $\hat{\theta}_n$ denotes the maximum likelihood estimator, $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

Proof: Notice that

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} \xrightarrow{P} E_{\theta_0} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} = -K(f_{\theta_0}, f_\theta) \tag{2}$$

The value of $-K(f_{\theta_0}, f_\theta)$ is strictly negative for $\theta \neq \theta_0$ by the identifiability of the model. Therefore, since $\hat{\theta}_n$ is the maximizer of the left hand side of Equation (2),

$$P(\hat{\theta}_n \neq \theta_0) = P\left(\max_{\theta \neq \theta_0}(\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}) > 0\right) \leq \sum_{\theta \neq \theta_0} P\left(\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0\right) \to 0. \tag{3}$$

5.2 Asymptotic normality

In the simplest form of the central limit theorem, we consider a sequence $X_1, X_2, ...X_n$ of independent and identically distributed (univariate) random variables with mean μ and finite variance σ^2. In this case, the central limit theorem states that

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

5.2.1 Delta Method

In this section, we wish to consider the asymptotic distribution of, say, some function of $\bar{X_n}$. In the simplest case, the answer depends on results already known: Consider a linear function $h(t) = at + b$ for some known constants a and b. Clearly $\mathrm{E}(h(\bar{X_n})) = a\mu + b = h(\mu)$ by the linearity of the expectation operator. Therefore, it is reasonable to ask whether $\sqrt{n}(h(\bar{X_n}) - h(\mu))$ tends to some distribution as $n \to \infty$. But the linearity of $h(t)$ allows one to write

$$\sqrt{n}(h(\bar{X_n}) - h(\mu)) = a\sqrt{n}(\bar{X_n} - \mu)$$

We conclude that

$$\sqrt{n}(h(\bar{X_n}) - h(\mu)) \Rightarrow \mathcal{N}(0, a^2\sigma^2)$$

None of the preceding development is especially deep; one might even say that it is obvious that a linear transformation of the random variable T_n alters its asymptotic distribution by a constant multiple. Yet what if the function $h(t)$ is nonlinear? It is in this nonlinear case that a strong understanding of the argument above, as simple as it may be, pays real dividends. For if T_n is consistent for θ (say), then we know that, roughly speaking, T_n will be very close to θ for large n. Therefore, the only meaningful aspect of the behavior of

$h(t)$ is its behavior in a small neighborhood of θ. And in a small neighborhood of θ, $h(\theta)$ may be considered to be roughly a linear function. Formally we use the Taylor expansion to obtain the following result:

Theorem 5 (First Order Delta Method) *If*

$$\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \tau^2) \tag{4}$$

then

$$\sqrt{n}(h(T_n) - h(\theta)) \Rightarrow \mathcal{N}(0, \tau^2(h'(\theta))^2)$$

provided $h'(\theta)$ exists and is not zero.

Proof: Step 1: It follows from equation (4) that $T_n \to \theta$ in probability.
Step 2: Consider the Taylor expansion of h around θ.

$$h(x) = h(\theta) + (x - \theta)(h'(\theta) + r)$$

where $r \to 0$ as $x \to \theta$.
Define R_n as the remainder in

$$h(T_n) = h(\theta) + (T_n - \theta)(h'(\theta) + R_n)$$

By step 1, $T_n \to \theta$ in probability.
Hence $R_n \to 0$ in probability.
This implies $h'(\theta) + R_n \to h'(\theta)$ in probability.

Step 3: The result follows by applying Slutsky's theorem to $\sqrt{n}(h(T_n) - h(\theta))$.
$\sqrt{n}(h(T_n) - h(\theta)) = \sqrt{n}(T_n - \theta) \times (h'(\theta) + R_n)$.
Let $Y_n = (h'(\theta) + R_n)$ and $X_n = \sqrt{n}(T_n - \theta)$ as above.
$X_n \Rightarrow X$ and $Y_n \to c$ in probability where $c = h'(\theta), X \sim \mathcal{N}(0, \tau^2)$.
By Slutsky's theorem, $\sqrt{n}(h(T_n) - h(\theta)) = Y_n X_n \Rightarrow cX$.
The distribution of cX is $\mathcal{N}(0, \tau^2(h'(\theta))^2)$.

Example 1(Exponential Rate) Let $X_i, i = 1, 2, \cdots, n$ be independent Exponential(λ) random variables and let $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then by CLT,

$$\sqrt{n}(T_n - \lambda) \Rightarrow \mathcal{N}(0, \lambda^2)$$

Suppose we are now interested in the large sample behavior of the estimate $\frac{1}{T_n}$ of the rate $h(\lambda) = \frac{1}{\lambda}$.
Since $h'(\lambda) = -\frac{1}{\lambda^2}$, it follows from Theorem 5 that

$$\sqrt{n}(\frac{1}{T_n} - \frac{1}{\lambda}) \Rightarrow \mathcal{N}(0, \left(-\frac{1}{\lambda^2}\right)\lambda^2 = \frac{1}{\lambda^2})$$

Example 2 (Binomial Variance) Let $X_i, i = 1, 2, \cdots, n$ be independent Bernoulli random variables and let $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then by CLT,

$$\sqrt{n}(T_n - p) \Rightarrow \mathcal{N}(0, p(1-p))$$

Suppose we are now interested in the large sample behavior of the estimate $T_n(1-T_n)$ of the variance $h(p) = p(1-p)$.
Since $h'(p) = 1-2p$, it follows from Theorem 5, when $p \neq 1/2$, that

$$\sqrt{n}(T_n(1-T_n) - p(1-p)) \Rightarrow \mathcal{N}(0, (1-2p)^2 p(1-p))$$

What happens when $h'(\theta) = 0$?

Theorem 6 (Second Order Delta Method) *If*

$$\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \tau^2) \quad \text{and} \quad h'(\theta) = 0 \tag{5}$$

then

$$n(h(T_n) - h(\theta)) \Rightarrow \frac{1}{2}\tau^2 h''(\theta)\chi_1^2$$

Proof Consider the Taylor expansion of $h(T_n)$ around $h(\theta)$ upto the second term.

$$h(T_n) = h(\theta) + (T_n - \theta)h'(\theta) + \frac{1}{2}(T_n - \theta)^2(h''(\theta) + R_n)$$

where $R_n \to 0$ as $T_n \to \theta$.

Step 1: It follows from equation (5) that $T_n \to \theta$ in probability.
Hence $R_n \to 0$ in probability. This implies $h''(\theta) + R_n \to h''(\theta)$ in probability.

Step 2: $\frac{1}{\tau^2} n(T_n - \theta)^2 \Rightarrow \chi_1^2$.
This follows from equation (5) after dividing by τ and squaring a standard normal random variable.

Step 3: The result follows by applying Slutsky's theorem to $n(h(T_n) - h(\theta))$.
$n(h(T_n) - h(\theta)) = n(T_n - \theta)^2 \times (h''(\theta) + R_n)$ since $h'(\theta) = 0$.
Let $Y_n = \tau^2(h''(\theta) + R_n)$ and $X_n = \frac{1}{\tau^2} n(T_n - \theta)^2$.
$X_n \Rightarrow X$ and $Y_n \to c$ in probability where $c = \tau^2 h''(\theta), X \sim \chi_1^2$.
By Slutsky's theorem, $n(h(T_n) - h(\theta)) = Y_n X_n \Rightarrow cX$.
The distribution of cX is $\tau^2 h''(\theta)\chi_1^2$.

Example 3'(Binomial Variance at $p = 1/2$) For $h(p) = p(1-p)$, we have at $p = 1/2$, $h'(1/2) = 0$ and $h''(1/2) = -2$. Hence from theorem 6, at $p = 1/2$,

$$n\left[T_n(1-T_n) - \frac{1}{4}\right] \Rightarrow -\frac{1}{4}\chi_1^2 \tag{6}$$

Although the equation (6) might appear strange, note that $T_n(1-T_n) \leq 1/4$, so the left side is always negative. An equivalent form is

$$4n\left[\frac{1}{4} - T_n(1-T_n)\right] \Rightarrow \chi_1^2$$

We now present a result on multivariate Delta method without proof.

Theorem 7 (Multivariate Delta Method) *Let* $(X_{1\nu}, \cdots, X_{s\nu}), \quad , \nu = 1, \cdots, n$ *be* n *independent* s*-tuples of random variables with* $\mathrm{E}(X_{i\nu}) = \xi_i$ *and* $\mathrm{Cov}(X_{i\nu}, X_{j\nu}) =$

σ_{ij}. Let $\bar{X}_i = \sum_{\nu=1}^{n} X_{i\nu}/n$, and suppose that h is a real valued function of s arguments with continuous first partial derivatives. Then

$$\sqrt{n}\left[h(\bar{X}_1, \cdots, \bar{X}_s) - h(\xi_1, \cdots, \xi_s)\right] \Rightarrow \mathcal{N}(0, \upsilon^2), \quad \text{where} \quad \upsilon^2 = \sum_{i=1}^{s}\sum_{j=1}^{s} \sigma_{ij} \frac{\partial h}{\partial \xi_i}\frac{\partial h}{\partial \xi_j}$$

Example 4 (Variance of Variance estimator) Suppose $X_1, \cdots, X_n$ are iid random variables with mean μ and variance σ^2. We are interested in the joint distribution of $s^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$, the estimator of σ^2. Denoting $\mathrm{E}(X^k)$ by m_k, we have

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= m_1 \\
\mathrm{E}(\bar{X}^2) &= m_2 \\
\mathrm{Cov}(\bar{X}, \bar{X}^2) &= (m_3 - m_1 m_2)/n \\
\mathrm{Var}(\bar{X}) &= (m_2 - m_1^2)/n \\
\mathrm{Var}(\bar{X}^2) &= (m_4 - m_2^2)/n
\end{aligned}$$

The parameter of interest is $\sigma^2 = h(m_1, m_2) = m_2 - m_1^2$. The derivatives of h are $\frac{\partial h}{\partial m_1} = -2m_1$ and $\frac{\partial h}{\partial m_2} = 1$.

$$\sqrt{n}\left[h(\bar{X}, \bar{X}^2) - h(m_1, m_2)\right] \Rightarrow \mathcal{N}(0, \upsilon^2), \quad \text{where}$$

$$\begin{aligned}
\upsilon^2 = \mathbf{D}h\Sigma\mathbf{D}h^T &= \begin{pmatrix} -2m_1 & 1 \end{pmatrix} \begin{pmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{pmatrix} \begin{pmatrix} -2m_1 \\ 1 \end{pmatrix} \\
&= -4m_1^4 + 8m_1^2 m_2 + m_4 - m_2^2 - 4m_1 m_3
\end{aligned}$$

The central limit theorem and the delta method will prove very useful in deriving asymptotic distribution results about functions of sample moments.

Example 9 (Distribution of sample T statistic) Suppose $X_1, X_2, \cdots, X_n$ are iid with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Define $s_n^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$, and let

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n}.$$

Letting $A_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ and $B_n = \sigma/s_n$, we obtain $Tn = A_n B_n$. Therefore, since $A_n \Rightarrow \mathrm{N}(0,1)$ by the central limit theorem and $B_n \xrightarrow{P} 1$ by the weak law of large numbers, Slutsky's theorem implies that $T_n \Rightarrow \mathcal{N}(0,1)$. In other words, T statistics are asymptotically normal under the null hypothesis.

5.2.2 Asymptotic Normality of MLE

It will be necessary to review a few facts regarding Fisher information before we proceed. For a density (or mass) function $f_\theta(x)$, we define the Fisher information function to be

$$I(\theta) = E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}^2 \tag{7}$$

If $\eta = g(\theta)$ for some invertible and differentiable function $g(\Delta)$, then since

$$\frac{d}{d\eta} = \frac{d\theta}{d\eta}\frac{d}{d\theta} = \frac{1}{g'(\theta)}\frac{d}{d\theta} \tag{8}$$

by the chain rule, we conclude that

$$I(\eta) = \frac{I(\theta)}{\{g'(\theta)\}^2} \tag{9}$$

Loosely speaking, $I(\theta)$ is the amount of information about θ contained in a single observation from the density $f_\theta(x)$.
Suppose that $f_\theta(x)$ is twice differentiable with respect to θ and that the operations of differentiation and integration may be interchanged in the following sense:

$$E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\} = E_\theta \left\{ \frac{\frac{d}{d\theta} f_\theta(X)}{f_\theta(X)} \right\} = \int \frac{d}{d\theta} f_\theta(X) dx = \frac{d}{d\theta} \int f_\theta(X) dx = \frac{d}{d\theta} 1 = 0 \tag{10}$$

$$E_\theta \left\{ \frac{d}{d\theta} \frac{\frac{d}{d\theta} f_\theta(X)}{f_\theta(X)} \right\} = E_\theta \left\{ \frac{\frac{d^2}{d\theta^2} f_\theta(X)}{f_\theta(X)} \right\} - I(\theta) = \frac{d^2}{d\theta^2} \int f_\theta(X) dx - I(\theta) = -I(\theta) \tag{11}$$

Equations (10) and (11) give two additional expressions for $I(\theta)$. From Equation (10) follows

$$I(\theta) = \mathrm{Var}_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\} \tag{12}$$

and Equation (11) implies

$$I(\theta) = -E_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(X) \right\}. \tag{13}$$

In many cases, Equation (13) is the easiest form of the information to work with. Equations (12) and (13) make clear a helpful property of the information, namely that for independent random variables, the information about θ contained in the joint sample is simply the sum of the individual information components. In particular, if we have an iid sample from $f_\theta(x)$, then the information about θ equals $nI(\theta)$. The reason that we need the Fisher information is that we will show that under certain regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}\left\{ 0, \frac{1}{I(\theta_0)} \right\}, \tag{14}$$

where $\hat{\theta}_n$ is the MLE.

Example 1 (Poisson case) Suppose that $X_1, \cdots, X_n$ are iid Poisson(θ_0) random variables. Then the likelihood equation has a unique root, namely $\hat{\theta}_n = \bar{X}_n$, and we know that by the central limit theorem $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \theta_0)$. However, the Fisher information for a single observation in this case is

$$-E_\theta \left\{ \frac{d^2}{d\theta^2} \log f_\theta(X) \right\} = E_\theta \frac{X}{\theta^2} = \frac{1}{\theta} \tag{15}$$

Thus, in this example, equation (14) holds.

Rather than stating all of the regularity conditions necessary to prove Equation (12), we work backwards, figuring out the conditions as we go through the proof. The first step is to expand $l'(\hat{\theta}_n)$ in a power series around θ_0:

$$l'(\hat{\theta}_n) = l'(\theta_0) + (\hat{\theta}_n - \theta_0) l''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l'''(\theta_n^*) \tag{16}$$

for some θ_n^* between $\hat{\theta}_n$ and θ_0. Clearly, the validity of Equation (16) hinges on the existence of a continuous third derivative of $l(\theta)$. Rewriting equation (16) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{n}\{l'(\hat{\theta}_n) - l'(\theta_0)\}}{l''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0) l'''(\theta_n^*)} = \frac{\frac{1}{\sqrt{n}}\{l'(\theta_0) - l'(\hat{\theta}_n)\}}{-\frac{1}{n} l''(\theta_0) - \frac{1}{2n}(\hat{\theta}_n - \theta_0) l'''(\theta_n^*)} \tag{17}$$

Let's consider the pieces of Equation (17) individually. If the MLE is consistent, then $l'(\hat{\theta}_n) \xrightarrow{P} 0$. If Equation (10) holds and $I(\theta_0) < \infty$, then

$$\frac{1}{\sqrt{n}} l'(\theta_0) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta} \log f_{\theta_0}(X_i) \right) \Rightarrow \mathcal{N}(0, I(\theta_0)) \tag{18}$$

by the central limit theorem and Equation (12). If Equation (11) holds, then

$$\frac{1}{n} l''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_{\theta_0}(X_i) \xrightarrow{P} -I(\theta_0) \tag{19}$$

by the weak law of large numbers and Equation (13). Finally, we would like to have the term involving $l'''(\theta_n^*)$ disappear, so clearly it is enough to show that $\frac{1}{n} l'''(\theta)$ is bounded in probability in a neighborhood of θ_0. Putting all of these facts together gives a theorem.

Theorem 8 *Let $\hat{\theta}_n$ denote a consistent root of the likelihood equation. Assume also that $\mathrm{l}'''(\theta)$ exists and is continuous, that equations (10) and (11) hold, and that $\frac{1}{n}\mathrm{l}'''(\theta)$ is bounded in probability in a neighborhood of θ_0. Then if $0 < I(\theta_0) < \infty$, (14) holds.*

The theorem is proved by noting that under the stated regularity conditions, $l'(\hat{\theta}_n) \xrightarrow{P} 0$ so that the numerator in (17) converges in distribution to $\mathcal{N}\{0, I(\theta_0)\}$ by Slutsky's theorem. Furthermore, the denominator in (17) converges to $I(\theta_0)$, so another application of Slutsky's theorem gives the desired result.

Sometimes, it is not possible to find an exact zero of $l'(\theta)$. One way to get a numerical approximation to a zero of $l'(\theta)$ is to use Newton's method, in which we start at a point θ_0 and then set

$$\theta_1 = \theta_0 - \frac{l'(\theta_0)}{l''(\theta_0)}. \tag{20}$$

Ordinarily, after finding θ_1 we would set θ_0 equal to θ_1 and apply Equation (20) iteratively. However, we may show that by using a single step of Newton's method, starting from a $\sqrt{n}$-consistent estimator of θ_0, we may obtain an estimator with the same asymptotic distribution as $\hat{\theta}_n$. The proof of the following theorem is left as an exercise:

Theorem 9 *Suppose that $\tilde{\theta}_n$ is any $\sqrt{n}$-consistent estimator of θ_0 (i.e., $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is bounded in probability). Then under the conditions of Theorem 7, if we set*

$$\delta_n = \tilde{\theta}_n - \frac{\mathrm{l}'(\tilde{\theta}_n)}{\mathrm{l}''(\tilde{\theta}_n)} \tag{21}$$

then

$$\sqrt{n}(\delta_n - \theta_0) \Rightarrow \mathcal{N}(0, \frac{1}{I(\theta_0)}) \tag{22}$$

5.3 Relative efficiency

we have considered various cases where the distribution of estimators converged at rate $\sqrt{n}$ to the normal distribution. If there are multiple estimators of the same parameter with this property, then all of them are $\sqrt{n}$ consistent. We can use the asymptotic variance as a means of comparing such estimators. This is the idea of asymptotic relative efficiency.

Definition 2 *If two estimators W_n and V_n satisfy*

$$\sqrt{n}[V_n - \theta] \Rightarrow \mathcal{N}(0, \sigma_V^2)$$
$$\sqrt{n}[W_n - \theta] \Rightarrow \mathcal{N}(0, \sigma_W^2)$$

The **asymptotic relative efficiency(ARE)** *of V_n with respect to W_n is*

$$\mathrm{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2} \tag{23}$$

Example 1(ARE of Poisson Estimators) Suppose $X_1, \cdots, X_n$ are iid Poisson(λ), and we are interested in estimating $\tau = P_\lambda(X_1 = 0) = \exp(-\lambda)$. For example number of customers who come into a bank in a given time period is modeled as

a Poisson random variable and we are interested in the probability that no one will enter the bank in one time period. A natural (but somewhat naive) estimator comes from defining $Y_i = I(X_i = 0)$. The Y_is are iid Bernoulli($\exp(-\lambda)$) and hence it follows that

$$\sqrt{n}(\bar{Y}_n - \exp(-\lambda)) \Rightarrow \mathcal{N}(0, \exp(-\lambda)(1 - \exp(-\lambda)))$$

Additionally,the MLE of $\exp(-\lambda)$ is $\hat{\tau} = \exp(-\hat{\lambda})$ where $\hat{\lambda} = \bar{X}_n$ is the MLE of λ. Using the Delta method, we have

$$\sqrt{n}(\hat{\tau} - \tau) \Rightarrow \mathcal{N}(0, \lambda \exp(-2\lambda))$$

The ARE of $\bar{Y}_n$ wrt the MLE is

$$\text{ARE}(\bar{Y}, \exp(-\bar{X})) = \frac{\lambda \exp(-2\lambda)}{\exp(-\lambda)(1 - \exp(-\lambda))} = \frac{\lambda \exp(-\lambda)}{(1 - \exp(-\lambda))}$$

Examination of this function shows that it is strictly decreasing with a maximum of 1 at $\lambda = 0$ and tailing off rapidly (< 0.1 when $\lambda = 4$) to 0 as $\lambda \to \infty$. So in this case the MLE is better in terms of ARE.

5.4 Asymptotic Bias and Efficiency

(CB 470-471)

There are two ways in which we can look at the bias as sample size goes to infinity. We can look at the finite sample bias $\text{Bias}(T_n)$ and take the limit as $n \to \infty$. This is called the limiting bias. We can also look for a suitably scaled version of the estimator converges in distribution to a non-degenerate random variable and look at the bias of that limiting distribution. This is the asymptotic bias. Here are the precise definitions:

Definition 3 *An estimator T_n of $\tau(\theta)$ is unbiased in the limit, if* $\lim_{n\to\infty} \text{E}(T_n) = \tau(\theta)$.

Definition 4 *For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \Rightarrow \mathcal{H}$. The estimator T_n is asymptotically unbiased if the expectation of $\mathcal{H}$ is zero.*

Example 1 (Asymptotically biased estimator) Let $X_1, \cdots, X_n$ are iid $U(0, \theta)$.

$$\text{The MLE of} \quad \theta \quad \text{is} \quad X_{(n)} \tag{24}$$

$$P(X_{(n)} \leq a) = (a/\theta)^n \quad \text{and} \quad \text{E}(X_{(n)}) = \theta \tag{25}$$

Hence $P(n(\theta - X_{(n)}) \leq a) = P(X_{(n)} \geq \theta - a/n) = 1 - (1 - a/n\theta)^n \to 1 - e^{-a/\theta}$. Thus $n(\theta - X_{(n)}) \Rightarrow \text{Exp}(\frac{1}{\theta})$. The expectation of the limiting random variable is not zero. So $X_{(n)}$ is not asymptotically unbiased. From (25) $X_{(n)}$ is unbiased in the limit.

Similar concepts exist for efficiency, which concerned with the asymptotic variance of the estimator.

Definition 5 *For an estimator T_n, if $\lim_{n\to\infty} k_n \mathrm{Var}(T_n) = \tau^2 < \infty$, where k_n is a sequence of constants, then τ^2 is called the limiting variance.*

Definition 6 *For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \Rightarrow \mathcal{H}$. Then $Var(H)$ is called the asymptotic variance of T_n.*

In most cases these two are the same. But in complicated cases, this may not hold. It is always the case that the asymptotic variance is smaller than the limiting variance (Lehmann and Casella Sec 6.1).

Example 2 Let us consider the mean $\bar{X}_n$ of n iid normal observations with mean μ and variance σ^2. Suppose we are interested in estimating $\frac{1}{\mu}$ and we use the estimator $T_n = \frac{1}{\bar{X}_n}$. For each finite n the distribution of $\sqrt{n}\bar{X}_n$ is $\mathcal{N}(0, \sigma^2)$.

$$\mathrm{Var}(\sqrt{n}T_n) = \infty, \quad \text{by direct integral of} \quad \frac{1}{x^2} \quad \text{with respect to the normal pdf.} \tag{26}$$

So, the limiting variance of T_n is infinity. On the other hand, by Delta method,

$$\sqrt{n}(T_n - \frac{1}{\mu}) \Rightarrow \mathcal{N}(0, \frac{\sigma^2}{\mu^4})$$

So the asymptotic variance of T_n is $\frac{\sigma^2}{\mu^4}$.

In the spirit of the Cramer Rao lower bound, there is an optimal asymptotic variance.

Definition 7 *A sequence of estimators W_n is asymptotically efficient for a parameter $\tau\theta$ if $\sqrt{n}(W_n - \tau(\theta) \Rightarrow \mathcal{N}(0, \nu(\theta)$ and*

$$\nu(\theta) = \frac{(\tau'(\theta)^2)}{\mathrm{E}_\theta((\frac{\partial}{\partial\theta}\log f(X \mid \theta))^2)} = \frac{(\tau'(\theta)^2)}{I(\theta)}, \tag{27}$$

that is the asymptotic variance of W_n achieves the Cramer-Rao lower bound.

For a long time it was believed that if

$$\sqrt{n}(W_n - \tau(\theta) \Rightarrow \mathcal{N}(0, \nu(\theta), \tag{28}$$

then

$$\nu(\theta) \geq \frac{(\tau'(\theta)^2)}{I(\theta)} \tag{29}$$

under regularity conditions on the densities. This belief was shattered by the example (due to hodges; see LaCam 1953) below:

Example 3 (Superefficient Estimator): Let $X_1, \cdots, X_n$ be iid $\mathcal{N}(\theta, 1)$ and the parameter of interest is θ. In this case, $h(\theta) = \theta$, and

$$\begin{aligned} I(\theta) &= \mathrm{E}_\theta((\frac{\partial}{\partial\theta}\log f(X \mid \theta))^2) \\ &= \mathrm{E}_\theta((\frac{\partial}{\partial\theta}\frac{1}{2}(X-\theta)^2)^2) \\ &= \mathrm{E}_\theta(X-\theta)^2 \\ &= 1 \end{aligned}$$

Thus equation(29) reduces $\nu(\theta) \geq 1$. Now consider the sequence of estimators

$$T_n = \begin{cases} \bar{X} & \text{if} \quad \mid \bar{X} \mid \geq 1/n^{1/4} \\ a\bar{X} & \text{if} \quad \mid \bar{X} \mid < 1/n^{1/4} \end{cases}$$

$$\text{Then,} \quad \sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \nu(\theta)), \tag{30}$$

$$\text{where} \quad \nu(\theta) = 1 \quad \text{when} \quad \theta \neq 0 \quad \text{and} \quad \nu(\theta) = a^2 \quad \text{when} \quad \theta = 0. \tag{31}$$

If $a < 1$, inequality (29) is violated at $\theta = 0$.

This phenomenon is quite common and is called superefficiency. There will typically exist estimators satisfying (28) but with $\nu(\theta)$ violating (29) at least for some values of θ. However, it was shown by LaCam(1953) that for any sequence of estimators satisfying (28), the set S of points of super-efficiency has Lebesgue measure zero.

5.5 Results and concepts from probability

1. Convergence almost surely(a.s), convergence in probability(P), convergence in distribution(d).
2. a.s.$\Rightarrow$ P$\Rightarrow$ d. But not the other way around.
3. Strong and weak laws of large numbers.
4. Central Limit Theorem
5. Chebyshev and Jensen inequalities
6. Continuous mapping Theorem: (pg 24 od Serfling) g is a continuous function. Then,
   (a) $X_n \Rightarrow X$ implies $g(X_n) \Rightarrow g(X)$.
   (b) $X_n \stackrel{P}{\rightarrow} X$ implies $g(X_n) \stackrel{P}{\rightarrow} g(X)$
   (c) $X_n \stackrel{a.s.}{\rightarrow} X$ implies $g(X_n) \stackrel{a.s.}{\rightarrow} g(X)$
7. Slutsky's Theorem:(pg 19 of Serfling) $X_n \Rightarrow X$ and $Y_n \stackrel{P}{\rightarrow} c$, where c is a constant. Then
   (a) $X_n + Y_n \Rightarrow X + c$
   (b) $X_n Yn \Rightarrow cX$
   (c) $X_n / Yn \Rightarrow X/c$ provided $c \neq 0$.

6 Elements of hypothesis testing

[CB8.3,BD4.2-4.3]

6.1 Introduction

Hypothesis testing begins with an assumption, called a hypothesis, that we make about a population parameter.
The bottom line in hypothesis testing is when we ask whether a population like we think this one is would be likely to produce a sample like the one we are looking at.

In hypothesis testing, we must state the assumed or hypothesized distribution of the population before we begin sampling.
The assumption we wish to test is called the **null hypothesis** (H). In parametric inference this will be in terms of a finite number of parameters.
Whenever we reject the hypothesis, the conclusion we draw is called **alternative hypothesis** (K).
Note: Null hypotheses are either rejected, or else there is insufficient evidence to reject them. (i.e., we don't accept null hypotheses.)

Reality ↓ / Test Result →	Do not reject H_0	Reject H_0
H_0 is true	Correct!	Type I Error: rejecting a true null hypothesis P(Type I error) = α
H_0 is false	Type II Error: not rejecting a false null hypothesis P(Type II error) = 1-β	Correct!

Type I Error: rejecting a true null hypothesis.
Max value of P(Type I error)=α Significance level of the test
Type II Error: not rejecting a false null hypothesis.
P(Type II error) = 1-β=1-Power of the test

Definition 1 *The* **power** *of a test ϕ against the alternative θ is the probability of rejecting H when θ is true and is denoted by $\beta(\theta, \phi)$.*

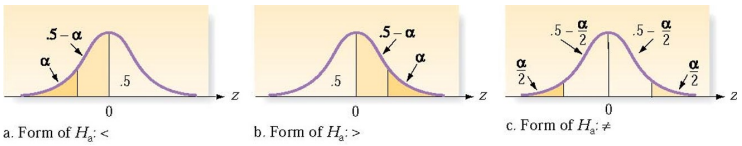
Example 1: The null hypothesis is that the battery has an average life of 300 days, with the alternative hypothesis being that the battery life is more than 300 days. You are the quality control engineer for the battery manufacturer.
(a) Would you rather make a Type I or a Type II error?

(b) Based on your answer to part (a), should you use a high or a low significance level?

Testing procedure: Fix H, K, α. Obtain sample. Calculate a test statistic based on the sample. If the test statistic has a low probability (fixed at α) when H is true, then H is rejected. Otherwise H is not rejected.

One and two sided alternative hypotheses. The null hypothesis is usually stated as an equality. The alternative hypothesis can be either an equality or an inequality.

One and two tailed tests. Depending on the type of the alternative the rejection region can be right-tailed, left-tailed or two-tailed.



Example 2: A drug will be released in the market only if it's efficacy is more than 30%. What are the null and alternative hypotheses? Which is appropriate, a one-tailed or a two-tailed test?

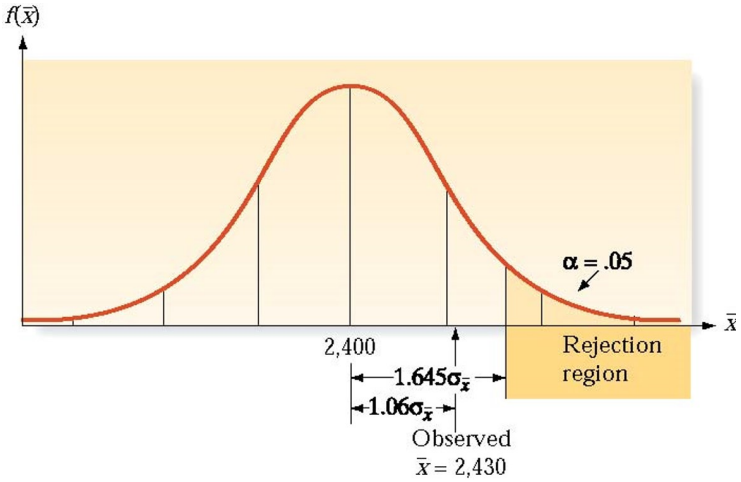


Figure 1: Example 3

Example 3: In figure 6.1, we have a sample of size n from a normal population with unknown mean μ. H: $\mu = 2400$, K: $\mu > 2400$. At $\alpha = 0.05$, we reject H for values of $\bar{X} > 2400 + 1.645\sigma_{\bar{X}}$. In this case, the observed value of $\bar{X}$ is 2430, which is $1.06\sigma_{\bar{X}}$. Hence we fail to reject H.

Parametric set-up: Data: $X \in \mathcal{X}, X \sim P_\theta, H : \theta \in \Theta_0, K : \theta \in \Theta_1$. If Θ_0 consists of a single point, we call it a **simple null**, otherwise, a **composite**

null. Similarly, with K.

In example 2, $H : \theta = \theta_0, K = \theta > \theta_0$, then we have a simple null vs a composite alternative. If we allow $H : \theta \leq \theta_0$, then we have a composite null. In most cases (Monotone Likelihood Ratio situations), the solutions to both problems are the same. In this example with $H : \theta = \theta_0$, it is reasonable to reject H if $X=$ number of cases in which the drug is effective in n trials, is "much" larger than what would be expected by chance if H is true and the value of θ is θ_0.
Thus, we reject H if X exceeds or equals some integer, say k.

Critical region or rejection region denotes the values of the test statistic X for which we reject H. In this example the critical region C is $\{X : X > k\}$. This is equivalent to specifying a test function $\phi : \mathcal{X} \to \{0, 1\}$, where 1 denotes rejection.

Thus $P(typeIerror) = P_{\theta=\theta_0}(X \geq k)$
and $P(typeIIerror) = P_\theta(X < k), \theta < \theta_0$.
k is called the critical value.
The power is obtained as $\sum_{i=k}^{n} \binom{n}{k} \theta^i (1-\theta)^{n-i}$. A plot of the function for $n = 10, \theta_0 = 0.3, k = 6$ is in figure 4.1.1 below (taken from BD).

Note that in this example the power at $\theta = \theta_1 > 0.3$ is the probability that the level 0.05 test will detect an improvement of the recovery rate from 0.3 to θ_1. When θ_1 is 0.5, a 67% improvement, this probability is only .3770. What is needed to improve on this situation is a larger sample size n. One of the most important uses of power is in the selection of sample sizes to achieve reasonable chances of detecting interesting alternatives

Also, the power function is increasing. It follows that the level and size of the test are unchanged if instead of $\Theta_0 = \{\theta_0\}$ we used $\Theta_0 = [0, \theta_0]$. That is,

$$\alpha(k) = sup\{P_\theta(X \geq k) : \theta < \theta_0\} = P_{\theta_0}(X \geq k).$$

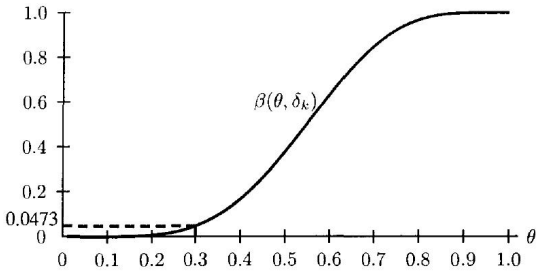


Figure 4.1.1. Power function of the level 0.05 one-sided test δ_k of $H : \theta = 0.3$ versus $K : \theta > 0.3$ for the $\mathcal{B}(10, \theta)$ family of distributions. The power is plotted as a function of $\theta, k = 6$ and the size is 0.0473.

6.2 Neyman-Pearson Theory

We start with the problem of testing a simple hypothesis $H : \theta = \theta_0$ versus a simple alternative $K : \theta = \theta_1$. Simple **likelihood ratio statistic** is defined by

$$L(x, \theta_0, \theta_1) = \frac{p(x, \theta_0)}{p(x, \theta_1)}$$

where $p(x, \theta)$ is the density (pdf) or frequency (pmf) function of the random vector X.
We call ϕ_k a likelihood ratio or Neyman-Pearson (NP) test (function) if for some $0 < k < \infty$ we can write the test function ϕ_k as

$$\phi_k(x) = \begin{cases} 1 & \text{if} \quad L(x, \theta_0, \theta_1) < k \\ 0 & \text{if} \quad L(x, \theta_0, \theta_1) > k \end{cases} \tag{1}$$

with $\phi_k(x)$ any value in (0,1) if equality occurs.

Because we want results valid for all possible test sizes α in $[0, 1]$, we consider randomized tests ϕ, which are tests that may take values in $(0, 1)$. If $0 < \phi(x) < 1$ for the observation vector x, the interpretation is that we toss a coin with probability of heads $\phi(x)$ and reject H iff the coin shows heads.

Theorem 1 *(Neyman-Pearson Lemma)*

1. *If $\alpha > 0$ and ϕ_k is a size α likelihood ratio test, then ϕ_k is MP in the class of level α tests.*
2. *For each $0 < \alpha < 1$ there exists an MP size α likelihood ratio test provided that randomization is permitted, $0 < \phi(x) < l$,for some x.*
3. *If ϕ is an MP level α test, then it must be a level α likelihood ratio test; that is, there exists k such that $P_\theta(\phi_k(x) \neq \phi(x), L(X, \theta_0, \theta_1 \neq k) = 0$ for $\theta = \theta_0$ and $\theta = \theta_1$*

It follows from the Neyman-Pearson lemma that an MP test has power at least as large as its level; that is,

Corollary 1 *If ϕ is an MP level α test, then $E_{\theta_1}\phi(x) > \alpha$ with equality iff $p(x, \theta_0) = p(x, \theta_1) \forall x$.*

In example 2, suppose the alternative is $\theta_1 = 0.5$. As before $\theta_0 = 0.3$. Thus we have a simple null vs simple alternative situation where the model is $X \sim Bin(n, \theta)$. The likelihood ration is

$$L(X, \theta_0, \theta_1) = \frac{\binom{n}{X}(0.3)^X(0.7)^{n-X}}{\binom{n}{X}(0.5)^X(0.5)^{n-X}} = (3/7)^X(7/5)^n$$

$L < k$ is equivalent to $X > (n \log(7/5) - k)/\log(7/3) = k_1$ (say). So the test that rejects the null hypothesis for large values of X is MP in the class of level α tests by NP lemma.

In order to determine the test explicitly given $\alpha = 0.05$ and n=10, we find the highest k_1 such that $P(X > k_1) < \alpha$.
From R, 1-pbinom(4,10,0.3)=0.1502683=$P(X > 4)$
and 1-pbinom(5,10,0.3)=0.04734899=$P(X > 5)$.
So, k_1=5 and a=$(\alpha - P(X > k))/P(X = 5)$=0.02575813. So the test function is

$$\phi(x) = \begin{cases} 1 & \text{if} X > 5 \\ 0.02575813 & \text{if} X = 5 \\ 0 & \text{if} X < 5 \end{cases} \quad (2)$$

That is, reject H if $X > 5$ and with probability a if $X = 5$.
For $\theta = 0.5$, the power is

$$\begin{aligned} \beta(\theta, \phi) &= P(X < 5) + aP(X = 5) \\ &= 1 - pbinom(5, 10, 0.5) + a * dbinom(5, 10, 0.5) \\ &= 0.383292 \end{aligned}$$

6.3 UMP tests and MLR families

Now we want to consider the case of composite null $H : \theta \in \Theta_0$ vs composite alternative $K : \theta \in \Theta_1$

Definition 2 *A level α test $\phi*$ is uniformly most powerful (UMP) for H vs K if*

$$\beta(\theta, \phi*) \geq \beta(\theta, \phi) \forall \theta \in \Theta_1$$

for any other level α test ϕ.

Definition 3 *The family of models $\{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}$ is said to be a monotone likelihood ratio (MLR) family if for $\theta_1 < \theta_2$ the distributions P_{θ_0} and P_{θ_2} are distinct and the ratio $p(x, \theta_2)/p(x, \theta_1)$ is an increasing function of a statistic $T(x)$.*

In example 2, for $\theta_1 < \theta_2$,

$$p(x, \theta_2)/p(x, \theta_1) = \frac{\theta_2^X (1 - \theta_2)^{n-X}}{\theta_1^X (1 - \theta_1)^{n-X}} = \left(\frac{\theta_2(1 - \theta_1)}{\theta_1(1 - \theta_2)} \right)^X \left(\frac{1 - \theta_2}{1 - \theta_1} \right)^n$$

is increasing in X and the model is MLR in $T(X) = X$.

Result: Consider the one-parameter exponential family model

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

If $\eta(\theta)$ is strictly increasing in $\theta \in \Theta$, then this family is MLR. Example 2 is of this form with $T(x) = x$ and $\eta(\theta) = \log(\theta/(1 - \theta))$.

Define the Neyman-Pearson (NP) test function

$$\delta_t(x) = \begin{cases} 1 & \text{if} \quad T(x) > t \\ 0 & \text{if} \quad T(x) < t \end{cases} \tag{3}$$

with $\delta_t(x)$ any value in (0,1) if $T(x) = t$. Consider the problem of testing $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ with $\theta_0 < \theta_1$. If $\{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}$, is an MLR family in $T(x)$, then $L(x, \theta_0, \theta_1) = g(T(x))$ for some increasing function g. Thus, δ_t equals the likelihood ratio test $\phi_{g(t)}$ and is MP. Because δ_t does not depend on θ_1 it is UMP at level $\alpha := E_{\theta_0}\delta_t(X)$ for testing $H : \theta = \theta_0$ versus $K : \theta > \theta_0$.

Theorem 2 *Suppose $\{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}$ is an MLR family in $T(x)$. Then*

1. *For each $t \in (0, \infty)$, the power function $\beta(\theta) = E_\theta\delta_t(X)$ is increasing in θ.*
2. *If $E_{\theta_0}\delta_t(X) = \alpha > 0$, then δ_t is UMP level α for testing $H : \theta \leq \theta_0$ versus $K : \theta > \theta_1$ for $\theta_1 > \theta_0$.*

6.4 Unbiased tests

Definition 4 *A test ϕ is unbiased if $\beta_\phi(\theta) \geq \alpha$ for all $\theta \in \Theta_1$ and $\beta_\phi(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.*

Remark: If ϕ is a UMP level α test, then ϕ is unbiased. Proof: compare ϕ with the trivial test function $\tilde{\phi} \equiv alpha$.

Definition 5 *A uniformly most powerful unbiased level α test is a test $\tilde{\phi}$ for which $E_\theta\tilde{\phi} \geq E_\theta\phi$ for all $\theta \in \Theta_1$ and for all unbiased level α tests ϕ.*

That is, $\tilde{\phi}$ is uniformly (for all $\theta \in \Theta_1$) most powerful ($E_\theta\tilde{\phi} \geq E_\theta\phi$) among all unbiased tests ϕ.

Theorem 3 *Consider testing $H : \theta = \theta_0$ versus $K : \theta \neq \theta_0$ in a one parameter exponential family with natural parameter θ and natural sufficient statistic T. The test ϕ with $E_{\theta_0}\phi(T) = \alpha$ given by*

$$\phi(T(x)) = \begin{cases} 1 & \text{if} \quad T(x) > c_2 \quad \text{or} \quad T(x) < c_1 \\ 0 & \text{if} \quad c_1 < T(x) < c_2 \end{cases} \tag{4}$$

with $\phi(T(x))$ any value in $\gamma \in (0, 1)$ if $T(x) = c_i, i = 1, 2.$ is UMPU for H versus K.

7 Likelihood Ratio and related tests

[CB8.2, CB10.3, BD4.9]

Definition 1 *The likelihood ratio test statistic for testing* $H : \theta \in \Theta_0$ *vs* $K : \theta \in \Theta_1$ *is*

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid x)}{\sup_{\theta \in \Theta} L(\theta \mid x)}$$

where $\Theta = \Theta_0 \bigcup \Theta_1$.

A likelihood ratio test is of the form

$$\phi(x) = \begin{cases} 1 & \text{if} \quad \lambda(x) \leq c \\ 0 & \text{if} \quad \lambda(x) > c \end{cases} \tag{1}$$

The value of c is determined from the level of the test such that $P_H(\lambda \leq c) = \alpha$.
Example 1: Consider testing $H : \mu = \mu_0$ vs $K : \mu \neq \mu_0$ where $X_1, \cdots, X_n$ are iid $\mathcal{N}(\mu, 1)$.
For the numerator, $\sup_{\theta \in \Theta_0} L(\theta \mid x) = \frac{1}{(2\pi)^{n/2}} exp(-\frac{1}{2} \sum_{i=1}^n (x_i - \mu_0)^2)$
The sup in the denominator is attained at $\theta = \bar{x}$ which is mle.
Hence $\sup_{\theta \in \Theta} L(\theta \mid x) = \frac{1}{(2\pi)^{n/2}} exp(-\frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2)$

$$\lambda(x) = exp(-\frac{1}{2}(\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{x})^2))$$

$\lambda(x) < c \iff \sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{x})^2 > c_1 \iff \mid \bar{x} - \mu_0 \mid > c_2$.
Thus the likelihood ration test rejects H when $\bar{x}$ differs from μ_0 by a large amount. The amount is determined from the constraint given by the level of the test, that is, $P_{\mu_0}(\mid \bar{x} - \mu_0 \mid > c_2) = \alpha$.

Exercise: Find the likelihood ration test for $H : \theta \leq \theta_0$ vs $K : \theta > \theta_0$ when $X_1, \cdots, X_n$ are iid from the exponential distribution with pdf $f(x \mid \theta) = exp(-x + \theta) I(x > \theta)$.

7.1 Large Sample Distribution of LRT

Let $X_1, \cdots, X_n$ be iid with density $f(x, \theta)$. We are interested in testing $H : \theta = \theta_0$ against $K : \theta \neq \theta_0$, where θ is of dimension k, using a likelihood ratio test. To carry out the test, we need to determine the appropriate critical value c. Recall that c is determined by the requirement that $P_H(\lambda(x) < c) = \alpha$. In order to determine the critical value, we thus need to determine the distribution of $\lambda(X)$ when the null hypothesis is true. We now develop a large sample approximation to solve this problem.

Let $\hat{\theta} = \text{argmax}_\theta L(\theta)$ denote the mle, and write the maximized likelihood ratio statistic as

$$\lambda(x) = \frac{L(\theta_0)}{L(\hat{\theta})} \tag{2}$$

Define the statistic $\xi_{LR}(x) = -2\ln(\lambda(x)) = 2(l(\hat{\theta}) - l(\theta_0))$ where $l(\theta) = \ln L(\theta)$. Since ξ_{LR} is a monotonic decreasing transformation of λ, the LR test can be implemented by rejecting the null hypothesis when $\xi_{LR}(x)$ is large.

To find the approximate distribution of $\xi_{LR}(X)$ under the null hypothesis, write

$$l(\theta_0) = l(\hat{\theta}) + (\theta_0 - \hat{\theta})' \frac{\partial l(\hat{\theta})}{\partial \theta} + \frac{1}{2}(\theta_0 - \hat{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) \tag{3}$$

where $\tilde{\theta}(\omega)$ is between θ_0 and $\hat{\theta}(\omega)$. Since mle is the root of the likelihood equation, $\frac{\partial l(\hat{\theta})}{\partial \theta} = 0$. We have

$$\begin{aligned} \xi_{LR} &= -(\theta_0 - \hat{\theta})' \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) && (4) \\ &= \sqrt{n}(\theta_0 - \hat{\theta})' \left(-\frac{1}{n} \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta} \right) \sqrt{n}(\theta_0 - \hat{\theta}) && (5) \end{aligned}$$

Proceeding as in our derivations of the properties of the maximum likelihood estimator,

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &\Rightarrow \mathcal{N}(0, I(\theta_0)^{-1}) && (6) \\ -\frac{1}{n} \frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} &\xrightarrow{P} I(\theta_0) && (7) \end{aligned}$$

so that by Slutsky and the Continuous Mapping Theorem,

$$\xi_{LR} \overset{H_0}{\Rightarrow} \chi^2_k \tag{8}$$

An asymptotically justified level $1 - \alpha$ confidence set based on the LR statistic is hence of the form

$$\theta^* \mid (\hat{\theta} - \theta^*)' \hat{V}^{-1} (\hat{\theta} - \theta^*) < c \tag{9}$$

where $\hat{V} = \left(-\frac{\partial^2 l(\tilde{\theta})}{\partial \theta \partial \theta'} \right)^{-1}$ and c solves $P(\chi^2_k > c) = \alpha$. This confidence set may be recognized as the interior of an ellipse centered at $\theta = \hat{\theta}$. In the one-dimensional case, we obtain a confidence interval $(\hat{\theta} - c^* \hat{V}^{-1/2}, \hat{\theta} + c^* \hat{V}^{-1/2})$ where c^* is the positive number that solves $P(\mathcal{N}(0,1) > c^*) = \alpha/2$.

7.2 Wald statistic

A close cousin of the LR statistic is the Wald statistic

$$\xi_W = \sqrt{n}(\hat{\theta} - \theta_0) \left(-\frac{1}{n} \frac{\partial^2 l(\hat{\theta})}{\partial \theta \partial \theta'} \right) \sqrt{n}(\hat{\theta} - \theta_0) \tag{10}$$

which differs from ξ_{LR} only because the estimated information matrix is evaluated at $\hat{\theta}$ rather than $\tilde{\theta}$. Note that we can compute the Wald statistic without doing any computations under the null hypothesis.

Since both $\hat{\theta}$ and $\tilde{\theta}$ converge in probability to θ_0 under the null hypothesis, $\xi_W - \xi_{\mathrm{LR}} \stackrel{P,H_0}{\to} 0$

The motivation of the Wald statistic is that under the null hypothesis, the difference between the estimator $\hat{\theta}$ and θ_0 satisfies $\sqrt{n}(\hat{\theta}-\theta_0) \Rightarrow \mathcal{N}(0, I(\theta_0)^{-1})$and $-\frac{1}{n}\frac{\partial^2 l(\hat{\theta})}{\partial\theta\partial\theta'}$ consistently estimates $I(\theta_0)^{-1}$. Under the alternative, $|| \hat{\theta} - \theta_0 ||$ is large and we reject.

7.3 Lagrange Multiplier statistic

Another approximation to ξ_{LR} is given by the Lagrange Multiplier test statistic

$$\begin{aligned} \xi_{\mathrm{LM}} &= \sqrt{n}S_n(\theta_0)\left(-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1}\sqrt{n}S_n(\theta_0) && (11)\\ &= \left(n^{-1/2}\sum_{i=1}^{n} s_i(\theta_0)\right)'\left(-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1}\left(n^{-1/2}\sum_{i=1}^{n} s_i(\theta_0)\right) && (12)\end{aligned}$$

with the advantage that we do not need to compute $\hat{\theta}$ in order to compute ξ_{LM}.

Since under the null hypothesis $n^{-1/2}\sum_{i=1}^{n} s_i(\theta_0) \Rightarrow \mathcal{N}(0, I(\theta_0))$ and $-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta'} \stackrel{P}{\to} I(\theta_0)$ we also find $\xi_{\mathrm{LM}} \stackrel{H_0}{\Rightarrow} \chi^2_k$.

7.4 Pearson's chi-square

[Lehman 5.5, Ferguson 9,10, Rao 6b]
Let $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ be iid from a multinomial$_k(1, \underline{p})$ distribution, where $\underline{p}$ is a k-vector with nonnegative entries that sum to one. That is,

$$P(\underline{X}_i = e_j) = p_j \quad \text{for all} \quad 1 \le j \le k \tag{13}$$

where $e_j =$ the k vector with 1 at the j-th position and 0's everywhere else.

Note that the multinomial distribution is a generalization of the binomial distribution to the case in which there are k categories of outcome instead of only 2. Also note that we ordinarily do not consider a binomial random variable to be a 2-vector, but we could easily do so if the vector contained both the number of successes and the number of failures. Equation (13) implies that the random vector $\underline{X}_i$ has expectation $\underline{p}$ and covariance matrix

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix} \tag{14}$$

Using the Cramer-Wold device, the multivariate central limit theorem implies

$$\sqrt{n}(\bar{\underline{X}}_n - \underline{p}) \Rightarrow \mathcal{N}_k(\underline{0}, \Sigma). \tag{15}$$

Note that the sum of the j-th column of Σ is $p_j - p_j(p_1 + \cdots + p_k) = 0$, which is to say that the sum of the rows of Σ is the zero vector, so Σ is not invertible.

We wish to derive the asymptotic distribution of Pearson's chi-square statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}, \tag{16}$$

where n_j is the random variable that is the j-th component if $n\underline{\bar{X}}_n$, the number of successes in the j-th category for trials $1, \cdots, n$. We will discuss two different ways to do this. One way avoids dealing with the singular matrix Σ, whereas the other does not.

In the first approach, define for each i, $\underline{Y}_i = (\underline{X}_{i1}, \cdots, \underline{X}_{ik-1})$. That is, let $\underline{Y}_i$ be the $k-1$-vector consisting of the first $k-1$ components of $\underline{X}_i$. Then the covariance matrix of $\underline{Y}_i$ is the upper-left $(k-1) \times (k-1)$ submatrix of Σ, which we denote by Σ^*. Similarly, let $\underline{p}^*$ denote the vector $(p_1, \cdots, p_{k-1})$. First, verify that Σ^* is invertible and that

$$\Sigma^{*-1} = \begin{pmatrix} \frac{1}{p_1} + \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2} + \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}} + \frac{1}{p_k} \end{pmatrix} \tag{17}$$

Second, verify that

$$\chi^2 = n(\underline{\bar{Y}}_n - \underline{p}^*)^t (\Sigma^*)^{-1} (\underline{\bar{Y}}_n - \underline{p}^*) \tag{18}$$

The facts in equations (17) and (18) are checked in exercise 1. If we now define

$$\underline{Z}_n = \sqrt{n}(\Sigma^*)^{-1/2}(\underline{\bar{Y}}_n - \underline{p}^*), \tag{19}$$

then clearly the central limit theorem implies $\underline{Z}_n \Rightarrow \mathcal{N}_{k-1}(\underline{0}, I)$. By definition, the χ^2_{k-1} distribution is the distribution of the sum of the squares of $k-1$ independent standard normal random variables. Therefore,

$$\chi^2 = (\underline{Z}_n)^t \underline{Z}_n \Rightarrow \chi^2_{k-1}, \tag{20}$$

which is the result that leads to the familiar chi-square test.

In a second approach to deriving the limiting distribution (20), we use some properties of projection matrices.

Definition 2 *A matrix P is called a projection matrix if it is idempotent; that is, if $P^2 = P$.*

The following lemmas, to be proven in exercise 2, give some basic facts about projection matrices.

Lemma 1 *Suppose P is a projection matrix. Then every eigenvalue of P equals 0 or 1. Suppose that r denotes the number of eigenvalues of P equal to 1. Then if $Z \sim \mathcal{N}_k(\underline{0}, P)$, then, $Z^t Z \sim \chi^2_r$.*

This can be derived from the Fisher-Cochran Theorem.

Lemma 2 *The trace of a square matrix equals the sum of its eigenvalues. For matrices A and B whose sizes allow them to be multiplied in either order,* $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

Define $\Gamma = \mathrm{diag}(\underline{p})$. Clearly, equation (15) implies

$$\sqrt{n}\Gamma^{-1/2}(\underline{\bar{X}}_n - \underline{p}) \Rightarrow \mathcal{N}_k(\underline{0}, \Gamma^{-1/2}\Sigma\Gamma^{-1/2}). \tag{21}$$

Since Σ may be written in the form $\Gamma - \underline{p}\underline{p}^t$,

$$\Gamma^{-1/2}\Sigma\Gamma^{-1/2} = I - \Gamma^{-1/2}\underline{p}\underline{p}^t\Gamma^{-1/2} = I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t \tag{22}$$

clearly has trace $k-1$; furthermore, $(I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t)(I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t) = I - 2\sqrt{\underline{p}}\sqrt{\underline{p}}^t + \sqrt{\underline{p}}\sqrt{\underline{p}}^t\sqrt{\underline{p}}\sqrt{\underline{p}}^t = I - \sqrt{\underline{p}}\sqrt{\underline{p}}^t$ because $\sqrt{\underline{p}}^t\sqrt{\underline{p}} = 1$, so the covariance matrix (22) is a projection matrix.

Define $\Delta_n = \sqrt{n}\Gamma^{-1/2}(\underline{\bar{X}} - \underline{p})$. Then we may check (exercise 2) that

$$\chi^2 = (\Delta_n)^t\Delta_n \tag{23}$$

Therefore, since the covariance matrix (22) is a projection with trace $k-1$, Lemma 1 and Lemma 2 prove that $\chi^2 \Rightarrow \chi^2_{k-1}$ as desired.

8 Confidence intervals

[CB9, BD4.4-4.5]
Let $X_1, \cdots, X_n$ be a random sample from a distribution P_θ which belongs to a family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.

Definition 1 *For fixed $\alpha \in (0,1)$, a random interval $[T_1(X), T_2(X)]$, where $P(T_1 < T_2) = 1$, such that $P_\theta(T_1 \leq \theta \leq T_2) = 1 - \alpha \forall \theta \in \Theta$, is called a $100(1-\alpha)\%$ Confidence Interval (CI) for θ. Random variables T_1 and T_2 are called the lower and upper limit, respectively; $1-\alpha$ is called the confidence coefficient.*

$T_1(X)$ and $T_2(X)$ are rvs, hence their value may be different for different realizations of the sample X . For some observed samples x, the interval $[T_1(x), T_2(x)]$ may not be covering the true unknown parameter θ. If the sampling procedure is repeated a large number of times, then the proportion of samples for which the interval actually covers θ should be approximately equal to $1-\alpha$. Note that here the interval is random, while θ is an unknown constant. $1-\alpha$ is usually taken to be 0.9, 0.95 or 0.99.

8.1 Pivotal method for finding CI

Definition 2 *Let $X = (X_1, \cdots, X_n)$ be a random sample from a distribution $P_\theta, \theta \in \Theta$. A function $Q(X, \theta)$ is called a pivot for θ if its distribution is completely known.*

Example 1: Let $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_0^2), \sigma_0^2$ known.
Then $Q(X, \mu) := (\bar{X} - \mu) \sim \mathcal{N}(0, \sigma_0^2/n)$ is a pivot for μ.
It is a function of the sufficient statistic $\bar{X}$.

Example 2: Let $X_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both parameters are unknown.
Then $Q(X, \mu) := (\bar{X} - \mu) \sim \mathcal{N}(0, \sigma^2/n)$ is not a pivot for μ since its distribution depends on σ.
However, $Q_1(X, \mu) := \sqrt{n}(\bar{X} - \mu)/S \sim t_{n-1}$ is a pivot for μ, where $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$.
Also, $Q_2(X, \sigma^2) := (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ is a pivot for σ^2.
Note, that these pivots are functions of sufficient statistics, which is usually the case.

To construct a CI we choose such values, say a and b, such that for a given $\alpha \in (0,1)$ we have $P_\theta(a \leq Q(X, \theta) \leq b) = 1 - \alpha \forall \theta \in \Theta$. Note that a and b are non-random since the distribution of Q is free of parameters. If Q is a strictly monotonic and continuous function of θ then this can be written as $P(T_1(X; a, b) \leq \theta \leq T_2(X; a, b)) = 1 - \alpha \forall \theta \in \Theta$. Then the CI for θ is $[T_1(X; a, b), T_2(X; a, b)]$.

In eg 2, Q_1 is strictly monotonic and continuous function of the parameter of interest μ. Note a and b can be chosen to be the p-th and q-th quantiles of the t distribution with $n-1$ df, such that $q-p=1-\alpha$. The most common choice is $q=1-\alpha/2$ and $p=\alpha/2$. Then a $(1-\alpha)$ level CI for μ is $\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$.

Exercise: Show that, in the setting of example 2, using Q_2 as pivot, a $(1-\alpha)$ level confidence interval for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right).$$

8.2 Approximate confidence intervals

If we cannot find an exact pivot, we will use an asymptotic pivot. This will often be based on the maximum likelihood estimator, which has an asymptotic normal distribution, i.e., $\sqrt{n}(\hat{\theta}-\theta) \Longrightarrow \mathcal{N}(0, 1/I(\theta))$ Here AN stands for asymptotically normal. Hence, an asymptotic pivot is a function is obtained as Q with its approximate distribution as follows:

$$Q = (X, \theta) = \sqrt{nI(\theta)}(\hat{\theta}-\theta) \sim \mathcal{N}(0,1).$$

Usually $I(\theta)$ will depend on the parameter. Then we use further approximation by substituting all the unknown parameters by their estimates (preferably consistent) to obtain $I(\hat{\theta})$. Therefore, for large n, we obtain

$$Q = (X, \theta) = \sqrt{nI(\hat{\theta})}(\hat{\theta}-\theta) \sim \mathcal{N}(0,1).$$

Example 3: Suppose that $X_i \overset{iid}{\sim} Poisson(\lambda)$ random variables. There is no obvious pivot in this case. The maximum likelihood estimator of λ is $\hat{\lambda} = \bar{X}$, and, for large n, we know that $\sqrt{n}(\hat{\lambda}-\lambda) \Rightarrow \mathcal{N}(0, \lambda)$. Thus, an approximate 95% confidence interval for λ is $\bar{X} \pm 1.96\sqrt{X/n}$.

Exercise: Use the above method to find an approximate CI for binomial parameter p.

8.3 Duality of CI and Tests

There is a duality between confidence intervals and hypothesis tests.

Example 1: Let $X_1, \ldots, X_n$ be a random sample from a normal distribution having unknown mean μ and known variance σ_0^2. We are interested in testing $H: \mu = \mu_0$ versus $K: \mu \neq \mu_0$. At significance level α, consider the following test: Reject H if $\mid \bar{X} - \mu_0 \mid > \sigma_0 z_{\alpha/2}/\sqrt{n}$, and do not reject otherwise. Thus the test accepts H when

$$-\sigma_0 z_{\alpha/2}/\sqrt{n} < \bar{X} - \mu_0 < \sigma_0 z_{\alpha/2}/\sqrt{n}.$$

The latter statement is also equivalent to a $(1-\alpha)$ level CI for μ_0 given by

$$\bar{X} - \sigma_0 z_{\alpha/2}/\sqrt{n} < \mu_0 < \bar{X} + \sigma_0 z_{\alpha/2}/\sqrt{n}.$$

In other words, the CI consists precisely of all those values of μ_0 for which the null hypothesis $H : \mu = \mu_0$ is accepted. The theorems below shows that the duality between CIs and hypothesis tests holds more generally.

Theorem 1 *Suppose that for every value θ_0 in Θ there is a test at level α of the hypothesis $H : \theta = \theta_0$. Denote the acceptance region of the test by $A(\theta_0)$. Then the set $C(X) = \theta : X \in A(\theta)$ is a $(1 - \alpha)$ confidence region for θ.*

Theorem 2 *Suppose that $C(X)$ is a $(1 - \alpha)$ confidence region for θ. Then an acceptance region for a test at level α of the hypothesis $H : \theta = \theta_0$ is $A(\theta_0) = X|\theta_0 \in C(X)$.*