Key Definitions: Sufficient, Complete, and Ancillary Statistics. Basu's Theorem.

1. Sufficient Statistics[1]: (Intuitively, a sufficient statistics are those statistics that in some sense contain all the information about θ)

A statistic T(**X**) is called sufficient for θ if the conditional distribution of the data **X** given T(**X**) = t does not depend on θ (i.e. P(**X** = **x** | T(**X**) = t) does not depend on θ).

How do we show T(X) is a sufficient statistic? (Note 1 pg 9)

Factorization Theorem:

In a regular parametric model $\{f(.,\theta):\theta\in\Theta\}$, *a statistic T(**X**)with range R is sufficient for* θ *iff there exists a function* $g: T\times\Theta\to\Re$ *and a function h defined on* $\aleph$ *such that* $\boxed{f(x,\theta)=g(T(x),\theta)h(x)\ \ \forall\, x\in\aleph \text{ and } \theta\in\Theta}$[2]

Sufficient Statistics in Exponential Families:

$p(x,\theta)=h(x)\exp\left[\sum_{j=1}^{k}\eta_j(\theta)T_j(x)-B(\theta)\right]$ or $p(x,\theta)=h(x)C(\theta)\exp\left[\sum_{j=1}^{k}\eta_j(\theta)T_j(x)\right]\rightarrow T(X)=(T_1(X),...,T_k(X))$ a sufficient stat. for θ

2. Ancillary Statistics: (Intuitively, ancillary statistics are those statistics that are in some sense uninformative about θ.)

A statistic S(X) is ancillary for a parameter θ if the distribution of S(X) does not depend on θ. It is first order ancillary if E(S(X)) does not depend on θ.

ANCILLARY STATISTICS ARE IRRELEVENT FOR INFERENCE ABOUT $\theta\rightarrow$ its realized values don't help us pin down θ since the realized values don't depend on θ.

Note: To show ancillarity, use the definition!

Intuition for The Need for Completeness:

For some sufficient statistics, T(X), if r(T(X)) is ancillary for some function r → T(X) contains some extraneous information about θ. We want to Eliminate this. "Completeness" of T(X) means that for all r(T) such that r is ancillary for θ, there are only constant functions, i.e. r(T(X)) = c. That is r does not depend on T(X). So, T(X) a complete statistic means that the only function r that makes r(T(X)) ancillary to θ are trivial, constant functions. It turns out that this is too strong, and we need only that if any function g is such that g(T) is first order ancillary, then g(T) = 0. So, the only way that a function of T can be uninformative about θ is if the function itself is constant/trivial function.

3. Complete Statistics[3]: (Intuitively, a complete statistic contains all the relevant, and no extraneous, information for inference about θ.)

A statistic T: $X\to T$ is complete if for every measurable real valued function g (.) defined on T such that $\mathrm{E}_\theta(g(T))=0\ \ \forall\,\theta\in\Theta\Rightarrow g(T)=0$

Complete Statistics in Exponential Families:

Let $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ *be an iid sample from an exponential family distribution with probability (mass or density) function of the form*

$$p(x,\theta)=h(x)\exp\left[\sum_{j=1}^{k}\eta_j(\theta)T_j(x)-B(\theta)\right]=C(\theta)h(x)\exp\left[\sum_{j=1}^{k}\eta_j(\theta)T_j(x)\right]$$

Then, the statistic $T(X)=\left(\sum_{i=1}^{n}T_1(X_i),...,\sum_{i=1}^{n}T_k(X_i)\right)$ *is complete if the set* $\{(\eta_1(\theta),...,\eta_k(\theta)):\theta\in\Theta\}$ *contains an open set in* R^k.

(Note: k is NOT the dimension of θ, but comes from the exponential form of the data.)

(We can do this with log likelihoods as well: e.g. Bernoulli distribution)

4. Basu's Theorem: (Intuitively, since complete statistics contain no extraneous information for inference about θ, it is independent of ancillary stat of θ)

If T(X) is a complete and (minimal) sufficient statistic, then T(X)is independent of every ancillary statistic

5. Theorem (minimal sufficiency and complete, sufficient statistics)[4]**:**

If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic.

---

[1] **Sufficient statistics have 3 main uses:**
A. **Decision Theory**: We should base decision rules on sufficient statistics
B. **Dealing with Nuisance Parameters in Hypothesis Testing**: We should find sufficient statistics for the nuisance parameters and condition decision rules on them
C. **Unbiased Estimation** (Lehman-Scheffe Theorem, Rao-Blackwell theorem): We should look for unbiased estimators that are functions of sufficient statistics

[2] We need to be able to write the prob. Dist of the data as a function of all the data and then a function of just the parameters

[3] Complete statistics are useful in **verifying the uniqueness of unbiased estimators**. If $T_1(S(X))$ is an unbiased estimator of θ and S(X) a complete statistic, then $T_1(S(X))$ must be unique (pg. 20 Note 1)

Pf: Suppose not. There exists $T_2(S(X))$ s.t. $E[T_2(S(X))]=\theta$. Define $h(T(X))=T_1(S(X))-T_2(S(X))\rightarrow E(h(T(X)))=0\rightarrow h(T(X))=0$ everywhere by completeness of $\rightarrow T_1(S(X))=T_2(S(X))$ almost everywhere (i.e. everywhere except those x with measure 0) $\rightarrow T_1(S(X))$ is a unique unbiased estimator of θ

[4] For purposes of this class, we will always assume that a minimal sufficient statistic exists. Therefore complete and sufficient statistics are minimal sufficient.

Example: How do we show minimal sufficient

Let $x_1,...,x_n$ *iid Bernoulli*(θ).

$$P(X \mid \theta) = \prod_i \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} = \exp\left\{\left(\sum_i x_i\right)\ln\theta + \left(n - \sum_i x_i\right)\ln(1-\theta)\right\}$$

$$\therefore T(\mathbf{X}) = \sum_i x_i \text{ is sufficient for } \theta.$$

To show min *suff , show for any other suff stat* $S(X)$, $\exists$ *a function r s.t.* $T = r(S(X))$

Pf : Let $S(X)$ *be any sufficient stat for* θ. *By factorization,* $f(X \mid \theta) = g(S(X),\theta)h(X)$

$$\Rightarrow g(S(X),\theta)h(X) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$$

$$\Rightarrow \ln g(S(X),\theta) + \ln h(X) = \left(\sum_i x_i\right)\ln\theta + \left(n - \sum_i x_i\right)\ln(1-\theta)$$

$$\Rightarrow \sum_i x_i = \frac{\ln g(S(X),\theta) + \ln h(X) - n\ln(1-\theta)}{\ln\frac{\theta}{1-\theta}}$$

Example: How do we show not complete?

Find non – trivial function $g(T)$ *s.t.* $E(g(T)) = 0$

(See 270B PS2 #11c)

Note on Sufficiency and Decision Theory:

Sufficiency can be given a clear operational interpretation in the decision theory setting. We base decision rules on sufficient statistics because, if T(**X**) is sufficient, we can, for any decision rule $\delta(x)$ find a randomized decision rule $\delta^*(x)$ depending only on T(X) that does as well as $\delta(x)$ in the sense of having the same risk function; i.e.

$$R(\theta,\delta(.)) = R(\theta,\delta^*(.))$$

Note: Randomized means $\delta^*(x)$ can be generated from the value *t* of T(X) and a random mechanism not depending on θ

Note2: So, optimally we want to use test statistics that are sufficient statistics.

Best Unbiased Estimators: UMVUE, and Finite Sample Variance Bounds

Unbiased Estimator: Consider the parameter space Θ for a parametric family $\{P_\theta : \theta \in \Theta\}$ where P_θ is a distribution over the sample space $\aleph$. Consider a mapping $g : \Theta \to \Im$ and a statistic $\phi : \aleph \to \Im$. We say that $\phi(.)$ is unbiased for g(θ) if $E_\theta(\phi(\mathrm{X})) = g(\theta) \quad \forall \theta \in \Theta$ (i.e. no matter what the "true" theta is)

Minimum Variance Unbiased Estimators (UMVUE) and "Best" Unbiased Estimators (Note 2 p.27)

1. UMVUE Estimators: An unbiased estimator ϕ of a quantity g(θ) is UMVUE if ϕ has finite variance and for every unbiased estimator d(X) of g(θ) we have $Var_\theta(\phi(\mathrm{X})) \le Var_\theta(d(\mathrm{X})) \quad \forall \theta \in \Theta$

2. Best Unbiased Estimators are Unique: *If* $\phi(\mathrm{X})$ is the best unbiased estimator of g(θ), then it is unique.

Sufficient Statistics, Complete Statistics, and UMVUE (Note 2 p.28): Unbiased estimators that are functions of sufficient and complete statistics are UNIQUE UMVUE!

1. **Theorem (Uniqueness of UMVUE)**:
If $\phi(\mathrm{X})$ is a "best" unbiased estimator (UMVUE) of g(θ), then it is ***unique****.*

2. **Rao-Blackwell**: (**conditioning an unbiased estimator by a sufficient stat always leads to a better statistic** → So, if we're interested in unbiased statistics, we need only look for unbiased statistics that are functions of sufficient statistics!)
*Let h(**X**) be any unbiased estimator of* g(θ), *and let T(**X**) be a sufficient statistic for θ. Define $\phi(T) = \mathrm{E}_\theta[h(X) \mid T] = \mathrm{E}[h(X) \mid T]$ by suff of T*
Then, $\boxed{E_\theta[\phi(T)] = g(\theta) \text{ and } Var_\theta(\phi(T)) \le Var_\theta(h(X))}$

3. **Theorem**: (Unbiased Estimator that is a function of a complete statistic is unique!)
*Suppose that $\phi(T)$ is an unbiased estimator of g(θ) and is a function of a complete statistic T(**X**). All other unbiased estimators that are functions of the (same) complete statistic are equal almost everywhere (i.e. except where the probability measure is 0)*

4. **Lehman-Scheffe**: (An **unbiased estimator that is a function of a complete and sufficient statistic is the unique UMVUE!**)
*Let T(**X**) be a complete sufficient statistic for a parameter θ and $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the* ***unique best unbiased estimator of its expected value (i.e. it's UMVUE of its expected value)****.*

Variance Bounds for Unbiased Estimators[5]: Fisher Information and Cramer-Rao Lower Bound

Fisher Information Matrix for Parameter θ (from a single data point *x* drawn from the distribution p(x, θ))

$$\boxed{I_1(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p(x_i,\theta)\right)\left(\frac{\partial}{\partial\theta}\log p(x_i,\theta)\right)'\right] = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta\,\partial\theta'}\log p(x_i,\theta)\right]} \text{ (6)}$$

1. **Cramer-Rao Inequality (Finite/Asymptotic Min Variance of Unbiased Estimators of linear combination of** θ, $\psi(\theta)$, **is given by the CRLB**.
*Suppose that T(**X**) is a real valued statistic and* $E_\theta[T(X)] = \psi(\theta)_{1x1}$ *and let* $\frac{\partial\psi(\theta)}{\partial\theta}$ *denote the dx1 vector of partial derivatives and suppose that* $I(\theta)^{-1}$ *is nonsingular and the conditions of the above theorem hold. Then, for all* $\theta \in \Theta$,

$$Var(T(X)) \ge \frac{\partial\psi(\theta)}{\partial\theta}'\, I(\theta)^{-1} \frac{\partial\psi(\theta)}{\partial\theta}' = \frac{\partial\psi(\theta)}{\partial\theta}'\, \frac{I_1(\theta)^{-1}}{n} \frac{\partial\psi(\theta)}{\partial\theta}' \quad \text{if } X_i \text{ i.i.d. } p(\cdot,\theta)\ ^{7}$$

(To get asymptotic CRLB, use $I_1^{-1}(\theta)$ if iid**)**

Note: Even if we have a UMVUE, it might not necessarily reach CRLB. But, an unbiased estimator that attains CRLB is UMVUE.

Note2: For a family $\{P_\theta : \theta \in \Theta\}$, *where* θ_{nx1} (*vector of nx*1 *unknown parameters*). $I^{-1}(\theta)$ *will be nxn matrix.*

Note3: if iid, $\frac{I(\theta)}{N} = I_1(\theta) \Rightarrow I^{-1}(\theta) = \frac{I_1^{-1}(\theta)}{N}$

[5] Sometimes finding UMVUE may be difficult. But if we can show that an unbiased estimator attains the CRLB, then we KNOW it is UMVUE.
[6] The RHS equality is only true under regularity conditions (See Note2 page. 30 and 31)
[7] If X_i iid sample where $X_1 \sim p(\cdot,\theta)$, then the model of the data X = (X_1,…,Xn) has information matrix I(θ) = nI_1(θ) → I^{-1}(θ)=I_1(θ) / n

TO GET CRLB, FIND THE LOG-LIKELIHOOD EXPRESSION AND TAKE SECOND DERIVATIVES!

Example 1: Gaussian Regression Model

$$y \mid X \sim N(X\beta,\sigma^2 I) \Rightarrow p(y \mid X;\beta,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2}(y-X\beta)'(\sigma^2 I)^{-1}(y-X\beta)\right\} = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right\}$$

$$\Rightarrow \log p(y \mid X;b,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y-Xb)'(y-Xb)$$

Here, $\theta = (\beta,\sigma^2)'_{k+1x1} \Rightarrow \frac{\partial^2}{\partial\theta\partial\theta'}\log p(y \mid X;\theta) = \begin{bmatrix} \underset{(KxK)}{\frac{\partial^2 \log p(y \mid X;\theta)}{\partial\beta\partial\beta'}} & \underset{(Kx1)}{\frac{\partial \log p(y \mid X;\theta)}{\partial\beta\partial\sigma^{2,}}} \\ \underset{(1XK)}{\frac{\partial \log p(y \mid X;\theta)}{\partial\sigma^2\partial\beta'}} & \underset{(1x1)}{\frac{\partial \log p(y \mid X;\theta)}{\partial\sigma^2\partial\sigma^{2,}}} \end{bmatrix}_{(K+1xK+1)}$

$$\frac{\partial^2 \log p(y \mid X;\theta)}{\partial\beta\partial\beta'} = \frac{\partial}{\partial\beta'}\left(\frac{1}{\sigma^2}X'(y-X\beta)\right) = -\frac{1}{\sigma^2}X'X \Rightarrow -E\left(-\frac{1}{\sigma^2}X'X\right) = \frac{1}{\sigma^2}X'X$$

$$\frac{\partial^2 \log p(y \mid X;\theta)}{\partial\beta\partial\sigma^2} = \frac{\partial}{\partial\sigma^2}\left(\frac{1}{\sigma^2}X'(y-X\beta)\right) = -\frac{1}{\sigma^4}X'(y-X\beta) \Rightarrow -E\left(-\frac{1}{\sigma^4}X'(y-X\beta)\right) = \frac{1}{\sigma^4}X'E(y-X\beta) = 0$$

$$\frac{\partial^2 \log p(y \mid X;\theta)}{\partial\sigma^2\partial\sigma^{2,}} = \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y-X\beta)'(y-X\beta)\right) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y-X\beta)'(y-X\beta) \Rightarrow -E\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y-X\beta)'(y-X\beta)\right) = \frac{n}{2\sigma^4}$$

$$\therefore I(\theta)^{-1} = \begin{bmatrix} \frac{1}{\sigma^2}(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Example 2: Univarite Normal RV's

$$x_i \sim N(\mu,\sigma^2) \Rightarrow p(x_i;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}$$

$$\Rightarrow \log p(x_i;\mu,\sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x_i-\mu)^2$$

$$\theta = (\mu,\sigma^2),\ \frac{\partial^2}{\partial\theta\partial\theta'}\log p(x_i;\theta) = \begin{bmatrix} \frac{\partial^2 \log p(x_i;\theta)}{\partial\mu^2} & \frac{\partial \log p(x_i;\theta)}{\partial\mu\partial\sigma^2} \\ \frac{\partial \log p(x_i;\theta)}{\partial\mu\partial\sigma^2} & \frac{\partial \log p(x_i;\theta)}{\partial\sigma^2\partial\sigma^2} \end{bmatrix}$$

$$\frac{\partial \log p(x_i;\theta)}{\partial\mu} = \frac{1}{\sigma^2}(x_i-\mu)$$

$$\frac{\partial^2 \log p(x_i;\theta)}{\partial\mu^2} = -\frac{1}{\sigma^2} \Rightarrow E\left(\frac{\partial \log p(x_i;\theta)}{\partial\mu^2}\right) = -\frac{1}{\sigma^2}$$

$$\frac{\partial^2 \log p(x_i;\theta)}{\partial\mu\partial\sigma^2} = -\frac{1}{\sigma^4}(x_i-\mu) \Rightarrow E\left(\frac{\partial^2 \log p(x_i;\theta)}{\partial\mu\partial\sigma^2}\right) = 0$$

$$\frac{\partial \log p(x_i;\theta)}{\partial\sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x_i-\mu)^2$$

$$\frac{\partial^2 \log p(x_i;\theta)}{\partial\left(\sigma^2\right)^2} = \frac{1}{2\sigma^4} - \frac{1}{4\sigma^6}(x_i-\mu)^2 \Rightarrow E\left(\frac{\partial^2 \log p(x_i;\theta)}{\partial\left(\sigma^2\right)^2}\right) = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}E\left[(x_i-\mu)^2\right] = -\frac{1}{2\sigma^4}$$

$$I_1(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log p(x_i;\theta)\right] = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \Rightarrow I_1^{-1}(\theta) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \text{ and } I^{-1}(\theta) = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}$$

So, for any unbiased estimator of μ, *i.e.* $\psi(\theta) = (1,0)'$, *has* min var *of* $\begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{\sigma^2}{N}$

MLE:

1. **Setup:**

Suppose a iid sample of Yi's are drawn from a parametric family $\{P_\theta : \theta \in \Theta\}$ of distributions on the support of Y (whose prob. on Y is > 0)

Suppose further that the data is generated from some **true** θ_0.

Estimate θ_0 by the value in Θ (i.e. R^n) that maximizes the (log) likelihood function

That is, we can characterize θ_0 as the θ s.t.

$Q(P_\theta, \theta) = \theta - \arg\max_{b\in\Theta} E_{P_\theta}[\log p(Y,b)] = 0$

$\Rightarrow \quad \theta_0 = \arg\max_{b\in\Theta} E_{P_\theta}[\log p(Y,b)]$ (by analogy principle)

2. **Method:**

$$\hat{\theta}_0 = \arg\max_{b\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p(y_i, b)$$

3. **Identification: Show unique maximization of** $E_{P_\theta}(\log p(Y,b))$ at b = θ_0 **(the true parameter value)**

Under what circumstances is $E_{P_\theta}(\log p(Y,b))$ uniquely maximized at b = θ_0?

Kullback-Kiebler Divergence/Information Criterion must be satisfied! (Note 3 pg. 6)

$$K(b,\theta) = E_\theta\left[\log\left(\frac{p(Y,\theta)}{p(Y,b)}\right)\right] > 0 \quad \forall b \neq \theta$$

(See 270B HW2 #6)
(Notes 2 p. 6)

4. **Invariance Property of MLE**

Suppose we observe an iid sample from a parametric model $\{P_\theta : \theta \in \Theta\}$ and we want to estimate a parameter $\lambda_0 = h(\theta_0)$. If $\hat{\theta}_{MLE}$ is MLE of θ_0, then $h(\hat{\theta}_{MLE})$ is MLE of λ_0.

5. **Consistency of MLE**

In MLE, the log-likelihood functions will be concave (at least after re-parameterization), and therefore it is easiest to apply the second consistency theorem. **See above notes on consistency of M-Estimators**.

Under appropriate smoothness conditions on f, MLE from an iid sample is a consistent estimator.
If we have explicit form of the MLE, it is sufficient to show Bias → 0 and Variance → 0. [8]

6. **Asymptotic Normality of MLE**

Theorem 16: (Note 3 page 21)
Under "regularity conditions" (See above results on M-Estimators), a consistent MLE for θ is asymptotically normal with mean θ and variance $I_1(\theta)^{-1}$. (Note: These results are same as M-Est results, but with MLE (under regularity), the Hessian and the variance of the score are both Fisher Information![9]

$$\sqrt{n}\left(\hat{\theta}_{n,MLE} - \theta\right) \xrightarrow{D} N(0, I_1(\theta)^{-1})$$

7. **Asymptotic Efficiency of MLE**

Under regularity conditions, we showed that Avar($\hat{\theta}_{n,MLE}$) = $I_1(\theta)^{-1}$.

It can be shown that a GMM estimator of θ, $\hat{\theta}_{n,GMM}$, based on some moment conditions such that $E[m(W_i,\theta)] = 0$, we can show that :

$$\text{Avar}(\hat{\theta}_{n,GMM}) \geq I_1(\theta)^{-1} = \text{Avar}(\hat{\theta}_{n,MLE})$$

The MLE estimator is asymptotically at least as efficient as ANY GMM estimator for θ. And since all estimators are GMM estimators (when viewed as solutions to FOC's being the moment conditions), this implies that MLE is at least as efficient as any other M-Estimator.

MLE does not always exist and MLE need not be unique
MLE need not be unbiased or consistent (remember the MLE for Unif(0,b))

---

[8] **How to Show Consistency** (i.e. $P(|Y_n - \mu| > \varepsilon) \xrightarrow{P} 0$**?)** By Chebychev we know $P(|Y_n - \mu| > \varepsilon) \leq \frac{E[(Y_n-\mu)^2]}{\varepsilon^2} = \frac{Var(Y_n-\mu)+[E(Y_n-\mu)]^2}{\varepsilon^2} = \frac{Var(Y_n)+Bias^2}{\varepsilon^2}$

→ **Show Var(Y_n) → 0 and Bias → 0** (sufficient but not necessary)

[9] $-E[H(W,\theta)] = E[s(W,\theta)s(W,\theta)']$ *bc* $-E\left[\frac{\partial^2 \log p(Y_i \mid Z_i,\theta)}{\partial b\,\partial b'}\right] = E\left[\frac{\partial \log p(Y_i \mid Z_i,\theta)}{\partial b}\frac{\partial \log p(Y_i \mid Z_i,\theta)}{\partial b}'\right] = I_1(\theta)$

Conditional MLE:

Idea: **As long as we can separate f(y,z) = f(y|z)f(z), then all we care about is f(y|z). (Assuming that the parameters of interest do not affect f(z)!) Then, when we conditional on Z's we can treat them as constants. and the conditional and unconditional MLE's are the same (since f(z) does not contain the parameters of interest).**

Explanation:

Previously, in the Gaussion Linear Regression Model and Binary Choice Logit Model (HW1) we assumed nonstochastic regressors, i.e. regressors are **fixed known constants and only y was stochastic**. This allowed us to get nice results, i.e. these models belonged to an exponential family.

e.g. Consider Yi independently distributed, Yi ~N(m_i, s^2) where $m_i = b_1 + b_2 z_i$, and z_i a sequence of known constants.
We can show here that for the whole data **Y**, p(Y| b) is in the exponential family. (HW1 #3)

$$\text{Log } p(\mathbf{Y}; b) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - b_1 - b_2 Z_i)^2$$

However, in economics these models are not applicable! In experimental settings the assumption of fixed regressors might be reasonable because the researcher can choose the regressors ahead of time (i.e. how much dosage to give and then observe the response y); however, in economics data this is not true.

If, however, we interpret economics models as being conditional on the realized values of the regressors (i.e. p(y,b) = p(y|x, b)), the results still go through! This is the idea of CMLE. Once we have stochastic regressors, we formulate the parametric model such that instead of having to specify the joint distribution of the data, p(y,z | b), we specify the distribution of the outcome conditional on the covariates, **i.e. we specify p(y|z; b) → this is the distribution we care about anyways!** We make the additional assumption (usually) that the distribution of z is not a function of the parameters of interest, so the unconditional and conditional MLE estimates will be the same.

Remember, in the non-stochastic case, we only have to make parametric assumptions about p(y|b) where b is function of z's → here, the idea is similar, we make assumptions about p(y|z, b).

e.g. Suppose we observe i.i.d. $W_i = \{Y_i, Z_{i\,(dx1)}\}$ and we specify the conditional distribution Y|Z ~ N(b'Z, s^2) : TREAT Z's as constants

$$\text{Log } p(\mathbf{Y}\,|\mathbf{Z}; b) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - b' Z_i)^2$$

In this formulation, we get the same results as before!

Hypothesis Testing

Basic Set-Up: Statistician observes data **X** from a distribution $P \in \{P_\theta : \theta \in \Theta\}$ and must decide whether $\theta \in \Theta_H$ *or* $\theta \in \Theta_K$ $(\Theta_H \cap \Theta_K = 0)$.

Statistician then chooses a decision rule / statistic $\delta(.) : \aleph \to A = \{0,1\}$ to either accept or reject the null hypothesis $\theta \in \Theta_H$.

Concepts and Definitions:

1. **Simple vs. Composite Hypothesis:**
   Two structural possibilities for Θ_H and Θ_K : if Θ_H singleton, then we call Θ_H and H "**Simple**" (similarly for K)
   if Θ_H contains more than 1 element, then we call Θ_H and H "**Composite**" (similarly for K)
2. **Test:** $\delta_c(\mathbf{X}) = I(T(\mathbf{X}) > c)$ (where f(X) is a function of the observed data)
   Reject the null if $\delta_c(\mathbf{X}) = I(T(\mathbf{X}) > c) = 1$, **Do not reject the null** if $\delta_c(\mathbf{X}) = I(T(\mathbf{X}) > c) = 0$

   Ex: *Let* $M_n = Max(y_1,...,y_N)$, $\phi_c(Y_1,...,Y_N) = I(M_n > c)$

3. **Power Function of the test** $\delta(.)$**:** $\beta_{\delta_c}(\theta) \equiv P_{\theta_0}(\delta_c(\mathbf{X}) = 1)$ **(i.e. probability of rejection the null when truth is** θ_0**)**
   (so, if θ_0 is the null, then we call it type I error, if θ_0 is alternative then it's power of test)

Intuition: The power function says, based on this test, what is the probability that the test will "reject" if the truth is at θ_0.

If $H_0 : \theta = \theta_0$, then $\beta_{\delta_c}(\theta_0)$ is the **probability of a Type I error**

If $H_K : \theta = \theta_0$, then $\beta_{\delta_c}(\theta_0)$ is the **probability of rejecting correctly**

Ex: From above, $\beta_{\phi_c}(\theta) \equiv P(\phi_c = 1) = P(M_n > c) = P(Max_i y_i > c) = 1 - P(Max_i y_i < c) = 1 - \left(\frac{c}{\theta}\right)^N$ (*assume* y_i *iid Uni*$[0,\theta]$)

Power of the Test: $\beta_{\delta_c}(\theta) \equiv P(\delta_c(\mathbf{X}) = 1)$ *when* $\theta_0 \in \Theta_k$ (Probability of rejecting the null correctly)

Size of the Test: $Sup_{\theta \in \Theta_0} \beta_{\delta_c}(\theta)$ (Largest probability of rejecting the null incorrectly, largest probability of committing Type I error).

Power + Size = Probability of rejecting the null using the test δ(X).

Level of a Test: $\phi_c(\mathbf{X})$ *is a level* α *test* $\Leftrightarrow E_\theta[\phi_c(\mathbf{X})] = P(\phi_c(\mathbf{X}) = 1) \le \alpha$ *for* $\theta \in \Theta_H$ (Type I error rate $\le$ α: Size is at most α)

Power function, Size of Test, and Power of Test:

1. Define the null and alternative hypotheses and form a test.
   Test is: Reject the null if $T(X) > c$ and do not reject the null if $T(X) < c$
   Based on this test, we calculate a power function.
2. To find size of the test, see the largest value of the power function over the range of $\theta \in \Theta_0$
3. To find power of the test for a particular alternative, θ', find $\beta_{\delta_c}(\theta')$ **(i.e. what is the probability of rejecting the null** correctly when the truth is $\theta' \in \Theta_K$.

Graph

P-Value:
Observed size or significance probability of the test. P-Value is defined as the smallest level of significance at which a researcher using the statistic T would reject the null.

Note: It is characteristic for tests that satisfy the monotone likelihood ratio property that the power function is increasing over the parameter space.

Neyman-Pearson Lemma

Motivation: Since we cannot simultaneously reduce Type I and Type II errors, our goal is to fix maximum Type I error rate and find the test of the highest power. (i.e. maximize power subject to size constraint)

For testing simple null vs. simple alternatives, NP Lemma provides a solution to the problem of **finding the level α test that has the highest possible power**.

NP Theorem: (LR Tests are MP Level α Test for testing simple null vs. simple alternative)

Consider testing H: $\theta = \theta_0$ against K: $\theta = \theta_1$ where the pdf or pmf corresponding to θ_i is p(x, θ_i) (i.e. p(x, θ_1) is the prob of observing the data/test statistic value, given that the "truth" is θ_1). Consider the LR test function (we reject if LR >k, don't reject if LR <k, indifferent if LR = k):

$$\phi_k(x) = \begin{cases} 1 & if \ \dfrac{p(x,\theta_1)}{p(x,\theta_0)} > k \\ 0 & if \ \dfrac{p(x,\theta_1)}{p(x,\theta_0)} < k \quad (^{10}) \\ any\ value\ in\ (0,1) & if \ \dfrac{p(x,\theta_1)}{p(x,\theta_0)} = k \end{cases}$$

(we call $\frac{p(x,\theta_1)}{p(x,\theta_0)}$ the **simple** LR statistic. Used for 1-D Parameter cases.)

Then

1. *If $\alpha > 0$ and ϕ_k is a level α test, i.e. $E_\theta(\phi_k(x)) \le \alpha$, then ϕ_k is **the Most Powerful** test in the class of level α tests.*
2. ***For each** $\alpha \in [0,1]$ **there exists a MP size α LR test** of the form $\phi_k(x)$ <u>provided that randomization is permitted</u> (i.e. 0< $\phi_k(x)$ <1 for some **x**)*
3. *If a test $\tilde{\phi}$ is a MP level α test, then it must be a level α LR Test. That is, there exists k s.t. for $\theta \in \{\theta_0, \theta_1\}$, $P_\theta(\tilde{\phi}(x) \neq \phi_k(x)) = 0$*

➔ ***(2&3): Most Powerful Level α Tests are LR Tests (for testing simple null vs. simple alternative)!***

Neyman Pearson Test

Procedure for constructing a level α test Neyman Pearson Test (simple vs. simple)

1. Figure out all the possible outcomes
2. Calculate likelihood under the null and under the alternative for each outcome, and find LR = f(alternative)/f(null)
3. Order the outcomes in terms of smallest LR to largest LR (largest LR most unlikely).
4. Pick the ones with biggest LR to be in the rejection region – i.e. ones that are more likely under alternative than the null. Pick them such that their probabilities under the null sum up to α.
   (i.e. the likelihood we reject if null is true is at most α)
5. Translate the test in terms of the likelihood ratios (i.e. we should be able to say: reject if lr > k).
6. Power of the test? Sum up the corresponding probabilities of the rejection region under the alternative.

Example:

Suppose $\{x_i\}_{i=1}^{10} \sim Binomial(10,\theta)$.

$H_0 : \theta = .3 \quad H_a : \theta = .5$

Outcomes/# of "Successes"	1	2	3	4	5	6	7	8	9	10	
Likelihood Under Alternative: p = 0.5	0.0098	0.0439	0.1172	0.2051	0.2461	**0.2051**	**0.1172**	**0.0439**	**0.0098**	**0.0010**	**0.3770** Power
Likelihood Under Null: p = 0.3	0.1211	0.2335	0.2668	0.2001	0.1029	**0.0368**	**0.0090**	**0.0014**	**0.0001**	**0.0000**	**0.0473** Size
Simple LR	0.0807	0.1882	0.4392	1.0248	2.3911	**5.5793**	**13.0184**	**30.3762**	**70.8779**	**165.3817**	

[10] The last case only applies only in the discrete case (if x discrete). If x continuous, then we only have LR≥k or LR<k

Monotone Likelihood Ratio Models (For 1-models with 1-dimensional parameter)

Motivation: The above theorem establishes the "optimality" of LR statistic for testing simple null vs. simple alternative.

Clearly, **any strictly increasing function of an optimal statistic is also optimal**.

Another way to look at this is, **T(X) is equivalent to LR statistic as long as LR is a monotone function of T(X)**.

Distributions with this property are "Monotone Likelihood Ratio Models": e.g. Exponential Family

Example1: Consider $X = (X_1,\ldots,X_n)$ iid from the parametric family $\left\{N\left(\mu,\sigma^2\right): \mu \in \Re, \sigma^2 \text{ known}\right\}$

$H_0 : \mu = 0 \quad v.s. \; H_a : \mu = v > 0$

$$LR = \frac{p(\mathbf{X} \mid v,\sigma^2)}{p(\mathbf{X} \mid \mu,\sigma^2)} = \frac{\left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - v)^2\right\}}{\left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - 0)^2\right\}} = \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - v)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i)^2\right\}} = \exp\left\{\frac{1}{2\sigma^2}\left(\sum_i {x_i}^2 - (x_i - v)^2\right)\right\}$$

$$= \exp\left\{\frac{1}{2\sigma^2}\left(\sum_i {x_i}^2 - \left({x_i}^2 - 2x_i v + v^2\right)\right)\right\} = \exp\left\{\frac{1}{2\sigma^2}\left(\sum_i 2x_i v - v^2\right)\right\} = \exp\left\{-\frac{nv^2}{2\sigma^2} + \frac{v}{\sigma^2}\sum_i x_i\right\}$$

$$= \exp\left\{-\frac{nv^2}{2\sigma^2} + \frac{vn}{\sigma^2}\left(\frac{1}{n}\sum_i x_i\right)\right\} = h(\bar{X}) \quad \textit{for } h \textit{ strictly increase}$$

LR strictly increase in $\bar{X}$, so LR and $\bar{X}$ will generate the same family of critical regions!
Re *ject for big values of $\bar{X}$!*

To construct a level α test, find c such that $P_0\left(\bar{\mathbf{X}} > c\right) = \alpha$ (*under the null distribution*)

MLR Family: T(X) Equivalent to LR

Def: *The family of models $\{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \Re$* **(i.e. 1-dimentsional parameter)** *is said to be a* ***monotone likelihood ratio family*** *if the distributions P_{θ_0} and P_{θ_1} are distinct and the ratio $\frac{P(x,\theta_1)}{P(x,\theta_0)}$ is a monotone function of some function/statistic T(X) of the data in the same direction for all pairs of θ_0 and θ_1.*

If, for $\theta_1 > \theta_0$ the ratio is an increasing function of the statistic T(X), the family is said to have increasing MLR in T(X) [or decreasing in –T(X)]
the ratio is a decreasing function of the statistic T(X), the family is said to have decreasing MLR in T(X) [or decreasing in –T(X)]

Prop: 1-Parameter Exponential Family is MLR in T(X), the Sufficient Statistic

One – parameter exponential: $p(x,\theta) = h(x)\exp\{\eta(\theta)T(x) - B(\theta)\}$: $\frac{p(x,\theta_1)}{p(x,\theta_0)} = \exp\{(\eta(\theta_1) - \eta(\theta_0))T(x) - (B(\theta_1) - B(\theta_0))\}$

→ LR is strictly increasing in $T(x)$

Thus, the exponential family (with 1 parameter) is MLR Increasing/Decreasing in T(x) – the sufficient statistic!

MLR: Finding LR-Equivalent Tests that are UMP for testing 1-sided null vs. 1-sided alternative:

Motivation: Having MLR property allows us to find equivalent tests that are equivalent to the LR test that have the additional property of being UMP for testing 1-sided/simple null versus 1-sided alternative hypothesis (since the test T(x) does not depend on the alternative θ_1 but the LR statistic $L(x,\theta_0,\theta_1)$ does!) See following theorem for the result.

Theorem (MLR in T(X) and UMP):

Suppose that the parametric family $\{P_\theta : \theta \in \Theta\}$ *with 1-dimensional parameter* (i.e. $\Theta \subseteq \Re$) *is MLR increasing in T(**X**).*
Then,

1. $For\ each\ t \in (0,\infty), the\ power\ function\ \beta_{\delta_t}(\theta)$ *is increasing in* θ $for\ all\ \theta \in \Theta$
2. $If\ E_0[\delta_t(x)] = \alpha > 0\ then\ \delta_t(.)\ is\ UMP\ level\ \alpha\ for\ testing\ H: \theta \le \theta_0\ vs.\ K: \theta > \theta_0$

Similarly, that the parametric family $\{P_\theta : \theta \in \Theta\}$ *with 1-dimensional parameter* (i.e. $\Theta \subseteq \Re$) *is MLR decreasing in R(**X**).*
Then,

1. $For\ each\ t \in (0,\infty), the\ power\ function\ \beta_{\delta_t}(\theta)$ *is decreasing in* θ $for\ all\ \theta \in \Theta$
2. $If\ E_0[\delta_t(x)] = \alpha > 0\ then\ \delta_t(.)\ is\ UMP\ level\ \alpha\ for\ testing\ H: \theta \ge \theta_0\ vs.\ K: \theta < \theta_0$

Why Most Powerful for Testing 1-Sided vs. 1-Sided?

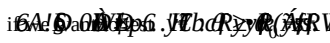
Consider the problem of testing a simple H: $\theta = \theta_0$ against K: $\theta = \theta_1$ where $\theta_1 > \theta_0$

Suppose a parametric family $\{P_\theta : \theta \in \Theta\}$ family is with 1-dimensional parameter (i.e. $\Theta \subseteq \Re$) is increasing MLR in T(**X**). Then, the LR statistic $L(x,\theta_0,\theta_1)$ will be some function h(T(**X**)) for some increasing function h(.). Therefore, the test $\delta_t(.)$ will be equivalent to the likelihood ratio test at level $E_{\theta_0}[\delta_t(X)]$, and therefore by Neyman-Pearson, will be Most Powerful against the alternative $\theta = \theta_1$ (where $\theta_1 > \theta_0$) (and since it's a simple alternative, the test will be trivially UMP).

Since the test T(**X**) does not depend on θ_1 (bc your decision rules are always the same no matter what the alternative is, as long as $\theta_1 > \theta_0$), the test $\delta_t(.)$ will in fact be **Uniformly Most Powerful for testing H against K:** $\theta > \theta_0$

Usefulness:

So, if we want to test $H: \theta \le \theta_0$ *vs.* $K: \theta > \theta_0$, find T(**X**) such that *P* is MLR increasing in T(**X**) → T(X) is UMP level $E_0[\delta_t(x)]$ test.



(Generalized) Likelihood Ratio Test (applies to composite nulls or composite alternatives and multidimensional unknown parameters): 2-Sided Tests

Why Useful: So far we've only considered one-parameter problems and have found that even in such restricted settings there are often no UMP tests. **LR tests are intuitive and efficient procedures that can be used when no optimal methods are available**, and are quite natural for **multidimensional parameters**.

The generalized LR test is used in situations in **which the hypotheses are not simple**. Such tests are not generally optimal, but they are typically non-optimal in situations for which no optimal test exists, and they usually perform reasonably well. Generalized likelihood ratio tests have wide utility, and play the same role in testing as MLE does in estimation.

Define[11]**:** $\Lambda = \dfrac{\sup_{\theta\in\Theta_M} p(X;\theta)}{\sup_{\theta\in\Theta_H} p(X;\theta)}$ *where* $\Theta_M = \Theta_H \cup \Theta_K$

Procedure:

1. Calculate the unrestricted/unconstrained estimator: MLE $\hat{\theta}$ over the entire parameter space $\Theta_M = \Theta_H \cup \Theta_K$ (the maintained hypothesis, to be more precise)**.**
2. Calculate the restricted/constrained estimator: MLE $\tilde{\theta}$ over the parameter space Θ_H
3. Form Λ → We want to reject for large values of Λ
4. Find a function that is strictly increasing on the range of Λ such that h(Λ (**X**)) has a simple form and a tabled distribution under H. **(i.e. the "null distribution")**
   Because h(Λ (**X**)) is equivalent to Λ (**X**), we're effectively specifying a LR test through the test statistic h(Λ (**X**)).

   Note: Most commonly we choose the statistic **h(Λ (X)) = 2 log Λ (X)** which has an asymptotic chi-squared distribution for a certain class of hypotheses.

   Note2: This generalized LR test also gives us an "efficient" way to test 2-sided alternative hypotheses for the case of a one-dimensional parameter family. (The notion of efficiency is asymptotic)

Examples:
Example 1: T-Statistic in Normal Model is equivalent to LR statistic

$\{x_i\}$ *iid* $N(\mu,\sigma^2)$

$H: \mu \le \mu_0 \quad K: \mu > \mu_0$

$$\Lambda = \frac{\sup_{\theta\in\Theta_M} p(\mathbf{x}|\theta)}{\sup_{\theta\in\Theta_H} p(\mathbf{x}|\theta)} \qquad \text{we know } \hat{\mu}_m = \bar{X} \text{ and } \hat{\sigma}_m = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2,\ \hat{\mu}_c = \mu_0 \text{ and } \hat{\sigma}_c = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2$$

$$= \frac{\left[\frac{1}{\sqrt{2\pi}\hat{\sigma}_m}\right]^n \exp\left\{\frac{-1}{2\hat{\sigma}_m^2}\sum(x_i-\bar{x})^2\right\}}{\left[\frac{1}{\sqrt{2\pi}\hat{\sigma}_c}\right]^n \exp\left\{\frac{-1}{2\hat{\sigma}_c^2}\sum(x_i-\bar{x})^2\right\}} = \frac{\left[\frac{1}{\hat{\sigma}_m}\right]^n \exp\left\{\frac{-n}{2}\right\}}{\left[\frac{1}{\hat{\sigma}_c}\right]^n \exp\left\{\frac{-n}{2}\right\}} = \left[\frac{\hat{\sigma}_c^2}{\hat{\sigma}_m^2}\right]^{n/2}$$

$$\hat{\sigma}_c^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i+\bar{x}-\bar{x}-\mu_0)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 - 2(x_i-\bar{x})(\bar{x}-\mu_0)+(\bar{x}-\mu_0)^2 = \hat{\sigma}_m^2 - 2\frac{1}{n}\sum_{i=1}^{n}\left(x_i\bar{x}-\bar{x}^2-\mu_0 x_i+\mu_0\bar{x}\right)+(\bar{x}-\mu_0)^2$$

$$= \hat{\sigma}_m^2 + (\bar{x}-\mu_0)^2$$

$$\Lambda^{2/n} = \left[\frac{\hat{\sigma}_c^2}{\hat{\sigma}_m^2}\right] = \left[\frac{\hat{\sigma}_m^2+(\bar{x}-\mu_0)^2}{\hat{\sigma}_m^2}\right] = \left[1+\frac{(\bar{x}-\mu_0)^2}{\hat{\sigma}_m^2}\right] = \left[1+\frac{(\bar{x}-\mu_0)^2}{\frac{n-1}{n}s^2}\right] = \left[1+\frac{1}{\sqrt{n-1}}\left(\frac{\bar{x}-\mu_0}{s}\right)^2\right] = \left[1+\frac{1}{\sqrt{n-1}}(T_n)^2\right]$$

LR increasing in T_n

---

[11] Whereas previously, we used $LR(x) = \dfrac{\sup_{\theta\in\Theta_K} p(X;\theta)}{\sup_{\theta\in\Theta_H} p(X;\theta)}$ **. However, the distribution properties of LR are complicated so we replace with Λ whose distribution is easier to obtain (after transformation).** This is an inessential change. And, in most of the cases we will consider where $p(\mathbf{x}|\theta)$ is a continuous function of θ and Θ_H is of smaller dimension than $\Theta = \Theta_H \cup \Theta_K$ so that $LR(x) = \Lambda(x)$. (unexplained)

Example 2: Let $X_1,\ldots,X_n$ iid $N(\mu,\sigma^2)$ σ *known* . Derive a test statistic to test $H_0:\mu=0$ *vs.* $H_A:\mu\neq 0$

Here, $\hat{\mu}_c=\bar{X}$ *and* $\hat{\mu}_u=0$ $(\text{since } H_0:\mu=0)$

$$\Lambda=\frac{p(\mathbf{X}\mid\hat{\mu}_u,\sigma^2)}{p(\mathbf{X}\mid\hat{\mu}_c,\sigma^2)}=\frac{\left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N\exp\left\{-\frac{1}{2\sigma^2}\sum_i\left(x_i-\bar{\mathbf{x}}\right)^2\right\}}{\left[\frac{1}{\sqrt{2\pi}\sigma}\right]^N\exp\left\{-\frac{1}{2\sigma^2}\sum_i\left(x_i-0\right)^2\right\}}=\exp\left\{\frac{1}{2\sigma^2}\left(\sum_i {x_i}^2-\left({x_i}^2-2x_i\bar{\mathbf{x}}+\bar{\mathbf{x}}^2\right)\right)\right\}$$

$$=\exp\left\{\frac{1}{2\sigma^2}\left(\sum_i 2x_i\bar{\mathbf{x}}-\bar{\mathbf{x}}^2\right)\right\}=\exp\left\{-\frac{n\bar{\mathbf{x}}^2}{2\sigma^2}+\frac{\bar{\mathbf{x}}}{\sigma^2}\sum_i x_i\right\}=\exp\left\{-\frac{n\bar{\mathbf{x}}^2}{2\sigma^2}+\frac{n\bar{\mathbf{x}}^2}{\sigma^2}\right\}=\exp\left\{\frac{n\bar{\mathbf{x}}^2}{\sigma^2}\right\}$$

Λ *strictly increase in* $T(\mathbf{X})=\frac{n\bar{\mathbf{x}}^2}{\sigma^2}=\left(\frac{\sqrt{n}\bar{\mathbf{x}}}{\sigma}\right)^2$

Under H_0, $\frac{\sqrt{n}\bar{\mathbf{x}}}{\sigma}\sim N(0,1)\Rightarrow\left(\frac{\sqrt{n}\bar{\mathbf{x}}}{\sigma}\right)^2\sim ChiSq(1)$ *by CMT*

Thus, pick c s.t. $P_{\theta_0}\left(\frac{\sqrt{n}\bar{\mathbf{x}}}{\sigma}>c\right)\leq\alpha$

Example 3: Let $X_1,\ldots, X_n$ be a random sample from an exponential distribution with density function $f(x\mid\theta)=\theta\exp(-\theta x)$.

Derive a likelihood ratio test for $H_0:\theta=\theta_0$ *vs.* $H_A:\theta\neq\theta_0$ and show that the rejection region is of the form $\{x:\bar{x}\exp(-\theta_0\bar{x})\leq c\}$

Here, $\hat{\theta}_c=\theta_0$

What is $\hat{\theta}_u$?

$$P(X\mid\theta)=\theta^n\exp\left(-\theta\sum_i x_i\right)\Rightarrow\ln p(X\mid\theta)=n\log\theta-\theta\sum_i x_i\Rightarrow FOC:\frac{n}{\theta}-\sum_i x_i=0$$

$$\hat{\theta}_u=\left(\bar{\mathbf{x}}\right)^{-1}$$

$$\Lambda=\frac{p(\mathbf{X}\mid\hat{\theta}_u)}{p(\mathbf{X}\mid\hat{\theta}_c)}=\frac{\left(\bar{\mathbf{x}}\right)^{-n}\exp\left(-\frac{1}{\bar{\mathbf{x}}}\sum_i x_i\right)}{\theta_0^n\exp\left(-\theta_0\sum_i x_i\right)}=\frac{1}{\bar{\mathbf{x}}^n\theta_0^n}\frac{\exp(-n)}{\exp\left(-\theta_0\sum_i x_i\right)}=\left(\bar{\mathbf{x}}\theta_0\right)^{-n}\exp\left(\theta_0\sum_i x_i-n\right)=\left(\bar{\mathbf{x}}\theta_0\right)^{-n}\exp\left(n\left(\theta_0\bar{\mathbf{x}}-1\right)\right)$$

$$=\left[\frac{1}{\bar{\mathbf{x}}\theta_0}\exp\left(\theta_0\bar{\mathbf{x}}-1\right)\right]^n$$

Re *ject for* $\Lambda=\left[\frac{1}{\bar{\mathbf{x}}\theta_0}\exp\left(\theta_0\bar{\mathbf{x}}-1\right)\right]^n>c_1$ *for some* c_1 *that gives us desired* α

$$\Leftrightarrow\left[\bar{\mathbf{x}}\theta_0\exp\left(1-\theta_0\bar{\mathbf{x}}\right)\right]^n<\frac{1}{c_1}\Leftrightarrow\left[\bar{\mathbf{x}}\theta_0\exp\left(1-\theta_0\bar{\mathbf{x}}\right)\right]<\left(\frac{1}{c_1}\right)^{1/n}\Leftrightarrow\left[\bar{\mathbf{x}}\exp\left(1-\theta_0\bar{\mathbf{x}}\right)\right]<\frac{1}{\theta_0}\left(\frac{1}{c_1}\right)^{1/n}$$

$$\Leftrightarrow\left[\bar{\mathbf{x}}\exp\left(-\theta_0\bar{\mathbf{x}}\right)\right]<\frac{1}{e\theta_0}\left(\frac{1}{c_1}\right)^{1/n}$$

Hypothesis Testing in Large Samples

Concepts and Definitions

Setup: We observe data **X** from *P* and want to test H: $P\ H: P \in P_0$ *vs.* $K: P \in P_1$

- **Asymptotic Significance Level:** If $\lim Sup_{n\to\infty} P_P\left(\delta_n(\mathbf{X})=1\right) \le \alpha$ *for all* $P \in P_0$, then the test $\delta_n(.)$ is said to have asymptotic significance level α.
- **Limiting Size**: If $\lim_{n\to\infty} Sup_{P\in P_0} P_P\left(\delta_n(\mathbf{X})=1\right)$ exists, then it is called the limiting size of $\delta_n(.)$.
- **Consistency**: The test $\delta_n(.)$ is consistent if $\lim_{n\to\infty} P_P\left(\delta_n(\mathbf{X})=1\right)=1$ *for all* $P \in P_1$

  (If we had infinite data, we would reject the null with probability 1 if truth is the alternative!)

Example1: Parametric Example

Suppose $\{X_i\}$ *iid* $N(\theta,\sigma^2)$, σ^2 *KNOWN*

$H: \theta = \theta_0$ *vs.* $K: \theta > \theta_0$

$$\textit{Test Statistic}:\ \delta_n(\mathbf{X}) = I\left(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) \quad \left[\textit{why?}\ \frac{\bar{X}-\theta_0}{\sigma/\sqrt{N}} \sim N(0,1).\ \text{Re ject if}\ \frac{\bar{X}-\theta_0}{\sigma/\sqrt{N}} > z_{1-\alpha} \Leftrightarrow I\left(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right)\right]$$

$$\textit{Size of Test}: \beta_{\delta_n}(\theta_0) = P\left(\delta_n(\mathbf{X})=1\right) = P\left(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) = P\left(\frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right) = \alpha$$

$$\textit{For any } \theta > \theta_0: \beta_{\delta_n}(\theta) = P\left(\delta_n(\mathbf{X})=1\right) = P\left(\bar{X} - \theta > (\theta_0 - \theta) + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) = P\left(\frac{\bar{X}-\theta}{\sigma/\sqrt{n}} > \frac{\sqrt{n}(\theta_0-\theta)}{\sigma} + z_{1-\alpha}\right) = 1 - \Phi\left(\frac{\sqrt{n}(\theta_0-\theta)}{\sigma} + z_{1-\alpha}\right)$$

$$\textit{As } n\to\infty,\ \theta > \theta_0 \Rightarrow \frac{\sqrt{n}(\theta_0-\theta)}{\sigma} + z_{1-\alpha} \to -\infty, \Phi\left(\frac{\sqrt{n}(\theta_0-\theta)}{\sigma} + z_{1-\alpha}\right) \to 0$$

$$\textit{Thus, } \beta_{\delta_n}(\theta) = P\left(\delta_n(\mathbf{X})=1\right) \to 1.$$

Example 2: Non-Parametric Example

Suppose $\{X_i\}$ *iid* P, σ^2 *KNOWN*, $E(P) = \theta_0$. *P is a distribution with finite first* 2 *moments*

$H: \theta = \theta_0$ *vs.* $K: \theta > \theta_0$

$$\textit{Test Statistic}:\ \delta_n(\mathbf{X}) = I\left(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right)$$

$$\textit{Size of Test}: \beta_{\delta_n}(\theta_0) = P\left(\delta_n(\mathbf{X})=1\right) = P\left(\bar{X} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) = P\left(\frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\right)$$

$$\textit{For any } \theta > \theta_0: \beta_{\delta_n}(\theta) = P\left(\delta_n(\mathbf{X})=1\right) = P\left(\bar{X} - \theta > (\theta_0 - \theta) + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\right) = P\left(\sqrt{n}\left(\frac{\bar{X}-\theta}{\sigma}\right) > \frac{\sqrt{n}(\theta_0-\theta)}{\sigma} + z_{1-\alpha}\right)$$

$$\textit{Since first 2 moments exists, we can invoke CLT}: \sqrt{n}\left(\frac{\bar{X}-\theta}{\sigma}\right) \Rightarrow_D N(0,1)$$

$$\textit{Thus, as } n\to\infty, \beta_{\delta_n}(\theta) \to 1 - \Phi\left(\lim \frac{\sqrt{n}(\theta_0-\theta)}{\sigma} + z_{1-\alpha}\right) = 1$$

$$\textit{Thus, } \beta_{\delta_n}(\theta) = P\left(\delta_n(\mathbf{X})=1\right) \to 1.$$

Confidence Region and Confidence Sets