Properties of Random Variables

We have defined Random Variables thoroughly in the Probability Theory notes, and now we will discuss some commonly referenced properties of random variables. The following do not necessarily exist for all random variables, but if they do exist they are defined as follows:

The expected value, also called the mean of a random variable (denoted $E[X]$ or sometimes μ):

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx \text{ if } X \text{ has a density function}$$

$$E[X] = \sum_{i=1}^{\infty} x_i P(X = x_i) \text{ if } X \text{ has no density function}$$

if X has a density function, we often call it <u>continuous</u>, and if not, we often call it <u>discrete</u>.

The variance of X, often denoted $Var(X)$:

$$Var(X) \equiv E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Properties of Expectation and Variance

Linearity: $E[X+Y] = E[X] + E[Y]$

iff X and Y are independent: $Var(X+Y) = Var(X) + Var(Y)$

Multiplication by a scalar: $E[aX] = a\,E[X]$
$Var(aX) = a^2\,Var(X)$, e.g. $Var\left(\frac{\sum X}{\sqrt{n}}\right) = \frac{1}{n} Var(\sum X)$

Adam's Rule: $E[X] = E[E[X|Y]]$

Eve's Rule:
$Var(X) = E[Var(X|Y)] + Var(E[Y|X])$

Jensen's Inequality: $E[g(X)] \geq g(E[X])$
if g is convex

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,dx \quad \text{or}$$

$$\sum_{i=1}^{\infty} g(x)\,P(X=x) \quad \text{if } X \text{ is discrete}$$

Cauchy-Schwartz Inequality: $Cov(X,Y)^2 \leq Var(X)\,Var(Y)$

Linearity of Expectation Example

For $X_1, \ldots, X_n$ iid, prove $E\left[\frac{\bar{X}_{n-1}}{\bar{X}_n}\right] = 1$

$$E\left[\frac{\bar{X}_{n-1}}{\bar{X}_n}\right] = E\left[\frac{\frac{X_1+X_2+\cdots+X_{n-1}}{n-1}}{\frac{X_1+X_2+\cdots+X_n}{n}}\right] = \frac{n}{n-1} E\left[\frac{X_1+\cdots+X_{n-1}}{X_1+\cdots+X_n}\right]$$

$$= \frac{n}{n-1}\left(E\left[\frac{X_1}{X_1+\cdots+X_n}\right] + E\left[\frac{X_2}{X_1+\cdots+X_n}\right] + \cdots + E\left[\frac{X_{n-1}}{X_1+\cdots+X_n}\right]\right)$$

Since $X_1, \ldots, X_{n-1}$ are iid, this is

$$= \frac{n}{n-1}\left(E\left[\frac{X_1}{X_1+\cdots+X_n}\right] + E\left[\frac{X_1}{X_1+\cdots+X_n}\right] + \cdots + E\left[\frac{X_1}{X_1+\cdots+X_n}\right]\right)$$

$$= \frac{n}{n-1} \cdot n-1 \cdot E\left[\frac{X_1}{X_1+\cdots+X_n}\right] = n E\left[\frac{X_1}{X_1+\cdots+X_n}\right] = E\left[\frac{n X_1}{X_1+\cdots+X_n}\right]$$

Again, since $X_1, \ldots, X_n$ are iid, this is

$$= E\left[\frac{X_1+X_1+\cdots+X_1}{X_1+X_2+\cdots+X_n}\right] = \left(E\left[\frac{X_1}{X_1+X_2+\cdots+X_n}\right] + \cdots + E\left[\frac{X_1}{X_1+\cdots+X_n}\right]\right)$$

$$= \left(E\left[\frac{X_1}{X_1+\cdots+X_n}\right] + E\left[\frac{X_2}{X_1+\cdots+X_n}\right] + \cdots + E\left[\frac{X_n}{X_1+\cdots+X_n}\right]\right)$$

$$= E\left[\frac{X_1+X_2+\cdots+X_n}{X_1+X_2+\cdots+X_n}\right] = E[1] = 1$$

Parameters

Often, for random variables with PdF's and/or PMF's or CdF's of the same functional form, we use <u>parameters</u> as placeholders for numbers. For example, the <u>Normal distribution</u> has PdF

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Where μ and σ^2 are the parameters.

Most of <u>Parametric Inference</u> (inference with parameters) is focused on assuming we have some realizations of a random variable with known functional form and unknown parameters and using the realizations to learn about the parameters

For the normal distribution, $E[X] = \mu$ and $Var(X) = \sigma^2$

Common Distributions

Normal Distribution:

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

$$E[X] = \mu \qquad Var(X) = \sigma^2$$

Uniform Distribution:

$$f(x \mid a, b) = \frac{1}{b-a}$$

$$E[X] = \frac{b+a}{2} \qquad Var(X) = \frac{(b-a)^2}{12}$$

Binomial Distribution: distribution of the number of heads of n coin flips

$$P(X = x \mid n, p) = \binom{n}{x} p^x (1-p)^{1-x}$$

$$E[X] = np \qquad Var(X) = np(1-p)$$

Poisson Dist:

$$P(X = x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$$E[X] = \lambda \qquad Var(X) = \lambda$$

Exponential Dist:

$$f(x \mid \beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$$

$$E[X] = \beta \qquad Var(X) = \beta$$

Can also be parameterized equivalently with $\lambda = \frac{1}{\beta}$

Moment Generating Functions

The moment generating function generates the moments! The MGF of a r.v. X is defined as a function of t:

$$M_X(t) = E[e^{tX}]$$

The MGF, like the CDF, uniquely identifies the distribution of a r.v.. The nth moment of X is defined as $E[X^n]$, and it can be generated by taking the nth derivative of $M_X(t)$ w.r.t. t and setting $t=0$:

$$M_X^{(n)}(0) = E[X^n]$$

The moment generating function of the sum of two independent random variables is the product of their moment generating functions. This fact is often used to derive the distribution of the sum of two r.v.s

$$M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t)$$

MGF Example

The normal random variable can be shown to have MGF $M_X(t) = E[e^{tX}] = \exp(\mu_X t + \frac{\sigma_X^2 t^2}{2})$ where X has parameters μ_X and σ_X^2

If Y is an independent normal with mean μ_Y and variance σ_Y^2, then Y also has MGF $M_Y(t) = E[e^{tY}] = \exp(\mu_Y t + \frac{\sigma_Y^2 t^2}{2})$

Then the MGF of $X+Y$ is

$$M_{X+Y}(t) = M_X(t) M_{Y(t)} = \exp(\mu_X t + \frac{\sigma_X^2 t^2}{2}) \exp(\mu_Y t + \frac{\sigma_Y^2 t^2}{2})$$

$$= \exp(\mu_X t + \frac{\sigma_X^2 t^2}{2} + \mu_Y t + \frac{\sigma_Y^2 t^2}{2})$$

$$= \exp((\mu_X + \mu_Y)t + \frac{(\sigma_X^2 + \sigma_Y^2)t^2}{2})$$

Thus $X+Y$ has normal distribution with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$

This fact, that sums of iid normals are normal with summed variances and means, is recurrent

Covariance & Correlation

The covariance of two random variables X and Y is defined as

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY]-E[X]E[Y] \end{aligned}$$

For an n-dimentional random vector $\vec{X}$, the variance-covariance matrix Σ has $var(X_i)=\Sigma_{ii}$ and has $cov(X_i, X_j)=\Sigma_{ij}$ $\forall i,j \leq n$, Σ is thus $n \times n$

The correlation between X and Y is defined as

$$Corr(X,Y)=\frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{var(Y)}}$$

The $\sqrt{var(X)}$ is often called the standard deviation of X

Transformation of Variables

Often, we want to re-express a pair of random variables in a different way, for example, express X, Y as U, V where $U = XY$ and $V = \frac{X}{Y}$. This is done through transformation of variables when X and Y have a pdf.

First, solve for X and Y in terms of U and V. e.g. $X = \frac{U}{Y}$, $V = \frac{U/Y}{Y}$, $Y^2 = \frac{U}{V}$, $X = \frac{U}{\sqrt{U/V}}$

Then, the joint PDF of U and V is

$$f_{U,V}(u,v) = f_{X,Y}\left(x\left(\begin{smallmatrix}\text{in } U \text{ and}\\ V \text{ terms}\end{smallmatrix}\right), y\left(\begin{smallmatrix}\text{in } U \text{ and}\\ V \text{ terms}\end{smallmatrix}\right)\right) |J|$$

Where J is the Jacobian:

$$|J| = \left| \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} \right|$$

Transformation of Variables

Example:

$$f_{X,Y} = \begin{cases} 4xy & 0<x<1,\ 0<y<1 \\ 0 & \text{o.w.} \end{cases}$$

$U = X+Y \qquad V = X-Y$

Solve for X and Y: $X = \frac{U+V}{2} \qquad Y = \frac{U-V}{2}$

$$|J| = \left|\det\begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}\right| = \left|\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}\right| = \left|-\frac{1}{4}-\frac{1}{4}\right| = \left|-\frac{1}{2}\right|$$

$= \frac{1}{2}$, $f_{U,V}(u,v) = 4\left(\frac{u+v}{2}\right)\left(\frac{u-v}{2}\right)\cdot\frac{1}{2}$

$= \frac{u^2-v^2}{8}$. Since $0<x<1$ and $0<y<1$,

$0 < \frac{u+v}{2} < 1 \Rightarrow -u < v < 2-u$

$0 < \frac{u-v}{2} < 1 \Rightarrow -u < -v < 2-u$

$0 < w < 2$

for $0 < w \le 1$

$-w < v < w$

for $1 < w < 2$

$-(2-w) < v < 2-w$

IID Random Sample

The random variables $X_1, \ldots, X_n$ are called a <u>random sample</u> of size n from the population $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of of each X_i is the same function $f(x)$. This is equivalent to the statement that $X_1, \ldots, X_n$ are i.i.d, independent identically distributed. Often, instead of stating the pdf, the functional form of the distribution the X's are sampled from is stated with an unknown parameter.
For example: let $X_1, \ldots, X_n$ be an i.i.d. random sample from a Normal (μ, σ^2)

An i.i.d random sample from pdf $f(x)$ or PMF $P(X=x)$ has joint pdf or PMF

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \quad \text{or} \quad P(X_1=x_1, \ldots, X_n=x_n) = \prod_{i=1}^{n} P(X_i=x_i)$$

In general, if the pdf/pmf can be factored into $f(x_1)f(x_2)\ldots$ then $X_1, X_2, \ldots$ are independent r.v.s

Statistics & Estimators

A <u>statistic</u> is a function that maps the <u>result</u> of a random sample to the real line. A statistic of a sample $\vec{X}$ is often notated $T(\vec{X})$

Examples of statistics:

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

Sample variance: $\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$

An <u>estimator</u> is a statistic used to estimate some quantity of interest called an <u>estimand</u>

Often, the estimand is a parameter of the distribution we are sampling from

For normal: $\bar{X}$ is an estimator of μ
$\hat{\sigma}_X^2$ is an estimator of σ_X^2

Order Statistics

The order statistics of a random sample $X_1, \ldots, X_n$ are the sampled values placed in ascending order. These are denoted $X_{(1)}, \ldots, X_{(n)}$

The order statistics are random variables that satisfy $X_{(1)} \leq \ldots \leq X_{(n)}$. $X_{(1)}$ is the minimum and $X_{(n)}$ is the maximum

If n is odd, the sample median is $X_{(\frac{n+1}{2})}$, the middle order statistic

If n is even, the sample median is $\frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}$, the average of the middle two order statistics

The range is $X_{(n)} - X_{(1)}$

The pth quantile is the minimum order statistic such that $\frac{p}{100}$ proportion of the sampled values are less than that order statistic

Order Statistics

Let $X_1, \dots, X_n$ be an i.i.d r.s. from a discrete distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < \dots$ are the possible values of X in ascending order (e.g. for bin(2, 0.5), this is $x_1 = 0, x_2 = 1, x_3 = 2$). Let $X_{(1)}, \dots, X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1-p_i)^{n-k} \qquad (1)$$

where $P_0 = 0, P_1 = P_0 + p_1, P_2 = P_1 + p_2, P_3 = P_2 + p_3, \dots$

Also,

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} \left[p_i^k (1-p_i)^{n-k} - P_{i-1}^k (1-p_{i-1})^{n-k} \right]$$

Pf of (1): For each x_i the probability that any given observation is below it is p_i, so Y, the number of observations less than or equal to x_i is binomially distributed with $n = n$ and $p = p_i$. As a result, the probability $X_{(j)} \le x_i$, the probability j of n are $\le x_i$, is $P(Y \le x_i)$

Order Statistics Continuous

When $X_1, \ldots, X_n$ is an iid random sample from a continuous distribution with PdF $f_X(x)$ and CdF $F_X(x)$, the PdF of the jth order statistic $X_{(j)}$ is

$$F_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)\left[F_X(x)\right]^{j-1}\left[1-F_X(x)\right]^{n-j}$$

This comes from $F_{X_{(j)}}(x)$, which is, utilizing the same exact logic as in the discrete case:

$$F_{X_{(j)}}(x) = P(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k}\left[F_X(x)\right]^{k}\left[1-F_X(x)\right]^{n-k}$$

the PdF is obtained by simply differentiating the rhs of the above and doing some algebra

Order Statistics Continuous

$$\frac{d}{dx}\sum_{k=j}^{n}\binom{n}{k}[F_x(x)]^k[1-F_x(x)]^{n-k}$$

$$=\sum_{k=j}^{n}\binom{n}{k}\Big(kF_x(x)^{k-1}f_x(x)[1-F_x(x)]^{n-k}-[F_x(x)]^k[1-F_x(x)]^{n-k-1}f_x(x)(n-k)\Big)$$

$$=\sum_{k=j}^{n}\binom{n}{k}kF_x(x)^{k-1}f_x(x)[1-F_x(x)]^{n-k}-\sum_{k=j}^{n}[F_x(x)]^k[1-F_x(x)]^{n-k-1}f_x(x)(n-k)\binom{n}{k}$$

$$=\binom{n}{j}jF_x(x)^{j-1}f_x(x)[1-F_x(x)]^{n-j}+\sum_{k=j+1}^{n}\binom{n}{k}kF_x(x)^{k-1}f_x(x)[1-F_x(x)]^{n-k}-\binom{n}{n}(n-n)F_x(x)^n[1-F_x(x)]^{n-n-1}f_x(x)-\sum_{k=j}^{n-1}[F_x(x)]^k[1-F_x(x)]^{n-k-1}f_x(x)(n-k)\binom{n}{k}$$

$$=\binom{n}{j}jF_x(x)^{j-1}f_x(x)[1-F_x(x)]^{n-j}+\left(\sum_{k=j+1}^{n}\binom{n}{k}kF_x(x)^{k-1}f_x(x)[1-F_x(x)]^{n-k}\right)-\left(\sum_{k=j}^{n-1}\binom{n}{k}(n-k)F_x(x)^kf_x(x)[1-F_x(x)]^{n-k-1}\right)$$

Now, $\forall j, j+1$, $\frac{n!}{j!(n-j)!}(n-j)F_x(x)^jf_x(x)[1-F_x(x)]^{n-j-1}=$
$\frac{n!}{(j+1)!(n-(j+1))!}jF_x(x)^{(j+1)-1}f_x(x)[1-F_x(x)]^{n-(j+1)}$
So the two sums cancel and since $\frac{n!}{j!(n-j)!}j=\frac{n!}{(j-1)!(n-j)!}$,
we have the result:

$$=\frac{n!}{(j-1)!(n-j)!}f_x(x)[F_x(x)]^{j-1}[1-F_x(x)]^{n-j}$$

Sufficiency

A statistic $T(\vec{X})$ of an i.i.d. sample $X_1, \ldots, X_n$ assumed to come from a distribution with unknown parameter θ is said to be sufficient for θ if any inference about θ depends on X only through $T(\vec{X})$.

That is, if two sample points X and Y are observed such that $T(x) = T(y)$ then inference about θ is the same whether $X = x$ or $Y = y$ is observed.

The conditional distribution of the sample X given the value of $T(\vec{X})$ does not depend on θ.

If you can factor the likelihood function of the data, the PDF or PMF considered as a function of θ, $L(\theta | \vec{X})$, into a function of the chosen $T(\vec{X})$ statistic and θ and a function of only the data, $g(\theta | T(\vec{X}))\, h(\vec{X})$, then $T(\vec{X})$ is sufficient.
If the likelihood divided by the pdf of $T(\vec{X})$ does not contain θ, $T(\vec{X})$ is sufficient.

If θ is multi dimensional, parameters $\theta_1, \ldots, \theta_k$ are all unknown, and $T(\vec{X})$ is a k-dimensional sufficient statistic, $T(\vec{X})$ is said to be minimal sufficient

Completeness

A statistic of a random sample $\vec{X}$ assumed to be i.i.d. coming from a distribution with unknown parameter θ is said to be <u>complete</u> (or <u>the family of distributions</u>) is said to be complete) if $\forall g \; E[g(T(\vec{X}))]=0$ implies $g(x)=0$, the only way to make a function of the sufficient statistic that <u>always</u> has expectation zero is to "cheat" by just setting it equal to zero.

A distribution is said to be part of the "exponential family" if its Pdf/Pmf for a single observation can be manipulated into the following form, and if it is/can, then a result about completeness follows:

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right)$$

Where $\theta=(\theta_1,\dots,\theta_k)$, if θ is k-dimensional. Then $T(\vec{X}) = \left(\sum_{i=1}^{n} t_1(x_i), \sum_{i=1}^{n} t_2(x_i), \dots, \sum_{i=1}^{n} t_k(x_i)\right)$ is complete <u>unless</u> $w_1(\theta),\dots,w_k(\theta): \theta \in \Theta$ contains no open sets in $\mathbb{R}^k$.

Completeness can equivalently be defined as $E_\theta[g(T)]=0$ for all θ implies $P_\theta(g(T)=0)=1 \; \forall \theta$

Basu's Theorem

An Ancillary statistic is a statistic of an i.i.d. random sample $\vec{X}$, $S(\vec{X})$, that has a distribution that does not depend on θ. For example, if we are sampling from a Uniform $(\theta-1, \theta+1)$, the statistic $X_{(n)} - X_{(1)}$ does not depend on θ so $S(\vec{X}) = X_{(n)} - X_{(1)}$ is ancilliary.

Basu's theorem says if $T(\vec{X})$ is a complete and minimal sufficient statistic, then $T(\vec{X})$ is independent of every ancillary statistic

Note that the word "minimal" is unnecessary since complete statistics are minimal by definition

UMVUE

A statistic used as an estimator of θ, $T(\vec{X})$, is unbiased if $E[T(\vec{X})] = \theta$.

If $W^*(\vec{X})$ is an unbiased estimator of θ, and $W^*(\vec{X})$ has less variance than any other unbiased estimator of θ $W(\vec{X})$, that is

$$Var_\theta(W^*(\vec{X})) \leq Var_\theta(W(\vec{X})) \quad \forall \theta, W_{unbiased}$$

then we call $W^*(\vec{X})$ the uniform minimum variance unbiased estimator of θ, or the UMVUE of θ

MLE & Log-Likelihood

The <u>likelihood</u> $L(\theta|\vec{X})$ is the joint PDF $L(\theta|\vec{X}) = f(\vec{x}|\theta)$ or PMF $L(\theta|\vec{X}) = P(\vec{X}=\vec{x}|\theta)$ of an i.i.d. random sample as a function of the parameter θ. Since the function depends on random variables, the likelihood is a "random function".

Often, we are concerned with finding the θ that, given $X_1, \ldots, X_n$, maximizes $L(\theta|\vec{X})$. This value for θ is called the <u>maximum likelihood estimator</u> or MLE of θ, $\hat{\theta}_{MLE}$.

Since $\log(x)$ is monotone, the θ that maximizes $L(\theta|\vec{X})$ also maximizes $\log(L(\theta|\vec{X}))$. $\log(L(\theta|\vec{X}))$ is typically easier to work with, and is called the <u>Log-Likelihood</u>, notated $\ell(\theta|\vec{X})$.

Often, MLE is found by equating the score to zero, checking boundaries (if any), and checking to see that second deriv of log-L is negative. i.e. Use calculus to find a global max of log-L.

Score Function

The derivative of the Log-likelihood $\ell(\theta|\vec{X})$ w.r.t. θ, $\frac{d}{d\theta}\ell(\theta|\vec{X})$, is called the Score-function. When $\theta = \hat{\theta}_{MLE}$, the score is always zero, since any maximizer (or minimizer) of a function must have a tangent of slope zero.

For any value of θ, the score is a random function because $L(\theta|\vec{X})$ is a random function depending on $\vec{X}$.

The expected value of the score across $\vec{X}$ is a function of θ but always is equal to zero,
$E_\theta\left[\frac{d}{d\theta}\log L(\theta|\vec{X})\right] = 0 \quad \forall\,\theta$

This is true because: $E_\theta\left[\frac{d}{d\theta}\log L(\theta|\vec{X})\right] =$

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\log\left(f(\vec{X}|\theta)\right)f(\vec{X}|\theta)\,dx = \int_{-\infty}^{\infty}\left(\frac{\frac{\partial}{\partial\theta}f(\vec{X}|\theta)}{f(\vec{X}|\theta)}\right)f(\vec{X}|\theta)\,dx$$

$$= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(\vec{X}|\theta)\,dx = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}f(\vec{X}|\theta)\,dx = \frac{\partial}{\partial\theta}1 = 0$$

$E[\text{score}] = 0$ for discrete r.v.s too, just replace $\int_{-\infty}^{\infty}$ with $\sum_{x_i}$

Note that $g(x) = \log f(x) \Rightarrow g'(x) = \frac{f'(x)}{f(x)}$ by chain rule

Fisher Information

Recall that the score is a random function, $\frac{d}{d\theta}\ell(\theta|\vec{X})$. The Fisher information $I(\theta)$ is the Variance of the score,

$$I(\theta) \equiv E\left[\left(\frac{d}{d\theta}\ell(\theta|\vec{X}) - \underbrace{E\left[\frac{d}{d\theta}\ell(\theta|\vec{X})\right]}_{=0}\right)^2\right]$$

$$= E\left[\left(\frac{d}{d\theta}\ell(\theta|\vec{X})\right)^2\right]$$

The Fisher information can also be shown to be equivalent to the negative of the Hessian, which is the second derivative of the log-likelihood:

Proof: $E[-H] = E\left[-\frac{\partial^2}{\partial\theta^2}\log\left(f(\vec{X}|\theta)\right)\right]$

$$= E\left[-\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\log\left[f(\vec{X}|\theta)\right]\right)\right] = E\left[-\frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}f(\vec{X}|\theta)}{f(\vec{X}|\theta)}\right)\right]$$

$$= E\left[\frac{\frac{\partial}{\partial\theta}f(\vec{X}|\theta)\frac{\partial}{\partial\theta}f(\vec{X}|\theta) - f(\vec{X}|\theta)\frac{\partial^2}{\partial\theta^2}f(\vec{X}|\theta)}{f(\vec{X}|\theta)^2}\right]$$

$$= E\left[\left(\frac{\frac{\partial}{\partial\theta}f(\vec{X}|\theta)}{f(\vec{X}|\theta)}\right)^2\right] - E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(\vec{X}|\theta)}{f(\vec{X}|\theta)}\right]$$

Fisher Information Continued

Recall that $\frac{d}{dx}\log(f(x)) = \frac{f'(x)}{f(x)}$, so

$$= E\left[\left(\frac{\partial}{\partial\theta}\log\left(f(\vec{X}|\theta)\right)\right)^2\right] - \int_{-\infty}^{\infty} x\,\frac{\frac{\partial^2}{\partial\theta^2}f(\vec{X}|\theta)}{f(\vec{X}|\theta)}\,dx$$

Integrate $\int_{-\infty}^{\infty} x\,\frac{\frac{\partial^2}{\partial\theta^2}f(\vec{X}|\theta)}{f(\vec{X}|\theta)}\,dx$ by parts:

$u = \frac{\partial^2}{\partial\theta^2}f(\vec{X}|\theta)$, $V = \frac{x}{f(\vec{X}|\theta)}$, $\int uV\,dx = u\int V\,dx - \int u'\left(\int V\,dx\right)dx$

Since $\int V\,dx = \int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2}f(\vec{X}|\theta) = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}f(\vec{X}|\theta)\right)dx$

$$= \frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(\vec{X}|\theta)\,dx\right) = \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty}f(\vec{X}|\theta)\,dx\right)\right) = \frac{\partial^2}{\partial\theta^2}(1) = 0$$

this becomes $= u\cdot 0 - \int u'\cdot 0\,dx = 0$, thus

$\int_{-\infty}^{\infty} x\,\frac{\frac{\partial^2}{\partial\theta}f(\vec{X}|\theta)}{f(\vec{X}|\theta)}\,dx = 0$, thus

$$E[-H] = E\left[\left(\frac{\partial}{\partial\theta}\ell(\theta|\vec{X})\right)^2\right] = I(\theta)$$

Fisher Information Matrix

The Fisher information, the variance of the score for parameter θ, has a natural extension to multiple parameters: For multiple parameters, the information matrix is the variance covariance matrix of the scores of these parameters.

For parameters $\theta_1, \theta_2, \ldots, \theta_n \in \vec{\theta}$ each entry $I_{i,j}$ is $\text{cov}(\text{score}(\theta_i), \text{score}(\theta_j)) =$

$$E\left[\left(\frac{\partial}{\partial \theta_i} \ell(\vec{\theta} \mid \vec{X})\right)\left(\frac{\partial}{\partial \theta_j} \ell(\vec{\theta} \mid \vec{X})\right)\right]$$

Additionally, since information matrix is the negative expectation of the Hessian:

$$I_{i,j} = -E\left[\frac{\partial^2 \ell(\vec{\theta} \mid \vec{X})}{\partial \theta_i \partial \theta_j}\right]$$

Fisher Scoring

Suppose we want to find $\hat{\theta}_{MLE}$ of a an i.i.d. sample from a distribution for which the correct $W(\vec{X}) = \hat{\theta}_{MLE}$ cannot be solved for analytically, or we are too lazy to determine it analytically. We can use a taylor expansion and $\ell(\theta|\vec{x})$ and $I(\theta)$ to approximate $\hat{\theta}_{MLE}$ arbitrarily closely. This is called Fisher-Scoring. If we decide not to replace H with $I(\theta)$, this is called the Newton-Raphson method.

Recall from calc 2 that a taylor expansion of a function $F(x)$ approximates the function's value at a point 'a' with $f(x) \approx f(a) + f'(a)(x-a) + \ldots$
Using the 'first order' expansion to approximate $f(x)$:
$f(x) \approx f(a) + f'(a)(x-a)$ (linear approximation at a)

If we are interested in finding the root of $f(x)$ (as we are in this case for reasons to be explained soon) we set $f(x)=0$: $0 \approx f(a) + f'(a)(x-a)$
$x \approx a - \frac{f(a)}{f'(a)}$

Fisher Scoring Continued

We can use the approximation $X \approx a - \frac{f(a)}{f'(a)}$ to define an iterative procedure that finds the root of $f(X)$ by repeatedly replacing a with the previous approximation of the root:

$X_{n+1} = X_n - \frac{f(X_n)}{f'(X_n)}$. If you run this for many iterations, and f isn't something really really weird, after a while $X_{n+1} \approx X_n \approx X$ s.t. $f(X) = 0$

Now recall that the MLE has to maximize the likelihood, so its first derivative must be zero. Thus, $\hat{\theta}_{MLE}$ can be approximated using the above Newton-Raphson technique where $f(x) = \frac{d}{d\theta} \ell(\theta|\vec{X})$, the Score

$$\theta_{n+1} = \theta_n - \frac{\frac{d}{d\theta} \ell(\theta_n|\vec{X}) \longrightarrow \text{Score}}{\frac{d^2}{d\theta^2} \ell(\theta_n|\vec{X}) \longrightarrow \text{Hessian}}$$

Since the Score and Hessian are both random this can be unstable, so replace H with $E[H|\theta=\theta_n]$ to get Fisher Scoring

$$\theta_{n+1} = \theta_n - \frac{\frac{d}{d\theta} \ell(\theta_n|\vec{X})}{E[H|\theta=\theta_n]} \Rightarrow \theta_{n+1} = \theta_n + \frac{\frac{d}{d\theta} \ell(\theta_n|\vec{X})}{I(\theta_n)}$$

Cramer-Rao Lower Bound

Let $X_1, \ldots, X_n$ be a sample with PdF $f(x|\theta)$, and let $W(\vec{X})$ be an estimator of θ with finite variance and expectation satisfying

$$\frac{d}{d\theta} E[W(\vec{X})] = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} W(\vec{x}) f(\vec{x}|\theta) \, d\vec{x}$$

Then $$\text{Var}(W(\vec{X})) \geq \frac{\left(\frac{d}{d\theta} E[W(\vec{X})]\right)^2}{I(\theta)}$$

The rhs is known as the Cramer-Rao lower bound because it is the lowest possible variance of an estimator $W(\vec{X})$

When we constrain $W(\vec{X})$ to be unbiased $E[W(\vec{X})] = \theta$, then

$$\text{Var}(W(\vec{X})) \geq \frac{\left(\frac{d}{d\theta}\theta\right)^2}{I(\theta)} \Rightarrow \text{Var}(W(\vec{X})) \geq \frac{1}{I(\theta)}$$

If $W(\vec{X})$ is unbiased, its variance is bounded by the inverse of the variance of the score

Cramer-Rao Lower Bound

Proof: Recall $\text{Cov}(U,V)^2 \leq \text{Var}(U)\text{Var}(V)$ and $\frac{d}{dx}\log(f(x)) = \frac{f'(x)}{f(x)}$.

Let $\vec{X}$ be an iid r.s. from a distribution with parameter θ and let $W(\vec{X})$ be an estimator of θ.

Let $U = W(\vec{X})$, $V = \frac{d}{d\theta}\ell(\theta \mid \vec{X})$ the score

$$\text{Cov}(U,V)^2 \leq \text{Var}(U)\text{Var}(V) \Rightarrow \text{Var}(W(\vec{X})) \geq \frac{\text{Cov}\left(W(\vec{X}), \frac{d}{d\theta}\ell(\theta|\vec{X})\right)^2}{\text{Var}\left(\frac{d}{d\theta}\ell(\theta|\vec{X})\right)}$$

$\text{Var}\left(\frac{d}{d\theta}\ell(\theta|\vec{X})\right) = I(\theta)$ by definition, and

$$\text{Cov}\left(W(\vec{X}), \frac{d}{d\theta}\ell(\theta|\vec{X})\right) = E\left[W(\vec{X})\frac{d}{d\theta}\ell(\theta|\vec{X})\right] - E[W(\vec{X})]\underbrace{E\left[\frac{d}{d\theta}\ell(\theta|\vec{X})\right]}_{0}$$

$$= \int_{-\infty}^{\infty} W(\vec{x}) f(\vec{x}|\theta)\left(\frac{\partial}{\partial\theta}\log(f(\vec{x}|\theta))\right)dx$$

$$= \int_{-\infty}^{\infty} W(\vec{x}) \cancel{f(\vec{x}|\theta)}\left(\frac{\frac{\partial}{\partial\theta}f(\vec{x}|\theta)}{\cancel{f(\vec{x}|\theta)}}\right)dx = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}W(\vec{x})f(\vec{x}|\theta)dx$$

$$= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} W(\vec{x})f(\vec{x}|\theta) = \frac{\partial}{\partial\theta}E[W(\vec{X})] \quad \text{q.e.d.}$$

Attainment & Lehman-Scheffé

Let $\vec{X}$ be an i.i.d. r.s. from $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of CRLB. If $W(\vec{X})$ is any unbiased estimator of $\tau(\theta)$ (a function of θ, usually just $\tau(\theta)=\theta$) then $W(\vec{X})$ attains CRLB for unbiased estimators if and only if:

$$\exists a(x) \text{ s.t. } a(\theta)[W(\vec{X}) - \tau(\theta)] = \frac{\partial}{\partial\theta} l(\theta|\vec{X})$$

Lehman-Scheffé Theorem:
Again, let $\vec{X}$ be an i.i.d. r.s. from a distribution with unknown parameter θ. Suppose $T(\vec{X})$ is an estimator of $\tau(\theta)$, and suppose $T(\vec{X})$ is a function of a complete and sufficient statistic for θ and is unbiased for $\tau(\theta)$ ($E[T(\vec{X})] = \tau(\theta)$). Then $T(\vec{X})$ is UMVUE for $\tau(\theta)$.

Further: $T(\vec{X})$ will also be a function of $\hat{\theta}_{MLE}$.
tldr; Unbiased + Complete + Sufficient = UMVUE

Rao-Blackwell & Uniqueness Theorem

Let $\vec{X}$ be i.i.d. rs from dist w/ unknown θ. Let W be any unbiased estimator of $\tau(\theta)$, and let $T(\vec{X})$ be a sufficient statistic for θ. Define $\phi(T(\vec{X})) = E[W(\vec{X}) \mid T(\vec{X})]$. Then $E[\phi(T(\vec{X}))] = \tau(\theta)$ and

$$\mathrm{Var}(\phi(T(\vec{X}))) \leq \mathrm{Var}(W(\vec{X})) \quad \forall \theta$$

Proof: Using Adam's rule and Eve's rule,

$$E[\phi(T(\vec{X}))] = E[E[W(\vec{X}) \mid T(\vec{X})]] = E[W(\vec{X})] = \tau(\theta)$$

$$\mathrm{Var}(W(\vec{X})) = \mathrm{Var}(E[W(\vec{X}) \mid T(\vec{X})]) + E[\mathrm{Var}(W(\vec{X}) \mid T(\vec{X}))]$$

$$= \mathrm{Var}(\phi(T(\vec{X}))) + E[\mathrm{Var}(W(\vec{X}) \mid T(\vec{X}))]$$

$\geq \mathrm{Var}(\phi(T(\vec{X})))$. This is the <u>Rao-Blackwell Theorem.</u>

Uniqueness Theorem

Suppose $W(\vec{X})$ is UMVUE. Then $W(\vec{X})$ is unique.

Hypothesis Testing

A hypothesis is a statement, often about a population parameter of an assumed distribution

The two complementary hypotheses in a testing procedure are called the null hypothesis and the alternative hypothesis, denoted H_0 and H_1

If θ is the population parameter, the general format of the test is usually

$H_0: \theta \in \Theta_0$ $\quad$ $H_1: \theta \in \Theta_0^c$

You can also use a hypothesis test to test the hypothesis that $X_1, X_2, \ldots, X_n$ come from a particular distribution, e.g. Normal

$H_0: X_1, \ldots, X_n \sim \text{Normal}$ $\quad$ $H_a: X_1, \ldots, X_n \not\sim \text{Normal}$

You can also test other estimands of interest, e.g.

$H_0: E[X] = 0$ $\quad$ $H_a: E[X] \neq 0$

Hypothesis Procedures

A hypothesis testing procedure is a rule that specifies for which sample values the decision is made to reject H_0 and accept H_1 as true and for which sample values this rejection of H_0 is not possible.

The subset of the sample space for which H_0 is rejected is the "rejection region" and the subset for which H_0 is not rejected is the "acceptance region"

Typically, a hypothesis test is specified in terms of a test statistic of the sample, $T(X_1, \ldots, X_n)$

Types of Error

A hypothesis test can make one of two kinds of error.

A <u>type I error</u> takes place when the null hypothesis is rejected even when it is actually true. False rejection.

A <u>type II error</u> takes place when the null hypothesis is accepted even when it is actually false. False Acceptance.

Since the null hypothesis doesn't depend on the data, the probability of making a type I for any particular test can be calculated explicitly without seeing the data.

For this reason, the goal of hypothesis testing is to minimize type II error while holding the type I error rate fixed at some level α, $0 < \alpha < 1$

Likelihood Ratio Tests

Likelihood Ratio Testing (LRTs) tests based on the ratio of the maximum likelihood under the null hypothesis to the overall maximum likelihood. The test statistic is written

$$\lambda(X) = \frac{\sup_{\theta_0} L(\theta \mid X)}{\sup_{\theta} L(\theta \mid X)}$$

And the test must have a rejection region of the form $\{X: \lambda(X) \leq c\}$, $0 \leq c \leq 1$

Likelihood Ratio Test Example

Let $X_1, \ldots, X_n$ be a r.s. from exponential population with pdf $f(x|\theta) = e^{-x+\theta}$ for $x \geq \theta$, $f(x|\theta) = 0$ $x < \theta$. Test $H_0: \theta \leq \theta_0$.

The likelihood function is thus

$$L(\theta|x) = \begin{cases} e^{-\sum x_i + n\theta} & \theta \leq x_{(1)} \\ 0 & \theta > x_{(1)} \end{cases}$$

Since L is an increasing function of θ, the denominator of the LRT uses the largest possible value of θ, $\theta = x_{(1)}$. The numerator will also be as high as possible, which is either $x_{(1)}$ or θ_0. So

$$\lambda(x) = \begin{cases} 1 & x_{(1)} \leq \theta_0 \\ \dfrac{e^{-\sum x_i + n\theta_0}}{e^{-\sum x_i + n x_{(1)}}} & x_{(1)} > \theta_0 \end{cases}$$

When $\theta_0 > x_{(1)}$, the null $\theta \leq \theta_0$ includes all θ less than $x_{(1)}$, so it includes all possible θ, so immediately fail to reject. Reject when θ_0 is too small.

Power

The probability of making a type 2 error under the alternative hypothesis usually necessitates a particular assumption of an "alternative distribution"

If we are testing $\mu = \mu_0$ for $X_1, \ldots, X_n$ iid normal$(\mu, 1)$, the type II error is substantially different if $\mu = \mu_0 + 1$ vs $\mu = \mu_0 + 10$.

The power function of a hypothesis test with rejection region R is the function of θ defined by

$$B(\theta) = P_\theta(X \in R) = \begin{cases} \text{Prob of type I error} & \theta \in \Theta_0 \\ 1 - \text{prob of type II error} & \theta \in \Theta_0^c \end{cases}$$

Ideally, a test should have probability near 0 (e.g. 0.05) $\forall\ \theta \in \Theta_0$ and probability near 1 $\forall\ \theta \in \Theta_0^c$

Neyman-Pearson Lemma

If C is a class of tests for $H_0: \theta \in \Theta_0$ vs $\theta \in \Theta_0^c$, a test in C with power function $\beta(\theta)$ is a <u>uniformly most powerful</u> test in C if $\beta(\theta) \geq \beta'(\theta)$ $\forall \theta \in \Theta_0^c$, $\forall \beta'$ of C

Consider testing $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$ where the pdf or pmf corresponding to θ_i is $f(x|\theta_i)$, $i = 0, 1$, using a test with rejection region R satisfying

a. $x \in R$ if $f(x|\theta_1) > k f(x|\theta_0)$ and $x \in R^c$ if $f(x|\theta_1) < k f(x|\theta_0)$ for some $k \geq 0$, and b. $\alpha = P_{\theta_0}[X \in R)$

I. Any test that satisfies a & b is a uniformly most powerful level α test

II. If there exists a test satisfying a & b with $k > 0$, then every UMP level α test is a size α test (satisfies b.), except perhaps $P_{\theta_0}(X \in A) = P_{\theta_1}(X \in A) = 0$

Sufficient Neyman-Pearson Lemma

Consider the hypothesis problem posed in the Neyman pearson. Suppose $T(X)$ is a sufficient statistic for θ and $g(t|\theta_i)$ is the pdf or pmf of T corresponding to θ_i, $i = 0, 1$. Then any test based on T with rejection region S (a subset of the sample space of T) is a UMP level α test if it satisfies

$$t \in S \quad \text{if} \quad g(t|\theta_1) > k\, g(t|\theta_0)$$

and

$$t \in S^c \quad \text{if} \quad g(t|\theta_1) < k\, g(t|\theta_0)$$

for $k \geq 0$ where $\alpha = P_{\theta_0}(T \in S)$

ie. A UMP level α test based on a sufficient statistic is just as good as one based on the whole sample

Karlin - Rubin Lemma

Consider testing $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$. Suppose T is a sufficient statistic for θ and the family of PDFs or PMFs $\{g(t|\theta): \theta \in \Theta\}$ of T has a <u>monotone likelihood ratio</u>.

That is, for every $\theta_2 > \theta_1 \in \Theta$ for $g(t|\theta): \theta \in \Theta$, the ratio $g(t|\theta_2)/g(t|\theta_1)$ is monotone (nonInc or nonDec) func of t on $t: g(t|\theta_1) > 0$ or $g(t|\theta_2) > 0$ with $\frac{c}{0}$ defined as ∞ if $c > 0$

Then for any t_0 the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test where $\alpha = P_{\theta_0}(T > t_0)$

P-Values

A P-Value $P(X)$ is a test statistic s.t. $0 \leq P(X) \leq 1$ for every sample point X, and s.t. the type I error of the test $P(X) \leq \alpha$ is less than or equal to α.

That is, a valid P-value has $\forall \alpha, 0 \leq \alpha \leq 1$

$$P_\theta(P(X) \leq \alpha) \leq \alpha$$

An easy way to construct a level α test based on $P(X)$ is $P(X) \leq \alpha$. Since a p-value can be anywhere from 0 to 1, it gives more information than "Accept" or "Reject"!

If $W(X)$ is a test statistic s.t. large values of W give evidence H_1 is true, then let

$$P(x) = \sup_{\theta \in \theta_0} P_\theta(W(X) \geq W(x))$$

(X: random var; x: realizations)

Probability of a result "as significant or more significant" than the one you obsvd

Confidence Interval

An Interval Estimate of a real-valued parameter θ is any pair of functions $L(X_1, \ldots, X_n)$ and $U(X_1, \ldots, X_n)$ of a sample that satisfy $L(\vec{x}) \leq U(\vec{x})$ $\forall x$. The inference $L(\vec{x}) \leq \theta \leq U(\vec{x})$, written $[L(x), U(x)]$ is an interval estimate.

The most common interval estimate is known as a Confidence Interval, and is associated with both a test and a sig. level α. The CI or Confidence Set for a set of observations $\vec{X}$ is the set of θ_0's for which the level α test of θ_0 would not result in a rejection of the null hypothesis. This is also called a $100\alpha\%$ CI.

In other words, a $100\alpha\%$ CI is the set of "plausible" null hypothesized θ_0s would not result in the null being rejected.

Pivot

Often, a $100\alpha\%$ CI can be found using a pivot. A pivot is a random variable whose distribution does not depend on any parameters.

Many distributions have the location-scale type, and these have simple pivots.

Form of PDF	Type of Pdf	Pivot
$f(x-\mu)$	Location	$\bar{X}-\mu$
$\frac{1}{\sigma}f\left(\frac{x}{\sigma}\right)$	Scale	$\bar{x}/\sigma$
$\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$	Location-Scale	$\frac{\bar{x}-\mu}{\sigma}$

The CDF can be found by putting the $\frac{\alpha}{2}$ and $\frac{1-\alpha}{2}$ % quantiles of the pivot distribution on each side and solving for the parameter(s).

For example, $X_1 \dots X_n \sim N(\mu, 4)$ $\bar{X}=10$, $\alpha=0.05$, pivot distribution is $N(0,1)$, so

$$-1.96 \le \frac{10-\mu}{2/\sqrt{n}} \le 1.96 \Rightarrow \frac{1.96\cdot 2}{\sqrt{n}} + 10 \ge \mu \ge \frac{-1.96\cdot 2}{\sqrt{n}} + 10$$

Unbiasedness & Length

In the previous notes, we have split α evenly between both sides of the distribution of the test statistic T, but this is not always optimal (although it often is).

One way to define an 'optimal' interval is through its length. For example, a 95% interval that is $(1, 2)$ is probably more helpful than one that is $(1.2, 4)$. Finding the interval of minimum length depends on the observed sample. Instead, we often minimize the <u>Expected Length</u>, $E[U(\vec{X})] - E[L(\vec{X})]$

Sometimes, it is more convinient to simply find the shortest Pivotal interval rather than the shortest interval.

A confidence set is said to be <u>unbiased</u> if $P_\theta(\theta' \in C(\vec{X})) \leq 1 - \alpha \quad \forall \ \theta \neq \theta'$

Pivoting a CDF

Another common pivot, other than location-scale, is to pivot using the CDF. Let T be a statistic with continuous CDF $F_T(t|\theta)$. If $F_T(t|\theta)$ is decreasing function of θ $\forall t$, solve for θ for

upper: $F_T(t|\theta_U(t)) = \frac{\alpha}{2}$, Lower: $F_T(t|\theta_L(t)) = 1 - \frac{\alpha}{2}$

If $F_T(t|\theta)$ is an increasing func of θ $\forall t$, do

upper: $F_T(t|\theta_U(t)) = 1 - \frac{\alpha}{2}$, Lower: $F_T(t|\theta_L(t)) = \frac{\alpha}{2}$

Central Limit Theorem

Suppose $X_1, \ldots$ is a sequence of i.i.d. random variables with $E[X] = \mu$ and $\text{Var}(X) = \sigma^2 < \infty$.

Then $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \text{Normal}(0, 1)$

Convergence in distribution: $X_1, \ldots$ converges in distribution to X, $X_n \xrightarrow{d} X$, if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x) \quad \forall x \text{ s.t. } F_X(x) \text{ is cont.}$$

Proof: First, let us rewrite the theorem explicitly. For $X_1, X_2, \ldots$ iid var $< \infty$, WLOG replace each r.v. $X_1, \ldots$ with $\frac{X_1 - \mu}{\sqrt{\sigma^2}}, \frac{X_2 - \mu}{\sqrt{\sigma^2}}, \ldots$, call these $Y_1, \ldots$. Obviously, $Y_1, Y_2, \ldots$, have mean zero and variance 1. CLT says

$$\sqrt{n}\,\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}} \xrightarrow{d} N(0,1), \text{ that is}$$

$$\lim_{n \to \infty} F_{\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}}(y) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

Proof of Central Limit Theorem

Recall from the MGF notes that the normal satisfies $X_1 \sim N + X_2 \sim N \Rightarrow X_1 + X_2 \sim N(w_1+w_2, \sigma_1^2 + \sigma_2^2)$ for independent X_1, X_2

Recall also that the MGF uniquely identifies a distribution (two rvs with the same MGF have the same distribution). Recall that the normal has MGF $\exp(ut + \frac{\sigma^2}{2}t^2)$, where $M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} F(x)dx$. Thus, $N(0,1)$ has MGF $\exp(\frac{t^2}{2})$. As a result, we can show (1) by showing (2) below:

$$\lim_{n\to\infty} M_{\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}}(t) = \exp\left(\frac{t^2}{2}\right)$$

Call $\frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}$ to be Z_n. By properties of MGFs (summing indep vars multiplies MGFs) it must be true that

$$M_{\sqrt{n} Z_n}(t) = \left(M_Y(t)\right)^n \Rightarrow M_{Z_n}(t) = \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Proof of Central Limit Theorem

$M_Y(t)$ has a taylor series expansion about zero given by

$$M_Y(t) = M_Y(0) + t M_Y'(0) + \frac{1}{2} t^2 M_Y''(0) + \varepsilon$$

Where $\frac{\varepsilon}{t^2}$ has a t and goes to zero as $t \to 0$.

Since $E[Y] = 0$, $M'(0) = 0$, and since $E[Y^2] = Var(Y) + E[Y]^2 = Var(Y) = 1$, so as $n \to \infty$, $\frac{t}{\sqrt{n}} \to 0$, thus

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty} \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n =$$

$M(0) = E[X^0] = 1$, $tM'(0) = tE[X] = 0$, $\frac{1}{2}t^2 M_Y''(0)$

$$\lim_{n\to\infty} \left(1 + 0 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 \cdot 1 + \varepsilon\right)^n$$

$$= \lim_{n\to\infty} \left(1 + \frac{t^2}{2n} + \varepsilon\right)^n = \lim_{n\to\infty} \left(1 + \frac{t^2/2}{n} + \varepsilon\right)$$

Where all terms in ε are $t^3 = \frac{t^3}{n^{3/2}}$ or greater and disappear in the limit. Thus,

Proof of CLT

$$\lim_{n\to\infty}\left(1+\frac{t^2/2}{n}+\varepsilon\right)^n = \lim_{n\to\infty}\left(1+\frac{t^2/2}{n}\right)^n \quad \text{and}$$

by definition of e^x this is

$$\lim_{n\to\infty}\left(1+\frac{t^2/2}{n}\right)^n = e^{t^2/2}$$

Thus $M_{Z_n(t)} \to e^{t^2/2}$ as $n\to\infty$

Thus a set of r.v.s $Y_1, \ldots, Y_n$ that are standardized $\left(\frac{X-E[X]}{\sqrt{Var(X)}}\right)$ and each X has $Var(X)<\infty$, it is true that $\bar{Y} \xrightarrow{d} N(0,1)$, since

$$M_{\bar{Y}}(t) \to M_{N(0,1)}(t), \text{ and}$$

convergence of MGF is equivalent to convergence in distribution, since both MGF and distribution f uniquely identify a probability distribution.

Lemma for CLT Proof

Let X be a standard normally distributed r.v., that is, it has pdf $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

Then the MGF is given by $E[e^{tX}]$

$$= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2tx)} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}((x-t)^2 - t^2)} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} e^{t^2/2} dx = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx$$

Since inside the integral is the Normal$(t, 1)$ PDF, it integrates to 1, so as a result we have

$$e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx = e^{t^2/2}$$

Thus standard normal has MGF $e^{t^2/2}$

Wald & Score Interval

In large samples, any distribution of finite variance has mean approaching a normal, $\bar{X} \to N(\mu, \frac{\sigma^2}{n})$. Wald and Score intervals are based on this large sample approximation.

Let $\hat{\theta}$ be a point estimate of θ based on the mean. Since $\hat{\theta}$ is a statistic and is based on the mean, as $n \to \infty$ $\hat{\theta} \to N(\theta, Var(\hat{\theta}))$. We want to construct the wald interval as

$$\hat{\theta} \pm \underbrace{Z_{\alpha/2}}_{\text{CDF of normal dist at } \frac{\alpha}{2}, \text{ eg } -1.96 \text{ for } \alpha = 0.05} SE(\hat{\theta})$$

Where Standard Error (SE) is just the Standard Deviation of the estimator $\hat{\theta}$, for $\bar{X}$ is $\sqrt{\frac{Var(X)}{n}} = \frac{SD(x)}{\sqrt{n}}$. For Wald, the $SD(\hat{\theta})$ is estimated too, leading to the below Wald Interval

$$\hat{\theta} \pm Z_{\frac{\alpha}{2}} \widehat{SE}(\hat{\theta})$$

Wald & Score Interval

The Score interval uses the same normal approximation, but is "smart" enough to use the true variance rather than the approximation based on $\hat{\theta}$. This is done by making a normal approximation to the score function, $\frac{\partial}{\partial\theta}\ell(\theta|\vec{x})$. This can be done because, for a set of i.i.d. r.v.s $X_1,\dots,X_n$, the joint pdf is $\prod(f(x_i|\theta))$, and thus the log likelihood is $\sum\log(f(x_i|\theta))$ and the derivative is $\frac{\partial}{\partial\theta}\sum\log(L(\theta|x_i)) = \sum\frac{\partial}{\partial\theta}\log(L(\theta|x_i))$. So the joint score function $\frac{\partial}{\partial\theta}\ell(\theta|\vec{x})$ is the sum of i.i.d. r.v.s $\frac{\partial}{\partial\theta}\log(L(\theta|x_i))$ and thus the CLT applies. The score interval is given by solving

$$\left(\frac{\partial}{\partial\theta}\ell(\hat{\theta}|\vec{x})\right)\Big/\sqrt{I(\theta)} = \pm z_{\alpha/2} \quad \text{for } \theta$$

recalling that $I(\theta)$ is just $\mathrm{Var}\left(\frac{\partial}{\partial\theta}\ell(\theta|\vec{x})\right)$. Solve individually for upper and lower intervals.

Wald & Score Example

Let $X_1, \dots, X_n$ be iid bernoulli. Find Wald and Score intervals for p.

$$P(X=x) = p^x (1-p)^{1-x} \quad \text{for } X=0 \text{ or } X=1$$

Wald: $\hat{p} \pm Z_{\alpha/2} \widehat{SE}(\hat{p})$

$\hat{p}$ is just $\frac{\sum X}{n} = \bar{X}$, and $SD[\bar{X}]$ is known to be $\sqrt{p(1-p)}$, so $SE(\bar{X}) = \frac{SD(\bar{X})}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$, filling $\hat{p}$ for p wald interval is

$$\bar{X} \pm Z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

Score: $\frac{\frac{\partial}{\partial\theta} \ell(\theta | \vec{x})}{I(\theta)} = \pm Z_{\alpha/2}$, first, let's find score and info.

$$\frac{\partial}{\partial\theta} \log\left(\prod p^{x_i} (1-p)^{1-x_i} \right) = \frac{\partial}{\partial p} \sum x_i \log(p) + (1-x_i) \log(1-p)$$

$$= \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = \frac{\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p}$$

Wald & Score Examples

Fisher info $I(p)$ is also $E[-\text{second deriv}]$, so

$$= -E\left[\frac{\partial}{\partial p}\left(\frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p}\right)\right] = E\left[\frac{\sum x_i}{p^2} + \frac{n-\sum x_i}{(1-p)^2}\right]$$

$$= \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n(1-p)}{p(1-p)} + \frac{np}{p(1-p)} = \frac{n}{p(1-p)}$$

So score interval lower is

$$\frac{\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p}}{\sqrt{\frac{n}{p(1-p)}}} = z_{\alpha/2} \Rightarrow \frac{\frac{n\bar{x}(1-p)}{p(1-p)} - \frac{n(1-\bar{x})p}{p(1-p)}}{\sqrt{\frac{n}{p(1-p)}}} = z_{\alpha/2}$$

$$\Rightarrow \frac{n(\bar{x} - \bar{x}p - p + \bar{x}p)}{\sqrt{np(1-p)}} = z_{\frac{\alpha}{2}} \Rightarrow z_{\frac{\alpha}{2}} = \frac{n(\bar{x} - p)}{\sqrt{np(1-p)}}$$

$$\Rightarrow \bar{x} - p = z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \Rightarrow p = \bar{x} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

Thus, $p_L = \bar{x} - z_{\alpha/2}\sqrt{\frac{p_L(1-p_L)}{n}}$, $p_U = \bar{x} + z_{\alpha/2}\sqrt{\frac{p_U(1-p_U)}{n}}$

Since p is unknown, the score interval here requires an iterative/numerical method to find p_L and p_U

Extended Example

Many concepts from foregoing notes are illustrated in the below example for the normal distribution:

Suppose for all parts $X_1, \ldots, X_n$ are normally distributed with unknown mean μ and known variance σ^2.

a. Show that the sample mean $\bar{X} = \frac{\sum x_i}{n}$ and sample variance $S^2 = \frac{1}{n-1}\sum (x_i - \bar{X})^2$ are independent, without Basu's theorem

b. Show that $\bar{X}$ is UMVUE for μ

c. Derive the score & Fisher info of $\vec{X}$, and show that $\bar{X}$ achieves the CRLB

d. Find LRT of $\mu = \mu_0$ against $\mu \neq \mu_0$

e. Use NP-lemma to find the UMP test

f. Find a CI for μ based on pivot

Part A

Show that $\bar{X} \perp S^2$

Since the normal is a location-scale family, assume WLOG that $\mu = 0$ $\sigma^2 = 1$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left((X_1-\bar{X})^2 + \sum_{i=2}^{n}(X_i-\bar{X})^2\right)$$

$$= \frac{1}{n-1}\left(\left(\sum_{i=2}^{n}(X_2-\bar{X})\right)^2 + \sum_{i=2}^{n}(X_i-\bar{X})^2\right)$$

Thus S^2 can be written as a function only of $X_2-\bar{X}, X_3-\bar{X}, \ldots,$ and we can show these are independent of $\bar{X}$.

The joint pdf of $X_1, \ldots, X_n$ is given by

$$f(X_1, \ldots, X_n) = \frac{1}{(2\pi)^{n/2}} e^{-(1/2)\sum_{i=1}^{n} x_i^2}, \quad -\infty < x_i < \infty$$

We transform $Y_1 = \bar{X}$, $Y_2 = X_2 - \bar{X}, \ldots, Y_n = X_n - \bar{X}$

Since the transformation is linear, we just need to write $\sum_{i=1}^{n} X_i^2$ in terms of Y's and scale the pdf.

Part A

$$\sum x_i^2 = \left(Y_1 - \sum_{i=2}^{n} Y_i\right)^2 + \sum_{i=2}^{n} \left(Y_i + Y_1\right)^2$$

$$\left(\bar{X} - \left(X_2 - \bar{X}\right) - \left(X_3 - \bar{X}\right) + \dots\right)^2 + \left(\left(X_2 - \bar{X} + \bar{X}\right)^2 + \left(X_3 - \bar{X} + \bar{X}\right)^2 + \dots\right)$$

$$\left(n\bar{X} - \sum_{i=2}^{n} X_i\right)^2 + \left(X_2^2 + X_3^2 + X_4^2 + \dots\right)$$

$$X_1^2 + \overbrace{X_2^2 + X_3^2 + \dots + X_n}$$

The pdf can then be rewritten

$$\frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\left(\left(Y_1 - \sum_{i=2}^{n} Y_i\right)^2 + \sum_{i=2}^{n} \left(Y_i + Y_1\right)^2\right)}$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\left(Y_1 - \sum_{i=2}^{n} Y_i\right)^2} e^{-\frac{1}{2}\sum_{i=2}^{n} \left(Y_i + Y_1\right)^2}$$

$$= \frac{1}{2\pi^{n/2}} e^{-\frac{1}{2}\left(Y_1^2 - \sum_{i=2}^{n} Y_1 Y_i - \left(\sum_{i=2}^{n} Y_i\right)^2\right)} e^{-\frac{1}{2}\left(\sum_{i=2}^{n} Y_i^2 + \sum_{i=2}^{n} Y_i Y_1 + \sum_{i=2}^{n} Y_1^2\right)}$$

$$= \frac{1}{2\pi^{n/2}} e^{-\frac{1}{2} n Y_1^2} e^{-\frac{1}{2}\left(\sum_{i=2}^{n} Y_i^2 - \left(\sum_{i=2}^{n} Y_i\right)^2\right)}$$

Thus $Y_1 = \bar{X}$ is independent of $Y_2, \dots, Y_n$, and S^2 can be written as a function only of $Y_2, \dots, Y_n$. Therefore $\bar{X}$ and S^2 are independent.

Part B

Show that $\bar{X}$ is UMVUE of μ

Step 1. Show that $\bar{X}$ is Unbiased

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = E\left[\frac{X_1}{n}\right] + \dots + E\left[\frac{X_n}{n}\right]$$

$$= \frac{\mu}{n} + \dots + \frac{\mu}{n} = n\left(\frac{\mu}{n}\right) = \mu$$

Step 2. Show that $\bar{X}$ is Sufficient

$$f(\vec{X}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2 + \mu x_i - \mu^2}{\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\sum_{i=1}^{n} \frac{-x_i^2}{\sigma^2} + \frac{\mu x_i}{\sigma^2} - \frac{\mu^2}{\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\sum_{i=1}^{n} \frac{-x_i^2}{\sigma^2}} \; e^{\sum_{i=1}^{n} \frac{\mu x_i}{\sigma^2} - \frac{\mu^2}{\sigma^2}}$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\sum_{i=1}^{n} \frac{-x_i^2}{\sigma^2}}}_{h(\vec{x})} \; \underbrace{e^{\frac{\mu n\bar{x} - \mu^2 n}{\sigma^2}}}_{g(\mu|\bar{x})}$$

Part B

Step 3. Show $\bar{X}$ is complete for μ

Exp Family: $f(x|\theta) = h(x)\, c(\theta) \exp\left(\sum_{i=1}^{k} w_i(\theta)\, t_i(x)\right)$

For parameters $\theta_1, \ldots, \theta_k$. In this case, σ^2 is known, we only are concerned with $\theta_1 = \mu$, and the statistic $\sum X_i$ is sufficient (and therefore $\bar{X}$ is sufficient) if $f(x|\theta)$ is in the Exp Family

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{x^2}{\sigma^2} - \frac{x\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{x^2}{\sigma^2}} e^{\frac{\mu^2}{\sigma^2}} \exp\left(-\frac{x\mu}{2\sigma^2}\right)$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{h(x)}\, \underbrace{e^{\frac{x^2}{\sigma^2}}}_{}\, \underbrace{e^{\frac{\mu^2}{\sigma^2}}}_{c(\mu)} \exp\left(\underbrace{-\mu}_{w(\mu)} \cdot \underbrace{\frac{x}{2\sigma^2}}_{t(x)}\right)$$

$w(\mu) = -\mu$ which contains open sets in $\mathbb{R}$ (it is not just a constant)

Therefore $\sum X_i$ (& $\bar{X}$) are suff for $\bar{X}$

thus $\bar{X}$ is complete & unbiased so $\bar{X}$ is UMVUE

Part C

First, Derive Score eq & Fisher info

$$L(w|\vec{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x_i-w)^2}{2\sigma^2}\right)$$

$$\ell(w|\vec{X}) = n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{n} \frac{(x_i-w)^2}{2\sigma^2}$$

$$\frac{d}{dw}\ell(w|\vec{X}) = -\sum_{i=1}^{n} \frac{2(x_i-w)}{2\sigma^2} = \sum_{i=1}^{n} \frac{(w-x_i)}{\sigma^2}$$

$$E\left[\left(\frac{d}{dw}\ell(w|\vec{X})\right)^2\right] = E\left[\sum_{i=1}^{n} \frac{(x_1-w)^2}{\sigma^4}\right]$$

$$= \frac{1}{\sigma^4} E\left[\sum_{i=1}^{n} (x_i-w)^2\right] = \frac{1}{\sigma^4} n\sigma^2 = \frac{n}{\sigma^2}$$

Thus, $I(w) = \frac{n}{\sigma^2}$. Thus CRLB is $\frac{1}{I(w)} = \frac{1}{n/\sigma^2} = \frac{\sigma^2}{n}$

$$Var(\bar{X}) = Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{1}{n^2} Var(X_1+\cdots+X_n)$$

Since $X_1+\cdots+X_n$ is the sum of independent variables,

$$= \frac{1}{n^2} Var\left(N(nw, n\sigma^2)\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

thus $\bar{X}$ achieves the CRLB for w

Part D

LRT $\mu = \mu_0$ vs $\mu \neq \mu_0$

LRT format: reject if $c \geq \frac{L_{H_0}}{L_{H_a}}$

$$\frac{\prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu_0)^2}{2\sigma^2}\right)}{\prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu_{MLE})^2}{2\sigma^2}\right)} = \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i-\mu_{MLE})^2-(x_i-\mu_0)^2\right)\right)$$

Now, we just need to derive μ_{MLE}. Luckily, this is easy to do by simply setting the previously derived score equation to zero.

$$\frac{d}{d\mu}\ell(\mu|\vec{x}) = \sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} \overset{set}{=} 0$$

$$\Rightarrow \sum_{i=1}^{n}(x_i-\mu_{MLE}) = 0 \Rightarrow \left(\sum_{i=1}^{n} x_i\right) - n\mu_{MLE} = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n\mu_{MLE} \Rightarrow \mu_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \mu_{MLE} = \bar{X}, \text{ so}$$

the LRT is just

reject if $c \geq \exp\left(\sum_{i=1}^{n}\left((x_i-\bar{x})^2-(x_i-\mu_0)^2\right)\right)$

Part E

Step 1. Show the LRT is UMP

Use Suff. NP-Lemma: a test based on a sufficient statistic is UMP if for pdf of S

$t \in S$ if $g(t|\theta_1) > k\, g(t|\theta_0)$

and

$t \in S^c$ if $g(t|\theta_1) < k\, g(t|\theta_0)$

For $k \geq 0$ where $\alpha = P_{\theta_0}(T \in S)$

Since $\bar{X}$ is normal with mean μ_0 / μ_{MLE} variance $\frac{\sigma^2}{n}$, it has pdf under the null

$$g(t|\mu_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu_0)^2}{2\sigma^2}\right)$$

$$g(t|\mu_{MLE}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\bar{x})^2}{2\sigma^2}\right)$$

The test is then

Part E

$t \in S$ if $\exp\left(-\frac{(t-\bar{x})^2}{2\sigma^2} - \frac{(t-\mu_0)^2}{2\sigma^2}\right) > k$

and

$t \in S^c$ if $\exp\left(-\frac{(t-\bar{x})^2}{2\sigma^2} - \frac{(t-\mu_0)^2}{2\sigma^2}\right) < k$

for $k \geq 0$ where $\alpha = P_{\theta_0}(T \in S)$

Since the SUFF statistic t is $\bar{X}$, this is

$\bar{x} \in S$ if $\exp\left(-\frac{(\bar{x}-\mu_0)^2}{2\sigma^2}\right) > k$

and

$\bar{x} \in S^c$ if $\exp\left(-\frac{(\bar{x}-\mu_0)^2}{2\sigma^2}\right) < k$

for $k \geq 0$ where $\alpha = P_{\theta_0}(\bar{X} \in S)$

Now, we just need to show that $\alpha = P_{\mu_0}(\bar{x} \in S)$

$$P_{\mu_0}(\bar{x} \in S) = P_{\mu_0}\left(\exp\left(-\frac{(\bar{x}-\mu_0)^2}{2\sigma^2}\right) > k\right)$$

$$= P_{\mu_0}\left(\pm\sqrt{2\sigma^2 \ln\left(\tfrac{1}{k}\right)} + \mu_0 \lessgtr \bar{x}\right) = P_{\mu_0}(c^+ \leq \bar{x},\ c^- \geq \bar{x}) \text{ for some } c^+, c^-$$

Part E

The rejection probability under the null hypothesis $P(c^+ \le \bar{x}, c^- \ge \bar{x})$ is known since $\bar{X}$ has normal distribution. So the UMP α-level test can be formulated as

$$\bar{X} \in R \quad \text{if} \quad c\, z_{\frac{\alpha}{2}} \le \bar{X} \le c\, z_{1-\frac{\alpha}{2}}$$

Part F

Bc $\bar{X}$ is also normal, and σ^2 is known, a CI for μ can be found by pivoting the standard normal distribution.

$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim Z$, so the 95% CI can be found with the CDF of Z, $z_{\frac{\alpha}{2}}$, $z_{1-\frac{\alpha}{2}}$

$$\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} \le \mu \le \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$$