Skip the review on pages 1-7. If a review were desired, a better one could be found elsewhere.

A probability space consists of 3 parts

Ω) The sample space, which is a set containing all possible outcomes,

F) A collection of Events, each of which is a subset of Ω,

$\mathbb{P}$) A function that assigns probabilities to events in F.

The collection of events, F, satisfies these properties:

1) If $A_i \in F$, then $\bigcup_{i=1}^{\infty} A_i \in F$

2) If $A \in F$, then $A^c \in F$

3) $F \neq \emptyset$

In other words, F is a sigma algebra.

Example: Show that if $A, B \in F$, then $A \cap B \in F$

The function, $\mathbb{P}$, satisfies these properties:

1) $\mathbb{P}(\Omega) = 1$.

2) If A_i are disjoint, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

3) $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in F$.

Example: Show that $\mathbb{P}(\emptyset) = 0$.

A random variable is a function whose domain is the sample space and whose codomain is $\mathbb{R}$: $X : \Omega \to \mathbb{R}$.

In order to be a random variable, $\{\omega \in \Omega \mid X(\omega) \leq t\}$ must be an element of F for all $t \in \mathbb{R}$

The cummulative distribution function, $F(t)$, is defined as $F(t) = \mathbb{P}(X \leq t)$
i.e. as $F(t) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \leq t\})$.

A c.d.f., $F(t)$, satisfies

1) $\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to \infty} F(t) = 1$,

2) $s \le t \Rightarrow F(s) \le F(t)$,

3) $F(t)$ is right continuous, i.e.,
$\lim_{t \to a+} F(t) = F(a)$ for all $a \in \mathbb{R}$

and $\lim_{t \to a-} F(t)$ exists and is finite $\forall a \in \mathbb{R}$.

It can also be shown that if a function, F, satisfies 1, 2, and 3 above, then there exists a probability space $(\Omega, F, \mathbb{P})$ and a random variable X on that space such that F is the cdf of X.

There are 2 main types of distributions that we will discuss.

A r.v. is discrete if $\{x \in \mathbb{R} \mid \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\}) > 0\}$ is finite or countably infinite, and $\sum \mathbb{P}(X=x) = 1$.

Examples: Binomial, Poisson, negative binomial, & geometric.

$X \sim BIN(n,p)$ if $\mathbb{P}(X=x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{0,1,\dots,n\}, \\ 0 & o/w. \end{cases}$

$X \sim POI(\lambda)$ if $\mathbb{P}(X=x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & x \in \{0,1,2,\dots\}, \\ 0 & o/w. \end{cases}$

$X \sim NB(r,p)$ if $\mathbb{P}(X=x) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & x \in \{r, r+1, \dots\}, \\ 0 & o/w. \end{cases}$

$X \sim GEO(p)$ if $\mathbb{P}(X=x) = \begin{cases} (1-p)^{x-1} p & x \in \{1,2,\dots\}, \\ 0 & o/w. \end{cases}$

A r.v. is continuous if there exists a function, f, such that $F(t) = \int_{-\infty}^{t} f(s)\,ds$ for all $t \in \mathbb{R}$.

Note that $f \geq 0$ and $\int_{-\infty}^{\infty} f(s)\,ds = 1$.

Question: Is f unique? No.

Is F unique? Yes.

Note that if F is differentiable, then $F'(t)$ is a density function for X.

Examples: Normal, exponential, uniform.

$X \sim N(\mu, \sigma^2)$ if $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$.

$X \sim EXP(\lambda)$ if $f(t) = \begin{cases} \frac{1}{\lambda} e^{-t/\lambda} & t \geq 0, \\ 0 & o/w. \end{cases}$

$X \sim UNIF(a,b)$ if $f(t) = \begin{cases} \frac{1}{b-a} & t \in (a,b), \\ 0 & o/w. \end{cases}$

Question: Are there random variables that are neither discrete nor continuous? Yes.

Example: Suppose your car is in a crash and that the damage to your car is $X \sim EXP(5{,}000)$. If your deductible is \$1000, then the insurance company will pay $Y = \max(0, X-1000)$.

Y is not discrete because the set of possible outcomes is uncountable: $[0, \infty)$.

Y is not continuous because there is no function, f, such that $F(t) = \int_{-\infty}^{t} f(s)\,ds \quad \forall t \in \mathbb{R}$.

If X is discrete, then the expected value is given by $E(X) = \sum_{\{x \mid \mathbb{P}(X=x)>0\}} x\,\mathbb{P}(X=x)$ provided that the sum converges absolutely, i.e., $\sum |x|\,\mathbb{P}(X=x) < \infty$. If the sum does not converge absolutely, then the expected value does not exist.

Similarly, the expected value of a continuous r.v. is $E(X) = \int_{\mathbb{R}} x f(x)\,dx$ provided that $\int_{\mathbb{R}} |x| f(x)\,dx < \infty$.

Example: Suppose that $\mathbb{P}(X = i(-1)^i) = \frac{6}{(\pi i)^2}$ for $i = 1, 2, \ldots$

Then $\sum_{i=1}^{\infty} i(-1)^i \frac{6}{(\pi i)^2} = \sum_{i=1}^{\infty} (-1)^i \frac{6}{\pi^2 i} = \log(2)$, but

the expectation does not exist.

The variance of a r.v., X, is its second central moment, $E(X - E(X))^2$, provided that the expected value exists.

Markov's Inequality states that if $X \geq 0$, $a > 0$ and $E(X)$ exists, then $\mathbb{P}(X \geq a) \leq \frac{E(X)}{a}$. Prove this!

Chebyshev's inequality states that if $E(X)$ exists, $\mathrm{var}(X) \notin \{0, \infty\}$, and $k > 0$, then $\mathbb{P}(|X - \mu| \geq k) \leq \frac{\mathrm{var}(X)}{k^2}$. Prove this!

The joint cdf, $F(x,y)$, describes the joint distribution of two random variables, X and Y. $F(x,y)$ is defined as $F(x,y) = \mathbb{P}(X \le x, Y \le y)$.

For discrete r.v.'s, this looks like

$$F(x,y) = \sum_{\{(s,t) \mid s \le x,\ t \le y,\ \text{and } \mathbb{P}(X=s, Y=t) > 0\}} \mathbb{P}(X=s, Y=t)$$

For continuous r.v.'s, this looks like

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

When the distribution of a random variable is derived from it's joint distribution with another r.v., then it is referred to as the marginal distribution.

$$F_X(x) = \mathbb{P}(X \le x) = \lim_{y \to \infty} F(x,y)$$

$$F_Y(y) = \lim_{x \to \infty} F(x,y)$$

Example: Look at how this behaves for jointly continuous r.v.'s:

$$F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(s,t)\, dt\, ds \quad \text{and} \quad f_X(x) = \int_{-\infty}^{\infty} f(x,t)\, dt$$

Two random variables, X and Y, are independent if $\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y) \quad \forall\, x, y \in \mathbb{R}$.

Equivalently, $f_{X,Y}(x,y) = f_X(x) f_Y(y) \quad \forall\, x, y \in \mathbb{R}$

Example: Show that the second definition implies the first.

Start here. Suppose X has density function, f, and $Y = h(X)$.

Question: How can we find the density of Y?

The first method we'll discuss is the cdf technique and is best illustrated with an example:

Example (6.2.1): Suppose $F_X(x) = \begin{cases} 1 - e^{-2x} & x > 0, \\ 0 & o/w. \end{cases}$

Let $Y = e^X$. Then if we want $F_Y(y)$:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(e^X \le y) = \begin{cases} \mathbb{P}(X \le \log(y)) & y > 0, \\ 0 & y \le 0. \end{cases}$$

$$= \begin{cases} F_X(\log(y)) & y > 0, \\ 0 & o/w. \end{cases}$$

We can then also find the density:

$$f_Y(y) = F'(y) = \begin{cases} f_X(\log(y)) \frac{1}{y} & y > 0, \\ 0 & o/w. \end{cases} = \begin{cases} 2/y^3 & y > 1, \\ 0 & o/w. \end{cases}$$

Example (6.2.2) Let $Y = X^2$.

Then $\mathbb{P}(Y \le y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$,

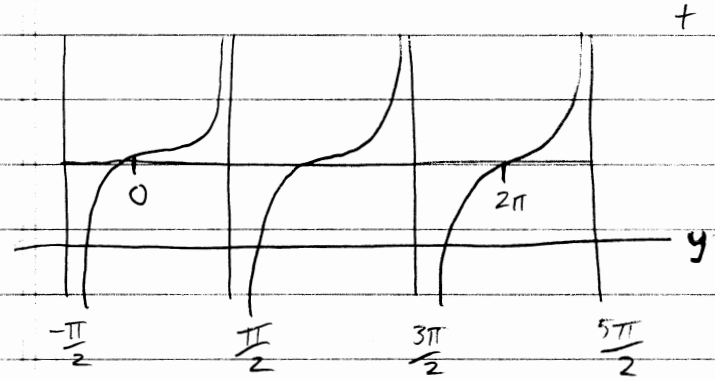
and $f_Y(y) = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{2\sqrt{y}}$

Example (6.2.3) Let $\Theta \sim \text{UNIF}(0, 2\pi)$

Let $Y = \tan(\Theta)$.

Assume for a moment that $y < 0$.

Then $F_Y(y) = \mathbb{P}(\tan(\Theta) \le y) = \mathbb{P}\left(\frac{\pi}{2} \le \Theta \le \tan^{-1}(y) + \pi\right)$
$+ \mathbb{P}\left(\frac{3\pi}{2} \le \Theta \le \tan^{-1}(y) + 2\pi\right)$



$= \tan^{-1}(y) + \pi - \frac{\pi}{2}$
$+ \tan^{-1}(y) + 2\pi - \frac{3\pi}{2}$

$= \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(y)$

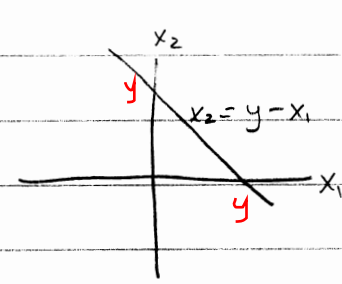
This can also be shown to be true for $y > 0$.

We can also find the density to be
$f_Y(y) = \frac{1}{\pi(1+y^2)} \quad \forall y \in \mathbb{R}$

Example: (6.2.4) Let $X_i \sim$ iid EXP(1).

Let $Y = X_1 + X_2$.

Then $F_Y(y) = \mathbb{P}(X_1 + X_2 \le y) = \int_0^y \int_0^{y-x_1} e^{-(x_1+x_2)}\, dx_2\, dx_1$



Evaluate the integral and differentiate to obtain the density function for Y.

Example: Let $X \sim UNIF(0,2)$

Let $Y = \begin{cases} 1 & \text{if } X > 1, \\ 0 & \text{o/w} \end{cases}$

i.e. $Y = h(X)$ where $h(x) = 1\{x > 1\}$.

Let $W = \max(0, X-1)$

We see that $Y \sim BER(\frac{1}{2})$ and W is neither discrete nor continuous; it is zero with probability $\frac{1}{2}$ but conditional on $W > 0$, it has a $UNIF(0,1)$ distribution.

Find the derivative of $F_W(w)$ and note that you do <u>NOT</u> get a density function — it integrates to $\frac{1}{2}$ instead of 1.

The above examples show that we must be somewhat cautious.

The next method is the transformation method.

Let $Y = h(X)$ and assume h is one-to-one.

Note that the nasty example above came from using transformations that were <u>not</u> 1-to-1.

We will use the notation $y(x)$ and $h(x)$ interchangably when $y = h(x)$. Also, $x(y)$ will be used in place of $h^{-1}(y)$ whenever convenient (and h^{-1} exists).

Theorem (6.3.1) If X is discrete, $Y = h(X)$, and h is 1-to-1 Then $f_Y(y) = f_X(x(y))$.

Equivalently, $f_Y(y) = f_X(h^{-1}(y))$.

Proof: $f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(h(X) = y) = \mathbb{P}(X = h^{-1}(y)) = \mathbb{P}(X = x(y))$

$$= f_X(x(y)) \quad \square$$

Example: (6.3.1) Let $X \sim GEO(p)$ so that $f_X(x) = \begin{cases} (1-p)^{x-1} p & x = 1, 2, \ldots \\ 0 & o/w. \end{cases}$

Let $Y = X - 1$. Note that Y is often also called geometric. If X is the number of trials until a success, then Y is the number of failures before a success.

By the above theorem, we have $f_Y(y) = f_X(y+1)$

$$= \begin{cases} (1-p)^y p & y = 0, 1, \ldots \\ 0 & o/w. \end{cases}$$

Lets try the cdf method for determining $f_Y(y)$:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X-1 \le y) = \mathbb{P}(X \le y+1) = F_X(y+1).$$

We need to compute $F_X(x) = \sum_{t=1}^{\lfloor x \rfloor} (1-p)^{t-1} p$

$$= \begin{cases} \dfrac{p\left(1-(1-p)^{\lfloor x \rfloor}\right)}{p} & x \ge 1 \\ 0 & o/w \end{cases}$$

$$= \begin{cases} 1-(1-p)^{\lfloor x \rfloor} & x \ge 1, \\ 0 & o/w. \end{cases}$$

Thus $F_Y(y) = F_X(y+1) = \begin{cases} 1-(1-p)^{\lfloor y+1 \rfloor} & y \ge 0 \\ 0 & o/w \end{cases}$

For a discrete r.v. the mass function is given by $f_Y(y) = \lim_{\delta \to 0+} \left(F_Y(y) - F_Y(y-\delta) \right)$

$$= F_Y(y) - \lim_{\delta \to 0+} F_Y(y-\delta).$$

Thus in our case, $f_Y(y) = \begin{cases} \left[1-(1-p)^{y+1}\right] - \left[1-(1-p)^{y}\right], & y = 1, 2, \ldots \\ \left[1-(1-p)^{1}\right] - [0] & y = 0 \\ 0 & o/w \end{cases}$

$$= \begin{cases} p(1-p)^y & y = 0, 1, 2, \ldots \\ 0 & o/w \end{cases}$$

As we have seen, one method may be easier than the other depending on the distribution of X, the transformation h, and what is already known.

Theorem (6.3.2) If X is continuous, $Y = h(X)$, and h is 1-to-1, then $f_Y(y) = f_X(x(y)) \left| \frac{d}{dy} x(y) \right|$, as long as $\frac{d}{dy} x(y)$ is continuous and never zero on $B = \{y \mid f_Y(y) > 0\}$. B can also be written as $B = \{h(x) \mid f_X(x) > 0\}$.

Proof: Because $\frac{d}{dy} x(y)$ is continuous and non-zero on B, $x(y)$ is monotonic on B.

We will prove the result when h is increasing and B is an interval.

We use the cdf method:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(h(X) \le y) = \mathbb{P}(X \le h^{-1}(y)) = \mathbb{P}(X \le x(y))$$
$$= F_X(x(y)) \quad \text{for } y \in B. \quad (\text{If } y \notin B, \text{ then } F(y) \in \{0,1\} \text{ and } f_Y(y) = 0)$$

Thus, $f_Y(y) = f_X(x(y)) \cdot \frac{d}{dy} x(y) = f_X(x(y)) \left| \frac{d}{dy} x(y) \right| \cdot 1\{y \in B\}$.

We've slightly cheated on the proof. If B is not an interval (e.g. $[0,1] \cup [2,3]$) then $x(y)$ may not be monotonic.

However, the result is easily shown on each interval separately.

Note that $\frac{d}{dy} x(y)$ is called the "Jacobian" of the transformation.

Example: Let $F_X(x) = \begin{cases} 1 - e^{-2x} & x > 0, \\ 0 & x \leq 0. \end{cases}$

Let $Y = e^X$.

We found $f_Y(y)$ previously using the cdf method (Example 6.2.1). We now use the transformation method.

$$f_Y(y) = f_X(x(y)) \left| \frac{d}{dy} x(y) \right| = \begin{cases} 2e^{-2\log(y)} \left|\frac{1}{y}\right| & y > 1, \\ 0 & o/w. \end{cases}$$

$$= \begin{cases} \frac{2}{y^3} & y > 1, \\ 0 & y \leq 1. \end{cases}$$

Example: Let $X \sim N(\mu, \sigma^2)$ and let $Y = e^X$. Find the density of $Y \sim LOGN(\mu, \sigma^2)$.

Note that $h^{-1}(y) = \log(y)$ which is differentiable with continuous, non-zero derrivative on $B = (0, \infty)$.

Thus $f_Y(y) = f_X(\log(y)) \left|\frac{1}{y}\right| 1\{y \in B\} = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\log(y) - \mu)^2} \frac{1}{y} & y > 0, \\ 0 & o/w. \end{cases}$

Example: Let $X \sim PAR(1,1)$ so that $f(x) = \begin{cases} \frac{1}{(1+x)^2} & x > 0 \\ 0 & o/w \end{cases}$

Let $Y = \log(X)$

Then $h^{-1}(y) = e^y$, which is differentiable with non-zero continuous derrivate on $B = \mathbb{R}$.

Thus $f_Y(y) = f_X(e^y)\, e^y = \frac{e^y}{(1+e^y)^2}$.

Theorem (6.3.3) If X is a continuous r.v., then $F(X) \sim UNIF(0,1)$

Proof: We prove the result only in the special case where F is invertible.

Let $U = F(X)$.

Then $F_U(u) = \mathbb{P}(F(X) \le u) = \begin{cases} 0 & u \le 0, \\ \mathbb{P}(X \le F^{-1}(u)) & u \in (0,1), \\ 1 & u \ge 1. \end{cases}$

Since F and hence F^{-1} are monotonically increasing.

$= \begin{cases} 0 & u \le 0, \\ F(F^{-1}(u)) = u & u \in (0,1), \\ 1 & u \ge 1. \end{cases}$

Theorem (6.3.4) Let $G(u) = \min\{x \mid u \le F(x)\}$ and $U \sim UNIF(0,1)$. Then $X = G(U) \sim F$.

Proof left as exercise 5.
Note that $G = F^{-1}$ if F is invertible.

This theorem allows us to generate data according to any distribution by first generating data according to a $UNIF(0,1)$ distribution.

Before we move on to joint transformations, let's do one more example.

Example: Let $X \sim UNIF(-1,4)$.
Let $Y = X^4$.
Find the density of Y.

$h(x) = x^4$ is not 1-to-1 on $(-1,4)$, so we will use the cdf method. Later we will learn how to use the transformation method in cases like this.

$$F_Y(y) = \mathbb{P}(X^4 \le y) = \begin{cases} \mathbb{P}\left(-y^{1/4} \le X \le y^{1/4}\right) & y \ge 0, \\ 0 & \text{o/w.} \end{cases}$$

$$= \begin{cases} F_X(y^{1/4}) - F_X(-y^{1/4}) & y \ge 0, \\ 0 & \text{o/w.} \end{cases}$$

Thus $f_Y(y) = \begin{cases} f_X(y^{1/4})\frac{1}{4}y^{-3/4} + f_X(-y^{1/4})\frac{1}{4}y^{-3/4} & y \geq 0 \\ 0 & \text{o/w} \end{cases}$

$$= \begin{cases} \frac{1}{5}\cdot\frac{1}{4}y^{-3/4} + \frac{1}{5}\cdot\frac{1}{4}\cdot y^{-3/4} & y \in [0,1) \\ \frac{1}{5}\cdot\frac{1}{4}y^{-3/4} & y \in [1, 4^4] \\ 0 & \text{o/w} \end{cases}$$

$$= \begin{cases} \frac{1}{10}y^{-3/4} & y \in [0,1), \\ \frac{1}{20}y^{-3/4} & y \in [1, 4^4], \\ 0 & \text{o/w}. \end{cases}$$

Theorem: (6.3.6) If $\underline{X} = (X_1, X_2, \ldots, X_k)$ is continuous
$\underline{Y} = (Y_1, Y_2, \ldots, Y_k)$
$\underline{Y} = h(\underline{X})$, that is, $Y_1 = h_1(X_1, X_2, \ldots X_k)$, etc.
h is 1-to-1

Then $f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{x}(\underline{y}))\left|\det\left(\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}}\right)\right| \mathbb{1}\{y \in B\}$

where $\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_k} \\ \vdots & & & \\ \frac{\partial x_k}{\partial y_1} & \frac{\partial x_k}{\partial y_2} & \cdots & \frac{\partial x_k}{\partial y_k} \end{bmatrix}$ and B is the range of the transformation.

as long as $\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}}$ is continuous and non-zero over the range of the transformation

Theorem (6.3.5) If $\underline{X}$ is discrete, $\underline{Y} = h(\underline{X})$, and h is 1-to-1, then $f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{x}(\underline{y}))$

Proof: $f_{\underline{Y}}(\underline{y}) = \mathbb{P}(\underline{Y} = \underline{y}) = \mathbb{P}(h(\underline{X}) = \underline{y}) = \mathbb{P}(\underline{X} = h^{-1}(\underline{y})) = \mathbb{P}(\underline{X} = \underline{x}(\underline{y}))$

$$= f_{\underline{X}}(\underline{x}(\underline{y})). \qquad \square$$

Proof of 6.3.6: Let B be the range of the transformation h on $\underline{X}$. Let $D \subseteq B$. Let C be the preimage of D under h: $C = \{\underline{x} \mid h(\underline{x}) \in D\}$.

Then, $\mathbb{P}(\underline{Y} \in D) = \int_D \cdots \int f_{\underline{Y}}(\underline{y})\, dy_1 dy_2 \cdots dy_k$

$$= \mathbb{P}(\underline{X} \in C) = \int_C \cdots \int f_{\underline{X}}(\underline{x})\, dx_1 dx_2 \cdots dx_k$$

$$= \int_D \cdots \int f_{\underline{X}}(\underline{x}(\underline{y})) \left| \det\left(\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}} \right) \right| dy_1 \cdots dy_k$$

by a change of variables $\underline{y} = h(\underline{x})$. $\square$

It can be helpful to do the proof in this way for a univariate transformation first.

Example: Let $X_1, X_2 \sim$ iid $EXP(1)$.
Let $Y_1 = X_1$
$Y_2 = X_1 + X_2$
Find $f_{\underline{Y}}(\underline{y}) = f_{Y_1,Y_2}(y_1, y_2)$.

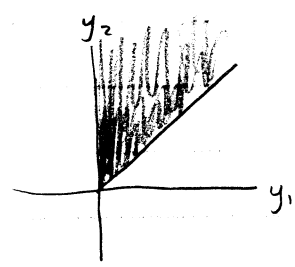
First, solve for $\underline{x}$:

$$x_1(\underline{y}) = y_1$$
$$x_2(\underline{y}) = y_2 - y_1$$

Next find the Jacobian and its determinant:

$$\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

and $\left| \det\left(\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}} \right) \right| = 1.$



Thus $f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{x}(\underline{y})) \left| \det\left(\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}} \right) \right|$

$$= \begin{cases} e^{-y_1 - (y_2 - y_1)} & \text{if } y_1 > 0,\ y_2 - y_1 > 0, \\ 0 & \text{o/w.} \end{cases}$$

$$= \begin{cases} e^{-y_2} & y_1 > 0, \quad y_2 > y_1, \\ 0 & o/w. \end{cases}$$

Example: Now find the marginals.

$$f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 = \begin{cases} \int_{y_1}^{\infty} e^{-y_2}\, dy_2 & y_1 > 0, \\ 0 & o/w. \end{cases}$$

$$= \begin{cases} e^{-y_1} & y_1 > 0 \\ 0 & y_1 \leq 0 \end{cases}$$

$$f_{Y_2}(y_2) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1, y_2)\, dy_1 = \begin{cases} \int_0^{y_2} e^{-y_2}\, dy_1 & y_2 > 0 \\ 0 & y_2 \leq 0 \end{cases}$$

$$= \begin{cases} y_2 e^{-y_2} & y_2 > 0 \\ 0 & y_2 \leq 0 \end{cases}$$

Example (6.3.13)

$$Y_1 = X_1 - X_2$$
$$Y_2 = X_1 + X_2$$
$$X_1, X_2 \sim \text{iid } EXP(1)$$

Find $f_{Y_1,Y_2}(y_1,y_2)$.

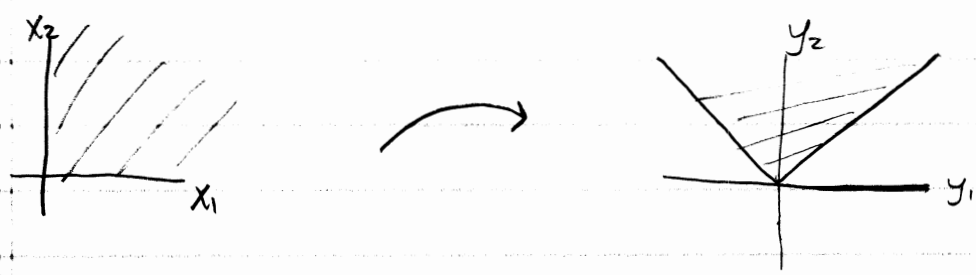
$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(x_1(\underline{y}), x_2(\underline{y}))\left|\det\left(\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}}\right)\right|$$

where $f_{X_1,X_2}(x_1,x_2) = \begin{cases} e^{-x_1-x_2} & x_1, x_2 > 0, \\ 0 & \text{o/w}. \end{cases}$

$$x_1(\underline{y}) = \frac{y_1+y_2}{2} \qquad x_2(\underline{y}) = \frac{y_2-y_1}{2}$$

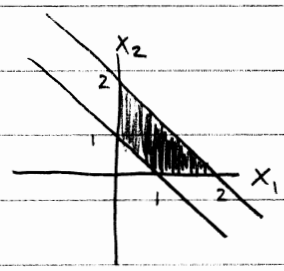
$$\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Thus $f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} e^{-y_2}\cdot\frac{1}{2} & y_2 > 0,\ -y_2 < y_1 < y_2 \\ 0 & \text{o/w} \end{cases}$



Note that one easy way to determine the range of the transformation is to transform the boundary, which, under a continuous transformation, will map to the boundary of the range.

Depending on time and interest, we may
- find the marginal for Y_1: $f_{Y_1}(y_1) = \frac{1}{2}e^{-|y_1|}$
- find the marginal for Y_2: $f_{Y_2}(y_2) = y_2 e^{-y_2} \mathbb{1}\{y_2 > 0\}$
- find $\mathbb{P}(Y_2 \in (1,2))$
  - using the marginal: $\int_1^2 y e^{-y}\,dy$
  - using the joint density of Y_1, Y_2: $\int_1^2 \int_{-y_2}^{y_2} \frac{1}{2} e^{-y_2}\,dy_1 dy_2$
  - using the joint density of X_1, X_2: $\int_0^1 \int_{1-x_1}^{2-x_1} e^{-x_1-x_2}\,dx_2\,dx_1 + \int_1^2 \int_0^{2-x_1} e^{-x_1-x_2}\,dx_2\,dx_1$



Let's now consider transformations that are not 1-to-1 over $A = \{x \mid f(x) > 0\}$.

It is often possible to partition A into disjoint subsets $A_1, A_2, \ldots A_J$ such that h is 1-to-1 on each A_j

Then $f_Y(y) = f_X(x^{(1)}(y)) + f_X(x^{(2)}(y)) + \ldots + f_X(x^{(J)}(y))$ (for X discrete)

where $x^{(j)}(y)$ is the inverse of h on A_j

Example: suppose X is discrete with $f(x) = \begin{cases} 1/5 & x \in \{-2,-1,0,1,2\} \\ 0 & o/w. \end{cases}$

Let $Y = X^2$

Then $f_Y(4) = f_X(x^{(1)}(4)) + f_X(x^{(2)}(4))$

where $A_1 = \{0, 1, 2\}$
$A_2 = \{-2, -1\}$

Thus $f_Y(4) = f_X(2) + f_X(-2)$

$= \frac{1}{5} + \frac{1}{5}$

$= 2/5$

A similar extention exists when X is continuous:

$$f_Y(y) = \sum_j f_X(x^{(j)}(y)) \cdot \left| \det\left(\frac{d}{dy} x^{(j)}(y)\right)\right| \quad \text{for } y \in B$$

This can also extend to joint transformations in the obvious way.

Example (6.3.8) Let $X \sim UNIF(-1,1)$
$Y = X^2$

$A_1 = (-1, 0)$ $\quad x^{(1)}(y) = -\sqrt{y}$
$A_2 = [0, 1)$ $\quad x^{(2)}(y) = \sqrt{y}$

$$f_Y(y) = f_X(-\sqrt{y}) \cdot \left|-\tfrac{1}{2} y^{-\frac{1}{2}}\right| + f_X(\sqrt{y}) \cdot \left|\tfrac{1}{2} y^{-1/2}\right| \quad \text{for } y \in (0,1)$$

$$= \frac{1}{2\sqrt{y}} \cdot 1\{y \in (0,1)\}$$

If there is time and interest, rework the previous example with $X \sim UNIF(-1, 2)$.

It is common to have a r.v. that is defined as a sum of other r.v.'s. If we choose a nice transformation we can find the distribution of the sum.

Example: (6.4.3) $X_i \sim GAM(1, \alpha_i)$ are independent.

$$Y_1 = \frac{X_1}{X_1+X_2+X_3}, \quad Y_2 = \frac{X_2}{X_1+X_2+X_3}, \quad Y_3 = X_1+X_2+X_3$$

Note that $f_{X_i}(x_i) = \frac{1}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} e^{-x_i} 1\{x_i > 0\}$.

$$x_1 = y_1 y_3 \qquad x_2 = y_2 y_3 \qquad x_3 = y_3 - y_1 y_3 - y_2 y_3$$

$$\frac{\partial \underline{x}(\underline{y})}{\partial \underline{y}} = \begin{bmatrix} y_3 & 0 & y_1 \\ 0 & y_3 & y_2 \\ -y_3 & -y_3 & (1-y_1-y_2) \end{bmatrix}$$

$$f_{Y_1,Y_2,Y_3}(y_1,y_2,y_3) = \frac{1}{\Gamma(\alpha_1)}\frac{1}{\Gamma(\alpha_2)}\frac{1}{\Gamma(\alpha_3)} (y_1 y_3)^{\alpha_1 - 1} (y_2 y_3)^{\alpha_2 - 1} \left[y_3(1-y_1-y_2)\right]^{\alpha_3 - 1}$$

$$\cdot\, e^{-y_3} 1\{y_1 > 0\} 1\{y_2 > 0\} 1\{y_3 > 0\}$$

$$\cdot\, 1\{y_1 + y_2 < 1\}$$

An alternative method is with moment generating functions.

Recall that $M_X(t) = E(e^{tX})$.

Also recall that a moment generating function uniquely determines the associated distribution

Theorem (6.4.1) If X_i are independent and $Y = \sum_{i=1}^{n} X_i$, then $M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$.

proof: $M_Y(t) = E\left(e^{t\sum_{i=1}^{n} X_i}\right) = E\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} E\left(e^{tX_i}\right)$

$$= \prod_{i=1}^{n} M_{X_i}(t) \qquad \square$$

Example: Let $X_1 \sim GAM(1, \pi)$ $X_2 \sim GAM(1, e)$.
Let $Y = X_1 + X_2$. Assume X_1, X_2 are independent.
Then $M_Y(t) = M_{X_1}(t) M_{X_2}(t) = \left(\frac{1}{1-t}\right)^{\pi} \left(\frac{1}{1-t}\right)^{e}$

$$= \left(\frac{1}{1-t}\right)^{\pi+e} \sim GAM(1, \pi+e)$$

Example (6.4.4) Let $Y = X_1 + X_2$ with $X_1 \sim BIN(n, 1/3)$
$X_2 \sim BIN(m, 1/3)$

Assume X_1, X_2 are independent.

Then $M_Y(t) = M_{X_1}(t) M_{X_2}(t) = \left(\frac{1}{3}e^t + (1-\frac{1}{3})\right)^n \cdot \left(\frac{1}{3}e^t + (1-\frac{1}{3})\right)^m$

$$= \left(\frac{1}{3}e^t + \frac{2}{3}\right)^{n+m} \sim BIN(n+m, \tfrac{1}{3})$$

Example: (6.4.5) If X_1, X_2 are independent with $X_1 \sim POI(5)$
$X_2 \sim POI(17)$

Then $M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t) = e^{5(e^t-1)} e^{17(e^t-1)}$

$$= e^{(5+17)(e^t-1)}$$

$$\sim POI(22)$$

Example: (6.4.7) If X_1, X_2 are independent with $X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2 \sim N(\mu_2, \sigma_2^2)$

Then $M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t) = e^{\mu_1 t + \sigma_1^2 t^2/2} e^{\mu_2 t + \sigma_2^2 t^2/2}$

$$= e^{(\mu_1+\mu_2)t + (\sigma_1^2+\sigma_2^2)t^2/2} \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$$

Suppose $X_i \sim$ iid F.

We can think about $X_1, X_2, \ldots, X_n$ as a sample of size n from a population with distribution F.

Because of independence, we would need to be sampling <u>with</u> replacement.

However, when the size of the population is large, then the choice of sampling with vs. without replacement is inconsequential

Example Consider a population of 4 balls — half red and half blue.

Sample without replacement and let

$$X_i = \begin{cases} 1 & \text{if the } i^{\underline{th}} \text{ ball is Red,} \\ 0 & \text{o/w.} \end{cases}$$

Sample <u>with</u> replacement and let

$$Y_i = \begin{cases} 1 & \text{if the } i^{\underline{th}} \text{ ball is Red,} \\ 0 & \text{o/w.} \end{cases}$$

$\mathbb{P}(X_2 = 1 \mid X_1 = 1) = \frac{1}{3} \neq \frac{1}{2} = \mathbb{P}(X_2 = 1)$

So X_i's are not independent, but

$\mathbb{P}(Y_2 = 1 \mid Y_1 = 1) = \frac{1}{2} = \mathbb{P}(Y_2 = 1)$

So Y_i's are independent.

Now, consider a population of size $2n$.

Then, $\mathbb{P}(X_2 = 1 \mid X_1 = 1) = \frac{n-1}{2n-1}$,
$\mathbb{P}(X_2 = 1) = 1/2$.

Notice that $\frac{n-1}{2n-1} \longrightarrow \frac{1}{2}$ as n gets large.

You can think about this as the X_i's becoming less dependent as n gets large.

In practice we nearly always assume independence when sampling from a <u>large</u> population without replacement.

Suppose you have random sample of size n from a population with distribution F, i.e. $X_1, X_2, \ldots, X_n \sim$ iid F

Question: What is the joint density function?

$f(x_1, x_2, \ldots x_n)$.

Let $X_{k:n}$ represent the kth largest observation from the sample.

In particular, $X_{1:n}$ is the smallest and $X_{n:n}$ is the largest.

Note that $X_{k:n}$ is a random variable.

In practice, when we take a sample, what we get is the outcome $(x_1, \ldots, x_n)$ of a random sample $(X_1, X_2, \ldots, X_n)$.

Example: Let $X = \begin{cases} 1 & \text{if heads,} \\ 0 & \text{o/w.} \end{cases}$

based on the toss of a fair coin.

Once we toss the coin (if it lands on heads) then we will have $x = 1$

It is not appropriate to say $X = 1$ because X is a r.v..

If a coin is tossed n times with
$$X_i = \begin{cases} 1 & \text{if heads on } i^{\text{th}} \text{ toss,} \\ 0 & \text{o/w.} \end{cases}$$

then $(X_1, X_2, \ldots, X_n)$ is a random sample from a $BER(\frac{1}{2})$ population and $(x_1, x_2, \ldots, x_n)$ is the outcome.

Remember r.v.'s are functions $X(H) = 1$
$X(T) = 0$

but outcomes are just constants.

Example If you toss a coin 5 times you may get $(x_1, x_2, x_3, x_4, x_5) = (1, 1, 0, 1, 0)$ as the outcome of the random sample.

Example: Suppose $X_i \sim POI(5)$.
Then $(X_1, \ldots, X_n)$ is a random sample of size n.
$X_{n:n} = \max\{X_1, \ldots, X_n\}$ is a r.v.
Suppose you get an outcome of $(8, 5, 7, 3, 4)$.
Then $x_{n:n} = \max\{x_1, \ldots, x_n\} = \max\{8, 5, 7, 3, 4\} = 8$.

Theorem (6.5.1) If $(X_1, \ldots, X_n)$ is a random sample from a continuous pdf, f, then the joint density of the order statistics is

$$n!\, f(y_1) f(y_2) \cdots f(y_n)\, \mathbb{1}\{y_1 < y_2 < \ldots < y_n\}$$

Prove this for discrete random variables first. This provides great intuition for the continous case.

Proof: ($n=2$) Let $X_1, X_2 \sim$ iid F

Then Let $Y_1 = X_{1:2}$

$Y_2 = X_{2:2}$

Note that the transformation is not 1-to-1 because $(2,3)$ and $(3,2)$ both map to $(2,3)$.

We partition $A = \{(x_1, x_2) \mid f(x_1, x_2) > 0\}$ into

$A_1 = \{(x_1, x_2) \mid x_1 < x_2 \text{ and } f(x_1, x_2) > 0\}$

$A_2 = \{(x_1, x_2) \mid x_1 \geq x_2 \text{ and } f(x_1, x_2) > 0\}$

Then
$$f_{Y_1, Y_2}(y_1, y_2) = \sum_{i=1}^{2} f_{X_1, X_2}\left(x_1^{(i)}(\underline{y}), x_2^{(i)}(\underline{y})\right) \cdot \left|\det\left(\frac{\partial \underline{x}^{(i)}(\underline{y})}{\partial \underline{y}}\right)\right| \cdot \mathbb{1}\{y_1 < y_2\}.$$

Note that $x_1^{(1)}(\underline{y}) = y_1$ $\quad$ $x_1^{(2)}(\underline{y}) = y_2$

$x_2^{(1)}(\underline{y}) = y_2$ $\quad$ $x_2^{(2)}(\underline{y}) = y_1$

Finish this calculation to complete the proof in the $n=2$ case.

Example (6.5.1) $X_1, X_2, X_3 \sim iid \quad f(x) = 2x\, 1\{x \in (0,1)\}$

Find the joint density of the order statistics, $X_{1:3}, X_{2:3}, X_{3:3}$.

$$f_{\underline{Y}}(y_1, y_2, y_3) = f(y_1) f(y_2) f(y_3) \cdot 3! \cdot 1\{y_1 < y_2 < y_3\}$$

$$= 48 y_1 y_2 y_3\, 1\{0 < y_1 < y_2 < y_3 < 1\}$$

Now, if we want the distribution of, say, the median, we can integrate to get the marginal.

$$f_{X_{2:3}}(y_2) = \int_{y_2}^{\infty} \int_{-\infty}^{y_2} f(y_1) f(y_2) f(y_3) \cdot 3!\, dy_1\, dy_3$$

$$= \int_{y_2}^{\infty} 3!\, F(y_2)\, f(y_2) f(y_3)\, dy_3$$

$$= 3!\, F(y_2) f(y_2) \left[1 - F(y_2)\right]$$

This result can be generalized as follows:

Theorem (6.5.2) If $X_1, \ldots, X_n$ is a random sample from a continuous distribution, then

$$f_{X_{k:n}}(y) = n \binom{n-1}{k-1} F(y)^{k-1} \left[1 - F(y)\right]^{n-k} f(y).$$

Theorem (6.5.3) For any (not just continuous) distribution, the cdf of the kth order statistic from a random sample is

$$F_{X_{k:n}}(y) = \sum_{i=k}^{n} \binom{n}{i} F(y)^i \left(1 - F(y)\right)^{n-i}.$$

Example: Let $X_1, \ldots, X_n \sim$ iid $EXP(1)$.
Find the cdf and the density of $X_{1:n}$.

$$\begin{aligned} F_{X_{1:n}}(y) = \mathbb{P}(X_{1:n} \le y) &= 1 - \mathbb{P}(X_{1:n} > y) \\ &= 1 - \prod_{i=1}^{n} \mathbb{P}(X_i > y) \\ &= 1 - \left[1 - F(y)\right]^n \\ &= \begin{cases} 1 - e^{-yn} & y > 0 \\ 0 & y \le 0 \end{cases} \end{aligned}$$

$$f_{X_{1:n}}(y) = \begin{cases} ne^{-yn} & y > 0, \\ 0 & y \le 0. \end{cases}$$

Note that theorem 6.5.3 would have been unpleasant, but the cdf technique is easy to use directly.

Note that our result matches theorem 6.5.2

Example: Repeat for $X_{n:n}$ where $f(x) = 2x\,\mathbb{1}\{x \in (0,1)\}$

Chapter 7

A sequence of random variables X_n "converges in distribution" to Y if $\lim_{n\to\infty} F_{X_n}(x) = F_Y(x)$ for all x where F_Y is continuous.

Example: (7.2.1) Let $X_1, \ldots, X_n \sim$ iid UNIF$(0,1)$. Find $X_{n:n}$ and show that it converges to 1.

$$F_{X_{n:n}}(y) = \mathbb{P}(X_{n:n} \le y) = \left[\mathbb{P}(X_1 \le y)\right]^n = \begin{cases} y^n & y \in (0,1), \\ 1 & y \ge 1, \\ 0 & y \le 0. \end{cases}$$

$$\xrightarrow[\text{as } n\to\infty]{} \begin{cases} 1 & y \ge 1, \\ 0 & y < 1. \end{cases}$$

Example: (7.2.2) The cdf of $X_{1:n}$ from an EXP(1) population is

$$F_{X_{1:n}}(y) = \begin{cases} 1 - e^{-yn} & y > 0, \\ 0 & y \le 0. \end{cases}$$

$$\longrightarrow \begin{cases} 1 & y > 0 \\ 0 & y \le 0 \end{cases}$$

Note that the limit is <u>not</u> a cdf.
(It's not right-continuous)

Does this mean that $X_{1:n}$ does not converge or that it doesn't converge to a r.v.?

No. It converges to 0. Note in the definition of convergence that $\lim_{n\to\infty} F_{X_n}(x)$ need not equal $F_X(x)$ at points of discontinuity.

Example (7.2.3) Let $X_1, \ldots, X_n$ be a random sample from PAR(1,1) so that $F(x) = \begin{cases} 1 - \frac{1}{1+x} & x > 0, \\ 0 & o/w. \end{cases}$

Then $F_{nX_{1:n}}(y) = \begin{cases} 1 - \left(1 + \frac{y}{n}\right)^{-n} & y > 0, \\ 0 & o/w. \end{cases} \to \begin{cases} 1 - e^{-y} & y > 0 \\ 0 & o/w \end{cases}$

Example (7.2.4) Consider $X_{n:n}$ from PAR(1,1). $f(x) = \frac{1\{x>0\}}{(1+x)^2}$.

$F(x) = \begin{cases} 1 - \frac{1}{1+x} & x > 0, \\ 0 & x \le 0. \end{cases}$

$F_{X_{n:n}}(y) = \begin{cases} \left(\frac{y}{1+y}\right)^n & y > 0, \\ 0 & o/w. \end{cases} \to 0.$

Thus $X_{n:n}$ does <u>not</u> converge to a r.v. because $\lim_{y\to\infty} F_{X_{n:n}}(y) = 0 \neq 1$.

Note that the cdf converging pointwise to zero is quite different that the random variable converging to zero in distribution.

Example 7.25:

Consider $\frac{1}{n} X_{n:n}$ from the previous example

$$F_{\frac{1}{n}X_{n:n}}(y) = \mathbb{P}\left(\tfrac{1}{n} X_{n:n} \le y\right) = F_{X_{n:n}}(ny) = \begin{cases} \left(\frac{ny}{1+ny}\right)^n & y > 0, \\ 0 & \text{o/w.} \end{cases}$$

$$\longrightarrow \begin{cases} e^{-1/y} & y > 0, \\ 0 & \text{o/w.} \end{cases}$$

Example 7.2.7

Find the limiting distribution of $\bar{X}_n$, where $X_i \sim \text{iid } N(\mu, \sigma^2)$. Recall that $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

$$F_{\bar{X}_n}(y) = \mathbb{P}(\bar{X}_n \le y) = \mathbb{P}\left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{y-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi\left(\frac{y-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$\longrightarrow \begin{cases} 0 & y < \mu, \\ \frac{1}{2} & y = \mu, \\ 1 & y > \mu. \end{cases}$$

Thus $\bar{X}_n$ converges in distribution to μ.

Theorem (7.3.1) If moment generating functions $M_n(t) \to M(t)$ in a neighborhood of 0, then the associated r.v.'s converge in distribution: $Y_n \xrightarrow{d} Y$.

Example (7.3.1) Let $X_i \overset{iid}{\sim} BER(p)$.
Let $Y_n = \sum_{i=1}^{n} X_i \sim BIN(n,p)$

Let $p = p_n = \frac{\mu}{n}$.

Then, $M_{Y_n}(t) = \left[\frac{\mu}{n} e^t + \left(1 - \frac{\mu}{n}\right)\right]^n$

$$= \left[1 + \left(e^t - 1\right)\frac{\mu}{n}\right]^n$$

$$\to e^{\mu(e^t - 1)}$$

Thus $Y_n \xrightarrow{d} POI(\mu)$.

Example 7.3.2 (Bernouli LLN)

Let $X_i \sim BER(p)$

$$M_{\bar{X}_n}(t) = \left(p e^{t/n} + 1 - p\right)^n \to e^{pt}$$

Note that e^{pt} is the MGF of p.

Thus $\bar{X}_n \xrightarrow{d} p$

It can be shown (example 7.3.3) that if $Y_n = \dfrac{\bar{X}_n - p}{\sqrt{\frac{1}{n} p(1-p)}}$ then $M_{Y_n}(t) \to e^{t^2/2}$

Thus $Y_n \xrightarrow{d} N(0,1)$.

Theorem (7.3.2) (The Central Limit Theorem)

If $X_i \sim iid(\mu, \sigma^2)$ with $\mu, \sigma^2 < \infty$, then

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$

Proof: For this proof, we assume that the MGF of the sequence exists.

Let $m(t) = M_{X-\mu}(t)$.

A Taylor series for $m(t)$ gives

$$m(t) = m(0) + m'(0)t + \frac{m''(\xi(t))t^2}{2} \quad \text{for some } 0 \le \xi(t) \le t.$$

$$= m(0) + \frac{m''(\xi(t))\,t^2}{2}$$

$$= m(0) + \frac{\sigma^2 t^2}{2} + \left(\frac{m''(\xi(t)) - \sigma^2}{2}\right) t^2$$

Now, let $Z_n = \dfrac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$.

Then $M_{Z_n}(t) = M_{\frac{\Sigma(X_i-\mu)}{\sigma\sqrt{n}}}(t) = M_{\Sigma(X_i-\mu)}\left(\frac{t}{\sigma\sqrt{n}}\right)$

$$= \prod_{i=1}^{n} M_{X_1-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) = \left[M_{X_1-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

$$= \left[m\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = \left[m(0) + \frac{\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2}{2} + \frac{\left[m''\left(\xi\left(\frac{t}{\sigma\sqrt{n}}\right)\right) - \sigma^2\right]\left(\frac{t}{\sigma\sqrt{n}}\right)^2}{2}\right]^n$$

$$= \left[m(0) + \frac{t^2}{2n} + \frac{d(n)}{n}\right]^n \qquad \text{where } d(n) \to 0$$

$$= \left[1 + \frac{t^2}{2n} + \frac{d(n)}{n}\right]^n$$

$$\longrightarrow e^{t^2/2}, \text{ the MGF for } N(0,1). \qquad \square$$

Example (7.3.4) Let $X_i \sim$ iid UNIF$(0,1)$

Then $\dfrac{\bar{X}_{12} - \frac{1}{2}}{\frac{1}{12}} \overset{\text{approx}}{\sim} N(0,1)$

The text suggests that this approximation is so close that some computer algorithms may use this to generate normal data.

Why not use $F^{-1}(X_1)$ instead? (see theorem 6.3.4)

Other downsides?

Note that $\dfrac{\bar{X}_{12} - \frac{1}{2}}{\frac{1}{12}} \leq 6$ and ≥ -6.

Is that a problem?

Example (7.4.1) Suppose your friend claims she makes 50% of her free throws.

Suppose she will shoots 20 times so that $Y \sim BIN(20, \frac{1}{2})$ represents the number of shots she will make.

Calculate that $\mathbb{P}(Y \geq 9) = 0.7483$

Estimate that $\mathbb{P}(Y \geq 9) \approx$ ~~0.8133~~ 0.6726

We can improve the estimate with a continuity correction:

$$\mathbb{P}(Y \geq 9) = \mathbb{P}(Y > 8.5) = \mathbb{P}\left(\frac{Y-10}{\sqrt{\frac{20}{4}}} > \frac{-1.5}{\sqrt{\frac{20}{4}}}\right)$$

$$\approx 1 - \Phi(-.67)$$

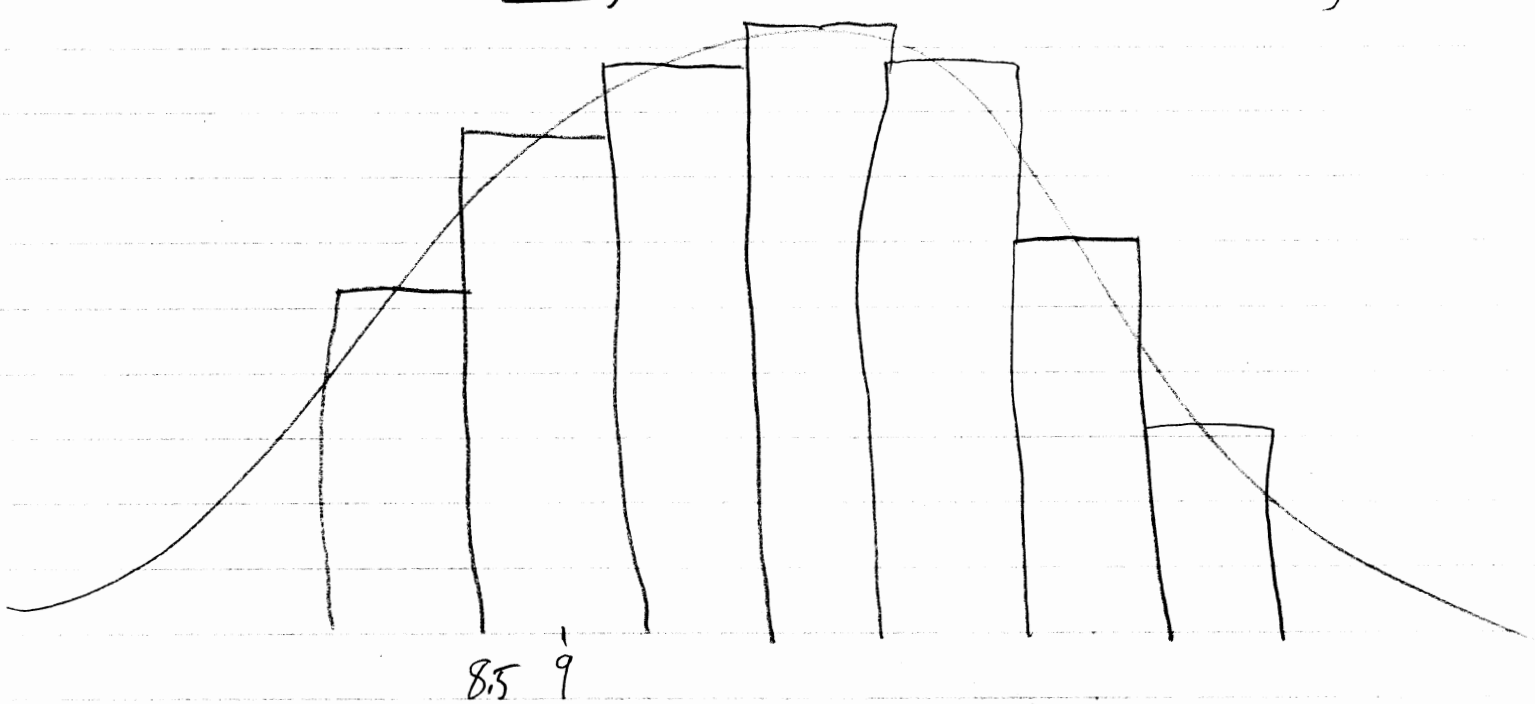
$$= 0.7486,$$

which is correct to 3 decimal places!

This is quite an improvement.

Should we always use a continuity correction?

This is a <u>rough</u> visualization of continuity correction:



Theorem (7.5.1) If $X_1, \ldots, X_n$ is a random sample from a continuous distribution, $f(x_p) \neq 0$, where x_p is the pth percentile, and $\frac{k_n}{n} \to p$ (with $k_n - np$ bounded)

then $X_{k_n:n}$ is asymptotically normal with mean x_p and variance c^2/n.

In other words, $$\frac{X_{k_n:n} - x_p}{\frac{c}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$

Note that c^2 is defined as

$$c^2 = \frac{p(1-p)}{\left[f(x_p)\right]^2}$$

Why can't we say that $X_{k_n:n} \xrightarrow{d} N\left(x_p, \frac{c^2}{n}\right)$?

Why can't $f(x_p) = 0$?

Example: Let $X_i \sim$ iid $EXP(1)$.
(7.5.2) Let $k_n = \begin{cases} n/2 & \text{if } n \text{ is even} \\ \frac{n+1}{2} & \text{if } n \text{ is odd} \end{cases}$

Then $X_{k_n:n}$ is the sample median.

$\frac{k_n}{n} \to \frac{1}{2}$ and $f(x_{.5}) = f(\log(2)) = \frac{1}{2}$ since

$$\frac{1}{2} = \int_0^m e^{-x}\,dx = -e^{-x}\Big|_0^m = 1 - e^{-m}$$

$$\Rightarrow e^{-m} = \frac{1}{2}$$

$$\Rightarrow -m = -\log(2)$$

$$\Rightarrow m = \log(2)$$

We calculate $c^2 = \dfrac{\frac{1}{2}(1-\frac{1}{2})}{(\frac{1}{2})^2} = 1$

$$\Rightarrow c = 1$$

Thus $X_{k_n:n}$ is asymptotically normal with mean $\log(2)$ and variance $\frac{1}{n}$.

That is, $\dfrac{X_{k_n:n} - \log(2)}{\frac{1}{\sqrt{n}}} \xrightarrow{d} N(0,1)$

Definition: X_n converges in probability to X if $\mathbb{P}(|X_n - X| > \varepsilon) \to 0 \quad \forall\ \varepsilon > 0$

Theorem: Convergence in probability implies convergence in distribution, i.e.

$$X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X.$$

Theorem: If X is a constant, c, then

$$X_n \xrightarrow{d} c \implies X_n \xrightarrow{\mathbb{P}} c$$

Note that together these theorems show that $X_n \xrightarrow{d} c$ iff $X_n \xrightarrow{\mathbb{P}} c$ for any <u>constant</u> c.

Theorem: (7.6.2) Law of Large Numbers

If $X_i \sim iid\ (\mu, \sigma^2)$ with μ, σ^2 finite, then $\bar{X}_n \xrightarrow{d} \mu$

proof: Recall the Chebyshev inequality:

$$\mathbb{P}(|X - E(X)| > k\sigma) \leq \frac{1}{k^2} \quad \text{for } k > 0, \sigma^2 < \infty.$$

Let $K_n = \frac{\varepsilon\sqrt{n}}{\sigma}$

Then $\mathbb{P}\left(|\bar{X}_n - \mu| > \varepsilon\right) = \mathbb{P}\left(|\bar{X}_n - \mu| > \frac{\sigma}{\sqrt{n}} \frac{\varepsilon\sqrt{n}}{\sigma}\right)$

$\leq \frac{\sigma^2}{\varepsilon^2 n^2} \to 0$

Thus $\bar{X}_n \xrightarrow{\mathbb{P}} \mu \Rightarrow \bar{X}_n \xrightarrow{d} \mu$. $\square$

Theorem (7.6.3) If $\frac{(X_n - m)}{\frac{c}{\sqrt{n}}} \xrightarrow{d} N(0,1)$,

then $X_n \xrightarrow{\mathbb{P}} m$.

Provide some intuition about why this is true.

Theorems 7.5.1 and 7.6.3 together show that $X_{K_n:n} \xrightarrow{\mathbb{P}} x_p$ if $\frac{K_n}{n} \to p$ and the assumptions of the theorems hold.

Theorem (7.7.2) If $X_n \xrightarrow{\mathbb{P}} c$ and g is continuous at c, then $g(X_n) \xrightarrow{\mathbb{P}} g(c)$.

proof: Let $\varepsilon > 0$. Choose $\delta > 0$ s.t. $|x - c| < \delta \Rightarrow |g(x) - g(c)| < \varepsilon$.
Then $\mathbb{P}\left(|g(X_n) - g(c)| > \varepsilon\right) \leq \mathbb{P}\left(|X_n - c| > \delta\right) \to 0$.

$\square$

Note that the above proof works for multivariate functions.

Theorem (7.7.3) Assume $X_n \xrightarrow{\mathbb{P}} c$ and $Y_n \xrightarrow{\mathbb{P}} d$

Then 1) $aX_n + bY_n \xrightarrow{\mathbb{P}} ac + bd$

2) $X_n Y_n \xrightarrow{\mathbb{P}} cd$

3) $\frac{X_n}{c} \xrightarrow{\mathbb{P}} 1$

4) $\frac{1}{X_n} \to \frac{1}{c}$ if $c \neq 0$

5) $\sqrt{X_n} \to \sqrt{c}$ if $X_n \geq 0$.

Proof: A trivial application of theorem 7.7.2

Theorem (7.7.4) Slutsky's Theorem

Assume $X_n \xrightarrow{\mathbb{P}} c$ and $Y_n \xrightarrow{d} Y$

Then 1) $X_n + Y_n \xrightarrow{d} c + Y$

2) $X_n Y_n \xrightarrow{d} cY$

3) $Y_n / X_n \xrightarrow{d} Y/c$ $c \neq 0$

Example (7.7.2) Let $X_i \sim$ iid BER(p)

Let $\hat{p}_n = \overline{X}_n$.

By the CLT, $\dfrac{\hat{p}_n - p}{\sqrt{\dfrac{p(1-p)}{n}}} \xrightarrow{d} N(0,1)$

Note $\hat{p}_n \xrightarrow{\mathbb{P}} p$ (LLN)

$\Rightarrow \hat{p}_n(1-\hat{p}_n) \xrightarrow{\mathbb{P}} p(1-p)$ (theorem 7.7.2)

$\dfrac{\sqrt{p(1-p)}}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{\mathbb{P}} 1$ (theorem 7.7.3 part 3)

Then by Slutsky,

$$\frac{\hat{p}_n - p}{\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n)}{n}}} \xrightarrow{d} N(0,1)$$

Theorem (7.7.6) If $\dfrac{Y_n - m}{\frac{c}{\sqrt{n}}} \xrightarrow{d} N(0,1)$

and if $g'(c)$ exists & is non-zero at m,

then $$\frac{g(Y_n) - g(m)}{\left(\frac{c\, g'(m)}{\sqrt{n}}\right)} \xrightarrow{d} N(0,1).$$

Proof: Let $u(y) = \begin{cases} \dfrac{g(y) - g(m)}{y - m} - g'(m) & y \neq m, \\ 0 & y = m. \end{cases}$

Note that u is continuous at m with $u(m) = 0$

Now,

$$\frac{g(Y_n) - g(m)}{\left(\frac{c\, g'(m)}{\sqrt{n}}\right)} = \left(\frac{Y_n - m}{\frac{c}{\sqrt{n}}}\right) \frac{g'(m) + u(Y_n)}{g'(m)}$$

By theorem 7.7.2, $\dfrac{g'(m) + u(Y_n)}{g'(m)} \xrightarrow{d} 1.$

Then By Slutsky's theorem, the result is established. $\square$

The text uses $|c\,g'(m)|$ in the denominator of the above theorem.

Are absolute values necessary here?

Chapter 8: Statistics and Sampling Distributions

This chapter defines a 'statistic' and demonstrates some properties of the sample mean, $\bar{X}$, and the sample variance, S^2.

Def: A 'statistic' is a funtion of a sample.

Thus it is a random variable whose outcome is completely determined by the outcome of the random sample.

Example: The sample mean $\bar{X}_n$ is defined as

$$\bar{X}_n = t(X_1, X_2, \ldots, X_n) \quad \text{where}$$

$$t(x_1, x_2, \ldots, x_n) = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}.$$

Example: The sample variance S^2 is

$$S^2 = t(X_1, X_2, \ldots, X_n) \quad \text{where}$$

$$t(x_1, x_2, \ldots, x_n) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n-1}$$

Theorem (8.2.1) If $X_i \sim iid\,(\mu, \sigma^2)$, then

$$E(\bar{X}_n) = \mu \quad \text{and} \quad var(\bar{X}_n) = \frac{\sigma^2}{n}$$

Prove this theorem!

Def: A statistic, T, is "unbiased" for a parameter, θ, if $E(T) = \theta$

Note that the above theorem states that $\bar{X}$ is an unbiased estimator for μ.

Thus, on average, outcomes $\bar{x}$ of $\bar{X}$ will be μ, but any particular $\bar{x}$ may be quite different from μ.

However, the variance shinks an n increases, suggesting a smaller and smaller chance that $\bar{x}$ of obtaining an $\bar{x}$ that is far from μ.

Example (8.2.2) Let $X_1, \ldots, X_n \sim$ iid $BER(p)$

Let $\hat{P} = \bar{X}$ be the sample proportion.

Then $E(\hat{P}) = p$ and $var(\hat{P}) = \frac{p(1-p)}{n}$

Note that $\hat{P}$ is an unbiased estimator for p.

Theorem (8.2.2) If $X_i \sim$ iid (μ, σ^2), then

$$E(S^2) = \sigma^2$$

$$var(S^2) = \frac{\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4}{n} \qquad n > 1$$

where $\mu_4 = E(X-\mu)^4$ is the 4th central moment of X_1.

Proof that $E(S^2) = \sigma^2$:

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

$$= \frac{1}{n-1} E\left[\left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2\right] = \frac{1}{n-1}\left[\sum E(X_1^2) - n E(\bar{X}^2)\right]$$

$$= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - \frac{1}{n}\sum_{i,k} E(X_i X_k)\right] = \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - \frac{1}{n}\left(n(\sigma^2+\mu^2) + (n^2-n)\mu^2\right)\right]$$

$$= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - (\sigma^2+\mu^2+(n-1)\mu^2)\right]$$

$$= \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2. \qquad \square$$

This space intentionally left blank...

... okay, it was a mistake.

The distribution of a statistic is called a "Sampling Distribution".

Example: Find the sampling distribution of $\bar{X}_n$, where $X_i \sim$ iid $N(\mu, \sigma^2)$

$$M_{\bar{X}}(t) = M_{\sum_{i=1}^{n} X_i}\left(\frac{t}{n}\right) = \prod_{i=1}^{n} M\left(\frac{t}{n}\right) = \left(e^{\mu \frac{t}{n} + \sigma^2 \left(\frac{t}{n}\right)^2 \frac{1}{2}}\right)^n$$

$$= e^{\mu t + \left(\frac{\sigma}{\sqrt{n}}\right)^2 t^2 \frac{1}{2}}$$

Thus $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Example: $X_i \sim$ independent $N(\mu_i, \sigma_i^2)$
(8.3.1)

Find the distribution of $Y = \sum_{i=1}^{n} X_i$

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n}\left(e^{\mu_i t + \sigma_i^2 t^2 \frac{1}{2}}\right)$$

$$= e^{t\left(\sum_{i=1}^{n} \mu_i\right) + \frac{t^2}{2} \sum \sigma_i^2}$$

Thus $Y \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$.

Example (8.3.1) It is claimed that the lifetime of a certain brand of battery is $N(60, 36)$.

Let $X_1, X_2, \ldots, X_{25}$ be a random sample from these batteries.

$\bar{X} \sim N\left(60, \frac{36}{25}\right)$ if the claim is true

If the claim is true, then

$$\mathbb{P}(\bar{X} > c) = \mathbb{P}\left(\frac{\bar{X}-60}{\frac{6}{5}} > \frac{c-60}{\frac{6}{5}}\right) = 1 - \Phi\left(\frac{c-60}{\frac{6}{5}}\right)$$

What value of c would make this .95?

$$\frac{c-60}{\frac{6}{5}} = z_{.05} = -1.645$$

$$\Rightarrow c = -1.645\left(\frac{6}{5}\right) + 60 = 58.026$$

What would you conclude if $\bar{x} = 56$?

$$\mathbb{P}(\bar{X} > 56) = 1 - \Phi\left(\frac{-4}{\left(\frac{6}{5}\right)}\right) = 1 - \Phi\left(\frac{-20}{6}\right) = 1 - \Phi(-3.33)$$

$$= 1 - .0004 = .9996$$

A special case of the Gamma distribution is

$GAM(2, \frac{v}{2})$ i.e. $\chi^2(v)$.

Theorem (8.3.2) Let $Y \sim \chi^2(v)$.

Then $M_Y(t) = \left(\frac{1}{1-2t}\right)^{v/2} = (1-2t)^{-v/2}$

$E(Y) = v$

$E(Y^2) = v(v+2)$

$var(Y) = v(v+2) - v^2 = 2v$

Theorem (8.3.3) If $X \sim GAM(\theta, k)$, then $\frac{2X}{\theta} \sim \chi^2(2k)$

Proof: $M_{\frac{2X}{\theta}}(t) = M_X\left(\frac{2t}{\theta}\right) = \left(\frac{1}{1-\theta\left(\frac{2t}{\theta}\right)}\right)^k = \left(\frac{1}{1-2t}\right)^k$

Thus $\frac{2X}{\theta} \sim GAM(2, k) \sim \chi^2(2k)$.

Example (8.3.3) Suppose you have a nice table that tells you that $\chi^2_{.10}(4) = 1.06$ and that a solar panels lifetime is $GAM(3,2)$.

For advertising, we'd like to say that 90% of solar panels will last c years or longer. Find c.

$$.10 = P(X \le c) = P\left(\frac{2X}{3} \le \frac{2c}{3}\right)$$

Since $\frac{2X}{3} \sim \chi^2(4)$, $\frac{2c}{3} = \chi^2_{.10}(4) = 1.06$

Thus, $c = \frac{3}{2}(1.06) = 1.59$ years.

Theorem: (8.3.4) Suppose $X_i \sim$ indep. $\chi^2(\nu_i)$

Then $\sum_{i=1}^n X_i \sim \chi^2\left(\sum_{i=1}^n \nu_i\right)$

proof: $M_{\Sigma X_i}(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n \left(\frac{1}{1-2t}\right)^{\nu_i/2} = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}\sum_{i=1}^n \nu_i}$. $\square$

Theorem: (8.3.5) If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$

proof: $M_{Z^2}(t) = E\left(e^{tZ^2}\right) = \int_{\mathbb{R}} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int e^{(t-\frac{1}{2})x^2}\, dx$

Let $\left(t-\frac{1}{2}\right)x^2 = \frac{-x^2}{2\sigma^2}$

Then $2\sigma^2 = \frac{-1}{\left(t-\frac{1}{2}\right)} \Rightarrow \sigma^2 = \frac{+1}{1-2t}$

Thus the integral becomes:

$$\frac{1}{\sqrt{2\pi}}\int e^{-\frac{x^2}{2\sigma^2}}\,dx = \sigma\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}}\,dx = \sigma = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}$$

Thus $Z^2 \sim \chi^2(1)$.

Corollary (8.3.2) If $X_i \sim iid\ N(\mu, \sigma^2)$,

then $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$

and $\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \sim \chi^2(1)$

Can the above corollary be used to find the distribution of $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$?

Dependence of terms is a problem.

Theorem: (8.3.6) If $X_1, \ldots, X_n \sim iid\ N(\mu, \sigma^2)$ then

1: $\bar{X}$ is independent of $\{X_i - \bar{X} \mid i = 1, 2, \ldots, n\}$

2: $\bar{X}$ is independent of S^2

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

Note that $\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left((x_i-\bar{x})+(\bar{x}-\mu)\right)^2}{\sigma^2}$

$$= \sum_{i=1}^{n} \left[\frac{(x_i-\bar{x})^2 + (\bar{x}-\mu)^2 + 2(x_i\bar{x} - \mu x_i - \bar{x}^2 + \mu\bar{x})}{\sigma^2} \right]$$

$$= \left(\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\sigma^2} \right) + \frac{n(\bar{x}-\mu)^2}{\sigma^2} + 0$$

Thus the joint density of $X_1, \ldots, X_n$ is

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \cdot \exp\left[\frac{-1}{2\sigma^2} \sum (x_i-\mu)^2 \right]$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \cdot \exp\left[\frac{-1}{2\sigma^2} \left(\sum_{i=1}^{n} (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 \right) \right]$$

Now consider the joint transformation:

$$Y_1 = \bar{X}, \quad Y_i = X_i - \bar{X} \quad \text{for } i = 2, 3, \ldots, n.$$

Then, $x_1 - \bar{x} = -\sum_{i=2}^{n} (x_i - \bar{x}) = -\sum_{i=2}^{n} y_i$

so, $\sum_{i=1}^{n} (x_i-\bar{x})^2 = \left(-\sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n} (y_i)^2$

Now $f_{\underline{Y}}(y_1, \ldots, y_n) = \dfrac{|J|}{(2\pi)^{n/2}\sigma^n} \cdot \exp\left[\dfrac{-1}{2\sigma^2}\left(\left(\sum_{i=2}^{n} y_i\right)^2 + \left(\sum_{i=2}^{n} y_i^2\right) + n(y_1-\mu)^2\right)\right]$.

Also, $x_1 = y_1 - \sum_{i=2}^{n} y_i$ and $x_i = y_i + y_1$ for $i = 2, 3, \ldots, n$,

Thus J is constant and $f_{\underline{Y}}$ can be written as a product of a function of only y_1 and a function of only $(y_2, y_3, \ldots, y_n)$.

Thus Y_2 is independent of Y_1

(i.e. $X_2 - \bar{X}$ is independent of $\bar{X}$)

Why does this mean that $\bar{X}$ is independent of $X_i - \bar{X}$ for $i = 1, 2, 3, \ldots, n$ (even $i = 1$)?

1 is now proven.

To prove 2, note that S^2 is a function only of $\{X_i - \bar{X} \mid i = 1, 2, \ldots, n\}$ which are independent of $\bar{X}$.

For 3, Recall that $\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \dfrac{n(\bar{X}-\mu)^2}{\sigma^2} + \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$

$$V_1 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} - \frac{n(\bar{X}-\mu)^2}{\sigma^2} = V_2 - V_3$$

Note that V_1 is independent of V_3 so that

$$M_{V_2}(t) = M_{V_1+V_3}(t) = M_{V_1}(t)\, M_{V_3}(t)$$

$$\Rightarrow M_{V_1}(t) = \frac{M_{V_2}(t)}{M_{V_3}(t)} = \frac{\left(\frac{1}{1-2t}\right)^{n/2}}{\left(\frac{1}{1-2t}\right)^{1/2}} = \left(\frac{1}{1-2t}\right)^{\frac{n-1}{2}}$$

since $V_2 \sim \chi^2(n)$ and $V_3 \sim \chi^2(1)$.

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

Example: Suppose it is claimed that a certain brand of battery has a lifetime that is $N(60, 36)$.

Suppose we will assume that this is true unless S^2 from our sample of 25 is < 54.63.

If the claim is true, what is the probability that we will reject it erroneously?

$$\mathbb{P}(S^2 < 54.63) = \mathbb{P}\left(\frac{(24)S^2}{36} < \frac{24}{36}(54.63)\right)$$

$$= \mathbb{P}\left(\chi^2(24) < \frac{24}{36}(54.63)\right) = 0.05$$

if claim is true

This should be 0.95. This problem makes more sense if the inequalities are flipped.

Recall that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ when $X_1, \ldots, X_n \sim^{iid} N(\mu, \sigma^2)$.

We can use $\bar{X}$ as an estimator for μ and to make inferences about μ, but not knowing σ means we don't know the distribution of $\bar{X}$. This could prevent us from making inferences about μ.

We will see that $\dfrac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ has a distribution that is not influenced by σ.

Definition: If $Z \sim N(0,1)$ and $V \sim \chi^2(v)$ are independent,

then $$\frac{Z}{\sqrt{\frac{V}{v}}} \sim t(v)$$

where $t(v)$ represents a "t-distribution" with n degrees of freedom

Describe the shape of the density function for small and large values of v.

Theorem: (8.4.2) $\operatorname{var}(T) = \frac{v}{v-2}$ where $T \sim t(v)$ and $v > 2$.

proof: $E(T) = 0$ by symmetry.

$$E(T^2) = E\left(\frac{Z^2}{\left(\frac{V}{v}\right)}\right) = E(Z^2)\,E\left(\frac{v}{V}\right) = E\left(\frac{v}{V}\right)$$

$$= v\,E\left(\frac{1}{V}\right) = v\int \frac{1}{x}\,\frac{1}{2^{v/2}\Gamma\left(\frac{v}{2}\right)}\,x^{v/2-1}\,e^{-x/2}\,dx$$

$$= v\left(\frac{1}{2}\right)\frac{\Gamma\left(\frac{v-2}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}\int \frac{1}{2^{\frac{v-2}{2}}\Gamma\left(\frac{v-2}{2}\right)} x^{\frac{v-2}{2}-1} e^{-x/2}\, dx$$

$$= v\left(\frac{1}{2}\right)\frac{\Gamma\left(\frac{v-2}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}$$

$$= v\left(\frac{1}{2}\right)\frac{\Gamma\left(\frac{v-2}{2}\right)}{\left(\frac{v-2}{2}\right)\Gamma\left(\frac{v-2}{2}\right)} = \frac{v}{v-2}$$

since $\Gamma(x+1) = x\Gamma(x)$.
(i.e. $\Gamma(x) = (x-1)\Gamma(x-1)$)

Note that for positive integers n, $\Gamma(n) = (n-1)!$

Theorem (8.4.3) $\quad \dfrac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$ $\quad$ if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

Proof: FROM theorem 8.3.6, $\bar{X}$, S are independent.

$$t(n-1) \sim \frac{Z}{\sqrt{\frac{V}{v}}} = \frac{\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)}{\sqrt{\frac{\left(\frac{(n-1)S^2}{\sigma^2}\right)}{(n-1)}}} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}. \quad \square$$

Def: If $V_1 \sim \chi^2(v_1)$ and $V_2 \sim \chi^2(v_2)$ are independent, then

$$\frac{\left(\frac{V_1}{v_1}\right)}{\left(\frac{V_2}{v_2}\right)} \sim F(v_1, v_2)$$

where $F(v_1, v_2)$ represents an "F-distribution" with v_1 "numerator" and v_2 "denominator" degrees of freedom.

Describe how the density function looks for various values of the parameters.

Theorem (8.4.5) $E(X) = \frac{v_2}{v_2 - 2}$ where $X \sim F(v_1, v_2)$

$$\text{var}(X) = \frac{2 v_2^2 (v_1 + v_2 - 2)}{v_1 (v_2 - 2)^2 (v_2 - 4)}$$

Proof of part 1: $E(X) = \frac{v_2}{v_1} E(V_1) E\left(\frac{1}{V_2}\right) = v_2 E\left(\frac{1}{V_2}\right) = \frac{v_2}{v_2 - 2}$

Note that we use the calculation of $E\left(\frac{1}{V_2}\right)$ from the proof of Theorem 8.4.2.

Example: Let $\star$ be the 5th percentile of $F(v_1, v_2)$.

Find the 95% of $F(v_2, v_1)$ in terms of $\star$.

Note that if $X \sim F(v_1, v_2)$, then $\frac{1}{X} \sim F(v_2, v_1)$.

$$.05 = \mathbb{P}(X \leq \star) = \mathbb{P}\left(\frac{1}{X} \geq \frac{1}{\star}\right) = 1 - \mathbb{P}\left(\frac{1}{X} < \frac{1}{\star}\right)$$

$$\Rightarrow .95 = \mathbb{P}\left(\frac{1}{X} < \frac{1}{\star}\right).$$

Thus $\frac{1}{\star}$ is the 95th percentile of $F(v_2, v_1)$.

Example: Develop the idea of a confidence interval. (Write out the general steps)

$X_1, \ldots, X_{16} \sim$ iid $N(\mu_1, \sigma_1^2)$

$Y_1, \ldots, Y_{21} \sim$ iid $N(\mu_2, \sigma_2^2)$

Suppose we want to estimate $\frac{\sigma_1^2}{\sigma_2^2}$.

$$\frac{(16-1)S_1^2}{\sigma_1^2} \sim \chi^2(16-1)$$

$$\frac{(21-1)S_2^2}{\sigma_2^2} \sim \chi^2(21-1)$$

$$\Rightarrow \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(15,20)$$

$$\Rightarrow .90 = \mathbb{P}\left(f_{.05}(15,20) < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < f_{.95}(15,20)\right)$$

$$= \mathbb{P}\left(.43 < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < 2.20\right)$$

$$= \mathbb{P}\left(.43\frac{S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < 2.20\frac{S_2^2}{S_1^2}\right)$$

Suppose you do the experiment and calculate $s_1^2 = 2$, $s_2^2 = 4$ from the outcome.

What can we say about $\left(.43\left(\frac{4}{2}\right), 2.20\left(\frac{4}{2}\right)\right)$ $= (.86, 4.40)$?

The above interval is a 90% confidence interval for $\frac{\sigma_2^2}{\sigma_1^2}$.

Is σ_1^2 different than σ_2^2?

First answer this from the CI, then develop the idea of a hypothesis test.

Steps for creating a CI

1) Pick an expression involving a statistic and the unknown parameter of interest.

2) Write a probability equation using the above expression.

3) Solve inside $\mathbb{P}$ for the parameter of interest.

4) Write down the random interval

5) Plug in the outcome to obtain the CI.

Steps for doing a hypothesis test:

1) choose a "null hypothesis", H_0.

2) choose a <u>statistic</u> whose distribution is known under H_0 such that the distribution is quite different if H_0 is not true.

3) Choose you type I error rate (this is called the alpha-level or the size of the test)

4) Determine the critical region

5) Do the experiment and plug the outcome into the statistic...

6) Reject H_0 if the outcome of the statistic falls in the critical region.

Chapter 9 — Point Estimation

Occassionally a few axioms or assumptions will allow us to completely determine the distribution of a random variable.

Ex: Toss a coin 5 times and let X be the number of heads. If the coin is fair, $X \sim BIN(5, \frac{1}{2})$

However we often cannot determine the exact distribution but we might know the family of distributions that it comes from.

For example, we might know that a population is normally distributed but not know the parameters, μ and σ^2.

Other times, we may not know anything about the distribution of a population, but we can still estimate parameters such as the mean and median.

In this chapter we assume the distribution family is known.

Let $X_1, \ldots, X_n$ be a random sample from a population with density function belonging to the family $\{f(x;\theta) \mid \theta \in \Omega\}$.

Note that Ω is the set of possible values for the parameter and that θ can be vector-valued.

For example: $\left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \;\middle|\; \mu \in \mathbb{R},\ \sigma^2 > 0 \right\}$

Ω is called the "parameter space."

Definition: A <u>statistic</u> used to estimate the value of the parameter, θ, or a function thereof, $\tau(\theta)$, is called an "estimator" of θ or $\tau(\theta)$ respectively.

The outcome of the estimator is called an estimate. Thus an estimator is a random variable and its outcome is an estimate.

We will often use $\hat{\theta}$ to represent an estimator or an estimate of the parameter, θ, e.g. $\hat{\mu}$ and $\hat{\sigma}^2$ are estimators / estimates of μ and σ^2 respectively.

The first method for creating an estimator is called the "method of moments."

A method of moments estimator (MME) is derived by setting the population moments equal to sample moments and then solving for the parameter.

Example (9.2.1). If we are interested in the population mean, μ, we set

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(no solving required)}$$

and toss a hat on the parameter to denote that it is an estimator:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

If there are more parameters in need of estimating, we move to the second moment, and so on.

Thus $\hat{\sigma}^2 + \hat{\mu}^2 = \dfrac{\sum_{i=1}^{n} X_i^2}{n}$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{\sum X_i^2}{n} - \hat{\mu}^2$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{\sum X_i^2}{n} - \bar{X}^2 \quad \text{(using the MME for } \mu\text{)}$$

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

Notice that we haven't specified the distribution family because the above MMEs are valid for any distribution.

Example (9.2.2)

$$f(x;\theta,\eta) = \begin{cases} \frac{1}{\theta} e^{-\frac{(x-\eta)}{\theta}} & x > \eta, \\ 0 & x \le \eta. \end{cases}$$

Find the MME for θ, η.

(1) $\hat{\theta} + \hat{\eta} = \bar{X}$

(2) $\hat{\theta}^2 + (\hat{\theta} + \hat{\eta})^2 = \dfrac{\sum_{i=1}^{n} X_i^2}{n}$

$$\hat{\theta}^2 + \bar{X}^2 = \frac{\sum X_i^2}{n}$$

$$\Rightarrow \hat{\theta}^2 = \frac{\sum X_i^2}{n} - \bar{X}^2$$

$$\Rightarrow \hat{\theta} = \sqrt{\frac{\sum X_i^2}{n} - \bar{X}^2}$$

and $\hat{\eta} = \bar{X} - \sqrt{\frac{\sum X_i^2}{n} - \bar{X}^2}$

Example (9.2.3). Suppose we want to estimate $\mathbb{P}(X \geq 0)$

$$\mathbb{P}(X \geq 0) = \int_{\max(0,\eta)}^{\infty} \frac{1}{\theta} e^{-\frac{(x-\eta)}{\theta}} \, dx$$

$$= \int_{\max(\frac{-\eta}{\theta},0)}^{\infty} e^{-u} \, du \qquad \text{by substitution of } u = \frac{x-\eta}{\theta}$$

$$= e^{-\max(\frac{-\eta}{\theta},0)}$$

$$= \begin{cases} e^{\eta/\theta} & \eta < 0 \\ 1 & \eta \geq 0 \end{cases}$$

Since $\mathbb{P}(X \geq 0)$ is a function of θ, κ and we already have estimators for those, An estimator for $\mathbb{P}(X \geq 0)$ can be obtained by plugging in $\hat{\theta}, \hat{\kappa}$ for θ, κ:

$$\widehat{\mathbb{P}(X \geq 0)} = \begin{cases} e^{\hat{\kappa}/\hat{\theta}} & \hat{\kappa} < 0, \\ 1 & \hat{\kappa} \geq 0. \end{cases}$$

where $\hat{\theta}, \hat{\kappa}$ are as obtained before.

Example (9.2.5) Let $X_1, X_2 \sim$ iid $BIN(1, p) \sim BER(p)$.

Suppose it is known that $p \in \{.2, .3, .8\}$.

The MME for $p = E(X_1)$ is $\bar{X}$. However,

$\bar{X}$ can only take on the values $\{0, \frac{1}{2}, 1\}$,

so perhaps this is not an ideal estimator.

Consider the joint density function

$$f(x_1, x_2; p) = \begin{cases} p^{x_1+x_2}(1-p)^{2-x_1-x_2} & x_1, x_2 \in \{0,1\}, \\ 0 & o/w. \end{cases}$$

for various values of x_1, x_2, p:

	(x_1, x_2)			
p	(0,0)	(0,1)	(1,0)	(1,1)
.2	.64	.16	.16	.04
.3	.49	.21	.21	.09
.8	.04	.16	.16	.64

Note that if $p = .2$, then $(0,0)$ is the outcome with the highest probability.

If $p = .8$, then $(1,1)$ has the highest probability.

However we are not in the situation of knowing p and computing probabilities of various outcomes.

Instead, we want to guess the value of the parameter based on the outcome of the experiment.

Suppose the outcome of the experiment is $(1,0)$.

Which parameter would provide the highest probability of getting the outcome $(1,0)$?

We can formulate a function whose domain is the set of possible outcomes and whose range is the parameter space:

$$\hat{p} = \begin{cases} .3 & \text{if } (x_1, x_2) \in \{(0,1),(1,0)\}, \\ .2 & \text{if } (x_1, x_2) \in \{(0,0)\}, \\ .8 & \text{if } (x_1, x_2) = (1,1). \end{cases}$$

We could do a similar problem with continuous r.v.'s, but it wouldn't make sense to ask, "which parameter would provide the highest probability of getting the observed outcome" because the probability is zero for all parameters.

Instead, we replace the term "probability" with "likelihood" and we can continue by defining the likelihood of an outcome as the joint density evaluated at the outcome. Thus the likelihood is a function of the parameter.

In the last example we saw that for a fixed outcome (x_1, x_2) the likelihood function was

$$f(x_1, x_2; p) = \begin{cases} p^{x_1+x_2}(1-p)^{2-x_1-x_2} & p \in \{.2, .3, .8\}, \\ 0 & \text{o/w.} \end{cases}$$

Note that since x_1, x_2 is a fixed outcome, i.e. it is in $\{(0,0), (0,1), (1,0), (1,1)\}$, we don't need to specify the value of the joint density for other values of (x_1, x_2).

However, we should specify its value for $p \notin [0,1]$. Remember, we are considering this a function of p.

The method of maximum likelihood is to choose the parameter value that would provide the highest likelihood of the observed outcome as the estimate of the parameter.

Thus $\hat{\theta} \in \underset{\theta \in \Omega}{\operatorname{argmax}} \; f(x_1, x_2, \ldots x_n ; \theta)$

is a maximum likelihood estimator (MLE).

Note that if multiple parameters maximize the likelihood, then "the" MLE is not unique.

It is also possible that no argument in the parameter space, Ω, maximizes the likelihood. In this case, no MLE exists.

Example (9.2.6) Suppose $X_1, X_2, \ldots, X_n \sim$ iid $POI(\lambda)$

The likelihood function, $L(\lambda)$, is

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

We want to find λ that maximize $L(\lambda)$ (or equivalently and often easier $\log[L(\lambda)]$).

$$\log(L(\lambda)) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) = \sum_{i=1}^{n} -\lambda + x_i \log(\lambda) - \log(x_i!)$$

$$= -n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!)$$

Since $\log(L(\lambda))$ is a differentiable function of λ, we differentiate to find argmax.

Note that the parameter space is $\Omega = \{\lambda \mid \lambda > 0\}$.

$$\frac{d}{d\lambda} \log(L(\lambda)) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i$$

A critical point, λ, is a solution of

$$-n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0.$$

$$\Leftrightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Now choose your favorite method to verify that $\bar{x}$ maximizes $\log(L(\lambda))$.

$\bar{X}$ is the MLE for λ.

Theorem (9.2.1)

— The Invariance Property —

If $\hat{\theta}$ is the MLE of θ, the $u(\hat{\theta})$ is the MLE of $u(\theta)$.

Proof of case where u is 1-to-1:

$$L(\theta) = L\left(u^{-1}(u(\theta))\right) = L\left(u^{-1}(\gamma)\right) \quad \text{where } \gamma = u(\theta)$$

The LHS is maximized at $\hat{\theta}$.

Let $\hat{\gamma} \in \operatorname{argmax} L\left(u^{-1}(\gamma)\right)$ so that $\hat{\gamma}$ is the MLE for $\gamma = u(\theta)$.

The proof is complete if we can show that $\hat{\gamma} = u(\hat{\theta})$:

$$L\left[u^{-1}\left(u(\hat{\theta})\right)\right] = L(\hat{\theta}). \quad \square$$

Example (9.2.7)

Let $X_1, \ldots, X_n \sim$ iid $EXP(\lambda)$.

$$L(\lambda) = \begin{cases} \prod_{i=1}^{n} \frac{1}{\lambda} e^{-x_i/\lambda} & \lambda > 0, \\ 0 & \text{o/w.} \end{cases}$$

$$\log(L(\lambda)) = \sum_{i=1}^{n} -\log(\lambda) - \frac{x_i}{\lambda} = -n\log(\lambda) - \frac{\sum x_i}{\lambda}$$

$$\frac{d}{d\lambda}\log(L(\lambda)) = \frac{-n}{\lambda} + \frac{\sum x_i}{\lambda^2} = 0$$

$$\Rightarrow -n\lambda + \sum x_i = 0$$

$$\Rightarrow \lambda = \frac{\sum x_i}{n} \quad \text{is a critical point.}$$

$\bar{X}$ is the MLE for λ.

Now, if we want the MLE for $\mathbb{P}(X \geq 1) = e^{-1/\lambda}$, we simply plug in $\hat{\lambda}$:

The MLE for $e^{-1/\lambda}$ is $e^{-1/\bar{x}}$.

Example (9.2.10) Suppose $X_i \sim$ iid $N(\mu, \sigma^2)$

Then $L(\mu, \sigma^2) = \prod\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$

and $\ell = \text{Log}(L) = \sum\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$

$$= C - \frac{n}{2}\log(\sigma^2) - \frac{\sum(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \ell}{\partial \mu} = \frac{+2\sum(x_i-\mu)}{2\sigma^2} = 0 \Rightarrow \sum x_i = n\mu \Rightarrow \mu = \bar{x}$$

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{-n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i-\mu)^2 = 0$$

Replace μ with $\bar{x}$ since we already saw that μ must be $\bar{x}$ at a local maximum:

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i-\bar{x})^2 = 0$$

$$\Rightarrow -n\sigma^2 + \sum(x_i-\bar{x})^2 = 0$$

$$\Rightarrow \sigma^2 = \frac{\sum(x_i-\bar{x})^2}{n}$$

Now, one must verify that this maximizes ℓ.

The bivariate second derivative test is this:

$$D = \det\begin{pmatrix} \ell_{xx} & \ell_{xy} \\ \ell_{yx} & \ell_{yy} \end{pmatrix} > 0 \text{ implies local extremum.}$$

$D > 0$ and $\ell_{xx} > 0$ implies local min.

$D > 0$ and $\ell_{xx} < 0$ implies local max.

$\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{\sum(x_i-\bar{x})^2}{n}$ do maximize the likelihood and are therefore MLEs.

~~Note that the MLE for σ^2 is not the MME.~~

Actually, the MM and ML estimators are the same in this case.

Example (9.2.11) Suppose $X_i \sim$ iid with

$$f(x;\eta,\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{(x-\eta)}{\theta}} & x \geq \eta \\ 0 & o/w \end{cases}$$

Note: $\theta > 0$ and $\eta \leq x_{1:n}$

$$L = \prod \left(\frac{1}{\theta} e^{-\frac{(x_i-\eta)}{\theta}} \right) 1\{\eta \leq x_{1:n},\ \theta > 0\}$$

$$\ell = \sum \left(-\log(\theta) - \frac{(x_i-\eta)}{\theta} \right) 1\{\eta \leq x_{1:n},\ \theta > 0\}$$

$$= \left(-n\log(\theta) - \frac{\sum x_i}{\theta} + \frac{n\eta}{\theta} \right) 1\{\eta \leq x_{1:n},\ \theta > 0\}$$

Assume $\eta \leq x_{1:n}$, $\theta > 0$.

$$\frac{\partial \ell}{\partial \eta} = \frac{n}{\theta}$$

$$\frac{\partial \ell}{\partial \theta} = \frac{-n}{\theta} + \frac{\sum x_i}{\theta^2} - \frac{n\eta}{\theta^2} = 0$$

$$\Rightarrow -n\theta + \sum x_i - n\eta = 0$$

$$\Rightarrow \theta = \frac{\sum x_i}{n} - \eta = \bar{x} - \eta$$

Since $\frac{\partial \ell}{\partial \eta}$ cannot be zero, we check if ℓ is maximized on the boundary.

For fixed θ, it is clear that L is largest when $\eta = x_{1:n}$.

Thus $x_{1:n}$ is the MLE for η.

Plugging this in to $\frac{\partial \ell}{\partial \theta}$, we see that

$\bar{x} - x_{1:n}$ is the MLE for θ.

We now have two methods for defining estimators and need some criteria to determine which is 'best'.

Definition (9.3.1) If $E(T) = \tau(\theta)$ $\forall \theta \in \Omega$, T is an unbiased estimator of $\tau(\theta)$.

Example (9.3.1) Suppose $X_i \sim iid\ N(\mu, 3^2)$.

We have seen that $E(\bar{X}) = \mu$ $\forall \mu \in \mathbb{R}$.
Thus $\bar{X}$ is an unbiased estimator of μ.

Let $\tau(\mu)$ be the 95th percentile of $N(\mu, 3^2)$
Thus $\tau(\mu) = \mu + z_{.95}(3) = \mu + 3(1.65) = \mu + 4.95$

$E(\bar{X} + 4.95) = \mu + 4.95$; therefore $\bar{X} + 4.95$ is an unbiased estimator for the 95th percentile.

Example (9.3.2): Recall that if $X \sim GAM(\theta, k)$, then $\frac{2X}{\theta} \sim \chi^2(2k)$.

Recall that $EXP(\theta) \sim GAM(\theta, 1)$

Thus if $X_i \sim$ iid $EXP(\theta)$, then $\frac{2X_i}{\theta} \sim$ iid $\chi^2(2)$

and $\sum \frac{2X_i}{\theta} \sim \chi^2(2n)$

so that $\frac{2n}{\theta}\bar{X} \sim \chi^2(2n)$

and $E\left(\frac{2n}{\theta}\bar{X}\right) = 2n$

so that $E(\bar{X}) = \theta$ and $\bar{X}$ is an unbiased estimator for θ. (Note that $\bar{X}$ is always unbiased for EX.)

Suppose now that you want to estimate $\frac{1}{\theta}$. Since $\bar{X}$ is the MLE for θ, $\frac{1}{\bar{X}}$ is the MLE for $\frac{1}{\theta}$.

However, it can be shown (equation 8.3.6 from theorem 8.3.2) that $E\left(\frac{\theta}{2n}\frac{1}{\bar{X}}\right) = \frac{1}{2}\frac{\Gamma(n-1)}{\Gamma(n)} = \frac{1}{2(n-1)}$

$\Rightarrow$ $E\left(\frac{1}{\bar{X}}\right) = \left(\frac{n}{n-1}\right)\frac{1}{\theta}$. Thus $\frac{1}{\bar{X}}$ is a biased estimator for $\frac{1}{\theta}$. The bias is $E\left(\frac{1}{\bar{X}}\right) - \frac{1}{\theta} = \frac{1}{\theta}\left(\frac{n}{n-1} - 1\right)$

$= \left(\frac{1}{n-1}\right)\frac{1}{\theta}$

and this is small for large n.

Also, in this case, we can create an unbiased estimator for $\frac{1}{\theta}$ by multiplying the MLE by $\frac{n-1}{n}$.

$\left(\frac{n-1}{n}\right)\frac{1}{\bar{X}}$ is not an MLE, but it is unbiased for $\frac{1}{\theta}$. This trick of starting with the MLE and multiplying by a constant is how we obtain the sample variance as an unbiased estimator of the population variance (in a normal population).

Unfortunately, it is not always possible to create an unbiased estimator in this way.

Example (9.3.3) Let $X_i \sim$ iid $EXP(\theta)$.
We have seen that $\bar{X}$ is unbiased for θ.

It can be shown using the cdf technique that $X_{1:n} \sim EXP(\theta/n)$.

Thus $n X_{1:n}$ is unbiased for θ.

Which estimator is better?

$var(n X_{1:n}) = n^2\left(\frac{\theta}{n}\right)^2 = \theta^2$,

$var(\bar{X}) = \frac{\theta^2}{n}$.

The question of which unbiased estimator is best leads to the definition of the

"Uniformly Minimum Variance Unbiased Estimator" (UMVUE)

An estimator T^* for $\tau(\theta)$ is a UMVUE if it is unbiased and $var(T^*) \leq var(T)$ $\forall\, \theta \in \Omega$ and all unbiased estimators, T, of $\tau(\theta)$.

It seems desirable to have a UMVUE, but how would you even know if you did?

The "Cramer-Rao Lower Bound" (CRLB), if it exists, gives a lower bound on the variance of an unbiased estimator. If your estimator has a variance equal to the CRLB, then yours is a UMVUE.

Let $u(x_1, \ldots, x_n; \theta) = \frac{\partial}{\partial \theta} \log f(x_1, \ldots, x_n; \theta)$

We will show that

1) $E(U) = 0$ where $U = u(X_1, \ldots, X_n; \theta)$

2) $\tau'(\theta) = E(TU)$ where $\tau(\theta) = E(T)$.

3) $\mathrm{var}(U) = EU^2$ and $\mathrm{cov}(U, T) = E(UT)$

4) $\left(\mathrm{Cov}(T, U)\right)^2 \leq \mathrm{var}(T)\,\mathrm{var}(U)$

5) $$\frac{[\tau'(\theta)]^2}{E\left[\frac{\partial}{\partial \theta} \log f(X_1, \ldots X_n; \theta)\right]^2} \leq \mathrm{var}(T).$$

6) $$\frac{[\tau'(\theta)]^2}{nE\left[\frac{\partial}{\partial \theta} \log f(X; \theta)\right]^2}$$

1: $E(U) = \int \cdots \int u(x_1, \ldots, x_n; \theta) f(x_1, \ldots, x_n; \theta)\, dx_1, \ldots, dx_n$

$= \int \cdots \int \frac{\partial}{\partial \theta} f(x_1, \ldots, x_n; \theta)\, dx_1, \ldots, dx_n$

$= \frac{\partial}{\partial \theta} \int \cdots \int f(x_1, \ldots, x_n; \theta)\, dx_1 dx_2 \ldots dx_n$

$= \frac{\partial}{\partial \theta} 1 = 0$

2: $\tau'(\theta) = \frac{d}{d\theta} E(T) = \frac{d}{d\theta} \int \cdots \int t(x_1, \ldots, x_n) f(x_1, \ldots, x_n; \theta) \, dx_1, dx_2 \ldots dx_n$

$$= \int \cdots \int \frac{\partial}{\partial \theta} t(x_1, \ldots, x_n) f(x_1, \ldots, x_n; \theta) \, dx_1 \ldots dx_n$$

$$= \int \cdots \int t(\underline{x}) \quad u(\underline{x}; \theta) \quad f(\underline{x}; \theta) \quad d\underline{x}$$

$$= E(TU)$$

3: Due to (1), $E(U^2) = var(U)$ and $E(UT) = cov(U, T)$

In particular $\tau'(\theta) = cov(U, T)$

4: Since $\rho \in [-1, 1]$, $\frac{cov(U,T)^2}{var(U)\, var(T)} \leq 1$

$$\Rightarrow cov(U,T) \leq var(U)\, var(T)$$

5: $$\Rightarrow E(UT) \leq E(U^2)\, var(T)$$

$$\Rightarrow \frac{E(UT)}{E(U^2)} \leq var(T)$$

$$\Rightarrow \frac{\tau'(\theta)}{E\left[\frac{\partial}{\partial \theta} \log f(\underline{X}; \theta)\right]^2} \leq var(T)$$

6: Since $X_1, \ldots, X_n$ is a random sample,

$$E(U^2) = \operatorname{var}(U) = \operatorname{var}\left(\frac{\partial}{\partial\theta} \sum \log\left[f(X_i;\theta)\right]\right)$$

$$= \sum \operatorname{var}\left(\frac{\partial}{\partial\theta} \log\left[f(X_i;\theta)\right]\right)$$

$$= n \operatorname{var}\left[\frac{\partial}{\partial\theta} \log f(X_1;\theta)\right]^2$$

$$= n\, E\left[\frac{\partial}{\partial\theta} \log f(X_1;\theta)\right]^2$$

Thus (from (5)), $\dfrac{\tau'(\theta)}{n\, E\left[\frac{\partial}{\partial\theta} \log f(X_1;\theta)\right]^2} \le \operatorname{var}(T)$.

Example: (9.3.4) Let $X_i \sim$ iid $EXP(\theta)$
Find the CRLB.

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

$$\frac{\partial}{\partial\theta} \log\left[f(x;\theta)\right] = \begin{cases} \frac{\partial}{\partial\theta}\left(-\log(\theta) - \frac{x}{\theta}\right) & x > 0, \\ \text{undefined} & \text{o/w.} \end{cases}$$

$$= \begin{cases} \frac{-1}{\theta} + \frac{x}{\theta^2} & x > 0 \\ \text{!!} & x \le 0 \end{cases}$$

$$E\left[\frac{X-\theta}{\theta^2}\right]^2 = \frac{E(X-\theta)^2}{\theta^4} = \frac{\theta^2}{\theta^4} = \frac{1}{\theta^2}$$

Thus, the CRLB is $\dfrac{1}{n/\theta^2} = \dfrac{\theta^2}{n}$

which we saw was the variance of $\bar{X}$.

Thus $\bar{X}$ is a UMVUE for θ. ☺

Example (9.3.5) Find a UMVUE for θ in $GEO(\theta)$

$$f(x;p) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \ldots, \\ 0 & o/w. \end{cases}$$

$$\sum \frac{\partial}{\partial p} \log\left[f(X;p)\right] = \sum \frac{\partial}{\partial p}\left(\log(p) + (X-1)\log(1-p)\right)$$

$$= \sum\left(\frac{1}{p} + \left(\frac{X-1}{1-p}\right)(-1)\right)$$

Thus the form of a UMVUE must be

$$a\sum\left(\qquad\qquad\right) + b$$

or equivalently, $c\bar{X}+d$.

$c=1$ and $d=0$ to have an unbiased estimator

Thus if there is an unbiased estimator that has a variance equal to the CRLB, it must be $\bar{X}$.

Theorem (9.3.1) If T is an estimator of $\tau(\theta)$ which achieves the CRLB, then only linear function of $\tau(\theta)$ will admit such an estimator.

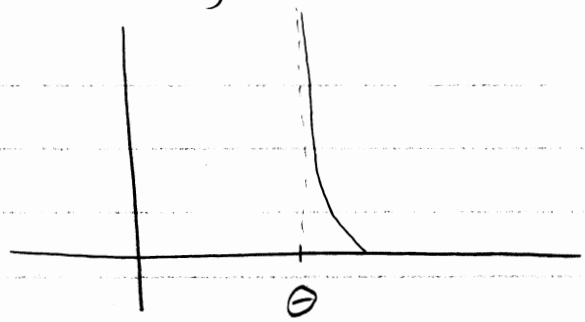
Definition: If T_1, T_2 are unbiased estimators of $\tau(\theta)$, then the "relative efficiency" of T_1 compare to T_2 is $\dfrac{var(T_2)}{var(T_1)}$.

Definition: An unbiased estimator, T_1, is "efficient" if $\dfrac{var(T_1)}{var(T_2)} \leq 1$ for any unbiased estimator T_2 and any $\theta \in \Omega$.

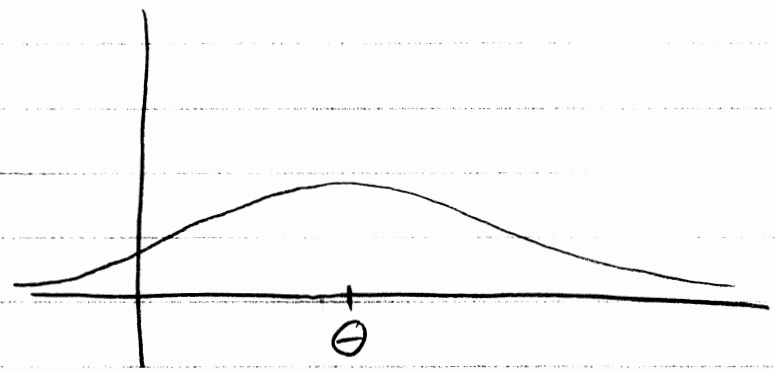
Definition The efficiency of an unbiased estimator is its relative efficiency compared to an efficient estimator.

Example

density of estimator 1 for θ



Density of estimator 2 for θ



Estimator 2 is unbiased, but is it better?

Definition The "mean square error" (MSE) of an estimator is $E(T-\tau(\theta))^2$

In most cases, I would prefer the estimator with the smaller MSE.

Is there a situation in which you might prefer the unbiased estimator with higher MSE?

Theorem (9.3.2) $MSE = var(T) + (bias(T))^2$

Question: Is it possible to find an estimator that has "uniformly minimum MSE"?

Consider an estimator that is always 5. If the parameter is 5, this estimator has the smallest MSE, but if the true parameter is not 5...

For any particular value of the parameter, the degenerate estimator of the same value has an MSE of 0.

Why not use this as the estimator?

Example: (9.3.9) Note that for $f(x;\theta,\eta) = \begin{cases} \frac{1}{\theta} e^{-\frac{(x-\eta)}{\theta}} & x > \eta \\ 0 & x \le \eta \end{cases}$

Suppose θ is known to be 1.
the MLE is $\hat{\eta}_1 = X_{1:n}$
and the MME is $\hat{\eta}_2 = \bar{X} - 1$

It can be shown that $\bar{X} - \eta \sim GAM(\frac{1}{n}, n)$ using the MGF technique.

It can be shown that $X_{1:n} - \eta \sim EXP(\frac{1}{n})$ using the cdf technique.

Thus $MSE(X_{1:n}) = E(X_{1:n} - \mu)^2 = E(Y^2)$
where $Y \sim EXP(\frac{1}{n})$

Since $var(Y) = E(Y^2) - (EY)^2$,
$E(Y^2) = var(Y) + (EY)^2 = (\frac{1}{n})^2 + (\frac{1}{n})^2 = \frac{2}{n^2}$

$MSE(\bar{X}-1) = E(\bar{X}-1-\mu)^2$

or using Theorem 9.3.2, $= var(\bar{X}-1) + [E(\bar{X}-1-\mu)]^2$

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

$$= \frac{1}{n}$$

Definition (9.4.1) An estimator T_n is "consistent" for $\tau(\theta)$ if

$$T_n \xrightarrow{\mathbb{P}} \tau(\theta)$$

Definition (9.4.3) An estimator T_n is "asymptotically unbiased" for $\tau(\theta)$ if

$$\lim_{n\to\infty} E(T_n) = \tau(\theta)$$

Example The MLE of σ^2 in a normal population is $\frac{\Sigma(X_i-\bar{X})^2}{n}$, which is biased but asymptotically unbiased and even consistent.

Example: $nX_{1:n}$ is an unbiased estimator for θ in an $EXP(\theta)$ population, but it is not consistent

Recall that $var(nX_{1:n}) = \theta^2$ so it is <u>not</u> possible that $nX_{1:n} \xrightarrow{\mathbb{P}} \theta$.

Def: An asymptotically unbiased estimator $T_1^{(n)}$ "asymptotically efficient" if

$$\frac{var(T_1^{(n)})}{var(T_2^{(n)})} \leq 1 \qquad \forall \text{ assymptotically unbiased } T_2^{(n)} \quad \forall\ \theta \in \Omega$$

Def: The asyptotic relative efficiency of $T_1^{(n)}$ compare to $T_2^{(n)}$ is $\lim_{n \to \infty} \frac{var(T_2^{(n)})}{var(T_1^{(n)})}$

Even when no estimator exists which is unbiased and meets the CRLB, it is often possible to find an estimator that is asymptotically efficient.

EXAMPLE: Show that $\frac{1}{\bar{X}}$ is asymptotically efficient for $\frac{1}{\theta}$ in an exponential population.

1) $\frac{1}{\bar{X}} \xrightarrow{\mathbb{P}} \frac{1}{\theta}$ so $\frac{1}{\bar{X}}$ is consistent and thus asymptotically unbiased.

$$\text{CRLB} = \frac{(\tau'\theta)^2}{n E\left[\frac{\partial}{\partial\theta} \log f(X;\theta)\right]^2} = \frac{\frac{1}{\theta^4}}{n E\left[\frac{\partial}{\partial\theta}\left(-\log(\theta) - \frac{X}{\theta}\right)\right]^2}$$

$$= \frac{\frac{1}{\theta^4}}{n E\left[-\frac{1}{\theta} + \frac{X}{\theta^2}\right]^2}$$

$$= \frac{\frac{1}{\theta^4}}{n E\left(\frac{X-\theta}{\theta^2}\right)^2} = \frac{\frac{1}{\theta^4}}{n \frac{\theta^2}{\theta^4}} = \frac{1}{n\theta^2}$$

Recall that $\frac{2n}{\theta}\bar{X} \sim \chi^2(2n)$

$$\text{var}\left(\frac{1}{\bar{X}}\right) = E\left(\frac{1}{\bar{X}}\right)^2 - \left(E\frac{1}{\bar{X}}\right)^2$$

which (from Theorem 8.3.2) is

$$\left(\frac{n}{n-1}\right)^2 \frac{1}{(n-2)\theta^2}.$$

To check asymptotic efficiency,

$$\frac{\text{CRLB}}{\text{var}\left(\frac{1}{\bar{X}}\right)} = \frac{\frac{1}{n\theta^2}}{\left(\frac{n^2}{(n-1)^2(n-2)\theta^2}\right)} \longrightarrow 1.$$

Thus $\frac{1}{\bar{X}}$ is asymptotically efficient.

ASYMPTOTIC PROPERTIES of an MLE:

If $\frac{\partial}{\partial\theta} \log f(x;\theta)$ exists and $\{x \mid f(x;\theta)\}$ doesn't depend on θ and if (other required derivatives),

then
1) $\hat{\theta}_n$ exists and is unique
2) $\hat{\theta}_n$ is consistent
3) $\hat{\theta}_n$ is asymptotically normal with asymptotic mean, θ, and asymptotic variance equal to the CRLB
4) $\hat{\theta}_n$ is asymptotically efficient.

Chapter 10 — Sufficiency & Completeness

Def: A statistic, S, is "sufficient" for θ if $f_{X|S}(\underline{x}|s)$ does not depend on θ.

You can think of a sufficient statistic containing all of the information from the sample about the unknown parameter.

That is, if you know the outcome of a sufficient statistic, then knowing the outcome of the entire sample doesn't add any knowledge about θ.

Def: If $\underline{S}$ is vector-valued, then we say that $\underline{S}$ is "jointly sufficient" for θ. Note that $\underline{\theta}$ may also be vector-valued.

Ex: The entire sample is a set of jointly sufficient stats.

Ex: The order statistics are also jointly sufficient.

Def: A set of jointly sufficient statistics is "minimal" if it is a function of every other set of sufficient statistics.

Example: If $\bar{X}$ is sufficient for θ and $(X_1, X_2, \ldots, X_n)$ are jointly sufficient, then $\bar{X}$ is a function of $(X_1, X_2, \ldots, X_n)$ but not the other way around.

Ex: (10.1.1) A coin is tossed n times.
Let $X_i = \mathbb{1}\{i^{th} \text{ toss is heads}\}$.

It seems like knowing ΣX_i may tell us all available information about the unknown parameter p.

Let's confirm that $S = \Sigma X_i$ is sufficient.

$$f_{\underline{X}|S}(\underline{x}|s) = \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, S = s)}{\mathbb{P}(S = s)}$$

$$= \begin{cases} \dfrac{p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}}{\binom{n}{s} p^s (1-p)^{n-s}} & x_i \in \{0,1\} \text{ and } \Sigma x_i = s, \\ 0 & \text{o/w.} \end{cases}$$

$$= \begin{cases} 1/\binom{n}{s} & x_i \in \{0,1\} \text{ and } \Sigma x_i = s, \\ 0 & \text{o/w.} \end{cases}$$

Since this does not depend on p, S is sufficient.

In fact, it is minimal sufficient.

Theorem (10.2.1) The Factorization Criterion

S is sufficient for θ iff

$$f_{\underline{X}}(\underline{x};\theta) = g(s;\theta)\,h(\underline{x}) \quad \text{for some } g, h.$$

In other words, if the joint density of the sample can be factored into a function of s and θ only and a function of $\underline{x}$ only, the S is sufficient.

Ex: Repeat example 10.1.1 with this theorem.

$$f(\underline{x};\theta) = \underbrace{p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}}_{g(\Sigma x_i;\, p)} \underbrace{1\{x_i \in \{0,1\}\}}_{h(\underline{x})}$$

Ex (10.2.3): Let $X_i \sim$ iid $\text{Unif}(0,\theta)$

$$f(\underline{x};\theta) = \prod \frac{1}{\theta} 1\{x_i \in (0,\theta)\} = \frac{1\{x_{1:n} > 0,\ x_{n:n} < \theta\}}{\theta^n}$$

Thus by the factorization criterion, $X_{1:n}, X_{n:n}$ are jointly sufficient for θ.

Question: Are they minimal sufficient?

Question: What is the minimal sufficient statistic?
It is $X_{n:n}$ also by the factorization criterion.

Example (10.2.4) Let $X_i \sim$ iid $N(\mu, \sigma^2)$

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

$$= (2\pi\sigma^2)^{n/2} e^{(-1/2\sigma^2)\left[\sum (x_i-\mu)^2\right]}$$

$$= (2\pi\sigma^2)^{n/2} e^{-\frac{1}{2\sigma^2}\left[\sum x_i^2 - 2\mu \sum x_i + n\mu^2\right]}$$

Thus $\sum X_i^2$ and $\sum X_i$ are jointly sufficient for μ, σ^2.

Note that there is a one-to-one function between the above sufficient statistics and the MLEs $\bar{X}$ and $\frac{\sum (X_i - \bar{X})^2}{n}$.

Ex: (10.2.5) Let $X_i \sim$ iid $UNIF(\theta, \theta+1)$

$$f(\underline{x}; \theta) = 1\{x_{1:n} > \theta, \ x_{n:n} < \theta + 1\}$$

Thus $X_{1:n}, X_{n:n}$ are jointly sufficient for θ.

Question: Are they minimal suffient? (Yes.)

Theorem (10.3.1) If S is sufficient for θ and $\hat{\theta}$ is the <u>unique</u> MLE, then $\hat{\theta}$ is a function of S.

Question: If $\hat{\theta}$ is the unique MLE and is sufficient, then is $\hat{\theta}$ necessarily minimal sufficient?

Yes, the theorem states it is a function of <u>every</u> sufficient statistic.

Theorem (10.3.4) If T is unbiased for θ and S is sufficient for θ, then

1) $T^* = E(T|S)$ is unbiased
2) T^* is a function of S
3) $var(T^*) \leq var(T) \quad \forall \theta \in \Omega$
   and $var(T^*) < var(T)$ for some θ
   unless $T^* = T$ with $\mathbb{P} = 1$.

This theorem says that the unbiased estimator with the smallest variance is a function of the sufficient statistic.

Def: A family of density functions, $f(x;\theta)$, is "complete" if

$$E(u(X)) = 0 \Rightarrow \mathbb{P}(u(X) \neq 0) = 0$$

Def: A statistic, T, is complete if its density could be any member of a complete family.

It can be shown that the following families are complete: BIN, POI, EXP, GAM, N.

Thus if a statistic has distribution $N(\mu, \sigma^2)$ where μ, σ^2 are the unknown parameters of the population, then the statistic is complete. E.g. $\bar{X}$.

Theorem (10.4.1)

If 1) $X_i \sim iid \quad f(\underline{x};\theta)$

2) S is complete and sufficient

3) $T = t(S)$ is a function of S and unbiased

Then T is a UMVUE.

Proof: Rao-Blackwell (10.3.4) shows that the unbiased estimator with the lowest variance is a function of S.

So, we need only consider functions of S.

Let T_2 be a function of S and unbiased for θ.

Then $E(T_2 - T) = 0$

$\Rightarrow T_2 - T = 0$ with probability 1 because

$T_2 - T$ is a function S, which is complete.

Thus $T_2 = T$ with prob. 1 and $var(T_2) = var(T)$. $\square$

Note that we also showed that there can only be one unbiased function of a complete statistic (where T_1 and T_2 are considered equal if $\mathbb{P}(T_1 \neq T_2) = 0$).

Example (10.4.1) $X_i \sim$ iid $POI(\lambda)$

$$f(\underline{x};\lambda) = \prod \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \mathbb{1}\{x_i \in \{0,1,2,\ldots\}\}$$

$$= e^{-n\lambda} \lambda^{\Sigma x_i} \prod \left(\frac{\mathbb{1}\{x_i \in \{0,1,2,\ldots\}\}}{x_i!} \right).$$

Thus ΣX_i is sufficient for λ.

Recall that $\Sigma X_i \sim POI(n\lambda)$.

Suppose $E(u(Y)) = 0 \quad \forall \ \theta > 0$ where $Y \sim POI(\theta)$.

That is, $\displaystyle\sum_{y=0}^{\infty} u(y) \frac{e^{-\theta}\theta^y}{y!} = 0 \quad \forall \ \theta > 0.$

Note that $\dfrac{e^{-\theta}\theta^y}{y!} > 0$ for $y \in \{0,1,\ldots\}$

We could choose coefficients $u(y)$ to make the sum zero, but the only coefficients that will make it zero $\forall \ \theta > 0$ are zeros.

Thus the POISSON family is complete

$\dfrac{\Sigma X_i}{n} = \bar{X}$ is unbiased for θ and hence a UMVUE.

This can be verified with the CRLB.

Thus by Theorem 9.3.1, a non-linear function of λ will not admit a CRLB estimator.

Example: $E(\bar{X}^2) = \lambda^2 + \frac{\lambda}{n}$

Thus $E\left(\bar{X}^2 - \frac{\bar{X}}{n}\right) = \lambda^2$

$\bar{X}^2 - \frac{\bar{X}}{n}$ will not achieve the CRLB,

but since it is an unbiased function of a complete suffient statistic, it is a UMVUE for λ^2.

Def: If 1) $f(x;\theta) = c(\theta)\, h(x) \exp\left(\sum_{i=1}^{k} q_i(\theta)\, t_i(x)\right)$

2) $\Omega = \{\underline{\theta} \mid a_i \le \theta_i \le b_i \text{ for } i = 1, 2, \ldots k\}$

3) $A = \{x \mid f(x;\theta) > 0\}$ does not depend on θ

4) $q_i(\theta)$ are non-trivial, functionally independent, continuous functions of θ

(continuous case) 5a) $t_i'(x)$ are linearly independent & continuous on A.

(discrete case) 5b) $t_i(x)$ are non-trivial and linearly independent

then the density is a member of the Reg. Exp. Class.

~~Density families in the REC are complete.~~

That's not quite true. If the population is a member of the REC, then the sufficient statistic from a random sample is complete.

Example (10.4.4) $X_i \sim iid \; N(\mu, \sigma^2)$

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

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

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\mu^2/2\sigma^2} e^{\frac{-1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x}$$