MGTECON 603 Section Notes

Cameron Taylor

Fall 2019

Introduction

These are section/discussion session notes for MGTECON 603. In each section I go over a few problems related to the lecture notes for about an hour and leave the rest of time for questions. Some of the exercises are taken from previous section notes which had contributions from Breno Vieira and Evgeni Drynkin.

Section 1 - Probability Spaces, Random Variables, and Transformations of Random Variables

1. (Sigma Algebra) Consider $X = \{1, 2, 3, 4\}$. What is the smallest sigma algebra that contains $\{1\}$? What is the smallest sigma algebra that contains $\{1\}$ and $\{2, 3\}$?

   Solution: In any sigma algebra we must have empty set and X so always put those in. Now consider $\{1\}$. Sigma algebras must be closed under unions and complements. Taking a union of $\{1\}$ with anything else considered is already in there so we are fine. How about the complement? The complement is $\{2, 3, 4\}$ so this must be in this sigma algebra. Do we need anything else? If we take complements and unions of all the elements $\emptyset, \{1\}, \{2, 3, 4\}, X$ we see that everything is closed, so this is the smallest!

   Now consider the same procedure for when adding in $\{2, 3\}$. First the complement $\{1, 4\}$ must be in there. The union $\{1, 2, 3\}$ must also be in there. The complement of this set $\{4\}$ must be in there. Then we must also add $\{2, 3, 4\}$. Anything else? Nope we are good!

2. (Independence) Suppose that $f_{X,Y}(x, y) = \mathbf{1}\{(x, y) \in S\}$ for some $S \subseteq \mathbb{R}^2$ that is closed and convex. First, what does this tell us about the area of S? Second, does this imply that X and Y are independent?

Solution: The formal way to do this is to look at whether the joint density can be represented as a product of marginals - like independence in probability. The intuitive way to answer this question is to draw pictures. Suppose that S is the unit-square. Then if we were to compute probabilities can show independence. However, what if we deviate from the unit-square? In particular, consider the example where S is the disc of radius $1/\sqrt{\pi}$ centered at 0. Then to find the marginals, one needs to integrate out the joint distribution. We will go over this more in Lecture 5.

S is formally defined as

$$S = \Big\{(x, y) \in \mathbb{R}^2 \ : \ x^2 + y^2 \leq \frac{1}{\pi}\Big\}.$$

Suppose that the joint distribution of X and Y admits a uniform distribution on S. Then to find the marginal densities we "integrate" out the other variable. In particular, for X we get that

$$\begin{aligned} f_X(x) &= \int_{y \in \mathbb{R}} f_{X,Y}(x, y) dy \\ &= \int_{\{y \ : \ x^2+y^2 \leq \frac{1}{\pi}\}} 1 dy \\ &= \int_{y \in \left[-\sqrt{\frac{1}{\pi}-x^2}, \sqrt{\frac{1}{\pi}-x^2}\right]} dy \\ &= 2\sqrt{\frac{1}{\pi} - x^2} \end{aligned}$$

and so, we have that the marginal density of X is

$$f_X(x) = 2\sqrt{\frac{1}{\pi} - x^2}, \ x \in [-\frac{1}{\sqrt{\pi}}, \frac{1}{\sqrt{\pi}}].$$

By symmetry, the marginal pdf of Y is the symmetric/the same.

Now note that

$$f_X(x) f_Y(y) = 4\sqrt{\frac{1}{\pi} - x^2}\sqrt{\frac{1}{\pi} - y^2}, \ x, y \in [-\frac{1}{\sqrt{\pi}}, \frac{1}{\sqrt{\pi}}]$$

which is NOT the same as the original joint density, and so they are not independent.

3. (CDFs) Show that the CDF is right continuous. Why can't we prove that it is left continuous?

Solution: Right continuous means that a limit approaching "from the right" passes through. To do this we will utilize the fact that probability functions are "continuous" in sets. Just assume this is true (look up proving continuity of probability measure if you're interested).

To figure this out recall that F_X is a function. What is $F_X(x)$? It is $P(X \leq x) = P((-\infty, x])$. So how do we get a limit from the right. In this case, consider the sequence of sets

$$A_n = (-\infty, x + 1/n]$$

Then $x \in A_n$ for all n. Thus $\lim_n A_n = (-\infty, x] := A$. Then we can take the limit from the right as:

$$\begin{aligned}
\lim_{u \downarrow x} F_X(u) &= \lim_n P(A_n) \\
&= P(\lim_n A_n) \\
&= P((-\infty, x]) \\
&= F_X(x)
\end{aligned}$$

using continuity of the probability function.

So why doesn't this work for left-continuous? Consider trying something similar. Let $B_n = (-\infty, x - 1/n]$. Then $\lim B_n = (-\infty, x)$ and so doing the same argument we get

$$\begin{aligned}
\lim_{u \uparrow x} F_X(u) &= \lim_n P(B_n) \\
&= P(\lim_n B_n) \\
&= P((-\infty, x)) \\
&= F_X(x) + P(X = x)
\end{aligned}$$

and if there is a "mass point" or mass at x, then these will differ and we will have a discontinuity.

4. (Integration and Differentiation) Leibniz rule and differentiating under an integral in general is a powerful tool. When does Leibniz rule and passing an a derivative through an integral fail?

   Solution: In full mathematical generality, we cannot pass a derivative through an integral. The reason is that we cannot pass a limit through an integral. For our purposes, the only real challenge is when we look at improper integrals, integrating over the whole real numbers. In general if we integrate over a finite range, as long as that finite range is well behaved

and everything is smooth and differentiable we won't have problems. What goes wrong integrating over the whole real line?

Consider $f(x, \theta)$. Then

$$\int \frac{\partial f(x,\theta)}{\partial \theta} dx = \int \lim_{h \to 0} \frac{f(x, \theta + h) - f(x, \theta)}{h} dx$$

while

$$\frac{\partial}{\partial \theta} \int f(x, \theta) dx = \lim_{h \to 0} \int \frac{f(x, \theta + h) - f(x, \theta)}{h} dx$$

Thus everything comes down to passing limits through integrals.

Consider the following sequence of functions:

$$f_n(x) = \frac{1}{2n} \mathbf{1}\{x \in [-n, n]\}$$

defined over the real numbers. Let's draw this.

Then, if we integrate over this for any n, we have that the solution is 1 since the area is $2n$. However, $f_n(x)$ converges to the function 0 on the real line (in fact it converges uniformly!) but then the integral of this limit is 0!

The issue in this specific example is that the support or area that we integrate over (the real numbers) is unbounded. If we only consider finite integrals, then this would work.

There is another general issue: the convergence needs to be fast enough. In general, one needs *uniform convergence* rather than *pointwise convergence*.

So how do we get around these issues? In general, one of the main tools we use is *Lebesgue Dominated Convergence Theorem*: essentially if we can find a function to bound the function of interest over the whole real line that is integrable, then we are good to go. In this way, we don't need to worry about integrating over the whole real line. See Theorem 2.4.2 in Casella and Berger.

In the differentiation case this amounts to finding a bound for the derivative (the thing we are integrating).

See Example 2.4.5 in Casella and Berger for more details and some practice.

5. (Integration by Parts) Integration by parts is a classic integration technique/strategy.

   The formula for integration by parts is

$$\int_a^b u(x)v'(x)dx = [u(x)v(x)]_a^b - \int_a^b u'(x)v(x)dx$$

We will do two examples

(a) Use IBP to integrate $\int_1^e \log(x)dx$

Solution: Let $u(x) = \log(x)$ and $v'(x) = 1$. Then $u'(x) = 1/x$ and $v(x) = x$. So

$$\int_1^e \log(x)dx = [\log(x)x]_1^e - \int_1^e \frac{x}{x}dx = \log(e)e - \log(1)1 - (e-1) = 1$$

(b) Use IBP to integrate $\int_0^\infty \lambda x e^{-\lambda x}dx$

Solution: Let $u(x) = x$ and $v'(x) = \lambda e^{-\lambda x}$. Then $v(x) = -e^{-\lambda x}$. So

$$\begin{aligned}\int_0^\infty \lambda x e^{-\lambda x}dx &= [-xe^{-\lambda x}]_0^\infty + \int_0^\infty e^{-\lambda x}dx \\ &= 0 - 0 + [-\frac{e^{-\lambda x}}{\lambda}]_0^\infty \\ &= \frac{1}{\lambda}\end{aligned}$$

using the fact that $xe^{-\lambda x}$ converges to 0 (can use L'Hopitals Rule!).

Section 2 - Expected Value, Transformations, MGFs, and Standard Distributions

1. (Transformation of Random Variable) When the transformation is monotone, we can use the tools that we learned in lecture to derive the pdf directly. Just make sure to be careful with the support and ensuring monotonicity! I will now show ways to derive transformations without using the general formula. This involves manipulating the CDF.

   (a) Let X be a uniform random variable on $[-1, 1]$. Let $Y = X^2$. Find the pdf of Y.

   Solution: Note that $g(X) = X^2$ is not monotonic on the support. So we need a different method.

   Clearly the support of Y will be $[0, 1]$. The CDF of X is $F_X(x) = \frac{x+1}{2}$. Consider the

CDF of Y. Let $y \in [0,1]$ be in the support. Then the CDF is

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(Y \leq y\right) \\ &= \mathbb{P}\left(X^2 \leq y\right) \\ &= \mathbb{P}\left(X \in [-\sqrt{y}, \sqrt{y}]\right) \\ &= \mathbb{P}\left(X \leq \sqrt{y}\right) - \mathbb{P}\left(X \leq -\sqrt{y}\right) \\ &= \frac{\sqrt{y}+1}{2} - \frac{-\sqrt{y}+1}{2} \\ &= \sqrt{y} \end{aligned}$$

Thus the CDF is $F_Y(y) = \sqrt{y}$ for $y \in [0,1]$ and so, since the random variable is continuous, the PDF can be found by differentiating giving

$$f_Y(y) = \frac{1}{2\sqrt{y}}, y \in [0,1]$$

(b) Let X have pdf $\frac{1}{2}e^{-|x|}$, $x \in \mathbb{R}$. Let $Y = |X|^3$. Find the pdf of Y.

Solution: We will use a similar method to above.

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(|X^3| \leq y\right) \\ &= \mathbb{P}\left(X \in [-y^{1/3}, y^{1/3}]\right) \\ &= \mathbb{P}\left(X \leq y^{1/3}\right) - \mathbb{P}\left(X < -y^{1/3}\right) \\ &= \mathbb{P}\left(X \leq y^{1/3}\right) - \mathbb{P}\left(X \leq -y^{1/3}\right), \text{ since } \mathbb{P}\left(X = -y^{1/3}\right) = 0 \\ &= F_X(y^{1/3}) - F_X(-y^{1/3}) \\ &= \frac{1}{2}\Big(\int_{-\infty}^{y^{1/3}} e^{-|x|}dx - \int_{-\infty}^{-y^{1/3}} e^{-|x|}dx\Big) \\ &= \frac{1}{2}\Big(\int_{0}^{y^{1/3}} e^{-x}dx - \int_{-y^{1/3}}^{0} e^{x}dx\Big) \\ &= \frac{1}{2}\Big(1 - e^{-y^{1/3}} + 1 - e^{-y^{1/3}}\Big) \\ &= 1 - e^{-y^{1/3}} \end{aligned}$$

Since this is also continuous, we can get the pdf by differentiating.

$$\frac{dF_Y(y)}{dy} = \frac{1}{3}y^{-2/3}e^{-y^{1/3}}$$

and so

$$f_Y(y) = \begin{cases} \frac{1}{3}y^{-2/3}e^{-y^{1/3}}, & y \in (0, +\infty) \\ 0, & \text{o.w.} \end{cases}$$

2. (Expectations 1: Linearity of Expectation) You learned in class that $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$. The reason is because the integral is linear. Another very useful property of random variables is that for *any* random variables X and Y, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. This is true regardless of their joint distribution!

   Why is this true? Consider (X, Y) with joint distribution $f_{XY}(x, y)$. Then

$$\begin{aligned} \mathbb{E}[X + Y] &= \int\int (x + y) f_{XY}(x, y) dx dy \\ &= \int\int x f_{XY}(x, y) dx dy + \int\int y f_{XY}(x, y) dx dy \\ &\qquad \text{using linearity of the integral} \\ &= \int\int x f_{XY}(x, y) dy dx + + \int\int y f_{XY}(x, y) dx dy \\ &\qquad \text{swapping the integration orders} \\ &= \int x f_X(x) dx + \int y f_Y(y) dy \\ &\qquad \text{since integrating the joint density yields the marginal density} \\ &= \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

   The key step is the *linearity of the integral* and understanding that integrating out a joint distribution leads to a marginal distribution. The same thing works for sums so we can use this for discrete sums, too. It is easy to see how we could extend this to multiple random variables $X_1, \ldots, X_n$ with any relationship.

3. (Expectations 2: Law of Total Expectation) It is often useful to break down a random variable into a partition of the sample space to compute its expectation. The formula is that for any (finite or countable) $\{A_i\}_i$ partition of the sample space,

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X|A_i]\mathbb{P}(A_i)$$

   Here is an example: Suppose that Y has the following form: with probability p I draw a normal random variable with mean μ. With probability $1 - p$ I then flip a coin. If the coin is heads I draw from uniform $[0, a]$. If the coin is tails I set the variable equal to 0. What is

$\mathbb{E}[Y]$?

The distribution is a nightmare but expectation is simple. There are essentially three events $A_1 =$ draw a normal, $A_2 =$ draw a uniform random variable and $A_3 =$ set the variable to 1. The expectations in these cases can be calculated by my description (more on conditional expectations later)

$$\mathbb{E}[Y|A_1] = \mu, \mathbb{E}[Y|A_2] = a/2, \mathbb{E}[Y|A_3] = 1$$

Then the only thing left to do is to calculate the probabilities. It is not hard to show that $\mathbb{P}(A_1) = p$, $\mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1-p}{2}$. Then the expectation is

$$\mathbb{E}[Y] = \mu p + \frac{1-p}{2}(\frac{a}{2} + 1)$$

4. (MGF Practice: Binomial and Poisson) There is a useful relationship Binomial and Poisson variables. In general, binomial probabilities are hard to compute but Poisson probabilities are much easier. A Binomial random variable can be approximated by a Poisson random variable when n is large and np is small. The appropriate Poisson parameter is $\lambda = np$.

   One can show this approximation using MGFs. Recall that MGF equivalence is "essentially" the same as distribution equivalence (modulo some technicalities). Recall the MGF of the binomial distribution with parameters n and p:

   $$M_X(t) = [p\exp(t) + (1-p)]^n$$

   Let $\lambda = np$ be the candidate Poisson parameter. Then we can write this as

   $$M_X(t) = [(\lambda/n)\exp(t) + 1 - (\lambda/n)]^n = \left[1 + \frac{\lambda}{n}(\exp(t) - 1)\right]^n$$

   Now a classic result from analysis states that $(1 + \frac{y}{n})^n \to_n \exp(y)$. So if we take $n \to \infty$ in the MGF we get

   $$\lim_n M_X(t) = \exp\{\lambda(e^t - 1)\}$$

   which is in fact the MGF of the Poisson with parameter λ! It is a useful exercise to calculate the Poisson MGF by hand.

   This is a common use of an MGF - showing convergence in distribution because it characterizes a distribution.

5. (Relationship between Bernoulli and Binomial) There is an important connection between Bernoulli and Binnomial random variables. In particular, the following is true: If $X_1, \ldots, X_n$

are independent Bernoulli random variables with parameter p, then $X_1 + \cdots + X_n$ is a Binomial distribution with parameters (n, p).

First let's show the distribution equivalence. We will use the MGF again. The Bernoulli MGF is

$$M_{X_i}(t) = 1 - p + p\exp(t)$$

Then letting $Y = \sum_i X_i$ we see that the MGF of Y is

$$\begin{aligned}
\mathbb{E}[\exp\{tY\}] &= \mathbb{E}[\exp\{tX_1 + \cdots tX_n\}] \\
&= \mathbb{E}[\exp\{tX_1\} \cdots \exp\{tX_n\}] \\
&= \mathbb{E}[\exp\{tX_1\}] \cdots \mathbb{E}[\exp\{tX_n\}] \\
&\qquad \text{using independence} \\
&= \prod_i M_{X_i}(t) \\
&= (1 - p + p\exp(t))^n
\end{aligned}$$

which is the MGF fo the binomial. The only possibly unfamiliar thing we used here is that the expectation of the product of two independent random variables is the product of their expectations. We will look at this more generally and closely when we cover joint random variables. Note also though that in our proof, we showed in general that the MGF of the sum of independent random variables is the product of the MGFs.

The more important thing, in my opinion, is that this allows us to calculate the expectation and variance easily. Recall linearity of expectation. There is also linearity of the variance operator when the random variables are *independent*. Thus if Y is Binomial with parameters (n, p) then

$$\mathbb{E}[Y] = \mathbb{E}[X_1 + \cdots + X_n] = np$$

and

$$\text{Var}(Y) = \text{Var}(X_1 + \cdots + X_n) = np(1 - p)$$

and this is a useful way to memorize these properties.

Section 3 - Multiple Random Variables, Inequalities, Conditional Expectations

1. (Exponential Family Practice: Multinomial Distribution) Recall the definition of an exponential family by pdfs

$$f(x|\theta) = h(x)c(\theta)\exp\{\sum_{i=1}^{K}\omega_i(\theta)t_i(x)\}$$

First, we will practice defining this. Consider the *multinomial distribution*. This distribution generalizes the binomial distribution to multiple outcomes. It describes the probability of different outcomes occurring over many independent and identical trials. For example, can model number of different types of role of a die. A common application in economics would be to model the number of times consumers or firms make a certain purchase or strategic decision assuming that decision errors or taste shocks are independent and identical across time.

The multinomial distribution is defined by three parameters (n, k, p) where n is the number of trials (like binomial) k is the number of categories, and p is a k-probability vector specifying the probabilities of each trial so that $\sum_{j=1}^{k} p_j = 1$. Let the outcomes be y_j for category j. The pdf is

$$\mathbb{P}\left(y_1 = x_1, \ldots, y_k = x_k\right) = \frac{n!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}$$

if $\sum_j x_j = n$ (0 otherwise). (Remember that $0! = 1$.)

How to express in exponential form. It is not hard to see that we should separate out the two parts. First letting $h(x) = \frac{1}{x_1! \cdots x_n!}$ and $c(\theta) = n!$ we are already close. Next note that $p_j^{x_j} = \exp\{x_j \log p_j\}$ and so we can write the second part as

$$\prod_j \exp\{x_j \log p_j\} = \exp\{\sum_j x_j \log p_j\}$$

and so letting $K = k$ and $x_j = t_j(x)$ and $\omega_j(\theta) = \log p_j$ gives us the exponential parameterization.

A good extra exercise: express this pdf as a function of Γ functions.

2. (Exponential Family Practice: Means) Here we will derive general properties of some functions of means of exponential families and apply them. See Casella-Berger p. 112-113 for more information.

Suppose that θ is one dimensional. I claim that

$$\mathbb{E}[\sum_i \omega_i'(\theta)t_i(X)] = -\frac{c'(\theta)}{c(\theta)}$$

How do we get this? Consider differentiating the density function with respect to θ. First rewrite the density as

$$f(x|\theta) = h(x)\exp\{c^*(\theta) + \sum_i \omega_i(\theta)t_i(x)\}$$

This gives

$$\partial f/\partial\theta = f(x|\theta)\Big(c^{*'}(\theta) + \sum_i \omega_i'(\theta)t_i(x)\Big)$$

Then note that

$$\mathbb{E}[\partial f/\partial\theta] = \partial\mathbb{E}[f]/\partial\theta = 0$$

if we can pass the derivative through the integral using Leibniz rule. Thus, integrating over the equation gives

$$\int \Big(c^{*'}(\theta) + \sum_i \omega_i'(\theta)t_i(x)\Big)f(x|\theta)dx = \mathbb{E}[c^{*'}(\theta)] + \mathbb{E}[\sum_i \omega_i'(\theta)t_i(x)]$$

and then noting that $c^*(\theta) = \log(c(\theta))$ so that the derivative is $\frac{c'(\theta)}{c(\theta)}$ gives us the result.

Why is the useful? An application: finding the mean and variance of Poisson random variables. Suppose $X \sim$ Pois(λ). Find the mean using the above formula. Note that the pdf is

$$f(x|\theta) = \frac{\lambda^x e^{-\lambda}}{x!}$$

and we can express in exponential form as

$$f(x|\theta) = \frac{1}{x!}\exp\{x\log(\lambda) - \lambda\}$$

so that $c(\theta) = 1, \omega_1(\theta) = \log(\lambda), \omega_2(\theta) = \lambda$ and $t_1(x) = x$ and $t_2(x) = -1$. Then applying the result and noting that $c'(\theta) = 0$ we get that

$$0 = \mathbb{E}[\frac{1}{\lambda}x - 1]$$

which gives us that

$$\mathbb{E}[x] = \lambda.$$

3. (Multiple Random Variable Practice) Suppose $f_{XY}(x,y) = K,\ x \in [0,1]$ and $y \in [0,x]$.

   (a) Find K

$$\begin{aligned}
\int_0^1 \int_0^x K dy dx &= 1 \\
\int_0^1 K \cdot x dx &= 1 \\
\frac{x^2}{2} \cdot K|_0^1 &= 1 \\
K &= 2
\end{aligned}$$

   (b) Plot the support of the joint distribution: Triangle

   (c) What's $E[Y|X]$?

   Next, first find f_X and f_Y.

$$\begin{aligned}
f_X &= \int_0^x 2dy \\
&= 2x, x \in [0,1]
\end{aligned}$$

   Verify: $\int_0^1 2x dx = 1$, ok.

$$\begin{aligned}
f_{Y|X} &= f_{XY}/f_X \\
&= 2/2x \\
&= 1/x, y \in [0,x]
\end{aligned}$$

$$\begin{aligned}
E[Y|X] &= \int_0^x y(1/x)dy \\
&= y^2/2x|_0^x \\
&= x/2
\end{aligned}$$

   (d) Recall that conditional expectations are random variables. What is the expectation of $\mathbb{E}[Y|X]$?

$$\mathbb{E}[\mathbb{E}[Y|X]] = \int_0^1 \frac{x}{2} dx = \frac{1}{4}$$

(e) What's $E[X|Y]$?

$$\begin{aligned} f_Y &= \int_y^1 2dx \\ &= 2(1-y),\ 0<y<1 \end{aligned}$$

$$\begin{aligned} f_{X|Y} &= 2/2(1-y) \\ &= 1/(1-y), x \in [y,1] \end{aligned}$$

$$\begin{aligned} E[X|Y] &= \int_y^1 x/(1-y)dx \\ &= \frac{1-y^2}{2(1-y)} \\ &= \frac{(1+y)(1-y)}{2(1-y)} \\ &= (1+y)/2 \end{aligned}$$

4. (Conditional Expectation Practice) Prove that if X and Y are independent then $\mathbb{E}[Y|X] = \mathbb{E}[Y]$.

   First recall conditional expectation definition:

$$\mathbb{E}[Y|X] = \int y f_{Y|X}(y|x)dy$$

   Recall what independence means in joint densities. It means that $f_{XY}(x,y) = f_X(x)f_Y(y)$ (if and only if!). Then can write conditional density as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = f_Y(y)$$

   and so the conditional expectation is

$$\mathbb{E}[Y|X] = \int y f_Y(y)dy = \mathbb{E}[Y].$$

Section 4 - Order Statistics, Method of Moments and MLE

1. (Uniform Order Statistics Distribution) Recall the formula for the CDF of order statistic k from class is:

$$F_{X_{(k)}}(x) = \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j}$$

where we use the binomial formula basically to compute the overall probabilities of the n variables falling in this order.

What does this look like for a (standard) uniform distribution? For $x \in [0, 1]$: $F(x) = x$ and so

$$F_{X_{(k)}}(x) = \sum_{j=k}^{n} \binom{n}{j} x^j (1 - x)^{n-j}$$

Then we can find the pdf by differentiating:

$$\begin{aligned}
f_{X_{(k)}}(x) &= \sum_{j=k}^{n} \frac{n!}{j!(n-j)!} (j x^{j-1}(1-x)^{n-j} + (n-j)x^j(1-x)^{n-j-1}) \\
&= n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k} - n\binom{n-1}{k} x^k (1-x)^{n-k-1} \\
&+ n\binom{n-1}{k} x^k (1-x)^{n-k-1} - n\binom{n-1}{k+1} x^{k+1}(1-x)^{n-k-2} \\
&+ \cdots - n\binom{n-1}{n-1} x^{n-1} + n x^{n-1} - 0 \\
&= n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k}
\end{aligned}$$

by deleting all the adjacent terms after the first term and thus we can use this telescoping sum term to derive the pdf.

Does this look familiar? There is a distribution called the *Beta Distribution* which has parameters (α, β) and has pdf

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}$$

recalling that $\Gamma(\alpha) = (\alpha - 1)!$ when α is an integer. Then note that

$$n\binom{n-1}{k-1} = \frac{n!}{(k-1)!(n-k)!} = \frac{\Gamma(k + n - k + 1)}{\Gamma(k)\Gamma(n - k + 1)}$$

and then we see that

$$X_{(k)} \sim \text{Beta}(k, n-k+1)$$

Thus, if we know some properties of the Beta distribution we can use these to understand properties of uniform order statistics. If $X \sim \text{Beta}(\alpha, \beta)$ then $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$. Let's use this to compute the expectation of the maximum and minimum of N iid standard uniform random variables. They are order stats $X_{(1)}$ and $X_{(N)}$ respectively. Then $X_{(1)} \sim \text{Beta}(1, N)$ and $X_{(N)} \sim \text{Beta}(N, 1)$ and so $\mathbb{E}[X_{(1)}] = \frac{1}{N+1}$ and $\mathbb{E}[X_{(N)}] = \frac{N}{N+1}$.

2. (MoM Example: Uniform with unknown endpoints) The MoM estimator essentially says give me a number of parameters, give me some moments, and let me solve for the parameters given the moments. Then we can use the analogy principle to form the estimator.

   Consider the following example: we have a random sample from a variable distributed as uniform $U[a, b]$ where both a and b are unknown. How do we get the MoM estimator? As suggested by the idea, we should compute some moments. Let's compute the first and second moment of the random variable. First note that if $X \sim U[a, b]$ then

$$\mathbb{E}[X] = \frac{a+b}{2}$$

   and

$$\mathbb{E}[X^2] = \int_a^b \frac{x^2}{b-a} dx = \frac{1}{3}\frac{b^3-a^3}{b-a} = \frac{a^2+ba+b^2}{3}$$

   Thus we have two moments. Using the analogy principle we put in the sample analogues of these to get a system

$$\bar{X} = \frac{a+b}{2}$$
$$\bar{X^2} = \frac{a^2+ba+b^2}{3}$$

   This has solution

$$\hat{a} = \bar{X} - \sqrt{\bar{X}^2 + 2\bar{X} - 3\bar{X^2}}$$
$$\hat{b} = \bar{X} + \sqrt{\bar{X}^2 + 2\bar{X} - 3\bar{X^2}}$$

   which gives us a MoM estimator.

   While using the first K moments for K parameters is generally the standard "MoM estimator" your estimator does not have to be based on standard moments. You'll more about this when you explore simulated method of moments and GMM.

3. (MLE Example 1: Normal unknown mean, known variance) MLEs are usually straightforward: they are maximization problems and so our calculus tools will usually come in handy. Suppose that we have a random sample of size N of normal variables with unknown mean μ and known variance σ^2. Then the PDF for one draw is

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

So we can derive the likelihood as the product of the N draws as

$$f(X_1, \ldots, X_N) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(X_i-\mu)^2}{2\sigma^2}\} = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\{-\sum_{i=1}^{N} \frac{(X_i-\mu)^2}{2\sigma^2}\}$$

which is also the likelihood function of the parameter μ given the data:

$$L(\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\{-\sum_{i=1}^{N} \frac{(X_i-\mu)^2}{2\sigma^2}\}$$

In almost all MLE situations it is easier to work with the log-likelihood. In this case the log-likelihood is

$$l(\mu) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(X_i-\mu)^2}{2\sigma^2}$$

To maximize this function set the derivative with respect to μ to 0:

$$\frac{1}{\sigma^2}\sum_i (X_i - \mu) = 0$$

and this gives

$$\mu = \frac{1}{N}\sum_i X_i = \bar{X}$$

and so the proposed MLE is $\hat{\mu}_{\text{MLE}} = \bar{X}$. How do we know? We can check the SOC: the second derivative of l is

$$-\frac{N}{\sigma^2} < 0$$

as long as $\sigma > 0$. Thus the function is strictly concave so the FOC is sufficient since the function is smooth over the real numbers.

How would this change with σ^2 unknown? To maximize the log-likelihood we would also need to maximize over σ^2 and so we would need that partial as well and solve for the maximizer using multivariate methods. This is a good exercise to practice the important opti-

mization techniques in statistics.

4. (MLE Example 2: Uniform with 1 unknown endpoint) Consider a random sample of size N from a uniform distribution $U[0, \theta]$ where θ is unknown. Let's find the MLE for θ.

   First the pdf is
   $$f(X_i) = \frac{1}{\theta}\mathbf{1}\{X_i \in [0, \theta]\}.$$
   Then the likelihood is
   $$L(\theta) = \prod_i \frac{1}{\theta}\mathbf{1}\{X_i \in [0, \theta]\}$$
   We will not go to log-likelihood now. Ignore the indicators and suppose we were a bit more sloppy. Then we would write this as
   $$L(\theta) = \frac{1}{\theta^N}$$
   Then it seems to maximize this we should set θ as small as possible to maximize this function - i.e. $\theta \to 0$.

   This argument is wrong because we have the indicators. In particular, suppose that we have $X_1 = 1, X_2 = 2$. Then the log likelihood for choosing $\theta = \epsilon$ very small gives
   $$L(\epsilon) = 0$$
   since both $X_1, X_2 > \epsilon$. Instead suppose we choose $\theta = 1$. Then the first indicator is 1 but the second is 0 so
   $$L(1) = (1)(0) = 0$$
   If we choose $\theta = 2$ then we get that
   $$L(2) = (\frac{1}{2})^2 = 1/4$$
   using this we should be able to compute the MLE generalizing to N data points in general.

Section 5 - Evaluating Estimators

1. (MSE Example) Recall that MSE is defined as Variance + Bias^2 where these things depend on the true parameter value (i.e. are taken with respect to the true parameter value). It is a function of the true parameter.

Consider the case where we have a random sample from a normal distribution with parameters (μ, σ^2) where both μ and σ^2 are unknown. Focus on estimating σ^2 for now. There are two candidate estimators that we are considering: one is the MLE

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_i (X_i - \bar{X})^2$$

and the other is the "biased-corrected" MLE

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

Let's compare these estimators by the MSE criteria. As motivated above, we have that

$$\mathbb{E}[\hat{\sigma}^2_{MLE}] = \frac{n-1}{n}\sigma^2$$

and

$$\mathbb{E}[S^2] = \sigma^2$$

Now we compute the variances. To do this computation note that $\mathbb{E}[\bar{X}] = \mu$. As well note that

$$\frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

a chi-squared r.v. with $n-1$ degrees of freedom. The variance of a χ^2 distribution as a function of the d.o.f. is $2n$ so the variance is $2(n-1)$. From this we can easily derive that

$$\text{Var}(\hat{\sigma}^2_{MLE}) = \frac{2(n-1)\sigma^4}{n^2}$$

and

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

using properties of the variance operator. Note that all these expectations and variances are *functions of the true parameter* σ^2.

Then the MSEs are repsectively

$$\begin{aligned}
\text{MSE}(\hat{\sigma}^2_{MLE}) &= (\mathbb{E}[\hat{\sigma}^2_{MLE}] - \sigma^2)^2 + \text{Var}(\hat{\sigma}^2_{MLE}) \\
&= \frac{\sigma^4}{n^2} + \frac{2(n-1)\sigma^4}{n^2} \\
&= \frac{(2n-1)\sigma^4}{n^2}
\end{aligned}$$

and

$$\begin{aligned}\text{MSE}(S^2) &= (\mathbb{E}[S^2] - \sigma^2)^2 + \text{Var}(S^2) \\ &= 0 + \frac{2\sigma^4}{n-1} \\ &= \frac{2\sigma^4}{n-1}\end{aligned}$$

Now note that

$$\frac{2n-1}{n^2} < \frac{2n}{(n-1)n} < \frac{2}{n-1}$$

and so the MSE of the MLE is strictly smaller *for all true values* of the parameters.

2. (Sufficient Statistic and Rao-Blackwellization Example) Recall the definition of a sufficient statistic: it is a statistic $T(X)$ s.t.

$$f(X|\theta, T) = f(X|T)$$

It is "sufficient" because once we include it, there is not extra information about θ contained in the distribution.

It is useful because we can improve unbiased estimators using the Rao-Blackwell theorem by conditioning on the sufficient statistic and improving it. In particular, Rao-Blackwellization (1) does not change the mean (it was unbiased so that is good) and (2) lowers the variance. Both of these make it at least as good an estimator in the MSE sense.

Here is an example: Let $X_1, \ldots, X_n \sim \text{Pois}(\lambda)$ where we want to estimate λ. First, I claim that $\sum_i X_i$ is a sufficient statistic for λ. This can be shown with the exponential family results we learned in class but I also want to show it using the Factorization Theorem. In general, I find the useful ways to prove sufficient are through Theorems 6.2.2 and 6.2.6 (Factorization Theorem) in Casella-Berger. They basically say the same thing: we can split the density appropriately. In this case, the joint pdf of the draws is

$$\begin{aligned}\prod_i f(X_i|\lambda) = \prod_i \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} &= \lambda^{\sum_i X_i} e^{-n\lambda} \frac{1}{\prod_i X_i!} \\ &= g(\sum_i X_i|\lambda) h(X_1, \ldots, X_n)\end{aligned}$$

and so by factorization the sum of the values this is a sufficient statistic.

Next, I show an example of Rao-Blackwellization. Consider an unbiased estimate, $\hat{\lambda}_1 = X_1$ the first draw. This is unbiased because $\mathbb{E}[X_1] = \lambda$. Now we will RB this.

To do this, we need to take the expectation X_1 with respect to the sum to get the estimator:

$$\hat{\lambda}_{RB} = \mathbb{E}[X_1 | \sum_i X_i = t]$$

To compute this consider

$$\mathbb{E}[\sum_i X_i | \sum_i X_i = t] = t$$

and expanding the left side is

$$\sum_i \mathbb{E}[X_i | \sum_i X_i = t] = t$$

and then using the fact that each term inside is identically distributed thus must each be the same and so we get that $\mathbb{E}[X_i | \sum_i X_i = t] = t/n$ and so

$$\hat{\lambda}_{RB} = \bar{X}$$

3. (Complete Statistic Example) A statistic is complete if for any (measurable) function g, if $\mathbb{E}_\theta[g(T)] = 0$ for all $\theta \in \Theta$ then $\mathbb{P}_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$.

   In the exponential family

   $$f(x|\theta) = h(x)c(\theta)\exp\{\sum_{i=1}^{K} \omega_i(\theta)t_i(x)\}$$

   we have that $T = \left(\sum_{i=1}^n t_1(X_i), \ldots, \sum_{i=1}^n t_K(X_i)\right)$ is sufficient for θ and is complete if $\{(\omega_1(\theta), \ldots, \omega_K(\theta)) : \theta \in \Theta)\}$ contains an open set in $\mathbb{R}^K$.

   I will now show that this open set property is important. Consider a normal distribution with parameters (μ, σ^2). In this case, we have that $t_1(x) = x$ and $t_2(x) = x^2$. If we allow

   $$\Theta_0 = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

   then Θ_0 contains an open set in $\mathbb{R}^2$; for example the open unit circle centered at $(10, 10)$.

   However, if we restrict the parameter space we may not get this open set condition. For example if we set

   $$\Theta_1 = \{(\mu, \sigma^2) : \mu^2 = \sigma^2, \mu > 0\}$$

this does not have an open set in $\mathbb{R}^2$ (it is a parabola). Now consider

$$g(T) = 2(\sum_i X_i)^2 - (n+1)\sum_i X_i^2$$

This is a function of the sufficient statistics. Note that $\mathbb{E}[X_i^2] = 2\mu^2$. Also note that

$$\begin{aligned}
\mathbb{E}[g(T)] &= 2\mathbb{E}[(\sum_i X_i)^2] - (n+1)\mathbb{E}[\sum_i X_i^2] \\
&= 2\Big(\sum_i \sum_j \mathbb{E}[X_i X_j]\Big) - 2(n+1)n\mu^2 \\
&= 2\Big(2n\mu^2 + (n^2-n)\mu^2) - 2(n+1)n\mu^2 \\
&\qquad \text{by counting terms} \\
&= 0
\end{aligned}$$

Now we just need for $g(T)$ to not always be 0. This is true: $g(T)$ is not generically 0. For example take an interval around 1 for each X_i for $i = 1, \ldots, n$. This has positive probability. For $X_i = 1$ we get

$$g(T) = 2n^2 - (n+1)n = n^2 - n \neq 0$$

4. (CRLB Uniform Example) Suppose that $X_1, \ldots, X_n \sim U[0, \theta]$. Then we learned in class that $\frac{1+n}{n} X_{(n)}$ is the UMVU. What is its variance? It is clearly equal to

$$\text{Var}(\frac{1+n}{n} X_{(n)}) = \frac{(n+1)^2}{n^2}\text{Var}(X_{(n)})$$

and the computation for finding $\text{Var}(X_{(n)})$ simply uses the form of the density: $f(x|\theta) = nx^{n-1}/\theta^n$. Then

$$\mathbb{E}[X_{(n)}] = \int_0^\theta nx^n/\theta^n dx = \frac{n}{n+1}\theta$$

and

$$\mathbb{E}[X_{(n)}^2] = \int_0^\theta nx^{n+1}/\theta^n dx = \frac{n}{n+2}\theta^2$$

and so

$$\text{Var}(X_{(n)}) = \frac{n}{(n+1)^2(n+2)}\theta^2$$

and so

$$\text{Var}(\frac{1+n}{n} X_{(n)}) = \frac{\theta^2}{n(n+2)}$$

What is the CRLB? For unbiased estimators it is

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n\mathbb{E}[(\frac{\partial}{\partial\theta}\log f(X_i|\theta))^2]}$$

Here

$$\log f(X_i|\theta) = -\log(\theta)\mathbf{1}\{X_i \in [0,\theta]\}$$

and so if $X_i < \theta$ this is the derivative is

$$1/\theta$$

so that the CRLB is

$$\theta^2/n$$

But

$$\frac{\theta^2}{n(n+2)} \leq \frac{\theta^2}{n}$$

so clearly it does not apply. What goes wrong? The condition about differentiating under the integral. Because the support of the distribution depends on the parameter, we are unable to differentiate under the integral.

Section 6 - Midterm Review and Hypothesis Testing Introduction

1. (Midterm Review) Any questions/concerns on the midterm? Let's go over the last question because it is a bit challenging.

2. (Hypothesis Testing Power: Binomial Example) Suppose that our null is $H_0 : \theta \in \Theta_0$ which makes the alternative $H_1 : \theta \in \Theta_1$ where $\Theta_1 = \Theta \setminus \Theta_0$. Our testing procedures tells us which hypothesis to choose. In particular as a function of the data X, we specify a region R s.t. we reject H_0 if $X \in R$.

   Then we can draw the table for decision making in hypothesis testing.

   Mathematically, type 1 error is $\mathbb{P}_\theta(X \in R)$ if $\theta \in \Theta_0$ and type 2 error is $1 - \mathbb{P}_\theta(X \in R)$ if $\theta \in \Theta_1$.

   Then we define a test's power as a function of θ: $\beta(\theta) = \mathbb{P}_\theta(X \in R)$.

   What is the ideal power function? 1 when $\theta \in \Theta_1$ and 0 when $\theta \in \Theta_0$. So qualitatively, want the power functions close to 1 when $\theta \in \Theta_1$ and close to 0 when $\theta \in \Theta_0$.

For example, suppose that $X \sim \text{Bin}(5, \theta)$ and we want to test whether $\theta \leq 0.5$. Then X is a vector of length 5 of 0's and 1's.

We will consider a few different tests. First suppose that $R = \{0, 1\}^5$. Then I always reject the null hypothesis. Thus

$$\beta_0(\theta) = 1$$

and this doesn't really satisfy the good properties of power intuition. It also seems like a bad way to test.

Now consider a test in which we reject if all the outcomes are successes. Then

$$\beta_1(\theta) = \theta^5$$

which does satisfy some of the requirements - it is close to 0 when $\theta \leq 1/2$ and closer to 1 when $\theta > 1/2$. Note that the Type 1 error probability is extremely low $(1/2)^5 < 0.05$ but the Type 2 error probability is still quite high: at $\theta = 3/4$ the probability of a Type 2 error is $1 - (3/4)^5 > 0.75$ which is very high.

Finally consider rejecting if $X \in \{3, 4, 5\}$. Then the power function is

$$\beta_2(\theta) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \theta^5$$

This will have a smaller Type 2 error but the Type 1 error will be much higher. Thus there is a balance that we want to strike.

3. (Hypothesis Testing Power: Normal Power Calculation Example) The power will typically depend on the sample size n and so experimenters and researchers can select n to have appropriate power properties.

   Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known and we want to test $H_0 : \theta \leq \theta_0$. A relatively standard test (one that will be derived in class) will be to reject H_0 if

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c$$

Thus we have that the power function is

$$\begin{aligned}\beta(\theta) &= \mathbb{P}_\theta\left(\frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}} > c\right) \\ &= \mathbb{P}_\theta\left(\frac{\bar{X}-\theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}_\theta\left(Z > c + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right)\end{aligned}$$

where Z is a standard normal random variable.

Suppose that we want to have a maximum Type 1 Error probability of 0.1 and we would like the Type 2 error probability to be 0.2 if $\theta \geq \theta_0 + \sigma$. Since $\beta(\theta)$ is increasing this will be true if

$$\beta(\theta_0) = 0.1 \text{ and } \beta(\theta_0 + \sigma) = 0.8$$

Now by choosing $c = 1.28$ we get $\mathbb{P}(Z > 1.28) = 0.1$ independent of n. So the binding constraint for n is

$$\beta(\theta_0 + \sigma) = \mathbb{P}\left(Z > 1.28 - \sqrt{n}\right) = 0.8$$

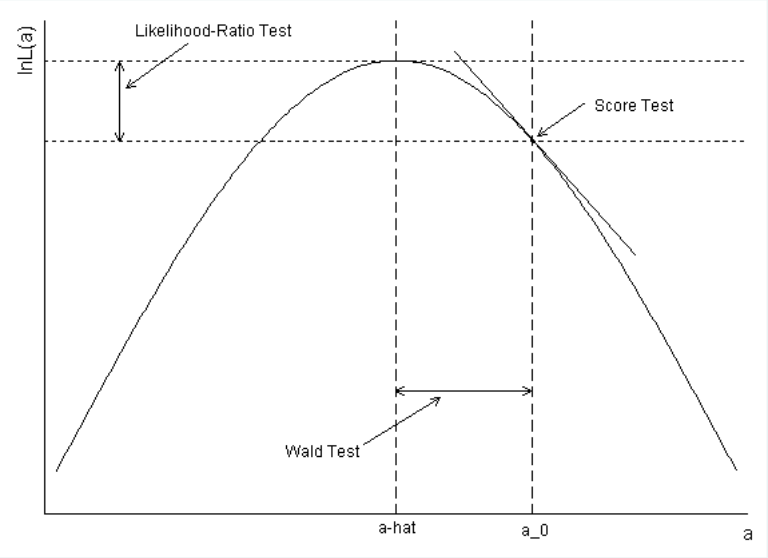
and using normal cdf computation tricks this shows that $n \geq 5$ must be chosen.

Section 7 - Hypothesis Testing

1. (Optimal Tests) Recall last time we talked about the power function $\beta(\theta) = \text{prob reject}(\theta)$. We say that a test has level α if $\beta(\theta) \leq \alpha, \forall\theta \in \Theta_0$. The size of a test is $\sup_{\theta\in\Theta_0}\beta(\theta)$. The Neyman approach to picking tests is to fix a class of tests with level α and the minimize the probability of type 2 errors (for all $\theta \in \Theta_1$; so minimize the max type 2 error). In particular, we pick the UMP once restricting the type 1 error in a class: a test in class C is UMP if $\beta(\theta) \geq \beta^*(\theta)$ for all $\theta \in \Theta_1$ for every β^* corresponding to a test in C.

   How do we find a UMP? In simple null and alternative cases it is not very difficult using the Neyman-Pearson Lemma: the test that rejects iff $f(X|\theta_1) > kf(X|\theta_0)$ is UMP level α test where $\alpha = \beta(\theta_0)$.

   Unfortunately, UMP is not that useful of a concept because in many cases it does not exist. In the simple cases (both null and alternative are simple hypotheses) we went over in class it does, but if the alternative hypothesis has the form $\theta \neq \theta_0$ one can show that the UMP does not exist within a class of level α tests. Thus we focus on more practical forms of tests.



To see the non-existence of a UMP in a case that we care about, look at Example 8.3.19 in Casella-Berger. Because of these technical problems, in the likelihood case, we focus on some other tests. In class we saw that when we have simple hypotheses the LRT yields the UMP.

2. (Trinity of Test Intro)

   This figure is a great way to understand these tests. Consider testing $H_0 : \theta = \theta_0$.

   Recall the definitions from class

   - LR: $T_{LR} = 2(l(\theta_{ML}) - l(\theta_0)$. Figure shows that this compares the value of the log-likelihood using the fact that MLE maximizes this.
   - LM/Score test: $T_{LM} = \frac{l'(\theta_0)^2}{-l''(\theta_0)}$. Figure shows that this compares derivative of log-likelihood at proposed value. Note: no estimation needed!
   - Wald: $T_W = \frac{(\theta_{ML}-\theta_0)^2}{(-l''(\theta_{ML}))^{-1}}$. This basically looks at distance standardized by variance of ML (recognize this formula?). Figure shows that this is comparing ML estimator and true value directly.

   Importantly showed in class that these tests are asymptotically equivalent using Taylor expansions. Their asymptotic distributions are all χ^2. We will talk more about asymptotic distributions in the coming weeks.

   If the results in finite samples differ, it may be a warning size that sample size is too small to apply asymptotic approximations or model is misspecified.

3. (Trinity of Test Examples)

Suppose that we observe $N = 20$ trials of Bernoulli draws with parameter p and we observe $x = 10$ successes. Our null is $H_0 : p = 0.29$. Let's compute the test statistics for each test and compare.

[might want to fudge around numbers]

To do this, we first need the likelihood function. In this case it will be

$$L(p|X) = p^x(1-p)^{N-x}$$

where $x = \sum_i X_i$ the number of observed succeses. Our tests will use the MLE so we derive it now. The log-likelihood is

$$l(p|X) = x\log(p) + (N-x)\log(1-p)$$

and maximizing this it is not hard to see that $\hat{p}_{ML} = \frac{x}{N} = 0.5$.

Knowing $N = 20$ and using a confidence level of 95% and using the chi-squared distribution with 1 degree of freedom, we get that the critical value we will reject at in general will be 3.84.

(a) (LRT) The test stat is $2(l(0.5) - l(0.29)) = 2(10\log(0.5) + 10\log(0.5) - 10\log(0.29) - 10\log(0.71)) = 3.88$. So we reject.

(b) (LMT) To find the test stat we need to find the derivatives. Here the first derivative is

$$l'(p|X) = \frac{x}{p} - \frac{N-x}{1-p}$$

and the second derivative is

$$l''(p|X) = -\frac{x}{p^2} - \frac{N-x}{(1-p)^2}$$

Then the test stat is

$$\frac{(\frac{10}{0.29} - \frac{10}{0.71})^2}{\frac{10}{0.29^2} + \frac{10}{0.71^2}} = 3.0$$

so we do not reject.

(c) (Wald) The test stat is

$$\frac{0.21^2}{(\frac{10}{0.5^2} + \frac{10}{0.5^2})^{-1}} = 3.528$$

so also do not reject.

Section 8 - Hypothesis Testing and Convergence Concepts

1. (UMP, Karlin-Rubin, Unbiased Tests) Recall last time we talked about how finding UMP tests are challenging in general unless hypotheses are simple. The Neyman-Pearson lemma helps up in simple cases.

   Similar to the NP Lemma, the Karlin-Rubin Theorem is helpful for certain kinds of tests. The theorem states:

   Theorem: When testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, if we have a sufficient statistic T for θ and the family of pdfs/pmfs $\{g(t|\theta) : \theta \in \Theta\}$ of T has a *monotone likelihood ratio* then the test $T > t_0$ can be made into a UMP level α test.

   Clearly we need to know what MLR is: $g(t|\theta_2)/g(t|\theta_1)$ monotone for every $\theta_2 > \theta_1$. A good exercise: when does an exponential family distribution have an MLR? Remember, that g is the pdf of the sufficient statistic!

   How about for two sided testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$? As pointed out before, in general UMP tests do not exist in the full class of tests. However, they do often exist in the class of **unbiased tests** where unbiasedness is defined as $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \inf_{\theta \in \Theta_1} \beta(\theta)$. We won't explore this theory too much but it's useful to know. Many of the tests that we use in economics and similar fields are UMP in the class of unbiased tests.

2. (Interval Estimation) We can use the tools of hypothesis testing, particularly how we choose to reject hypotheses, as a way to form interval estimators for parameters θ. An interval estimator is $[L(X), U(X)]$ which is a "random interval" (the bounds defining the region and shape are random variables). The most important concept is the **coverage probability** which is the probability that the interval covers the parameter. NOT the probability that the true parameter falls in the interval, because θ is just a scalar number.

   How do we form interval estimators? The general rule is *test inversion* which means we essentially "invert" optimal tests or tests that have good properties. We then can match the size of the test to the complementary coverage probability. Thus size α tests lead to $1 - \alpha$ coverage probability interval estimators. For more formal justification see CB Theorem 9.2.2.

   Let's do an example. Suppose that $X_1, \ldots, X_n$ is a random sample from a normal distribution with $\theta = \mu$ and σ^2 is known. We want to form a $(1 - \alpha)$ confidence interval for the test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Among unbiased tests, the UMP level α test is to reject the null if

   $$|\bar{X} - \theta_0| > z_\alpha \frac{\sigma}{\sqrt{n}}$$

Then we can write the "acceptance region" as

$$\{X : |\bar{X} - \theta_0| \leq z_\alpha \frac{\sigma}{\sqrt{n}}\}$$

and simplifying this out becomes

$$\{X : \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \theta_0 + \frac{\sigma}{\sqrt{n}}\}$$

To invert, we express in terms of θ:

$$\{\theta_0 \in \Theta : \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \theta_0 + \frac{\sigma}{\sqrt{n}}\}$$

which can be written as

$$\{\theta_0 \in \Theta : \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \theta_0 \leq \bar{X} + \frac{\sigma}{\sqrt{n}}\}$$

and so $L(X) = \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$ and $U(X) = \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$ summarize our interval estimator.

3. (Convergence Concepts) We moved onto asymptotic concepts this week which are very important in econometrics. The main concepts we learned about were

   - Convergence in Probability: $\lim \mathbb{P}(|X_n - X| \geq \epsilon) = 0$ for all $\epsilon > 0$.
   - Almost Sure Convergence: $\mathbb{P}(\lim_n |X_n - X| > \epsilon) = 0$ for all $\epsilon > 0$.
   - Convergence in Distribution: $\lim_n F_{X_n}(x) = F_X(x)$ (at all cty points)

   Adding some more for completeness:

   - Convergence in r-th mean (we did quadratic mean in class): $\lim_n \mathbb{E}[(X_n - X)^r] = 0$.

   What is their relationship?

$$\begin{array}{ccccc} \xrightarrow{L^s} & & \underset{s>r\geq 1}{\Rightarrow} & \xrightarrow{L^r} & \\ & & & \Downarrow & \\ \xrightarrow{a.s.} & \Rightarrow & \xrightarrow{p} & \Rightarrow & \xrightarrow{d} \end{array}$$

   Main tools: **Law of Large Numbers** (weak for cvg in prob; strong for as cvg); **Central Limit Theorem** (cvg in dist)

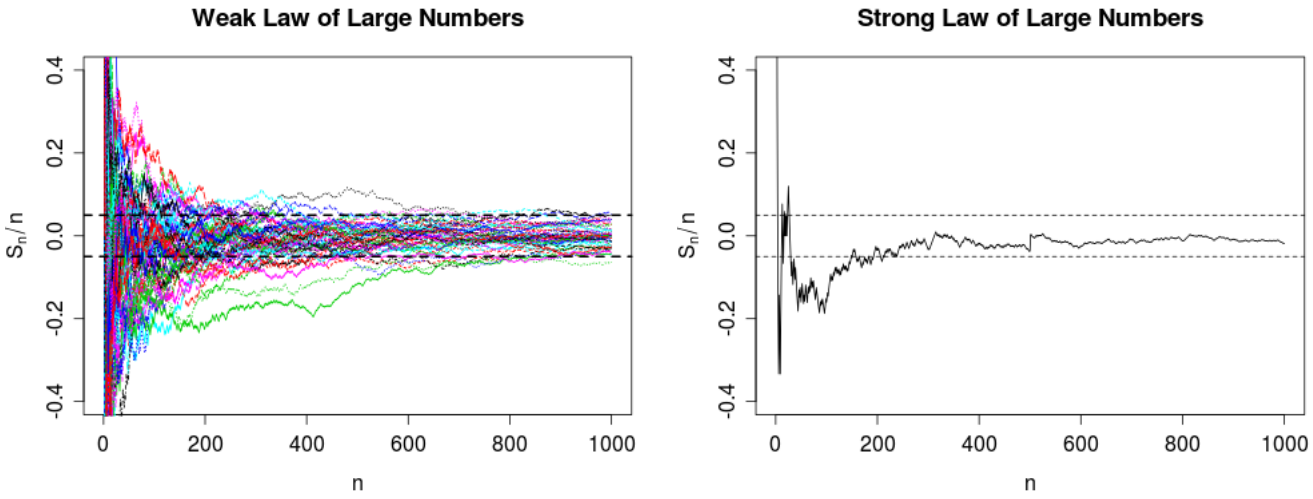


Figure 1: Source - Statistics Stack Exchange (Link)

4. (Cvg in Probability vs. Almost Sure Convergence) We did an example in class on the difference between these two concepts but I want to go over one more conceptual clarification because the equations do not provide much intuition.

   Consider a simple example $X_j \in \{-1, 1\}$ with equal probability and we look at $\bar{X}(n) = S_n/n$ which should converge to 0.

   WLLN: Pick some $\epsilon > 0$ small. The WLLN says that we can make the number of paths across samples of $\bar{X}(n)$ in the range $(-\epsilon, \epsilon)$ arbitrarily close to 1 by picking n large enough. It does not guarantee though that there will not be any paths that fall outside of $(-\epsilon, \epsilon)$ for n large enough.

   SLLN: Pick some $\epsilon > 0$ small. The SLLN says that there exists some N s.t. for all $n > N$ $\bar{X}(n) \in (-\epsilon, \epsilon)$ with probability 1. Thus if the SLLN holds here, the average will never fail for large enough n.

   See Figure for more details.

5. (Bias vs. Consistency) Consistency seems similar to bias but they are different. One is about finite samples and one is about asymptotics.

   For example, we learned that in the $U[0, \theta]$ case the MLE is $X_{(n)}$ the max of the sample, which is biased but is consistent. The reason is consistent is because under general conditions (which likely covered next quarter) the MLE is consistent.

   Suppose that $X_1, \ldots, X_n \sim N(\mu, 1)$. Another example of a biased but consistent estimator is $\frac{n-1}{n}\bar{X}$.

What about unbiased but not consistent? Consider in the same case the estimator of μ, X_1. Clearly not consistent since it has a positive probability of being away from the true value for all large n.

There is a way to understand the relationship between the two:

Result: An estimator will be consistent if it is asymptotically unbiased and its variance converges to 0.

Proof. We will use Chebychev. Note that

$$\mathbb{P}\left(|X_n - \theta| \geq \epsilon\right) = \mathbb{P}\left((X_n - \theta)^2 \geq \epsilon^2\right) \leq \frac{\mathbb{E}[(X_n - \theta)^2]}{\epsilon^2}$$

by Chebychev. Then expanding out the numerator yields

$$\frac{\text{Bias}^2(\theta, n) + \text{Var}(\theta, n)}{\epsilon^2}$$

since this is the classic MSE calcuation. And so if both terms go to 0 then we get converge in probability. □

Section 9 - Asymptotic Distributions and Bayesian Estimation

1. (Asymptotic MLE Distribution Example) Consider $X_1, \ldots, X_n \sim \text{Ber}(p)$. Then the MLE is $\hat{p} = \bar{X}$. Then by direct calculation

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

Suppose we wanted to find the asymptotic distribution of $\hat{p}$. There are two ways to do this. One easy way to do it is to look at

$$\sqrt{n}(\hat{p} - \mathbb{E}[\hat{p}]) \to_d N(0, n\text{Var}(\hat{p})) = N(0, p(1-p))$$

where $\mathbb{E}[\hat{p}] = p$. Thus, we have an asymptotic distribution.

The other way is to use the MLE property. We know that the variance is $I(p)^{-1}$ where $I(p)$ is the Fisher information. What is the Fisher Information?

$$I(p) = \mathbb{E}[-l''(p)]$$

Here, the log-likelihood for one data point is

$$l(p) = X_i \log(p) + (1 - X_i) \log(1 - p)$$

with

$$l''(p) = -\frac{X_i}{p^2} - \frac{1 - X_i}{(1 - p)^2}$$

and then

$$\mathbb{E}[-l''(p)] = \frac{p}{p^2} + \frac{1 - p}{(1 - p)^2} = \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)}$$

and so $I(p)^{-1} = p(1 - p)$ and thus by the MLE theorem

$$\sqrt{n}(\hat{p} - p) \to_d N(0, p(1 - p))$$

Note that in general, when we reach this variance, by the ***CRLB***, this is *asymptotically efficient*. So this is quite a good asymptotic estimator! This makes MLE quite attractive, especially in empirical analysis where we use and refer to asymptotic properties of our estimators.

2. (Asymptotic Hypothesis Testing Example) Recall the trinity of tests. In all cases, these tests will be the same. Theorem 10.3.3 in CB shows that the LRT statistic converges to a *Chi-squared distribution* with degrees of freedom equal to the number of free parameters in the null hypothesis. Similarly, the classic Wald and LM tests will have Chi-squared distributions as well. We went over some examples in lecture in class.

   What about other asymptotic tests based on asymptotic normal distributions? Clearly, for us to use asymptotic normal distributions, to compare to a non-parameter dependent value, we can use the standard normal distribution $N(0, 1)$.

   A common type of test is a *Wald test*, related to the Wald in the trinity of tests. This test has the form

$$Z_n = \frac{W_n - \theta_0}{S_n}$$

   where W_n is an estimator of θ_0 and S_n estimates the asymptotic variance of W_n. If this is the case and we can apply a CLT to W_n, and S_n converges in probability to the variance, using Slutsky's theorem we get that Z_n will be a standard normal variable.

   Consider testing $H_0 : p \leq p_0$. Consider the MLE $\hat{p}$. Then the standard deviation of this estimator is $\sigma_n = \sqrt{p(1 - p)/n}$. We cannot form an estimator based on p, the true value, so we will use $S_n = \sqrt{\hat{p}(1 - \hat{p})/n}$. Then it can be shown that $\sigma_n / S_n \to_p 1$. Thus, by utilizing

CLT and Slutsky's theorem we get that

$$W = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \to N(0,1)$$

Then we form and reject the null by replacing p with p_0 and reject W if $W > z_\alpha$ where z_α is the α quantile of the normal distribution.

3. (Bayesian Example) Bayesian estimation follows Bayes' Rule: for events A and B

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

The density analogy is

$$f_{X|Y} = \frac{f_{XY}}{f_Y} = \frac{f_{Y|X} f_X}{f_Y}$$

This gives us the posterior density, the main thing we use for estimation.

Given a $f_{Y|X}$ and f_X we can find the posterior distribution using these formulas. In particular, the methodology usually recognizes that f_Y is a constant and instead look at the numerator to find the *kernel* of the distribution. Another way to do Bayesian modeling in a convenient way is to use *conjugate priors*: families in which the posterior is in the same family as the prior.

Let's do an example: Gamma prior and Exponential likelihood. Suppose that $\theta \sim \text{Gamma}(\alpha, \beta)$ and $X_i|\theta \sim \text{Exp}(\theta)$ i.i.d for $i = 1, \ldots, n$.

Consider one observation. Then the kernel looks like

$$\begin{aligned} f(\theta) \propto f_{X_i|\theta}(X_i) f_\theta(\theta) &= \theta \exp\{-\theta X_i\} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp\{-\beta\theta\} \\ &\propto \theta^{\alpha+1-1} \exp\{-\theta(X_i + \beta)\} \end{aligned}$$

which is also a Gamma distribution! If we move to n observations we get

$$\begin{aligned} f(\theta) \propto \prod_i f_{X_i|\theta}(X_i) f_\theta(\theta) &= \prod_i (\theta \exp\{-\theta X_i\}) \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp\{-\beta\theta\} \\ &\theta^{\alpha+N-1} \exp\{-\theta(\sum_i X_i + \beta)\} \end{aligned}$$

which is a Gamma$(\alpha + N, \beta + \sum_i X_i)$.

4. (Numerical Methods for Bayesian Analysis) What do we do if we don't have conjugate priors? How can we draw from the posterior? There are some neat techniques developed to

deal with these cases. I will go over one: **Gibbs sampling**.

The idea of Gibbs Sampling is to build off of ideas from chains (Markov-Chain-Monte-Carlo Methods). The idea behind chains is to find $\theta_{k+1} \sim f(\theta|\theta_k, \text{data})$ in such a way that if $\theta_k \sim p(\theta|\text{data})$ then $\theta_{k+1} \sim p(\theta|\text{data})$. Then we need an initializer $\theta_0 \sim p(\theta|\text{data})$. In many cases, though, wherever we start, as $k \to \infty$, we will get convergence. So after some large K we will be drawing from the the posterior.

For Gibbs Sampling, instead of directly sampling the full vector, we split up the the vector of parameters θ into separate parts $\theta = (\theta_1, \ldots, \theta_M)$ and sample from conditional distributions for θ_m given $-m$.

For example suppose that we have a draw of N 0-1 data points. Let's treat N as uncertain following a Poisson distribution with parameter λ. The number of successes is Binomial with parameters N and parameter θ. Finally, the prior of θ is Beta with parameters α and β.

The joint distribution of (θ, X, N) has a really complicated form. No analytic-closed form. So how could we draw from the posterior? To do this we need to find the conditional distributions. Here

$$
\begin{aligned}
f(X|\theta, N) &\propto \binom{N}{X}\theta^X(1-\theta)^{N-X} \propto \text{Bin}(N, \theta) \\
f(\theta|X, N) &\propto \theta^{\alpha+X-1}(1-\theta)^{\beta+N-X-1} \propto \text{Beta}(\alpha + X, \beta + N - X) \\
f(N|\theta, X) &\propto \binom{N}{X}\frac{\lambda^N}{N!}(1-\theta)^{N-X} \propto \frac{(\lambda(1-\theta))^{N-X}}{(N-X)!} \propto \text{Poi}(\lambda(1-\theta)) \text{ for } n = x, x+1, \ldots
\end{aligned}
$$

How to do the sampling? Pick $(X^{(0)}, \theta^{(0)}, N^{(0)})$. Then for $k = 1, 2, \ldots$ *sample*

$$
\begin{aligned}
X^{(k)} &\sim \text{Bin}(N^{(k-1)}, \theta^{(k-1)}) \\
\theta^{(k)} &\sim \text{Beta}(\alpha + X^{(k)}, \beta + N^{(k-1)} - X^{(k)}) \\
N^{(k)} = X^{(k)} + z,\ z &\sim \text{Poi}(\lambda(1-\theta^{(k)}))
\end{aligned}
$$

and repeat until convergence.

Why does this work? There is some deep theory behind it using markov chains that I won't go into. For more information, take Guido's metrics class or Bayesian statistics courses.

Section 10 - OLS

1. (Linear Algebra Review) This week we did OLS. To fully understand the derivations behind OLS, linear algebra is necessary. Let's review some important and basic concepts for linear

algebra.

- (Transpose of Multiplication) An important operator is transpose. If A and B are matrices then $(AB)' = B'A'$.
- (Inverse of Multiplication) An important operator is inverse. If A and B are invertible matrices then $(AB)^{-1} = B^{-1}A^{-1}$
- (Definitiveness) The quadratic form of a matrix is $x'Ax$. Based on the sign of this object for all non-zero x of the appropriate dimensions, we say that A is [something]-definite where something could be "positive", "negative" etc. This idea is useful in Gauss-Markov theorem statement.
- (Matrix Calculus and Algebra):

$$\frac{\partial Ax}{\partial x} = A, \ \frac{\partial x'Ax}{\partial x} = (A + A')x, \ \frac{\partial x'Ax}{\partial A} = xx'$$

  these are useful for solving for the OLS solution and other related problems.
- (Idempotent Matrices) A (square) matrix A is idempotent if $AA = A$. Consider the projection matrix and residual makers

$$P_X = X(X'X)^{-1}X'$$

$$M_X = I - P_X$$

  How to check this? First note that these matrices are symmetric.

$$\begin{aligned} P_X P_X &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= X(X'X)^{-1}IX' \\ &= P_X \end{aligned}$$

  and similar for M_X.
- (OLS Coefficient) In general we can express the OLS coefficient in a model of the form

$$y = X\beta + \epsilon$$

  as $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y$.

2. (Helpful Properties of Matrix Notation for OLS Objects) There very important objects in OLS are y, the dependent variable observations, $\hat{y}$, the predicted values based on the linear

model, and e the residuals. Note that

$$\begin{aligned} e &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= (I - X(X'X)^{-1}X')y \\ &= M_X y \end{aligned}$$

and

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= yX(X'X)^{-1}X'y \\ &= P_X y \end{aligned}$$

and so

$$y = \hat{y} + e = P_X y + M_X y$$

so that the observed data consist of the projection of y onto the column space of X plus the residual. This gives a useful way to represent and compare these three objects.

3. (t-tests: An Example of Differentiating Normal and more General Models) One common test run on OLS model coefficients is $H_0 : \beta_j = 0$. These are usually called t-tests. Consider the normal linear model. In this case, $\hat{\beta}$ is the MLE and has a multivariate normal distribution *in finite samples* and so the t-stat

$$t_j = \hat{\beta}_j / \text{s.e.}(\hat{\beta}_j)$$

actually does have an exact t distribution.

However, in general we do not like the strong normal linear model assumptions. Instead, we prefer the more general exogeneity assumptions. In particular, under the assumptions that ϵ is not necessarily normal we have that $\hat{\beta}_j$ is not exactly normal and so t_j is not exactly t. However, as shown, the asymptotic distribution of t_j is now a standard normal distribution, and so this is what we base our tests on.

4. (OLS as MoM) We motivated OLS as a maximum likelihood estimator of a normal linear model and also as a linear statistical optimization problem. There is another way to motivate using some of the methods we have learned - as a method of moments estimator. To do this consider we need a moment. Recall that $\mathbb{E}[\epsilon|X] = 0$ implies that $\mathbb{E}[X'\epsilon] = 0$ (or that their covariance is 0 if we normalize $\mathbb{E}[\epsilon] = 0$). Then the method of moments estimator based on

this replacing $\epsilon = y - X\beta$ can be written as

$$\mathbb{E}[X'(y - X\beta)] = 0$$

and expanding this gives us

$$\mathbb{E}[X'y] - \mathbb{E}[X'X]\beta = 0$$

since β is a constant. Thus solving for β we have

$$\beta = (\mathbb{E}[X'X])^{-1}\mathbb{E}[X'y]$$

and the sample analogue to this is

$$\hat{\beta}_{\text{mom}} = (\frac{1}{n}X'X)^{-1}\frac{1}{n}X'y = \hat{\beta}_{\text{ols}}$$

and thus, this is a method of moments estimator!

This solidifies the intuition that the important assumptions in OLS for recovering the parameters really only have to do with exogeneity conditions, not distributional assumptions, since the method of moments estimator is completely distribution free. Also, it shows that strict exogeneity is not strictly necessary, and that $\mathbb{E}[X'\epsilon] = 0$ is sufficient in this linear model. You will see more on this next quarter.

5. (Review for Final) Leave extra time for questions on final.