Math 481 crash course

Summer 2024

Disclaimer: The following notes are NOT comprehensive, and are being written to help keep people on track with the course. You still need to take notes in lecture and read the textbook.

Review of prerequisite materials

Definition 1. A *sample space* is a set S containing *sample points*, i.e elements $s \in S$. The sample points are the possible outcomes of an experiment or measurement process.

Examples of sample spaces

- $S = \{H, T\}$ for flipping a coin once
- $S = \{HH, HT, TH, TT\}$ for flipping a coin twice
- $S = \{1, .., 6\}$ for rolling a dice once
- $S = \{1, ..., 6\} \times \{1, ..., 6\}$ for rolling two die
- $S = [0, 1]$ for partially filling a volume 1 container with water and recording the total volume

Definition 2. A subset of the sample space $E \subset S$ is called an *event*. Events A and B are *mutually exclusive* if $A \cap B = \emptyset$.

Example If the sample space is $S = \{1, ..6\}$, the events $Even = \{2, 4, 6\}$ and $Odd = \{1, 3, 5\}$ are mutually exclusive

Definition 3. A *probability measure* on S is a function $P : \mathcal{P}(S) \to [0, 1]$ satisfying

1. $P(S) = 1$
2. If $\{A_i\}_{i \in I}$ is a collection of mutually exclusive events (i.e $A_i \cap A_j = \emptyset$ for any $i \neq j$), then

$$P(\bigcup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$$

The notation $\mathcal{P}(S)$ is the powerset of S, i.e the set of all subsets/events of S.

Definition 4. A *random variable* (RV) X is a function $X : S \to \mathbb{R}$. A random variable is a *discrete random variable* if the range of X, $X(S)$ consists of finitely many or countably infinitely many points. Otherwise X is a continuous random variable.

Examples of RV's

- Flip a coin n times. The sample space is $S = \{L_1...L_n | L_i = H \text{ or } T\}$, the set of n length strings made of heads and tails. Let X be the random variable counting the number of heads of a sample point. More explicitly,

$$X(L_1...L_n) = k$$

  where k is the number of $L_i = H$.

- Roll two die. The sample space is $S = \{1, ..., 6\} \times \{1, ...6\} = \{(i, j) | i, j = 1, ..., 6\}$. Let Y be the random variable that records the sum of the faces of the die: that is, $Y((i, j)) = i + j$.

A random variable X on a sample space S naturally splits up the sample space into a set of mutually exclusive events:

$$S = \bigcup_{x \in \mathbb{R}} X^{-1}(x).$$

Often times we write $X = x$ as notation for $X^{-1}(x) = \{s \in S | X(s) = x\}$, so we can rewrite this as

$$S = \bigcup_{x \in X} (X = x).$$

If S is equipped with a probability measure P, then we can record the probabilities of these events $X = x$ in a function.

Definition 5. If X is a discrete random variable, then $f(x) := P(X = x)$ is the *probability distribution function* (PDF) of X. If X is a continuous random variable and f is non-negative function satisfying

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

for all real constants $a < b$, then we say that $f(x)$ is the *probability density function* (also abbreviated PDF) of X. Given two discrete random variables X and Y, their *joint PDF* is $f(x, y) = P(X = x, Y = y)$. Given two continuous random variables, $f(x, y)$ is the *joint PDF* of X and Y if for all subset $A \subset \mathbb{R}^2$,

$$P((X, Y) \in A) = \int \int_A f(x, y)dxdy.$$

The joint PDF is defined also for more than two random variables in the obvious way.

Definition 6. If the joint PDF $f(x, y)$ of two random variables X and Y factors into the product of the individual PDFs $f(x, y) = f_X(x)f_Y(y)$, we say that that the random variables are *independent*.

Definition 7. The *expectation value* of a discrete random variable X with PDF $f(x)$ is

$$E[X] := \sum_{x \in X(S)} xf(x).$$

The *expectation value* of a continuous random variable X with PDF $f(x)$ is

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

If $g : \mathbb{R} \to \mathbb{R}$ is a function, and $X : S \to \mathbb{R}$ is an RV we can form a new RV $g(X) := g \circ X : S \to \mathbb{R}$. It is not hard to show that

$$E[g(X)] = \sum_{x \in X(s)} g(x)f(x)$$

in the discrete case or

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

in the continuous case.

Definition 8. The r^{th} *moment of a random variable* X is

$$\mu'_r := E[X^r].$$

Using the above discussion about expectations of $g(X)$ with $g = x^r$ we can compute the moments of a RV. The first moment μ'_1 is called the *mean* of X and is denoted $\mu := \mu'_1 = E[X]$ (notice that this is also just the expectation value of X – it has way too many names, sorry).

Definition 9. The r^{th} *moment about the mean* of a random variable X is

$$\mu_r := E[(X - \mu)^r].$$

The second moment about the mean $\mu_2 = E[(X - \mu)^2]$ is called the *variance* of X and is denoted $\sigma^2 = \sigma_X^2 = Var(x)$.

Simple algebra shows that $\sigma^2 = \mu'_2 - \mu^2 = E[X^2] - E[X]^2$.

Definition 10. The moment generating function (MGF) of a RV X is $M_X(t) := E[e^{tX}]$.

Moment generating functions are very important because of the theorem that says that if X and Y are RV's with equal MGFs, then X and Y also have equal PDFs, and if X_n is a sequence of RVs such that the MFGs of X_n approaches the MGF of some RV Y, then the PDFs of the X_n approach the PDF of Y. It is obvious that the converse of this theorem is true, so we can think of this as an equivalence of MGFs and PDFs.
Now we discuss some special PDFs and the associated means, variances and MFGs of RV's which possess such PDFs. Please note that many different random variables can all possess the same PDF, so when we say that a random variable has (Bernoulli, binomial, poisson, Chi-square, normal, ...) distribution, we are just saying that we know the PDF of that RV: it is NOT complete information about that RV, i.e the pdf does not specify the function $X : S \to \mathbb{R}$.

Special PDFs

- If X is a discrete random variable with range equal $X(S) = \{0, 1\}$ and PDF given by $f(0) = \theta$, $f(1) = 1 - \theta$, or more succinctly

$$\boxed{f(x) = \theta^x(1 - \theta)^{1-x}}$$

, then we say that X is a Bernoulli random variable with parameter θ.

- If X is a discrete random variable with range equal to $X(S) = \{0, ...n\}$ and PDF given by

$$\boxed{b(x, n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}},$$

then we say that X is a binomial random variable with paramter θ. The mean and variance are $\boxed{\mu = n\theta}$ and $\boxed{\sigma^2 = n\theta(1-\theta)}$. The MGF of a binomial random variable is

$$\boxed{M_X(t) = (e^t\theta + 1 - \theta)^n}$$

- If X is a discrete random variable with range equal to $X(S) = \{0, 1, 2, ...\} = \mathbb{Z}_{\geq 0}$ and PDF given by

$$\boxed{f(x, \lambda) = \frac{\lambda^x, e^{-\lambda}}{x!}}$$

we say that X is a Poisson random variable with parameter λ. The MGF of a Poisson random variable is

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

- If X is a continuous random variable with PDF

$$\boxed{g(x, \theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}}$$

for $x > 0$ and 0 elsewhere, then we say that X is an exponential random variable with parameter θ. The MGF of an exponential random variable is

$$\boxed{M_X(t) = (1-\theta t)^{-1}}$$

- If X is a continuous random variable with PDF

$$\boxed{f(x, \nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\nu/2}x^{\frac{\nu-2}{2}}e^{-\frac{x}{2}}}$$

for $x > 0$ and 0 for $x \leq 0$ then X is a Chi-square random variable with ν degrees of freedom. The MFG of a Chi-square random variable is

$$\boxed{M_X(t) = (1-2t)^{-\frac{\nu}{2}}}$$

- If X is a continuous random variable with PDF

$$\boxed{\mathcal{N}(\mu, \sigma)(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}$$

then X is a normal random variable with mean μ and variance σ^2. For $\mu = 0$, $\sigma = 1$, this is the standard normal distribution. The MGF of a normal random variable is

$$\boxed{M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}}$$

If X has standard normal distribution, then X^2 has Chi-square distribution with one degree of freedom. Important strategies for figuring out the PDF of a random variable:

- Calculate the CDF and differentiate. This works for showing that X^2 is Chi-square with one d.o.f for example if X is standard normal with PDF f and CDF $F(x)$: for $x \geq 0$ we have

$$P(X^2 \leq x) = P(-\sqrt{x} \leq X \leq \sqrt{x}) = F(\sqrt{x}) - F(-\sqrt{x})$$

  and taking derivatives while using the chain rule we see that the PDF of X^2 is given by

$$F(\sqrt{x})' - F(-\sqrt{x})' = \frac{1}{2\sqrt{x}} f(\sqrt{x}) + \frac{1}{2\sqrt{x}} f(-\sqrt{x}) =$$

$$\frac{1}{2\sqrt{x}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\sqrt{x}^2} + \frac{1}{2\sqrt{x}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\sqrt{x}^2} = \frac{1}{\sqrt{2\pi x}} e^{-\frac{1}{2}x}$$

- The transformation technique in one or multiple variables
- Identifying the MGF of the RV with some well known MGF and concluding that the PDF's must be the same by the theorem stated above.

Chapter 8

8.1

Definition 11. The *population* is our sample space in a statistics problem.

We usually will assume that the population is infinite. In light of various limiting theorems about discrete random variables, this is actually a pretty useful assumption to make calculations.

Definition 12. A collection of random variables $X_1, ..., X_n : S \to \mathbb{R}$ is *identically distributed* if all of the X_i have the same PDF, $f_{X_1} = f_{X_2} = ... = f_{X_n}$.

The definition above along with the definition of independence (suitably generalized to n instead of just two variables) allows us to define the following fundamental concept.

Definition 13. A *random sample* is a collection of independent and identically distributed (i.i.d) random variables $X_1, ..., X_n$.

Given a random sample on our population, we are interested in knowing information about various statistics associated to it:

Definition 14. Given a random sample $X_1, ..., X_n$, a *statistic* is a new random variable Y which is a function of the random sample

$$Y = u(X_1, ..., X_n)$$

Famous examples of statistics that we are most concerned with in this course (though we will see some others) are

Definition 15. The *sample mean* of a random sample $X_1, ..., X_n$ is

$$\overline{X} := \frac{1}{n}(X_1 + ... + X_n).$$

Definition 16. The *sample variance* of a random sample $X_1, ..., X_n$ is

$$S^2 := \frac{1}{n-1}((X_1 - \overline{X})^2 + ... + (X_n - \overline{X})^2)$$

The $n-1$ in the definition of sample variance will make more sense after the next chapter. We also define $S = \sqrt{S}$, the positive square root of the sample variance, to be the *sample standard deviation.*

8.2

The mean and variance of the random variable $\overline{X}$ are easy to compute.

$$E[\overline{X}] = E[\frac{1}{n}(X_1 + ... + X_n)] = \frac{1}{n}(E[X_1] + ... + E[X_i]) = \frac{1}{n}(\mu + ... + \mu) = \mu$$

where $\mu = E[X_1] = .. = E[X_n]$ (recall that the X_i, being a random sample, are identically distributed, hence all of the same PDF, and hence all have the same mean μ).

$$Var(\overline{X}) = Var(\frac{1}{n}(X_1 + ..._+ X_n)) = \frac{1}{n^2}(Var(X_1) + ... + Var(X_n)) = \frac{1}{n^2}(\sigma^2 + ... + \sigma^2) = \frac{\sigma^2}{n}$$

where $\sigma^2 = Var(X_1) = ...Var(X_n)$. In the second equality we made use of the formula

$$Var(\sum a_i X_i) = \sum a_i^2 Var(X_i)$$

if the X_i are independent, which is of course the case for X_i comprising a random sample. There is a more general formula for the variance of a sum involving covariances that we discussed in class but have not recorded here. Applying Chebyshev's theorem (not proved here) to the sample mean yields the following:

Let $c > 0$ be any constant. Then

$$P(|\overline{X} - \mu| \leq c) \geq 1 - \frac{\sigma^2}{nc^2}.$$

This is often called the Law of large numbers. Taking the limit as $n \to \infty$ shows us that the sample mean concentrates about μ more and more as we take larger sample sizes of our population. Now we recall the Central Limit Theorem, a cornerstone of modern statistics.

The Central Limit Theorem
Statement: Let $X_1, ..., X_n$ be a random sample from a population with mean μ and variance σ^2. As $n \to \infty$, the PDF's of the sequence of random variables

$$Z_n := \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

approaches the PDF of a normal distribution.
Proof outline: By the MGF theorem it suffices to show that $M_{Z_n}(t) \to M_{\mathcal{N}(0,1)}(t)$ as $n \to \infty$. We showed this by recalling that $M_{\mathcal{N}(0,1)}(t) = e^{\frac{1}{2}t^2}$ and hence it suffices to show that $ln(M_{Z_n}(t))$ approaches $\frac{1}{t^2}$. This was accomplished by Taylor expanding both $M_{Z_n}(t)$ and $ln(1+x)$ and grouping together terms by their power of t before concluding that all besides the t^2 term die in the $n \to \infty$ limit. Details in textbook.

Some important terminology: Sometimes, a problem will say that a random sample "comes from a [] population" for [] = normal, exponential, ... All this means is that the PDFs of the random sample (which, remember, are all the same PDF) are equal to that of a normal, exponential or whatever else distribution. The CLT isn't even needed if we take our sample from a normal population; in this case the sample mean isn't just approximately normal, it is normal:

If $X_1, ..., X_n$ is a random sample from a normal population with mean μ and variance σ^2 then $\overline{X}$ is normal with parameters μ and $\frac{\sigma^2}{n}$. This is proved by looking at MGFS.

0.1 8.4

Recall the PDF of a Chi square distributed RV χ^2 with ν degrees of freedom: $f(x, \nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu-2}{2}} e^{-\frac{x}{2}}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. Recall also its MGF:

$$M_{\chi^2}(t) = (1 - 2t)^{-\frac{\nu}{2}}.$$

At the end of the review section we showed that if X is an RV with standard normal distribution, then X^2 is Chi square with one d.o.f. More generally:

If $X_1, ..., X_n$ are independent RV's with standard normal distribution, then $X_1^2 + ... + X_n^2$ is Chi-square distributed with n degrees of freedom.
Proof: If $X_1, ..., X_n$ are independent, then $X_1^2, ..., X_n^2$ are also independent (this is more generally true about $g_1(X_1), ..., g_n(X_n)$ for measurable functions g_i; this is again an independent set of RVs if the X_i are independent). The MGF of a sum of independent RV's is the product of the MGFs of those variables, so we get

$$M_{X_1^2+...+X_n^2}(t) = M_{X_1^2}(t)...M_{X_n^2}(t) = (1 - 2t)^{\frac{-n}{2}}.$$

The RHS is the MGF of a Chi-square distributed variable with n d.o.f, and hence by the correspondence theorem between MGF's and PDFS we are done.

An exercise left for the homework is to show the following mild generalization.

If $X_1, ..., X_n$ are Chi-square RVs with $\nu_1, ..., \nu_n$ d.o.f respectively, then $X_1 + ... + X_n$ is Chi-square with $\nu_1 + ... + \nu_n$ d.o.f.

What does this have to do with sample statistics?

If $\overline{X}$ and S^2 are the sample mean and variance of a normal population with mean μ and variance σ^2, then

- $\overline{X}$ and S^2 are independent
- The random variable $\frac{(n-1)S^2}{\sigma^2}$ is Chi square distributed with $\nu = n - 1$ degrees of freedom.

The first part of this theorem is outlined in exercise 31 of the book. I still suggest that you try to work through it, and will offer substantial extra credit (about 1 quiz worth) to anyone who turns in a full write up it. Assuming the first part, by taking MGFs of the algebraic identity (another homework problem)

$$\sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma})^2 = \frac{(n-1)S^2}{\sigma^2} + n(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}})^2$$

, we are able to conclude (how? it is worth thinking through this and formalizing it) that $\frac{(n-1)S^2}{\sigma^2}$ is Chi-square with $n - 1$ d.o.f.

8.5

Definition 17. If Y has Chi-square distribution with ν degrees of freedom, and Z has standard normal distribution, then

$$T := \frac{Z}{\sqrt{\frac{Y}{\nu}}}$$

is called a *t-distribution with ν degrees of freedom.*

It is possible to calculate the PDF of a t-distribution using the multivariable change of variables technique. We just state the result here, but students should be aware that they are responsible for understanding how to derive such a distribution (calculating appropriate Jacobians and using the transformation technique). I might fill this in later.

The PDF of T is given by

$$\boxed{f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})}(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}}$$

Why do we care about the t-distribution? If we take a random sample from a normal population, we know that $\frac{\overline{X}-\mu}{\sigma}$ has standard normal distribution, but σ is hard to know precisely. So often times, statisticians instead try to measure

$$\frac{\overline{X}-\mu}{\sqrt{S^2}} = \frac{\frac{\overline{X}-\mu}{\sigma}}{\sqrt{\frac{S^2}{\sigma^2}}}$$

which – by the above theorem combined with the theorems about $\overline{X}$ being normally distributed with paramters μ and $\frac{\sigma^2}{n}$ and $\frac{(n-1)S^2}{\sigma^2}$ being Chi-square distributed with $n-1$ d.o.f – has t-distribution with $n-1$ degrees of freedom.

8.7

Let $u_r(x_1, ..., x_n)$ be the function that outputs the r^{th} largest x_i. Then given any random sample we can define the following statistics.

Definition 18. The r^{th} *order statistic* is the random variable defined by

$$Y_i := u_i(X_1, ..., X_n).$$

Definition 19. If $n = 2m+1$ is odd, Y_{m+1}, the $(m+1)^{th}$ order statistic is called the *sample median* and is denoted $\widetilde{X} := Y_{m+1}$. If $n = 2m$ is even, then $\widetilde{X} := \frac{1}{2}(Y_m + Y_{m+1})$ is instead the definition of the *sample median*.

The PDF of the r^{th} order statistic Y_r can be derived as follows. Say that $f(x)$ is the PDF of the X_i, and $F(x) = \int_{-\infty}^{x} f(x)dx$ is the CDF of X_i. Below we use that $P(X_i \geq x) = 1 - F(x)$. To find the $g_r(y)$, the PDF of Y_r, we take the derivative of $G_r(y)$, the CDF of Y_r.

$$g_r(y) = G_r'(y) = \frac{d}{dy}P(Y_r \leq h) = \frac{d}{dy}\Big[P[(Y_r \leq y) \wedge (Y_{r+1} \geq y)] + P[(Y_{r+1} \leq y) \wedge (Y_{r+2} \geq y)] +$$

$$... + P[(Y_{n-1} \leq y) \wedge (Y_n \geq y)] + P[Y_n \leq y]\Big] = \frac{d}{dy}\sum_{i=r}^{n}\binom{n}{i}F(y)^i(1-F(y))^{n-i}$$

$$= \Big[\sum_{i=r}^{n-1}\frac{n!}{(i-1)!(n-i)!}f(y)F(y)^{i-1}(1-F(y))^{n-i} - \frac{n!}{i!(n-i-1)!}F(y)^i f(y)(1-F(y))^{n-i-1}\Big]$$

$$+ f(y)nF(y)^{n-1} =$$

$$\Big[\sum_{i=r}^{n-1}\frac{n!}{(i-1)!(n-i)!}f(y)F(y)^{i-1}(1-F(y))^{n-i} - \sum_{j=r+1}^{n}\frac{n!}{(j-1)!(n-j)!}f(y)F(y)^{j-1}(1-F(y))^{n-j}\Big]$$

$$+ f(y)nF(y)^{n-1} =$$

$$\frac{n!}{(r-1)!(n-r)!}f(y)F(y)^{r-1}(1-F(y))^{n-r} - \frac{n!}{(n-1)!0!}f(y)F(y)^{n-1} + f(y)nF(y)^{n-1} =$$

$$\boxed{\frac{n!}{(r-1)!(n-r)!}f(y)F(y)^{r-1}(1-F(y))^{n-r}}$$

As an example, let's compute the PDF of the first and n^{th} order statistics of a random sample $X_1, ..., X_n$ drawn from an exponential population with parameter θ. Recall that this means that the PDF's of each X_i are $f(x) = \frac{1}{\theta}e^{\frac{-x}{\theta}}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. First let's compute $g_1(y)$, the PDF of $Y_1 = min(X_1, ..., X_n)$. For $y > 0$ we get

$$g_1(y) = \frac{n!}{0!(n-1)!}f(y)[\int_{-\infty}^{y} f(x)dx]^0[\int_{y}^{\infty} f(x)dx]^{n-1} = \frac{ne^{-\frac{y}{\theta}}}{\theta}[\int_{y}^{\infty} \frac{1}{\theta}e^{-\frac{x}{\theta}}dx]^{n-1} =$$

$$\frac{ne^{-\frac{y}{\theta}}}{\theta^n}(\int_{\frac{y}{\theta}}^{\infty} e^{-u}\theta du)^{n-1} = \frac{n}{\theta}e^{-\frac{y}{\theta}}(-e^{-u}|_{\frac{y}{\theta}}^{\infty})^{n-1} = \boxed{\frac{n}{\theta}e^{-\frac{ny}{\theta}}}$$

And for $y \leq 0$, $g_1(y) = 0$. We leave the computation of $g_n(y)$ and $g_{m+1}(y)$, the PDFS of Y_n and the sample median, for $n = 2m + 1$ as exercises THAT ARE WELL WORTH DOING. They are quite similar to the computation of $g_1(y)$.

Chapter 10

10.1

Definition 20. A *population parameter* is a variable associated to the pdf of a random sample.

Examples of population parameters

- If the random sample is drawn from a normal population with pdf $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ then σ and μ are population parameters.
- If the random sample consists of binomial random variables with success rate θ and pdf $\binom{n}{x}\theta^x(1-\theta)^{n-x}$, θ is a population parameter.
- If the random sample is drawn from an exponential population with pdf $\frac{1}{\theta}e^{-\frac{x}{\theta}}$, θ is a population parameter.
- If the random sample is drawn from a uniform population with pdf $\frac{1}{\beta-\alpha}$ for $\alpha < x < \beta$ and 0 otherwise, then α and β are population parameters.

Definition 21. The process of using a sample statistic $\hat{\Theta}$ to estimate the value of a population parameter θ is known as *point estimation*, and the sample statistic $\hat{\Theta}$ used to estimate is called a *point estimator*.

There are various desirable properties that a point estimator might have, and there are a couple of methods that one can use to produce point estimators – these topics will occupy the rest of this chapter.

10.2

Definition 22. An *unbiased estimator* for a population parameter θ is a point estimator $\hat{\Theta}$ for θ such that $E[\hat{\Theta}] = \theta$.

If $\hat{\Theta}$ is an estimator for θ then we can more generally define the *bias* of $\hat{\Theta}$ to be $E[\hat{\Theta}] - \theta$. A weaker but more widely applicable notion than unbiasedness is asymptotic unbiasedness.

Definition 23. A point estimator Θ for θ is *asymptotically unbiased* if in the limit as n (the size of the random sample) goes to infinity, the bias of Θ tends to 0.

Unbiased estimators for uniform populations Consider a random sample drawn from a uniform population with $\alpha = 0$ and β the population parameter being measured. We claim that $\frac{n+1}{n}Y_n$ is an unbiased estimator for β. The pdf of Y_n (see 8.7 for the formula we are about to use) is given by

$$g_n(y) = n \cdot \frac{1}{\beta}(\int_0^\beta y^n dy)^{n-1} = \frac{n}{\beta^{n-1}}$$

for $0 < y < \beta$ and 0 elsewhere. Hence

$$E(\frac{n+1}{n}Y_n) = \frac{n+1}{n} \cdot \frac{n}{\beta^n} \int_0^\beta y^n dy = \beta$$

We can now show why in the definition of sample variance we divided by $n-1$ instead of n. It is because this definition makes S^2 into a unbiased estimator of σ^2:

$$E[S^2] = \frac{1}{n-1}E[\sum_{i=1}^{n}(X_i - \overline{X})^2)] = \frac{1}{n-1}E[\sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2)] =$$

$$\frac{1}{n-1}E[\sum_{i=1}^{n}(X_i - \mu)^2] - nE[(\overline{X} - \mu)^2] = \frac{1}{n-1}(n\sigma^2 - n \cdot \frac{\sigma^2}{n}) = \sigma^2$$

10.3

Definition 24. If $\hat{\Theta}$ is an unbiased estimator for a population parameter θ such that if $\hat{\Theta}'$ is any other unbiased estimator for θ, then

$$Var(\hat{\Theta}) < Var(\hat{\Theta}')$$

then we say that $\hat{\Theta}$ is the *minimum variance unbiased estimator* for θ or *the best estimator* for θ.

Since the variance is a measure of how spread an unbiased estimator is away from the mean (which is the population parameter it is measuring), it is obviously desirable to produce a best estimator for θ in the sense of the above definition. This seems like a daunting task as there are infinitely many unbiased estimators for a given population parameter. This situation can sometimes be addressed through the use of the Cramer-Rao inequality, whose proof we did not cover in this course.

Cramer-Rao inequality Say that the PDF of our random sample is $f(x)$, and that f depends implicitly on the population parameter θ. Also assume that $\frac{\partial}{\partial\theta}\int_{-\infty}^{y} f(x)dx = \int_{-\infty}^{y}\frac{\partial}{\partial\theta}f(x)dx$ for all values of y. Then if $\hat{\Theta}$ is an unbiased estimator for θ and X is any of the random variables in the random sample,

$$Var(\hat{\Theta}) \geq \frac{1}{n \cdot E[(\frac{\partial ln(f(X))}{\partial\theta})^2]}$$

We can use the Cramer-Rao inequality to demonstrate that certain unbiased estimators are actually best estimators.

The best estimator for the mean of a normal population Given a random sample from a normal population with pdf $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, we calculate the right hand side of the Cramer inequality with $\theta = \mu$.

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

$$\frac{d}{d\mu}(ln(f(x)) = \frac{x-\mu}{\sigma^2} \implies$$

$$E[(\frac{d}{d\mu}(ln(f(X)))^2] = \frac{1}{\sigma^2}E[(\frac{X-\mu}{\sigma})^2] = \frac{1}{\sigma^2} \implies$$

$$\frac{1}{n \cdot E[(\frac{\partial ln(f(X))}{\partial\theta})^2]} = \frac{\sigma^2}{n} = Var(\overline{X})$$

Since the variance of $\overline{X}$ is equal to the RHS of the Cramer-Rao inequality and $\overline{X}$ is an unbiased estimator for μ, it follows that $\overline{X}$ has smaller variance than any other unbiased estimator for μ and hence is a best estimator for μ. Technically we should also check that $\frac{d}{d\mu}\int_{-\infty}^{y} f(x)dx = \int_{-\infty}^{y}\frac{d}{d\mu}f(x)dx$ so that the assumptions of Cramer-Rao are met, but we leave this as an exercise.

Given two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ for a population parameter θ, we can compare their relative efficiencies by examining the ratio of their variances: we say that $\hat{\Theta}_2$ is *relatively* $\frac{Var(\hat{\Theta}_1)}{Var(\hat{\Theta}_2)} \cdot 100\%$ *efficient as an estimator for* θ as $\hat{\Theta}_1$.

Given a uniform population with $\alpha = 0$, we showed before that $\frac{n+1}{n}Y_n$ is an unbiased estimator of β. We also have that $\overline{X}$ is an unbiased estimator of β since $E[2\overline{X}] = 2 \cdot \frac{\alpha+\beta}{2} = \beta$. Using the pdf of Y_n derived in 8.7 applied to a uniform population, we can show that

$$E[Y_n^2] = \frac{n}{n+2}\beta^2$$

and hence

$$Var(Y_n) = E[Y_n^2] - E[Y_n]^2 = \frac{n}{n+2}\beta^2 - (\frac{n}{n+1}\beta^2) \implies$$

$$Var(\frac{n+1}{n}Y_n) = (\frac{n+1}{n})^2(\frac{n}{n+2}\beta^2 - (\frac{n}{n+1}\beta^2)) = \frac{\beta^2}{n(n+2)}$$

We also know that if X_i is an element of the random sample then $Var(X_i) = \frac{\beta^2}{12}$ (basic integration exercise with pdf of uniform population), and hence $Var(2\overline{X}) = \frac{4}{n}Var(X_i) = \frac{\beta^2}{3n}$. Therefore

$$\frac{Var(\frac{n+1}{n}Y_n)}{Var(2\overline{X})} = \frac{3}{n+2}$$

so for e.g $n = 10$, $2\overline{X}$ is 25% efficient as an estimator compared to $\frac{n+1}{n}Y_n$, and as n gets larger this efficiency gets even lower – we conclude that $\frac{n+1}{n}$ is a much more efficient estimator of β than $2\overline{X}$.

10.4

The next property of estimators that we discussed is called consistency.

Definition 25. An estimator $\hat{\Theta}$ for a parameter θ is called *consistent* if for any $c > 0$,

$$\lim_{n\to\infty} P(|X - \theta| < c) = 1.$$

We note that Chebyshev's theorem implies that an unbiased estimator $\hat{\Theta}$ with the property that $\lim_{n\to\infty} Var(\hat{\Theta}) = 0$ then $\hat{\Theta}$ is a consistent estimator. In fact, more is true: if $\hat{\Theta}$ is only asymptotically unbiased with $\lim_{n\to\infty} Var(\hat{\Theta}) = 0$ then $\hat{\Theta}$ is still a consistent estimator.

Using the fact that $\frac{(n-1)S^2}{\sigma^2}$ is chi-square distributed with $n-1$ d.o.f for a random sample from a normal population, we can calculate that $Var(S^2) = \frac{2\sigma^4}{n-1}$. Since we know that S^2 is an unbiased estimator for σ^2 and the variance of S^2 (for a random sample from a normal population clearly goes to 0 by the formula above, we see that S^2 is a consistent estimator of σ when we sample from a normal population.

10.7

Here we describe a simple method, called the method of moments, to write down an estimate for a collection of population parameters $\theta_1, ..., \theta_k$. First we need a definition

Definition 26. Let $x_1, ..., x_n$ be the observed values of some random sample $X_1, ..., X_n$. Then the r^{th} *sample moment* is

$$m_r' := \frac{\sum_{i=1}^n x_1^r}{n}.$$

The method of moments simply consists of equating m_r' with the r^{th} moment $\mu_r' = E[X_i^r]$ for $r = 1, ..., k$ and solving for the population parameters $\theta_1, ..., \theta_k$. We illustrate this with two examples.

Let's use the method of moments to estimate α for a uniform population with $\beta = 1$. In this case we just need to set the first sample moment $m_1' = \bar{x} = \frac{x_1+...+x_n}{n}$ equal to the first moment $\mu_1' = \frac{\alpha+\beta}{2} = \frac{\alpha+1}{2}$,. Solving for α we see that $\boxed{\hat{\alpha} = 2\bar{x} - 1}$. We put a hat on α to indicate that this equation is an estimate for α using the observed data $x_1, ..., x_n$.

Given a random sample from a Γ population, let's use the method of moments to estimate both α and β. Recalling that the MGF of a Γ RV is $(1-\beta)^{-\alpha}$, by taking two derivatives we can calculate that $\mu_1' = \alpha\beta$ and $\mu_2' = \alpha(\alpha+1)\beta^2$. Therefore the method of moments consists of solving the two equations for α and β:

$$m_1' = \overline{x} = \alpha\beta$$

$$m_2' = \alpha(\alpha+1)\beta^2$$

. Note that $m_2' - (m_1)'^2 = \alpha\beta^2$ and hence $\boxed{\frac{m_2' - (m_1')^2}{m_1'} = \hat{\beta}}$ while $\boxed{\frac{m_1'^2}{m_2' - (m_1')^2} = \hat{\alpha}}$.

10.8

Finally we describe a method for producing estimators which are asymptotically unbiased and sufficient (we didn't cover sufficiency, but it is another desirable property of an estimator) called the *method of maximum likelihood*. The method consists of

1. Thinking of the jpdf of our random sample $f(x_1, ..., x_n)$ as only a function $L(\theta_1, ..., \theta_k)$ of the population parameters $\theta_1, .., \theta_k$ that we are trying to write down estimators for. THIS IS JUST A NOTATION SHIFT THERE IS NO NEW FUNCTION. We call this newly notated function the "likelihood function."
2. Find the value of θ that maximizes L in terms of the x_i. This can be done using calculus or some other optimization technique.
3. Promote θ_k to $\hat{\Theta}_k$, the estimator for θ_k, by promoting the x_i to the X_i, the elements of the random sample.

We call $\hat{\Theta}_k$ the *maximum likelihood estimator for* θ_k. Let's see a couple of examples of this method.

Problem: Given x "successes" in n trials, find the maximum likelihood estimate of the parameter θ of the corresponding binomial distribution.
Solution In this case the pdf is $b(x;n,\theta) = \binom{n}{x}\theta^x(1-\theta^{n-x}$. We rename this pdf the likelihood function and denote it $L(\theta)$. Maximizing L as a function of θ is equivalent to maximizing $ln(L(\theta))$ since ln is monotone, so we work with $\frac{d}{d\theta}ln(L(\theta)) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$. Setting this equal to 0 and solving for θ yields $\theta = \frac{x}{n}$. Promoting x to X and θ to $\hat{\Theta}$, we see that that the max likelihood estimator for θ is $\hat{\Theta} = \frac{X}{n}$.

Problem: Find the max likelihood estimator of $\beta > 0$ when we draw a random sample of size n from a uniform population with $\alpha = O$.
Solution: The jpdf of such a random sample is $L(\beta) = f(x_1,...,x_n) = \frac{1}{\beta^n}$ for $0 \leq x_1,...,x_n \leq \beta$ and 0 otherwise. Clearly making β smaller makes the likelihood function larger so long as β is larger than the all of the x_i. It follows that the largest value the likelihood function can take for fixed values of the x_i is when $\beta = max(x_1,...,x_n)$. Promoting the x_i to the RV's X_i from the random sample gives $max(X_1,...,X_n) = Y_n$, the n^{th} order statistic, as the maximum likelihood estimator for β.

Problem: Given a random sample $X_1,...,X_n$ from a normally distributed population with parameters μ and σ^2, find the max likelihood estimators of μ and σ^2.
Solution: The jpdf, which we denote by $L(\mu,\sigma^2)$ to indicate that we are thinking of it as our likelihood function, is

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

We again use the trick of maximizing the log of $L(\mu,\sigma^2)$ instead since ln is an increasing function. Taking partial derivatives of $ln(L(\mu,\sigma^2))$ with respect to μ and σ^2 yields

$$ln(L(\mu,\sigma^2))_\mu = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu)$$

$$ln(L(\mu,\sigma^2))_{\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \mu^2).$$

Setting both of these equations equal to 0 and solving for μ and σ^2 in terms of the x_i and then promoting the x_i to X_i yields the max likelihood estimators

$$\hat{M} := \frac{1}{n}(X_1 + ... + X_n) = \overline{X}$$

$$\hat{\Sigma^2}L := \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{n-1}{n}S^2.$$

Notice that the max likelihood estimator of σ^2 is not an unbiased estimator of σ^2.