Stats 135, Fall 2019
Lecture 2, Friday, 8/30/2019

1 Review

Last time, we took the dichotomous case $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, and we have the sample mean:

$$\hat{p} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

as an (unbiased) estimate of the unknown parameter p.
We also found:

$$E(\hat{p}) = E(\overline{x}) = E(x) = p$$

and

$$\sigma_{\hat{p}}^2 = \text{SE of } (\hat{p})^2 = \text{var}(\overline{x}) = \frac{\overbrace{\text{var}(x)}^{p(1-p)}}{n}\left[\frac{N-n}{N-1}\right]$$

There are 2 estimates of SE of $(\hat{p})$: (1) a **conservative** estimate, and (2) a **bootstrap** estimate.

Topics Today:

- §7.3.3 : Confidence Intervals (CI) for $\mu = E(x)$ (or p in the dichotomous case).

  As an example in the dichotomous case, a 68% CI for p is

  $$\hat{p} \pm (\text{SE of } \hat{p})$$

  and a 95% CI for p is

  $$\hat{p} \pm 1.96\,(\text{SE of } \hat{p})\,.$$

  We can approximate SE of $\hat{p}$ to find a conservative or bootstrap confidence interval of p.
- §7.3.1 The expectation and variance of the sample mean.

2 Normal Approximation

First consider the normal approximation to the sampling distribution of $\overline{x}$:

Example: Consider a population of $N = 393$ hospitals. Let $x :=$ the number of patients discharged from the ith hospital, and let:

$$\mu = 814.6$$
$$\sigma = 590,$$

and a SRS (simple random sample) of $n = 50$ is taken. From our box (with μ, σ), we draw:

$$X_1, \ldots, X_{50}$$

and $\hat{\mu} = \overline{x}$. We want to find $P(|\overline{x} - \mu| > 100)$. The picture we have is a distribution centered at $\mu = 814.6$, and we have (from yesterday):

$$\sigma_{\overline{x}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = \frac{590}{\sqrt{50}}\sqrt{\frac{393-50}{393-1}} = 77.95$$

What's the chance of being at the tail? Take $\mu \pm \sigma_{\overline{x}}$.

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overbrace{\overline{x}}^{914.6} - 814.6}{77.95} = 1.28,$$

and this gives a standard normal curve, with bounds at -1.28 and 1.28.
The c.d.f. gives the tail end is

$$1 - \Phi(1.28)$$

The standard normal curve has c.d.f.

$$\Phi(z) = P(Z \leq z) = \texttt{pnorm(z)},$$

where in R we can find this using the command `pnorm(z)`.
And we have:

$$P\left(|\overline{x} - \mu| > 100\right) = 2\left(1 - \Phi(1.28)\right) = 2(.1) = .2$$

The question was posed in class what do we mean by a normal approximation? Lucas mentions that this curve is approximately normal, which follows from the Central Limit Theorem. So we just model it by the standard normal model. Given the mean and variance at the beginning, we obtain another SE $\sigma_{\overline{x}}$ which is written in terms of constants in our box. Usually we don't know these constants, and we would have to approximate this.

Example: Dichotomous Case In the hospital example, $n = 50$. Let p be the proportion of hospitals with fewer than 1000 discharges. We assign 0 if the hospital has more than or equal to 1000 discharges, and 1 if the hospital is less than 1000 discharges.
Suppose we know that $p = .65$. Find the tail-end probability:

$$P\left(|\hat{p} - p| > .13\right).$$

We use our finding from yesterday that:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}\left[\frac{N-n}{N-1}\right]} = \frac{.65 \cdot .35}{50}\left[\frac{N-n}{N-1}\right] = 0.063$$

$$P\left(|\hat{p} - p| > .13\right) = 2\left[1 - \Phi\left(\frac{.13}{.063}\right)\right] = 2\left(1 - \texttt{pnorm}(2.06)\right) = 0.039$$

3 Confidence Intervals (CI)

A Confidence Interval for a population parameter θ is a random interal, calculated from the sample that contains θ with some specified probability.
For $0 \leq \alpha \leq 1$, let $z(\alpha)$ be the number such that the area under the standard normal curve to the right of $z(\alpha)$ is α.
Then

$$P\left(-z\left(\frac{\alpha}{2}\right) \leq z < z\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha,$$

just by the definition of $z(\alpha/2)$. For example, if $\alpha = 0.05$, then

$$z\left(\frac{\alpha}{2}\right) = z\left(0.025\right) = \texttt{qnorm}(1 - .025) \approx 1.959964 = 1.96$$

and to find the point (z-value) such that this area is satisfied, we can use `qnorm` in R.
If $z = \frac{\overline{x}-\mu}{\sigma_{\overline{x}}}$, then $\overline{x}$ is approximately normal (so we normalize it). Then we take:

$$\begin{aligned} 1-\alpha &= P\left(\frac{\overline{x}-\mu}{\sigma_{\overline{x}}} \in \left[-z\left(\frac{\alpha}{2}\right), z\left(\frac{\alpha}{2}\right)\right]\right) \\ &= P\left(\overline{x}-\mu \in \left[-z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}, z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}\right]\right) \\ &= P\left(\mu-\overline{x} \in \left[-z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}, z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}\right]\right), \end{aligned}$$

since the interval is symmetric about zero, and this equals:

$$= P\left(\mu \in \left[\overline{x} - z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}}, \overline{x} + z\left(\frac{\alpha}{2}\right)\right]\right),$$

and we call this a $(1-\alpha)100\%$ confidence interval of μ.

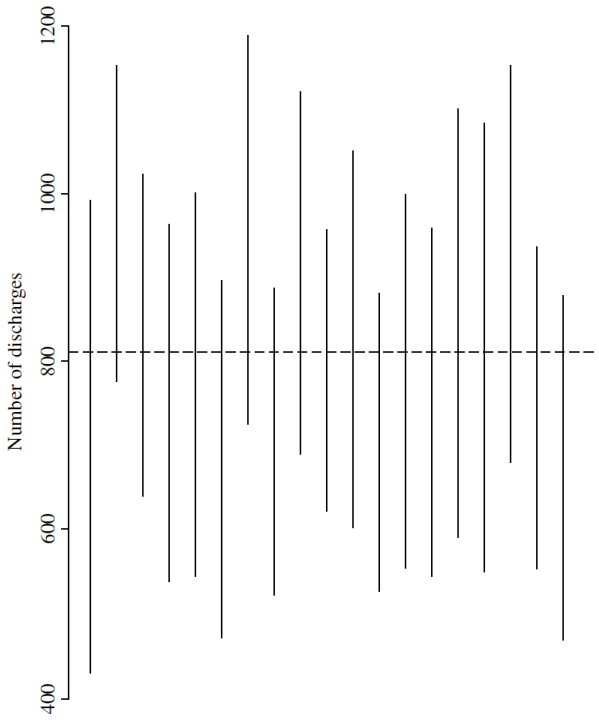


FIGURE **7.4** Vertical lines are 20 approximate 95% confidence intervals for μ. The horizontal line is the true value of μ.

The confidence interval is a variable. There is a 95% (95 out of 100) chance that our interval contains the true parameter μ. The width of all these are the same, but the center is different.

3.1 Example

Consider $N = 393$ hospitals, and let x_i be the number of patients discharged from the ith hospital. Take $\mu = 814.6$ and $\sigma = 590$. Take a simple random

sample (SRS) of $n = 50$. We showed:

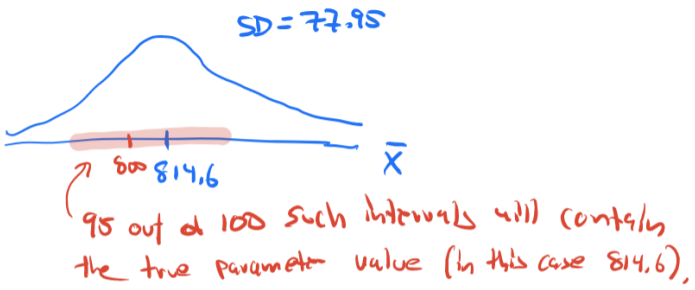
$$\sigma_{\overline{x}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = 77.95$$

Suppose $\overline{x} = 800$. Find the 95% CI for μ and interpret the result.

Solution. $\alpha = 0.05$, so take:

$$\overline{x} \pm z\left(\frac{\alpha}{2}\right)\sigma_{\overline{x}} = 800 \pm 1.96(77.95) = [647.2, 952.8],$$

and we can draw a picture of our distribution centered around 814.6, and that our confidence interval centered at 800 **contains** 814.6.



To interpret results, we can say that 95 out of 100 such intervals will contain 814.6. □

4 §7.3.1 Expectation, Variance of the Sample Mean

Definition: Unbiased Estimator -

We say that an estimator $\hat{p}$ of p is unbiased if $E(\hat{p}) = p$.

Example: If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, then $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of p since:

$$E(\overline{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n} \cdot np = p$$

Let $X_1, \ldots, X_n$ be iid with $E(x) = \mu$ and $\text{var}(x) = \sigma^2$. Show that

$$E\left(\overline{x}^2\right) = \frac{\sigma^2}{n} + \mu^2.$$

To see this, notice:

$$\text{var}(y) = E(y^2) - E(y)^2,$$

so

$$E\left(\overline{x}^2\right) = \text{var}(\overline{x}) + E\left(\overline{x}^2\right),$$

and because this is i.i.d., we have:

$$\text{var}(\overline{x}) = \text{var}\left(\frac{1}{n}\sum_{X_i}\right) = \frac{1}{n^2}n \cdot \text{var}(x) = \frac{1}{n}\sigma^2$$

We will use this in the the next lecture to prove a theorem that the sample variance is an unbiased estimator of the true population variance r^2.

Lecture ends here.

Stats 135, Fall 2019
Lecture 3, Wednesday, 9/4/2019

1 Review

Adam Lucas hinted that our last result from lecture 2 will be used in the next lecture for a proof, which briefly gloss over now.
Let $X_1, \dots, X_n$ be iid with $E(x) = \mu$, with $\text{Var}(x) = \sigma^2$.

Theorem 1.1. The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is an unbiased estimator of the true population variance r^2.

Proof. Note that $E(\sum x_i) = \sum E(x_i)$ and $E(cx) = cE(x)$. Recall that $\text{Var}(x) = x = E(x^2) - \underbrace{E(x)^2}_{\mu^2}$ which implies

$$E(x^2) = \sigma^2 + \mu^2.$$

Now, $\text{Var}(\overline{x}) = E(\overline{x}^2) - E(\overline{x})^2$ implies

$$E(\overline{x}^2) = \frac{\sigma^2}{n} + \mu^2. \tag{1}$$

We show that $E(S^2) = \sigma^2$. Consider:

$$\begin{aligned}
E\left(\sum(x_i - \overline{x})^2\right) &= E\left(\sum(x_i^2 - 2x_i\overline{x} + \overline{x}^2)\right) \\
&= E\left(\sum x_i^2 - \sum 2x_i\overline{x} + \sum \overline{x}^2\right) \\
&= E\left(\sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2\right) \quad \text{because } \overline{x} \text{ is a constant} \\
&= E\left(\sum x_i^2 - 2\overline{x}n\overline{x} + n\overline{x}^2\right) \quad \text{because } \sum x_i = n\overline{x} \\
&= E\left(\sum x_i^2 - n\overline{x}^2\right) \\
&= \sum E(x_i^2) - nE(\overline{x}^2) \\
&= \sum\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \quad \text{by (1) above} \\
&= n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2} \\
&= (n-1)\sigma^2
\end{aligned}$$

Hence

$$\frac{1}{n-1}E\left(\sigma(x_i - \overline{x})^2\right) = \sigma^2,$$

as required. □

Adam Lucas notes that he's not particularly interested in the proof, but we should be able to understand every step of it.
A question arises in the audience: why, intuitively is it $\frac{1}{n-1}$? Lucas notes that $1/n$ is a little bit too small, and to make it a little bit bigger, we use $1/(n-1)$. He says 'honestly it's just what we need to make it unbiased', and it simply appears from the algebra; he laments he does not have a super intuitive explanation for this outside of the algebraic steps of the proof.

2 §7.3.1: Expectation and Variance of Sample Mean

We now prove:

$$\text{Var}(\overline{x}) = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right],$$

if $x_1, \ldots, x_n$ are identically distributed with mean μ and variance σ^2.
We give a different proof from what was given in the book (we will follow Pitman page 441-443).

Proof. Assume $x_1, \ldots, x_n$ are SRS (simple random samples, without replacement), each with mean μ and variance r^2.
For example, x_i can be the annual income of the ith household in the US.
We take a look at the sum of these incomes:

$$T := x_1 + \cdots + x_n$$

and then

$$\text{Var}(T) = \text{Var}(x_1 + \cdots + x_n) = \sum_{i,j=1}^{n} \text{Cov}(x_i, x_j),$$

and we can break up this sum into where $i = j$ and $i \neq j$, where equality simply nets variance:

$$\text{Var}(T) = \sum_{j=1}^{n} \text{Var}(x_i) + \sum_{i \neq j} \text{Cov}(x_i, x_j).$$

Because all these x_i are identically distributed, this will simply give:

$$= \underbrace{n \cdot \text{Var}(x_i)}_{\sigma^2} + n(n-1)\text{Cov}(x_1, x_2),$$

and we employ a trick to find out this covariance. Recall that n is a sample of some population, so assume there are N of these (i.e. US households). The trick we use is to consider $n := N$ (our sample will be the entire population). In this case, the variance of T will simply be 0 because there will be no variation when we take our sample to be the entire population! So we have:

$$\text{Var}(T) = 0 \implies 0 = N\sigma^2 + N(N-1)\text{Cov}(x_1, x_2),$$

which implies

$$\text{Cov}(x_1, x_2) = \frac{-N\sigma^2}{N(N-1)} = \boxed{\frac{-\sigma^2}{N-1}}.$$

Now for general n, we have:

$$\text{Var}(T) = n\sigma^2 + n(n-1)\left(\frac{-\sigma^2}{N-1}\right) = n\sigma^2\left[\frac{N-n}{N-1}\right],$$

so

$$\text{Var}(\overline{x}) = \text{Var}\left(\frac{T}{n}\right) = \frac{1}{n^2}\text{Var}(T) = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right].$$

□

Remark: Also, we have:

$$E\left(\overline{x}^2\right) = \mathrm{Var}(\overline{x}) + E(\overline{x})^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right] + \mu^2$$

3 §7.3.2: Estimation of Population Variance

We showed that the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is an unbiased estimator of the true population variance r^2 (i.e. $E(s^2) = \sigma^2$).
However, this is only true if $x_1, \ldots, x_n$ are iid. More generally, if $x_1, \ldots, x_n$ is a SRS (simple random sample), then

$$\left(1 - \frac{1}{N}\right) S^2$$

is an unbiased estimator of r^2. We won't prove this in class, but see lecture notes.

Theorem 3.1. Let $x_1, \ldots, x_n$ be SRS with mean μ and variance σ^2. Then with the 'fudge factor' $\left(\frac{N-1}{N}\right)$, we have:

$$E\left[\left(\frac{N-1}{N}\right)\frac{1}{n-1}\left(\sigma(x_i - \overline{x})^2\right)\right] = \sigma^2.$$

With this, we essentially finish Chapter 7. The results we need to know are highlighted in Lucas' notes, from page 214 in Rice.

Population Parameter	Estimate	Variance of Estimate	Estimated Variance
μ	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$	$s^2_{\overline{X}} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$
p	$\hat{p}$ = sample proportion	$\sigma^2_{\hat{p}} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$	$s^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1 - \frac{n}{N}\right)$
τ	$T = N\overline{X}$	$\sigma^2_T = N^2\sigma^2_{\overline{X}}$	$s^2_T = N^2 s^2_{\overline{X}}$
σ^2	$\left(1 - \frac{1}{N}\right)s^2$		

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

4 Chapter 8: Estimation of Parameters, Fitting of a Probability Distribution

For motivation, the idea is that we start with observed and recorded data in the real world. We may believe our data set obeys some probability distribution (say Poisson) for a certain parameter θ. We wish to find an estimate for our parameter θ, and we usually write the estimate as $\hat{\theta}$ which has good properties: perhaps it is an unbiased estimator, or other properties.
The canonical example is radioactive decay. Suppose in an experiment we observe α-partical decay for 12,070 seconds. We break this time interval (axis) into 10 second intervals and count the number of arrivals in each interval.
Performing the experiment and counting the total of 10,129 α-particle arrivals, we let X be the number of arrivals in a 10-second interval. Then X is a good candidate for a Poisson distribution because α-particles obey the three properties that a Poisson process has. Let λ be the rate of arrival in 10 seconds.
(1) λ is constant (Americium has a long half-life)
(2) the numbers (counts) of arrivals in disjoint intervals are independent
(3) the arrivals do not coincide (no simultaneous arrivals)
Hence we can model this as 1207 iid Poisson(λ) random variables (RV). Recall that $X \sim \text{Poisson}(\lambda)$ implies:

$$\pi_k = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

We don't know λ, so we need to estimate. We're told there are 10,129 arrivals, so we take the average:

$$\hat{\lambda} := \frac{10,129 \text{ arrivals}}{1207 \text{ seconds}} = 8.392 \text{ arrivals / 10 sec intervals.}$$

We have 1207 independent intervals, so what is the probability that an interval gets 3 arrivals? This is simple: We know X is the number of arrivals and $X \sim \text{Poisson}(8.4)$, so:

$$p = P(X = 3) = \frac{e^{-8.4}(8.4)^3}{3!} = 0.022$$

Now we need to know the number of intervals that get 3 arrivals. The probability of getting 3 arrivals in any interval is given by the probability 0.022. We can think of this as a bunch of Bernoulli trials (either get 3 or not), where the chance of getting 3 is 0.022.
Let Y be the number of intervals that get 3 arrivals. Then we have:

$$Y \sim \text{Binomial}(1207, .022),$$

so

$$E(Y) = np = 1207 \cdot 0.022$$

using this, we fill out this table for expected counts:

n	Observed	Expected
0–2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1
	1207	1207

For example, to get $n = 4$, we have:

$$56.5 = \text{Binomial}\left(1207, \frac{e^{-8.4}(8.4)^4}{4!}\right)$$

In Chapter 9, we perform a χ-squared test to see how well our model fits the data. Next time we'll look at the method of moment estimating (§8.4).

Stats 135, Fall 2019
Lecture 4, Friday, 9/6/2019

Last time, we showed the highlighted part of the formula table. To wrap up, Chapter 7 was about the population mean, estimating μ, p, σ^2 and looking at the SE of them. Now in Chapter 8, we're going to generalize this. We have some data that fits a given distribution, and we need an estimator that should have some nice properties.
Further, we motivated why we estimate parameters of a probability model (as in the hospital Poisson example).

1 §8.4 Method of Moment (MOM) estimators

An estimator should converge to the true value (this is called **consistency**). There are different notions and definitions of convergence of a random variable (we will focus on the definition for Probability).
We'll first review the Gamma (Γ) distribution in the α-particle example. Suppose that $X \sim \text{Gamma}(\text{r}, \lambda)$, where X is the time to the rth arrival of a Poisson process. Here, r is the rth particle, and λ is the rate of arrival of α-particles in the Poisson process. This has a density that we should know:

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x},$$

and recall:

$$\Gamma(r) = (r-1)!, \ r \in \mathbb{Z}^+$$

and

$$\mathbb{E}(x) = \frac{r}{\lambda}$$
$$\text{Var}(x) = \frac{r}{\lambda^2}.$$

Now for Method of Moment estimators (MOM), if we want to estimate l parameters $(\theta_1, \ldots, \theta_l)$ of a probability distribution $f(x \mid \theta_1, \ldots, \theta_l)$ from iid sample $x_1, \cdots, x_n$ from this distributions, there are 3 steps:

1.1 (Step 1)

We compute the first l moments (where moments are the kth expectation):

$$\mu_k = \mathbb{E}(x^k), \quad k = 1, \ldots, l.$$

Now the RHS is given by the integral:

$$\mu_k = \int_{-\infty}^{\infty} x^k f(x \mid \theta_1, \ldots, \theta_l) \, dx.$$

This does depend on what these θ_i are. For i in `1 : k`, we'll say these μ_i are functions g_i on the arguments θ_j. That is, we have the family of equations:

$$\mu_1 = g_1(\theta_1, \ldots, \theta_l)$$
$$\mu_2 = g_2(\theta_1, \ldots, \theta_l)$$
$$\vdots$$
$$\mu_l = g_l(\theta_1, \ldots, \theta_l).$$

Example: Let $X \sim \text{Poisson}(\lambda)$, with $l = 1$ and let X be the number of arrivals in 10 second intervals. The average rate of arrivals is just λ:

$$\mu_1 = \mathbb{E}(x) = \lambda.$$

Example: Suppose $X \sim \text{Gamma(r}, \lambda)$ now with $l = 2$. Then,

$$\begin{aligned}\mu_1 &= \mathbb{E}(x) = \frac{r}{\lambda} \\ \mu_2 &= \mathbb{E}(x^2) = \underbrace{\text{Var}(x)}_{r/\lambda^2} + \underbrace{\mathbb{E}(x)^2}_{(r/\lambda)^2} = \frac{r + r^2}{\lambda^2}.\end{aligned}$$

1.2 (Step 2)

Now we use algebra to invert the above system of equations (require h to be a continuous function of $\mu_1, \ldots, \mu_l$):

$$\begin{aligned}\theta_1 &= h_1(\mu_1, \ldots, \mu_l) \\ \theta_2 &= h_2(\mu_1, \ldots, \mu_l) \\ &\vdots \\ \theta_l &= h_l(\mu_1, \ldots, \mu_l)\end{aligned}$$

Example: In the Poisson case, then we simply have:

$$\mu_1 = \lambda \implies \lambda = \mu_1.$$

We wrote μ_1 in terms of the parameter, and in step 2 we wrote the parameter in terms of the moment. Done!

Example: In the Gamma case, we have:

$$\begin{aligned}\mu_1 &= \frac{r}{\lambda} \\ \mu_2 &= \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2} = \frac{\mu_1}{\lambda} + \mu_1^2 \\ \implies \frac{\mu}{\lambda} &= \mu_2 - \mu_1^2,\end{aligned}$$

so this gives:

$$\begin{aligned}\lambda &= \frac{\mu_1}{\mu_2 - \mu_1^2} \\ r = \lambda\mu_1 &= \frac{\mu_1^2}{\mu_2 - \mu_1^2}\end{aligned}$$

1.3 (Step 3)

Now we insert into $(*)$ the estimator for the moments $\mu_1, \ldots, \mu_l$. We call these **sample moments**.

The first moment is the mean, so the first sample moment is the sample mean:

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_1 \\ \hat{\mu}_2 &= \frac{1}{n}\sum_{i=1}^{n} x_1^2 \\ \mathbb{E}(x^2) \qquad &\vdots \\ \hat{\mu}_l &= \frac{1}{n}\sum_{i=1}^{n} x_i^l. \end{aligned}$$

We will show in homework that these are unbiased estimators of $\mu_1, \ldots, \mu_l$.
Now we have:

$$\begin{aligned} \hat{\theta}_1 &= h_1(\hat{\mu}_1, \ldots, \hat{\mu}_l) \\ \hat{\theta}_2 &= h_2(\hat{\mu}_1, \ldots, \hat{\mu}_l) \\ &\vdots \\ \hat{\theta}_l &= h_l(\hat{\mu}_1, \ldots, \hat{\mu}_l), \end{aligned}$$

where we essentially just replace the non-hats with hats. We call these the **MOM estimators** for $\theta_1, \ldots, \theta_l$. In practice, these are usually not the best estimators, but they are simple (just algebraic) so we talk about it now.
For example, we have:
Poisson:

$$\lambda = \mu_1 \implies \hat{\lambda} = \hat{\mu}_1 = \hat{x}$$

and for Gamma:

$$\begin{aligned} \hat{\lambda} &= \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \\ \hat{r} &= \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \end{aligned}$$

2 Showing MOM estimators are Consistent

This is the most fundamental property that we want, which is to say that

$$\hat{\theta}_{MOM} \xrightarrow{p} \theta,$$

where we write p to mean convergence in the probability sense.
We first take a short digression to prove the **Weak Law of Large Numbers**. Our book doesn't go into this, so Lucas wants us to have this little missing piece.
We want to have Markov's inequality:

Theorem 2.1. (Markov's Inequality) :
For $x \geq 0$, $c > 0$, then we the tail probability gives:

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}.$$

We won't prove this in lecture (Adam diverts the proof to Pitman).

Example: Let $x_1, \ldots, x_n$ be iid with mean μ and variance σ^2. Then the sample mean is:

$$\overline{x}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and note that we know this is unbiased and we know the variance:

$$\mathbb{E}(\overline{x}_{(n)}) = \mu, \qquad \text{Var}(\overline{x}_{(n)}) = \frac{\sigma^2}{n}.$$

Let $\epsilon > 0$. Use Markov's inequality to give an upper bound for

$$\mathbb{P}(\underbrace{|\overline{x}_{(n)} - \mu|}_{\text{nonrandom var}} \geq \underbrace{\epsilon}_{c>0})$$

So we have:

$$\begin{aligned}
\mathbb{P}(|\overline{x}_{(n)} - \mu| \geq \epsilon) &= \mathbb{P}[(\overline{x}_n - \mu)^2 \geq \epsilon^2] \\
&\leq \frac{\mathbb{E}\left[(\overline{x}_{(n)} - \mu)^2\right]}{\epsilon^2} \\
&= \frac{\text{Var}(\overline{x}_{(n)})}{\epsilon^2} \\
&= \frac{\sigma^2}{n\epsilon^2},
\end{aligned}$$

where in the first line we square both sides because both are positive, and we notice we have n in the denominator, so taking a distribution of $\overline{x}_{(n)} - \mu$ is centered at 0 and the tail area gets smaller as $n \to \infty$. More precisely, the area is bound by:

$$\frac{\sigma^2}{n\epsilon^2} \to 0, \text{ as } n \to \infty.$$

Definition: Convergence in Probability -

In friendly words, this is the idea that the probability of an unusual outcome gets smaller and smaller as n grows.
To say that one random variable converges in probability to another, as n grows larger, it will be very unlikely that their difference will be any greater, than say ϵ.
For example, take a sequence $X_1, X_2, \ldots$ of random variables. We say this sequence converges in probability to a random variable X if $\forall_{\epsilon>0}$, the probability:

$$\mathbb{P}(|x_n - x| > \epsilon) \to 0, \text{ as } n \to \infty.$$

Example of Weak Law of Large Numbers: The Weak Law of Large Numbers says that for $x_1, \ldots, x_n$ iid with mean μ and variance σ^2, we have:

$$\overline{x}_{(n)} \xrightarrow{p} \mu,$$

where μ is the constant random variable that takes the value μ with probability 1.

This follows directly from Markov's inequality.
We already derived:

$$\mathbb{P}(|\overline{x}_{(n)} - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 \cdot n} \to 0, \text{ as } n \to \infty.$$

This proves the Weak Law of Large numbers. We can generalize this to higher moments. Take $x_1, \ldots, x_n$ iid with mean μ and variance σ^2. This says:

$$\hat{\mu}_k \xrightarrow{p} \mu_k,$$

where

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} x^k, \text{ and } \mu_k = \mathbb{E}(x^k).$$

Now we want to show consistency.

Definition: Consistency -

An estimator $\hat{\theta}$ of a parameter θ is **consistent** if $\hat{\theta} \xrightarrow{p} \theta$.

We argue that MOM is consistent. Adam Lucas notes that there is a highly-believable theorem:

Theorem 2.2. Suppose random variables $x_1, x_2, \ldots$ converge in probability to a random variable x and h is a **continuous** function. Then $h(x_1), h(x_2), \ldots$ converge in probability to $h(x)$.

Lucas states that to make our weekend complete, consider that if h is continuous (as in the Method of Moments case), then the estimator is **consistent**. In other words, our MOM estimator:

$$\hat{\theta}_{MOM} = h(\hat{\mu}_1, \ldots, \hat{\mu}_l) \xrightarrow{p} \theta = h(\mu_1, \ldots, \mu_l),$$

which is very clear to see.

Lecture ends here.

Stats 135, Fall 2019
Lecture 5, Monday, 9/9/2019

CLASS ANNOUNCEMENTS: Quiz 1 on Friday. There will be 3 questions total: 2 problems similar to those in HW 1 and 2, and 1 MOM calculation.

1 Review

Last time, in §8.4, we found that MOM estimators are **consistent** when h is continuous. That is,

$$\hat{\theta}_{MOM} = h(\hat{\mu}_1, \ldots, \hat{\mu}_l).$$

Because the sample moment converges in probability to the true moment,

$$\underbrace{\hat{\mu}_k}_{\frac{1}{n}\sum_{i=1}^n x^k} \xrightarrow{p} \underbrace{\mu_k}_{\mathbb{E}(X^2)}$$

(that is the sample parameter converges in probability to the true parameter) via generalized weak law of large numbers, we have:

$$\hat{\theta} = h(\hat{\mu}_1, \ldots, \hat{\mu}_l) \xrightarrow{p} \theta = h(\mu_1, \ldots, \mu_l)$$

when h is continuous.

Topics Today:

- Example of MOM calculation
- p 264 : nonparametric bootstrap for 95% confidence interval (done in R last Friday in lab)
- p 262 : finding SE of $(\hat{\theta})$ by hand

2 §8.4 MOM

Recall that the density of Gamma is given by:

$$f(x) = \frac{\lambda^r}{\Gamma(r)} \underbrace{x^{r-1}e^{-\lambda x}}_{\text{variable}},$$

where $\Gamma(r) = r - 1$ when $r \in \mathbb{Z}^+$.

$$\int_0^\infty f(x)\, dx = 1 \implies \int_0^\infty \frac{\lambda^r}{\Gamma(r)} x^{r-1}e^{-\lambda x}\, dx = 1$$
$$\implies \boxed{\int_0^\infty x^{r-1}e^{-\lambda x}\, dx = \frac{\Gamma(r)}{\lambda^r}},$$

which follows from a useful identity when r is an integer (so we don't need integration by parts).
Consider an iid sample of random variables with density

$$f(x|\sigma) = \frac{1}{2\sigma} e^{\left(\frac{-|x|}{\sigma}\right)}, \quad \sigma > 0.$$

We want to find the MOM estimator $\hat{\sigma}$ of σ, so we calculate the first moment $\mathbb{E}(x)$. That is,

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} xf(x|\sigma)\ dx = \int_{-\infty}^{\infty} \overbrace{x\frac{1}{2\sigma}e^{(-\frac{|x|}{\sigma})}}^{g(x)}\ dx = 0$$

where $g(-x) = -g(x)$ shows g is an odd function, and hence this integral equals zero. More precisely,

$$\int_{-\infty}^{0} xf(x|\sigma)\ dx = -\int_{0}^{\infty} xf(x|\sigma)\ dx$$

Our function is 0, so this isn't helpful. We try the next moment, μ_2, $\mathbb{E}(x^2)$. Lucas tasks us to compute this.

$$\begin{aligned}
\mathbb{E}(x^2) &= \int_{-\infty}^{\infty} x^2 f(x|\sigma)\ dx \\
&= \int_{-\infty}^{\infty} x^2 \frac{1}{2\sigma} e^{(-\frac{|x|}{\sigma})}\ dx \\
&= 2\frac{1}{2\sigma}\int_{0}^{\infty} x^2 e^{-\frac{1}{\sigma}x}\ dx \\
&= \frac{1}{\sigma}\cdot\frac{\Gamma(3)}{(1/\sigma)^3} = \boxed{2\sigma^2}.
\end{aligned}$$

Here we used the boxed formula we found above, $\int_0^\infty x^{r-1}e^{-\lambda x}\ dx = \frac{\Gamma(r)}{\lambda^r}$, with $r = 3$ and $\lambda = \frac{1}{\sigma}$.
Recall that the second step to this Method of Moments calculation is to rewrite the parameter in terms of the moments. That gives us:

$$\sigma = \sqrt{\frac{\mu_2}{2}},$$

and for Step 3, we put in the sample moment (put hats on) to get:

$$\hat{\sigma} = \sqrt{\frac{\hat{\mu}_2}{2}},$$

where $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

3 Nonparametric Bootstrapping and Confidence Interval of Population Parameter

See p284-285 in Rice. Recall that to calculate the 95% CI of a population mean **by hand**, we need two things:

- a Simple Random Sample (We need an SRS so that we know $\text{Var}(\overline{x})$, and we have a formula for that. It may be unrealistic to assume sampling without replacement.)
- a large sample size (so that the distribution is approximately normal so that we know the 2.5th and 97.5th quantiles of $\overline{x}$.)

However, it is often the case that the sample is small and perhaps we don't have an SRS. Instead, we perform a bootstrap technique. Instead, we can find a 95% CI by bootstrapping in R **without any assumptions on the sampling distribution**, which is useful.

To talk about this, Lucas gives some background on the 95% CI of a population parameter. We did this in detail for the mean before, and now we want to generalize. Let θ be a (generalized) parameter of our population (for example it could be the population average, population median, or the population SD, etc).
Let $\hat{\theta}$ be an estimator of θ. Of course, $\hat{\theta}$ is a random variable, and θ is an unknown constant. We look at:

$$\hat{\theta} - \theta,$$

which is also a random variable. Its distribution is **not necessarily normal**. Suppose we want to compute points a, b that give the 2.5th and 97.5th percentiles of $\hat{\theta} - \theta$. In other words,

$$\mathbb{P}(a < \hat{\theta} - \theta < b) = 95\%.$$

As justification for this formatting, Lucas notes that we can compute $a, b, \hat{\theta}$ from our sample. It only takes a little algebra to get:

$$\begin{aligned} &\iff \mathbb{P}(-\hat{\theta} + a < -\theta < -\hat{\theta} + b) = 95\% \\ &\iff \mathbb{P}(\hat{\theta} - b < \theta < \hat{\theta} - a) = 95\%, \end{aligned}$$

which we call the 95% CI of θ. This distribution of $\hat{\theta} - \theta$ is only known to 'Tyche', the god of fortune (we don't know anything about θ), so there is no way for us to know what a, b are. Lucas notes that this is okay, because we are going to make a simulation (nonparametric bootstrap) and estimate these. What this is is that we'll take a sample once from our population and compute our estimator for that sample, and set this value to θ. Once we have this, we will sample from that same sample (same size) "many many times" **with replacement** to fill out our distribution. We can approximate $\hat{\theta} - \theta$ via bootstrapping:

4 Bootstrapping a 95% CI of θ:

(Step 1) Take a sample $x_1, \ldots, x_n$ of size n one time from your population. We will calculate $\hat{\theta}$ (which is θ above).

(Step 2) Now we **resample** from this same sample (size n), say $B = 1000$ times, but now **with replacement**. Now we compute the estimator of θ each time. This will give us a list of $B = 1000$ numbers, $\theta_1^*, \theta_2^*, \ldots, \theta_B^*$.

(Step 3) Now we subtract $\hat{\theta}$ (from step 1) from each θ_i^*, which gives us the list:

$$\theta_1^* - \hat{\theta},\ \theta_2^* - \hat{\theta},\ \ldots,\ \theta_B^* - \hat{\theta}$$

This is $\hat{\theta} - \theta$ above. Lucas apologizes there are multiple different $\hat{\theta}$ 'running around'. This will fill our distribution for $\hat{\theta} - \theta$ (and we can sketch the graph).

(Step 4) Now we find the 25th and 975th largest values from your (ordered) list of 1000 resulting numbers in step 3. This is an approximation of the 2.5th and 97.5th percentile of $\hat{\theta} - \theta$.

(Step 5) The 95% CI of θ then is:

$$(\hat{\theta} - b, \hat{\theta} - a),$$

where $\hat{\theta}$ is our original sample estimator of θ, and a is our 25th largest number in the list of $\theta_i^* - \hat{\theta}$, and b is our 975th largest. The question arises for a rule of thumb for the value of B, and Lucas holds off answering because we'll talk about why this method works and produces good results.

5 §8.4: SE of $\hat{\theta}$

Recall that θ is a parameter **determining** the distribution of x. So the $SD(x) = \theta$ is a continuous function of θ, so we can write $\sigma(\theta)$.
For example, if $x \sim \text{Poisson}(\lambda)$, then the $SD(\lambda) = \sqrt{\lambda}$, and so:

$$\sigma(\lambda) = \sqrt{\lambda},$$

which is a continuous function. Now because σ is continuous, then we've found that

$$\hat{\theta} \xrightarrow{p} \theta \implies \sigma(\hat{\theta}) \xrightarrow{p} \sigma(\theta).$$

So for large n, unknown $\sigma(\theta)$ is very closely approximated by $\sigma(\hat{\theta})$, which is known. We don't know the true SD of x, but we can bootstrap an estimator.

Lecture ends here.

We'll look at an example of this at the start of Wednesday's lecture.

Stats 135, Fall 2019
Lecture 6, Wednesday, 9/11/2019

CLASS ANNOUNCEMENTS: Quiz 1 on Friday will have 3 problems: 2 problems similar to HW1,2 or up to lecture 5. There will be 1 problem of MOM calulation.

1 Review:

Last time, we covered §8.4. We showed how to compute SE by hand. To find $\text{Var}(\hat{\theta})$, we often need to know $\sigma^2 = \text{Var}(x)$ which is a function of θ. We have $\sigma^2(\theta) \approx \sigma^2(\hat{\theta})$.

Example:
Given an iid sample, we collect data:

$$x_1 4, x_2 = 7, x_3 = 4, x_4 = 2, x_5 = 3,$$

which follows a Poisson(λ) distribution. Then we find a MOM estimator of λ and approximate the SE of our estimate.

Solution.

$$\begin{aligned} X &\sim \text{Poisson}(\lambda) \\ \mu_1 &= \mathbb{E}(X) = \lambda \\ \lambda = \mu_1 &\implies \hat{\lambda} = \overline{x} = 4 \end{aligned}$$

Using properties of the Poisson distribution (variance is simply λ), we have:

$$\begin{aligned} SE(\hat{\lambda}) &= \sqrt{\text{Var}(\hat{\lambda})} \\ &= \sqrt{\text{Var}(\overline{x})} \\ &= \sqrt{\frac{\text{Var}(x)}{n}} \\ &= \sqrt{\frac{\lambda}{n}} \\ &\approx \sqrt{\frac{\hat{\lambda}}{n}} \\ &= \sqrt{4/5}. \end{aligned}$$

Alternatively, at the 3rd equality we could have taken the sample variance instead of $\hat{\lambda}$ as an estimate of $\text{Var}(x)$ anove. That is,

$$S^2 = \cdots$$

□

Topics Today:

- Empirical cdf
- §8.4 Example
- §8.4, 8.4.6.

2 Empirical CDF (p 378 Rice)

This is actually in Chapter 10, but we talk about it briefly for justification of the bootstrap method.

(insert example here)

Facts about ECDF.
(1) F_n is an unbiased estimator of F.
(2) $\text{Var}(F_n) \to 0$ as $n \to \infty$.
So for large sample size, the empirical cdf is a good approximation of the population cdf. Now when we bootstrap, we take from the 'staircase' as opposed to the smooth curve population.

3 Another example of MOM estimator and computing the SE

This is #4ab from Chapter 8. Suppose X is a discrete random variable with:

$$\begin{aligned}
\mathbb{P}(X=0) &= \frac{2}{3}\theta \\
\mathbb{P}(X=1) &= \frac{1}{3}\theta \\
\mathbb{P}(X=2) &= \frac{2}{3}(1-\theta) \\
\mathbb{P}(X=3) &= \frac{1}{3}(1-\theta).
\end{aligned}$$

Then the following 10 iid observations are taken, giving us:

$$3, 0, 2, 1, 3, 2, 1, 0, 2, 1.$$

We are tasked to find the MOM estimator of θ and approximate the SE of our estimate. We have one parameters, so our first step is to compute the first moment:

$$\begin{aligned}
\mathbb{E}(X) &= 0 \cdot \mathbb{P}(X=0) + 1 \cdot \mathbb{P}(X=1) + 2 \cdot \mathbb{P}(X=2) + 3 \cdot \mathbb{P}(X=3) \\
&= \frac{1}{3}\theta + 2 \cdot \frac{2}{3}(1-\theta) + 3 \cdot \frac{1}{3}(1-\theta) \\
&= \frac{7}{3} - 2\theta
\end{aligned}$$

This implies:

$$\begin{aligned}
\theta &= \frac{1}{2}\left(\frac{7}{3} - \mu_1\right) \\
\hat{\theta} &= \frac{1}{2}\left(\frac{7}{3} - \underbrace{\hat{\mu}_3}_{=\overline{X}=\frac{3}{2}}\right) = \frac{5}{12}.
\end{aligned}$$

Next we find the variance $\text{Var}(\hat{\theta})$. From our finding earlier,

$$\begin{aligned}
\text{Var}(\hat{\theta}) &= \text{Var}\left(\frac{1}{2}\left(\frac{7}{3} - \overline{X}\right)\right) \\
&= \frac{1}{4}\text{Var}(\overline{X}) \\
&= \frac{1}{4}\frac{\text{Var}(X)}{10}
\end{aligned}$$

Recall that in our previous example, we had a Poisson distribution, and we knew the variance is just λ. Now here we have an unknown distribution, and we don't know the variance of the top of our head. There are two ways to find the variance.
(1) Approximate $\text{Var}(X)$ by s^2. Taking this approach, we get that $SE(\hat{\theta}) = .171$ (done in R).

```
> sqrt( 1/(4*10) * var( c(3,0,2,1,3,2,1,0,2,1)) )
[1] 0.1707825
```

(2) Analytically compute $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, in terms of θ. Lucas jokes that he takes out his Pitman book for this. Here we get that $SE(\hat{\theta}) = .173$.

4 Yet Another Example

Let $X \sim \text{Exponential}(\lambda)$. Fact: $\mathbb{E}(X) = \frac{1}{\lambda}$. We are tasked to find $\hat{\lambda}$ and $SE(\hat{\lambda})$.
Then

$$\mu_1 = \frac{1}{\lambda} \implies \lambda = \frac{1}{\mu_1} \implies \boxed{\hat{\lambda} = \frac{1}{X}}$$

Here we don't have the variance of a linear combination of μ_1, so here we need the δ-method instead.

5 δ-method

Consider a random variable X with a known mean μ and known SD σ. We have some function $Y = g(X)$, smooth around the mean, for example it can be $1/X$. That is, $g'(\mu) \neq 0$. The idea comes from Taylor expansion. Taylor had the idea that all smooth functions are locally linear. We'll make a linear Taylor approximation around μ (truncation) here. Take:

$$Y = g(X) \approx g(\mu) + g'(\mu)(X - \mu),$$

and this is a good approximation when $(X - \mu)$ is very small (and hence exponentials of them, the truncated terms, are even smaller). In other words, this is a good approximation when $X \approx \mu$.
If we can write it this way, then applying Var across the equation gives:

$$\begin{aligned}\text{Var}(Y) &\approx \text{Var}\left(g(\mu) + g'(\mu)(X - \mu)\right) \\ &= \left(g'(\mu)\right)^2 \underbrace{\text{Var}(X)}_{\sigma^2}.\end{aligned}$$

Now we may ask: what random variables have this property? That is, what random variables have a small SD? Surely, the normal or uniform distributions won't work. Instead, take $\overline{X}$ with large n, which Lucas notes is very 'pointy' around the mean.

Theorem 5.1. (δ-method). Let $X_1, \ldots, X_n$ be iid with mean μ and $SD = \sigma$. Take g to be smooth μ, where $g'(\mu) \neq 0$. Then

$$\text{Var}(g(\overline{X})) \approx \left(g'(\mu)\right)^2 \cdot \frac{\sigma^2}{n}$$

Now for us, take: $\hat{\theta} := g(\overline{X})$. Let's finish the earlier example.

Example: $X \sim \text{Exponential}(\lambda)$. Then the estimator is $\hat{\lambda} = \frac{1}{\overline{X}}$, so let:

$$g(\overline{X}) := \frac{1}{\overline{X}},$$

in the δ-method. By the theorem, this method gives

$$\text{Var}(g(\overline{X})) \approx (g'(\mu))^2 \cdot \frac{\sigma^2}{n}.$$

All we need to do is compute the derivative and plug it in, evaluated at μ.

$$g'(X) = \frac{-1}{X^2} \implies (g'(\mu))^2 = \left(\frac{-1}{\mu^2}\right)^2 = \lambda^4,$$

where we used $\mu = \frac{1}{\lambda}$.
So the variance of our MOM estimator is:

$$\text{Var}(\hat{\lambda}) = \text{Var}(g(\overline{X})) = \lambda^4 \cdot \frac{\frac{1}{\lambda^2}}{n}.$$

We don't know λ, so we plug in $\hat{\lambda} = \frac{1}{\overline{X}}$. Then

$$SE(\hat{\lambda}) \approx \frac{\lambda}{\sqrt{n}} \approx \frac{\frac{1}{\overline{X}}}{\sqrt{n}} = \frac{1}{\sqrt{n} \cdot \overline{X}}.$$

Example: This one is problem 52 from Chapter 8, and it is a bit more complicated. Let's say we have $X_1, \ldots, X_n$ iid random variables with density $f(X|\theta) = (\theta + 1)X^\theta$, where $0 \leq X \leq 1$. We're tasked to find $\hat{\theta}$ and use the δ-method to approximate $SE(\hat{\theta})$.

Solution. First, we find our MOM estimator $\hat{\theta}$. We have only one parameter, so we need only the first moment:

$$\mathbb{E}(X) = \int_0^1 (\theta + 1)X^{\theta+1}\, dx = \frac{\theta + 1}{\theta + 2} X^{\theta+2}|_0^1 = \frac{\theta + 1}{\theta + 2}$$

Now this is our $\mu = \frac{\theta+1}{\theta+2}$. We do some algebra on

$$\begin{aligned} \mu\theta + 2\mu &= \theta + 1 \\ \mu\theta - \theta &= 1 - 2\mu \\ \theta(\mu - 1) &= 1 - 2\mu \\ \theta &= \frac{1 - 2\mu}{\mu - 1} \end{aligned}$$

Now plugging in $\mu = \overline{X}$ gives:

$$\hat{\theta} = \frac{1 - 2\overline{X}}{\overline{X} - 1}.$$

Notice this is a function of $\overline{X}$, so we should be able to use the δ method. We'll do this during the next lecture. $\square$