Stats 135, Fall 2019

Lecture 1, Wednesday, 8/28/2019

Topics Today:

- Announcements
- Syllabus
- Ice Breaker
- Introduction
    1. Parameter Estimation and Standard Error
    2. Sampling Distribution

CLASS ANNOUNCEMENTS:

Lab starts this Friday, and it will be focused on R. The assignment for Friday is to make sure to get R and RStudio working before the first lab. `library('datacomputing')`

OH: MWF 9-10AM in SLC in the large room, Atrium. Large on emails, so feel free to email alucas(at)berkeley.edu

134 is a prerequisite, and 133 is a corequisite (R)

Familiarity with linear algebra (matrix operations, inverse of a matrix, possibly eigenvalues) will be necessary at the end of the course (chapter 14, multiple regression).

Familiary with Moment Generating Functions as covered in Stats 134.

The textbook assumes familiarity with multivariable calculus in particular Lagrange Multipliers (see Khan Academy for the necessary background.)

Textbook: John Rice, Mathematical Statistics and Data Analysis, 3rd Edition. We will cover the second-half of this textbook, from Chapter 7,8,9,11,12,13,14.

Lecture notes and summary video is on b-courses/pages after lecture. These will be no longer than 15 minutes, which will be simply the main points.

Grading: 4 Quizzes (in Section, Sep 13, Sep 27, Nov 1, Nov 15) and **1 Midterm (Oct 16)**. Piazza participation (top 10) up to 1% for participation.
Grading:

- 25% Midterm, Clobbered
- 40% Final
- 15% weekly assignments, drop lowest
- 20% section quizzes, drop lowest

The distribution will be something like 30% A, 30% B, 30% C.

We break out into ice-breakers to discuss the relationship between **population** and **sample**. We say that from a sample, we want to infer something about the population. We learn probability first to give us a language to **inverse** the process of taking a sample.

1 Introduction: Parameter Estimation, Standard Error (SE)

We'll start with Chapter 7 and move super fast through it. We'll focus on Parameter Estimation and Standard Error (SE).

Example: Say that a coin lands heads with probability p. It is tossed 100 times and lands heads 45 times. What can we say about p?

Solution. We can estimate p with $\hat{p}$ and find a **standard error** for $\hat{p}$. We assign

$$\begin{aligned} 1-p &\mapsto 0 \\ p &\mapsto 1 \end{aligned}$$

Take $x_1, \ldots, x_{100}$ and we an write:

$$\hat{p} = \overline{x} = \frac{1}{100}\sum_{i=1}^{100} x_i = \frac{45}{100} = .45.$$

We say that $\hat{p}$ is a random variable (RV), and it has a distribution called the **sampling distribution**.
We can make a picture of this distribution (a la Normal distribution). Our distribution is a sum (average) of expectations, so by the Central Limit Theorem, we can expect the distribution to be approximately normal.
The center of this distribution has:

$$E(\hat{p}) = E\left(\frac{1}{100}\sum_{i=1}^{100} x_i\right)$$

To get an **unbiased estimator**, we take:

$$\text{dist } \hat{p} = \frac{1}{100} \cdot 100 \underbrace{E(x_i)}_{p} = p$$

□

1.1 Standard Error

What is the Standard Deviation (SD) of $\hat{p}$? We call this the Standard Error (SE) of $\hat{p}$.

$$\text{Var}\left(\hat{p}\right) = \text{Var}\left(\frac{1}{100}\sum_{i=1}^{100} x_i\right) = \left(\frac{1}{100}\right)^2 100\text{Var}(x_i)$$

X is just a draw from a box, so it takes on a Bernoulli distribution:

$$\begin{aligned} x &\sim \text{Bernoulli}(p) \\ \text{Var}(p) &= p(1-p) \\ (\text{ SE of } \hat{p})^2 &= \frac{p(1-p)}{100} \end{aligned}$$

However, we don't know p, so we can try to:

- find the SE analytically (exactly)
- approximate the SE

We expect $p(1-p)$ to be a downward-facing parabola with zeros at 0,1 and has its max at 0.5. This tells us something about the shape of the SE.
We'll say:

$$p(1-p) \leq \frac{1}{4} \implies \text{SE of } \hat{p} \leq \sqrt{\frac{1/4}{100}} = .05,$$

and we call this a **conservative estimate** of $\hat{p}$ (an upper bound). Another way to approximate the SE is called the **bootstrap estimate**. If our sample is extremely (sufficiently) large, then our approximation is close to our true probability and our sample is a good representation of the population.
In conclusion:

$$\text{SE of } \hat{p} \approx \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{100}} = \frac{\sqrt{(.45)(.55)}}{\sqrt{100}} = .0497$$

and to summarize, we're looking at a **dichotomous case (box of 0,1)**. We take p to be a proportion of 1 in the box. We draw a sample size n $(x_1, \ldots, x_n)$ with N numbers, with $\hat{p} = \overline{x}$. Then the bootstrap estimate SE of $\hat{p}$ is

$$\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}},$$

and as a conservative SE of $\hat{p} = \frac{.5}{\sqrt{n}}$.
What happens if we draw from the box without replacement? We call this case a **Simple Random Sample, SRS**. Now, $x_1, \ldots, x_n$ are **dependent**, so this messes up our previous move of distributing the variance across all variables. To deal with this, we use a **correction factor**.

Definition: Correction Factor -

$$\frac{N-n}{N-1}$$

The thing to note is that if $N >> n > 1$, then this correction factor is close to 1.

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

See the table in page 214 of Rice.

Example: Consider a box of 4 0's and 1 1. We say that the probability of drawing 1 is $p = \frac{1}{5}$. Our variance is then:

$$r^2 = p(1-p) = \frac{1}{5} \cdot \frac{4}{5} = \frac{4}{25}$$

Suppose we draw two without replacement. Then our estimator is just the proportion of 1s in our sample:

$$p := \overline{x}$$

Adam's task for us is to list all samples of size 2, and for each, record the proportion of 1s in each sample. Call this $\hat{p}$.

Solution. We have $\binom{5}{2}$ samples of 2, so we'll make a histogram of 10 elements.
First case: No 1s. There are $\binom{4}{2}$ ways to do this without replacement, and all have $\hat{p} = 0$ (no 1s).
Adam gives:

$$E(\hat{p}) = 0 \cdot \frac{6}{10} + \frac{1}{2} \cdot \frac{4}{10} = \frac{1}{5}$$
$$\text{Var}(\hat{p}) = E(\hat{p}^2) - E(\hat{p})^2 = 0^2 \left(\frac{6}{10}\right) + \left(\frac{1}{2}\right)^2 \frac{4}{10}$$

□

Lecture ends here.

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.

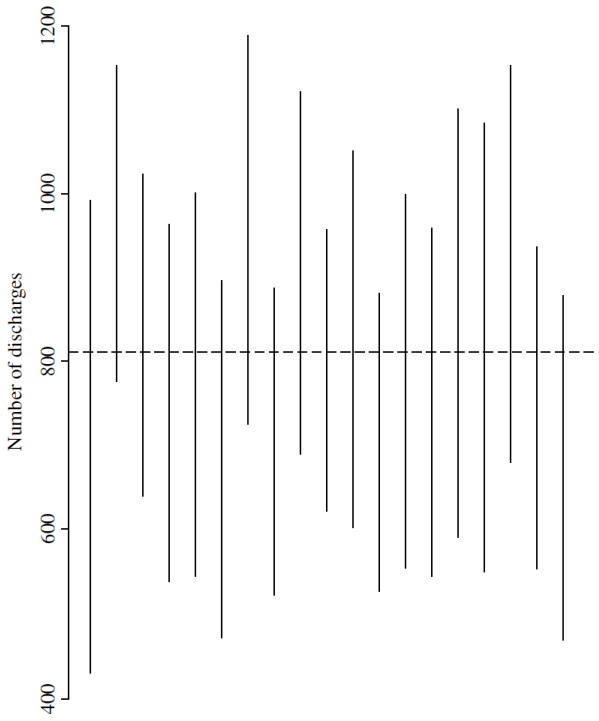


FIGURE **7.4** Vertical lines are 20 approximate 95% confidence intervals for μ. The horizontal line is the true value of μ.

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:

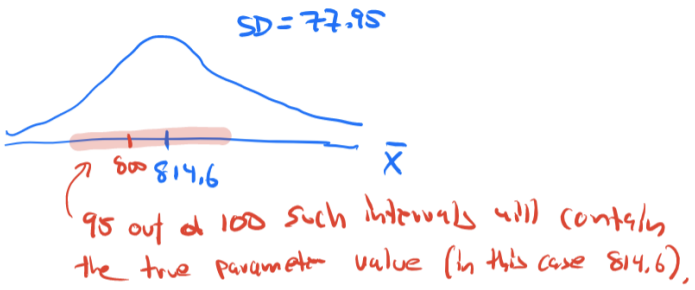
$$\sigma_{\overline{x}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = 77.95$$

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

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

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

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

To show that the sample variance S^2 is an unbiased estimator, 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 notes posted by Lucas.

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 the tail probability has the bound:

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

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 $\text{Var}(x)$ above. 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. $\square$

Stats 135, Fall 2019

Lecture 7, Friday, 9/13/2019

CLASS ANNOUNCEMENTS: RStudio Lab demo today in-class, and Quiz today in lab section.

Stat 135 lec 7

Last time

Sec 4.6 Delta Method

Thm (δ-method)

$X_1, \ldots, X_n$ iid mean μ, SD $= \sigma$

g smooth around μ, $g'(\mu) \neq 0$

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

ex* *St 2 chap 6

Let $X_1, \ldots, X_n$ be iid RVs w/ density $f(x|\theta) = (\theta+1)x^{\theta}$, $0 \le x \le 1$

Find $\hat{\theta}$ and use the δ-method to approx $SE(\hat{\theta})$.

Find $\hat{\theta}$:

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

$$\mu = \frac{\theta+1}{\theta+2} \Rightarrow \mu\theta + 2\mu = \theta + 1$$

$$\Rightarrow \mu\theta - \theta = 1 - 2\mu$$

$$\Rightarrow \theta(\mu - 1) = 1 - 2\mu$$

$$\theta = \frac{1-2\mu}{\mu-1}$$

$$\Rightarrow \boxed{\hat{\theta} = \frac{1-2\bar{X}}{\bar{X}-1}}$$

Note

$$E(X^2) = \frac{\theta+1}{\theta+3}$$

Use δ-method to approx $SE(\hat{\theta})$!

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

$$g(x) = \frac{1-2x}{x-1}$$

$$g'(x) = \frac{1}{(x-1)^2}$$

$$\Rightarrow g'(\mu) = \frac{1}{(\mu-1)^2}$$

$$\sigma^2 = E(X^2) - E(X)^2 = \left(\frac{\theta+1}{\theta+3}\right) - \left(\frac{\theta+1}{\theta+2}\right)^2$$

$$\Rightarrow \text{Var}(\hat{\theta}) \approx (g'(\mu))^2 \frac{\sigma^2}{n}$$

$$= \frac{1}{(\mu-1)^4} \frac{\sigma^2}{n}$$

$$= \frac{1}{\left[\left(\frac{\theta+1}{\theta+2}\right) - 1\right]^4} \frac{\left[\left(\frac{\theta+1}{\theta+3}\right) - \left(\frac{\theta+1}{\theta+2}\right)^2\right]}{n}$$

Now plug in $\hat{\theta} = \frac{1-2\bar{X}}{\bar{X}-1}$ for θ

This you can calculate from the data.

Today sec 8.5

① - examples of MLE — by hand / in R

② - property of MLE ~ equivariance

① Sec 8.5 Maximum Likelihood estimator (MLE)

Suppose RV $X_1, \ldots, X_n$ have joint density $f(x_1, \ldots, x_n | \theta)$.

Given observed data values

$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ fixed numbers

we form a likelihood function

$lik(\theta) = f(x_1, \ldots, x_n | \theta)$ ← function of θ for fixed $x_1, \ldots, x_n$

The max value of $lik(\theta)$ represents the value of θ that maximizes the likelihood of observing your data.

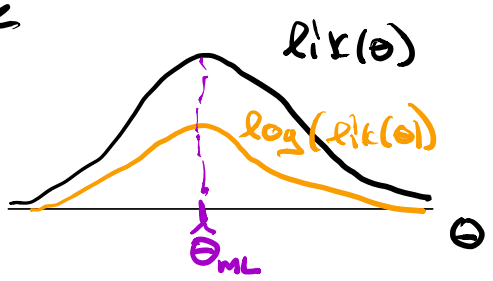
If x_i are iid

$$lik(\theta) = \prod_{i=1}^{n} f(x_i | \theta)$$

Taking the derivative of $lik(\theta)$ will involve the product rule. So we take the log of the likelihood so that we are finding the derivative of a sum.

Note that if $\hat{\theta}_{ML}$ is the value of θ that maximizes $lik(\theta)$, then $\hat{\theta}_{ML}$ also is the value of θ that maximizes $\log(lik(\theta))$ since log is a monotonically increasing function

Picture



$$l(\theta) = \log(lik(\theta)) = \sum_{i=1}^{n} \log(f(x_i|\theta))$$

To maximize $l(\theta)$ we take the derivative, set equal to zero, and solve for θ

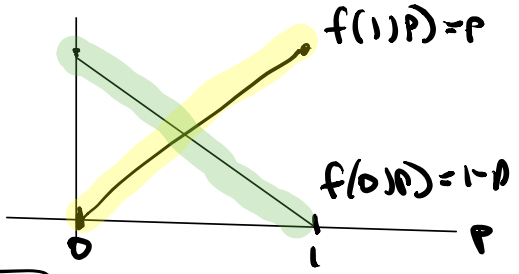
$\hat{\theta}_{ML}$ is the θ that solves $l'(\theta)=0$

ex $X_1 \sim Ber(p)$ (have single observation 0,1)

$$lik(p) = f(x|p) = p^x(1-p)^{1-x}$$

$f(1|p)=p$

$f(0|p)=1-p$



$$\boxed{\hat{p}_{ML} = \bar{x}}$$

ex $x_1, \ldots, x_n$ iid $exp(\lambda)$

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

$$lik(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

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

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\lambda}_{ML} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \boxed{\frac{1}{\bar{x}}}$$

$\hat{\lambda}_{ML}$

Ex # 16

Consider an i.i.d sample RV w/ density

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

Find $\hat{\sigma}_{ML}$

$$lik(\sigma) = \prod_{i=1}^{n} f(x_i|\sigma) = \left(\frac{1}{2\sigma}\right)^n \exp\left(-\frac{1}{\sigma}(|x_1| + \cdots + |x_n|)\right)$$

$$l(\sigma) = n\log\left(\frac{1}{2\sigma}\right) - \frac{1}{\sigma}\sum_{i=1}^{n}|x_i|$$

$$= -n\log 2 - n\log\sigma - \frac{1}{\sigma}\sum_{i=1}^{n}|x_i|$$

$$l'(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^2}\sum_{i=1}^{n}|x_i| = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}|x_i| = \frac{n}{\sigma}$$

$$\Rightarrow \boxed{\hat{\sigma}_{ML} = \frac{\sum_{i=1}^{n}|x_i|}{n}}$$

found $\hat{\sigma}_{MM} = \sqrt{\frac{\sum x_i^2}{n}}$

EXAMPLE C *Gamma Distribution*

Since the density function of a gamma distribution is

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}, \qquad 0 \le x < \infty$$

the log likelihood of an i.i.d. sample, $X_1, \ldots, X_n$, is

$$l(\alpha, \lambda) = \sum_{i=1}^{n}[\alpha \log \lambda + (\alpha - 1)\log X_i - \lambda X_i - \log \Gamma(\alpha)]$$

$$= n\alpha \log \lambda + (\alpha - 1)\sum_{i=1}^{n} \log X_i - \lambda \sum_{i=1}^{n} X_i - n \log \Gamma(\alpha)$$

formula we want to optimize

The partial derivatives are

$$\frac{\partial l}{\partial \alpha} = n \log \lambda + \sum_{i=1}^{n} \log X_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$\frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i$$

Setting the second partial equal to zero, we find

$$\hat{\lambda} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} X_i} = \frac{\hat{\alpha}}{\overline{X}}$$

But when this solution is substituted into the equation for the first partial, we obtain a nonlinear equation for the mle of α:

$$n \log \hat{\alpha} - n \log \overline{X} + \sum_{i=1}^{n} \log X_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0$$

see in-class exercise on b-course>.

This equation cannot be solved in closed form; an iterative method for finding the roots has to be employed. To start the iterative procedure, we could use the initial value obtained by the method of moments.

Computationally finding MLE for Gamma paramaters

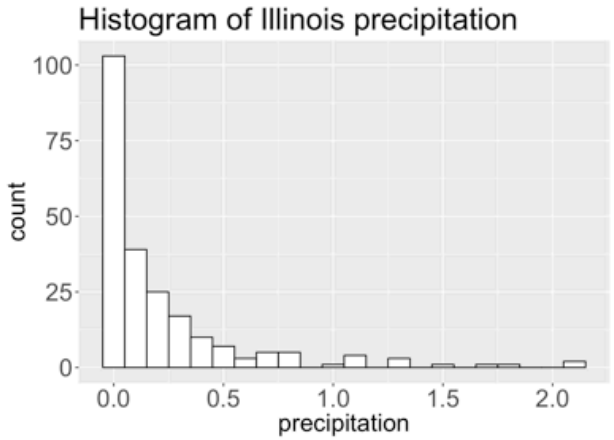
We take the data from 227 storms in Illinois between 1960,and 1964

```
#here is the precip data

precip <- c(0.020, 0.001, 0.001, 0.120, 0.080, 0.420, 1.720, 0.050, 0.010, 0.010,0.003,0.001,0.
003, 0.270, 0.001, 0.060, 0.050, 2.130, 0.040, 1.100, 0.020, 0.001,0.140, 0.080, 0.210, 0.070,
0.320, 0.240, 0.290, 0.001, 0.290, 1.130, 0.003,
0.010, 0.190, 0.002, 0.010, 0.040, 0.002,0.070, 0.450, 0.010, 0.180, 0.670, 0.003, 0.010, 0.040
, 0.002, 0.490, 0.020, 0.020, 0.340, 0.140, 0.370, 0.330,
0.330, 0.350, 0.010, 0.500, 0.760, 1.060, 0.002 ,0.060, 0.160 ,0.270, 0.250,
0.290, 0.020, 0.050, 0.460, 0.070, 0.410, 0.020, 0.080, 0.210, 0.010, 0.440,
0.020, 0.050 ,0.110, 1.500, 0.003, 0.180 ,0.010, 0.002, 0.240, 0.010, 0.750,
0.010, 0.140 ,0.130, 0.010, 0.010 ,0.270, 0.450, 1.780, 0.250, 0.240, 0.004,
0.210, 0.170, 0.830, 0.150, 0.030, 0.030, 0.500 ,0.040, 0.090 ,0.040 ,0.060,
0.060, 0.120, 0.003, 0.003, 0.400, 0.020, 0.510, 0.003, 0.020, 0.020, 0.020,
0.010, 0.001, 0.140, 0.001, 0.100, 0.010 ,1.090 ,0.010, 0.002, 0.001, 0.840,
0.030, 0.350, 0.070, 0.001, 0.002, 0.002, 0.200, 0.060, 0.140, 0.010, 0.020,
0.020, 0.002, 0.001, 0.550, 0.130, 0.190, 2.100, 0.090, 0.350, 0.790, 0.320,
1.350,0.170,0.020,0.002, 0.010,0.250,0.230,0.170,0.010,0.020,0.001,
0.010,0.020,0.110,0.210,1.260,0.010,0.730,0.100,0.090,0.007, 0.360,
0.770,0.210,1.270,0.070,0.080,0.160,0.260,0.010,0.230,0.080,0.020,
0.010,0.290,0.010,0.010,0.070,0.400,0.002, 0.003, 0.010,0.090,0.160,
0.040,0.270,0.730,0.410,0.030,0.120,0.030,1.040,0.060,0.090,0.730,
0.040,0.160,0.590,0.003, 0.002, 0.020,0.004, 0.010,0.001, 0.060,0.620,
0.010,0.520,0.110,0.003, 0.600,0.002, 0.050)

rain <- data.frame(precip)

#make histogram
rain %>% ggplot(aes(x=precip)) +
  geom_histogram(binwidth = .1,col="black",fill="white") +
  labs(title="Histogram of Illinois precipitation",x="precipitation",y="count") + theme(
    axis.text = element_text(size = 20),
    plot.title = element_text(size = 25),
    axis.title=element_text(size=20))
```



First we find the MLE estimate of α and λ. This will be the initial value of our optimization.

```
alpha_hat_MOM <- mean(precip)^2/(mean(precip^2)-mean(precip)^2)
alpha_hat_MOM
```

```
## [1] 0.3779155
```

```
lambda_hat_MOM <- mean(precip)/(mean(precip^2)-mean(precip)^2)
lambda_hat_MOM
```

```
## [1] 1.684175
```

```
fun <-  function(para, x) {
    alpha <-  para[1]
    lambda <-  para[2]
    -sum(alpha * log(lambda) + (alpha - 1) * log(x) - lambda * x - log(gamma(alpha)))


mle <-  optim(par = c(alpha_hat_MOM, lambda_hat_MOM),  fn = fun, x=precip)
#par=Initial values for the parameters to be optimized over.
#fn=    A function to be minimized (or maximized), with first argument the vector of parameters
over which minimization is to take place. It should return a scalar result.
#x= further arguements to pass to fn.
#method=the method to be used. The default is a method that only uses values of the function. I
t is very robust to initial values but slow. My favorite is method="BFGS" which is Newton's met
hod which use a combination of the function value and its derivative. It is very fast. Type ?op
tim in the console to learn more.

#Note: the function optim returns a list including "par" the best set of parameters found.


alpha.mle <-  mle$par[1]
lambda.mle <-  mle$par[2]
alpha.mle
```

method = "BFGS" is Newton's method.

```
## [1] 0.4408386
```

```
lambda.mle
```

```
## [1] 1.964841
```

Note that the MLE and MOM estimates give different values.

Stats 135, Fall 2019
Lecture 8, Monday, 9/16/2019

1 Review

Last time, we went over §8.5, the method of Maximum Likelihood, which is an optimization problem and typically involves calculus. There are some technical assumptions we need to make about our random variables in order for an MLE to exist. These are called **regularity assumptions**. Our text treats these very lightly, and Lucas emphasizes the same points as the book:

R0. The pdfs are distinct (i.e. $\theta \neq \theta' \implies f(x_i \mid \theta) \neq f(x_i \mid \theta')$) That is, if we have two different parameters, then their pdfs are distinct.

R1. The pdfs have common support for all θ. Support is where the function equals zero. (i.e. $\{x \mid f(x \mid \theta) \neq 0\} = \{x \mid f(x \mid \theta') \neq 0\}$.

R2. The true parameter θ_0 is an interior point in the parameter space.

Theorem 1.1. Let θ_0 be the true parameter. Under assumptions (R0) and (R1),

$$\lim_{n\to\infty} \ldots$$

In other words, for large enough sample size, we can find the MLE.

For example, we look at the normal distribution with the following log-likelihood:

$$l(\mu, \sigma) = -n\log(\sigma) - \frac{n}{2}\ldots$$

Now unlike in Gamma, we can actually solve for our MLE estimators. We simply have:

$$\hat{\mu} = \overline{X}$$

and

$$\hat{\theta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

where we do NOT have an $n-1$ in the denominator. Often we want an estimator for σ^2, and this is very easy via the **equivariance** property of MLE estimators. That is,

$$\overline{\sigma}^2 = \overline{\sigma}^2_{ML} = \frac{1}{n}\sum(X_i - \overline{X})^2.$$

Topics Today:

- §8.5 - Equivariance
- §8.5 - Consistency
- §6.2 - t distribution and $\chi^2(n)$ distribution

In the interest of time Lucas notes that we'll go through consistency rather quickly, so that we may do t-distribution and $\chi^2(n)$ distributions more carefully (as they may be in exams or homework).

2 Equivariance

A nice property of the MLE $\hat{\theta}_{ML}$ is **equivariance** in that

$$\hat{g(\theta)} = ...$$

Theorem 2.1. If $\hat{\theta}_{ML}$ is a MLE of θ and g is a function (need not be monotonic) then $g(\hat{\theta}_{ML})$ is an MLE of $g(\theta)$.

3 Consistency

Recall that this means that our estimator converges in probability to the true parameter. The picture that we should have in our heads is that we have a list of parameters

$$\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n,$$

where for larger sample sizes, our estimator will be more and more pointed and converge to the true value θ_0. Lucas puts it as that we become more and more certain of what that value is.

Theorem 3.1. (Consistency.) Assume that $X_1, \ldots, X_n$ satisfy ergularity conditions (R0), (R1), (R2), where θ_0 is the true parameter, and further assume that $f(x \mid \theta)$ is differentiable with respect to θ. Then $l'(\theta) = 0$ has at least a solution $\hat{\theta}_n$ with the property that $\hat{\theta}_n \xrightarrow{p} \theta$.
If $l'(\theta) = 0$ has a unique solution, then it is a consistent estimator θ, so we need not check that our parameter has a maximum.

If n is infinitely large, there still may be several solutions. We would check each solution and see which gives the maximum likelihood. Lucas notes that for our purposes, we need only understand how this proof works (we will not be held responsible for the proof's details), but for time we will put it in the appendix and skip onwards.

4 §6.2: t distribution and χ^2 distribution

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Motivation: In chapter 7, we showed that for large n,

$$\overline{x} \pm 1.96 z(.025)$$

is a 95% confidence interval for μ. We needed the averages to be approximately normal. However, to talk about a confidence interval for small n, we need a t distribution to make this precise. Furthermore, to find a 95% CI for σ^2, we will need a χ^2 distribution.

The goal for today is to provide background on these two distributions.

4.1 χ^2 distribution

We define this by starting off with a standard normal distribution:

$$Z \sim N(0,1).$$

Now define:

$$Z^2 \sim \chi^2_1,$$

where the subscript 1 is the 'degree(s) of freedom', here showing that there is only one variable.
Now we should be able to verify: $Z^2 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$. If we know the density for Z, we can find the density of $g(Z)$ via the change of variable probability theorem. Lucas will add the details to this later.

Definition: χ^2_n **-**

Let $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$.
Then:

$$Z_1^2 + \cdots + Z_n^2 \sim \chi^2_n,$$

which is the chi-squared with n degrees of freedom.

Fact:

$$Z_1^2 + \cdots + Z_n^2 \sim \text{Gamma}(\frac{\text{n}}{2}, \frac{1}{2}).$$

To see this, recall that the sum of n iid Gamma(r, α) is also Gamma(nr, α).

Lucas adds:

$$\mathbb{E}(\chi^2_k) = k$$

As $k \to \infty$, this χ^2_k distribution looks more and more normal.

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

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

so

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \sim \chi^2_n.$$

We see (this is very important) that:

$$\frac{n-1}{\sigma^2} s^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Xi - \overline{X})^2 \sim \chi^2_{n-1}.$$

Somehow, we lose a degree of freedom. Why is this the case? Basically, what this says is that the standard **variance is a multiple** of χ^2_{n-1}. Lucas notes that this is a simple way to think about the sample variance.

Intuitively, we would expect to lose a degree of freedom, whereas

$$\{X_1 - \mu, X_2 - \mu, \ldots, X_n - \mu\}$$

are independent. However,

$$\{X_1 - \overline{X}, x_2 - \overline{X}, \ldots, X_n - \overline{X}\}$$

ARE dependent, because if we add them all up and distribute the sum, we have:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - n \cdot \overline{X} = 0.$$

There is a dependency between these guys, so it somewhat makes sence that we'll lose one degree of freedom. It's a good thing to make this rigorous. We provide a sketch of a more rigorous proof.

Theorem 4.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

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

Proof. The proof of this is given in Rice Chapter 6 on page 197. Basically,

$$\begin{aligned}
\frac{(n-1)s^2}{\sigma^2} &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} - \frac{\overline{X} - \mu}{\sigma} \right)^2 \\
&= \sum_{i=1}^{n} \left(Z_i - \overline{Z} \right)^2, \text{ where } Z_i := \frac{x_i - \mu}{\sigma^2}, \text{ and } \overline{Z} := \frac{1}{n} \sum Z_i = \frac{\overline{X} - \mu}{\sigma} \\
&= \sum_{i=1}^{n} Z_i^2 - 2\overline{Z} \underbrace{\sum_{i=1}^{n} Z_i}_{=n\overline{Z}} + \underbrace{\sum_{i=1}^{n} \overline{Z}^2}_{=n\overline{Z}^2} \\
&= \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2.
\end{aligned}$$

Then rearranging gives:

$$\sum_{i=1}^{n} Z_i^2 = \frac{(n-1)s^2}{\sigma^2} + n\overline{Z}^2,$$

where the LHS is a chi-squared distribution with n degrees of freedom. In other words, $\chi^2_n = \text{Gamma}(\frac{n}{2}, \frac{1}{2})$. Now for the second term on the RHS, $n\overline{Z}^2$, recall that $\overline{Z} = \frac{\overline{X} - \mu}{\sigma}$. Then $\sqrt{n}\overline{Z}$ takes on the standard normal. Hence $n\overline{Z}^2$ takes on the $\chi^2_1 = \text{Gamma}(\frac{1}{2}, \frac{1}{2})$ distribution.
Additionally, by Theorem A, p. 195 of Rice, we have that:

$$\overline{X} \text{ and } s^2 \text{ are independent random variables for } X_i \sim N(\mu, \sigma^2).$$

It follows that $\frac{(n-1)s^2}{\sigma^2}$ and $n\overline{Z}^2$ are independent. Now because the sum of independent gamma is gamma, and $\sum_{i=1}^{n} Z_i^2$ is Gamma($\frac{n}{2}, \frac{1}{2}$), we conclude:

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

must be Gamma($\frac{n-1}{2}, \frac{1}{2}$) $= \chi^2_{n-1}$.

□

Lecture ends here.

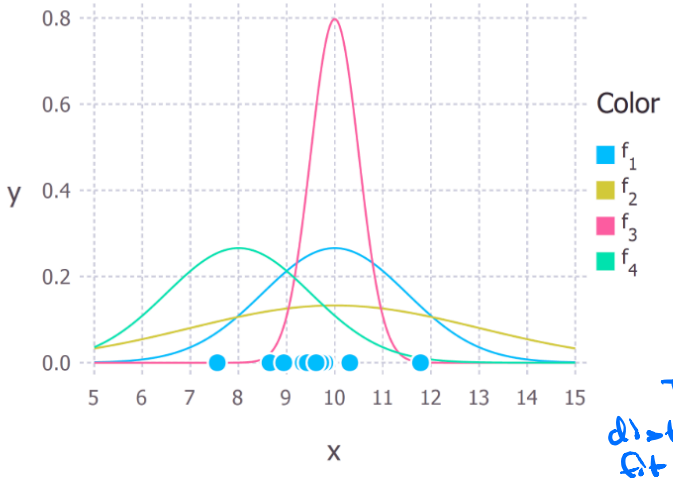
Next time, we'll look at the t-distribution.

Stats 135, Fall 2019

Lecture 9, Wednesday, 9/19/2019

CLASS ANNOUNCEMENTS: RStudio demonstration in-class today.

stat 135 lec 9



Which normal distribution corresponds to our data?

Answer

$f_1 = N(10, 2.25)$

$\hat{\mu}_{ML} = 10$

$\hat{\sigma}^2_{ML} = 2.25$

The ML estimator is the distribution parameters that best fit your data.

Last time For $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$\hat{\mu} = \bar{X}$ ← ML estimator

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

$\frac{n-1}{\sigma^2} S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$

$\Rightarrow \frac{n\hat{\sigma}^2}{\sigma^2} = \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

Today

0. Sec 6.3 t-distribution
1. Sec 8.5.3 Find 95% CI for μ, σ^2 for $N(\mu, \sigma^2)$
2. Sec 8.5.3 Parametric bootstrap in R (to find 95% CI)
3. Sec 8.5.2 large sample properties of $\hat{\theta}_{ML}$ (to find 95% CI)

t-distribution (see Chap 6)

def^n $Z \sim N(0,1)$

$U \sim \chi^2_{n-1}$

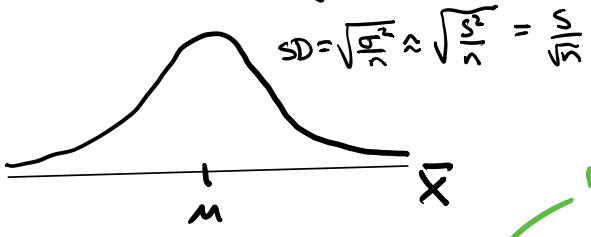
$$\frac{Z}{\sqrt{U/n-1}} \sim t_{n-1}$$

($n-1$: d.f.)

Let $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where μ, σ^2 are unknown.

the t-distribution comes up when making a CI for μ.

$\bar{X}$ has sampling distribution



NOT $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

but $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$

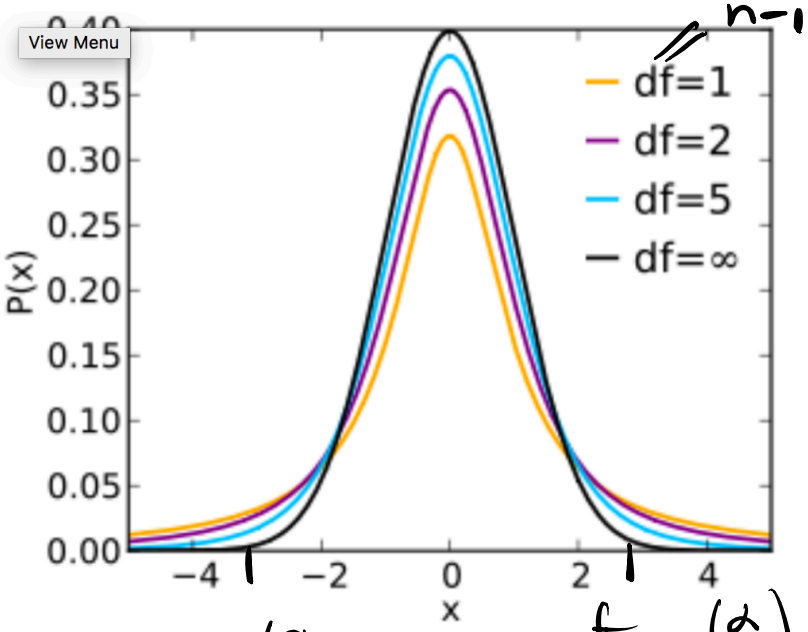
"normalize $\bar{X}$"

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/n-1}} \sim t_{n-1}$$

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = Z$

$\sqrt{\frac{(n-1)s^2}{\sigma^2}/n-1} \leftarrow \sqrt{U/n-1}$

$U \sim \chi^2_{n-1}$



$t_{n-1}\left(\frac{\alpha}{2}\right)$ ← $\frac{\alpha}{2}$ quantile of t_{n-1} $\qquad$ $t_{n-1}\left(\frac{\alpha}{2}\right)$

Sec 8.8.3 **95% CI for μ and σ^2 of $N(\mu, \sigma^2)$** where $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

95% CI of μ:

We have,

$$P\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} \in \left[-t_{n-1}\left(\frac{\alpha}{2}\right), t_{n-1}\left(\frac{\alpha}{2}\right)\right]\right) = 1-\alpha$$

$$\Rightarrow P\left(\mu \in \left[\bar{X} - t_{n-1}\left(\frac{\alpha}{2}\right)\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1}\left(\frac{\alpha}{2}\right)\frac{S}{\sqrt{n}}\right]\right) = 1-\alpha$$

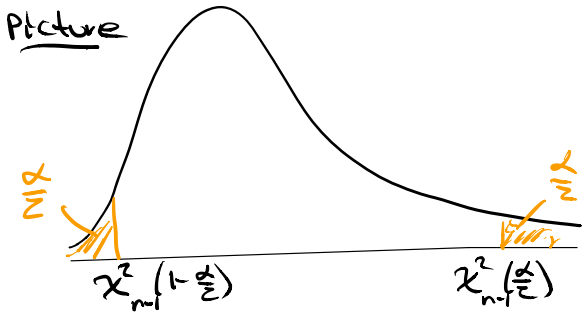
↑ $(1-\alpha)100\%$ CI for μ.

Here $\frac{S}{\sqrt{n}}$ is an estimate of the SE of $\hat{\mu}_{ML}$.

Next 95% CI of σ^2:

Recal $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

Picture



$\frac{n}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-1}$

we have,

$$P\left(\frac{n}{\sigma^2}\hat{\sigma}^2 \in \left[\chi^2_{n-1}(1-\tfrac{\alpha}{2}), \chi^2_{n-1}(\tfrac{\alpha}{2})\right]\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{1}{\sigma^2} \in \left[\frac{1}{n\hat{\sigma}^2}\chi^2_{n-1}(1-\alpha), \frac{1}{n\hat{\sigma}^2}\chi^2_{n-1}(\tfrac{\alpha}{2})\right]\right) = 1-\alpha$$

Note that $3 \in [3,4] \Rightarrow \frac{1}{3} \in [\frac{1}{4}, \frac{1}{2}]$

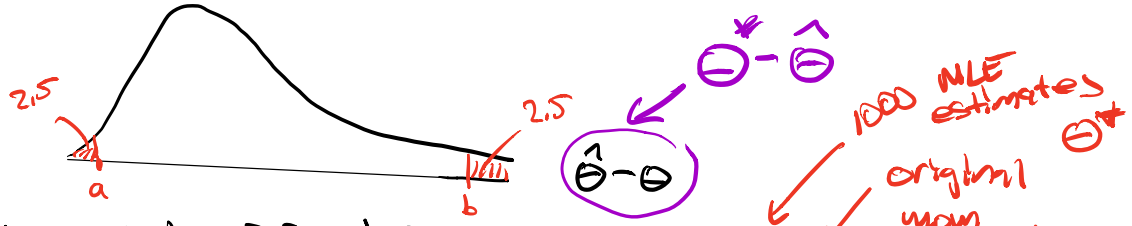
So

$$P\left(\sigma^2 \in \left[n\hat{\sigma}^2 \cdot \frac{1}{\chi^2_{n-1}(\frac{\alpha}{2})}, n\hat{\sigma}^2 \frac{1}{\chi^2(1-\frac{\alpha}{2})}\right]\right) = 1-\alpha$$

↖ $(1-\alpha)100\%$ CI for σ^2

(2) Sec 8.5.3 Parametric bootstrap in R (to find 95% CI)

To make a 95% CI recall from lecture 5:



let a, b be 2.5 and 97.5 percentile of $\hat{\theta} - \theta$

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

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

↖ This is a 95% CI of θ.

In R,

$b = \text{quantile}(\theta^* - \hat{\theta}, 1 - \frac{\alpha}{2})$

$a = \text{quantile}(\theta^* - \hat{\theta}, \frac{\alpha}{2})$

$[\hat{\theta} - b, \hat{\theta} - a] = \hat{\theta} - \text{quantile}(\theta^* - \hat{\theta}, c(1-\frac{\alpha}{2}, \frac{\alpha}{2}))$

$\hat{\theta}$ is a constant. → $\text{quantile}(\theta^*, c(1-\frac{\alpha}{2}, \frac{\alpha}{2})) - \hat{\theta}$

$= 2\hat{\theta} - \text{quantile}(\theta^*, c(1-\frac{\alpha}{2}, \frac{\alpha}{2}))$

1a) The parametric bootstrap of estimator SE

In a previous lab you saw the nonparametric bootstrap where you resample from the sample to approximate the se of some location parameter such as the sample mean. The parametric bootstrap may be used to find the se of a MOM or MLE estimate of a distribution parameter. Below we compare and contrast these approaches for finding the se of the MOM approximation to the Poisson rate parameter λ.

In the parametric bootstrap we find a sample then we look at the distribution of the sample and guess a model. We use MOM (or MLE) to find estimator of parameters. We repeatedly simulate model with estimated parameters and find se of sampling distribution.

Step 1. Get a sample make a histogram and guess a model that fits the histogram.

Step 2. Estimate parameter using MOM estimate. Overlay the histogram from step 1 with a frequency plot for the model distribution with the estimated parameter.

Step 3. Use model with estimated parameter to generate sampling distribution and find its SD.

Discuss with a neighbor the following code.

Note we generally don't know the SD of the ML estimator. In this simple example $SD(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{\lambda}{\sqrt{n}} = \frac{1}{\sqrt{n}}$ For this reason we need to bootstrap the SE even if we assume our distribution is Pois(1).

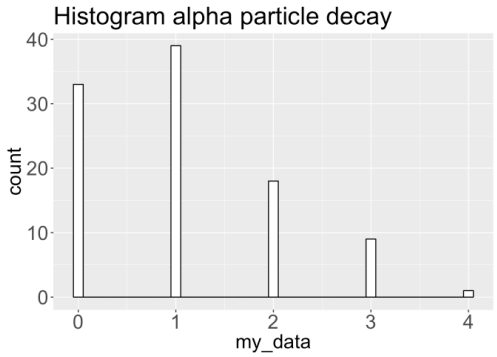
Step 1: visualize data

```
#pop is Pois(1)
sample_size <- 100
B <- 1000 #number samples

my_data <- rpois(sample_size,lambda=1) #my sample
df.my_data<- data.frame(my_data) #make data frame
head(df.my_data)
```

```
##   my_data
## 1       1
## 2       2
## 3       1
## 4       0
## 5       1
## 6       1
```

```
#make histogram
df.my_data %>% ggplot(aes(x=my_data)) +
  geom_histogram(binwidth = .1,col="black",fill="white") +
  labs(title="Histogram alpha particle decay",x="my_data",y="count") +
  theme(
    axis.text = element_text(size = 20),
    plot.title = element_text(size = 25),
    axis.title=element_text(size=20))
```



Based shape of distribution we decide to model as Pois(λ).

Step 2: model data as probability distribution with MLE (or MOM) parameter estimate

We saw in class MLE predicts: $\hat{\lambda} = \bar{X}$.

```
#We find estimated parameters
lambda_hat <- mean(my_data)
lambda_hat
```

```
## [1] 1.06
```

We check for fit of our model by generating random numbers with parameter lambda_hat.

```
theory <- rpois(sample_size, lambda=lambda_hat)
data.df <- data.frame(my_data,theory)
head(data.df)
```

```
##   my_data theory
## 1       1      0
## 2       2      2
## 3       1      1
## 4       0      0
## 5       1      0
## 6       1      2
```

```
#make histogram
data.df %>% ggplot() +
  geom_histogram(aes(x=data.df$my_data), binwidth = .1,col="black",fill="white") +
  geom_freqpoly(aes(x=data.df$theory),col="red",show.legend = TRUE) +
  labs(title="Theory versus data",x="my_data",y="count") +
  theme(
    axis.text = element_text(size = 20),
    plot.title = element_text(size = 20),
    axis.title=element_text(size=15))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Step 3: Find SE

```
find_lambda <- function(sample_size){
data <- rpois(sample_size, lambda=lambda_hat)
mean(data)
}

lambda_vector <-  replicate(B, find_lambda(sample_size))
sd(lambda_vector)
```

```
## [1] 0.1056974
```

```
#expect 1/sqrt(sample_size)
```

1b) The nonparametric bootstrap of estimator SE

In the nonparametric bootstrap we find a sample then we resample that sample many times.

Discuss with a neighbor the following code.

```
sample_size <- 100
B=1000
my_data <- rpois(sample_size,lambda=1)  #my sample
head(my_data)
```

```
## [1] 2 0 0 0 0 0
```

```
#we will run this function many times
find_lambda_np <- function(){
 resample <- my_data %>% sample(replace=TRUE)
  mean(resample)
}

lambda_vec <-  replicate(B, find_lambda_np())
se.lambda=sd(lambda_vec)
se.lambda
```

```
## [1] 0.1040743
```

```
#expect 1/sqrt(sample_size)
```

To experiment cut and paste the above code in the console and run.

```
library(ggplot2)
library(dplyr)
```

2) 95% CI of parameters of λ of Pois(λ) distribution using parametric bootstrap

Next we find the $(1-\alpha)100$ % CI for alpha=.05 (i.e. 95% CI)

Discuss with a neighbor the following code where we calculate the CI of λ.

```
alpha=.05
CI.lambda= 2*lambda_hat -quantile(lambda_vec,c(1-alpha/2,alpha/2))
as.vector(CI.lambda)
```

```
## [1] 0.99 1.41
```

Sec 8.5.2 Large sample theory for MLE

Thm - p277 Rice

The MLE $\hat{\Theta}_{ML}$ for θ is asymptotically (i.e. for large sample size n) unbiased and normal.

More precisely, for large n,

$$\hat{\Theta}_{ML} \approx N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$ where

$I(\theta_0)$ is the Fisher info at the value θ_0 of θ.

Fisher info (FI)

What does FI measure?

Let's say a RV X and param θ are related with some known density function $f(x|\theta)$

The FI answers the question:

How useful is the RV X in determining θ.

$\Rightarrow$ lets look at $\log f(x|\theta)$ for two circumstances where x is really informative and really uninformative about θ.

Informative $X \sim N(\mu, 1)$

$$\ell(\mu) = \log f(x|\mu)$$

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

$$= \log\left(\frac{1}{\sqrt{2\pi}}\right) - (x-\mu)^2$$

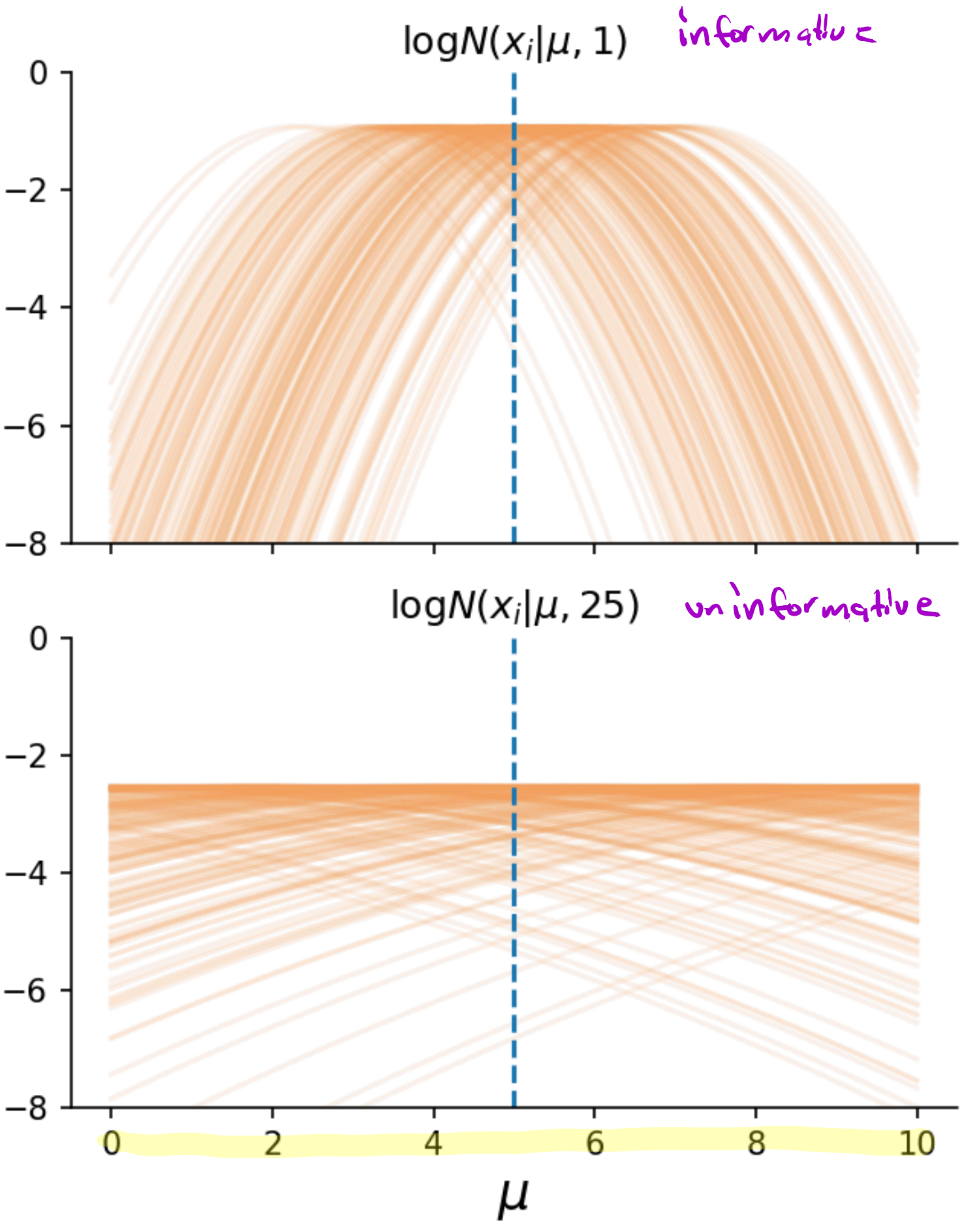
uninformative

$$X \sim N(\mu, 25)$$

$$\ell(\mu) = \log\left(\frac{1}{\sqrt{2\pi \cdot 25}}\right) - \left(\frac{x-\mu}{5}\right)^2$$

We plot $\ell(\mu)$ for 1000 individual x for both $N(\mu, 1)$ and $N(\mu, 25)$.

Suppose $\mu = 5$.


$\log N(x_i|\mu, 1)$
informative
0
−2
−4
−6
−8
$\log N(x_i|\mu, 25)$
uninformative
0
−2
−4
−6
−8
0
2
4
6
8
10
μ

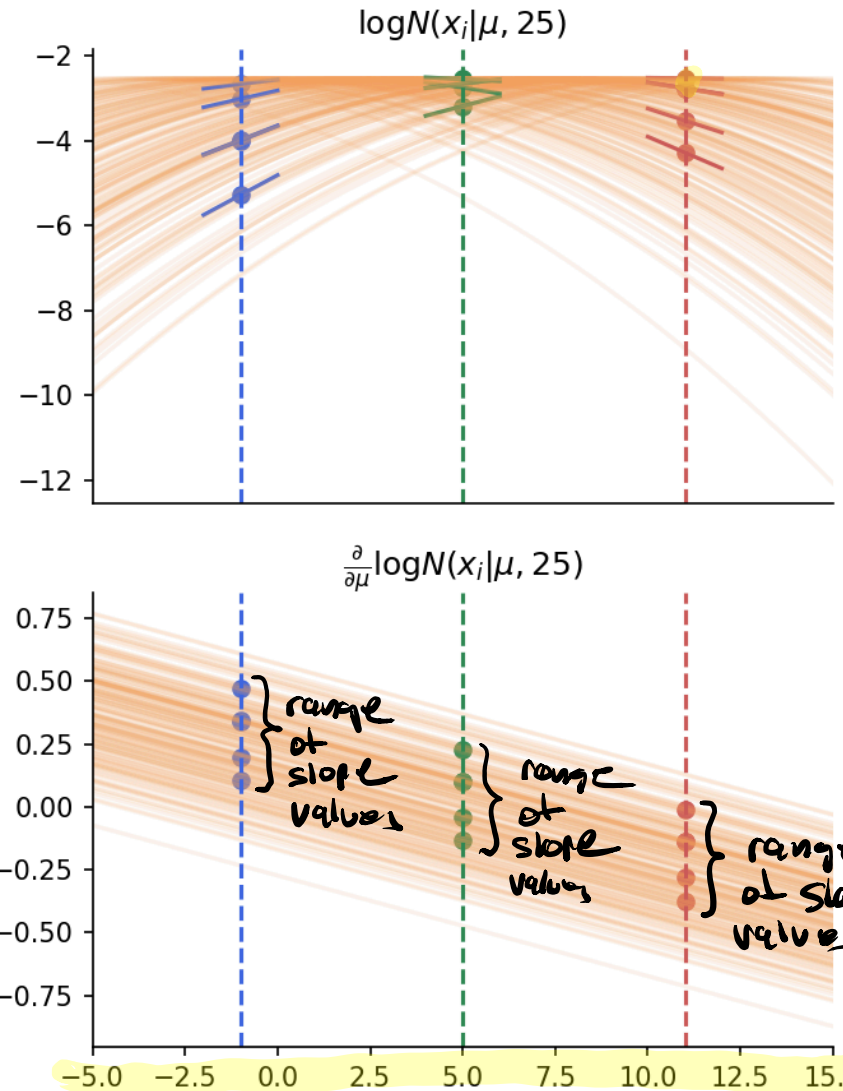
Each curve is providing their own vote of the true parameter location (peak).

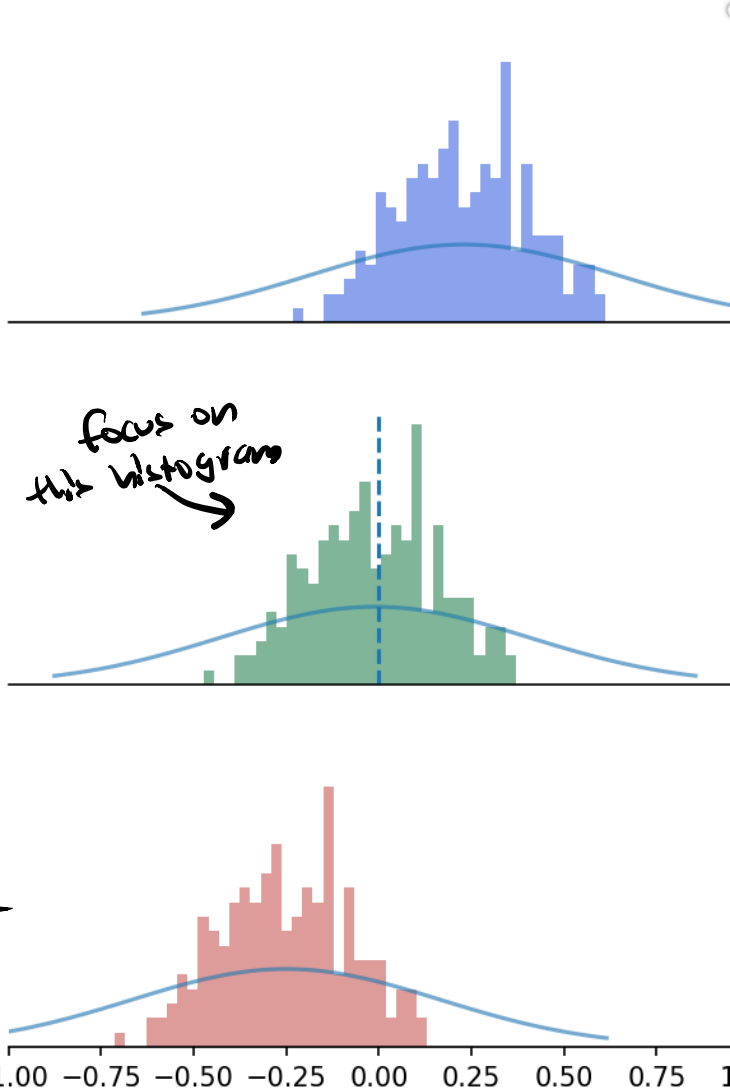
How do we measure that a curve has a tighter peak?

It is done by looking at the slopes of the curves (called the <u>score function</u>)

$$\text{Score}_X(\theta) = \frac{\partial}{\partial \theta} \log\left(f(x|\theta)\right)$$

We pick a few values of θ and look to evaluate the score function and make a histogram


$\log N(x_i|\mu, 25)$
−2
−4
−6
−8
−10
−12
$\frac{\partial}{\partial\mu}\log N(x_i|\mu, 25)$
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
range of slope values
range of slope values
range of slope values
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0


focus on this histogram
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75

Notice

(1) We will concentrate on the middle (green) histogram of scores. The above pictures is only for $N(5,25)$. If we made the pictures for $N(5,1)$ the middle histogram would have been wider since there would have been very steep uphill slopes and steep downhill slopes.

(2) The middle (green histogram) is centered at zero. This is true for both $N(5,1)$ and $N(5,25)$.

$$\text{i.e } E\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\Big|_{\theta=\theta_0}\right) = 0$$

so the spread of the score at $\theta=\theta_0$ is a measure of how informative x is in estimating θ_0.

Large spread of the histogram of the score at $\theta=\theta_0$ is more informative, x has a large FI if the histogram of of the score at $\theta=\theta_0$ has big variance.

Def^n FI is the variance of the middle histogram above.

$$I(\theta_0) = \text{var}\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\Big|_{\theta=\theta_0}\right)$$

Fisher info

Stats 135, Fall 2019
Lecture 10, Friday, 9/20/2019

1 Review

Last time, we covered §8.5.2. We found the key result that the MLE has very nice asymptotic properties. That is, for $X_1, \ldots, X_n$ iid with density $f(x \mid \theta)$, we have:

$$\hat{\theta}_{ML} \sim N\left(\theta, \frac{1}{nI(\theta)}\right),$$

for large n, where $I(\theta)$ denotes Fisher information. In other words, the MLE is approximately normal and is unbiased.
Lucas shows some log-likelihood distribution plots, where histograms for more informative $X \sim N(\mu, 1)$ would be more spread out. When the variance is small, it is very easy to see the location of the max (μ). Of course, all these trajectories are different because they are samples from the population.
We look at the slope (derivative) of the log-likelihood, which we call the "score". We look at it at the true value.

Definition: Fisher information -

Fisher information, $I(\theta)$ is the amount of information that a single observation X has to estimate the max of the log likelihood, $l(\theta)$. That is,

$$I(\theta) := \text{Var}(l'(\theta))$$

Topics Today:

- Alternative definition and examples of F.I. (Fisher information)
- Computing the CI using large sample theory and showing that $(1-\alpha)100\%$ CI is

$$\hat{\theta}_{ML} \pm z\left(\frac{\alpha}{2}\right)\sqrt{\frac{1}{nI(\theta)}}$$

- §8.7, Cramer Rao (CR) inequality

2 Definition of Fisher Information

We define the Fisher Information to be the variance of the score at the true parameter value.

$$I(\theta_0) = \text{Var}\left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\big|_{\theta=\theta_0}\right)$$

Now because the expectation of the score function at θ_0 is zero, we can write the Fisher information as:

$$I(\theta_0) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\left(f(x \mid \theta)\right)\big|_{\theta=\theta_0}\right)^2\right]$$

because $\text{Var}(A) = \mathbb{E}(A^2) - (\mathbb{E}(A))^2$, and if $\log f(x \mid \theta)$ is twice-differentiable (see Rice, lemma A on page 276), then

$$I(\theta_0) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log\left(f(x \mid \theta)\right)\big|_{\theta=\theta_0}\right]^2$$

That is, the Fisher information is the average curvature of the log-likelihood over all values of $x \in X$ (all our sample trajectories). Lucas notes that basically, the Fisher information has two equivalent interpretations:

1. Variance of score function at the true parameter θ_0.
2. negative of the average curvature of the log-likelihood θ_0 over all $x \in X$.

Lucas says that for our purposes, we will refer ('go-to') the second definition.

3 Cramér–Rao Inequality (Cramér–Rao Lower Bound, CRLB)

Let $X_1, \ldots, X_n$ be iid random variables with density $f(x \mid \theta)$. Now two things are true.
(1) The MLE is asymptotically normal, unbiased, with variance:

$$\hat{\theta}_{ML} \approx N\left(\theta, \frac{1}{nI(\theta)}\right)$$

for large n.
(2) The Cramer Rao inequality holds. Let the ML estimator, $\hat{\theta}_{ML}$, be unbiased. Then

$$\text{Var}\left(\hat{\theta}_{ML}\right) \geq \frac{1}{nI(\theta)}.$$

We won't prove this in this class (Lucas jokes this is deferred for our next graduate class in Statistics).

Definition: Efficient -

An unbiased estimator where variance achieves the CR lower bound is called **efficient.**

Now from point (1) above, we see that the ML estimator $\hat{\theta}_{ML}$ is asymptotically efficient.

Example: Take $X_1, \ldots, X_n$ iid with Binomial(1, p), where $0 < p < 1$. We wish to find $I(p) := -\mathbb{E}[l''(p)]$.
Recall that

$$f(x \mid p) = p^x(1-p)^{1-x},$$

so taking logs of both sides yields:

$$l(p) = x\log p + (1-x)\log(1-p),$$

and now taking successive derivatives gives:

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

To continue, we need to take the expectation:

$$\begin{aligned}\mathbb{E}[l''(p)] &= \frac{-\mathbb{E}(x)}{p^2} - \frac{(1-\mathbb{E}(x))}{(1-p)^2}, \text{ where } \mathbb{E}(x) = p \\ &= \frac{-p}{p^2} - \frac{(1-p)}{(1-p)^2} \\ &= -\frac{1}{p} - \frac{1}{1-p} \\ &= \boxed{\frac{-1}{p(1-p)}}\end{aligned}$$

Hence:

$$\boxed{I(p) = -\mathbb{E}(l''(p)) = \frac{1}{p(1-p)}}$$

Now because we want to find the SE of our ML estimator of p, $\hat{p}_{ML}$, we need to find $\frac{1}{nI(p)}$. We assume that n is large and use the asymptotic property. We have:

$$nI(p) = \frac{n}{p(1-p)} \implies \frac{1}{nI(p)} = \boxed{\frac{p(1-p)}{n}},$$

and hence

$$\text{SE of } (\hat{p}_{ML}) = \sqrt{\frac{1}{nI(p)}} = \sqrt{\frac{p(1-p)}{n}}$$

Lucas notes that this used the asymptotic property. Now we try to get this sort of result directly without this assumption that n is large and compare. The Cramer-Rao inequality states that what we found above is a lower bound of what we will directly compute. Hence what we will find must be greater than or equal to our above expression.

3.1 Direct Computation:

To find $\hat{p}_{ML}$ directly, again consider $X_1, \ldots, X_n$ iid Binomial(1, p).
In other words,

$$\text{lik}(p) = f(X_1, \ldots, X_n \mid p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i},$$

and

$$l(p) = \log \text{lik}(p) = \cdots$$

Lucas skips the details and mentions that it's not surprising that we simply get:

$$\boxed{\hat{p}_{ML} = \overline{X}},$$

which is generally reasonable and Lucas skips proof (or provides supplementally in notes).
Now we want to find SE of $(\hat{p}_{ML})$. Take:

$$\text{Var}(\overline{X}) = \frac{p(1-p)}{n},$$

which implies

$$\text{SE of } (\hat{p}_{ML}) = \sqrt{\frac{p(1-p)}{n}},$$

which is precisely the Cramer-Rao lower bound. We showed the same exact expression as above with the assumption (MLE property).
Hence the estimator of $\hat{P}_{ML}$ is approximately:

$$\text{SE of } (\hat{p}_{ML}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}.$$

4 Confidence Intervals

Now the approximate $(1-\alpha)100\%$ confidence interval for p is

$$\begin{aligned}
&\hat{p}_{ML} \pm z\left(\frac{\alpha}{2}\right)\sqrt{\frac{1}{nI(p)}} \\
&\approx \overline{X} \pm z\left(\frac{\alpha}{2}\right)\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}
\end{aligned}$$

A question arises in the audience as to when we would use this estimator or the bootstrap method.
Bootstrapping is to use our sample as opposed to our population. Here, we consider that we are using the equivariance property to use $\overline{X}$ as an estimator for p.

5 Information in a Random Sample

Lucas skips through this very quickly but provides the algebraic derivations later.
The Fisher Information for the entire n sample is

$$I_n(\theta) = nI(\theta),$$

which can be shown very simply by taking the variance of a sum of iid variables.
Hence the FI in an iid random sample is simply n times the FI in a single observation. So for large n,

$$\hat{\theta} \sim N\left(\theta, \frac{1}{I_n(\theta)}\right),$$

where $I_n(\theta) = nI(\theta)$.

6 Example

Let $X_1, \ldots, X_n$ be a random iid sample with density $f(x \mid \theta) = \theta x^{\theta-1}$, for $0 < x < 1$, and $\theta > 0$.
(a) Find $\hat{\theta}_{ML}$.

Solution.

$$\begin{aligned} f(X_1, \ldots, X_n \mid \theta) &= \left(\theta X_1^{\theta-1}\right)\left(\theta X_n^{\theta-1}\right)\cdots\left(\theta X_n^{\theta-1}\right) \\ l(\theta) &= n\log(\theta) + (\theta - 1)\log(X_1 \cdots X_n) \\ l'(\theta) &= \frac{n}{\theta} + \log(X_1 \cdots X_n) = 0 \\ \implies -\log(X_1 \cdots X_n) &= \frac{n}{\theta}, \end{aligned}$$

which ultimately gives us:

$$\hat{\theta}_{ML} = \frac{-n}{\log(X_1 \cdots X_n)}$$

□

(b) Find $I_n(\theta)$.

Solution.

$$\begin{aligned} I_n(\theta) &= -\mathbb{E}[l''(\theta)] \\ &= -\mathbb{E}\left(\frac{-n}{\theta^2}\right) = \frac{+n}{\theta^2}, \end{aligned}$$

which gives us:

$$I_n(\theta) = \boxed{\frac{n}{\theta^2}} = \frac{1}{\theta^2}$$

□

(c) What distribution is $\hat{\theta}_{ML}$ approaching as $n \to \infty$?
Normal, with:

$$\hat{\theta} \sim \text{Normal}\left(\theta, \frac{1}{\text{I}_\text{n}(\theta)}\right)$$

which gives:

$$\hat{\theta} \approx N\left(\hat{\theta}, \frac{1}{I_n(\hat{\theta})}\right) = \boxed{N\left(\hat{\theta}, \frac{\hat{\theta}^2}{n}\right)}$$

(d) Find an approximate $(1-\alpha)100\%$ confidence interval for θ when n is large.

$$\begin{aligned} &\hat{\theta} \pm z\left(\frac{\alpha}{2}\right)\sqrt{\frac{\hat{\theta}^2}{n}} \\ &= \boxed{\frac{-n}{\log(X_1, \ldots, X_n)} \pm z\left(\frac{\alpha}{2}\right)\frac{\sqrt{n}}{\log(X_1, \ldots, X_n)}} \end{aligned}$$

Lecture ends here.

Stats 135, Fall 2019
Lecture 11, Monday, 9/23/2019

CLASS ANNOUNCEMENTS: Quiz 2 this Friday will focus on §8.5, 8.7.:

- Calculate MLE
- calculate the asymptotic variance of MLE (FI)
- 95% of MLE of parameter θ

There won't be any sufficiency on the quiz, which is something we'll start today.

1 Review

Last time, we established that $\hat{\theta}_{ML}$ has good properties:

- equivariance $\hat{g(\theta)} = g(\hat{\theta})$ (true for MOM if g is continuous)
- consistency $\hat{\theta} \xrightarrow{p} \theta$ (true for MOM)
- asymptotic: unbiased, normal, efficient

Recall that the Cramer Rao inequality says:

2 Mean Square Error (MSE) of an estimator

The MSE is used to measure how good our estimator is.

Definition: $MSE(\hat{\theta})$ **-**

The Mean Square Error of a parameter is:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right] \\ &= \iiint \cdots \int \left[\hat{\theta}(x_1,\ldots,x_n)-\theta\right]^2 f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)\; dx_1 \cdots\; dx_n \end{aligned}$$

We can think of the MSE as the average distance of $\hat{\theta}$ from θ.

When we calculate this, it'll be a function of θ, which we won't know. The main theorem that is helpful is:

Theorem 2.1.

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \left[\mathrm{Bias}(\hat{\theta})\right]^2,$$

where $\mathrm{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$.

Proof.

$$MSE(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right],$$

where Adam Lucas jokes that we then follow our algebraic heart and foil this and complete this for homework (to save time). □

The important picture to have in our heads is that we'll make two estimators that have the same MSE. Let one be an unbiased estimator for θ, so that it has high variance and low bias. We can also have the opposite situation: low variance (pointy at the center) but high bias (θ far off-center). Lucas notes there's a trade-off between variance and bias for the same value of MSE.

Example: These two estimators may have the same MSE. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, so that:

$$\mathbb{E}(X_1) = \theta, \text{Var}(X_1) = \theta(1-\theta).$$

Now we may ask: between $\tilde{\theta} = \overline{X}$ and $\hat{\theta} = X_1$, which has the smaller MSE? Notice that $\mathbb{E}(\tilde{\theta}) = \mathbb{E}(\overline{X}) = \theta$, and $\mathbb{E}(\hat{\theta}) = \mathbb{E}(X_1) = \theta$, so $\tilde{\theta}, \hat{\theta}$ are both unbiased.
Now we compare their variances:

$$\text{Var}(\tilde{\theta}) = \text{Var}(\overline{X}) = \frac{\theta(1-\theta)}{n}$$
$$\text{Var}(\hat{\theta}) = \text{Var}(X_1) = \theta(1-\theta),$$

so their MSEs are:

$$MSE(\tilde{\theta}) = \text{Var}(\overline{X}) = \frac{\theta(1-\theta)}{n}$$
$$MSE(\hat{\theta}) = \text{Var}(X_1) = \theta(1-\theta),$$

so it implies:

$$MSE(\tilde{\theta}) \leq MSE(\hat{\theta}).$$

Conclusion: Among all unbiased estimators, if MSE is the most important factor to us, then for large sample size, the MLE has the smallest possible MSE because it is efficient (that is, it achieves the Cramer Rao Lower Bound).

Note that the CR inequality only applies for unbiased estimators, and if we allow biased estimators, we might get an even smaller MSE than what we would find from unbiased estimators. That is, if we care to minimize Mean Square Error, we may want to consider biased estimators.

3 §8.8: Sufficiency

Lucas wants to motivate this section on sufficiency. There are two primary motivations for sufficient estimators.
(a) For one, once we collect all of our data, we can form a **sufficient** statistic and then throw away our data (we need not keep the large storage of data). That is, we only need the sufficient statistic to estimate the parameter θ.
(b) We can make an estimator $\hat{\theta}$ better (for example, lower MSE) by taking the conditional expectation $\tilde{\theta}(T)$ given a sufficient statistic T:

$$\tilde{\theta}(T) = \mathbb{E}\left[\hat{\theta}(X_1, \dots, X_n) \mid T\right],$$

called the Rao-Blackwell theorem.
Before we talk about a sufficient statistic, we define a statistic.

Definition: Statistic -

We say that a **statistic** $T(X_1, \ldots, X_n)$ is a function of our data **only**.

For example, a statistic could be the average $\overline{X}$, or the third element X_3, or the minimum $\min(X_1, \ldots, X_n)$, or an ordered statistic $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. However, for example, $\overline{X} - \mu$ is NOT a statistic, because it involves the parameter μ.

Definition: Sufficient Statistic -

We say that T is a **sufficient** statistic for θ if the conditional distribution of

$$X_1, \ldots, X_n \mid T$$

dependent on T no longer depends on θ.

Lucas notes this is quite abstract. For example, take:

$$f(X_1, \ldots, X_n \mid T = t, \theta) = f(X_1, \ldots, X_n \mid T = t),$$

where essentially that $T = t$ contains all the data, so conditioning on θ is the same as not conditioning on θ. That is, on the RHS, the conditional density does not depend on θ.
First, recall Bayes' Rule, which gives:

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

and for densities, Bayes' Rule gives:

$$f(X_1, \ldots, X_n \mid T = t, \theta) = \frac{\overbrace{f(X_1 = x, \ldots, X_n = x_n, T = t}^{} \mid \theta)}{f(T = t \mid \theta)},$$

where Lucas notes that the overbraced portion is truly overkill because T is a function of the $X_1, \ldots, X_n$. In other words, determining all the $X_i = x_i$ makes it so $T = t$ does not matter. Here's a simple example:

Example: Let $X_1, X_2 \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then

$$\begin{aligned} f(X_1, X_2 \mid \theta) &= \mathbb{P}(X_1 = x_1 \mid \theta) \cdot \mathbb{P}(X_2 = x_2 \mid \theta) \\ &= \underbrace{\theta}_{x_1 + x_2} \underbrace{(1 - \theta)}_{2 - x_1 - x_2}. \end{aligned}$$

Now if $T := X_1 + X_2$, then

$$f(t \mid \theta) = \mathbb{P}(X_1 + X_2 = t \mid \theta) = \sum_{(x_1, x_2): x_1 + x_2 = t} \mathbb{P}(x_1, x_2),$$

Then we can make a table for this example. The question is if $T = X_1 + X_2$ is a sufficient statistic. Then the right-most table should not depend on θ.

(x_1, x_2)	$T = X_1 + X_2$	$f(x_1, x_2 \mid T = t, \theta) = \frac{f(x_1, x_2 \mid \theta)}{f(t \mid \theta)}$
$(0,0)$	$t = 0$	$\frac{f(0,0\mid\theta)}{f(x_1+x_2=0\mid\theta)} = \frac{(1-\theta)(1-\theta)}{(1-\theta)(1-\theta)} = 1$
$(1,0)$	$t = 1$	$\frac{f(1,0\mid\theta)}{f(x_1+x_2=1\mid\theta)} = \frac{\theta(1-\theta)}{\theta(1-\theta)+(1-\theta)\theta} = \frac{1}{2}$
$(0,1)$	$t = 1$	$\frac{f(0,1\mid\theta)}{f(x_1+x_2=1\mid\theta)} = \frac{(1-\theta)\theta}{\theta(1-\theta)+(1-\theta)\theta} = \frac{1}{2}$
$(1,1)$	$t = 2$	1

and because these conditionals do not depend on θ, this is an example of a sufficient statistic.

Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$. Then recall:

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

and so

$$f(X_1, \ldots, X_n \mid \theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n X_i}}{(X_1!)(X_2!)\cdots(X_n!)}$$

We will show that $T := X_1 + \cdots + X_n$ is a sufficient statistic. Lucas notes that instead of having to worry about all the data we can have, we worry only about the sums they can have. This is a reduction in storage.
Now we recall that the sum of n independent poisson distributions is poisson, so:

$$T \sim \text{Poisson(n}\theta),$$

so

$$f(t|\theta) = \frac{e^{-n\theta}(n\theta)^t}{t!}.$$

Now we write:

$$\begin{aligned} f(X_1, \ldots, X_n \mid T = t, \theta) &= \frac{f(X_1, \ldots, X_n \mid \theta)}{f(t|\theta)} \\ &= \frac{\frac{e^{n\theta}\theta^{\sum X_i}}{(X_1!)\cdots(X_n!)}}{\frac{e^{-n\theta}n^t\theta^t}{\cdots}}, \end{aligned}$$

which proves this definition of T is sufficient.

Lecture ends here.

Stats 135, Fall 2019
Lecture 12, Wednesday, 9/25/2019

1 Review

Last time, we went over §8.8 and sufficiency. Recall our definition that T is a sufficient statistic for θ if the conditional distribution of $X_1, \dots, X_n \mid T$ does not depend on θ.
Essentially, it allows data reduction (we need not store all the data).
We worked through the example of $X_1, X_2 \sim$ iid Bernoulli(θ), and the sufficient statistic of the sample data $T := \overline{X}$. We computed the conditional densitites $f(X_1, X_2 \mid T = t, \theta)$ and we saw that these are not dependent on θ anymore. We are specifying (fixing) θ.
Lucas notes that obviously setting $S := X_1$ (only the first data point) does not give a sufficient statistic. Suppose $(X_1 = 0, X_2 = 0)$, so that $S = 0$. Then

$$f(X_1, X_2 \mid S = 0, \theta) = \frac{f(X_1, X_2 \mid \theta)}{f(S)} = \frac{(1-\theta)(1-\theta)}{(1-\theta)(1-\theta) + (1-\theta)\theta},$$

which certainly is a function of θ.

2 Inventing Shorthand Notation

We'll write $X^n := (X_1, \dots, X_n)$ and $Y^n := (Y_1, \dots, Y_n)$.

Theorem 2.1. (Factorization Theorem)
T is sufficient if and only if

$$f(X^n \mid \theta) = g\left(T(X^n), \theta\right) h(X^n).$$

See Thm A in the book. From this we see that an MLE estimate of θ optimizes $g(T(X^n), \theta)$ and hence is a function of T. In other words, we don't need to know all of our data; only $T(X^n)$. Recall that as a function of θ, the MLE is the θ that maximizes $f(X^n \mid \theta)$, and this is the same as maximizing $g(T(X^n), \theta)$. Notice that this function depends on $T(X^n)$ (not necessarily all the data).
This is the Corollary A (p. 309) of Rice, which states:

If T is sufficient for θ, the MLE is a function of T.

Example: Let $X^n \overset{iid}{\sim}$ Bernoulli(θ).
Then

$$g(X^n|\theta) = \underbrace{\theta^{\sum X_i}(1-\theta)^{n-\sum X_i}}_{g(T(X^n,\theta))} \cdot \underbrace{1}_{h(X^n)},$$

where $T(X^n) = \sum X_i$. This proves $T := \sum X_i$ is sufficient for this example. If, for example, we want to show that $\sum X_i$ is a sufficient statistic, we could use this factorization theorem and say that $\sum X_i$ is only a part of this function g. Lucas notes that we may have a T in mind, otherwise it's an argument.

Remark: If we define $T := (X_1, \dots, X_n)$, taking T to be all our data, this would trivially be sufficient by factorization. But $T := \sum X_i$ is better as it leads to a data reduction.

We will eventually get to a minimal sufficient statistic.

2.1 A more complicated example

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then

$$\begin{aligned} f(X^n|\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i-\mu)^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2}. \end{aligned}$$

Lucas says we need to do some massaging. The trick here is that we can write

$$\sum(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu)^2,$$

where this follows from some algebra. Now given that we can do this, we see:

$$\begin{aligned} f(X^n|\mu) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-\mu)^2\right]} \\ &= \underbrace{e^{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}}}_{g(T(X^n),\mu)} \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum(X_i-\overline{X})^2}{2\sigma^2}}}_{h(X^n)} \end{aligned}$$

This implies that $T := \overline{X}$ is a sufficient statistic.

Remark: Note that we CANNOT factor $f(X^n|\mu)$ as:

$$f(X^n|\mu) = \underbrace{e^{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}} e^{-\frac{\sum(X_i-\overline{X})^2}{2\sigma^2}}}_{g(\overline{X},\mu)} \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}}_{h(X^n)}.$$

In other words, because this left factor isn't a function of only $\overline{X}$ and μ (it is also a function of $S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$), then $T := (\overline{X}, S^2)$ is a sufficient statistic for $\theta = \mu$ (with σ^2 known and constant), but it is a higher-dimensional statistic than $T = \overline{X}$, which is known as a minimal statistic (lowest-dimensional possible).

2.2 Yet Another Example

Suppose σ^2 is NOT known. Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Then

$$f(X_1, \dots, X_n \mid \mu, \sigma^2) = \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{(n-1)S^2}{2\sigma^2}\right) \cdot \exp\left(\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right)}_{g(\overline{X},S^2,\mu,\sigma^2)} \cdot \underbrace{1}_{h(X^n)}$$

Recall that given our data, h has to be a number. This factorization shows that

$$T := \left(\overline{X}, S^2\right)$$

is a sufficient statistic for $\theta = (\mu, \sigma^2)$.
Here we need $T = (\overline{X}, S^2)$ since σ^2 is not a constatnt.

2.3 And more examples

Example: Consider $\hat{\theta}_{ML} = \overline{X}$ for $X^n \sim$ Bernoulli(θ) with $T := \overline{X}$. Then

$$\hat{\theta}_{ML} = \overline{X},$$

and our ML estimator is our sufficient statistic.

Example: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.
$T = (\overline{X}, S^2)$, $\hat{\theta}_{ML} = \left(\overline{X}, \frac{n-1}{n} S^2\right)$.
The question arises as to why we do not set the sufficient statistic simply to the ML estimator.
Lucas notes that we have an entire family of estimators that are sufficient and this is to introduce that concept.

3 Partitions

Definition: Likelihood Parition (LP) -

We say that likelihoods $f(X^n|\theta)$ and $f(y^n|\theta)$ are equivalent if

$$f(X^n|\theta) = f(Y^n|\theta)$$

for some constant c that may depend on $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, but not θ. This is called a likelihood partition (LP).

Let's take a look at our familiar example.

Example: Take $X_1, X_2 \sim iid$Bernoulli(θ).
$(1-\theta)\theta$ and itself for $(0,1)$ and $(1,0)$ form a partition because they are multiples (equal) to one another. However, $(1-\theta)\theta$ is not a multiple of $(1-\theta)^2$ or θ^2, so those are in separate partitions.

Definition: Statistic Partition (SP) -
Statistics $T(X^n)$, $T(Y^n)$ are equivalent if

$$T(X^n) = Y(y^n).$$

This is called a statistic partition (SP).

The likelihood partition is the "coarsest" partition (the minimal sufficient statistic), whereas taking all our data gives the "finest" partition (every data entry is in a separate cell).
One way to think about it is to start with our LP as our minimal sufficient statistic, and we get other sufficient statistics by 'adding lines' into our partition.

Definition: -

We say that a statistic $T(X^n)$ is sufficient for θ if and only if SP $\subseteq$ LP, which is to say that SP is contained in (finer than) LP, and that we get SP from LP by adding lines.

Now we can define:

Definition: Minimal Sufficiency -

A statistic is a minimal sufficient statistic (MSS) if SP = LP (in that we get the same partition).

Let's look at a more complicated example, 3 bernoullis (8 possible values). We found that $T_0 := X^3, T_3 := \frac{\sum X_i}{30}, T_4 := (T_3, X_1)$ are all sufficient, where T_3 is minimally sufficient. We arrive at this by comparing against LP.
Now we want to end with:

4 How to find a minimal sufficient stistic

§8.8
We defined that T is MSS (minimal sufficient statistic) if SP = LP.
In other words, T is MSS if the following is true:

$$T(X^n) = T(Y^n) \iff \frac{f(Y^n|\theta)}{f(X^n|\theta)}$$

does not depend on θ. On the left we have equivalence of statistics, and on the right we have equivalence of likelihoods.
Find a T such that these are equal exactly when the ratio on the right does not depend on θ. We'll look at an example.

Example: Take $X_1, X_2 \sim$ Bernoulli(θ). Find MSS.
Write the ratio of the likelihoods:

$$\begin{aligned}\frac{f(Y_1, Y_2|\theta)}{f(X_1, X_2|\theta)} &= \frac{\theta^{\sum(Y_i)}(1-\theta)^{(2-\sum Y_i)}}{\theta^{(\sum X_i)}(1-\theta)^{(2-\sum X_i)}} \\ &= \theta^{(\sum Y_i - \sum X_i)}(1-\theta)^{(\sum X_i - \sum Y_i)},\end{aligned}$$

and we have this ratio not depend on θ exactly when these sums are equal, so that the exponents are zero. Hence

$$T := \sum X_i$$

is a MSS.

Lecture ends here.

Stats 135, Fall 2019
Lecture 13, Monday, 9/30/2019

1 Review

Last time, we worked through §9.2, the Neyman-Pearson Paradigm. We have the decision function

$$d(x) = \begin{cases} 1, & \text{if reject } H_0 \\ 0, & \text{if accept } H_0 \end{cases}$$

A simple hypothesis has null and alternative distribution is specified. There are two types of error:

type 1 error: $\alpha = \mathbb{P}_0(d(x) = 1)$, level of significance
type 2 error: $\beta = \mathbb{P}_1(d_(x) = 0)$.

We set a cap for α and want to minimize β, given that α.
Many believe that a type 1 error is worse than a type 2 error. We will see that α and β are negatively correlated so lowering one type of error makes the other type larger.
Hence we fit α at a tolerable level (say $\alpha = 0.05$) and design our experiment to minimize β with that fixed value of α. We call α the **significance level** of the test.

Topics Today:

- Likelihood Ratio Test
- Power of a test
- p-value

2 Likelihood Ratio Test (LR)

A **test statistic (TS)** is a function of your sample that leads you to a decision whether to accept or reject the null (hypothesis). Usually we choose a TS that has a distribution that we know.
Typically we'll work with the χ-squared distribution, but we'll first work with some basic examples.

Example: Suppose there are 2 coins, where coin 0 has $p = .5$ of landing heads, and coin 1 has $p = .7$ of landing heads. We don't know which coin we have.
Let H_0 (our null hypothesis) is that we have a fair coin (coin 0). Our alternative H_1 is that we have coin 1 (biased with $p = .7$). To test this, we flip our coin 3 times. Let X be the number of heads in 3 tosses. Then consider the following table:

x	$f_0(x)$	$f_1(x)$	$\Lambda(x) = \frac{f_0(x)}{f_1(x)}$
3	$\binom{3}{3}(.5)^3 = .125$	.343	.364
2	$\binom{3}{2}(.5)^3 = .375$	.411	.850
1	.375	.189	1.984
0	.125	.027	4.630

Our acceptance region is in the bottom half (high LR), and the rejection region is in the top half (low LR). We reject the the null H_0 for small Λ.

Definition: Likelihood Ratio Test -

For a fixed α, the LRT (likelihood ratio test) says to reject H_0 if $\Lambda < c$ (for some c we need to specify).
The cutoff c is a function of α since

$$\alpha = \mathbb{P}_0(d(x) = 1) = \mathbb{P}_0(\Lambda < c).$$

For $\alpha = .5$,

$$.5 = \mathbb{P}_0(\Lambda < c).$$

Now from the table above, $\mathbb{P}_0(\Lambda < 1.984) = .125 + .375 = .5$ implies $c = 1.984$.
Then the LRT is to reject H_0 if $\Lambda < 1.984$, or (because we know $x \sim$ Binomial), we reject H_0 if $x > 1$.

Definition: Rejection Region, Acceptance Region -

The region of values of Λ in which we reject H_0 is called the **rejection region**.

Definition: Power of a test -

We define

$$\text{Power} = 1 - \beta = 1 - \mathbb{P}_1(d(x) = 0) = \mathbb{P}_1(d(x) = 1).$$

Lucas notes that for powers, the subscript of $\mathbb{P}$ on will be the same (1) as the equality. In the previous example,

$$\begin{aligned} \text{Power} = \mathbb{P}_1(d(x) = 1) &= \mathbb{P}_1(\Lambda < 1.984) \\ &= .343 + .441 \\ &= .784 \end{aligned}$$

To summarize the flow, we fixed α and found the cutoff using the null distribution. Then, we find the power using the alternative distribution. Let's do another example.

Example: (Finding the rejection and acceptance region of an LRT).
Suppose under the null that a random variable X has a uniform distribution on $[0, 1]$, and under the alternative, it has density $f(x) = 2x$, for $0 < x < 1$.

(a) What is the LRT at $\alpha = .1$ level of significance? Doing so, we get the cutoff, which we can use to solve the next problem.

We draw a picture (plot) of the densities, where H_1 takes on a line with slope 2, and H_0 is the uniform density on $[0, 1]$. Then

$$\alpha = \mathbb{P}_0(d(x) = 1) = .1,$$

so LRT rejects H_0 if $\Lambda = \frac{f_0(x)}{f_1(x)} < c$. That is,

$$\alpha = .1 = \mathbb{P}_0\left(\frac{1}{2x} < c\right)$$
$$.1 = \mathbb{P}_0\left(x > \frac{1}{2c}\right) \implies \frac{1}{2c} = .9 \implies c = \frac{1}{1.8} = \boxed{\frac{5}{9}}$$

Then we conclude that LRT rejects H_0 if $\Gamma < \frac{5}{9}$ or reject H_0 if $x > .9$.

Later in the course, we'll draw the distribution with respect to Λ, instead of the distribution with respect to x.

(b) What is the power of the test?

We found that .9 is the magical cutoff, and so the power, $1 - \beta$ is the area from .9 to 1 under the alternative distribution. That is,

$$\begin{aligned}\text{Power} &= \mathbb{P}_1(d(x) = 1) \\ &= \mathbb{P}(X > .9) \\ &= \int_{.9}^{1} 2x \, dx = x^2\Big|_{.}^{} 9^1 = 1 - .81 = .19.\end{aligned}$$

3 Neyman Pearson Lemma

Lucas notes this is an important lemma.

Lemma 3.1. Suppose that H_0, H_1 are simple hypothesis (that is, we know the distributions for both of them). Suppose that the LRT rejects H_0 with significance level α. Then any other test for which the significance level is α has less power.

This is a great result. This means that the likelihood ratio test that we have (that we will reject when $\Lambda < c$) is as good as we can get. We may be able to cook up other tests, but they will not be as good. So for simple hypothesis, likelihood ratio tests are "the bomb," as given by Lucas. In other words, LRT is the best we can do for hypothesis testing (HT) of simple hypothesis.

4 p-values

We want to quickly talk about p-values for a test statistic, as it's quite easy.

Definition: p-value -

A test statistic's p-value is the probability that it is as or more extreme than what we observe.

Let's look at a tricky example. Suppose a TS (test statistic), T, has the standard normal distribution. If the test rejects for large $|T|$, what is the p-value if we observe -1.5? (In other words, we reject if $|T| < c$ for $c > 0$). This means that the p-value is

$$\begin{aligned} p\text{- value} &= \mathbb{P}_0(T < -1.5 \text{ or } T > 1.5) \\ &= 2(1 - \Phi(1.5)) \\ &= 2(.068) \\ &= .1336. \end{aligned}$$

The important thing is to note that if the likelihood ratio test says to reject for large negative values, then the p value is the probability that T is less than -1.5 (in other words, we would have one tail). That is, p-value= $\mathbb{P}_0(T < -1.5)$. But in our situation, we consider both tails and double this value.

Lecture ends here.

Stats 135, Fall 2019
Lecture 14, Wednesday, 10/2/2019

1 Review

Last time, we went through §9.2 and defined two types of errors:

$$\alpha = \mathbb{P}_0(d(x) = 1), \text{significance level, often .05}$$
$$\beta = \mathbb{P}_1(d(x) = 0),$$

and we set that LRT (likelihood ratio test) gives: reject H_0 if

$$\Lambda = \frac{f_0(x)}{f_1(x)} < c,$$

for some c determined by (fixed, pre-determined) α.
We defined the **power** (which we want to maximize) as

$$1 - \beta = \mathbb{P}_1(d(x) = 1)$$

Then the Neyman Pearson Lemma (NPL) just says that we'll reject half the line. It says that LRT gives the most powerful test between simple hypothesis (taking the rejection region of $\Lambda < c$).

Topics Today:

- Review LRT and Power
- "Uniformly most power test" is a LRT, most powerful, even for non-simple hypothesis. This usually doesn't exist, but there are cases where it does.
- Power curve: For simple hypotheses, we can compute the power (for the alternative). However, for nonsimple hypotheses, there will be a power curve instead.

2 Review Likelihood Ratio Test (LRT)

Example: Let $X \sim$ Exponential(λ) (like waiting times for something to happen), with

$$f(x) = \lambda e^{-\lambda x}, x > 0.$$

Then take $H_0 : \lambda = 1$ and $H_1 : \lambda = 2$, with $\alpha := .05$.
Our cutoff will look something like $x < c$ by our sketch of the graphs. We can tell that the rejection region will be .

(a) Lucas tasks us to find the LRT.

LRT says to reject H_0 if $X < k$. By drawing the picture and knowing the cut-off is on the left, take

$$\alpha = \mathbb{P}_0(X < k) = \int_0^k e^{-x}\, dx = 1 - e^{-k}$$
$$\implies e^{-k} = 1 - \alpha = 0.95$$
$$\implies k = -\log(.95) = 0.0513$$

Equivalently, by definition, in terms of $\Lambda = \frac{f_1(x)}{f_0(x)}$, we have:

$$\begin{aligned}
\alpha &= \mathbb{P}_0(\Lambda < c) \\
&= \mathbb{P}_0\left(\frac{1}{2}e^{-x} < c\right) \\
&= \mathbb{P}_0\left(x < \underbrace{\log(2c)}_{=:k}\right) \\
&= \int_0^k e^{-x}\, dx \\
&= -e^{-x}\Big|_0^k \\
&= 1 - e^{-k} \\
\implies e^{-k} &= 1 - \alpha = .95 \\
\implies k &= -\log(.95) = 0.0513
\end{aligned}$$

(b) For $\alpha = .04$, what is the maximum power that we can achieve, among all hypothesis tests?

LRT is most powerful by NPL (Neyman Pearson Lemma), with

$$\text{Power} = \mathbb{P}_1(d(x) = 1) = \int_0^{.0513} 2e^{-2x}\, dx = -e^{-2x}\Big|_0^{0.0513} = 1 - e^{-2(.053)}$$

3 §9.2.3: Uniformly Most Powerful Test (UMPT)

Definition: Uniformly Most Powerful Test -

The NPL requires H_0, H_1 to be simple. If the null is simple and the alternative is composite (for example, say $H_0 : \theta = 2, H_1 : \theta > 2$), then a likelihood ratio test (LRT) that is most powerful for **every** simple alternative is called **uniformly most powerful test**.

This test must have the same rejection region for every single possible alternative. We ask, does this exist often? Lucas says this is a start to considering composite alternatives.
Otherwise, we would use a generalized likelihood ratio test, which is much more complicated.

Example: Let X be a single observation from a probability density function

$$f(x|\theta) = \frac{1}{2}\theta \exp(-\theta|x|),$$

which we call the Laplace distribution. Find the most powerful test for $\alpha = .05$, testing $H_0 : \theta = 1$ and $H_1 : \theta = \frac{1}{2}$.
From the plot, we gain insight to reject the null if $|x| > k$. We need to find the appropriate k. So we have

$$\begin{aligned}
\alpha = \mathbb{P}_0(|x| > k) = 2\mathbb{P}_0(X > k) &= \cancel{2}\int_k^\infty \frac{\cancel{1}}{\cancel{2}}e^{-x}\, dx \\
&= -\exp(-x)\Big|_k^\infty = \exp(-k),
\end{aligned}$$

so

$$e^{-k} = .05 \implies k = \log(.05) = 3,$$

where log denotes the natural log. Hence the likelihood ratio test tells us to reject H_0 if $|x| > 3$.

Next, let's see if this test is UMP (uniformly most powerful) for composite alternative, say:

$$H_0 : \theta = 1$$
$$H_1 : \theta = b, 0 < b < 1$$

We want to reject H_0 if $|x| > k$. The question is if this is a function of b? Consider:

$$\begin{aligned}\alpha = \mathbb{P}_0(|x| > k) &= 2\mathbb{P}_0(x > k) \\ &= 2\int_k^\infty \frac{1}{2}\exp(-x)\ dx \\ &= -\exp(-x)\big|_k^\infty \\ &= \exp(-k) = .5,\end{aligned}$$

where $k = 3$ and is not a function of b. Hence the rejection region is the same for all b in $0 < b < 1$. So the likelihood ratio test that we have uniformly most powerful (UMP) over this range of values of b!.
Now we may ask, what happens when $b > 1$? We'll check this quickly. When $b > 1$, our rejection region changes:

$$\begin{aligned}\alpha = \mathbb{P}(|x| < k) &= 2\mathbb{P}_0(x < k) \\ &= 2\int_0^k \frac{1}{2}\exp(-x)\ dx \\ &= -\exp(-x)\big|_0^k \\ \implies e^{-k} &= .95 \\ \implies k = -\log(.95) &= \boxed{.0513}.\end{aligned}$$

Now we're going to reject the null H_0 if $|x| < .0513$, and we conclude that the test is uniformly most powerful for $0 < b < 1$ as well as $b > 1$, but not for $b < 0$ or $b = 1$.

4 Power Curve

Example: Let's find the power curve for the Laplace example. Recall that

$$\text{Power} = P_1(d(x) = 1),$$

and take $H_0 : \theta = 1$ and $H_1 :$ any $\theta > 0$.
Then for $0 < \theta < 1$,

$$\text{Power} = \mathbb{P}_1(|x| > 3) = 2\int_3^\infty \frac{\theta}{2}\exp(-\theta x)\ dx = \exp(-3\theta),$$

once we've solved this.
Consider then $\theta > 1$, so that the rejection region is now inside. Then

$$\text{Power} = \mathbb{P}_1(|x| < .051) = 2\int_0^{.051} \frac{\theta}{2}\exp(-\theta x)\ dx = 1 - \exp(-.051\theta)$$

Now these are both functions of θ. This gives the power curve. We'll analyze this more in coming days.

Lecture ends here.

Stats 135, Fall 2019
Lecture 15, Friday, 10/4/2019

1 Review

Last time we worked through the Likelihood Ratio Test (LRT) for a simple hypothesis where

$$\Lambda = \frac{f_0(x|\theta)}{f_1(x|\theta)}$$

,
We consider an alternative version: the generalized likelihood ratio test:
Let $\Omega = \{\theta_0, \theta_1\}$ and let

$$\tilde{\Lambda} = \frac{f_0(x|\theta)}{\max_{\theta\in\Omega} f(x|\theta)},$$

where the denominator is just the likelihood at the MLE. Then we would reject if $\tilde{\Lambda} < c$. Notice that these two are equivalent, where small values of Λ corresponds to small values of $\tilde{\Lambda}$.

Topics Today:

- Duality of Confidence Interval and Hypothesis Test
- Generalized Likelihood Ratio Test (GLRT)

2 Duality of CI and HT

In words, the test statistics (TS) for a lever α hypothesis test (HT) with null $H_0 : \theta = \theta_0$ lies in the acceptance region if and only if the null value θ_0 lies in a $100(1-\alpha)$ confidence interval for θ. Hence we see this connection between confidence intervals and hypothesis tests. This is the duality, and we'll demonstrate this through an example.

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where σ^2 is known. Take the null $H_0 : \mu = \mu_0$ and the alternative $\mu \neq \mu_2$. Now fix α level of significant, and we sill see later today that the acceptance region of GLRT for this hypothesis test is

$$\overline{X} = \mu_0 \pm z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$$

We draw a picture of x bar under the null, centered at mu_0.
Notice that mathematically, the following two inequalities are equivalent.

$$\mu_0 - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu_0 + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}},$$

which defines the acceptance region, and

$$\overline{X} - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \mu_0 < \overline{X} + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}},$$

which gives the $100(1-\alpha)\%$ confidence interval for μ.
It follows that we accept for the null when $\overline{X}$ is in the acceptance region or equivalently when the null parameter μ_0 is in the $100(1-\alpha)\%$ confidence interval for μ. This is certainly obvious for our example, and for generalizing, see proof of A,B of §9.3.

3 Generalized Likelihood Ratio Test (GLRT)

Most hypothesis tests are not simple, and the UMPT often doesn't work in the range of alternative values that we need. Then in this case, we use a GLRT.
Let $\Omega := \omega_0 \oplus \omega_1$, the direct sum (disjoint union). Consider the null $H_0 : \theta \in \omega_0$ and the alternative $H_1 : \theta \in \omega_1$.
Then define

$$\Lambda := \frac{\max_{\theta\in\omega_0}(\text{lik}(\theta))}{\max_{\theta\in\Omega}(\text{lik}(\theta))}.$$

Notice the denominator is just simply $\text{lik}(\hat{\theta}_{ML})$.
Then the GLRT tells us to reject the null H_0 if $\Lambda < c$, and otherwise accept.
Example:
We would like to test $H_0 : \mu = \mu_0$, where μ_0 is fixed. Take the alternative to be $H_0 : \mu \neq \mu_0$. Here, $\omega_0 = \{\mu_0\}$ and $\omega_1 = \mathbb{R} - \{\mu_0\}$. This is so that $\Omega = \mathbb{R}$. Then there is just one element, so

$$\begin{aligned}\max_{\theta\in\omega_0}(\text{lik}(\theta)) &= \text{lik}(\mu_0) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2\right),\end{aligned}$$

and so we write

$$\begin{aligned}\max_{\theta\in\Omega}(\text{lik}(\theta)) &= \text{lik}(\overbrace{\hat{\theta}_{ML}}^{\overline{X}}) \\ &= \cancel{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n (X_i-\overline{X})^2\right)\end{aligned}$$

Then we have

$$\Lambda = \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n \left[(X_i-\mu_0)^2 - (X_i-\overline{X})^2\right]\right)$$

so then

$$\Lambda < c \iff -2\log\Lambda > -2\log c,$$

which gives

$$-2\log\Lambda = \frac{1}{\sigma^2}\sum_{i=1}^n \left[(X_i-\mu_0)^2 - (X_i-\overline{X})^2\right],$$

which after some algebra gives

$$\frac{n}{\sigma^2}(\overline{X}-\mu_0)^2.$$

Hence $-2\log\Lambda > k$, which gives

$$\frac{n}{\sigma^2}(\overline{X}-\mu_0)^2 > k,$$

which is a chi square with 1 degree of freedom: $\left(\frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)^2 = z^2 \sim \chi_1^2$.

It turns out that asymptotically, for large n, we will have a χ-square distribution. So the Generalized Likelihood Ratio Test says to reject H_0 if

$$\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > \chi_1^2(\alpha).$$

This converts a two-sided test into a one-sided test.
We find that the acceptance region is

$$\mu_0 - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu_0 + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}.$$

However, this only works for this particular example with $X_1, \ldots, X_n \sim N(\mu, \sigma)$.
Now more generally,

Theorem 3.1. (Thm A Rice p. 341):
If the likelihood function of X is smooth, then the null distribution of $-2\log\Lambda$ approaches $\chi^2_{\dim\Omega - \dim\omega_0}$.

In our last example, we had $\Omega = \mathbb{R} - \dim 1$ and $\omega_0 = \{\mu_0\} - \dim 0$.

4 Applications

Suppose a pollster wants to know the fraction of the population θ that supports a particular legislative bill. They want to test if $\theta = \frac{1}{2}$ versus $\theta \neq \frac{1}{2}$. What is the GLRT at a 5% level of significance?

We have $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then the likelihood is given by
Recall that $\hat{\theta}_{ML} = \overline{X}$. Then $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta \neq \frac{1}{2}$ gives:

$$\boxed{\Gamma = \frac{\left(\frac{1}{2}\right)^n}{\overline{X}^{n\overline{X}}(1-\overline{X})^{n-n\overline{X}}}},$$

which we can find from data.
The GLRT tells us to reject the null for $-2\log\Lambda > k$ and accept otherwise.
For large n,

$$-2\log\Lambda \sim \chi^2_{1-0},$$

under the null distribution. Then

$$k = \chi_1^2(.05) = \text{qchisq}(.95, \text{df} = 1) = 3.84.$$

Hence the GLRT tells us to reject H_0 if $-2\log\Lambda > 3.84$, and accept otherwise.
The bias then is $\mathbb{E}(\hat{)}$

Stats 135, Fall 2019
Lecture 16, Friday, 10/4/2019

1 Review

Last time we worked through the Likelihood Ratio Test (LRT) for a simple hypothesis where

$$\Lambda = \frac{f_0(x|\theta)}{f_1(x|\theta)}$$

,
We consider an alternative version: the generalized likelihood ratio test:
Let $\Omega = \{\theta_0, \theta_1\}$ and let

$$\tilde{\Lambda} = \frac{f_0(x|\theta)}{\max_{\theta\in\Omega} f(x|\theta)},$$

where the denominator is just the likelihood at the MLE. Then we would reject if $\tilde{\Lambda} < c$. Notice that these two are equivalent, where small values of Λ corresponds to small values of $\tilde{\Lambda}$.

Topics Today:

- Duality of Confidence Interval and Hypothesis Test
- Generalized Likelihood Ratio Test (GLRT)

2 Duality of CI and HT

In words, the test statistics (TS) for a lever α hypothesis test (HT) with null $H_0 : \theta = \theta_0$ lies in the acceptance region if and only if the null value θ_0 lies in a $100(1-\alpha)$ confidence interval for θ. Hence we see this connection between confidence intervals and hypothesis tests. This is the duality, and we'll demonstrate this through an example.

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where σ^2 is known. Take the null $H_0 : \mu = \mu_0$ and the alternative $\mu \neq \mu_2$. Now fix α level of significant, and we sill see later today that the acceptance region of GLRT for this hypothesis test is

$$\overline{X} = \mu_0 \pm z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}$$

We draw a picture of x bar under the null, centered at mu_0.
Notice that mathematically, the following two inequalities are equivalent.

$$\mu_0 - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu_0 + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}},$$

which defines the acceptance region, and

$$\overline{X} - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \mu_0 < \overline{X} + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}},$$

which gives the $100(1-\alpha)\%$ confidence interval for μ.
It follows that we accept for the null when $\overline{X}$ is in the acceptance region or equivalently when the null parameter μ_0 is in the $100(1-\alpha)\%$ confidence interval for μ. This is certainly obvious for our example, and for generalizing, see proof of A,B of §9.3.

3 Generalized Likelihood Ratio Test (GLRT)

Most hypothesis tests are not simple, and the UMPT often doesn't work in the range of alternative values that we need. Then in this case, we use a GLRT.
Let $\Omega := \omega_0 \oplus \omega_1$, the direct sum (disjoint union). Consider the null $H_0 : \theta \in \omega_0$ and the alternative $H_1 : \theta \in \omega_1$.
Then define

$$\Lambda := \frac{\max_{\theta\in\omega_0}(\text{lik}(\theta))}{\max_{\theta\in\Omega}(\text{lik}(\theta))}.$$

Notice the denominator is just simply $\text{lik}(\hat{\theta}_{ML})$.
Then the GLRT tells us to reject the null H_0 if $\Lambda < c$, and otherwise accept.
Example:
We would like to test $H_0 : \mu = \mu_0$, where μ_0 is fixed. Take the alternative to be $H_0 : \mu \neq \mu_0$. Here, $\omega_0 = \{\mu_0\}$ and $\omega_1 = \mathbb{R} - \{\mu_0\}$. This is so that $\Omega = \mathbb{R}$. Then there is just one element, so

$$\begin{aligned}\max_{\theta\in\omega_0}(\text{lik}(\theta)) &= \text{lik}(\mu_0) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2\right),\end{aligned}$$

and so we write

$$\begin{aligned}\max_{\theta\in\Omega}(\text{lik}(\theta)) &= \text{lik}(\overbrace{\hat{\theta}_{ML}}^{\overline{X}}) \\ &= \cancel{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n (X_i-\overline{X})^2\right)\end{aligned}$$

Then we have

$$\Lambda = \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n \left[(X_i-\mu_0)^2 - (X_i-\overline{X})^2\right]\right)$$

so then

$$\Lambda < c \iff -2\log\Lambda > -2\log c,$$

which gives

$$-2\log\Lambda = \frac{1}{\sigma^2}\sum_{i=1}^n \left[(X_i-\mu_0)^2 - (X_i-\overline{X})^2\right],$$

which after some algebra gives

$$\frac{n}{\sigma^2}(\overline{X}-\mu_0)^2.$$

Hence $-2\log\Lambda > k$, which gives

$$\frac{n}{\sigma^2}(\overline{X}-\mu_0)^2 > k,$$

which is a chi square with 1 degree of freedom: $\left(\frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)^2 = z^2 \sim \chi_1^2$.

It turns out that asymptotically, for large n, we will have a χ-square distribution. So the Generalized Likelihood Ratio Test says to reject H_0 if

$$\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > \chi_1^2(\alpha).$$

This converts a two-sided test into a one-sided test.
We find that the acceptance region is

$$\mu_0 - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu_0 + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}.$$

However, this only works for this particular example with $X_1, \ldots, X_n \sim N(\mu, \sigma)$.
Now more generally,

Theorem 3.1. (Thm A Rice p. 341):
If the likelihood function of X is smooth, then the null distribution of $-2\log\Lambda$ approaches $\chi^2_{\dim\Omega - \dim\omega_0}$.

In our last example, we had $\Omega = \mathbb{R} - \dim 1$ and $\omega_0 = \{\mu_0\} - \dim 0$.

4 Applications

Suppose a pollster wants to know the fraction of the population θ that supports a particular legislative bill. They want to test if $\theta = \frac{1}{2}$ versus $\theta \neq \frac{1}{2}$. What is the GLRT at a 5% level of significance?

We have $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then the likelihood is given by
Recall that $\hat{\theta}_{ML} = \overline{X}$. Then $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta \neq \frac{1}{2}$ gives:

$$\boxed{\Gamma = \frac{\left(\frac{1}{2}\right)^n}{\overline{X}^{n\overline{X}}(1-\overline{X})^{n-n\overline{X}}}},$$

which we can find from data.
The GLRT tells us to reject the null for $-2\log\Lambda > k$ and accept otherwise. For large n,

$$-2\log\Lambda \sim \chi^2_{1-0},$$

under the null distribution. Then

$$k = \chi_1^2(.05) = \text{qchisq}(.95,\ \text{df} = 1) = 3.84.$$

Hence the GLRT tells us to reject H_0 if $-2\log\Lambda > 3.84$, and accept otherwise.

Lecture ends here.

Stats 135, Fall 2019

Lecture 17, Monday, 10/7/2019

Absent (out of state) - See attached written notes by Lucas for today's and Wednesday's lectures.

CLASS ANNOUNCEMENTS: Today marks the end of the materials that will be on the Midterm exam.

stat 135 lec 17

Last time sec 9.4 GLRT

$H_0: \theta \in \omega_0$
$H_1: \theta \in \omega_1$
$\Omega = \omega_0 \cup \omega_1$

Fix α level of significance

We reject H_0 if $\Lambda = \dfrac{\max_{\theta \in \omega_0} \{lik(\theta)\}}{\max_{\theta \in \Omega} \{lik(\theta)\}} < c$

otherwise accept H_0.

Equivalently,

GLRT: reject H_0 if $-2\log\Lambda > K$ else accept H_0,

where

$-2\log\Lambda \sim \chi^2_{d\text{im}\,\Omega - d\text{im}\,\omega_0}$ and $K = \chi^2_{d\text{im}\,\Omega - d\text{im}\,\omega_0}(\alpha)$

Discussion Question (T, F explain)

The GLR, Λ, is always ≤ 1 ?

True, the set in the denom includes the set in the numerator so the max of denom $\geq$ max of num.

End of material on midterm.

Sec 11.2 2 sample z,t test
1 sample z,t test.

Sec 11.2 Comparing 2 indep samples

Picture

$X \sim N(\mu_x, \sigma^2)$ — draw n w/ replacement

$Y \sim N(\mu_y, \sigma^2)$ — draw m w/ replacement

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

since linear combinations of indep normal is normal and variances add.

If σ^2 is known

$100(1-\alpha)\%$ CI for $\mu_x - \mu_y$ is

$$\bar{x} - \bar{y} \pm z\left(\frac{\alpha}{2}\right)\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

If σ^2 not known an unbiased estimate of σ^2 is the "pooled sample variance".

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Chk $E(S_p^2) = \dfrac{(n-1)E(S_x^2) + (m-1)E(S_y^2)}{n+m-2}$ ($E(S_x^2) = \sigma^2$, $E(S_y^2) = \sigma^2$)

$= \sigma^2$ ✓

So S_p^2 is an unbiased estimator of σ^2.

We use this particular unbiased estimator because of the following theorem.

Thm A (2 sample t test)

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma^2)$

$Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma^2)$

The T.S.

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

$S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}} = S_{\bar{X}-\bar{Y}}$

Proof/ recall $\frac{Z}{\sqrt{U/n}} \sim t_n$ where $\left.\begin{matrix} Z \sim N(0,1) \\ U \sim \chi^2_n \end{matrix}\right\}$ indep.

and $\frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1}$

$\frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{m-1}$

} indep since X and Y are indep.

$$\Rightarrow \frac{(n-1)S_x^2}{\sigma^2} + \frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

let $Z = \dfrac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$

$$U = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

then $\dfrac{Z}{\sqrt{U/_{n+m-2}}} \sim t_{n+m-2}$ by defn.

since Z and U are independent since $\bar{X}$ is indep of S_X^2 and $\bar{Y}$ is indep of S_Y^2 (see lec 8).

It remains to show

$$\frac{Z}{\sqrt{U/_{n+m-2}}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\frac{1}{\sigma}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}} \cdot \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$= \underbrace{\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}_{Z} \cdot \underbrace{\frac{1}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2(n+m-2)}}}}_{\frac{1}{\sqrt{U/_{n+m-2}}}}$$

□

From thm A:

$$S_{\bar{x}-\bar{y}} = S_p\sqrt{\frac{1}{n}+\frac{1}{m}}$$

Cor A

A $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$ is $\bar{x}-\bar{y} \pm t_{m+n-2}\left(\frac{\alpha}{2}\right) \cdot S_{\bar{x}-\bar{y}}$

Hypothesis testing for 2 sample problem using t-test.

$H_0: \mu_x - \mu_y = 0$
$H_1: \mu_x - \mu_y \neq 0$ (2 sided alternative)

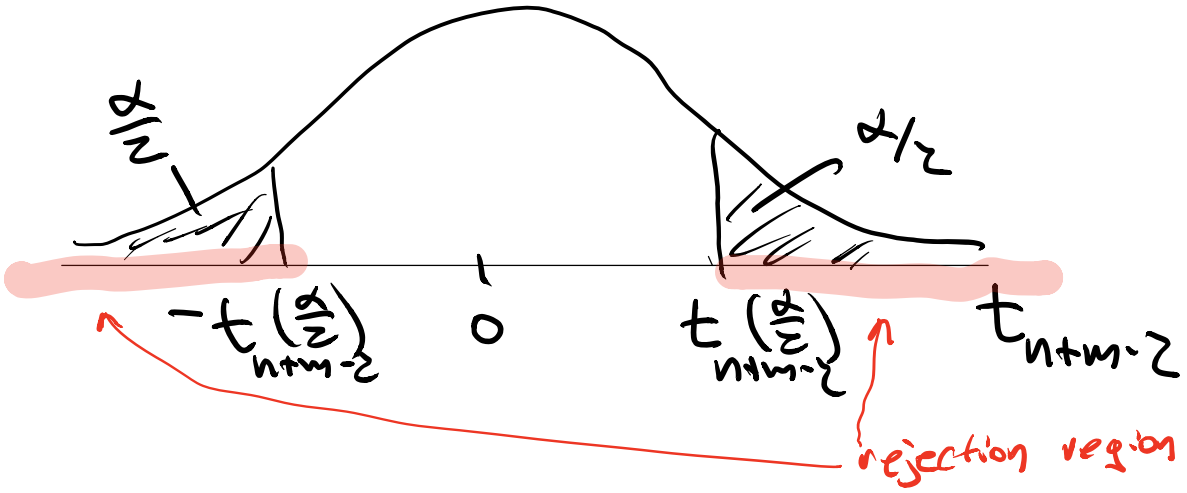
$\mu_x - \mu_y > 0$
$\mu_x - \mu_y < 0$ (1 sided alternatives)

For α level of significance

The T.S. used to make a decision whether to reject the null, assuming the null is true, is

$$t = \frac{\bar{X} - \bar{Y} - 0}{S_{\bar{X}-\bar{Y}}} \sim t_{n+m-2}$$

Picture (2 sided alternative)



We reject H_0 if

$$|t| > t_{n+m-2}^{(\frac{\alpha}{2})} \quad \text{or}$$

equivalently

$$0 \notin (\bar{X} - \bar{Y}) \pm t_{n+m-2}\left(\frac{\alpha}{2}\right) S_{\bar{X}-\bar{Y}}$$

by duality of HT and CI.

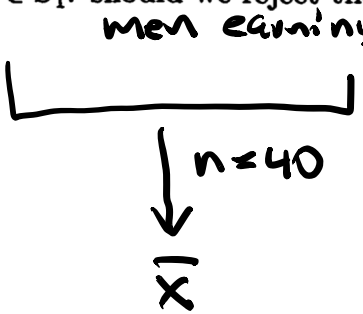
Note if n, m are large the t-test is approximately a z-test.

Fact The t, z test is equivalent to the GLRT (proof in book p 426).

1 Sample Z test

This is a simple variation of a 2 sample t test where we have just 1 large sample.

Problem (gender pay gap). The average weekly earnings for a female social worker is \$670. Do men in the same positions have average weekly earnings that are higher than those for women? A random sample of $n = 40$ male social workers showed sample mean $\overline{X} = \$725$ and sample standard deviation $s = \$102$. We want to test the hypotheses: $H_0 : \mu = 670$, $H_1 : \mu > 670$, using a significance level of $\alpha = 0.01$. Note that $n \geq 30$, which is a threshold above which we can consider $s \simeq \sigma$, and $\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim \mathcal{N}(0,1)$, approximately. Proceed as follows: **(a)** compute the "z-score", i.e. the value of the test statistic $z = (\overline{X} - \mu)/(s/\sqrt{n})$, under the null hypothesis H_0. **(b)** Find the acceptance and rejection regions S_0 and S_1 respectively. **(c)** Determine whether $z \in S_1$: should we reject the null hypothesis? **(d)** Finally, compute the p-value of the test statistic.



$H_0: \mu = 670$

$H_1: \mu > 670$

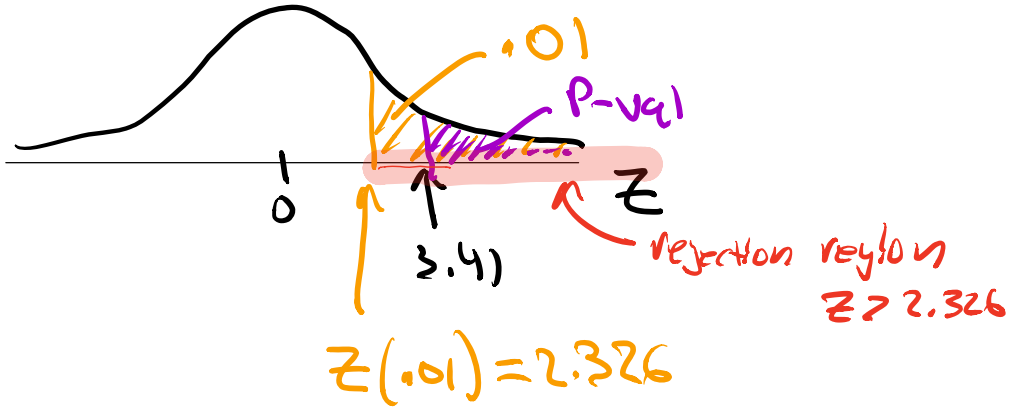
$\alpha = .01$

$\bar{x} = 725$

$S = 102$

$n = 40$

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{725 - 670}{102/\sqrt{40}} = 3.41$$



$3.41 > 2.326 \Rightarrow$ reject null

$p\text{-val} = 1 - \Phi(3.41) = .0003 < .01$

reject null.