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 $\hat{p}$.

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

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:

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

□

Lecture ends here.