API-201 ABC REVIEW SESSION #7

Friday, October 28

Table of Contents

Lecture Recap

Confidence intervals

Suppose there is some mean μ that we want to measure. We usually don't have access to information for all the population, so we estimate μ by computing the mean $\hat{\mu}$ in a random sample.

The sample mean $\hat{\mu}$ is random, as it may produce a different estimate if we apply it to a different sample. The **sampling distribution** is the distribution of the sample mean $\hat{\mu}$. In other words, it is a probability distribution formed by the estimates we obtain from calculating the mean for different samples from the population of interest.

The sample mean $\hat{\mu}$ is a random variable, so it has an expected value and a standard deviation. The **Central Limit Theorem** states that for a large enough sample (n > 30), the distribution of $\hat{\mu}$ is approximately normal:

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

The sampling distribution allows us to construct a **95% confidence interval around our point estimate**. 95% of all possible confidence intervals will contain the true value of the population parameter.

If the sampling distribution is normal, then we can construct a 95% confidence interval around $\hat{\mu}$ by using the mean and standard deviation of $\hat{\mu}$. However, given that we don't know μ, we use both the sample proportion and the estimated standard error instead, such that:

$$CI = \hat{\mu} \pm 2SD(\hat{\mu})$$

where:

$$SD(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

Constructing confidence intervals

Recall that an **estimate** is our best guess of the true value of the population parameter. There are multiple types of parameters that we could be interested in estimating. For example:

1. Proportion: Proportion of vaccinated people that got Covid.
2. Mean: Average consumption for people that had access to a microfinance program in India.
3. Difference in proportions: Difference in Covid infection rates between those who got vaccinated and those who didn't.
4. Difference in means: Difference in average consumption between treatment and control groups of a microfinance program in India.

We can use the sampling distribution of our estimator to tell us how confident we are in our estimate. The table below contains the information we need to construct a confidence interval for each of these parameters:

Population parameter	Sample parameter	Mean of sampling distribution	Standard deviation of sampling distribution
p	$\hat{p}$	$E(\hat{p}) = p$	$\sqrt{\frac{p(1-p)}{n}}$
μ	$\hat{\mu}$	$E(\hat{\mu}) = \mu$	$\frac{\sigma}{\sqrt{n}}$
$p_1 - p_2$	$\hat{p_1} - \hat{p_2}$	$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
$\mu_1 - \mu_2$	$\hat{\mu_1} - \hat{\mu_2}$	$E(\hat{\mu_1} - \hat{\mu_2}) = \mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

For example, if we wanted to estimate the difference in proportions, the 95% confidence interval has the same form as before:

$$CI = \hat{p_1} - \hat{p_2} \pm 2SD(\hat{p_1} - \hat{p_2})$$

Plugging in the standard deviation of the difference in proportions, we get:

$$CI = \hat{p_1} - \hat{p_2} \pm 2\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

Statistical significance

Hypothesis testing consists in posing a hypothesis concerning the value of a population parameter, drawing a sample from the population *under the assumption this hypothesis is true*, and then assessing the estimate with respect to the hypothesis posed. The observed value is **statistically significant** if it is unlikely under the null hypothesis. In other words, it doesn't fall within the middle 95% of the null sampling distribution.

For example, suppose the Scranton branch of a large paper company wants to test that its outgoing shipment of letter-sized paper is 8.5 inches wide on average. The company randomly samples 100 pieces of paper and finds that the average width is 8.45 inches with a standard deviation of 0.1 inches.

- The null hypothesis is $\mu = 8.5$.
- $n = 100$

- $\hat{\mu} = 8.45$
- $\hat{\sigma} = 0.1.$

The standard error of the mean is $\frac{\hat{\sigma}}{\sqrt{n}} = \frac{0.1}{\sqrt{100}} = 0.01$ Therefore the 95% confidence interval is $[8.43, 8.47]$.

The branch found that the mean width in the sample is less than 8.5, but this could just be the result of sample fluctuations. They want to know whether we have enough evidence to reject the null hypothesis that $\mu = 8.5$, so they ask: "if the average size truly is 8.5 inches, how unlikely is it to observe this sample mean?"

An estimate is statistically significant if it is at least two standard deviations away from the null hypothesis, so that we reject only the most extreme 5 percent of values. The confidence interval consists of those points within two standard deviations of the estimate. So an estimate is statistically significant if its confidence interval does not contain the null hypothesis.

In the paper example, the confidence interval does not contain 8.5, so the branch **rejects this null hypothesis.** If instead the confidence interval was wide enough to include 8.5, they would **fail to reject the null hypothesis**. This could either be because the null hypothesis is true, or because the sample size is not large enough to reject it. For example, if the true mean was $\mu = 8.4999$, the null hypothesis is technically false but because the mean is so similar, we would need to have a very large sample size to reject the null hypothesis with regularity.

The **p-value** is the probability of obtaining an estimate as extreme or more extreme than the one we obtained assuming our null hypothesis is true. A result is statistically significant if the p-value is less than 0.05. In other words, if the null hypothesis were true, we are unlikely to have obtained our result.

We compute p-values using z-scores. The z-score tells us how many standard deviations our estimate is from the null hypothesis:

$$z = \frac{x - \mu_0}{s}$$

where μ_0 is the null hypothesis and s is the standard error.

We convert z-scores to p-values with the **pnorm** function in R. For any number z, pnorm(z) returns the probability that the normal random variable is less than z standard deviations below its mean. In order to get both of the "tails" on the extremes of our estimate, we need to make sure we use a negative Z-score and double the value to account for both tails. So if our z-score is z, we can calculate the p-value as `2 * pnorm(-abs(z))`.

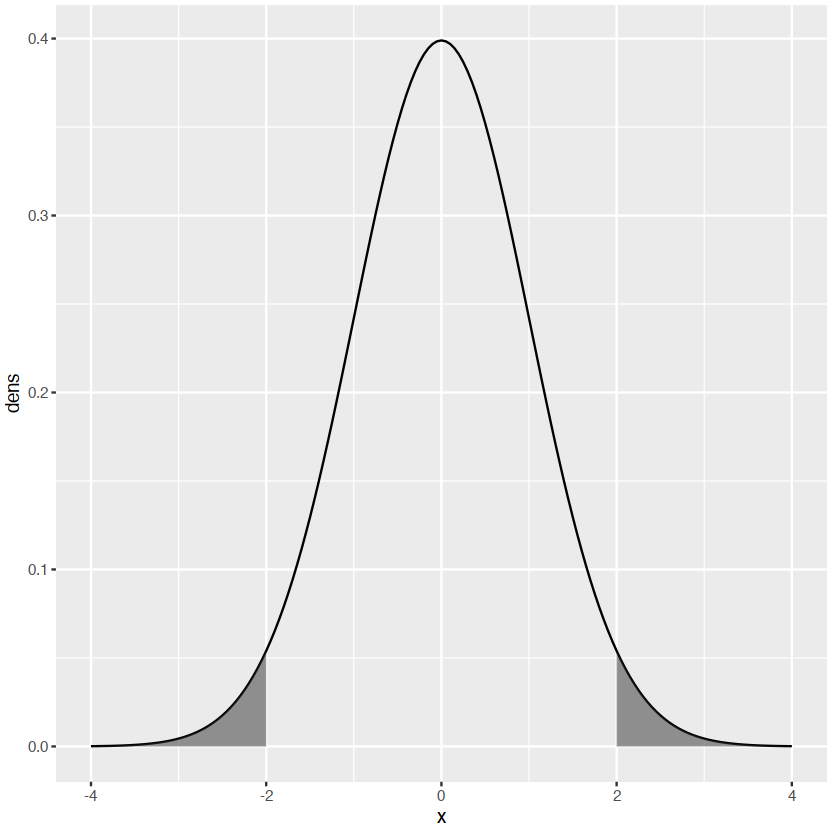
Calculating p-values in R

Suppose that we were testing whether the microfinance has an effect on consumption. The null hypothesis is that the difference in means is equal to 0, and the null sampling distribution is plotted below. If you find a z-score equal to 2, what is the p-value associated with that z-score? Is your estimate of the difference in means significant?

```
# Plot a standard normal distribution
library(tidyverse)

data <- tibble(x = seq(-4, 4, .05), dens = dnorm(x))

ggplot(data, aes(x = x, y = dens)) +
    geom_line() +
    geom_area(data = filter(data, x <= -2), alpha = .5) +
    geom_area(data = filter(data, x >= 2), alpha = .5)
```




```
# Pnorm
2 * pnorm(-abs(2))
```

0.0455002638963584

Exercise: Project STAR Part 2

From last week: The Project STAR (for Student-Teacher Achievement Ratio) was designed to determine the effect of smaller class size in the earliest grades on short-term and long-term pupil performance (source). Over 7,000 students in 79 schools across the state of Tennessee were randomly assigned into one of three interventions: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide). Classroom teachers were also randomly assigned to the classes they would teach. The interventions were initiated as the students entered school in kindergarten and continued through third grade.

In this exercise, we are going to use data from Project STAR to assess whether there is a statistically significant impact of class sizes on learn about the pupils involved in the project through visualization and measure the association between classroom size and student achievement.

Unlike last class, the data has been aggregated at the teacher level so that scores are averages across all students taught by that teacher.

Download the data using this link.

Data Dictionary

- `teacher_id`: kindergarten teacher ID
- `class_type`: kindergarten class type; S - Small, R - Regular/Large
- `reading_score`: average kindergarten reading score
- `math_score`: average kindergarten math score

1. Upload the Excel file `STAR_teachers.xlsx` to Google Colab and use `read_excel` to read its first worksheet as a new table called `star_teachers`. Examine the first 10 rows of the data.

```
library(tidyverse)
library(readxl)

# Your answer here!

# START
star_teachers <- read_excel(path = "STAR_teachers.xlsx", sheet = 1)
head(star_teachers, 10)
# END
```

A tibble: 10 × 4

2. Calculate the number of teachers and the mean and variance of reading score by class type. Which class type has a higher average reading score? Which class type has greater variance in reading score?

```
# Your answer here!

# START
star_teachers %>%
    group_by(class_type) %>%
    summarize(n = n(),
              mean = mean(reading_score),
              var = var(reading_score))
# END
```

A tibble: 2 × 4

class_type	n	mean	var
<chr>	<int>	<dbl>	<dbl>
R	196	445.8346	362.3407
S	126	451.6607	548.6109

3. Suppose $\hat{\mu}_R$ is the sample mean of the reading score in regular classes and $\hat{\mu}_S$ is the sample mean in small classes. Using your results from (2), calculate the difference in sample means and the standard error of $\hat{\mu}_S - \hat{\mu}_R$.

Recall that $SE(\hat{\mu}_S - \hat{\mu}_R) = \sqrt{\frac{\hat{\sigma}_S^2}{n_S} + \frac{\hat{\sigma}_R^2}{n_R}}$ where $\hat{\sigma}_S^2$ and $\hat{\sigma}_R^2$ are the sample variances and n_S and n_R are the sample sizes.

```
# Your answer here!

# START
diff <- 451.7 - 445.8
se <- sqrt(548.6 / 126 + 362.3 / 196)
c("Difference in Means" = diff, "Standard error" = se)
# END
```

Difference in Means: 5.89999999999998 Standard error: 2.49046936173151

4. What is the 95% confidence interval of $\mu_S - \mu_R$?

```
# Your answer here!

# START
c("Lower Bound" = diff - 2 * se,
  "Upper Bound" = diff + 2 * se)
# END
```

Lower Bound: 0.919061276536956 Upper Bound: 10.880938723463

5. What is the Z-score corresponding to the null hypothesis $\mu_S - \mu_R = 0$?

```
# Your answer here!

# START
z <- (diff - 0) / se
c("Z-score" = z)
# END
```

Z-score: 2.36903135234636

6. What is the p-value corresponding to the Z-score? Is the difference in means statistically significant?

```
# Your answer here!

# START
p <- 2 * pnorm(-abs(z))
c("p-value" = p)
c("Reject null hypothesis?" = p < .05)
# END
```

p-value: 0.0178347415382212
Reject null hypothesis?: TRUE