Lecture 14: Hypothesis Testing

Chapter 4.3

Previously: Confidence Intervals

If we know the sampling distribution of $\overline{x}$ is Normal with

- mean equal to the true unknown population mean μ
- standard error $SE = \frac{s}{\sqrt{n}}$

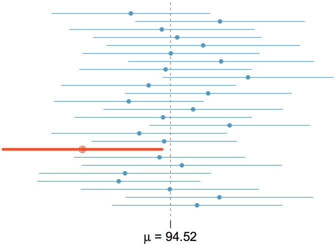
then we can use the Normal model build confidence intervals.

i.e.

$$[\overline{x} - z^* SE,\ \overline{x} + z^* SE] = \left[\overline{x} - z^* \frac{s}{\sqrt{n}},\ \overline{x} + z^* \frac{s}{\sqrt{n}}\right]$$

i.e. where z^* is the value from the normal table that sets the confidence level. Ex: $z^* = 1.96$ sets a 95% confidence interval.

Previously: How to Interpret a Confidence Interval



Goals for Today

- Introduce Hypothesis Testing Framework
- Testing Hypotheses Using Confidence Intervals
- Types of Errors
- Testing Hypotheses Using p-Values

Statistical Hypothesis Testing

(For now) A hypothesis is a claim about a population parameter.

A hypothesis test is a method for using sample data to decide between two competing hypotheses about the population parameter:

- A null hypothesis H_0.
  i.e. the status quo that is initially assumed to be true, but will be tested.
- An alternative hypothesis H_A.
  i.e. the challenger.

Examples

- We flip a coin many times and start to suspect that it is biased:
  - H_0: the coin is fair. i.e. the probability of heads is $p = 0.5$
  - H_A: the coin is not fair. i.e. $p \neq 0.5$
- From book: The average 10 mile run time for the Cherry Blossom Run in 2006 μ_{2006} was 93.29 min. Researchers suspect μ_{2012} was different:
  - H_0: the average time was the same. i.e. $\mu_{2012} = 93.29$
  - H_A: the average time was different. i.e. $\mu_{2012} \neq 93.29$

Crucial Concept: Conclusions of Hypothesis Tests

There are two potential outcomes of a hypothesis test. Either we

- reject H_0 in favor of H_A
- fail to reject H_0

Note the difference between accepting H_0 & failing to reject H_0

- "accepting H_0" is saying we are sure H_0 is true
- "failing to reject H_0" is saying something not as strong: we do not have enough evidence to reject H_0.

Analogy: US Criminal Justice System

In a recent trial, George Zimmerman was found not guilty by the jury. The jury's verdict does NOT make any statement about the defendant being innocent, rather that there was not enough evidence to prove beyond a reasonable doubt that they were guilty.

Analogy: US Criminal Justice System

Let's compare criminal trials to hypothesis tests:

Truth:

- Truth about the defendant: innocence vs guilt
- Truth about the hypothesis: H_0 or H_A

Decision:

- Verdict: not guilty vs guilty
- Test outcome: "Do not reject H_0" vs "Reject H_0"

Testing Hypotheses Using Confidence Intervals

Back to example: The average race time μ_{2006} for 2006 was 93.29 min. Researchers suspect μ_{2012} was different:

- H_0: the average time was the same. i.e. $\mu_{2012} = 93.29$
- H_A: the average time was different. i.e. $\mu_{2012} \neq 93.29$

93.29 is called the null value μ_0 (mu-naught) since it represents the value of the parameter if the null hypothesis is true.

They take a sample of size $n = 100$ times from 2012 and find that $\overline{x} = 95.61$ and $s = 15.78$

The average time $\overline{x} = 95.61$, our estimate of μ_{2012}, is greater than 93.29. Is that enough to say that the times are different?

Testing Hypotheses Using Confidence Intervals

Recall that a 95% confidence interval for the population mean μ based on a sample of points $x_1, \ldots, x_n$ is

$$\left[\overline{x} - 1.96 \times \frac{s}{\sqrt{n}} \, , \overline{x} + 1.96 \times \frac{s}{\sqrt{n}}\right] = [92.45 \, , 98.77]$$

Testing Hypotheses Using Confidence Intervals

Since the 2006 null value 93.29 falls in the range of plausible values, we cannot say the null hypothesis is implausible. i.e. we fail to reject the null hypothesis that the 2006 and 2012 times are the same.

Again, we are NOT saying that the 2006 and 2012 times are the same. Just that there is insufficient evidence to suggest otherwise.

Decision Errors

Hypothesis tests will get things right sometimes and wrong sometimes:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	OK	Type I Error
	H_A true	Type II Error	OK

Two kinds of errors:

- Type I Error: a false positive
- Type II Error: a false negative

Decision Errors

- There is a trade-off between these two error rates: procedures with lower type I error rates typically have higher type II error rates and vice versa
- In other words, there is almost never a perfect test that makes no type I errors while making no type II errors
- Some sort of balance between the two is required

Example: US Criminal Justice System

Defendants must be proven "guilty beyond a reasonable doubt" i.e. in theory they would rather let a guilty person go free, than put an innocent person in jail.

So let:

- H_0: the defendant is innocent
- H_A: the defendant is guilty

thus rejecting H_0 corresponds to a guilty verdict. i.e. putting them in jail

In this case:

- A type I error is putting an innocent person in jail (considered worse)
- A type II error is letting a guilty person go free.

Example: Airport Screening

An example of where type II error is much more serious: airport screening.

Let:

$$\begin{aligned} H_0 &: \quad \text{passenger X does not have a bomb/weapon} \\ H_A &: \quad \text{passenger X has a bomb/weapon} \end{aligned}$$

Failing to reject H_0 when H_0 is false corresponds to not "patting down" passenger X when they really have a bomb/weapon. This is disastrous!

Hence the long lines at airport security.

Significance Level

Hypothesis testing is built around rejecting or failing to reject the null hypothesis
i.e. we do not reject H_0 unless we have strong evidence.

As a general rule of thumb, for those cases when H_0 is true, we do not want to incorrectly reject H_0 more than 5% of the time. In this case $\alpha = 0.05 = 5\%$ is the significance level.

Using the procedure to create 95% confidence intervals earlier, we expect it to miss the true population parameter 5% of the time. This corresponds to $\alpha = 0.05$.

p-Values

Thought experiment: Say you flip a coin you think is fair 1000 times. Thus you expect 500 heads. Now say you observe

- 501 heads? Do you think the coin is biased?
- 525 heads? Do you think the coin is biased?
- 900 heads? Do you think the coin is biased?

Intuitively, a p-value quantifies how extreme an observation is given the null hypothesis.

The smaller the p-value, the more extreme the observation, where the meaning of extreme depends on the context.

p-Value Definition

The p-value or observed significance level is the probability of observing a test statistic as extreme or more extreme (in favor of the alternative) as the one observed, assuming H_0 is true.

It is NOT the probability of H_0 being true. This is the most common misinterpretation of the p-value.

Example: Exercise 4.28 on Page 177 on Sleep

A poll found that college students sleep about 7 hours a night. Researchers suspect that Reedies sleep more. They use a sample of $n = 110$ Reedies to investigate this claim at an $\alpha = 0.05$ level.

Let μ be the true # of hours Reedies sleep a night:

- $H_0 : \mu = \mu_0 = 7$
- $H_A : \mu > 7$

Example: Exercise 4.28 on Page 177 on Sleep

Researchers find that $\overline{x} = 7.42$ and $s = 1.75$. Before we proceed, we check the 3 conditions

1. Independence: the sample size $n = 110$ is less than 10% of 1,453 (Reed enrollment)
2. The sample size is greater than 30
3. The distribution of the $n = 110$ observations (Figure 4.14) is not too skewed.

Example: Exercise 4.28 on Page 177 on Sleep

Question to keep in mind: What if the null hypothesis were true (i.e. the null value $\mu_0 = 7$)?

How likely are we to observe $\overline{x} = 7.42$ or something more extreme in favor of the alternative, i.e. greater? Using the z-score of $\overline{x}$ and plots.

Remember in general

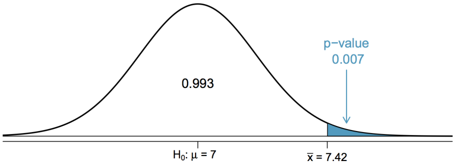
$$z = \frac{x - \mu}{\sigma}$$

So in our case for $\overline{x}$

$$z = \frac{\overline{x} - \mu \text{ when } H_0 \text{ is true}}{SE} = \frac{\overline{x} - \mu_0}{SE} = \frac{7.42 - 7}{\frac{1.75}{\sqrt{110}}} = 2.47$$

Example: Exercise 4.28 on Page 177 on Sleep

If the null hypothesis were true, then $\overline{x}$ would have come from the following nearly normal distribution. The p-Value is 0.007 since:



Example: Exercise 4.28 on Page 177 on Sleep

Correct interpretation: If the null hypothesis is true, the probability of observing a sample mean $\overline{x} = 7.42$ or greater from a sample of size $n = 110$ is only 0.007.

The p-value quantifies how strongly the data favor H_A over H_0. A small p-value corresponds to sufficient evidence to reject H_0 in favor of H_A.

Final decision. Since we set $\alpha = 0.05$ beforehand and the p-value $0.007 < 0.05 = \alpha$, we reject the null hypothesis. i.e. based on evidence, we believe Reedies sleep more than 7 hours a night.

Next Time

- More Hypothesis Testing
- Central Limit Theorem