Reminders

• Last HW and Last quiz on Thursday
• My office hours will be from 11-1 on Wed. this week
• If you won’t be around during the final week to take the Final Project, please email me ASAP to arrange for a time for you to take it.
I would like to estimate the true proportion of all UW students who are in fraternities and sororities. I take a simple random sample of 100 UW students, and learn that 12 of them are in a fraternity or sorority. **What is the mean of the sampling distribution for the statistic I calculate from my sample?**
Chapter 22, Part 1: What is a Test of Significance?

Aaron Zimmerman
STAT 220 - Summer 2014

Department of Statistics
University of Washington - Seattle
Motivating Example (Cobb and Gehlbach 2004)

- Kristen Gilbert was a nurse at the VA hospital in Northamption, MA from March 1990-February 1996
- Though she was regarded as a skilled and competent nurse, her colleagues noticed a suspicious pattern of deaths by cardiac arrest during the shifts that she worked at the hospital
- Her grand jury trial has become a famous example of a test of significance
Motivating Example

• At first, colleagues attributed the deaths during Kristen Gilbert’s shifts to random chance.
• But then they noticed that supplies of epinephrine (which can prompt cardiac arrest) were disappearing from the storage area.
• One of her colleagues then reported that she was offered a bottle of epinephrine by Nurse Gilbert, leaving little doubt about who was stealing the supplies.
• However, nobody ever saw Kristen Gilbert give a patient a lethal injection, and nobody could give a good motive for why she would want to kill her patients.
Motivating Example

• Nonetheless, Assistant US Attorney William Welch convened a grand jury in 1998 to hear the evidence against Gilbert.

• The purpose of a grand jury trial is to decide whether there’s enough evidence against the defendant to bring him or her to trial.

• As you all know, defendants are innocent until proven guilty, so the job of the prosecutor at a grand jury trial is to demonstrate that there is enough evidence against the defendant to justify bringing her to trial.

• Pretend that you are the prosecutor at Kristen Gilbert’s grand jury trial. What data would you collect to argue that there’s enough evidence against her to justify bringing her to trial?
Motivating Example

Display 1: The pattern of deaths, by year and by shift.

In each set of three bars, left = night (midnight - 8 am), middle = day (8am - 4 pm), right = evening (4pm - midnight).
Motivating Example

• Clearly, there is a suspicious pattern of deaths during the shifts that Kristen Gilbert worked at the VA.
• However, her defense team had an obvious response: these extra deaths during her shifts could just be due to random chance!
• Again, imagine that you are the prosecutor in this grand jury trial. Can you think of a way to argue that the extra deaths during Kristen Gilbert’s shifts couldn’t just be due to random chance, and are actually very suspicious?
Motivating Example

• Prosecutors went back through one year of data and calculated the proportion of shifts that Kristen Gilbert worked that had a death in her ward.

• She worked 257 days that year, and a patient died during her shift on 40 of those days. So, the sample proportion for Kristen Gilbert’s shifts is:

\[ \hat{p} = \frac{40}{257} = .156 = 15.6\% \]

• Key question: is this proportion of deaths “abnormal”? Or are the higher proportions of deaths reasonably possible by random chance.
Motivating Example

- Prosecutors then went through all the data from the shifts that Kristen Gilbert did not work that year.
- There were 1384 other shifts that year, and someone died during 34 of those shifts.
- In other words, the proportion of deaths in other shifts is:

  \[ p = \frac{34}{1384} = .025 = 2.5\% \]

- Remember, a patient died during 15.6\% of Kristen Gilbert’s shifts at the hospital. How likely is it that this could have happened by random chance, given that a patient died during only 2.5\% of the other shifts?
Motivating Example

- The **key** to answering this question is to **think about** the sampling distribution of the sample proportion.

- Suppose there was nothing suspicious going on during Kristen Gilbert’s shifts. Then we’d expect the sample proportion to come from a sampling distribution with mean $p = .025$ and standard error:

$$
\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.025(.975)}{257}} = .0097
$$
Motivating Example

- Where is Kristen Gilbert’s sample proportion in this sampling distribution?

- Common sense conclusion: if there was really only a 2.5% chance of a patient dying during one of Kristen Gilbert’s shifts, then it’s EXTREMELY unlikely that a patient would have died during 15.6% of Kristen Gilbert’s shifts!
Motivating Example

- As statisticians, we can go one step farther: exactly how likely would it be to see a result this “extreme” if the true probability of a patient dying was .025?
- The prosecutors actually calculated this number: it turns out that there’s less than 1 in 100,000,000 chance that we would get a sample proportion this extreme if there was nothing suspicious going on.
Tests of significance

- The evidence presented at Kristen Gilbert’s grand jury trial is an example of a test of significance.

- We are going to go back through this example and add some terminology to the concepts we discussed.

- Don’t let all the terminology throw you off! If you understood the basic logic that the prosecutors used in the trial, then you already understand tests of significance!
Tests of significance

• A test of significance always starts with a “claim” that needs to be tested.
  ★ In a jury trial, the “claim” is that the defendant is innocent

• The claim being tested in a statistical test is called the **null hypothesis** (H₀).

• The test is designed to test the strength of the evidence against the null hypothesis.

• In Kristen Gilbert’s grand jury trial, we can write the null hypothesis in two ways.
  ★ *(in words)* H₀: The probability of a patient dying during one of Kristen Gilbert’s is the same as in all the other shifts at the hospital
  ★ *(mathematically)* H₀: p = .025
Tests of significance

- The statement we hope or suspect is true instead of $H_0$ is called the **alternative hypothesis** ($H_a$ or $H_1$).
- A **one-sided alternative** means that we’re seeking evidence that the truth is greater than OR less than some value.
  - This is the case in Kristen Gilbert’s trial: we suspect that the true probability that a patient dies is greater than .025.
  - $H_a : p > .025$
- A **two-sided alternative** means that we’re seeking evidence that the truth is different (either greater than or less than) some value.
  - If we were just checking to see if the true probability of a patient dying is different than in all the other shifts, the alternative hypothesis would be $H_a : p \neq .025$
Tests of significance

Here are the null and alternative hypotheses in Kristen Gilbert's grand jury trial:

- $H_0 : p = 0.025$
- $H_a : p > 0.025$

Key idea #1: The null hypothesis expresses the idea that an observed difference is due to chance. The alternative hypothesis claims that the difference is real. Both hypotheses are claims about parameters.
Tests of significance

• Again, here are the null and alternative hypotheses in Kristen Gilberts grand jury trial:
  \[ H_0 : p = 0.025 \]
  \[ H_a : p > 0.025 \]

• The prosecution’s job was to show that there is good reason to doubt the null hypothesis, so they calculate a p-value

• The **p-value (P)** of the test is the probability of observing an outcome as or more extreme than what we actually observed if the null hypothesis were true.

  * The smaller the p-value is, the stronger the evidence against \( H_0 \) provided by the data.

  * The prosecutions p-value was 1/100,000,000. In other words, there’s a .000000001 chance that we would see a result this extreme if the null hypothesis was true.
Tests of significance

• In practice, researchers set a “cutoff point” for the p-value that they believe would give sufficient evidence to “reject the null hypothesis”

• This “cutoff point” for P is called the **significance level** ($\alpha$) (alpha).

  ★ If we choose $\alpha = .05$, for example, we are requiring that the data give evidence against $H_0$ so strong that it would happen no more than 5% of the time when $H_0$ is true

  ★ If we choose $\alpha = .01$, we are insisting that the data give evidence against $H_0$ so strong that it would happen no more than 1% of the time when $H_0$ is true

• If the p-value is as small or smaller than $\alpha$, we say that the data are **statistically significant at level** $\alpha$

• At any reasonable significance level, the data in Kristen Gilbert’s trial are statistically significant
Tests of significance

- There are only two possible conclusions to a test of significance!

- If the p-value is statistically significant, then you can say, “I have enough evidence to reject the null hypothesis.”

- If the p-value is not statistically significant, then you can say, “I do not have enough evidence to reject the null hypothesis.”

- You can never, ever say that you have proven the null hypothesis. You either have enough evidence to reject it, or you don’t have enough evidence to reject it.
Tests of significance

- The most common significance level to choose is .05
- Other frequently chosen significance levels are 0.1, and 0.01

Key idea #2: The p-value (P) of a test is the chance of getting a big test statistic – assuming the null hypothesis to be right. P is not the chance of the null hypothesis being right.
Charles Tart ran an experiment at the Univ. of California, Davis to test for ESP in humans.

He used a machine called “Aquarius”. It’s a random number generator that generates one of four target numbers but it doesn’t tell anyone what it’s picked. A person can then push one of four buttons (one for each target) indicating which target they think Aquarius has chosen. Aquarius then lights up the correct answer and rings a bell if the subject guessed correctly.

Tart selected 15 subjects who claimed they were clairvoyant.
ESP Example

- Each of the 15 subjects made 500 guesses on Aquarius (total of 7500 guesses)
- Out of all of the trials, 2006 correct guesses were made.
- If monkeys were guessing, we would anticipate that they would get about $\frac{1}{4} \times 7500 = 1875$ correct guesses.

Question: Is the surplus of correct answers (2006 - 1875 = 131 surplus correct answers) evidence to suggest that these self-proclaimed clairvoyants actually exhibit ESP?
Framework of a test of significance

• Before I start, I’m going to set the significance level to $\alpha = .05$
  ★ In other words, I’m requiring that the data give evidence against the standard claim of NO ESP ($H_0$) that’s so strong that it would happen no more than 5% of the time when $H_0$ is true

• Next, I determine my null and alternative hypotheses
  ★ $H_0 : p = 0.25$ (clairvoyants are as good as monkeys at guessing)
  ★ $H_a : p > 0.25$ (clairvoyants are better than monkeys - possibly because they have ESP)
Finally, I take the sample proportion $\hat{p} = \frac{2006}{7500} = 0.267$ and calculate how likely it is that the clairvoyants would get a result this extreme if they really don’t have ESP (the p-value).

I calculate a p-value of 0.0002. In other words, there is a 2 in 10,000 chance that the proportion of correct guesses would be this extreme (this large) if they had a 1/4 chance of guessing correctly.
Framework of a test of significance

- Since my significance level is $\alpha = .05$, and my p-value is smaller ($P=.0002$), this test is statistically significant. So, we reject the null hypothesis that the clairvoyants are as good as monkeys at guessing in favor for the alternative that they’re better than monkeys.

- Remember: this doesn’t mean their claims are true! It just means that we have enough evidence to show that this sort of result would only happen very very very rarely if the null was true.
Framework of a test of significance

- So, with such a small p-value (.0002), it’s hard to explain away the surplus correct guesses as chance variation
- BUT, this doesn’t prove that ESP exists
- It could be that Aquarius wasn’t a very good random number generator
- Or the machine could have been giving subtle clues as to which target it had picked

Key idea #3: A test of significance can only tell you that a difference is there. It cannot tell you the cause of the difference.
Practice

• A drug company develops an AIDS treatment that they hope will reduce the proportion of AIDS patients who die within 50 years. In a randomized control trial, 35% of patients in the control group died within 5 years. The drug company would like to show that the proportion of patients who die within 5 years in the treatment group is less than this.

• What is the null hypothesis for this experiment?
• What is the alternative hypothesis for this experiment?
A drug company develops an AIDS treatment that they hope will reduce the proportion of AIDS patients who die within 50 years. In a randomized control trial, 35% of patients in the control group died within 5 years. The drug company would like to show that the proportion of patients who die within 5 years in the treatment group is less than this. **What is the null hypothesis for this experiment? What is the alternative hypothesis for this experiment?**

- \( H_0 : p = 0.35 \)
- \( H_a : p < 0.35 \)
Practice

• It turns out that 28% of the patients in the treatment group died within 5 years. The drug company calculates that the p-value for the experiment is .014. What does this p-value mean?

• Before the trial, the drug company set the significance level of the test at $\alpha = 1\% = .01$. What is the conclusion of this experiment?
Practice

• It turns out that 28% of the patients in the treatment group died within 5 years. The drug company calculates that the p-value for the experiment is .014. **What does this p-value mean?**
• There is a .014 chance (14 in 1000 chance) that we would observe results as extreme (as small) as we did if the null hypothesis was true.
• **Before the trial, the drug company set the significance level of the test at $\alpha = 1\% = .01$. What is the conclusion of this experiment?**
• Since the p-value is **larger** than the significance level, we **fail to reject the null hypothesis** and conclude that the **differences we observe could be due to random chance** alone. So, we don’t have enough evidence to suggest that the treatment group has a statistically lower percent of people dying within 5 years.
Homework

• The final HW is up on the website
• After today you can:
  • Read pages 481-489
  • Do problems 22.1, 22.8, 22.11, 22.12, 22.13, 22.14, 22.15, 22.16, 22.17, 22.18