CHAPTERS 4-6: Hypothesis Tests

Read sections 4.3, 4.5, 5.1.5, 6.1.3

Confidence Interval vs. Hypothesis Test (4.3):

- The purpose of a confidence interval is to estimate the value of a parameter.
- The purpose of a Hypothesis Test is to decide whether or not a claim (or hypothesis) about the value of a parameter is plausible.

Hypotheses are statements about parameters not statistics! (4.3.1)

- The **Null Hypothesis**, H_0, is a statement of equality signifying no difference or no effect.
- The **Alternative Hypothesis**, H_a is a statement of inequality which you are trying to find evidence for.

Hypothesis Testing for the population mean μ of a quantitative variable

1. Hypotheses: Choose from one of the following sets of hypotheses where μ_0 is the hypothesized value of μ:

	One-sided test		One-sided test		Two-sided test
1.	$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0$	2.	$H_0 : \mu = \mu_0$ $H_a : \mu < \mu_0$	3.	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$

   NOTE: Perform Steps 2-4 assuming that H_0 is true and determine if the data suggests that H_0 is not true.

2. Assumptions (4.5.1):
   (a) The data must be a SRS
   (b) If you sample a finite population without replacement, then $.05N \geq n$.
   (c) The sampling distribution of $\overline{X}$ must be at least approximately normal (so the data is normal OR $n \geq 15$ for symmetric non-normal data OR $n \geq 30$ for skewed data). If H_0 is true, then $\overline{X} \dot{\sim} N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$.

3. Test Statistic (4.5.2 and 5.1.5):
   - If σ is known, the test statistic is $Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ where $Z \sim N(0,1)$ when H_0 is true.
   - If σ is not known, the test statistic is $T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$ where $T \sim t(n-1)$ if H_0 when true.

4. The p-value is the probability of obtaining a sample statistic that is as extreme or more extreme than what was actually observed assuming that H_0 is true (4.3.4 and 4.3.5).

   - A small p-value indicates that the statistic you observed would rarely occur if H_0 were true. Thus, a small p-value is evidence that H_0 is not true.
   - The smaller the p-value, the stronger the evidence against H_0 and in favor of H_a.
   - Computing the p-value depends on H_a:

Alternative Hypothesis	p-value if σ is known	p-value if σ is unknown
$H_a : \mu > \mu_0$	$P(Z > z)$	$P(T > t)$
$H_a : \mu < \mu_0$	$P(Z < z)$	$P(T < t)$
$H_a : \mu \neq \mu_0$	$2P(Z > \lvert z \rvert)$	$2P(T > \lvert t \rvert)$

   - Observe that the p-value can be written in terms of $\overline{X}$. For example, for upper tailed tests,

$$p\text{-value} = P(\overline{X} > \bar{x}) = P(Z > z).$$

5. Decision (4.3.6): Given a predetermined *significance level* α:

   - **Reject** H_0 if p-value $\leq \alpha$. In this case we say that the results are *statistically significant*.
   - **Fail to Reject** H_0 if p-value $> \alpha$. In this case we say that the results are not statistically significant.

   Commonly-used significance levels are .01, .05 and .1. The significance level α is the probability cut-off value for how *rare* you are requiring the sample result to be, if H_0 were actually true.

6. Conclusion: a statement of how much evidence there is for the alternative hypothesis.

   - If you reject H_0, then you conclude "The evidence suggests H_a".
   - If you fail to reject H_0, then you conclude "The evidence fails to suggest H_a".

   In each case, specify H_a in terms of the problem.

EXAMPLE:
A soil scientist is interested in studying the pH level in the soil for a certain field. Is there evidence that the mean pH level is acidic (that is, less than 7)? To test this hypothesis, she plans on using a significance level of $\alpha = 5\%$. She examines a random sample of 50 soil samples and measures their pH levels. She finds that the sample mean pH level is $\bar{x}$=6.14 and that the sample standard deviation is $s = 0.48$.

1. Hypotheses:

2. Check assumptions:

3. Test statistic:

4. p-value:

5. Decision:

6. Conclusion:

Statistical Significance:

Important points to remember:

- Specify α, the level of significance of the test **before looking at the data!** If you set the significance level after looking at the data, then you are **data mining**, and your results are not legitimate.
- A hypothesis test answers the question: If H_0 is true, how likely is it that we observe the statistics that we get? The p-value quantifies this so we can reject our assumption that H_0 is true or not.
- Not finding enough evidence to reject H_0 **does not imply** that H_0 is true! That is: "The absence of evidence is not the evidence of absence!" - anon stat nerd.
- Recall the Scientific Method from the Chapter 1 notes:

  1. Observe some phenomenon
  2. State a hypothesis explaining the phenomenon
  3. Collect data
  4. Test: Do the data support the hypothesis?
  5. Conclusion. If the test fails, go back to step 2.

  Steps 1 and 2 are driven by human inquisitiveness and intuition. Step 3 should be carried out according to the Experimental and/or Sampling Designs outlined in Chapter 2. Confidence Intervals and Hypothesis Tests allow us to perform Steps 4 and 5.

- **The Six Steps in Hypothesis Testing** can be inserted into steps 4 and 5 of the Scientific method. So we can renumber the six steps in hypothesis testing to emphasize this relationship:

  4. Test: Do the data support the hypothesis which explains the phenomenon?
     - 4.1 State H_0 and H_a with respect to the parameter of interest.
     - 4.2 Check the assumptions necessary so that the test is valid.
     - 4.3 Compute the test statistic.
     - 4.4 Compute the p-value.
     - 4.5 Make a decision about H_0.
  5. Draw a conclusion.

- Statistical significance is different than practical significance! (4.5.5)

Hypothesis Testing for a population proportion p that describes a categorical variable

1. Hypotheses: Choose from one of the following sets of hypotheses where p_0 is the hypothesized value for p.

	One-sided test		One-sided Test		Two-sided test
1.	$H_0 : p = p_0$ $H_a : p > p_0$	2.	$H_0 : p = p_0$ $H_a : p < p_0$	3.	$H_0 : p = p_0$ $H_a : p \neq p_0$

   NOTE: Perform Steps 2-4 assuming that H_0 is true and determine if the data suggests that H_0 is not true.

2. Assumptions:
   (a) The data must be a SRS
   (b) If you sample a finite population without replacement, then $.05N \geq n$.
   (c) The sampling distribution for $\hat{p}$ must be approximately normal (so $np_0 \geq 10$ and $n(1-p_0) \geq 10$). Thus, if H_0 is true, $\hat{p} \dot{\sim} N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$.

3. Test Statistic: The test statistic is $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ where $Z \dot{\sim} N(0,1)$ if H_0 is true.

4. p-value:

Alternative Hypothesis	p-value
$H_a : p > p_0$	$P(Z > z)$
$H_a : p < p_0$	$P(Z < z)$
$H_a : p \neq p_0$	$2P(Z > \lvert z \rvert)$

5. and 6. Make a Decision and give a Conclusion

EXAMPLE: When tossing a coin and recording heads or tails for each toss, is the coin is unfair (i.e. not 50% / 50%). A nerdy coin tosser observes 4,933 heads out of 10,000 tosses.

1. Hypotheses:

2. Check assumptions:

3. Test statistic:

4. p-value:

5. Decision:

6. Conclusion:

Errors in Hypothesis Testing (4.3.3):

- A Type I Error is rejecting H_0 when H_0 is actually true (rejecting H_0 when you shouldn't). So you are saying there is a difference in the world (between μ and μ_0) that really does not exist.

$$P(\text{Type I Error}) = P(\text{Rejecting } H_0 | H_0 \text{ is true}) = \alpha$$

- A Type II Error is failing to reject H_0 when H_a is actually true (failing to reject H_0 when you should). So you are failing to detect a difference between μ and μ_0 which really does exist.

$$P(\text{Type II Error}) = P(\text{Failing to Reject } H_0 | H_a \text{ is true}) = \beta$$

- Power is the probability of rejecting H_0 when you should. So this is the probability of detecting the true difference between μ and μ_0.

$$\text{Power} = P(\text{rejecting } H_0 | H_a \text{ is true}) = 1 - \beta$$

TRUTH TABLE

	H_0 is true	H_a is true
Reject H_0		
Fail to Reject H_0		

The Trade-offs:

1. Increasing the significance level α decreases β and increases Power=$1-\beta$.
2. Increasing n increases Power.
3. Decreasing σ increases Power.
4. Greater distances between μ and μ_0 increases Power.

NOTE: The value a researcher chooses for α depends on the seriousness of the Type I Error relative to the Type II Error. For example, if the Type I Error is more serious than a Type II Error, then the researcher would prefer a smaller α (to minimize the chance of making a Type I Error).

Statistical Significance Does Not Imply Practical Importance!
A statistically significant result (i.e. you rejected H_0 and concluded H_a) may NOT be of any practical importance.

EXAMPLE:
Is the mean ketchup content in 16-ounce Ketchup Bottles less than what the label says?

$$H_0: \ \mu = 16$$
$$H_a: \ \mu \neq 16$$

From a SRS of 144 16oz bottles, the sample mean content is $\bar{x}$=16.01, and $s = 0.03$.

The test statistic is $t = \frac{16.01 - 16}{0.03/\sqrt{144}} = 4$, the p-value < 0.0005, and so we Reject H_0 in favor of H_a.

The evidence suggests that the ketchup packer is giving consumers too much ketchup (on average)! The shock, the sham, the injustice in the world!

NOTE: A very small difference can be declared "statistically significant" simply due to a large sample size.

EXAMPLE:
Is one mite medication better than another to to eradicate mite infestations in bee hives?

H_0: The two treatments have the same mean mite survival rate

H_a: The two treatments have different mean mite survival rates

From a SRS of 4 beehives treated with *Mites are Us*, 80% of the mites survived. From a SRS of 5 beehives treated with *Ain't no Mites no More*, 50% survived.

We get a large p-value (we'll see how to conduct this test on two samples in Chapter 11) and so we Fail to Reject H_0

NOTE: A large difference can be declared "not statistically significant" simply because the sample size is too small.

BE SKEPTICAL!

- If an article claims that a result is "statistically significant", you should ask the questions: "What was the difference you detected as significant?"; "Is the difference of any practical importance?"; "Is the set of analyses appropriate? large enough sample? random sample?"
- If an article claims their result was "not statistically significant", they may try to imply that means there is no real difference, which may not be true! It could be they simply had too small of a sample size to detect the difference that truly exists.

Using Confidence Intervals to Conduct Hypothesis Tests (4.3.2)

You can use a $100C\%$ confidence interval where $C = 1 - \alpha$ to conduct two-sided hypothesis tests at the α significance level.

To make a **decision** regarding H_0

- Reject H_0 if the hypothesized value under H_0 is NOT in the CI.
- Fail to Reject H_0 if the hypothesized value in H_0 is in the CI.

EXAMPLE revisited: When tossing a coin and recording heads or tails for each toss, is the coin is unfair (i.e. not 50% / 50%). A nerdy coin tosser observes 4,933 heads out of 10,000 tosses.

1. Hypotheses:

2. Check assumptions:

3. and 4. Construct a CI corresponding to a significance level of $\alpha = .05$ (instead of calculating a test statistic and p-value).

5. Decision:

6. Conclusion:

R code

```
>#  pH Example, test for mu
> t.test(pH,mu=7,conf.level=.95,alternative="less")

        One Sample t-test

data:  pH
t = -12.669, df = 49, p-value < 2.2e-17
alternative hypothesis: true mean is less than 7
95 percent confidence interval:
     -Inf 6.258825
sample estimates:
mean of x
 6.14


> #coin toss Example, test for pi
> prop.test(4933,10000,p=.5,conf.level=.99,alternative="two.sided",correct=F)

        1-sample proportions test without continuity correction

data:  4933 out of 10000, null probability 0.5
X-squared = 1.7956, df = 1, p-value = 0.1802
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.4804307 0.5061782
sample estimates:
     p
0.4933
```

Exercises

Testing means on p209: 4.17, 4.19, 4.23 - 4.27 odd; p257: 5.3, 5.7, 5.9
Testing Proportions on p.315: 6.15, 6.17
Testing with CIs on p210: 4.21; p259: 5.11
Testing errors on p212: 4.29, 4.31, 4.32 (correct answers are TTFTT), 4.47.