Name _____

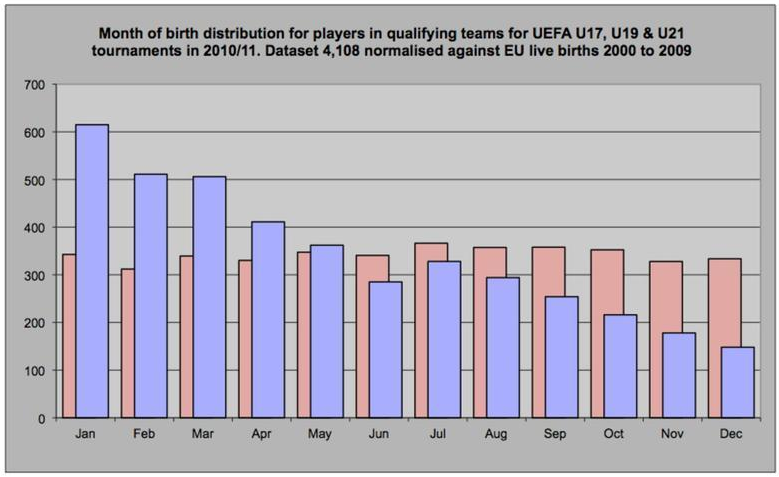
Chapter 9 Learning Objectives	**Section**	**Related Example on Page(s)**	**Relevant Chapter Review Exercise(s)**	**Can I do this?**
State the null and alternative hypotheses for a significance test about a population parameter.	9.1	540	R9.1	
Interpret a *P*-value in context.	9.1	543, 544	R9.5	
Determine if the results of a study are statistically significant and draw an appropriate conclusion using a significance level.	9.1	546	R9.5	
Interpret a Type I and a Type II error in context, and give a consequence of each.	9.1	548	R9.3, R9.4	
State and check the Random, 10%, and Large Counts conditions for performing a significance test about a population proportion.	9.2	555	R9.4	
Perform a significance test about a population proportion.	9.2	559, 562	R9.4	
Interpret the power of a test and describe what factors affect the power of a test.	9.2	565, discussion on 568	R9.3	
Describe the relationship among the probability of a Type I error (significance level), the probability of a Type II error, and the power of a test.	9.2	565	R9.3	
State and check the Random, 10%, and Normal/Large Sample conditions for performing a significance test about a population mean.	9.3	575	R9.2, R9.6, R9.7	
Perform a significance test about a population mean.	9.3	580, 583	R9.6	
Use a confidence interval to draw a conclusion for a two-sided test about a population parameter.	9.2, 9.3	563, 585	R9.5, R9.6	
Perform a significance test about a mean difference using paired data.	9.3	586	R9.7	

9.1 Significance Tests

Read 537

Suppose we take the temperature of a random sample of 200 healthy adults and their $\bar{x}$ = 98.35ºF. What are the two explanations for why their sample mean is not 98.6 ºF?

Suppose we toss a coin 200 times and the proportion of heads is $\hat{p}$ = 58% instead of 50%.



The "relative age effect" suggests that people born in certain months (or other time periods) are under-represented in some groups. The graph shows the distribution of birth month for some European youth soccer tournament teams in 2010/11 (front bars) against the distribution of birth month for the general population from 2000-2009 (rear bars). This seems to indicate that players who are a few months older than their peers tend to make up a larger-than-expected proportion of those teams and that the players who were born later in the year tend to make up less of the teams than we might expect. These European leagues use January 1 as the cut-off date when determining a player's eligibility to compete.

Suppose that in this sample of 4108 players, only 544 (13%) were born in October, November, and December. Is this *convincing* evidence that the true proportion *p* of *all* European youth soccer players born in October, November, and December is smaller than 3/12 (or 25%)?

Give two explanations for why the sample proportion was below 3/12 (or 25%).

Note that we could also estimate this probability by using the sampling distribution of $\hat{p}$ and that we could answer this question with a confidence interval for p.

Read 539–541

What is the difference between a null and an alternative hypothesis? What notation is used for each?

DEFINITION: Null hypothesis H_0, alternative hypothesis H_a

The claim we weigh evidence against in a statistical test is called the **null hypothesis (H_0)**. Often the null hypothesis is a statement of "no difference."

The claim about the population that we are trying to find evidence *for* is the **alternative hypothesis (H_a)**.

Common mistake when stating hypotheses:

Writing words or symbols that refer to samples or statistics.

For each of the following scenarios, define the parameter of interest and state appropriate hypotheses.

(a) The soccer player data from the previous page.

(b) Tim is an engineer who is responsible for quality control in the manufacture of certain parts of fighter jets. Tim knows that the mean diameter of a certain rivet hole is supposed to be $\mu = 0.250$ inches with a standard deviation of $\sigma = 0.003$ inches. He is hoping that a newly developed drill bit will cut these rivet holes so that the diameter is more consistent (less variable).

What is the difference between and one-sided and a two-sided alternative hypothesis? How can you decide which to use?

DEFINITION: One-sided alternative hypothesis and two-sided alternative hypothesis

The alternative hypothesis is **one-sided** if it states that a parameter is *larger than* the null hypothesis value or if it states that the parameter is *smaller than* the null value. It is **two-sided** if it states that the parameter is *different from* the null hypothesis value (it could be either larger or smaller).

HW #12: page 551 (2–10 even)

9.1 *P*-values and Conclusions

Read 541–544

What is a *P*-value?

DEFINITION: *P*-value

The probability, computed assuming H_0 is true, that the statistic (such as $\hat{p}$ or $\bar{x}$) would take a value as extreme as or more extreme than the one actually observed, in the direction specified by H_a, is called the ***P*-value** of the test.

In the youth soccer example, the *P*-value = $P(\hat{p} \leq 0.13 | p = 0.25) \approx 0$. Interpret this value.

Alternate Example: *A better golf club?*

When Tim was testing a new drill bit, the hypotheses were $H_0: \sigma = 0.003$ versus $H_a: \sigma < 0.003$ where σ = the true standard deviation of the diameters of the rivet holes made with the new drill bit. Based on a sample of holes made with the new drill bit, the standard deviation was $s_x = 0.002$ inch.

(a) What are the two explanations for why $s < 0.003$?

(b) A significance test using Tim's sample data produced a *P*-value of 0.97. Interpret the *P*-value in this context.

Read 544–547

The two possible conclusions for a significance test:

Reject H_0 and conclude that we have significant (or convincing) evidence that the H_a is true.	Fail to reject H_0 and conclude that we do not have significant (or convincing) evidence that the H_a is true. It's possible this may be written using a double negative: Fail to reject H_0 and conclude that we do not have significant (or convincing) evidence that the H_0 is not true.

Common errors that students make in their conclusions:

"Accepting" H_0.

Making statements about samples or statistics (past tense often implies a reference to samples).

When are the results of a study statistically significant?

DEFINITION: Statistically significant at level α

If the *P*-value is smaller than alpha, we say that the results of a study are **statistically significant at level α**. In that case, we reject the null hypothesis H_0 and conclude that there is convincing evidence in favor of the alternative hypothesis H_a.

That Greek lower-case letter alpha, α, is called the **<u>significance level</u>**. α is chosen before conducting a hypothesis test (or significance test).

A student decided to investigate whether students at his school prefer the taste of a certain name-brand bottled water to a certain store brand bottled water. After collecting data, the student performed a significance test using the hypotheses $H_0: p = 0.5$ versus $H_a: p > 0.5$ where p = the true proportion of students at the school who prefer the name-brand water. The resulting P-value was 0.067. What conclusion would you make at each of the following significance levels?

(a) $\alpha = 0.10$

(b) $\alpha = 0.05$

What should be considered when choosing a significance level? *See page 547 in book.*

- *How plausible is* H_0? If H_0 represents an assumption that the people you must convince have believed for years, strong evidence (a very small P-value) will be needed to persuade them.
- *What are the consequences of rejecting* H_0? If rejecting H_0 in favor of H_a means making an expensive change of some kind, you need strong evidence that the change will be beneficial.

HW #14: page 551 (1–17 odd)

9.1 Errors / 9.2 Significance Tests for a Population Proportion

Read 547–550

In a jury trial, what two errors could a jury make?

In a significance test, what two errors can we make?

DEFINITION: Type I error and Type II error

If we reject H_0 when H_0 is true, we have committed a **Type I error**.

If we fail to reject H_0 when H_a is true, we have committed a **Type II error**.

Which error is worse? It depends; we can't say one is always worse.

	H_0 is really true	H_0 is really false
Decide to reject H_0	Type I error "false alarm" prob is α	Correct decision prob is power
Decide to fail to reject H_0	Correct decision	prob is β Type II error "a miss"

P(correctly rejecting H_0)+P(Type II)=1
power + β =1

Describe a Type I and a Type II error in the context of the youth soccer example.

If there has been an error made, which one could it be? Explain.

What is the probability of a Type I error?

The significance level α of any fixed-level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is actually true. Consider the consequences of a Type I error before choosing a significance level.

What can we do to reduce the probability of a Type I error? Are there any drawbacks to this?

Read 554–557

What are the three conditions for conducting a significance test for a population proportion?

- **Random:** The data come from a well-designed random sample or randomized experiment.
  - **10%:** When sampling without replacement, check that $n \leq \frac{1}{10}N$.
- **Large Counts:** Both np_0 and $n(1 - p_0)$ are at least 10.

Note the difference in the large n condition:

Confidence Interval	**Hypothesis Test**
use p-hat so we get the *observed* numbers of successes and failures	use p0 so we get the *expected* numbers of successes and failures

What is a test statistic? What does it measure? Is the formula on the formula sheet?

A **test statistic** measures how far a sample statistic diverges from what we would expect if the null hypothesis H_0 were true, in standardized units. That is,

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

It's set up like the z score that we did first semester.

Read 557–560

What are the four steps for conducting a significance test? What is required in each step?

State: What *hypotheses* do you want to test, and at what *significance level*? Define any *parameters* you use.

Plan: Choose the appropriate inference *method*. Check *conditions*.

Do: If the conditions are met, perform *calculations*.

- Compute the **test statistic**.
- Find the ***P*-value**.

Conclude: Make a *decision* about the hypotheses in the context of the problem.

What test statistic is used when testing for a population proportion? Is this on the formula sheet?

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

If the data won't support H_a (such as if p-hat is lower than the hypothesized value of p and the H_a says ">"), don't bother conducting the test.

According to a 2015 random sample conducted by the Pew Research Center, 852 of 1343 Facebook users reported getting their news from Facebook. Is this convincing evidence that the majority of Facebook users use the social media site to get news?

HW #15: page 552 (19, 23, 25–28), page 570 (31–39 odd)

9.2 Two-sided tests for a proportion

Read 562–564

When the accounting firms audit company financial records for fraud, they can use a test based on Benford's law. Benford's law states that the distribution of first digits in many real-life sources of data is not uniform. In fact, when there is no fraud, about 17.6% of the numbers in financial records begin with the digit 2. However, if the proportion of first digits that are 2 is significantly different from 0.176 in a random sample of records, an auditor would conduct a much more thorough investigation of the company. Suppose that a random sample of 300 expenses from a company's financial records results in only 38 expenses that begin with the digit 2. Should auditors do a more thorough investigation of this company?

Describe a Type I and Type II error in this context.

We can use a confidence interval to decide between two hypotheses whenever H_a is two-sided (when it has $\neq$).
The advantage is that the confidence interval gives a set of plausible values for p or μ.
We will <u>not</u> use a CI to decide between two hypotheses when H_a is one-sided (when it has < or >).

Alternate Example: *Benford's law and fraud*
A 95% confidence interval for the true proportion of expenses that begin with the digit 2 for the company in the previous Alternate Example is (0.089, 0.164). Does the interval provide convincing evidence that the company should be investigated for fraud?

HW #16: page 571 (41–49 odd, 63)
9.1 Type II Errors and the Power of a Test
Can you use your calculator for the Do step? Are there any drawbacks to this method?

Read 565–569
What the power of a test?

> The **power** of a test against a specific alternative is the probability that the test will reject H_0 at a chosen significance level α when the specified alternative value of the parameter is true.

The only calculation of power you are responsible for is this relationship between power and beta, β, the P(Type II error):

Power + P(Type II error) =1 or Power + β =1

In the potato example on page 565 of the book, a shipment of potatoes is rejected if there is evidence that more than 8% of the shipment is blemished. So, H_0: p=0.08 and H_a: p>0.08 and then suppose that the true proportion of blemished potatoes in some shipment is $p = 0.10$.
This means that we should reject H_0 because $p = 0.10 > 0.08$.

(a) Will the inspector be more likely to find convincing evidence that $p > 0.08$ if he looks at a small sample of potatoes or a large sample of potatoes? How does sample size affect power?

As n increases, power _____.

(b) Will the inspector be more likely to find convincing evidence that $p > 0.08$ if he uses $\alpha = 0.10$ or $\alpha = 0.01$? How does the significance level affect power?

As the significance level, α, increases, power _____.

(c) Suppose that a second shipment of potatoes arrives and the proportion of blemished potatoes is $p = 0.50$. Will the inspector be more likely to find convincing evidence that $p > 0.08$ for the first shipment ($p = 0.10$) or the second shipment ($p = 0.50$)? How does "effect size" affect power?

As the effect size increases, power _____.

(d) Note also that standard deviation affects power—if the standard deviation increases, power decreases.

An analogy to understand statistical power: Looking for a tool in a basement
(Adapted from John Hartung, SUNY HSC Brooklyn)

You send someone into the basement to find a tool. He comes back and says "it isn't there". What do you conclude? Is the tool there or not? There is no way to be sure.	You set up an experiment or survey to look for evidence of an effect/difference/change. Our results come back to say that there is no effect/difference/change. What do you conclude? Is the effect/difference/change there or not? There is no way to be sure.
Power = probability that the person would have found the tool, if the tool really is in the basement.	Power = probability that significant evidence of the effect/difference/change would have been found, if the effect/difference/change really exists.

	More likely to find the tool	More likely to correctly reject H_0	
How long did he spend looking?	Long time spent looking	Large sample size	Sample size, n
How big is the tool?	Large tool, like snow shovel	Large effect	Effect size
How messy is the basement?	Organized basement	Small standard deviation	Standard deviation
How trustworthy is the person?	High level of honesty	Large significance level	Significance level, α

(e) Suppose that the true proportion of blemished potatoes is $p = 0.11$. If $\alpha = 0.05$, the power of the test is 0.76. Interpret this value.

(f) What is the probability of a Type II error for this test? Interpret this value.

In the Benford's Law and fraud example earlier (H_0: p=0.176 and H_a: p≠0.176), suppose that $p = 0.25$. That is, 25% of all financial records at this company begin with the digit 2. When $\alpha = 0.05$, the power of the test is 0.58.
(a) Interpret this value.

(b) How can auditors increase the power of their test?

(c) For what values of p would the power of the test be greater than 0.58, assuming everything else stayed the same?

HW #17 page 572 (51–57 odd, 59–62)
9.3 Significance Tests for a Population Mean
Read 574–579
What are the three conditions for conducting a significance test for a population mean?

- **Random**: The data come from a well-designed random sample or randomized experiment.
  - **10%**: When sampling without replacement, check that $n \leq \frac{1}{10}N$.
- **Normal/Large Sample**: The population has a Normal distribution or the sample size is large ($n \geq 30$). If the population distribution has unknown shape and $n < 30$, use a graph of the sample data to assess the Normality of the population. Do not use t procedures if the graph shows strong skewness or outliers.

What test statistic do we use when testing a population mean? Is the formula on the formula sheet?

$$t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

How do you calculate *P*-values using the *t* distributions?

Read 579–582

Abby noticed that her internet at home seemed slower than she would expect, considering that she pays for 20Mbps service. To investigate, she randomly selected 18 different times during the next month and ran an internet speed test. Here are the download speeds she obtained (Mbps):

21.17 19.06 17.96 18.21 19.55 20.33 20.99 18.99 18.73
19.95 19.33 19.80 19.98 18.70 20.08 19.5 19.37 19.83

(a) Do these data provide convincing evidence at the $\alpha = 0.10$ level that Abby's internet service is slower than advertised, on average?

(b) Given your conclusion in part (a), which kind of mistake—a Type I or a Type II error—could you have made? Explain what this mistake would mean in context.

HW #18 page 573 (54–58 even), page 595 (65, 69, 73)
9.3 Two-sided tests for μ
Read 582–583
Can you use your calculator for the Do step? Are there any drawbacks?

Read 583–586

Many US homes have water supply lines made of copper tubing and a common size is to have 0.625 inch exterior diameter. Fittings made to connect this tubing has to have openings wide enough so that the tubing can be inserted but not so wide that the tubing is not secure and prone to leaking. These fittings are supposed to have an interior diameter of 0.627 inches, but the actual diameter varies a little. To ensure that the fittings are being made correctly, a worker inspects a random sample of 50 fittings every hour, measuring their interior diameter. One sample had a mean of 0.6267 with a standard deviation of 0.0018 and a standard error of 0.00025.

(a) Interpret the standard deviation and the standard error provided.

(b) What are the two explanations for why $\bar{x}$ = 0.6267?

(c) Do these data give convincing evidence that the mean interior diameter of fittings produced this hour is not 0.627 mm? Use a significance test with α = 0.05 to find out.

(d) Calculate a 95% confidence interval for μ. Does your interval support your decision from (c)?

HW #19: page 597 (75, 77, 79, 83)

9.3 Paired Data and Using Tests Wisely

Read 586–589

Kelli and Tim decided to investigate which was faster at a certain fast food restaurant: the drive through or the counter inside. To collect their data, they randomly selected 12 times during a week, went to the same fast food restaurant, and bought the same item. However, one of them used the drive through and the other ordered inside. To decide which each of them would use, they flipped a coin. If it was heads, Kelli used the drive through and Tim went inside. If it was tails, Kelli went inside and Tim used the drive through. They each recorded the time in seconds it took them from the end of their order to getting their food. Carry out a test to see if there is convincing evidence that the drive through is faster.

Time for Drive through (seconds)	Time for Going inside (seconds)
332	347
221	475
502	455
411	536
147	179
285	345
153	230
361	259
348	328
253	352
320	344
382	395

Read 592–593
What is the difference between statistical and practical significance?

What is the problem of multiple tests?

Suppose that 20 significance tests were conducted and in each case the null hypothesis was true. If we are using a 5% significance level, each individual test has a 0.95 probability of avoiding a Type I error. What is the probability that we avoid a Type I error in all 20 tests?

What is the probability that we make a Type I error in at least one of these 20 tests?

HW #20: page 588 (85–93 odd, 95–102)
Chapter 9 Review
Read 602–603
FRAPPY! *2009B #5 (bottle filling machine, t-test, sim of SD test)*
HW #21: page 604 Chapter 9 Review Exercises
Review Chapter 9
HW #22: page 605 Chapter 9 AP Statistics Practice Test
Chapter 9 Test