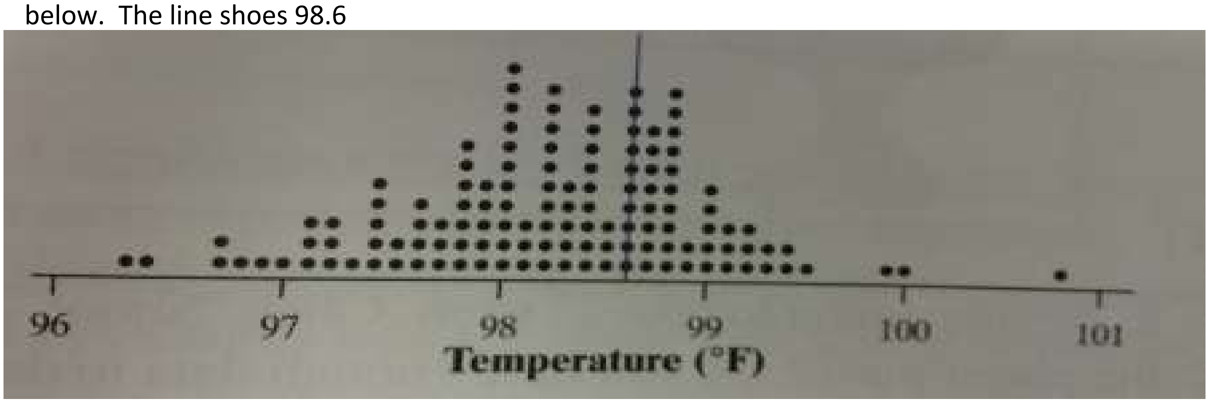
Chapter 11 – Testing a Claim

Sometimes when you are sick, your forehead feels really warm. You might have a fever. How can you find out whether you do? By taking your temperature of course. But what temperature should the thermometer show if you are healthy? Is this temp the same for everyone?

Several years ago, researchers conducted a study to determine whether the "accepted" value for normal body temp, 98.6 F, is accurate. They used an oral thermometer to measure the temps of a random sample of healthy men and women 18 to 40. As is often the case, researchers did not provide their original data.

Allen Shoemaker, from Calvin College, produced a data set with the same properties as the original temperature readings. His data set consists of one oral temp reading for each of the 130 randomly chosen healthy 18 to 40 year olds. A dotplot of Shoemaker's temperature data is shown below. The line shoes 98.6



Give two reasons why the mean temperature was $\bar{x} = 98.25\ ^\circ F$ instead of 98.6

The true average is less than 98.6

The true average is 98.6 and we got a sample mean this low by chance.

62.3% of the temperature readings were less than 98.6. Why is $\hat{p} = .623$ instead of .50?

The true proportion of people with temps less than 98.6 is more than 50%

The true proportion is 50% and we got a sample proportion this high by chance.

Let's look at a free throw shooter

http://bcs.whfreeman.com/ips4e/cat_010/applets/TestSignificance.html

So in the free throw shooter applet, the interest is the proportion, p, of free throws that the player will make if he shoots forever. Our claim, we will say H_0 ,is that they will make .80. If the player only makes

40% we have strong evidence that H_0 is false, but what if he shoots 76%, what does that say about long term percent. It could be less than 80%, but maybe not. Sometimes we like to say $p < .80$ instead of $p = .8$

11.1 – Significance Tests (The Basics)

Significance Test – formal procedure for using observed data to decide between two competing claims (hypotheses).

These usually deal with p or μ (population mean or population proportion)

Stating Hypotheses

- Must begin with a careful statement of the claims we want to compare

In free throw example, the player claimed that his long run proportion of made free throws is $p = .80$. this what we want to seek evidence AGAINST. We call it null hypotheses and represent it H_0. The claim we hope to show to be true or suspect to be true is the alternative hypothesis. This is shown as H_a. It our case with the basketball player, we suggest H_a: $p < .80$

Definitions

- Null Hypothesis – The claim we weigh evidence against in a statistical test.
- Alternative Hypothesis – The claim about the population that we are trying to find evidence for

In the basketball example, our alternative hypothesis is **one sided** because we are interested only in whether the player is overstating his free-throw shooting ability.

Here is a two sided example:

At the Hawaii Pineapple Company, managers are interested in the size of the pineapples grown in the company's field. Last year, the mean weight of the pineapples harvested from one large field was 31 ounces. A different irrigation system was installed in this field after the growing season. Managers wonder how this change will affect the mean weight of pineapples grown in the field this year. State appropriate hypotheses for performing a significance test.

SOLUTION: the parameter of interest is the mean weight, μ, of all pineapples grown in the field this year. They want to know if it is different than last year, so

H_0: $\mu = 31$ H_a: $\mu \neq 31$

DEFINITION:

- A **one sided test** states that it is either larger than or smaller than the null hypothesis.
- A **two sided test** states that it is different that the null (larger OR smaller) not equal to.

Your hypotheses always refer to a POPULATION….NEVER A SAMPLE!!!!

Never use p hat or x bar!!!

Let's try some

1. A recent study on "The relative age effect and career success: Evidence from corporate CEOs" (Economics Letters 117 (2012)) suggests that people born in June and July are under-represented in the population of corporate CEOs. This "is consistent with the 'relative-age effect' due to school admissions grouping together children with age differences up to one year, with children born in June and July disadvantaged throughout life by being younger than their classmates born in other months." In their sample of 375 corporate CEOs, only 45 (12%) were born in June and July. Is this *convincing* evidence that the true proportion *p* of *all* corporate CEOs born in June and July is smaller than 2/12?

2. Mike is an avid golfer who would like to improve his play. A friend suggests getting new clubs and lets Mike try out his 7-iron. Based on years of experience, Mike has established that the mean distance that balls travel when hit with his old 7-iron is $\mu = 175$ yards with a standard deviation of $\sigma = 15$ yards. He is hoping that this new club will make his shots with a 7-iron more consistent (less variable), and so he goes to the driving range and hits 50 shots with the new 7-iron.

Conditions for Significance Tests

These are the same conditions that we checked when working with confidence intervals

- Simple Random Sample
- Normality
    - For means n > 30
    - For proportions np > 10 and n(1 – p) > 10
- Independent
    - 10(sample) < population

When thinking about the null hypothesis, it is treated like a criminal….innocent until proven guilty. We assume that it is innocent and we are trying to provide convincing evidence that it is not. The null hypothesis states the claim that we are seeking evidence against. The probability that measures the **strength** of the evidence against H_0 and in favor of H_a is called a **P-value**

Definition: A P-value is the probability of getting a statistic at least as extreme as the one observed in a study, assuming the null hypothesis is true.

- Small p values are evidence against H_0 because they say that the observed result is unlikely to occur when H_0 is true. Large P-values **fail to give convincing evidence** against H_0 and are in favor of H_a because they say that the observed result is likely to occur by chance alone when H_0 is true.

---- In the CEO Example, Give two explanations for why the sample proportion was below 2/12.

1. June and July kids are disadvantaged.

2. There is no disadvantage—the lower percentage was due to chance alone.

In the CEO example, the *P*-value = $P(\hat{p} \leq 0.12 \mid p = 2/12) = 0.008$. Interpret this value.

Assuming that the birthdays of corporate CEOs are uniformly distributed throughout the year, there is a 0.008 probability of getting a sample proportion as small or smaller than 0.12 by chance alone.

---- back to the golf example

When Mike was testing a new 7-iron, the hypotheses were $H_0: \sigma = 15$ versus $H_a: \sigma < 15$ where σ = the true standard deviation of the distances Mike hits golf balls using the new 7-iron. Based on a sample of shots with the new 7-iron, the standard deviation was $s_x = 13.9$ yards.

(a) What are the two explanations for why $s < 15$?

(b) A significance test using the sample data produced a *P*-value of 0.28. Interpret the *P*-value in this context.

If the true standard deviation is 15 yards, then there is a probability of 0.28 that the sample standard deviation would be 13.9 yards or smaller by chance alone.

Another example:

Calcium is a vital nutrient for healthy bones and teeth. The National Institutes of Health recommends a calcium intake of 1300 mg per day for teenagers. The NIH is concerned that teenagers aren't getting enough calcium. Is this true?

Researchers want to perform a test of:

$H_0 : \mu = 1300$

$H_a : \mu < 1300$

Where μ is the true mean daily calcium intake in the population of teenagers. They ask a random sample of 20 teens to record their food and drink consumption for 1 day. The researchers then compute the calcium intake for each student. Data analysis reveals that $\bar{x} = 1198\ mg$ and $s_x = 411\ mg$. After checking that conditions were met, researchers performed a significance test and obtained a P-value of .1404.

a) Explain what it would mean for the null hypothesis to be true in this setting.

b) Interpret the P-value in context.

Assuming that the mean daily calcium intake is 1300mg, there is a .1404 probability of getting a sample mean of 1198mg or less by chance in a random sample of 20 teens.

The final step in a significance test is to draw a conclusion about the claims you were testing. You either **reject H_0** or **fail to reject H_0**. If our sample result is too unlikely to have happened by chance then we'll reject the null and say that there is convincing evidence for H_a. Otherwise we will fail to reject H_0 and say there is NOT convincing evidence for H_a.

Just like the trial example Guilty means reject H_0, Innocent means fail to reject H_0

- ***P-value is small → reject*** H_0 → [redacted] H_a [redacted]
- ***P-value is large → fail to reject*** H_0 → [redacted] H_a [redacted]

There is no rule for how small a P-value is required to reject H_0, but we can compare it with a fixed value called the **significance level….known as α (alpha).**

If we choose α of .05, we are requiring that the data give evidence against H_0 so strong that it would happen less than 5% of the time by chance when H_0 is true.

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

Example:

A company has developed a new deluxe AAA battery that is supposed to last longer than its regular AAA battery. However, these new batteries are more expensive to produce, so the company would like to be convinced that they really do last longer. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. The company selects an SRS of 15 new batteries and uses them continuously until they are completely drained. The sample mean lifetime is $\bar{x} = 33.9\ hours$. A significance test is performed using the hypotheses

H_0 : μ=30 hours H_a: $\mu > 30$ hours, where μ is the true mean lifetime of the new deluxe AAA batteries. The resulting p-value is .0729

What conclusion would you make for each of the following significance levels? Justify your answer.

a) $\alpha = .10$ b) $\alpha = .05$

a) Because our p value is less than alpha, we reject H_0. We have convincing evidence that the deluxe lasts longer than 30 hours on average.

b) Because p value is greater than alpha, we fail to reject H_0. We do not have convincing evidence that the deluxe lasts longer than 30 hours on average.

Page 693 #3 – 6

Page 698 #7, 8, 10, 11, 12

Page 701 # 13, 14, 15, 16