Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

0. Sampling students, continued.
1. Estimating the mean for a quantitative variable, t-test and song time example.
2. Significance level, type I and type II errors.
3. Power.
4. Confidence Intervals for a proportion and the dog sniffing cancer example.
5. CIs for a proportion and the Affordable Care Act example.

Reminder: lectures July 7 and 9 will be pre-recorded with links on course website.
The course website is http://www.stat.ucla.edu/~frederic/13/sum25

Start reading chapter 4.

0. Sampling Students, continued.

- What would happen if we took all possible random samples of 30 students from this population?
  - The averages of the statistics would match the parameters exactly
- Statistics computed from SRSs cluster around the parameter.
- Why is this an unbiased sampling method?
  - There is no tendency to over or underestimate the parameter.
- The sampling method and statistic you choose determine if a sampling method is biased.
- A sample mean of a simple random sample is an unbiased estimate of the population mean. Same for proportions instead of means.

Sampling Students

- We can *generalize* when we use simple random sampling because it creates:
  - A sample that is representative of the population.
  - A sample statistic that is unbiased and thus close to the parameter for large n.

Sampling Students

- If the researcher at the College of the Midwest uses 75 students instead of 30 with the same early morning sampling method will it be less biased?
- No. Selecting more students *in the same manner* doesn't fix the tendency to oversample students who live on campus.
- A smaller sample that is random is actually more accurate.

Sampling Students

- What is an advantage of a larger sample size?
  - Less sample to sample variability.

1. Inference for a Single Quantitative Variable. t-test and song time example.

Section 2.2

https://www.youtube.com/watch?v=ho7796-au8U

Example 2.2: Estimating Elapsed Time

- Students in a stats class (for their final project) collected data on students’ perception of time
- Subjects were told that they’d listen to music and be asked questions when it was over.
- 10 seconds of the Jackson 5’s “ABC” and subjects were asked how long they thought it lasted
- Can students accurately estimate the length?

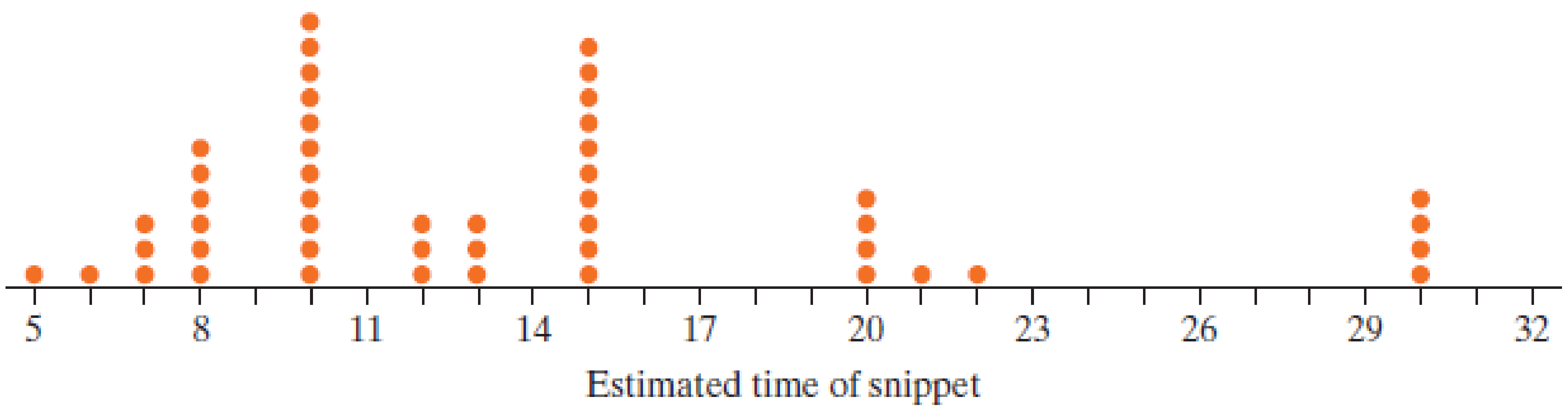
Hypotheses

Null Hypothesis: People will accurately estimate the length of a 10 second-song snippet, on average. ($\mu = 10$ seconds)

Alternative Hypothesis: People will not accurately estimate the length of a 10 second-song snippet, on average. ($\mu \neq 10$ seconds)

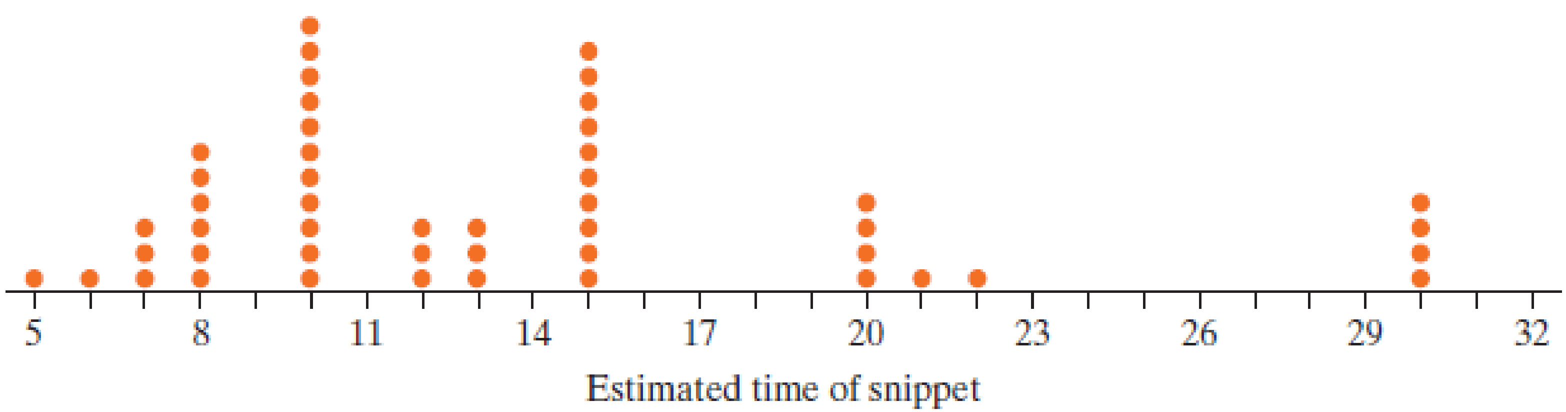
Estimating Time

- A sample of 48 students on campus were subjects and song length estimates were recorded.
- What does a single dot represent?
- What are the observational units? Variable?



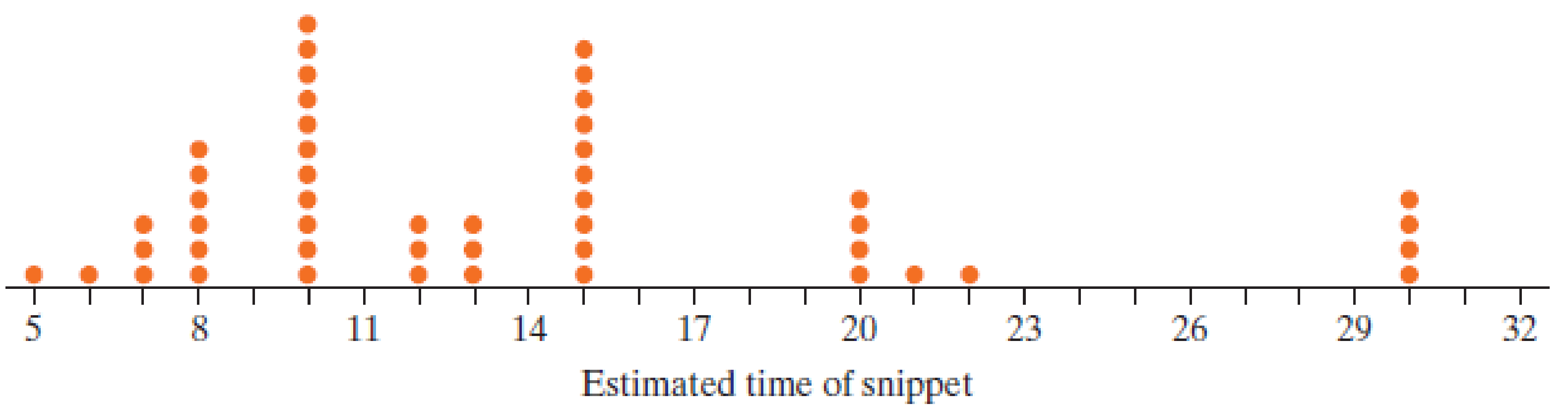
Skewed, mean, median

- The distribution obtained is not symmetric, but is **right skewed.**
- When data are skewed right, the **mean** gets pulled out to the right while the **median** is more resistant to this.



Mean vs Median

- The mean is 13.71 and the median is 12.
- How would these numbers change if one of the people that gave an answer of 30 seconds actually said 300 seconds?
- The standard deviation is 6.5 sec. Also not resistant to outliers.



Inference

- **H_0**: $\mu = 10$ seconds
- **H_a**: $\mu \neq 10$ seconds
- Our problem now is, how do we develop a null distribution?
  - Here we don't have population data that reflects our null hypothesis where $\mu = 10$ seconds.
  - All we have is our sample of 48.

Population?

- We need to come up with a large data set that we think our population of time estimates might look like **under a true null.**
- We might assume the population is skewed (like our sample) and has a standard deviation similar to what we found in our sample, but has a mean of 10 seconds.
- The book recommends using an applet for this. We could use *R*, or do a (theory-based) t-test.

Theory-Based Test

- Using simulations to create a population each time we want to run a test of significance is extremely time consuming and cumbersome.
- The null distribution that we developed can be predicted with theory-based methods.
- We know it will be centered on the mean given in the null hypothesis.
- We can also predict its shape and its standard deviation.

t-distribution

- The shape is very much like a normal distribution, but slightly wider in the tails and is called a t-distribution.
- The t-statistic is the standardized statistic we use with a single quantitative variable that looks approximately normal, when the sample size is small, and the statistic can be found using the formula:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The $s/\sqrt{n}$ (standard deviation of our sample divided by the square root of the sample size) is called the standard error and is an estimate for the standard deviation of the null distribution.

Here $t = \frac{13.71 - 10.0}{6.5/\sqrt{48}} = 3.95.$

p-value = 2*(1-pt(3.95,df=47)) = 0.000261.

Validity Conditions

- The observations must be independent.
- The population must be normally distributed!
- The book says you need the sample size to be at least 20 for the t-test, but this is not technically true. The whole point of the t-test is you can use it even when your sample size is small, provided the two assumptions above hold.

But it is often hard to have any idea if the population is normal without having at least 20 observations.

Estimating Time

Formulate Conclusions.

- Based on our small p-value, we can conclude that our subjects did not accurately estimate the length of a 10-second song snippet and in fact they significantly overestimated it.
- How far can we generalize this?

Summary

- When we test a single quantitative variable, our hypothesis has the following form:
  - **H_0**: μ = some number
  - **H_a**: $\mu \neq$ some number, $\mu <$ something or $\mu >$ something.
- We can get our data (or mean, sample size, and SD for our data) and use the Theory-Based Inference to determine the p-value.
- The p-value we get with this test has the same general meaning as from a test for a single proportion.

2. Significance level, type I and type II errors

Section 2.3

Significance Level

- We think of a p-value as telling us something about the strength of evidence from a test of significance.
- The lower the p-value the stronger the evidence.
- Sometimes it makes sense to think of this in more black and white terms. Either we reject the null or not.

Significance Level

- The value that we use to determine how small a p-value needs to be to provide convincing evidence whether or not to reject the null hypothesis is called the **significance level**.
- We reject the null when the p-value is less than or equal to the significance level.
- The significance level is often represented by the Greek letter alpha, α.

Significance Level

- Typically we use 0.05 for our significance level. There is nothing magical about 0.05. We could set up our test to make it
  - harder to reject the null (smaller significance level say 0.01) or
  - easier (larger significance level say 0.10).

Type I and Type II errors

- In medical tests:
  - A type I error is a false positive. (conclude someone has a disease when they don’t.)
  - A type II error is a false negative. (conclude someone does not have a disease when they actually do.)
- These types of errors can have very different consequences.

Type I and Type II Errors

TABLE 2.9 A summary of Type I and Type II errors

		What is true (unknown to us)	
		Null hypothesis is true	Null hypothesis is false
What we decide (based on data)	Reject null hypothesis	Type I error (false alarm)	Correct decision
	Do not reject null hypothesis	Correct decision	Type II error (missed opportunity)

-

Type I and Type II errors

TABLE 2.10 Type I and Type II errors summarized in context of jury trial

		What is true (unknown to the jury)	
		Null hypothesis is true (defendant is innocent)	Null hypothesis is false (defendant is guilty)
What jury decides (based on evi-dence)	Reject null hypothesis (Jury finds defendant guilty)	Type I error (false alarm)	Correct decision
	Do not reject null hypothe-sis (Jury finds defendant not guilty)	Correct decision	Type II error (missed opportunity)

The probability of a Type I error

- The probability of a type I error, when the null hypothesis is true, is the significance level.
- Suppose the significance level is 0.05. If the null is true we would reject it 5% of the time and thus make a type I error 5% of the time.
- If you make the significance level lower, you have reduced the probability of making a type I error, but have increased the probability of making a type II error.

The probability of a Type II error

- The probability of a type II error is more difficult to calculate.
- In fact, the probability of a type II error is not even a fixed number. It depends on the value of the true parameter.
- The probability of a type II error can be very high if:
  - The true value of the parameter and the value you are testing are close.
  - The sample size is small.

3. Power.

- Power is 1 – P(Type II error). Usually expressed as a function of μ.
- Recall Type I and Type II errors.
  - A type I error is a false positive. Rejecting the null when it is true.
  - A type II error is a false negative. Failing to reject the null when the null is false.

Power

- The probability of rejecting the null hypothesis when it is false is called the **power** of a test.
- Power is 1 minus the probability of type II error.
- We want a test with high power and this is aided by
  - A large effect size, i.e. true μ far from the parameter in the null hypothesis.
  - A large sample size.
  - A small standard deviation.
  - Significance level. A higher sign. level means greater power. The downside is that you increase the chance of making a type I error.

4. Estimation and confidence intervals.

Chapter 3

Chapter Overview

- So far, we can only say things like
  - “We have strong evidence that the long-run frequency of death within 30 days after a heart transplant at St. George's Hospital is greater than 15%.”
  - “We do not have strong evidence kids have a preference between candy and a toy when trick-or-treating.”
- We want a method that says
  - “I believe 68 to 75% of all elections can be correctly predicted by the competent face method.”

Confidence Intervals

- Interval estimates of a population parameter are called **confidence intervals**.
- We will find confidence intervals three ways.
  - Through a series of tests of significance to see which proportions are plausible values for the parameter.
  - Using the standard error (the standard deviation of the simulated null distribution) to help us determine the width of the interval.
  - Through traditional theory-based methods, i.e. formulas.

Statistical Inference: Confidence Intervals

Section 3.1

Can Dogs Sniff Out Cancer?

Section 3.1

Can Dogs Sniff Out Cancer?

Sonoda et al. (2011). Marine, a dog originally trained for water rescues, was tested to see if she could detect if a patient had colorectal cancer by smelling a sample of their breath.

- She first smells a bag from a patient with colorectal cancer.
- Then she smells 5 other samples; 4 from normal patients and 1 from a person with colorectal cancer
- She is trained to sit next to the bag that matches the scent of the initial bag (the "cancer scent") by being rewarded with a tennis ball.

Can Dogs Sniff Out Cancer?

In Sonoda et al. (2011). Marine was tested in 33 trials.

- Null hypothesis: Marine is randomly guessing which bag is the cancer specimen (π = 0.20)
- Alternative hypothesis: Marine can detect cancer better than guessing (π > 0.20)

π represents her long-run probability of identifying the cancer specimen.

Can Dogs Sniff Out Cancer?

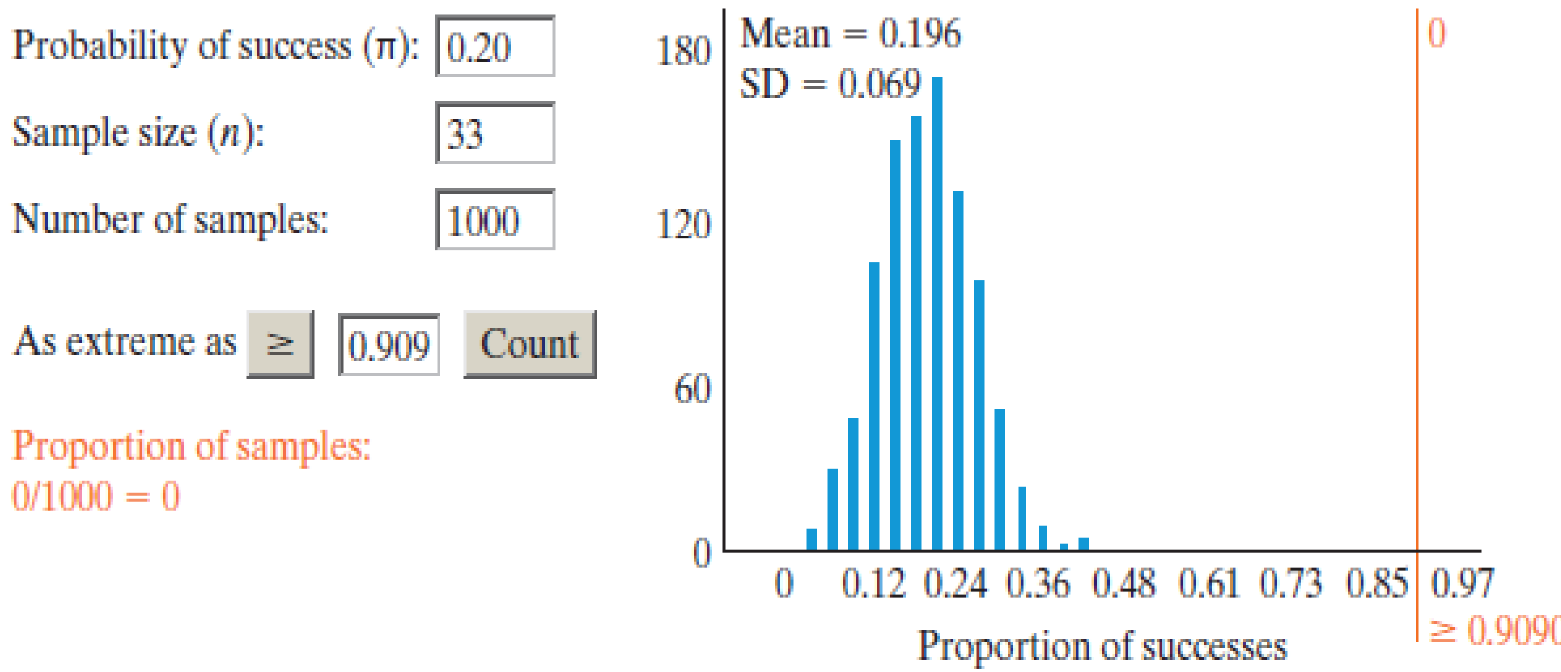
- 30 out of 33 trials resulted in Marine correctly identifying the bag from the cancer patient
- So our sample proportion is

$$\hat{p} = \frac{30}{33} \approx 0.909$$

- Do you think Marine can detect cancer?
- What sort of p-value will we get?

Can Dogs Sniff Out Cancer?

Our sample proportion lies more than 10 standard deviations above the mean and hence our p-value ~ 0.

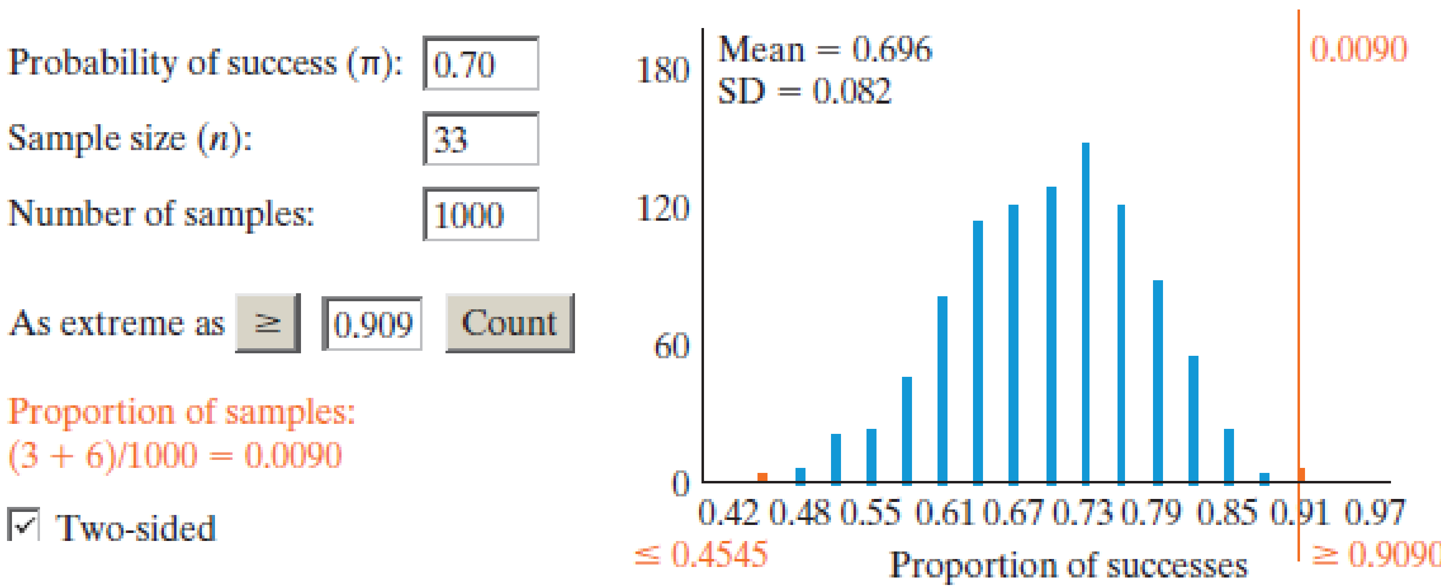


Can Dogs Sniff Out Cancer?

- Can we estimate Marine’s long run frequency of picking the correct specimen?
- Since our sample proportion is about 0.909, it is plausible that 0.909 is a value for this frequency. What about other values?
- Is it plausible that Marine’s frequency is actually 0.70 and she had a lucky day?
- Is a sample proportion of 0.909 unlikely if $\pi = 0.70$?

Can Dogs Sniff Out Cancer?

- H_0: $\pi = 0.70$ H_a: $\pi \neq 0.70$
- We get a small p-value (0.0090) so we can essentially rule out 0.70 as her long run frequency.

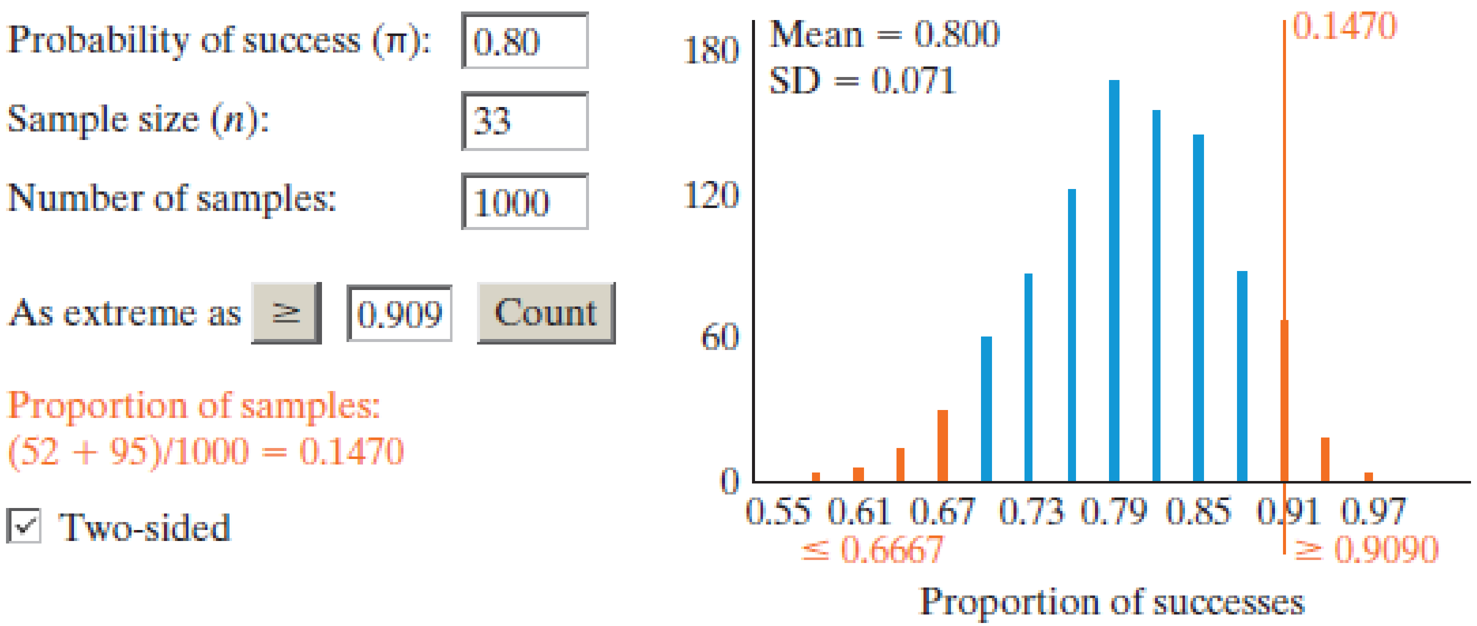


Can Dogs Sniff Out Cancer?

- What about 0.80?
- Is 0.909 unlikely if π = 0.80?

Can Dogs Sniff Out Cancer?

- H_0: $\pi = 0.80$ H_a: $\pi \neq 0.80$
- We get a large p-value (0.1470) so 0.80 is a ***plausible*** value for Marine’s long-run frequency.



Developing a range of plausible values

- If we get a small p-value (like we did with 0.70) we will conclude that the value under the null is not plausible. This is when we reject the null hypothesis.
- If we get a large p-value (like we did with 0.80) we will conclude the value under the null is plausible. This is when we can’t reject the null.

Developing a range of plausible values

- One could use software (like the one-proportion applet the book recommends) to find a range of plausible values for Marine's long term probability of choosing the correct specimen.
- We will keep the sample proportion the same and change the possible values of π.
- We will use 0.05 as our cutoff value for if a p-value is small or large. (Recall that this is called the **significance level**.)

Can Dogs Sniff Out Cancer?

- It turns out values between 0.761 and 0.974 are plausible values for Marine’s probability of picking the correct specimen.

Probability under null	0.759	0.760	0.761	0.762	…….	0.973	0.974	0.975	0.976
p-value	0.042	0.043	**0.063**	**0.063**		**0.059**	**0.054**	0.049	0.044
Plausible?	No	No	Yes	Yes	……… Yes	Yes	Yes	No	No

Can Dogs Sniff Out Cancer?

- (0.761, 0.974) is called a ***confidence interval*.**
- Since we used 5% as our significance level, this is a 95% confidence interval. (100% – 5%)
- 95% is the ***confidence level*** associated with the interval of plausible values.

Can Dogs Sniff Out Cancer?

- We would say we are 95% confident that Marine's probability of correctly picking the bag with breath from the cancer patient from among 5 bags is between 0.761 and 0.974.
- This is a more precise statement than our initial significance test which concluded Marine's probability was more than 0.20.
- Sidenote: We do not say P{π is in (.761, .974)} = 95%, because π is not random. The *interval* is random, and would change with a different sample. If we calculate an interval this way, then P(interval contains π) = 95%.

Confidence Level

- If we increase the confidence level from 95% to 99%, what will happen to the width of the confidence interval?

Can Dogs Sniff Out Cancer?

- Since the confidence level gives an indication of how sure we are that we captured the actual value of the parameter in our interval,
  to be more sure our interval should be wider.
- How would we obtain a wider interval of plausible values to represent a 99% confidence level?
  - Use a 1% significance level in the tests.
  - Values that correspond to 2-sided p-values larger than 0.01 should now be in our interval.

5. 1.96SE and Theory-Based Confidence Intervals for a Single Proportion and ACA example.

Section 3.2

Introduction

- Section 3.1 found confidence intervals by doing repeated tests of significance (changing the value in the null hypothesis) to find a range of values that were plausible for the population parameter (long run probability or population proportion).
- This is a very tedious way to construct a confidence interval.
- We will now look at two others way to construct confidence intervals [1.96SE and Theory-Based].

The Affordable Care Act

Example 3.2

The Affordable Care Act

- A November 2013 Gallup poll based on a random sample of 1,034 adults asked whether the Affordable Care Act had affected the respondents or their family.
- 69% of the sample responded that the act had no effect. (This number went down to 59% in May 2014 and 54% in Oct 2014.)
- What can we say about the proportion of all adult Americans that would say the act had no effect?

The Affordable Care Act

- We could construct a confidence interval just like we did last time.
- We find we are 95% confident that the proportion of all adult Americans that felt unaffected by the ACA is between 0.661 and 0.717.

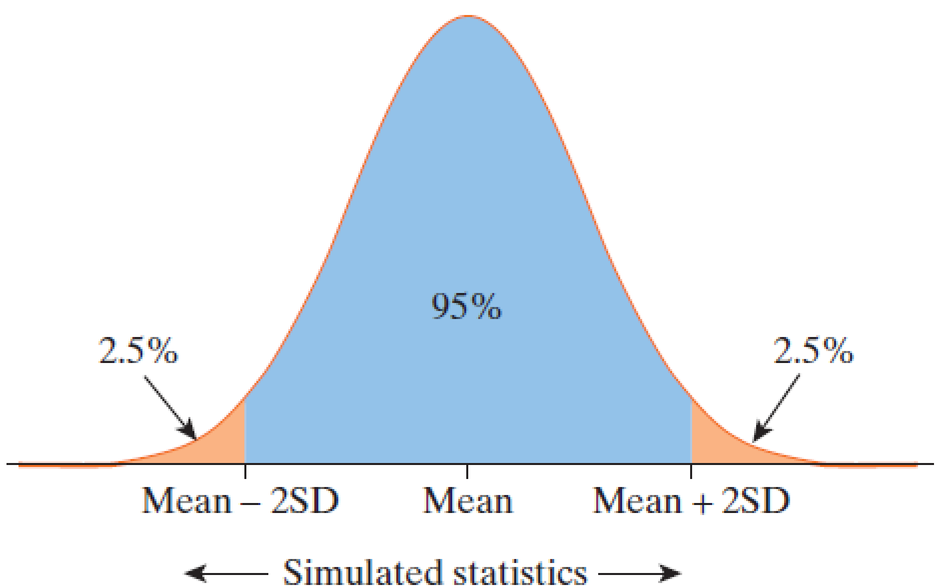
Probability under null	0.659	0.660	0.661	…………	0.717	0.718	0.719
Two-sided p-value	0.0388	0.0453	0.0514	…………	0.0517	0.0458	0.0365
Plausible value (0.05)?	No	No	Yes	…………	Yes	No	No

Short cut?

- The method we used last time to find our interval of plausible values for the parameter is tedious and time consuming.
- Might there be a short cut?
- Our sample proportion should be the middle of our confidence interval.
- We just need a way to find out how wide it should be.

1.96SE method

- When a statistic is normally distributed, about 95% of the values fall within 1.96 standard errors of its mean with the other 5% outside this region



1.96SE method

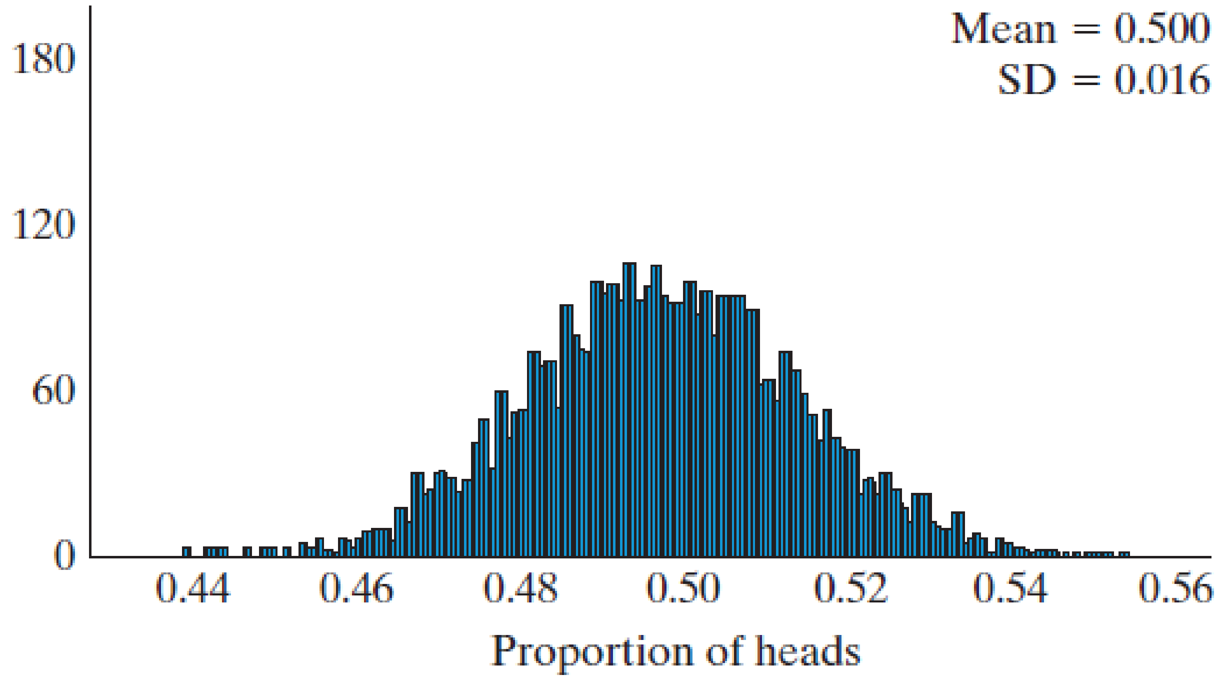
- So we could say that a parameter value is plausible if it is within 1.96 standard errors from our best estimate of the parameter, our observed sample statistic.
- This gives us the simple formula for a 95% confidence interval of

$$\boldsymbol{\hat{p} \pm 1.96SE}$$

Note that your book calls this the 2SD method but it really should be called the 1.96SE method.

Where do we get the SE?

- Null distribution for ACA with π = 0.5.



1.96SE method

- Using the 1.96SE method on our ACA data we get a 95% confidence interval

$$0.69 \pm 1.96(0.016)$$
$$0.69 \pm 0.031$$

- The $\pm$ part, like 0.031 in the above, is called the **margin of error.**
- The interval can also be written as we did before using just the endpoints; (0.659, 0.721)
- This is approximately what we got with our range of plausible values method (a bit wider).

Theory-Based Methods

- The 1.96SE method only gives us a 95% confidence interval
- If we want a different level of confidence, we can use the range of plausible values (hard) or theory-based methods (easy).
- The theory-based method is valid provided there are at least 10 successes and 10 failures in your sample.

Theory-Based Methods

- With theory-based methods we use normal distributions to approximate our simulated null distributions.
- Therefore we can develop a formula for confidence intervals.

$$\hat{p} \pm multiplier \times \sqrt{\hat{p}(1-\hat{p})/n}.$$

For a 95% CI, the book suggests a multiplier of 2. Actually people use 1.96, not 2. This comes from a property of the normal distribution.

qnorm(.975) = 1.96.

qnorm(.995) = 2.58.

- Let's check out this example using the theory-based method.
- Remember 69% of 1034 respondents were not affected.

$\widehat{p} \pm multiplier \times \sqrt{\hat{p}(1-\hat{p})/n}$

= 69% $\pm$ 1.96 x $\sqrt{.69(1-.69)/1034}$

= 69% $\pm$ 2.82%.

With 2 instead of 1.96 it would be 69% $\pm$ 2.88%.