Announcements:

- New use of clickers: to test for understanding. I will give more clicker questions, and randomly choose five to count for credit each week.
- Discussion this week is not for credit – question/answer, practice problems.
- Chapter 9 practice problems now on website
- Today: Sections 9.1 to 9.4
- **Homework (due Wed, Feb 27):**
  Chapter 9: #22, 26, 40, 144

Chapter 9

Understanding Sampling Distributions: Statistics as Random Variables

Recall: Sample Statistics and Population Parameters

A **statistic** is a numerical value computed from a sample. Its value may differ for different samples.
e.g. sample mean $\bar{x}$, sample standard deviation s, and sample proportion $\hat{p}$.

A **parameter** is a numerical value associated with a population. Considered fixed and unchanging.
e.g. population mean μ, population standard deviation σ, and population proportion p.

Statistical Inference

Statistical Inference: making conclusions about population parameters on basis of sample statistics.

See picture on board in lecture

Two most common procedures:

Confidence interval: an interval of values that the researcher is fairly sure will cover the true, unknown value of the population parameter.

Hypothesis test: uses sample data to attempt to reject (or not) a hypothesis about the population.

The Plan for the Rest of the Quarter

- We will cover statistical inference for five situations; each one has a parameter of interest.
- For each of the five situations we will identify:
  - The parameter of interest
  - A sample statistic to estimate the parameter
- For each of the five situations we will learn about:
  - The *sampling distribution* for the sample statistic
  - How to construct a *confidence interval* for the parameter
  - How to *test hypotheses* about the parameter

How (Statistical) Science Works

The Big Five Parameters (See Table on page 317)

Parameter Name and Description	Symbol for the Population Parameter	Symbol for the Sample Statistic
For Categorical Variables		
One population proportion (or probability)	p	$\hat{p}$
Difference in two population proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
For Quantitative Variables		
One population mean	μ	$\bar{x}$
Population mean of paired differences (dependent)	μ_d	$\bar{d}$
Difference in two population means (independent)	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$

How (Statistical) Science Works

Curiosity about something → Question about a parameter → Collect data appropriate for that parameter → Make inferences about the value of the parameter

Example:

Curiosity: Do a majority of voters favor stricter gun control?

Parameter: p = proportion of *population* of registered voters who do favor stricter gun control. What is the value of p?

Collect data: Ask a random sample of registered voters. Sample statistic = proportion of the *sample* who favor stricter gun control

Make inferences: Use the *sample proportion* to compute a 95% confidence interval for the *population proportion* (parameter)

Structure for the rest of the Quarter

Parameter name and description	Sampling Distribution	Confidence Interval	Hypothesis Test
For Categorical Variables:	Chapter 9	Chapter 10	Chapter 12
One population proportion or binomial probability	Today & Fri.	Mon, Feb 25	Mon, Mar 4
Difference in two population proportions	Friday	Mon, Feb 25	Wed, Mar 6
For Quantitative Variables:	Chapter 9	Chapter 11	Chapter 13
One population mean	Fri, March 8	Mon, Mar 11	Wed, Mar 12
Population mean of paired differences (paired data)	Fri, March 8	Mon, Mar 11	Wed, Mar 12
Difference in two population means (independent samples)	Fri, March 8	Mon, Mar 11	Wed, Mar 12

For Situation 4, we need "Paired Data"

Paired data (or ***paired samples***): when pairs of variables are collected. Only interested in population (and sample) of ***differences***, and not in the original data.

Here are ways this can happen:

- **Each person (unit) measured twice**. Two measurements of same characteristic or trait made under different conditions.
- Similar **individuals are paired** prior to an experiment. Each member of a pair receives a different treatment. Same response variable is measured for all individuals.
- **Two different variables** are measured for each individual. **Interested in amount of difference** between two variables.

Situations 2 and 5: Independent Samples

Two samples are called **independent samples** when the measurements in one sample are not related to the measurements in the other sample.

Here are ways this can happen:

- **Random samples** taken separately from two populations and same response variable is recorded.
- **One random sample** taken and a variable recorded, but units are **categorized** to form two populations.
- Participants **randomly assigned** to one of two treatment conditions, and same response variable is recorded.

Familiar Examples Translated into Questions about Parameters

Situation 1.

Estimate/test the proportion *falling into a category of a categorical variable* OR *a* ***binomial success probability***

Example research questions:
What proportion of American adults believe there is extraterrestrial life? In what proportion of British marriages is the wife taller than her husband?

Population parameter: p = proportion in the <u>population</u> falling into that category.

Sample estimate: $\hat{p}$ = proportion in the <u>sample</u> falling into that category.

Data Example for Situation 1

Question: What *proportion* (p) of all households with TVs watched the Super Bowl? *Get a confidence interval for p*. (Hypothesis test of no use in this example – nothing of interest to test!)

Population parameter:

p = proportion of the *population* of all US households with TVs that watched it.

Sample statistic:

Nielsen ratings, random *sample* of n = 25,000 households.

X = *number* in *sample* who watched the show = 11,510.

$\hat{p} = \frac{X}{n} = \frac{11,510}{25,000}$ = 0.46 = *proportion* of *sample* who watched. This is called "p-hat."

Familiar Examples

Situation 2.

Compare two population proportions using independent samples of size n_1 and n_2. ***Estimate*** *difference;* ***test*** *if 0.*

Example research questions:

- How much difference is there between the proportions that would quit smoking if taking the antidepressant buproprion (Zyban) versus if wearing a nicotine patch?
- How much difference is there between men who snore and men who don't snore with regard to the proportion who have heart disease?

Population parameter: $p_1 - p_2$ = difference between the two population proportions.

Sample estimate: $\hat{p}_1 - \hat{p}_2$ = difference between the two sample proportions.

Data Example for Situation 2

Question: Is the population proportion favoring stricter gun control laws the same now as it was in April 2012?

- Get a *confidence interval* for the population difference.
- *Test* to see if it is statistically significantly different from 0.

Population parameter:

$p_1 - p_2$ = *population* difference in proportions where p_1 is the proportion now, and p_2 was the proportion in April 2012

Sample statistic: Based on CBS News Poll, n_1 and n_2 each about 1,150; 53% favor now and only 39% did in April 2012.

Difference in *sample* proportions is $\hat{p}_1 - \hat{p}_2$ = .53 – .39 = +.14

This is read as "p-hat-one minus p-hat-two"

Note that the parameter and statistic can range from −1 to +1.

Familiar Examples

Situation 3.

Estimate the population mean *of a quantitative variable.*
Hypothesis test *if there is a logical null hypothesis value.*

Example research questions:

- What is the mean time that college students watch TV per day?
- What is the mean pulse rate of women?

Population parameter: μ = population mean for the variable

Sample estimate: $\bar{x}$ = sample mean for the variable

Data Example for Situation 3

Question: Airlines need to know the average weight of checked luggage, for fuel calculations. Estimate with a confidence interval, and test to see if it exceeds airplane capacity.

Population parameter:

μ = mean weight of the luggage for the *population* of all passengers who check luggage.

Sample statistic: Study measured n = 22,353 bags; $\bar{x}$ = 36.7 pounds (st. dev. = 12.8)

Source:
http://www.easa.europa.eu/rulemaking/docs/research/Weight%20Survey%20R20090095%20Final.pdf

Familiar Examples

Situation 4.

Estimate the population mean *of paired differences for quantitative variables, and* ***test*** *null hypothesis that it is 0.*

Example research questions:

- What is the mean difference in weights for freshmen at the beginning and end of the first quarter or semester?
- What is the mean difference in age between husbands and wives in Britain?

Population parameter: μ_d = population mean of differences

Sample estimate: $\bar{d}$ = mean of differences for paired sample

Data Example for Situation 4

Question: How much different on average would IQ be after listening to Mozart compared to after sitting in silence?

- Find *confidence interval* for population mean difference μ_d
- *Test null hypothesis* that $\mu_d = 0$.

Population parameter:

μ_d = *population* mean for the difference in IQ *if* everyone in the population were to listen to Mozart versus silence.

Sample statistic: For the experiment done with n = 36 UCI students, the mean difference for the sample was 9 IQ points. $\bar{d}$ = 9, read "d-bar"

Familiar Examples

Situation 5.

Estimate *the difference between two population means for quantitative variables and* **test** *if the difference is 0.*

Example research questions:

- How much difference is there in mean weight loss for those who diet compared to those who exercise to lose weight?
- How much difference is there between the mean foot lengths of men and women?

Population parameter: $\mu_1 - \mu_2$ = difference between the two population means.

Sample estimate: $\bar{x}_1 - \bar{x}_2$ = difference between the sample means, based on independent samples of size n_1 and n_2

Data Example for Situation 5

Question: Is there a difference in mean IQ of 4-year-old children for the population of mothers who smoked during pregnancy and the population who did not? If so, how much?

- Find *confidence interval* for population difference $\mu_1 - \mu_2$
- *Test null hypothesis* that $\mu_1 - \mu_2 = 0$.

Population parameter:

$\mu_1 - \mu_2$ = difference in the mean IQs for the two *populations*

Sample statistic: Based on a study done at Cornell, the difference in means for two *samples* was 9 IQ points.

$\bar{x}_1 - \bar{x}_2 = 9$, Read as "x-bar-one minus x-bar-two."

Sampling Distributions: Some Background

Notes about statistics and parameters:

- Assuming the sample is representative of the population, the *sample statistic* should represent the *population parameter* fairly well. (Better for larger samples.)
- But… the sample statistic will have some error associated with it, i.e. it won't necessarily *exactly* equal the population parameter. Recall the "margin of error" from Chapter 5!
- Suppose repeated samples are taken from the same population and the sample statistic is computed each time. These sample statistics will *vary,* but in a *predictable way*. The possible values will have a *distribution*. It is called the **sampling distribution** for the statistic.

Rationale And Definitions For Sampling Distributions

Claim: A *statistic* is a special case of a *random variable*.

Rationale: When a sample is taken from a population the resulting numbers are the outcome of a *random circumstance*. That's the definition of a random variable.

Super Bowl example:

- A random circumstance is taking a random sample of 25,000 households with TVs.
- The resulting number (statistic) is the *proportion of those households that watched the Super Bowl.* (*0.46, or 46%*)
- *A different sample would give a different proportion.*

Rationale, Continued

Remember: a random variable is a number associated with the outcome of a random circumstance, which can change each time the random circumstance occurs.

Example: For each different sample of 25,000 households that week, we could have had a different sample proportion (sample statistic) watching the Super Bowl.

- Therefore, a sample statistic is a *random variable*.
- Therefore, a sample statistic has a pdf associated with it.
- The pdf of a sample statistic can be used to find the probability that the sample statistic will fall into specified intervals when a new sample is taken.

Sampling Distribution Definition

Statistics as Random Variables

Each new sample taken →
value of the sample statistic will change.

The distribution of possible values of a statistic for repeated samples of the same size from a population is called the **sampling distribution** of the statistic.

More formal definition: A sample statistic is a random variable. The probability density function (pdf) of a sample statistic is called the **sampling distribution** for that statistic.

Sampling Distribution for a Sample Proportion

Let p = population proportion of interest
or binomial probability of success.

Let $\hat{p}$ = sample proportion or proportion of successes.

If numerous random samples or repetitions of the same size n are taken, the distribution of possible values of $\hat{p}$ is **approximately** a **normal** curve distribution with

- **Mean** = p
- **Standard deviation** = s.d.($\hat{p}$) = $\sqrt{\dfrac{p(1-p)}{n}}$

This approximate distribution is **sampling distribution of** $\hat{p}$ **.**

Conditions needed for the sampling distribution to be approx. normal

The sampling distribution for $\hat{p}$ can be applied in ***two situations***:

Situation 1: A random sample is taken from a population.

Situation 2: A binomial experiment is repeated numerous times.

In each situation, ***three conditions*** must be met:

1: *The Physical Situation*
There is an actual population or repeatable situation.

2: *Data Collection*
A random sample is obtained or the situation repeated many times.

3: *The Size of the Sample or Number of Trials*
The size of the sample or number of repetitions is relatively large, *np* and *np(1-p)* must be at least 5 and preferable at least 10.

Motivation via a Familiar Example

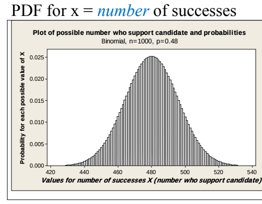
Suppose 48% (p = 0.48) of a *population* supports a candidate.

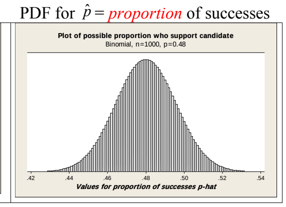
- In a poll of 1000 randomly selected people, what do we expect to get for the *sample proportion* who support the candidate in the poll?
- In the last few lectures, we looked at the pdf for X = the *number* who support the candidate. X was binomial, and also X was approx. normal with mean = 480 and s.d. = 15.8.
- Now let's look at the pdf for the *proportion* who do.

  $\hat{p} = \frac{X}{n} =$ where X is a binomial random variable.
- We have seen picture of possible values of X.

  Divide all values by n to get picture for possible $\hat{p}$.

PDF for x = *number* of successes PDF for $\hat{p}$ = *proportion* of successes





What's different and what's the same about these two pictures?

Everything is the same except the values on the x-axis!
On the left, values are *numbers* 0, 1, 2, to 1000
On the right, values are *proportions* 0, 1/1000, 2/1000, to 1.

Recall: Normal approximation for the binomial

For a *binomial* random variable X with parameters n and p with np and $n(1-p)$ at least 5 each:

- X is *approximately* a *normal* random variable with:

  mean $\mu = np$ standard deviation $\sigma = \sqrt{np(1-p)}$

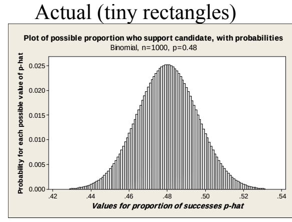
NOW: Divide everything by n to get similar result for $\hat{p} = \frac{X}{n}$

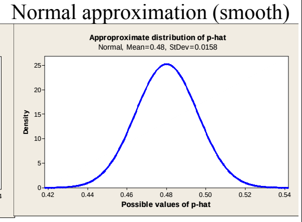
- $\hat{p}$ is *approximately* a *normal* random variable with:

  mean $\mu = p$ standard deviation $\sigma = \sqrt{\frac{p(1-p)}{n}}$

So, we can find probabilities that $\hat{p}$ will be in specific intervals if we know n and p.

Actual (tiny rectangles) Normal approximation (smooth)





For example, to find the probability that $\hat{p}$ is at least 0.50:
Could add up areas of rectangles from .501, .502, …, 1000
but that would be too much work! P($\hat{p}$ > 0.50)

$$\approx P(z > \frac{0.50 - .48}{.0158}) = P(z > 1.267) = .103$$

Sampling Distribution for a Sample Proportion, Revisited

Let p = population proportion of interest
or binomial probability of success.

Let $\hat{p}$ = sample proportion or proportion of successes.

If numerous random samples or repetitions of the same size n are taken, the distribution of possible values of $\hat{p}$ is **approximately** a **normal** curve distribution with

- **Mean** = p
- **Standard deviation** = $\text{s.d.}(\hat{p}) = \sqrt{\dfrac{p(1-p)}{n}}$

This approximate distribution is **sampling distribution of** $\hat{p}$ **.**

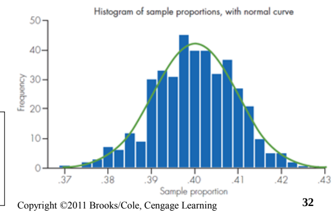
31

Example 9.4 *Possible Sample Proportions Favoring a Candidate*

Suppose 40% all voters favor Candidate C. Pollsters take a sample of n = 2400 voters. Rule states the sample proportion who favor X will have approximately a normal distribution with

$$\text{mean} = p = 0.4 \text{ and s.d.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{2400}} = 0.01$$

Histogram at right shows sample proportions resulting from simulating this situation 400 times.



Empirical Rule: Expect
68% from .39 to .41
95% from .38 to .42
99.7% from .37 to .43

32

A Final Dilemma and What to Do

In practice, we don't know the true population proportion p, so we cannot compute the **standard deviation** of $\hat{p}$,

$$\text{s.d.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \ .$$

Replacing p with $\hat{p}$ in the standard deviation expression gives us an estimate that is called the **standard error of** $\hat{p}$.

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ .$$

The *standard error* is an excellent approximation for the *standard deviation*. We will use it to find *confidence intervals,* but will not need it for sampling distribution or hypothesis tests because we assume a specific value for p in those cases.

33

CI Estimate of the Population Proportion from a Single Sample Proportion

CBS Poll taken this month asked "*In general, do you think gun control laws should be made more strict, less strict, or kept as they are now?*

Poll based on n = 1,148 adults, 53% said "more strict."

Population parameter is p = proportion of *population* that thinks they should be more strict.

Sample statistic is $\hat{p}$ = .53

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.53(.47)}{1148}} = .015$$

If $\hat{p}$ = 0.53 and n = 1148, then the standard error is 0.015.
Sample $\hat{p}$ = .53 is 95% certain to be within 2 standard errors of population p, so p is probably between .50 and .56.

34

Preparing for the Rest of Chapter 9

For all 5 situations we are considering, the sampling distribution of the sample statistic:

- Is approximately normal
- Has mean = the corresponding population parameter
- Has standard deviation that involves the population parameter(s) and thus can't be known without it (them)
- Has standard error that doesn't involve the population parameters and is used to estimate the standard deviation.
- Has standard deviation (and standard error) that get smaller as the sample size(s) n get larger.

Summary table on page 353 will help you with these!

35