This week: Chapter 9 (will do 9.6 to 9.8 later, with Chap. 11)
Understanding Sampling Distributions:
Statistics as Random Variables

ANNOUNCEMENTS:

- Shandong Min will give the lecture on Friday.
- See website for different office hours Fri, Mon, Tues.
- New use of clickers: to test for understanding. I will give many more clicker questions, and randomly five to count for credit each week.
- Homework from today and Friday is due *Monday*, Nov 8. Homework to be assigned Monday is not due.
- Midterm in one week. You are allowed *two* sheets of notes.

HOMEWORK: Due *Mon* 11/8, Chapter 9: #15, 25, 37, 44

Chapters 9 to 13: Statistical Inference
See picture drawn on board.

Five situations we will cover for the rest of this quarter:

Table 9.1 Population Parameters and Sample Statistics for the Big Five Scenarios

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	$\overline{x}$
Population mean of paired differences (dependent)	μ_d	$\overline{d}$
Difference in two population means (independent)	$\mu_1 - \mu_2$	$\overline{x}_1 - \overline{x}_2$

For each parameter we will:

- Learn how to find a confidence interval for its true value
- Test hypotheses about its true value

EXAMPLES OF EACH OF THE 5 SITUATIONS

One proportion: Binomial situation with n and p

Question: What proportion of households watched Dancing with the Stars the week of Oct 18? Get a confidence interval.

Population parameter:
p = proportion of the *population* of all US households that watched it.

Sample statistic:
Nielsen ratings measure n = 25,000 households.
X = *number* in *sample* who watched the show = 3,075.

$$\hat{p} = \frac{X}{n} = \frac{3,075}{25,000} = .123 = \textit{proportion of sample who watched}$$

This is called "p-hat."

Difference in two proportions: Compare two population proportions using *independent* samples of size n_1 and n_2.

Question: What is the difference in the proportion of smokers who would quit if wearing a nicotine patch versus placebo?
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 who would quit if everyone were to use each type of patch (nic.-plac.)

Sample statistic:
Difference in the proportions in the *sample* who did quit

$$\hat{p}_1 - \hat{p}_2 = .46 - .20 = .26$$

This is read as "p-one-hat minus p-two-hat"
Note that the parameter and statistic can range from −1 to +1.

One mean: Population mean for a quantitative variable.

Question: An airline would like to know the average weight of checked luggage per passenger, for fuel calculations.
Get a confidence interval for the population mean.
There is *no* logical value to test, so we would not do a test.

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

Sample statistic:
Collect a sample of *n* observations. For instance, suppose they sample 100 passengers and find the mean is 30 pounds.
$\bar{x} = 30 =$ the mean for the *sample* of 100 passengers

Remember this is read as "x-bar"

Mean for paired differences: Population mean for the *difference* in two quantitative measurements in a matched pairs situation.

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

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 was 9 IQ points.
$\bar{d} = 9 =$ the mean difference for the *sample* of 36 students
Read as "d-bar."

Difference in two means: Comparing two population means when *independent* samples of size n_1 and n_2 are available.

Question: What is the difference in mean IQ of 4-year-old children for the population of mothers who smoked during pregnancy and the population who did not?
Get a confidence interval for the difference.
Test to see if the difference is stat signif. different from 0.

Population parameter:
$\mu_1 - \mu_2$ = difference in the means for the two *populations*

Sample statistic: Based on a study done at Cornell, the difference in means for two *samples* was 9 IQ points.
$\bar{x}_1 - \bar{x}_2 =$ difference in the means for the two *samples* $= 9$

Read as "x-bar-one minus x-bar-two."

GOAL: Estimate and test *parameters* based on *statistics*.
Get confidence intervals and do hypothesis tests

SOME LOGICAL NOTES:

1. Assuming the sample is representative of the population, the *sample statistic* should represent the *population parameter* fairly well. (Better for larger samples.)

2. But the sample statistic will have some error associated with it, i.e. it won't equal the parameter exactly. Recall the "margin of error" from Chapter 3!

3. If 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*, i.e. they will have a *distribution*. It is a *pdf* for the statistic. It is called a **sampling distribution** for the statistic.

RATIONALE AND DEFINITION FOR SAMPLING DISTRIBUTIONS

Rationale:

- When a sample is taken from a population the resulting numbers are the outcome of a *random circumstance*.

Dancing with the Stars example:
A random circumstance is taking a random sample of 25,000 households with TVs. The resulting number is the *proportion of those households that were watching Dancing with the Stars that week = .123 (or 12.3%)*

- Remember that 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 would have had a different sample proportion (sample statistic) watching the show.

- 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.

Definition:
The pdf of a sample statistic is called the **sampling distribution** for that statistic.

Example: The *random variable* is $\hat{p}$ = sample proportion = sample statistic.

The pdf of $\hat{p}$ will be defined next. It is the distribution of *possible* sample proportions in this scenario.

We already know the pdf for X = *number* of households out of 25,000 that are watching the show. It is *binomial* with n = 25,000 and p = true proportion of households in US that watched.

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* $\hat{p}$ 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.

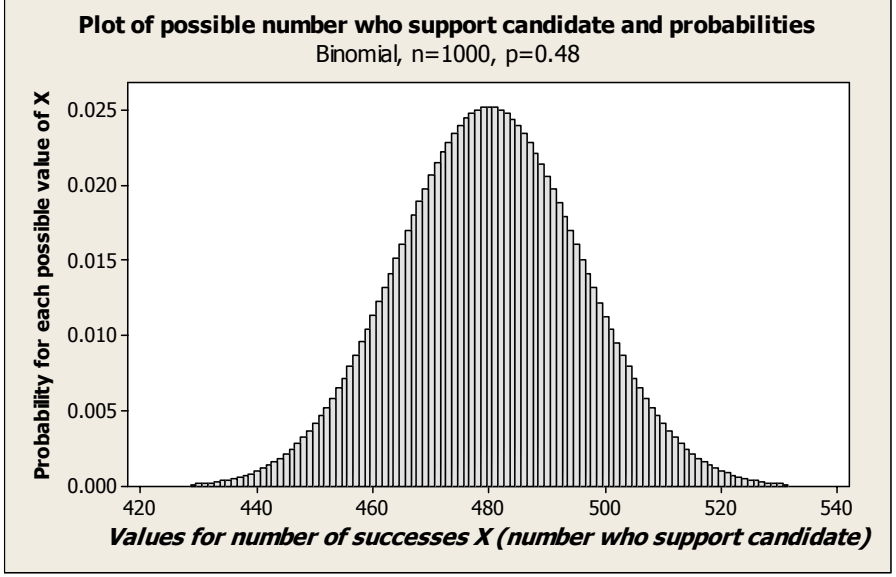
Now let's look at the pdf for the *proportion* who do.

$\hat{p} = \frac{X}{n}$ where X is a binomial random variable.

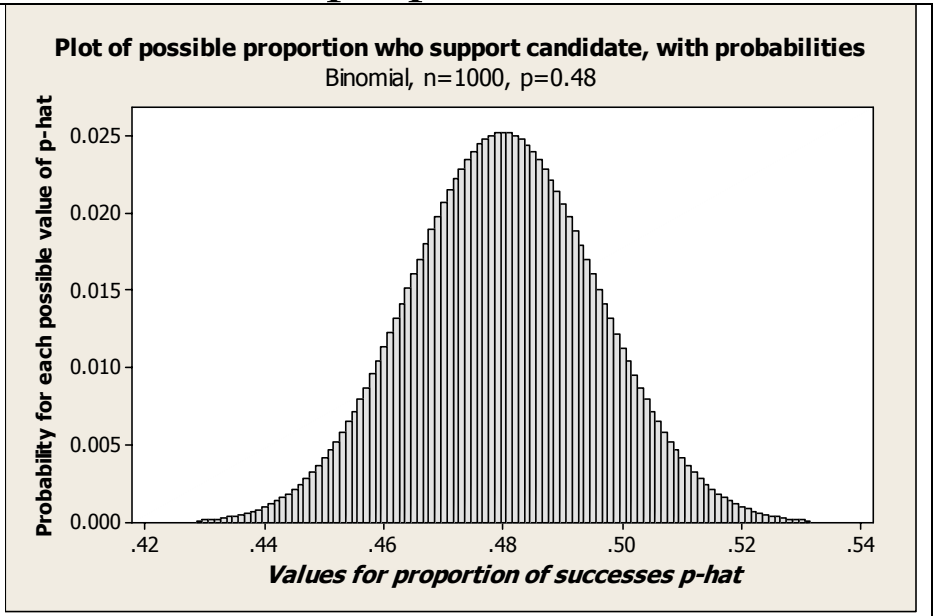
We have seen picture of possible values of X.

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 the 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)

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

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

The Sampling Distribution for a Sample Proportion $\hat{p}$

1. <u>The physical situation</u>: binomial.

   Actual population with fixed proportion w/trait or opinion (e.g. polls, TV ratings, etc.)

   OR

   A repeatable situation with fixed probability of a certain outcome (e.g. birth is a boy, probability of heart attack if one takes aspirin)

2. <u>The Experiment</u>

   Random sample of n from the population, $\hat{p}$ = proportion w/trait

   OR

   Repeat situation n times, $\hat{p}$ = proportion with specified outcome

3. <u>Sample size requirement</u>:

   In either case, must have np and $n(1-p)$ at least 5, prefer at least 10.

Assuming the above conditions are met, the distribution of *possible* values of $\hat{p}$ is approximately normal with:

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

The resulting normal distribution is called the ***sampling distribution of*** $\hat{p}$

<u>Notation</u>:

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

But suppose p is unknown (which is will be if we are estimating it!) Then instead we approximate the s.d. using

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

= *estimate* of the standard deviation of $\hat{p}$

This result is also called "the normal curve approximation rule for sample proportions"

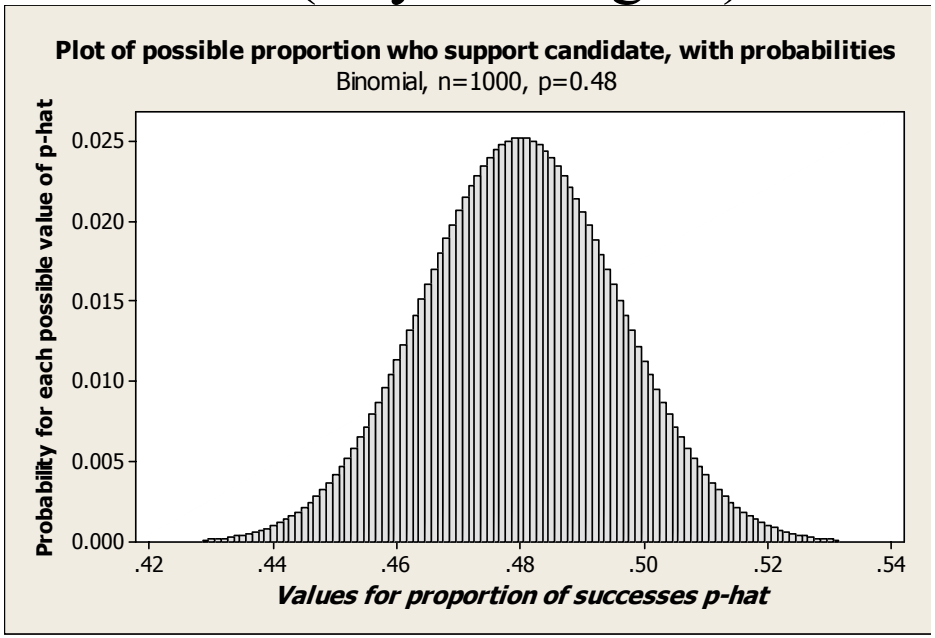
For the poll example:

Poll of $n = 1000$ people, where the *true population proportion* $p = 0.48$.

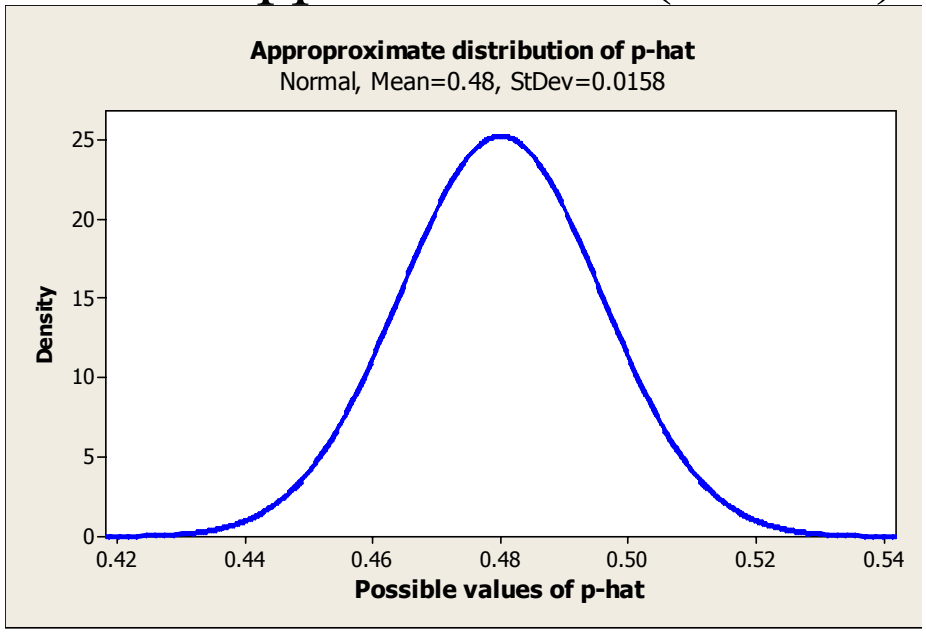
The distribution of *possible* values of $\hat{p}$ is approximately normal with

mean $\mu = p = 0.48$ and s.d. $\sigma = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.48(1-.48)}{1000}} = 0.0158$

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!

$P(\hat{p} > 0.50)$

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

Going back to the "Big Picture"

- The *sampling distribution* for $\hat{p}$ describes the distribution of possibilities for it if we were to take millions of samples of size *n* and compute $\hat{p}$ each time. It tells us what ranges we can expect $\hat{p}$ to fall in, and with what probability.

- To find the sampling distribution, we would need to know the true value of the parameter p.

- In practice, we don't know the true value of the parameter p. In fact the whole point of statistical inference is to estimate the parameter, or test for possible values of it.

- BUT, the standard deviation (or standard error) of the sampling distribution tells us how far the sample statistic is likely to fall from the parameter p, even if we don't know what that value of p is.

- For example, in our poll of $n = 1000$, we know that the standard deviation of $\hat{p}$ is about .0158 (or .016). So, (from the Empirical Rule) we know that for approximately 68% of all samples $\hat{p}$ will be within one standard deviation = .016 of the true parameter p. We can use that to estimate p!

- For instance, if $\hat{p}$ is 0.45, we can be 68% certain that the true p is somewhere in the range of 0.45 ± .016 or between 0.434 and 0.466.

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 pages 382-383 will help you with these!

New Example

In 2005, according to the Census Bureau, 67% of all children in the United States were living with 2 parents. (Includes step-parents and adoptive parents, but not foster parents.)

In our class, there are about 180 of you who participate in clicker questions. Are you a representative sample for this question? If so, what should we expect the class proportion to be?

$n = 180$

$p = .67$

The sampling distribution of $\hat{p}$ is approximately normal with mean = .67 and standard deviation = $\sqrt{\frac{(.67)(.33)}{180}} = .035$

Clicker question (not for credit, answers anonymous)

In 2005, were you living with 2 parents? (Step parents and adoptive parents count, but foster parents do not.)

A. Yes B. No