

Announcements:

Next Monday, Nov 9, there will be 5 clicker questions because Wed is a holiday, and Fri will be guest lecture by Yan He. So, all of that week's clicker questions will be on Monday. Most of them will be review.

Homework:

9.51, 9.87



Chapter 9,
Sections 4, 5, 9

Sampling Distributions for Proportions: One proportion or difference in two

Copyright ©2004 Brooks/Cole, a division of Thomson Learning, Inc., updated by Jessica Utts, Nov 2009

Understanding Dissimilarity Among Samples

Key: Need to understand what kind of dissimilarity we should expect to see in various samples from the same population.

- Suppose knew most samples were likely to provide an answer that is within 10% of the population answer.
- Then would also know the population answer should be within 10% of whatever our specific sample gave.
- => Have a good guess about the *population value* based on just *one sample value*.

Copyright ©2004 Brooks/Cole, a division of Thomson Learning, Inc., updated by Jessica Utts, Nov 2009

3

Statistics and 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 .*

Copyright ©2004 Brooks/Cole, a division of Thomson Learning, Inc., updated by Jessica Utts, Nov 2009

4

Sampling Distributions

Each new sample taken =>
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.*

Many statistics of interest have sampling distributions that are *approximately normal* distributions

9.4 Sampling Distribution for One Sample Proportion

- Suppose (unknown to us) **40% of a population carry the gene** for a disease, ($p = 0.40$).
- We will take a **random sample of 25** people from this population and count $X =$ **number with gene**.
- Although we *expect* on average to find 10 people (40%) with the gene, we know the number will *vary* for different samples of $n = 25$.
- In this case, X is a **binomial random variable** with $n = 25$ and $p = 0.4$.

Many Possible Samples

Four possible random samples of 25 people:

Sample 1: $X = 12$, proportion with gene $= 12/25 = 0.48$ or 48%.

Sample 2: $X = 9$, proportion with gene $= 9/25 = 0.36$ or 36%.

Sample 3: $X = 10$, proportion with gene $= 10/25 = 0.40$ or 40%.

Sample 4: $X = 7$, proportion with gene $= 7/25 = 0.28$ or 28%.

Note:

- Each sample gave a different answer, which did not always match the population value of 40%.
- Although we cannot determine whether one sample will accurately reflect the population, statisticians have determined how far off from the truth they will be, on average.

The Normal Curve Approximation Rule for Sample Proportions

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

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

The Normal Curve Approximation Rule for Sample Proportions

Normal Approximation Rule 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:

Condition 1: The Physical Situation

There is an actual population or repeatable situation.

Condition 2: Data Collection

A random sample is obtained or situation repeated many times.

Condition 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.

Copyright ©2004 Brooks/Cole, a division of Thomson Learning, Inc., updated by Jessica Utts, Nov 2009

9

Examples for which Rule Applies

- **Polls:** to estimate proportion who favor a candidate; units = all voters.
- **Television Ratings:** to estimate proportion of households watching TV program; units = all households with TV.
- **Consumer Preferences:** to estimate proportion of consumers who prefer new recipe compared with old; units = all consumers.
- **Testing ESP:** to estimate probability a person can successfully guess which of 5 symbols on a hidden card; repeatable situation = a guess.

Copyright ©2004 Brooks/Cole, a division of Thomson Learning, Inc., updated by Jessica Utts, Nov 2009

10

Example: Belief in evolution

Gallup Poll. Feb. 6-7, 2009. $N=1,018$ adults nationwide. Margin of error given as $\pm 3\%$.

"Now, thinking about another historical figure: Can you tell me with which scientific theory Charles Darwin is associated?" Options rotated

Correct response (Evolution, natural selection, etc.)	55%
Incorrect response	10%
Unsure/don't know	34%
No answer	1%

Copyright ©2004 Brooks/Cole, a division of Thomson Learning, Inc., updated by Jessica Utts, Nov 2009

11

Example, continued

"In fact, Charles Darwin is noted for developing the theory of evolution. Do you, personally, believe in the theory of evolution, do you not believe in evolution, or don't you have an opinion either way?"

(Poll based on $n = 1018$ adults)

Believe in evolution	39%
Do not believe in evolution	25%
No opinion either way	36%
No answer	1%

Copyright ©2004 Brooks/Cole, a division of Thomson Learning, Inc., updated by Jessica Utts, Nov 2009

12

Example, continued

Let p = *population proportion* who believe in evolution.

Our observed $\hat{p} = .39$, from sample of 1018.

Based on samples of $n = 1018$, \hat{p} comes from a distribution of possible values, which is approximately normal with mean $\mu = p$ and standard deviation $\sigma = \sqrt{\frac{p(1-p)}{1018}}$

Based on this, can we use \hat{p} to estimate p ?

Estimating the Population Proportion from a Single Sample Proportion

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

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

In practice, we only take one random sample, so we only have one sample proportion \hat{p} . Replacing p with \hat{p} in the standard deviation expression gives us an estimate that is called the **standard error of \hat{p}** .

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

If $\hat{p} = 0.39$ and $n = 1018$, then the standard error is 0.0153. So the *true proportion* who believe in evolution is **almost surely** between $0.39 - 3(0.0153) = 0.344$ and $0.39 + 3(0.0153) = 0.436$.

Independent Samples

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

- **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.

Parameter 2: Difference in two population proportions, based on independent samples

Example research questions:

- How much difference is there between the proportions that would quit smoking if wearing a nicotine patch versus if wearing a placebo patch?
- How much difference is there in the proportion of UCI students and UC Davis students who are an only child?
- Were the proportions believing in evolution the same in 1994 and 2005?

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.

General format for all sampling distributions in Chapter 9

The sampling distribution of the sample estimate (the sample statistic) is:

- Approximately normal
- Mean = population parameter
- Standard deviation is called the *standard deviation* of _____, where the blank is filled in with the name of the statistic (p-hat, x-bar, etc.)
- The estimated standard deviation is called the *standard error of* _____.

Sampling distribution for the difference in two proportions $\hat{p}_1 - \hat{p}_2$

- Approximately normal
- Mean is $p_1 - p_2$ = true difference in the *population* proportions
- Standard deviation of $\hat{p}_1 - \hat{p}_2$ is

$$s.d.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Standard Error of the Difference Between Two Sample Proportions

$$s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Are more UCI than UCD children an only child?

$n_1 = n_2 = 173$ from each campus (amazing, it was same!)

UCI: 19 of the 173 students were an only child = $\hat{p}_1 = .110$

UCD: 14 of the 173 students were an only child = $\hat{p}_2 = .081$

So, $\hat{p}_1 - \hat{p}_2 = .110 - .081 = .029$

and $s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{.11(1-.11)}{173} + \frac{.08(1-.08)}{173}} = .031$

Suppose population proportions are the same, so true difference $p_1 - p_2 = 0$

Then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is:

- Approximately normal
- Mean = population parameter = 0
- The estimated standard deviation is .031
- *Observed* difference of .029 is $z = .94$ standard errors above the mean of 0.
- See picture on board; area above .029 = .1736

Standardized Statistics for sampling distributions

Recall the general form for standardizing a random variable x when it has a normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

For all 5 parameters we will consider, we can find where our *observed* sample statistic falls if we hypothesize a specific number for the population parameter:

$$z = \frac{\text{sample statistic} - \text{population parameter}}{\text{s.d.}(\text{sample statistic})}$$

Example 9.15 *Unpopular TV Shows*

Networks cancel shows with low ratings. Ratings based on random sample of households, using the sample proportion \hat{p} watching show as estimate of population proportion p . If $p < 0.20$, show will be cancelled.

Suppose in a random sample of 1600 households, 288 are watching (sample proportion is $288/1600 = 0.18$). Is it likely to see $\hat{p} = 0.18$ even if p were 0.20 (or higher)?

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.18 - 0.20}{\sqrt{\frac{0.20(1-0.20)}{1600}}} = -2.00$$

The sample proportion of 0.18 is about 2 standard deviations *below* the mean of 0.20.

Case Study 9.1 *Do Americans Really Vote When They Say They Do?*

Election of 1994:

- *Time Magazine Poll*: $n = 800$ adults (two days after election), **56% reported that they had voted.**
- Info from Committee for the Study of the American Electorate: **only 39% of American adults had voted.**

If $p = 0.39$ then sample proportions for samples of size $n = 800$ should vary approximately normally with ...

$$\text{mean} = p = 0.39 \text{ and s.d.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.39(1-0.39)}{800}} = 0.017$$

Case Study 9.1 *Do Americans Really Vote When They Say They Do?*

If respondents were telling the truth, the sample percent should be no higher than $39\% + 3(1.7\%) = 44.1\%$, nowhere near the reported percentage of 56%.

If 39% of the population voted, the **standardized score** for the reported value of 56% is ...

$$z = \frac{0.56 - 0.39}{0.017} = 10.0$$

It is virtually **impossible** to obtain a standardized score of 10.