Math 183
Statistical Methods

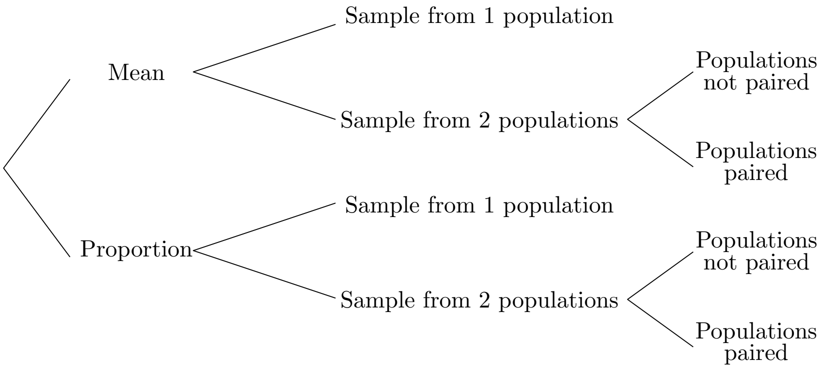
Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 6 (beginning)

- Inference for proportions
- Conditions for inferring proportions
- Estimate the sample size needed for a given margin or error

Statistics in the Large



First Things First

The statistical infrastructure around inference for proportions is nearly identical to that for means.

Reason: Proportions are a type of mean:

$$\hat{p} = \frac{\text{number with feature}}{\text{total number studied}} = \frac{0+1+1+0+1+\cdots+1}{n} = \text{ an average,}$$

where you assigned each Success to "1", and each Failure to "0".

Notation:

p:	in population	(like μ)
$\hat{p}$:	in sample	(like $\bar{x}$)

Big Question: What does the sampling distribution look like when we work with proportions?

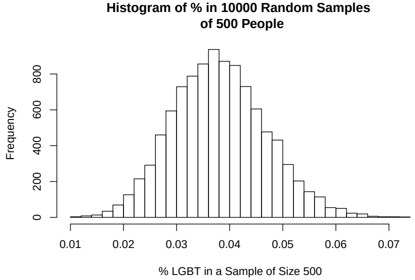
The Sampling Distribution for Proportions

Population: San Diego Population

Parameter: p_{LGBT}

(actually is 3.9%, the goal is to find this percentage. Pretend we don't know it.)

Sample: 500 people randomly selected.



Sampling distribution:
The histogram of a statistic for many different samples.

It turns out that:

Shape: Normal Distribution:

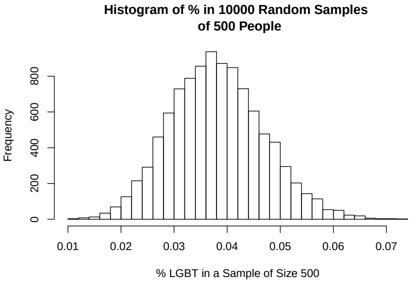
$Center_{model}$: p_{LGBT}

SE_{model}: $\sqrt{\dfrac{p_{LGBT} \times q_{LGBT}}{n}}$

Our First C.I. for Proportions

You ask 200 random UCSD students if they identify as LGBT and 8 say they do. Find a 90% C.I. for the true proportion of UCSD students that self-identify as LGBT.

$$\hat{p} = \frac{8}{200} = 0.04.$$



$$\text{C.I.: } \hat{p} \pm z^* \times SE$$

$$= 4\% \pm 1.645 \times \sqrt{\frac{pq}{n}}$$

Approximate p and q by $\hat{p}$ and $\hat{q}$

$$\text{C.I.: } \simeq 4\% \pm 1.645 \times \sqrt{\frac{4 \times 96}{200}}$$

$$= (1.72\%, 6.28\%).$$

You can work with decimals instead of percentages:

$$\text{C.I.: } \simeq 0.04 \pm 1.645 \times \sqrt{\frac{0.04 \times 0.96}{200}} = (0.172, 0.628).$$

But Wait! What About Those Pesky Conditions?

The sampling distribution for a proportion will only be approximately $N\left(p, \sqrt{\frac{pq}{n}}\right)$ when two conditions are met:

- The sample is independent.
  (usually met through the Randomization and the $< 10\%$ Conditions).
- The sample has at least 10 successes and failures.
  (this ensures the idea you are studying has appeared enough times to be given a fair estimate).

Note: Our last example had 8 "Successes" and 192 "Failures", so we should <u>not</u> have built the C.I.

Your Turn!

In a recent Gallup poll, 810 out of 1012 randomly-selected U.S. adults favored creating higher fuel-efficiency standards for automobiles. What is the parameter of this study?

1. 810/1012
2. $810/1012 \times 100$
3. The proportion of Americans who favor higher fuel-efficiency standards for automobiles
4. The proportion of U.S. adults who favor higher fuel-efficiency standards for automobiles

Answer: 4.

Your Turn!

Which of the following calculations correctly finds the 70% C.I. for the parameter. Note that $810/1012 = 0.8$ and $z^* = 1.04$.

1. $.8 \pm 1.04 \times \sqrt{\frac{80 \times 20}{1012}}$
2. $80 \pm 1.04 \times \sqrt{\frac{.8 \times .2}{1012}}$
3. $.8 \pm 1.04 \times \sqrt{\frac{.8 \times .2}{1012}}$
4. $80 \pm 1.04 \times \sqrt{\frac{80 \times 20}{1012}}$

Answer: 3. and 4. (in decimals and % respectively)

Don't mix % with decimals
(as we see in 1. and 2.)

Your Turn!

Are we allowed to do inference given our set up?
(Recall that 810 out of 1012 randomly-selected U.S. adults favored creating higher fuel-efficiency standards for automobiles)

1. No, we don't meet the Nearly Normal condition
2. No, we don't meet the Independence condition
3. No, we don't meet the 10 Successes/10 Failures conditions
4. Yes

Answer: 4. Yes!
Here, we have 810 successes and 202 failures. Adults were chosen randomly with <10% of U.S. adults sampled.

Note that there is no (explicit) "Nearly Normal" condition for proportion inference.

Your Turn!

We are 70% confident that the percent of U.S. adults who favor higher fuel-efficiency standards is between 78.7% and 81.3%.
What does the % on the numbers 78.7 and 81.3 refer to?

1. What fraction of people in our sample do/don't support new standards
2. What fraction of U.S. adults do/don't support new standards
3. The probability that the true parameter value lies in this interval
4. What fraction of C.I.'s built in this way capture the true parameter value

Answer: 2.

Your Turn!

We are 70% confident that the percent of U.S. adults who favor higher fuel-efficiency standards is between 78.7% and 81.3%.
What does the % on the numbers 70 refer to?

1. What fraction of people in our sample do/don't support new standards
2. What fraction of U.S. adults do/don't support new standards
3. The probability that the true parameter value lies in this interval
4. What fraction of C.I.'s built in this way capture the true parameter value

Answer: 4.

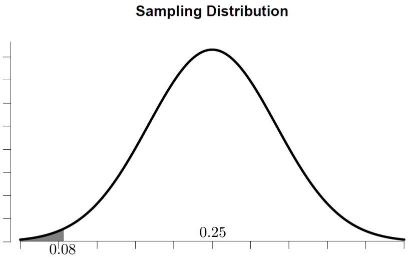
Hypothesis Testing for a Proportion

"Death Postponement": The theory that people will somehow delay their death until after an important life event (e.g., birthday, wedding of a child, etc...).

Let p be the percentage of people that die in the three-month window before their birthdays.

H_0: Death postponement is nonsense: $p = 1/4$.
H_A: Death postponement is real: $p < 1/4$.

Researchers looked at 747 deaths in Salt Lake City and found 60 deaths occurred in the three-month window before a person's birthday. *(Newsweek, 3/6/1978)*



Assuming H_0, the universe should give us sample from

$$N\left(p, \sqrt{\frac{pq}{n}}\right) \simeq N\left(0.25, \sqrt{\frac{0.25 \times 0.75}{747}}\right)$$
$$= N(0.25, 0.0158)$$

Our data gives $\hat{p} = \dfrac{60}{747} \simeq 0.08$.

```
> pnorm(0.08,0.25,0.0158)
[1] 2.673155e-27
```

Shading the area to the left of 0.08 gives a p-value of $2.67 \cdot 10^{-27} \ll 0.05$.

We reject H_0 in favor of H_A.

Choosing Sample Size for Proportions

You have been tasked with measuring Donald Trump's approval rating. You have to report a 95% C.I. with a margin of error around 2%. Online, you see his recent rating is around 39%.

Recall that a C.I. has the form $\hat{p} \pm z^* \times SE = \hat{p} \pm$ margin error.

We need $1.96 \times \sqrt{\dfrac{pq}{n}} = 2$.

Our best guess at p is the recent data of $\hat{p} = 39\%$. This yields

$$1.96 \times \sqrt{\frac{39 \times 61}{n}} = 2, \Leftrightarrow \sqrt{n} = \frac{1.96 \times \sqrt{39 \times 61}}{2}$$
$$\Leftrightarrow n = 2284.792$$

We need to ask $n = 2285$ people.

Working With Even Less Info

A new political candidate has joined the race and you must approximate the percentage of people that plan to vote for her. How large a sample must you draw if you 95% C.I. has a margin of error of 3%?

Recall that C.I. has the for $\hat{p} \pm z^* \times SE$

We need $\sqrt{n} = \dfrac{1.96\sqrt{pq}}{0.03}$, but we know nothing about p since the candidate is brand new to polling.

To make sure n is large enough, we choose the worst case for p.
(The value that makes $pq = p(1-p)$ maximal).
This value is $p = 0.5$, which yields

$$\sqrt{n} = \frac{1.96\sqrt{0.5 \times 0.5}}{0.03},$$

so that $n = 1068$.

Does Extra-Sensory Perception Exist?

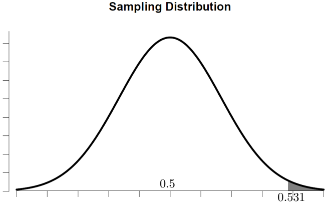
In a 2011 article, Daryl Benn claims to have found evidence for Extra-Sensory Perception (ESP). Participants had to choose which of two curtains on a computer screen had an erotic picture behind it. They were able to do this 829 out of 1560 times.
Do these data suggest the ability to perceive erotica beyond what we expect from random chance?

H_0: ESP does not exist with erotic pictures.
H_A: ESP allows for better-than-random perception of erotic imagery.

Let p be the percentage of erotic pictures identified by those claiming to have ESP. We have

$$H_0\colon\ p = 0.5$$
$$H_A\colon\ p > 0.5$$

In this study, $\hat{p} = \dfrac{829}{1560} = 0.531$.



Under H_0, we are on the sampling distribution

$$N\left(0.5, \sqrt{\frac{0.5 \times 0.5}{1500}}\right) \simeq N(0.5, 0.01266).$$

```
> pnorm(0.531, 0.5, 0.01266, lower.tail=F)
[1] 0.007169492
```

Since $p = 0.007 < 0.05$, we reject H_0. These data are strong enough to move to the alternative saying that ESP exists!!
Might be a false positive (Type I error)? Need to be reproduced to be validated.

Such a study is part of the field of *Parapsychology*. For more info on such studies, see

a conference of Chris French

Remark: C. French and D. Bem aren't best friends... (link)