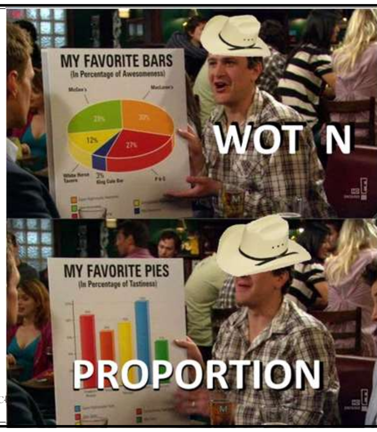


1

SOC424 – Statistics w/ Dr. Ellis Godard

Estimating Proportions

Inline for Today

- **Pause**
  - Breathe :)
  - Back-patting
  - Questions?
- **Summary**
  - Descriptions vary by level of measurement
  - So do inferences (estimates about populations)
  - Instead of point estimates, we want to estimate intervals.
  - We have some % confidence that the pop. parameter is in it
- **Moving on**
  - *Nothing* last lecture was really *new;* you'd seen or done all of it
  - Three (3) ideas today are, marked w/ lightning bolts:
    - See slides 12, 16, & 28

Outline for Today

- **Reminders about Missings & CI interps**
- **Interpreting Confidence Intervals**
- **Confidence About Proportions**
  - Proportions as Interval Variables
  - Confidence Intervals
  - Confidence Coefficients
- **Examples**
- **Lab Exercise** (2 parts, two, 1 plus another)
  - You must do BOTH for credit
  - Both involve a lil' math, YAY! ☺

Last Reminders about Missing Data

- **How do I recode to take care of missing values.**
  - You don't. EVER. Do NOT use ~~TRANSFORM - RECODE~~ for missings.
  - **I gave 7 solutions for missings – recoding *wasn't* one!**
  - Recoding *doesn't address* missings. See slide 6 from 9/30.
  - Systematically waning support: see slide 26, Indices lecture!
- **Cases w/ missing values don't require any action**
  - They're already identified as missing – that's how they're counted
- **Invalid Values counted as valid - that's trouble!**
  - See the list of questions to ask yourself, in the Indices lecture.
  - Look in the FAQ for "I don't understand missin values".
  - Look at the "Missing" column in the Variable View.
  - DO NOT RECODE. That's not even *relevant*.

Prose for Confidence Intervals

- **Numbers alone aren't meaningful**
  - *Not* "0.75 and 1.53"
    - It's the middle (the "interval") that matters, not just the bounds
  - *Not* just "0.75 to 1.53"
    - What *is* that?
  - *Not* just "confidence interval is 0.75 to 1.53"
    - What does that *mean?*
- **Use terms anyone could understand**
  - In general:
  - "We are 95% confident that the population parameter falls within this range."
  - For that specific example:
    "We're 95% confident that, in the population at large, the average number of reality shows that each individual has seen (from that list of 6) falls somewhere between 0.75 and 1.53."

Conf. Int. for Proportions?

- Mean and standard deviation can only be computed for *interval* variables (e.g. age in years, income in dollars – NOT in ranges).
- Conf. Interval for interval variables use both:

$$\bar{Y} \pm (z^* \hat{\sigma}_{\bar{Y}}) = \bar{Y} \pm (z^* \frac{s}{\sqrt{n}})$$

- What do we do if our variables are either *ordinal* or *nominal* (a.k.a. *categorical*)?
  - Central Tendency: mode or *median* – **no mean!**
  - Dispersion: *range* & *variation ratio* – **no std dev!**

Only Intervals have Means

- **Interval data has means**
  - Numbers (counts) can have an arithmetic average
  - Z scores count # of std deviations
  - Confidence intervals are ranges, of width 2*z, around mean
- **Nominals and Ordinals <u>don't have means</u>**
  - Can't arithmetically average
  - Can't count # of std deviations
    - Measures distance *from* mean
    - Calculated *based* on the mean
  - Can't calculate CIs around a *mean* using *standard deviations*
- **But… Nominals & Ordinals *do* have proportions**
  - The % that is any value – Yes, Male, Black, Very Satisfied

Equation for a Proportion

- A proportion is a special case of a mean.
- Define Y_i=1 if the ith observation is in the category of interest (e.g. a Democrat in the first example or "disloyal" Democrat in the 2nd) and define Y_i=0 if the ith observation is not in the category. Then,

$$\hat{\pi} = p = \frac{\sum_{1}^{n} Y_i}{n}.$$

Nominal Proportion

Consider a hypothetical relative frequency distribution for the nominal variable Political Party for a sample of 500 voters:

Value	f	r.f.
Democrat	265	0.53
Republican	220	0.44
Other	15	0.03
	500	1.00

- Based on this information, we could say that the proportion of Democrats (versus Republicans and other) is 0.53, or 53%.
- Be sure to use *valid* percent!

Ordinal Proportion

Consider a hypothetical frequency distribution for strength of party loyalty (Do you support your party's position …) among a sample of 1,000 registered Democrats:

Value	f	r.f.
Most of the Time	630	0.63
Some of the Time	300	0.30
Rarely	50	0.05
Never	20	0.02
	1,000	1.00

Based on this information, we could say that the proportion of "Disloyal" Democrats (who either rarely or never support their party's position) is 0.07, or 7%.

Proportions are Interval

- They're about a value of a nominal or ordinal variable
- But the intervals between them are equal and consistent
- They range from 0% (0.0) to 100% (1.0)
  - Remember: percents represent "hidden decimals"
  - 50% = 50 per 100 = 50/100 = 0.50
- We can add them – and average them
- A *set* of proportions has a mean and standard deviation

Population vs. Sample Proportion

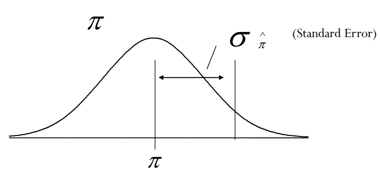
- Let π (pi) equal the proportion of the defined population classified in some specific category.
- The best (least biased and most efficient) estimate of the population proportion π is the sample proportion p(or $\hat{\pi}$, as in the text).

Means of Sample Proportions

- Proportions vary *intervally* among *samples*
  - Each sample has a proportion female, Athiest, tall, etc.
    - Sample proportion
  - Can arithmetically average those proportions
    - For example, take the % female in each and every class
    - The collection of % from every possible sample is a sampling distribution
  - There's a hypothetical std. deviation to all of them
    - The standard deviation of a sampling distribution is the standard error
  - With mean & std dev, we can calculate Zs & CIs
    - of the population mean proportion – the average of all the sample proportions, which would be the population proportion
  - So, *almost* nothing new today… ☺

Sampling Distribution of a Proportion

If we calculated the sample proportion for all possible samples of size n, the distribution of these sample proportions would be approximately normally distributed around the population proportion .



Sampling Variance for "pi"

For nominal or ordinal variables, we use proportions rather than means, and the formula for variance of proportions:

$$\sigma_\pi^{\ 2} = \frac{\pi(1-\pi)}{n}$$

For example, the variance for the percent that's female, in a class of 10 students that is 50% female, would be…

$$\sigma_\pi^{\ 2} = \frac{0.50(1-0.50)}{10} = \frac{0.50(0.50)}{10} = \frac{0.25}{10} = .025$$

"But Professor Godard, you said not to use the variance…"

Except as a step to a standard deviation…

Standard Error for a Proportion

For nominal or ordinal variables, the formula for variance of proportions:

$$\sigma_\pi^{\ 2} = \frac{\pi(1-\pi)}{n}$$

The variance = the standard deviation squared, so…

The square root of each side gives the standard deviation of proportions:

$$\sqrt{\sigma_\pi^{\ 2}} = \sigma_\pi = \sqrt{\frac{\pi(1-\pi)}{n}}$$

That standard deviation is of a *sampling distribution*, so is a *standard error*

We *estimate* that standard error by using the <u>sample</u> proportion as an estimate of the population proportion:

$$\hat{\sigma}_\pi = \sqrt{\frac{p(1-p)}{n}}$$

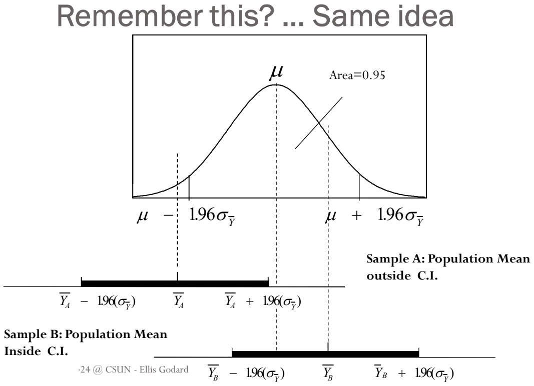
Largest possible

- When p & q are both 50%
  - .5 x .5 = .25
  - But 90% female -> 0.9(1-0.9) = 0.9(0.1) = 0.09
- Use that (.25) when don't know p & 1-p (aka q)
  - It's more conservative than any other guess
  - Larger std error makes a wider conf. interval, which gives looser claims about reality
- Review examples on p.137 in the text
  - More in next lecture…

Conf. Interval for a Proportion

C.I. for a Mean	C.I. for a Proportion
$\bar{Y} \pm z_{\alpha/2}\hat{\sigma}_{\bar{Y}}$	$p \pm z_{\alpha/2}\hat{\sigma}_p$
or	or
$\bar{Y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$	$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

Remember this? … Same idea



Prose for Proportions

- If we drew repeated samples of the same size from a population and calculated the 95 percent confidence intervals for each sample's proportion

…then…

- we would expect that the population ***proportion*** would fall in the confidence interval 95 percent of the time.

Serious Errors to Avoid

- **What you're predicting:**
  - Don't confuse mean and proportion
  - Don't confuse sample mean & population mean
    - We are **<u>100%</u>** confident what the <u>sample</u> mean is! ☹
- **Calculating the interval**
  - **Mean:** Use the standard error (measuring dispersion in the sampling distribution) not sample standard deviation (measuring the dispersion in the sample)
  - **Proportion:** Use the standard error for a proportion (which is *not* the std. dev. divided by the sqrt rt of n)
- **Where the p's & Conf. Coefficients are:**
  - Don't confuse the area in the tails with the area *between* the tails
  - Draw pictures!

What proportion of CSUN students have consumed marijuana?

- Point estimate: 14 from sample of 26 = 53.8% or .538
- A 95% confidence interval would be:

$$\pi \pm \left(z * \sqrt{\frac{p*(1-p)}{n}}\right) = .538 \pm \left(1.96 * \sqrt{\frac{.538*(1-.538)}{26}}\right)$$

$$= .538 \pm \left(1.96 * \sqrt{\frac{.2486}{26}}\right) = .538 \pm \left(1.96 * \sqrt{.009562}\right)$$

$$= .538 \pm (1.96 * .0977) = .538 \pm (.1916) = \{3463 \longrightarrow .7296\}$$

- Assuming our sample was random & unbiased, we can be 95% confident that between 34.63% & 72.96% of CSUN students have smoked pot

AP Style Guide on Polls

Do not exaggerate poll results. In particular, with pre-election polls, these are the rules for deciding when to write that the poll finds one candidate is leading another:

-- If the difference between the candidates is more than twice the **sampling error** margin*, then the poll says one candidate is leading.

-- If the difference is less than the **sampling error** margin, the poll says that the race is close, that the candidates are "about even." (Do not use the term "statistical dead heat," which is inaccurate if there is any difference between the candidates; if the poll finds the candidates are tied, say they're tied.)

-- If the difference is at least equal to the **sampling error** but no more than twice the sampling error, then one candidate can be said to be "apparently leading" or "slightly ahead" in the race.

* They mean standard error (oops)

Proportion Example from SPSS

Last lab was C.I. for the interval index MUSIC

Could also do nominal, e.g. proportion who like folk

Point estimate (*valid* %) is 65.4%, on 1044 valid cases

Interval estimate:

$$p \pm \left(z * \sqrt{\frac{p*(1-p)}{n}} \right) \quad = .654 \pm \left(1.96 * \sqrt{\frac{.654*(.346)}{1044}} \right)$$

$$= .654 \pm .0288 = \{.6251 - .6829\}$$

Interpretation: We can be 95% confident that, in the population as whole, between 62.5% and 68.3% percent like folk.

And note: Since this interval does *not* include 50%, we're 95% confident that a *majority* of the population like folk music.

Proportion Example from SPSS

But what if the point estimate was, e.g., 47.8%

Interval estimate:

$$p \pm \left(z * \sqrt{\frac{p*(1-p)}{n}} \right) \quad = .478 \pm \left(1.96 * \sqrt{\frac{.478*(.522)}{1429}} \right)$$

$$= .478 \pm .026 = \{.452 - .530\}$$

Interpretation: We can be 95% confident that, in the population as whole, between 45.2 and 53.0 percent like folk.

Note that this includes the *possibility* that more than half like folk

But we can't be *confident* that a majority do.

It's possible that *less* than a majority do – *too close to call*

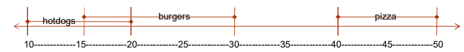
Proportion Example w/ Attendance

- What percent is attending today? – 89% (0.89)
- That's a point estimate for the average percent present across the semester, w/ each class meeting as a different sample, and the entire semester as a population
  - Each day is different – some have higher attendance, some lower
  - Some days are outlier samples (unusual) – first day, last day, bad weather, etc.
- Could calculate a confidence interval, to predict what percent attend across the entire semester (the population)
  - Standard error: square root of (.89 * (1-.89))/41
  - CI: .89 +/- 1.96 * (that ^)

Going Further w/ C.I.'s

3

- Compare them, e.g. if we estimate that:
  - 10-20% like hot dogs
  - 15-30% like burgers
  - 40-50% like pizza….



- If any overlap, can't conclude there's any difference
  - Both parameters could be same value (e.g. 19% like hotdogs & burgers)
  - Could be the reverse of what it appears (19% hotdogs, 16% burgers)
  - Rest of the CI doesn't matter (e.g. burger *mostly* above hotdog? irrelevant)
- But if the intervals *don't* overlap, can conclude parameters differ
  - We're 95% confident that pizza is more popular than either one

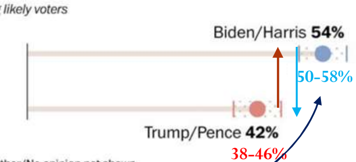
Presidential Polling Example

Wash Post, 10/11/20

Biden leads Trump among likely voters nationally

Q: If the presidential election were held today, for whom would you vote/lean toward?

Among likely voters



Note: Other/No opinion not shown.

Source: Oct. 6-9, 2020, Washington Post-ABC News poll among 725 likely voters with an error margin of +/- 4 percentage points.

Lab Hints

- **There are *two* parts, that count as *one* lab.**
  - This is not the first time that this has happened. It's the 3rd.
  - You need to at least TRY the full lab, to get credit.
  - Some of you still need to catch on ☹
- **Remember, percentages are interval**
  - Not the variables – the percentages
  - Percentages for a values, don't change the level of msrmnt
    - A nominal variable is still nominal; an ordinal still ordinal
  - They include two "hidden decimals":
    50% = 50 per 100 = 50/100 = 0.50

Lab, Part I (*not* in SPSS)

- If the sample proportion is 50% and the standard *error* is 10%, construct *and interpret*…

1. a 90% confidence interval
2. a 95% confidence interval
3. a 95.44% confidence interval

Lab, Part II (*partly* in SPSS)

- Get %'s from freq. distributions in SPSS
- *Don't* get S.E. or C.I. from SPSS! Assumes interval!

4. Calculate a 95% confidence interval for the proportion who like Rap
5. Calculate a 95% confidence interval for the proportion who like Opera
6. Calculate a 95% confidence interval for the proportion who like Bluegrass
7. Compare those intervals – draw a picture, and make conclusions based on how the intervals compare to each other!