STAT:2010/4200
Statistical Methods and Computing

Robustness of t procedures
Inference for Proportions

Lecture 18
Apr. 6, 2020

Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

Robustness

- Recall: we use t confidence intervals and one-sample t hypothesis tests when we assume that values in the population of interest follow a normal distribution.
  - t confidence levels and significance levels are exacltly correct if the population distribution is exactly normal.
  - But no population is *exactly* normal, so...
  - Critical issue: how strongly are t procedures affected by small and large violations of the assumption of normality of the population?
- A procedure to compute a confidence interval or significance test is called *robust* if the confidence level or p-value does not change much when the assumptions of the procedure are violated.

- Samples from normal distributions usually have very few outliers.
  - Outliers suggest that data are not a sample from a normal population.
- t procedures are strongly influenced by outliers in the sample data
  - because they are based on $\bar{x}$ and s
- But t procedures are quite robust against violations of the normality assumption when there are no outliers.
  - especially if population distribution is roughly symmetrical
  - and especially if sample size is large
    - ∗ Recall that Central Limit Theorem says that sampling distribution of $\bar{x}$ becomes approximately normal if sample size is "large enough" even if population distribution is not normal.

Rules of thumb regarding one-sample t procedures

- Always plot sample data to check for skewness and outliers before using t procedures, especially for small samples.
- Sample size less than 15: Use t procedures if the sample data look close to normal. If not, or if outliers are present, get help regarding an appropriate alternative to t.
- Sample size $\geq$ 15: t procedures can be used except in the presence of outliers or very strong skewness.
- Sample size $\geq$ about 40: t procedures can be used even when distribution of sample data is clearly skewed.

- Assumption that data are a SRS from the population generally is more critical than normality assumption for use of t procedures.

Rules of thumb regarding two-sample t procedures

- 2-sample t procedures are *more* robust than 1-sample t procedures, especially if distributions are asymmetric
- If
  - sample sizes are equal in the 2 samples
  - and distributions in the two *populations* are similar in shape

  then 2-sample t procedures are quite accurate even for $n_1 = n_2$ as small as 5.
- for different shapes of population distributions, larger samples needed
- for rules of thumb, use those for 1-sample t procedures, but replace "sample size" with "sum of sample sizes"

Inference about a population proportion: some examples

- In New York City, a study was conducted to evaluate factors associated with the need for special education in children. In a random sample of 45 children enrolled in the public school special ed program, 5 had mothers who had had more than 12 years of schooling. That's 11.1% of the sample.

  Based on these data, what can we say about the percent of all New York City children requiring special ed whose mothers have more than 12 years of schooling? We want to estimate a single population proportion.

- Is the advice given by a physician during a routine physical exam effective in encouraging patients to quit smoking? A study looked at 114 current smokers whose doctors talked to them during a routine exam about the hazards of smoking and encouraged them to quit smoking. A second group of 96 current smokers who had a routine exam were given no advice regarding smoking. All patients were given a follow-up exam. 11 of the 114 patients (9.6%) who had received the advice reported that they had quite smoking, while 7 of the 96 others (7.3%) reported that they had quit smoking.

  Is this significant evidence that physician advice makes a difference in patients' quitting smoking? We wish to compare two population proportions.

The point estimate of an unknown population proportion

- wish to estimate unknown population parameter p
  - proportion of a population that has some outcome
  - designate the outcome a "success"
- in first example, population of interest is NYC children requiring special ed
- p is proportion whose mothers had more than 12 years of schooling

- statistic that estimates p is *sample proportion* $\hat{p}$

$$\begin{aligned}\hat{p} &= \frac{number\ of\ successes\ in\ sample}{number\ of\ observations\ in\ sample}\\ &= \frac{5}{45}\\ &= 0.111\end{aligned}$$

Inference for a population proportion

What would happen if we took many different samples and computed $\hat{p}$ from each one?

If we:

- Choose a simple random sample of size n from a large population that contains an unknown population proportion p of "successes"
- Compute the **sample proportion** of successes

$$\hat{p} = \frac{number\ of\ successes\ in\ sample}{n}$$

Then:

- As the sample size increases, the sampling distribution of $\hat{p}$ becomes **approximately normal**
- The mean of the sampling distribution is the true p.
- The *standard deviation* of the sampling distribution is

$$\sqrt{\frac{p\,(1-p)}{n}}$$

How large must n be in order for the normal approximation to the distribution of $\hat{p}$ to be reasonably accurate?

- Rule of thumb: both $n\,p$ and $n\,(1-p)$ should equal at least 10.
  Note: Up until about 10 years ago, the recommendation was that both $n\,p$ and $n\,(1-p)$ should equal at least 5. However, the more conservative rule with 10 is now considered safer.

Another consideration: The formula for the standard deviation of $\hat{p}$ is not accurate unless the population is much larger than the sample.

- Rule of thumb: The population must be at least 10 times as large as the sample.

Inference about proportions is still possible when the rules of thumb are not satisfied, but more

elaborate methods are needed.

Example:

- Suppose that in fact 12% of all NYC children requiring special ed had mothers with more than 12 years of schooling.
- The study sampled 45 children.
- What is the probability that at least 8% of such a sample have mothers with more than 12 years of schooling?

The rules of (old) thumb are met, so we'll proceed for now:

- There are at least 10×45 children in NYC who require special ed.
- $np = 45(.12) = 5.4$
- $n(1-p) = 45(.88) = 39.6$

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.12)(0.88)}{45}} = 0.048$$

We want the probability that $\hat{p}$ is 0.08 or greater.

Standardize $\hat{p}$, producing a z statistic.

$$z = \frac{\hat{p} - 0.12}{0.048}$$

The probability that we want is

$$\begin{aligned} P(\hat{p} \geq 0.08 &= P(\frac{\hat{p} - 0.12}{0.048} \geq \frac{0.08 - 0.12}{0.048} \\ &= P(Z \geq -0.833) \\ &= 1 - 0.203 \\ &= 0.797 \end{aligned}$$

Inference about p

To do inference about a population proportion p, we use the z statistic that results from standardizing the sample proportion $\hat{p}$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

This z statistic has approximately the standard normal distribution with mean 0 and standard deviation 1 as long as the rules of thumb obtain.

Of course we don't know the value of the unknown p, so:

- To test the null hypothesis

  $$H_0 : p = p_0$$

  just replace p by p_0 in the z statistic and in rule of thumb 2.

- In a confidence interval for p, we have no specific value to substitute. In large samples, $\hat{p}$ will be close to the true p. So:
  - Replace p by $\hat{p}$ in rule of thumb 2.
  - Replace p by $\hat{p}$ to estimate the standard deviation of $\hat{p}$. This gives the *standard error of $\hat{p}$*.

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Assumptions for inference about a proportion

- The data are a simple random sample from the population of interest.
- The population is at least 10 times as large as the sample.
- For a test of $H_0 : p = p_0$, the sample size n is large enough that
  - $np_0 \geq 10$
  - $n(1 - p_0) \geq 10$
- For a confidence interval, n is large enough that
  - the count of the number of successes $n\hat{p} \geq 10$
  - the count of the number of failures $n(1 - \hat{p}) \geq 10$

Example: the study of children requiring special ed

We wish to use the data to compute a c.i. for the proportion of NYC children requiring special ed whose mothers had more than 12 years of schooling. Are the assumptions met?

- What is the population of interest, and is it at least 10 times as large as the sample?

- The counts of "Yes" (5) and "No" (40) responses are both ≥ 10?
  No! But let's continue just to see how the procedure works. This is a case where we really should use SAS to get the exact results instead of using this normal approximation.

Computing the confidence interval using the normal approximation

Suppose we want a two-sided 99% confidence interval for p. Table A (and previous experience) tells us that the normal critical value $z^* = 2.58$.

$$
\begin{aligned}
\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.111 \pm 2.58 \sqrt{\frac{(0.111)(0.889)}{45}} \\
&= 0.111 \pm 0.121 \\
&= (0.00, 0.232)
\end{aligned}
$$

We are 99% confident that the percent of NYC children requiring special ed whose mothers have at least 12 years of schooling about 0% and 23.2%.

The plus-four confidence interval for p

- previously accepted rules of thumb for sample sizes needed for accuracy of normal-theory confidence intervals for proportions aren't reliable
  - confidence level actually may be *smaller* than claimed – bad!
- quick-and-dirty solution: the plus-four confidence interval:
  - add 2 successes and 2 failures to the true counts in your sample
  - then use the normal-theory approximation with augmented data
- better solution: use the exact confidence interval calculation in SAS