

STAT22000 Autumn 2013 Lecture 24

Yibi Huang

November 22, 2013

8.1 Inference for a Single Proportion

Lecture 24 - 1

Lecture 24 - 2

8.1 Inference for a Single Proportion

- ▶ Large-sample confidence interval for p
- ▶ Wilson's "Plus four" confidence interval for p
- ▶ Significance test for a single proportion
- ▶ Choosing a sample size

Inference for Proportions

Suppose we are interested in the proportion p of individuals with some characteristic of a certain population. We may

1. Draw a **simple random sample of size n** .
2. Let X be the count of "success" in the sample. (Here a "success" is an observation with the characteristic of interest)
3. Estimate the unknown true **population proportion p** with the **sample proportion $\hat{p} = X/n$**

What is the **sampling distribution of \hat{p}** ?

- ▶ The exact distribution of X is $Bin(n, p)$.
- ▶ The Central Limit Theorem tells us when n is sufficiently large

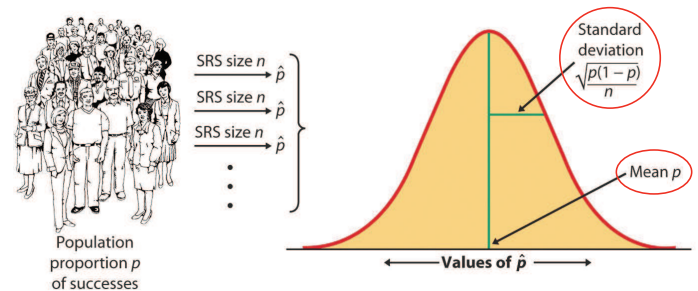
$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Lecture 24 - 3

Sampling Distribution of \hat{p}

When we say $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, what does it mean?

It is the distribution of \hat{p} when we repeatedly draw many SRSs of the same size n .



Lecture 24 - 4

Conditions that CLT Applies to \hat{p}

- ▶ The data used for the estimate are an SRS from the population studied.
- ▶ The population size is at least 10 times as large as the sample size.
 - ▶ This is for insuring that X is Binomial
- ▶ The sample size n is large enough.

How large the sample size n needs to be ?

- ▶ It depends in part on the value of p and the test conducted.

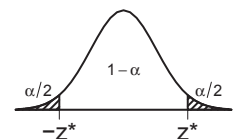
Lecture 24 - 5

Large-Sample Confidence Interval for p

An approximate $(1 - \alpha)100\%$ CI for the population proportion p is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is chosen so that $P(Z > z^*) = \alpha/2$ for $Z \sim N(0, 1)$.



Remark: The precise margin of error should be $z^* \sqrt{p(1-p)/n}$. However, since p is unknown, we replace it with \hat{p} . This makes the confidence intervals for p are not very accurate, i.e., the actual confidence levels are often lower than the nominal $(1 - \alpha)100\%$.

Use this method only when $n\hat{p}$ and $n(1 - \hat{p})$ are both at least 15.

Lecture 24 - 6

Example: Side Effects of Pain Relievers (1)

Arthritis is a painful, chronic inflammation of the joints, so many arthritis patients rely on pain relievers, like Ibuprofen. However, Ibuprofen may induce side effects (like dizziness, muscle cramp, allergy, or even seizure) on some patients.

A study interviewed 440 arthritis patients taking Ibuprofen, and found 23 had experienced side effects. Suppose the 440 patients is a SRS from the population of arthritis patients taking Ibuprofen.

Find a 90% confidence interval for the population proportion p of arthritis patients who suffer some adverse symptoms.

Lecture 24 - 7

Example: Side Effects of Pain Relievers (2)

The sample proportion is $\hat{p} = \frac{23}{440} \approx 0.052$.

z^*	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	...
	50%	60%	70%	80%	90%	95%	96%	98%	99%	...
	Confidence level C									

From Table D, we find the z^* for 90% confidence level is 1.645. So a 90%-CI for p is

$$\begin{aligned} \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &\approx 0.052 \pm 1.645 \sqrt{\frac{0.052 \times (1-0.052)}{440}} \\ &\approx 0.052 \pm 0.017 = (0.035, 0.069) \end{aligned}$$

Conclusion: With a 90% confidence, between 3.5% and 6.9% of arthritis patients taking this pain medication experience some adverse symptoms.

Lecture 24 - 8

Problems of the Classic CIs for p

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- ▶ If **no success**, or **no failure** in the sample, then $\hat{p} = 0$ or 1. In that case, the width of the CI for p is 0 because $\sqrt{\hat{p}(1-\hat{p})/n} = 0$.
- ▶ Likewise, when p is close to 0 or 1, $\sqrt{\hat{p}(1-\hat{p})/n}$ tend to underestimate $\sqrt{p(1-p)/n}$. The classical CIs based on \hat{p} may not achieve the nominal confidence level, $1 - \alpha$.

Edwin Bidwell Wilson (1927) suggested the following adjustment, known as **Wilson's estimate**, or the **plus-four estimate**

$$\tilde{p} = \frac{X + 2}{n + 4} = \frac{\text{number of "success" + 2}}{\text{number of all observations + 4}}$$

We simply add two "phantom" successes and two "phantom" failures to the data.

Lecture 24 - 9

Wilson's "Plus-Four" Confidence Interval for p

And an approximate level C confidence interval is:

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

- ▶ Use this method when \hat{p} is close to 0 or 1, in particular when

$$n\hat{p} \leq 15 \quad \text{or} \quad n(1-\hat{p}) \leq 15$$

but we still requires that $n\hat{p}$ and $n(1-\hat{p})$ both > 10 to use Wilson's CI.

Lecture 24 - 10

Example: Side Effects of Pain Relievers (3)

We now use the "plus four" method to calculate the 90% confidence interval for the population proportion of arthritis patients who suffer some "adverse symptoms."

The "plus four" estimate of p is $\tilde{p} = \frac{23+2}{440+4} = \frac{25}{444} \approx 0.056$

An approximate 90% CI for p using the "plus four" method is:

$$\begin{aligned} \tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} &\approx 0.056 \pm 1.645 \sqrt{\frac{0.056 \times (1-0.056)}{444}} \\ &\approx 0.056 \pm 0.018 = (0.038, 0.074) \end{aligned}$$

Conclusion: With a 90% confidence, between 3.8% and 7.4% of arthritis patients taking this pain medication experience some adverse symptoms.

Lecture 24 - 11

Why Not Using t for Proportions?

If the sample size n is large enough to apply CLT, n is usually a few hundreds or a few thousands.






The t_{n-1} distribution is very close to normal when $n > 99$, and therefore it is justified to do normal-based inference for proportions.

Lecture 24 - 12

Hypothesis testing for a population proportion

Suppose we want to test $H_0 : p = p_0$ for some fixed value p_0 .
Under H_0 ,

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0, 1)$$

	Two-tailed test	One-tailed test	
H_a	$p \neq p_0$	$p < p_0$	$p > p_0$
P-value		 	 

Here n should be at least so large that $np_0 \leq 10$, and $n(1 - p_0) \leq 10$.

Lecture 24 - 13

Remark. Recall that for confidence intervals, we use

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

to estimate the SD of \hat{p} , but for hypothesis testing we use

$$\sqrt{\frac{p_0(1 - p_0)}{n}}$$

Why?

- ▶ Recall by CLT when n is large $\hat{p} \sim N\left(p, \sqrt{\frac{p(1 - p)}{n}}\right)$
- ▶ When constructing CIs for p , p is unknown, so we estimate $\sqrt{\frac{p(1 - p)}{n}}$ by $\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$
- ▶ Under $H_0 : p = p_0$, p is known to be p_0 . There is no need to estimate p and the SD $\sqrt{\frac{p(1 - p)}{n}}$ is simply $\sqrt{p_0(1 - p_0)/n}$.

Lecture 24 - 14

Example: Children's Play Preference (1)

- ▶ An observational study was conducted at Chicago Children's Museum to determine the age at which a child's preferred play partner switched from gender-neutral to a same-sex peer
- ▶ Let p be the proportion of children who preferred to interact with a same-sex peer
- ▶ If no preference, then $p = 1/2$
- ▶ For 6-year old children 59 of 97 preferred to interact with a same-sex peer ($\hat{p} = 59/97 \approx 0.61$)
- ▶ Test to see if there was a preference for interaction with a same-sex peer for the 6-year old children ($\alpha = 0.05$)

Lecture 24 - 15

Example: Children's Play Preference (2)

We want to test $H_0 : p = 1/2$ versus $H_a : p \neq 1/2$

- ▶ Is the z test appropriate?
Check $np_0 > 10$ and $n(1 - p_0) > 10$
(Yes; $np_0 = 97(0.5) = 48.5 > 10$ and $n(1 - p_0) = 48.5 > 10$)
- ▶ Test statistic
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.61 - 0.5}{\sqrt{0.5(1 - 0.5)/97}} = \frac{0.11}{0.0508} = 2.17$$
- ▶ P-value is $2P(z > 2.17) = 2(0.015) = 0.03 < \alpha = 0.05$
- ▶ Conclude: 6-year old children prefer to interact with same-sex peers

Lecture 24 - 16

Choosing a Sample Size

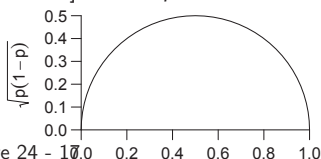
How large the sample size n need to be to make the margin of error of a $(1 - \alpha)$ CI $\leq m$?

$$\text{margin of error} = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq m \Rightarrow n \geq \left(\frac{z^*}{m}\right)^2 \hat{p}(1 - \hat{p})$$

Problem: \hat{p} is UNKNOWN before we get the data. Need to make a guess for p^* . How to choose p^* ?

1. Conduct a small pilot study, or use prior studies or knowledge to get a range for possible values of p . Choose the bound that is closer to 0.5. E.g., if possible range of p is $[0.1, 0.2]$, choose $p^* = 0.2$.
if possible range of p is $[0.85, 0.95]$, choose $p^* = 0.85$ s.

2. The most **conservative** approach is to choose $p^* = .5$.



Lecture 24 - 17

Example – Sample size calculation for a proportion

A 1993 survey reported that 72.1% of freshmen responding to a national survey were attending the college of their first choice. Suppose that $n = 500$ students responded to the survey.

1. Find a 95% CI for the proportion p of college freshmen attending their first choice college.
2. Suppose that given the CI, we want to conduct a survey which has a margin of error of 1% (i.e. $m = 0.01$) with 95% confidence? How many people should we interview?

Lecture 24 - 18

Example – Sample size calculation for a proportion

The two-sided 95% confidence interval is:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.721 \pm 1.96 \sqrt{\frac{0.721(1-0.721)}{500}} \\ = (0.682, 0.760)$$

We have good reason to believe p is in that range. For a sample size calculation, we choose the p closest to .5, that is $p = 0.682$

$$\left(\frac{z^*}{m}\right)^2 p(1-p) = \left(\frac{1.96}{.01}\right)^2 \times 0.682(1-0.682) = 8331.508$$

The required sample size is 8332. We need a much larger sample size than the original study because we want a smaller margin of error.

Lecture 24 - 19

Gallup Polls

A Gallup poll pre-election survey based on a sample of 1,000 people estimates a 65% vote for the Democratic candidate in a certain election. True or false, and explain:

The margin of error for a 95% CI can be figured as follows —

$$1.96 \times \frac{\sqrt{0.65 \times 0.35}}{\sqrt{1,000}} \times 100\% \approx 3\%.$$

False. The Gallup poll isn't based on a **simple random sample**, so the formulas don't apply.

Lecture 24 - 21

True or False (Cont'd)

- ▶ There are about 68% probability for the percentage of Democrats in the population to be in the range $54.3\% \pm 1.6\%$.

False. The population percentage either is in this particular interval, or is not in this particular interval — there is no element of chance involved.

- ▶ If another survey organization takes a SRS of 1,000 persons, there is about a 95% probability that the percentage of Democrats in their sample will be in the range $54.3\% \pm 3.2\%$.

False. Again, we are interested in the population, not the sample.

Moreover, the probability depends on the actual p , and may not be 95%.

Lecture 24 - 23

What's Wrong?

A box contains 10,000 marbles, of which some are red and the others blue. To estimate the percentage of red marbles in the box, 100 are drawn at random without replacement. Among the draws, 1 turns out to be red. The percentage of red marbles in the box is estimated as 1%.

True or false, and explain: An approximate 95% confidence interval for the percentage of red marbles in the box is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.01 \pm 1.96 \sqrt{\frac{0.01 \times 0.99}{100}} \\ \approx 0.01 \pm 0.02 = 1\% \pm 2\%.$$

Answer. False. The sample size (number of draws) is too small for the normal approximation to be applied here because

$$n\hat{p} = 100 \times 0.01 = 1 < 10.$$

When p is close to 0 or 1, a large sample is required.

Lecture 24 - 20

True or False

A simple random sample of 1,000 persons is taken to estimate the percentage of Democrats in a large population. It turns out that 543 of the people in the sample are Democrats. The sample percentage \hat{p} is $(543/1,000) \times 100\% = 54.3\%$. The SE for the sample percentage of Democrats is figured as 1.6%. True or false, and explain:

- ▶ $54.3\% \pm 3.2\%$ is a 95%-C.I. for the percentage of Democrats in the population.

True.

- ▶ $54.3\% \pm 3.2\%$ is a 95%-C.I. for the percentage of Democrats in the sample.

False. The sample percentage is known (but not what you're interested in.) You don't need a C.I. for it.

Lecture 24 - 22

Describe the Population and Parameter Clearly

A psychologist administers a test of passivity to 100 students in his class and finds that 20 of them score over 50. He concludes that approximately 20% of all students would score over 50 on this test. He recognizes that this estimate may be off a bit, and he estimates the margin of error for a 95% CI as follows:

$$1.96 \times \frac{\sqrt{0.2 \times 0.8}}{\sqrt{100}} \times 100\% \approx 8\%.$$

What does statistical theory say?

- ▶ Watch out for this guy!
- ▶ What population is he talking about?
- ▶ Why are his students like a simple random sample from the population?
- ▶ Until he can answer these questions, don't pay too much attention to the calculations.

Lecture 24 - 24