STAT 22000 Lecture Slides
Inference for One and Two Proportions

Yibi Huang
Department of Statistics
University of Chicago

Outline

This set of slides covers Section 6.1 and 6.2 in the text.

- Large-sample confidence interval for a single proportion
- Choosing a sample size
- Hypothesis testing for a single proportion
- Large-sample CI for the difference of two proportions
- Hypothesis testing for the equality of two proportions

Inference for a Single Proportion

Inference for Proportions

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

1. Draw *simple random sample* of size n
2. Let X be the count of "successes" 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* $\widehat{p} = X/n$

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

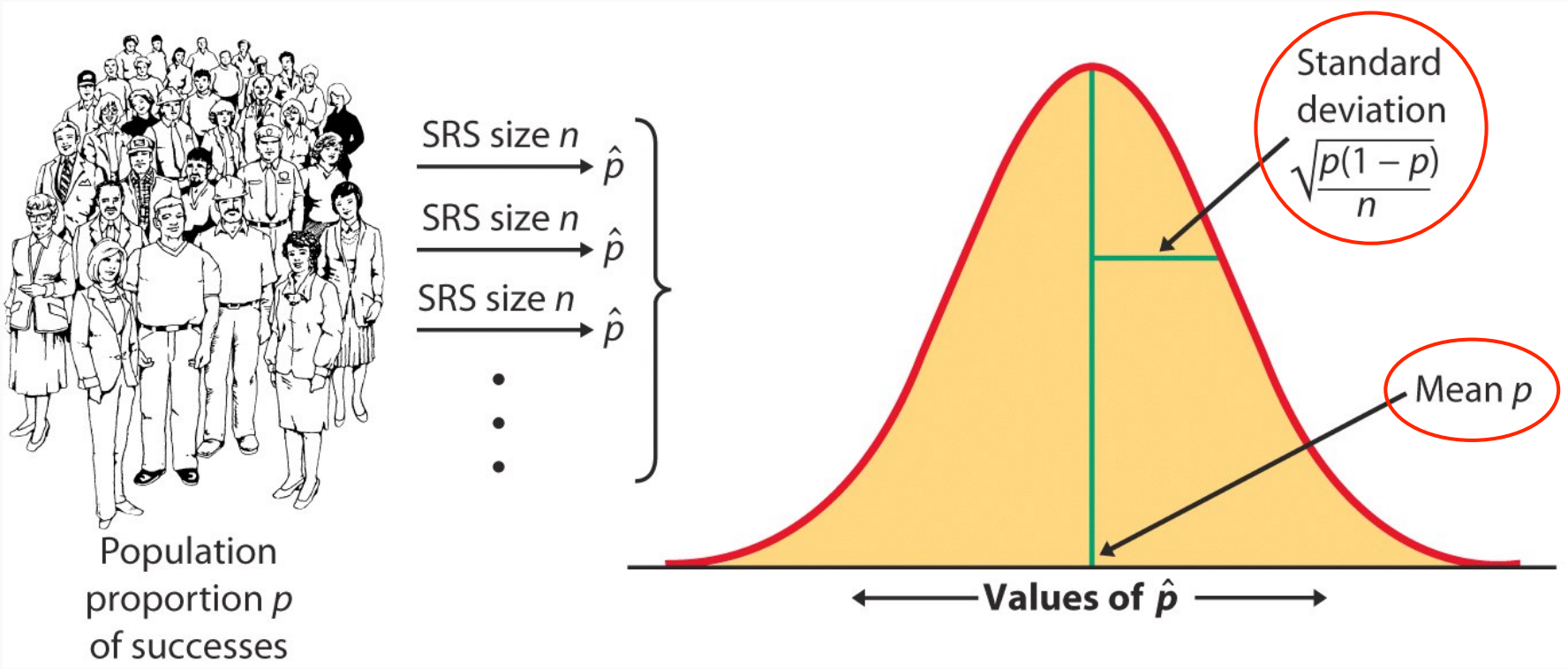
- The exact distribution of X is $Bin(n, p)$.
- Normal approximation to Binomial tells us that when n is sufficiently large

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

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.



Large-Sample Confidence Interval for p

An approximate CI for the population proportion p is

$$\widehat{p} \pm z^* SE \quad \text{where} \quad SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

where

Confidence level	90%	95%	99%
z^*	1.645	1.960	2.576

Remark: The exact SE should be $\sqrt{p(1-p)/n}$, but the unknown p is replaced with the estimate $\widehat{p}$. This large-sample CI is not very accurate, meaning the actual confidence level often falls below the nominal level.

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.

Example: Side Effects of Pain Relievers (2)

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

The z^* for a 90% CI is 1.645. So a 90%-CI for p is

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

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

Why Not Using t for Proportions?

If the sample size n is large enough to apply normal approximation to binomial, 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.

Choosing a Sample Size

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

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

But $\widehat{p}$ 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 $p^* = 0.85$.
2. The most **conservative** approach is to choose $p^* = 0.5$ since the margin of error is the largest when $\widehat{p} = 0.5$.

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?

Example – Sample Size Calculation for a Proportion

The two-sided 95% confidence interval is:

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

We have good reason to believe p is in that range. For a sample size calculation, we choose the p closest to 0.5, that is $p = 0.682$

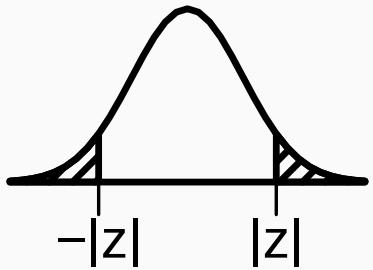
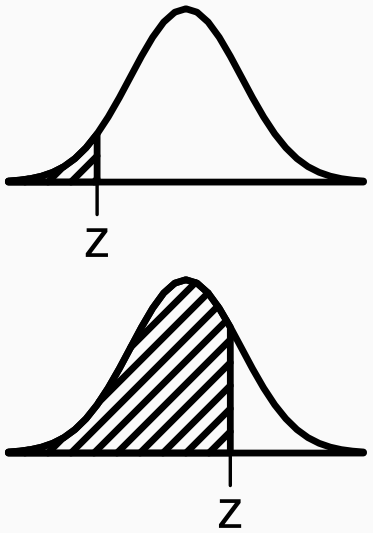
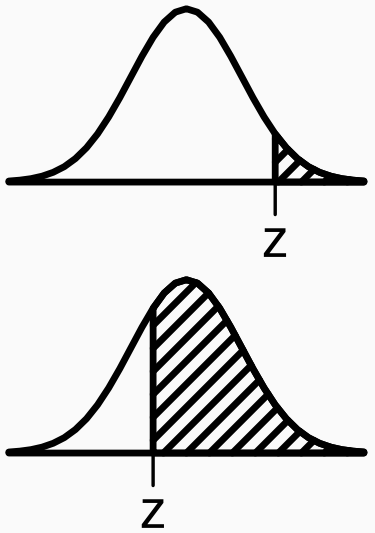
$$\left(\frac{z^*}{m}\right)^2 p(1-p) = \left(\frac{1.96}{0.01}\right)^2 \times 0.682(1-0.682) \approx 8331.51$$

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.

Hypothesis Testing for a Proportion

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

Under H_0, $z = \dfrac{\widehat{p} - p_0}{SE} \dot{\sim} N(0,1)$, where $SE = \sqrt{\dfrac{p_0(1-p_0)}{n}}$

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

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

Remark. Recall that for confidence intervals, we use

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

but for hypothesis testing we use

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

Why?

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

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

- For 5-year old children, 78 of 162 preferred to interact with a same-sex peer (48%)
- For 6-year old children, 59 of 97 preferred to interact with a same-sex peer (61%)

Under the *null hypothesis* of *no preference*, the probability that a child select a same-sex peer is $p = 0.5$

We want to test if 6-year old children had a preference interacting with a same-sex peer.

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 whether $np_0 > 10$ 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

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.61 - 0.5}{\sqrt{0.5(1 - 0.5)/97}} = \frac{0.11}{0.0508} = 2.17$$

- 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 rather than gender-neutral

Conditions Required for Using the Large-Sample CI and Tests for Proportions

- The observations are (nearly) i.i.d. from the population studied.
  - If SRS, the sample size is at most 10% of the population size.

- The sample size n is large enough. A rule of thumb is that
  - to use the large-sample CI:
    $n\widehat{p}$ and $n(1-\widehat{p})$ need to be both ≥ 10
  - to use the large-sample test for the H_0 of $p = p_0$:
    np_0 and $n(1-p_0)$ need to be both ≥ 10

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

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

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

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

Comparing Two Proportions

Comparing Two Proportions

Choose an SRS of size n_1 from a large population having proportion p_1 of successes and an independent SRS of size n_2 from another population having proportion p_2 of successes.

Population	Population Proportion	Sample Size	Count of Successes	Sample Proportion
1	p_1	n_1	X_1	$\widehat{p}_1 = X_1/n_1$
2	p_2	n_2	X_2	$\widehat{p}_2 = X_2/n_2$

Large Sample Confidence Intervals for $p_1 - p_2$

When n_1 and n_2 are both large,

$$\hat{p}_1 - \hat{p}_2 \dot{\sim} N\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

An approximate $(1-\alpha)100\%$ CI for $p_1 - p_2$ is

$$\text{estimate} \pm z^* \text{SE}$$

where

$$\text{estimate} = \hat{p}_1 - \hat{p}_2, \quad \text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Use this method only when the number of successes and the number of failures in both samples are at least 10, i.e.,

$$n_1\hat{p}_1, \quad n_1(1-\hat{p}_1), \quad n_2\hat{p}_2, \quad n_2(1-\hat{p}_2) \text{ all } \geq 10.$$

Example: Aspirin and Heart Attacks (1)

The Physicians' Health Study was a 5-year randomized study published testing whether regular intake of aspirin reduces mortality from cardiovascular disease[1].

- Participants were male physicians 40-84 years old in 1982 with no prior history of heart attack, stroke, and cancer, no current liver or renal disease, no contraindication of aspirin, no current use of aspirin
- Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo.
- Response: whether the participant had a heart attack (including fatal or non-fatal) during the 5 year period.

[1] *Source*: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262-64,1988.

Example: Aspirin and Heart Attacks (2)

Result:

	Heart Attack?		
Group	Yes	No	Sample Size
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\Rightarrow \hat{p}_1 = \frac{189}{11034} \approx 0.0171$

$\Rightarrow \hat{p}_2 = \frac{104}{11037} \approx 0.0094$

The z^* for a 99% CI is 2.58, so the 99% CI for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= 0.0171 - 0.0094 \pm 2.58 \sqrt{\frac{0.0171(1-0.0171)}{11034} + \frac{0.0094(1-0.0094)}{11037}}$$

$$= 0.0077 \pm 0.0040 = (0.0037, 0.0117)$$

Example: Aspirin and Heart Attacks (3)

Conclusion:

- As the 99% CI does not contain 0, the incidence rate of heart attack was significantly lower in aspirin group than in the placebo group
- Can we claim that taking aspirin every other day is effective in reducing the chance of heart attack?
  Yes, because it was a randomized, double-blind, placebo-controlled experiment.

Testing the Equality of Two Proportions (1)

While we test

$$H_0 : p_1 = p_2$$

the SE for $\hat{p}_1 - \hat{p}_2$ under H_0 is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where p is the common value of p_1 and p_2.

How to estimate the common p?

Testing the Equality of Two Proportions (2)

If there is no difference in the proportion of successes between the two populations, we can combine the samples, giving the *pooled estimate* of p

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1}{n_1 + n_2}\hat{p}_1 + \frac{n_2}{n_1 + n_2}\hat{p}_2$$

So the SE for testing $H_0 : p_1 = p_2$ is

$$SE = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and the z-statistic for testing $\mathrm{H}_0 : p_1 = p_2$ is

$$z = \frac{\text{estimate}}{\text{SE}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Under H_0, the Z-statistic is approximately $N(0, 1)$ provided that

$$n_1\hat{p}, \quad n_1(1 - \hat{p}), \quad n_2\hat{p}, \quad n_2(1 - \hat{p}) \text{ all } \geq 10.$$

Example: Aspirin and Heart Attacks (4)

Group	Sample Size	Heart Attack
Placebo	11034	189
Aspirin	11037	104

For testing $H_0 : p_1 = p_2$,

$$\begin{aligned}
\hat{p}_1 - \hat{p}_2 &= \frac{189}{11034} - \frac{104}{11037} \approx 0.0077 \\
\hat{p} &= \frac{189 + 104}{11034 + 11037} \approx 0.0132 \\
\text{SE} &= \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\
&\approx \sqrt{0.0132(1-0.0132)\left(\frac{1}{11034} + \frac{1}{11037}\right)} \approx 0.00154 \\
z\text{-statistic} &= \frac{\hat{p}_1 - \hat{p}_2}{\text{SE}} \approx \frac{0.0077}{0.00154} \approx 5.001
\end{aligned}$$

Example: Aspirin and Heart Attacks (4)

The 2-sided p-value is $2P(Z > 5.001) < 0.0004$ by normal table and is 0.00000057 by R.

Not surprisingly, we are getting strong evidence that the two probabilities are different.

Example: Partisanship 2015

A Gallop poll in 2015 based on a random sample of 12137 adults in U.S. (aged ≥ 18), found that 29% self-identified as Democrats, 26% as Republicans, and 45% as independent or other. *True or False and explain*: a 95% confidence interval for the difference of proportions of American adults self-identified as Democrats and Republicans $p_D - p_R$ is

$$0.29-0.26 \pm 1.96\sqrt{\frac{0.29(1-0.29)}{12137}+\frac{0.26(1-0.26)}{12137}} = (0.019, 0.041)$$

- How many samples are there? One or two?
- The two sample percentages, 29% and 26%, are calculated based on the same sample. They were not independent, but negatively correlated. The more people identified as Democrats, the fewer identified as Republicans. One cannot use a two-sample CI here.

---

http://http://news.gallup.com/poll/188096/democratic-republican-identification-near-historical-lows.aspx

Example: Partisanship 2015 v.s. 2011

Continue the previous example. Another survey of 15,000 American adults in 2011 found that 35.3% identified as Democrats, 34.0% as Republicans, and 30.7% as independent or other. Assume both surveys in 2011 and 2015 were both based on simple random samples. Can we test whether there were more American adults self-identified as independent or other in 2015 than in 2011 using a two-sample z-test for proportions?

Yes, the percentages identified as independent or others in 2011 and in 2015 were based on two independent samples.

Example: Partisanship 2015 v.s. 2011

For testing $H_0 : p_{2011} = p_{2015}$,

$$\begin{aligned}
\hat{p}_{2015} - \hat{p}_{2011} &= 0.450 - 0.307 \approx 0.143 \\
\hat{p} &= \frac{0.450 \times 12137 + 0.307 \times 15000}{12137 + 15000} \approx 0.371 \\
\text{SE} &= \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\
&\approx \sqrt{0.371(1-0.371)\left(\frac{1}{12137} + \frac{1}{15000}\right)} \approx 0.00590 \\
z\text{-statistic} &= \frac{\hat{p}_{2015} - \hat{p}_{2011}}{\text{SE}} \approx \frac{0.143}{0.00590} \approx 24.2
\end{aligned}$$

As the z-statistic is huge, there is super strong evidence that there were a higher percentages of American adults self-identified as independent or other in 2015 than in 2011.

Exercise 6.24 Heart Transplant (p.317)

The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might benefit from a new heart. Patients were randomly assigned to receive a transplant (treatment group) or not (control group). The table below shows how many patients survived and died in each group.

	control	treatment
alive	4	24
dead	30	45

Can we test whether the survival rate is the same in the two groups using a two-sample z-test for proportions?

No. Since the success-failure condition is not met, the two-sample z-statistic won't be approx. normal.

B. Turnbull et al. "Survivorship of Heart Transplant Data" In: *J. of the Amer. Stat. Assoc.* 69 (1974), pp. 74-80.

Summary: Standard Errors

	one sample	two samples
mean	$\frac{s}{\sqrt{n}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ if $\sigma_1 \neq \sigma_2$ $s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ if $\sigma_1 = \sigma_2$ where $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$
proportion (CIs)	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
proportion (tests)	$H_0 : p = p_0$ $\sqrt{\frac{p_0(1-p_0)}{n}}$	$H_0 : p_1 = p_2$ $\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$

A Larger Population Does NOT Require a Larger Sample

- All the SEs depend on the sample size only, not the population size.
- The relative size of a sample to the population size doesn't matter. It is the absolute size of a sample that matters.
- A larger population does NOT require a larger sample!