Statistical Inference

Inference for Categorical Variables

Behnam Bahrak

Spring 2020

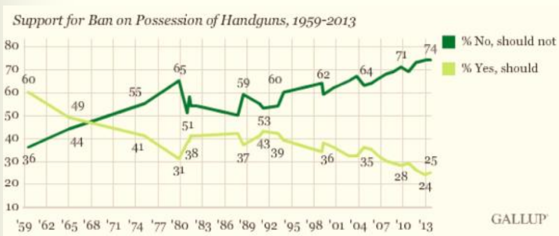
Dataset

- In early October 2013, a Gallup poll asked "Do you think there should or should not be a law that would ban the possession of handguns, except by the police and other authorized persons?"

(a) No, there should not be such a law
(b) Yes, there should be such a law
(c) No opinion

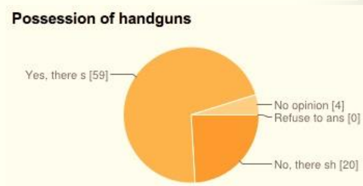
Dataset

US



International Students

No, there should not be such a law	**20**	24%
Yes, there should be such a law	**59**	71%
No opinion	**4**	5%



	success	n	$\hat{p}$
US	257	1028	0.25
International Students	59	83	0.71

Research Question

- How do international students and the American public at large compare with respect to their views on laws banning possession of handguns?

parameter of interest

Difference between the proportions of **all** international students and **all** Americans who believe there should be a ban on possession of handguns.

$p_{Intl} - p_{US}$

point estimate

Difference between the proportions of **sampled** international students and **sampled** Americans who believe there should be a ban on possession of handguns.

$\hat{p}_{Intl} - \hat{p}_{US}$

Estimating the difference between two proportions

point estimate ± margin of error

$$(\hat{p}_1 - \hat{p}_2) \pm z^{\star} SE_{(\hat{p}_1 - \hat{p}_2)}$$

- Standard error for difference between two proportions, for calculating a confidence interval:

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

Conditions for inference

Conditions for inference for comparing two independent proportions:

1. ***Independence:***
   - **within groups:** sampled observations must be independent within each group
     - random sample/assignment
     - if sampling without replacement, $n < 10\%$ of population
   - **between groups:** the two groups must be independent of each other (non-paired)
2. ***Sample size/skew:*** Each sample should meet the success-failure condition:
   - $n_1 p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
   - $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

Example

- Using a 95% confidence interval, estimate how international students and the American public at large compare with respect to their views on laws banning possession of handguns.

	$succ.$	n	$\hat{p}$
US	257	1028	0.25
International	59	83	0.71

1. Independence:
   - Sampled Americans independent of each other, sampled international students may not be.
2. Sample size / skew:
   - We can assume that the sampling distribution of the difference between two proportions is nearly normal.

Example

$$\hat{p}_{Intl} - \hat{p}_{US} \pm z^* SE$$

$$= (0.71 - 0.25) \pm 1.96 \sqrt{\frac{0.71 \times 0.29}{83} + \frac{0.25 \times 0.75}{1028}}$$

$$= 0.46 \pm 1.96 \times 0.0516$$

$$= 0.46 \pm 0.10$$

$$= (0.36, 0.56)$$

	$succ.$	n	$\hat{p}$
US	257	1028	0.25
International	59	83	0.71

Does the order matter?

remember $\underbrace{(\hat{p}_1 - \hat{p}_2)}_{\text{can be - or +}} \pm z^{\star} \underbrace{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}_{\text{always +}}$

$$
\begin{aligned}
P_{Intl} - P_{US} &= \\
&= (0.71 - 0.25) \pm 0.10 \\
&= 0.46 \pm 0.10 \\
&= (0.36, 0.56)
\end{aligned}
$$

$$
\begin{aligned}
P_{US} - P_{Intl} &= \\
&= (0.25 - 0.71) \pm 0.10 \\
&= -0.46 \pm 0.10 \\
&= (-0.56, -0.36)
\end{aligned}
$$

Example

- Based on the confidence interval we calculated, should we expect to find a significant difference (at the equivalent significance level) between the population proportions of international students and the American public at large who believe there should be a law banning the possession of handguns?

$$(p_{Intl} - p_{US}) = (0.36, 0.56)$$

$H_0: p_{Intl} - p_{US} = 0$

0 0.36 0.56

Reject H_0

Dataset

- A SurveyUSA poll asked respondents whether any of their children have ever been the victim of bullying. Also recorded on this survey was the gender of the respondent (the parent). Below is the distribution of responses by gender of the respondent.

	Male	Female
Yes	34	61
No	52	61
Not Sure	4	0
Total	90	122
$\hat{p}$	0.38	0.50
	34/90	61/122

H_0: $p_{male} - p_{female} = 0$

H_A: $p_{male} - p_{female} \neq 0$

✓ check conditions

✓ calculate test statistic & p-value

Working with one proportion ($\hat{p}$ vs. p)

	observed confidence interval	expected hypothesis test
success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

Working with two proportion ($\hat{p}$ vs. p)

	observed confidence interval	expected hypothesis test
success-failure condition	$n_1\hat{p}_1 \geq 10$ $\quad n_2\hat{p}_2 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $\quad n_2(1-\hat{p}_2) \geq 10$	$H_0 : p_1 = p_2$
standard error	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	

Pooled Proportion

$H_0 : p_1 = p_2 \quad = ?$

Pooled proportion:

$$\hat{p}_{pool} = \frac{total\ successes}{total\ n} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

Example

➢ Calculate the estimated pooled proportion of males and females who said that at least one of their children has been a victim of bullying.

$$\hat{p}_{pool} = \frac{34 + 61}{90 + 122}$$

$$\approx 0.45$$

	Male	Female
Yes	34	61
No	52	61
Not Sure	4	0
Total	90	122
$\hat{p}$	0.38	0.50

Working with two proportion ($\hat{p}$ vs. p)

	observed confidence interval	expected hypothesis test
success-failure condition	$n_1\hat{p}_1 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $n_2\hat{p}_2 \geq 10$ $n_2(1-\hat{p}_2) \geq 10$	$n_1\hat{p}_{pool} \geq 10$ $n_1(1-\hat{p}_{pool}) \geq 10$ $n_2\hat{p}_{pool} \geq 10$ $n_2(1-\hat{p}_{pool}) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$

What about means?

parameter of interest: μ

$$H_0 : \mu = null\ value$$

$$SE = \frac{s}{\sqrt{n}}$$

μ doesn't appear in SE

parameter of interest: p

$$H_0 : p = null\ value$$

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

p appears in SE

Example

➢ Are conditions for inference met for conducting a hypothesis test to compare the two proportions?

	Male	Female
Total	90	122
$\hat{p}$	0.38	0.50
$\hat{p}_{pool}$	0.45	

1. Independence:
✓ within groups: random sample & n < 10%
Sampled males independent of each other, sampled females are as well.
✓ between groups:
No reason to expect sampled males and females to be dependent.

2. Sample size / skew: ✓ Males: 90 × 0.45 = 40.5 and 90 × 0.55 = 49.5
✓ Females: 122 × 0.45 = 54.9 and 122 × 0.55 = 67.1

We can assume that the sampling distribution of the difference between two proportions is nearly normal.

Example

- Conduct a hypothesis test, at 5% significance level, evaluating if males and females are equally likely to answer "Yes" to the question about whether any of their children have ever been the victim of bullying.

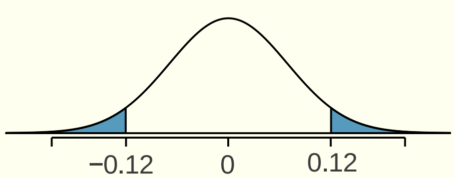
	Male	Female
Total	90	122
$\hat{p}$	0.38	0.50
$\hat{p}_{pool}$	0.45	

H_0: $p_{male} - p_{female} = 0$ $\qquad$ H_A: $p_{male} - p_{female} \neq 0$

$$\hat{p}_{male} - \hat{p}_{female} \sim N\left(mean = 0\,, SE = \sqrt{\frac{0.45 \times 0.55}{90} + \frac{0.45 \times 0.55}{122}} \approx 0.0691\right)$$

point estimate $= \hat{p}_{male} - \hat{p}_{female} = 0.38 - 0.50 = -0.12$

Example



	Male	Female
Total	90	122
$\hat{p}$	0.38	0.50
$\hat{p}_{pool}$	0.45	

point estimate $= -0.12$

null value $= 0$

SE $= 0.0691$

$$Z = \frac{-0.12 - 0}{0.0691} \approx -1.74$$

p-value $= P(|Z| > 1.74) \approx 0.08$