Two-Sample Proportion Inference

Modeling a difference in proportions

Download the section 12.Rmd handout to `STAT240/lecture/sect12-two-proportions`.

Download the file `chimpanzee.csv` to `STAT240/data`.

Material in this section is covered by Chapter 12 on the notes website.

Let's return to the chimpanzee data.

What is the difference in prosocial choices made by chimpanzee C with vs without a partner?

Parameter of interest: $p_{partner} - p_{nopartner}$, or $p_1 - p_2$

$$\text{point estimate} \;\pm\; \text{critical value} \;\times\; \text{standard error}$$

- For $p_1 - p_2$, the point estimate is $\hat{p}_1 - \hat{p}_2$

What about the other parts? Consider the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

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

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

A difference of independent normals is also normal.

$$\hat{p}_1 - \hat{p}_2 \ \dot{\sim} \ N\left(p_1 - p_2, \text{ SE of difference}\right)$$

$$\text{SE} \ = \ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Since we have a normal sampling distribution, we can build a Z CI.

We need to estimate the standard error. The **Agresti-Coffe** adjustment works like Agresti-Coull.

2 successes and 2 failures are distributed across *both* groups.

$$\hat{p}_{1AC} = \frac{X_1 + 1}{n_1 + 2}, \quad \hat{p}_{2AC} = \frac{X_2 + 1}{n_2 + 2}$$

For chimpanzee C:

$$\hat{p}_{1AC} = \frac{57+1}{90+2} = 0.63$$

$$\hat{p}_{2AC} = \frac{17+1}{30+2} = 0.56$$

The 95% AC interval is (-0.13, 0.27).

We have not done inference on p_1 or p_2, just the difference.

We are 95% confident that the difference in the % of prosocial choices is between (-0.13, 0.27).

- Negative: More prosocial without a partner
- Positive: More prosocial with a partner
- Ours covers 0

In general, an AC interval for a difference in proportions is

$$\hat{p}_{1AC} - \hat{p}_{2AC} \quad \pm$$

$$z_{\alpha/2} \times \sqrt{\frac{\hat{p}_{1AC}(1 - \hat{p}_{1AC})}{n_{1AC}} + \frac{\hat{p}_{2AC}(1 - \hat{p}_{2AC})}{n_{2AC}}}$$

Build and interpret a 95% CI for the difference in prosocial behavior for Chimpanzee B.

- $\hat{p}_1$: proportion of prosocial choices with a partner
- $\hat{p}_2$: proportion of prosocial choices without a partner

Make sure to use the A-C adjustment.

We can also use the normal to test a difference in proportions.

Is the probability of chimpanzee C making the prosocial choice higher when there is a partner?

Let's use $\alpha = 0.05$.

We have hypotheses

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_A : p_1 > p_2$$

$$H_0 : p_1 - p_2 = 0 \quad \text{versus} \quad H_A : p_1 - p_2 > 0$$

We build a test statistic based on $\hat{p}_1 - \hat{p}_2$.

From before:

$$\hat{p}_1 - \hat{p}_2 \;\dot{\sim}\; N\Big(p_1 - p_2,\ \text{SE of difference}\Big)$$

$$\text{SE} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

We will standardize this to create a Z test statistic. But, we can simplify things.

$p_1 = p_2$ implies that there is a **common proportion** p. This is the overall rate of prosocial choices.

Under H_0,

$$\hat{p}_1 - \hat{p}_2 \ \dot{\sim} \ N\left(0, \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Estimate p with the overall observed prosocial rate for C:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{57 + 17}{90 + 30}$$

We use $\hat{p}$ for our standard error, giving test statistic

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

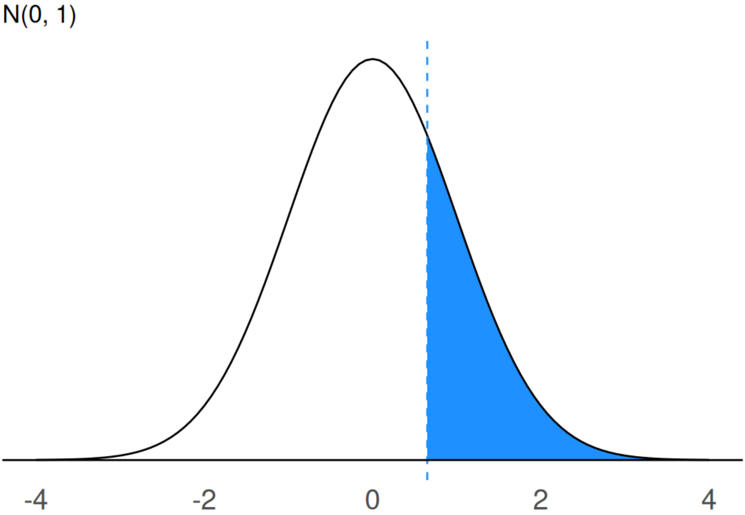
which has value $z_{obs} = 0.65$.

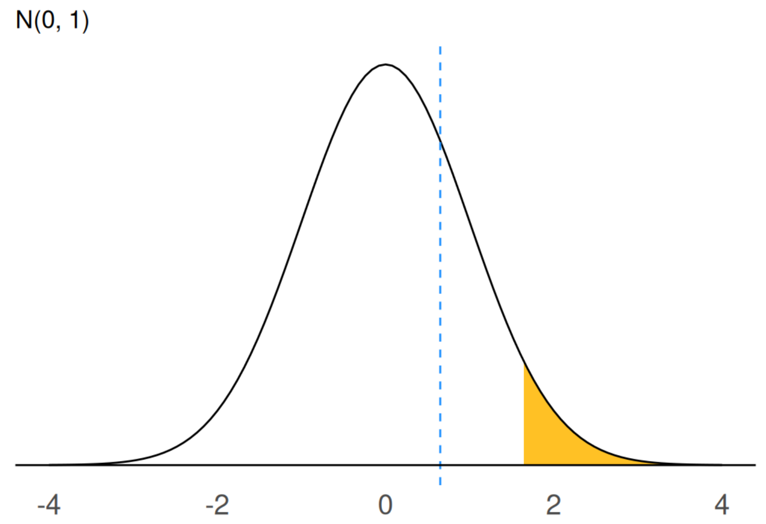
Since we have one-sided hypotheses

$$H_0 : p_1 - p_2 = 0 \quad \text{versus} \quad H_A : p_1 - p_2 > 0$$

our p-value is the area above 0.65 on $N(0, 1)$.

We get a large p-value of 0.258 and fail to reject the null.


N(0, 1)
-4
-2
0
2
4



Or, use a rejection region with the 95th percentile.

Perform a hypothesis test of

$$H_0 : p_1 - p_2 = 0 \quad \text{versus} \quad H_A : p_1 - p_2 \neq 0$$

for chimpanzee B, with $\alpha = 0.05$.

- Use Z test statistic + null distribution
- Now we have *two-sided* hypotheses.

The general hypothesis testing procedure is:

- Write **hypotheses** about parameter
- Identify **null distribution**
- Calculate **test statistic**
- Calculate **p-value** on the null