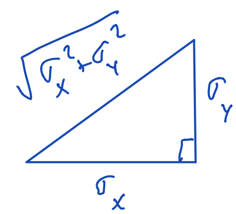
Two-Proportion inference

Thomas Scofield

October 26, 2021



First, a fact: Related to CLT (and other facts from that day)

Theorem: Suppose X and Y are independent variables, and both are normally distributed, with $X \sim \text{Norm}(\mu_X, \sigma_X)$ and $Y \sim \text{Norm}(\mu_Y, \sigma_Y)$. Then their difference $X - Y$ also has a normal distribution, with $(X - Y) \sim \text{Norm}(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2})$.

Two-Proportion context

Imagine you have two groups/populations in mind, and you take *independent* samples, one of size n_1 from Group 1, and one of size n_2 from Group 2. The variable you measure is binary categorical (sex, Christian or not?, have a certain gene or not?). The proportions of *successes* are

- p_1, p_2, – parameters in the two populations
- $\widehat{p}_1$, $\widehat{p}_2$, in the two samples

Note that

- $\widehat{p}_1$, $\widehat{p}_2$ should be independent, since the samples are.
- If the rules-of-thumb
$$n_1p_1 \geq 10 \qquad \text{and} \qquad n_1(1-p_1) \geq 10$$
are met, then
$$\hat{p}_1 \sim \text{Norm}\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right).$$
— From Sections 6.1–6.3
- Likewise, if
$$n_2p_2 \geq 10 \qquad \text{and} \qquad n_2(1-p_2) \geq 10$$
then
$$\hat{p}_2 \sim \text{Norm}\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right).$$

Under these conditions, the theorem tells us

By Theorem
$$\underbrace{\widehat{p}_1 - \widehat{p}_2}_{\text{difference}} \sim \text{Norm}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

This is a statement about the sampling distribution for $\widehat{p}_1 - \widehat{p}_2$—that (under conditions) it is approximately normal. Thus, the *spread* of that sampling distribution is rightly called the **standard error** of $\widehat{p}_1 - \widehat{p}_2$:

$$\text{SE}_{\widehat{p}_1 - \widehat{p}_2} = \underbrace{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}_{\text{Formerly, estimated this quantity using bootstrapping / randomization}}.$$

6.8 Confidence Intervals for $p_1 - p_2$

It's going to be the usual thing:

critical value

$$(\text{point estimate}) \pm (z^*)(\text{SE}_{\widehat{p}_1-\widehat{p}_2})$$

or, adapting to our situation (and the fact that we do not know the values of p_1, p_2):

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

Examples:

1. **One True Love** (see Example 6.19). Here (summarized data)

$$\widehat{p}_f = \frac{363}{1412} \doteq 0.257 \qquad \text{and} \qquad \widehat{p}_m = \frac{372}{1213} \doteq 0.307.$$

You try a 98% CI.

2. **Scolding Crows** (see Data 6.3). Here (summarized data)

$$\widehat{p}_1 = \frac{158}{444} \doteq 0.356 \qquad \text{and} \qquad \widehat{p}_2 = \frac{109}{922} \doteq 0.118.$$

   Group 1 represents the "taggers".

3. **KidsFeet** (available when Mosaic package is loaded). Here, we have raw data on variables `biggerfoot` and `domhand`.

Details

1. $\hat{p}_1 - \hat{p}_2 = 0.257 - 0.307 = -0.05$

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{(0.257)(1-0.257)}{1412} + \frac{(0.307)(0.693)}{1213}} = 0.01762$$

approx.

level of confidence - say 95% $\Longrightarrow$ $z^* = 1.96$.

Then our 95% CI

$$(-0.05) \pm (1.96)(0.01762) \quad \text{or} \quad (-0.085, -0.015)$$

Carrying out the **Scolding Crows** example, we have

- point estimate

```
pointEst <- 158/444 - 109/922
pointEst
```

```
## [1] 0.2376346
```

- standard error

```
se = sqrt(158/444*(1 - 158/444)/444 + 109/922*(1-109/922)/922)
se
```

```
## [1] 0.02508647
```

- z*-value

```
zstar <- qnorm(.99)
zstar
```

```
## [1] 2.326348
```

And the 98% CI is

```
pointEst + c(-1,1)*zstar*se
```

```
## [1] 0.1792747 0.2959944
```

If we use `prop.test()` as a one-stop-shopping method to solve (saving us from the individual calculations)

```
prop.test(c(158,109), c(444,922), conf.level=.98)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c out of c158 out of 444109 out of 922
## X-squared = 106.11, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 98 percent confidence interval:
##  0.1776063 0.2976629
## sample estimates:
##    prop 1    prop 2
## 0.3558559 0.1182213
```