ST 516: Foundations of Data Analytics

Difference in Proportions

Difference in Proportions

- Standard Errors
- Example

Difference in Proportions

When we wish to compare the population proportions for two different populations, it's reasonable to use sample proportions from each of the populations.

Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ are two independent samples of binary data.

- Here, the first subscript denotes the population, 1 or 2; and the second subscript denotes the observations, $1, 2, \ldots, n_1$ for population 1, and $1, 2 \ldots, n_2$ for population 2.
- We will assume that the population proportions are π_1 and π_2 for populations 1 and 2, respectively.
- Finally, we calculate

$$\hat{\pi}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} X_{1j} \text{ and } \hat{\pi}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j}$$

Difference in Proportions

When both of n_1 and n_2 are large, and when the two samples are independent of each other, then the quantity $\hat{\pi}_1 - \hat{\pi}_2$ has a sampling distribution that is approximately Normal.

- The mean of the sampling distribution of $\hat{\pi}_1 - \hat{\pi}_2$ is $\pi_1 - \pi_2$
- The standard deviation of the sampling distribution of $\hat{\pi}_1 - \hat{\pi}_2$ is

$$\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

- Just as in the one-sample case, we will estimate this standard deviation in slightly different ways depending on whether we are performing a hypothesis test or creating a confidence interval.

Standard Errors

Suppose that we wish to test the hypotheses

$$H_0: \quad \pi_1 = \pi_2$$
$$H_A: \quad \pi_1 \neq \pi_2$$

Under the null hypothesis, the two population proportions are assumed to be the same.

This suggests that we use a *pooled* estimate of the sample proportion:

$$\hat{\pi}_c = \frac{\sum_{j=1}^{n_1} X_{1j} + \sum_{j=1}^{n_2} X_{2j}}{n_1 + n_2}$$

Using this, we calculate the standard error of $\hat{\pi}_1 - \hat{\pi}_2$ to be

$$SE_{\hat{\pi}_1 - \hat{\pi}_2} \sqrt{\frac{\hat{\pi}_c(1-\hat{\pi}_c)}{n_1} + \frac{\hat{\pi}_c(1-\hat{\pi}_c)}{n_2}}$$

Standard Errors

For calculating a confidence interval for $\pi_1 - \pi_2$, we make no assumption about their equality.

Therefore, for the confidence interval we use

$$SE_{\hat{\pi}_1 - \hat{\pi}_2} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

There are small sample procedures, like the exact test, that we can use when n_1 and n_2 are not large, but we defer discussion of these to a later time.

Example

Suppose that an internet retailer collects two samples of data—whether or not site visitors made a purchase under two different website designs.

1. What are the two populations?

Example

Suppose that an internet retailer collects two samples of data—whether or not site visitors made a purchase under two different website designs.

1. What are the two populations? Visits under the two different website designs

2. What are the binary observations?

Example

Suppose that an internet retailer collects two samples of data—whether or not site visitors made a purchase under two different website designs.

1. What are the two populations? Visits under the two different website designs
2. What are the binary observations? the sale/no-sale results of the visits
3. Is one website design better than the other in terms of increased sales?

Example

Suppose that an internet retailer collects two samples of data—whether or not site visitors made a purchase under two different website designs.

1. What are the two populations? Visits under the two different website designs
2. What are the binary observations? the sale/no-sale results of the visits
3. Is one website design better than the other in terms of increased sales? We'll test

$$H_0: \quad \pi_1 = \pi_2$$
$$H_A: \quad \pi_1 \neq \pi_2$$

Example

The table below shows the results from this study

	Sale?	
	Yes	No
Website Design A	132	410
Website Design B	161	621

These data result in a 95% confidence interval for $\pi_1 - \pi_2$ of (-0.008,0.084). There is not enough evidence to say that the population proportions are different.

Is there *practical significance* in these results?

Using R

To perform this test in R, use the commands

```
X <- c(132,161)
n <- c((132+410),(161+621))
prop.test(X,n,correct=FALSE)
```

The R output is:

```
##       data:  X out of n
##       X-squared = 2.6347, df = 1, p-value = 0.1046
##       alternative hypothesis: two.sided
##       95 percent confidence interval:
##       -0.008262485  0.083582650
##       sample estimates:
##         prop 1    prop 2
##       0.2435424 0.2058824
```