ST 518: Data Analytics II

Proportions

Vitamin C

Recall the Vitamin C example from the previous lecture:

	Outcome		
	Cold	No Cold	Totals
Placebo	335	76	411
Vitamin C	302	105	407
Totals	637	181	818

The important question of interest here is whether the proportion of people who got colds among the Vitamin C takers is less than the proportion of people who got colds among the placebo takers.

Vitamin C

At the end of the last lecture, we talked about an estimate for the proportion of cold getters among placebo takers:

$$\hat{p}_1 = \frac{335}{411} = 0.815.$$

Similarly, an estimate for the proportion of cold getters among Vitamin C takers is:

$$\hat{p}_2 = \frac{302}{407} = 0.742.$$

Now, these estimates are different, but we'd like to know if those differences are reflected in the general population.

Important note: This was a study based on a sample of volunteers—not a random sample. Therefore, if we are to draw inference to a population, we would have to make a *non-statistical* argument about how well this volunteer sample represents that population.

Differences in Proportions

For the Vitamin C example, it's reasonable to consider the hypotheses:

$$\begin{aligned} H_0: &\quad p_1 = p_2 \\ H_1: &\quad p_1 \neq p_2 \end{aligned}$$

where p_1 and p_2 are the population proportions of cold getters among placebo and Vitamin C takers, respectively.

To evaluate these hypotheses, we need to know about the sampling distributions of $\hat{p}_1 - \hat{p}_2$.

- Recall that the sampling distribution of a statistic is the theoretical histogram of that statistic calculated from repeated samples (of the same size) from the population of interest.

Differences in Proportions

You may recall that for $X \sim \text{bin}(n, p)$ and for large n, the sampling distribution of $\hat{p} = X/n$ is Normal with mean p and variance $np(1-p)$.

Now we have to consider two issues:

1. How large is "large?"
2. What's the sampling distribution of the difference in two proportions?

A good metric to use for determining if the sample size is large enough is to check that $n\hat{p} > 5$ and $n(1-\hat{p}) > 5$.

For the placebo takers

$$
\begin{aligned}
411(0.815) &= 335 > 5 \\
411(1-0.815) &= 76 > 5
\end{aligned}
$$

Differences in Proportions

It turns out that for two samples, if both of n_1 and n_2 are large, then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is Normal with mean $p_1 - p_2$ and we have an expression for the variance.

- You can evaluate whether n_1 and n_2 are large enough in a 2×2 table by verifying that all entries in the table are at least 5.
- In R you can perform the test for a difference in proportions using the function **prop.test**
- It's also instructive to compare this approach to the chi-squared test **chisq.test**

Some R Work

```
> VitC = matrix(c(335,76,302,105),2,2,byrow=T)
> prop.test(VitC)

2-sample test for equality of proportions with continuity
correction

 data:  VitC
 X-squared = 5.9196, df = 1, p-value = 0.01497
 alternative hypothesis: two.sided
 95 percent confidence interval:
  0.01391972 0.13222111
 sample estimates:
    prop 1    prop 2
 0.8150852 0.7420147
```

Comparison with Chi-Squared Test

We'll leave it to you to verify that the R command:

chisq.test(VitC)

gives you the identical inference.

A chi-squared test for homogeneity, is essentially a test to see whether the proportion of "successes" is the same in two or more groups.

And that's exactly what the difference in proportions test checks.

Next steps: Odds and odds ratios are another way to compare properties of categorical/count data distributions.