Lecture 12: Inference for Two Proportions
STAT 630, Fall 2021

Let $X_1, X_2, \cdots, X_{n_1}$ be independent random variables from Bern(p_1), and $Y_1, Y_2, \cdots, Y_{n_2}$ be independent random variables from Bern(p_2). Assume that the two samples are independent of each other.

The sample proportions are defined as:
$\hat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ and $\hat{p}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$

Find the expectation and variance for the difference between two proportions:

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$
$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + (-1)^2 \cdot Var(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

According to the central limit theorem, the sampling distribution for the difference between two proportions can be approximated by a normal distribution:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

The approximation should only be used if $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$. That is, there are at least that 10 “successes” and “failures” for each group.

$1-\alpha$ confidence interval for $p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Conditions:

- $n_i \hat{p}_i \geq 10$ and $n_i(1-\hat{p}_i) \geq 10$ for $i = 1, 2$
- The data are independent within and between groups. Generally this is satisfied if the data come from two independent random samples, or if the data come from a randomized experiment.

Ex1: A 5-year study was conducted to evaluate the effectiveness of fish oils on reducing cardiovascular events, where each subject was randomized into one of two treatment groups. We'll consider heart attack outcomes in these patients:

	heart attack	no event	Total
fish oil	145	12788	12933
placebo	200	12738	12938

(a) Is this study an experiment or an observational study?
Experiment

(b) Are the conditions for inference satisfied?
Yes, the "success" / "failure" condition is satisfied since the counts in each cell of the contingency table are greater than 10. That is, there are more than 10 outcomes (hear attack / no event) for the fish oil and placebo groups. The independence condition is satisfied since the data come from a randomized experiment.

(c) Calculate a 95% confidence interval for the difference in heart attack rates of patients that take fish oils and patients that take a placebo. Let's use R to do this:

```
n_fish <- 12933
phat_fish <- 145 / 12933; phat_fish

## [1] 0.01121163

n_placebo <- 12938
phat_placebo <- 200 / 12938; phat_placebo

## [1] 0.01545834

z_crit <- qnorm(0.975); z_crit

## [1] 1.959964

SE <- sqrt(phat_fish*(1-phat_fish)/n_fish + phat_placebo*(1-phat_placebo)/n_placebo)
ci_lower <- phat_fish - phat_placebo - z_crit * SE
ci_upper <- phat_fish - phat_placebo + z_crit * SE
round(c(ci_lower, ci_upper), 4)

## [1] -0.0070 -0.0015
```

(d) Interpret the confidence interval in the context of the data.

A 95% confidence interval for $p_1 - p_2$ is $(-0.007, -0.0015)$.

We are 95% confident that the heart attack rate for patients that take fish oils is between 0.7% and 0.15% lower than patients that take a placebo.

The result is *statistically significant* since 0 is not inside the confidence interval. However, the result may not be *practically significant* since the effect fish oils have on reducing heart attacks is small (less than a 0.15% reduction in the heart attack rate when compared to taking a placebo).

Hypothesis test for the difference between two proportions

Null and alternative hypothesis:
H_0: $p_1 = p_2$ (the two population proportions are the same)
H_A: $p_1 > p_2$ or $p_1 < p_2$ or $p_1 \neq p_2$

Test statistic: Assuming H_0 is true, the two proportions are equal: $p_1 = p_2 = p$. Under this assumption, the variance is given by

$$\begin{aligned} Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \\ &= \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \end{aligned}$$

We estimate p with the pooled proportion:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Conclusively, the z-test statistic is given by

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

Once the z-test statistic is computed, the p-value can be computed to decide whether to reject or not reject the null hypothesis.

The conditions are the same as for confidence intervals except we use the pooled proportion $\hat{p}$ when checking the number successes and failures (i.e., $n_i\hat{p} \geq 10$ and $n_i(1-\hat{p}) \geq 10$ for $i = 1, 2$).

Ex2: Are young people more concerned with environmental issues than older people? The Pew Research Center for the People and the Press conducted a survey in 2009 asking a random sample of adults whether or not there is solid evidence of global warming. Of the 197 people aged 18–29 years old who responded, 126 said yes, compared to 223 out of 406 people aged 30–39 years old. Does this indicate that a higher proportion of younger people believe there is evidence of global warming? [1]

$H_0 : p_1 = p_2$
$H_A : p_1 > p_2$

$\hat{p}_1 = 126/197 = 0.64$, $n_1 = 197$
$\hat{p}_2 = 223/406 = 0.55$, $n_2 = 406$
The pooled proportion: $\hat{p} = (126 + 223)/(197 + 406) = 0.58$

$$z = \frac{0.64 - 0.55}{\sqrt{0.58(0.42)(1/197 + 1/406)}} = 2.1$$

p-value $= P(Z > 2.1) =$ `1 - pnorm(2.1)` $= 0.018$

Since the p-value < 0.05, we reject the null hypothesis. The data provide strong evidence that the population proportion of 18-29 year olds that believe in global warming is higher than 30-39 year olds.

[1] `http://www.pewresearch.org/2009/10/22/fewer-americans-see-solid-evidence-of-global-warming/`