Unit III

ECON 3406

Dr. Josh Martin

Differences of Two Proportions

- Like with $\hat{p}$, the difference of two sample proportions $\hat{p}_1 - \hat{p}_2$ can be modeled using a normal distribution (when conditions are met).

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

- This standard error comes from the fact that variances of independent variables **add**, even when subtracting.

Differences of Two Proportions: Standard Errors

- When we talk about the spread of an estimate, we're really talking about **variance** (the square of the standard error).
- If two random variables **A** and **B** are independent, then:

$$\mathrm{Var}(A - B) = \mathrm{Var}(A) + \mathrm{Var}(B)$$

- This might seem counterintuitive — but remember:
  - Even if you're subtracting two noisy measurements, the **uncertainty (noise)** from both still adds up.
  - Think of it like using two shaky rulers. Subtracting doesn't cancel the shakiness — it just combines it!

Differences of Two Proportions: Simulation Setup

We'll use simulation to understand the sampling distribution of sample proportions for two independent groups.

- Group 1 has a true proportion $p_1 = 0.5$
- Group 2 has a true proportion $p_2 = 0.4$
- Each group has $n = 500$ individuals per sample
- We'll repeat this sampling 1,000 times to observe variation in sample means

Differences of Two Proportions: Simulation

```r
set.seed(1)       # ensures reproducibility
B <- 1000          # number of simulations
n <- 500           # sample size per group
p1 <- 0.5          # true proportion in group 1
p2 <- 0.4          # true proportion in group 2

# Create empty vectors to store simulated means and S
mean_x1 <- mean_x2 <- numeric(B)
sd_x1 <- sd_x2 <- numeric(B)

# Loop to simulate samples for both groups
for (i in 1:B) {
  # Generate random binary outcomes for group 1 (succ
  x1 <- rbinom(n, size = 1, prob = p1)
  mean_x1[i] <- mean(x1)  # sample proportion for gro
  sd_x1[i]   <- sd(x1)    # sample SD for group 1

  # Repeat for group 2
  x2 <- rbinom(n, size = 1, prob = p2)
```

Comparing Theoretical and Empirical Standard Errors

- We can compare:
  - the **theoretical** standard error (from the formula)
  - the **empirical** standard error (from our simulations)

Comparing Theoretical and Empirical Standard Errors

```
# Theoretical standard error for a sample proportion
sqrt(p1*(1-p1)/n)
```

```
[1] 0.02236068
```

```
# Empirical standard error from simulated data
mean(sd_x1) / sqrt(n)
```

```
[1] 0.0223613
```

- The first line gives the theoretical SE: $\sqrt{p(1-p)/n}$
- The second line gives the empirical SE, based on simulated SDs
- These values should be nearly identical, validating the normal approximation for large n

Sampling Distribution for One Group (p1 = 0.5)

- We can visualize the distribution of sample proportions across simulations, and overlay a 95% confidence interval around the true mean.

Sampling Distribution for One Group (p1 = 0.5)

Sampling Distribution for One Group (p1 = 0.5)

- Roughly 5% of simulated sample proportions should fall outside this interval — confirming the 95% confidence level's interpretation.

```
# Calculate the proportion of estimates that fall outs
mean(ifelse(mean_x1 >= ub_x1 | mean_x1 <= lb_x1, 1, 0)
```

```
[1] 0.052
```

Sampling Distribution for One Group (p2 = 0.4)

Sampling Distribution for One Group (p2 = 0.4)

- Again, around 5% of simulated estimates will fall outside the interval.
- The spread is slightly narrower than for Group 1 because the variance is smaller.

```
# Share of points outside the 95% CI
mean(ifelse(mean_x2 >= ub_x2 | mean_x2 <= lb_x2, 1, 0)
```

```
[1] 0.048
```

Combining Two Proportions

- Now that we understand the sampling variation of each group separately, we can combine them just as we would when estimating a **difference in proportions**:

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2}$$

- This formula reflects that variances add, even though we're subtracting proportions.

Simulated Differences Between Groups

```r
# Compute simulated differences
diff <- mean_x1 - mean_x2

# Theoretical SE for the difference in proportions
sqrt(p1*(1-p1)/n + p2*(1-p2)/n)
```

```
[1] 0.03130495
```

```r
# Empirical SE estimate (not exact but illustrative)
mean(sd_x1 + sd_x2) / sqrt(n + n)
```

```
[1] 0.03129505
```

Simulated Differences Between Groups

Simulated Differences Between Groups

```
# Check coverage rates for both formulas
mean(ifelse(diff >= ub_diff_1 | diff <= lb_diff_1, 1,
```

```
[1] 0.037
```

```
mean(ifelse(diff >= ub_diff_2 | diff <= lb_diff_2, 1,
```

```
[1] 0.793
```

Differences of two proportions: Example 1

- Consider an experiment involving patients who underwent cardiopulmonary resuscitation (CPR) following a heart attack and were subsequently admitted to a hospital.
  - Patients were randomly assigned to either a **treatment group** (received a blood thinner) or a **control group** (no blood thinner).
  - The outcome of interest was **survival for at least 24 hours**.

Differences of two proportions: Example 1

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Differences of two proportions: Example 1

- Create and interpret a **90% confidence interval** of the difference for the survival rates in the CPR study.
  - $p_t - p_c = 0.35 - 0.2 = 0.13$
  - $SE = \sqrt{\frac{0.35(1-0.35)}{40} + \frac{0.22(1-0.22)}{50}} \approx 0.095$
  - $0.13 \pm 1.645 \times 0.095 = (-0.027, 0.287)$

Differences of two proportions: Computing in R

```r
pt <- 14/40
pc <- 11/50
nt <- 40
nc <- 50

point_est <- pt - pc
se <- sqrt((pt * (1 - pt) / nt) + (pc * (1 - pc) / nc)

z <- data.frame(
  sig_level = c(0.01, 0.05, 0.1),
  z_score = c(2.45, 1.95, 1.645)
)

z$min <- point_est - z$z_score * se
z$max <- point_est + z$z_score * se
z
```

```
  sig_level z_score         min       max
1      0.01   2.450 -0.10396538 0.3639654
2      0.05   1.950 -0.05621734 0.3162173
3      0.10   1.645 -0.02709104 0.2870910
```

Differences of two proportions: Visualizing Confidence Intervals

Differences of two proportions: Interpretation

- We are **90% confident** that blood thinners change the 24-hour survival rate by between -3 and 29 percentage points for patients similar to those in the study.
- Because **0% is within this range**, the evidence is inconclusive — we cannot determine whether blood thinners help or harm heart attack patients who have undergone CPR.

Differences of Two Proportions: Example 2

- A 5-year clinical trial evaluated whether **fish oil supplements** reduce the risk of **heart attacks**.
- Each participant was randomly assigned to one of two groups:
  - **Fish Oil group**
  - **Placebo group**
- We'll examine heart attack outcomes across both groups.

Differences of Two Proportions: Example 2

Group	Heart Attack	No Event	Total
Fish Oil	145	12,788	12,933
Placebo	200	12,738	12,938

Differences of Two Proportions: Example 2

- Construct a **95% confidence interval** for the effect of fish oil on heart attack incidence among patients represented by this study.
- Interpret the interval in context:
  - What does the direction and width of the interval suggest?
  - Is there evidence that fish oil has a meaningful effect on heart attack risk?

Differences of two proportions: Computing in R

```r
nt <- 12933
nc <- 12938

pt <- 145 / nt
pc <- 200 / nc

point_est <- pt - pc
se <- sqrt((pt * (1 - pt) / nt) + (pc * (1 - pc) / nc)

z <- data.frame(
  sig_level = c(0.01, 0.05, 0.1),
  z_score = c(2.45, 1.95, 1.645)
)

z$min <- point_est - z$z_score * se
z$max <- point_est + z$z_score * se
```

Differences of two proportions: Visualizing Confidence Intervals

Differences of two proportions: Interpretation

- The **point estimate** for the effect of fish oil is approximately **−0.0043**, meaning heart attacks occurred **0.43 percentage points less often** in the fish-oil group than in the placebo group.
- We are **90% confident** that fish oil changes the heart-attack rate by between **−0.66 and −0.19 percentage points** for patients similar to those in the study.
- Because this interval **does not include 0**, the reduction in heart-attack risk is **statistically significant** at the 10% (and even 5% an 1%) level.

Practical vs. Statistical Significance

- While statistically significant, the **effect size is extremely small** — roughly **0.4 fewer heart attacks per 100 individuals**.
- In a large clinical sample, even minor effects can reach significance if variability is low.
- From a **practical** standpoint, such a small reduction may **not justify** the cost, side effects, or adherence burden of treatment.

More on Two-Proportion Hypothesis Tests

- When conducting a two-proportion hypothesis test, the null hypothesis is typically: $H_0 : p_1 - p_2 = 0$
- However, there are cases where we may want to test for a *specific* difference other than zero.
  - For example, suppose we want to test whether: $H_0 : p_1 - p_2 = 0.10$
- In contexts like these, we use the sample proportions

$$\hat{p}_1$$

and

$$\hat{p}_2$$

to check the success–failure condition and to construct the standard error.

Differences of Two Proportions: Example 3

- A drone quadcopter company is considering a new manufacturer for rotor blades.
- The new manufacturer is more expensive but claims that their higher-quality blades are **3% more reliable**, meaning that 3% more blades pass inspection compared to the current supplier.
- Set up the appropriate hypotheses for this test:
  - $H_0 : p_{\text{highQ}} - p_{\text{standard}} = 0.03$
  - $H_A : p_{\text{highQ}} - p_{\text{standard}} \neq 0.03$

Differences of Two Proportions: Example 3 (Data)

- A quality control engineer collects samples of 1,000 blades from each manufacturer:
  - Current supplier: 899 blades pass inspection
  - Prospective supplier: 958 blades pass inspection
- Using these data, evaluate the hypotheses above at a significance level of 5%.

Compute the Point Estimate and Standard Error

```r
p_us <- 958 / 1000
p_them <- 899 / 1000
point_est <- p_us - p_them

# Standard error for independent samples
se <- sqrt( p_us * (1 - p_us) / 1000 +  p_them * (1 -

p_us; p_them
```

```
[1] 0.958
```

```
[1] 0.899
```

```r
point_est
```

```
[1] 0.059
```

```r
se
```

```
[1] 0.01144705
```

Compute Confidence Intervals for Various Significance Levels

```r
z <- data.frame(
  sig_level = c(0.01, 0.05, 0.10),
  z_score = c(2.45, 1.95, 1.645)
)

z$min <- (point_est - z$z_score * se) - 0.03
z$max <- (point_est + z$z_score * se) - 0.03
z
```

```
  sig_level z_score          min        max
1      0.01   2.450 0.0009547225 0.05704528
2      0.05   1.950 0.0066782486 0.05132175
3      0.10   1.645 0.0101695994 0.04783040
```

Visualizing Confidence Intervals

Compute and Visualize the z-Statistic

```r
z <- (point_est - 0.03) / se

set.seed(1)
sim <- rnorm(1000, mean = 0.03, sd = se)

# Probability of observing a value this extreme or lar
1 - mean(ifelse(point_est >= sim, 1, 0))
```

```
[1] 0.004
```

```r
# p-value (right-tailed)
1 - pnorm(z)
```

```
[1] 0.005648044
```

Visualizing the Sampling Distribution

Example 3: Conclusion

- From the standard normal distribution:
  - The right-tail area is approximately 0.004
  - Doubling for a two-tailed test gives $p = 0.008$
  - Since $p = 0.008 < 0.05$, we **reject the null hypothesis**
- We find statistically significant evidence that the higher-quality blades have a pass rate greater than 3% higher than the standard blades — exceeding the company's claims.

Chi-Squared Distributions: Introduction

- χ = the greek letter for “chi” (pronounced like “kai”)
- The χ^2 distribution is a continous probability distribution that is widely used in statistical inference.
  - Closely related to the standard normal distribution
- If a variable Z has the standard normal distribution, then Z^2 has the χ^2 distribution with one **degree of freedom**

Chi-Squared Distributions: Histograms

Chi-Squared Distributions: Definition

- If $Z_1, Z_2, \ldots, Z_k$ are independent standard normal variables, then…
  - $Z_1^2 + Z_2^2 + \ldots + Z_k^2$
  - …has a χ^2 distribution with k **degrees of freedom**.

Degrees of Freedom: Concept

- A **degree of freedom (df)** represents the number of **independent pieces of information** available to estimate something.
- Whenever we calculate a statistic, we "use up" some information.
  - For example, once we estimate the sample mean, one data point can be perfectly predicted from the others.
  - So for a sample of size n, only $(n-1)$ observations are *free to vary* when computing the sample variance.

Degrees of Freedom: Intuition

- Average (mean): $\bar{x} = \frac{1}{n}\sum_i^n x_i = \frac{x_1 + \ldots + x_n}{n}$
- if $i = 4, x_1 = 8, x_2 = 10, x_3 = 12$ and $\bar{x} = 10$...
  - ... then $x_4 = 10$
- So even though we had four data points, only three were free to vary — the fourth is determined by the mean.
- That's why when calculating the sample variance, we divide by n instead of $n - 1$: one degree of freedom has been "used up" in estimating the mean.

Degrees of Freedom: Motivation

- **Why it matters:**
  - Degrees of freedom tell us **how much independent information** our test or estimate is based on.
  - They affect the **shape of sampling distributions** (like t, F, and χ^2), which in turn changes the critical values and p-values we use.
  - More degrees of freedom → more information → the distribution becomes narrower and more normal-looking.
- **Big idea:**
  - Degrees of freedom link **sample size**, **uncertainty**, and the **reliability of inference** — they remind us that every time we estimate something, we "spend" information.

Chi-Squared Distributions: Properties

- mean: $\mu = k$
- variance: $\sigma^2 = 2k$
- mode occurs at $\mu - 2$

```r
set.seed(1)
x <- rnorm(1000, mean = 0, sd = 1)
x2 <- x^2
mean(x2)
```

```
[1] 1.070115
```

```r
var(x2)
```

```
[1] 2.291541
```

Chi-Squared Distributions: Multiple Degrees of Freedom

Chi-Squared Distributions: Multiple Degrees of Freedom

```r
n_loops <- 2
z <- list()
for(i in 1:n_loops){
  set.seed(i)
  x <- rnorm(1000, mean = 0, sd = 1)
  x2 <- x^2
  if(i == 1){
    z2 <- x2
  }else{
    z2 <- z2 + x2
  }
}
mean(z2); n_loops
```

```
[1] 2.103046
```

```
[1] 2
```

```r
var(z2); 2*n_loops
```

```
[1] 4.281053
```

```
[1] 4
```

Chi-Squared Distributions: Multiple Degrees of Freedom

```r
n_loops <- 20
z <- list()
for(i in 1:n_loops){
  set.seed(i)
  x <- rnorm(1000, mean = 0, sd = 1)
  x2 <- x^2
  if(i == 1){
    z2 <- x2
  }else{
    z2 <- z2 + x2
  }
}
mean(z2); n_loops
```

```
[1] 20.08944
```

```
[1] 20
```

```r
var(z2); 2*n_loops
```

```
[1] 40.06536
```

```
[1] 40
```

Chi-Squared Distributions: Multiple Degrees of Freedom