Public Policy 529
Comparison of Two Groups

Part 1

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

October 25, 2023

Outline

1. Recap Difference of Means Test

2. Difference of Proportions

3. Comparing Means With Dependent Samples

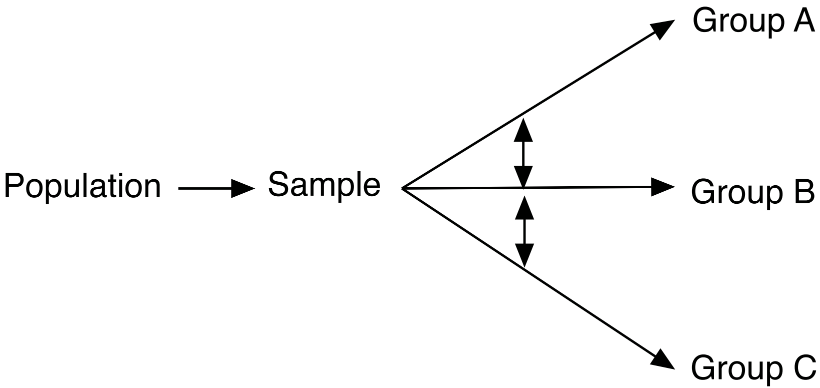
Outline

1. Recap Difference of Means Test

2. Difference of Proportions

3. Comparing Means With Dependent Samples

Our Goal When Performing Comparisons



We want to create comparison groups with different values of the independent variable. Test whether the dependent variable also differs across these groups.

The Big Picture

- We are interested in whether there is a difference between means from two populations: μ_1 and μ_2.
- We estimate the difference using samples to obtain $\bar{y}_1$ and $\bar{y}_2$.
- From these samples we calculate the difference: $\bar{y}_2 - \bar{y}_1$. This is a statistic that has a standard error.
- We test whether this statistic is different from H_0 with sufficient confidence.

Infant Mortality and Democracy Example

Regime Type	Mean	s	n	se
Non-democracy	40.0	28.2	75	3.26
Democracy	20.7	21.9	101	2.18
Overall	28.9	26.5	176	2.00

We can perform a t-test for the difference between the means.

Steps for Difference of Means Test

1. Determine type of comparison.
   - Do we have independent or dependent samples?
   - Can we assume the variances of the two populations are equal? (usually, no)
2. Calculate the difference of means.
3. Calculate the standard error (the formula depends on the answer to the questions above).
4. Calculate degrees of freedom (formula depends) and critical value of test statistic.
5. Calculate test statistic, p-value, and apply decision rule.

Testing for the Difference of Means (Infant Mortality)

Regime Type	Mean	se
Non-democracy	40.0	3.26
Democracy	20.7	2.18
Difference	19.3	?

These are independent samples. Can we assume infant mortality has equal variance in the populations of democracies and non-democracies? Probably not wise.

The Standard Error of the Difference

- The standard error of the difference is a function of the standard errors for each mean.

- For independent samples, and no assumption of equal variances, the basic formula is:

$$
\begin{aligned}
se_{\textsf{diff}} &= \sqrt{(se_1)^2 + (se_2)^2} \\
&= \sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2} \\
&= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\end{aligned}
$$

Applying the Formula

Quick version:

$$\begin{aligned} se_{\text{diff}} &= \sqrt{(se_1)^2 + (se_2)^2} \\ &= \sqrt{3.26^2 + 2.18^2} \\ &= 3.92 \end{aligned}$$

Full version:

$$\begin{aligned} se_{\text{diff}} &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{28.2^2}{75} + \frac{21.9^2}{101}} \\ &= \sqrt{10.60 + 4.75} \\ &= 3.92 \end{aligned}$$

Degrees of Freedom for Independent Samples/Unequal Variances

- The correct degrees of freedom is somewhere in the range from $\min(n_1 - 1, n_2 - 1)$ to $n_1 + n_2 - 2$.

- The best formula is quite complex. Let your software do it.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

- Using $\min(n_1 - 1, n_2 - 1)$ is a safe approximation, though it is likely too conservative.

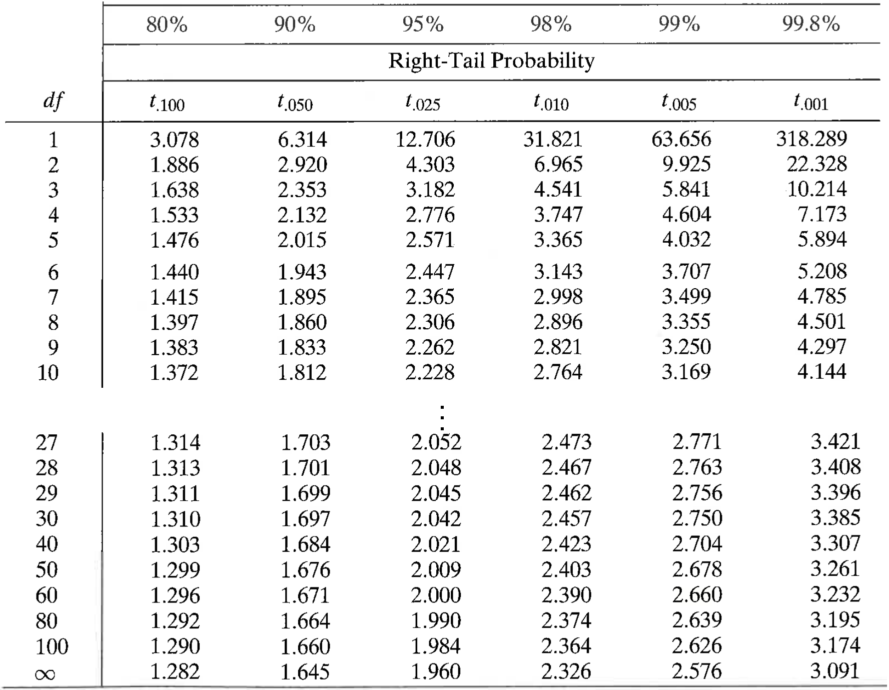
Confidence Interval

In our example, $\min(n_1 - 1, n_2 - 1)$ is 74. The long formula produces 174 df.

Using the t-table to find the critical value for a 95% confidence interval, we follow the 60 df line to be safe. The value is 2.000.

$$\begin{aligned} ci &= (\bar{y}_2 - \bar{y}_1) \pm t\sqrt{(se_1)^2 + (se_2)^2} \\ &= (40.0 - 20.7) \pm 2.0\sqrt{3.26^2 + 2.18^2} \\ &= 19.3 \pm 2.0 \cdot 3.92 \\ &= 19.3 \pm 7.8 \end{aligned}$$

The 95% confidence interval for the difference of means is 11.5 to 27.1.



	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
			⋮			
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.091

Significance Test

Recall that:

$$\begin{aligned} H_0 : \mu_2 - \mu_1 &= 0 \\ H_A : \mu_2 - \mu_1 &\neq 0 \end{aligned}$$

Our test statistic is:

$$t = \frac{\text{estimate} - H_0}{\text{se of the estimate}} = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se_{\text{diff}}} = \frac{19.3}{3.92} = 4.92$$

We can reject H_0 with very high confidence. (Find the p-value)

Using R

Performing a difference of means test with independent samples and unequal variances. A dichotomous variable makes the categories.

- Use the t.test function. The default is unequal variances. To change that, use the argument `var.equal = TRUE`.

  ```
  t.test(interval_var ~ dichotomous_var)
  ```

- To get t-statistic associated with a particular degrees of freedom and α:

  qt($\alpha/2$, *df*, lower.tail = F)

Using Stata

Performing a difference of means test with independent samples and unequal variances.

- The ttest command with the option unequal (for unequal variances). Unfortunately, the default is for equal variances.

  ```
  ttest interval_var, by(dichotomous_var) unequal
  ```

- To get t-statistic associated with a particular degrees of freedom and α:

  display invttail(***df***, $\alpha/2$)

See guide posted on Canvas.

Summary

- The basic procedures for significance tests and confidence intervals are the same.
- We just need to get the right standard error for our test statistics and find the correct degrees of freedom.
- Next: tests with dependent samples, difference of sample proportions, and other methods.

Outline

1. Recap Difference of Means Test

2. Difference of Proportions

3. Comparing Means With Dependent Samples

Difference of Proportions

We want to know whether there is a difference between π_1 and π_2, the proportions from two populations.

- We estimate the difference using samples to obtain $\hat{\pi}_1$ and $\hat{\pi}_2$.
- The difference, $\hat{\pi}_2 - \hat{\pi}_1$, is a sample statistic with a standard error.
- We test whether it is different from H_0 with sufficient confidence.
- Rough rule of thumb: we need at least 10 cases in each category in each sample.

Working Example: Education Program

Suppose we are examining the results of a program to boost the rate of high-school graduation. The data:

	Participant?		
Graduated?	Yes	No	Total
Yes	83	102	185
No	10	19	29
Total	93	121	214
Proportion	.892	.843	.864

Finding the Correct Standard Error

Key: the formula for the standard error of the difference differs for confidence intervals and hypothesis tests.

- The same thing is true with one-sample proportions.
- For confidence intervals, we follow the general formula:

$$se_{\textsf{diff}} = \sqrt{(se_1)^2 + (se_2)^2}$$

- For hypothesis tests, we construct a standard error based on the theoretical proportion represented by the null hypothesis.

Standard Error for Confidence Intervals

Recall that the formula for the standard error of a sample proportion is:

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Therefore,

$$\begin{aligned} se_{\textsf{diff}} &= \sqrt{(\hat{\sigma}_{\hat{\pi}_1})^2 + (\hat{\sigma}_{\hat{\pi}_2})^2} \\ &= \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \end{aligned}$$

Formula for the Confidence Interval

Since, for proportions, we use a z-statistic for samples that are (relatively) large, the formula for the confidence interval for the difference of proportions is:

$$\begin{aligned} c.i. &= (\hat{\pi}_2 - \hat{\pi}_1) \pm z(se_{\mathsf{diff}}) \\ &= (\hat{\pi}_2 - \hat{\pi}_1) \pm z\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \end{aligned}$$

Example: Finding the Confidence Interval

For the example of the program to boost high school graduation rates, the 95% confidence interval would be:

$$
\begin{aligned}
c.i. &= (\hat{\pi}_2 - \hat{\pi}_1) \pm z\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \\
&= (.892 - .843) \pm 1.96\sqrt{\frac{.892(1-.892)}{93} + \frac{.843(1-.843)}{121}} \\
&= .049 \pm .09
\end{aligned}
$$

The confidence interval for the difference is -.04 to .14. Note: the interval includes 0. What does that tell us?

Standard Error for Hypothesis Tests

- With significance tests, we use the null hypothesis as a baseline assumption.
- Typically, with a difference of proportions test, the null hypothesis is that the proportions are the same.
- Accordingly, the standard error should reflect the baseline assumption.
- We need to calculate $\hat{\pi}$, the proportion consistent with the null hypothesis.

Standard Error for Hypothesis Tests

- For $\hat{\pi}$, we use the proportion calculated from the combined samples.
- In other words, combine the 2 samples into 1, and calculate the overall proportion $\hat{\pi}$.
- Use $\hat{\pi}$ in place of $\hat{\pi}_1$ and $\hat{\pi}_2$ in the formula for the standard error.

$$se_0 = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}$$

Example: Finding $\hat{\pi}$

What is the baseline assumption for the graduation rate if the program made no difference?

	Participant?		
Graduated?	Yes	No	Total
Yes	83	102	185
No	10	19	29
Total	93	121	214
Proportion	.892	.843	.864

The overall proportion that graduated, $\frac{185}{214} = .864$, is $\hat{\pi}$.

Example: Finding se_0

Next, use $\hat{\pi} = .864$ in the formula for the standard error.

$$\begin{aligned} se_0 &= \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} \\ &= \sqrt{\frac{.864(1-.864)}{121} + \frac{.864(1-.864)}{93}} \\ &= .047 \end{aligned}$$

Example: Significance Test

We can now calculate the z-statistic for the difference of proportions test. Let $H_0 = 0$, for no difference.

$$\begin{aligned} z &= \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{se_0} \\ &= \frac{.892 - .843}{.047} \\ &= 1.049 \end{aligned}$$

With a z-statistic of 1.049, we cannot reject the null hypothesis of no difference. 1.049 is less than the critical value of 1.96.

Using Software for Difference of Proportions

In Stata:

```
prtest varname, by(categorical_variable)
```

Things are a bit trickier in R, which will do a χ^2 test:

```
prop.test(c(#cat1, #cat2), c(n1, n2))
```

- Where `#cat1` is the number of cases in the category of interest within sample 1, and `n1` is the size of sample 1.
- Taking the square root of the resulting χ^2-statistic will provide the z-statistic when we have 2 categories.

See handout on Canvas for comparisons of two groups in Stata or the R Help document.

Outline

1. Recap Difference of Means Test

2. Difference of Proportions

3. Comparing Means With Dependent Samples

Dependent Samples: The Basics

We have dependent samples when each observation in sample 1 has a counterpart observation in sample 2.

- e.g. pre- and post-experiment measurements of a single sample.
- e.g. studies involving sets of twins.

With dependent samples, we control for other factors that might explain differences.

We use the matched pairs to convert a two-sample problem into a one-sample problem.

Matched Pairs

Each observation is the difference between the paired items in the samples (of size n).

$$y_{di} = (y_i \text{ in sample } 2) - (y_i \text{ in sample } 1)$$

This gives us n measured differences: $y_{d1}, y_{d2} \ldots y_{dn}$

The mean of these differences, $\bar{y}_d$, is our statistic of interest.

Mathematically, $\bar{y}_d$ is the same number as $\bar{y}_2 - \bar{y}_1$: the mean of the differences = the difference of means.

Example: Dependent Samples

Basically, y_d is treated like a single sample.

	y_2	y_1	y_d
Pair 1	4	2	2
Pair 2	3	3	0
Pair 3	5	3	2
Pair 4	6	7	-1
Means	4.5	3.75	0.75

Note: $\bar{y}_2$ - $\bar{y}_1 =$ .75. The sample standard deviation of y_d is 1.5.

The standard error of $\bar{y}_d$

We then calculate the standard error of $\bar{y}_d$ with the familiar formula:

$$\hat{\sigma}_{\bar{y}_d} = \frac{s_d}{\sqrt{n}}$$

where s_d is the (usual) sample standard deviation of y_d:

$$\begin{aligned} s_d &= \sqrt{\frac{(y_{d1} - \bar{y}_d)^2 + (y_{d2} - \bar{y}_d)^2 + \ldots + (y_{dn} - \bar{y}_d)^2}{n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(y_{di} - \bar{y}_d)^2}{n-1}} \end{aligned}$$

Confidence Intervals for Paired Differences

At this point, there is nothing different from what we already know. The formula for the confidence interval is:

$$\bar{y}_d \pm t(se_d)$$

$$\bar{y}_d \pm t\left(\frac{s_d}{\sqrt{n}}\right)$$

Significance Test for Paired Differences

We use the regular formula for the test statistic:

$$t = \frac{\text{estimate } - H_0}{\text{se of estimate}}$$

$$= \frac{\bar{y}_d - 0}{\hat{\sigma}_{\bar{y}_d}}$$

With a sufficiently large test statistic, we would reject the null hypothesis of no difference between the paired samples.

Example: Lead Contamination in Water

Suppose we want to test whether measures to reduce the level of lead in drinking water have been working.

- We take a sample of tap water from 250 homes.
- We come back a few months later to take another sample from exactly the same homes.
- These are dependent samples because they consist of a set of matched pairs.
- We can then create a single sample that contains the differences between the two lead level measurements for each house.

Lead Example continued

Suppose that the mean of sample of differences is -3.2 parts per billion with a standard deviation of 23.7.

$$\begin{aligned} t &= \frac{\bar{y}_d - 0}{\hat{\sigma}_{\bar{y}_d}} \\ &= \frac{-3.2 - 0}{\frac{23.7}{\sqrt{250}}} \\ &= -2.13 \end{aligned}$$

With 249 degrees of freedom, the critical value of the t-statistic is 1.97. We could reject the null hypothesis that there is no difference in lead levels.

Using Software

In Stata, the ttest command can also be used for paired samples. The specific form differs.

```
    ttest varname1 = varname2
```

To use this method, the data for the two samples are stored in separate variables.

Things work in an equivalent manner in R:

```
    t.test(data$varname1, data$varname2)
```