Public Policy 529
Comparison of Two Groups
Part 2

Jonathan Hanson

Gerald R. Ford School of Public Policy
University of Michigan

October 30, 2023

Recap: Comparing Two Groups

- Typically, we test whether two means, or two proportions, are different from each other with statistical confidence.
- Our baseline assumption is that the difference between these means or proportions is 0.
- If the difference of means/proportions in our samples is large enough, we reject the null hypothesis of no difference.
- The difficulty is getting the right standard errors and degrees of freedom for these tests.

Difference of Means Tests

Key questions to ask:

- Are the samples dependent or independent?
- Can we assume the variances of the populations that produce these samples are equal?

The answers to these questions determine the formula we use for our standard errors and how we calculate degrees of freedom.

Difference of Proportions Tests

Key questions to ask:

- Are the samples dependent or independent?
- Is the sample large enough to use the z distribution? i.e. are there at least 10 cases in each category?

If not, we need to use a small-sample method, like Fisher's Exact test.

Outline

1. Means: Independent Samples Assuming Equal Variances
2. Small-Sample Tests for Proportions
3. Non-Parametric Tests for Small-Sample Means
4. Controlled Comparison of Means

Outline

Mean Comparison: Equal Variances

- This method makes the assumption that the variances of the populations that produce our samples are equal.
- Under the assumption, if one sample is very small, pooling with the larger sample may help obtain a better estimate of the standard error.
- Additionally, making this assumption permits simplification of our degrees of freedom calculation:

$$df = n_1 + n_2 - 2$$

Meaning of Equal Variances

- When we draw a sample from a population, we estimate the variance σ^2 with s^2.
- If we draw samples from two populations, we then have two estimated variances:

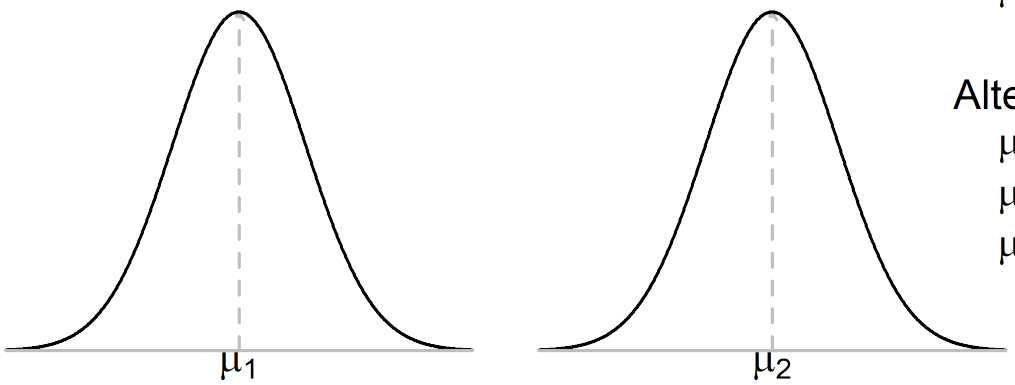
$$s_1^2 \longrightarrow \sigma_1^2$$

$$s_2^2 \longrightarrow \sigma_2^2$$

- But if $\sigma_1^2 = \sigma_2^2$, then s_1^2 and s_2^2 are estimating the same σ^2.
- We thus pool, or combine, our two sample variances to increase our degrees of freedom for estimating σ^2.

Two sample t: equal variances assumed

population distributions



Null hypothesis:
$\mu_1 = \mu_2$

Alternative hypothesis
$\mu_1 \neq \mu_2$ (two sided)
$\mu_1 > \mu_2$ (right sided)
$\mu_1 < \mu_2$ (left sided)

sample data

36	41	46	35
40	41	43	39

68	59	62	64
58	56	67	51

$\bar{y}_1 = 40.1$ $\quad \bar{y}_2 = 60.6$
$s_1 = 3.6$ $\quad s_2 = 5.8$
$n_1 = 8$ $\quad n_2 = 8$

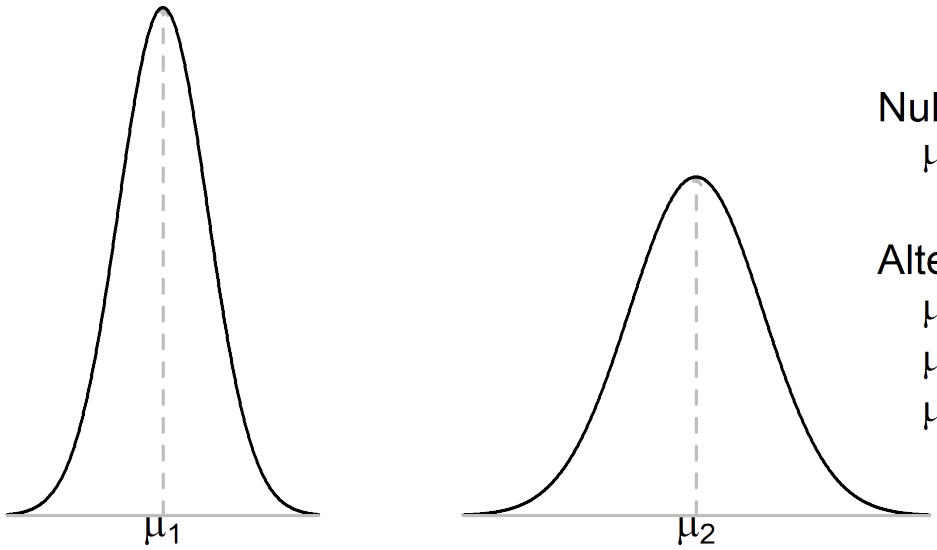
$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$CI = (\bar{y}_1 - \bar{y}_2) \pm t^* \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Source: statkat.com

Two sample t: equal variances not assumed

population distributions



Null hypothesis:
$\mu_1 = \mu_2$

Alternative hypothesis
$\mu_1 \neq \mu_2$ (two sided)
$\mu_1 > \mu_2$ (right sided)
$\mu_1 < \mu_2$ (left sided)

sample data

37	41	37	46
41	37	42	43

62	59	66	62
58	51	64	60

$\bar{y}_1 = 40.5$ $\quad \bar{y}_2 = 60.2$
$s_1 = 3.3$ $\quad s_2 = 4.6$
$n_1 = 8$ $\quad n_2 = 8$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$CI = (\bar{y}_1 - \bar{y}_2) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Source: statkat.com

Should We Make this Assumption?

- We don't actually know the population variances, so we could be wrong when assuming they are equal.
- Being wrong creates more problems when n is small and when samples are very different in size.

  $\Rightarrow$ If the assumption is wrong, it can worsen the very problem we hope to address!
- Advice: do not assume equal variances unless there is good reason to believe the assumption is valid.

Pooling the Sample Variances

- The purpose, in the end, is to calculate the standard error of the difference.
- Rather than use the standard deviations from each sample, s_1 and s_2, we combine the samples to create a pooled estimator s.
- We then use s in place of s_1 and s_2 in the formula for the standard error of the difference.

$$se_{diff} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Calculating the Pooled Estimator s

s^2 is a weighted average of s_1^2 and s_2^2, where the weights are the degrees of freedom for each sample, and we divide by the total degrees of freedom.

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

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

With s, we have what we need for calculating the standard error of the difference.

Confidence Intervals and Hypothesis Tests

The formulas do not change. We just apply the correct calculation for the standard error:

$$se_{diff} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Confidence intervals:

$$(\bar{y}_2 - \bar{y}_1) \pm t(se_{diff})$$

Significance tests:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - H_0}{se_{diff}}$$

The degrees of freedom are $n_1 + n_2 - 2$.

Example: Mean Temperature in Ann Arbor

Suppose you are testing whether the mean temperature in Ann Arbor differed in 1900 and 2000.

- You have a small random sample of temperature readings from 1900: $\bar{y}_1 = 47$, $s_1 = 14$, and n_1=15.
- For 2000, you have temperature readings for every day: $\bar{y}_2 = 50$, s_2=11, and n_2=366 (leap year).
- It may be realistic to think that variance in temperature is equal for these two "populations."
- Plus, with the small size of sample 1, we may gain information by pooling the standard deviations.

Calculate the Pooled s and se_{diff}

$$
\begin{aligned}
s &= \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \\
&= \sqrt{\frac{(15-1)14^2+(366-1)11^2}{15+366-2}} \\
&= \sqrt{\frac{2744+44165}{379}} = 11.13
\end{aligned}
$$

$$
se_{diff} = s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 11.13\sqrt{\frac{1}{15}+\frac{1}{366}} = 2.93
$$

Perform the Difference of Means Test

We have $n_1 + n_1 - 2 = 379$ degrees of freedom, so the critical value of t for $\alpha = .05$ is 1.966. Let's calculate the test statistic:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - \mathsf{H}_0}{se_{diff}} = \frac{(50 - 47) - 0}{2.93} = \frac{3}{2.93} = 1.02$$

Since $|t| < 1.966$, we fail to reject the null hypothesis that there is no difference in mean temperature.

If we had assumed unequal variances, our t-statistic would have been 0.82.

Using Software

- In Stata, the `ttest` command by default applies the assumption of equal variances to independent-sample tests.

  `ttest` *interval_var*, `by(`*categorical_var*`)`

- In R, just add the argument `var.equal = TRUE` to the `t.test` function.

  `t.test(`*interval_var* $\sim$ *categorical_var*, `var.equal = TRUE)`

- Since we get the maximum possible degrees of freedom, our t-statistic is generally larger when assuming equal variances.

- But this does not make it the right approach!

Outline

Small-Sample Tests with Proportions

- For hypothesis tests involving a single proportion with a small sample, we use the binomial distribution.
- For small-sample tests involving a comparison of proportions, we use Fisher's Exact Test.
- The calculations are complicated, and usage is uncommon, so we will just learn to do it with software.
- Rule of thumb: use this test if one or more of the categories in the two samples has fewer than 10 cases.

Fisher's Exact Test Setup

- Assumptions: we have random samples in which observations fall into one of two categories.
- Hypotheses:

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

- We obtain a p-value, the sum of all probabilities at least as unlikely as the observed probability, under the scenario that H_0 is true.

Example: Sleep-Deprived MA Students

Suppose a study compared MPP/MPA students versus other young professionals on the likelihood of getting "insufficient sleep."

- In a sample of 13 MPP and MPA students, 11 reported insufficient sleep.
- In sample of 8 young professionals, 3 reported insufficient sleep.

The null hypothesis would be that young professionals of all kinds are equally likely to get insufficient sleep.

Fisher's Exact Test Calculations

- Find all possible joint frequency distribution tables that are consistent with the row and column totals from our sample.
- Then, find the proportion of these tables that reflect more extreme non-independence than does the table in our sample.
- This proportion is an exact p-value.
- Note: if the number of rows or columns is too large, R and Stata will have trouble doing the calculations.
- There are too many possible tables, so the software reports an error message.

Running a Fisher's Exact Test (Stata)

The `tabi` command can create a row by column table. The option `exact` performs the test.

```
. tabi 11 3 \ 2 5, exact
```

row	col 1	2	Total
1	**11**	**3**	**14**
2	**2**	**5**	**7**
Total	**13**	**8**	**21**

```
          Fisher's exact =                 0.056
1-sided Fisher's exact =                 0.041
```

Running a Fisher's Exact Test (R)

Make a table (data matrix) with the `rbind()` function, then use the `fisher.test() function`.

```
> mytable <- rbind(c(11,3), c(2,5))
> mytable
     [,1] [,2]
[1,]   11    3
[2,]    2    5
> fisher.test(mytable)

        Fisher's Exact Test for Count Data

data:  mytable
p-value = 0.05552
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.8235417 127.5923429
sample estimates:
odds ratio
  8.033314
```

Evaluating the Test

- The two-sided p-value is .056.
- The two-sided p is not double because the distribution is not symmetric. It is a hypergeometric distribution.
- We cannot reject H_0 at $\alpha = .05$.
- Note: the option `exact` also works with the regular `tabulate` command.

Outline

Non-Parametric Tests

- The tests we have used so far all require CLT assumptions or the assumption that the population has a normal distribution.
- If we have means of small samples from populations with highly skewed distributions, we cannot perform a t-test.
- Instead, we can use non-parametric tests that need no assumptions about the shape of the population distribution.
- For example, some tests are based on ordinal rankings of the cases from high to low

Example: Wilcoxon-Mann-Whitney Test

- This is rank test for independent samples.
- It involves combining the samples, ranking the data ordinally, and computing a test statistic (U).
- Basically, U_1 counts the total number of times in which an observation from sample 1 is ranked higher than one from sample 2 compared with the expected sum of rankings.

Using Software

- We have an outcome variable of interest (`outcome`) and a dichotomous variable that makes categories (`catvar`).
- This is implemented in Stata with the `ranksum` command.

```
ranksum outcome, by(catvar)
```

- In R:

```
wilcox.test(outcome ~ catvar, data=dataset_name)
```

Example: Democracy and GDP per capita

- Suppose I want to test whether GDP per capita differs across democracies and non-democracies in a small sample of cases.
- Since GDP per capita is highly skewed, using mean comparisons could be problematic with small samples.
- The Wilcoxon-Mann-Whitney test ranks the combined samples in order of GDP per capita. The test examines relative rankings for the regime-type categories.
- The data are for 24 countries in Central and South America in 1975, 12 of which were democracies.

Country	Democracy in 1975?
Mexico	No
Brazil	No
Colombia	Yes
Venezuela	Yes
Argentina	Yes
Peru	No
Chile	No
Ecuador	No
Uruguay	No
Dominican Republic	Yes
Guatemala	Yes
Trinidad-Tobago	Yes
Costa Rica	Yes
Jamaica	Yes
Bolivia	No
Panama	No
Honduras	No

```
. ranksum GDPcap, by(Democracy)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

Democracy	obs	rank sum	expected
0	**12**	**158**	**150**
1	**12**	**142**	**150**
combined	**24**	**300**	**300**

```
unadjusted variance        300.00
adjustment for ties          0.00
                           ------
adjusted variance          300.00

Ho: GDPcap(Democr~y==0) = GDPcap(Democr~y==1)
                z =   0.462
    Prob > |z| =   0.6442
```

Interpretation: there is not a detectable difference between democracies and non-democracies in GDP per capita.

Wilcoxon Tests for Paired Samples

- The Wilcoxon signed-rank test is a non-parametric test for paired samples.
- The data from each sample are combined and ranked ordinally. The mean ranking of each sample is calculated.
- These means are distributed normally when n gets large. We get a z statistic.
- The null hypothesis is that the mean rankings are the same.
- See the `signrank` command in Stata.
- R: `wilcox.test(data$var1, data$var2, paired=TRUE)`

Outline

Adding a Control Variable

- Last week, we went over the format for a comparison of means table involving two variables (e.g. x and y).
- We can add another categorical variable (e.g. z) to the analysis to control for the effect of that variable.
- Specifically, we can obtain the mean of y for all the combinations of x and z to see how the mean changes when holding the value of z constant.

Format of the Mean Comparison Table

Category of x	Mean of y	Frequency
Value 1	Mean 1	% cases (# cases)
Value 2	Mean 2	% cases (# cases)
Value 3	Mean 3	% cases (# cases)
Total	Mean all	100% (# cases)

A Controlled Comparison of Means

	Category of z		
Value of x	Value 1	Value 2	Total
Value 1	$\text{Mean}_{1,1}$ (# cases)	$\text{Mean}_{1,2}$ (# cases)	$\text{Mean}_{x=1}$ (# cases)
Value 2	$\text{Mean}_{2,1}$ (# cases)	$\text{Mean}_{2,2}$ (# cases)	$\text{Mean}_{x=2}$ (# cases)
Value 3	$\text{Mean}_{3,1}$ (# cases)	$\text{Mean}_{3,2}$ (# cases)	$\text{Mean}_{x=3}$ (# cases)
Total	$\text{Mean}_{z=1}$ (# cases)	$\text{Mean}_{z=2}$ (# cases)	Mean_{all} (# cases)

Where the cells of the table contain the mean value of y for each combination of categories.

Example: Controlled Comparison of Means

Suppose we are interested in the effect of democracy on a life expectancy in a country.

- We can measure life expectancy levels from various sources, such as the World Bank. Scholars have created different dichotomous measures of democracy.
- Given that national wealth also affects life expectancy, we should control for level of wealth to get a better sense of the effect of democracy.

A Controlled Comparison of Means

Mean Level of Life Expectancy

	Level of GDP/capita			
Democracy?	Low	Medium	High	Total
No	58.8	67.4	72.6	64.5
	(27)	(23)	(11)	(61)
Yes	63.1	73.0	78.3	72.5
	(28)	(31)	(43)	(102)
Total	61.0	70.6	77.1	69.5
	(55)	(54)	(54)	(163)

Interpretation

- By comparing means along a particular column/row, we hold one variable constant and examine the effect of the other variable.
- We see that the effect of democracy is similar across levels of GDP/capita (though it matters less when GDP/capita is Low).
- This is mostly an additive relationship: both variables matter, and the effect of one variable is close to the same for each value of the other variable.
- By adding the control variable, we created comparison groups that are more similar in composition, except for the key variable of interest.

Other Types of Relationshps

We found an additive relationship. What else might we have found?

- Spurious Relationship: after adding the control variable, we might have found that Democracy made no difference at all. The apparent differences were due to GDP per capita.
- Interactive Relationship: the effect of Democracy depends upon the level of GDP per capita, and vice-versa.

Multivariate Analysis

- This is a simple way to do multivariate analysis, but it quickly becomes unmanageable if we want to add more variables.

  e.g. adding another control variable with two categories would require a table with 7 columns (counting columns for subtotals).

- Linear regression is much more useful in these circumstances (assuming y is interval-level).

Overall Summary for Comparison of Two Groups

- In the end, we perform very familiar tests with t- and z-statistics. We just need to get the right standard error and degrees of freedom.
- The key factors: Are we dealing with means or proportions? Are the samples independent? Can we assume equal variances? Are the samples large or small?
- We need to know which formulas to apply based upon our answers to these questions.