

Lecture 4.1 - Comparing Groups

Student

2025-04-25

Table of contents

Comparing Gropus	3
Setting Expectations	3
Summarizing the Data	4
Analyze	4



Comparing Gropus

Setting Expectations

Today we are going to continue working with a dataset on wages in the U.S. collected by the US Department of Labor: `us.dol.wages.csv`

As before, we will first assume that the dataset approximately represents all workers in the U.S. – statistics of the mean/proportions can serve as stand-ins for the true population parameters.

The data columns in this dataset are:

- **exp**: Years experience
- **wks**: Weeks worked per year
- **bluecol**: Is the job a blue collar job
- **ind**: Works in the manufacturing industry
- **south**: Is the person working in the American South
- **married**: Is the person married
- **sex**: What is the sex of the person
- **union**: Is the person in a union
- **ed**: Number of years of education
- **black**: Is the person Black
- **lwage**: Log of wages per week
- **wage**: Wage of the person per week

First, form a hypothesis about the data based on **categorical** variable subgroups with respect to the response variable **wage**. You might think that union member earn more than non-union members, for example, or that males earn more than females.

1. Write out your hypothesis about the data, stating the null hypothesis and alternative hypothesis and write a justification for your hypothesis.
2. Then, make a general statement about whether you expect the difference to be substantively large or small. Do you expect men to earn a little more than women or a lot more, for example?
3. Based on a sample size of 100 and the nature of the hypotheses, select an appropriate alpha value.
4. Create a sample of size 100. Remember, you can do this with the following command:

```
us.dol.sample <- us.dol.wages %>%  
  slice_sample(n=100, replace=TRUE)
```

Summarizing the Data

5. Make a simple table that includes summary data (mean, standard deviation, range, n) from for each side of your categorical variable with respect to the response variable (wage).

You can get the information by filtering the data and summarizing it according to the following example of not married people in the dataset:

```
us.dol.sample %>%  
  filter(married=="no") %>%  
  summarize(mean.wage.married = mean(wage),  
            sd.wage.married = sd(wage),  
            range.wage.married = range(wage))
```

6. Also make a second table with summary data (mean, standard deviation, range) for the overall values of your response variable (wage) from your sample.
7. Write a few sentences interpreting your summary statistics. Does the difference between the subgroups on the variables of interest seem large or small compared to the range of your response variable?

Analyze

8. Check the conditions for a t test – does the data support using a t test?
9. Calculate the p value of the difference to test your hypotheses by hand, do not use R's t test function. To choose the appropriate degree of freedom, you can use the textbook shortcut of number of degrees of freedom of the smaller of the n .

Remember, the formula for the $t_{difference}$ is: $t = \frac{\bar{y}_1 - \bar{y}_2}{S_e}; S_e = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

10. Make a table with five columns, the first two columns being the mean salary of the two groups (for example, union/non-union), the third column is the size of the difference, the fourth column being the confidence interval for the difference, and the fifth column is the p value associated with whether the difference is statistically significantly different from zero.
11. Now let's use R to conduct t tests on the data. The general form of a t test in R is: `t.test(groupa, groupb)` for a two group comparison or `t.test(group, mu=<your hypothesis>)` for a one group comparison where `mu` is your null hypothesis. Find the equivalent piece of information for each cell in your table. How close were your results calculated by hand to the result calculated by R?

Hint: to quickly create a subset, you can use the following command:

```
females.sample <- subset(us.dol.sample, sex=="female")
males.sample <- subset(us.dol.sample, sex=="male")

t.test(females.sample$wage, males.sample$wage)
```

13. Now find the “true” population difference from the entire dataset. Did the confidence interval for your difference cover the entire population difference?

Interpret

14. Make some comments regarding whether the results surprised you or confirmed what you thought about the world.
15. Also make some notes regarding whether the difference in means that you found were substantively significant.
16. Write up a profile of an ‘average’ person in this dataset and describe what features this person has that cause them to earn more or less than someone in the opposite demographic categories that they are in.
17. Finally, note whether you think there are any other factors we should consider when evaluating your hypotheses or if you think there are any problems with just using a t test to evaluate the difference.