Estimating Differences in Means in Paired Samples

Here we investigate estimating the difference in the means between two *paired samples*.

With **paired samples**, we have a single set of observations, but two measures or two variables for each observation. Obtaining two or more measures from each observation to pair can result from the following situations:

1. **Across-time measures:** The same dependent variable is measured for each individual but at two different time periods. For example, a sample of individuals may have their income measures once in 2013 and again in 2014, and a researcher could ask whether there was a change in average income from one year to the next.
2. **Different conditions measures:** The same dependent variable is measured for each individual, but under two different conditions, or before and after some treatment. For example, a sample of high school students may have test scores measured before introducing them to a new curriculum and afterwards. A researcher could ask whether the curriculum affected test scores.
3. **Related topics measures:** Two slightly different variables are measured for each individual. For example, foreign language students may take separate exams for writing proficiency and speaking proficiency. A researcher could ask whether students are more proficient with writing in a foreign language versus speaking.

Testing differences in *means* between paired samples is appropriate when the variables are measured at the interval or ratio scale.

Example: The Centers for Disease Control and Prevention (CDC) maintains data on motor vehicle fatalities by State, Age, and Gender. In our example, a dataset with 50 observations, one for each U.S. state, the motor vehicle occupant fatality rate per 100,000 members of the population. The dataset includes separate variables for the following age groups: 0-20, 21-34, 35-54, and 55+. The dataset also includes variables for the mortality rate for women as a whole and men as a whole.

1. Download the dataset

The code below downloads a csv file available on my personal website, http://murraylax.org/datasets/, then reads it into an R dataset that we name `fatalities`.

```
download.file(
  url="http://murraylax.org/datasets/vehiclefatalities.csv",
  dest="vehiclefatalities.csv")
fatalities <- read.csv("vehiclefatalities.csv");
```

2. Computing Means

The function `t.test` computes a number of statistics and statistical tests for a difference between two means, including sample estimates for the differences in the means, a confidence interval, and a hypothesis test. In the code below, we call the function to compare the means of variables `Age21.34` and `Age.35.54`, instruct the function that these are *paired* samples, and assign all the resulting output to a new variable we call `fatalstats`.

```
fatalstats <- t.test(x=fatalities$Age.21.34,
                     y=fatalities$Age.35.54,
                     paired=TRUE,
                     alternative="two.sided",
                     conf.level=0.95)
```

The first two parameters, `x` and `y`, into the function `t.test` are the two variables that we are comparing. As each variable is a member of the dataset named `fatalities`, we access each one by first naming the dataset, typing a dollar sign (`$`), then specifying which variable in the dataset we are referring to.

The third parameter specifies that we want a two tailed test. We do a two tailed test because our research question simply asked if there is a *difference* between the average fatality rate for the two groups, not whether a specific variable was *greater* than the other. Consistent with the research question and the two-tailed test is the not-equal sign in the alternative hypothesis. The last parameter `conf.level=0.95` will generate output that will be useful later for computing a 95% confidence interval.

The output of `t.test` that we assigned to variable `fatalstats` is a list which includes an item called `estimate`. The `estimate` item is equal to the mean of the `x` variable (the 21-34 age group) minus the mean of the `y` variable (the 35-54 age group). We report this item with the following code:

```
fatalstats$estimate
```

```
## mean of the differences
##                   4.625
```

Here we see a positive number, equal to `4.625`, which means the fatality rate for the 21-34 age group is higher than the 35-54 age group, by amount of 4.625 people per 100,000 in the population. If we wish to compute the mean for each age group, we can use the `mean()` function for each of the variables as follows:

```
mean(fatalities$Age.21.34, na.rm=TRUE)
```

```
## [1] 13.70435
```

```
mean(fatalities$Age.35.54, na.rm=TRUE)
```

```
## [1] 8.970455
```

The parameter `na.rm=TRUE` tells the function `mean` to ignore missing values, which are coded with NA (i.e. not available). We can see here that the mean fatality rate for the 21-34 age group is approximately 13.704 per 100,000 and the mean fatality rate for the 35-54 age group is 8.970 per 100,000. The difference is equal to the estimate found above, 4.625.

3. Calculate a 95% Confidence Interval

The confidence interval is a range of values for difference between the population means of the two variables, based on our samples estimates of the means and an estimate for the margin of error due to random sampling.

The output to the call to `t.test` above also included a confidence interval, in an item called `conf.int`. Let's call this item to report our confidence interval:

```
fatalstats$conf.int
```

```
## [1] 3.674189 5.575811
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval places the mean difference between fatality rates of the 21-34 age group and the 35-54 age group in the range 3.674 and 5.576. We can say with 95% confidence that this interval estimate includes the true difference in population means.

4. Two-Tailed Paired Samples T-test

Let us test the hypothesis that the vehicle fatality rate 21-34 age group is different than the vehicle fatality rate for 35-54 age group. The null and alternative hypotheses are given as follows:

Null hypothesis: $\mu_{21/34} - \mu_{35/54} = 0$
Alternative hypothesis: $\mu_{21/34} - \mu_{35/54} \neq 0$

Notice that the alternative hypothesis includes a $\neq$ sign which implies that this is a two-tailed test. We are not specifying that a particular age group should be higher than the other. We are only testing whether the population means are *different* from one another.

The output to the call to `t.test` above also included a paired samples t-test. If we call our return value, summary information from the test is output to the screen.

```
fatalstats
```

```
##
##  Paired t-test
##
## data:  fatalities$Age.21.34 and fatalities$Age.35.54
## t = 9.8097, df = 43, p-value = 1.541e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.674189 5.575811
## sample estimates:
## mean of the differences
##                   4.625
```

We can see from above that the p-value is 1.5e-12 which is much smaller than a significance level of 5%. We can say confidently that there is statistical evidence that the average fatality rate for the 21-34 age group is different than the 35-54 age group.

5. One-Tailed Independent Samples T-Test

Suppose our intuition tells us that the 21-34 age group may have a higher fatality rate than the 35-54 age group, because they have less experience driving and less maturity may lead to more dangerous decisions. To test this intuition, suppose a researcher is interested in instead testing the following one-tailed hypotheses:

Null hypothesis: $\mu_{21/34} - \mu_{35/54} = 0$
Alternative hypothesis: $\mu_{21/34} - \mu_{35/54} > 0$

The *greater-than* symbol implies that this is a one-tailed.

The code that we ran above did conduct a hypothesis test, but we need to call the function again to specify that this is a one-tailed test. The relevant call to `t.test` is given by,

```r
t.test(x=fatalities$Age.21.34,
       y=fatalities$Age.35.54,
       paired=TRUE,
       alternative="greater",
       conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  fatalities$Age.21.34 and fatalities$Age.35.54
## t = 9.8097, df = 43, p-value = 7.707e-13
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.832424      Inf
## sample estimates:
## mean of the differences
##                   4.625
```

The p-value is 7.7e-13 which is much smaller than a significance level of 5%. Not coincidentally, this p-value is exactly half of the p-value we found in the two tailed test above. Given the low p-value, we can say confidently that there is statistical evidence that the average fatality rate for the 21-34 age group is *greater than* the 35-54 age group.