Estimating Differences in Means - Independent Samples

R Tutorials for Applied Statistics

1 Introduction

In this tutorial we investigate estimating the differences in means between two *independent samples*.

With **independent samples**, we have two groups of observations, distinguishable by some measured characteristic to divide into these groups, and no member of one group is also in the other group. The outcome of the observations in one group must be *independent* of the outcome of the observations in the other group.

Testing differences in *means* between two independent samples is appropriate when a variable measured from two independent samples are in the same units and at the interval or ratio scale.

Example: Current Population Survey from 2004 that includes data on average hourly earnings, marital status, gender, and age for thousands of people. A part of it is available for download from the textbook website for Stock and Watson's *Introduction to Econometrics*.

Our goal with this example is to estimate the difference in average hourly earnings between men and women. The gender variable is our measured characteristic that will divide our sample into two independent samples.

2 Download the Data

The code below downloads and loads into memory the data set.

```r
load(url("http://murraylax.org/datasets/cps04.RData"))
```

The data frame `df` contains a variable called `ahe`, which stands for average hourly earnings, and a variable `female` which is equal to 1 if the observation is for a female and equal to 0 if for a male.

3 Calculate Means

The function `t.test` computes a number of statistics and statistical tests for a difference between two means, including sample estimates for each mean, a confidence interval, and a hypothesis test. In the code below, we call the function and assign all the resulting output to a new object we call `ahestats`.

```
ahestats <- t.test(ahe ~ female, data=df, conf.level=0.95, alternative="two.sided")
```

The first parameter, `ahe ~ female`, is a *formula* that says we are interested in the outcome variable `ahe`and how it is different for different values for the explanatory variable, `female`. The second parameter, `data=df`, tells the function in what data frame the variables `ahe` and `female` can be found.

The next parameter, `conf.level=0.95`, will generate output that will be useful later for computing a 95% confidence interval. This parameter is optional, and the default value for `conf.level` if nothing is specified is equal to 0.95. Therefore, even if this had been completely omitted, the function call would have performed identically. Still, it is useful to include this for readability, as people may not have memorized default values for optional parameters.

The final parameter, `alternative="two.sided"` is also an optional parameter, and the default value is what we have assigned, “two.sided”. This means `t.test()` will compute a two-tailed hypothesis test and confidence interval. Confidence intervals are almost always reported as two-sided (i.e. two-tailed). We learn more about the difference between two-tailed and one-tailed hypothesis tests in the sections below.

The output of `t.test` that we assigned to the object `ahestats` is a list which includes an item called `estimate`. The `estimate` item includes the mean of each of the groups defined by `female`. Report this item with the following code:

```
ahestats$estimate
```

```
## mean in group 0 mean in group 1
##        17.77269        15.35860
```

We can see from above that men in our sample have average hourly earning equal to $17.77 and women have average hourly earnings equal to $15.36.

4 Calculate a 95% Confidence Interval

The confidence interval is an estimate of the lower and upper bounds for our estimate of the difference between the population means for our two independent groups. These estimated bounds are based on the estimated sample means and an estimate for the margin of error due to random sampling.

The output to the call to `t.test` above also includes a confidence interval, in an item called `conf.int`. Let us call this item to report our confidence interval:

```
ahestats$conf.int
```

```
## [1] 2.039770 2.788425
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval for the difference between average hourly earnings between men and women is between $2.04 and $2.79. We can say with 95% confidence that this interval estimate includes the true difference in population means.

5 Two-Tailed Independent Samples T-Test

An **independent samples t-test** lets us determine whether there is evidence that the mean of the first group is different than the mean of the second group in the population. The typical two-tailed test considers the following null and alternative hypotheses:

Null hypothesis: $\mu_0 - \mu_1 = 0$μ0−μ1=0
Alternative hypothesis: $\mu_0 - \mu_1 \neq 0$μ0−μ1≠0

Notice that the alternative hypothesis includes a $\neq$≠ sign which implies that this is a two-tailed test. We are not explicitly testing which group has a larger average hourly earnings. We are only testing whether the population means are different from one another.

The output to the call to `t.test` above also includes an independent samples t-test. If we call our return value from the R console, summary information from the test is output to the screen.

```
ahestats
```

```
##
##  Welch Two Sample t-test
##
## data:  ahe by female
## t = 12.642, df = 7792.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.039770 2.788425
## sample estimates:
## mean in group 0 mean in group 1
##        17.77269        15.35860
```

We can see from above that the p-value is less than 2.2e-16 (i.e. 2×10^{-16}2×10−16) which is much smaller than a significance level of 0.05 or 5%. We can say confidently that there is statistical evidence that the average hourly earnings is different for men and women.

6 One-Tailed Independent Samples T-Test

Suppose a politician claims that men earn on average more than $2.00 per hour more than women. The null and alternative hypotheses for testing this claim are given by the following:

Null hypothesis: $\mu_0 - \mu_1 = 2.00$μ0−μ1=2.00
Alternative hypothesis: $\mu_0 - \mu_1 > 2.00$μ0−μ1>2.00

In the hypotheses above, μ_0μ0 denotes the mean hourly earnings for men (female=0), and μ_1μ1 denotes the mean hourly earnings for women (female=1). The alternative hypothesis has a *greater-than* symbol, because the claim we are testing suggested that men on average make more than $2.00 per hour than women, when we subtract group 1 (women) from group 0 (men), the result should be *greater than* 2.00. This is therefore a one-tailed test.

The code that we ran above did conduct a hypothesis test, but we need to call the function again to give it the specifics that we want a one-tailed test and that we have a value of $2.00 in the hypotheses.

The relevant call to `t.test` is given by,

```r
t.test(ahe ~ female, data=df, alternative="greater", mu=2.00)
```

```
##
##  Welch Two Sample t-test
##
## data:  ahe by female
## t = 2.1685, df = 7792.6, p-value = 0.01507
## alternative hypothesis: true difference in means is greater than 2
## 95 percent confidence interval:
##  2.099964      Inf
## sample estimates:
## mean in group 0 mean in group 1
##        17.77269        15.35860
```

The p-value is 0.015, which is less than 0.05, so we can say at the 5% significance level, we found sufficient statistical evidence that on average men earn more than $2.00 per hour more than women.