

Estimating the Population Mean

R Tutorials for Applied Statistics

The **population mean** is a measure of the center or “average” value in the whole population of a variable measured at the interval or ratio level.

The **sample mean** is a sample estimate of the population mean. It is a the same measure of center, obtained from a sample. The variable in your sample must be measured at the interval or ratio level.

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*.

1 Download the Dataset

The code below downloads the data set and assigns the dataset to a variable we create and call `cps04`.

```
cps04 <- read.csv(url("http://murraylax.org/datasets/cps04.csv"))
```

The dataset `cps04` contains a variable called `ahe`, which stands for average hourly earnings.

2 Compute the Sample Mean

```
mean(cps04$ahe)
```

```
## [1] 16.77115
```

The sample estimate for average hourly earnings for U.S. workers in 2004 is \$16.77. This is not necessarily the population mean. Like every statistic, it includes a margin of error due to random sampling error.

3 Compute a 95% Confidence Interval

The confidence interval is a range of values for the population mean, based on our estimate of the sample mean, and an estimate for the margin of error due to random sampling.

The function `t.test` computes a number of statistics and statistical tests for a variable, including a confidence interval. In the code below, we use the function to compute our confidence interval and assign all the resulting output to a new variable we call `ahestats`.

```
ahestats <- t.test(cps04$ahe, conf.level = 0.95)
```

The output of `t.test` that we assigned to variable `ahestats` is a list which includes an item called `conf.int`.

Let's call this item to report our confidence interval:

```
ahestats$conf.int
```

```
## [1] 16.57902 16.96328  
## attr(,"conf.level")  
## [1] 0.95
```

The confidence interval for average hourly earnings for U.S. workers in 2004 is \$16.58 - \$16.96. We can say with 95% confidence that this interval estimate includes the true population mean.

4 One Sample T-Test (One-tailed)

Suppose a politician claimed that the average earnings of American workers was more than \$16.50 per hour. We know that the sample estimate is larger from above, but let's test the hypothesis that the *population mean is more than* \$16.50.

The appropriate statistical procedure is the **One-sample T-test for a Mean** which tests whether a single population mean is equal to or different than a particular value. Our null and alternative hypotheses for our one-sample t-test is given by the following:

Null hypothesis: $\mu = 16.50$

Alternative hypothesis: $\mu > 16.50$

The `t.test` function can also compute the one-sample t-test using the following code:

```
t.test(cps04$ahe, mu=16.50, alternative = "greater")
```

```
##  
## One Sample t-test  
##  
## data: cps04$ahe  
## t = 2.7665, df = 7985, p-value = 0.002839  
## alternative hypothesis: true mean is greater than 16.5  
## 95 percent confidence interval:  
## 16.60992 Inf  
## sample estimates:  
## mean of x  
## 16.77115
```

The output of the test reveals a p-value equal to 0.00028. Since this is below 5%, we reject the null hypothesis and conclude that we do have statistical evidence that the population mean is greater than \$16.50.

5 Two-Tailed Test

The previous example is a **one-tailed** test. That is, it involved an alternative hypothesis that looked for statistical evidence that the population parameter was in a particular direction away from the null hypothesized value (in the case above, *greater than* the null hypothesis).

A two tailed test instead tests an alternative hypothesis that simply says the population parameter is *different than* the null hypothesized value, leaving the possibility that it may be less than or may be greater than the value.

Let's test the following two-tailed hypotheses:

Null hypothesis: $\mu = 16.50$

Alternative hypothesis: $\mu \neq 16.50$

Notice the \neq sign in the alternative hypothesis.

We use the `t.test` function again to compute the one-sample t-test using the following code:

```
t.test(cps04$ahe, mu=16.50, alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  cps04$ahe
## t = 2.7665, df = 7985, p-value = 0.005679
## alternative hypothesis: true mean is not equal to 16.5
## 95 percent confidence interval:
##  16.57902 16.96328
## sample estimates:
## mean of x
##  16.77115
```

We can see from the output that the p-value is equal to 0.0057. Since this is below 5%, we reject the null hypothesis and conclude that we do have statistical evidence that the population mean *is different than* \$16.50.