

# Chapter 11

## Estimating Means with Confidence

### 1. Determining the t-multiplier for a confidence interval

In Example 11.4 we need to find the multiplier  $t^*$  values for 24 degrees of freedom and 95% or 99% confidence. This is almost identical to what was done in Chapter 10 to find the multiplier  $z^*$ , but we only need to specify the degrees of freedom and not the mean nor standard deviation. The R function `qt( p, df )` is used where  $p$  is the percentile and  $df$  is the degrees of freedom. So for the 95% confidence, the multiplier  $t^*$  is found by typing the command `qt( 0.975, 24 )` and for the 99% confidence, `qt( .995, 24 )`. The commands and R output are shown below. Recall that for the 95% confidence multiplier, we need that value of  $t^*$  such that 2.5% of the area beneath the t-distribution density curve is to the right of  $t^*$  or equivalently 97.5% is to the left.

```
> qt( 0.975, 24 )
[1] 2.063899
> qt( 0.995, 24 )
[1] 2.796940
```

### 2. Constructing a confidence interval for a single mean

Example 11.5 of Section 11.2 constructs the 95% confidence interval for the mean forearm length of men using a random sample of  $n=9$  men. We will first show a long way to calculate the confidence interval by explicitly utilizing the formula  $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ . Next we will do the equivalent with the R function `t.test()`. Either way, we will first have to enter the 9 forearm lengths into a variable which we will call `x`. The R commands and output follow showing the interval to be 24.33 to 26.67 inches. (Note: Multiple commands can be used in the same line of commands as done below with `mean()`, `qt()`, and `sd()`.)

```
> x <- c(25.5, 24, 26.5, 25.5, 28, 27, 23, 25, 25)
> mean(x) + qt( .975, 8) * sd(x)/sqrt(9)
[1] 26.66891
> mean(x) - qt( .975, 8) * sd(x)/sqrt(9)
[1] 24.33109
>
> t.test( x, conf.level=0.95 )
```

```
One Sample t-test
data: x
t = 50.3061, df = 8, p-value = 2.7e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
24.33109 26.66891
sample estimates:
mean of x
25.5
```

### 3. Checking the conditions before finding a confidence interval for the mean

Example 11.7 of Section 11.2 explores whether liberal art majors or more technical majors sleep more. The liberal art majors are from an introductory statistics course (Stat10, n=25) while the more technical majors are from a different larger introductory statistics course (Stat13, n=148). The data are found in UC Davis1.RData.

We want to calculate confidence intervals for the mean hours of sleep for each group. For a t-confidence interval to be valid either the population needs to be normally distributed or the sample size needs to be large. The Stat13 class is large so issues of normality are not of concern. There are only 25 Stat10 students, however, so we should check that the distribution of the sample is reasonably close to a normal distribution. This will involve quickly graphing the data to check for any extreme outliers or skewness. We will use boxplots and a histogram to do this.

First, we need to import the data into R and create vectors for the number of hours slept by the Stat10 and Stat13 students. (Do not forget to attach the data frame as shown in the below R commands. ) The column "class" lists Stat10 students as "L" and Stat13 students as "N". We will extract the number of hours of sleep ("Sleep") for each group into `stat10` and `stat13` vectors using the logical R syntax `==` which is the equivalent to the question "is equal to?". For example, the command `stat10 <- Sleep[Class=="L"]` will assign to the vector `stat10` those values from the `Sleep` column where the corresponding class value is equal to L. The functions `length()`, `mean()`, and `sd()` will be used to get quick numerical descriptive statistics of each group. The R commands to be typed and the R output follow.

```
> load("C:/RData/UCDavis1.RData")
> ucdavis1 <- edit(ucdavis1)
> names( ucdavis1 )
[1] "Sex" "TV" "computer" "Sleep" "s" "a" "h" "m" "d"
[10] "e" "g" "Class"
> attach( ucdavis1 )
> stat10 <- Sleep[Class=="L"]
> stat13 <- Sleep[Class=="N"]
>
> length( stat10 )
[1] 25
> mean( stat10 )
[1] 7.66
> sd( stat10 )
[1] 1.344123
> length( stat13 )
[1] 148
> mean( stat13 )
[1] 6.8125
> sd( stat13 )
[1] 1.732756
```

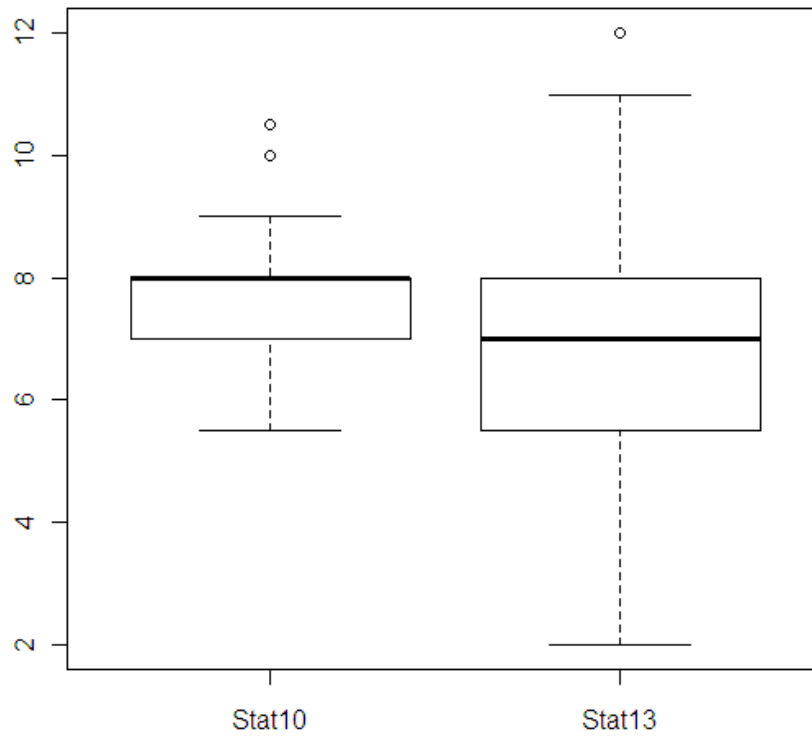
We next do the commands

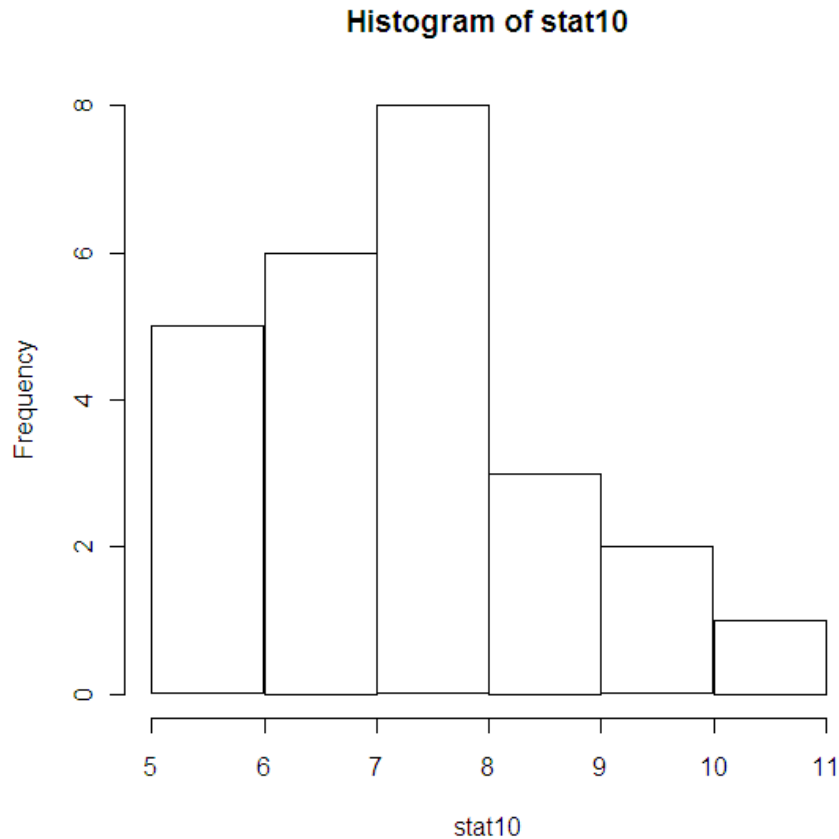
```
boxplot( stat10, stat13, names=c("Stat10", "Stat13") )
```

and

```
hist( stat10 )
```

to see the following boxplot and histograms. The hours of sleep for the Stat10 students is slightly skewed, but with a sample size of 25 not being terribly small (getting close to 30) we will not be too critical.





#### 4. Computing a confidence interval for a single mean

In the above numerical descriptive statistics we saw that the Stat10 students averaged 7.7 hours of sleep compared to 6.8 hours for the Stat13 students. Let us treat these as samples from a larger population and find a confidence interval for the different population means.

Assuming conditions are valid for the confidence intervals, the next step is easy. We simply use the `t.test()` function. The default is for the 95% confidence intervals, so the “`conf.level=0.95`” option is actually not necessary. The R command lines and output that follow give us 7.11 to 8.21 hours for the Stat10 students and 6.53 to 7.09 hours for the Stat13 students. There is no overlap suggesting that a population of Stat10 students would have a greater mean number hours of sleep than a population of Stat13 students.

```
> t.test( stat10, conf.level=0.95 )
```

```
One Sample t-test
```

```
data: stat10
```

```
t = 28.4944, df = 24, p-value < 2.2e-16
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
7.105173 8.214827
```

```

sample estimates:
mean of x
      7.66

> t.test( stat13 )

      One Sample t-test
data:  stat13
t = 47.8299, df = 147, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  6.531022 7.093978
sample estimates:
mean of x
      6.8125

```

## 5. Finding paired differences from raw data

As done with hours of sleep, we need to extract the computer and TV times for only the Stat10 students into `stat10computer` and `stat10TV` vectors. The R commands follow and the 25 differences. The students spent 5.36 hours a week more behind computers than watching TV.

```

> stat10computer <- computer[Class=="L"]
> stat10TV <- TV[Class=="L"]
> difference <- stat10computer - stat10TV
> difference
[1] 28.0 18.5 -4.0 8.0 4.0 -20.0 21.0 19.0 -12.0 -5.0 -5.0
[12] 2.0 40.0 -1.0 -12.0 10.0 5.0 0.0 35.0 -10.5 -14.0 1.0
[23] 5.0 14.0 7.0
> mean(difference)
[1] 5.36

```

## 6. Checking the conditions before finding a confidence interval for paired data

Creating a confidence interval for the mean of paired differences is essentially the same as creating a confidence interval for a single mean. Thus with only 25 differences, we should check the data to see if we can reasonably assume the population is normally distributed. Like before, we will do a boxplot and histogram of the differences.

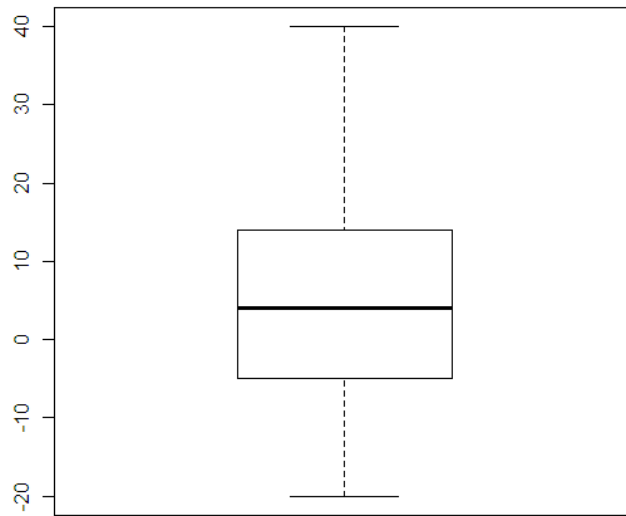
The R commands and graphs follow.

```

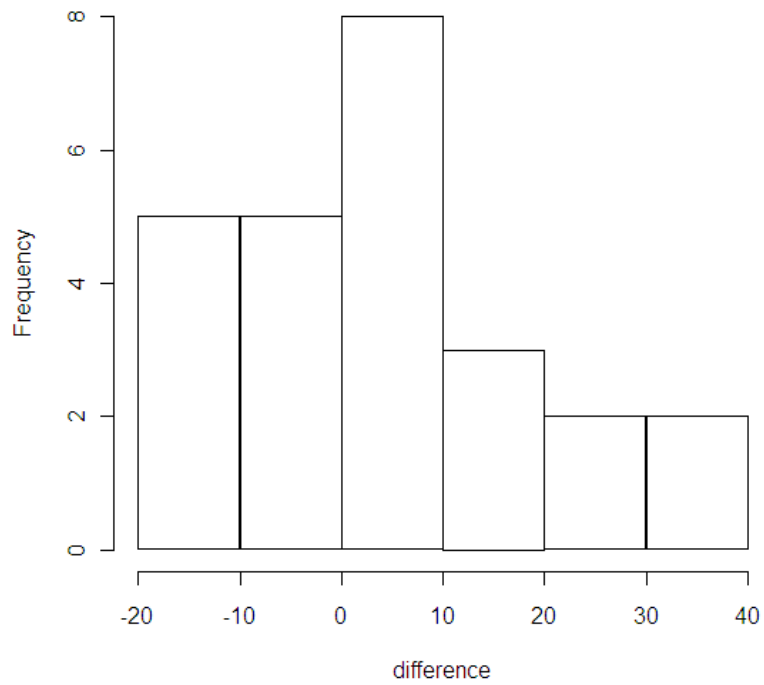
boxplot( difference, main="Computer-TV")
hist( difference )

```

**Computer-TV**



**Histogram of difference**



Neither graph shows serious outliers or skewness, so conditions are acceptable to calculate a confidence interval for the mean difference.

## 7. Calculating a confidence interval for paired data

There are several ways of calculating the confidence interval for paired data. We can treat the differences as a single random variable and simply calculate the confidence interval for a single mean. This would be just like section 2 of this chapter's lab manual and we use the vector `difference`. The R commands for two different methods are given below, but no output.

```
> mean(difference) + qt(0.95, 24)*sd(difference)/sqrt(25)
> mean(difference) - qt(0.95, 24)*sd(difference)/sqrt(25)
> t.test( difference, conf.level=0.9 )
```

The `t.test()` function has a special option to work with paired data. Using the `paired=T` option allows you to skip the calculation of difference and instead simply provide the two vectors of paired data. Below is the R command and the output. The 90% confidence interval is for between 0.14 to 10.58 more hours of computer than TV a week.

```
> t.test( stat10computer, stat10TV , paired=T, conf.level=0.9 )

Paired t-test
data: stat10computer and stat10TV
t = 1.7582, df = 24, p-value = 0.09146
alternative hypothesis:true difference in means is not equal to 0
90 percent confidence interval:
0.1442597 10.5757403
sample estimates:
mean of the differences
                    5.36
```

### Example 11.14, Section 11.4: General confidence interval for the difference between two means (independent samples)

We will continue to use the `UCDavis1.RData` dataset for this example. The number of hours slept by males and females will be compared. The sample will consist of only the non liberal art students; i.e., Stat13 students. Males and females are to be considered two independent samples.

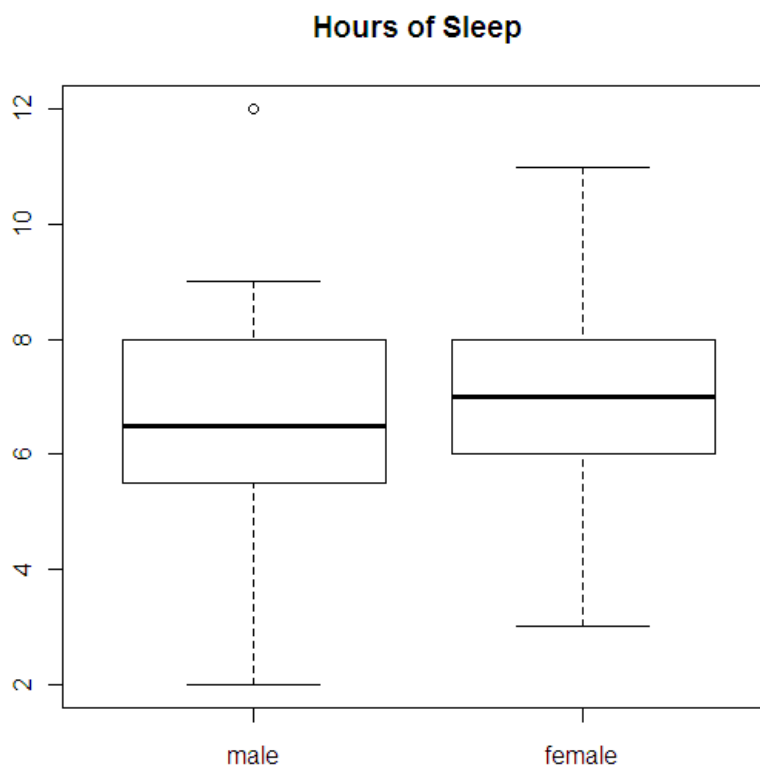
The data will be extracted to vectors using the logical `==` and the logical `&` which is equivalent to “and”. For example, the R commands `sleepmale <- Sleep[ Class=="N" & Sex=="M"]` will extract only those values from `Sleep` where `Class` is equal to `N` and `Sex` is `M`. To get vectors of male and female hours of sleep, to see their respective mean number hours of sleep, and to count the number of males and females type the following R code.

```
> sleepmale <- Sleep[ Class=="N" & Sex=="M"]
> sleepfemale <- Sleep[ Class=="N" & Sex=="F"]
> mean( sleepmale )
[1] 6.553846
> mean( sleepfemale )
[1] 7.01506
> length(sleepmale)
[1] 65
> length(sleepfemale)
[1] 83
```

## 1. Checking conditions before computing a confidence interval for the difference between two independent means.

The distributions of male and female sleep times need to be normal for the t-confidence interval to be valid unless the sample sizes are large. The sample sizes are 65 and 83, so issues of normality are not critical for this example. Nevertheless, we will do a quick boxplot inspection to convince ourselves that the data are not terribly skewed or have extreme outliers. The R command and graph follow which assures us that the data are fairly symmetric and without any extreme outliers.

```
> boxplot( sleepmale, sleepfemale, names=c("male","female"),  
+         main="Hours of Sleep")
```



## 2. Calculating the confidence interval for the difference between two independent means – pooled and unpooled versions.

The pooled version of the confidence interval for the difference between two independent mean assumes that the two populations share a common (same) variance. The sample variances for the males and females ( $\text{var}(\text{sleepmale})$ ;  $\text{var}(\text{sleepfemale})$ ) are 2.83 and 3.08. The difference may simply be a consequence of sampling variation. Assuming the population variances are the same, it is sensible to combine (pool) the sample variances of both groups into a single variance estimate of the common variance. The degrees of freedom will be  $n_1 + n_2 - 2 = 65 + 83 - 2 = 146$ . The R command is simple,



but we do need to specify that we are assuming equal variances since the default is unequal variances. The command and R output follows.

```
> t.test( sleepfemale, sleepmale, var.equal=T )

      Two Sample t-test

data:  sleepfemale and sleepmale
t = 1.6158, df = 146, p-value = 0.1083
alternative hypothesis:true difference in means is not equal to 0
95 percent confidence interval:
-0.1029025  1.0253307
sample estimates:
mean of x mean of y
 7.015060  6.553846
```

The above R output gives a confidence interval of -0.103 to 1.025 for the mean number of hours more sleep female students get compared to male students.

The unpooled estimate does not assume the different populations have equal variances. The t-confidence interval formula is fairly simple,  $\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ , however the degrees of freedom for  $t^*$  is complicated using Welch's approximation. Fortunately R can quickly provide the result as shown below.

```
> t.test( sleepfemale, sleepmale, var.equal=F )

      Welch Two Sample t-test

data:  sleepfemale and sleepmale
t = 1.6243, df = 140.142, p-value = 0.1066
alternative hypothesis:true difference in means is not equal to 0
95 percent confidence interval:
-0.1001752  1.0226034
sample estimates:
mean of x mean of y
 7.015060  6.553846
```

The confidence interval using the unpooled method, -0.100 to 1.023, is practically the same as the results from the pooled method. Note how the degrees of freedom, 140.142, is not an integer for the unpooled method.