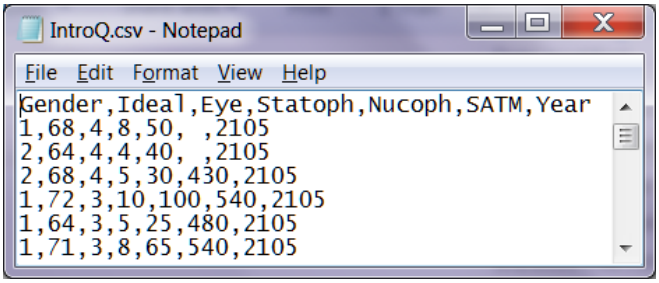
Independent Samples *T* Tests with R

The data we shall use here were collected from students in my introductory statistics classes from 1983 through Spring, 2015. Here is a description of the survey.

The data were in an SPSS file, but I wrote them from SPSS to a csv file. A csv file is a plain text file that uses a comma as the delimiter. At first R did not want to work with this csv file. I discovered that this was because I had SPSS set to use Unicode, but R was assuming the file was in locale code. I went back into SPSS and changed the encoding to locale and all was well after that. Below, on the left, is a snapshot of the first few lines of the csv file. On the right is how the data appear in R.




```
> introq
  Gender Ideal Eye Statoph Nucoph SATM Year
1      1  68.0   4     8.0     50   NA 2105
2      2  64.0   4     4.0     40   NA 2105
3      2  68.0   4     5.0     30  430 2105
4      1  72.0   3    10.0    100  540 2105
```

Using commas as delimiters has the advantage of making it easier to deal with missing data. Look at the data for the first two subjects, above, left. There is just white space between the commas marking off the scores for SATM. Those two subjects were missing data on SATM. On the right, notice that R replaces missing values with its missing values code, "NA."

Here is the code used to read in the csv file:

```
introq <- read.table("C:/Users/Vati/Documents/StatData/IntroQ/IntroQ.csv", header=TRUE, sep=",")
```

I intent to use the "psych" package, so I activate it: `library(psych)`

Now for some basic descriptive statistics comparing men with women on height of ideal mate:

```
describeBy(introq$Ideal, introq$Gender)
```

```
group: 1
  vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
1    1 539 71.43 3.25     72   71.67 2.97  55  80    25 -0.97     2.82 0.14
------------------------------------------------------------
group: 2
  vars   n  mean   sd median trimmed  mad min max range skew kurtosis   se
1    1 180 66.59 3.27     66   66.42 2.97  55  78    23 0.44     1.23 0.24
```

Not surprisingly, the mean height of female students' ideal mates is greater than that of male students' ideal mates.

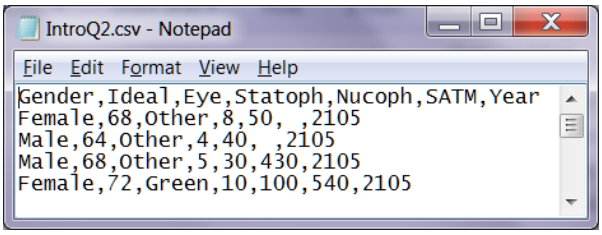
Now for an independent samples *t* test, comparing the two genders on height of ideal mate.

```
t.test(introq$Ideal ~ introq$Gender)

        Welch Two Sample t-test

data:  Ideal by Gender
t = 17.218, df = 305.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.282968 5.388255
sample estimates:
mean in group 1 mean in group 2
       71.43006        66.59444
```

It would be easier to deal with the output if I did not need to remember which numeric code stands for male and which for female. Fortunately, SPSS will, upon request, write value labels to the csv file, so I went back to SPSS and exported to csv with that request.




```
    Gender Ideal   Eye Statoph Nucoph SATM Year
1   Female  68.0 Other     8.0     50   NA 2105
2     Male  64.0 Other     4.0     40   NA 2105
3     Male  68.0 Other     5.0     30  430 2105
4   Female  72.0 Green    10.0    100  540 2105
```

As you can see, above, the csv file now has value labels, rather than numeric values, for the categorical variables. When I read in the new csv file and run the *t* test again, I get

```
        Welch Two Sample t-test

data:  Ideal by Gender
t = 17.218, df = 305.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.282968 5.388255
sample estimates:
mean in group Female   mean in group Male
           71.43006             66.59444
```

Notice that, by default, R does a separate variances *t* test. This is, IMHO, a good idea, but if you want a pooled variances *t* test, you can get it this way. Even though the sample sizes here differ quite a bit, the sample variances are nearly identical, so I am comfortable with the pooled test.

```
t.test(introq$Ideal ~ introq$Gender, var.equal=TRUE)

        Two Sample t-test

data:  Ideal by Gender
t = 17.274, df = 717, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.286033 5.385189
sample estimates:
mean in group Female   mean in group Male
           71.43006             66.59444
```

Since the unit of measure for our scores is well known to folks in the US, it is perfectly reasonable to use the simple difference between means (4.84 inches) as the effect size estimate. We might want to convert that to cm for folks more familiar with the metric system. Suppose, however, that our scores were estimates of political conservativism from the Überarschloch scale. Unless we are very familiar with that scale, we are not going to know whether a difference of 4.84 is a small difference or a large difference. In that case, it would be best to use Cohen's *d*, the standardized difference between group means. Here I shall use the "lsr" package to get Cohen's d

```
library(lsr)
cohensD(introq$Ideal~ introq$Gender)
```

```
[1] 1.487092
```

To put a confidence interval around *d*, I am going to use the "compute.es" package.

```
install.packages("compute.es")
```

```
Installing package into 'C:/Users/Vati/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/compute.es_0.2-4.zip'
Content type 'application/zip' length 272294 bytes (265 KB)
downloaded 265 KB
package 'compute.es' successfully unpacked and MD5 sums checked
```

```
library(compute.es)
des(d=1.487092, n.1=539, n.2=180)
```

```
Mean Differences ES:

 d [ 95 %CI] = 1.49 [ 1.3 , 1.67 ]
  var(d) = 0.01
  p-value(d) = 0
```

Presenting the Results

Students in Professor Karl's undergraduate statistics classes completed a brief survey. They were asked to indicate their sex/gender and the height, in inches, of their ideal mate. Mean height of ideal mate was significantly greater for female students (M = 71.43, SD = 3.25, n = 539) than for male students (M = 66.59, SD = 3.27, n = 180), $t(717) = 17.27$, $p < .001$, d = 1.49, 95% CI [1.30. 1.67].

- Wuensch's R Lessons