We continue our last project where we used R to find p-values. In this project we learn to find p-values for two-samples tests of mean and we see how to find p-values when we have summarized data.

1 Two sample tests

Last time we saw that when we wanted to do a t-test for a population mean when we had the data, we could use the `t.test()` function with the syntax

```
            t.test(x, mu = ..., alt = ...)
```

where `x` contained the data, `mu =` had a number for the mean under the null hypothesis, and `alt =` specified the alternative as one of `less`, `greater` or `two.sided`.

To do a two-sample test of equivalence of means is just as easy. For this, we need to specify the two samples and the alternative. Notice the null hypothesis is always just

$$H_0 : \mu_1 = \mu_2$$

so it need not be specified.

For example, this question was found on the web:

Women who are union members earn $2.50 per hour more than women who are not union members. (The Wall Street Journal, July 26, 1994). Suppose independent samples of 15 unionized women and 20 non-unionized women in manufacturing have been selected and the following hourly wage rates are found (Anderson, et al., 1998, p. 394).

Union Workers (n1= 15):

```
22.40   18.90   16.70   14.05   16.20   20.00   16.10   16.30   19.10   16.50
18.50   19.80   17.00   14.30   17.20
```

Nonunion Workers (n2= 20):

```
17.60   14.40   16.60   15.00   17.65   15.00   17.55   13.30   11.20   15.90
19.20   11.85   16.65   15.20   15.30   17.00   15.10   14.30   13.90   14.50
```

Question: Does there appear to be any difference in the mean wage rate between these groups?

To answer this, we enter in the data, and then use `t.test`

```
> x = scan()
1: 22.40  18.90  16.70  14.05  16.20  20.00  16.10  16.30  19.10  16.50
11: 18.50  19.80  17.00  14.30  17.20
16:
Read 15 items
> y = scan()
1: 17.60  14.40  16.60  15.00  17.65  15.00  17.55  13.30  11.20  15.90
11: 19.20  11.85  16.65  15.20  15.30  17.00  15.10  14.30  13.90  14.50
21:
Read 20 items
> t.test(x,y,alt="two.sided")

        Welch Two Sample t-test
```

```

data:  x and y
t = 3, df = 28, p-value = 0.005829
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.68 3.67
sample estimates:
mean of x mean of y
       18        15
```

Notice the small *p*-value indicating we should **reject** the null hypothesis.

We used `scan()` instead of `c()` to enter in the data as it allow me to cut and paste from the web site. You too could cut and paste from the pdf file if you wanted to do this example. Or you could type this in as in

```
> x = c(22.4,18.9,16.7,14.05,16.2,20,16.1,16.3,19.1,16.5,18.5,19.8,
+ 17,14.3,17.2)
```

Technically, we should have specified that we are assuming the variances are equal. If this is so, the *p*-values will be smaller, although not necessarily by much. For this example we would have

```
> t.test(x,y,alt="two.sided",var.equal=T)

        Two Sample t-test

data:  x and y
t = 3.0, df = 33, p-value = 0.004654
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.72 3.64
sample estimates:
mean of x mean of y
       18        15
```

Notice the difference in the degrees of freedom and the *p* value.

2 Using other commands when we have summarized data

We know how to use `prop.test()` and `t.test()` to do the significance tests covered in the class. They are pretty easy to use *if* you have all of the data. If you don't then you can still compute *p*-values, but the commands are different.

To compute a *p*-value the first thing to do is find the observed value of the test statistic. In the example above we have

```
   n     xbar   sd
x  15    17.5  2.2
y  20    15.4  2.0
```

The test statistic is then

```
> sp = sqrt( ((15-1)*2.2^2 + (20-1)*2^2)/(15 + 20 -2))
> T = (17.5 - 15.4)/( sp*sqrt(1/15 + 1/20))
> T
[1] 2.95
```

The test is two sided so we need to add the probability that a t-distribution with $15+20-2=33$ degrees of freedom is more than 2.95 or less than -2.95. We could use a table, but the “p”-functions do this for us. For example the probability we are less than -2.95 is

```
> pt(-2.95, df = 33)
[1] 0.0029
```

So, by symmetry, the p-value is twice this

```
> 2 * pt(-2.95, df = 33)
[1] 0.0058
```

The difference from the previous answer is due to rounding error.

For the normal distribution, the function to use is `pnorm()`. For example, to find the probability that a normal is bigger than 1.2 is done with

```
> 1 - pnorm(1.2)
[1] 0.115
```

3 Problems

For each, identify the two hypotheses, what test to use and find the p-value.

3.1 Suppose the following data is the price of a slice of pizza in Manhattan

```
1.75 2.15 2.25 1.85 2.15 2.25 1.65 2.50 1.75 1.85
```

Do a significance test to see if this is different than the price of a subway ride ($2).

3.2 Sam Sleepresearcher hypothesizes that people who are allowed to sleep for only four hours will score significantly lower than people who are allowed to sleep for eight hours on a cognitive skills test. He brings sixteen participants into his sleep lab and randomly assigns them to one of two groups. In one group he has participants sleep for eight hours and in the other group he has them sleep for four. The next morning he administers the SCAT (Sam’s Cognitive Ability Test) to all participants. (Scores on the SCAT range from 1-9 with high scores representing better performance). SCAT scores

```
8 hours sleep group (X) 5 7 5 3 5 3 3 9
4 hours sleep group (Y) 8 1 4 6 6 4 1 2
```

3.3 In 1879, A. A. Michelson made 100 determinations of the velocity of light in air using a modification of a method proposed by the French physicist Foucault. The numbers are in km/sec, and have had 299,000 subtracted from them.

The currently accepted ”true” velocity of light in vacuum is 299,792.5 km/sec. Do a test to see if these numbers are consistent with $\mu = 792$ or different.

```
850 740 900 1070 930 850 950 980 980 880 1000 980 930 650 760
810 1000 1000 960 960
```

3.4 Michelson did another run (even more actually). Do a two sample test to see if the next run has the same population mean as the last one or a different one.

```
960 940 960 940 880 800 850 880 900 840 830 790 810 880 880 830
800 790 760 800
```

3.5 The data between two types of processes give the following data on yields.

```
variable  count   mean   sd
yldA       13     549.3  168
yldB       16     555.7  104
```

Assume the variances are equal. Does the data indicate a difference in the means?

3.6 The effect of a calcium treatment on blood pressure is measured and the data is summarized by following difference in systolic pressure

```
n   xbar    sd
21  2.24   7.69
```

Do a *t*-test to determine if the mean change in systolic blood pressure is greater than 2.