228.371 Computer Lab: Confidence Intervals and Hypothesis Tests

Semester One 2015 - Week 1

Instructions: Read each section and try the commands. Then try the Stream worksheet questions suggested to test your knowledge. The worksheet is "adaptive" which means if you get an answer wrong, you can try again. This quiz is to help you monitor your progress, it does not count toward your mark.

Note that because of fonts, especially for symbols like quotation marks, cutting and pasting commands from this document occasionally will not work - you may have to retype.

1 Confidence intervals for proportions

To create confidence intervals for proportions, we can use the `R` function `prop.test`.

This function has many possible arguments, but to produce 95% confidence intervals for proportions the syntax is simple. Suppose we have done a survey about beverages, and 73 out of 100 people have responded "yes" to the question "have you had any coffee today?" The function is used this way:

```
prop.test(73, 100)
```

The output produced is:

```
        1-sample proportions test with continuity correction

data:  73 out of 100, null probability 0.5
X-squared = 20.25, df = 1, p-value = 6.795e-06
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.6303929 0.8116352
sample estimates:
   p
0.73
```

If all we are interested in is the confidence interval, we should ignore all the lines above where it says "95% confidence interval." You will notice this confidence interval is different from the one we learned about in class. It is not symmetric around the sample estimate. This is because rather than using the formula *estimate* ± *multiplier* × *std.error*, it is constructed by considering the test of whether $p = p_0$ vs $p \neq p_0$, and including in the confidence interval all p_0 where the null hypothesis of this test would not be rejected at the $1 - 0.95 = 0.05$ significance level. The difference is small for a large sample size. Both confidence intervals are interpreted in the same way: values within the confidence interval are considered plausible values for the population parameter.

Notice the default value for the confidence interval produced is 95%. If we want to change the confidence level used, we have to supply it as an argument, for example:

```
prop.test(73, 100, conf.level=0.98)
```

Also notice that R will produce a confidence interval even when it is not appropriate, i.e. $n\hat{p} < 10$ or $n(1 - \hat{p}) < 10$. Suppose we had only asked 10 people to participate in our coffee survey, and 7 had consumed coffee:

```
prop.test(7,10)
```

It will also produce an answer even when the data we give it are not consistent with estimating a proportion from a sample - e.g. we say a fractional number of people drank coffee:

```
prop.test(7.3, 10)
```

Try worksheet question 1.

2 Test for a single proportion

As you may have gathered, the function `prop.test()` also performs tests of proportions. The data is entered the same way as for a confidence interval. The default test has the null hypothesis $p = 0.5$ and is two sided. We can change these settings by giving arguments. For instance, suppose for our survey of 1000 people about coffee drinking, we are interested in testing whether the population proportion drinking coffee daily is less than 75%. Suppose 730 people respond 'yes'. We use the following code:

```
prop.test(730, 1000, p=0.75, alternative="less")
```

Note that a "one sided" confidence interval is given in this case; it will cover the population parameter 95% of the time, but will only bound it from above.

The test statistic used in R is the square of the one we have learned about; consequently it has a Chi-squared distribution rather than a normal distribution (which is why it is labelled X-squared in the output). It also incorporates a continuity correction, i.e. a slight modification to account for the fact that we are using continuous reference distributions, but our observed values must have the form $k/730$.

Without the continuity correction, (command given below) it is exactly equivalent to the test we have studied; i.e. it always gives the same p-value.

```
prop.test(730, 1000, p=0.75, alternative="less", correct=FALSE)
```

If you like, compute the test statistic we have learned about in class, and find the probability that we would observe a test statistic at least this negative if the null hypothesis were true (hint: use the `pnorm` function). You will see the test statistic is the square root of the one obtained from `prop.test()` with `correct=FALSE`, and the p-value is the same for both methods.

In general, using the continuity correction (the default) is slightly more accurate and is preferred.

Try worksheet questions 2-4

3 Difference in proportions

To create a confidence interval for the difference between two population proportions, enter the number of successes and the number of trials as vectors with one entry for each population. For instance, perhaps our coffee survey was conducted in Wellington and Auckland. We may wish to compare the two groups and create a confidence interval for the difference in proportions. Suppose the Wellington survey had 500 participants, and 378 of them had consumed coffee that day; Auckland also had 500 participants, and 352 had coffee. The commands

```
successes <- c(378, 352)
trials    <- c(500, 500)
prop.test(successes, trials)
```

produces a 95% confidence interval for the difference in proportions (Wellington-Auckland). What is your interpretation of the interval?

The default test when a vector of values is given for the number of successes and number off trials is a two sided test of whether the difference in sample proportions is zero.

Try worksheet question 5

4 t-distribution

We have learned that the sampling distribution of

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is a t-distribution when $\bar{x}$ has a roughly normal distribution. The t-distribution has heavier tails than the normal; this accounts for the extra variability introduced by estimating σ by s. The t-distribution is therefore used in constructing tests and confidence intervals for means.

Just as for the normal, uniform, and binomial distributions, we can find densities, tail probabilities, and quantiles for the t-distribution, and generate random numbers from the t-distribution, using the functions `dt()`, `pt()`, `qt()` and `rt()`. The degrees of freedom must be given as an argument. Try the following

1. Generate 100 random samples from a t-distribution from 1 degree of freedom and make a histogram. How does the picture differ from what you got when a normal distribution is used?

2. Plot a line showing the t-density with three degrees of freedom and the standard normal density on the same plot:

    ```
    x <- seq(-3,3, .01)
    plot (x, dnorm(x), type='l', lty=2)  # will give a dashed line
    lines (x, dt(x,3))                    # will add a solid line
    ```

Try worksheet question 6.

5 Tests and Confidence Intervals: single mean

Like the confidence intervals and tests for proportions, the confidence intervals and tests for a single mean are produced using the same function; this function is t.test. The argument given is the vector of data. Use the command `data(trees)` to load some data on cherry tree heights. The variable we will consider is called `Height`. The following would produce a 95% confidence interval for the average height of a tree from this population:

```
t.test(trees$Height)
```

Verify this is the same interval you get using the formula we discussed in class (use the `qt()` function to find the multiplier).

The default test uses the hypothesis pair $H_0 : \mu = 0$, $H_A : \mu \neq 0$. As for `prop.test`, the null value and alternative hypothesis can be changed via arguments. Suppose we are interested in determining whether the population of mature cherry trees has average height less than 79. The relevant hypothesis test can be performed with:

```
t.test(trees$Height, mu=79, alternative='less')
```

Try worksheet questions 7 and 8.