NRES 798 — Lab 7

Hypothesis tests

1 The ants data set

One way of entering the data from the example in Chapter 5 is:

```
> ants <- data.frame(habitat = c(rep('forest',6), rep('field',4)),
+ nests = c(9,6,4,6,7,10,12,9,12,10))
```

What does this do? Look-up the functions that you do not understand in the *Help*.

Do it, and list the data frame (`ants <Enter>`). These are counts of ant nests in 10 quadrats of 1 m^2, 6 in forests and 4 in fields.

Do a `summary`. Note that `habitat` was interpreted as a *factor*, in *R* terminology, while `nests` is a numerical variable.

Rows can be extracted by indexing, e.g., `ants[4:7, ]`, or `ants[c(4, 2, 7), ]`, or `ants[ants$nests > 10, ]`. Same for columns: `ants[, 2]`, or `ants[, 'nests']`. Or using list notation, `ants$nests`. Try it.

Sometimes it might be more convenient to have the counts in two simple vectors:

```
> forest <- ants$nests[ants$habitat == 'forest']
> field <- ants$nests[ants$habitat == 'field']
```

When working with data frames, some typing can be saved by making the names directly available:

```
> attach(ants)
> forest <- nests[habitat == 'forest']
> detach()
```

Forgetting to `detach` can lead to confusion, a safer alternative is:

```
>  forest <- with(ants, nests[habitat == 'forest'])
```

Here are some ways of visualizing the data as a box plot:

```
> boxplot(forest, field)
> boxplot(nests ~ habitat, ants)
> plot(nests ~ habitat, ants)
```

(remember that in *RStudio* pressing *Tab* gives a short description of function arguments, and *F1* gives the full *Help*). The `nests` $\sim$ `habitat` part is a simple example of a *formula*, used in many of the *R* statistical functions. Here it means plot `nests` (the dependent variable, on the left-hand side of the squiggle) over `habitat`.

2 Testing for differences between fields and forests

Remember the steps:

1. What is tested. H_0: fields and forests have a common normal distribution. Alternative: the means are different.
2. Test statistic. The F-ratio. We will not define it here, just note that under H_0 it tends to 1, and is larger under the alternative. The value calculated from our sample is 8.78. Is this large enough to reject H_0? Let's see...
3. Sampling distribution (if H_0 is true). An F distribution with parameters $m-1$ and $n-m$, where m is the number of groups and n is the number of observations.
4. Find the p-value for the observed test statistic. In this instance, the probability of the F-ratio being larger than 8.78 (if H_0 is true). How likely would that be? Calculate using the CDF of the F distribution, `pf` (you may want to draw a picture).

5. Compare to the significance level(s) α. Assign stars. How many?

Of course, there are easier ways of doing all that in *R*. But everything should be done "by hand" (we cheated a little) at least once. One way is

```
> oneway.test(nests ~ habitat, ants, var.equal=T)
```

Compare to your previous results. Look-up `oneway.test`, and make sure you understand what the parameters that we used are. Remember that you can always specify arguments by name instead of by position, e.g., `data = ants`.

Or we can use a sledgehammer, the general analysis of variance function

```
> aov(nests ~ habitat, ants)
> summary(aov(nests ~ habitat, ants))
```

Again, look it up. Both these tests use the F-ratio, we will see later how to do it with a t-test.

With *R*, the steps become: (1) Formulate the model, hypothesis, what to test? (2) Which test to use? (3) Find the relevant function and figure out how to call it. (4) Interpret the results.

3 More on the ants example

We know from before that this kind of quadrat counts are likely to be more Poisson that normal. And the Poisson mean and variance are the same, equal to the rate parameter λ. So assuming normals with different means and the same variance (in the alternative) might not be all that great. One way of making normal approximations better is to use variable transformations. For Poisson random variables, it is known that the square root is closer to normal, and its variance is more independent from the mean.

We will look at variable transformations sometime. For now, repeat `oneway.test` using the square root of the counts. Hint: `ants$sqrt <- sqrt(ants$nests)`. Any difference?

How can we know how much that matters in this example? How close the Poisson and the normal might be in this case? These days we do not need to guess. Plot the Poisson PDF for $\lambda = 8$, close to our mean and variance, over the range of our data: `plot(0:15, dpois(0:15, 8), type='h')`. Overlay a normal with the same mean and variance: `curve(dnorm(x, 8, sqrt(8)), add=T)`. What do you think?

For hypothesis testing, the upper tail area is more relevant. Do the same, but plotting the CDF's (use `type='s'` for `ppois` this time).

4 *t*-tests

The most common hypothesis tests may be tests for means. And of those, the *t*-tests are the most popular.

4.1 Example data

In order to show all the *t*-test varieties, let us use the CO2 data set included with *R*. For a description enter `?CO2`, and do a *summary*. Depending on your installation, you might have to load it with `data(CO2)` first.

We will use only the `uptake` and `Treatment` variables for the plants from Quebec: `co2 <- CO2[CO2$Type == 'Quebec', ]` (you could also drop the unused columns if you want).

4.2 Checking

First thing to do with any real data is (a) inspect the structure of the data, (b) check for gross measurement or recording errors and outliers, and (c) see if the model assumptions are reasonable.

For (a) and (b), let's simulate a number recorded with the decimal point in the wrong place: `co2.bad <- co2; co2.bad$uptake[9] <- co2.bad$uptake[9] * 10`.

For `co2$uptake` and `co2.bad$uptake`, do `summary`, `plot`, `boxplot`, `hist`. Can you detect the outlier? Anything suspicious with the "good" `co2` data? What is on the x-axis of the `plot` graph?

For (c), we will assume a normal distribution. A good way of checking for approximate normality is to plot the sample quantiles (sorted observation values) over the theoretical quantiles of the normal distribution. A normal sample would approximate a straight line, with slope and intercept depending on the mean and variance (think this over when you have some time!). The graph is produced with `qqnorm`. A line through the 1st and 3rd quartiles, for comparison, is added by `qqline`. Try it.

A formal hypothesis test for normality is the Shapiro-Wilk test (`shapiro.test`). Try that too. This is one case where we do not want to reject H_0.

Actually, this normality is not all that important. All we need is for the means to be approximately normal, and the Central Limit Theorem makes that easier.

4.3 One-sample t-test

Here H_0 is that the population is normal, with a given mean μ. The alternative may be that the real population mean is greater than μ, smaller than μ, or different from μ (two-sided). Typically one tests against $\mu = 0$, but for purposes of demonstration let's test if the uptake mean is greater than 40.

The test statistic is the difference between the sample mean and μ, divided by the sample standard error:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \, .$$

If H_0 is true, this t has a Student t distribution with $n-1$ degrees of freedom.

Carry out the test using this information and the t PDF (`?Distributions`).

All the t tests are produced in *R* with the function `t.test`. See the *Help*, and use that to check your results above. Hint: `x` is `co2$uptake`, there is no `y`, `mu = 40`, and `paired = FALSE`.

4.4 Two-sample t-test

H_0 is that the two groups (here *chilled* and *nonchilled*) belong to the same normal population. The alternatives may be that the population mean of

the first group is smaller than, larger than, or different from that of the second group. The variances may be assumed equal or not.

The test statistic is the difference between the two sample means, divided by the quadratic mean of the sample standard errors:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \, .$$

This is if the variances are not assumed the same, otherwise the whole sample variance s^2 is used instead of s_1^2 and s_2^2.

If H_0 is true, this t has a Student t distribution with $n_1 + n_2 - 2$ degrees of freedom.

The `t.test` function can be used with `x` and `y` being the values of `uptake` for *chilled* and *unchilled*, respectively. Somewhat easier is to use the formula version: `t.test(uptake ~ Treatment, co2)`. Try it.

This is the same as the ants problem (assuming equal variances). Apply this t-test to the ants data set, and compare the conclusions to those from the F test. The t-test can be used with unequal variances, but the F-ratio test can be used with more than two groups.

4.5 Paired t-test

When feasible, it is more efficient to apply the treatments to the same experimental units, or to reasonably homogeneous groups (or strata), and compare within them (why?). Although not true, for an example we can pretend that the chilling treatment was applied to the same 21 plants, so that we have two uptake measurements for each plant, with and without chilling. The comparison can then be done using the differences for each plant. Another example would be counting ant nests in different sites.

The hypotheses are the similar to those in the two-sample test, except that there can be differences among units/groups/strata. The test statistic, is like the one-sample test applied to the differences. The `t.test` function is used with `x` and `y` being the two measurements (two treatments), and `paired` set to `TRUE`.

When you have time, look at the nonparametric competitor, the Wilcoxon tests (`wilcox.test`). Most of the common and not so common hypothesis tests available are listed by `apropos('*.test')`.