Lecture 4: Classical Statistical Tests

Trivia:
Who was the “Student” behind Student’s T-Test? Where did he work?
http://blogs.sas.com/content/jmp/2013/10/07/celebrating-statisticians-william-sealy-gosset-a-k-a-student/

What We’ve Covered So Far:

1. Backbone of R: assigning variables, saving/exporting objects, getting help, simple calculations, using new packages
2. Creating and subsetting matrices and data frames
3. Standard sample statistics (mean, variance, range, etc.)
4. Generating different types of tables
5. Making various plots (box, scatter, dot, bar)
6. Conditional indexing to select certain data (necessary skill)
7. if and ifelse() statements, renaming variables, missing values
8. Functional syntax and how to write your own functions
9. Various additional topics not graded (fancy graphics, merging datasets)

The rest of the course will use the tools we have learned so far, but the focus will be on applying and interpreting various statistical tests in R.

Outline:
In this lecture we will discuss how to conduct basic (classical) statistical tests in R:
1. One-sample tests
   (Is my sample different from some hypothesized value? What’s the plausible range of values for the population statistic, given my sample?)
2. Two-sample tests  
   (Are my two (sub) populations the same?)

3. Tests on more than two samples  
   (Are my $k$ populations/treatments/exposures the same?)

**Objectives:**
Students will be able to:
1. Use R to perform one-sample tests using t-tests, Wilcoxon signed-rank test
2. Use R to perform two-sample tests using t-tests, paired t-tests, and Wilcoxon tests
3. Use R to perform one-way analysis of variance (ANOVA) and Kruskal-Wallis tests
4. **Interpret test results.**

**A note on testing and inference**
My first statistics teacher told me “You can’t do inference without a model.” What he meant by that is that any time you make a decision (inference) about your data, usually by way of a p-value ($p<0.05$ or whatever), you are doing so in the context of a statistical model, a “null” hypothesis. Ultimately, you choose what that model is.

In plainer language, in any statistical test, you’re comparing your observed data to hypothetical data that could come from a particular statistical model. You’re asking the question: “If this model were the true data-generating mechanism (for example, ‘if the true population mean of my data were 200’), how likely would it be that I would see the data I actually did see?”

That is why we never accept the null hypothesis. The null hypothesis is the very basis of our model. We assumed it to begin with, so we can’t then say definitively “the null is true” if we fail to reject the alternative hypothesis (usually the case when we find a p-value greater than 0.05, or some prespecified threshold).

So we reject the null, but we never accept the null—we only fail to reject it.

If you’re scratching your head at this, don’t fret. It takes a while to become comfortable with this type of thinking. Statistical inference is in the broader field of decision theory, so it is not an easy topic.
4.1 One-Sample Test
Suppose we have a sample of continuous data. This could be blood pressure or height in inches or a microarray luminosity score.

The questions we might want to answer are:
- What is the mean value of my sample? Of my population the sample comes from?
- Is the mean value significantly different from a given or theoretic value?
- Is the mean a useful characterization of these data?

A one-sample test compares the mean from a sample to a hypothesized value that is pre-specified in your null hypothesis (so by you).

4.1.1 Parametric One-sample T-test

**Boston Data and Assumption Checking**
A one-sample t-test is a test based on the normality and independence assumptions (in probability jargon, “IID”: independent, identically-distributed random variables). Checking these assumptions before analyzing data is necessary.

We will use a dataset of a study of Boston housing, which includes several potential explanatory variables. The general question of what factors determine housing values (well, at least in Boston 1970!) is of interest. It contains 506 census tracts in the Boston Standard Statistical Metropolitan Area (SMSA) in 1970. This dataset is available within the MASS library and can be access via the name Boston. Here is a description of the dataset:

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>crim</td>
<td>Per capita crime rate by town</td>
</tr>
<tr>
<td>zn</td>
<td>Proportion of residential land zoned</td>
</tr>
<tr>
<td>indus</td>
<td>Proportion of non-retail business acres per town</td>
</tr>
<tr>
<td>chas</td>
<td>Charles River dummy variable (1 if tract bounds river, 0 otherwise)</td>
</tr>
<tr>
<td>nox</td>
<td>nitrogen oxide concentration (parts per 10 million)</td>
</tr>
<tr>
<td>rm</td>
<td>average number of rooms per dwelling</td>
</tr>
<tr>
<td>age</td>
<td>proportion of owner-occupied units built prior to 1940</td>
</tr>
<tr>
<td>dis</td>
<td>weighted mean of distances to five Boston employment centers</td>
</tr>
<tr>
<td>rad</td>
<td>index of accessibility to radial highways</td>
</tr>
<tr>
<td>tax</td>
<td>full-value property-tax rate per $10,000</td>
</tr>
<tr>
<td>ptratio</td>
<td>pupil-teacher ratio by town</td>
</tr>
</tbody>
</table>
As usual, we begin with a set of single sample plots along with some summary statistics.

```r
> summary(rm)
     Min.  1st Qu.   Median     Mean   3rd Qu.     Max. 
```

This `summary()` function gives us six pieces of information about our variable `rm`. The mean and median are close to each other and so we expect that the variable `rm` is symmetric.

Here are some plots to look more closely at these data.

```r
> par(mfrow=c(2,2))
> plot(rm)
> hist(rm, main= "Histogram of number of Rooms")
> qqnorm(rm, main= "QQ-plot for number of Rooms")
> qqline(rm)
> boxplot(rm)
```
These plots give a good indication of the normality (or lack thereof) of the average number of rooms per tract.

Alternatively, we may conduct a formal statistical test, the Shapiro-Wilk test, to see whether the data come from a normal distribution.

$H_0$: The data are sampled from a normal distribution

$H_1$: The data are not sampled from a normal distribution

(notice it doesn’t say anything else: there is no claim as to the specific alternative distribution—Gamma or Poisson or whatever).
A small p-value (usually < 0.05, by convention), as shown below, means that we **reject**
the null hypothesis, and conclude that the data is not normally distributed.

> shapiro.test(rm)

```
Shapiro-Wilk normality test
data:  rm
W = 0.9609, p-value = 2.412e-10
```

The function outputs a test statistic (W), and a p-value. At the 0.05 α-level we reject the
null hypothesis that the `rm` variable was sampled from a normal distribution (W = 0.9609,
p-value = 2.412e-10).

(Note: I have not displayed the formula for this test because it is rather complicated.)

The assumption of normality for a t-test is actually quite flexible. We’re really interested
in the normality of our sample to decide if the mean is even **useful** as a summary statistic
of the data.

The mean in the plot above is 9.85. We can test the mean and get a confidence
interval for it, but is the mean even a useful measurement when the data look like
this?
The next assumption is *independence*. In this case, the data were collected from different census tracts and we **assume** that each census tract is independent from each other, and hence the number of rooms can be assumed to be independent as well. Is this a reasonable assumption?

1. **One-sample t-test for the mean $\mu$**

Suppose we are interested in testing whether the average number of rooms per dwelling in Boston in 1970 equals 6. The assumptions for a one-sample t-test are:

1. Independent observations
2. Sample drawn from a Normal distribution

Test statistic (picture from Wikipedia):

\[
t = \frac{\bar{X} - \mu}{s / \sqrt{n}}
\]

We can use `t.test()` function in R. R performs a two-tailed test by default, which is what we need in our case.

\[
> \text{t.test(rm, mu=6)}
\]

One Sample t-test

data:  rm
t = 9.1126, df = 505, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 6
95 percent confidence interval:
6.223268 6.346001
sample estimates:
mean of x
6.284634

H₀: The mean of the number of rooms per dwelling is equal to 6
H₁: The mean of the number of rooms per dwelling is not equal to 6

The point estimate of the population mean is 6.28, and the 95% confidence interval is from 6.223 to 6.346. The hypothesis testing p-value is smaller than the standard 0.05 ( <2.2e-16, t = 9.11, df = 505), which leads us to reject the null hypothesis that the mean number of rooms per dwelling is equal to 6. Thus we conclude that the average number of rooms per dwelling in Boston does not equal 6. (That’s all we can say given our hypothesis!)

What’s a confidence interval??

If the null is true (I’m sampling from the distribution I think I am), then the probability my t-statistic is between two values that define an area under the probability density of 95% is—by construction—95%.

\[ P \left( -t_{\alpha/2} < T < t_{\alpha/2} \right) = 0.95 \]

which by the definition of the t-statistic means:

\[ P \left( -t_{\alpha/2} < \frac{\overline{X} - \mu}{s/\sqrt{n}} < t_{\alpha/2} \right) = 0.95 \]

Then I can rearrange this statement algebraically to yield

\[ P \left( \overline{X} - \frac{s}{\sqrt{n}} * t_{\alpha/2} < \mu < \overline{X} + \frac{s}{\sqrt{n}} * t_{\alpha/2} \right) = 0.95 \]

So now I have an interval for my true mean, which, if the null hypothesis is true, means that 95% of the time, my true population mean will fall within a given computed confidence interval.

This just gives a plausible range of values for the population mean. A CI is not saying there’s a 95% chance the population mean is in my interval (the value either is or is not in that interval): it’s saying that if I were to repeat my sampling a large number of times, 95% of the generated CIs would contain the true value.
Here’s a plot of many confidence intervals generated from random data about true mean $\mu$.

Exercise 1. Location, location, location! People are always concerned about location when looking for a home. Suppose we are interested in the weighted distance to five Boston employment centers and would like to know if the average distance is around 3.5 miles.

(a) Conduct a $t$-test for testing whether the average distance is around 3.5.
(b) What is the mean and median distance?
(c) Is the distance normally distributed? Use plots and a formal normality test to decide.
(d) Is the one-sample t-test appropriate for these data?

4.1.2 Non-parametric Wilcoxon Signed-Rank Test

A quick overview of parametric vs. nonparametric testing: [http://www.mayo.edu/mayo-edu-docs/center-for-translational-science-activities-documents/berd-5-6.pdf](http://www.mayo.edu/mayo-edu-docs/center-for-translational-science-activities-documents/berd-5-6.pdf)

So what if the normality assumption fails? Here is where the non-parametric test comes in. The Wilcoxon Signed rank test can compare the median to a hypothesized median value. For example, in our case,

```r
> wilcox.test(dis, mu=3.5)

   Wilcoxon signed rank test with continuity correction
data:  dis
V = 67110, p-value = 0.3661
alternative hypothesis: true location is not equal to 3.5

H_0: The median weighted distance is equal to 3.5
H_1: The median weighted distance is not equal to 3.5

We fail to reject the null hypothesis that the median weighted distance is equal to 3.5.
V = 67110, p-value = 0.3661.

The `wilcox.test()` function conducts a two-sided test by default but a one-sided test is also available by changing the `alternative` argument. The `alternative = “less”` option would test the null hypothesis that the median weighted distance is less than 3.5; the `alternative = “greater”` option would test the null hypothesis that the median weighted distance is more than 3.5. The output for `wilcox.test()` is compact but other information such as confidence intervals can be requested if necessary.

Note that the Wilcoxon signed-rank test still assumes independence, although it relaxes the normality assumption.

```r
wilcox.test(dis, mu = 3.5, alternative="less")
wilcox.test(dis, mu = 3.5, alternative= "greater")
```
Exercise 2. (a) Conduct a Wilcoxon signed-rank test for determining whether the median number of rooms is significantly different from 6. (b) Compare the result with section on the one-sample test of the mean.

4.2 Two-sample Test
Given the variation within each sample, how likely is it that our two sample means were drawn from populations with the same average? A better way to answer this question is to work out the probability that our two samples were indeed drawn from populations with the same mean. If this probability is very low (i.e. small p-value), then we can be reasonably certain that the means really are different from one another.

An example question here could be: “Do men and women have different average systolic blood pressures?”

We’ll start with “paired data.”

4.2.1 Two-sample Paired Test
Paired tests are used when there are two measurements on the same experimental unit. The paired t-test has the same assumptions of independence and normality as a one-sample t-test.

Let us look at a data set on weight change (anorexia), also from the MASS library. The data are from 72 young female anorexia patients. The three variables are treatment (Treat), weight before study (Prewt), and weight after study (Postwt). Here we are interested in finding out whether there is a placebo effect (i.e. patients who do not get treated gain some weight in the study).

```r
> detach(Boston)### important
> attach(anorexia)
> ?anorexia
> dif <- Postwt - Prewt
> dif.Cont <- dif[which(Treat=="Cont")]
```
Exercise 3. (a) Apply the `summary()` function to variable `dif.Cont` and comment on the summary statistics. (b) Create plots to test the normality assumption. (c) Conduct a formal test for normality. (d) Does the normality assumption for variable `dif.Cont` hold?

Conducting a “paired” t-test is virtually identical to a one-sample test on the element-wise differences. Both the parametric pair-wise t-tests and non-parametric Wilcoxon signed-rank tests are shown below.

```r
> t.test(dif.Cont)

    One Sample t-test

data:  dif.Cont
    t = -0.2872, df = 25, p-value = 0.7763
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:  
   -3.676708  2.776708  
sample estimates:  
    mean of x  
        -0.45

H₀: There is, on average, no difference in mean weight before and after the study period for those in the Control group.
H₁: There is, on average, some difference in mean weight before and after the study period for those in the Control group.

We fail to reject the null hypothesis at the 0.05 level (t = -0.29, df = 25, p-value = 0.7763), and conclude that there is insufficient evidence to conclude a difference in mean weight before and after the study in the Control group. The sample mean difference is equal to -0.45 with a 95% confidence interval of (-3.67, 2.77).

> wilcox.test(dif.Cont)

    Wilcoxon signed rank test with continuity correction

data:  dif.Cont
    V = 150, p-value = 0.7468
alternative hypothesis: true location is not equal to 0

H₀: The median difference in weight before and after the study period for those in the Control group is equal to 0.
H₁: The median difference in weight before and after the study period for those in the Control group is not equal to 0.

We thus fail to reject the null hypothesis (V = 150, p-value = 0.7468) that there is no difference in the median birth weight before and after the study in the Control group.

It is not necessary to create the derived difference variable as we did here (dif.Cont). Instead, you may turn on the paired option in the R command as follows:

```r
> t.test(Postwt[which(Treat=="Cont")],
Prewt[which(Treat=="Cont")], paired=TRUE)
```

```r
> wilcox.test(Prewt[which(Treat=="Cont")],
Postwt[which(Treat=="Cont")], paired=TRUE)
```

**Exercise 4.**
There are three treatment levels in the *anorexia* dataset.

Conduct an appropriate test to determine whether *any* treatment is effective.

(Hint: Create a new variable called *trt* using conditional indexing—use an `ifelse()` statement—that is designated 0 if the patient was not given treatment, and 1 if the person received FT or CBT).

---

**4.2.2 Parametric Two-sample T-test**
Now, we will analyze the *Pima.tr* dataset. The US National Institute of Diabetes and Digestive and Kidney Diseases collected data on 532 women who were at least 21 years
old, of Pima Indian heritage and living near Phoenix, Arizona, who were tested for
diabetes according to World Health Organization criteria. One simple question is whether
the plasma glucose concentration is higher in diabetic individuals than it is in non-
diabetic individuals. So now we have two populations of interest.

To do this, we will perform a two-sample t-test, which makes the following assumptions:

1. Independent observations,
2. Normal distribution for each of the two groups, and crucially,
3. Equal variance for each of the two groups

The statistic (assuming equal population variances) is

\[ t_{\text{two-sample}} = \left( \bar{Y}_1 - \bar{Y}_2 - D_0 \right) / \left[ S_p^2 \left( 1/n_1 + 1/n_2 \right) \right] \sim T_{(n_1 + n_2 - 2)} \]

(usually \( D_0 \) is just 0)

\[ S_p^2 \text{(pooled variance)} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \]

> detach(anorexia)
> attach(Pima.tr)
> ?Pima.tr

The syntax for this command is pretty straightforward:

> t.test(glu ~ type)

    Welch Two Sample t-test

    data:  glu by type
    t = -7.3856, df = 121.756, p-value = 2.081e-11
    alternative hypothesis: true difference in means is not equal to 0
    95 percent confidence interval:
    -40.51739 -23.38813
    sample estimates:
    mean in group No mean in group Yes
    113.1061          145.0588
H₀: The mean glucose for those who are diabetic is the same as those who are not diabetic.
H₁: The mean glucose for those who are diabetic is not the same as those who are not diabetic.

Based on our pre-specified p-value threshold of 0.05, we reject the null hypothesis that the mean glucose for those who are diabetic is the same as those who are not diabetic (t = -7.39, df = 121.76, p-value < 2.081e-11). The average glucose for those who are diabetic is 145.06 and for those who are not diabetic is 113.11. The 95% confidence interval for the difference in glucose between the two groups is (-40.52, -23.38).

One thing to remember about the t.test() function is that it assumes the variances are different by default. The argument var.equal=T can be used to accommodate the scenario of homogeneous variances.

(The unequal variances formula is known as Satterthwaite’s formula—the degrees of freedom are approximated in the case of unequal variances.)

cf.

> t.test(glu ~ type, var.equal=T)

In other words, we need to determine if the different groups have the same spread. As we did in the normality checking, we can collect information from summary statistics, plots and formal tests and then make a final judgment call.

Exercise 5. Are the variances of the plasma glucose concentration the same between diabetic individuals and non-diabetic individuals? Use the summary statistics and plots to support your argument.
Comparison of Variance. R provides the `var.test()` function for testing the assumption that the variances are the same, this is done by testing to see if the ratio of the variances is equal to 1. The test of variances is called the same way as `t.test()`:

```r
> var.test(glu ~ type)

  F test to compare two variances
data:  glu by type
F = 0.7821, num df = 131, denom df = 67, p-value = 0.2336
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.5069535 1.1724351
sample estimates:
  ratio of variances
           0.7821009
```

\( H_0 \): The variance in glucose for diabetics is equal to the variance in glucose for non-diabetics.

\( H_1 \): The variance in glucose for diabetics is not equal to the variance in glucose for non-diabetics.

We fail to reject the null hypothesis that the variance in glucose is equal to the variance in glucose for non-diabetics (\( F_{131,67} = 0.7821 \), p-value = 0.2336). The ratio of the variances is estimated to be 0.78 with a 95% confidence interval of (0.51, 1.17). Notice we did not reject the null, and the 95% CI contains 1.

So here for our t-test we would use the `var.equal=T` option.

4.2.3 Non-parametric Wilcoxon Test

To perform a nonparametric equivalent of a 2 independent sample t-test we use the Wilcoxon rank sum test. To perform this test in R we need to put the formula argument into the `wilcox.test` function, or provide two vectors for the test. The script below shows one example:

```r
> wilcox.test(glu ~ type)
```
Wilcoxon rank sum test with continuity correction
data:  glu by type
W = 1894, p-value = 2.240e-11
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(glu[type=="Yes"],glu[type=="No"])
  # alternative way to call the test

**Hypotheses:**

H$_0$: The median glucose for those who are diabetic is the same as the median of glucose for those who are not diabetic.

H$_1$: The median of glucose for those who are diabetic is not the same as the median of glucose for those who are not diabetic.

We reject the null hypothesis that the median glucose for those who are diabetic is equal to the median glucose for those who are not diabetic ($W = 1894$, p-value $= 2.24e-11$).

### 4.3 Tests for more than two samples

In this section, we consider comparisons among more than two groups parametrically, using analysis of variance (ANOVA), as well as non-parametrically, using the Kruskal-Wallis test.

#### 4.3.1 Parametric Analysis of Variance (ANOVA)

To test if the means are equal for more than two groups we perform an analysis of variance test. An ANOVA test will determine if the grouping variable explains a significant portion of the variability in the dependent variable. If so, we would expect that the mean of your dependent variable will be different in each group. The assumptions of an ANOVA test are as follows:
1. Independent observations
2. The dependent variable follows a normal distribution in each group
3. Equal variance of the dependent variable across groups

Here, we will use the *Pima.tr* dataset. According to National Heart Lung and Blood Institute (NHLBI) website ([http://www.nhlbisupport.com/bmi/](http://www.nhlbisupport.com/bmi/)), BMI can be classified into 4 categories:

- Underweight: < 18.5
- Normal weight: 18.5 ~ 24.9
- Overweight: 25 ~ 29.9
- Obesity: >= 30

(Online app to calculate BMI: [http://www.nhlbi.nih.gov/guidelines/obesity/BMI/bmicalc.htm](http://www.nhlbi.nih.gov/guidelines/obesity/BMI/bmicalc.htm))

Exercise 6. (a) Create a categorical variable *bmi.new* to categorize the continuous *bmi* variable into four classes based on the definition shown above. Note that we have very few underweight individuals, so collapse underweight and normal weight into “Normal/under weight” (b) Report the number of individuals in each category. (c) Calculate the average glucose concentration in each category.

An Aside:
In this *Pima.tr* dataset the BMI is stored in numerical format, so we need to categorize BMI first since we are interested in whether categorical BMI is associated with the plasma glucose concentration. In the Exercise, you can use an “if-else-” statement to create the *bmi.cat* variable. Alternatively, we can use `cut()` function as well. Since we have very few individuals with BMI < 18.5, we will collapse categories “Underweight” and “Normal weight” together.

```r
> bmi.label <- c("Underweight/Normalweight", "Overweight", "Obesity")
> summary(bmi)

> bmi.break <- c(18, 24.9, 29.9, 50)
> bmi.cat <- cut(bmi, breaks=bmi.break, labels = bmi.label)
```
> table(bmi.cat)
bmi.cat
Underweight/Normal weight       Overweight       Obesity
       25                43                132

> tapply(glu, bmi.cat, mean)
          Normal/under weight  Overweight  Obesity
                108.4800 116.6977   129.2727

Back to ANOVA…
Suppose we want to compare the means of plasma glucose concentration for our four BMI categories. We will conduct analysis of variance using `bmi.cat` variable as a factor.

> bmi.cat <- factor(bmi.cat)
> bmi.anova <- aov(glu ~ bmi.cat)

Before looking at the result, you may be interested in checking each category’s glucose concentration average. One way it can be done is using the `tapply()` function. But alternatively, we can also use another function.

> print(model.tables(bmi.anova, "means"))

Tables of means
Grand mean
123.97

  bmi.cat
          Underweight/Normal weight  Overweight  Obesity
                108.5          116.7   129.3

  rep
        25.0          43.0   132.0

Apparently, the glucose level varies in different categories. We can now request the ANOVA table for this analysis to check if the hypothesis testing result matches our observation in summary statistics.

> summary(bmi.anova)

       Df  Sum Sq  Mean Sq   F value    Pr(>F)
bm.cat    2   11984    5992 6.2932482 0.0022417 **
Residuals 197 187575   952

H₀: The mean glucose is equal for all levels of `bmi` categories.
H_1: At least one of the bmi categories has a mean glucose that is not the same as the other bmi categories.

We reject the null hypothesis that the mean glucose is equal for all levels of bmi categories (F_{2,197} = 6.29, p-value = 0.002242). The plasma glucose concentration means in at least two categories are statistically significantly different.

Naturally, we will want to know which category pair has different glucose concentrations. This type of analysis is often called contrasts. One way to answer this question is to conduct several two-sample tests and then adjust for multiple testing using the Bonferroni correction.

Performing many tests will increase the probability of finding one of them to be significant; that is, the p-values tend to be exaggerated (our type I error rate increases). A common adjustment method is the Bonferroni correction, which adjusts for multiple comparisons by changing the level of significance \( \alpha \) for each test to \( \alpha / \text{(# of tests)} \).

If we were performing 10 tests, to maintain a level of significance \( \alpha \) of 0.05 we adjust for multiple testing using the Bonferroni correction by using 0.05/10 = 0.005 as our new level of significance.

A function called pairwise.t.test computes all possible two-group comparisons.

```r
> pairwise.t.test(glu, bmi.cat, p.adj = "none")

   Pairwise comparisons using t tests with pooled SD

data:  glu and bmi.cat

 Underweight/Normalweight Overweight
Overweight 0.2910 -
Obesity 0.0023 0.0213

P value adjustment method: none
```
From this result we reject the null hypothesis that the mean glucose for those who are obese is equal to the mean glucose for those who are underweight/normal weight (p-value = 0.0023). We also reject the null hypothesis that the mean glucose for those who are obese is equal to the mean glucose for those who are overweight (p-value = 0.0213). We fail to reject the null hypothesis that the mean glucose for those who are overweight is equal to the mean glucose for those who are underweight (p-value = 0.2910).

We can also make adjustments for multiple comparisons, like so:

```r
> pairwise.t.test(glu, bmi.cat, p.adj = "bonferroni")

Pairwise comparisons using t tests with pooled SD
data: glu and bmi.cat

                      Underweight/Normal weight Overweight
Overweight          0.8729
Obesity            0.0069   0.0639
```

P value adjustment method: bonferroni

However, the Bonferroni correction is very conservative. Here, we introduce an alternative multiple comparison approach using Tukey’s procedure:

```r
> TukeyHSD(bmi.anova)

Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = glu ~ bmi.cat)

$bmi.cat

                      diff      lwr      upr     p adj
Overweight-Underweight/Normalweight 8.217674 -10.1099039 26.54525 0.5407576
Obesity-Underweight/Normal weight  20.792727  4.8981963 36.68726 0.0064679
Obesity-Overweight             12.575053 -0.2203125 25.37042 0.0552495
```

From the pairwise comparison, what do we find regarding the plasma glucose in the different weight categories?

It is important to note that when testing the assumptions of an ANOVA, the `var.test` function can only be performed for two groups at a time. To look at the assumption of equal variance for more than two groups, we can use side-by-side boxplots:

```r
> boxplot(glu~bmi.cat)
```
To determine whether or not the assumption of equal variance is met we look to see if the spread is equal for each of the groups.

We can also conduct a formal test for homogeneity of variances when we have more than two groups. This test is called **Bartlett’s Test**, which assumes normality. The procedure is performed as follows:

```r
> bartlett.test(glu~bmi.cat)

Bartlett test of homogeneity of variances

data:  glu by bmi.cat
Bartlett's K-squared = 3.6105, df = 2, p-value = 0.1644
```

H₀: The variability in glucose is equal for all bmi categories.
H₁: The variability in glucose is not equal for all bmi categories.
We fail to reject the null hypothesis that the variability in glucose is equal for all \textit{bmi} categories (Bartlett’s K-squared = 3.6105, df = 2, p-value = 0.1644). So using ANOVA here is fine.

### 4.3.2 Non-parametric Kruskal-Wallis Test

ANOVA is a parametric test and it assumes normality as well as homogeneity of variance (and what else?).

What if the assumptions fail? Here, we introduce its counterpart on the non-parametric side, the \textit{Kruskal-Wallis Test}.

As in the Wilcoxon two-sample test, data are replaced with their ranks.

```r
> kruskal.test(glu~bmi.cat)

Kruskal-Wallis rank sum test

data:  glu by bmi.cat
Kruskal-Wallis chi-squared = 12.7342, df = 2, p-value = 0.001717
```

H$_0$: The distribution of glucose is the same for each bmi category.
Ha: The distribution of glucose is not the same for each bmi category.

We see that we reject the null hypothesis that the distribution of glucose is the same for each bmi category at the 0.05 $\alpha$-level. ($\chi^2 = 12.73$, p-value = 0.001717).

**Exercise 7**

Conduct an appropriate test (check the normality and equal variance assumptions) to determine if the plasma glucose concentration levels are the same for the non-diabetic individuals across different age groups. People are classified into three different age groups: group1: < 30; group2: 30-39; group3: $\geq$ 40.
Recap:
- Normality Testing
- One- and Two-Sample tests: $t$, Wilcoxon, paired $t$
- Single factor ANOVA & Kruskal-Wallace tests
- Bartlett’s Test for equal variances

Reading:
1. VS Chapter 8.3

Assignment:
1. Homework 3 due and Homework 4 assigned.
2. Select your dataset and study question for your project, and next week bring your project proposal to class.