CAS R Workshop

Hypothesis Testing

Nathaniel T. Stevens

October 13, 2015

Introduction

In the context of data-driven decision-making, a hypothesis is a statement of interest that one wishes to prove or disprove using collected data. A hypothesis is tested by comparing one's observed data with a hypothesized statistical distribution. For this task we define the "null hypothesis" denoted H_0 and the "alternative hypothesis" denoted H_A.

By convention, the null hypothesis is assumed true unless we have sufficient evidence to disprove it. Within this framework, based on the observed data, we either reject the null hypothesis, or we fail to reject it.

An analogy can be drawn between hypothesis testing logic and criminal sentencing in the courtroom: in today's justice system a defendant is assumed innocent until proven guilty. One cannot prove for certain that the defendant is guilty, but with enough evidence the jury can be convinced that the defendant is guilty. Similarly, we can never prove for certain that H_0 is false, but with enough evidence against it, we can choose to reject it.

However, just like in the courtroom, we can fall victim to wrongful conviction. In making such decisions, we can make two types of errors based on whether the null hypothesis H_0 is true or false. The two types of errors are as follows:

- Type I Error: Based on the observed data we reject H_0 when it is in fact true
- Type II Error: Based on the observed data we accept H_0 when it is in fact false

In terms of the courtroom analogy a Type I Error is equivalent to sentencing an innocent person, and a Type II Error is equivalent to letting a criminal go free.

Clearly we would like to reduce the likelihood of committing either type of error. We denote the probability of committing a Type I Error as α, which is sometimes referred to as the *significance level* of the test. A common choice of α is 0.05. We denote the probability of committing a Type II Error as β. A closely related quantity is $1-\beta$ which is referred to as the *power* of the test. The power is the probability that we correctly reject the null hypothesis when it is indeed false. Clearly we would like the power of the test to be large.

For purposes of illustration, suppose that someone postulates that the mean of a population is equal to 100, while others believe the that the population mean is not equal to 100. Formally, this hypothesis would be stated as follows:

$$H_0: \mu = 100 \text{ vs. } H_A: \mu \neq 100$$

If the data provide sufficient evidence, we reject H_0 in favor of H_A. If we reject H_0, we conclude that the data suggests μ is different from 100; it could be larger or it could be smaller. Such a hypothesis is considered "two-sided" because the alternative hypothesis, H_A, is two-sided. If, however, one believed that the population mean was actually larger than 100, the corresponding hypothesis statement would be the following:

$$H_0: \mu = 100 \text{ vs. } H_A: \mu > 100$$

In this "one-sided" hypothesis test, a rejection of H_0 means that the data suggests μ is larger than 100.

While the content of a hypothesis statement will change by context, and from one problem to another, the general framework is consistent. That is, null and alternative hypotheses must be defined, the alternative can be one or two-sided, and with enough evidence we reject H_0 in favor of H_A.

So how much evidence is enough evidence to reject H_0?

To answer this question, we must consider p-values. Strictly speaking, the p-value is defined as *the probability of observing a result at least as extreme as that actually observed, if the null hypothesis is indeed true*. Loosely speaking, it can be thought to quantify the likelihood that H_0 is true. Thus, small p-values suggest H_0 is false, and the smaller the p-value, the more evidence there is against H_0. Often we choose 0.05 as a cut-off for rejecting H_0. Specifically:

- If p-value ≤ 0.05, we reject H_0
- If p-value > 0.05, we fail to reject H_0

The calculation of the p-value is based on a comparison of a *test statistic* to a *null distribution*. If the test statistic seems extreme relative to the null distribution, the corresponding p-value will be small; in fact, the more extreme the test statistic, the smaller the p-value. The exact nature of the test statistic and the null distribution will depend on the type of hypothesis being tested, but the general framework persists globally.

Example Data Sets

For illustration, we will use data sets already stored in R that can be accessed at any time. Specifically we will be using the `iris` and `chickwts` data sets. For more information on these datasets, use the following commands:

```
? iris
? chickwts
```

We will also be using the `drpscores` data set, which we will manually load into R later.

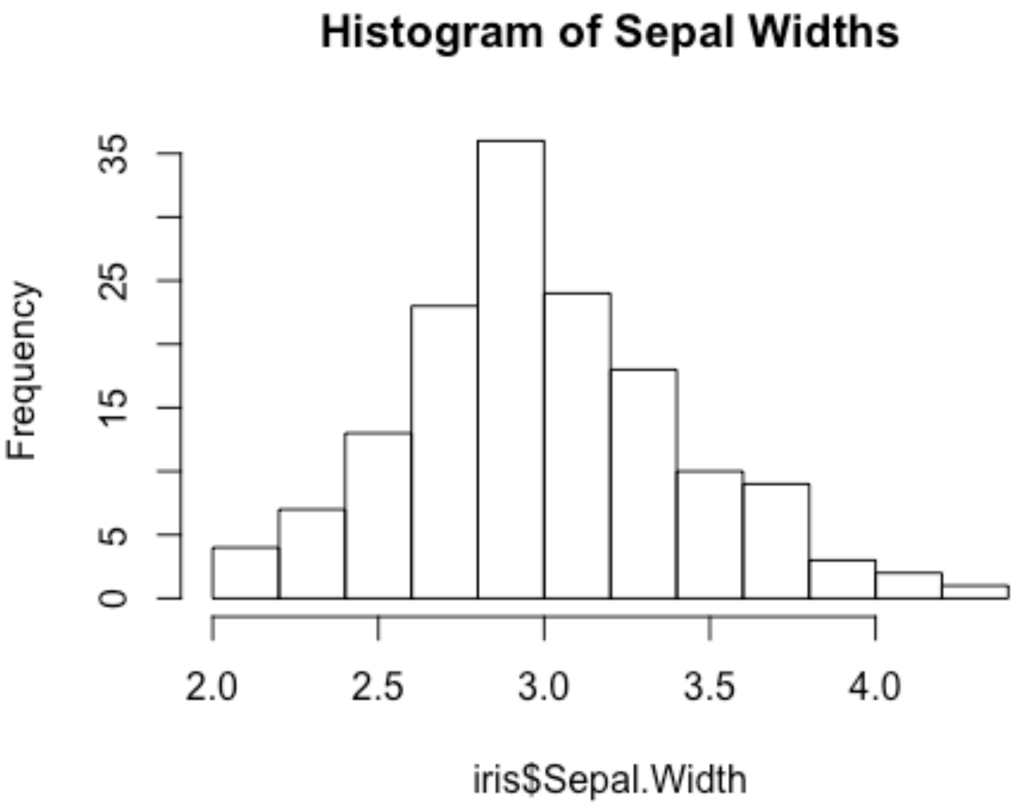
One-Sample t-Tests

Here we compare the population mean to some hypothesized value. For illustration we use the `iris` data. Let's visualize this data:

```
summary(iris)

##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  

hist(iris$Sepal.Width, main = "Histogram of Sepal Widths")
```



Commonly one-sample hypothesis tests are carried out using a **t-test**. Use the following command for information on the `t.test()` function:

```
? t.test
```

Suppose we it is assumed that the true population mean is 3, and wish to establish whether it is in fact different from 3. Formally, we state this hypothesis as

$$H_0: \mu = 3 \text{ vs. } H_A: \mu \neq 3$$

Let's test this hypothesis:

```
t.test(iris$Sepal.Width, alternative = "two.sided", mu = 3, conf.level
= 0.95)

##
##  One Sample t-test
##
## data:  iris$Sepal.Width
## t = 1.611, df = 149, p-value = 0.1093
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  2.987010 3.127656
## sample estimates:
## mean of x
##  3.057333
```

Because p-value > 0.05, we fail to reject that hypothesis, suggesting that the population mean could be 3. Suppose, instead, we wish to test whether the population mean is 4 or not. Formally, we state this hypothesis as

$$H_0: \mu = 4 \text{ vs. } H_A: \mu \neq 4$$

Let's test this hypothesis:

```
t.test(iris$Sepal.Width, alternative = "two.sided", mu = 4, conf.level
= 0.95)

##
##  One Sample t-test
##
## data:  iris$Sepal.Width
## t = -26.488, df = 149, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  2.987010 3.127656
## sample estimates:
## mean of x
##  3.057333
```

Now we have p-value < 0.05, and so we reject $H_0: \mu = 4$ in favor of $H_A: \mu \neq 4$.

Two-Sample t-Tests: Independent Samples

Here we compare the means of two independent populations. In terms of notation, a two-sided version of this hypothesis may be stated as:

$$H_0\colon \mu_1 = \mu_2 \text{ vs. } H_A\colon \mu_1 \neq \mu_2$$

To illustrate such a comparison we consider the `drpscores` data. First let us load this data into R and manipulate it in such a way that accomodates hypothesis testing.

```
#View the data
setwd("/Users/ntstevens/Documents/R Workshop")
data <- read.csv("drpscores.csv", header = T)
data
```

```
##    Treatment Response
## 1    Treated       24
## 2    Treated       43
## 3    Treated       58
## 4    Treated       71
## 5    Treated       43
## 6    Treated       49
## 7    Treated       61
## 8    Treated       44
## 9    Treated       67
## 10   Treated       49
## 11   Treated       53
## 12   Treated       56
## 13   Treated       59
## 14   Treated       52
## 15   Treated       62
## 16   Treated       54
## 17   Treated       57
## 18   Treated       33
## 19   Treated       46
## 20   Treated       43
## 21   Treated       57
## 22   Control       42
## 23   Control       43
## 24   Control       55
## 25   Control       26
## 26   Control       62
## 27   Control       37
## 28   Control       33
## 29   Control       41
## 30   Control       19
## 31   Control       54
## 32   Control       20
## 33   Control       85
## 34   Control       46
## 35   Control       10
## 36   Control       17
## 37   Control       60
## 38   Control       53
## 39   Control       42
## 40   Control       37
## 41   Control       42
## 42   Control       55
## 43   Control       28
## 44   Control       48
```

```
#Subset the data by treatment group
treated <- subset(data,data$Treatment=="Treated")
control <- subset(data,data$Treatment=="Control")
```

```
t.test(treated$Response,control$Response, alternative = "two.sided",0,
conf.level = 0.95)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  treated$Response and control$Response
## t = 2.3109, df = 37.855, p-value = 0.02638
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.23302 18.67588
## sample estimates:
## mean of x mean of y 
##  51.47619  41.52174
```

Since p-value < 0.05, we reject H_0 and conclude that the average reading score in the two treatment groups is indeed different. Intuition tells us that the mean DRP score in the treatment group should be higher than in the control group. As such, let us more appropriately define a one-sided alternative:

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 > \mu_2$$

The following code tests this hypothesis.

```
t.test(treated$Response,control$Response, alternative = "greater",0, co
nf.level = 0.95)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  treated$Response and control$Response
## t = 2.3109, df = 37.855, p-value = 0.01319
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.691293      Inf
## sample estimates:
## mean of x mean of y 
##  51.47619  41.52174
```

Since p-value < 0.05, we reject H_0 and conclude that the average reading score in the treatment group is indeed larger than in the control group, indicating efficacy of the directed reading condition.

Two-Sample t-Tests: Dependent Samples

Also known as **"paired" t-tests**, dependent samples t-tests compare the means of two dependent groups. Typically dependent groups arise when two measurements are taken on the same individual or individuals who are paired in some way (i.e., before/after tests, twin studies etc.). The statement of such a hypothesis is no different than in the independent case; what differs is the definition of the test statistic and the null distribution.

To illustrate this idea we move back to the `iris` data. Here we compare a plant's sepal width to it's sepal length in accordance with

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 \neq \mu_2$$

```
t.test(iris$Sepal.Length,iris$Sepal.Width, alternative = "two.sided", 0
, paired = T, conf.level = 0.95)

## 
##  Paired t-test
## 
## data:  iris$Sepal.Length and iris$Sepal.Width
## t = 34.815, df = 149, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.627874 2.944126
## sample estimates:
## mean of the differences 
##                   2.786
```

Let's contrast this output with that of an independent sample comparison (note that this is inapropriate in this setting).

```
t.test(iris$Sepal.Length,iris$Sepal.Width, alternative = "two.sided", 0
, paired = F, conf.level = 0.95)

## 
##  Welch Two Sample t-test
## 
## data:  iris$Sepal.Length and iris$Sepal.Width
## t = 36.463, df = 225.68, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.63544 2.93656
## sample estimates:
## mean of x mean of y 
##  5.843333  3.057333
```

In both cases p-value < 0.05, leading to a rejection of H_0. However, we see that the test statistic and the null distribution differ between the two cases.

As we can see, `t.test()` is very powerful. We use the same function for all sorts of tests; all that changes is the inputs we provide.

Tests of Normality

Often it will be of interest to determine whether your data is normally distributed. Several formal and informal methods may be used to test this. In general, the hypothesis may be stated as

$$H_0\text{: my data is normally distributed vs. } H_A\text{: my data is not normally distributed}$$

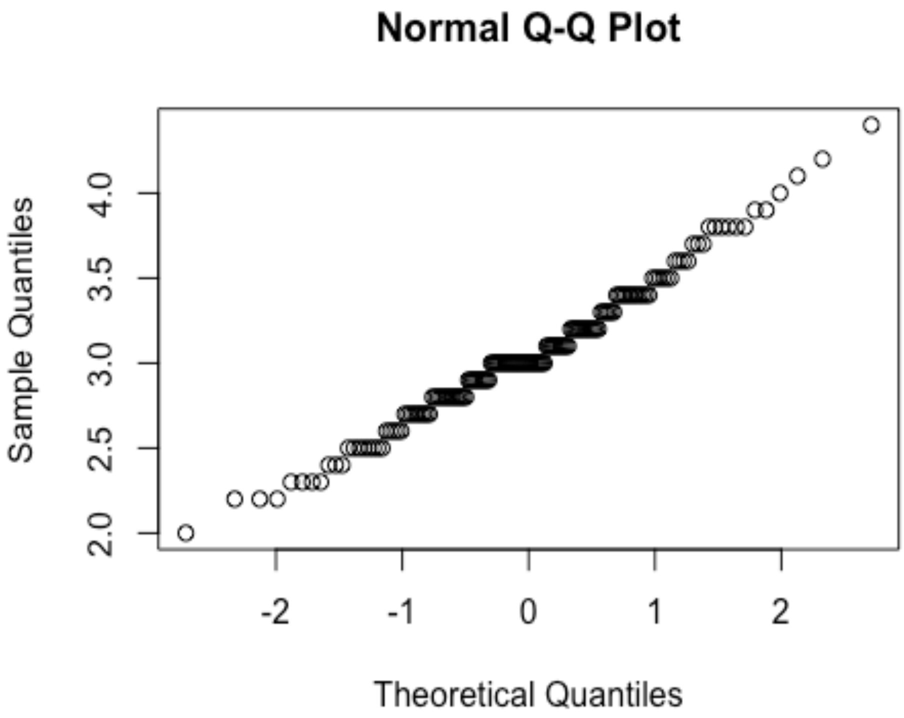
To illustrate these ideas, we return to the `iris` data. We begin by describing two informal methods of checking for normality. The first is to simply construct a histogram of the data and assess the extent to which it is "bell-shaped" and symmetric. Let's check this for the sepal width data:

```
hist(iris$Sepal.Width, main = "Histogram of Sepal Widths")
```

We see that this looks relatively bell-shaped, and close to symmetric.

Another method of graphically assessing normality is with a QQ-plot (QQ stands for "quantile-quantile"). With this plot we compare the sample quantiles to that of a normal distribution. If the sample data is indeed normally distributed, we would expect these quantiles to be linearly related, and so the QQ-plot should visually resemble a straight line. Let's check this for the sepal width data:

```
? qqnorm
qqnorm(iris$Sepal.Width)
```



Both of these methods seem to indicate that the sepal width data is roughly normally distributed.

Let us now formally check this with a **Shapiro-Wilk Test** or a **Kolmogorov-Smirnov Test**. We try both of these in turn:

```
? shapiro.test
shapiro.test(iris$Sepal.Width)

##
##  Shapiro-Wilk normality test
##
## data:  iris$Sepal.Width
## W = 0.98492, p-value = 0.1012
```

We see that the Shapiro-Wilk test is simple to call, and simple to interpret. We find that there is not enough evidence to reject normality (p-value > 0.05). Let us see if the Kolmogorov-Smirnov test agrees.

```
? ks.test
ks.test(iris$Sepal.Width, rnorm(1000, mean = mean(iris$Sepal.Width), sd
= sd(iris$Sepal.Width)))

## Warning in ks.test(iris$Sepal.Width, rnorm(1000, mean = mean(iris
## $Sepal.Width), : p-value will be approximate in the presence of ties

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  iris$Sepal.Width and rnorm(1000, mean = mean(iris$Sepal.Width
), sd = sd(iris$Sepal.Width))
## D = 0.091333, p-value = 0.2266
## alternative hypothesis: two-sided
```

This test also tells us that it is reasonable to believe that the sepal width data is normally distributed (we do not have sufficient evidence to reject H_0).

Comparing Two Distributions

On the face of it, the Shapiro-Wilk test seemed easier to use than the Kolmogorov-Smirnov test, when check for normality. While this is true for checking normality, the Kolmogorov-Smirnov test is actually more powerful in that it can be used to compare any two distributions. We may frame this hypothesis as follows:

$$H_0\text{: sample 1 and sample 2 have the same distribution}$$
$$\text{vs.}$$
$$H_A\text{: sample 1 and sample 2 do not have the same distribution}$$

As an illustration, perhaps we are interested in determing whether the distribution of sepal widths is the same as the sepal lengths. To investigate this, we use the following command:

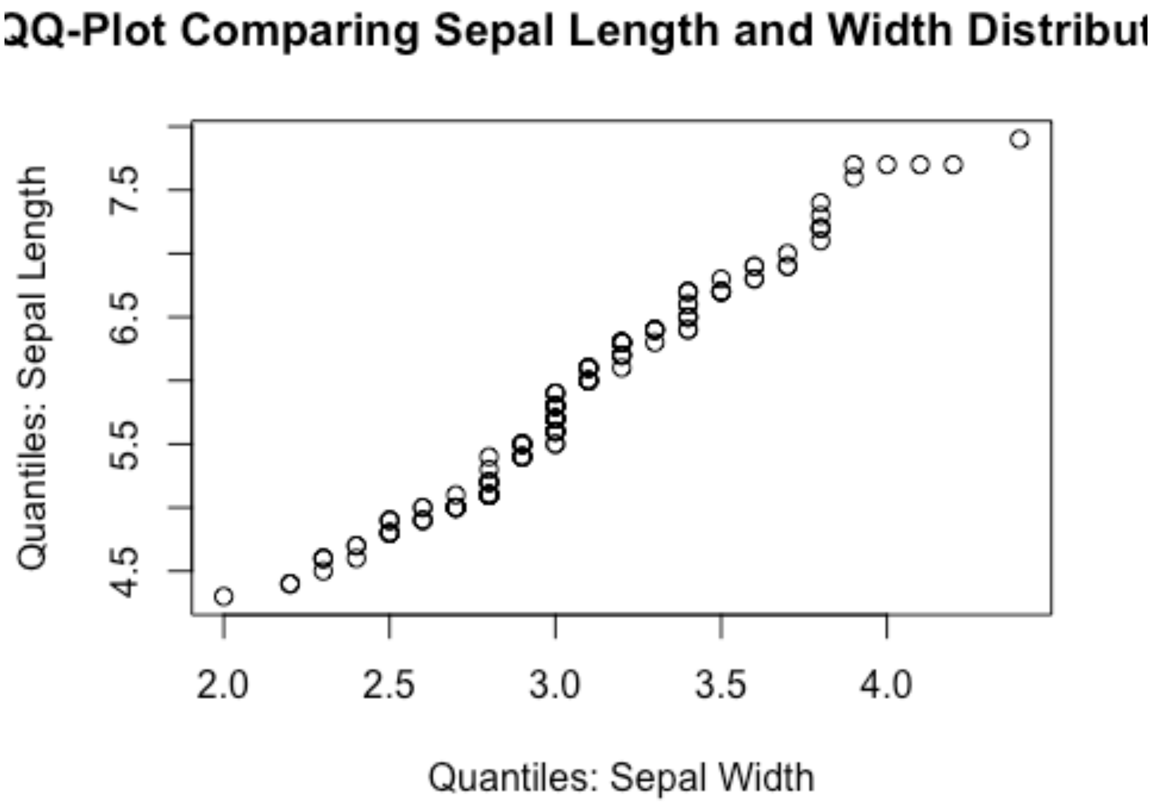
```
ks.test(iris$Sepal.Width, iris$Sepal.Length)

## Warning in ks.test(iris$Sepal.Width, iris$Sepal.Length): p-value wil
l be
## approximate in the presence of ties

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  iris$Sepal.Width and iris$Sepal.Length
## D = 0.99333, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

We can similarly use a QQ-plot to informally compare two distributions using the following command:

```
? qqplot
qqplot(iris$Sepal.Width,iris$Sepal.Length, main = "QQ-Plot Comparing Se
pal Length and Width Distributions", xlab = "Quantiles: Sepal Width", y
lab = "Quantiles: Sepal Length")
```



By both of these methods we conclude that the sepal widths and lengths do not follow the same distribution.

Multiple Group Comparisons

In some cases we may have several groups we wish to compare in some way. For example, we may want to compare the means across all groups:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$\text{vs.}$$
$$H_A: \text{at least one of the } \mu_i\text{'s is different}$$

Or we may wish to compare the variances across all groups:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$
$$\text{vs.}$$
$$H_A: \text{at least one of the } \sigma_i^2\text{'s is different}$$

We respectively achieve these goals with **ANOVA Tests** and **Bartlett's Tests**. We present each of these in turn. For both, let's consider the `chickwts` data set. To begin, let's do a quick summary of this data:

```
summary(chickwts)
```

```
##       weight             feed   
##  Min.   :108.0   casein   :12  
##  1st Qu.:204.5   horsebean:10  
##  Median :258.0   linseed  :12  
##  Mean   :261.3   meatmeal :11  
##  3rd Qu.:323.5   soybean  :14  
##  Max.   :423.0   sunflower:12
```

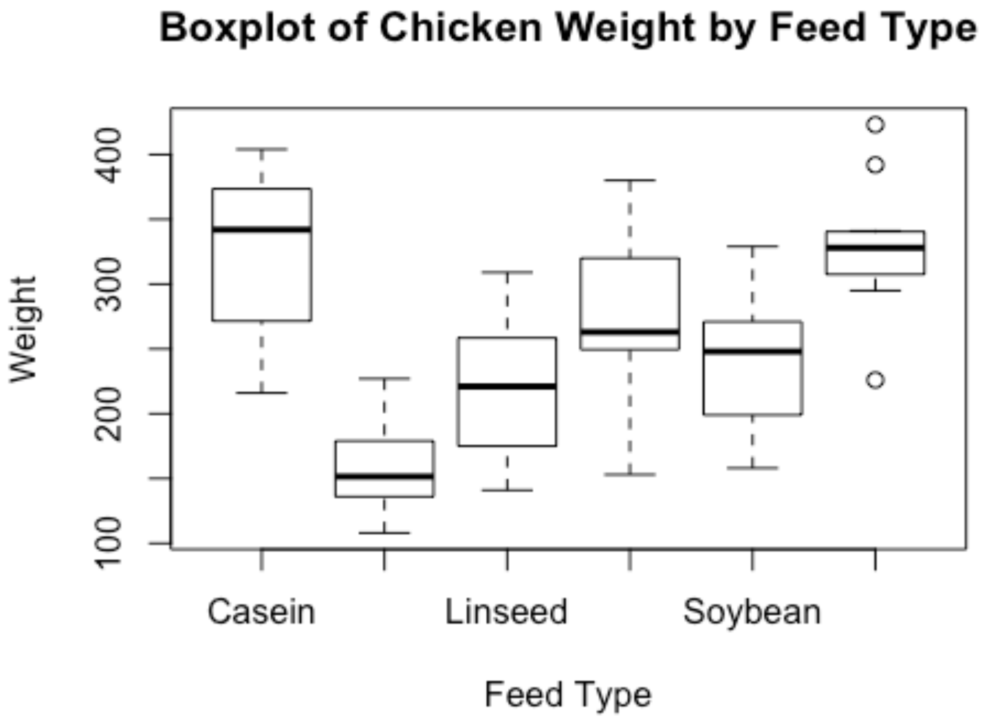
```
weight <- chickwts$weight
feed <- chickwts$feed
tapply(weight,feed,mean)
```

```
##    casein horsebean   linseed  meatmeal   soybean sunflower 
##  323.5833  160.2000  218.7500  276.9091  246.4286  328.9167
```

```
tapply(weight,feed,sd)
```

```
##    casein horsebean   linseed  meatmeal   soybean sunflower 
##  64.43384  38.62584  52.23570  64.90062  54.12907  48.83638
```

```
boxplot(weight[feed=="casein"],weight[feed=="horsebean"],weight[feed=="
linseed"],weight[feed=="meatmeal"],weight[feed=="soybean"],weight[feed=
="sunflower"], main = "Boxplot of Chicken Weight by Feed Type", xlab =
"Feed Type", ylab = "Weight", names = c("Casein","Horsebean","Linseed",
"Meatmeal","Soybean","Sunflower"))
```



Now let us formally compare the group means:

```
? anova
anova(lm(weight~feed))
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## feed       5 231129   46226  15.365 5.936e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now let us formally compare the group variances:

```
? bartlett.test
bartlett.test(weight,feed)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight and feed
## Bartlett's K-squared = 3.2597, df = 5, p-value = 0.66
```