Statistics Case Study: Cross Fell Grazing 2013

CK

19 Sept 2017

Time run: 2017-09-21 15:13:59

# Natural England Statistics Case Study

## Analysis of Variance in a grazing impacts study at Cross Fell SSSI

**Study:** *Survey of Grazing Impact on Cross Fell Montane Heath and Analysis of Change 2003-2013, Martin D., Natural England, 2014*

## Background

Cross Fell is part of Kirkland Fell, one of five contiguous upland commons in eastern Cumbria that came into Countryside Stewardship Scheme (CSS) Agreement Between 2000 and 2003. Under the terms of the agreement, stocking levels were reduced and shepherding introduced to further limit grazing pressure on sensitive habitats. As part of the project, the area of montane heath on the summit plateau of Cross Fell was surveyed in October 2003. Attributes of the vegetation condition and species composition were recorded. The survey was repeated in 2005, 2008, 2010 and 2013.

## Field Method

Surveyors measured vegetation height of a variety of taxa, including bilberry *Vaccinium myrtillus* on five sucessive surveys on montane grassland at Cross Fell in the North Pennines. The surveys took place in 2003, 2005, 2008, 2010 and 2013. Survey quadrats were randomly located each year. There were different numbers of quadrats in each survey year.

|  |  |
| --- | --- |
| year | n.total |
| 2003 | 42 |
| 2005 | 44 |
| 2008 | 20 |
| 2010 | 31 |
| 2013 | 56 |

Quadrats were placed at the random points that fell within montane heath on Cross Fell. The same set of random points was used in 2003 and 2005, but new sets generated in each of 2008, 2010 and 2013.

There is a record for each height measurement per year (although the quadrats are not labelled). Up to four bilberry height measurements were taken in each quadrat.

The number of measurements taken in each year is as follows:

|  |  |
| --- | --- |
| year | measurements |
| 2003 | 21 |
| 2005 | 53 |
| 2008 | 44 |
| 2010 | 40 |
| 2013 | 70 |

## Data analysis and results

One of the desired outcomes of the intervention was a significant increase in bilberry height, as an indicator of favourable condition. The bilberry height data was therefore analysed for differences in the mean height between years.

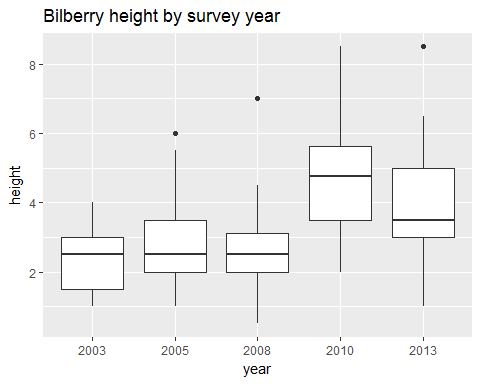
### Data exploration

**For each year** the following summary values of **vegetation height** were calculated:

Bilberry vegetation height summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| year | mean | median | sd | n |
| 2003 | 2.40 | 2.50 | 0.94 | 21 |
| 2005 | 2.88 | 2.50 | 1.15 | 53 |
| 2008 | 2.58 | 2.50 | 1.21 | 44 |
| 2010 | 4.66 | 4.75 | 1.73 | 40 |
| 2013 | 3.68 | 3.50 | 1.52 | 70 |

Plotted this looks as follows:



### Testing assumptions

The data was tested for normality using the Anderson Darling test. Only the measurements in 2005 were not normally distributed (A = 1.2507, p = 0.0027). The other years' measurements were normally distributed (2003: A = 0.4746, p = 0.216; 2008: A = 0.7769, p = 0.0403; 2010: A = 0.3875, p = 0.3717; 2013: A = 0.6164, p = 0.1046)

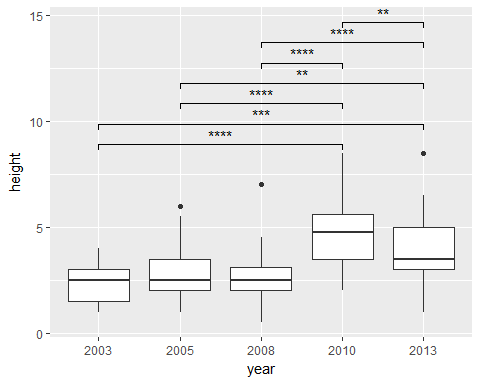
The data were tested for homoscedasticity (homogeneity of variances) using the Bartlett Test. The variances were not found to be homogeneous (Bartlett's K-squared = 14.7949104, p = 0.005146). The data was found not to meet the requirements of homoscedasticity required for ANOVA so non-parametric testing was required.

### Comparing differences in means

The mean difference vegetation height in each taxon was tested using the Kruskal Wallis rank sum test test. There was a significant difference between the means (Kruskal-Wallis chi-squared = 50.3392199, p = 3.067429210^{-10}).

Having found significant differences, post-hoc testing using pairwise was carried out to identify and characterise the change. There earliest significant increase in vegetation height occurs in 2010. There was a significant increase in mean height of 2.26cm between 2003 (mean = 2.4cm) and 2010 (mean = 4.66cm) (p = 1.8e-06), and in fact also between 2010 and all previous survey years (2005: height increase = 1.27, p = 0.00045; 2008: height increase = 1.79, p = 6.2e-07). Although the mean height of bilberry subsequently decreased by -0.98cm between 2010 (mean = 4.66cm) and 2013 (mean = 3.68cm) (p = 0.00584), over the entire survey period there is an overall increase of 0.8cm between 2003 (mean = 2.4cm) and 2013 (mean = 3.68cm) (p = 0.00191).

This is shown in the graph below:



## Understanding the statistics

### Data exploration

#### Mean

* What question does it answer: What is the average value of my data?
* How do I read it: It is the average of all the values.
* How does it work: It is the sum of all the values, divided by the number of values.

#### Median

* What question does it answer: What is the middle value of my data?
* How do I read it: It's just the middle value.
* How does it work: All the values are lined up in order and the middle value is found.

#### standard deviation, sd, sigma,

* What question does it answer: How spread out are the values?
* How do I read it: The greater the standard deviation, the wider the values are spread around the mean.
* How does it work: The standard deviation is the square root of the variance. The variance is the average of the squared differences from the mean.

#### n

* What question does it answer: What is the size of my sample?
* How do I read it / how does it work: It is simply the number of observations.
* Tips: If n is very small (e.g. below 5), then most statistics are unlikely to be accurate.

#### boxplot (box and whisker plot)

* What does it show: It shows the distribution of the data.
* How do I read it: First understand how it is drawn (below). Then you can use it to:
  + get an impression of how evenly the data is distributed (is the boxplot symmetrical),
  + whether the data is skewed (is the median closer to one end of the box than the other),
  + whether there are outliers,
  + whether the variances of two samples are similar (similar sized boxes).
* How does it work: There are many ways of drawing these but the most common is as follows:
  + the middle line is the *median*,
  + the box goes from the first to the third quartile (i.e. it encloses the middle half of the data), known as the *inter-quartile range*,
  + the lines extend to the highest and lowest values,
  + points show *outliers*, which are usually defined as values which are further than 1.5 times the inter-quartile range from the first or third quartile.

### Testing assumptions

#### p value

* What question does it answer: Is the statistic I have calculated *statistically significant*?
* How do I read it: In most cases, if p is smaller than 0.05 then you have found a statistically significant result. It means that you have a smaller than 1 in 20 chance of your *null hypothesis* being true purely by chance. In most cases in ecology the null hypothesis (*H0*) is 'no difference'. However, it depends on the test you are carrying out and you should always check. It also depends on the *significance level* (alpha or ) that has been decided beforehand. In ecology is usually 0.05.
* How does it work: It is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true – the definition of ‘extreme’ depends on how the hypothesis is being tested.
* Tips: Always carefully check what the *null hypothesis* is.

#### Anderson Darling test for test for the composite hypothesis of normality

* What question does it answer: Is the data normally distributed?
* How do I read it: If p is greater than 0.05 then the distribution is normal.
* How does it work: The test compares the data to a normal distribution with the same mean and standard deviation and calculates the probability of getting a more extreme result (i.e. more different to normal) by chance.

#### Bartlett test for homogeneity of variances

* What question does it answer: How different is the spread (variance) of each sample? Are they homogeneous?
* How do I read it: If the p-value is smaller than 0.05 then the variances are very different and not homogeneous.
* How does it work: The null hypothesis is that the variance is the same for all samples. The test compares the variances to each other and calculates the probability of getting a more extreme result (i.e. more different from each other) by chance.

### Comparing means

#### Kruskal-Wallis

* What question does it answer: How different are the means of our samples from each other?
* How do I read it: If the p-value is smaller than 0.05 then at least two of the samples are significantly different from each other. It will not tell you which two samples.
* How does it work: The null hypothesis is that the samples are drawn from the same population, i.e. the true mean is the same for each survey. The test lines up all the values in order and assigns ranks to them. It then calculates the probability of getting a more extreme result (i.e. more different from each other) by chance.
* Tip: You can use this test if your data is not normally distributed or if the variances are not homogeneous.

#### pairwise Wilcoxson rank sum tests (also known as Mann-Whitney U test)

* What question does it answer: Which pairs of samples are significantly different from each other.
* How do I read it: The test will return a p-value for combination of two samples. If the p-value is smaller than 0.05 for a particular pairing then the two samples are significantly different from each other.
* How does it work: The null hypothesis is that the samples are drawn from the same population, i.e. the true mean is the same for each survey. The test lines up all the values in order and assigns ranks to them. It then calculates the probability of getting a more extreme result (i.e. more different from each other) by chance.
* Tip: You cannot use this test unless you have already found that there is a significant difference between at least two samples. If there is not, the test may suggest that there are significant differences, but this is likely to be a false-positive.

## How to do it (in R)

### Packages

# Imports the required packages into R  
library(tidyverse) # Loads the following core packages:   
 # ggplot2, for data visualisation  
 # dplyr, for data manipulation  
 # tidyr, for data tidying  
 # readr, for data import  
 # purrr, for functional programming  
 # tibble, for tibbles, a simpler type of data frame  
library(knitr) # turns rmarkdown document to html, pdf or word  
library(ggpubr) # helps make publication-ready graphs  
library(broom) # cleans up code output from common functions  
library(nortest) # functions for testing normality  
library(lattice) # for faceting graphs

### Load data

# reads an excel file and writes it to a new object: ht.data  
ht.data <- read.csv("ht.data.csv", header = TRUE)

### Inspect data

# look at the data table  
View(ht.data)

# inspect data classes  
str(ht.data) # shows structure of data

## 'data.frame': 228 obs. of 2 variables:  
## $ year : int 2003 2003 2003 2003 2003 2003 2003 2003 2003 2003 ...  
## $ height: num 2.5 4 4 2 2.5 3.5 3 2 2.5 1 ...

Year is an integer; height is a numeric value. That's fine, but for now, we'd like to treat 'year' as a categorical variable to make it easier to group our data.

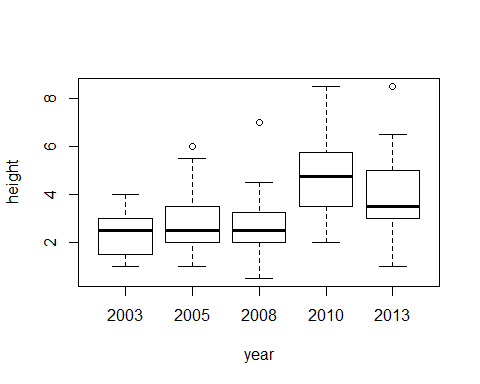
# turn 'year' into a factor  
ht.data$year <- as.factor(ht.data$year)

### Explore data

#summarise the data  
summary(ht.data)

## year height   
## 2003:21 Min. :0.500   
## 2005:53 1st Qu.:2.000   
## 2008:44 Median :3.000   
## 2010:40 Mean :3.336   
## 2013:70 3rd Qu.:4.500   
## Max. :8.500

#graph the data  
plot(ht.data) #R selects plot type (usually chooses well)



# Calculate means using package `dplyr`  
ht.means <- # create a new object  
 ht.data %>% # take the height data  
 group\_by(year) %>% # group it by year  
 summarise( # calculate summary statistics for each group  
 mean = mean(height), # its mean  
 median = median(height), # its median  
 sd = sd(height), # the standard deviation  
 n = length(height) # the number of observations  
 )   
  
ht.means #print the result

## # A tibble: 5 x 5  
## year mean median sd n  
## <fctr> <dbl> <dbl> <dbl> <int>  
## 1 2003 2.404762 2.50 0.9436505 21  
## 2 2005 2.877358 2.50 1.1515277 53  
## 3 2008 2.579545 2.50 1.2149418 44  
## 4 2010 4.662500 4.75 1.7297825 40  
## 5 2013 3.678571 3.50 1.5228589 70

### Check assumptions for statistical tests

Here we want to know whether the data meets the assumptions required for 'parametric' tests which are generally more likely to find a significant difference if there is one.

The most powerful method for testing whether there is a significant change in the height of each taxon between survey years is Analysis of Variance (ANOVA). For us to be able to use this our data must comply with teh following assumptions:

* **Independence of cases**: this is met because each quadrat was selected at random each year.
* **Equality of variances (homoscedasticity)**: we will test for this.
* **Normality**: we will test for this if the variances are equal.

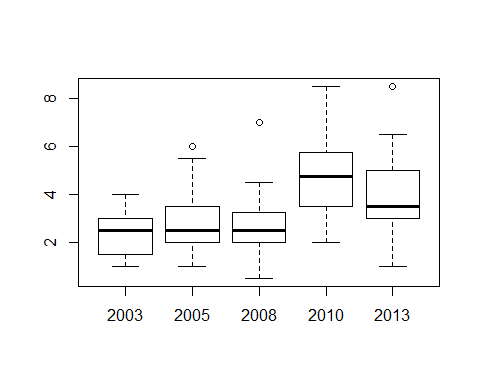
#### Plot data

The first step is to 'eyeball' the data. Key points to look out for

1. Outliers: look for individual points on the box plots. Lots of outliers means the data is unlikely to meet homoscedasticity or normality assumptions.
2. Skewness: High skewness means data is unlikely to be normally distributed. Density plot will be skewed and boxplots will be asymmetrical.
3. Unequal variance: Unequal box sizes will indicate that variances are not equal. If there is little data, then this may not be reliable, as variances can appear unequal when they aren't.

##### Boxplots

#Make a series of boxplot grouped by a factor  
boxplot(height ~ year, data = ht.data)

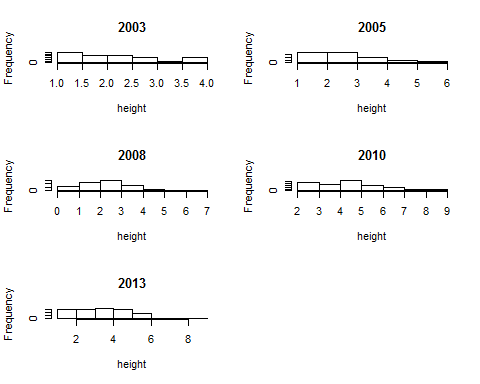


The variances (box sizes) appear to be different in the later years. The median line in some of the boxes is very off-centre, suggesting a skewed distribution. There are few outliers. The 2010 results are considerably higher than previous results; its box hardly overlaps with the previous years'.

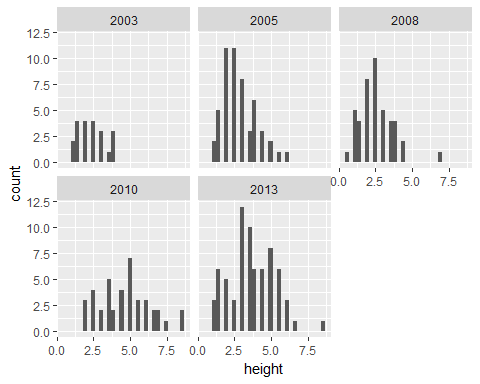
##### Histograms

This plots the number of observations for a particular value.

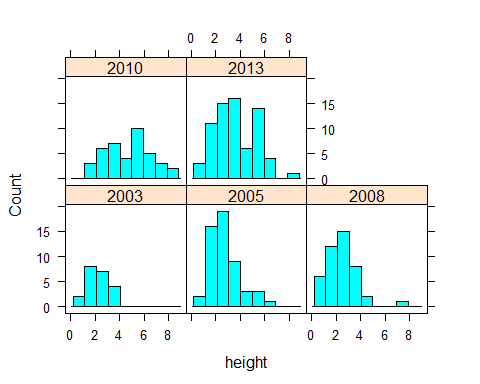
# histograms of height by year in base R  
par(mfrow=c(3,2))  
hist(ht.data$height[which(ht.data$year == "2003")], main = "2003", xlab = "height")  
hist(ht.data$height[which(ht.data$year == "2005")], main = "2005", xlab = "height")  
hist(ht.data$height[which(ht.data$year == "2008")], main = "2008", xlab = "height")  
hist(ht.data$height[which(ht.data$year == "2010")], main = "2010", xlab = "height")  
hist(ht.data$height[which(ht.data$year == "2013")], main = "2013", xlab = "height")  
par(mfrow=c(1,1))



# histograms of height by year - ggplot  
ggplot(ht.data, aes(x= height)) +  
 geom\_histogram() +  
 facet\_wrap(~year)



# histograms of height by year - lattice  
histogram(~height|year, data = ht.data, type = "count")

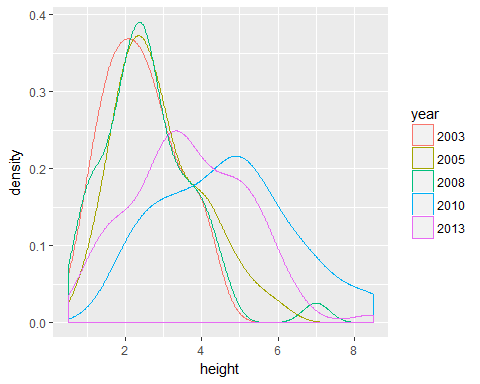


Here we can see more clearly how the variances differ (some of the histograms are bunched, some are spread out).

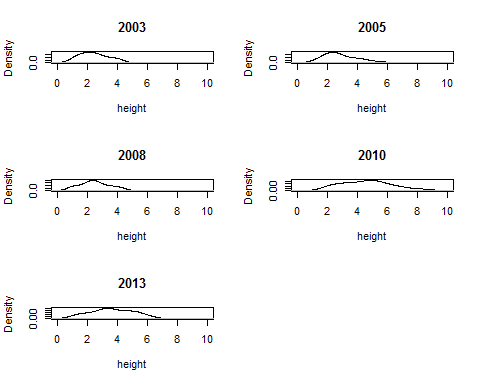
##### Density plots

Kernel density estimation (KDE) is a way to estimate the probability density function of a random variable. Each position on the density plot gives you estimated probabilty for a given value. The plot as a whole can be used to judge the distribution of the data.

#density plot of height by year - ggplot  
ggplot(ht.data, aes(x= height)) +  
 geom\_density(aes(colour = year))



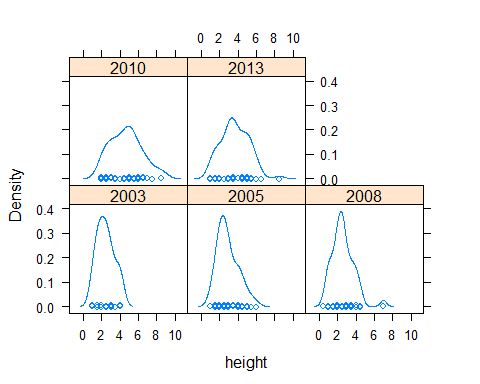
#density histogram and smoother of height by year - base R  
par(mfrow=c(3,2))  
plot(density(ht.data$height[which(ht.data$year == "2003")]),  
 main = "2003", xlab = "height", xlim = c(0,10))  
plot(density(ht.data$height[which(ht.data$year == "2005")]),  
 main = "2005", xlab = "height", xlim = c(0,10))  
plot(density(ht.data$height[which(ht.data$year == "2008")]),   
 main = "2008", xlab = "height", xlim = c(0,10))  
plot(density(ht.data$height[which(ht.data$year == "2010")]),   
 main = "2010", xlab = "height", xlim = c(0,10))  
plot(density(ht.data$height[which(ht.data$year == "2013")]),   
 main = "2013", xlab = "height", xlim = c(0,10))  
par(mfrow=c(1,1))



#density histogram and smoother of height by year - ggplot  
ggplot(ht.data, aes(x= height)) +  
 geom\_histogram(aes(y=..density..), fill = "grey") +   
 geom\_density(colour = "red") +  
 facet\_wrap(~year)



#density plots and 'rug marks' of height by year - lattice  
densityplot(~height|year, data = ht.data)



Here we can see that the distributions do not appear to be normal. The graphs lack long 'tails' which we would expect from the classic 'bell-curve' shape of a normal distribution.

#### Test for normality

We now carry out a formal test whether the data are normally distributed. First we will check the distribution of the combined dataset:

# Anderson Darling test for the composite hypothesis of normality (package: nortest)  
ad.test(x = ht.data$height)

##   
## Anderson-Darling normality test  
##   
## data: ht.data$height  
## A = 3.2442, p-value = 3.776e-08

#if p > 0.05 then the data is normally distributed

The p value is much lower than 0.05. We therefore have to reject the null hypothesis that the data is normally distributed.

Now we will test each survey year in turn:

#run Anderson Darling test for normality on each year in turn - base R  
norm.2003 <- ad.test(ht.data$height[which(ht.data$year == "2003")])  
norm.2003

##   
## Anderson-Darling normality test  
##   
## data: ht.data$height[which(ht.data$year == "2003")]  
## A = 0.47457, p-value = 0.216

norm.2005 <- ad.test(ht.data$height[which(ht.data$year == "2005")])  
norm.2005

##   
## Anderson-Darling normality test  
##   
## data: ht.data$height[which(ht.data$year == "2005")]  
## A = 1.2507, p-value = 0.002676

norm.2008 <- ad.test(ht.data$height[which(ht.data$year == "2008")])  
norm.2008

##   
## Anderson-Darling normality test  
##   
## data: ht.data$height[which(ht.data$year == "2008")]  
## A = 0.77685, p-value = 0.04034

norm.2010 <- ad.test(ht.data$height[which(ht.data$year == "2010")])  
norm.2010

##   
## Anderson-Darling normality test  
##   
## data: ht.data$height[which(ht.data$year == "2010")]  
## A = 0.38752, p-value = 0.3717

norm.2013 <- ad.test(ht.data$height[which(ht.data$year == "2013")])  
norm.2013

##   
## Anderson-Darling normality test  
##   
## data: ht.data$height[which(ht.data$year == "2013")]  
## A = 0.61644, p-value = 0.1046

# put results in a dataframe - base R  
norm <- data.frame(year = c(2003, 2005, 2008, 2010, 2013),   
 statistic = c(norm.2003$statistic,   
 norm.2005$statistic,   
 norm.2008$statistic,   
 norm.2010$statistic,   
 norm.2013$statistic),   
 p.value = c(norm.2003$p.value,  
 norm.2005$p.value,   
 norm.2008$p.value,   
 norm.2010$p.value,   
 norm.2013$p.value))  
   
#run Anderson Darling test for normality on each year in turn - dplyr and broom  
norm <- # create a new object to hold our results  
 ht.data %>% # start with the height data  
 group\_by(year) %>% # group it by year  
 do( # do the following functions on each group element  
 tidy( # extracts test results into a table  
 ad.test(.$height) #Anderson Darling test  
 )  
 )   
norm

## # A tibble: 5 x 4  
## # Groups: year [5]  
## year statistic p.value method  
## <fctr> <dbl> <dbl> <fctr>  
## 1 2003 0.4745657 0.215987885 Anderson-Darling normality test  
## 2 2005 1.2506867 0.002676464 Anderson-Darling normality test  
## 3 2008 0.7768502 0.040338390 Anderson-Darling normality test  
## 4 2010 0.3875231 0.371720318 Anderson-Darling normality test  
## 5 2013 0.6164418 0.104594747 Anderson-Darling normality test

Only the p value for 2005 is below 0.05, which means it has a normal distribution. None of the other samples have a normal distribution.

##### Try exponential transform

#run Anderson Darling test for normality on each year in turn - dplyr and broom  
norm <- # create a new object to hold our results  
 ht.data %>% # start with the height data  
 group\_by(year) %>% # group it by year  
 do( # do the following functions on each group element  
 tidy( # extracts test results into a table  
 ad.test(exp(.$height)) #Anderson Darling test on log transformed data  
 )  
 )   
norm

## # A tibble: 5 x 4  
## # Groups: year [5]  
## year statistic p.value method  
## <fctr> <dbl> <dbl> <fctr>  
## 1 2003 2.192589 8.825722e-06 Anderson-Darling normality test  
## 2 2005 8.459569 7.399365e-21 Anderson-Darling normality test  
## 3 2008 12.808119 3.700000e-24 Anderson-Darling normality test  
## 4 2010 8.478665 5.255041e-21 Anderson-Darling normality test  
## 5 2013 18.273584 3.700000e-24 Anderson-Darling normality test

hmm, its all normal now??? Need to ask SH.

#### Test for homogeneity of variances

Testing whether the variances are the same:

Bartlett Test of Homogeneity of Variances: **if the p-value is smaller than 0.05 then data not suitable for ANOVA**

#run Bartlett test for homogeneity of variances   
bart <- bartlett.test(height ~ year, data = ht.data)  
bart

##   
## Bartlett test of homogeneity of variances  
##   
## data: height by year  
## Bartlett's K-squared = 14.795, df = 4, p-value = 0.005146

Here the null hypothesis is that at least two variances are not equal. The p value is well below 0.05 which means that we reject that hypothesis.

Since the data is not normally distributed and the variances are not equal, we cannot use a parametric test (e.g. ANOVA) to compare the means.

### Compare means

#### Kruskal-Wallis rank sum test

This test estimates the probability that all the samples were drawn from the same population, i.e. that the true mean is the same for each survey.

#run Kruskal Wallis test  
krusk <- kruskal.test(height ~ year, data = ht.data)  
krusk

##   
## Kruskal-Wallis rank sum test  
##   
## data: height by year  
## Kruskal-Wallis chi-squared = 50.339, df = 4, p-value = 3.067e-10

The null hypothesis is that the samples are drawn from the same population, so if p is smaller than 0.05 we must reject that hypothesis. Here, p is much smaller than 0.05 so we can conclude that there is a significant difference between at lest two of the mean vegetation heights.

#### Pairwise Wilcoxson / Mann-Whitney U tests

In order to determine between which survey years there were significant differences we can carry out **pairwise** tests, i.e. test the difference between each pair of samples. This was not possible earlier because there is a high likelihood of false positives.

pairwise.wilcox.test(x = ht.data$height, g = ht.data$year, paired = FALSE)

##   
## Pairwise comparisons using Wilcoxon rank sum test   
##   
## data: ht.data$height and ht.data$year   
##   
## 2003 2005 2008 2010   
## 2005 0.38333 - - -   
## 2008 0.66002 0.43356 - -   
## 2010 1.5e-05 5.6e-06 6.8e-07 -   
## 2013 0.00269 0.00956 0.00052 0.02334  
##   
## P value adjustment method: holm

If the p value is below 0.05 then the differences in means are statistically significant. Here we have quite a few significant differences between 2010 and all previous years, and 2013 and all previous years.

We can graph this using the package ggpubr:

#run test again using function ggpubr::compare\_means() to prepare brackets for graphic output.   
wilc.pairw <- compare\_means(formula = height ~ year,   
 data = ht.data,   
 method = "wilcox.test", paired = FALSE)  
wilc.pairw <- filter(wilc.pairw, p < 0.05) #select only significant results.  
  
#generate list of brackets  
my\_comparisons <- list()   
for (i in 1:nrow(wilc.pairw)) {  
 my\_comparisons[[i]] <- c(wilc.pairw$group1[i], wilc.pairw$group2[i])  
 }  
  
#prepare plot  
p <- ggplot(ht.data, aes(x = year, y = height)) +  
 geom\_boxplot()  
  
# Add pairwise comparisons p-value to plot  
p + stat\_compare\_means(comparisons = my\_comparisons,   
 label = "p.signif",   
 hide.ns = FALSE)

