ST1050

Class 6

- More on Histograms and Boxplots
- Testing for differences in means

We continue with the Wolf dataset:

Variable name	Definition of the variable
Individual	= the ID of each individual (1-178)
Sex	= the sex of each individual (M=male, F=female)
Population	= the population that each individual belongs to (1=boreal forest, lightly hunted, 2=tundra-taiga, heavily hunted, 3=boreal forest, heavily hunted).
Colour	= coat colour of each individual (D=dark, W=light, blank=missing data)
Cpgmg	= concentration of cortisol in a hair sample [units=pg/mg of hair]
Tpgmg	= concentration of testosterone in a hair sample, males only [units=pg/mg of hair]
Ppgmg	= concentration of progesterone in a hair sample, females only [units=pg/mg of hair]

Recall:
wolf = read.csv('~/Desktop/wolf_hormone_data_for_dryad.csv')

 # Subset the wolf data frame and remove unwanted levels- we are not including the wolves that were culled as part of a control program.

wolf.sub = **subset**(wolf, Population**!=**3)

 # Make a 'Hunting' variable, which is a factor
wolf.sub**$**Hunting = 'Heavy' # setting up a vector of the right size quickly
wolf.sub**$**Hunting[wolf.sub**$**Population**==**1] = 'Light'

wolf.sub**$**Hunting = **as.factor**(wolf.sub**$**Hunting)

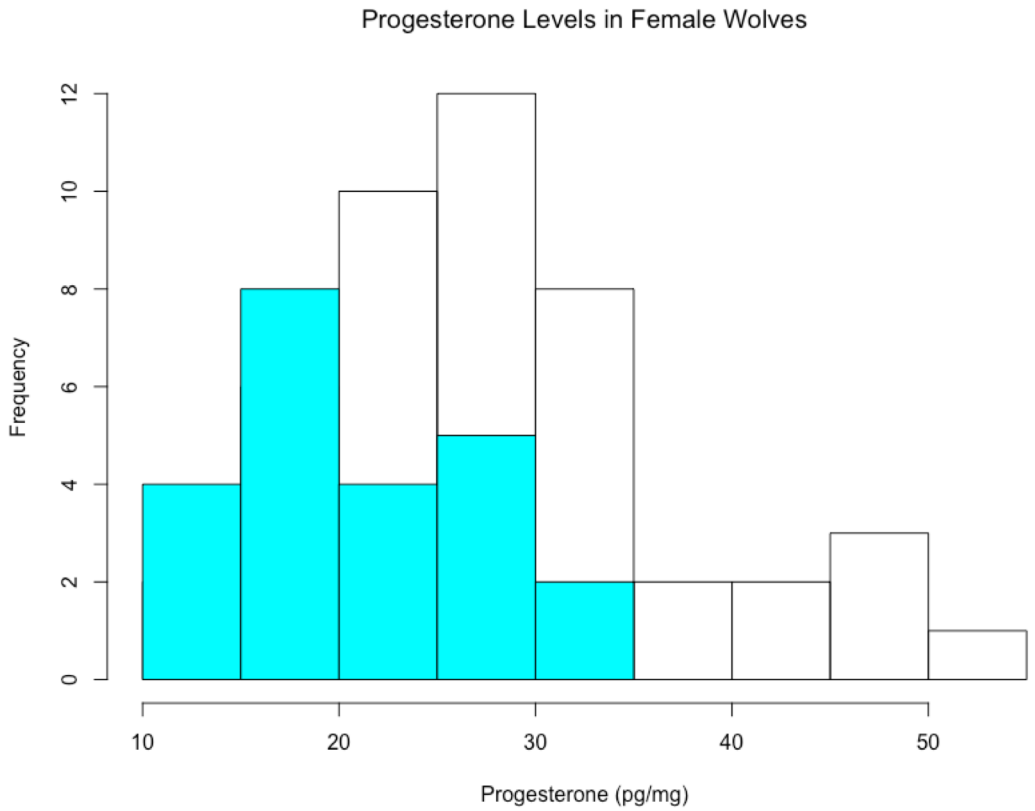
We also set up the following variables for simplifying commands:

Population = wolf.sub$Population
Sex = wolf.sub$Sex
Cpgmg = wolf.sub$Cpgmg
Tpgmg = wolf.sub$Tpgmg
Ppgmg = wolf.sub$Ppgmg #error last time on this- was set as Tpgmg
Hunting = wolf.sub**$**Hunting #same as population but a factor

Consider the Progesterone difference between Females that are Heavily Hunted and those that are Lightly hunted:

```
> Light_F=subset(wolf.sub,Hunting=='Light' & Sex=='F')
> Heavy_F=subset(wolf.sub,Hunting=='Heavy' & Sex=='F')

> hist(Heavy_F$Ppgmg,main='Progesterone Levels in Female Wolves',xlab='Cortisol (pg/mg)')
> hist(Light_F$Ppgmg,add=T,col=5).  Add=T allows us to put these together in one histogram.
```



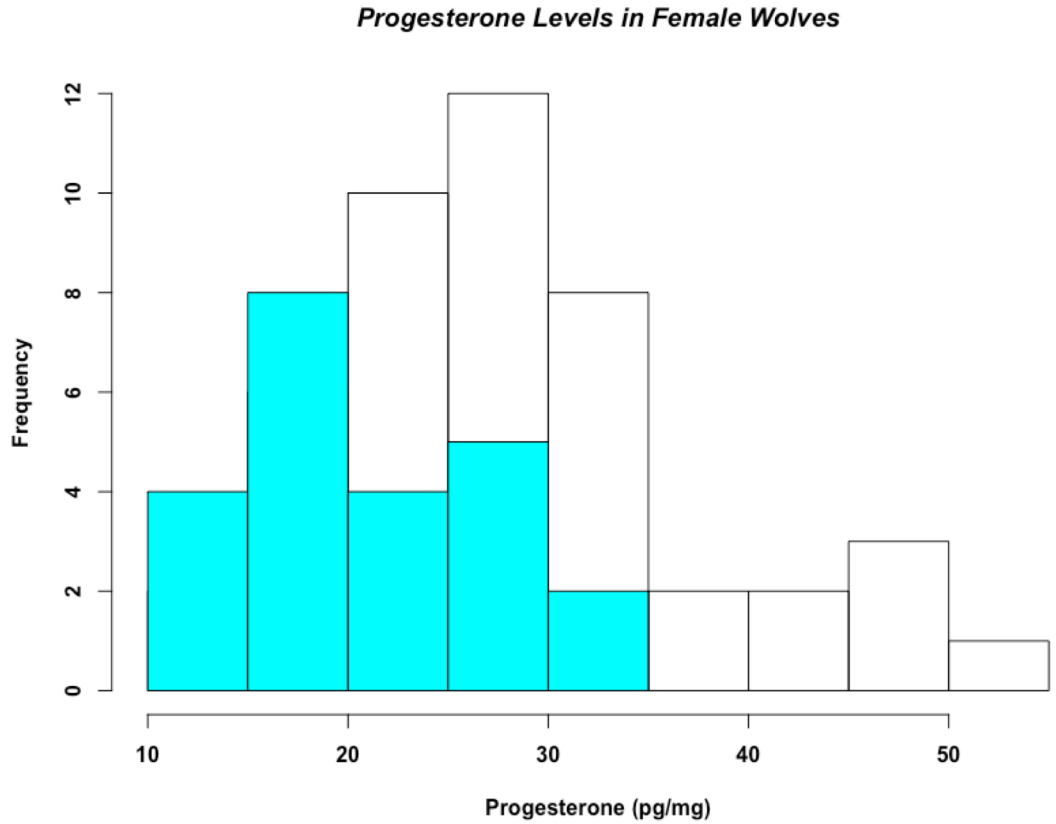
Progesterone level in Female Wolves Grouped according to their Level of Hunting (cont'd)

It is important when presenting data to get a plot as clear as possible- to do this we will add a few changes to the basic histogram function.

Having all text in bold would be much easier to read- we can do this using 'font' arguments of the 'par' function. The options for these can be found under help for 'par'.
The fonts settings are:
1=Plain; 2=Bold; 3=Italic; 4=Italic and Bold

We try:
> par(font.lab=2,font.main=4,font.axis=2)

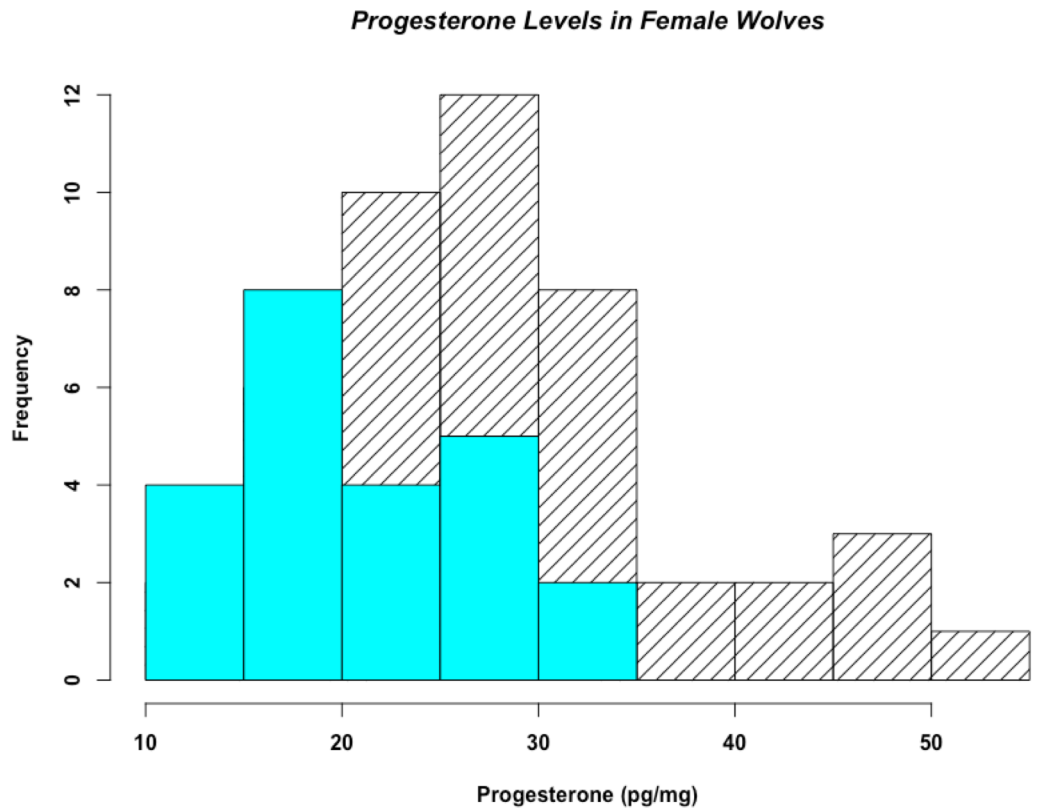
Using the same histogram commands we now get this:



Progesterone level in Female Wolves Grouped according to their Level of Hunting (cont'd)

The Heavy group are a little difficult to see in that histogram, and we want to add 'hatching' (diagonal lines) to improve visibility- this is done using the 'density' argument of hist:

> hist(Heavy_F$Ppgmg,main='Progesterone Levels in Female Wolves',xlab='Progesterone (pg/mg)',**density=10**)
> hist(Light_F$Ppgmg,add=T,col=5)
One concern- what is the distribution for the Heavy group at Progesterone under20? It is hidden here.



Progesterone level in Female Wolves Grouped according to their Level of Hunting (cont'd)

To get around the hidden data we reverse the colour scheme so the solid colour is put on the histogram first. (There are other ways to deal with this problem- such as making the colours more transparent.) We also add a legend.

```
> hist(Heavy_F$Ppgmg,main='Progesterone Levels in Female Wolves',xlab='Progesterone (pg/mg)',col=5)
> hist(Light_F$Ppgmg,,density=10,add=T)
```

```
> legend("topright", c("Heavy", "Light"), bty = "n", angle = c(0, 45), density = c(NA, 30), fill=c("cyan", "black"))
```

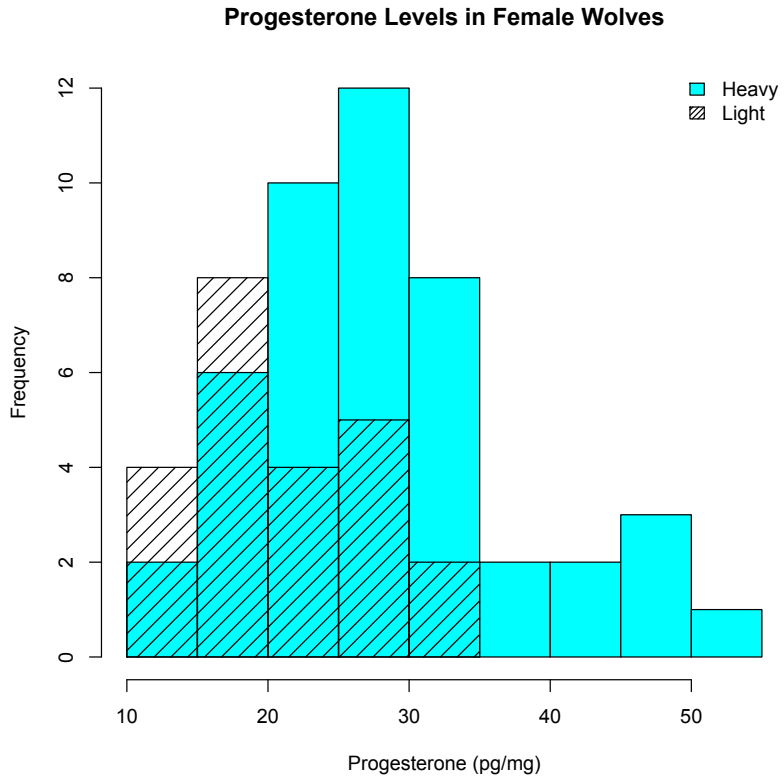
position: can be specified as a vector or a keyword for position.

bty: do you want a box around the legend or not (Y/N).

Angle: the angle for lines that will be drawn- this, combined with density, is to make the colour square looks like the diagonal lines in the boxplot.

density: whether to have a solid box or lines (only needed if you need to have lines instead of solid boxes).

fill: fill for the colour squares. Without fill or density there would be no colour squares.



The histogram suggests that the Heavily Hunted Females might have a higher progesterone level than the Lightly hunted ones.

We'll examine that question.
The summary() function is a good place to start:

```
> summary(Light_F$Ppgmg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  13.19   16.73   19.91     21.36   25.55     34.82      1

> summary(Heavy_F$Ppgmg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  12.76   21.55   27.02     28.15   31.89     53.28      2
```

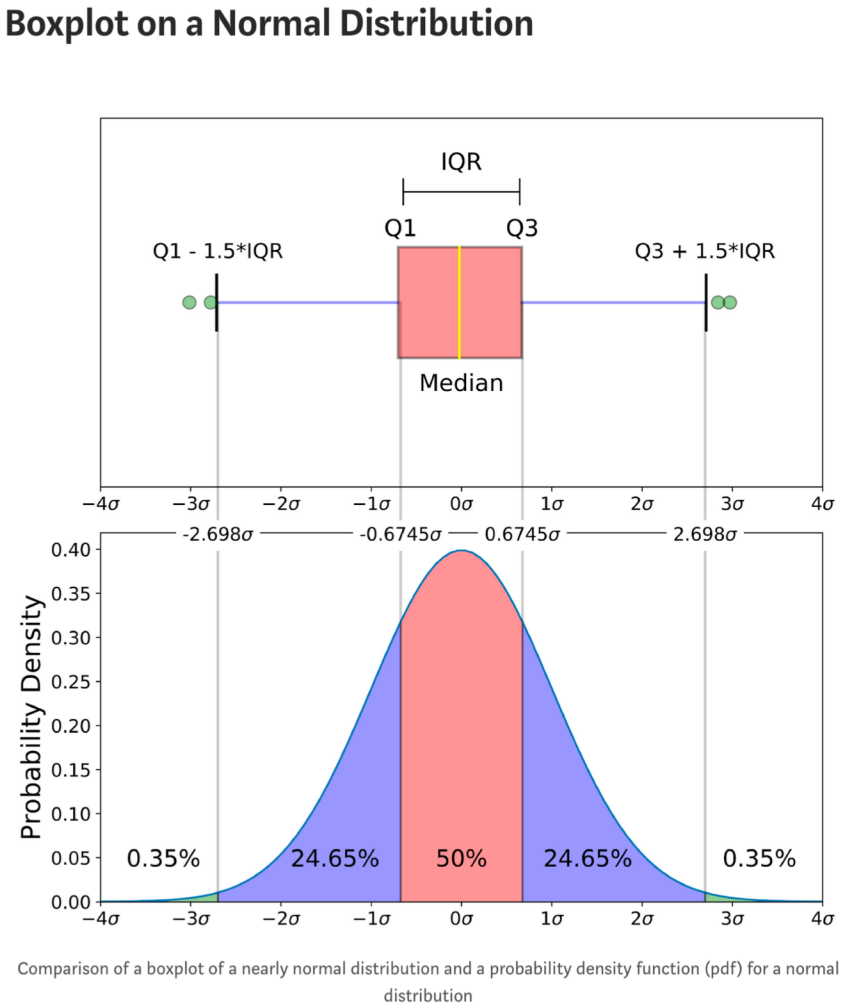
The mean of progesterone for the lightly hunted group is 21.36, while the mean is 28.15 for the heavily hunted group.
That seems like a large difference.

We look at the data using 'Box Plots' which were discussed last week.

Recall:
A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles.
The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally.
Box Plots have the advantage of taking up less space than many other plots, which is useful when comparing distributions between many groups or datasets.

Here are the types of observations one can make from viewing a Box Plot:

- What the key values are, i.e.: the average, median 25th percentile etc.
- If there are any outliers and what their values are.
- Is the data symmetrical.
- How tightly is the data grouped.
- If the data is skewed and if so, in what direction.



Comparison of a boxplot of a nearly normal distribution and a probability density function (pdf) for a normal distribution

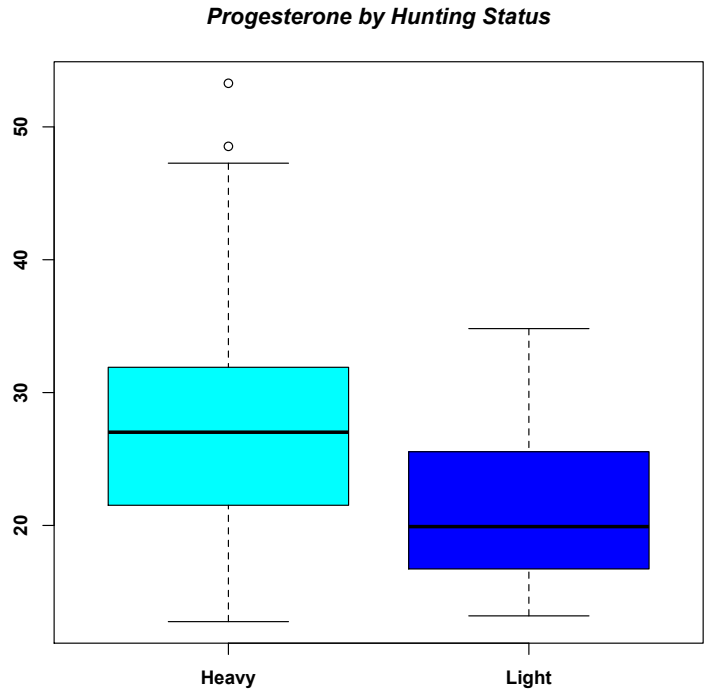
https://datavizcatalogue.com/methods/box_plot.html

https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Do Heavily Hunted and Lightly hunted female wolves have different Progesterone levels?

Using boxplots on our question:

```
> par(font.lab=2,font.main=4,font.axis=2)
> boxplot(Ppgmg ~ Hunting,main='Progesterone by Hunting Status',col=c(5,4))
```



We would like to see whether this difference is 'statistically significant'- basically whether it is meaningful or not.

The t-test

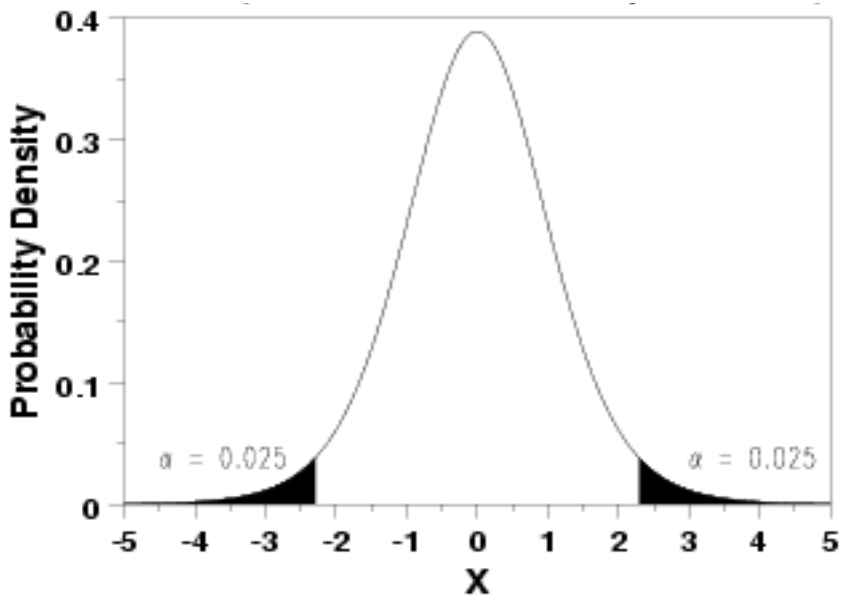
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

The Numerator is the Signal

For the 2-sample **t-test**, the **numerator** is the difference between the means of the two samples. For example, if the mean of group 1 is 10, and the mean of group 2 is 4, the difference is 6. The default null hypothesis for a 2-sample **t-test** is that the two groups are equal.

The Denominator is the Noise

The denominator is the noise. The equation in the denominator is a measure of variability known as the standard error of the mean. This statistic indicates how accurately your sample estimates the mean of the population. A larger number indicates that your sample estimate is less precise because it has more random error. The formula appears complicated because it allows for different sample sizes in the 2 groups.



The t-Value: The Ratio of Signal to Noise

Relatively large signals and low levels of noise produce larger t-values. If the signal does not stand out from the noise, it's likely that the observed difference between the sample estimate and the null hypothesis value is due to random error in the sample rather than a true difference at the population level.

To judge what we consider 'large' we compare to a distribution similar to the normal distribution.

We will discuss distributions and tests in more detail after the mid-term.

To calculate the t-test to test the difference in progersterone between Lightly and Heavily hunted wolves we would look at:

$$(X_1 - X_2)/ \text{standard error}$$

$$= (28.15 - 21.36)/ \text{standard error}$$

In R there is a function to perform a t-test:

```
> t.test(Heavy_F$Ppgmg,Light_F$Ppgmg)

data:  Heavy_F$Ppgmg and Light_F$Ppgmg
t = 3.614, df = 61.452, p-value = 0.0006089
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.030893 10.536549
sample estimates:
mean of x mean of y
 28.14870  21.36498
```

We conclude that there is a highly significant difference between the two groups.

Summary:

- When faced with new data the best place to start understanding it is with some simple graphics-scatterplots, histograms/stem-leaf and boxplots are all very useful.
- It is worth putting in the effort to make the plots as clear as possible. Proper labelling, scaling and aesthetics are important parts of that.
- It is often useful to 'test' a question using statistical techniques. R has a wide array of statistical functions. We have given a small example of this (t-test). Later we will go into some detail of statistical methods in R.