ST1050

Class 5

Histograms, Stemleaf and Boxplots

Histograms and Box and whisker plots

(based on Introduction to Data Visualisation in R, Dr J Yearsley, UCD)

WOLF.CSV

This file is a text file of comma separated variables.

The data in this file are from the publication:
Bryan H, Smits J, Koren L, Paquet P, Musiani M, Wynne-Edwards K (2014) Heavily hunted wolves have higher stress and reproductive steroids than wolves with lower hunting pressure. Functional Ecology 29(3): 347-356.

This dataset includes measurements of cortisol, testosterone, and progesterone in wolf hair samples collected from hunters in the tundra-taiga and northern boreal forest of Canada. Additional samples were collected from wolves killed as part of a control program in the boreal forest (population 3).

This dataset has seven variables:

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]

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)

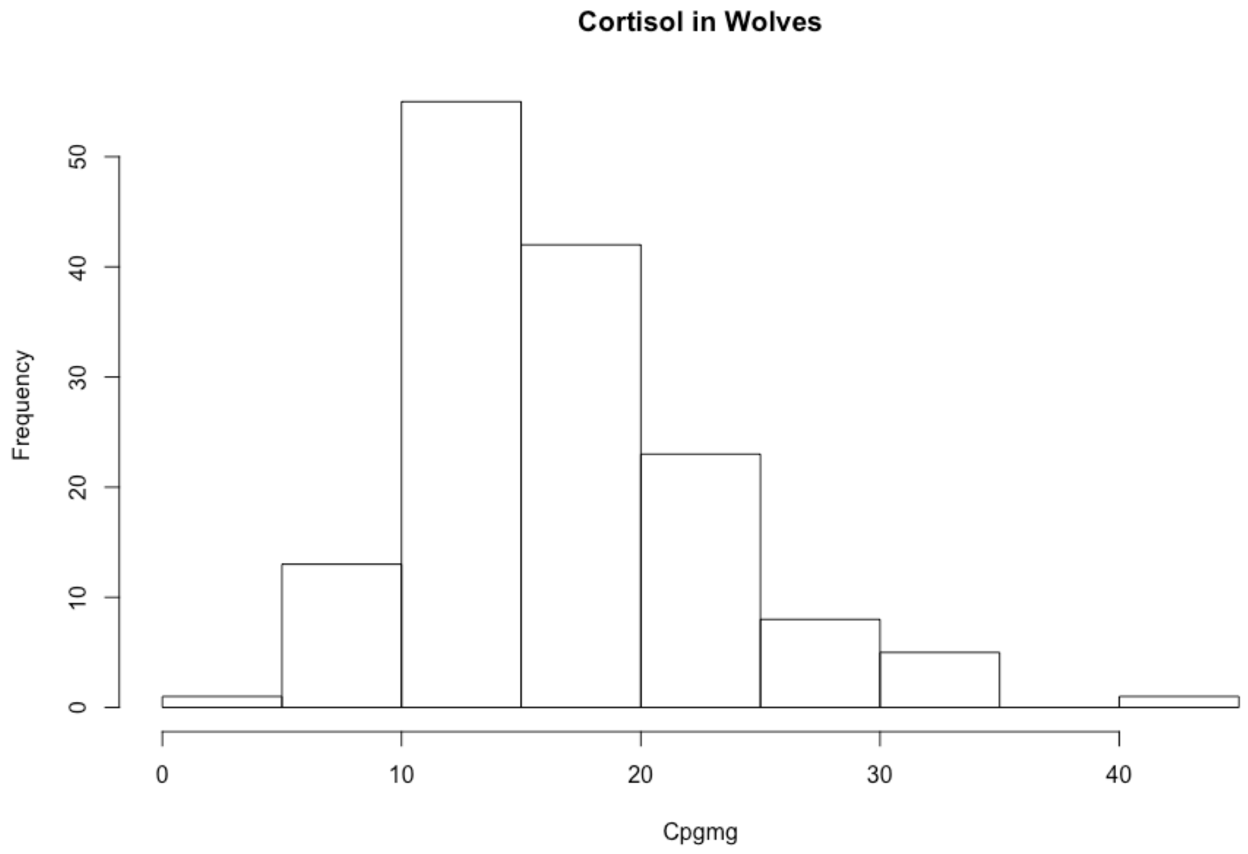
We also set up the following variables for simplifying commands:

Population = wolf.sub$Population
Sex = wolf.sub$Sex
Cpgmg = wolf.sub$Cpgmg
Tpgmg = wolf.sub$Tpgmg
Tpgmg = wolf.sub$Ppgmg
Hunting = wolf.sub**$**Hunting (same as population but a factor)

A **histogram** is a representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) and was first introduced by Karl Pearson.

A histogram only considers one variable. Later we will study 'barcharts'; these look similar to histograms but are actually quite different and can consider more than one variable.

```
> hist(Cpgmg,main='Cortisol in Wolves')
```



We wish to contrast cortisol levels (Cpgmg) in males and females. First we see what the frequency distribution of males and females is:

```
> table(Sex)
Sex
 F  M  U
72 76  0
```

We should get rid of the (empty) 'U' category in the Sex variable-
we can do this using the 'droplevels()' function.

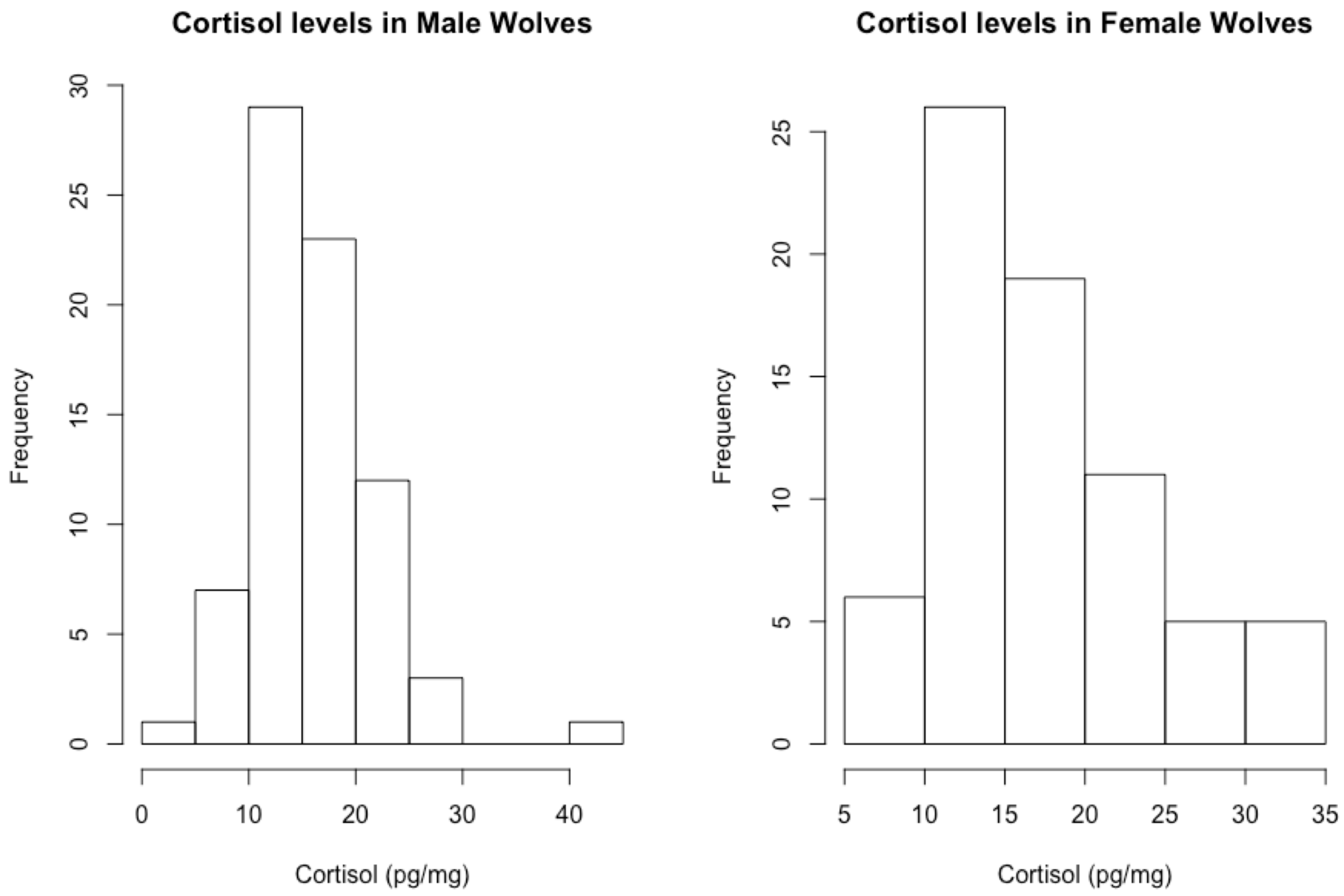
```
wolf.sub=droplevels(subset(wolf.sub, Sex!='U'))
```

```
> table(wolf.sub$Sex)

 F  M
72 76
```

NB: The 'local' variable Sex will now need to be rewritten:

```
Sex = wolf.sub$Sex
```

```
> par(mfrow=c(1,2))
> Cpgmg_m=Cpgmg[Sex=='M']
> Cpgmg_f=Cpgmg[Sex=='F']
> hist(Cpgmg_m,main='Cortisol levels in Male Wolves',xlab='Cortisol (pg/mg)')
> hist(Cpgmg_f,main='Cortisol levels in Female Wolves',xlab='Cortisol (pg/mg)')
```



Advantages of considering histograms for variable description:

- Gives a good idea of the frequency distribution
- Outliers are easily spotted (although in a large data set that can be confusing)
- It very quickly gives an idea of the data since it is visual.

Disadvantages:

- Have to be a bit careful with outliers in large datasets
- Only one variable can be considered- and only numerical variables.
- Changing the groupings of the bars can change the way the histogram looks quite a bit
- Histograms display the number of values within an interval and not the actual values- unlike stem-leaf plots.

Stem-Leaf plot:

A **Stem and Leaf Plot** is a table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit). Scale controls how long the plot is.

```
> sort(Cpgmg)

4.75
6.37
7.61
7.93
7.93
8.00
8.19
8.84
8.91
9.10
9.17
9.43
9.95 etc
```

```
> stem(Cpgmg,scale=1)

  The decimal point is at the |

   4 | 8
   6 | 4699
   8 | 0289124
  10 | 00123568133446666888
  12 | 0011222334556678022235579
  14 | 0000233458892233344899
  16 | 0223355680223444678
  18 | 114689168899
  20 | 0004455114579
  22 | 245767
  24 | 0146627
  26 | 34838
  28 | 5
  30 | 2
  32 | 229
  34 | 0
  36 |
  38 |
  40 | 4
```

```
> stem(Cpgmg,scale=3)

  The decimal point is at the |

   4 | 8
   5 |
   6 | 4
   7 | 699
   8 | 0289
   9 | 124
  10 | 00123568
  11 | 133446666888
  12 | 0011222334556678
  13 | 022235579
  14 | 000023345889
  15 | 2233344899
  16 | 022335568
  17 | 0223444678
  18 | 114689
  19 | 168899
  20 | 0004455
  21 | 114579
  22 | 2457
  23 | 67
  24 | 01466
  25 | 27
  26 | 348
  27 | 38
  28 |
  29 | 5
  30 | 2
  31 |
  32 | 22
  33 | 9
  34 | 0
  35 |
  36 |
  37 |
  38 |
  39 |
  40 | 4
```

Boxplots

A broad indication of a quantitative variable's distribution can be seen by plotting quantiles (e.g. 0%, 25%, 50%, 75% and 100% quantiles correspond to minimum, 1st quartile, median, 3rd quartile and the maximum).

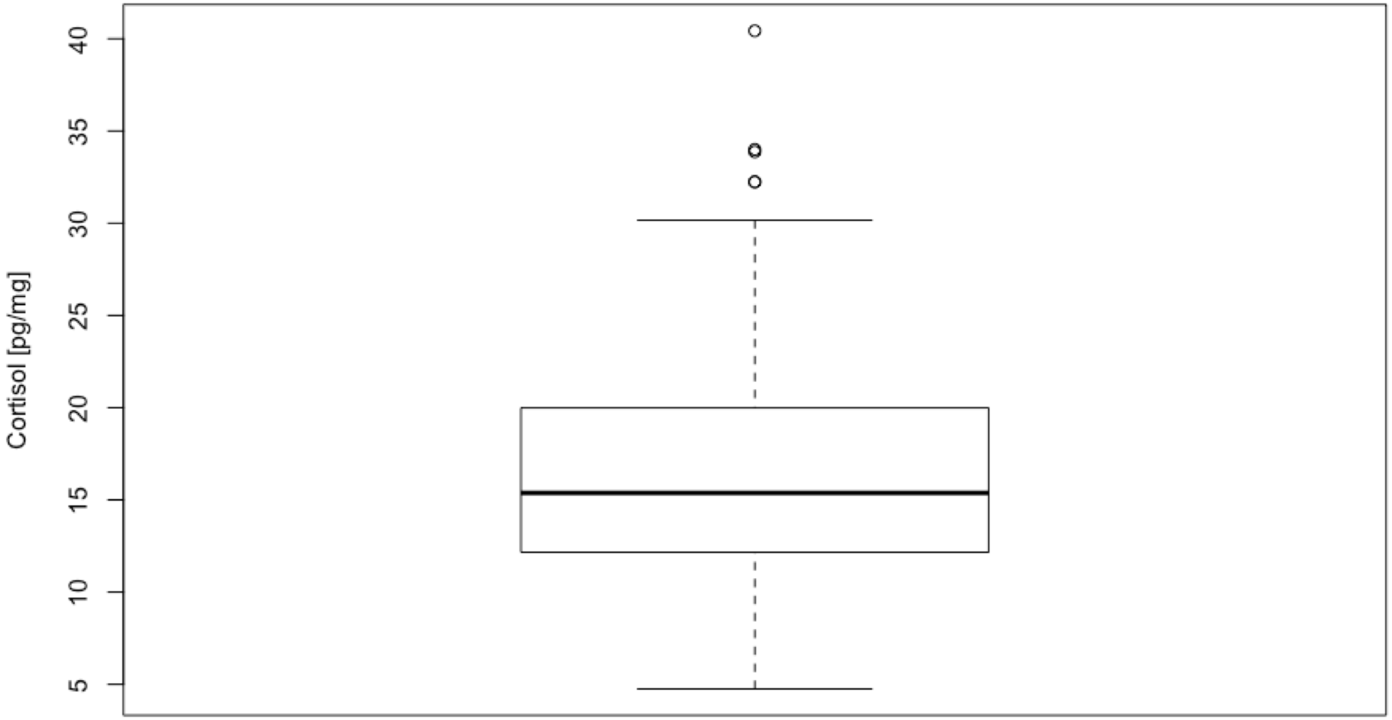
Quantiles can be calculated using the `quantile()` function:

```
# We will plot the 0%, 5%, 25%, 50%, 75%, 95% and 100% quantiles for the Cpgmg variable in the wolf data frame.

> quantile(Cpgmg, probs=c(0, 0.05, 0.25, 0.5, 0.75, 0.95, 1.0) )
     0%      5%     25%     50%     75%     95%    100%
  4.7500  8.8645 12.1600 15.3750 19.9750 27.6380 40.4300
```

Quantiles are commonly represented on a **box and whiskers plot**. The boxplot() functions can be used for this.

Box and whiskers plot for the Cpgmg variable in the wolf data frame

```
boxplot(Cpgmg, ylab='Cortisol [pg/mg]')
```



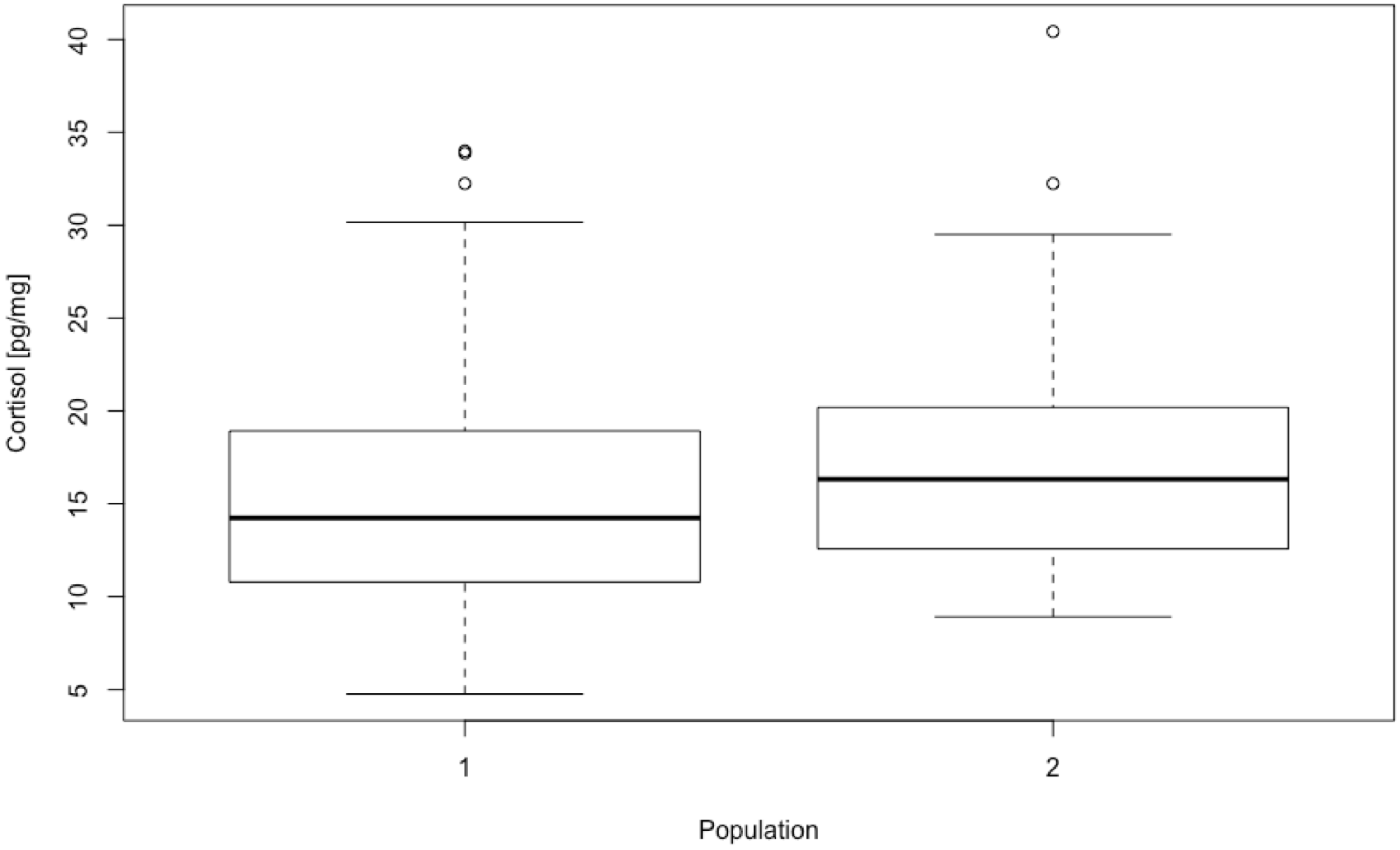
The box and whiskers plot displays:

the median as the central bar in the box
the 25% quantile as the lower end of the box
the 75% quantile as the upper end of the box outliers as individual points
whiskers extend to 1.5 times the inter-quartile range

Box and whisker plots show less information than a histogram but they can be used to easily plot the distributions from several variables.

For example, we can compare the distributions from the two populations in the wolf.sub data frame.

```
# Box and whiskers plot for the Cpgmg variable from the two populations separately

boxplot(Cpgmg~Population, data=wolf.sub, ylab='Cortisol [pg/mg]', xlab='Population')
```



Using a formula to specify a plot

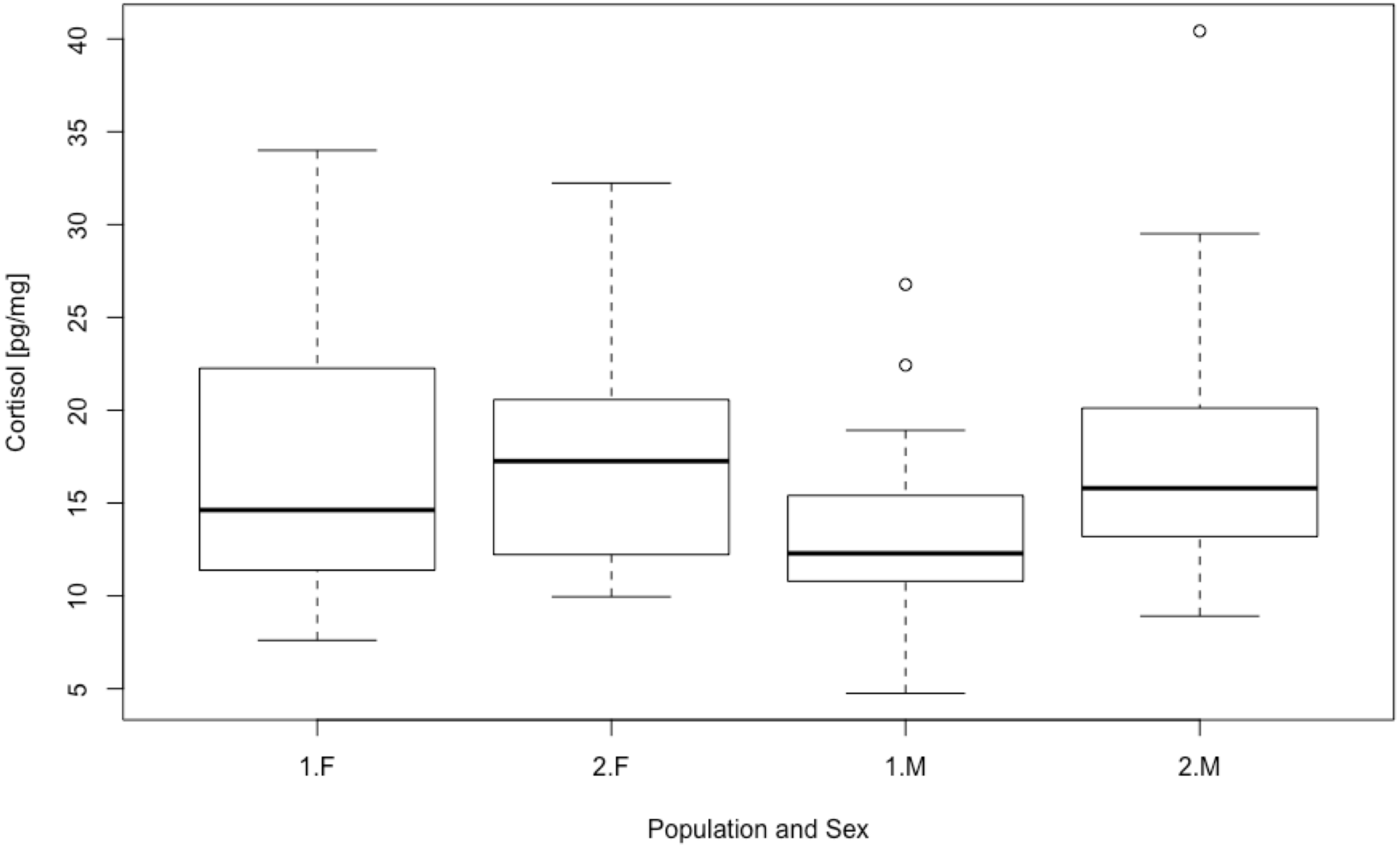
In the box and whiskers plot above we used the formula
`Cpgmg ~ Population` to specify which data to plot.

The **~** symbol (called a tilde) identifes a formula.

To the left of the ~ is the variable for the y-axis,
To the right of the ~ is the variable(s) for the x-axis.

You can put more than one variable on the x-axis. Here is an example of plotting box and whisker plots of cortisol for different populations and different sexes.

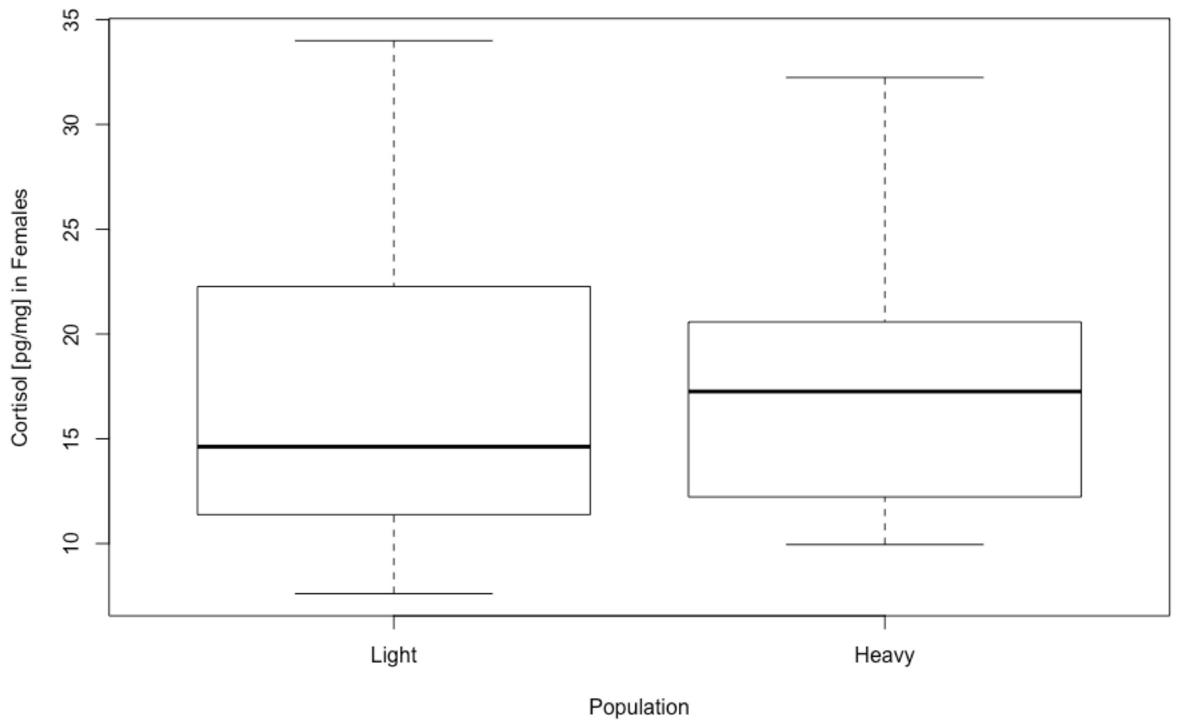
```
# Box and whiskers plot for the Cpgmg variable in the wolf data frame

boxplot(Cpgmg ~ Population + Sex, data=wolf.sub, ylab='Cortisol [pg/mg]',
     xlab='Population and Sex')
```



Plotting a subset of a data frame

Using this formula notation to specify a plot makes it very easy to plot subsets of a data frame by using the `subset()` function. We also add clearer description to the population groups here using 'names'. Here is an example of the code for a box and whiskers plot where Cpgmg in just females is plotted:

```
# Box and whiskers plot for the Cpgmg variable for just females
boxplot(Cpgmg ~ Population, data=subset(wolf.sub, Sex=='F'), ylab='Cortisol [pg/mg] in Females',
xlab='Population',names=c('Light','Heavy')))       #Light corresponds to population=1.
```



Plotting a subset of a data frame(cont'd)

```
> par(mfrow=c(1,2))
> # Box and whiskers plot for the Tpgmg (i.e. testosterone levels) variable (males only):
> boxplot(Tpgmg ~ Population, data=subset(wolf.sub, (Hunting=='Heavy'  &Sex=='M')), ylab='Testosterone in Heavily Hunted', xlab='Males Only')
> boxplot(Tpgmg ~ Population, data=subset(wolf.sub, (Hunting=='Light'  &Sex=='M')), ylab='Testosterone in Lightly Hunted', xlab='Males Only')
```