Week 5: Exploring Data

Session 1

Spring 2020

iClicker Question 1

Which of the following should I use to read the file `mydata.csv` into a data frame called `dat` in R?

A `dat = read.csv(mydata.csv)`

B `read.csv(mydata.csv)`

C `dat = read.csv("mydata.csv")`

D `read.csv(mydata.csv, row.names = 1)`

E `read.csv("mydata.csv")`

iClicker Question 2

Which symbol do we use to represent the **sample mean**?

A σ

B $\bar{s}$

C $\bar{x}$

D μ

E $\bar{m}$

iClicker Question 3

Which symbol do we use to represent the **population mean**?

A σ

B $\bar{s}$

C $\bar{x}$

D μ

E $\bar{m}$

iClicker Question 4

Which plot type is most appropriate to show the **distribution** of a set of measurements?

A scatterplot

B boxplot

C barchart

D histogram

E pie chart

i>clicker®

iClicker Question 5

Which of the following lines of code will make a scatterplot of the dataframe with length on the x-axis and mass on the y-axis?

```
##       length     width mass
## 1 0.7209039 2.8777226   37
## 2 0.8757732 1.6052823   29
## 3 0.7609823 0.4713699   19
```

A `plot(dat$mass, dat$length)`

B `scatter(dat$length, dat$mass)`

C `boxplot(dat$length, dat$mass, type = "p")`

D `dotplot(dat$length, dat$mass)`

E `plot(dat$length, dat$mass)`

Announcements

Trying a different slide layout today

Graphical exploration

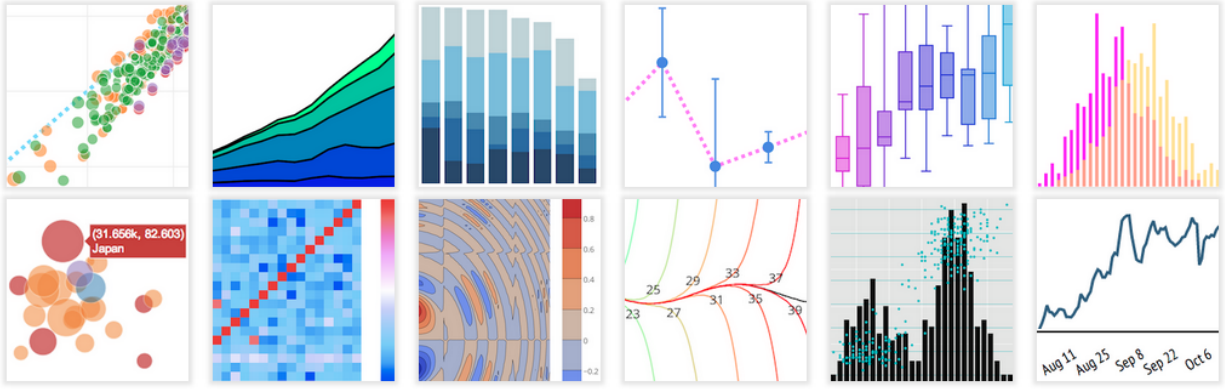
Why use graphs?

Graphical exploration

Two main reasons to use graphs:

1. Inform how to analyze the data

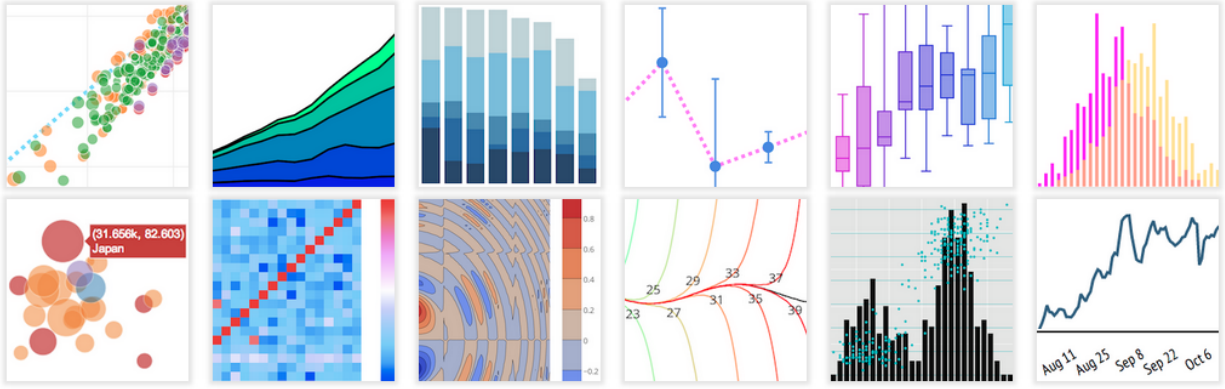
2. Presentation of the data



Graphical exploration

Two main reasons to use graphs:

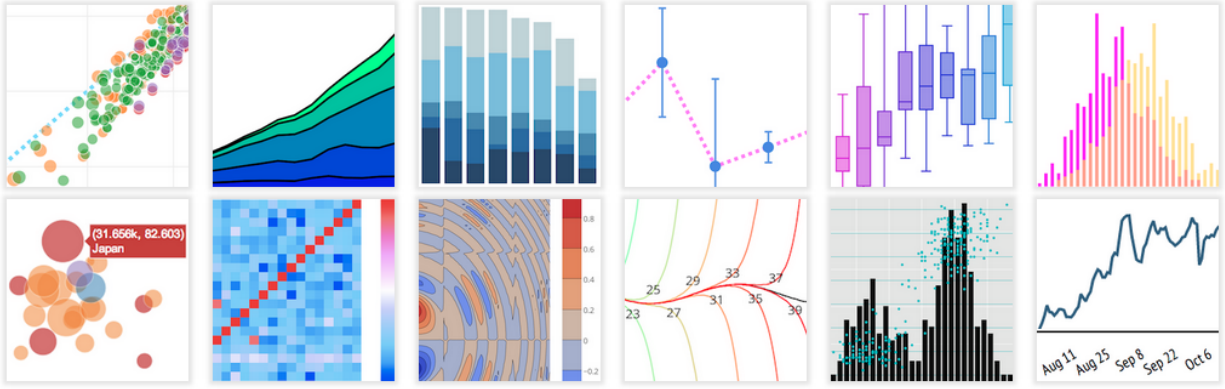
1. Inform how to analyze the data
   - visualization
   - identify patterns
   - choose appropriate statistical test
2. Presentation of the data



Graphical exploration

Two main reasons to use graphs:

1. Inform how to analyze the data
   - visualization
   - identify patterns
   - choose appropriate statistical test
2. Presentation of the data
   - summarize results
   - communicate results
   - publish results



Types of graphs - *Exploratory*

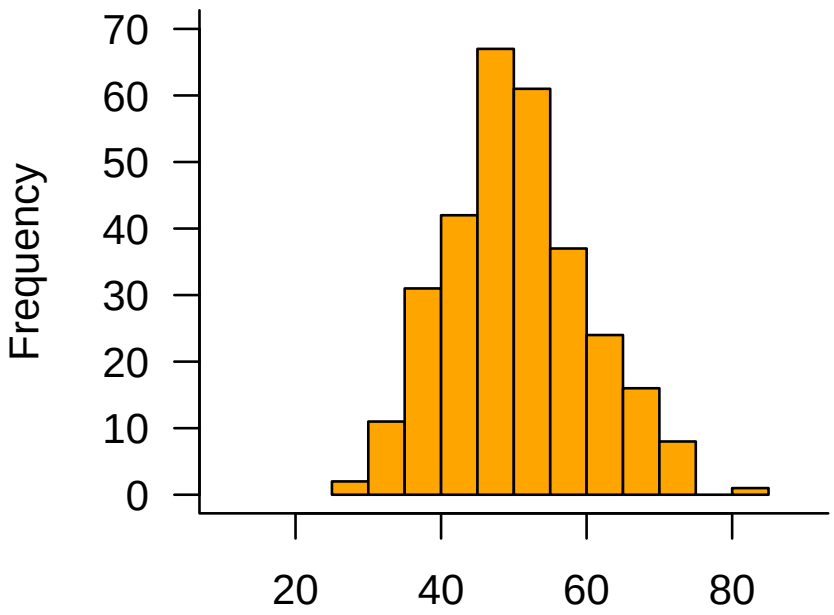
Exploratory graphs help understand the distribution of the data:

- are the data normally distributed
  - important assumption in statistics
  - determines how data are analyzed
- what is the central tendency
- what is the spread
- general summaries of the data

Exploratory: *Histogram*

- width of bars are defined data bins or intervals
- height of bars represent bin-specific frequencies

```
hist(values)
```

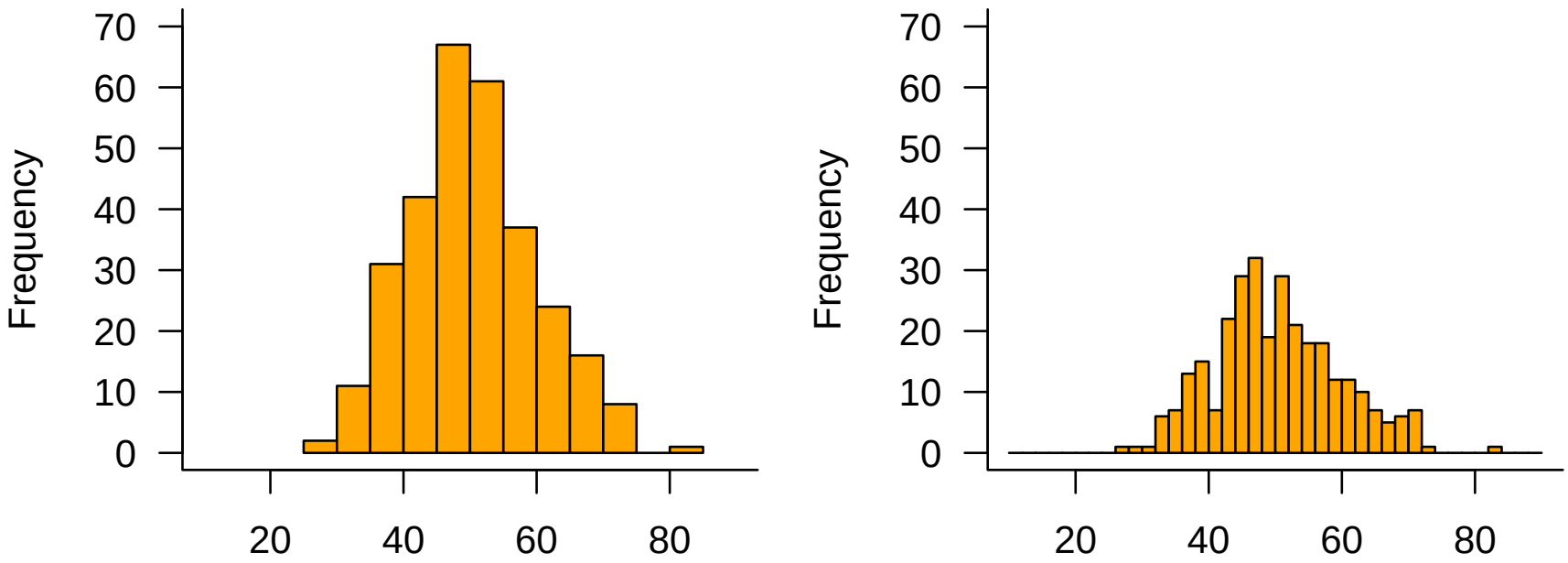


Exploratory: *Histogram*

- width of bars are defined data bins or intervals
- height of bars represent bin-specific frequencies

You can change the number and widths of the bins.

```
hist(values, breaks=seq(10,90,2))
```

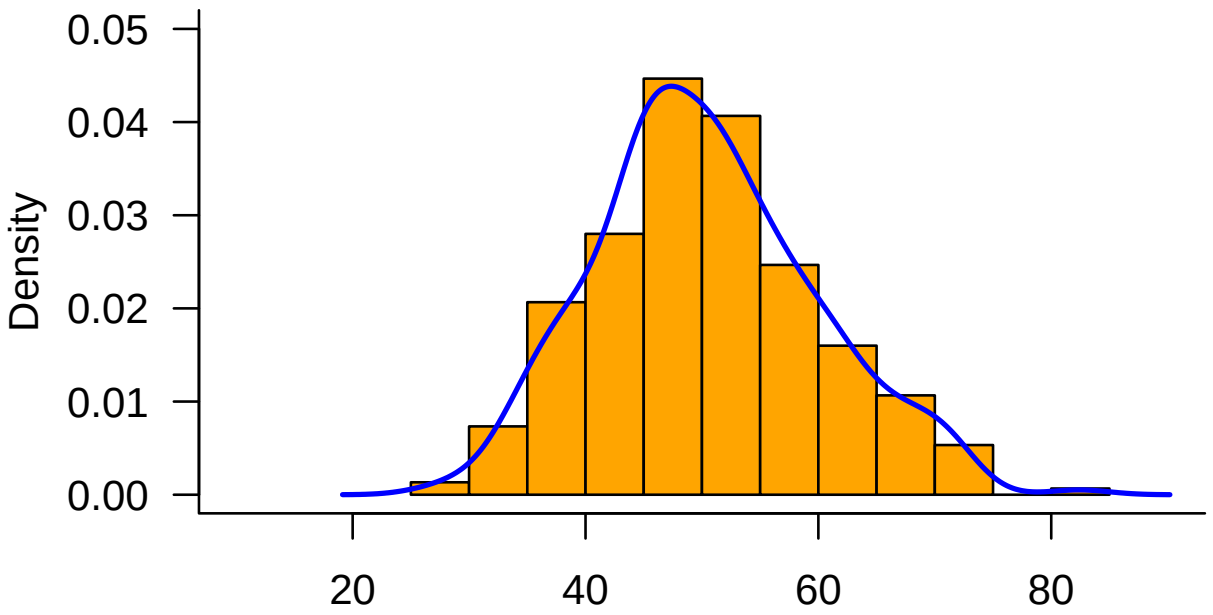


Exploratory: *Histogram + Density Plot*

A *density plot*: smoothed version of histogram

- To overlay on a histogram, tell `hist()` to plot the *probability* version of the histogram:

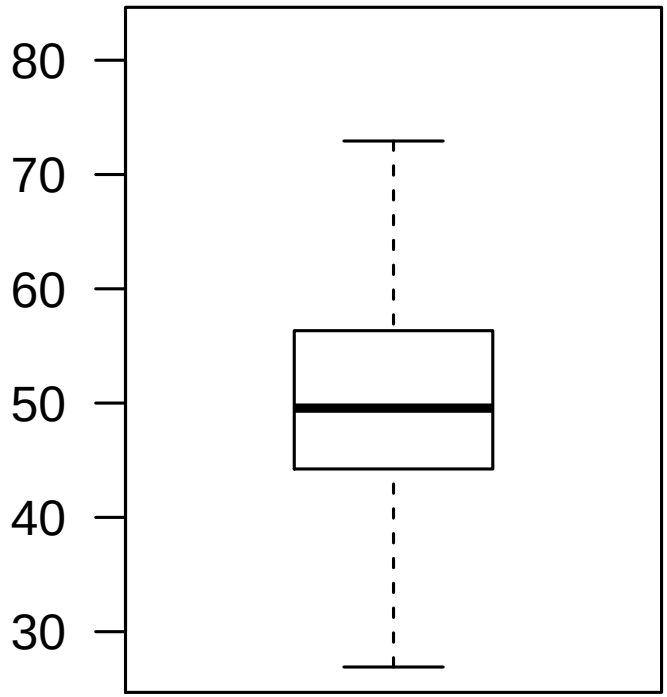
```
hist(values, probability=TRUE)
lines(density(values))
```



Exploratory: *Box-whisker/Box plot*

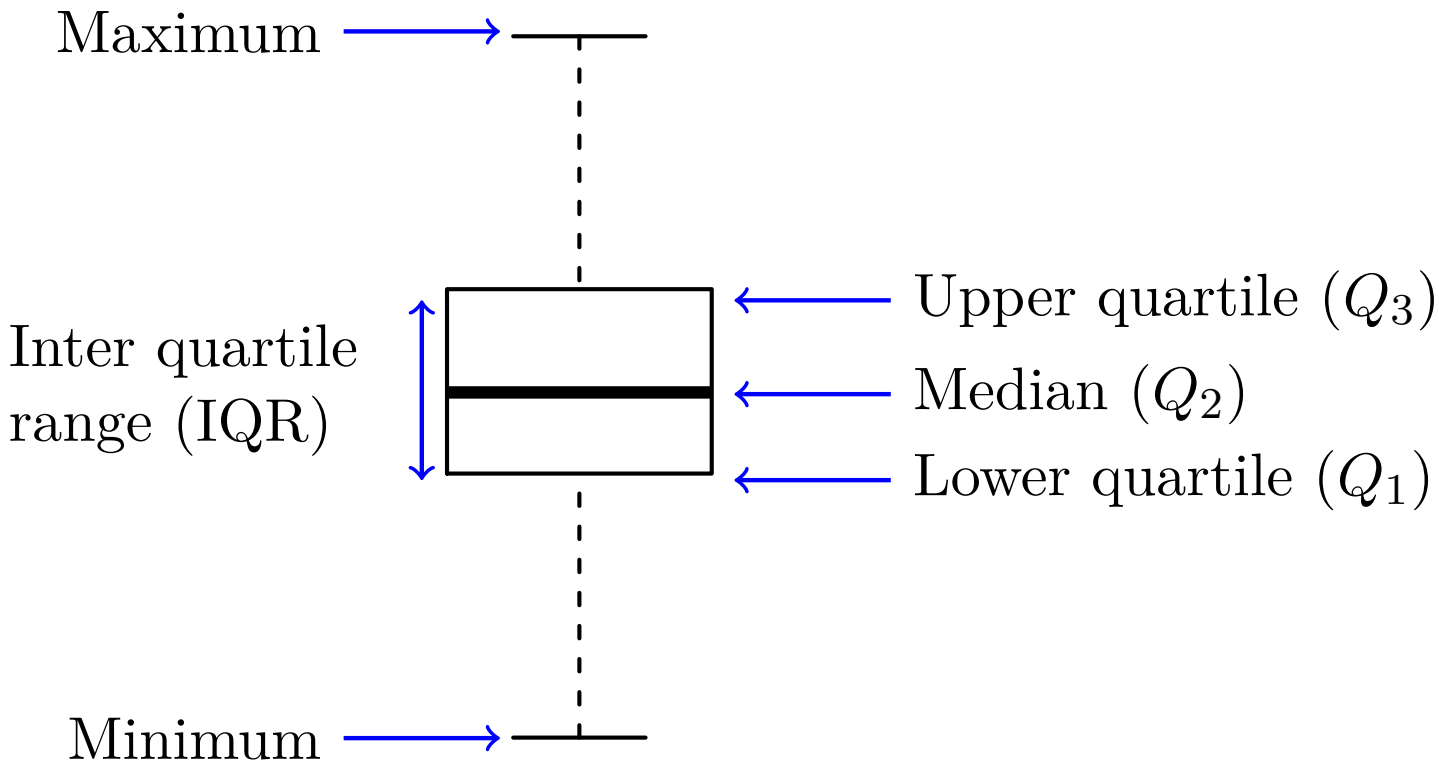
- distribution
- outliers
- symmetry or skewness

```
boxplot(values)
```



Exploratory: *Box-whisker/Box plot*

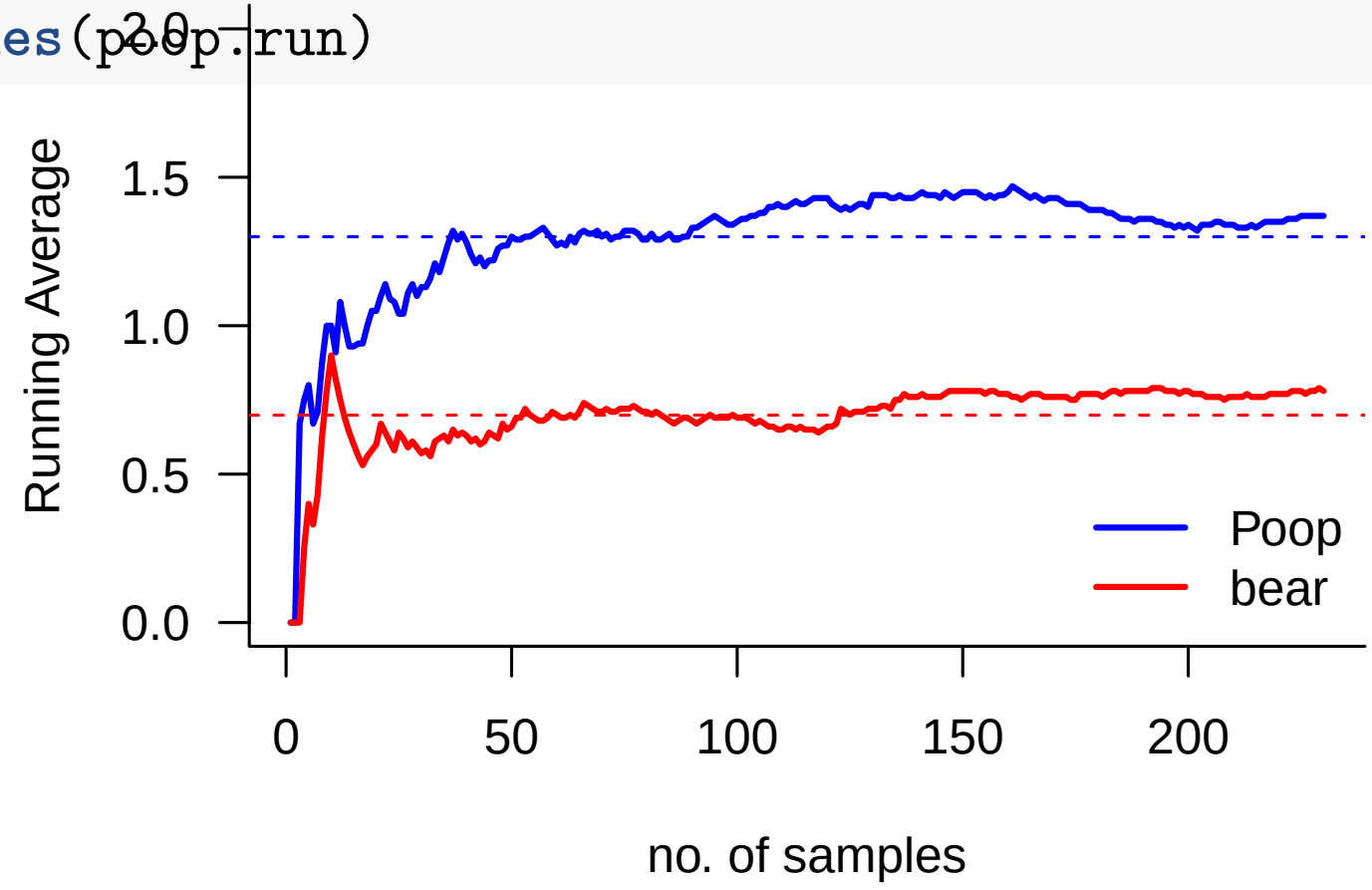
- R: `boxplot(x) # x is data`



Exploratory: *Line graph*

Line graph is a useful plot for running average or time series data

```
plot(bear.run, type="l") #"l": line, "p": points, "b": both
lines(poop.run)
```



Differences

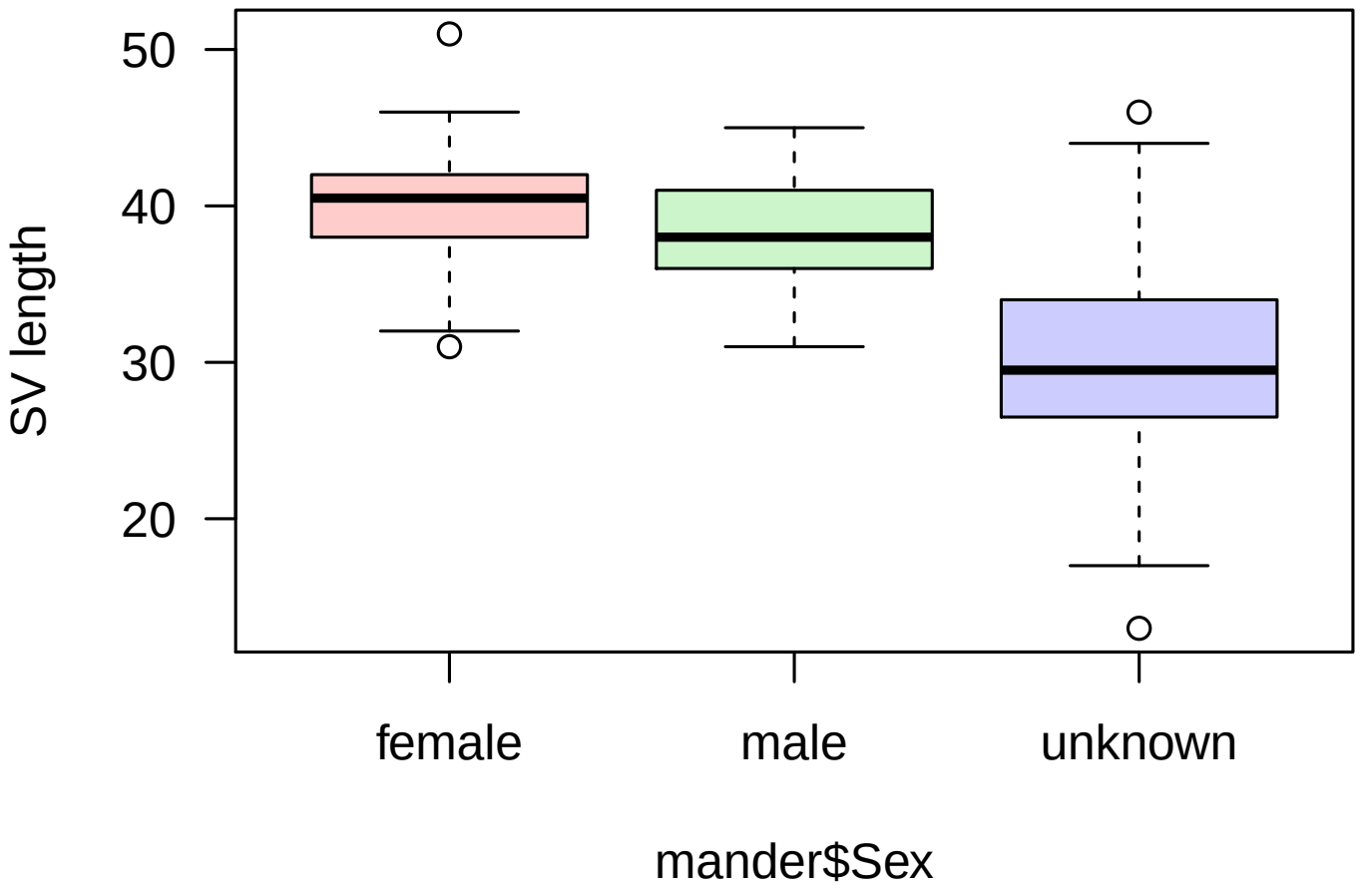
To visualize differences between groups

- box-whisker plots
  - compares averages
  - compares distribution
- bar charts
  - compares averages

Differences: *Box-whisker plot*

Compare salamander snout-vent lengths be three sexes:

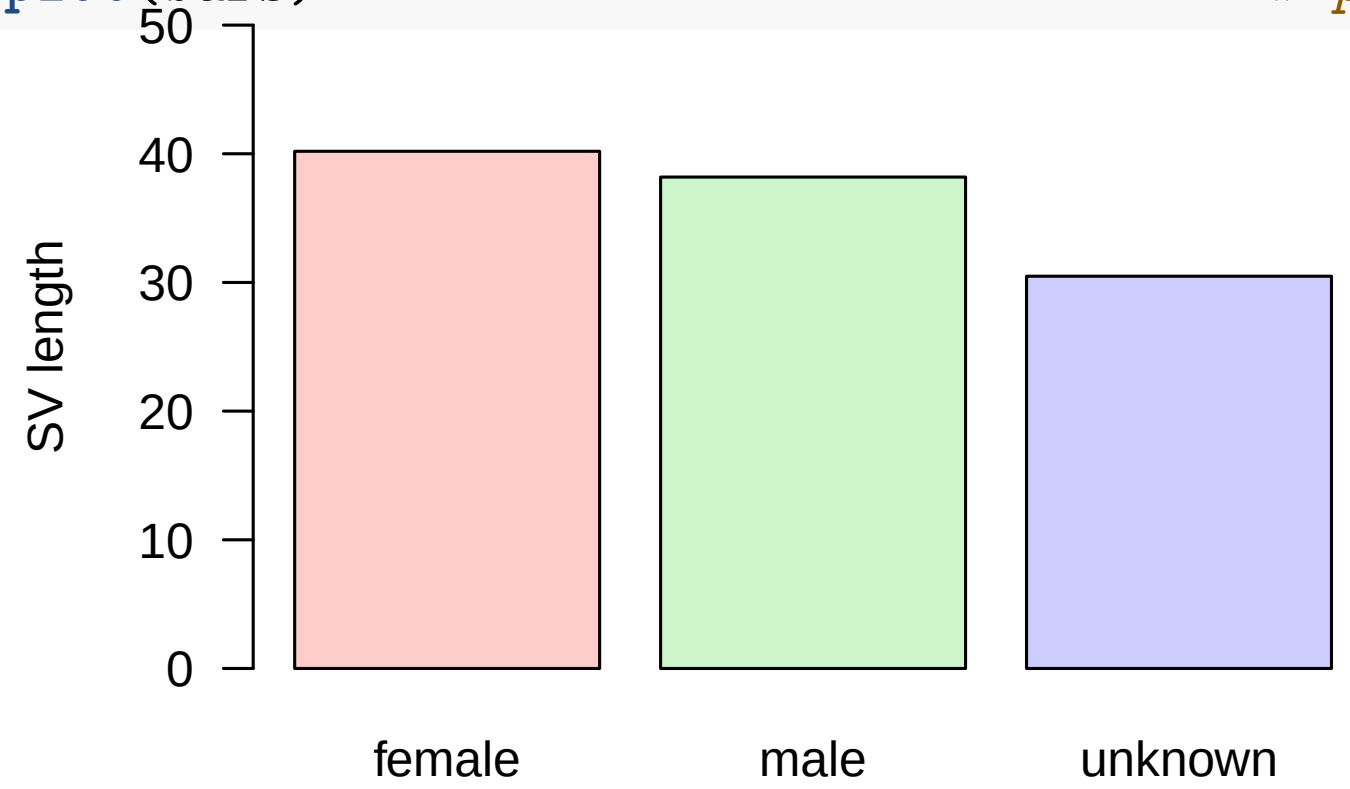
```
boxplot(mander$SVL ~ mander$Sex) #formula notation
```



Differences: *Bar chart*

Compare salamander snout-vent lengths be three sexes:

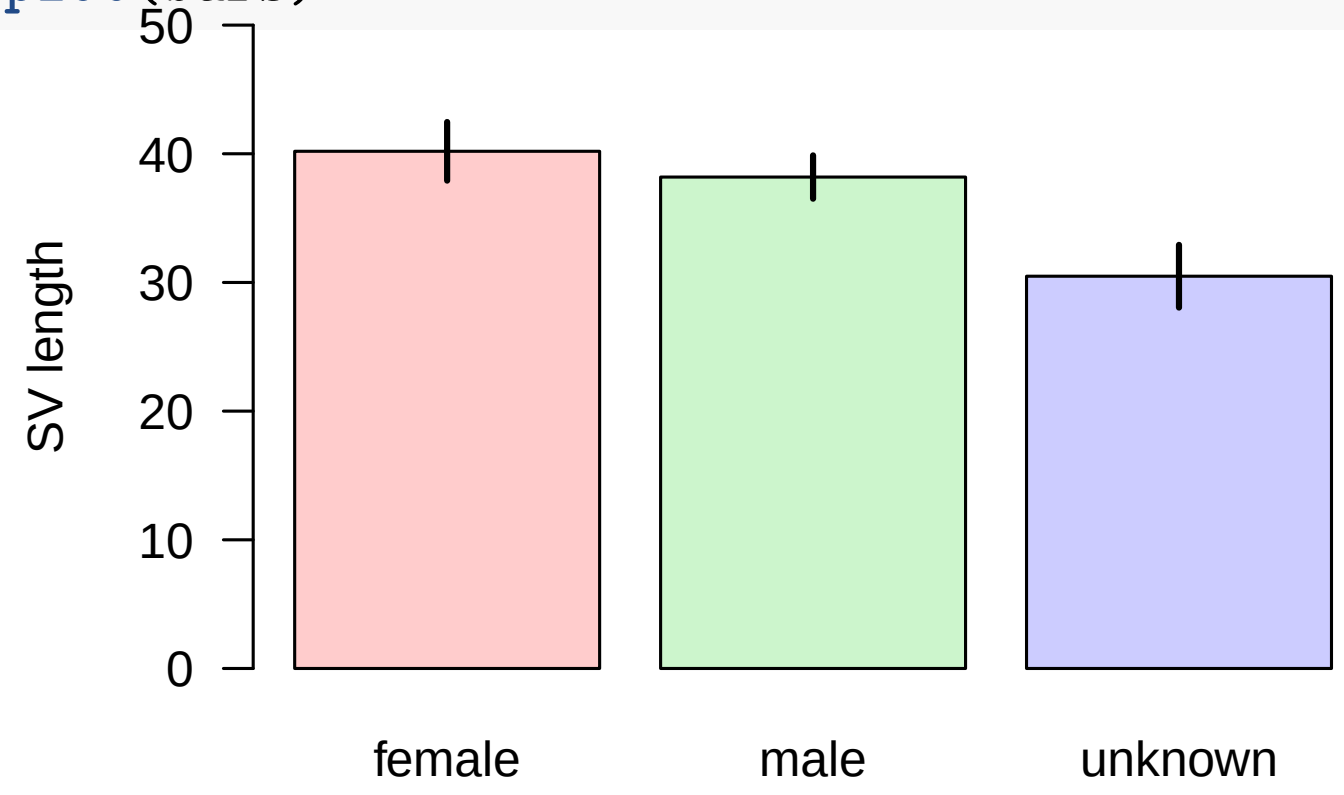
```
bars <- tapply(mander$SVL,mander$Sex,mean) #create matrix
barplot(bars)                              # plot it
```



Differences: *Bar chart* with associated error

Compare salamander snout-vent lengths be three sexes:

```
bars <- tapply(mander$SVL,mander$Sex,mean)
barplot(bars)
```



Links

Two main approaches for relationships between data:

1. Correlations
2. Associations

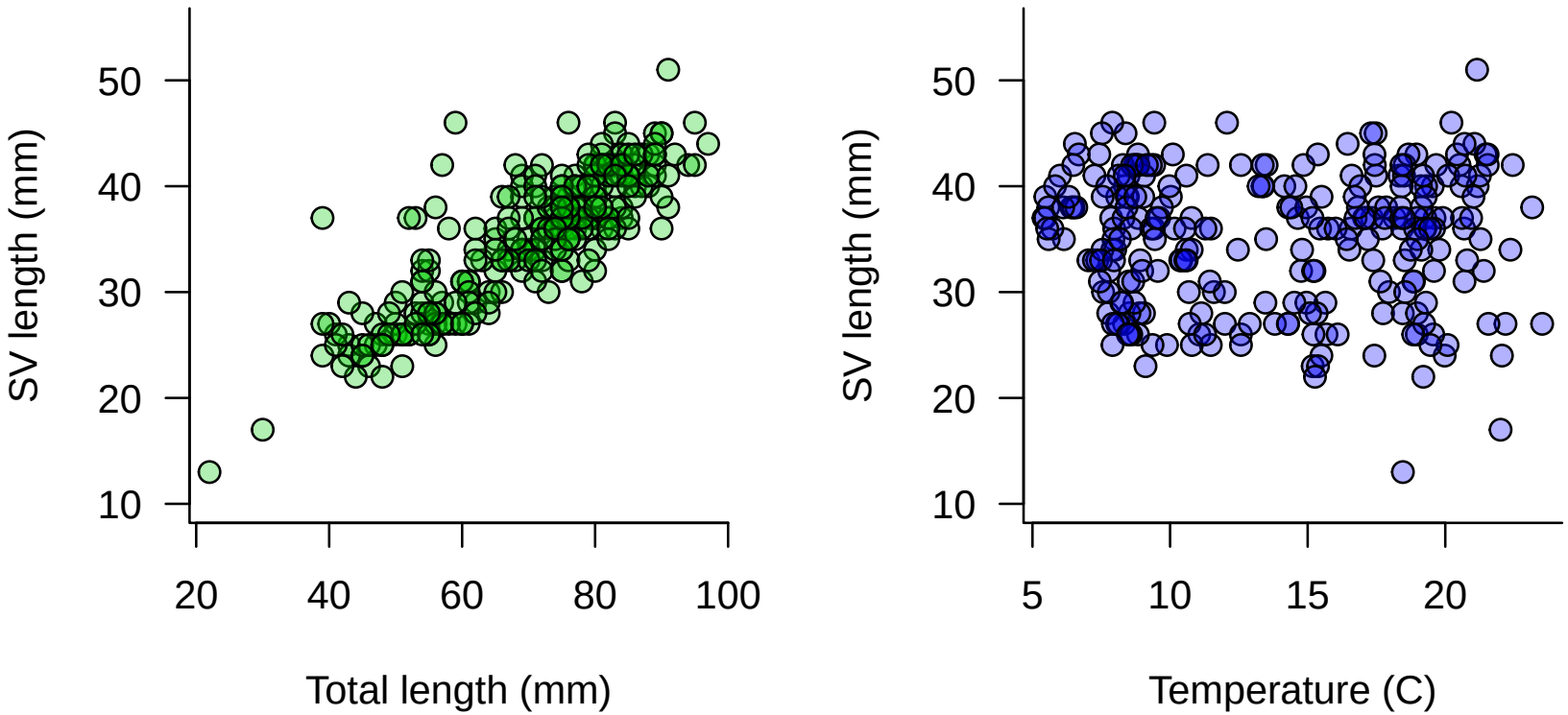
Links

Two main approaches for graphing relationships between data:

1. Correlations
   - two numeric variables
     - *de*pendent variable (of primary interest: y-axis)
     - *inde*pendent variable (explanatory variable: x-axis)
   - how one variable is related to another
   - *scatter plots*

Links: *Scatter plot*

```
plot(x,y) # x and y are numeric vectors
```



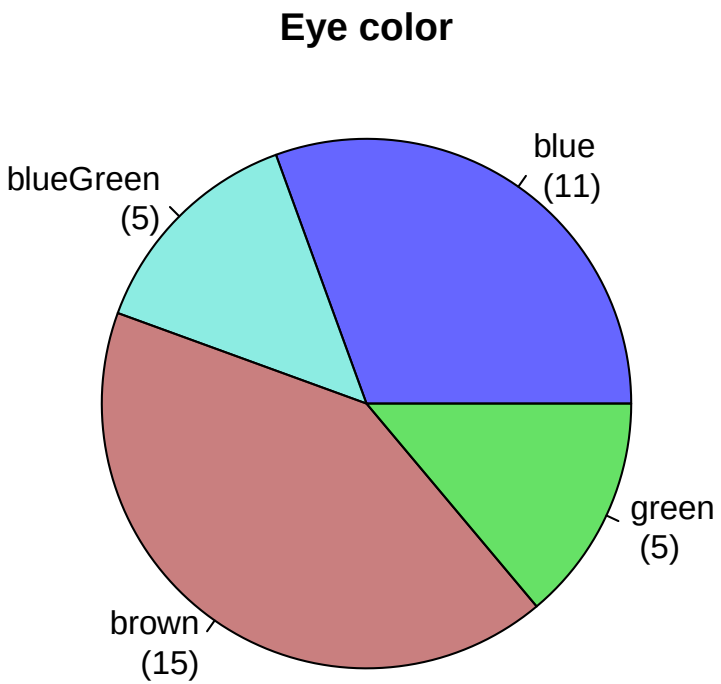
Links

Two main approaches for graphing relationships between data:

2. Associations
   - categorical data
   - summarize categories
     - counts
     - proportions
     - by rows and/or columns of a table
   - *pie charts* for single categories
   - *bar graphs* for several categories

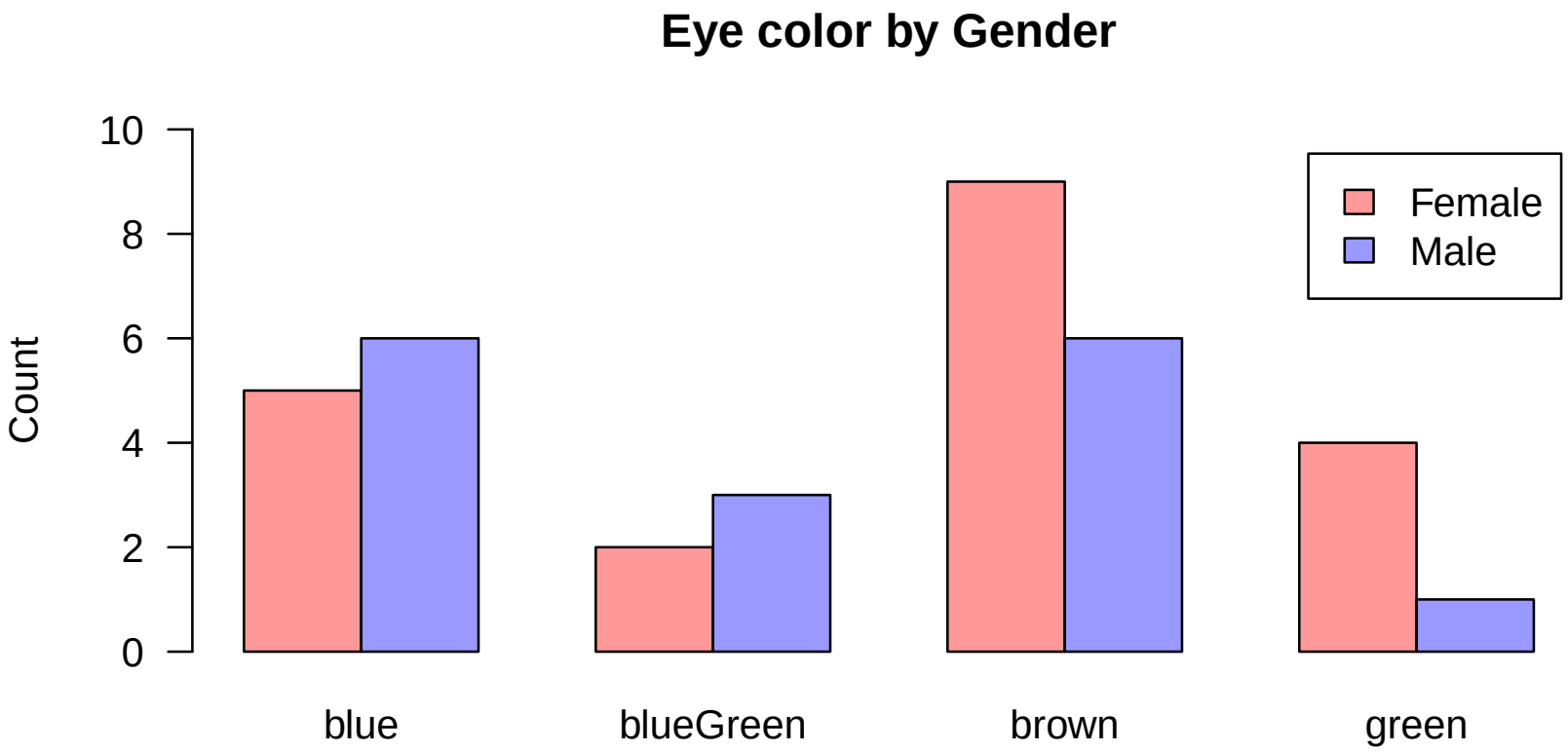
Links: *Pie chart*

```
pietab <- table(classData$Eyes)
pie(pietab) #(number of people with each eye color)
```



Links: *Bar chart*

```
bartab <- table(classData$Gender,classData$Eyes)
barplot(pietab, beside=TRUE) #(number of each gender with
```



Some graphics pointers

In summary, graphs are a useful data visualization tool

- summarizing
- understanding
- describing
- presenting/communicating

Some graphics pointers

In summary, graphs are a useful data visualization tool

- summarizing
- understanding
- describing
- presenting/communicating

BUT we must label the well or they are useless!

- label both axes
- provide a main title for your graph
- avoid clutter
- make it readable
- *I expect graphs to be propery labeled from now on*!

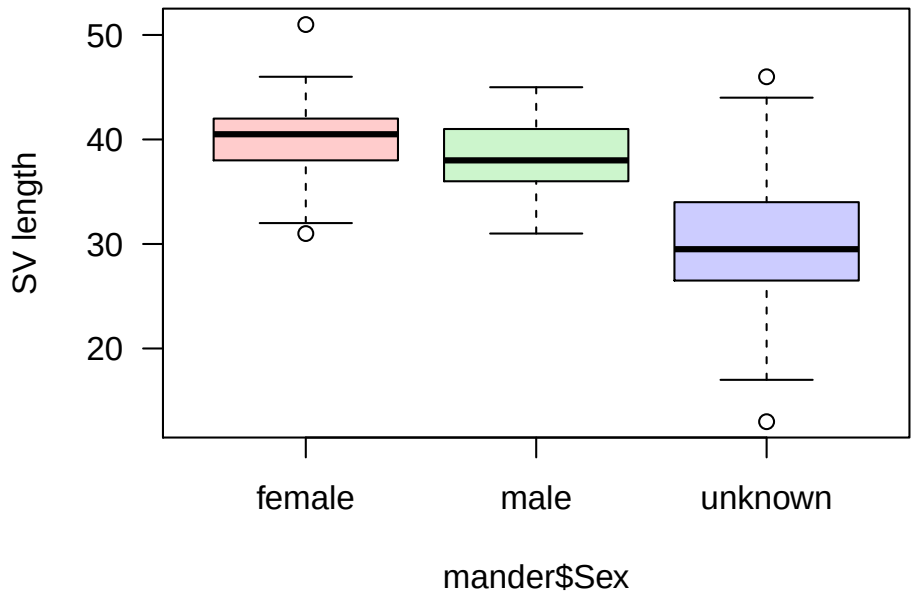
Some graphics pointers

In summary, graphs are a useful data visualization tool

Purpose	Graph Type
Illustrating *distribution*	Histogram, Density plot
	Box(-whisker) plot
Illustrating *differences*	Bar chart, Box plot
Illustrating *correlations*	Scatter plot
Illustrating *associations*	Pie chart, Bar chart
Illustrating *sample size*	Line plot of running avg

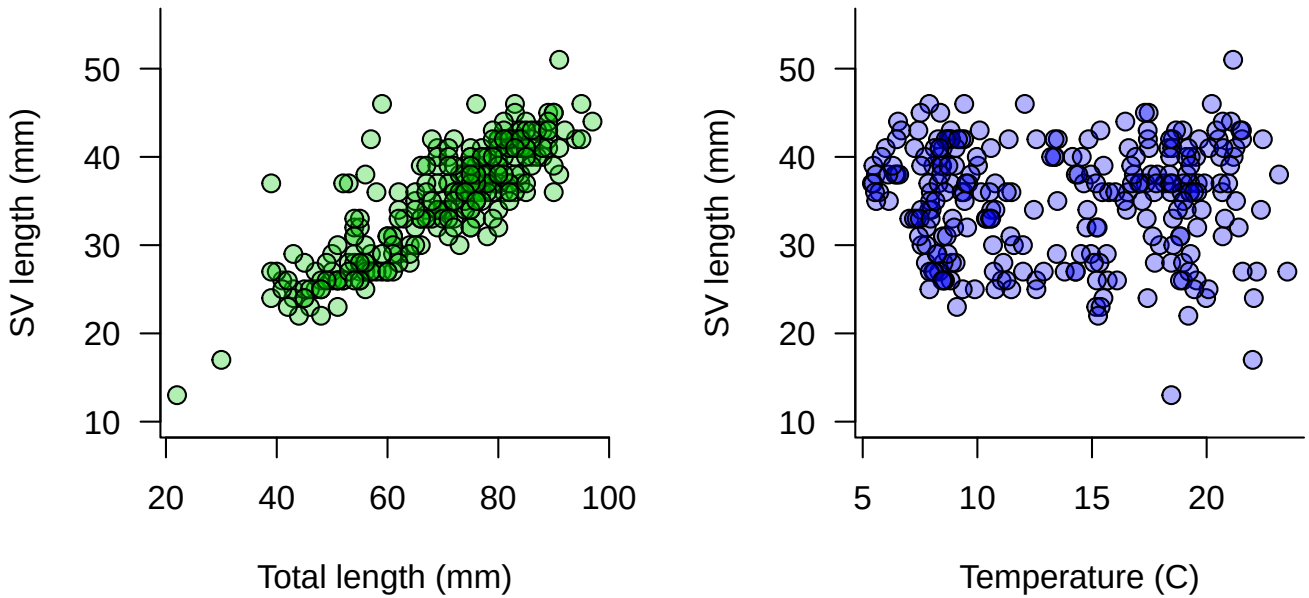
Beyond graphs, Towards statistics

- Graphs are powerful tools that provide insight and understanding of the patterns and relationships in the data.
- Don't give us the answer though:
  - are differences *significant*?
  - are associations *significatnt*?



Beyond graphs, Towards statistics

- Graphs are powerful tools that provide insight and understanding of the patterns and relationships in the data.
- Don't give us the answer though:
  - are differences *significant*?
  - are associations *significatnt*?



Beyond graphs, Towards statistics

- Graphs are powerful tools that provide insight and understanding of the patterns and relationships in the data.
- Don't give us the answer though:
  - are differences *significant*?
  - are associations *significatnt*?
- Statistics is the tool we use to formally answer these questions!
  - the differences *are/are not* significant!
  - are associations *are/are not* significant!