Week 6: Graphical Data Exploriation

Session 2

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 of the following lines of code will make a scatterplot with a custom **title**?

```
##       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, col = 3)`

B `plot(dat$mass, dat$length, type = "l")`

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

D `plot(dat$mass, dat$length, main = "data plot")`

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

iclicker®

iClicker Question 3

Which is the best location for me to type my R code for in-class and individual activities?

A A text file in Word or another word processor

B The RStudio console pane

C The RStudio code editor pane

D The RStudio Environment pane

i>clicker®

Announcements

- Error in Tuesday's title slide is now fixed. You can re-download the correct version on Moodle.

Follow-up questions from the Chapter 5 homework

Question 1: What is the difference between parametric and a non-parametric distribution?

- The book's explanation of *parametric distribution* is slightly misleading. The reading characterizes the Normal distribution as a parametric distribution, however there are many *parametric distributions*.
- I prefer to use the term *theoretical distributions* when I'm describing *parametric distributions* to emphasize that they are precisely mathematically defined.
- The behavior of *parametric distributions* is governed by 1 or more *parameters* in the probability functions.
- What is a **non-parametric distribution**?

Follow-up questions from the Chapter 5 homework

Question 3: In R, what would be returned if you compute the mean value of a vector containing the following values: 4, 9, 2, 13, NA, 9? Please describe your methods.

- Question from me: Is there a benefit to having `mean()` fail when you pass an NA?

For Today

- Are there any general questions about data exploration, numeric or graphical?
- Short lecture
- Group activity: graphical data exploration: rarefaction data.
- We'll start to use *inference* in the coming weeks!

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 them 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 properly labeled from now on*!

Some graphics pointers

In summary, graphs are a useful data visualization tool

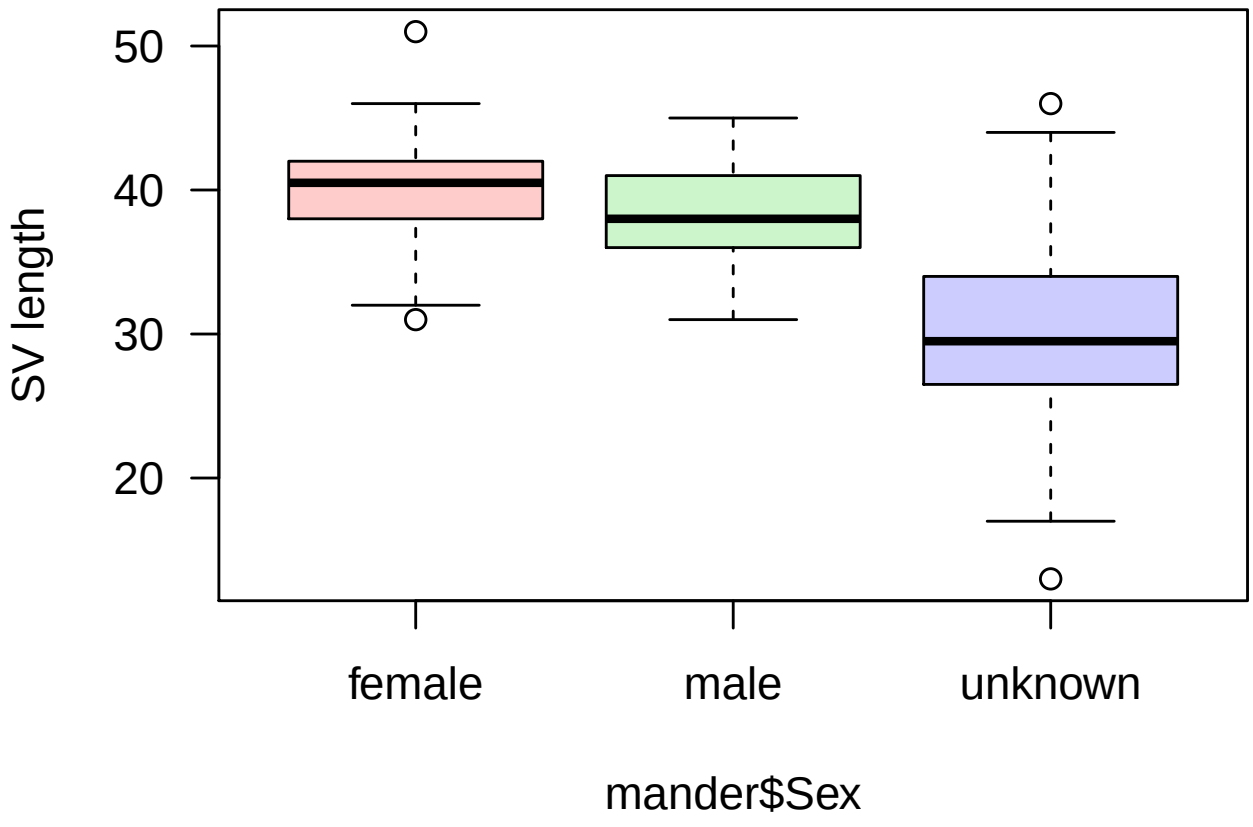
Purpose	Graph Type
Illustrating *distribution*	Histogram, Density plot
	Box(-whisker) plot
Illustrating *differences*	Bar chart, Box plot
Illustrating *correlations*	Scatter plot
Illustrating *associations*	Pie chart, Bar chart
Illustrating *sample size*	Line plot of running avg

Beyond graphs, Towards statistics

- Graphs are powerful tools that provide insight and understanding of the patterns and relationships in the data.
- Graphs alone don't give us the complete answer. We need to **quantify** the relationships we see in our plots.

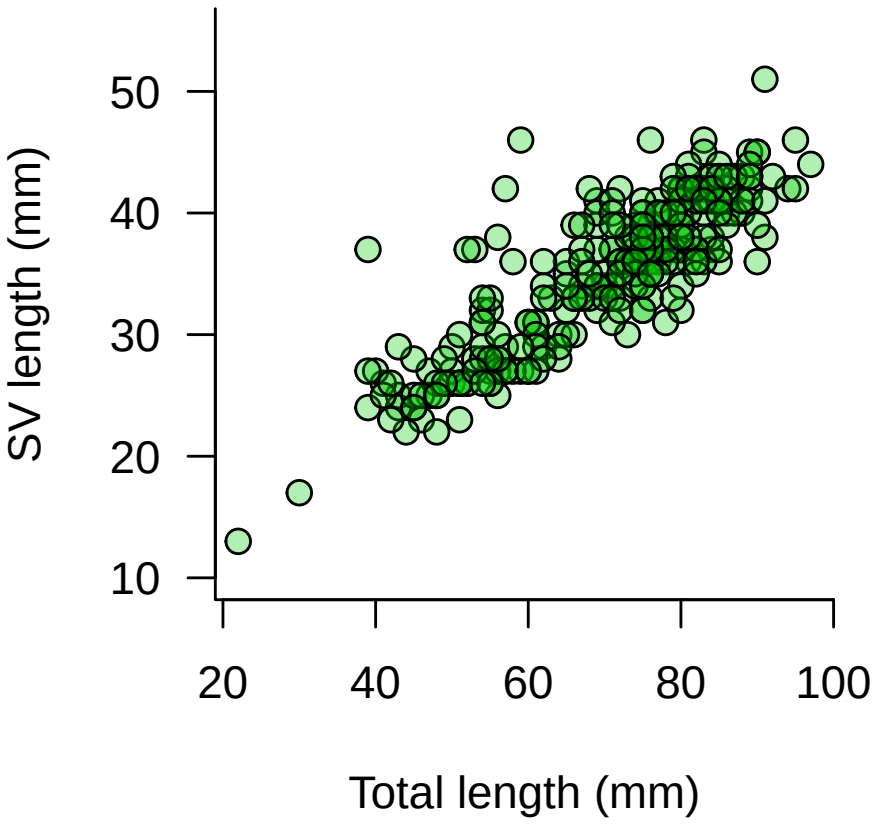
Beyond graphs, Towards statistics

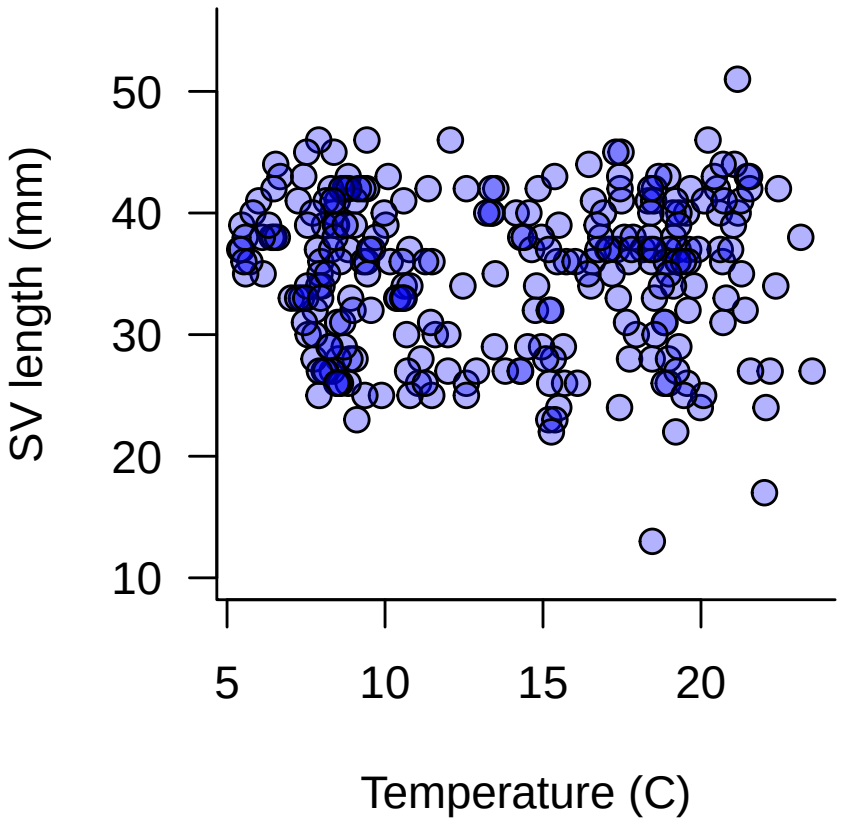
- How can we **quantify** our evidence for relationships?
  - Are differences between groups *significant*?
  - Are differences between groups *meaningful*?



Beyond graphs, Towards statistics

- How can we **quantify** our evidence for relationships?
  - Are associations between two variables *significant*?
  - Are associations between two variables *meaningful*?





Beyond graphs, Towards statistics

- Graphs are powerful tools that provide insight and understanding of the patterns and relationships in the data.
- Graphs alone don't give us the complete answer. We need to **quantify** the relationships we see in our plots.
- Statistics is the tool we use to formally answer these questions:
  - Are differences *significant*?
  - Are associations *significant*?

Group graphical activity

- Plotting the rarefaction data
- See the instructions on Moodle