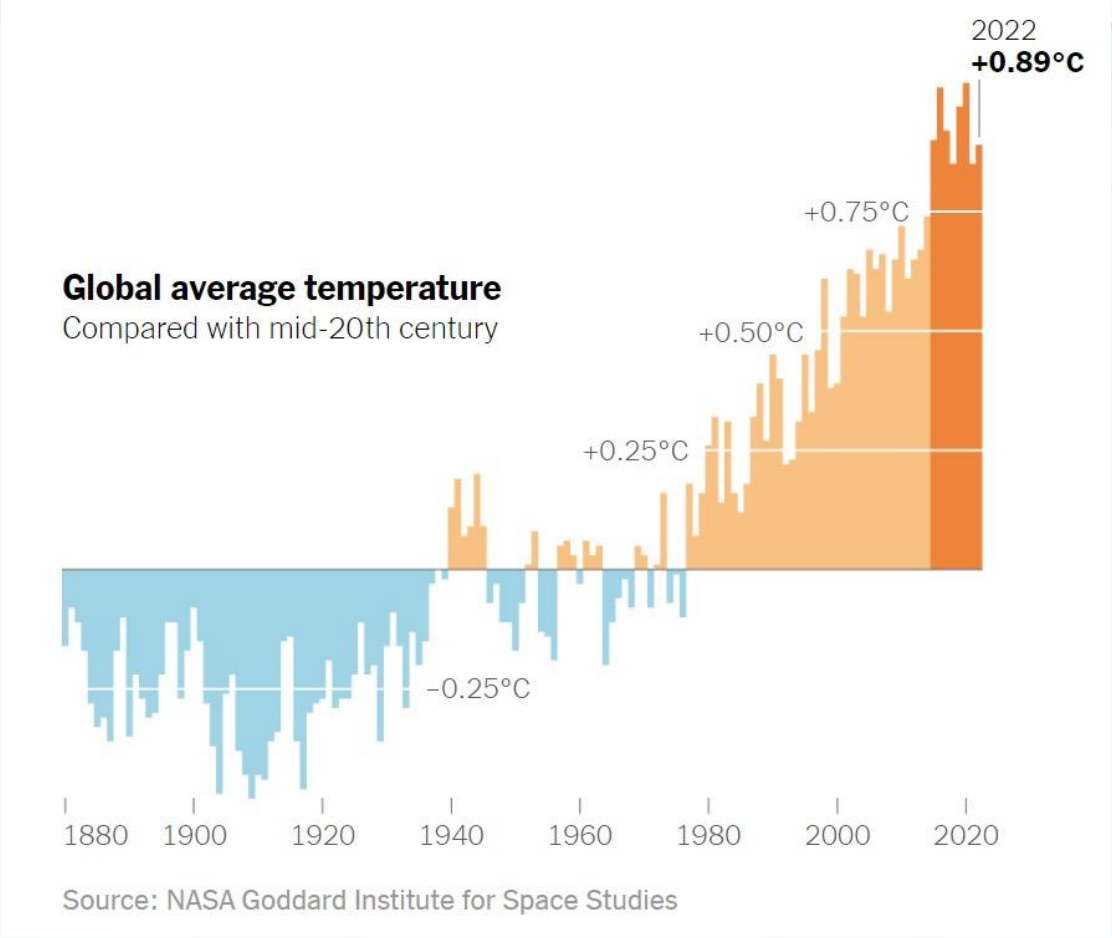
Lecture 4

Summarizing & Visualizing Data Part ??

agenda

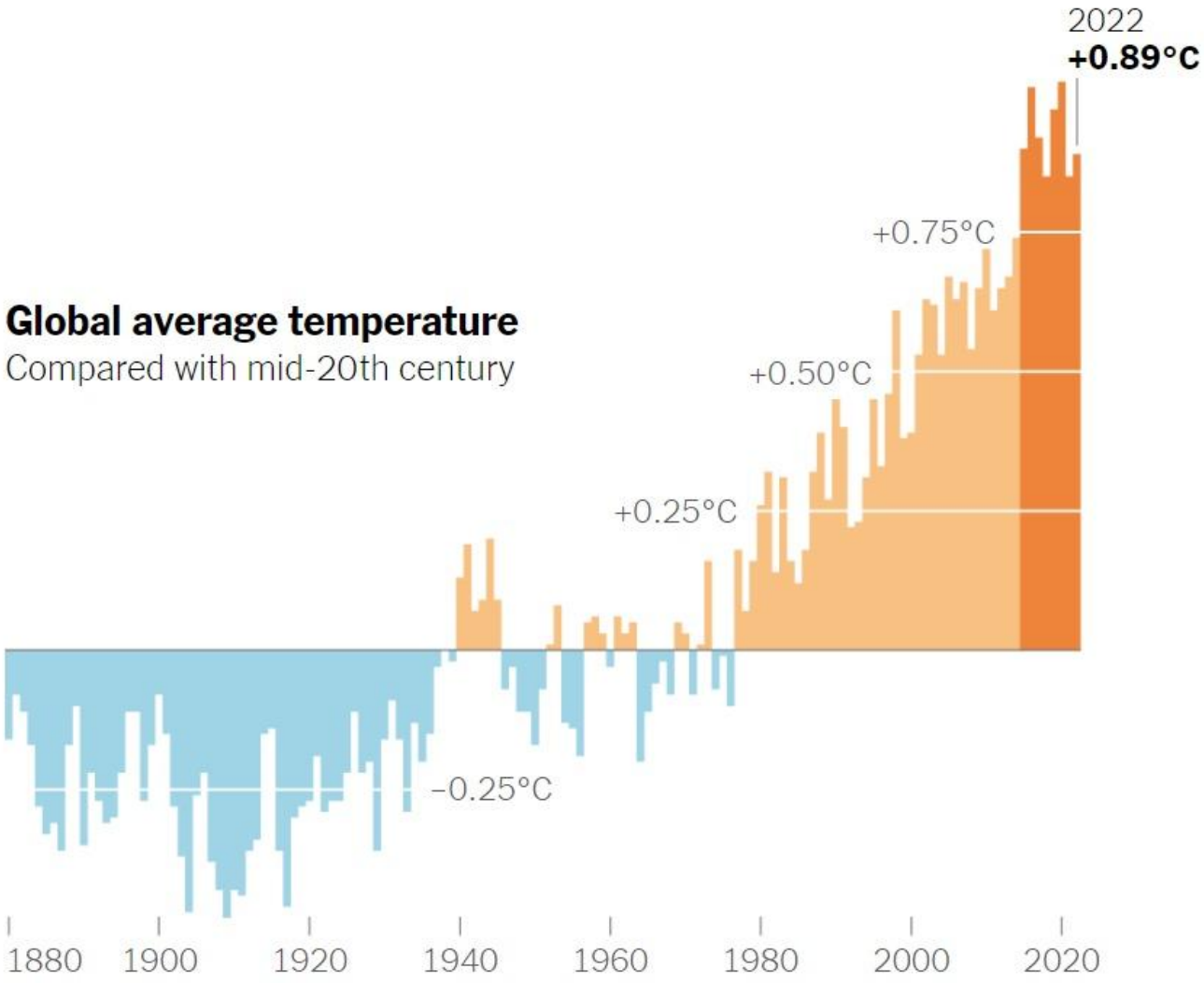
Announcements	HW (extension requests), “ratify” grading guidelines, navigating Canvas
Activity	What’s going on in this graph? NYT
Review	Descriptive statistics for categorical and quantitative variables
R	Code along: statistics, and welcome to the tidyverse
Wrap-up	reminders, sum up

What's going on in this graph?



Get into groups of 2-3 and answer the following questions

1. What do you notice?
2. What do you wonder?
3. How does this relate to you and your community?
4. What's going on in this graph? Create a catchy headline that captures the graph's main idea.



Global average temperature
Compared with mid-20th century

Source: NASA Goddard Institute for Space Studies

Categorical statistics

01

frequency

the number of times an event occurs

02

relative frequency

the number of times an event occurs divided by the total number of outcomes

03

contingency table

matrix or two-way table that displays the frequency of two variables

In R

`summary()` of one or more factor variable

`table()` of one or more factor variables

Contingency tables

transmission

joint ("and")

# of cylinders

	Automatic	Manual	Total
4 cylinder	3	8	**11**
6 cylinder	4	3	**7**
8 cylinder	12	2	**14**
Total	**19**	**13**	**32**

marginals

Plotting a single (factor) categorical variable

plot or barplot

displays frequencies/relative frequencies for categorical variables

```
plot(factor variable, col, border,
     ylim,...)

barplot(table(factor variable),
     col, border, ylim,...)
```

- col = color of bars
- border = color of bar outline
- ylim = c(lower limit, upper limit)

= c(0, 10)

Plotting two (factor) categorical variables

table(data$col1, data$col2)

```
barplot(table(factor variables),
    beside, col, border, ylim,...)
```

side-by-side barplot

displays frequencies/relative frequencies for two or more categorical variables side-by-side

- beside =
  - TRUE: side by side
  - FALSE: (default) stacked
- col = c("color 1", "color 2")
- border = c("color 1", "color 2")
- ylim = c(lower limit, upper limit)

Quantitative statistics

center

01

mean

the "balancing point" of a distribution (arithmetic average)

02

median

the "middle" of a distribution 50% below, 50% above ("Q2")

03

mode

the highest point of a distribution (number that occurs most often)

spread

04

variance

average squared deviations from the mean

05

standard deviation

square root of variance

06

range

distance between the minimum and maximum value of a set of numbers

Calculating the 3 m's

mean

add up all the numbers and divide by the total

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\mu = \frac{x_1 + \cdots + x_N}{}$$

median

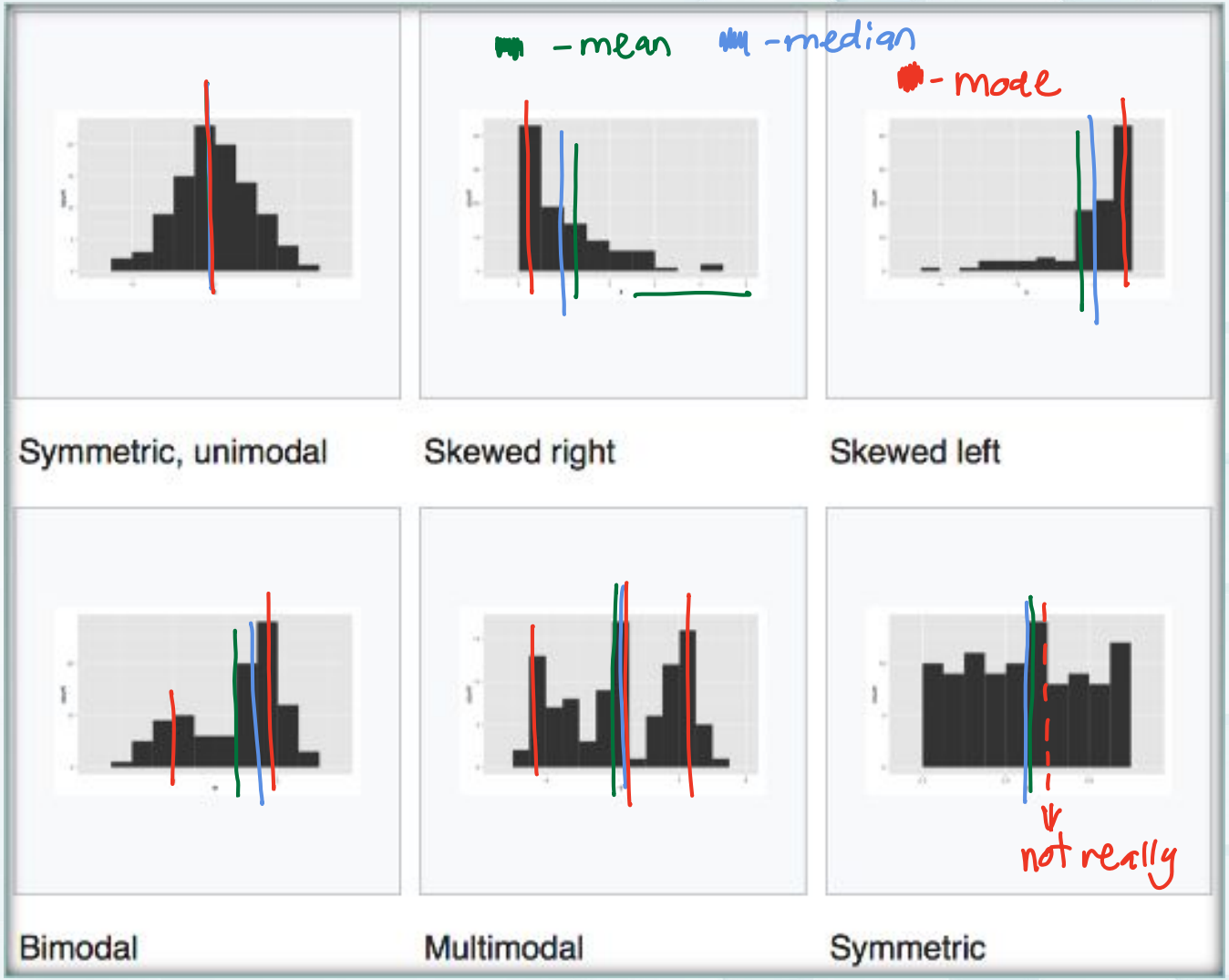
order data values from smallest to largest and pick the number in the middle (if there are two numbers, take the average)

mode

See which value occurs the most often or what is the highest point in the distribution

Measures of central tendency

Identify where the **mean**, **median**, and **mode** could be on these plots



What happens to the mean if we add or multiply every number by a constant?

$$x_1, x_2, x_3, x_4$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

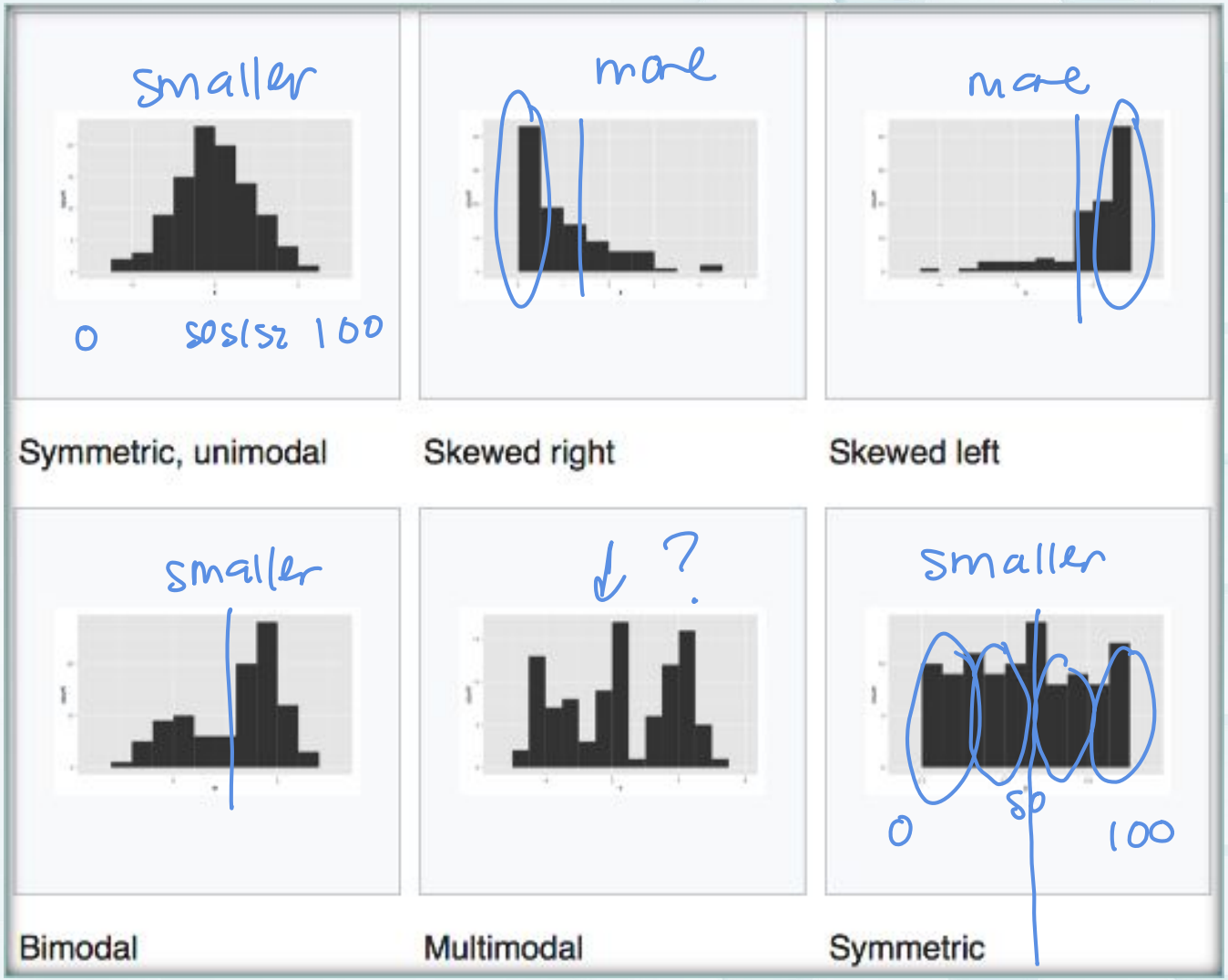
$$2\bar{x} = \frac{2x_1 + 2x_2 + 2x_3 + 2x_4}{4}$$

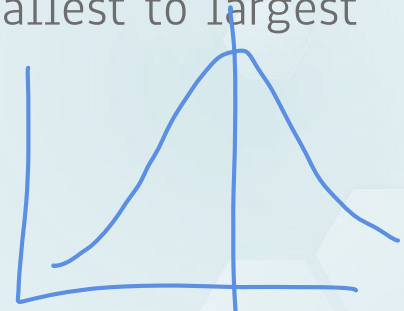
$$+3 \rightarrow \bar{x} + 3$$

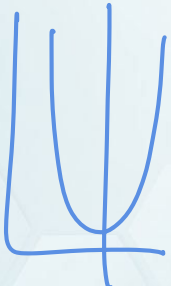
$$s^2 = \frac{\sum_{i=1} (x_i - \bar{x})^2}{n-1}$$

Measures of spread

Compare the variances of these plots and rank from smallest to largest







What happens to the sd/var if we add or multiply every number by a constant?

s^2 :

$+c$ $\xrightarrow{\text{shift}}$ changes nothing

$\times c$ stretching/shrink changes by c^2

$(2x_i - \bar{x})^2 \rightarrow 4(x_i - \bar{x})^2$

Measures of position

Percentiles (Quantiles)

Identifies the percent of observations *below* a certain value, e.g., 80th percentile in height

Q_1 & Q_3

First and third **quartiles**

25th and 75th percentiles

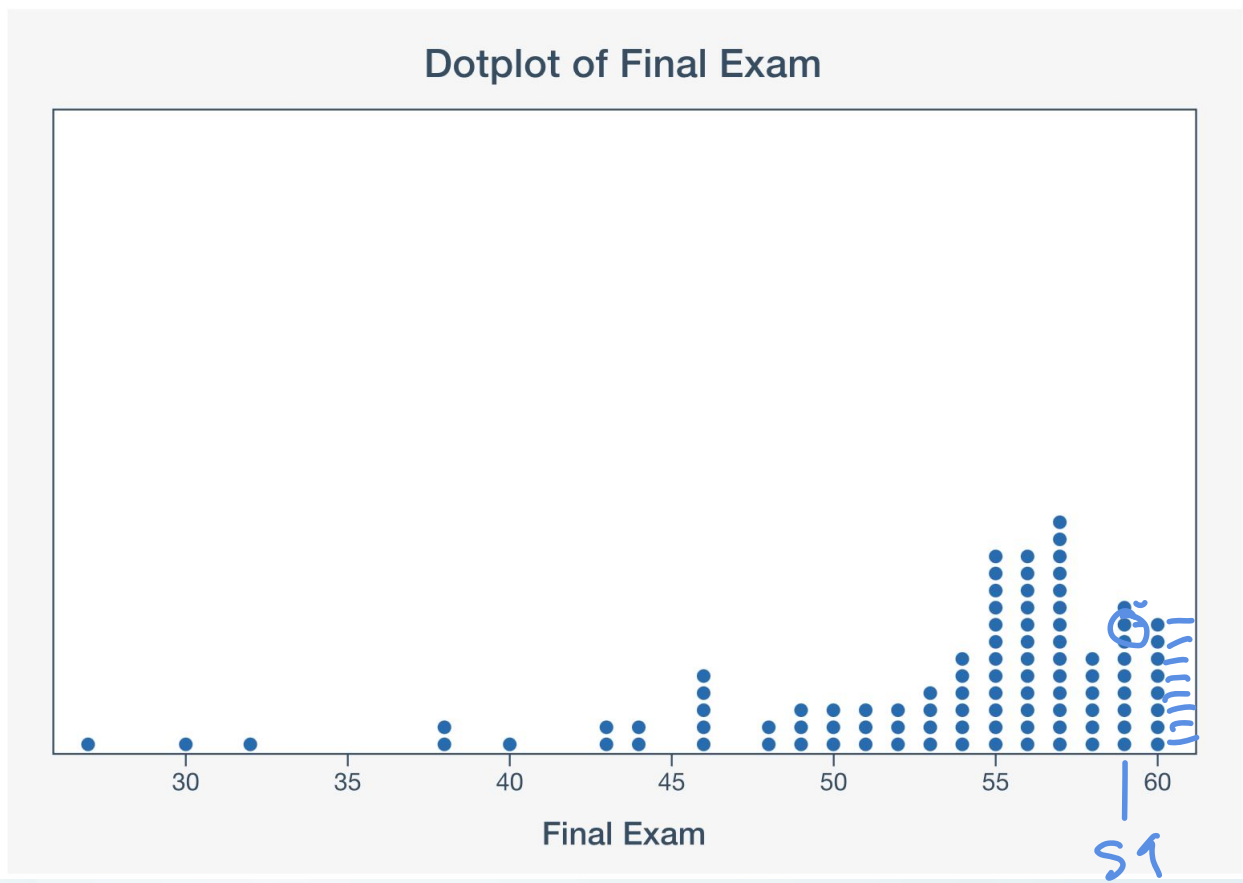
IQR

Interquartile Range

= Q3 - Q1

middle 50%

Percentiles



There are 100 dots on this plot, the 90th percentile will fall around the 90th and 91st points. The point that is 10th from the top is 59. Thus, the 90th percentile in this sample is a score of 59 points.

Measures of position in R

Percentiles (Quantiles)

```
quantile(x, probs = [#])
```

- x = set of numbers
- probs = one or more numbers between 0 and 1
  - e.g., 0.80 or c(0.2, 0.3, 0.97)

Q_1 & Q_3

```
Q1 = quantile(x, probs = 0.25)
Q2 = quantile(x, probs = 0.75)
```

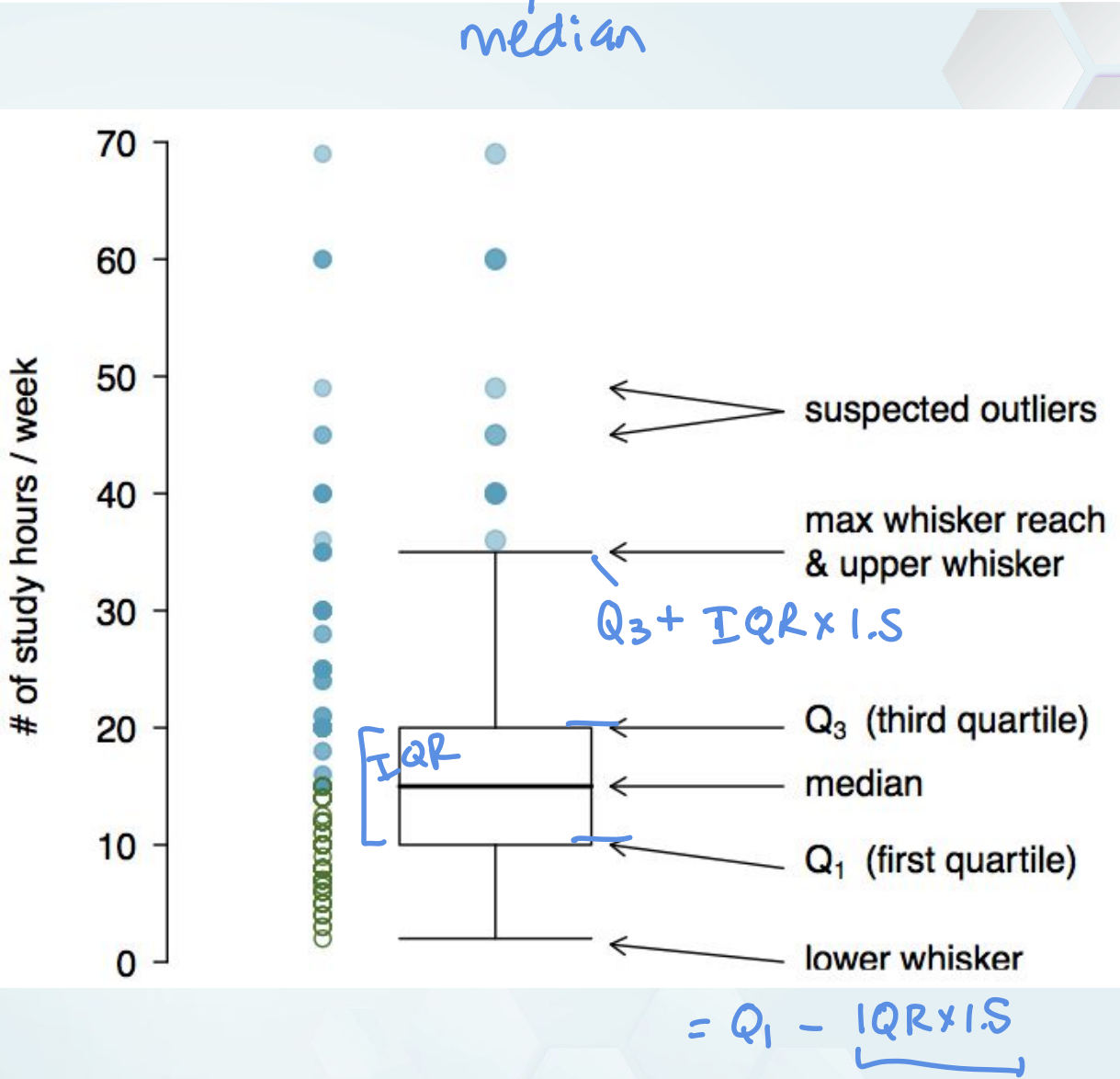
OR

```
quantile(x)
```

quantile() by itself computes 0%, 25%, 50%, 75%, 100%

Box Plots

Percentage of time spent taking notes versus doing activities in class



Plotting a single (numeric) quantitative variable

box plot

A plot which summarizes a dataset using five statistics (1st and 3rd quartiles, median, whiskers, and outliers)

- 1st and 3rd quartiles (Q1 and Q3): 25th and 75th percentiles
- Interquartile range (IQR): |Q3-Q1|
- Whiskers: Q1-1.5xIQR and Q3+1.5xIQR
- Outliers: extreme observations (beyond whiskers)

Plotting a single (numeric) quantitative variable

box plot

A plot which summarizes a dataset using five statistics (1st and 3rd quartiles, median, whiskers, and outliers)

```
boxplot(x, col, border, plot,
    horizontal, ...)
```

- x = quantitative variable
- plot =
  - TRUE: (default) boxplot is produced
  - FALSE: summary statistics are produced
- horizontal:
  - TRUE: boxplot is horizontal
  - FALSE: (default) boxplot is vertical

Plotting a single (numeric) quantitative variable

histogram

graphical representation of the distribution of numerical data

Describing Histograms:

1) Shape: patterns described in previous slide
2) Center: mean, median, mode
3) Spread: sample sd/var, range

Plotting a single (numeric) quantitative variable

```
hist(x, breaks, freq, col,
  border, xlim, ylim,...)
```

histogram

graphical representation of the distribution of numerical data

- breaks = # of bars
- freq =
  - TRUE: (default) frequency plot
  - FALSE: relative frequency plot

Let's practice!

Download recent_grads.txt from Week 3 Page on Canvas
Open RStudio