

CS 397F
Introduction to Data Science

Week 2: Statistical Inference

Gordon Anderson

What can we infer from a set of observations?

- Infer:

“To infer is to make a well informed guess — if you see your mom’s bag on the table, you might infer that she’s home”. (vocabulary.com)

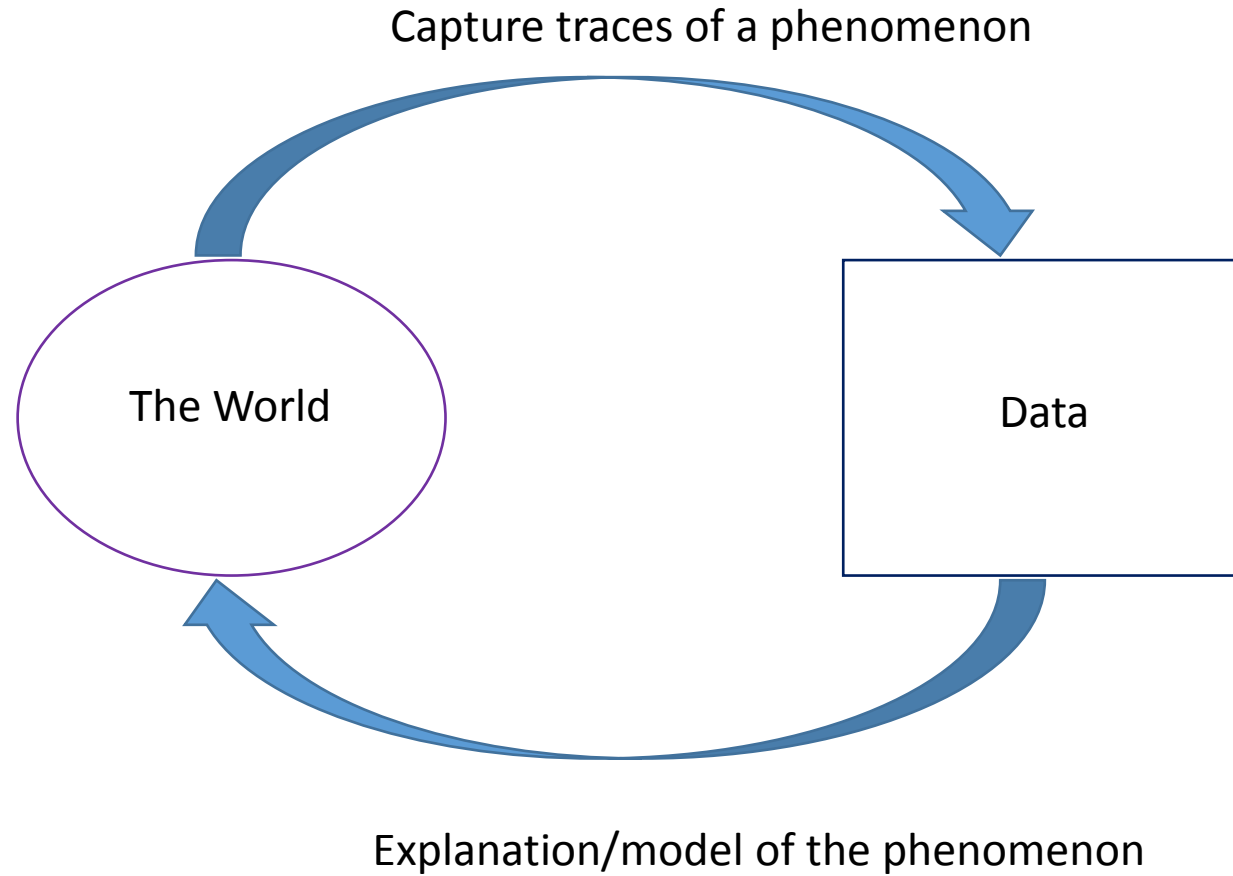
- Why do we have to infer/guess anything?

Because the world is stochastic (it varies).

- Why can’t we make a calculation and make a definite conclusion?

Because of variance in the world and bias in our methods.

Statistical inference



The Process:

- Given: the question you want answered has been defined.

Do:

- Step 1: make observations- collect data.
- Step 2: analysis- make models, evaluate them.
- Step 3: draw conclusion, ask: has the question been answered?
- Step 4: communicate results- graphs, charts.

Note that there often will be a need to go back to a previous step(s)!

Example 1

- What is the height of students in CS397F?
- Let's say we measure these heights (in cm):
167, 170, 155, 186, 160, 163, 158, 157, 166

The heights vary (of course). Therefore, we need statistics to help describe these observations.

Wait just a minute!

- There are 35 students in CS379F, and we only measured 9.
- The population is 35, we sampled 9 from that population.
- Our sample should represent the population, but it might not.
- Our sample is probably *biased*. (we'll talk about unbiased estimators)
- We could perform our sample many times- drawing subjects at random.
- Also- what about other semesters of CS397F? Isn't that the real population?
- Remember: data is not objective!

And now a brief statistical interlude...

Descriptive stats:

- Given these data (our sample):

167, 170, 155, 186, 160, 163, 158, 157, 166

Calculate descriptive statistics:

min 155

max 186

range 31

mean 164.7

median 163

What do these statistics say about the data?

Descriptive stats:

The data sorted in ascending order:

155, 157, 158, 160, 163, 166, 167, 170, 189

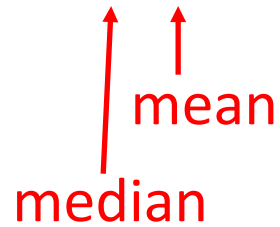


Diagram illustrating the mean and median of the sorted data. The data values are 155, 157, 158, 160, 163, 166, 167, 170, 189. The mean is indicated by a red arrow pointing to the value 166, and the median is indicated by a red arrow pointing to the value 163.

min 155

max 186

range 31

mean 164.7

median 163

Descriptive stats:

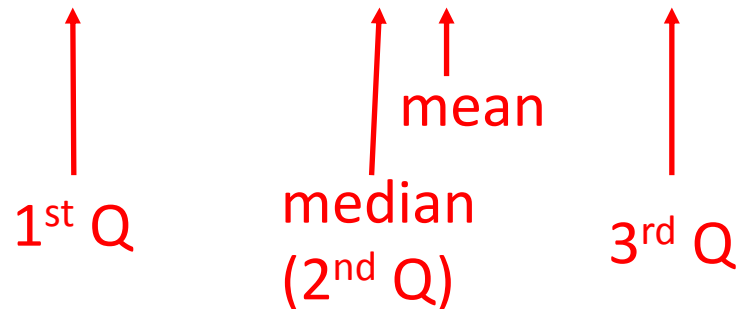
min 155, max 186, range 31, mean 164.7, median 163

In addition to the above, we want to know about the *distribution* of the heights- how they are “scattered” over their range.

We calculate the *quartiles* to get a sense of how many data points lie in the bottom 25%, middle 50%, top 25%.

```
> heights <- c(167, 170, 155, 186, 160, 163, 158, 157, 166)
> summary(heights)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
155.0  158.0  163.0  164.7  167.0  186.0
```

155, 157, 158, 160, 163, 166, 167, 170, 189

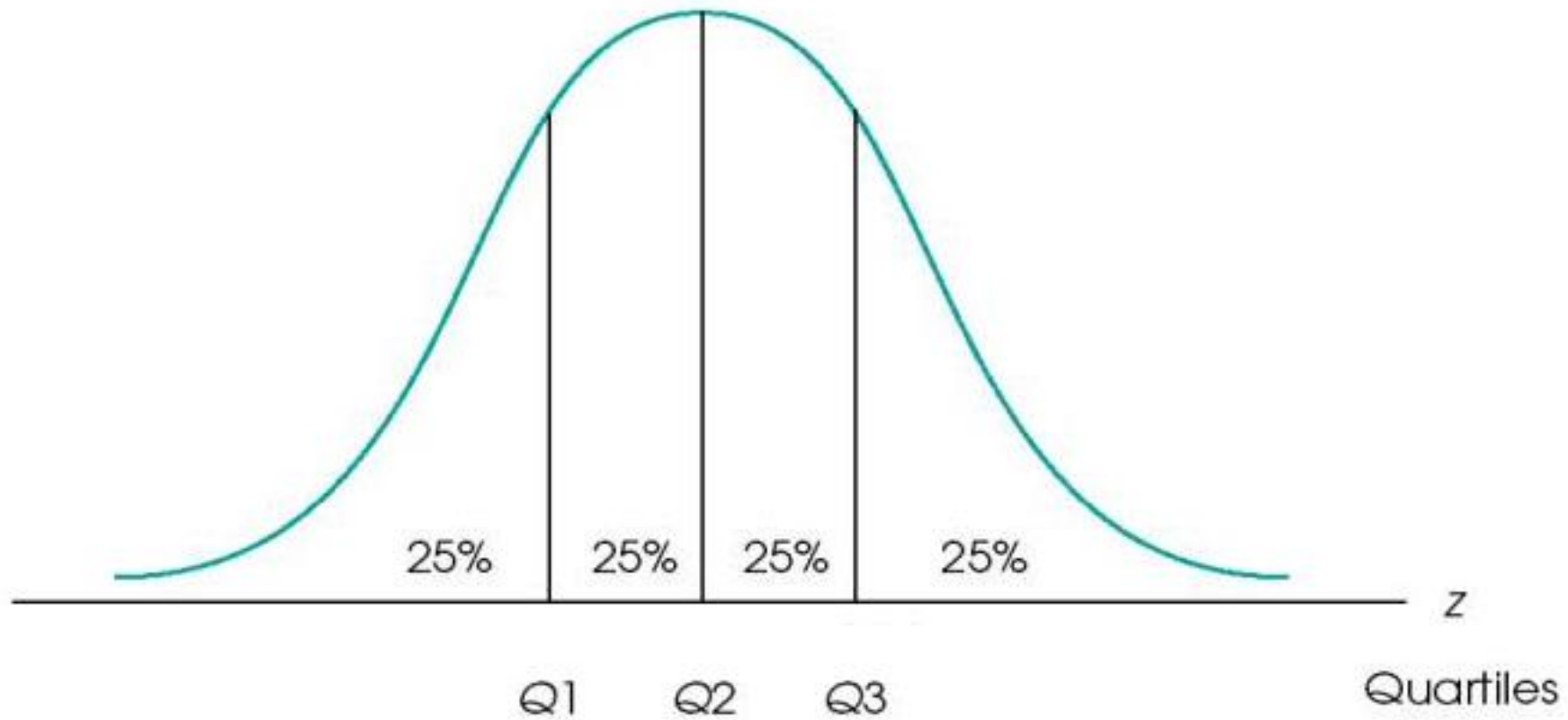

 A diagram showing the distribution of heights. The sorted list of heights is 155, 157, 158, 160, 163, 166, 167, 170, 189. Red arrows point from labels below to specific values in the list: '1st Q' points to 158, 'median (2nd Q)' points to 163, 'mean' points to 164.7 (between 163 and 166), and '3rd Q' points to 166.

|-----25%-----|-----50%-----|-----25%-----|

The percentage of heights within each quartile.

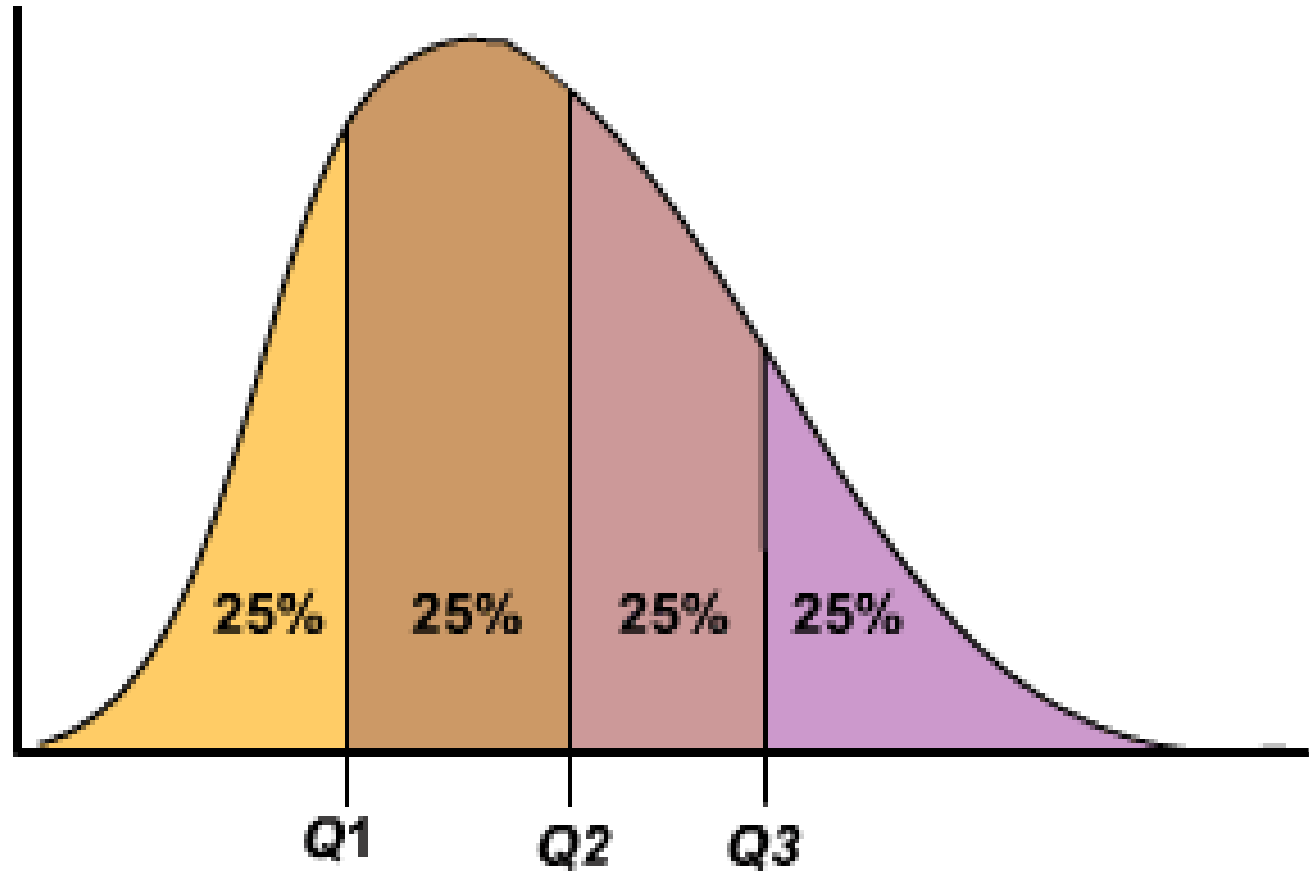
The middle two quartiles are called the interquartile range-IQR

Quartiles in general:



Quartiles:

A “skewed”
distribution.



Variance and Standard Deviation

Variance is a measure of how widely dispersed the data points are relative to the distribution *mean*.

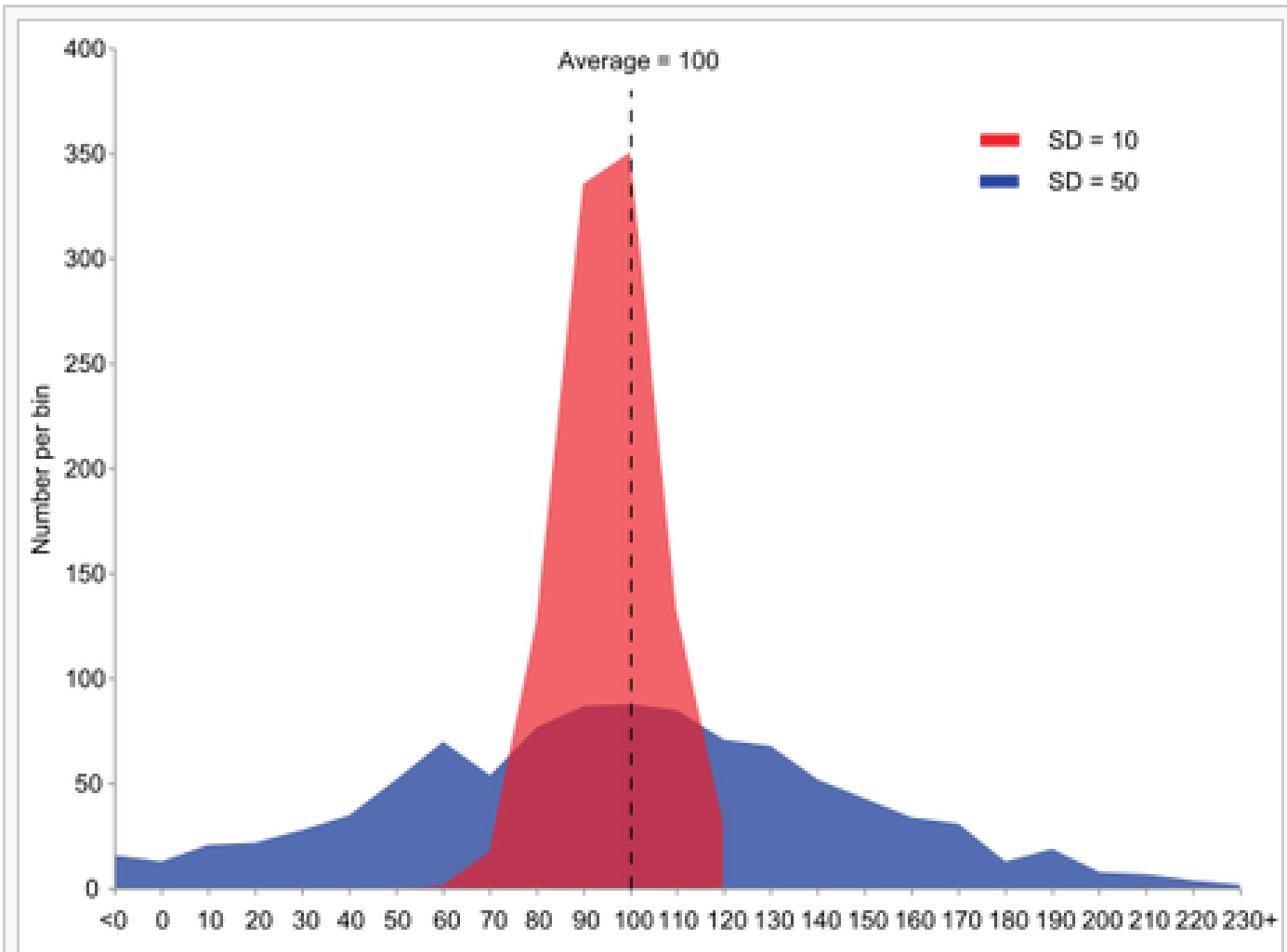
Standard Deviation: also a measure of dispersal of data, based on the variance of the data. SD is in the same units as the data.

Notation: SD, or σ , or sometimes s .

(Note that the SD is the square root of the variance, so variance is σ^2 in many texts).

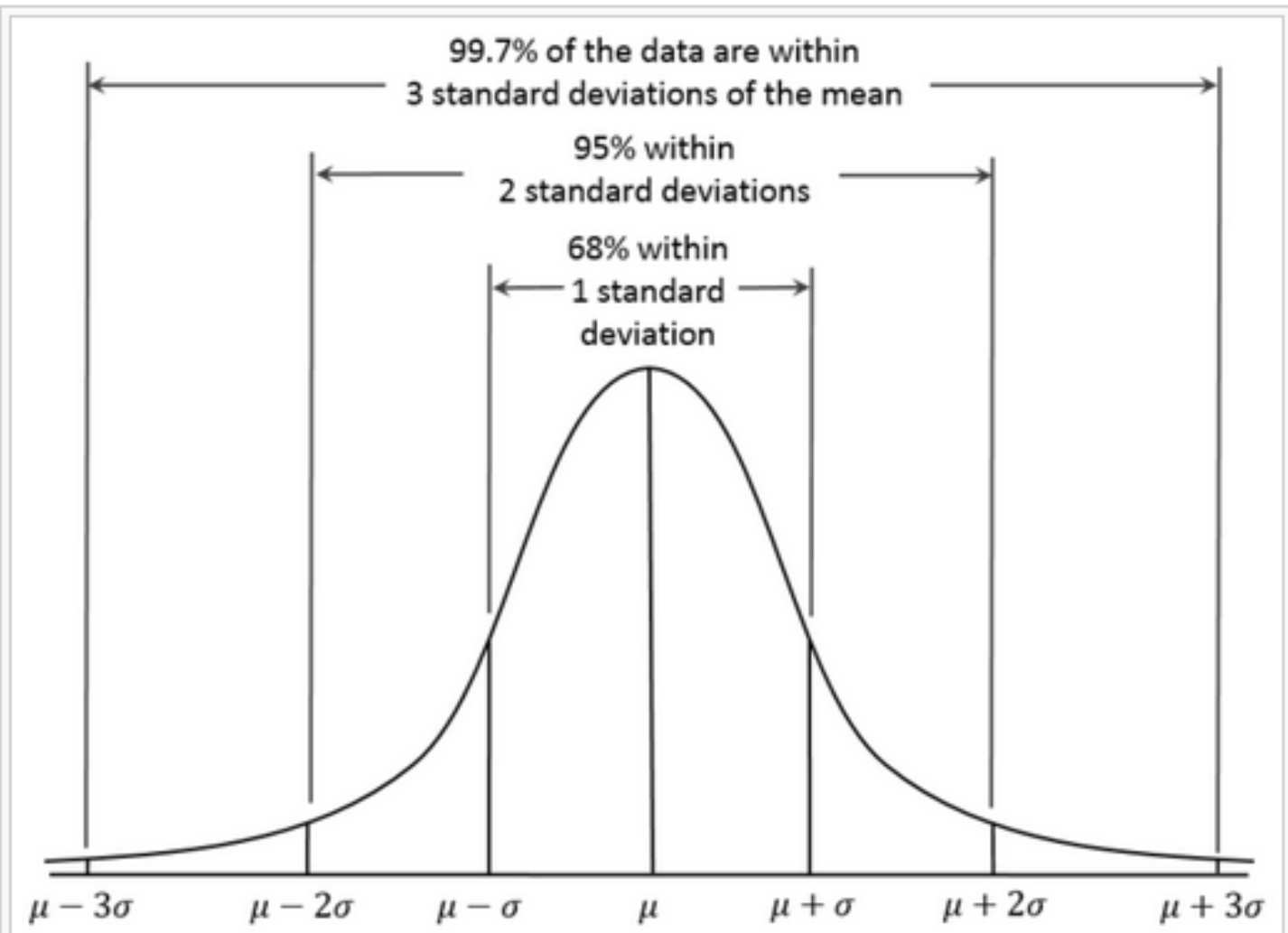
Distributions and variance.
(from Wikipedia).


What would you infer about
these two distributions
if they were heights?



Example of samples from two populations with the same mean but different standard deviations. Red population has mean 100 and SD 10; blue population has mean 100 and SD 50.

SD in general:
(Wikipedia)



For the [normal distribution](#), the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%. 

Variance and Standard Deviation

Our height data:

```
> heights <- c(167, 170, 155, 186, 160, 163, 158, 157, 166)
> summary(heights)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 155.0   158.0   163.0   164.7   167.0   186.0
> sd(heights)
[1] 9.433981
```

We can say that 68% of the students are within what range?

mean – SD, mean + SD, so: 155 and 174 cm

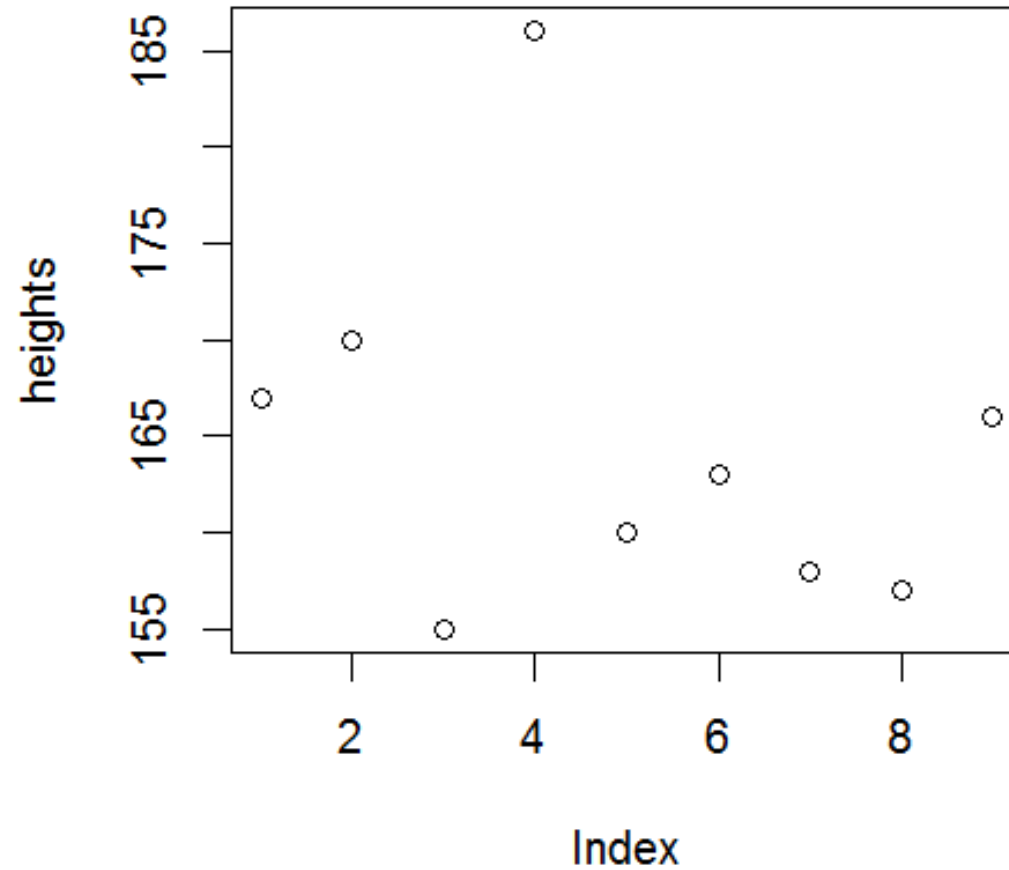
Calculating $\pm 1SD$ (68%) in R code:

```
> mean <- mean(heights)
> SD<-sd(heights)
> low<- mean-SD
> hi<-mean+SD
> low
[1] 155.2327
> hi
[1] 174.1006
```

Viewing the data

```
> plot(heights)
```

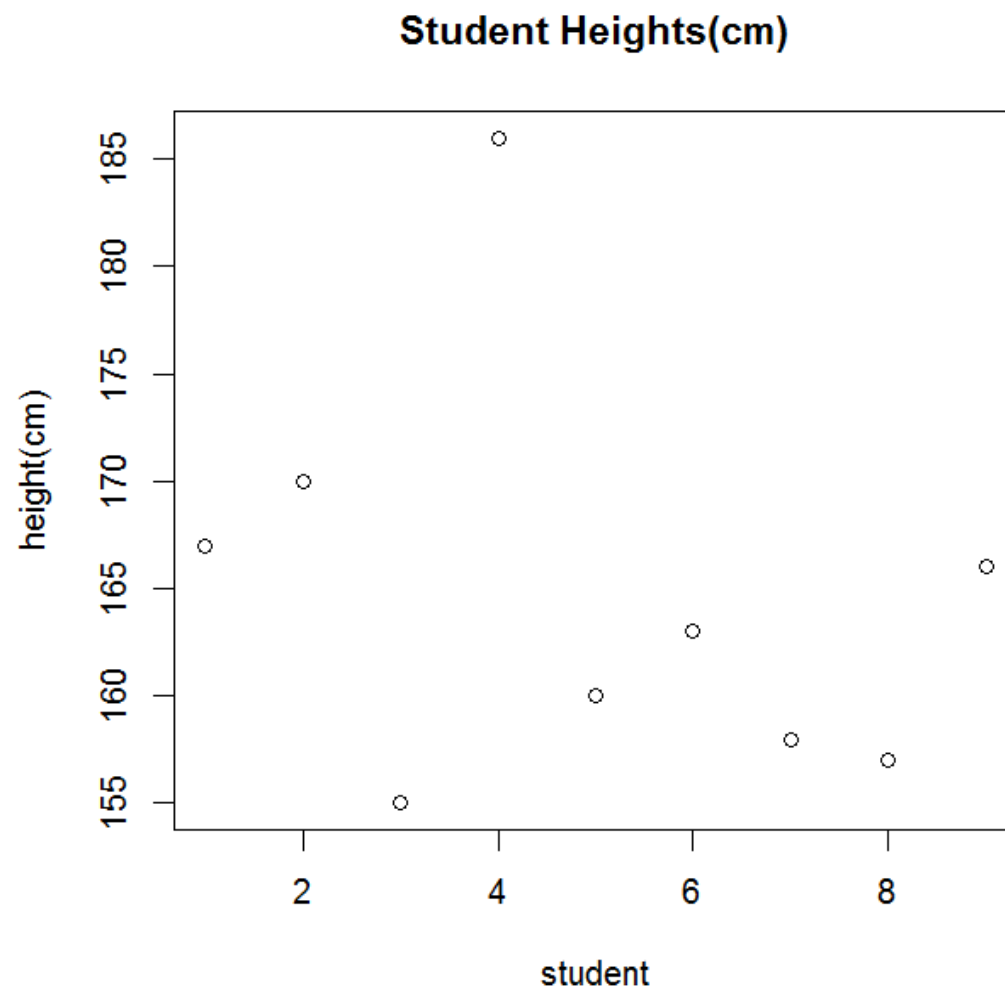
Basic scatter plot.



Viewing the data

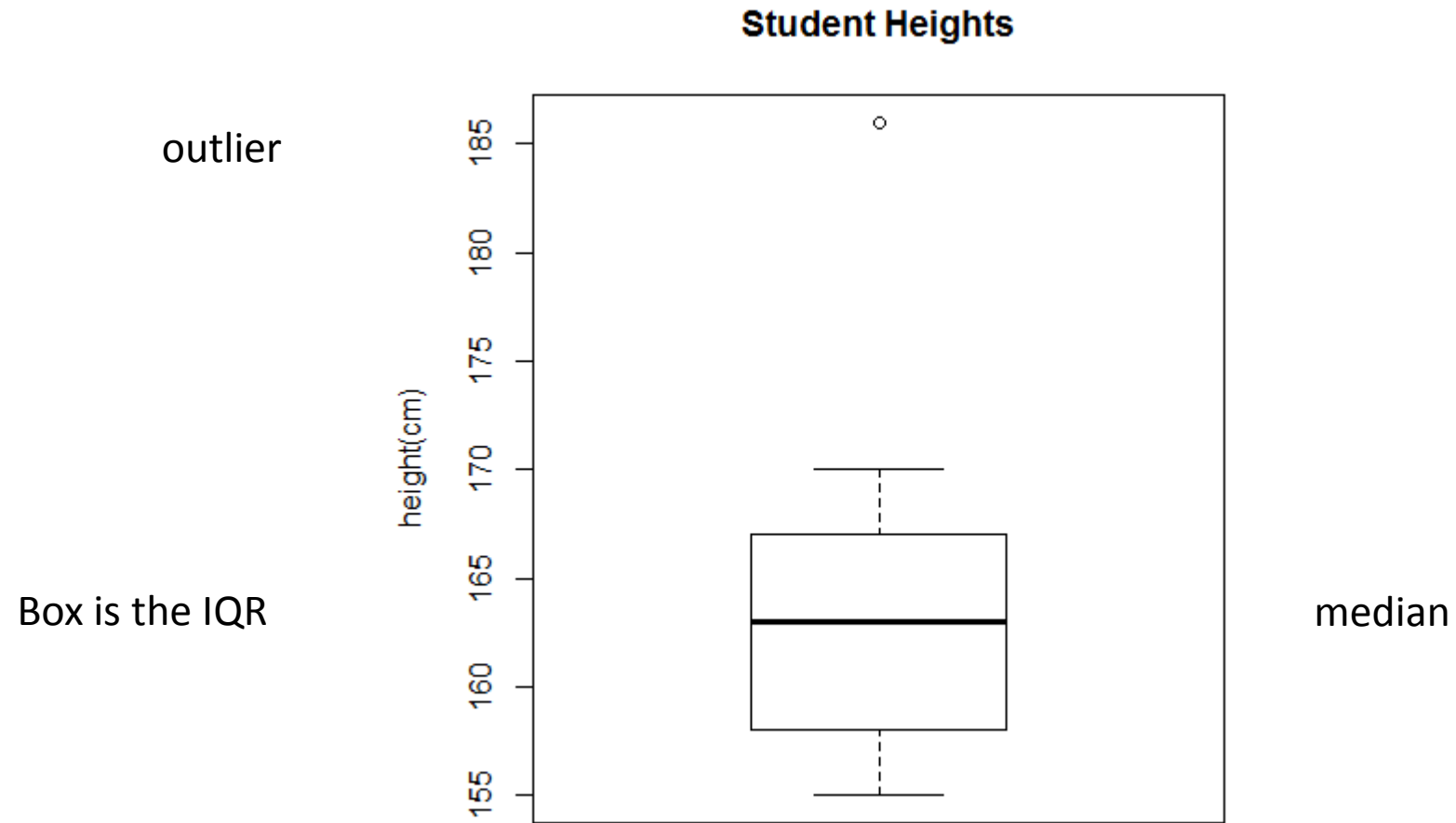
```
> plot(heights, main="Student Heights(cm)", ylab="height(cm)", xlab="student")
```

Add a title and axis labels.



Boxplot

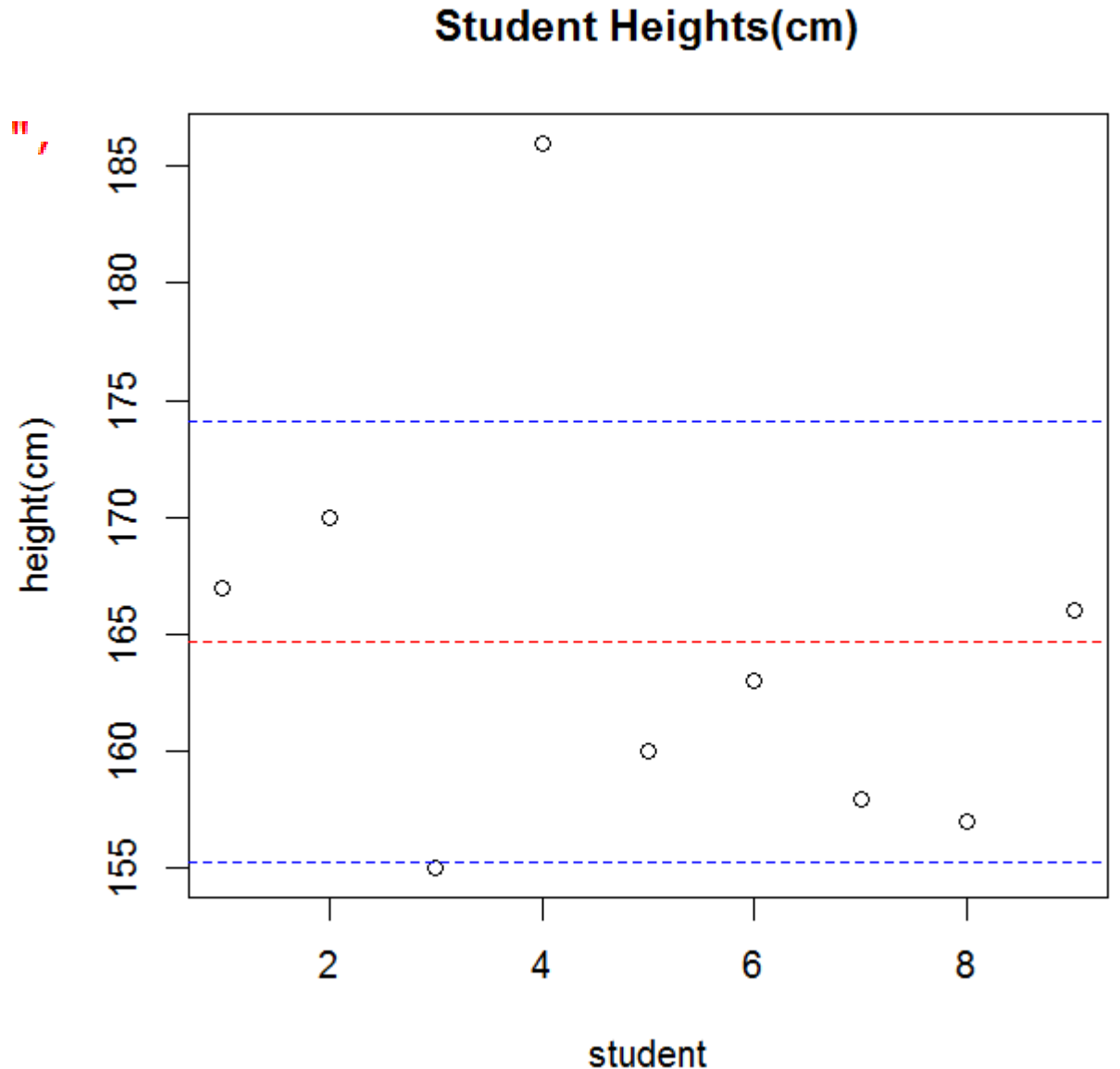
```
boxplot(heights, main="Student Heights", ylab="height(cm)")
```



Viewing the data

```
> plot(heights, main="Student Heights(cm)",  
+ ylab="height(cm)", xlab="student")  
> abline(h=mean,col=2,lty=2)  
> abline(h=hi,col=4,lty=2)  
> abline(h=low,col=4,lty=2)
```

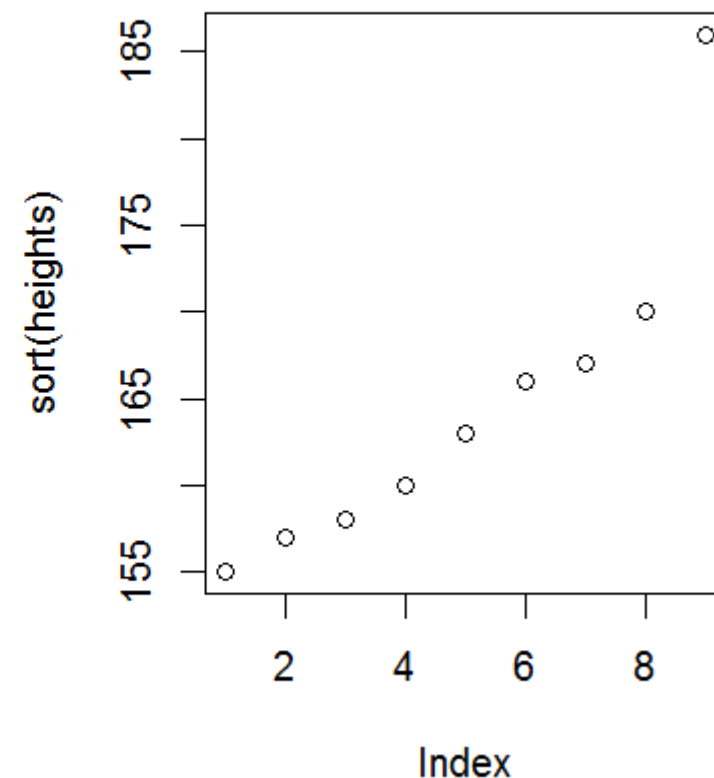
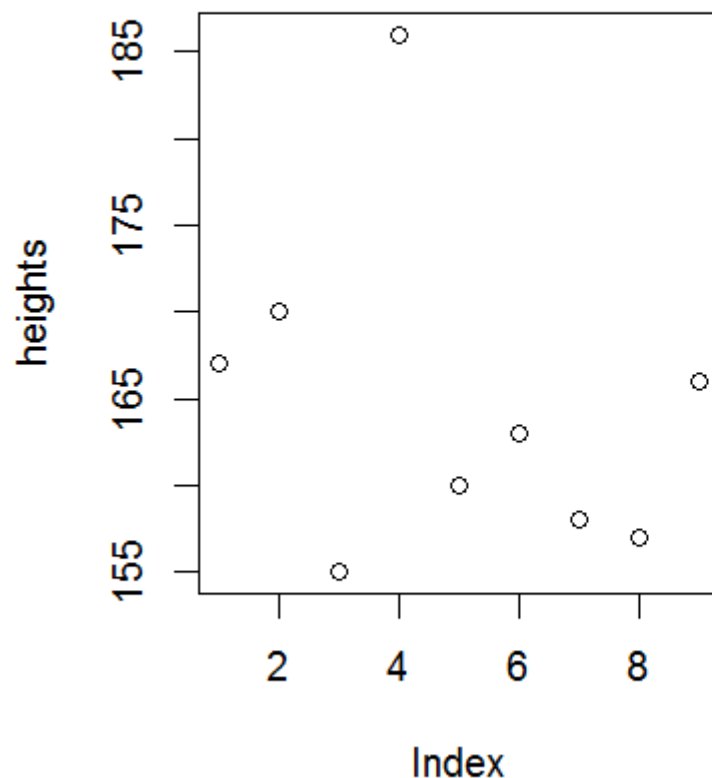
Add lines for mean and mean \pm SD. (remember that we calculated hi and low previously).



Viewing the data

```
> #scatter plots  
> par(mfrow=c(1,2))  
> #unsorted  
> plot(heights)  
> #sorted  
> plot(sort(heights))
```

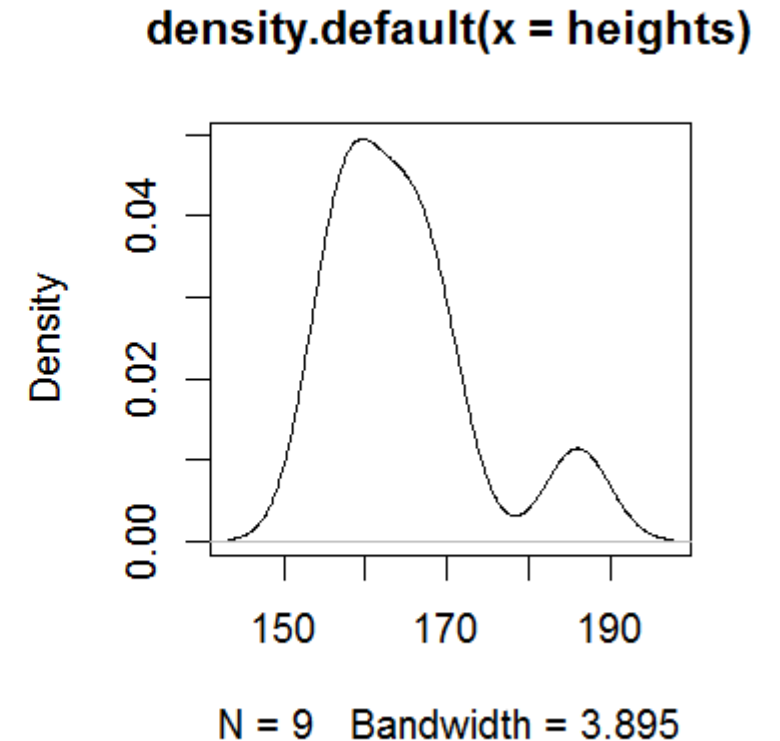
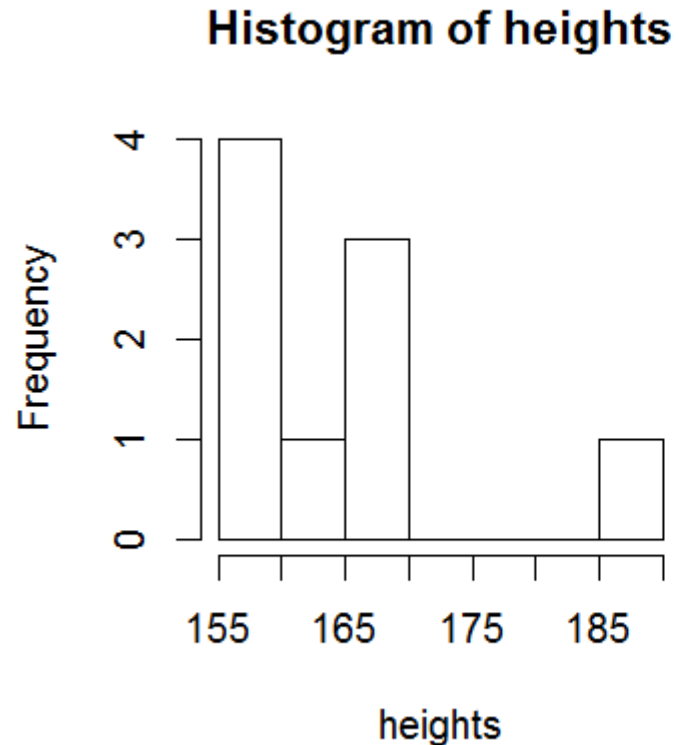
Multiple plots.



Viewing the data

```
> #density and histogram  
> par(mfrow=c(1,2))  
> #histogram  
> hist(heights)  
> #density  
> plot(density(heights))
```

Density and histogram.



Note: The bandwidth is a measure of how closely you want the density to match the distribution.

We'll talk more about this when we discuss model fitting and smoothing.

For more on density/bandwidth:

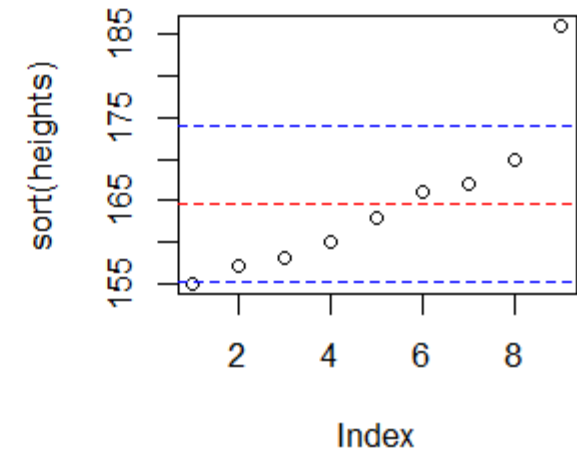
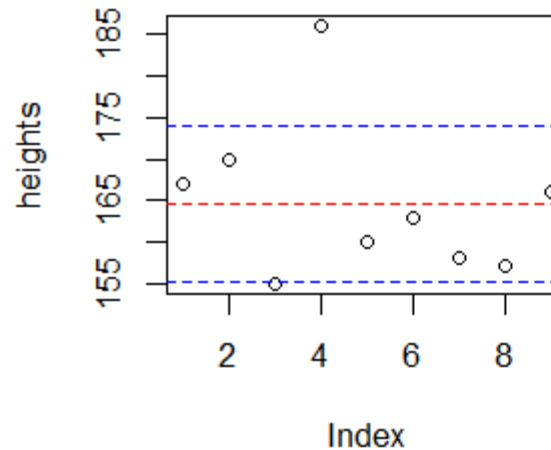
https://en.wikipedia.org/wiki/Kernel_density_estimation

Well, that was more than a brief interlude,
including some data visualization.

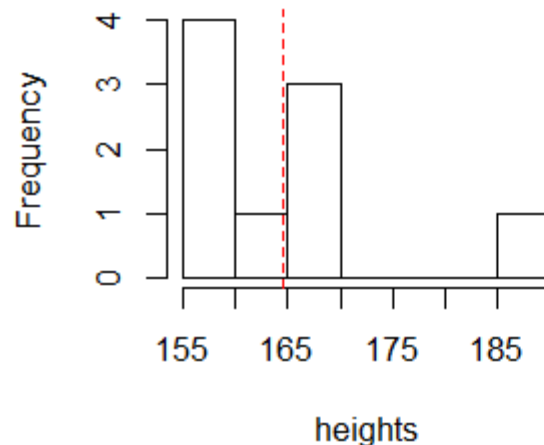
But, we just did some exploratory data
analysis, or EDA!

What did our EDA tell us about heights?

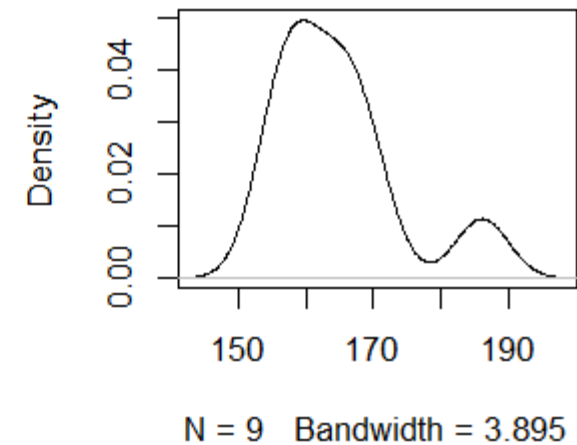
```
> par(mfrow=c(2,2))
> plot(heights)
> abline(h=mean,col=2,lty=2)
> abline(h=hi,col=4,lty=2)
> abline(h=low,col=4,lty=2)
> plot(sort(heights))
> abline(h=mean,col=2,lty=2)
> abline(h=hi,col=4,lty=2)
> abline(h=low,col=4,lty=2)
> hist(heights)
> abline(v=mean,col=2,lty=2)
> plot(density(heights))
```



Histogram of heights



density.default(x = heights)



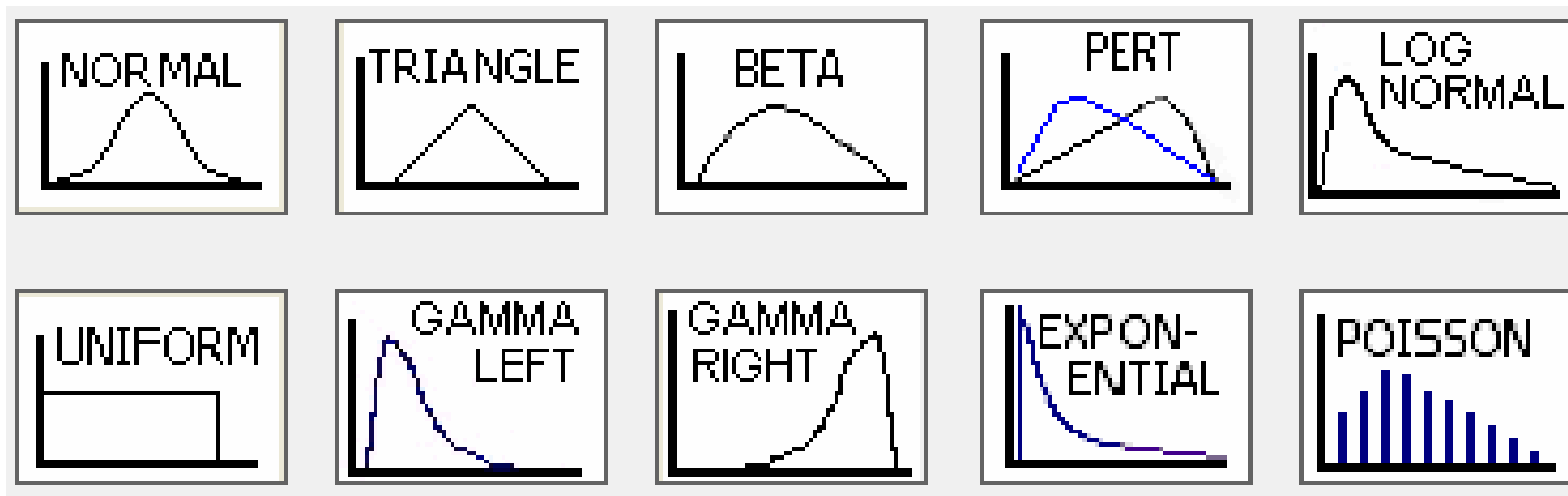
Didn't really have a specific question.
A very small sample (vs a population).
Still, we do know something.

Now, about distributions:

- The data (any data) has a distribution. You can plot frequency vs values to see it, via histogram or density plot.
- We would like to model the data's distribution so we can make predictions, etc.
- We want to model the data's distribution with a *probability* distribution.
- Then, we can state what the probability of any value occurring is!

“Standard” distributions.

- The world tends to generate data distributions that resemble some common forms:

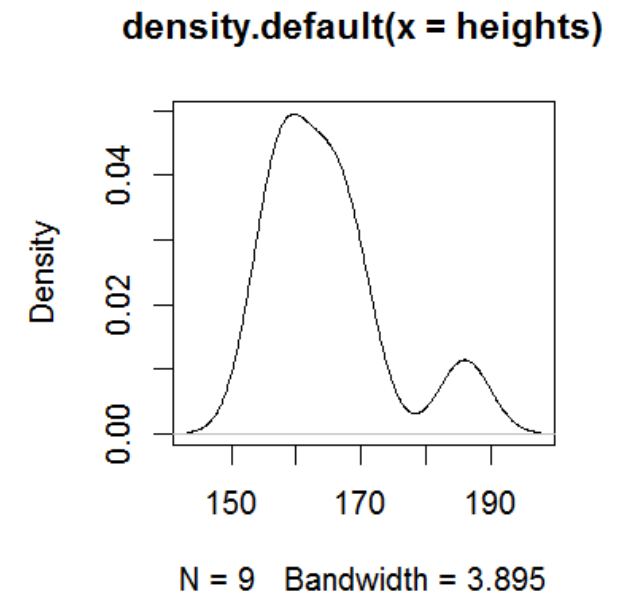
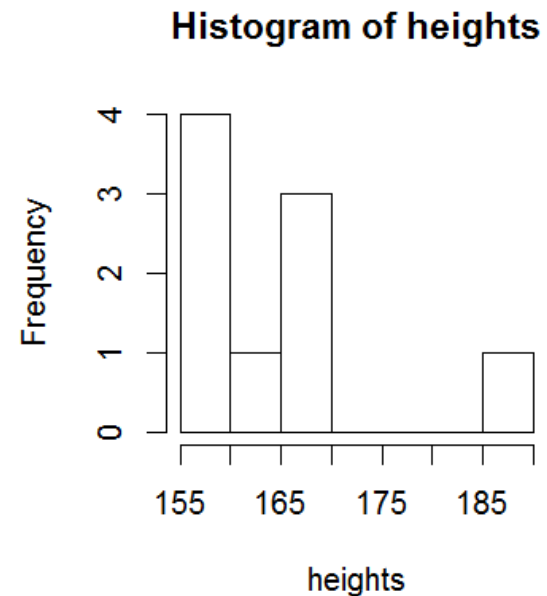
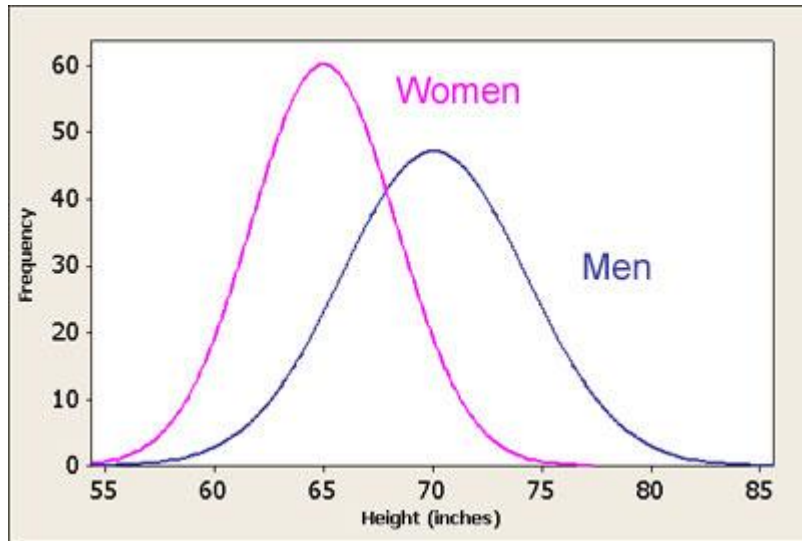


Random variable

- A random variable, RV, can take on a set of values with a certain probability.
- Let X be an RV. Then, there is some function p that maps X to a real value- a probability in $[0,1]$.
- So, $p(X=x_i)$ is the probability that X takes on some value x_i .
- Often, we assume the data conforms to one of the “standard” distributions. We can then make inferences about the data and its underlying, generative process.

Our height example

- Our RV was height in centimeters.
- We didn't fit a model or choose a distribution, but we could have.



A typical height distribution vs our data. We would guess a normal dist.

Modeling

Modeling is not an exact science- we choose models based on previous experience and by trial and error:

pick model, fit, evaluate fit, repeat

There are tons of modeling techniques, and we will be learning and practicing them all semester.

Always be thinking about what biases you are dealing with. In our example, we have such a small sample that our results would be uncertain.

More inference...

First, a review of the inference process:

- Given: the question you want answered has been defined.

Do:

- Step 1: make observations- collect data.
- Step 2: analysis- make models, evaluate them.
- Step 3: draw conclusion, ask: has the question been answered?
- Step 4: communicate results- graphs, charts.

Note that there often will be a need to go back to a previous step(s)!

Example 2:

- How many people use the walking path next to my house?
- Step 1: make observations.
 - Must choose what and how to observe (therefore introducing bias).
 - In DS, a choice means you are making an assumption.
 - Data scientist: are the assumptions reasonable/defendable?
 - Consider possible errors in the data gathering technique you chose.

What/how would you implement step1?

Example 2

- How many people use the walking path next to my house?
- Step 2: analysis.
 - What algorithms to use?
 - What assumptions am I making by choosing specific algorithms?
 - As before, choosing means assumptions and therefore bias.
 - The goal is to make a generalization about the specific observations- discover a pattern in the data that represents the underlying process of interest.

How could we model these data?

Example 2

- How many people use the walking path next to my house?
- Step 3: draw conclusion, ask: has the question been answered?

What do we conclude about the number of people using the path?

Does the conclusion answer the question?

Did we discover any unexpected results/aspects of the data?

Usually more questions are created!

Is the question a good one in the first place?

Example 2

- How many people use the walking path next to my house?
- Step 4: communicate results- graphs, charts.

Who is our audience?

Technical/non-technical, a mixture?

Usually we will create several of each kind of view of the results:

tables

charts and graphs