جامعة نيويورك أبوظبي

NYU | ABU DHABI

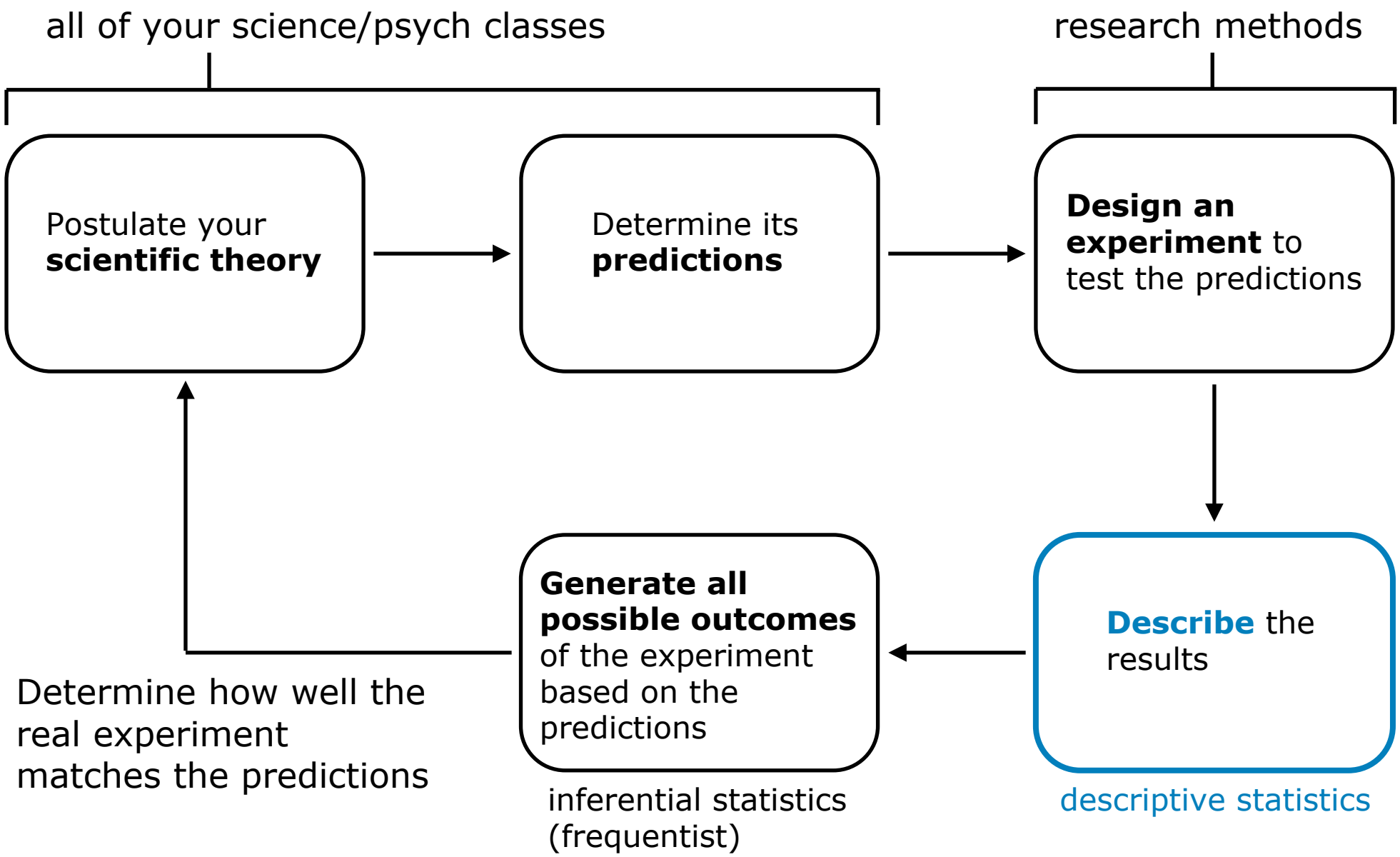
PSYCH-UH 1004Q: Statistics for Psychology

Class 3: Describing our results - Frequencies

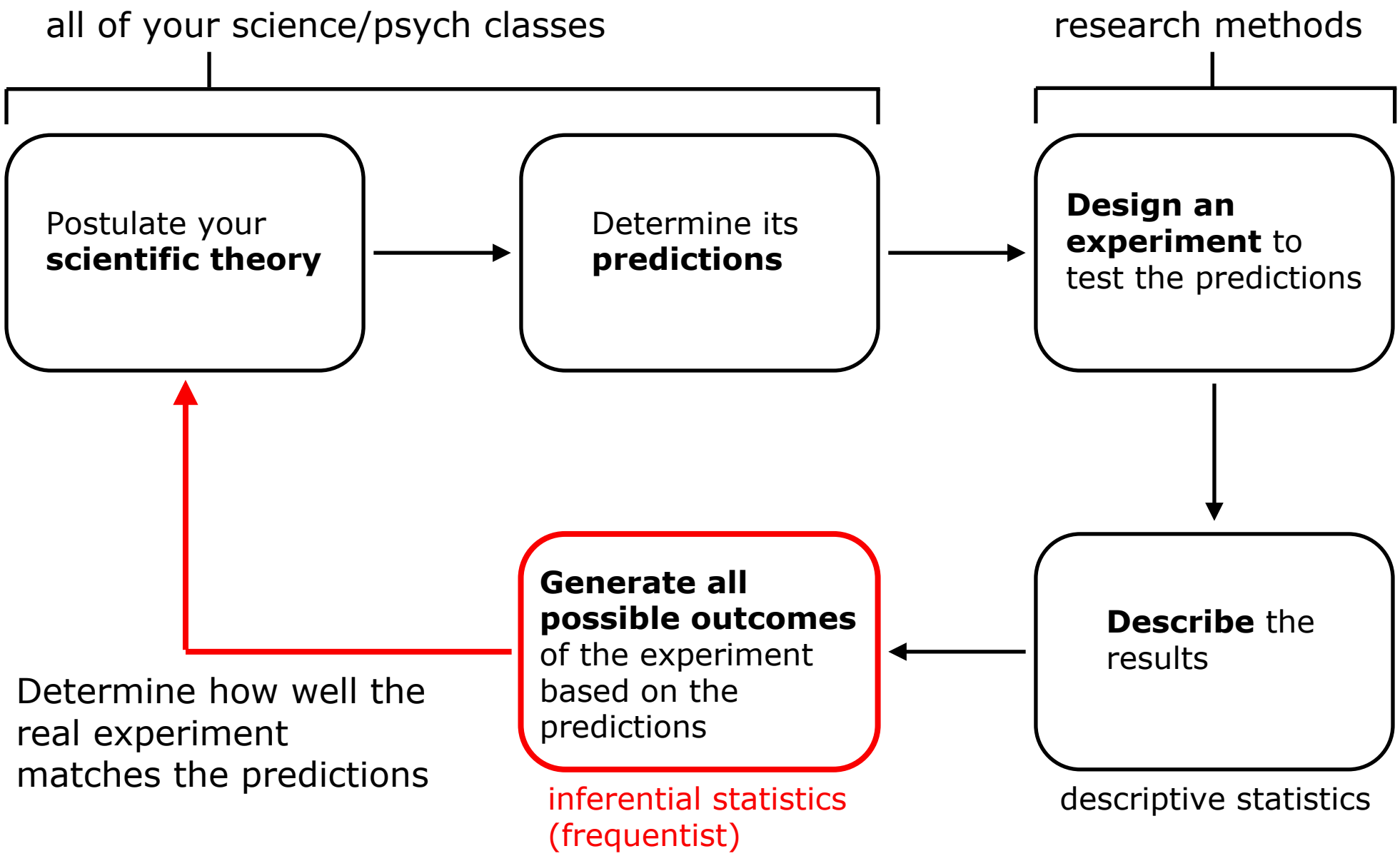
Prof. Jon Sprouse
Psychology

Why are we learning about frequency?

Our immediate goal: describing our data



Our longterm goal: evaluating our theory



Frequency and our immediate goal

Frequency: How often an event occurs.

Why does frequency matter for descriptive statistics (describing our results)?

Let's say the dependent variable in your experiment is measured on a scale from 1-5. And let's say you have 25 scores:

1 2 1 2 3 2 2 1 5 4 5 5 4 5 4 4 4 4 1 1 5 2 3 4 5

```
scale <-c(1:5)

scores <- sample(scale, 25, replace=T)
```

One way to organize is to arrange it into an ordered array based on the scores:

1 1 1 1 1 2 2 2 2 2 3 3 4 4 4 4 4 4 4 5 5 5 5 5 5

```
sort(scores)
```

But this can be impractical with large data sets. So we can instead count the frequency of each score:

score →	**1**	**2**	**3**	**4**	**5**
frequency →	5	5	2	7	6

```
table(scores)
```

Frequency and our longterm goal

The foundation of inferential statistics is probability. We want to be able to make statements about the probability of our data given the predictions of our theory.

But what is probability? Here is a neutral answer you have probably seen (or given!) to a question like this.

Probability: A mathematical statement about how likely an event is to occur. It takes a value between 0 and 1, where 0 means the event will never occur, and 1 means the event is certain to occur. (You can also think of it as a percentage 0% to 100%)

But what does it mean to say an event is likely? That is just restating the word probability (=probable, = likely). We want to add some content to this.

Frequentist probability: Probability is the long-run relative frequency of an event. In other words, the proportion of times the event would occur if the phenomenon were repeated a very large (ideally infinite) number of times. It is also called objective probability.

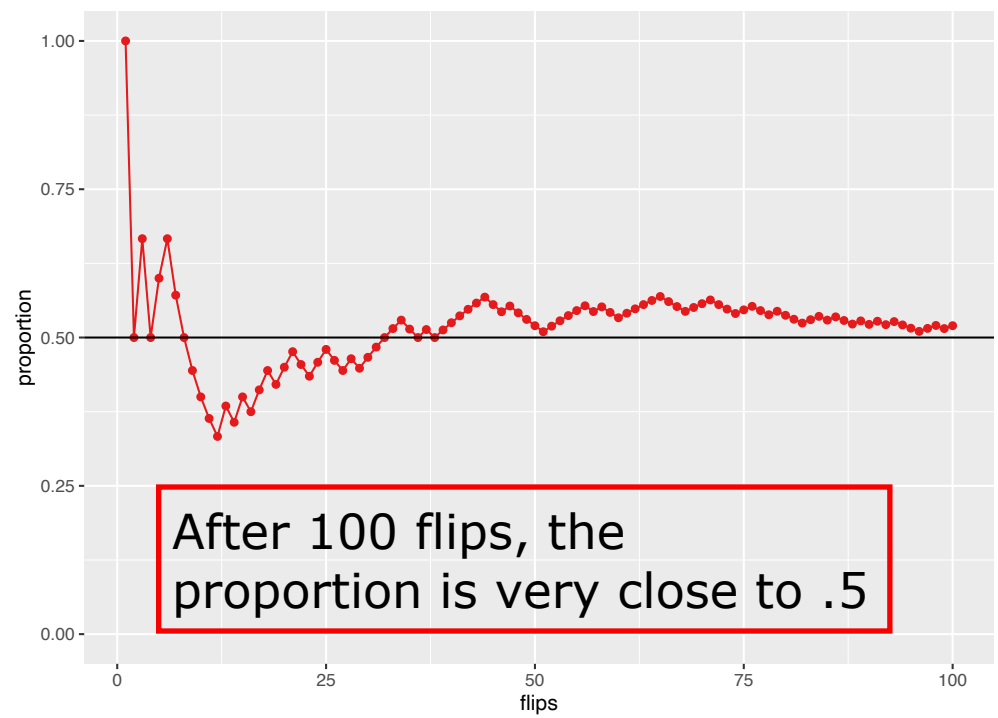
An example of frequentist probability

What is the probability that a balanced coin will land heads up when flipped?

You know the answer - 0.5. But how did you come up with that answer? Let's use the **frequentist definition** of probability.

The long run relative frequency requires a large number of repetitions of the process. So let's simulate flipping a coin 100 times:

```
library(tidyverse)

coin <- c("heads", "tails")

flips <- c(1:100)

simulations <- sample(coin, max(flips), replace=T)

heads <-
data.frame(proportion=cumsum(simulations=="heads")
/flips, flips)

p<-ggplot(heads, aes(x=flips, y=proportion)) +
        geom_line(color='#e41a1c')+
        geom_point(color='#e41a1c')+
        geom_hline(yintercept=.5)+
        ylim(0,1)
```



There is another type - subjective probability

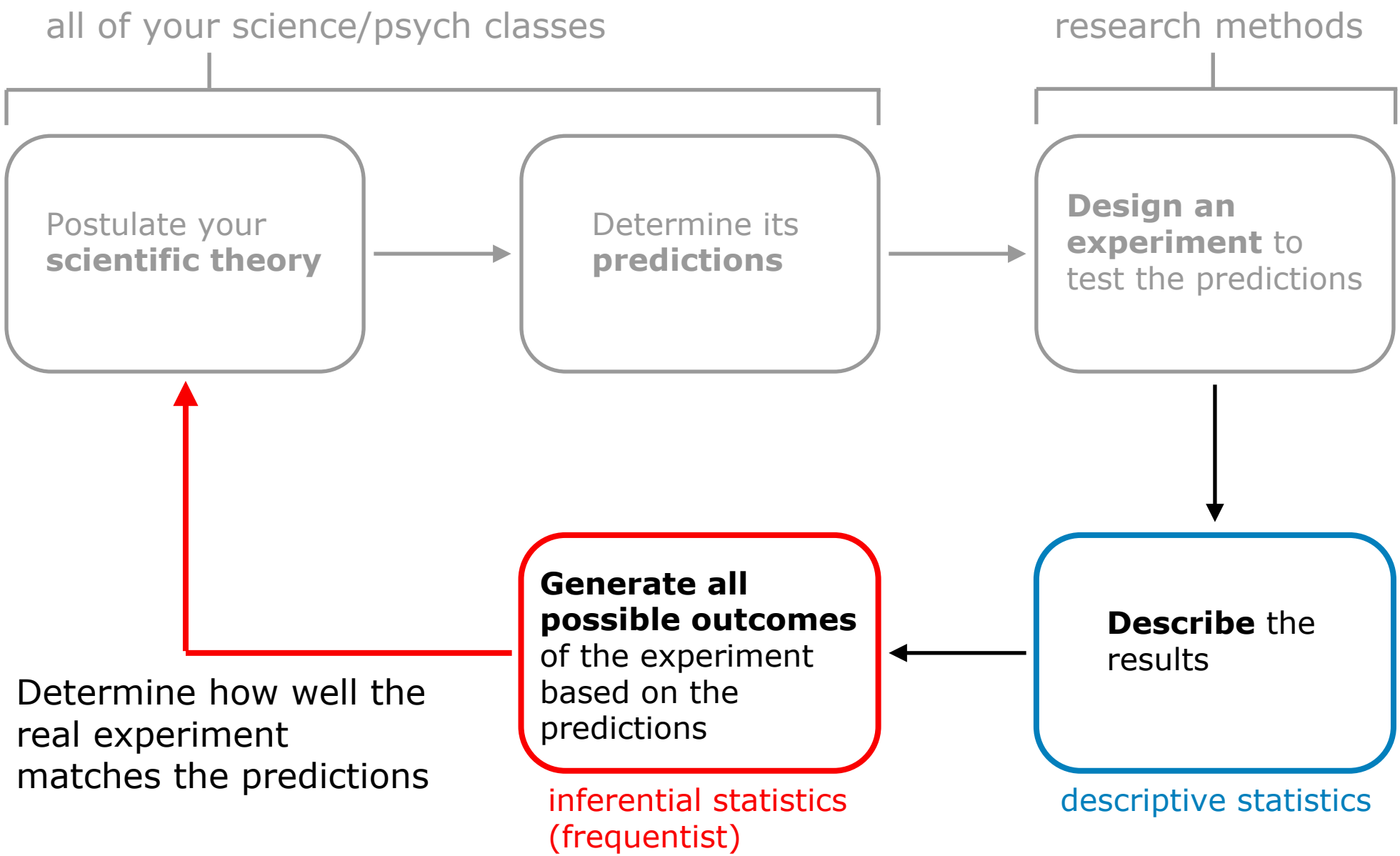
Subjective probability: Probability is the likelihood that an individual assigns to an event. In other words, it is about the belief that an individual has in a specific outcome.

We tend to use subjective probability when we say things like "there is a 10% chance that it will rain tomorrow". We are reporting our beliefs. We tend to use subjective probability when discussing events that cannot be repeated (like "tomorrow").

This is not to say that there is no way to think about singular events using frequentist probability. We could simulate 100 hypothetical "tomorrows", each one starting with conditions based on today, using a sophisticated computer (this is how weather forecasting works). The issue is just that we cannot actually repeat the event in the real world (unlike a coin flip).

There is a type of statistics that explores subjective probability. It is called Bayesian statistics, after Reverend Thomas Bayes (1701-1761). I am very much a fan of Bayesian statistics, and I encourage you to learn about it **later**. For this course, your first course in statistics, you should focus on frequentist statistics (using frequentist/objective probability).

Frequencies are the foundation of both of our goals in this course



OK, now let’s learn about frequency so we can describe our data!

(We will look at probability and inferential statistics a bit later in the course.)

Frequency **Distribution**

unsorted	sorted
1	5
2	5
1	5
2	5
3	5
2	5
2	4
1	4
5	4
4	4
5	4
5	4
4	4
5	3
4	3
4	2
4	2
4	2
1	2
1	2
5	1
2	1
3	1
4	1
5	1

On the left is a very simple data set. It has 25 scores in it. There are 5 possible values of the scores, 1 through 5. Each value occurs some number of times.

The second column is just the data set sorted by value, with some colors added so that the frequency of each score can be seen.

Frequency just means how often something occurs. It is a count. We can count the frequency of any object or event. In this case, it is our types of scores.

A **distribution** tells us how often each value occurs. Distributions can do this with frequency (a frequency distribution) or with probability (a probability distribution).

We can show a distribution with a list like this, or with a **table**, or with a **plot**. We will build frequency tables and plots now!

unsorted	sorted
1	5
2	5
1	5
2	5
3	5
2	5
2	4
1	4
5	4
4	4
5	4
5	4
4	4
5	3
4	3
4	2
4	2
4	2
1	2
1	2
5	1
2	1
3	1
4	1
5	1

Frequency tables

scores	f	cf	rf	crf
5	6	25	.24	1
4	7	19	.28	.76
3	2	12	.08	.48
2	5	10	.20	.40
1	5	5	.20	.20

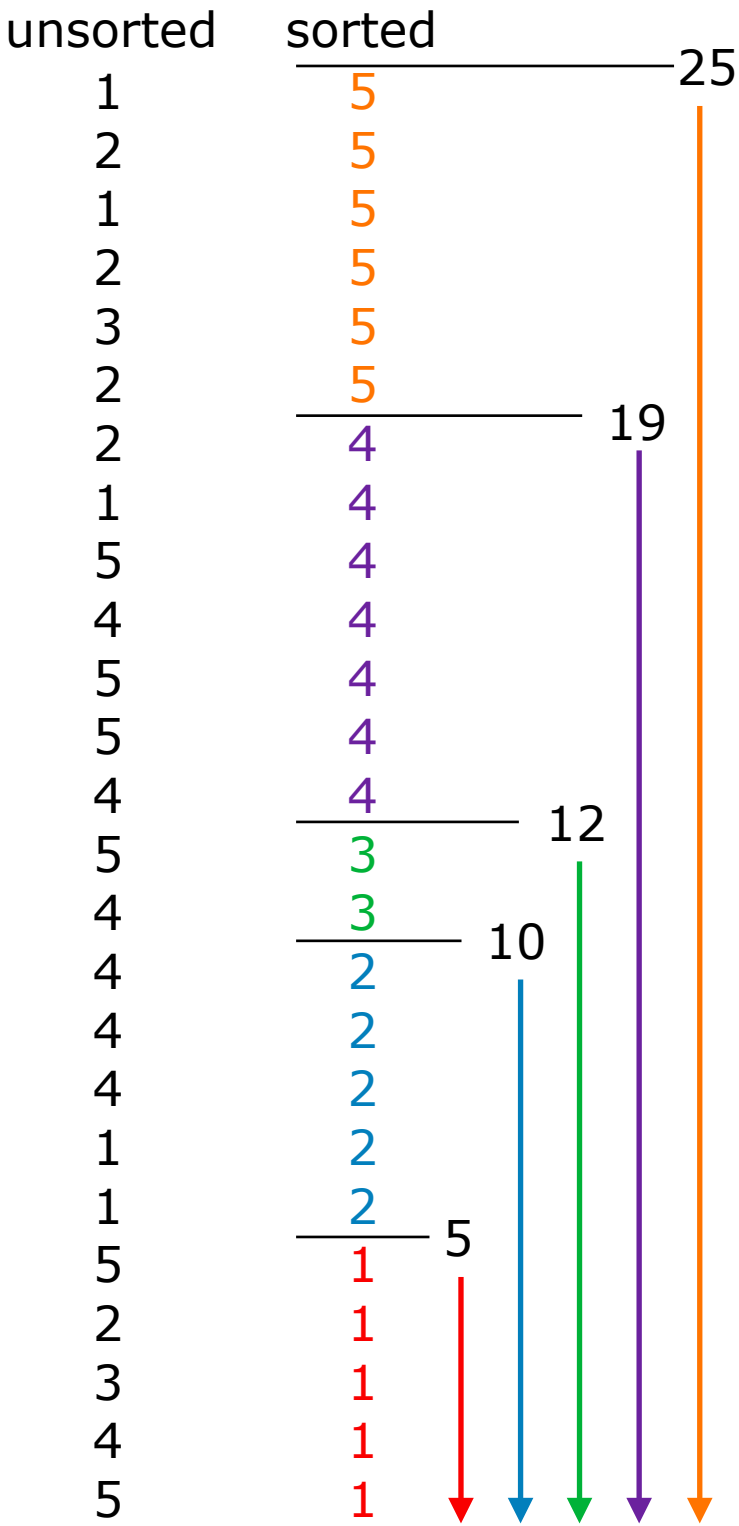
To see the power of describing our data using frequencies, we will assume that our data is **at least ordinal** (interval and ratio will be similar). The first step is to sort the data set according to its order (you can choose the direction that makes the most sense to your theory).

unsorted	sorted
1	5
2	5
1	5
2	5
3	5
2	5
2	4
1	4
5	4
4	4
5	4
5	4
4	4
5	3
4	3
4	2
4	2
4	2
1	2
1	2
5	1
2	1
3	1
4	1
5	1

Frequency tables

scores	f	cf	rf	crf
5	6	25	.24	1
4	7	19	.28	.76
3	2	12	.08	.48
2	5	10	.20	.40
1	5	5	.20	.20

Frequency is just the count of each score. I have colored the scores to make it easy to verify these counts in our toy example.



Frequency tables

scores	f	cf	rf	crf
5	6	25	.24	1
4	7	19	.28	.76
3	2	12	.08	.48
2	5	10	.20	.40
1	5	5	.20	.20

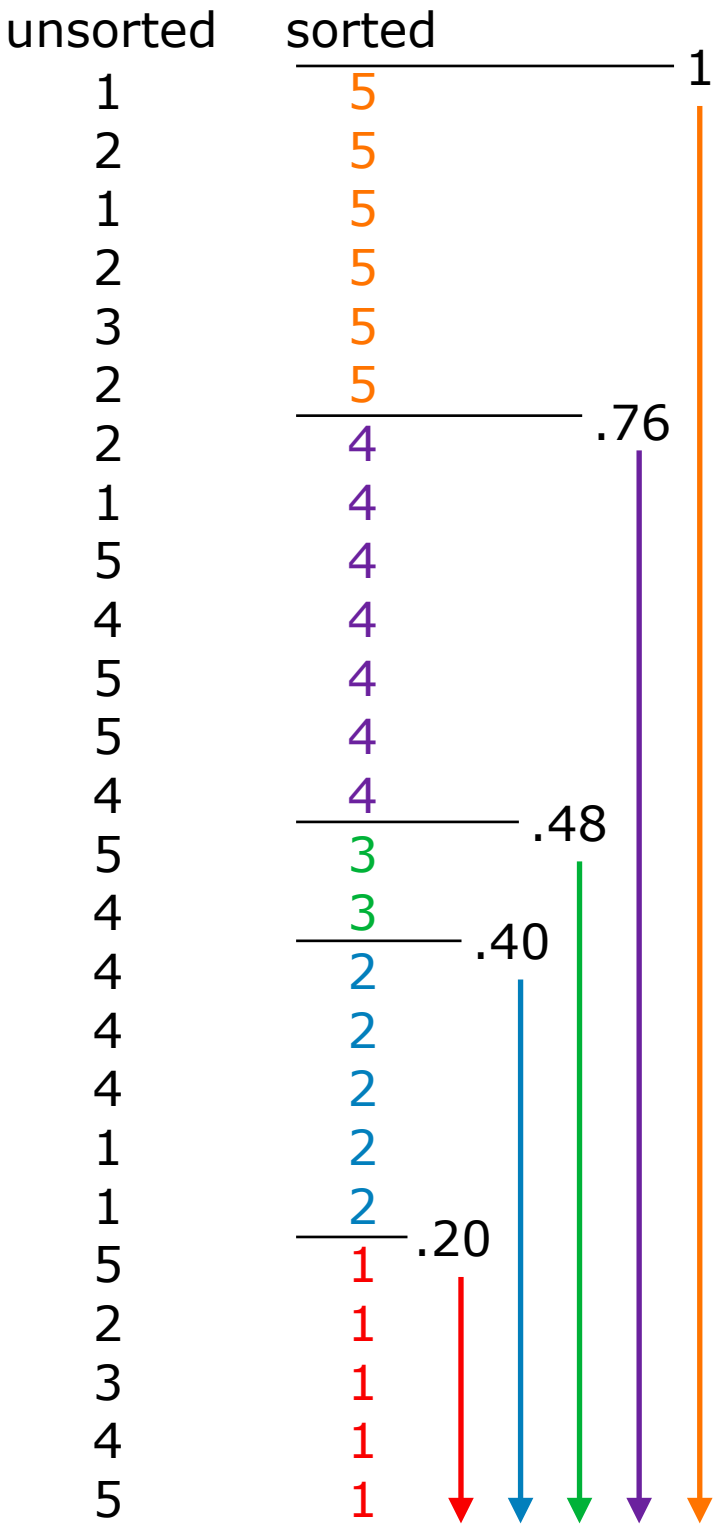
Cumulative frequency is the sum of frequencies at and below the score. This is where the choice of order matters. This shows us the count of scores equal to or less than the chosen score.

unsorted	sorted	
1	5	
2	5	
1	5	6/25 =.24
2	5	
3	5	
2	5	
2	4	
1	4	
5	4	
4	4	7/25 =.28
5	4	
5	4	
4	4	
5	3	2/25 =.08
4	3	
4	2	
4	2	
4	2	5/25 =.20
1	2	
1	2	
5	1	
2	1	
3	1	5/25 =.20
4	1	
5	1	

Frequency tables

scores	f	cf	rf	crf
5	6	25	.24	1
4	7	19	.28	.76
3	2	12	.08	.48
2	5	10	.20	.40
1	5	5	.20	.20

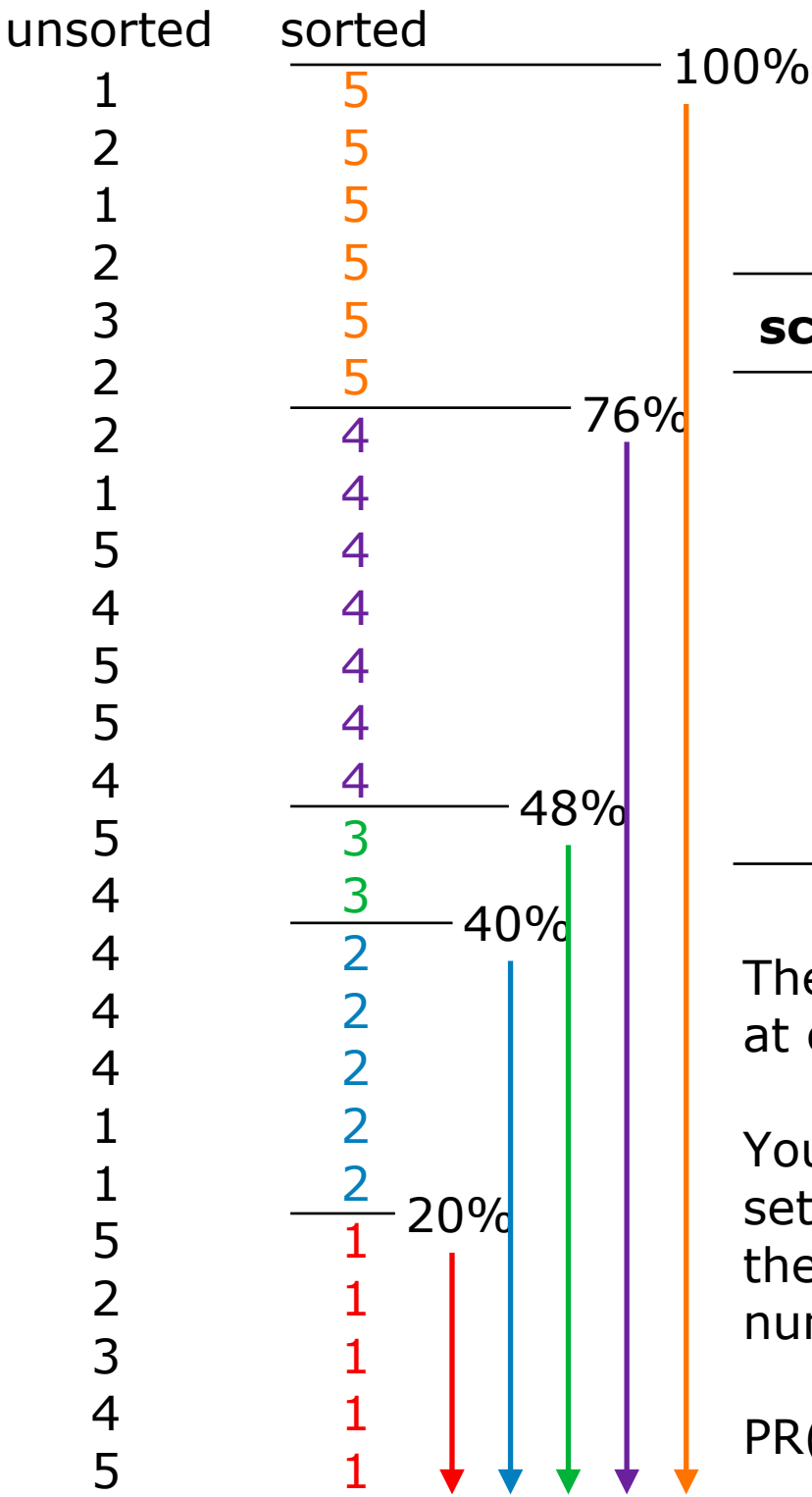
Relative frequency is the frequency of each score divided by the total number of scores (in this case, 25). This shows us the proportion of the total data set represented by each score.



Frequency tables

scores	f	cf	rf	crf
5	6	25	.24	1
4	7	19	.28	.76
3	2	12	.08	.48
2	5	10	.20	.40
1	5	5	.20	.20

Cumulative relative frequency is the sum of the relative frequencies at and below the score. This shows us the proportion of scores that are equal to or less than the chosen score.



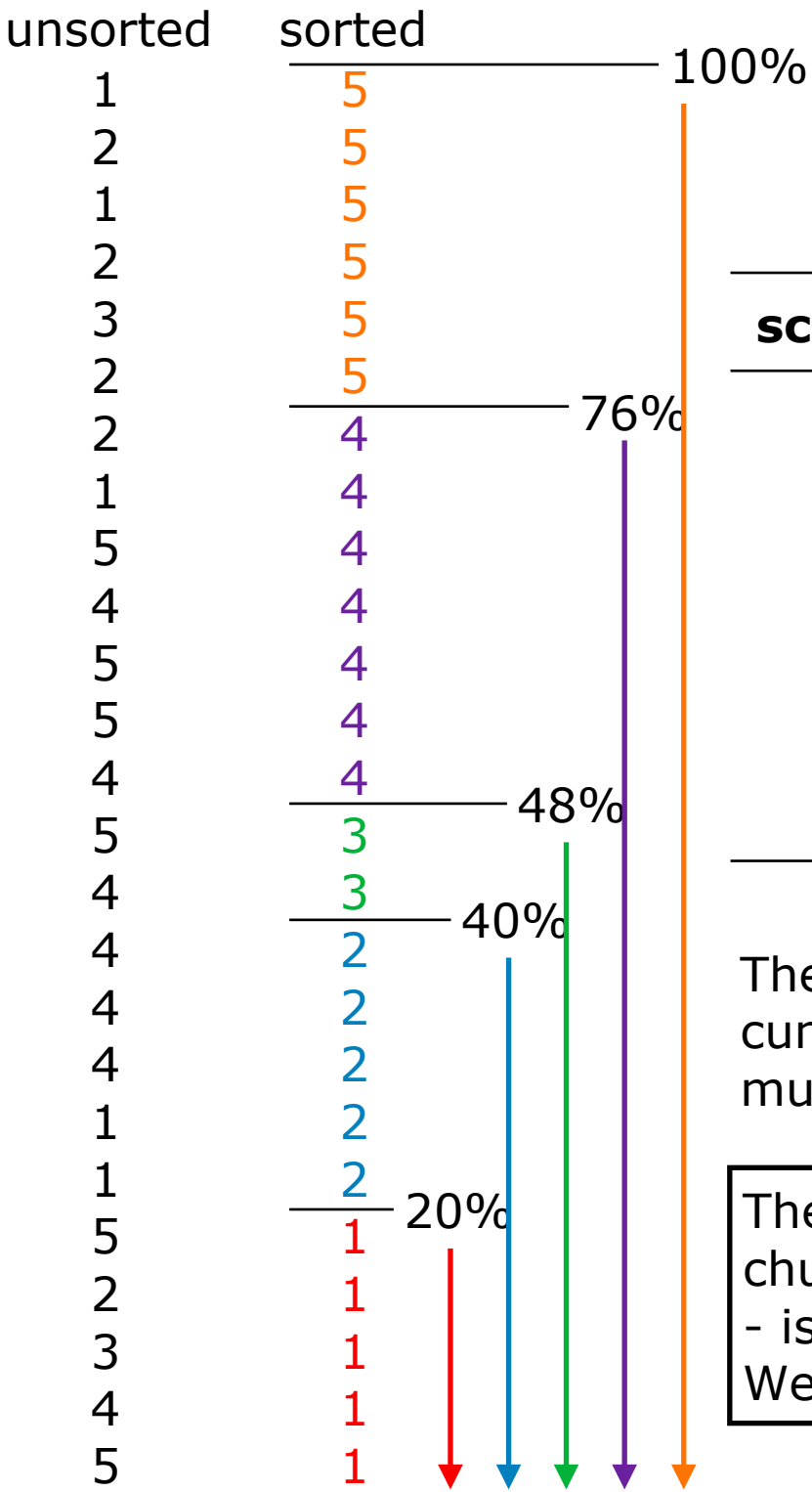
Percentile Rank

scores	f	cf	rf	crf	pr
5	6	25	.24	1	100
4	7	19	.28	.76	76
3	2	12	.08	.48	48
2	5	10	.20	.40	40
1	5	5	.20	.20	20

The **Percentile Rank** is the percentage of scores at or below the chosen score.

You can calculate it by hand by sorting the data set, counting the number of scores at or below the current score, and dividing by the total number of scores and multiplying by 100:

$$PR(score) = \frac{\#\text{ at or below score}}{\text{total }\#\text{ of scores}} \times 100$$



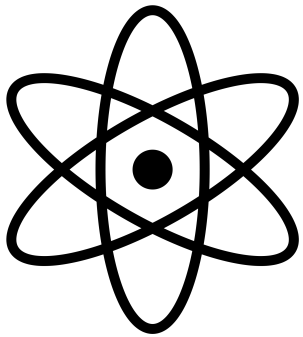
Percentile Rank and crf

scores	f	cf	rf	crf	pr
5	6	25	.24	1	100
4	7	19	.28	.76	76
3	2	12	.08	.48	48
2	5	10	.20	.40	40
1	5	5	.20	.20	20

The **Percentile Rank** is obviously related to the cumulative relative frequency - it is just the crf multiplied by 100 (into percentages).

The idea of splitting a distribution into two chunks - scores above and below a critical score - is an incredibly important concept in statistics. We will practice this more later today!

Why do we care?



Right now our goal is just to learn how to calculate frequency, relative frequency, cumulative relative frequency, etc.

So it is going to feel like a list of concepts. That is ok. Just focus on the concepts, and try not to worry about how to apply them to science yet. We will get there.

As we keep moving through the course, you will see how these are the foundational concepts for what we do in frequentist statistic.

One thing to also keep in mind is that what we want to do is become so familiar with the mathematical concepts of frequency that using it to answer our scientific questions takes no thought whatsoever. So, if once these concepts start feeling easy and boring, you are ready to move on to the deeper parts of statistics!

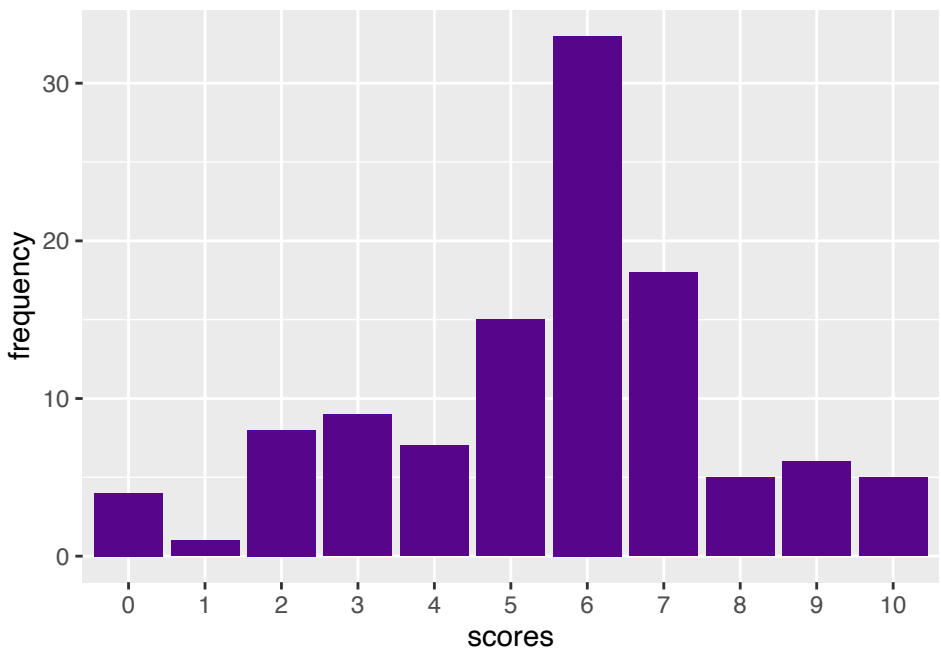
Frequency tables become more useful for larger data sets

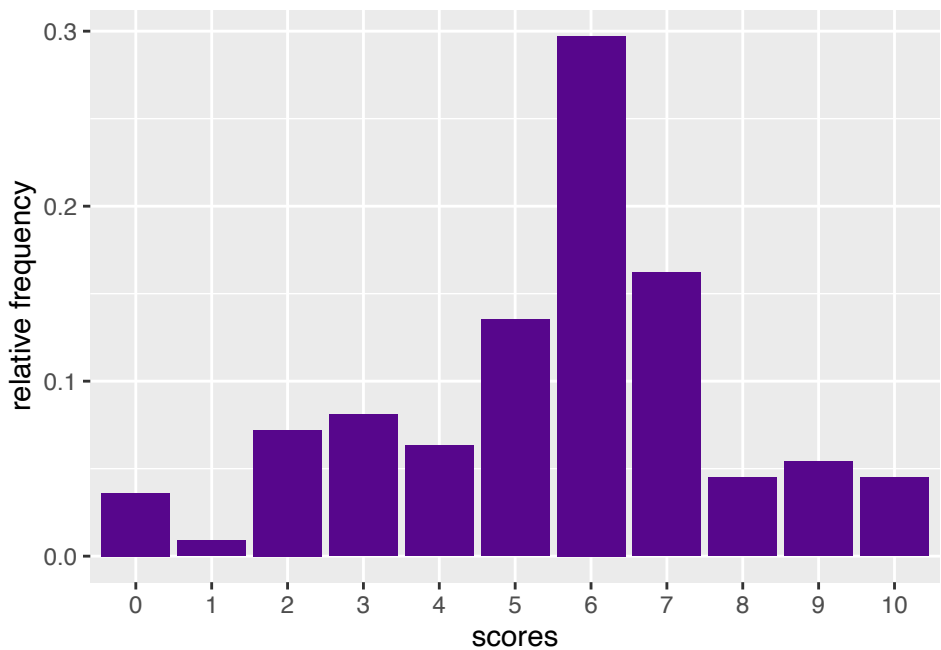
10	10	10	10	10	9
9	9	9	9	9	8
8	8	8	8	7	7
7	7	7	7	7	7
7	7	7	7	7	7
7	7	7	7	6	6
6	6	6	6	6	6
6	6	6	6	6	6
6	6	6	6	6	6
6	6	6	6	6	6
6	6	6	6	6	6
6	5	5	5	5	5
5	5	5	5	5	5
5	5	5	5	4	4
4	4	4	4	4	3
3	3	3	3	3	3
3	3	2	2	2	2
2	2	2	2	1	0
0	0	0			

score	f	cf	rf	crf	pr
10	5	111	.045	1	100
9	6	106	.054	.955	95.5
8	5	100	.045	.901	90.1
7	18	95	.162	.856	85.6
6	33	77	.297	.694	69.4
5	15	44	.135	.396	39.6
4	7	29	.063	.261	26.1
3	9	22	.081	.198	19.8
2	8	13	.072	.117	11.7
1	1	5	.009	.045	4.5
0	4	4	.036	.036	3.6

Plots can be even more useful

Bar plots are used to show the frequency of discrete responses. We put the possible responses on the x-axis and the frequency measure on the y-axis. Because the possible responses are discrete, there is a space between the bars (to show that the values are discrete).





Pro tip: Always label your axes with words that a human can understand. R will name them after the factors in your data set. If those are not easy to understand, you have to rename them (either in the data set or in the plot directly).

Continuous variables require grouping

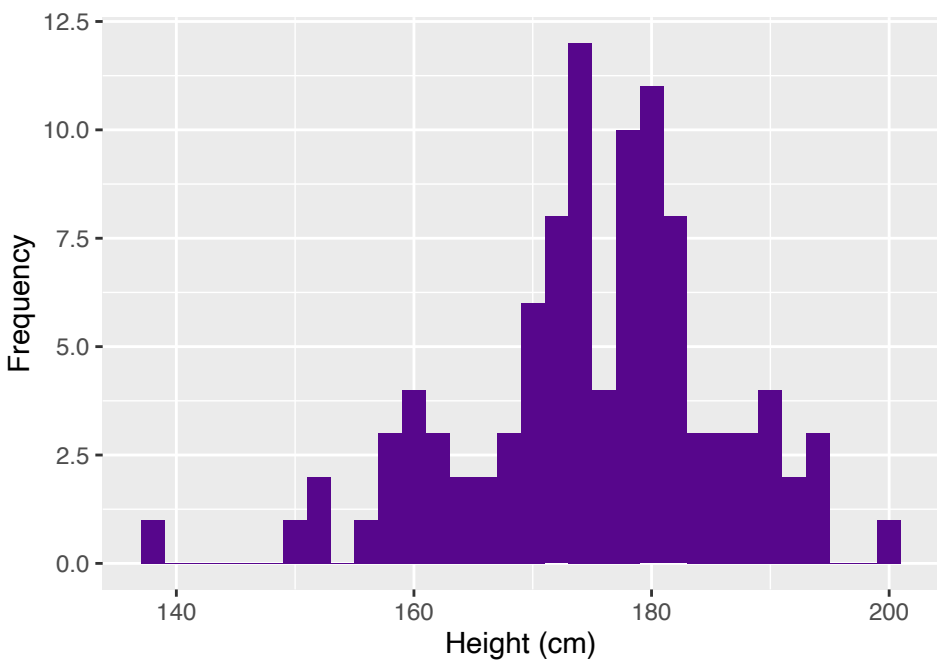
This is a simulated data set of human heights in cm. It is a simulated sample of 100 participants. Height is a **continuous variable**, so there is an infinite number of possible values. If we want to count frequencies, we need to create **bins** that contain a range of values. Here are two possible options: bins of size 2cm and bins of size 5 cm. (There are lots of other options!)

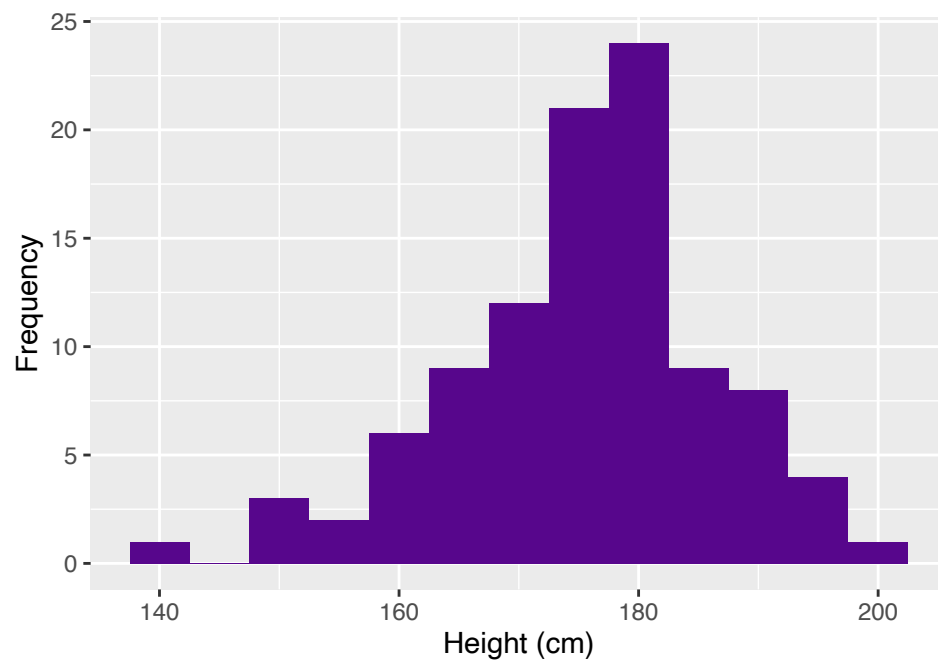
range	freq	range	freq
136 - 138	1	168 - 170	2
138 - 140	0	170 - 172	8
140 - 142	0	172 - 174	11
142 - 144	0	174 - 176	6
144 - 146	0	176 - 178	8
146 - 148	0	178 - 180	10
148 - 150	0	180 - 182	11
150 - 152	2	182 - 184	5
152 - 154	1	184 - 186	4
154 - 156	0	186 - 188	2
156 - 158	2	188 - 190	4
158 - 160	3	190 - 192	3
160 - 162	3	192 - 194	4
162 - 164	3	194 - 196	0
164 - 166	3	196 - 198	0
166 - 168	3	198 - 200	1

range	freq
135-140	1
140-145	0
145-150	0
150-155	3
155-160	5
160-165	8
165-170	6
170-175	25
175-180	18
180-185	18
185-190	8
190-195	7
195-200	1

Histograms for plotting continuous variables

Histograms are used to show the frequency of continuous variables. We put the values of the variable on the x-axis and the frequency measure on the y-axis. Because there are an infinite number of possible values for the variable, we have to create bins ourselves to make them countable. Because the bins are chunking up a continuous variable, the bars of a histogram touch each other!

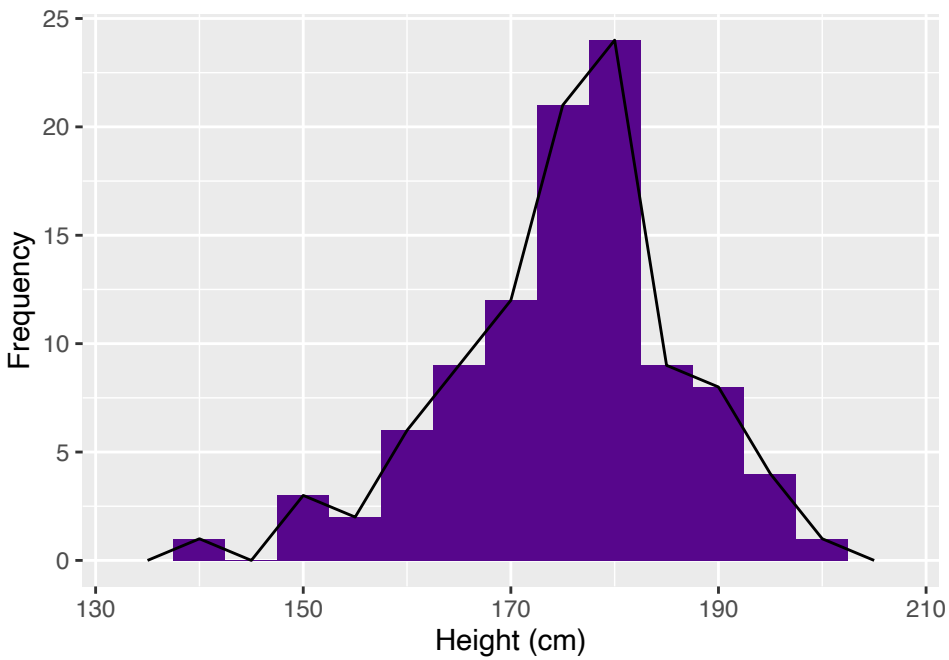


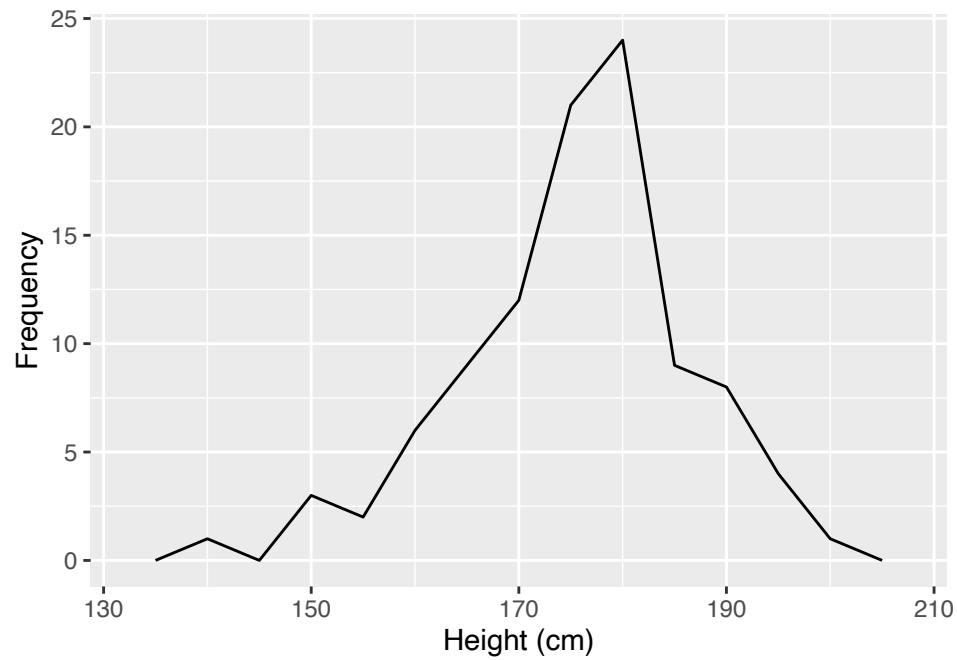


The size of the bins matters. You can see that the basic shape is the same between the two, but the details change. There are no hard rules for choosing bin sizes. Ideally, your theory will help guide you to meaningful bin sizes.

Frequency polygon

Frequency polygons are used to show the shape of the distribution of scores. They are created by choosing a point (usually the center) on each bar, and drawing a straight line from point to point.



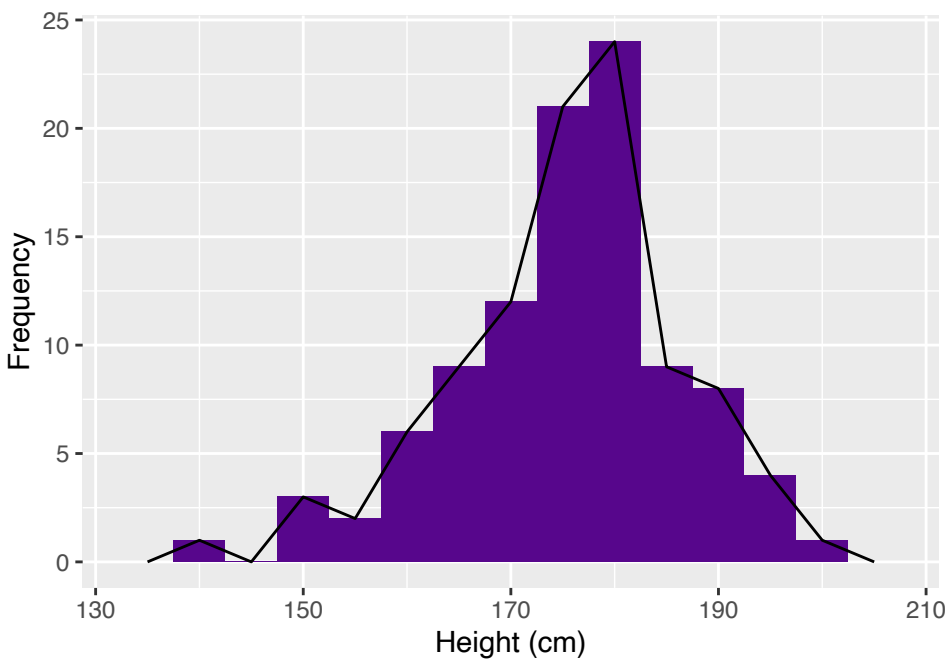


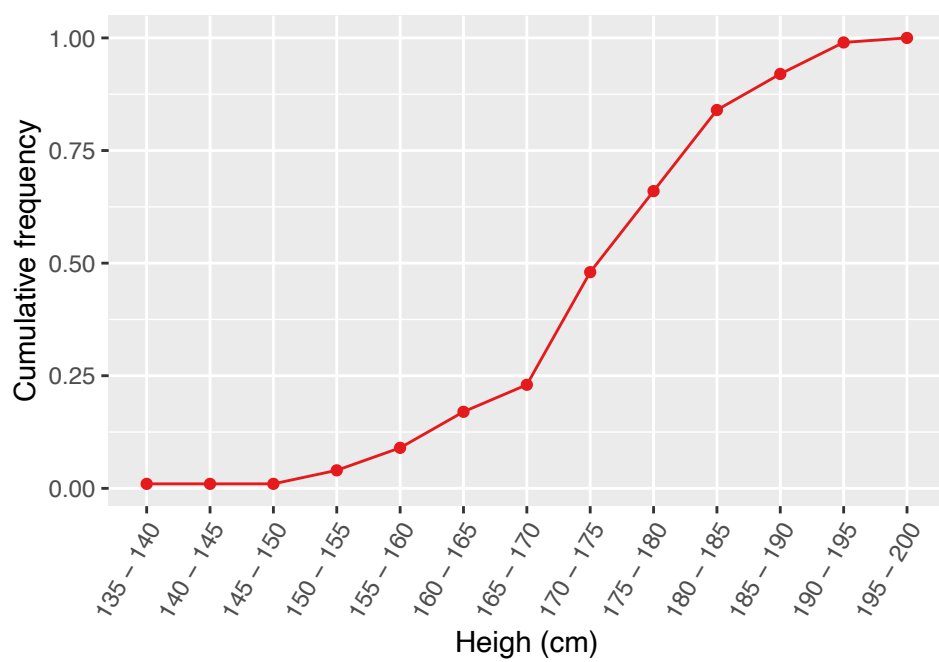
Frequency polygons are not used all that often in the daily life of a scientist. But they are a stepping stone toward probability distributions, which are very important. So, it is useful seeing how you can begin to see the shape of a distribution from a frequency plot like a histogram overlaid with a frequency polygon.

Cumulative frequency polygon

Cumulative frequency polygons are used to show the shape of the cumulative frequency distribution of scores. They are created by plotting the cumulative frequencies for each score or bin.

I have placed the frequency histogram beside the cumulative frequency polygon so you can see their relationship.





Again, these are not used all that often in daily science. But there are some types of experiments where the critical measure is the cumulative response, so it is worth knowing about these!

Identifying the percentile rank of a value

(PR is one example of one of the most fundamental concepts in statistics - splitting a distribution into two chunks at a critical score.)

Identifying the percentile rank of a score

We will use distributions constantly in statistics. And one of the most common tasks we will do is identify the percentile rank of a score in a distribution. Let's do a basic example:

data set =

2	4	6	8	10	12	14	16	18	20	22	24	26	28
30	32	34	36	38	40	42	44	46	48	50	52	54	56
58	60	62	64	66	68	70	72	74	76	78	80	82	84
86	88	90	92	94	96	98	100						

What is the **percentile rank** of the value 92?

It is 92! 92% of the scores are equal to or less than 92, and 8% are greater.

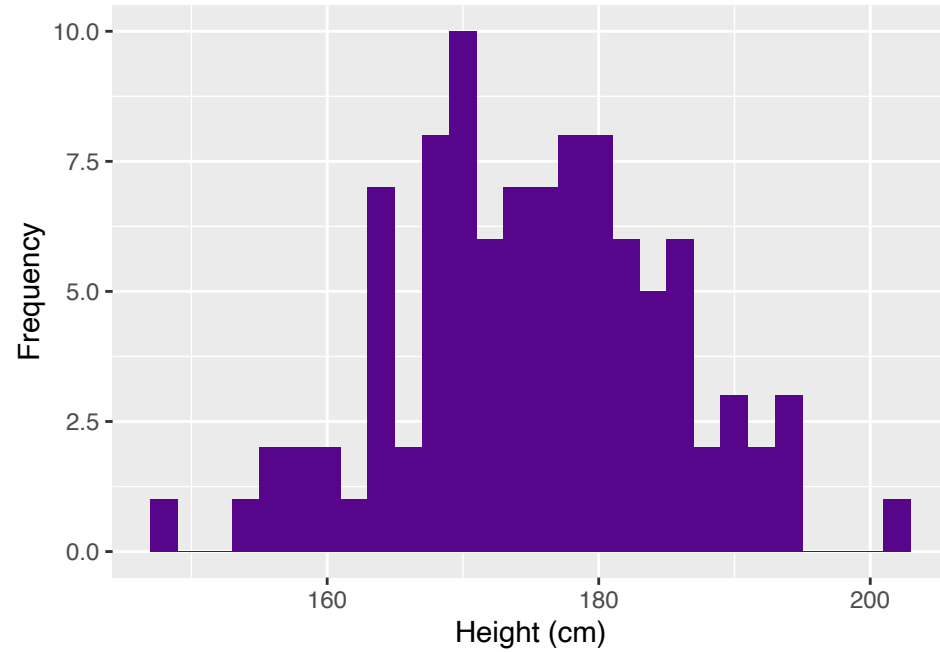
But how did we calculate that? We count the scores that are equal to or less than the target scores, and divide by the total number of scores:

$$\frac{\text{\# scores equal to or less than 92}}{\text{\# scores}} = \frac{46}{50} = .92 \times 100 = 92\%$$

A more complicated example

Let's do the same thing with a new simulation of height. In this one, I will round the height to the nearest cm for convenience.

Let's find the percentile rank of 173 by hand:

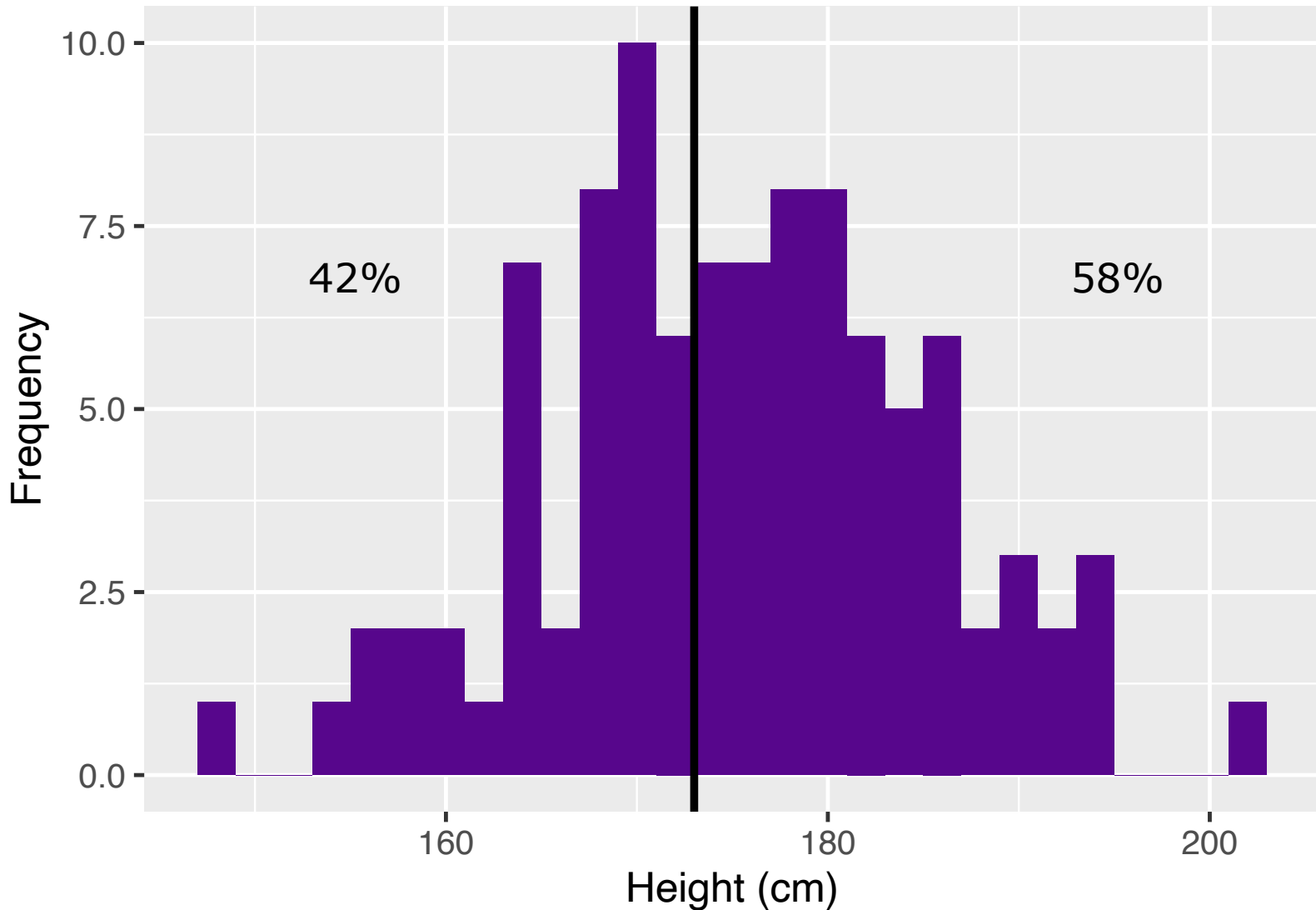


148 154 156 157 158 158 160 161 163 164 164 164 164 164 165 165 166
167 168 168 168 169 169 169 169 169 170 170 170 170 170 170 171 171
171 171 172 172 173 173 173 173 174 174 174 174 175 175 175 176 176
176 176 177 177 177 178 178 179 179 179 179 179 179 180 180 180 180
181 181 181 181 182 182 182 183 183 183 184 184 185 185 185 186 186
186 186 187 187 189 189 190 191 191 192 192 195 195 195 202

$$\frac{\text{\# scores equal to or less than 173}}{\text{\# scores}} = \frac{42}{100} = .42 \times 100 = 42\%$$

The visual consequence

If we plot our distribution of height scores, and plot a line at 173, this line divides the distribution into 42% on the left and 58% on the right.



Identifying the critical value that divides the distribution based on a percentile

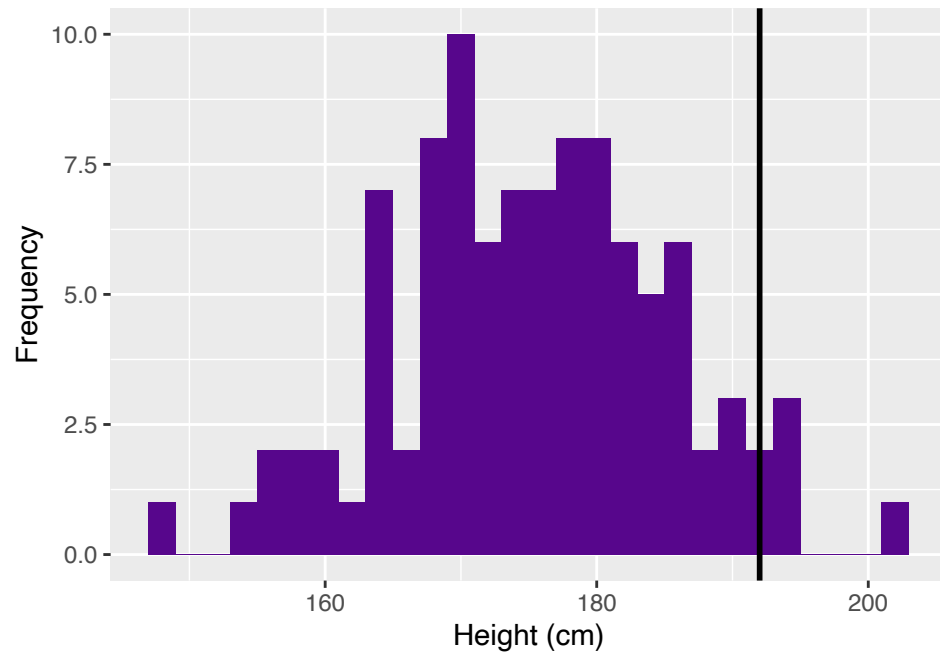
(Another one of the most common tasks in statistics)

The other direction - percentiles

In the previous section we working in this order: (i) choose a value, (ii) look up the percentile rank of the value. But we also need to be able to do the other direction: (i) choose a percentile, and look up the value of that percentile.

For example, if we look at our height data set, we can ask which value is equal to or greater than 95% of the values. The answer is 192.

(192 is also at the 96th percentile. If we asked what percentile 192 is, we'd get 96th. But we asked specifically about the 95th percentile, which means the value that is equal to or greater than 95% of values.)



148 154 156 157 158 158 160 161 163 164 164 164 164 164 165 165 166
167 168 168 168 169 169 169 169 169 170 170 170 170 170 170 171 171
171 171 172 172 173 173 173 173 174 174 174 174 175 175 175 176 176
176 176 177 177 177 178 178 179 179 179 179 179 179 180 180 180 180
181 181 181 181 182 182 182 183 183 183 184 184 185 185 185 186 186
186 186 187 187 189 189 190 191 191 192 192 195 195 195 202

The general idea - quantiles

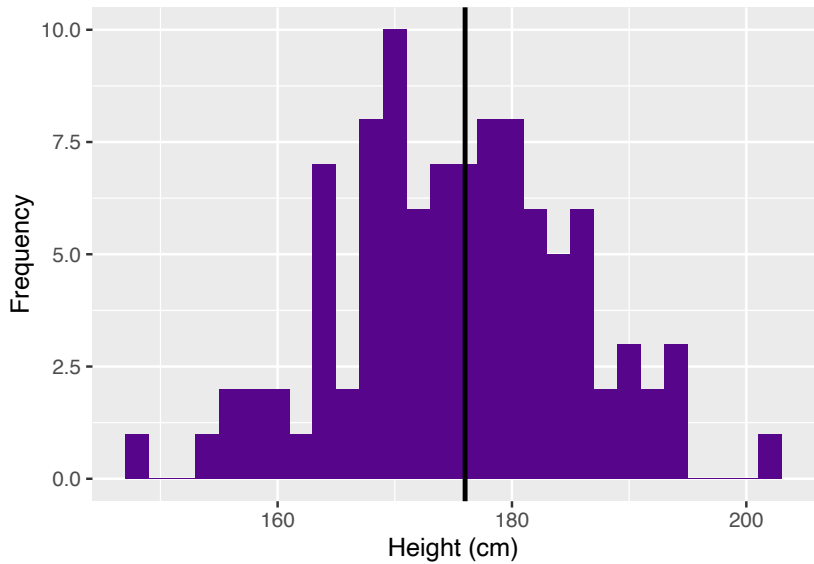
Generalizing the idea: **quantiles**

We will primarily work with percentiles in this course. But percentiles are a specific instance of a more general idea - dividing the distribution into equal chunks:

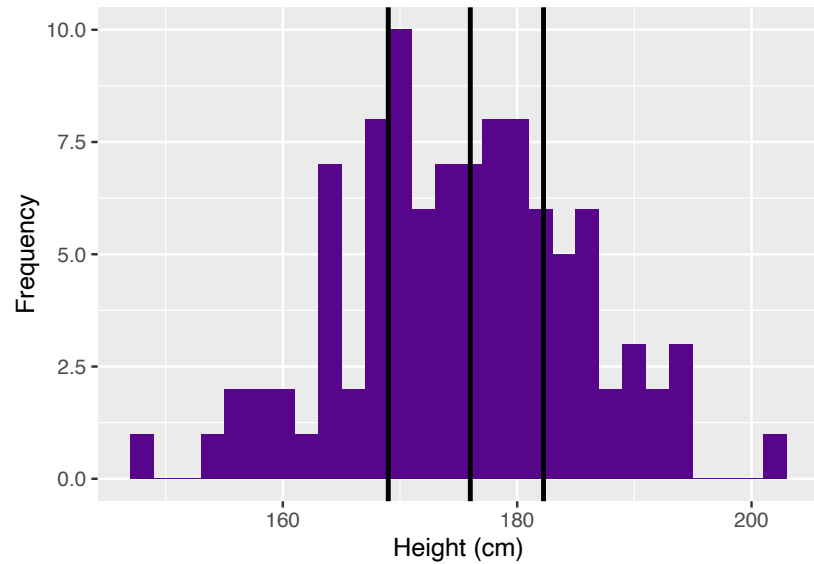
quantile: A value that divides a distribution into equal (adjacent) subgroups. Basically, the quantiles are the "cut points".

You can choose to divide a distribution into whichever number of subgroups you want. The number of quantiles (cut points) will be 1 fewer than the number of groups.

2 50% chunks, 1 quantile (cut point)



4 25% chunks, 3 quantiles (cut points)



Special quantiles: percentiles, deciles, quartiles

Though you can define quantiles using any number of chunks that you want, there are three special ones that appear frequently: percentiles, deciles, quartiles.

percentile: Quantiles (cut points) that divide the distribution into 100 equal chunks (1% each). Technically, there are 99 percentiles.

decile: Quantiles (cut points) that divide the distribution into 10 equal chunks (10% each). Technically, there are 9 deciles.

quartile: Quantiles (cut points) that divide the distribution into 4 equal chunks (25% each). Technically, there are 3 quartiles.

The R function for finding quantiles is quantile(). Its default behavior is to give you quartiles, but you can tell it to give you other cut points. It also allows unequal cut points - so it is a bit more general than the narrow statistical concept of quantile as equal cuts.

Some sloppiness in terms

Though the technical definition of a quantile (percentile, quartile, etc) is that they are the cut points dividing the distribution, sometimes people use these terms to label the chunks (or intervals). This is sloppy, but it is natural. Sometimes you want to talk about the chunks/intervals!

1st quartile: Either the cut point for the 25th percentile, or the interval between the 0th and 25th percentile.

3rd quartile: Either the cut point for the 75th percentile, or the interval between the 50th and 75th percentile.

My best advice is to use context to tell you whether the person is describing the cut points or the intervals.