Plan for the week

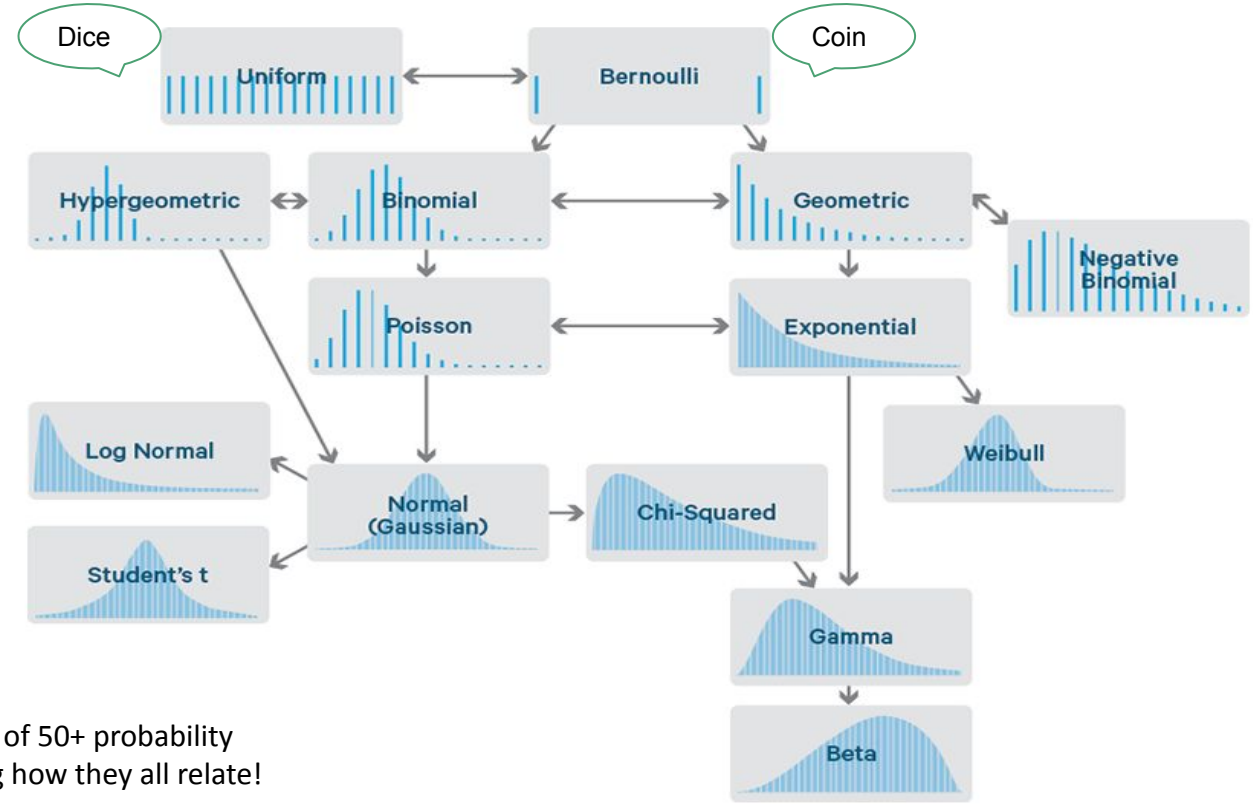
- M: Probability distributions & histograms
- W: Measures of dispersion
  - Variance, standard deviation, covariance, correlation
- F: Section
  - Review HW 0
  - Visualization Tips, with special guest `ggplot`

Probability Distributions

Probability

- This is not a math class, or an applied math class, or a statistics class; but it is a computer science course!
- Still, probability, which is a math-y concept underlies much of what we will do in this course.
- You might not know it, but you are likely already at least somewhat familiar with probability.
  - If you flip a coin, what is the chance that it will turn up heads?
  - If you roll a die, what is the chance you will roll a 6?

There are many, many model probability distributions



Here's a link to a map of 50+ probability distributions, showing how they all relate!

Features of Probability Distributions

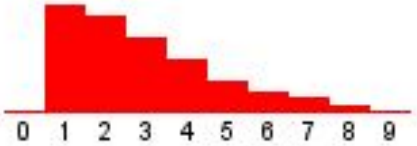
- The center is the mean, median, or mode.
- The spread is the variability of the data:

Less spread

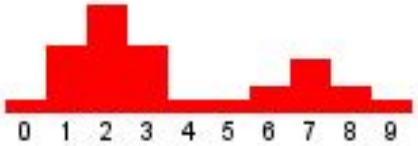
More spread

- Shape can be described by symmetry, skewness, number of peaks (modes), etc.

Symmetric, unimodal, bell-shaped

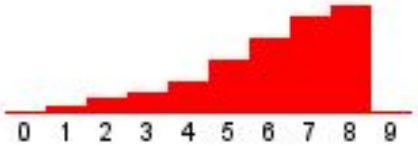


Skewed right

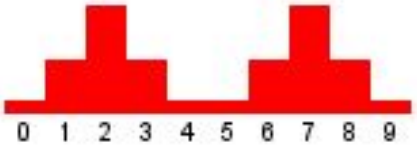


Non-symmetric, bimodal

Uniform



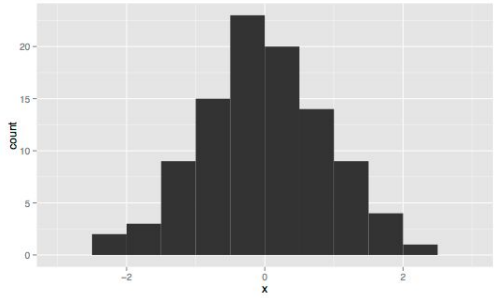
Skewed left



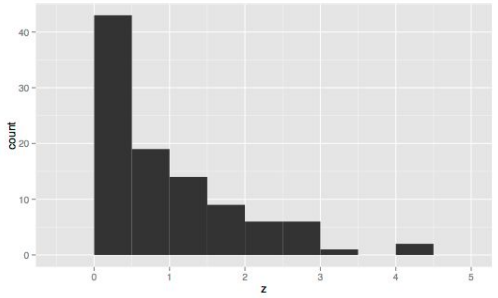
Symmetric, bimodal

Image source

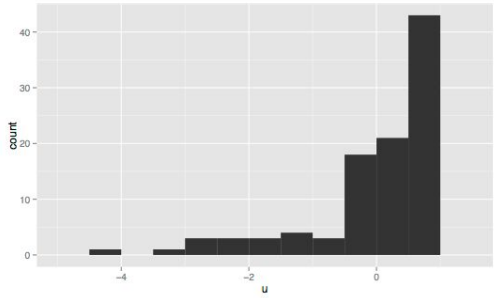
Descriptive statistics



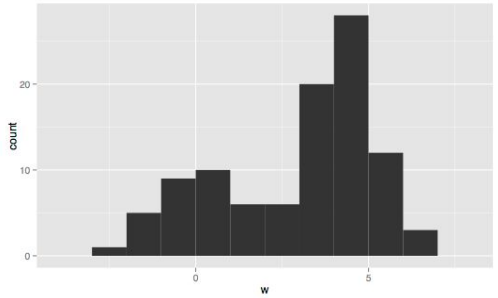
Symmetric, unimodal



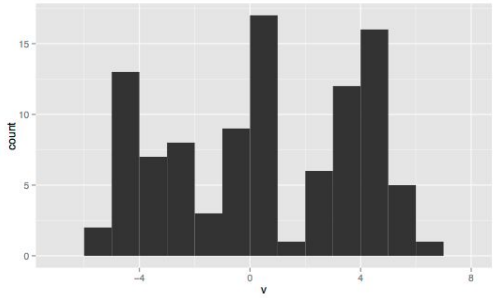
Skewed right



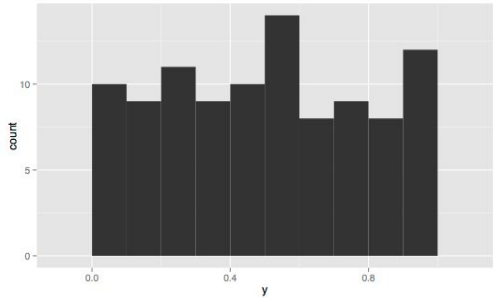
Skewed left



Bimodal



Multimodal

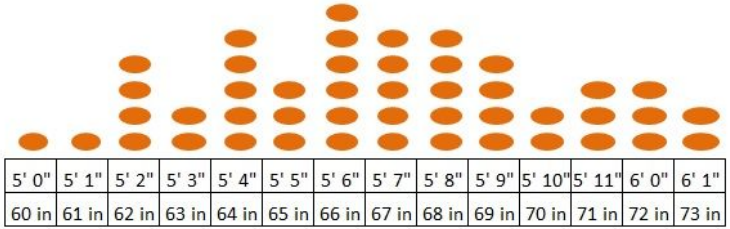


Symmetric

Image Source

Frequency Distributions

A statistics class at Simon Fraser University



5' 0"	5' 1"	5' 2"	5' 3"	5' 4"	5' 5"	5' 6"	5' 7"	5' 8"	5' 9"	5' 10"	5' 11"	6' 0"	6' 1"
60 in	61 in	62 in	63 in	64 in	65 in	66 in	67 in	68 in	69 in	70 in	71 in	72 in	73 in

Image source



- 2016 TA Andreas loves M&M'S; he once ate a bag of 55 M&Ms in less than 10 seconds!
- M&M'S have one variable, color, which has six possible values (outcomes): brown, red, yellow, green, blue, orange
- The bag Andreas inhaled contained 17 brown M&M'S, 18 red M&M'S, 7 yellow M&M'S, 7 green M&M'S, 2 blue M&M'S, 4 orange M&M'S

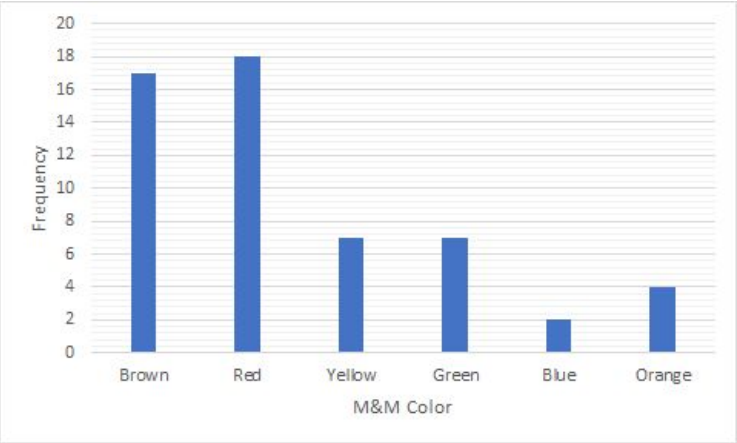
Frequency distribution for Andreas' bag of M&M'S

- A count of the number of times each outcome occurs is called a(n absolute) frequency distribution.
- This information can be conveyed in a table or in a plot.

TABLE

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

PLOT



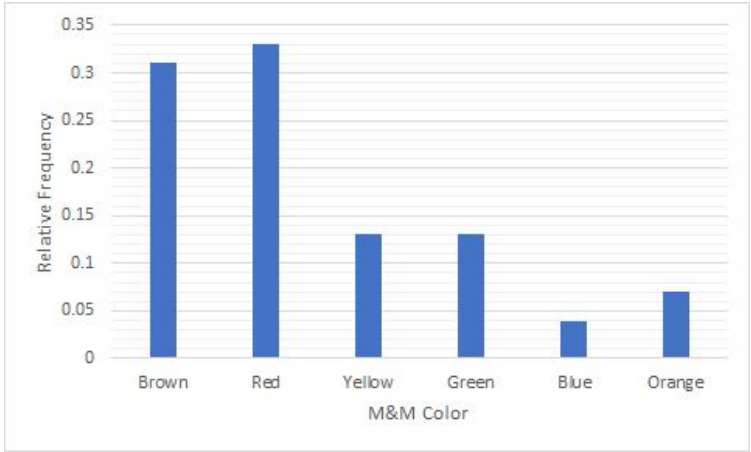
Relative frequency distribution for Andreas' M&M'S

- The proportion of times each outcome occurs is a relative frequency distribution.
- Count the number times each outcome occurs, and then divide the individual counts by the total count. This last step is called normalizing.

TABLE

Color	Frequency
Brown	17/55 = .31
Red	18/55 = .33
Yellow	7/55 = .13
Green	7/55 = .13
Blue	2/55 = .04
Orange	4/55 = .07

PLOT



What is a Probability Distribution?

- At the M&M factory, the machine is putting some number of each color of M&M into each bag.
  - The machine operates with some variability.
  - Sometimes, there are a lot of a color you love, and other times there are not so many of that same color.
- A distribution is a collection of outcomes and their relative counts, or proportions, which we interpret as their likelihoods.
  - Outcomes: The colors of M&Ms.
  - Likelihoods: The proportion with which we expect to see each outcome: i.e., each color M&M.

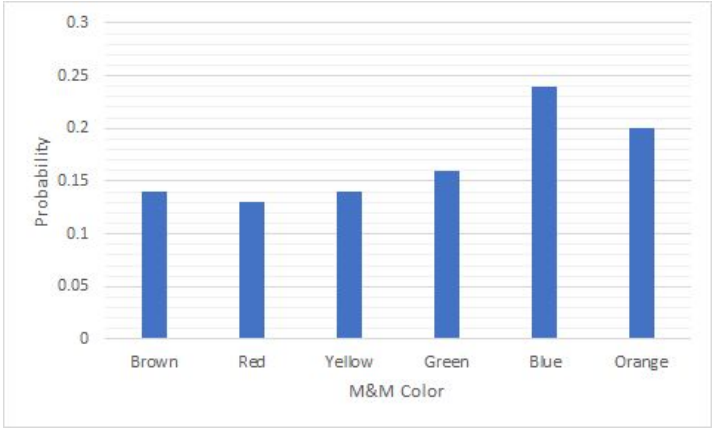
Probability distribution for M&M'S

- A probability distribution describes the chance of each possible outcome.
- Probabilities are never negative and their sum across outcomes is always 1.
- Thus, each outcome's probability is always bounded between 0 and 1, inclusive.

TABLE

Color	Probability
Brown	.14
Red	.13
Yellow	.14
Green	.16
Blue	.24
Orange	.2

PLOT

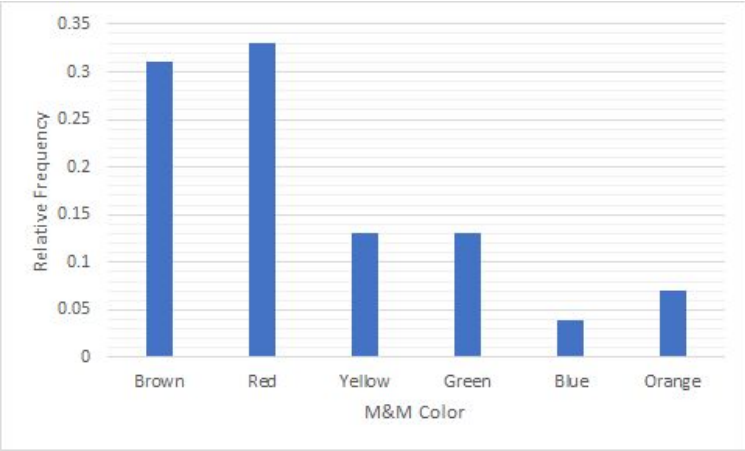


Source

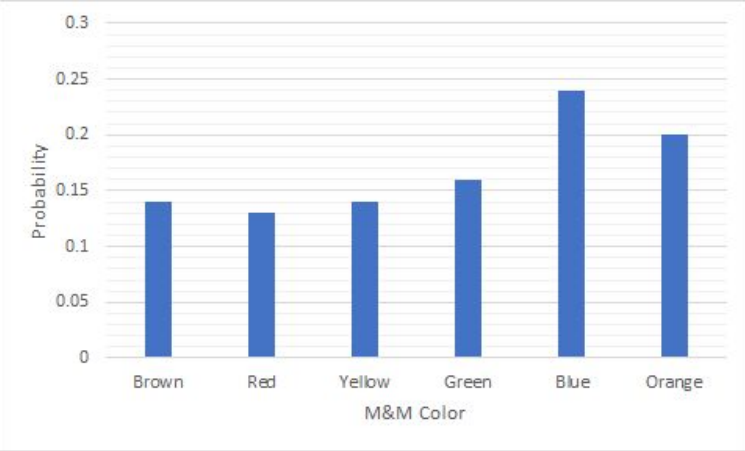
Sample vs. true distribution

- Relative frequency distribution of Andreas' bag
- "True" probability distribution set by M&M'S manufacturers

SAMPLE



TRUE



Error

- Blue M&Ms are the most common color, with probability 0.24.
- In the sample, blue was observed with relative frequency 0.10.
- Error can be calculated as (.10 - .24)/.24 x 100% = -58.33%.
- There were 58.33% fewer blue M&M’s than expected.

The Law of Large Numbers

- The relative frequency distribution of one bag (i.e., a sample) can differ from the true probability distribution.
- But in general, very large bags of M&M'S will mimic the proportions set by M&M'S manufacturers.
  - This is called the Law of Large Numbers.
  - Imagine flipping a fair coin 10 times; you might see 7 heads.
  - But by flipping the coin 100 times, you'll likely see closer to 50 heads (e.g., 45 heads).
- The average of the relative frequency distributions of more and more (small) bags of M&M's should approach the true underlying distribution.

Histograms

Population study

- The population under study is the 75 students in our class.
- We asked you how many languages you speak (besides English).
- Here were your responses:

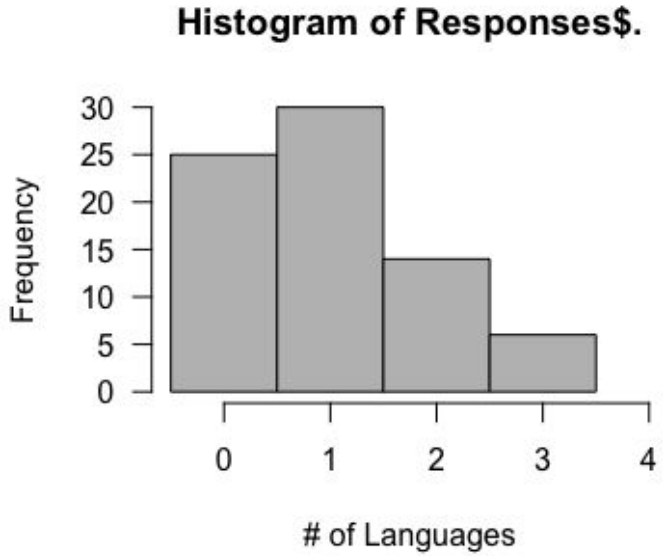
1 3 1 1 0 1 1 1 0 2 1 1 1 0 0 1 0 0 1 0 0 0 3 1 1 3 1 2 3 0 1 0 1 2 1 2 0 1
2 1 2 0 1 1 1 2 0 1 2 1 2 2 0 1 0 0 3 0 3 0 1 2 1 2 0 1 0 1 0 0 1 0 0 2 2

- Each response is a measurement, or an outcome

Histograms of discrete, quantitative data

- A histogram is a plot of a frequency distribution, when the data are numerical (this description is necessary, not sufficient; formal definition coming soon)

# Languages	Frequency
0	25
1	30
2	14
3	6
Grand Total	**75**



Histograms of continuous, quantitative data

- Frequency distributions can also be made of continuous data, by clumping similar values into bins (or buckets)
- Frequency distributions of binned, quantitative data can be displayed in tables, or they can be plotted, as histograms

Frequency table of (two) crew teams' weights

- This data set contains the weights in pounds of the crews participating in the Oxford Cambridge boat race in 1992
- The first table (raw data) has the first 15 rows of these data
- The second table (frequency table) was created by binning the data into intervals of size 10

Raw Data

	Weight
1	188.5
2	183.0
3	194.5
4	185.0
5	214.0
6	203.5
7	186.0
8	178.5
9	109.0
10	186.0
11	184.5
12	204.0
13	184.5
14	195.5
15	202.5

Frequency table

	bin	frequency
1	105 - 115	2
2	115 - 125	0
3	125 - 135	0
4	135 - 145	0
5	145 - 155	0
6	155 - 165	0
7	165 - 175	1
8	175 - 185	6
9	185 - 195	4
10	195 - 205	4
11	205 - 215	1

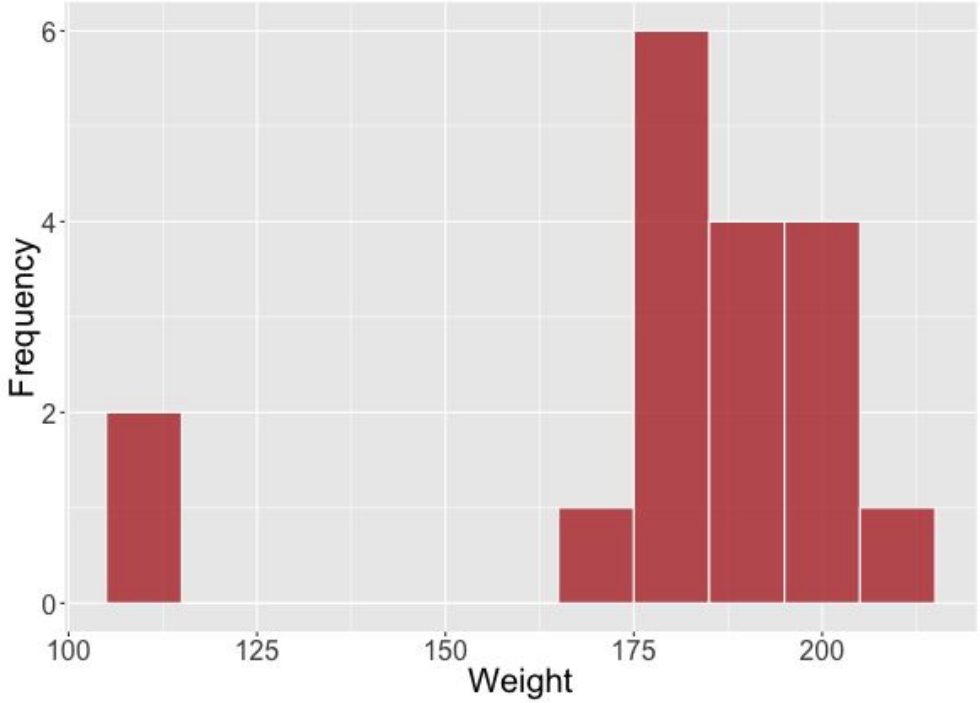
Histogram of the weights of (two) crew teams

Raw Data

	Weight
1	188.5
2	183.0
3	194.5
4	185.0
5	214.0
6	203.5
7	186.0
8	178.5
9	109.0
10	186.0
11	184.5
12	204.0
13	184.5
14	195.5
15	202.5

Frequency table

	bin	frequency
1	105 - 115	2
2	115 - 125	0
3	125 - 135	0
4	135 - 145	0
5	145 - 155	0
6	155 - 165	0
7	165 - 175	1
8	175 - 185	6
9	185 - 195	4
10	195 - 205	4
11	205 - 215	1



Relative vs. Absolute Frequencies

- We obtain relative frequencies by dividing absolute frequencies by the sample size. This is called normalization.
- So, relative frequencies are proportions; they tell us what percentage of the data set falls into each bin.
- Relative frequencies are easier to compare with one another.

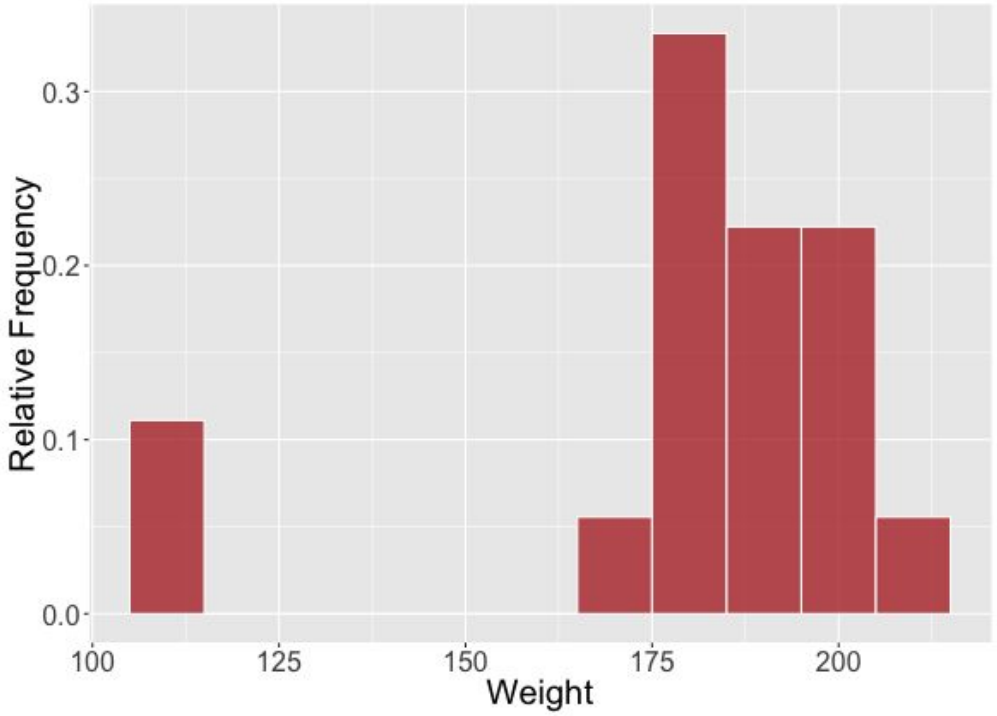
Normalized histogram of the teams' weights

Raw Data

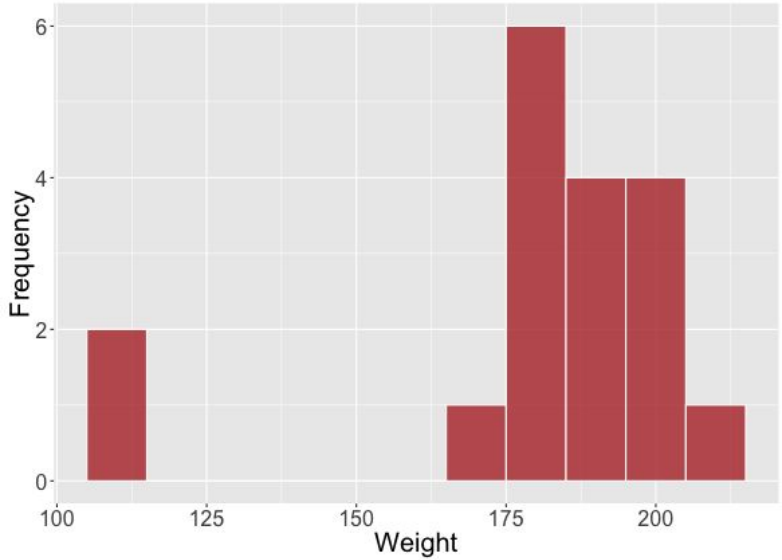
	Weight
1	188.5
2	183.0
3	194.5
4	185.0
5	214.0
6	203.5
7	186.0
8	178.5
9	109.0
10	186.0
11	184.5
12	204.0
13	184.5
14	195.5
15	202.5

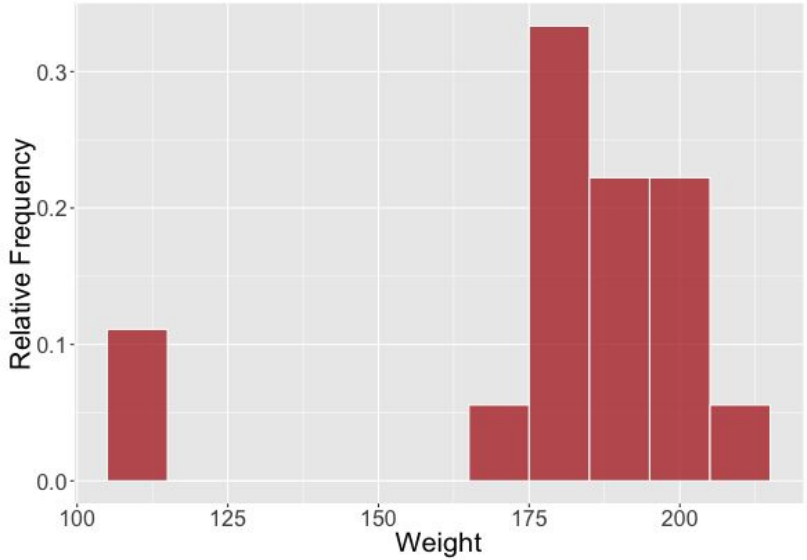
Relative Frequency table

	range	relative_frequency
1	105 - 115	0.0111
2	115 - 125	0.0000
3	125 - 135	0.0000
4	135 - 145	0.0000
5	145 - 155	0.0000
6	155 - 165	0.0000
7	165 - 175	0.0056
8	175 - 185	0.0333
9	185 - 195	0.0222
10	195 - 205	0.0222
11	205 - 215	0.0056



Normalized vs. unnormalized histograms





Observe that the shape of the histograms is the same, whether we plot absolute (left) or relative (right) frequencies. Only the scale (y-axis) differs.

Histograms

- Histograms are bar charts for continuous, quantitative data
  - Each bar is associated with a range of neighboring values (i.e., buckets or bins)
- Histograms depict
  - Center
  - Spread
  - Skewness
  - Outliers
  - Modes

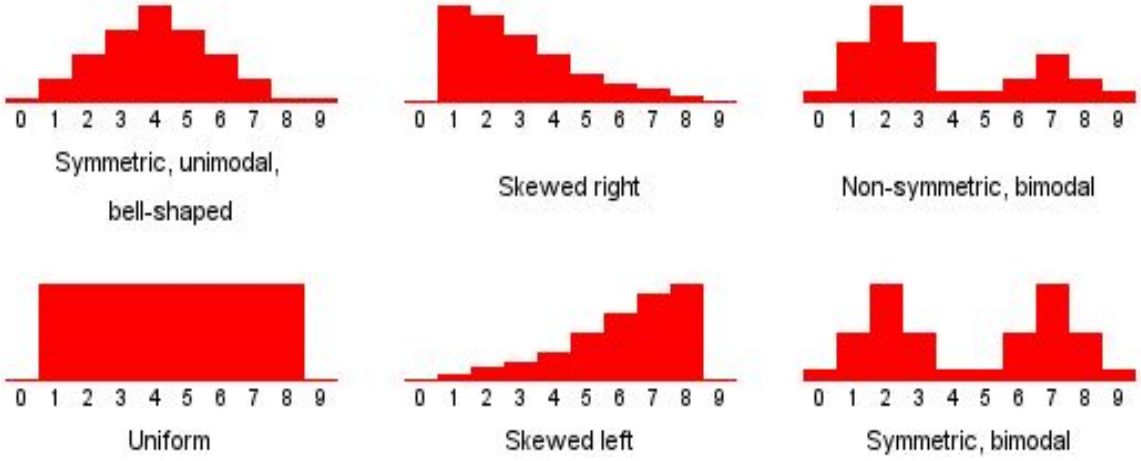


Image source