Basic terminology

- **Cases** are the objects described by a set of data
- A **variable** is a characteristic of a case

Example: Books by J.K. Rowling:

Book	Number of Pages
Harry Potter & Sorceror's Pet Rock	223
Harry Potter & Chamber of Teapots	251
Harry Potter & Prisoner of Extended Family	217
...	...

Types of variables

- A **categorical variable** is a grouping attribute
- A **quantitative variable** is a numerical attribute

Student	Year	School	Midtm1	Midtm2	Hwk	Final
Smith	1	A&S	78	85	88	86
Johnson	3	Busi	82	90	65	72
Hardy	1	A&S	65	88	98	82
Klein	CE	PubHlth	85	89	93	74
Parry	2	Jour	77	98	76	57
Watkins	4	A&S	45	76	56	81
Allen	3	Busi	87	82	90	88
Abbas	2	Busi	97	84	88	91

Types of quantitative variables

- A **discrete** quantitative variable is from a countable list of numbers (usually counting)
- A **continuous** quantitative variable is from *any* number in a certain range (usually measuring)

Types of quantitative variables

- A **discrete** quantitative variable is from a countable list of numbers (usually counting)
- A **continuous** quantitative variable is from *any* number in a certain range (usually measuring)

Discrete examples: Points scored in soccer matches, number of stores in a region, ...

Continuous examples: Weight of food products, carbon dioxide ppm in the air

Types of quantitative variables

- A **discrete** quantitative variable is from a countable list of numbers (usually counting)
- A **continuous** quantitative variable is from *any* number in a certain range (usually measuring)

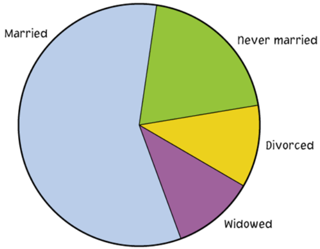
Discrete examples: Points scored in soccer matches, number of stores in a region, . . .

Continuous examples: Weight of food products, carbon dioxide ppm in the air

Warning: Distinction not always obvious. Depends on setting.

Describing data with charts: categorical data

Categorical data: Bar graph (left) or pie chart (right)

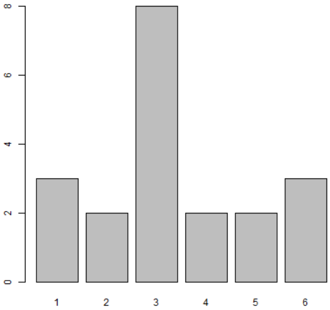


Q: Can we display quantitative data with a pie chart or bar graph?

Describing data with charts: quantitative data

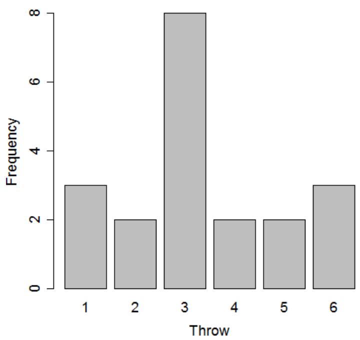
Discrete quantitative data: Bar graph will work and is best if number of possible outcomes is small.

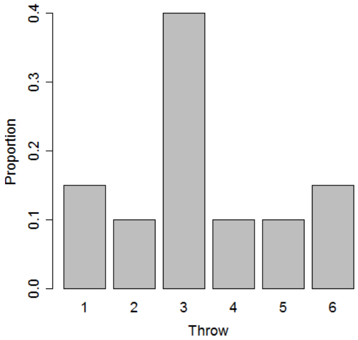
Example: Dice rolls.



Describing data with charts: quantitative data

Two types of bar graphs: Frequency bar graph (left) and proportion bar graph (right)





Describing data with charts: histograms

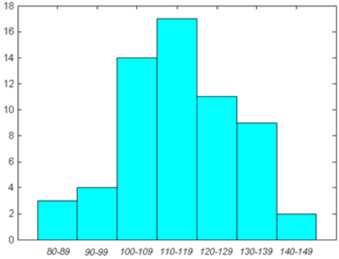
General quantitative data: A **histogram** is a *continuous* barplot for *ranges* of a variable.

TABLE 1.1

IQ test scores for 60 randomly chosen fifth-grade students

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

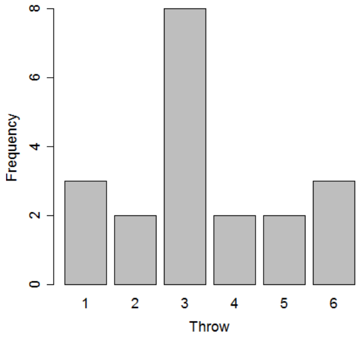
Class	Count
80 – 89	3
90 – 99	4
100 – 109	14
110 – 119	17
120 – 129	11
130 – 139	9
140 - 149	2

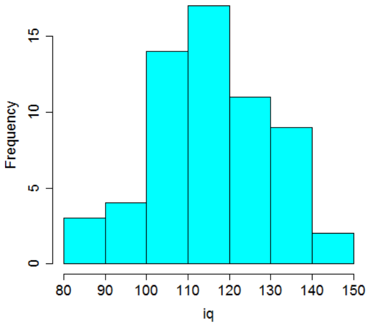


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
-
-

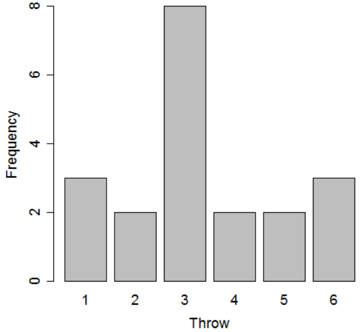


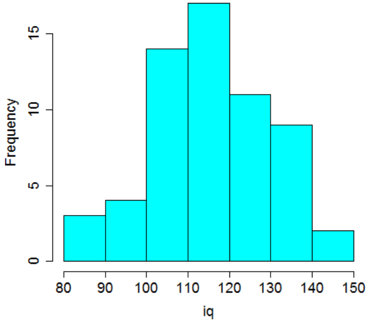


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
- Bar graphs have spaces between columns; histograms do not
-

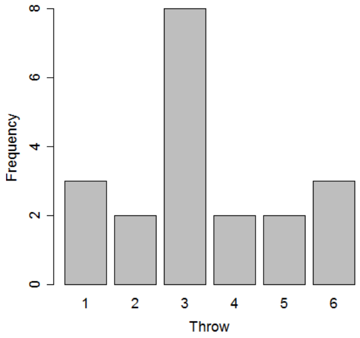


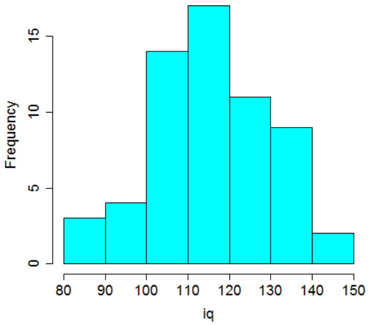


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
- Bar graphs have spaces between columns; histograms do not
- Both have frequency versions

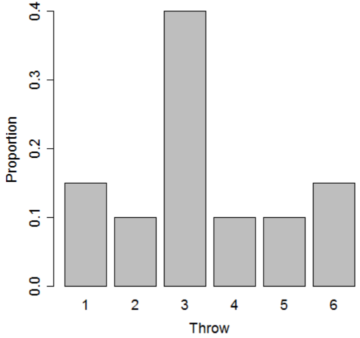


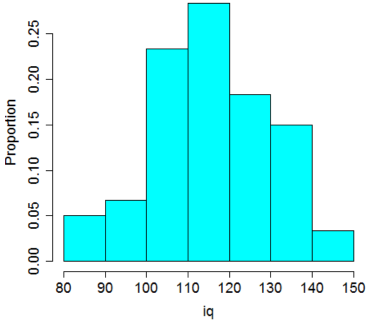


Describing data with charts: bar graphs vs. histograms

Things to notice:

- Bar graphs separate by *value*; histograms separate by *range*.
- Bar graphs have spaces between columns; histograms do not
- Both have frequency versions and both have proportion versions





Describing data with numbers

First, notation:

- Unknown variables usually written with letters: x, y, z, etc.
- If you know I have 10 variables, but you don't know what they are, you could write:

$$x_1, x_2, \ldots, x_{10}$$

  as placeholders.
- Number of variables (or cases) usually written "n" for "number"

Describing data with numbers

Summation notation:

- If you wanted to write the sum, you could write:

$$x_1 + x_2 + \ldots + x_{10}$$

- **Summation notation**:

$$\sum x_i \quad \text{just means} \quad x_1 + x_2 + \ldots + x_{10}$$

- Also holds for expressions involving each x_1, x_2, etc.:

$$\sum x_i^2 \quad \text{just means} \quad x_1^2 + x_2^2 + \ldots + x_{10}^2$$

Describing data with numbers: measures of center

Mean or average: sum of the variables divided by n:

$$\text{Mean} = \bar{x} = \frac{\sum x_i}{n}$$

Median: center *ordered* data point

- Order the data
- If n is odd, the median is the middle data point
- If n is even, the median is the average of the two middle data points

Mode: most occuring data point

Describing data with numbers: example

Example: Randomly sampled years in school from this class:

$$1, 1, 2, 1, 1, 1, 3, 2, 1, 1$$

Describing data with numbers: example

Example: Randomly sampled years in school from this class:

$$1, 1, 2, 1, 1, 1, 3, 2, 1, 1$$

Mean = 1.4, Median = 1, Mode = 1.

Describing data with numbers: quartiles

Quartiles are numbers that divide data into fourths.

Median is second quartile. First and third quartiles are:

- **Q1**: The median of data to the left of the median
- **Q3**: The median of data to the right of the median

Example: Randomly sampled years in school from this class:

$$1, 1, 1, 1, \quad 1, \quad 1, 2, 2, 3$$

Describing data with numbers: quartiles

Quartiles are numbers that divide data into fourths.

Median is second quartile. First and third quartiles are:

- **Q1**: The median of data to the left of the median
- **Q3**: The median of data to the right of the median

Example: Randomly sampled years in school from this class:

$$1, 1, 1, 1, \quad 1, \quad 1, 2, 2, 3$$

Q1 $= 1$; Q3 $= 2$

Describing data with numbers: *percentiles*

The p-th **percentile** of the data has $p\%$ of the data below it

- To find p-th percentile:
  1. Order data
  2. Count up until you have *no more than* $p\%$ of the data
  3. The number you stop on is the p-th percentile
- Note: Q1$\approx p_{.25}$, Q2$\approx p_{.50}$, and Q3$\approx p_{.75}$

Describing data with numbers: five number summary

The **five number summary** is the following list:

Minimum, Q1, Median (Q2), Q3, Maximum

Gives concise but informative "image" of the data

Example: Randomly sampled years in school from this class:

$$1, 1, (1), 1, 1, \quad 1, \quad 1, 2, (2), 2, 3$$

Five number summary is 1, 1, 1, 2, 3

Describing data with numbers: measures of spread

Definitions first, explanations later:

- **Variance** of the data (written s^2):

$$\frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_{10} - \bar{x})^2}{n-1}$$

- **Standard error** of the data (written s):

$$\sqrt{\frac{(\sum x_i - \bar{x})^2}{n-1}}$$

Describing data with numbers: measures of spread

$$\text{Variance} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_{10} - \bar{x})^2}{n - 1}$$

...essentially an average of the *squared* distances from the mean.

- So more *spread out* data will have higher variance

$$\text{Standard Error} = \sqrt{Variance}$$

- Since sum terms are squared, this is on the scale of the data

Measures of spread: example

Measures of spread: example

Luke vs.

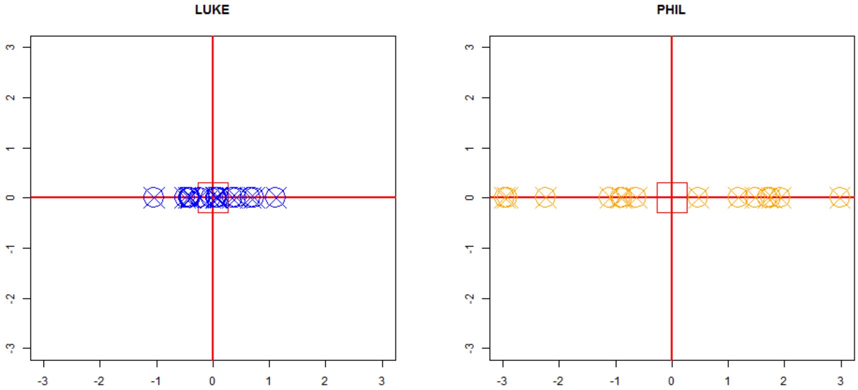
Measures of spread: example

Luke vs. Phil

Measures of spread: example

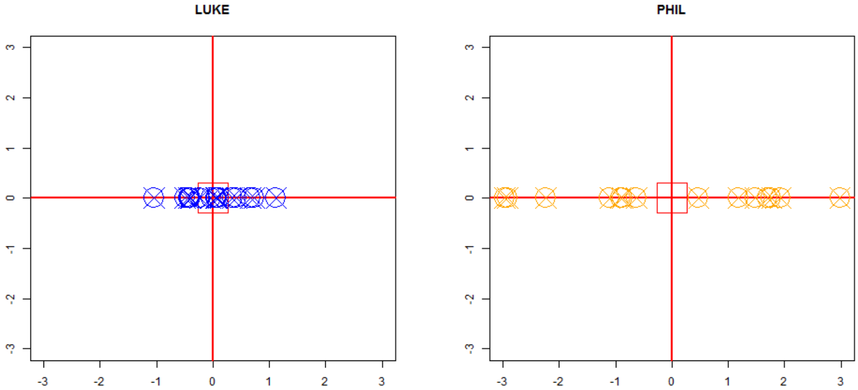
Luke vs. Phil (guy who thinks Luke is kind of a derpwad)

Measures of spread: example



Which has the higher variance?

Measures of spread: example



Luke var.=0.32; Phil var.=3.48! But Luke mean = 0.03 and Phil mean = -0.07: almost the same!

Describing data with numbers: data transformations

- Sometimes of interest to transform data
- **Linear transformation**: equation of the form

$$x_{\text{new}} = a + bx$$

- We are taking each data point x to a new data point x_{new}

Example: Celsius to Fahrenheit:

$$x_{\text{new}} = \frac{9}{5}x + 32$$

Describing data with numbers: data transformations

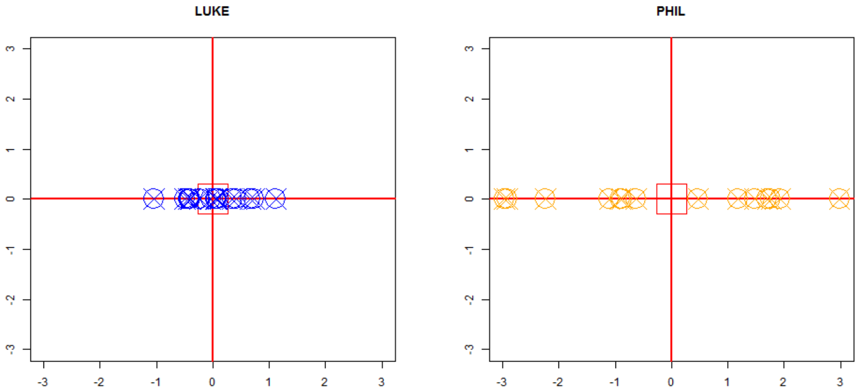
$$x_{\text{new}} = a + bx$$

Facts about data transformations:

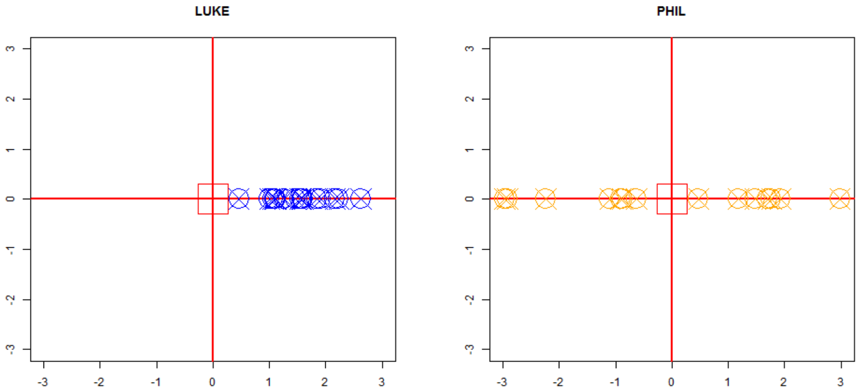
- Mean $\bar{x}_{\text{new}}$ is equal to $b\bar{x} + a$
- Standard deviation s_{new} is equal to $b \cdot s$
- Variance s^2_{new} is equal to $b^2 \cdot s$

Why no a for variance? Shifting the data will shift the center (mean). But the distances from the mean do not change.

Data transformation: illustration



Data transformation: illustration



Describing data with numbers: review

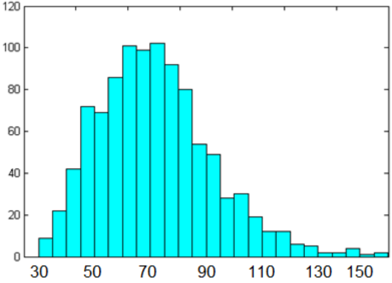
- **Measures of center**: Mean, median, mode
- **Measures of distribution**: Q1, Q2, min, max (five-number summary)
- **Measures of spread**: Variance, standard deviation

Combining concepts: modes and histograms

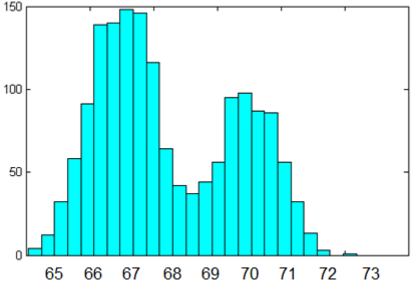
Terminology: *Unimodal* data vs. *bimodal* data

- ⋆ In general, we can say *multi*-modal

Unimodal data:



Bimodal data:

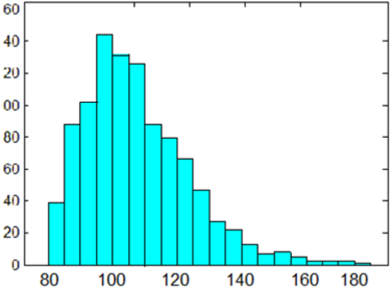


Combining concepts: distribution shape

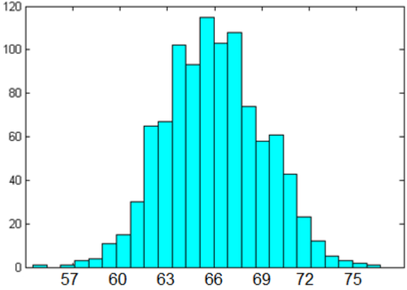
Terminology: *Skewed* data vs. *symmetric* data

- ⋆ Skew is in the direction of the "longer" side

Right-skewed data:



Symmetric data:



Combining concepts: mean, median, skew

Why do we need two measures of spread, mean and median? They can be quite different for skewed data.

Example: Recall, mean of the following is 1.4 and median is 1:

$$1, 1, 2, 1, 1, 1, 3, 2, 1, 1$$

Now, add two seniors to the class:

$$1, 1, 2, 1, 1, 1, 3, 2, 1, 1, 4, 4$$

How do the mean and median change? This is a good example of a skewed distribution; mean and median give very different info.

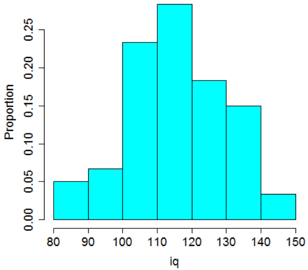
Combining concepts: mean, median, skew

Facts in general:

- In right-skewed data, mean is larger than median
  - ⋆ High values pull mean up; median stays the same
- In left-skewed data, mean is smaller than median
  - ⋆ Low values pull mean down; median stays the same

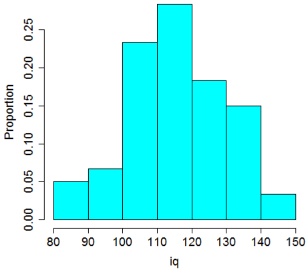
Combining concepts: percentiles and histograms

In-class exercise: Consider the histogram of IQ scores:



Combining concepts: percentiles and histograms

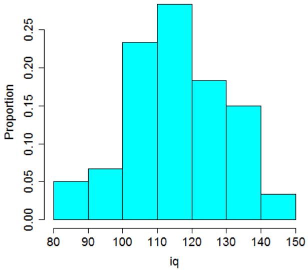
In-class exercise: Consider the histogram of IQ scores:



1. Between which two numbers is the median?

Combining concepts: percentiles and histograms

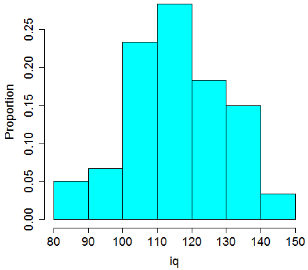
In-class exercise: Consider the histogram of IQ scores:



1. Between which two numbers is the median? A: 110 and 120

Combining concepts: percentiles and histograms

In-class exercise: Consider the histogram of IQ scores:

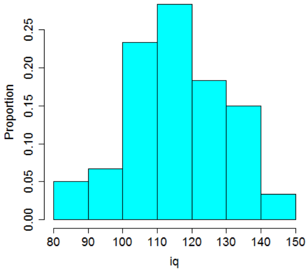


1. Between which two numbers is the median? A: 110 and 120

2. Between which two numbers is the 30% percentile?

Combining concepts: percentiles and histograms

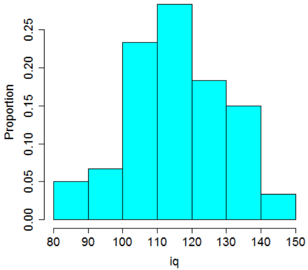
In-class exercise: Consider the histogram of IQ scores:



1. Between which two numbers is the median? A: 110 and 120
2. Between which two numbers is the 30% percentile? A: 100 and 110

Combining concepts: percentiles and histograms

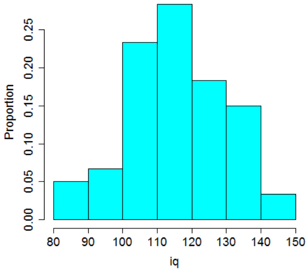
In-class exercise: Consider the histogram of IQ scores:



1. Between which two numbers is the median? A: 110 and 120
2. Between which two numbers is the 30% percentile? A: 100 and 110
3. Betwen which two numbers is the 90% percentile?

Combining concepts: percentiles and histograms

In-class exercise: Consider the histogram of IQ scores:



1. Between which two numbers is the median? A: 110 and 120
2. Between which two numbers is the 30% percentile? A: 100 and 110
3. Betwen which two numbers is the 90% percentile? A: 130 and 140