Week 2: Examining Numerical Data

Professor Kathryn Jacobs
Today's music theme: More 80's hits!

Quantitative Variables: General

"Numerical variables"

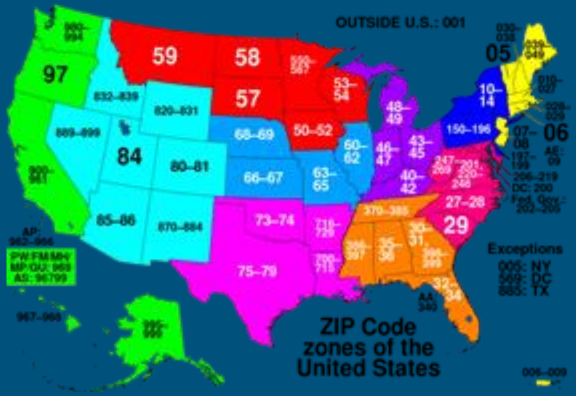
Mathy math

Measured

Not all numbers will count

Name	Company	Company _Number	Serving	Calories	Fat	Sodium	Carbs	Fiber	Sugars	Protein
AppleJacks	K	2	1.00	117	.6	143	27	.5	15.0	1.0
Boo Berry	G	1	1.00	118	.8	211	27	.1	14.0	1.0
Cap'n Crunch	Q	3	.75	144	2.1	269	31	1.1	16.0	1.3
Cinnamon Toast Crunch	G	1	.75	169	4.4	408	32	1.7	13.3	2.7
Cocoa Blasts	Q	3	1.00	130	1.2	135	29	.8	16.0	1.0
Cocoa Puffs	G	1	1.00	117	1.0	171	26	.8	14.0	1.0
Cookie Crisp	G	1	1.00	117	.9	178	26	.5	13.0	1.0
Corn Flakes	K	2	1.00	101	.1	202	24	.8	3.0	2.0
Corn Pops	K	2	1.00	117	.2	120	28	.3	15.0	1.0
Crispix	K	2	1.00	113	.3	229	26	.1	3.0	2.0
Crunchy Bran	Q	3	.75	120	1.3	309	31	6.4	8.0	1.3
Froot Loops	K	2	1.00	118	.9	150	26	.8	12.0	2.0
Frosted Mini-Wheats	K	2	1.00	175	.8	5	41	5.0	10.0	5.0
Golden Grahams	G	1	.75	149	1.3	359	33	1.3	14.7	2.7
Honey Nut Clusters	G	1	1.00	214	2.7	249	46	2.8	17.0	4.0
Honey Nut Heaven	Q	3	1.00	192	3.7	216	38	3.5	13.0	4.0
King Vitaman	Q	3	1.50	80	.7	173	17	.9	4.0	1.3
Kix	G	1	1.30	87	.5	205	20	.8	2.3	1.5
Life	Q	3	.75	160	1.9	219	33	2.7	8.0	4.0

Quantitative Variables: General

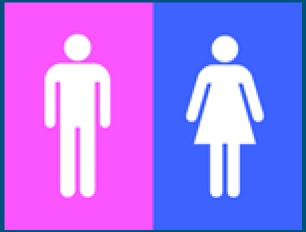


Numerical variables that are NOT quantitative:

Zip codes

Football jersey numbers

Male = 1, Female = 0



Quantitative Variables: THINK

3 examples of numerical variables that ARE quantitative

3 examples of numerical variables that ARE NOT quantitative

Quantitative Variables: PAIR

3 examples of numerical variables that ARE quantitative

3 examples of numerical variables that ARE NOT quantitative

Quantitative Variables: SHARE

3 examples of numerical variables that ARE quantitative

3 examples of numerical variables that ARE NOT quantitative

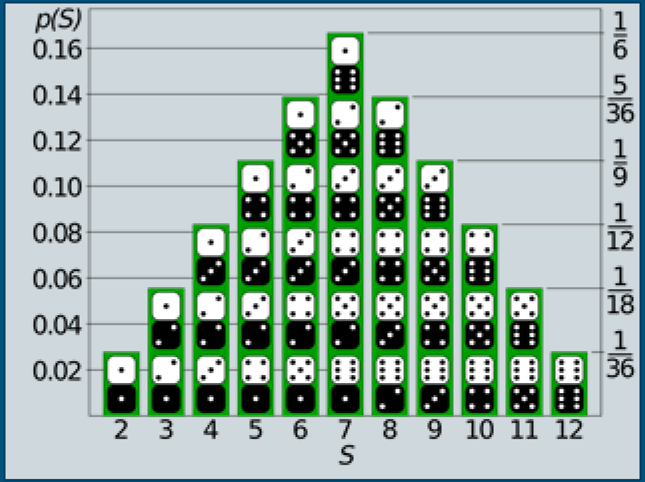
Distributions

What is a distribution?

“How likely different values are for a given variable”

How are different values *distributed* in the population

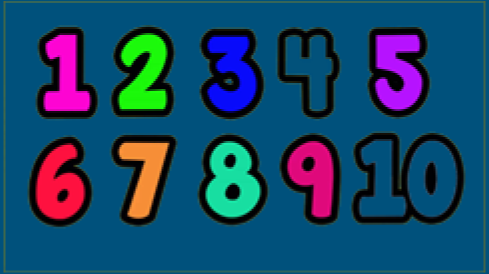
We can find this information using graphs, plots, and number summaries

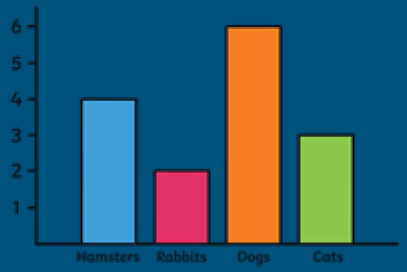


To Do

Summary Statistics

- Center: mean, median, mode
- Spread: standard deviation, range, IQR
- Percentiles
- 5 number summary

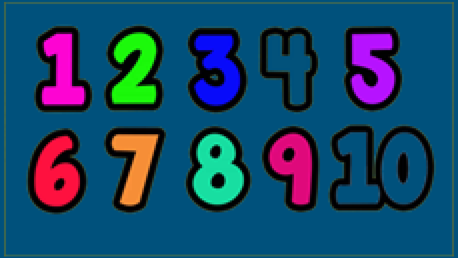




Visualization

- Dot plots
- Frequency tables
- Stem-and-leaf plots
- Histograms
- Box plots

Summary Statistics



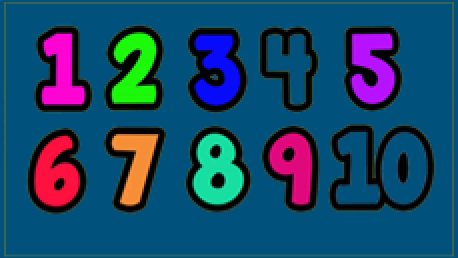
Center

Spread

Percentiles

5- number summary

Summary Statistics



Center

Spread

Percentiles

5- number summary

Measures of center: Mean, Median, Mode

Mean: mathematical average

Median: center number if numbers are ordered smallest-largest

Mode: most common number

Measures of center: Mean

Parameter: μ (mew)

Statistic: xbar

Add all numbers together, divide by total sample size/population

Population

$$\mu = \frac{\sum x}{N}$$

Sample

$$\bar{x} = \frac{\sum x}{n}$$

Measures of center: Median

Parameter: η (eta)

Statistic: $\tilde{x}$

1, 3, 3, **6**, 7, 8, 9

Median = $\underline{\underline{6}}$

1, 2, 3, **4**, **5**, 6, 8, 9

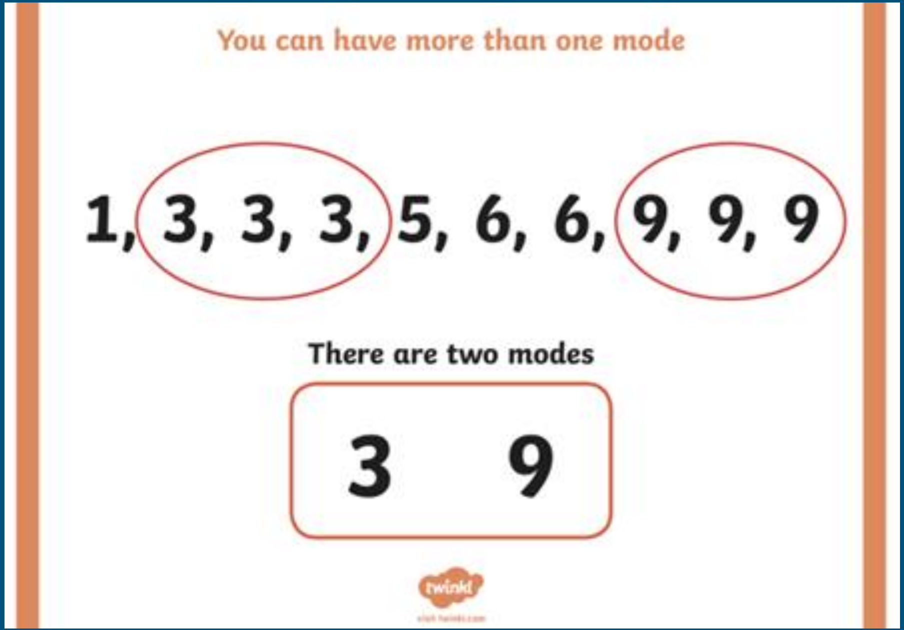
Median = $(4 + 5) \div 2$

= $\underline{\underline{4.5}}$

Measures of center: Mode

No symbol, we just say mode

Uniform, unimodal, bimodal, multimodal



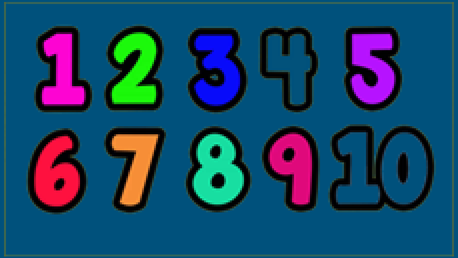
Measures of center: Resistance

Is mean, median, or mode best?

Depends on our data!

Some things, like extreme values, will affect some measures more than others

Summary Statistics

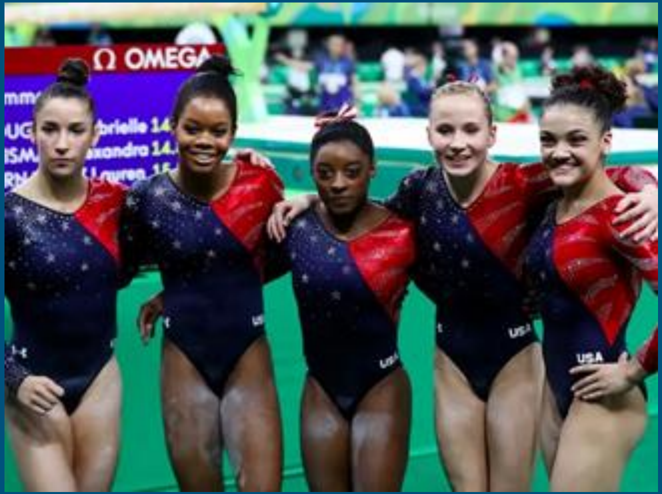


- Center
- Spread
- Percentiles
- 5- number summary

Spread: General

Spread describes how *varied* our distribution is

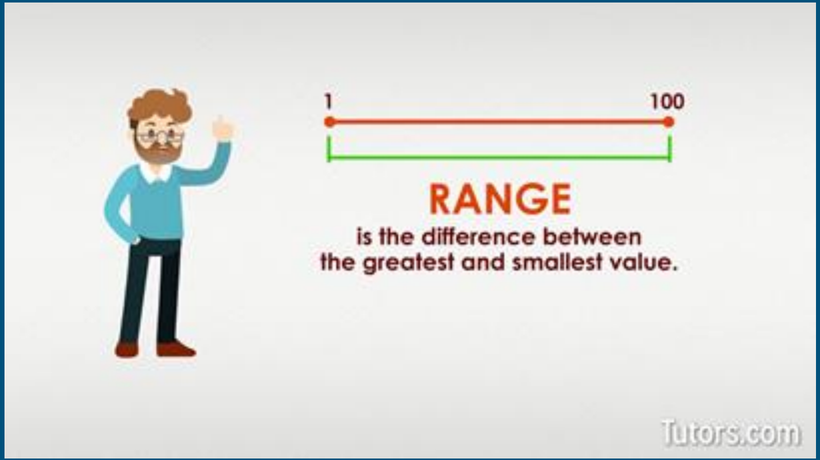
Are most of our values close together, or spread out?



Spread: Range

Very simple!

Largest value - smallest value



Spread: Standard deviation

Representation of just how varied a variable is: “average distance of a data point from the mean”

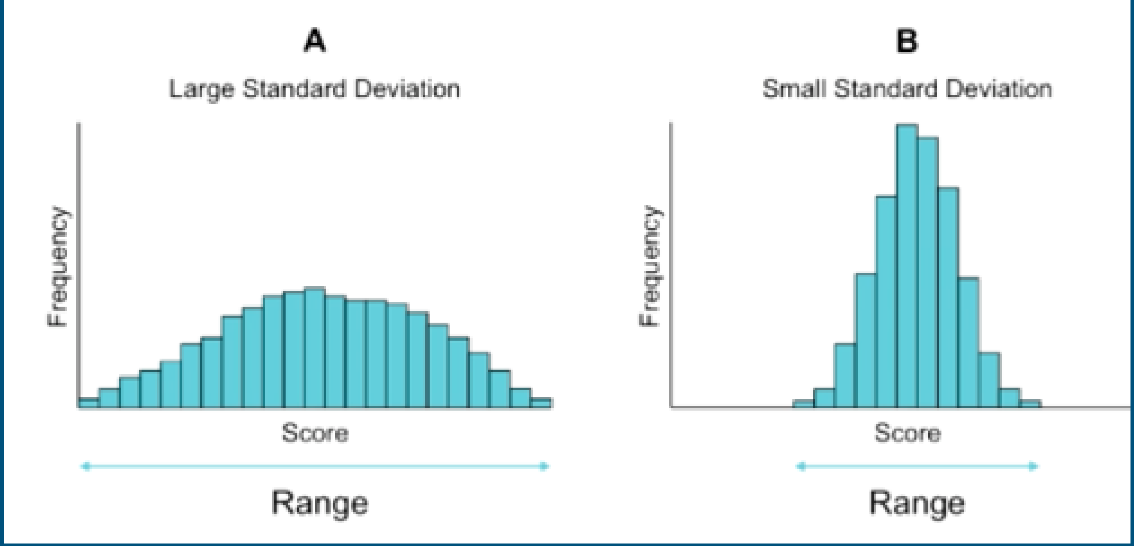
Large SD:

extremely varied data

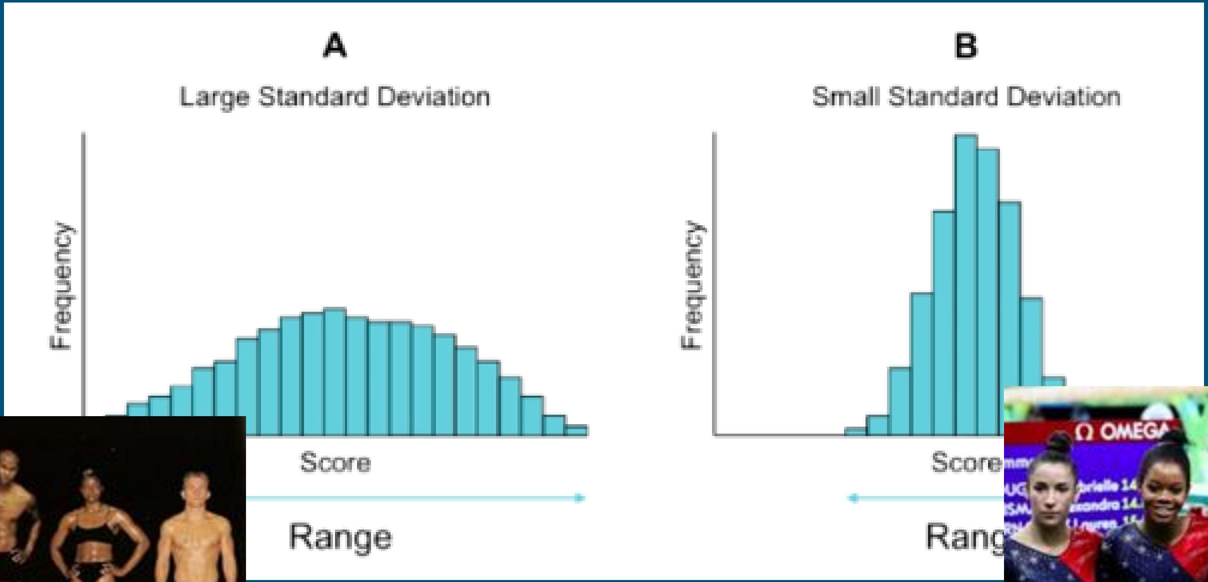
Small SD:

Data all clusters

close to mean



Spread: Standard deviation



Spread: Standard deviation

Population	Sample
$\sigma = \sqrt{\dfrac{\Sigma(x_i-\mu)^2}{n}}$	$S = \sqrt{\dfrac{\Sigma(x_i-\overline{x})^2}{n-1}}$
μ - Population Average xi - Individual Population Value n - Total Number of Population	x - Sample Average x_i - Individual Population Value n - Total Number of Sample

Standard deviation: Practice

Let's calculate the standard deviation of this data set:

[1, 1, 3, 5, 5]

Mean = 3

How extreme is the value of 5?

Sample

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

$\bar{x}$ - Sample Average
x_i - Individual Population Value
n - Total Number of Sample

Spread: Variance

Variance = standard deviation squared

Represents TOTAL amount of variation within a data set

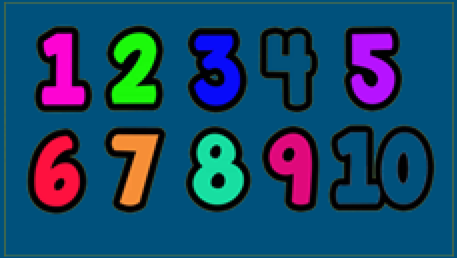
Doesn’t mean much by itself- used in formulas for other things

Population	Sample
$\sigma^2 = \frac{\Sigma(x_i-\mu)^2}{n}$	$S^2 = \frac{\Sigma(x_i-\overline{x})^2}{n-1}$
μ - Population Average xi - Individual Population Value n - Total Number of Population σ^2 - Variance of Population	X - Sample Average x_i - Individual Population Value n - Total Number of Sample S^2 - Variance of Sample

Notation: for reference

Name	Population Parameters	Sample Statistics
Mean	μ	$\overline{X}$
Median	η	$\tilde{X}$
Mode	No symbol	No symbol
Range	R	R
Variance	σ^2	s^2
Standard Deviation	σ	s
Sample Size	N	n
Estimates	$\hat{\sigma}$	n/a

Summary Statistics



Center

Spread

Percentiles

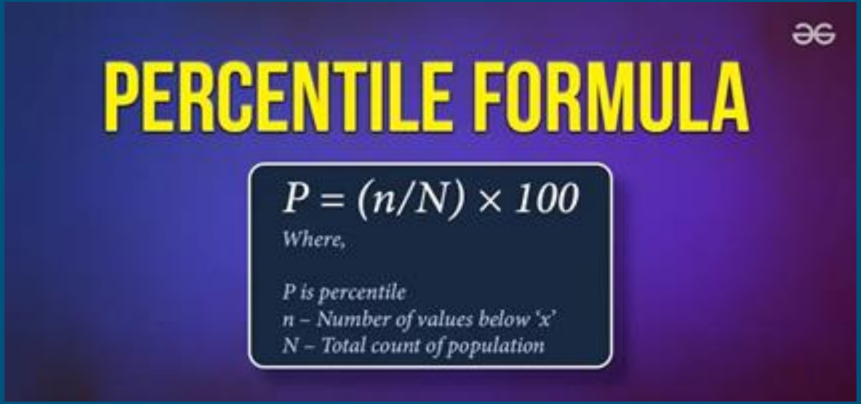
5- number summary

Percentiles

Compare a single value to the entire data set

“This number is bigger than 80% of the rest of the data”

Median = 50th percentile



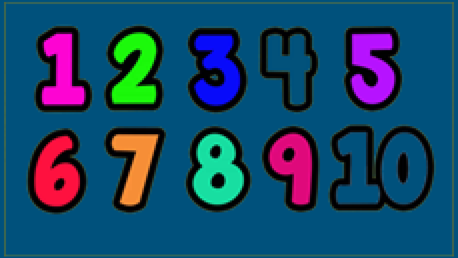
Percentiles: Example

For the following sample, what percentile is a value of 4?

[3, 9, 4, 5, 5, 8, 2]

1. Order smallest-largest
2. Count values smaller than target number
3. Divide by total sample size

Summary Statistics



Center

Spread

Percentiles

5- number summary

5-number summary

5 Number Summary

Min = Smallest number

Q1 = Median of the first half of the data

Q2 = Median

Q3 = Median of the second half of the data

Max = Largest number

5-number summary

1. Order numbers smallest - largest
2. Find min and max
3. Find median
4. (numbers below median) median (numbers above median)
5. Find Q1 and Q3 using numbers in (parentheses)
6. [min, Q1, median, Q3, max]

IQR = Interquartile range = Q3 - Q1

5-number summary: IQR

Another way to measure spread of distribution

Q3 - Q1

Practice

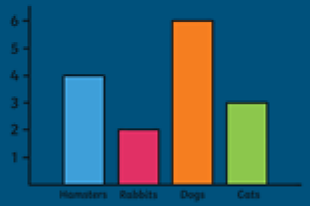
Find the mean, median, mode, 5-number summary, and IQR for the following data set:

[3, 5, 7, 7, 2, 4, 2, 2, 8, 6, 5, 7, 7, 7, 4]

What number is at the 80th percentile?

[2, 2, 2, 3, 4, 4, 5, 5, 6, 7, 7, 7, 7, 7, 8]

Visualization



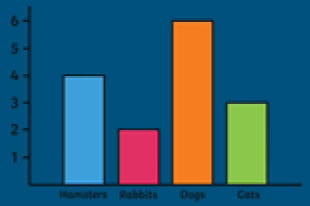
Dot plots

Frequency tables

Stem-and-leaf plots

Histograms

Box plots



Visualization

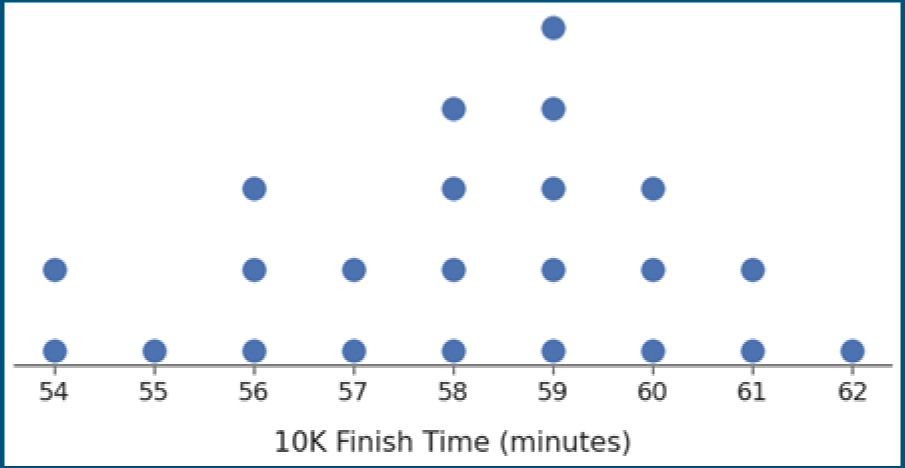
Dot plots

Frequency tables

Stem-and-leaf plots

Histograms

Box plots



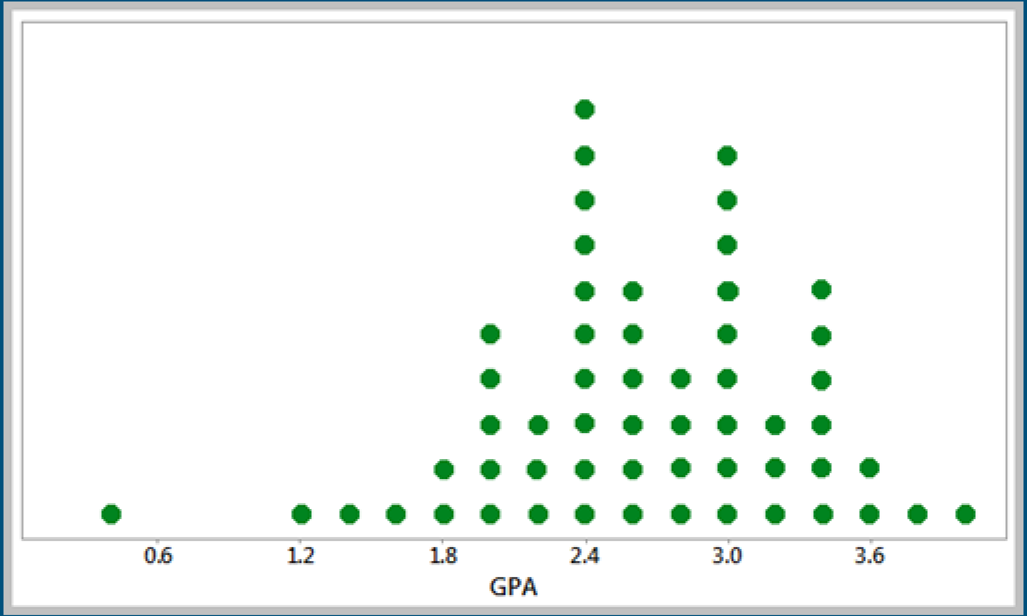
Dot Plots

Each dot represents 1 case

Only discrete data

Can be used to approximate mode(s)

Does this distribution have a mode?



Visualization

Dot plots

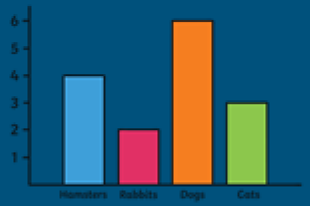
Frequency tables

Stem-and-leaf plots

Histograms

Box plots

Score	Frequency
6	2
7	3
8	7
9	7
10	1



Frequency tables

Let’s construct a frequency table

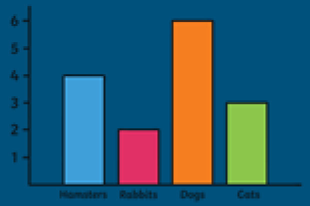
5, 7, 3, 10, 18, 10, 10, 5, 13, 13, 18

Frequency tables: Grouped data

What if we don't have any repeat values, or have lots and lots of values?

Age Group	Frequency	Percent
21-25	87	43.1
26-30	43	21.3
31-35	25	12.3
16-20	17	8.4
36-40	15	7.4
46-50	4	2.0
51-55	4	2.0
41-45	3	1.5
56-60	3	1.5
61+	1	0.5

Visualization



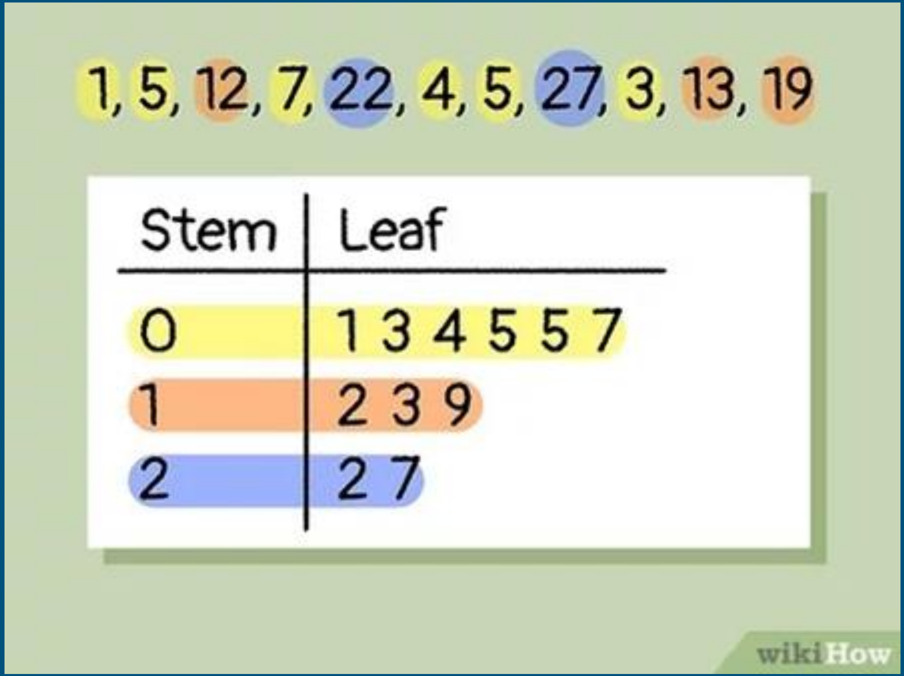
Dot plots

Frequency tables

Stem-and-leaf plots

Histograms

Box plots



Stem and Leaf plots

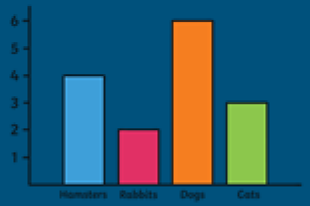
Good for small data sets

Data ordered - usually smallest to largest

What is our mode here?

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

Visualization



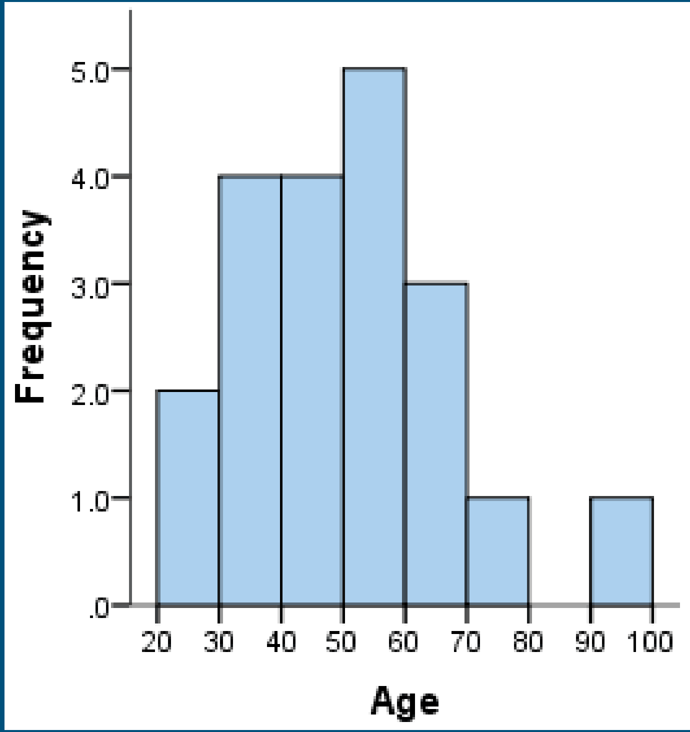
Dot plots

Frequency tables

Stem-and-leaf plots

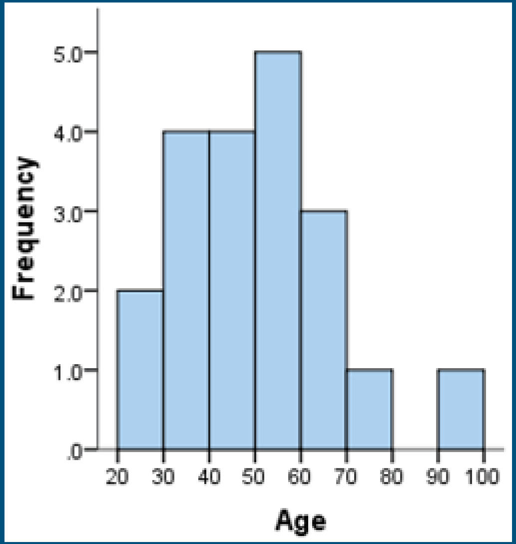
Histograms

Box plots



Histograms: General

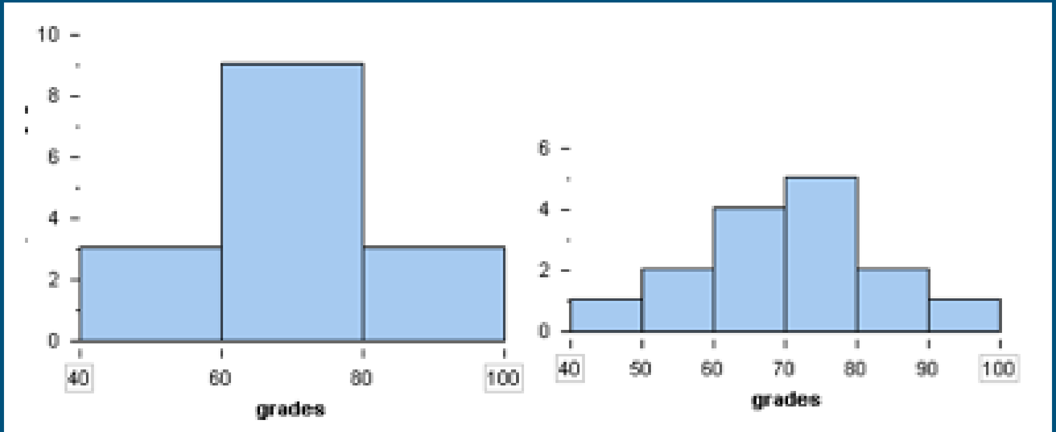
- Group continuous variables into ranges, or *bins*
- Frequency of cases in each bin is added up, and graphed along y axis



Histograms: Bin Width

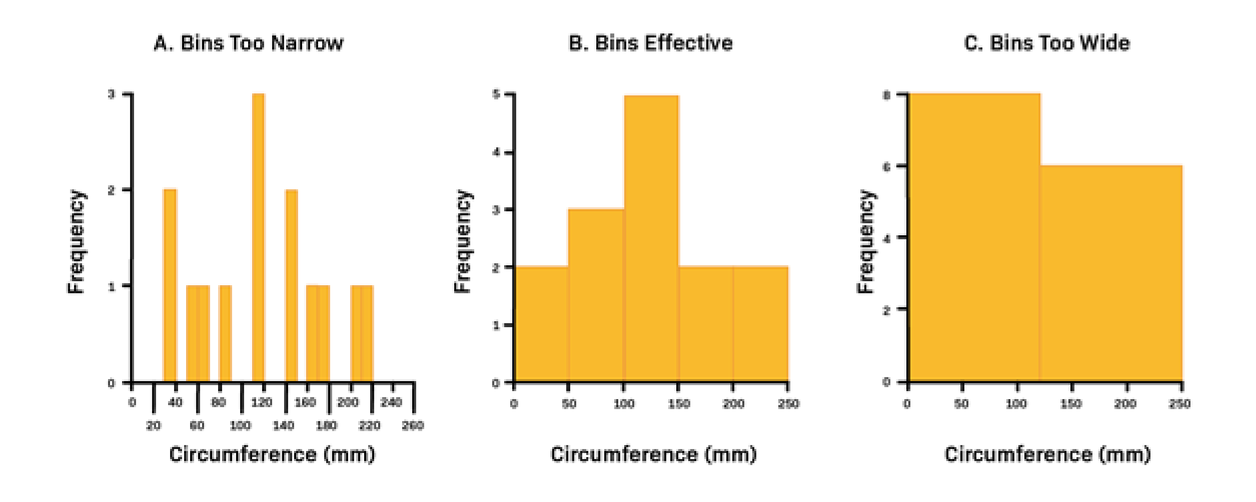
In general, more bins = better

Narrower bins give us more information



Histograms: Bin Width

Sometimes if bins are too narrow, we lose information

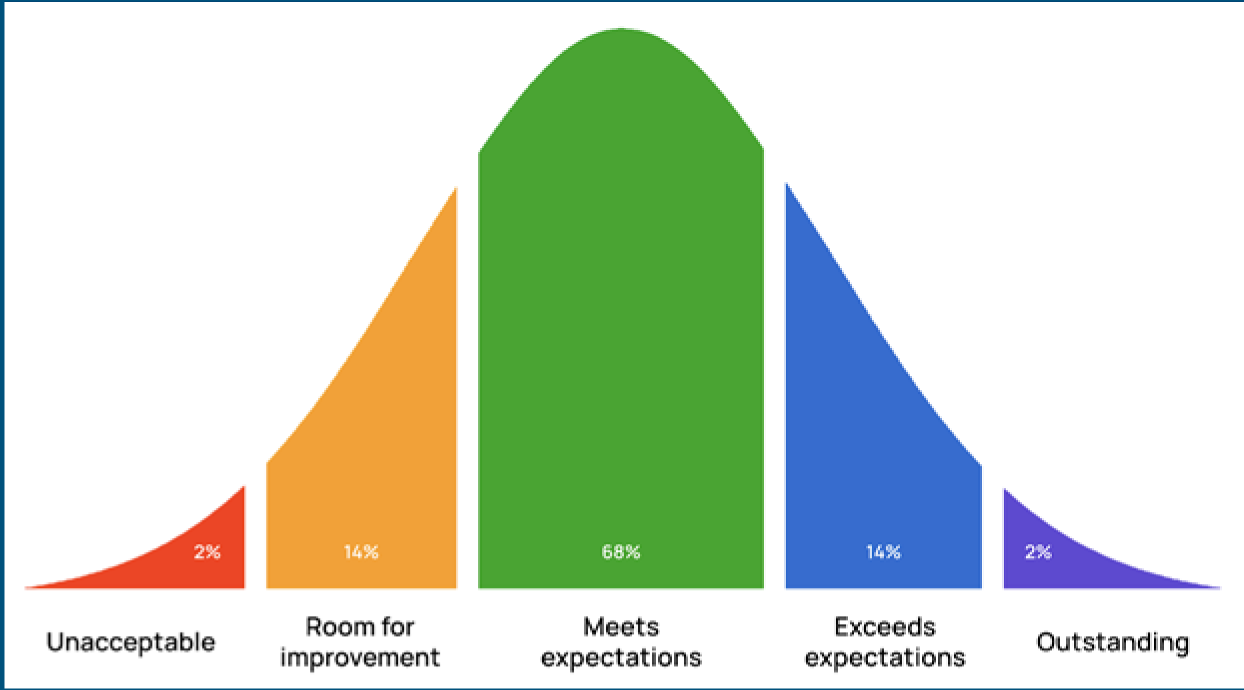


Histograms from: Frequency tables

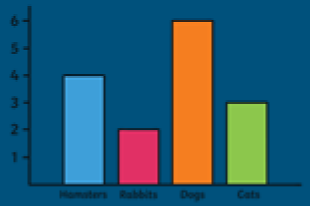
Let's make a histogram!

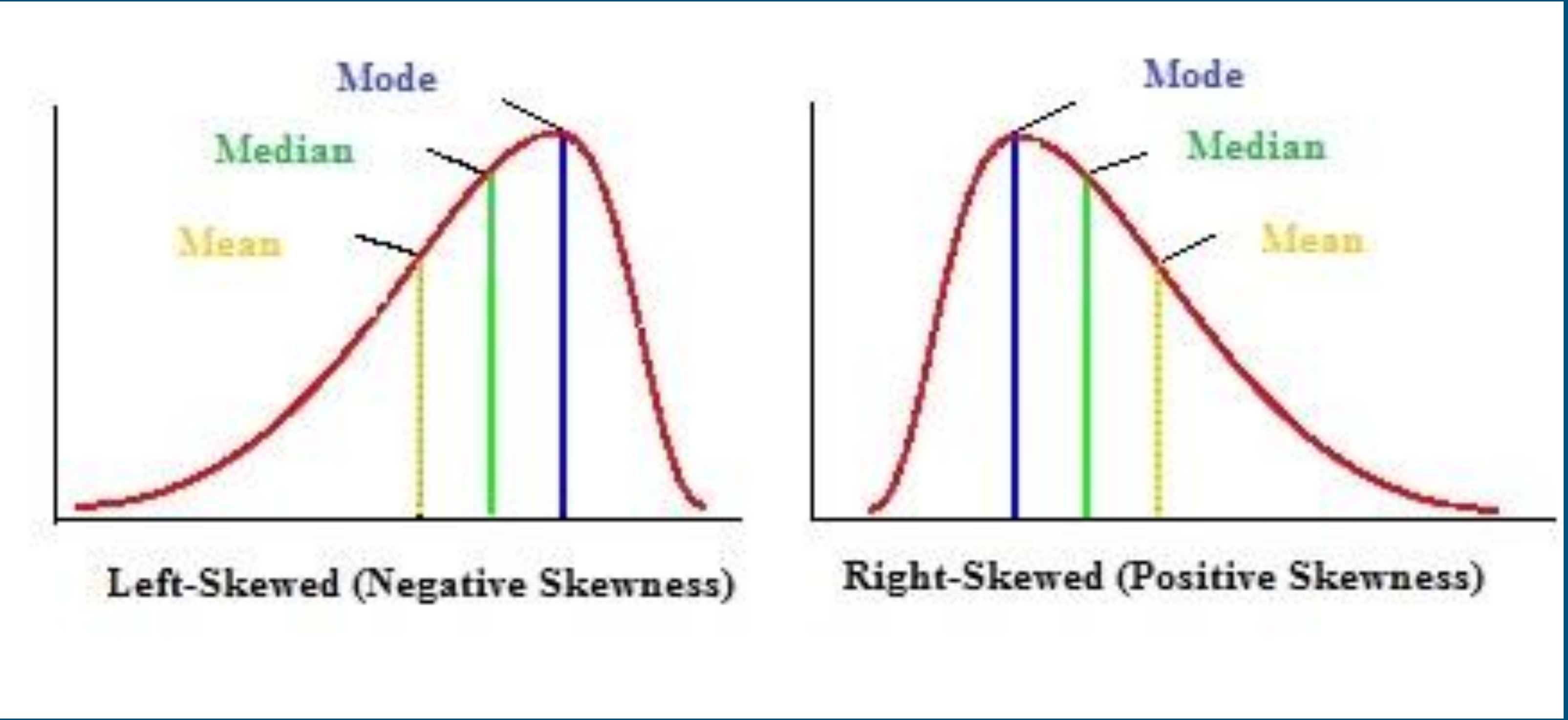
Age Group	Frequency	Percent
21-25	87	43.1
26-30	43	21.3
31-35	25	12.3
16-20	17	8.4
36-40	15	7.4
46-50	4	2.0
51-55	4	2.0
41-45	3	1.5
56-60	3	1.5
61+	1	0.5

Histograms: Normal curve

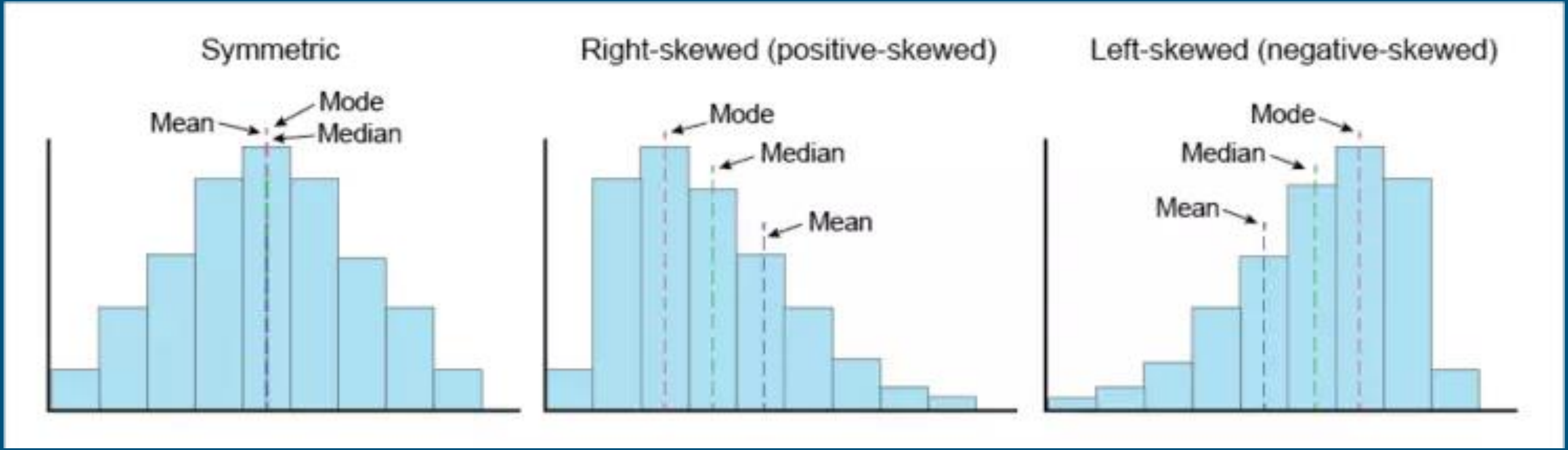


Histograms: Skew

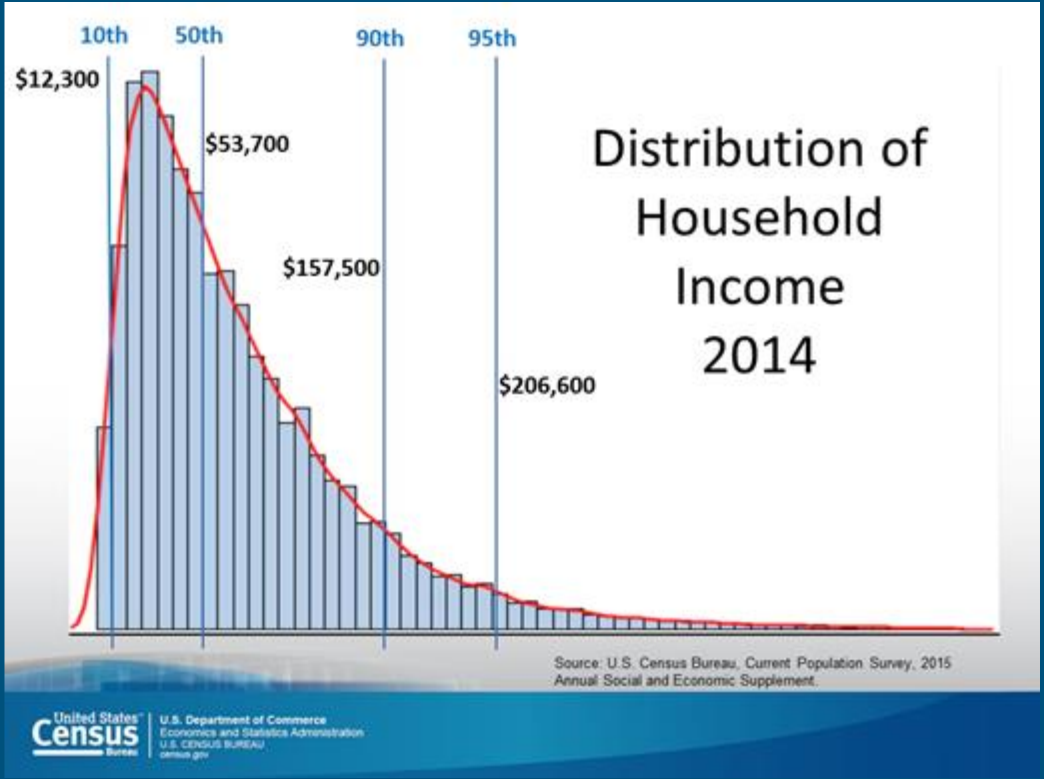




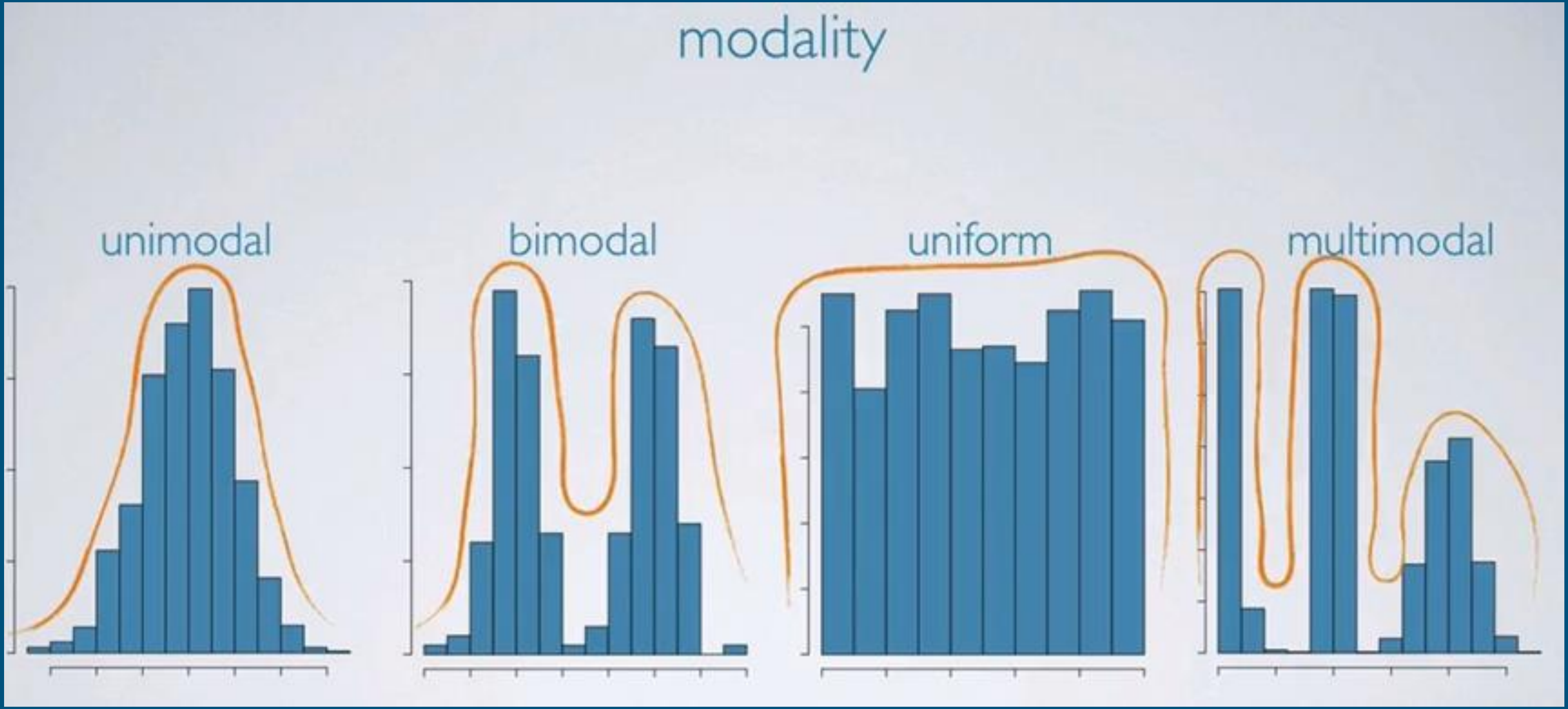
Histograms: Skew



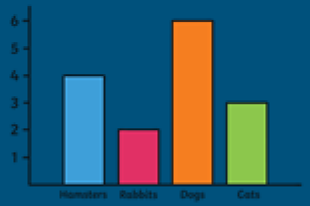
Histograms: Skew



Histograms: Modality



Visualization



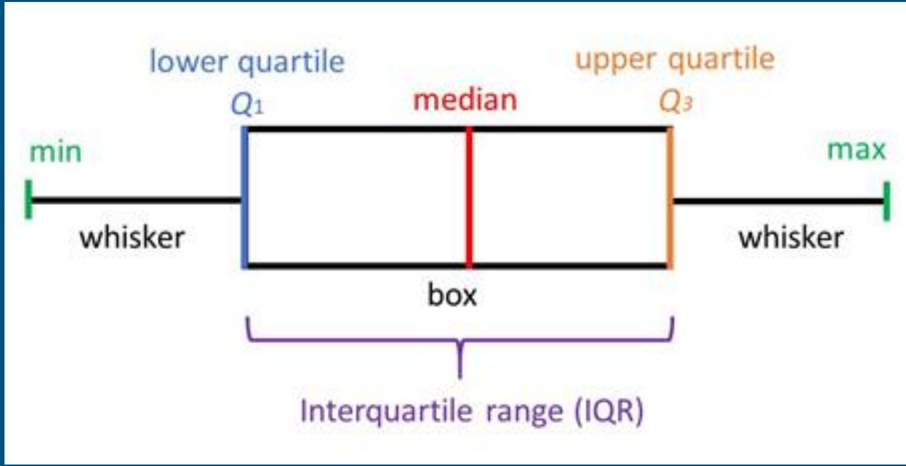
Dot plots

Frequency tables

Stem-and-leaf plots

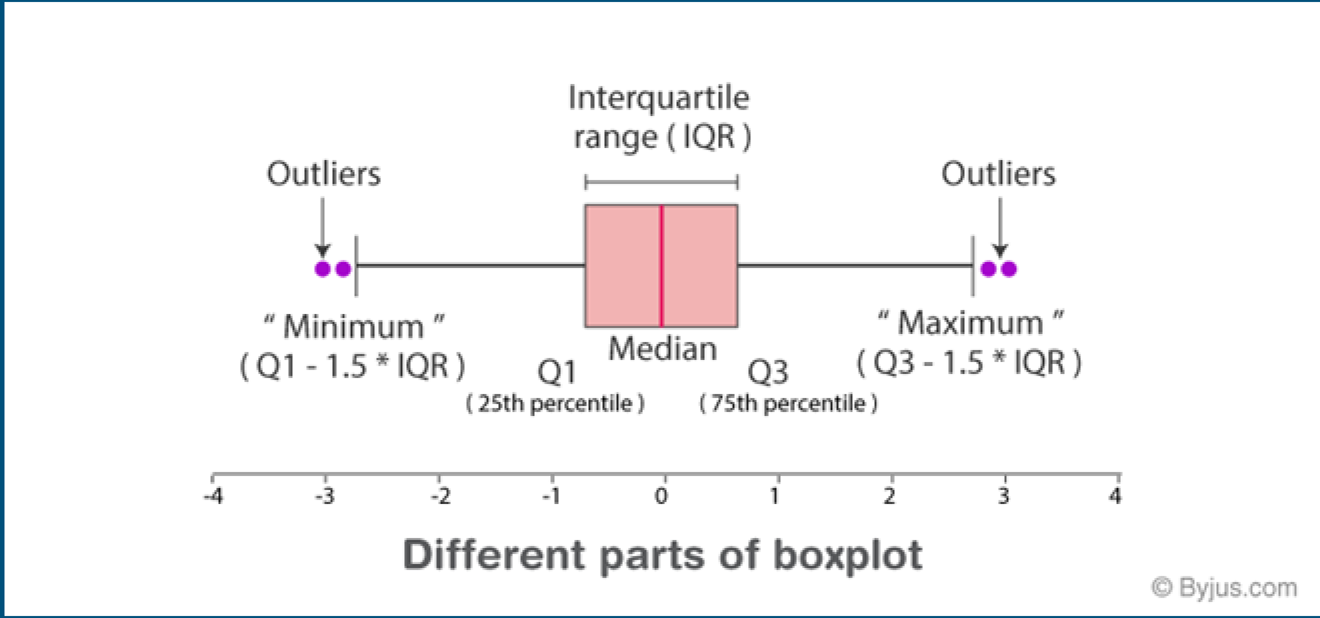
Histograms

Box plots



Box plots

Visualization of the 5-number summary, mostly



Outliers: Box plot

IQR x1.5 = length of whiskers

Anything outside of whiskers is an outlier

Outliers are considered statistically “extreme”

Outliers: Robust or no?

If a measure is *robust*, it means that it is not greatly affected by outliers

- Robust: median, IQR
- Not robust: mean, standard deviation

For symmetric data sets with no large outliers, better to use mean and SD

For skewed data sets or those with outliers, use median and IQR