Descriptive Statistics

ANA 500 – Foundations of Data Analytics

Module 1: Lecture 2

Descriptive statistics-- “Getting to know your data”

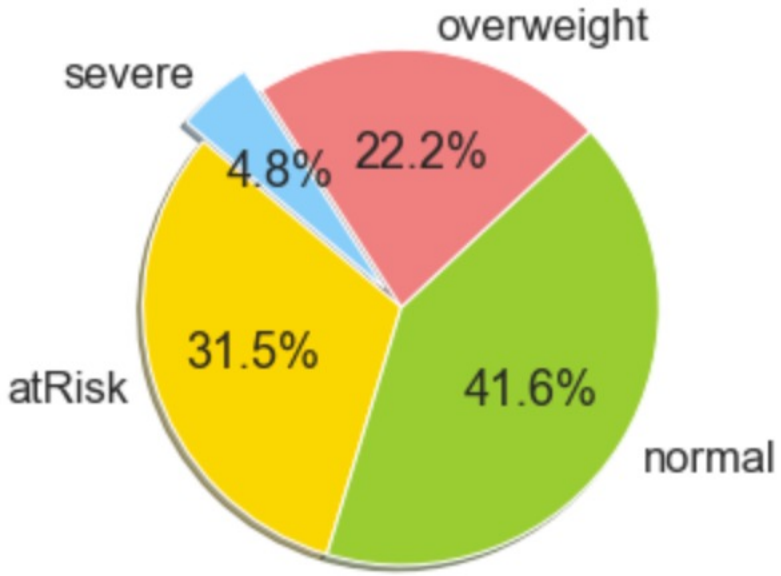
- Describe the characteristics of a dataset
- Can involve a single variable or multiple variables
- Numerical, graphical, tabular
- **Measure of Central Tendency**
  - Mean, Median, & Mode
- **Measures of Variability**
  - Range, Variance, & Standard deviation

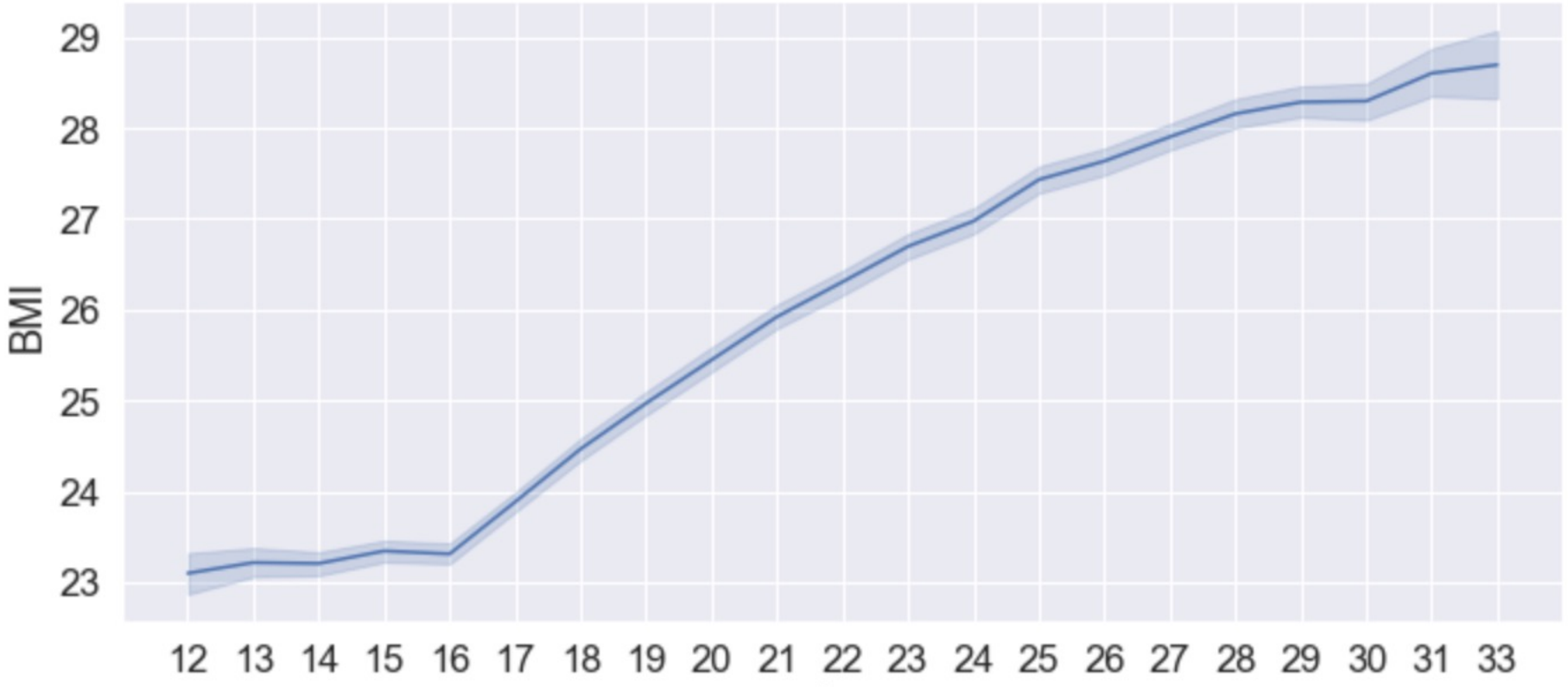
ID	Sex	Race	Age13BMI	Age14BMI	Age15BMI	Age16BMI	Age17BMI	Age18BMI	Age19BMI	Age20BMI	Age21BMI	Age22BMI	Age23BMI
1	2	4	22.71	23.49	23.49	24.33	25.53	24.28	23.96	25.06	25.06	24.59	24.43
2	1	2		21.14	22.15	21.93	24.21	24.75	25.53	26.63	28.19	29.12	
3	2	2				17.16	15.66	17.43	20.18		0.00		
4	2	2	35.14	34.75	33.84	33.84	36.58	43.16	35.67	45.73	42.07	51.03	51.21
5	1	2		22.46	23.17	23.40	22.60	27.44	26.63	28.25	29.86	28.73	27.44
6	2	2		25.96	23.52	25.61	25.75	25.75	24.89	27.10	29.18	31.93	27.46
7	1	2			28.58	28.74	27.02	30.27	31.01	33.96			39.13
8	2	4	20.53	20.53	21.29	22.81	20.53	21.93	23.18	23.49	22.71	24.12	21.29
9	1	4		23.63	22.24	24.41	25.10	26.50	24.41	24.41	25.10	25.80	26.50
10	1	4				15.55	15.94	17.43	16.50	17.94	17.43		18.99
11	2	2		20.60	21.63	21.11	21.26	21.61	23.17	26.78	29.23	30.65	30.11
12	1	2	34.61	35.51	34.70	35.51	31.32	37.59	35.24	36.02		40.35	38.74
13	1	2					18.84	23.41	24.09	25.42	20.80	21.63	20.80
14	1	2	27.44										
15	2	2			18.39	18.18	18.38	16.40	17.17	18.18	18.38	18.18	17.57
16	1	2		20.09	23.09	26.12	25.73	24.39	24.39	25.16	25.55	21.91	26.96
17	2	2	31.18	33.59	35.35	27.40				35.90			
18	1	1		19.73	20.83	22.71	20.36			21.41	21.14	21.29	21.14
19	1	1				16.72	25.57	20.25	20.80	21.80	20.99	20.80	23.40
20	1	1	21.41	20.92	18.83			23.67	23.09				
21	1	2		20.36	17.94	17.43	20.67	19.53	20.22	20.22	22.32	20.92	23.01
22	1	2		28.50	24.37	26.58	24.37	25.84	31.75	31.75	32.49		
23	2	2			18.81	17.76	18.99	16.97	19.53	20.31	21.61	22.22	23.23
24	1	2				15.94	18.29	20.54		20.34		23.40	22.05
25	2	2				30.18	33.66	39.32	34.33	34.33	37.76	38.96	34.33
26	1	1	25.02	24.13	23.40	26.63	24.21	25.82	24.37	31.01	34.70	31.15	37.12
27	1	1	24.28		23.63	24.33	22.81		26.63	22.81	23.49	24.21	22.71
28	2	1			20.78	19.77	20.98	21.93	25.06		24.33	25.84	26.61
29	2	4			19.49	20.37	21.26	-0.34					
30	1	4	20.80	22.31									
31	1	4		18.07	19.14	20.80	20.80	21.14	20.36	20.53	20.53	20.53	19.77

Example: descriptive statistics

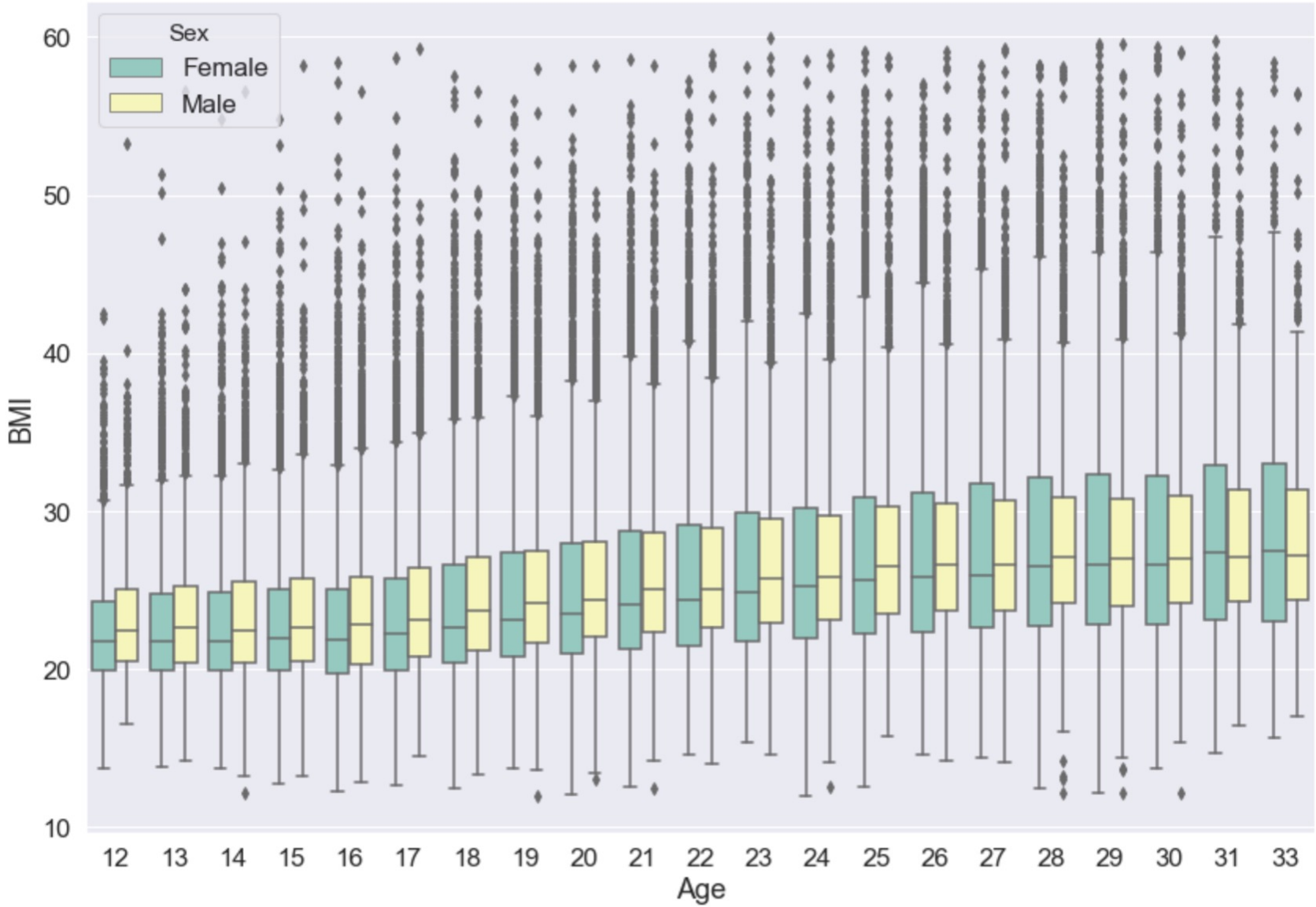
		BMI.13	BMI.14	BMI.15	BMI.16	BMI.17	BMI.18	BMI.19	BMI.20	BMI.21	BMI.22	BMI.23	BMI.24	BMI.25	BMI.26	BMI.27	BMI.28	BMI.29	BMI.30	BMI.31	BMI.33
N	Valid	3186	4760	6237	7701	7702	7541	7415	7284	7189	7168	7179	7214	7206	7190	7049	6897	5541	4080	2614	1308
	Missing	5798	4224	2747	1283	1282	1443	1569	1700	1795	1816	1805	1770	1778	1794	1935	2087	3443	4904	6370	7676
Mean		23.23	23.21	23.35	23.32	23.92	24.51	24.99	25.48	25.99	26.34	26.72	27.02	27.47	27.70	27.99	28.22	28.35	28.37	28.67	28.72
Median		22.16	22.15	22.31	22.31	22.71	23.12	23.67	24.13	24.51	25.00	25.11	25.75	25.99	26.45	26.57	26.63	26.78	26.87	27.26	27.30
Std. Deviation		4.57	4.52	4.76	4.99	5.20	5.54	5.63	5.82	6.16	6.16	6.23	6.31	6.52	6.63	6.78	6.83	6.97	6.96	7.07	7.04
Skewness		1.641	1.494	1.555	1.554	1.834	1.967	1.658	1.778	1.854	1.685	1.448	1.442	1.368	1.43	1.466	1.332	1.356	1.393	1.319	1.22
Kurtosis		4.91	3.683	4.59	4.382	8.118	8.497	4.762	6.835	7.602	6.152	3.322	3.685	2.852	3.488	3.913	2.8	2.947	2.968	2.578	1.857
Minimum		13.82	12.2	12.8	12.28	12.63	12.44	12.02	12.05	12.5	14.01	14.64	12.02	12.55	14.25	14.08	12.15	12.13	12.19	14.73	15.66
Maximum		61.24	56.58	67.81	62.65	92.87	87.83	78.34	92.22	97.65	92.45	71.55	81.92	72.8	79.71	79.6	75.22	83.01	73.51	68.14	62.65

Example: descriptive statistics



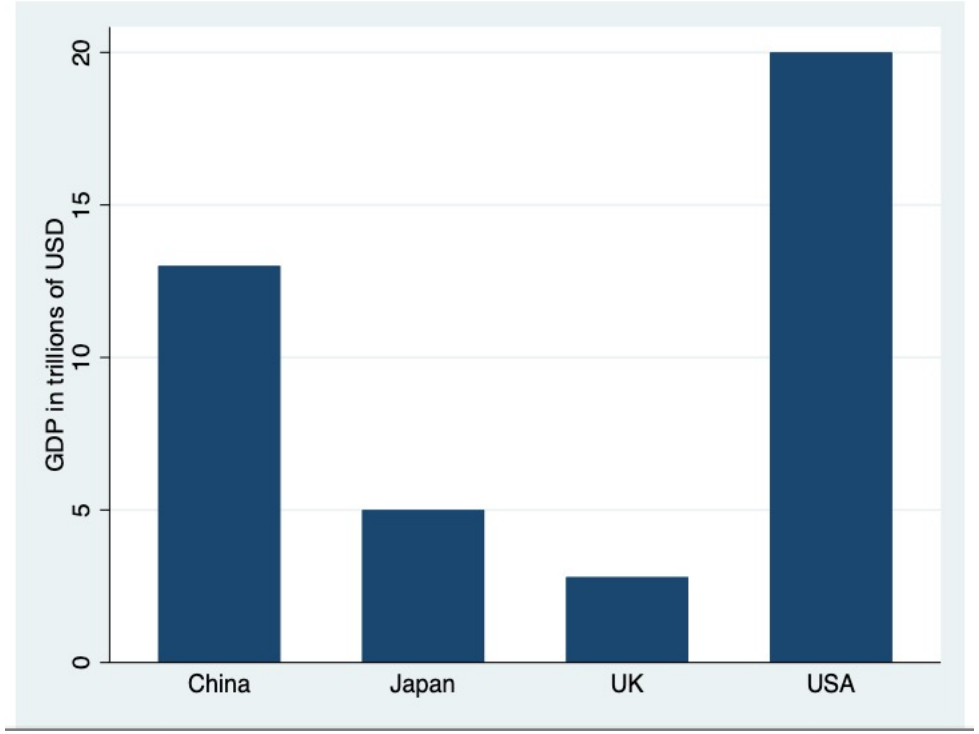


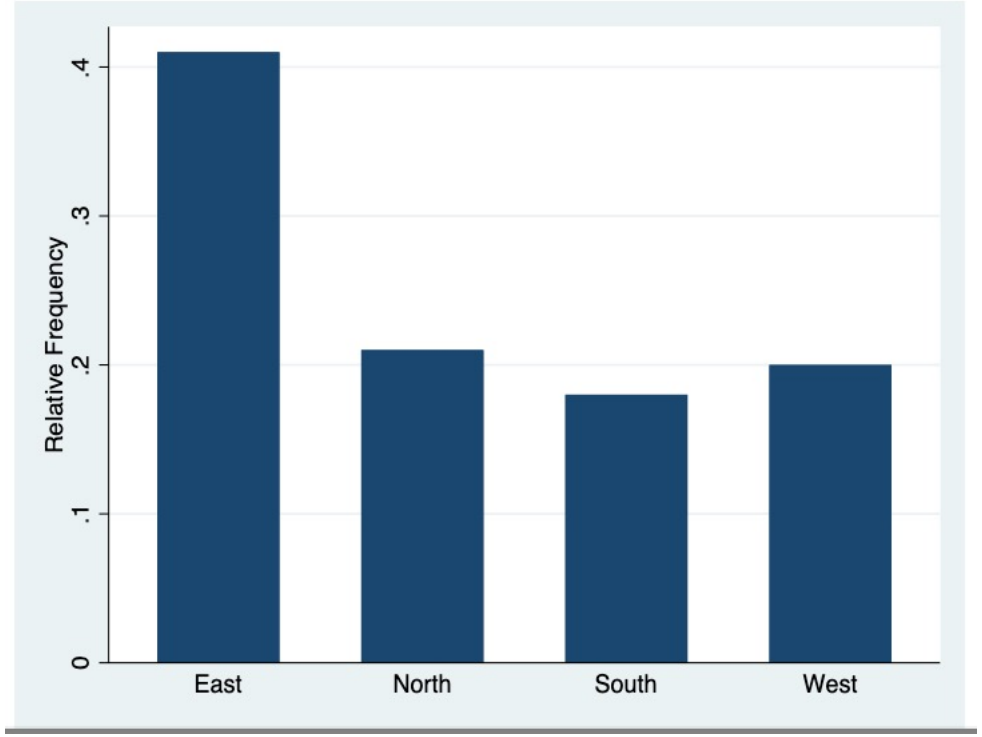
Example: descriptive statistics



Bar chart

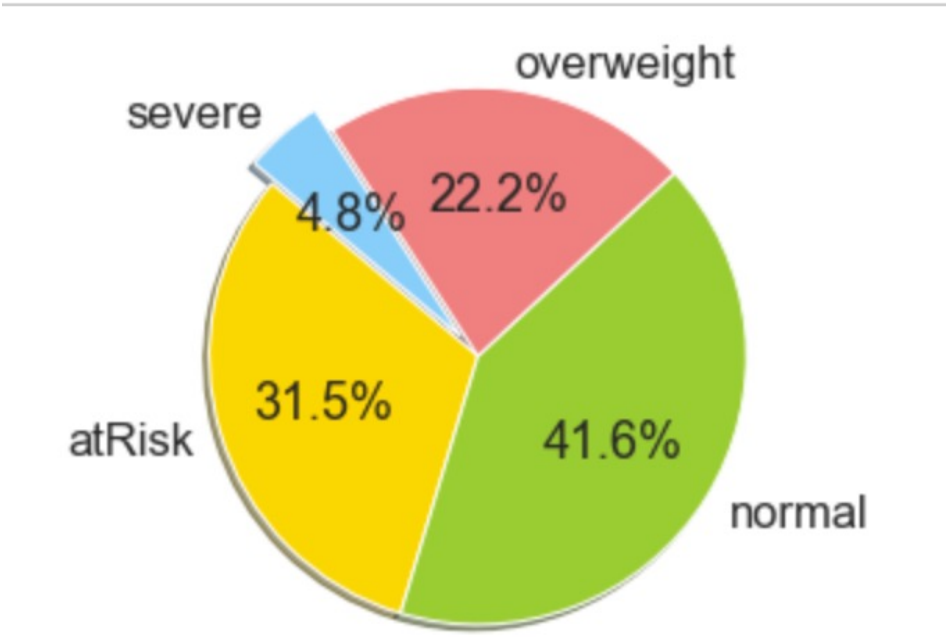
- Helpful for describing categorical data
- Categories on the x-axis. Outcome of interest measured on the y-axis.
- Sometimes the y-axis measures the relative frequency of the categories





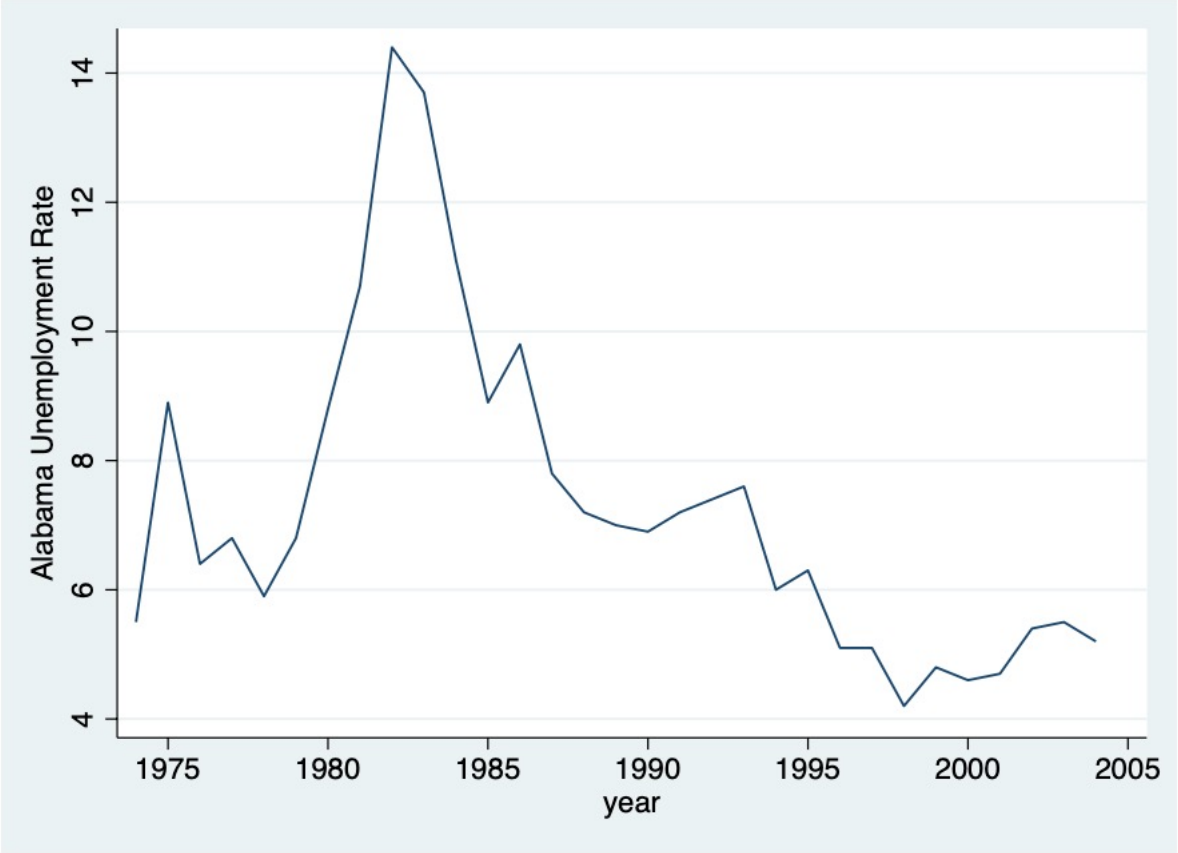
Pie chart

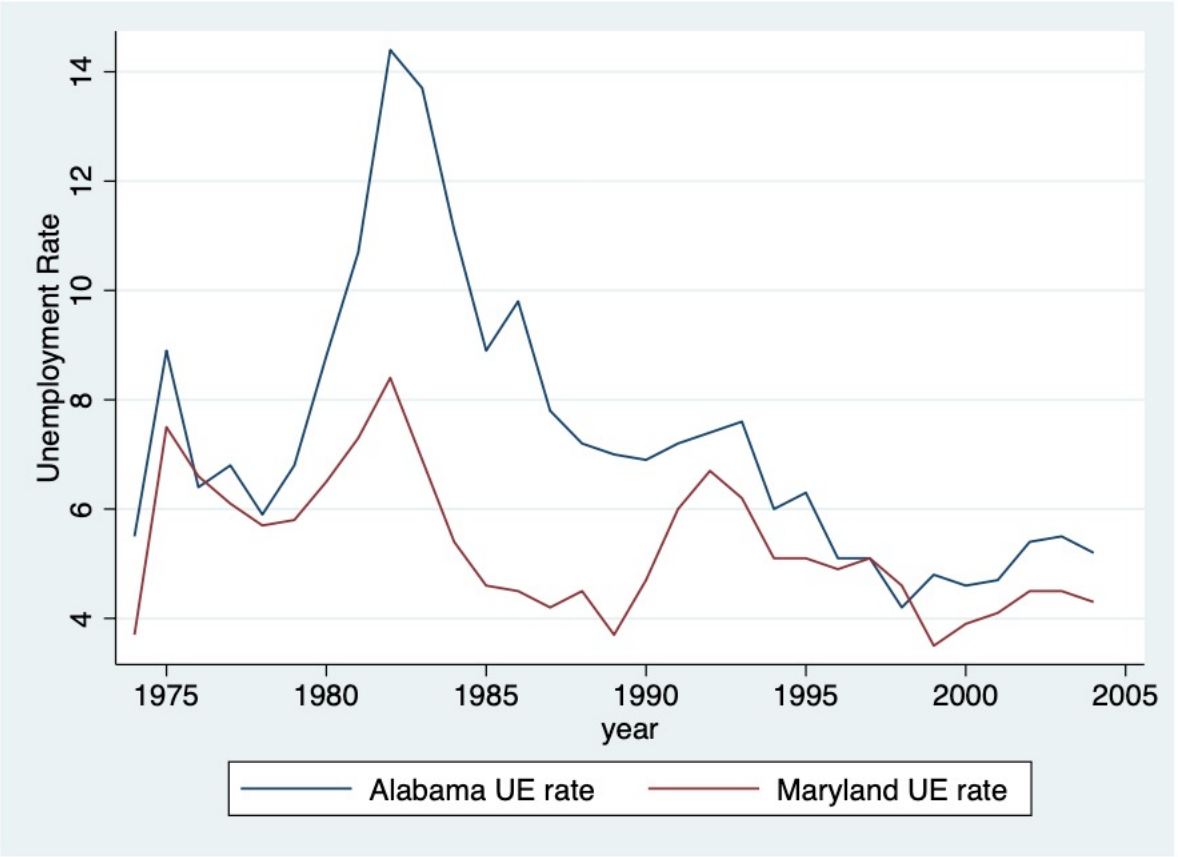
- Best to order the percentages
- A smaller number of categories works the best
- All observations must fit into a mutually exclusive group.



Time series graph (line graph)

- Helpful to see how a variable behaves over time

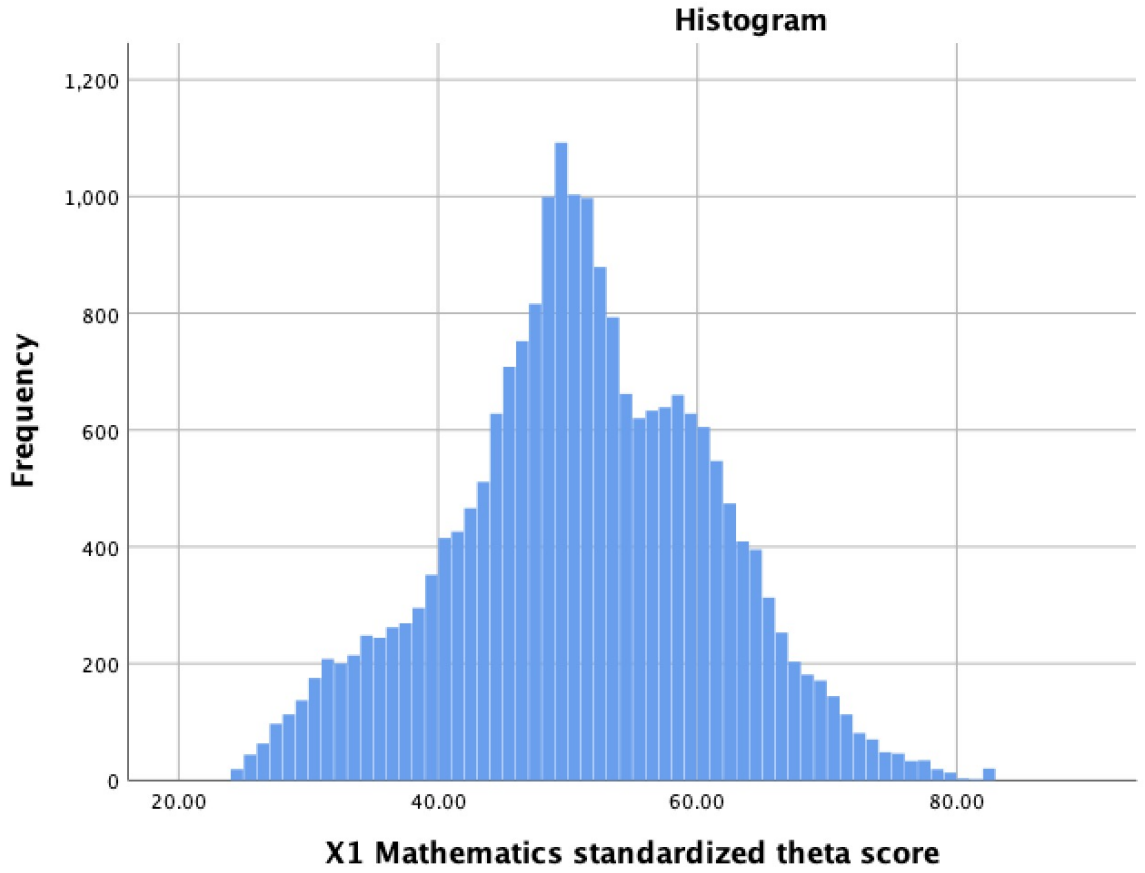




Histogram

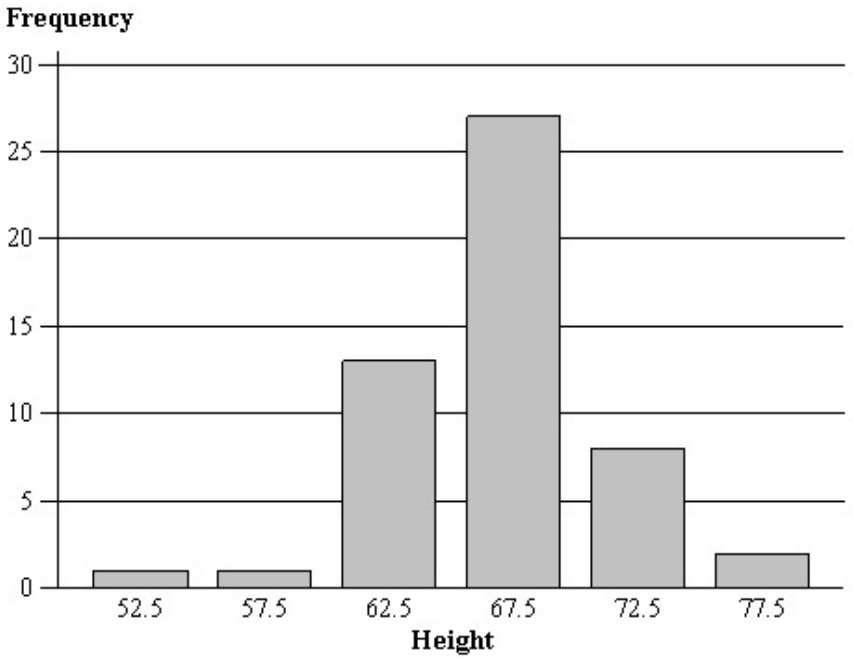
- Useful for describing the center and spread of a continuous variable.
- Adjoining intervals (boxes or bins) along the x-axis and the relative frequency of the values within each interval is measured on the y-axis.
- First interval contains smallest values
- Y axis can be “frequency” or “percentage”

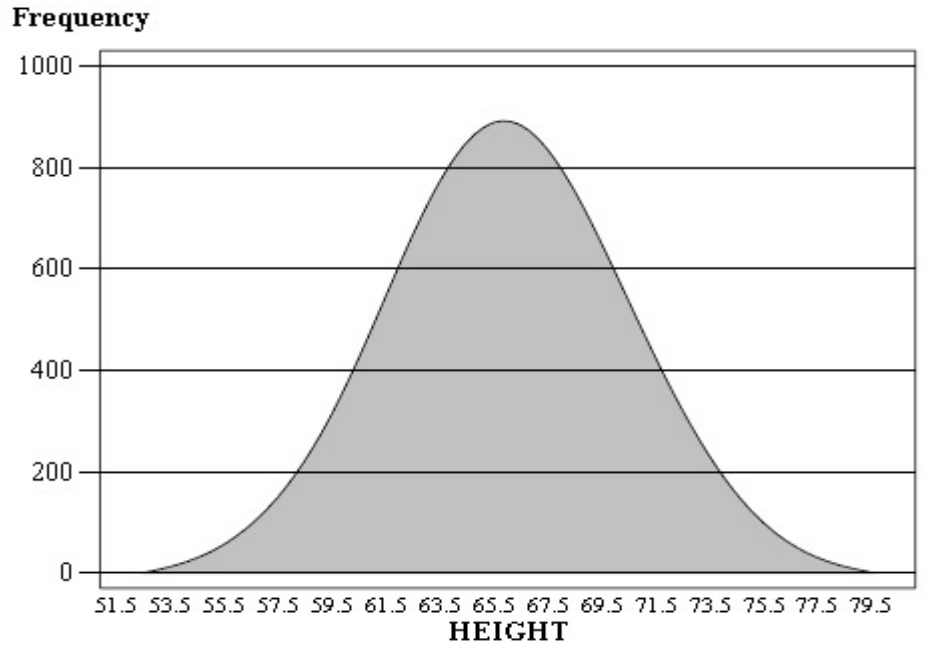
- Let *f* = frequency a values occurs
- Let *n* = sample size
- Relative frequency = *f* / *n*



Histogram

If the class intervals were made to be very small and we have a large data set, we would have a smooth line.
This smooth line can be thought of as a probability function.
Probability functions will be covered later in semester.





Percentiles - Measure of location

- Range from 1 to 99
- Indicate the percentage of scores that a given value is higher or greater than.
  - suppose that Lily scored at the 85th percentile on a standardized test such as the SAT. As she scored the 85th percentile this indicates that Lily scored better than 85% of the people on the exam.

Percentiles - Measure of location

Percentiles can also be thought of as dividing scores into separate groups

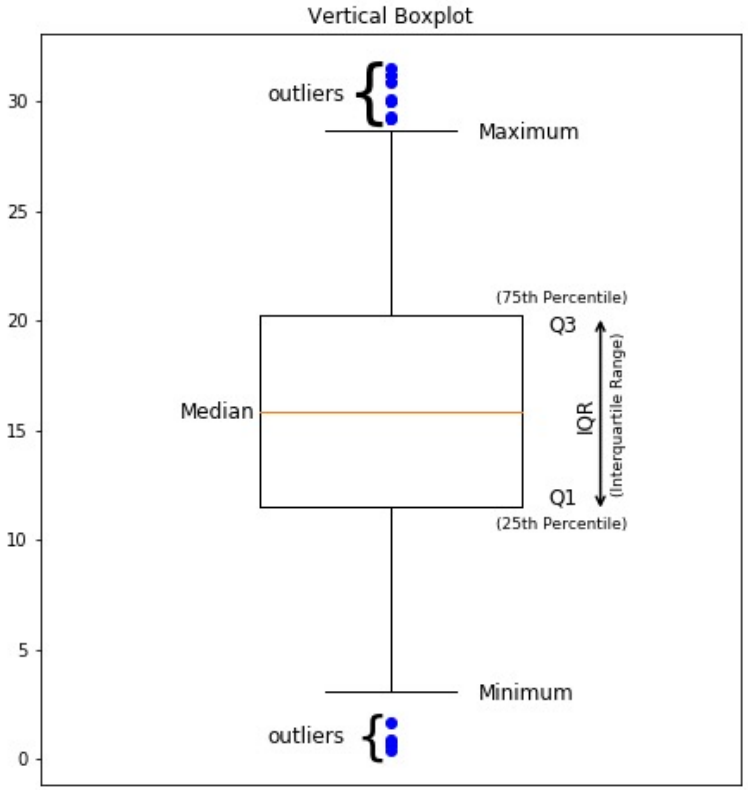
- the 25th percentile that indicates once again that the person who scored the 25th percentile they scored better than 25% of the examinees. Now that exact percentile is also known as Q1 or the first quartile.

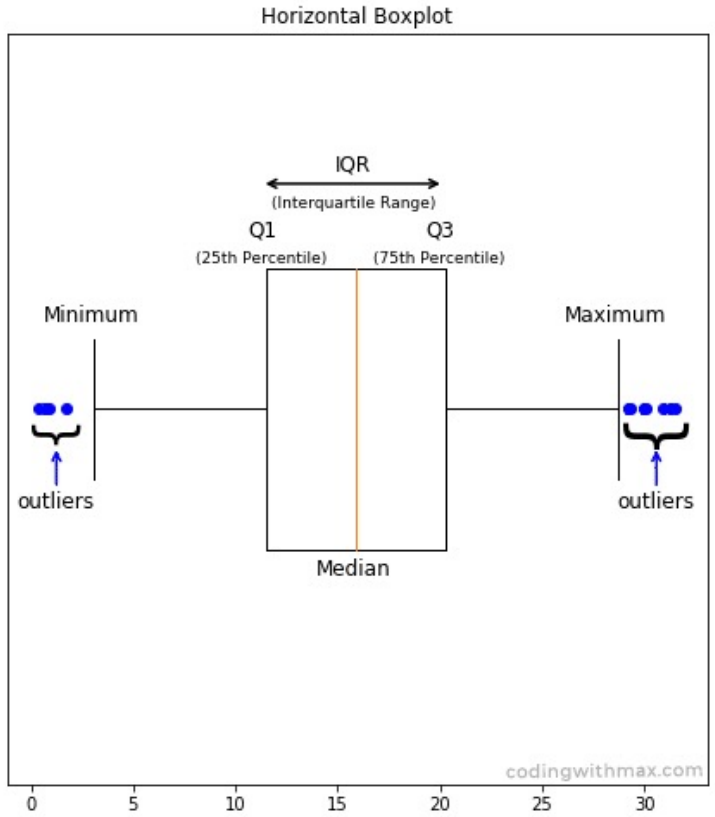
- The 50th percentile indicates that the person who scored there, scored better than 50% of the examinees. And the 50th percentile or Q2 is also equal to the median as it splits the distribution exactly in half.

- The 75th percentile indicates that the person who scored there did better than 75% of the examinees and this is known as Q3 or the third quartile.

Boxplots

- Boxplots are helpful graphical device for showing the location and spread of data.
- The “box” represents the IQR, the middle 50% of the data.
- The line in the middle of the box represents the median (50th percentile)
- help identify potential outliers.





Measures of central tendency

- Helpful to determine center of data
- The mean, median, and mode.

Mean

Most commonly called the “average.”

Add up the values for each case and divide by the total number of cases.

use the notation $\bar{x}$ to denote a sample mean

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency.

Measures of central tendency

Mean = (20+15+33+59+28)/5 = 31

- Median = 28
- The mean can be a misleading measure of central tendency when data are skewed.

Person	Income
A	20
B	15
C	33
D	59
E	28

Measures of central tendency

- Mean = (20+15+33+59+28+300)/6 = 75.83
- Median = (28+33)/2 = 30.5
- Recall the mean (31) and median (28) from the previous slide • Positive and negative skew

Person	Income
A	20
B	15
C	33
D	59
E	28
F	300

Median

The middle value when a variable's values are ranked in order; the point that divides a distribution into two equal halves.

When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it.

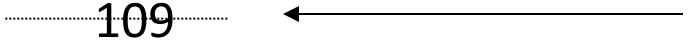
The 50^{th} percentile.

The median is unaffected by outliers, making it a better measure of central tendency, better describing the "typical person" than the mean when data are skewed.

Median

Class A--IQs of 13 Students

89
93
97
98
102
106
109
110
115
119
128
131
140



Median = 109

(six cases above, six below)

Median

If the first student were to drop out of Class A, there would be a new median:

89
93
97
98
102
106
109
110
115
119
128
131
140

Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

Mode

The most common data point is called the mode.

The combined IQ scores for Classes A & B:

80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115 119 120

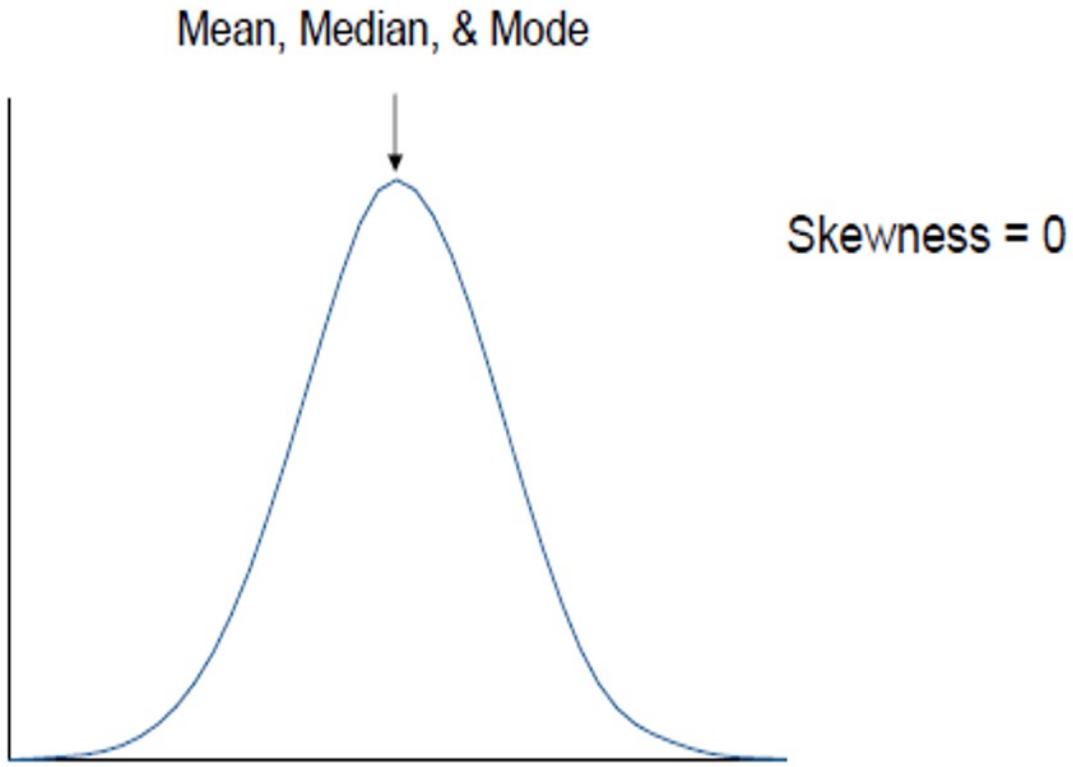
127 128 131 131 140 162

BTW, It is possible to have more than one mode!

Skewness: the degree of asymmetry

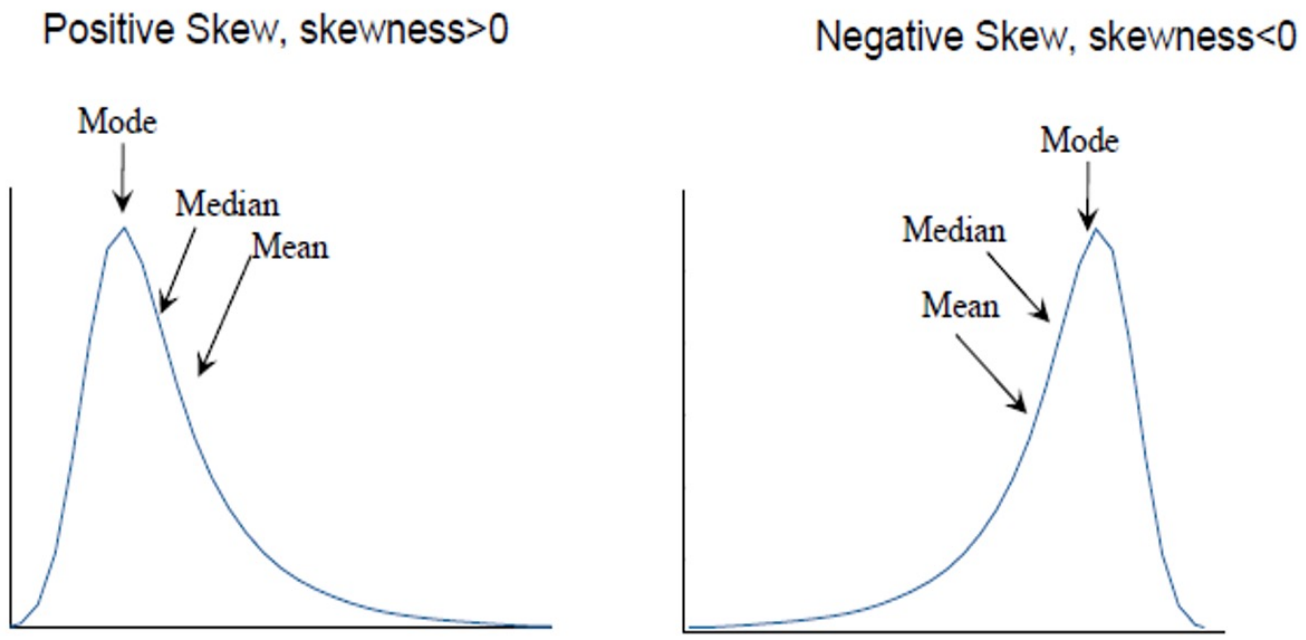
Symmetric distribution: has the same shape on both sides of the center.

The mean, Median, and mode are equal only when the distribution is symmetrical and unimodal.



Skewness: the degree of asymmetry

A measure of “skewness” will indicate the extent and the degree of asymmetry about the mean.



General rule of thumb: you can assume normality if -1 < skewness < 1

From: Huck, S. (2004) Reading Statistics & Research (4th ed.). Allyn & Bacon.

Measures of Variability (Dispersion)

Need additional measure(s) to indicate the degree to which individual observations are clustered around or, equivalently, deviate from that mean value.

- **Range**
- **Variance**
- **Standard Deviation**

Range

- The difference between the two most extreme data points (maximum–minimum).
- Sensitive to outliers

Class A--IQs of 13 Students	
102	115
128	109
131	89
98	106
140	119
93	97
110	

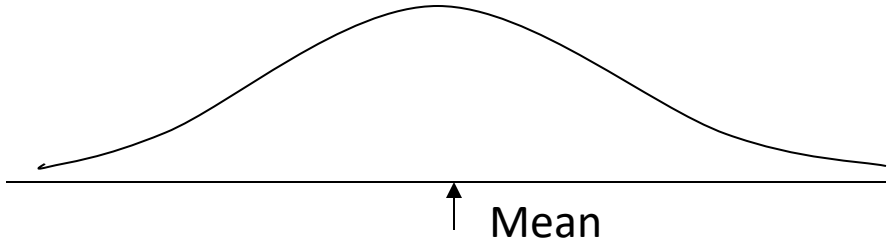
Class A Range = 140 - 89 = 51

Variance

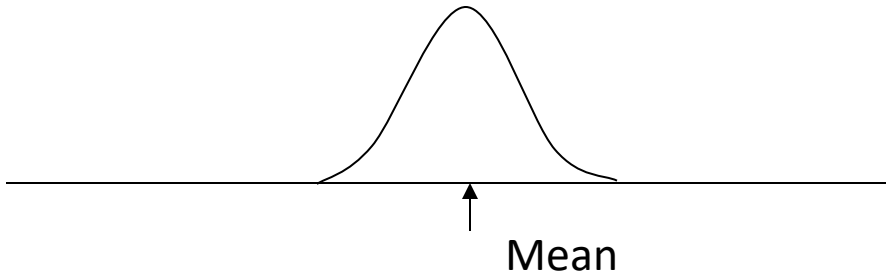
A measure of the spread of the recorded values on a variable. A measure of dispersion. (Sensitive to outliers)

- How far observations typically fall from the mean

The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



Variance (Dispersion)

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

s^2 = sample variance

x_i = value of i^{th} element

$\bar{x}$ = sample mean

n = sample size

Example:

$x_1 = 15$

$x_2 = 13$

$x_3 = 17$

$x_4 = 7$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

$$= \frac{15 + 13 + 17 + 7}{4}$$

$$= \frac{52}{4}$$

$$= 13$$

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

$$= \frac{(15-13)^2 + (13-13)^2 + (17-13)^2 + (7-13)^2}{4-1}$$

$$= \frac{2^2 + 0^2 + 4^2 + -6^2}{3}$$

$$= \frac{4 + 0 + 16 + 36}{3}$$

$$= \frac{56}{3}$$

$$= 18.67$$

Standard deviation

- Standard deviation: $S = \sqrt{S^2}$
- Standard deviation gives a measure of the spread of the data and can also be used to determine how far an observation is from the mean.
- For most variables, most values are within 2 standard deviations of the mean.
- Chebyshev's inequality: $(1 - 1/k^2)$ of a variable's values must be within k standard deviations from the mean, for $k>1$.

Standardization

- It is difficult to compare variables measured in different units.
- A variable can be **standardized** to convert its units into standard deviations from the mean.
- A variable is standardized by subtracting the mean from each value and dividing by the standard deviation.
- $z = \frac{x - \bar{x}}{s}$
- A standardized variable has a mean of 0 and a standard deviation of 1.
- Is used to improve interpretability of variables

Standardization

how do we interpret this 1.71 right here? What that means is the value of 65 is 1.71 standard deviations above the mean

x	z
21	(21 – 29.33) / 20.83 = -0.40
43	(43 – 29.33) / 20.83 = 0.66
12	(12 – 29.33) / 20.83 = -0.83
65	(65 – 29.33) / 20.83 = 1.71
23	(23 – 29.33) / 20.83 = -0.30
12	(12 – 29.33) / 20.83 = -0.83
$\bar{x}$ =29.33 $s_x = 20.83$	$\bar{z}$ =0 $s_z = 1$

example of -0.83 means that 12 is 0.83 standard deviations below the mean.

Descriptive statistics

- Remember, random sampling makes our statistics (e.g. $\bar{x}$) random variables.
- A statistic's variation can be calculated.
- A statistic's standard deviation is called its standard error.