Econ 1620: Introduction to Econometrics

Week 1, Lecture 2 & Week 2, Lecture 1

Descriptive statistics for numerical data (ASW ch. 3)

Dimitra Politi -- Brown University -- Spring 2015

Getting a "feel" for your data

- Suppose you interview a large number of Brown alumni, and you record various characteristics: age, gender, income, GPA and courses they took at Brown…

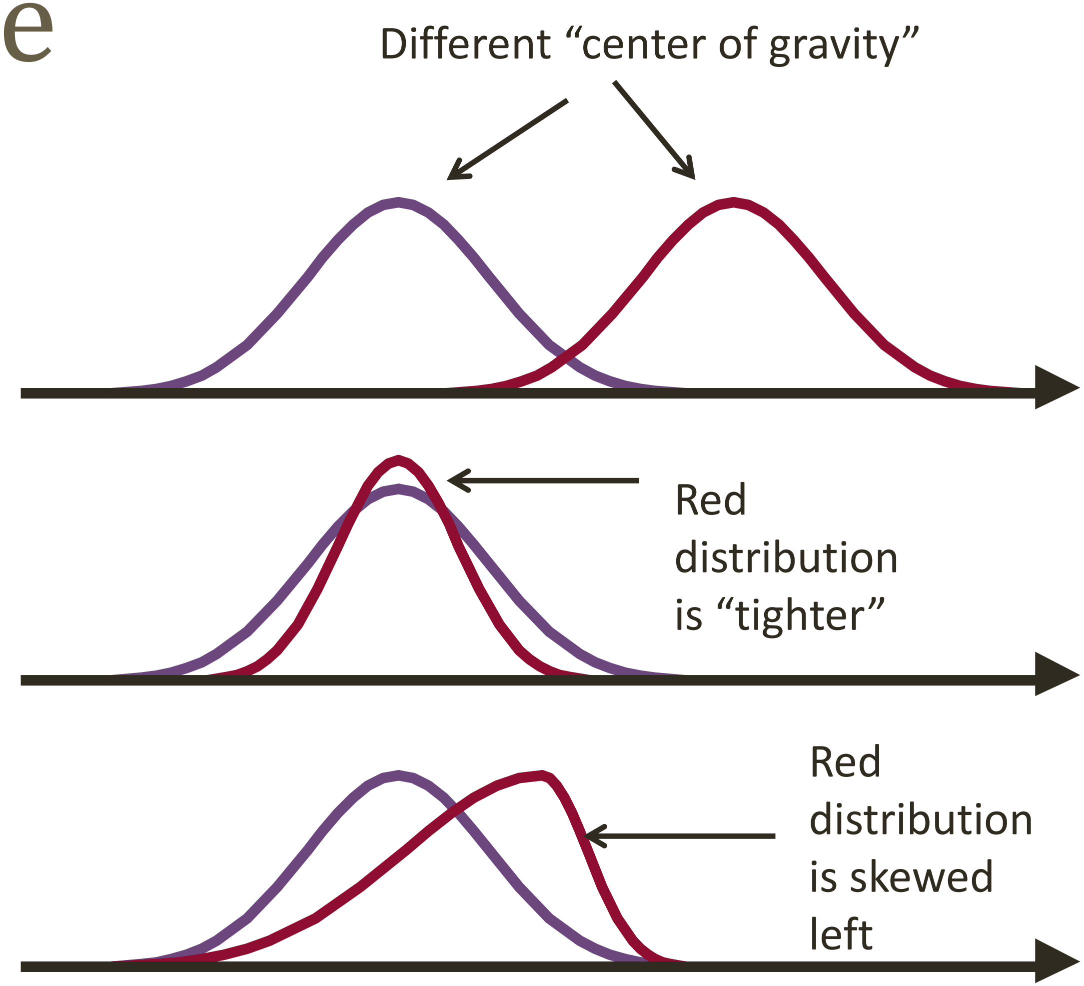
- Let's focus on income. How would you summarize its distribution?
  - What's the center of gravity? Do numbers tend to lump around one value? → measures of **central tendency**: mean, median, mode…
  - How "tight" is the distribution around its center? Are numbers tightly packed around the average, or are they more spread out? → measures of **dispersion**: range, interquartile range, variance and standard deviation, coefficient of variation…
  - What is the **shape** of the distribution? Is it symmetric or skewed? Are the tails thinner or fatter? → quantiles, skewness, kurtosis
  - Is income **related** to other characteristics, such as age or GPA at Brown? → covariance, correlation …

Central tendency, dispersion, and shape

Central Tendency (Location)

Variation (Dispersion)

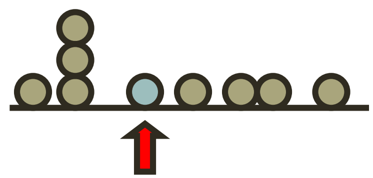
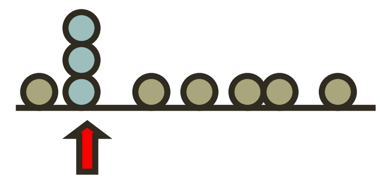
Shape



Central Tendency

Overview

Central Tendency

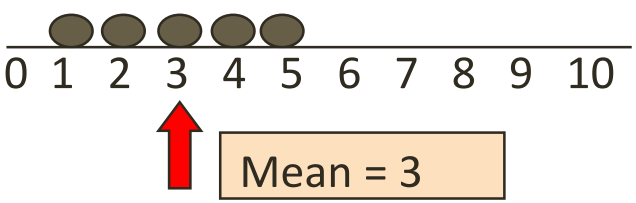
Mean	Median	Mode
$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$		
Arithmetic average	Midpoint of ranked values	Most frequently observed value

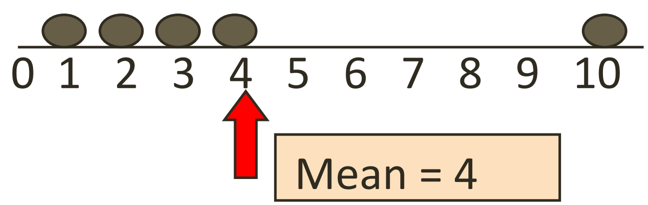
The arithmetic mean

- The (arithmetic) mean is the most common measure of central tendency.

Population values

- For a population of size N: $\mu = \frac{\sum_{i=1}^{n} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$ population size
- For a sample of size n: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ sample values, sample size
- When we only have access to a sample, the sample mean (or average) is a good estimator (proxy) for the population mean.
- The mean is affected by extreme values, or outliers, which is why sometimes the **median** might be a better measure.



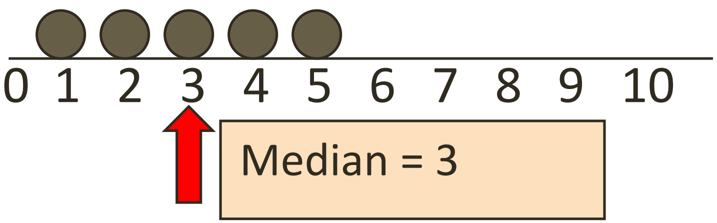


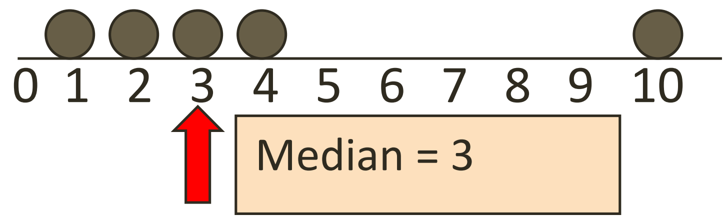
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

The median

- The median is the "middle" number, the number that splits the sample (or population) in half. 50% of observations are below the median.
- The median is not affected by extreme values.





- How do we locate the median? First, you have to order the data.
  $Median\ position = \frac{n+1}{2}\ position\ in\ the\ ordered\ data$
  - Note that $\frac{n+1}{2}$ is the <u>position</u>, not the <u>value</u> of the median.
  - If the # of observations is odd, the median is the middle number
  - If the # of observations is even, the median is the average of the two middle numbers.

Mean or median?

- In general, they measure different things!

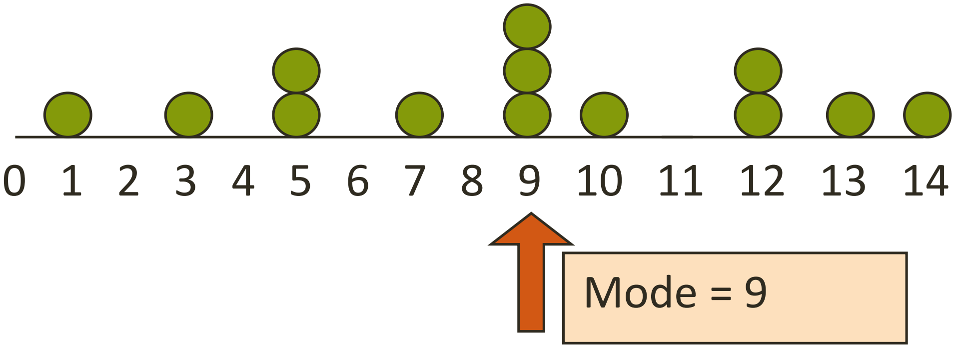
- If the distribution is symmetric, the mean is equal to the median (in the population. In our sample this might not be true, but both the sample mean and sample median will be proxies for the same population value.)

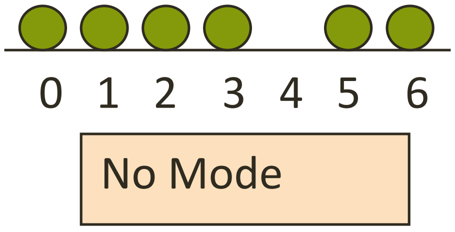
- When we conduct statistical inference:
  - The sample mean is a more precise estimator (proxy) than the median if the variable is normally distributed (more on the normal distribution in later lectures…).
  - The median is a more robust measure: if there is error in our data, the median will be less affected than the mean.

- The median is the measure of location most often reported for annual income and property value data. A few extremely large incomes or property values can inflate the mean.

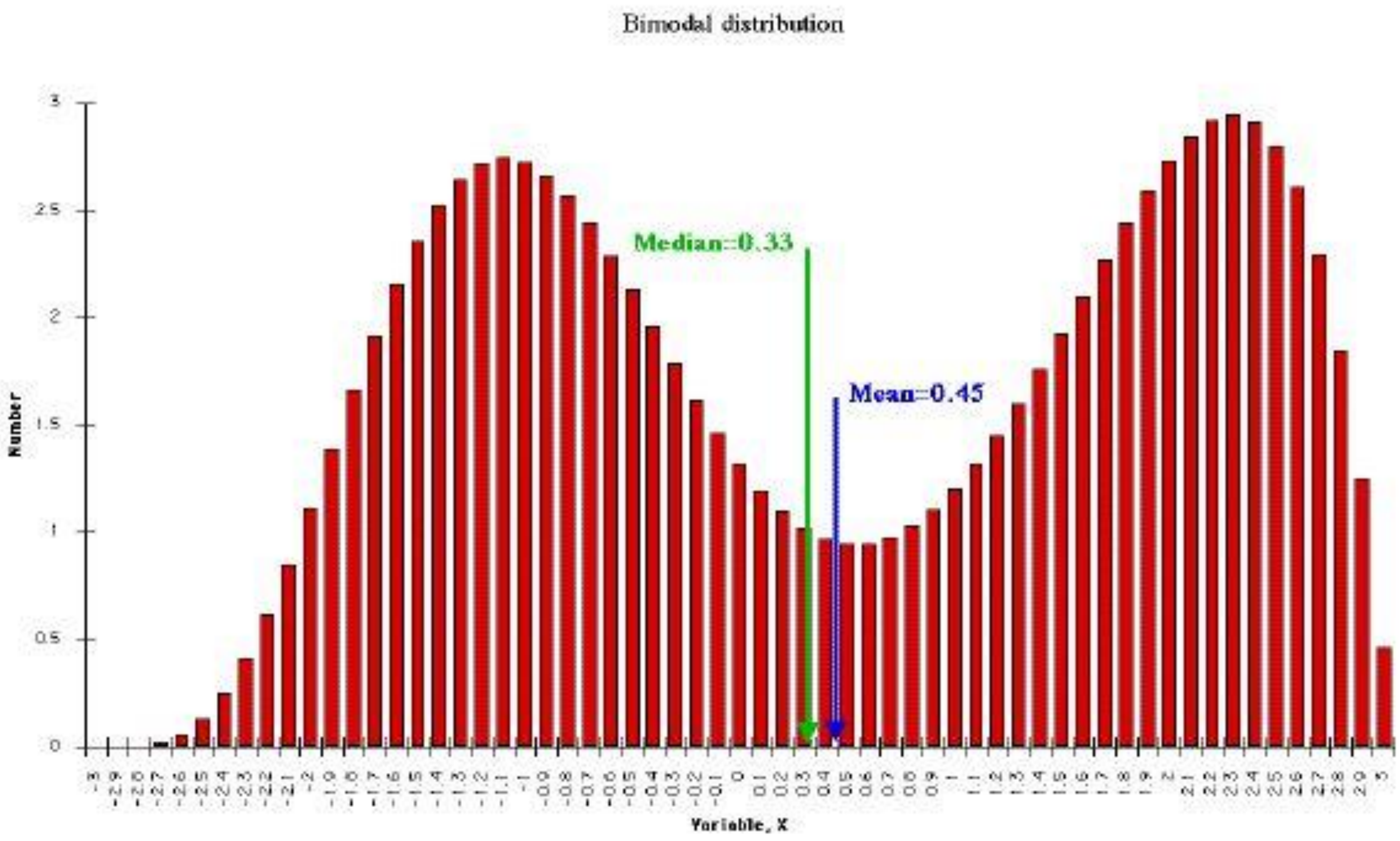
The mode

- The mode is another measure of central tendency of a distribution. It is the value that occurs most often.
- It is not affected by extreme values.
- There may be no mode, two modes (bimodal distribution), or several modes (multimodal distribution).
- If the distribution is symmetric (e.g. normal distribution), mean = median = mode.





A bimodal distribution



Central tendency: example

- You have data on property prices for 5 houses on a hill by the beach.

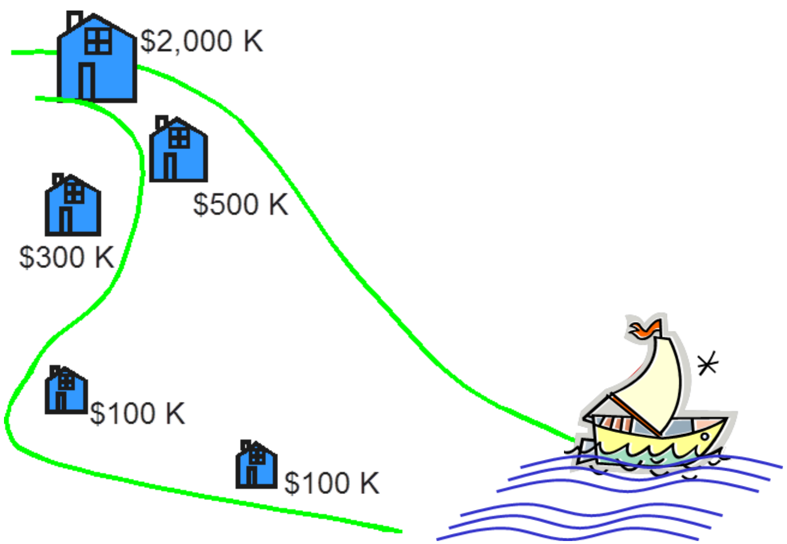
House 1:	$2,000,000
House 2:	$500,000
House 3:	$300,000
House 4:	$100,000
House 5:	$100,000
Sum of values:	$3,000,000

- Summary statistics:

$$\boldsymbol{Mean} = \frac{\$3{,}000{,}000}{5} = \$600{,}000$$

$\boldsymbol{Median} = \$300{,}000$: middle value of ranked data

$\boldsymbol{Mode} = \$100{,}000$: most frequent value



Question (food for thought)

You have a sample of 5 workers. Their hourly wages are the following: 10$, 15$, 20$, 25$, 30$. When inputing the data into Stata, your secretary inadvertently inputs randomly one of the wage in cents instead of in dollars. You use these data to calculate the mean and the median. Which of the following statements is incorrect?

A) The calculated mean is always different from the true mean.

B) The calculated median is always different from the true median.

C) When the mean and the median are incorrect, the error of the mean is larger than the error of the median.

D) The true median and the true mean are equal.

Weighted average (or mean)

- Sometimes your data comes grouped into classes, with w_i values in the i^{th} class.
- For example, you might be interested in the age distribution in a group of 50 students, and instead of individual data on age, you have the following information:

Age	Number of students
18	20
19	22
20	4
21	3
25	1

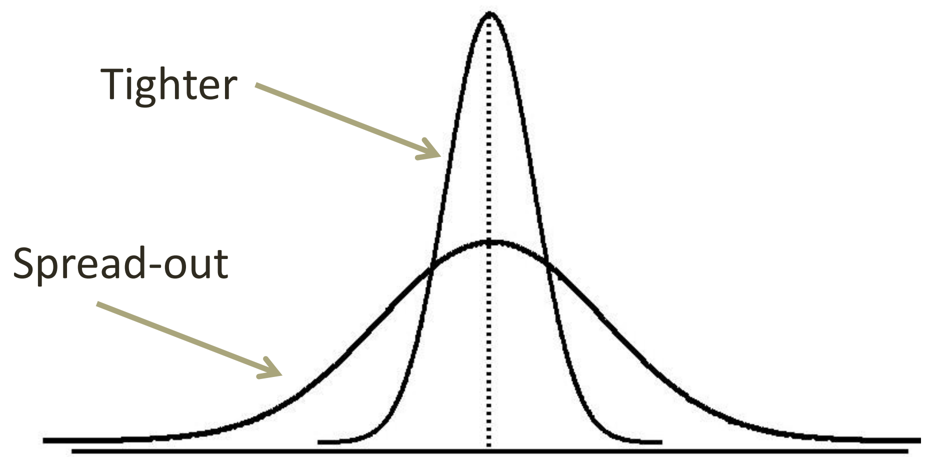
- To calculate the average age in the group, you need to weight each age i by the number of students if that age group (w_i), and divide by your sample size ($n = \sum w_i = 50$). The resulting average is a weighted mean.

- $\bar{X} = \frac{\sum_{i=1}^{n} w_i \cdot X_i}{n} = \frac{w_1 \cdot X_1 + w_2 \cdot X_2 + \cdots + w_n \cdot X_n}{n}$

Dispersion (or variability)

- Dispersion
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

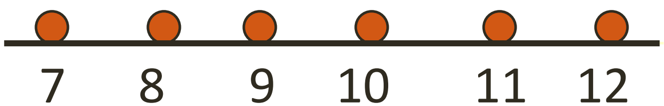
- Measures of variation give information on the spread or variability of the data values.



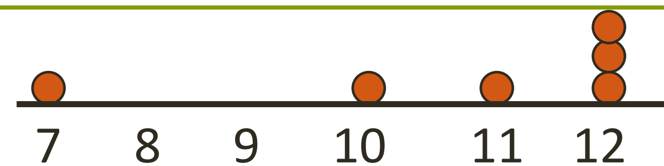
Same center,
different variation

Range: as simple as it gets

- Range is simply the difference between the largest and smallest observation in your sample: $Range = X_{max} - X_{min}$
- The range is a very crude measure, though: it ignores the way in which the data are distributed, and it is sensitive to outliers.
- That’s why sometimes we use the interquartile range instead.



Range = 12 - 7 = 5



Range = 12 - 7 = 5

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

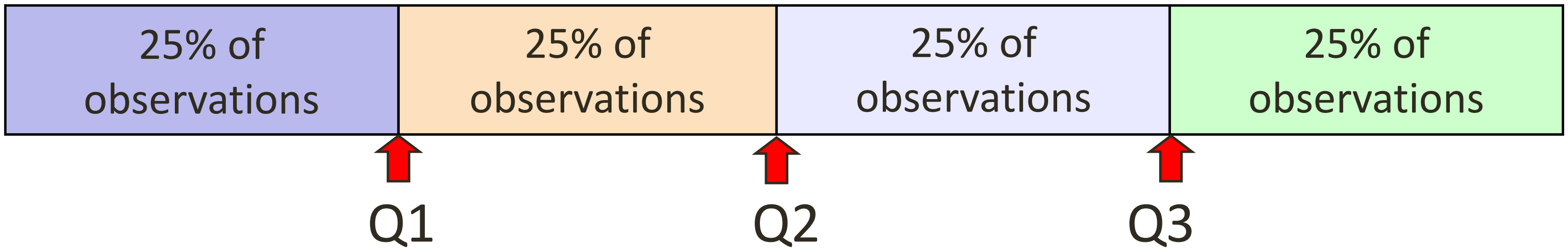
Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

Quartiles

- Quartiles split the ordered/ranked sample into four equal parts.

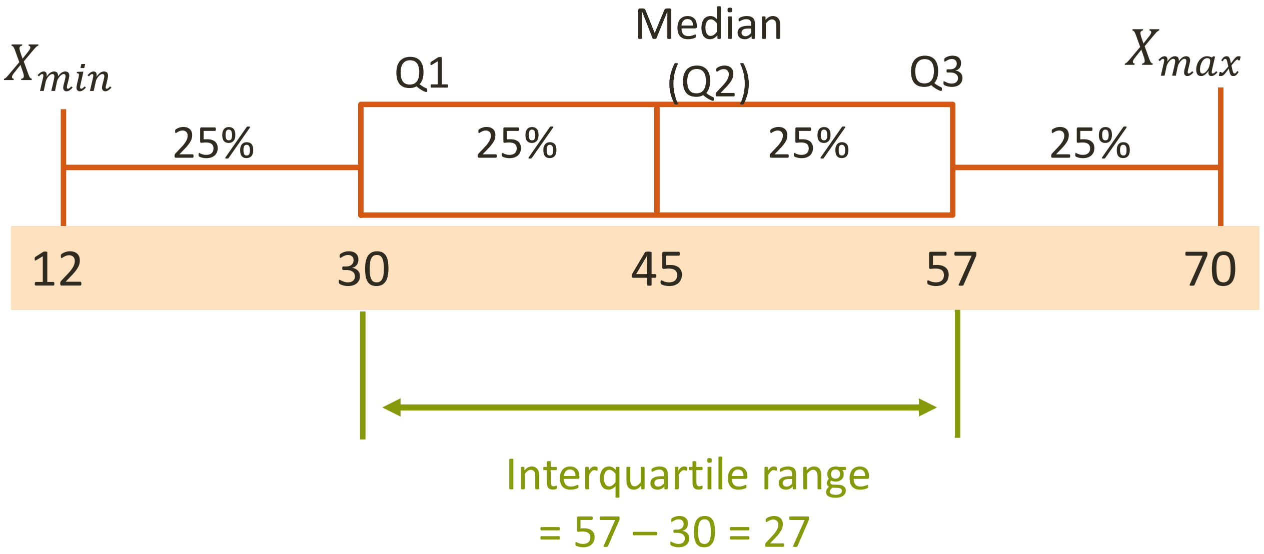


- Q_1: value for which 25% of observations are smaller.
- Q_2: value for which 50% of observations are smaller, aka median!
- Q_3: value for which 75% of observations are smaller.

- First quartile position: $Q_1 = 0.25 \cdot (n+1)$ [remember to rank the data first!]
- Second quartile position: $Q_2 = 0.50 \cdot (n+1)$ [same as median]
- Third quartile position: $Q_1 = 0.75 \cdot (n+1)$
  - If the position is in-between two values, take the average.

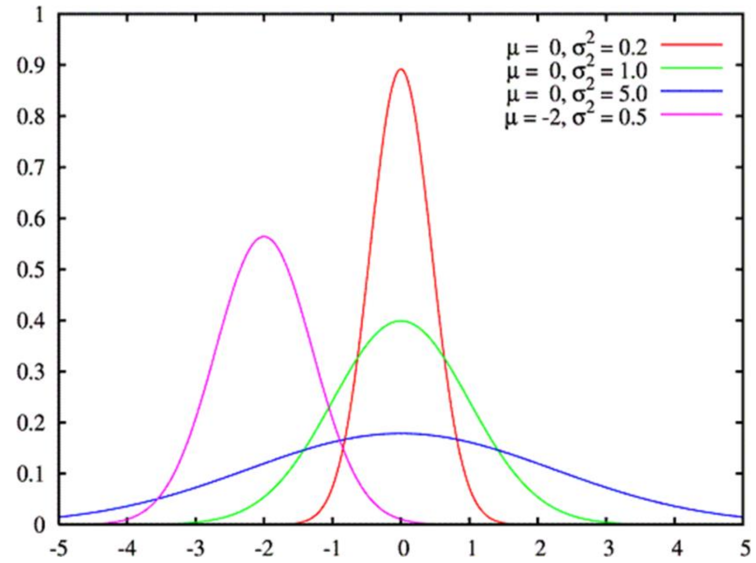
Interquartile range

- Interquartile range is the range of the middle 50% of our data.
- To find it, we eliminate values lying below the first quartile and over the third quartile, and calculate the range of what is left: IQR = 3rd quartile – 1st quartile = $Q_3 - Q_1$
- Example:



Variance

The variance and its close relative, the standard deviation, are two very common measures to describe the “spread” of “dispersion” of values in a distribution: they measure how far, on average, the values are from the mean (or center) of a distribution.



- If you have access to data from the whole **population**:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

- If you only have a **sample**, then:

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

μ: Population mean
$\bar{X}$: sample mean
N: population size
n: sample size
X_i: i^{th} value of variable X

Notes on variance

- Look at the formula for the variance: it is the average **square deviation** of data values from the mean, so the variance is measured in squared units of X.

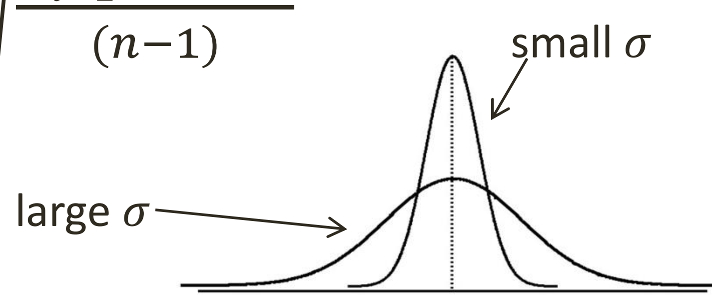
- By squaring the distance between a value and the mean we give extra weight to values far from the mean.→ The variance is very sensitive to outliers (extreme values).

- Note that to get the sample variance, we divide by $(n-1)$, rather than n. Why do we do that?
  - It turns out that by dividing by $(n-1)$ we get a better estimator (proxy) for the population variance (more on that later…).
  - The intuition is that we "use up" some of the information in our sample to calculate the sample mean, which we need to calculate the sample variance.
  - To reflect that, we divide by a smaller number, otherwise we would be underestimating the variance.
  - Note that as our sample size gets large, this becomes insignificant, because $(n-1) \approx n$ when $n \to \infty$.

Standard deviation

- Because the variance is measured in units of the square of X, we often measure the spread of a distribution by its square root, the standard deviation.
- The standard deviation is the most common measure of dispersion of a distribution, and has the same units as our original data.

- **Population** standard deviation: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$
- **Sample** standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{(n-1)}}$



Standard deviation example

Sample data (X_i): 10 12 14 15 17 18 18 24

$n = 8 \qquad Mean = \bar{X} = 16$

Sample standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{(n-1)}}$

$$s = \sqrt{\frac{(10-16)^2 + (12-16)^2 + \cdots + (24-16)^2}{n-1}} =$$

$$= \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{X})^2 + \cdots + (24-\bar{X})^2}{8-1}} =$$

$$= \sqrt{\frac{126}{7}} = \boxed{4.2426}$$

→ A measure of the "average" scatter around the mean

Questions (food for thought)

- The following data represent scores on a 15 point aptitude test: 8, 10, 15, 12, 14, and 13.

1. Subtract 5 from every observation and compute the sample mean for the original data and the new data.
2. Subtract 5 from every observation and compute the sample variance for the original data and the new data.

- The following ten scores were obtained on a 20-point quiz: 4, 5, 8, 9, 11, 13, 15, 18, 18, and 20. The teacher computed the usual descriptive measures of center (central tendency) and variability (dispersion) for these data, and then discovered an error was made. One of the 18s should have been a 16. Which of the following measures, calculated on the corrected data, would change from the original computation?

  a. the median b. the mean

  c. the range d. the interquartile range

Coefficient of variation

- The standard deviation is measured in units of X, so it is more intuitive than the variance.
- However, when we want to compare two variables measured in different units, we're stuck! We need a measure of **relative variation**. → coefficient of variation
- The coefficient of variation measures dispersion relative to the mean. It is a percentage, so it makes comparisons easy!

$$CV = \left(\frac{s}{\bar{\bar{X}}}\right) \cdot 100\%$$

Example:

Stock A	Stock B
Average price last year = $50	Average price last year = $100
Standard deviation = $5	Standard deviation = $5
$CV_A = \left(\frac{s}{\bar{\bar{X}}}\right) \cdot 100 = \frac{5}{50} \cdot 100 = 10\%$	$CV_B = \left(\frac{s}{\bar{\bar{X}}}\right) \cdot 100 = \frac{5}{100} \cdot 100 = 5\%$

Note that $s_A = s_B$, but stock B is less variable relative to its price.

Shape of a distribution: quantiles

- Sometimes we want to know more about a distribution than its center of gravity and its degree of dispersion around the mean.
- We defined the median as the value that splits the sample in half, and quartiles as values that split the sample in quarters.

- **Quantiles** generalize this idea: the q^{th} quantile of a data set is a value such that at least $q\ percent$ of our sample items take on this value or less, and at least $(1-q)\ percent$ of our sample items take on this value or more.
- If q is a round number, we can it a *percentile*.

- For example, if your midterm score is in the 74th percentile, it means that 74% of students who took the midterm got grades lower or equal to yours, and 26% of students got grades higher or equal to yours.

Quantiles (continued)

- Arrange your data in ascending order (just like you did to calculate quartiles).
- Compute index i, the <u>position</u> of the q^{th} quantile: $i = \frac{q}{100} \cdot (n+1)$
- If i is an integer, inte q^{th} percentile is the value in the i^{th} position. If i is not an integer, interpolate.
- Example: 80th percentile: $i = q/100 \cdot (n+1) = 80/100 \cdot 71 = 56.8$

The value in the 56th position is **535**. Since the “exact” position is 56.8, we add 80% of the distance between 535 and the next value, 549: So the 80th percentile is: $535 + 0.8(549 - 535) = \mathbf{546.2}$

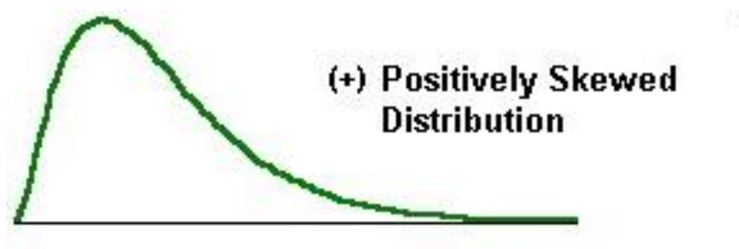
425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

- At least 80% of items take on a value of 546.2 or less.
- 56/70 = 0.8 or 80%.
- At least 20% of items take on a value of 546.2 or more.
- 14/70 = 0.2 or 20%.

Degree of symmetry: Skewness

- Skewness is an important measure of the shape of a distribution.
- The formula for skewness is complex: $skewness = \frac{\frac{1}{N} \cdot \sum_{i+1}^{N} (X-\mu)^3}{\sigma^3}$
- Skewness measures the degree of symmetry in a distribution. If a distribution is symmetric, then its skewness is equal to 0.

  - If the distribution has a right tail, we say that it is skewed right, or positively skewed. In this case, usually $mean > median$.



  - If the distribution has a left tail, we say that it is skewed left, or negatively skewed. In this case, usually $mean < median$.

Shape of a distribution: Kurtosis

- Kurtosis is a measure of the “peakedness” of a distribution.
- The formula for kurtosis is also complex:

$$kurtosis = \frac{\frac{1}{N} \cdot \sum_{i+1}^{N} (X-\mu)^4}{\sigma^4}$$

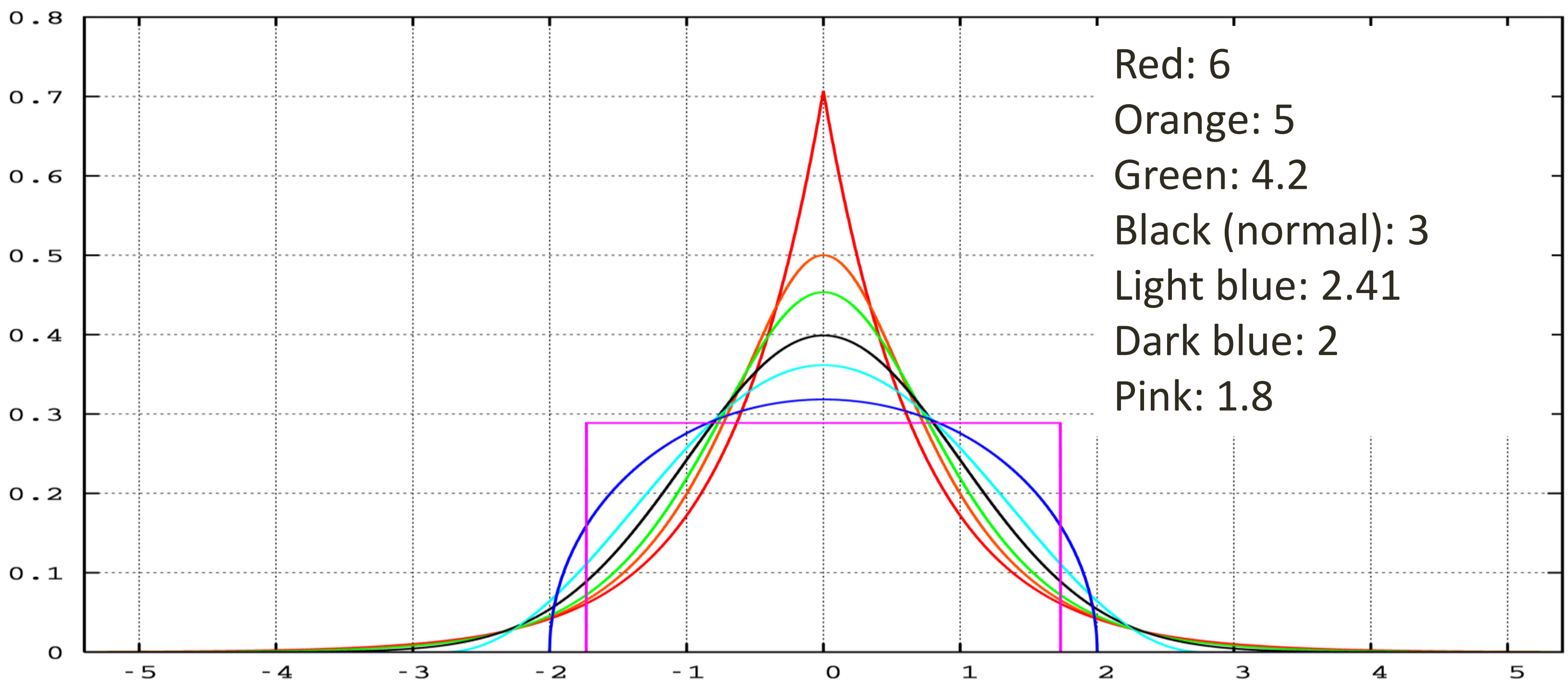
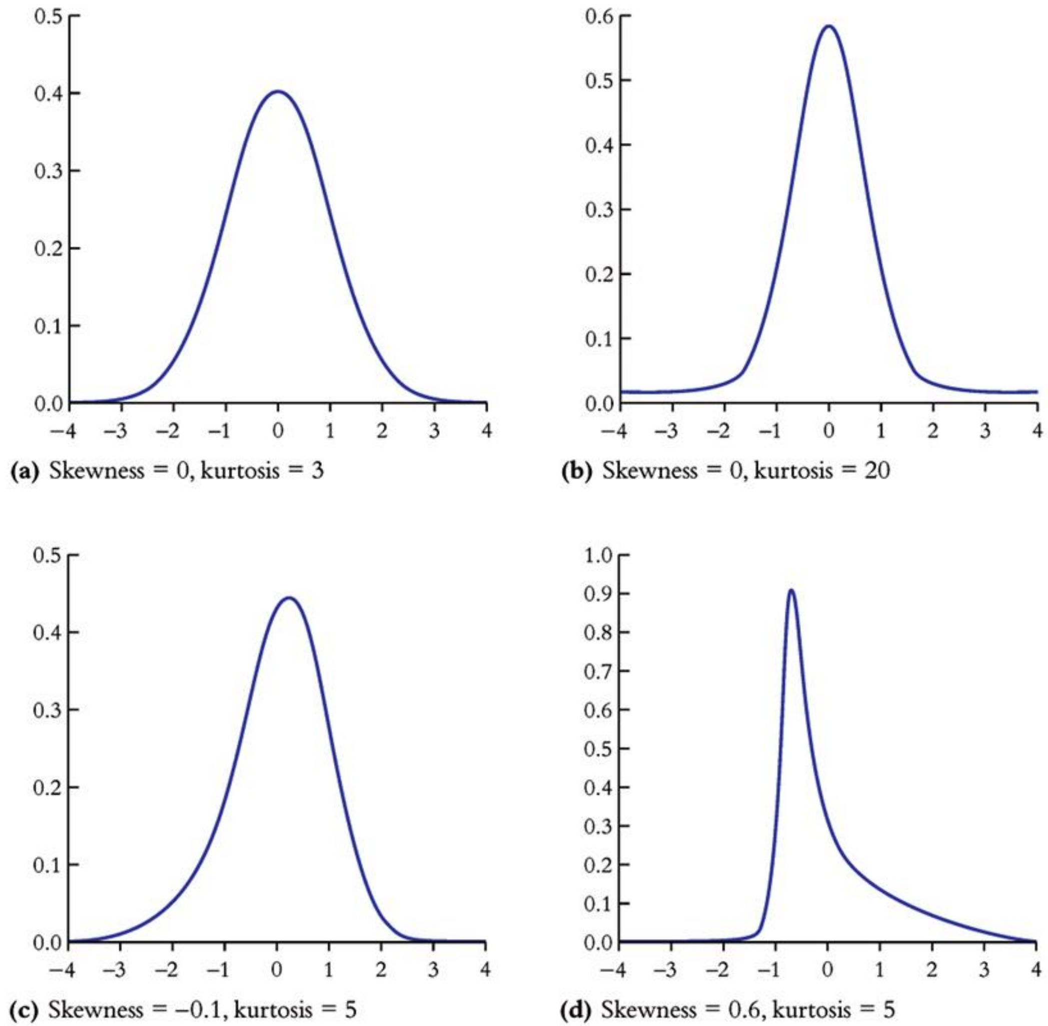


FIGURE 2.3 **Four Distributions with Different Skewness and Kurtosis**



All of these distributions have a mean of 0 and a variance of 1. The distributions with skewness of zero (a and b) are symmetric; the distributions with nonzero skewness (c and d) are not symmetric. The distributions with kurtosis exceeding 3 (b–d) have heavy tails.

Relationships between variables

- When we have a dataset that we would like to analyze, apart from looking at each variable (characteristic) separately, we might also want to describe the relationship between some variables.
- For example, we might want to see if higher incomes are associated with higher levels of schooling, or age.
- Two measures are particularly important here:

covariance and correlation

- Both measures tell us if two variables tend to move together in our data: higher values for one imply higher values for the other. Careful: NO CAUSAL EFFECT IS IMPLIED!
- The correlation coefficient is unit-free, so it makes comparisons easy.

Covariance

- Covariance in the population:

$$cov(X,Y) = \sigma_{XY} = \frac{1}{N} \cdot \sum_{i=1}^{N} \left((X_i - \mu_X) \cdot (Y_i - \mu_Y) \right)$$

- Sample covariance:

$$cov(X,Y) = s_{XY} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left((X_i - \bar{X}) \cdot (Y_i - \bar{Y}) \right)$$

Notice that here, as in the case of the sample variance, we divide by $(n-1)$ and not by (n)

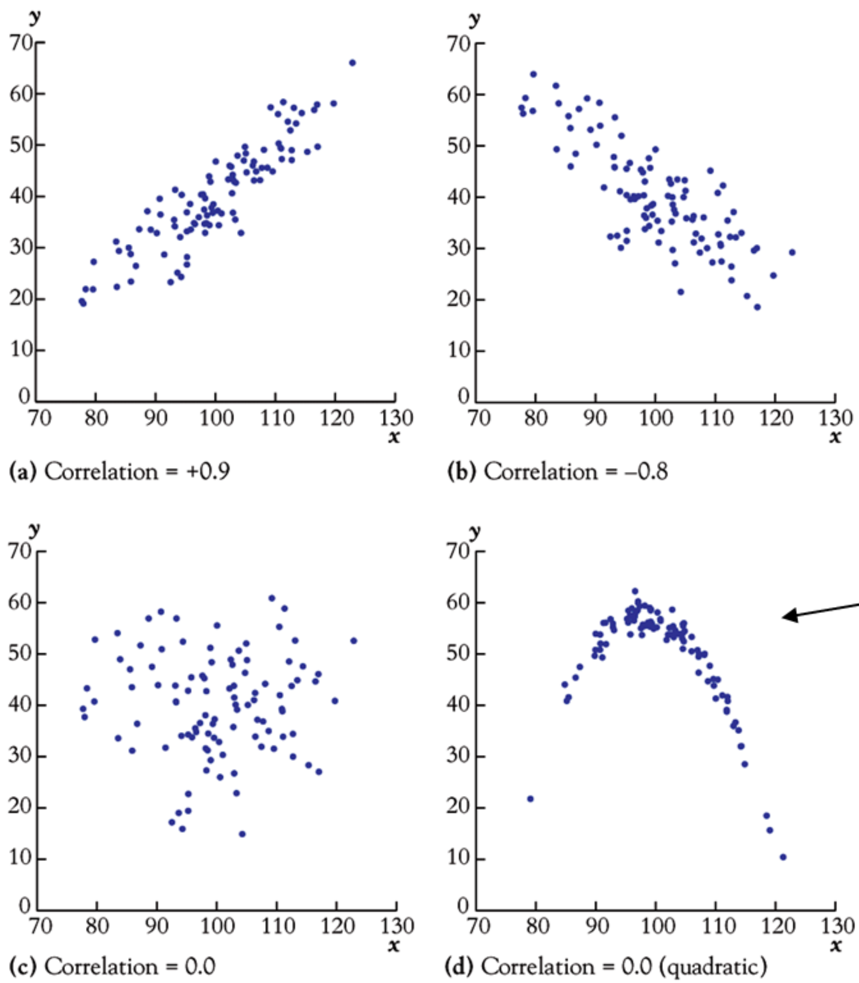
- $cov(X,Y) > 0 \rightarrow$ X and Y tend to move in the same direction
- $cov(X,Y) < 0 \rightarrow$ X and Y tend to move in opposite directions
- $cov(X,Y) = 0 \rightarrow$ X and Y are uncorrelated, there is no linear relationship between them.

Coefficient of correlation

- Population correlation coefficient: $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$
- Sample correlation coefficient: $r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$

- The correlation coefficient:
  - measures the strength of a **linear** relationship
  - takes values between -1 and 1: $|r_{XY}| \leq 1$. The closer to -1, the stronger the negative linear relationship. The closer to 1, the stronger the positive linear relationship.
  - has no units of measurement, so we can compare across different pairs of variables.
  - What will r_{XY} be if $X = Y$ for all i? What about if $X = -Y$? What would the scatterplot of X and Y look like? What can we tell about the slope?

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets



The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between X and Y. In Figure 3.3c, X is independent of Y and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.

Notice that in figure (d), correlation is 0, although clearly the two variables are related. Remember that r_{XY} only picks up **linear** relationships, so there could be a quadratic relationship that we are not measuring if we only look at correlation.