Quantitative Methods

Serena DeStefani – Lecture 4 - 7/8/2020

Announcements

- HW2 (due Tuesday)
- HW3 (due Thursday)
- Posted over the weekend
- Office hours on Monday after class

Announcements

- Office hours: Monday and Tuesday after class

- Midterm in two weeks
  - The normal model
  - Z-scores
  - Regression
  - Confidence intervals?
  - Surveys - Probability

Review: Shifting, Scaling, and *z*-Scores

- Converting to *z*-scores:

$$z = \frac{y - \bar{y}}{s}$$

  - Subtract the mean $\bar{y} - \bar{y} = 0$
  - Divide by the standard deviation $s/s = 1$
  - Shape ?
  - Center ?
  - Spread?

Example: SAT and ACT Scores

- How high does a college-bound senior need to score on

the **ACT** in order to make it into the top quarter of equivalent of SAT scores for a college with middle 50% between 1530 and 1850?

- SAT: Mean = 1500, Standard Deviation = 250
- ACT: Mean = 20.8, Standard Deviation = 4.8

- **Plan:** Want ACT score for upper quarter. Have $\bar{y}$ and s
- **Variables:** Both are quantitative. Units are points.

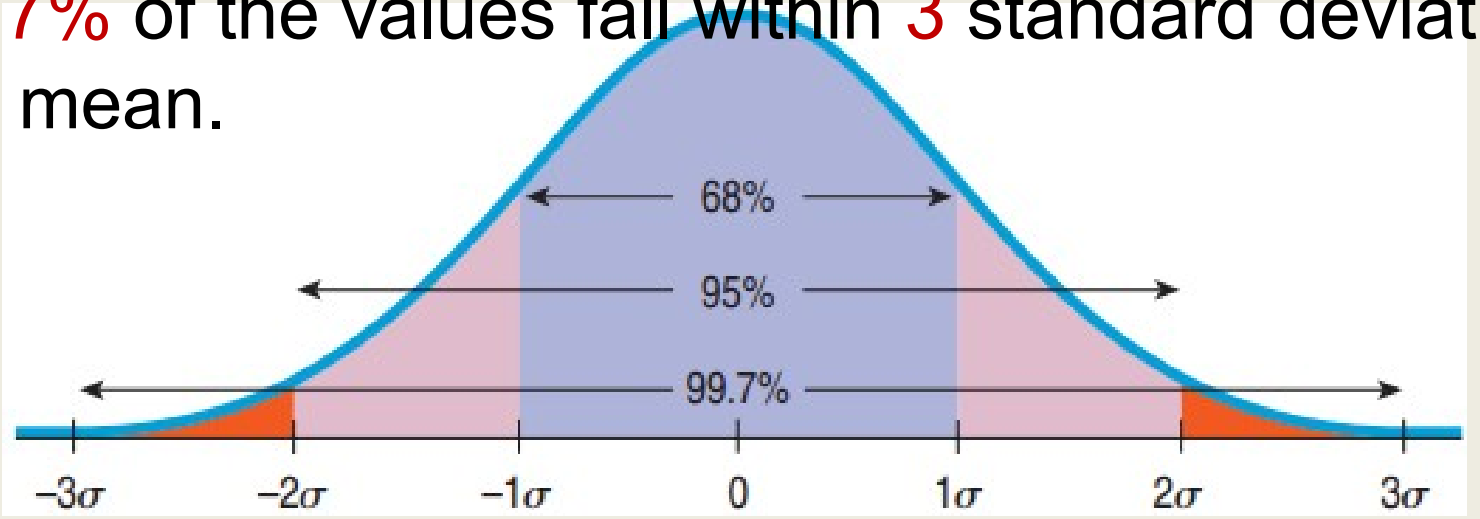
Show →Mechanics: Standardize the Variable

- It is known that the middle 50% of SAT scores are between 1530 and 1850, $\bar{y} = 1500$, $s = 250$
- The top quarter starts at 1850.
- Find the *z*-score: $z = \dfrac{1850 - 1500}{250} = 1.40$
- For the ACT, 1.40 standard deviations above the mean:

$$20.8 + 1.40(4.8) = 27.52$$

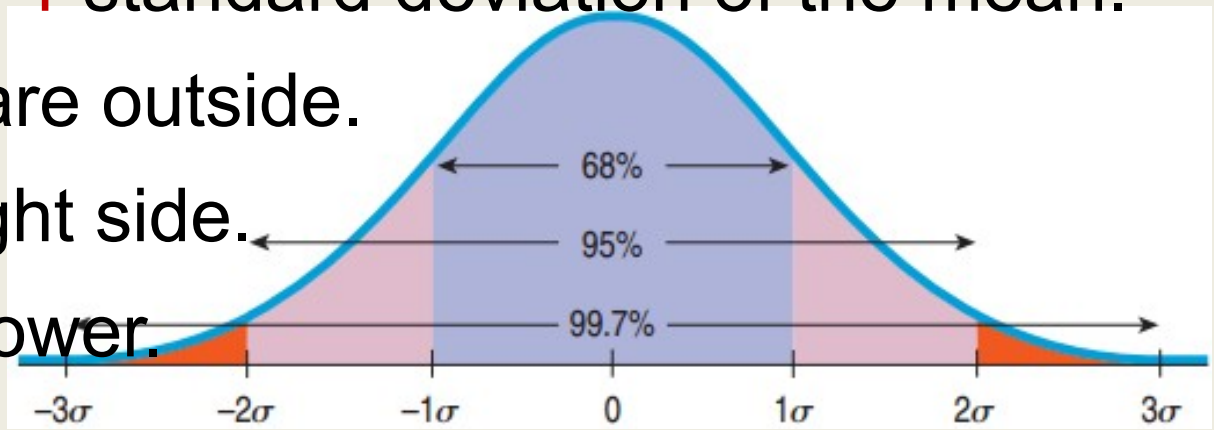
The 68-95-99.7 Rule

- 68% of the values fall within 1 standard deviation of the mean.
- 95% of the values fall within 2 standard deviations of the mean.
- 99.7% of the values fall within 3 standard deviations of the mean.



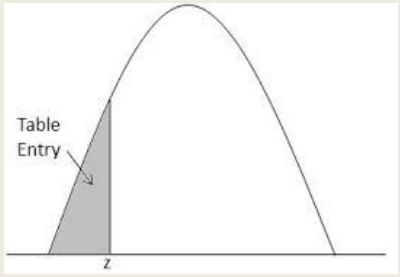
Example of the 68-95-99.7 Rule

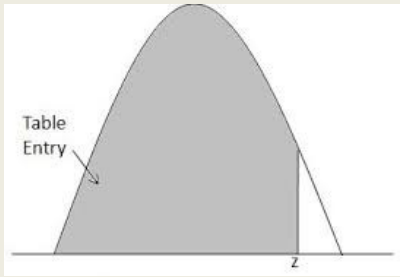
- In the 2010 winter Olympics men's slalom, Li Lei's time was 120.86 sec, about 1 standard deviation slower than the mean. Given the Normal model, how many of the 48 skiers were slower?
- About 68% are within 1 standard deviation of the mean.
- 100% – 68% = 32% are outside.
- "Slower" is just the right side.
- 32% / 2 = 16% are slower.
- 16% of 48 is 7.7.
- About 7 are slower than Li Lei.



Review: What if z is not −3, −2, −1, 0, 1, 2, or 3?

- We will use a table. It gives you the percentile to the left





- **Example:** Where do you stand if your SAT math score was 680? $\mu = 500$, $\sigma = 100$
- Note that the z-score is not an integer:

$$z = \frac{680 - 500}{100} = 1.8$$

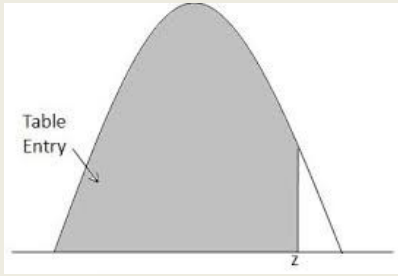
The Z table

Look for the z-score on the table: 1.8

Look for the second decimal place.

Result: 0.9641

96.4% of SAT scores are below 680.

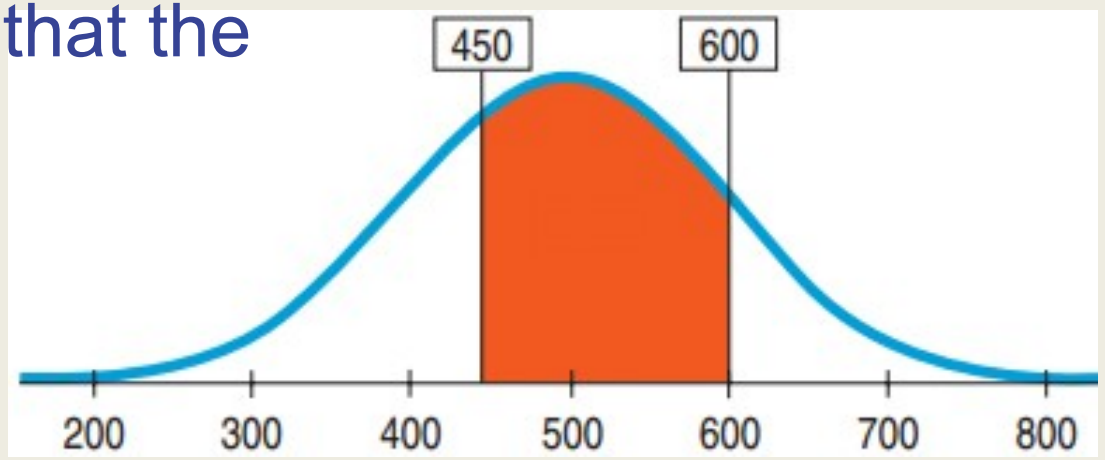


A Probability Involving "Between"

- What is the proportion of SAT scores that fall between 450 and 600? $\mu = 500$, $\sigma = 100$

- **Plan:** Probability that x is between 450 and 600
  = Probability that $x < 600$ – Probability that $x < 450$

- **Variable:** We are told that the Normal model works.
  $N(500, 100)$



A Probability Involving "Between"

- What is the proportion of SAT scores that fall between 450 and 600? $\mu = 500$, $\sigma = 100$
- z = (600-500)/100 = 1 z=(450-500)/100= -0.5

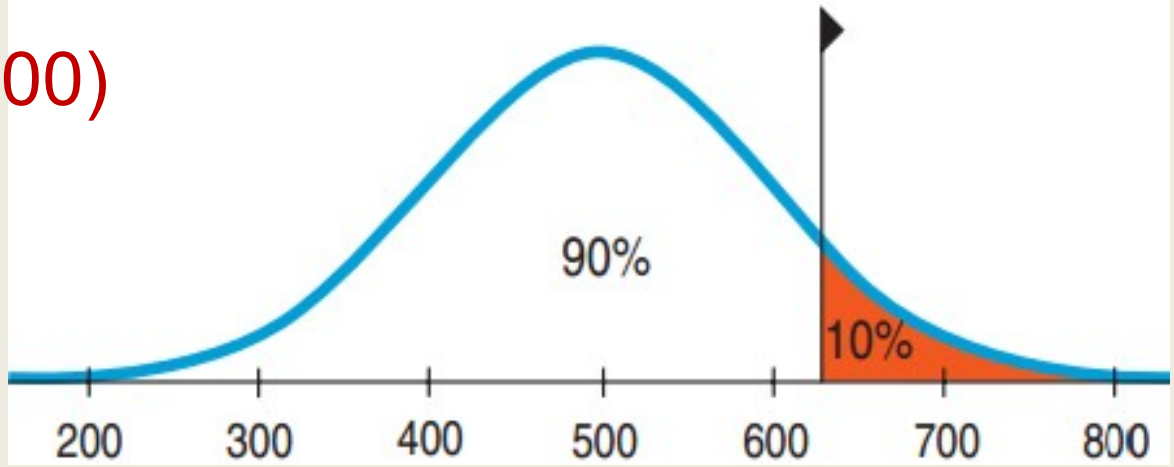
- Probability that x is between 450 and 600
  = Probability that x < 600 – Probability that x < 450
  Look for z-scores on the sides of the table, get percentage in the middle of it
  = 0.8413 – 0.3085 = 0.5328
- Conclusion: The Normal model estimates that about 53.28% of SAT scores fall between 450 and 600.

From Percentiles to Scores: *z* in Reverse

- Suppose a college admits only people with SAT scores in the top 10%. How high a score does it take to be eligible? $\mu = 500$, $\sigma = 100$

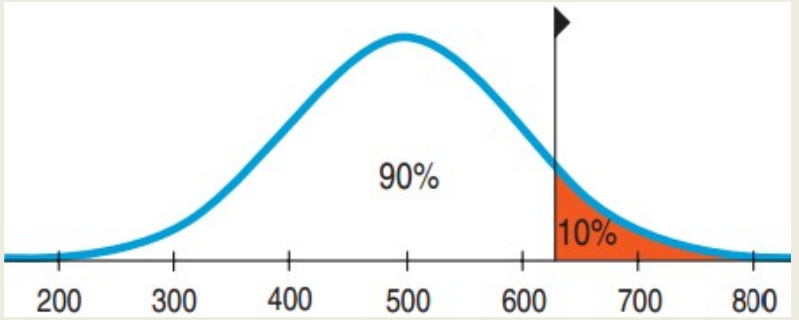
  - **Plan:** We are given the probability and want to go backwards to find *x*.

  - **Variable:** *N*(500, 100)



From Percentiles to Scores: *z* in Reverse

- Suppose a college admits only people with SAT scores in the top 10%. How high a score does it take to be eligible? $\mu = 500$, $\sigma = 100$
- Look for percentage in middle of table, get z-score on the sides of it
- z = 1.29



- (x-500)/100 = 1.29
- x= 1.29*100 +500 = 629
- Conclusion: Because the school wants the SAT Verbal scores in the top 10%, the cutoff is 629.

Quality control: Underweight Cereal Boxes

- Based on experience, a manufacturer makes cereal boxes that fit the Normal model with mean 16.3 ounces and standard deviation 0.2 ounces, but the label reads 16.0 ounces. What fraction will be underweight?

- **Plan:** Find Probability that $x < 16.0$
- **Variable:** $N(16.3, 0.2)$

Underweight Cereal Boxes

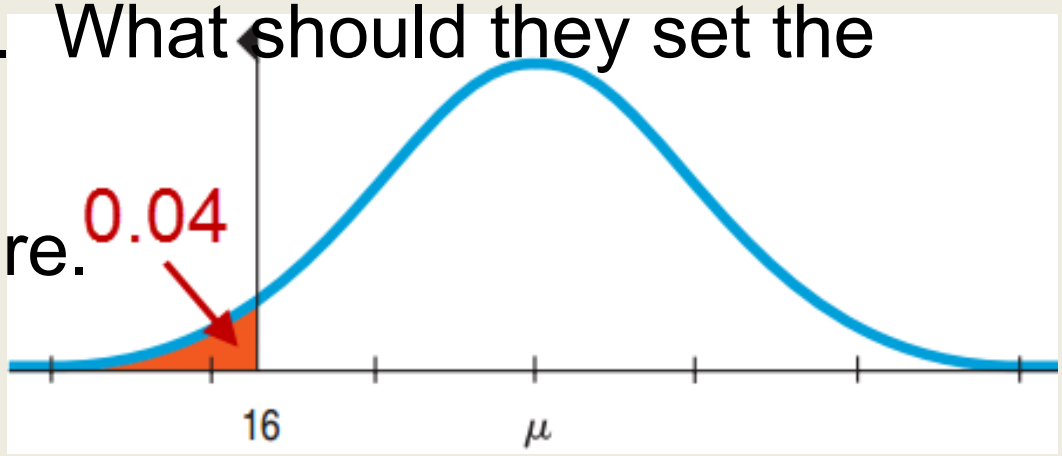
- What fraction of the cereal boxes will be underweight (less than 16.0)?
  $\mu = 16.3$, $\sigma = 0.2$

- $z = (16.0-16.3)/0.2 = -1.5$
- Probability $x < 16.0 = 0.0668$

- Conclusion: I estimate that approximately 6.7% of the boxes will contain less than 16.0 ounces of cereal.

Underweight Cereal Boxes Part II

- Lawyers say that 6.7% is too high and recommend that at most 4% be underweight. What should they set the mean at? $\sigma = 0.2$



- Mechanics: Sketch a picture. What do we need?
- $z = -1.75$
- Find $16 + 1.75(0.2)$
  $= 16.35$ ounces

- Conclusion: The company must set the machine to average 16.35 ounces per box.

Underweight Cereal Boxes Part III

- The CEO vetoes that plan and sticks with a mean of 16.2 ounces and 4% weighing under 16.0 ounces. She demands a machine with a lower standard deviation. What standard deviation must the machine achieve?

- **Plan:** Find σ such that Probability $x < 16.0 = 0.04$.
- **Variable:** $N(16.2, ?)$

Underweight Cereal Boxes Part III

- What standard deviation must the machine achieve? $N(60.2, ?)$
- From before, $z = -1.75$

$$-1.75 = \frac{16.0 - 16.2}{\sigma}$$

- $1.75\sigma = 0.2, \quad \sigma = 0.114$
- Conclusion: The company must get the machine to box cereal with a standard deviation of no more than 0.114 ounces. The machine must be more consistent.

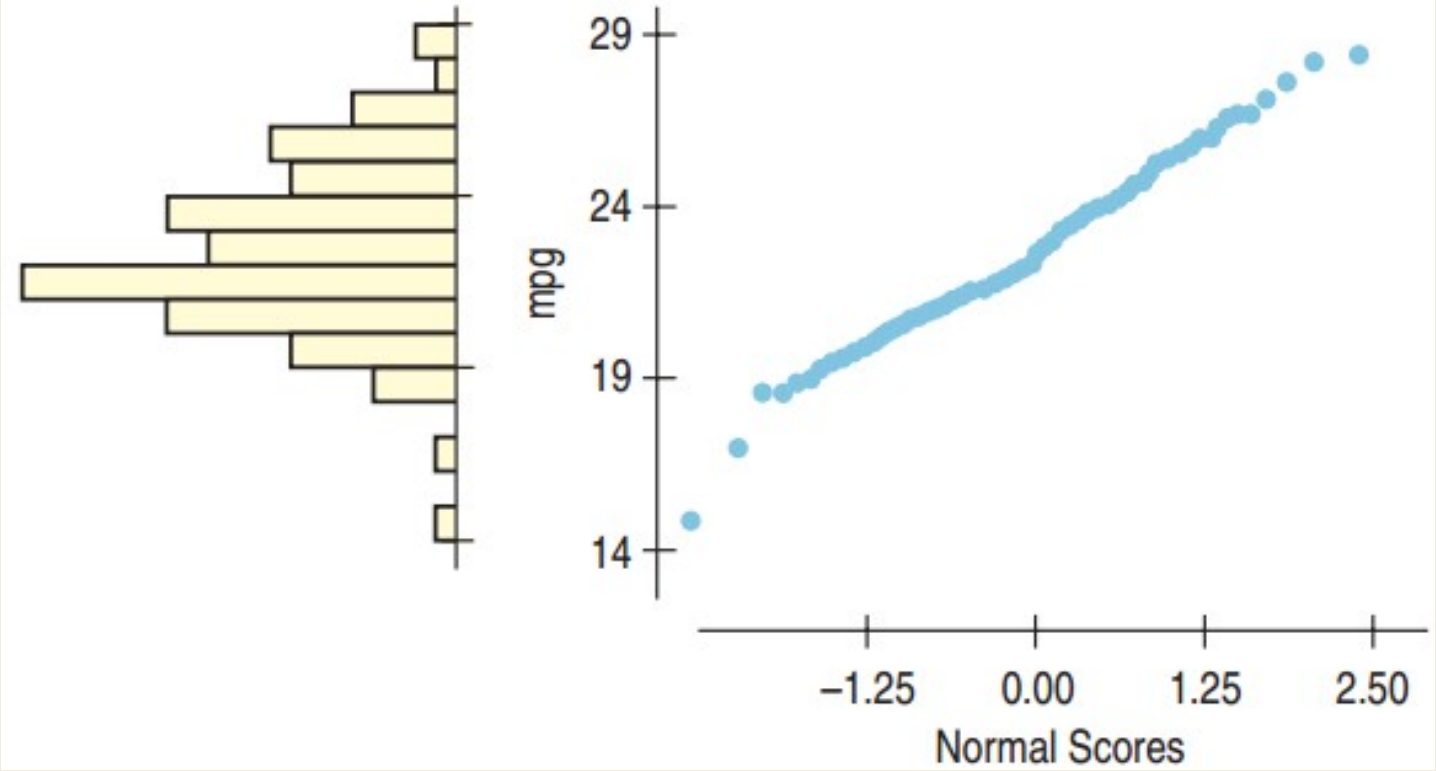
5.5

Normal Probability Plots

Checking if the Normal Model Applies

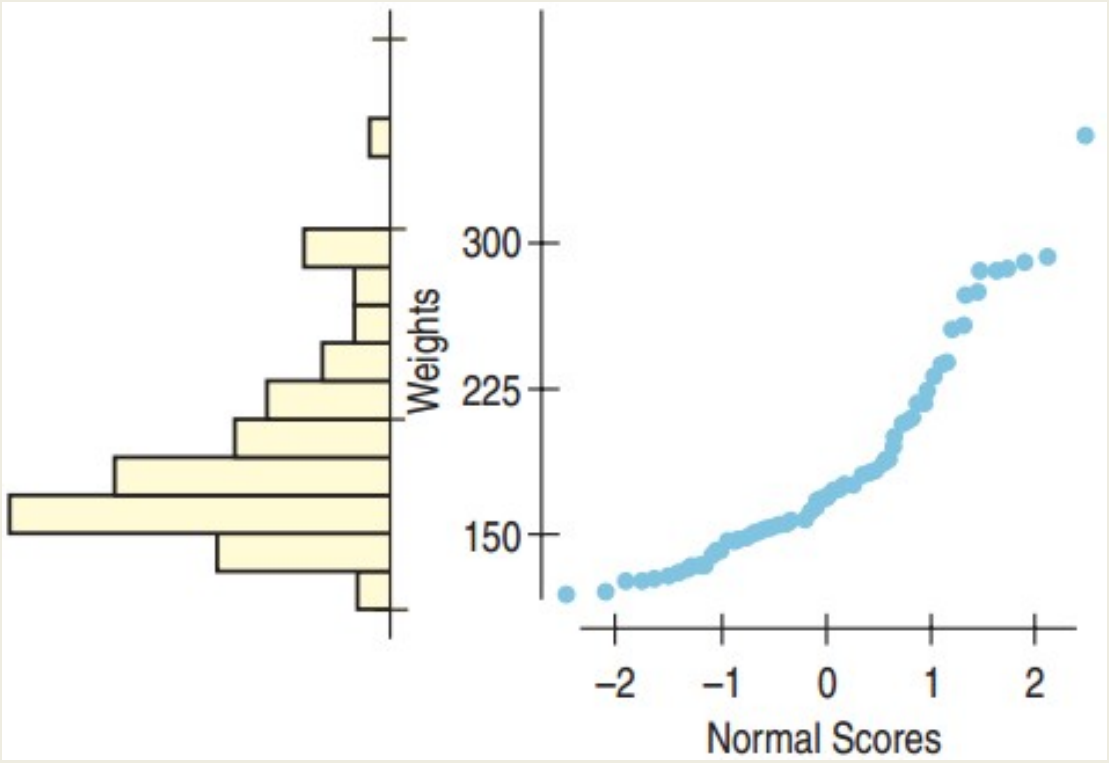
- A histogram will work, but there is an alternative method.
- Instead use a Normal Probability Plot.
  - Plots each value against the *z*-score that would be expected had the distribution been perfectly normal.
  - If the plot shows a line or is nearly straight, then the Normal model works.
  - If the plot strays from being a line, then the Normal model is not a good model.

The Normal Model Applies



- The Normal probability plot is nearly straight, so the Normal model applies. Note that the histogram is unimodal and somewhat symmetric.

The Normal Model Does Not Apply



- The Normal probability plot is not straight, so the Normal model does not apply applies. Note that the histogram is skewed right.

What Can Go Wrong

- Don't use the Normal model when the distribution is not unimodal and symmetric.
  - Always look at the picture first.
- Don't use the mean and standard deviation when outliers are present.
  - Check by making a picture.
- Don't round your results in the middle of the calculation.
  - Always wait until the end to round.
- Don't worry about minor differences in results.
  - Different rounding can produce slightly different results.

Chapter 6

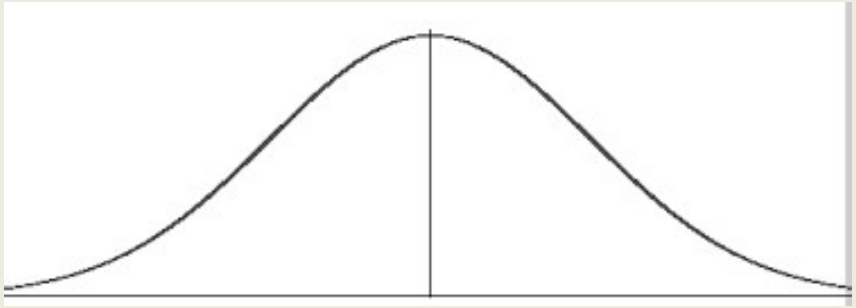
Scatterplots, Association, and Correlation

History of Statistics

- Demography
- Astronomy: how is measurement error distributed?
- Gambling: what are the odds of coin tosses?

- → study of binomial distribution
- → Gauss (1809) and Laplace (1812) find the equation of the Normal Distribution → the Law of Errors

- There was some variability in astronomy, but we solved it!

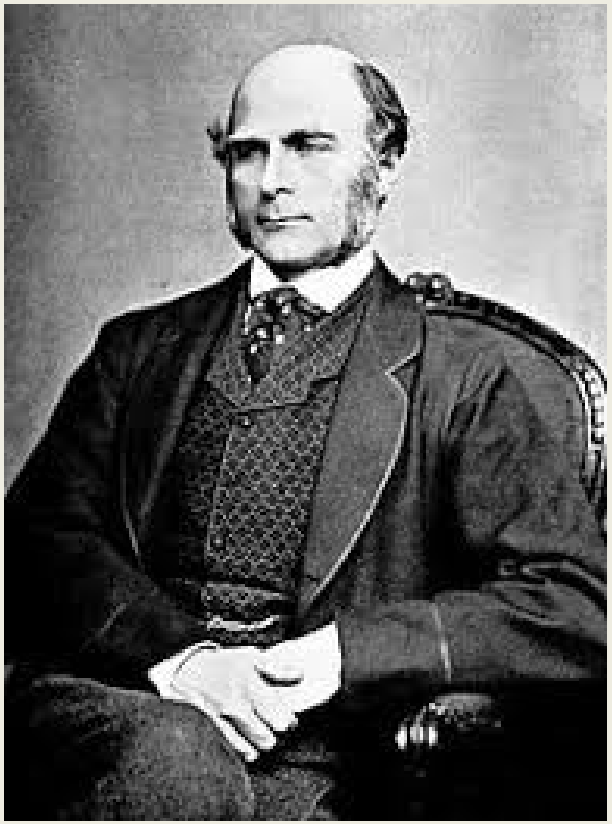


History of Statistics

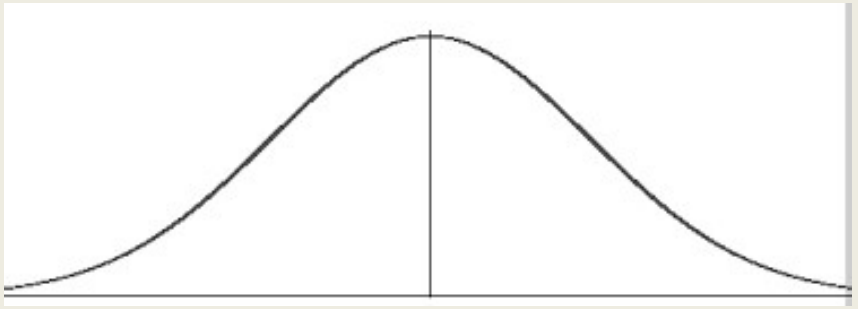
- Then… not much else happened for some time
- Then in 1859 Darwin published “The Origin of Species”
- Problem of variation and hereditability
- How to bring order into chaos?

History of Statistics: Galton

- Sir Francis Galton (1822 – 1911)
- Darwin's half cousin
- Studied variability and hereditary traits
- Problem: how to compare different features? → by their variation
- Idea: features that vary together
- are "co-related"



Galton



- Law of error

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the law of frequency of error. The law would have been personified by the Greeks if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob, the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason." (1889)

Galton

- Law of error

- **Regression** to the mean (1877)

- How could one compare different measures of anthropometric variable ?

- Compare them by their variation – on scales based on their own variability

Galton

- Compare variables by their variation – on scales based on their own variability

- **Correlation**: “two variables are said to correlated when variation in one is accompanied on the average by more or less variation on the other, and *in the same direction*”

History of Statistics: Pearson

- Karl Pearson (1857 – 1936)

- Galton's student

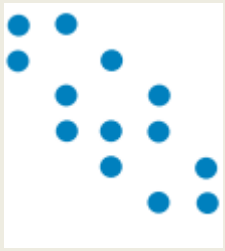
- Studied correlation in the context of variability and hereditary traits
- Came up with the **correlation coefficient**

6.1

Scatterplots

The Direction of the Association

- Negative Direction: As one goes up, the other goes down.

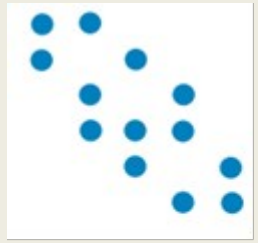


- Positive Direction: As one goes up, the other goes up also.

- No Direction:

Form

- Linear: The points cluster near a straight line.

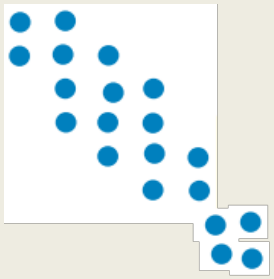


- Gently curves in a direction. May be able to straighten with a transformation.

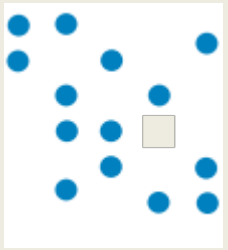
- Curves up and down. Difficult to straighten

Strength of the Relationship

- Strong Linear Relationship:



- Moderate Linear Relationship:

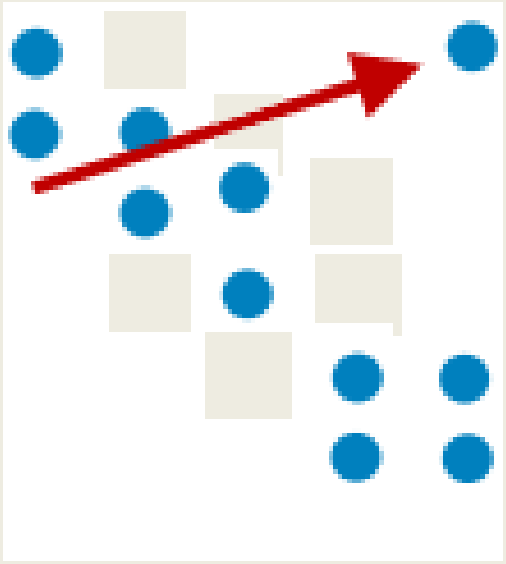


- No Linear Relationship:

Outliers

- An outlier is a point on a scatterplot that stands away from the overall pattern of the scatterplot.

- Outliers are almost always interesting and always deserves special attention.



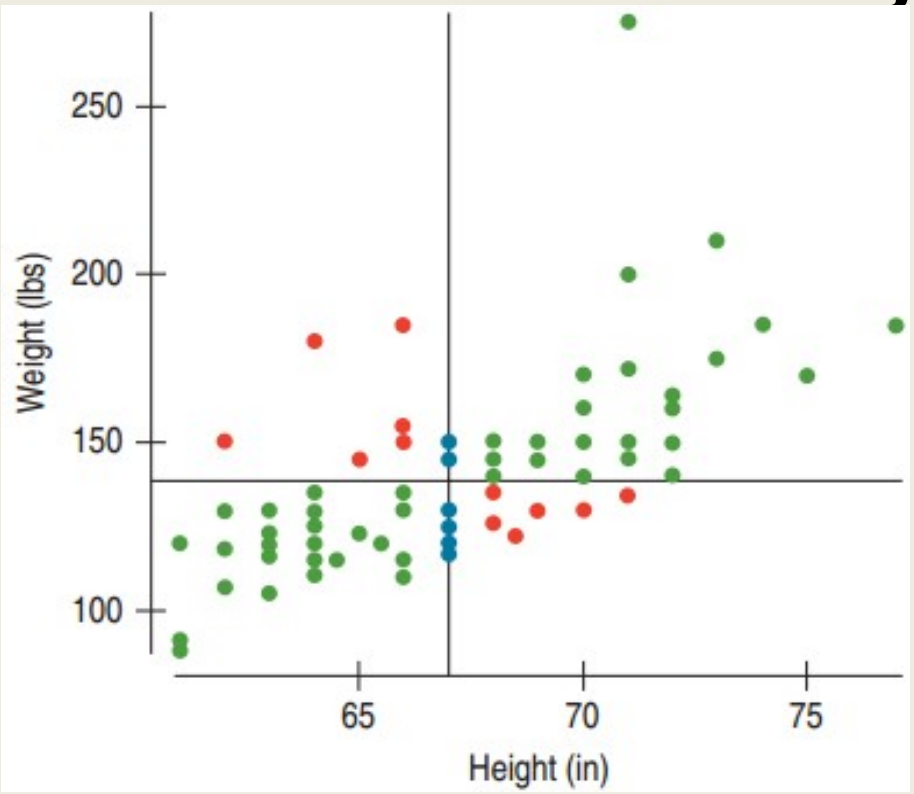
6.2

Correlation

Height and Weight

- How strong is the association between height and weight?

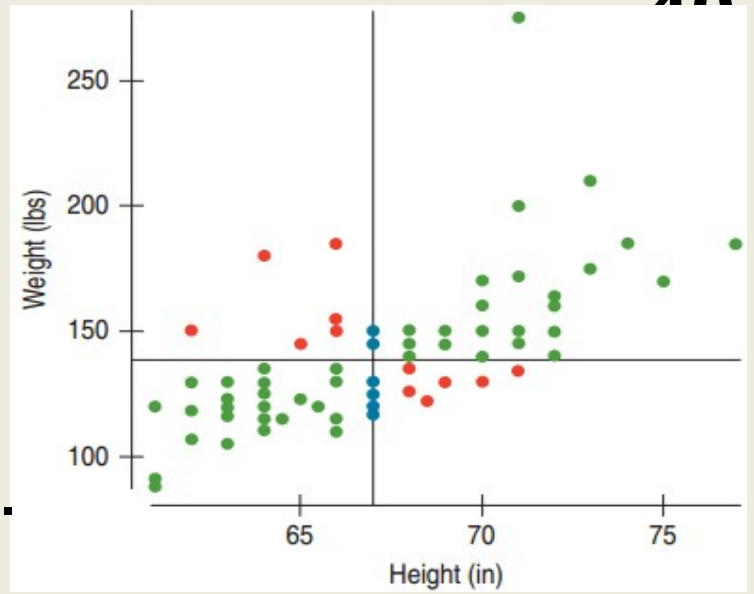
- It looks **positive**. How do we measure it?



- Positive association means: below/above average height predicts below/above average weight…

- Green → +, Red → -, Blue → No Association

Correlation



- For the green dots: *z*-scores have the same sign, so multiplying the z-scores produces a positive value.

- For the red dots: *z*-scores have opposite signs, so multiplying the *z*-scores produces a negative value.

- Define the correlation coefficient by an almost average product of the z-scores: $r = \dfrac{\sum z_x z_y}{n-1}$

Assumptions and Conditions for Correlation

- To use *r*, there must be a true underlying linear relationship between the two variables.

- The variables must be quantitative.

- The pattern for the points of the scatterplot must be reasonably straight.

- Outliers can strongly affect the correlation. Look at the scatterplot to make sure that there are no strong outliers.

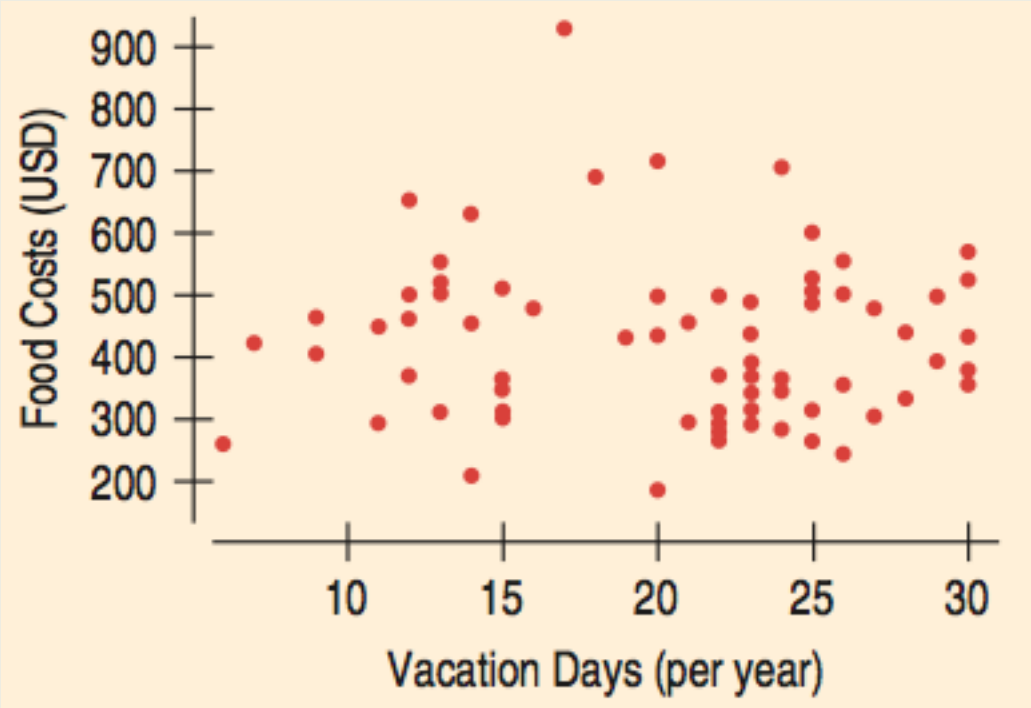
Example: Clothes and Food Revisited

- The scatterplot indicates a straight-line pattern. The variables are both quantitative ($), and there are no strong outliers away from the linear pattern.

- The correlation of $r = 0.614$ represents a strong positive association.

Example: Vacation and Food Revisited

- The scatterplot indicates that there may be no underlying linear relationship between vacation days and food costs.



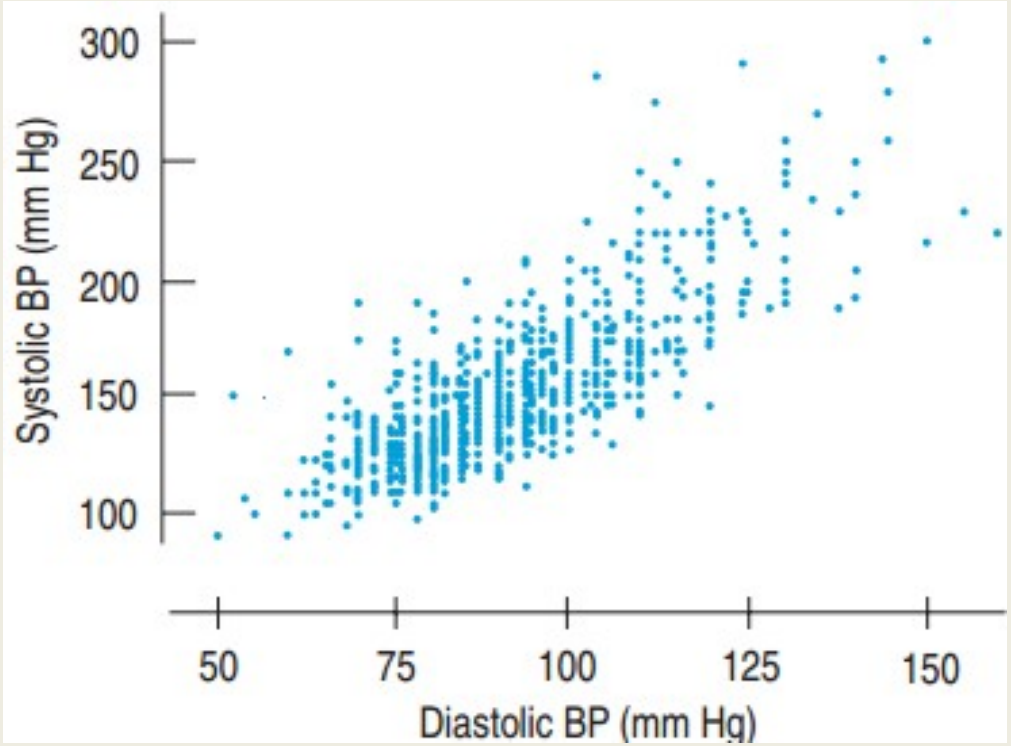
- The correlation is $r = -0.040$ supports that there may be no linear association between the two.

Systolic and Diastolic Blood Pressure

- Plan: Examine the relationship between the two types of blood pressure.

- Variables: Systolic (SBP) and Diastolic (DBP) blood pressure measured for each of 1406 people from Framingham, MA.

Blood Pressure Continued



- Plot:
  - ✓ Both variables are quantitative.
  - ✓ The scatterplot is quite straight.
  - ✓ There are no extreme outliers.
- Therefore, correlation is a suitable measure of association.

Blood Pressure Continued

- **Conclusion:** The correlation of 0.792 indicates that there is a positive association between systolic and diastolic blood pressure. There is a tendency for DBP to be high when SBP is high.

Properties of Correlation

- $r > 0 \rightarrow$ positive association
- $r < 0 \rightarrow$ negative association
- $-1 \leq r \leq 1$, with $r = -1$ only if the points all lie exactly on a negatively sloped line and $r = 1$ only if the points all lie exactly on a positively sloped line.
- Interchanging x and y does not change the correlation.
- r has no units.

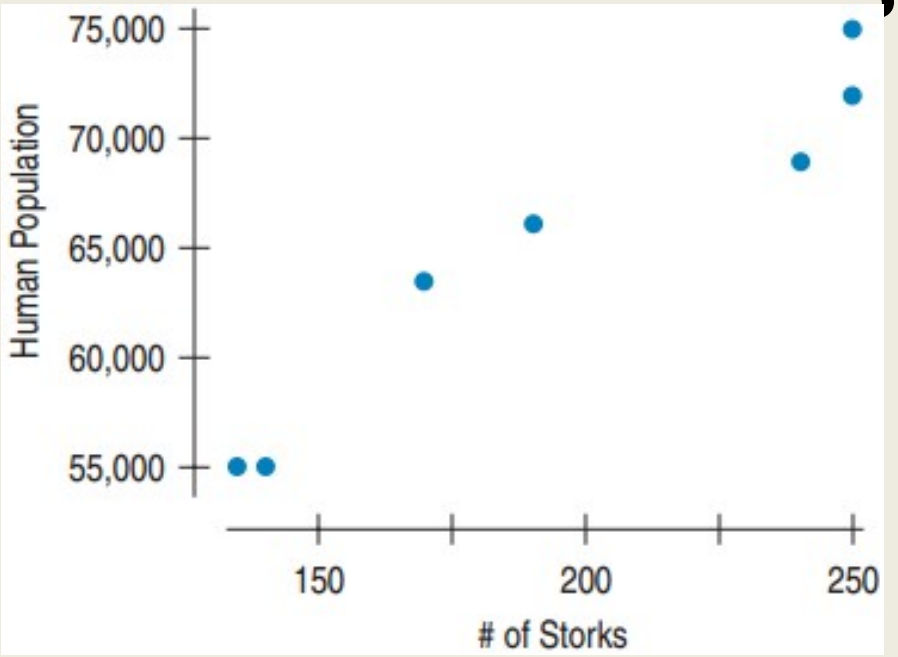
Properties of Correlation Continued

- Changing the units of *x* or *y* does not affect *r*.
  - Measuring in dollars, cents, or Euros will all produce the same correlation.
- Correlation measures the strength of the linear association between the two variables.
- Correlation is sensitive to outliers. An extreme outlier can cause a dramatic change in r.
- The adjectives *weak*, *moderate*, and *strong* can describe correlation, but there are no agreed upon boundaries.

6.3

Warning:
Correlation ≠ Causation

Storks and Babies



- There is a clear positive association between the number of storks and the population.
- This does not prove that an increase in storks has caused an increase in babies being born.
- Causation is in reverse. Storks nest on house chimneys, so the increased population has increased nesting sites.

Reasons for Correlation?

- Causation is a possibility, but more must be done to prove causation.
- The causation could be in reverse (*y* causes *x*)
- A lurking variable may cause both.

How to Report Correlation

- Bad: Raising salaries increases productivity.
- Good: Employees with higher salaries **tend to be** more productive.

- Bad: $r = -0.99$. This proves that drinking more red wine lowers cholesterol.
- Good: There is a strong negative association between red wine consumption and cholesterol level.

Correlation Tables

- Often surveys or experiments contain many quantitative variables. A table can be used to show the correlation for each pair of variables.
- Why do all the diagonals have 1.000?

	Assets	Sales	Market Value	Profits	Cash Flow	Employees
Assets	1.000					
Sales	0.746	1.000				
Market Value	0.682	0.879	1.000			
Profits	0.602	0.814	0.968	1.000		
Cash Flow	0.641	0.855	0.970	0.989	1.000	
Employees	0.594	0.924	0.818	0.762	0.787	1.000

Cautions about Correlation Tables

	Assets	Sales	Market Value	Profits	Cash Flow	Employees
Assets	1.000					
Sales	0.746	1.000				
Market Value	0.682	0.879	1.000			
Profits	0.602	0.814	0.968	1.000		
Cash Flow	0.641	0.855	0.970	0.989	1.000	
Employees	0.594	0.924	0.818	0.762	0.787	1.000

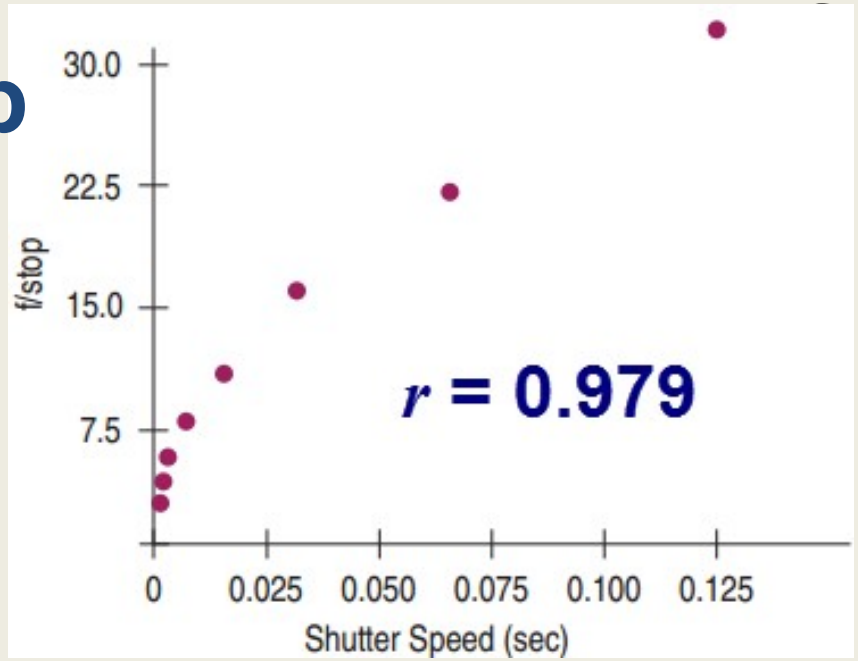
- Don't blindly read a correlation table without first looking at each of the many scatterplots to check for linearity.

- You must also check for outliers.

6.4

Straightening Scatterplots

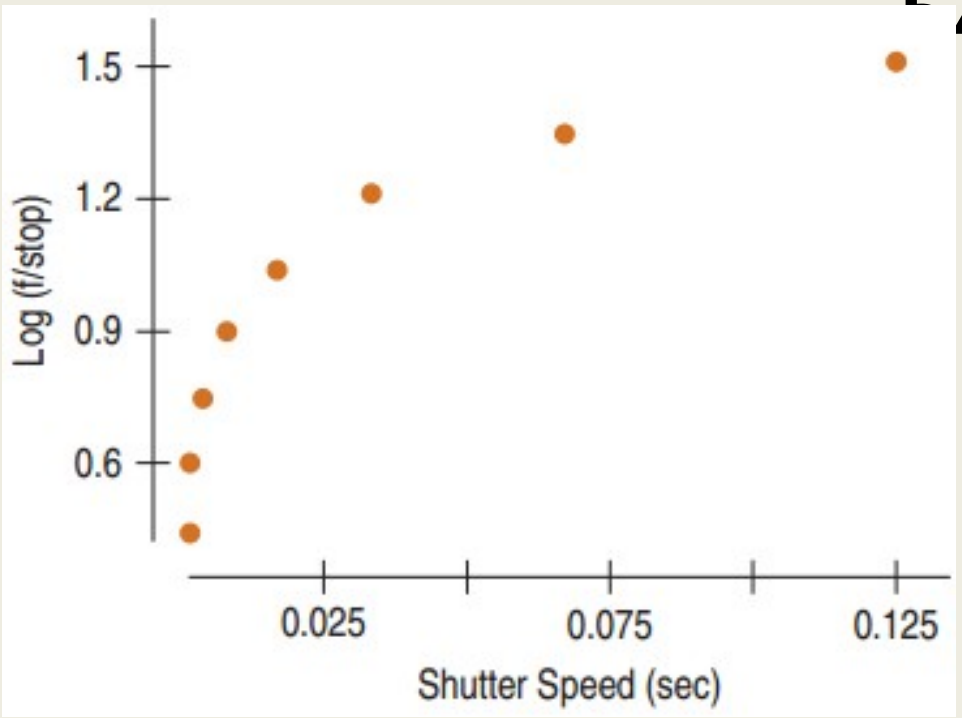
Shutter Speed and f/stop



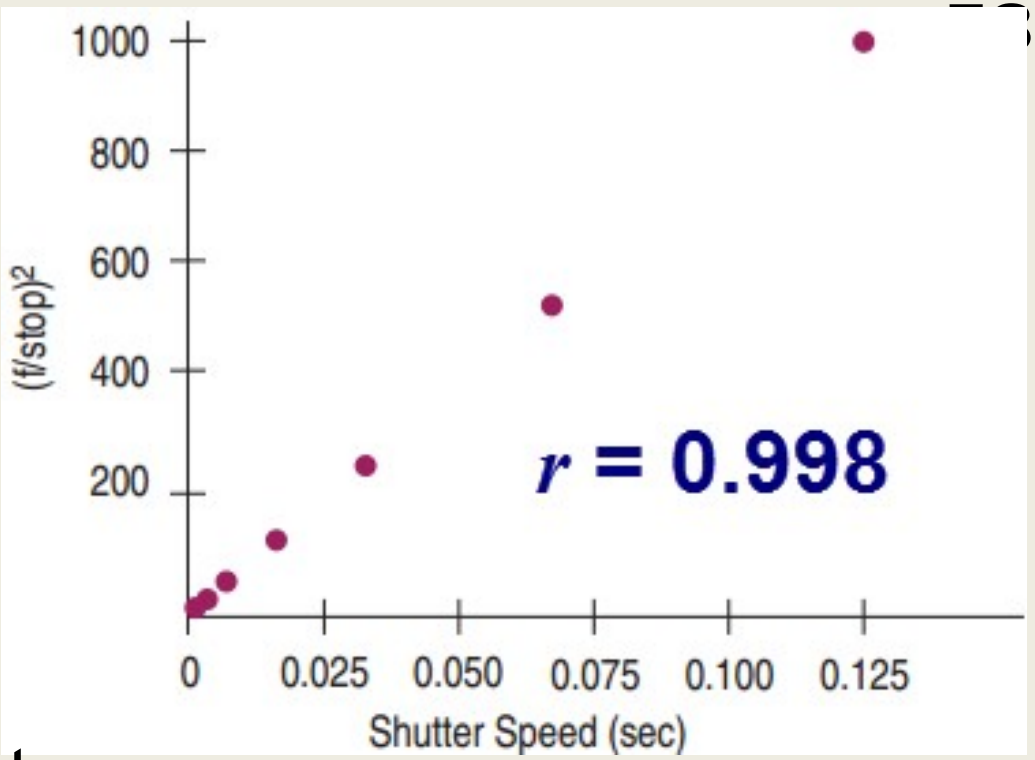
- What's wrong with concluding that since $r = 0.979$ that shutter speed and f/stop are strongly correlated?
- The scatter plot indicates that the two variables are not linearly related.
- **Solution:** Use a transformation, much like we did for histograms.

Re-expressing with log?

- How about the log transformation?
- This is even less linear!



Re-Expressing



- The shutter speed and the **square** of the f/stop are linearly related.

- Now we can conclude that there is a very strong correlation between shutter speed and the **square** of the f/stop.

Guidelines for Re-Expressions

- Scatterplot bends downwards. ➢ y^2
- Scatterplot is linear. ➢ No Change
- For data that is a count ➢ $y^{1/2}$
- For data that is always positive ➢ $\log y$
- If nothing else seems to work try ➢ $y^{-1/2}$
- For ratios such as miles per gallon ➢ $1/y$

What Can Go Wrong?

- Don't say "correlation" when you mean "association."
  - Correlation implies a **linear** relationship. Association means any relationship.
- Don't correlate categorical variables.
  - It makes no sense to say *car model* and *personality type* are correlated.
- Don't confuse correlation with causation.
  - Correlation only implies general tendencies.

What Can Go Wrong?

- Make sure the association is linear.
  - Always look at the scatterplot to check.
- Don't assume the association is linear just because the correlation coefficient is high.
  - Always look at the scatterplot to check.
- Beware of outliers!
  - $r = 0.5$, but there is no correlation between shoe size and IQ.