Quantitative Methods

Serena DeStefani – Lecture 3 - 7/8/2020

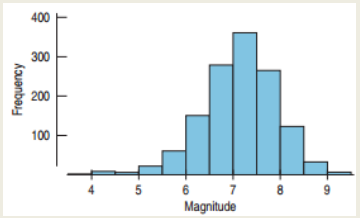
Announcements

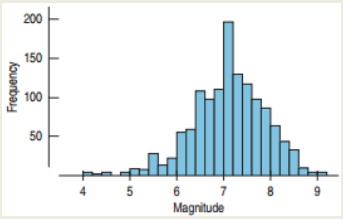
- Datacamp due tomorrow before class
- HW2 (due Tuesday)
- HW3 (due Thursday)
- Posted over the weekend
- Office hours on Monday after class

- Today: CH 5

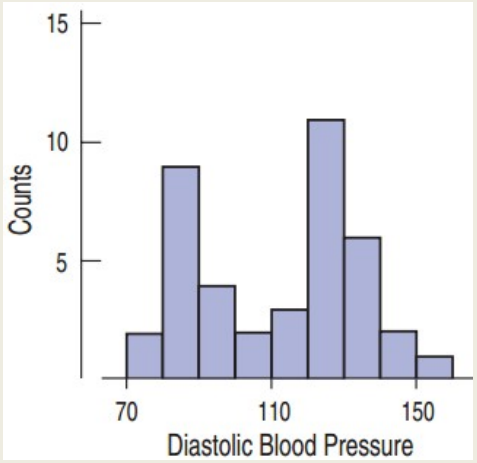
Review

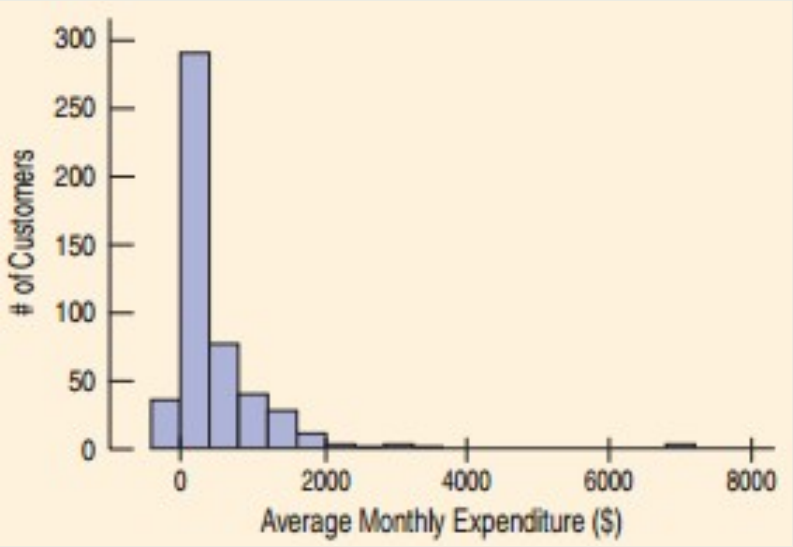
- How to display quantitative data: histograms
  - bin width





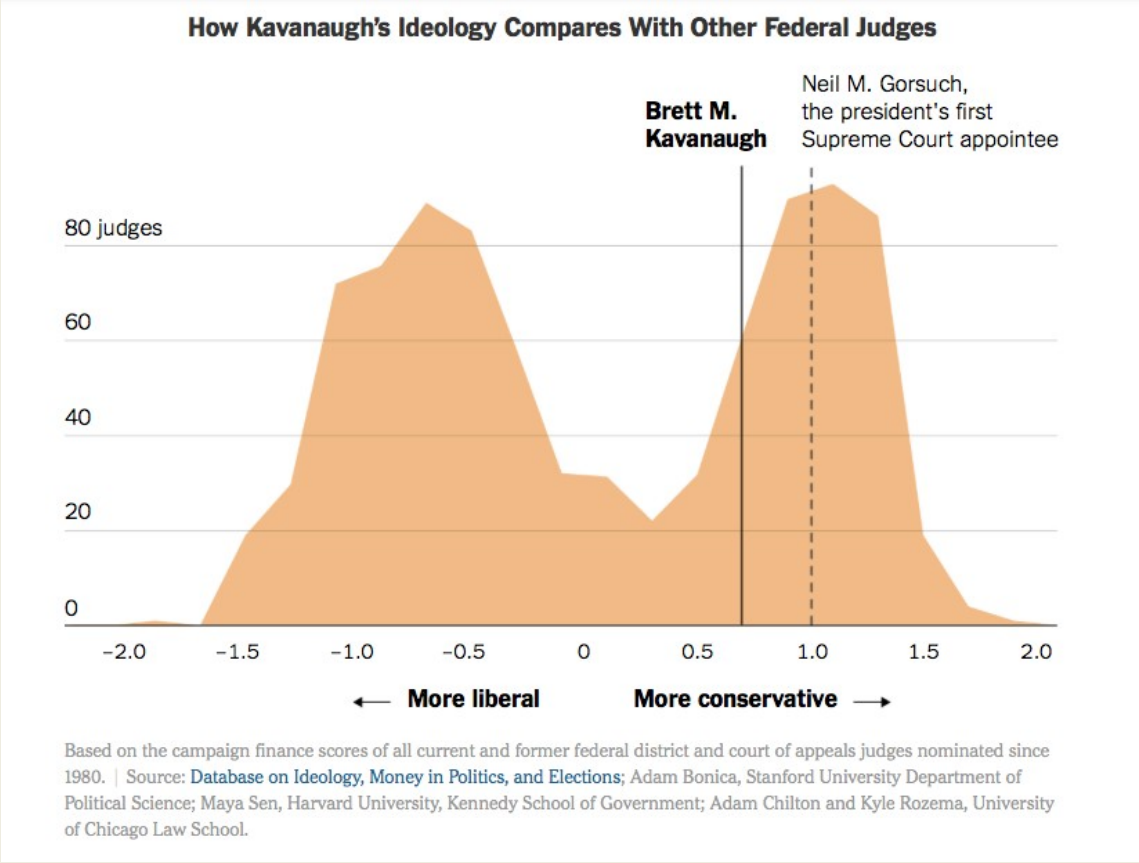
  - shape: mode, simmetry, skew
  - outliers





Review

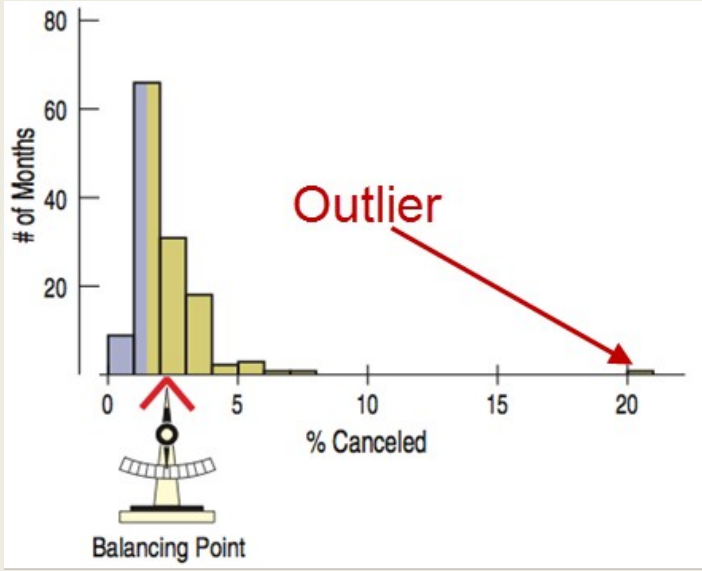
- Example of bimodal distribution from The New York Time



Review

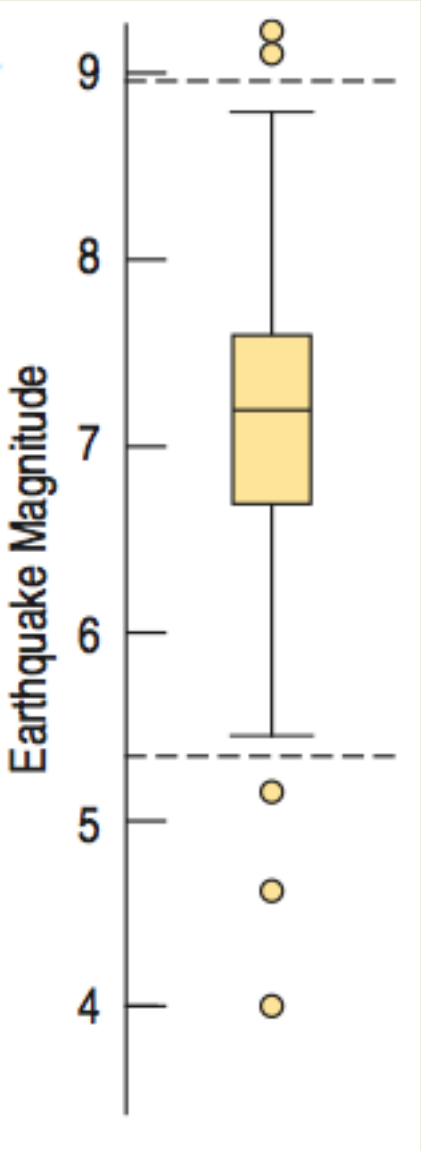
- How to display quantitative data: histograms
  - CENTER: median, mean
  - Median: the center of data values
  - not sensible to outliers
  - Mean: average
  - sensible to outliers
- Mode: value that appears most often

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



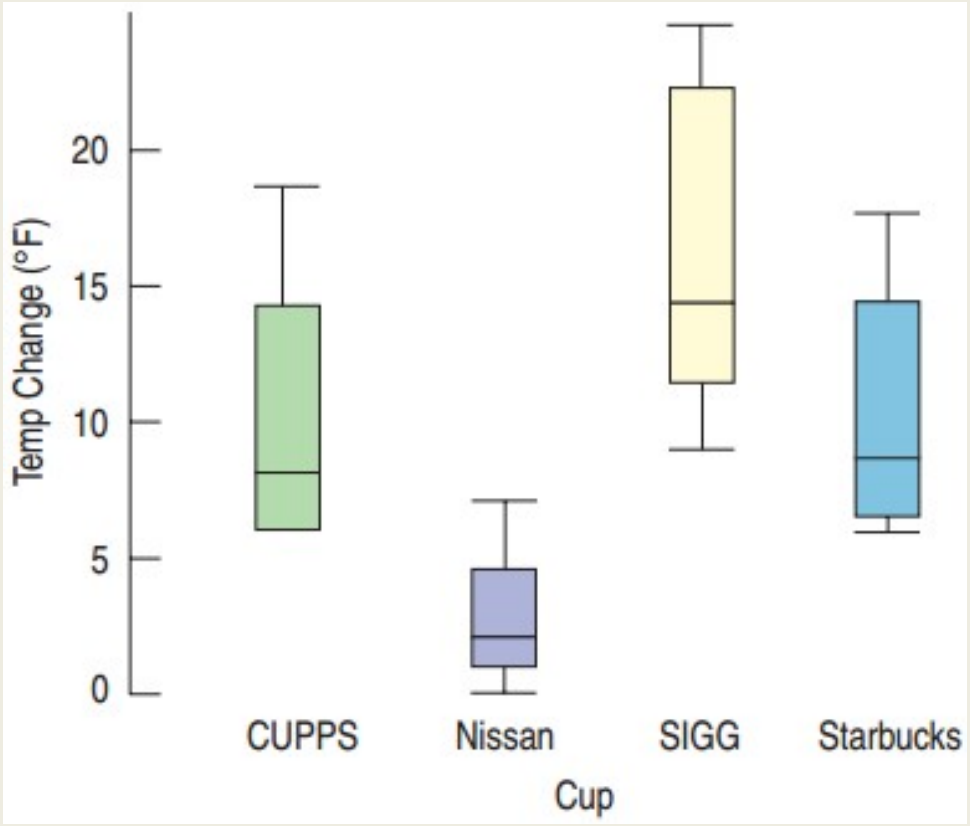
Review

- How to display quantitative data: boxplots
  - SPREAD: range, IQR
  - range: sensible to outliers
  - five numbers summary
  - IQR
  - boxplot



Mechanics

	Min	Q1	Median	Q3	Max	IQR
CUPPS	6°F	6	8.25	14.25	18.50	8.25
Nissan	0	1	2	4.50	7	3.50
SIGG	9	11.50	14.25	21.75	24.50	10.25
Starbucks	6	6.50	8.50	14.25	17.50	7.75



Review

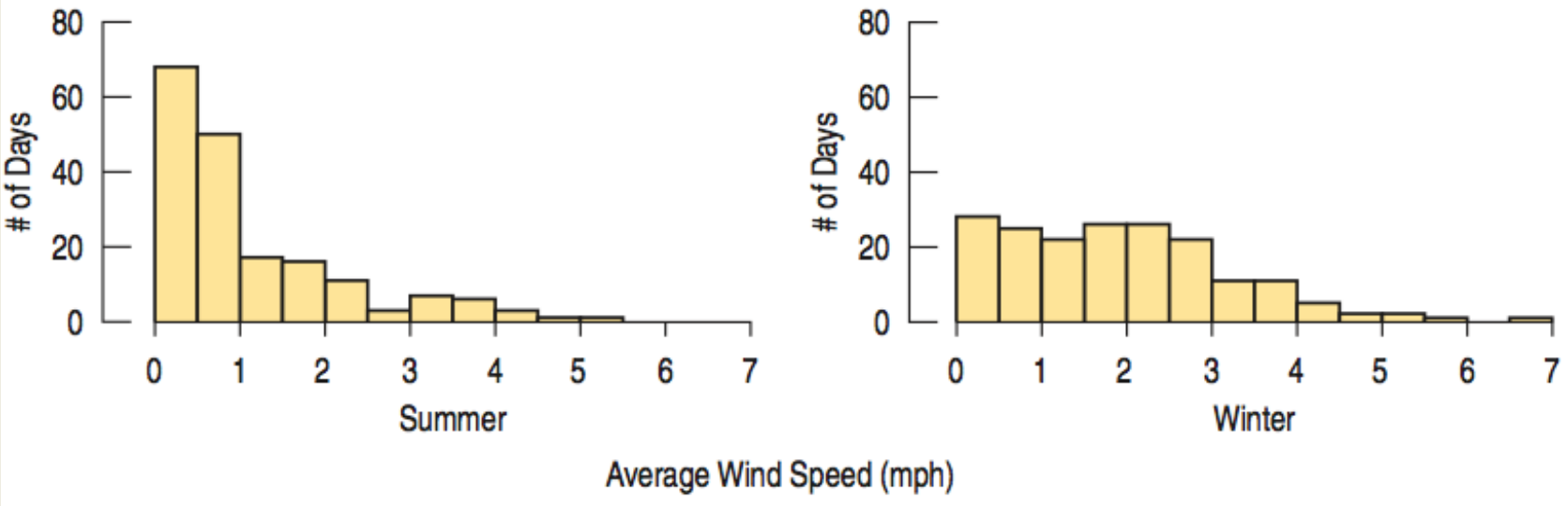
- How to display quantitative data:
  SPREAD as variance and standard deviation

  - The variance is a measure of how far the data is spread out from the *mean*.
  - The standard deviation is the square root of the variance
  - The sd expresses the average distance form the mean

$$s^2 = \frac{\sum (y - \bar{y})^2}{n-1} \qquad s = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$$

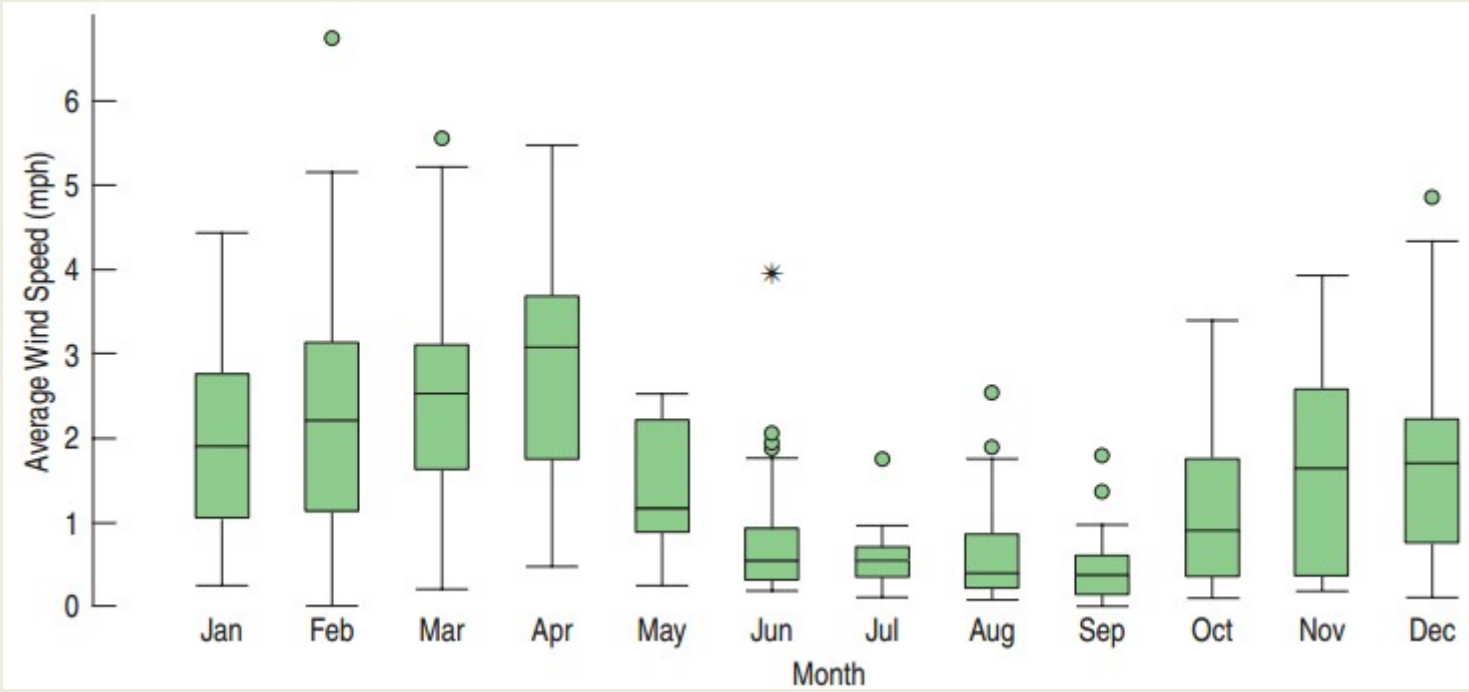
Review

- How to **compare** quantitative data: histograms and boxplots

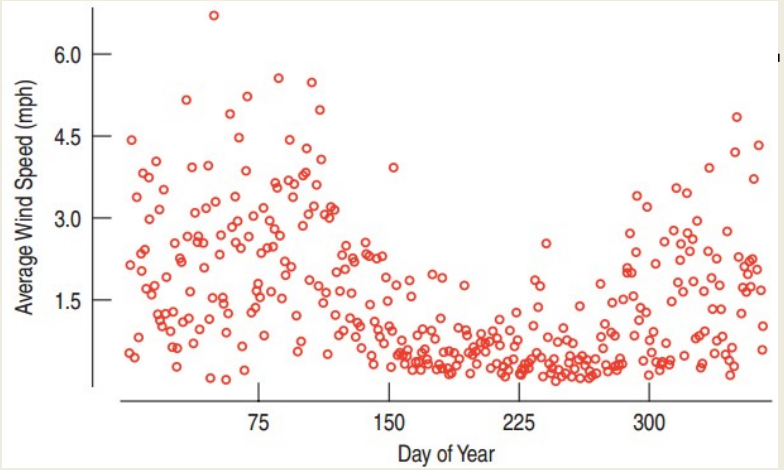


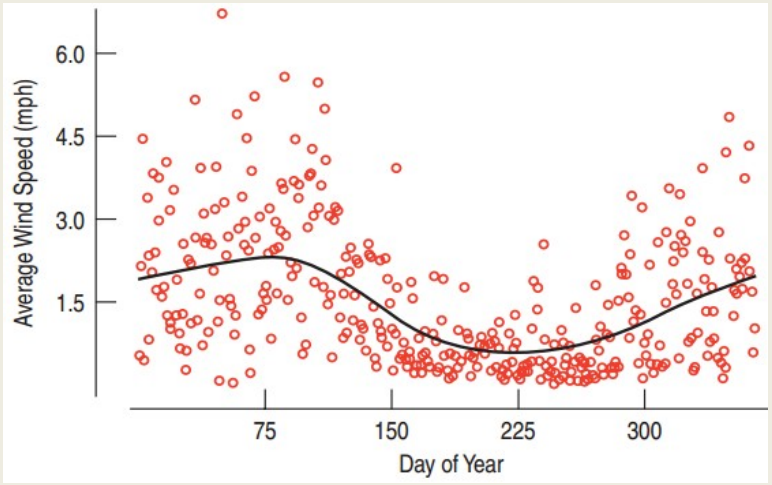
Review

- How to **compare** quantitative data: histograms and boxplots



Review: Timeplots





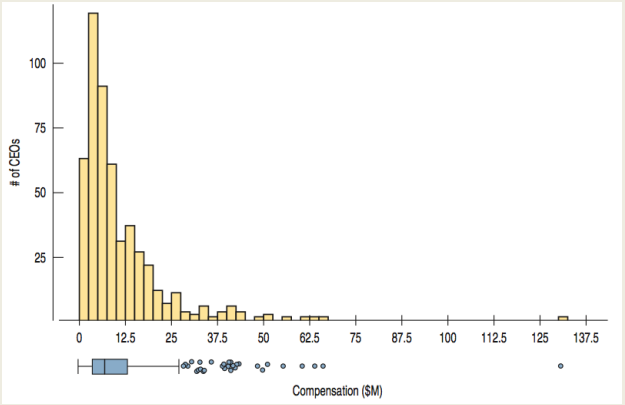
- Timeplots display every data value on a timeline.

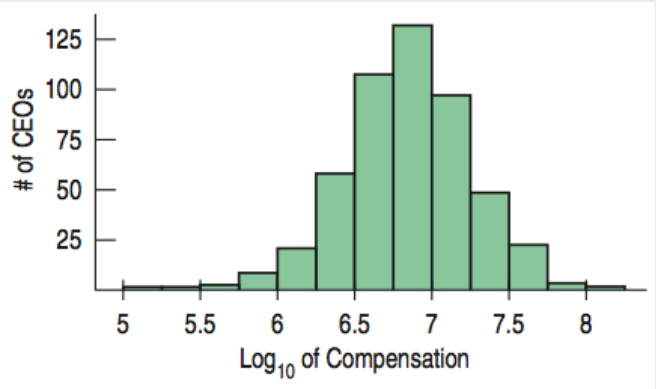
- Great for spotting trends - easier with a **lowess** curve

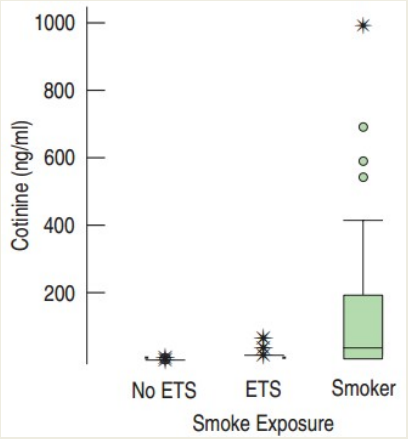
- We can look into the future…
- …but we can never predict a single event, an extraordinary event or the stock market (random)

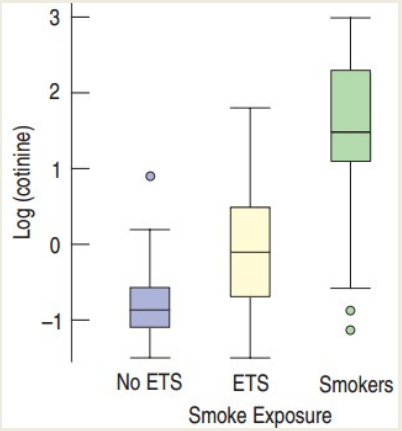
Review: Transformation of the Data

- Taking logarithm of the salaries makes histogram much easier to interpret.









Review: Comparing graphs

- Avoid inconsistent scales.
  - Don't try to compare one thing measured in feet to another measured in meters.

- Label Clearly.
  - Variables should be identified, and axes labeled.

Review: Outliers

- Check to see if there may have been an error in the data collection or data input.
- Check to see if there was an extraordinary outcome.

- Beware of Outliers!
  - If the outliers are errors, remove them.
  - Otherwise, considering presenting with and without the outliers.
  - Always comment on them

Chapter 5

The Standard Deviation as a Ruler and the Normal Model

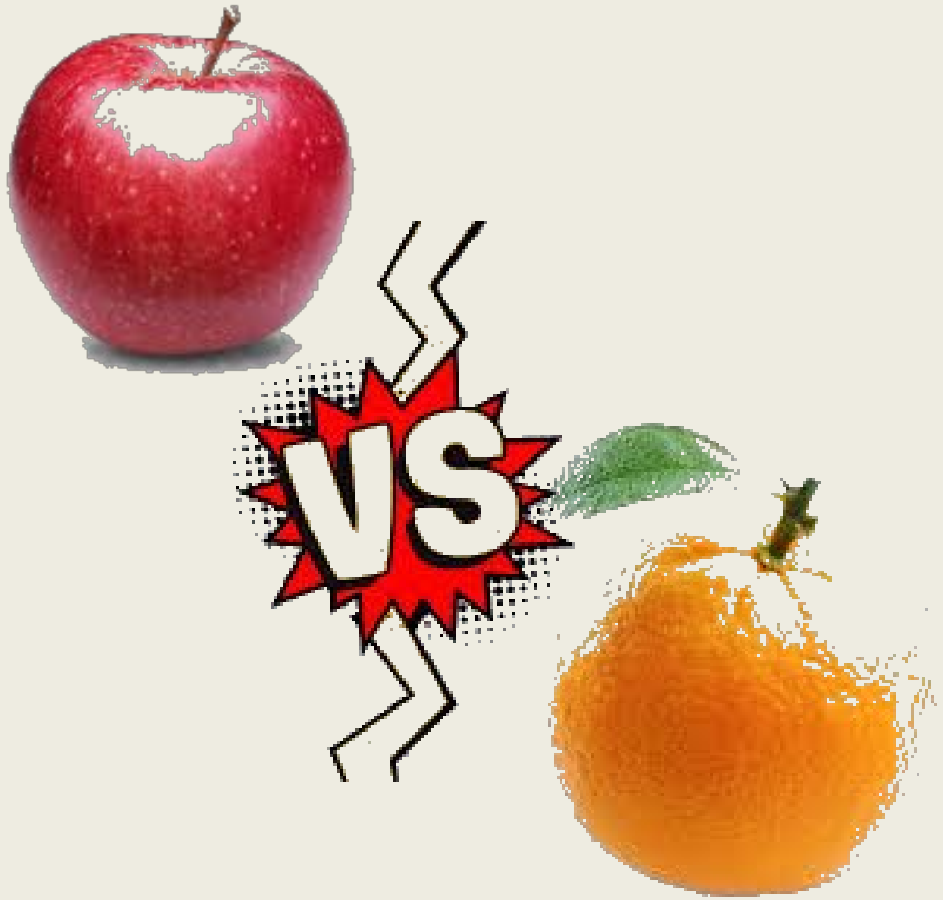
5.1

Standardizing with *z*-Scores

Apples and oranges

Apples and oranges

A comparison of apples and oranges occurs when two items or groups of items are compared that cannot be practically compared. The idiom, comparing apples and oranges, refers to the apparent differences between items which are popularly thought to be incomparable or incommensurable, such as apples and oranges.
Wikipedia



Comparing Athletes

- Chernova took the gold in the 2012 Olympics with a long jump of 6.545 m
- About 0.5 m farther than the mean distance.

- Jessica Ennis won the 200 m run with a time of 22.83 s
- More than 2 s faster than average.

- Whose performance was more impressive?

How Many Standard Deviations Above?

•The standard deviation helps us compare.

	Long Jump	200 m
Mean (all contestants)	5.91 m	24.48 s
SD	0.56 m	0.80 s
n	35	36
Chernova	6.54 m	23.67 s
Ennis	6.48 m	22.83 s

- Is there an even more precise way to calculate these?
- We can use a “z-score”

The z-Score

- In general, to find the distance between the value and the mean in standard deviations:

1. Subtract the mean from the value.
2. Divide by the standard deviation.

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

- This is called the *z*-score.

The z-score

- The *z*-score measures the distance of the value from the mean in standard deviations.

- Positive z-score?
- Negative z-score?
- Small z-score?
- Large z-score?

How Many Standard Deviations from Mean?

	Long Jump	200 m
Mean (all contestants)	5.91 m	24.48 s
SD	0.56 m	0.80 s
n	35	36
Chernova	6.54 m	23.67 s
Ennis	6.48 m	22.83 s

How Many Standard Deviations from Mean?

- Chernova's long jump

$$z = \frac{6.54 - 5.91}{0.56} \approx 1.1$$

- Ennis's 200 m run

$$z = \frac{22.83 - 24.48}{0.80} \approx -2.1$$

	Long Jump	200 m
Mean (all contestants)	5.91 m	24.48 s
SD	0.56 m	0.80 s
n	35	36
Chernova	6.54 m	23.67 s
Ennis	6.48 m	22.83 s

- Ennis's winning time is a little more impressive.
- Judges could assign points based on standard deviations from mean and this system would have a correlation of 0.99 with the one currently used!

How Many Standard Deviations from Mean?

- $-1 < z < 1$: Not uncommon
- $z = \pm 3$: Rare
- $z = 6$: Shouts out for attention!

5.2

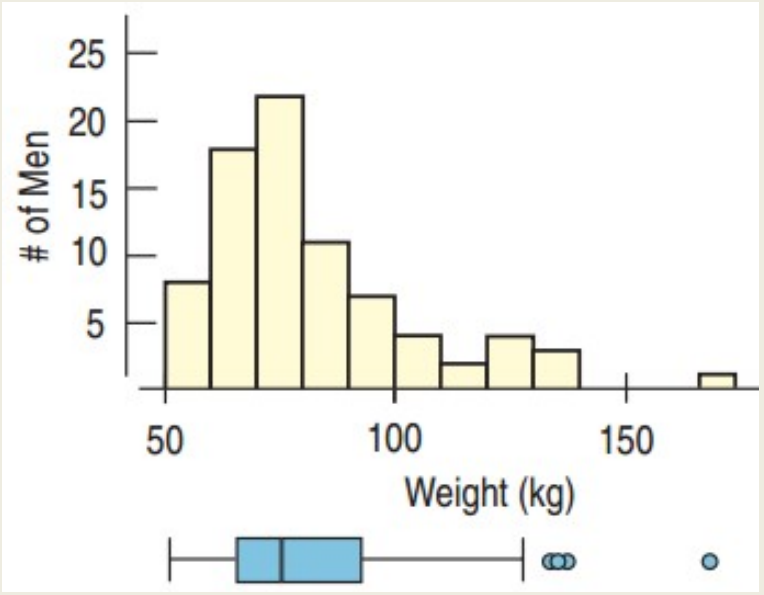
Shifting and scaling

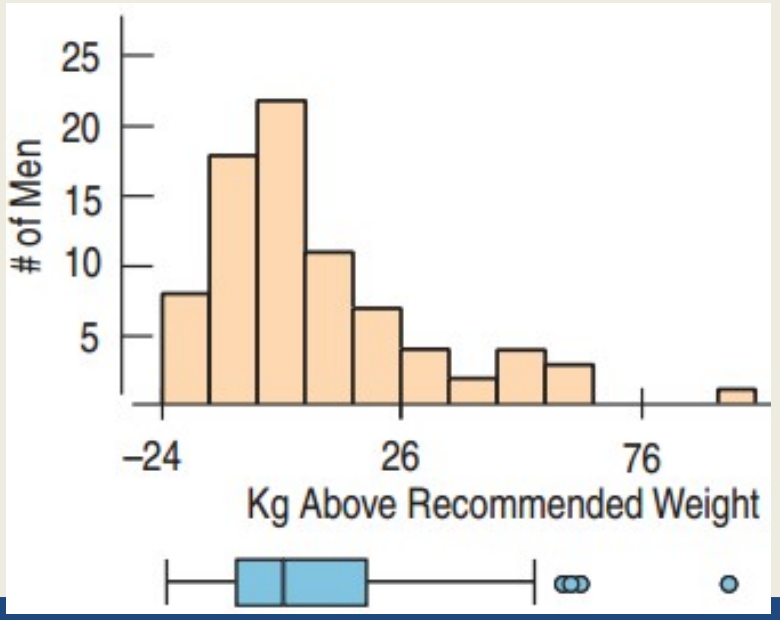
National Health and Examination Survey

- Who? 80 male participants between 19 and 24 who measured between 68 and 70 inches tall
- What? Their weights in kilograms
- When? 2001 – 2002
- Where? United States
- Why? To study nutrition and health issues and trends
- How? National survey

Shifting the Distribution

- Mean: 82.36 kg
- Maximum Healthy Weight: 74 kg
- How are shape, center, and spread affected when 74 is ***subtracted*** from all values?
  - Shape and spread are unaffected.
  - Center is shifted by 74.



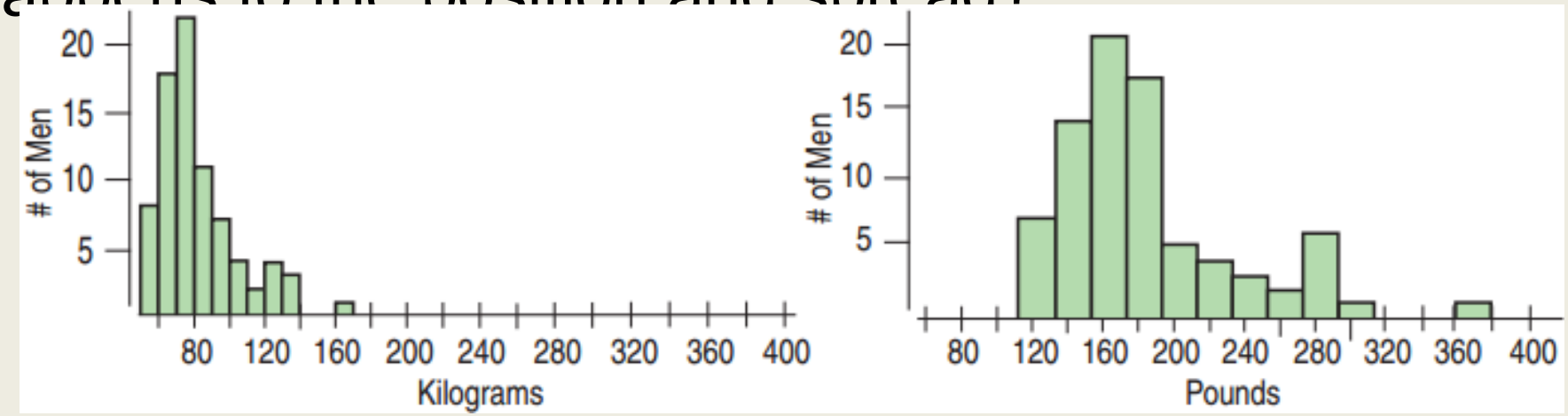


Rules for Shifting

- If the same number is subtracted or added to all data values, then:
  - The measures of the spread – standard deviation, range, and IQR – are all unaffected.
  - The measures of position – mean, median, and mode – are all changed by that number.

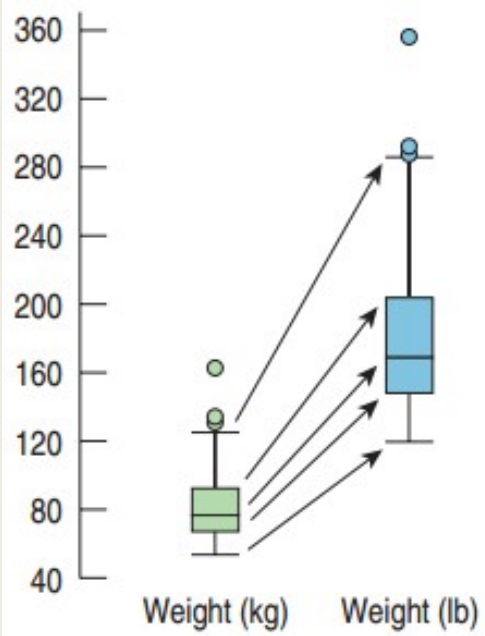
Rescaling

- If we multiply all data values by the same number, what happens to the position and spread?



- To go from kg to lbs, multiply by 2.2.
- The mean and spread are also multiplied by 2.2.
- And how about the shape?

How Rescaling Affects the Center and Spread



	Weight (kg)	Weight (lb)
Min	54.3	119.46
Q1	67.3	148.06
Median	76.85	169.07
Q3	92.3	203.06
Max	161.5	355.30
IQR	25	55
SD	22.27	48.99

- When we multiply (or divide) all the data values by a constant, all measures of position and all measures of spread are multiplied (or divided) by that same constant.

Example: Rescaling Combined Times in the Olympics

- The mean and standard deviation in the men's combined event at the Olympics were 168.93 seconds and 2.90 seconds, respectively.

- If the times are measured in minutes, what will be the new mean and standard deviation?
  - Mean: 168.93 / 60 = 2.816 minutes
  - Standard Deviation: 2.90 / 60 = 0.048 minute

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:

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

Conclusion

- To be in the top quarter of applicants for that specific college in terms of combined SAT scores, a college-bound senior would need to have an ACT score of at least 27.52.

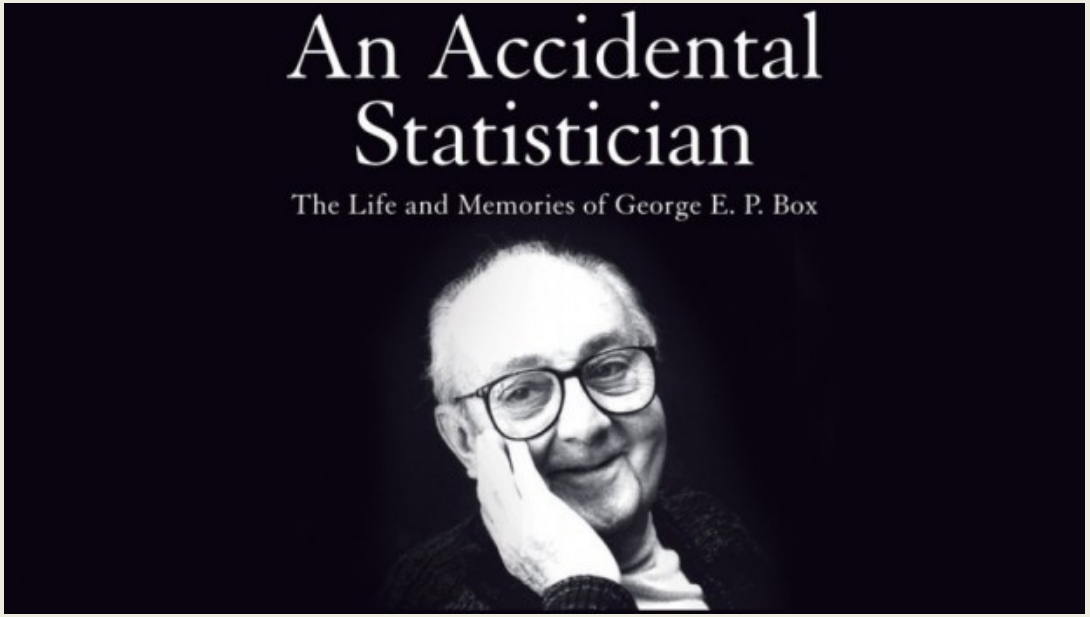
5.3

Normal Models

Models

- "All models are wrong, but some are useful."

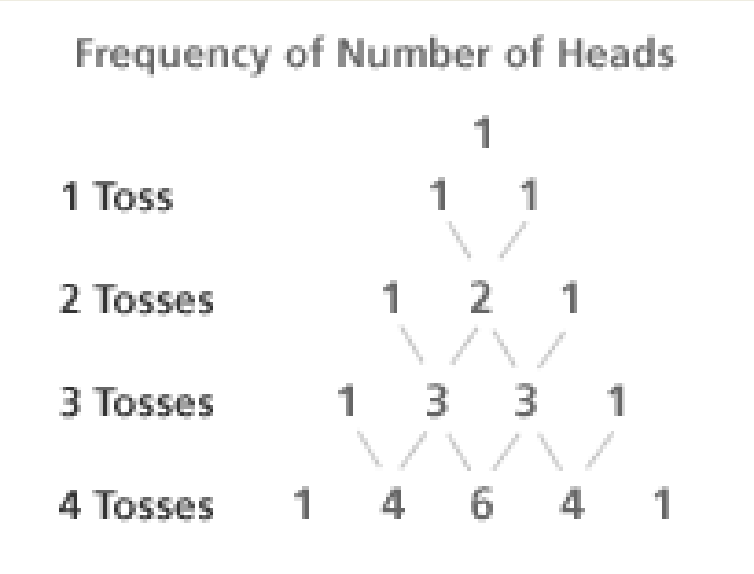
  George Box, statistician



Review

From the binomial expansion to the normal distribution

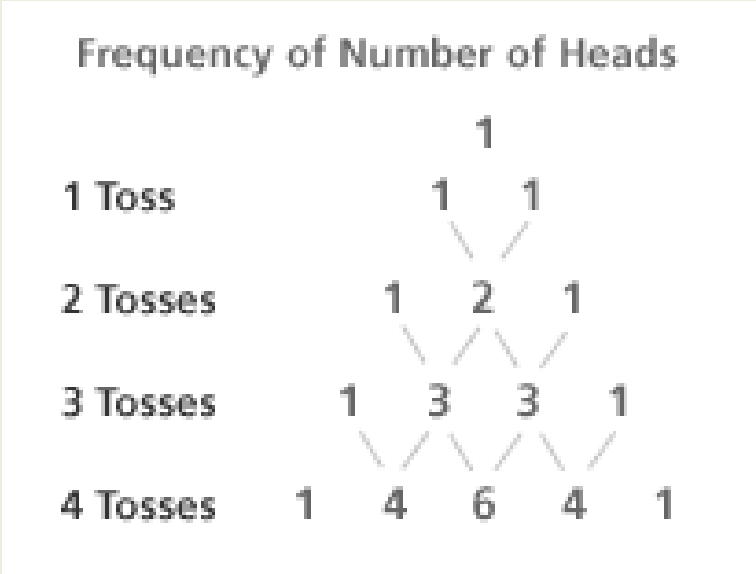
We can plot the frequency of getting heads on an histogram



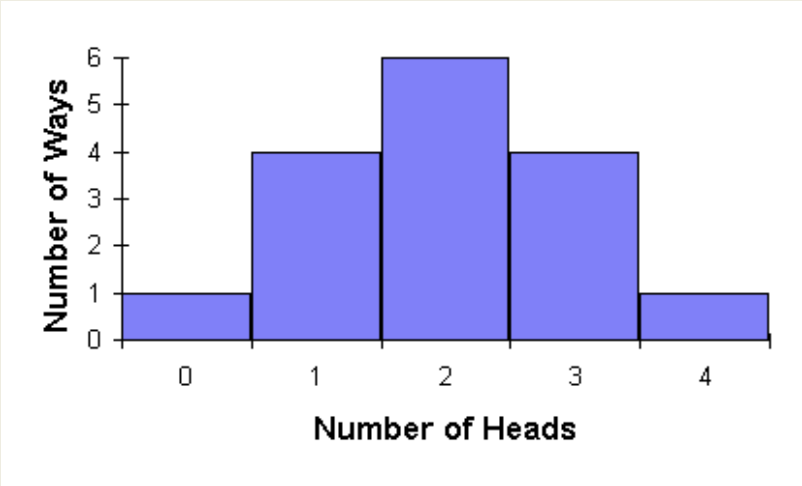
Review

From the binomial expansion to the normal distribution

We can plot the frequency of getting heads on an histogram



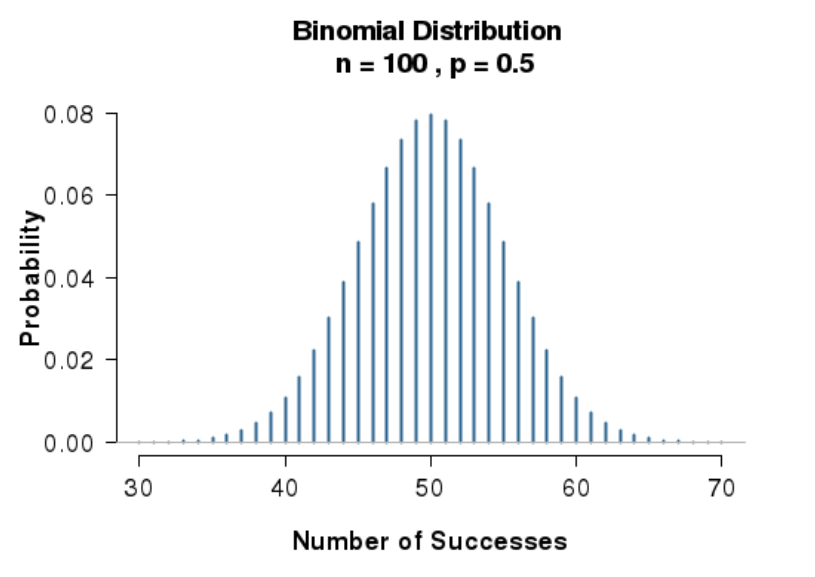
$$(\tfrac{1}{2} + \tfrac{1}{2})^4 = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$$



Review

From the binomial expansion to the normal distribution

The more coin tosses I make, the more this histogram will resemble a curve:



See simulation at:
https://shiny.rit.albany.edu/stat/binomial/

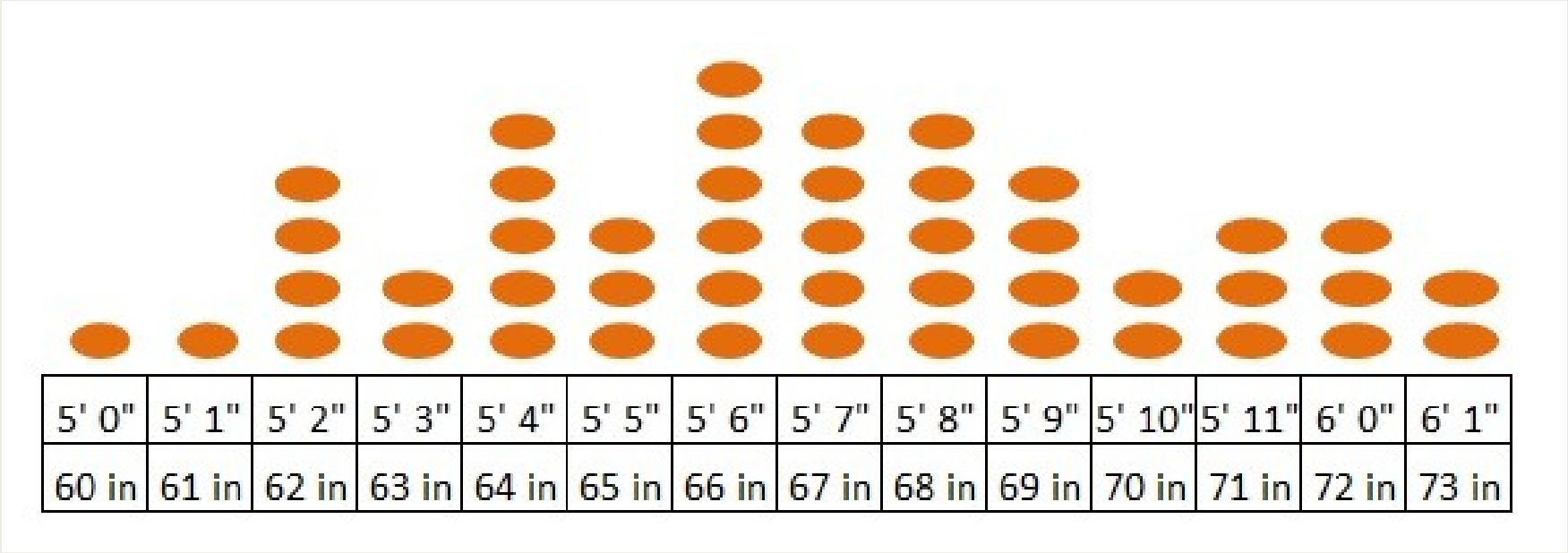
What follows a normal model?

We saw that the probability of getting heads in the course of a(n) (infinite) number of coin tosses approximates a normal model.

What other measures approximates a normal model?

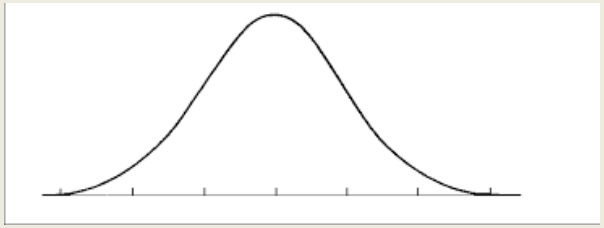
Students' heights!

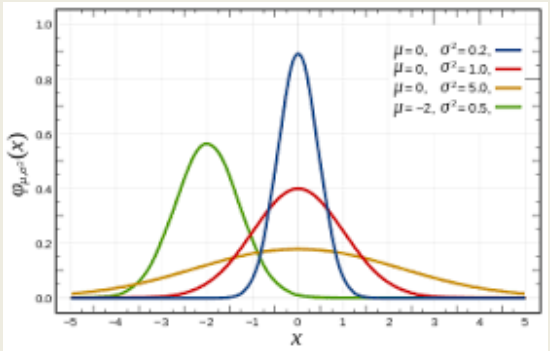
Students' heights!



5' 0"	5' 1"	5' 2"	5' 3"	5' 4"	5' 5"	5' 6"	5' 7"	5' 8"	5' 9"	5' 10"	5' 11"	6' 0"	6' 1"
60 in	61 in	62 in	63 in	64 in	65 in	66 in	67 in	68 in	69 in	70 in	71 in	72 in	73 in

The Normal Model





- Bell Shaped: unimodal, symmetric
- A Normal model for every value of mean and standard deviation.

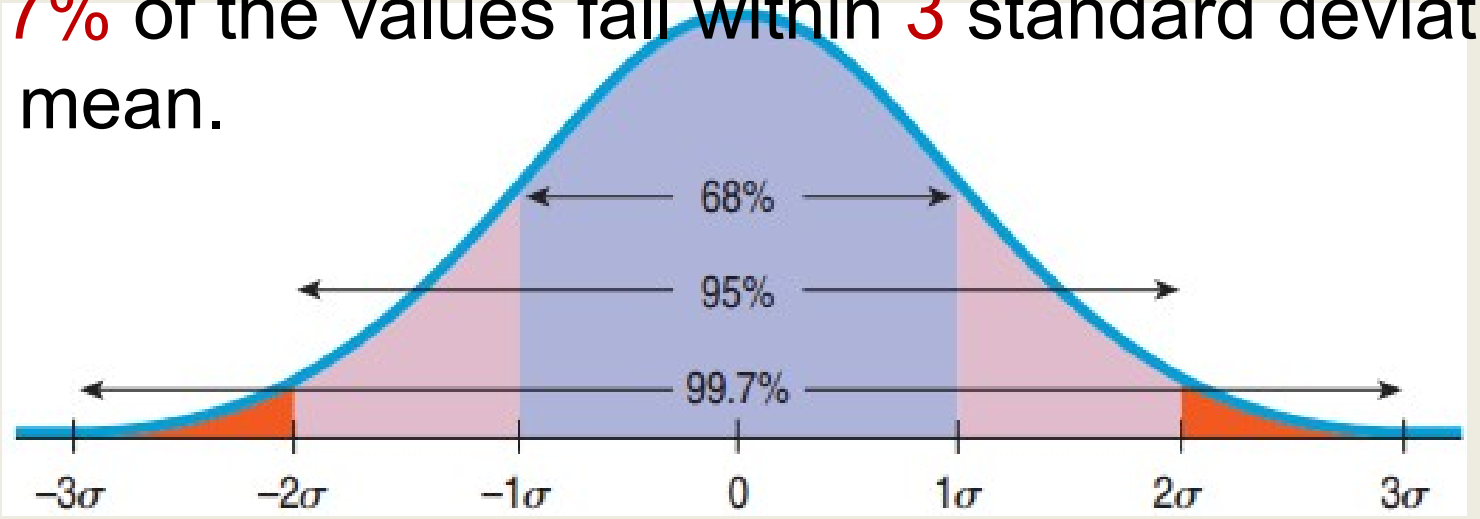
- μ (read "mew") represents the **population** mean.
- σ (read "sigma") represents the **population** standard deviation.
- $N(\mu, \sigma)$ represents a Normal model with mean μ and standard deviation σ.

Parameters and Statistics

- Parameters: Numbers that help specify the model
  - μ, σ
- Statistics: Numbers that summarize the data
  - $\bar{y}$, s, median, mode
- $N(0, 1)$ is called the standard Normal model, or the standard Normal distribution.
- The Normal model should only be used if the data is approximately symmetric and unimodal.

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.

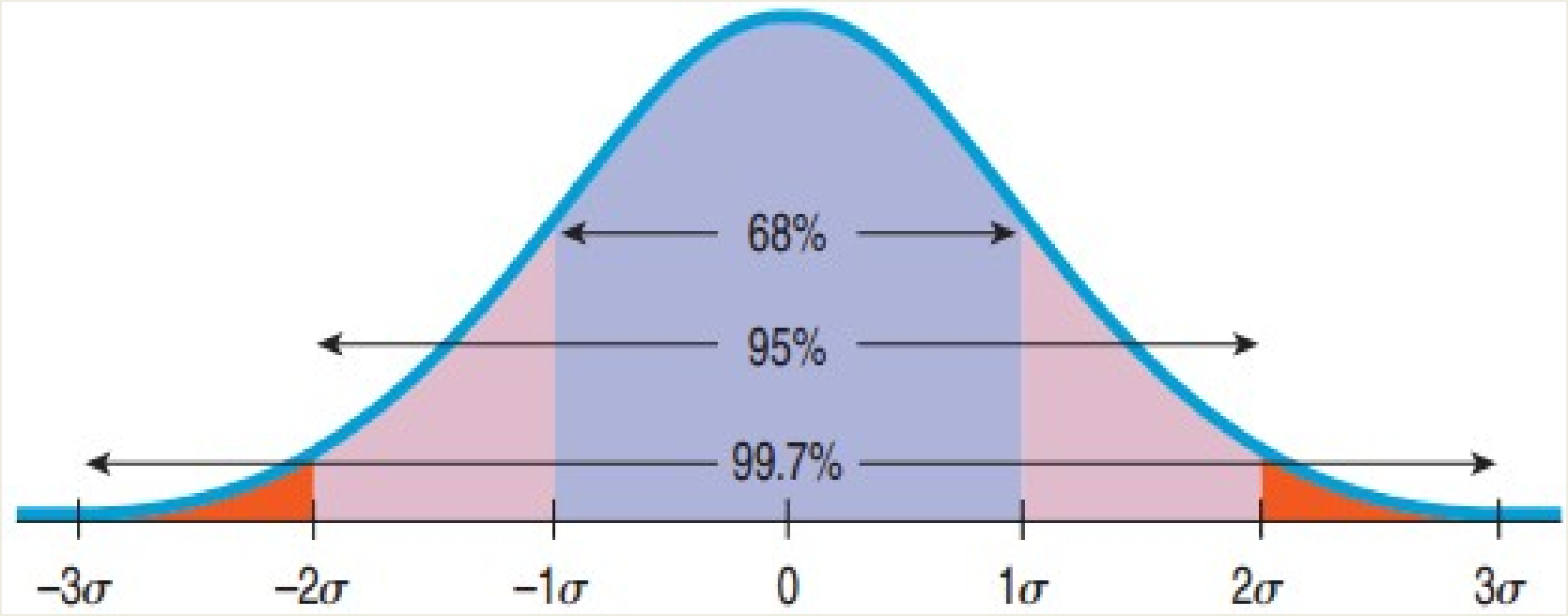


Probability is the area under the curve

- Remember the binomial expansion? Let's sketch it.

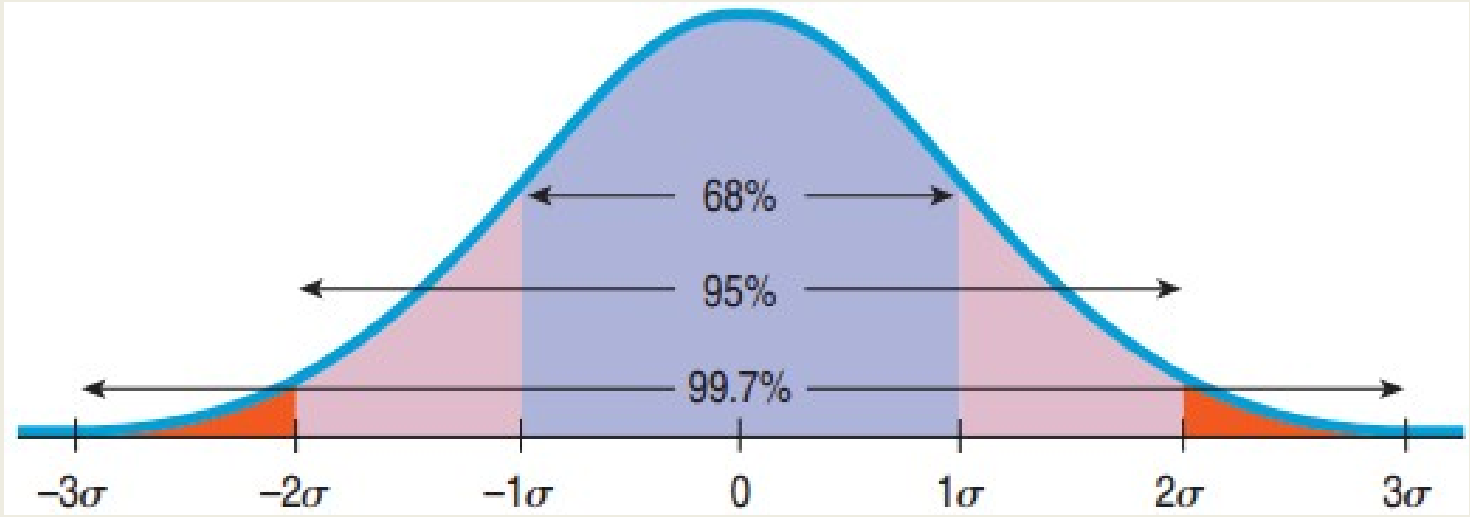
$$(\tfrac{1}{2} + \tfrac{1}{2})^4 = \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16}$$

Probability an an area under the curve



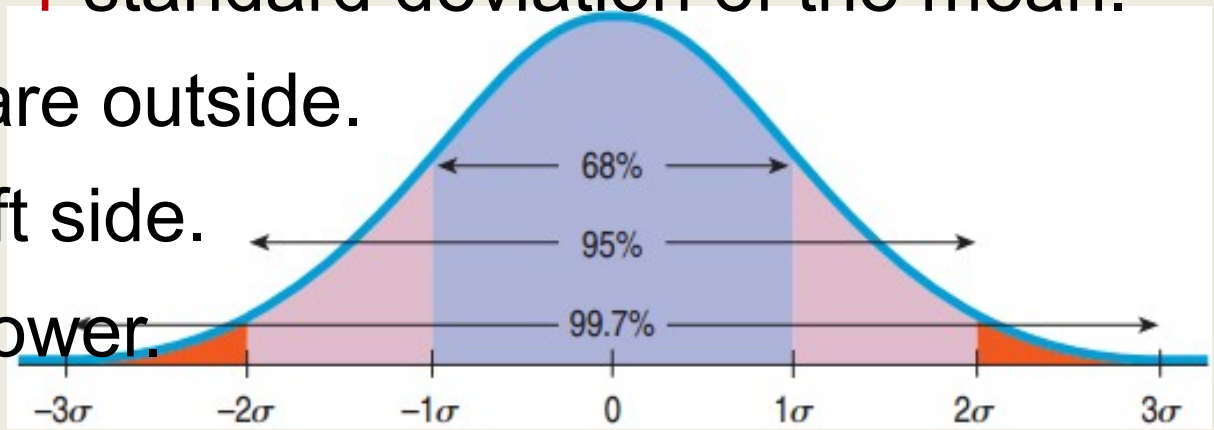
How Many Standard Deviations from Mean?

- $-1 < z < 1$: Not uncommon
- $z = \pm 3$: Rare
- $z = 6$: Shouts out for attention!



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 left side.
- 32% / 2 = 16% are slower.
- 16% of 48 is 7.7.
- About 7 are slower than Li Lei.



Three Rules For Using the Normal Model

- When data is provided, first make a histogram to make sure that the distribution is symmetric and unimodal.

- Then sketch the Normal model.

Working With the 68-95-99.7 Rule

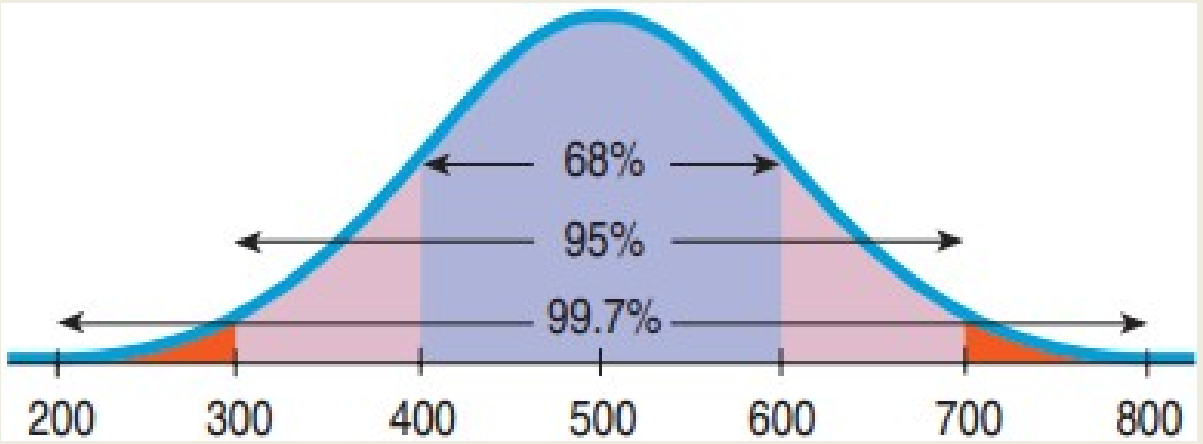
- Each part of the SAT has a mean of 500 and a standard deviation of 100. Assume the data is symmetric and unimodal.
- If you earned a 600 on one part of the SAT how do you stand (in percentage) among all others who took the SAT?

- **Plan:** The variable is quantitative and the distribution is symmetric and unimodal. Use the Normal model $N(500, 100)$.

Mechanics

- **Mechanics:**

Make a picture.



Conclusion:

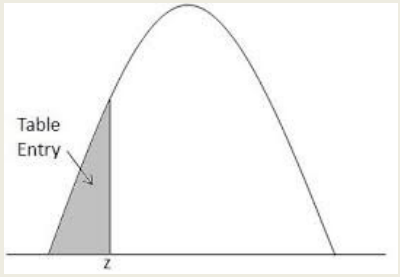
- 68% lies within 1 standard deviation of the mean.
- 100% - 68% = 32% are outside of 1 standard deviation of the mean.
- Above 1 standard deviations is half of that.
  - 32% / 2 = 16%
- Your score is higher than 84% of all scores on this test.

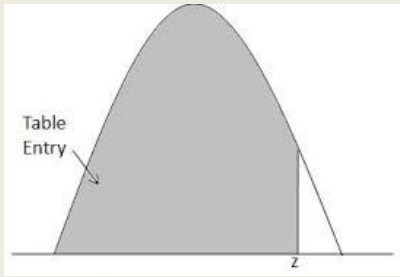
5.4

Finding Normal Percentiles

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





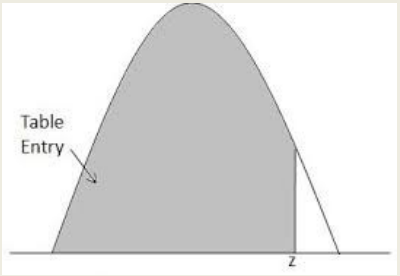
- **Example:** Where do you stand if your SAT math score was 680? $\mu = 500$, $\sigma = 100$
- Note that the z-score is not an integer:

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

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.

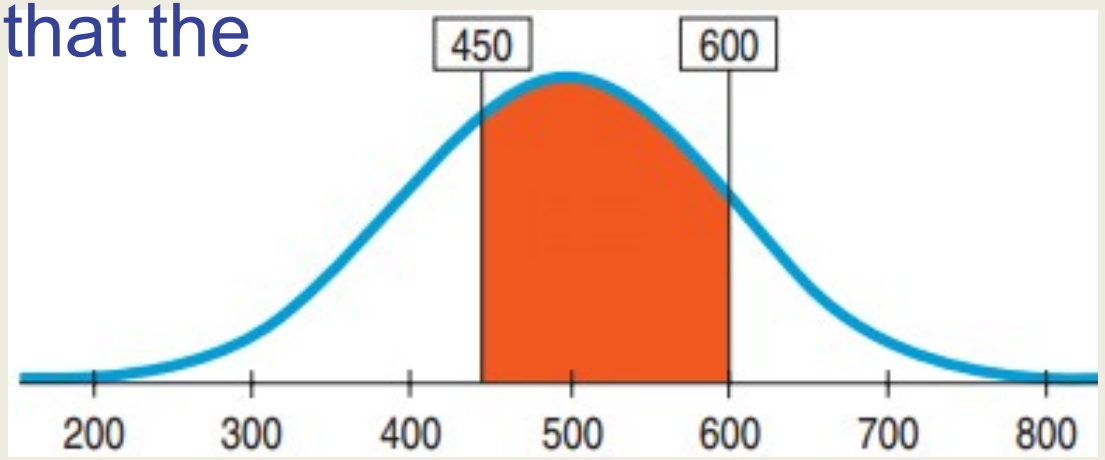
Look for z-scores on the sides of the table, get percentage in the middle of the table

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

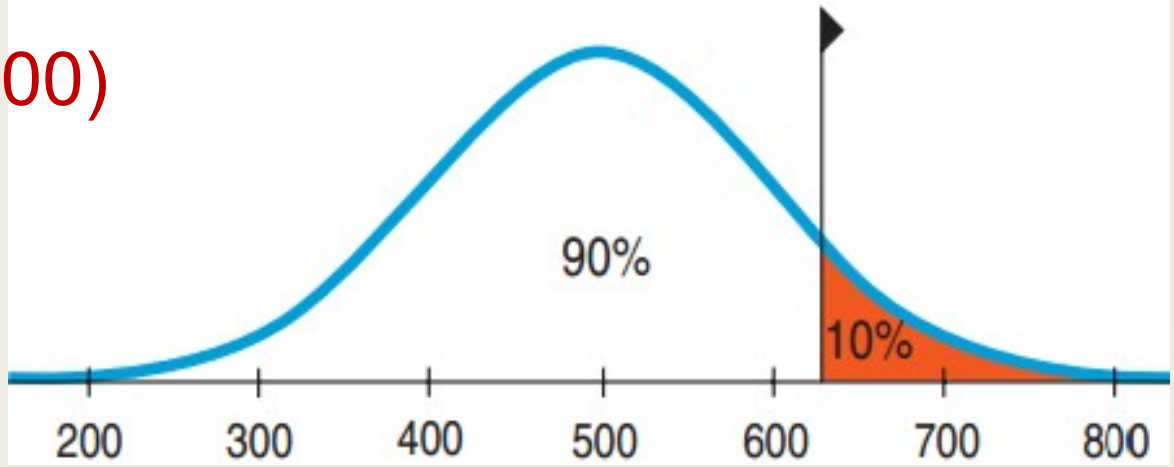
- Probability that x is between 450 and 600
  = Probability that x < 600 – Probability that x < 450
  Look for z-scores on the sides of the table, get percentage in the middle of the table
  = 0.8413 – 0.3085 = 0.5328
- Conclusion: The Normal model estimates that about 53.28% of SAT scores fall between 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? $\mu = 500$, $\sigma = 100$

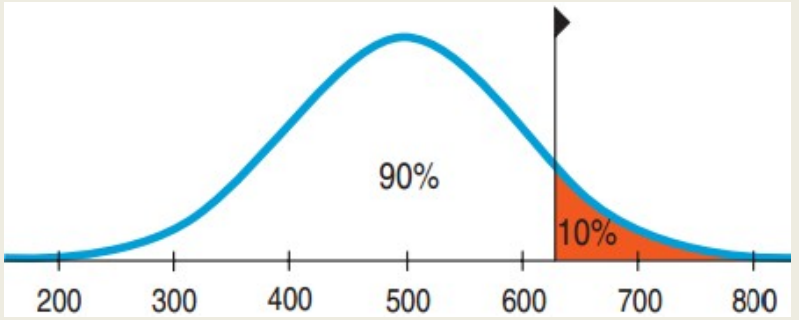
  - **Plan:** We are given the probability and want to go backwards to find *x*.

  - **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

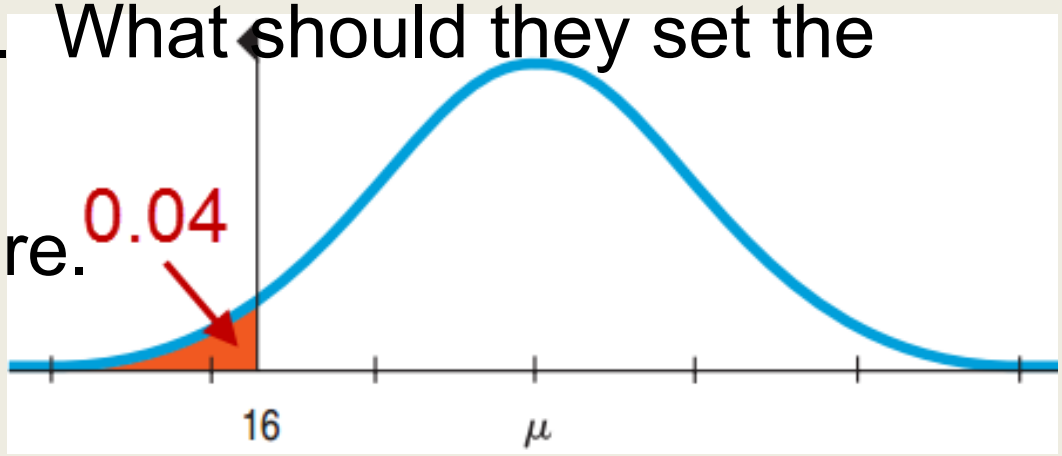
- What fraction of the cereal boxes will be underweight (less than 16.0)?
  $\mu = 16.3$, $\sigma = 0.2$

- $z = (16.0-16.3)/0.2 = -1.5$
- 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 $16 + 1.75(0.02)$
  $= 16.35$ 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.

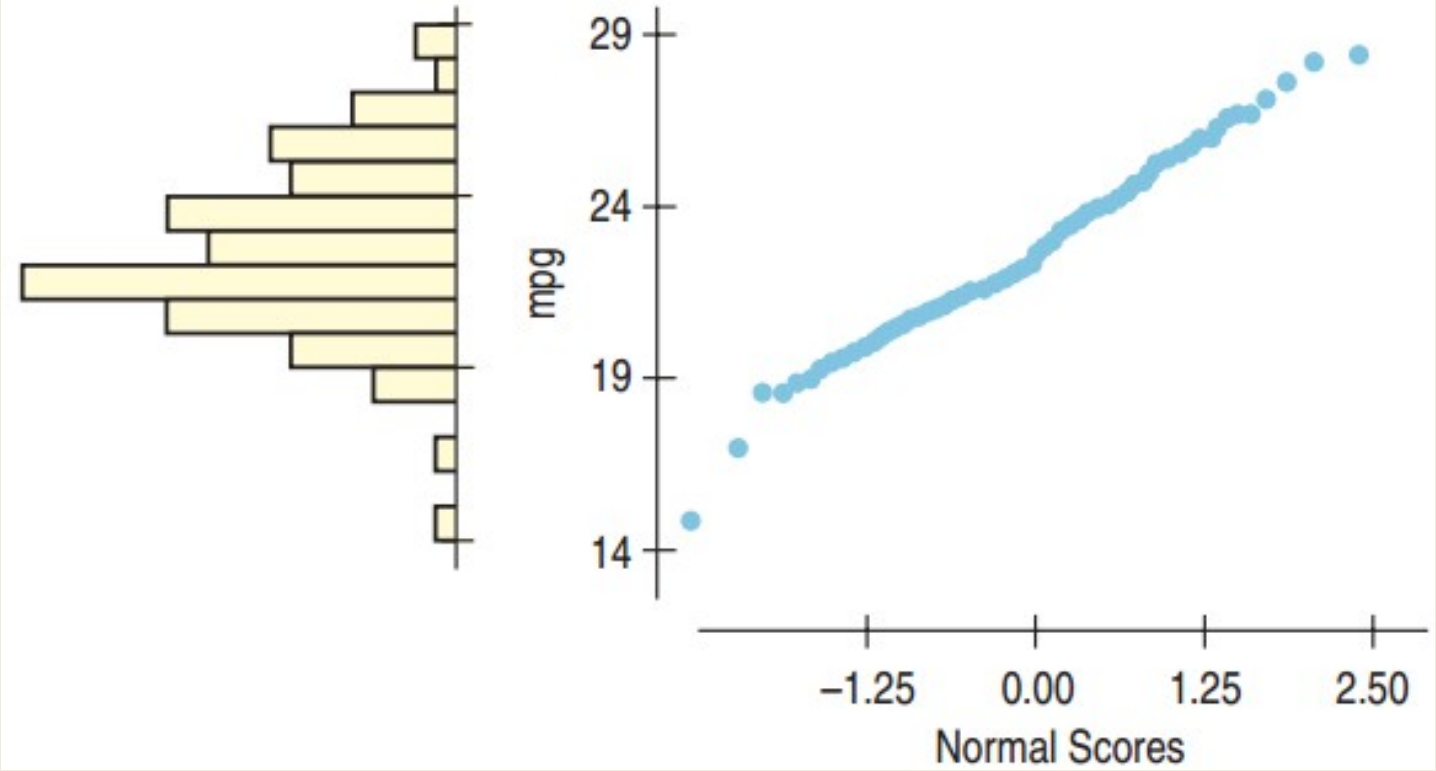
5.5

Normal Probability Plots

Checking if the Normal Model Applies

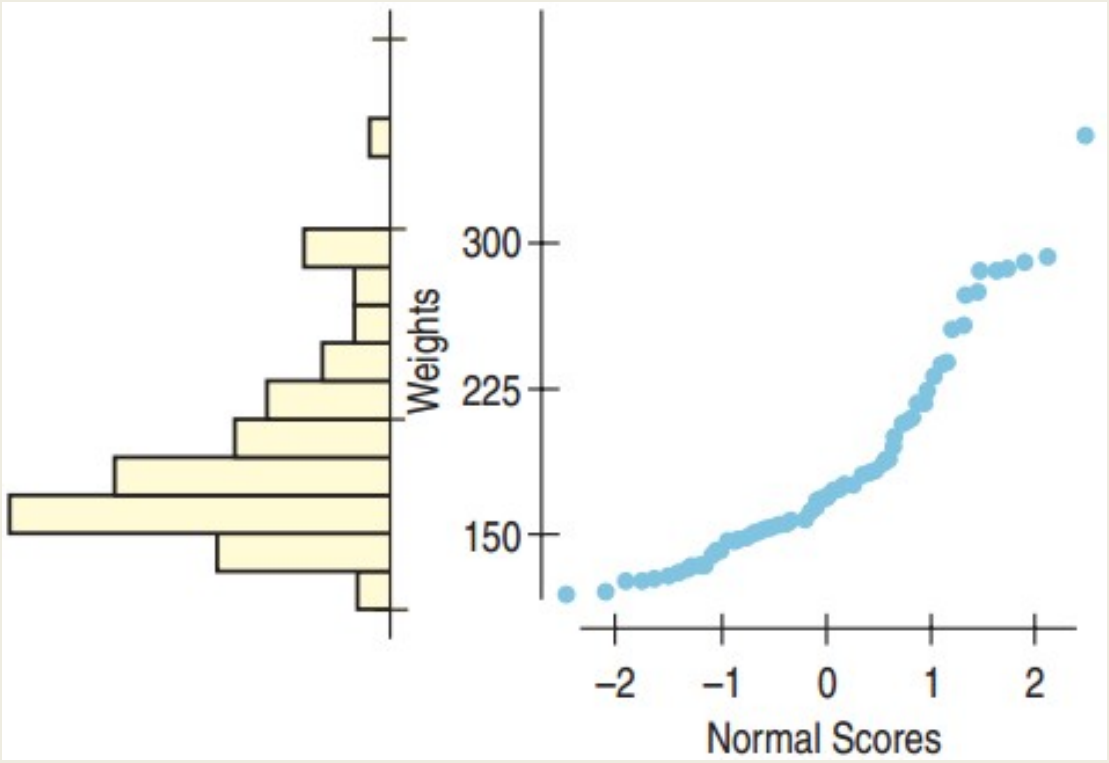
- A histogram will work, but there is an alternative method.
- Instead use a Normal Probability Plot.
  - Plots each value against the *z*-score that would be expected had the distribution been perfectly normal.
  - If the plot shows a line or is nearly straight, then the Normal model works.
  - If the 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

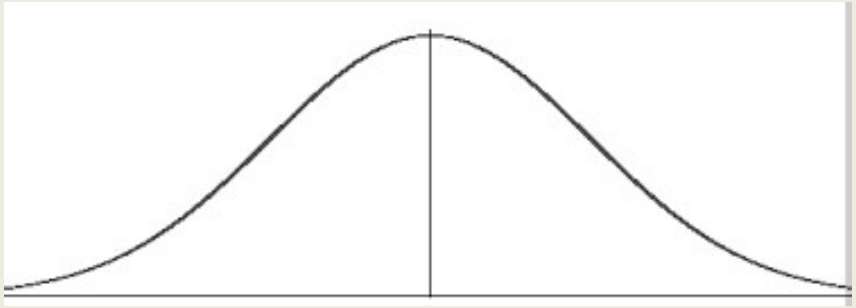
Scatterplots, Association, and Correlation

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?

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*”