Quantitative Methods

Serena DeStefani – 7/22/2020

Review II

Chapter 5

The Standard Deviation as a Ruler and the Normal Model

How Many Standard Deviations Above?

- The standard deviation helps us compare.
- Chernova's long jump was more than 1 standard deviation better than the 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

- Ennis's winning time in the 200 m was more than 2 standard deviations faster than than the mean.

Is there an even more precise way to calculate these?

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.
- A positive *z*-score indicates the value is above the mean.
- A negative *z*-score indicates the value is below the mean.
- A small *z*-score indicates the value is close to the mean when compared to the rest of the data values.
- A large *z*-score indicates the value is far from the mean when compared to the rest of the data values.

How Many SDs 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!

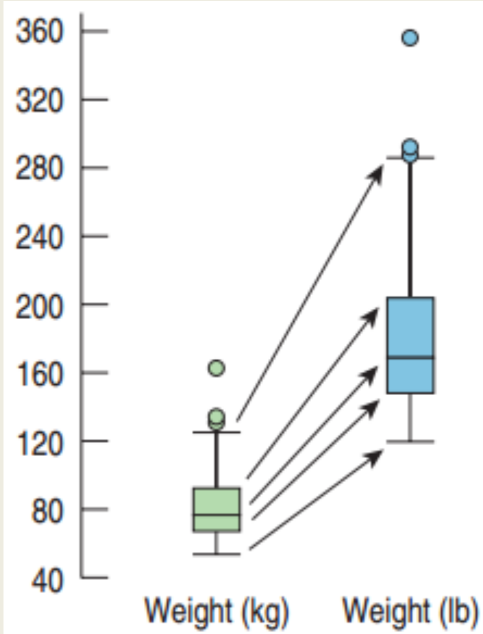
How Many SDs from Mean?

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

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



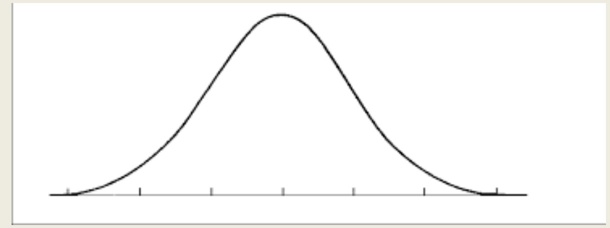
	Weight (kg)	Weight (lb)
Min	54.3	119.46
Q1	67.3	148.06
Median	76.85	169.07
Q3	92.3	203.06
Max	161.5	355.30
IQR	25	55
SD	22.27	48.99

- 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.

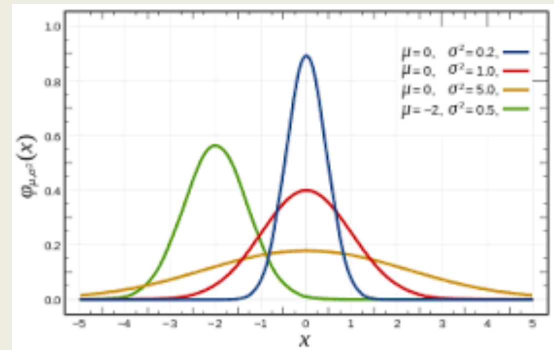
Can we apply a Normal Model to our data?

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

The Normal Model



- Bell Shaped: unimodal, symmetric



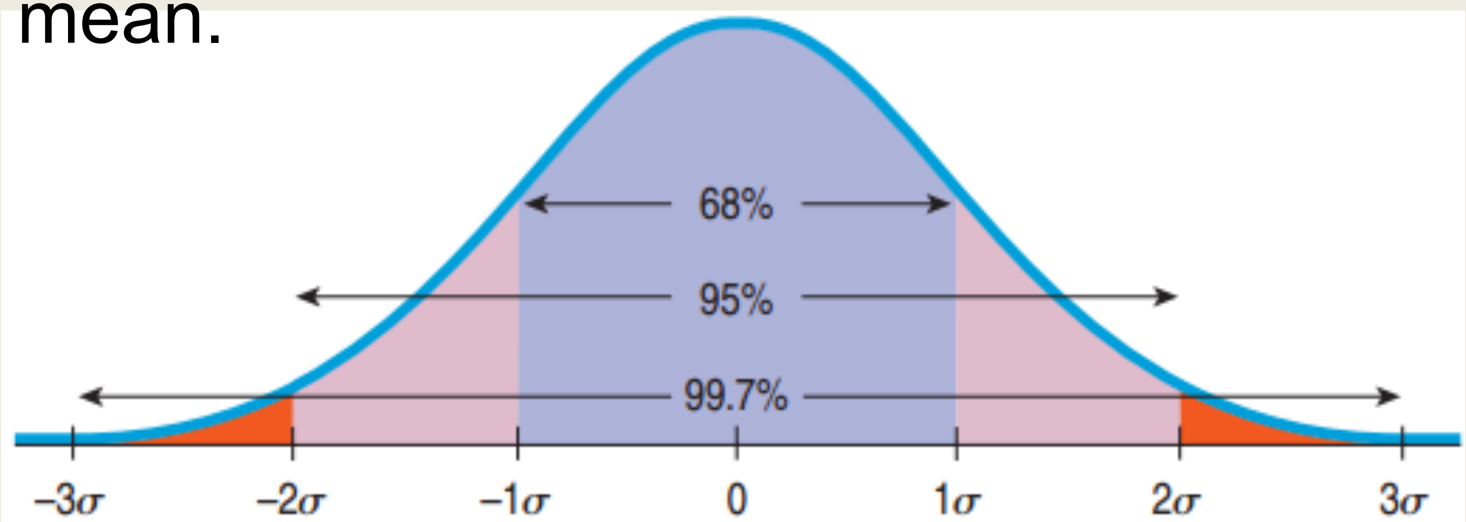
- A Normal model for every mean
- and standard deviation.

  - 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.

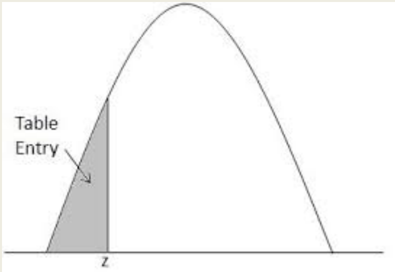
The 68-95-99.7 Rule

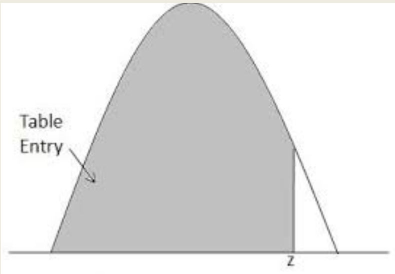
- 68% of the values fall within 1 standard deviation of the mean.
- 95% of the values fall within 2 standard deviations of the mean.
- 99.7% of the values fall within 3 standard deviations of the mean.



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: in this case, 0 (1.8=1.80)

Result: 0.9641

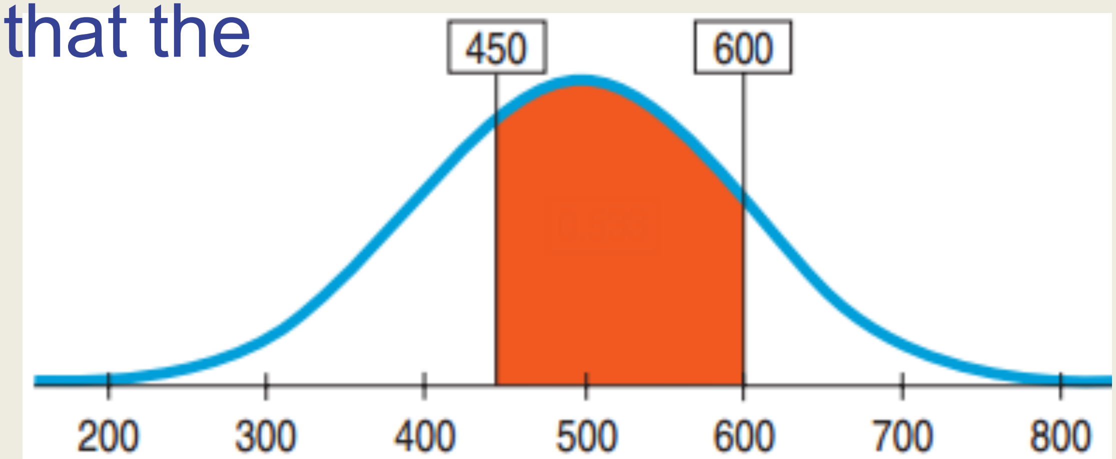
96.4% of SAT scores are below 680.

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.
  $N(500, 100)$

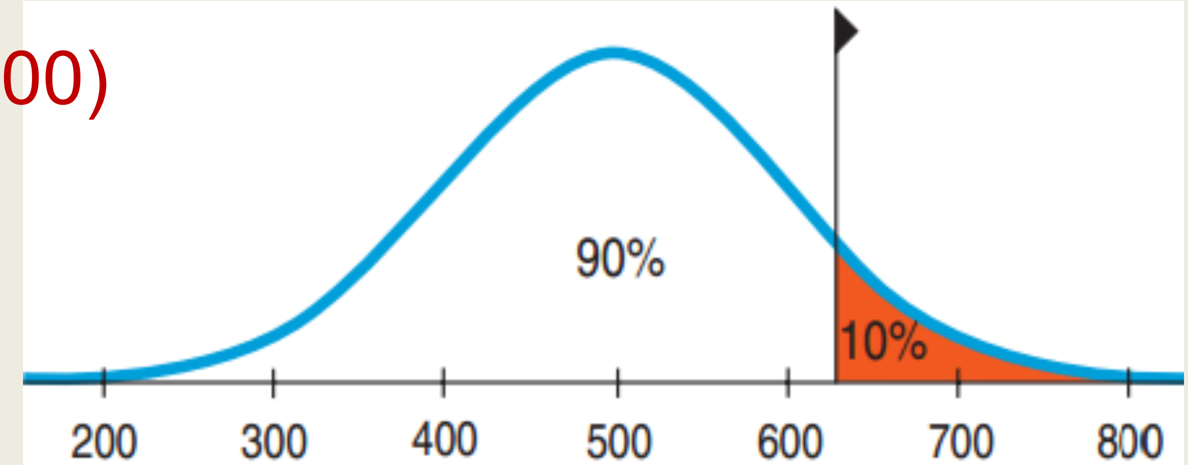


From Percentiles to Scores - *z* in Reverse

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



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 (less than x<16.0)?

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

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? $\sigma = 0.2$

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:** $N(16.2, ?)$

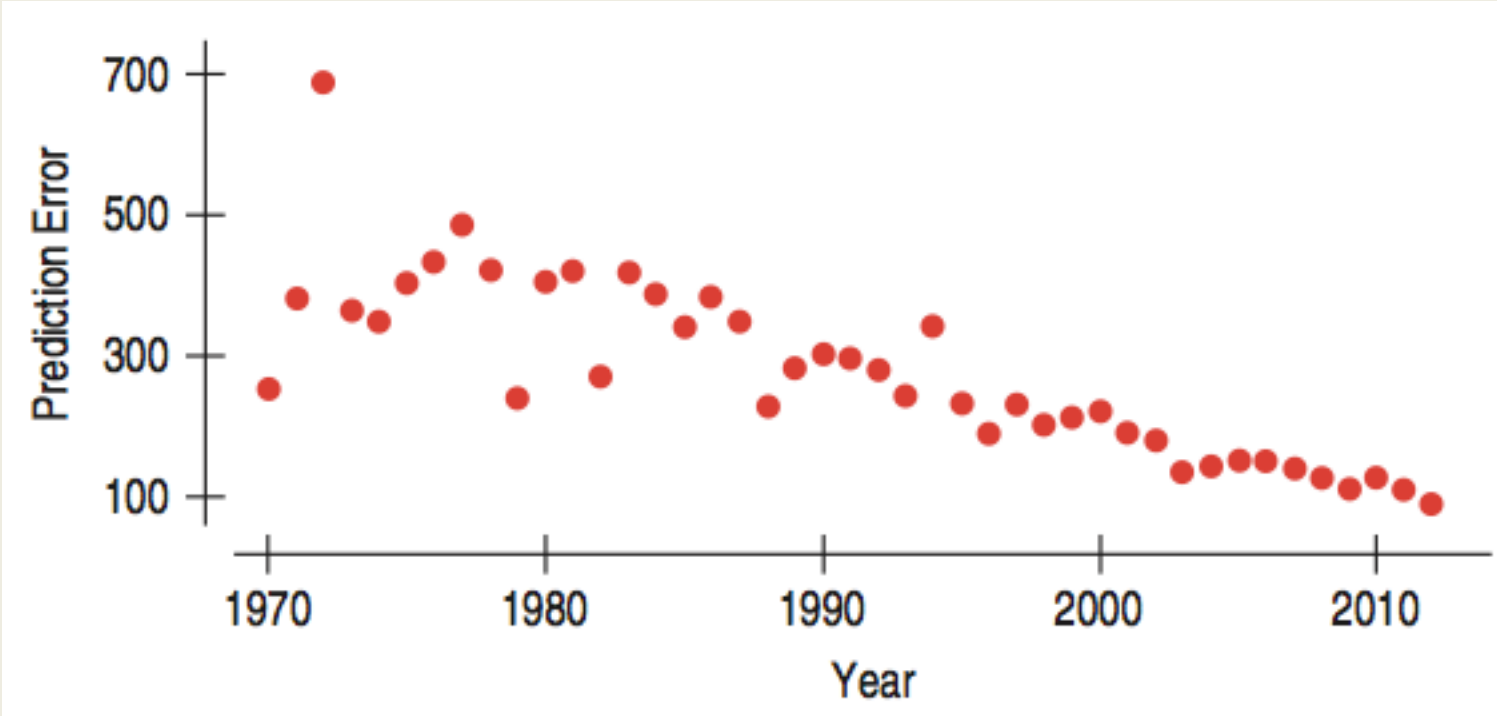
Underweight Cereal Boxes Part III

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

Chapter 6

Scatterplots, Association, and Correlation

Scatterplot of Hurricane Predictions



- Scatterplots exhibit the relationship between two variables.
- Used for detecting patterns, trends, relationships, and extraordinary values

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

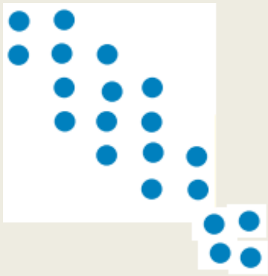
- Linear: The points cluster near a straight line.

- 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.

Roles of Variables

- Response Variable (y): The variable of interest. It is what we want to predict.
- Explanatory or Predictor Variable (x): The variable that we use to provide information or a prediction of the response variable.
- Choosing the response variable and the explanatory variable depends on how we think about the problem.

Properties of Correlation

- $r > 0 \rightarrow$ positive association
- $r < 0 \rightarrow$ negative association
- $-1 \le r \le 1$, 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.

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.

Correlation ≠ Causation

- 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.
  - Number of gray hairs and number of wrinkles are strongly correlated, but dyeing hair black does not undo wrinkles. Age is the lurking variable that causes both to increase.

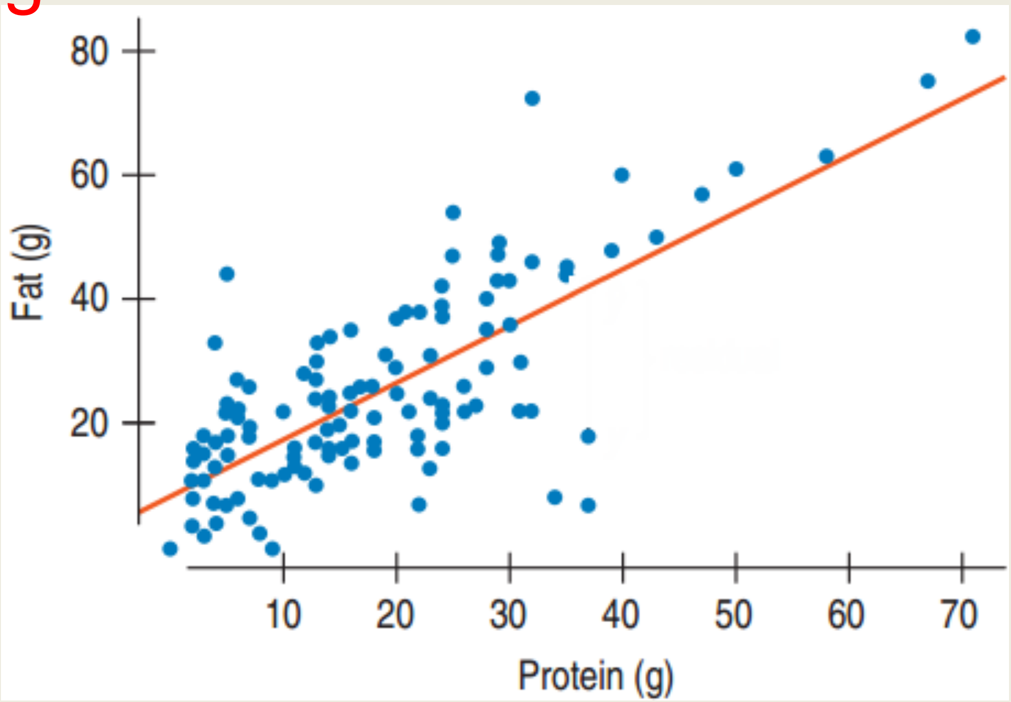
Chapter 7

Linear Regression

The Linear Model

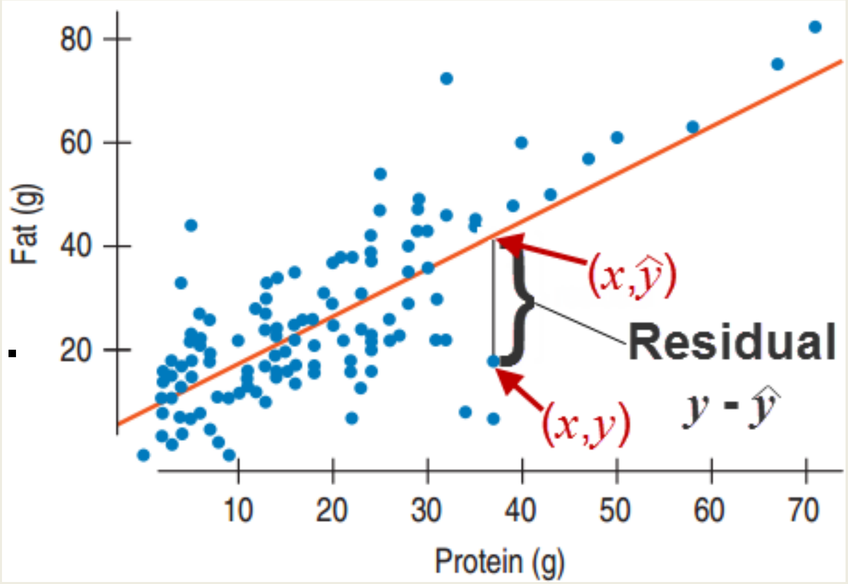
Fat and Protein at Burger King

- The correlation is 0.76.
- This indicates a strong linear fit, but how do we choose the line?
- The line should be "closest" to the points.



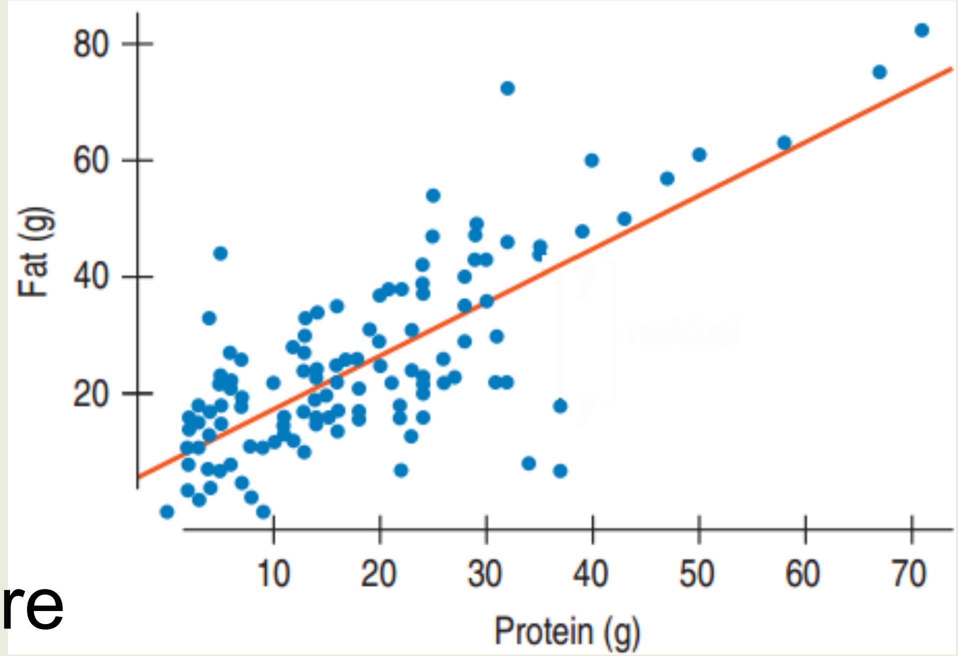
The Residual

- $\hat{y}$ is the value on the line
- It is called the **predicted value**.



- For each point (x,y) look at the point $(x,\hat{y})$ on the line with the same *x*-coordinate.
- The residual is defined by $y - \hat{y}$
- The residual is the difference between the observed value and the predicted value.

The Line of Best Fit



- The best fitting line will have small residuals.
- High negative residuals are just as “bad” as high positive residuals.
- Squaring all residuals makes them all positive.
- The line of best fit is the line for which the sum of the squares of the residuals is the smallest, also called the least squares line.

What's the equation of the Line of Best Fit?

Line from Algebra

- $y = mx + b$

Line of Best Fit

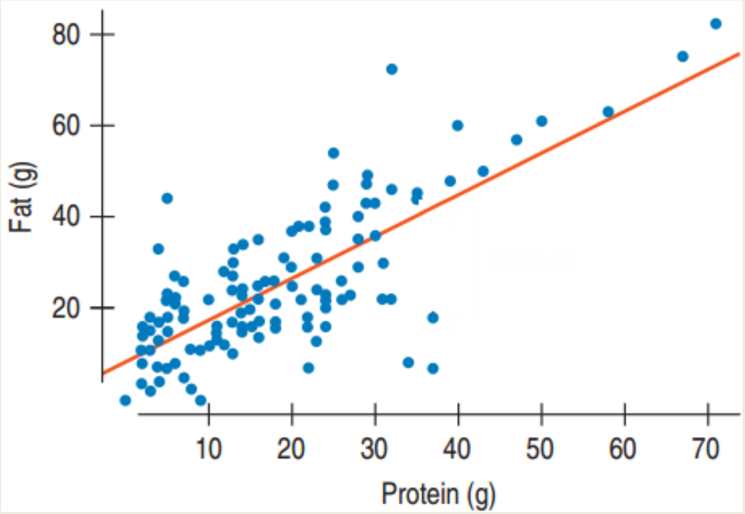
- $\hat{y} = b_0 + b_1 x$
- b_1 is the slope: how rapidly $\hat{y}$ changes with respect to x.
- b_0 is the y-intercept: The value of $\hat{y}$ when x is 0.

Interpreting the Line of Best Fit

Protein and Fat

- $\widehat{Fat} = 8.4 + 0.91\ Protein$
- What's the slope? What does it mean?
- Slope = 0.91: A Burger King item with one more gram of protein is expected to have 0.91 additional grams of fat.
- What's the y-intercept? What does it mean?
- *y*-intercept = 8.4: A Burger King item with no grams of protein is expected to have 8.4 grams of fat. In reality the two items with no protein also have no fat.

Slope and Correlation



Formula $$b_1 = r\frac{s_y}{s_x}$$

- Since the standard deviations are always positive, the slope and the correlation always have the same sign.

- The correlation has no units, but the slope has units of y per units of x.

  - For the Burger King example, the units for the slope are grams of fat per grams of protein.

The *y*-Intercept

$$\hat{y} = b_0 + b_1 x$$

The *y*-intercept and the slope are related by

$$\bar{y} = b_0 + b_1 \bar{x}$$

- The point corresponding to the means of *x* and *y*: $(\bar{x}, \bar{y})$ will always lie on the line of best fit.
- Given the mean of *x*, the mean *y*, and the slope, we can find the *y*-intercept: $b_0 = \bar{y} - b_1 \bar{x}$

Finding the Regression Equation

PROTEIN $\bar{x} = 18.0$ gr, $s_x = 13.5$ gr

FAT $\bar{y} = 24.8$ gr, $s_y = 16.2$ gr

$r = 0.76$

Conditions for Using Regression

The line of best fit is also called the least squares line or the regression line. Only use the regression line to make predictions if:

- The variable must be Quantitative.
- The relationship is Straight Enough.
- There should be no Outliers.

Finally, always check if the prediction is reasonable.

Residuals Revisited

- The residual is the difference between the y value of the data point and the $\hat{y}$ value found by plugging the x value into the least squares equation.

$$\text{Residual} = y - \hat{y}$$

- To find the residual:

1. Plug x into the least squares equation to get $\hat{y}$.
2. Subtract what you get from y to produce the residual.

Residual Example

$$\text{Residual} = y - \hat{y}$$

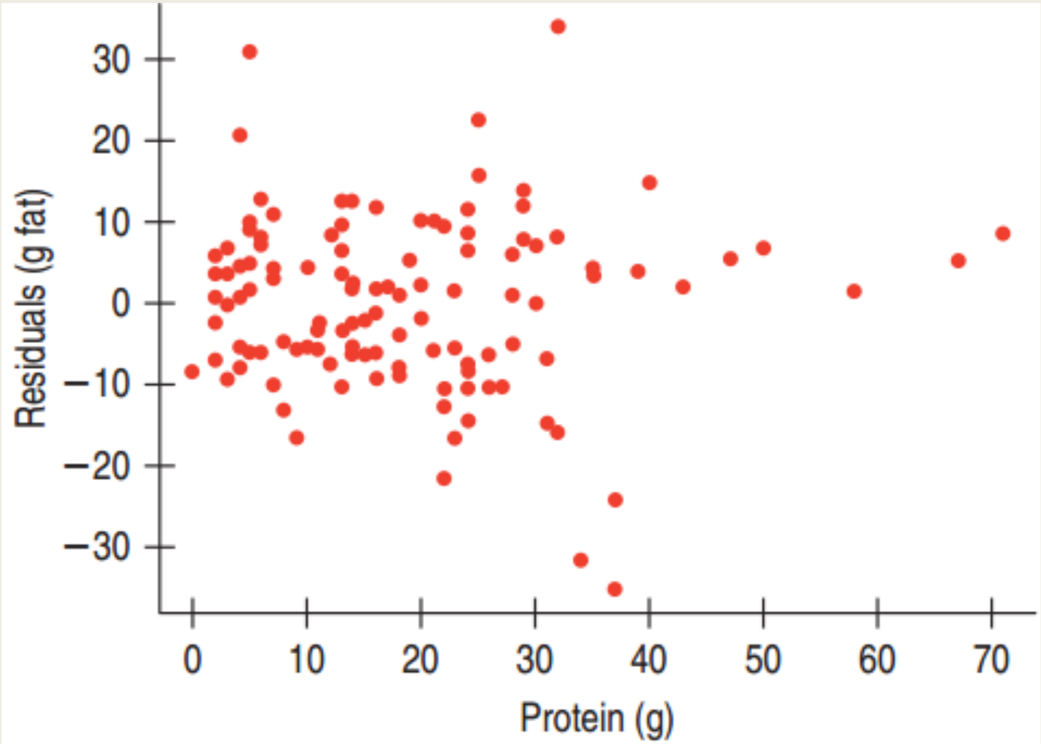
- That data that compared central pressure and maximum wind speed had $\hat{y} = 1024.464 - 0.968x$
- Hurricane Katrina's central pressure was x = 920 millibars and the maximum wind speed was y = 150 knots.
- Plugging in 920 gives

  $\hat{y}$ = 1024.464 – 0.968(920) = 133.90
- The residual is

  Residual = 150 – 133.90 = 16.1 kts

A Good Regression Model: Residual plot

- The regression model is a good model if the residual scatterplot against has no interesting features.
  - No direction
  - No shape
  - No bends
  - No outliers
  - No identifiable pattern



Comparing the Variation of *y* with the Variation of the Residuals

r = −1 or 1

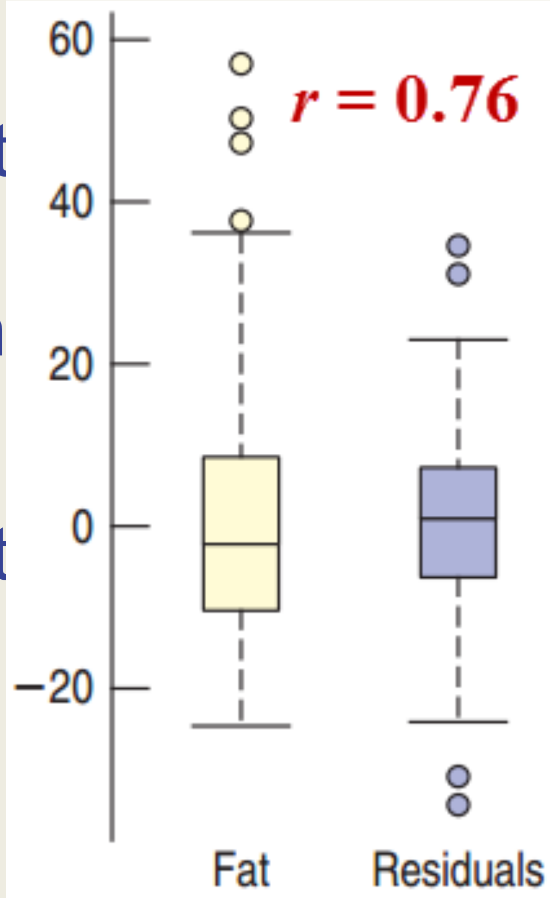
- **Perfect Correlation**: all points on a straight line
- The residuals are all 0. There is no variation of the residuals.

r = 0

- **No Correlation**: we always predict the same value (the mean)
- The regression line is horizontal through the mean.
- The residuals are the *y* values minus the mean.
- The variation of the residuals would be the same
- as the variation of the original *y* values.

Variation of *y* and the Variation of the Residuals

- $R^2 = 0.76^2 = 0.58$
  - 58% of the variability in fat content in Burger King's menu items is accounted for by the variation in protein content.
  - 42% of the variability in fat content is left in the residuals.
    - Other factors such as how the food is prepared account for this remaining variability.



When is R^2 Big Enough

- R^2 provides us with a measure of how useful the regression line is as a prediction tool.
- If R^2 is close to 1, then the regression line is useful.
- If R^2 is close to 0, then the regression line is not useful.
- What "close to" means depends on who is using it.
  - **Good Practice:** Always report R^2 and let the researcher decide.

Leverage and Influential Points

- A data point whose *x*-value is far from the mean of the rest of the *x*-values is said to have high leverage.
- Leverage points have the potential to pull strongly on the regression line.
- A point is influential if omitting it from the analysis changes the model enough to make a meaningful difference.
- Influence is determined by

1. The residual
2. The leverage

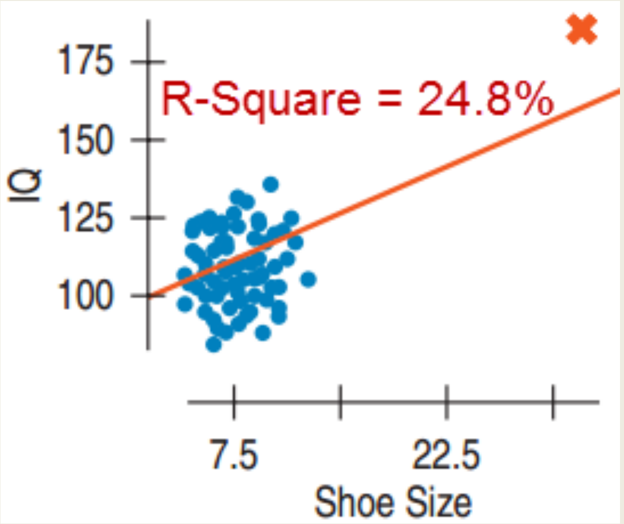
Shoe Size and IQ: Bozo the Genius Clown

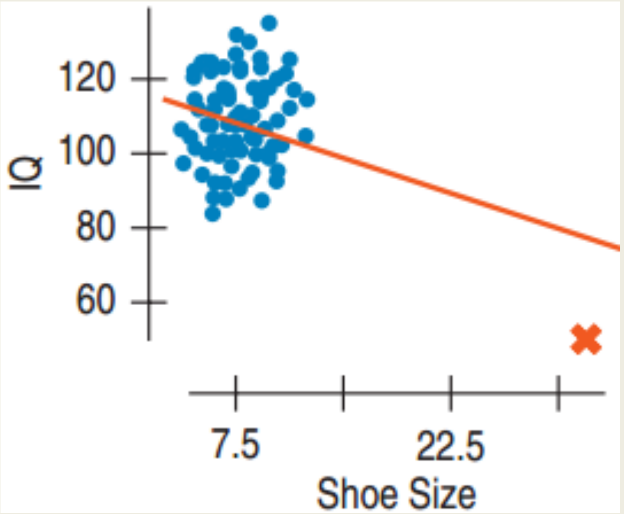
Model that Includes Bozo

- Almost all the variation accounted for by the model is from one point.
- After removing the outlier $R^2 = 0.7\%$
- Bozo is an **influential** point.

What if Bozo had an IQ of 50?

- The slope would go from 0.96 IQ point/shoe size to −0.69





Chapter 11

Sample Surveys

Idea 1: Examine a Part of the Whole

The Goal

- Learn about the entire group of individuals (called the population)

The Problem

- It is usually <u>impossible</u> to collect data on the entire population.

The Compromise

- Collect data on a smaller group of individuals (called a sample) selected from the population.

Idea 2: Randomize

Can we list the characteristics of the population and ensure we represent them all without bias?

- Race, age, ethnicity, income, marital status, work type, family size, …
- The list would go on forever. There are more types of people than the number of people.

Randomizing can lead to a representative sample.

- Randomizing protects us from the influences of all the features of the population.
- On average, the sample will look like the population.

Idea 3: It's the Sample Size

If you need 100 students to get a random sample at the university, how many Americans would you need to achieve the same level of randomness from the entire U.S.A.?

- Answer: 100
- It is the number of individuals, not the percent of individuals that matters.
- The number of individuals in the sample is called the sample size.

Representative Sampling

Since we can't take a true census, we want to compute statistics that reflect the parameters.

- A sample that does the above is called a representative sample.
- Biased samples tend to not be representative.
  - The statistic tends to be much higher or much lower than the parameter.

Random But Not Representative

Random

- Suppose there are 100 men and 100 women in a class. Flip a coin.
  - **Heads:** Choose the 100 men.
  - **Tails:** Choose the 100 women.

- Every student has an equally likely chance of being chosen. Randomness was achieved.

- This will **not** produce a **representative sample**.

Simple Random Sampling

SRS

- Order the students from 1 to 200.
- Use a computer to randomly select 20 numbers from 1 to 200.
- Select the students with the chosen numbers.

Simple Random Sampling (SRS) is when <u>every combination</u> has an equally likely chance to be selected.

- SRS is the standard which all other sampling techniques are measured.
- Statistical theory is based on SRS.

Sampling Frame and Sampling Variability

The **sampling frame** is the list of all individuals from which the sample is drawn.

The sample to sample differences are called the **sampling variability** (or sampling error).

Mistakes

- Voluntary Response Sample
- Convenience Sampling
- Use a bad sampling frame
- Undercoverage
- Non response bias
- Response bias

Chapter 12

Experiments and Observational Studies

Observational Studies

Observational Studies

- Researchers don’t assign choices.
- Passively observe participants
- Good for discovering relationships related to rare outcomes
- Bad for establishing cause-and-effect relationships
- Tough to handle lurking variables

How Experiments Work

- Identify the explanatory variable(s), called the factor(s).
- Identify the response variable.
- Select subjects or participants (if human) or experimental units (if not human).
- Decide on the levels to choose for each factor.
  - Music program or no music program
  - Sleep hours: 4, 6, or 8
- The combination of specific levels from all factors that a subject receives is called its treatment.

Assigning Participants to Treatments

- Don’t let them choose.
- Don’t assign based on what’s best for each.
- **Randomly** assign participants into groups. Each group receives a different treatment.
- Only through random assignment can a cause-and-effect relationship be established.
- What ethical dilemmas might this introduce?

Control and Randomization

1. Control
   - Make all conditions as similar as possible for all treatment groups.
   - Control allows us to isolate the one thing that is being studied. Helps avoid lurking variables
2. Randomize
   - Equalizes the effects of variation that we cannot control
   - Distributes the uncontrollable factors equally

Control what you can, randomize the rest.

Replicate and Block

Replicate

- Apply each treatment to a number of subjects.
- Repeat the entire experiment on an entirely different population of experimental units.

Block

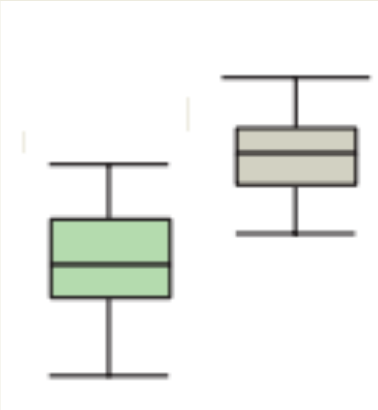
- Group similar individuals together and randomize within each of these blocks.
- Blocking helps account for the variability due to the difference between blocks.

Statistical Significance

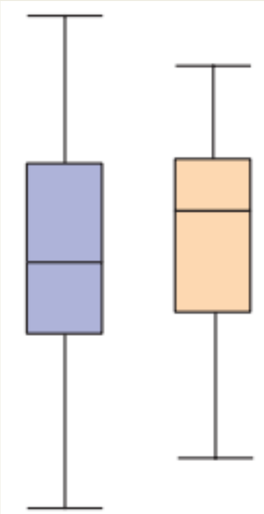
A difference is called **statistically significant** if the difference is greater than what we would expect from random chance... when repeating the same experiment over and over

- Flip a coin 100 times:
  - 54 heads is not statistically significant since it would not be surprising to observe this outcome.
  - 94 heads is statistically significant since it would be surprising to observe this outcome.

Statistical Significance



Statistically significant since the medians of each are outside the typical values of the other.



Not statistically significant since the medians of each are within the typical values of the other.

Random Samples and Random Treatments

- **Surveys** use a random group of participants.
- **Experiments** find a homogeneous group, separate them into random subgroups for treatment.
- <u>Experiments do not use a random sample from the general population.</u>
- Beware of stating that the participants from the experiment represent the larger population.

Control Groups and Control Treatments

Control

- Does eating ten carrots a day help you lose weight?

- Find 200 participants and randomly select 100 of them to eat ten carrots a day.

- The other 100 are the control group.

- **Not** eating ten carrots a day is the control treatment.

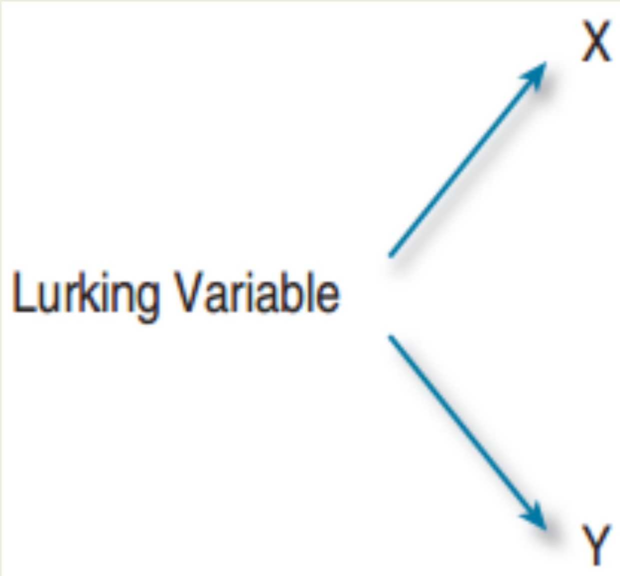
Blinding

- Single-blinding involves the participants not knowing whether they are in the control or treatment group.
- Double-blinding means neither the participant nor the person handing out the soda knows the label.

Placebos

- A placebo is a “fake” treatment that looks like the treatment being tested.
- Just telling a patient that they are being treated can aid recovery.
- This is called the placebo effect.
- Use a placebo for effective blinding.

Lurking and Confounding

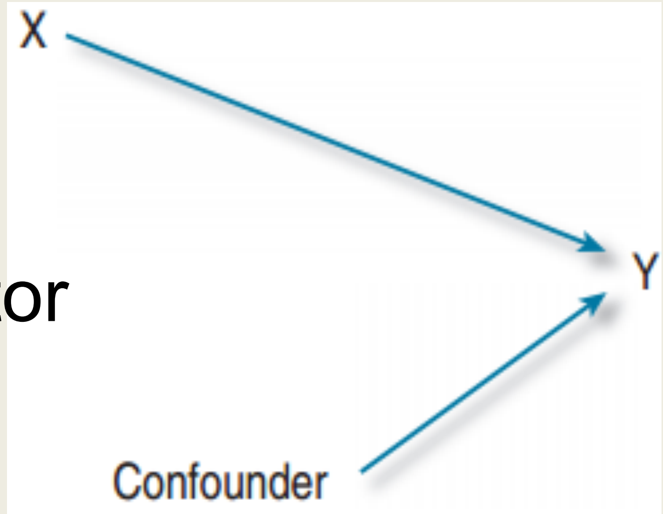


Lurking Variable

- Associated with both *x* and *y*
- Makes it appear that *x* causes *y*

Confounding Variable

- Associated in a noncausal way with a factor
- Affects the response
- Can’t tell if the cause was the factor or confounding variable



Formula sheet

$Range = Max - Min$
$IQR = Q3 - Q1$
Outlier Rule-of-Thumb: $y < Q1 - 1.5 \times IQR$ or $y > Q3 + 1.5 \times IQR$

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

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

$$z = \frac{y - \bar{y}}{s} \text{ (data based)}$$

$$r = \frac{\Sigma z_x z_y}{n - 1}$$

$$\hat{y} = b_0 + b_1 x \qquad \text{where } b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

$$P(\mathbf{A}) = 1 - P(\mathbf{A}^C)$$

$$P(\mathbf{A} \text{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B})$$

$$P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B} \mid \mathbf{A})$$

$$P(\mathbf{B} \mid \mathbf{A}) = \frac{P(\mathbf{A} \text{ and } \mathbf{B})}{P(\mathbf{A})}$$

If **A** and **B** are independent, $P(\mathbf{B} \mid \mathbf{A}) = P(\mathbf{B})$

Random Phenomena Vocabulary

Trial

- Each occasion in which we observe a random phenomena (the thing that is happening)

Outcome

- The value of the trial for the random phenomena (what could happen)

Event

- The combination of the specific trial's <u>outcomes</u> (what actually happened)

Sample Space

- The collection of all possible outcomes

Flipping Two Coins

Trial

- The flipping of the two coins

Outcome

- Heads or tails for each coin flip

Event

- HT, for example

Sample Space

- **S** = {HH, HT, TH, TT}

The Law of Large Numbers

- If you flip a coin once, you will either get 100% heads or 0% heads.

- If you flip a coin 1000 times…
- …you will probably get close to 50% heads.

The **Law of Large Numbers** states that for many trials, the proportion of times an event occurs settles down to one number.

•This number is called the empirical probability.

The Nonexistent Law of Averages

Wrong

- If you flip a coin 6 times and get 6 heads, then you are due for a tail on the next flip.
- You put 10 quarters in the slot machine and lose each time. You are just a bad luck person, so you have a smaller chance of winning on the 11th try.

- There is no such thing as the Law of Averages for **short** runs.

Theoretical Probability

American Roulette

- 18 Red, 18 Black, 2 Green
- If you bet on Red, what is the probability of winning?

Theoretical Probability

- $P(\mathbf{A}) = \dfrac{\text{\# of outcomes in } \mathbf{A}}{\text{\# of possible outcomes}}$
- $P(\text{red}) = \dfrac{18}{38}$

Rules: 1, 2 and 3

Rule 1: $0 \le P(\mathbf{A}) \le 1$

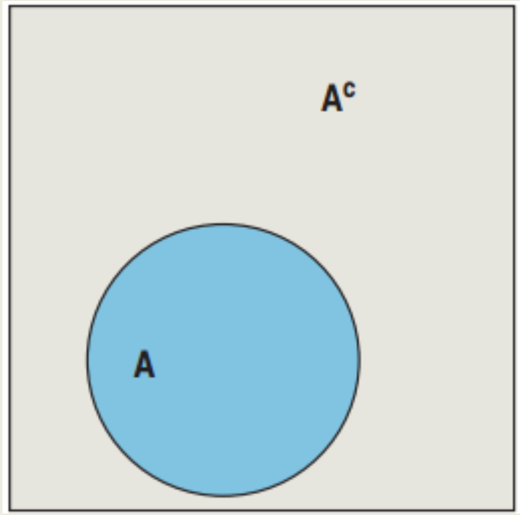
That's how we define probability

Rule 2: $P(\mathbf{S}) = 1$

The set of all possible outcomes has probability 1

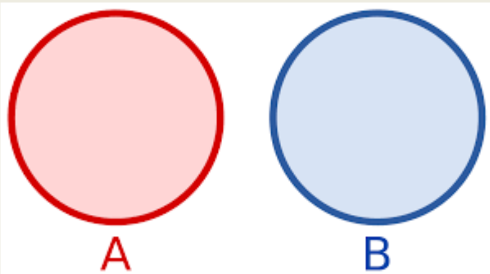
The Rule of Complements: $P(\mathbf{A}^C) = 1 - P(\mathbf{A})$

$\mathbf{A}^C$ is the event of "A not happening".



Events

Disjoint



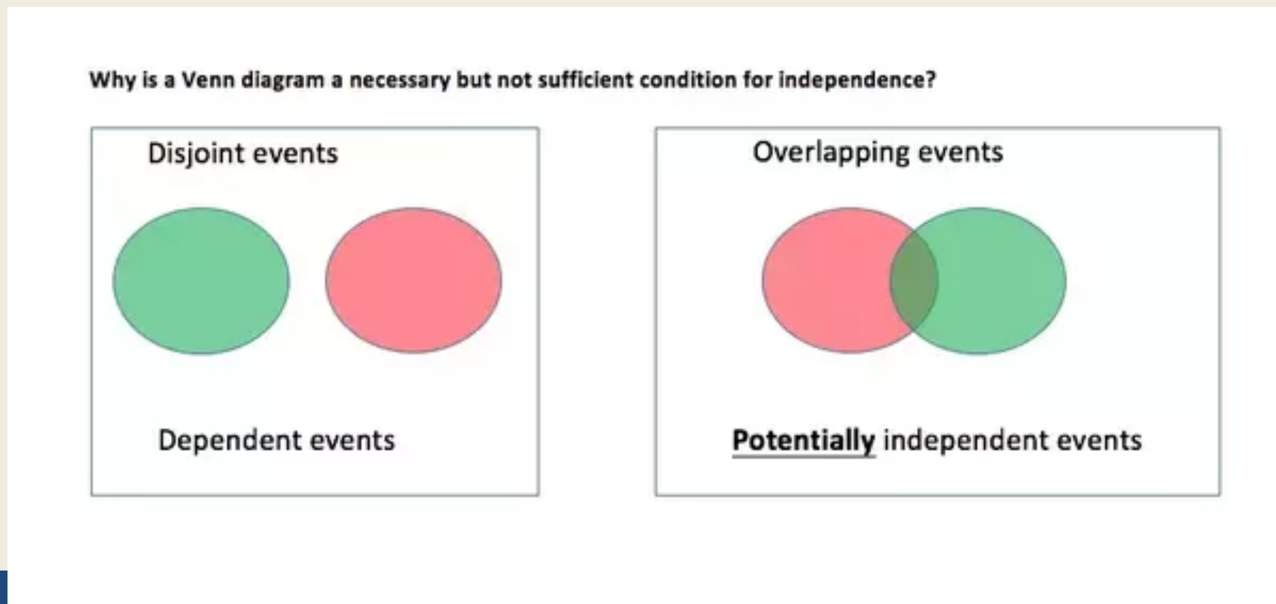
Independent and Dependent

Independent Events

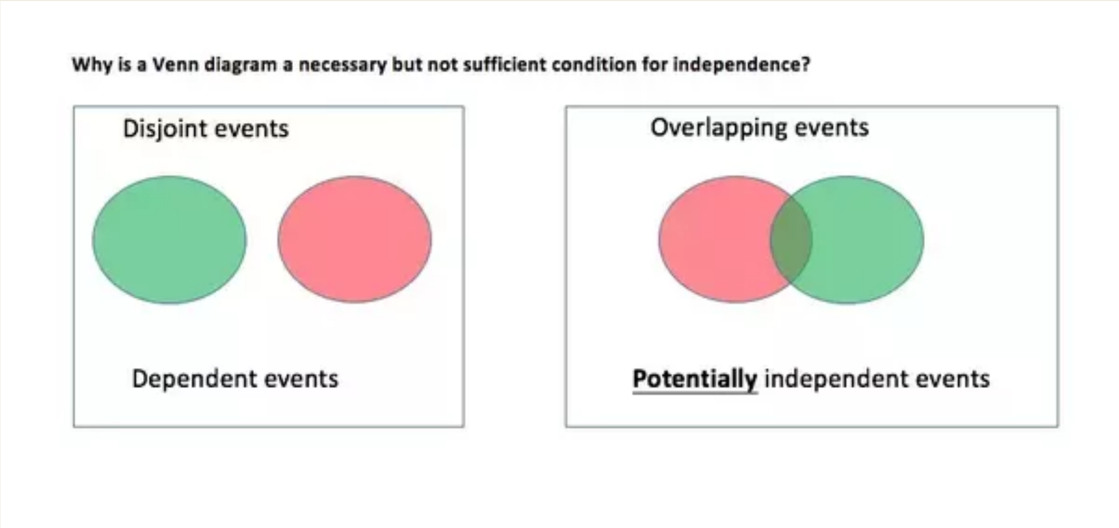
Two or more events that occur in a sequence. If the outcome of any event **does not** affect the possible outcomes of the other event(s), then the events are independent.

Dependent Events

Two or more events that occur in a sequence. If the outcome of any event **changes** the possible outcomes of the other event(s), then the events are dependent.



Events



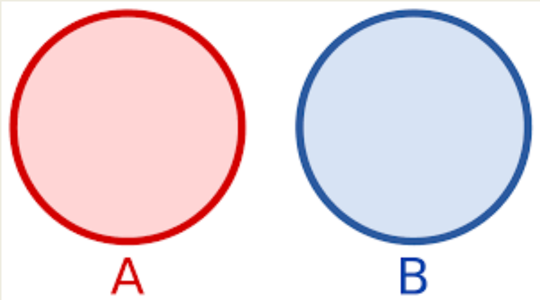
	Disjoint	Overlapping
Dependent	YES	YES
Independent	DOES NOT EXIST	YES

Rule 4: The Addition Rule

Suppose

P(sophomore) = 0.2 and P(junior) = 0.3

- Find P(sophomore OR junior)
- Solution: 0.2 + 0.3 = 0.5
- This works because sophomore and junior are **disjoint events**. They have no outcomes in common.



The Addition Rule

- If **A** and **B** are **disjoint** events, then

$$P(\mathbf{A} \text{ OR } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$$

Rule 5: The Multiplication Rule

The probability that an Atlanta to Houston flight is on time is 0.85.

- If you have to fly every Monday, find the probability that your first two Monday flights will be on time.

Multiplication Rule: For **independent** events **A** and **B**:

$$P(\mathbf{A} \text{ AND } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$$

- $P(1^{st} \text{ on time AND } 2^{nd} \text{ on time})$

$$= P(\text{1st on time}) \times P(\text{2nd on time})$$

$$= 0.85 \times 0.85$$

$$= 0.7225$$

Red Light AND Green Light AND Yellow Light

Find the probability that the light will be red on Monday, green on Tuesday, and yellow on Wednesday.

- The multiplication rule works for more than 2 events.
- P(red Mon. AND green Tues. AND yellow Wed.)

  = P(red Mon.) × P(green Tues.) × P(yellow Wed.)

  = 0.61 × 0.35 × 0.04

  = 0.00854

At Least One Red Light

Find the probability that the light will be red at least one time during the week.

- Use the Complement Rule.
- P(at least 1 red)

  $= 1 - P(\text{no reds})$

  $= 1 - (0.39 \times 0.39 \times 0.39 \times 0.39 \times 0.39 \times 0.39 \times 0.39)$

  ≈ 0.9986

Chapter 14

Probability Part 2

The General Addition Rule

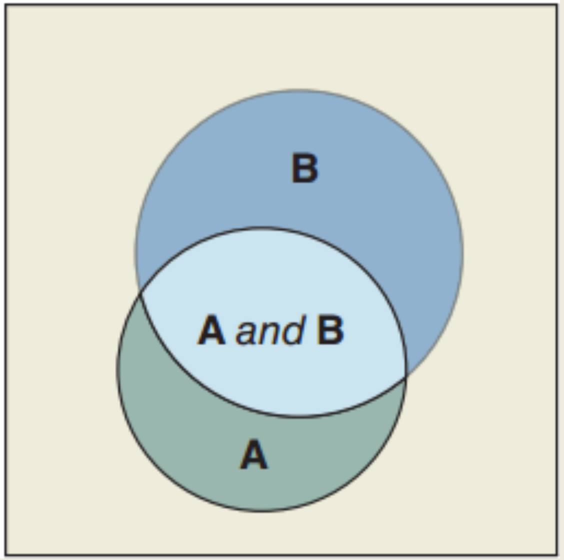
$P(\mathbf{A} \; or \; \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \; and \; \mathbf{B})$

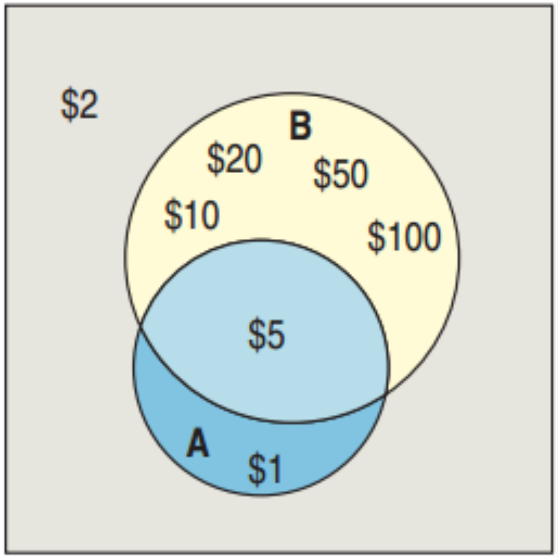
- **The General Addition Rule in words:** Add the probabilities of the two events and then subtract the probability of their intersection.

P(odd amount *or* bill with a building)

$= P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B}\}$

$= P(\{\$1, \$5\}) + P(\{\$5, \$10, \$20, \$50, \$100\}) - P(\{\$5\})$





Facebook or Twitter?

71% use Facebook, 18% Twitter, 15% both

- What is the probability that a randomly selected person:

1. Uses either Facebook or Twitter?
2. Uses either Facebook or Twitter, but not both?
3. Doesn’t use Facebook or Twitter?

- Plan:
  - **A** = {uses Facebook}
  - **B** = {uses Twitter}

Contingency Table

		Goals			
		Grades	Popular	Sports	Total
Sex	Boy	117	50	60	227
	Girl	130	91	30	251
	Total	247	141	90	478

A table that displays the results of two categorical questions is called a **contingency table**.

- P(girl) = 251/478 = 0.525
- P(girl *and* popular) = 91/478 = 0.190
- P(sports) = 90/478 = 0.188

Conditional Probability

		Goals			
		Grades	Popular	Sports	Total
Sex	Boy	117	50	60	227
	Girl	130	91	30	251
	Total	247	141	90	478

- What if we knew the chosen person was a girl? Would that change the probability that the girl's goal was sports?
- Yes! We write *P*(sports | girl)
- Only look at Girl row: *P*(sports | girl) = 30/251 = 0.120
- Find the probability of selecting a boy given the goal is grades.
- *P*(boy | grades) = 117/247 = 0.474

Conditional Probability Formula

		Goals			
		Grades	Popular	Sports	Total
Sex	Boy	117	50	60	227
	Girl	130	91	30	251
	Total	247	141	90	478

Probability of **B** *Given* **A**:

- $P(\mathbf{B} \mid \mathbf{A}) = \dfrac{P(\mathbf{A} \textit{ and } \mathbf{B})}{P(\mathbf{A})}$
- Example:

$$P(\text{girl} \mid \text{popular}) = \frac{P(\text{girl } \textit{and} \text{ popular})}{P(\text{popular})}$$

$$= \frac{91/478}{141/478}$$

$$= \frac{91}{141} = 0.65$$

The General Multiplication Rule

- For **A** and **B** **independent**, we had:

  $$P(\mathbf{A}\ \textit{and}\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$$

- Rearranging the conditional probability equation, we get the **General Multiplication Rule**:

  $$P(\mathbf{A}\ \textit{and}\ \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B} \mid \mathbf{A})$$

- Equivalently,

  $$P(\mathbf{A}\ \textit{and}\ \mathbf{B}) = P(\mathbf{B}) \times P(\mathbf{A} \mid \mathbf{B})$$

Definition of Independence

- Events **A** and **B** are independent if knowing **A** happened does not change the probability of **B**. In symbols:

  $$\mathbf{A} \text{ and } \mathbf{B} \text{ are independent} \leftrightarrow P(\mathbf{B} \mid \mathbf{A}) = P(\mathbf{B})$$

- Equivalent formulas for independence:
  - $P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{A})$
  - $P(\mathbf{A} \textit{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$

Grades and Girl Independent?

		Goals			
		Grades	Popular	Sports	Total
Sex	Boy	117	50	60	227
	Girl	130	91	30	251
	Total	247	141	90	478

- Determine if the "goal of good grades" and gender are independent.

- Are the "goal of sports" and gender independent?

Independent ≠ Disjoint

Disjoint events cannot be independent.

- Consider the events:
  - Course grade A
  - Course grade B
  - Disjoint: You can't get both.
  - Not independent: $P(\mathbf{A} \mid \mathbf{B}) = 0 \neq P(\mathbf{A})$
  - A and B are disjoint (also called mutually exclusive) but not independent.

Conditional probability
Probability of **B** *Given* **A**:

$$P(\mathbf{B} \mid \mathbf{A}) = \frac{P(\mathbf{A} \text{ and } \mathbf{B})}{P(\mathbf{A})}$$

Events

DEFINITIONS	OR ADDITION more general → P $P(\mathbf{A} \text{ or } \mathbf{B}) =$	AND MULTIPLICATION more restrictive → P $P(\mathbf{A} \text{ and } \mathbf{B}) =$
Disjoint dependent (mutually exclusive) $P(\mathbf{A} \text{ and } \mathbf{B}) = 0$	Addition rule $P(\mathbf{A}) + P(\mathbf{B})$ because $P(\mathbf{A} \text{ and } \mathbf{B}) = 0$	NA
Overlapping independent $P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$ $P(\mathbf{B} \mid \mathbf{A}) = P(\mathbf{B})$ $P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{A})$	General addition rule $P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B})$ If specified as exclusionary: $P(\mathbf{A}) + P(\mathbf{B}) - 2P(\mathbf{A} \text{ and } \mathbf{B})$	Multiplication rule $P(\mathbf{A}) \times P(\mathbf{B})$ because $P(\mathbf{B} \mid \mathbf{A}) = P(\mathbf{B})$ $P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{A})$
Overlapping dependent $P(\mathbf{A} \text{ and } \mathbf{B}) \neq P(\mathbf{A}) \times P(\mathbf{B})$ $P(\mathbf{B} \mid \mathbf{A}) \neq P(\mathbf{B})$ $P(\mathbf{A} \mid \mathbf{B}) \neq P(\mathbf{A})$	General addition rule $P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B})$ If specified as exclusionary: $P(\mathbf{A}) + P(\mathbf{B}) - 2P(\mathbf{A} \text{ and } \mathbf{B})$	General multiplication rule $P(\mathbf{A}) \times P(\mathbf{B} \mid \mathbf{A})$ $P(\mathbf{B}) \times P(\mathbf{A} \mid \mathbf{B})$

Midterm

*

Midterm tomorrow during class time

Have calculator, z-table

Exam: 40 questions total

A few questions are worth more than one point

Problems: boxplots, contingency table, z-scores, regression, probability

Best way to study: **review HW and class exercises**

Review lectures slides

How to approach exam: time yourself (3 min per point)

Final grades will be curved