

## Variables

- **Categorical / Qualitative**
  - Classifies subject by an attribute or characteristic.
  - Hair color, type of professor, make of car
- **Quantitative**
  - Gives numerical measures of subjects.
  - Weight, height, response time, number of miles traveled to work

## Quantitative Variables

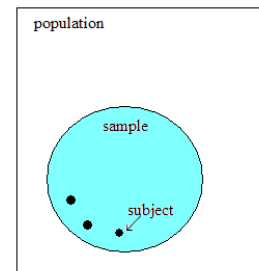
- **Discrete**
  - A countable number of whole-numbered values (no decimals).
    - # of people entering a shop per hour (whole number)
    - # of living grandparents (0,1,2,3,4 only)
    - # of spades in a poker hand (0,1,2,3,4,5 only)
    - # of balls a juggler is currently juggling
- **Continuous**
  - Can take on any numerical value (including decimals) on an interval.
    - The weight of an athlete (150, 150.01, 181.312, etc)
    - The time taken to complete a lap (in seconds, minutes, etc.)
    - The current speed of an airplane (in mph, Mach, etc.)
    - The speed of an angry fire ant (in cm/second, say)

## Examples (HW 1.1-2.1)

- Which of these are categorical or quantitative? For the latter, which are discrete or continuous?
- Length of an earthworm (in mm)  
**Quantitative, continuous**
- Region of U.S. (Southeast, West, etc.)  
**Categorical**
- Literary genre  
**Categorical**
- Number of times in one month the Creswell fire alarm goes off  
**Quantitative, discrete**

## Important Terms

- **Population**
  - Total set of subjects in which we are interested
- **Sample**
  - A subset of the population for which we have data
- **Subject**
  - Entities we measure (individuals)



## Important Terms

- **Parameter**
  - A numerical value summarizing the population data.
  - Ex: number of freshmen out of all STAT 2000 students
- **Statistic**
  - A numerical value summarizing the sample data.
  - Ex: number of freshmen out of a sample of 100 STAT 2000 students
- Parameter & Population both begin with P
- Statistic and Sample both begin with S

## Notation

We use different letters for population parameters versus sample statistics.

$$\begin{array}{ll} \mu = \text{population mean} & \bar{x} = \text{sample mean} \\ \sigma = \text{population st. dev.} & s = \text{sample st. dev.} \\ p = \text{population proportion} & \hat{p} = \text{sample proportion} \end{array}$$

## Example (HW 1.1 – 2.1)

- A college dean wants to know the average age of the faculty. She takes a random sample of 10 faculty members and averages their ages.
- Population = all faculty members
- Sample = the 10 faculty members selected
- Subject = an individual faculty member
- Parameter = average age of all faculty members
- Statistic = average age of the 10 selected

## Descriptive vs. Inferential

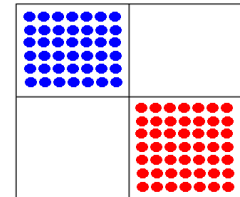
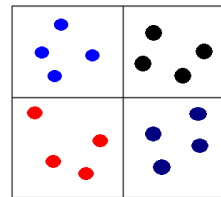
- Descriptive Statistic
  - Summary of the data in the sample.
    - Majority of students in a sample of 1000 attend UGA football games
- Inferential Statistic
  - A conclusion or prediction about the population based on the sample data.
    - Majority of all UGA students attend UGA football games, based on the sample

## Sampling Methods

- Simple Random Sampling
  - Each subject everywhere has an equally likely chance of being selected
  - Often done with a random number table
  - Choosing a company somewhere in the U.S.
- Systematic
  - Selecting every “k-th” subject
  - Surveying every 10<sup>th</sup> person we meet downtown
- Convenience
  - Individuals are easily found (e.g. internet surveys)
  - Often the “laziest” way, so less reliable answers

## Sampling Methods

- Stratified Sampling
  - Taking **some** subjects from all possible groups
- Cluster Sampling
  - Taking **all** subjects from some possible groups



## Sampling (HW 4.1-4.4)

- A researcher takes 3 possible classifications of companies, each of which contains 1000 businesses, and draws 100 random subjects from all three. What type of sampling is this?
  - Stratified
- Suppose instead she draws 200 businesses at random from the whole population of companies. What type of sampling is this?
  - Simple Random Sampling
- The same researcher instead randomly selects 2 of the 3 possible classifications and then surveys all businesses in those groups. What type of sampling is this?
  - Cluster
- Suppose instead she gets an alphabetical list of all these companies, starts with #4, and selects every 100<sup>th</sup> after that for her sample.
  - Systematic

## Random Table (HW 4.1-4.4)

- A study will assign subjects numbered 1 – 8 into one of two groups, four in each. Use the table to decide who goes into the first group. Start with the top left, and answer in numeric order.

30494 17011  
22368 46573

Skip 0 and 9 because only eight subjects  
Skip the second 4 because already picked  
Persons 1, 3, 4, 7 will go into the first group.  
So persons 2, 5, 6, 8 will be the 2<sup>nd</sup> group.

## Frequencies

$$\text{proportion} = \frac{\text{frequency}}{\text{total number of observations}}$$

$$\text{percentage} = \left( \frac{\text{frequency}}{\text{total number of observations}} \right) \times 100$$

Example: 18 cookies out of a random sample of 32 are chocolate chip

$$\text{proportion} = \frac{18}{32} = .5625$$

$$\text{percentage} = \left( \frac{18}{32} \right) \times 100 = 56.25\%$$

## Frequencies (HW 2.1-2.2)

- Results from the question of how many children a family has had. Fill in the answers.

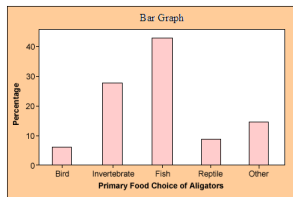
# Children	0	1	2	3
Count	786	460	662	489

- Proportion      .32791   .19191   .27618   .20401
- Percentage      32.791% 19.191% 27.618% 20.401%
- Total number = 2397 families

## Types of Charts (Categorical)

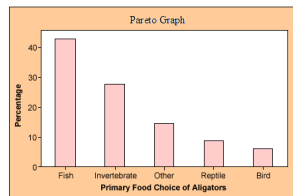
### Bar Graph

- Categories on horizontal axis, frequency on vertical axis, height of rectangle is frequency



### Pareto Graph

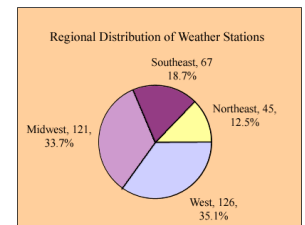
- A bar graph arranged with bars in descending order of frequency



## Types of Charts (Categorical)

### Pie Chart

- A circle divided into slices, with each slice representing a category of a variable
- Size of a slice represents overall percentage
- To determine mode, easier to use a bar chart



## Categorical Data (HW 2.2)

- Consider the following table of 240 animal tracks that were found in a certain park:

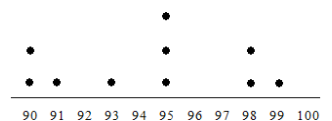
Raccoon	Opossum	Squirrel	Deer	Chipmunk
83	19	90	28	20

- What proportion of tracks were not from raccoons or deer?
  - We have  $240 - 83 - 28 = 129$ , so  $129/240 = .5375$
- If we were to make a pie chart, which animals would have the largest and smallest slices?
  - Largest: Squirrel      Smallest: Opossum
- Can we find the mean, median, mode, and range from this data? If so, find them. We can't compute the mean, median, or range since this data is categorical. The best we can do is the mode, which is "Squirrels", since it has the highest frequency.

## Types of Charts (Quantitative)

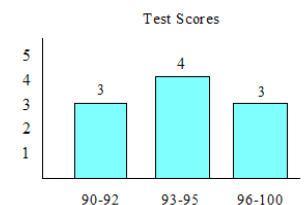
### Dot Plot

- Places a dot for every data value above a number line



### Histogram

- A bar graph for quantitative data

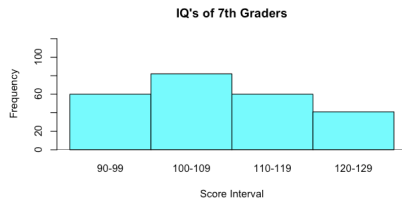


### A's on a Test

90 90 91 93 95  
95 95 98 98 99

## Histogram Interpretation (HW 2.2)

- How many total students sampled?  
 $60 + 80 + 60 + 40 = 240$
- Which class has highest / lowest frequency? What are those frequencies?
- Highest: "100-109" with 80  
Lowest: "120-129" with 40
- How many students have an IQ between 100 and 119?  $80 + 60 = 140$



## Stem-And-Leaf Plot

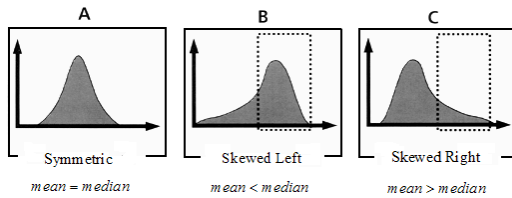
- A bar chart on its side
- "Stem" is all digits except the last one
- Last digit is the "leaf"
- Ascending order
- No commas
- If nothing in a row, write the row, but leave it blank

```

19 | 9
20 |
21 | 00
22 | 35558
23 | 25
    
```

Example (HW 2.1-2.2)  
eBay selling prices  
199 210 210 223 225  
225 225 228 232 235

## Skewness



## Outliers

- The mean is sensitive to outliers.
- The median is resistant to outliers.
- When outliers are present, best to use median as measure of central tendency.
- Examples:
  - Earthquake magnitudes on the Richter Scale (skewed right since some, but very few, big earthquakes)
  - Ages of MENSA members at the time they joined (skewed left since most were adults, but a few children had high enough IQs)

## Outliers Example

- Miles traveled on public transportation  
0 0 3 0 0 0 9 0 5 0  
Mean = 1.7      Median = 0
- Now introduce a new data point: 90  
0 0 3 0 0 0 9 0 5 0 90  
Mean = 9.72727      Median = 0

## Mean & Median (HW 2.3-2.4)

- The number of trains a British person takes to get from one town to another in England can be modeled as follows. 200 Brits were sampled, and the results are listed. Compute the mean and median. What can you say about the distribution's shape?

Trains Taken	Frequency
1	30
2	102
3	36
4	32
Total	200

## Mean & Median (HW 2.3-2.4)

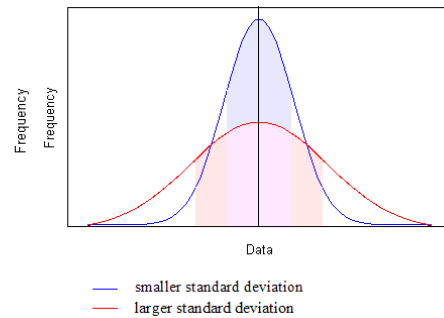
Trains Taken	Frequency
1	30
2	102
3	36
4	32
Total	200

$$\text{mean} = \frac{(1 \times 30) + (2 \times 102) + (3 \times 36) + (4 \times 32)}{200} = 2.35$$

- For the median, find half the total count (about 100), so we need to find where person # 100 is.
- It's not in Row 1 since we have the first 30 only
- After Row 2, we have 30 + 102 = 132 people
- **Median = 2 since person # 100 falls in row 2**
- **Mean > median => somewhat skewed right**

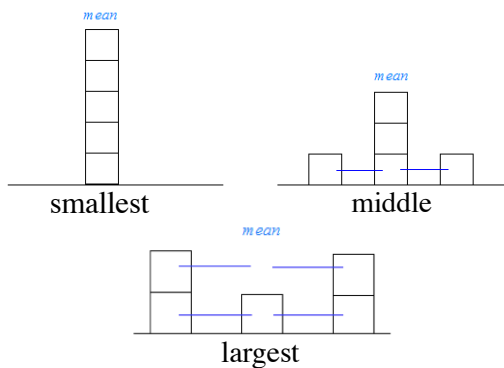
## Standard Deviation

- The average distance between any data point and the mean of the data.
- Measures how much/little the data distribution is spread out.
- Which has larger and smaller st. dev.?



## Standard Deviation (HW 2.3 - 2.4)

Which has largest and smallest standard deviation?



## StatCrunch Commands

### Summary Stats

1. Enter data in one column
2. Stat > Summary Stats > Columns
3. Select column var1
4. Calculate

### Regression

1. Enter data in two columns (same order)
2. STAT > REGRESSION > SIMPLE LINEAR
3. Select columns var1 and var2
4. Calculate

## Summary Stats Example From StatCrunch

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
var1	15	7.0933332	5.6920953	2.3858113	0.6160138	6.9	7.5	3.7	11.2	4.7	9.2

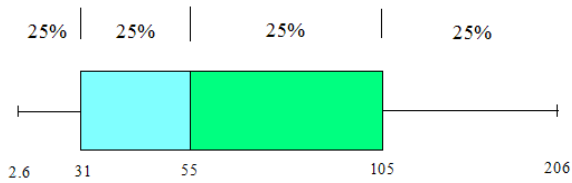
- **Mean = 7.09333**
  - Average of the data set
- **Standard Deviation = 2.38581** (average spread in data set)
- **Q1 = 4.7** (25% of data lie below this)
- **Median (sometimes Q2) = 6.9**
  - 50% of data lie below (and above) this value.
- **Q3 = 9.2** (75% of data lie below this)
- **Range = 7.5**
  - Difference between maximum (11.2) and minimum (3.7)

## Box-Plot (HW 2.5-2.6)

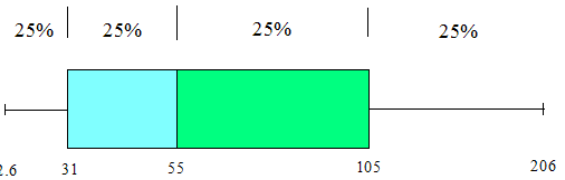
### Distribution of taxes (in cents)

- Minimum = 2.6      Q3 = 105
- Q1 = 31      Maximum = 206
- Median = 55
- What proportion of states have taxes...
  - Greater than 31 cents?
  - Greater than \$1.05 (105 cents) ?

## Box-Plot (HW 2.5-2.6)



- Greater than 31 cents: **.75**
- Greater than \$1.05: **.25**



- Between what two values are the middle 50% of the data found?
  - The quartiles: **31 and 105**
- Find and interpret the interquartile range.
  - $IQR = Q3 - Q1 = 105 - 31 = 74$
  - The range for the middle half of the data.

## New Box-Plot (HW 2.5-2.6)

### Computer Drive Use (in kilobytes)

- Min = **4**                      Q3 = **1105**
- Q1 = **256**                      Max = **320,000**
- Median = **530**
- Is this bell-shaped or skewed?
- Use the  $1.5 * IQR$  rule to test for outliers.



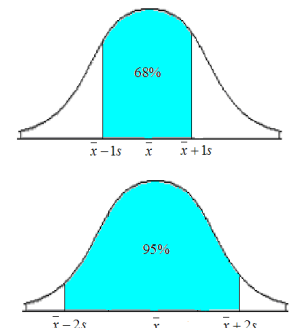
- Notice the median is closer to the left side of the middle box, suggesting right-skewness
- The right line (from Q3 to highest point) is much longer than the left line
- So the distribution is skewed right.

## Box-Plot (HW 2.5-2.6)

- **Skewed right**
- $1.5 * IQR = 1.5 (Q3 - Q1) = 1.5 (1105 - 256) = 1273.5$
- $Q1 - 1.5 * IQR = 256 - 1273.5 = -1017.5$
- **Because there are no points beneath this cutoff, we have no lower outliers.**
- $Q3 + 1.5 * IQR = 1105 + 1273.5 = 2378.5$
- **Because the max is greater than this cutoff ( $320,000 > 2378.5$ ), we have an upper outlier.**

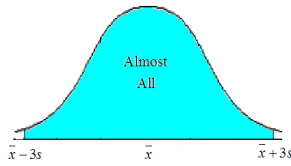
## Empirical Rule

- **Only used for bell-shaped distributions**
- Within one standard deviation from the mean, we have 68% of all data points.
- Within two standard deviations from the mean, we have 95% of all data points.



## Empirical Rule

- Within three standard deviations from the mean, we have almost all data points.
- Anything else is an outlier.



### SUMMARY

- 1 s: 68%
- 2 s: 95%
- 3 s: Almost all

## Example (HW 2.3-2.4)

- The weight of a zebra is bell-shaped with an average of 700 pounds and a standard deviation of 70 pounds.
- Give an interval within which about 95% of the data fall.

$$\bar{x} = 700 \quad s = 70$$

95% means we go left and right 2 deviations

$$\text{Lower Limit: } \bar{x} - 2s = 700 - (2 \times 70) = 560$$

$$\text{Upper Limit: } \bar{x} + 2s = 700 + (2 \times 70) = 840$$

So the interval is (560,840)

## Example (HW 2.3-2.4)

- The weight of a zebra is bell-shaped with an average of 700 pounds and a standard deviation of 70 pounds.
- Approximately what percentage of the data is between 630 and 770?

Notice  $700 - 630 = 70$  and  $770 - 700 = 70$

We have therefore gone out 70 units, which is 1 deviation from the mean.

By the Empirical Rule, 1 deviation has about 68% of the data.

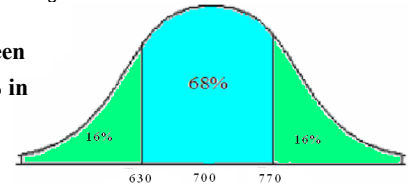
- Find the weight of a zebra that is three standard deviations above the mean.

$$\bar{x} + 3s = 700 + (3 \times 70) = 910$$

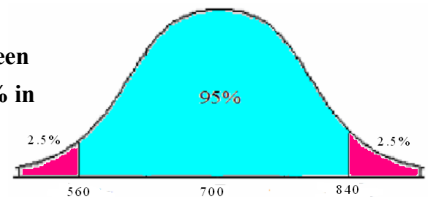
## Example (HW 2.3-2.4)

- Approximately what percentage of the data is between 560 and 770?

We have 68% between 630 and 770 (so 16% in both tails)



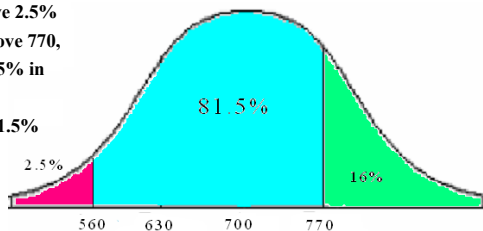
We have 95% between 560 and 840 (so 2.5% in both tails)



## Example (HW 2.3-2.4)

- Approximately what percentage of the data is between 560 and 770?

That means we have 2.5% below 560, 16% above 770, and therefore 81.5% in the middle.  
 $100 - 16 - 2.5 = 81.5\%$



## Z-Score

$$Z = \frac{(\text{value}) - (\text{mean})}{(\text{standard deviation})} = \frac{\text{value} - \bar{x}}{s} \quad \text{or} \quad \frac{\text{value} - \mu}{\sigma}$$

- A z-score is the number of standard deviations above/below the mean the data point lies.
  - If negative: data point is below mean
  - If positive: data point is above mean
- Data point is an outlier if...
  - Z-score > 3, or
  - Z-score < -3

## Z-Score (HW 2.5-2.6)

- For 261 heights, the mean was 65.8 inches and the standard deviation was 3.0 inches. The shortest person in this sample had a height of 56 inches.
- Calculate the z-score for this person.

$$Z = \frac{(\text{data point}) - (\text{mean})}{(\text{st. dev.})} = \frac{56 - 65.8}{3.0} = -3.26667$$

- Interpret the Z-score.  
This person's height is 3.26667 standard deviations below the mean (because it's negative). It's less than -3, so it's an unusual observation, an outlier.

## Z-Score (HW 2.5-2.6)

- For 261 heights, the mean was 65.8 inches and the standard deviation was 3.0 inches.
- What is the Z-score for someone whose height is 2.4 standard deviations below the mean?

$$z = -2.4 \text{ (negative because below mean)}$$

- Find the height corresponding to the above Z-score.

$$-2.4 = \frac{x - 65.8}{3.0} \Rightarrow x = (-2.4)(3.0) + 65.8 \Rightarrow x = 58.6 \text{ in}$$

## Percentiles

- The 20th percentile, for example, is the "cutoff" such that 20% of the subjects have a score falling beneath that cutoff
- So, x% of subjects fall beneath the xth percentile
- Example: We have 200 subjects. To find the number falling beneath the 20th percentile, we take 20% of 200, which is  $200 * .20 = 40$ .
- Therefore 40 subjects (out of 200) fall below the 20th percentile.
- QUESTION**
- For 200 subjects, how many fall above the 45th percentile?
  - $200 * .45 = 90$  fall below the 45th percentile
  - Therefore  $200 - 90 = 110$  fall above
  - Alternatively,  $200 * (1 - .45) = 110$  (just a different method)
- For 200 subjects, what is the percentile for the person who's 52nd from the top? The 52nd from the top is the  $200 - 52 = 148$ th person, so  $148/200 = .74$ . Therefore it's the 74th percentile
- Answer: 74 (a whole number, so round to nearest if necessary)

## Variable Types

- Response**
  - Determined by another variable
  - y-variable, on the vertical axis (scatter plots)
- Explanatory**
  - Explains or affects the response variable
  - x-variable, on the horizontal axis (scatter plots)
- A contingency table is a table that relates two categorical variables
  - Explanatory variable on the side
  - Response variable on the top

## Variables (HW 3.1)

- Consider a study in which you are interested in any connections between gender and preference for dessert (chocolate cake or ice cream)
- Explanatory: **gender**
- Response: **dessert preference**
- Could your gender possibly determine your preference for dessert? (Sounds reasonable)
- Could your preference for dessert possibly determine your gender? (Rubbish!)

	Good Adjustment	Bad Adjustment	Total
Orientation	72	14	86
No Orientation	28	45	73
Total	100	59	159

- This is a chart of students that took freshmen orientation and students that did not, and whether they adjusted well or poorly to college
- 86 / 159 did orientation
- 59 / 159 adjusted poorly
- 14 / 86 is the proportion of the "orientation students" that did not adjust well (as opposed to all students surveyed)

## Relative Risk (HW 3.1)

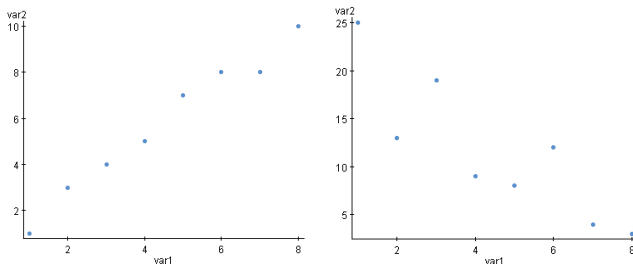
relative risk =  $\frac{\text{conditional proportion for one group (larger number)}}{\text{conditional proportion for another group (smaller number)}}$

- Relative risk tells us how many times more likely the outcome is for one group than the other group.
- The following three facts therefore follow:
  - Relative risk  $\geq 1$ .
  - When the numerator and denominator proportions are very similar, relative risk will be very close to 1.
  - However, when the numerator is quite a bit larger, then relative risk will be quite a bit greater than 1.

	Good Adjustment	Bad Adjustment	Total
Orientation	72	14	86
No Orientation	28	45	73
Total	100	59	159

- Find the proportion of "orientation-students" that adjusted well.
 
$$72/86 = 0.83721$$
- Find the proportion of "no-orientation-students" that adjusted well.
 
$$28/73 = 0.38356$$
- Find the relative risk of adjusting well to college for both groups of students. [Look at the "Good Adjustment proportions"](#)
  - Larger / Smaller =  $0.83721 / 0.38356 = 2.18272$
- Students that [did orientation](#) are 2.18272 times more likely to adjust well to college than students that [did not do orientation](#).

## Scatter Plots



Strong, + correlation

Weak, - correlation

## Correlation (r) (HW 3.2 - 3.4)

- $-1 \leq r \leq 1$
- If r is positive, then so is the slope. (Same if r's negative)
- Closer r is to 1 (or -1), strong correlation
- Closer r is to 0, weak correlation
- r is unitless
- r does not change if we flip variables
- r measures only LINEAR relationship
- A strong correlation is **not** proof that one variable causes the other
- Which of the following has the strongest and weakest correlation?

.80      .67      -.34      .11      -.92

Strongest: -.92 (closest to a 1)

Weakest: .11 (closest to 0)

## Lurking Variables Example

- x = # of ounces of coffee drunk the day before an exam
- y = score on that exam
- Strong correlation does **not** prove that drinking more coffee causes an exam score to increase (there could be lurking variables)
  - Number of hours reviewing
  - GPA

## Least-Squares Regression

$$\hat{y} = a + bx$$

- x = given data point
- $\hat{y}$  = predicted response
- a = intercept
  - Predicted response when x = 0
  - May not always have a practical interpretation!
- b = slope
  - Slope is how much the predicted response increases (or decreases) for every unit increase in x

## Regression (HW 3.2-3.4)

- We want to predict average monthly car insurance payments ( $y$ ), given the number of accidents ( $x$ ) the client has had within the past three years.
 
$$\hat{y} = 137.11 + 39.82x$$
- What's the predicted payment for someone who's had 2 accidents?
 
$$\hat{y} = 137.11 + 39.82(2) = 216.75$$
- Interpret the slope and intercept.
  - For every additional accident, payment is expected to increase by \$39.82
  - The expected payment for someone with no accidents is \$137.11
- Is correlation positive or negative?
  - Positive because the slope's positive

## Example (HW 3.2-3.4)

- The predicted number of visitors in Destin during the summer is to be modeled.
- For every 1 degree (in Fahrenheit) in temperature, the predicted number of beach visitors increases by 265. The y-intercept is 15,000.
- Using this information, write down the regression equation.

$$\hat{y} = 15000 + 265x$$

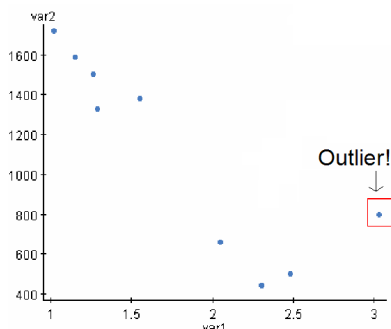
## Regression (HW 3.2-3.4)

- A shop owner wants to assign a new price for dog biscuit packets. She is curious how the price per packet ( $x$ , in dollars) affects the number sold per day ( $y$ ). She studies previous years' data and gets  $\hat{y} = 98 - 18x$ .
- Interpret the slope.
  - For every dollar increase in price, the number of dog biscuit packets sold per day is expected to decrease by 18.
- Interpret the intercept.
  - Literally: when price is \$0 (free!), the number sold per day is about 98 packets
  - Nonsense, so intercept has no interpretation here

## Regression (HW 3.2-3.4)

- We want to predict the number of misprints ( $y$ ) in a novel that's  $x$  pages long (in hundreds). For instance,  $x = 2.5$  is a 250 page novel. The regression equation is  $\hat{y} = 5.1 + 3.2x$
- Interpret the intercept (choose the best answer):
  - For every additional 100 pages, the predicted number of misprints goes up by 5.1.
  - The number of misprints in a novel 0 pages long is about 5.1.
  - The intercept has no practical interpretation.
- Interpret the slope (choose the best answer):
  - For every additional 3.2 pages, the predicted number of misprints goes up by 1.
  - A novel 400 pages long can be expected to have 3.2 more misprints than a novel 300 pages long.
  - The slope has no practical interpretation.

## Spotting an Outlier



## Regression Output

### Simple linear regression results:

Dependent Variable: var2

Independent Variable: var1

var2 = 30.761627 - 0.13430233 var1 ← Regression Equation

Sample size: 5

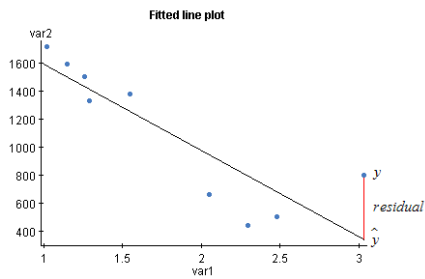
R (correlation coefficient) = -0.9956 ← Correlation

R-sq = 0.9911769 ← **Don't Use!**

Estimate of error standard deviation: 0.6068089

- The two bolded lines above are what you should use
- Use R (and not R-sq) for correlation

## Residuals



$$\begin{aligned} \text{residual} &= \text{observed} - \text{predicted} \\ &= y - \hat{y} \end{aligned}$$

## Residual (HW 3.2-3.4)

- The car insurance question again:  $\hat{y} = 137.11 + 39.82x$
- The predicted payment for someone with 2 recent accidents was \$216.75. Suppose someone with 2 accidents had an actual payment of \$201. Compute this person's residual.

$$y = 201 \quad \hat{y} = 216.75 \quad y - \hat{y} = -15.75$$

- Negative because actual was less (below the regression line)
- The model is based on people with between 0 and 6 accidents. Can we use it to predict the payment for someone with 13 recent accidents?
  - No—the model is linear only between  $x = 0$  and 6. Who knows what happens outside that range? (This is extrapolation)

## Extrapolation (HW 3.2-3.4)

- This is a valid prediction for years between 1900 and 2000
- But not safe to use to predict the year 3000
- You can't predict outside the interval

