

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

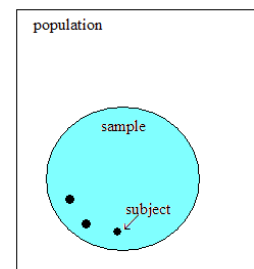
- **Discrete**
 - A countable number of whole-numbered values (no decimals).
 - # of people entering a shop per hour (whole number)
 - # of living grandparents (0,1,2,3,4 only)
 - # of spades in a poker hand (0,1,2,3,4,5 only)
 - # of balls a juggler is currently juggling
- **Continuous**
 - Can take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

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)
- Region of U.S. (Southeast, West, etc.)
- Literary genre
- Number of times in one month the Creswell fire alarm goes off

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

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 =
- Sample =
- Subject =
- Parameter =
- Statistic =

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

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

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

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

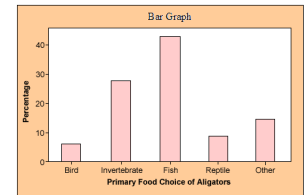
Frequencies (HW 2.1-2.2)

- 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
- Percentage

Types of Charts (Categorical)

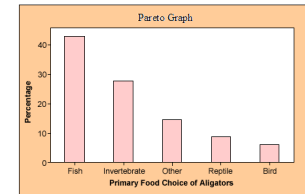
Bar Graph

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



Pareto Graph

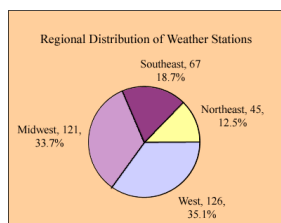
- A bar graph arranged with bars in descending order of frequency



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



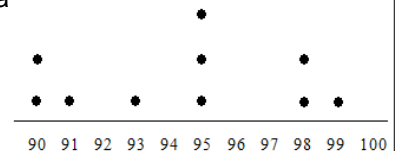
Types of Charts (Quantitative)

Dot Plot

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

A's on a Test

90 90 91 93 95
95 95 98 98 99



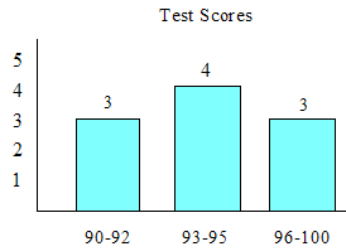
Types of Charts (Quantitative)

Histogram

- A bar graph for quantitative data

A's on a Test

90 90 91 93 95
95 95 98 98 99

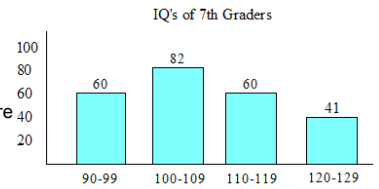


Histogram Interpretation (HW 2.2)

- How many total students sampled?

- Which class has highest / lowest frequency? What are those frequencies?

- How many students have an IQ between 100 and 119?



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

Example (HW 2.1-2.2)

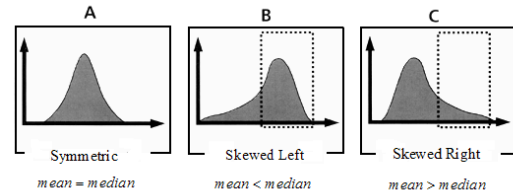
eBay selling prices

199 210 210 223 225
225 225 228 232 235

```

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

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 residents at a retirement home (skewed left since some, but very few, young people live there)

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 cups of tea a British person drinks per day 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?

Cups of Tea	Frequency
1	30
2	102
3	36
4	32
Total	200

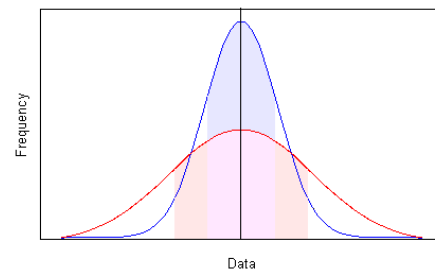
Standard Deviation

- The average distance between any data point and the mean of the data.
- Measures how much/little the data distribution is spread out.

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

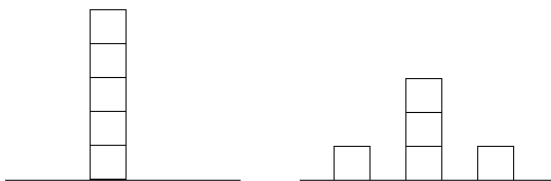
Standard Deviation

Which has larger and smaller st. dev.?



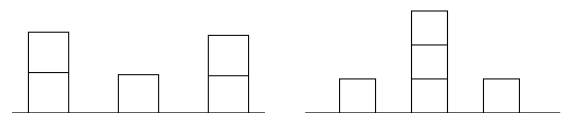
Standard Deviation

Which has larger and smaller st. dev.?



Standard Deviation

Which has larger and smaller st. dev.?



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

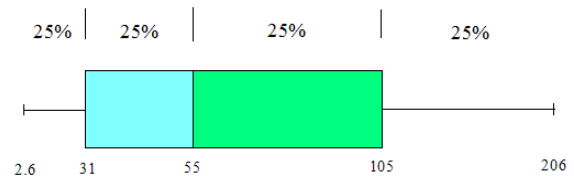
Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
var1	15	7.0933332	5.6920953	2.3858113	0.6160138	6.9	7.5	3.7	11.2	4.7	9.2

- 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) ?



- Between what two values are the middle 50% of the data found?
- Find and interpret the interquartile range.

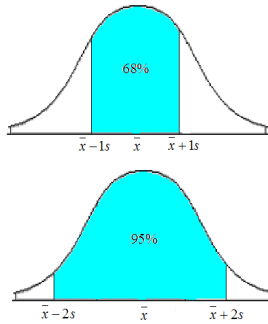
New Box-Plot (HW 2.5-2.6)

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.

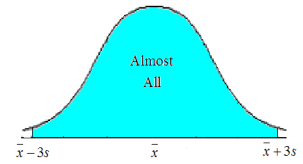
Empirical Rule

- **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.

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?
- Find the weight of a zebra that is three standard deviations above the mean. Would this be an unusual observation?

Z-Score

$$Z = \frac{x - \bar{x}}{s} = \frac{(\text{data point}) - (\text{mean})}{(\text{st. dev.})}$$

- 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 female 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.
- Interpret the Z-score.

Z-Score (HW 2.5-2.6)

- For 261 female 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 2.0 standard deviations above the mean?
- Find the height corresponding to the above Z-score. (Hint: height is a data point, which is x)

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 x th 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?

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)

- Study between gender and views on fighting terrorism
- Response?
- Explanatory?

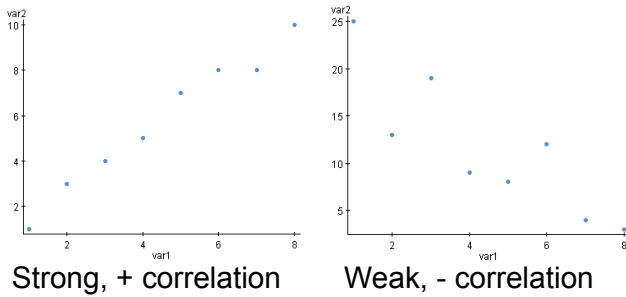
	Good Adjustment	Bad Adjustment	Total
Orientation	72	14	86
No Orientation	28	45	73
Total	100	59	159

- 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
- 45 / 159 did not do orientation and also adjusted poorly
- 72 / 100 is the proportion of students adjusting well to college that did orientation (conditional)

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

- Find the proportion of students that did not do orientation.
- Find the proportion of “orientation-students” that adjusted well.
- Find the proportion of “no-orientation-students” that adjusted well.
- Does there appear to be an association between taking/not taking orientation and adjustment to college?

Scatter Plots



Correlation (r)

- $-1 \leq r \leq 1$
- If r is positive, then so is the slope
- If r is negative, then so is the slope
- 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

Correlation (HW 3.2-3.3)

- Which of the following has the strongest and weakest correlation?

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

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?
- Interpret the slope and intercept.
- Is correlation positive or negative?

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.

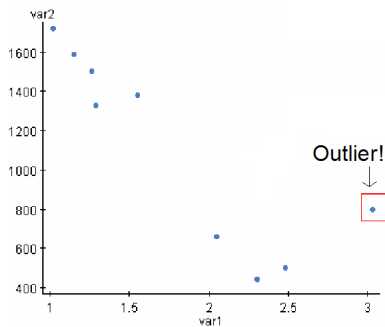
Regression (HW 3.2-3.4)

- A shop owner wants to assign a new price for dog biscuit packets. He is curious how the price per packet (x , in dollars) affects the number sold per day (y). He studies previous years' data and gets: $\hat{y} = 98 - 18x$
- Interpret the slope.
- Interpret the intercept.

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):
 1. For every additional 100 pages, the predicted number of misprints goes up by 5.1.
 2. The number of misprints in a novel 0 pages long is about 5.1.
 3. The intercept has no practical interpretation.
- Interpret the slope (choose the best answer):
 1. For every additional 3.2 pages, the predicted number of misprints goes up by 1.
 2. A novel 400 pages long can be expected to have 3.2 more misprints than a novel 300 pages long.
 3. The slope has no practical interpretation.

Spotting an Outlier



Regression Output

StatCrunch Output:

$\text{var2} = 2206.1917 - 615.97797\text{var1}$

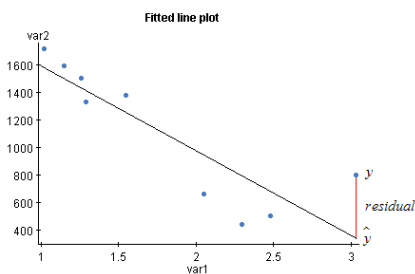
Sample size: 9

R (correlation coefficient) = -0.8648

R-sq = 0.74791104

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

Residuals



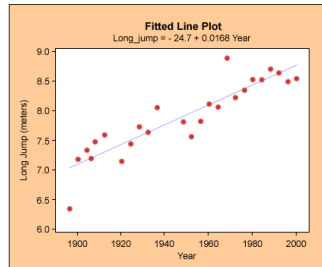
$$\begin{aligned} \text{residual} &= \text{observed} - \text{predicted} \\ &= y - \hat{y} \end{aligned}$$

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.
- 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?

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



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