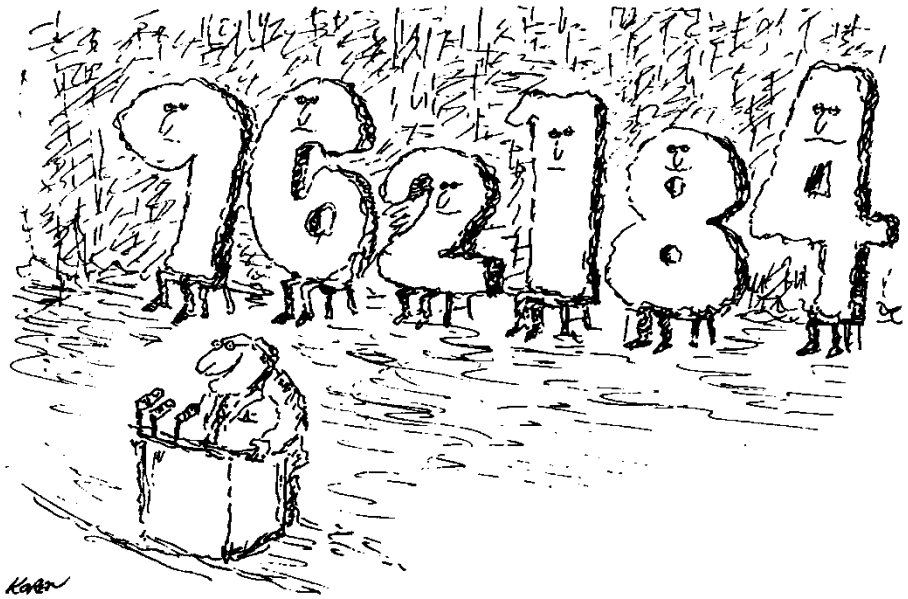
Statistics Review



"Tonight, we're going to let the statistics speak for themselves."

TJHSST

Summer Semester

Review & Self-Study for Research

Stat 1 KEY

Discussion 1: Measures of Central Tendency

Fourth graders weight data:

Using the previous results from above, answer the following questions:

1. What is the mean weight of these fourth graders? **67.68**

   What is the median weight? **68** What is the mode? **60**

2. Which do you think is the most representative number for these weights, the mean, median or mode? Explain.

   The mean takes into account every single student so it is very representative. In addition, the median is very close to the mean so it confirms that the mean is a good representative number. The mode is so far from the mean and median that it is not a good representative number.

Exercises for Discussion 1

1. If you wanted to estimate the total amount spent on junk food for a week by your class, would you prefer to know the daily mean, median or mode amount spent by the class on one day? Explain.

   MEAN. You could multiply the mean by the number of students to find the total.

2. If you wanted to know if you read more or fewer books per month than most people in the class, would you prefer to know the mean, median or mode? Explain.

   MEDIAN. If you read more books than the median number, then you know your read more than 50% of the class.

3. The Reston Town Center skating rink is ordering new skates. Which would be more useful to know, the mode, mean or median skate size? Explain.

   MODE. We need to know the skate sizes that are most frequently used.

4. You want to know which Virginia county has a large portion of people with low incomes. Which is most helpful to know for each county: the mean, mode or median income? Explain.

   MEDIAN. This would tell you what income half the residents are below. The mean could be skewed by a few very wealthy peoples.

5. (Taken from Statistics and Information Organization: Math Resource Program by University of Oregon) A manufacturing company boasts that they pay an average salary of $30,000 to their employees. Study the chart below and answer the following questions:

   a) Is the company telling the truth? To help you decide, find each of the following:

   mean salary **$26,440** median salary **$22,000** mode salary **$18,000**

   The company is not telling the truth. All measure of central tendency, even the mean which is affected by the "extremely" high salaries, are lower the $30,000

   b) Which do you think is a more representative number for these salaries, the mean, median or mode? Explain.

   The median is the most representative. It is not overly affected by high salaries but recognizes their existence.

Discussion #2: Measures of Spread or Variability

Example 1: For the set S = {1, 3, 4, 6, 7, 9}, the mean is 5. The absolute deviation is just the absolute value of each deviation. The deviations, absolute deviations, and squared deviations for the first four numbers are given. Complete the table and answer questions a-c without a calculator.

x	$x - \mu$	deviation	absolute deviation	squared deviation
1	1 – 5	-4	4	16
3	3 – 5	-2	2	4
4	4 – 5	-1	1	1
6	6 – 5	1	1	1
7	7 – 5	**2**	**2**	**4**
9	9 – 5	**4**	**4**	**16**

a) What is the sum of the deviations? Will this always happen? Why or why not? **The sum of the deviations is 0. It will always be zero, because the mean is the average of the data points and the deviations measure the distance from the mean so the deviations to the left of the mean should balance the deviations on the right.**

b) The mean absolute deviation is calculated by finding the mean of all the absolute deviations. Find the mean absolute deviation (MAD) for the data set in the example above. **2.333**

c) Calculate the variance and standard deviation. **variance = 7, standard deviation = 2.646**

Exercises for Discussion 3

1. **Class A**
Mean = 86.286; Median = 87; Range = 93-77=16
IQR=92-82=10

x	$x - \mu$	**Absolute Deviation** $\lvert x - \mu \rvert$	**Squared Deviation** $\lvert x - \mu \rvert^2$
77	**-9.286**	**9.286**	**86.224**
77	**-9.286**	**9.286**	**86.224**
77	**-9.286**	**9.286**	**86.224**
82	**-4.286**	**4.286**	**18.367**
85	**-1.286**	**1.286**	**1.653**
85	**-1.286**	**1.286**	**1.653**
86	**-0.286**	**0.286**	**0.082**
88	**1.714**	**1.714**	**2.939**
90	**3.714**	**3.714**	**13.796**
91	**4.714**	**4.714**	**22.224**
92	**5.714**	**5.714**	**32.653**
92	**5.714**	**5.714**	**32.653**
93	**6.714**	**6.714**	**45.082**
93	**6.714**	**6.714**	**45.082**
	SUM =	**70**	**474.857**
AVE	**SUM/14=**	**5 (MAD)**	**33.917 (VARIANCE)** σ^2
			STANDARD DEVIATION $\sigma = 5.824$

CLASS	B	**C**
mean	86.357	84.429
median	87.5	87.5
range	23	44
IQR	18	17
MAD	7.735	9.816
variance	74.515	159.102
standard deviation	8.632	12.614

b) All three classes had a similar test average. However, class A was the least spread out, with most students performing close to the average. Class C had the most variability, with students performing significantly better and worse than the class average.

Discussion #3: Stem-and-Leaf Plots, Dotplots, and Box –and-Whisker Plots

Exercises for Discussion 3

1. Make a stem-leaf plot of the fat content.
   a) Use this plot to help find the mean, median and mode for fat content. Confirm your answers by using your calculator.
   b) Make a stem-leaf plot of the carbohydrate content.
   c) Use this plot to help find the mean, median and mode of the carbohydrate content.

FAT CONTENT:

1	**1 2 5 8 8 8 8 8**	***n*=15**
2	**0 1 3 3 6 8**	
3	**1**	

mean: 20 g
median: 18 g
mode: 18 g

CARBOHYDRATE CONTENT:

1	**2 5 7 9**	***n*=15**
2	**8 9**	
3	**3 5 4 6 8**	
4	**2 2 6**	
5	**0**	

mean: 31.733 g
median: 34 g
mode: 42 g

d) Are there any outliers for the fat content? Justify your answer using the method above.

$IQR = 42 - 19 = 23$

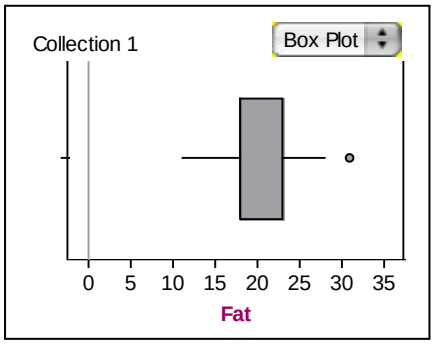
$(1.5)(23) = 34.5$

$Q1 - 34.5 = -18.5$

$Q3 - 34.5 = 76.5$

ok range for outliers: $[-18.5, 76.6]$ **There are no data points below -18.5 nor above 76.5. Therefore there are no outliers**

e) Draw a modified box plot for the fat content of fast foods, indicating any outliers that you find.



Note that for the FAT content there is an outlier.
The "ok" interval is [14.5, 26.5], using the formula for outliers.
Therefore the data entry, 31g, is an outlier.

2. Refer to the plots below.

   a) Which class had the higher median? **The classes had the same median of 3.**

   b) What was the interquartile range for each class?

IQR_{JR} = 3.5-2.5 = 1.0
IQR_{SR} = 3.3-2.6 = 0.7

c) Estimate each of the classes' best and worst grade point averages. Are there any outliers? Explain.
Best JR = 4.0 Worst JR = 0.75
Best SR = 4.0 Worst SR = 1.5

JR outlier interval [1.0, 5.0] Therefore 0.75 is an outlier
SR outlier interval [1.55, 4.35] Therefore there are no outliers.

Discussion 4: Quantitative Data, Frequency Tables, and Histograms

Exercises for Discussion 5

1. For the following examples, determine whether the data is continuous or discrete:
   a) Population in Fairfax County, Virginia **DISCRETE**
   b) Weight of newspapers collected for recycling on a single day at TJ **CONTINUOUS**
   c) Score on a math test **DISCRETE**
   d) GPA **CONTINUOUS**

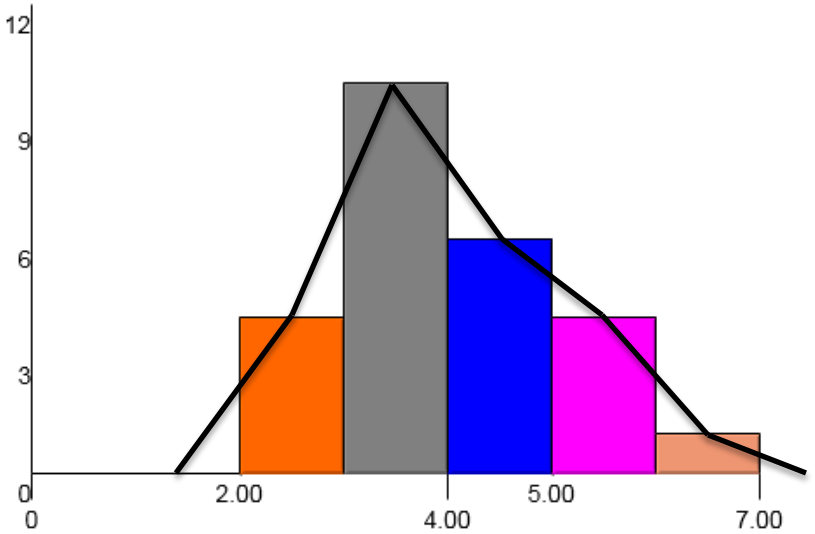
2. For the following set of continuous data, determine an appropriate number of classes and set class limits. Then, set up a frequency table to organize the data.

Low: 2.4
High: 6.2

Answers will vary.

Interval	Class Limits	Tally	Frequency	Relative Frequency
1	2.0-< 3.0	|||	3	3/25
2	3.0-<4.0	|||| ||||	10	10/25
3	4.0-<5.0	|||| |	6	6/25
4	5.0-<6.0	||||	5	5/25
5	6.0-<7.0	|	1	1/25

Using the data from the frequency table, design a histogram and then a frequency polygon of the magnetite data.



Magnetite Data

What shape does the frequency polygon appear to have? Explain. **The frequency polygon is skewed slightly to the right.**

Discussion 5: The Scatter Plot

Exercises for Discussion 5:

1. Soft Drinks. The following is a plot over time showing how many 12 ounce soft drinks the average person in the U. S. drank each year from 1945 to 1980.

   a) About how many soft drinks did the average person drink in 1950? **about 105** in 1970? **about 220**

   b) About how many six-packs of soft drinks did the average person drink in 1980? **410/6, approx 68 six packs**

   c) About how many soft drinks did the average person drink per week in 1950? **105/52, approx 2**
   in 1980? **410/52, approx 8**

   d) If the trend in the plot continued, about how many 12 ounce soft drinks did the average person drink in the year 2000? **Trend: in 10 years, about 150 more. In 200, about 700.**

   e) In what year did soft drink consumption start to "take off"? Can you think of any possible reason for this phenomenon? **Around 1961. Diet sodas were introduced and aluminum cans were starting to be used.**

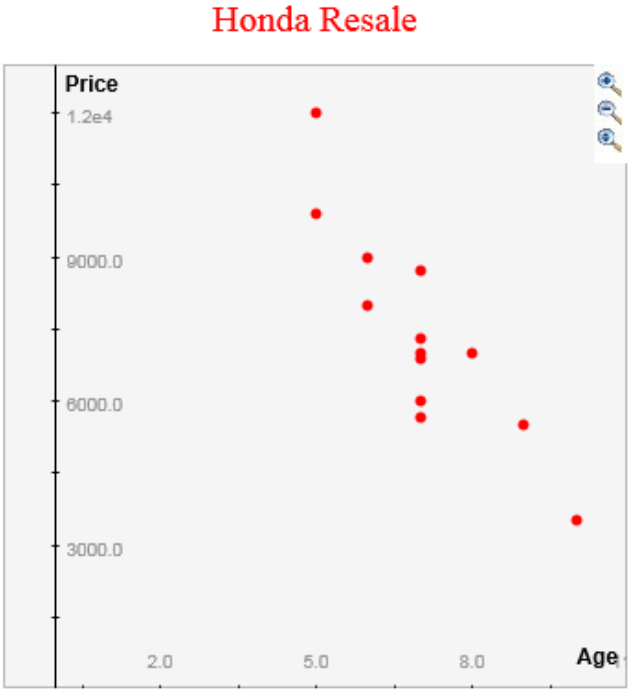
Discussion 6: Correlation

Exercises for Discussion 10:

1. Honda Resale. The following data indicates the resale value of 13 Honda Accords and the age of the car at the time of resale.

x: Age	*y*: Price
5	11995
5	9900
6	9000
6	8000
7	6900
7	6000
7	8700
7	5650
7	7300
7	7000
8	6999
9	5500
10	3500

   a) Construct the scatter plot for the data. Sketch it on your paper. Indicate the scale and labels.



   b)

c) Before finding the correlation coefficient, predict what you think *r* will be and explain why. **r should be close to -1 because there is a fairly strong negative correlation.**

d) Describe the relationship between the two variables. **Age and price are related in a negative linear relationship.**

e) What is the linear regression equation? $y = -1345.458x + 16836.978$

f) Find the correlation coefficient, *r*. Is there a high linear correlation between these two variables? Explain. **r = -0.87777. This is a fairly strong negative linear relationship.**

Discussion 7: The Normal Distribution

Exercises for Discussion 7:

1. The SAT's are designed so that the distribution of scores would appear as shown:
   a) What is the mean score? $\mu = 500$

   b) What is the standard deviation of the scores? $\sigma = 100$

   c) What percentage of scores were between
      I. 500 and 600 **34%**
      II. 500 and 700 **47.5%**
      III. 500 and 800 **49.9%**

   d) Some colleges do not admit applicants whose scores are less than 600. According to this distribution, what percentage of students would be expected to have scores of 600 or more?

      > 600 84% of scores
      < 600 16% of scores

   e) The most competitive colleges require scores over 700. What percentage of students would be considered by these colleges? **2.5% of scores**

For each of the following problems apply the empirical rule to answer the questions. **Draw a curve for each problem.**

2. A ketchup company has fixed the weight of a bottle at 16 oz., with a standard deviation of 0.5 oz. The curve depicting the weights is bell-shaped. Approximately what percentage of bottles will be:
   a) greater than 15 oz.? **97.5%**
   b) greater than 17 oz.? **2.5%**
   c) less than 14 oz.? **< 0.1%**
   d) less than 13 oz.? **< 0.1%**
   e) between 15 and 17 oz.? **95%**

4. Quarters

   a) Which curve do you think shows the weights of the newly minted quarters--**A**
   which curve the coins after five years--**B**
   which curve after ten years--**C**

   b) What happens to the average weight of the coins as time passes? **Mean weight decreases over time**

   c) What happens to the standard deviation of the weight of the coins as time passes? **Standard deviation increases over time**

Discussion 8: Standard Scores or *z* Scores

Exercises for Discussion 8:

1. Transfer students to a new high school are sometimes given a standardized test with a mean of 100 and a standard deviation of 20. To three decimal places, convert the raw scores of the following students to z scores:
   Alice--105 Bob--72 Carol--142 David--133 Elliott—95

   $Z_{Alice} = 0.25$ $Z_{Bob} = -1.4$ $Z_{Carol} = 2.1$ $Z_{David} = 1.65$ $Z_{Elliott} = -0.25$

2. John weighs 220 pounds; his dog Fido weights 90 pounds. If human males weigh an average of 160 pounds with a standard deviation of 20 pounds, and all dogs of Fido's breed have an average weight of 80 pounds with a standard deviation of 5 pounds, how do John and Fido compare, relative to their populations, with respect to weight?

   $Z_{John} = 0.25$ $Z_{Fido} = 0.25$ John is more overweight than his dog.

Discussion 9: The Standard Normal or *z* Distribution

Exercises for Discussion 9:

1. For a standard normal distribution, find z if
   a) $P(Z < z) = 0.0668$ **z = invNorm(0.0668) = -1.5**
   b) $P(Z > z) = 0.9861$ **z = invNorm(1-.9861) = -2.2**

2. Use your calculator to find the following probabilities, assuming a normal distribution.
   a) $P(X < 3.5)$ when $\mu = 5$, $\sigma = 1$. **P(z<-1.5) = 0.0668**
   b) $P(X > 130)$ when $\mu = 110$, $\sigma = 25$ **P(z>0.8) = 0.212**
   c) $P(14.2 < X < 15.0)$ when $\mu = 14$, $\sigma = 0.5$. **P(.4<z<2) = 0.322**

3. One part of a test administered to adults is an exercise in manual dexterity. The average time for the test is 165 seconds, with a standard deviation of 21 seconds. Assume the relative frequency distribution of the times needed to complete the test is approximately normal. What proportion of the adult population can complete the test in
   a) more than 190 seconds **P(x>190) = P(z>1.190) = 0.117**
   b) between 140 and 160 seconds **P(140<x<160) = P(-1.1905<z<-0.238) = 0.289**

4. If the mean of a normal distribution is 10 feet and the standard deviation is 2 feet, for what values of *y* is it true that $P(Y < y) = 0.025$? **z = invNorm(0.025) = -1.9599 = (y – 10)/2 so y = 6.080**

5. If the mean of a normal distribution is 20.5 inches and the standard deviation is 0.2 inches, find *x* such that $P(X > x) = 0.7995$. **z = invNorm(1-.7995) = -0.8398 = (x – 20.5)/.2 so x = 20.332**

6. If *X* is a continuous random variable that can be modeled as normal with a mean of 100 cm and a standard deviation of 10 cm, find *x* so that
   a) $P(X > x) = 0.0049$ **z = invNorm(1-.0049) = 2.583 = (x – 100)/10 so x = 125.828**
   b) $P(X < x) = 0.9332$ **z = invNorm(.9332 = 1.5 = (x – 100)/10 so x = 115**