AP Statistics Summer Assignment

Part 1: ***The Joy of Stats*** with Professor Hans Rosling

The following videos can be found on the website gapminder.org. The creator, Professor Hans Rosling, states on the site: "Gapminder is a non-profit foundation based in Stockholm. Our goal is to replace devastating myths with a fact-based worldview. Our method is to make data easy to understand…" Rosling uses innovative methods to present and display data and statistics.

Your assignment is to watch the video and **answer the questions below**.
(Feel free to watch more of the videos, most are short and interesting.)

Go to the website http://www.gapminder.org/videos/ and watch the video *The Joy of Stats*

1) What was the mean number of correct answers given by Swedish students to the questions:
'Which country has the highest child mortality rate?'

2) In displaying data on global health, what does the size of a country's bubble indicate?

3) What are two benefits of public statistics?

4) The word 'statistics' is derived from what word?

5) The first systematic collection of statistics is what document?

6) Along with average, when reporting data, what other value is important?

7) Which distribution models the number of buses that appear in a given hour?

8) Who used a polar area graph to display data?

9) What analytical method explores meaning and relationships within data?

10) What is a zettabyte?

11) Other than the internet, identify 3 technologies used to gather massive amounts of data.

12) How many bananas were eaten worldwide in the time it took to watch the video?

Part 2: An introduction to and review of the first 3 chapters of the Statistics course.

Use a separate sheet of paper for responses.

CHAPTER 1: EXPLORING DATA

FOR ALL DISPLAYS, LABEL THE AXES AND TITLE THE GRAPH.

When asked to **describe a distribution** of data, look at a display (graph) and give the **overall pattern**, which includes the following 4 items:

Center: mean or median (often we report the value that divides the data into 2 equally sized groups, meaning half the values are smaller and half are larger.)

Spread: Described as: "from (smallest value) to (largest value)."

Shape: key words to consider: symmetrical, skewed, single peaked, many peaked, etc.

Deviations: An individual observation that deviates from the overall pattern of the graph is an outlier (or extreme value). Use best judgement to determine whether outliers are present.

Display 1: STEMPLOT (or stem-and-leaf plot)

Consider the following when making a stemplot:

-Each stem should have an equal number of possible leaves (equal intervals)

-Don't want too few stems (values are clustered) or too many stems (values are too spread out). 5 stems is a good minimum.

Example 1: The values below indicate the number of home runs hit by Babe Ruth, Hank Aaron, and Barry Bonds for the first 22 seasons.

Ruth

0	4	3	2	11	29	54	59	35	41	46
25	47	60	54	46	49	46	41	34	22	6

Aaron

13	27	26	44	30	39	40	34	45	44	24
32	44	39	29	44	38	47	34	40	20	12

Bonds

16	25	24	19	33	25	34	46	37	33	42
40	37	34	49	73	46	45	45	5	26	28

a) Construct a stemplot for the number of home runs hit by each player.

b) Describe the **overall pattern** of each stemplot (Center, spread, shape) and deviations (outliers).

c) **Centers:** Find the **mean** number of home runs hit in a year for each player.
Find the **median** number of home runs for the first 22 seasons for each player.

d) Eliminate the highest number of home runs for each player. Find the mean for each player for the remaining 21 seasons. **Compare** the change in mean among the three players.

e) Eliminating the highest number of home runs for each player, find the median for each player for the remaining 21 seasons. **Compare** the change in median among the three players.

f) Shapes: Which player has the most symmetric distribution? Which player's distribution seems to be the most skewed?

g) Compare centers: For which player are the mean and median for all 22 seasons the closest? For which player are the mean and median for all 22 seasons the farthest apart?

Display 2: HISTOGRAM

Example 2: States differ widely with respect to the percentage of college students who are enrolled in public institutions. The U.S. Department of Education provided the accompanying data on this percentage for the 50 U.S. states for fall 1999.

Percentage of College Students Enrolled in Public Institutions

95	81	85	72	73	74	79	84	89	63
91	86	89	92	87	90	83	84	89	96
87	85	76	84	80	95	75	81	73	82
81	77	75	70	55	56	87	88	82	81
84	76	80	56	55	43	52	62	80	82

a) Construct a *frequency table* (choose equally sized intervals such that you have at least 5 intervals)

Intervals	Frequency

b) Display the information in a histogram.

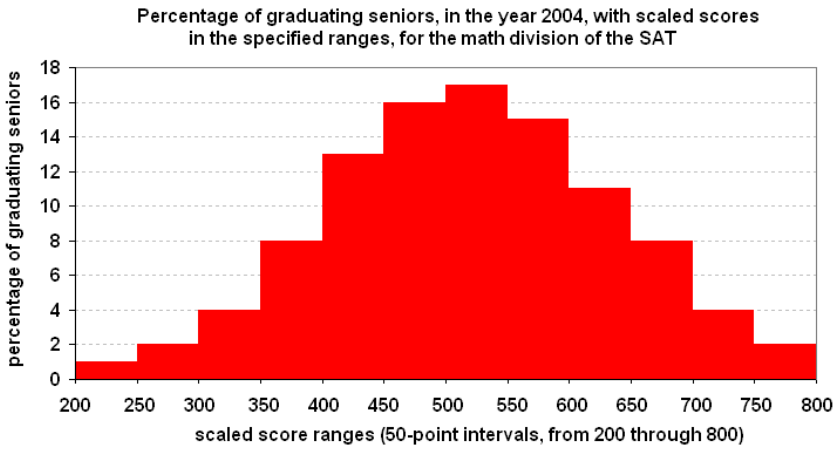
c) Describe the distribution. Include all 4 criteria as listed earlier.

CHAPTER 2: MODELING DISTRIBUTIONS OF DATA (Density Curves and the Normal Distribution)

Sometimes a smooth curve can be used to model and describe the **overall pattern** in a display. This smooth curve is called a **density curve**.

Example 3:

a) Draw a smooth curve that gives the overall pattern of the distribution in the graph of Math SAT scores.



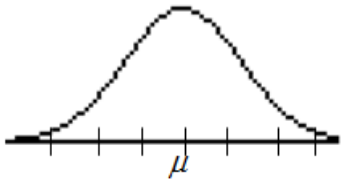
b) Answer the following questions:
(Use the Example: What *proportion* of students scored below a 400? Answer: Total the frequencies of the bars in that range (200-400): 1% + 2% + 4% + 8% = 15% of students)

What proportion of students scored:

i) between 400 and 600? ii) between 550 and 700? iii) 700 or higher?

Normal Curve

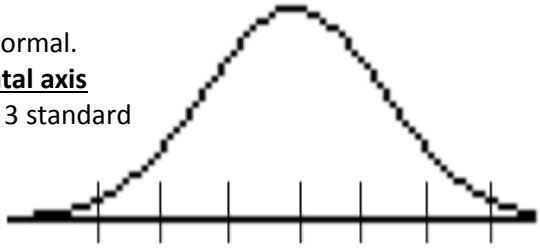
A normal curve is a density curve that is **symmetric** and **bell shaped** (single peaked).
A normal curve is described by its mean and standard deviation (S.D.)



Mean: determines the location of the curve, specifically of the peak.
Notation: μ (Greek letter mu)

S.D.: determines the spread of the curve. There are approximately **3 standard deviations** in each direction above and below the mean. Notation: σ (Greek letter sigma)
(Each tick mark on the horizontal axis is 1 standard deviation)

Example 4: The distribution of scores of an IQ (intelligence quotient) test is Normal. The mean score is 100 with standard deviation 15. **Label the horizontal axis of the normal curve** that represents IQ with values for the mean and 3 standard deviations in each direction from the mean.



Empirical Rule (68 -95 – 99.7 Rule): All Normal distributions obey a common rule:

68% of the observations fall within 1 standard deviation of the mean (in the interval $\mu \pm \sigma$)
95% of the observations fall within 2 standard deviations of the mean (in the interval $\mu \pm 2\sigma$)
99.7% of the observations fall within 3 standard deviations of the mean (in the interval $\mu \pm 3\sigma$)

Example 4: A different IQ test has scores that are normally distributed with $\mu = 100$ and $\sigma = 16$.

a) Draw and label the horizontal axis of a normal curve.

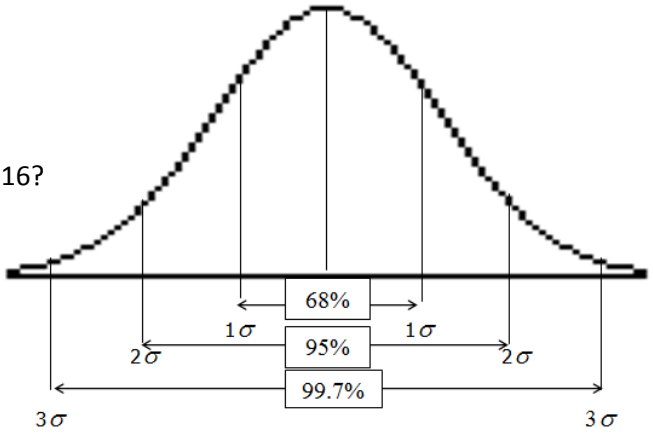
b) What proportion of the population has scores between 84 and 116?

c) What proportion has scores between 100 and 132?

d) What proportion has scores between 68 and 84?

e) What proportion has scores above 116?

f) What proportion has scores below 52?



Example 5: A normal distribution of Introduction to Biology Final exam scores at a large university has a mean of 77 and a standard deviation of 7.

a) Sketch and label a normal curve, showing the mean and 3 standard deviations in each direction.

b) Use the 68-95-99.7 rule to find the proportion of exam scores that are in the given interval:

i) Between 70 and 84

ii) Between 63 and 77

iii) Between 70 and 91

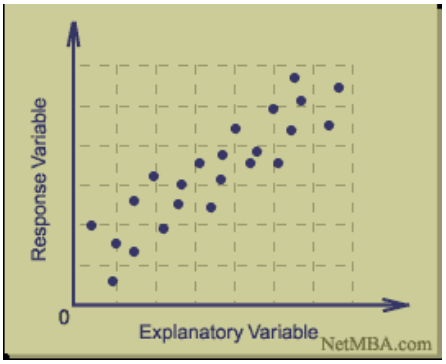
iv) More than 84

v) Less than 63

vi) At least 91

CHAPTER 3: DESCRIBING RELATIONSHIPS (Comparing 2 Quantitative Variables)

A **scatterplot** is sued to display the relationship between two ***quantitative variables*** measured on the same individuals. Each individual in the data appears as a point in the plot.



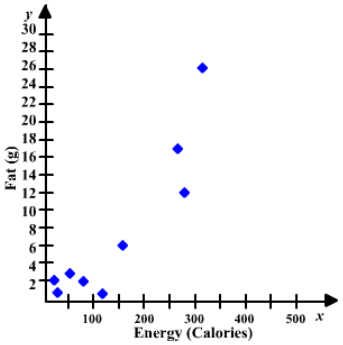
Explanatory variable on x-axis
Response variable on y-axis

Example 6:
The table and graphs display the amount of fat, amount of fiber, and the number of calories in a 4-oz serving of vegetables.

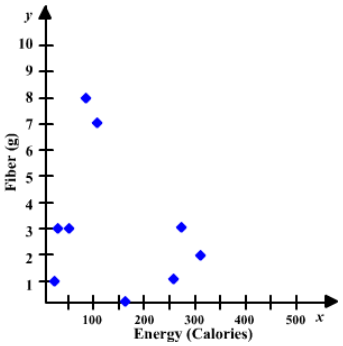
Nutritional Values for 4oz servings

Vegetable Item	Fat (grams)	Fiber (grams)	Energy (calories)
Martian Cabbage	3	3	52
Butter Casserole	26	2	328
Mauve Beans	2	8	82
Broccoli Tofu Casserole	12	3	276
Jalapeno Corn Husks	17	1	261
Rancid Eggplant Curry	1	7	123
Mixed Tuber Salad	2	1	21
Four-times-baked Potato	6	0	158
Steamed Mush	1	3	35

A) Comparing amount of ***FAT*** with the Number of calories



B) Comparing amount of ***FIBER*** with the number of calories



a) What are the explanatory variable and response variable in graph (A)?

b) What are the explanatory variable and response variable in graph (B)?

c) Which graph seems to display a *stronger relationship* between the explanatory and response variables? Justify your choice with an explanation.

When asked to *describe* the relationship between an explanatory and a response variable, explain:

1) The **Form** of the association: linear or non-linear.

2) The **Direction** of the association: Positive (increasing) or negative (decreasing).

3) The **Strength** of the association (use 1 or 2 words) weak, moderate, strong, or a combination of these (ex: moderately strong)

d) Describe the relationship between the explanatory and response variables in graph (A).

e) Describe the relationship between the explanatory and response variables in graph (B).

Example 7: ***SCATTERPLOT***

a) Construct a scatterplot showing the relationship among SAT Math and Verbal scores for the students in a Calculus class. Label axes appropriately.
Let the math score be the explanatory variable (x-axis).
Let the verbal score be the response variable (y-axis).

Math score	**Verbal score**	**Math score**	**Verbal score**
660	520	740	670
590	600	790	570
800	590	710	580
680	600	670	600
660	570	690	580
620	670	690	570
650	590	800	570
600	560	710	560
660	660	740	630
760	610		

b) Describe the scatterplot. (i.e. describe the relationship between the two variables.)