9-22-15

Exam 2 Material Starts

BOX PLOTS

Just like histograms, box plots are used as a way to visually represent numerical data. They do this through selected percentiles which are given special names. ***Quartiles*** divide a data set into quarters:

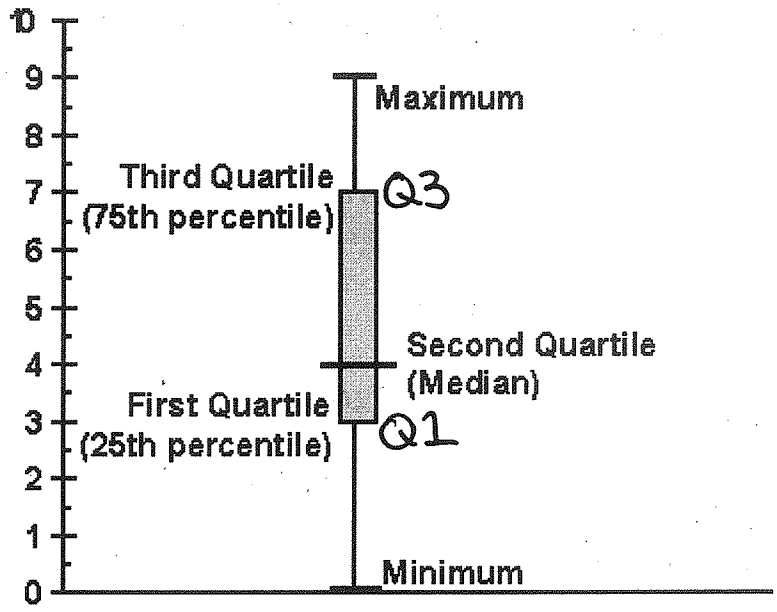
Q1- 1st quartile	=	25th percentile
Q2- 2nd quartile	=	50th percentile - median
Q3- 3rd quartile	=	75th percentile

The second quartile is also the median.

When comparing the spread of various data sets, the ***interquartile range*** **(IQR)** is often used as an alternative to the standard deviation. The interquartile range is the range of the middle 50% of the data:

$$IQR = 75^{th} \text{ percentile} - 25^{th} \text{ percentile } (Q3-Q1)$$

Box plots have lines extending from them to indicate the variability outside Q1 and Q3 called whiskers, so a lot of times they are called box and whisker plots. Outliers may also be plotted as individual points. Here is a simple boxplot with no outliers:



The bottom of the box is always Q1, the top of the box is always Q3, and the band inside the box is always the median (Q2).

Example 1: Draw a boxplot of this data: Listed below are the final exam scores of 12 random students from last semester: 60, 68, 75, 85, 86, 87, 88, 90, 92, 92, 95, 100

outlier → ↑ 87.5 = median

Step 1: Find the median, Q1, Q3, and the IQR. We've already discussed how to find the median of a list of numbers. To find Q1 & Q3, find the median of the lower half & upper half of the data.

Median (Q2)= 2 middle #'s (87 + 88) median = $\frac{87+88}{2}$ = 87.5

Q1= (25th percentile) 60, 68, 75, 85, 86, 87 Q1 = $\frac{75+85}{2}$ = 80

Q3= (75th percentile) 88, 90, 92, 92, 95, 100 Q3 = 92

IQR= Q3 - Q1 = 92 - 80 = 12

Step 2: Check for outliers. The ends of the whiskers represent different values depending on if the data has outliers or not. Outliers are data points that "lie outside" most of the other values in the data set. You can calculate outliers mathematically like this:

*Low outliers < Q1 - 1.5*IQR* → Q1 - 1.5(IQR) = 80 - 1.5(12) = 62

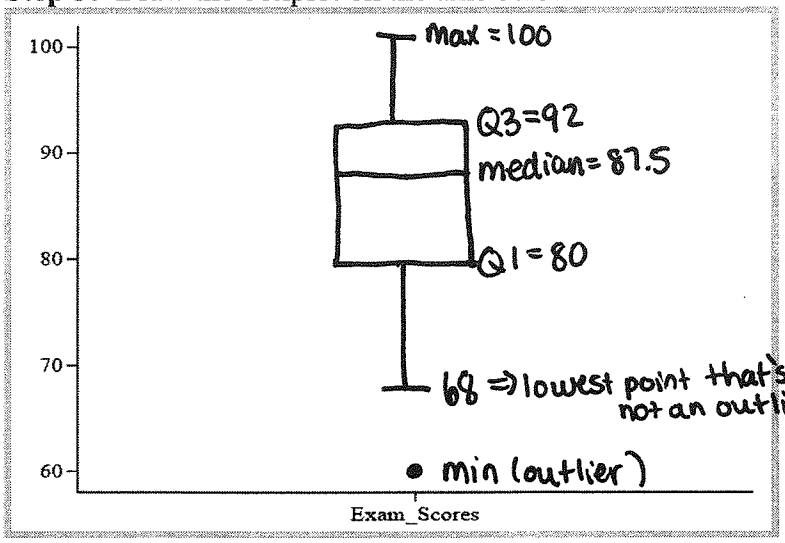
*High outliers > Q3 + 1.5*IQR* → Q3 + 1.5(IQR) = 92 + 1.5(12) = 110

If the dataset has no outliers, the ends of the whiskers represent the minimum and maximum values of the dataset (like on the previous page).

✳ If the dataset does have outliers, the outliers are plotted as single dots & the ends of the whiskers represent the highest and lowest data points that are not outliers (within 1.5*IQR of Q1 and Q3).

Does the dataset of exam scores have any outliers? Yes ⇒ 60

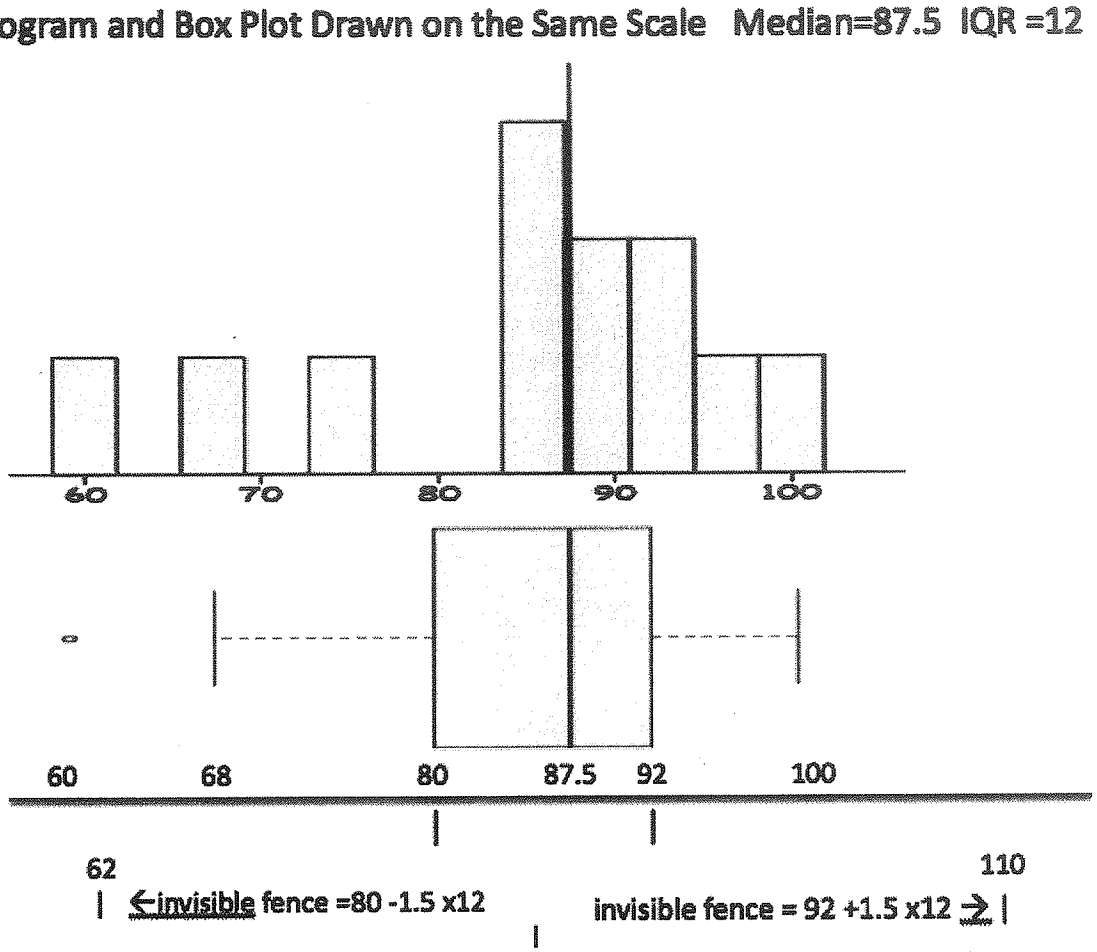
Step 3: Draw the boxplot on the axis below:



example:
1, 2, 3, (4), 5, 6, 7
↓ median

* If there's an odd # of numbers, don't include the median in your calculations for Q1 + Q3.

Here, we can see the histogram of exam scores alongside the boxplots. Notice that they are both drawn with the same scale.



Example 2: Here are two side by side boxplots the amount of time it takes men and women to get ready in the morning. Answer the following questions below with men, women, or both.

Which group has a bigger range of times? F

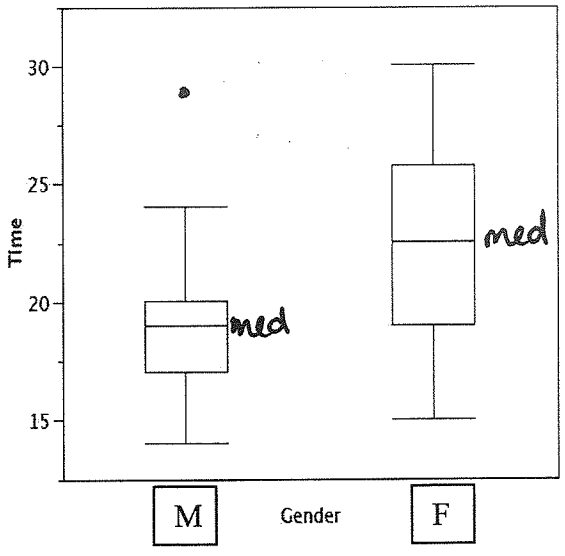
Which group has the smallest time? M

Which group has the largest time? F

Which group has outliers? M

Which group has a smaller IQR? M
Q3-Q1

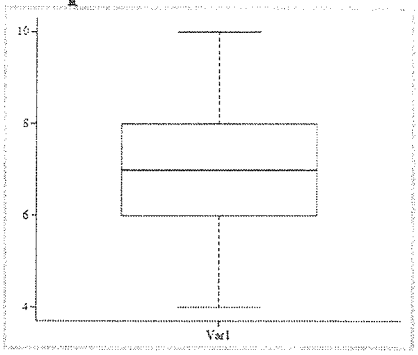
Which group has a larger median? F



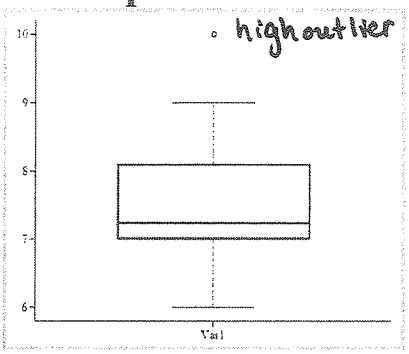
Notice that the statistics that are easy to read on the boxplots are the median and IQR, not the average and the SD.

Example 3: Matching Histograms and Boxplots:
Below are 3 different boxplots (A, B, and C). Match them to the correct histogram.

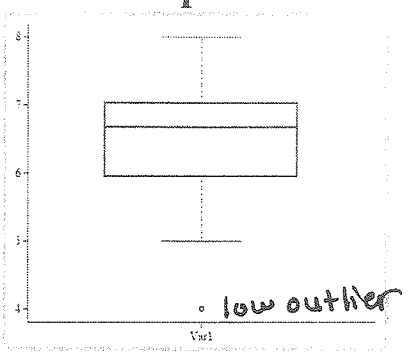
Boxplot A



Boxplot B

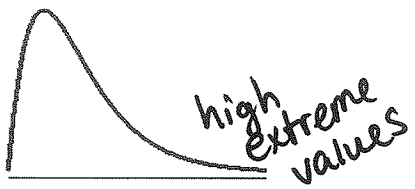


Boxplot C



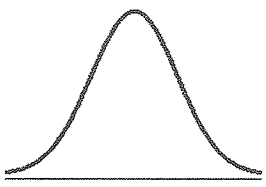
Boxplot B

Long Right-Hand Tail



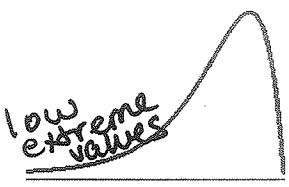
Boxplot A

Symmetric Distribution



Boxplot C

Long Left-Hand Tail



Remember from Chapter 4: p. 35

Fill in the blanks with >, <, *or* =:

If the histogram is symmetrical then: (boxplot) average = median

If the histogram has a long right-hand tail (extreme large values) then: (boxplot) average > median

If the histogram has a long left-hand tail (extreme small values) then: (boxplot) average < median

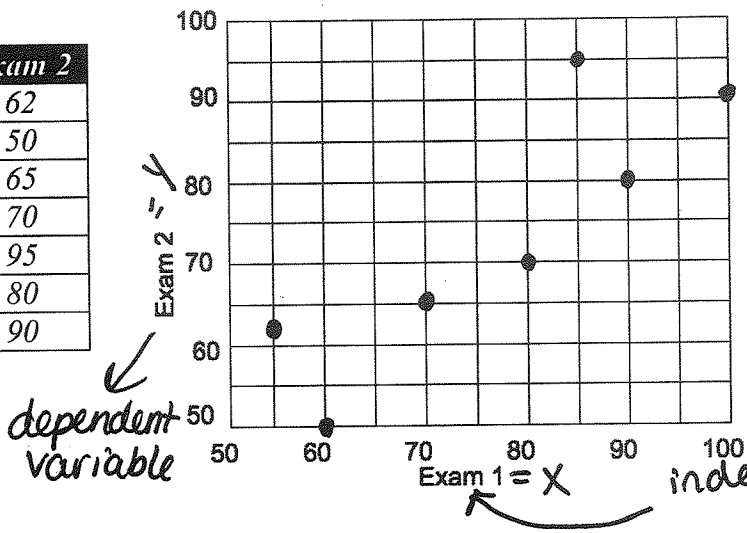
***The same rules apply to the matching boxplots.**

PART III—LINEAR REGRESSION
CHAPTER 7 — CORRELATION

SCATTER PLOT-- A graph used for showing the relationship between 2 variables. One variable is assigned to the x-axis and the other is assigned to the y-axis. The convention is to call the x variable the *independent* variable and the y variable the *dependent variable.* Usually the independent variable is thought to influence the dependent variable.

Example 1: Construct a scatter plot for the Exam 1 and Exam 2 scores of 7 students:

Exam 1	*Exam 2*
55	*62*
60	*50*
70	*65*
80	*70*
85	*95*
90	*80*
100	*90*



There is a **positive association** between x and y when the pattern of the points slopes upward.
As x increases, y tends to increase. makes sense w/ exam scores

There is a **negative association** between x and y when the pattern of the points slopes downward.
As x increases, y tends to decrease.

Is there a positive or negative association between exam 1 and exam 2 in the example above?

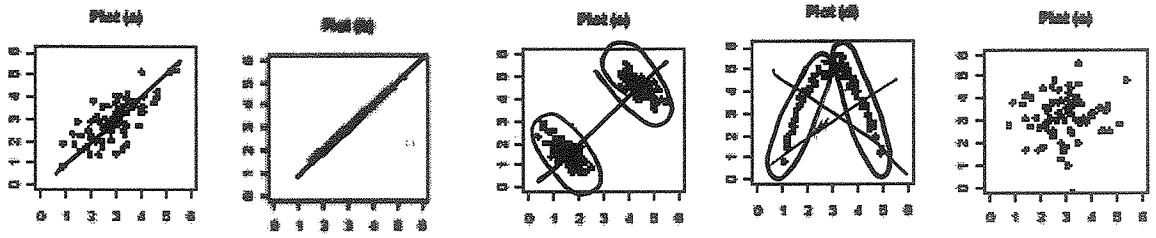
positive

For the following pairs of variables state whether you think the association is positive, negative, or neither.

- Height and weight positive
- Weight of a car and how many miles per gallon it gets negative
- Years of Education and Income positive
- Height and GPA among college students neither ($r=0$)
- Temperature in Fahrenheit and temperature in Celsius. positive ($r=+1$)
- # right and # wrong on a test negative ($r=-1$)

- **CORRELATION COEFFICIENT (*r*)**- measures the strength of the linear association between X and Y. It measures how tightly points are clustered around a line. (It does not measure clustering around a curve). It is relevant when the scatter plot forms a "football-shaped" cloud. How closely the points hug a line

Example 2: Is r appropriate summary statistic for the plots below? If not, explain why not.



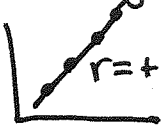
Yes Yes No (2 groups) No (curve) Yes

r would correctly tell us there is no pattern or relationship

The correlation coefficient is always between –1 and 1.
The closer the points hug a line with a positive slope, the closer *r* is to +1. The closer the points hug a line with a negative slope the closer *r* is to -1. If there is no association between *x* and *y* then the correlation coefficient is 0 and the scatter plot has a basically round pattern.

A correlation of 1 or -1 means you can perfectly predict one variable knowing the other. Given x, I can tell you exactly what y is.

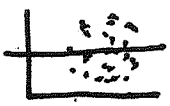
Examples:
- Age and date of birth
- Height in inches and height in centimeters.



all points lie on line

A correlation of 0 means that knowing one variable gives you no information about the other. Knowing x tells you nothing about y

Examples:
- Weight and ACT scores of college freshman.
- Attendance in Stat 100 and Number of Pets Owned



r=0 blob of points

Which of the variables on the previous page have a perfect correlation of 1 or -1? Which have correlations close to 0?

5 Summary Statistics to Describe 2 Variables

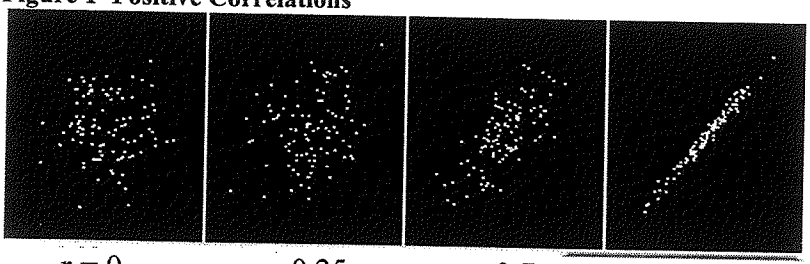
ave of x ave of y
SD of x r SD of y

STATISTICS OF THE "CLOUD" SCATTER PLOT

- Point of averages ($Avex$, $Avey$) is at the center of the cloud
- SD_X measures the horizontal spread of the cloud
- SD_Y measures the vertical spread of the cloud

The above statistics tell us the size and location of the box the cloud is in but nothing about what shape the cloud is. They tell us nothing about the relation between x and y. Look at figures 1 and 2 below. There are 100 points in each of the 8 scatter plots. The average and SD of the x's and y's are the same in all the plots but the correlation coefficients are different.

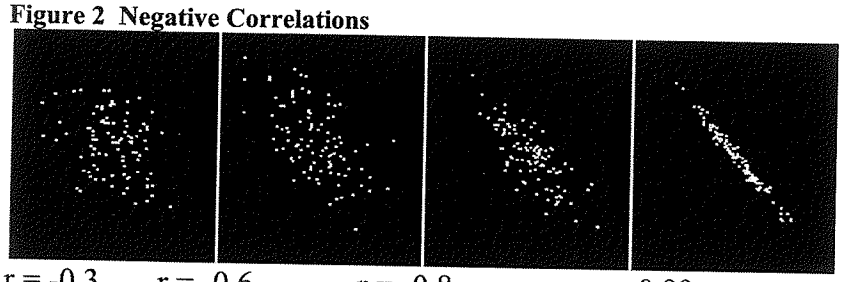
Figure 1 Positive Correlations



r = 0 r = 0.25 r = 0.7 r = 0.99

The closer r gets to +1, the closer the points form a straight line w/ a positive slope

Figure 2 Negative Correlations



r = -0.3 r = -0.6 r = -0.8 r = -0.99

Example 1: **Match the correlations with the plots**
(see http://www.istics.net/Correlations/)