Exam 1 Math 1040 Introduction to Statistics Fall 2011

Name_____

You may use your calculator and the provided page from the textbook formula card. No other notes are permited. Please carefully and completely show your work to receive full credit. This exam is worth 15% of your final grade.

---

1) Google has 322,900,000 stockholders (September 2011) and a poll is conducted by randomly selecting 45 stockholders from each of the 50 states. The number of shares held by each sampled stockholder is recorded.

(a) Are the values obtained **discrete** or **continuous**? (*please circle your answer*)

(b) Identify the level of measurement (**nominal, ordinal, interval, ratio**) for the sample data. (*please circle your answer*)

(c) Which type of sampling (**random, systematic, convenience, stratified, cluster**) is being used? (*please circle your answer*)

(d) If the average (mean) number of shares is calculated from the collected data, is the result a **statistic** or a **parameter**? (*please circle your answer*)

(e) Would you consider this an **observational study** or an **experiment**? *(please circle your answer)*

(f) What type of bias is introduced by mailing a questionnaire that stockholders could complete and mail back? **Sampling bias**, **non-response bias**, or **response bias**. *(please circle your answer and **Briefly explain**.)*

Find the mean of the data summarized in the given frequency distribution.

2) The manager of a bank recorded the amount of time each customer spent waiting in line during peak business hours one Monday. The frequency distribution below summarizes the results. **Find the mean waiting time.** Round your answer to one decimal place.

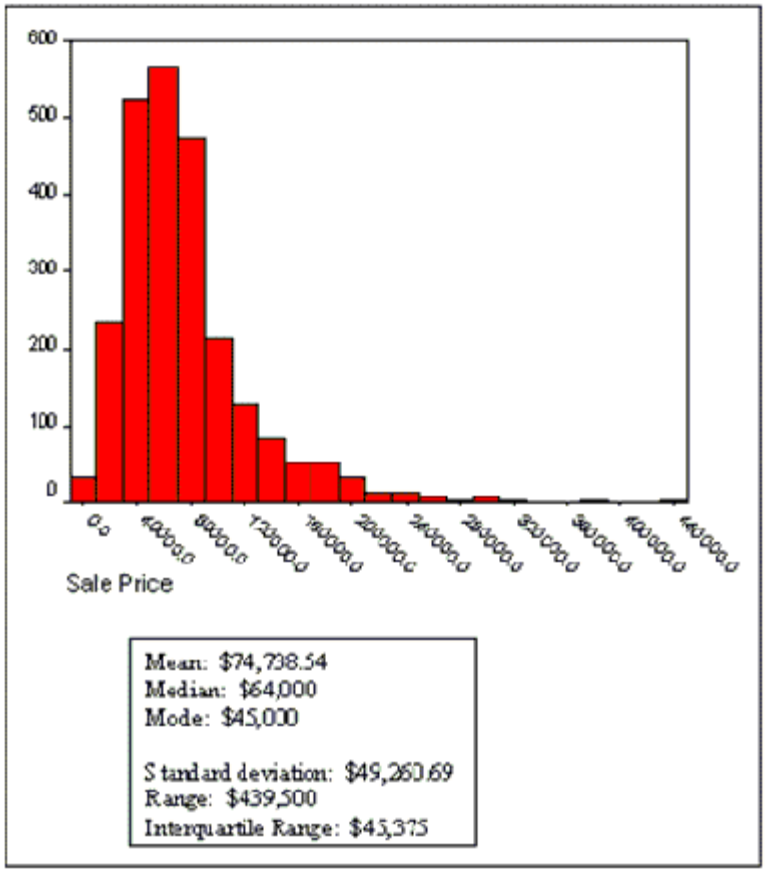
Waiting time (minutes)	Number of customers
0 - 3	15
4 - 7	11
8 - 11	14
12 - 15	7

3) Draw a sketch of two histograms (or smooth curves) that have the same shape and center, but where the first one has a larger standard deviation than the second one.

4) Which of the following statistics is most sensitive to outliers?

A) Interquartile Range
B) Standard Deviation
C) Median
D) Mode

In January 2003, it was reported that 2440 homes were sold in the Twin Cities area in December of 2002. The sale prices of each home are displayed in the histogram, and measures of center and spread associated with this data set are also displayed.



For problems 5 and 6, choose the correct option.

5) The distribution of home sale prices basically looks bell-shaped, with two outliers.
   A) Agree, it looks pretty symmetric if you ignore the outliers.
   B) Agree, most distributions are bell-shaped.
   C) Disagree, it looks more skewed to the left.
   D) Disagree, it looks more skewed to the right.
   E) Disagree, it looks more bimodal.

6) The median home sale price tells you that a majority of homes sold for about $64,000.
   A) Agree, the median is an average and that is what an average tells you.
   B) Agree, $64,000 is representative of the data.
   C) Disagree, a majority of homes sold for more than $64,000.
   D) Disagree, the median tells you only that 50% sold for less than $64,000 and 50% sold for more.

It's time for playoff baseball! These are the hitting statistics for the top 11 MLB (Major League Baseball) players ranked by Batting Average as of 9/28/2011. (from http://mlb.mlb.com/stats/).

RK	Player	Team	Pos	G	AB	R	H	2B	3B	HR	RBI	BB	SO	SB	CS	AVG ▾
1	Cabrera, M	DET	1B	160	568	110	195	47	0	30	105	108	89	2	1	.343
2	Young, M	TEX	DH	158	627	87	212	41	6	11	106	47	78	6	2	.338
3	Gonzalez, A	BOS	1B	158	628	108	212	45	3	27	117	71	119	1	0	.338
4	Reyes, J	NYM	SS	126	537	101	181	31	16	7	44	43	41	39	7	.337
5	Braun, R	MIL	LF	149	559	109	187	38	6	33	111	58	93	32	6	.335
6	Martinez, V	DET	C	144	537	75	175	40	0	12	102	46	51	1	0	.326
7	Kemp, M	LAD	CF	160	598	114	194	33	4	38	124	74	158	40	11	.324
8	Ellsbury, J	BOS	CF	157	655	119	211	46	5	32	105	52	97	38	15	.322
9	Pence, H	PHI	RF	153	602	83	188	38	5	22	96	53	124	8	2	.312
10	Votto, J	CIN	1B	161	599	101	185	40	3	29	103	110	129	8	6	.309
11	Ortiz, D	BOS	DH	145	521	84	160	40	1	29	96	77	83	1	1	.307

7) Make a histogram for the number of homeruns (the column is titled HR) hit by the top 11 MLB players. Create your histogram with 5 classes, use a class width of 7 home runs, and use 5 as the lower class limit of the first class. Be sure to label both axes and give your histogram a title!

8) The number of games played by each of the 11 MLB players (from the column labeled G) are given below (in order from smallest to largest) are:

126 144 145 149 153 157 158 158 160 160 161

Report the five-number summary for the number of games played by the top 11 MLB hitters, giving the name of each statistic and its' value.

name	_____	_____	_____	_____	_____
value	_____	_____	_____	_____	_____

Construct a boxplot for the number of games played.

9) Report the mean number of games played by the top 11 MLB hitters:

Report the standard deviation of the number of games played by the top 11 MLB hitters:

10) Find the **correlation coefficient** and the **equation of the least-squares regression line** for predicting the number of homeruns hit by the top 11 MLB hitters using the number of games played as the explanatory variable.

11) The distributions of SAT and LSAT scores are both approximately normal and symmetric. Georgette took both tests (at different times) and would like to know on which test her performance was better. Use the data given on each test to decide which score was better, relative to other people who took each test. ***Briefly explain your answer.***

Test	Georgette's Score	Mean Score	Standard Deviation
SAT	775	998	203
LSAT	135	150	9

12) Suppose two researchers wanted to determine if aspirin reduced the chance of a heart attack.

Researcher 1 studied the medical records of 500 patients. For each patient, he recorded whether the person took aspirin every day and if the person had ever had a heart attack. Then he reported the percentage of heart attacks for the patients who took aspirin every day and for those who did not take aspirin every day.

Researcher 2 also studied 500 people. He randomly assigned half (250) of the patients to take aspirin every day and the other half to take a placebo every day then after a certain length of time he reported the percentage of heart attacks for the patients who took aspirin every day and for those who did not take aspirin every day.

Suppose that both researchers found that there is a statistically significant difference in the heart attack rates for the aspirin users and the non-aspirin users and that aspirin users had a lower rate of heart attacks. ***Can both researchers conclude that aspirin caused the reduction in the rate of heart attacks?***

A) Yes, because aspirin users had a lower heart attack rate in both studies.

B) Yes, because aspirin is known to reduce heart attacks.

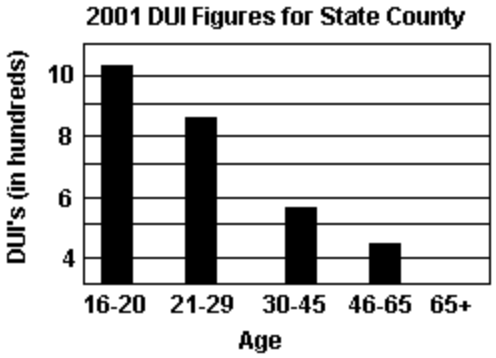
C) No, only researcher 1 can conclude this.

D) No, only researcher 2 can conclude this.

13) Suppose your company is filling cereal boxes and you are monitoring the weights of the boxes for quality control purposes. Why is a small standard deviation (SD) of weights good? Please answer briefly with a short paragraph.

<u>Briefly</u> analyze this graphic, discussing how well it depicts the data and noting any ways it might be misleading. Please use complete sentances.

14)

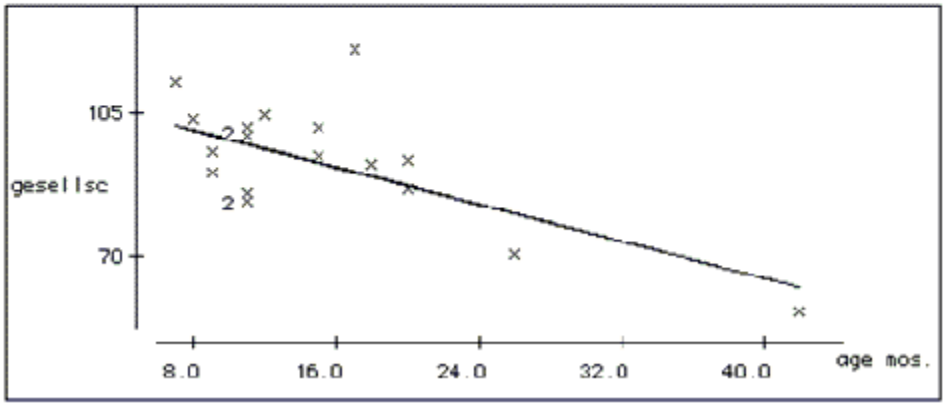


In a study of whether a relationship exists between a child's aptitude and the age at which he/she first speaks, researchers recorded the age (in months) of a child's first speech and the child's score on an aptitude test. The data for these 21 children is shown in the table.

The least squares line for predicting aptitude score from age at first speech turns out to be $\hat{y}$ **= 102.951 - 0.701 x** We could also write it as **score = 102.951 - 0.701 * age.** The value of the correlation coefficient is **r = - 0.397**. The scatterplot displays this relationship.

child	1	2	3	4	5	6	7	8	9	10	11
age	15	2	10	9	15	20	18	11	8	20	7
score	95	71	83	91	102	87	93	100	104	94	113

child	12	13	14	15	16	17	18	19	20	21	
age	9	10	11	11	10	12	42	17	11	10	
score	96	83	84	102	100	105	57	121	86	100	



15) Which child seems to be the most influential observation? *(**circle the point on the scatterplot**)*

16) Does the correlation coefficient represent significant linear correlation? Carefully explain your decision process as well as state your conclusion.

17) What would the least squares line predict for the aptitude score of a child who first spoke at 20 months?

18) The following data represent the living situation of newlyweds in a large metropolitan area and their annual household income. What percent of people who own their own home make between $35,000 and $50,000 per year? Carefully show the details of your computation and respond with a complete secntance.

	< $20,000	$20-35,000	$35-50,000	$50-75,000	> $75,000
Own home	31	52	202	355	524
Rent home	67	66	52	23	11
Live w/family	89	69	30	4	2

19) A county real estate appraiser wants to develop a statistical model to predict the appraised value of houses in a section of the county called East Meadow. One of the many variables thought to be an important predictor of appraised value is the number of rooms in the house. Consequently, the appraiser decided to fit the simple linear regression model, $\hat{y} = \beta_0 + \beta_1 x$, where y = appraised value of the house (in thousands of dollars) and x = number of rooms. Using data collected for a sample of n = 74 houses in East Meadow, the following regression equation was obtained:

$\hat{y} = 74.80 + 22.75x$

Give a practical interpretation of the estimate of the slope of the least squares line.

A) For each additional room in the house, we estimate the appraised value to increase $22,750.

B) For each additional dollar of appraised value, we estimate the number of rooms in the house to increase by 22.75 rooms.

C) For each additional room in the house, we estimate the appraised value to increase $74,800.

D) For a house with 0 rooms, we estimate the appraised value to be $74,800.