Name _____

AP Statistics Midterm Review

Exploring Data

1) Mini-tab data for monthly rate of return on Wal-Mart stock from 1973 to 1991 is shown on the right.

```
Mean   =  3.064
Standard deviation  =  11.49

N  = 228   Median  =  3.4691
Quartiles  =  -2.950258,  8.4511

Decimal point is 1 place to the right of the colon

Low:  -34.04255  -31.25000  -27.06271  -26.61290

 -1 : 985
 -1 : 444443322222110000
 -0 : 999988777666666665555
 -0 : 444444443333333222222222221111111100
  0 : 000001111111111112222223333333344444444
  0 : 555555555555555555556666666666777777788888888888899999
  1 : 000000001111111112223333444 4
  1 : 55566667889
  2 : 011334

High:  32.01923  41.80531  42.05607  57.89474  58.67769
```

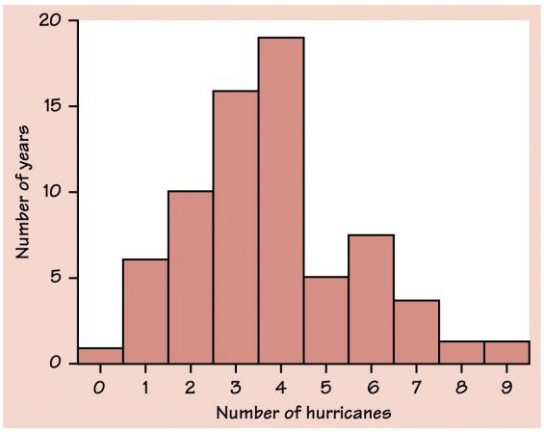
(a) Give the five number summary for the data presented.

(b) Describe in words the distribution of the data. (SOCS!)

(c) Find the IQR. Are there any outliers based on the outlier formula? Based on the output, does it appear the Minitab program uses this formula?

(d) What does the term "resistant" mean? Which statistics are considered resistant?

2) The distribution of the annual number of hurricanes in the US over a 70 year period is given on the right.



(a) Give a description of the data.

(b) What would be the appropriate measures to use for center and spread based on your answer to (a)?

(c) About where does the center of the data lie?

Undergraduate college enrollment, fall 1995 (thousands of students)

Age	2-year full-time	2-year part-time	4-year full-time	4-year part-time
under 18	41	125	75	45
18 to 24	1378	1198	4607	588
25 to 39	428	1427	1212	1321
40 and up	119	723	225	605
Total	1966	3472	6119	2559

Source: Digest of Education Statistics 1997, accessed on the National Center for Education Statistics Web site, http://www.ed.gov/NCES.

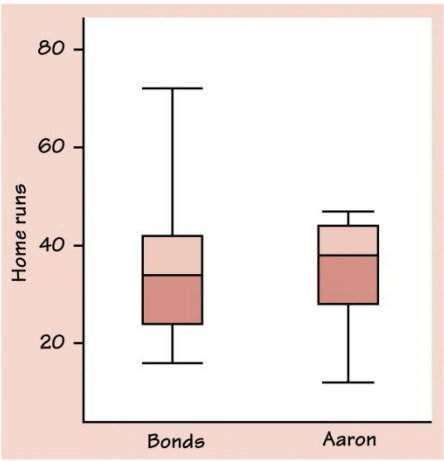
* NOTE: Rounding of thousands may cause slight error in totals!

3) What percent of all undergraduate students were 18 to 24 years old in the fall of the academic year?

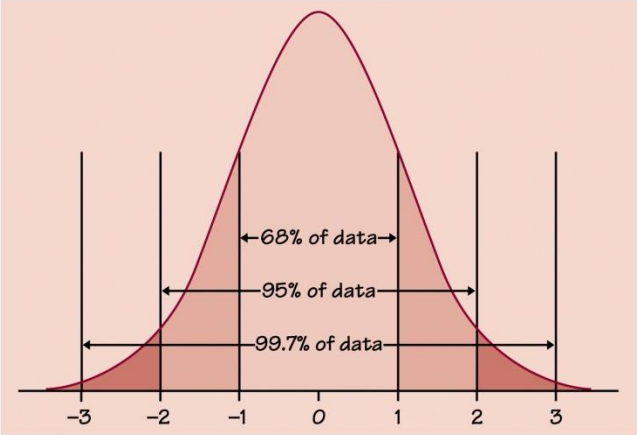
4) What proportion of 2-year full-time students were 40 and up?

5) Find the marginal distribution of age among all undergraduate students, first in counts and then percents. Make a bar graph of the distribution in percents. Comment on what you see.

6) The box-plots on the right show the career homeruns distributions for Barry Bonds and Hank Aaron. Describe and compare the distributions.



Modeling Distributions of Data



The Weschler Adult Intelligence Scale scores for teenagers is described by $N(110, 25)$.

7) What percent of teenagers have scores between 85 and 135?

8) What proportion of teenagers have scores lower than 80?

9) What score is considered the cutoff for the 90^{th} percentile?

10) What score would a teenager need to be in the top 1% of all teenagers?

11) What is the z-score for a student who scores 105? Describe what this z-score means.

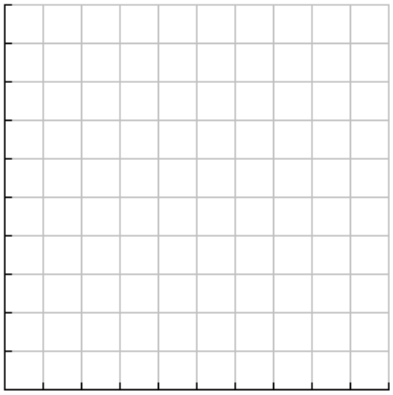
12) What area of a normal curve would correspond to $-1.3 < Z < 2.4$?

Describing Relationships (Least Squares Regression)

Wine consumption and heart disease

Country	Alcohol from wine (liters/year)	Heart disease death rate (per 100,000)	Country	Alcohol from wine (liters/year)	Heart disease death rate (per 100,000)
Australia	2.5	211	Netherlands	1.8	167
Austria	3.9	167	New Zealand	1.9	266
Belgium/Lux.	2.9	131	Norway	0.8	227
Canada	2.4	191	Spain	6.5	86
Denmark	2.9	220	Sweden	1.6	207
Finland	0.8	297	Switzerland	5.8	115
France	9.1	71	United Kingdom	1.3	285
Iceland	0.8	211	United States	1.2	199
Ireland	0.7	300	West Germany	2.7	172
Italy	7.9	107			

Source: M. H. Criqui, University of California, San Diego, reported in the *New York Times*, December 28, 1994.



13) Construct a scatter-plot to the right and label axes. Describe the relationship between the variables below. Assume alcohol consumption is the explanatory variable.

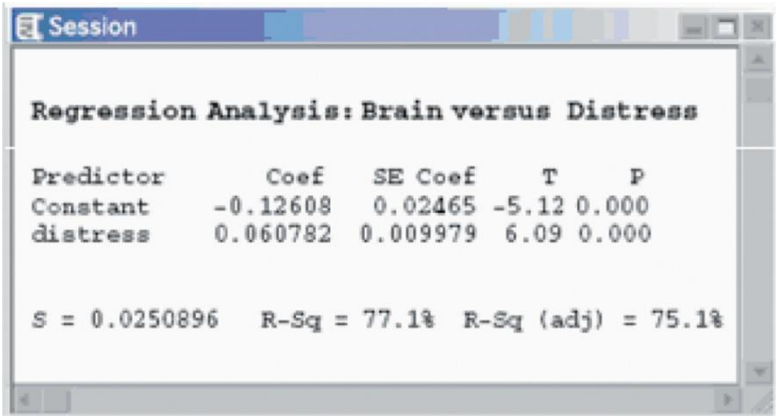
14) Determine the LSRL equation for predicting heart disease death rate from wine consumption below. Show the line on your scatter-plot.

15) Interpret the correlation coefficient and the coefficient of determination based on the data. Be very careful how you state the interpretations.

16) Based on your regression model in #15.

(a) Predict the heart disease death rate from a country with a annual alcohol from wine consumption of 10.1 liters/year of alcohol from wine. Are you confident in prediction?

(b) Show a residual plot (including labels!) for this problem below. Comment on the plot – is a linear model appropriate? Could any data point be considered influential?



17) The figure above shows computer output for a linear regression between measured brain activity in a human and a measure of social distress a person feels from exclusion. Both variables are quantitative to allow for a regression.

(a) What is the linear regression equation found by the computer? Be sure to label the explanatory and response variables.

(b) What is the correlation coefficient between the variables?

18) In an economics class, the correlation between students' total scores prior to the final exam and their score on the final examination score is $r = 0.7$. The mean and standard deviation of prior total scores are 280 and 30, respectively. The mean and standard deviation for final exam scores are 75 and 8, respectively. What is the equation for the LSRL?

Designing Studies

19) Two treatments are considered for breast cancer detected in early stages, total removal of the breast or removal of just the tumor and nearby lymph nodes. A medical team examines the records of 25 hospitals and examines the length of time of survival after surgery of women who have had either treatment.

(a) What are the explanatory and response variables?

(b) Is this an observational study or an experiment?

(c) Is there a confounding variable present? If so, explain.

20) An experiment compares two new varieties of corn with higher lysine content, opaque-2 and floury-2, to see if roosters/chickens grow larger with the new types of corn. Researchers mix each type of corn with soybeans at three protein levels, 12%, 16% and 20% to create diet mixes for the chickens. They feed each diet to 10 one-day old birds and record weight gains after 21 days.

(a) What are the experimental units and response variable?

(b) How many factors are there? How many treatments? Use a diagram to describe the experiment. How many experimental units are required?

(c) How would you create a randomized design for the experiment?

(d) Male chicks (baby roosters?) grow faster than females. How could you account for this in the experimental design? What is this process called?

21) We want to know if people like Coke or Pepsi better. So we set up a double-blind taste test where each subject tastes both. Describe what this means. Is this an experiment or observational study? What type of observational study or experiment is this?

Probability: What Are The Chances?

Age and marital status of women (thousands of women)

	Age			
	18–29	30–64	65 and over	Total
Married	7,842	43,808	8,270	59,920
Never married	13,930	7,184	751	21,865
Widowed	36	2,523	8,385	10,944
Divorced	704	9,174	1,263	11,141
Total	22,512	62,689	18,669	103,870

Source: Data for 1999 from the *2000 Statistical Abstract of the United States.*

22) What is P(Married)?

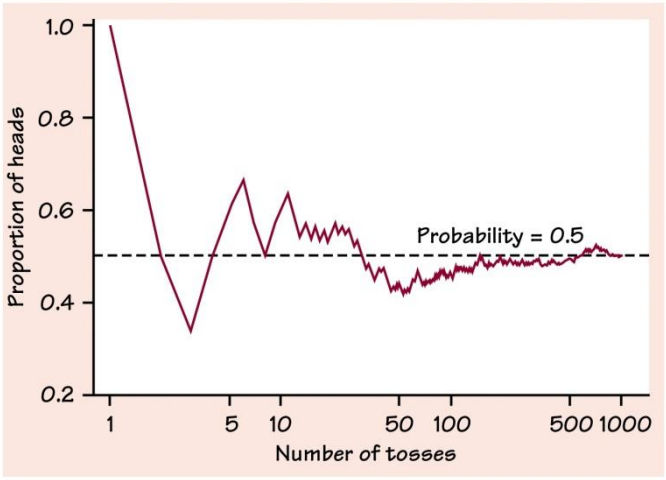
23) What is P(Age 30-64)?

24) What is P(Married\Age 30-64) ?

25) Are the events Married and Age 30-64 independent? Justify your answer.

26) Are the events Married and Age 30-64 disjoint? Justify your answer.

27) If a fair coin is tossed many times, as *n* increases, the proportion of heads should approach 0.5. What important statistical idea does this represent?



28) Sadly, the probability that a dropped piece of buttered bread will land on the buttered side is 0.72. Yuck!

(a) If three pieces of buttered bread are dropped, what is the probability at least one will land on the buttered side?

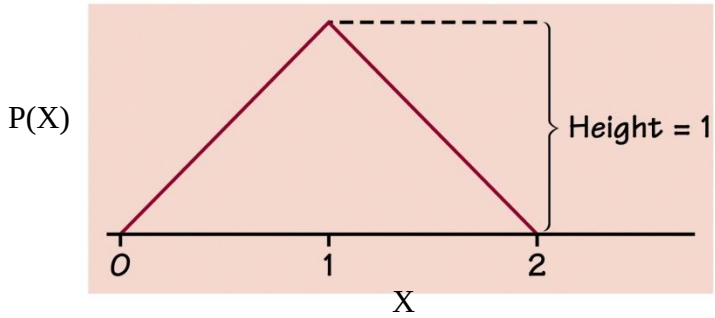
(b) Describe how you would simulate this using a random digit table.

29) John is considering either surgery for heart bypass or medical management of his heart condition. The event "A" is that John survives at least 5 years and has a good quality of life during that time. The doctor gives him the following information:

* Under medical management M, P(A\M) = 0.7
* Probability is 0.05 John will not survive bypass surgery, P(B)
* Probability is 0.10 John survives surgery but with serious complications, P(C)
* Probability is 0.85 John survives surgery with no complications, P(D)
* If John survives surgery with complications, P(A\C) = 0.73.
* If John survives with no complications, P(A\D) = 0.76.

Calculate overall P(A) assuming John chooses surgery. Should he choose surgery or medical management based on the doctor's information?

Random Variables



A probability distribution of a continuous random variable.

30) $P(X \leq 0.5) =$

31) $P(X < 0.5) =$

32) $P(0.5 < X \leq 1.5) =$

33) $P(1.0 \leq X \leq 2.5) =$

Assume an NFL football score simulator is designed to give the following discrete distribution of points by an average football team in a single quarter of the game:

X	0	3	7	10	14
P(X)	0.1	0.1	0.4	0.2	?

34) What is the missing probability?

35) What is the expected mean (μ_X) and standard deviation (σ_X) of a score by a team in a quarter?

36) What is the expected mean (μ_X) and standard deviation (σ_X) of a score by a team in a game? (Assume each quarter score is independent).

37) Unfortunately, the Lions μ_X = 22.1 and σ_X = 3.1 for an entire game. By how much should the Lions expect to lose to an average NFL team? What is the standard deviation of the distribution of the difference between an average NFL team and the Lions score?

38) If the Lions draft the #1 running back, Las Vegas assumes their revised mean score (Y) would be modeled by the equation $Y = 1.5X - 2.1$. What are the new values of μ_Y and σ_Y for the Lions?

A family has seven children. Assume the genetics of the parents is such that the chance of having a girl is 0.6, and the probability of having a boy is 0.4. Assume they are independent events.

39) What is the probability that all seven children are boys?

40) What is the probability that exactly 3 children are girls?

41) What is the probability that at most 2 children are girls?

42) What is the mean number of girls expected of this family? What is the standard deviation of the count of girls?

43) Would it be accurate to assume a normal distribution with the mean and standard deviation calculated above? Why or why not?

Suppose Miguel Cabrera expects to hit for his lifetime average next year for the Tigers, which happens to be 0.321. (This is his total numbers of hits / total number of at-bats).

44) What is the probability he will get a hit on his first at bat in a game?

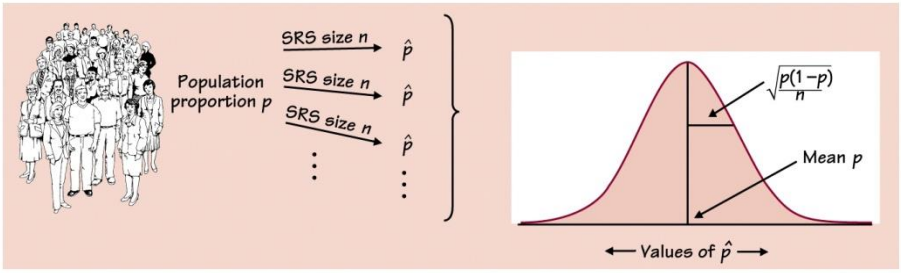
45) What is the probability he will get his first hit on his fifth at-bat in the game?

46) What is the probability it will take at most 3 at-bats to get a hit in a game?

47) If Miguel gets 620 at bats next year, how many hits should he expect for the season?

48) How many at bats should we expect before he gets a hit next season?

Sampling Distributions

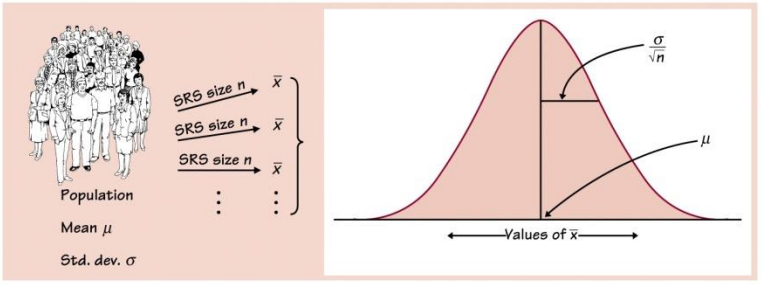


49) Voter registrations show 68% of registered Troy voters are Republicans. To test this you randomly phone 150 registered Troy voters, and 73% are found to be Republican.

(a) Use proper notation to indicate if 68% a parameter or statistic? What about the 73%?

(b) What are the mean and standard deviation of the sampling distribution?

(c) What is the probability of obtaining a SRS of 150 Troy voters in which 73% or more are registered Republicans.

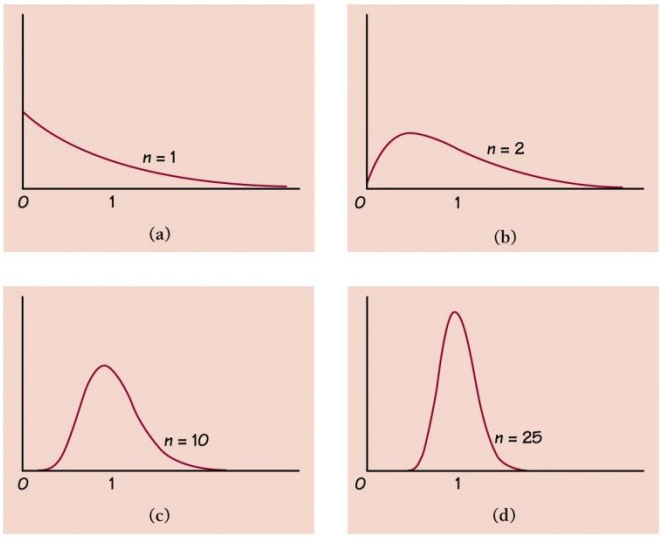


50) A sulfur compound (DMS) is sometimes found in milk from cows eating near onion fields. Researchers want to find the lowest concentration of DMS detectable to humans. Previous studies show that this threshold follows a normal distribution with $\mu = 26$ and $\sigma = 7$. Here are results from 10 randomly chosen subjects (in micrograms/liter):

28	40	28	33	20	31	29	27	17	21

(a) What is the mean threshold based on this sample? (Use proper notation)

(b) What is the probability of getting a sample mean even farther away from μ than this?



51) Suppose a density curve for a distribution of a population is very non-normal, with $\mu = 1$, as shown in the figure above. Many sample means of different sizes ($n = 2, n = 10, n = 25$) from this population are observed. As n increases the shape becomes more normal. What important statistical idea does this represent?

Estimating with Confidence

52) The City of Troy is wondering whether residents would favor a millage to support road repairs in the city. A preliminary SRS of 300 potential Troy voting residents finds that 162 would vote in favor of the millage.

(a) Check required assumptions to create a confidence interval of passing the millage.
(b) Construct a 95% confidence interval. Should the City be 95% confident the millage will pass? Justify your answer.
(c) Assuming the sample proportion remains the same as found in part (a). How large would a SRS is required so the City could be 95% confident of the millage passing?

53) A member of the Troy Colt bowling team usually bowls 400 games in the course of a year. A random sample of a member of the Troy Colt Bowling team shows the following scores for 12 games:

145 141 123 162 170 155 167 177 148 163 175 169

(a) Construct a 99% confidence interval for the mean score for this bowler.
(b) What condition was violated when creating this interval? Do you feel confident in the interval that was created?