Review for AP Exam and Final Exam Name _____

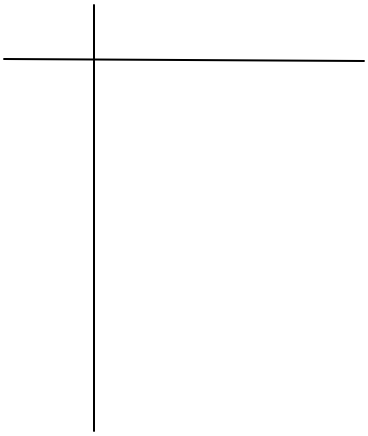
Topic I. Exploratory Analysis of Data

1. The data in the chart below shows the survival times in days for guinea pigs after they were injected with tubercle bacilli in a medical experiment.

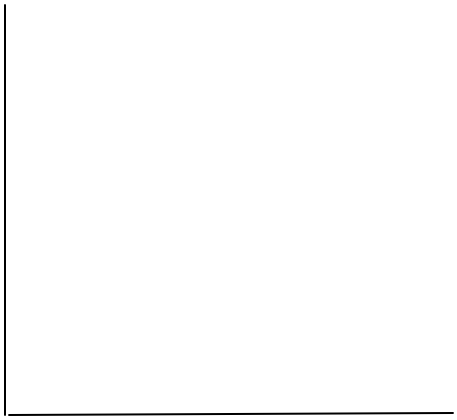
43	45	53	56	56	57	58	66	57	73	74	79	80	80	81	81	81	82	82	83
83	84	88	89	91	91	92	97	99	99	100	101	102	102	102	103	104	107	108	109
113	114	118	121	123	126	128	137	138	139	144	147	156	162	174	178	179	184	191	198
211	214	243	249	329	380	403	511	522	508	510	514	520	520	521	530	530	533	540	541

a. Create the following charts and graphs for the data in the chart above:

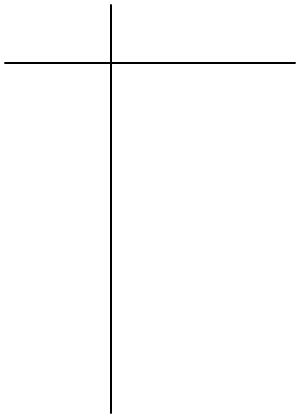
Frequency Table



Histogram



Stem and Leaf Plot



b. Discuss the main features of the histogram. Center, spread, clusters, gaps, outliers, shape.

c. Find the following values for the data.

Measures of **Center**: Median _____ Mean _____

Measures of **Spread**: Range _____ IQR _____

Standard Deviation _____ Variance _____

Measures of **Position**: Q1 _____ Q2 _____ Q3_____

Min _____ Max _____

The 70 Percentile _____

c. Find the standardized scores (z-scores) for 80 days and 520 days.

d. List the 5-number summary and create the modified box plot for the data.

e. Identify any outliers by using the IQR method.

f. If the data were changed in the following ways, which one of the summary measures would change and how would they change?

Change the max days to 1000 _____

Trim the data by 10% _____

Change the unit of measures by dividing every piece of data by 100 _____

2. The following quiz scores are from 2 different classes for an AP Stats test in chapter 1.

4th Hour	48	76	82	96	92	84	100	98	96	76	92	72	88	82	66	58	78	81	78
78	92	92	78	84	52	70	84	88	92	84									

5th Hour	90	96	78	94	94	88	86	96	86	82	90	87	88	76	92	94	80	82	88
84	86	80	86	72	96	90													

a. Create back-to-back box-plots (on the same scale) and compare them on the following:

Spread: Center: Clusters:

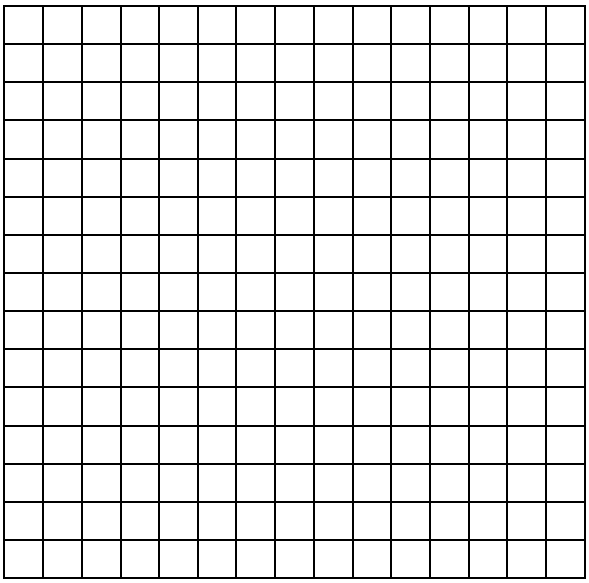
Gaps: Outliers: Shape:

3. Is there a correlation between test anxiety and exam score performance? Data on x = score on a measure of test anxiety and y = exam score are given in the table below.

X = test anxiety	23	14	14	0	7	20	20	15	21
Y = score on exam	43	59	48	77	50	52	46	51	51

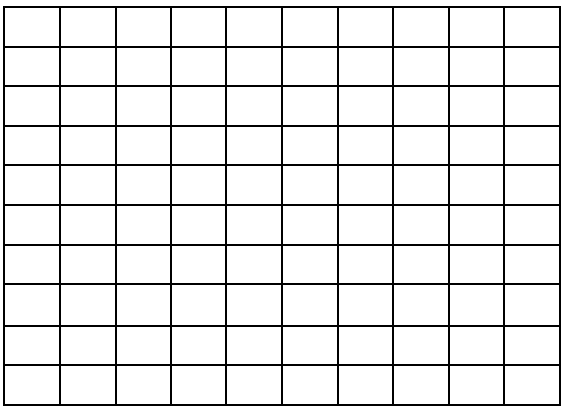
a. Which one of the variables is the explanatory and which is the response variable?

b. Construct a scatter plot and comment on the features of the plot. (Overall pattern, deviations, direction, form, strength)



c. Find the correlation coefficient, the coefficient of determination and the LSRL.

d. Construct a residual table and the residual plot.



e. Comment on the relationship between test anxiety and test scores based upon the analysis you performed.

f. If we were to add the data point (5,100) how would it affect the LSRL? What is this point called?

4. The sample correlation coefficient between annual raises and teaching evaluations for a sample of 353 college faculty was found to be r = .11.

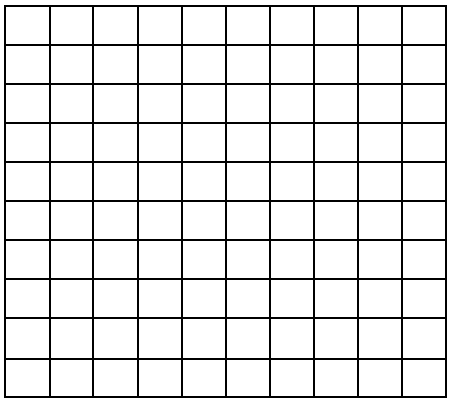
   a. Interpret this value.

   b. If a straight line were fit to the data using least squares regression, what proportion of variation in raises could be attributed to the approximate linear relationship between raises and evaluation?

5. Each year the FBI issues a report that provides information about crimes in the United States. The following table gives the total number of violent crimes in the United States for the year 1984 to 1994.

Year (x)	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
No. of violent crimes (y) (thousands)	1273	1329 ( )	1489 ( )	1484 ( )	1566 ( )	1646 ( )	1820 ( )	1912 ( )	1932 ( )	1923 ( )	1864 ( )

   a. Plot the data. Observe that there is a pattern but that several points don't fit the pattern. Which points don't fit?

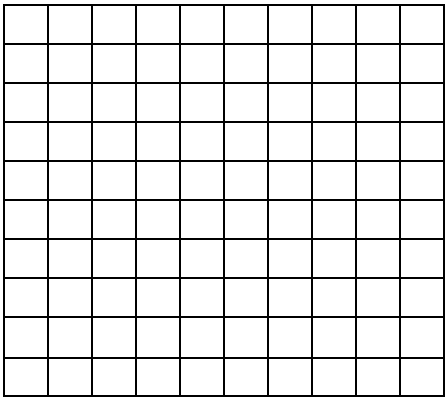


b. Are violent crimes increasing linearly or exponentially? Calculate the ratios and put into the table, where you see the ( ). Are the ratios approximately constant and greater than 1? What is the average ratio for the first eight data points?

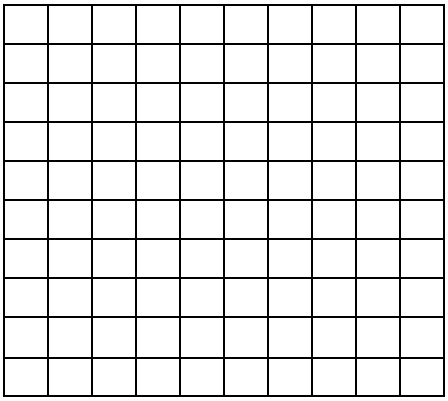
b. You decide to discard the last three points and develop and exponential model for the years 1984 to 1991. Delete these points and transform the remaining data to achieve a linear scatterplot. Put the years (x) and the transformed values for y in the table below.

Year									

c. Plot the transformed data and the residual plot for the transformed plot. Perform a least squares regression on the transformed points and record the correlation coefficient, coefficient of determination and LSRL.



d. Perform the inverse transformation and record the equation that model the data for the years 1994 to 1991.



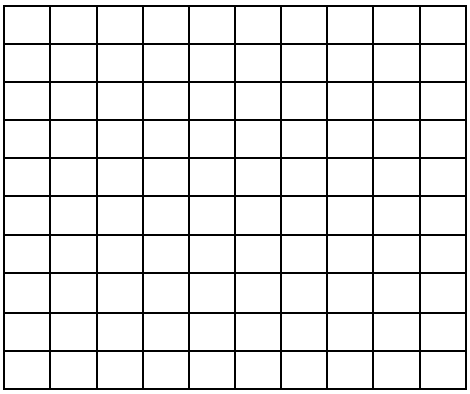
e. Use the exponential model from part d to predict the number of violent crimes in 1986.

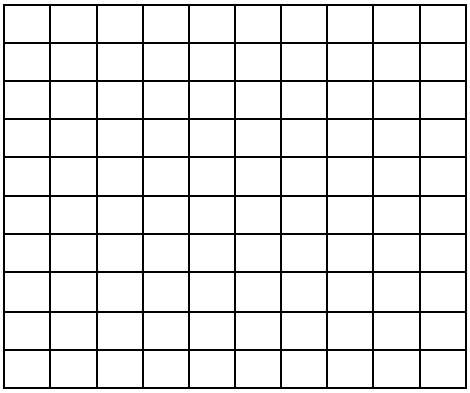
f. 1986 produces the largest residual. What is the residual for this year?

6. In physics class, the intensity of a 100-watt light bulb was measured by a sensing device at various distances from the light source, and the following data was collected. Note that a candela (cd) is an international unit of luminous intensity.

Distance	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Intensity (candelas)	.2965	.2522	.2055	.1746	.1534	.1352	.1145	.1024	.0923	.0832	.0734

a. Plot the data. Based on the pattern of the points, propose a model for the data. Then use a transformation followed by a linear regression and then an inverse transformation to construct a model.





b. Describe the relationship between the intensity and the distance from the light source.

7. The following table reports Census Bureau data on undergraduate students in U.S. colleges and universities in the fall of 1991.

UnderGraduate College enrollment by age of students – Fall 1991 (thousands of students)

Age	2-yr Full-time	2-yr part-time	4-yr full time	4-yr part-time	Totals
15-17	44	4	79	0	
18-21	1345	456	3869	159	
22-29	489	690	1358	494	
30-44	287	704	289	627	
>=45	49	209	62	160	
Totals					GT ( )

a. Fill in the "totals" in the table above. What is the grand total (GT) of students who were enrolled in colleges and universities in the fall of 1991?

b. What percent of all undergraduate students were 18-21 years old in the fall of the 1991?

c. Find the percent of the undergraduates enrolled in each of the four types of programs who were 18-21 years old. Make a bar chart to compare these percents.

d. The 18-21 group is the "traditional" age group for college students. Briefly summarize what you have learned from the data about the extent to which this group predominates in different kinds of college programs.

Topic II. Sampling and Experimentation: Planning and Conducting a Study

8. Define these terms:
    a. Census

    b. Population

    c. Sample

    d. Survey

    e. Simple Random Sample (SRS)

    f. Bias in a sample

    g. Confounding

    h. Stratified random sample

    i. Cluster Sample

    j. Block design

    k. Experiment

    l. Observational study

9. The Ministry of Health in the Canadian Province of Ontario wants to know whether the national health care system is achieving its goals in the province. Much information about health care comes from patient records but that source doesn't allow us to compare people who use health services with those who don't. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a random sample of 61,239 people who in the Province of Ontario.

    a. What is the population for this sample survey? What is the sample?

    b. The survey found the 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. Do you think these estimates are close to the truth about the entire population? Why or why not?

    c. Is this an experiment or an observation study? How can you tell?

10. What are the characteristics of a well-designed and well-conducted study?

11. Elaine is enrolled in a self-paced course that allows three attempts to pass an examination on the material. She does not study and has 2 out of 10 chances of passing on any one attempt by pure luck. What is Elaine's likelihood of passing on at least one of the three attempts? (Assume the attempts are independent because she takes a different exam at each attempt.)

    a. Explain how you would use random digits to simulate one attempt at the exam. Elaine will of course stop taking the exam as soon as she passes.

    b. Simulate 50 repetitions. What is your estimate of Elaine's likelihood of passing the course?

    c. A more realistic model for Elaine's attempts to pass an exam would be as follows: On the first try she has a probability 0.2 of passing. If she fails on the first try, her probability on the second try increases to 0.3 because she learned something from the first try. If she fails on the first 2 attempts, the probability of passing on the third attempt is 0.4. She will stop as soon as she passes. The course rules force her to stop after three attempts. Explain how to simulate one repetition of Elaine's tries on the exam with this new approach.

    d. Simulate 50 repetitions and estimate the probability that Elaine eventually passes the exam with the approach in part c.

12. Can aspirin help prevent heart attacks? The Physicians' Health Study, a large medical experiment involving 22,000 male physicians, attempted to answer this question. One group of about 11,000 physicians took an aspirin every second day, while the rest took a placebo. After several years the study found that subjects in the aspirin group had significantly fewer heart attacks than the subjects in the placebo group.

    a. Identify the experimental subjects, the factor and its levels, and the response variable in the health study.

    b. Use a diagram to outline a completely randomized design for the health study.

13. A mortgage lender routinely places advertisements in a local newspaper. The advertisements are of three different types: one focusing on low interest rates, one featuring low fees for first-time buyers, and one appealing to people who may want to refinance their homes. The lender would like to determine which advertisement format is most successful in attracting customers to call for more information. Describe be an experiment that would provide the information needed to make this determination. Be sure to consider extraneous factors such as the day of the week that the advertisement appears in the paper, the section of the paper in which the advertisement appear, daily fluctuations of the interest rate and so forth. What role does randomization play in your design? Diagram the design.

Topic III Anticipating Patterns: Exploring Random Phenomena using Probability and Simulation

14. Probability is a measure of how likely an event is to occur. Match one of the probabilities that follow with each statement about an event.

0	0.01	0.3	0.6	0.99	1.00

a. The sun will rise in the west in the morning.

b. Thanksgiving will be on Thursday next year.

c. An event is very unlikely, but it will occur vary rarely.

d. The event will occur most of the time. Very rarely will it not occur.

e. Give an example of where the other 2 probabilities may occur.

15. What is the formula used for each of the following probabilities:

a. Addition Rule b. Multiplication Rule c. Conditional Probability

16. The type of medical care a patient receives may vary with the age of the patient. A large study of women who had a breast lump investigated whether or not each woman received a mammogram and a biopsy when the lump was discovered. Here are some probabilities estimated by the study. The entries in the table are the probabilities that both of two events occur; for example: 0.321 is the probability that a patient is under 65 years of age and the tests were done.

	Tests	Done
	Yes	No
Age Under 65	.321	.124
Age 65 and Over	.365	.190

a. What is the probability that a patient in this study is under 65?

b. Is 65 or over?

c. What is the probability that the tests were done for a patient? That they were not done?

d. Are the events A = (patient was 65 or older) and B= (the tests were done) independent? Were the tests omitted on older patients more or less frequently that would be the case if testing were independent of age?

17. Here are the counts (in thousands) of earned degrees in the United States in a recent year, classified by level and by the sex of the degree recipient:

	Bachelor's	Master's	Professional	Doctorate	Total
Female	616	194	30	16	
Male	529	171	44	26	
Total					

a. If you choose a degree recipient at random, what is the probability that the person you choose is a woman?

b. What is the conditional probability that you choose a woman, given that that person chosen received a professional degree?

c. Are the events "choose a woman" and "choose a professional degree recipient" independent? How do you know?

18. Consolidated Builders has bid on two large construction projects. The company president believes that the probability of winning the first contract (event A) is 0.6, that the probability of winning the second (event B) is 0.4 and the joint probability of winning both jobs (event A and B) is 0.2 .

a. Draw the Venn diagram that illustrates the relationship between events A and B.

b. Find the following probabilities:

P(A or B) P(A and B) P(A, and Not B)

P(Not A, and B) P(not A and not B)

19. What is the difference between discrete and continuous random variables?

20. Let x be the number of courses for which a randomly selected student at a certain university is registered. The probability distribution of x appears in the accompanying table.

X	1	2	3	4	5	6	7
P(x)	0.02	0.03	0.09	0.25	0.40	0.16	0.05

a. What is P(x = 4)? b. What is P(x <=4)?

c. What is the probability that the selected student is taking at most five courses?

d. What is the probability that the selected students is taking at least five courses?

e. Calculate P(3<=x<=6) and P(3<x<6). Explain why the two probabilities are different.

f. Find the mean, standard deviation and variance of the random variable x.

21. You have two scales for measuring weights in a chemistry lab. Both scales give answers that vary a bit in repeated weightings of the same item. If the true weight of a compound is 2.00 grams, the first scale produces readings X that have mean 2.000 grams and standard deviations 0.002 grams. The second scale's readings Y have mean 2.001 grams and standard deviation of 0.001 grams.

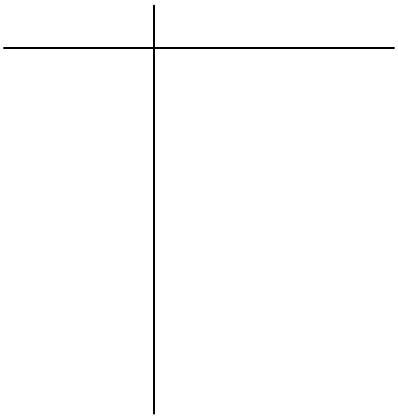
a. What are the mean and standard deviation of the difference y – x between the readings? (The readings X and Y and independent.)

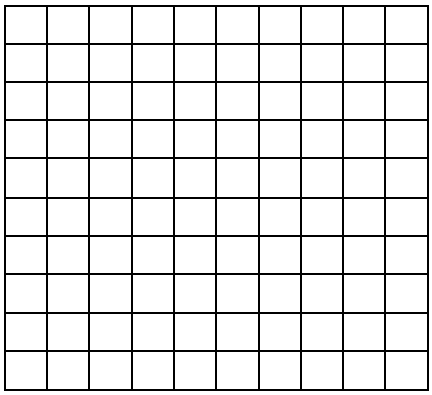
b. You measure once with each scale and average the readings. Your result is Z = (X + Y)/2. What are the mean and standard deviation of Z?

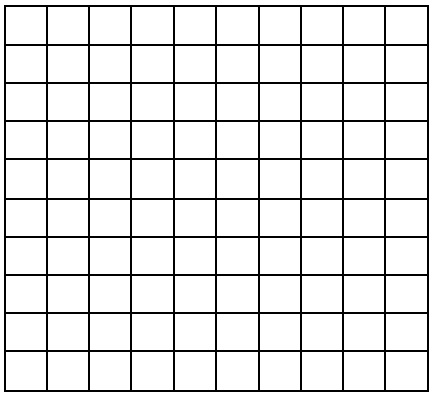
22. Among employed women, 25% have never been married. You select 10 employed women at random.

a. The number in your sample who have never been married has a binomial distribution. What are n and p?

b. Create a binomial distribution table, a probability distribution table and a cumulative distribution table for this data.







c. What is the probability that exactly 2 of the 10 women in your sample has never been married?

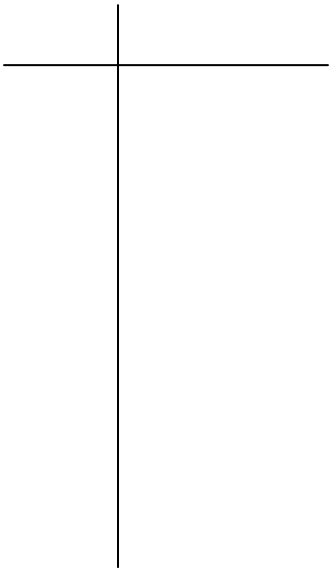
d. What is the probability that 2 or fewer have never been married?

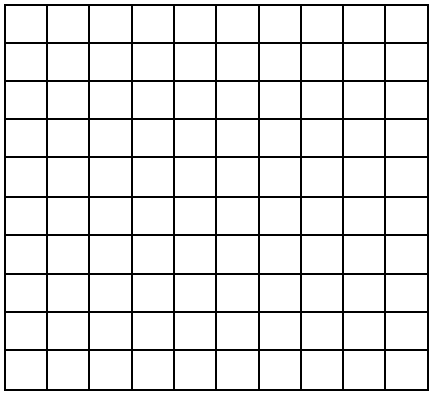
e. What is the mean and standard deviation for this binomial distribution?

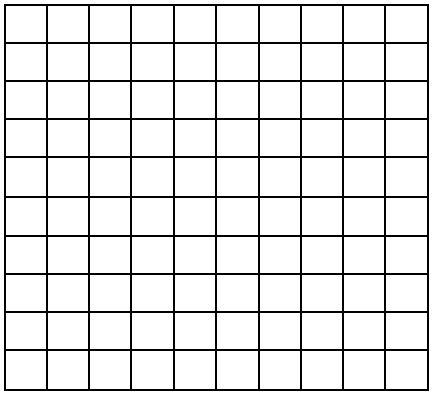
23. A basketball players makes 80% of his free throws. We put him on the free throw line and ask him to shoot free throws until he misses one. Let X = number of free throws the player takes until he misses.

a. What assumptions do you need to make in order for the geometric model to apply? With these assumptions, verify that X has a geometric distribution. What actions constitutes "success" in this context?

b. Create a geometric distribution table for x values from 1 to 10. Create a pdf and a cdf.







c. What is the probability that the player will make 5 shots before he misses?

d. What is the probability that he will make at most 5 shots before he misses?

e. What is the mean of this geometric distribution?

24. The area under the curve for a normal distribution is represented by a bell-shaped curve.

a. What are the properties of a normal distribution? Sketch a normal curve.

25. A certain population of whooping cranes that migrate between Wisconsin and Florida every year has a SRS taken. The sample of 15 male cranes were weighed before they left Wisconsin to begin their trip. The mean weight of the 15 males was found to be 22.7 pounds with a standard deviation of 2.3 pounds. Why is this population considered a normal distribution?

a. What is the probability that a random selected male crane weights less than 20 pounds? Sketch the curve and put in all the appropriate values. Write the probability statement.

b. What is the probability that a random selected male crane weights more than 25 pounds?

c. What is the probability that a random selected male crane weights between 21 and 26 pounds?

d. When these cranes reach Florida, another random sample of 25 male cranes is weighted and measured. The mean weight is recorded at 19.5 pounds with a standard deviation of 1.7 pounds. Using this sample statistics, make a prediction about another sample of 25 from the same population, what is the probability that the mean of the samples will be between 15 and 22 pounds?

e. What is the probability that the sampling distribution of 25 cranes would have a mean greater than 23 pounds?

f. What is the probability that the sampling distribution would be less than 18 pounds?

26. The Helsinki Heart Study asks whether the anti-cholesterol drug gemfibrozil will reduce heart attacks. In planning such an experiment, the researchers must be confident that the sample sizes are large enough to enable them to observe enough heart attacks. The Helsinki study plans to give gemfibrozil to 2000 men and a placebo to another 2000 men. The probability of a heart attack during the 5-year period of the study for men this age is about 0.04. We can think of the study participants as an SRS from a large population, of which the proportion $p = 0.04$ will have heart attacks.

    a. What is the mean number of heart attacks that the study will find in one group of 2000 men if the treatment doesn't change the probability of 0.04?

    b. What is the probability that the group will suffer at least 75 heart attacks? Sketch the curve, show all the work and write the probability statement.

27. Children in kindergarten are sometimes given the Ravin Progressive Matrices Test (RPMT) to assess their readiness for learning. Experience at Southward Elementary School suggests that the RPMT scores for its kindergarten pupils have a mean of 13.6 with a standard deviation of 3.1. The distribution is close to normal. Mr. Brown has 22 children in his kindergarten class this year.

    a. What is the probability that class's mean score will be less than 12.0?

    b. Mr. Brown suspects that the class RPMT scores will be unusually low because the test was interrupted by a fire drill. He wants to find the level L such that there is only a probability of 0.05 that the mean score of his class fall below L. What is this value of L. (Hint: this requires you to find the z-score and then convert to the x-score.)

28. Explain what is meant by the Law of Large Numbers. How does this law apply to sampling distributions?

29. What is the Central Limit Theorem? How is the CLT used in sampling distributions?

Topic IV: Statistical Inference: Estimating Population Parameters and Hypotheses Testing

30. Estimating Population Parameters.
    a. Why is an unbiased statistic generally preferred over a biased statistic for estimating a population characteristic?

    b. Does unbiasedness alone guarantee that the estimate will be close to the true value? Explain.

    c. A random sample of 12 four-year old red pine trees was selected and the diameter (in) of each tree's main stem was measured. The resulting observations are as follows:

    11.3 10.7 12.4 15.2 10.1 12.1 16.2 10.5 11.4 11.0 10.7 12.0

    Find the point estimate that can be used to estimate the true population mean.

    Find the point estimate that can be used to estimate the true population standard deviation.

    Find the point estimate that be used to estimate the true population proportion of trees whose diameter is greater than the average.

31. What is meant by the standard error of a population parameter? What are the standard errors for the following:

    Population Mean Population Proportion Population Variability

    Difference between Two Population Means Difference between Two Population Proportions

32. What is the general form of all confidence intervals?

33. Suppose that a random sample of 50 bottles of a particular brand of cough medicine is selected and the alcohol content of each bottle is determined. Let mu denote the average alcohol contend for the population of all bottles of the brand under the study. Suppose that the sample mean is 8.2 grams with a standard deviation of 1.5 grams.

    a. Find the 95% confidence interval for the mean alcohol content of the cough medicine. Report the margin of error and show all the work.

    b. Explain in words any layman can understand what the 95% confidence interval means.

    c. Would the 90% confidence interval be narrower or wider? Explain why.

    d. The manufacturer claims that the alcohol content is 8.0 grams per bottle. Perform a hypothesis test to test the manufacturer's claim.

34. Retailers report that the use of cents-off coupons in increasing. The Scripps Howard News Service reported that proportion of all households that use coupons as 0.77. Suppose that this estimate was based on a random sample of 800 households.

    a. Construct the 99% confidence interval for the true population proportion. Show all work.

    b. The manager of the retail store in reporting to his superiors claims that the true proportion of customers that use coupons is 80%. Test the manager's claim.

35. Explain the relationship between the t-test and the z-test in hypothesis testing.

36. What is meant by margin of error in a confidence interval?

37. Are girls less inclined to enroll in science courses than boys? One recent study asked randomly selected 4^{th}, 5^{th} and 6^{th} graders how many science courses they intend to take in high school. The following data was obtained:

	n	Mean	Standard Deviation
Males	203	3.42	1.49
Females	224	2.42	1.35

a. Calculate a 99% confidence interval for the difference between males and females in mean number of science courses planned. Interpret your interval.

b. The science teacher at the high school these students plan on attending claims that there is no difference in the number of courses boys and girls take. Test the science teacher's claim.

38. Techniques for processing poultry were examined by a manufacturer of canned chicken. Whole chickens were chilled 0, 2, 8 and 24 hours before being cooked and canned. To determine whether the chilling time affected the texture of the canned chicken, samples were evaluated by trained testers. One characteristic of interest was hardness. Each mean is based on 36 ratings.

Chilling Time

	0 hour	2 hour	8 hour	24 hour
Mean Hardness	7.52	6.55	5.70	5.65
Standard Deviation	.96	1.74	1.32	1.50

a. Does the data suggest that there is a difference in mean hardness for chicken chilled 0 hours before cooking and chicken chilled 2 hours before cooking? Use a significance level of 0.05.

b. Does the data suggest that there is a difference in mean hardness for chicken chilled 8 hours before cooking and chicken chilled 24 hours before cooking?

c. Use a 90% confidence interval to estimate the difference in mean hardness for chicken chilled 2 hours before cooking and chicken chilled 8 hours before cooking.

d. If a Type I error were made in part a, what would this mean? What are the consequences?

e. If a Type II error were made in part b, what would this mean? What are the consequences?

39. The discharge of industrial wastewater into rivers affects water quality. To assess the effect of a particular power plant on water quality, 24 water specimens were taken 16 km upstream and 4 km downstream of the plant. Alkalinity (mg/L) was determined for each specimen, resulting in the summary quantities in the table below.

Location	n	Mean	Standard Deviation
Upstream	24	75.9	1.83
Downstream	24	183.6	1.70

a. Does the data suggest that the true mean alkalinity is higher downstream than upstream by more than 50 mg/L? Perform a hypothesis test. Show all steps.

b. Find the 90% confidence interval for the mean difference. Does this confirm your conclusion in the hypothesis test?

40. The article "Softball Sliding Injuries" provided a comparison of breakaway bases (designed to reduce injuries) and stationary bases. Consider the accompanying data. Does the use of breakaway bases reduce the proportion of games in which a player suffers a sliding injury? Perform the test at a 1% significance test.

	Number of Games Played	Number of Games Where a Player Suffered a Sliding Injury
Stationary Bases	1250	90
Breakaway Bases	1250	20

41. The color vision of birds plays a role in their foraging behavior. Birds use color to select and avoid certain types of food. The authors of the article "Color Avoidance in Northern Bobwhites" studied the pecking behavior of 1-day-old bobwhites. In an area painted white, they inserted four pins with different colored heads. The color of the pin chosen on the birds first peck for each of 33 bobwhites, resulting in the accompanying table. Does this data provide evidence of color preference? Test at the 15 significance level.

Color	Blue	Green	Yellow	Red
First Peck Frequency	16	8	6	3

42. Do women have different patterns of work behavior than men? The article "Workaholism in Organizations: Gender Differences" attempts to answer this question. Each person in a random sample of 423 graduates of a business school in Canada were polled and classified by gender and workaholism type, resulting in the accompanying table:

Workaholism	Female	Male
Work Enthusiasts	20	41
Workaholics	32	37
Enthusiastic Workaholics	34	46
Unengaged Workers	43	52
Relaxed Workers	24	27
Disenchanted Workers	37	30

a. Test the hypothesis that gender and workaholism type are independent.

b. The author writes "women and men fell into each of the six workaholism types to a similar degree." Does the outcome of the test you performed in part a support this conclusion? Explain.

43. It is certainly plausible that workers are less likely to quit their jobs when wages are high than when they are low. The paper "Investigating the Causal Relationship Between Quits and Wages" presented the accompanying data on x = average hourly wages and y = quit rate (number of employees per 100 who left jobs during 1986.) Each observation is for a different industry.

x	8.20	10.35	6.18	5.37	9.94	9.11	10.59	13.29	7.99	5.54	7.50	6.43	8.83	10.93	8.80
y	1.4	.7	2.6	3.4	1.7	1.7	1.0	.5	2.0	3.8	2.3	1.9	1.4	1.8	2.0

The following is the Minitab output:

Predictor	Coef	Stdev	t-ratio	p
Constant	4.8615	0.5201	9.35	0.0000
Wage	-0.34655	0.05866	-5.91	0.0000

s= 0.4862 R-sq = 72.9% R-sq(adj) = 70.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	8.2507	8.2507	34.90	0.0000
Error	13	3.0733	0.2364		
Total	14	11.3240			

a. Identify the slope and y-intercept for the LSRL for average hourly wages and quit rate.

b. What is the LSRL?

c. What values do the values for slope and y-intercept model for the population?

d. Find the 95% confidence interval for the slope of the line.

e. Test the hypothesis that there is a linear relationship between average hourly wages and quit rate.