

STAT 217
Assignment #1

Note: answers may vary slightly due to rounding and whether or not you use the computer or the tables.

MINITAB INSTRUCTIONS

CALC ⇒ Probability Distributions ⇒ normal.

Finding the area above and below a Z-value under the Standard Normal Curve

For a given Z-value, you are required to find a probability. In the dialog box, which corresponds to the Normal distribution, you have three choices:

- Probability**
- Cumulative probability**
- Inverse cumulative probability**

Click **Cumulative probability**. This will calculate the cumulative probability associated with a specific Z-value (or the area under the Standard Normal Curve to the **left** of a specific Z-value.)

The middle of the dialog box has 2 options:

- Mean**
- Standard Deviation**

The default values for the **Mean** and **Standard Deviation** are **0** and **1**, respectively. There is no need to change these values, so just leave these as is.

In the bottom portion of the dialog box, select **Input constant**. It is in this box that you enter a specific Z-value. Once you have completed this, either press return or “click” on **OK**. In the upper portion of your screen, or the command module, MINITAB will return the area to the **left** of the Z-value you have entered. Note that when this routine is employed, the probability returned is **ALWAYS THE AREA TO THE LEFT OF** the Z-value entered above or $P(Z < z)$

For practice, try question 1 using this routine.

1. Given that Z is a standard normal random variable, compute the following probabilities:

- (a) $P(-0.72 \leq Z \leq 0)$ (.2642)
- (b) $P(-0.35 \leq Z \leq 0.35)$ (.2736)
- (c) $P(0.22 \leq Z \leq 1.87)$ (.3822)
- (d) $P(Z \leq -1.02)$ (.1539)
- (e) $P(Z \geq -0.88)$ (.8106)
- (f) $P(Z \geq 1.38)$ (.0838)
- (g) $P(-0.34 \leq Z \leq 2.33)$ (.6232)

Finding a Z-value for a given area (or probability under the Standard Normal Curve)

For a given probability, you are required to find a Z-value that corresponds to this probability. This requires the use of the **Inverse cumulative probability** routine in the dialog box employed above.

“Click” on the circle which corresponds to **Inverse cumulative probability** and just as was done previously, do not touch the box labeled **Mean** and **Standard Deviation**.

This routine needs an area, and will subsequently find the Z-value which matches up with the area entered. Just as was done above, move your mouse down to the bottom portion of the dialog box and “click” on the circle which corresponds to **Input constant**. Previously you entered a Z-value here. But now you want to

find a Z-value for a given area, or probability. So the number you will enter in the **Input constant** box is a probability, or an area to the left of the Z-value in question. Once you have entered the correct probability, either press return or “click” on **OK**. MINITAB will return a Z-value in the command module on the upper portion of your screen.

A good rule-of-thumb in these types of problems is to draw your standard normal curve and piece together the areas given. The Z-value will be given when you specify the area to the left of that value. Practice this routine on question 2.

2. Given that Z is a standard normal random variable, determine Z_o if it is known that:

- (a) $P(-Z_o \leq Z \leq Z_o) = 0.90$ (1.645)
- (b) $P(-Z_o \leq Z \leq Z_o) = 0.10$ (.1257)
- (c) $P(Z \geq Z_o) = 0.20$ (.842)
- (d) $P(-1.66 \leq Z \leq Z_o) = 0.25$ (-.529)
- (e) $P(Z \leq Z_o) = 0.40$ (-.253)
- (f) $P(Z_o \leq Z \leq 1.80) = 0.20$ (.720)

3. Every year around Halloween in the bustling metropolis of Red Deer, many street signs are defaced. The average repair cost per sign is \$68.00 and the standard deviation is \$12.40.

- (a) What is the probability that a defaced sign’s repair cost will exceed \$85.00? (.0852)
- (b) In one particular year, 36 signs were vandalized. What is the probability that the average repair cost of these signs is between \$67.00 and \$70.00? (0.5184)

4. According to a article in the January 1991 issue of *Health* magazine, the cost of a root canal ranges from \$200 to \$700. Suppose the mean cost for root-canal therapy is \$450 and the standard deviation is \$125. If a sample of 100 dentists from across the country was taken, what is the probability that the mean cost of a root-canal will fall between \$425 and \$475 (.9554)

5. The time a recreational skier takes to go down a downhill course has a normal distribution with a mean of 12.3 minutes and a standard deviation of 0.4 minutes.

- (a) What is the probability that the skier will take between 12.1 and 12.6 minutes to complete a run on the course? (.4649)
- (b) What is the maximum time (in minutes) the skier must have for the time to be classified “among the fastest 10% of his times”? (11.7874)
- (c) The times for a random sample of 4 of the skier’s runs is considered. What is the probability that the average time for this sample is more than 12.75 minutes? (0.0122)

6. The time it takes Bob to get from home to his office follows a normal distribution. The probability that it takes him less than 3 minutes is 0.345. The probability that it takes him more than 10 minutes is 0.01. Find the average time and variance (μ and σ^2) of this normal distribution. (4.0246, 6.5978)

7. “Your Eyes”, a daily eye ware retail store, serves an average of 14.3 customer per day. Assume that the distribution of the number of customer served per day has a standard deviation of 5.9. What is the probability that the average number of customers served per day will be:

- (a) at least 15, based on a random sample of 50 days? (.2008)
- (b) less than 14, based on a random sample of 50 days? (0.3598)

8. Electrical connectors last on average 18.2 months with a standard deviation of 1.7 months. Assume that the life of the connectors is normally distributed.

- (a) The manufacturer agrees to replace, free of charge, any connectors that fail within 17 months of installation. What percentage of the connectors can he expect to have replaced free of charge? (.2404)
 - (b) The manufacturer does not want to have to replace more than 2.5% of the connectors free of charge. What should he set the life at for free replacement? (14.868)
 - (c) This question will not be on the quiz, but could be asked as a bonus question. What is the probability that the total lifetime of 24 connectors will exceed 38 years? ($z = 2.3054$, prob. ~ 0.010)
9. A random sample of size 64 is taken from a normal population with $\mu = 51.4$ and $\sigma = 6.8$. What is the probability that the mean of the sample will
- (a) exceed 52.9? (0.0388)
 - (b) fall between 50.5 and 52.3? (0.7104)
 - (c) be less than 50.6? (0.1734)

The directions for the t distribution for MINITAB is the same as the directions for the standard normal. The only difference is that you have to also plug in degrees of freedom.

Note: If you want to calculate the mean and standard deviation of a data set,

1. input all the data into one column.
2. Click on the header Calc>Column Statistics.
3. Click on the statistic that you are interested in (mean, st.dev etc)
4. Type the column in which the data is in, in the input variable box (or click on the input variable box and then double click on the column where the data is located.
5. Hit enter or click on OK

Note: You should familiarize yourself with some of the other functions of MINITAB. Check them out. You may find some time saving techniques.

Confidence intervals

1. In developing patient appointment schedules, a medical center desires to estimate the mean time a staff member spends with each patient. How large a sample should be taken if the precision of the estimate is to be ± 2 minutes at a 95% level of confidence? How large a sample is needed for a 99% level of confidence? Use a planning value for the population standard deviation of 8 minutes. [62, 107]
2. A simple random sample of five people provided the following data on ages: 21, 25, 20, 18, and 21. Develop a 95% confidence interval for the mean age of the population being sampled. State any assumptions you must make in your method. [17.8349, 24.1651]
3. The time (in minutes) taken by a biological cell to divide into two cells has a normal distribution. From past experience, the standard deviation can be assumed to be 3.5 minutes. When 16 cells were observed, the mean time taken by them to divide was 31.2 minutes. Estimate the true mean time for a cell division using a 98 percent confidence interval. [29.1645, 33.2355]
4. Based on a random sample of 100 cows of a certain breed, a confidence interval for estimating the true mean yield of milk is given by $41.6 < \mu < 44.0$. If the yield of milk of a cow may be assumed to be normally distributed with a standard deviation of 5, what was the level of confidence used? [98.36%]
5. When 16 cigarettes of a particular brand were tested in a laboratory for the amount of tar content, it was found that their mean content was 18.3 milligrams with a standard deviation of 1.8 milligrams. Set a 90 percent confidence interval for the mean tar content in cigarettes of this brand. (Assume that the amount of tar in a cigarette is normally distributed.) [17.5111, 19.0889]
6. In 10 half-hour programs on a TV channel, Mary found that the number of minutes devoted to commercials were 6, 5, 5, 7, 5, 4, 6, 7, 5, and 5. Set a 95% confidence interval for the true mean time devoted to commercials during a half-hour program. Assume that the amount of time devoted to commercials is normally distributed. [4.8049, 6.1951]

7. A random sample of 16 servings of canned pineapple has a mean carbohydrate content of 49 grams. If it can be assumed that population is normally distributed with a variance of 4 grams, find a 98 percent confidence interval for the true mean carbohydrate content of a serving. [47.835, 50.165]
8. An archaeologist found that the mean cranial width of 17 skulls was 5.3 inches with a standard deviation of 0.5 inches. Using a 90% confidence level, set a confidence interval for the true mean cranial width. Assume that the cranial width is normally distributed. [5.0883, 5.5117]
9. It is suspected that a substance called actin is linked to various movement phenomena of non-muscle cells. In a laboratory experiment when eight fertilized eggs were incubated for 14 days the following amounts (mg) of total brain actin were obtained: 1.2, 1.4, 1.5, 1.2, 1.4, 1.7, 1.5, 1.7. Assuming that brain-actin amount after 14 days of incubation is normally distributed,
 - (a) Find a 95 percent confidence interval for the true mean brain-actin amount. [1.2890, 1.6111]
 - (b) How can we decrease/increase the error? Assume that the variability does not change from the data given above.
10. An economist wants to estimate the mean income for the first year of work for a college graduate who has had the profound wisdom to take a statistics course. How many such incomes must be found if we want to be 95% confident that the sample mean is within \$500 of the true population mean? Assume that a previous study has revealed that for such incomes, $\sigma = 6250$. [601]
11. If we want to estimate the mean weight of plastic discarded by households in one week, how many households must we randomly select if we want to be 99% confident that the sample mean is within 0.250 lb of the true population means when preliminary results show that the standard deviation is 1.067? [121]
12. Wawanesa Mutual Insurance Company wants to estimate the percentage of drivers who change tapes or CDs while driving. A random sample of 850 drivers results in 544 who change tapes or CDs while driving.
 - (a) Find the point estimate of the percentage of all drivers who change tapes or CDs while driving. [64.0%]
 - (b) Find a 90% interval estimate of the percentage of all drivers who change tapes or CDs while driving. [61.29% < p < 66.71%]
13. In a study of store checkout scanners, 1234 items were checked and 20 of them were found to be overcharges.
 - (a) Using the sample data, a confidence interval for the proportion of all such scanned items that are overcharges was found to be from 0.00915 to 0.02325. What was the level of confidence that was used? [~95% level of confidence]
 - (b) Find the sample size necessary to estimate the proportion of scanned items that are overcharges. Assume that you want 99% confidence that the estimate is in error by no more than 0.005.
 - (i) Use the sample data as a pilot study [4228]
 - (ii) Assume, instead, that we do not have prior information on which to estimate the value of \hat{p} . [66307]
14. A hotel chain gives an aptitude test to job applicants and considers a multiple-choice test question to be easy if at least 80% of the responses are correct. A random sample of 6503 responses to one particular question includes 84% correct responses. Construct the 99% confidence interval for the true percentage of correct responses. Is it likely that the question is really easy? [82.83% < p < 85.17%, yes]

STAT 217

Assignment #2

Confidence intervals, Hypothesis Tests, Type I, Type II error

Note: Answers may vary slightly due to rounding.

Each time you perform a hypothesis test, state any assumptions that must be made for your method; find the p-value of your test statistic; indicate whether your decision could result in a Type I or a Type II error.

- The manufacturer of an over-the-counter pain reliever claims that its product brings pain relief to headache sufferers in less than 3.5 minutes, on average. In order to be able to make this claim in its television advertisements, the manufacturer was required by a particular television network to present statistical evidence in support of the claim. The manufacturer reported that for a random sample of 50 headache sufferers, the mean time to relief was 3.3 minutes and the standard deviation was 66 seconds.
(a) Does this data support the manufacturer's claim? Test using a 5% significance level. [-1.2856, NO]
(b) In general, do large p-values or small p-values support the manufacturer's claim? Explain

- The maximum acceptable level for exposure to microwave radiation is an average of 10 microwatts per square centimeter. It is feared that a large television transmitter may be polluting the air nearby by pushing the level of microwave radiation above the safe limit.
(a) Set up the null and alternative hypotheses needed to gain evidence to support his contention.
(b) The following random sample of 9 observations on \bar{X} , the number of microwatts per square centimeter, taken at locations near the transmitter:

9 11 14 10 10 12 13 8 12

Can the null hypothesis be rejected at a 0.10 significance level? What practical conclusion can be drawn? [1.5492, yes]

- The average total blood protein in a healthy adult is 7.25 grams per decilitre. A series of 8 blood tests were run on a particular patient over several days giving the following results:

7.23 7.25 7.28 7.29 7.32 7.26 7.27 7.24

At a 5% significance level, do these test results indicate this patient has the total blood protein level of a healthy adult? [1.6951, yes]

- The effects of drug and alcohol on the nervous system have been the subject of considerable research. Suppose a research neurologist is testing the effects of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to a neurological stimulus, and recording its response time. The neurologist knows the response time for rats not infected with the drug (the control mean) is 1.2 seconds. She wishes to test whether the response time for drug-injected rats differs from that control mean. Her sample of 100 rats give a mean of 1.05 seconds and a standard deviation for 0.5 seconds. Perform a test of her hypothesis at a 1% significance level. [-3.0, RHo]

- In a survey, 1039 adults were asked "How much respect and confidence do you have in the public school system?" The results, reported in the Toronto Star (Sept. 26, 1988), are shown below:

Responses	A great deal	Quite a lot	Some	Very little	No opinion
Percentages	12%	30%	35%	13%	10%

Estimate with a 90% confidence the proportion of all adults who had "a great deal" or quite a lot" of respect for the public school system. Interpret this interval [0.3948, 0.4452]

- Of the 200 individuals interviewed, 80 said they were concerned about fluorocarbon emissions in the atmosphere. Obtain a 99% confidence interval estimate for the true proportion of individuals who are concerned. [0.3108, 0.4892] Interpret this interval.

- As part of a study of post-secondary education, a random sample of the graduating classes of colleges and universities is to be selected to estimate their expected success in finding employment. It is desired to estimate the success rate to within ± 0.01 , with a confidence of 95%. No reliable planning value for the success rate is available.

- (a) What is a conservatively large sample size to meet the precision requirements? [9604]
- (b) It was finally decided to select 2500 graduating students for the sample. Of these, 1141 were successful in finding immediate employment. Estimate the true success rate of this graduating class in a 99% confidence interval. Interpret the meaning of the interval. [.4307, .4821]

8. The reputations (and hence the sales) of many businesses can be severely damaged by shipments of manufactured items that contain an unusually large percentage of defectives. A manufacturer of alkaline batteries wants to be reasonably certain that fewer than 5% of its batteries are defective. Suppose 300 batteries are randomly selected from a very large shipment. Each is tested and 10 defective batteries are found. Will this sample provide sufficient evidence to the manufacturer that this shipment will be satisfactory at a 1% significance level? [-1.3272, No]

9. A marketing research organization wishes to estimate the proportion of television viewers who watch a particular prime-time situation comedy on December 14. The proportion is expected to be approximately 0.30. At a minimum, how many viewers should be randomly selected to ensure that a 95% confidence interval for the true proportion of viewers will have a width of at most 0.01? [32270]

10. A sporting goods manufacturer who produces both white and yellow tennis balls claims that more than 75% of all tennis balls sold are yellow. A marketing study of the purchases of white and yellow tennis balls at a number of stores showed that of 470 cans sold, 410 were yellow and 60 were white.

(a) Is there sufficient evidence to support the manufacturer's claim at a 1% significance level? [6.1213, yes]

(b) Calculate the probability of a Type II error if, in fact, 80% of the tennis balls sold are yellow. [.4247]

11. Consider the following hypothesis test:

$$H_0: \mu = 2400$$

$$H_a: \mu > 2400$$

The rejection region has been defined as $R: \text{Sample mean} > 2446.5$

Complete the table below indicating whether the given (population mean, sample mean) pair would result in acceptance or rejection of the null hypothesis and which type of error, if any, would result.

Population Mean	Sample Mean	Ho True or False	Accept or Reject Ho	Type of Error
2401	2400			
2401	2450			
2400	2450			
2400	2401			

12. Management of a shopping centre believes that on weekends people spend more than an hour and a half on average at the centre. To see if this is the case, a parking survey will be taken for a random sample of 100 cars on weekends. It is assumed that the standard deviation of parking times is 30 minutes. Assume that testing is carried out at the 2.5% significance level. Determine the power of the test if the true mean parking time (for the population) is 97 minutes. (0.6443)

13. The manufacturer of Grin toothpaste claims that children under the age of 10 years that use their toothpaste regularly have, on average, less than 2 cavities. A random sample of 25 children had an average of 1.94 cavities with a sample standard deviation of 0.13. Assume cavity rate is normally distributed. Do the data support the manufacturer's claim?

(a) Carry out the test with a 5% significance level [t=-2.3077, Rho]

(b) Determine the p-value for the test in (a) [0.0150]

(c) Using a 5% significance level, find the probability of making a type II error if the true mean for cavities is 1.96. [0.5678]

14. A random sample of 324 adults shows that 120 smoke. At the 1% significance level

(a) test the claim that more than 1/3 of all adults smoke. [z = 1.4167, Aho]

- (b) What's the probability of making a type I error? (0.01)
- (c) Using a 1% significance level, find the power of the test if the true population proportion was 40%. [0.5932]

15. Assume that you are using a significance level of $\alpha = 0.05$ to test the claim that $\mu < 2$ and that your sample consists of 50 randomly selected values.

- (a) Find β given that the population actually has a normal distribution with $\mu = 1.5$ and $\sigma = 1$. [0.0294]
- (b) Find the power of the test and describe what it means. [0.9706]

STAT 217
Assignment #3
Chi-squared, F-test, and two sample hypothesis testing

1. Suppose that a machine dispenses sand into bags. A random sample of 100 bags is taken from a new machine and the standard deviation is 14kg. The population standard deviation is known to be 18kg. Test at the 5% significance level that the new machine is better. [59.8889, reject]

2. Last year, the mean number of books borrowed per cardholder at a major university was 18.2 books per semester with a st.deviation of 4.2. A random sample of 25 cardholders showed the following results for this semester: $s^2=6.17$. The library administration would like to know whether this semester's variance has changed from last semester's at the 1% significance level. What assumptions were made?

[8.39, reject]

3. Business schools A and B reported the following summary of GMAT verbal scores:

	N	\bar{x}	s^2
A	11	34.75	48.59
B	13	33.74	44.68

- (a) At the 5% significance level, is the variance of school A greater than school B? [1.0875, Aho]
 (b) Construct a 95% confidence interval for the true population standard deviation ratio. Interpret. [0.5681, 1.9841]

4. 16 alkaline and 16 heavy-duty Radio shack batteries were placed individually in a circuit consisting of two flashlight bulbs wired in parallel, a switch, a battery holder, and a Hewlet Packard 427-A analog DC voltmeter. Each battery was drained to a reference failure voltage of 0.9 volts and the time to failure (min) was measured.:

Alkaline				Heavy-duty			
105	141	147	158	29	22	22	27
140	143	108	125	26	17	22	23
116	139	146	134	23	27	23	24
140	149	142	140	22	25	22	26

Assume the populations of times to failure is normally distributed.

- (a) Test the hypothesis at the 0.05 level that the two types of batteries have the same population variances. [$F_{calc}=26.9141$, Rho]
 (b) Find the 95% confidence interval for the ratio of standard deviations. [3.0677, 8.7729]

5. Two marathon training procedures are tried for comparison purposes. Their efficacy is to be determined in a marathon race. Assume race times are normally distributed with unequal variances.. The follow results were observed:

	No. In sample	mean race time	s.d of race times
Procedure 1	15	150 min	12 min
Procedure 2	26	170 min	15 min

- (a) Test that procedure 2 is greater than procedure 1 at $\alpha = 0.025$ [4.6812, Rho]
 (b) Find a 90% confidence interval for the difference in the mean race times. [12.4762, 27.5238]
 (c) Find a 90% confidence interval for the difference in the mean race times assuming population variances are the same. [12.3519, 27.6281]

6. A large brokerage house wants to see if the percent of new accounts valued at over \$50,000 has changed over the year. A random sample of 900 accounts opened last year showed that 27 were over \$50,000 in size. A random sample of 1,000 accounts opened this year showed 44 that were in excess

of \$50,000. Does that data indicate that there has been a change in the percent of new accounts valued at over \$50,000?

- (a) Carry out the appropriate statistical test at a 1% significance level. [z=-1.6231, Aho]
- (b) Determine the minimum level of significance to conclude that the data indicates that there has been a change. [p-value=0.1046]

7. The occupational Health and Safety Act is not popular with Management because of the cost of implementing its requirements. However, some sources claim it has been effective in reducing industrial accidents. Listed are the man-hours lost due to accidents at 6 randomly selected plants.

Plant number	1	2	3	4	5	6
Before OHSA	38	64	42	70	58	30
After OHSA	31	58	43	65	52	29

- (a) Is there sufficient evidence to indicate that The Occupational Health and Safety Act has been effective in reducing lost-time accidents? Test at $\alpha = 0.10$. [t=3.038, rHo]
 - (b) Find the P-value in (a). [0.0144]
 - (c) What assumptions did you make in (a)? [random sample from a normal population]
8. Let p_1 denote the percentage of people who were unemployed in March, and let p_2 denote the percentage of people who were unemployed in August. Suppose that during late March, the government instituted policies designed to lower the unemployment rate. We want to test whether the policies were effective at $\alpha = 0.05$. In March, in a random sample of 1000 people, 75 were unemployed. During April, in an independent random sample of 1000 people, 65 were unemployed. What is your conclusion? [z=0.8765, Aho]
9. A producer of barley wanted to know which of two fertilizers, A or B, was better for growing barley. To study this, seven plots of land were selected. Each plot was divided in half. For each plot, fertilizer A was assigned to one half and fertilizer B to the other half using the toss of a coin. The yield of barley in bushels per acre after a period of time gave the following data:

plot	1	2	3	4	5	6	7
A	42	38	41	43	42	38	41
B	40	42	36	39	41	40	36

Does the above data indicate one of the fertilizers is more effective than the other? Use $\alpha = 0.05$
[T=1.186, Aho]

10. A USA Today study reported the longest average workweeks for non-supervisory employees in private industry to be mining (45.4 hours) and manufacturing (42.3 hours). The same article reported the shortest average workweeks to be retail trade (29 hours) and services (32.3 hours). A study conducted in the state of Illinois found the following results:

Industry	N	Ave hours per week	Standard deviation
Mining	15	47.3	5.5
Manufacturing	10	43.5	4.9

Assuming that there is a statistically significant difference in the variability of hours per week worked in the mining and manufacturing industries

- (a) Construct a 95% confidence interval estimate of the mean difference in hours per week worked for non-supervisory employees. Interpret this interval. [-.9543, 8.5543]

- (b) Does the Illinois data suggest that non-supervisory employees in the mining industry, on average, work more hours per week than their manufacturing industry counterparts? Regulate the probability of type I error at 5 percent. [T=1.808, Rho]
- (c) What is the p-value of your results in (b)? [0.0428]

11. Of a random sample of 100 stocks on the Toronto Stock Exchange, 32 made a gain today. A random sample of 100 stocks on the Montreal Stock Exchange showed 27 stocks making a gain.

- (a) Construct a 95% confidence interval, estimating the difference in the proportion of stocks making a gain on the two exchanges. [-0.0762, 0.1762]
- (b) Does the data suggest there is a statistically significant difference between the proportion of stocks making gains on the two exchanges? Use $\alpha = 0.05$. [z=0.7764, Aho]

12. The residents of Cardston have complained for many years that they pay higher traffic fines than the residents of their adversary, Raymond. A local newspaper hired a statistician to determine whether there was any merit to this argument. The statistician randomly selected 25 residents from Cardston and 25 residents from Raymond, all of whom had received a speeding ticket. The fine paid by each Cardston and Raymond resident were recorded. The summary statistics are as follows:

Town	Mean	Standard deviation
Cardston	\$60.00	\$12
Raymond	\$50.00	\$10

- (a) Construct a 95% confidence interval, estimating the mean difference in traffic fines paid by Cardston residents and Raymond residents. [3.5519, 16.4481]
- (b) Using the p-value approach, does the data provide sufficient evidence to suggest that Cardston residents pay more, on average, in traffic fines than Raymond residents? Use $\alpha = 0.05$. [0.00121, Rho]

13. A researcher wanted to know if the percentage of drinking drivers increased during later evening hours. She decided to compare the percentage of drinking drivers from 7:00 to 9:00 pm with the percentage from 10:00 to 12:00am. The randomly selected 250 drivers from the first time period and 200 from the second time period. Each driver in both samples was given a breath test and assigned to the category drinking driver or non-drinking driver. From the early period, 12 percent were drinking drivers, and 19 percent from the later period. Does this data suggest that there is a statistically significant increase in the percentage of drinking drivers in the later time period? Use $\alpha = 0.05$. Determine the p-value of your results. How convincing (or unconvincing) is the data?

[z=-2.0276, Rho, 0.0197]

14. In a province wide poll of 2000 men and 2010 women, 980 and 1025 women report that they are opposed to the death penalty in ALL circumstances.

- (a) Find a 90% confidence interval estimate for the difference in the proportion of men opposed to the death penalty and corresponding proportion of women. [-0.046, 0.006]
- (b) Test the claim at a level of significance of 5% that men are less likely to favour the death penalty than women. What is your conclusion? [z=1.26, Aho]

15. Independent random samples taken at two local liquor stores provide the following information regarding customers' purchases.

Store	Bob's Booze	Len's Liquor Emporium
N	46	39
Mean	\$52.40	\$61.75
s ²	\$9.50	\$7.25

Assuming that there does exist a difference in the variability between the two liquor stores, does the data indicate that the average purchases at Len's exceed those at Bob's by more than \$5, on average? Test at $\alpha = 0.05$. [z=6.9441, Rho]

Stat 217 Assignment #4

One-Way ANOVA

A firm developing a new citrus-flavoured soft drink conducted an experiment to study customer preferences for the colour of the drink. Four colours were considered: colourless, pink, orange, and lime green. Twenty test localities, which were similar in sales potential and representative of the target market for this product, were selected. Each colour was randomly assigned to five localities for test marketing. The number of cases sold per 1000 population during the test period are recorded below:

Colourless	Pink	Orange	Lime Green
26.5	31.2	27.9	30.8
28.7	28.3	25.1	29.6
25.1	30.8	28.5	32.4
29.1	27.9	24.2	31.7
27.2	29.6	26.5	32.8

Are the sales the same for all colours of the drink? Test at the 5% significance level.

Minitab

Method 1

1. Enter the data into the MINITAB worksheet one sample per column- *ie.* The 5 values from the 'colourless' sample are entered in c1, the 5 values of the 'pink' sample in c2, and so forth.
2. From the MENU BAR, select STAT>ANOVA>ONEWAY(UNSTACKED)
When the dialog box appears, list all the columns in which you entered data.
3. Note the table and other details are printed in the SESSION window.

Method 2

1. Enter all the sample values into one column of the worksheet. In a corresponding position in a second column, enter the number of the sample (1,2,3, etc.) from which each value came. For instance, if you entered all the 'colourless' observations, then pink, then orange, then lime green in column c1, c1 would contain 26.5, 28.7, 25.1, 29.1, 27.2, 31.2, 28.3, ... 32.8. You would enter into c2: 1, 1, 1, 1, 1, 2, 2, ..., 4. NOTE: These entries in c2 are most efficiently accomplished with CALC>MAKE PATTERNED DATA>SIMPLE SET OF NUMBERS. In the dialog box, store patterned data in c2. FROM 1st VALUE box type **1**, TO LAST VALUE box type **4**, LIST EACH VALUE box type **5**, LIST THE WHOLE SEQUENCE box type **1**.
2. From the MENU BAR, select STAT>ANOVA>ONEWAY
As the RESPONSE, specify the column in which you entered the observations, as the FACTOR specify the column in which you placed the sample number.
3. The output appears in the SESSION window.

1. A large consulting firm hires a West Coast university to provide an M.B.A program for its employees. The basic statistics course is taught at four locations of the firm. After completion of the course, standardized tests are given to the participating employees at each location. The results are:

Observation	Location			
	A	B	C	D
1	96	65	66	60
2	88	74	90	72
3	92	77	88	66
4	75	82	73	75
5	81	70	85	78
6	62	78	87	68
7	86	84	94	
8	86		74	
9	90			
10	85			

Can it be concluded that there is no difference in test results between location? Use $\alpha = 0.05$
 { $F_{calc} = 4.130 > 2.9680$, Rho}

2. The sales research division of a large corporation is conducting research on sales methods for selling one of the products of the corporation. The division has designed a completely randomized one-factor analysis of variance model to investigate the efficiency of three sales methods. The responses are measured in units of \$100 sales, and are listed below.

Response	Sales Method		
	A	B	C
1	20	13	31
2	23	18	28
3	22	16	48
4	22	29	28
5	36	14	30
6	45		25
7	21		
8	26		

Are there differences between the three different sales methods? Use $\alpha = 0.10$. What assumptions did you make? { $F_{calc} = 3.899 > 2.668$, Rho, Sales are normally distributed and share a common variance}

3. A county employs 3 assessors who are responsible for determining the values of residential property in the county. To see whether or not these assessors differ in their appraisal, 5 houses were selected and each assessor determined the market value of each house. The data (assessors are the treatment since your main concern is to see if there is difference in appraisals) was then analyzed using a two-way ANOVA routine giving the following (partially completed) ANOVA table.

(a) Complete the ANOVA table

Source	SS	DF	MS	F
Treatment	45.9			
Block	141.7			
<u>Error</u>				
Total	250.8			

- (b) Is there any indication of a difference between appraisors? Use a 5% significance level for testing.
 { $F_{calc} = 2.905 < 4.459$, Aho}

Two-Way ANOVA

The data in the following table represent the milliequivalents of sodium excreted by six subjects 2 hours after treatment with one of four diuretics assigned at random by a clinician over a 6-day period. Using

significance level of 0.05, analyze the data to determine whether or not there are any differences between patients and any differences in the effectiveness of diuretics.

Subjects	Treatments (Diuretics)			
	A	B	C	D
1	3.9	30.6	25.2	4.4
2	5.6	30.1	33.5	7.9
3	5.8	16.9	25.5	4.0
4	4.3	23.2	18.9	4.4
5	5.9	26.7	20.5	4.2
6	4.3	10.9	26.7	4.4

{F_B = 1.5758 < 2.90, Aho depending on rounding}
(F_{Tr} = 37.153 > 3.287, Rho)}

Method

- Enter all of the data into one column. You should first enter all the values for treatment A, then treatment B, and so forth.
- In the second column, enter from which treatment each value came from (*ie.* You would enter A, A, A, A, A, A, B, B, ..., D) These entries are most efficiently accomplished with CALC>MAKE PATTERNED DATA > TEXT VALUES
Store patterned data in **C2**
In the dialog box, type A B C D (Note: leave a space between each letter)
In the list each value box, type **6**
In the list the whole sequence box, type **1**.
- In the third column, we want to enter the subject number (*ie.* 1,2,3,4,5,6,1,2,3,4,5,6,...
Go to CALC>MAKE PATTERNED DATA>SIMPLE SET OF NUMBERS
 - Store patterned data in **C3**
 - From 1st value, type **1**
 - To last value, type **6**
 - In steps of, type **1**
 - List each value, type **1**
 - List the whole sequence, type **4**

This will assign the numbers 1 through 6 four times.

- From MENU BAR select STAT>ANOVA>TWO-WAY
 - Response, type **C1**
 - Row factor, type **C2** (MINITAB takes row factor to represent treatment effect)
 - Column factor, type **C3** (MINITAB takes column factor to represent block effect)
Note: do not take "row factor" and "column factor" literally
- Output appears in the SESSION window.
- For the following data, present the ANOVA table. What conclusions can you draw from the two F tests? Use an $\alpha = 0.05$

Treatment	Block			
	1	2	3	4
1	14	8	10	4
2	7	7	4	2
3	12	6	16	6

{F_B = 3.231 < 4.7571, Aho}
{ F_{Tr} = 3.462 < 5.1433, Aho}

- Three loaves of bread, each made according to a different recipe, are baked in one oven at the same time. Because of possible uncontrolled variations in oven performance, each baking is treated as a block. This procedure is repeated five times and the following measurements of density are obtained.

Recipe	Block				
	1	2	3	4	5

1	.95	.86	.71	.72	.74
2	.71	.85	.62	.72	.64
3	.69	.88	.51	.73	.44

- (a) How should the three oven positions of the three loaves be selected for each trial?
{Randomize the position of the loaves in the oven}
- (b) Perform an analysis of variance for these data using a 5% significance level
{FBI = 5.31 > 3.837, Rho}
{FTr = 3.92 < 4.46, Aho}
- (c) Find the p-value for the above Fcalculated value for Treatment.

Non-Parametric Tests

6. In 1995, the median age of Canadian residents was 34 years, as reported by the Census Bureau. A random sample taken this year of 10 Canadian residents yielded the following ages, in years,

40 60 12 55 34 43 47 37 9 34

At the 5% significance level, do the data provide sufficient evidence to conclude that the median age of today's Canadian residents has increased over the 1995 median age of 34? {W+ = 36.5, Aho}

7. Twenty years ago, the U.S. Bureau of Justice Statistics reported that the mean educational attainment of jail inmates was 10 years. Ten current inmates are randomly selected and found to have the following educational attainments, in years,

14 10 5 6 8 10 10 8 9 9

Assume that the educational attainments of current jail inmates have a symmetric, non-normal distribution. At the 10% significance level, do the data provide sufficient evidence to conclude that this year's mean educational attainment has decreased? {W+ = 9, Rho}

8. Several batches of fruit flies are exposed to each treatment, and the mortality percent is recorded as a measure of toxicity. The following data are obtained:

Treatment 1	40	28	31	38	43	46	29	18
Treatment 2	36	49	56	25	37	30	41	

Determine if the data strongly indicate different toxicity levels among the treatments at $\alpha = 0.05$ (assume non-normal distributions, but similar distributions). {W1=62, Aho}

9. Two critics rate the service at seven award winning restaurants on a continuous 0 to 10 scale. Is there a difference between the critics' ratings at a 0.05 significance level? {W+ = 23, or W+ = 5, Aho}

Restaurant	1	2	3	4	5	6	7
Critic 1	6.1	5.2	8.9	7.4	4.3	9.7	5.5
Critic 2	7.3	5.5	9.1	7.0	5.1	9.8	5.7

10. 13 people were given a pill. Their blood pressure was measured before and after they took a pill. At the 5% significance level, determine if blood pressure has decreased after taking the pill. {W+ = 83, right tailed test or W+ = 9.5, left tailed test, Rho}

Before	70	80	72	76	76	76	72	78	82	64	74	92	74
After	68	72	62	70	58	66	68	52	64	72	74	60	74

11. Use the appropriate nonparametric method to perform the test in question 1. (KW=9.68>7.815, Rho)
12. Use the appropriate nonparametric method to perform the test in question 2. (KW=7.56>4.605, Rho)

STAT 217
Assignment #5

Goodness of fit test.

Computer instruction

Enter observed values in column 1 and expected values in column 2.

Click on Calc<Calculator.

Type c3 in the box Store Results.

Type the formula $(c1-c2)**2 / c2$ in the large box. Hit enter.

Click on Calc<column statistics

Click on sum.

Type c3 in the input variable box.

1. An official of a plastics industry claimed that the industry employed 30% white women, 5% minority women, 50% white men, and 15% minority men. To test the claim, an affirmative action committee randomly sampled 150 employees and obtained the following information:

Category	observed
White females	40
Minority females	15
White males	80
Minority males	15

Test the official's claim at a 5% level of significance [10.89, Rho]

2. A computer science major claimed to have written a program that would randomly generate integers from 1 to 100. The program generated the following data. Use a 5% level of significance to test the claim. [6.40, Aho]

Integers	Observed
1-10	6
11-20	6
21-30	13
31-40	9
41-50	13
51-60	11
61-70	8
71-80	12
81-90	10
91-100	12

3. Given below are the frequencies observed from 310 tosses of a die. Do these data cast doubt on the fairness of the die at the 5% significance level? [13.225, Rho]

Face No.	1	2	3	4	5	6
Frequency	38	61	54	65	55	37

4. A shipment of assorted nuts is labeled as having 45% walnuts, 20% hazelnuts, 20% almonds, and 15% pistachios. By randomly picking several scoops of nuts from this shipment, an inspector find the following counts.

Counts	Walnuts	Hazelnuts	Almonds	Pistachios	Total
	92	69	32	42	235

Could these findings be a strong basis for an accusation of mislabeling? Test at the 5% significance level. [18.165, Rho]

Chi-Square Tests of Independence

Minitab will perform all necessary calculations for chi-square tests on contingency tables, presenting the expected values, the value of the test statistic, degrees of freedom and the p-value.

1. Enter the observed frequencies into rows and columns just as they are given in the contingency table.
 2. Go to the main header and click on **STAT>Tables>Chi-Square Test**
 3. In **Columns containing the tables**, enter the columns at which your contingency table is contained.
 4. Click on **OK**.
1. Over the years pollsters have found that the public's confidence in big business has been closely tied to the economic climate of the country. When businesses are growing and employment is increasing public confidence is high. When the opposite occurs, public confidence is low. In one study, Harvey Kahalas (1981) explored the relationship between confidence in big business and job satisfaction. He hypothesized that there is a relationship between the level of confidence and job satisfaction and that this relationship holds true for both union and nonunion workers. To test his hypotheses he used the sample data given in the tables below:

Union Members

Job Satisfaction

Confidence in Major Corporations	Very Satisfied	Moderately Satisfied	Little dissatisfied	Very Dissatisfied
A great deal	30	19	6	6
Only some	99	77	20	9
Hardly any	38	32	14	15

NonUnion Members

Job Satisfaction

Confidence in Major Corporations	Very Satisfied	Moderately Satisfied	Little dissatisfied	Very Dissatisfied
A great deal	111	52	13	5
Only some	246	142	37	18
Hardly any	73	51	19	9

Perform a hypothesis test on each of these data sets at the 5% significance level. Does the data support Kahalas's theory? [13.359, Rho] [8.298, Aho]

2. A personnel administrator provided the following data as an example of hiring to fill 12 positions from among 40 male and 40 female applicants.

Applicant	Selected	Not Selected	Total
Male	7	33	40
Female	5	35	40

Does this sample indicate a selection bias in favour of males? Use a p-value in your conclusion. [0.392, 0.5312]

3. Applicants for public assistance are allowed an appeals process when they feel unfairly treated. At such a hearing, the applicant may choose self-representation or representation by an attorney. The appeal may result in an increase, decrease, or no change of the aid recommendation. Court records of 320 appeals cases provided the following data.

Amount of Aid

Type of Representation	Increased	Unchanged	Decreased
Self	59	108	17
Attorney	70	63	3

Are the patterns of the appeals decision significantly different between the two types of representation? Test at $\alpha = 0.05$ [15.734, Rho]

4. A survey was conducted by sampling 400 persons who were questioned regarding union membership and attitude toward decreased national spending on social welfare programs. The cross-tabulated frequency counts are presented.

	Support	Indifferent	Opposed
Union	112	36	28
NonUnion	84	68	72
Total	196	104	100

Can these observed differences be explained by chance or are there real differences of attitude between the populations of members and non-members at the 5% significance level? [27.847, Rho]

5. In a Study of possible genetic influence of parental hand preference, a sample of 400 children was classified according to each child's handedness and the handedness of the biological parents. Do these findings demonstrate an association between the handedness of parents and their biological offspring at the 5% significance level? [10.653, Rho]

Handedness of Biological Offspring

Parents' Handedness	Right	Left	Total
Father x Mother			
Right x Right	303	37	340
Right x Left	29	10	39
Left x Right	16	6	22
Total	348	52	401

6. In a genetic study of chromosome structures, 132 individuals are classified according to the type of structural chromosome aberration and carriers in their parents. The following counts are obtained.

Carrier

Type of Aberration	One Parent	Neither Parent	Total
Presumably innocuous	28	19	47
Substantially unbalanced	35	50	85
Total	63	69	132

Test the null hypothesis that type of aberration is independent of parental carrier. Use p-value [4.106, Rho]

The F-distribution and Simple Linear Regression

Note: there may be slight differences in the answers due to rounding.

1. Here is a set of data showing the historic yearly rates of return in seven randomly selected years, for Stock Y and the New York Stock Exchange Index (the predictor variable).

Year	1	2	3	4	5	6	7
Stock Y	2.0%	7.9%	-6.0%	-9.5%	13.5%	7.5%	1.2%
NYSE Index	4.9%	13.0%	-2.5%	-10.6%	11.0%	14.5%	4.3%

- (a) Write down the linear regression model expressing the yearly rate of return on Stock Y as a linear function of the yearly rate of return of the NYSE Index.
- (b) Estimate the intercept and slope term in the model. (Note: the slope term is referred to as Stock Y's "beta", or β . This is a measure that stock analysts use to evaluate the past performance of a stock. Stocks possessing β 's greater than 1 tend to have larger expected rates of return compared to stocks with smaller β 's. [$\beta_0 = -1.7470$, $\beta_1 = 0.8332$])

Minitab instructions

1. Enter Stock Y data in column C1
2. Enter NYSE Index data in column C2
3. Click on **STAT>Regression>Regression**
4. Enter C1 in the **Response** box
5. Enter C2 in the **Predictor** box
6. Click on **Graphs**, click on **residuals versus fits**, click **OK**
7. Click **OK**

You will now get a graph of the residuals vs fits for the data. There will also be a printout of the regression equation and the ANOVA table on the screen. If you want to see a scatter plot of the data with the fitted line,

Click on **STAT>Regression>Fitted Line Plot**

Enter C1 in **Response (Y)**

Enter C2 in **Predictor (X)**

Highlight **Linear** for **Type of Regression**.

Click **OK**.

- (c) Construct a 95% confidence interval estimate for Stock Y's β (β_1). Interpret the meaning of this interval. [$0.4560 \leq \beta_1 \leq 1.2105$]
- (d) Is the rate of return on Stock Y positively related to the rate of return on the NYSE Index? Test at a level of significance of 0.05. [$T=5.676$, reject the null hypothesis]
- (e) Construct an analysis of variance table for the above regression. In addition, perform the same test in (d) using a different test. (again, $\alpha = 0.05$). Are the results in (c) and (d) consistent?
- (f) Find the standard error of the regression and interpret its significance. [$Se^2 = 10.563$]
- (g) Find the coefficient of determination and interpret its meaning. [$r^2 = 0.8656$]
- (h) Find a 94% confidence interval estimate for the mean rate of return on Stock Y if the rate of return on the NYSE Index is 4.6%. [$-0.8916, 5.063$]
- (i) Find a 99% confidence interval estimate for this year's rate of return on Stock Y if the New York Stock Exchange Index has a rate of return of 8.1%. (or a 99% prediction interval). Interpret this interval. Would you invest in this stock, based on your interval? [$-9.1316, 19.1354$]
- (j) Find the coefficient of correlation between the rate of return on Stock Y and the rate of return on the NYSE [$r = +0.930$]
2. The Director of Management Information Systems at a conglomerate must prepare his long-range forecasts for the company's 3-year budget. In particular, he must develop staffing ratios to predict the

number of managers and project leaders based on the number of programmers. The results of a sample of the electronic data processing staffs of 10 companies within the industry are displayed below.

# of applications Programmers	15	7	20	12	16	20	10	9	18	15
# of Managers and Project leaders	6	2	10	4	7	8	4	6	7	9

- Find the regression coefficients. State the least squares linear regression equation. [$\hat{y} = -0.0885 + 0.45x$]
- Interpret the meaning of the slope and intercept.
- Compute Se^2 and interpret this value. [$Se = 1.42$]
- Compute the coefficient of determination and interpret its meaning in this problem. [$r^2 = 0.7018$]
- At the 0.05 level of significance, is there a linear relationship between the number of managers and the number of application programmers? Use T-test [$T = 4.339$, Rho]
- At the 0.05 level of significance, test for the appropriateness of the simple linear regression model. Use F-test [$F = 18.834$, Rho]
- Set up a 95% confidence interval estimate of the true population slope. [$0.2109 \leq \beta_1 \leq 0.6891$]
- Set up a 95% confidence interval estimate of the true population intercept. [$-3.6374 \leq \beta_0 \leq 3.4604$]
- Set up a 95% confidence interval estimate of the average number of managers at companies where there are 10 programmers. [$2.9711, 5.8559$]
- Set up a 95% prediction interval estimate of the number of managers for a particular company in which there are 10 programmers. [$0.8354, 8.1716$]
- Construct a residual plot of the above data. What can you conclude from this residual plot? Does the linear model seem appropriate? Explain.

3. High salaries for presidents and high executives of charitable organizations have been in the news from time to time. Consider the information in the table below for the United Way in 10 major cities in Canada.

<u>City</u>	<u>Salary of President</u>	<u>Money Raised (per capita)</u>
Ottawa	\$161,396	\$17.35
Montreal	\$189,808	\$15.81
Toronto	\$201,490	\$16.74
Winnipeg	\$171,798	\$31.49
Halifax	\$108,364	\$15.51
St. John's	\$126,002	\$23.87
Regina	\$146,641	\$15.89
Saskatoon	\$155,192	\$9.32
Edmonton	\$169,999	\$29.84
Vancouver	\$143,025	\$24.19

- Find the least-squares regression equation that expresses the presidents' annual salary as a linear function of the amount of money raised (per capita). Interpret the meaning of the slope term in the context of the question. [$\hat{y} = 152657 + 235.71x$]
- This past year, the City of Lethbridge (with a population of approximately 70,000) raised a total of 1.9 million dollars. Estimate the salary of the president of the United Way Lethbridge Chapter. [$x = 27.14$, $\hat{y} = \$159,024.95$]
- Find the ANOVA table. What percentage of the variation in presidents' salary is explained by the fact that some raised more money per capita than others? [$r^2 = 0.0035$, very small]
- Is there a significant linear relationship between the president's salary and per capita money raised? Use the p-value approach and interpret the p-value in the context of the question. [$t = 0.1677$, Aho]
- Does there appear to be a significant linear relationship between the amount of money raised per capita and the presidents' salary? Conduct this test using both the t-test and the F-test. What is your conclusion? [$F = 0.0281$, Aho]

- (f) This past year the United Way in Calgary raised \$34.94 per capita. Construct a 99% confidence interval for the mean salary of the president of the United Way in Calgary.
[\$107,862.93, \$213,922.71]
- (g) Construct a 95% prediction interval for the salary of the president of the United Way in Calgary.
[\$74,227.12, \$247,588.52]
- (h) Estimate, with 90% level of reliability, the average (mean) salary of a United Way president who raised \$29.00 per capita.
[\$130,189.84, \$188,795.56]

4. The following is a MINITAB output for a random sample of 8 employees at Tackey Toy Manufacturing Company. The company wanted to see if there was a relationship between aptitude test results and output (dozens of units produced).

The regression equation is
Output = 1.03 + 5.14 test results

Predictor	Coef	St.Dev	T	P
Constant		2.070		
Test result		0.2831		

S=1.695 R-sq= R-sq(adj)=97.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression					
Residual Error					
Total		968.00			

- (a) Fill in the above tables and find R-sq.
- (b) State the least squares regression equation. [$\hat{y} = 1.03 + 5.14x$]
- (c) What percentage of the variation in output is explained by the fact that some had higher test results on the aptitude tests than others? [$r^2 = 0.982$, very large]
- (d) Is there a significant linear relationship between output and aptitude test results? Use the p-value approach and interpret the p-value in the context of the question. [$t=18.16$, Rho, p-value~0]
- (e) Does there appear to be a significant linear relationship between the aptitude test results and output? Conduct this test using both the t-test and the F-test. What is your conclusion? [$t=18.16$, $F=329.61$, Rho]