MCDA5520 Statistics and Data Analysis

Practice Questions Solutions

80 Points

Section I:
Descriptive Statistics

Please observe the following instructions: 1) Answer each question on a **fresh new page**. 2) Ensure that all questions are clearly labeled. 3) **Clearly label** all Excel printouts. 4) Submit only the **relevant portion** of your Excel printout. (5) Submit neat assignments. Illegible assignments will not be marked.

1. The Highway Loss Data Institute's Injury and Collision Loss Experience report rates car models on the basis of the number of insurance claims filed after accidents. Index ratings near 100 are considered average. Lower ratings are better, indicating a safer car. Shown are ratings for 20 mid-size cars and 20 small cars. Question 1 **Total = 20 pts**

Midsize	81	91	93	127	68	81	60	51	58	75
	100	103	119	82	128	76	68	81	91	82
Small	73	100	127	100	124	103	119	108	109	113
	108	118	103	120	102	122	96	133	80	140

Midsized Cars

5	1 8
6	0 8 8
7	5 6
8	1 1 1 2 2
9	1 1 3
10	0 3
11	9
12	7 8

Small Cars

7	3
8	0
9	6
10	0 0 2 3 3 8 8 9
11	3 8 9
12	0 2 4 7
13	3
14	0

a) Find the mean, median, and mode of the number of accidents filed for small and mid-sized cars.
8 pts (breakdown = 2, 1, 1 for each group)

$$\bar{x}_{midsize} = \frac{\sum x}{n} = \frac{1715}{20} = 85.75 \qquad \bar{x}_{Small} = \frac{\sum x}{n} = \frac{2198}{20} = 109.9$$

$$Median = x_{\frac{1}{2}(20+1)} = x_{10.5} = 81.5 \qquad Median = x_{\frac{1}{2}(20+1)} = x_{10.5} = 108.5$$

$$Mode = 81 \qquad Mode = 100$$

b) Find the range, inter-quartile range, standard deviation and coefficient of variation of the number of accidents filed for small and mid-sized cars. **10 pts (breakdown = 1, 1, 2, 1 for each group).**

MidSize Cars

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{155839 - \frac{(1715)^2}{20}}{20-1}} = 21.49$$

$$cv = \frac{s}{\overline{x}} = \frac{21.49}{85.75} = .25$$

$$Range = x_{\max} - x_{\min} = 128 - 51 = 77$$

$$Q_1 = x_{\frac{1}{4}(n+1)} = x_{5.25} = 68 + .25(75-68) = 69.75$$

$$Q_3 = x_{\frac{3}{4}(n+1))} = x_{15.75} = 93 + .75(100-93) = 98.25$$

$$IQR = 98.25 - 69.75 = 28.5$$

Small Cars

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{246708 - \frac{(2198)^2}{20}}{20-1}} = 16.46$$

$$cv = \frac{s}{\overline{x}} = \frac{16.46}{109.9} = .1497$$

$$Range = x_{\max} - x_{\min} = 140 - 73 = 67$$

$$Q_1 = x_{\frac{1}{4}(n+1)} = x_{5.25} = 100 + .25(102-100) = 100.5$$

$$Q_3 = x_{\frac{3}{4}(n+1))} = x_{15.75} = 120 + .75(122-120) = 121.5$$

$$IQR = 121.5 - 100.5 = 21$$

c) Using the information in a) and b), offer a viewpoint about the safety of mid-size cars in comparison to small cars. Which type of vehicle would you recommend to a safety conscious buyer? **2 pts**

Mid size cars have a smaller average number of claims, while small cars have a small standard deviation of the number of claims. II would recommend midsized cars based on the smaller average number of claims. And recognizing that the standard deviation is not dramatically higher than that of small cars.

2. The Nielson Home Technologies Report provided information about home technology and its usage by persons age 12 and older. The following data are hours of personal computer usage during one week for a sample of 50 users.
Question 2 Total = 20 pts.

Hours

4.1	1.5	10.4	5.7	3.0	5.9	3.4	6.1	1.6	3.7
3.1	4.8	2.0	4.2	11.1	14.8	5.4	4.1	3.9	3.5
4.1	4.1	8.8	3.3	6.2	5.6	4.3	10.3	7.1	7.6
10.8	2.8	9.5	0.7	4.4	12.9	12.1	9.2	4.0	5.7
7.2	6.1	5.7	3.9	6.1	5.9	4.7	3.1	3.7	3.1

a) Draw an ordered stem and leaf plot for the data and describe the shape of the plot. **3 pts**

0	7
1	5 6
2	0 8
3	0 1 1 1 3 4 5 7 7 9 9
4	0 1 1 1 1 2 3 4 7 8
5	4 6 7 7 7 9 9
6	1 1 1 2
7	1 2 6
8	8
9	2 5
10	3 4 8
11	1
12	1 9
13	
14	8

b) Find the mean, median, mode, standard deviation and coefficient of variation for the data.
7 pts (breakdown = 1,1,1,3,1)

$$\bar{x}_{midsize} = \frac{\sum x}{n} = \frac{285.3}{50} = 5.7$$

$$Median = x_{\frac{1}{2}(50+1)} = x_{25.5} = 4.75$$

$$Mode = 4.1$$

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{2107.31 - \frac{(285.3)^2}{50}}{50-1}} = 3.128$$

$$cv = \frac{s}{\bar{x}} = \frac{3.128}{5.7} = .548$$

c) Assuming that the data was drawn from a bell-shaped curve, within what limits would you expect:

- 67% of the values to fall?
- 95% of the values, and,
- 99.9% of the data to fall?

Empirical Rule	Lower Limit	Upper Limit
(67%) $\bar{x} \pm s$	2.578	8.83
(95%) $\bar{x} \pm 2s$	-0.55	11.96
(99%) $\bar{x} \pm 3s$	-3.68	15.09

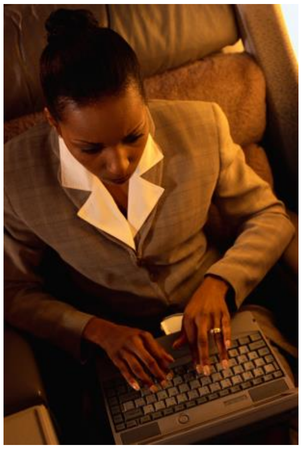
d) compute the actual proportion of the sample that fall within the limits stipulated by the empirical rule. Is there reason to believe that the data came from a bell-shaped curve? **4 pts (break = 3, 1)**

Lower Limit	Upper Limit	Actual Proportion (count)
2.578	8.83	37/50 = 74%
-0.55	11.96	47/50 = 94%
-3.68	15.09	100%

The mean, media and mode for this distribution appear to be quite different. Further, the distribution appears to be skewed to the right. There is evidence that the data may not have come from a bell-shaped curve.

3. In January 2003, the American worker spent an average of 77 hours logged on to the internet while at work (CNBC, March 15, 2003). Assume times are normally distributed with a standard deviation of 20 hours.

Question Total 20 pts



a) What is the probability that a randomly selected worker spent less than 50 hours on the internet? **3pts**

$$P(X \le 50) = P(Z \le z)$$

$$z = \frac{x-\mu}{\sigma} = \frac{50-77}{20} = -1.35$$

$$P(Z \le 1.35) = 0.5 - 0.4115 = 0.0885$$

$$Ans\, 8.85\%$$

b) What percentage of workers spent between 60 and 85 hours? **4 pts**

$$P(60 \le X \le 85) = P(z_1 \le Z \le z_2)$$
$$z_1 = \frac{x-\mu}{\sigma} = \frac{60-77}{20} = -.85$$
$$z_2 = \frac{x-\mu}{\sigma} = \frac{85-77}{20} = .40$$
$$P(-0.85 \le Z \le 0.4) = 0.3023 + 0.1554 = .4577$$
$$Ans\ 45.77\%$$

c) A person is classified as a heavy user if he or she is in the upper 20% of usage. How many hours must a worker log on to be considered a heavy user? **3 pts**

In this case we know that 20% of all users spend more than that amount of time on the internet.

$$P(X \ge x) = P(Z \ge z) = 0.2$$
$$\text{Find the z that corresponds to 20\% in the tail.}$$
$$z = 0.84 \quad \Rightarrow z = \frac{x-\mu}{\sigma} = \frac{x-77}{20} = 0.84$$
$$x = 77 + 16.8 = 93.8\ hours$$
$$Ans\ 93.8\ \text{hours}$$

d) What is the probability that a random sample of 40 workers will spend an average of 87 hours or more logged on to the internet? **4 pts**

Here we are interested in the sampling distribution of the mean.

$$P(\overline{X} \ge \overline{x}) = P(Z \ge z)$$
$$z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} = \frac{87-77}{20/\sqrt{40}} = \frac{10}{1.58} = 6.32$$
$$P(Z \ge 6.32) < 0.001\ \text{(since the largest va lue inthe table is 3.09.}$$
$$Ans\ \text{less that } 0.1\%$$

e) Reflecting on your answer in d), is this a highly likely (highly probable) event? Please interpret your answer. **2 pt**

The probability of this event occurring is extremely small. It is a rare even occurring less than once in 1000 cases.

f) An internet analyst projects by the end of 2005, the top 20% of internet users at work will exceed 120 hours. Assuming the standard deviation hasn't changed, what is the new mean time spent on the internet by an American worker? **4 pts**

$$P\left(X \geq 120\right) = P(Z \geq z) = 0.20$$

$$z = \frac{x-\mu}{\sigma} \Rightarrow 0.84 = \frac{120-\mu}{20}$$

$$\mu = 120 - 0.84 * 20 = 103.2$$

$$\mathit{Ans}\ \text{The new average is } 103.2 \text{ hours}$$

4. In a recent poll of 1000 teens across North America, 37.3% sent text messages. It is believed that 35% of all teens using cell-phones send text messages. **Question Total = 15 pts**

a) What is the probability that in a random sample of 1000 teens, less than 30% send text messages? **4 pts**

$$p = 0.373, \quad \pi = 0.35, n = 1000$$

$$P\left(p \leq 0.30\right) = P(Z \leq z)$$

$$z = \frac{p-\pi}{\sqrt{\frac{\pi\left(1-\pi\right)}{n}}} = \frac{0.30-0.35}{\sqrt{\frac{0.35\left(1-0.35\right)}{1000}}} = \frac{-0.05}{0.0151} = -3.31$$

$$P(Z \leq -3.31) < 0.001$$

$$\mathit{Ans}\ \text{less than } 0.1\%$$

b) What is the probability that the sample proportion of teens sending text messages in a random sample of 1000 would be within $\pm$5% of the population proportion. **5 pts**

$$p = 0.373, \quad \pi = 0.35, n = 1000$$

$$P(0.3 \le p \le 0.4) = P(z_1 \le Z \le z_2)$$

$$z_1 = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.30-0.35}{\sqrt{\frac{0.35(1-0.35)}{1000}}} = \frac{-0.05}{0.0151} = -3.31$$

$$z_2 = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.40-0.35}{\sqrt{\frac{0.35(1-0.35)}{1000}}} = \frac{0.05}{0.0151} = 3.31$$

$$P(-3.31 \le Z \le 3.31) \ge 1-0.002 \ge 0.998$$

Ans greater th an 99.8%

c) What is the probability that in a random sample of 1000 teens, 37.3% or more send text messages? **4 pts**

$$p = 0.373, \quad \pi = 0.35, n = 1000$$

$$P(p \ge 0.373) = P(Z \ge z)$$

$$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.373-0.35}{\sqrt{\frac{0.35(1-0.35)}{1000}}} = \frac{0.023}{0.0151} = 1.52$$

$$P(Z \ge 1.52) = 0.5 - .4357 = 0.0643$$

Ans 6.43%

d) Given that the statistic in c) was actually observed, do you believe there is evidence that the true proportion of teens sending text messages has increased?**2 pts**

The chance of observing this sample is 6.43% . This tells us that if the true proportion of teens sending text messages is 35%, we should see such a sample 64 times in 1000 cases, not a very common occurrence. So why did we see something that should not be a common occurrence, perhaps the proportion of teens sending text messages could now be higher than 35%.

5. Surf the web and find 3 interesting news items that makes use of statistics. Write one short summary paragraph on each story. You can search websites the focus on health trends, technology, music, automobiles, finance, marketing, etc. Please provide the url reference for your story. **5 pts**

Click this link for an example: http://content.nejm.org/cgi/content/short/336/7/453 . You can't use this story however.

Section II:
Sampling Distributions

1. **Define the following terms: 14 marks total**
   a. **Parameter 2 marks**
   A parameter is a measure of a characteristic of a population. For example population mean, population proportion. Once determined the value will not change.

   b. **Statistic 2 marks**
   A statistic is a measure of a characteristic of a sample. For example sample mean, sample proportion. The value of a statistic will change depending upon the sample selected.

   c. **Sampling distribution. 2 marks**
   Is the distribution of all sample means (sampling distribution of means) or sample proportions (sampling distribution of proportions) that we could have obtained from the population.

   d. **State the First Limit and Central Theorems 2 marks**
   First limit theorem states that if the original distribution of individuals is normal, it does not matter what sample size we take, the sampling distribution will also be normal.
   Central Limit theorem states that if our sample size is large enough (n>=30), regardless of the shape of the original distribution of individuals, the sampling distribution will be normal.

   e. **Point estimate 2 marks**
   When we use a sample statistic to estimate a population parameter, we refer to it as obtaining a point estimate for the population

   f. **Interval estimate 2 marks**
   Rather than use one point (obtained from a sample) to estimate a population parameter, we construct an interval of a certain width around the sample value and use that interval to estimate the location of the population parameter.

   g. **Margin of error 2 marks**
   The amount that is added and subtracted to the point estimate to determine the endpoints of the confidence intervals.

2. **Towers Perrin, a New York human resources consulting firm, conducted a survey of 1100 employees at medium-sized and large companies to determine how dissatisfied were employees with their jobs *(The Wall Street Journal,* January 29, 2003). A total of 473 employees indicated they strongly disliked their current work experience. 18 marks total**
   a. **What is the sample statistic of interest? 2 marks**
   Sample statistic of interest is p; the proportion of people in sample who are dissatisfied with their jobs. It is equal to 43%.

   b. **What can we say about the nature of the sampling distribution of the sample statistic in (a)? 2 marks**
   The sampling distribution of proportions is normal. This is because our sample size is greater than 30 and hence we were able to use the Central Limit Theorem.

c. Compute a *95%* confidence interval for the proportion of the population of employees who strongly dislike their current work experience? 5 marks

Let X = # of people in sample who are dissatisfied with their jobs= 473

p = Proportion of people in sample who are dissatisfied with their jobs = 473/1100 = .43

q = 1 - .43 = .57

n = 1100

$$p \pm z\sqrt{\frac{pq}{n}}$$

$$.43 \pm (1.96)\sqrt{\frac{.43(.57)}{1100}}$$

$$.43 \pm .029257$$

$$.40 \text{ to} .46$$

We can be 95% confident that the true Proportion of all employees working in medium and large size companies who dislike their jobs lies somewhere between 40% and 46%.

d. The common view among such firms is that 35% of employees will not like their jobs. Is there any reason to believe that job satisfaction is on the decline? Please explain. 2 marks

Based on our confidence interval, we have concluded that we are 95% confident the true Proportion of all employees working in medium and large size companies who dislike their jobs lies somewhere between 40% and 46%. If the previous view was that this proportion was 35% it implies that job dissatisfaction is on the rise or job satisfaction is on the decline.

e. Towers Perrin would like to estimate the true proportion of employees who do not like their jobs with a margin of error of 1%. How many more employees need to be surveyed? 5 marks

$$\left.\begin{array}{l} 1.\ n > 30 \\ 2.\ pq \text{ problem} \\ 3.\ \text{one sample} \end{array}\right\} \Rightarrow p \pm z\sqrt{\frac{pq}{n}} \quad \text{useful here.}$$

95% CI. $z = \pm 1.96$

E=RHS of Equation=Margin of Error.

$$E = z\sqrt{\frac{pq}{n}}$$

$$.01 = 1.96\sqrt{\frac{(.43)(.57)}{n}}$$

$$\frac{.01}{1.96} = \sqrt{\frac{.2451}{n}}$$

$$\frac{.2451}{n} = \left(\frac{.01}{1.96}\right)^2 = .000026031$$

$$n = \frac{.2451}{.000026031}$$

$$n = 9416$$

We already have a sample of 1100 employees. In order to obtain a margin of error of 1%, we need to survey (9416-1100)= 8316 more employees.

f. Towers Perrin estimates that it costs employers one-third of an hourly employee's annual salary to find a successor and as much as 1.5 times the annual salary to find a successor for a highly compensated employee. What message did this survey send to employers? 2 marks

The information given implies that it is very expensive to replace highly compensated employees. The confidence interval calculated in part c show that job dissatisfaction is on the rise. Therefore it is important for companies to find ways to keep their employees satisfied otherwise it will cost them a lot of money in the long run.

3. Audience profile data collected at the ESPN Sports Zone Web site showed that *26%* of the users were women *(USA Toda)* January 21, 1998). Assume that this percentage was based on a sample of 400 users. 10 marks total

a. At *98%* confidence, what is the margin of error associated with the estimated proportion of users who are women? 3 marks

Let X = # of women in our sample of people who use the ESPN website.
p = Proportion of women in our sample of people who use the ESPN website = .26
q = 1 - .26 = .74
n = 400

$$\left.\begin{array}{l} 1.\ n > 30 \\ 2.\ pq \text{ problem} \\ 3.\ \text{one sample} \end{array}\right\} \Rightarrow p \pm z\sqrt{\frac{pq}{n}} \quad \text{useful here.}$$

98% CI. $z = \pm 2.33$

E=RHS of Equation=Margin of Error.

$$E = z\sqrt{\frac{pq}{n}}$$

$$E = (2.33)\sqrt{\frac{.26(.74)}{400}}$$

$$E = .0511$$

Therefore margin of error is 5.11%.

b. What is the *98%* confidence interval for the population proportion of ESPN Sports Zone Web site users who are women? 3 marks
Two years ago, ESPN Sports Zone estimated the proportion of women using the website at 15%. Is there sufficient reason to believe the proportion of women using the website has increased? Please explain. How might ESPN Sports Zone use this data?

$$p \pm z\sqrt{\frac{pq}{n}}$$

$$.26 \pm (2.33)\sqrt{\frac{.26(.74)}{400}}$$

$$.26 \pm .0511$$

$$.21 \text{ to } .31$$

We can be 98% confident that the true Proportion of women using the ESPN website lies somewhere between 21% and 31%.

Based on the above confidence interval we can conclude that the proportion of women using the ESPN website has increased. We are 98% confident that this proportion lies somewhere between 21% and 31%.

c. How large a sample should be taken if the desired margin of error is .03 and the confidence level is 98%? 4 marks

$$\left.\begin{array}{l} 1.\ n > 30 \\ 2.\ pq \text{ problem} \\ 3.\ \text{one sample} \end{array}\right\} \Rightarrow p \pm z\sqrt{\frac{pq}{n}} \quad \text{useful here.}$$

98% CI. $z = \pm 2.33$

E=RHS of Equation=Margin of Error.

$$E = z\sqrt{\frac{pq}{n}}$$

$$.03 = 2.33\sqrt{\frac{(.26)(.74)}{n}}$$

$$\frac{.03}{2.33} = \sqrt{\frac{.1924}{n}}$$

$$\frac{.1924}{n} = \left(\frac{.03}{2.33}\right)^2 = .0000166$$

$$n = \frac{.1924}{.0000166}$$

$$n = 1161$$

4. A recent study examined the buying preferences of Trinidadian university students for two types of Indian dishes, Dhall and Roti. 225 students were sampled and asked to consider the two types of food. 12 marks total

a. It is believed that students are indifferent to Dhall and Roti. What proportion of students should chose Roti over Dhall? 2 marks

If the students are indifferent it means they have no preference to Roti or Dhall. Therefore the proportion of students who chose Roti over Dhall is 50% or 0.5.

b. In the above survey, 135 students actually expressed a preference for Roti. Construct a 95% confidence interval for the true proportion of students who would choose Roti. 5 marks

Let X = # of students in our sample of 225 who prefer Roti over Dhall = 135

p = Proportion of students in our sample of 225 who prefer Roti over Dhall = 135/225 = 0.6

q = 1 - .6 = .4

n = 225

95% CI. $z = \pm 1.96$

$$p \pm z\sqrt{\frac{pq}{n}}$$

$$.60 \pm (1.96)\sqrt{\frac{.60(.40)}{225}}$$

$$.60 \pm .0640$$

$$.54 \text{ to } .66$$

We can be 95% confident that the true Proportion of all Trinidadain students who choose Roti over Dhall lies somewhere between 54% and 66%.

c. Using your interval in (b), would you conclude that students are indifferent to Dhall and Roti? 2 marks

Based on the above confidence interval I would conclude that the students are not are indifferent to Dhall and Roti.

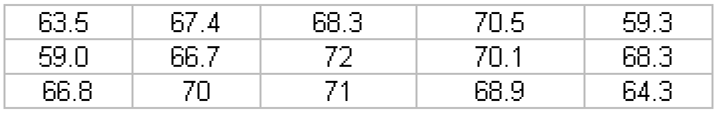
d. What is the nature of the sampling distribution for this problem? 2 marks

The sampling distribution of proportions is normal. This is because our sample size is greater than 30 and hence we were able to use the Central Limit Theorem.

e. What theorem if any is being applied here? 1 marks

Central Limit Theorem.

5. A clothing manufacturer is interested in knowing the mean height of men in Mexico. They believe that knowing this will help them in their clothes design and manufacturing. They wish to estimate the mean within ± 0.20 inches with 95 percent confidence. Before actually sampling the population, they randomly selected 15 men and measured their heights (in inches) shown as follows. 10 marks total



63.5	67.4	68.3	70.5	59.3
59.0	66.7	72	70.1	68.3
66.8	70	71	68.9	64.3

a. Using the sample data, find a 95% confidence interval for the true mean height of men in Mexico. 5 marks

$$\text{MEAN} = \overline{X} = \frac{\sum X}{n} = \frac{1006.1}{15`} = 67.07 inches$$

$$SS = \sum X^2 - \frac{(\sum X)^2}{n} = 67704.37 - \frac{(1006.1)^2}{15}$$

$$= 67704.37 - 67482.48 = 221.889$$

$$\text{STD. DEV} = \text{s} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{221.889}{14}} = 3.98 inches$$

$$n = 15$$

$$\overline{X} = 67.07$$

$$s = 3.98$$

$$95\%\ \text{CI} \Rightarrow \text{t} = \pm 2.1448 \ \ \text{df} = \text{n - 1} = 15 - 1 = 14$$

$$\overline{X} \pm t \ {}^{s}\!/\!_{\sqrt{n}}$$

$$67.07 \pm (2.1448)\, {}^{3.98}\!/\!_{\sqrt{15}}$$

$$67.07 \pm 2.20$$

$$64.87 \text{ inches to } 69.27 \text{ inches}$$

b. What assumptions do we have to make for the confidence interval in (a) to be valid? 2 marks

Because our sample size was less than 30, we could not use CLT. Therefore we were forced to use the FLT. In order for us to use the FLT we had to make the assumption that the distribution of individual values (i.e. the height of men in Mexico) was normally distributed.

c. Using the sample data above, how many more men should be sampled to obtain a desired margin of error of $\underline{+}$ 0.20 at a 95% level of confidence? 3 marks

note: because we are trying to determine sample size, we do not how many degrees of freedom to use (df=n-1). Therefore we use the z value rather than the t value.

$$E=t\left(s/\sqrt{n}\right)$$

$$E=z\left(s/\sqrt{n}\right)$$

$$0.2=(1.96)\left(3.98/\sqrt{n}\right)$$

$$0.2=\frac{7.8008}{\sqrt{n}}$$

$$\sqrt{n}=\frac{7.8008}{0.2}$$

$$\sqrt{n}=39.004$$

$$n=(39.004)^2$$

$$n=1521.312$$

$$n\approx 1522$$

We already have 15 men so the extra men needed to obtain a margin of error of 0.2 would be (1522-15) = 1507 more men.

5. An educational organization in Canada is interested in estimating the mean number of minutes per day that children between the age of 6 and 18 spend watching television per day. The organization selected a random sample of n = 200 children between the age of 6 and 18 and recorded the number of minutes of TV that each person watched on a particular day. The mean time was 191.3 minutes with a standard deviation of 21.5 minutes. 10 marks total

a. Find a 90% confidence interval for the true mean number of minutes that a child between the age 6 and 18 watch per day. 5 marks

$$n=200$$

$$\overline{X}=191.3\,\text{min}$$

$$s=21.5\,\text{min}$$

$$90\%\ \text{CI}\Rightarrow \text{z}=\pm 1.645$$

$$\overline{X}\pm \text{z}\ s/\sqrt{n}$$

$$191.3\pm(1.645)\,21.5/\sqrt{200}$$

$$191.3\pm 2.50$$

$$188.8\ \text{min to}\ 193.8\ \text{min}$$

b. If the leaders of the organization wish to develop an interval estimate with 90 percent confidence, what will the margin of error be? 3 marks

$$E{=}z\left({}^{s}\!\big/\!_{\sqrt{n}} \right)$$

E is the margin of error

$$E = (1.645)\left({}^{21.5}\!\big/\!_{\sqrt{200}} \right)$$

$$E = \frac{35.3675}{\sqrt{200}}$$

$$E = 2.50$$

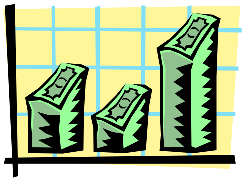
c. What assumptions are necessary for the interval in (a) be valid? 2 marks

Random sampling and sample size greater than 30, but these are already given in the question.

110
PT

Section III:
Interval Estimation and Hypothesis Testing

1: Examine the data on the total annual sales (in billions of dollars) for 25 industrial corporations. The data are shown below. **Total 11 pts**



27.0	29.4	17.6	12.0	10.2
96.9	24.2	15.9	12.0	10.1
86.6	21.7	15.4	11.9	10.1
63.4	21.7	15.2	11.7	10.0
60.0	20.3	15.0	11.4	9.9

$$\sum x = 639.6$$
$$\sum x^2 = 30500$$
$$x = 639.6/25 = 25.58$$
$$s^2 = \sqrt{\frac{1}{25-1}\left[30500 - \frac{639.6^2}{25}\right]} = 24.27$$

a. Find a 95% confidence interval for the mean annual sales for industrial corporations. **(4 pts)**

Since the sample size < 30 and the population SD is unknown, use t-distribution.

$$\bar{x} \pm t\frac{s}{\sqrt{n}}; \quad df = n - 1 = 24,\ t_{.025,24} = 2.064$$
$$25.58 \pm 2.064\left(\frac{24.27}{\sqrt{25}}\right) = 25.58 \pm 10.02$$
$$\$15.56 \le \mu \le \$35.60 billion$$

b. What assumptions are necessary for the analysis in (a) to be valid? **(2 pts)**

- The population is normal
- Sample is randomly chosen

c. Draw a stem and leaf plot of the data. Is there apparent support for your assumptions in (b)?

(2pts)

We can truncate the numbers by dropping off the decimal place, or round the decimals. This stem and leaf plot is based on rounded values.

0	9
1	00001222255568
2	022497
3	
4	
5	
6	03
7	
8	7
9	7

This stem and leaf is highly skewed to the right and therefore non-symmetrical. It is not representative of a normal population.

d. It is desirable to reduce the error in the estimate by half. What sample size is needed assuming we still want a 95% confidence interval? **(3 pts)**

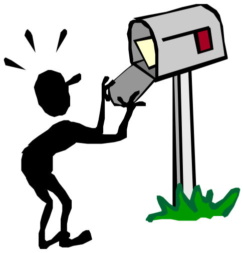
$$AE = z\frac{s}{\sqrt{n}} \Rightarrow n = \left[\frac{z \times s}{AE}\right]^2$$

$$n = \left[\frac{1.96 \times 24.27}{5.01}\right]^2 \approx 91$$

$$Ans: n = 91\ companies$$

For those who use the t value, n is approximately100

2: A mail order company in Dartmouth is attempting a direct marketing strategy on one of its new products. A previous survey of 240 people resulted in 84 willing to purchase the product. The product sells for $14.99. **Total 7 pts**



a. Give a 98% confidence interval for the expected gross revenue if the company will market the product to 10,000 Metro residents. **(4 pts)**

Expected revenue = (the proportion of people that will buy the product) x (number of people targeted) x (selling price)

To get a confidence interval for the revenue we must first get a confidence interval for the true proportion of people (p or π in Lee's case) that will buy the product.

$$p = \frac{84}{240} = .35$$

$$p \pm z\sqrt{\frac{p(1-p)}{n}} \Rightarrow .35 \pm 2.33\sqrt{\frac{.35(.65)}{240}}$$

$$.35 \pm 0.072$$

$$We\ are\ 98\%\ confident\ 0.278 \le \pi \le 0.422$$

$$98\%\ CI\quad for\ revenue\ is\ given\ by:$$

$$(.278)(14.99)(10{,}000) \le E(revenue) \le (.422)(14.99)(10{,}000)$$

$$(\$41{,}672.20) \le E(revenue) \le (\$63{,}257.80)$$

b. What sample size is needed to estimate the true proportion of people willing to purchase the product with an allowable error of ±5% . **(3 pts)**

The allowable error AE=0.05

$$AE = z\sqrt{\frac{p(1-p)}{n}} \Rightarrow n = \frac{z^2(p(1-p))}{AE^2}$$

$$n = \frac{2.33^2(.35(.65))}{0.05^2} \approx 495$$

Answer: 495 persons.

3: A *chewing cycle* is defined as an upward movement followed by a downward movement of the chin. Clinicians have found that the chewing cycles of normal children differ from the chewing cycles of children with eating difficulties. In one study (*The American Journal of occupational Therapy*, May.1984),the number of chewing cycles required for a "normal" preschool child to swallow a bite of graham cracker was found to have a mean of 15.0. In a recent study, a random sample of 20 normal preschool children were found to require a mean of 12 chewing cycles with a standard deviation of 2.5 cycles.

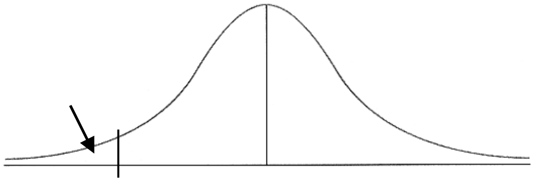
Total 30 points.

a. Find the probability that a random sample of 20 "normal" preschool children will have a mean of 12 chewing cycles or less for a bite of graham cracker. **(3 pts)**

$$P(\bar{x} \le 12) = P(T \le t_o)$$

$$t_o = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{12 - 15}{2.5/\sqrt{20}} = -5.37$$

$$P(\bar{x} \le 12) = P(T \le -5.37) \approx 0$$



$$t_o = -5.37$$

b. Find a 95% confidence interval for the true mean number of chewing cycles required to chew swallow a bite of a graham cracker. **(4 pts)**

$$\bar{x} \pm t\frac{s}{\sqrt{n}}; \quad df = n - 1 = 19, t_{.025,19} = 2.093$$

$$12 \pm 2.093\left(\frac{2.5}{\sqrt{20}}\right) = 12 \pm 1.17$$

$$10.83 \le \mu \le 13.17\ cycles$$

c. What assumptions are needed for the interval in (b) to be valid? **(2 pts)**

Normal population
Randomly chosen sample

d. Is there sufficient evidence that the mean number of chewing cycles necessary to swallow a bite of graham cracker for a "normal" preschool child has changed? **(2 pts)**

Any value in the interval 10.83 to 13.17 is smaller than 15. Hence, there is strong evidence that the mean number of chewing cycles is less than 15, and thus has changed. We are 95% confident about that and there is a 5% chance of being in error.

e. In a separate random sample of 25 preschool children thought to have eating difficulties, a mean of 22 cycles with a standard deviation of 6.4 was observed. Find a 90% confidence interval for the difference between the mean number of chewing cycles required to chew and swallow a bite of graham cracker for the two groups of children. **(5 pts)**

$$\bar{x}_n = 12, s_n = 2.5, \bar{x}_d = 22, s_d = 6.4$$

$$s_p^2 = \frac{(n_d - 1)s_n^2 + (n_n - 1)s_d^2}{n_d + n_n - 2} = \frac{(25-1)6.4^2 + (20-1)2.5^2}{25 + 20 - 2} = 25.623$$

$$df = 25 + 20 - 2 = 43;\ t_{.05,43=1.6811}$$

$$(\bar{x}_d - \bar{x}_n) \pm t\sqrt{s_p^2\left(\frac{1}{n_d} + \frac{1}{n_n}\right)} = (22 - 12) \pm 1.6811\sqrt{25.623\left(\frac{1}{25} + \frac{1}{20}\right)}$$

$$10 \pm 2.553$$

$$We\ are\ 90\%\ confident\ that:$$

$$7.447 \le (\mu_d - \mu_n) \le 12.553$$

We are 90% confident that children with eating difficulties require and average of 7.447 to 12.553 more chewing cycles to chew and swallow a bite of graham cracker.

f. What assumptions are necessary if any for the confidence interval in (e) to be valid? **(2 pts)**

a. Populations are normal
b. Variances are the same
c. Samples are randomly chosen

g. Using the information in (e), is there sufficient evidence that preschool children with eating difficulties require a greater number of cycles to chew and swallow a graham cracker? Assume a 5% significance level.

Step 1: Hypotheses 2 pts

H_0 : Children with eating difficulties do not require more chewing cycles

H_1 : Children with eating difficulties require more chewing cycles

$\mathrm{H}_0 : \mu_D - \mu_n \leq 0$

$H_1 : \mu_D - \mu_n > 0$

Step 2: Test Statistic 2 pts

$$t_0 = \frac{\left(\overline{X}_D - \overline{X}_n\right) - \left(\mu_D - \mu_n\right)}{\sqrt{S_p^2\left(\frac{1}{n_D} + \frac{1}{n_n}\right)}}$$

Step 3: Decision Rule 2 pts

$\mathrm{t}_{\mathrm{crit}} = t_{.10,43} = 1.6811$

If $\mathrm{t}_{\mathrm{o}} > 1.6811$ Reject H_0 or if p-value $< 5\%$ reject H_0.

Step 4: Analysis 4 pts

$$\mathrm{S}_{\mathrm{p}}^2 = \frac{24(6.4)^2 + 9(2.5)^2}{43} = 25.62$$

$$t_0 = \frac{(22-12) - 0}{\sqrt{25.62\left(\frac{1}{20} + \frac{1}{25}\right)}} = \frac{10}{1.52} = 6.57$$

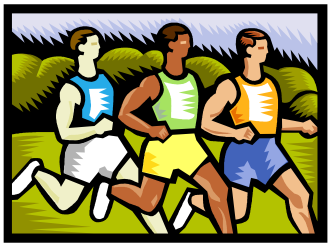
$p - value = P(t \geq 6.57) < .005$

Step 5: Decision/Conclusion 2 pts

Since $\mathrm{t}_0 >> 1.6811$, we Reject H_0.

We conclude there is strong evidence that children with eating difficulties require more chewing cycles...

We have less than 1/2% chance of commiting type I error.



4. The following experiment was conducted to compare two coatings designed to improve the durability of the soles of jogging shoes. A 1/8 inch layer of coating 1 was applied to one of a pair of shoes and a layer of equal thickness of coating 2 was applied to the other shoe. Ten joggers were given pairs of shoes treated in this manner and were instructed to record the number of miles covered in each shoe before the 1/8 inch coating was worn through in any one place. The results are listed in the accompanying table. **Total 10 pts**

Jogger	1	2	3	4	5	6	7	8	9	10
Coating 1	892	904	775	435	946	853	780	695	825	750
Coating 2	985	953	775	510	895	875	895	725	858	812

a. Develop a 90% confidence interval for the difference between the mean number of miles covered by the two types of coatings before being worn through.

This is a paired-difference (matched pairs) test. First we compute the difference in miles of wear for the two coatings 6 ***pts***

Obs	1	2	3	4	5	6	7	8	9	10
diff	93	49	0	75	-51	22	115	30	33	62

$$t_{.05,9} = 1.833$$

$$\overline{x_d} \pm t\frac{s_d}{\sqrt{n}} \Rightarrow 42.8 \pm (1.833)\frac{47.73}{\sqrt{10}}$$

$$42.8 \pm 27.67$$

$$15.13 \le \mu_d \le 70.46$$

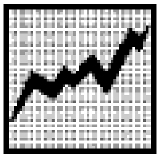
b. There is sufficient evidence that coating 2 lasts longer than coating 1. We are 90% confident that coating 2 provides on average between 15.13 and 70.46 more miles of wear. ***2 pts***

c. Please ignore Question c, as we did not do a hypothesis test

d. Explain why this experimental design is more useful than using two independent groups of joggers, with one group receiving shoes with coating 1 and the other receiving shoes with coating 2.
2 pts

If the samples were independent, there would be several reasons why the coatings lasted differently. By matching coatings according to joggers, the matched pairs would have been subjected to the same conditions of use. Hence the primary explanation for a difference is the difference in the quality of the coatings.

5: Please answer the following questions **Total 10 points**



a. Outline the phases of a hypothesis test. **(4 pts)**

The typical hypothesis test has 5 phases:

I: Statement of hypotheses
II: The test statistic
III: The decision rule for measuring evidence
IV: The analysis
V: The decision and conclusion from the data analysis

b. Define the following terms:

- Null hypothesis and alternate hypothesis **(2 pts)**

The null hypothesis H(0) - is what we assume is true for the purposes of the test. It is set in opposition to H(1).
The alternate hypothesis H(1) - is the claim that we are attempting to prove.

- Type I and Type II Errors
  (2 pts)

Type I error is rejecting the null hypothesis when it is true.
Type II error is failing to reject the null hypothesis when it is false

- P-value (or observed significance level) **(2 pts)**

The p-value or observed significance level is the likelihood of observing the sample evidence or evidence that is even stronger assuming that the null hypothesis is true.

6: The metropolitan airport commission is considering the establishment of limitations on the extent of noise pollution around a local airport. At the present time the noise level per jet takeoff in one neighborhood near the airport is approximately normally distributed with a mean of 100 decibels and a standard deviation of 6 decibels. **Total 17 pts**

a. What is the probability that a randomly selected jet will generate a noise level greater than 108 decibels in this neighborhood? **(3 pts)**

$$\mu = 100db,\ \sigma = 6db$$

$$P(X \geq 108) = P(Z \geq z)$$

$$z = \frac{x-\mu}{\sigma} = \frac{108-100}{6} = 1.33$$

$$P(z \geq 1.33) = .5 - A(1.33) = .5 - 0.4082 = .0918$$

$$Ans : 9.18\%$$

b. Suppose a regulation is passed that requires jet noise in this neighborhood to be lower than 105 decibels 95% of the time. Assuming the standard deviation of the noise distribution remains the same, how much will the mean noise level have to be lowered to comply with the regulation? **(3 pts)**

$$x = 105db,\ \sigma = 6db$$

$$P(X \leq 105) = P(Z \leq z) = .95$$

$$1.64 = \frac{105-\mu_{new}}{6} = \frac{105-\mu_{new}}{6}$$

$$\mu_{new} = 105 - (1.64)(6) = 95.16db$$

$$\text{Re}\,duction = 100 - 95.16 = 4.84db$$

$$Ans : 4.84db$$

c. After the regulation was passed, a random sample of 50 jets yielded a mean noise level of 92 decibels with a standard deviation of 4 decibels. At a 5% significance level, is there sufficient evidence that the true mean noise level produced by the jets complies with the regulation implied in (c)?

Step 1: Hypotheses **(2 pts)**

H_0 :Mean noise level not in compliance

H_a :Mean noise level in compliance

$H_0 : \mu \geq 95.16$

$H_a : \mu < 95.16$

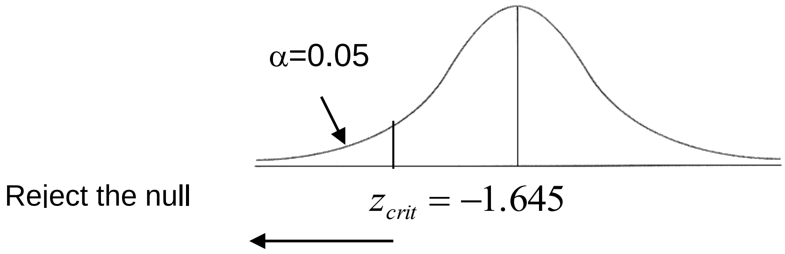
Step 2: Test Statistic **(2 pts)**

The sampling distribution of $\bar{x}$ is normally distributed (first limit theorem). We assume that □□is known.

$$z_o = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Note to markers: Some students will use s instead of □□□□That is OK.

Step 3: Decision Rule **(2 pts)**



If $z_o \leq -1.645$, reject the null hypothesis, otherwise, fail to reject it.
Alternatively, if the p-value or observed significance is smaller than 5% reject the null hypothesis.

Step 4: Analysis
(3 pts)

$$z_o = \frac{92 - 95.16}{6 / \sqrt{50}} = -3.72$$

The p-value is $P(z \leq -3.72) = .5 - A(-3.72) = .5 - .49990 = .0001$

Step 5: Decision/Conclusion **(2 pts)**

Since the test statistic (-3.72) << -1.645, since the p-value << 0.05, reject the null hypothesis. We conclude there is strong evidence that the current average noise level complies with the new regulation.

7: The percentage of body fat can be a good indicator of an individual's energy metabolic status and general health. In a recent study conducted by the Brazilian Health Commission of the percentage of body fat of college students, two groups of healthy male students from urban and rural colleges in Eastern Brazil, were randomly and independently selected. The percentage of body fat was measured in each group, and the results summarized in the table below. **Total 25 pts**

	Urban Students	Rural Students
Sample Size	193	188
Mean	12.07	11.04
Std. Dev. 3.04		2.63

a) If the we believe that there is no difference between the true mean percentage of body fat for urban and rural students, what is the probability that random samples of 193 urban students and 188 rural students will yield an **absolute** difference between the sample means of 1.03% or higher? **6 pts**

$$P(|\bar{x}_U - \bar{x}_R| \geq 1.03) = P((\bar{x}_U - \bar{x}_R) \leq -1.03) + P((\bar{x}_U - \bar{x}_R) \geq 1.03)$$

$$\Rightarrow P(Z \leq z_1) + P(Z \geq z_2)$$

$$z_2 = \frac{(\bar{x}_U - \bar{x}_R) - (\mu_U - \mu_R)}{\sqrt{\frac{s_U^2}{n_U} + \frac{s_R^2}{n_R}}} = \frac{1.03 - 0}{\sqrt{\frac{3.04^2}{193} + \frac{2.63^2}{188}}} = 3.54$$

$$z_1 = -3.54$$

$$P(|\bar{x}_U - \bar{x}_R| \geq 1.03) = 2P(z \geq 3.54) = 2(.5 - .4998) = 2(.0002) = .0004$$

Answer: 0.04%

b) Would you refer to this sample as a rare sample? Please explain. **1 pt**

This sort of sample should occur one 4 times in every 10,000. Hence this is a rare occurrence.

c) Find a 98% confidence interval for the difference between the true mean percentage of body fat for urban and rural students. Please interpret the interval. **6 pts**

$$\left(\bar{X} - \bar{X}_R\right) \pm Z\sqrt{\frac{S_u^2}{n_u} + \frac{S_R^2}{n_R}} \Rightarrow (12.07 - 11.04) \pm 2.33\sqrt{\frac{3.04^2}{193} + \frac{2.63^2}{188}}$$

$$\Rightarrow 1.03 \pm 2.33(.29099) \Rightarrow 1.03 \pm 0.678$$

$$0.352 \leq \mu_u - \mu_R \leq 1.708$$

We are 98% Confident that urban students have an average between 0.35% and 1.708% more body fat than rural students.

d) Does the study provide sufficient evidence that there is a difference between the mean percentage of body fat for urban and rural students in Eastern Brazil. Use a significance level =0.02. **10 pts**

$$\text{Step 1: Hyp } \mathrm{H}_0 : \textit{Mean} \text{ percentage body fat are the same}$$

$$\mathrm{H}_1 : \text{Mean percentage body fat are different}$$

$$\mathrm{H}_0 : \mu_u - \mu_R = 0$$

$$H_1 : \mu_u - \mu_R \neq 0$$

$$\text{Step 2: Test Statistic}$$

$$Z_0 = \frac{\left(\bar{X}_u - \bar{X}_R\right) - \left(\mu_u - \mu_R\right)}{\sqrt{\frac{S_u^2}{n_u} + \frac{S_R^2}{n_R}}}$$

$$\text{Step 3: Decision Rule}$$

$$\text{If } |\mathrm{Z}_0| > 2.33 \text{ Reject } \mathrm{H}_0 \text{ or if p-value} < 2\%$$

$$\text{Step 4: Analysis}$$

$$Z_0 = \frac{(12.07 - 11.04) - 0}{\sqrt{\frac{3.04^2}{193} + \frac{2.63^2}{188}}} = 3.54$$

$$p - value = \mathrm{P}(\mathrm{Z} \geq 3.54) + \mathrm{P}(\mathrm{Z} \leq \text{-}3.54)$$

$$= .0004 \text{ (same as (a)).}$$

$$\text{Step 5: Since } \mathrm{Z}_0 >> 2.33 \text{ or since p-value} << .02, \text{ reject } \mathrm{H}_0.$$

We have strong evidence that the mean body fat of urban and rural students are different.

e) What is the observed significance of the test (or p-value). Give an interpretation of the p-value. **2 pts**

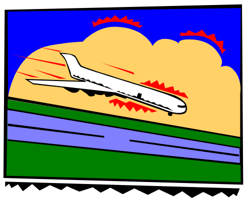
The p-value or observed significance implies that if the students have the same mean body fat, then there is only a 0.04% chance of observing the sample evidence of evidence that is stronger.

110

Section IV:
Proportions and Analysis of Variance

Please observe the following instructions: 1) Answer each question on a **fresh new page**. 2) Ensure that all questions are clearly labeled. 3) **Clearly label** all Excel printouts where appropriate. 4) Submit only the **relevant portion** of your Excel printout. (5) Submit neat assignments. Illegible assignments will not be marked.

1: The metropolitan airport commission is considering the establishment of limitations on the extent of noise pollution around a local airport. At the present time the noise level per jet takeoff in one neighborhood near the airport is approximately normally distributed with a mean of 100 decibels and a standard deviation of 6 decibels. Regulations were passed that requires a mean noise level of 95 decibels. Prior to the regulation, 120 out of a random sample of 300 aircraft met the regulation standard. In the first year after the regulation was imposed, 250 out of a random sample of 450 aircraft met the standard. **TOTAL 27 points**



a) Find a 95% confidence interval for the difference between the true proportion of aircraft before and after the regulation that meet the new noise standard. *4 pts*

$$(p_A - p_B) \pm Z\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}} \Rightarrow p_B = \frac{120}{300} = .4 \quad p_A = \frac{250}{450} = .56$$

$$(.56 - .4) \pm 1.96\sqrt{\frac{(.56)(.44)}{450} + \frac{(.4)(.6)}{300}} = .16 \pm (1.96)(.036)$$

$$= .16 \pm .072$$

Between 8.8% to 23.2% more jets are complying with the regulation . We are 95% confident about that .

b) At a 5% significance level, is there sufficient evidence that a greater number of aircraft now meet the desired noise level of 95 decibels? ***10 pts***

Step 1 : Hyp 2 pts

H_0 : Compliance is not higher after the regulation

H_1 : compliance is higher after the regulation

$\mathrm{H}_0 : \pi_A - \pi_B \leq 0$

$H_1 : \pi_A - \pi_B > 0$

Step 2 : Test Statistic 2 pts

$$Z_0 = \frac{(p_A - p_B) - (\pi_A - \pi_B)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \qquad \bar{\mathrm{p}} = \frac{\mathrm{x}_\mathrm{A} + x_B}{n_A + n_B}$$

Step 3 : Decision Rule 1 pt

If $Z_0 > 1.64$ Reject H_0 3 pts

Step 4 : Analysis

$$\bar{\mathrm{p}} = \frac{120 + 250}{450 + 300} = \frac{370}{750} = .49$$

$$Z_0 = \frac{(.56 - .4) - 0}{\sqrt{.49(.51)\left(\frac{1}{300} + \frac{1}{450}\right)}} = \frac{.16}{.037} = 4.29$$

$p - value = P(Z \geq 4.29) < 0.001$ (actual value from Excel is .00003)

Step5 : Decision/C onclusion 2 pts

Since $\mathrm{Z}_0 >> 1.64$ Reject H_0. There is strong evidence that a higher percentage of jets comply with the regulation .

c) What is the observed significance (p-value) of the test? ***1 pt***

Observed significance $P(Z \geq 4.29)$<.0.001

d) What error do we risk making given our conclusion in (b)? ***1 pt***

Risk of type I error (rejecting the null hypothesis when it is true) which is very small.

e) **The FAA will consider an aggressive campaign to go after violators if the difference between the true proportion of aircraft meeting the regulation standard before and after the enactment of the standard did not improve by at least 10% in the first year following the new regulation. At a 5% significance level, is there a need for the FAA to get tough with violators?** ***10 pts***

$\text{Step 1: Hyp} \quad 2\text{pts}$

$\text{H}_0: \text{ An aggressive campaign is not needed}$

$\text{H}_1: \text{An aggressive campaign is needed}$

$\text{H}_0: \pi_A - \pi_B \le .10$

$H_1: \pi_A - \pi_B > .10$

$\text{Step 2: Test Statistic} \quad 2\text{pts}$

$$Z_0 = \frac{(p_A - p_B) - (\pi_A - \pi_B)}{\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}}$$

$\text{Step 3: Decision Rule} \quad 1\,\text{pt}$

$\text{If } Z_0 > 1.64 \text{ Reject H}_0$

$\text{Step 4: Analysis} \quad 3\,\text{pts}$

$$Z_0 = \frac{(.56 - .4) - .10}{\sqrt{.\frac{(.56)(.44)}{450} + \frac{(.4)(.6)}{300}}} = \frac{.06}{.0367} = 1.63$$

$p - value = P(Z \ge 1.63) = .0516$

$\text{Step5: Decision/C onclusion} \quad 2\text{ pts}$

$\text{Since Z}_0 < 1.63 \text{ fail to reject H}_0.$ We do not have sufficient evidence of a need for an aggressive campaign.

f) **What is the observed significance (p-value) of the test?** ***1 pt***

Observed significance $P(Z \ge 1.63)$ = .0516 or 5.16%.

If we ignored the significance level alpha and rejected the null hypothesis anyway, our risk of being in error (type I) is 5.16%.

2: A hospital in Nova Scotia would like to determine whether there is a relationship between the level of satisfaction of workers with working conditions and their job category. A random sample of 250 employees were surveyed. The results are summarized in the table below.

TOTAL *22 pts*

	Nurses	*Support Staff*	*Doctors*	*Total*
Satisfactory	40	80	20	140
Indifferent	20	40	20	80
Poor	15	5	10	30
Total	75	125	50	**250**

a) Find a 98% confidence interval for the true proportion of all workers that is satisfied with working conditions? *4 pts*

Based on the sample data, 140 out of 250 workers are satisfied with working conditions.

$$p = \frac{140}{250} = ..56$$

$$p \pm z\sqrt{\frac{p(1-p)}{n}} \Rightarrow 0.56 \pm 2.33\sqrt{\frac{.56(.44)}{250}}$$

$$0.56 \pm 0.073$$

$$0.487 \leq \pi \leq 0.633$$

We are 98% confident that between 48.7% and 63.3% of all employees are satisfied with working conditions.

b) Is there sufficient evidence that support staff are more satisfied with working conditions than nursing staff? Use a 90% confidence interval.

4 pts

We must first find a confidence interval for the difference between the true proportion of support staff and nurses that are satisfied with working conditions.

$$(p_s - p_n) \pm Z\sqrt{\frac{p_s(1-p_s)}{n_A} + \frac{p_n(1-\hat{p}_n)}{n_B}} \Rightarrow p_s = \frac{80}{125} = .667 \quad p_n = \frac{40}{75} = .533$$

$$(.667 - .533) \pm 1.64\sqrt{\frac{(.667)(.333)}{125} + \frac{(.533)(.467)}{75}} = .134 \pm .117$$

$$.017 \leq \pi_s \text{ - } \pi_n \leq .251$$

We are 90% confident that the difference between the proportion of support staf and nurses who find their working conditions satisfactory lies between 1.7% and 25.1% We have a 10% chance of being wrong.

c) **Using your answer in (b) do you think that job category and worker satisfaction are related? Please explain. Just saying yes or no is not acceptable.**

2 pts

We note that every value in the interval 1.7% to 25.1% is positive, implying that $\pi_s > \pi_n$. Hence satisfaction is related to job category. If you are a support staff, you are more likely to be satisfied with conditions than nurses.

d) **At a 5% significance level, is there sufficient evidence of a difference between the level of satisfaction with working conditions for doctors and nurses?**

<u>10 pts</u>

Step 1: Hypotheses

H_0 : Satisfaction levels for doctors and nurses are the same

H_1 : Satisfaction levels for doctors and nurses are different

$\mathrm{H}_0 : \pi_n - \pi_d \le 0$

$H_1 : \pi_n - \pi_d > 0$

Step 2 : Test Statistic

$$Z_0 = \frac{(p_n - p_d) - (\pi_n - \pi_d)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_n} + \frac{1}{n_d}\right)}} \qquad \bar{\mathrm{p}} = \frac{\mathrm{x_n} + x_d}{n_n + n_d}$$

Step 3 : Decision Rule

This is a two - tailed test. $\alpha = .05$ and must be divided between the tails.

If $Z_0 > 1.96$ or $Z_0 < -1.96$ Reject H_0

Step 4 : Analysis

$$p_n = \frac{40}{75} = 0.533; \ p_d = \frac{20}{50} = 0.4$$

$$\bar{\mathrm{p}} = \frac{40+20}{75+50} = \frac{60}{125} = .48$$

$$Z_0 = \frac{(.533 - .4) - 0}{\sqrt{.42(.52)\left(\frac{1}{75} + \frac{1}{50}\right)}} = \frac{.133}{.091} = 1.46$$

$p - value = 2 \times P(Z \ge 1.46) = 2(.5 - .4279) = .1442$

*Step*5 : *Decision / Conclusion*

Since $\mathrm{Z}_0 < 1.96$ we fail to reject H_0. There is insufficient evidence of a difference between the satisfaction levels of doctors and nurses.

e) Looking at your answer in (d) is there any indication that job category and level of satisfaction are related? Please explain. ***2 pts***

The analysis in (d) does not provide sufficient evidence that job category and satisfaction level are related.

3. An Agricultural Lab in Mexico is testing the effect of two types of fertilizers on the growth rate of mango seedlings. A random sample of 20 seedlings were given fertilizer A, and 25 seedlings were given fertilizer B. The increase in the height of the seedlings over a three-week period was measured, and the results summarized as follows: **TOTAL *10 pts***

	Fertilizer A	Fertilizer B
Sample size	20	25
Sample mean	50.5 mm	57.5 mm
Sample standard deviation	13.2 mm	8.5 mm

A statistician wants to perform a t-test to determine whether fertilizer B results in a larger mean growth rate for the seedlings over the three-week period. To do so, she must assume equal population variances. Determine whether the assumption of equal variances is reasonable. Use a 5% significance level

Step 1: Hypotheses **2 pts**

H_0: Variances are the same
H_1: Variances are not the same
H_0: $\sigma_A^2 / \sigma_B^2 = 1$
H_1: $\sigma_A^2 / \sigma_B^2 \neq 1$

Step 2: test Statistic **2 pts**

$F = \dfrac{S_A^2}{S_B^2}$ (place the larger sample standard deviation over the smaller one)

$df_A = 19$; $df_B = 24$

Step 3: Decision Rule **2 pts**

$\dfrac{\alpha}{2} = .025$ $F_{Crit} = F_{.025,\,19,\,24} = 2.345$

If the calculated F > 2.345, reject H_o, otherwise do not reject H_o
If p-value < .05 reject H_0

Step 4: Analysis **2 pts**

$F = 13.2^{2}/8.5^{2} = 2.412$

p-value = 2 x P(F>2.412)

0.02 < p-value < 0.05

Step 5: Decision/Conclusion **2 pts**

Since 2.412 > 2.345 reject H_o and conclude that the assumptions of equal variances is not a reasonable assumption.

4. High school students planning to attend university were randomly assigned to watch one of four videos about Dalhousie University. Each video differed in the particular aspect of university life of students that was emphasized: athletics, academics, social life, or art and culture. After watching a video, each student was given a questionnaire that measured his or her desire to attend Dalhousie University. Student's answers were analyzed to provide a measure of their desire to attend Dalhousie. (Scores reflect the percentage of questions indicating a strong preference to attend Dalhousie. High scores reflect a greater desire to attend Dalhousie.) A total of 16 students were shown one of the videos. **TOTAL *20 pts***

Athletics	Social Life	Academics	Art/Cultural
68	89	74	76
56	78	82	71
69	81	79	69
70	77	80	65

(a) At the 1% level of significance, do these data suggest that the type of activity emphasized in a university video affects the desire of prospective students to attend Dalhousie? (Do all calculations by hand.) ***12 pts***

STEP 1: HYPOTHESES 2 PTS

H_0: The mean effect for the types of videos are the same
H_a: The mean effect of the types of videoes are not all the same
H_0: $\mu_1=\mu_2=\mu_3=\mu_4$
$H_{a:}$ at least one mean is different

Step 2: Test Statistic
1 pt
F_0=MSB/MSW

Step 3: Decision Rule **2 pt**

df_1 =k-1 = 3; df_2 = n-k = 16-4 = 12
$F_{crit} = F_{.01,3,12 = 5.95}$

If F_0>5.95 reject H_0 or if p-value $P(F \geq F_0)$ <1% reject H_0.

Step 4: Analysis **5 pts**

(Please note that I have done my calculations in Excel. Not everyone will do that.)

	Athletics	*Social Life*	*Academics*	*Art/Cultural*	
	68	89	74	76	
	56	78	82	71	
	69	81	79	69	
	70	77	80	65	**Total**
Sample Size	4	4	4	4	16
Sum (x)	263	325	315	281	1184
Sum (x2)	17421	26495	24841	19803	88560
SS	128.75	88.75	34.75	62.75	944

SOV	**SS**	**DF**	**MS**	**F(observed)**	**F(crit)**	**P-value**
Groups	629	3	209.6666667	7.987301587	5.952545	0.003419
Within	315	12	26.25			
Total	944	15				

SST = 944
SSW = 315
SSB = 944 – 315 = 629

MSB = SSB /(k - 1) = 629 /3 = 209.67
MSW = SSE /(n - k) = 315 /(16 - 4) = 26.25

$$F = \frac{MSB}{MSW} = \frac{209.76}{26.25} = 7.99$$

Using F table (3 degrees of freedom numerator and 12 denominator), p-value is <.01 *(Actual p-value = .003419 (from Excel))*

Step 5: Decision/Conclusion **2 pts**

Since $F_0 > 5.95$ or because p-value $< \alpha = .01$ we reject H_0. We conclude that there is sufficient evidence that the mean effect of the types of videos are different. We could commit type I error, but the chance is < 1%.

b) What is the observed significance (p-value) of the test? ***1 pt***

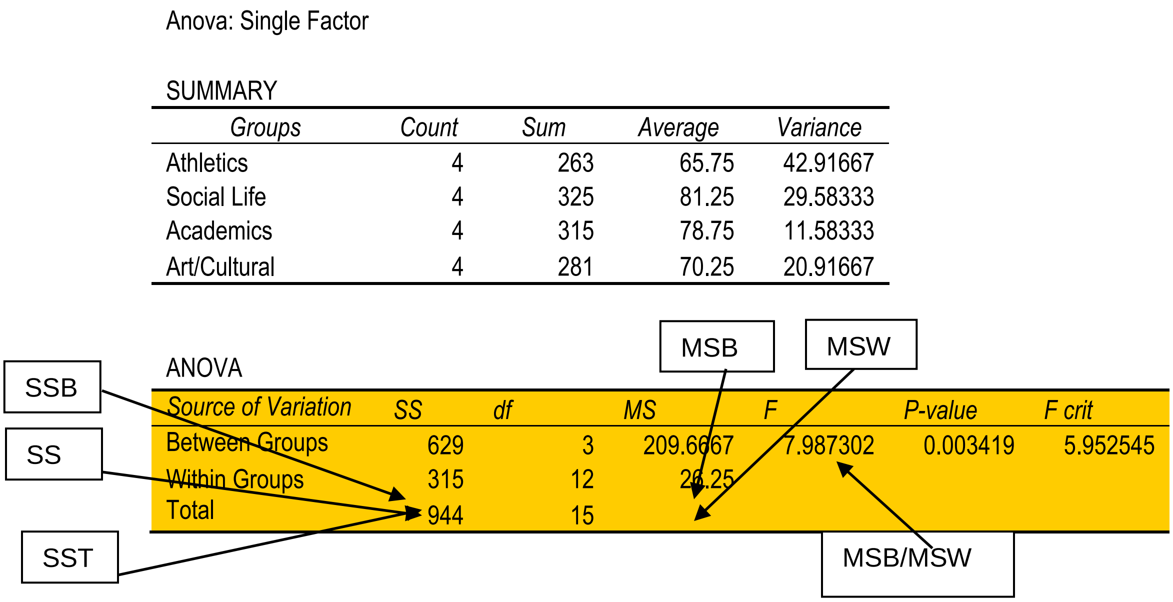
The observed significance of the test is the same as the p-value which is < 1% or exactly equal to 0.3419%.

c) What assumptions are necessary for the above analysis to be valid? ***2 pts***

- Samples are independently and randomly chosen
- Populations are normally distributed
- Population variances are equal

d) Use Excel to answer part (a). Interpret the results by annotating your printout. Hand in your annotated printout. [Note: Your results for this part of the question must be printed out on a single sheet of computer paper with all excess perforated paper removed. Failure to follow these instructions will result in a mark of zero.]
5 pts

Excel Output



Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Athletics	4	263	65.75	42.91667
Social Life	4	325	81.25	29.58333
Academics	4	315	78.75	11.58333
Art/Cultural	4	281	70.25	20.91667

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	629	3	209.6667	7.987302	0.003419	5.952545
Within Groups	315	12	26.25			
Total	944	15				

5. A manager in a rapidly growing industry wishes to study the effects of different production systems on worker output. Traditionally, all staff has used an assembly line for production. Senior management has been most reluctant to change the existing system. Within his own branch, however, the manager wishes to gather evidence on whether or not a change in the system of production can affect worker output. Over the past 6 months three different production systems – an assembly line in which each worker does only one task, a system in which each worker does many different tasks (and, therefore, requiring more expertise), and a system combining elements of both a single tasking and a multiple tasking system - have been tried with the results presented in the table below. Each of the 9 individuals in the office worked under each of the three production systems (for periods of 1 month each). Each person started on the single task system, then did the combination system, and finished with the multiple-task system. The following data represent output for each worker (in hundreds of units produced) achieved during the last month on the job under each production scheme.

TOTAL *31 pts*

SYSTEM USED

Worker	**Combination-Single-Multiple Tasks**	**Multiple Tasks**	**Single Task (Assembly Line)**
1	256	224	269
2	239	254	284
3	222	273	294
4	207	285	290
5	228	237	247
6	241	277	278
7	212	261	263
8	216	228	229
9	236	234	236

(a) Is there evidence that the three production systems differ significantly in their effect on worker output? Use $\alpha = .05$ and find the *p*-value. (Do all calculations by hand.)

13 pts

STEP 1: HYPOTHESES

2 PTS

H_0: Mean output for the three production systems do not differ
H_a: Mean output for the three production systems are not all the same
H_0: $\mu_1=\mu_2=\mu_3$
H_a: at least one mean is different

Step 2: Test Statistic

1 pt

F_0=MSB/MSW

Step 3: Decision Rule 2 pt

$Df_1 = k-1 = 2$; $df_2 = (b-1)(k-1) = (2 \times 8) = 16$
$F_{crit} = F_{.05,2,16} = 4.458$

If $F_0>4.458$ reject H_0 or if p-value $P(F \geq F_0) <5\%$ reject H_0.

Step 4: Analysis 6 pts

	Worker	Combination-Single-Multiple Tasks	Multiple Tasks	Single Task (Assembly Line)	Blocks		
					k	mean(i)	SS(i)
	1	256	224	269	3	249.67	1.81
	2	239	254	284	3	259.00	306.70
	3	222	273	294	3	263.00	597.37
	4	207	285	290	3	260.67	416.15
	5	228	237	247	3	237.33	400.59
	6	241	277	278	3	265.33	811.26
	7	212	261	263	3	245.33	37.93
	8	216	228	229	3	224.33	1808.93
	9	236	234	236	3	235.33	551.26
Groups	**b**	**9**	**9**	**9**		**SSBL**	**4932**
	mean(j)	228.56	252.56	265.56	**SSB**		**SST**
	SS(j)	3721	121	2500	6342		16974.67
	Grand mean	248.89					

SOV	SS	DF	MS	F	Fcrit	Pvalue
Groups	6342	2	3171	8.900012	3.63	0.0025
Blocks	4932	8	616.5	1.730324	3.12	0.1668
Error	5700.67	16	356.2917			
Total	16974.67	26				

$$SST = \sum_i \sum_j x_{ij}^2 - \frac{\left(\sum_i \sum_j x_{ij}\right)^2}{n} = 16974.67$$

$$SSB = b\sum\left(\bar{x}_j - \bar{\bar{x}}\right)^2 = 6342$$

$$SSBL = k\sum\left(\bar{x}_i - \bar{\bar{x}}\right)^2 = 4932$$

SSW = SST - SSB - SSBL = 16974.67-6342-4932=5700.67

MSB = SSB/(k-1) = 6342/2=3171
MSBL = SSBL/(b-1) = 4932/8=616.5
MSW = SSW/(b-1)(k-1) = 5700.67/16=356.29

F_B = MSB/MSW = 3171/356.27=8.9

Using *F* table (2 degrees of freedom numerator and 16 denominator), *p*-value is < 0.01 *(Actual p-value = .0025 from Excel)*
Step 5: Decision/Conclusion 2 pts

Since 8.9 > 3.63 or since p-value < 5% reject H_0. We conclude the mean output for the three production systems are not the same.

(b) Is there evidence that the workers differ significantly in their mean production output? Use α = .025 and find the *p*-value. (Do all calculations by hand.) ***<u>8 pts</u>***

STEP 1: HYPOTHESES

2 PTS

H_0: The workers do not differ in their mean production output
H_a: At least one worker's mean production output differ from the rest
H_0: $\mu_1=\mu_2=\ldots=\mu_9$
$H_{a:}$ at least one mean is different

Step 2: Test Statistic **1 pt**
F_0=MSBL/MSW

Step 3: Decision Rule **1 pt**
$Df_1 = b-1 = 8$; $df_2 = (b-1)(k-1) = (2 \times 8) = 16$
$F_{crit} = F_{.025,8,16} = 3.12$

If $F_0>3.12$ reject H_0 or if p-value $P(F \geq F_0) < 2.5\%$ reject H_0.

Step 4: Analysis **2 pts**
NB: We do not need to re-compute the data. All of the work was already done in answering part (a).

F_{BL}= MSBL/MSW = 1.73
p-value = $P(F\geq1.73)$. Based on our table p-value > 0.05. *(Actual value is 0.1668 from Excel)*

Step 5: Decision/Conclusion 2 pts
Since 1.73 < 3.12 or since the p-value > 2.5%, we conclude that there is insufficient evidence that at least one worker's mean output differs from the rest.

(c) Can you see any advantages/disadvantages to the use of the type of design used in this study? Discuss in depth. Be clear and concise. Avoid the use of jargon.

<u>4 pts</u>

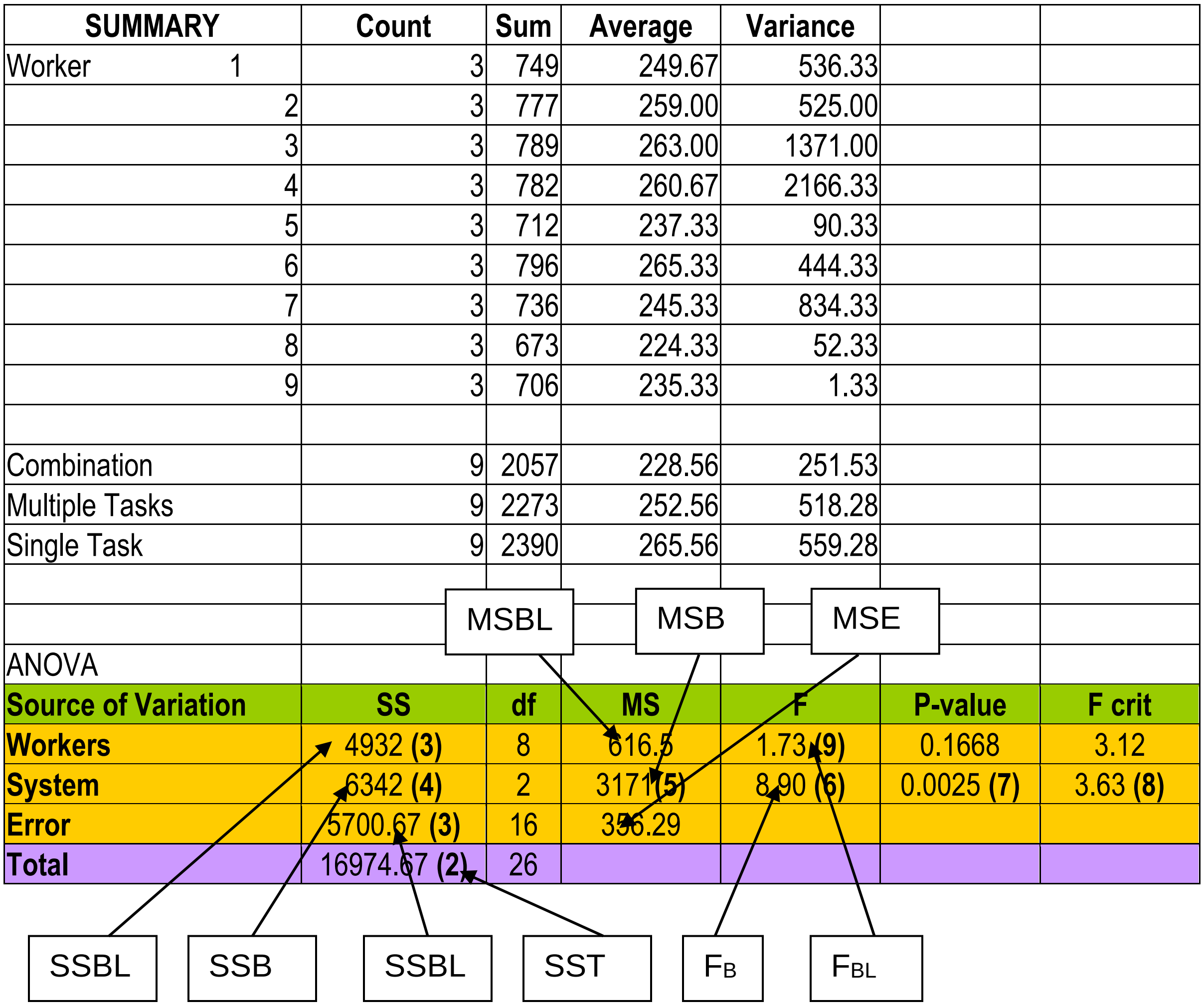
- The block design removes the variation due to the blocks from SSW. As a result, we can better represent the variation due to treatments relative to the error (SSW). It helps us to better detect the effect of the treatments.
- It can be quite effective with even small sample sizes
- One potential problem is the order effects. (The order in which the workers use the production system). Ordering can affect learning, and fatigue. Moving from treatment to another can cause learning which may impact performance with other treatments. As well, moving from one treatment to another depending on the experiment can wear out the subjects (cause fatigue). Fatigue can affect performance.

(d) **Discuss an alternative way this study could have been conducted that would not have had the same major disadvantages.** ***1 pt***

Perform a completely randomized design, i.e., use independent samples.

e) Use Excel to answer part (a). Interpret the results by annotating your printout. Hand in your annotated printout. [Note: Your results for this part of the question must be printed out on a single sheet of computer paper with all excess perforated paper removed. Failure to follow these instructions will result in a mark of zero.]

5 pts



SUMMARY	Count	Sum	Average	Variance		
Worker 1	3	749	249.67	536.33		
2	3	777	259.00	525.00		
3	3	789	263.00	1371.00		
4	3	782	260.67	2166.33		
5	3	712	237.33	90.33		
6	3	796	265.33	444.33		
7	3	736	245.33	834.33		
8	3	673	224.33	52.33		
9	3	706	235.33	1.33		
Combination	9	2057	228.56	251.53		
Multiple Tasks	9	2273	252.56	518.28		
Single Task	9	2390	265.56	559.28		
ANOVA						
Source of Variation	**SS**	**df**	**MS**	**F**	**P-value**	**F crit**
Workers	4932 **(3)**	8	616.5	1.73 **(9)**	0.1668	3.12
System	6342 **(4)**	2	3171 **(5)**	8.90 **(6)**	0.0025 **(7)**	3.63 **(8)**
Error	5700.67 **(3)**	16	356.29			
Total	16974.67 **(2)**	26				

Section V:
Contingency Tables and Regression Analysis

90 Points

1: A hospital in Nova Scotia would like to determine whether there is a relationship between the level of satisfaction of workers with working conditions and their job category. A random sample of 250 employees were surveyed. The results are summarized in the table below. **(TOTAL = 12 points)**

	Nurses	*Support Staff*	*Doctors*	*Total*
Satisfactory	40	80	20	140
Indifferent	20	40	20	80
Poor	15	5	10	30
Total	75	125	50	**250**

f) Is there clear evidence that employee category and the level of satisfaction with working conditions are related? Use a 1% significance level.

Step 1: Hyp H_0: Job category and job satisfaction are unrelated (2 pts)

H_1 : Job category and job satisfaction are related

$H_0 : \pi_{ij} = \pi_i \cdot \pi_j$

$H_1 : \pi_{ij} \neq \pi_i \cdot \pi_j$

Step 2: Test Statistic (2 pts)

$$X^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Step 3: Decision Rule (1 pt)

df $= (r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$

$X^2_{.01,4} = 13.27$

If $X_0^2 \geq 13.28$ Reject H_0

Step 4: Analysis (4 pts)

$$X^2 = \frac{(40-42)^2}{42} + \frac{(80-70)^2}{70} + \ldots\ldots + \frac{(10-15)^2}{15} =$$

$$= 0.95 + 1.428 + 2.286 + 0.667 + 0 + 1 + 4 + 6.667 + 2.667 = 18.81$$

$$p - value = P(X^2 \geq 18.81) < 0.005$$

Step5: Decision/Conclusion (1 pt)

Since $X_o^2 > 13.28$ reject H_0. There is sufficient evidence that job category and satisfaction are related.

g) Compute and interpret the observed significance of the test.

2 pts

If job category and job satisfaction are unrelated, then we should see such a sample less than 5 times in 1000.

2. A supermarket chain wanted to know whether there was any relationship between the price (in dollars) set by the chain for its in-house brand of coffee and demand (measured in pounds of coffee). Eight stores in the chain that had nearly equal past histories of demand for this brand of coffee were used in the study. Eight different prices were randomly assigned to the stores (one price per store) and an identical ad campaign was run in each market. The number of pounds of coffee sold during the following week was recorded for each store (see below). **(TOTAL = 58 points)**

Store	**Demand**	**Price**
A	1120	$ 3.00
B	999	$ 3.10
C	932	$ 3.20
D	884	$ 3.30
E	807	$ 3.40
F	760	$ 3.50
G	701	$ 3.60
H	688	$ 3.70

Store	Demand (Y)	D2 (Y2)	Price (X)	P2 (X2)	DxP (XY)
A	1120	1254400	3.00	9	3360
B	999	998001	3.10	9.61	3096.9
C	932	868624	3.20	10.24	2982.4
D	884	781456	3.30	10.89	2917.2
E	807	651249	3.40	11.56	2743.8
F	760	577600	3.50	12.25	2660
G	701	491401	3.60	12.96	2523.6
H	688	473344	3.70	13.69	2545.6
Sum	***6891***	***6096075***	***26.8***	***90.2***	***22829.5***
Mean	***861.375***		***3.35***		

(a) What is the coefficient of correlation in this situation? Explain what it means. ***(3 pts)***

$$SS_X = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 90.2 - \frac{26.8^2}{8} = .42$$

$$SS_Y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} = 6096075 - \frac{6891^2}{8} = 160339.88$$

$$SS_{XY} = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} = 22829.5 - \frac{(6891)(26.8)}{8} = -255.35$$

$$r = \frac{SS_{XY}}{\sqrt{SS_X\, SS_Y}}$$

$$= \frac{-255.35}{\sqrt{(.42)(160339.88)}} = -.9840$$

There is a high negative correlation between demand and price. As the price goes up, demand for this brand of coffee falls. The correlation is close to the maximum possible correlation of r = ±1. This means that the two variables, demand and price, are very closely related. Increasing the price decreases demand; decreasing the price increases demand for coffee.

(b) Is there a significant relationship between demand and price? (Do a complete analysis to justify your conclusions here.) Explain. Use the classical method of testing hypotheses and $\alpha = 5\%$.
(10 pts = 2, 2, 1, 3, 2)

To find out if this correlation is significant, we must perform a hypothesis test.

1. *Hyps:* H_0: r = 0 There is no relationship (i.e., no correlation) between demand and price

H_1: $r \neq 0$ There is a significant relationship (i.e., correlation) between demand and price

2. *Test:*

$$t_{sample} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

3. *Decision Rule*

$$\alpha = .05 \qquad df = n-2 = 8-2 = 6$$
$$t_{critical} = \pm 2.447$$

4. *Analysis*

$$t_{sample} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$
$$= \frac{-.9840\sqrt{8-2}}{\sqrt{1-.9840^2}}$$
$$= \frac{-2.410}{.1782} = -13.52$$

5. *Conclusion:* Reject H_0. There is a significant relationship (i.e., correlation) between demand and price. There is a high positive relationship between them, the higher the price, the higher demand becomes. In making this conclusion, we might be making a Type I error. The probability of making such an error is $\alpha = 5\%$.

(c) Does price predict demand for this coffee? Explain. Assume α = 5% and use the classical method of testing hypotheses.
(10 pts = 2, 2, 1, 3, 2)

Some students will use the formula: $s_e = \sqrt{\frac{SSE}{n-2}}$

where SSE = $SSE = S_y - \frac{SS_{xy}^2}{SS_x}$ = $SSE = 160339.88 - \frac{(-255.35)^2}{0.42}$ =5093.154

$s_e = \sqrt{\frac{5093.154}{6}}$ =29.13

Prof Mishra's students may use the following formulas:

$$s_e = \sqrt{\frac{\sum Y^2 - b_0 \sum Y - b_1 \sum XY}{n-2}} = \sqrt{\frac{6096075 - (2898.11)(6891) - (-607.98)(22829.5)}{8-2}} = 29.09$$

$$s_{b_1} = s_e \sqrt{\frac{1}{SS_X}} = 29.09\sqrt{\frac{1}{.42}} = 44.89$$

1. *Hyps:* H$_0$: $\beta_1 = 0$ There is no slope (and therefore we cannot use the regression line to predict)
H$_1$: $\beta_1 \neq 0$ There is a definite slope (and therefore we can use the regression line to predict)

2. *Test:*

$$t_{sample} = \frac{b_1 - \beta_1}{s_{b_1}} \quad \text{or} \quad t_0 = \frac{b_1 - \beta_1}{s_e \big/ \sqrt{SS_x}}$$

3. *Decision Rule*

$$\alpha = .05 \qquad df = n - 2 = 8 - 2 = 6$$
$$t_{critical} = \pm 2.447$$

4. *Analysis:*

$$t_{sample} = \frac{b_1 - \beta_1}{s_{b_1}}$$
$$= \frac{-607.98 - 0}{44.89} = -13.54$$

5. *Conclusion:* Reject H_0. There is very strong evidence suggesting that the slope deviates from zero. Therefore, we are justified in using the regression equation to predict demand as a function of price. In making this conclusion, we might be making a Type I error. The probability of making such an error is α = 5%.

(d) What is the least squares prediction equation (i.e., the regression equation) here for predicting demand? (*3 pts)*

$$b_1 = \frac{SS_{XY}}{SS_X} = \frac{-255.35}{.42} = -607.98 \quad b_0 = \bar{Y} - b_1\bar{X} = 861.375 - (-607.98)3.35 = 2898.11$$

$$\hat{Y} = b_0 + b_1 X_1$$

$$\hat{Y} = 2898 - 608.0 X_1 \quad or\ equivalently$$

$$DEMAND = 2898.11 - 607.98\, PRICE$$

(e) Are you justified in using the regression equation to make predictions? Explain.***(1 pts)***

Yes. In part (c), we were able to show that the slope of the regression line is significant. Hence the regression line can be used to make predictions, since price helps to explain demand..

(f) Is there a straight-line relationship between price and demand for this coffee? Explain.***(1 pts)***

Yes, there is a straight line relationship between price and demand. In part (b) the correlation coefficient was found to be significant.

(g) The chain had been contemplating selling this coffee for $3.99 per pound. Use the regression equation to predict what demand in one of these stores would have been at this price.

(2 pts)

$$DEMAND = 289.11 - 607.98 PRICE$$
$$= 2898.11 - 607.98(3.99) = 472.27\, pounds$$

Obviously, coffee is very price sensitive. Demand drops off rapidly as price increases.

(h) In making your prediction for this price of $3.99, did you violate any conditions which must be adhered to when using regression techniques to make predictions? Explain briefly. *(2 pts)*

The price of coffee for which we generated predicted demand was \$3.99. The problem is that this price is outside the range of coffee prices from which we generated the regression equation (coffee prices for our sample ranged from \$3.00 to \$3.70 per pound). Within the range of \$3.00 to \$3.70 per pound, the prediction equation is certainly a straight-line relationship (as required for use of Pearson r). Outside this range of prices, however, we have no idea what the shape of the relationship between demand and price takes on. The relationship might actually be curvilinear, and not linear at all. The only way to find out is to take a larger sample that includes coffee prices that include the \$3.99 price. We are therefore on very shaky grounds if we try to predict demand for a price of \$3.99 per pound. Our prediction is likely to be quite wrong.

(i) What is the coefficient of determination in this situation? Explain what it means. ***(2 pts)***

$r^2 = -.984^2 = .9683$ = 96.83% of the total variation in demands can be attributed to differences in the price set for coffee (i.e., coffee is very price sensitive). (NB- Note that demand varies considerably from month to month. Since r^2 = 96.83%, we can account for 96.83% of this variation in demand by knowing the price set for the coffee.) The coefficient of determination is, therefore, a measure of the magnitude of effect. r^2 is also called explained variation (that is, it is the variation in Y scores explained by your predictor).

(j) What assumptions did you have to make to answer part (e)? ***(4 pts)***

Errors are:

1. Normally distributed
2. Independent of one another
3. Variance of errors is constant across values of the predictors X
4. Average error = 0

(k) What is the percentage of total variation in demand that is not explained by the prediction equation (SSE) (i.e., what is the error variation here)? ***(2 pts)***

Error variation = unexplained variation = $1 - r^2 = 1 - .9683 = .0317$ = 3.17% of the total variation among demand scores cannot be accounted for by knowing the price. Recall that r^2 is the coefficient of determination which represents the explained variation.

(l) What do the coefficients in the regression equation mean? Explain what each of the 2 coefficients mean. ***(2 pts)***

For price, the regression coefficient (i.e., slope) is -607.98 which means that for every extra dollar that our coffee costs, demand for the coffee will decrease by 607.98 pounds.

The intercept = 2898.11 means that if coffee is priced at 0 cost demand will be quite high at 2898 pounds.

(m) What is the standard error of estimate in this example?
(2 pts)

29.14

(n) What is the 95% confidence interval for demand if a store were to price its coffee at $3.00 per pound?
(3 pts)

95% CI for X = $3.00

$$\hat{Y} = 2898.11 - 607.98X \ = 2898.11 - 607.98(3.00) = 1074.17$$

$$t = \pm 2.447 \qquad \text{for } \alpha = 5\% \text{ and } df \ = 6$$

$$s_e = 29.09 \qquad \text{from part (c)}$$

$$\sqrt{1+\frac{1}{n}+\frac{(X-\overline{X})^2}{SS_X}} = \sqrt{1+\frac{1}{8}+\frac{(3-3.35)^2}{.42}} = 1.19024$$

$$\hat{Y} \pm t\ s_e\sqrt{1+\frac{1}{n}+\frac{(X-\overline{X})^2}{SS_X}}$$

$$1074.17 \ \pm \ (2.447)(29.09)(1.19024)$$

$$1074.17 \pm 84.73$$

989.44 to 1158.90 pounds of coffee

We can be 95% confident that, if we were to sell this coffee at $3.00 per pound, then we would sell somewhere between 989 and 1159 pounds in total. 5% chance we are wrong.

(o) What is the 95% confidence interval for average demand of all stores offering their coffee at $3.00 per pound?
(3 pts)

95% CI for mean demand when price X = $3.00

$\hat{Y} = 2898.11 - 607.98X = 2898.11 - 607.98(3.00) = 1074.17$

$t = \pm 2.447$ for $\alpha = 5\%$ and $df = 6$

$s_e = 29.09$ from part (c)

$$\sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{SS_X}} = \sqrt{\frac{1}{8} + \frac{(3 - 3.35)^2}{.42}} = .6455$$

$$\hat{Y} \pm t\ s_e \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{SS_X}}$$

$1074.17 \pm (2.447)(29.09)(.6455)$

1074.17 ± 45.95

1028.22 to 1120.12 pounds of coffee

We can be 95% confident that, if we were to sell this coffee at $3.00 per pound, then we would sell in each store an average of somewhere between 989 and 1159 pounds. 5% chance we are wrong.

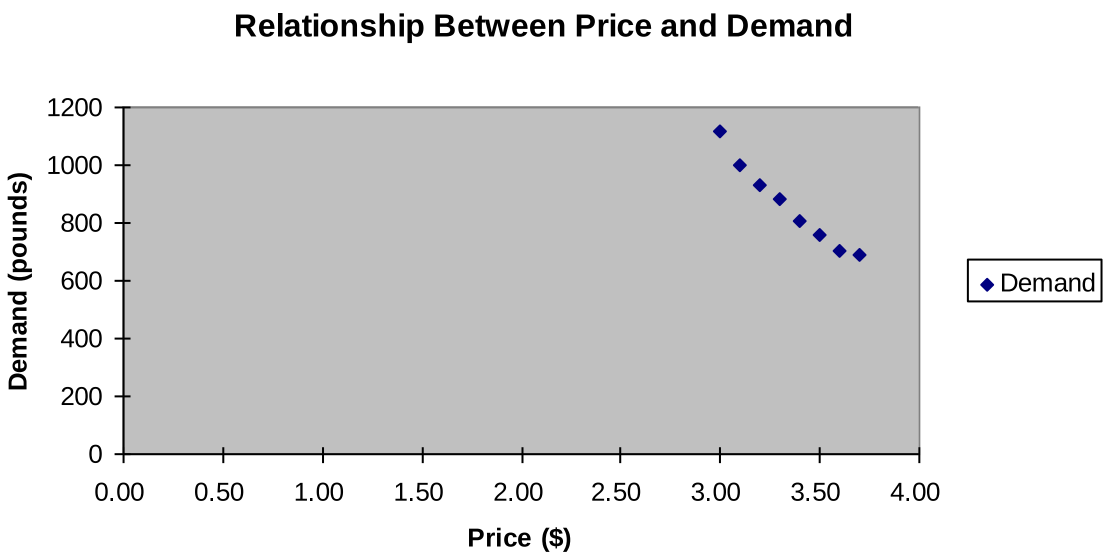
(p) Draw the scatter plot showing the relationship between demand and price. What conclusions can you draw from this diagram? ***(3 pts)***

From the scatter plot below, it is clear that

(1) The relationship is linear (and, therefore, Pearson r can be used here to assess the degree of relationship between price and demand).

(2) Most data points fall very close to a straight line falling through the middle of the data points (this is the regression line). Thus, there is very little error here and consequently the correlation is very high (close to +1.00).

(3) The relationship is negative, i.e., as price increases, demand falls off. Therefore, both Pearson r and the slope must also be negative.


Relationship Between Price and Demand

(q) **Reanalyze the data in this question using Excel. Annotate the printout to show r, r^2, the regression equation, the t and F tests, sample size, etc.**
(5 pts)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98398946
R Square	0.968235258
Adjusted R Square	0.962941134
Standard Error	29.13518709
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	155246.7202	155246.7	182.8887	1.01374E-05
Residual	6	5093.154762	848.8591		
Total	7	160339.875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2898.095238	150.9563691	19.19823	1.29E-06
Price	-607.97619	44.9565697	-13.5236	1.01E-05

The table shows all of the relevant values:

Intercept = 2898.09
Slope = -607.98

Both the slope and the intercept are significant from their t-values.
Further, the f-value is quite strong for a goodness of fit test for the regression.

3. A local express delivery service bases its charge for shipping a package on package weight and distance shipped. The company recently conducted a study to investigate the relationship between the cost of shipment to the delivery company (in dollars) and one of the variables that control the shipping charge to the customer -- package weight (in kilos). Use $\alpha = 5\%$. **(TOTAL = 20 points)**

Package	**Weight**	**Cost**
1	5.9	$ 2.60
2	3.2	$ 3.90
3	4.4	$ 8.00
4	6.6	$ 9.20
5	0.75	$ 4.40
6	0.7	$ 1.50
7	6.5	$ 14.50
8	4.5	$ 1.90
9	0.6	$ 1.00
10	7.5	$ 14.00

Excel printout:
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.73449203
R Square	0.53947854
Adjusted R Square	0.48191336
Standard Error	3.64642371
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	124.61	124.609	9.3716	0.0156
Residual	8	106.37	13.296		
Total	9	230.98			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.3781	2.1962	0.1721	0.8676
Weight	1.4076	0.4598	3.0613	0.01555

Use Excel to answer the following questions: Students are expected to get all of their data from the Excel analysis)

(a) Is the model useful for predicting shipping cost? Explain. Assume α = 5% and use the p-value method of testing hypotheses. ***(2 pts)***

For the model to be useful in making predictions, the independent variable weight must help to explain cost. We must examine the p-value associated with the coefficient b_1. The p-value = 0.0156 or 1.56%. It is significant at 5% level, hence the model is useful in predicting shipping cost.

(b) What is the least squares prediction equation (i.e., the regression equation) here for predicting cost? ***(2 pts)***

SHIPPING COST = 1.41 (WEIGHT) + 0.3781

(c) Are you justified in using the regression equation to make predictions? Explain. ***(2 pts)***

To answer this question, we must perform a goodness of fit test or ANOVA for the regression. The p-value associated with the test (1.56%) is also significant at a 5% significance level implying that the regression line can be used to predict shipping cost.

(d) Is there a straight-line relationship between shipping cost and shipping weight? Explain. ***(2 pts)***

The coefficient of linear correlation is 0.734, which is reasonably strong. The p-value associated with a test for correlation is 1.56% which implies that weight is significant at a 5% level and even at a 2% level. Hence there is sufficient evidence the cost and weight are linearly related.

(e) Use the regression equation to predict the shipping cost of a package that weighs 5 kilos. ***(2 pts)***

Shipping cost = 1.41 x 5 + 0.38 = $7.43.

(f) In making your prediction in part (e), did you violate any conditions which must be adhered to when using regression techniques to make predictions? Explain briefly. ***(2 pts)***

No conditions were violated. 5 kilos is within the range of the weights used in the sample to fit the regression line.

(g) What is the coefficient of determination in this situation? Explain what it means in this situation. ***(2 pts)***

The coefficient of determination r^2 =0.539. This means that the regression line can account for 53.9% of the variation in the values of the cost.

(h) What is R in this example? Explain what it means. ***(2 pts)***

The coefficient of correlation r = 0.734, which measures the strength of the linear relationship between cost and weight.

(i) What is the percentage of total variation in cost that is not explained by the prediction equation (i.e., what is the error variation here)? ***(2 pts)***

The unexplained variation (SSE) is 46.1%. SSY=SSR+SSE.

(j) What do the coefficients in the regression equation mean? Explain what each of the 2 coefficients mean. ***(2 pts)***

Slope: Every additional kilo will increase shipping cost by $1.41
Intercept: The fixed cost associated with shipping a package is $0.38. One may say that a package with no weight will cost $0.38, which does not make sense.