

Grant's Tutoring

BASIC STATISTICS 2

Volume 2 of 2

September 2011 edition



This volume covers the topics taught
after your midterm exam.

Learn What You Need to Know
Know What You Need to Learn

While studying this book, why not hear Grant explain it to you?

Contact Grant for info about purchasing **Grant's Audio Lectures**. Some concepts make better sense when you hear them explained.

Better still, see Grant explain the key concepts in person. Sign up for **Grant's Weekly Tutoring** or attend **Grant's Exam Prep Seminars**. Text or Grant (204) 489-2884 or go to **www.grantstutoring.com** to find out more about all of Grant's services. **Seminar Dates will be finalized no later than Sep. 25 for first term and Jan. 25 for second term.**

HOW TO USE THIS BOOK

I have broken the course up into lessons. Do note that the numbering of my lessons do not necessarily correspond to the numbering of the units in your course outline. Study each lesson until you can do all of my lecture problems from start to finish without any help. Then do the Practise Problems for that lesson. If you are able to solve all the Practise Problems I have given you, then you should have nothing to fear about your exams.

I also recommend you purchase the *Multiple-Choice Problems Set for Basic Statistical Analysis II (Stat 2000)* by Dr. Smiley Cheng available at The Book Store. The appendices of my book include complete step-by-step solutions for all the problems and exams in Cheng's book. Be sure to read the "Homework" section at the end of each lesson for important guidance on how to proceed in your studying.

You also need a good, but not expensive, scientific calculator. Any of the makes and models of calculators I discuss in Appendix A are adequate for this course. I give you more advice about calculators at the start of Lesson 1. **Appendix A in this book shows you how to use all major models of calculators.**

I have presented the course in what I consider to be the most logical order. Although my books are designed to follow the course syllabus, it is possible your prof will teach the course in a different order or omit a topic. It is also possible he/she will introduce a topic I do not cover. **Make sure you are attending your class regularly! Stay current with the material, and be aware of what topics are on your exam. Never forget, it is your prof that decides what will be on the exam, so pay attention.**

If you have any questions or difficulties while studying this book, or if you believe you have found a mistake, do not hesitate to contact me. My phone number and website are noted at the bottom of every page in this book. "Grant's Tutoring" is also in the phone book. **I welcome your input and questions.**

Wishing you much success,

Grant Skene

Owner of Grant's Tutoring and author of this book

FORMULA SHEET

A formula sheet is included in your exams. Check your course syllabus and compare it to the formula sheet I use below in case the formula sheet in your course has changed.

$$1. \quad SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{with df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

$$2. \quad SE(\bar{x}_1 - \bar{x}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{with df} = n_1 + n_2 - 2 \quad \text{if } \sigma_1^2 = \sigma_2^2$$

$$\text{where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$3. \quad SSG = \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2$$

$$4. \quad \text{Poisson Distribution:} \quad P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad k = 0, 1, 2, \dots$$

$$5. \quad t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$6. \quad SE_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}, \quad s_e = \sqrt{MSE}$$

$$7. \quad SE_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

$$8. \quad SE_{\hat{\mu}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$9. \quad SE_{\hat{y}} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$10. \quad SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{if } p_1 = p_2 \quad \text{where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$11. \quad SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad \text{if } p_1 \neq p_2$$

STEPS FOR TESTING A HYPOTHESIS

- Step 1.** State the null and alternative hypotheses (H_0 and H_a), and so determine if the test is 2-tailed, upper-tailed, or lower-tailed.
- Step 2.** Use the given α (always use $\alpha = 5\%$ if none is given) to get the **critical value** (z^* , t^* , F^* , etc. depending on the hypothesis you are testing) from the appropriate table and state the **rejection region**.
- Step 3.** Compute the **test statistic** (z , t , F , etc. depending on the hypothesis you are testing) using the appropriate formula, and see if it lies in the rejection region.
- Step 4.** (Only if specifically asked to do so.) Compute the **P-value**.

Draw a density curve (z -bell curve, t -bell curve, F right-skewed curve, etc. depending on the test statistic you have computed), mark the test statistic (found in Step 3), and shade the area as instructed by H_a . That area is the P -value.

Remember, a P -value is very handy to know if you are asked to make decisions for more than one value of α .

Reject H_0 if P -value $< \alpha$.

- Step 5.** State your conclusion.

Either: Reject H_0 . There is statistically significant evidence that the alternative hypothesis is correct. (Replace the underlined part with appropriate wording from the problem that says H_a is correct.)

Or: Do not reject H_0 . There is no statistically significant evidence that the alternative hypothesis is correct. (Replace the underlined part with appropriate wording from the problem that says we are not convinced that H_a is correct.)

KEY FORMULAS TO MEMORIZE

Lesson 7. To compute the sample proportion, $\hat{p} = \frac{x}{n}$.

The mean and standard deviation of \hat{p} are $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Confidence Interval for p : $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Test statistic for $H_0: p = p_0$ is: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

Sample size determination when estimating p : $n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$.

Confidence Interval for $p_1 - p_2$: $(\hat{p}_1 - \hat{p}_2) \pm z^* SE(\hat{p}_1 - \hat{p}_2)$.

Test statistic for $H_0: p_1 = p_2$ is: $z = \frac{\hat{p}_1 - \hat{p}_2}{SE(\hat{p}_1 - \hat{p}_2)}$.

The formulas for the standard error of $\hat{p}_1 - \hat{p}_2$, $SE(\hat{p}_1 - \hat{p}_2)$, are given on the Formula Sheet given on your exams (page 1 of my book).

Lesson 8. In a two-way table $df = (r - 1) \times (c - 1)$.

In goodness-of-fit, $df = k - 1 - 1$ more for each estimated parameter.

In a two-way table:

Expected count for a cell = $\frac{(\text{The cell's Row Total}) \times (\text{The cell's Column Total})}{\text{The Grand Total}}$

In a goodness-of-fit test:

Expected count is found using the given probability distribution

Expected count = each probability \times the total of the observed counts

To compute each cell's chi-square: $\chi_{\text{cell}}^2 = \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}}$

The chi-square test statistic is: $\chi^2 = \sum \chi_{\text{cell}}^2$

SUMMARY OF KEY CONCEPTS IN LESSON 7

❖ If we need to compute a sample proportion, \hat{p} , we are usually given a value of n followed by x , the number of yeses in our question of interest. Then $\hat{p} = \frac{x}{n}$.

❖ The distribution of \hat{p} :

▪ The mean of $\hat{p} = \mu_{\hat{p}} = p$.

▪ The standard deviation of $\hat{p} = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

▪ If $np \geq 10$ and if $n(1-p) \geq 10$, then \hat{p} has an approximately normal distribution.

❖ If we want to find the probability the sample proportion \hat{p} is above, below or between some given amount(s), we can bet our Rule of Thumb will tell us \hat{p} is approximately normal, so we can use a **\hat{p} -bell curve** to compute the approximate probability.

▪ The standardizing formula for \hat{p} is $z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$.

❖ **A confidence interval for p is** $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

❖ When you are asked to determine the sample size necessary to achieve a desired margin of error in a confidence interval for a *proportion*, use this formula:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

▪ If you have no idea what to use for p^* , use **the conservative estimate**.
Let $p^* = 50\% = .5$.

❖ If you want to test a hypothesis about a *proportion*, p , use this formula:

▪ **The test statistic for $H_0: p = p_0$ is** $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

❖ If we are interested in comparing two proportions, p_1 and p_2 , we must first estimate these two proportions.

▪ If our first sample is size n_1 , and x_1 are the number who said “yes” to our question, then $\hat{p}_1 = \frac{x_1}{n_1}$ is our estimate of p_1 . If our second sample is size n_2 , and x_2 are the

number who said “yes” to our question, then $\hat{p}_2 = \frac{x_2}{n_2}$ is our estimate of p_2 .

▪ If $n_1 p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2 p_2 \geq 5$, and $n_2(1-p_2) \geq 5$ then it is reasonable to assume $\hat{p}_1 - \hat{p}_2$, the difference between the two sample proportions, has an approximately normal distribution, allowing us to use z-scores.

❖ The test statistic formula for $H_0: p_1 = p_2$ is $z = \frac{\hat{p}_1 - \hat{p}_2}{SE(\hat{p}_1 - \hat{p}_2)}$.

❖ A confidence interval for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z * SE(\hat{p}_1 - \hat{p}_2)$.

❖ Be sure to use the correct formula for $SE(\hat{p}_1 - \hat{p}_2)$, the standard error of $\hat{p}_1 - \hat{p}_2$. These formulas are given on your formula sheet, so you do not have to memorize them.

▪ A hypothesis test for the difference in proportions will always assume $p_1 = p_2$ since that is what the null hypothesis believes. Consequently, you must compute the pooled sample proportion, $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$, and proceed to work out the

pooled standard error:

$$\bullet SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

▪ A confidence interval for the difference between two proportions, $p_1 - p_2$, will never use the pooled standard error since it does not believe the two proportions are equal. Confidence intervals will use the standard error:

$$\bullet SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

LECTURE PROBLEMS FOR LESSON 7

For your convenience, here are the 11 questions I used as examples in this lesson. Do not make any marks or notes on these questions below. Especially, do not circle the correct choice in the multiple choice questions. You want to keep these questions untouched, so that you can look back at them without any hints. Instead, make any necessary notes, highlights, etc. in the lecture part above.

- 1.** In Big City only 35% of the voters in the last election were in favour of a one-time levy to cover the cost of sewer upgrades. During the current campaign a random sample of 575 voters will be selected. Assume the opinion has not changed since the last election.

(a) What is the mean and standard deviation of the number of voters sampled in favour of the levy?

- (A)** 0.35; 0.020 **(B)** 0.65; 0.020 **(C)** 201.25; 11.44
(D) 373.75; 11.44 **(E)** 350; 13.46

(See the solution on page 386.)

(b) What is the mean and standard deviation of the proportion of voters sampled in favour of the levy?

- (A)** 0.35; 0.020 **(B)** 0.65; 0.020 **(C)** 201.25; 11.44
(D) 373.75; 11.44 **(E)** 350; 13.46

(See the solution on page 386.)

(c) What is the probability the sample proportion will be between 30% and 40% in favour of the levy?

(See the solution on page 387.)

(d) What is the probability a random sample of 575 voters finds at least 225 in favour of the levy?

(See the solution on page 388.)

- 2.** In a random sample of 450 homes it was found, in 90 cases, at least one resident watches *Survivor*. Give a 95% confidence interval for the proportion of homes that watch this program.

- (A)** (.199, .201) **(B)** (.163, .237) **(C)** (.169, .231)
(D) (.156, .244) **(E)** (.160, .240)

(See the solution on page 391.)

3. A government claims at least two-thirds of the people who claim refugee status are approved. An independent organization randomly selected 320 refugee claimants whose case had already been adjudicated. Of these, 180 had been given refugee status. Justifying any methods you use, is this significant evidence to reject the government's claim at the 5% level? Include the hypotheses, test statistic, critical value, P -value, and your conclusion in your answer.

(See the solution on page 394.)

4. It is believed a majority of farmers own their farms. In testing this hypothesis, 600 randomly selected farms revealed 315 were owned by the farmers. At the 1% level of significance, the test statistic and critical value are, respectively,

(A) $\frac{.025}{\sqrt{\frac{(.525)(.475)}{600}}}$ and 1.960

(B) $\frac{.025}{\sqrt{\frac{(.525)(.475)}{600}}}$ and 2.326

(C) $\frac{.025}{\sqrt{\frac{(.525)(.475)}{600}}}$ and 2.576

(D) $\frac{.025}{\sqrt{\frac{(.5)(.5)}{600}}}$ and 2.326

(E) $\frac{.025}{\sqrt{\frac{(.5)(.5)}{600}}}$ and 2.576

(See the solution on page 395.)

5. From long experience, it is known a machine produces 30% defective tubes. After a modification is made to the machine, a random sample of 100 tubes is selected. It is found 22 of the sampled tubes are defective.

(a) In testing if the modification has improved the machine, what is the P -value for the test?

- (A) .0268 (B) .0536 (C) .2200 (D) .0401 (E) .0802

(See the solution on page 396.)

(b) A 90% confidence interval for the proportion of defective tubes after the modification is:

- (A) $.22 \pm 1.96 \sqrt{\frac{(.30)(.70)}{100}}$ (B) $.22 \pm 1.96 \sqrt{\frac{(.22)(.78)}{100}}$
(C) $.22 \pm 1.645 \sqrt{\frac{(.30)(.70)}{100}}$ (D) $.30 \pm 1.645 \sqrt{\frac{(.30)(.70)}{100}}$
(E) $.22 \pm 1.645 \sqrt{\frac{(.22)(.78)}{100}}$

(See the solution on page 397.)

6. In the last election, the mayor received 60% of the vote. He wishes to determine what proportion of the electorate plans to vote for him in the upcoming election. Assuming the percentage is about the same, how many voters should be polled to estimate the proportion within 2% with 95% confidence?

- (A) 61 (B) 106 (C) 172 (D) 122 (E) 2305

(See the solution on page 398.)

7. In order to estimate the unknown proportion of people who attend church on a regular basis, how large a sample do you need to draw if you want to construct a 90% confidence interval of width no more than 0.04?

- (A) 1692 (B) 1691 (C) 3301 (D) 3328 (E) 3394

(See the solution on page 399.)

8. A random sample of 500 14 year-old Canadian children found 65 were obese. Another random sample of 750 14 year-old Japanese children found 75 were obese.

(a) At the 5% level, is there significant evidence 14 year-old Canadian children are more likely to be obese than their Japanese counterparts?

(See the solution on page 404.)

(b) Construct a 90% confidence interval for the difference in proportions of obese Canadian children and obese Japanese children.

(See the solution on page 405.)

9. Two different methods of manufacture, casting and die forging, were used to make parts for an appliance. In service tests of 100 of each type it was found ten castings failed during the test, but only three forged parts failed. A 95% confidence interval for the difference between the proportions of the cast and forged parts that would fail under similar conditions would be:

$$(A) \left(\frac{10}{100} - \frac{3}{100} \right) \pm 1.96 \sqrt{\frac{(10/100)(90/100)}{100} - \frac{(3/100)(97/100)}{100}}$$

$$(B) \left(\frac{10}{100} - \frac{3}{100} \right) \pm 1.96 \sqrt{\frac{(10/100)(90/100)}{100} + \frac{(3/100)(97/100)}{100}}$$

$$(C) \left(\frac{10}{100} - \frac{3}{100} \right) \pm 1.645 \sqrt{\frac{(10/100)(90/100)}{100} + \frac{(3/100)(97/100)}{100}}$$

$$(D) \left(\frac{10}{100} - \frac{3}{100} \right) \pm 1.645 \sqrt{\left(\frac{13}{200} \right) \left(\frac{187}{200} \right) \left(\frac{1}{100} + \frac{1}{100} \right)}$$

$$(E) \left(\frac{10}{100} - \frac{3}{100} \right) \pm 1.96 \sqrt{\left(\frac{13}{200} \right) \left(\frac{187}{200} \right) \left(\frac{1}{100} + \frac{1}{100} \right)}$$

(See the solution on page 406.)

10. In order to test whether an experimental vaccine decreases the mortality rate for a particular cow disease, 400 infected cows were randomly selected and 200 of them were randomly chosen to receive the vaccine. Among those vaccinated, 20 still died of the disease while there were 30 deaths in the control group. What is the P -value for the appropriate hypothesis test?

(A) 0.1 (B) 0.1286 (C) 0.0643 (D) 0.1310 (E) 0.0655

(See the solution on page 408.)

11. A random sample of 1800 Manitoba families revealed 292 were below the national poverty line while 308 out of a random sample of 1650 Newfoundland families were below the national poverty line. Construct a 93% confidence interval for the difference in proportions of families below the poverty line in these two provinces.

(See the solution on page 410.)

SUMMARY OF KEY CONCEPTS IN LESSON 8

- ❖ If given a **two-way table**, one thing we can be asked to do is compute various proportions. To be able to do so, we must first compute the marginal totals (both the row totals and column totals) and the grand total.
 - A **marginal proportion** is the appropriate marginal total divided by the grand total.
 - A **joint proportion** is the appropriate cell's observed count divided by the grand total.
 - A **conditional proportion** is the appropriate cell's observed count divided by the appropriate marginal total.
- ❖ If we are given a **two-way table**, we will perform either a **chi-square test for independence** or a **chi-square test for homogeneity**.
 - If a two-way table summarizes two or more independent samples' response to one question, you are preparing for a test of homogeneity.
 - If a two-way table summarizes one random sample's responses to two different questions, you are preparing for a test for independence.
 - There are several ways we can phrase the hypotheses for a two-way table, but they all essentially same something like this:
 - H_0 : There is no relationship between the rows and columns.
 - H_a : There is a relationship between the rows and columns.
- ❖ The degrees of freedom for the chi-square two-way-table test are $(r - 1) \times (c - 1)$.
- ❖ The expected count for each cell =
$$\frac{(\text{The cell's Row Total}) \times (\text{The cell's Column Total})}{\text{The Grand Total}}$$
- ❖ The chi-square value in each cell:
$$\chi_{\text{cell}}^2 = \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}}$$
- ❖ The chi-square test statistic =
$$\chi^2 = \sum \chi_{\text{cell}}^2 = \sum \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}}$$

- ❖ If we are given a discrete probability distribution model (a list of categories together with the probability of each category), we can investigate whether a random sample appears to follow this distribution by performing a **chi-square goodness-of-fit test**.
 - The hypotheses will have this general form:
 - H_0 : The data does fit the suggested probability distribution.
 - H_a : The data does not fit the suggested probability distribution.
- ❖ The degrees of freedom for the chi-square goodness-of-fit test are $k - 1 - 1$ more for each estimated parameter where k is the number of cells in the distribution table.
- ❖ To compute the **expected counts in a goodness-of-fit problem**, first find the total of the observed counts, then multiply that total by each probability in the model to get the expected count for each cell.
- ❖ **The expected count in any cell must be at least 5 for a chi-square test to be reliable. If this is not the case in a goodness-of-fit problem, we must combine neighbouring cells together until the condition is satisfied.**
- ❖ **If the suggested probability model is a ratio**, add the numbers in the ratio together. Now, divide each number in the ratio by the total of the ratios to get the probabilities for each category in the model.
- ❖ **If the suggested model is a binomial distribution**, use Table C to determine the probabilities of $k = 0, 1, 2, \dots, n$ for the given parameters n and p .
 - **If p is not given in the binomial distribution**, estimate p from the given data.
(See question 9 above for an example of how to do this.)
- ❖ **If the suggested model is a Poisson distribution**, use the Poisson probability formula (given on your formula sheet) to determine the probabilities of $k = 0, 1, 2, \dots$ for the given parameter λ .
 - **If λ is not given in the Poisson distribution**, estimate λ from the given data.
(See question 11 above for an example of how to do this.)
- ❖ If you have to estimate p in a binomial goodness-of-fit problem, or if you have to estimate λ in a Poisson goodness-of-fit problem, you will lose one more degree of freedom because you have had to estimate a parameter. That will make your degrees of freedom $k - 2$.

LECTURE PROBLEMS FOR LESSON 8

For your convenience, here are the 11 questions I used as examples in this lesson. Do not make any marks or notes on these questions below. Especially, do not circle the correct choice in the multiple choice questions. You want to keep these questions untouched, so that you can look back at them without any hints. Instead, make any necessary notes, highlights, etc. in the lecture part above.

- 1. A random survey was conducted in which 180 bank clerks, 320 construction workers, and 240 school teachers were each asked if they were satisfied with their jobs. The responses are summarized below.

Table with 4 columns: Job, Satisfaction Level (Low, Moderate, High). Rows: Bank Clerk, Construction, Teacher.

- (a) Test the hypothesis job satisfaction is independent of type of job at the 5% level.
(b) Put bounds on the P-value for this problem.

(See the solution on page 435.)

- 2. A study of 500 members of a particular population was conducted to determine aspects of their consumer behaviour. One question asked was "Do you enjoy shopping?" The table at right summarizes the responses for men and women. In testing the hypothesis a person's attitude towards shopping is unaffected by their gender at a 1% level of significance:

Table with 4 columns: Gender, Never, Sometimes, Always. Rows: Men, Women.

- (a) What are the degrees of freedom and critical value for the test?
(b) What is the expected number of men who always enjoy shopping?
(c) What is the test statistic for this hypothesis?
(d) In testing the hypothesis the proportion of men who never enjoy shopping is higher than the proportion of women who never enjoy shopping, the test statistic would be:

(See the solution on page 436.)

3. A national survey was conducted to obtain information on alcohol consumption patterns of U.S. adults by marital status. A random sample of 1772 residents, 18 years old or older, yielded the data displayed in the following table. We want to see whether there is an association between marital status and alcohol consumption.

NOTE: (1) Expected values for some cells are enclosed in brackets.
 (2) Cell χ^2 values are given in the second row of some cells.
 (3) The sum of the given χ^2 values in the table is 63.49.

Marital Status	Drinks per month			Total
	Abstain	1-60	Over 60	
Single	67 (117.9) 21.95	213 () 2.49	74 () 18.78	354
Married	411 (390.6) 1.07	633 (633.5) .00	129 (148.9) 2.67	1173
Widowed	85 (47.6)	51 () 8.91	7 () 6.86	143
Divorced	27 (34)	60 (55.1) .44	15 (13.0) 0.32	102
Total	590	957	225	1772

- (a) The null hypothesis we usually test for data such as this is:
 (A) Alcohol consumption depends on marital status.
 (B) The four categories of marital status are equally likely for all alcohol consumption categories.
 (C) Married people drink less than unmarried people.
 (D) Alcohol consumption is independent of marital status.
 (E) None of the above.
- (b) The expected frequency for the (3,2) cell (number of widowed people who have 1-60 drinks per month) is:
 (A) 77.2 (B) 51 (C) 47.7 (D) 147.7 (E) 239.5
- (c) The degrees of freedom and 5% critical value for the appropriate test statistic are:
 (A) 12 & 21.03 (B) 2 & 5.99 (C) 11 & 19.68 (D) 6 & 14.45 (E) 6 & 12.59
- (d) The value of the appropriate test statistic is:
 (A) 18.27 (B) 30.83 (C) 64.06 (D) 94.32 (E) 81.75
 (See the solution on page 442.)

4. A survey was conducted to investigate whether there is a relationship between alcohol drinking and smoking. The information at right was compiled from a simple random sample of 600 individuals.

	Smoker	Non-Smoker
Drinker	193	165
Non-Drinker	89	153
Total	282	318

- (a) There are two possible techniques we could use to test our hypothesis. Assuming we are using a 5% level of significance, identify the two techniques and list their critical values.
- (b) Find the test statistics for both techniques. Is there a relationship between alcohol drinking and smoking?
- (c) Test the hypothesis that smokers are more likely to drink alcohol at the 5% level, including the test statistic, critical value and *P*-value in your analysis.
- (d) What is a 90% confidence interval for the proportion of people who are non-smokers?
- (e) What is a 98% confidence interval for the difference in the proportion of smokers who drink and the proportion of non-smokers who drink?

(See the solution on page 445.)

5. A researcher believes the blood types for a particular population will have this distribution:

Type A	Type B	Type AB	Type O
37%	13%	6%	44%

A random sample of 1000 people found 403 had Type A blood, 102 had Type B, 37 Type AB, and 458 Type O. Perform a chi-square goodness-of-fit test on this information.

(See the solution on page 455.)

6. An office manager has kept track of the number of employees who have called in sick on each day of the week for the past two years. Here are his records:

Day	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Sick Employees	123	96	71	77	102

Test this distribution for homogeneity, and include a *P*-value in your conclusion.

(See the solution on page 458.)

7. According to genetic theory, a hybrid generation of a certain plant should show a segregation of colours in the ratio *1 red: 3 pink: 1 white*. When such an experiment was performed with 200 plants, it resulted in 44 red, 110 pink, and 46 white flowers. In testing if these results are consistent with genetic theory at $\alpha = .05$, the critical value and test statistic are, respectively
- (A)** 5.99; 2.13 **(B)** 7.38; 2.13 **(C)** 7.38; 2.06 **(D)** 7.81; 2.06 **(E)** 5.99; 2.06

(See the solution on page 459.)

8. A class of 200 students each tossed five coins and recorded the number of heads as given below. Test the hypothesis this data came from a binomial distribution with parameters $n = 5$ and $p = 1/2$.

Number of Heads	Frequency
0	6
1	24
2	66
3	69
4	27
5	8

(See the solution on page 461.)

9. A statistician goes through four intersections with traffic lights on her way to work each day. Over a period of six months, she records the number of red lights she encounters on her way to work. The data are as follows:

# of Red Lights	# of Days
0	33
1	39
2	47
3	17
4	14

- (a)** In what proportion of the intersections did she encounter a red light?
- (b)** Conduct a chi-square goodness-of-fit test at the 2.5% level of significance to determine if the number of red lights encountered on her trip to work each day follows a binomial distribution.

(See the solution on page 463.)

- 10.** Consider the following distribution of the number of goals a certain hockey team scored in each of their 80 games last season. Test the hypothesis the number of goals per game has a Poisson distribution with parameter of $\lambda = 2.8$.

Number of Goals	Number of Games
0	4
1	9
2	19
3	23
4	13
5	6
6	3
7	2
8	0
9	1

(See the solution on page 470.)

- 11.** Stenographers are tested on their typing accuracy. A sample of 100 pages is randomly selected for a particular stenographer, and an inspector counts the number of typos on each page. The data are as follows:

# of Typos	# of Pages
0	31
1	41
2	20
3	5
4	2
5	0
6	1

- (a)** Compute the average number of typos per page in this sample.
(b) Conduct a chi-square goodness-of-fit test at the 1% level of significance to determine if the number of typos per page follows a Poisson distribution.

(See the solution on page 472.)

SUMMARY OF KEY CONCEPTS IN LESSON 9

- ❖ A regression analysis can be performed on (x, y) data pairs provided **both x and y are quantitative variables**.
- ❖ The **explanatory variable** is x and the **response variable** is y . We believe x can explain y 's response. We hope to use x to predict y .
- ❖ We use a **scatterplot** to see if x and y have either **a positive or a negative association** (or neither), and if the trend (if any) is **linear** or **nonlinear**.
 - A rising trend is a positive association (y gets larger as x gets larger).
 - A falling trend is a negative association (y gets smaller as x gets larger).
- ❖ If we do believe we have a linear trend, we can confirm it by computing the **correlation coefficient, r** .

- $$r = \frac{1}{n-1} \sum \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right]$$

- **r measures the strength and direction of the linear relationship between x and y .**
 - **$-1 \leq r \leq 1$; the closer r is to -1 or 1 , the stronger the linearity.**
 - **r has no units.**
 - r will be identical regardless of which of the two variables is considered x and which is considered y .
 - If you change your mind as to which units to use when measuring x and y , the value of r will not change; r is independent of the units used for x and y .
- ❖ The **least-squares regression equation** is $\hat{y} = a + bx$. This equation draws the best fit line through our scatterplot. This can also be called the “fitted model”.
 - **$b = r \frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$**
 - The **intercept, a** , is where the regression line intercepts the “ $x = 0$ ” line
 - **a is the predicted value of y when $x = 0$.**
 - The **slope, b** , is **Rise/Run**; as x runs 1 unit, y rises (or falls) b units.
 - **b is the amount y changes for each additional unit of x .**

- ❖ The **coefficient of determination** is r^2
 - r^2 is the percent of the y 's variation explained by the regression with x .
- ❖ Extending the regression line beyond the scatterplot in an attempt to predict y for an x value beyond the data range is an **extrapolation**. Extrapolations are unreliable because there is no guarantee the linear trend continues.
- ❖ **Residual** = $e = y - \hat{y}$
 - A residual is the difference between the observed value of y and the predicted value of y .
 - If x and y truly have a linear relationship, a residual plot will have no pattern.
 - If a residual plot has a pattern, x and y could not have a linear relationship; linear regression should not have been attempted.
 - The sum of the residuals is 0.
 - The sum of the squares of the residuals has been minimized. That is what we mean by “least-squares” regression or the method of “least-squares”.
 - The variance of the residuals is $s_e^2 = MSE = \frac{\sum(\text{residuals})^2}{n-2}$.
- ❖ The model for least-squares regression is “ $y_i = \alpha + \beta x_i + \varepsilon_i$ ”.
 - This also means that “ $\mu_y = \alpha + \beta x$ ”.
 - Least-squares regression has three parameters α , the true intercept; β , the true slope; and σ_ε , the standard deviation of the residuals. Our estimates for these parameters are a , b and s_e , respectively.
- ❖ When we find an unexpected correlation between x and y , it is undoubtedly due to a **lurking variable**.
 - A strong correlation between children's *test scores* and *shoe size* is because *age* is a lurking variable, for example.
- ❖ Any observation on a scatterplot that is far to the left or right of the cluster could be an **influential observation**. If removing that observation considerably changes the least-squares regression equation (changes the value of a and/or b), it is definitely an influential observation.

LECTURE PROBLEMS FOR LESSON 9

For your convenience, here are the 7 questions I used as examples in this lesson. Do not make any marks or notes on these questions below. Especially, do not circle the correct choice in the multiple choice questions. You want to keep these questions untouched, so that you can look back at them without any hints. Instead, make any necessary notes, highlights, etc. in the lecture part above.

1. A researcher believes that the fuel economy of a car can be predicted by its speed. A car was driven on an oval track for two hours at five different speeds and then had its fuel economy measured. The results, together with their means, were tabulated as shown below.

						Mean
Fuel Economy (mpg)	19	13	12	8	7	11.8
Speed (mph)	10	20	30	40	50	30

- (a) Identify the explanatory and response variables. *(See the solution on page 499.)*
- (b) Make a scatterplot and comment on its pattern (if any). *(Solution on page 501.)*
- (c) Compute the correlation coefficient and interpret it. *(Solution on page 502.)*
- (d) Compute the coefficient of determination and interpret it. *(Solution on page 505.)*
- (e) Determine the least-squares regression equation and draw its line on your scatterplot found in (b). *(Solution on page 508.)*
- (f) Interpret the slope of the least-squares regression line. *(Solution on page 509.)*
- (g) Predict the fuel economy at a speed of 35 mph, or explain why this is not a reasonable calculation. *(Solution on page 510.)*
- (h) Predict the fuel economy at a speed of 80 mph, or explain why this is not a reasonable calculation. *(Solution on page 510.)*
- (i) Without calculation, predict the fuel economy at a speed of 30 mph.
(See the solution on page 511.)
- (j) Compute the residual for the above data when the speed was 20 mph.
(See the solution on page 512.)
- (k) Compute the variance of the residuals, s_e^2 .
(See the solution on page 522.)

2. Which of the following statements does not contain a blunder?
- (A) A study found a strong positive correlation (0.89) between the colour of a person's shirt and the number of mosquito bites received in a 3 hour period.
 - (B) There is a strong negative correlation (-1.23) between the number of alcoholic beverages consumed the night before an exam and the mark achieved on the exam.
 - (C) The correlation between a man's leg length and his time to climb two flights of stairs is 0.32 minutes.
 - (D) The correlation between the height of a student and their GPA is -0.53 suggesting that taller students tend to have a lower GPA than shorter students.
 - (E) The correlation between elementary school children's shoe size and their score on a standard spelling test is 0.85 suggesting that having big feet causes you to spell better.

(See the solution on page 524.)

3. In each case below a regression analysis was performed on data. An alteration was made to the data, and a regression was done again. Which of the following alterations would give two different values of r , the correlation coefficient?
- (A) A regression on people's height (in inches) and weight (in pounds) was performed. Now the same data is converted into metric units (centimetres and kilograms, respectively), and the regression is performed again.
 - (B) A regression on height versus weight used height as the explanatory variable. Now the same data is used, but we use weight as the explanatory variable.
 - (C) A regression on automobile deaths for the years 1991, 1992, to 2004 was performed. Now the same data is used, but we decide to code 1991 as year 1, 1992 as 2, etc.
 - (D) A study on crop yield (bushels per acre) versus amount of fertilizer (litres per acre, l/a) was designed. Ten farms used no fertilizer; ten used 1 l/a; ten used 2 l/a; ten used 3 l/a; and ten used 4 l/a. The crop yields for the 50 farms were recorded and a regression analysis was performed. Now the same data is used, but instead of using the individual farms, the mean crop yield for the ten farms receiving the same amount of fertilizer is computed, giving us an average crop yield for no fertilizer, 1 l/a, 2 l/a, 3 l/a, and 4 l/a. A regression on mean crop yield versus amount of fertilizer is performed on these 5 values.
 - (E) None of the above.

(See the solution on page 526.)

4. Below is the *JMP*TM output for 12 random samples of a particular model of used car. The selling price of the particular car (in dollars) and the distance travelled according to its odometer reading (in thousands of kilometres) were recorded.

Variable	Mean	Std Dev	Correlation	Number
Odometer Reading	48	6.96	-0.87	12
Selling Price	6500	1997.52		

- (a) The appropriate least-squares regression equation would be:
- (A) $\hat{y} = 67.70 + 0.0030x$ (B) $\hat{y} = 18,485.12 - 249.69x$
 (C) $\hat{y} = -5485.12 - 249.69x$ (D) $\hat{y} = 67.70 - 0.0030x$
 (E) $\hat{y} = 28.5 - 0.0030x$

(See the solution on page 528.)

- (b) Which one of the following statements is TRUE?
- (A) A car with a lot of kilometres on its odometer tends to be more expensive.
 (B) 87% of the selling price is predicted by the regression equation.
 (C) -87% of the selling price is predicted by the regression equation.
 (D) The selling price drops by 25 cents per kilometre on the odometer.
 (E) The selling price drops by 250 dollars per kilometre on the odometer.

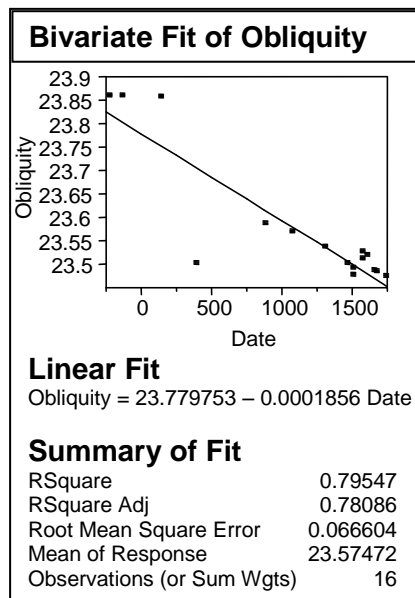
(See the solution on page 529.)

5. A least-squares regression equation for a random sample is $\hat{y} = 120 - 12.5x$. The explanatory variable is the amount of alcohol consumed (in ounces) and the response variable is the score on a physical coordination test. Participant A in the study consumed 2 ounces of alcohol and scored 100 on the test. Which of the following statements is false?
- (A) We would predict a person who drinks no alcohol would score 120 on the test.
 (B) For each additional ounce of alcohol consumed, we would predict the score would drop 12.5 points.
 (C) The residual for Participant A is 5.
 (D) The sum of the squares of the residuals has been minimized.
 (E) There is clearly a strong negative linear correlation between the score on a physical coordination test and the amount of alcohol consumed.

(See the solution on page 530.)

6. The angle of the earth's rotation is called its "obliquity" and is measured in degrees. Measurements have been made at various dates in earth's history. Let us assume that the measurements were accurate. A regression of the obliquity by date was performed by JMP™ and the results are at right. Which of the following statements are false?

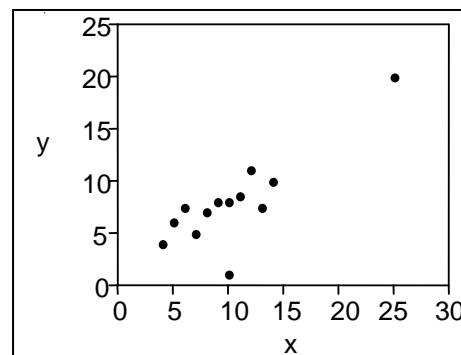
- (A) The measurement made about the year 490 is an outlier.
- (B) Each year the obliquity increases by 0.0001856 degrees.
- (C) The correlation between date and obliquity is -0.8919 .
- (D) The measurement at about the year 900 has a negative residual.
- (E) About 79.5% of the variation in the earth's obliquity can be explained by this regression line.



(See the solution on page 532.)

7. For the scatterplot shown at right, circle and label any point you believe is an outlier, influential observation, or both.

(See the solution on page 534.)



SUMMARY OF KEY CONCEPTS IN LESSON 10

- ❖ In **simple linear regression**, t has $n - 2$ degrees of freedom.
- ❖ There are four confidence intervals we can be asked to make in simple linear regression.
 - $a \pm t * SE_a$ is a confidence interval for the intercept, α .
 - $b \pm t * SE_b$ is a confidence interval for the slope, β .
 - $\hat{y} \pm t * SE_{\hat{y}}$ is a confidence interval for y , an individual observation, at a given value, x^* . This is also called the **prediction interval**.
 - $\hat{y} \pm t * SE_{\hat{\mu}}$ is a confidence interval for μ_y , the mean or average value of y , at a given value, x^* .
- ❖ There are four different hypothesis tests we can be asked to do in simple linear regression.
 - We can test the **hypothesis of zero correlation** ($H_0: \rho = 0$) using the test statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (given as #5 on your formula sheet).
 - We can test the **hypothesis of zero slope using the ANOVA F statistic**.
 - $H_0: \beta = 0$ vs. $H_a: \beta \neq 0$
 - $DFM = 1, DFE = n - 2$
 - $F = \frac{MSM}{MSE}$
 - We can test the **hypothesis of zero slope** ($H_0: \beta = 0$) **using t** with $df = n - 2$.
 - With t we can do an upper-tailed, lower-tailed, or two-tailed test.
 - $t = \frac{b}{SE_b}$
 - We can test the **hypothesis of zero intercept** ($H_0: \alpha = 0$) using t with $df = n - 2$.
 - With t we can do an upper-tailed, lower-tailed, or two-tailed test.
 - $t = \frac{a}{SE_a}$
- ❖ The two test statistics for zero slope are related: $(\text{slope's } t)^2 = F$.

- ❖ Both the test of zero correlation and zero slope answer the question, “Is there any evidence of a linear relationship between x and y ?” If you reject H_0 , you have significant evidence the correlation is not zero, or the slope is not zero. Either of these conclusions tell you there is significant evidence of a linear relationship between x and y . **The t test statistics for zero correlation and zero slope are identical in value.**
- ❖ The coefficient of determination = $r^2 = \frac{SSM}{SST}$. This is the percentage of y 's variation explained by the linear regression model. This can also be denoted R^2 .
- ❖ Given a **multiple linear regression** with p explanatory variables, x_1, x_2, \dots, x_p :
 - The model is $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$
 - This also means $\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$
 - The fitted model is $\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$
- ❖ Make sure you know the four conditions that must be satisfied for a multiple linear regression model to be reliable.
- ❖ In a multiple linear regression model the t distribution has $n - k$ degrees of freedom where k is the number of variables in the model, including y , the response variable.
- ❖ The ANOVA F test in multiple linear regression is doing a “Whole Model Test”.
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ vs.
 - H_a : at least one of the coefficients, β_i , is not 0.
 - $DFM = k - 1, DFE = n - k$
 - $F = \frac{MSM}{MSE}$
- ❖ We can test hypotheses for a specific coefficient in multiple linear regression using t with $n - k$ degrees of freedom, using the same test statistic formula as for simple linear regression.
- ❖ We can make confidence intervals for the intercept and multiple linear regression coefficients using the same formulas as we do for simple linear regression. Again, t has $n - k$ degrees of freedom.

LECTURE PROBLEMS FOR LESSON 10

For your convenience, here are the 4 questions I used as examples in this lesson. Do not make any marks or notes on these questions below. Especially, do not circle the correct choice in the multiple choice questions. You want to keep these questions untouched, so that you can look back at them without any hints. Instead, make any necessary notes, highlights, etc. in the lecture part above.

1. A researcher believes that the fuel economy of a car can be predicted by its speed. A car was driven on an oval track for two hours at five different speeds and then had its fuel economy measured. The results are tabulated below.

$y =$ Fuel Economy (mpg)	19	13	12	8	7
$x =$ Speed (mph)	10	20	30	40	50

The least-squares regression equation is $\hat{y} = 20.5 - 0.29x$.

You are also given $r = -0.9624$, $\bar{x} = 30$, $s_x = 15.8114$, $\bar{y} = 11.8$, $s_y = 4.7645$, and $s_e^2 = 2.2333$.

- (a) Test the hypothesis of zero correlation at the 1% level of significance. Include the hypotheses, test statistic, critical value, and P -value in your solution. Is there strong evidence of a linear relationship between fuel economy and speed?
(See the solution on page 554.)
- (b) Test the hypothesis of zero slope using t at the 1% level of significance. Include the hypotheses, test statistic, critical value, and P -value in your solution. Is there strong evidence of a linear relationship between fuel economy and speed?
(See the solution on page 557.)
- (c) Compare the values you got for the test statistic in parts (a) and (b).
(See the solution on page 557.)
- (d) Test the hypothesis of zero slope using F at the 1% level of significance. Include the hypotheses, test statistic, critical value, and P -value in your solution. Is there strong evidence of a linear relationship between fuel economy and speed?
(See the solution on page 560.)
- (e) Compare the values you got for the test statistic in parts (b) and (d).
(See the solution on page 561.)

Question 1 continued

1. (f) Test the hypothesis the slope is negative, using all appropriate methods, at the 5% level of significance. Include the hypotheses, test statistic, critical value, and *P*-value in your solution.

(See the solution on page 563.)

- (g) Make a 95% confidence interval for the slope, and interpret it.

(See the solution on page 563.)

- (h) Make a 95% confidence interval for the intercept.

(See the solution on page 564.)

- (i) Make a 95% confidence interval for the fuel economy of a car travelling at 25 mph, and interpret it.

(See the solution on page 565.)

- (j) Make a 95% confidence interval for the average fuel economy of a car travelling at 25 mph, and interpret it.

(See the solution on page 566.)

- (k) What assumptions for the simple linear regression model were necessary to enable us to trust the inferences in this question?

(See the solution on page 567.)

2. Below is the *JMP*TM output for 12 random samples of a particular model of used car. The selling price of the particular car (in dollars) and the distance travelled according to its odometer reading (in thousands of kilometres) were recorded.

Variable	Mean	Std Dev	Correlation	Number
Odometer Reading	48	6.96	-0.87	12
Selling Price	6500	1997.52		

- (a) Compute the appropriate least-squares regression equation for this problem.

(See the solution on page 568.)

- (b) Test the hypothesis the correlation coefficient is negative at the 1% level of significance.

(See the solution on page 569.)

- (c) Construct a 98% confidence interval for the slope. Interpret this interval.

(See the solution on page 571.)

3. A linear regression analysis was conducted by the city water department to predict the monthly water usage (in litres) based on the number of residents in a household. Ten randomly selected households had their water usage monitored for a month. A partial *JMP*TM output is shown below.

Bivariate Fit of Water Usage By Residents
--

Variable	Mean	Std Dev	Number
Residents	5.7	3.465705	10
Water Usage	1003.8	417.0438	

Linear Fit

Water Usage = \mathbf{a} + \mathbf{b} Residents

Summary of Fit

RSquare	\mathbf{c}
RSquare Adj	
Root Mean Square Error	\mathbf{d}
Mean of Response	1003.8
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	\mathbf{e}	\mathbf{f}	\mathbf{g}	\mathbf{h}
Error	\mathbf{i}	554666.6	\mathbf{j}	Prob>F
C. Total	\mathbf{k}	1565329.6		0.0051

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	452.65587	166.6488	\mathbf{l}	0.0264
Residents	96.691952	25.32551	\mathbf{m}	0.0051

Answer the questions (a) through (k) below.

3. (a) Determine the values for \mathbf{a} through \mathbf{m} in the outputs above.
(See the solution on page 577.)

- 3. (b)** State the model for the linear regression defining any symbols you use.
(See the solution on page 577.)
- (c)** List the three parameters for the linear regression model and their estimates in this problem. What is your fitted model for the regression?
(See the solution on page 578.)
- (d)** A 95% confidence interval for the intercept of the linear regression is:
(A) (38.29, 155.09) **(B)** (395.37, 509.94) **(C)** (394.26, 511.06)
(D) (68.36, 836.95) **(E)** (39.41, 153.98)
(See the solution on page 578.)
- (e)** A 95% confidence interval for the slope of the linear regression is:
(A) (38.29, 155.09) **(B)** (395.37, 509.94) **(C)** (394.26, 511.06)
(D) (68.36, 836.95) **(E)** (39.41, 153.98)
(See the solution on page 579.)
- (f)** A 95% prediction interval for the water usage of a household with 4 residents is:
(A) (-212.16, 220.16) **(B)** (-640.53, 648.53) **(C)** (394.26, 511.06)
(D) (194.90, 1483.95) **(E)** (623.26, 1055.58)
(See the solution on page 582.)
- (g)** A 95% confidence interval for the mean water usage of a household with 4 residents:
(A) (-212.16, 220.16) **(B)** (-640.53, 648.53) **(C)** (394.26, 511.06)
(D) (194.90, 1483.95) **(E)** (623.26, 1055.58)
(See the solution on page 582.)
- (h)** The test statistic and P -value for testing $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$ are, respectively:
(A) 14.58; 0.0051 **(B)** 3.818; 0.0051 **(C)** 3.818; 0.0026
(D) 2.716; 0.0264 **(E)** both (A) and (B).
(See the solution on page 586.)
- (i)** If we wished to test the hypothesis the correlation coefficient is positive at $\alpha = 5\%$, the critical value and test statistic, respectively, would be:
(A) 1.833; 3.818 **(B)** 1.860; 3.818 **(C)** 2.306; 3.818
(D) 3.818; 1.833 **(E)** 1.860; 2.716
(See the solution on page 587.)
- (j)** The test statistic and P -value for testing $H_0: \beta_0 = 0$ vs. $H_a: \beta_0 \neq 0$ at the 5% significance level are, respectively:
(A) 14.58; 0.0051 **(B)** 3.818; 0.0051 **(C)** 3.818; 0.0026
(D) 2.716; 0.0264 **(E)** both (A) and (B).
(See the solution on page 588.)
- (k)** Test the hypotheses the slope is 100 and the intercept is 500 at the 5% level of significance. i.e. Is it fair to say, for the population, $\hat{y} = 500 + 100x$?
(See the solution on page 591.)

4. A realtor in Winnipeg is trying to set up a regression model to identify the selling price of condominiums (“Price” is measured in thousands of dollars). She is using the explanatory variables “Area” (size of condo in hundred square feet), “Land” (value of the land in thousands of dollars), “Distance” (distance from city centre in kilometres), and “#Beds” (number of bedrooms). See the *JMP*TM output based on data from 20 condo sales on the following page.
- (a) State the model for the multiple linear regression being performed, defining any symbols you use. *(See the solution on page 603.)*
- (b) What is the fitted model that can be used to predict the selling price? Interpret the coefficients. *(See the solution on page 605.)*
- (c) Predict the selling price for a condo which is 1200 square feet, has a land value of \$30,000, is 3 km from the city centre, and has 2 bedrooms. *(See the solution on page 606.)*
- (d) What assumptions are being made? *(See the solution on page 607.)*
- (e) With reference to the “Correlations” matrix in the *JMP* output, is there any cause for concern? *(See the solution on page 608.)*
- (f) Identify the correlations for each explanatory variable with “Price” and interpret them. *(See the solution on page 609.)*
- (g) With reference to the residual plot in the *JMP* output, is there any cause for concern? *(See the solution on page 610.)*
- (h) State the coefficient of determination and interpret it. *(See the solution on page 610.)*
- (i) Explain what the “Analysis of Variance” is telling us. What conclusion can we make at the 5% level of significance? Include the hypotheses, degrees of freedom for F , critical value, test statistic, and P -value in your response. *(See the solution on page 612.)*
- (j) Explain what the t -ratios in the “Parameter Estimates” are telling us. What conclusions can we make about each coefficient at the 5% level of significance? Include the hypotheses, degrees of freedom, critical value, test statistics, and P -values in your response. *(See the solution on page 613.)*
- (k) Give a 95% confidence interval for the “Area” coefficient. *(See the solution on page 614.)*
- (l) Give a 95% confidence interval for the “#Beds” coefficient. *(See the solution on page 614.)*
- (m) Are there any improvements you could suggest for the model? *(See the solution on page 615.)*

Response Price

Whole Model

Summary of Fit

RSquare	0.934126
RSquare Adj	0.91656
Root Mean Square Error	5.348438
Mean of Response	50.1
Observations (or Sum Wgts)	20

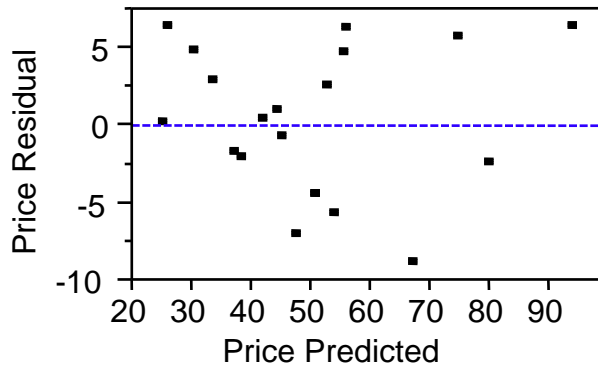
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	6084.7132	1521.18	53.1773
Error	15	429.0868	28.61	Prob > F
C. Total	19	6513.8000		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-10.46558	5.679468	-1.84	0.0852
Area	3.5085371	0.803183	4.37	0.0006
Land	0.7501672	0.56821	1.32	0.2066
Distance	-0.577782	0.219431	-2.63	0.0188
#Beds	8.5056949	2.386334	3.56	0.0028

Residual by Predicted Plot



Correlations

	Price	Area	Land	Distance	#Beds
Price	1.0000	0.9247	0.6325	-0.1696	0.8380
Area	0.9247	1.0000	0.5841	-0.0493	0.7853
Land	0.6325	0.5841	1.0000	-0.2057	0.4244
Distance	-0.1696	-0.0493	-0.2057	1.0000	0.1706
#Beds	0.8380	0.7853	0.4244	0.1706	1.0000

SUMMARY OF KEY CONCEPTS IN LESSON 11

- ❖ We use the **sign test** when we have a small random sample from a continuous population that is not normally distributed. We can also use the sign test for matched pairs data.
 - If $n < 15$, we can only use t to analyze a single sample or matched pairs sample if the population is normal. If the population is not normal, we will use the sign test.
- ❖ As long as our sample is random and our population is continuous, it is valid to use the sign test.
- ❖ The sign test hypothesizes the median, M , of a population.
 - If $H_0: M = M_0$, we will subtract M_0 from each data value.
 - For matched pairs data, we subtract the pairs; $H_0: M = 0$.
- ❖ **Discard any 0 values that come up in your differences!**
- ❖ Let n = the number of signs (discarding 0's), $p = .5$, and k = whatever sign came up less often. Use Table C or the binomial probability formula to compute $P(X \leq k)$. If you are doing a one-tailed, that probability is your P -value. If you are doing a two-tailed test, double this probability to get your P -value.

LECTURE PROBLEMS FOR LESSON 11

For your convenience, here are the 4 questions I used as examples in this lesson. Do not make any marks or notes on these questions below. Especially, do not circle the correct choice in the multiple choice questions. You want to keep these questions untouched, so that you can look back at them without any hints. Instead, make any necessary notes, highlights, etc. in the lecture part above.

- To determine whether an epilepsy drug is useful in treating children with severe learning problems, 10 children with a history of learning and behavioural problems are recruited. In a blind experiment, each child was given a placebo for 3 weeks and the epilepsy drug for the other 3 weeks. The order of the treatments was randomly determined where 5 of the children were given the drug first, the other 5 were given the placebo first. After each 3 week period, all 10 children were given an IQ test. The following results were recorded:

Child	1	2	3	4	5	6	7	8	9	10
IQ after Drug	113	113	101	119	111	122	121	103	110	126
IQ after Placebo	97	106	106	95	102	111	115	104	90	126

- Assuming the population is not normal, indicate what method should be used to analyze this data, and set up the appropriate hypothesis to determine if the drug improves IQ.
 - What is the P -value?
 - What is your conclusion at the .10 level of significance?
(See the solution on page 630.)
- The median salary of male professors at U of M is 65 thousand dollars. A random sample of 9 female professors produces salaries of 47, 58, 92, 58, 65, 73, 60, 57, and 58 thousand dollars.
 - Assuming the population is not symmetric, at the .05 level of significance, is there evidence female professors make a lower median salary?
 - If the population is actually normally distributed, how would that change your approach in part (a)? State the hypotheses and rejection region, but do not actually perform this test.
(See the solution on page 632.)

3. The median age of the rural residents in a country not unlike Canada was 32 in 1900 . A random sample of 12 rural residents from the same country in the year 2000 revealed the ages 66, 72, 45, 3, 10, 48, 52, 28, 37, 32, 93, 50. It is believed the median age increased during the last century. Assuming the population is strongly skewed, the P -value would be:

(A) 0.081 **(B)** 0.054 **(C)** 0.016 **(D)** 0.113 **(E)** 0.033

(See the solution on page 635.)

4. Two laboratories were asked to determine the bacterial count of 15 samples of water. Each sample was divided into two bottles and then one of these bottles was randomly selected to go to Lab A, the other was then sent to Lab B. You are given a table of their readings.

Sample	1	2	3	4	5	6	7	8
Lab A	151	75	129	180	232	166	53	41
Lab B	148	77	118	171	239	155	50	33

Sample	9	10	11	12	13	14	15
Lab A	99	161	185	112	117	60	92
Lab B	89	162	170	108	130	54	71

Let $\alpha = .10$. Does it appear the laboratories differ in their determination of bacterial counts?

- (a)** What is your conclusion if you use the sign test? Be sure to include the hypotheses and P -value in your answer.
- (b)** What other method could be used to study this data? State the hypotheses we would test using that method. Do not actually perform this test.
- (c)** Compare the assumptions we have to make to be able to trust the methods used in parts (a) and (b).

(See the solution on page 636.)

5. Twenty-five randomly selected students who took two Statistics Courses (Stats I and Stats II) had their final grades compared. Four students did better in Stats II, sixteen students did worse in Stats II, and the other five got the same grade in both courses. If we are testing the hypothesis that a student's grade would be the same in both courses, the P -value for the sign test is:

(A) 0.0020 **(B)** 0.0059 **(C)** 0.0040 **(D)** 0.0118 **(E)** cannot be determined

(See the solution on page 638.)

PREPARING FOR THE FINAL EXAM

- ❖ If you have done all of the homework from all 11 lessons, you are now ready to start preparing for your final exam. **Be sure to do all of the Final Exams** from the Smiley Cheng *Multiple-Choice Problems Set for Basic Statistical Analysis I (Stat 2000)* available in the Statistics section of the UM Book Store (but not the final exams obviously). **Note that the course has undergone changes in topics and philosophy over the years, so some questions in the old finals must be omitted. Again, I will send you more tips to help prepare for your exam if you have signed up for Grant's Updates.** (I prefer to wait until the exam is approaching to make sure I know tips are relevant.) I suggest you start with the most recent exams and work your way backwards. The more recent exams are probably more indicative of what your exam will be like. The exams from the 90s are probably too easy, as the midterm has definitely gotten harder over the years.
- ❖ Your main source of review for midterm material should be going over the actual midterm exam you wrote this term. If you are clear on how to answer those questions, I would say that is sufficient review of that half of the course. You will get additional midterm review by looking at the old final exams in the Cheng book as well.
- ❖ **The solutions to the final exams in *Smiley Cheng* are here in Appendix D of my book starting on page D-1.**
 - Please understand many of these solutions were written several years ago, and the course has undergone many changes during that time. Specifically, you may see me referring to different statistical tables (Cheng/Fu Tables 1, 2, etc.). You are expected to use the tables in Moore & McCabe to answer all of the questions. (In the case of Poisson, of course, you have no tables, so you will have to use the Poisson probability formula.)
 - **In all of the exams prior to 2007-2008**, I use the conservative method for the two-sample t test when the pooled method is not called for. This is essentially the generalized method, but where we use the smaller of $df = n_1 - 1$ and $df = n_2 - 1$ for our degrees of freedom rather than that horrible formula we use nowadays. Back then, the course used the conservative method rather than the generalized method. You would have to use that really complicated df formula given as #1 on the formula sheet these days (but, then again, they will probably compute the df for you in that case).

❖ **Omit the following questions from the old final exams:**

- **1997/98 Term 1:** Do not omit any questions.
- **1997/98 Term 2:** Do not omit any questions.
- **1998/99 Term 1:** OMIT question 11 and Part B, question 1.
- **1998/99 Term 2:** Do not omit any questions.
- **1999/00 Term 1:** OMIT questions 7 and 8, and Part B, question 2.
- **1999/00 Term 2:** OMIT questions 5, 6 and 7, and Part B, question 1.
- **2002/03 Term 1:** OMIT questions 6, 7 and 8, and Part B, question 1.
- **2002/03 Term 2:** OMIT questions 3 (but do 4) and 6 to 12 (all).
- **2003/04 Term 1:** OMIT questions 7, 8 and 22.
- **2003/04 Term 2:** OMIT questions 5, 6 (but do 7), 9, 10, and 12.
- **2005/06 Term 1:** OMIT questions 10, 13, 14, and 15.
- **2005/06 Term 2:** OMIT questions 6, 7, 8, and 9.
- **2006/07 Term 1:** OMIT questions 7, 11, 12, 13, and 14
- **2006/07 Term 2:** OMIT the entire exam
(it is the 05/06 Term 2 exam again by mistake)
- **2007/08 to 2010/11:** Do not omit any questions.

❖ **If your exam has a long answer section, be sure you do the long answer part first.** Time is sometimes an issue on the exam. If you are running out of time, you would rather be rushed as you are finishing off some multiple-choice questions (where you could always guess and hope) than feel rushed while trying to complete a more valuable long answer question. **A prepared student should have no fear of the long answer questions while there will undoubtedly be multiple-choice questions that will confuse any student.**

❖ **Never doubt yourself when answering a multiple-choice question.** If your answer is not one of the choices, simply select the closest choice and move on. Never waste your time redoing a question! If you have done it wrong, you are likely to still do it wrong the second time. You have other questions to do. Getting obsessed with one question, may mean not having time to answer two or three or more at the end. They are all worth the same marks, so leaving two or three blank at the end in order to vainly attempt to get one question right is just silly. If you have completed the exam, and still have time, by all means go back and try questions you had doubts about. Since you are now looking at the question fresh and with some distance, you have a much better chance of correcting your mistake (if you made one).

❖ **If the question is strictly theory, no math at all, you should never spend more than two minutes to make up your mind what choice to make.** Believe me. If you don't know the answer within one minute, they got you anyway, so just trust your gut, make a choice, and move on. That will buy you time to spend on the slower calculation questions.

APPENDIX A

HOW TO USE STAT MODES ON YOUR CALCULATOR

In the following pages, I show you how to enter data into your calculator in order to compute the mean and standard deviation. I also show you how to enter x, y data pairs in order to get the correlation, intercept and slope of the least squares regression line.

Please make sure that you are looking at the correct page when learning the steps. I give steps for several brands and models of calculator.

I consider it absolutely vital that a student know how to use the Stat modes on their calculator. It can considerably speed up certain questions and, even if a question insists you show all your work, gives you a quick way to check your answer.

If you cannot find steps for your calculator in this appendix, or cannot get the steps to work for you, do not hesitate to contact me. I am very happy to assist you in calculator usage (or anything else for that matter).

SHARP CALCULATORS

(Note that the EL-510 does not do Linear Regression.)

You will be using a "MODE" button. Look at your calculator. If you have "MODE" actually written on a button, press that when I tell you to press "**MODE**". If you find mode written above a button (some models have mode written above the "DRG" button, like this: "**MODE** **DRG**") then you will have to use the "**2ndF**" button to access the mode button; i.e. when I say "**MODE**" below, you will actually press "**2ndF** **MODE** **DRG**".

BASIC DATA PROBLEM

Feed in data to get the mean, \bar{x} , and standard deviation, s (which Sharps tend to denote "sx").

Step 1: Put yourself into the "STAT, SD" mode.

Press **MODE** **1** **0** (Screen shows "Stat0")

Step 2: Enter the data: 3, 5, 9.

To enter each value, press the "M+" button. There are some newer models of Sharp that have you press the "CHANGE" button instead of the "M+" button. (The "CHANGE" button is found close by the "M+" button.)

3 **M+**
DATA 5 **M+**
DATA 9 **M+**
DATA

You should see the screen counting the data as it is entered (Data Set=1, Data Set=2, Data Set=3).

Step 3: Ask for the mean and standard deviation.

RCL \bar{x}
4

We see that $\bar{x} = 5.6666\dots = 5.6667$.

RCL sx
5

We see that $s = 3.05505\dots = 3.0551$

Step 4: Return to "NORMAL" mode. This clears out your data as well as returning your calculator to normal.

MODE **0**

LINEAR REGRESSION PROBLEM

Feed in x and y data to get the correlation coefficient, r , the intercept, a , and the slope, b .

Step 1: Put yourself into the "STAT, LINE" mode.

Press **MODE** **1** **1** (Screen shows "Stat1")

Step 2: Enter the data:

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the x and y must be entered as a pair). The pattern is first x , press "STO" to get the comma, first y , then press "M+" (or "CHANGE") to enter the pair; repeat for each data pair.

3 **STO** 7 **M+**
(x,y) DATA

5 **STO** 10 **M+**
(x,y) DATA

9 **STO** 14 **M+**
(x,y) DATA

You should see the screen counting the data as it is entered (Data Set=1, Data Set=2, Data Set=3).

Step 3: Ask for the correlation coefficient, intercept, and slope. (The symbols may appear above different buttons than I indicate below.)

RCL r
÷

We see that $r = 0.99419\dots = 0.9942$.

RCL a
(

We see that $a = 3.85714\dots = 3.8571$.

RCL b
)

We see that $b = 1.14285\dots = 1.1429$.

Step 4: Return to "NORMAL" mode. This clears out your data as well as returning your calculator to normal.

MODE **0**

CASIO CALCULATORS

(Note that some Casios do not do Linear Regression.)

BASIC DATA PROBLEM

Feed in data to get the mean, \bar{x} , and standard deviation, s (which Casios tend to denote " $x\sigma_{n-1}$ " or simply " σ_{n-1} ").

Step 1: Put yourself into the "SD" mode.

Press "**MODE**" once or twice until you see "SD" on the screen menu and then select the number indicated. A little "SD" should then appear on your screen.

Step 2: Clear out old data.

SHIFT $\overset{\text{ScI}}{\text{AC}}$ **=** (Some models will have "ScI" above another button. Be sure you are pressing "ScI", the "Stats Clear" button. (Some models call it "SAC" for "Stats All Clear" instead of ScI.)

Step 3: Enter the data: 3, 5, 9.

To enter each value, press the "M+" button.

3 $\overset{\text{DT}}{\text{M+}}$ 5 $\overset{\text{DT}}{\text{M+}}$ 9 $\overset{\text{DT}}{\text{M+}}$ (You use the "M+" button to enter each piece of data.)

Step 4: Ask for the mean and standard deviation.

SHIFT $\overset{\bar{x}}{1}$ **=**

We see that $\bar{x} = 5.6666\dots = 5.6667$.

SHIFT $\overset{x\sigma_{n-1}}{3}$ **=**

We see that $s = 3.05505\dots = 3.0551$

(Some models may have \bar{x} and $x\sigma_{n-1}$ above other buttons rather than "1" and "3" as I illustrate above.)

If you can't find these buttons on your calculator, look for a button called "S. VAR" (which stands for "Statistical Variables", it is probably above one of the number buttons).

Press: **SHIFT** **S. VAR** and you will be given a menu showing the mean and standard deviation. Select the appropriate number on the menu and press "=" (You may need to use your arrow buttons to locate the \bar{x} or $x\sigma_{n-1}$ options.)

Step 5: Return to "COMP" mode.

Press **MODE** and select the "COMP" option.

LINEAR REGRESSION PROBLEM

Feed in x and y data to get the correlation coefficient, r , the intercept, a , and the slope, b .

Step 1: Put yourself into the "REG, Lin" mode.

Press "**MODE**" once or twice until you see "Reg" on the screen menu and then select the number indicated. You will then be sent to another menu where you will select "Lin". (Some models call it the "LR" mode in which case you simply choose that instead.)

Step 2: Clear out old data.

Do the same as Step 2 for "Basic Data".

Step 3: Enter the data.

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the x and y must be entered as a pair). The pattern is first x , first y ; second x , second y ; and so on. Here is the data we want to enter:

3 **,** 7 $\overset{\text{DT}}{\text{M+}}$ 5 **,** 10 $\overset{\text{DT}}{\text{M+}}$ 9 **,** 14 $\overset{\text{DT}}{\text{M+}}$

(If you can't find the comma button "**,**", you probably use the open bracket button instead to get the comma "**[(-)**". You might notice " $[x_D, y_D]$ " in blue below this button, confirming that is your comma.)

Step 4: Ask for the correlation coefficient, intercept, and slope. (The symbols may appear above different buttons than I indicate below.)

SHIFT $\overset{r}{(}$ **=**

We see that $r = 0.99419\dots = 0.9942$.

SHIFT $\overset{A}{7}$ **=**

We see that $a = 3.85714\dots = 3.8571$.

SHIFT $\overset{B}{8}$ **=**

We see that $b = 1.14285\dots = 1.1429$.

If you can't find these buttons on your calculator, look for a button called "S. VAR"

Press: **SHIFT** **S. VAR** and you will be given a menu showing the mean and standard deviation. Use your left and right arrow buttons to see other options, like " r ". Select the appropriate number on the menu and press "=".

Step 5: Return to "COMP" mode.

Press **MODE** and select the "COMP" option.

HEWLETT PACKARD HP 10B II

BASIC DATA PROBLEM

Feed in data to get the mean, \bar{x} , and standard deviation, s (which it denotes "Sx").

Step 1: Enter the data: 3, 5, 9.

To enter each value, press the " $\Sigma+$ " button.

$\boxed{3} \boxed{\Sigma+} \boxed{5} \boxed{\Sigma+} \boxed{9} \boxed{\Sigma+}$ (As you use the " $\Sigma+$ " button to enter each piece of data, you will see the calculator count it going in: 1, 2, 3.)

Step 2: Ask for the mean and standard deviation.

Note that by "orange" I mean press the button that has the orange bar coloured on it. The orange bar is used to get anything coloured orange on the buttons.

$\boxed{\text{orange}} \boxed{7}$
 \bar{x}, \bar{y}

We see that $\bar{x} = 5.6666\dots = 5.6667$.

$\boxed{\text{orange}} \boxed{8}$
 s_x, s_y

We see that $s = 3.05505\dots = 3.0551$

Step 3: "Clear All" data ready for next time.

$\boxed{\text{orange}} \boxed{\text{C}}$
 C ALL

LINEAR REGRESSION PROBLEM

Feed in x and y data to get the correlation coefficient, r , the intercept, a , and the slope, b .

Step 1: Enter the data:

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the x and y must be entered as a pair). The pattern is first x , first y ; second x , second y ; and so on.

$\boxed{3} \boxed{\text{INPUT}} \boxed{7} \boxed{\Sigma+}$

$\boxed{5} \boxed{\text{INPUT}} \boxed{10} \boxed{\Sigma+}$

$\boxed{9} \boxed{\text{INPUT}} \boxed{14} \boxed{\Sigma+}$

(As you use the " $\Sigma+$ " button to enter each pair of data, you will see the calculator count it going in: 1, 2, 3.)

Step 2: Ask for the correlation coefficient, intercept, and slope.

$\boxed{\text{orange}} \boxed{4} \boxed{\text{orange}} \boxed{\text{K}}$
 \hat{x}, r SWAP

We see that $r = 0.99419\dots = 0.9942$.

Note that the "SWAP" button is used to get anything that is listed second (after the comma) like " r " in this case.

The intercept has to be found by finding \hat{y} when $x=0$:

$0 \boxed{\text{orange}} \boxed{5}$
 \hat{y}, m

We see that $a = 3.85714\dots = 3.8571$.

The slope is denoted " m " on this calculator:

$\boxed{\text{orange}} \boxed{5} \boxed{\text{orange}} \boxed{\text{K}}$
 \hat{y}, m SWAP

We see that $b = 1.14285\dots = 1.1429$.

Step 3: "Clear All" data ready for next time.

$\boxed{\text{orange}} \boxed{\text{C}}$
 C ALL

TEXAS INSTRUMENTS TI-30X-II

(Note that the TI-30Xa does not do Linear Regression.)

BASIC DATA PROBLEM

Feed in data to get the mean, \bar{x} , and standard deviation, s (which it denotes "Sx").

Step 1: Clear old data.

$\boxed{2nd}$ $\boxed{\overset{STAT}{DATA}}$ Use your arrow keys to ensure "CLRDATA" is underlined then press $\boxed{\overset{ENTER}{=}}$

Step 2: Put yourself into the "STAT 1-Var" mode.

$\boxed{2nd}$ $\boxed{\overset{STAT}{DATA}}$ Use your arrow keys to ensure "1-Var" is underlined then press $\boxed{\overset{ENTER}{=}}$

Step 3: Enter the data: 3, 5, 9.

(You will enter the first piece of data as "X1", then use the down arrows to enter the second piece of data as "X2", and so on.)

\boxed{DATA} 3 $\boxed{\overset{ENTER}{=}}$ (X1 = 3)

$\boxed{\downarrow}$ $\boxed{\downarrow}$ 5 $\boxed{\overset{ENTER}{=}}$ (X2 = 5)

$\boxed{\downarrow}$ $\boxed{\downarrow}$ 9 $\boxed{\overset{ENTER}{=}}$ (X3 = 9)

Step 4: Ask for the mean and standard deviation.

Press $\boxed{STATVAR}$ then you can see a list of outputs by merely pressing your left and right arrows to underline the various values.

We see that $\bar{x} = 5.6666\dots = 5.6667$.

We see that $s = 3.05505\dots = 3.0551$

Step 5: Return to standard mode.

\boxed{CLEAR} This resets your calculator ready for new data next time.

LINEAR REGRESSION PROBLEM

Feed in x and y data to get the correlation coefficient, r , the intercept, a , and the slope, b .

Step 1: Clear old data (as in BASIC DATA PROBLEM at left).

Step 2: Put yourself into the "STAT 2-Var" mode.

$\boxed{2nd}$ $\boxed{\overset{STAT}{DATA}}$ Use your arrow keys to ensure "2-Var" is underlined then press $\boxed{\overset{ENTER}{=}}$

Step 3: Enter the data:

x	3	5	9
y	7	10	14

(You will enter the first x -value as "X1", then use the down arrow to enter the first y -value as "Y1", and so on.)

\boxed{DATA} 3 $\boxed{\overset{ENTER}{=}}$ $\boxed{\downarrow}$ 7 $\boxed{\overset{ENTER}{=}}$ (X1 = 3, Y1 = 7)

$\boxed{\downarrow}$ 5 $\boxed{\overset{ENTER}{=}}$ $\boxed{\downarrow}$ 10 $\boxed{\overset{ENTER}{=}}$ (X2 = 5, Y2 = 10)

$\boxed{\downarrow}$ 9 $\boxed{\overset{ENTER}{=}}$ $\boxed{\downarrow}$ 14 $\boxed{\overset{ENTER}{=}}$ (X3 = 9, Y3 = 14)

Step 4: Ask for the correlation coefficient, intercept, and slope.

Press $\boxed{STATVAR}$ then you can see a list of outputs by merely pressing your left and right arrows to underline the various values. **Note: Your calculator may have a and b reversed. To get a , you ask for b ; to get b you ask for a .** Don't ask me why that is, but if that is the case then realize it will always be the case.

We see that $r = 0.99419\dots = 0.9942$.

We see that $a = 3.85714\dots = 3.8571$.

We see that $b = 1.14285\dots = 1.1429$.

Step 5: Return to standard mode (as in BASIC DATA PROBLEM at left).

TEXAS INSTRUMENTS TI-36X

(Note that the TI-30Xa does not do Linear Regression.)

BASIC DATA PROBLEM

Feed in data to get the mean, \bar{x} , and standard deviation, s (which it denotes " σx_{n-1} ").

Step 1: Put yourself into the "STAT 1" mode.

$\boxed{3\text{rd}} \boxed{x \rightleftharpoons y}$ ^{STAT 1}

Step 2: Enter the data: 3, 5, 9.

To enter each value, press the " $\Sigma+$ " button.

$3 \boxed{\Sigma+} 5 \boxed{\Sigma+} 9 \boxed{\Sigma+}$ (As you use the " $\Sigma+$ " button to enter each piece of data, you will see the calculator count it going in: 1, 2, 3.)

Step 3: Ask for the mean and standard deviation.

$\boxed{2\text{nd}} \boxed{\bar{x}}$

We see that $\bar{x} = 5.6666\dots = 5.6667$.

$\boxed{2\text{nd}} \boxed{\sigma x_{n-1}}$

We see that $s = 3.05505\dots = 3.0551$

Step 4: Return to standard mode.

$\boxed{\text{ON}/\text{AC}}$ (Be careful! If you ever press this

button during your work you will end up resetting your calculator and losing all of your data. Use the $\boxed{\text{CE}/\text{C}}$ button to clear mistakes without resetting your calculator. I usually press this button a couple of times to make sure it has cleared any mistake completely.)

LINEAR REGRESSION PROBLEM

Feed in x and y data to get the correlation coefficient, r , the intercept, a , and the slope, b .

Step 1: Put yourself into the "STAT 2" mode.

$\boxed{3\text{rd}} \boxed{\Sigma+}$ ^{STAT 2}

Step 2: Enter the data:

x	3	5	9
y	7	10	14

Note you are entering in pairs of data (the x and y must be entered as a pair). The pattern is first x , first y ; second x , second y ; and so on.

$3 \boxed{x \rightleftharpoons y} 7 \boxed{\Sigma+}$

$5 \boxed{x \rightleftharpoons y} 10 \boxed{\Sigma+}$

$9 \boxed{x \rightleftharpoons y} 14 \boxed{\Sigma+}$

(As you use the " $\Sigma+$ " button to enter each pair of data, you will see the calculator count it going in: 1, 2, 3.)

Step 3: Ask for the correlation coefficient, intercept, and slope.

Note that this calculator uses the abbreviations "COR" for correlation, "ITC" for intercept and "SLP" for slope.

$\boxed{3\text{rd}} \boxed{\text{COR}}$

We see that $r = 0.99419\dots = 0.9942$.

$\boxed{2\text{nd}} \boxed{\text{ITC}}$

We see that $a = 3.85714\dots = 3.8571$.

$\boxed{2\text{nd}} \boxed{\text{SLP}}$

We see that $b = 1.14285\dots = 1.1429$.

Step 4: Return to standard mode.

$\boxed{\text{ON}/\text{AC}}$

TEXAS INSTRUMENTS TI-BA II Plus

Put yourself into the "LIN" mode.

press $\boxed{2\text{nd}} \boxed{8}$ ^{STAT} If "LIN" appears, great; if not, press $\boxed{2\text{nd}} \boxed{\text{ENTER}}$ ^{SET} repeatedly until "LIN" does show up. Then press $\boxed{2\text{nd}} \boxed{\text{CPT}}$ ^{QUIT} to "quit" this screen.

Note: Once you have set the calculator up in "LIN" mode, it will stay in that mode forever. You can now do either "Basic Data" or "Linear Regression" problems.

BASIC DATA PROBLEM

Feed in data to get the mean, \bar{x} , and standard deviation, s (which it denotes "Sx").

Step 1: Clear old data.

$\boxed{2\text{nd}} \boxed{7}$ ^{DATA} $\boxed{2\text{nd}} \boxed{\text{CE/C}}$ ^{CLR Work}

Step 2: Enter the data: 3, 5, 9.

(You will enter the first piece of data as "X1", then use the down arrows to enter the second piece of data as "X2", and so on. Ignore the "Y1", "Y2", etc.)

$\boxed{\text{DATA}} \boxed{3} \boxed{\text{ENTER}} \boxed{=}$ (X1 = 3)

$\boxed{\downarrow} \boxed{\downarrow} \boxed{5} \boxed{\text{ENTER}} \boxed{=}$ (X2 = 5)

$\boxed{\downarrow} \boxed{\downarrow} \boxed{9} \boxed{\text{ENTER}} \boxed{=}$ (X3 = 9)

Step 3: Ask for the mean and standard deviation.

Press $\boxed{2\text{nd}} \boxed{8}$ ^{STAT} then you can see a list of outputs by merely pressing your up and down arrows to reveal the various values.

We see that $\bar{x} = 5.6666\dots = 5.6667$.

We see that $s = 3.05505\dots = 3.0551$

Step 4: Return to standard mode.

$\boxed{\text{ON/OFF}}$ This resets your calculator ready for new data next time.

LINEAR REGRESSION PROBLEM

Feed in x and y data to get the correlation coefficient, r , the intercept, a , and the slope, b .

Step 1: Clear old data.

$\boxed{2\text{nd}} \boxed{7}$ ^{DATA} $\boxed{2\text{nd}} \boxed{\text{CE/C}}$ ^{CLR Work}

Step 2: Enter the data:

x	3	5	9
y	7	10	14

(You will enter the first x -value as "X1", then use the down arrow to enter the first y -value as "Y1", and so on.)

$\boxed{\text{DATA}} \boxed{3} \boxed{\text{ENTER}} \boxed{=}$ $\boxed{\downarrow} \boxed{7} \boxed{\text{ENTER}} \boxed{=}$ (X1 = 3, Y1 = 7)

$\boxed{\downarrow} \boxed{5} \boxed{\text{ENTER}} \boxed{=}$ $\boxed{\downarrow} \boxed{10} \boxed{\text{ENTER}} \boxed{=}$ (X2 = 5, Y2 = 10)

$\boxed{\downarrow} \boxed{9} \boxed{\text{ENTER}} \boxed{=}$ $\boxed{\downarrow} \boxed{14} \boxed{\text{ENTER}} \boxed{=}$ (X3 = 9, Y3 = 14)

Step 3: Ask for the correlation coefficient, intercept, and slope.

Press $\boxed{2\text{nd}} \boxed{8}$ ^{STAT} then you can see a list of outputs by merely pressing your up and down arrows to reveal the various values. We see that $r = 0.99419\dots = 0.9942$.

We see that $a = 3.85714\dots = 3.8571$.

We see that $b = 1.14285\dots = 1.1429$.

Step 4: Return to standard mode.

$\boxed{\text{ON/OFF}}$ This resets your calculator ready for new data next time.