Appendix B
Practice Exam

STAT 2000 Practice Exam
Distance and Online Education

This practice examination is similar in format to the final examination. That is, the total number of marks possible is 60, and the examination is divided into two parts as follows:

a) **PART A** is worth 35 marks and consists of 30 multiple choice questions. For each question, **CIRCLE THE LETTER** corresponding to the **BEST** answer out of the five possibilities. Only one letter should be circled; otherwise, the question will be marked wrong. The questions are of equal value. There is no correction made for guessing; therefore, all questions should be attempted.

b) **PART B** is worth 25 marks and consists of 5 long answer questions which are to be answered in the spaces provided on the examination paper.

The final examination is a three-hour, closed book, examination. A formulae page will be provided; it will be the same as the one that is attached to this practice examination. A copy of *Tables to accompany IPS* will be provided. You will also be permitted to use a calculator. We suggest that you write this practice examination under similar conditions.

Part A (35 marks)

1. A random sample of 15 employees at a large manufacturing company gave a 90% confidence interval of $8 \leq \mu \leq 12$ for the average number of years employees had worked for the company. This means:

   a. Nine out of ten employees worked for the company between 8 and 12 years.
   b. The true mean number of years employees worked for the company is between 8 and 12 years.
   c. If many samples of size 15 were taken and a similar 90% confidence interval were obtained for each sample, 90% of such intervals would contain the true value of μ.
   d. If many samples of size 15 were taken, for 90% of such intervals, $\bar{x}$ would fall between 8 and 12.
   e. 90% of samples would yield mean number of years worked for the company of between 8 and 12 years.

2. A sleep-producing drug was administered to 5 patients who recorded the following increases in sleep times (in minutes):

   26 32 18 12 17

   The margin of error for a 95% confidence interval is:

   a. 9.13 b. 9.85 c. 6.96 d. 4.41

   e. We need to know σ to calculate the margin of error.

3. The effect on a test of increasing n and keeping β and σ fixed is:

   a. increase the level of significance
   b. decrease the level of significance
   c. decrease the power
   d. increase the power
   e. decrease the probability of a Type II error

Questions 4 and 5 refer to the following:

In a survey of gas stations across Canada done on July 16, 2002, the price (in cents/liter) of regular gasoline was recorded from 10 randomly selected locations in western Canada, and 13 in Eastern Canada. The following results were obtained:

Western Canada	68.7	68.4	64.9	69.0	75.9	75.9	71.3	78.3	70.2	89.9			
Eastern Canada	67.4	73.9	69.9	75.7	66.4	69.5	67.4	72.8	77.8	69.0	73.0	70.8	68.0

From the following *JMP* output, it desired to study the difference in gas price between western and eastern Canada.

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Eastern	13	70.8923	3.50487	0.9721	68.774	73.010
Western	10	73.2500	7.16384	2.2654	68.125	78.375

t Test

Western-Eastern
Assuming equal variances

Difference	2.3577	t Ratio	1.040617
Std Err Dif	2.2657	DF	21
Upper CL Dif	7.0694	Prob > |t|	0.3099
Lower CL Dif	-2.3540	Prob > t	0.1549
Confidence	0.95	Prob < t	0.8451

t Test

Western-Eastern
Assuming unequal variances

Difference	2.3577	t Ratio	0.956408
Std Err Dif	2.4652	DF	12.30643
Upper CL Dif	7.7140	Prob > |t|	0.3573
Lower CL Dif	-2.9986	Prob > t	0.1786
Confidence	0.95	Prob < t	0.8214

4. A 90% confidence interval on the difference (western – eastern) in mean price between western and eastern Canada is:

a. (-3.00, 7.71) b. (-0.03, 4.87) c. (-2.04, 6.75)

d. (-1.54, 6.26) e. (-0.99, 5.70)

5. It is hypothesized that the mean price for regular gasoline was higher in Western Canada at the time. When testing this claim at the 10% level of significance, the P-value is:

a 0.1549 b. 0.1786 c. 0.2099 d. 0.3099 e. 0.3573

Questions 6 to 9 refer to the following

A study compared the effect of various training frequencies on the development of strength. Participants were randomly selected from each of four groups with a measurement of strength as follows:

Bi-weekly (A)	Weekly (B)	Twice-weekly (C)	Daily (D)
27	39	37	24
50	36	28	53
43	47	44	51
31	51	36	51
37		30	45
37		27	65
		44	

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Frequency				
Error			89.786	
Total		2343.6522		

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
Bi-weekly	6	37.5000	8.2401	3.3640
Daily	6	48.1667	13.5413	5.5282
Twice	7	35.1429	7.1281	2.6942
Weekly	4	43.2500	6.9462	3.4731

6. The hypotheses to be tested are:

a. $H_o: \mu_A = \mu_B = \mu_C = \mu_D$ versus $H_a: \mu_A \neq \mu_B \neq \mu_C \neq \mu_D$

b. $H_o: \mu_A = \mu_B = \mu_C = \mu_D$ versus H_a: Not all means are equal

c. $H_o: \bar{x}_A = \bar{x}_B = \bar{x}_C = \bar{x}_D$ versus $H_a: \bar{x}_A \neq \bar{x}_B \neq \bar{x}_C \neq \bar{x}_D$

d. H_o: $\bar{x}_A = \bar{x}_B = \bar{x}_C = \bar{x}_D$ versus H_a: Not all means are equal

e. H_o: $\mu_A = \mu_B = \mu_C = \mu_D$ versus H_a: The means are all different

7. The value for the test statistic is:

   a. 2.37 b. 26.10 c. 1.18 d. 7.10 e. 0.728

8. Using pooled estimate of the error variance, the standard error of the estimate of $\mu_D - \mu_B$ is:

   a. 6.53 b. 37.41 c. 7.44 d. 6.12 e. 55.29

9. Which of the following assumptions is NOT required for the appropriate test statistic?

   (Note: this question refers to the problem on the previous page.)

   a. The populations must have a common variance.
   b. The samples must be simple random samples.
   c. The samples must all have a normal distribution.
   d. The samples must be independent.
   e. The samples must come from populations which have normal distributions.

10. A random variable Y has the following probability distribution.

k	2	3	5	6
P(Y = k)	0.4	0.3	0.2	0.1

The expected value of Y is:

a. 4 b. 3.3 c. 3 d. 2 e. 0.25

11. Nancy claims that 60% of the email she receives is spam. If this is correct and she selects 5 of her email letters at random, what is the probability that more than 3 are spam?

    a. 0.0518 b. 0.2592 c. 0.0778 d. 0.3370 e. 0.1296

Questions 12 and 13 refer to the following

12. A Weight Clinic claims that everyone will lose weight after two weeks in it's program. The following table lists the weights, in kilograms, of randomly selected clients before and two weeks after beginning the program. We want to test the clinic's claim.

Client	1	2	3	4	5	6	7	8	9	10	11
Before	99	57	70	85	64	74	68	60	78	58	63
After	94	57	71	85	61	69	71	60	71	57	59

The P-value of the Sign test to test the claim is:

a. 0.0352 b. 0.1094 c. 0.1445 d. 0.2256 e. 0.2744

13. The data was also analyzed using a parametric test with a calculated value of 2.06. The P-value of the parametric test statistic is:

    a. 0.0197
    b. 0.0394
    c. 0.010 < P-value < 0.025
    d. 0.025 < P-value < 0.05
    e. 0.050 < P-value < 0.10

14. The number of parking tickets issued by a U of M campus policeman has a Poisson distribution with an average of 2.3 tickets per hour. What is the probability he issues at least 7 tickets in a 3 hour shift?

    a. 0.6136 b. 0.4647 c. 0.5353 d. 0.1489 e. > 0.80

Questions 15 and 16 refer to the following:
The amount of kitty litter that can be poured into a small container varies with a mean of 8 ounces and a standard deviation of 1 ounce. The amount that can be poured into a large container varies with a mean of 12 ounces and a standard deviation of 2 ounces. Let the random variable X represent the difference between the amounts that can be poured into the large container minus the amount that can be poured into a small container.

15. The mean of the random variable X is:

    a. 4 b. 8 c. 12 d. 16 e. 20

16. The standard deviation of the random variable X is:

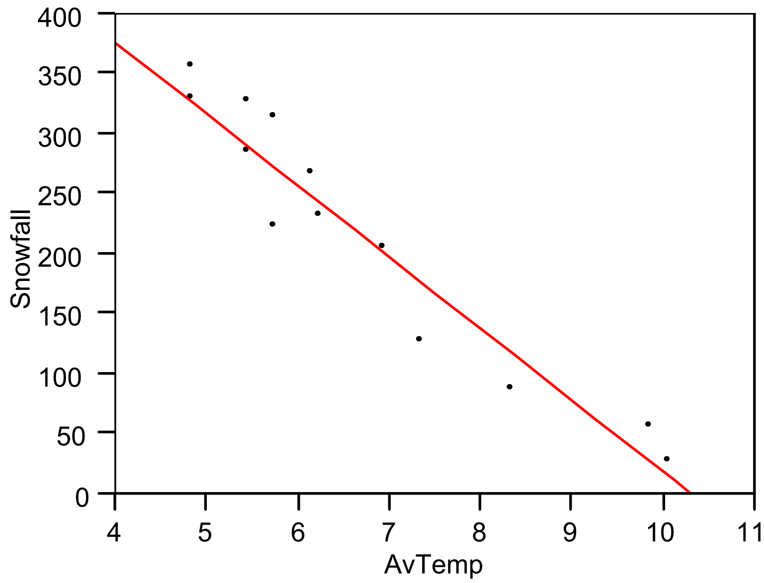
    a. 1 b. 2.24 c. 2.41 d. 3 e. 5

Questions 17 to 21 refer to the following:
The following data provide the Average temperature and Snowfall for large non-prairie Canadian cities for one year. Data given is **AvTemp** (^oC) and **Snowfall** (cm).

AvTemp (x):	4.8	5.4	6.1	6.9	5.4	6.2	5.7	7.3	9.8	10	8.3
Snowfall (y):	359	331	271	208	290	235	227	131	60	32	92

$\bar{x} = 6.9$ $\bar{y} = 203.27$ $\sum(x-\bar{x})^2 = 31.62$

Bivariate Fit of Snowfall By AvTemp



Linear Fit

Snowfall = 614.5 - 59.6 AvTemp

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model				
Error			1044	
C. Total		121,732		

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		614.5	40.82545		
AvTemp		-59.6			

17. The value of the t statistic used to test the hypothesis that the slope parameter is zero, i.e., $H_o:\beta = 0$ is:

    a. –1.84 b. -10.37 c. 15.05 d. 107.61 e. –3.46

18. For this data, the coefficient of determination is:

    a. 0.923 b. 0.077 c. 0.9606 d. 614.54 e. 0.278

19. A 90% prediction interval for the Snowfall of a city with a mean annual temperature of 5^oC is closest to:

a. (267.4, 365.6)
b. (252.2, 380.8)
c. (288.8, 344.2)
d. (265.8, 367.2)
e. (251.5, 381.5)

20. A 90% confidence interval for the mean Snowfall of all cities with a mean annual temperature of 5^oC is closest to:

a. (288.8, 344.2)
b. (289.7, 343.3)
c. (290.0, 343.0)
d. (296.3, 336.7)
e. (298.6, 334.4)

21. For this data we can conclude that:

a. an increase in Average annual temperature causes a decrease in Snowfall.
b. for every increase of 1^oC in Average annual temperature, Snowfall will decrease by 59.6 cm.
c. there is no linear relationship between Average annual temperature and Snowfall.
d. for every increase of 1^oC in Average annual temperature, Snowfall will decrease by an average 59.6 cm.
e. the true slope, β, of the least squares line is –59.6.

22. Consider a test of the hypothesis $H_0:\mu = 68$ vs. $H_1:\mu < 68$ for a population with $\sigma^2 = 108$, a sample of size 12, and a critical region of $\overline{Y} \leq 62.9$. The power of the test against the alternative $\mu = 59$ is:

a. 0.9032 b. 0.9554 c. 0.4032 d. 0.0968 e. 0.0446

Questions 23 and 24 refer to the following:
The following data give the price ($) of a certain CD and the corresponding weekly sales for a random sample of stores.

Store								
Price (x)	11	12	13	15	20	16	15	18
Sales (y)	26	30	31	22	16	28	23	24

Given: $\Sigma x = 120$ $\Sigma x^2 = 1864$ $\Sigma y = 200$ $\Sigma y^2 = 5166$ $\Sigma xy = 2924$
$r = -0.737$ Model SS = 90.25

23. The value of the test statistic used to test $H_o:\rho = 0$ is:

a. -0.737 b. -1.09 c. –1.37 d. -2.67 e. –3.08

24. The least squares regression equation to predict the number of CDs sold (Sales) based on the Price of the CD is:

a. $\hat{y} = 42.81 - 1.1875x$

b. $\hat{y} = 7.19 - 1.875x$

c. $\hat{y} = -10.6 - 2.04x$

d. $\hat{y} = 53.7 - 1.912x$

e. $\hat{y} = 129.40 - 2.04x$

Questions 25 to 28 refer to the following:
A random sample of donors at a U of M Blood donor clinic is selected and each donor classified according to Blood Type and the Year the student is enrolled in. JMP*IN* output is given below.

Note: (1) Expected values for **some cells** are given in **boldface** in the second row.

(2) Cell χ^2 values are given in the third row of some cells.

(3) The sum of the given χ^2 values in the table is 9.612.

Contingency Analysis of Blood Type By Year

Year By Blood Type

Count Expected Cell Chi^2	A	AB	B	O	Total
I	16 0.8000	24 1.2000	33 **35.2** 0.1375	7 **12.8**	80
II	10 0.1286	15 0.1929	32 **30.8** 0.0468	13 **11.2**	70
III	8 **9.6** 0.2667	16 **14.4** 0.1778	22 **26.4** 0.7333	14 **9.6** 2.0167	60
IV	6 **6.4** 0.0250	5 **9.6** 2.2042	23 **17.6** 1.6568	6 **6.4** 0.0250	40
Total	40	60	110	40	250

25. The null hypothesis we usually test for data such as this is:

a. The number of students with four Blood Types is the same in each Year.
b. The four Blood Types are equally likely in each Year.
c. There is no relationship between a student's Blood Type and his/her Year enrolled.
d. The number of students with four Blood Types varies from Year to Year.
e. The ratio of the four Blood Types varies from Year to Year.

26. If the null hypothesis is true, the expected cell count for the (1,2) cell (number of students in Year I with Blood Type AB) is:

a. 19.2 b. 15 c. 20 d. 15.63 e. 24

27. The degrees of freedom and 5% critical value for the χ^2 test is:

a. 9 & 19.02 b. 9 & 16.92 c. 15 & 25 d. 1 & 3.84 e. 16 & 26.3

28. The value of the chi-square statistic is:

a. 9.612 b. 14.67 c. 10.60 d. 10.23 e. 12.53

Questions 29 and 30 refer to the following situation:
A public opinion poll indicates that 136 out of 300 voters in rural Manitoba favour the Progressive Conservative party in the June election compared to 35% of the 380 voters in the city of Winnipeg. Let p_R represent the population proportion of rural voters who favour the Progressive Conservatives and p_C represent the proportion of Winnipeg voters who favour the Progressive Conservatives.

29. A 92% confidence interval for the true difference, $p_R - p_C$, is:

a. $$\frac{136}{320}-\frac{133}{380}\pm 1.96\sqrt{\frac{136}{320}\cdot\frac{184}{320}\cdot\frac{1}{320}+\frac{133}{380}\cdot\frac{247}{380}\cdot\frac{1}{380}}$$

b. $$\frac{136}{320}-\frac{133}{380}\pm 1.75\sqrt{\frac{136}{320}\cdot\frac{184}{320}\cdot\frac{1}{320}+\frac{133}{380}\cdot\frac{247}{380}\cdot\frac{1}{380}}$$

c. $$\frac{136}{320}-\frac{133}{380}\pm 1.96\sqrt{\frac{269}{700}\cdot\frac{431}{700}\cdot\left(\frac{1}{320}+\frac{1}{380}\right)}$$

d. $$\frac{136}{320}-\frac{133}{380}\pm 1.75\sqrt{\frac{269}{700}\cdot\frac{431}{700}\cdot\left(\frac{1}{320}+\frac{1}{380}\right)}$$

e. $$\frac{136}{320}-\frac{133}{380}\pm 1.75\sqrt{\frac{136}{320}\cdot\frac{184}{320}\cdot\frac{1}{320}+\frac{35}{380}\cdot\frac{345}{380}\cdot\frac{1}{380}}$$

30. The value of the test statistic and 6% critical region used to test the hypothesis

H_0: $p_R - p_C = 0$ versus H_a: $p_R - p_C \neq 0$ is:

a. $\left(\frac{136}{320}-\frac{133}{380}\right)\Big/\sqrt{\frac{136}{320}\cdot\frac{184}{320}\cdot\frac{1}{320}+\frac{133}{380}\cdot\frac{247}{380}\cdot\frac{1}{380}}$, $|z| > 1.88$

b. $\left(\frac{136}{320}-\frac{133}{380}\right)\Big/\sqrt{\frac{269}{700}\cdot\frac{431}{700}\left(\frac{1}{320}+\frac{1}{380}\right)}$, $|z| > 1.96$

c. $\left(\frac{136}{320}-\frac{35}{380}\right)\Big/\sqrt{\frac{171}{700}\cdot\frac{529}{700}\left(\frac{1}{320}+\frac{1}{380}\right)}$, $|z| > 1.88$

d. $\left(\frac{136}{320}-\frac{133}{380}\right)\Big/\sqrt{\frac{136}{320}\cdot\frac{184}{320}\cdot\frac{1}{320}+\frac{133}{380}\cdot\frac{247}{380}\cdot\frac{1}{380}}$, $|z| > 1.96$

e. $\left(\frac{136}{320}-\frac{133}{380}\right)\Big/\sqrt{\frac{269}{700}\cdot\frac{431}{700}\left(\frac{1}{320}+\frac{1}{380}\right)}$, $|z| > 1.88$

Part B (25 marks)

1. **5 marks**
   Suppose that over the years the cellulose content of hay in Manitoba has a mean of 142 mg/g with a standard deviation of 7 mg/g. An agronomist wants to determine if this year's hay has a higher mean cellulose content. To test his claim he decides to take a sample of 12 cuttings. If the average cellulose content in his sample exceeds 145 mg/g, he will conclude that this year's hay has a higher mean cellulose content.
   a. Clearly state the null and alternative hypotheses for the agronomist.
   b. What is the level of significance of the test?
   c. What is the power of the test against the alternative $\mu = 146$?

2. **5 marks**
   The number of goals scored in 2002/2003 for samples of teams that made the playoffs in the National Hockey League and of those that did not are given below along with some JMP*IN* output. Can we conclude that on average teams that made the playoffs scored more goals than those that missed the playoffs at a 5% level of significance?

Playoff Teams

Team	Goals
Boston	245
Tampa Bay	219
Anaheim	203
Colorado	251
Edmonton	231

NON-playoff Teams

Team	Goals
Los Angeles	203
Calgary	186
Chicago	207
Columbus	213

Oneway Analysis of Goals By Group
Means and Std Deviations

Level	Number	Mean	Std Dev
Non-playoffs	4	202.250	11.5866
Playoffs	5	229.800	19.4731

   a. What are the appropriate hypotheses to answer the question? (Note: Clearly define any notation you use.)
   b. What is the approximate P-value for the test?
   c. Clearly state your conclusion.

3. **5 marks**
In 2001, 41% of the students enrolled in the summer session were level I students, 31% were level II, 17% were level III, and 11% were level IV (or other). The following table gives the breakdown of students registered in the current 5.200 early in May.

Level	I	II	III	IV (or other)
Enrolment	22	41	23	14

Can we conclude that the enrolment ratio for this class is the same as that of all summer students in 2001?

a. State the appropriate hypotheses. (Define any notation you introduce.)

b. Complete the test using the P-value approach and clearly state your conclusions.

c. What assumption(s) were required for the test in (b)? Indicate which ones may not be true, if any.

4. **5 marks**

a. What are the conditions required for a Poisson random variable (i.e., define the Poisson setting)?

b. A professor wants to know if a certain student was just guessing on a MC statistics test or if the student had some knowledge. To test $H_0: p = 0.2$ versus $H_a: p > 0.2$, she selects 10 questions at random. She will reject H_0 if the student gets 6 or more questions right. What is the probability of a Type II error against the alternative $p = 0.5$?

5. **5 marks**
The following provides multiple regression output for a sample of breakfast cereals. The purpose of the study was to examine the relationship between the number of **calories** per serving and the amounts of **Fat**, **Tot Carbos** and **Sugars** contained in the cereal. (Some values have been deleted.)

Multiple linear regression results

Dependent Variable: Calories

Independent Variable(s): Fat, Tot Carbos, Sugars

Parameter estimates:

Variable	Estimate	Std. Err.	Tstat	P-value
Intercept	-18.88	2.1107		
Fat	11.34	0.0056		
Tot Carbos	4.53	0.0248		
Sugars	0.1265	0.0205		

Analysis of variance table for multiple regression model:

Source	**DF**	**SS**	**MS**	**F-stat**
Model				
Error		2452.8		
Total		44055.0		

a. State the multiple linear regression model appropriate for this study. State the estimates of each of the parameters in the model.

b. Complete a test of whether any of the explanatory variables are predictors of the response variable in the form expressed by the model.

c. Determine the value of R^2 and provide a carefully worded interpretation of what it means. Your interpretation should be understandable to someone with no knowledge of statistics and should be stated in the context of **<u>this</u>** problem.

~~We have 5 – 2 = 3 df.~~

~~[We lose 1 df for estimating μ and 1 df for requiring $\Sigma O_i = \Sigma E_i$]~~

~~$\chi^2 = (32\text{-}26.77)^2/26.77 + (40\text{-}40.16)^2/40.16 + . . . + (13\text{-}7.88)^2/7.88 = 6.65$~~

~~0.05 < P-value < 0.10. At a 5% level of significance, we fail to reject H_0. There is insufficient evidence to reject the claim that the distribution is Poisson.~~

~~Review problems for module 3~~

~~1. a~~	~~7. e~~	~~13. e~~	~~19. a~~
~~2. c~~	~~8. c~~	~~14. a~~	~~20. d~~
~~3. a~~	~~9. a~~	~~15. d~~	~~21. d~~
~~4. c~~	~~10. e~~	~~16. b~~	~~22. d~~
~~5. e~~	~~11. d~~	~~17. e~~	~~23. b~~
~~6. b~~	~~12. b~~	~~18. c~~	~~24. a~~

Practice exam answers

Answers to Part A:

1. c	6. b	11. d	16. b	21. d	26. a
2. b	7. a	12. c	17. b	22. a	27. b
3. b	8. d	13. d	18. a	23. d	28. e
4. c	9. c	14. c	19. e	24. a	29. b
5. b	10. b	15. a	20. b	25. c	30. e

Answers to Part B are on the next page.

Note: When preparing for the exam you should pay particular attention to the ASSUMPTIONS or CONDITIONS that need to be satisfied for a test to be valid. This is the last column on tables such as presented on pages 151 and 152.

The flow charts presented on pages 15 and 19 of module 1 provide good “decision trees” for choosing which test to use for testing means.

On the exam it is not uncommon for students to think they did well because they correctly “plugged numbers into a formula”. However, if you use the wrong method, you will lose most of the marks because little or none of the work will pertain to the correct solution. For example, you cannot use the normal approximation when the conditions for the normal approximation are not satisfied.

Keep in mind that on the assignments, the questions are divided into units so it should be obvious which test to use. On the final exam, as on the practice exam, you need to be able to discern from the question what the correct procedure is.

Long answer

1. a. H_o: μ = 142 H_a:μ > 142

   b. $$\alpha = P(\bar{x} > 145) = P(z > \frac{145-142}{7/\sqrt{12}}) = P(z > 1.48) = 0.0694$$

   c. $$P(\bar{x} > 145 / \mu = 146) = P(z > \frac{145-146}{7/\sqrt{12}}) = P(z > -.5) = 0.6915$$

2. a. Let μ_P and μ_N represent playoff and non-playoff teams, respectively.

   H_o: μ_P - μ_N = 0 H_a: μ_P - μ_N > 0

   b. We should use pooled t test as 19.47/10.08 < 2.

   $$s_p = \sqrt{\frac{4(19.47^2) + 3(10.08^2)}{7}} = 16.56$$ t = 2.48

   0.02 < P-value < 0.025

   c. We reject H_o and conclude teams that made the playoffs scored more goals.

3. a. Let p_I, p_{II}, p_{III} and p_{IV}, represent the four levels

   H_o: p_I, = .41, p_{II} = .31, p_{III} = .17; p_{IV} = .11
   H_a: At least one of above proportions is not valid this year

   b. Expected frequencies: N*.41 = 41, 31, 17 and 11 resp.

   χ^2 = 14.97 0.001 < P-value < 0.0025

   Reject H_o, conclude at least one proportion has changed.

   c. We have counts of frequencies for categories, all expected frequencies are ≥ 1 with at most 20% < 5. Satisfied here.

4. a. i. The random variable represents a count of the number of times an event occurs during a specific period (of time or any other unit of measurement).

   ii. The probability of occurrence of an event in a specific unit of time or space is common among all units.

   iii. The number of events that occur in one unit of time or space is independent of the number that occur in other units.

   iv. The expected number of events that occur during a single unit of time or space is given by μ.

   b. X is binomial with n = 10.

   P(Type II error) = P(reject H_o when H_o is false)
   = P(X ≥ 6 / p = 0.5)
   = P(X = 6) + P(X = 7) + … + P(X = 10)
   = .2051 + .1172 + .0439 + .0098 + .0010 = **0.3770**

   Note: You CANNOT use the normal distribution here.

5. a. Model is: $\hat{y} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ with $\varepsilon_I \sim N(0, \sigma)$

Estimates: -18.88, 11.34, 4.53, 0.126 and 12.38 respectively.

b. H_o: $\beta_1 = \beta_2 = \beta_3 = 0$

H_a: at least one $\beta_j \neq 0$

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat
Model	3	41,602.2	13,867.4	90.46
Error	16	2452.8	153.3	
Total	19	44,055.0		

P-value < 0.0001 We reject H_o, conclude that at least one $\beta_j \neq 0$.

c. $R^2 = \frac{41,602.2}{44,055} = 0.944$

94% of the variation in number of calories can be explained by variations in Fat, Tot Carbos, and Sugars.

Appendix E
Selected Formulae for STAT 2000

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

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

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

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

4. Poisson Probability Distribution:

   $$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad k = 0, 1, 2, \ldots$$

5. Test statistic for zero Correlation Coefficient: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

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

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

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

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

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

PART B

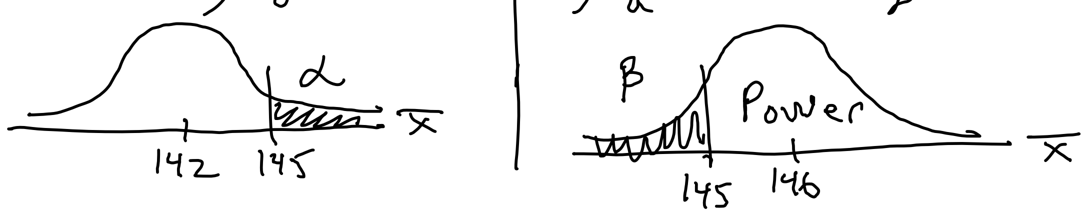
1. $\mu = 142 \quad \sigma = 7$

(a) $H_0: \mu = 142$ vs $H_a: \mu > 142$

$n = 12$, Reject H_0 if $\bar{x} > 145$

Find α = level of significance
= probability of Type I error

α	β
Reject H_0 if $\bar{x} > 145$	Do not reject H_0 if $\bar{x} \leq 145$
$\mu_0 = 142$	$\mu_a = 146$ (given in part (c))



(b) $\alpha = P(\text{Type I error})$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{145 - 142}{7/\sqrt{12}} \rightarrow z = 1.49$$

$\alpha = P(z > 1.49) = 1 - .93119 = .0681$

α = level of significance = .0681

(c) Power = $1 - \beta$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{145 - 146}{7/\sqrt{12}} \rightarrow z = -0.49$$

$\beta = .3121$

Power = $1 - .3121 = .6879$

2. (a) Let μ_1 = mean number of goals scored by Non-Playoff teams

μ_2 = mean number of goals scored by Playoff teams

If playoff scored more goals on average, then $\mu_1 < \mu_2$

$H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 < \mu_2$ non Playoff < Playoff

(b) $n_1 = 4$ $\bar{x}_1 = 202.25$ $s_1 = 11.5866$

$n_2 = 5$ $\bar{x}_2 = 229.8$ $s_2 = 19.4731$

$\frac{s_2}{s_1} = \frac{19.4731}{11.5866} < 2$ → use Pooled Method (Assume $\sigma_1^2 = \sigma_2^2$)

$df = n_1 + n_2 - 2 = 4 + 5 - 2 = 7$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{3(11.5866)^2 + 4(19.4731)^2}{7}$$

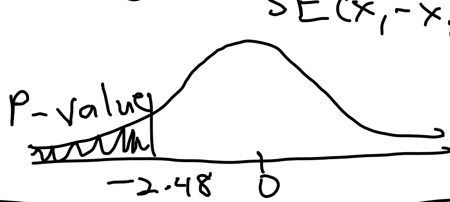
$s_p^2 = 274.222$ → $s_p = 16.5597$

$$SE(\bar{x}_1 - \bar{x}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 16.5597\sqrt{\frac{1}{4} + \frac{1}{5}}$$

$SE(\bar{x}_1 - \bar{x}_2) = 11.10855$

$$\text{test statistic} = t = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{202.25 - 229.8}{11.10855}$$

$t = -2.48$



P-value is between .02 and .025.

(c) Reject H_0 (P-value < 5%). There is statistically significant evidence Playoff teams score more goals on average.

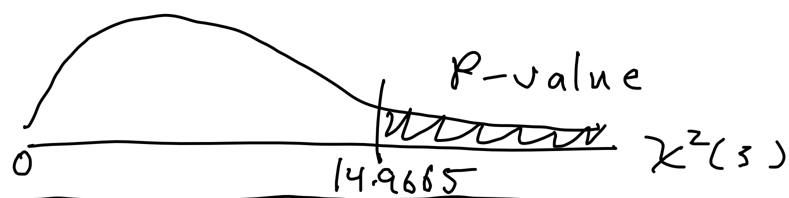
3. Chi-Square Goodness-of-Fit

Model	Obs	Exp	χ^2
I (.41)	22	$.41 \times 100 = 41$	$\frac{(22-41)^2}{41} = 8.8049$
II (.31)	41	$.31 \times 100 = 31$	$\frac{(41-31)^2}{31} = 3.2258$
III (.17)	23	$.17 \times 100 = 17$	$\frac{(23-17)^2}{17} = 2.1176$
IV (.11)	14	$.11 \times 100 = 11$	$\frac{(14-11)^2}{11} = 0.8182$
TOTAL	100	100	$\chi^2 = 14.9665$

(a) H_0: The current class has the same distribution as in 2001.

H_a: The current class does not have the same distribution.

(b) $df = 4 - 1 = 3$, test statistic $= 14.9665$



P-value is between .001 and .0025

Reject H_0 (P-value < 5%). There is statistically significant evidence that the current class does not have the same distribution as in 2001.

(c) We must assume this 5.200 class is a representatitive sample of all summer students. This is almost certainly not true which would make our conclusion in part (b) suspect.

4. (a) Look at the Answer Key

Also memorize the four conditions for a Binomial setting:

1. There are a fixed number of trials, n
2. Each trial has only two possible outcomes: Success or Failure
3. The probability of success is the same on each trial, p.
4. The trials are independent.

(b) $H_0: p = 0.2$ vs $H_a: p > 0.2$

$n = 10 \rightarrow$ Binomial $\quad \beta = P(\text{Type II error})$

α	β
Reject H_0 if $X \geq 6$	Don't reject H_0 if $X < 6$
$p_0 = 0.2$	$p_a = 0.5$
$n = 10$	$n = 10$
$X = 0, 1, 2 \ldots \boxed{6 \ldots 10}$ (α)	$X = \boxed{0, 1, \ldots 4, 5} \; 6 \ldots 10$ (β)

$\beta = P(X < 6)$ for Binomial wher $n = 10$, $p = .5$

$\beta = .0010 + .0098 + .0439 + .1172 + .2051 + .2461$

$\boxed{\beta = .6231}$ = Probability of Type II error

5. (a) Model:

$$y_i = \alpha + \beta_1 (\text{Fat})_i + \beta_2 (\text{Tot Carbos})_i + \beta_3 (\text{Sugars})_i + \varepsilon_i$$

where y = Calories, α = true intercept, β_1, β_2, and β_3 are the true coefficients for Fat, Tot Carbos, and Sugars, respectively and ε_i is the residual for any observation (Fat_i, Tot Carbos_i, Sugars_i, y_i)

Estimates for α is $a = -18.88$, β_1 is $b_1 = 11.34$, β_2 is $b_2 = 4.53$, and β_3 is $b_3 = 0.1265$

Prediction equation:

$$\hat{y} = -18.88 + 11.34(\text{Fat}) + 4.53(\text{Tot Carbos}) + 0.1265(\text{Sugars})$$

(b) Anova F-test   Note: n=20 was not given by mistake

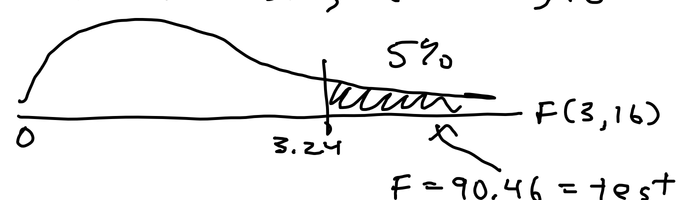
$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs H_a: at least one of $\beta_1, \beta_2, \beta_3 \neq 0$

k = # of variables = 4

	df	SS	MS
Model	$k-1 = 3$	41602.2	13867.4
Error	$n-k = 16$	2452.8	153.3
C. Total	$n-1 = 19$	44055.0	

$$F = \frac{13867.4}{153.3} = 90.46$$

Use $\alpha = 5\%$, df = 3,16 → $F^* = 3.24$



Reject H_0. There is statistically significant evidence that at least one of Fat, Tot Carbos or Sugars is a predictor of calories.

(c) $$R^2 = \frac{SSM}{SST} = \frac{41602.2}{44055.0} = .9443$$

94.43% of the variation in calories can be explained by this multiple linear regression model.