Student's Name:

SFU id #:

Instructions:

You are permitted 1 two-sided single page of notes

Calculators, cell phones and other electronics are not permitted.

Your bag and jacket (if applicable) must be against a wall of the room.

If a question asks you for a number that would involve some computation (+, -, x or ÷), you don't need to compute the value, just leave it as something like 55+100 or 12 ÷ 3

You can write hypotheses in you choice of symbols or words.

Unless otherwise specified use a 5% level of significance.

Anytime I ask you for a conclusion be sure to justify your answer with a p-value to get full marks.

page		maximum points on that page
2	-	9
3	-	7
4	-	4
5	-	4
6	-	3
7	-	5
8	-	10
11	-	13
12	-	8
14	-	8
15	-	6
16	-	5
total;		82

1. Final exam scores in a class have the following descriptive statistics:
min 55 max 190
Quartiles are 60, 70 and 100
Mean is 95

a) Draw the approximate shape of the distribution of grades. (3pts)

b)What proportion of students scored lower than 100? (1pt)

2. The grades on a midterm are uniformly distributed and you want to know about the population average grade. You obtain a random sample of 100 grades, find the sample mean is 79 and know that the population standard deviation (σ) is 20 and the standard error of the mean is 2. You want to make a 90% confidence interval for the mean.

a) What is the value from the table that you need for this interval? (2pts)

b) Describe the shape of the sampling distribution of the mean? Give a description or name of the distribution and use actual values (where applicable). (3pts)

3. A candy comes in 6 different colour varieties, and a researcher is wondering if the different colour candies are evenly distributed in the candy population. The researcher buys a really big box of candy and counts the frequency of each coloured candy in the box. The data was entered into SPSS and the output is given below. (Candy colours were entered into SPSS as group numbers from 1 to 6 instead of actual colour names.)

a) What are the null and research hypotheses? (2 pts)

b) What do you conclude? Justify your answer (3pts)

c) In the SPSS output, why are the values in the 'Expected N' and 'Observed N' different? (2pt)

Chi-Square Test

Frequencies

colour_number

	Observed N	Expected N	Residual
1.00	77	77.0	.0
2.00	67	77.0	-10.0
3.00	70	77.0	-7.0
4.00	88	77.0	11.0
5.00	85	77.0	8.0
6.00	75	77.0	-2.0
Total	462		

Test Statistics

	colour_number
Chi-Square	4.390[a]
df	5
Asymp. Sig.	.495

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 77.0.

4. Using the MTF data set we wish to examine the variable about lifetime cocaine usage. In the first part of our analysis SPSS gives us the following box.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
062B10A:#X COKE/LIFETIME * 062C03 :R'S SEX	13717.749[a]	92.6%	1096.254	7.4%	14814.003	100.0%

a) Why isn't N a whole number? In other words, why is N (Valid) 13717.749 when we can't have .749 of a person? (1pts)

b) There is a procedure that causes the lack of whole numbers above. Explain why that procedure is important to generalizing the results of the analysis to the population. (3pt)

5. Using the Best Places data we wish to test the following hypotheses:

H_o: There is no difference in mean unemployment rate between the 4 different geographic regions of the USA

H_r: There is a difference in mean unemployment rate between the 4 different geographic regions of the USA.

The SPSS output is below

ANOVA

Unemployment rate (Bureau of Labor Statistics)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	35.449	3	11.816	3.796	.011
Within Groups	1014.879	326	3.113		
Total	1050.328	329			

a) What is the research question? (1pt)

b) How many people were in the study (1pt)

c) What do you conclude? (2 pts)

6. (1pt) The t distribution more closely approximates the distribution of the normal curve when:
   a) the degrees of freedom decrease
   b) the degrees of freedom increase
   c) the sample size decreases
   d) the mean of the sample grows further from the mean of the population

7. (1pt) When using a T-test for a difference between means, when we retain the null hypothesis, we:
   a) claim that a significant difference exists between groups.
   b) have obtained a t-value greater than our critical t-value.
   c) have committed a Type 1 error.
   d) conclude that sampling error is responsible for our obtained difference.

8. (1pt) Choose the correct word to complete this sentence:

The larger the value of the test statistic, the _____ the p-value will be.

   a) smaller
   b) larger

9. At the bottom of the page is the correlation between several variables from the Best Places data set. Let's shorten the variable names as follows to make writing your answers easier:
P - Property Crimes per 100,000 **A - Average Alcoholic Drinks per month**
S - Suicides per 100,000 population **D - Days during the past 30 days where...**
The SPSS output is given below.

Correlations

		Property Crimes per 100,000 (Uniform Crime Reports, 2002)	Average alcoholic drinks per month (CDC BRFSS)	Suicides per 100,000 population (CDC mortality tables)	Days during the past 30 days where your mental health not good, including stress, depression, and problems with emotions (CDC BRFSS)
Property Crimes per 100,000 (Uniform Crime Reports, 2002)	Pearson Correlation	1	-.156	.195	.025
	Sig. (2-tailed)		.004	.000	.646
	N	331	331	331	331
Average alcoholic drinks per month (CDC BRFSS)	Pearson Correlation	-.156	1	.040	-.130
	Sig. (2-tailed)	.004		.471	.018
	N	331	331	331	331
Suicides per 100,000 population (CDC mortality tables)	Pearson Correlation	.195	.040	1	.072
	Sig. (2-tailed)	.000	.471		.189
	N	331	331	331	331
Days during the past 30 days where your mental health not good, including stress, depression, and problems with emotions (CDC BRFSS)	Pearson Correlation	.025	-.130	.072	1
	Sig. (2-tailed)	.646	.018	.189	
	N	331	331	331	331

a) In the bottom right corner box of the correlation matrix, the values are 1 and 331. Why is the correlation equal to 1? (2pts)

b) What can you say about the association between Property Crimes per 100,000 and Average alcoholic drinks per month? (3pts)

10. Draw a picture of what some data might look like if the correlation was - 0.5 (3pts)

11. Give a definition for the p-value. (4pts)

12. What are the values of the quartiles of a standard normal variable (I'm looking for the actual numbers not a definition)? (3pts)

13. We would like to model the property crime rate of cities in the US based on the stress level (stress index) of the city. The SPSS output is below. The questions to answer are on page 11. Feel free to remove this page if you like, the next page is blank.
To simplify feel free to use these abbreviations for the variable names:
S - Stress Index
C - Property Crime rate per 100,000 population.

Use this SPSS output for the questions on pages 10 and 11

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.439[a]	.193	.190	1050.2285

a. Predictors: (Constant), Stress Index (www.BestPlaces.net)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.675E7	1	8.675E7	78.653	.000[a]
	Residual	3.629E8	329	1102979.995		
	Total	4.496E8	330			

a. Predictors: (Constant), Stress Index (www.BestPlaces.net)

b. Dependent Variable: Property Crimes per 100,000 (Uniform Crime Reports, 2002)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3066.722	115.097		26.645	.000
	Stress Index (www. BestPlaces.net)	17.683	1.994	.439	8.869	.000

a. Dependent Variable: Property Crimes per 100,000 (Uniform Crime Reports, 2002)

Intentionally left blank

[13. Continued and using the SPSS output from page 9]
As stated earlier feel free to simplify using these abbreviations for the variable names:
S - Stress Index **C - Property Crime rate per 100,000 population.**

a)Which variables are independent and dependent in the model? (2pts)

b) Write down the model being considered. Use actual values from the SPSS output. (4 pts)

c) Can you conclude that the model is a significant predictor of C? Justify your answer (3pts)

d) How much of the variability in C is modeled by S? (1pt)

e)If the stress index increases by 1 unit, according to the model, by how much will C change? (1pt)

g) In the 'Coefficients' table of the output, this question refers to the term labelled '(Constant)'. When will our model predict a value of C that is equal to the constant (3066.722)? (2pts)

14. Define type 2 error. (2pts)

15. Define type 1 error. (2pts)

16. If you get a sample and make a 95% confidence interval for the mean, then get a new sample and make a new confidence interval and repeat the process with 100 different samples, how many of those intervals would you expect to contain the true population mean? (2pts)

17. When will one-way ANOVA and a t-test for the difference between two means give you the same result? (2pts)

(SPSS output for question 18 on the following page)

Commute_time * Regional location as defined by Bureau of the Census Crosstabulation

			Regional location as defined by Bureau of the Census				
			Northeast	Midwest	South	West	Total
Commute_time	short	Count	22	64	48	31	165
		% within Regional location as defined by Bureau of the Census	37.9%	75.3%	40.3%	47.7%	50.5%
	long	Count	36	21	71	34	162
		% within Regional location as defined by Bureau of the Census	62.1%	24.7%	59.7%	52.3%	49.5%
Total		Count	58	85	119	65	327
		% within Regional location as defined by Bureau of the Census	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	29.691[a]	3	.000
Likelihood Ratio	30.789	3	.000
Linear-by-Linear Association	.860	1	.354
N of Valid Cases	327		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 28.73.

18. A researcher wishes to know if commute time varies with region in the US. The data has been transformed into ordinal or nominal variables. The Cross-Tabulation output from SPSS is on page 13.

a) Does the proportion of people with a short commute vary with changes in regions? Justify your answer. (3pts)

b) What is the value of the test statistic? (1pt)

c) In the cross-tabulation table, what does the circled number 52.3% mean? (What I want to know is how is this number interpreted?) (2 pts)

d) How many respondents have a long commute? (2 pts)

19. A professor is trying to decide if one of two potential versions of an exam is more difficult than the other. The professor chooses 2 random samples of students and randomly assigns them to write either midterm version 1 or version 2. The SPSS output is below.

Group Statistics

	Midterm	N	Mean	Std. Deviation	Std. Error Mean
Score	1.00	96	74.3025	13.66171	1.39434
	2.00	92	50.3170	14.86797	1.55009

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Score	Equal variances assumed	1.309	.254	11.525	186	.000	23.98549	2.08118	19.87974	28.09125
	Equal variances not assumed			11.504	183.046	.000	23.98549	2.08494	19.87189	28.09910

a) Which midterm version has a higher mean score? (1pt)

b) Is it reasonable to assume that the two groups had the same variability? Justify your answer. (3 pts)

c)Is there evidence to suggest that one exam is harder than the other? Justify your answer and circle the p-value on the SPSS output. (3 pts)

20. Classify the measurement type in each of the following examples (1pt each):

a) What dorm you live in _____

b) Number of children in a family _____

c) Tuition in dollars _____

d) Attitudes toward premarital sex between consenting adults (always wrong, usually wrong, sometimes wrong, never wrong) _____

e) Racial categories _____

Congratulations, you've reached the end of the exam!