STAT 203

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	-	6
4	-	4
5	-	4
6	-	3
7	-	5
8	-	9
11	-	14
12	-	9
14	-	8
15	-	6
16	-	5
total;		81

1. A sample of 100 test scores have the following descriptive statistics:
min 45 max 100
Quartiles are 50, 55 and 80

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

b) Which will be higher, the mean or median (or will they be about the same)? Justify your answer (2pts)

c)What proportion of scores are below 55? (1pt)

d) What is the shape of the sampling distribution of the mean for this distribution? (1pts)

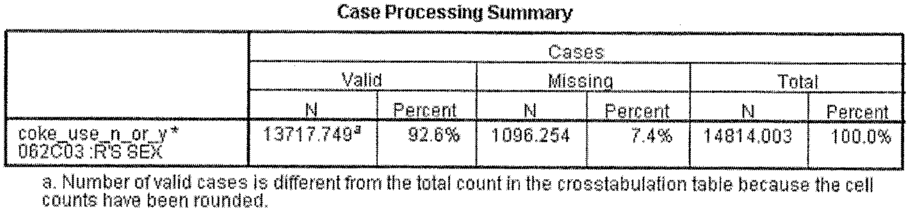
e) If you were to take a larger sample what do you expect to happen to the central tendency and variability of the sampling distribution of the mean? (2pts)

3. A researcher wishes to see if there is a difference in lifetime cocaine usage between genders through the Monitoring the Future questionnaire. The SPSS output is given below. The variable "coke_use_n_or_y" takes values 0 when there has never been any cocaine usage and 1 when cocaine was used at least once. The variable "R's Sex" takes values 1 for Males and 2 for Females.

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

b) What do you conclude? Justify your answer and circle the p-value that you used (3pts)

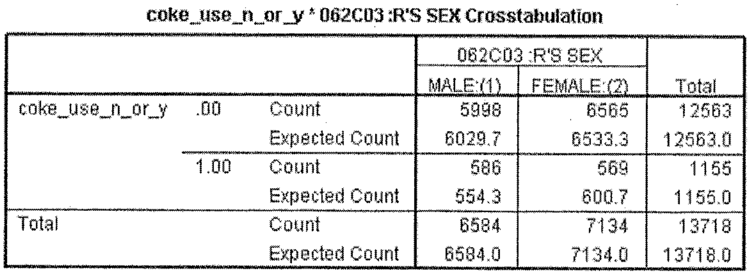
c) Name a type of plots that would be appropriate for the showing the counts of individuals from the lifetime cocaine use variable (1pt)



Case Processing Summary

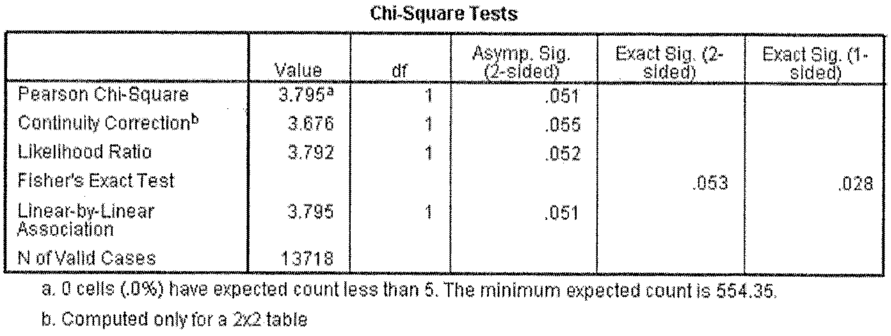
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
coke_use_n_or_y * 062C03 :R'S SEX	13717.749[a]	92.6%	1096.254	7.4%	14814.003	100.0%

a. Number of valid cases is different from the total count in the crosstabulation table because the cell counts have been rounded.



coke_use_n_or_y * 062C03 :R'S SEX Crosstabulation

			062C03 :R'S SEX		
			MALE:(1)	FEMALE:(2)	Total
coke_use_n_or_y	.00	Count	5998	6565	12563
		Expected Count	6029.7	6533.3	12563.0
	1.00	Count	586	569	1155
		Expected Count	554.3	600.7	1155.0
Total		Count	6584	7134	13718
		Expected Count	6584.0	7134.0	13718.0



Chi-Square Tests

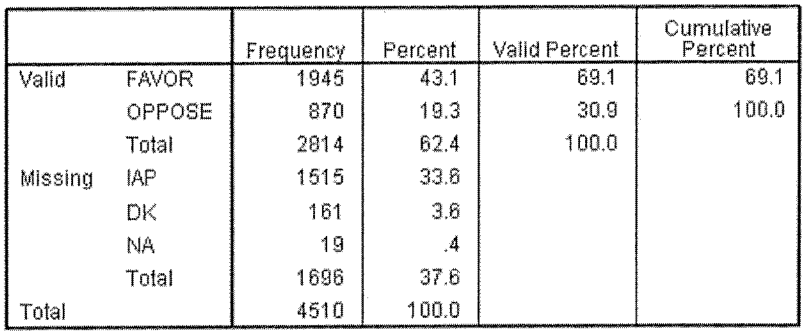
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.795[a]	1	.051		
Continuity Correction[b]	3.676	1	.055		
Likelihood Ratio	3.792	1	.052		
Fisher's Exact Test				.053	.028
Linear-by-Linear Association	3.795	1	.051		
N of Valid Cases	13718				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 554.35.

b. Computed only for a 2x2 table

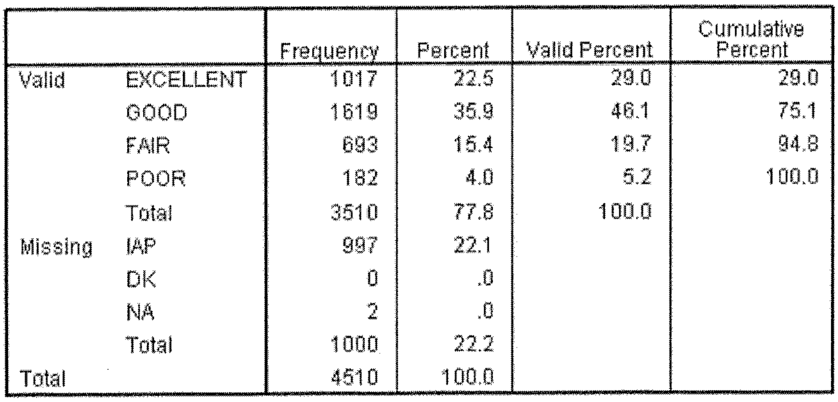
4. Two variables obtained from the General Social Survey (GSS) are summarized below. Questions in the GSS were recorded using a scantron multiple choice system similar to the way course evaluations are recorded. Categories NA, DK, and IAP are different types of missing responses.

FAVOR OR OPPOSE DEATH PENALTY FOR MURDER



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FAVOR	1945	43.1	69.1	69.1
	OPPOSE	870	19.3	30.9	100.0
	Total	2814	62.4	100.0	
Missing	IAP	1515	33.6		
	DK	161	3.6		
	NA	19	.4		
	Total	1696	37.6		
Total		4510	100.0		

CONDITION OF HEALTH



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	EXCELLENT	1017	22.5	29.0	29.0
	GOOD	1619	35.9	46.1	75.1
	FAIR	693	15.4	19.7	94.8
	POOR	182	4.0	5.2	100.0
	Total	3510	77.8	100.0	
Missing	IAP	997	22.1		
	DK	0	.0		
	NA	2	.0		
	Total	1000	22.2		
Total		4510	100.0		

a) **How many people gave valid answers when asked if they favour or oppose the death penalty for murder? (1pt)**

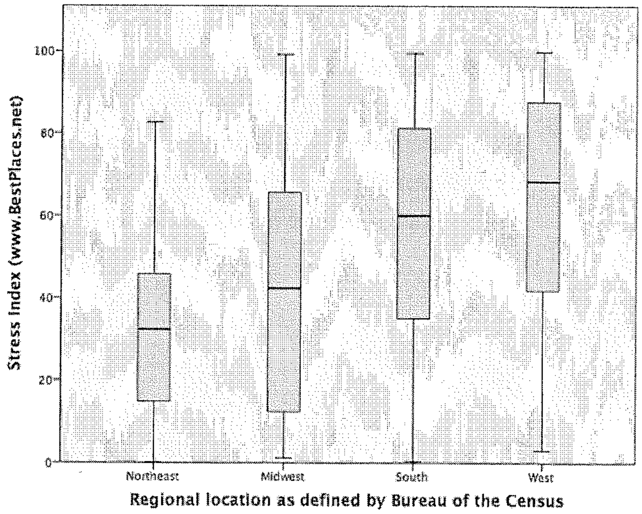
b) **What percent of respondents who answered the question listed their condition of health as poor? (1pt)**

c) **Give the mode for each of the variables? (2pts)**

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

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

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



ANOVA

Stress Index (www.BestPlaces.net)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	39773.851	3	13257.950	18.241	.000
Within Groups	237673.526	327	726.830		
Total	277447.378	330			

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

b) How many valid responses were available? (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 standard deviation 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 will ANOVA and a t-test comparing 2 means give the same conclusion?

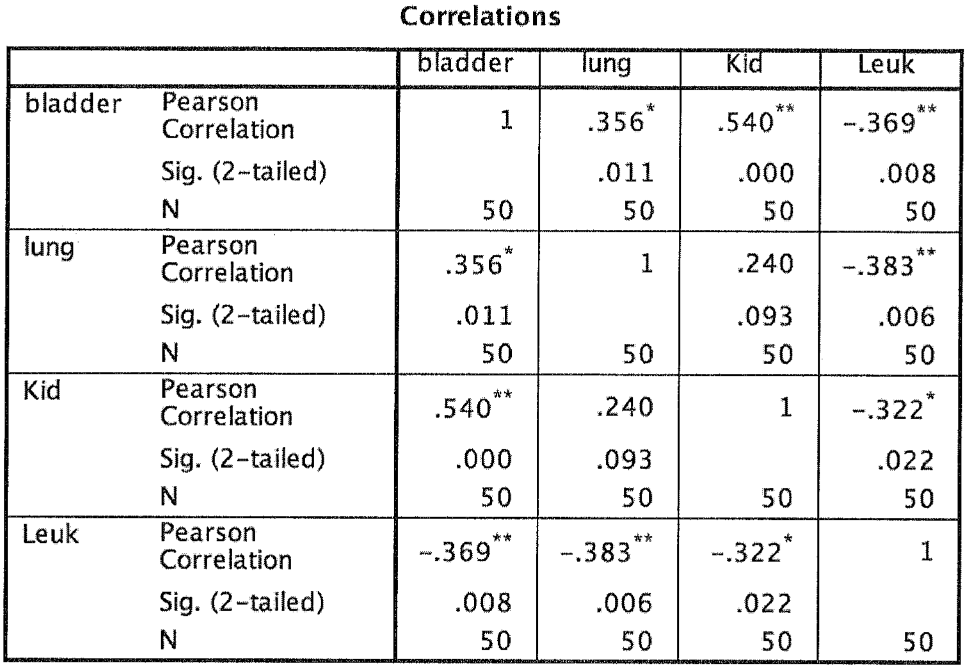
a) When using a matched pairs comparison with a 2-sided hypothesis test.
b) When using independent samples comparison with a 2 sided hypothesis test.
c) They don't. T-tests and ANOVA do different things.
d) When the 'degrees of freedom within' matches the test statistic.

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

A large difference between what is observed and what is expected compared to the variability in the sampling distribution will result in a _________ p-value

a) small
b) large

9. The SPSS output shows correlations between several variables from the smoking and cancer set.



Correlations

		bladder	lung	Kid	Leuk
bladder	Pearson Correlation	1	.356*	.540**	-.369**
	Sig. (2-tailed)		.011	.000	.008
	N	50	50	50	50
lung	Pearson Correlation	.356*	1	.240	-.383**
	Sig. (2-tailed)	.011		.093	.006
	N	50	50	50	50
Kid	Pearson Correlation	.540**	.240	1	-.322*
	Sig. (2-tailed)	.000	.093		.022
	N	50	50	50	50
Leuk	Pearson Correlation	-.369**	-.383**	-.322*	1
	Sig. (2-tailed)	.008	.006	.022	
	N	50	50	50	50

a) Which variables have significant positive linear relationships at the 1% significance level? (2pts)

b) What can you say about the association between Bladder and Leuk? (3pts)

10. Draw a picture of what Kid and bladder might look like given the SPSS output on page 7 (3pts)

11. A 95% confidence interval for a mean is from -2 to 12 with a sample mean of 5. What would you conclude is you then decided to test:
Ho: population mean is 12
Ha: population mean is larger than 12
(4pts for conclusion and p-value)

12. Between which 2 values is 60% of a standard normal distribution contained? (2pts)

13. We would like to model the lung cancer rates based on number of cigarettes smoked per capita. The SPSS output is below and on the next page. The questions to answer are on page 11. Feel free to remove this page if you like,

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.606[a]	.367	.354	2.30993028

a. Predictors: (Constant), cig

b. Dependent Variable: lung

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	148.450	1	148.450	27.822	.000[b]
	Residual	256.117	48	5.336		
	Total	404.567	49			

a. Dependent Variable: lung

b. Predictors: (Constant), cig

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.928	2.849		1.730	.090	-.801	10.657
	cig	.497	.094	.606	5.275	.000	.308	.687

a. Dependent Variable: lung

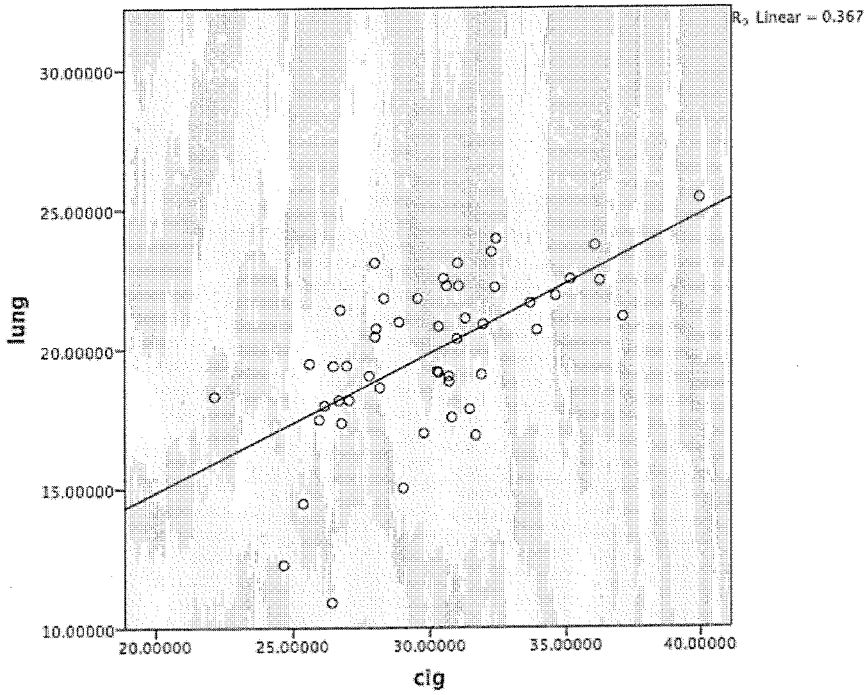
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	15.9386435	24.7849197	19.8577342	1.74057075	50
Residual	-7.1466236	4.24703836	.00000000	2.28623806	50
Std. Predicted Value	-2.252	2.831	.000	1.000	50
Std. Residual	-3.094	1.839	.000	.990	50

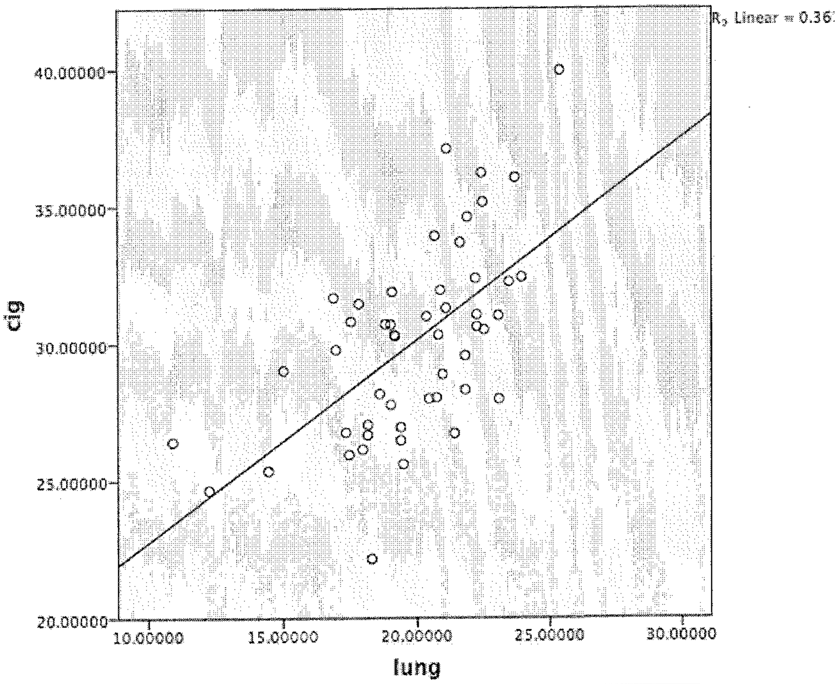
a. Dependent Variable: lung

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

Plot #1


R^2 Linear = 0.367
lung
cig
30.00000
25.00000
20.00000
15.00000
10.00000
20.00000
25.00000
30.00000
35.00000
40.00000

Plot #2


R^2 Linear = 0.36
cig
lung
40.00000
35.00000
30.00000
25.00000
20.00000
10.00000
15.00000
20.00000
25.00000
30.00000

[13. Continued and using the SPSS output from page 9]

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) Given your answer to the above should we use Plot #1 or Plot #2 on page 10 to visualize the data? (1pts)

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

d) How much of the variability in one variable can be explained by the least squares regression line? (1pt)

e)If the number of cigarettes increases by 1 unit, according to the model, by how many units will lung 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 that is equal to the constant? (2pts)

14. How do you reduce the chance of a type 2 error? (2pts)

15. How do you reduce the chance of a 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. Define p-value (3pts)

(SPSS output for question 18 on the following page)

Satisfied with the Amount of Fun * 062C03 :R'S SEX Crosstabulation

			062C03 :R'S SEX		Total
			MALE:(1)	FEMALE:(2)	
Satisfied with the Amount of Fun	COMP DIS:(1)	Count	36	35	71
		% within Satisfied with the Amount of Fun	50.7%	49.3%	100.0%
		% within 062C03 :R'S SEX	3.3%	3.0%	3.1%
	2	Count	38	62	100
		% within Satisfied with the Amount of Fun	38.0%	62.0%	100.0%
		% within 062C03 :R'S SEX	3.5%	5.2%	4.4%
	3	Count	85	83	168
		% within Satisfied with the Amount of Fun	50.6%	49.4%	100.0%
		% within 062C03 :R'S SEX	7.8%	7.0%	7.4%
	NEUTRAL:(4)	Count	178	204	382
		% within Satisfied with the Amount of Fun	46.6%	53.4%	100.0%
		% within 062C03 :R'S SEX	16.3%	17.2%	16.8%
	5	Count	215	259	474
		% within Satisfied with the Amount of Fun	45.4%	54.6%	100.0%
		% within 062C03 :R'S SEX	19.7%	21.9%	20.8%
	6	Count	308	291	599
		% within Satisfied with the Amount of Fun	51.4%	48.6%	100.0%
		% within 062C03 :R'S SEX	28.2%	24.6%	26.3%
	COMP SAT:(7)	Count	233	250	483
		% within Satisfied with the Amount of Fun	48.2%	51.8%	100.0%
		% within 062C03 :R'S SEX	21.3%	21.1%	21.2%
Total		Count	1093	1184	2277
		% within Satisfied with the Amount of Fun	48.0%	52.0%	100.0%
		% within 062C03 :R'S SEX	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.110[a]	6	.167
Likelihood Ratio	9.159	6	.165
Linear-by-Linear Association	1.249	1	.264
N of Valid Cases	2277		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 34.08.

18. A researcher wishes to know if the Males and Females differ in their satisfaction about the amount of fun they are having and obtains the variables "Sex" and "Satisfied with the amount of fun". The "Satisfied with the amount of fun" variable is coded so that 1 is Completely Dissatisfied and 7 is Completely Satisfied.

a) Is the median an appropriate measure for either variable? state why/ why not for both variables? (2pts)

b) The number 38.0% is circled in the table. Give a sentence interpreting that number in this table. (1pt)

c) What percent of those who are completely dissatisfied with the amount of fun are Males? (1pt)

d) What percent of Females are completely satisfied with the amount of fun? (1pt)

e) How many people answered both questions? (1pt)

f) What are the hypotheses that the researcher wishes to test? (2pts)

g) What do you conclude? (3pts)

19. In the General Social Survey a researches wishes to know if there is a difference between the average age between genders. The SPSS output is below.

Group Statistics

	RESPONDENTS SEX	N	Mean	Std. Deviation	Std. Error Mean
AGE OF RESPONDENT	MALE	2050	44.93	16.443	.363
	FEMALE	2445	45.69	16.627	.336

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
AGE OF RESPONDENT	Equal variances assumed	.002	.969	-1.523	4492	.128	-.755	.495	-1.726	.217
	Equal variances not assumed			-1.525	4372.555	.127	-.755	.495	-1.725	.216

a) Which gender has a higher mean? (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 sex has an older average age 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!