MATH 1203 – Practice Exam Final

This is a practice exam only. The actual exam may differ from this practice exam.
*In fact, there are **many more questions here than will be on the final exam**.*

1. Please state, in your own words, what the following terms mean

   a) Population
   b) Sample, random sample
   c) Mean, median, mode, range, variance, standard deviation, percentiles
   d) Heterogeneous distribution, homogeneous distribution
   e) Numerical variable, categorical, nominal, ordinal, continuous, discrete variable
   f) Central Limit Theorem
   g) Confidence Interval
   h) Hypothesis Test
   i) Contingency table and Chi-square test
   j) Scatter plot, least-square regression line, correlation coefficient, predictions
   k) Skewed Distribution, box plot
   l) t-score, z-score
   m) Proportion, probability of success, standard deviation of a proportion

2. Please decide if the following statements are true or false.

   a) To compute the variance, you must first compute the mean.

FALSE

   b) To draw a box plot, you need the highest and lowest values as well as the mean and the standard deviation.

FALSE

   c) The median is influenced by extremely large or extremely small values.

FALSE

   d) The standard deviation is the square root of the variance.

TRUE

   d) A crosstabs or contingency table predicts the value of the independent variable based on the dependent variable.

TRUE

   f) Suppose you compute the equation of a least-square regression line as $y = 2x + 3$ and the correlation coefficient $r = 0.8$, could that be possible.

YES (TRUE)

   g) A 90% confidence interval is *smaller* than a 95% confidence interval

TRUE

3. Pleease provide brief answers to the following questions:

   a) If you are using a **t-distribution** with **df = 10** for a **2-tail** statistical test at the **a = 0.05** level, then the corresponding number $\mathbf{t_a}$ will be what?

2.228

   b) If you are using a **t-distribution** with **df = 9** for a **2-tail** statistical test, and the number t_0 you compute is $\mathbf{t_0 = 2.45}$, whereas the number t_1 you look up is $\mathbf{t_1 = 2.262}$. What is your conclusion for the corresponding test?

REJECT H_0

   c) If you are using **z-distribution** for a **1-tail** statistical test at the usual 5% level of significance, the number z_0 you compute is $\mathbf{z_0 = 1.64}$, and the corresponding p-value for that value of z_0 is **p = 0.0505**. What is your conclusion for the corresponding test?

INCONCLUSIVE

   d) Someone is interested in designing a statistical test for the mean of a population. In deciding whether to use a test based on the ***t***-distribution or a test based on the standard normal distribution, what is the deciding factor?

SAMPLE SIZE

e) You are conducting a **2-tailed** statistical test for the population mean at the $\alpha = 0.05$ level. The null hypothesis is $H_0 = 17.1$, while the alternative hypothesis is $H_a > 17.1$. The sample size is large enough to use a normal distribution, and the statistics for the sample turns out to be $z_0 = 2.045$. From the standard normal table for the z-distribution you compute $P(z > 2.045) = 0.0202$. What is your conclusion?

REJECT H_0

f) A statistical test for the population mean at the $\alpha = 0.05$ level results in your rejection of the null hypothesis. Can the null hypothesis still be true? If so, what is the probability that the null hypothesis is true, even though you rejected it?

H_0 CAN BE TRUE WITH PROB AT MOST 5%

4. Below is a short segment from a (fictitious) survey questionnaire. How many variables can you identify? For each variable, state whether it is nominal, ordinal, or numeric.

3 variables (numeric, ordinal, nominal)

Please state your age: _____	The statistics class MATH 1203 is useful and interesting [ ] Strongly Agree [ ] Agree [ ] Disagree [ ] Strongly Disagree	In which area is your major or intended major field of study [ ] Social Sciences [ ] Natural Sciences [ ] Art and Literature [ ] Other

5. Use the frequency distribution listed below to answer the following questions:

RS HIGHEST DEGREE

	Freq.	Percent	Valid %	Cum %
LT HIGH SCHOOL	448	15.4	15.5	15.5
HIGH SCHOOL	1567	54.0	54.1	69.6
JUNIOR COLLEGE	187	6.4	6.5	76.0
BACHELOR	471	16.2	16.3	92.3
GRADUATE	224	7.7	7.7	100.0
Total	2897	99.8	100.0	
Missing	7	.2		
Total	2904	100.0		

a) What percentage of respondents have a High School degree?
54.1% (valid percent)

b) What percentage of respondents have *at most* a Junior College degree?
(100 – 16.3 – 7.7)%

c) What percentage of respondents have *at least* at Bachelor degree?
(16.3 + 7.7)%

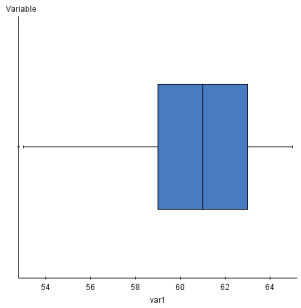
6. Suppose a random sample of size 9 taken from the GSS survey shows that the age for that sample are as follows:

53, 65, 54, 64, 59, 60, 61, 61, 63

a) Find the mode, the mean, and the median for this data
b) Find Q_1, Q_3, and the IQR for this data
c) Draw an approximate box plot for this data
d) Find the variance and the standard deviation for this data

Summary statistics:

Column	Mean	Variance	Std. Dev.	Median	Q1	Q3	IQR
var1	60	17.25	4.1533117	61	59	63	4



7. A random sample of size 20 selected from the GSS shows the following distribution for the number of children of the respondent.

NUMBER OF CHILDREN

		Frequency	Percent	Valid %	Cumulative %	
Valid	0	4	20.0	20.0	20.0	
	1	3	15.0	15.0	35.0	
	2	6	30.0	30.0	65.0	
	3	4	20.0	20.0	85.0	
	4	3	15.0	15.0	100.0	
	Total	20	100.0	100.0		

a) Is this distribution heterogeneous or homogeneous?

HETEROGENEOUS

a) Find the mean, mode, and median for this distribution, if possible

MEAN = makes no sense here
MEDIAN = 2 children
MODE = 2 children

b) Find Q_1, Q_3, and the IQR.

Q1 = 1, Q3 = 3, IQR = 3-1=2

8. The table below shows a contingency table for the variables "DEGREE" by "LIFEFUN". (Almost) each cell lists three numbers: the count (top), the row percentage (middle), and the column percentage (bottom).

RS HIGHEST DEGREE vs IS LIFE EXCITING OR DULL

	IS LIFE EXCITING OR DULL			Total	
	DULL	ROUTINE	EXCITING		
LT HIGH SCHOOL	34 11.2% 42.5%	162 53.3% 18.7%	108 35.5% 11.4%	304 100.0%	
HIGH SCHOOL	43 4.2% 53.8%	502 49.6% 57.9%	467 46.1% 49.2%	1012 100.0%	
JUNIOR COLLEGE	1 .8% 1.3%	55	64 53.3% 6.7%	120 100.0%	
BACHELOR	2 .7% 2.5%	101 33.0% 11.6%	203 66.3% 21.4%	306 100.0%	
GRADUATE		47 30.3% 5.4%	108 69.7% 11.4%	155 100.0%	
Total	80 100.0%	867 100.0%	950 100.0%	1897 100.0%	

a) One of the cells is missing the row and column percentages. What is the missing row percentage, and what is the missing column percentage?

ROW = 45.8%, COL = 6.3%

b) How many respondents with a High School degree as highest degree think that life is exciting?

46.1%

c) How many people thinking that life is routine have a Bachelor's degree

11.6%

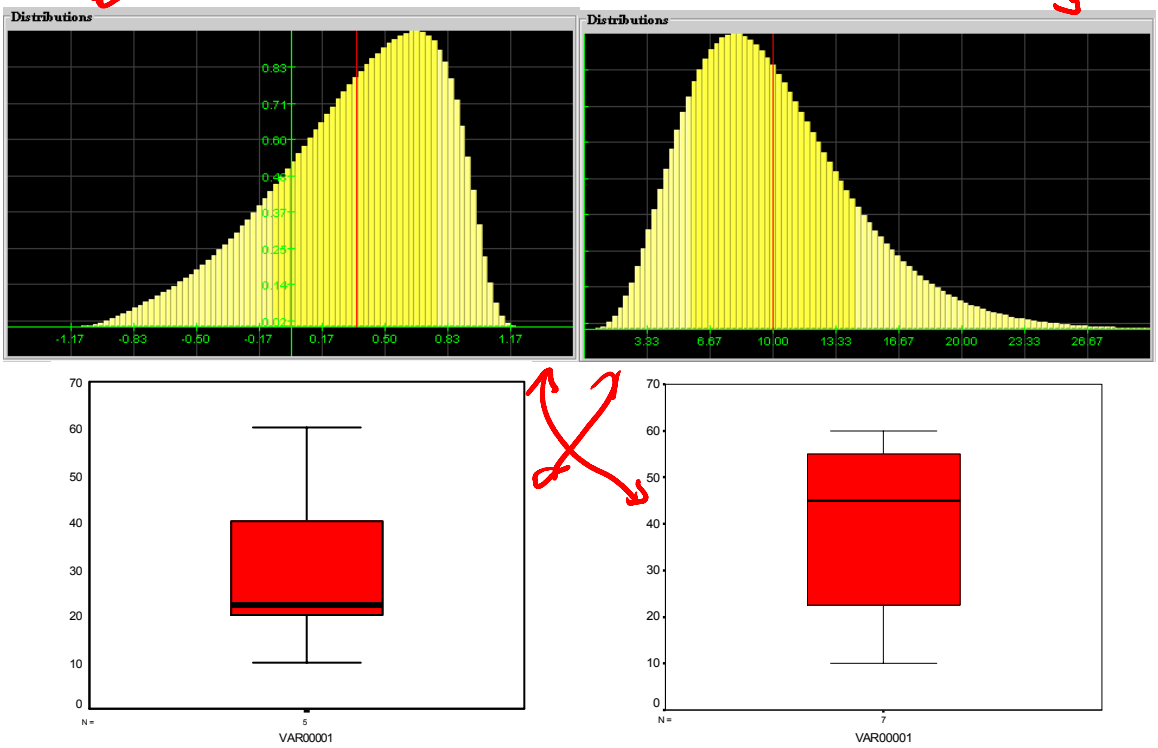
d) How many people who think that life is exciting have *at least* a Bachelor's degree

(21.4 + 11.4)%

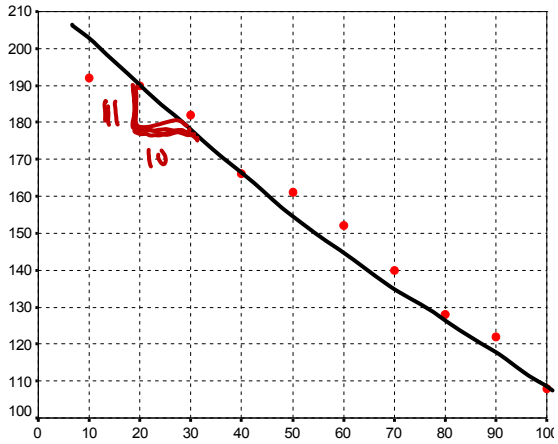
9. Please match the following statements to the distribution pictures below.

(a) Mean less than median

(b) median less than mean



10. When using StatCrunch to draw a "scatter plot, it comes up with the following picture:



a) Draw a "best-fit" line through this data.

b) Use the line to estimate the y-intercept and slope of the equation of the least-square regression line

y-intercept is about 215, slope is about 11/10 = 1.1 (approx)

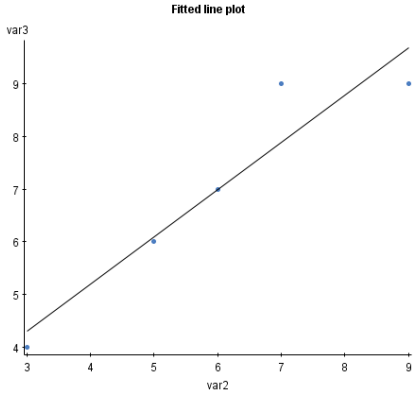
c) Look at the data and your line and estimate whether *r* would be close to -1, close to 0, or close to 1

r would be close to -1

11. Please consider the following results on a quiz, measuring scores before and after a certain lecture:

Before lecture: 5, 6, 7, 9, 3
After lecture: 6, 7, 9, 9, 4

a) Create a scatter plot representing this data, including a best-fit line for the data



b) Find the exact equation of the least-square regression line (use back page for computation, but show equation here)

Y = 0.9 x + 1.6

c) Compute Pearson's r (use back page for computation but show r here)

R = 0.9487

Recall the corresponding formulas:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \qquad slope = \frac{S_{xy}}{S_{xx}} \qquad \text{y} - \text{intercept} = \overline{Y} - slope \cdot \overline{X},$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \qquad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} \qquad S_{xx} = \sum xy - \frac{(\sum x)\sum(y)}{n}$$

12. Each score listed below comes from a sample with the indicated mean and standard deviation. Convert each one to a z-score and find the indicated probability (in percent). Note that drawing a picture will help to find the indicated probabilities (percentages).

a) X = 6 (mean 4, standard deviation 2),
$P(z \le z_0)$ **= P(z < 1) = 1 – 0.1587**

b) X = 9 (mean 6, standard deviation 1.5),
$P(z \le z_0)$ **= P(z < 2) = 1 – 0.0228**

c) X = 1.5 (mean 0, standard deviation 1),
$P(z \ge z_0)$ **= P(z > 1.5) = 0.0668**

d) X = 2 (mean 3, standard deviation 1),
$P(z \ge z_0)$ **= P(z > -1) = 1 – 0.1587**

d) If Z is a variable with mean 0 and standard deviation 1 (i.e. a "z-score"), then find
$P(-2 \le z \le 1)$ **= 1 – 0.1587 – 0.0228**

13. Consider the following sample data, selected at random from some population:

10, 8, 12, 10

a) What is your best guess for the unknown population mean?

10

b) Find the standard error for the sample mean.

0.8165

c) Find a 95% confidence interval for the unknown population mean.

From 7.4015436 to 12.598456

14. Using the GSS survey data to find the average number of hours that people watched TV in the US in 1996, we found that the descriptive statistics for the variable 'tvhours' as:

N = 1000, Mean = 2.96, Standard Deviation = 2.38

a) Find a 95% confidence interval for the average number of hours that all people in the US in 1996 watched TV.

From 2.812 to 3.107

b) If you used a larger sample (i.e. a sample with a larger N) would that improve your estimate for the population mean?

Yes, a larger sample would usually result in a narrower confidence interval

c) Now find a 90% confidence interval instead of a 95% one, and then a 99% confidence interval (this is a possible extra credit question).

The 90% interval is smaller, the 99% interval is wider

15. When using StatCrunch for a linear regression analysis of pre-test versus post-test scores, it shows the output:

Simple linear regression results:

Dependent Variable: PostTest
Independent Variable: PreTest
PostTest = -1.1083333 + 1.975 PreTest
Sample size: 15
R (correlation coefficient) = 0.9859
R-sq = 0.97191656
Estimate of error standard deviation: 1.5580642

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	-1.1083333	1.633227	≠ 0	13	-0.67861557	0.5093
Slope	1.975	0.09311215	≠ 0	13	21.210981	<0.0001

a) Find the exact equation of the least-square regression line

Y = 1.975 x – 1.108

b) What is the correlation coefficient r

R = 0.9719

c) Predict the post-test score of someone with a pretest score of 38.

73.942

d) Do you think your prediction is accurate? Justify your answer

Accurate because the correlation coefficient is close to 1

16. Using StatCrunch, we have computed a contingency table for *degree category* versus *income category* variables from our GSS survey. We have also computed a particular statistics for this contingency table, as shown below.

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	96	349.11243	<0.0001

Are the two variables related or independent of one another. ?

P is very small, so DEPENDENT

Strictly speaking, you also need **what** additional information to make sure your test applies
You need to make sure all expected values are 5 or more.

17. The lifetimes (in months) of ten automobile batteries of a particular brand are:

    22 17 20 21 17 23

    Estimate the mean lifetime for all batteries, using a 95% confidence interval.

From 17.345114 to 22.654886

18. A large supermarket chain sells longhorn cheese in one-pound (= *16* ounces) packages. As a city inspector you weigh *100* randomly selected packages of cheese and note that the sample mean is *15.6* ounces, with a standard deviation of *2.0* ounces. You therefore suspect that the chain is miss-labeling the cheese and that the actual weight of a package is less than *16* ounces. Use this data to test your suspicion against the null hypothesis that the average weight of a package is *16* ounces. Use $\alpha = 0.05$.

Z = 2 so that p = 2*P(z > 2) = 2*0.0228 < 0.05, so REJECT H_0, i.e. suspicion is confirmed

19. The caffeine content of a random sample of *81* cups of black coffee dispensed by a new machine is measured. The mean and standard deviation for the sample are *110 mg* and *5.0* mg, respectively. The manufacturer of the machine claims that the average caffeine content per cup is *109 mg*. Do you believe that the manufacturer's claim is valid or invalid?

Z = 1.8, which results in p 0.05, so INCONCLUSIVE (i.e. manufacturer could be right, I don't know)

20. A test was conducted to determine the length of time required for a student to read a specified amount of material while a low-level music was playing to see if students were distracted by the noise. All students were instructed to read at the maximum speed at which they could still comprehend the material. Fourteen students took the test, with the following results (in minutes):

    25, 18, 27, 29, 20, 19, 25, 24, 32, 21, 24, 20, 24, 28

    The average reading time for students in a quiet environment is 22 minutes. Use an appropriate statistical test to determine whether noise is indeed distracting students.

    Hint: Using the above numbers we find that the sample mean is 24 minutes, while the sample standard deviation is 4.1.

T = 1.825, T_a = 2.160 so INCLONCLUSIVE

21. To test the research hypothesis that teacher expectation can improve student performance, two groups of students were compared. Teachers of the experimental group were told that their students would show large IQ gains during the test semester, while teachers of the control group were told nothing. At the end of the semester, IQ change scores were calculated with the following results:

	Mean	**Standard Deviation**	**Sample Size**
Experimental	16.5	14.2	49
Control	7.0	13.1	64

    Find a 95% confidence interval for the difference of the average population scores.

WE DID NOT COVER CONFIDENCE INTERVALS FOR DIFFERENCE OF MEANS

22. Researchers are comparing the attitudes of male college students toward their fathers with their attitudes toward their mothers. 100 subjects were selected for study and they described their attitude on a scale from 1 (poor) to 10 (excellent). The data for the samples is summarized as follows:

	Sample Size	**Mean**	**Standard Deviation**

Attitude toward Father	100	8.4	2.2
Attitude toward mother	100	7.8	3.1

Test whether the male students' attitudes toward their fathers differ from their attitudes towards their mothers, on average.

Z = 1.578 which results in p > 0.05 so this test is inconclusive

22'. The Ford Motor Company claims that the average Miles per Gallon (MPG) rating of all cars in their product line is 24 MPG, which is the minimum required by law. You, as EPA commissioner of New Jersey, have doubts about that figure. Therefore you select a random sample of 398 cars and measure their MPG. Then you use StatCrunch to conduct a test for the Mean of 24. StatCrunch comes up with the following output:

Hypothesis test results:

μ : mean of Variable

H_0 : μ = 24

H_A : μ ≠ 24

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Miles Per Gallon	23.517588	0.39186713	397	-1.2310603	0.219

Would you contest the assertion made by the Ford Motor Company or not? Use $\alpha = 0.05$, as usual. Make sure to state the null and alternative hypothesis that StatCrunch is making when conducting the test.

P = 0.219 > 0.05 so the test is inconclusive. I would not contest the claim.

23. Using the 1996 GSS survey, we compare the income levels of participants with their party affiliation to decide whether the two variables are related or not. In other words, we are interested in knowing whether, as a general rule, people with higher income levels tend to be republican, while people with less income tended to be democrat. Using StatCrunch to conduct a "Chi-Square" test, we obtain the following results:

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	168	194.85733	0.0764

What is your conclusion?

P = 0.0765 > 0.05, so the test is inconclusive (again).

24. During a psychological convention someone claimed that the average American adult watches approximately 3 hours of TV per week. You want to dispute that claim, so you use our GSS survey to test the null hypothesis of the average number of hours watching TV is equal to 3. StatCrunch comes up with the following output:

Hypothesis test results:

μ : mean of Variable

H_0 : μ = 3

H_A : μ ≠ 3

Std. Dev. not specified

Variable	n	Sample Mean	Std. Err.	Z-Stat	P-value
HOURS PER DAY WATCHING TV	1324	2.981873	0.07308211	-0.24803454	0.8041

Would you contest the assertion made at the convention or not?

No. Since p > 0.05 the test is inconclusive, which means I would not contest the assertion.

25. We are interested in which person people would have voted for, if they had voted, in 2004. In particular, we want to know if the majority would have voted for or against Georg Bush. We use our GSS data and define a proportion variable to mean 1 if a person would have voted for Bush, and 0 if not. With the help of StatCrunch we conduct a test for propoprion Pi = 0.5 and find the following output:

Hypothesis test results:
Outcomes in : WOULD HAVE VOTED FOR IN 2004 = Bush
Success : 1
p : proportion of successes
H_0 : p = 0.5
H_A : p ≠ 0.5

Variable	Count	Total	Sample Prop.	Std. Err.	Z-Stat	P-value
WOULD HAVE VOTED FOR IN 2004 = Bush	195	629	0.3100159	0.019936306	-9.529554	<0.0001

What is your conclusion?

P is small, so that we reject H_0. Thus, there is a majority one way or another, and clearly it goes against Bush.

26. For the same setup as in the previous question, we have used StatCrunch to compute the confidence interval for Pi, the probability of success. We find:

95% confidence interval results:
Outcomes in : WOULD HAVE VOTED FOR IN 2004 = Bush
Success : 1
p : proportion of successes
Method: Standard-Wald

Variable	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
WOULD HAVE VOTED FOR IN 2004 = Bush	195	629	0.3100159	0.018441074	0.27387205	0.34615973

What does this mean and how does it connect to your result in the previous question.

This means that the unknown prob. of success is between 27% and 34%, which does not include 50%, so again a majority was against Bush.

27. We suspect a coin to be not fair. Suppose we flip that coin 200 times and we come up with 94 heads, 106 tails. Based on this evidence, do you think the coin is unfair?

We do a test about the proportion.

Z = (94/200 - 0.5) / sqrt(94/200 * 106/200 / 200) = -0.85, which would result in a small value of p < 0.05, so the test is inconclusive, which means the coin just might be fair.

28. We conduct a survey to ask people if they are for or against Hydraulic fracturing in a particular county. The survey asked 265 people, 116 came out for the practice, 149 against. Compute a 95% confidence interval for the probability of voting for hydraulic fracturing. If you were to advise a congress person to represent her district accurately, would you advise her to vote for or against the practice?

Consider S to come out FOR. Then the confidence interval is from 0.378 to 0.497. That interval does not include 0.5 (barely) so I would conclude that the majority is against