MATH 1203 – Practice Exam 2

This is a practice exam only. The actual exam may differ from this practice exam.

1. Please state the Central Limit Theorem as discussed in class.

Select samples of size n from a population with unknown distribution, mean μ and std dev. σ. Then the sample means $\bar{x}$ are normal with mean μ and std. dev $\sigma/\sqrt{N}$.

2. Please state, in your own words, what the following terms mean
   - Contingency Table
   - Chi-Square Test
   - Least-Square Regression
   - Confidence Interval
   - p-value of a Chi-square Test
   - Correlation Coefficient r
   - Scatter Plot
   - Least Square Regression line
   - P(event) = 0
   - P(z < 1), where z has a standard normal distribution
   - Degree of freedom df for a t-distribution
   - When to use a t distribution and when to use a normal distribution
   - P(x > 55) if x is N(50, 12)
   - Standard error

$\}$ see note ←

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

F
- If the p-value of a Chi-Square test is close to one, the association between two variables is strong. If r is close to 1 or $p < 0.05$

T
- Both the p-value of a Chi-Square test and the correlation coefficient r tell you whether two variables are related, but the correlation coefficient r carries even more information. r tells you strength and direction of relation

T
- A Chi-Square test is appropriate for categorical variables, a regression analysis is appropriate for two numeric variables.

T
- The expected value in a cell of a contingency table tells you how many items would fall in that cell if the two variables were independent of one another.

F
- If an expected value in any cell of a contingency table is less than 5, then the two variables are dependent. it means the chi-square test is not applicable

F
- Suppose you compute the equation of a least-square regression line as y = -2 x + 3 and the correlation coefficient r = 0.8, could that be possible? if slope is neg, r should be negative

F
- If r = 0.8, it means that two variables are strongly related in such a way that as x gets larger, the corresponding y gets ~~smaller.~~ larger

F
- P(z < 2) = 0.02211 F (= 0.9773)

F
- If X is N(10, 2) and X = 11, then the corresponding z-value is 2.1. $z = \frac{x-\mu}{\sigma} = \frac{11-10}{2} = 1/2 \neq 2.1$

F
- If X is N(95, 10) then P(X > 105) = 0.21~~3~~41 0.1586

T
- A 95% confidence interval means that you are 95% certain that the true population mean is contained in the computed interval.

F
- A 99% confidence interval is ~~*smaller*~~ wider than a 90% confidence interval.

T
- If a sample of size 100 has mean 123 and std. dev. 12, then the standard error is 1.2 $12/\sqrt{100} = \frac{12}{10} = 1.2$

4. Compute the following probabilities:

total possible HH, HT, TH, TT

- In tossing one coin twice, find P(HH) or P(exactly one head) or P(no head) or P(at least one head)

P(HH)= 1/4, P(one H)= 1/2, P(no head)= 1/4, P(at least one H)= 3/4

- In throwing two dice, find P(sum is 4) or P(sum = 1) or P(sum is 4 or more)

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

P(sum=4)= 3/36

P(sum=1)= 0

P(sum ≥4)= $\frac{36-3}{36} = \frac{33}{36}$

- In drawing one card randomly from a standard 52-card deck, find P(card is Ace)

there are 4 aces in a deck ⇒ P(ace)= 4/52

5. A (hypothetical) frequency distribution for the age of people in a survey, the categories have the following probabilities:

Category	Probability
0 – 18	0.15
19-40	0.25
41-65	0.3
65 and older	0.3
	1.0

- One number is missing – what is that number? 0.3

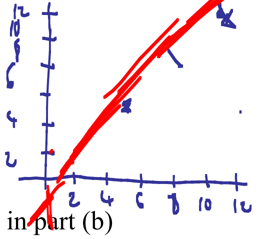
- What is the chance that a randomly selected person is 40 years or younger? P(...)= 0.25+0.15 = 0.4

6. Please consider the following data:

x **Father's schooling (years):**	8, 11, 5, 12
y **Respondent's schooling (years):**	8, 12, 5, 15

a) find the mean for both variables $\bar{x} = (8+11+5+12)/4 = 36/4 = 9$, $\bar{y} = (8+12+5+15)/4 = 40/4 = 10$

b) create a scatter plot representing this data



c) draw a best-fit line through the scatter plot in part (b)

d) find the exact equation of the least-square regression line $y = 1.367x - 2.3$ (used phone)

e) compute the correlation coefficient r

r=0.9829

f) predict the highest year of schooling for someone who's father completed 14 years of school.

$y = 1.367 \cdot 14 - 2.3 = 16.838$

Recall the corresponding formulas (or use StatCrunch)

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

7. The following scores were obtained as part of a sample with mean 10 and standard deviation 2. For each score, find the appropriate z-score: X = 10, X = 14, X = 6, X = -1. Then, for each z-score found, find the probabilities of obtaining a score *less than or equal* to the computed z-score. Note: in mathematical notation this means that we want to find $P(z \le z_0)$, where z_0 is the computed z-score.

$z = \frac{X - \mu}{\sigma}$

$\frac{10-10}{2} = 0$ $\qquad P(z \le 0) = 0.5$

$\frac{14-10}{2} = 2$ $\qquad P(z \le 2) = 0.9773$

$\frac{6-10}{2} = -2$ $\qquad P(z \le -2) = 0.0227$

$\frac{-1-11}{2} = -6$ $\qquad P(z \le -6) \approx 0$

8. Each score listed below comes from a sample with the indicated mean and standard deviation. Find the indicated probability (in percent).

- X is normal with mean 3, standard deviation 1.5, find $P(x \le 6)$ $\approx \underline{0.9773}$
- X is normal with mean 3, standard deviation 3, find $P(x \ge 9)$ $= \underline{0.0227}$
- X is normal with mean 0, standard deviation 2, find $P(1 < x < 2)$ $= P(X \le 2) - P(X \le 1) =$
  $= \underline{0.8413 - 0.6914}$
- X is normal with mean 3, standard deviation 1, find $P(x \ge 2)$
  $= \underline{0.8413}$

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

12, 16, 5, 19

a) What is your best guess for the unknown population mean? The sample mean $(12+16+5+19)/4 = 52/4 = \underline{13}$

b) Find the standard error for the sample mean.

std error $= S/\sqrt{N} = \frac{6.05}{\sqrt{4}} = \underline{3.025}$

c) Do you need to use the t distribution to find the multiplier?

Yes because it's a small sample $(n < 30)$

d) Find a 95% confidence interval for the unknown population mean. multiplier = $P(t \geq t_0) = 0.025$, $df = n-1 = 3$

$t_0 = 3.18$ ⇒ $\boxed{13 \pm 9.601}$ done

e) Find a 99% confidence interval for the population mean. Explain why this interval differs from the previous one.

multiplier: $P(t \geq t_0) = 0.005$, $df = 3$ ⇒ mult is 5.84 ⇒ 13 ± 17.668

f) If you were to compute a 90% confidence interval, would it be wider or narrower than the previous two?

99% wants to more sure ⇒ cover more possible numbers ⇒ widder

10. Find the following probabilities, assuming z is N(0, 1).

$P(z < 1.1)$ = 0.8643

$P(z > -1.2)$ = 0.8849

$P(z > 1.3)$ 0.0968

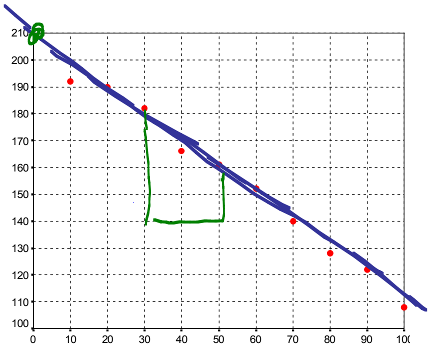
$P(z < -1.6)$ = 0.0548

$P(1.2 < z < 2.1)$ = $P(z \leq 2.1) - P(z \leq 1.2) = 0.9821 - 0.8849$

$P(-2.1 < z < -1.2)$ = $P(z \leq -1.2) - P(z \leq -2.1) = 0.1150 - 0.0179$

$P(-1.2 < z < 2.1)$ = $P(z \leq 2.1) - P(z \leq -1.2) = 0.9821 - 0.1150$

11. 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-int: ≈210, slope $(180-160)/(30-50) = \frac{20}{-20} \approx -1$ ⇒ $y = -1 \cdot x + 210$ (approx.)

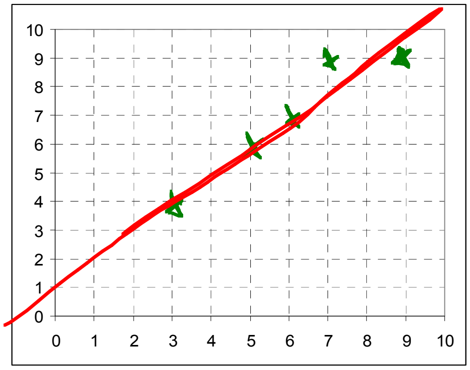
c) Look at the data and your line and estimate whether *r* would be close to -1, close to 0, or close to 1

r is very close to -1

12. 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

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



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

  $y = 0.9x + 1.6$ (used phone)

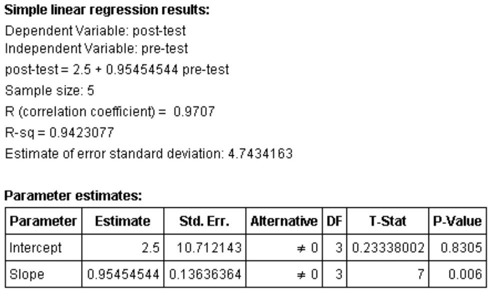
- Compute the correlation coefficient r

  $r = 0.9487$

- Predict the “after lecture” score for a “before lecture” score of 8.

  $y = 0.9 \cdot 8 + 1.6 = 8.8$

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



Simple linear regression results:

Dependent Variable: post-test
Independent Variable: pre-test
post-test = 2.5 + 0.95454544 pre-test
Sample size: 5
R (correlation coefficient) = 0.9707
R-sq = 0.9423077
Estimate of error standard deviation: 4.7434163

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	2.5	10.712143	≠ 0	3	0.23338002	0.8305
Slope	0.95454544	0.13636364	≠ 0	3	7	0.006

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

$y = 0.9545x + 2.5$

b) What is the correlation coefficient, and what does it mean

$r = 0.9707$, data is close to line

c) Predict the post-test score of someone with a pretest score of 77.

$y = 0.9545 \cdot 77 + 2.5 = 75.99$

d) Do you think your prediction is accurate? Justify your answer using the correlation coefficient

r is close to 1, so I think it's an accurate prediction

14. The table below shows a contingency table for the variables “DEGREE” by “RACE”. Each cell lists three numbers: the count, the row, and the column percentage.

DEGREE					
	HIGHEST DEGREE	70.5%	23.0%	6.5%	100.0%
	RESPONDENT	13.5%	25.8%	19.2%	15.5%
	HIGHEST DEGREE	81.9%	13.6%	4.5%	100.0%
	RESPONDENT	54.7%	53.4%	47.0%	54.1%
	HIGHEST DEGREE	85.0%	12.8%	2.1%	100.0%
	RESPONDENT	6.8%	6.0%	2.6%	6.5%
	HIGHEST DEGREE	83.9%	9.1%	7.0%	100.0%
	RESPONDENT	16.8%	10.8%	21.9%	16.3%
	HIGHEST DEGREE	86.6%	7.1%	6.3%	100.0%
	RESPONDENT	8.3%	4.0%	9.3%	7.7%
	HIGHEST DEGREE	81.0%	13.8%	5.2%	100.0%
	RESPONDENT	100.0%	100.0%	100.0%	100.0%

a) Which of the two variables is independent, which is the dependent variable? Race is indep., Degree is dep.

b) Which number is the count, the row, and the column percentage

top = count, middle = row, bottom = col

c) Compute the *expected value* for the cell "Whites with a High School degree"

count = 1283, exp. value = 1567 · 2347 / 2897 = 1269.5

d) How many *Blacks* have a high school degree, in percent?

col %: answer = 53.4%

e) How many people with a college degree, ~~graduate~~ or bachelor, are *White*, in percent?

16.8 + 8.3 = 25.1%

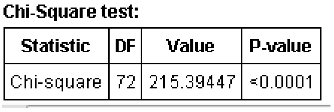
f) How many *Blacks* have **at most** a junior college degree, in percent?

100 − 10.8 − 4.0 = 85.2%

15. Consider the contingency table for religious preference versus political opinion, using our GSS survey below.
    a) Compute the row percentage in the "Liberal and Catholic" cell, as well as column percentage and expected value.

	PROTES					
Opinion						

b) Using *StatCrunch*, we conducted a Chi-Square test with the output as follows:



Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	72	215.39447	<0.0001

What is your conclusion?

p < 0.05 ⇒ variables are related!

c) Why should you compute and double-check all expected values in that table before finalizing your conclusion?

If an expected value is less than 5, test is not valid

16. To investigate whether a relation exists between affiliation with a particular political party and the opinion on gun permits we used *StatCrunch* to create the following contingency table.

GUN PERMITS					

a) Based on that table, do you think there is strong evidence that the two variables associated, using common sense?

No, everything sees evenly distributed

b) Based on your analysis in part (a), what do you think might be the p-value of a Chi-Square test for this data?

$p > 0.05$

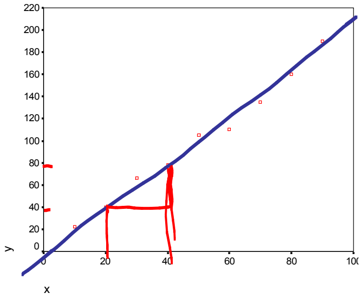
17. Use the GSS survey data to find the average number of siblings for people in the US in 1996 with reasonable accuracy. Note: Using *StatCrunch* we found that the descriptive statistics for the variable 'sibs' is as follows:

N	2897 ← large sample
Mean:	3.86
Standard Deviation:	3.52

std err $= 3.52/\sqrt{2897} = 0.065 \Rightarrow 1.96 \cdot 0.065 = 0.1281$

$\Rightarrow 3.84 \pm 0.1281$ is 95% conf. interval

18. 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

$y = \frac{78-40}{40-20}x - 5 \approx 1.9x - 5$

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

$r \approx 0.9$ close to 1

19. Suppose you were asked to compute a *95%* confidence interval. The resulting interval, however, turned out to be too large to be of use to your client. What could you do to achieve a smaller confidence interval?

Redo survey with a larger sample, or compute 90%-conf. interval, which will be smaller.

20. The lifetimes (in years) of ten automobile batteries of a particular brand are:

2.4	1.9	2.0	2.1	1.8
2.3	2.1	2.3	1.7	2.0

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

sample mean: 2.06

standard error: 0.0718

DF = 9,

From 1.897 to 2.222

21. A test was conducted to determine the length of time required for a student to read a specified amount of material. All students were instructed to read at the maximum speed at which they could still comprehend the material. Sixteen students took the test, with the following results (in minutes):

25, 18, 27, 29, 20, 19, 25, 24, 32, 21, 24, 19, 23, 28, 31, 22

Estimate the mean length of time required for all students to read the material, using a 95% confidence interval.

sample mean: 24.1875

std.error: 1.0809

from 21.88 to 26.49

22. The caffeine content of a random sample of *90* cups of black coffee dispensed by a new machine is measured. The mean and standard deviation for the sample are *110 mg* and *6.1* mg, respectively.

a) Compute a *90% confidence interval* for the true population mean caffeine content per cup dispensed by the machine.

$$\bar{X} \pm 1.645 \cdot \frac{S}{\sqrt{N}} = 110 \pm 1.058$$

b) If you would compute a *99% confidence interval* for the true population, would it be wider or narrower than the 90% confidence interval ? (You do **not** actually have to compute this interval to answer the question).

wider

c) Another person selected a random sample of 900 instead of 90 cups, and the mean and standard deviation of this larger sample turned out to be 110mg and 6.1mg as well. That person uses her data to compute a 90% confidence interval. Would the 90% confidence interval for the larger sample size be wider or narrower than the 90% confidence interval for the smaller sample size ?

90% - interval for larger sample would be narrower! (smaller)