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.

If x is a variable with some unknown distribution with mean μ and std. dev. σ, and we select samples of size N and compute their mean $\bar{x}$, then $\bar{x}$ is 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 Intervals
   - 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

See lecture notes

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.
   - 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.
   - 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.
   - 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? No
   - 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 If X is N(10, 2) and X = 11, then the corresponding z-value is 2.1.
   - F If X is N(95, 10) then P(X > 105) = 0.21341
   - 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* than a 90% confidence interval.

$\frac{105-95}{10} = 1$, $P(z>1) = 0.1587$

4. Compute the following probabilities:
   - In tossing one coin twice, find P(HH) or P(exactly one head) or P(no head) or P(at least one head). $= 3/4$

All outcomes: TT, TH, HT, HH $\Rightarrow$ $P(HH) = 1/4$, $P(\text{ex. one H}) = 2/4$, $P(\text{no head}) = 1/4$

   - 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

$3/36$ $\quad$ 0 $\quad$ $33/36$

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

- One number is missing – what is that number? 0.3 (numbers must add to 1)
- What is the chance that a randomly selected person is 40 years or younger?

0.15 + 0.25 = 0.4

6. Please consider the following data:

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

$\Sigma x^2 = 354$

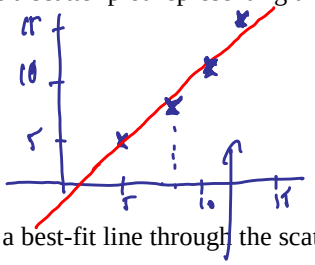
$\Sigma y^2 = 458$

$\Sigma xy = 401$

a) find the mean for both variables

$\bar{X} = 36/4 = 9$, $\bar{Y} = 40/4 = 10$

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

$\Sigma x^2 = 354$, $\Sigma y^2 = 458$, $\Sigma xy = 401$

$S_{xx} = 354 - \frac{36^2}{4} = 30$

$S_{yy} = 458 - \frac{40^2}{4} = 58$

$S_{xy} = 401 - \frac{36 \cdot 40}{4} = 41$

$m = 41/30 \quad b = -2.3$

$\Rightarrow \boxed{y = 1.36x - 2.3}$

e) compute Pearson's r

$r = \frac{41}{\sqrt{30 \cdot 58}} = 0.9929$

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

$y = 1.36 \cdot 14 - 2.3 = 16.74$

Recall the corresponding formulas:

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

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, use the table at the end to 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.

$\frac{10-10}{2} = 0 = z$, $P(z<0) = 0.5$

$\frac{14-10}{2} = 2 = z$, $P(z<2) = 1-.0228$

$\frac{6-10}{2} = -2 = z$, $P(z<-2) = 0.0228$

$\frac{-1-10}{2} = -5.5 = z$, $P(z<-5.5) \approx 0$ (can't lookup in table)

8. 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).

- X is normal with mean 3, standard deviation 1.5, find $P(x \le 6)$

$\frac{6-3}{1.5} = \frac{3}{1.5} = 2$ $P(z<2) = 1-0.0228$

- X is normal with mean 3, standard deviation 3, find $P(x \ge 9)$

$\frac{9-3}{3} = \frac{6}{3} = 2$ $P(z>2) = 0.0228$

- X is normal with mean 0, standard deviation 2, find $P(1 < x < 2)$ $= P(0.5 < x < 1) = 0.3085 - 0.1587$

0.3085 0.1587

- X is normal with mean 3, standard deviation 1, find $P(x \ge 2)$

$\frac{2-3}{1} = -1$ $P(z>-1) = 1-0.1587$

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?

$\bar{X} = 13$

b) Find the standard error for the sample mean.

$S = 6.05 \Rightarrow$ standard error $S/\sqrt{n} = \frac{6.05}{2} = 3.025$

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

$13 \pm 1.96 \cdot 3.025 = 13 \pm 5.929 = \begin{cases} 13+5.929 = 18.929 \\ 13-5.929 = 7.071 \end{cases}$

From 7.071 to 18.929

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

$13 \pm 2.54 \cdot 3.025 = 13 \pm 7.6935$ From 5.3165 to 20.6835

more certainty requires more choices i.e. wider interval

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

narrower

10. The table on page 592 in our text book can be used to compute probabilities for a variable z, assumed to have the Standard Normal Distribution N(0, 1). Use that table to find the following probabilities and shade the parts in the probability distribution that corresponds to the probability you computed.

$P(z < 1.1) = 1 - 0.1357$ (0.1357)

$P(z > -1.2) = 1 - 0.1151$ (0.1151)

$P(z > 1.3) = 0.0968$ (0.0968)

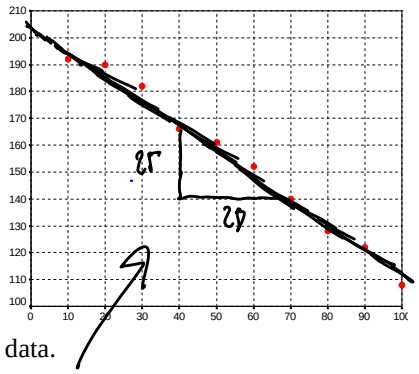
$P(z < -1.6) = 0.0548$ (0.0548)

$P(1.2 < z < 2.1) = 0.1151 - 0.0179$ (0.1151, 0.0179)

$P(-2.1 < z < -1.2) = 0.1151 - 0.0179$ (0.0179, 0.1151)

$P(-1.2 < z < 2.1) = 1 - (0.0179 + 0.1151)$ (0.1151, 0.0179)

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

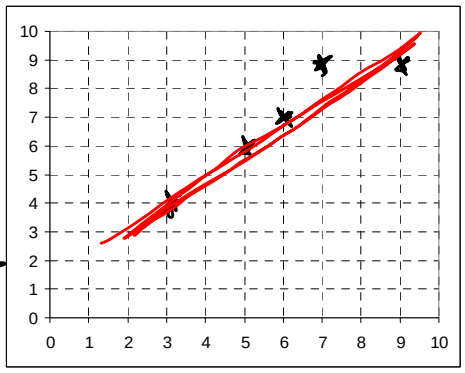
y-intercept ≈ 205, slope $\approx \frac{-25}{21} \approx -0.99$

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

$r \approx -0.9$ because data lines up nicely and slope is negative

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

X **Before lecture:** 5, 6, 7, 9, 3
Y **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

$S_{xx} = 20,\ S_{yy} = 18,\ S_{xy} = 18,\ y = 0.9x + 1.65$

- Compute the correlation coefficient r (use back page for computation but show r here)

0.947

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

$y = 0.9 \cdot 8 + 1.65 = 8.85$

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

$y = 0.9545 \cdot x + 2.5$

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

$r = 0.9707$ (strong positive relation between X and y)

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

$y = 0.9545 \cdot 77 + 2.5 = 75.3$

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

Yes, because r is close to 1.

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.

RS HIGHEST DEGREE * RACE OF RESPONDENT Crosstabulation

			RACE OF RESPONDENT			
			WHITE	BLACK	OTHER	Total
RS HIGHEST DEGREE	LT HIGH SCHOOL	Count	316	103	29	448
		% within RS HIGHEST DEGREE	70.5%	23.0%	6.5%	100.0%
		% within RACE OF RESPONDENT	13.5%	25.8%	19.2%	15.5%
	HIGH SCHOOL	Count	1283	213	71	1567
		% within RS HIGHEST DEGREE	81.9%	13.6%	4.5%	100.0%
		% within RACE OF RESPONDENT	54.7%	53.4%	47.0%	54.1%
	JUNIOR COLLEGE	Count	159	24	4	187
		% within RS HIGHEST DEGREE	85.0%	12.8%	2.1%	100.0%
		% within RACE OF RESPONDENT	6.8%	6.0%	2.6%	6.5%
	BACHELOR	Count	395	43	33	471
		% within RS HIGHEST DEGREE	83.9%	9.1%	7.0%	100.0%
		% within RACE OF RESPONDENT	16.8%	10.8%	21.9%	16.3%
	GRADUATE	Count	194	16	14	224
		% within RS HIGHEST DEGREE	86.6%	7.1%	6.3%	100.0%
		% within RACE OF RESPONDENT	8.3%	4.0%	9.3%	7.7%
Total		Count	2347	399	151	2897
		% within RS HIGHEST DEGREE	81.0%	13.8%	5.2%	100.0%
		% within RACE OF RESPONDENT	100.0%	100.0%	100.0%	100.0%

a) Which of the two variables is independent, which is the dependent variable?

race is independent, degree is dependent

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"

$\frac{1567 \cdot 2347}{2897} = 1269.5$

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

53.4% (col %)

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

6.8% + 16.9% + 9.3% = 31.9%

f) How many *Blacks* have **at most** a junior college degree, in percent? 25.9% + 53.4% + 6% = 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.

Political Opinion * RS RELIGIOUS PREFERENCE Crosstabulation

Count

		RS RELIGIOUS PREFERENCE					
		PROTESTANT	CATHOLIC	JEWISH	NONE	OTHER	Total
Political Opinion	Liberal	329	160	34	132	40	695
	Moderate	588	274	20	102	60	1044
	Conservative	658	221	14	76	31	1000
Total		1575	655	68	310	131	2739

row % = $\frac{160}{695}$ = 23% col % = $\frac{160}{655}$ = 24.4% expected value = $\frac{695 \cdot 655}{2739}$ = 166.2

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?

$p < 0.05 \Rightarrow$ there is an association between race & degree

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

to check that none is less than 5.

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.

FAVOR OR OPPOSE GUN PERMITS * Party Affiliation Crosstabulation

% within FAVOR OR OPPOSE GUN PERMITS

		Party Affiliation				
		Democrat	Independent	Republican	Other	Total
FAVOR OR OPPOSE GUN PERMITS	FAVOR	35.7%	36.5%	26.5%	1.3%	100.0%
	OPPOSE	23.4%	39.5%	34.7%	2.4%	100.0%
Total		33.5%	37.0%	28.0%	1.5%	100.0%

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

looks evenly distributed, so guess no association

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:

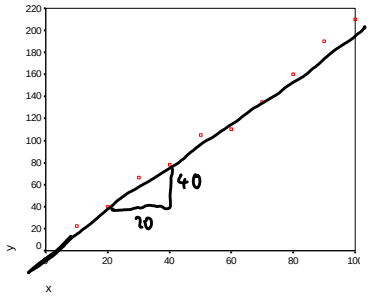
N	2897
Mean:	3.86
Standard Deviation:	3.52

choose 95% confidence interval.

standard error: $3.52/\sqrt{2897} = 0.065$, multiplier 1.96.

From $3.86 - 1.96 \cdot 0.065 = 3.86 - 0.13 = 3.73$ to $3.86 + 0.13 = 3.99$

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

$b = -10$, $m = \frac{40}{20} = 2$, $y = 2x - 10$

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

$r \approx 0.9$