Stat 1040
Final Exam Review Problems

1. In the game of chess, the first few moves play a very important role in determining the final outcome. Five different opening strategies are highly favored by chess experts. To determine whether one or more of these strategies is most preferred by grad masters in international competition, a random sample of 100 grand masters is taken, and each is asked which of the strategies he or she would prefer to employ. A summary of their responses is given below:

Strategy	A	B	C	D	E
Observed Frequency	17	27	22	15	19
Expected	20	20	20	20	20

Test the hypothesis that there is no preference between these strategies by grand masters in international competition.

Null: no preference; just chance variation

χ^2 test, df = 4

$$\chi^2 = \text{sum of } \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$= \frac{(17-20)^2}{20} + \frac{(27-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(19-20)^2}{20}$$

$$= 4.4$$

p-value: .35 or 35%

Don't reject, could be just chance.

2. Many scientists believe that alcoholism is linked to social isolation. One measure of social isolation is marital status, i.e., whether a person is married or not. To test the notion that alcoholics are socially isolated, 280 adults were randomly selected and each was classified as a diagnosed alcoholic, undiagnosed alcoholic, or nonalcoholic and categorized according to his or her marital status. A summary of the responses is shown in the table. Can you conclude that there is a relationship between the marital status and alcoholic classifications?

	Diagnosed	Undiagnosed	Nonalcoholic	
Married	21 (33)	37 (41)	58 (42)	116
Not Married	59 (47)	63 (59)	42 (58)	164
	80	100	100	280

Null: variables are independent

χ^2 test for independence; $df = (2-1)(3-1) = 2$

Expected frequencies

Married & Diagnosed: $\frac{80}{280} \times 116 = 33$

Married & Undiagnosed: $\frac{100}{280} \times 116 = 41$

$$\chi^2 = \frac{(21-33)^2}{33} + \frac{(37-41)^2}{41} + \frac{(58-42)^2}{42} + \frac{(59-47)^2}{47} + \frac{(63-59)^2}{59} + \frac{(42-58)^2}{58}$$

$\simeq 14.88$; p value $= .0006$ Reject null

3. A study reported in the New England Journal of Medicine on January 20, 2010 concerns a new treatment for *relapsing-remitting multiple sclerosis* (MS). More than 1,000 subjects participated in the study with about half being randomly assigned to the treatment group and the remainder to the control group. Subjects in the treatment group received daily doses of the drug *fingolimod*; those in the control group received a placebo. At the end of the trial the incidence of relapsing was about 57% lower in the treatment group than in the control group.

a. What kind of a study is this? Randomized, controlled, blind

b. What is a *placebo*, and what does the presence of a placebo in this study tell us about the study? A placebo resembles the treatment & is neutral. The subjects don't know which group.

c. Why is randomization important? How does it affect your conclusions with regards to the results of a study?

Randomization eliminates bias, the T group is like the C group as much as possible. Any differences will be a result of treatment.

4. I took measurements of my diastolic blood pressure each day for a week. The values I obtained were: 102, 95, 108, 103, 92, 97, and 117.

92, 95, 97, 102, 103, 108, 117

a) Compute the average and median of these values and compare them.

Median = 102 , AV = 102

b) Compute the SD of the 7 measurements.

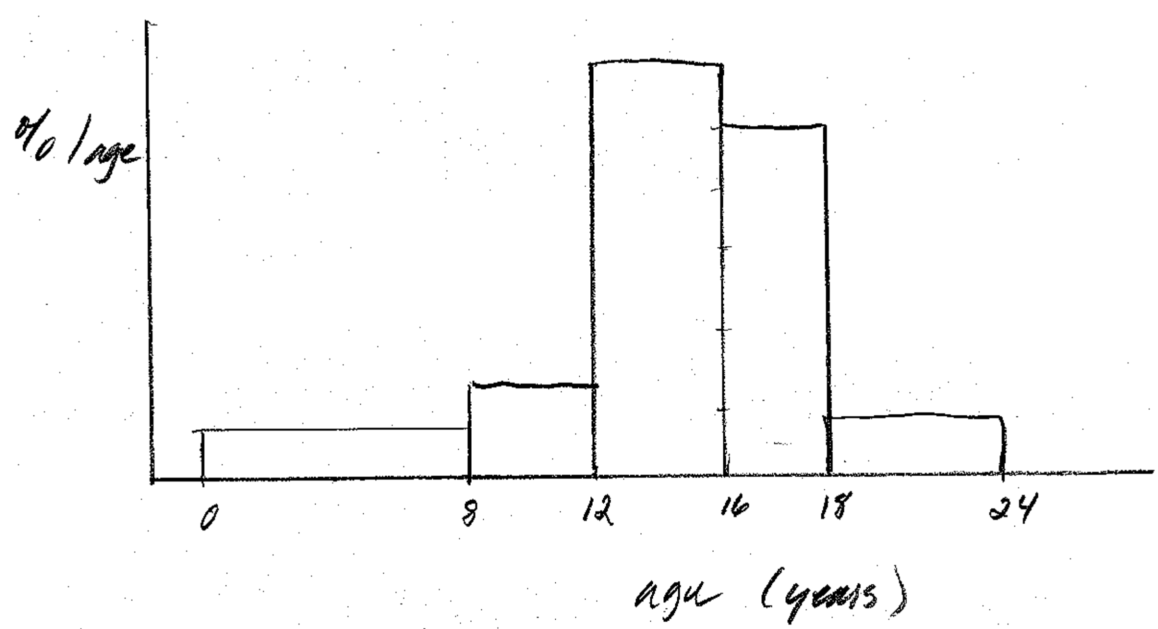
−10, −7, −5, 0, 1, 6, 15 $\sqrt{\frac{100+49+25+1+36+225}{7}} = 7.89$

c) My wife also measured her diastolic blood pressure once a day for a week. Her measurements averaged 64.3 with an SD of 8.2. If we combined our two sets of measurements, the SD of the combined group would be (circle one): about 8.2 , less than 8.2 , or quite a bit more than 8.2 ?

5. Data were obtained from the *Current Population Survey* on the educational levels of 250 persons living in California and aged between 25 and 55 years of age. In this study, educational level was defined as years of school (including college) completed. The table below summarizes the data (left-endpoint convention). Draw the histogram.

h = % / width

Educational level (years)	Number of Persons	Percent of Persons	Width of Class Interval	Height of Histogram Bar
0 - 8 years	25	10	8	1.25
8 - 12 years	25	10	4	2.5
12 - 16 years	120	48	4	12
16 - 18 years	50	20	2	10
18 - 24 years	30	12	6	2
Total	250			

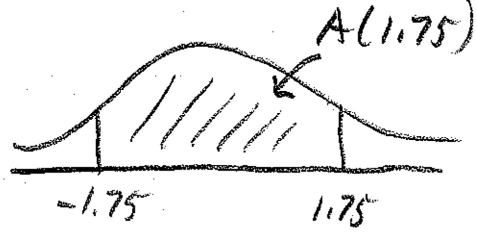


6. Many universities (including USU) require students to take the *Graduate Records Examination* (*GRE*) before being admitted to graduate school. The *Verbal* GRE is measured on a scale that runs from 200 to 800 points. Over the last 4 years the average Verbal GRE score was 460, the standard deviation was 120, and the histogram of Verbal GRE scores looked like the normal curve.

a. Approximately what percentage of students got 670 or higher on the Verbal GRE?

Normal approximation :

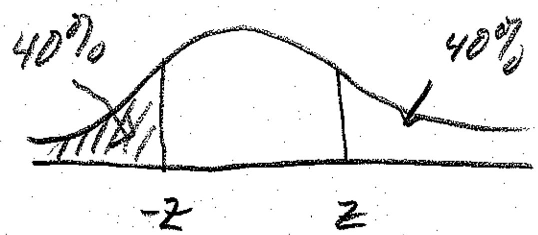
$$\frac{670-460}{120} = 1.75$$



$A(1.75) = 92\%$

4 %

b. The School of Graduate Studies at USU requires students to score above the 40th percentile on the Verbal GRE. What is the score that represents the 40th percentile?



$A(z) = 20\%$

$z \approx .53$

$$\frac{\text{score} - 460}{120} = .53$$

$$\text{score} = (120)(.53) + 460$$

$$\approx \boxed{524}$$

7. Systolic and diastolic blood pressure measurements tend to be quite highly correlated. For a sample of men aged 20—29 the following summary information was obtained:

X	Average systolic blood pressure = 120	SD = 12	r = 0.8
y	Average diastolic blood pressure = 85	SD = 10	

The scatter plot for these data was football shaped. Predict the average diastolic blood pressure for all the men whose systolic blood pressure was 135. Predict the diastolic blood pressure for a randomly chosen man whose blood pressure was 135.

Regression estimate

1. independent variable is systolic bp
2. 135
3. $\frac{135-120}{12} = 1.25$ (standard units)
4. multiply by r $(.8)(1.25) = 1$
5. multiply by SD_y $10 \times 1 = 10$
6. add AV_y $85 + 10 = \boxed{95}$

8. In a study, reading comprehension is tested for a large group of third grade students, once at the beginning of the school year and once at the end of the school year. During the school year, the students work on reading comprehension skills. The following results were obtained. The scatter plot for these data was football shaped.

X	Beginning of year AV = 75	SD = 15	
y	End of year AV = 80	SD = 17,	r = 0.6

Find the equation of the regression line for predicting the end-of-year score from the beginning-of-year score.

point : (AV_x, AV_y)

$(75, 80)$

slope $= r \times \frac{SD_y}{SD_x}$

$= (.6) \cdot \frac{17}{15}$

$= .68$

$y - y_1 = m(x - x_1)$

$y - 80 = .68(x - 75)$

$y = .68x + 29$

For those students who scored 65 on the beginning-of-year test, what percentage scored 75 or higher on the end-of-year test?

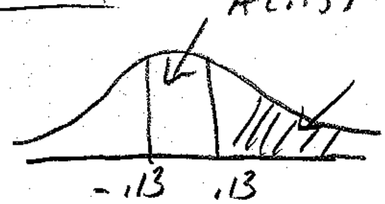
New AV = regression estimate $= (.68)65 + 29 \approx \boxed{73.2}$

New SD is r.m.s. error $= \sqrt{1 - r^2} \times SD_y$

$= (.8)(17) = \boxed{13.6}$

$A(.13) = 10\%$

$\frac{75 - 73.2}{13.6} = .13$



45%

9. For each state, a student finds data on the average amount of tax money spent per student and average student performance on a standardized test. The correlation coefficient is 0.7. This value is (choose one)

(a) ~~An accurate measure~~ of the association between these two variables.

(b) Misleadingly high.

(c) Misleadingly low.

Ecological correlation!

10. Gamers often use a 10-sided dice called a D-10 which has sides labeled 0, 1, 2, … , 9. Suppose I roll three D-10 dice.

a) How many possible outcomes are there? 10

b) What is the chance that *all three* dice show numbers that are 3 or higher? $\frac{7}{10} \times \frac{7}{10} \times \frac{7}{10}$

c) What is the chance that *none* of the dice shows numbers that are 3 or higher? $\frac{3}{10} \times \frac{3}{10} \times \frac{3}{10}$

d) What is the chance that *at least one* of the dice shows a number 3 or higher?

$$1 - \frac{3}{10} \times \frac{3}{10} \times \frac{3}{10}$$

11. A deck of cards has four suits (diamonds, hearts, clubs, and spades) with 13 cards in each suit: 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, king, and ace. I shuffle the deck of cards thoroughly so that the cards are in a completely random order. Then I draw two cards from the deck *without replacement.*

a) What is the chance that the first card is a diamond? $\frac{13}{52}$

b) What is the chance that the second card is a diamond given that the first card is a diamond? $\frac{12}{51}$

c) Are the events {*the first card is a diamond*} and {*the second card is a diamond*} (i) independent (ii) mutually exclusive (iii) neither independent nor mutually exclusive? Circle one options and briefly explain your answer?

d) Are the events {*the* first *card is a diamond*} and {*the* first *card is a club*} (i) independent (ii) mutually exclusive (iii) neither independent nor mutually exclusive? Circle one option and briefly explain your answer?

12. Gamers often use a 10-sided dice called a D-10 which has sides labeled 0, 1, 2, …, 9. Suppose I roll a D-10 die 100 times.

a) Construct a box model for this experiment. $\boxed{0,1,2,3,4,5,6,7,8,9} \rightarrow$ Draw 100 times

b) Compute the expected value (EV) and the SE for the *sum* of the 100 rolls of the D-10.

EV for sum = Box AV × (100) = (4.5)(100) = 450

SE for sum = Box SD × $\sqrt{100}$ = (2.87)10 = 28.7

c) Construct a box model for counting the number of times the number 0 turns up.

$\boxed{9 \text{ 0s}, 1} \rightarrow$ Draw 100 times & consider the sum of draws.

d) Compute the expected value (EV) and the SE for the *number of times* 0 turns up in the 100 rolls of the D-10.

EV for sum = Box AV × (100) = $\frac{1}{10} \times 100 = 10$

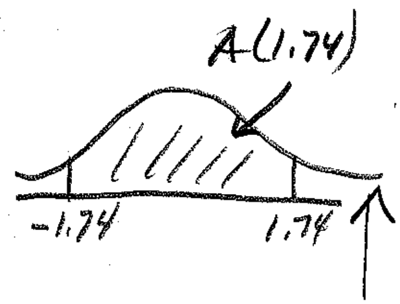
SE for sum = Box SD × $\sqrt{100}$

$= \sqrt{\frac{1}{10} \times \frac{9}{10}} \times 10 = \frac{3}{10} \times 10 = 3$

13. Gamers often use a 10-sided dice called a D-10 which has sides labeled 0, 1, 2, … , 9. Suppose I roll a D-10 die 100 times.

a) Compute the chance that the *sum of the draws* is more than 500.

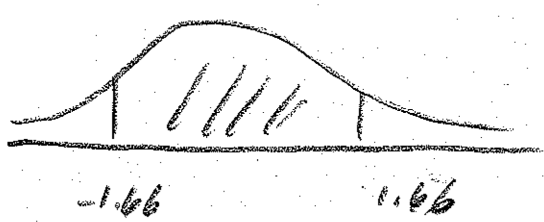
Normal ≈ $\frac{500-450}{28.7} = 1.74$



A(1.74) ≈ 92% 4%

b) What is the chance that the *number of times* 0 *turns up* is between 5 and 15?

$\frac{5-10}{3} = -1.66$, $\frac{15-10}{3} = +1.66$



A(1.66) ≈ 90%

14. One hundred draws will be made at random with replacement from the box:

[-2 , -1 , 0 , 1 , 5]

Box AV $= \frac{3}{5}$

Box SD $= 2.42$

The sum of the positive numbers will be around __60__, give or take __24__ or so.

EV for sum $= \frac{3}{5} \times 100 = 60$

SE for sum $=$ Box SD $\times \sqrt{100} = (2.42)\sqrt{100} = 24.2$

15. A roulette wheel has 38 slots numbered 0, 00, 1, 2, 3, … , 36. All 38 outcomes are equally likely. The *Topline* bet in roulette is to a bet on the numbers 0, 00, 1, 2, and 3. If you bet \$1 on the Topline and any one of the five numbers turns up you win \$6; otherwise you lose your \$1. Suppose that I bet \$1 on the Topline 380 times.

Chance of winning is $\frac{5}{38}$

a. Construct a box model for the total amount I win over the 380 games.

$\boxed{5\ \boxed{6}\text{s},\ 33\ \boxed{-1}\text{s}}$ → Draw 380 times with replacement + consider sum of draws.

b. What is the chance that I don't lose money over these 380 games? That is, what is the chance the *sum* of the amounts I win or lose over the 380 games is greater than or equal to \$0?

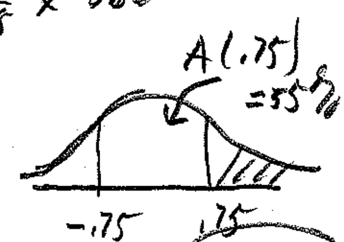
Box AV $= \frac{-3}{38}$, EV $= \frac{-3}{38} \times 380 = -30$

Box SD $= [5-(-1)]\sqrt{\frac{5}{38} \cdot \frac{33}{38}} = 2.03$

SE $= (2.03)\sqrt{380} = 39.6$

$\frac{0-(-30)}{39.6} = .75$

A(.75) = 55%



≈ 22½ %

c. Now construct a box model that counts the *number of times* I win in the 380 games.

$\boxed{5\text{-}1\text{s},\ 33\text{-}0\text{s}}$ → Draw 380 + consider sum.

Box AV $= \frac{5}{38}$, Box SD $= \sqrt{\frac{5}{38} \cdot \frac{33}{38}} = .34$

EV for sum $= \frac{5}{38} \times 380 = 50$, SE for sum $= (.34)\sqrt{380} = 6.63$

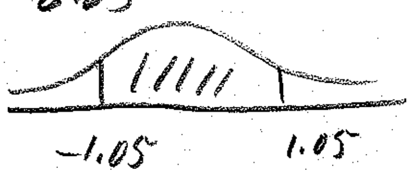
d. What is the chance I win 15% or more of the games?

15% of 380 = 57

$\frac{57-50}{6.63} = 1.05$

A(1.05) ≈ 71%

14½ %



(You could find EV for % & SE for %)

16. The population of Utah is estimated to be about 2,843,000 and the population of California is estimated to be about 36,962,000 which is about 13 times the population of Utah. The Current Population Survey (CPS) is supposed to measure the unemployment rates in all 50 states with equal accuracy. How large should the CPS sample be for California compared to the sample for Utah? *Briefly* explain your answer. The sample size should be about the same. Accuracy is determined by the size of sample. (sample size are both small compared to populations)

17. I take a simple random sample of 300 persons aged 25 and over in Cache Valley and I find that 75 of them have 4-year ("bachelors") college degrees. Construct a 95% confidence interval for the *percentage* of persons aged 25 and older in Cache Valley that has a 4-year college degree.

$\boxed{?\ 1s\ ,\ ?0s}$ → Draw 300 + consider % 1s drawn.

EV for sample % = population %.

$\frac{75}{300} \times 100\% = 25\%$

$$\text{SE for \% is} = \frac{\text{Box SD} \times \sqrt{300}}{300} \times 100\% \approx \frac{\sqrt{(.25)(.75)}\,\sqrt{300}}{300} \times 100\%$$

$= 2.5\%$

$25\% \pm 2(2.5\%)$

25% ± 5%

18. Modern laser altimeters are amazingly accurate and can measure the heights of mountains to accuracies within inches. Fifty independent measurements are made of the height of a famous mountain (you can guess which one!). The average of the measurements is 348,144 inches and the SD is 99 inches. Compute a 95% confidence interval for the true height of the mountain (at the time the measurements were made). Did you have to assume the measurements followed the normal curve? *Briefly* explain your answer.

$\boxed{\text{measurements}}$ → Draw 50 + consider the AV of draws.

EV for AV of draws = Box AV

$$\text{SE for AV of draws} = \frac{\text{Box SD} \times \sqrt{50}}{50} \approx \frac{99\sqrt{50}}{50} = 14$$

$348{,}144 \pm 2(14)$

348,144″ ± 28″

The measurements don't have to follow the normal curve but not be too outlier.

19. A high school teacher working at an inner-city high school is concerned about the time students spend working a after school jobs. She randomly selects 15 of her school's students and finds the average time spent working at after school jobs is 13.1 hours, with an SD of 11.5 hours. The national average is 10.7 hours. Assuming that the hours worked follow the normal curve, test to see whether the inner-city high school students work longer, on average, than those in the nation as a whole.

$\boxed{\text{hours worked / week}} \rightarrow$ Draw 15 & consider AV of draws

Null: Box AV = 10.7 hrs., Alt: Box AV > 10.7, t-test df = 14

EV for AV of draws = 10.7

SE for AV of draws = $\frac{\text{Box SD}\sqrt{15}}{15} \approx \frac{SD^{+}\sqrt{15}}{15} = \sqrt{\frac{15}{14}}\cdot\frac{(11.5)\sqrt{15}}{15}$

SE for AV ≈ 3.07

$\frac{13.1 - 10.7}{3.07} = .78$

p-value ≈ 22%. Don't reject; could be just do to chance.

20. Scores on the Verbal Graduate Record Examination Test were recorded. For 68 randomly selected women from a given population, the average score was 538.82 with an SD of 114.16. For 86 randomly selected men from the same population, the average score was 525.23 with and SD of 97.23. Test to see if the population average of the women is higher than that of the men for population. State the Null and Alternate Hypotheses and clearly state your conclusion.

$\boxed{\text{GRE scores}}$ Women $\rightarrow$ Draw 68 & consider AV of draws. $\boxed{\text{GRE scores}}$ Men $\rightarrow$ Draw 86 & consider AV of draws

Null: AV of women = AV of men; Alternate: AV of women > AV of men

Test statistic is the difference in AV of draws & follows z-curve.

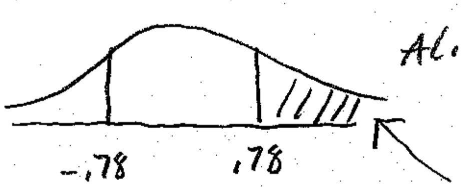
SE for AV of women = $\frac{\text{Box SD}\sqrt{68}}{68} \approx \frac{(114.16)\sqrt{68}}{68} = 13.84$

SE for AV of men = $\frac{\text{Box SD}\sqrt{86}}{86} \approx \frac{(97.23)\sqrt{86}}{86} = 10.48$

SE for diff = $\sqrt{(13.84)^2+(10.48)^2} = 17.36$

$\frac{538.82 - 525.23}{17.36} = .78$

A(.78) ≈ 56%



p-value = 22%.

Don't reject; could be chance variation.