Stat 2470, Final Exam, Spring 2015 Name KEY

Instructions: Show all work. State any formulas used. If you use the calculator, you should say which function you used, and what you entered into it, as well as any output. I can only give partial correct for incorrect answers if I have something to grade.

1. The contingency table below shows data displays data from a survey cross-categorizing people by marijuana usage and political party. Determine if there is a relationship between political party and marijuana usage, or if the two conditions are independent. Clearly state your null and alternative hypotheses and your conclusion in the context of the problem. (15 points)

Marijuana Usage Level

Political Views		Never	Rarely	Frequently
	Liberal	479	173	119
	Independent	172	45	85
	Conservative	214	47	15

χ^2-test

$\chi^2 = 64.654$

$p = 3.04 \times 10^{-13} < .05$

H_0: marijuana use and political affiliation are independent

H_a: they are dependent

reject H_0

they are dependent

2. Below is a table of data for the sex of kittens in a sample of litters of 4 kittens. The data collected is below. The number of girls born in each litter should be distributed binomially with $p = 0.5$. Conduct a hypothesis test to see if this data fits that model. (15 points)

Number of girls in the litter	*0*	*1*	*2*	*3*	*4*
Number of litters	13	30	94	50	18
	$1/16$	$4/16$	$6/16$	$4/16$	$1/16$
	12.8125	51.25	76.875	51.25	12.8125

= 205

$$\chi^2 = \sum \frac{(obs-exp)^2}{exp} = .0027439 + 8.001295 + 3.8148 + .0304878 + 2.1003$$

$= 13.94955\ldots$

$\chi^2cdf(13.9496, E99, 4) = .00745766 < .05$

this data does not appear to fit a binomial distribution

3. Conduct an appropriate hypothesis test comparing two different age groups and their stance duration (ms) to determine if older people have a shorter stance duration than younger people. Clearly state the appropriate hypotheses and compare to a 5% significance level. (15 points)

Age Group	Sample Average	Sample Standard Deviation	Sample Size
Older	756	88	38
Younger	811	64	46

2SampTTest Stats

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 < \mu_2$

$\bar{x}_1 = 756$

$Sx_1 = 88$

$n_1 = 38$

$\bar{x}_2 = 811$

$Sx_2 = 64$

$n_2 = 46$

$\Rightarrow$ $t = -3.214$

$p = .00101 < .05$

reject Ho

there does appear to be sufficient evidence to think that older people do have shorter stance durations

4. Data shown in the table below represent individual student test scores after reading the book only, or after seeing a 30-minute video lecture on the same material. Determine if watching the video helped students improve their test scores by more than 5 points. Conduct an appropriate hypothesis test at the 10% significance level. (15 points)

SUBJECT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
READING	26	20	15	33	22	11	13	13	16	14	22	11	15	7	15	26
VIDEO	27	24	18	36	30	16	21	28	28	20	25	19	16	29	14	22

Paired TTest

$H_0: \delta = 5$

$H_a: \delta > 5$

$L2 - L1 \Rightarrow L3$

TTest Data

$\mu_0 = 5$

List: L_3

$\mu > \mu_0$

$\Rightarrow$ $t = .546$

$p = .2964 > .10$

fail to reject Ho.

There does not seem to be sufficient evidence to think that the vdo watching improved scores by more than 5 points

5. Conduct an ANOVA test on the following datasets. Clearly state the hypotheses and the result on the test in the context of the problem. The data represents samples of the heights (in) of women in various sports at a particular high school. (15 points)

Gymnastics	**57**	**59**	**62**	**61**	**62**	**58**	**62**	**63**	**60**
Soccer	69	69	64	67	70	65	67	58	64
Volleyball	68	70	69	67	66	73	67	68	69
Baseball	59	69	76	51	70	71	63	72	59

ANOVA (L_1, L_2, L_3, L_4)

$F = 4.7677$

$p = .00738$ ← $< .05$

Factor $df = 3$

$SS = 310$

$MS = 103.3$

Error $df = 32$

$SS = 693.56$

$MS = 21.67$

$Sxp = 4.655$

$H_o: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: at least one $\mu_i \neq \mu_j$ for $i \neq j$

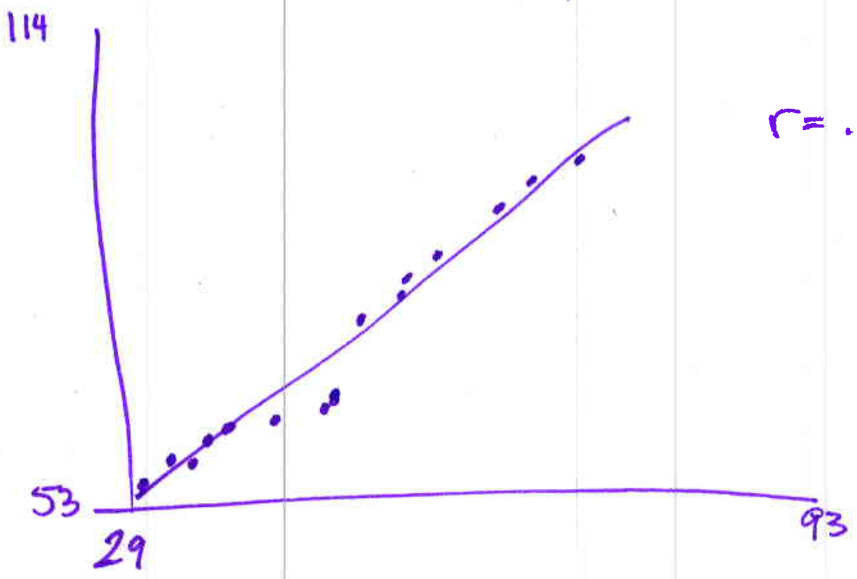
reject Ho

at least one mean is different than the others

6. Use the data to find a linear regression model. Plot the data on a scatterplot along with the regression equation. State the correlation for the model. (15 points)

x	**35**	**38**	**43**	**44**	**46**	**53**	**58**	**59**	**62**	**68**	**68**	**70**	**77**	**80**	**88**
y	61	65	67	69	70	72	73	74	88	91	93	96	101	103	106

$y = .9309x + 26.759$



$r = .9688$

7. What proportion of the change in y (in ~~#8~~ #6) is explained by the relationship with x? (6 points)

$r^2 = .93859$

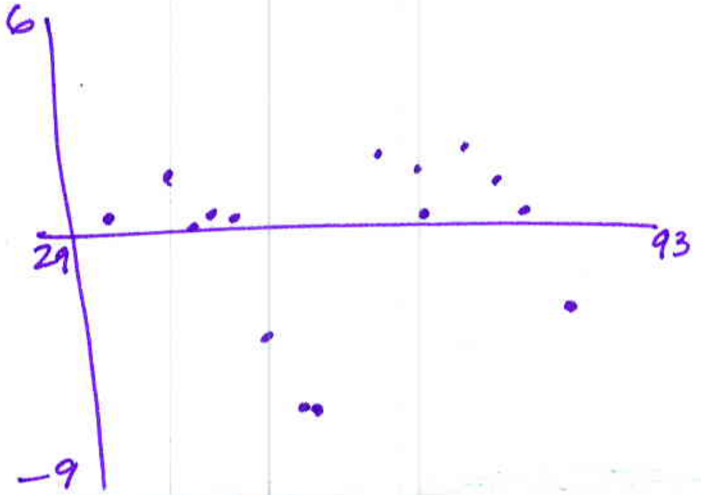
$\approx 94\%$

8. Use the linear regression equation to plot a residual plot and describe any potential problems with the model. Sketch the graph of the residuals vs. x here. (15 points)

$Y_1(L_1) \rightarrow L_3$

$L_2 - L_3 \rightarrow L_4$

plot L_1 vs. L_4



looks fairly random
no clear pattern

9. Let y be the error percentage for subjects reading a 4-digit liquid crystal display, and let x_1 be the level of backlight, x_2 be the character subtense, x_3 be the viewing angle and x_4 be the level of ambient light. The model to fit the data was $y = 1.52 + 0.02x_1 - 1.5x_2 + 0.02x_3 - 0.0006x_4 + \epsilon$.

a. Estimate the mean value of the error percentage when $x_1 = 20, x_2 = 0.8, x_3 = 40, x_4 = 120$. (7 points)

$\hat{y} = 1.448$

b. Interpret the meaning of β_4 when all other variables are held constant. (7 points)

as the level of ambient light increases, the error rate decreases by .0006 % per illumination unit

10. A discrete joint probability mass function is shown in the table below.

		y				
	$f(x,y)$	0	1	2	3	4
	0	0.04	0.07	0.07	0.09	0.13
x	1	0.11	0.07	0.06	0.05	0.05
	2	0.05	0.05	0.06	0.06	0.04

a. Find the marginal distribution function $f_X(x)$. (8 points)

X	0	1	2
p(x)	.4	.34	.26

b. Find $E(Y)$. (10 points)

$$0(.04+.11+.05)+1(.07+.07+.05)+2(.07+.06+.06)+3(.09+.05+.06)$$
$$+4(.13+.05+.04)=2.05$$

c. Find ~~$f_{X|Y}(x|y)$~~ for $X = 2$. (10 points) $\frac{f(x,y)}{f(x)}$

$f_{Y|X}(y|x)$; $X=2$

Y	0	1	2	3	4
p(y)	$\frac{5}{26}$	$\frac{5}{26}$	$\frac{6}{26}$	$\frac{6}{26}$	$\frac{4}{26}$
	$\approx .1923$	$\approx .1923$	$\approx .2308$	$\approx .2308$	$\approx .1538$

11. A sample of 35 items is selected, with 14 of them from type A, and 21 of them from type B. Suppose this data is distributed as a binomial distribution. Find the maximum likelihood function and use it to find an estimate for p, the probability that an element in the population belongs to type A. (15 points)

$$L(p) = p^{14}(1-p)^{21}$$

$$\frac{dL}{dp} = 14p^{13}(1-p)^{21} + p^{14} \cdot 21(1-p)^{20}(-1)$$

$$= p^{13}(1-p)^{20}\left[14(1-p) - 21p\right] = 0$$

$$14 - 35p = 0$$

$$14 = 35p$$

$$\frac{14}{35} = p \text{ or } \boxed{\frac{2}{5} = p}$$

12. Suppose that Gallup wishes to conduct a Presidential poll to determine who in the front-runner in an upcoming election is. They ask 1549 people whether they prefer the Democrat or the Republican and they find that 52% of respondents chose the Republican candidate. Conduct a hypothesis test on these results to determine if this is sufficiently strong evidence to support the claim that more Americans support the Republican candidate than the Democratic one? Clearly state your null and alternative hypotheses, and interpret the results in the context of the problem. (15 points)

1PropZTest

$H_0: p = .5$

$H_a: p > .5$

$p_0 = .5$

$x = .52 * 1549 = 805.48 \Rightarrow 805$

$n = 1549$

$\Rightarrow z = 1.5499$

$p = .06058 > .05$

fail to reject Ho

There is not enough evidence to think that more 50% of Americans favor the Republican Candidate.

13. Construct a 99% confidence interval for the proportion of Americans who support the Republican candidate in the poll in #12. (8 points)

1PropZInt

X=805
n=1549
C-level: .99

(.48699, .55239)

14. What conditions must be satisfied to use a Z-test instead of a T-test when conducting hypothesis testing (or constructing a confidence interval)? (8 points)

σ must be known, normally distributed and/or sample size greater than 40.

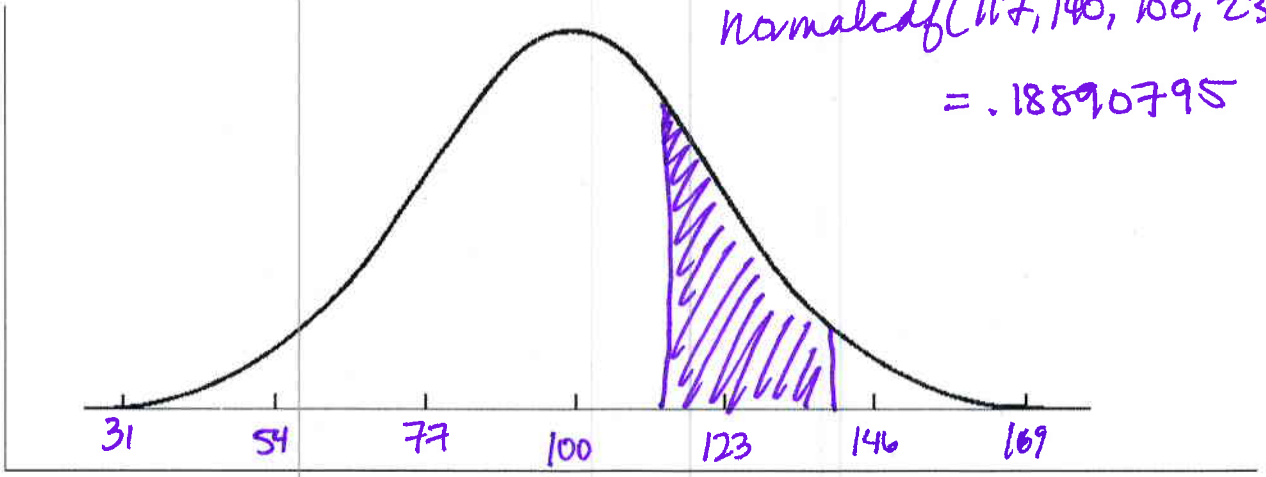
otherwise, use t-test

15. Find the value of x or z, or the indicated probability as needed and graph each situation on the normal curve shown. (7 points each)

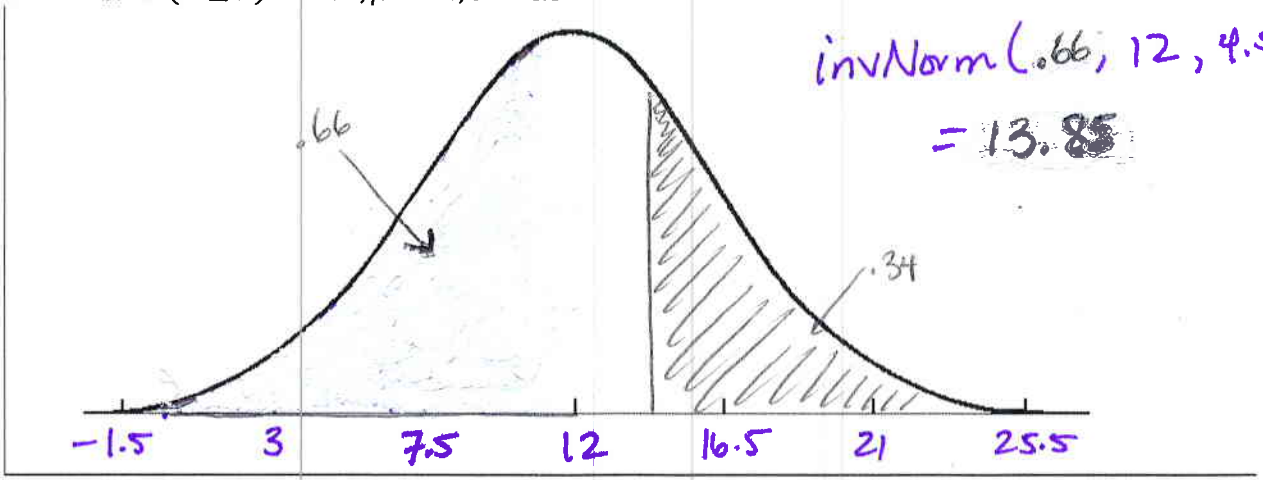
a. $P(117 \leq X \leq 140), \mu = 100, \sigma = 23$

normalcdf(117, 140, 100, 23)

= .18890795



b. $P(X \geq x) = 0.34, \mu = 12, \sigma = 4.5$



invNorm(.66, 12, 4.5)

= 13.85

16. Find the value of k that will make $f(x) = \begin{cases} kx^2, & -1 \leq x \leq 2 \\ 0, & otherwise \end{cases}$ a legitimate probability distribution. (10 points)

$$\int_{-1}^{2} kx^2\,dx = \frac{k}{3}x^3\Big|_{-1}^{2} = \frac{k}{3}[8-(-1)] = \frac{k}{3}(9) = 3k = 1$$

$$k = \frac{1}{3}$$

17. Use the distribution in #16, and the value of k you found, to find the expected value of the distribution. (8 points)

$$\int_{-1}^{2} \frac{1}{3}x^3\,dx = \frac{1}{12}x^4\Big|_{-1}^{2} = \frac{1}{12}[16-(1)] = \frac{15}{12} = \frac{5}{4}$$

18. For each of the distributions given below, find the a) equation of the distribution, and b) $P(X = 9)$. (8 points each)

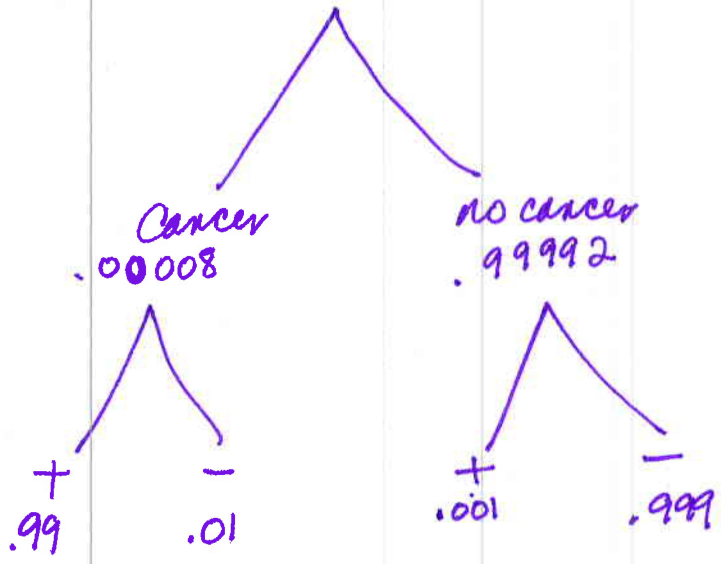
i. A hypergeometric distribution with a population of $N = 500$, a sample size $n = 15$, and the number of success in the population is $M = 135$.

$$\frac{\binom{135}{x}\binom{365}{15-x}}{\binom{500}{15}} , \qquad \frac{\binom{135}{9}\binom{365}{6}}{\binom{500}{15}} \approx .005217$$

ii. A Poisson distribution with $\mu = 5$.

$$p(x) = \frac{e^{-5}5^{x}}{x!} \qquad p(9) = \frac{e^{-5}5^{9}}{9!} = .0362655$$

19. A certain type of cancer occurs in only 0.008% of the population. There is a blood test for this cancer which correctly detects the cancer 99% of the time, and correctly detects the absence of cancer 99.9% of the time. Use Bayes' Theorem and a tree diagram to find the probability of actually having cancer given that you've received a positive blood test. (10 points)



positive tests $= (.00008)(.99) + (.99992)(.001) = .00107912$

chance of cancer given positive test

$$= \frac{(.00008)(.99)}{.00107912} = .07339$$ only about a 7% chance

20. Answer the following questions about the dataset shown here.

42	63	46	53	53	33	25	23	56	56
64	46	36	49	78	56	51	54	68	55
60	52	45	41	54	64	36	48	49	43

a. Find the five-number summary. Determine if any of the datapoints in our set constitute outliers. Are they mild or extreme outliers? Use this information to sketch (to scale) a boxplot. (10 points)

1 Var Stats

min = 23
$Q_1 = 43$
Med = 51.5
$Q_3 = 56$
Max = 78

$Q_3 - Q_1 = 13$

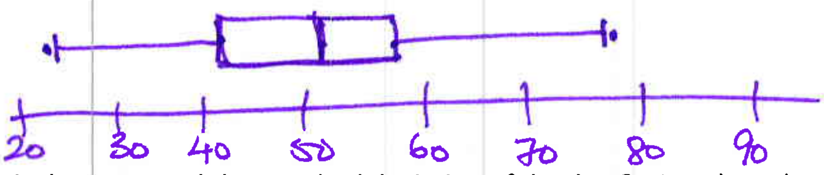
1.5 IQR = 19.5

43 − 19.5 = 23.5 ⇒ 23 is a mild outlier

56 + 19.5 = 75.5 ⇒ 78 is a mild outlier

3 IQR = 39

43 − 39 = 4
56 + 39 = 95 } no extreme outliers



b. What is the mean and the standard deviation of the data? How does the mean compare to the median? (8 points)

$\bar{x} = 49.967$

$S_x = 12.1043$

mean is a bit less than median but pretty close

boxplot looks pretty symmetric