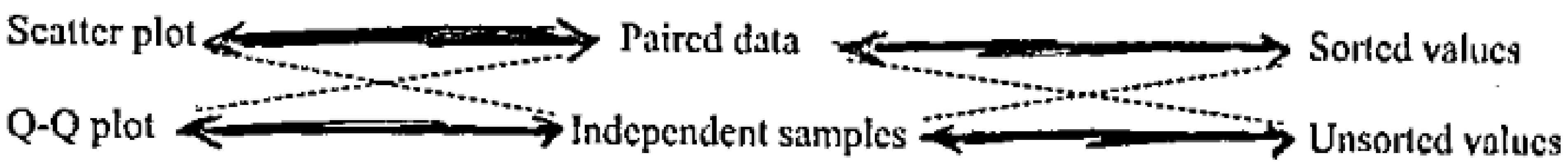
MATH 440B/SCHILLING NAME Erick Castillo.

FINAL EXAMINATION
150 POINTS: SHOW ALL WORK

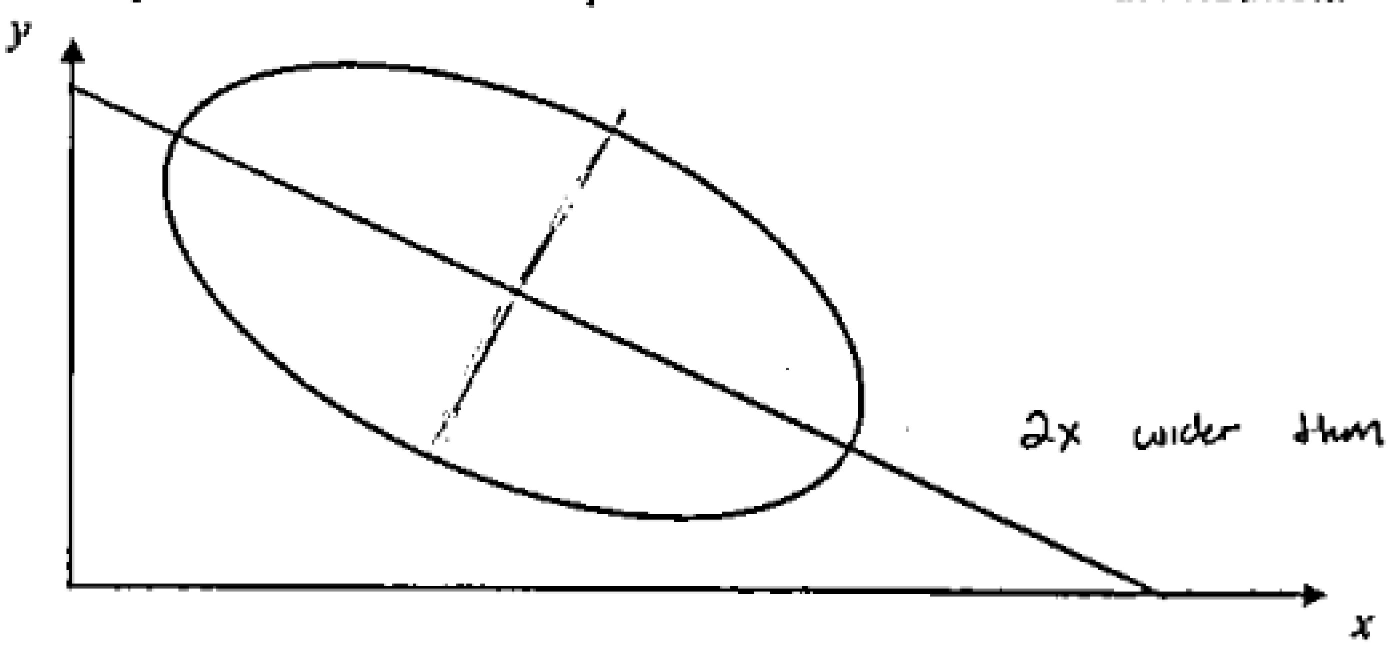
*****PLEASE UPLOAD YOUR EXAM AS A SINGLE DOCUMENT*****

PART I: SHORT PROBLEMS

1. Darken in four of the lines below that connect items that go together:



2. The ellipse below represents a contour of a particular bivariate normal distribution.



2x wider than longer.

(a) As carefully as possible, sketch the regression function $E(Y|x)$ through the ellipse.

(b) Use the method shown in class to give a reasonable estimate of ρ.

$$r \approx \frac{(D/d)^2 - 1}{(D/d)^2 + 1} \approx \frac{(1/2)^2 - 1}{(\frac{1}{2})^2 + 1} \approx \frac{\frac{1}{4} - 1}{\frac{1}{4} + 1} \approx \frac{-\frac{3}{4}}{\frac{5}{4}} \approx \boxed{-\frac{3}{5}}$$

3. We showed in class that for simple linear regression, the mean square error (= variance) of a predicted value $\hat{Y}$ for a given value x of the predictor variable is given by $\frac{\sigma^2}{n}\left[\frac{(x-\bar{x})^2}{s_x^2}+1\right]$.

(a) Show that a prediction made for a value of x that is three standard deviations away from the mean is 10 times less accurate (as measured by mean square error) than a prediction made at the mean itself.

$$\frac{\sigma^2}{n}\left[\frac{(3s-\bar{x})^2}{s_x^2}+1\right] \geq \frac{\sigma^2}{n}\left[\frac{(\bar{x}-\bar{x})^2}{s_x^2}+1\right] = \frac{\sigma^2}{n}$$

where $3s$ is an x value 3 sd's away from the mean.

As desired.

(b) There is typically little or no data whose x values are as far away from the mean as 3 sd's. What is the word that describes a prediction made at such a value?

EXTRAPOLATION!

4. A local hospital is interested in learning whether resistance to the COVID-19 vaccination is associated with educational level. They collect the following data from patients who do not have COVID and have not been vaccinated:

Educational level	Definitely plan to get vaccinated	Not sure or do not plan to get vaccinated	
College degree	10	3	13
No college degree	6	6	12
	16	9	25

(a) Give the name of an appropriate test to assess whether this data gives good evidence that patients with a college degree are more likely to have definite plans to be vaccinated than patients without a college degree.

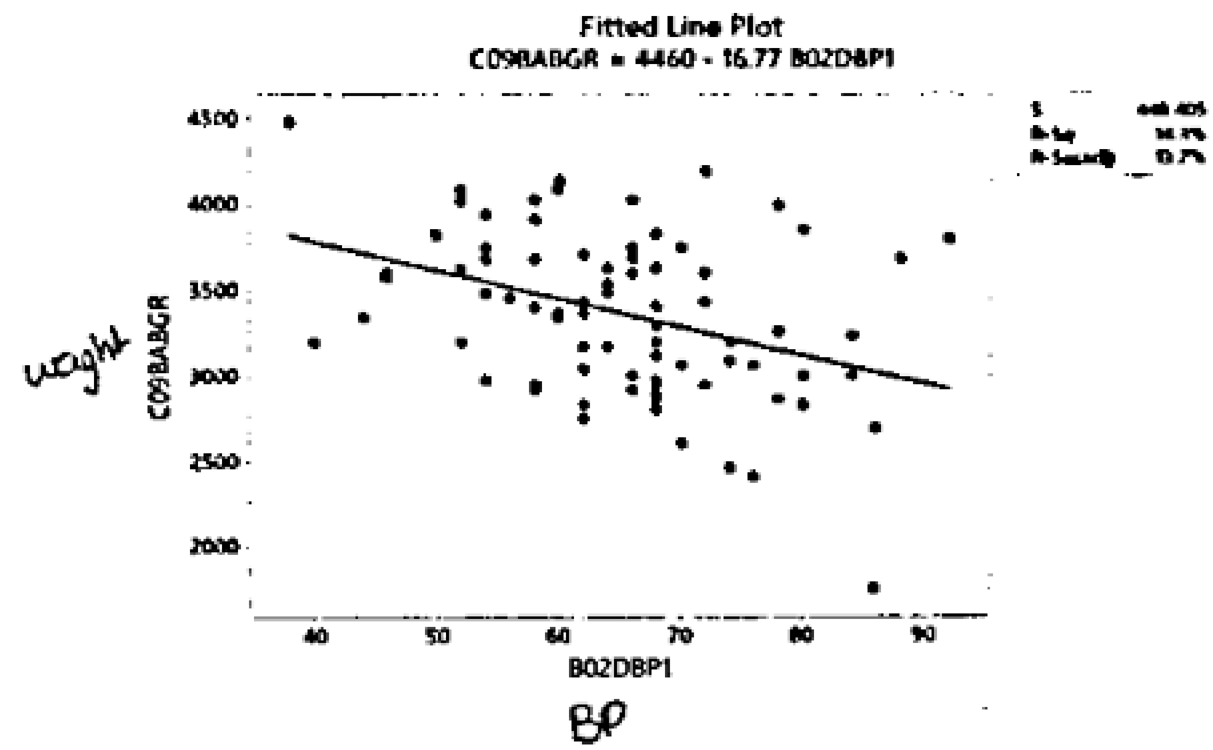
Fisher's Exact Test.

(b) Give a specific mathematical expression for the p-value of the test.

$$p\text{-value}: \frac{\binom{13}{10}\binom{12}{6}}{\binom{25}{16}} + \frac{\binom{13}{11}\binom{12}{5}}{\binom{25}{16}} + \frac{\binom{13}{12}\binom{12}{4}}{\binom{25}{16}} + \frac{\binom{13}{13}\binom{12}{3}}{\binom{25}{16}}$$

one-sided

5. The graph below shows the results of an analysis that looked at the association between x = mother's diastolic blood pressure during pregnancy and y = weight of her newborn baby.



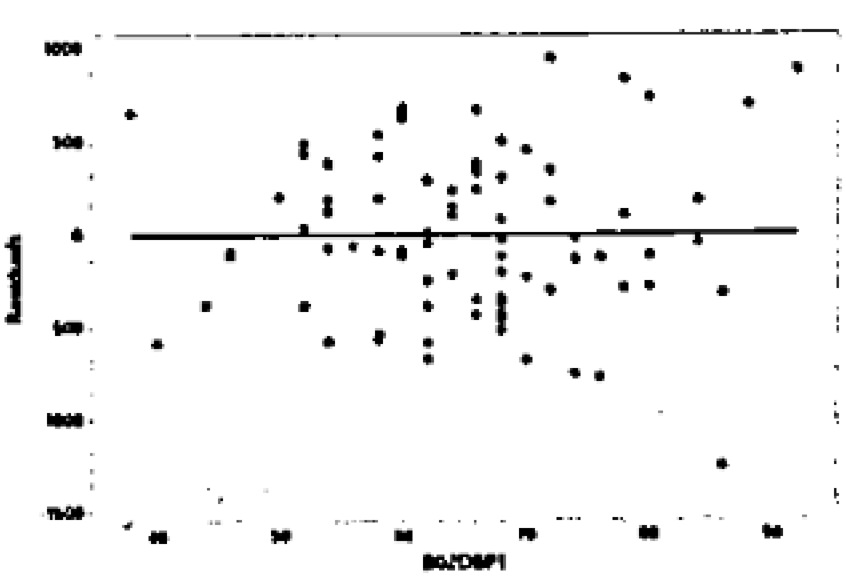
(a) Find the correlation coefficient between the two variables.

$r = \sqrt{R^2} = 0.3782$

(b) Explain how the given value for s relates to the graph.

S states that that predictions of baby's weights are accurate at about 500 units. So 67% of all observations would fall 448.41 units from the regression line.

(c) Below is a residual plot for this analysis. Comment on what it indicates. You don't need to say too much.



The errors mostly appear to be random. Notice that after a BP of 70 the variance of the final residuals seem to increase.

PART II: LONGER PROBLEMS

INSTRUCTIONS: WORK FOUR OF THE REMAINING FIVE PROBLEMS. CROSS OUT THE ENTIRE WORKSPACE OF THE PROBLEM YOU DON'T WISH TO HAVE GRADED. YOU WILL NOT GET CREDIT FOR ALL SIX. EACH PROBLEM IS WORTH THE SAME.

6. A sample of ten married couples was given a survey that assessed how happy they were with their relationship; the possible score range is from 0 to 100. Here are the results:

Husband	87	54	58	47	9	48	78	84	38	70
Wife	78	53	47	38	2	44	67	81	43	61
Differance	9	1	11	9	7	4	11	3	-5	9
Rank ±	7	1	9.5	7	5	3	9.5	2	-4	7

(a) Use the signed-rank test to test the hypothesis that the population distributions of scores for husbands and wives is identical, against the alternative that the population distributions of scores for husbands and wives differ in location. Give bounds for the p-value, and indicate whether the null hypothesis would be rejected at $\alpha = 1\%$. Use Table 9, p. A24.

Two sided, $\alpha = 0.01$ w/ $n=10$, 3

$W^- = -4$.

Because $|-4| > 3$, we say that the test is not significant. Use FTR H0.

Bounds for the p-value: $0.01 < p < 0.02$.

(b) Why should the signed-rank test be used here rather than the rank sum test?

Because the samples are dependent. These people are married

(c) What is the crucial assumption that must be made about the data?

This is a non-parametric method. This does not assume any distribution about the data. This method is preferable for smaller samples.

(d) Give a reason why the signed-rank test is preferable to the sign test here.

The signed-rank test considers the rank/magnitude making it a more powerful test.

(e) Give a reason why the signed-rank test is preferable to the t test here.

The t-test depends on the assumption that normality holds.

Also the t-test is sensitive to outliers.

The signed-rank test overcomes both these faults.

7. Suppose that $X_1, X_2, \ldots, X_n$ are an i.i.d. sample from the negative binomial distribution

$$\binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \ldots$$

(a) Show that the beta distribution Beta(a,b) is a conjugate prior, and determine the posterior density. Recall that the density of a beta random variable is proportional to $p^{a-1}(1-p)^{b-1}$.

Begin by getting Likelihood.

$$L(p) = \prod_{i=1}^{n} \binom{x_i-1}{r-1} p^r (1-p)^{x_i-r}$$

$$= \prod \binom{x_i-1}{r-1} p^{nr} (1-p)^{\sum x_i - nr}$$

Beta Prior:

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}$$

$$\text{Post} \propto \text{Prior} \times \text{Like} \Rightarrow \prod \binom{x_i-1}{r-1} p^{nr} (1-p)^{\sum x_i - nr} * f(p)$$

$$\propto p^{nr+a-1} (1-p)^{\sum x_i - nr + b - 1} \Rightarrow \boxed{\text{Beta}(nr+a, \sum x_i + b - nr)}$$

Because the post is beta, the Beta(a,b) is a conjugate prior

(b) Give the posterior mean.

Post mean: $$\frac{nr+a}{nr+a-(\sum x_i + b - nr)} \Rightarrow \boxed{\frac{nr+a}{2nr+a-b-\sum x_i}}$$

~~8. A random sample $X_1, \ldots, X_7$ is drawn from a symmetric distribution whose form is otherwise completely unknown. The following estimator of the center of symmetry θ is proposed:~~

$$\cancel{\hat{\theta} = [X_{(1)} + X_{(7)} + 2(X_{(2)} + X_{(6)}) + 3(X_{(3)} + X_{(5)}) + 4X_{(4)}]/16,}$$

~~where the $X_{(i)}$'s are the ordered values of the sample. Suppose $\hat{\theta} = 16.8$ and that $B = 600$ bootstrap samples are drawn. Suppose the ordered resampled values $\hat{\theta}_j^*$ are 13.2, 13.5, 13.5, 13.6, 13.8, 14.1, 14.3, 14.3, 14.4, 14.6, 14.7, ..., 22.0, 22.0, 22.0, 22.1, 22.1, 22.3, 22.5, 22.8, 22.8. Find a 99% bootstrap confidence interval for θ.~~

SKIP 8

Parts a and b should be switched!
I solved a with $\hat{\lambda}_{MLE}$ and b with $\lambda = 4$
Thanks :)

9. It is postulated that the number of customers that visit a particular ATM between 12am and 5am should follow a Poisson distribution with parameter $\lambda = 4$. The data shown below were collected for 50 consecutive nights. The mle was found to be $\hat{\lambda} = \bar{x} = 3.52$.

	No. of gamma rays	0	1	2	3	≥4
	Count	1	6	11	8	26
$\hat{\lambda}_{MLE}$ →	probability	0.0296	0.1042	0.1834	0.2152	0.4676
↳	Expected #	1.539	5.418	9.537	11.1904	24.3152
$\lambda = 4$ →	probs	0.0183	0.0733	0.1465	0.1954	0.5665
↳	Expected #	0.9524	3.809	7.619	10.159	29.458

52

(a) Use Pearson's χ^2 test to test whether the data fits the Poisson distribution ($p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0, 1, 2, \ldots$) with parameter $\lambda = 4$. You can put part of your work above, below the table.

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = \frac{(1-1.539)^2}{1.539} + \ldots + \frac{(26-24.3152)^2}{24.3152}$$

$$= 0.1888 + 0.0625 + 0.2244 + 0.9096 + 0.1167$$

$$\boxed{\chi^2 = 1.502} \qquad df = 5-1-1 = 3$$

Because df is so much larger than the test statistic, we can say that the Poisson model fits the data well.

(b) Test whether the data fits any Poisson distribution. Use the fact that the m.l.e. of λ is $\bar{x} = 3.52$.

$$\chi^2 = \frac{(1-0.9524)^2}{0.9524} + \ldots + \frac{(26-29.458)^2}{29.458}$$

$$= 0.0023 + 1.2603 + 1.5003 + 0.4588 + 0.4059$$

$$\boxed{\chi^2 = 3.628} \qquad df = 4$$

This model with $\lambda = 4$ fits the data pretty well as well.

so $\hat{\vec{B}} = \begin{bmatrix} 1/4(y_{11}+y_{12}+y_{21}+y_{22}) \\ 1/4(y_{11}+y_{12}-y_{21}-y_{22}) \\ 1/4(y_{11}-y_{12}+y_{21}-y_{22}) \end{bmatrix}$

10. Consider the following general linear model:

$$Y=\begin{bmatrix} y_{11}\\ y_{12}\\ y_{21}\\ y_{22}\end{bmatrix},\quad X=\begin{bmatrix}1&1&1\\1&1&-1\\1&-1&1\\1&-1&-1\end{bmatrix},\quad \beta=\begin{bmatrix}\mu\\ \alpha_1\\ \beta_1\end{bmatrix},\quad e=\begin{bmatrix}e_{11}\\ e_{12}\\ e_{21}\\ e_{22}\end{bmatrix},$$ with the $\{e_{ij}\}$~ i.i.d. $N(0,\sigma^2)$.

(a) Find expressions for the least squares estimator of β.

$\hat{B} = (X'X)^{-1}X'\vec{Y}$

$(X'X)^{-1} = \left(\begin{bmatrix} 1&1&1&1\\ 1&1&-1&-1\\ 1&-1&1&-1 \end{bmatrix} \begin{bmatrix} 1&1&1\\ 1&1&-1\\ 1&-1&1\\ 1&-1&-1 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 4&0&0\\ 0&4&0\\ 0&0&4 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{4}&0&0\\ 0&\frac{1}{4}&0\\ 0&0&\frac{1}{4} \end{bmatrix}$

$X'\vec{Y} = \begin{bmatrix} 1&1&1&1\\ 1&1&-1&-1\\ 1&-1&1&-1 \end{bmatrix} \begin{bmatrix} Y_{11}\\ Y_{12}\\ Y_{21}\\ Y_{22} \end{bmatrix} = \begin{bmatrix} Y_{11}+Y_{12}+Y_{21}+Y_{22}\\ Y_{11}+Y_{12}-Y_{21}-Y_{22}\\ Y_{11}-Y_{12}+Y_{21}-Y_{22} \end{bmatrix}$ Answer above

(b) Find the variances of $\hat{\mu}$, $\hat{\alpha}_1$ and $\hat{\beta}_1$.

$Var(\hat{\mu}) = Var(\hat{\alpha}_1) = Var(\hat{\beta}_1) = \frac{1}{4}\sigma^2$

(c) Show that these three estimators are independent.

$Cov(\hat{\mu},\hat{\alpha}_1) = Cov(\hat{\mu},\hat{\beta}_1) = Cov(\hat{\alpha}_1,\hat{\beta}_1) = 0.$

Thus the three estimators are independent.

(d) Now show how this represents the two-way ANOVA model $Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ with $I = 2$, $J = 2$ and $K = 1$. To do this,

(i) Use the ANOVA constraint equations to find expressions for α_2 and β_2 in terms of the parameters in β above.

$\sum \alpha_i = 0$ and $\sum \beta_j = 0$. $\quad \beta_1 + \beta_2 = 0$

$\beta_2 = -\beta_1$

$\alpha_1 + \alpha_2 = 0$.

$\boxed{\beta_2 = \frac{1}{4}(y_{12}+y_{22}-y_{11}-y_{21})}$

$\frac{1}{4}(y_{11}+y_{12}-y_{21}-y_{22}) + \alpha_2 = 0.$

$\boxed{\alpha_2 = \frac{1}{4}(y_{22}+y_{21}-y_{11}-y_{12})}$