METHODS QUALIFYING EXAM

August 2010

Student's Name
INSTRUCTIONS FOR STUDENTS:
1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the
UPPER RIGHT HAND CORNER of EACH PAGE of your solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer Questions I, II, and III but select only ONE of Questions IV and V to answer.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. Be sure to hand in all of your exam. No additional material will be accepted once the exam has ended and you have left the exam room.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.
8. You are to answer Questions I, II, and III and then select ONE of Questions IV and V in this exam.
I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.
Student's Signature
INSTRUCTIONS FOR PROCTOR:
Immediately after the student completes the exam, \mathbf{fax} the student's solutions to $\mathbf{979\text{-}845\text{-}6060}$ or email to $\mathbf{longneck@stat.tamu.edu}$ Do not send the questions, just send the student's solutions.
(1) I certify that the time at which the student started the exam was
and the time at which the student completed the exam was
(2) I certify that the student has followed all the INSTRUCTIONS FOR STUDENTS listed above.
(3) I certify that the student's solutions were faxed to 979-845-6060 or
emailed to longneck@stat.tamu.edu.

Proctor's Signature_____

PROBLEM I. Part A:

For the following experiment, provide the following information:

- 1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
- 2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
- 3. Identify each of the factors as being Fixed or Random;
- 4. Describe the Experimental Units and Measurement Units;
- 5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures;
- 6. A partial ANOVA Table containing just Sources of Variation (SV) and Degrees of Freedom (DF);

Commercial cheese is manufactured by bacterial fermentation of Pasteurized milk. Selected bacteria is added to the milk to implement the fermentation, referred to as starter cultures. However, some *Wild* bacteria may also be present in cheese, nonstarter bacteria, which may alter the desired quality of the cheese. Thus, cheese manufactured under seemingly identical conditions in two cheese making facilities may produce cheese of differing quality due to the present of different indigenous nonstarter bacteria. To test the impact of two nonstarter bacteria, R50 and R21, on cheese quality, the nonstarter bacteria was added to the cheese to see if it impacted the quality of the cheese.

The researchers decided to use four types of nonstarter bacteria: a control (no nonstarter bacteria added), addition of R50, addition of R21, and addition of a blend of R50 and R21. Twelve containers of cheeses were made, three of each of the four types of nonstarter bacteria, with the type of bacteria randomly assigned to the cheese containers. Each of the 12 containers of cheese was then divided into four portions. The four portions are then randomly assigned to one of four aging times: 1 day, 28 days, 56 days, and 84 days. At the end of the specified aging period, the cheese is measured for total free amino acids. The researcher was particularly interested in the bacterial effects and their interaction with aging times.

		Days						
Bacteria	Container	1	28	56	84			
Control	1	0.637	1.250	1.697	2.892			
	2	0.549	0.794	1.601	2.922			
	3	0.604	0.871	1.830	3.198			
R50	1	0.678	1.062	2.032	2.567			
	2	0.736	0.817	2.017	3.000			
	3	0.659	0.968	2.409	3.022			
R21	1	0.607	1.228	2.211	3.705			
	2	0.661	0.944	1.673	2.905			
	3	0.755	0.924	1.973	2.478			
R50+R21	1	0.643	1.100	2.091	3.757			
	2	0.581	1.245	2.255	3.891			
	3	0.754	0.968	2.987	3.322			

PROBLEM I. Part B:

For each of the following questions, select **ONE** letter from the list on the next page which is the **BEST** solution to each of the following situations.

SITUATION:

- (1) A CRD with three factors: F_1 -fixed levels, F_2 -fixed levels, F_3 -fixed levels, was conducted. The experimenter obtained the following results from the AOV F-tests: $F_1 * F_2 * F_3$ is significant, $F_1 * F_2$ -not significant, $F_1 * F_3$ -significant, $F_2 * F_3$ not significant, and F_1 , F_2 , F_3 are all not significant. She wants to determine which pairs of means are different across the levels of F_1 .
- (2) A three factor experiment is run with Factor F_1 -fixed, Factor F_2 -fixed having levels nested within the levels of Factor F_1 , Factor F_3 -fixed crossed with Factor F_1 . The interaction between factors F_1 and F_3 was not significant and the interaction between factors $F_2(F_1)$ and F_3 was not significant. The researcher was interested in determining which pairs of means are different across the levels of F_1 .
- (3) A RCBD with three factors: F_1 -fixed, F_2 -fixed, F_3 -random, was conducted. The experimenter obtained the following results from the AOV F-tests: $F_1 * F_2 * F_3$ is not significant, $F_1 * F_3$ -not significant, $F_1 * F_3$ -significant, $F_2 * F_3$ -significant, and F_1 , F_2 , F_3 are all not significant. She wants to determine if there are pairwise differences in the levels of F_1 .
- (4) In a quality control experiment, the production engineer was interested in evaluating factors which may have caused a high defective rate in a product. There are five Rates, F₁, at which a platinum coating is applied to the product, with levels, 1.0, 1.1, 1.2, 1.3, 1.4 mm/second. The second factor, F₂, is four Types of Machines used to apply the coating to the product, with levels, M1, M2, M3, M4. The third factor, F₃, was the Operators the coating machines. Twenty operators of the coating machines were randomly selected from the workforce. Each operator applied the coating to 80 units of the product, four units for each combination of a Rate and Type of Machine. There was significant evidence of a 3-factor interaction and all 2-factor interactions were found to be significant. The company wants to know if the mean defective rate, μ_{ijk}, increased as the Rate, F₁, of applying the coating was increased.
- (5) An experiment is designed to investigate plant growth involving a factorial treatment structure with factor F_1 , the temperature in a growth chamber, $15^{\circ}C$, $20^{\circ}C$, $30^{\circ}C$, $35^{\circ}C$ crossed with factor F_2 , four brands of growth stimulants at three dose levels: 0 ml/mg, 10 ml/mg, 15 ml/mg, factor F_3 . The experiment was conducted as a completely randomized design with 10 flowers randomly assigned to each of the 36 treatments. The experimenter determined from the AOV F-tests that only the following effects were significant: $F_1 * F_3$, $F_1 * F_2$, F_2 , F_3 . The researcher wants to determine the temperature that yields the maximum mean growth.

TECHNIQUE:

- A. Trend analysis using Scheffe contrasts
- B. Trend analysis using Bonferroni contrasts
- C. Trend analysis in the levels of F_1 averaged over levels of the other factors
- D. Trend analysis in the levels of F_1 separately at each level of the other factors
- E. Trend analysis in the levels of F_1 separately at each level of F_2 but averaged over the other factors
- F. Scheffe's test for contrast differences
- G. Dunnett's comparison technique
- H. Dunnett's comparison technique to all combinations of the factors
- I. Dunnett's comparison technique applied to the levels of factor F_1 separately at each level of the other factors
- J. Dunnett's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors
- K. Tukey's comparison technique
- L. Tukey's comparison technique to all combinations of the factors
- M. Tukey's comparison technique applied to the levels of factor F_1 separately at each level of the other factors
- N. Tukey's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors
- O. Hsu's comparison technique
- P. Hsu's comparison technique applied to the levels of factor F_1 separately at each level of the other factors
- Q. Hsu's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors
- R. Hsu's comparison technique applied to all combinations of the factors
- S. Nothing new is learned beyond the results of the F-tests from the AOV table.
- T. Comparison of marginal means is not appropriate.
- U. None of the above methods are appropriate.

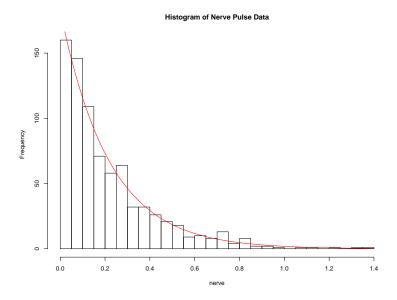
PROBLEM II.

Consider a set of observations $Y_1, Y_2, ..., Y_{400}$, which are assumed to be independent and identically distributed with a mean μ and variance σ^2 . NOTE that due to the large sample size involved here, you may use normal-distribution tables (distributed with this examination) for any probability calculations or decision rules required below.

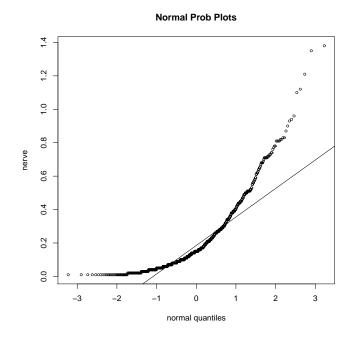
- 1. Consider the hypotheses $H_o: \mu = 12$ versus the two-sided alternative hypothesis $H_1: \mu \neq 12$.
 - a. Write down a general formula for the t test statistic commonly used for this hypothesis test.
 - b. Write down the decision rule for this hypothesis test. Use $\alpha = 0.05$.
- 2. For this part of the question, you may assume that σ is known with $\sigma = 1$.
 - a. Calculate the power of your test for the following six values of the true parameter μ : 11.9, 11.95, 11.975, 12.025, 12.05, 12.1.
 - b. Use your results from part 3.a to sketch a power curve for your test. Be sure to label your axes clearly.
- 3. An agonomist reviews your work from parts 1. through 2. and comments, "The power you have for $\mu = 12.025$ is lower than I was hoping to get. How can I increase it?" Answer your agronomist's question by providing at least two ways in which the power can be increased.
- 4. The 400 observations considered above represent the weight (in grams) of pecans (a type of nut). However, (unknown to you) the 400 pecans were actually collected from 10 trees, with 40 pecans picked from each tree. Also, your agronomist admits that within a given tree, pecan weights cannot be considered independent, and will have a strong *positive* correlation, due to common genetic and environmental factors. Given this additional information, answer the following questions without carrying out additional calculations.
 - a. How will this positive correlation within trees affect the expectation of the variance estimator of the sample mean that you used in part 1.a.?
 - b. Suppose you ignored the positive correlation within trees and proceeded to use the test you proposed in part 1. Will the actual values of the power of the test be larger or smaller than the values you calculated in part 2.? Explain.
- 5. In light of your answer to part 4., your agronomist states, "OK, I see that it's wrong to use the test statistics from part 1. to evaluate the hypotheses. What should I do instead?" Answer your agronomist's question by presenting a standard testing method that will account appropriately for the correlated data described in part 4. Be sure to give clear, explicit statements of both your test statistic formula and your decision rule.

PROBLEM III.

We consider a data set consisting of 799 waiting times between successive pulses along a nerve fiber. The data appear in the following histogram with a smooth curve superimposed on the histogram.



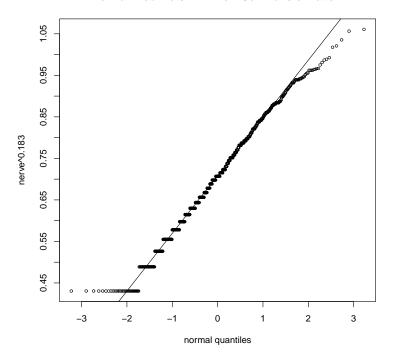
a.) QQ plots and goodness-of-fit tests are often used to assess distributional assumptions. The following normal quantile plots were produced in R using the command qqnorm. Also, the command shapiro.test was used on the data set. Use following plot and the results of the R command shapiro.test to discuss the assumption of normality for the pictured data. If the data are nonnormal, describe the manner in which the data are nonnormal.



Shapiro-Wilk normality test

data: nerve W = 0.8271, p-value < 2.2e-16 b.) The Box-Cox transformation is often used to produce a data set that is better fit by the normal distribution. For this data set, the value of the Box-Cox parameter that produced the best fit was $\hat{\lambda} = 0.183$. Describe carefully the transformation that corresponds to this value. Then use the following plot and results of the R command shapiro.test to discuss the assumption of normality for the transformed data.

Normal Prob Plots with Box-Cox Transformation



Shapiro-Wilk normality test

data: nerve 0.183 W = 0.9897, p-value = 2.269e-05

c.) Looking for a transformation to normality may not be appropropriate for this data set. Discuss some reasons why the exponential distribution with cumulative distribution function

$$F(x) = \begin{cases} 1 - e^{-\beta x} & x > 0\\ 0 & x \le 0 \end{cases}$$

or the Weibull distribution with cumulative distribution function

$$F(x) = \begin{cases} 1 - e^{-(x/\alpha)^{\gamma}} & x > 0\\ 0 & x \le 0 \end{cases}$$

might be more reasonable for the above data.

d.) The normal qq-plot in parts a. and b. was used to assess the fit of the normal distribution to the data. Discuss similar plots that could be used to assess the fit of the exponential or the Weibull distributions to the data. Be sure to explain why these plots are appropriate.

6

PROBLEM IV.

The World Health Organization defines low birth weight as a baby weighing less than 2500 grams at birth. This data was taken from Stat Labs: Mathematical Statistics Through Applications by Deb Nolan and Terry Speed, University of California, Berkeley. This data is available at http://www.statsci.org/datasets.html.

We want to determine if gestation, age, weight, height and the mother smoking can be use to predict low birth weight.

Variable	Description
bwt	1 if less the 2500 grams; 0 otherwise
Gestation	Length of pregnancy in days
age	mother's age in years
height	mother's height in inches
weight	Mother's prepregnancy weight in pounds
smoke	Smoking status of mother 0=not now, 1=yes now

1. Since the dependent variable (low birth weight – bwt) is a one (1) or a zero (0), multiple linear regression is not appropriate. Please give some reasons why multiple linear regression is not appropriate.

- 2. What is the logit equation for the model that predicts low birth weight form the predictors given above?
- 3. Given the graphics on pages 10, 11, 12, and 13; the researcher decided to use a log transformation on the predictor variables. Does this seem reasonable? Please explain you answer.
- 4. Suppose that your model for low birth weight contains the following predictor variables:

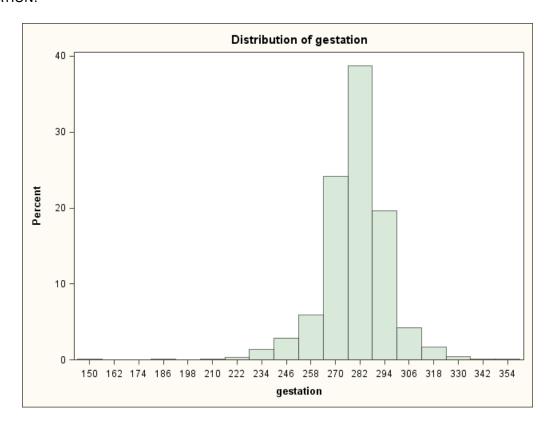
smoke					
LOG_	GEST				
LOG_	AGE				
LOG	HEIGHT				
LOG	_WEIGHT				

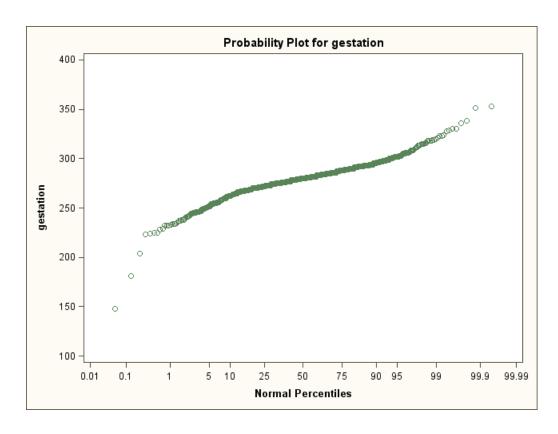
Given the table below the researcher wants to delete all three of the non-significant variables all at once. Is this the correct approach? Explain your answer.

Analysis of Maximum Likelihood Estimates										
				Standard						
Parameter		DF	Estimate	Error	Chi-Square	Pr > ChiSq				
Intercept		1	109.0	18.7220	33.8971	<.0001				
smoke	0	1	-1.2378	0.3058	16.3821	<.0001				
LOG_GEST		1	-18.4168	2.1783	71.4834	<.0001				
LOG_AGE		1	0.8558	0.6826	1.5719	0.2099				
LOG_HEIGHT		1	-2.1324	3.9408	0.2928	0.5884				
LOG_WEIGHT		1	-0.4369	1.0929	0.1598	0.6893				

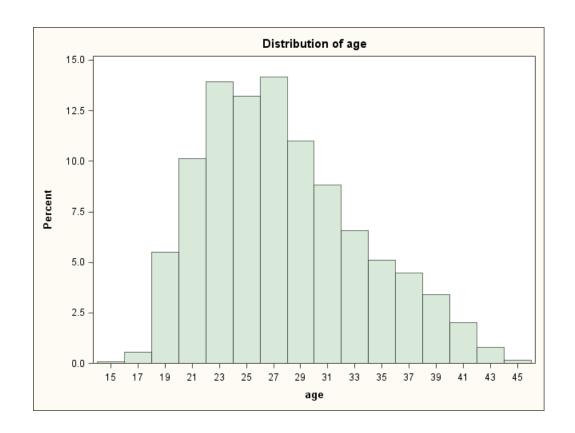
5. Suppose that your model for low birth weight has just smoke and LOG_GEST as predictors. Page 14 has the marginal model plots. Do these indicate a valid model? What transformation would you suggest?
6. After an appropriate transformation, the marginal model plots on page 15 were obtained. Do these indicate a valid model? Explain your answer.
7. If the probability of a low birth weight baby is .437 for a log gestation of 5.5 for a woman who smokes and the probability of a low birth weight baby is .204 for a log gestation of 5.5 for a woman who does not smokes, what is the odds ratio of getting a low birth weight baby?
8. If the probability of a low birth weight baby is .559 for a log gestation of 5.48 for a woman who smokes and the probability of a low birth weight baby is .295 for a log gestation of 5.48 for a woman who does not smokes, what is the odds ratio of getting a low birth weight baby?
9. A researcher found that the odds ratio in parts 7 and part 8 are identical. How can you explain this phenomenon?

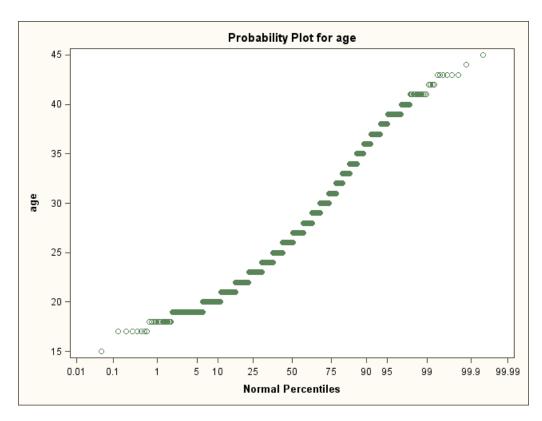
GESTATION:



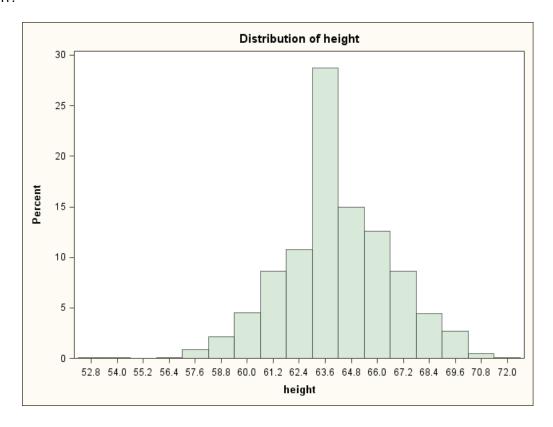


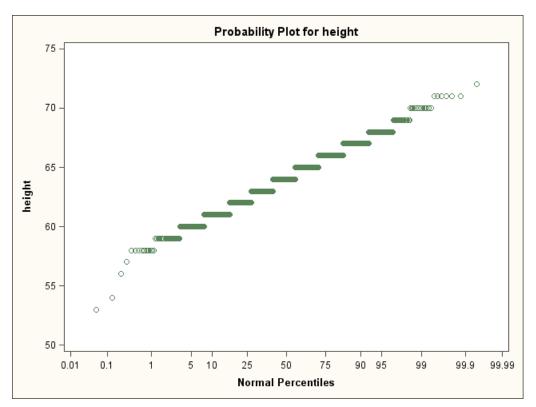
AGE:



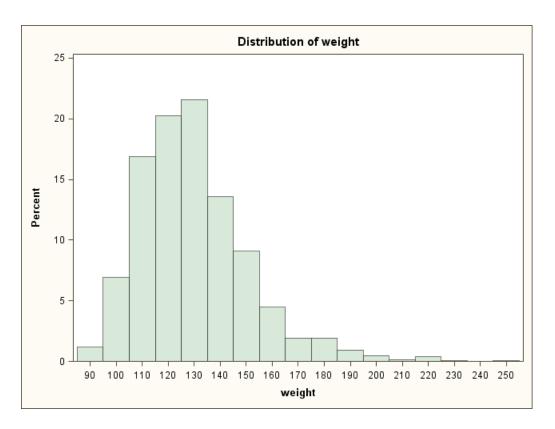


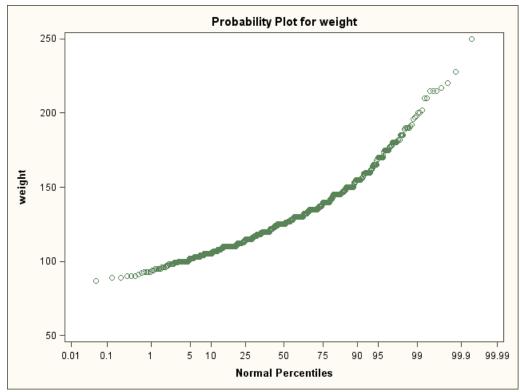
HEIGHT:

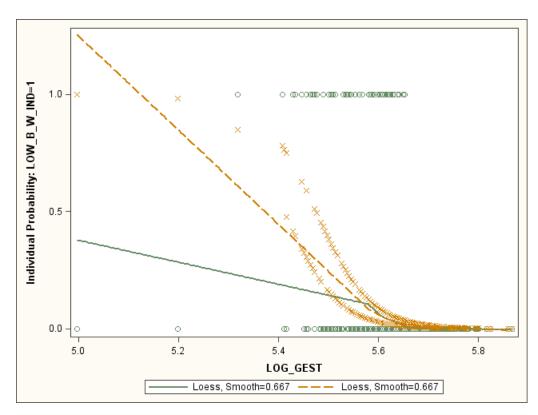


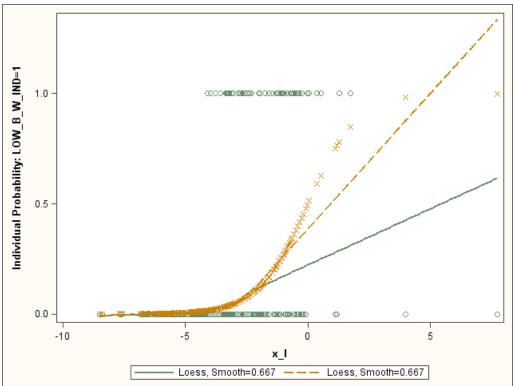


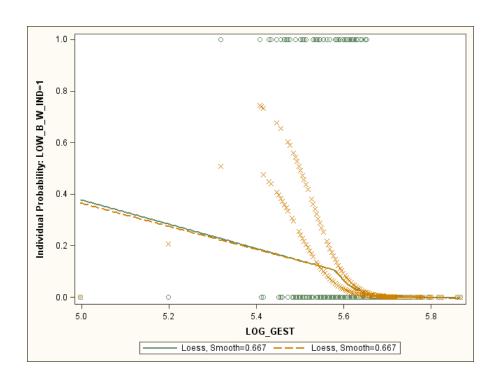
WEIGHT:

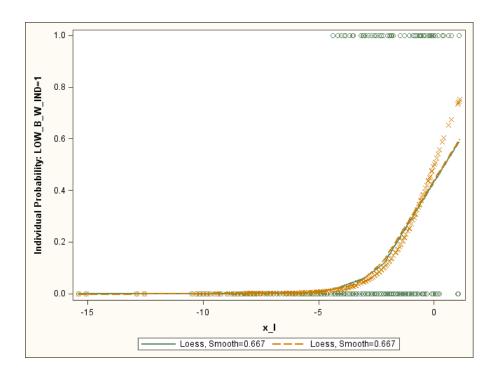












PROBLEM V.

Let

$$Y_1 = \alpha_1 + \varepsilon_1$$

$$Y_2 = 2\alpha_1 - \alpha_2 + \varepsilon_2$$

$$Y_3 = \alpha_1 + 2\alpha_2 + \varepsilon_3$$

where ϵ_1,ϵ_2 and ϵ_3 are iid $N(0,\sigma^2)$ random variables. Derive the F statistic for testing

$$H_0: \alpha_1 = \alpha_2$$
.

Note: "Derive the F statistic" means to produce an expression which is a function only of real numbers and the random variables Y_1, Y_2 , and Y_3 . You may (and are encouraged to) introduce simplifying notation in your derivation. However, be sure to carefully define, in terms of real numbers and Y_1, Y_2 , and Y_3 , all notation that you introduce.

Hint: Consider the usual multiple linear regression model, $Y = X\beta + \epsilon$. Y is an $n \times 1$ vector of response variables, X is an $n \times p$ matrix (of rank p) of predictor variables, β is a $p \times 1$ vector of unknown parameters and ϵ is an $n \times 1$ vector of unobservable independent and identically normally distributed random variables, each with mean zero and variance σ^2 . Either of two (equivalent) forms of the F statistic for testing the null hypothesis $H_0: A\beta = c$ versus $H_1: A\beta \neq c$, where A is a known $q \times p$ matrix of rank q, are

$$F = \frac{\left(RSS_H - RSS\right)\!/q}{RSS\!/\!\left(n - p\right)} = \frac{\left(A\hat{\beta} - c\right)^T\!\left[A\!\left(X^TX\right)^{\!-1}A^T\right]^{\!-1}\!\left(A\hat{\beta} - c\right)\!/\!q}{RSS\!/\!\left(n - p\right)} \ ,$$

where RSS and RSS_H are the residual sum of squares for the least squares fit of the unconstrained model and the model constrained by H_0 : $A\beta = c$, respectively.

Tables of the Normal Distribution

			Probability Content from 0.00 to Z							
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713									
2.0	0.9772									
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861									
2.3	0.9893									
2.4	0.9918									
2.5	0.9938									
2.6	0.9953									
2.7	0.9965									
2.8	0.9974									
2.9			0.9982							
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

	Far Right Tail Probabilities	1				
Z P{Z to	oo} Z P{Z to oo}	z	P{Z to 00}	2	Z P{	[Z to oo}
2.0 0.022	75 3.0 0.001350	4.0	0.00003167	!	5.0 2.	.867 E-7
2.1 0.017	86 3.1 0.0009676	4.1	0.00002066	!	5.5 1.	.899 E-8
2.2 0.013	90 3.2 0.0006871	4.2	0.00001335	- [(6.0 9.	866 E-10
2.3 0.010	72 3.3 0.0004834	4.3	0.00000854	i (6.5 4.	.016 E-11
2.4 0.008	20 3.4 0.0003369	4.4	0.000005413	j :	7.0 1.	.280 E-12
2.5 0.006	21 3.5 0.0002326	4.5	0.000003398	j :	7.5 3.	191 E-14
2.6 0.004	661 3.6 0.0001591	4.6	0.000002112	į į	8.0 6.	221 E-16
2.7 0.003	467 3.7 0.0001078	4.7	0.000001300	į į	8.5 9.	480 E-18
2.8 0.002	555 3.8 0.00007235	4.8	3 7.933 E-7	j :	9.0 1.	129 E-19
2.9 0.001	866 3.9 0.00004810	4.9	4.792 E-7	j :	9.5 1.	.049 E-21