MS In-Class Comprehensive Exam (Statistics) 18 August 2017

<u>Instructions</u>: This exam is closed notes and closed book though you may use a calculator. Please answer these questions on separate sheets of paper **making sure to include your name at the top of each page, number the pages, and only write on ONE SIDE of the page**. **For each question (there are four) please use a new sheet of paper.** Due to the length of the exam, complete sentences are not necessary. Please look over the entire exam before beginning. Good luck!

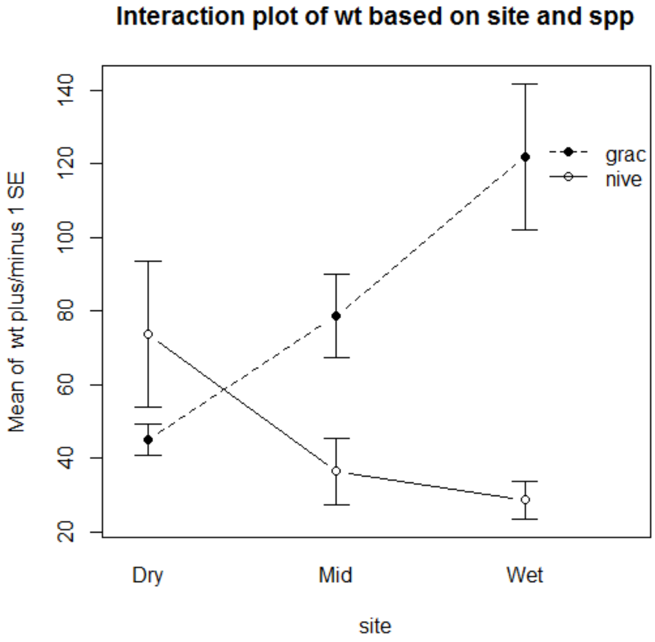
****

1) **Potentilla** [40pts]. An ecologist is interested in the distribution of plants in alpine tundra (a type of biome that does not contain trees due to their altitude) at Cumberland Pass in the Colorado Rockies. In one part of this study, individual plants of two species (*Potentilla nivea*—a small plant occurring only in dry ridgetops and *Potentilla gracilis*—a larger plant that naturally occurs in only moist sites) were excavated and then transplanted to new sites (dry, mid, and wet). The questions of interest to the ecologist are:

   1) Did the transplant site have any effect on the growth of Potentilla averaged over the two species? If so, what is the nature of those differences?
   2) Do the two species respond differently to site (for example does *P. nivea* grow better at one site than *P. gracilis*)? If so, what is the nature of those differences?

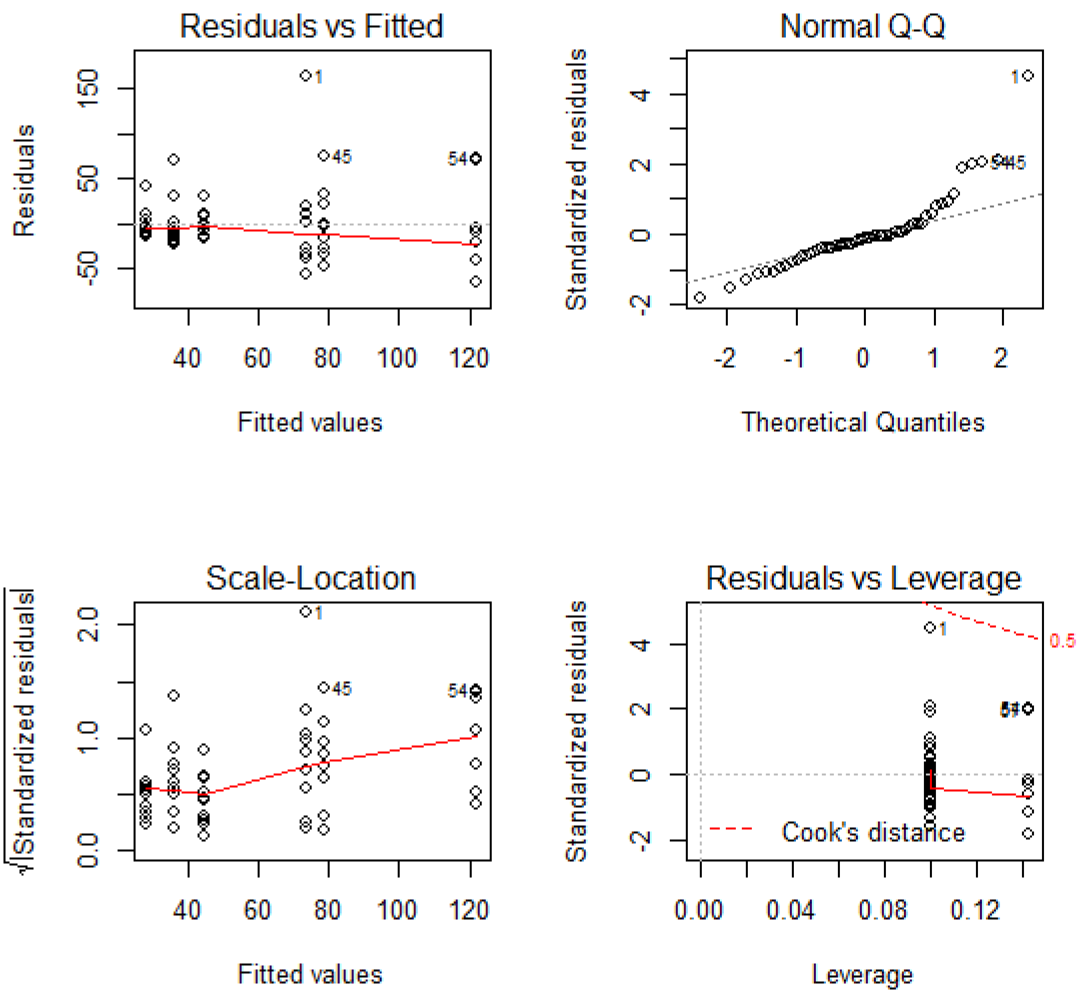
   Initially the ecologist started with 10 plants for each species for each site (that is for *P. nivea* 10 plants were transplanted to the dry site, 10 to the mid, and 10 to the wet and similarly for *P. gracilis*) for a total of 60 plants. Unfortunately, some seedlings died so the number of plants that were harvested three years later varied between 5 and 10. For each plant the dry mass of the year's growth (in mg) was measured.

   a) Below is an interaction plot with site on the X-axis, species as the plotting factor, and dry mass (weight) as the response. Standard error bars are also included. Carefully explain what this plot indicates. [4pts]



   b) Set up the model of interest to the researchers carefully defining each parameter. [6pts]

c) Before jumping into answering the ecologist's research questions let's first assess the adequacy of the model. Based on the diagnostic plots below, briefly explain if you have any concerns regarding this model making sure to reference the appropriate plot as needed. [6pts]



d) After discussing some concerns with you, the ecologist decides to log transform the response. Report the F-statistics and p-values for the two research questions of interest to the ecologist if possible and appropriate. You'll notice that someone provided you with several ANOVA tables associated with Type I, Type II, and Type III sums of squares in the output handout. When reporting your F-statistics and p-values indicate which type of SS you are using and why. Also, if it is not reasonable or possible to provide an F-statistic and p-value explain why. [8pts]

e) There are two ways that the ecologist could obtain point estimates, test statistics, and p-values to answer the second questions of interest. The first is to conduct a separate analysis for each species. The second option is to combine the data for both species and use linear contrasts to examine the differences between species within a given site. Will these two approaches provide the same point estimates? p-values? Explain. [4pts]

f) Which option from the previous question do you recommend and why? [3pts]

This study was actually conducted at two sites—Cumberland Pass and Pennsylvania Mountain. For our purposes we can think of this as conducting two different experiments one at each location. The purpose of conducting this second study was to examine if the findings from Cumberland Pass are similar to those found at Pennsylvania Mountain.

g) This new model can be thought of as a three-way factorial with location as the third factor. Assuming that none of the 120 plants died and that we want to include all possible interactions, sketch the skeleton ANOVA table for this study making sure to include df. [5pts]

h) In the ANOVA table from above each effect has an associated F-test.
   i. What F-test would be most appropriate to use to test whether the difference between the two species, averaged over the sites, is similar or different for the two experiments (locations)? [2pts]
   ii. What F-test would you use to test if the pattern of interactions between species and sites is similar or different at the two experiments (locations)? [2pts]

2) **Pine Trees** [60pts]. An ecologist studying diseased pine trees takes data (a_i, Y_i), $i = 1, 2, \ldots, n,$ where a_i is the known size of the i^{th} randomly selected area and Y_i is the number of diseased pine trees in that area. Suppose we model the data by $Y_i \sim Poisson(a_i\theta)$, where the $Y_i s$ are assumed to be independent.

a) Explain why the $Y_i s$ are not a random sample as defined by Casella and Berger. Justify your response with appropriate evidence. [2pts]

b) Find the joint probability mass function. Show all work and state any assumptions and theorems used. [6pts]

c) Derive the maximum likelihood estimator (MLE) of θ^2. Show all work and state any assumptions and theorems used. (Note: You do not need to use the second derivative test to verify it is a maximum.) [9pts]

d) Use the joint probability mass function to show that $S = \sum_{i=1}^{n} Y_i$ is a complete, sufficient statistic for θ. Justify how we know this statistic is complete and sufficient. [6pts]

e) Use MGFs to derive the distribution of $S = \sum_{i=1}^{n} Y_i$. If S follows a known distribution, state the distribution, along with its parameters. Show all work and state any assumptions and theorems used. [9pts]

f) Verify that $T = \dfrac{S^2 - S}{\left(\sum_{i=1}^{n} a_i\right)^2}$ is the uniformly minimum variance unbiased estimator (UMVUE) of θ^2. Show and explain all work, being sure to explain why it is the UMVUE. [8pts]

g) How would you determine which estimator (MLE or UMVUE) to use for estimating θ^2? Discuss the criteria and/or process you would use, why you would use it (e.g., what information would the criteria and/or process provide), and how you would use the information provided to decide between the two estimators. (Note: You do not need to do any calculations to answer this question.) [10pts]

h) Another researcher proposes to model the number of diseased pine trees per acre in the i^{th} randomly selected area, Y_i, $i = 1, 2, \ldots, n,$ as a Poisson random variable with mean $\lambda > 0$. However, the mean varies from area to area, and its random behavior is modeled by a $Gamma(\alpha, \beta)$ distribution, i.e., $\Lambda \sim Gamma(\alpha, \beta)$.

Compare this modeling approach to the previous one where it was assumed $Y_i \overset{independent}{\sim} Poisson(a_i\theta)$, $i = 1, 2, \ldots, n.$ Which approach would you recommend using to model these data? Explain why you prefer that choice and not the other, being sure to discuss similarities and/or differences between the two approaches' assumptions and parameters within the context of the problem. [10pts]

3) **Diabetes** [30pts]. The prevalence of Type 2 diabetes in the population of American Indians (Native Americans) and Native Alaskans is estimated to be higher than in any other racial/ethnic population in the US. Because of high prevalence of this disease and its debilitating complications, it is of interest to identify what factors contribute to the prevalence of diabetes in the American Indian and Native Alaskan population.

In one such study by the National Institute of Diabetes and Digestive and Kidney Diseases, 768 adult female Pima Indians living near Phoenix were sampled. All participants received a diabetes test and 268 tested positive for diabetes.

a) Let Y_i be the diabetes status for woman *i* where $Y_i = 1$ if the woman tests positive for diabetes and 0 otherwise. Of interest is the proportion of women testing positive for diabetes. Carefully explain what distribution $D = \sum Y_i$ follows and what assumptions you are making in context of the problem. [7pts]

b) According to epidemiologists, diabetes is at an epidemic level if more than 15% of a population has diabetes. The epidemiologists calculate $P\left(\sum_{i=1}^{n} Y_i \geq 268 \middle| p = .15, n = 768\right) = P\left(D \geq 268 \middle| p = .15, n = 768\right) < 0.0001.$

i. Set up (but do not evaluate) the expression used to obtain the probability the epidemiologists calculated. [6pts]

ii. Explain (or interpret) what this probability represents in context of the problem. [6pts]

iii. You are asked to report to the tribal council whether diabetes should be considered an epidemic. Does the probability calculated by the epidemiologists answer the question of interest? If so, explain if there is evidence to suggest a diabetes epidemic. If not, explain what information you would need to answer the question. [5pts]

c) The epidemiologists calculate $\hat{p} = 0.349,$ but they are unsure about which estimate of variability to report with this value. Should the epidemiologists report the standard deviation of diabetes status or the standard error? Explain what each of these estimates of variability represent and why they should report the estimate you chose and not the other. [6pts]

4) **Diabetes Revisited** [40pts] Recall that the major purpose of the study in Problem 3 is to investigate what factors are related to the prevalence of developing Type 2 diabetes. In this study, the following variables were also collected:

Variable Name	Description
`Pregnant`	Number of times pregnant
`Glucose`	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
`Diastolic`	Diastolic blood pressure (mm Hg)—the bottom number in a blood pressure reading
`Triceps`	Triceps skin fold thickness (mm)
`Insulin`	2 hour serum insulin (mu U/ml)
`Bmi`	Body mass index (weight in kg/(height in meters squared))
`Diabetes`	Diabetes pedigree function
`Age`	Age in years
`Test`	Test of whether the patient shows signs of diabetes (0=negative, 1 = positive)

Previous research indicates that age, BMI, and blood pressure often are useful in predicting whether or not a person has diabetes. Let's consider using these three explanatory variables. For the following questions there is output included in the handout that may or may not be useful.

a) Write out the hypothesized (or population) model of interest. You do not need to define the parameters. [4pts]

b) When looking at the output provided, notice that the option link="logit" was used. Explain what a link function is and why a logit link function was (likely) used. [5pts]

c) Provide an interpretation for the coefficient associated with BMI in context of the problem. [5pts]

d) Yolanda, a Pima Indian not part of the original study, is a 54 year old woman with diastolic blood pressure of 100 mm Hg and a BMI of 38. What is the probability that she tests positive for diabetes? Note that you do NOT need to simplify your answer but show your work! [4pts]

e) Researchers are interested in identifying at which BMI the probability of having diabetes exceeds 10% for the "average" woman. In this sample average diastolic blood pressure is 69.11 and the average age is 33.24. Calculate the BMI for the "average" woman such that the probability of testing positive for diabetes is 10%. Note that you do NOT need to simplify your answer but show your work! [4pts]

f) What assumption is being made regarding the variance of the response? Explain, using the output provided, if it appears that this assumption is reasonably satisfied. [6pts]

g) Based on previous research, the effect of one of the explanatory variables on the presence of diabetes depends on the values of the other variables. Because of this you recommend that all two-way and three-way interactions be included in the model and decide to conduct a hypothesis test of whether or not the interactions are needed in the model. To test whether the inclusion of the interactions terms is necessary, we typically use a likelihood ratio test.

   i. Set up the hypotheses being tested when testing if the interaction terms are useful in modeling the probability of having diabetes. [4pts]

   ii. The test statistic associated with testing the hypothesis above is -2(log likelihood of the null model – log likelihood of the alternative model). Based on large sample theory, what distribution (including parameters) does this test statistic follow? [4pts]

   iii. If possible calculate this test statistic (at this point it would be very helpful to know that residual deviance = -2log likelihood). If not possible, explain what additional information is needed and how you would obtain that information. [4pts]

1) Potentilla - Output

```
> pont.anova2<-lm(lnwt~spp*site,data=pontilla.cp)
> summary(pont.anova2)

       Call:
       lm(formula = lnwt ~ spp * site, data = pontilla.cp)

       Residuals:
          Min     1Q Median     3Q    Max 
       -1.228 -0.358 -0.008  0.272  1.422 

       Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
       (Intercept)       3.7680     0.1676  22.485  < 2e-16 ***
       sppnive           0.2800     0.2370   1.181 0.242898    
       siteMid           0.5020     0.2370   2.118 0.039052 *  
       siteWet           0.9520     0.2612   3.645 0.000626 ***
       sppnive:siteMid  -1.1690     0.3352  -3.488 0.001012 ** 
       sppnive:siteWet  -1.7620     0.3527  -4.996 7.24e-06 ***
       ---
       Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       Residual standard error: 0.5299 on 51 degrees of freedom
       Multiple R-squared:  0.4845,    Adjusted R-squared:  0.4339 
       F-statistic: 9.585 on 5 and 51 DF,  p-value: 1.712e-06


> #type I SS
> anova(pont.anova2)

       Analysis of Variance Table

       Response: lnwt
                 Df  Sum Sq Mean Sq F value    Pr(>F)    
       spp        1  5.9133  5.9133 21.0564 2.934e-05 ***
       site       2  0.0858  0.0429  0.1527    0.8588    
       spp:site   2  7.4604  3.7302 13.2827 2.273e-05 ***
       Residuals 51 14.3224  0.2808                      
       ---
       Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> #type II SS
> Anova(pont.anova2,type=2)

       Anova Table (Type II tests)

       Response: lnwt
                  Sum Sq Df F value    Pr(>F)    
       spp        5.9269  1 21.1049 2.882e-05 ***
       site       0.0858  2  0.1527    0.8588    
       spp:site   7.4604  2 13.2827 2.273e-05 ***
       Residuals 14.3224 51                      
       ---
       Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #type III SS
> Anova(pont.anova2,type=3)

       Anova Table (Type III tests)

       Response: lnwt
                    Sum Sq Df  F value    Pr(>F)
       (Intercept) 141.978  1 505.5651 < 2.2e-16 ***
       spp           0.392  1   1.3959  0.242898
       site          3.808  2   6.7799  0.002449 **
       spp:site      7.460  2  13.2827 2.273e-05 ***
       Residuals    14.322 51
       ---
       Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4) Diabetes Revisited - Output

```
> summary(pima)

           pregnant         glucose        diastolic         triceps
        Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
        1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
        Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
        Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
        3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
        Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00

           insulin           bmi           diabetes           age
        Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
        1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
        Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
        Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
        3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
        Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00

             test
        Min.   :0.000
        1st Qu.:0.000
        Median :0.000
        Mean   :0.349
        3rd Qu.:1.000
        Max.   :1.000


> m1<-glm(test~age+diastolic+bmi,data=pima,
+ family=binomial(link="logit"))
> summary(m1)

       Call:
       glm(formula = test ~ age + diastolic + bmi, family = binomial(link = "logit"),
           data = pima)

       Deviance Residuals:
           Min       1Q   Median       3Q      Max
       -2.0174  -0.8893  -0.5755   1.0987   2.6928

       Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
       (Intercept) -5.070673   0.538328  -9.419  < 2e-16 ***
       age          0.049073   0.007213   6.803 1.02e-11 ***
       diastolic   -0.009120   0.004695  -1.943   0.0521 .
       bmi          0.103889   0.012942   8.027 9.98e-16 ***
       ---
       Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       (Dispersion parameter for binomial family taken to be 1)

           Null deviance: 993.48  on 767  degrees of freedom
       Residual deviance: 871.89  on 764  degrees of freedom
       AIC: 879.89

       Number of Fisher Scoring iterations: 4
```