**Texas A&M University**

**August 2015 Diagnostic Exam Study Group**

**Discussion of Previous Exams**

This document is for reviewing and discussing the results of previous diagnostic exams, which are stored on the eCampus page. Try to keep discussion for each exam on separate pages for simplicity’s sake. I’ve just put a few here; add more as needed. Thanks!

Try to keep your questions/contributions to the discussion clear by signing your “posts” (perhaps with color-coding), and indenting your responses beneath the appropriate post.

This doc is only for discussing previous diagnostic exams; the other Google doc is for housing study notes for potentially relevant course material.

**Grading format (told from Aug 2014 Diag. survivor) = graded in A/B/C/D per problem and overall a Pass/Fail grade. In case you flunk one out of four problems, it would technically still be fine to pass. - Monica Huynh 6/10/2015**

**NOTE: Since Categorical Data Analysis is now an elective, there doesn’t seem to be much point in reviewing 2013 or earlier qual exams, which tend to be a lot heavier on categorical data. -Sean**

I think it might be good to also use this document for any problem from a pre 2013 test which is worth discussing. -Sen Zhang 08/02/2015

Jan 2012 started. Also, a table of contents might be helpful at this point, if anyone is inspired to create one. - Chris Shannon 8/11/15

**Jan 2013 exam started further down. --Sean W. 8/12/15**

**Jan. 2015 Exam Discussion**

**Part 1 (started by Sean Wisnieski)**

**from STAT 642 Handout 9 pg 24 and 25 Monica Huynh 7/7/15**

**1.** I completed the table as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| SOURCE | DF | MS | EMS |
| Thickness | 2 | 0.0263 |  |
| Type | 2 | 319.1202 |  |
| Thickness\*Type | 4 | 0.1152 |  |
| Machine(Type) | 6 | 1.4935 |  |
| Thickness\*Machine(Type) | 12 | 0.0878 |  |
| Error | 27 | 0.0445 |  |

**2.** Let be the *l*th observation within the *ijk*th treatment combination, be the fixed effect of the *i*th thickness, be the fixed effect of the *j*th machine type, be the random effect of the *k*th machine within the *j*th machine type, and be an error which is iid N(0,) and independent of . Then, our model is

.

additional model assumptions: Greg L 7/15/15

distribution properties: (642 HO9, pg 5)

are idd N(0,)

are idd N(0,)

are idd N(0,)

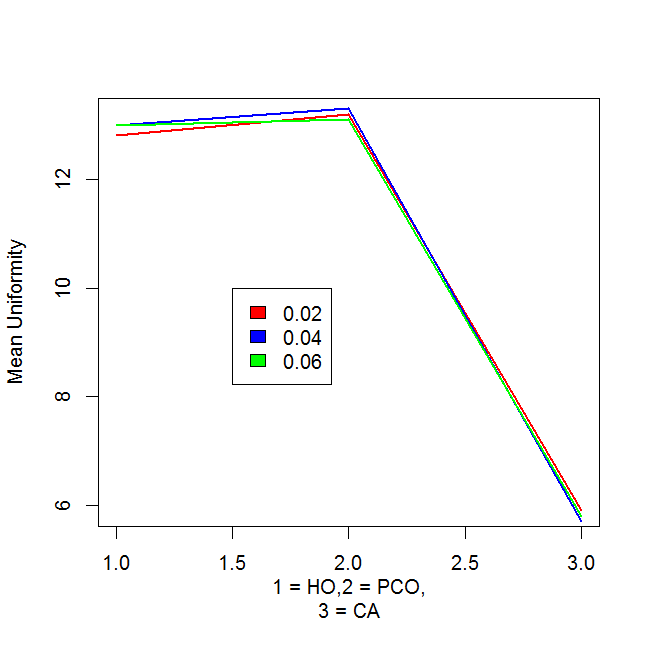
Monica 7/22/15

(Kuehl top of 239: “The interaction effects are assumed random when one of the factors involved is a random effect.”)

**3.** This is equivalent to testing a thickness-by-type interaction. If there is no true interaction, then this is equivalent to testing , in which case . Then, we can test the null using

with *df* = {4,12}, which is clearly *p* > .05. We therefore fail to reject and conclude that the mean uniformity of the windings produced by the three types of Machines are consistent across the three thicknesses. **(This feels wrong, just from looking at the table. Can someone verify that I did this step correctly? --Sean)**

**This looks right to me, Sean, I also made the graph below of the mean uniformity of the windings vs machine type for the three different thicknesses - the lines look parallel to me so I would conclude no interaction just looking at the graph - Greg L**



**4.** If there is no effect due to machine type, then = 0, in which case . Then, we can test the null using with *df* = {2,6}, which is clearly *p* < .001. We therefore reject = 0 and conclude that there is a significant effect of machine type.

**5.** **a)** = 0.0445

0.0878 = ⇒ 0.0878 = ⇒

1.4935 = ⇒ = 0.234

= 0.0445 + + 0.234 = 0.30015

/ {} = 0.148

/ {} =0.072

/ {} =0.78

( Chris Shannon See Kuehl 247 for an example of estimating the variance components and their percentages.)

b) Recall that our model is .

Then, .

Hence, (recalling that fixed effects are discarded because the variance of a constant is 0)

, the square root of which is .

c) Again recalling our model above:

and

.

Then, - and

(once again discarding fixed effects)

-]

the square root of which is .

d) Again recalling our model above:

and

.

Then, . Remembering as always to discard fixed effects from the variance (since the contribution of a constant to the variance is 0), then

. Then, .

**SEAN 7/29/15:**

**SAS code for future reference:**

**data jan2015q1;**

**input type $ mach $ thick $ y;**

**cards;**

**HO 1 .02 12.30**

**HO 1 .02 12.59**

**HO 2 .02 13.46**

**HO 2 .02 14.00**

**HO 3 .02 12.35**

**HO 3 .02 12.06**

**HO 1 .04 13.16**

**HO 1 .04 13.00**

**HO 2 .04 13.29**

**HO 2 .04 13.62**

**HO 3 .04 12.50**

**HO 3 .04 12.39**

**HO 1 .06 12.87**

**HO 1 .06 12.92**

**HO 2 .06 13.46**

**HO 2 .06 13.82**

**HO 3 .06 12.73**

**HO 3 .06 12.15**

**PCO 4 .02 13.01**

**PCO 4 .02 12.63**

**PCO 5 .02 13.46**

**PCO 5 .02 13.92**

**PCO 6 .02 13.15**

**PCO 6 .02 13.20**

**PCO 4 .04 12.74**

**PCO 4 .04 12.68**

**PCO 5 .04 13.84**

**PCO 5 .04 13.75**

**PCO 6 .04 13.46**

**PCO 6 .04 13.57**

**PCO 4 .06 12.47**

**PCO 4 .06 12.15**

**PCO 5 .06 13.62**

**PCO 5 .06 13.28**

**PCO 6 .06 13.36**

**PCO 6 .06 13.42**

**CA 7 .02 5.47**

**CA 7 .02 5.96**

**CA 8 .02 5.75**

**CA 8 .02 5.68**

**CA 9 .02 6.24**

**CA 9 .02 6.15**

**CA 7 .04 5.73**

**CA 7 .04 5.64**

**CA 8 .04 5.60**

**CA 8 .04 5.65**

**CA 9 .04 5.92**

**CA 9 .04 5.64**

**CA 7 .06 5.01**

**CA 7 .06 5.62**

**CA 8 .06 5.80**

**CA 8 .06 5.71**

**CA 9 .06 6.19**

**CA 9 .06 6.23**

**;**

**run;**

**proc glm data=jan2015q1;**

**class thick type mach;**

**model y = thick|type|mach(type);**

**random mach(type);**

**run;**

**Part 2 (started by Sean Wisnieski)**

**from STAT 641 - Monica Huynh 7/7/15**

**1.** 95% CI:

**2.** I estimated *n* = 266 at least, using to generate an initial estimate of *n* = 264, then refined to 266 when it was apparent that such an interval was too small to generate *t*-based intervals of width under 4. Nevertheless, *n* = 265 generates an interval of approximately (but barely over) 4 days’ width.

I’m getting n= 66 using equation on Handout 11 page 26. The plus sign in your formula is throwing me off or my math is off :) - Monica 7/18/15



**(This looks more correct; I have no idea where I got that original formula from! Nevertheless the principle of getting an initial estimate using the formula, then trying it out to make sure the width of the intervals don’t exceed , should be followed). --Sean W.**

I think the discrepancy may be due to the interpretation of the delta in the formula on pg 26 of HO 11. The formula is used to calculate the number of units needed to insure that the estimate of the parameter is within Δ units of the true value of θ with 100(1 − α)% confidence. In that case, I think Δ is 2 instead of 4, which gives a value of n = 264. Greg L 7/20/15

**(Oh, yes, it should definitely be 2 -- the intervals extend on both sides of the mean, after all. --Sean 7/21/15)**

**Thanks for the clarification. Monica 7/21/15**

**3.** The survival times are extremely nonnormal, as witnessed by both the Shapiro-Wilk *p*-value (*p* < .001 despite the small sample size) and the very nonnormal quantile reference plot, which implies a fairly skewed distribution. This is to be expected: survival times are usually gamma-distributed.

**4.** (If I understand the notes right): We can draw 10,000 bootstrap samples with replacement from the data and calculate the mean for each, then order the resulting set of means from smallest to largest. We can then pull the 250th and 9750th means in the ordered set (corresponding to the 2.5th and 97.5 percentile), supply those means into the pivot, and solve for mu to derive *t*-values.

(Demonstrative R code follows; note the percentiles this produces DO NOT match what the exam gives in the following question, although is kinda close; that’s probably fine.

x = c(0.2,0.8,1.4,2.0,2.3,2.6,3.5,3.6,4.8,5.0,5.3,5.4,5.6,6.7,8.0,8.8,13.1,

15.4,16.1,18.9,26.1,32.4,38.7,42.3,68.7)

n = length(x)

thest = mean(x)

B = 9999

thestS = numeric(B)

thestS = rep(0,times =B)

for (i in 1:B)

thestS[i] = mean(sample(x,replace=T))

RS= sort(thestS-thest)

LRS = RS[250]

URS = RS[9750]

thL = thest-URS

thU = thest-LRS

lcl.t = sqrt(n)\*(thL - thest)/sd(x)

ucl.t = sqrt(n)\*(thU - thest)/sd(x)

lcl.t

ucl.t

Your Rcode shows that the standard deviation is calculated from the original data - sd(x). Should it be the standard deviation of the 250 and 9750th bootstrap sample? Monica 7/18/15

**Sean 7/20/15: Yeah, Monica, that makes more sense, and gets t-values much closer to the ones given in the exam output. Revised code below, for reference’s sake:**

**x = c(0.2,0.8,1.4,2.0,2.3,2.6,3.5,3.6,4.8,5.0,5.3,5.4,5.6,6.7,8.0,8.8,13.1,**

**15.4,16.1,18.9,26.1,32.4,38.7,42.3,68.7)**

**n = length(x)**

**thest = mean(x)**

**B = 9999**

**mm = NULL**

**T = NULL**

**i = NULL**

**for (i in 1:B) {**

**mm = sample(x,replace=TRUE)**

**T[i] = (sqrt(n)\*(mean(mm) - thest))/sd(mm)**

**}**

**T = sort(T)**

**LT = T[250]**

**UT = T[9750]**

**c(LT,UT)**

**Greg L 7/20/15: was wondering if here we should do the parametric bootstrap since the question states that the distribution is a gamma? In that case, we would use the original data to get the mle estimates for the parameters, then generate 10k samples of n=25 random values from a gamma with those parameters. I gave that a try (the data is actually not a great fit to a gamma) and got t0.025 = - 3.02 and t0.975 = 1.46.**

**Sean W. 7/20/15: That seems reasonable, since it’s the only way to incorporate the mentioning of the gamma distribution. But like you say, the data’s not a great fit and using bootstrap method yields *t*-values closer to what are indicated in the exam. Anyone else have thoughts? Maybe either way would work.**

**Should we calculate the test statistics t= , instead of mean from the bootstraped sample? The questions is asking about the percentile of the distribution of t, not the distribution of mean. I think we should construct a distribution of the test statistics: t instead of the mean. Jingjing 8/6/2015**

**Yes; that’s what my revised code (in dark red above) does. --Sean W. 8/11/15**

**5.** I just used and solved for , getting (7.075, 25.9125).

**I am getting different answer for this one. -3.75\*16.55/ 1.75\*16.55/ , and the interval that I got is (1.0875,19.3). I double checked the calculation. Jingjing 8/6/2015**

**Let’s walk through my steps and see where we diverge. Starting with , we multiply all sides by 16.55 to get , then divide both sides by 5 to get , then subtract 13.5 (x-bar) from both sides to get . Then we just divide both sides by -1 and flip the signs around to get the limits I got above: (7.7075, 25.9125). --Sean W. 8/11/15**

The result Jingjing has comes from using the typical Wald Interval: . - I don’t feel this method works, and I have the same results as Sean. D.Bartkowiak 8/12/15

**I agree with Sean as well. The quantiles are for the distribution of the t statistic shown so it should be used to get the CI for the mean. -Oscar P 8/12/15**

**6.** Huynh 7/18/15 - Sign Test Handout 12

Null hypothesis: u~ u0~

Alternative hypothesis: u~ u0~

p = P[Xi >0] = P[Yi > u0~] < 0.5

Xi = Yi  - u0~

S+ = # of positive Xi = 19

Reject null hypothesis if S+ Bn,1- Suge

Bn,1- = qbinom(,n,0.5) =qbinom(0.05,25,0.5) =8

19 greater than 8 so can not reject null hypothesis that the new treatment has increased the survival time in comparison to the previous treatment

**Sean W. 7/20/15: I think it should be qbinom(1-a,n,0.5), which would give 17 rather than 8 (still fail to reject null).**

I think if you use qbinom(1-a,n,0.5) which is H1 on page 36, you would actually reject the null and accept the alternative hypothesisu~ u0~. I used the setup for H2 on page 36. Monica 7/21/2015

pvalue= P[B(n,0,5) S+ ] = pbinom(S+ ,n,0.5) = pbinom(19,25,0.5) = 0.9979613

**According to HO12, page 53**

**H0: u~ u~>3.**

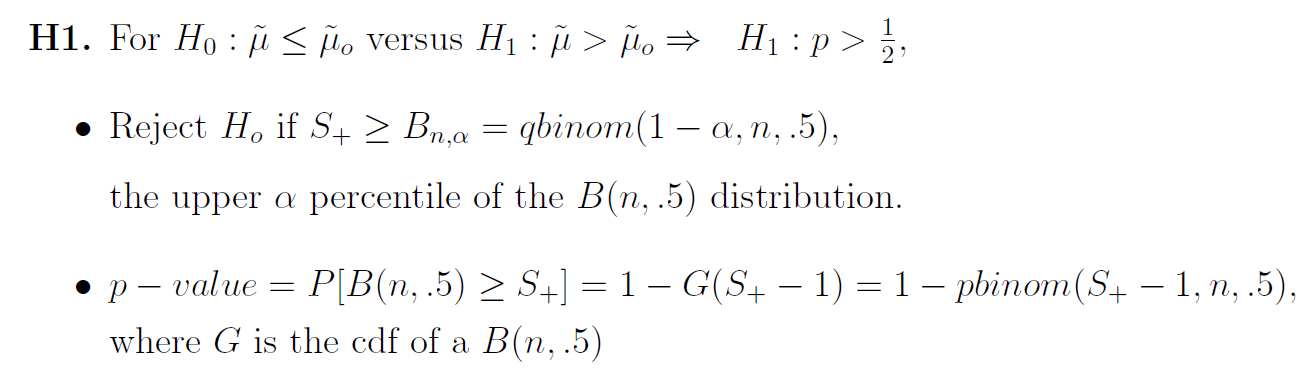
**reject region: Y > qbinom(.95, n=25, p0=0.5)= 17, Y=17 does not greater than 17 so fail to rejec**

**t H0. State that the new treatment has not increased the survival time compared to previous treatment.**

**p-value = p[Y >=17]=1-p[Y < 16]=1-pbinom(16,25,0.5)= 0.05388 → fail to reject H0.**

**[It looks like this discussion has moved on, but after reviewing the discussion above and checking back with HO12, it looks like we are supposed to reject HO. I get:**

**S+ = 19 and qbinom(1-0.05, 25, 0.5) = 17, with the decision-rule to reject H0 if S+ >= 18. Here’s a snip from the handout for setup H1 - and it does look like we need to use setup H1 because the question is asking us if the survival time has increased. Hence the null hypothesis should be u <= u0:**

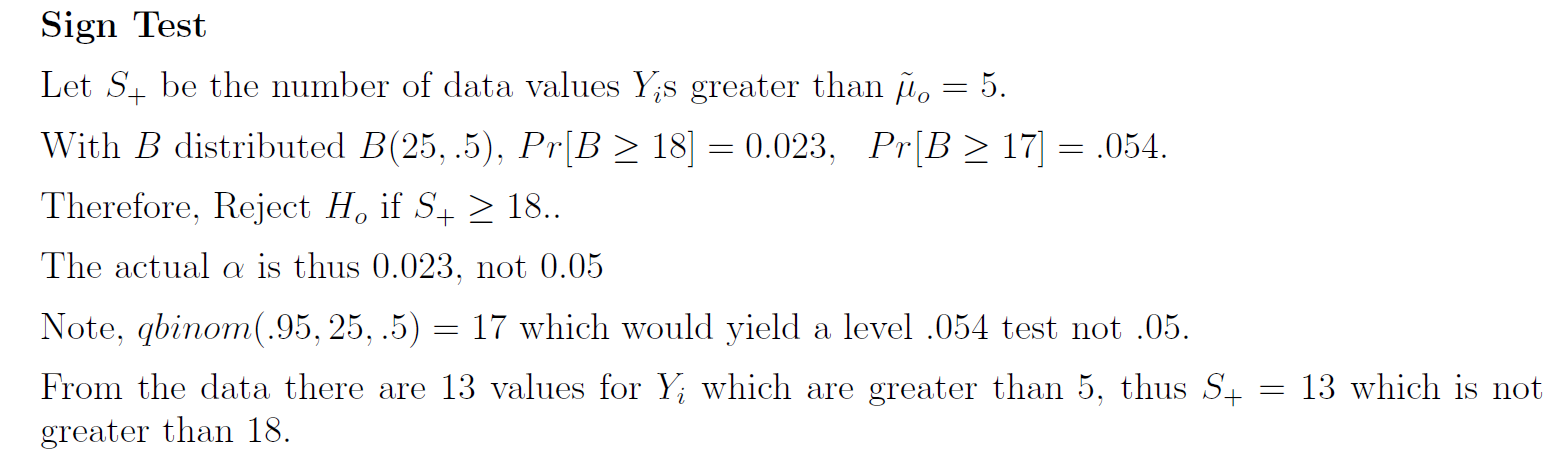


**Just to clarify my use of qbinom(1-0.05,25,0.5) = 17:**

**1-pbinom(17-1,25,0.5) = 0.054**

**1-pbinom(18-1,25,0.5) = 0.022**

**As per the example on page 43 in HO12, we should use 18 if (using >=, or 17 if using >) as our critical value because of the binomial dist being discrete:**



**- Chris Shannon 8/15/15]**

**Part 3 (started by Sean Wisnieski, 7/17/15)**

1. I believe this question is just asking for the expressions
   1. -D.Bartkowiak (7/20/15)

**(These should be X-bar rather than just x, right? --Sean W., 7/30/15)**

(I do not believe so as you would arrive at a single value if you used the mean value of the predictor -D.Bartkowiak, 8/7/15)

**2.**

**a.** is the mean response for subjects receiving treatment level when *x* is held

constant at 0.

**b.** is the difference between the mean response for subjects receiving treatment level and those receiving treatment level when *x* is held constant at 0.

**c.** is the difference between the mean response for subjects receiving treatment level and those receiving treatment level when *x* is held constant at 0.

**d.** is the average increase in the response for each one-unit increase in *x*, assuming the subject is receiving treatment level .

**e.** is the difference in the average increase in the response (for subjects receiving treatment level , compared to those receiving treatment level ) for each one-unit increase in *x*.

**f.** is the difference in the average increase in the response (for subjects receiving treatment level , compared to those receiving treatment level ) for each one-unit increase in *x*.

**3.** The slopes are equal iff the interaction between the treatment level and covariates is nonsignificant, i.e., if the treatment effects do not depend on the value of the covariate. Therefore, a useful test of the equality of the slopes is

**4.**

**(It would seem part 4 of this exam was cut off of this copy -- I posted in the discussion board asking about it, since the cover page says there are four problems, but have not yet received a response. --Sean W., 7/20/15)**

**Aug. 2014 Exam Discussion**

**Part A (started by Sean Wisnieski, 7/21/15) (Chris Shannon - Stephanie Lowe and I disagree with the nested factor. Number 4 and 5 are correct.)**

**1.** This looks like an RCBD with a split-plot treatment assignment. Containers are both the block and the whole-plot EU; one-quarter portions are the split-plot EU.

**2.** 4 x 4 x 3 factorial treatment structure, with container nested within sweetener and crossed with blending time. **(alternative below)**

**3.** Sweetener [S] - Fixed.

Container(Sweetener) [C(S)] - Random.

Blending Time [B] - Fixed.

**4.** Ice cream container is the (whole-plot) EU; one-quarter portion is the MU (and split-plot EU).

**5.** RV - texture score.

Covariates - none.

Subsampling - none.

Repeated Measures - none.

**6.**

|  |  |  |
| --- | --- | --- |
| SOURCE | DF | EMS |
| S | 3 |  |
| C(S) | 8 |  |
| B | 3 |  |
| S\*B | 9 |  |
| C(S)\*B = Error | 24 |  |

**7.**

**Alternative answer to Part A 1 - 7 (Chris Shannon on 7/26/15)**

**Not sure I agree with this -- separate containers are assigned to each sweetener (i.e., it’s not just the same three containers), so isn’t it nested? --Sean 7/27/15   
According to O Kuehl on page 243 (Ch 7.3), the levels of one factor will contain different levels of another. “In certain types of studies the levels of one factor, B, will not be identical across all levels of another factor, A. Each level of factor A will contain different levels of factor B. The levels of B are said to be nested within the levels of A.” But that doesn’t seem to be the case here, as the container itself is not regarded as a factor. Instead, the split plot seems closer to what is being asked here.**

**The motivation for the split plot design is when one factor “requires more experimental material for its evaluation than a second factor in factorial experiments.” (O Kuehl, Ch 14 page 469) And he gives some examples, which include applying different treatments to subplots within a whole plot and, for example, temperature control chambers with different temperatures in each one applying other treatments within each chamber.**

**He says, “If one factor, such as temperature, humidity, or photoperiod, requires environmental control chambers, then each level of that factor requires a separate chamber. A second treatment factor such as culture media for plants and even a third treatment factor can be included within each of the chambers, thereby achieving an economical use of the chambers to study more than one factor at a time. The chambers represent whole plots, and the subplots are units to which the second factor levels are applied inside each chamber.” (O Kuehl, page 471) -- Chris Shannon 7/27/15 @ 18:15 CST**

**(Hmmm. I didn’t cross check it again with your design, but it looks like we get the same degrees freedom and same E(MS) using either design. I wonder if either answer would be accepted? - Chris Shannon 7/27/15 @ 18:20 CST)**

looking at the example on 642 Handout 2 page 2, the levels of containers only have meaning if we know which sweetener the containers are selected from so wouldn’t it be nested? Monica 8/9/15

**When I did this problem before checking the study guide, I arrived at the same design as Chris. I just don’t see how the containers are being used as a factor. I was thinking that the 12 initial containers were supposed to be homogenous with respect to each other. In any case, we are not told of any difference between the containers assigned to each sweetener that would indicate levels to be concerned with. (I was thinking that any initial variation in the containers should be handled by an error term and not by factors) -Oscar P. 8/9/15**

**Looking back over the notes, I came upon the example on HO 12, pg. 15 about the CRD split-plot design. I believe that the containers in this problem are comparable to the classrooms in the example in the HO. -Oscar P. 8/10/15** Thanks! -Monica 8/10

1. **Completely randomized split plot design. (See O Kuehl ch. 14, page 480, table 14.7 first table.)**
2. **4 X 4 factorial treatment.**
3. **Fixed: Sweetener (S), Blending time (B); Random: Container (whole plot)**
4. **Experimental units: Containers; Measurement Units: portion of container.**
5. **Response variable: texture score; Covariates: none; Subsampling: None; Repeated measures: None**

**6)**

|  |  |  |
| --- | --- | --- |
| **Source Variation** | **df** | **E(MS)** |
| **Sweetener (S)** | **3** |  |
| **Error(1)** | **8** |  |
| **Blend time (B)** | **3** |  |
| **SxB** | **9** |  |
| **Error(2)** | **24** |  |
| **Total** | **47** |  |

**7)**

**is the general mean.**

**is the fixed effect of the ith level of sweetener.**

**is the random container effect with mean 0 and variance .**

**is the fixed effect of the jth level of blending time.**

**is the interaction effect of sweetener and blending time.**

**is the random experimental error with mean 0 and variance .**

**Necessary conditions:**

1. **and are independent, normally distributed with constant variance.**

**Part B (started by Sean Wisnieski, 7/20/15)**

1. **M 2)Q 3)N 4)E 5)P Sen Zhang, 08/02/2015**

**I got the same answers for all but #2. Why use Hsu’s procedure here when trying to determine which pairs of means are different across the levels of F1? I was guessing Tukey’s, but am not that confident in my answer either. The nested factor was confusing me. G Long 8/3/2015**

**I chose N for #2, Tukey’s applied to levels of F1 averaged over levels of other factors. I think the issue is that interaction effects between fixed level factors are not significant. Gabrielle G 8/3/15**

**I agree with both of you. N is the right answer. Thanks.**

**Sen 08/04/2015**

In 4, would you average over values if the 3-way interaction and all the 2-way interactions were significant? -D.Bartkowiak, 8/7/15.

**I chose S for #3 because the researcher “wants to determine if there are pairwise differences in the levels of F1”, which I interpret as the presence of simple effects. Since the F-test for F1 is not significant, there cannot be any pairwise differences. Along with D. Bartkowiak, I also chose D for #4 because of the presence of significant interactions. Since the “mean defective rate” is given as u\_ijk, and since there are interactions, you would produce a number of trend plots with those interactions against F1. - Chris Shannon 8/8/15.**

**I agree with Chris on #3 being S since it was already determined that there were no main effects for F1. However, for #4 I thought it would be E. My reasoning was that the operators were random meaning you can automatically average over them even if an interaction is present. You are then left with the two other factors which interact leaving you with having to do the comparison separately for the levels of F2.**

**-Oscar P. 8/9/15**

**I agree with Oscar’s analysis for #4 and now believe E to be correct - Chris Shannon 8/9/15.**

**Problem II.**

**1.**

**2.** RSS = 0.23471, MSE = 0.0782366667, therefore = 0.2797082.

**3.** is found in the diagonal of , which gives and .

I think the parameter variances you found here are using standard deviation instead of variance from #2. If I multiply by .278 I get your results, but if I multiply by .078 I get .086 and .01 respectively.

Gabrielle Griffin, 7/22/15

**Yes, good catch -- thanks! Corrected in code below as well. --Sean 7/27/15**

**I am not sure if this has already been addressed here, but we are looking for** and **so you need to take the square root of the variances found above- Oscar P. 8/12/2015**

**4.** When *x* = 3, we estimate a mean response of . Then, the 95% CI is   
 (1.11191, 1.90809). Therefore we can say with 95% confidence that the mean response at *x* = 3 falls somewhere between 1.11191 and 1.90809.

**5.** Similar to 5, but with formula slightly changed: =   
 = = (0.5349, 2.4852). Therefore we can say with 95% confidence that, when *x* = 3, a random realization *y* will fall somewhere between 0.5349 and 2.4852.

(I did the above by hand/calculator, but it can be verified with the following code:

*y <- c(0.34,1.28,1.16,2.11,2.66)*

*x.simple <- c(1,2,3,4,5)*

*x <- matrix(c(1,1,1,1,1,1,2,3,4,5),ncol=2,byrow=F)*

*beta <- solve(t(x) %\*% x) %\*% t(x) %\*% y*

*beta # regression coefficients*

*y.pred <- x %\*% beta*

*e <- y - y.pred*

*RSS <- sum(e^2) # residual sum of squares*

*MSE <- RSS / 3 # mean square error (sigma-squared)*

*RMSE <- sqrt(MSE) # root mean square error (sigma)*

*# variance-covariance matrix for regression estimates*

*MSE \* solve(t(x) %\*% x)*

*# proper regression analysis - verify that things add up*

*obs.model <- lm(y~x.simple)*

*summary(obs.model)*

*# confidence/prediction intervals*

*testvalue <- data.frame(x=3)*

*predict(obs.model,testvalue,interval="confidence")*

*predict(obs.model,testvalue,interval="predict")*

**The last two predict functions give me error message. Anyone successfully execute this code? --- Sen 08/04/2015**

**Sorry, I changed variable names a few times in the course of writing this code and never ran it “fresh.” I’ve corrected the errors and updated the code above so it should run correctly now. Thank you for catching that. --Sean 8/11/15**

# Question III (Chris Shannon on 7/26/2015)

## Part A

### 1)

H0: , HA: , p-value =

DO NOT Reject H0.

For this one, I got the p-value as 1 - P(X <= 5) = 1 - 0.939 = 0.061, therefore we do not reject H0. The p-value is the probability of a result as extreme or more extreme than the observed data, which is the probability that x is 6 or greater, which is 1 minus the probability that x is 5 or less. G Long 8/5/15

Thanx, G Long. I’ve made the changes. -- Chris Shannon 8/6/15

### 2)

Mid-p-value:

For this one, I get P(x = 6) = 0.982 - 0.939 = 0.043 and P(x =>7) = 1 - 0.982 = 0.018, so the mid-p value = 0.0395. The result is the same - reject H0. G Long 8/5/15

Thanx, G Long. Good catch. I should have used 0.982 = P(X <=6). Changes made - Chris S. 8/6/15

### 3)

where t(0.975, 15-1) = 2.145

**(Shouldn’t we prefer Wilson CI’s? --Sean 7/27/15)**

**(Not sure. It’s a bit complex, though. Would we be expected to remember that for the test? -- Chris 7/27/15 @ 18:53)**

**(It is very complex, yes…. hard to believe we’d be expected to remember it; but the notes say explicitly that a CI of this sort is inappropriate since the assumptions for it are not met. --Sean 7/27/15)**

**(But then np = 15(0.4) = 6 > 5 and n(1-p) = 15(0.6) = 9 > 5, so this could work. I guess the problems come about especially when the probability is in the tails - Chris 7/27/15)**

**(Fair enough; professors have indicated in the past that grading on this is not so tough, so we could simply present the Wald CIs as above and make a note that something like the Wilson CIs may yield more accurate results. --Sean 7/27/15)  
(Making that caveat with Wald CI sounds safest. -- Chris 7/27/15)**

What about using the A-C interval y\*=y+2,n\*=n+4,p\*=y\*/n\*? D.Bartkowiak 8/9/15

**Isn’t A-C recommended only when n>=40? -O. Pena 8/9/15**

## Part B

### 1)

H0: , HA: ,

DO NOT Reject H0.

### 2)

**Why use 0.4? --Sen 08/04/2015**

**Yes, 0.4 is a mistake, carried over from part A of the problem. Changing to 0.4 under the sqrt sign to 0.23 and updating the result. -- Chris 08/04/2015**

### 3)

When n is sufficiently large a binomial distribution approximates a normal distribution.

# Question IV (Chris Shannon on 7/26/2015)

### 1)

One major concern that potentially threatens the validity of the model fit by Miller (2013) is the low Adjusted R-squared: 0.45. This tells us that less than half the variability is explained by the model. This can lead to improper predictions and unexpected outcomes.

I would also say that variance is not constant, which is needed to make inferences from the model. GabrielleG 072915

I also put that there was a significant interaction between Day/Night and temperature as shown in the graph which is not accounted for in the model G Long 8/13/15

### 2)

The following is a recommendation: **Add interaction terms.**

1. Remove variables with high p-values step-by-step until I have a set of main effects whose p-values are significant.
2. Start adding interactions from the main effects.
   1. Add the most significant interaction effects one-by-one, adding the most significant effect first and then the next most significant others.
3. For each interaction added, reassess the fit of all the variables and then if some variables are no longer significant, they may be dropped.
4. Repeat steps 2 and 3 until no more significant interactions are found.

### 3)

Keep Saturday, Sunday, Tuesday, Thursday, Friday (because of their high means and high Q3s)  
Keep all months except October

Keep Bobblehead promotion and Opponents from Large Metro Areas.

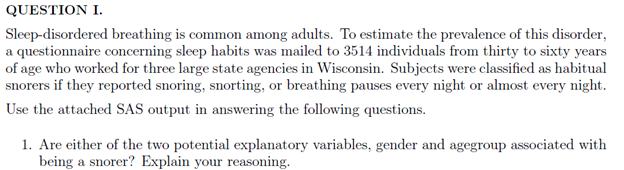
Keep Attendance vs. Temperature

# 

**Aug. 2013 Exam Discussion**

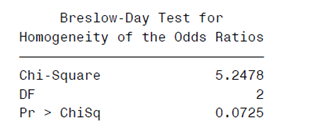
Discuss.

G Long - question 1



Yes, both age group and gender appear to be significant factors since the chi square statistic is greater than the critical value in both cases





Since the p-value > 0.05, we do not reject the null hypothesis that the odd ratios differ

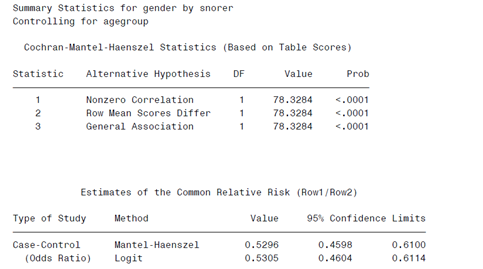
For the 30 to 39 age group, the 95% CI for the odds ratio is [0.4734 0.7646]

For the 40 to 49 age group, the 95% CI for the odds ratio is [0.4516 0.6981]

For the 50 to 59 age group, the 95% CI for the odds ratio is [0.2913 0.529]

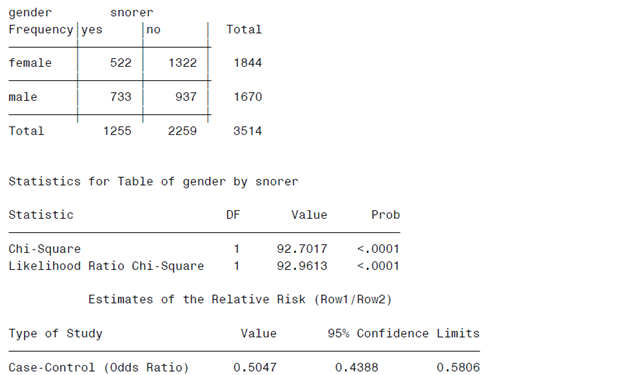
Since the confidence intervals all overlap, we conclude that the odds ratio for being a snorer between women and men do not differ significantly





There is significant evidence of conditional dependence of snoring and gender, controlling for age group



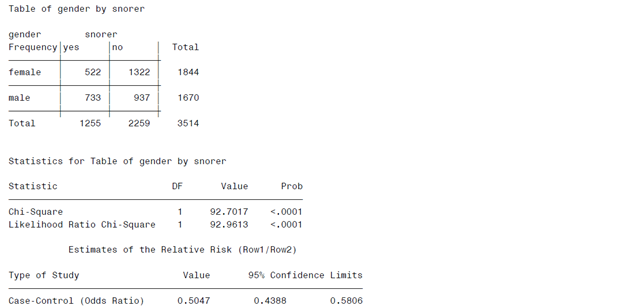


Per the chi-square test and the odds ratio test, there does appear to be evidence of dependence of snoring and gender, ignoring the effect of age group



Since we do not reject the hypothesis that the odds ratios differ, it is more appropriate to use the test in part 4. This will result in greater power to reject the null if it is in fact false





Ψ\_hat = 522\*937/(1322\*733) = 0.5047

Log(Ψ\_hat) = -0.6836959

MSE(Log(Ψ\_hat)) = (1/(n11)+1/(n12)+1/(n21)+1/(n22))^.5 = 0.0714397

CI (Log(Ψ\_hat)) = 0.5047 +/- 1.96\*0.0714397 = [-0.8237177, -0.543674]

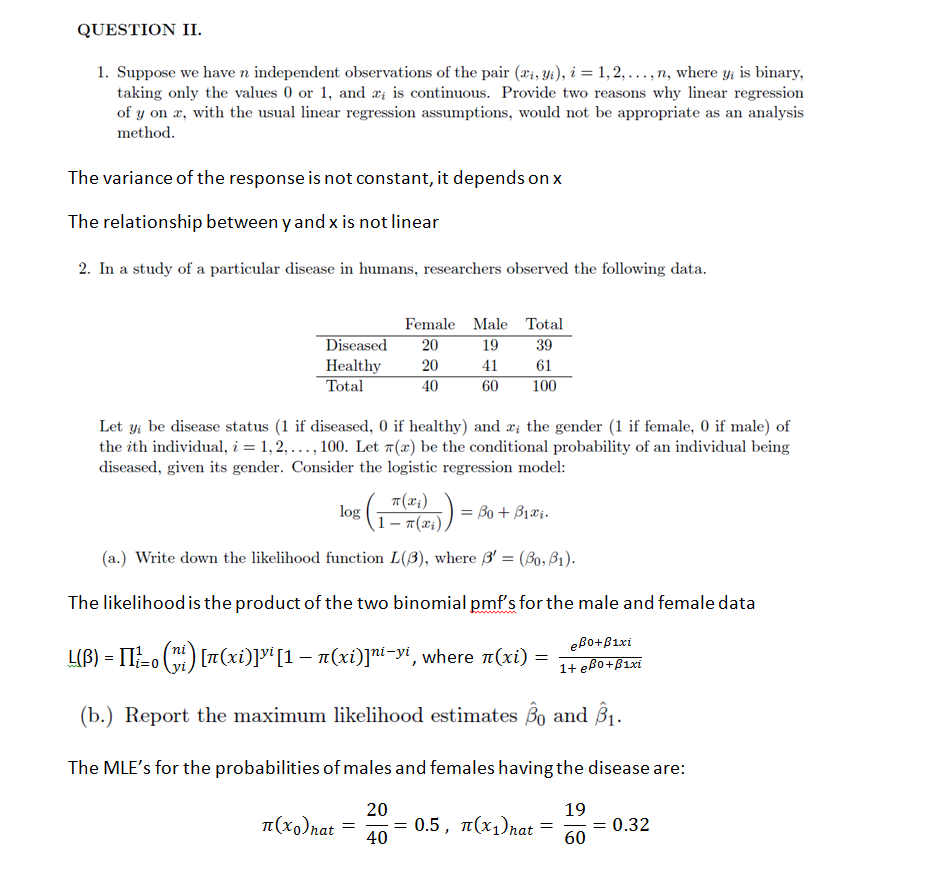
CI(Ψ\_hat) = [exp(-0.8237177), exp(0.543674)] = [0.4387973, 0.5806111]

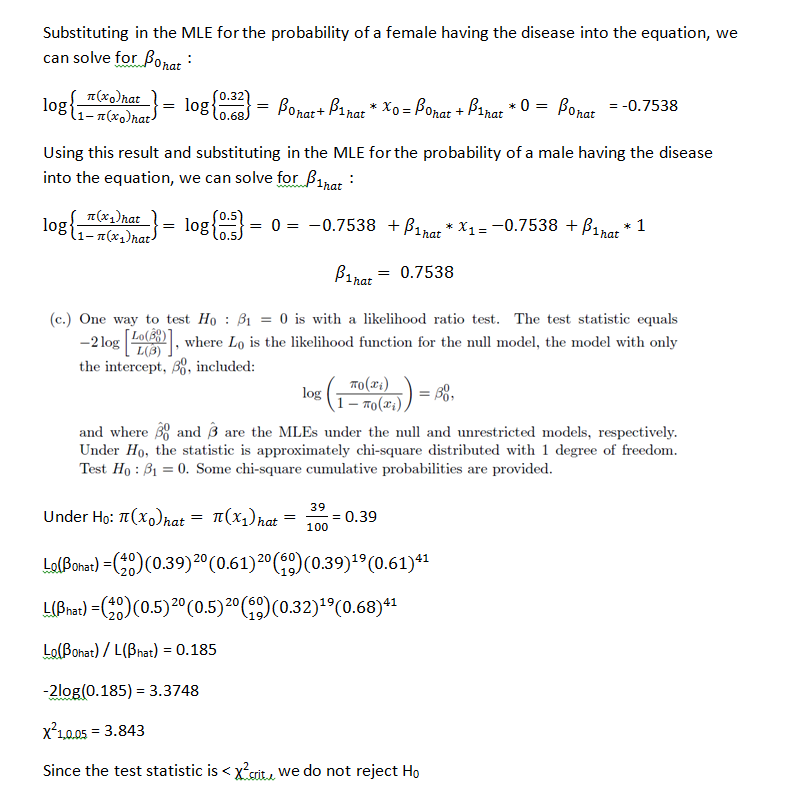
**(For a common odds ratio, wouldn’t you use the Cochran Mantel-Haenszel Statistic [0.4598, 0.6100] or corresponding logit statistic [0.4604, 0.6113] for an assumed common odds ratio? Chris Shannon - 7/28/15)**

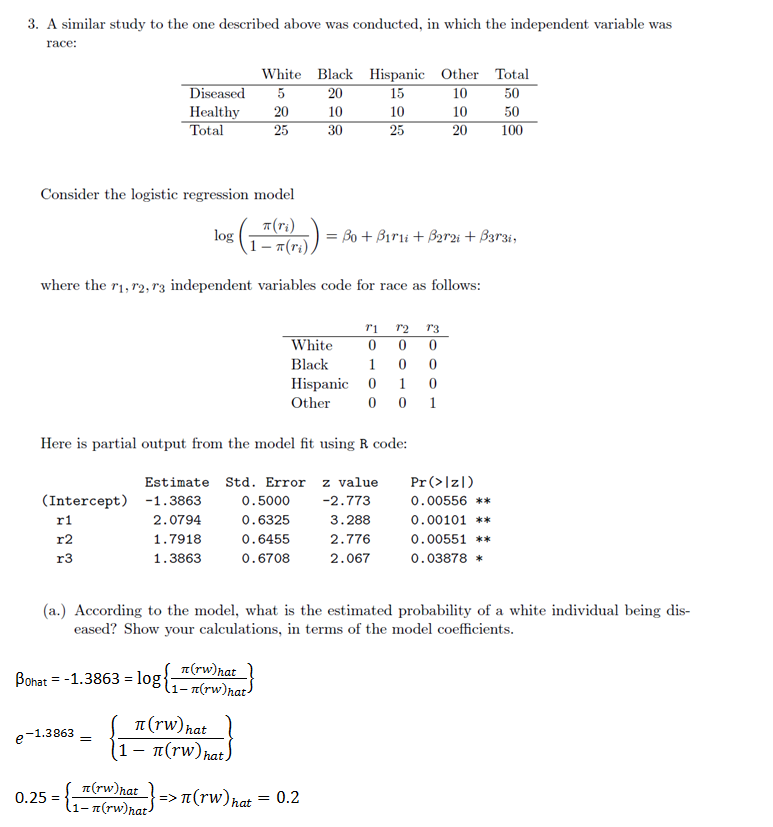


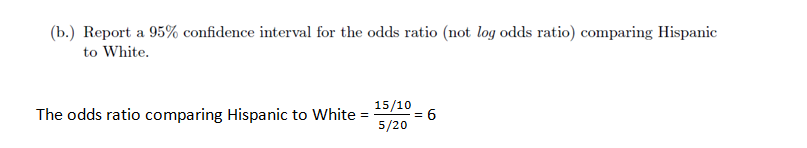
Simpson’s paradox refers to a situation where a lurking variable (one not considered in the data analysis) leads to an incorrect interpretation of the data. For example, if we averaged the data over the age group variable, it might lead us to conclude that gender was a significant predictor of snoring, but if we looked at the data separately for each age group, we might find that gender was not a significant predictor within each age category. That does not appear to be the case here – gender is significant both overall and within each age group.

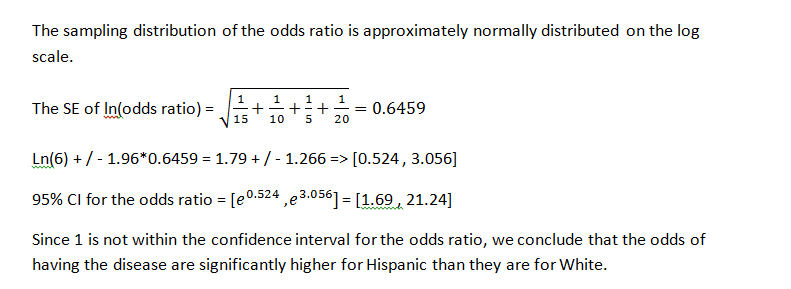
G Long - question 2







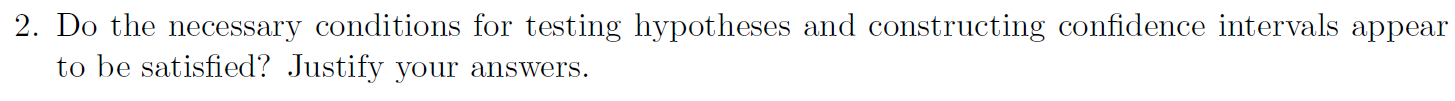




# Question IV (Chris Shannon - 7/29/2015)

### 

|  |  |
| --- | --- |
| Term | Definition |
|  | General mean. |
|  | Fixed effect of i-th level of food. |
|  | Fixed effect of j-th level of method. |
|  | Fixed effect of ij-th level of food\*method. |
|  | Random effect of kth-level of laboratory in method, distributed iid . |
|  | Random effect of ik-th-level of food\*laboratory in method, distributed iid . |
|  | Random effect of experimental error. |



### C1 Normality:

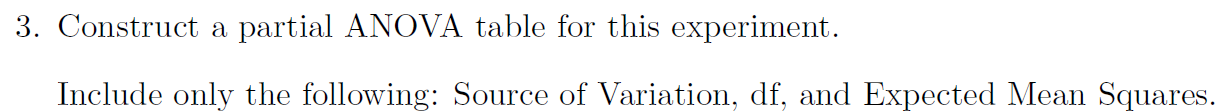
The QQ-plot and histogram both suggest normality, with some slight left-skewing in the data set.

### C2 Equal Variance:

The predicted vs. residuals plot seems random, which suggests equal variance condition is met. No patterns stand out that might indicate unequal variance.

### C3 Independence:

Independence is guaranteed by the randomization design.



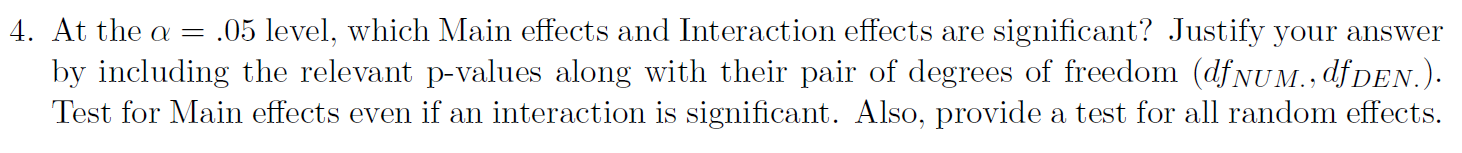
|  |  |  |
| --- | --- | --- |
| Source Variation | DF | E(MS) |
| Food | (3-1) = 2 |  |
| Method | (4-1) = 3 |  |
| Food\*Method | (3-1)(4-1) = 6 |  |
| Lab/Method | 4(5-1) = 16 |  |
| Food\*Lab/Method | 4(3-1)(5-1) = 32 |  |
| Error | 3\*4\*5(2-1) = 60 |  |

degree of freedom of Lab(Method)=4(5-1) = 16 Jingjing Liu 8/6/15

Whoops, yes. Corrected. Leaving it red. Thanx! Chris Shannon 8/7/15

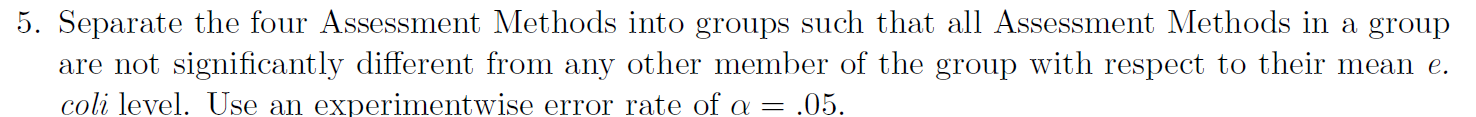
(Rule-Table for constructing E(MS) - not asked for on exam - see O Kuehl, Ch 7.6, page 255 - 258)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | F | F | R | R |
|  |  | 3 | 4 | 5 | 2 |
| sv name | sv | i | j | k | l |
| Food |  | 0 | 4 | 5 | 2 |
| Method |  | 3 | 0 | 5 | 2 |
| F\*M |  | 1 | 1 | 5 | 2 |
| Lab/M |  | 3 | 1 | 1 | 2 |
| F\*L/M |  | 1 | 1 | 1 | 2 |
| Error |  | 1 | 1 | 1 | 1 |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Effect | Hypothesis | Rule | Test statistic | p-value |
| Food |  | If F0 > F(0.95, 2,60) then reject H0 |  | 0.0002 |
| Method |  | If F0 > F(0.95, 3,60) then reject H0 |  | 0.0014 |
| Food \* Method |  | If F0 > F(0.95, 6,60) then reject H0 |  | < 0.0001 |
| Lab/Method |  | If F0 > F(0.95, 16,60) then reject H0 |  | 0.3705 |
| Food\*Lab/Method |  | If F0 > F(0.95, 32,60) then reject H0 |  | 0.2734 |

I believe these denominator df’s should be 32, 16, 32, 32, 60 from top to bottom. -D.Bartkowiak 8/15/2015



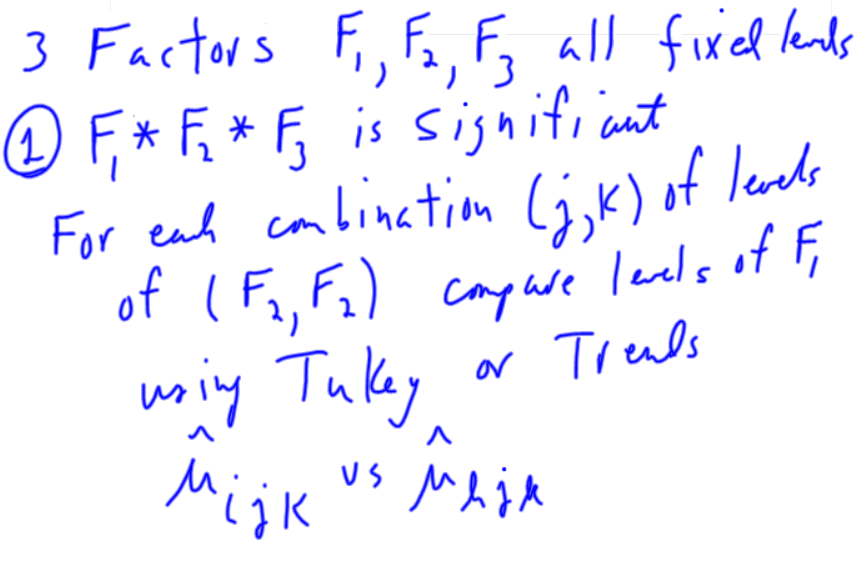
Since the interaction between Method \* Food is significant, I think we should group the Method at each level of Food.

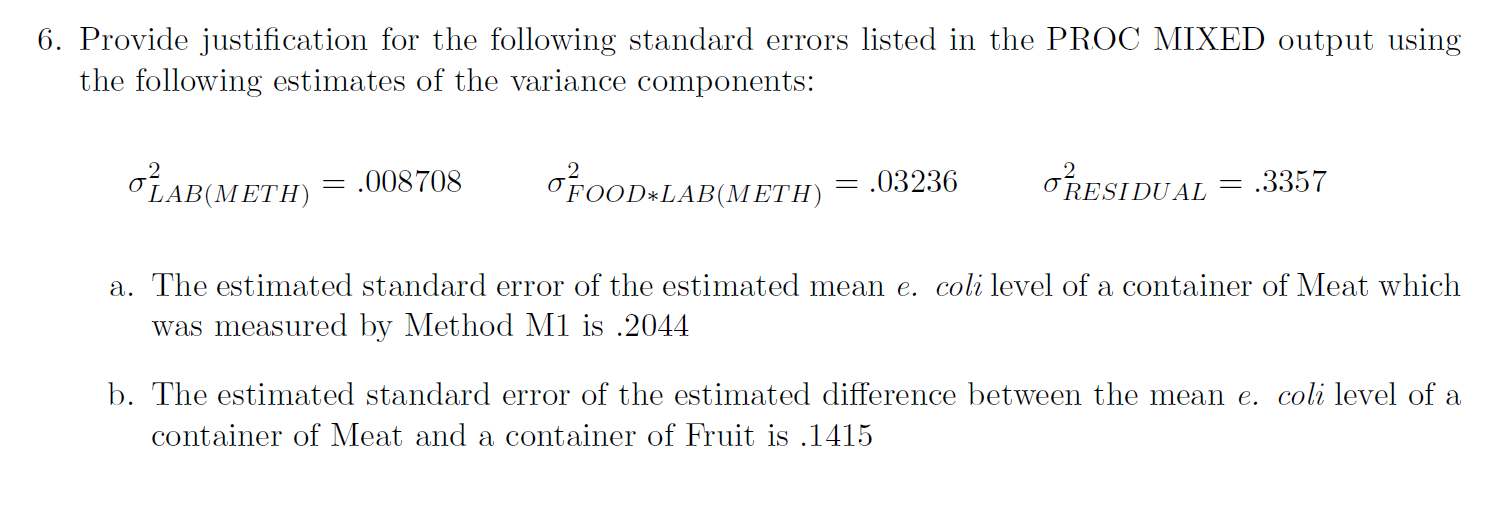
My answer: for Meat, Group 1: {M1, M3}, {M2, M3}, {M2,M4}

for Fruit: Group 1 {M1,M2, M3, M4}

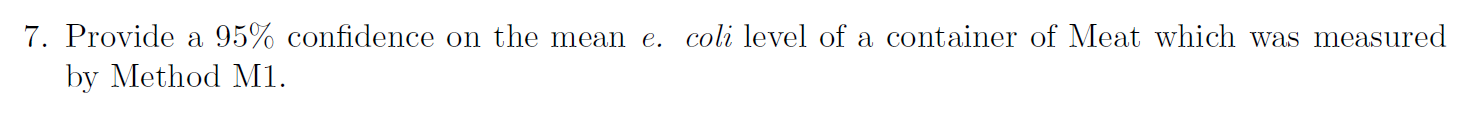
for VEG: Group 1: {M3,M4}, {M4,M2,M1} Jingjing Liu 8/6/2015

Jingjing, I think your answer would also be accepted. But the way I read the question is that they are asking us to compare the marginal means of Method, which is essentially telling us that we are not interested (for the purposes of the question) in the conditional means. Anyone else have an opinion about this? - Chris Shannon 8/7/15

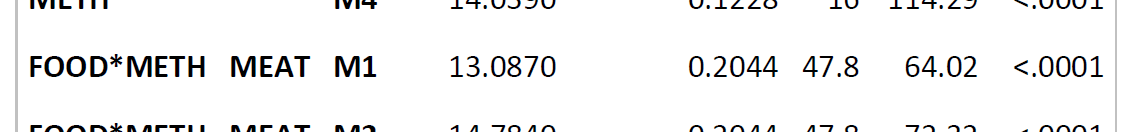
Further update: I agree with your answer now. Found this in lecture 30 from Dr. Longnecker:  
  








Take the estimate (13.0870), standard error (0.2044), and degrees of freedom (47.8) from the SAS LS Means printout:



And then calculate the confidence interval. To calculate the t-value from the chart for 48 degrees freedom, take t(0.975, 40) - t(0.975,60) = 2.021 - 2.000 = 0.021, multiply this by the proportion of the difference between 40 and 60 degrees of freedom that constitutes 48 degrees freedom, and then subtract from t(0.975,40) =>

So,

**2014 Jan**

Problem I

Part A

1. Type of Randomization

RCBD with split-split-plot design I think it’s actually CRD with split-split-plot. If Field were a blocking factor, it would have to be included as a source of variation in the ANOVA table, but there aren’t enough total df to include it. Also, in a complete block design each block would receive every treatment, which doesn’t happen here because of each field only gets one variety of wheat; if anything, it would be a BIBD, but it doesn’t meet all of the criteria. (Kate Mullan 8/10/15)

**Kate makes a great point about this. I think it is a CRD as well. It just seems that many times field is used as a blocking factor that it is easy to jump to that conclusion. -Oscar Pena 8/10/15**

2. Type of Treatment Structure

Varieties (2) \* Treatments (3) \* Amount (2) crossed factorial structure

3. Identify each of the factors as being Fixed or Random

Fields-random, blocking factor

Varieties-Fixed

Treatments-Fixed

Amount-Fixed

I think field should be nested within variety, Field(variety). Jingjing 8/14/2015

4. Describe the Experimental Units and Measurement Units.

EU for Varieties is Field; EU for Treatment is tract; EU for Amount is half of tract.

MU is half of tract

5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures.

Response Variable is yield of wheat. No Covariates, No SubSampling, No Repeated Measures.

6. An ANOVA Table with just the following information: Sources of Variation and Degrees of Freedom.

|  |  |
| --- | --- |
| Sources of Variation | Degrees of Freedom` |
| Varieties | 2-1=1 |
| Treatments | 3-1=2 |
| Amount | 2-1=1 |
| Varieties\*Treatments | 1\*2=2 |
| Varieties\*Amount | 1\*2=2 |
| Treatments\*Amount | 2\*1=1 |
| Varieties\*Treatments\*Amount | 1\*2\*1=2 |
| Error | 2\*3\*2=12 |

I also have error for each split - see page 480 in Kuehl text

error1 df=(r-1)(a)=(4-1)(2)=6

error2 df=(a)(r-1)(b-1)=12

error3 df=(ab)(r-1)(c-1)=18

total df=n-1=47

GabrielleG 080915

I got the same answer as Gabrielle, Jingjing 08/14/15

Part B

1.C 2.P 3.M 4.P 5.S 6.I

**I obtained the same responses for Part B -Oscar P. 8/10/15**

**I also got the same responses for Part B, which is good -- I was a little alarmed by the duplicate answer! --Sean 8/11/15**

Problem II

1. β0=1.15235, β1=-2.62637, β2=-0.55213, β3=5.02318

2. Δy= β1 when z=0

95% CI -2.62637±1.96\*0.16003

3. β1+ β3 ???? => -2.62637 + 5.02318 = 2.39681 - Chris Shannon 8/16/2015

4. Adjusted R squared measures the power of the linear fitting which means how good the linear fit is. R-squared = Explained variation / Total variation  
Simon Sheather in his book on Regression (page 137, in chapter 5), states that because adding irrelevant variables increases R2, one should use R2adj to compensate because RSS/(n-p-1) is an unbiased estimator of var(ei) while SST/(n-1) is an unbiased estimator of var(Yi) when b1 = b2 = … = bp = 0. - Chris Shannon 8/16/2015

5. H0: β0=β1=β2=β3=0; Ha: at least one in β0,β1,β2,β3 is not equal to zero.

I think the F test on the bottom output should not include the intercepts, just the predictors.

H0: β1=β2=β3=0; Ha: at least one in β1,β2,β3 is not equal to zero.

To get the degrees of freedom for the F test

df2 = n-p-1 = 250-3-1 = 246

df1 = 3 (# parameters that distinguish between Ho and Ha) Monica 8/15/2015

6. Equal variance assumption is not met???? The variance looks pretty good; the x vs. resids plot looks random and the z vs. resids boxes are about the same size. The first scatter supports linearity and the histogram suggests normality. I wasn’t sure about independence of errors--can we evaluate that using these plots?

**I don’t think we can evaluate the independence of the errors with these plots. I think from the plots we don’t see any assumptions being violated. This is ok as the question states “If any”.- Oscar P. 8/10/15**

**I agree that none of the required assumptions (normality and equal variance) have not been violated. Might want to say there is not enough information to check independence. - Chris S. 8/16/2015**

7. (0.5035)2 ??? Leave it unsquared. The question asks for an estimate for sigma, not sigma squared. (Kate Mullan 8/10/15)

**I believe part 1 for problem II requires a qualitative answer for each of the regression coefficients such as Bo is the mean response when x=0 and z=0. Part 2 looks fine to me. Part 3 should be correct as well. For part 4, should you also mention that the adjusted R-squared accounts for the number of predictor variables in the model as well? Part 5 looks fine. -Oscar P. 8/10/2015**

Problem III

Part A

1. Xbar~N(20,5/4), Ybar~N(24,16/25)

2. W~N(-4,5/4+16/25); P(Xbar < Ybar )=ɸ(Ybar)- ɸ(Xbar)

3.

4. K=, m = 20

I just noticed that I had the wrong SDs for Xbar and Ybar. Nm I worked it out. Var(X-bar)=Var(X)/n=25/20=5/4, etc.

Isn’t the Std deviation of X-bar = sqrt(25/20) = 5/sqrt(20)? G Long 8/15/15

Hey G Long, yes that is the Std. Dev of X-bar, the notation however is usually N(mean, variance), that is why it is written as 5/4. -Oscar P. 8/15/15

Can you provide a reference on how you solved 3 and 4, or show some of your work? I had trouble with these.

GabrielleG 081015

For #3, I looked at the wikipedia page for chi square distribution.

**Hey Gabrielle, here is a small explanation for problem 4.**

**An r.v. that has a t-distribution has the following form where X~N(0,1) and Y~ with n degrees of freedom. m=20 because that is the number of terms in the summation for U which corresponds to the chi-square portion of the t-distribution.**

**K can be found by first standardizing the numerator to be distributed N(0,1). Since 20 was subtracted from the mean, then it is already centered. Now you need to divide by the standard deviation of X-bar-20 which is sqrt(5/4) to finish standardizing it. U however is not distributed chi-squared as is. U is sum of all terms (Xi-X-bar)2, and each Xi-X-bar must be distributed N(0,1). Since X~N(20, 25) while X-bar~N(20, 5/4) this tells us that X-Xbar ~ N(0, 25+5/4) so we must first divide X-Xbar by the appropropriate standard deviation in order to make it N(0,1). From this we have**

.

**This gives a value for K of . l**

Isn’t the SD of X-bar sqrt(5/4) for standardizing, the numerator of part 4 (X-bar - 20)? - DBartkowiak 08/10/15

David is right about that. I have corrected the whole thing. -Oscar P 8/10/15

**For reference you can use Probability and Statistics by Evans and Rosenthal. The latter part of Ch. 4 has the forms for distributions like T, Chi-square, F and others.**

**-Oscar P 8/10/15**

I got different answer for problem 2. P(xbar<ybar) = P[(xbar-ybar) <0]= P[W<0]=P[(W+4/sqrt(1.89)) < 0]= ɸ(2.9).

problem 3: var(xi-xbar) var(xi) + var(xbar), since xi and xbar are correlated. I used the trick to write xbar= xi/20 + (the part that contains xi and the part that does not contain xi).

xi-xbar= xi - (xi/20)- ( ) = -

thus var(xi-xbar)= \* var(xi) + \*19\*var(xi)= \* 25 + \*19\*25=23.75

Then is distributed at N(0,1)

thus U/23.75 is distributed at chisq(20).

thus P[U>50] = P[> ] = P[> ] = 1- G(2.105), G is the cdf of chisq(20). Jingjing 08/14/15

Where did you get the idea to do this for problem 3? I just realized that what you are doing is just Var(Xi)- Var(X-bar) since (25 - 5/4) = 23.75 Oscar P -8/15/15

Part B

1. U~N(0, 61) and V~N(-8, 23)
2. C1=⅙, C2=2, C3=¼, C4= -3, C5=2
3. C1=⅔, C2=2, C3=2, C4= -3, C5=2, C6=C7=1

Question IV

1. The expected value of Y is a linear combination of the betas; E(Y) is linearly related to the betas.
2. Least squares criterion: Minimize the sum of squares of vertical distances between observed and predicted y-values. I.e., minimize RSS = [not sure what formulas should be included]
3. Gauss-Markov: assumption of normality not required for LSE’s of betas to be unbiased
4. Assumption of normality not required (??? not sure about this)
5. All assumptions must be true to conduct hypothesis tests. Or is normality not needed for large enough samples? (I’m looking at the flow chart on Sheather page 103, which suggests that normality is only needed for small n.)

\*\*\*If anyone has a strong grasp/good resource on which assumptions are and are not necessary for certain tasks (calculating unbiased estimators, constructing confidence/prediction/tolerance intervals, conducting hypothesis tests), please let me know! [katemullan1@gmail.com](mailto:katemullan1@gmail.com) Thanks :)

For 2. I was going to use minimize RSS=sum(ei^2) = e’e = (y-xb)’(y-xb) = y’y-y’xb-b’x’y-b’x’xb =y’y-2y’xb-b’x’xb

dRSS/db = -2x’y+2x’xb → 2x’y=2x’xb → (x’x)^-1(x’y)=b-hat though I don’t know how necessary this is to show/derive. D.Bartkowiak 8/10/15

You’re probably right to include all of that.

Here is what I put for 3, 4 and 5: G Long 8/16/15

3) I had that the errors must have an expected value of zero for beta-hat to be unbiased. Therefore independence and homoskedasticity are not required, nor do they have to be normally distributed.

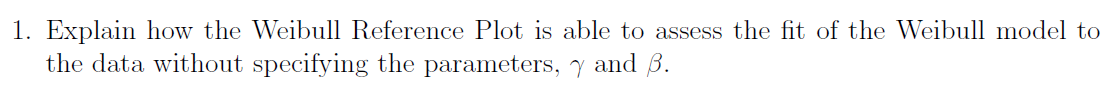
4) The constant variance assumption is also required for S to be an unbiased estimate of sigma-squared, normality is not required

5) All the assumptions are required. Although, I think you are correct about this not being required for large n

January 2012

Started Chris Shannon - 8/10/2015

# Question I

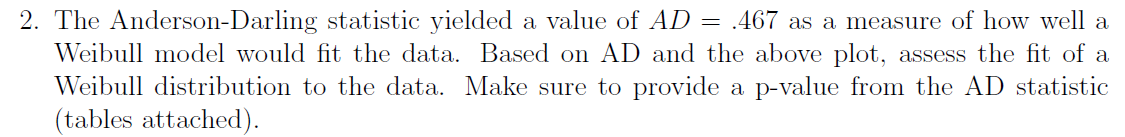


If the log-transformed observed values belong to a Weibull distribution, then a straight line will approximate the points in the plane given in the plot above according to the equation:

log(Vi) = 3.87 + 0.1099Q(ui)

Because the log-transformed Weibull distribution is a location-scale distribution, the intercept and slope of the line will approximate the transformed parameters of the log-Weibull distribution if the observed data is from a Weibull distribution.

From here onwards we solve for ui to get the CDF to show that the regression intercept and slope are parameters from a Weibull distribution.



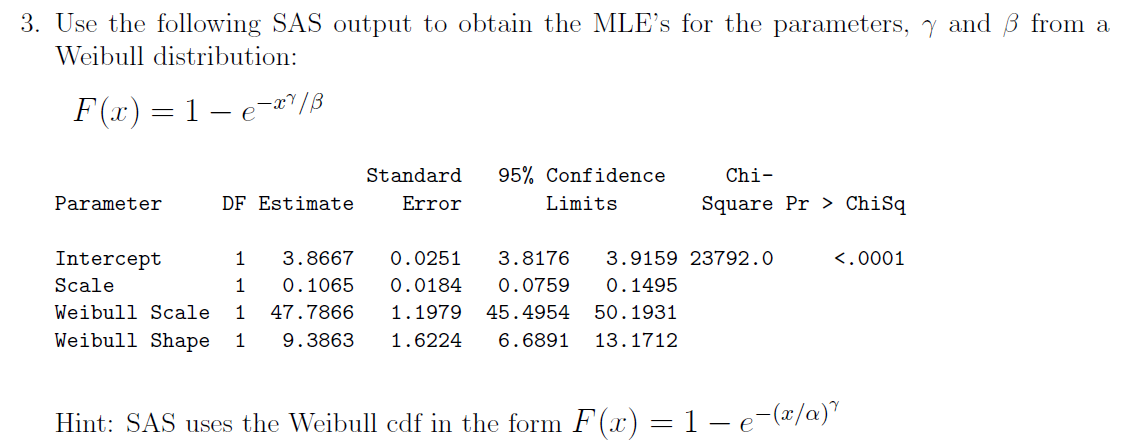
The above plot shows that the data is somewhat close to the straight line, and the Anderson-Darling statistics does not give any evidence of a deviation from the Weibull distribution. The test statistic is:

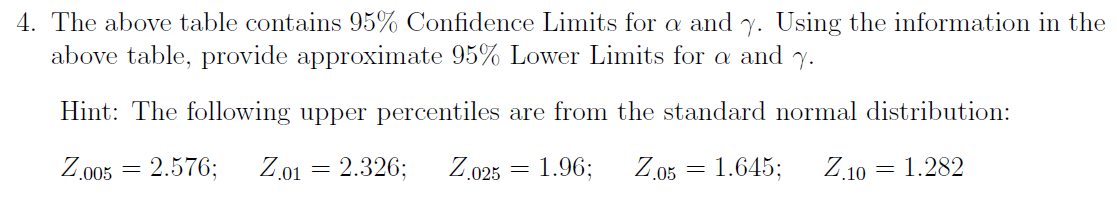
AD\* = 0.467 < ADcrit= 2.492  
p-value = P(0.467 < 2.492) = 0.2019 + [(0.467-0.45)/0.05] x (0.2532 - 0.2019) = **0.2193**

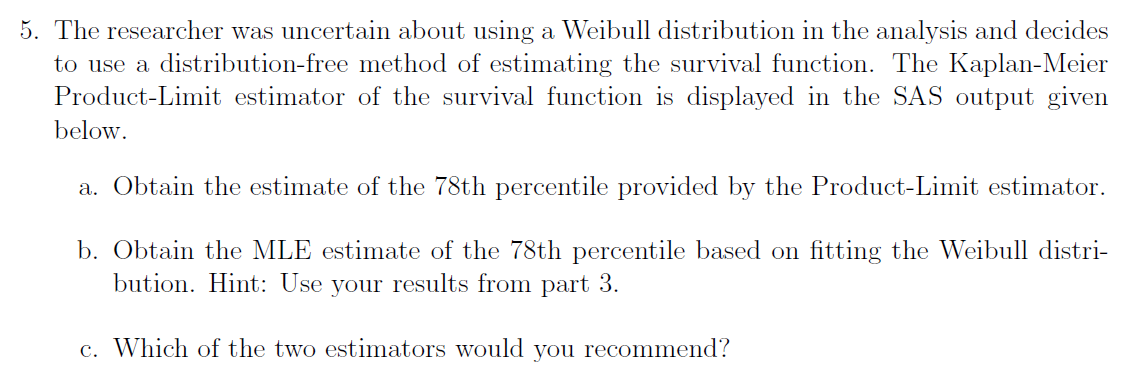
I get p-value = 0.2193, no change to the conclusion though. -D.Bartkowiak 8/12/2015

**Yes, error in punching in the numbers on the calculator. Thank you for catching it. I changed the answer from 0.2036 to 0.2193 - Chris Shannon 8/12/2015**

Since the p-value of 0.2193 > 0.05, we conclude H0, that the observed values are from a Weibull distribution.







1. ; solve for V.
2. I would recommend the Kaplan-Meier Product-Limit estimator, because it is based on the data, because the Weibull reference plot shows that the data only loosely fit a Weibull distribution, and because the difference between the two estimators given the range of the data is large.

# Question II

1)

|  |  |
| --- | --- |
| C1 Normality | Yes, the condition is satisfied.  Shapiro-Wilk, Kolmogorov-Smirnov, Cramer von Mises, and Anderson-Darling test statistics all have p-values above 0.15 > 0.05 significance level. Additionally, the QQ Normal plot shows that the residuals closely fit a straight line. |
| C2 Equal Variance | Yes, the condition is satisfied.  The Brown-Forsythe test for constant variance returns a test statistic of 0.50 with a p-value of 0.85, so we do not reject the null hypothesis of the variance being constant. |
| C3 Independence | Yes, guaranteed by the randomization design. |

2)

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **dfnum** | **dfdenom** | **p-value** |
| Q | 2 | 12 | < 0.0001 |
| M\*Q | 4 | 12 | < 0.0001 |
| R(M) | 6 | 12 | 0.0016 |

3) (Copy-and-pasted from SAS printout included with exam.)



4) Since there is a significant interaction between M and Q, we should separate the levels of Grade (Q) at each level of Method (M):

M1: {Q1,Q2}, {Q3}

M2: {Q1}, {Q2}, {Q3}

M3: {Q1}, {Q2,Q3}

5)

CI95%=

6)

1. SS Quality
2. SS Quality X Manufacturer
3. SS Run(Manufacturer)

# Question III

1)

1. Chi-squared with 5 degrees freedom, since a Chi-squared by definition is
2. Student’s t with 6 degrees of freedom, since
3. F-distribution with 1 and 6 degrees of freedom, since F is a Chi-squared over a Chi-Squared.
4. t-distribution with one df or Cauchy distribution.
5. F-distribution with 2 and 3 degrees freedom.

2)

1. E(U) = E(2X + 3Y - 5) = 2E(X) + 3E(Y) - E(5) = 2(2) + 3(-3) - 5 = 4 - 9 - 5 = -10  
   Var(U) = Var(2x + 3Y - 5) = 4Var(X) + 9Var(Y) = 4(8) + 9(5) = 77  
   **U ~ N(-10,77)**  
   E(V) = E(X - CY) = E(X) - CE(Y) = 2 - 3C  
   Var(V) = Var(X - CY) = Var(X) + C2Var(Y) = 8 + 5C2  
   **V ~ N(2 + 3C, 8 + 5C2) Since E(Y) = -3, shouldn’t E(V) = 2 + 3C? --Sean W., 8/13/15 Yes, it should. Thanx for catching it! Changes made throughout problem. Chris Shannon 8/13/15**
2. U and V are by definition jointly normal:  
     
   Therefore if we show that the correlation of U and V = 0, then U and V are independent.  
     
     
     
    (If E(UV)-E(U)E(V) = 0, then we need not worry about )  
     
     
     
     
    since X and Y are independent   
     
     
     
    start reducing the E(X2) and E(Y2) terms  
     
     
    now we have E(UV) in terms of all known values  
     
   ; and  
      
   So,   
   The value of C that makes U and V independent is
3. Since , C1..4 must be chosen to make X and Y standard normal random variables.  
     
     
     
   C1 = ⅛, C2  = -2, C3 = ⅕, C4 = 3, C5 = 2
4. , C1 =, C2 = -2, C3 = 1, C4 = 3, C5 = 1, C6 = 1
5. , C1 = ⅝, C2 = -2, C3 = 2, C4 = 3, C5 = 2, C6 = 1, C7 = 1

# Question IV

1)

1. It is not correct to test Y with a Shapiro-Wilk test because Y is a linear combination of fixed effects and the the residual error term, which is distributed as N(0,1). It is correct to test the residual error with Shapiro-Wilk.
2. Inxn is the identity matrix of dimensions nxn.
3. There p+1 unknown parameters, which include β0 for the intercept and β1 … βp for slopes of the predictors describing the relationship of X with Y.
4. Would it make sense to replace Y with to arrive at  
    the result in part f? -D.Bartkowiak 8/13/15  
    Yeah, I think so. - Chris Shannon 8/13/15
5. Yes, is an unbiased estimator of since

2) I’d disagree with this conclusion as R-Squared is not an indicator of model validity and is not invariant under transformation. If Model 2 is invalid (if it violates one of the assumptions for MLR) then it does not matter how much larger the value of R-Squared is compared to Model 1.

D.Bartkowiak 8/13/15

(Moving on to other things - will appreciate others’ insights into the answer on this one.) - Chris Shannon 8/12/15)

**Jan. 2013 Exam Discussion**

**Problem I**

1. is the mean response for subjects when (i.e., for wild-type males with average consumption).  
    is the difference in means between wild-type males of average consumption and wild-type females of average consumption.  
    is the difference in means between wild-type males of average consumption and knockout males of average consumption.  
    is the increase in the mean of *Y* for each one-unit increase in food consumption from the average for wild-type males.  
    is the difference in the increase in the mean of *Y* for each one-unit increase in food consumption from the average for wild-type females and the same increase for wild-type males.
2. Recall that . (Let’s call the resulting matrix *V*).  
     
   Thus,   
     
     
   Hence, .  
     
   Then, the CIs are given by   
     
     
   **I got different answer using =, with v’=[0 1 0 0 -1]. However, my =**

**Then, the CIs are given by 0.8142 (-1.94, 3.57) Jingjing 081515**

**Ironically this was the answer I got originally, using -2Cov instead of +2Cov! --Sean W. 8/16/15**

**Jingjing and Sean, I also got (-1.94,3.57). Since the interval includes 0, we have insufficient evidence of a genotype\*gender effect, right? Gabrielle G 081615**

1. There is some evidence that the knockout treatment increases time spent exercising (by approximately 1 minutes, *p* < .05). Note, however, that variable selection methods produce questionable *p*-values. As such, this finding should be considered provisional until and unless the reduced model is successfully fitted to data from a new sample, ideally gathered under identical experimental conditions.

**Problem II**

1. For a distribution *X* ~ Poisson(), E(*X*) = Var(*X*) = . Thus, is effectively equivalent to the usual (distribution-free) confidence intervals. Additionally, because the sample size is relatively large, the distribution of sample means is approximately normal, so may be used.  
     
   (This actually does work reasonably well. R code follows:

**i = NULL**

**m = NULL**

**n = 48**

**for(i in 1:10000){**

**m[i] = mean(rpois(n,2.04))}**

**xbar = mean(m)**

**se = sqrt(xbar/n)**

**z = qnorm(.975)**

**moe = se\*z**

**lcl = xbar - moe**

**ucl = xbar + moe**

**(length(m[m<lcl])+length(m[m>ucl]))/10000**

1. No. Severe overdispersion is present. If this were Poisson distributed, the mean should be close to the variance, but the variance is over 4 times the size of the mean, so Poisson is a poor fit.
2. Good fit to the specified distribution. However, in this case, , so we reject the null and conclude the data are a poor fit to the Poisson distribution.
3. Residual variance looks nonconstant, which is fatal for prediction intervals regardless of sample size.

**Problem III**

1. a)   
   b)   
   c)
3. a)   
   b)   
   c)