Summer 2015 homework 4 solution

1. (14 points) Accelerated Life Test Experiment:

   a. Let F_{ij} be the time to failure of the jth TV under temperature i. The normality of the distribution of the F_{ij}'s is evaluated using the residuals from the fitted model, $\hat{e}_{ij} = F_{ij} - \bar{F}_{i\cdot}$. From the normal probability plot and the stem-leaf plot, the distribution of $\hat{e}_{ij}$ is nearly symmetric. The p-value from the Shapiro-Wilk p-value = .0896. Therefore, we conclude that the distribution of the residuals appears to be approximately normally distributed. The four temperatures have the following values for the sample means and standard deviations:

```
Level of
 T N    F Mean    F Std Dev
40 6 3955.66667 2021.46300
45 6 1925.50000  689.20149
55 6 1043.00000  344.64881
70 6  647.00000   73.42752
```

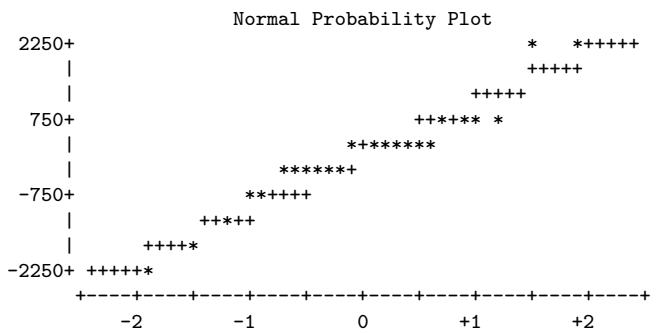
   The p-value from the BFL test is p-value $< .0001$ which would imply that there is significant evidence of a difference in the four populations variances which is confirmed by examining the very large differences in the four sample standard deviations, S_i, 73.4 to 2021.5.

```
Brown and Forsythe's Test for Homogeneity of  Variances

Source DF  Sum of Squares Mean Square   F Value  p-value
T       3     10612311     3537437       23.43   <.0001
Error  20      3019636      150982

  Variable: RESID
Tests for Normality
Test                Statistic        p Value
Shapiro-Wilk        W 0.92835     Pr < W     0.0896

   Stem Leaf                     #  Boxplot
      2 24                       2     0
      1                                |
      1                                |
      0 5689                     4     |
      0 0000123                  7  +--+--+
     -0 442210                   6  +-----+
     -0 76                       2     |
     -1                                |
     -1 85                       2     0
     -2 0                        1     0
        ----+----+----+----+
```



b. Using the following R code, the estimated slope from the regression of log(sd) on log(mean) is obtained.

```
mean = c(3955.7,1925.5,1043,647)
sd   = c(2021.46,689.20,344.65,73.43)
lm(log(sd)~log(mean))

Coefficients:
(Intercept)    log(mean)
     -6.583        1.729
```

Therefore, $\hat{\beta} = 1.73$ which yields the transformation: $y = F^{1-1.73} = F^{-.73}$. In most cases, the transformation $y = F^{-1} = 1/F$ works nearly as well and the reciprocal transformation is easier for non-mathematical consultees to understand.

c. Using the following R code, the boxcox transformation is obtained.

```
d41 = matrix(0,24,2)
f = matrix(0,24,1)

f = c(1953, 2135, 2471, 4727, 6134, 6314, 1190, 1286, 1550, 2125, 2557, 2845,
      651, 817, 848, 1038, 1361, 1543, 511, 651, 651, 652, 688, 729)

tmp = c(rep("40",6),rep("45",6),rep("55",6),rep("70",6))
tmp = as.factor(tmp)
d41 = data.frame(f,t)
summary(lm(f~t))
anova(lm(f~tmp))

library(MASS)
boxcox(f~tmp,lambda=seq(-3, 3,.01))
boxcox(f~tmp,lambda=seq(-1.5, 0,.01))
boxcox(f~tmp,lambda=seq(-.8,-.5,.01))
```

From the R output the value of θ is approximately -.63 with a 95% confidence interval which includes -1

d. Using the boxcox transformation $y = F^{-.635}$ and the reciprocal transformation, $y = 1/F$, we obtain the following results:

Transform	SW p-value	Residual Plots	BFL p-value
$y = F^{-.635}$	.5719	Symmetric with plotted points near a straight line	.1106
$y = 1/F$	.8021	Symmetric with plotted points near a straight line	.1929

Both transformations yield data which appears to have normally distributed residuals with equal variances based on the Shapiro-Wilk test, Brown-Forsythe-Levene test, and the residual plots.

e. Using the original data and the transformation $y = 1/F$, we obtain the following results:

```
Dependent Variable: F - Time to Failure

Source          DF  Sum of Squares   Mean Square   F Value   Pr > F
Model            3   39183995.13      13061331.71   11.15    0.0002
Error           20   23427428.83       1171371.44
Corrected Total 23   62611423.96

Source DF   Type III SS   Mean Square   F Value   Pr > F
T      3    39183995.13   13061331.71   11.15     0.0002


Dependent Variable: Y = 1/F - Reciprocal of Time to Failure

Source          DF  Sum of Squares   Mean Square    F Value  Pr > F
Model            3   5.4071936E-6     1.8023979E-6   32.85   <.0001
Error           20   1.0972338E-6     5.4861689E-8
Corrected Total 23   6.5044274E-6

Source DF  Type III       SS Mean Square   F Value   Pr > F
T      3    5.4071936E-6   1.8023979E-6     32.85    <.0001
```

Based on the two AOV's, we would reach the same conclusion, there is significant evidence that the four temperatures have a difference in their mean time to failure. However, the conclusion is somewhat stronger from the reciprocal transformation than from the original data, a p-value$<$ .0001 vs a p-value of .0002. The

validity of the p-value from the original data is questionable because there was significant evidence that the original data did not have equal variances.

f. Both transformations, $y = 1/F$ and $y = F^{-.635}$, generated Tukey HSD groupings of the four treatment means as follows:

$G_1 = \{40°C,\ 45°C\}$; $G_2 = \{55°C\}$; $G_3 = \{70°C\}$

- The original data had considerably different groupings:

$G_1 = \{40°C\}$; $G_2 = \{45°C,\ 55°C,\ 70°C\}$ but of course, these groupings are of questionable validity because the original data did not have equal variances.

g. Using $\alpha_{PC} = 1 - (.05)^{1/3} = .01695$ and the data from the reciprocal transformation, $y = 1/F$, there is significant evidence (p-value< .0001) of a positive linear trend in the mean wrinkle resistances across the four temperatures. The quadratic and cubic trends were not significant with p-values of .3737 and .9589, respectively.

- Using the data from the Box-Cox transformation, $y = F^{-.635}$, similar p-values were obtained for the three trends: Linear p-value < .0001, Quadratic p-value = .1335, and Cubic p-value = .8437.
- Evaluating the trends using the **untransformed data**, the p-values for the three trends were Linear p-value< .0001, Quadratic p-value = .0143, Cubic p-value =.1821, thus there was significant evidence of both a linear and quadratic trend. These conclusions are different from the conclusions that were observed in the transformed data.

2. (6 points)

a. Using the following R code or using the SAS program from Handout 5,

```
d41 = matrix(0,24,2)
f = matrix(0,24,1)
f = c( 4953, 5135, 5471, 7727, 9134, 9314,  4190, 4286, 4550, 5125, 5557, 5845,
3651, 3817, 3848, 4038, 4361, 4543, 3511, 3651, 3651, 3652, 3688, 3729)
f = f-3000
tmp = c(rep("40",6),rep("45",6),rep("55",6),rep("70",6))
tmp = as.factor(tmp)
d41 = data.frame(f,t)
kruskal.test(f,tmp,f~tmp)

Kruskal-Wallis chi-squared = 18.2785, df = 3, p-value = 0.0003853
```

We can thus conclude there is very significant evidence (p-value = 0.0004) of a location difference in the four wrinkle resistance populations.

b. The rank based multiple comparison procedure yields the following results: The p-value from the Kruskal-Wallis test was p-value = .0004.

The mean ranks are $\bar{R}_{1.} = 20.33$ for $40°C$, $\bar{R}_{2.} = 16.00$ for $45°C$, $\bar{R}_{3.} = 9.50$ for $55°C$ and $\bar{R}_{4.} = 4.177$ for $70°C$.

Using the Miller procedure, two pairs of Strains are said to be different if

$|\bar{R}_{i.} - \bar{R}_{h.}| > q(.05, t, \infty)\sqrt{\frac{t(n+1)}{12}} = 3.633\sqrt{\frac{4(24+1)}{12}} = 10.49$

with $q(.05, 3, \infty) = qtukey(.95, 4, 100000) = 3.633$ or use the value from the Tukey table in the textbook.

The Miller rank procedure finds

$\bar{R}_{1.} - \bar{R}_{2.} = 20.33 - 16.00 = 4.33 < 10.49 \Rightarrow$ Not significant
$\bar{R}_{1.} - \bar{R}_{3.} = 20.33 - 9.50 = 10.83 > 10.49 \Rightarrow$ Significant
$\bar{R}_{1.} - \bar{R}_{4.} = 20.33 - 4.17 = 16.16 > 10.49 \Rightarrow$ Significant
$\bar{R}_{2.} - \bar{R}_{3.} = 16.00 - 9.50 = 6.50 < 10.49 \Rightarrow$ Not significant
$\bar{R}_{2.} - \bar{R}_{4.} = 16.00 - 4.17 = 11.83 > 10.49 \Rightarrow$ Significant
$\bar{R}_{3.} - \bar{R}_{4.} = 9.50 - 4.17 = 5.33 < 10.49 \Rightarrow$ Not significant

The yields the following groupings of the temperatures on the basis of similar mean wrinkle resistance are

$G_1 = \{40°C,\ 45°C\}$; $G_2 = \{45°C,\ 55°C\}$; $G_3 = \{55°C,\ 70°C\}$.

c. The groupings using the ranks are different from the groupings obtained from the untransformed data but are nearly the same as the groupings from the transformed data with the exception that the rank procedure did not separate the mean from $55°C$, from the means of $45°C$ and $70°C$.

3. (12 points) Female Moth Egg Count Problem:

   a. The three Strains of moth have the following values for the sample means and standard deviations:

STRAIN	N	Mean	Std Dev
USDA	15	368.000000	265.621374
FIELD	15	181.266667	210.991018
RESIST	15	90.800000	118.106610

   The Poisson distribution has the relationship of $\sigma_i^2 = \mu_i$. From the sample means and standard deviations, the sample variances would be much larger than the sample means for all three strains. Therefore, an overdispersed Poisson may be possible but the regular Poisson distribution would not be a reasonable fit to the data.

   b. From the SAS output for the fit of a Poisson model to the data, the "Scaled Deviance/DF = 200.3217" which is not very close to 1.0. Therefore, the results of the Poisson analysis would not be valid.

   • Using the Overdispersed Poisson model, the SAS output has the "Scaled Deviance/DF = 1.0170" which is very close to 1.0. Therefore, the results of the Overdispersed Poisson analysis would appear to be valid.

   • From the SAS output

Contrast	NumDF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq	Type
S2 vs S1	1	42	4.93	0.0318	4.93	0.0264	LR
S2 vs S3	1	42	2.33	0.1340	2.33	0.1265	LR
S3 vs S1	1	42	13.67	0.0006	13.67	0.0002	LR

   Based on the SAS output using $\alpha_{PC} = .05/3 = .0167$, there is significant evidence of a difference in the mean egg count of the USDA and RESIST strains but not between USDA and FIELD nor between FIELD and RESIST.

   c. Because there was an indication that the original data was not normally distributed, the Runs Test will be used to evaluate correlation in the data. The following R code will yield the results for the runs test:

```
 N1=c(448,906, 28,277,634 ,48,369,137,29,522,319,242,261,566,734 )
N2=c(211,276,415,787, 18,118,  1,151, 0,253, 61,  0,275,  0,153 )
N3=c(0,9,  143,  1, 26,127,161,294, 0,348,  0, 14, 21 , 0,218 )
N = c(N1,N2,N3)
data3 = matrix(N,nrow=3,byrow=T)
S = matrix(0,3,15)
meansN = matrix(0,3,15)

for (i in 1:3) {

  meansN[i] = mean(data3[i,])
   S[i,] = data3[i,]-meansN[i]
  }

n.neg  = rep(0,3)
n.pos  = rep(0,3)
for (i in 1:3) {
  n.neg[i]  = length(S[i,][S[i,]<0])
  n.pos[i]  = length(S[i,][S[i,]>0])
  }

numb.runs  = rep(1,3)
for (i in 1:3) {
  for (j in 2:15) {
    if (sign(S[i,j]) != sign(S[i,j-1])) {numb.runs[i]  = numb.runs[i] + 1}
    }
  }

Sruns.result  = as.data.frame(cbind(numb.runs, n.pos, n.neg))
names(Sruns.result)  = c("No. runs", "N+", "N-")
```

The critical values n_{Lower}, n_{Upper} are from Table A.30(a) on page 54 in HO 5:

With N equal to the number of runs, N+ the number of positive differences $y_i - \bar{y}_.$, and N- the number of negative differences, we obtain the following results

N	N+	N-	n_{Lower}	n_{Upper}
9	7	8	4	13
6	6	9	4	13
8	6	9	4	13

For all three strains, there is not significant evidence of correlation in the data because N fell between n_{Lower} and n_{Upper} in all three cases.

4. (18 points) Problem 5.4 from the Kuehl's textbook:

   a. i. $Y_{ij} = \mu + A_i + e_{ij},\ i = 1, \cdots, 8,\ j = 1, 2, 3, 4$, where μ is overall mean, A_i is the random effect of the selected lots and e_{ij} is the random effect of the selected samples within lots.

   • Distributional conditions: $A_i \sim iid\ N(0, \sigma_A^2),\ e_{ij} \sim iid\ N(0, \sigma_e^2)$, A_i and e_{ij} are independent.

   ii. $t = 8,\ r = 4$:

Source	df	SS	MS	EMS	F	$Pr > F$
TRT	7	13696.46875	1956.63839	$\sigma_e^2 + 4\sigma_A^2$	8.46	$< .0001$
Error	24	5548.25000	231.17708	σ_e^2		
Total	31	19244.71875				

   For testing $H_0 : \sigma_A = 0$ vs $H_1 : \sigma_A > 0$, the AOV table has $p - value < 0.0001$. Thus reject H_0 and conclude that there is significant evidence ($p - value < 0.0001$) of variation in the Aflatoxin concentration between lots.

   b. Variance for lots= σ_A^2, Variance for samples within lots= σ_e^2. Using the AOV–matching moment method, we obtain

$$\hat{\sigma}_e^2 = MSE = 231.17708, \qquad \hat{\sigma}_A^2 = \frac{MS_{TRT} - MSE}{r} = 431.3767.$$

   • Because this was a one factor random treatment experiment with equal replications per treatment, the REML estimators from PROC MIXED are identical to the AOV-MOM's.

   c. $\hat{\sigma}_y^2 = \hat{\sigma}_A^2 + \hat{\sigma}_e^2 = 662.5538$.

   d. i. Proportion due to variation between lots: $\frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_y^2} = 0.6511$ (65.11%).

   ii. Proportion due to variation within lots: $\frac{\hat{\sigma}_e^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_y^2} = 0.3489$ (34.89%).

   e. $\hat{\sigma}_y = \sqrt{\hat{\sigma}_y^2} = 25.7401$.

   f. The variance components can be used to determine (1) the number of samples needed per lot to achieve a specified power of a test of hypotheses and (2) optimum allocations of samples to lots based on future costs.

5. (10 points)

   a. Because the levels of the treatment factor Lot Number were randomly selected, it would not be appropriate to test for differences in the means of the eight Lot Numbers.

   b. This problem wants you to determine the number of reps, r, for a fixed number of random treatment levels, t.

   With $\alpha = .01$, $\gamma_o = .90$, $t = 8$, and $\tau_o = \frac{\sigma_A^2}{\sigma_e^2} = 2$, we obtain the value of the number of reps, r, as follows:

   i. $\nu_1 = t - 1 = 7,\ \nu_2 = t(r-1) = 8(r-1),\ \ \lambda = \sqrt{1 + r\tau_o} = \sqrt{1 + 2r}$

   ii. Power at λ is given by $\gamma(\lambda) = 1 - pf(qf(1 - .01, 7, 8(r-1))/(1 + r\tau_o), 7, 8(r-1))$ using the R-function for the cdf of an F-distribution.

   iii. Iteratively determine the smallest value of r such that $\gamma(\lambda) \geq .90$

iv. Using either Table X in the textbook, the SAS program, **repsize,randomeffects,fixedt.sas**, or the following R code we obtain:

```
r=seq(2,8,1)
power = 1-pf(qf(1-.01,7,8*(r-1))/(1+2*r),7,8*(r-1))
out = cbind(r,power)
out
     r     power
[1,] 2 0.3834528
[2,] 3 0.7659997
[3,] 4 0.8999179
[4,] 5 0.9503449
[5,] 6 0.9724452
[6,] 7 0.9833860
[7,] 8 0.9893302
```

r	$\nu_1 = t-1$	$\nu_2 = t(r-1)$	$F_{.01,7,8(r-1)}$	$\lambda = \sqrt{1+2r}$	$\gamma(\lambda)$
2	7	8	2.8661	2.236	$1 - G_{7,8}(2.763) = .3835$
3	7	16	2.7587	2.646	$1 - G_{7,16}(1.522) = .7660$
4	7	24	2.6896	3.000	$1 - G_{7,24}(1.165) = .8999$
5	7	32	2.6415	3.317	$1 - G_{7,32}(.982) = .9503$

From the above results, we have that $r \geq 5$.

6. (8 points) This problem wants you to determine the number of random treatment levels, t, for a fixed number of reps.

With $\alpha = .01$, $\gamma_o = .90$, $r = 5$, and $\tau_o = \frac{\sigma_A^2}{\sigma_e^2} = (2.1)^2/(2)^2 = 1.1025$, we obtain the value of the number of circuits, t, as follows:

i. $\nu_1 = t-1,\ \nu_2 = t(r-1) = t(5-1),\quad \lambda = \sqrt{1+r\tau_o} = \sqrt{1+5(1.1025)} = 2.55196$

ii. Power at τ_o is given by $\gamma(\tau_o) = 1 - pf(qf(1-.01, t-1, t(5-1))/(1+r\tau_o), t-1, t(5-1))$ using the R-function for the cdf of an F-distribution.

iii. Iteratively determine the smallest value of r such that $\gamma(\tau_o = 1.1025) \geq .90$

iv. Using either Table X in the textbook, the SAS program, **trtsize,randomeffects,fixedr.sas**, or the following R code we obtain:

```
t=seq(2,12)
power = 1-pf(qf(1-.01,t-1,t*(5-1))/(1+5*(1.1025)),t-1,t*(5-1))
out_t = cbind(t,power)

      t     power
[1,]  7 0.7719918
[2,]  8 0.8271899
[3,]  9 0.8700052
[4,] 10 0.9028619
[5,] 11 0.9278456
```

t	$\nu_1 = t-1$	$\nu_2 = t(5-1)$	$F_{.01,t-1,t(5-1)}$	$\lambda = \sqrt{1+2(5)}$	$\gamma(2.552)$
7	6	28	3.528	2.552	$1 - G_{6,28}(.5417) = .7720$
8	7	32	3.258	2.552	$1 - G_{7,32}(.5003) = .8272$
9	8	36	3.052	2.552	$1 - G_{8,36}(.4686) = .8700$
10	9	40	2.888	2.552	$1 - G_{9,40}(.4434) = .9029$
11	10	44	2.754	2.552	$1 - G_{10,44}(.4228) = .9278$

From the above results, we have that $t \geq 10$.

7. (32 points, 4 each) Short Answer Questions

a. The power will be increased but also the probability of a Type I Error will be much larger than the nominal value.

(Hence the procedure would too often state that there is strong evidence that the research hypothesis is true when in fact it is false.)

b. Regression of S_i on $\bar{y}_i$ yields $R^2 = .92$ and p-value=.0023, which is an indication of nonequal variances. Use transformation $X_{ij} = y^{1-\hat{\beta}_1} = \bar{y_{ij}}^{.5} = 1/\sqrt{y_{ij}}$

c. The ordering of X_{ij} values is reversed from the ordering of y_{ij} values. Apply Hsu's procedure to X_{ij}'s values with "best is smallest".

d. The t treatment population distributions have same location-scale family with equal scale but different location parameters. Data iid within treatments and independent between treatments. Conditions are identical with standard AOV model except non-Normal distributions are allowed.

e. When the treatments have random levels, comparisons of the t selected means is not of interest. Question of interest is the size of differences in population of treatments, not in the t selected treatments.

f. The ratio MS_{SAL}/MS_{CON} is correct ratio only when sample sizes are equal. For unequal sample sizes, use the ratio of MS_{SAL} to a linear combination of MS_{CON} and MSE

g. Declare a data value an outlier, if

$$|\hat{e}_{ij}| \geq \hat{\sigma_e}\sqrt{1 - \frac{1}{n_i}}t_{.0005,df_E} = \sqrt{9}\sqrt{1 - \frac{1}{10}}(3.52) = 10.02$$

Data value is outlier if corresponding residual is less than -10.02 or greater than 10.02. Thus, no outliers in the 50 data values.

h. With equal sample sizes and data from populations having mildly different variances, then distribution of MS_{TRT}/MSE would be very nearly an F-distribution. When sample sizes are unequal and inversely related to the size of the treatment population variances, the distribution of MS_{TRT}/MSE would not be close to an F-distribution.