Priyasri Sankaran
Project 1

The primary objective of the Study on the efficacy of Nosocomial Infection Control(SENIC project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. (Reference: Special Issue, \The SENIC project",

1. (a) Test whether or not the mean infection risk (variable 4) is the same in the four geographic regions American Journal of Epidemiology 111 (1980), 465-653.)(variable 9); use α = :05.

Assume that ANOVA model is applicable. State the alternatives, conclusion.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ The mean infection is the same in the four geographic regions

H_a: Not all means are equal in the four geographic regions

$Yij = \mu i + \varepsilon ij$, εij iid ~ $N(0, \sigma^2)$, i = 1, 2, 3, 4, j = 1, 2, ...113.

From the ANOVA table F statistics=2.71, we calculated $F_{(0.05,3,109)}$ critical=2.6789.

$F_{stat} > F_{critical}$. The P value=0.0484<0.05. We reject the null hypothesis and conclude that not all means are equal in the four geographic regions are the same.

One-Way ANOVA for Infection Risk across Geographic Regions

The GLM Procedure

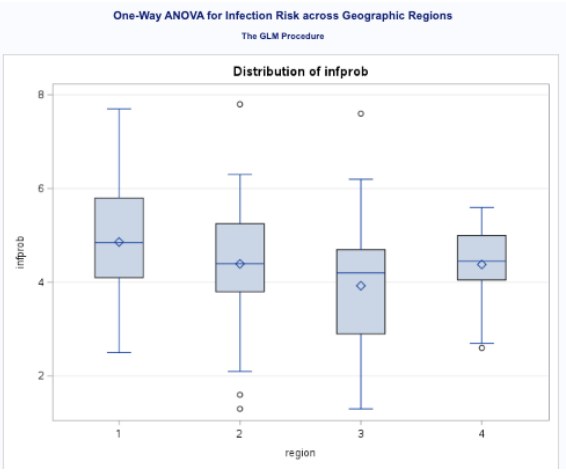
Dependent Variable: infprob

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.9969393	4.6656464	2.71	0.0484
Error	109	187.3828837	1.7191090		
Corrected Total	112	201.3798230			

R-Square	Coeff Var	Root MSE	infprob Mean
0.069505	30.10765	1.311148	4.354867

Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	13.99693932	4.66564644	2.71	0.0484

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	13.99693932	4.66564644	2.71	0.0484



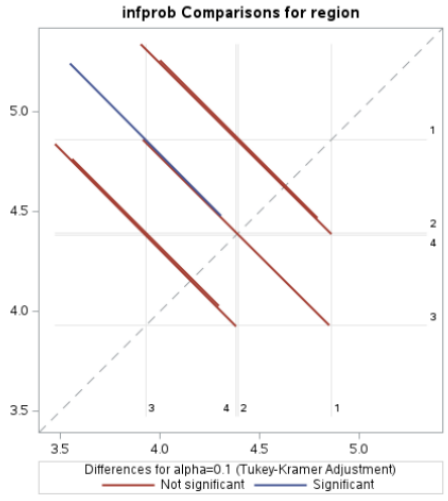
1)b) Obtain confidence intervals for all pairwise comparisons between four regions, use the Tukey procedure and a 90 percent family confidence coefficient. Interpret your result and state your findings. Prepare a line plot of the estimated factor level means and underline all nonsignificant comparisons.

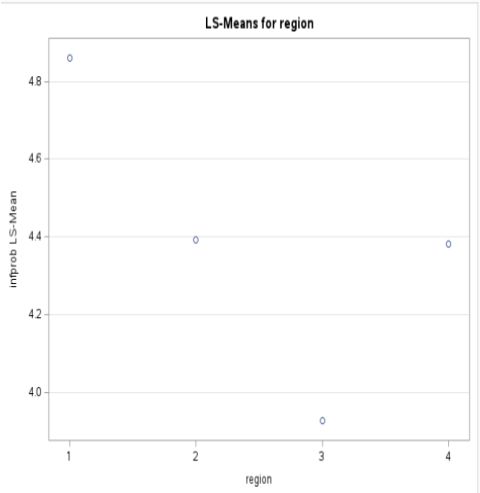
Interpretation and Findings: We performed Tukey's procedure. We obtained the below result. We are 90% confidence that the true value lies between the intervals. Please see the below result.

We also noticed that these regions 1-2,1-4,2-1,2-4 and 2-3, contains zero indicating it's not significant. Region 1-3 do not contain zero, indicating a significant difference between the regions.

Line plot: The line plot is also shows the significant and not significant regions are marked.

Least Squares Means for Effect region				
i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.466964	-0.319842	1.253770
1	3	0.933687	0.172095	1.695280
1	4	0.479464	-0.473405	1.432333
2	3	0.466723	-0.267275	1.200721
2	4	0.012500	-0.918461	0.943461
3	4	-0.454223	-1.363975	0.455529





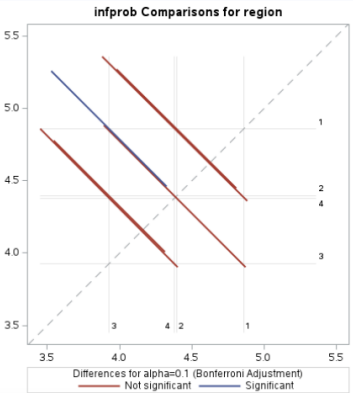
(c) For the same family confidence coefficient, try a different pairwise comparison procedure. Interpret your result and state your findings.

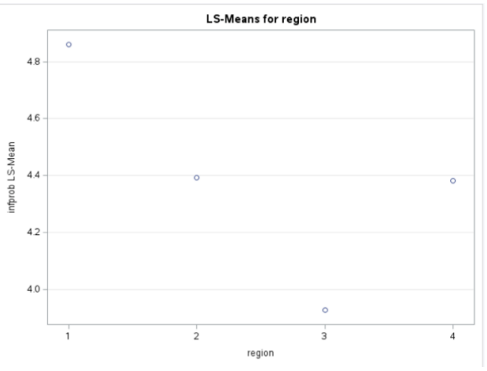
Interpretation and Findings: We tried the pairwise comparison procedure in a Bonferroni procedure at 10% significant level. We are 90% confidence that these interval contains the true value.

We obtained the below result. The region 1-2,1-4,1-3,2-4 and 3-4, on these intervals contain zero in them. Indicating that they are not significant. The region 1-3 , interval contains zero, indicating a significant difference.

Line plot: The line plot is also shows the significant and not significant regions are marked.

Least Squares Means for Effect region				
i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.466964	-0.358020	1.291948
1	3	0.933687	0.135140	1.732235
1	4	0.479464	-0.519641	1.478570
2	3	0.466723	-0.302890	1.236336
2	4	0.012500	-0.963634	0.988634
3	4	-0.454223	-1.408119	0.499673





2)The effect of average age of patient (variable 3) on mean infection risk (variable 4) is to be studied. For purposes of this ANOVA study, average age is to be classified ed into four categories: under 50, 50-54.9, 55.0-59.9, 60.0 and over. Assume that ANOVA model is applicable. Test whether or not the mean infection risk differs for the four age groups. Control the risk at α=0.10. State the alternatives and conclusion.

H_0:The mean infection risk is the same across all age groups.

H_a: The mean infection risk is not the same across all the are groups.

F_{stat}=0.56 <$F_{(0.10,3,109)}$ Critical= 2,1347 and the P-value = 0.6412 >0.10.

Decision: We do not reject null hypothesis.

Conclusion: We conclude that there is not enough evidence that the mean infection risk is the same across all age groups at 10% significant level.

The GLM Procedure

Dependent Variable: infprob

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3.0677159	1.0225720	0.56	0.6412
Error	109	198.3121071	1.8193771		
Corrected Total	112	201.3798230			

R-Square	Coeff Var	Root MSE	infprob Mean
0.015233	30.97323	1.348843	4.354867

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age_group	3	3.06771587	1.02257196	0.56	0.6412

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age_group	3	3.06771587	1.02257196	0.56	0.6412

3)Conduct a test of whether or not mean length of stay (variable 2) is the same in the four geographic regions. Then do the following questions.

ANOVA for Length of Stay Across Regions

The GLM Procedure

Dependent Variable: stay

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	103.5541834	34.5180611	12.31	<.0001
Error	109	305.6561972	2.8041853		
Corrected Total	112	409.2103805			

R-Square	Coeff Var	Root MSE	stay Mean
0.253059	17.35608	1.674570	9.648319

Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	103.5541834	34.5180611	12.31	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	103.5541834	34.5180611	12.31	<.0001

H_0:Mean length of stay is same in all four geographic region

H_a: Mean length of stay not the same in all four geographic region

$Yij = \mu. + \tau i + \epsilon ij$

Yij = Length of stay

τi =region

Decision:

F_{stat} value =12.31

$F_{(0.05,3,109)}$ critical value=2.6879

$F_{stat} > F_{critical}$ and

the P-value=0.0001<0.05

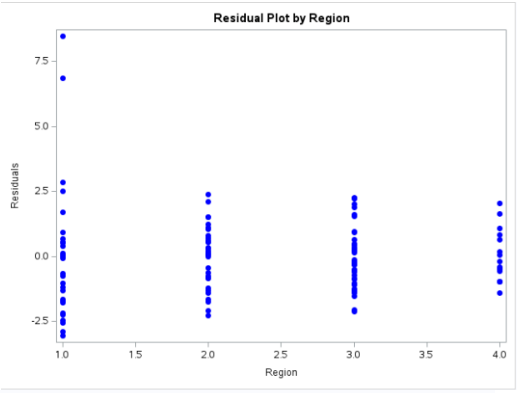
We reject the null hypothesis.

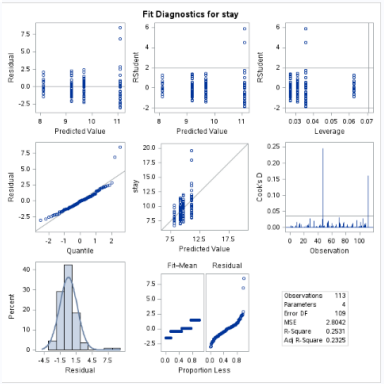
Conclusion:

We conclude that the mean length of stay is same in all four regions.

(a) Obtain the residuals and prepare aligned residual dot plots by region. Are any serious departure from ANOVA model?

From the residual plots by the regions. The output plot shows the variance of constant for the regions. We also notice that for the region 1, we see 2 outliers. Please refer the below results.





ANOVA for Length of Stay Across Regions

The GLM Procedure

Dependent Variable: stay

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	103.5541834	34.5180611	12.31	<.0001
Error	109	305.6561972	2.8041853		
Corrected Total	112	409.2103805			

R-Square	Coeff Var	Root MSE	stay Mean
0.253059	17.35608	1.674570	9.648319

Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	103.5541834	34.5180611	12.31	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	103.5541834	34.5180611	12.31	<.0001

(b) Examine by means of the Brown-Forsythe test whether or not the geographic region error variances
are equal.

H_0: The geographic region error variance among the groups are equal

H_a: The geographic region error variance among the groups are not equal

Brown-Forsythe Test	Levene's Test
P-Value=0.0064<0.05 F value=4.33 small F value	P-Value=0.0171 F value=3.54 small F value
Decision: We reject the null at 5% significant level	Decision: We reject the null at 5% significant level
Conclusion: The geographic region error variance among the groups are not equal.	Conclusion: The geographic region error variance among the groups are not equal.

ANOVA Model Fitting
Brown-Forsythe Test for Equal Variance Across Regions

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of stay Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
region	3	17.2864	5.7621	4.33	0.0064
Error	109	145.2	1.3319		

Levene's Test for Homogeneity of Variance

The GLM Procedure

Levene's Test for Homogeneity of stay Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
region	3	649.6	216.5	3.54	0.0171
Error	109	6665.3	61.1491		

3)c)For each geographic region, calculate Yi and si. Examine the three relations found in the table on page 791 and determine the transformation that is the most appropriate one here. What do you conclude?

We obtained the mean and standard deviation. Please see the below result.

We also calculated the metrics for transformation decision by region. The result shows s2 has a very close values and looks stable than the other regions. Si/Yi bar

Means and Standard Deviations for Length of Stay by Region

The MEANS Procedure

Analysis Variable : stay			
region	N Obs	Mean	Std Dev
1	28	11.0889286	2.6696155
2	32	9.6834375	1.1929378
3	37	9.1913514	1.2249879
4	16	8.1137500	1.0031210

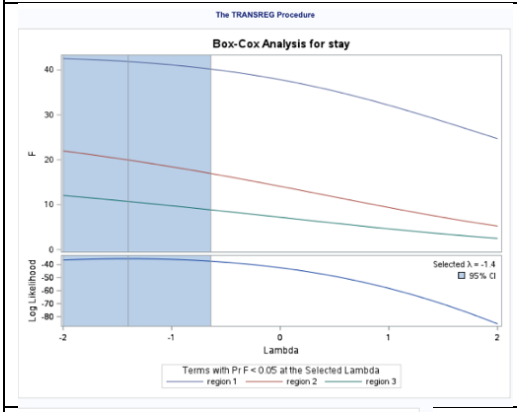
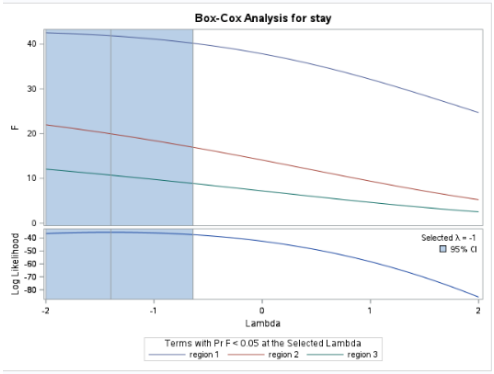
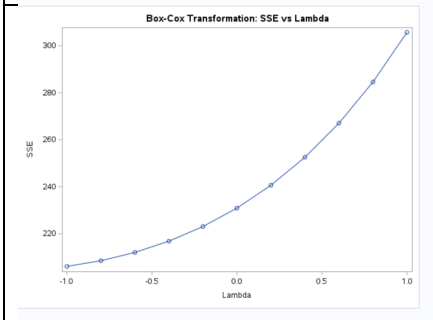
Calculated Metrics for Transformation Decision by Region

Obs	region	s1	s2	s3
1	.	0.37868	0.19811	0.020533
2	1	0.64270	0.24075	0.021710
3	2	0.14696	0.12319	0.012722
4	3	0.16326	0.13328	0.014500
5	4	0.12402	0.12363	0.015237

(d) Use the Box-Cox procedure to and an appropriate power transformation of Y . Evaluate SSE for the values of given in Table 18.6. Does $\lambda = -1$, a reciprocal transformation, appear to be reasonable, based on the Box-Cox procedure?

Yes, the reciprocal transformation appeared to be reasonable based on the Box-Cox procedure. We obtained λ value= -1.4, which is very close to $\lambda = -1$, a reciprocal transformation.

Please refer to the below table and with values and graph. We observed that the SSE is minimum around -1. We can conclude that $\lambda = -1$, this is reasonable based on Box-Cox(Y*=1/Y) transformation.

Box-Cox without convenient	Box-Cox with convenient
	
	**Box-Cox SSE Results for Different Lambda Values** Obs / lambda / sse 1 / -1.0 / 206.148 2 / -0.8 / 208.554 3 / -0.6 / 212.085 4 / -0.4 / 216.874 5 / -0.2 / 223.087 6 / 0.0 / 230.930 7 / 0.2 / 240.652 8 / 0.4 / 252.564 9 / 0.6 / 267.042 10 / 0.8 / 284.550 11 / 1.0 / 305.656

(e) Use the reciprocal transformation Y'= 1/Y to obtain transformed response data. Fit ANOVA model to the transformed data and obtain the residuals.
We did the reciprocal transformation Y'= 1/Y to obtain transformed response data. Fitted the ANOVA model to the transformed data and obtained the residuals. Please see the below result. First 10 values are shown. The smaller residual indicates the better fit. The negative residuals indicate a larger discrepancy between the predicted and the actual value.

Obs	region	stay_reciprocal	predicted	residual
1	4	0.14025	0.12494	0.015316
2	2	0.11338	0.10486	0.008520
3	3	0.11990	0.11066	0.009242
4	4	0.11173	0.12494	-0.013205
5	1	0.08929	0.09424	-0.004954
6	2	0.10246	0.10486	-0.002400
7	3	0.10331	0.11066	-0.007357
8	2	0.08945	0.10486	-0.015413
9	3	0.11534	0.11066	0.004678
10	1	0.11312	0.09424	0.018883

The GLM Procedure

Dependent Variable: stay_reciprocal

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.01034953	0.00344984	14.79	<.0001
Error	109	0.02542843	0.00023329		
Corrected Total	112	0.03577796			

R-Square	Coeff Var	Root MSE	stay_reciprocal Mean
0.289271	14.27848	0.015274	0.106971

Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	0.01034953	0.00344984	14.79	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	0.01034953	0.00344984	14.79	<.0001

(f) Examine by means of the Brown-Forsythe test whether or not the geographic region variances for the transformed response variable are equal. use $\alpha = .01$.
H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: not all the variances are equal
F_{stat} value= 0.97, P-value = 0.41> 0.01. $F_{(0.01,3,109)}$ critical value=3.97. F_{stat} < F critical.
Decision :We do not reject the null hypothesis at $\alpha = 0.01$
Conclusion: There is not enough evidence to conclude that the variances are different across the geographic regions.

Brown-Forsythe Test for Equal Variance Across Regions

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of stay_reciprocal Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
region	3	0.000246	0.000082	0.97	0.4100
Error	109	0.00924	0.000085		