**Module Assignment**

**Module 9**

**QMB-6304 Foundations of Business Statistics**

**#Debjani Sarma.**

**rm(list = ls())**

**library(rio)**

**library(car)**



**Preprocessing**

1. Load the file “6304 Module 9 Assignment Data.xlsx” into R. This data set includes several variables on 437 counties in six midwestern states. This will be your master data set

**master.states = import("6304 Module 9 Assignment Data.xlsx")**

**set.seed(24173877)**

1. Create a single data frame for your analysis. This will be your primary data set. It will meet the following characteristics:
   1. Includes state, popdensity, density.category, percollege, and inmetro variables from the master (N=437) data set.

**mydata = master.states[, c("state", "popdensity", "density.category", "percollege", "inmetro")]**

* 1. Be a random sample of n=250, with each state making up 20% of the total sample. For example, 20% of 250 observations will be from the state of Illinois, 20% from Indiana, etc. Use your U number as the random number seed.

**> mydata.il = mydata[mydata$state == "IL", ]**

**> mydata.il = mydata.il[sample(1:nrow(mydata.il),50), ]**

**> mydata.in = mydata[mydata$state == "IN", ]**

**> mydata.in = mydata.in[sample(1:nrow(mydata.in),50), ]**

**> mydata.mi = mydata[mydata$state == "MI", ]**

**> mydata.mi = mydata.mi[sample(1:nrow(mydata.mi),50), ]**

**> mydata.oh = mydata[mydata$state == "OH", ]**

**> mydata.oh = mydata.oh[sample(1:nrow(mydata.oh),50), ]**

**> mydata.wi = mydata[mydata$state == "WI", ]**

**> mydata.wi = mydata.wi[sample(1:nrow(mydata.wi),50), ]**

**> mydata = rbind(mydata.il, mydata.in, mydata.mi, mydata.oh, mydata.wi)**

**> rm(mydata.il)**

**> rm(mydata.in)**

**> rm(mydata.mi)**

**> rm(mydata.oh)**

**> rm(mydata.wi)**

* 1. The variables state, density.category, and inmetro will be factors.

**> mydata$state = as.factor(mydata$state)**

**> mydata$density.category = as.factor(mydata$density.category)**

**> mydata$inmetro = as.factor(mydata$inmetro)**

**Analysis**

Using your primary data set:

1. Show the results of an str() command.

**> str(mydata)**

**'data.frame': 250 obs. of 5 variables:**

**$ state : Factor w/ 5 levels "IL","IN","MI",..: 1 1 1 1 1 1 1 1 1 1 ...**

**$ popdensity : num 35.7 69.3 35.3 26.9 43.7 ...**

**$ density.category: Factor w/ 5 levels "100 to 249","250 to 749",..: 5 3 5 5 5 5 5 1 5 3 ...**

**$ percollege : num 14.1 15.5 22 20.7 11.7 ...**

**$ inmetro : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...**

1. Show the results of the table() command on the state variable.

**> table(mydata$state)**

**IL IN MI OH WI**

**50 50 50 50 50**

1. Determine if percollege has an equal variance across all five states. Briefly interpret your results. If you determine there is a difference in variances across the states, discuss where is/are the differences.

**> percollege.levenetest = leveneTest(percollege ~ state, data = mydata)**

**> percollege.levenetest**

**Levene's Test for Homogeneity of Variance (center = median)**

**Df F value Pr(>F)**

**group 4 0.3449 0.8474**

**245**

The null hypothesis suggests that percollege has equal variance across all five states.

With p=0.8474, we fail to reject the null hypothesis , which means we failed to find a difference across all five states.

**> aggregate(percollege ~ state, data = mydata, var)**

**state percollege**

**1 IL 32.81501**

**2 IN 29.11697**

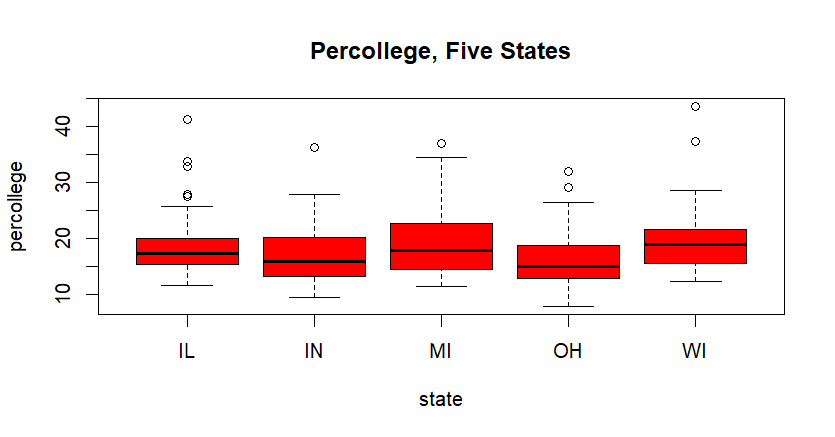
**3 MI 37.33591**

**4 OH 31.05126**

**5 WI 32.67329**

Here, we calculated the variance of percollege for each 5 states. After calculating variance, we see that all the states have different variances percollege.

**> boxplot(percollege ~ state, data = mydata,main="Percollege, Five States", col="red")**



Based on the boxplot, we see there is a difference in heights , including height and dept of the whiskers across all states.

So, yes there is a difference which the leveneTest failed to pick up, which could be due to small sample size.

1. Conduct a one-way analysis of variance with percollege as the dependent variable and state as the independent variable. Plot the results of a Tukey HSD test. Briefly explain the results shown in the plot, stating between which pairs of states do/do not show significant population mean differences in percollege. Make sure factor level names can be clearly and completely read on the appropriate axis of your plot.

**> anovatest = aov(percollege ~ state, data = mydata)**

**> summary(anovatest)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**state 4 435 108.8 3.338 0.011 \***

**Residuals 245 7987 32.6**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**> anovatukey = TukeyHSD(anovatest)**

**> anovatukey**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = percollege ~ state, data = mydata)**

**$state**

**diff lwr upr p adj**

**IN-IL -1.7196934 -4.8578693 1.418482423 0.5595738**

**MI-IL 0.6924303 -2.4457456 3.830606104 0.9739946**

**OH-IL -2.4367867 -5.5749626 0.701389121 0.2090921**

**WI-IL 0.8409479 -2.2972280 3.979123710 0.9477668**

**MI-IN 2.4121237 -0.7260522 5.550299527 0.2180963**

**OH-IN -0.7170933 -3.8552691 2.421082545 0.9704516**

**WI-IN 2.5606413 -0.5775346 5.698817133 0.1677388**

**OH-MI -3.1292170 -6.2673928 0.008958863 0.0510604**

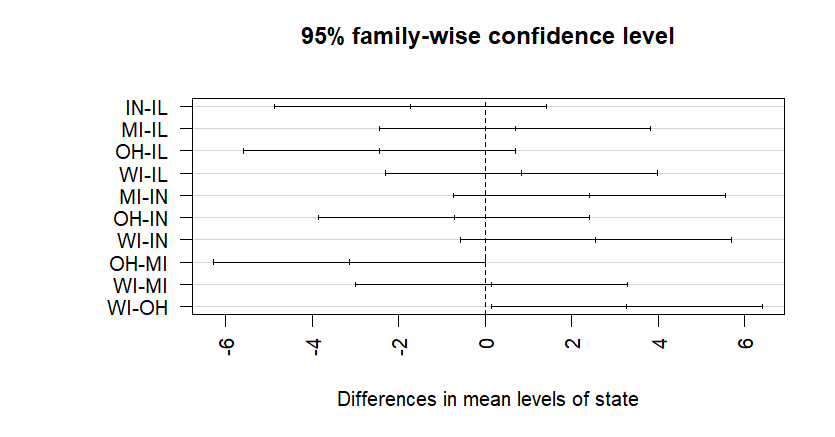
**WI-MI 0.1485176 -2.9896582 3.286693452 0.9999352**

**WI-OH 3.2777346 0.1395587 6.415910434 0.0357480**

**> par(mar = c(5.1,8,4.1,2.1))**

**> plot(anovatukey, las = 2)**

**> par(mar = c(5.1,4.1,4.1,2.1))**



With p-value>0.05 , we reject the null hypothesis in anova test, indicating there are significant differences in the means of percollege across the states.

There are no significant differences between most state pairs due to p-values > 0.05. The significant differences are between WI-OH (with p = 0.0357) and OH-MI (with p = 0.05106), which are on the verge of significance but do not show definitive significant mean differences across all state pairs.

The confidence intervals containing 0 in the Tukey HSD suggest that for many comparisons, we cannot confidently conclude there is a meaningful difference.

1. Repeat Steps 3 and 4 above using percollege as the dependent variable and density.category as the independent variable. Again, briefly explain your analysis results and make sure category names can be clearly and completely read on the appropriate axis of your plot.

**> percollege.density.category = leveneTest(percollege ~ density.category, data = mydata)**

**> percollege.density.category**

**Levene's Test for Homogeneity of Variance (center = median)**

**Df F value Pr(>F)**

**group 4 3.0667 0.01722 \***

**245**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**> aggregate(percollege ~ density.category, data = mydata, var)**

**density.category percollege**

**1 100 to 249 40.75994**

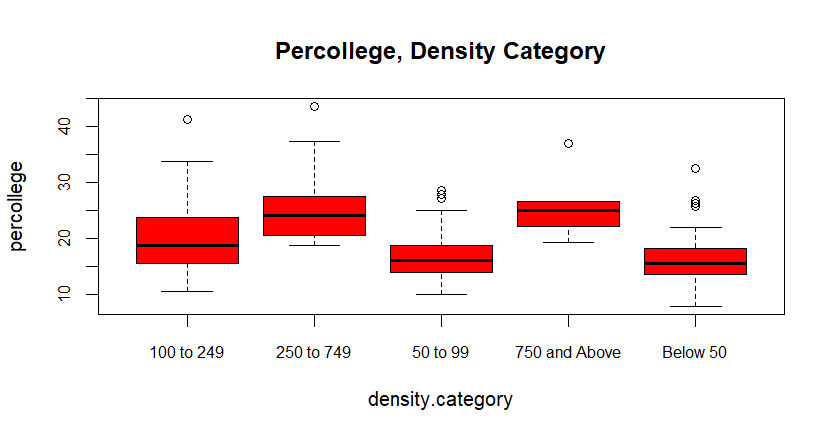
**2 250 to 749 41.87131**

**3 50 to 99 16.89753**

**4 750 and Above 32.95574**

**5 Below 50 16.49464**

**> boxplot(percollege ~ density.category, data = mydata,main="Percollege, Density Category", col="red", cex.axis=0.8)**



As per above levengeTest, the p-value=0.01722, is significant indicating that the variance in percollege differs significantly across the density categories

Using aggregate function, we showed the difference mean percollege across density category.

Moreover, using boxplot, we see there is a difference in heights , including height and dept of the whiskers of percollege across all density category.

**> anovatest.density.category = aov(percollege ~ density.category, data = mydata)**

**> summary(anovatest.density.category)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**density.category 4 2550 637.6 26.61 <2e-16 \*\*\***

**Residuals 245 5872 24.0**

**---**

**> anovatukey.density.category = TukeyHSD(anovatest.density.category)**

**> anovatukey.density.category**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = percollege ~ density.category, data = mydata)**

**$density.category**

**diff lwr upr p adj**

**250 to 749-100 to 249 5.6104002 2.26750920 8.9532911 0.0000626**

**50 to 99-100 to 249 -3.3599604 -5.86705221 -0.8528686 0.0026069**

**750 and Above-100 to 249 5.3783580 -0.07987359 10.8365895 0.0556388**

**Below 50-100 to 249 -4.0510300 -6.46766347 -1.6343964 0.0000642**

**50 to 99-250 to 749 -8.9703606 -12.06725697 -5.8734642 0.0000000**

**750 and Above-250 to 749 -0.2320422 -5.98508919 5.5210048 0.9999657**

**Below 50-250 to 749 -9.6614301 -12.68556231 -6.6372979 0.0000000**

**750 and Above-50 to 99 8.7383184 3.42718490 14.0494519 0.0000930**

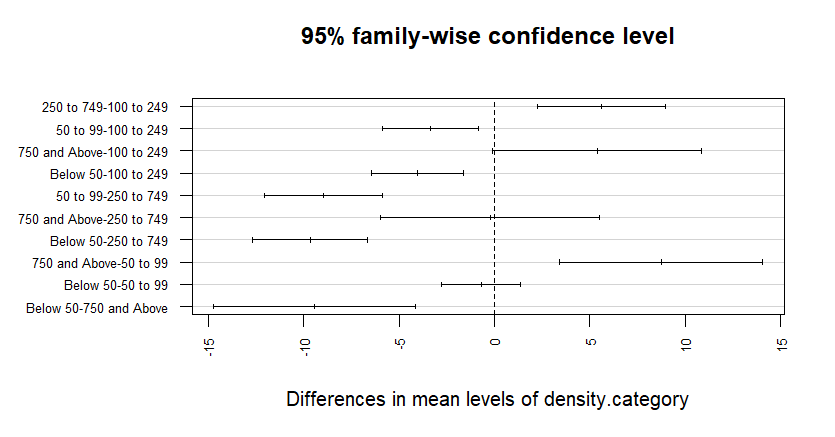
**Below 50-50 to 99 -0.6910695 -2.75406851 1.3719295 0.8887664**

**Below 50-750 and Above -9.4293879 -14.69842457 -4.1603513 0.0000158**

**> par(mar = c(5.1,8,4.1,2.1))**

**> plot(anovatukey.density.category, las = 2, cex.axis=0.65)**

**> par(mar = c(5.1,4.1,4.1,2.1))**



Here anova test,using percollege as the dependent variable and both state and inmetro as the independent variables is implemented.

With p-value<2e-16, it is very significant, we reject the null hypothesis , which indicates that there is no difference in percollege across density category.

Below are the categories based on their p-values, which are significant and non-significant:

**Significant:**

250 to 749-100 to 249

50 to 99-100 to 249

750 and Above-100 to 249

Below 50-100 to 249

50 to 99-250 to 749

750 and Above-50 to 99

Below 50-250 to 749

Below 50-750 and Above

**Not significant:**

750 and Above-250 to 749

Below 50-50 to 99

1. Conduct a two-way ANOVA using percollege as the dependent variable and both state and inmetro as the independent variables. Plot the results of a Tukey HSD test to show whether/where there are differences in percollege. Briefly explain the results shown in the plot, stating if state and inmetro together appear to show significant mean differences in percollege. Make sure the names of levels of independent variables can be clearly and completely read on the appropriate axis of your plot. Be sure to include and interpret an appropriate test for equality of variances.

**> anovatest.inmetro = aov(percollege ~ state + inmetro, data = mydata)**

**> summary(anovatest.inmetro)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**state 4 435 108.8 4.528 0.00152 \*\***

**inmetro 1 2123 2122.9 88.338 < 2e-16 \*\*\***

**Residuals 244 5864 24.0**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

Here anova test,using percollege as the dependent variable and both state and inmetro as the independent variables is implemented.

**State**: p = 0.00152 indicates a significant difference in means across states.

**Inmetro**: p < 2e-16 indicates a significant difference between inmetro levels..

**> anovatukey.inmetro = TukeyHSD(anovatest.inmetro)**

**> anovatukey.inmetro**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = percollege ~ state + inmetro, data = mydata)**

**$state**

**diff lwr upr p adj**

**IN-IL -1.7196934 -4.4142254 0.9748385 0.4030404**

**MI-IL 0.6924303 -2.0021017 3.3869622 0.9549370**

**OH-IL -2.4367867 -5.1313187 0.2577452 0.0973033**

**WI-IL 0.8409479 -1.8535841 3.5354798 0.9119048**

**MI-IN 2.4121237 -0.2824083 5.1066556 0.1032791**

**OH-IN -0.7170933 -3.4116253 1.9774387 0.9490116**

**WI-IN 2.5606413 -0.1338907 5.2551732 0.0713482**

**OH-MI -3.1292170 -5.8237489 -0.4346850 0.0137370**

**WI-MI 0.1485176 -2.5460144 2.8430496 0.9998810**

**WI-OH 3.2777346 0.5832026 5.9722665 0.0084381**

**$inmetro**

**diff lwr upr p adj**

**1-0 6.161371 4.85219 7.470551 0**

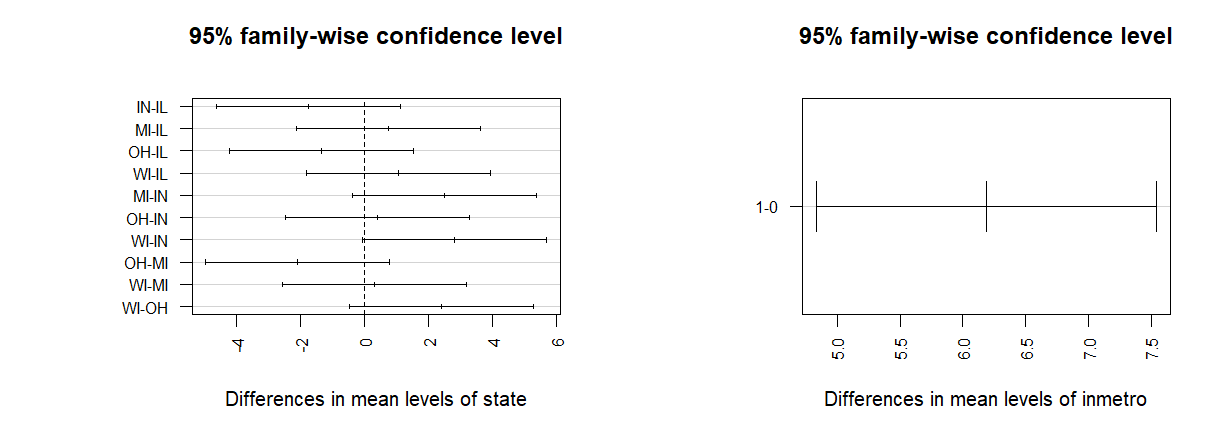
**1-0 6.188077 4.831588 7.544566 0**

**>**

**> par(mar = c(5.1,8,4.1,2.1))**

**> plot(anovatukey.inmetro, las = 2, cex.axis=0.8)**

**> par(mar = c(5.1,4.1,4.1,2.1))**

****

The Tukey HSD results show significant differences between inmetro levels and between some state pairs (e.g., OH-MI and WI-OH).

While the results of the Tukey test are significant for state and inmetro, the confidence intervals for some pairwise comparisons (like OH-MI and WI-OH) include 0, suggesting the effect is not always perfect.

**> percollege.inmetro = leveneTest(percollege ~ inmetro, data = mydata)**

**> percollege.inmetro**

**Levene's Test for Homogeneity of Variance (center = median)**

**Df F value Pr(>F)**

**group 1 19.9 1.238e-05 \*\*\***

**248**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**> percollege.state = leveneTest(percollege ~ state, data = mydata)**

**> percollege.state**

**Levene's Test for Homogeneity of Variance (center = median)**

**Df F value Pr(>F)**

**group 4 0.3449 0.8474**

**245**

**>**

**> aggregate(percollege ~ inmetro, data = mydata, var)**

**inmetro percollege**

**1 0 16.95924**

**2 1 47.09798**

**> aggregate(percollege ~ state, data = mydata, var)**

**state percollege**

**1 IL 32.81501**

**2 IN 29.11697**

**3 MI 37.33591**

**4 OH 31.05126**

**5 WI 32.67329**

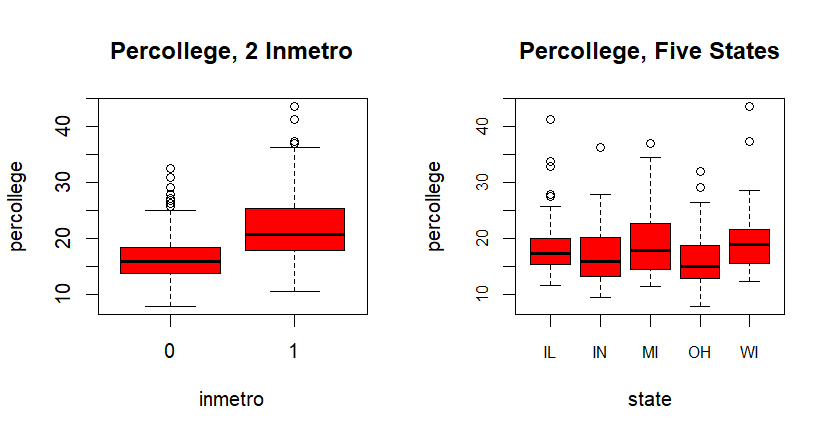
**>**

**> par(mfrow = c(1,2))**

**> boxplot(percollege ~ inmetro, data = mydata,main="Percollege, 2 Inmetro", col="red")**

**> boxplot(percollege ~ state, data = mydata,main="Percollege, Five States", col="red", cex.axis = 0.8)**

**>**



**Equality of Variance:**

Inmetro:

As per above levengeTest, the p-value=1.238e-05, which is statistically significant, indicating that we reject the null hypothesis. This indicates that there is a difference in mean percollege across inmetro. Using aggregate, we see the difference for inmetro 0 and 1 are 16.95924and 47.09798 respectively.

Using boxplot, we see there is a difference in heights , including height and dept of the whiskers of percollege across inmetro.

State:

As per above levengeTest, the p-value=0.8474, which is not statistically significant, indicating that we fail to reject the null hypothesis. This indicates that there is not enough evidence to prove there is a difference in mean percollege across states.

However, using aggregate, we see the difference for all states.

Moreover, using boxplot, we see there is a difference in heights , including height and dept of the whiskers of percollege across all states.

The leveneTest might not have picked up the differnce due to small sample size.

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above instructions and 2) the results of those instructions. The first line of your script file should be a “#” comment line showing your name as it appears in Canvas. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed and submitted by the time stated on Canvas. No collaboration of any sort is allowed on this assignment.**