ANOVA

Khadija Sitabkhan

2021-11-10

library(Stat2Data)

## Warning: package 'Stat2Data' was built under R version 4.0.3

library(agricolae)

## Warning: package 'agricolae' was built under R version 4.0.4

library(tidyr)

## Warning: package 'tidyr' was built under R version 4.0.5

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

1. Nationally, the abuse of methamphetamine has become a concern, not only because of the effects of drug abuse, but also because of the dangers associated with the labs that produce them. A random sample of a total of 11 counties in Iowa (stratified by size of county – small, medium, or large) produced the following ANOVA table relating the number of methamphetamine labs to the size of the county. Use this table to answer the following questions:

One-way ANOVA: meth labs versus type

1. Fill in the values missing from the table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F | P |
| type | 2 | 39.01 | 19.505 | 4.9286 | 0.04028 |
| Error | 8 | 31.66 | 3.9575 |  |  |
| Total | 10 | 70.67 |  |  |  |

1. What does the MS for the county type tell you?

*The variability in mean methamphetamine labs for different county size*

1. Find the p-value for the F-test in the table. [hint: **carefully** use the pf() function in R]

pf(4.9286,2,8,lower.tail = F)

## [1] 0.04028158

1. Describe the hypotheses tested by the F-test in the table, and using the p-value from part (c), give an appropriate conclusion.

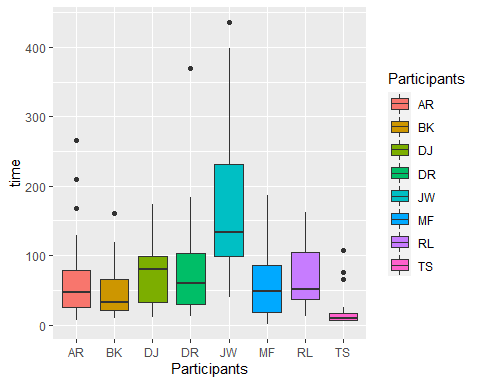
*The null hypothesis is average methamphetamine labs for the 3 county sizes. We observe that it is same and we have evidence against null hypothesis as per p-value. Hence, we can say there is a significance difference between average of methamphetamine between 3 county sizes.*

1. A group of friends who participate in a “fantasy baseball” league became curious about whether some of them take significantly more or less time to make their selections in the “fantasy draft” through which they select players. In the data (FantasyBaseball), there are records: the times (in seconds) that each of the eight friends (identified by their initials) took to make their 24 selections in 2008.

library(Stat2Data)  
data(FantasyBaseball)

1. Produce boxplots and calculate descriptive statistics to compare the selection times for each participant. Comment on what they reveal. Also identify (by initials) which participant took the longest and which took the shortest time to make their selections.

FBaseBall <- FantasyBaseball %>% pivot\_longer(DJ:MF,names\_to = "Participants", values\_to = "time")   
ggplot(FBaseBall)+geom\_boxplot(aes(x=Participants,y=time,fill=Participants))



Descriptive Statistics

FBaseBall %>% group\_by(Participants)%>% summarise(mean = mean(time),sd = sd(time))

## # A tibble: 8 x 3  
## Participants mean sd  
## <chr> <dbl> <dbl>  
## 1 AR 68.3 66.9  
## 2 BK 48.0 39.3  
## 3 DJ 69.6 41.6  
## 4 DR 80.1 75.8  
## 5 JW 164. 104.   
## 6 MF 63.8 56.0  
## 7 RL 67.1 44.6  
## 8 TS 19.3 25.8

*From the data above we can say that ANOVA on this data can not be performed as the assumption of constants variance is not fulfilled since the largest standard deviation(sd) is far greater than twice of the smallest sd. JW took longest time and TS took least time to make selection.*

1. Conduct a one-way ANOVA to assess whether the data provide evidence that averages as far apart as these would be unlikely to occur by chance alone if there really were no differences among the participant in terms of their selection times. For now, assume that all conditions are met. Report the ANOVA table, test statistics, and p-value. Also summarize your conclusion.

anova <- aov(time ~ Participants, data = FBaseBall)   
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Participants 7 287196 41028 10.89 1.79e-11 \*\*\*  
## Residuals 184 693126 3767   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*The null hypothesis can be rejected since the The p-value is very small, so we can say that the observed differences in mean is unlikely due to chance.*

1. Use Fisher’s LSD procedure to assess which participants’ average selection times differ significantly from which others.

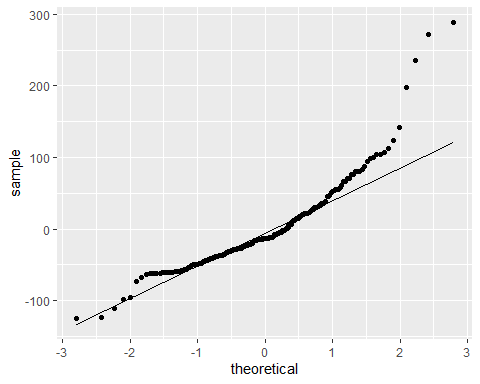
print(LSD.test(anova, trt = "Participants", p.adj = "none"))

## $statistics  
## MSerror Df Mean CV t.value LSD  
## 3766.99 184 72.52083 84.63196 1.972941 34.95591  
##   
## $parameters  
## test p.ajusted name.t ntr alpha  
## Fisher-LSD none Participants 8 0.05  
##   
## $means  
## time std r LCL UCL Min Max Q25 Q50 Q75  
## AR 68.29167 66.89153 24 43.574107 93.00923 7 266 26.00 46.5 79.25  
## BK 47.95833 39.25112 24 23.240774 72.67589 9 161 20.75 33.0 65.25  
## DJ 69.62500 41.61502 24 44.907441 94.34256 11 174 32.25 79.5 99.00  
## DR 80.12500 75.83467 24 55.407441 104.84256 13 369 29.75 59.5 103.50  
## JW 163.87500 104.16555 24 139.157441 188.59256 39 436 98.50 134.0 231.25  
## MF 63.83333 55.99664 24 39.115774 88.55089 1 187 18.75 49.0 86.50  
## RL 67.12500 44.55120 24 42.407441 91.84256 13 162 36.75 51.5 105.25  
## TS 19.33333 25.82999 24 -5.384226 44.05089 5 107 6.00 9.5 17.00  
##   
## $comparison  
## NULL  
##   
## $groups  
## time groups  
## JW 163.87500 a  
## DR 80.12500 b  
## DJ 69.62500 b  
## AR 68.29167 b  
## RL 67.12500 b  
## MF 63.83333 b  
## BK 47.95833 bc  
## TS 19.33333 c  
##   
## attr(,"class")  
## [1] "group"

*Fisher’s LSD gives an idea about the different groups that could be formed from the data. It is observed that JW and TS have significant difference in there selection time so they are categorized in different group. DR,DJ,AR,RL,MF and BK are in one group (group b), JW is in group ‘a’ and ‘BK’,‘TS’ are in group ‘c’. Participants in the same group have no significant difference. Participants from different groups have significant difference.*

1. Produce the normal probability plot of the residuals for the ANOVA model in (b), and comment on the appropriateness of the ANOVA model for these data.

FBaseBall$Res= anova$residuals   
ggplot(FBaseBall, aes(sample = Res)) + geom\_qq() + geom\_qq\_line()



*It is seen that the points do not fit on line properly in the normal probability plot, This implies that the data is not normal. Since the normality assumption is false, Annova cannot be used on this data*

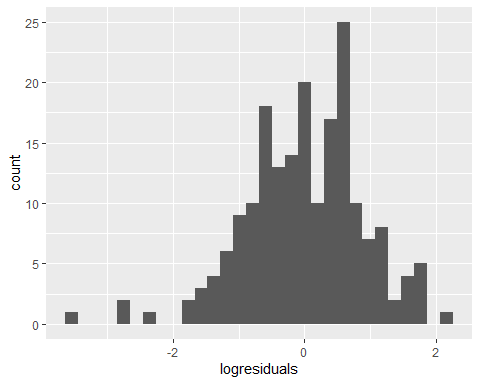
1. Transform the selection times using the natural log. Repeat your analysis of the data (b,c) and report your findings. Check the conditions necessary for conducting an ANOVA.

loganova <- aov(log(time) ~ Participants, data = FBaseBall)   
   
# display the results   
summary(loganova)

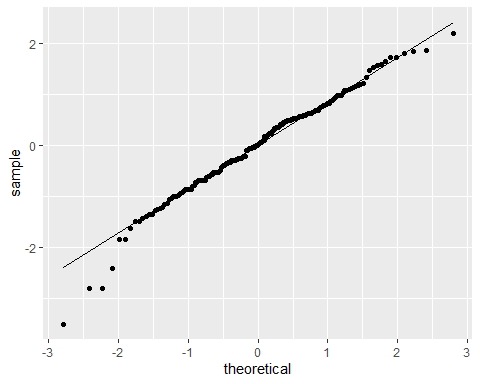
## Df Sum Sq Mean Sq F value Pr(>F)   
## Participants 7 78.75 11.250 12.99 1.54e-13 \*\*\*  
## Residuals 184 159.37 0.866   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# store residuals   
FBaseBall$logresiduals <- loganova$residuals  
## histogram of residuals   
ggplot(FBaseBall,aes(x=logresiduals)) + geom\_histogram()

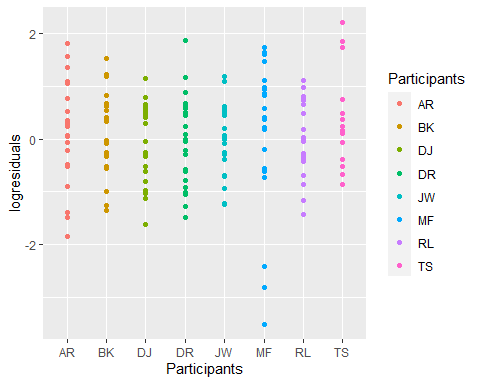
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## QQ plot of residuals   
ggplot(FBaseBall,aes(sample = logresiduals)) + geom\_qq() + geom\_qq\_line()



## Residuals vs Groups   
ggplot(FBaseBall,aes(x=Participants,y=logresiduals, color=Participants)) + geom\_point()



*The null hypothesis can be rejected on the basis of the p-value. hence we can say that there significant difference in the average selection time between participants. After using a log transformation, the histogram seen is normally distributed and normality assumption is adequate. Since the conditions are met, Anova can be performed.*

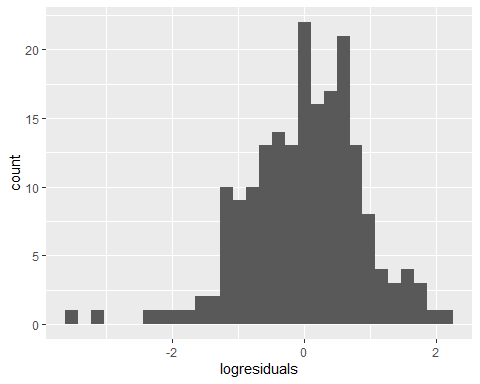
1. Using the transformed data in (e), conduct a two-way ANOVA, and check the conditions necessary for conducting an ANOVA.

anova\_two <- aov(log(time) ~ Participants + Round, data = FBaseBall)   
   
# display the results   
summary(anova\_two)

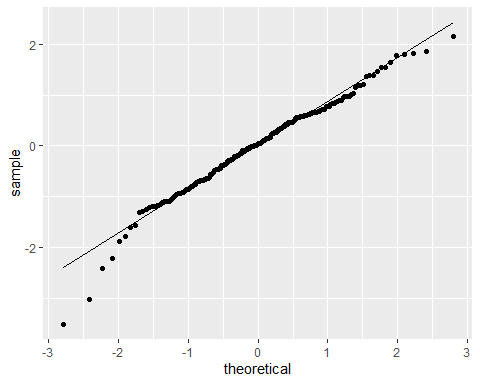
## Df Sum Sq Mean Sq F value Pr(>F)   
## Participants 7 78.75 11.250 13.87 2.29e-14 \*\*\*  
## Round 1 10.95 10.948 13.50 0.000313 \*\*\*  
## Residuals 183 148.42 0.811   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# store residuals   
FBaseBall$logresiduals <- anova\_two$residuals   
   
## histogram of residuals   
ggplot(FBaseBall,aes(x=logresiduals)) + geom\_histogram()

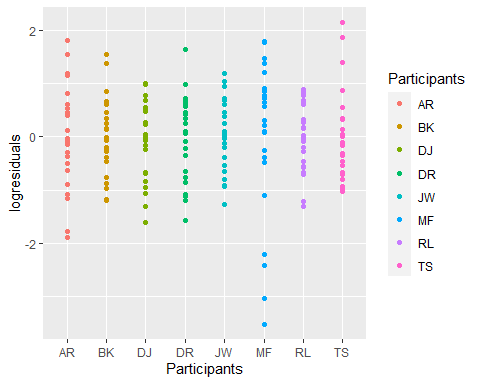
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## QQ plot of residuals   
ggplot(FBaseBall,aes(sample = logresiduals)) + geom\_qq() + geom\_qq\_line()



## Residuals vs Groups   
ggplot(FBaseBall,aes(x=Participants,y=logresiduals, color=Participants)) + geom\_point()



*Since the histogram follows normal distribution and qq-plot is roughly linear so we can say that the necessary conditions are met to conduct the two way anova on the data.*

1. List the conditions, under which the following statement is true:

* “In one-way ANOVA, MSGroups/MSE has a F distribution with parameters K-1 and n-K”

Here, we have K groups and n observations.

*Conditions for one-way anova :*

*1) The data should be randomly selected from the population as the sample should be the representation of the population.*

*2) Normality assumption check - The residuals of the sample should follow normal distribution and qq-plot should be linear.*

*3) The largest standard deviation should be smaller than the twice of the smallest standard deviation.*