Exercise 3 Test of Assumptions

Ayop, Kyla M. and Garridos, Charlene P.

2024-02-02

Contents

1 R Library

```
library(tidyverse)
library(car)
library(ggplot2)
```

2 Problem

Questionnaire Color. In an experiment to investigate the effect of color paper (blue, green, orange) on response rates for questionnaires distributed by the "windshield method" in supermarket parking lots, 15 representative supermarket parking lots were chosen in a metropolitan area and each color was assigned at random to five of the parking lots. The response rates (in percent) follow.

COLOR	RESPONSE RATE
BLUE	28 26 31 27 35
GREEN	34 39 25 31 29
ORANGE	31 25 27 29 28

3 Dataset

```
color=rep(c('blue','green','orange'), each=5)
response =c(28,26,31,27,35, 34,39,25,31,29, 31,25,27,29,28)
color=as.factor(color)
ques=data.frame(color,response)
ques
```

```
##     color response
## 1    blue       28
## 2    blue       26
## 3    blue       31
## 4    blue       27
## 5    blue       35
## 6   green       34
## 7   green       39
## 8   green       25
## 9   green       31
## 10  green       29
## 11 orange       31
## 12 orange       25
## 13 orange       27
## 14 orange       29
## 15 orange       28
```

4 Analysis of Variance

```r
#ANOVA
anova=aov(response~color, data=ques)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## color        2  32.93   16.47   1.072  0.373
## Residuals   12 184.40   15.37
```

Interpretation

Hypothesis and Test Statistic:

> H_O: There are no differences in the mean responses rate percentage among the different color papers.
>
> $$[\mu_1 = \mu_2 = \mu_3]$$
>
> H_a: At least one color paper gave a different response rate percentage.
>
> $$[\mu_i \neq \mu_j, i \neq j]$$
>
> Test statistic: One-way analysis of Variance at $\alpha = 0.05$.

Decision Rule: Reject H_o at $\alpha = 0.05$ if $F_c > F_t = F_{2,12}(0.05) = 3.89$. Otherwise fail to reject H_o.

Decision: Since $F_c = 1.07 < F_t = 3.89$, then we fail to reject H_o.

The results of the test shows the independent variable color with the model error residuals, with the degrees of freedom 2 and 12, respectively.

Moreover, the total variability in the data is measured by adding up sum of squares due to treatment which is equal to 32.93 that is between treatment and the sum of squares due to error equal to 184.40 which is within treatments, hence, the total variability is equal to 217.33.

The mean square of treatment is equal to 16.47 is the measure of variability between groups and this is how much of the overall variability in the data which is greater than the mean square of error equal to 15.37 that measures the unexplained variability within groups that cannot be accounted for by the treatments.

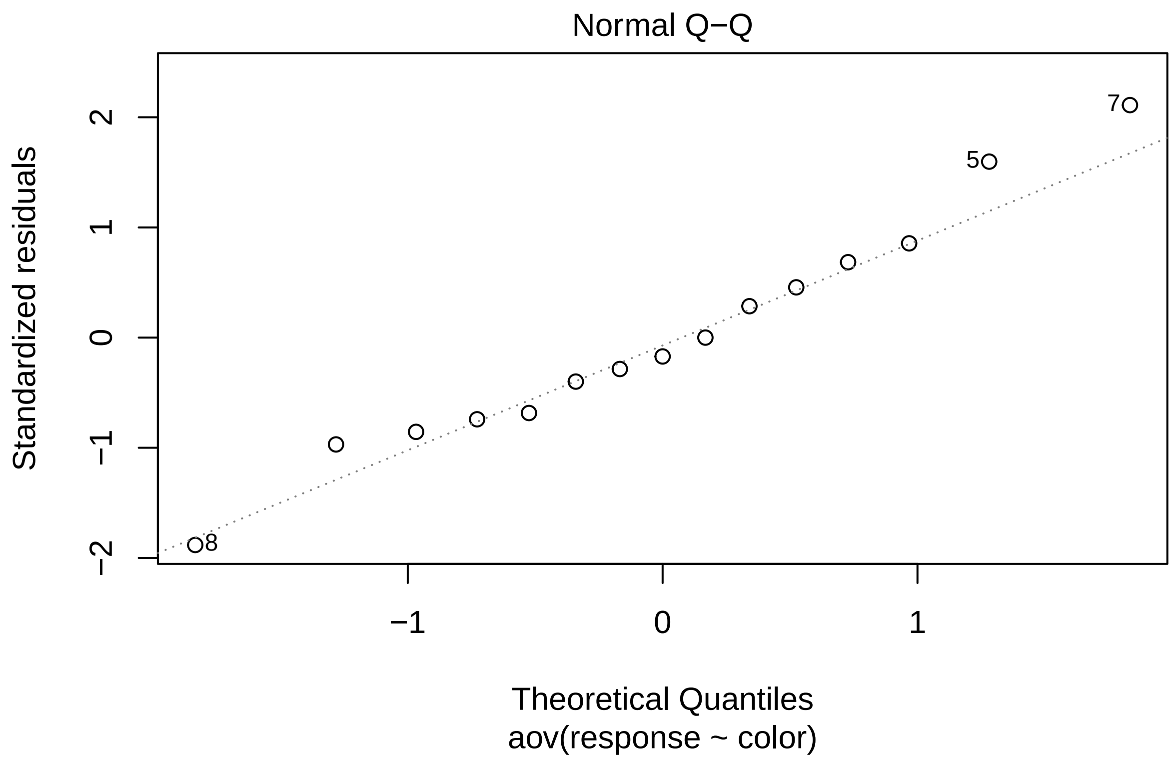
The F value is equal to 1.072 is the test statistic from the F test, since the F value is small,more likely the variation associated with the independent variable is purely from to chance. Along with Pr(>F) equal to 0.373 which is the p-value of the F statistic.

Conclusion: Since $F_c = 1.07 < F_t = 3.89$, and $p - value = 0.373 > \alpha = 0.05$, then we fail to reject H_o. Thus, we fail to reject the null hypothesis at 5% level of significance. There is no sufficient evidence to support the claim that at least one color paper gave a different response rate percentage.

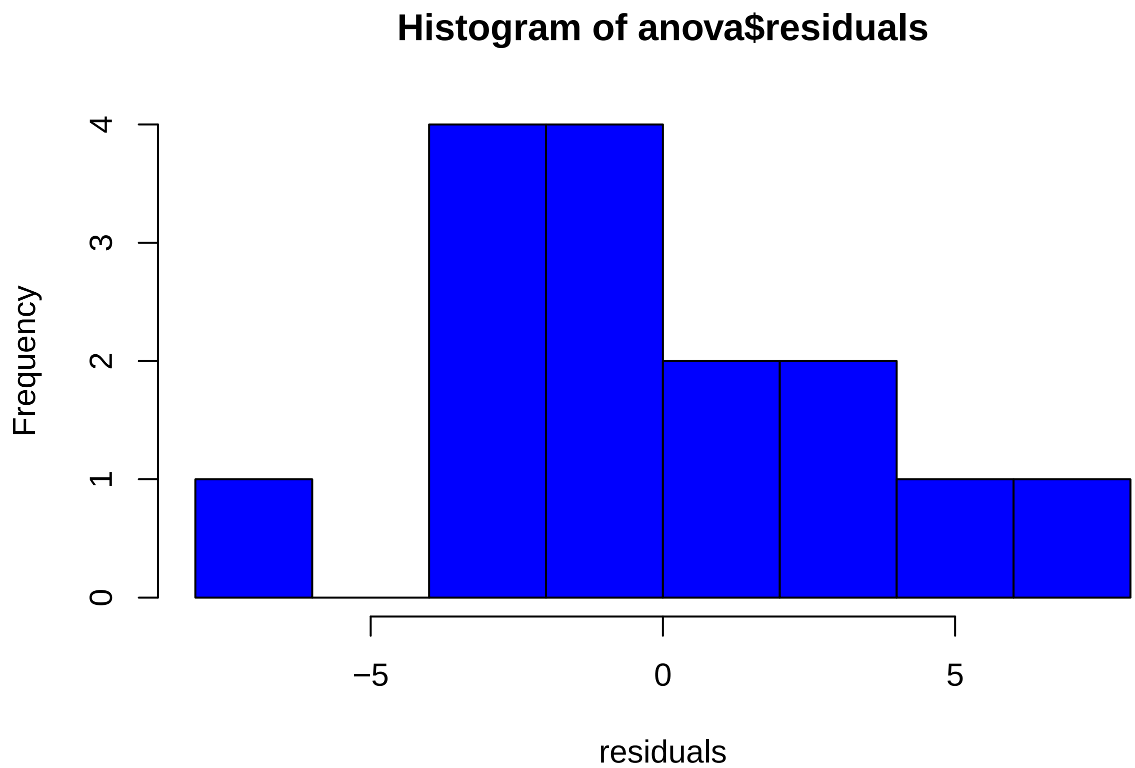
5 Normality assumption

5.1 Normality of the residuals

```
plot(anova,which=2) # gives normal QQ plot
```




```
hist(anova$residuals, xlab = "residuals", col = "blue") # normal probability plot
```



Interpretation

Hypotheses: $H_o : X_i's$ comes from a normal distribution. $H_a : X_i's$ do not come from a normal distribution

The residuals plotted in the normal probability plot most of it fall along a straight line, which suggests that the normality assumption is valid. The tails, particularly the upper tail, show departures from the fitted line.Hence, the normal plot of residuals for this model indicates a fairly normal distribution of the residuals as it is quite close to a straight line.

In addition, the histogram plot of residual vs frequency suggests a nearly symmetrical bell-shape pattern, which is consistent with a sample from a normal distribution. But, visually there is one extreme outlier (with a value less than -5).

5.2 Shapiro-wilk Test

```
shapiro.test(anova$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova$residuals
## W = 0.97638, p-value = 0.9387
```

Interpretation

The Shapiro-Wilk test is used as a method in detecting non-normality of this problem since the sample size is less than 50. Utilizing $\alpha = 0.05$.

Hypotheses: H_0: X_i's come from a normal distribution. H_a: X_i's do not come from a normal distribution.

Decision Rule: Reject the null hypothesis if $W_c < W_{a(N)} = 0.2543$. Otherwise, do not reject H_o.

Decision: Since $W_c = 0.97638$ not less than $W_{a(N)} = 0.2543$, then H_o is not rejected. Moreover, we can utilize the p-value $= 0.9387 > \alpha = 0.05$, which also shows that H_o is not rejected.

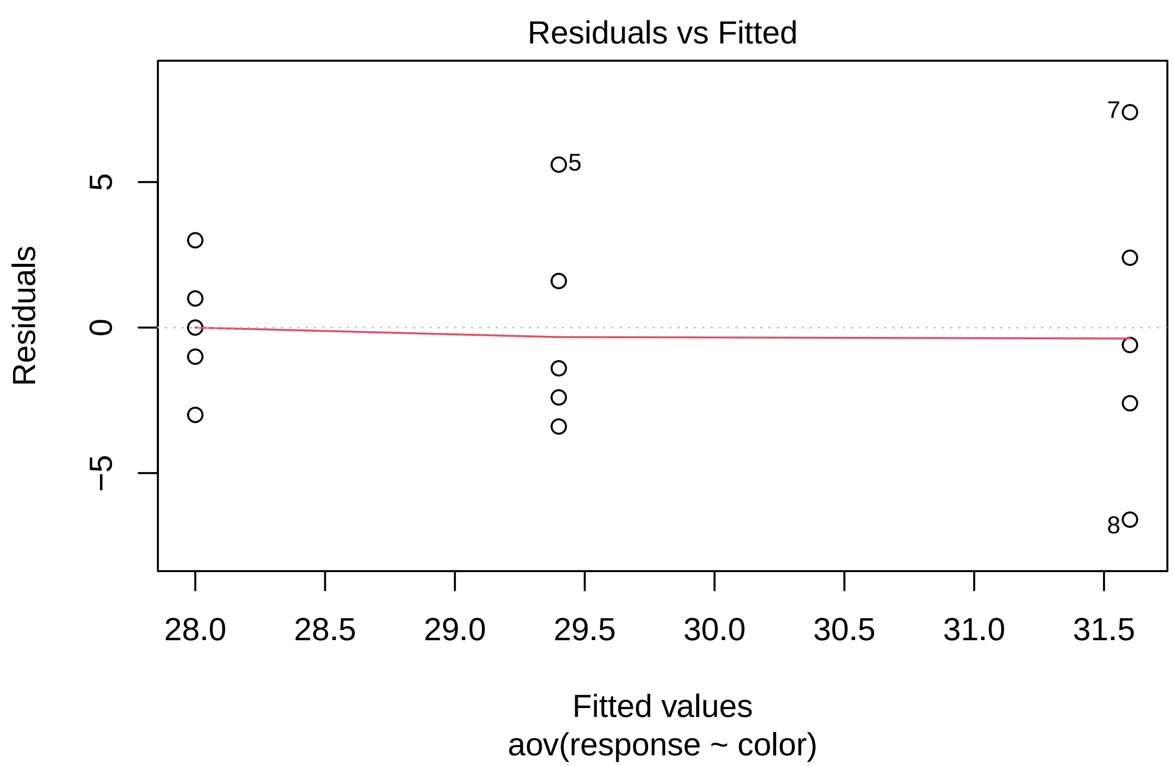
Conclusion: This implies that the distribution of the data are not significantly different from normal distribution with respect with its residuals. The large p-value indicates the data set is normally distributed and assume that the X_i's come from a normal distribution.

6 Equal Variances

Homogeneity of variances

6.1 Residual vs fitted plot

```
plot(anova,which=1) # give residual vs fitted values
```



Interpretation

The residuals-versus-predicted figure seems funnel-shaped. This demonstrates that the variance of the initial observations is not constant. The residuals depicted in the normal probability plot do not form a straight line, implying that the normality assumption is invalid. A data transformation is recommended, such as $\sqrt{Y}$ correction or log Y correction.

6.2 Bartletts test

```
bartlett.test(response~color)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  response by color
## Bartlett's K-squared = 2.4277, df = 2, p-value = 0.297
```

Interpretation

Bartlett's test (Snedecor and Cochran, 1983) is used to test if k samples have equal variances. Set $\alpha = 0.05$.

Hypothesis:

In words

H_o : The variances are homogeneous. vs. H_a : The variances are heterogeneous.

In symbols

$$H_o : \sigma_1^2 = \sigma_2^2 = \sigma_a^2 \text{ vs } H_a : \sigma_i^2 \neq \sigma_j^2, i \neq j$$

Decision Rule: Reject H_o if $X_o^2 > X_{a,a-1}^2$ where $X_{a,a-1}^2$ is the upper α percentage point of the chi-square distribution with α -1 degrees of freedom. Otherwise, do not reject H_o.

Decision: Since $X_o^2 = 2.4277 < X_{a,a-1}^2 = 5.991$, then H_o is not rejected. In addition to that, the $p-value = 0.297 > \alpha = 0.05$. This also implies that we failed to reject the null hypothesis.

Conclusion: Therefore, we do not have enough evidence to reject the null hypothesis. Hence, the data suggests implies the homogeneity of variances.

6.3 Levene's test

```
leveneTest(anova)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2   0.943 0.4165
##       12
```

Interpretation

Levene's test is consists of conducting a standard ANOVA F statistic equality of means. Using $\alpha = 0.05$.

Hypotheses: H_o: The three variances are homogeneous. H_a: The three variances are heterogeneous.

Decision Rule: Reject the null hypothesis if $F_c > F_t = 3.89$. Otherwise, do not reject H_o.

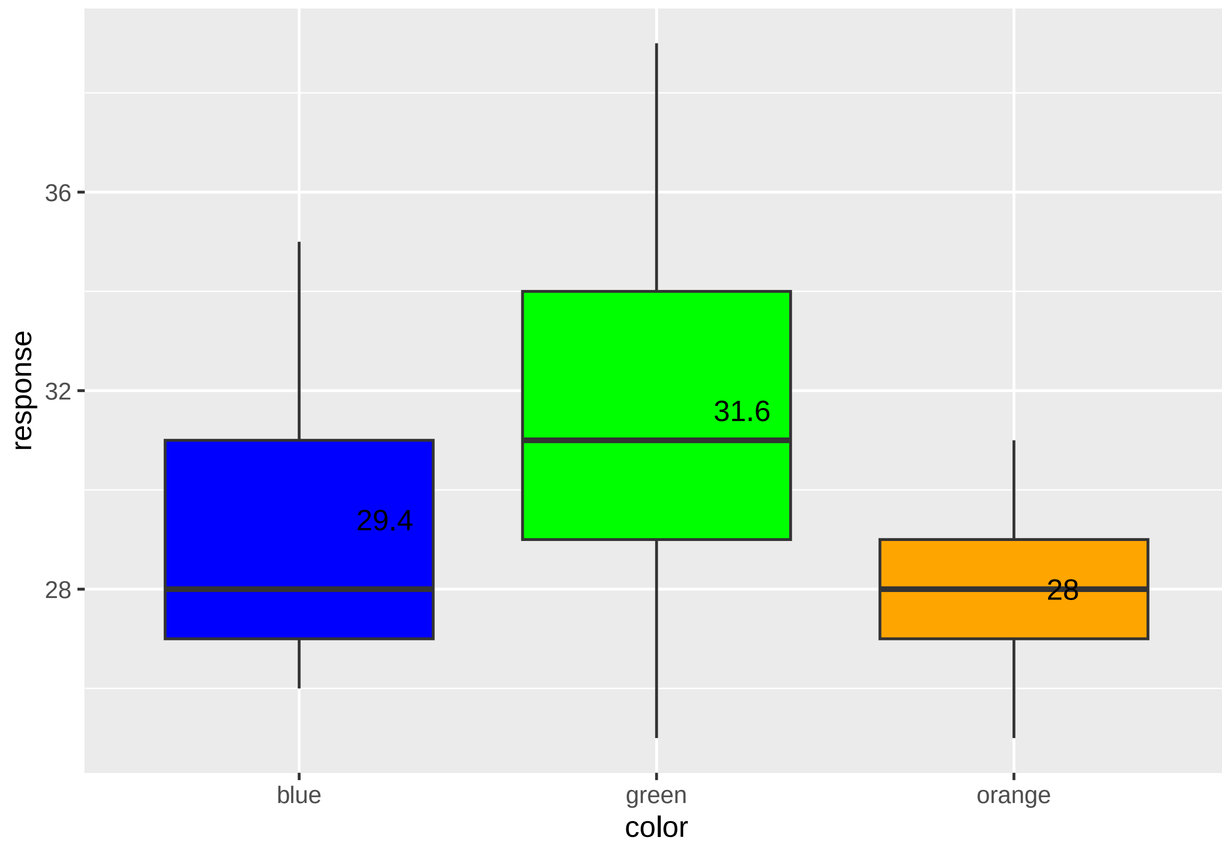
Decision: Since $F_c = 0.943$ not greater than $F_t = 3.89$, then H_o is not rejected. Moreover, we can utilize the $p - value = 0.4165 > \alpha = 0.05$, which also shows that Ho is not rejected.

Conclusion: Thus, all three variances are homogeneous,then the variances are not significantly different from each other. This is the same conclusion reached by analyzing the plot of residual versus fitted values. The homogeneity of assumption of the variance is met.

7 Outliers

```
ggplot(ques, aes(x = color, y = response, fill = color)) +
  geom_boxplot(show.legend = FALSE, fill = c("blue", "green", "orange")) +
  stat_summary(aes(label = round(after_stat(y), 1)),
               geom = "text",hjust = -1)
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
outlierTest(anova)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 7 2.548283           0.027078      0.40617
```

Interpretation

According to the results stating "No Studentized residuals with Bonferonni p-value less than 0.05", there were no outliers found. Since observations are considered outliers if their Bonferroni p is less than 0.05.

The largest studentized residual identified is located in row 7 and has been listed. In this case, 0.40617 is greater than 0.05, so, with Bonferroni correction, it might not be considered a significant outlier.

8 Question: Are the results of the analysis in exercise 2 valid? Why?

Through diagnostic analysis, we select multiple statistical models that best fit the data and are also suitable statistical models. Residual measures are also applied to satisfy the assumptions of the chosen statistical model and then investigate whether or not it is possible to alter the statistical model in some way (through the transformation of the variables or adding and eliminating variables) to make the model fit the data better.

The assumptions of normality, homogeneity of variances, and independence are met. Hence, the analysis results were valid and reliable since no assumption was violated. However, in practice, these assumptions will usually not hold. Thus, it is unwise to rely on the ANOVA until the validity of these assumptions has been checked. Moreover, if one or more of the assumptions are violated, then ANOVA may not be the most powerful test available, and the results of the analysis may be incorrect or misleading. So conducting the other or alternative test is necessary.

Therefore, in testing the assumptions of the ANOVA test results of the Questionnaire Color problem, the following assumptions are met: response variable residuals are normally distributed (or approximately normally distributed), variances of populations are equal, and no outliers are evident, then the results of the analysis in Exercise 2 are valid and reliable.