Doing Tests of Assumptions

Aishat Alli

2023-04-23

Q1: This file describes a study in which designers used Adobe Illustrator or Adobe InDesign to create a benchmark set of classic children's illustrations. The amount of time they took was recorded, in minutes. How many subjects took part in this study?

```
designtime = read.csv("designtime.csv")
```

```
summary(designtime)
```

Ans: 60 subjects took part in the study

```
##     Subject          Tool                Time       
##  Min.   : 1.00   Length:60          Min.   : 98.19  
##  1st Qu.:15.75   Class :character   1st Qu.:149.34  
##  Median :30.50   Mode  :character   Median :205.54  
##  Mean   :30.50                      Mean   :275.41  
##  3rd Qu.:45.25                      3rd Qu.:361.99  
##  Max.   :60.00                      Max.   :926.15  
```

Q2: Create a boxplot of the tasktime data for each tool. At a glance, which of the following conclusions seems to be the most likely?

```
install.packages("plyr")
```

Ans: Illustrator has a higher median task time than InDesign, with dissimilar variances

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(plyr)
ddply(designtime, ~ Tool, function(data) summary(data$Time))
```

```
##          Tool     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 1 Illustrator 169.8026 258.1377 363.7149 394.4744 451.3768 926.1523
## 2    InDesign  98.1917 128.3364 148.7572 156.3502 176.8919 230.4602
```

```
ddply(designtime, ~ Tool, summarise, Time.mean=mean(Time), Time.sd=sd(Time))
```

```
##          Tool Time.mean   Time.sd
## 1 Illustrator  394.4744 171.84720
## 2    InDesign  156.3502  33.40872
```

Q3: Conduct a Shapiro-Wilk test on Time response for each of the tools; a common test of normality. To the nearest ten-thousandth (four digits), what is the p - value of this test for illustrator?

```
shapiro.test(designtime[designtime$Tool == "Illustrator",]$Time)
```

Ans: Illustrator p - value = 0.01129,With the p-value being lesser than 0.05 , we reject the null hypothesis. Therefore, we can say that the time spent using Illustrator deviate from normality. For, InDesign p - value = 0.2553 , we can say that the time spent using Indesign follows a normal distribution, thereby accepting the null hypothesis

```
##
##  Shapiro-Wilk normality test
##
## data:  designtime[designtime$Tool == "Illustrator", ]$Time
## W = 0.90521, p-value = 0.01129
```

```
shapiro.test(designtime[designtime$Tool == "InDesign",]$Time)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  designtime[designtime$Tool == "InDesign", ]$Time
## W = 0.95675, p-value = 0.2553
```

Q4: Conduct a Shapiro- Wilk normality test on the residuals of Time by Tool. To the nearest ten-thousandth (four digits), what is the W value displayed? Hint: Use aov to fit a model and then run shapiro.test on the model residuals.

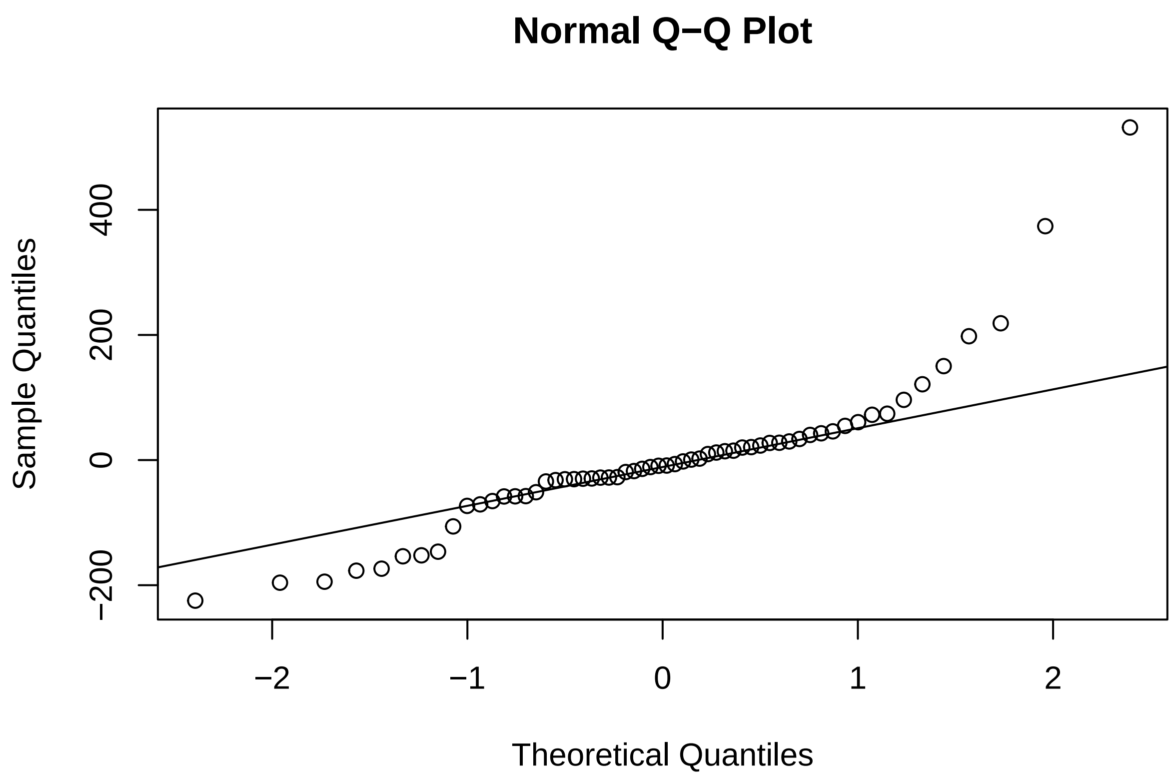
```
m = aov(Time ~ Tool, data=designtime) # fit model
```

```
shapiro.test(residuals(m)) # test residuals
```

Ans: W value = 0.85077 In light of the normality tests, I would conclude that the residuals of time do violate normality since the p value is significant thereby rejecting the null hypothesis. Also, from the QQ plot, there is a pretty big deviation from this line.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m)
## W = 0.85077, p-value = 3.211e-06
```

```
qqnorm(residuals(m)); qqline(residuals(m)) # plot residuals
```



Q5: Conduct a Brown-Forsythe test of homoscedasticity to test for homogeneity of variance. To the nearest hundredth two digits, what is the F statistic for the test? Hint: Use the car library and its levene test function with center = median

```
install.packages("car")
```

Ans: The f statistics is 20.08 and p-value = 3.545e-05, we can see that we have a violation of homogeneity of variance with the p -value being statistically significant rejecting the null hypothesis.

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(Time ~ Tool, data=designtime, center=median) # Brown-Forsythe test
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value    Pr(>F)
## group  1  20.082 3.545e-05 ***
##       58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Due to the difference in variance, I'm going to use a Welch t test indicating that the variance is not equal. And by doing that we don't have to conform to the assupmtion of homogeneity

of variance.

```
t.test(Time ~ Tool, data = designtime, var.equal = FALSE)
```

We can see that in fact we see a statistical difference in design time by tools there by rejecting the null hypothesis and accepting the alternate hypothesis.

```
##
##  Welch Two Sample t-test
##
## data:  Time by Tool
## t = 7.4502, df = 31.189, p-value = 2.077e-08
## alternative hypothesis: true difference in means between group Illustrator and group InDesign is not
## 95 percent confidence interval:
##  172.9527 303.2956
## sample estimates:
## mean in group Illustrator    mean in group InDesign
##                  394.4744                  156.3502
```

Q6: Fit a lognormal distribution to the Time response of each design tools. Conduct a Kolmogorov-Smirnov goodness-of-fit test. To the nearest ten-thousandth (four digits), what is the exact p - value of the test for Illustrator data? Hint: Use the MASS library and its fitdistr function with "lognormal" to acquire a fit estimate. Then use ks.test with "plnorm" passing the acquired fit values as meanlog and sdlog. Request an exact fit.

```
install.packages("MASS")
```

Ans: The exact p value for Illustrator data is 0.9344 and the exact p value for InDesign data is 0.8958, the p-value is not significant, meaning that there is no detectable significant departure from log normality. Also ,since it's a time measurement, task time can fall log normally.

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(MASS)
```

```
fit = fitdistr(designtime[designtime$Tool == "Illustrator",]$Time, "lognormal")$estimate
ks.test(designtime[designtime$Tool == "Illustrator",]$Time, "plnorm", meanlog=fit[1], sdlog=fit[2], exa
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  designtime[designtime$Tool == "Illustrator", ]$Time
## D = 0.093358, p-value = 0.9344
## alternative hypothesis: two-sided
```

```
fit = fitdistr(designtime[designtime$Tool == "InDesign",]$Time, "lognormal")$estimate
ks.test(designtime[designtime$Tool == "InDesign",]$Time, "plnorm", meanlog=fit[1], sdlog=fit[2], exact=
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  designtime[designtime$Tool == "InDesign", ]$Time
## D = 0.10005, p-value = 0.8958
## alternative hypothesis: two-sided
```

Q7: Create a new column that is the log-transformed Time response. Compute the mean of this log-transformed response for each drawing tool. To the nearest hundredth (two digits), what is the mean of the log-transformed response for InDesign?

```
designtime$logTime = log(designtime$Time) # log transform
```

```
library(plyr)
ddply(designtime, ~ Tool, function(data) summary(data$logTime))
```

Ans: The mean of the log-transformed response for InDesign is 5.03

```
##          Tool     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 1 Illustrator 5.134637 5.551303 5.896325 5.894288 6.112203 6.831039
## 2    InDesign 4.586922 4.854655 5.002285 5.030470 5.175538 5.440078
```

```
ddply(designtime, ~ Tool, summarise, logTime.mean=mean(logTime))
```

```
##          Tool logTime.mean
## 1 Illustrator     5.894288
## 2    InDesign     5.030470
```

```
shapiro.test(designtime[designtime$Tool == "Illustrator",]$logTime)
```

After transforming the time data by taking the log of time so it complies with the assumptions of ANOVA, we can now test for the Shapiro-Wilk test. We can see what was significant on time is no longer significant on log time, indicating we don't significantly deviate from a normal distribution. Hence rejecting the alternate hypothesis.

```
##
##  Shapiro-Wilk normality test
##
## data:  designtime[designtime$Tool == "Illustrator", ]$logTime
## W = 0.9829, p-value = 0.8962
```

```
shapiro.test(designtime[designtime$Tool == "InDesign",]$logTime)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  designtime[designtime$Tool == "InDesign", ]$logTime
## W = 0.97617, p-value = 0.7172
```

Q8 Conduct an independent-samples t-test on the log-transformed Time response. Use the Welch version for unequal variances. To the nearest hundredth(two digits), what is the t statistic for the test?

```
t.test(logTime ~ Tool, data=designtime, var.equal=TRUE)
```

Ans: t statistic for the test is 10.23 Here we have a significant P value, but now, one that we can trust a bit more because we have transformed our data.

```
##
##  Two Sample t-test
##
## data:  logTime by Tool
```

```
## t = 10.23, df = 58, p-value = 1.327e-14
## alternative hypothesis: true difference in means between group Illustrator and group InDesign is not
## 95 percent confidence interval:
##  0.6947957 1.0328408
## sample estimates:
## mean in group Illustrator    mean in group InDesign
##                  5.894288                  5.030470
```