BT1101

Lab 4: Statistical Measures, Probability Distributions and Data Modelling

Installing and loading packages

```r
# install required packages if you have not (suggested packages: rcompanion, rstatix, Rmisc, dplyr, tidyr, rpivotTable, knit
r, psych)
# install.packages("dplyr") #only need to run this code once to install the package
# load required packages
# library("xxxx")

library("rcompanion") #this package is required for transformTukey function
library("rstatix")
```

```r
library("Rmisc")
```

```r
library("dplyr") #need to call the libary before you use the package
```

```r
library("tidyr")
library("rpivotTable")
library("knitr")
library("psych")
```

- It is important to be able to code and produce your Rmarkdown output file independently

Part 1

Tutorial 4 Part 1 (For lab session)

- Dataset required: `Sales Transactions.xlsx`

`Sales Transactions.xlsx` contains the records of all sale transactions for a day, July 14. Each of the column is defined as follows:

- `CustID` : Unique identifier for a customer
- `Region` : Region of customer's home address
- `Payment` : Mode of payment used for the sales transaction
- `Transction Code` : Numerical code for the sales transaction
- `Source` : Source of the sales (whether it is through the Web or email)
- `Amount` : Sales amount
- `Product` : Product bought by customer
- `Time Of Day` : Time in which the sale transaction took place.

In the last tutorial, you were tasked to help the store manager develop dashboards that will enable him to gain better insights of the data.

In this tutorial, you will use the data to conduct sampling estimation and hypotheses testing. Where necessary, check the distribution for the variables and for the presence of outliers.

If the answer is greater than 1, round off to 2 decimal places. If the answer is less than 1, round off to 3 significant numbers. When rounding, also take note of the natural rounding points, for example, costs in dollars would round off to 2 decimal places.

Loading datasets into R

```r
#put in your working directory folder pathname ()

#import excel file into RStudio
library(readxl)
setwd("C:/nbox/Soc Acad Courses/AY2022 BT1101/Data")
#import xlsx file into RStudio
ST <- read_excel("Sales Transactions.xlsx", col_types = c("numeric", "text", "text", "numeric", "text", "numeric", "text",
"date"), skip = 2)
head(ST)
```

```
## # A tibble: 6 × 8
##   `Cust ID` Region Payment `Transaction Code` Source Amount Product
##       <dbl> <chr>  <chr>                <dbl> <chr>   <dbl> <chr>
## 1     10001 East   Paypal            93816545 Web      20.2 DVD
## 2     10002 West   Credit            74083490 Web      17.8 DVD
## 3     10003 North  Credit            64942368 Web      24.0 DVD
## 4     10004 West   Paypal            70560957 Email    23.5 Book
## 5     10005 South  Credit            35208817 Web      15.3 Book
## 6     10006 West   Paypal            20978903 Email    17.3 DVD
## # … with 1 more variable: `Time Of Day` <dttm>
```

Coding Practice

Q1.(a) Computing Interval Estimates

Using the sale transaction data on July 14,

- i. compute the 95% and 99% confidence intervals for the mean of `Amount` for DVD sale transactions. Which interval is wider and how does a wider interval affect type 1 error?
- ii. compute the 90% confidence interval for proportion of DVD sale transactions with sales amount being greater than $22. Could the company reasonably conclude that the true proportion of DVD sale transactions with sales amount greater than $22 is 30%?
- iii. compute the 95% prediction interval for `Amount` for sales of DVD. Explain to the store manager what this prediction interval mean?

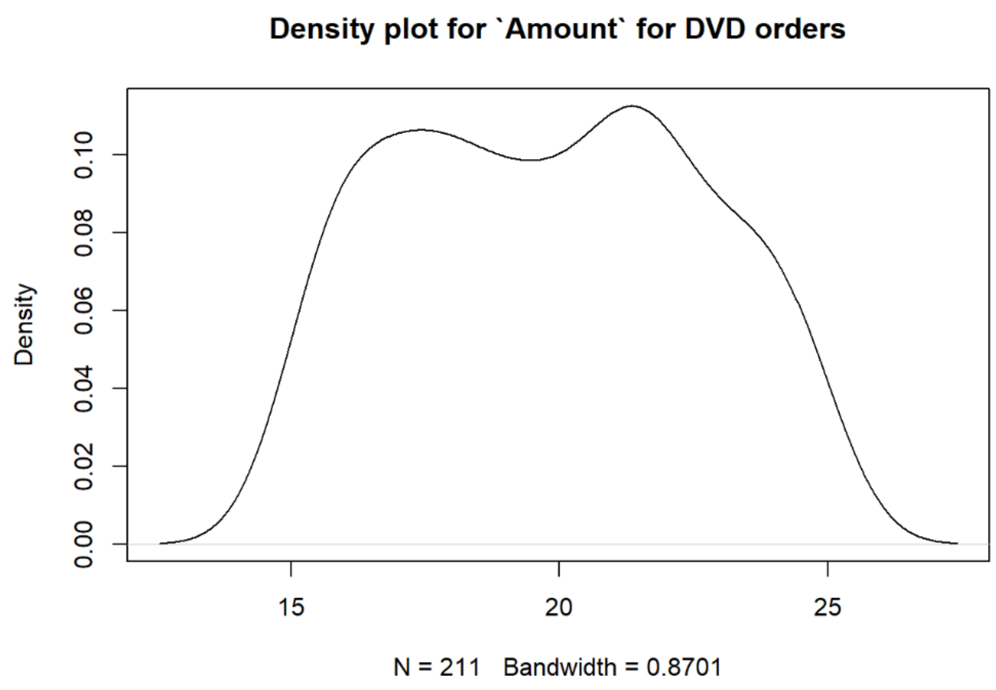
Tutorial Discussion:
What would you do to compute the interval estimates for Book Sales instead of DVD sales?

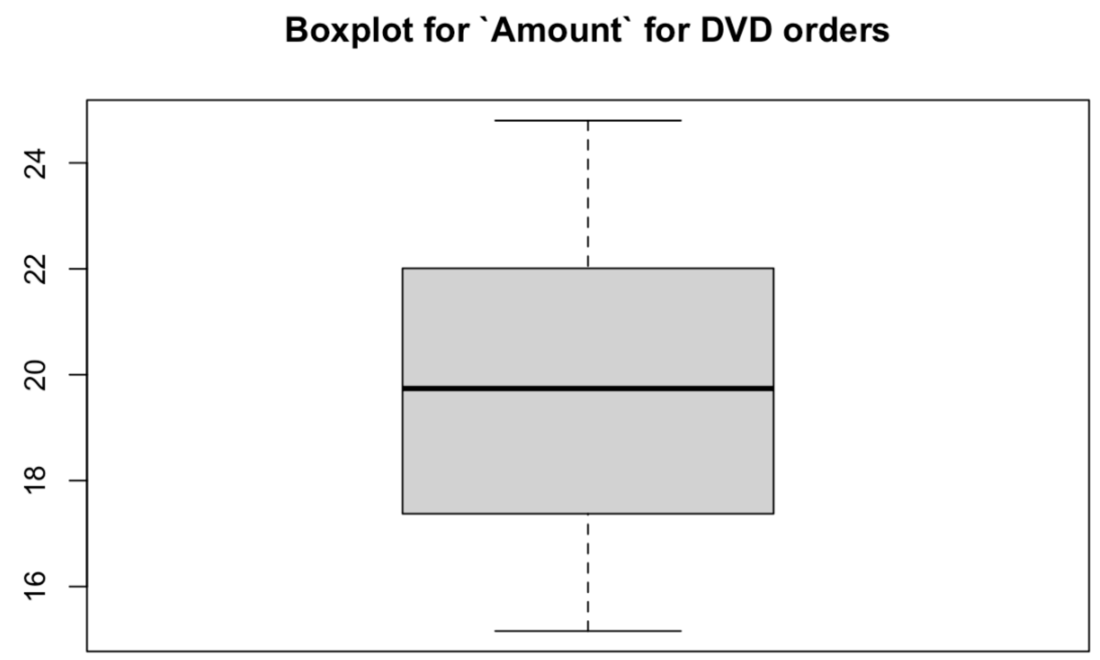
Recall the outlier analyses done on Amount for books in last tutorial.

Q1(a) - Customer Dashboard

(i) Compute the 95% and 99% confidence intervals for the mean of Amount for DVD sale transactions. Which interval is wider and how does a wider interval affect type 1 error?

```
dfD<-ST%>%filter(Product=="DVD")


# Previously outlier analyses has already been done for `Amount` data. So we can recap them here
plot(density(dfD$Amount),main="Density plot for `Amount` for DVD orders")
```






```
boxplot(dfD$Amount, main="Box plot for 'Amount' for DVD orders")
```

Q1(a) - Customer Dashboard

(i) Compute the 95% and 99% confidence intervals for the mean of Amount for DVD sale transactions. Which interval is wider and how does a wider interval affect type 1 error?

Formula for confidence intervals with unknown population standard deviation:

$$\bar{x} \pm t_{\alpha/2,n-1}(s/\sqrt{n})$$

```
# i)
# compute manually 95% CI for mean DVD `Amount`
uCIamt95<- mean(dfD$Amount) - qt(0.025,df=nrow(dfD)-1)*sd(dfD$Amount)/sqrt(nrow(dfD))
lCIamt95 <- mean(dfD$Amount) + qt(0.025,df=nrow(dfD)-1)*sd(dfD$Amount)/sqrt(nrow(dfD))
print(cbind(lCIamt95, uCIamt95), digits=4)
```

```
##      lCIamt95 uCIamt95
## [1,]    19.44     20.2
```

Q1(a) - Customer Dashboard

(i) Compute the 95% and 99% confidence intervals for the mean of Amount for DVD sale transactions. Which interval is wider and how does a wider interval affect type 1 error?

```
#compute manually 99% CI for mean DVD `Amount`
uCIamt99<- mean(dfD$Amount) - qt(0.005,df=nrow(dfD)-1)*sd(dfD$Amount)/sqrt(nrow(dfD))
lCIamt99 <- mean(dfD$Amount) + qt(0.005,df=nrow(dfD)-1)*sd(dfD$Amount)/sqrt(nrow(dfD))
print(cbind(lCIamt99, uCIamt99), digits=4)
```

```
##      lCIamt99 uCIamt99
## [1,]    19.32    20.33
```

- The 99% interval is wider —> which should make sense intuitively, since we are more confident that the true mean of **Amount** falls within this range!
- Type I error = α —> the probability of incorrectly rejecting when the null hypothesis is true. 99% CI has a lower Type I error.

Q1(a) - Customer Dashboard

(ii) Compute the 90% confidence interval for proportion of DVD sale transactions with sales amount being greater than $22. Could the company reasonably conclude that the true proportion of DVD sale transactions with sales amount greater than $22 is 30%?

```
# ii) compute 90% CI for proportion DVD (Amount>22)

d22<- dfD %>% filter(Amount>22)
pd22<-nrow(d22)/nrow(dfD)
lCIpd22 <- pd22 + (qnorm(0.05)*sqrt(pd22*(1-pd22)/nrow(dfD)))
uCIpd22 <- pd22 - (qnorm(0.05)*sqrt(pd22*(1-pd22)/nrow(dfD)))
print(cbind(lCIpd22, uCIpd22),digits=3)
```

```
##      lCIpd22 uCIpd22
## [1,]   0.202     0.3
```

- Yes, as the proportion of 0.3 falls within the 90% confidence interval.

Q1(a) - Customer Dashboard

(iii) Compute the 95% prediction interval for Amount for sales of DVD. Explain to the store manager what this prediction interval means.

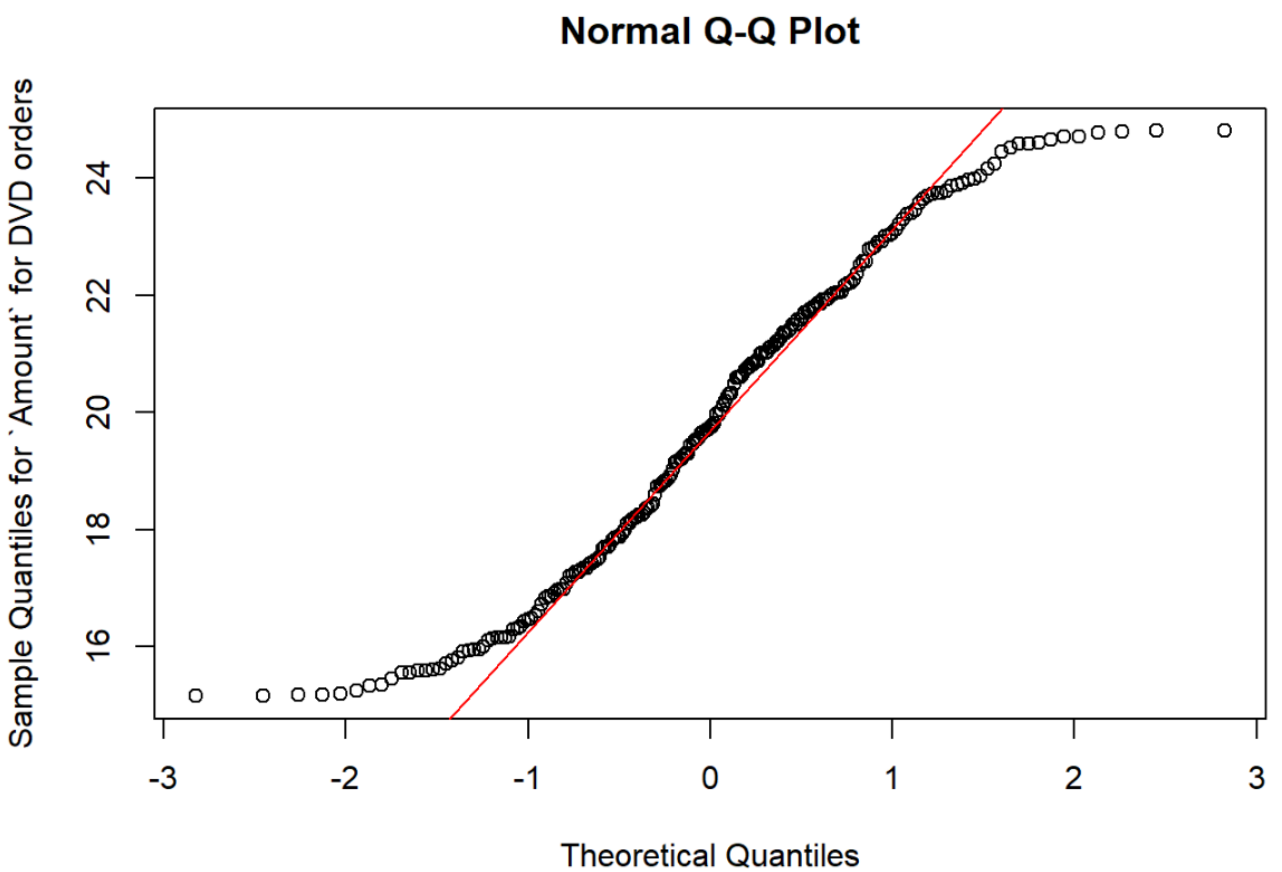
In order for prediction intervals to be valid, our variable should be normally distributed. Hence, you **must** check for normality before proceeding to compute prediction intervals.

```
# iii) compute 95% prediction interval for `Amount` for DVD
# check normality
qqnorm(dfD$Amount,
       ylab="Sample Quantiles for `Amount` for DVD orders")
qqline(dfD$Amount,col="red")
```

```
shapiro.test(dfD$Amount)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dfD$Amount
## W = 0.95635, p-value = 4.703e-06
```

Is this data (sufficiently) normally distributed?



Q1(a) - Customer Dashboard

(iii) Compute the 95% prediction interval for `Amount` for sales of DVD. Explain to the store manager what this prediction interval means.

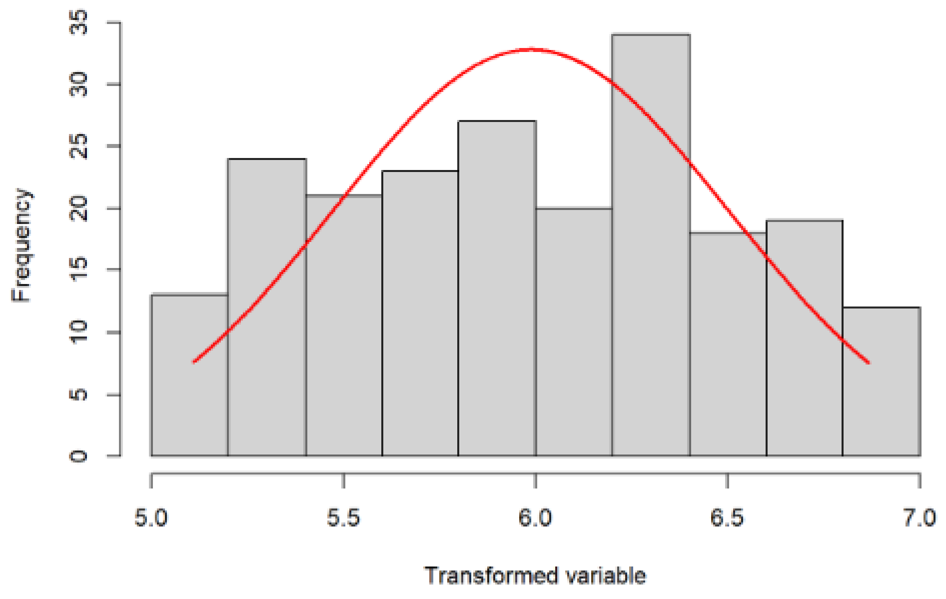
Transformations such as the **Tukey ladder of powers transformation** can reshape distributions to more closely approximate a normal distribution.

```
#transform data to normal distribution using transformTukey
dfD$Amt.t = transformTukey(dfD$Amount, plotit=TRUE)
```

```
##
##     lambda      W Shapiro.p.value
## 425    0.6 0.9566       4.959e-06
##
## if (lambda >  0){TRANS = x ^ lambda}
## if (lambda == 0){TRANS = log(x)}
## if (lambda <  0){TRANS = -1 * x ^ lambda}
```

The output tells you which transformation was applied to your variable, based on the calculated lambda value.

Compare the new value of W (test statistic for the Shapiro-Wilk test) to the original value on the previous slide. What do you notice?



Q1(a) - Customer Dashboard

(iii) Compute the 95% prediction interval for `Amount` for sales of DVD. Explain to the store manager what this prediction interval means.

```
#using x ^ lambda where lambda = 0.6
mnamt.t <- mean(dfD$Amt.t)
sdamt.t <- sd(dfD$Amt.t)
lPI.amtt <- mnamt.t + (qt(0.025, df = (nrow(dfD)-1))*sdamt.t*sqrt(1+1/nrow(dfD)))
uPI.amtt <- mnamt.t - (qt(0.025, df = (nrow(dfD)-1))*sdamt.t*sqrt(1+1/nrow(dfD)))
cbind(lPI.amtt, uPI.amtt)
```

Note that output of the Turkey transformation was stored in **dfD$Amt.t** (i.e., we are working with a transformed variable).

```
##      lPI.amtt uPI.amtt
## [1,] 4.972433 7.001642
```

```
#reverse transform; comments below is to derive the formula
# y= x^lamda
# y = x^0.6
# x = y^(1/0.6)
lPI.amt <- lPI.amtt^(1/0.6)
uPI.amt<- uPI.amtt^(1/0.6)

print(cbind(lPI.amt,uPI.amt),digits=4) # reverse transform
```

```
##      lPI.amt uPI.amt
## [1,]   14.49   25.63
```

However, we should report the 95% prediction interval for Amount to the store manager in its **original** units. The transformed variable does not provide meaningful information for business decisions.

Hence, we have to **reverse** the previously applied transformation.

Can you explain what the prediction interval means in simple English?

Coding Practice

Q1.(b) Hypothesis Testing

The store manager would like to draw some conclusions from the sample sales transaction data. He would like to retain all the data for the analyses. Please help him to set up and test the following hypotheses.

- i. The proportion of book sales transactions with `Amount` greater than $50 is at least 10 percent of book sales transactions.
- ii. The mean sales amount for books is the same as for dvds.
- iii. The mean sales amount for rare books is greater than mean sales amount for normal books. Rare books are books where `Amount` is greater than 100, while normal books are those where `Amount` is less than or equal to 100. (Hint: Create a new categorical variable to group the books into Rare vs Normal types.)
- iv. The mean sales amount for dvds is the same across all 4 regions.

Q1(b) - Hypothesis testing

(i) Hypothesis: The proportion of book sales transactions with Amount greater than \$50 is at least 10 percent of book sales transactions.

One-sample test for proportion —> hence we use the **z statistic.**

- H0: $proportion \geq 0.1$
- H1: $proportion < 0.1$

```
# i)
# compute z-statistic for proportion.
book<-ST %>% filter(Product=="Book")
bk50<- book %>% filter(Amount>50)
pbk50<-nrow(bk50)/nrow(book)
pbk50
```

```
## [1] 0.2030651
```

```
z <- (pbk50 - 0.10) / sqrt(0.1*(1-0.1)/nrow(book))
z
```

```
## [1] 5.550227
```

Q1(b) - Hypothesis testing

(i) Hypothesis: The proportion of book sales transactions with Amount greater than $50 is at least 10 percent of book sales transactions.

```
#compute critical value
cv95<-qnorm(0.05)
cv95
```

```
## [1] -1.644854
```

```
z<cv95
```

```
## [1] FALSE
```

Since the test statistics z is larger than the critical value (left tail), we fail to reject the null hypothesis. Therefore, the data shows evidence that "proportion of 50 book order amount greater than 50 is at least 10%"

Q1(b) - Hypothesis testing

(ii) Hypothesis: The mean sales amount for books is the same as for dvds.

Two-sample test for means with independent samples —> hence we use a **t-test.**

- H0: $\mu_{books} = \mu_{dvds}$
- H1: $\mu_{books} \neq \mu_{dvds}$

```
# ii)
t.test(Amount~Product, data=ST)
```

```
##
##  Welch Two Sample t-test
##
## data:  Amount by Product
## t = 8.0304, df = 260.96, p-value = 3.344e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  27.47079 45.31916
## sample estimates:
## mean in group Book  mean in group DVD
##           56.21559           19.82062
```

The t-statistic is 8.03. Since $p < 0.05$, we can conclude that there is a significant difference between the mean sales amount for books and DVDs.

(Note — report the test statistic and p-value when stating the results of a test.)

Q1(b) - Hypothesis testing

(ii) Hypothesis: The mean sales amount for books is the same as for dvds.

```
#try to run with WELCH ANOVA to compare results with t test)
wa.out.t <- ST %>% welch_anova_test(Amount~ Product)
ga.out.t <- games_howell_test(ST, Amount ~ Product)
wa.out.t
```

ANOVA vs WELCH ANOVA

```
## # A tibble: 1 × 7
##   .y.        n statistic   DFn   DFd        p method
## * <chr>  <int>     <dbl> <dbl> <dbl>    <dbl> <chr>
## 1 Amount   472      64.5     1  261. 3.34e-14 Welch ANOVA
```

```
ga.out.t
```

```
## # A tibble: 1 × 8
##   .y.    group1 group2 estimate conf.low conf.high    p.adj p.adj.signif
## * <chr>  <chr>  <chr>     <dbl>    <dbl>     <dbl>    <dbl> <chr>
## 1 Amount Book   DVD       -36.4    -45.3     -27.5 1.49e-13 ****
```

```
# try to run with ANOVA to compare with t test
aov.t <- aov(ST$Amount ~ ST$Product)
summary(aov.t)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## ST$Product    1  154548  154548   52.15 2.09e-12 ***
## Residuals   470 1392966    2964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA can identify a difference among the means of multiple populations

Q1(b) - Hypothesis testing

(iii) Hypothesis: The mean sales amount for rare books is greater than mean sales amount for normal books. Rare books are books where `Amount` > 100, while normal books are those where `Amount` <= 100. (Hint: Create a new categorical variable to group the books into Rare vs Normal types.)

Two-sample test for means with independent samples —> hence we use a **t-test.**

- H0: $\mu_{rare} \leq \mu_{normal}$
- H1: $\mu_{rare} > \mu_{normal}$

```
# iii)
# Create categorical variable for book type
book$bktype <- NA
book$bktype[book$Amount>100] <- "Rare"
book$bktype[book$Amount<=100] <- "Normal"
book$bktype <- as.factor(book$bktype)
levels(book$bktype)
```

```
## [1] "Normal" "Rare"
```

Q1(b) - Hypothesis testing

(iii) Hypothesis: The mean sales amount for rare books is greater than mean sales amount for normal books. Rare books are books where `Amount` > 100, while normal books are those where `Amount` <= 100. (Hint: Create a new categorical variable to group the books into Rare vs Normal types.)

Two-sample test for means with independent samples —> hence we use a **t-test.**

- H0: $\mu_{normal} \geq \mu_{rare}$
- H1: $\mu_{normal} < \mu_{rare}$

```
# compare amount
t.test(Amount~bktype, alternative="less", data=book)
```

```
##
##  Welch Two Sample t-test
##
## data:  Amount by bktype
## t = -40.548, df = 52.22, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -170.5442
## sample estimates:
## mean in group Normal   mean in group Rare
##             20.09216            197.98302
```

How do you know that **t.test** is comparing $\mu_{normal} < \mu_{rare}$ rather than the other way around?

The t-statistic is -40.55. Since p < 0.05, we can conclude that the mean sales amount for normal books is significantly less than the mean sales amount for rare books.

Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

More than two sample test for means —> hence we use an **ANOVA.**

- H0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H1: At least one μ is different from the others.

Assumptions of ANOVA

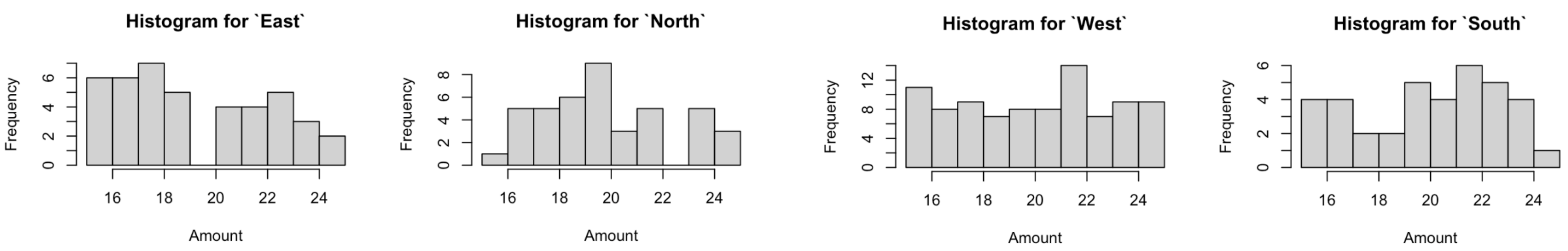
In order for the results of the ANOVA to be reliable, the groups in our data must typically satisfy three assumptions:

1. Data points are independent
2. Within each group, data is normally distributed
3. These distributions have equal variances.

Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

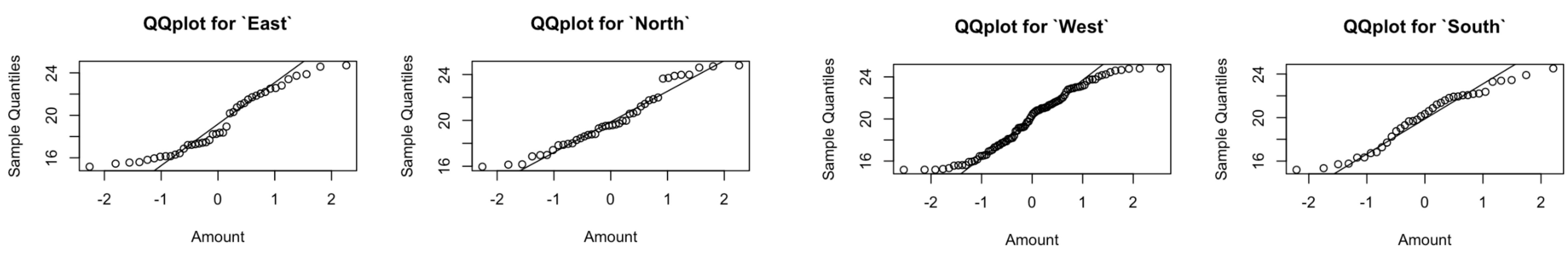
```
# iv) Check if ANOVA assumptions are met
# check normality
par(mfcol=c(2,2))
ST.dvd <- ST %>% filter(Product=="DVD")
E<-ST.dvd %>% filter(Region=="East")
W<-ST.dvd %>% filter(Region=="West")
N<-ST.dvd %>% filter(Region=="North")
S<-ST.dvd %>% filter(Region=="South")

# plot histogram
hist(E$Amount, main="Histogram for `East`", xlab="Amount")
hist(W$Amount, main="Histogram for `West`", xlab="Amount")
hist(N$Amount, main="Histogram for `North`", xlab="Amount")
hist(S$Amount, main="Histogram for `South`", xlab="Amount")
```



Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

```
# plot qqplots
par(mfcol=c(2,2))
qqnorm(E$Amount, main="QQplot for `East`", xlab="Amount")
qqline(E$Amount)
qqnorm(W$Amount, main="QQplot for `West`", xlab="Amount")
qqline(W$Amount)
qqnorm(N$Amount, main="QQplot for `North`", xlab="Amount")
qqline(N$Amount)
qqnorm(S$Amount, main="QQplot for `South`", xlab="Amount")
qqline(S$Amount)
```

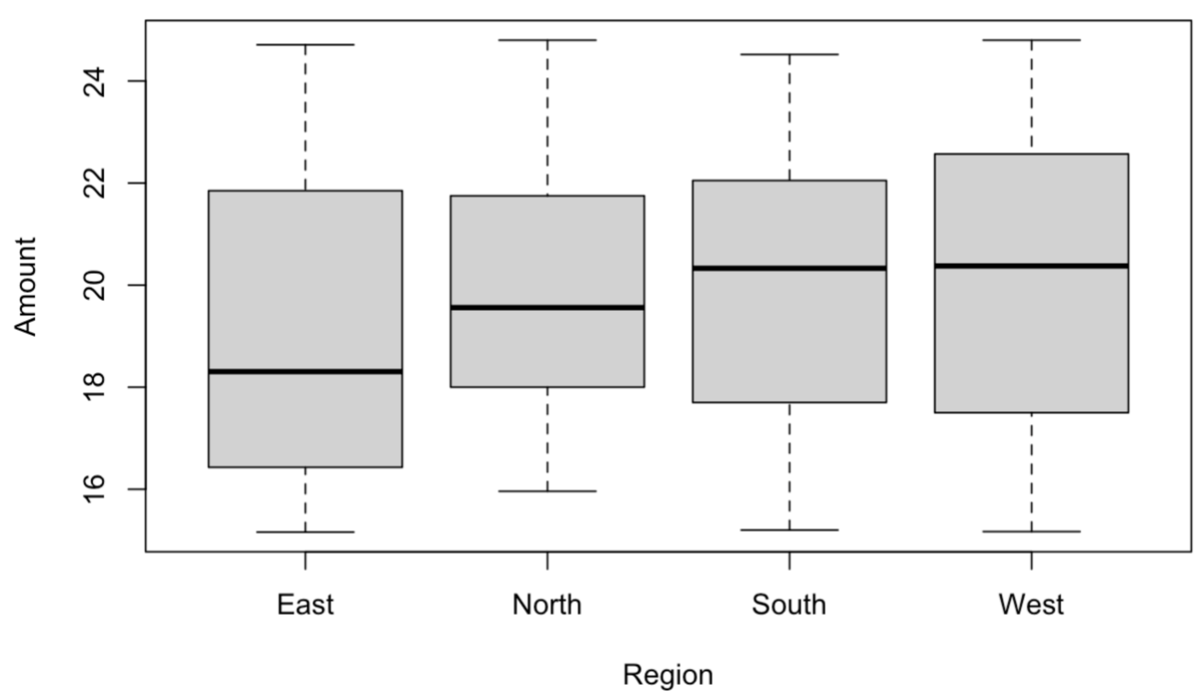


- Conduct a test of normality and at least one method of visual inspection when assessing whether data is normally distributed
- Based on these results, do you think the data is sufficiently normally distributed?

Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

```r
# plot boxplots
boxplot(ST.dvd$Amount ~ ST.dvd$Region)

# Shapiro Wilk Test
lapply(list(E,W,N,S),
       function(sa)
         {
         shapiro.test(sa$Amount)
         })
```




```
## [[1]]
##
##  Shapiro-Wilk normality test
##
## data:  sa$Amount
## W = 0.91567, p-value = 0.004389
##
##
## [[2]]
##
##  Shapiro-Wilk normality test
##
## data:  sa$Amount
## W = 0.94898, p-value = 0.001448
##
##
## [[3]]
##
##  Shapiro-Wilk normality test
##
## data:  sa$Amount
## W = 0.94309, p-value = 0.03669
##
##
## [[4]]
##
##  Shapiro-Wilk normality test
##
## data:  sa$Amount
## W = 0.94666, p-value = 0.07532
```

Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

```
# check sample sizes across regions
table(ST.dvd$Region)
```

```
##
##  East North South  West
##    42    42    37    90
```

```
# check equal variance assumption
fligner.test(Amount~ Region, ST.dvd)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Amount by Region
## Fligner-Killeen:med chi-squared = 3.584, df = 3, p-value = 0.31
```

If group sizes are similar, ANOVA is fairly robust to unequal variances.

However, that is not the case for our data — hence, we conduct a more than two-sample test for variances.

Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

```
# Conduct Anova, or directly perform Welch Anova
aov.amt<-aov(ST.dvd$Amount ~ as.factor(ST.dvd$Region)) #note the group variable should be a factor
summary(aov.amt)
```

```
##                           Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(ST.dvd$Region)   3   20.7   6.898   0.866   0.46
## Residuals                207 1648.8   7.965
```

```
TukeyHSD(aov.amt)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ST.dvd$Amount ~ as.factor(ST.dvd$Region))
##
## $`as.factor(ST.dvd$Region)`
##                     diff        lwr      upr     p adj
## North-East   0.8064285714 -0.7887125 2.401570 0.5579913
## South-East   0.7760617761 -0.8720884 2.424212 0.6151432
## West-East    0.7763650794 -0.5896321 2.142362 0.4562708
## South-North -0.0303667954 -1.6785170 1.617783 0.9999609
## West-North  -0.0300634921 -1.3960607 1.335934 0.9999333
## West-South   0.0003033033 -1.4272374 1.427844 1.0000000
```

Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

```
# Welch ANOVA
wa.out1 <- ST.dvd %>% welch_anova_test(Amount~ Region)

# games howell test does not assume normality and equal variances
gh.out1 <- games_howell_test(ST.dvd, Amount ~ Region)
wa.out1
```

```
## # A tibble: 1 × 7
##   .y.        n statistic   DFn   DFd     p method
## * <chr>  <int>     <dbl> <dbl> <dbl> <dbl> <chr>
## 1 Amount   211       0.8     3  94.6 0.495 Welch ANOVA
```

What do post-hoc tests such as the TukeyHSD and Games-Howell test indicate?

```
gh.out1
```

```
## # A tibble: 6 × 8
##   .y.    group1 group2  estimate conf.low conf.high p.adj p.adj.signif
## * <chr>  <chr>  <chr>      <dbl>    <dbl>     <dbl> <dbl> <chr>
## 1 Amount East   North   0.806      -0.773      2.39 0.541 ns
## 2 Amount East   South   0.776      -0.889      2.44 0.614 ns
## 3 Amount East   West    0.776      -0.668      2.22 0.497 ns
## 4 Amount North  South  -0.0304     -1.58       1.52 1     ns
## 5 Amount North  West   -0.0301     -1.34       1.28 1     ns
## 6 Amount South  West    0.000303   -1.42       1.42 1     ns
```

Q1(b) - Hypothesis testing

(iv) Hypothesis: The mean sales amount for dvds is the same across all 4 regions.

More than two sample test for means —> hence we use an **ANOVA.**

- H0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H1: At least one μ is different from the others.

Results

Since p>0.05, we do not reject the null hypothesis that the mean sales amount for DVDs is the same across all 4 regions.