National University of Singapore-APAC (nus)

Institution: National University of Singapore  
ID: e1032495

Start:

*#install.packages("psych")*

library("tidyverse") #pivot\_wider()

library("dplyr")

library("tidyr")

library("rpivotTable")

library("knitr")

library("psych") #kurtosi()/corr.test

library("rcompanion") *# transformTukey function*

library("rstatix") *# for Welch ANOVA & Games Howell test*

library("Rmisc") *# CI(*, ci=0.95)   
library("ggplot2")

library("TTR") *# SMA(, n=)*  
library("tsbox") #ts\_ts(df)  
library("car") #car::vif to check multicollinearity  
library("factoextra") #See Kmeans cluster  
library(lpSolve) #Linear optimisation

Outlier

*# Conduct Shapiro-Wilk Test*

shapiro.test(BD$Savings)

*# Since data is not normally distributed, we can use Chebyshev's Theorem to identify data that is within 3 sd of the mean. Then use Z-scores > +3 or < -3 for outlier identification*

Plot1 <- boxplot(BD$Savings, horizontal=TRUE) *#applying 1.5\*IQR to the left of Q1 or right of Q3*

Plot1 <- boxplot(BD$Savings, horizontal=TRUE, range=3 ) *#* *3 times the IQR from Q3*

plot1$out #extract the outliers

spread:

BD3 <- BD %>% group\_by(`Loan Purpose`,Gender) %>% tally()

BD3.spread<- BD3 %>% spread(key=`Loan Purpose`,value=n)

BD3.spread[is.na(BD3.spread)]<-0 *#convert NA to 0 value*

kable(BD3.spread, caption = "Contingency Table for Gender & Loan Purpose")

Pivot wider: Transform into wide for LM

mobile %>%

  select(code,continent,year,mobile\_subs) %>%

  pivot\_wider(names\_from=year, values\_from=mobile\_subs)

Missing Data:

Missing <- MC[!complete.cases(MC),]

# summary(MC)

x<-x[!is.na(x)]

Piechart:

*#Create table for employment type*DBp1 <- DB %>% count(employment\_type, sort = TRUE)  
kable(DBp1, caption = "Employment Type")

*#Create pie chart for employment type*

slice.employment <- DBp1$n  
employment.piepercent <- 100\*round(DBp1$n/sum(DBp1$n),2)  
label <- DBp1$employment\_type  
label <- paste(label, ",", sep ="")  
label <- paste(label, employment.piepercent)  
label <- paste(label, "%", sep = "")  
pie(slice.employment, labels=label, col=c("blue","cyan","dodgerblue","deepskyblue"), radius=1, main="Employment Type")

Line Graph

plot(covid\_total$date, covid\_total$new\_cases\_smoothed\_per\_million, xlab = 'Date', ylab = 'Value', main = 'Leading vs Lagging', type = 'l') + lines(covid\_total$date, covid\_total$new\_cases\_smoothed\_per\_million, col= 'red')

lines(covid\_total$date, covid\_total$search\_vol, col= 'blue')

legend('topleft', legend=c("new\_cases\_smoothed\_per\_million", "search\_vol"), col=c("red", "blue"), lty=1.2, cex=0.8)

ggplot

sub1 <- covid\_total %>% filter(date >= '2020-01-26' & date <= '2020-06-26')

ggplot(data = sub1, aes(x = date)) + geom\_line(aes(y= new\_cases\_smoothed\_per\_million, colour = "New cases smoothed per million")) + geom\_line(aes(y= search\_vol, colour = "Search volume")) + scale\_color\_manual(name = "Legend", values = c("New cases smoothed per million" = "blue", "Search volume" = "red")) + theme\_bw() + labs(y = "Index", title = "New cases and search volume")

Barplot:

LoanFreq<-BD%>%count(`Loan Purpose`)

Loanbar <- LoanFreq$n

*# Horizontal*

barplot(Loanbar,names.arg=LoanFreq$`Loan Purpose`,col="blue", main="Frequency of Loan Purpose", cex.names = 0.8, xlim=c(0,120), xlab="No. of Loans", horiz=TRUE, las=1)

# Base R: Stacked

matrix1 <- as.matrix(total[c(1,2), c(2:7)])

barplot(matrix1, names.arg = col\_names[2:7], col = c('cyan','violet'), main = 'Barplot of Campaign responses')

legend("right", cex=0.5, fill=c('cyan','violet'), c('Not accepted', 'Accepted'))

# ggplot2

plot2 <- ggplot(sub2, aes(x= Paid, y=Mean, fill=Paid)) +

geom\_bar(stat="identity", color="black", position=position\_dodge()) + geom\_errorbar(aes(ymin=Mean+SE, ymax=Mean-SE), width=.2, position=position\_dodge(.9))   
 + labs(title="Paid Status on Lifetime.Post.Consumers", x="Paid", y = "Lifetime.Post.Consumers")  
 + theme\_classic()   
 + scale\_fill\_manual(values=c('#1B9E77','#D95F02'))

Histogram:

h1 <- hist(mark$`Total Amount Spent`, main="Histogram of Total Expenditure",xlab="Customer Expenditure",ylab="No. of Customers", col=c("darkorange"),xlim=c(0,2600), labels=TRUE)

Extracting table from histogram:

total\_exp<-cut(mark$`Total Amount Spent`,h1$breaks)  
t1<-table(total\_exp)  
kable(t1, caption = "Frequency distribution by Expenditure")

Contingency Table / Replacing NA with 0:

BD3 <- BD %>% group\_by(`Loan Purpose`,Gender) %>% tally()

*#same as BD2 <- BD%>%count(`Loan Purpose`,Gender)*

BD3.spread<- BD3 %>% spread(key=`Loan Purpose`,value=n)

BD3.spread[is.na(BD3.spread)]<-0 *#convert NA to 0 value*

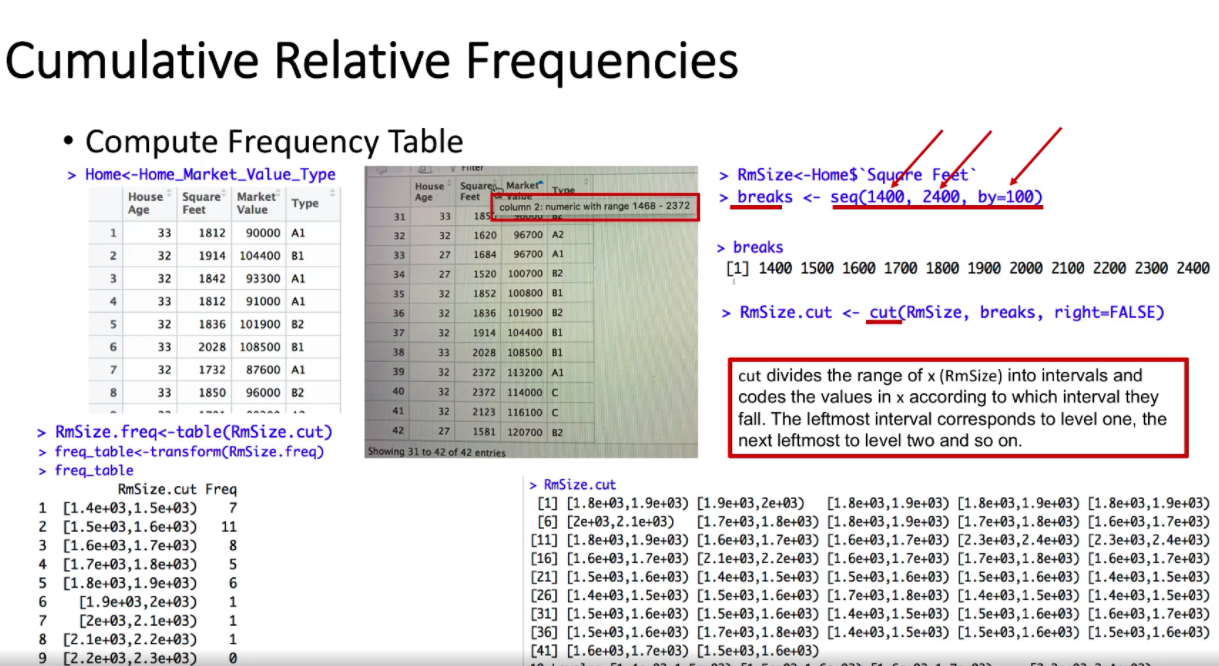
kable(BD3.spread, caption = "Contingency Table for Gender & Loan Purpose")

Rpivottable:

rpivotTable(mark, rows = c("Education","Marital Status"), cols="Complain", aggregatorName = "Count", height = “auto”)

Pareto:

mark3 <- mark %>% arrange(desc(`Total Amount Spent`))  
mark3$Percentage<-mark3$`Total Amount Spent`/sum(mark3$`Total Amount Spen`)  
mark3$Cumulative<-cumsum(mark3$Percentage)  
which(mark3$Cumulative>0.8)[1]  
# compute percentage ofcustomers with top 80% savings (86 is value)  
86/nrow(BD)



“Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables.

Yes. The correlation coefficient is appx 0.53, which we can conclude that there is a moderately strong positive linear relationship between tenure and satisfaction. In addition, the covariance coefficient is 2.07, which indicate a positive covariance and a direct relationship between tenure and satisfaction. This means that on average, as the tenure increases, the satsifaction is expected to increase as well.

Table of measures:

*# Use describe() function in psych package to generate the descriptive statistics*

dfage <- describe(BD$Age, IQR=TRUE)

dfsavings <- describe(BD$Savings, IQR=TRUE)

df.desc1<-rbind(dfage,dfsavings)

df.desc1$trimmed <- df.desc1$mad <- df.desc1$se <- NULL *# remove se, mad and trimmed if not needed*

df.desc1$vars[1]<-"Age"

df.desc1$vars[2]<-"Savings"

kable(df.desc1, row.names = FALSE, caption = "Descriptive Statistics for Age and Savings - using psych package")

Check if data is normally distributed

qqnorm(D$Demand.imp)

qqline(D$Demand.imp)

Confidence Interval for mean

*#library("Rmisc")*

*#compute 95% CI for mean Age using Rmisc::CI()*

ci.age<-CI(BD$Age, ci=0.95)

ci.age

*#To get the CI, you can print the lower and upper limit*

print(cbind(ci.age[3],ci.age[1]), digits=4)

Confidence Interval for proportion

n.bd=nrow(BD)

*#compute 95% CI for proportion (Age>50)*

age50<- BD %>% filter(Age>50)

p50=nrow(age50)/nrow(BD)

lCIp50 <- p50 + (qnorm(0.025)\*sqrt(p50\*(1-p50)/n.bd))

uCIp50 <- p50 - (qnorm(0.025)\*sqrt(p50\*(1-p50)/n.bd))

print(cbind(lCIp50, uCIp50),digits=3)

1. The 95% CI for mean age is [33.34, 35.45]. We are 95% “confident” that this interval contains the true population mean. More generally, with repeated random sampling from the same population (infinitely), 95% of such constructed confidence intervals would contain the true population mean age.

Prediction Interval (need normal distribution)

plot(density(BD$Age),main="Density plot for Age") *# distribution looks a little right skewed*

qqnorm(BD$Age,

ylab="Sample Quantiles for Age")

qqline(BD$Age,

col="red")

shapiro.test(BD$Age)

*#transform data to normal distribution using transformTukey*

BD$Age.t = transformTukey(BD$Age, plotit=TRUE)

*#using -1 \* x ^ lambda where lambda = -0.65*

mnage.t <- mean(gpa2$colgpa)

sdage.t <- sd(gpa2$colgpa)

uPI.aget <- mnage.t - (qt(0.005, df = (nrow(gpa2)-1))\*sdage.t\*sqrt(1+1/nrow(gpa2)))

lPI.aget <- mnage.t + (qt(0.005, df = (nrow(gpa2)-1))\*sdage.t\*sqrt(1+1/nrow(gpa2)))

cbind(lPI.aget, uPI.aget)

*# lPI.aget uPI.aget*

*# [1,] -0.1573027 -0.05346039*

*#reverse transform; comments below is to derive the formula*

*# y=-1\*x^lamda*

*# -y = x^-0.65 = 1/(x^0.65)*

*# x^0.65 = -1/y*

*# x = (-1/y)^(1/0.65)*

lPI.age2 <- (-1/lPI.aget)^(1/0.65)

uPI.age2<- (-1/uPI.aget)^(1/0.65)

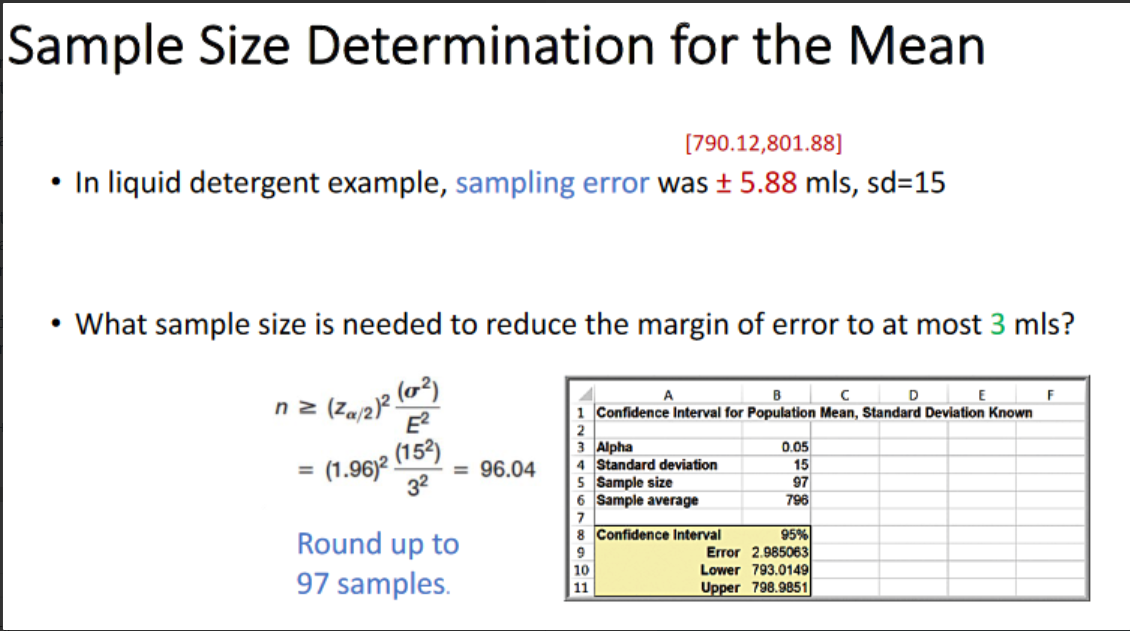
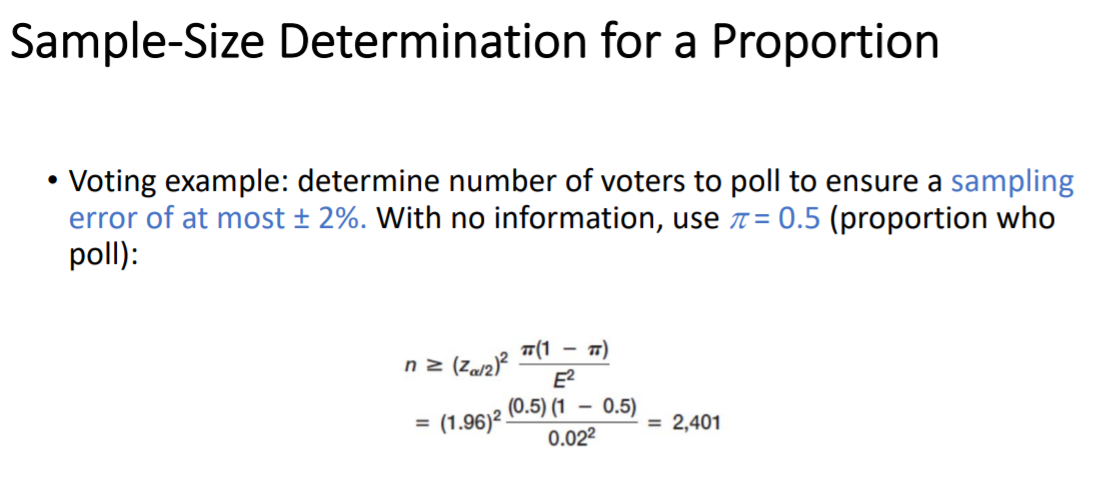
cbind(lPI.age2,uPI.age2) *# reverse transform*

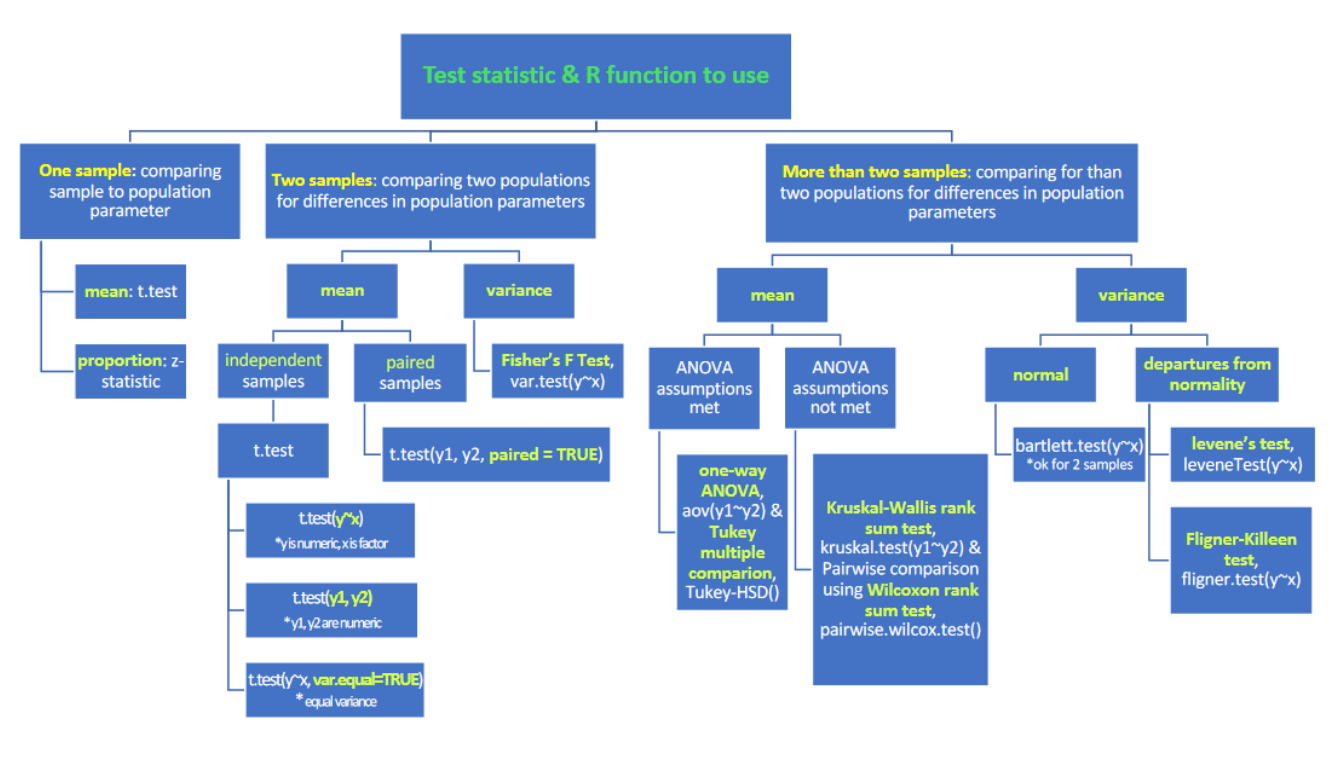
*# if lambda above 0*

lPI.total2 <- (lPI.totalt)\*\*(1/0.2)  
uPI.total2<- (uPI.totalt)\*\*(1/0.2)  
cbind(lPI.total2,uPI.total2)

99% prediction interval for colgpa is [0.9551664 4.350205]. This means that given the observed GPA of students, the GPA for a new student will lie within this interval of [0.9551664 4.350205] with 99% level of confidence. More generally, with repeated random sampling, 99% of such constructed predictive intervals would contain the GPA of a new student.

Sample Size Determination



Hypothesis Testing

One Sample T test

|  |  |
| --- | --- |
| *#using t.test function*  t.test(BD$Age,  alternative="two.sided",  mu=35,  conf.level = 0.95) | *#part ii*  t.test(BD$Age,  alternative="less",  mu=40,  conf.level = 0.95) |

Calculating Proportion: One sample T test

*# compute z-statistic for proportion*

age50<- BD %>% filter(Age>50)

p50 <- nrow(age50)/nrow(BD)

z <- (p50 - 0.18) / sqrt(0.18\*(1-0.18)/nrow(BD))

z

## [1] -3.97715

*#compute critical value*

cv.age50<-qnorm(0.05)

cv.age50

## [1] -1.644854

z<cv.age50

## [1] TRUE

From our results (z-statistic=-3.98 & z-critical=-1.64), the z-statistic is lying in the lower critical region. Thus we have sufficient evidence to reject H0 and accept that proportion of Age is statistically less than 0.18 at the 5% level of significance.

Two-Sample T test

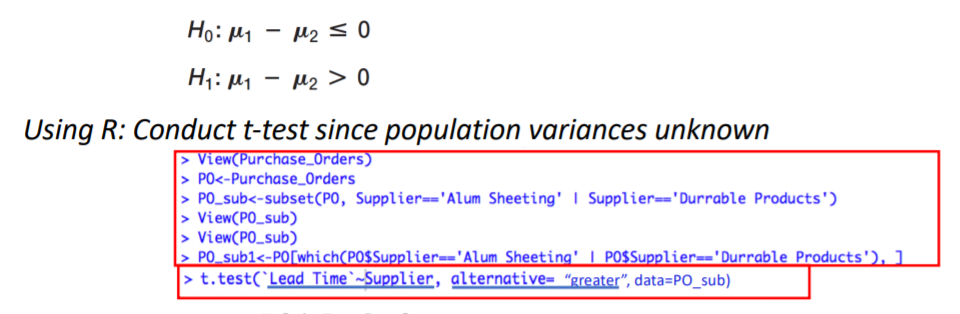
*#conduct required t.test on Age*

t.test(usecar$Age, newcar$Age)

# If comparison needed

Df1 <- subset(PO, supplier == ‘Alum sheeting’ | supplier == ‘durable’

t.test(`Lead Time`~Supplier, alternative = “greater”/”less”, data = df1)



Note: If Variances are equal, set t.test(var.equal=TRUE)

Note: If paired, set t.test(paired=TRUE)

ANOVA Assumptions

Null (H0): Means of the different age groups are equal Alternate (H1): At least one mean of an age group is signifcantly different from the others

i) measures are randomly and independently obtained (assume this is true; random sample) - ii) normally distributed (no but ANOVA is quite robust to deviations from normality) - iii) equal variances across populations (we need to check if sample sizes are equal, if not then we need to check this)

Step1: Check for normal distribution

Step 2: *# Since sample sizes are not equal, check if equal variance assumption is met*

fligner.test(Age ~ `Loan Purpose`, BD.less)

Step 3:

BD.less$`Loan Purpose`<- as.factor(BD.less$`Loan Purpose`)

aov.age<- welch\_anova\_test(BD.less$Age ~ BD.less$`Loan Purpose`) *#note the group variable should be a factor*

summary(aov.age)

Step4:

TukeyHSD(aov.age) OR

gh.out <- games\_howell\_test(BD.less, Age ~ Loan) #does not assume normality and equal variances

Test if variances are equal:

If 2 samples: Fisher’s F-test (comparing variances)

1. Var.test(outcome ~ `Loan Purpose`, BD.less)

H0: VA - VD = 0 ; H1: VA - VD ≠ 0

If more than 2:

1. fligner.test(Age ~ `Loan Purpose`,BD.less)

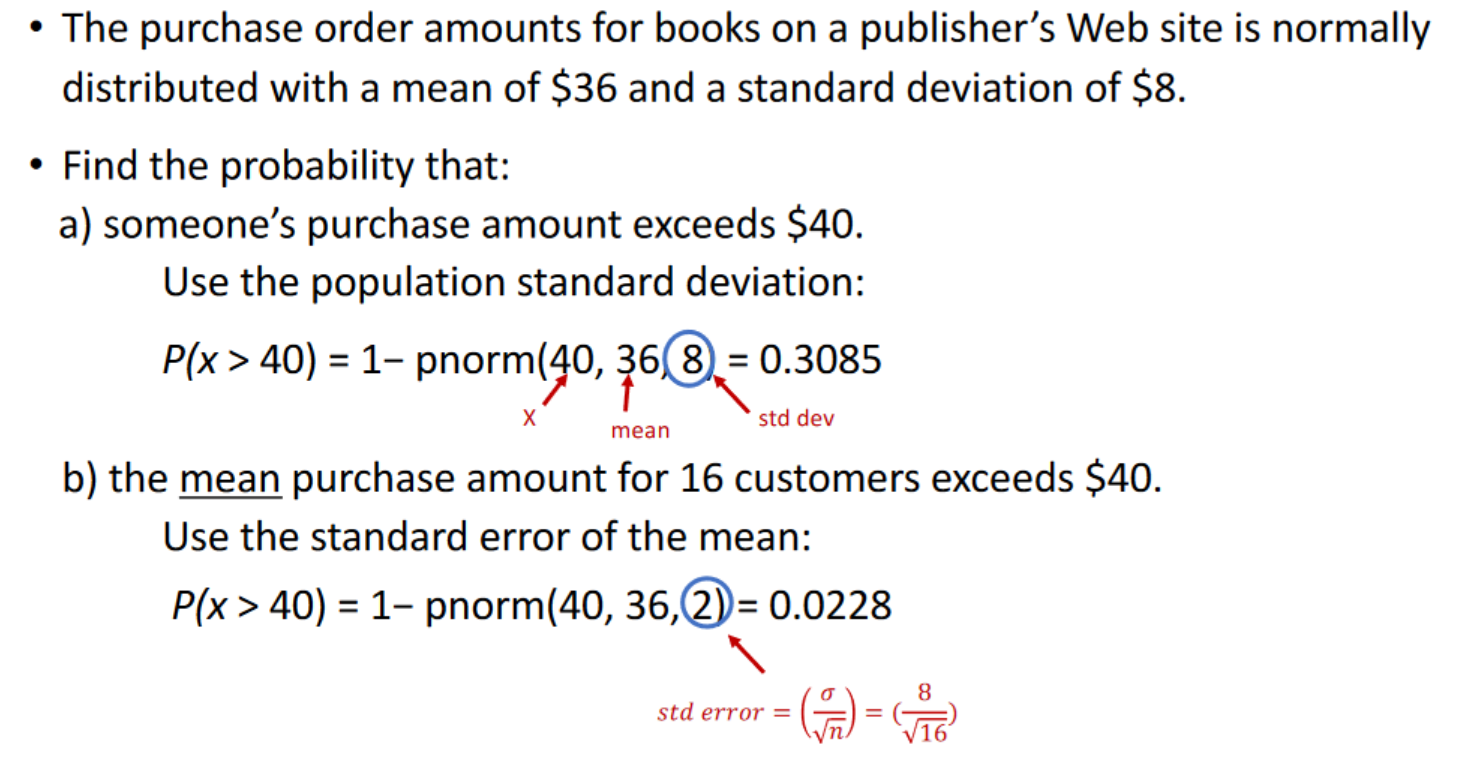
Sampling Error: occurs because samples are only a subset of population. Error depends on the size of the sample relative to population

* Increase sample size

Non Sampling Error: occurs when samples are not representative of the population ( poor sample design, wrong population frame etc

* unbiased sampling

Calculate standard error of mean



Z score (standardised normal curve)

(qnorm(0.025) or (qnorm(1-0.025)

P score (standardised normal curve)

pnorm(75,60,15,lower.tail=FALSE)

* Compute sample sizes needed to ensure a confidence interval for means and proportions with a specified margin of error

**Type Errors**

1. **Type I Error** ( p = α = P( rejectingH0 | H0 is true) ) H0 is True, but you rejected H0. ( E.g Wrongly convicting a person )
2. **Type II Error** ( p = β = P(not rejecting H0| H0 is false) ) H1 is True, but you rejected H1. (Person is guilty, wrongly let go ) H0 is false, but you fail to reject h0.

Generally, as α increases, β will decrease.

α is actually **level of significance / false positive**. From this, if you p<0.05, and you reject H0, there is only a 5% probability that you are giving a false positive.

β **(false negative)** is depends on parameter on how close is it to the test. If the true parameter is actually close to the test, the sample has a high chance of being on the wrong side of the test, falsely supporting that H0. Hence, the further parameter is from the hypothesized value, the smaller the β value.

Power of test shows the probability of not committing a type II error. Power of test = 1 - β. Your power of test needs to be significant.

Power of test is highly sensitive to sample size Small samples → Low Power can increase by taking large samples Large samples are needed for small α

**Difference between Type 1 & Type 2 Errors**

1. The value of alpha is set by the modeller, and the beta cannot be specified in advance, and depends on the value of the population parameter.
2. Increase in sample size will decrease the beta, and improve the power of test. Changing sample size will not affect the type 1 error.

Linear Regression

Assumptions of linear regression:

1. Linear relationship between DV and IV
2. Errors are **normally distributed**, and there are **no outliers**

* To check: examine histogram of residuals ( ) (can be extracted from lm() objects)
* Or, scatterplot of X against Y

1. Homoscedasticity (variation about regression line is constant for all values of the independent variable)

* To check: examine scatterplots of residual

1. Independence of errors
2. Multicollinearity

* Multicollinearity occurs when two or more regressors are "collinear" with each other, that is, they are very highly correlated with each other (e.g., r > .7 as a rule of thumb).
* inflated standard errors, low T values, high P values

**Residual standard error:** Unexplained variance: The lower, the better. - Residual standard errors are in the same units as the dependent variable (“Y”).

**R squared**: R2 (“R-squared”, “coefficient of determination”) measures the proportion of the variance explained by the model. - R2 varies between 0 and 1 (“unit-less”) - A value of 1 indicates a perfect fit; all data points lie on the line. The larger the value of R2, the better the fit.

**F-statistic** indicates whether the model fits our data better than an intercept-only model with no independent variables

**Residual plot:**

> eruption.lm = lm(eruptions ~ waiting, data=faithful)   
> eruption.res = resid(eruption.lm)

#We now plot the residual against the observed values of the variable waiting.

> plot(faithful$waiting, eruption.res,   
+     ylab="Residuals", xlab="Waiting Time",   
+     main="Old Faithful Eruptions")   
> abline(0, 0)                  # the horizon

# Use mtcars, run simple lm()

data(mtcars)

fit1 <- lm(mpg ~ hp, mtcars)

mtcars$predicted <- predict(fit1) # get predictions if needed

mtcars$residuals <- residuals(fit1) # get residuals

Relevelling a factor

new\_order\_data <- factor(factor\_data,levels = c("East","West","North"))

Munging

# munging, then plotting

titanic\_table <- titanic\_train %>% group\_by(Survived, Sex) %>% summarize(number=n()) %>% mutate(Survived = factor(Survived, levels=c(0,1), labels=c("No", "Yes")))

ggplot(titanic\_table, aes(x=Sex, y=number, fill=Survived)) + geom\_bar(stat = "identity") + theme\_bw()

Logistic Regression

fit\_log1 <- glm(Survived ~ Sex + Age, family="binomial", titanic\_train)

(b0): Log-odds when x1 and x2 are both zero.   
(b1): Expected increase in log-odds of event when x1 becomes 1, holding x2 (Age) constant.   
(b2): Expected increase in log-odds of event per unit-increase of x2 (Age), holding x1 (Sex) constant.

Predicting Values

Method 2: use the predict() function with manual input > fit\_log1 <- glm(Survived ~ Sex + Age, family="binomial", titanic\_train)

> predict(fit\_log1, newdata = data.frame(Sex="male", Age=35)) -1.378553

Method 2+ : use the predict() function with a dataset # predicting on a held-out "test" set (from library(titanic))

> predict(fit\_log1, newdata = titanic\_test)

Equation: Lifetime.post = 2.65936 + 0.05990Inspiration + 0.12915Product

Residual standard error: 0.3879

Intercept: xxx

The Linear Model shows that there should be an expected mean of 2.65936 Lifetime post consumers when its an Action (reference group) and unpaid post, which is significant at 0.05 alpha level.

There would be a 0.05990 and 0.12915 increase in post consumers when the post is an Inspiration or Product respectively. However, the P value for Inspiration is >0.05 hence we do not have sufficient evidence to support that Category is significant in affecting Lifetime.Post.Consumers.

On the other hand, if the post was paid, one would expect a 0.09181 increase in post consumers. The p Value for paid category is also <0.05 hence we have sufficient evidence to say that post consumer is affected by paid status. The Multiple R-squared is 0.02752, which shows that 0.02752 of the variance is explained by this linear model, which is considered quite little in comparison with xxcxx, indicating its a bad linear model thus supporting the fact that Category does not affect Lifetime Posts alot.

Ggplot

sub2 <- d2 %>% group\_by(Paid) %>% summarize(Mean = mean(Lifetime.Post.Consumers), SE = sd(Lifetime.Post.Consumers)/sqrt(n()))

sub2$Paid <- as.factor(sub2$Paid)

plot2 <- ggplot(sub2, aes(x= Paid, y=Mean, fill=Paid)) +

geom\_bar(stat="identity", color="black",

position=position\_dodge()) +

geom\_errorbar(aes(ymin=Mean+SE, ymax=Mean-SE), width=.2,

position=position\_dodge(.9)) + labs(title="Paid Status on Lifetime.Post.Consumers", x="Paid", y = "Lifetime.Post.Consumers")+

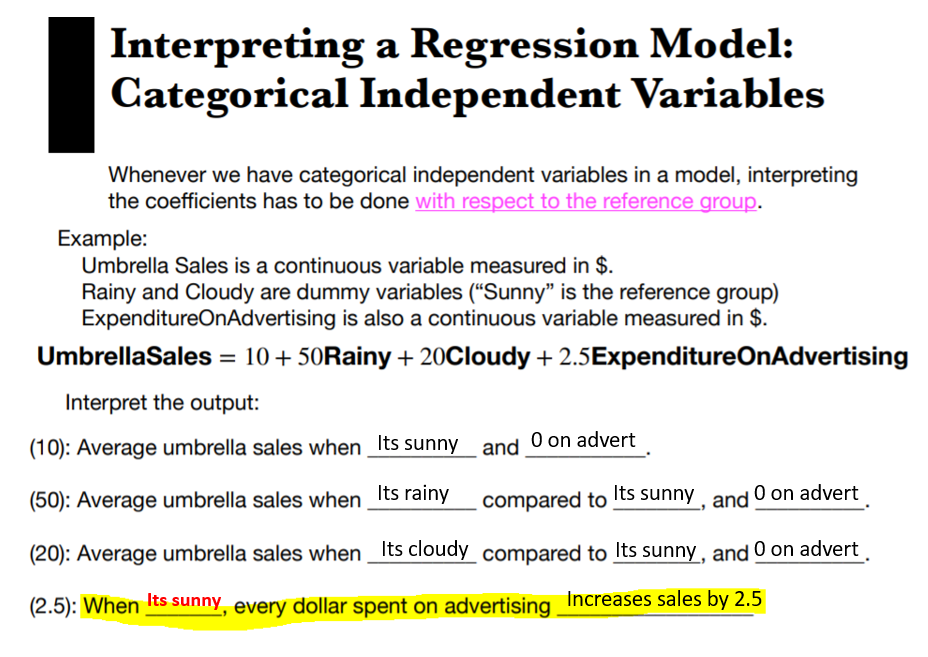
theme\_classic() +

scale\_fill\_manual(values=c('#1B9E77','#D95F02'))

dummy variable

d2$Weekendornot <- ifelse(d2$Post.Weekday == 'Saturday' |d2$Post.Weekday == 'Sunday', 1, 0 )

At any point in time, when the 65 mph law is in effect and seatbelt law is not in effect, the average total number of accidents decreases by 1318.83, compared to if the 65mph law was not in effect.



**Time Series**

* Trends refer to a gradual upwards/downwards movement of a time series over time
* Seasonal effects refer to effects that occurs/repeats at a fixed interval
* Cyclical effects refer to longer-term effects that don’t have a fixed interval/length

# **removing unused columns**

d1\_wide\_HELDOUT <- d1\_wide[,114:119] # HOLDING OUT values in 2018 and 2019  
d1\_wide <- d1\_wide[,2:113] # keeping values up to and including 2016

# **convert to a `ts` object:**

d1\_ts = ts(unlist(d1\_wide[1,1:ncol(d1\_wide)], use.names=F), frequency=4, start = c(1990, 1))

# **Split to training and test (Must be in TS alr)**   
ap\_train <- window(AirPassengers, start=1949, end=1958.917)  
ap\_test <- window(AirPassengers, start=1959, end=1960.917)

**# Need to lag?**

df$x\_lagged = dplyr::lag(df$x,1) # lagging by 1

#**Predict (Holt or LM)**

hw1 <- HoltWinters(d1\_sub, beta = F, gamma = F)  
plot(hw1)  
hw\_pred <- predict(hw1, n.ahead = 6/12)

#**RMSE**

sqrt(mean(unlist((hw\_pred[1:6] - as.vector(d1\_wide\_HELDOUT))\*\*2), na.rm = T))  
sqrt(mean((ap\_pred - ap\_test)^2))

**Interaction/Moderation**

lm(y ~ x + m + m\*x, df)

b3 is the difference in slopes (on previous day’s haze), between mothers and non-mothers.

Possible interpretations: (i) The third coefficient (on age\_group60 and above) is the expected difference in percentage of online shoppers between age groups \*at year 0\*. That is, at year 0, the model predicts there to be a 8335 percentage point more usage in the 60-and-older group, compared to the 25-to-34 year old group. (ii) This is a time-series regression, and the fourth coefficient on the interaction term is the expected difference in \*slopes\* (rate of change) of percentage of online shoppers in the 60-and-above group as compared to the 25-to-34 group. In other words, compared to the 25-to-34 group, the rate of increase of online shopping among older adults above 60 is 4.16 percentage points less per year.

Is the sub model better than full model?

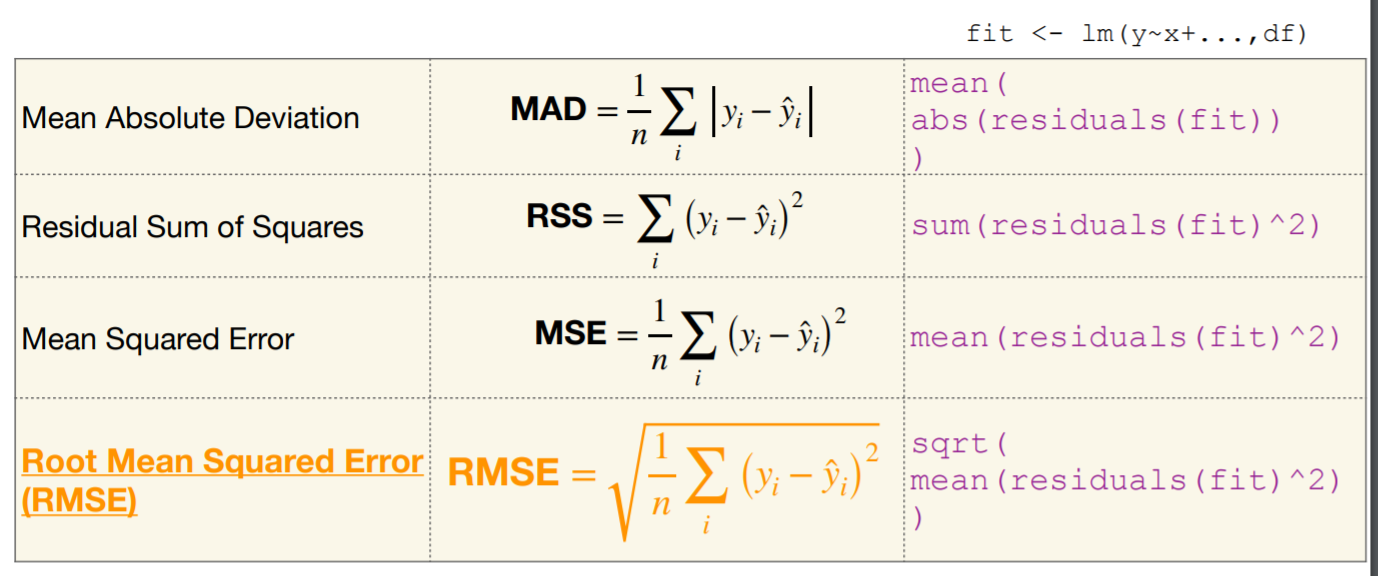
Full\_model <- lm(y ~ x1 + x2, df1)

Sub\_model <- lm(y ~ x1, df1)

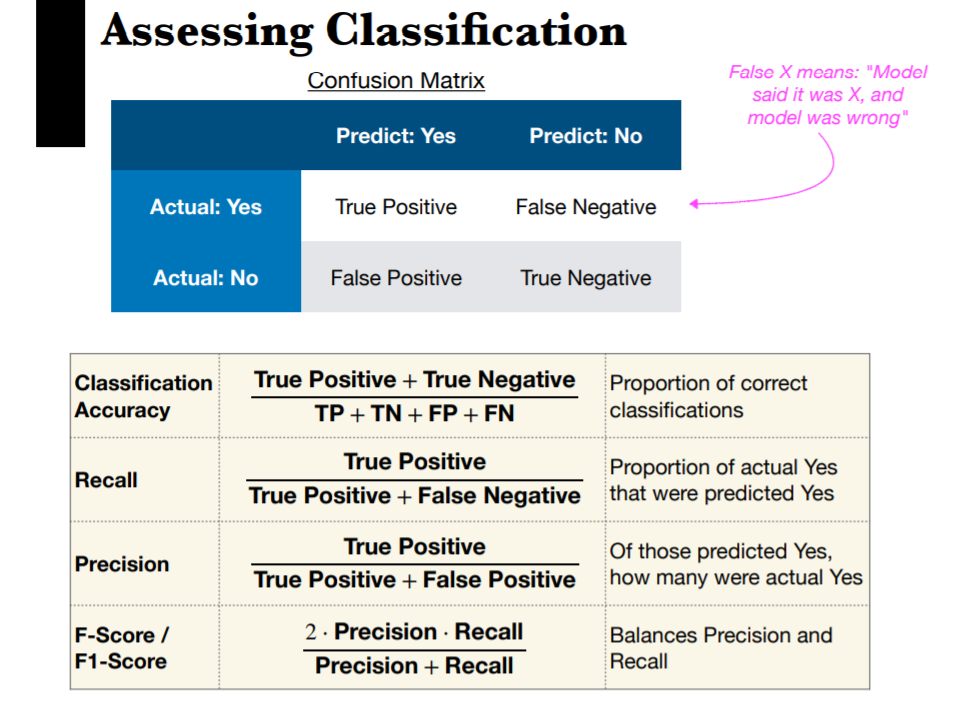
# Is the full model significantly better?

# **To use ANOVA, make sure submodel is nested within model**

Anova(full\_model, sub\_model)



* Low R2 doesn’t mean the a model is unable to capture trends in the data
* Think about effect size of Paid variable in comparison to the range of the Lifetime.Post.Consumers variable



NOTE: ONLY FOR Logistic Regression (GLM)

*# Store the probabilities back into the Test set*

test$prediction <- predict(glm, test , type="response")

test$prediction[test$prediction >= 0.5] <- 1  
test$prediction[test$prediction < 0.5] <- 0

table(test$Bankrupt, test$prediction)

## 0 1

## 0 2016 8

## 1 61 9

TP: 9   
FP: 8  
FN: 61  
TN: 2016

Accuracy = (9+2016)/2094   
Precision = TP/(TP+FP)  
Recall = TP/(TP+FN)  
F1 = 2\*P\*R/(P+R)

vector <- c(ca,precision,recall,f1)  
names(vector) <- c('Classification Accuracy', 'Recall', 'Precision', 'F1-score')

vector

Data Mining

**Principal Component Analysis (PCA)**

# **SELECT ONLY continuous variables**

data(mtcars)

mtcars1 = mtcars %>% select("cyl", "disp", "hp", "drat", "wt", "qsec", "gear", "carb")

# **running a Principal Component Analysis**  
pca1 <- prcomp(mtcars1, center=T, scale=T) #scale to standardise (value 1)

# **examining the output**  
summary(pca1)

# **examining loadings/b1 on first 2 PCs**  
pca1$rotation[,1:2]

# **extracting the PCs for LM**

d2 <- d2 %>% mutate(  
pc1 = pca1$x[,"PC1"],  
pc2 = pca1$x[,"PC2"],  
pc3 = pca1$x[,"PC3"],  
pc4 = pca1$x[,"PC4"])

# **use in an lm()**  
summary(lm(mpg ~ pc1, mtcars))

**K means clustering**

# **nstart to choose 10 diff initialisations**

km\_obj <- kmeans(df, <number\_of\_clusters>, nstart=10)

# **We can visualise using factoextra::fviz\_cluster**

**#Note: Fviz computes 2 PCA and plots it**

library(factoextra)

fviz\_cluster(km\_obj, mtcarsX)

**Elbow method**

mtcars1 = mtcars %>% select("cyl", "disp", "hp", "drat", "wt", "qsec", "gear", "carb")

wss <- rep(NA, 10)

for(k in c(1:10)) {

wss[k] = kmeans(mtcars1, k, nstart=10)$tot.withinss

}

plot(wss, type="b", xlab="Number of clusters", ylab="Total within-cluster sum of squares")

**Cross Validation (Train/Test/Validation Set)**

# split data randomly into train, valid and test (60:20:20)

set.seed(1) # for reproducibility

df$partitionNum <- sample(1:3, size=nrow(df),prob=c(0.6,0.2,0.2), replace=T)

df$partition <- factor(df$partitionNum, levels=c(1,2,3), labels=c("Train", "Valid", "Test"))

df\_train <- df %>% filter(partition=="Train")  
df\_valid <- df %>% filter(partition=="Valid")  
df\_test <- df %>% filter(partition=="Test")

Multicollinearity. It occurs when two or more regressors are very highly correlated with each other. In this case, several independent variables within each group are highly correlated with each other since they have similar definitions. This undermines the statistical significance of an independent variable, because the independent variables are not truly independent from one another. Multicollinearity reduces the precision of the estimated coefficients and the p values of the linear regression.

Linear Optimisation

Let:  
X1 = Number of Fashion gloves to produce/ number of rolls cut into pattern j  
X2 = Number of Winter gloves to produce  
X3 = Number of Work gloves to produce

Maximize total profit using decision variables $X\_1$, $X\_2$, $X\_3$ | Profit = 30 $X\_1$ + 50 $X\_2$ + 90 $X\_3$

--- | ---

Subject to |   
Budget Constraint | 15$X\_1$ + 20$X\_2$ + 45$X\_3$ $\leq$ 8000  
Space Constraint | 0.15$X\_1$ + 0.35$X\_2$ + 0.25$X\_3$ $\leq$ 1000  
Constraint 1 | $X\_1$ + $\quad$ + $\quad$ $\geq$ 300  
Constraint 2 | $\quad$ + $X\_2$ + $\quad$ $\geq$ 300  
Constraint 3 | $\quad$ + $\quad$ + $X\_3$ $\geq$ 300  
Non-Negativity Constraint 1 | $X\_1$ + $\quad$ + $\quad$ $\geq$ 0  
Non-Negativity Constraint 2 | $\quad$ + $X\_2$ + $\quad$ $\geq$ 0  
Non-Negativity Constraint 3 | $\quad$ + $\quad$ + $X\_3$ $\geq$ 0

# **SELECT ONLY continuous variables**

data(mtcars)

mtcars1 = mtcars %>% select("cyl", "disp", "hp", "drat", "wt", "qsec", "gear", "carb")

# **running a Principal Component Analysis**  
pca1 <- prcomp(mtcars1, center=T, scale=T) #scale to standardise (value 1)

# **examining the output**  
summary(pca1)

# **examining loadings/b1 on first 2 PCs**  
pca1$rotation[,1:2]

**# defining parameters**

objective.fn <- c(30, 50, 90)

# ncol <- number of decision vars

const.mat <- matrix(c(15, 20, 45,  
 0.15, 0.35, 0.25,  
 1,0,0,  
 0,1,0,  
 0,0,1),  
 ncol=3 , byrow=TRUE)

const.dir <- c("<=", "<=", ">=", ">=", ">=")  
const.rhs <- c(8000, 1000, 300, 300, 300)

**# solving model**lp.solution <- lp("max", objective.fn, const.mat, const.dir, const.rhs, compute.sens=TRUE) #Linear

lp.solution <- lp("max", objective.fn, const.mat, const.dir, const.rhs, int.vec = c(1,2,3), compute.sens=FALSE) #Integer, sensitivity not accurate

#binary.vec = c(1:24) if the decisions are binary

**# to print out the optimal objective function value**lp.solution

**# to print out the values of the decision variables (X1, X2)**lp.solution$solution

**# Shadow prices**lp.solution$duals

**# sensitivity analysis on coefficients**lp.solution$sens.coef.tolp.solution$sens.coef.from

The optimal solution, given the constraints, is to cut 73 rolls into pattern 1, 210 rolls into pattern 2, and 56 rolls into pattern 4. Do not cut any rolls into any of the other patterns (pattern 3, 5, 6). This will yield a minimum scrap of 1527 inches of metal, while satisfying the demand for next week

What are the optimal holdings of the two stocks in the portfolio? The optimal holdings of the two stocks in the portfolio is 63777 shares of 5X and 1449 shares of EG, and the optimal total return of the portfolio would be $256258.8

**Binding Constraint**: changing the constraint will change the optimal solution

**Sensitivity Explanation**:

This means that the optimal solution remains the same if:

* The profit from tables (coefficient on X1) lies between $225 and $630
* The profit from sofas (coefficient on X2) lies between $364 and $1000
* The profit from beds (coefficient on X3) lies between $ (-Infinity) and $681.48
* **Objective function:** Our goal of linear optimization is to maximize/minimize an objective function which is a linear function of both our decision variables x, given parameters θ.
* **Constraints:** The restrictions or limitations on the decision variables. They usually limit the value of the decision variables. Binding & non-binding
* **Non-negativity:** For some linear programs, the decision variables should take non-negative values. This means the values for decision variables should be greater than or equal to 0
* **Decision variables:** The variables controlled by decision maker. For example: production quantity, goods to consume, stocks to buy in, etc.
* **Shadow price:** The marginal change in the optimal objective function value that occurs if the righthand side of a constraint is changed.
* **Feasible solution:** Any solution that satisfies all the given constraints.
* **Feasible region:** The set or region of decision variable satisfying all constraints

Constraint1: 1.767442

An increase of 1 dollar in budget above $80,000 would result in an increase of appx $1.76 in profits. This is logical since increasing the budget will allow the increase production of more gloves (regardless of which) which will in turn increase profits.

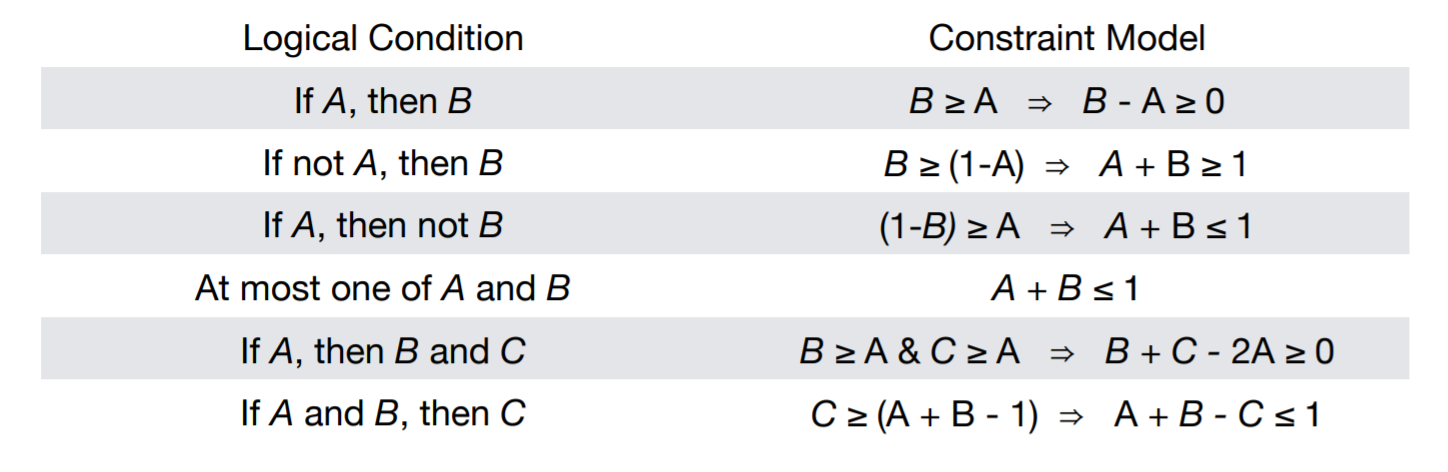
Constraint2: 41.860465  
An increase in 1 cubic-meter of space will result in ain increase of appx $41.9 in profits. This is also logical since increase space will allow increase in capacity of production, hence increasing profits.

Constraint3: -2.790698 If there is a new requirement to produce one more unit of fashion gloves above the current 300, the expected profits will decrease by -$2.79. This means that at this given constraint of 301 gloves and above, the fashion gloves are not profitable to produce since it decreases profits. This also shows that fashion gloves >= 300 is a binding constraint since the optimal solution is to produce 300 gloves (bare minimum).

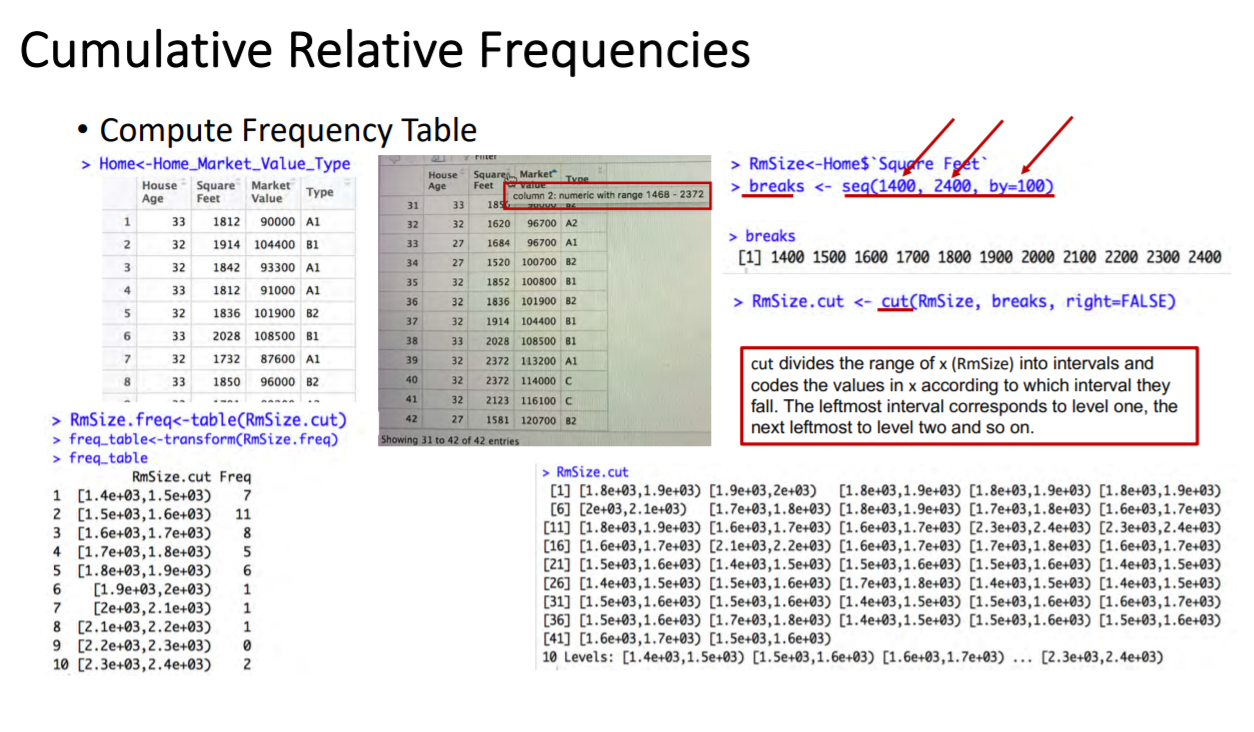
Constraint4&5: 0 A requirement to step up production of winter or work gloves by 1 unit respectively will not change the profits. This is because the optimal solution for winter and work gloves are 2k+ and 681, which are way above the demanded 300 minimum. Hence, if 301 winter or work gloves are required, it doesnt change the optimal solution since its already included in the feasible region.

Binding constraints: (1) Budget constraint <= 80,000 (2) Space constraint <= 1,000 (3) Fashion glove constraint >= 300

Non-binding: (1) Winter & Work gloves constraint >= 300 (2) Non-negativity constraints for all 3 gloves



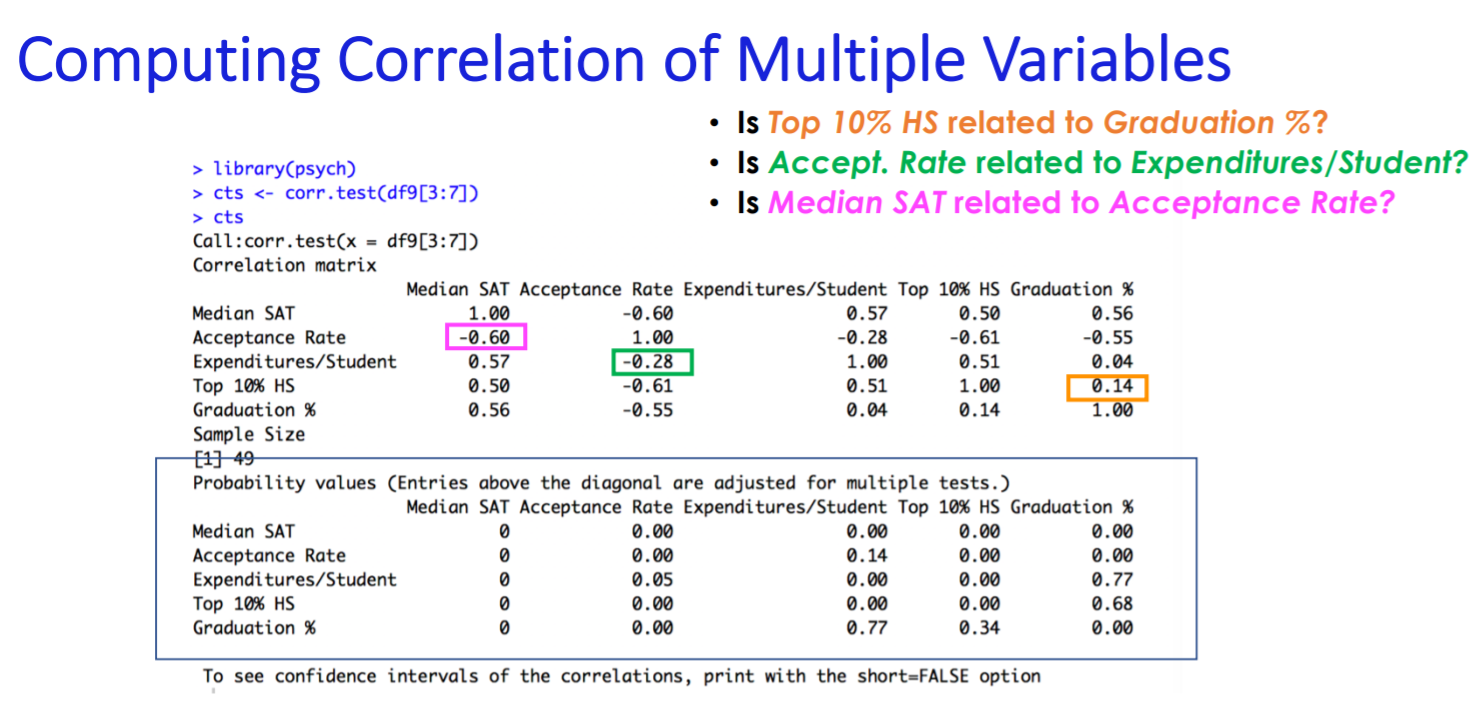
Appendix



Plotting Cumulative Relative Frequency:

Plot of Cumulative Frequencies (Ogive) – Week 3 Notes

Multiple correlation/ Corr.test



Density Plot

*# Other graphical methods to check for normality: 1) Density plot 2) Q-Q plot*

*# Density Plot (smoothed version of the histogram)*

plot(density(D$Demand.imp))

**Calculating Confidence Interval from Linear Regression**

Confidence Interval : sample statistic + margin of error

Margin of error (ME = critical value \* standard error):

1. Standard Error: Given in LM
2. T critical Value: qt(0.995, df, lower.tail=False) two tailed remember to divide alpha by 2

**Choosing best SMA window using RMSE**

For (k in 1:10){

Df1$new\_cases\_sma <- SMA(df1$new\_cases, n = k)]

Print(k)

Print(sqrt(mean((df1$new\_cases\_sma - df1$new\_cases)^2), na.rm = T))

}

**Lagging variables by country**

Recall if you lag the whole column by 1, the first row for each country will be the last row of the previous country. Have to group\_by!

Df2 <- df2 %>%

group\_by(country)%>%

mutate(new\_cases\_lag = dplyr::lag(new\_cases,n=1,default = NA))

**Assumptions of GLM**

1. Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
2. GLM requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.
3. GLM requires no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.
4. GLM assumes linearity of independent variables and log odds. although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
5. GLM typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10\*5 / .10).

|  |
| --- |
| ```{r preparation, echo=TRUE }  ``` |