MIDTERM1

B01537135

2022-10-18

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

#QUESTION 1

#1a

```
data1<- filter(data, Order==1, PureRTs<5000)
```

μ represents the mean RT on random words

#H0: $\mu = 844$msec #H1: $\mu > 844$msec

Test statistic is $\mu = 844$msec

```
data1_normalword <- filter(data1, Stimulus=="Normal")
```

```
data1_randomword <- filter(data1, Stimulus=="Random")
```

```
sd(data1_randomword$PureRTs)
```

```
## [1] 507.7824
```

```
mean(data1_randomword$PureRTs)
```

```
## [1] 1076.394
```

1076.39 - 844 / 507.78/sqrt38

z = 2.821, p = 0.00239

```
1-pnorm(2.821)
```

```
## [1] 0.00239371
```

Since p=0.002 at alpha=0.05, we can reject the null hypothesis. The p value measures the probability of getting a more extreme value than the one from the experiment, since it is smaller than our level of significance, this value falls into the right tail and we can reject the null.

We can conclude that at the given level of significance there is enough evidence to support the alternative claim, which is that the time it took to start typing random words is higher than that of normal words.

#1b

```
data1_bigramword <- filter(data1, Stimulus=="Bigrams")
```

```
sd(data1_bigramword$PureRTs)
```

```
## [1] 433.0271
```

```
mean(data1_bigramword$PureRTs)
```

```
## [1] 924.9543
```

1.96(433.03/sqrt38)

(924.95 - 137.68, 924.95 + 137.68)

787.27, 1062.63

We are 95% confident that the true mean value falls between the confidence interval (787.27, 1062.63).

#1c # Since the test statistic (844msec) falls within the confidence interval, there is not significant evidence that it takes more time to start typing for a bigram compared to a normal word. # This on its own does not support the claim that claim that the mean reaction time differs significantly depending on whether the word is a "normal" word or bigram

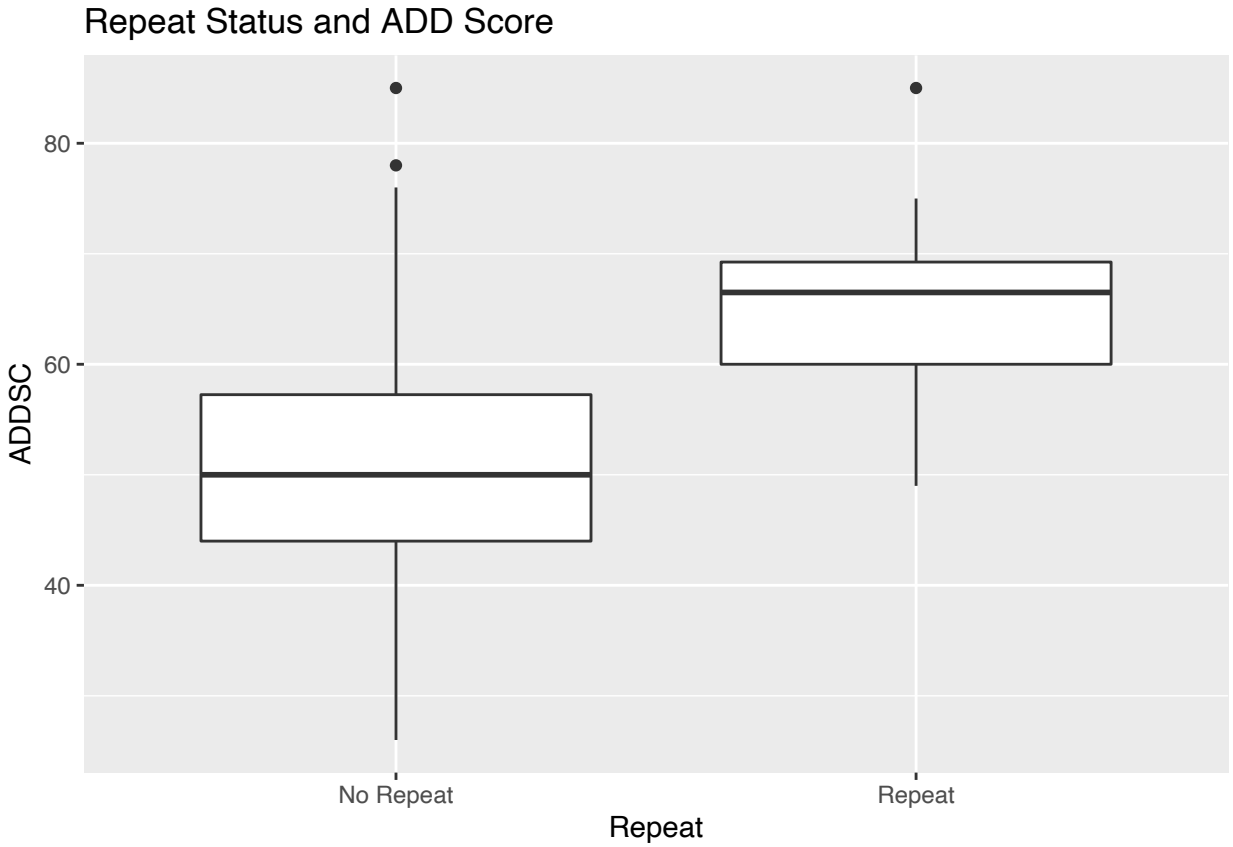
#QUESTION 2

```
add <- read.table("https://www.uvm.edu/~statdhtx/fundamentals9/DataFiles/Add.dat", header=TRUE)
```

#2a

```
add <- add%>%
mutate(Repeated=ifelse(Repeat==0,"No Repeat","Repeat"))
```

```
ggplot(add)+
  geom_boxplot(aes(as.character(Repeated), ADDSC))+
  labs(title = "Repeat Status and ADD Score", x = "Repeat", y = "ADDSC")
```



Looking at the boxplot, it seems as though those who repeated a grade (coded by 1) had higher ADD test scores, and therefore, exhibit more ADD-like behaviors than those who did not repeat a grade. The IQR is much higher in those who repeated the grade, as is the lowest possible score for ADD test (around 50 compared to around 10 for non repeaters). This means that looking at this data alone, if a child has repeated a grade his score is higher than that of a child who has not. There are higher-end outliers in both cases.

#2b

```
add1 <- add%>%
  select(Dropout, IQ, GPA)
```

```
add1 <- add1 %>%
  mutate(Dropout=ifelse(Dropout==0,"Finished", "Dropped Out"))
```

```
IQGPADropout <- add1 %>%
  group_by(Dropout) %>%
  summarize(MeanIQ= mean(IQ), MeanGPA= mean(GPA))
IQGPADropout
```

```
## # A tibble: 2 x 3
##   Dropout     MeanIQ MeanGPA
##   <chr>        <dbl>   <dbl>
## 1 Dropped Out   89.4    1.98
## 2 Finished     102.     2.52
```

#2c

```
addcontin <- add %>%
  mutate(ADDScore=ifelse(ADDSC <= 60, "<=60", ">60"))
```

```r
addtable <- table(addcontin$ADDScore, addcontin$Repeat)
colnames(addtable) <- c("No Repeat", "Repeat")
addmargins(addtable)
```

```
##
##        No Repeat Repeat Sum
##   <=60        63      3  66
##   >60         13      9  22
##   Sum         76     12  88
```

P(repeated a grade), P(R) = 12/88

P(ADD score greater than 60), P(S) = 22/88

If independent, P(R) x P(S) = P(R and S)

0.1364 x 0.25 = 0.1023

The equation is false, therefore the events are not independency and rely on one another to some degree.

QUESTION 3

#a P(republican or lean towards GOP n college degree) # P(college | republican) = P(republican) * P(college | republican) # 0.44 * 0.3 = 0.132 = 13.2%

#b P(college degree) # P(college degree) = P(college | republican) * P(republican) + P(college | democrat) * P(democrat) + P(college | independent) * P(independent) # (0.49 * 0.41) + (0.44 * 0.30) + (0.07 * 0.355) = 0.35775 = 35.76%

#c P(democrat or lean towards democratic party given college degree) # P(democrat | college) = P(democrat n college) / P(college) # (0.49 * 0.41) / 0.35775 = 0.5616 = 56.16%

QUESTION 4

```r
help <- read_csv("https://raw.githubusercontent.com/SCosta352/Math240/master/help.csv")
```

```
## Rows: 453 Columns: 88
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): substance, racegrp
## dbl (86): id, e2b1, g1b1, i11, pcs1, mcs1, cesd1, indtot1, drugrisk1, sexris...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

#4a P(patient is female | reported more than 10 drinks per day in the last 30 days)

```r
helpcontin1 <- help %>%
  mutate(i1=ifelse(i1 <= 10, "<=10", ">10"))
```

```
helpcontin1 <- helpcontin1%>%filter(!is.na(i1))
helpcontin1$i1
```

```
##   [1] ">10"  ">10"  "<=10" "<=10" "<=10" "<=10" ">10"  ">10"  ">10"  ">10"
##  [11] "<=10" ">10"  ">10"  ">10"  ">10"  "<=10" "<=10" "<=10" "<=10" ">10"
##  [21] ">10"  "<=10" ">10"  "<=10" "<=10" "<=10" "<=10" ">10"  "<=10" ">10"
##  [31] "<=10" "<=10" "<=10" "<=10" ">10"  "<=10" "<=10" ">10"  "<=10" ">10"
##  [41] "<=10" "<=10" ">10"  ">10"  "<=10" ">10"  ">10"  "<=10" "<=10" "<=10"
##  [51] ">10"  ">10"  "<=10" ">10"  ">10"  ">10"  "<=10" "<=10" "<=10" ">10"
##  [61] ">10"  ">10"  "<=10" ">10"  "<=10" ">10"  ">10"  "<=10" ">10"  "<=10"
##  [71] ">10"  "<=10" "<=10" ">10"  "<=10" ">10"  "<=10" "<=10" ">10"  "<=10"
##  [81] "<=10" "<=10" ">10"  "<=10" "<=10" "<=10" "<=10" "<=10" ">10"  ">10"
##  [91] ">10"  "<=10" ">10"  ">10"  "<=10" ">10"  "<=10" "<=10" ">10"  ">10"
## [101] "<=10" ">10"  ">10"  "<=10" ">10"  "<=10" "<=10" ">10"  ">10"  ">10"
## [111] ">10"  ">10"  "<=10" "<=10" ">10"  "<=10" ">10"  ">10"  ">10"  ">10"
## [121] ">10"  ">10"  "<=10" ">10"  "<=10" "<=10" ">10"  "<=10" ">10"  ">10"
## [131] ">10"  "<=10" "<=10" "<=10" ">10"  ">10"  "<=10" ">10"  "<=10" "<=10"
## [141] ">10"  ">10"  ">10"  "<=10" "<=10" "<=10" ">10"  "<=10" ">10"  ">10"
## [151] "<=10" "<=10" ">10"  ">10"  ">10"  ">10"  "<=10" ">10"  ">10"  "<=10"
## [161] "<=10" ">10"  ">10"  "<=10" ">10"  ">10"  ">10"  ">10"  ">10"  "<=10"
## [171] ">10"  "<=10" "<=10" ">10"  ">10"  "<=10" ">10"  "<=10" "<=10" ">10"
## [181] "<=10" "<=10" ">10"  ">10"  ">10"  ">10"  ">10"  "<=10" "<=10" "<=10"
## [191] "<=10" ">10"  "<=10" ">10"  ">10"  "<=10" ">10"  ">10"  "<=10" ">10"
## [201] ">10"  "<=10" ">10"  ">10"  ">10"  "<=10" "<=10" ">10"  ">10"  ">10"
## [211] "<=10" ">10"  ">10"  ">10"  ">10"  "<=10" ">10"  "<=10" ">10"  ">10"
## [221] ">10"  "<=10" ">10"  ">10"  "<=10" "<=10" ">10"  "<=10" ">10"  "<=10"
## [231] ">10"  ">10"  "<=10" ">10"  ">10"  "<=10" ">10"  "<=10" ">10"  "<=10"
## [241] ">10"  ">10"  ">10"  ">10"  "<=10" "<=10" "<=10" "<=10" ">10"  ">10"
## [251] "<=10" "<=10" "<=10" ">10"  ">10"  "<=10" "<=10" ">10"  "<=10" "<=10"
## [261] "<=10" ">10"  ">10"  "<=10" "<=10" "<=10" "<=10" ">10"  ">10"  "<=10"
## [271] "<=10" ">10"  ">10"  "<=10" "<=10" "<=10" ">10"  ">10"  ">10"  "<=10"
## [281] ">10"  ">10"  ">10"  "<=10" "<=10" ">10"  ">10"  ">10"  "<=10" "<=10"
## [291] ">10"  "<=10" ">10"  ">10"  ">10"  "<=10" ">10"  "<=10" ">10"  "<=10"
## [301] ">10"  ">10"  "<=10" ">10"  ">10"  ">10"  ">10"  "<=10" ">10"  ">10"
## [311] ">10"  ">10"  "<=10" ">10"  "<=10" "<=10" ">10"  ">10"  "<=10" ">10"
## [321] "<=10" "<=10" "<=10" "<=10" "<=10" "<=10" "<=10" "<=10" ">10"  "<=10"
## [331] "<=10" "<=10" "<=10" ">10"  ">10"  ">10"  ">10"  "<=10" "<=10" ">10"
## [341] "<=10" "<=10" "<=10" "<=10" ">10"  ">10"  "<=10" "<=10" ">10"  ">10"
## [351] ">10"  "<=10" ">10"  ">10"  "<=10" "<=10" "<=10" ">10"  ">10"  ">10"
## [361] ">10"  ">10"  ">10"  ">10"  "<=10" "<=10" "<=10" "<=10" "<=10" ">10"
## [371] ">10"  ">10"  ">10"  ">10"  "<=10" ">10"  "<=10" "<=10" "<=10" "<=10"
## [381] ">10"  "<=10" "<=10" ">10"  ">10"  "<=10" ">10"  ">10"  ">10"  ">10"
## [391] ">10"  ">10"  ">10"  "<=10" ">10"  ">10"  ">10"  "<=10" ">10"  ">10"
## [401] ">10"  ">10"  "<=10" ">10"  ">10"  ">10"  ">10"  ">10"  ">10"  "<=10"
## [411] ">10"  ">10"  ">10"  "<=10" "<=10" "<=10" "<=10" ">10"  "<=10" ">10"
## [421] ">10"  "<=10" ">10"  ">10"  ">10"  "<=10" ">10"  "<=10" ">10"  "<=10"
## [431] ">10"  ">10"  ">10"  "<=10" "<=10" ">10"  "<=10" ">10"  ">10"  ">10"
## [441] ">10"  ">10"  ">10"  ">10"  "<=10" "<=10" ">10"  "<=10" "<=10" ">10"
## [451] "<=10" ">10"  ">10"
```

```
helptable1 <- table(helpcontin1$female, helpcontin1$i1)
rownames(helptable1) <- c("Male", "Female")
addmargins(helptable1)
```

```
##
##          <=10 >10 Sum
##   Male    145 201 346
##   Female   59  48 107
##   Sum     204 249 453
```

48 / 249 = 0.1927 = 19.28%

#4b P(patient reported cocaine as primary substance of abuse)

```
substancecount<-table(help$substance)
addmargins(substancecount)
```

```
##
## alcohol cocaine  heroin     Sum
##     177     152     124     453
```

P(cocaine primary substance of use) = 152/453 = 0.3355 = 33.55%

#4c.i P(10 patients selected, more than half say primary substance of abuse is cocaine)

```
1- pbinom(5, 10, 0.3355)
```

```
## [1] 0.07880439
```

= 0.0788

#4c.ii If 1200 selected, how many expected to report primary substance of abuse is coaine # E(x) = n*p* # *1200* 0.3355 = 402.6 # 403 people

QUESTION 5

#5a Percent of total runners in wave 1 # P(wave 1) = P(155 < X <= 198) = P(X <= 198) - P(X <= 155) # for 155, z = 155-238 / 47 = -1.77 # for 198, z = 198-238 / 47 = -0.85 # using z-score table z(-1.77) = 0.384 # using z-score table z(-0.85) = 0.1977 # So P(-1.77 <= z <= 0.85) = 0.1977 - 0.384 = 0.1593

#5b P(selected runner is in wave 2 or less) # P(x <= 244) # for 244, z = 244-238 / 47 = 0.13 # using z-score table z(0.13) = 0.548 # P(z <= 0.13) = 0.54776 = 54.776%

#5c P(group of 30 runners has average of 200 min or less) # for 200, z = 200-238 / 47 / sqrt(30) = -4.43 # using z-score table z(-4.43) = close to zero (not on table) # so likelihood of group of 30 being under 200 min is close to 0

#5d What time separates fastest 10% of runners # for x, z = 0.10, on table body, around -1.28 # x-238 / 47 = -1.28 # x = 178 minutes # 178 minutes separates the fastest 10% of the runners

QUESTION 6

```
babies <- read_csv("https://raw.githubusercontent.com/SCosta352/CLPS0900/main/babies3")
```

```
## Rows: 96 Columns: 153
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (7): study_code, dob, dot1, dot2, dot3, stim, module
```

```
## dbl (146): id, exp1, exp2, exp3, exp4, exp5, female, dad, train, Baseline_Pr...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
babies1 <- filter(babies, exp2==1)
```

```
babies1 <- babies1 %>%
    select(Baseline_Proportion_Gaze_to_Singer, Test_Proportion_Gaze_to_Singer)
```

```
babies1 <- babies1%>%
  mutate(D=Test_Proportion_Gaze_to_Singer-Baseline_Proportion_Gaze_to_Singer)
```

```
D_bar <- mean(babies1$D)
s_D <- sd(babies1$D)
D_bar
```

```
## [1] 0.0250904
```

```
s_D
```

```
## [1] 0.1649521
```

#6a # H0: μD = 0 # H0: μD > 0 # where μD is μT - μB, looking at the test - the baseline gaze proportion # using right tailed test, at alpha = 0.05

#6b

```
t <- (D_bar-0)/(s_D/sqrt(32))
t
```

```
## [1] 0.8604484
```

test statistic is D_bar, the difference between means, assuming null so 0

standardized test statistic is t value = 0.8604484

#6c

```
1 - pt(t, 31)
```

```
## [1] 0.198074
```

associated p value = 0.198

#6d # Given the p value is around 0.2, 0.198 > 0.05, we fail to reject the null hypothesis that the mean difference is 0. At alpha = 0.05, there is not enough evidence to support the claim that the mean difference is above 0 - which would mean more time is spent looking at the familiar object in the test phase (since μD = μT - μB). This indicates babies may have not felt more connected to the object presence when hearing the familiar melody.