STAT511 Final Exam KEY

Fall 2019, Sections 001/801

General Comment: Remember that we do NOT accept the null hypothesis! In other words, absence of evidence is not evidence of absence. While grading, I saw many written conclusions that (incorrectly) suggested we could conclude "equality" or "no assocation" (between proportions, means, etc) when $p > 0.05$. Instead, we can say something like "we do not have evidence of an association".

Exercise

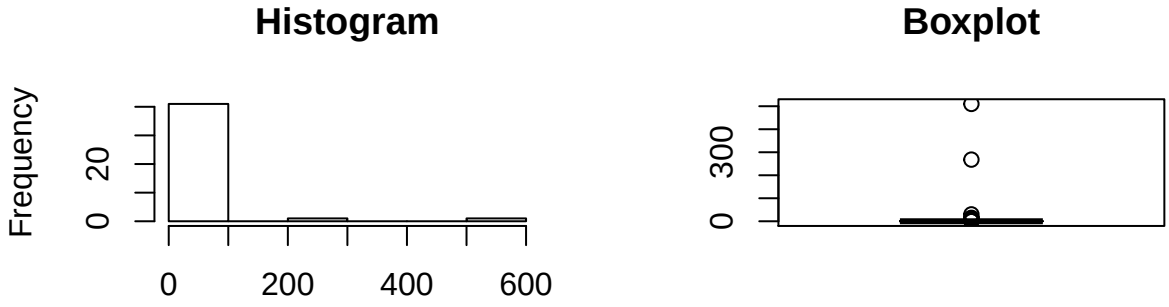
For this data we consider treating the response as either continuous or binary (CACS_score or CACS_01). Because of the large number of zeroes, I would feel comfortable using CACS_01. But the trade off is some loss of information.

We also consider treating the predictor as either continuous or binary (METh or Group). Sometimes using a grouping variable makes the results easier to interpret. The trade off is that you would need to justify the cutoff used.

Q1 (6pts)

Group	n	PropCACS	meagAge	minMETh	medMETh	maxMETh
ATH	25	0.32	50.36	73	117	209
CON	18	0.11	49.61	8	26	44

Q2 (4 pts)




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   20.18    0.00  510.00
```

Q3 (2 pts)

The analyst might consider treating CACS as binary because the distribution of CACS_score is very skewed and most of the values (>75%) are zero.

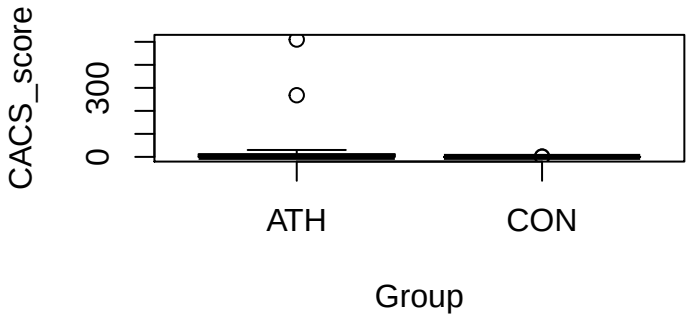
Q4 (4 pts)

Of females, 5% are positive for CACS.
Of males, 39% are positive for CACS.
-2pts for 2.3%, 20.9%.

Q5 (6 pts)

Based on Fisher's Exact test (p = 0.011), we conclude that there is an association between CACS and Sex.
A higher proportion of males are positive for CACS.
FET is preferred here due to small sample size.
Note: I know the R output for FET includes odds ratio and CI, but I would not report these.
-2pts for chi-square/Z-test/Proportions test (p = 0.022). All of these (equivalent) options generated a warning "Chi-squared approximation may be incorrect".
-3pts for t-test (p = 0.006)

Q6 (2 pts)



Q7 (6 pts)

Based on Wilcoxon rank-sum test (p = 0.064 or 0.070), we do not have evidence of an association between CACS score and Group.
CACS score is higher for athletes.
Wilcoxon test (equivalent to Kruskal-Wallis) is prefered here due to skew/non-normality.
-2pts for two-sample t-test or ANOVA (p = 0.201 or 0.137).

Q8 (6 pts)

OR = 1.027, 95% CI = (1.010, 1.051)
A **one hour increase** in METh is associated with a **multiplicative increase** of 1.027 (2.7%) in the **odds** of CACS.

Q9 (2 pts)

The Groups are defined based on METh. Hence METh and Group give "redundant" information.

Chocolate

Q10 (6 pts)

n = 83-97 per group.
-2 pts for n = 6 - 7 per group (using SE instead of sd).
-2 pts for n = 65-77 per group (corresponding to one-sided alternative).

Q11 (4 pts)

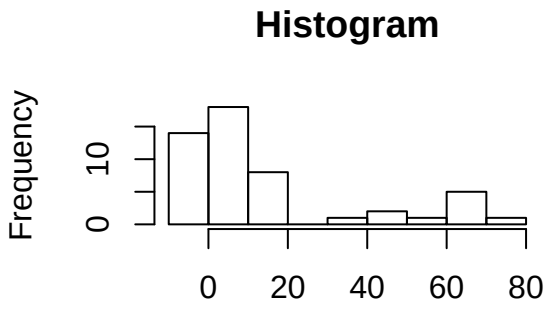
SE = 0.214 (based on TS = 2.947)
OK for SD = 0.855
-2 pts for SE = 0.242 or SD = 0.968 (based on TS = 2.602)

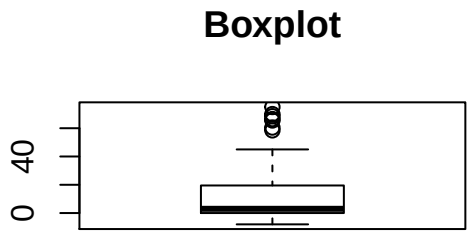
Note: Calculating sample size based on paired t-test would yield n = 17! This considerably lower than required sample size based on two-sample t-test.

Guns

Q12 (4 pts)

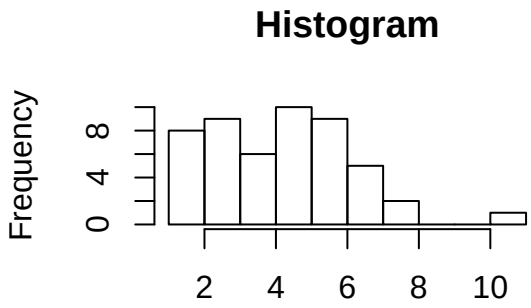
Arizona has the lowest BradyScore (-8).
California has the highest BradyScore (+75).

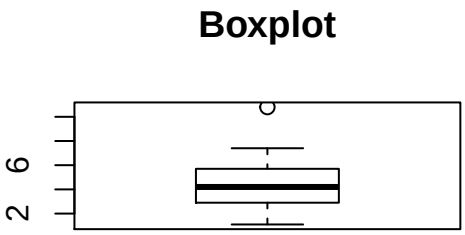




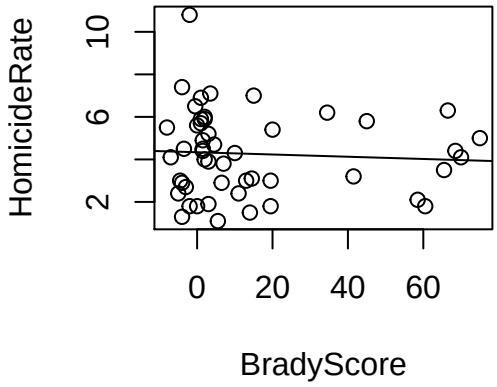
Q13 (4 pts)

New Hampshire has the lowest Homicide Rate (1.1).
Louissiana has the highest Homicide Rate (10.8).





Q14A (4 pts)



Q14B (2 pts)

slope = -0.005
p-value = 0.659
95% CI = (-0.029, 0.0185)

Q14C (2 pts)

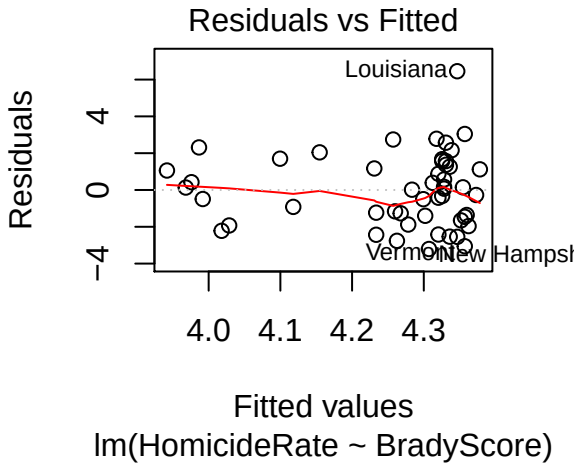
We do not have evidence of an association between Homicide Rate and Brady Score.

Q14D (2 pts)

R2 = 0.004
-1 pt for adjusted R2 = -0.017.

Q15A (4 pts)

Louissiana has **more** homicides than we would have expected based on the model.



Q15B (4 pts)

Based on the Bonferroni adjusted p-value = 0.027, we have evidence that Louissiana is an outlier. Bonferroni adjustment is appropriate here because Louissiana was identified **after** looking at the data.

Q15C (4 pts)

slope = -0.001
p-value = 0.899
95% CI = (-0.022, 0.019)
Still no evidence of association.

Q16A (4 pts)

```
##              HomicideRate PerUrban Poverty MedAge PerDgr
## HomicideRate         1.00     0.06    0.65  -0.09  -0.44
## PerUrban             0.06     1.00   -0.30  -0.28   0.45
## Poverty              0.65    -0.30    1.00  -0.06  -0.72
## MedAge              -0.09    -0.28   -0.06   1.00   0.08
## PerDgr              -0.44     0.45   -0.72   0.08   1.00
```

Q16B (2 pts)

Poverty has the strongest correlation with Homicide Rate.

Q17A (2 pts)

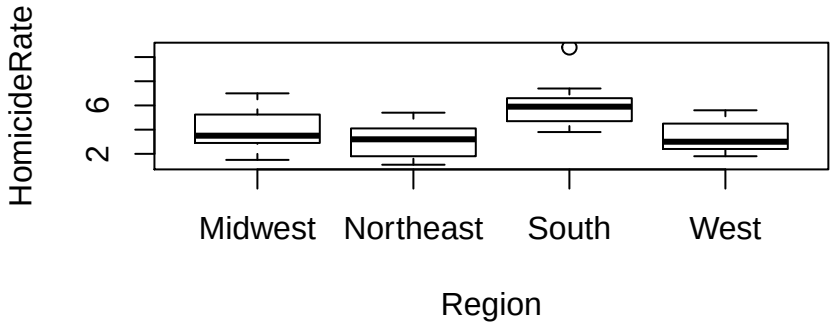
slope = 0.019
p-value = 0.048
95% CI = (+0.0002, 0.038)

Note: After accounting for Poverty, we find evidence of an association between Homicide Rate and Brady Score. But surprisingly, the association is positive?!?

Q17B (2 pts)

R2 = 0.466
-1 pt for adjusted R2 = 0.443.

Q18A (2 pts)



Q18B (4 pts)

term	df	sumsq	meansq	statistic	p.value
Region	3	70.8587	23.619567	9.125552	7.52e-05
Residuals	46	119.0613	2.588289	NA	NA

Q18C (6 pts)

Based on Tukey adjusted pairwise comparisons, we have evidence that the South has **higher** average homicide rate as compared to all other US Regions.

level1	level2	estimate	std.error	df	statistic	p.value
Midwest	Northeast	0.9833333	0.7094212	46	1.3861065	0.5143132
Midwest	South	-1.9812500	0.6143768	46	-3.2248126	0.0120124
Midwest	West	0.5653846	0.6440419	46	0.8778694	0.8162602
Northeast	South	-2.9645833	0.6703400	46	-4.4225068	0.0003357
Northeast	West	-0.4179487	0.6976294	46	-0.5990985	0.9318161
South	West	2.5466346	0.6007223	46	4.2392879	0.0006006

Appendix

```
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
Exercise <- read.csv("C:/hess/STAT511_FA11/Exams-2019/Final/Exercise.csv")
str(Exercise)
library(tidyverse)
library(kableExtra)
library(broom)
#Q1
SumStats <- Exercise %>%
            group_by(Group) %>%
            summarize(
              n = n(),
              PropCACS = mean(CACS_01),
              meagAge = mean(Age),
              minMETh = min(METh),
              medMETh = median(METh),
              maxMETh = max(METh)
            )
kable(SumStats, digits = 2)
#Q2
par(mfrow=c(1, 2))
hist(Exercise$CACS_score, main = "Histogram", xlab = "")
boxplot(Exercise$CACS_score, main = "Boxplot")
summary(Exercise$CACS_score)
#Q4-5
ExTable <- table(Exercise$Sex, Exercise$CACS_01)
ExTable
prop.table(ExTable, 1)
chisq.test(ExTable)
fisher.test(ExTable)
t.test(CACS_01 ~ Sex, data = Exercise)
#Q6
par(mfrow=c(1,1))
boxplot(CACS_score ~ Group, data = Exercise)
#Q7
aggregate(CACS_score ~ Group, FUN = median, data = Exercise)
aggregate(CACS_score ~ Group, FUN = mean, data = Exercise)
wilcox.test(CACS_score ~ Group, data = Exercise)
library(coin)
wilcox_test(CACS_score ~ Group, distribution = "exact", data = Exercise)
t.test(CACS_score ~ Group, data = Exercise)
t.test(CACS_score ~ Group, var.equal = TRUE, data = Exercise)
#Q8
ExModel <- glm(CACS_01 ~ METh, family = binomial, data = Exercise)
exp(coef(ExModel))
exp(confint(ExModel))
#Q10
#SD Milk
0.39*sqrt(16)
```

```r
#SD Dark
0.36*sqrt(16)
power.t.test(delta = 0.63, sd = 1.56, power = 0.80,
             type = "two.sample")
power.t.test(delta = 0.63, sd = 1.44, power = 0.80,
             type = "two.sample")
power.t.test(delta = 0.63, sd = 0.39, power = 0.80,
             type = "two.sample")
power.t.test(delta = 0.63, sd = 0.36, power = 0.80,
             type = "two.sample")
power.t.test(delta = 0.63, sd = 1.56, power = 0.80,
             type = "two.sample", alternative = "one.sided")
power.t.test(delta = 0.63, sd = 1.44, power = 0.80,
             type = "two.sample", alternative = "one.sided")
#Q11
TS <- qt(0.995, df = 15)
TS
SE <- 0.63/TS
SE
#Not required, just for illustration
SD <- SE*sqrt(16)
SD
power.t.test(delta = 0.63, sd = SD, power = 0.80,
             type = "one.sample")
GunData <- read.csv("C:/hess/STAT511_FA11/Exams-2019/Final/GunData.csv",row.names = 1)
str(GunData)
#Q12
par(mfrow=c(1,2))
hist(GunData$BradyScore, main = "Histogram", xlab = "")
boxplot(GunData$BradyScore, main = "Boxplot")
par(mfrow=c(1,2))
hist(GunData$HomicideRate, main = "Histogram", xlab = "")
boxplot(GunData$HomicideRate, main = "Boxplot")
#Q14
par(mfrow=c(1,1))
GunModel1 <- lm(HomicideRate ~ BradyScore, data = GunData)
plot(HomicideRate ~ BradyScore, data = GunData); abline(GunModel1)
summary(GunModel1)
confint(GunModel1)
#Q15A
plot(GunModel1, which = 1)
#Q15B
library(car)
outlierTest(GunModel1)
#Q15C
GunModel2  <- update(GunModel1, subset = -18)
summary(GunModel2)
confint(GunModel2)
#Q16
SmallData <- subset(GunData, select=HomicideRate:PerDgr)
round(cor(SmallData),2)
#Q17
GunModel3 <- lm(HomicideRate ~ BradyScore + Poverty, data = GunData)
```

```r
summary(GunModel3)
confint(GunModel3)
#Q18
boxplot(HomicideRate ~ Region, data = GunData)
GunModel4 <- lm(HomicideRate ~ Region, data = GunData)
kable(tidy(anova(GunModel4)))
library(emmeans)
emout <- emmeans(GunModel4, ~ Region)
kable(tidy(pairs(emout)))
```