STAT 511A Final Exam

Kathleen Wendt

12/18/2019

I have not given, received, or used any unauthorized assistance on this exam.

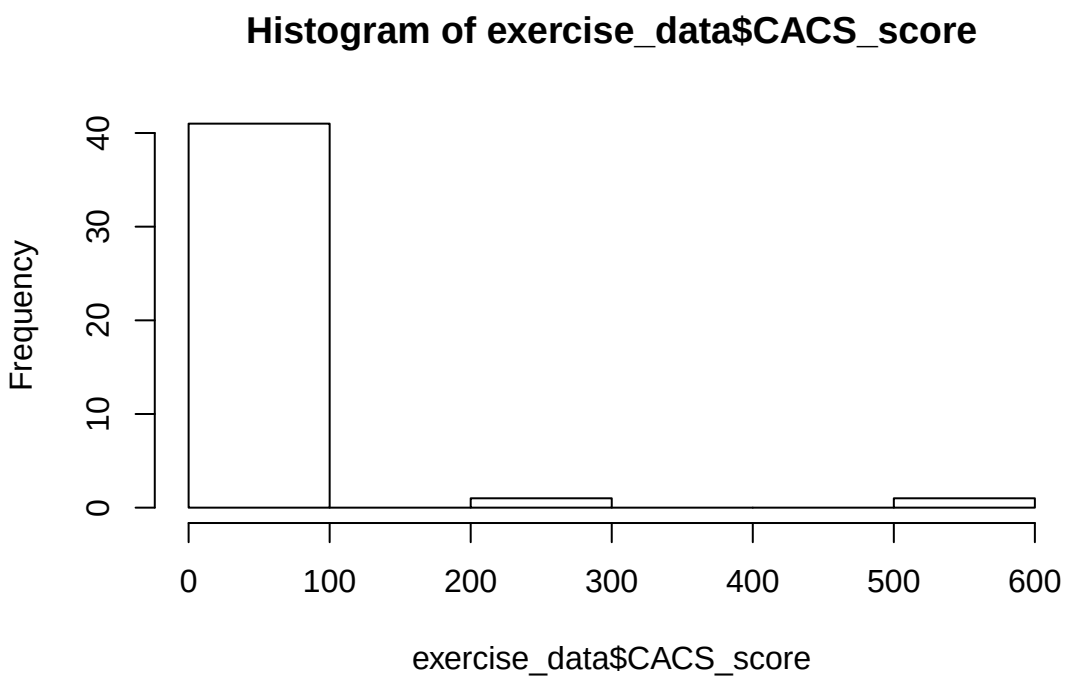
Kathleen E Wendt

Part A: Exercise and coronary artery calcium (CAC)

Question 1: CAC table

Group	n	cac_pos	mean_age	min_meth	med_meth	max_meth
ATH	25	8	50.36000	73	117	209
CON	18	2	49.61111	8	26	44

Question 2: CAC graph




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

Question 3: Rationale for `CACS_01`

Most subjects had a coronary artery calcium (CAC) score of 0, leading to high skewness and low variability. Furthermore, some analyses that might be of interest for these data could require a categorical predictor.

Question 4: CAC by sex

0.05 of females tested positive for coronary artery calcium.

0.3913043 of males tested positive for coronary artery calcium.

Question 5: Difference in CAC proportions

Research question

Is there a sex difference in proportions of those with coronary artery calcium?

Approach

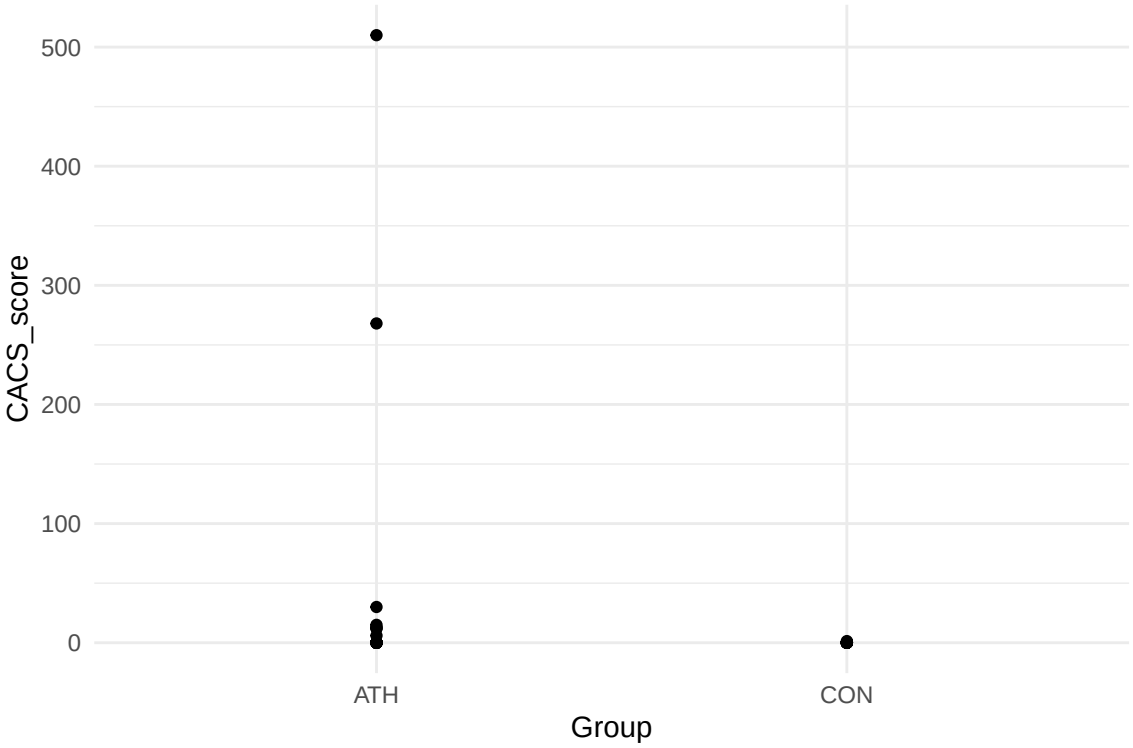
- Predictor: `Sex` (categorical)
- Response: `CACS_01` (categorical)
- Selected test: Fisher's Exact Test

	cac	no_cac
females	0.0500000	0.9500000
males	0.3913043	0.6086957

Conclusion

Based on Fisher's Exact Test for equality of proportions with small sample sizes, there is a difference in proportion of coronary artery calcium (CAC) by sex, such that males have a higher proportion of CAC compared to females. The corresponding p-value is 0.0112321, which is less than $\alpha = .05$.

Question 6: CAC graph by group



Question 7: Difference in CAC means

Research question

Is there a difference between athletes and non-athletes in the level of coronary artery calcium?

Approach

- Predictor: `Group` (categorical)
- Response: `CACS_score` (continuous)
- Selected test: Wilcoxon rank sum test (exact; non-parametric)

Conclusion

Based on the Wilcoxon rank sum test using the exact distribution, there is no evidence to suggest a difference between groups (athletes and non-athletes) in mean coronary artery calcium level, although the group of athletes has a higher mean level (descriptively). The corresponding p-value is 0.06466, which is above $\alpha = .05$.

Question 8: Logistic regression

Research question

Does the number of metabolic equivalent hours (METh) predict the presence of CAC?

Approach

- Predictor: `METh` (continuous)
- Response: `CACS_01` (categorical)
- Selected test: Logistic regression

Conclusion

Logistic regression yielded a model-based estimate of odds ratio of 1.0265684 with a corresponding 95% confidence interval of (1.0097901, 1.0512333). Every one unit increase in metabolic equivalent hours per week (METh) multiplies the odds of coronary artery calcium by 1.0265684. METh was a significant predictor of CAC, $p = 0.0075166 < \alpha = .05$.

Question 9: METh & Group

One variable (`Group`) was derived from the other (`METh`). Subjects were split into group (athletes vs. non-athletes) based on their metabolic equivalent hours (METh) per week from their physical activity diaries; the control group had METh below 60, and the athlete group had METh equal to or above 60. A formal test of these variables would not reveal any new information.

Part B: Chocolate

Question 10: Sample size for two-sample design

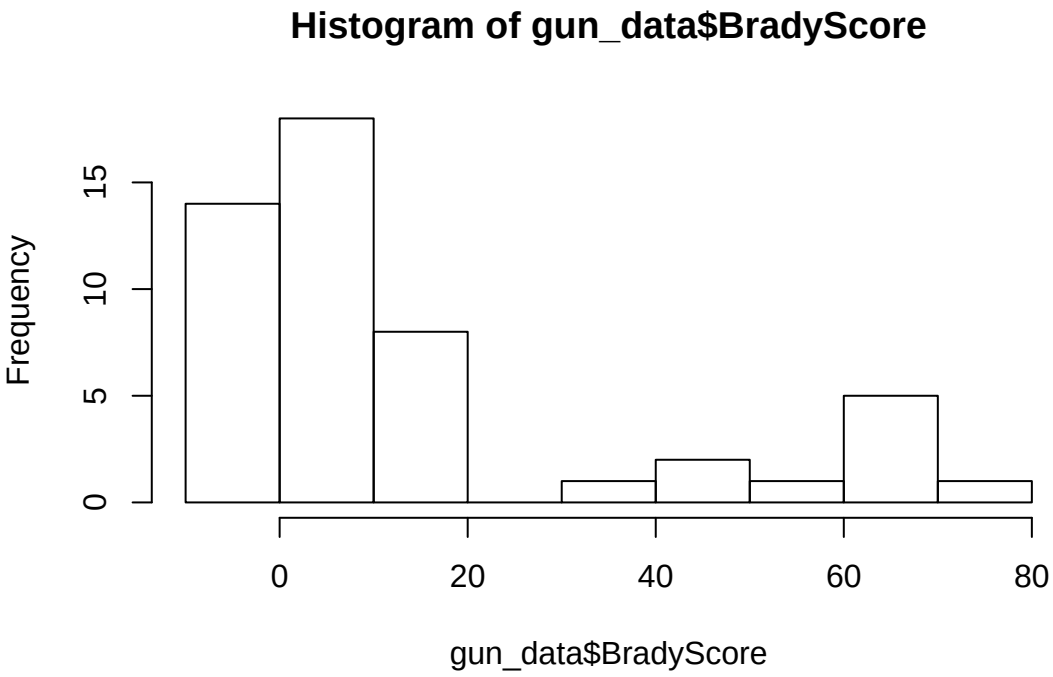
90 subjects per group are needed to achieve 80% power for a two-sample, two-sided t-test.

Question 11: Standard error of difference in means

Based on $df = 15$ and two-sided p-value of 0.01, the test statistic is approximately 2.947; therefore, the standard error of the difference in means is 0.2137767, and the corresponding sample standard deviation of the differences in means is 0.8551069.

Part C: Guns

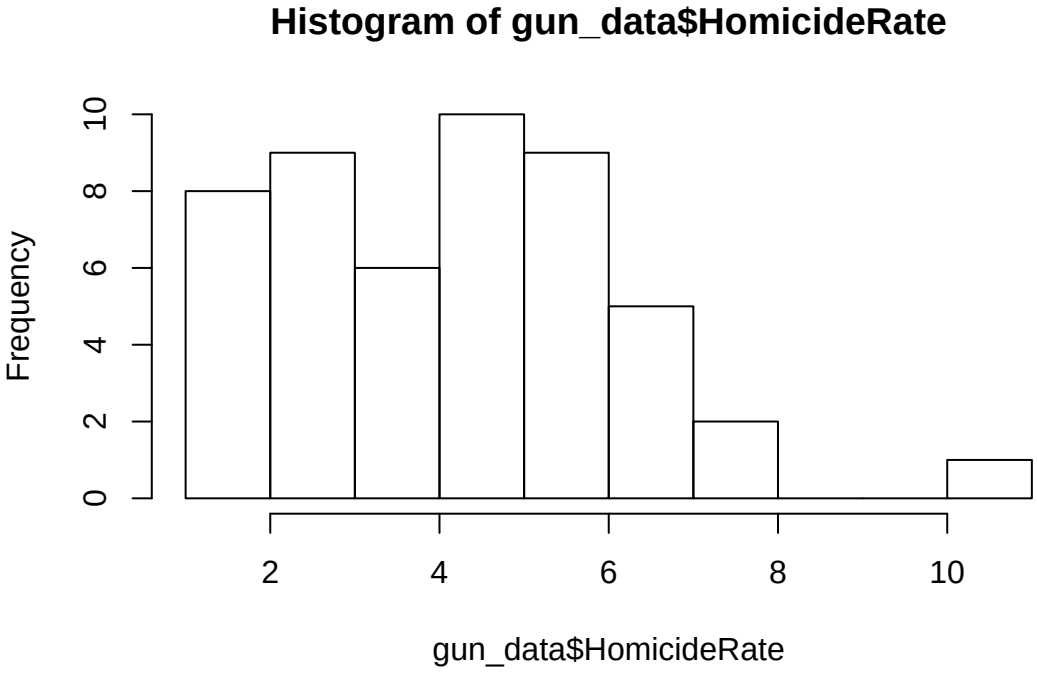
Question 12: Brady score



Arizona has the lowest Brady score (-8), indicating a low level of gun restriction.

California has the highest Brady score (75), indicating a high level of gun restriction.

Question 13: Homicide rate

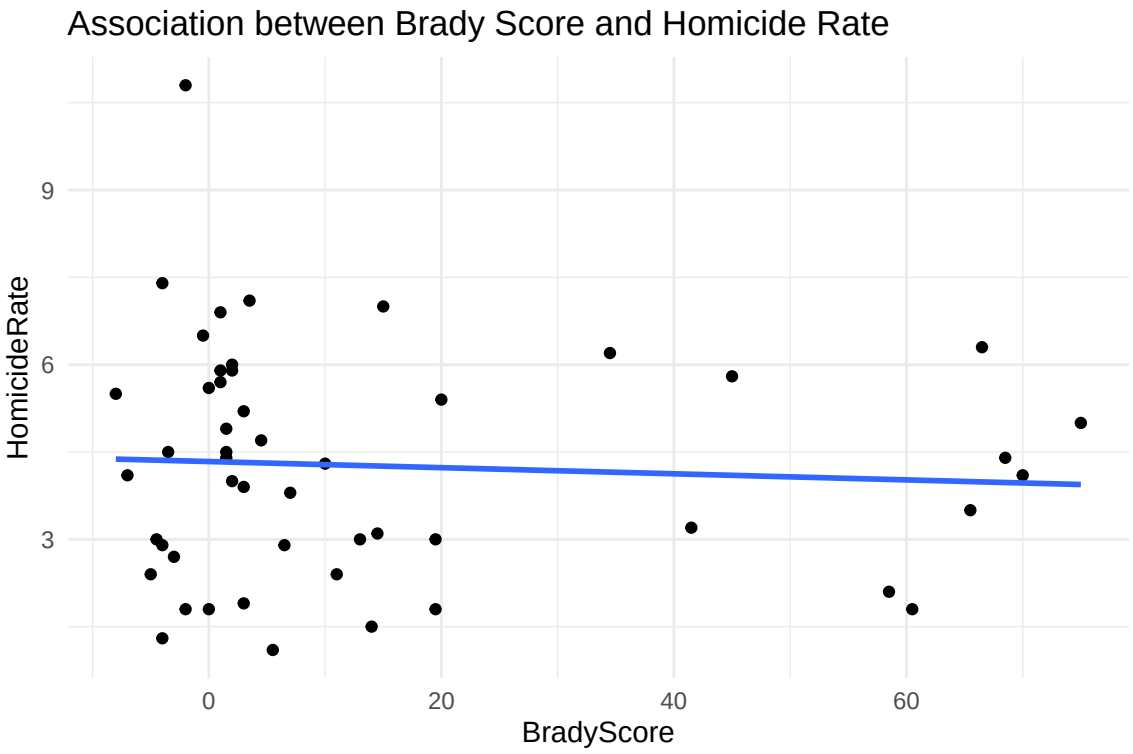


New Hampshire has the lowest homicide rate (1.1 homicides per 100,000).

Louisiana has the highest homicide rate (10.8 homicides per 100,000).

Question 14: Simple linear regression

14A: Scatterplot



14B: Slope estimate

The model-based estimate of the slope is -0.0052566 with a corresponding p-value of 0.6591869.

14C: Conclusion

We cannot conclude there is a linear association between homicide rate and Brady score. There is no evidence to suggest that the level of gun restriction, as measured by Brady score, is related to homicide rate in the United States of America, excluding Washington, D.C.

14D: Multiple R-squared

As indicated by the multiple R^2 value, 0.004086 of the variation in homicide rate is explained by the linear model of Brady score and homicide rate.

Question 15: Checking assumptions

15A: Outlying residual

Louisiana has the largest magnitude residual. There are *more* homicides (10.8 per 100,000) in Louisiana than expected based on the model (4.347 per 100,000).

15B: Outlier test

The Bonferonni adjusted p-value for observation 18 (Louisiana) is $0.0270186 < \alpha = 0.05$. There is evidence to suggest Louisiana is an outlier in terms of homicide rate. A Bonferonni adjustment is appropriate because multiple tests (one for each state) were conducted before confirming the outlier.

15C: Drop Louisiana and re-run analysis

After removing Louisiana from the gun data and re-running the linear model regressing Brady score on homicide rate, the slope estimate is -0.0013543, and the corresponding p-value is 0.8986654. Based on this updated model, we still cannot conclude there is a linear association between gun restriction (Brady score) and homicide rate. Removing the outlier (Louisiana) did not change the ultimate conclusions of this model.

Question 16: Homicide rate and demographic variables

16A: Correlation matrix

rowname	HomicideRate	PerUrban	Poverty	MedAge	PerDgr
HomicideRate	NA	0.0606973	0.6473680	-0.0944459	-0.4358440
PerUrban	0.0606973	NA	-0.3029615	-0.2759823	0.4469987
Poverty	0.6473680	-0.3029615	NA	-0.0634999	-0.7240876
MedAge	-0.0944459	-0.2759823	-0.0634999	NA	0.0818834
PerDgr	-0.4358440	0.4469987	-0.7240876	0.0818834	NA

16B: Strongest correlate

Poverty is the most strongly (positively) correlated with homicide rate, $r = 0.647368$.

Question 17: Multiple regression

17A: Slope estimate

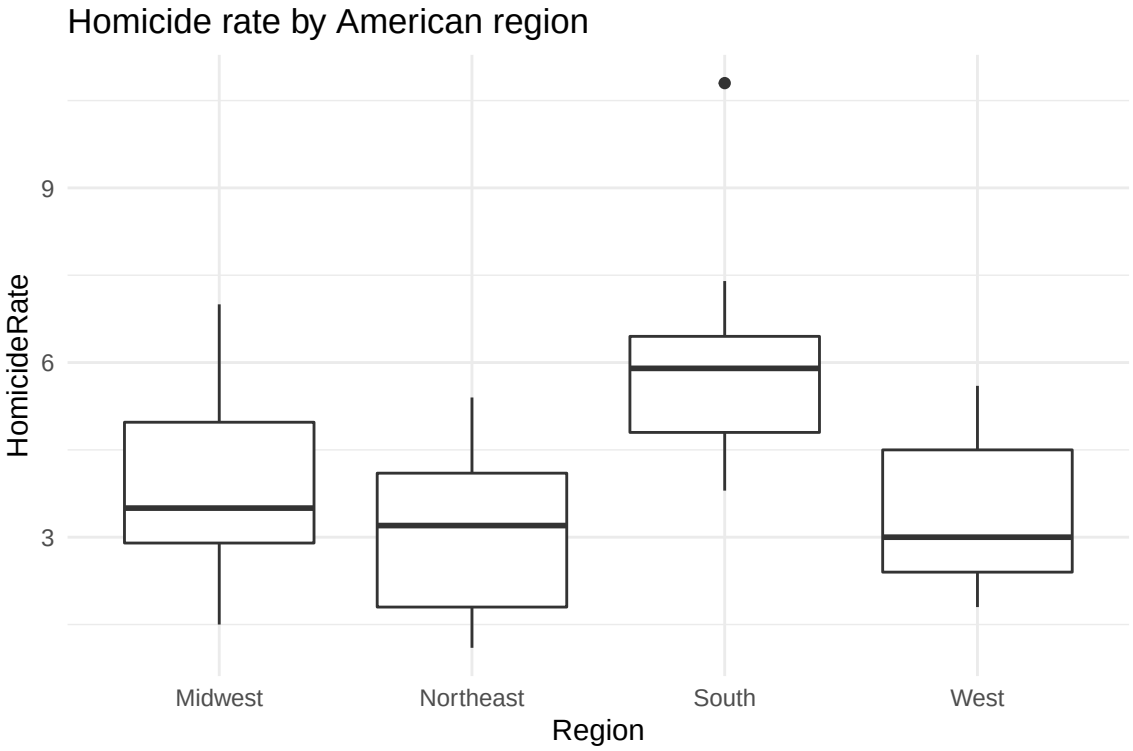
The multiple regression model yielded a slope estimate of 0.0132861 and corresponding p-value of 0.3351374 for Brady score.

17B: Multiple R-squared

As indicated by the multiple R^2 value, 0.5064 of the variation in homicide rate is explained by the multiple regression model.

Question 18: One-way ANOVA

18A: Box plot



18B: ANOVA table

```
## Analysis of Variance Table
##
## Response: HomicideRate
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Region     3  70.859 23.6196  9.1256 7.521e-05 ***
## Residuals 46 119.061  2.5883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

18C: Pairwise comparisons

Based on Tukey-adjusted pairwise comparisons, the South has a higher homicide rate compared to each of the three other regions:

- Midwest, $p = 0.0120$
- Northeast, $p = 0.0003$
- West, $p = 0.0006$

There are no differences between the Midwest, Northeast, and West in homicide rate.

Appendix

```r
# set global options
library(knitr)
knitr::opts_chunk$set(fig.width = 6,
                      fig.height = 4,
                      fig.path = "figs/",
                      echo = FALSE,
                      warning = FALSE,
                      message = FALSE)
# load packages
library(tidyverse)
library(kableExtra)
library(broom)
library(car)
library(coin)
library(corrr)
library(emmeans)
# A. read exercise data
exercise_data <- readr::read_csv("data/Exercise.csv")
# 1. create cacs summary table
kableExtra::kable(
  exercise_data %>%
  dplyr::group_by(Group) %>%
  dplyr::summarize(n = n(),
            cac_pos = sum(CACS_01),
            mean_age = mean(Age),
            min_meth = min(METh),
            med_meth = median(METh),
            max_meth = max(METh)) %>%
  dplyr::ungroup())
# 2. create histogram of cacs
hist(exercise_data$CACS_score)
# 2. calculate cacs number summary
summary(exercise_data$CACS_score)
# 4. calculate female proportion
female_prop <- sum(exercise_data$Sex == "F"
                & exercise_data$CACS_01 == 1) /
  sum(exercise_data$Sex == "F")
# 4. calculate male proportion
male_prop <- sum(exercise_data$Sex == "M"
                & exercise_data$CACS_01 == 1) /
  sum(exercise_data$Sex == "M")
# 5. set up proportion table
cacs_sex_prop <- matrix(c(1, 19, 9, 14),
                    nrow = 2,
                    byrow = TRUE)
rownames(cacs_sex_prop) <- c("females", "males")
colnames(cacs_sex_prop) <- c("cac", "no_cac")
# 5. create prop table to check proportions
kableExtra::kable(prop.table(cacs_sex_prop, margin = 1))
# 5. conduct exact test for equality of proportions with small sample sizes
cacs_fisher <- broom::tidy(fisher.test(x = cacs_sex_prop))
```

```r
# 6. create summary plot for cacs by group
exercise_data %>%
  ggplot(aes(x = Group, y = CACS_score)) +
  geom_point() +
  theme_minimal()
# 7. coerce group to factor
exercise_data$Group <- as.factor(exercise_data$Group)
# 7. check assumption of equal variances
car::leveneTest(CACS_score ~ Group,
                data = exercise_data,
                center = "median")
# 7. check assumption of normality
exercise_data %>%
  ggplot(aes(x = CACS_score)) +
  facet_wrap("Group") +
  geom_histogram() +
  theme_minimal()
shapiro.test(exercise_data$CACS_score)
# 7. conduct nonparametric wilcoxon rank sum test
coin::wilcox_test(CACS_score ~ Group,
                  data = exercise_data,
                  distribution = "exact")
# 8. conduct logistic regression
cacs_glm <- glm(formula = CACS_01 ~ METh,
                family = binomial(link = "logit"),
                data = exercise_data)
# 8. review glm summary
summary(cacs_glm)
# 8. create tidy glm object
cacs_glm_tidy <- broom::tidy(cacs_glm)
# 8. exponentiate for odds ratio estimate
cacs_odds <- broom::tidy(exp(cacs_glm$coefficients))
# 8. exponentiate for OR CI estimate
cacs_odds_ci <- broom::tidy(exp(confint(cacs_glm)))
# 10. power calculation for two-sample, two-sided t-test
power.t.test(delta = 0.63,
             sd = 1.5,
             sig.level = 0.05,
             power = 0.80,
             type = "two.sample",
             alternative = "two.sided")
# C. read gun data
gun_data <- readr::read_csv("data/GunData.csv")
# 12. create histogram for brady score
hist(gun_data$BradyScore)
# 12. filter to state with minimum brady score
min(gun_data$BradyScore)
gun_data %>% dplyr::filter(BradyScore == "-8")
# 12. filter to state with maximum brady score
max(gun_data$BradyScore)
gun_data %>% dplyr::filter(BradyScore == "75")
# 13. create histogram for homicide rate
hist(gun_data$HomicideRate)
```

```r
# 13. filter to state with minimum homicide rate
min(gun_data$HomicideRate)
gun_data %>% dplyr::filter(HomicideRate == "1.1")
# 13. filter to state with maximum homicide rate
max(gun_data$HomicideRate)
gun_data %>% dplyr::filter(HomicideRate == "10.8")
# 14. fit linear model regressing brady score on homicide rate
guns_lm <- lm(HomicideRate ~ BradyScore, data = gun_data)
guns_lm_tidy <- broom::tidy(guns_lm)
# 14a. plot brady scores and homicide rates with regression line
gun_data %>%
  ggplot(aes(x = BradyScore, y = HomicideRate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Association between Brady Score and Homicide Rate") +
  theme_minimal()
# 14d. check multiple R-squared value
summary(guns_lm)
# 15. check plot of residuals vs fitted values for guns_lm
plot(guns_lm, which = 1)
# 15a. extract observation 18
gun_data %>% slice(18)
# 15a. compare predicted and actual homicide rate for Louisiana
predict(guns_lm)
# 15b. conduct outlier test for Louisiana
gun_outlier <- car::outlierTest(guns_lm)
# 15c. drop Louisiana
gun_data_dropla <- gun_data %>% dplyr::filter(State != "Louisiana")
# 15c. re-run simple linear regression from Q14
guns_lm_nola <- lm(HomicideRate ~ BradyScore, data = gun_data_dropla)
# 15c. tidy lm
guns_lm_nola_tidy <- broom::tidy(guns_lm_nola)
# 16a. create correlation matrix of homicide rate with demo vars
gun_demo_corr <- gun_data %>%
  dplyr::select(HomicideRate, PerUrban, Poverty, MedAge, PerDgr) %>%
  corrr::correlate()
# 16a. kable correlation matrix
kableExtra::kable(gun_demo_corr)
# 17. fit multiple regression model
gun_mult_lm <- lm(HomicideRate ~ BradyScore
                  + PerUrban + Poverty + MedAge + PerDgr,
                  data = gun_data)
# 17. tidy multiple regression output
gun_mult_lm_tidy <- broom::tidy(gun_mult_lm)
# 17b. check multiple R-squared value
summary(gun_mult_lm)
# 18. create lm object with homicide rate and region
gun_reg_lm <- lm(HomicideRate ~ Region, data = gun_data)
# 18. fit and tidy anova model
gun_anova_tidy <- broom::tidy(anova(gun_reg_lm))
# 18a. create boxplot for homicide rate by region
gun_data %>%
  ggplot(aes(x = Region, y = HomicideRate)) +
```

```
  geom_boxplot() +
  ggtitle("Homicide rate by American region") +
  theme_minimal()
# 18b. show anova table
anova(gun_reg_lm)
# 18c. convert region to factor
gun_data <- gun_data %>% dplyr::mutate(Region = as.factor(Region))
# 18c. conduct tukey-adjusted pairwise comparisons
gun_emout <- emmeans(gun_reg_lm, pairwise ~ Region)
# 18c. examine contrasts
gun_emout$contrasts
# 18c. create compact letter display
CLD(gun_emout$emmeans)
```