B201A HW2

Lillian Chen

10/27/2021

1. A study investigating stopping distances (in feet) of automobiles traveling at different speeds yielded the following results:

Automobile speed n Mean Std. Deviation 10 m.p.h. 20 20.2 5.1 20 m.p.h. 20 36.3 8.4
30 m.p.h. 5 66.4 15.5

(a) Carry out a one-way analysis of variance (making the usual equal-variance assumption), and summarize whether the ANOVA F-test suggests that the hypothesis of equality of group means should be rejected at the $\alpha = 0.05$ level.

Solution: The ANOVA F-value comes out to $F = 68.1$ with a $p < 0.001$, indicating that the hypothesis of equality of group means should be rejected at the $\alpha = 0.05$ level.

```
autodata <- data.frame(speed = c("10mph", "20mph", "30mph"),
                       n = c(20,20,5),
                       means = c(20.2,36.3,66.4),
                       sd = c(5.1,8.4,15.5))

#using anovaMean function from package HH for one-way ANOVA with summary statistics
anovaMean(autodata$speed, autodata$n, autodata$means, autodata$sd, ylabel=c("automobile speed"))
```

```
## Warning in ybar - (ybar %*% n)/sum(n): Recycling array of length 1 in vector-array arithmetic is dep
##   Use c() or as.vector() instead.

## Analysis of Variance Table
##
## Response: automobile speed
##
## Terms added sequentially (first to last)
##                   Df Sum of Sq Mean Sq F value     Pr(F)
## automobile speed  2    9060.6  4530.3  68.056 6.661e-14 ***
## Residuals        42    2795.8    66.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Carry out two-sample t-tests comparing the 10-m.p.h. group to the 20-m.p.h. group assuming both equal and unequal variances. Comment on similarities and differences in the findings, and in particular on the extent to which you think that possible violation of the equal-variance assumption affects your interpretations.

Solution:

$H_0 : \mu_{10} = \mu_{20}$

$H_A : \mu_{10} \neq \mu_{20}$

For both the equal variances and unequal variances two-sample t-tests, the test statistic had a value of 7.326, and the critical value for a two-sided test at $\alpha = 0.05$ with 38 degrees of freedom was 2.024. In both the equal variance and unequal variance two-sample t-tests, we reject the null (TS > CV) and conclude that the means of the 10-m.p.h. group and the 20-m.p.h. group are significantly different from each other. Because the sample sizes of the two groups were equal, there were no differences in the findings since the denominator of the test statistic results in the same value in both calculations. Additionally, it is unlikely there is violation of the equal-variance assumption since the ratio of the variances was less than 3, following the rule of thumb that it is okay to use the equal variance assumption.

```
#assuming equal variances
(diff <- 36.3-20.2)
```

```
## [1] 16.1
```

```
(s_pooled <- sqrt((19*5.1^2+19*8.4^2)/(20+20-2)))
```

```
## [1] 6.948741
```

```
(TS <- diff/(s_pooled*sqrt(1/20+1/20)))
```

```
## [1] 7.326891
```

```
pt(TS,38)
```

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

```
#assuming unequal variances
(TS <- diff/sqrt(5.1^2/20 + 8.4^2/20))
```

```
## [1] 7.326891
```

```
pt(TS,38)
```

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

```
#t value for 0.05, 38 df
qt(1-0.05/2, 38)
```

```
## [1] 2.024394
```

(c) Carry out two-sample t-tests comparing the 20-m.p.h. group to the 30-m.p.h. group assuming both equal and unequal variances. Comment on similarities and differences in the findings, and in particular on the extent to which you think that possible violation of the equal-variance assumption affects your interpretations.

Solution:

$H_0 : \mu_{20} = \mu_{30}$

$H_A : \mu_{20} \neq \mu_{30}$

The test statistic for the equal variances two-sample t-test had a value of 6.018, and the test statistic for the unequal variances two-sample t-test had a value of 4.191. The critical value for a two-sided test at $\alpha = 0.05$ with 23 degrees of freedom was 2.069. In both the equal variance and unequal variance two-sample t-tests, we reject the null (TS $>$ CV) and conclude that the means of the 20-m.p.h. group and the 30-m.p.h. group are significantly different from each other. Both tests definitively rejected the null hypothesis, however the test for unequal variances yielded a test statistic lower in magnitude due to the difference in the denominator, which is based on the separate standard error for each group instead of taking the pooled estimate like in the equal variances scenario.

There is possible violation of the equal-variance assumption since the ratio of the variances between the groups was greater than 3, and following the rule of thumb, it is not okay to use the equal variance assumption and we must consider the unequal variances test.

```
#assuming equal variances
(diff <- 66.4-36.3)
```

```
## [1] 30.1
```

```
(s_pooled <- sqrt((4*15.5^2+19*8.4^2)/(5+20-2)))
```

```
## [1] 10.00356
```

```
(TS <- diff/(s_pooled*sqrt(1/5+1/20)))
```

```
## [1] 6.017855
```

```
pt(TS,23)
```

```
## [1] 0.9999981
```

```
#assuming unequal variances
(TS <- diff/sqrt(15.5^2/5 + 8.4^2/20))
```

```
## [1] 4.19116
```

```
pt(TS,23)
```

```
## [1] 0.9998252
```

```
#t value for 0.05, 23 df
qt(1-0.05/2, 23)
```

```
## [1] 2.068658
```

2. The following data give the fastest running times (in seconds) recorded in the modern Olympic games through 1984 for both men and women at various distances.

Distance (m) 100 200 400 800 1500 42195 (marathon)

X = Fastest time for women (sec) 11.0 21.8 48.8 113.5 236.6 8692.0

Y = Fastest time for men (sec) 9.9 19.8 43.8 103.0 212.5 7761.0

(a) Calculate the correlation between X and Y based on the entire set of results.

Solution: The correlation between X and Y based on the entire set of results is 1.

```
rundata <- data.frame(distance = c(100,200,400,800,1500,42195),
                      x = c(11.0, 21.8, 48.8, 113.5, 236.6, 8692.0),
                      y = c(9.9, 19.8, 43.8, 103.0, 212.5, 7761.0))

cor(rundata$x, rundata$y)
```

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

(b) Calculate the correlation between X and Y omitting the values for marathons. Comment on the similarity or difference comparing your answer to the answer in part (a).

Solution: The correlation between X and Y omitting the values for marathons is 0.9999848. This is very similar to the previous answer of 1 in part a). This may be because the last data point is so far out that it has an impact on further cementing the positive correlation observed in the first 5 distance results.

```
cor(head(rundata$x, -1), head(rundata$y, -1))
```

```
## [1] 0.9999848
```

(c) If you wanted to predict the fastest time for men and women over a distance of 1000 meters, would you be more inclined to rely on the entire sample or on the sample excluding the results from marathons? Explain your reasoning.

Solution: I would be more inclined to rely on the sample excluding the results from marathons. Since the correlation of the full sample is not too different from the truncated sample, the extreme values of the marathon data points are not errors in the data and removal of these data points would not aid in prediction. However, including the marathon data points will shift the mean of the distance further from 1000 m, meaning that the prediction interval may be wider than expected and the prediction for the fastest time for men and women over a distance of 1000 m will have more uncertainty when including the marathon data points vs excluding the marathon data points.

3. (Glantz and Slinker problem 2.2): Benzodiazepine tranquilizers (such as Valium) exert their physiological effects by binding to specific receptors in the brain. This binding then interacts with a neurotransmitter, γ-amino butyric acid (GABA) to induce changes in nerve activity. Because most direct methods of studying the effects of receptor binding are not appropriate for living human subjects, Hommer and coworkers (Archives of General Psychiatry, 1986, 43: 542-551) sought to study the effects of different doses of Valium on various readily measured physiological variables. They then looked at the correlations among these variables to attempt to identify those that were most strongly linked to the effect of the drug. Two of these variables were the sedation state S induced by the drug and the blood level of the hormone cortisol C. The data are presented below.

   Sedation score S Cortisol C, μg/dl 32 6.6 42 7.4 52 8.8 61 9.7 62 10.5 65 11.8 66 10.7

```
valiumdata <- data.frame(S = c(32,42,52,61,62,65,66),
                         C = c(6.6,7.4,8.8,9.7,10.5,11.8,10.7))
```

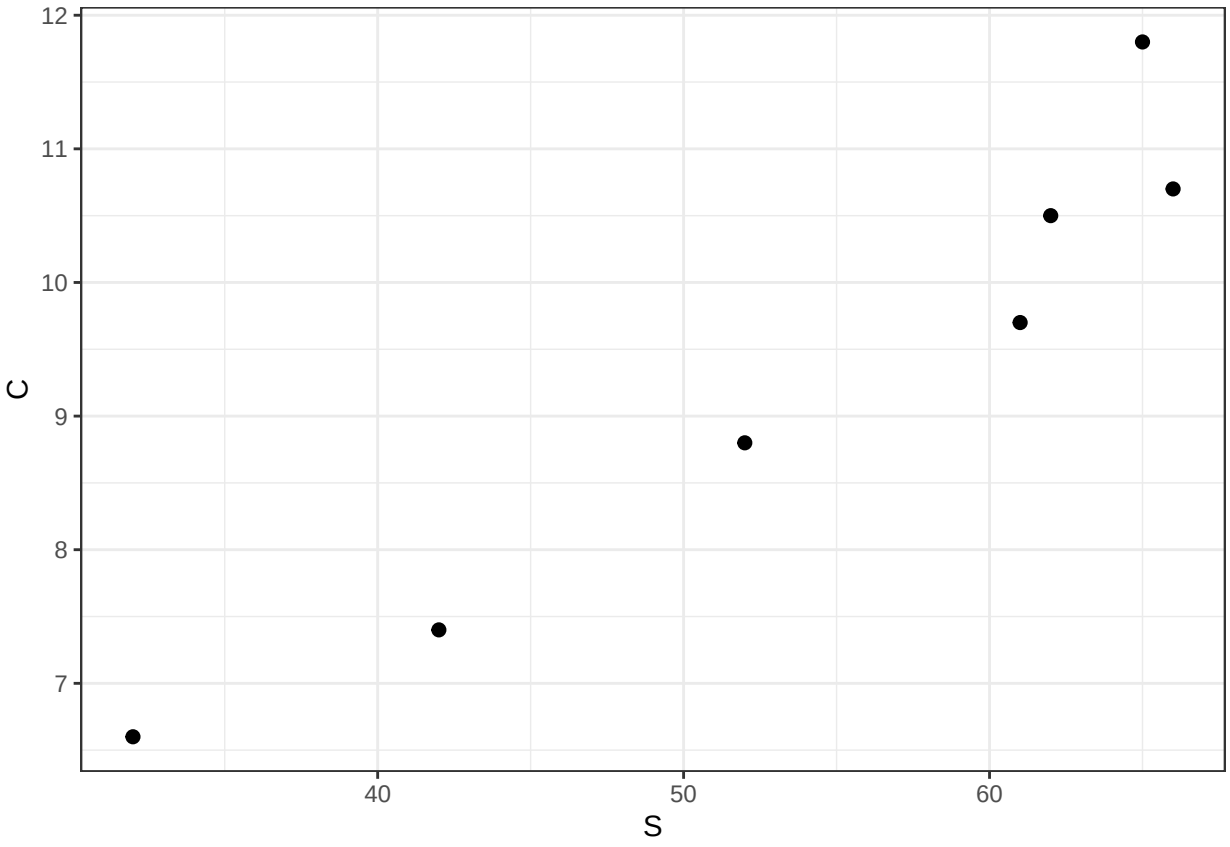
(a) Calculate the sample mean and sample variance of both S and C.

```
valiumdata %>%
  summarise(across(c(S,C), mean, .names = "{.col}.smean"),
            across(c(S,C), var, .names = "{.col}.svar"))
```

```
##    S.smean  C.smean   S.svar   C.svar
## 1 54.28571 9.357143 168.2381 3.489524
```

(b) Draw (or produce on a computer) a scatter-plot relating S and C.

```
ggplot(data=valiumdata, aes(x=S, y=C)) +
  geom_point(size=2) +
  theme_bw()
```



(c) Calculate the correlation between S and C.

Solution: The correlation between S and C is 0.9615376.

```
cor(valiumdata$S, valiumdata$C)
```

```
## [1] 0.9615376
```

(d) Using your judgment, comment on whether you think the association between S and C seems genuine in the sense that if you observed S and C on another 7 people it would likely yield a correlation in the same direction, or whether you think the estimated correlation based on the 7 different people might plausibly have a sign (i.e., positive or negative) in the opposite direction as the sign of correlation you calculated in part (c).

Solution: I believe that the association between S and C seems genuine since the correlation is extremely close to 1, indicating a near perfect positive correlation. If data was observed for S and C on another 7 individuals, it would likely yield a correlation in the same direction of being positively correlated.

4. Consider further the data from Problem 3.

   Sedation score S Cortisol C, μg/dl 32 6.6 42 7.4 52 8.8 61 9.7 62 10.5 65 11.8 66 10.7

(a) For the regression of S on C, calculate the least-squares estimate of the intercept, slope, and residual variance.

Solution: The least-squares estimates of the intercept and slope, respectively, are $\beta_0 = 1.8397$ and $\beta_1 = 0.1385$. The residual variance is the square of the residual standard error, which is $\sigma^2 = 0.3159$

```
lm <- lm(C ~ S, data=valiumdata)
summary(lm)
```

```
##
## Call:
## lm(formula = C ~ S, data = valiumdata)
##
## Residuals:
##        1        2        3        4        5        6        7
##  0.32898 -0.25582 -0.24062 -0.58694  0.07458  0.95914 -0.27934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.83965    0.98358   1.870 0.120358
## S            0.13848    0.01769   7.828 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5621 on 5 degrees of freedom
## Multiple R-squared:  0.9246, Adjusted R-squared:  0.9095
## F-statistic: 61.27 on 1 and 5 DF,  p-value: 0.0005458
```

```
(residual_sigma <-summary(lm)$sigma)
```

```
## [1] 0.5620695
```

```
(residual_var<- residual_sigma^2)
```

```
## [1] 0.3159222
```

(b) Calculate the standard error of the slope and the standard error of the intercept, and comment on whether each of these parameters is significantly different from 0.

Solution: The standard error of the intercept is 0.9836, and the standard error of the slope is 0.01769. The intercept for the regression of S on C is not significantly different from 0 at the $\alpha = 0.05$ level ($p = .12$). The slope for the regression of S on C is significantly different from 0 at the $\alpha = 0.05$ level ($p < .001$).

```
summary(lm)
```

```
##
## Call:
## lm(formula = C ~ S, data = valiumdata)
##
## Residuals:
##        1        2        3        4        5        6        7
##  0.32898 -0.25582 -0.24062 -0.58694  0.07458  0.95914 -0.27934
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.83965    0.98358   1.870 0.120358
## S            0.13848    0.01769   7.828 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5621 on 5 degrees of freedom
## Multiple R-squared:  0.9246, Adjusted R-squared:  0.9095
## F-statistic: 61.27 on 1 and 5 DF,  p-value: 0.0005458
```

(c) Comment on what the findings from (a) and (b) tell you about whether the correlation is significantly different from 0.

Solution: The residual variance from part a), also known as the unexplained variance, is much smaller than the sample standard deviation of C (by a factor of ~10), indicating that the model will have a close fit of the estimate to the actual data. The standard error of the slope in part b) has a significance at the $\alpha = 0.05$ level, indicating that the slope for the regression of S on C is significantly different from 0 and that there is a positive relationship for every 1-unit change in the predictor S. However, the findings from b) alone cannot define the strength of the linear relationship. Combining the significance of the slope and the small residual variance observed in part a), we may be able to conclude that the correlation is significantly different from 0.