RMHI/ARMP Problem Set 2

Your ID goes here [Word Count: XX]

Please put your answers here, following the instructions in the assignment description. Put your answers and word count tallies in the locations indicated; if none is indicated that means there is no word count for that question. Remember to knit as you go, and submit the knitted version of this on Canvas.

## Q1

**Q1a**

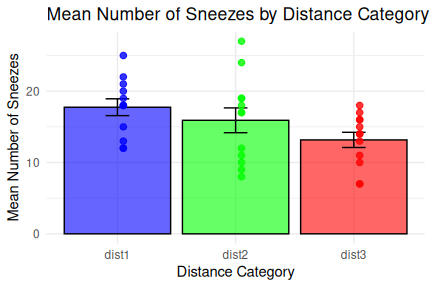
f\_five\_data=head(ds)  
cor.test(ds$distance,ds$sneezes)

##   
## Pearson's product-moment correlation  
##   
## data: ds$distance and ds$sneezes  
## t = -2.7082, df = 34, p-value = 0.01051  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.6586198 -0.1075823  
## sample estimates:  
## cor   
## -0.4212298

*ANSWER:* A Pearson’s product-moment correlation test was conducted to assess the null hypothesis that there the relationship between distance from the plant and sneezes is not significant against the alternative hypothesis that relationship is significant, as both variables were continuous. The test showed a p-value less than 0.05 suggesting leading to a rejection of null hypothesis and concluding there is significant relationship. The significant negative correlation indicated that as distance increased, sneezes decreased, suggesting the plant likely caused Foxy’s sneezing. [Word Count: 53]

**Q1b**

summary\_df <- ds %>%  
 group\_by(distCat) %>%  
 summarise(  
 mean\_sneezes = mean(sneezes),  
 se = sd(sneezes) / sqrt(n())  
 )  
  
colors <- c("blue", "green", "red")  
names(colors) <- levels(ds$distCat)  
  
ggplot(summary\_df, aes(x = distCat, y = mean\_sneezes, fill = distCat)) +  
 geom\_bar(stat = "identity", color = "black", alpha = 0.6) +  
 geom\_errorbar(aes(ymin = mean\_sneezes - se, ymax = mean\_sneezes + se),  
 width = 0.2) +  
 geom\_point(data = ds,   
 aes(x = distCat, y = sneezes, color = distCat),   
 position = position\_nudge(x = 0.05),   
 alpha = 0.8, size = 2, show.legend = FALSE) +  
 scale\_fill\_manual(values = colors) +  
 scale\_color\_manual(values = colors) +  
 theme\_minimal() +  
 labs(  
 title = "Mean Number of Sneezes by Distance Category",  
 x = "Distance Category",  
 y = "Mean Number of Sneezes"  
 ) +  
 theme(legend.position = "none")



The bar plot shows a decreasing trend in the mean number of sneezes across increasing distance. When closer to the plant (dist1) sneeze were more on average than those farther away (dist3). The error bars, representing standard error, do not overlap substantially between dist1 and dist3, suggesting the difference may be meaningful

**Q1c**

library(car)  
  
leveneTest(sneezes ~ distCat, data = ds)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 1.5643 0.2244  
## 33

model <- aov(sneezes ~ distCat, data = ds)  
shapiro.test(residuals(model))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(model)  
## W = 0.95864, p-value = 0.1952

*Note that the order of these assumptions (i.e., what assumption you put in 1 vs 2) does not matter! The words ASSUMPTION 1 and ASSUMPTION 2 do not contribute to word count.*

*ASSUMPTION 1:* Levene’s test evaluated homogeneity of variances, testing the null hypothesis of equal variances across groups. The test showed a p-value of 0.224. The p-value was greater than 0.05 and thus the null hypothesis was not rejected, concluding there wasz equal variances. [Word Count: 37]

*ASSUMPTION 2:* The Shapiro-Wilk test assessed normality of residuals. It tested the null hypothesis that residuals were normally distributed. The test showed W = 0.95864, p-value = 0.1952. Since the p-value exceeded 0.05, the null hypothesis was not rejected, indicating residuals were normally distributed. [Word Count: 33]

**Q1d**

anova\_model <- aov(sneezes ~ distCat, data = ds)  
summary(anova\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## distCat 2 127.7 63.86 2.876 0.0706 .  
## Residuals 33 732.8 22.21   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*ANSWER:* A one-way ANOVA test was carried to evaluate differences in the number of sneezes across the three categories of distances, since the assumptions of normality and homogeneity of variances were not violated in the hypothesis testscarried out in Q1c. The predictor was distance category, and the outcome was the number of sneezes. The test showed F(2, 33) = 2.876, p-value = 0.0706. The p-value being greater than 0.05 indicated no significant difference in sneezes across categories. The effect size, eta-squared, was calculated as 0.148, suggesting a moderate effect of distance on sneezing. [Word Count: 72]

**Q1e**

TukeyHSD(anova\_model)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = sneezes ~ distCat, data = ds)  
##   
## $distCat  
## diff lwr upr p adj  
## dist2-dist1 -1.833333 -6.554053 2.8873863 0.6112323  
## dist3-dist1 -4.583333 -9.304053 0.1373863 0.0584796  
## dist3-dist2 -2.750000 -7.470720 1.9707196 0.3377931

*ANSWER:* Post-hoc tests using Tukey’s HSD were conducted to compare mean sneezes across distance categories. The results showed p-values of 0.6112 (dist2 vs. dist1), 0.0585 (dist3 vs. dist1), and 0.3378 (dist3 vs. dist2). No differences were significant, as all p-values exceeded 0.05. [Word count: 34]

**Q1f**

*ANSWER:* The analysis in Q1a with significant correlation and anova test in Q1c-e (no significant post-hoc differences) lead to different conclusions. Q1a suggests a significant relationship, while Q1c-e does not confirm significant mean differences. The ANOVA is more appropriate, as it directly tests group means, whereas correlation assesses a linear trend, leading to a better study of the research question of category differences. [Word count: 71]

## Q2

**Q2a**

dl\_data\_five<-head(dl)  
shapiro.test(dl$lfb)

##   
## Shapiro-Wilk normality test  
##   
## data: dl$lfb  
## W = 0.98304, p-value = 0.5695

*ANSWER:* The assumption tested was normality of the lie detector scores for LFB. The null hypothesis that the data follows a normal distribution vs alternative hypothesis that the data does not follow a normal distribution was conducted using Shapiro-Wilk test on the variable lfb with 0.05 significance level. The test showed a p-value was greater than 0.05 hence the null hypothesis was not rejected.The p-value (0.5695) being greater than 0.05 indicated the scores were approximately normally distributed. [Word count: 71]

**Q2b**

t.test(dl$lfb,mu=50)

##   
## One Sample t-test  
##   
## data: dl$lfb  
## t = 2.2329, df = 59, p-value = 0.02936  
## alternative hypothesis: true mean is not equal to 50  
## 95 percent confidence interval:  
## 50.62827 61.47173  
## sample estimates:  
## mean of x   
## 56.05

mean(dl$lfb)

## [1] 56.05

rank\_biserial(dl$lfb, mu = 50)

## r (rank biserial) | 95% CI  
## --------------------------------  
## 0.32 | [0.04, 0.56]  
##   
## - Deviation from a difference of 50.

*ANSWER:* The null hypothesis that the lie was by chance against the alternative that the lie was significantly different from the mean was conducted using a one-sample t-test. The hypothesis was conducted to compare LFB’s lie detector scores to a chance level of 50 and 0.05 signifincance level, since the normality assumption was not violated in Q2a. The mean score was 56.05. The test showed a p-value less than 0.05, leading to rejection of null hypothesis and concluding a significant difference. The rank-biserial effect size was 0.32, suggesting a moderate effect. LFB’s scores were significantly higher than chance, implying greater honesty. [Word count: 67]

**Q2c**

clean\_data <- drop\_na(dl)  
shapiro.test(clean\_data$lfb - clean\_data$rainbow)

##   
## Shapiro-Wilk normality test  
##   
## data: clean\_data$lfb - clean\_data$rainbow  
## W = 0.95978, p-value = 0.1868

*ANSWER:* The assumption tested was that the difference between LFB and Rainbowz’s lie detector scores followed a normal distribution, against the alternative that they did not. The Shapiro-Wilk test was conducted on the difference scores (lfb - rainbow), using a significance level of 0.05. The test showed a p-value of 0.1868, which exceeds the significance value of 0.05, thus, the null hypothesis was not rejected, indicating that the assumption of normality was not violated and the differences were approximately normally distributed. [Word count: 41]

**Q2d**

t.test(clean\_data$lfb,clean\_data$rainbow)

##   
## Welch Two Sample t-test  
##   
## data: clean\_data$lfb and clean\_data$rainbow  
## t = -0.25495, df = 72.162, p-value = 0.7995  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -11.603624 8.972045  
## sample estimates:  
## mean of x mean of y   
## 56.07895 57.39474

mean(clean\_data$lfb)

## [1] 56.07895

mean(clean\_data$rainbow)

## [1] 57.39474

*ANSWER:* A Welch two-sample t-test was conducted to compare LFB and Rainbow’s lie detector scores, as the normality assumption was not violated in Q2c. The test evaluated the null hypothesis that the mean scores were equal versus the alternative hypothesis that they differed, using a significance level of 0.05. The mean for LFB was 56.07895, and for Rainbow, 57.39474. The test result showed a p-value greater than 0.05, thus the null hypothesis was not rejected. This led to the conclusion that there was no significant difference, suggesting the lie detector rated both similarly and that any lie was by chance. [Word count: 81]

## Q3

**Q3a**

head(dp)

## # A tibble: 6 × 3  
## person then now   
## <fct> <fct> <fct>   
## 1 person1 negative positive  
## 2 person10 negative positive  
## 3 person11 positive positive  
## 4 person12 positive negative  
## 5 person13 positive positive  
## 6 person14 positive positive

table\_dp <- table(dp$then, dp$now)  
table\_dp

##   
## positive negative  
## positive 21 23  
## negative 9 14

mcnemar.test(table\_dp)

##   
## McNemar's Chi-squared test with continuity correction  
##   
## data: table\_dp  
## McNemar's chi-squared = 5.2812, df = 1, p-value = 0.02156

*ANSWER:* The null hypothesis that the proportions of changes between the two years were equal against the alternative hypothesis that they differed significantly was conducted using a McNemar’s Chi-squared test using a significance level of 0.05. The table showed 21 stayed positive, 14 stayed negative, 23 shifted from positive to negative, and 9 from negative to positive. The test result showed a p-value of 0.02156, which was less than 0.05, thus the null hypothesis was rejected. The p-value being less than 0.05 indicated a significant change, suggesting people’s outlook worsened over time. [Word count: 88]

## Q4

**Q4a**

head(df)

## # A tibble: 6 × 3  
## month population amount  
## <dbl> <dbl> <dbl>  
## 1 1 999 40.7  
## 2 2 963 39.2  
## 3 3 1017 42.3  
## 4 4 994 39.0  
## 5 5 982 43.6  
## 6 6 1068 41.5

model <- lm(amount ~ month + population, data = df)  
summary(model)

##   
## Call:  
## lm(formula = amount ~ month + population, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5981 -1.1853 0.0528 1.7930 5.0668   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.905336 5.293817 2.249 0.0284 \*   
## month -0.235521 0.022598 -10.422 8.08e-15 \*\*\*  
## population 0.028949 0.005224 5.541 8.01e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.758 on 57 degrees of freedom  
## Multiple R-squared: 0.7988, Adjusted R-squared: 0.7917   
## F-statistic: 113.2 on 2 and 57 DF, p-value: < 2.2e-16

*ANSWER:* A multiple linear regression was conducted to evaluate if the outcome (food waste) was predicted by the two variables (month and population). The model showed p-value less than 0.001 with R-squared = 0.7988, showing a significant model explaining 79.88% of variance, only 20% of the variance was not explained by the model and was assumed to be random errors. The coefficient for month was -0.235521 (p = 8.08e-15), suggesting food waste decreased by 0.235521 units per month. The coefficient for population was 0.028949 (p = 8.01e-07), indicating food waste increased by 0.028949 units per person. This supported a decline in food waste over time suggesting less food availability. [Word count: 132]

**Q4b**

model\_interaction <- lm(amount ~ month \* population, data = df)  
summary(model\_interaction)

##   
## Call:  
## lm(formula = amount ~ month \* population, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.0943 -1.1785 -0.0276 1.6072 5.9282   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 38.2574242 9.0007921 4.250 8.15e-05 \*\*\*  
## month -1.1329585 0.2591702 -4.371 5.41e-05 \*\*\*  
## population 0.0015948 0.0092121 0.173 0.863181   
## month:population 0.0009498 0.0002734 3.474 0.000997 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.524 on 56 degrees of freedom  
## Multiple R-squared: 0.8345, Adjusted R-squared: 0.8256   
## F-statistic: 94.1 on 3 and 56 DF, p-value: < 2.2e-16

*ANSWER:* (i) A multiple linear regression with an interaction term was conducted. The month coefficient was -1.1329585, indicating food waste decreased by 1.1329585 units per month when population was zero. The interaction term (month:population) was 0.0009498, suggesting that as population increased, the rate of decrease in food waste per month slowed by 0.0009498 units per person. (ii) The R-squared was 0.8345, compared to 0.7988 in Q4a. The higher R-squared suggested the interaction model explained more variance, fitting the data better by capturing the combined effect of month and population [Word count: 106]

**Q4c**

model\_std <- lm.beta(model)  
model\_interaction\_std <- lm.beta(model\_interaction)  
summary(model\_std)

##   
## Call:  
## lm(formula = amount ~ month + population, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5981 -1.1853 0.0528 1.7930 5.0668   
##   
## Coefficients:  
## Estimate Standardized Std. Error t value Pr(>|t|)   
## (Intercept) 11.905336 NA 5.293817 2.249 0.0284 \*   
## month -0.235521 -0.680655 0.022598 -10.422 8.08e-15 \*\*\*  
## population 0.028949 0.361876 0.005224 5.541 8.01e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.758 on 57 degrees of freedom  
## Multiple R-squared: 0.7988, Adjusted R-squared: 0.7917   
## F-statistic: 113.2 on 2 and 57 DF, p-value: < 2.2e-16

summary(model\_interaction\_std)

##   
## Call:  
## lm(formula = amount ~ month \* population, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.0943 -1.1785 -0.0276 1.6072 5.9282   
##   
## Coefficients:  
## Estimate Standardized Std. Error t value Pr(>|t|)   
## (Intercept) 38.2574242 NA 9.0007921 4.250 8.15e-05 \*\*\*  
## month -1.1329585 -3.2742466 0.2591702 -4.371 5.41e-05 \*\*\*  
## population 0.0015948 0.0199359 0.0092121 0.173 0.863181   
## month:population 0.0009498 2.4784679 0.0002734 3.474 0.000997 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.524 on 56 degrees of freedom  
## Multiple R-squared: 0.8345, Adjusted R-squared: 0.8256   
## F-statistic: 94.1 on 3 and 56 DF, p-value: < 2.2e-16

*ANSWER:* The standardized coefficients for Q4a showed month as -0.680655 and population as 0.361876. In Q4b, month was -3.2742466, population was 0.0199359, and the interaction term was 2.4784679. The direction of month remained negative in both models, but population’s direction changed from positive to nearly neutral, and its magnitude decreased significantly. The interaction term’s coefficient in Q4b indicated that the effect of month on food waste depended on population size, with a positive standardized value suggesting a moderating effect. These models suggested that time consistently reduced food waste, but population’s role varied, with the interaction model revealing a more complex relationship between predictors and food waste. [Word count: 192]

**Q4d**

df$percapita <- df$amount / df$population  
model\_percapita <- lm(percapita ~ month, data = df)  
summary(model\_percapita)

##   
## Call:  
## lm(formula = percapita ~ month, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0118347 -0.0012461 0.0001381 0.0021610 0.0072626   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.131e-02 8.155e-04 50.66 < 2e-16 \*\*\*  
## month -2.457e-04 2.325e-05 -10.57 3.83e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.003119 on 58 degrees of freedom  
## Multiple R-squared: 0.6582, Adjusted R-squared: 0.6523   
## F-statistic: 111.7 on 1 and 58 DF, p-value: 3.834e-15

*ANSWER:* (i) One advantage of analyzing per capita food waste was simplifying the model by combining amount and population into one variable, reducing multicollinearity seen in Q4a-Q4b where population and month interacted. This provided a clearer view of food waste trends over time. (ii) One disadvantage was losing insight into the separate effects of population and amount. Q4b showed a significant interaction, indicating population moderated the effect of time, which per capita analysis obscured, potentially missing nuances in how population dynamics influenced food waste. [Word count: 113]

## Q5

**Q5a**

head(da)

## # A tibble: 6 × 2  
## year size   
## <fct> <fct>  
## 1 past small  
## 2 past small  
## 3 past small  
## 4 past small  
## 5 past small  
## 6 past small

size\_table <- table(da$year, da$size)  
size\_table

##   
## small medium large  
## past 13 35 54  
## present 17 35 28

test\_result <- chisq.test(size\_table)  
#cramersV(size\_table)  
test\_result

##   
## Pearson's Chi-squared test  
##   
## data: size\_table  
## X-squared = 6.2086, df = 2, p-value = 0.04486

*ANSWER:* A Chi-squared test was conducted to compare the distribution of animal sizes (small, medium, large) between past and present. The test conducted the null hypothesis that the distributions were the same against the alternative hypothesis that they differed, using a significance level of 0.05. The past counts for past and present in both small, medium and large are as shown. The test result showed a p-value of 0.04486, which was less than 0.05, this led to the rejection of the null hypothesis. The significant p-value suggested the distribution of animal sizes changed over time [Word count: 76]

**Q5b**

test\_result$stdres

##   
## small medium large  
## past -1.534856 -1.298743 2.414486  
## present 1.534856 1.298743 -2.414486

*ANSWER:* The adjusted residuals from the Chi-squared test were calculated. They showed values for past and present in small , medium and large as above. The large size residuals exceeded ±1.96, indicating a significant difference. This suggested the proportion of large animals decreased significantly from past to present, while small and medium sizes did not differ significantly. [Word count: 56]

## Q6

**Q6a**

head(dh)

## # A tibble: 6 × 4  
## person height income health  
## <fct> <fct> <fct> <dbl>  
## 1 person1 short rich 66  
## 2 person10 tall poor 61  
## 3 person100 short rich 73  
## 4 person101 tall poor 64  
## 5 person102 short rich 46  
## 6 person103 tall poor 44

table(dh$height)

##   
## short tall   
## 80 80

table(dh$income)

##   
## poor rich   
## 80 80

model\_dh <- aov(health ~ income \* height, data = dh)  
summary(model\_dh)

## Df Sum Sq Mean Sq F value Pr(>F)   
## income 1 1066 1066.1 5.289 0.02278 \*   
## height 1 1507 1506.8 7.476 0.00698 \*\*  
## income:height 1 294 294.3 1.460 0.22873   
## Residuals 156 31442 201.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*ANSWER:* A two-way ANOVA was conducted to evaluate the null hypothesis that income and height had no effect on health ratings against the alternative that suggested that they did with a significance level of 0.05, including their interaction. The predictors were income and height, with health as the outcome. The results showed income had a p-value of 0.02278, and height had a p-value of 0.00698, both significant. The interaction was not significant since it had a p-value of 0.22873. This indicated that both income and height independently influenced health, with richer or taller individuals both contributing to better health outcomes. [Word count: 101]

**Q6b**

eta\_squared(model)

## # Effect Size for ANOVA (Type I)  
##   
## Parameter | Eta2 (partial) | 95% CI  
## ------------------------------------------  
## month | 0.77 | [0.69, 1.00]  
## population | 0.35 | [0.19, 1.00]  
##   
## - One-sided CIs: upper bound fixed at [1.00].

*ANSWER:* The eta-squared values for the predictors were calculated. Income had an eta-squared of 0.032, and height had 0.046. Height accounted for the most variance in health, as its eta-squared (0.046) was higher than income’s (0.032), indicating a stronger effect on health ratings. [Word count: 27]

**Q6c**

*ANSWER:* The results suggested that in Otherland, health was influenced by both income and height. Taller and richer individuals likely had better health, as both factors were significant predictors. The lack of interaction indicated their effects were independent. It’s possible taller individuals had better access to resources, or height reflected better nutrition historically. Similarly, richer individuals might afford better healthcare. This suggested socio-economic and physical factors played a role in health disparities, potentially exacerbated by food shortages. [Word count: 99]

## Q7

**Q7a**

*ANSWER:* i) Option B is not possible.

1. Option A is not possible.
2. Option A is not possible.

**Q7b**

*ANSWER:* (i) Option B (χ²(6) = 12.22, p = .431) is not possible because a chi-squared test with 6 degrees of freedom should yield a larger test statistic for a p-value of .431, which is not significant. A value of 12.22 suggests a significant result, inconsistent with the high p-value. ii) Option A (t(23) = 0.162, p = .003) is not possible because a t-test with 23 degrees of freedom and a test statistic of 0.162 is too small to produce a p-value of .003, which indicates significance. A larger t-value is needed for such a low p-value. (iii) Option A (F(3,54) = -13.2, p < .001) is not possible because an F-statistic cannot be negative, as it represents a ratio of variances, which must be positive regardless of degrees of freedom (3, 54) or p-value significance. [Word Count: 142]

## Q8

*ANSWER:* My favourite character was Super Size — he brings calm, diplomacy, and perspective to the group. While others get defensive, he steps in to refocus everyone on collaboration. In a data-driven debate having someone like Super Size is exactly what a team needs.