HA2

Xinrui Hu

2023-02-17

1.1 Show descriptive statistics for each of the variables What is the mean subjective taste score?

```
library(faraway)
summary(cheddar)
```

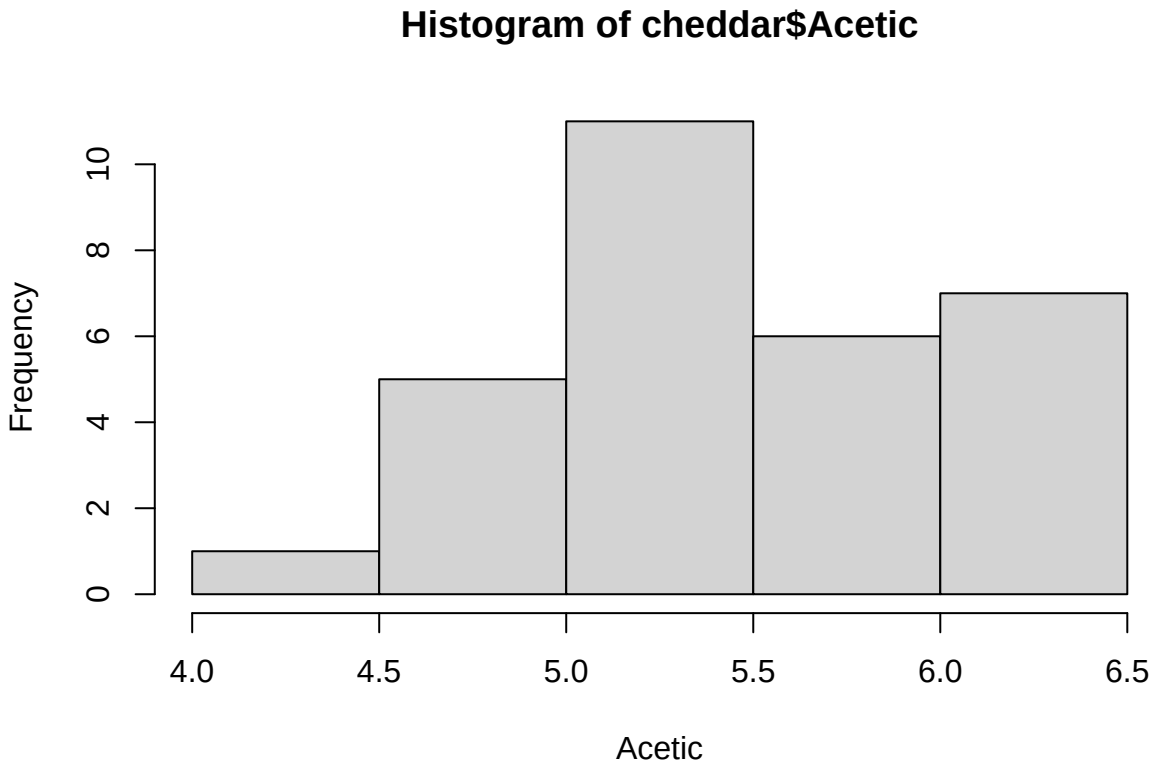
```
##      taste           Acetic           H2S              Lactic      
##  Min.   : 0.70   Min.   :4.477   Min.   : 2.996   Min.   :0.860  
##  1st Qu.:13.55   1st Qu.:5.237   1st Qu.: 3.978   1st Qu.:1.250  
##  Median :20.95   Median :5.425   Median : 5.329   Median :1.450  
##  Mean   :24.53   Mean   :5.498   Mean   : 5.942   Mean   :1.442  
##  3rd Qu.:36.70   3rd Qu.:5.883   3rd Qu.: 7.575   3rd Qu.:1.667  
##  Max.   :57.20   Max.   :6.458   Max.   :10.199   Max.   :2.010
```

```
# taste mean is 24.53
```

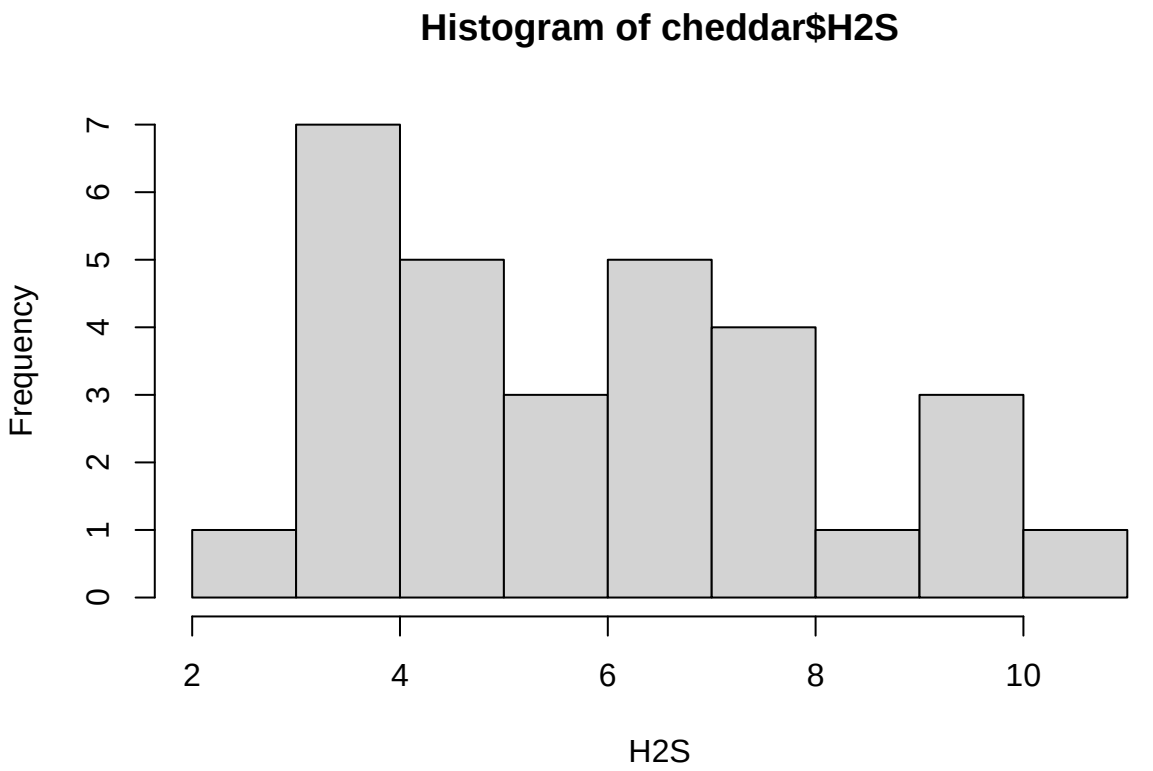
1.2 Show a histogram for each of the variables. Make sure to label the x-axis

```
hist(cheddar$taste, xlab = "Taste")
```

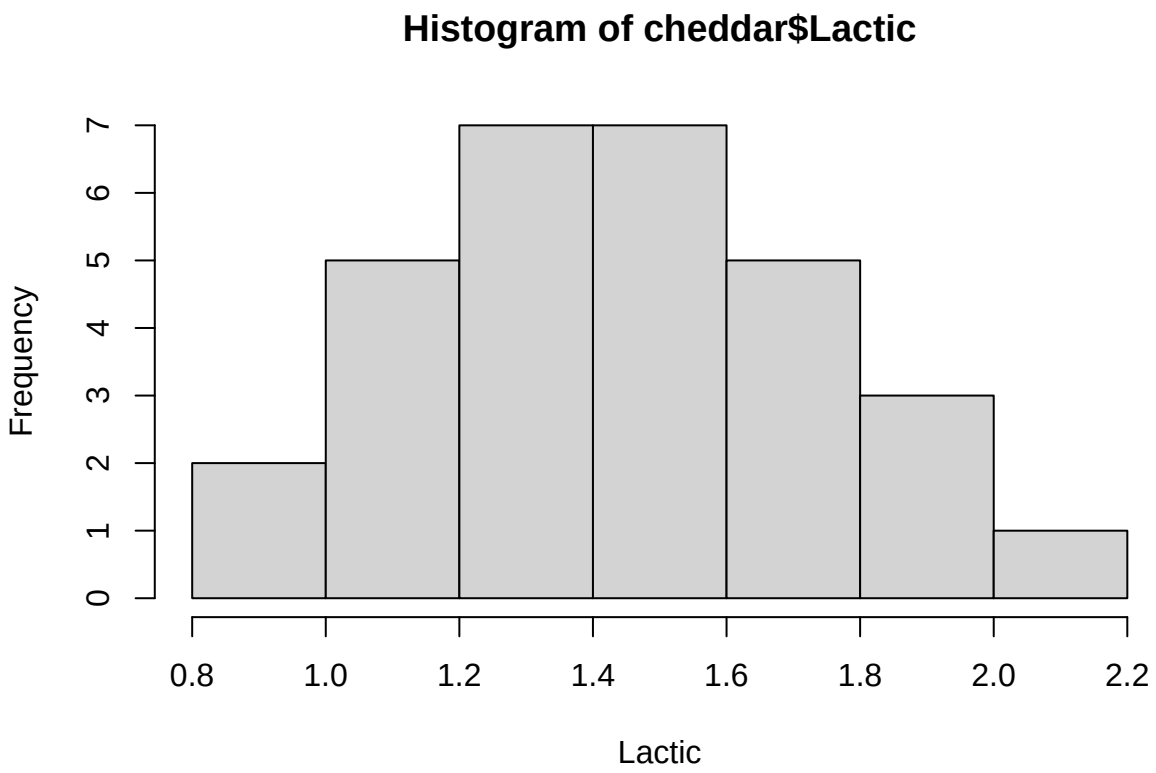
```r
hist(cheddar$Acetic, xlab = "Acetic")
```




```r
hist(cheddar$H2S, xlab = "H2S")
```



```
hist(cheddar$Lactic, xlab = "Lactic")
```



1.3 Fit a regression model with taste as the response and no predictors. What is the value of the intercept? What does it represent?

```
Tnull <- lm(cheddar$taste ~ 1)
Tnull
```

```
##
## Call:
## lm(formula = cheddar$taste ~ 1)
##
## Coefficients:
## (Intercept)
##       24.53
```

```
# It represents the value of the mean of the taste
```

1.4 Fit a regression model with taste as the response and Lactic as the only predictor.

```
summary(lm(taste ~ Lactic, cheddar))
```

```
##
## Call:
## lm(formula = taste ~ Lactic, data = cheddar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.9439  -8.6839  -0.1095   8.9998  27.4245
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -29.859     10.582  -2.822  0.00869 **
## Lactic        37.720      7.186   5.249 1.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.75 on 28 degrees of freedom
## Multiple R-squared:  0.4959, Adjusted R-squared:  0.4779
## F-statistic: 27.55 on 1 and 28 DF,  p-value: 1.405e-05
```

```
# It is statistically significant at 5% level
```

1.5 Calculate the p-value of the model you created in Question 1.4 using the anova function.

```
anova(lm(taste ~ Lactic, cheddar))
```

```
## Analysis of Variance Table
##
## Response: taste
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Lactic     1 3800.4  3800.4   27.55 1.405e-05 ***
## Residuals 28 3862.5   137.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.6 Fit a regression model with taste as the response and the three chemical contents as predictors. Identify the predictors that are statistically significant at the 5% level.

```
mall <- lm(taste ~ Lactic + H2S + Acetic, cheddar)
summary(mall)
```

```
##
## Call:
## lm(formula = taste ~ Lactic + H2S + Acetic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Lactic       19.6705     8.6291   2.280  0.03108 *
## H2S           3.9118     1.2484   3.133  0.00425 **
## Acetic        0.3277     4.4598   0.073  0.94198
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

```
# It is statistically significant at 5% level
```

1.7 Use the anova function to recalculate the significance of the H2S variable as shown in the output of Question 1.6:

```
model1 <- lm(taste ~ Lactic + Acetic, cheddar)
anova(model1,mall)
```

```
## Analysis of Variance Table
##
## Model 1: taste ~ Lactic + Acetic
## Model 2: taste ~ Lactic + H2S + Acetic
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     27 3676.1
## 2     26 2668.4  1    1007.7 9.8182 0.004247 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p value is smaller than 5% so we can not frop the variable H2S
```

1.8 Test the hypothesis that the coefficients of Acetic and Lactic both equal 0 when H2S is included in the model. Should we reject this hypothesis?

```
model2 <- lm(taste ~ H2S, cheddar)
anova(model2, mall)
```

```
## Analysis of Variance Table
##
## Model 1: taste ~ H2S
## Model 2: taste ~ Lactic + H2S + Acetic
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 3286.1
## 2     26 2668.4  2    617.73 3.0095 0.06674 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Fail to reject the null hypothesis since the p-value is greater than 5%
```

2.1 Convert sex into a factor and label and levels (male and female).

```
teengamb$sex <- factor(teengamb$sex,
                       levels = c(0,1),
                       labels = c("male", "female"))
head(teengamb)
```

```
##      sex status income verbal gamble
## 1 female     51   2.00      8    0.0
## 2 female     28   2.50      8    0.0
## 3 female     37   2.00      6    0.0
## 4 female     28   7.00      4    7.3
## 5 female     65   2.00      8   19.6
## 6 female     61   3.47      6    0.1
```

2.2 Fit a model with gamble as the response and income, verbal and sex as predictors. Which variables are statistically significant at the 5% level?

```
gmodel <- lm(gamble ~ income + verbal + sex, data = teengamb)
summary(gmodel)
```

```
##
## Call:
## lm(formula = gamble ~ income + verbal + sex, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.639 -11.765  -1.594   9.305  93.867
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1390    14.7686   1.634   0.1095
## income        4.8981     0.9551   5.128 6.64e-06 ***
## verbal       -2.7468     1.8253  -1.505   0.1397
## sexfemale   -22.9602     6.7706  -3.391   0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.43 on 43 degrees of freedom
## Multiple R-squared:  0.5263, Adjusted R-squared:  0.4933
## F-statistic: 15.93 on 3 and 43 DF,  p-value: 4.148e-07
```

```
# Income and sex are significant at the 5% level
# On average, the aggregate responses of female is 22.96 lower compared to male when we control for inc
```

2.3 Use the confint function to produce 95% confidence intervals for the coefficients based on the same model. Can you deduce which coefficients are significant at the level of 5% based on the intervals?

```
confint(gmodel, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept)  -5.644725 53.9226685
## income        2.971911  6.8242686
## verbal       -6.427846  0.9342123
## sexfemale   -36.614385 -9.3060548
```

```
#  If the interval does not include zero, we can conclude that the corresponding coefficient is statist
```