Project 4

AUTHOR
Raymond Fleming

```
  library(tidyverse)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2
──
✓ ggplot2 3.3.6     ✓ purrr   0.3.4
✓ tibble  3.1.8     ✓ dplyr   1.0.9
✓ tidyr   1.2.0     ✓ stringr 1.4.1
✓ readr   2.1.2     ✓ forcats 0.5.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```
  library(gsheet)
  library(ggplot2)
  library(DescTools)
  library(lindia)
  library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:DescTools':

    Recode

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
library(ggpubr)
library(gmodels)
```

```
Warning: package 'gmodels' was built under R version 4.2.2
```

```
Registered S3 method overwritten by 'gdata':
  method         from
  reorder.factor DescTools
```

Consider the Palmer penguin data, available through the `palmerpenguins` package in R. Note that if you are using R/RStudio on your own computer, you may need to install this package.

```
data <- palmerpenguins :: penguins
```

For all questions, assume $\alpha = 0.05$.

```
data <- na.omit(data)

data %>%

  summarize(

    NumObs = nrow (data),

    NumFemale = length (which (data$sex == "female")),

    NumMale = length (which (data$sex == "male")),

    PercentFemale = NumFemale / NumObs * 100,

    PercentMale = NumMale / NumObs * 100

  )
```

```
# A tibble: 1 × 5
  NumObs NumFemale NumMale PercentFemale PercentMale
   <int>     <int>   <int>         <dbl>       <dbl>
1    333       165     168          49.5        50.5
```

1. Consider the sex (*sex*) of the penguins.

1a. It is known that the split of biological sex in humans is 50/50. Perform the appropriate hypothesis test to determine if the Palmer penguins have the same split.

```
prop.test( x = length( which (data$sex == "female")),
           n = nrow(data),
           p = 0.5,
           alternative = "two.sided",
           correct = FALSE)
```

```
    1-sample proportions test without continuity correction

data:  length(which(data$sex == "female")) out of nrow(data), null probability
0.5
X-squared = 0.027027, df = 1, p-value = 0.8694
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4421534 0.5489403
sample estimates:
        p
0.4954955
```

Hypotheses

$H_0 : \pi_{Female} = 0.5$

$H_1 : \pi_{Female} \neq 0.5$

Test Statistic

$\chi^2 = 0.027$ or $z_0 = 0.164$

p-Value

$p = 0.869$

Conclusion

Fail to reject H_0 at $\alpha = .05$. There is not sufficient evidence to suggest that the proportion of female penguins is not equal to 0.5. Because for this study sex is binary, this implies that there is also not sufficient evidence to suggest that the proportion of male penguins is not 0.5.

1b. Find the 95% CI for the proportion of female penguins.

The 95% confidence interval for the proportion of female penguins is (0.442, 0.549)

2. Consider the species (*species*) of the penguins.

2a. Use the appropriate hypothesis test to determine if there is an even split of species in the dataset.

```
NumAdelie = length( which( data$species == "Adelie"))

NumGentoo = length( which( data$species == "Gentoo"))

NumChinstrap = length( which( data$species == "Chinstrap"))

counts <- c( NumAdelie, NumGentoo, NumChinstrap)

probs <- c( 1/3, 1/3, 1/3)

chisq.test( counts, p = probs)
```

```
    Chi-squared test for given probabilities

data:  counts
X-squared = 28.27, df = 2, p-value = 7.264e-07
```

Hypotheses

H_0 : There is an even split of penguin species in the dataset.

H_1 : There is not an even split of penguin species in the dataset.

Test Statistic

$\chi_0^2 = 28.27$

p-Value

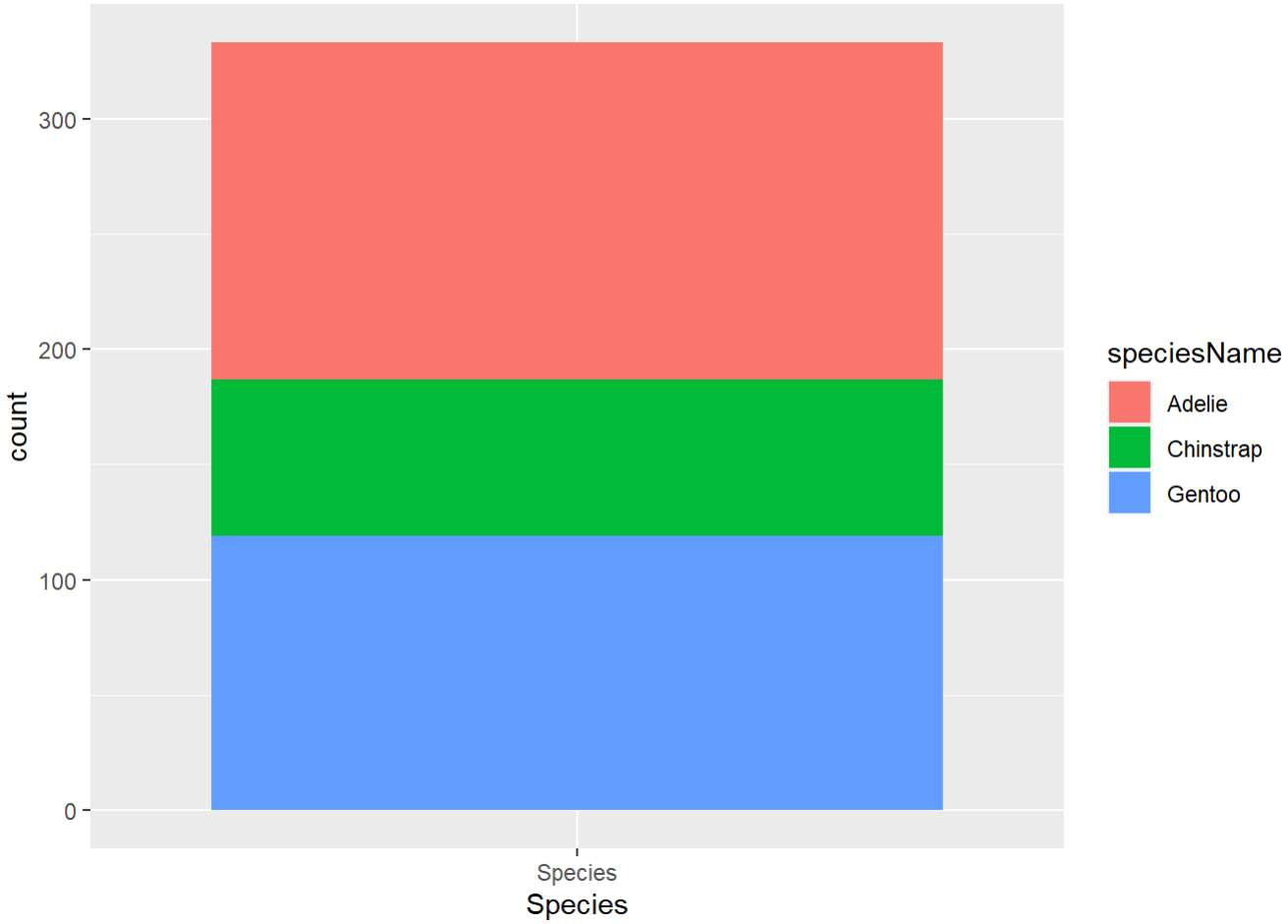
$p < 0.001$

Conclusion

Reject H_0 at $\alpha = .05$. There is sufficient evidence to suggest that there is not an even split of penguin species in the dataset.

2b. Construct a graph to accompany the test performed in 2a.

```
Species = c( "Species",
             "Species",
             "Species")

speciesName = c( "Adelie",
                 "Gentoo",
                 "Chinstrap")

count = c( NumAdelie,
           NumGentoo,
           NumChinstrap)

dataspecies = tibble( "Species" = Species,
                      "Species Color" = speciesName,
                      "count" = count)

ggplot( data = dataspecies, aes( x = Species,

                                 y = count,

                                 fill = speciesName)) +

  geom_bar( stat = "identity")
```



3. Consider sex and species.

3a. Use the appropriate hypothesis test to determine if sex depends on species.

```
Adelie <- data %>% filter( species == "Adelie" )

Gentoo <- data %>% filter( species == "Gentoo" )

Chinstrap <- data %>% filter( species == "Chinstrap" )

NumMaleAdelie = length( which( Adelie$sex == "male" ))

NumMaleGentoo = length( which( Gentoo$sex == "male" ))

NumMaleChinstrap = length( which( Chinstrap$sex == "male" ))

NumFemaleAdelie = length( which( Adelie$sex == "female" ))

NumFemaleGentoo = length( which( Gentoo$sex == "female" ))
```

```r
NumFemaleChinstrap = length( which( Chinstrap$sex == "female" ))

speciessex <- matrix( c( NumMaleAdelie, NumMaleGentoo, NumMaleChinstrap,
                         NumFemaleAdelie, NumFemaleGentoo, NumFemaleChinstrap ),
                         nrow =2,
                         ncol =3,
                         byrow =T )

chisq.test(speciessex)
```

```
    Pearson's Chi-squared test

data:  speciessex
X-squared = 0.048607, df = 2, p-value = 0.976
```

Hypotheses

H_0 : Sex does not depend on species

H_1 : Sex depends on species

Test Statistic

$\chi_0^2 = 0.049$

p-Value

$p = 0.976$

Rejection Region

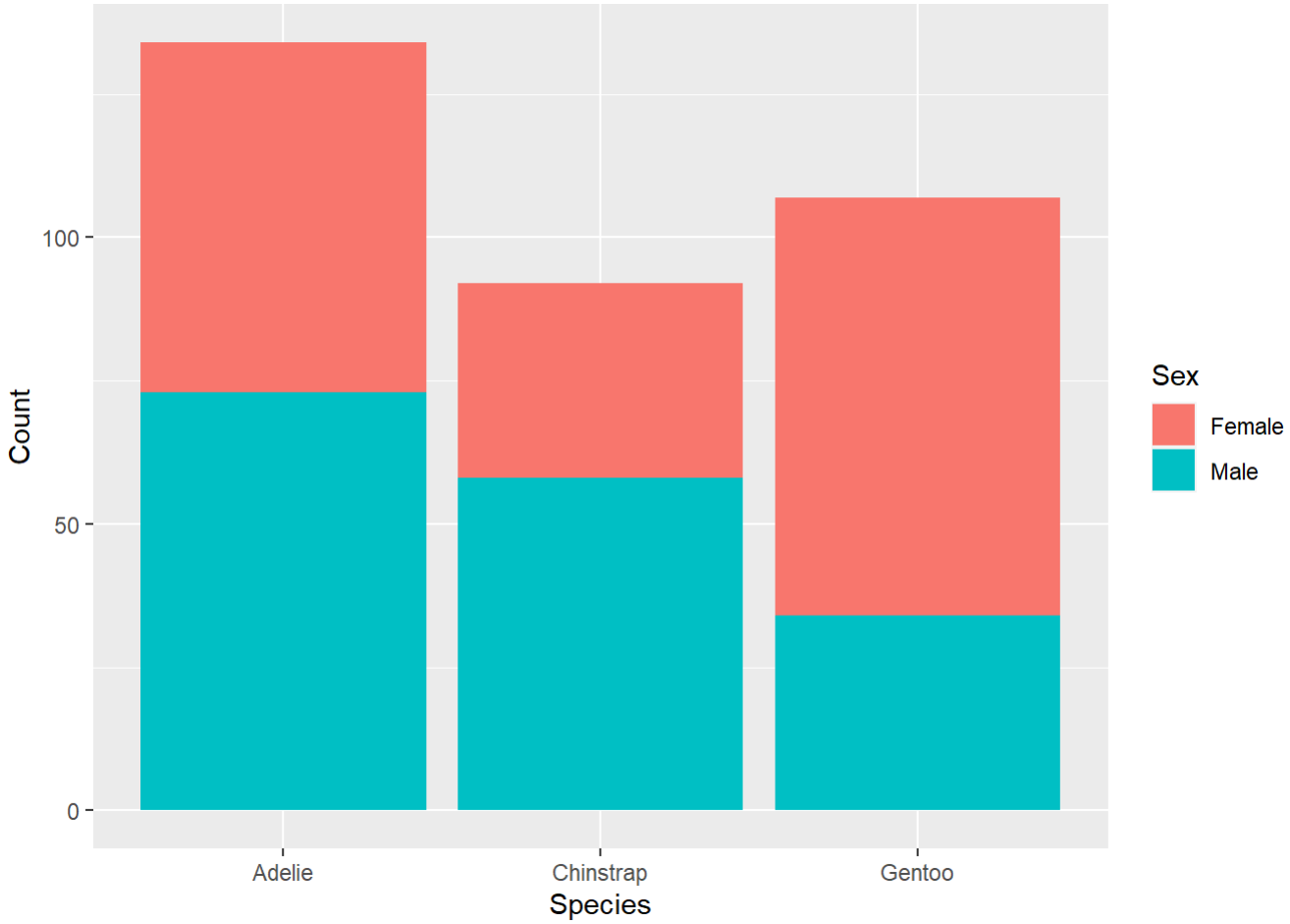
Reject H_0 if $p < \alpha$; $\alpha = 0.05$

Conclusion

Fail to reject H_0 at $\alpha = .05$. There is not sufficient evidence to suggest that sex depends on species.

3b. Construct a graph to accompany the test performed in 3a.

```
Species = c( "Adelie", "Adelie",
             "Gentoo", "Gentoo",
             "Chinstrap", "Chinstrap" )
Sex = c( "Male", "Female",
         "Male","Female",
         "Male", "Female" )
Count = c( NumMaleAdelie, NumMaleGentoo, NumMaleChinstrap,
           NumFemaleAdelie, NumFemaleGentoo, NumFemaleChinstrap )
PenguinsSexSpecies <- tibble( "Species" = Species,
                              "Sex" = Sex,
                              "Count" = Count )
ggplot( data = PenguinsSexSpecies, aes( x = Species,
                                        y = Count,
                                        fill = Sex ))+
  geom_bar( stat = "identity")
```



4. Suppose we want to predict the sex of the penguin.

4a. Construct the regression model that models sex as a function of bill length (*bill_length_mm*), body mass (*body_mass_g*), and flipper length (*flipper_length_mm*). Remember to state the resulting model.

```
m <- glm ( formula = sex ~ bill_length_mm +
                           body_mass_g +
                           flipper_length_mm,

       data = data,

       family = binomial )

summary( m )
```

```
Call:
glm(formula = sex ~ bill_length_mm + body_mass_g + flipper_length_mm,
```

```
    family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5691  -0.9023   0.1991   0.8757   2.3051

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        8.2569324  2.7636399   2.988  0.00281 **
bill_length_mm     0.1458945  0.0332279   4.391 1.13e-05 ***
body_mass_g        0.0029604  0.0004188   7.069 1.56e-12 ***
flipper_length_mm -0.1348614  0.0234043  -5.762 8.30e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 461.61  on 332  degrees of freedom
Residual deviance: 350.73  on 329  degrees of freedom
AIC: 358.73

Number of Fisher Scoring iterations: 4
```

The model for Sex based on Bill Length, Body Mass, and Flipper Length is

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 8.257 + 0.146 \text{ Bill Length} + 0.003 \text{ Body Mass} - 0.135 \text{ Flipper Length}$$

4b. Use the appropriate hypothesis tests to determine if there are any significant predictors of sex. You do not need to state all parts of the hypothesis tests, but you must justify your answer statistically.

At $\alpha = 0.05$ all predictors are significant predictors of sex. For Bill Length, Body Mass, and Flipper Length the p value is < 0.001 which is less than the 0.05 required to be statistically significant at this α level.

4c. Find the odds ratios for all of the predictors in the model.

```
round( exp( coefficients( m ) [-1] ), 3 )
```

```
   bill_length_mm       body_mass_g flipper_length_mm
            1.157             1.003             0.874
```

Predictor	Odds Ratio
Bill Length	**1.157**
Body Mass	**1.003**
Flipper Length	**0.874**

4d. Provide interpretations for each of the odds ratios.

For a 1mm increase in Bill Length, there is a 15.7% increase in the odds of being male

For a 1g increase in body mass, there is a 0.3% increase in the odds of being male

For a 1mm increase in flipper length, there is a 12.6% decrease in the odds of being male

4e. Construct the 95% confidence intervals for the odds ratios.

```
round( exp( confint( m )), 3)
```

```
Waiting for profiling to be done...
```

	2.5 %	97.5 %
(Intercept)	19.104	999243.586
bill_length_mm	1.086	1.237
body_mass_g	1.002	1.004
flipper_length_mm	0.833	0.913

Predictor	Confidence Interval
Bill Length	**(1.086, 1.237)**
Body Mass	**(1.002, 1.004)**
Flipper Length	**(0.833, 0.913)**

4f. Use the metrics discussed in the last lecture to determine if this is a good model.

```r
data <- data %>%

  mutate(p_hat = predict(m, type="response"))

data <- data %>%

  mutate(Predicted = ifelse(p_hat>0.5,1,0))

CrossTable(data$sex,data$Predicted,

           prop.r = FALSE,

           prop.c = FALSE,

           prop.t = FALSE,

           prop.chisq=FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|-------------------------|


Total Observations in Table:  333


             | data$Predicted
    data$sex |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
      female |       115 |        50 |       165 |
-------------|-----------|-----------|-----------|
        male |        45 |       123 |       168 |
-------------|-----------|-----------|-----------|
Column Total |       160 |       173 |       333 |
-------------|-----------|-----------|-----------|
```

True Positive: 123 male and predicted to be male.

True Negative: 115 female and predicted to be female.

False Positive: 50 female and predicted to be male.

False Negative: 45 male and predicted to be female.

The true positive rate(sensitivity) for this model is $\frac{truepositives}{truepositives+falsenegatives} = \frac{123}{123+45} = 0.732 = 73.2\%$

The true negative rate(specificity) for this model is $\frac{truenegatives}{truenegatives+falsepositives} = \frac{115}{115+50} = 0.697 = 69.7\%$

The positive predictive value for this model is $\frac{truepositives}{truepositives+falsepositives} = \frac{123}{123+50} = 0.711 = 71.1\%$

The negative predictive value for this model is $\frac{truenegatives}{truenegatives+falsenegatives} = \frac{115}{115+45} = 0.719 = 71.9\%$

The false discovery rate for this model is $\frac{falsepositives}{falsepositives+truepositives} = \frac{50}{50+123} = 0.289 = 28.9\%$

The true positive and true negatives rates for this model are balanced and reasonably high. The positive and negative predictive values are also similarly high and balanced, and the false discovery rate is reasonably low. This is a good model, and definitely a better model than the one created in the Logistic Regression activity.