Tarea del curso: The Analysis of Categorical and Count Data

Yan Carlos Leyva Labrador

11 de octubre de 2021

Practical Problems 1

In Costa Rica, the vampire bat Desmodus rotundus feeds on the blood of domestic cattle. If the bats respond to a hormonal signal, cows in estrous (in heat) may be bitten with a diferent probability than cows not in estrous. (The researcher could tell the diference by harnessing painted sponges to the undersides of bulls who would leave their mark during the night.)

```
##                       Estrous
## Bitten                 In estrous Not in estrous Total
##   Bitten by a bat              15              6    21
##   Not bitten by a bat           7            322   329
##   Total                        22            328   350
```

Answer:

The relative risk $\widehat{\psi}_{Bitten}$ is:

```
## [1] 37.27273
```

$\widehat{\psi}_{Bitten} = 37.27$

The asymptotic confidence interval based on the direct approach is given by:

```
##        Limit
##         Lower Limit Upper Limit
##   Value        5.87       68.68
```

Then the relative risk is less than 1. Also we cant that cows in estrous to be bitten with a different probability than cows not in estrous. Also the proportion crow in estrous bitten by a bat was 37.3 times higher for the crow Not in estrous bitten by a bat.

```
##  odds ratios for Estrous and Bitten
##
## [1] 115
```

```
##                                                                          2.5 %   97.5 %
## In estrous:Not in estrous/Bitten by a bat:Not bitten by a bat 34.39285 384.5276
```

```
## log odds ratios for Estrous and Bitten
##
## [1] 4.744932
```

```
##                                                                          2.5 %   97.5 %
## In estrous:Not in estrous/Bitten by a bat:Not bitten by a bat 3.537849 5.952015
```

Thus, $OR > 1$ and we have that probability to be bitten by a bat if the cows is in estrous is significantly higher than probability to be bitten by a bat if the cows is not in estrous.

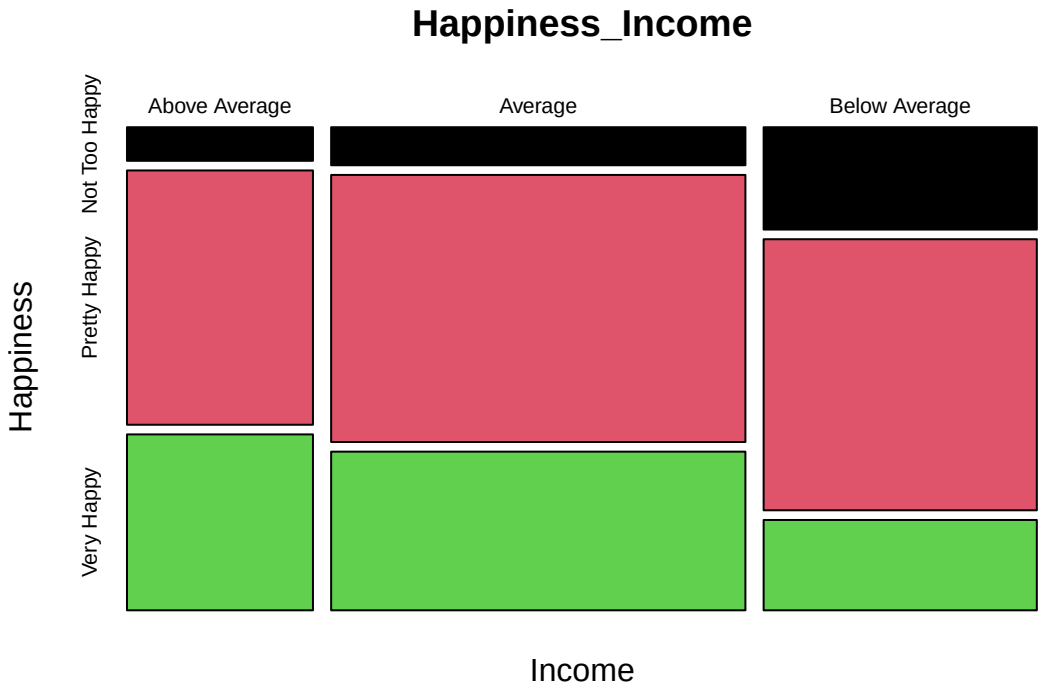
Practical Problems 2

(from Agresti) The following table shows data from the 2002 General Social Survey cross classifying a person's perceived happiness with their family income.

```
##                Happiness
## Income          Not Too Happy Pretty Happy Very Happy Total
##   Above Average            21          159        110   290
##   Average                  53          372        221   646
##   Below Average            94          249         83   426
##   Total                   168          780        414  1362
```

Are Happiness and Income independent?

Answer:




```
##                      X^2 df   P(> X^2)
## Likelihood Ratio 71.305  4 1.1990e-14
## Pearson          73.352  4 4.4409e-15
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.226
## Cramer's V        : 0.164
```

It is highly significant that the Happiness and the group of Income are not independent.

Practical Problems 3

To investigate the the rate of sprouted seeds in two different kinds of water rainwater, muddy water,and tap water were used to water 100 seeds each. Then they were checked and noted how many of those seeds sprouted. The result is presented in the following table:

```
##              Sprouted
## Water Type    Yes  No Total
##   Rain Water   64  36   100
##   Muddy Water  74  26   100
##   Tap Water    60  40   100
##   Total       198 102   300
```

The question is whether the type of the water has an effect on the probability for a seed to sprout.

Answer:

```
##
##  Pearson's Chi-squared test
##
## data:  Water_Sprouted
## X-squared = 4.6346, df = 2, p-value = 0.09854
```

There is a significant difference of the type of the water. Then the different water type has an effect on the probability for a seed to sprout.

Practical Problems 4

(From Agresti) The following table comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In 20 hospitals in London, UK, patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a noncancer control patient at the same hospital of the same sex and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

(a) Identify the response variable and the explanatory variable.
(b) Identify the type of study this was.
(c) Can you use these data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?
(d) Summarize the association, and explain how to interpret it.
(e) Apply a suitable test procedure.

```
##             Lung Cancer
## Have Smoked Not Cases Control
##       Yes         688     650
##       No           21      59
##       Total       709     709
```

Answer (a):

(-) Response variable: Lung Cancer

(-) Explanatory variable: Have Smoked

Answer (b):

The type of study was: Retrospective studies.

Answer (c)

if i need design a study to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer, the first that i think is: - Take a random sample of people and then put each people in a class (smokers, nonsmokers, lung cancer and not lung cancer) but in the population are few people whit lung cancer, then is necessary select a big sample of people to find people with lung cancer. This design not is good because require large expenditure on resources.

For the reason explain above, the best design to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer is the design expose in this exercise.

Answer (d)

This association is a contingency table that referee to Model II, in this table we have two group the firs group it is conformed by people whit lung cancer and the second group include the control person, also the people are classify according they smoking behavior.

Answer (e)

Steps for the construction of test:

Step 1: Select the hypothesis H_0 and H_1;
$H_0 : p_1 \leq p_2$
$H_1 : p_1 > p_2$

p_1 is the probability of a person smoker given that have lung cancer. p_2 is the probability of a person smoker given that is inside control group.

Step 2: Fix the level of the test or the 1st type error equal to α;
$\alpha = 0.05$

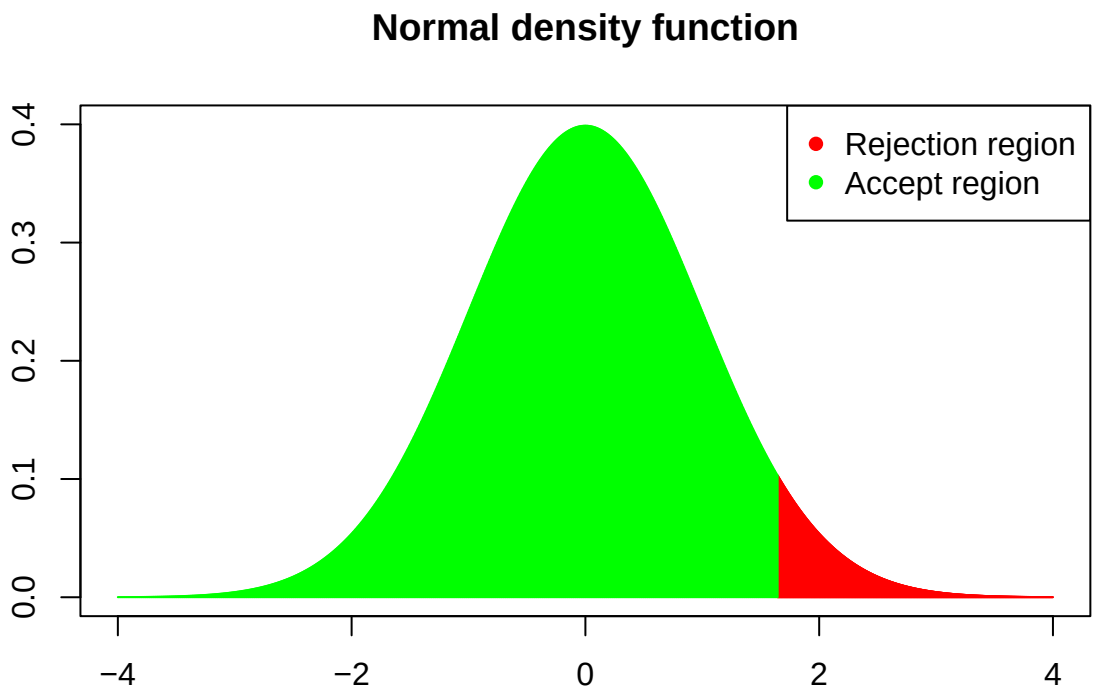
Step 3: Select the test statistic, t;
$t = \frac{\hat{\theta}}{se\left(\hat{\theta}\right)} \sim N\left(0, 1\right)$

Where: $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$ and $se\left(\widehat{\theta}\right) = \sqrt{\frac{\widehat{p}_1\left(1-\widehat{p}_1\right)}{n_1} + \frac{\widehat{p}_2\left(1-\widehat{p}_2\right)}{n_2}}$

Step 4: Determine the form of the rejection region W, depending on the behavior of t under H_1;
$W = \{t > Z_{1-0.05}\}$

Step 5: Explicitly compute the rejection region W according to α;

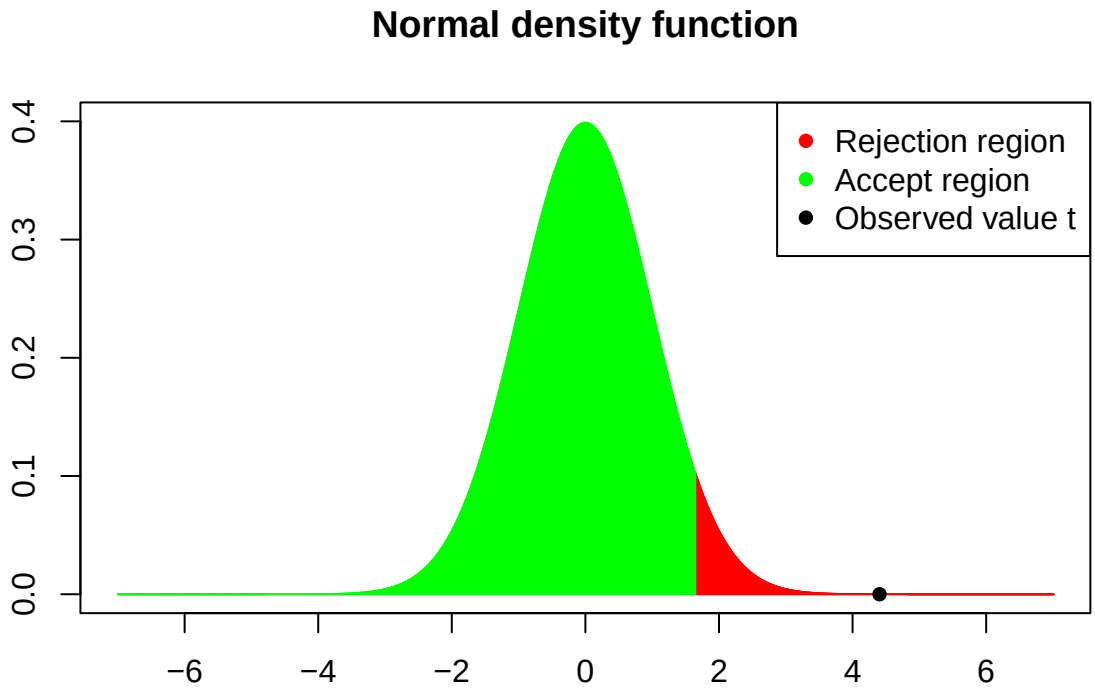
$W = \{(1.644854, \infty)\}$



Step 6: Compute the observed value, t, for the test statistic t;

```
##        Parameters
##                p1        p2      theta  sd(theta)      t
##   Value 0.9703808 0.9167842 0.05359661 0.01217137 4.4035
```

Step 7: According to t, decide whether to accept or not H_0.



As t it is in rejection region W ($t \in W$), we reject H_0 and accept H_1, therefore $p_1 > p_2$. Then we cant conclude that the to smoker increase the probability to get lung cancer.

Also the relative risk $\widehat{\psi}$ is:

```
## [1] 1.058462
```

$\widehat{\psi} = 1.058$

The asymptotic confidence interval based on the direct approach is given by:

```
##        Limit
##         Lower Limit Upper Limit
##   Value        1.03        1.09
```

Then the relative risk is less than 1. Also we cant ratify that the to smoker increase the probability to get lung cancer.

The OR is:

```
##  odds ratios for Lung Cancer and Have Smoked
##
## [1] 2.973773
```

```
##                              2.5 %   97.5 %
## Not Cases:Control/Yes:No 1.786737 4.949427
```

The conclusion is analogue to the case of relative risk.

Practical Problems 5

A company filling grass seed bags wants to evaluate their filling machine. The following distribution is advertised on their bags, where T1-T5 are different kinds of grass seeds:

```
##        Kind
##          T1   T2   T3   T4   T5
##   Prob. 0.5 0.25 0.15 0.05 0.05
```

The company wants to check if the seed distribution in the bags fits the advertised distribution. They take a sample of size 1000 and find the following summarized data:

```
##         Kind
##           T1  T2  T3 T4 T5
##   Count. 480 233 160 63 64
```

Does the seed distribution in the bags correspond correctly to the advertised distribution?

Answer

For Kind T1:

Steps for the construction of test:

Step 1: Select the hypothesis H_0 and H_1;
$H_0 : p = p_0$
$H_1 : p \neq p_0$

Step 2: Fix the level of the test or the 1st type error equal to α;
$\alpha = 0.05$

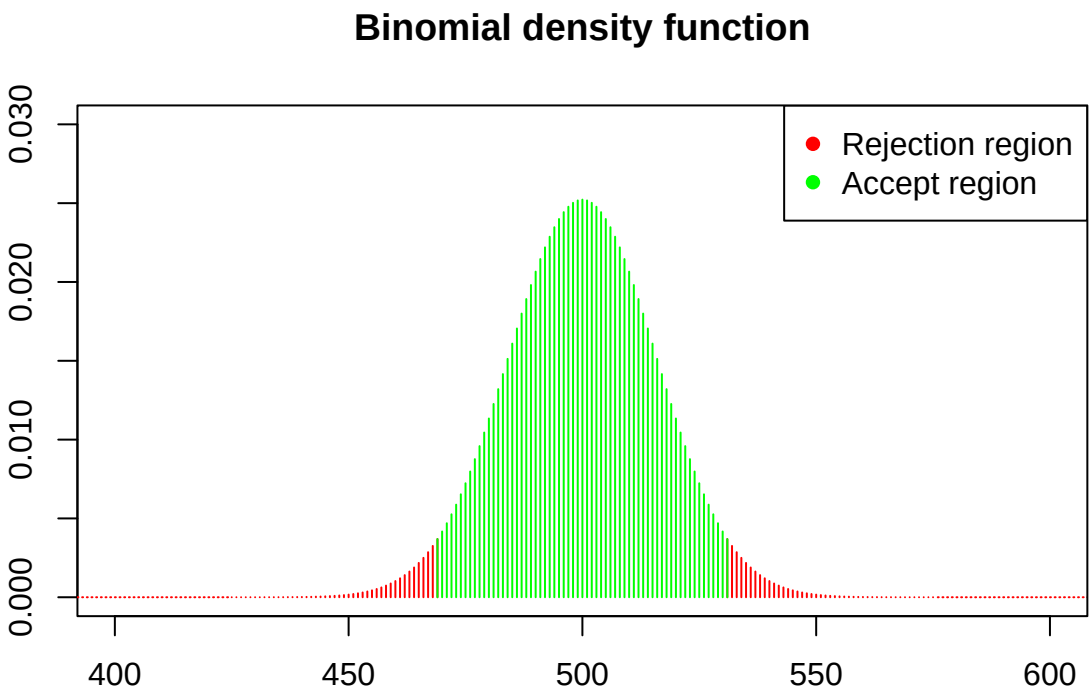
Step 3: Select the test statistic, t;
$t = \sum_{i=1}^{n} Y_i \sim B(n, p)$

Step 4: Determine the form of the rejection region W, depending on the behavior of t under H_1;
$W = \left\{ \left(-\infty, b_{\frac{\alpha}{2}}\right] \cup \left[b_{1-\frac{\alpha}{2}}, \infty\right) \right\}$

where b_α is the α quantile of the binomial distribution $B(n, p)$

Step 5: Explicitly compute the rejection region W according to α;

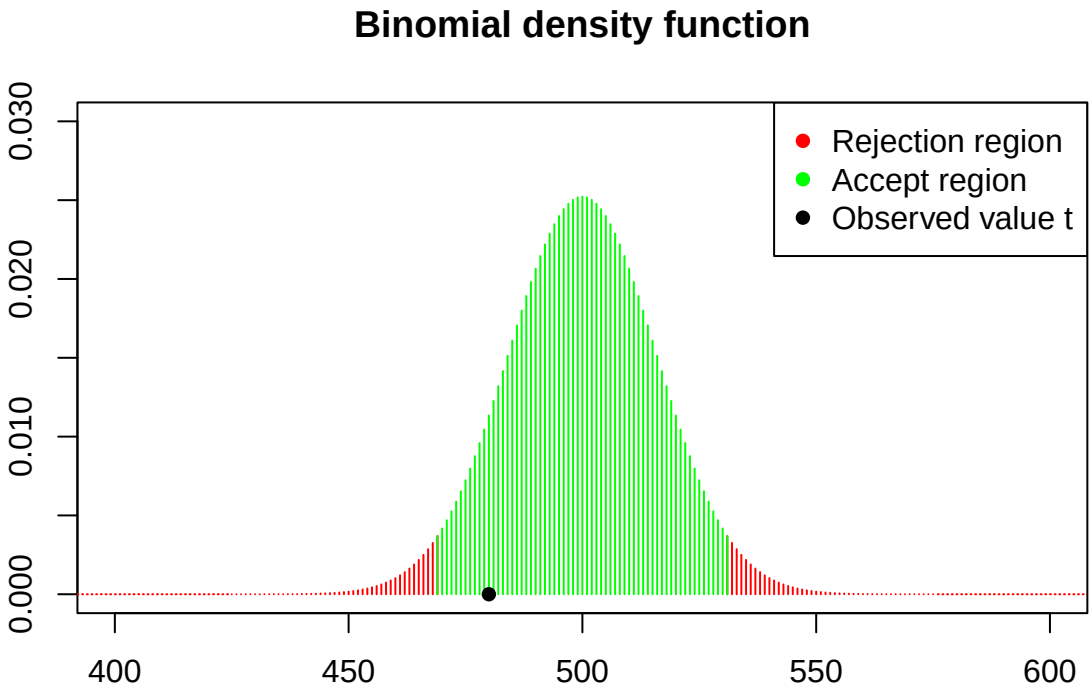
$W = \{(-\infty, 469] \cup [531, \infty)\}$



Step 6: Compute the observed value, t, for the test statistic t;

```
## [1] "t= 480"
```

Step 7: According to t, decide whether to accept or not H_0.



As t is not in rejection region W ($t \notin W$), we accept H_0, therefore $p = p_0$.

Then Yes, there is a correspondence between the distribution of the seed bags of the kind T1 and the advertised distribution

Is analogue for T2,T3,T4 and T5.

Use R function we have:

```
## [1] "For Kind T1"
##
##  Exact binomial test
##
## data:  Kind[1, i] and 1000
## number of successes = 480, number of trials = 1000, p-value = 0.2174
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4486329 0.5114851
## sample estimates:
## probability of success
##                   0.48
##
## [1] "Yes, there is a correspondence between the distribution of the seed "
## [1] "bags of the kind T1  and the advertised distribution"
## [1] " "
## [1] " "
## [1] " "
## [1] "For Kind T2"
```

```
##
##  Exact binomial test
##
## data:  Kind[1, i] and 1000
## number of successes = 233, number of trials = 1000, p-value = 0.2281
## alternative hypothesis: true probability of success is not equal to 0.25
## 95 percent confidence interval:
##  0.207116 0.260466
## sample estimates:
## probability of success
##                 0.233
##
## [1] "Yes, there is a correspondence between the distribution of the seed "
## [1] "bags of the kind T2  and the advertised distribution"
## [1] " "
## [1] " "
## [1] " "
## [1] "For Kind T3"
##
##  Exact binomial test
##
## data:  Kind[1, i] and 1000
## number of successes = 160, number of trials = 1000, p-value = 0.3758
## alternative hypothesis: true probability of success is not equal to 0.15
## 95 percent confidence interval:
##  0.1378052 0.1842168
## sample estimates:
## probability of success
##                  0.16
##
## [1] "Yes, there is a correspondence between the distribution of the seed "
## [1] "bags of the kind T3  and the advertised distribution"
## [1] " "
## [1] " "
## [1] " "
## [1] "For Kind T4"
##
##  Exact binomial test
##
## data:  Kind[1, i] and 1000
## number of successes = 63, number of trials = 1000, p-value = 0.06906
## alternative hypothesis: true probability of success is not equal to 0.05
## 95 percent confidence interval:
##  0.04874693 0.07988804
## sample estimates:
## probability of success
##                 0.063
##
## [1] "Yes, there is a correspondence between the distribution of the seed "
## [1] "bags of the kind T4  and the advertised distribution"
## [1] " "
## [1] " "
## [1] " "
## [1] "For Kind T5"
```

```
##
##  Exact binomial test
##
## data:  Kind[1, i] and 1000
## number of successes = 64, number of trials = 1000, p-value = 0.04958
## alternative hypothesis: true probability of success is not equal to 0.05
## 95 percent confidence interval:
##  0.04963309 0.08099507
## sample estimates:
## probability of success
##                  0.064
##
## [1] "No, there is not correspondence between the distribution of the seed "
## [1] "bags of the kind T5  and the advertised distribution"
## [1] " "
## [1] " "
## [1] " "
```

Practical Problems 6

Consider the example "Blood pressure" discussed in the slides.

(a) We can fit a logistic regression model taking as explanatory variables the values

```
##
## Pressure  [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
##    Value 111.5 121.5 131.5 141.5 151.5 161.5 176.5 191.5
```

(b) Fit the logistic regression and compare the fitted probabilities with the relative interval fre-quencies.

(c) In the data set *bpchol.r* also the cholesterol level is given. Fit a logistic regression that describes the effect of cholesterol on heart disease.

(d) Fit a logistic regression model that simultaneously describes the effects of cholesterol and blood pressure on heart disease.

Answer (a):

We can fit a logistic regression model with the variables:

```
##         Heart Disease
## Pressure Pressure Present Absent
##        1    111.5       3    153
##        2    121.5      17    235
##        3    131.5      12    272
##        4    141.5      16    255
##        5    151.5      12    127
##        6    161.5       8     77
##        7    176.5      16     83
##        8    191.5       8     35
```

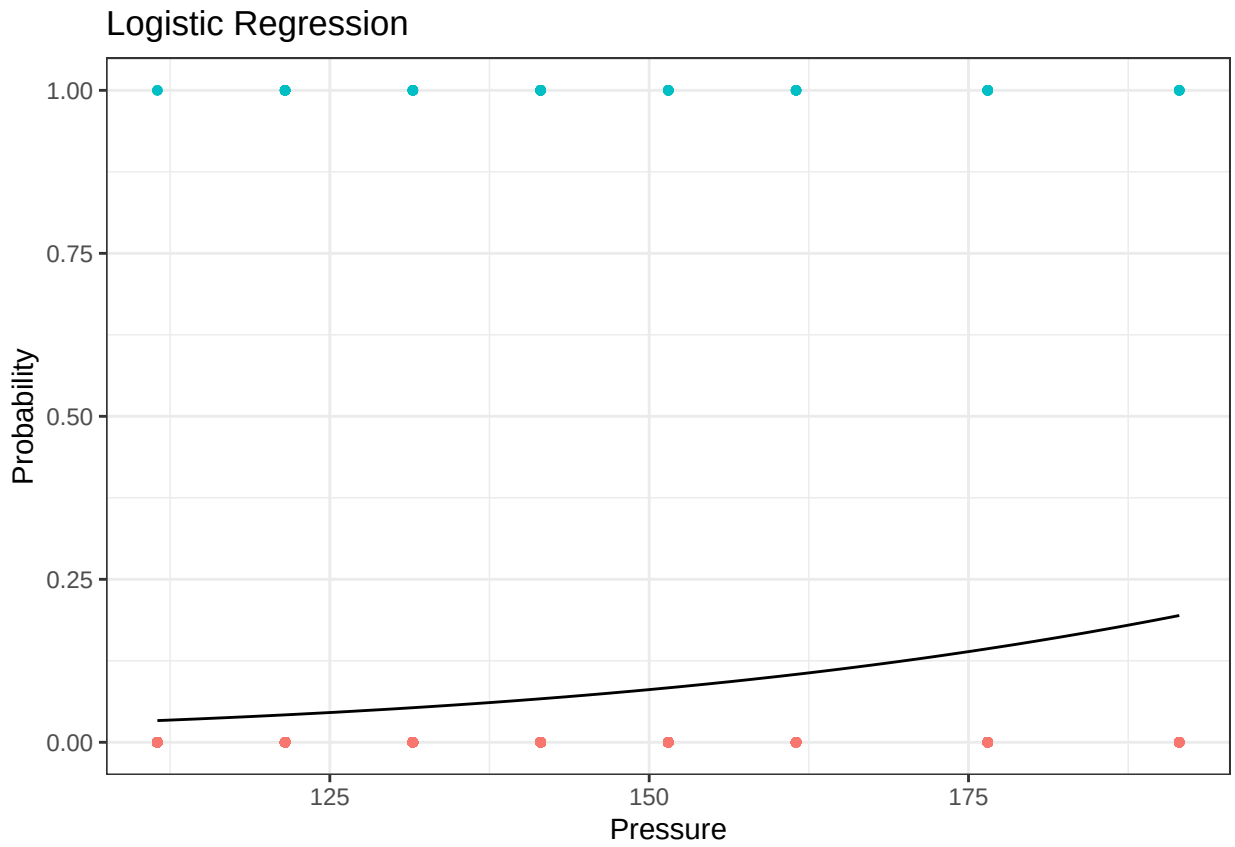
Answer (b):

Fit the logistic regression:

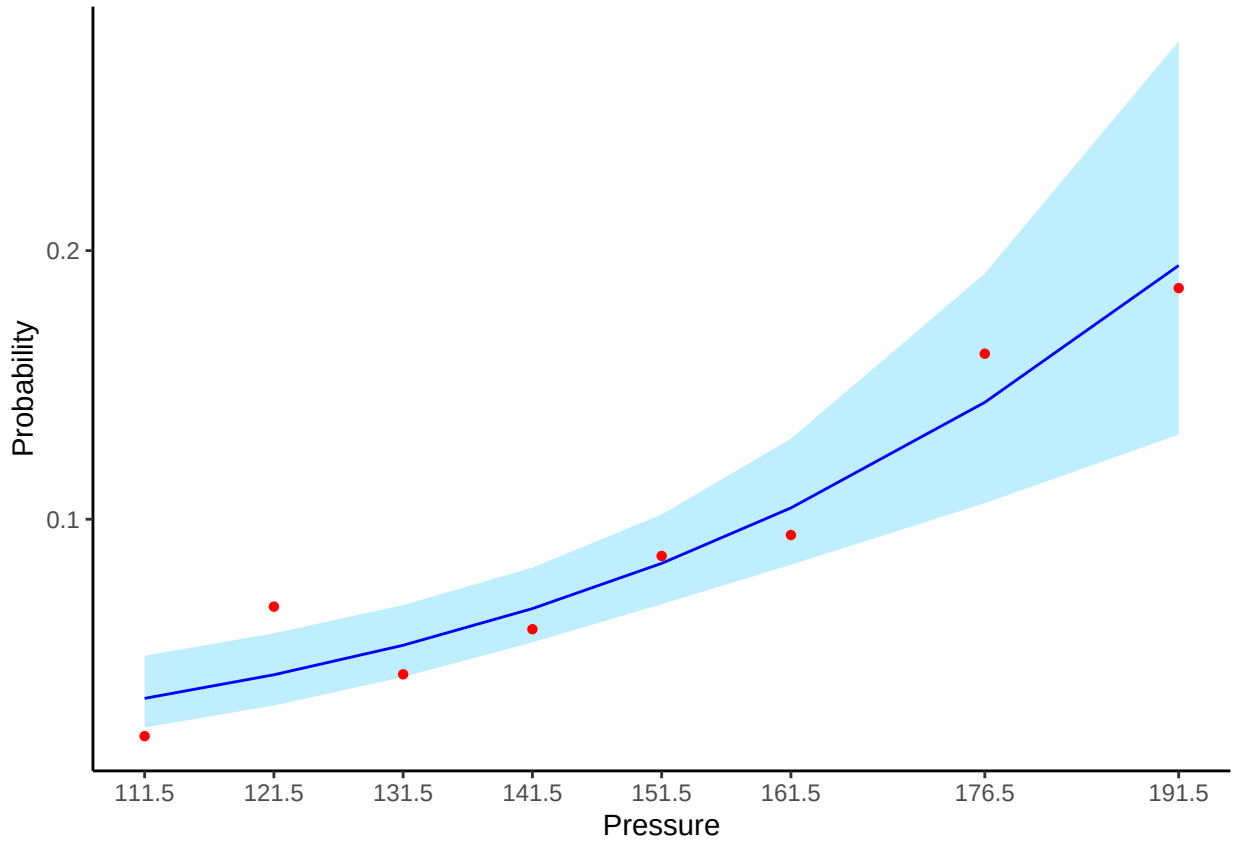
```
##
## Call:
## glm(formula = `Heart Disease` ~ Pressure, family = binomial,
##     data = Pressure_expand)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6576  -0.3716  -0.3302  -0.2933   2.6085
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.082033   0.724293  -8.397  < 2e-16 ***
## Pressure     0.024338   0.004843   5.025 5.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 668.83  on 1328  degrees of freedom
## Residual deviance: 644.72  on 1327  degrees of freedom
## AIC: 648.72
##
## Number of Fisher Scoring iterations: 5
```

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

```
##                  2.5 %      97.5 %
## (Intercept) -7.5165871 -4.67188476
## Pressure     0.0147868  0.03381264
```



Compare the fitted probabilities with the relative interval frequencies



Legend:

red point: relative frequencies

line blue: fitted probabilities

area light blue: confidence interval to 95%

We conclude that for almost all cases the relative frequencies is inside of confidence interval to 95%.

Answer (c):

The data set *bpchol.r* is not in the curse folder.

Answer (d):

The data set *bpchol.r* is not in the curse folder.

Practical Problems 7

On February 28, 1986, the space shuttle Challenger took on the 25th flight in NASA's space shuttle program. Less than 2 minutes into the flight, the spacecraft exploded, killing all on board. The space shuttle uses two booster rockets to help lift it into the orbit. Each booster rocket consists of several pieces whose joints are sealed with rubber O-rings, which are designed to prevent the release of hot gases produced during combustion. Each booster contains 3 primary O-rings (for a total of 6 for the orbiter). In the 23 previous flights the O-rings were examined for damage. One interesting question is the relationship of O-ring damage

to temperature, particularly since it was forecast to be cold on the morning of January 28, 1986 ($31°F$). There was a good deal discussion among the engineers the previous day as to whether the flight should go on as planned or not. Use the data given in SpaceShuttle.r to estimate failure probability depending on the temperature.

```
## # A tibble: 6 x 5
##   Temperature Erosion Blowby Total Failure
##         <dbl>   <dbl>  <dbl> <dbl>   <dbl>
## 1          53       3      2     5       1
## 2          57       1      0     1       1
## 3          58       1      0     1       1
## 4          63       1      0     1       1
## 5          66       0      0     0       0
## 6          67       0      0     0       0
```

Answer (a):

We can fit a logistic regression:

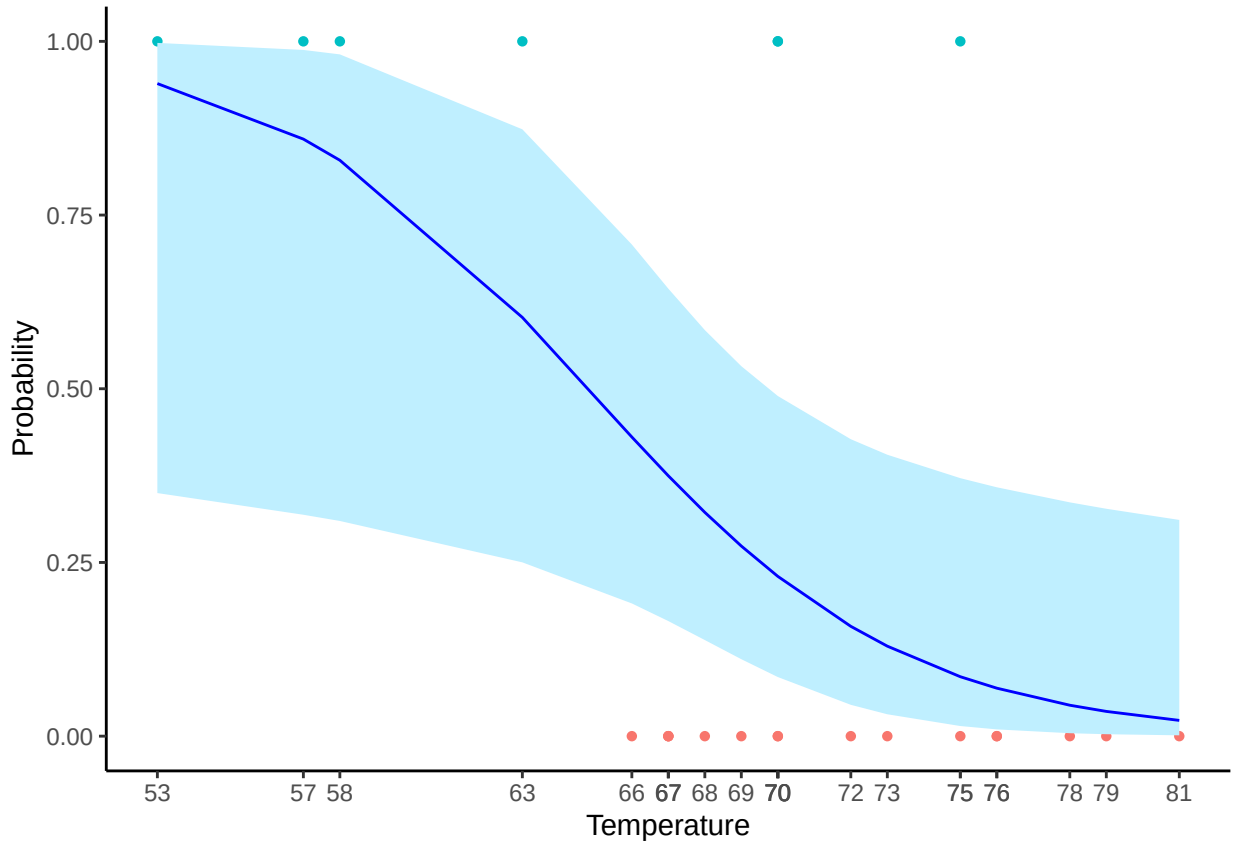
```
##
## Call:
## glm(formula = Failure ~ Temperature, family = binomial, data = SpaceShuttle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0611  -0.7613  -0.3783   0.4524   2.2175
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  15.0429     7.3786   2.039   0.0415 *
## Temperature  -0.2322     0.1082  -2.145   0.0320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.267  on 22  degrees of freedom
## Residual deviance: 20.315  on 21  degrees of freedom
## AIC: 24.315
##
## Number of Fisher Scoring iterations: 5
```

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

```
##                  2.5 %      97.5 %
## (Intercept)  3.3305848 34.34215133
## Temperature -0.5154718 -0.06082076
```

We cant to explain in this point, that, zero is not include within the coefficient confidence interval, that's one more proof that the coefficients are significant.

Plot:



Forecast to $31°F$.

```
##   Temperature Prediction Lower Limit Upper Limit
## 1          31  0.9996088   0.4816106   0.9999999
```

Then it can be stated that the probability of damage to $31°F$ is very high.