W271 Group Lab 1

Due 11:59pm Pacific Time Sunday Sep 12 2021

Aidan Jackson, Sandip Panesar, Devesh Khandelwal

Instructions (Please Read Carefully):

- 20 page limit (strict)
- Submit by the due date. **Late submissions will not be accepted.**
- Do not modify fontsize, margin or line_spacing settings
- One student from each group should submit the lab to their student github repo by the deadline; submission and revisions made after the deadline will not be graded
- Answers should clearly explain your reasoning; do not simply 'output dump' the results of code without explanation
- Submit two files:

  1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the code in your pdf file
  2. The R markdown (Rmd) file used to produce the pdf file. Knit to pdf, **do not knit to html and save as pdf**

  The assignment will not be graded unless **both** files are submitted

- Name your files to include all group members names. For example the students' names are Stan Cartman and Kenny Kyle, name your files as follows:

  – `StanCartman_KennyKyle_Lab1.Rmd`
  – `StanCartman_KennyKyle_Lab1.pdf`

- Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files
- All answers should include a detailed narrative; make sure that your audience can easily follow the logic of your analysis. All steps used in modeling must be clearly shown and explained
- For statistical methods that we cover in this course, use the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you must provide an explanation of why such libraries and functions are used and reference the library documentation. For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc
- For mathematical formulas, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file

- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity.

Investigation of the 1989 Space Shuttle Challenger Accident

Carefully read the Dalal et al (1989) paper (Skip Section 5).

Part 1 (25 points)

Conduct a thorough EDA of the data set, including univariate, bivariate and trivariate analysis. This should include both graphical and tabular analysis as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score. Since the report has a page-limit, you will have to be selective when choosing visuals to illustrate your key points, associated with a concise explanation of the visuals. This EDA should begin with an inspection of the given dataset; examination of anomalies, missing values, potential of top and/or bottom code etc.

First, the data set will be read-in.

```r
library(data.table)
library(ggplot2)
library(gridExtra)
library(dplyr)
library(mcprofile)

d <- read.csv("challenger.csv")
d <- data.table(d)
```

With the data loaded, it can be seen that there are 5 variables and 23 observations. Between these, there are 0 missing values. In total there is a small amount of information available, which is to be expected with rare events such as space shuttle launches. However, it will also provide more opportunities for intimate examination of the variables and their trends.

Based on the information in the accompanying paper, the meaning of each variable can be understood. `Flight` represents a generic index for each launch entry. `O.ring` represents the number of primary field O-rings which experienced a "thermal distress" event for each flight. Each shuttle has six primary field O-rings in total, which is shown by the variable `Number`. `Temp` and `Pressure` represent the O-ring launch temperature in Fahrenheit and pressure in psi respectively. With each of these variables well defined, it can be seen that none are top or bottom coded to censor certain specific values.

Because the total number of O-rings for each launch is constant at 6, it will not be used further. Instead, the O-ring variable will be examined to start.

```r
ggplot(d, aes(O.ring)) +
  geom_histogram(binwidth = 1, fill = "black", color='black', alpha=0.7) +
  ggtitle("Figure 1. O-Ring Distribution") +
  xlab("Number of O-Ring Thermal Distress Events") +
  ylab("Count")
```

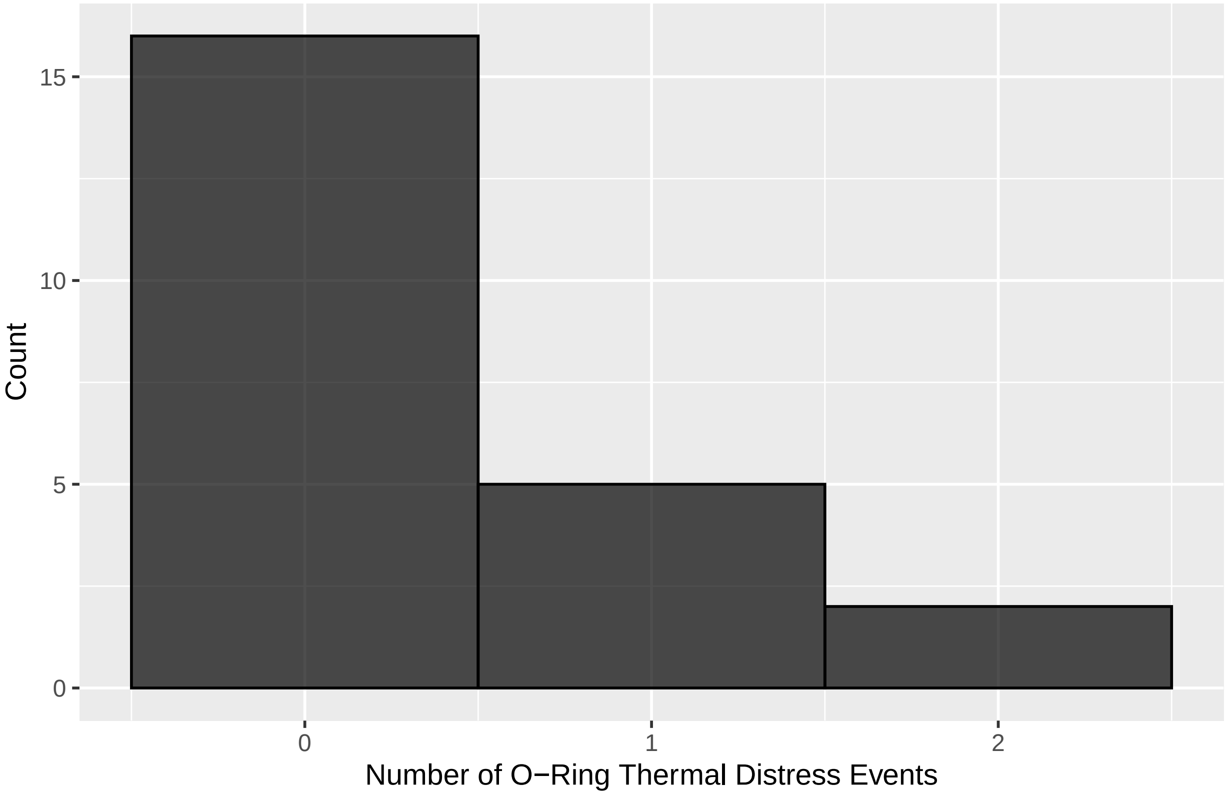


Figure 1. O−Ring Distribution

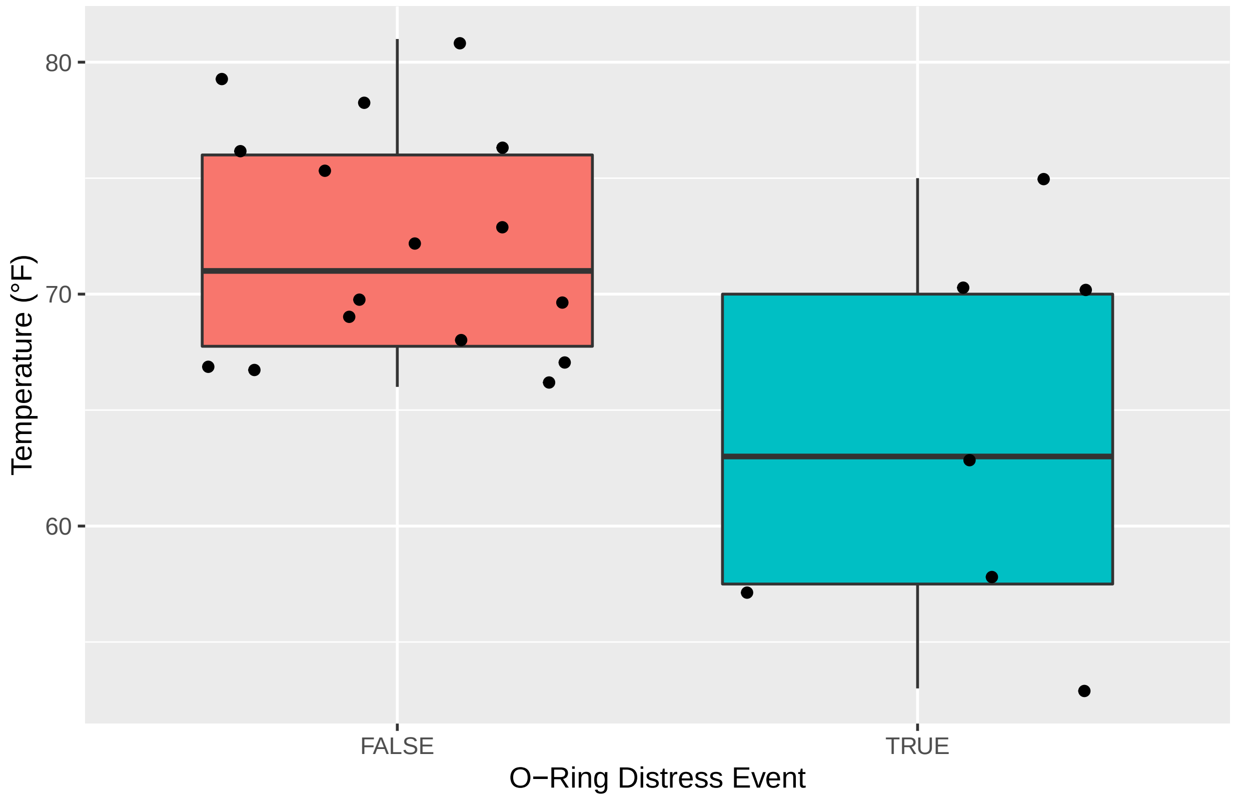
Shown above in *Figure 1.*, the majority of launches had O-rings which experienced no thermal distress events. About 5 launches had a single O-ring experience thermal distress, while even fewer had two. None of the recorded launches had greater than two O-rings distress events. With the majority of launches not experiencing any O-ring distress in this data set, those that experienced any distress at all can be grouped together for a more discrete analysis.

```r
d$distress <- d$O.ring > 0
```

To better differentiate between these two situations, a new Boolean variable is created which indicates whether a launch had at least one O-ring distress event or not. With two potential predictors of these failures, `Temp` and `Pressure`, it will be important to understand how they relate to the distribution of the outcome.

```r
ggplot(d, aes(factor(distress), Temp)) +
  geom_boxplot(aes(fill = factor(distress))) +
  theme(legend.position = "none") +
  geom_jitter() +
  ggtitle("Figure 2. Temperature vs. O-Ring Distress Event") +
  xlab("O-Ring Distress Event") +
  ylab("Temperature (°F)")
```

Figure 2. Temperature vs. O−Ring Distress Event



Demonstrated in *Figure 2*, O-Ring distress events generally occurred at lower launch temperatures. In fact, with so few total observations, it can be seen that over half of the launches with O-ring failures were at lower temperatures than the coldest launch with no O-ring failures. There is a single outlier that is notable, however, where a distress event occurred at a launch temperature of ~75°F. This temperature would be greater than average even among launches which experienced no distress, further demonstrating how it is extreme. Note that because of this value, along with generally having a wider distribution, there is greater temperature variance among launches which did have distress events compared to those which did not.

The other environmental variable in the data set, `pressure`, may also be examined in addition to temperature to see if insightful relationships emerge. The `O.ring` variable may also be re-visualized with these two to see if the most extreme environmental conditions are associated with greater O-ring distress events.

```
ggplot(data = d, aes(Temp, O.ring, color = Pressure)) +
  geom_point() +
  ggtitle("Figure 3. O-Ring Failures vs. Environmental Conditions") +
  xlab("Temperature (°F)") +
  ylab("O-Ring Distress Events") +
  labs(size = "Pressure (psi)") +
  scale_color_gradient(low = "#56B1F7", high = "#132B43")
```

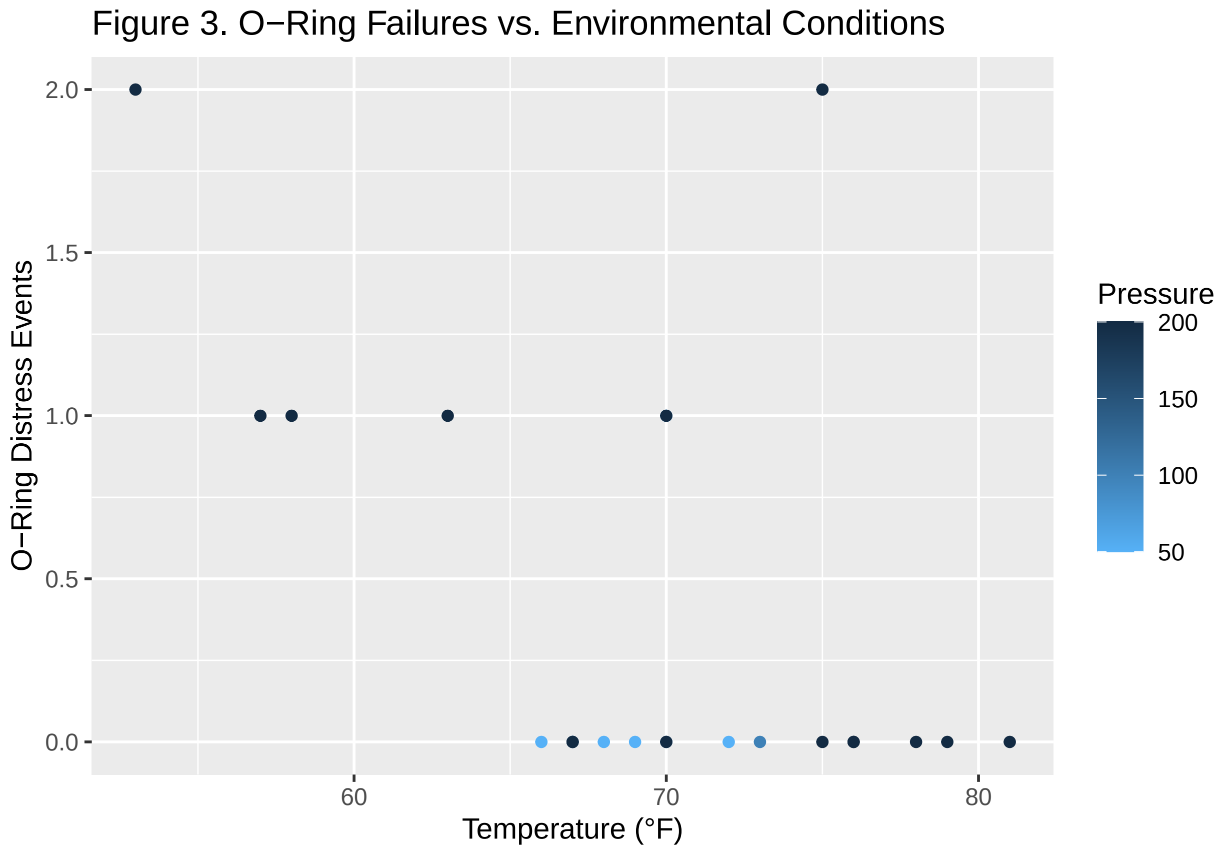


Figure 3 shows that generally the launches with a greater number of O-ring distress events occurred at lower temperatures and higher pressures. This also fits with conceptual information known about the O-rings, such as lower temperatures causing them to shrink in size and provide a less complete seal against gases. Higher pressures could then also cause more erosion/blowby from a greater amount of gas in the same space in the system. While all of the distress events did occur in the highest pressure category, the general trend does not seem as strong as with temperature. For example, the highest pressure rating of 200 psi was also the most common value of this variable generally. However, temperatures around the midpoint of the range ~65-75°F also had lower pressures, which may have helped them avoid any distress events.

The correlation between these variables can also be computed numerically for a more thorough understanding.

```
# generate correlation table without certain columns
cat("Table 1. Correlation of Numerical Variables\n")
```

```
## Table 1. Correlation of Numerical Variables
```

```
cor(d[ , !c("Flight","Number","distress")])
```

```
##                 Temp   Pressure     O.ring
## Temp      1.00000000 0.03981769 -0.5111264
## Pressure  0.03981769 1.00000000  0.2846663
## O.ring   -0.51112639 0.28466627  1.0000000
```

Table 1 shows the correlation between the three numerical variables which varied in the data set over the course of the launches, being `Temp`, `Pressure`, and `O.ring`. As suspected in *Figure 3*, pressure had a slight positive correlation with O-ring distress incidents. This supports the previous observation that higher pressures were associated with both launches with and without distress events, but lower pressures almost always were on flights without any distress events. Temperature has a stronger negative correlation with O-ring distress events. This continues to follow the trend shown in *Figures 2* and *3* where flights which experienced any distress were generally at colder temperatures. Finally, we can also see that temperature and pressure have a much smaller and slightly positive correlation with each other. This is in agreement with physical science, where gas enclosed in fixed volume will grow to higher pressures if the temperature is increased. Since the function of the O-ring is to provide a seal, i.e. a fixed volume, it would be reasonable that there is a slight positive correlation between the two.

```r
# turn data into a data.frame from data.table
# for later analysis
d <- data.frame(d)
```

Part 2 (20 points)

Answer the following from Question 4 of Bilder and Loughin Section 2.4 Exercises (page 129):

(a) The authors use logistic regression to estimate the probability an O-ring will fail. In order to use this model, the authors needed to assume that each O-ring is independent for each launch. Discuss why this assumption is necessary and the potential problems with it. Note that a subsequent analysis helped to alleviate the authors' concerns about independence.

The authors model the data based on a binomial distribution with n = 6 independent and identically-distributed (IID) O-rings on each flight, with each flight being IID as well. This distribution requires a single underlying relationship between the predictor variables, such as pressure and temperature, and the outcome of O-ring damage, for the model to fit. If the O-rings were not independent of each other, the outcome variable would not only be a function of these predictor variables but also of the damage of other O-rings. Consequently, there would not be just one relationship between the O-ring damage and the predictors, but multiple relationships depending on which O-ring was under consideration.

This is avoided with the assumption of independence, however it may not be completely valid. For example, the six primary O-rings are spread out across two rocket motors for each launch. It may be reasonable to expect that there could be small differences from rocket motor to rocket motor during each launch with respect to the stress the O-rings experience. For example, if a primary O-ring in one rocket experienced a certain amount of stress then that rocket's conditions may be expected put the other two O-rings under stress as well. The three O-rings in the other rocket, however, may have a different experience if that rocket performs differently. Therefore, in reality there is likely some relationship between the O-rings that is ignored for the sake of the model.

The authors also state that after each launch the solid rocket motors are recovered from the sea and examined for re-use. Recovery and examination was what allowed for the data set to be created and for the number of O-ring distress events to be known after the flights. However, it was not explicitly stated how many of the observations in the data set were from rockets which had been re-used. This might be a problem as a re-used motor could compromise the durability of other insulation and hence the stability of the O-ring. As a result, the probability of failure or success might differ between launches as well as between new and re-used rockets, violating the assumption of observations being identically distributed. This would also violate the assumption of independence between observations if the same rocket motor is used for some flights within the data set.

(b) Estimate the logistic regression model using the explanatory variables in a linear form.

There are two explanatory variables of interest, `Temp` and `Pressure`. First `Temp` is fit in a standalone model, and then another with both `Temp` and `Pressure`.

```
# Null hypothesis- coefficient of pressure is 0
model_h0 <-  glm(formula = distress ~ Temp, data = d ,
                 family = binomial(link = logit))

# Alt hypothesis- coefficient of pressure is not 0
model_h1 <- glm(formula = distress ~ Temp + Pressure, data = d ,
                family = binomial(link = logit))
```

(c) Perform LRTs to judge the importance of the explanatory variables in the model.

In Part B, two models were fit to test the hypothesis:

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

(where β_2 is the model coefficient of `Pressure`). Note that the statistical significance of `Temp` is demonstrated in Section 3-A.

A Likelihood Ratio (LR) test will be performed using the function *anova*. In general an LR test can be mathematically expressed as:

$$-2log(\Lambda) = -2log(\frac{L(\hat{\beta}^{(0)}|y_1,\ldots,y_n)}{L(\hat{\beta}^{(a)}|y_1,\ldots,y_n)} = -2\sum y_i log\left(\frac{\hat{\pi}_i^{(0)}}{\hat{\pi}_i^{(a)}}\right) + (1-y_i)log\left(\frac{1-\hat{\pi}_i^{(0)}}{1-\hat{\pi}_i^{(a)}}\right)$$

Where:

- $\hat{\pi}_i^{(0)}$ is the estimated probability of success under H_0
- $\hat{\pi}_i^{(a)}$ is the estimated probability of success under H_a

Code for the anova function is:

```
anova_obj <- anova(model_h0 , model_h1, test = "Chisq")
anova_obj
```

```
## Analysis of Deviance Table
##
## Model 1: distress ~ Temp
## Model 2: distress ~ Temp + Pressure
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     20.315
## 2        20     18.782  1   1.5331   0.2156
```

```
beta_est <- anova_obj$`Pr(>Chi)`[2]
```

From analyzing the anova outputs, we see that the probability of obtaining a non-zero value of coefficient of `Pressure` is 0.2156479, hence there is not enough evidence to reject the null-hypothesis.

(d) The authors chose to remove Pressure from the model based on the LRTs. Based on your results, discuss why you think this was done. Are there any potential problems with removing this variable?

In the previous LR test, there was not sufficient evidence to reject the null hypothesis, and thus the `Pressure` variable was found to be insignificant. This agrees with the authors' choice. To better understand this result, the confidence interval (CI) can also be examined. This helps to understand the uncertainty in this result with such a small number of samples in the data set.

```r
pressure_ci <- confint(object = model_h1 , parm = "Pressure", level = .95)
```

It can be seen that the 95% Wald CI is between -0.0057514 and 0.0322629. As zero lies within the CI, there is additional evidence that the `Pressure` variable is not statistically significant from both Wald and the previous LR test. There is a failure to reject the null-hypothesis, which states that $\beta_{Pressure} = 0$.

To see how this could effect the odds of a launch failure, and cause a potential problem in the model, consider a 10 psi increase in `Pressure` with all else constant.

```r
exp_pressure_ci <- exp(10*pressure_ci)
```

With every 10 psi increase in `Pressure` the odds of a launch failure changes between 0.9441084 and 1.3807523 times at the 95% confidence level. This re-demonstrates that `Pressure` is not significant, since the OR contains 1, but it does show that the majority of the predicted OR is greater than one. An OR greater than one would indicate that an increase in the variable would lead to greater odds of the outcome event occuring. As was seen before, there did appear to be a slight positive correlation between higher pressures and O-ring distress, which likely led to most of the OR being above 1 in the interval. In addition, `Pressure` is a physical phenomena related to `Temperature` in the system, and only considering one and not the other may be taking a unnecessary risk. However, due to being statistically insignificant, there is not evidence that the model would perform better with its inclusion and so it will be dropped.

Part 3 (35 points)

Answer the following from Question 5 of Bilder and Loughin Section 2.4 Exercises (page 129-130):

Continuing Exercise 4, consider the simplified model $logit(\pi) = \beta_0 + \beta_1 Temp$, where π is the probability of an O-ring failure. Complete the following:

(a) Estimate the model.

Above, we demonstrated that a model including the `Pressure` variable did not offer significantly more information compared to a model with only temperature as the dependent variable.

```r
fail_model <- glm(formula = distress ~ Temp, data = d ,
                  family = binomial(link = logit))
se <- sqrt(diag(vcov(fail_model)))
```

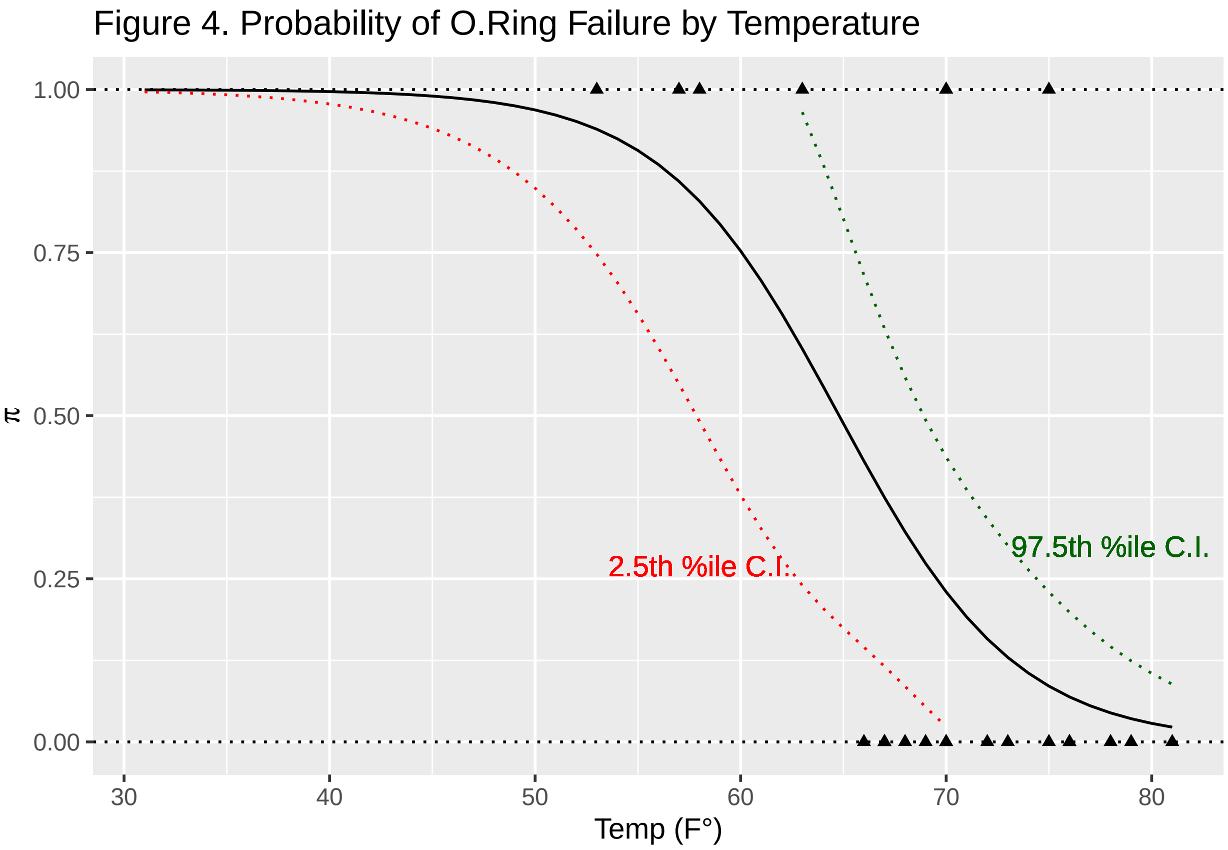
In this model, an increase in the temperature by 1 degree Farenheit will cause a -0.232 decrease in the log-odds of an O-ring failure event. This coefficient value is associated with a standard error of 0.108, which is significant at the 95% level. This demonstrates that `Temp` is a significant variable for predicting the outcome.

(b) Construct two plots: (1) π vs. Temp and (2) Expected number of failures vs. Temp. Use a temperature range of 31° to 81° on the x-axis even though the minimum temperature in the data set was 53°.

(c) Include the 95% Wald confidence interval bands for π on the plot. Why are the bands much wider for lower temperatures than for higher temperatures?

First, we construct a data frame of X-values (i.e. temperature in degrees Fahrenheit), that will be used to predict π. Then, we feed this data frame into the model using the predict function, with the `response` function specified and `se=TRUE` . This returns a corresponding vector of probabilities for failure based upon the existing `fail_model` which was previously specified:

```r
# Create a dataframe of new temperatures
temps <- data.frame(seq(31,81,1))
colnames(temps) = 'Temp'

# Create vector of predictions using model_h0 (i.e. the model without pressure)
temp_prob <- predict(object=fail_model, newdata=temps, type='response', se=TRUE)

# Pi values
temp_pis <- temp_prob$fit
# SEs
temp_ses <- temp_prob$se.fit

alpha <- .05

#2.5th %ile
lower_ci <- temp_pis - qnorm( p = 1-alpha/2)*temp_prob$se
#97.5th %ile
upper_ci <- temp_pis + qnorm( p = 1-alpha/2)*temp_prob$se

final_temp <- data.frame(temps$Temp, temp_pis, lower_ci, upper_ci)
colnames(final_temp) <- c('temp', 'pi', 'lwr', 'upr')

ggplot(data=final_temp, aes(x=temp)) +
  geom_line(aes(y=pi)) +
  geom_line(aes(y=lwr), color='red', linetype='dotted') +
  geom_line(aes(y=upr), color='darkgreen', linetype='dotted') +
  ylim(0,1) +
  ggtitle("Figure 4. Probability of O.Ring Failure by Temperature") +
  ylab(expression(pi)) +
  xlab("Temp (F°)") +
  geom_hline(yintercept=1.0, linetype='dotted') +
  geom_hline(yintercept=0.0, linetype='dotted') +
  geom_point(data = d, aes(x=Temp, y= ifelse(distress==TRUE, 1, 0) ),
             shape = 17) +
  geom_text(aes(x = 58, y = 0.27), label = "2.5th %ile C.I.", color='red') +
  geom_text(aes(x = 78, y = 0.30), label = "97.5th %ile C.I.", color='darkgreen')
```



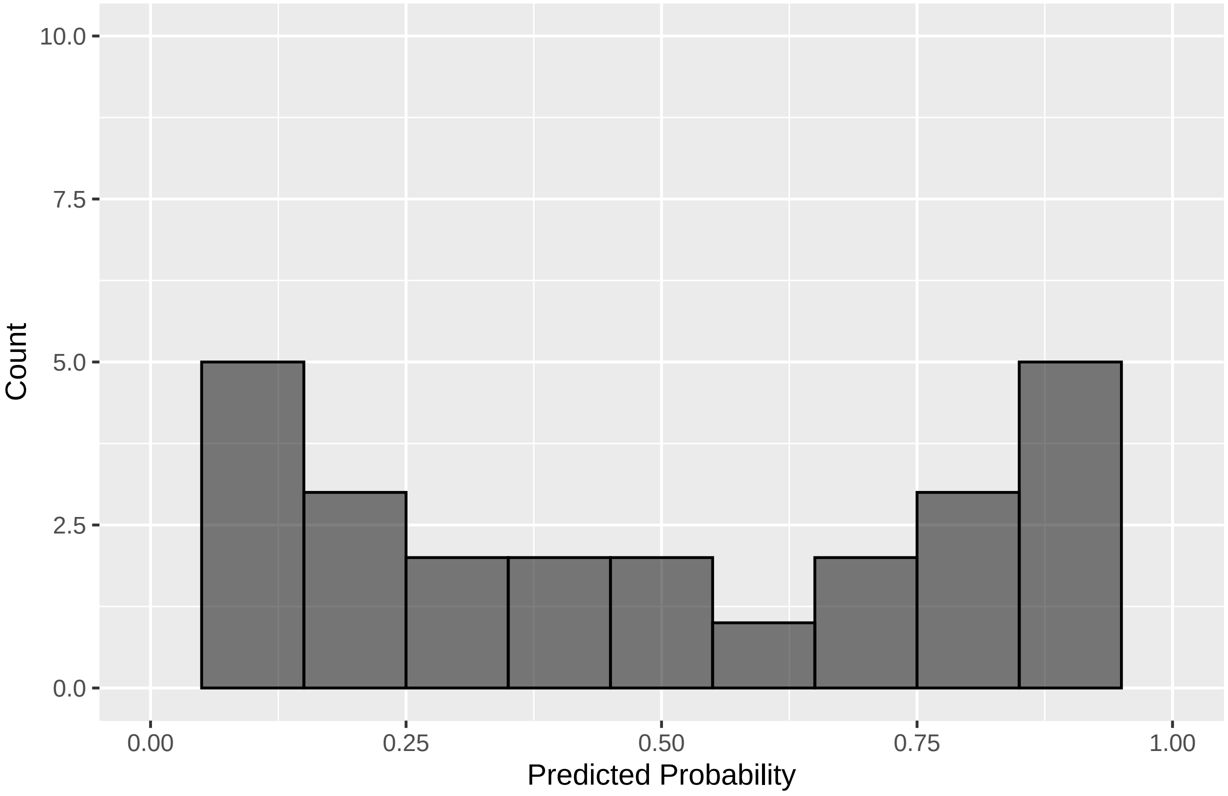
Shown above in *Figure 4*, the estimated probability of any O-ring failure is close to 1 at low temperatures, including the temperature at which the Challenger launched at.

From *Figure 4* we also observe that the width of the confidence interval (CI) increases gradually until ~55°F before decreasing. This is due to the fact that there are no ground truth observations below 50°F, and that the data is an approximation of the true distribution. Hence, there is a greater amount of sampling variance in $\hat{\pi}$ at these low temperatures.

```
ggplot(final_temp, aes(pi)) +
  geom_histogram(binwidth = .1, color='black', fill = "black", alpha=0.5) +
  xlim(0,1) +
  ylim(0,10) +
  ggtitle("Figure 5. Count Distribution of Predicted Probabilities") +
  xlab("Predicted Probability") +
  ylab("Count")
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

Figure 5. Count Distribution of Predicted Probabilities



Moreover, the Wald interval assumes the outcome variable follows a normal distribution and has a large sample size, which is not true and demonstrated by *Figure 5* above. Shown in *Figure 5*, $\hat{\pi}$ has fewer observations towards the center of its range, with more on the edges. This is almost the opposite of a standard normal distribution, where the median value is the most often observed. Hence, the resulting interval is broader in general.

Next, the expected number of failures as a function of temperature can be calculated.

```
# This function calculates the maximum number of failures at a specified probability of a sing
get_ring_failure <- function(est_prob){
  # This function is from the paper Dalal et al.
  mu = 6.0*(1-(1-est_prob)^(1/6))
  return(mu)
}
```

```
# Now calculate expected number of incidents

# Create vector of predictions using model_h0 (i.e. the model without pressure), also specify
temp_exps <- predict(object=fail_model, newdata=temps, type='response', se=TRUE)

# Pi values
pi_hat <- temp_exps$fit

# SEs
pi_hat_ses <- temp_exps$se.fit
```

```r
#calculating failure
failure_hat <- apply(X = data.frame(pi_hat), MARGIN = 1, FUN = get_ring_failure)

#creating a data frame
final_temp_exp <- data.frame(temps$Temp, pi_hat, failure_hat)
colnames(final_temp_exp) <- c('Temp', 'pi_hat', 'failure_hat')

#creating a subset of the original df to only retain relevant cols:
d_trunc <- d[,c("Temp","O.ring")]

#merging the data frames to create a final plot:
final_temp_exp = merge(x = final_temp_exp, y = d_trunc, by = "Temp",
                                 all.x = TRUE)

 ggplot() +
  geom_line(data=final_temp_exp, aes(x=Temp, y= failure_hat)) +
  ggtitle("Figure 6. Expected Number of O-Ring Incidents by Temperature") +
  ylab("No. Incidents") + xlab("Temp (F°)") +
   geom_point(data = final_temp_exp , aes(x=Temp, y= O.ring ), shape = 17) +
   labs(colour = "final_temp_exp$O.ring")
```

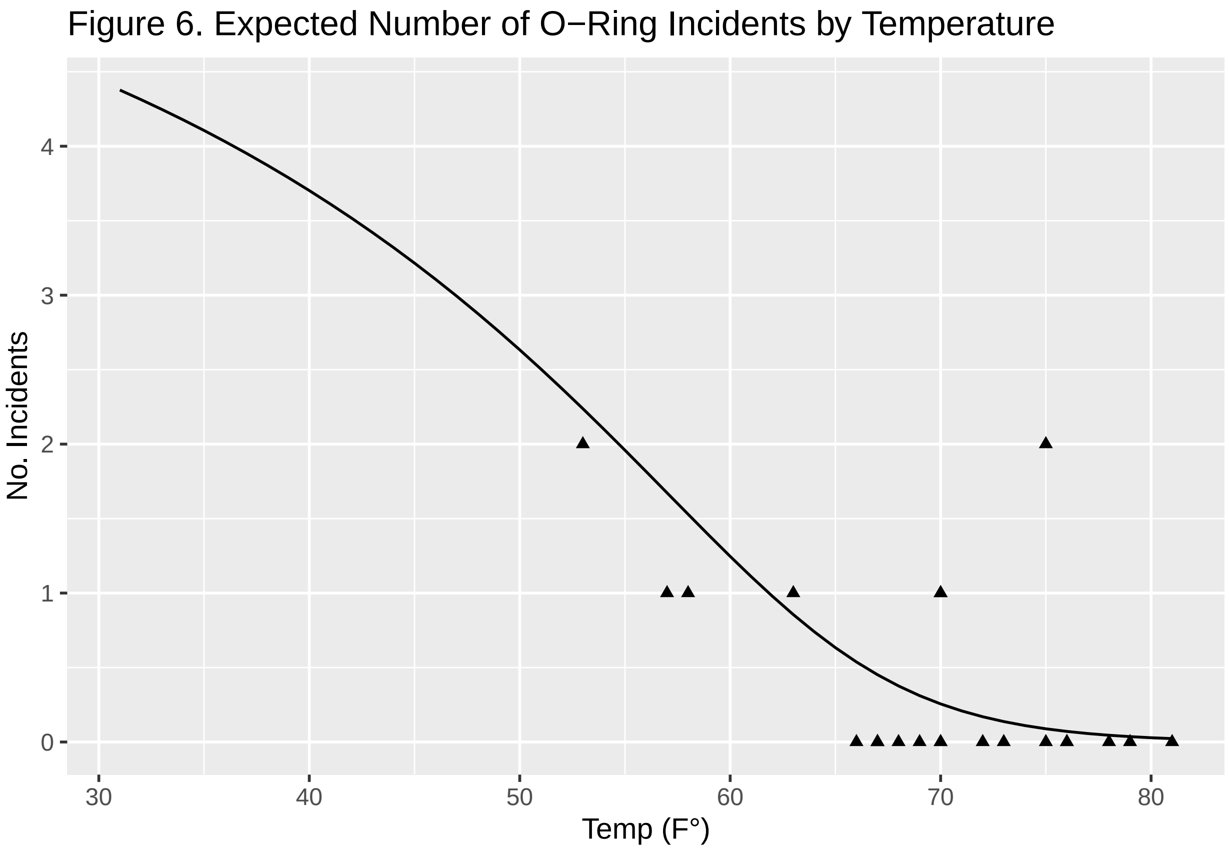


Figure 6 above shows that it is expected that more O-ring failures than previously observed would have occurred at the Challenger's launch temperature. This should have caused serious concern for

launching at the temperature of that day had this analysis been carried out ahead of time. The only point which does not fit this curve well is the observation with two O-ring failures at a temperature of about ~75°F.

(d) The temperature was 31° at launch for the Challenger in 1986. Estimate the probability of an O-ring failure using this temperature, and compute a corresponding confidence interval. Discuss what assumptions need to be made in order to apply the inference procedures.

We've already calculated the Wald intervals in the previous section. As shown before, a Wald test relies on following assumptions:

1. The underlying data is normally distributed.
2. We have a large sample size.

As we have seen in the previous subsection, $\hat{\pi}$ is not normally distributed and there is a small sample size. Therefore, in this section we'll look at estimating the CI with the Likelihood Ratio Test (LRT) using the model that we previously fitted. Unlike the Wald interval, the LRT assumes that the the test statistic has a `chi.sq` distribution with 1 degree of freedom.

```r
beta_0 <- fail_model$coefficients[1]
beta_1 <- fail_model$coefficients[2]

# Create a K-matrix for temp = 31 degrees

K <- matrix(data = c(1,31) , nrow = 1, ncol = 2)

# Calc -2log(lambda)
lc <- mcprofile(object=fail_model, CM=K)
ci_logit_profile <- confint(object=lc, level=0.95)

interval <- exp(ci_logit_profile$confint)/(1 + exp(ci_logit_profile$confint))

wald_int_profile <- wald(object=lc)

wald_ci_exp <- confint(wald_int_profile , level = .95)

wald_lower_exp <- wald_ci_exp$confint$lower
wald_upper_exp <- wald_ci_exp$confint$upper

wald_interval_lower <- exp(wald_lower_exp)/(1 + exp(wald_lower_exp))
wald_interval_upper <- exp(wald_upper_exp)/(1 + exp(wald_upper_exp))
```

For the LRT test and at 31°F, the odds of any `O.ring` failure has a 95% confidence interval of 0.8036982 and 1. The Wald test under the same conditions has the 95% confidence interval of any `O.ring` failure between 0.4816106 and 0.9999999. This demonstrates that the LRT is more conservative than the Wald and should be preferred.

(e) Rather than using Wald or profile LR intervals for the probability of failure, Dalal et al. (1989) use a parametric bootstrap to compute intervals. Their process was to (1) simulate a large number of data sets (n = 23 for each) from the estimated model of Temp; (2) estimate new models for each data set, say and (3) compute at a specific temperature of interest. The

> authors used the 0.05 and 0.95 observed quantiles from the simulated distribution as their 90% confidence interval limits. Using the parametric bootstrap, compute 90% confidence intervals separately at temperatures of 31° and 72°.27

First, the calculations will be set up.

```r
bs_data_1 <- data.frame(c(31))
colnames(bs_data_1) = 'Temp'

bs_data_2 <- data.frame(c(72.27))
colnames(bs_data_2) = 'Temp'

vector_pi <- NA
vector_lower_ci <- NA
vector_upper_ci <- NA

vector_pi2 <- NA
vector_lower_ci2 <- NA
vector_upper_ci2 <- NA


suppressWarnings(for(i in 1:1000) {
  #generating data from repeated sampling
  d_bootstrap <- sample_n(d, size=23, replace=TRUE)
  #modeling
  model <- glm(formula = distress ~ Temp, data = d_bootstrap,
               family = binomial(link = logit))
  prediction_1 <- predict(object=model,
                          newdata=bs_data_1, type='response', se=TRUE)
  vector_pi[i] <- prediction_1$fit
  #note that we are supposed to do a 90% CI
  vector_upper_ci[i] <- prediction_1$fit + 1.645*prediction_1$se.fit
  vector_lower_ci[i] <- prediction_1$fit - 1.645*prediction_1$se.fit
  #doing the same for 72 degrees
  prediction_2 <- predict(object=model,
                          newdata=bs_data_2, type='response', se=TRUE)
  vector_pi2[i] <- prediction_2$fit
  vector_upper_ci2[i] <- prediction_2$fit + 1.645*prediction_2$se.fit
  vector_lower_ci2[i] <- prediction_2$fit - 1.645*prediction_2$se.fit
})

temp_30_pis <- data.frame(data.frame(vector_pi),
                          data.frame(vector_lower_ci),
                          data.frame(vector_upper_ci))
colnames(temp_30_pis) = c('pi','lwr', 'upr')

temp_72_pis <-data.frame(data.frame(vector_pi2),
                         data.frame(vector_lower_ci2),
                         data.frame(vector_upper_ci2))
```

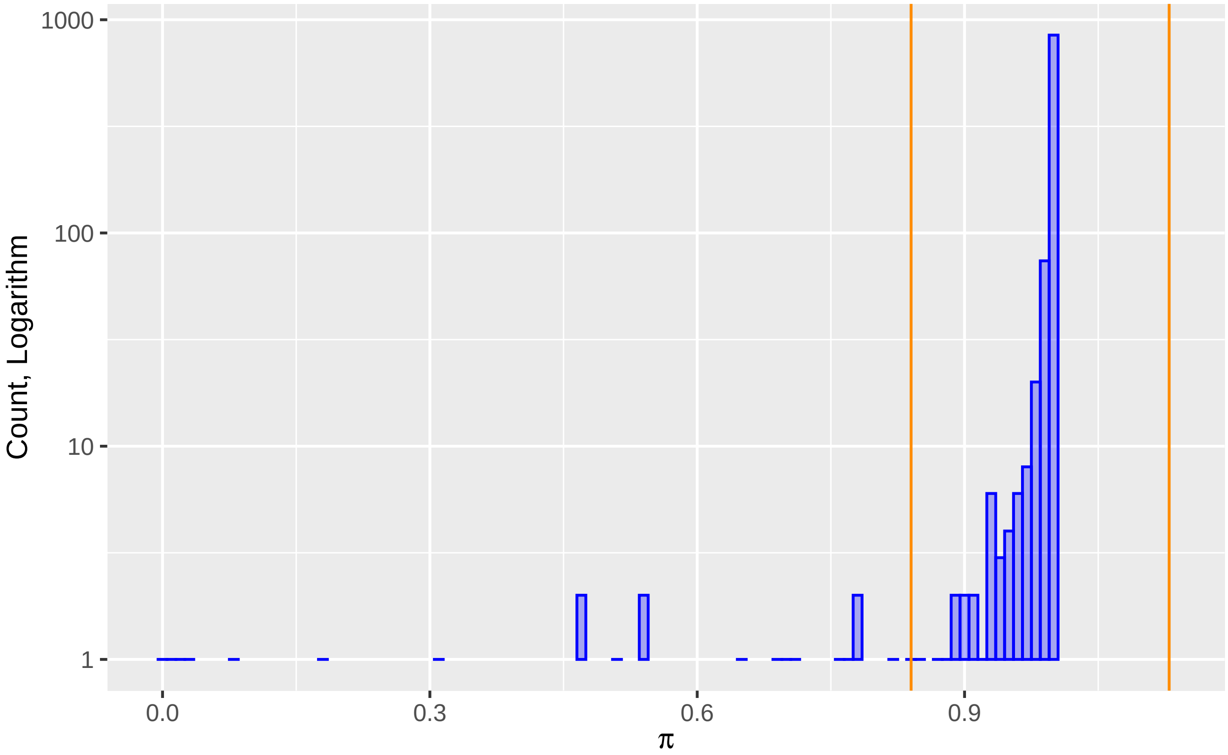
```
colnames(temp_72_pis) = c('pi','lwr', 'upr')

#calculating the standard deviations

pi_hat30_std_dev <- sd(temp_30_pis$pi)
pi_hat72_std_dev <- sd(temp_72_pis$pi)

#estimated mean
mu_hat_30 <- mean(temp_30_pis$pi)
mu_hat_72 <- mean(temp_72_pis$pi)

alpha <- .1
ci_30 <- mu_hat_30 + qnorm(p = c(alpha/2,1-alpha/2))*pi_hat30_std_dev
ci_72 <- mu_hat_72 + qnorm(p = c(alpha/2,1-alpha/2))*pi_hat72_std_dev
```

Next, the figures will be created.

```
# 30 degrees
plt_30 <- ggplot(data = temp_30_pis) +
aes(x = pi) +
geom_histogram(binwidth=0.01, fill="blue", color="blue", alpha=0.3) +
geom_vline(xintercept = ci_30 , color = 'darkorange') +
labs(title = "Figure 7. Histogram of Bootstrapped Probabilities of Failure at 31°F",
     caption = ("Vertical lines denote the 90% CI")) +
  ylab("Count, Logarithm") +
  xlab(expression(pi)) +
  scale_y_log10()

#72 degrees
plt_72 <- ggplot(data = temp_72_pis) +
aes(x = pi) +
geom_histogram(binwidth=0.01, fill="purple", color="purple", alpha=0.3) +
geom_vline(xintercept = ci_72 , color = 'darkorange') +
labs(title =
       "Figure 8. Histogram of Bootstrapped Probabilities of Failure at 72.27°F",
     caption = ("Vertical lines denote the 90% CI")) +
    ylab("Count") +
  xlab(expression(pi))

plt_30
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 80 rows containing missing values (geom_bar).
```

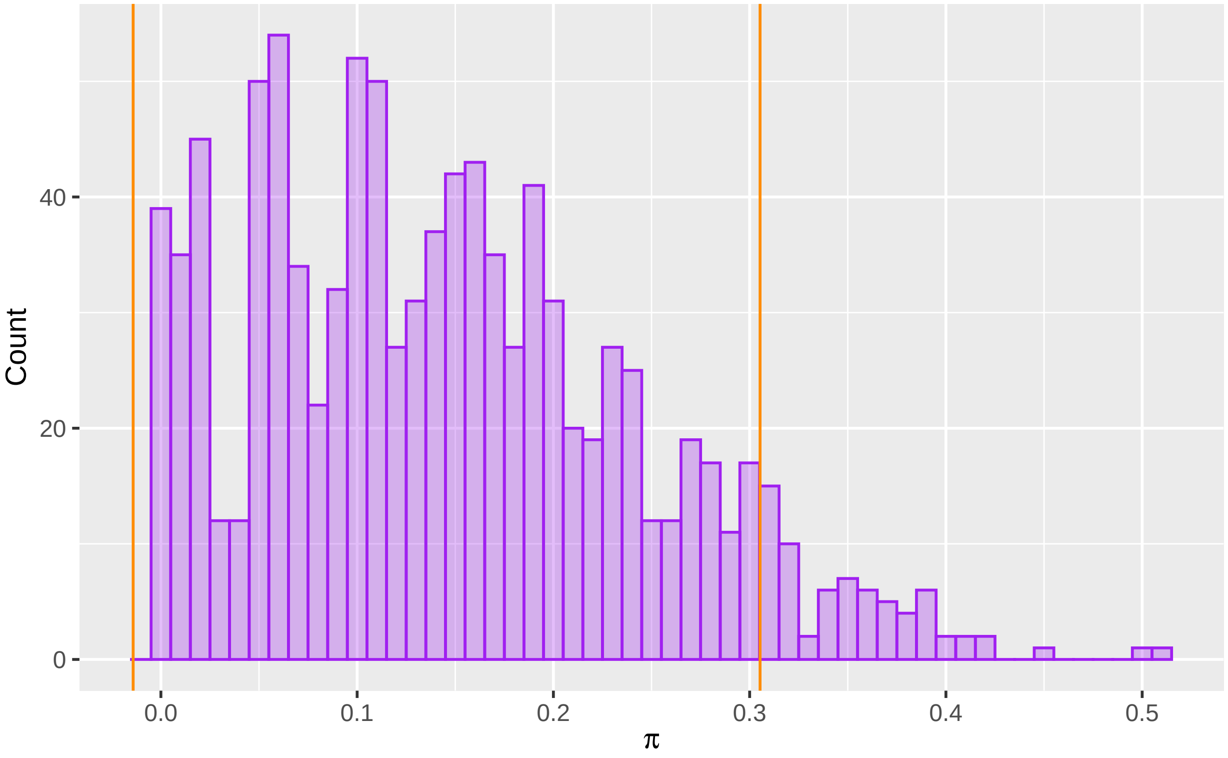
Figure 7. Histogram of Bootstrapped Probabilities of Failure at 31°F



Vertical lines denote the 90% CI

```
plt_72
```

Figure 8. Histogram of Bootstrapped Probabilities of Failure at 72.27°F



Vertical lines denote the 90% CI

We can visualize a histogram of the results of the bootstrap (100,000 trials each) for 31°F in **Figure 7** and for 72°F in **Figure 8**. Using data in these figures, we can calculate that:

- The probability of an `O.ring` failure at 30°F are between 0.840054 and 1.1296322 (with a 90% C.I.)
- The probability of an `O.ring` failure at 72.27°F are between -0.0141308 and 0.3053071 (with a 90% C.I.)

While these probabilities may be calculated to sometimes be less than one or greater than zero, it is known their true bounds are between [0, 1].

(f) Determine if a quadratic term is needed in the model for the temperature.

We'll create another glm model with a quadratic term and then compare the two models with an anova test as below:

```
fail_model_quad <- glm(formula = distress ~ Temp + I(Temp^2), data = d,
                       family = binomial(link = logit))
anova_mod <- anova(fail_model, fail_model_quad , test = "Chisq")

p_val<- anova_mod$`Pr(>Chi)`[2]
```

The p-value from the anova test is 0.3357766, which is fairly large and greater than 0.05. Thus, we conclude that there is not enough evidence that the quadratic model is an improvement over the the linear model.

Part 4 (10 points)

With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results, conduct model diagnostics and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case? Explain why.

```
linear_model <- glm(distress ~ Temp, data = d, family = gaussian())
```

Shown above, the linear model also finds that temperature is significant for the presence of a O-ring distress event. Like the previous models there is an inverse relationship between the two variables, where the likelihood of a distress event increases as the temperature decreases. Specifically, with all else constant, increasing the temperature by 1°F would decrease the probability of a distress event by ~3.74%.

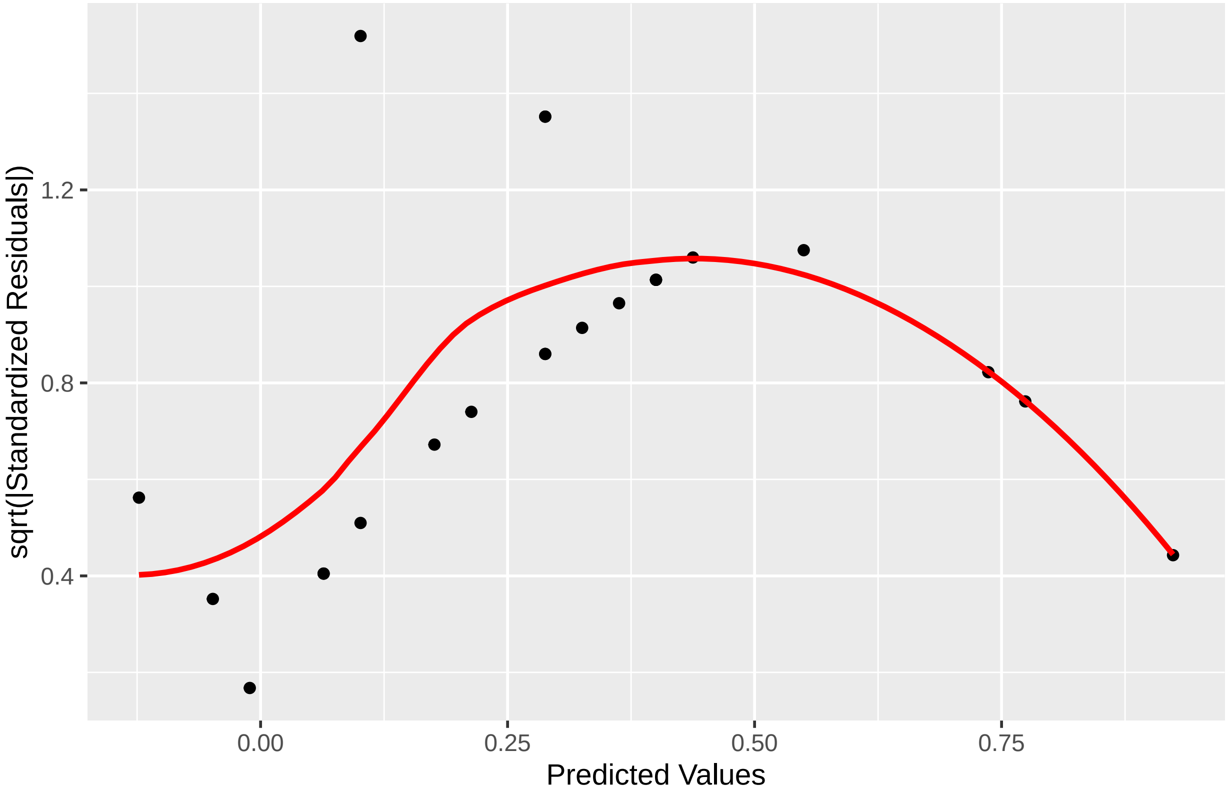
Even though a similar result is found, this model should not be used instead of the binary logistic regression model. This is because the linear model does not have bounds on the output, so that sometimes a probability greater than one or less than zero may be predicted. In this case, a probability of one is reached at about 50.96°F while a probability of zero is reached at about 77.71°F. Given that there were recorded launches both above and below these temperatures, this is an undesirable property of the model.

In addition, the linear model assumes homoscedasticity, or that the variance in the model's residuals is constant both over the range of the output as well as the domain of its inputs.

```
linear_model %>%
  ggplot(aes(x = linear_model$fitted.values,
             y = sqrt(abs(linear_model$residuals/sd(linear_model$residuals))))) +
  geom_point() +
  stat_smooth(color="red", se=FALSE) +
  labs(
    title = "Figure 9. Predicted Probabilities vs. Standardized Residuals",
    x = "Predicted Values",
    y = "sqrt(|Standardized Residuals|)")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Figure 9. Predicted Probabilities vs. Standardized Residuals



Shown in *Figure 9*, the standardized residuals are not constant over the range of the output variable in the model. This does not meet the homoscedasticity assumption, because the output variable is itself a function of the input variables. This result was expected, however, as it can be derived that the variance in a random variable from a binomial distribution is a function of that variable itself, or $Var(\pi) \propto \pi * (1 - \pi)$. It can be further seen that the maximum variance is observed around $\pi \approx 0.5$, which follows the prediction from the variance equation. Given that the homoscedasticity assumption is not met, there is further evidence of preferring the logistic model to the linear regression.

A final drawback of the linear model is that, as shown in its name, it only considers linear combinations of the input variables to predict the output variables. However, the logistic regression model also follows this linear combination form, and so is not any more or less desirable for this reason alone. However, given that in this scenario the previous two drawbacks were so large, the logistic regression model should be preferred instead.

Part 5 (10 points)

Interpret the main result of your final model in terms of both odds and probability of failure. Summarize the final result with respect to the question(s) being asked and key takeaways from the analysis.

The final model is the logistic regression which uses temperature as the single explanatory variable. The model is of the form

$$\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1(\text{Temp})$$

```
beta_0_hat <- fail_model$coefficients[1]
beta_1_hat <- fail_model$coefficients[2]
```

The intercept in the model, β_0, had the value 15.0429016 while the coefficient of `Temp` had the value -0.2321627. The coefficient on the `Temp` variable can be interpreted as decreasing the odds of any O-ring failures by 0.7928171 with every 1°F increase in temperature and all else constant.

This is a key finding. It underscores the importance git status of atmospheric temperature on the day of the launch, and that lower temperatures would increase the odds of an O-ring failure occurring. This is supported by Dalal et al. when they state the resiliency of an O-ring is highly dependent on temperature.

On the day of the launch the temperature was 31°F. At this point, the predicted the probability of any O-ring failure is 0.9996088, with a 95 % confidence interval (LRT) as 0.8036982 and 1.

Therefore, with the 95% confidence interval so high, at least one O-ring is very likely to fail under these conditions. If this was known, the launch could have been postponed for warmer weather when there was a lower probability of a failure. We can conclude that the catastrophe could have been avoided had NASA managers appropriately analyzed the data by factoring in both the successful and failed launches.