Space Shuttle Safety and the *Challenger*
Statistical Modeling for Critical Decisions

Aidan Jackson, Sandip Panesar, Devesh Khandelwal

Contents

Investigation of the 1986 Space Shuttle *Challenger* Accident

Summary

This exercise examines the Space Shuttle *Challenger*'s 1986 disaster using the previous safety record of the Space Shuttle Program. The probability of O-ring failures at various launch temperatures are modeled to make a prediction for the conditions on the morning of the disaster. It is found that the probability of a catastrophic failure is 34%, over three times higher than when using the methods of the original estimate. The Rogers Commission, tasked with investigating the accident, noted that this underestimate occurred due to omitting launch data where no O-ring damage was found.[2] The original estimate procedure was used by Morton Thiokol, a contractor to NASA and producer of the O-ring involved component that failed.

This work was originally completed as part of the W271 Statistical Methods for Discrete Response, Time Series, and Panel Data course in the Master of Information and Data Science program at University of California, Berkeley. The exercise in turn was based on a previous 1989 paper on the same topic, originally published in the Journal of the American Statistical Association and included in this repo as a pdf.[1]

Background

On January 28th, 1986, the *Challenger* launched from Cape Canaveral, Florida, for what was the 25th orbital mission of the Space Shuttle program. 59 seconds after launch, a tracking camera recorded a burning plume coming out of the side of the right solid rocket booster (SRB). As the leaking fuel continued to burn, gradually other parts of the shuttle stack began to disintegrate. 73 seconds after launch, still visible in the sky, the shuttle exploded as different fuel sources ignited. This was particularly shocking both for being the first American astronaut deaths in a launch mission[1], as well as its public broadcasting as part of the inaugural Teacher in Space Project.

In the ensuing investigation, the cause of the accident was found to be O-ring failure in the field-joint of the same right SRB.[2] O-rings are doughnut-shaped and commonly made of rubber, and provide a flexible seal between mechanical components. In a typical car engine, O-rings have diameters on the scale of inches. Inside the Space Shuttles' SRBs, O-ring diameter was 37.5 feet.[1]

Like many other applications, the SRBs' O-rings sealed against gas to prevent it from escaping where mechanical joints connected. Every Shuttle launch required 2 SRBs, with each SRB containing 3 field-joints that used O-ring seals. For redundancy, each field-joint had both a primary and a secondary O-ring. This resulted in 6 field-joint O-rings total per SRB, or 12 field-joint O-rings total per launch.

Redundancy is needed because of the potential for an O-ring to fail. There are two mechanisms which could cause this, known as erosion and blow by. Erosion has the potential to occur during normal operation, where hot gases sealed against the O-ring gradually disintegrate it. On the SRBs, putty was added to provide a protective barrier against this mechanism. Blow by occurs when an O-ring already fails to provide a complete seal, allowing high velocity gas to move past the O-ring and damage it further.

---

[1]Apollo 1 resulted in 3 astronaut deaths during pre-launch training on the ground. Soyuz 1 and 11 resulted in a combined 4 cosmonaut deaths due to hardware failures during each launch mission.

Although the secondary O-rings were a backup should distress and failure occur in a primary O-ring, before the first ever Space Shuttle launch it was discovered that additional problems could also occur. This was due to field-joint design which sometimes caused unintended rotation among different components, including the O-rings. While the primary O-rings would engage a seal reliably even with this rotation, the secondary O-rings specifically could sometimes slip and leave permanent space for gas leakage. If this occurred during a launch where the primary O-ring failed due to the previous distresses, then it was known the result would be a "catastrophic failure".[1]

Being made of rubber-like material, there are a variety of conditions which may impact the ability of any O-ring to provide a complete seal and suffer distress. One which was investigated for the Space Shuttle program was temperature, where colder conditions can cause an O-ring to shrink and lose its elasticity. During the late 1970s and early 1980s, the SRB engine with the O-rings was qualified for temperatures as low as 40°F.

Without any additional statistics, when the Challenger started its Winter morning launch it was 36°F.

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

Exploratory Data Analysis

As with any dataset, understanding what information is available is a prerequisite to deciding on what value to extract from it. Although the size of this dataset is small, with only 23 observations from previous Shuttle launches, the choice of how to treat it is equally as important and influential to later analysis.

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

grid.table(head(d), rows = NULL)
```

| Flight	Temp	Pressure	O.ring	Number
1	66	50	0	6
2	70	50	1	6
3	69	50	0	6
4	68	50	0	6
5	67	50	0	6
6	72	50	0	6

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 provides an opportunity to more thoroughly examine all aspects of the dataset.

To begin, the meaning of each variable must be properly understood. `Flight` represents a generic index for each launch. `O.ring` represents the number of primary field-joint O-rings which experienced a "thermal distress" event for each flight. A thermal distress event was defined as when either the effects of erosion or blow by could be observed in the recovered SRBs post launch. Although every launch had 12 field-joint O-rings in total, only six of these are primary while the second six are secondary. This was to design for redundancy, even though mechanisms were found which could lead to the immediate failure of the secondary O-rings should the primary fail as mentioned in **Background**. As confirmation of the total, the variable `Number` represents the count of primary field-joint O-rings and is constant at six for all launches. `Temp` and `Pressure` represent the recorded field-joint temperature in Fahrenheit and pressure in psi respectively. `Temp` is recorded to the nearest degree, while `Pressure` is recorded to the nearest 50 psi (e.g. 50, 100, 150, etc). With a much coarser grain of pressure measurements, it may be less meaningful since less specific information is recorded.

Being the cause of the accident, the `O.ring` variable will be examined to start. In this analysis, unless stated otherwise "O-ring" will refer specifically to a primary field-joint O-ring.

```
ggplot(d, aes(O.ring)) +
  geom_histogram(binwidth = 1, fill = "black", color='black', alpha=0.8) +
  ggtitle("Figure 1. O-Ring Distribution") +
  xlab("Count of O-Ring Thermal Distress Events per Launch") +
  ylab("Count")
```

Figure 1. O−Ring Distribution
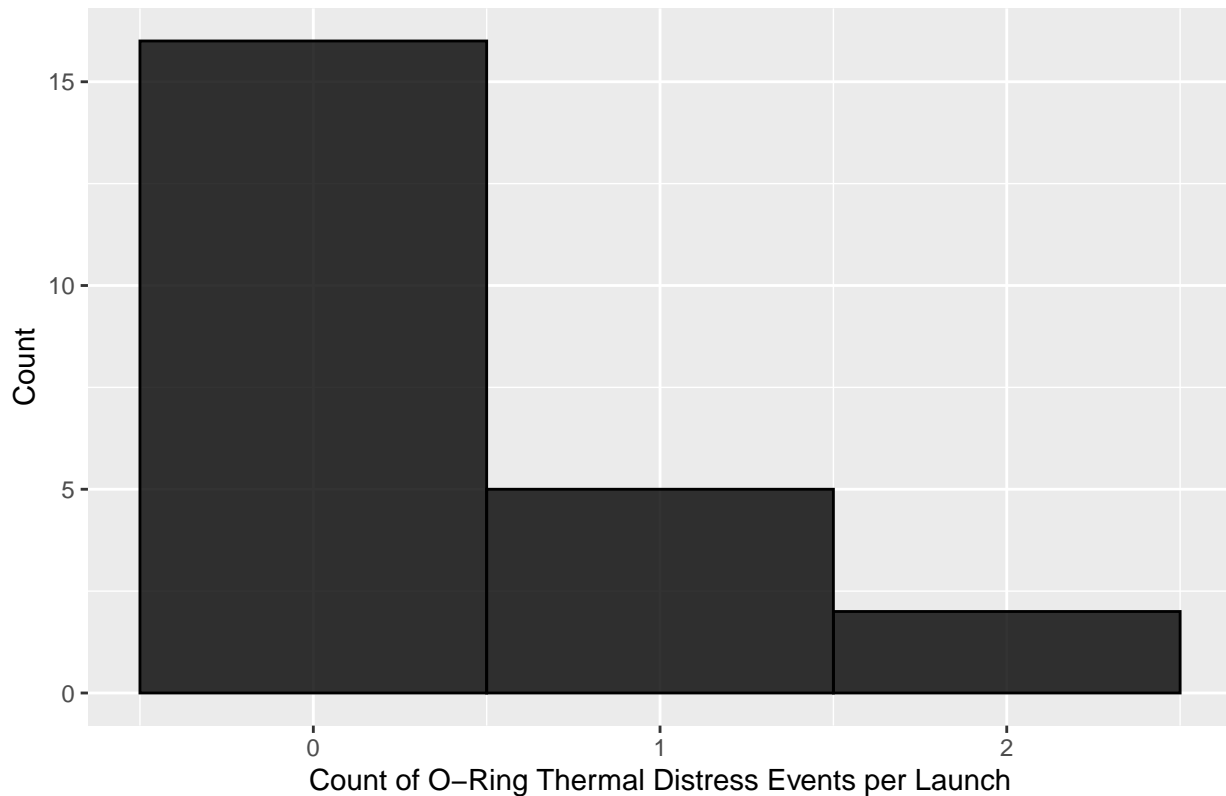
Shown above in **Figure 1**, the majority of launches had O-rings which experienced no thermal distress events. About 5 launches had a single O-ring experience thermal distress, while even fewer had two. None of the recorded launches had greater than two O-rings distress events. However, for a catastrophic failure to occur, all that was potentially needed would be a single primary O-ring failure to occur when its secondary backup had rotated and lost contact for a seal.

With the majority of launches not experiencing O-ring distress, those that experienced any distress at all can be grouped together for a more discrete analysis.

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

grid.table(tail(d), rows = NULL)
```

Flight	Temp	Pressure	O.ring	Number	distress
18	81	200	0	6	0
19	76	200	0	6	0
20	79	200	0	6	0
21	75	200	2	6	1
22	76	200	0	6	0
23	58	200	1	6	1

To better differentiate between these two situations, a new Boolean variable named `distress` 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.

```
ggplot(d, aes(factor(distress), Temp)) +
  geom_boxplot(aes(fill = factor(distress))) +
  scale_fill_brewer(palette = "Set1", direction = -1) +
  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
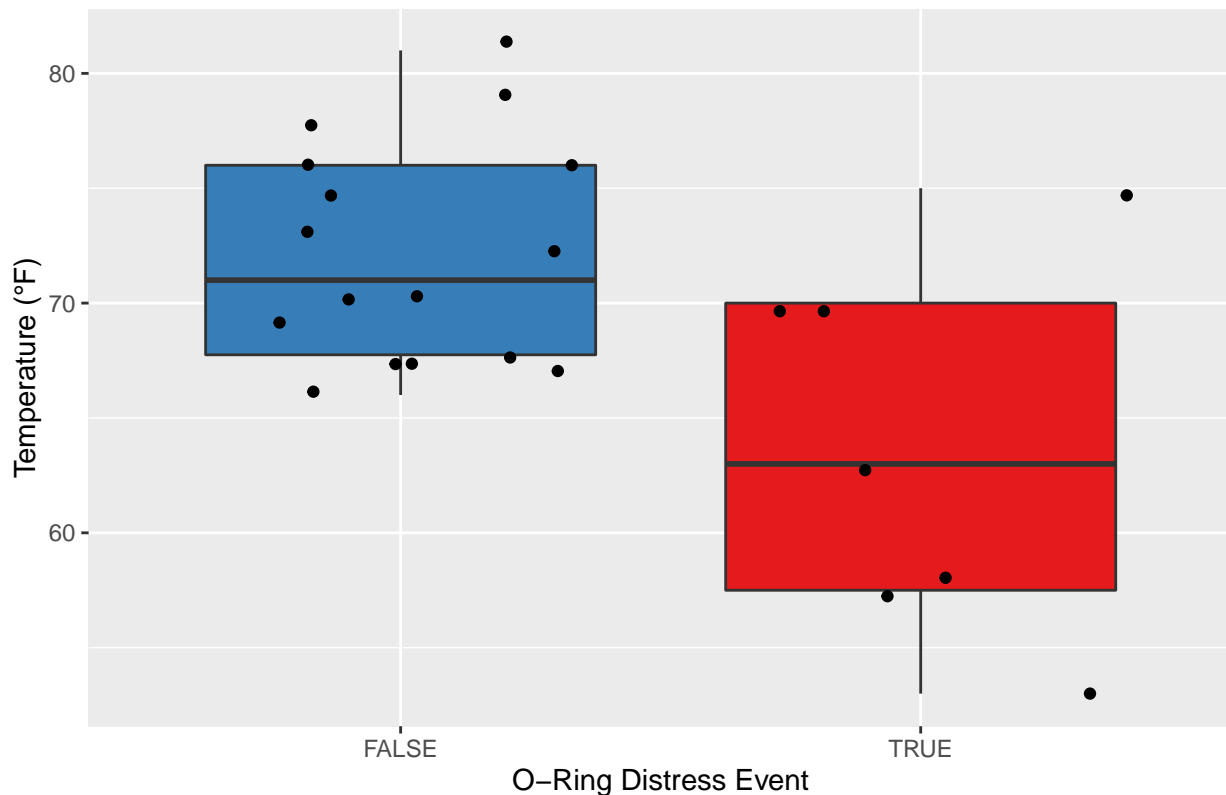
Demonstrated in **Figure 2**, O-Ring distress events generally occurred at lower launch temperatures. In fact, with so few total observations, it can be seen that over half of the launches with O-ring distress were at lower temperatures than the coldest launch with no O-ring distress There is a single 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. Note that because of this point, along with generally having a wider distribution, there is greater temperature variance among launches which had distress events compared to those which did not.

The other environmental variable in the data set, `Pressure`, may also be examined to see if insightful relationships emerge.

```
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 = "blue", high = "red") +
  scale_y_continuous(breaks = c(0, 1, 2)) +
  scale_x_continuous(breaks = c(50, 60, 70, 80))
```
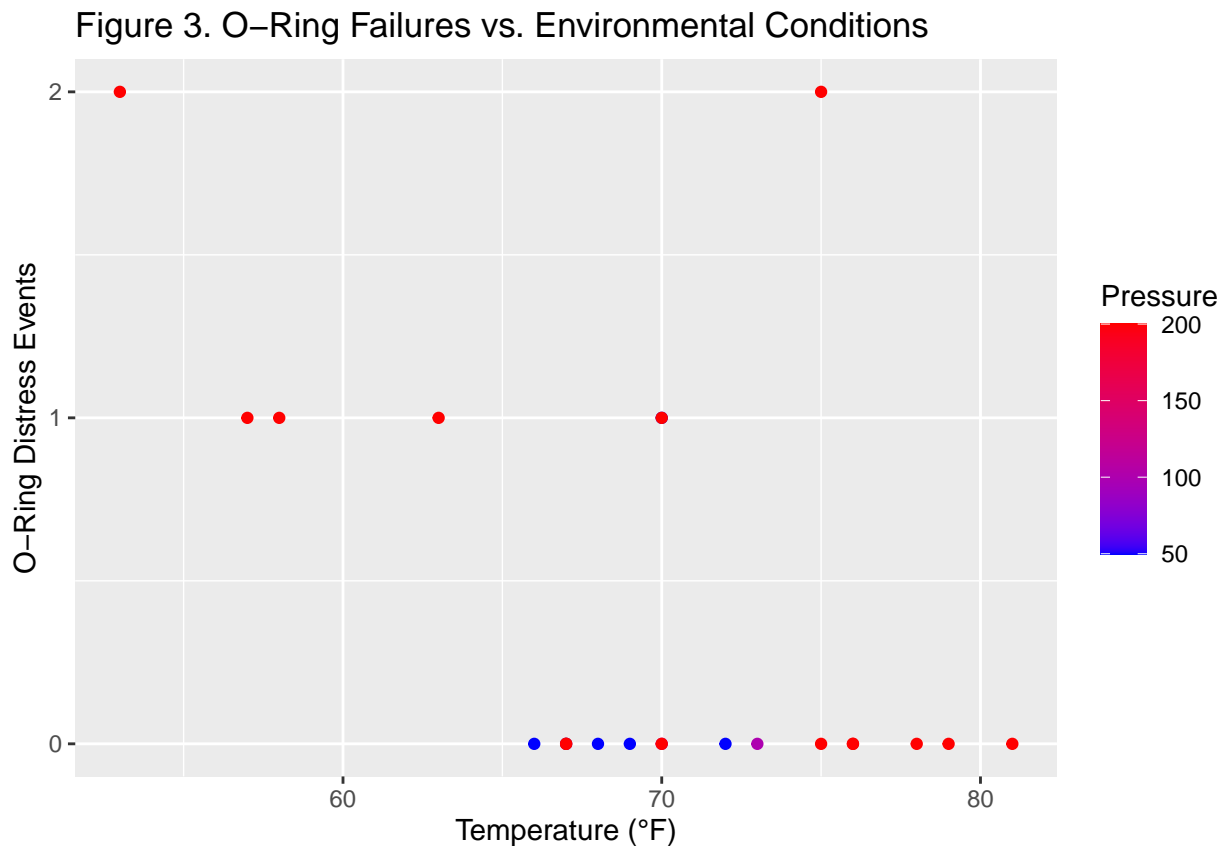

Figure 3. O–Ring Failures vs. Environmental Conditions

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 at the highest pressure measurement of 200 psi, the general trend does not seem as strong as with temperature. For example, 200 psi pressure measurement 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. Although pressure is binned into multiples of 50 psi, resembling a categorical or ordinal variable, the ratio of one pressure to another is still meaningful. For example, 200 psi is known to be four times as much pressure as 50 psi. A pressure of zero psi is also meaningful, meeting

the definition for a ratio variable. Since theoretically a pressure can take on any value, even if only multiples of 50 are recorded, it will still be treated as a continuous variable. The count of O-rings which experienced distress meets the same definition of a ratio variable but is by nature discrete and not continuous. For example, 1.5 O-rings cannot experience distress. Therefore, Spearman's rank correlation coefficient will be used.

```
cat("Table 1. Correlation of Numerical Variables\n")
```

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

```
# generate correlation table without certain columns
cor(d[ , !c("Flight","Number","distress")], method = "spearman")
```

```
##                Temp  Pressure      O.ring
## Temp      1.0000000 0.1541326 -0.4294734
## Pressure  0.1541326 1.0000000  0.2775559
## O.ring   -0.4294734 0.2775559  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 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 had 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.

Modeling

With the given context, there are different potential approaches to estimate the relationship between the environmental variables and O-ring distress. Given that each O-ring is discrete, a Poisson or Binomial distribution could be used depending on the scenario and modeling goals. The Poisson distribution describes processes which, among other things, occur over an interval at a constant average rate. This could be used to model the expected count of O-rings which fail for a series of Space Shuttle launches. However, the focus on the average over time is not particularly useful to estimate failure in a specific future launch. The Binomial distribution, on the other hand, describes processes which have a sequence of binary events, such as success or failure of multiple O-rings within a single Shuttle launch. This would be much more fitting for predicting a probability of failure on a single launch in the future, such as the *Challenger* at the start of its final mission. Because of this, only the Binomial distribution will be used in this investigation.

For both the Poisson and Binomial distributions, however, an assumption of independence is required. This would require that each Space Shuttle launch is independent and identically distributed

(IID) compared to all others. In general this should be true, although there may be minor reasons why this is violated. For example, over time Shuttle technology or launch procedures may have been improved based on prior experience. This would be equivalent to a distributional change over time. In addition, part of the reason SRBs were recovered were not just for observing O-ring damage, but also for refurbishment and potential re-use in later launches. Having sets of launches with the same SRBs could result in sets of correlated observations that are different from other flights. It could also be expected that re-used SRBs could in general cause greater O-ring distress than those being newly used. This would be a notable violation since it can involve the observed outcome of O-ring damage along with its relationship to environmental conditions. With either a Poisson or Binomial distribution, it must be assumed to be negligent. In the future, knowledge of which SRBs were re-used for which launches could be incorporated into modeling for a more accurate estimate.

These modeling options also require that in addition to each Shuttle launch being IID, each of the six O-rings for any given flight must be IID as well. If the O-rings were not independent, the damage of a given O-ring would not only be a function of the environmental 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 environment, but multiple relationships depending on which O-ring was under consideration.

Just like with the IID Shuttle launches, this is avoided with the assumption of independence however it may not be completely valid. For example, the six O-rings are spread out across two SRBs for each launch. It could be expected that there are small differences from between specific SRBs with respect to the stress the O-rings experience. For example, if an O-ring in one SRB had minor manufacturing differences that put more stress on its O-rings then these observations would be correlated. The three O-rings in the other SRB, however, may have their own correlation based on its own unique qualities. Therefore, in reality there is likely some relationship between the O-rings within a single SRB that is ignored for the sake of the model.

```
# Determine fraction of O-rings that failed for each launch
d$O.ring.fraction <- d$O.ring / d$Number
```

With the assumption of independence, two different uses of the Binomial distribution will be investigated. The first will model $n = 6$ IID O-rings for flight, each with an identical probability p of experiencing distress. The set of O-rings within each launch and the series of all Shuttle launches will then make up a Bernoulli process. This will be referred to as the "full" outcome throughout the rest of this report. The second will investigate only the probability of a launch experiencing *any* O-ring distress by use of the binary `distress` variable created during the EDA. This will model the reduced $n = 1$ Bernoulli trial of whether at least one O-ring was distressed on a launch, with a single probability p of this occurring. This is a special case of the Binomial distribution known as the Bernoulli distribution, and will be referred to as the "binary" outcome throughout this report.

An advantage of the former is that it uses all available information in the dataset, potentially estimating a more accurate relationship between O-ring failure and environmental conditions. It also can provide a more useful prediction for the outcome of the *Challenger* disaster, since any design-flaw failure in a secondary O-ring when the primary is distressed would cause a catastrophe. Having an estimate of a specific number of primary O-rings which fail would then be proportional to the probability of the catastrophe. The use of the binary `distress` outcome, however, only estimates if at least one O-ring will fail without regard to how many. Still, because of its simpler

nature it can potentially be more precise when the number of observations is so low. Therefore, the full model will be used to provide a primary estimate of the disaster's probability, with the binary model used as a spot-check against its results.

There are two explanatory variables of interest, temperature and pressure. To test whether either of these are significant, a series of three models will be created.

$$logit(\pi) = \beta_0$$

The first, shown above, will only include an intercept and serve as the null hypothesis where neither temperature nor pressure have a relationship with O-ring distress. This equivalent to $\beta_T = \beta_P = 0$.

As common with Binomial distributions, a logit link function is used to relate the dependent and independent variables. The logit is defined as $logit(\pi) = ln(\frac{\pi}{1-\pi})$ where π is the probability of the outcome of interest. In the full model, this is the probability of failure of a single O-ring while in the binary model it is the probability of failure in at least one O-ring in a launch.

$$logit(\pi) = \beta_0 + \beta_T * T$$

The second, shown above, will include temperature as the sole independent variable denoted by T. This is motivated by Morton Thiokol's pre-launch investigation of O-ring condition deteriorating at lower temperatures using the same dataset. Later, it will be shown how their mistakes in statistical modeling lead to an underestimate of the probability of catastrophe.

$$logit(\pi) = \beta_0 + \beta_T * T + \beta_P * P$$

Finally, the pressure P will be included along with T in the third model to see if both environmental conditions are also significantly related to O-ring distress.

```r
# Null hypothesis, no relationship between env and o-rings
binary_null <- glm(formula = distress ~ 1, data = d,
                   family = binomial(link = logit))

# First alternative hypothesis, only temp has relationship
binary_temp <- glm(formula = distress ~ Temp, data = d,
                   family = binomial(link = logit))

# Second alternative hypothesis, both env variables have relationship
binary_both <- glm(formula = distress ~ Temp + Pressure, data = d,
                   family = binomial(link = logit))

# Null hypothesis, no relationship between env and o-rings
binom_null <- glm(formula = O.ring.fraction ~ 1, data = d,
                  family = binomial(link = logit), weights = Number)

# First alternative hypothesis, only temp has relationship
binom_temp <- glm(formula = O.ring.fraction ~ Temp, data = d,
                  family = binomial(link = logit), weights = Number)
```

```
# Second alternative hypothesis, both env variables have relationship
binom_both <- glm(formula = O.ring.fraction ~ Temp + Pressure, data = d,
                  family = binomial(link = logit), weights = Number)
```

The same set of three models will be used for both the full and binary outcome variables, shown above. Note that for the full models, instead of creating $n = 6$ unique data points for the O-rings of each launch, the fraction of O-ring failures in a single launch was used as the dependent variable. When paired with the weights of each observation, which is constant at six O-rings for all launches, this produces the same point estimate of model coefficients.

A likelihood ratio test (LRT) will be used to determine whether a model associated with an alternative hypothesis is a significant improvement over the null hypothesis model. If the first alternative hypothesis is significant, i.e. involving only T, then the second alternative hypothesis including both T and P will be assessed against it.

LRTs compare the likelihood of the alternative hypothesis model against the null hypothesis model. In general, an LRT is 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:

- Λ is the likelihood ratio test statistic

- $L()$ is the likelihood function of the estimated model given the data, which in this case involves cross entropy between the predicted probabilities and binary true values.

- $\hat{\pi}_i^{(0)}$ is the estimated probability of success of Bernoulli trial i under the null hypothesis H_0

- $\hat{\pi}_i^{(a)}$ is the estimated probability of success of Bernoulli trial i under the alternative hypothesis H_a

```
alpha <- 0.05
significance_level <- alpha / 4
```

Because of these planned, repeated hypothesis tests, the Bonferroni correction will be applied to strengthen the level of significance required to reject the null hypothesis. Starting from $\alpha = 0.05$, with four comparisons between the alternative and null hypothesis models, the new p-value required for significance will be $\alpha = 0.05/4 = 0.0125$.

```
binary_temp_test <- anova(binary_null, binary_temp, test = "LRT")
binom_temp_test <- anova(binom_null, binom_temp, test = "LRT")
```

When performing the LRTs including only temperature in the alternative hypothesis, p-values of 0.0048 and 0.0132 are found for the binary and full models respectively. The binary model is definitively past the threshold required for significance of 0.0125, but the full model is very slightly above. While this would normally result in a failure to reject the null hypothesis, often times

the choice of interpretation is dependent on the testing context and broader considerations. In particular, methods for defining significance should be viewed as guidelines rather than hard and fast rules. In this case, the Bonferroni correction is known for being overly conservative in that it guarantees the family wise Type I error rate will *always* be less than the desired level of α. In particular, strict adherence to the Bonferroni correction is known for higher Type II error rates, or failure to reject a false null hypothesis. The significance of temperature in the binary model also suggests this could be the case if judging the borderline full model as non-significant. Therefore, for being nearly at the threshold and with these considerations, this result will still be taken as significant.

```
binary_pressure_test <- anova(binary_temp, binary_both, test = "LRT")
binom_pressure_test <- anova(binom_temp, binom_both, test = "LRT")
```

With the first pair of alternative hypotheses being failed to be rejected, they now will take the position of the null hypothesis for comparison against including pressure. For these LRTs, p-values of 0.2156 and 0.2145 are found for the binary and full models respectively. Unlike the previous borderline case, both of these are well above the Bonferroni corrected significance level as well as the original α. Therefore, the null hypothesis that pressure does not have a significant relationship with O-ring condition is failed to be rejected. Moving forward, the temperature-only models will be used for further investigation and prediction.

Before use of the models for specific tasks, the fitted parameters within the model should also be examined. In this case, β_T may be interpreted for the relationship between temperature and O-ring failure. When doing so with any link function other than the identity link, changes in an independent variable must account for additional transformation(s) to find the change in the dependent variable.

$$OR = \frac{Odds_{T+c}}{Odds_T} = \frac{\pi_{T+c}}{1 - \pi_{T+c}} * \frac{1 - \pi_T}{\pi_T} = \frac{exp(\beta_0 + \beta_T * (T + c))}{exp(\beta_0 + \beta_T * T)} = exp(\beta_T * c)$$

With a logit link function, changes in an independent variable are most commonly interpreted through the odds ratio (OR), with a derivation for this context shown above. Specifically, the example shown describes the OR as a function of a c °F increase in temperature. Previously, the coefficient of temperature β_T was estimated to be about -0.232 in the binary model and -0.116 in the full model. Both are negative, implying that generally as temperature increases then the estimated odds of O-ring failure decreases. Conversely, as temperature decreases then the estimated odds of O-ring failure increases.

```
c <- -10
binary_or <- exp(c*binary_temp[["coefficients"]][["Temp"]])
binom_or <- exp(c*binom_temp[["coefficients"]][["Temp"]])
```

For a 10°F decrease in temperature, the OR would be about 10.2 in the binary model and 3.2 in the full model. Put into context, this implies that with a 10°F lower temperature the odds of *any* O-ring failure increases by 10.2 and the odds of a specific O-ring failure increases by 3.2. This change is irrespective of the starting temperature, and as the change in temperature grows more negative then the odds of O-ring failure will continue to increase.

For future plotting, visualization of confidence intervals will sometimes be included. While the previous hypothesis tests were LRTs, when this is done the Wald confidence intervals will be used instead.

```r
# create sequence of temperatures to show on plot
temps <- data.frame(Temp = seq(30,82,1))
temp_challenger <- 36

# make predictions on the sequence of temperatures
binary_prob <- predict(object = binary_temp, newdata = temps,
                       type='response', se=TRUE)
binom_prob <- predict(object = binom_temp, newdata = temps,
                       type='response', se=TRUE)

# pi values
binary_pis <- binary_prob$fit
binom_pis <- binom_prob$fit

# 2.5th percentile
binary_lower_ci <- binary_pis - qnorm(p = 1 - alpha/2)*binary_prob$se.fit
binom_lower_ci <- binom_pis - qnorm(p = 1 - alpha/2)*binom_prob$se.fit
# 97.5th percentile
binary_upper_ci <- binary_pis + qnorm(p = 1 - alpha/2)*binary_prob$se
binom_upper_ci <- binom_pis + qnorm(p = 1 - alpha/2)*binom_prob$se

binary_df <- data.frame(temp = temps$Temp, pi = binary_pis,
                        lwr = binary_lower_ci, upr = binary_upper_ci)
binom_df <- data.frame(temp = temps$Temp, pi = binom_pis,
                       lwr = binom_lower_ci, upr = binom_upper_ci)

binary_color <- "darkgreen"
binom_color <- "blue"

ggplot() +
  geom_line(data = binary_df, aes(x = temp, y = pi, color = binary_color),
            show.legend = TRUE) +
  geom_line(data = binary_df, aes(x = temp, y = lwr),
            color = binary_color, linetype = "dotted") +
  geom_line(data = binary_df, aes(x = temp, y = upr),
            color = binary_color, linetype = "dotted") +
  geom_line(data = binom_df, aes(x = temp, y = pi, color = binom_color),
            show.legend = TRUE) +
  geom_line(data = binom_df, aes(x = temp, y = lwr),
            color = binom_color, linetype = "dotted") +
  geom_line(data = binom_df, aes(x = temp, y = upr),
            color = binom_color, linetype = "dotted") +
  ylim(0,1) +
  ggtitle("Figure 4. Probability of O-Ring Failure by Temperature") +
```

```
ylab(expression(pi)) +
xlab("Temperature (°F)") +
geom_hline(yintercept = 1.0, linetype = "dotted") +
geom_hline(yintercept = 0.0, linetype = "dotted") +
geom_vline(xintercept = temp_challenger, linetype = "solid") +
geom_point(data = d, aes(x = Temp, y = ifelse(distress==TRUE, 1, 0)),
           shape = 17) +
scale_color_identity(name = "Model",
                     labels = c("Binary", "Full"),
                     breaks = c(binary_color, binom_color),
                     guide = "legend")
```
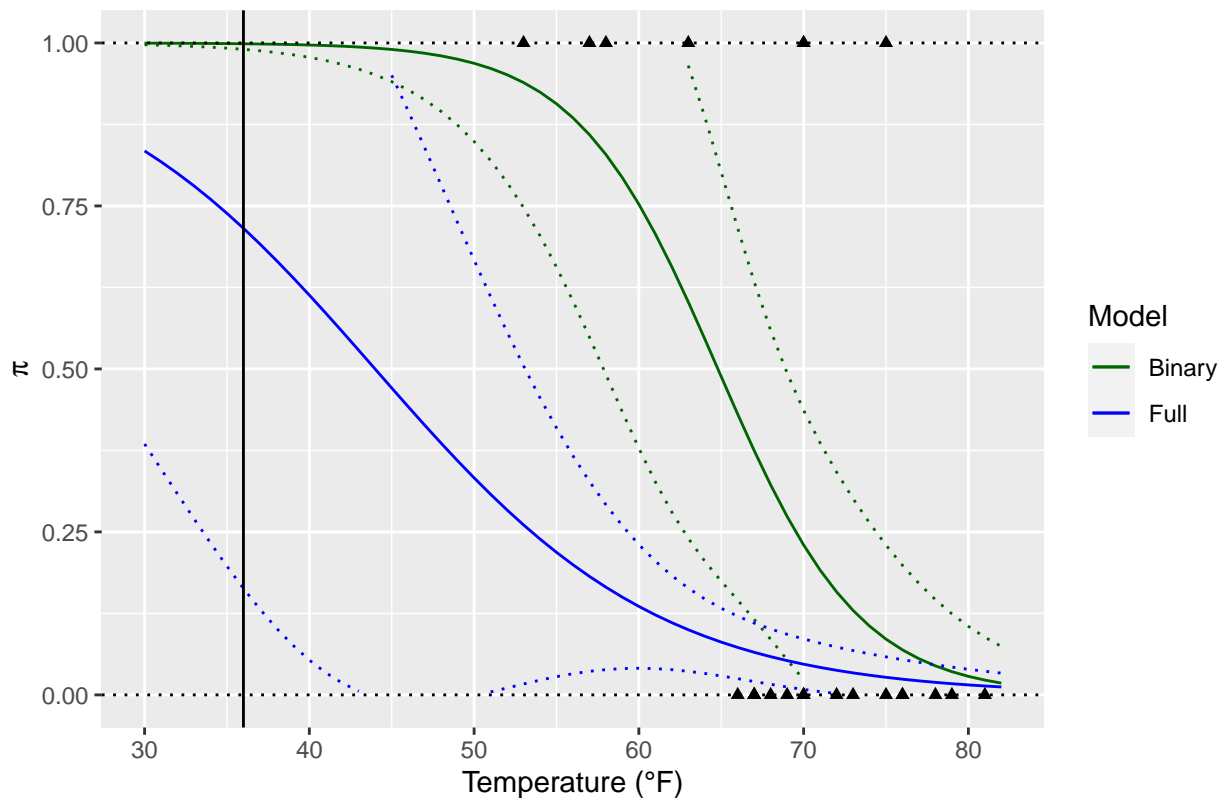

Figure 4. Probability of O−Ring Failure by Temperature

Figure 4 shows the probability of O-ring failure from both models over a range of temperatures. The Wald confidence intervals with $\alpha = 0.05$ are shown as dotted lines for each model. Previous Shuttle launches are also included depending whether they did or did not have at least one O-ring failure, which is the dependent variable in the binary model. Finally, the vertical line represents the launch temperature of the *Challenger* disaster of 36°F.

As described previously the binary case represents the probability that *any* of the six O-rings will fail during launch, while the full case represents the probability of a single O-ring within the six failing. Shown earlier with the OR, a decrease in temperature generally increases the probability of O-ring failure in both models. This is visualized here as the probabilities from both models increase at changing rates as temperature decreases. The binary model has a larger increase in probability at moderate temperatures than the full model, which is intuitive as it only tracks whether at least

14

one out of six O-rings will fail rather than the probability of a single specific O-ring.

At the launch temperature of the *Challenger*, the binary model predicts there is a near guarantee that at least one O-ring will fail. The full model likewise predicts a high probability of each individual O-ring failing at nearly 75%. With the O-rings being assumed IID as required for the Binomial distribution, the expected number of O-ring failures on a single launch is then given by:

$$E[Y] = n * p = 6 * \hat{\pi}(T)$$

where $n = 6$ from the six IID O-rings on each launch and $p = \hat{\pi}(T)$ is the estimated probability of an individual O-ring failing at a specific temperature as given by the full model. Again, by design only the full model can estimate how many O-rings on a given launch may fail as the binary model is only useful for understanding if *at least* one O-ring will fail, without regard to the total number.

The key feature of Morton Thiokol's original analysis was that all launches which had zero O-ring failures were omitted.[1] This was justified by assuming they had nothing to offer to an understanding why O-rings may fail, or if temperature could be related. However, to falsify the claim that temperature is related to O-ring failure, both successes and failures of O-rings must be considered. Attempting to do so only based on examples of O-rings which failed, or conversely O-rings which did not fail, is impossible. Especially when a causal effect is suspected, an analysis which attempts to do so does not meet fundamental criteria of science.

```
d_morton <- d[d$O.ring > 0, ]
binom_temp_morton <- glm(formula = O.ring.fraction ~ Temp, data = d_morton,
                         family = binomial(link = logit), weights = Number)
binom_prob_morton <- predict(object = binom_temp_morton, newdata = temps,
                             type='response', se=TRUE)
binom_pis_morton <- binom_prob_morton$fit
binom_lower_ci_morton <- binom_pis_morton -
  qnorm(p = 1 - alpha/2)*binom_prob_morton$se.fit
binom_upper_ci_morton <- binom_pis_morton +
  qnorm(p = 1 - alpha/2)*binom_prob_morton$se.fit
binom_df_morton <- data.frame(temp = temps$Temp, pi = binom_pis_morton,
                              lwr = binom_lower_ci_morton,
                              upr = binom_upper_ci_morton)
binom_color_morton <- "red"
```

An identical subset of the launch data as used by Morton Thiokol is created above. With this, the same structure of the full model can be fit using to determine how it may impact analysis.

```
mode <- function(v) {
  #' Function to find the statistical mode of a vector and return the result
  #' Inputs:
  #'  vector v: vector to find the mode of
  #' Outputs:
  #'  statistical mode of the vector v
  uniqv <- unique(v)
  return(uniqv[which.max(tabulate(match(v, uniqv)))])
```

```
}

num_O_rings <- mode(d$Number)

# estimated number of failed O-rings by temperature
binom_df$num_failed_est <- num_O_rings*binom_df$pi
binom_df$num_failed_lwr <- num_O_rings*binom_df$lwr
binom_df$num_failed_upr <- num_O_rings*binom_df$upr
binom_df_morton$num_failed_est <- num_O_rings*binom_df_morton$pi
binom_df_morton$num_failed_lwr <- num_O_rings*binom_df_morton$lwr
binom_df_morton$num_failed_upr <- num_O_rings*binom_df_morton$upr

ggplot() +
  geom_line(data = binom_df,
            mapping = aes(x = temp, y = num_failed_est, color = binom_color)) +
  geom_line(data = binom_df, aes(x = temp, y = num_failed_lwr),
            color = binom_color, linetype = "dotted") +
  geom_line(data = binom_df, aes(x = temp, y = num_failed_upr),
            color = binom_color, linetype = "dotted") +
  geom_line(data = binom_df_morton,
            mapping = aes(x = temp, y = num_failed_est,
                          color = binom_color_morton)) +
  geom_line(data = binom_df_morton, aes(x = temp, y = num_failed_lwr),
            color = binom_color_morton, linetype = "dotted") +
  geom_line(data = binom_df_morton, aes(x = temp, y = num_failed_upr),
            color = binom_color_morton, linetype = "dotted") +
  ggtitle("Figure 5. Expected Number of O-Ring Incidents by Temperature") +
  ylab("Number of O-Ring Failures per Launch") + xlab("Temperature (°F)") +
  geom_point(data = d, mapping = aes(x = Temp, y = O.ring), shape = 17) +
  geom_vline(xintercept = temp_challenger, linetype = "solid") +
  geom_hline(yintercept = num_O_rings, linetype = "dotted") +
  geom_hline(yintercept = 0, linetype = "dotted") + ylim(0, num_O_rings) +
  scale_color_identity(name = "Full Model",
                       labels = c("Subset Data", "Full Data"),
                       breaks = c(binom_color_morton, binom_color),
                       guide = "legend")
```
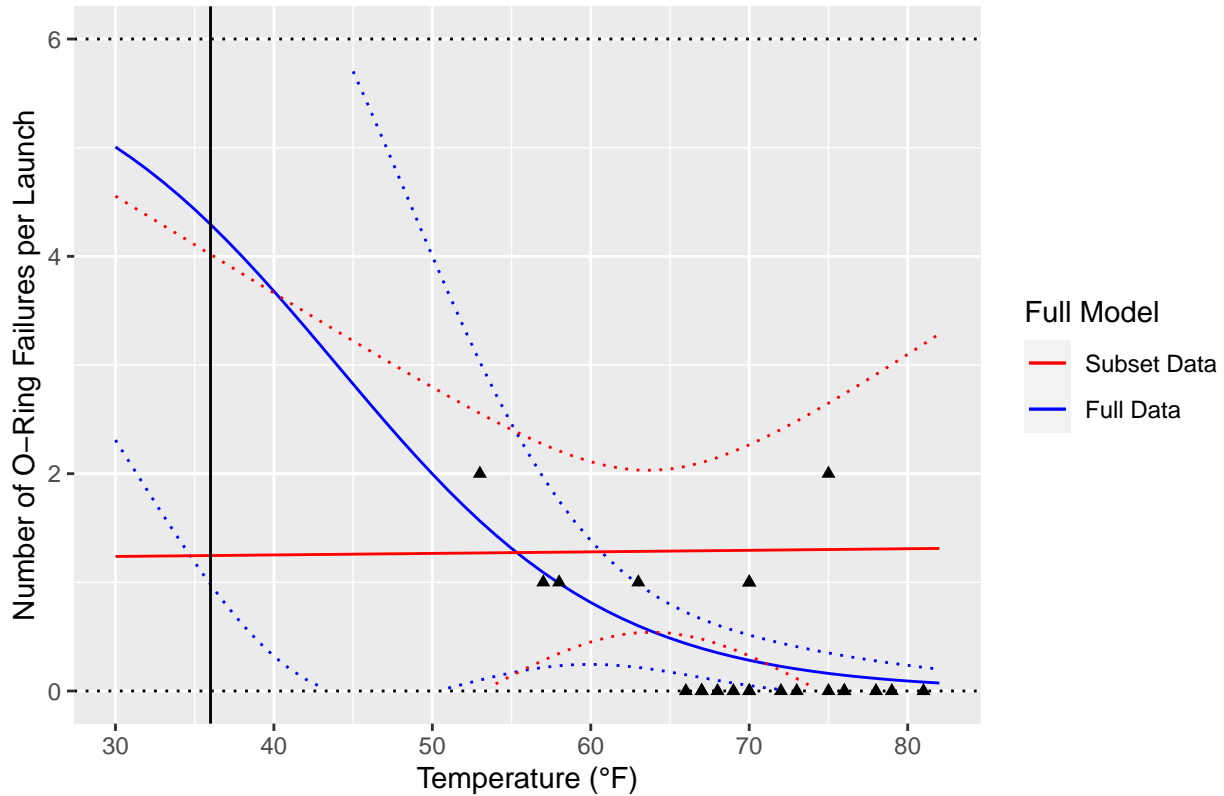
Figure 5 shows the number of predicted O-ring failures per launch over a range of temperatures, along with the Wald confidence intervals. The two curves are both generated from the full model, one which uses only the data subset Morton Thiokol considered and one on the full data. Like **Figure 4**, the vertical line indicates the launch temperature on the morning of the *Challenger* disaster of 36 °F.

Most notably, using only the subset data implies that temperature is not related to O-ring failure. Instead, the predicted number of failures is constant around one. At higher and lower temperatures the Wald confidence interval grows wider, but this is due to a lack of observations outside of the six which remain. This also highlights that in addition to being biased, this subset is also incredibly small. When already starting with rare events such as Space Shuttle launches, any further reduction in dataset size must be well motivated. As a result, with only six observations and two parameters in the model, the available degrees of freedom are nearly exhausted.

On the other hand, the full model with all data tracks previous Shuttle launches well. The only point which does not fit this curve is the observation with two O-ring failures at a temperature of about ~75°F. Even though this point is far from the predictions of the model, it does not appear to have skewed its fit.

Importantly, the model with all data predicts that at about four out of the six primary O-rings would fail at the launch temperature of the *Challenger*. This is an key result, as no previous launch had greater than two O-ring failures observed. As mentioned earlier, these six primary O-rings were intended to be redundant with six secondary O-rings. However, early in the Shuttle program it was known that a phenomena specific to the secondary O-rings could occur which rendered them useless at the start of each launch. If this happened at the same time as a primary O-ring failed, then it was already known that the result would be a catastrophe. Therefore, with at least four

primary O-rings expected to fail at launch, the *Challenger* disaster would only have been averted if all four of the secondary O-rings to these did not also start the launch non-functional.

Error Estimation

Still, this estimate must be qualified by the large confidence interval ranging from one to all six primary O-rings failing. In both **Figures 4** and **5**, the width of the Wald confidence intervals generally increases as temperature decreases in the full model. In the binary model of **Figure 4**, however, the confidence interval grows and then shrinks as temperatures decreases. In fact, at the launch temperature of 36°F the width is extremely small in the binary model. This represents high confidence that at least one O-ring will fail. However, this is likely an example of a "zero-width" interval being estimated due to failing to meet the Wald interval assumptions for a Binomial problem. The width of the confidence interval ranging to below zero and greater than one is a similar a demonstration of its failure in the current context.

```
binwidth <- 0.1

ggplot() +
  geom_histogram(data = binary_df, mapping = aes(x = pi, fill = binary_color),
    binwidth = binwidth, center = binwidth/2, alpha=0.5, show.legend = TRUE) +
  geom_histogram(data = binom_df, mapping = aes(x = pi, fill = binom_color),
    binwidth = binwidth, center = binwidth/2, alpha=0.5, show.legend = TRUE) +
  ggtitle("Figure 5. Distribution of Predicted Probabilities") +
  xlab(expression(hat(pi))) + ylab("Count") +
  scale_x_continuous(breaks = seq(0, 1, binwidth)) + ylim(0,30) +
  scale_fill_identity(name = "Model",
                      labels = c("Binary", "Full"),
                      breaks = c(binary_color, binom_color),
                      guide = "legend")
```

Figure 5. Distribution of Predicted Probabilities
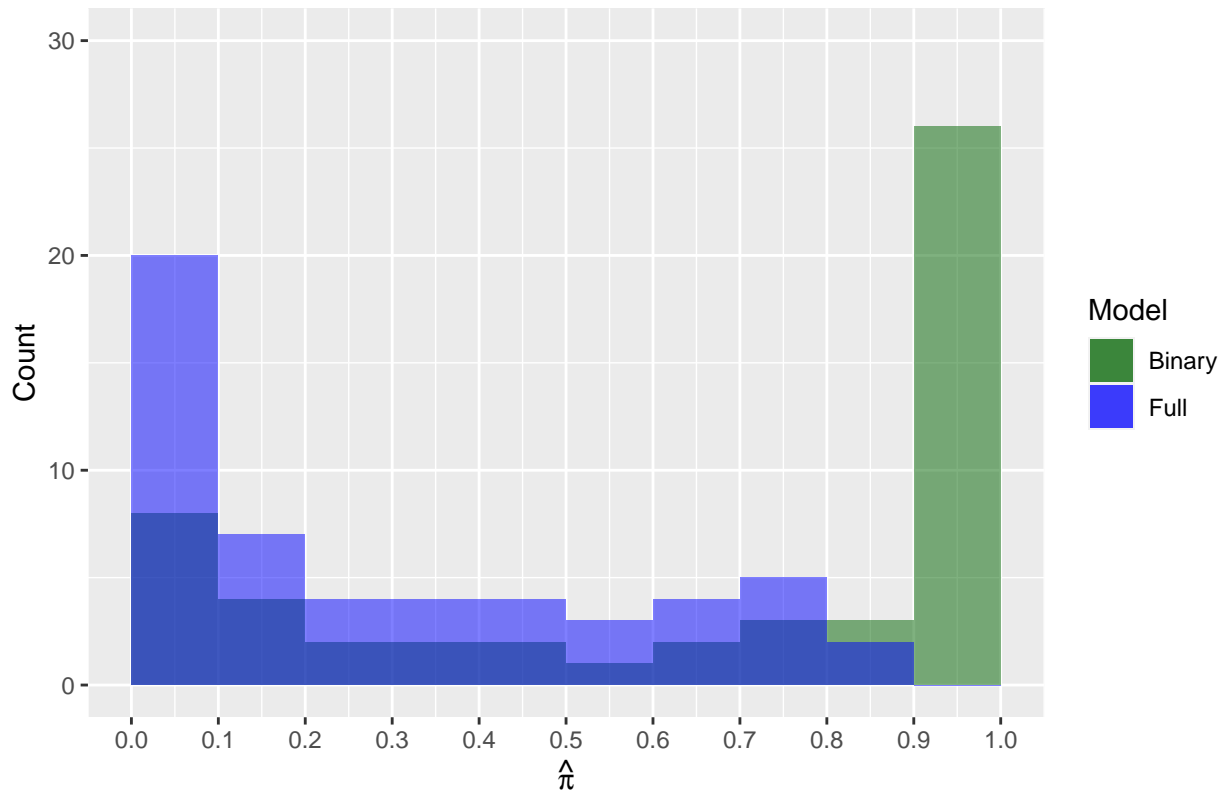
The validity of the Wald intervals can be examined via **Figure 5**. Specifically, the Wald interval assumes the outcome variable follows a normal distribution and has a large sample size. Earlier in the exercise, however, a Binomial distribution was assumed for a more appropriate and useful model. While Binomial distributions resemble normal distributions when n is large and p is near 0.5, these circumstances are not usually met and lead to "chaotic coverage" of the Wald interval.[3] Also as noted earlier, Space Shuttle launches are rare events and the overall size of the dataset is small at less than thirty observations.

The effect of these differences can be seen in **Figure 5** where $\hat{\pi}$ has fewer observations towards the center of its range, with a greater number on its edges. This is true for both the binary and full models. However, this is the opposite of a normal distribution where the median value is the most often observed. With all of these differences, the Wald intervals should not be considered precise estimates of the standard error.

As done with the previous hypothesis tests, the profile LR confidence interval can be constructed instead of the Wald. Unlike the Wald, the LR confidence interval assumes a χ^2 distribution in the outcome. This is much more easily approximated by the Binomial distribution the data was assumed to follow for modeling, as found when comparing a χ^2 distribution to **Figure 5**.

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

# create a K-matrix for challenger temperature
K <- matrix(data = c(1, temp_challenger) , nrow = 1, ncol = 2)
```

```r
# calculate -2log(lambda)
lc <- mcprofile(object = binom_temp, CM = K)

# calculate confidence interval of logit based on likelihood
ci_logit_profile <- confint(object = lc, level = 1 - alpha)

# calculate confidence interval of probability based on logit
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 = 1 - alpha)

wald_lower_logit <- wald_ci_exp$confint$lower
wald_upper_logit <- wald_ci_exp$confint$upper

wald_lower_pi <- exp(wald_lower_logit)/(1 + exp(wald_lower_logit))
wald_upper_pi <- exp(wald_upper_logit)/(1 + exp(wald_upper_logit))
```

At 36°F, the probability of an individual O-ring failing has a 95% LR confidence interval of 12.6% and 97.4%. The Wald confidence interval under the same conditions is 14.3% and 97.4%. While a small difference at this temperature, this demonstrates that the LR interval is more conservative than the Wald and should be preferred.

Dalal et al., the authors of the publication this exercise is based on, did not use Wald or LR methods to estimate error. Instead, they used a parametric bootstrap. Bootstrapping involves randomly varying a quantity of interest for many iterations, where a new outcome is calculated for each iteration. After many iterations, a distribution of possible results is generated. By the central limit theorem, this distribution is approximately normal. Once constructed, the distribution's variance can then be used as an estimate of uncertainty in the outcome as a function of the varied quantity.

Because one of the key features of this dataset is the low number of observations, at 23 total, it was assumed that this would be a primary contributor to error in a model's predictions. From this, a procedure was designed where each iteration would: - Sample the original dataset with replacement to create a new synthetic dataset of equal size (23 observations) - Fit a model with identical structure to the new dataset - Make a prediction using the new model at a temperature of interest

After a large number of iterations, the variance in the predictions would then be used to construct the confidence interval at the given temperature. In this context, an accurate confidence interval for the temperature of the *Challenger*'s final launch at 36°F is desired. Because no prior observation was from such a low temperature, this also represents an estimate of the out-of-sample uncertainty. For comparison, the procedure is also repeated at a temperature of 70°F to demonstrate an in-sample estimate.

```r
temp_in_sample <- 70
temps_of_interest <- data.frame('Temp' = c(temp_challenger, temp_in_sample))

vector_pi_challenger <- NA
vector_pi_in_sample <- NA
```

```r
for(i in 1:1000) {

  # generate new dataset from sampling with replacement
  d_i <- sample_n(d, size = nrow(d), replace=TRUE)

  # fit new model using new dataset
  model <- glm(formula = O.ring.fraction ~ Temp, data = d_i,
               family = binomial(link = logit), weights = Number)

  # make predictions using new model at temperatures of interest
  preds_i <- predict(object = model, newdata = temps_of_interest,
                     type='response', se=TRUE)

  vector_pi_challenger[i] <- preds_i$fit[1]
  vector_pi_in_sample[i] <- preds_i$fit[2]
}

# estimate the standard error
se_challenger <- sd(vector_pi_challenger)
se_in_sample <- sd(vector_pi_in_sample)

# estimate the mean
mean_challenger <- mean(vector_pi_challenger)
mean_in_sample <- mean(vector_pi_in_sample)

# estimate bootstrapped confidence intervals
ci_challenger <- mean_challenger + qnorm(p = c(alpha/2, 1-alpha/2))*se_challenger
ci_in_sample <- mean_in_sample + qnorm(p = c(alpha/2, 1-alpha/2))*se_in_sample
```

Similar to before, the probability of an individual O-ring failing at the launch temperature of the *Challenger* was estimated via parametric bootstrap to be 67.8%. The bootstrapped 95% confidence interval was 20.5% and 115.1%. At 70°F, the same method estimated the probability at 4.5% with a 95% confidence interval of -0.1% and 9.2%. Note that in both of these intervals, bounds extend beyond the limit of 0% and 100% similar to the Wald interval. Although bootstrapping has become popular in recent decades with the rise of computation power and simple application to many contexts, certain issues can still be encountered depending on the use case.

```r
df_bootstrap_challenger <- data.frame("pi" = vector_pi_challenger)
df_bootstrap_in_sample <- data.frame("pi" = vector_pi_in_sample)

plt_c <- ggplot(data = df_bootstrap_challenger) +
  aes(x = pi) +
  geom_histogram(binwidth = binwidth, fill = "blue", color = "blue",
                 center = binwidth/2, alpha = 0.3) +
  geom_vline(xintercept = ci_challenger, color = 'darkorange') +
  labs(title = "Figure 6a. 36°F") +
  ylab("") + xlab("") +
```

```
    scale_x_continuous(breaks = seq(0, 1, binwidth)) +
    theme(plot.title = element_text(size = 7),
          axis.text.x = element_text(size = 8, angle = 35))

plt_is <- ggplot(data = df_bootstrap_in_sample) +
  aes(x = pi) +
  geom_histogram(binwidth = binwidth/5, fill = "purple", color = "purple",
                 center = binwidth/(5*2), alpha = 0.3) +
  geom_vline(xintercept = ci_in_sample, color = 'darkorange') +
  labs(title = "Figure 6b. 70°F") +
  ylab("") + xlab("") +
  scale_x_continuous(breaks = seq(0, 1, binwidth/5)) +
  theme(plot.title = element_text(size = 7),
        axis.text.x = element_text(size = 8, angle = 35))

grid.arrange(plt_c, plt_is, nrow = 1,
             top = textGrob(
               label = "Figure 6. Bootstrapped Probabilities of Failure"),
             bottom = textGrob(label = expression(pi)),
             left = textGrob(label = "Count", rot = 90))
```


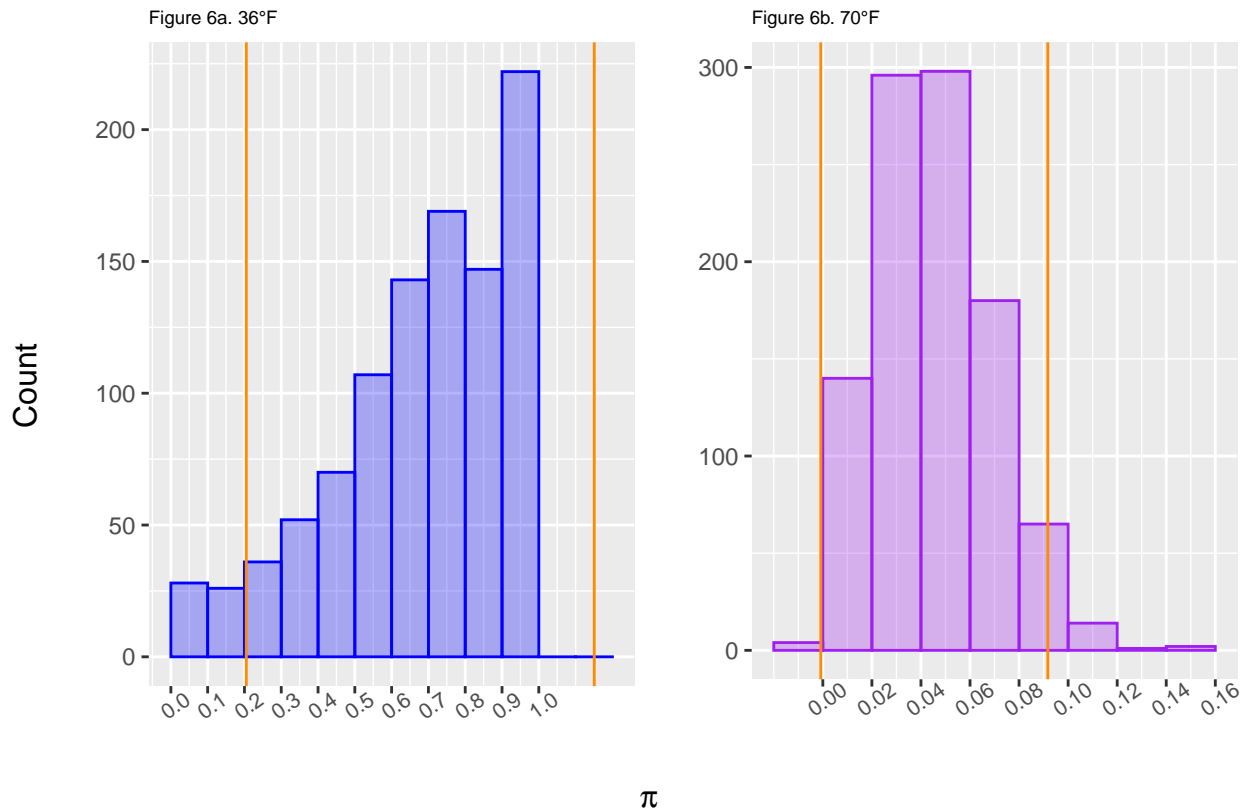Figure 6. Bootstrapped Probabilities of Failure

Figure 6 shows the bootstrapped probabilities of failure for a single O-ring at 36°F and 70°F. The confidence intervals determined from the distribution are also shown in orange for each temperature.

The "parametric" nature of the bootstrap can be observed as both resemble a normal distribution, although the boundaries at 0% and 100% add a slight skew. However, these are both much more similar to normal distribution than what is shown in **Figure 5**, suggesting the bootstrap is more reliable than the Wald intervals even though both have similar boundary problems.

Linear Regression

While a Binomial distribution was assumed for the outcome and used with a logit link function, a simpler approach may have been to use only standard linear regression with a Gaussian distribution and identity link function.

```
linear_model <- glm(O.ring.fraction ~ Temp, data = d,
                    family = gaussian(), weights = Number)


summary(linear_model)
```

```
##
## Call:
## glm(formula = O.ring.fraction ~ Temp, family = gaussian(), data = d,
##     weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.22894  -0.16102  -0.03486   0.04311   0.76223
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.616402   0.203252   3.033  0.00633 **
## Temp        -0.007923   0.002907  -2.725  0.01268 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.05557193)
##
##     Null deviance: 1.5797  on 22  degrees of freedom
## Residual deviance: 1.1670  on 21  degrees of freedom
## AIC: -38.503
##
## Number of Fisher Scoring iterations: 2
```

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 an O-ring failure increases as the temperature decreases. Specifically, with all else constant, decreasing the temperature by 10°F would increase the probability of an individual O-ring failure by ~7.9%.

Even though a similar result is found, this model should not be used instead of the Binomial model. Chiefly, 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. Although previous confidence intervals were found to have this problem, this was due to the calculation of the standard error and not the model itself. In this linear model, a probability of one is reached at about -48.41°F while a probability of zero is reached at about 77.8°F. Given that there were recorded launches both above and below these temperatures, this is a very undesirable property of the model.

In addition, the linear model assumes homoscedasticity. This means that the variance in the model's residuals should be 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 7. Predicted Probabilities vs. Standardized Residuals",
    x = "Predicted Values",
    y = "Standardized Residuals")
```

Figure 7. Predicted Probabilities vs. Standardized Residuals
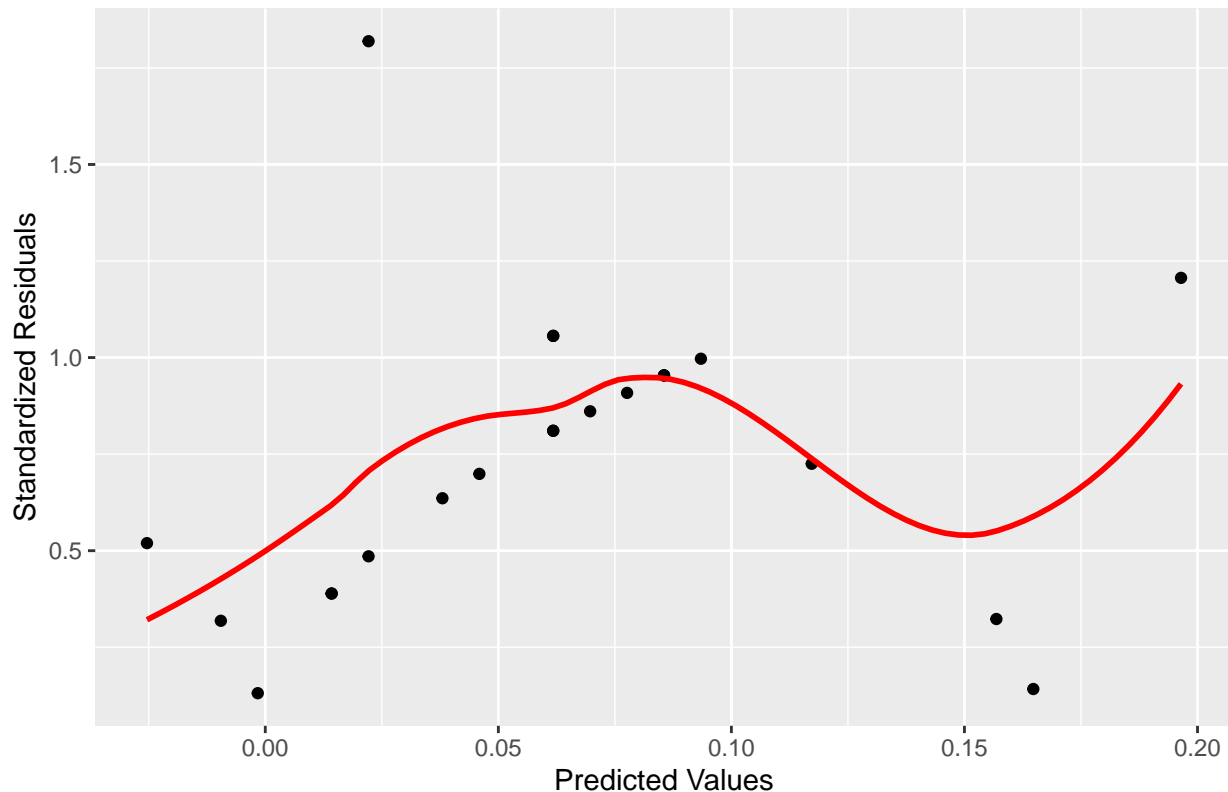
Shown in **Figure 7**, the standardized residuals are not constant over the range of the output variable in the model. This does not meet the homoscedasticity assumption, because the output variable is itself a function of the input variables. This result was expected, however, as it can be shown that the variance in a random variable from a Binomial distribution is a function of itself, or $Var(\pi) \propto \pi * (1 - \pi)$. It can be further seen in **Figure 7** that there is a general upward trend as

$\pi \to 0.5$, which follows from the variance relationship. Given that the homoscedasticity assumption is not met, there is further evidence of preferring the logistic model to the linear regression.

Summary

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

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

where $\hat{\pi}$ represents the probability of at least one O-ring failure per launch in the binary model and the probability of an individual O-ring failing in the full model.

```
beta_0_binary <- binary_temp$coefficients[1]
beta_1_binary <- binary_temp$coefficients[2]

beta_0_binom <- binom_temp$coefficients[1]
beta_1_binom <- binom_temp$coefficients[2]

beta_0_morton <- binom_temp_morton$coefficients[1]
beta_1_morton <- binom_temp_morton$coefficients[2]
```

In the binary model, the intercept $\hat{\beta}_0$ had the value 15.0429016 while the coefficient of temperature $\hat{\beta}_1$ had the value -0.2321627. The coefficient of temperature can be interpreted as increasing the odds of any O-ring failure by 1.261325 with every 1°F decrease in temperature and all else constant. Likewise, in the full model the intercept $\hat{\beta}_0$ had the value 5.0849772 while the coefficient of temperature $\hat{\beta}_1$ had the value -0.1156012. In this case, for every 1°F decrease in temperature with all else constant the odds of a single O-ring failing would be increased by 1.1225481.

This is a key finding. It underscores the importance 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 morning of the *Challenger*'s final launch the temperature was 36°F. At this point, the predicted the probability of at least one O-ring failure is 99.9%, with each individual O-ring having a predicted probability of failure of 71.6%. This corresponds to about four out of the six primary O-rings being expected to fail. As shown previously, the Wald, LR, and Bootstrapped 95% confidence intervals are large, ranging from only one to all six O-rings failing, due to a launch never before being attempted at such a low temperature.

The original estimate by Morton Thiokol omitted all launch data when no O-ring damage was observed under the assumption that it did not have anything to offer to an analysis of O-ring damage. This reduced the number of data points from 23 to only six. More importantly, such an action makes it fundamentally impossible to understand why some O-rings fail while others don't when only considering failed O-rings and not the entire group. If this was corrected, the relationship between temperature and O-ring damage could have been learned with the available data.

Dalal et al. further went on to estimate the probability of secondary O-ring failure with additional data. It was found that given failure in a primary O-ring, a secondary O-ring had a probability of failure equal to $p = 0.286 * 0.292 \approx 8.4\%$. This comes from the probability of erosion followed by the probability of blowby in a secondary O-ring, similar to what could occur in a primary O-ring. Taking this, the full model's estimate of a catastrophic failure for the *Challenger* launch would be $0.084 * 4 \approx 34\%$ given that only one secondary O-ring needed to fail out of the four behind the failed primary O-rings. Certain engineers within Morton Thiokol who had a similar belief and suggested this, but had their opinions dismissed. Using the model developed with the company's methods of dropping non O-ring damaged launches, the probability of catastrophe would only be estimated at 10.5%. Therefore, it would have been possible to predict that the launch temperature of the *Challenger* disaster was very likely to end in a catastrophic failure.

Had the launch been postponed only a few hours to reach that day's noon temperature of 66°F, the probability of an individual O-ring failure would have dropped to 7.3% and the expected number of O-ring failures for the launch to ~0.4. At a value less than one, it would be expected that no primary O-rings failed. In this scenario, the probability of a catastrophe would only have been $0.084 * 0.4 \approx 3.4\%$. Even this remains high, and had this flaw been focused on earlier then work could have proceeded to make the launch system safer and more reliable. It can be concluded that the catastrophe was avoidable had the program appropriately analyzed data by factoring in both the successful and failed launches, and that the O-ring flaw was inherent up to this point in Shuttle design.

Citations:

[1] Dalal, Siddhartha R., et al. "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure." Journal of the American Statistical Association, vol. 84, no. 408, [American Statistical Association, Taylor & Francis, Ltd.], 1989, pp. 945–57, https://doi.org/10.2307/2290069.

[2] Presidential Commission on the Space Shuttle *Challenger* Accident (1986), *Report of the Presidential Commission on the Space Shuttle Challenger Accident* (Vols. 1 & 2), Washington, D.C.: Author.

[3] Lawrence D. Brown, T. Tony Cai, Anirban DasGupta "Interval Estimation for a Binomial Proportion," Statistical Science, Statist. Sci. 16(2), 101-133, (May 2001)