Statistical Analysis of Alberta Wildfires from (2006 - 2024)

Group4: Johanna Enright - 30303537, Uyoyoumena Orona - 30295200, Agambir Bandesha - 30302510, F

Date: 2025-09-30

For example, here is our dataset in-text citation (Forestry & Parks, 2025) and the full citation will appear at the bottom.

Introduction

- written text about wildfires and why they are bad here
- a little bitty about how climate changing is increasing lightning strikes and making forests drier and more flammable here.
- a bit about wind and wildfire prediction here.

Hypotheses

- add our hypotheses here.

Data Visualization

- here is where we can put our code to generate plots for distributions and visualizations and such.

Statistical Analyses

Question 1: Is the proportion of lightning-caused wildfires from 2024 significantly greater than those from 2006?

Here we will acquire the proportion data for lightning caused wild fires in 2006 and 2024.

```
####### determining proportion for lightning caused fires in 2006 and 2024  #######

#2006
df_2006 <- df %>%  filter(YEAR == 2006)
n_2006 <- nrow(df_2006)
x_2006 <- df_2006 %>% filter(GENERAL_CAUSE == "Lightning") %>% nrow()  #number from 2006 caused by ligh
p_lightning_2006 <- x_2006/n_2006

#2024
df_2024 <- df %>%  filter(YEAR == 2024)
```

```r
n_2024 <- nrow(df_2024)
x_2024 <- df_2024 %>% filter(GENERAL_CAUSE == "Lightning") %>% nrow() #x is number of fires caused by l
p_lightning_2024 <- x_2024/n_2024

print(paste("The 2006 proportion is", round(p_lightning_2006,4) ," and the 2024 proportion is", round(p
```

```
## [1] "The 2006 proportion is 0.3823  and the 2024 proportion is 0.4528"
```

Next we will confirm that our data follows the assumptions necessary to use a parametric test. (i.e. that n is sufficiently large under the central limit theorem).

```r
## confirming it follows assumtions (np > 10) , (nq > 10)   # is sufficiently large.
n_2006*p_lightning_2006 >10 ; n_2006*(1 - p_lightning_2006) > 10 ## greater than 10
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```r
n_2024*p_lightning_2024 > 10 ; n_2024*(1 - p_lightning_2024) > 10
```

```
## [1] TRUE
```

```
## [1] TRUE
```

Confirmed here, all values of n are sufficiently large to use the central limit theorem tests.

Testing hypothesis $H_0 : p_{2024} \leq p_{2006}$ and $H_A : p_{2024} > p_{2006}$

```r
############################  CALCULATING Z-observed ######################################
# Zobs = ((phat1 - phat2) - (p1 - p2)) / sqrt(phat*(1-phat)*(1/n1 + 1/n2))

phat <- (x_2006 + x_2024)/(n_2006 + n_2024)
p1_min_p2 <- 0   #under the assumption of the null hypothesis

Zobs <- ((p_lightning_2024 - p_lightning_2006) - p1_min_p2)/sqrt(phat*(1-phat)*(1/n_2024 + 1/n_2006))
pvalue <- round(1 - pnorm(Zobs),5)
print(paste("The test statistic is:", round(Zobs,4), "The p-value is:", pvalue))
```

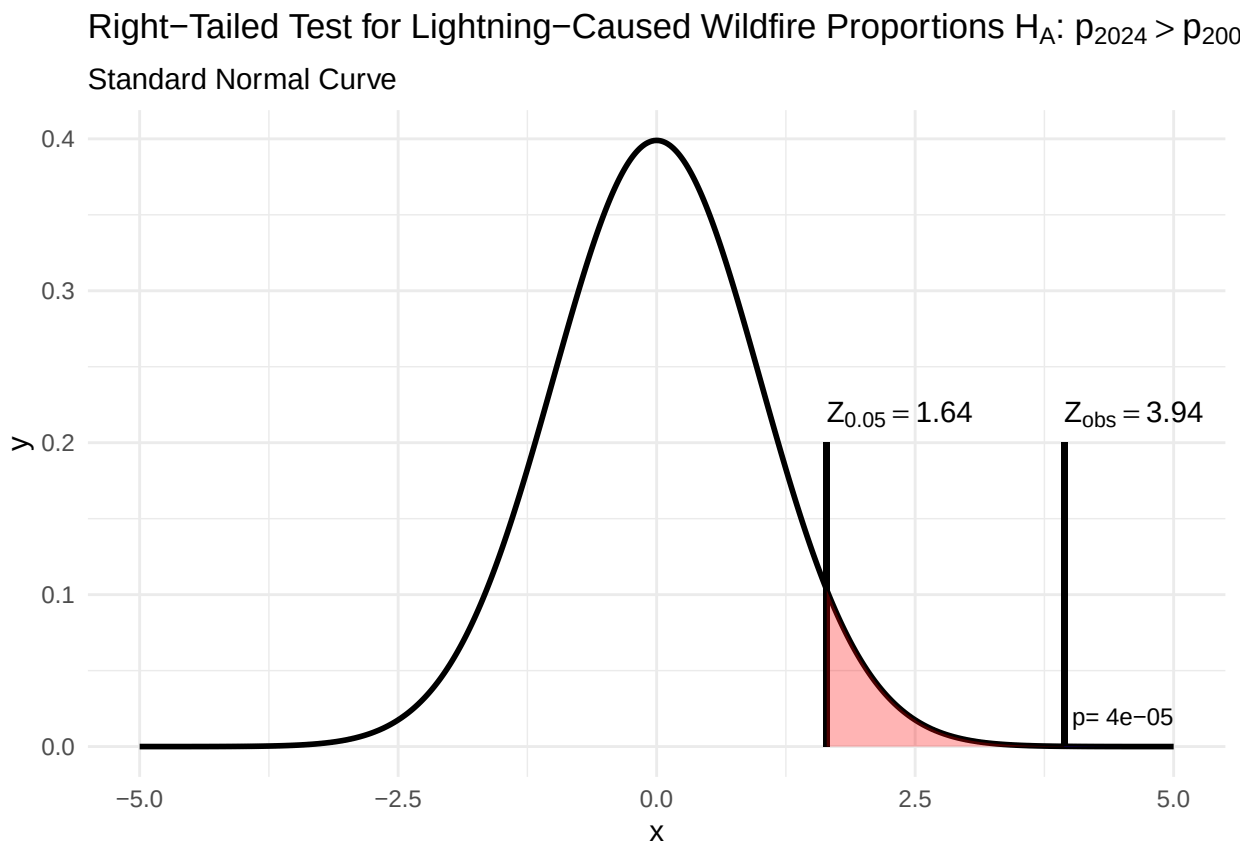
```
## [1] "The test statistic is: 3.9418 The p-value is: 4e-05"
```

Therefore, because our p-value is less than our alpha of 0.05, we can conclude that the proportion of lightning caused wildfires in 2024 is significantly greater than the proportion from 2006.

```r
#set up dataframe
x <- seq(-5, 5,0.001)
df_norm <- tibble(x = x, y = dnorm(x, mean = 0, sd = 1))

#labels for lines
labels_df <- tibble(x = c(qnorm(0.95), Zobs), y = 0.21,
                    label = c(bquote(Z[0.05] == .(round(qnorm(0.95),2))), bquote(Z["obs"] == .(round(Zobs,
```

```r
p <- df_norm %>%
      ggplot(aes(x = x, y = y)) +
            geom_line(linewidth = 1)+
            geom_segment(aes(x = qnorm(0.95), xend = qnorm(0.95), y = 0, yend = 0.2),
               linewidth = 1, color = "black") +
            geom_segment(aes(x = Zobs, xend = Zobs, y = 0, yend = 0.2),
               linewidth = 1, color = "black") +
            geom_text(data = labels_df, aes(x = x, y = y, label = label), vjust = 0, hjust = 0, color =
    #shade in areas
   geom_area(data = subset(df_norm, x > qnorm(0.95)), aes(x = x, y = y), fill = "red", alpha = 0.3) +
   geom_area(data = subset(df_norm, x > Zobs), aes(x = x, y = y), fill = "blue", alpha = 1) +     #shad
   annotate(geom = "text", x = 5, y = 0.02, label = paste("p=", pvalue), size = 3, hjust = 1) +
  ggtitle(bquote("Right-Tailed Test for Lightning-Caused Wildfire Proportions H"[A] * ": p"[2024] > "p"
          subtitle = "Standard Normal Curve") +
  theme_minimal()
print(p)
```



Next, we will confirm our findings with a 95% Agresti-Coull Confidence interval for the difference in proportions, $p_{2024} - p_{2006}$.

```r
################### USING THE AGRESTI-COULL (+1/+2 METHOD) 95% CI ################################
# (ptilde1 - ptilde2) +/-  (z (1 - alpha/2)) * sqrt((ptilde1*qtilde1/(n1 + 2)) + (ptilde2*qtilde2)/(n2
#p_tilde = (x + 1)/ (n + 2)

p_tilde_2024 <- (x_2024 + 1)/(n_2024 + 2)
```

```r
p_tilde_2006 <- (x_2006 + 1)/(n_2006 + 2)

ptilde_diff <- p_tilde_2024 - p_tilde_2006
z <- qnorm(1 - 0.05/2)
SE <- sqrt(((p_tilde_2024)*(1-p_tilde_2024)/(n_2024 + 2)) +
            ((p_tilde_2006)*(1-p_tilde_2006)/(n_2006 + 2))
                )

lb <- ptilde_diff - z * SE
ub <- ptilde_diff + z * SE
paste(round(lb, 4), "-", round(ub, 4))
```

```
## [1] "0.0353 - 0.1057"
```

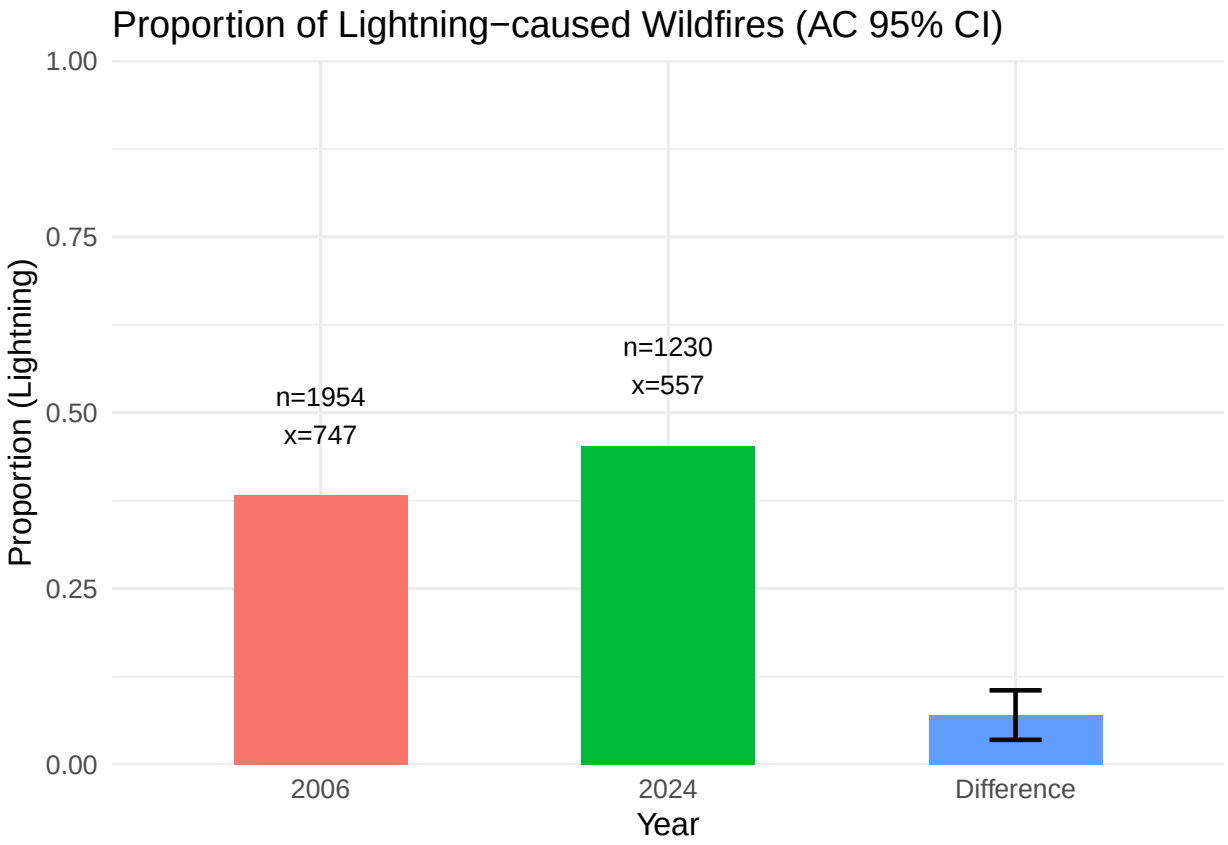
```r
# CONCLUSION : 2024 HAS A SIGNIFICANTLY HIGHER PROPORITON OF LIGHTNING-CAUSED WILDFIRES THAT 2006
```

Our 95% confidence interval for $p_{2024} - p_{2006}$ lies between 0.0353 - 0.1057. Both lower and upper bounds of our confidence interval are greater than 0, so this supports the above conclusion that the proportion of lightning caused wildfires from 2024 is significantly greater than that of 2006.

Here is a visualization for our 95% confidence interval.

```r
# Visualization
plot_df <- tibble(
  YEAR  = factor(c(2006, 2024, "Difference")),
  prop  = c(p_lightning_2006, p_lightning_2024, p_lightning_2024-p_lightning_2006),
  lower = c(NA, NA, lb),
  upper = c(NA, NA, ub),
  n     = c(n_2006, n_2024, NA),
  x     = c(x_2006, x_2024, NA)
)

ggplot(plot_df, aes(x = YEAR, y = prop, fill = YEAR)) +
  geom_col(width = 0.5, show.legend = FALSE) +
  geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.15, linewidth = 0.8) +
  geom_text(aes(label = ifelse(YEAR %in% c(2006, 2024), paste0("n=", n, "\nx=", x), NA)),
            vjust = -0.9, size = 3.5) +
  scale_y_continuous(limits = c(0, 1), expand = c(0, 0)) +
  labs(
    title = "Proportion of Lightning-caused Wildfires (AC 95% CI)",
    x = "Year",
    y = "Proportion (Lightning)"
  ) +
  theme_minimal(base_size = 12)
```



Question 2

- Agam and Mena, you guys can do your analyses here.

References

Forestry, & Parks. (2025). *Historical wildfire data: 2006 to 2024*. Open Government Dataset; Government of Alberta. https://open.alberta.ca/opendata/wildfire-data#summary