Shark Attack Analysis

AJ, Chase, Chloe

Shark Analysis

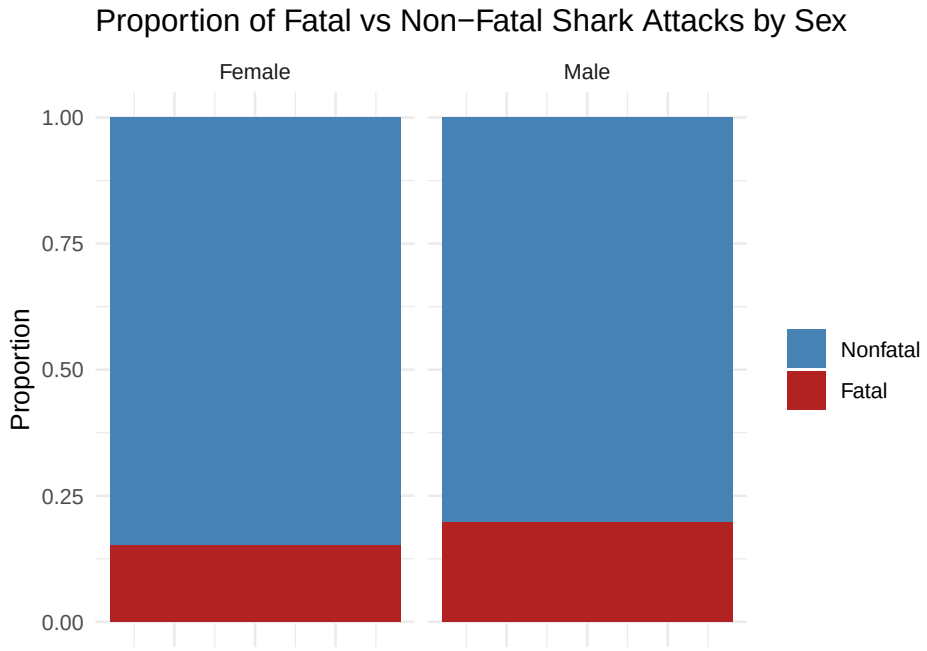
Introduction

Is there a difference between the fatality rate from shark attacks for men and women? We chose this research question because we are curious about the difference in proportions of fatal shark attacks among men and among women. Sharks have long been feared — especially since the movie *Jaws* came out in 1975. Are sharks really as dangerous as people think? After all, cows cause more annual deaths than sharks on average, and more commonplace things like cars are far more likely to kill you than a shark. We hope in this analysis to assuage fears (at least somewhat) concerning sharks.

Methods

Our data comes from Global Shark Attack File ("https://www.sharkattackfile.net/"). It is a compilation of shark attack information coming from many different sources. Each of our observations represent one occurrence of a shark attack. For each observation we have information about the sex of the victim and the fatality of the attack – whether the victim survived or was killed. Below is a graph that displays the differences in mortality rates for male and female victims and a table that displays the proportions of each category.

Sex	Fatal	Nonfatal	Fatality Rate (%)
Male	918	3718	19.8
Female	108	601	15.2

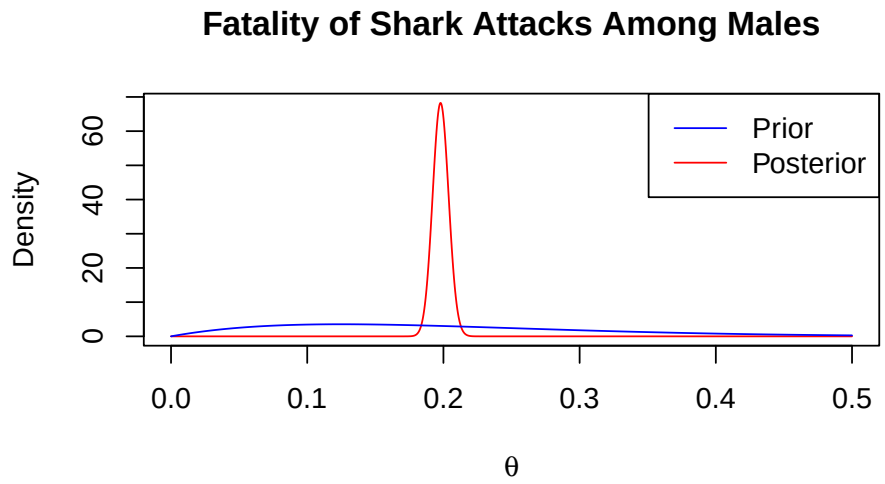


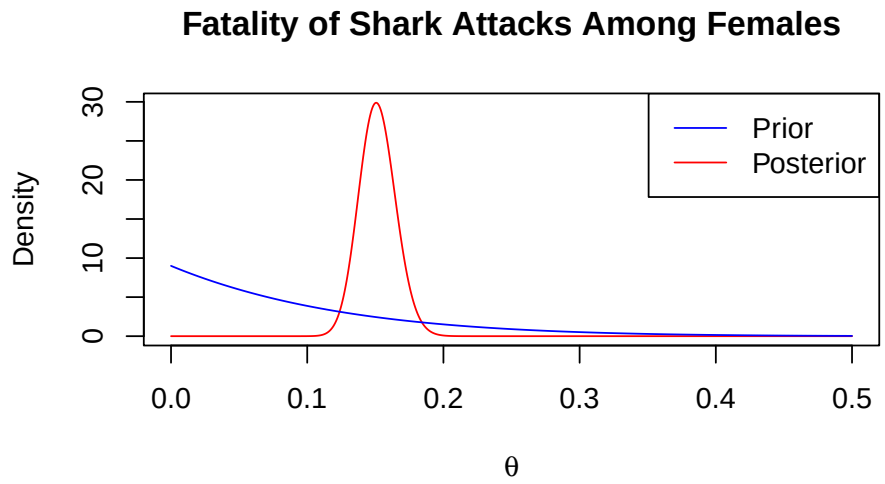
Our data lends itself to the binomial setting. We have a fixed number of events where each has a binary outcome, fatal or nonfatal. It is fairly reasonable to assume that our events are independent. It is unlikely that one shark attack has any sort of significant impact on another. The assumption of greatest concern in using a binomial likelihood is the constant fatality probability of each event. It is unlikely that the fatality of each shark attack is exactly the same. Ignoring this violation would normally result in underestimating the variability of our prediction for the probability of a shark attack being fatal. However, using beta-binomial Bayesian inference allows us to accurately convey our levels of uncertainty.

Our parameters of interest are θ_{female} and θ_{male}. These parameters represent the probability that a shark attack is fatal for each respective sex. With these parameters, we can use Monte Carlo simulation to generate a posterior distribution for $\theta_{female-male}$. This distribution allows us to answer our question of whether one sex is more likely to survive a shark attack than the other. Based on the likelihood of the data and our prior beliefs about shark attacks, we chose $\pi_{female} \sim beta(1, 9)$ for our prior for females. We chose a similar $\pi_{male} \sim beta(2, 8)$ prior for male fatality rates. Beta priors are conjugate with the binomial setting and they will make it easy for us to encode our prior beliefs. In our priors we assume that about 1/10 shark attacks will be fatal to women and about 2/10 shark attacks will be fatal for men, encoding a slight survival favoritism towards women. This favoritism is very weak overall, allowing room for our data to inform our posterior distribution.

Results

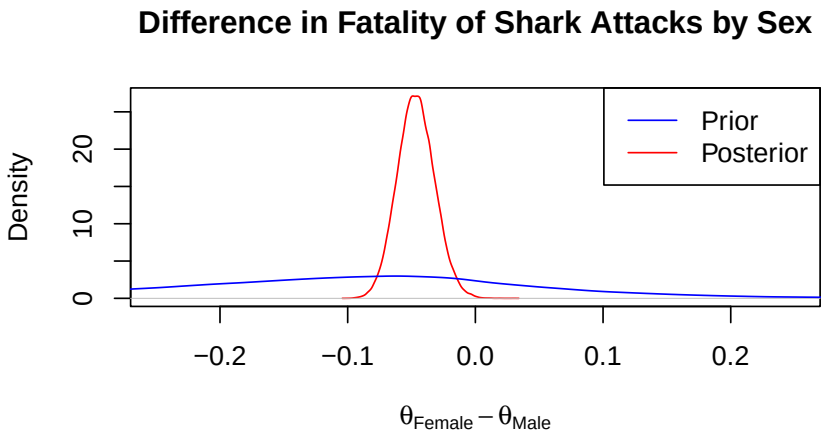
First of all, in order to preform our analysis we obtain the posterior distributions for θ_{male} and θ_{female}. Because we chose conjugate priors, we can obtain these through a simple process of adding the results of our data onto our priors. The graphs below display the prior and posterior distributions for θ_{male} and θ_{female}.





Through Monte Carlo simulation we obtain the posterior distribution for the difference of fatality rates between males and females from shark attacks. We plot this posterior distribution

below with its associated prior. We also obtained a 95% credible interval of between -0.074 and -0.017. This means that there is a 95% probability that the difference in proportions of fatal shark attacks for males and females is between -0.074 and -0.017.



Discussion

Our posterior results indicate a clear difference in fatality rates between men and women. The posterior distribution of $\theta_{female} - \theta_{male}$ is centered below zero, and the 95 percent credible interval stays entirely negative. This implies that, based on the data, women have a lower probability of dying from a shark attack than men. Despite using weak priors, the data overwhelmingly push the posteriors toward the observed male–female gap.

That said, several limitations temper the strength of this conclusion. The Global Shark Attack File aggregates incidents across many decades and regions, and reporting quality varies. The binomial model also assumes a constant fatality probability within each gender even though factors like shark species, activity, and location clearly affect outcomes. Ignoring these sources of heterogeneity likely understates uncertainty. Finally, the data itself may suffer from reporting and selection biases.

Future work could incorporate additional predictors or use a hierarchical model to better account for variation across attack types and conditions. Examining how fatality rates change over time or differ across regions would also help clarify whether the observed gender gap reflects biological, behavioral, or environmental differences. Overall, our analysis identifies a modest but credible difference in fatality probabilities, while leaving room for more detailed modeling to explain why that difference exists. So even though sharks are often treated like the super-villains of the ocean, the data suggests a far less deadly—though slightly more for males—reality.

Appendix

```
#set seed for reproducability
set.seed(2024)

#Read in data and clean Sex and Fatal Variables
#filter out data before 1900 as it is unreliable
sharks <- read_excel("sharks.xlsx") %>%
  select(1:14) %>%
  rename(Fatal = 'Fatal Y/N') %>%
  mutate(Year = as.numeric(Year),
         Sex = case_when(
           str_detect(tolower(Sex), "m") ~ "Male",
           str_detect(tolower(Sex), "f") ~ "Female",
           TRUE ~ NA_character_),
         Fatal = case_when(
           str_detect(tolower(Fatal), "n") ~ "No",
           str_detect(tolower(Fatal), "y") ~ "Yes",
           TRUE ~ NA_character_)) %>%
  filter(Year > 1900) %>%
  select(Year, Sex, Fatal)

#generate posterior distribution of data for males
a_male <- 2
b_male <- 8
a_star_male <- nrow(sharks %>% filter(Sex == 'Male')
                    %>% filter(Fatal == "Yes")) + a_male
b_star_male <- nrow(sharks %>% filter(Sex == 'Male')
                    %>% filter(Fatal == "No")) + b_male

#graph the prior and posterior distribution for males
theta <- seq(0,.5,by = .0001)
plot(theta, dbeta(theta, a_star_male, b_star_male), type='l', col="red",
     xlab=expression(theta), ylab="Density",
     main="Fatality of Shark Attacks Among Males")
lines(theta, dbeta(theta, a_male, b_male), col="blue")
legend("topright", legend = c("Prior", "Posterior"),
       col = c("blue", "red"), lty = 1)

#generate posterior distribution for females
a_female <- 1
b_female <- 9
```

```r
a_star_female <- nrow(sharks %>% filter(Sex == 'Female')
                      %>% filter(Fatal == "Yes")) + a_female
b_star_female <- nrow(sharks %>% filter(Sex == 'Female')
                      %>% filter(Fatal == "No")) + b_female

#graph prior and posterior distributions for females
theta <- seq(0,.5,by = .0001)
plot(theta, dbeta(theta, a_star_female, b_star_female), type='l', col="red",
     xlab=expression(theta), ylab="Density",
     main="Fatality of Shark Attacks Among Females")
lines(theta, dbeta(theta, a_female, b_female), col="blue")
legend("topright", legend = c("Prior", "Posterior"),
       col = c("blue", "red"), lty = 1)

#use Monte Carlo simulation to generate distribution for females minus males
diff_prior <- rbeta(100000, a_female, b_female) -
  rbeta(100000, a_male, b_male)
diff_post <- rbeta(100000, a_star_female, b_star_female) -
  rbeta(100000, a_star_male, b_star_male)

#graph prior and posterior distribution for females minus males
plot(density(diff_post), xlab=expression(theta[Female]-theta[Male]),
     main="Difference in Fatality of Shark Attacks by Sex",
     col="red", xlim=c(-.25, .25))
lines(density(diff_prior), col="blue")
legend("topright", legend = c("Prior", "Posterior"),
       col = c("blue", "red"), lty = 1)

#95% credible interval for difference in fatalities between females and males
quantile(diff_post, c(.025, .975))
```