Inferential Analysis

2024-05-01

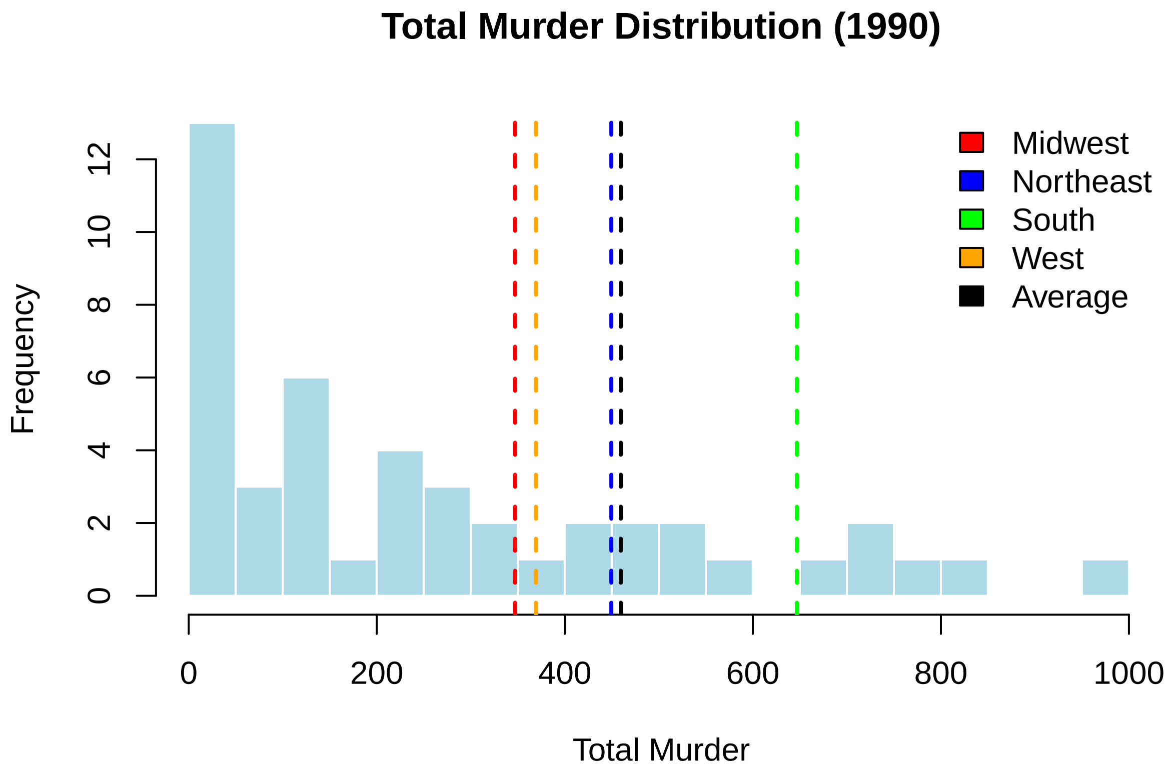
Is more crime of a specific type depending on the region i.e., is murder more apparent in the East or West Coast, and is there a correlation to that type of crime that occurs based on location? To explore this subquestion, we have to assume that no other factors are affecting the data; factors such as current events and the unemployment rate, for example, will be disregarded. I created subsets of my main dataset and organized each subset into regions. Each state was then categorized by region based on the state and after doing so, I created a histogram that contained "total crime" on the x-axis and the frequency on the y-axis. A problem I continued to run into was the outliers that completely skewed the dataset, causing the averages in the datasets to be an inaccurate description of the data as a whole so I created a filter with an outlier threshold that removes threshold using this function

In doing so, we can see the data more accurately and explore the different types of crime that occurred in different regions in 1990, 2009, and 2019.

When looking at the total murder distribution in 1990, 2009, and 2019, showing the distribution, we can see every single region lies on the graph. The average contains the average of every single region and is averaged out. When going through the findings, it seems that the South has a noticeable trend with specific murder as its crime, consistently being above the average in every year we decided on.

Histogram for 1990

```r
merged_data <- left_join(c, a, by = c("State"))

west <- subset(merged_data, merged_data$Region == "West")
northEast <- subset(merged_data, merged_data$Region == "Northeast")
midWest <- subset(merged_data, merged_data$Region == "Midwest")
south  <- subset(merged_data, merged_data$Region == "South")

merged_data1990 <- merged_data|>
  filter(Year == 1990)

outlier_threshold <- 3
mean_data <- mean(merged_data1990$Data.Totals.Violent.Murder)
sd_data <- sd(merged_data1990$Data.Totals.Violent.Murder)
filtered_data <- merged_data1990|>
  filter(Data.Totals.Violent.Murder >= mean_data - outlier_threshold * sd_data &
           Data.Totals.Violent.Murder <= mean_data + outlier_threshold * sd_data)
filtered_mean2019 <- mean(filtered_data$Data.Totals.Violent.Murder)

region_averages <- aggregate(Data.Totals.Violent.Murder ~ Region, data = merged_data1990, FUN = mean)

region_colors <- c("red", "blue", "green", "orange", "black")

hist(merged_data1990$Data.Totals.Violent.Murder,
     main = "Total Murder Distribution (1990)",
     xlab = "Total Murder",
     ylab = "Frequency",
     col = "lightblue",
     border = "white",
     xlim = c(min(merged_data1990$Data.Totals.Violent.Murder), 1000),
     breaks = 800
)
abline(v = filtered_mean2019, col = "black", lwd = 2, lty = 2)

for (i in seq_along(region_averages$Region)) {
  abline(v = region_averages$Data.Totals.Violent.Murder[i], col = region_colors[i], lwd = 2, lty = 2)
}
legend("topright", legend = c(region_averages$Region, "Average"), fill = c(region_colors, "black"), bty
```



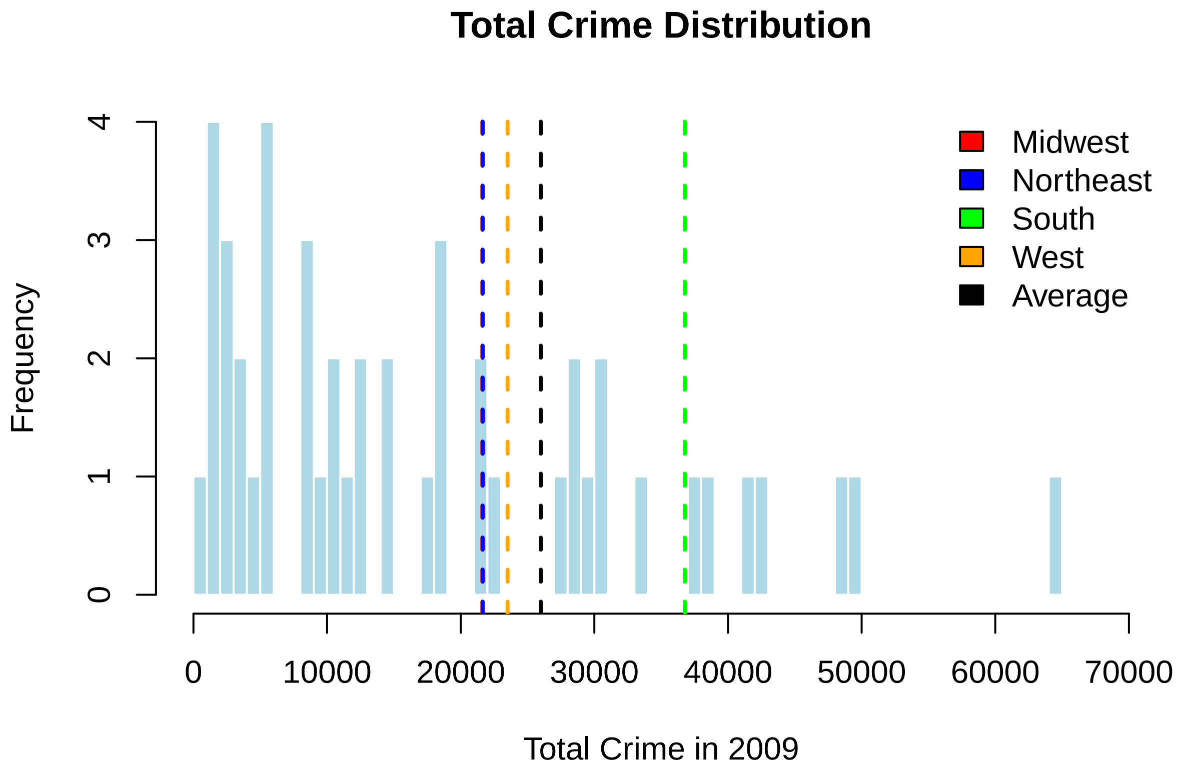
Histogram for 2009

```
merged_data2009 <- merged_data|>
  filter(Year == 2009)

midwest_mean <- mean(region_averages$Data.Totals.Violent.All)
```

```
## Warning in mean.default(region_averages$Data.Totals.Violent.All): argument is
## not numeric or logical: returning NA
```

```
outlier_threshold <- 3
mean_data <- mean(merged_data2009$Data.Totals.Violent.All)
sd_data <- sd(merged_data2009$Data.Totals.Violent.All)
```

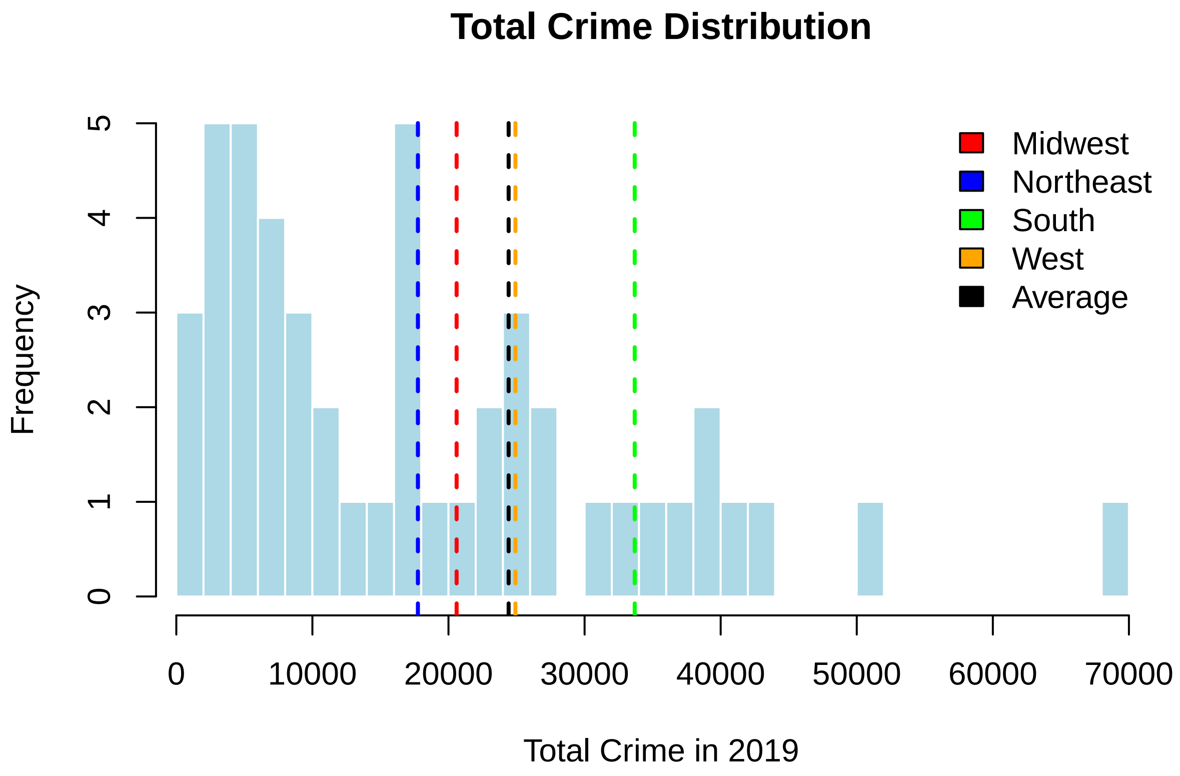
```
filtered_data <- merged_data2009 %>%
  filter(Data.Totals.Violent.All >= mean_data - outlier_threshold * sd_data &
           Data.Totals.Violent.All <= mean_data + outlier_threshold * sd_data)
filtered_mean2009 <- mean(filtered_data$Data.Totals.Violent.All)


region_averages <- aggregate(Data.Totals.Violent.All ~ Region, data = merged_data2009, FUN = mean)

region_colors <- c("red", "blue", "green", "orange", "black")

hist(merged_data2009$Data.Totals.Violent.All,
     main = "Total Crime Distribution",
     xlab = "Total Crime in 2009",
     ylab = "Frequency",
     col = "lightblue",
     border = "white",
     xlim = c(0, 70000),
     breaks = 1000
)
abline(v = filtered_mean2009, col = "black", lwd = 2, lty = 2)


for (i in seq_along(region_averages$Region)) {
  abline(v = region_averages$Data.Totals.Violent.All[i], col = region_colors[i], lwd = 2, lty = 2)
}
legend("topright", legend = c(region_averages$Region, "Average"), fill = c(region_colors, "black"), bty
```



Histogram for 2019

```
merged_data2019 <- merged_data|>
  filter(Year == 2019)

outlier_threshold <- 3
mean_data <- mean(merged_data2019$Data.Totals.Violent.All)
sd_data <- sd(merged_data2019$Data.Totals.Violent.All)
filtered_data <- merged_data2019 %>%
  filter(Data.Totals.Violent.All >= mean_data - outlier_threshold * sd_data &
           Data.Totals.Violent.All <= mean_data + outlier_threshold * sd_data)
filtered_mean2019 <- mean(filtered_data$Data.Totals.Violent.All)

region_averages <- aggregate(Data.Totals.Violent.All ~ Region, data = merged_data2019, FUN = mean)

region_colors <- c("red", "blue", "green", "orange", "black")

hist(merged_data2019$Data.Totals.Violent.All,
     main = "Total Crime Distribution",
     xlab = "Total Crime in 2019",
     ylab = "Frequency",
     col = "lightblue",
     border = "white",
     xlim = c(min(merged_data2019$Data.Totals.Violent.All), 70000),
     breaks = 800
)
abline(v = filtered_mean2019, col = "black", lwd = 2, lty = 2)
```

```
for (i in seq_along(region_averages$Region)) {
  abline(v = region_averages$Data.Totals.Violent.All[i], col = region_colors[i], lwd = 2, lty = 2)
}
legend("topright", legend = c(region_averages$Region, "Average"), fill = c(region_colors, "black"), bty
```


Total Crime Distribution
Frequency
Total Crime in 2019
🟥 Midwest
🟦 Northeast
🟩 South
🟧 West
■ Average