Midterm

Shivam Bajaj

2023-02-26

R Markdown

```
data <- read_csv("https://jlucasmckay.bmi.emory.edu/global/mckay_2021/S1.csv")
```

```
## Rows: 68 Columns: 191
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): Patient, Sex, Study Group, cohortClass, mds_updrs_iii_pheno, data...
## dbl (168): Age, PD Duration, TAka, TAkv, TAkd, TAlambda, TAkaPrime, TAkvPrim...
## lgl  (17): zofran_daily, duodopa_daily, ryatary_ld_daily, seleg_sublin_daily...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data <- select(data, Age, Sex, num_falls_6_mo)
data <- drop_na(data)
data <- filter(data, !grepl("bat112", row.names(data)))
print(data)
```

```
## # A tibble: 61 x 3
##      Age Sex    num_falls_6_mo
##    <dbl> <chr>           <dbl>
##  1    84 Female              2
##  2    53 Male                0
##  3    75 Female              4
##  4    59 Male                0
##  5    71 Female              0
##  6    69 Female              2
##  7    71 Male                0
##  8    66 Male                1
##  9    54 Female              0
## 10    73 Female              0
## # ... with 51 more rows
```

```
obs <- group_by(data, Sex)
obs <- summarize(obs, n = n())

print(obs)
```

```
## # A tibble: 2 x 2
##   Sex        n
##   <chr>  <int>
## 1 Female    31
## 2 Male      30
```

```
age_summary <- group_by(data, Sex)
age_summary <- summarize(age_summary, avg_age = mean(Age), sd_age = sd(Age))
print(age_summary)
```

```
## # A tibble: 2 x 3
##   Sex    avg_age sd_age
##   <chr>    <dbl>  <dbl>
## 1 Female    66.5   7.40
## 2 Male      68.1   7.53
```

```
falls_summary <- group_by(data, Sex)
falls_summary <- summarize(falls_summary, avg_falls = mean(num_falls_6_mo), sd_falls = sd(num_falls_6_mo
print(falls_summary)
```

```
## # A tibble: 2 x 3
##   Sex    avg_falls sd_falls
##   <chr>      <dbl>    <dbl>
## 1 Female      8.55    33.1
## 2 Male        0.7      2.00
```

```
overall_obs <- summarize(data, n = n())
print(overall_obs)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    61
```

```
overall_age_summary <- summarize(data, avg_age = mean(Age), sd_age = sd(Age))
print(overall_age_summary)
```

```
## # A tibble: 1 x 2
##   avg_age sd_age
##     <dbl>  <dbl>
## 1    67.3   7.44
```

```
overall_falls_summary <- summarize(data, avg_falls = mean(num_falls_6_mo), sd_falls = sd(num_falls_6_mo)
print(overall_falls_summary)
```

```
## # A tibble: 1 x 2
##   avg_falls sd_falls
##       <dbl>    <dbl>
## 1      4.69     23.8
```

```r
female_data <- filter(data, Sex == "Female")
male_data <- filter(data, Sex == "Male")

mean_age_female <- mean(female_data$Age)
mean_age_male <- mean(male_data$Age)

diff_means <- mean_age_female - mean_age_male
cat("Difference in sample means: ", diff_means)
```

```
## Difference in sample means:  -1.573441
```

```r
var_female <- var(female_data$Age)
var_male <- var(male_data$Age)
n_female <- length(female_data$Age)
n_male <- length(male_data$Age)
pooled_var <- ((n_female - 1) * var_female + (n_male - 1) * var_male) / (n_female + n_male - 2)
cat("\nPooled variance: ", pooled_var, "\n")
```

```
##
## Pooled variance:  55.71362
```

```r
t_stat <- diff_means / sqrt(pooled_var * (1/n_female + 1/n_male))
cat("Test statistic: ", t_stat, "\n")
```

```
## Test statistic:  -0.8230881
```

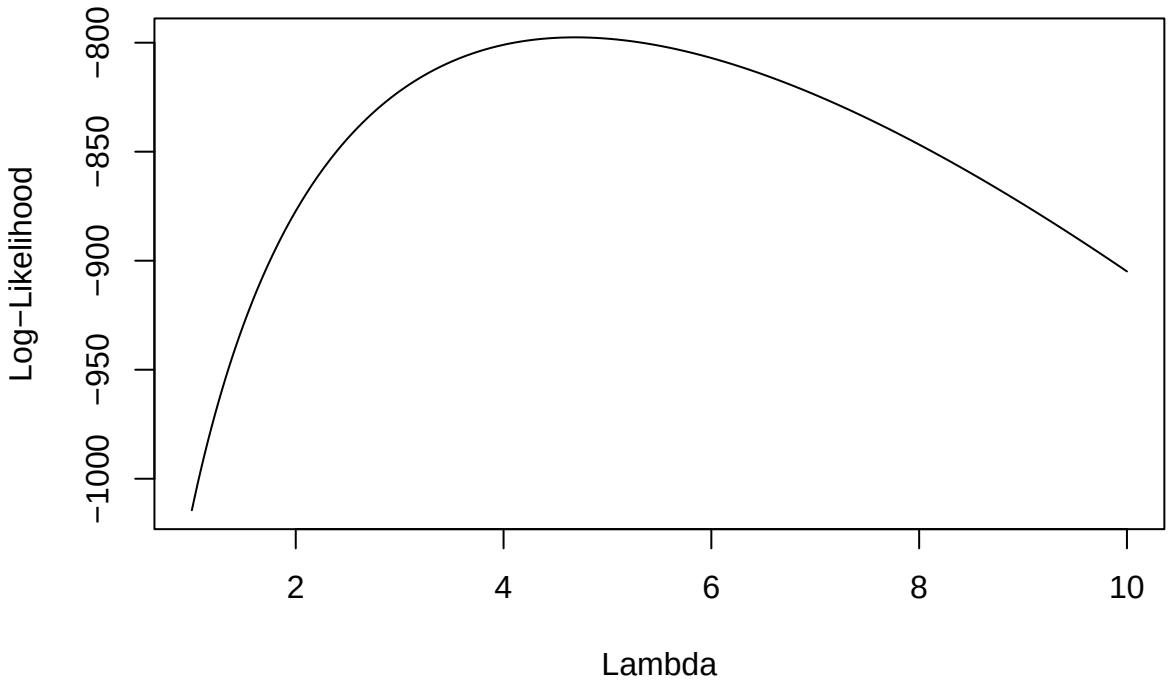
```r
df <- n_female + n_male - 2
p_value <- 2 * pt(abs(t_stat), df = df, lower.tail = FALSE)
cat("p-value: ", p_value, "\n")
```

```
## p-value:  0.4137724
```

```r
t_test_result <- t.test(Age ~ Sex, data = data, var.equal = TRUE)
cat("t-test p-value: ", t_test_result$p.value, "\n")
```

```
## t-test p-value:  0.4137724
```

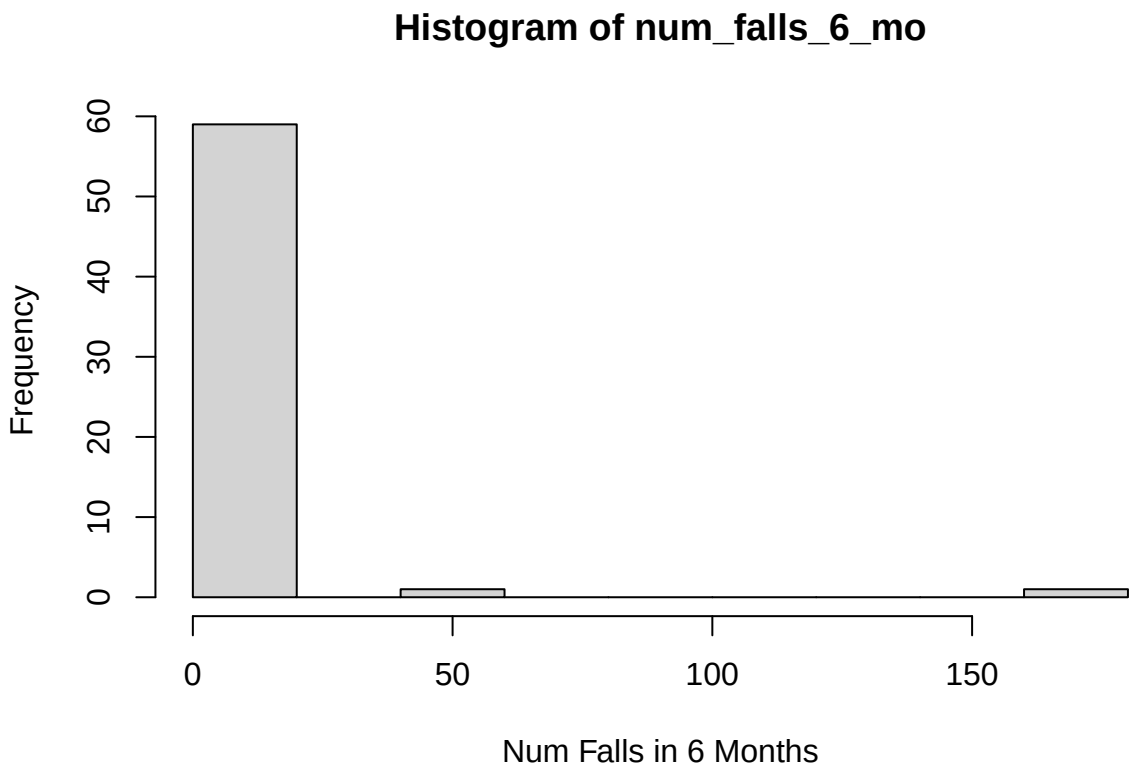
```r
lambda_seq <- seq(1, 10, by = 0.01)
loglik <- sapply(lambda_seq, function(lambda) {
  sum(dpois(data$num_falls_6_mo, lambda, log = TRUE))
})

plot(lambda_seq, loglik, type = "l", xlab = "Lambda", ylab = "Log-Likelihood")
```



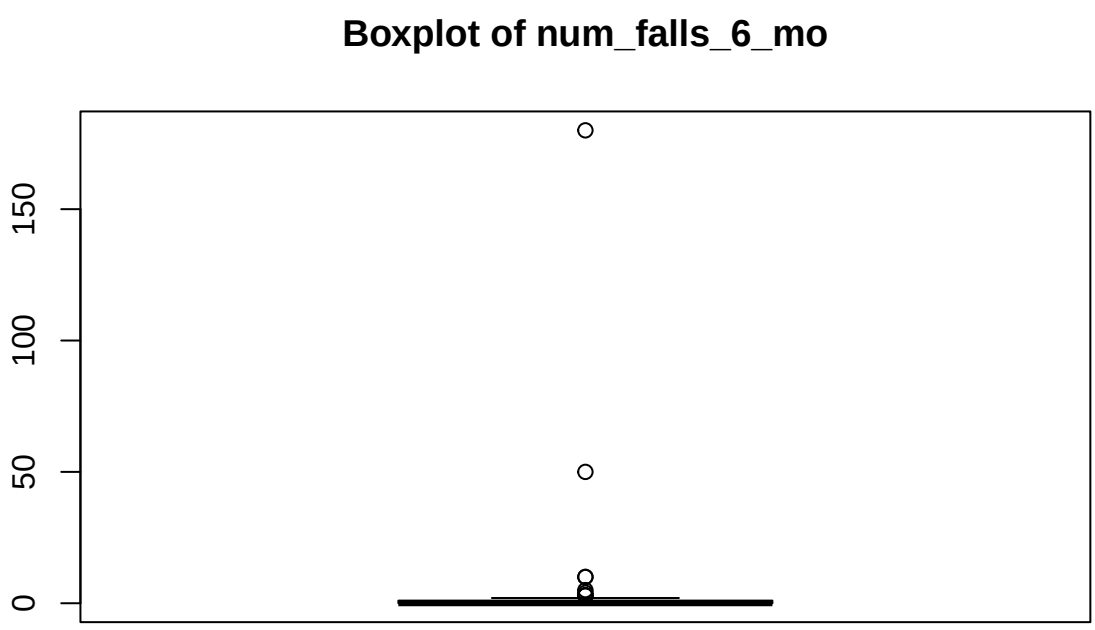

```
lambda_mle1 <- lambda_seq[which.max(loglik)]
cat("lambdaMLE1: ", lambda_mle1, "\n")
```

```
## lambdaMLE1:  4.69
```

```
hist(data$num_falls_6_mo, main = "Histogram of num_falls_6_mo", xlab = "Num Falls in 6 Months")
```

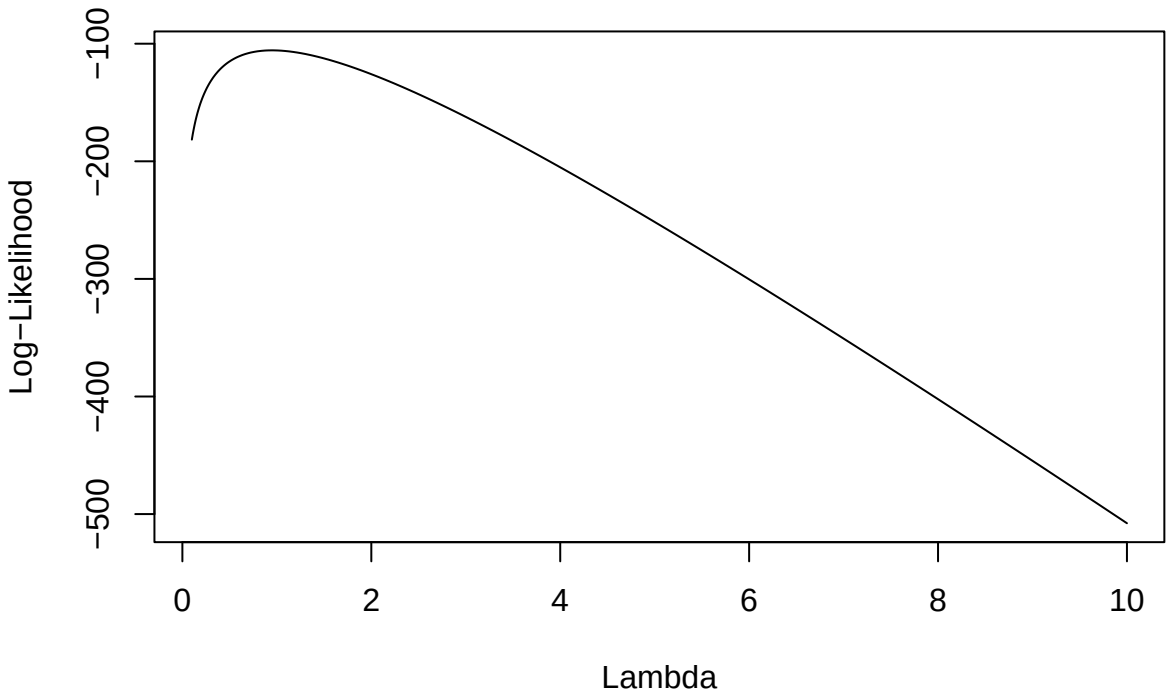



```
boxplot(data$num_falls_6_mo, horizontal = FALSE, main = "Boxplot of num_falls_6_mo", xlab = "Num Falls
```



Num Falls in 6 Months

```
data_clean <- data[data$num_falls_6_mo != 50 & data$num_falls_6_mo != 180,]

lambda_seq <- seq(0.1, 10, by = 0.01)
loglik <- sapply(lambda_seq, function(lambda) {
  sum(dpois(data_clean$num_falls_6_mo, lambda, log = TRUE))
})

plot(lambda_seq, loglik, type = "l", xlab = "Lambda", ylab = "Log-Likelihood")
```




```
lambda_mle2 <- lambda_seq[which.max(loglik)]
cat("lambdamle2: ", lambda_mle2, "\n")
```

```
## lambdamle2:  0.95
```

```
loglik1 <- sum(dpois(data_clean$num_falls_6_mo, lambda_mle1, log = TRUE))
loglik2 <- sum(dpois(data_clean$num_falls_6_mo, lambda_mle2, log = TRUE))

cat(loglik1,loglik2)
```

```
## -236.8935 -105.6501
```

```
#In comparison to the log likelihood value of -105.6501, the log likelihood value of -236.8935 is highe
```

```
total_campaign_expense = (365000*100)/12
cat(round(total_campaign_expense, digits = 0))
```

```
## 3041667
```

```
spending <- read_csv("https://jlucasmckay.bmi.emory.edu/global/bmi510/campaign-spending.csv")
```

```
## Rows: 433 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Representative, Office Running For
## dbl (1): Total Spent
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spending <- spending[, c(1, ncol(spending))]

col_names = names(spending)
new_row <- data.frame(r = "George Santos (R-NY)",t = total_campaign_expense)
names(new_row) <- col_names
spending <- rbind(spending, new_row)

spending
```

```
## # A tibble: 434 x 2
##    Representative           `Total Spent`
##    <chr>                            <dbl>
##  1 Val Demings (D-Fla)           79939789
##  2 Tim Ryan (D-Ohio)             56348529
##  3 Katie Porter (D-Calif)        28483084
##  4 Nancy Pelosi (D-Calif)        27776296
##  5 Kevin McCarthy (R-Calif)      26676447
##  6 Steve Scalise (R-La)          19963517
##  7 Adam Schiff (D-Calif)         18036600
##  8 Dan Crenshaw (R-Texas)        16095358
##  9 Ted Budd (R-NC)               15043283
## 10 Jim Jordan (R-Ohio)           12404151
## # ... with 424 more rows
```

```
spending$Rank <- rank(spending$`Total Spent`)

n_less_than_santos <- sum(spending$`Total Spent` < total_campaign_expense)
cat("Number of representatives that spent less than Santos: ", n_less_than_santos, "\n")
```

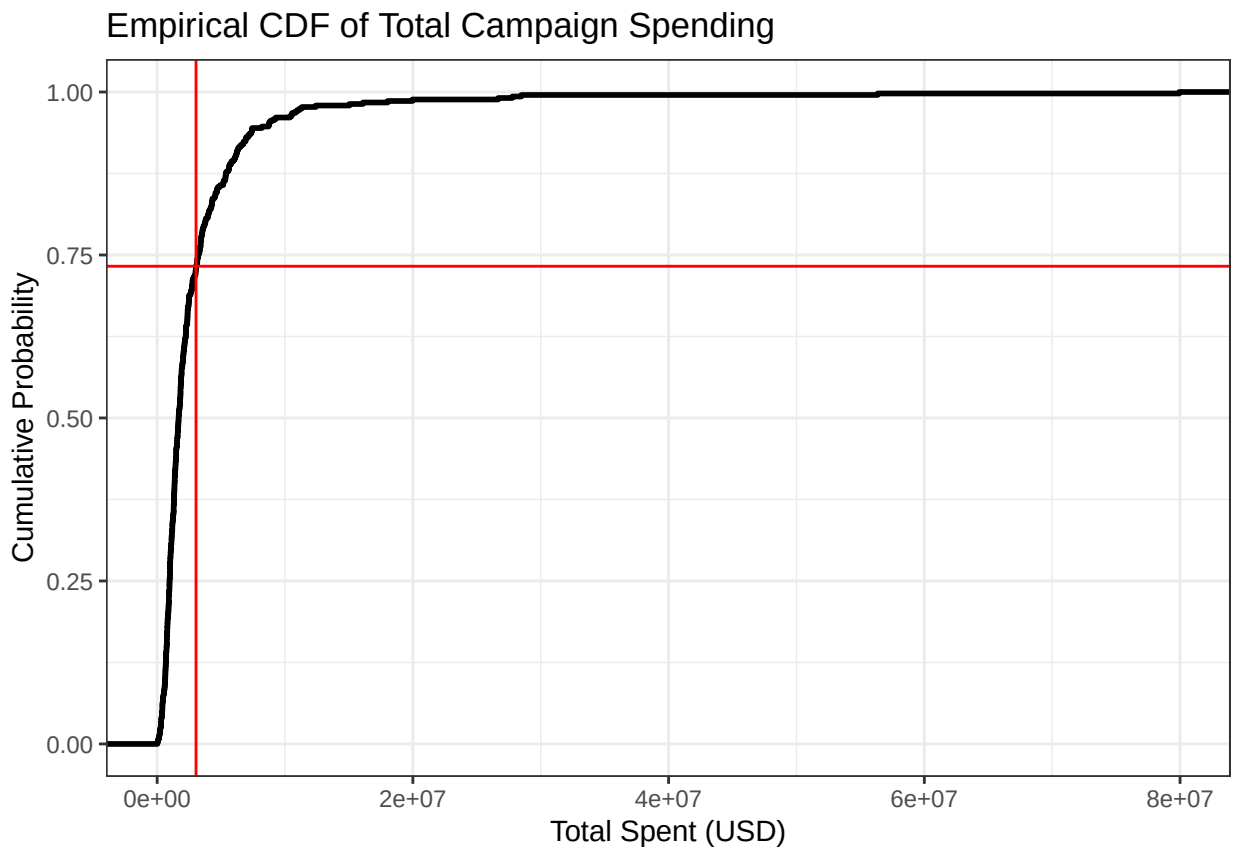
```
## Number of representatives that spent less than Santos:  317
```

```
prop_less_than_santos <- n_less_than_santos / nrow(spending)

cat("Proportion of representatives that spent less than Santos: ", round(prop_less_than_santos, 2), "\n
```

```
## Proportion of representatives that spent less than Santos:  0.73
```

```
ggplot(spending, aes(x = `Total Spent`)) +
  stat_ecdf(size = 1) +
  geom_vline(xintercept = total_campaign_expense, color = "red") +
    geom_hline(yintercept = ecdf(spending$`Total Spent`)(total_campaign_expense), color = "red") +
  labs(title = "Empirical CDF of Total Campaign Spending",
       x = "Total Spent (USD)",
       y = "Cumulative Probability") +
  theme_bw()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```

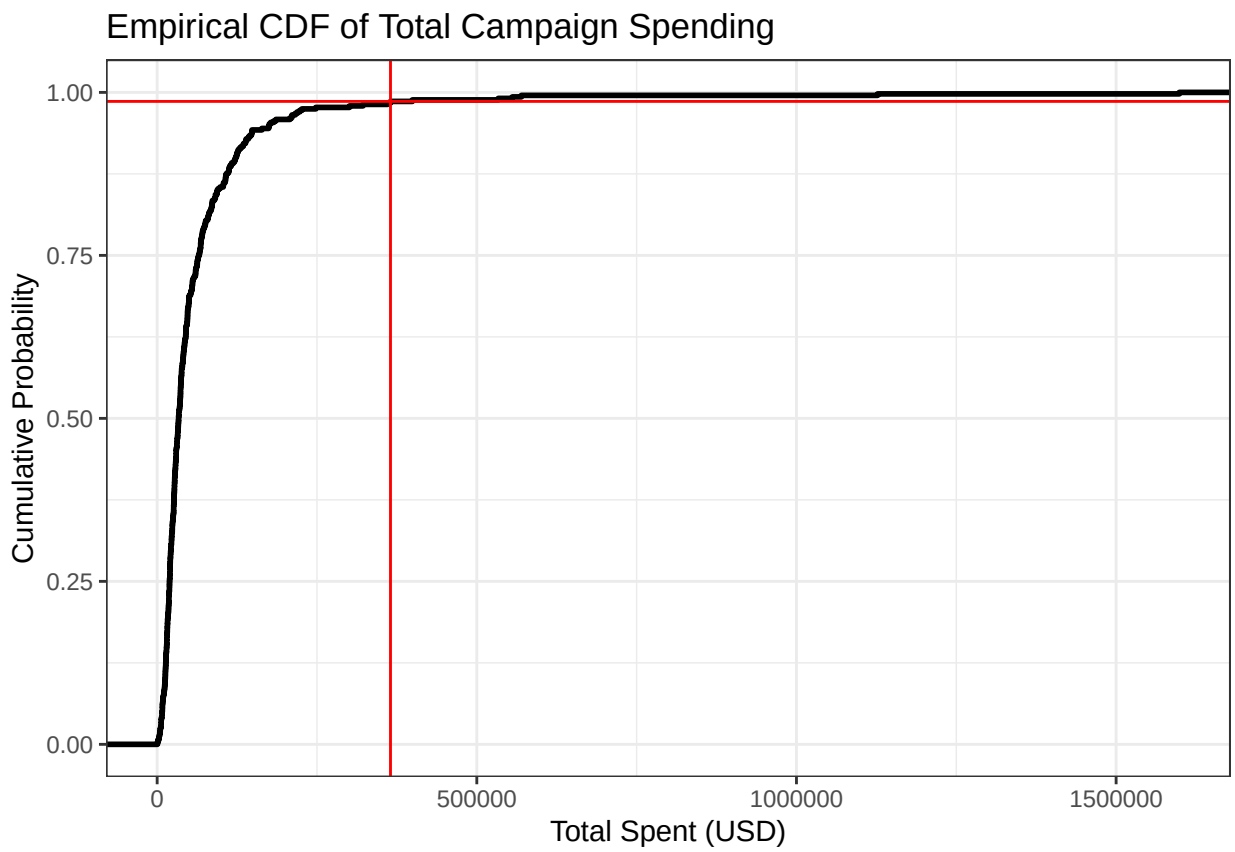



```
spending$Rank <- NULL

spending_undoc <- subset(spending, Representative != "George Santos (R-NY)")

spending_undoc$undocumented <- spending_undoc$`Total Spent` * 0.02
santos_spending <- round(total_campaign_expense * 0.12, 2)
santos_row <- data.frame(Representative = "George Santos (R-NY)", `Total Spent` = total_campaign_expens
print(santos_row)
```

```
##         Representative Total.Spent undocumented
## 1 George Santos (R-NY)     3041667       365000
```

```
#we assume that everyone else does not have documents worth 2% and santos does not have documents for 1

col_names = names(spending_undoc)
names(santos_row) <- col_names


spending_undoc <- rbind(spending_undoc, santos_row)
df_undoc_sorted <- spending_undoc[order(spending_undoc$undocumented), ]
df_totalspent_sorted <- spending_undoc[order(spending_undoc$`Total Spent`), ]
```

```
ggplot(spending_undoc, aes(x = `undocumented`)) +
  stat_ecdf(size = 1) +
  geom_vline(xintercept = santos_spending, color = "red") +
    geom_hline(yintercept = ecdf(spending_undoc$`undocumented`)(santos_spending), color = "red") +
  labs(title = "Empirical CDF of Total Campaign Spending",
       x = "Total Spent (USD)",
       y = "Cumulative Probability") +
  theme_bw()
```




```
quantile_santos <- ecdf(spending$`Total Spent`)(total_campaign_expense)
proportion_more_spending <- 1 - quantile_santos
num_more_spending <- round(proportion_more_spending * nrow(spending))

cat(num_more_spending,"had more total spendings \n")
```

```
## 116 had more total spendings
```

```
quantile_santos_undoc <- ecdf(spending_undoc$undocumented)(santos_spending)
proportion_more_spending_undocumented <- 1 - quantile_santos_undoc
num_more_spending_undocumented <- round(proportion_more_spending_undocumented * nrow(spending_undoc))

cat(num_more_spending_undocumented,"had more undocumented spendings \n")
```

```
## 6 had more undocumented spendings
```

```r
if (quantile_santos > quantile_santos_undoc) {
  atypicality <- "Santos' total spending is more atypical than his undocumented spending."
} else if (quantile_santos < quantile_santos_undoc) {
  atypicality <- "Santos' undocumented spending is more atypical than his total spending."
} else {
  atypicality <- "Santos' total and undocumented spending are equally atypical."
}

cat(atypicality)
```

```
## Santos' undocumented spending is more atypical than his total spending.
```