1 Introduction

This report records the analysis, and the predictions made on the wind speeds at the top of Arthur's Seat, based on the given historic wind speeds data. The following is assumed throughout the report.

Let $X_1, ..., X_n$, where $n = 1000$, be Independent Identically Distributed (i.i.d) random variables from a $Rayleigh(\sigma)$ distribution, where $\sigma > 0$ is an unknown parameter, and probability density function given by

$$f(x,\sigma) = \begin{cases} \frac{x}{\sigma^2}\exp\left(\frac{-x^2}{2\sigma^2}\right), & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{1}$$

2 Understanding the data

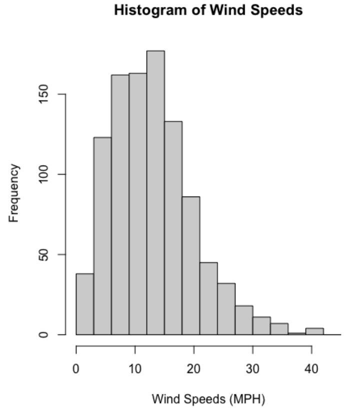

Figure 1: Histogram of Wind Data

```
> summary(wind_data)
        x
 Min.   : 0.130
 1st Qu.: 7.615
 Median :12.145
 Mean   :12.945
 3rd Qu.:16.960
 Max.   :41.870
>
> sd_ori <- sd(wind_data$x); sd_ori
[1] 6.997126
>
> IQR(wind_data$x)
[1] 9.345
>
```

Listing 1: Summary Statistics & S.d of Wind Data

Figure 1 and listing 1 shows the histogram, and numerical summaries (summary statistics, standard deviation, and Interquartile range (IQR)) of the set of wind speeds on the top of Arthur's seat. The wind speed data has a mean of 12.945mph, median of 12.145mph, standard deviation of 6.997126mph, Q1 of 7.615mph, Q3 of 16.960mph, and an IQR of 9.345mph. This means, the average wind speed on Arthur's Seat is about 12 to 13, and usually not over 16.9mph or under 7.6mph, the rest of the numbers are only in the rare cases.

From the histogram, it could be seen that the distribution is right-skewed, as expected since the wind speeds are modeled after the $Rayleigh(\sigma)$ distribution. This is also perfectly normal since wind speeds cannot be negative, hence a lower bound of 0mph, and it is possible to record wind speeds of greater than 16.960mph (3^{rd} quartile), only in some cases.

3 Maximum Likelihood Estimator (MLE) for σ

The likelihood function is

$$L(\sigma) = \prod_{n=1}^{n} f(x_n, \sigma) = \prod_{n=1}^{n} \left[\frac{x}{\sigma^2}\exp\left(\frac{-x^2}{2\sigma^2}\right)\right] \tag{2}$$

The log-likelihood function is

$$\ell(\sigma) = \log[L(\sigma)] = \left[\sum_{1}^{n} \log(x_i)\right] - 2n\log(\sigma) - \frac{1}{\sigma^2}\sum_{1}^{n}\left[\frac{x_i^2}{2}\right] \tag{3}$$

To obtain the MLE for σ, $\hat{\sigma}$, calculate the root of the first derivative of $\ell(\sigma)$, and solve for 0.

$$\frac{d}{d\sigma}\ell(\sigma) = -2n\left(\frac{1}{\sigma}\right) + 2\left(\frac{1}{\sigma^3}\right)\sum_{1}^{n}\left[\frac{x_i^2}{2}\right] = 0 \tag{4}$$

$$\implies \hat{\sigma} = \left(\frac{1}{n}\sum_{1}^{n}\left[\frac{x_i^2}{2}\right]\right)^{\frac{1}{2}} = \left(\frac{1}{2n}\sum_{1}^{n}x_i^2\right)^{\frac{1}{2}} \tag{5}$$

4 Fisher's Information $I(\sigma)$ for σ

$$\begin{aligned} nI(\sigma) = -E\left[\frac{d^2\ell}{d\sigma^2}\right] &= -E\left[\frac{d}{d\sigma}\left(-2n\left(\frac{1}{\sigma}\right) + 2\left(\frac{1}{\sigma^3}\right)\sum_{1}^{n}\left[\frac{x_i^2}{2}\right]\right)\right] \\ &= -E\left[\frac{2n}{\sigma^2} - \frac{6}{\sigma^4}\left(\frac{1}{2}\right)n(2\sigma^2)\right] \\ &= \frac{4n}{\sigma^2}, \quad \implies I(\sigma) = \frac{4}{\sigma^2} \end{aligned} \tag{6}$$

Using $Var(\hat{\sigma}) = \frac{1}{I(\sigma)}$, it is deduced that $Var(\hat{\sigma}) \approx \frac{\sigma^2}{4}$. That said, $\hat{\sigma} \sim Exp(\frac{2}{\sigma})$.

5 95% Confidence Interval (CI) for $\hat{\sigma}$

Using equations 5 and 6, a 95% confidence interval for $\hat{\sigma}$ is

$$\begin{aligned} [\sigma_L(\underset{\sim}{x}), \sigma_U(\underset{\sim}{x})] &= \hat{\sigma} \pm \left(Z_{\frac{\alpha}{2}} \times ese(\hat{\sigma})\right) \\ &= \hat{\sigma} \pm \left(1.96 \times \sqrt{(Var\,(\hat{\sigma})}\right) \end{aligned} \tag{7}$$

Through Simulations in R, $\hat{\sigma} = 10.40396$, and $\sqrt{Var\,(\hat{\sigma})} = 5.201978$ were calculated, hence, the 95% CI for $\hat{\sigma}$, $I = [\sigma_L(\underset{\sim}{x}), \sigma_U(\underset{\sim}{x})] = [0.2082665, 20.5996459]$.

6 Calculating & Estimating Y'

This section aims to Calculate, Predict and Estimate Y', the mean wind speeds for the next 1000 days.

6.1 Simulation of Y'

By letting $X_i' \sim$ Rayleigh($\hat{\sigma}$), and $Y' = \frac{1}{1000}\sum_{i=1}^{1000} X_i'$ be the predicted mean wind speeds for the next 1000 days, R was used to simulate the predicted wind speeds and estimate the distribution of Y'. The number of times of simulation was set at 10,000 times, and by using the `rrayleigh(n, scale = 1)` function from the `VGAM` package — where n is the number of observations, and scale being the Rayleigh($\hat{\sigma}$) parameter —, the histogram and numerical summaries of the observation is shown in figure 2 and listing 2.

Following the results, it could seen that within the next three years, the average mean speed for this period would stay in the range of 12.18mph and 13.85mph, with an interquartile range of 0.2902123mph.

Even the mean value of the simulated results, 13.04mph is not far off from the mean value of the current three years 12.945mph.

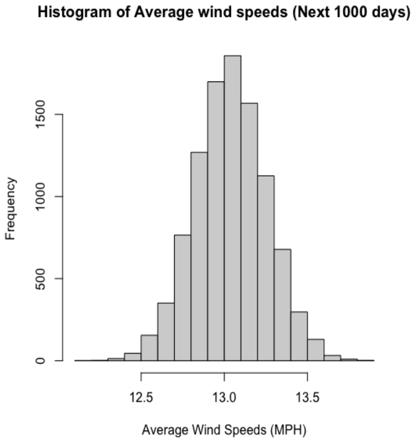

Figure 2: Histogram of Y'

```
> summary(Y)
 Min.   1st Qu.  Median   Mean    3rd Qu.  Max.
12.18   12.89    13.04    13.04   13.18    13.85
>
> sd(Y)
[1] 0.2154952
>
> IQR(Y)
[1] 0.2902123
>
> quantile(Y, c(0.025, 0.975))
    2.5%     97.5%
12.61386 13.46574
>
>
```

Listing 2: Summary Statistics of Y'

6.2 Larger Range of wind speeds in the Coming Three Years?

Researchers believe that the the variance of wind speeds will increase over the years, where $P[\text{sd}(Z') > \text{sd}(\underset{\sim}{x})]$, $\text{sd}(Z')$ is the standard deviation of the predicted sample of future wind speeds, $Z' = (X'_1, ..., X'_{1000})$, and $\text{sd}(\underset{\sim}{x})$ is the standard deviation of speeds in the current three years.

Given $P[\text{sd}(Z') > \text{sd}(\underset{\sim}{x})] = P[\text{sd}(X'_1, ..., X'_{1000}) > \text{sd}(\underset{\sim}{x})]$, and let variable $k = 0$, for each event $\{\text{sd}(X'_i) > \text{sd}(\underset{\sim}{x})\}$, where $i = 1, 2, ..., 1000$ that is true, we add 1 to the variable k. At the end of the simulation, to obtain the value of $P[\text{sd}(Z') > \text{sd}(\underset{\sim}{x})]$, we simply take $\frac{k}{\text{number of simulations}}$.

Through simulations in R, we obtain a value where $P[\text{sd}(Z') > \text{sd}(\underset{\sim}{x})] = 0.1342$, and $\frac{1342}{10000}$ simulations return a standard deviation that is greater than the standard deviation of the current three years.

6.3 Using the CI's of $\underset{\sim}{x}$, I, to check its Robustness

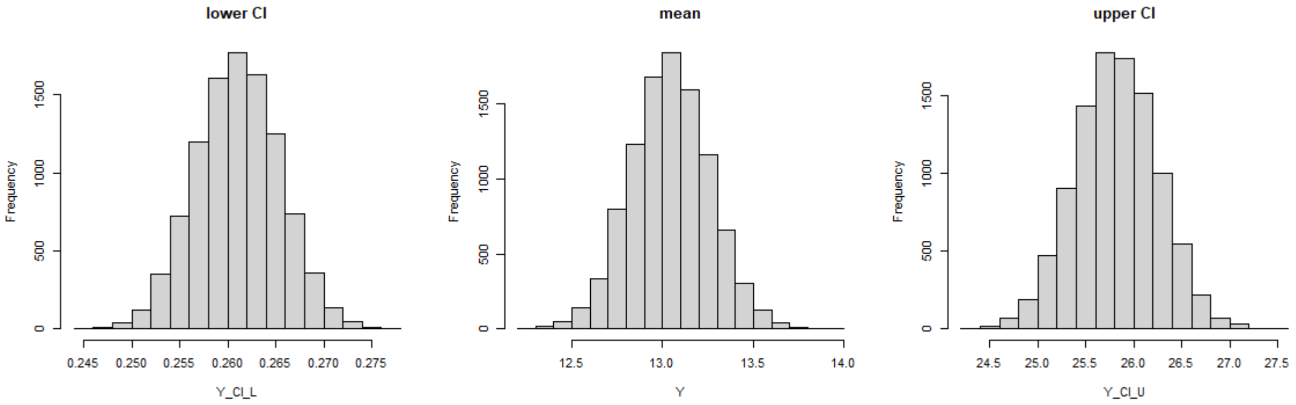

Figure 3: Individual Histograms of Mean and CI's of $\text{sd}(Z)$

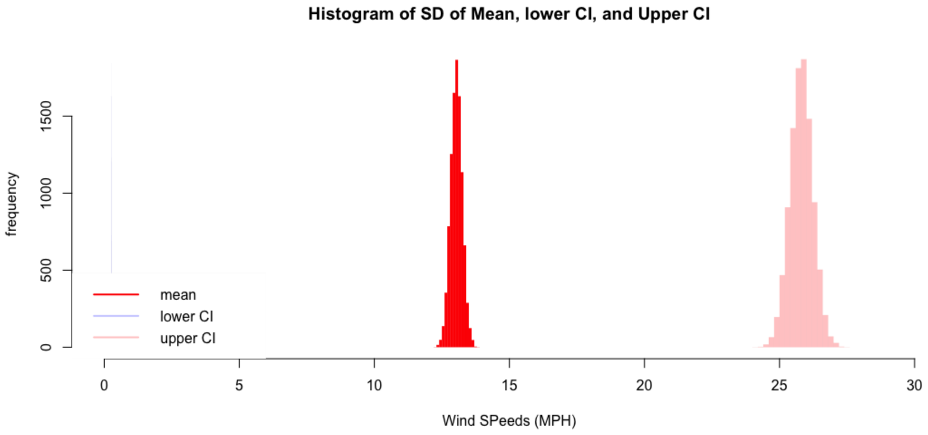

Figure 4: Histograms of Mean and CI's of sd(Z)

Figure 3 and 4 shows the histograms of CI's and Means of Z. Let $X^{'}_{i_{LowCI}} \sim$ Rayleigh$(\sigma_L(\underset{\sim}{x}))$, and $X^{'}_{i_{UpperCI}} \sim$ Rayleigh$(\sigma_U(\underset{\sim}{x}))$.

to check for robustness of Z, $P[\text{sd}(Z^{'}_{LowCI}) > \text{sd}(\underset{\sim}{x})]$, and $P[\text{sd}(Z^{'}_{UpperCI}) > \text{sd}(\underset{\sim}{x})]$ were calculated, where $Z^{'}_{LowCI}$ and $Z^{'}_{UpperCI}$ were predicted means of $[\sigma_L(\underset{\sim}{x}), \sigma_U(\underset{\sim}{x})] = [0.2082665, 20.5996459]$. Retrospectively, $Z^{'}_{LowCI} \sim$ Rayleigh$(\hat{\sigma}_{LowCi})$, and $Z^{'}_{UpperCI} \sim$ Rayleigh$(\hat{\sigma}_{UpperCi})$.

From the results from simulation in R, over 10000 iterations, results show $P[\text{sd}(Z^{'}_{LowCI}) > \text{sd}(\underset{\sim}{x})] = 0$, and $P[\text{sd}(Z^{'}_{UpperCI}) > \text{sd}(\underset{\sim}{x})] = 0.8407$. The explanation for the probabilities are simple. It is due to the positive-skewed, and non-negative nature of the distribution, where, the higher $\hat{\sigma}$ is, the higher the standard deviation yield. And since $\hat{\sigma}_{UpperCi} > \hat{\sigma}_{MLE} > \hat{\sigma}_{LowCI}$, it is only logical that $P[\text{sd}(Z^{'}_{UpperCI}) > \text{sd}(\underset{\sim}{x})] > P[\text{sd}(Z') > \text{sd}(\underset{\sim}{x})] > P[\text{sd}(Z^{'}_{LowCI}) > \text{sd}(\underset{\sim}{x})]$.

7 Conclusion

The prediction of wind speeds over the next 1000 days yielded very different results due to the large upper and lower bounds of $\hat{\sigma}$, caused by the large CI range. This means that the model and results are not robust towards the change in σ, hence the estimation of sigma within the model is very important and a slightest deviation in that will lead to very different results.

8 Appendices

8.1 R-Code

Link to GitHub: https://github.com/jonleesy/F21SA-Coursework
Recommend the Viewing of RAW code on GitHub, due to the formatting of LaTeX.

```
# Setting working directory to where .csv file is stored
# setwd("C:/Users/jonat/OneDrive - Heriot-Watt University/Own Uni/
#       PGY2/F21SA Statistical Modelling and Analysis/CW")
library(VGAM)
# Reading and storing csv
wind_data <- read.csv("wind.csv", header = TRUE, row.names = 1)
n <- length(wind_data$x)

# Question 1
# Max speed
max_speed <- max(wind_data$x)
# Histogram
par(mfrow=c(1,1), col.main ="black", col.lab=’black’)
hist(wind_data$x,
     main = "Histogram␣of␣Wind␣Speeds",
     xlab = "Wind␣Speeds␣(MPH)",
     breaks = seq(from = 0, to = round(max_speed) + 5, by = 3))
# Summary statistics
summary(wind_data)
sd_ori <- sd(wind_data$x); sd_ori
IQR(wind_data$x)
quantile(wind_data$x, c(0.025, 0.975))

# Question 2
# (1/2n * (sum of (x_i^2)))^(0.5)
sigma_hat_mle <- sqrt(sum(wind_data$x^2)/(2*n))

# Question 3
sqrt_inverse_i <- sqrt((sigma_hat_mle^2)/(4))

# Question 4
ci <- c(0,0)
z_025 <- qnorm(p = 0.025, lower.tail = FALSE)
var_change <- z_025*(sqrt_inverse_i)
ci[1] <- sigma_hat_mle - var_change
ci[2] <- sigma_hat_mle + var_change
ci_range <- ci[2] - ci[1]
ci

# Question 5, 6, 7
# Simulation by 10,000 times
m = 10000
# Y is an array of m number, where it stores the
### simulated mean value of wind speeds for the next
### 1000 days for each iteration (Q5)
Y <- numeric(m)
Y_CI_L <- numeric(m)
Y_CI_U <- numeric(m)
# SD_Y stores the binary values of 1, and 0, where 1,
### if the standard deviation of simulated values of
```

```r
###
P_SD_Y <- numeric(m)
P_CI_L <- numeric(m)
P_CI_U <- numeric(m)
for (i in 1:m){
  X_sigma <- rrayleigh(n, scale = sigma_hat_mle)
  X_CI_L <- rrayleigh(n, scale = ci[1])
  X_CI_U <- rrayleigh(n, scale = ci[2])
  Y[i] <- mean(X_sigma)
  Y_CI_L[i] <- mean(X_CI_L)
  Y_CI_U[i] <- mean(X_CI_U)
  if (sd(X_sigma) > sd_ori){
    P_SD_Y[i] <- 1
  }
  if (sd(X_CI_L) > sd_ori){
    P_CI_L[i] <- 1
  }
  if (sd(Y_CI_U) > sd_ori){
    P_CI_U[i] <- 1
  }
}
p_sd_greater <- sum(P_SD_Y)/length(P_SD_Y)
P_CI_L_greater <- sum(P_CI_L)/length(P_CI_L)
P_CI_U_greater <- sum(P_CI_U)/length(P_CI_U)

# Plotting stacked/overlapped histograms for all three Y, Y_CI_L, and Y_CI_U
par(mfrow=c(1,1), col.main ="white", col.lab='white')
c1 <- rgb(173,216,230,max = 255, alpha = 80, names = "lt.blue")
c2 <- rgb(255,192,203, max = 255, alpha = 80, names = "lt.pink")
opar <- par(lwd=0.01)
hgY <- hist(Y, plot = FALSE)
hgCIL <- hist(Y_CI_L, plot = FALSE)
hgCIU <- hist(Y_CI_U, plot = FALSE)
plot(hgY, col = 'red', xlim=c(0,30))
plot(hgCIL, add=TRUE, col=rgb(0,0,1,1/4), xlim=c(0,30))
plot(hgCIU, add=TRUE, col=rgb(1,0,0,1/4), xlim=c(0,30))
title(main = "Histogram␣of␣SD␣of␣Mean,␣lower␣CI,␣and␣Upper␣CI",
      xlab = "Wind␣SPeeds␣(MPH)", col.main="black", col.lab="black",
      ylab = "frequency")
legend ("bottomleft", legend =c("mean", "lower␣CI" , "upper␣CI"),
         col=c('red', rgb(0,0,1,1/4), rgb(1,0,0,1/4)), lty = 1, lwd =2)
par(opar)

# Individual Plots for the previous
par(mfrow=c(1,3), col.lab = "black", col.main = 'black')
# text(1,1,"First title",cex=2,font=2)
hist(Y_CI_L, main = "lower␣CI")
hist(Y, main = "mean")
hist(Y_CI_U, main = "upper␣CI")
```

References

[1] DataAnalytics.org.uk (2019) *Plot two (overlapping) histograms on one chart in R*. Available At: https://www.dataanalytics.org.uk/plot-two-overlapping-histograms-on-one-chart-in-r/ Accessed: 31 Oct 2021

[2] MIT Open Course Ware (No Date) *ML and MOM Estimates of Rayleigh Distribution Parameter*. Available At: https://ocw.mit.edu/ans7870/18/18.443/s15/projects/Rproject3_rmd_rayleigh_theory.html Accessed: 28 Oct 2021

[3] Goual. H, Et al., (2019) 'Validation of Burr XII inverse Rayleigh model via a modified chi-squared goodness-of-fit test', *Journal of Applied Statistics*, pp. 393-423. doi: 10.1080/02664763.2019.1639642