Assignment 1

Statistical Modelling: Theory and Practice

Laura Sans, Felix Pacheco, Begoña Bolós

10/4/2020

Project 1

Wind Power Forecast

```
setwd(wd)
raw_wp <- read.csv("project_data/tuno.txt", sep=" ")
```

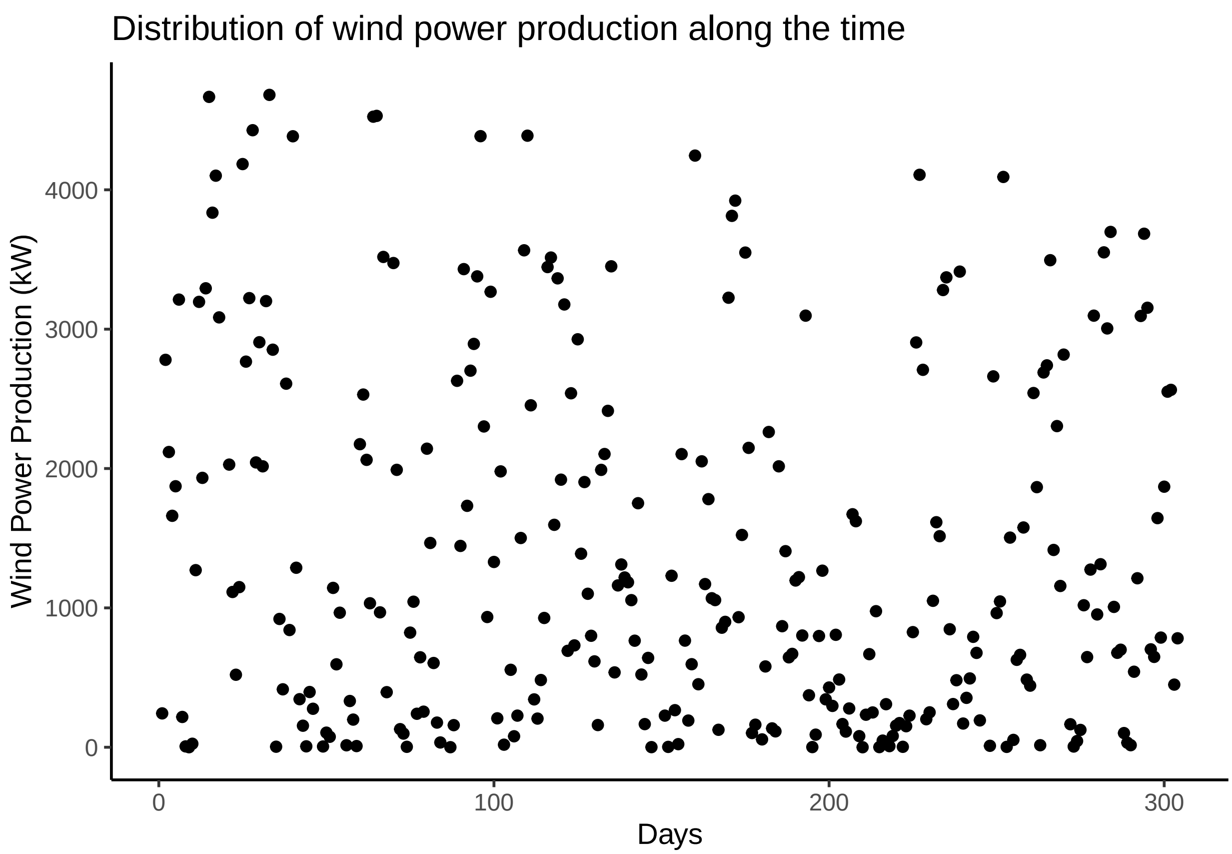
Summary statistics

```
summary(raw_wp)
```

```
##      r.day           month             day           pow.obs
##  Min.   :  1.00   Min.   : 1.000   Min.   : 1.00   Min.   :   0.123
##  1st Qu.: 78.75   1st Qu.: 3.000   1st Qu.: 8.00   1st Qu.: 254.158
##  Median :156.50   Median : 6.000   Median :15.00   Median : 964.123
##  Mean   :154.30   Mean   : 5.594   Mean   :15.47   Mean   :1381.196
##  3rd Qu.:229.25   3rd Qu.: 8.000   3rd Qu.:23.00   3rd Qu.:2196.579
##  Max.   :304.00   Max.   :10.000   Max.   :31.00   Max.   :4681.062
##       ws30             wd30
##  Min.   : 1.139   Min.   :0.000095
##  1st Qu.: 5.779   1st Qu.:2.474999
##  Median : 8.498   Median :4.079297
##  Mean   : 9.112   Mean   :3.602390
##  3rd Qu.:11.202   3rd Qu.:4.945443
##  Max.   :24.950   Max.   :6.274642
```

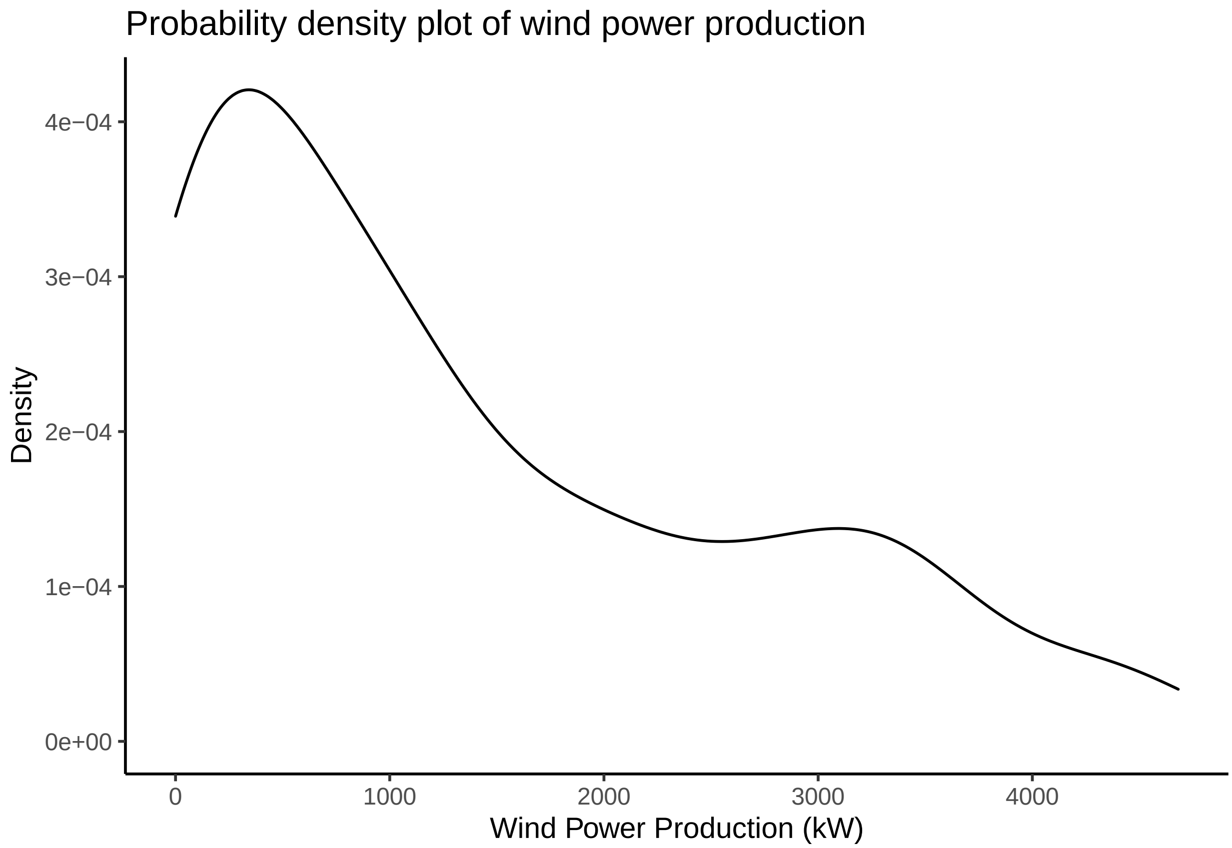
Distribution of wind power production along the time

```
ggplot(data=raw_wp, aes(x=r.day, y=pow.obs)) + geom_point()  + labs(title= "Distribution of wind power p
```



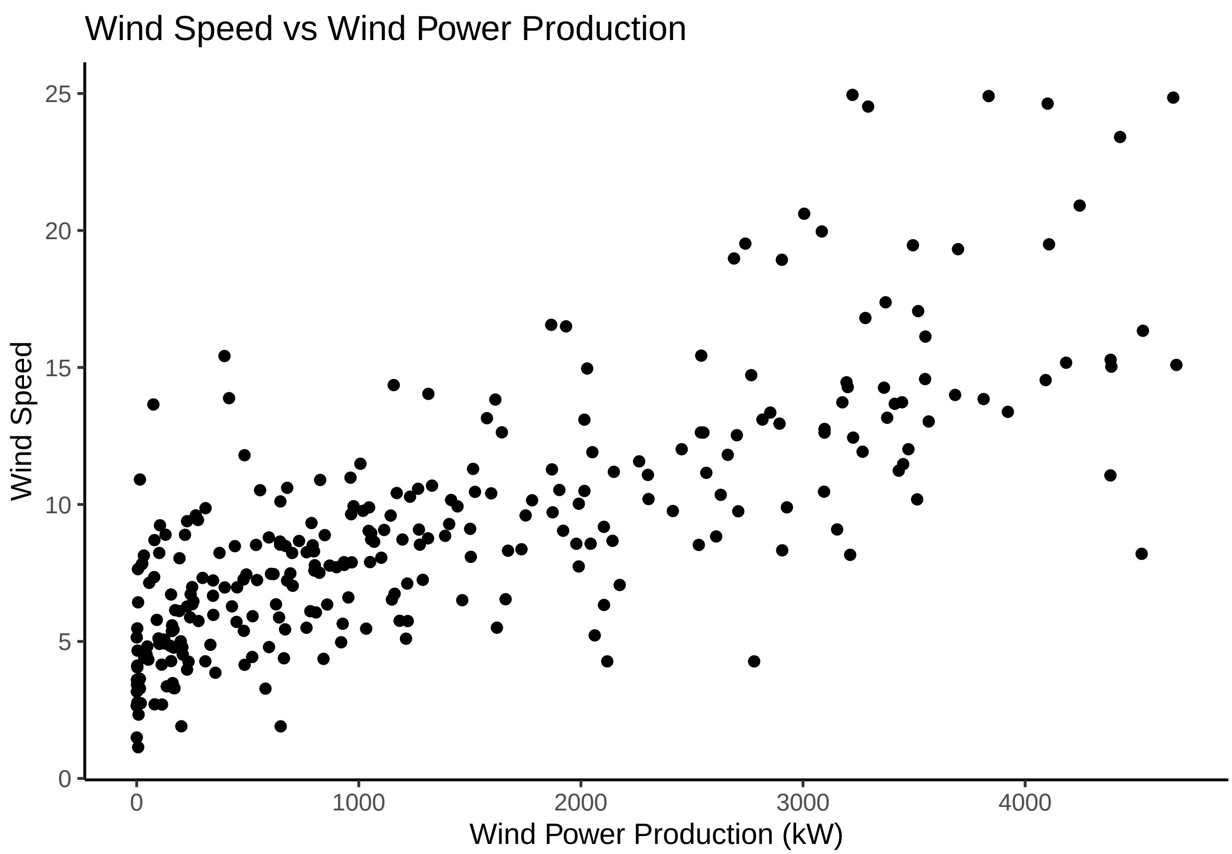
Probability density function of the Wind Power Production

```
ggplot(data = raw_wp, aes(x=pow.obs)) + geom_density()  + labs(title= "Probability density plot of wind
```



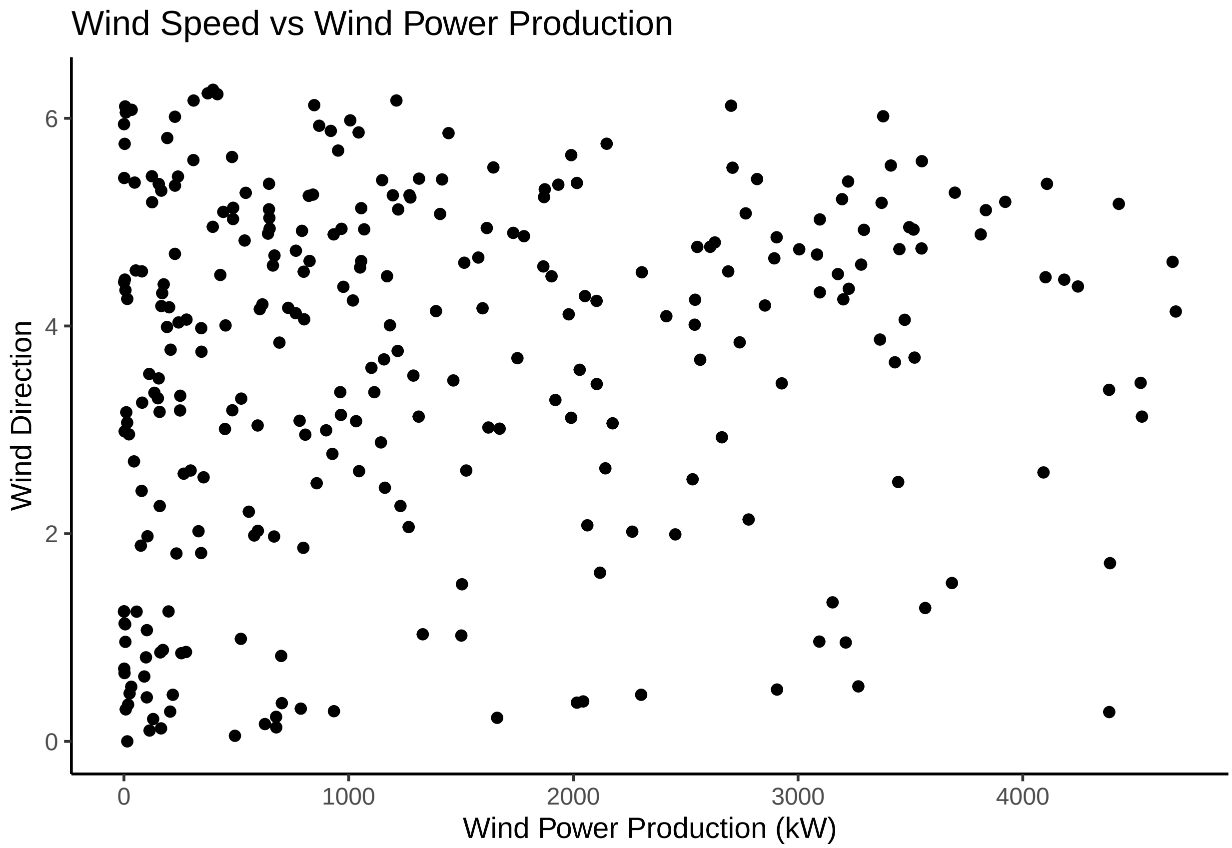
Wind Speed vs Wind Power Production

```
ggplot(data = raw_wp, aes(x=pow.obs, y=ws30)) + geom_point() + labs(title= "Wind Speed vs Wind Power Pro
```

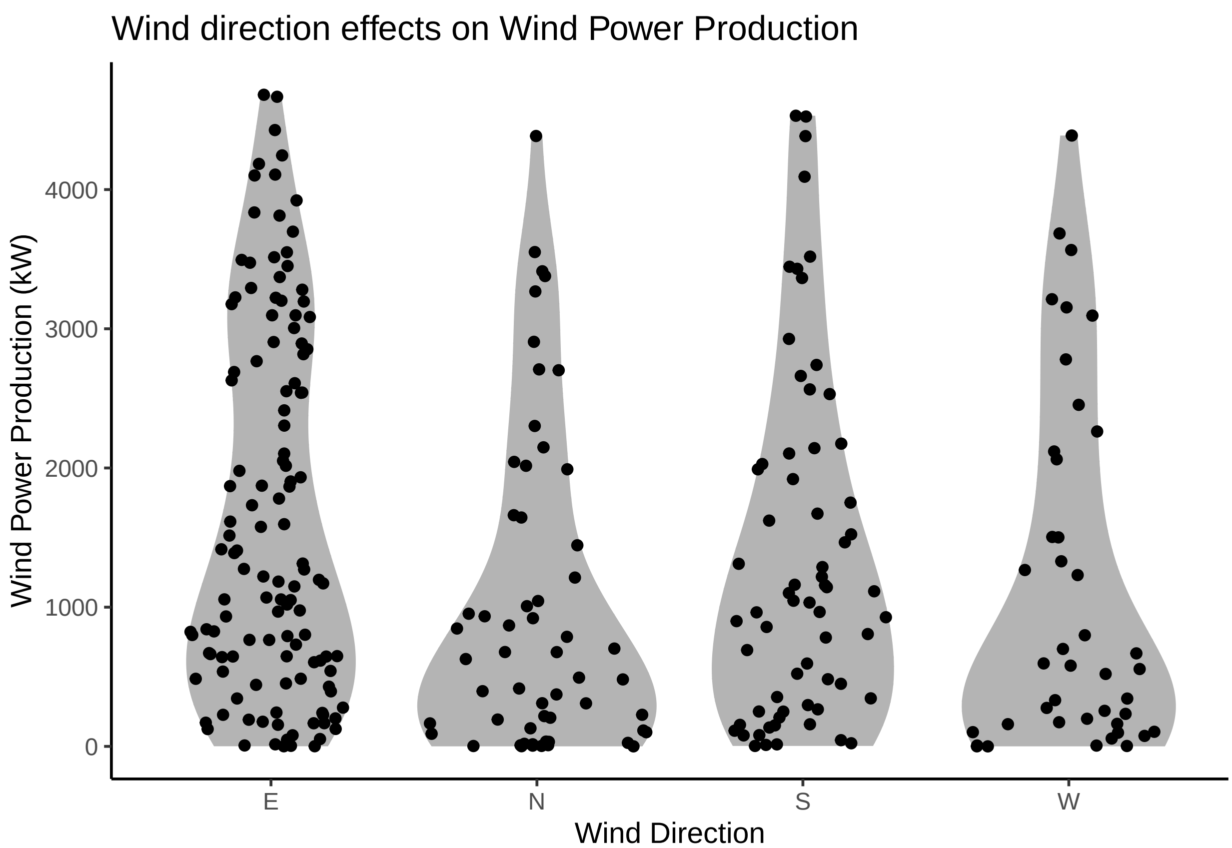


Wind direction vs Wind Power Production

```
# NOT SUPER INFORMATIVE
ggplot(data = raw_wp, aes(x=pow.obs, y=wd30)) + geom_point() + labs(title= "Wind Speed vs Wind Power Pro
```




```
# use case when

wp <- raw_wp %>%
  mutate(direction = case_when( pi/4 >= wd30  ~ "N",
                                (7*pi)/4 < wd30   ~ "N",
                                (3*pi)/4 >= wd30 & wd30 > pi/4  ~ "W",
                                (5*pi)/4 >= wd30 & wd30 > (3*pi)/4  ~ "S",
                                (7*pi)/4>= wd30 & wd30 > (5*pi)/4 ~ "E"))



ggplot(data=wp, aes(x=direction, y=pow.obs)) + geom_violin(color=NA, fill="black", alpha= 0.3, draw_quan
```



Project 2

We will first read the two dataset and store them as two variables.

Survival data (Both datasets)

```
setwd(wd)
raw_logistic <- read.csv("/Users/felix_pacheco/Desktop/DTU/semester3/stats/statistical_modelling/projec

raw_trial <- read.csv("/Users/felix_pacheco/Desktop/DTU/semester3/stats/statistical_modelling/project_d
raw_trial = raw_trial[c("time", "event", "tx")]
```

To start with the binary data, we will first compute the probabilities with a frequestist approach that we will compare to the bayesian approach (use of the likelihood.)

We first compute the probabilities using the frequestist approach, in which, a probability is assimilated to a frequency. To do so we simply compute the number of individuals having AIDS divided by the sample size. First when we consider the data without grouping by AZT treatment with probability (p__0), then with AZT treatment (p__1) and finally without treatment (p__2). The calculation can be found bellow :

$p_0 = (25 + 44)/(170 + 168) = 0, 204142$

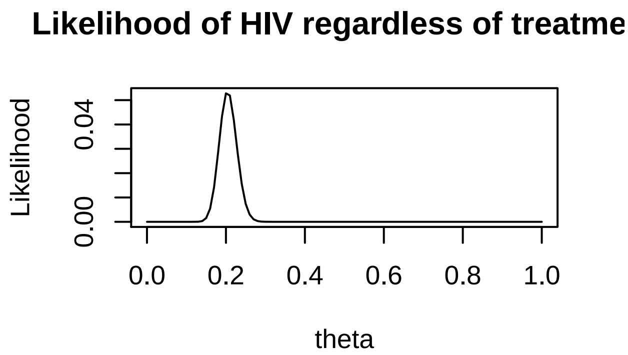
$p_1 = 25/170 = 0.1470588$

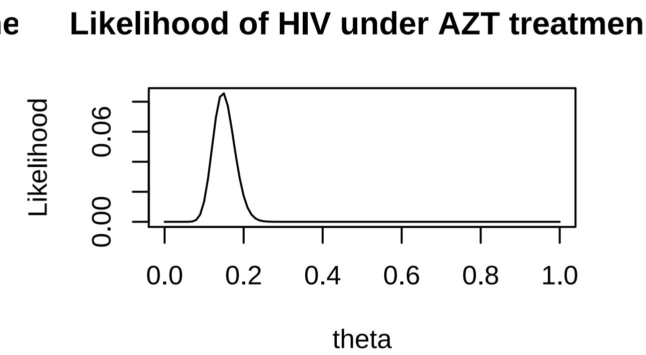
$p_2 = 44/170 = 0.2619048$

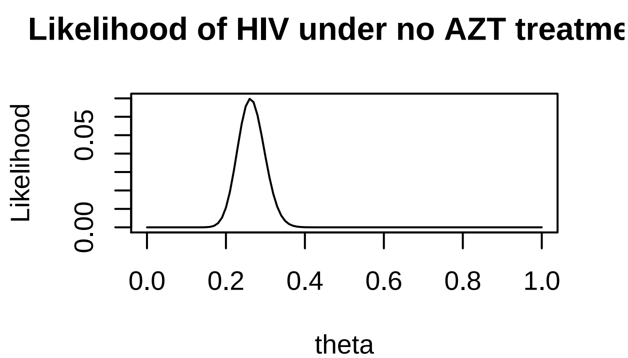
Apparently we would say that the treatment seems to have an effect but further tests should be performed to test the confidence of our hypothesis.

We will now, estimate the probabilities using likelihood approaches with the same groupings as before.

```r
# ---------- LIKELIHOOD FUNCTION ----------
Likelihood_binomial <- function(theta, x, n){
prod(dbinom(prob=theta,x=x, size=n))}
par(mfrow=c(2,2))
# ----------------------------------------

# ----- GROUPED DATA BINOMIAL FITTING -----
# Fit the binomial distribution to the data (same population joint groups).
# Binomial density parameters without grouping

n = sum(raw_logistic$n)
x = sum(raw_logistic$AIDS_yes)

# Plot the likelihood for theta [0,1] by 0.01
theta <- seq(0,1, by=0.01)
ll <- sapply(theta, FUN=Likelihood_binomial, x=x, n=n)
plot(theta,ll, type="l", main="Likelihood of HIV regardless of treatment", ylab="Likelihood")
# ----------------------------------------

# ----- NON-GROUPED DATA BINOMIAL FITTING -----
# Fit the binomial separately to the two distributions and test if there is a difference between groups

n_AZT = sum(raw_logistic$n[1])
n_no_AZT = sum(raw_logistic$n[2])
x_AZT = sum(raw_logistic$AIDS_yes[1])
x_no_AZT = sum(raw_logistic$AIDS_yes[2])

# Plot the likelihood for theta [0,1] by 0.01
theta <- seq(0,1, by=0.01)
ll_AZT <- sapply(theta, FUN=Likelihood_binomial, x=x_AZT, n=n_AZT)
ll_no_AZT <- sapply(theta, FUN=Likelihood_binomial, x=x_no_AZT, n=n_no_AZT)
plot(theta,ll_AZT, type="l", main="Likelihood of HIV under AZT treatment", ylab="Likelihood")
plot(theta,ll_no_AZT, type="l", main="Likelihood of HIV under no AZT treatment", ylab = "Likelihood")

# Test if there is a difference (We will refer to chapter 4.3)

# ----------------------------------------

# Estimate parameters in the model (p_0 probability of AIDS in control group, p_1 probability of AIDS i
```







We can observe that the likelihood maximum estimates (MLE) correspond to the frequestists approach. Additionally, we get a sense of how much uncertainty we are facing since we can see how the curvature of the function is.