BSTAT 3322 - Homework H03

by Craig W. Slinkman

September 11, 2015

This homework has two learning objectives. The first objective is to learn to use **RMarkdown**. The second objective is to learn to use R-package .

This homework must be a **Microsoft Word** document generated by the RStudio knit package. It is assumed that you have watched **R Markdown** videos on YouTube. We will cover **R Markdown** Tuesday.

The points for this homework are distributed as follows:

|  |  |  |
| --- | --- | --- |
| Question | Requirement | Percent |
| Overall | Professionalism | 20 |
| R Markdown | Word document | 10 |
| 1 | Getting data | 10 |
| 2 | Histograms | 10 |
| 3 | Descriptive statistics | 10 |
| 4 | Analyze results | 10 |
| 5 | Comparative boxplots | 10 |
| 6 | Scatterplot diagram | 10 |
| 7 | Analysis of scatterplot | 10 |
| Total |  | 100 |

You will submit your word document. Do not hide any your chunks.

# Windfarm location analysis

Windmills (Data file: **wm1** in R-package alr4 ) Energy can be produced from wind using windmills. Choosing a site for a wind farm, the location of the windmills, can be a multi-millionaire dollar gamble. If wind is inadequate at the site, then the energy produced over the lifetime of the wind farm can be much less than the cost of building and operation. Prediction of long-term wind speed at a candidate site can be an important component in the decision to build or not to build. Since energy produced varies as the square of the wind speed, even small errors can have serious consequences. The data in the file **wm1** provides measurements that can be used to help in the prediction process. Data were collected every 6 hours for the year 2002, except that the month of May 2002 is missing. The values are the calculated wind speeds in meters per second at a candidate site for building a wind farm. These values were collected at a tower erected on the site. The values and are wind speeds at a reference site, which is a nearby location for which wind speeds have been recorded over a very long time period. Airports sometimes serve as reference sites, but in this case, the reference data comes from the National Center for Environmental Modeling (NCAR, 2013). The reference is about 50 km southwest of the candidate site. Both sites are in the northern part of South Dakota. The data were provided by Mark Ahlstrom and Rolf Miller of WindLogics.

# 1. Getting the data

The data set **wm1** is stored in the **alr4** package. The steps outlined below will guide you through this process.

1. If you have not yet installed the R-package alr4 package you will need in install it. The following R-statement will install the **alr4** package for you.

install.packages( "alr4")

Be sure you get a conformation message that the package installed. You do not need to include this statement in your homework R Markdown document.

1. Now load the package **alr4** package with either the R-command library( alr4 ) or require( alr4).

require( alr4 ) # Make package functionality available.

## Loading required package: alr4  
## Loading required package: car  
## Loading required package: effects  
##   
## Attaching package: 'effects'  
##   
## The following object is masked from 'package:car':  
##   
## Prestige

Ignore the waning messages as the effects they are complaining about do not affect anything we will do on this class.

1. Load the data. We load the data in a package using the R-function data(). This statement is demonstrated below.

data( wm1 ) # Make the data available to R.

1. Now we use the R-functions head() and tail() to confirm that we have the correct dataset.

head( wm1 ) # Confirm data set.

## Date CSpd RSpd  
## 1 2002/1/1/0 6.9 5.9666  
## 2 2002/1/1/6 7.1 7.2176  
## 3 2002/1/1/12 7.8 7.9405  
## 4 2002/1/1/18 6.9 6.0174  
## 5 2002/1/2/0 5.5 6.1646  
## 6 2002/1/2/6 3.1 1.7687

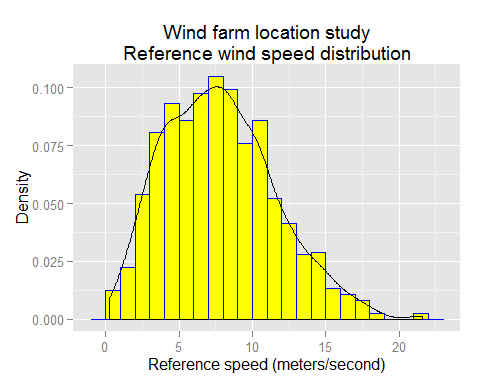
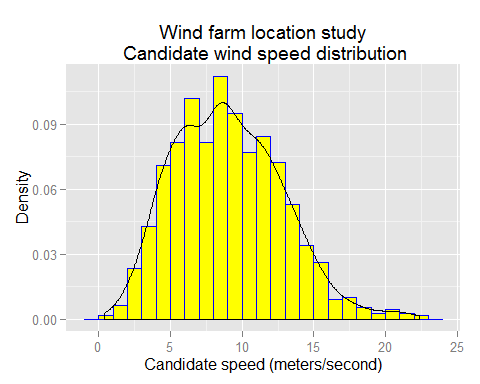
tail( wm1 )

## Date CSpd RSpd  
## 1111 2002/11/29/6 16.3 14.4395  
## 1112 2002/11/29/18 21.4 18.5025  
## 1113 2002/11/30/0 17.7 21.2215  
## 1114 2002/11/30/6 13.1 10.9579  
## 1115 2002/11/30/12 11.1 10.0234  
## 1116 2002/11/30/18 8.7 5.6299

# 2. Histograms

Draw histograms of the variables and . Your histograms should look like the histograms plotted in Figure 1. Note that you need not use the same and colors I used. Use your best judgement. If you are unsure ask your manager.

## Loading required package: ggplot2

# 3 Descriptive statistics

Use the R-functions summary() and sd() to provide numerical description of both variables in the **wm1** data set.

# 4 Analyze results

Provide a short descriptive summary of each variable. This must include the appropriate measures of location, dispersion, i.e., variability, and skewness of the sample distribution. The subsection below gives the rule for making this decision.

## 4.1 Selection

The appropriate guide for the selection of what measures to use for measures of location (centrality) and measures of dispersion (variability) are the two rules given below.

* If the shape of the distribution is **symmetric** and there are **no outliers** use **mean** as the measure location and the **standard deviation** as the measure of variability.
* Otherwise use the **median** as the measure of location and the **interquartile range** as the measure of variability.

## 5 Comparative boxplots of and

Suppose we need to use boxplots to compare the distribution of the speed of the reference site with the site. To do this we need to modify the shape of the data.

Wide format is where one or more variables essentially measure the same thing. For example in the **wm1** package the variables and both measure wind speed. They appear on the same data row in the table.

## Date CSpd RSpd  
## 1 2002/1/1/0 6.9 5.9666  
## 2 2002/1/1/6 7.1 7.2176  
## 3 2002/1/1/12 7.8 7.9405  
## 4 2002/1/1/18 6.9 6.0174  
## 5 2002/1/2/0 5.5 6.1646  
## 6 2002/1/2/6 3.1 1.7687

To do this we will use the package reshape2. Use the **Tools** menu to install the reshape2 package.

require( reshape2 ) # Load package to change the data from

## Loading required package: reshape2

# # Wide form to long form.

We know use the R-function melt. Use exactly as I specify below. That is, copy and paste it into your own script.

rswm1 <- # Change from wide form to  
 melt( wm1, # long form data.  
 variable.name = "Site",  
 value.name="Speed")

## Using Date as id variables

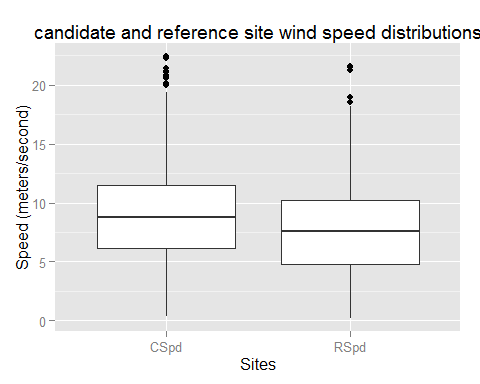
head( rswm1 ) # Display first 6 rows.

## Date Site Speed  
## 1 2002/1/1/0 CSpd 6.9  
## 2 2002/1/1/6 CSpd 7.1  
## 3 2002/1/1/12 CSpd 7.8  
## 4 2002/1/1/18 CSpd 6.9  
## 5 2002/1/2/0 CSpd 5.5  
## 6 2002/1/2/6 CSpd 3.1

tail( rswm1 ) # Display last 6 rows.

## Date Site Speed  
## 2227 2002/11/29/6 RSpd 14.4395  
## 2228 2002/11/29/18 RSpd 18.5025  
## 2229 2002/11/30/0 RSpd 21.2215  
## 2230 2002/11/30/6 RSpd 10.9579  
## 2231 2002/11/30/12 RSpd 10.0234  
## 2232 2002/11/30/18 RSpd 5.6299

Now produce a comparative boxplot with the as your variable. Follow the class standards for plots. Use the graph plotted in Figure 2 as your guide.

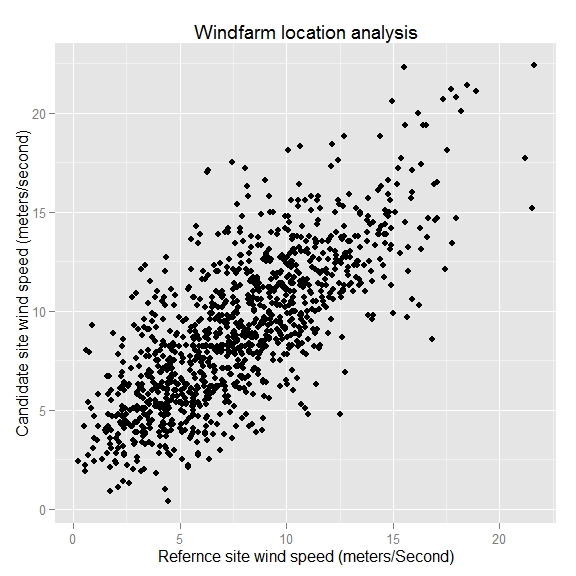


## 6 Additional analysis

Do the comparative boxplots add any information to you previous analysis in 4? If SO, what do the box plots add to your analysis?

## 7 Scatterplot diagram

Plot the Candidate speed versus the reference speed for the **wm1** data. Add rug plots to the scatterplot. Your scatter should look almost exactly like the one given Figure 3.



# 8 Analysis of scatterplot

Analyze the scatterplot diagram. Address the following issues.

1. Does there appear to be a linear relationship between the wind speed at the candidate site and the wind speed at the reference site? Justify your answer.
2. Is the relationship the wind speed at the candidate site and the wind speed at the reference site a positive or negative relationship? Justify your answer?
3. Is the relationship the wind speed at the candidate site and the wind speed at the reference site is the relationship between a weak, moderate, or a strong relationship? Justify your answer.