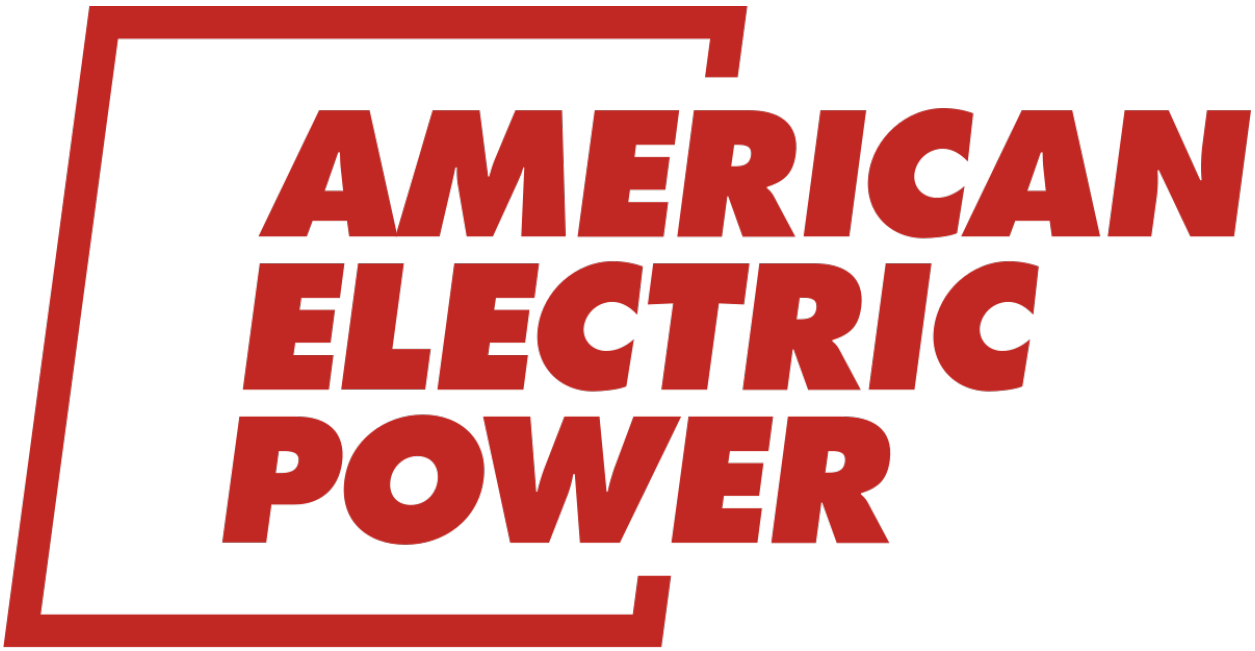
Mathematical methods of data processing. L-1

Yaroslav Kravchenko

Data pre-processing. Descriptive statistics. Exploratory analysis



Introduction

American Electric Power is a major investor-owned electric utility in the USA, delivering electricity to more than five million customers in 11 states. The aim of this paper is to outline the most significant steps of numerical data analysis on the example of overall power output(in MW) of all the AEP's facilities. The paper focuses on the following aspects of research:

1. Getting data ready and generalizing it
2. Using visualization techniques to make an assumption on the distribution type
3. Checking the assumption with one of the T-criteria methods

Time period to be analysed: 2004-12-31 01:00:00 - 2018-01-02 00:00:00 (14 years). Total amount of records: 121273. Data frame includes two columns:

* Data
* AEP_MW

Data summary

```
head(AEP_Data)
```

```
##              Datetime AEP_MW
## 1 2004-12-31 01:00:00  13478
## 2 2004-12-31 02:00:00  12865
## 3 2004-12-31 03:00:00  12577
## 4 2004-12-31 04:00:00  12517
## 5 2004-12-31 05:00:00  12670
## 6 2004-12-31 06:00:00  13038
```

Note, that ***AEP_Data*** is a data frame, whereas ***energyData***(=AEP_Data$AEP_MW) is a column(vector) and contains numerical values.

Quantiles, Mean, Variation range

```
quantile(energyData)
```

```
##    0%   25%   50%   75%  100%
##  9581 13630 15310 17200 25695
```

```
mean(energyData)
```

```
## [1] 15499.51
```

```
range(energyData)
```

```
## [1]  9581 25695
```

Measures of spread

Variance

```
var(energyData)
```

```
## [1] 6715349
```

Standard deviation

Defined as a square root of variance.

```r
sd(energyData)
```

```
## [1] 2591.399
```

Variation coefficient

```r
sd(energyData)/mean(energyData)*100
```

```
## [1] 16.71923
```

Interquartile range

3rd minus 1st quartile value.

```r
IQR(energyData)
```

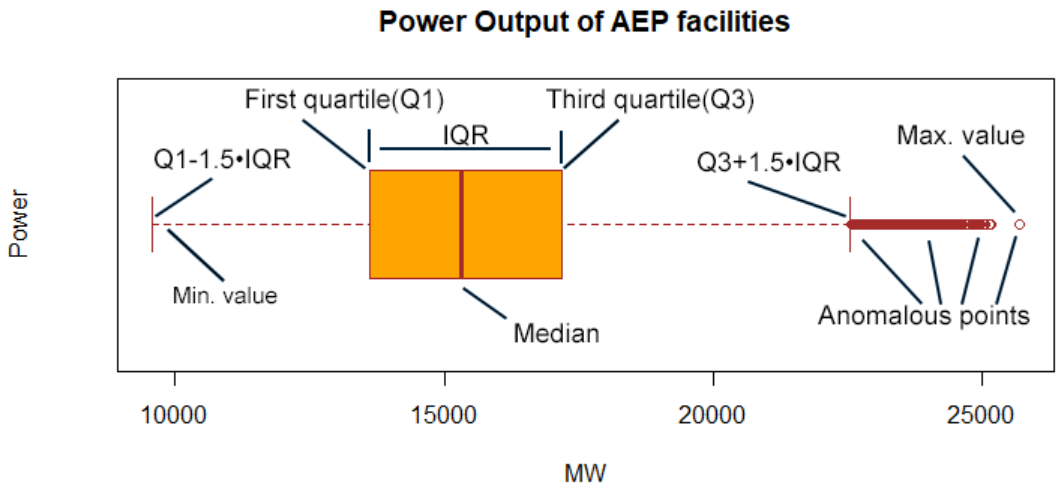
```
## [1] 3570
```

Five-point characteristic

Overall description

A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables.

```r
boxplot(AEP_Data$AEP_MW,
        main="Power Output of AEP facilities",
        ylab="Power",
        xlab="MW",
        col = "orange",
        border = "brown",
        horizontal = TRUE,
        notch = FALSE)
```



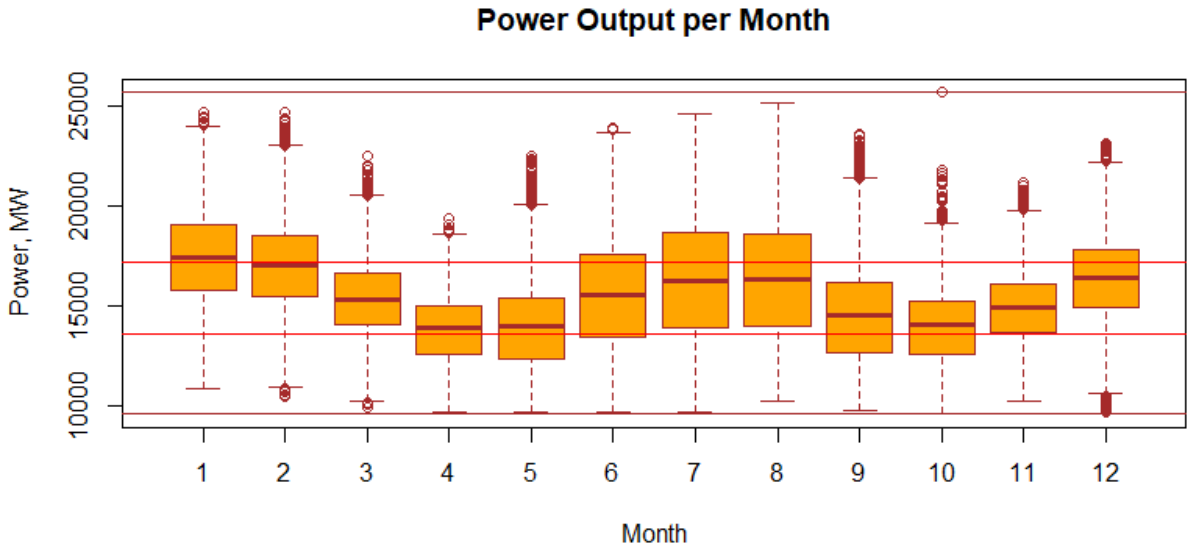
A few observations from the figure:

```
1. Minimal observed value of the researched random variable(RRV) is not anomalous, the maximal observed
2. Since mean is larger than median, the skewness of the RRV is going to be positive.
```

Boxplot monthly

```
d2_date = as.POSIXct(AEP_Data$Datetime , format = "%Y-%m-%d %H:%M:%S ")
AEP_Data$DATE = d2_date
month = lubridate::month(AEP_Data$DATE)
AEP_Data$Month = month

boxplot(AEP_Data$AEP_MW~Month, data=AEP_Data,
        main="Power Output per Month",
        ylab="Power, MW",
        col="orange",
        border="brown")
abline(h=quantile(energyData, c(0.25, 0.75)), col="red")
abline(h=range(energyData), col="brown")
```



The top- and bottom brown lines outline variation range. The two red lines in the lower part of the plot mark the 1st and 3rd quartiles. In December, February, and March anomalous observations of RRV appear below the $Q_1 - 1.5IQR$ point. The majority of anomalies though is above the $Q_3 + IQR$ point.

First and Ninth Decile

```
quantile(energyData, prob = c(0.1, 0.9))
```

```
##     10%     90%
## 12197.0 19064.8
```

Skewness and Excess kurtosis coefficient

The first is a measure of asymmetry of the probability distribution about its mean.
$+ \rightarrow$ tail to the left, - $\rightarrow$ to the right, $0 \rightarrow$ balance.
The latter measures the "fatness" of the tails of distribution.
$+ \rightarrow$ peak is plain, - $\rightarrow$ peak is sharp, $0 \rightarrow \gamma_2(X \sim \mathsf{Norm}(0,1)) = 0$.

```
skewness(energyData)
```

```
## [1] 0.3789887
```

```
kurtosis(energyData)
```

```
## [1] -0.2113649
```

The results might be visually observed in the next section.

Grouped histograms

Sturges' formula

The ***number of bins*** is calculated as following:

$$k = [\log_2(n)] + 1$$

where $[x]$ is the ceiling function.

```
buildFreq<-function(binsCount){
  ggplot(AEP_Data,aes(x=AEP_MW))+
    geom_histogram(bins=binsCount, color="white",
        aes(fill=..count..))+
    scale_x_continuous("Power, MW")+
    scale_y_continuous("Frequency")+
    scale_fill_gradient("Count", low="green", high="red")+
    labs(title = "Power Output in AEP")
}
k=nclass.Sturges(energyData)
buildFreq(k)
```

Scott's normal reference rule

The ***bin width*** is calculated using the formula:

$$h = \frac{3.49\hat{\sigma}}{n^{1/3}}$$

where $\hat{\sigma}$ is the sample standard deviation.
Though the above formula calculates the width of the bins, the return value of

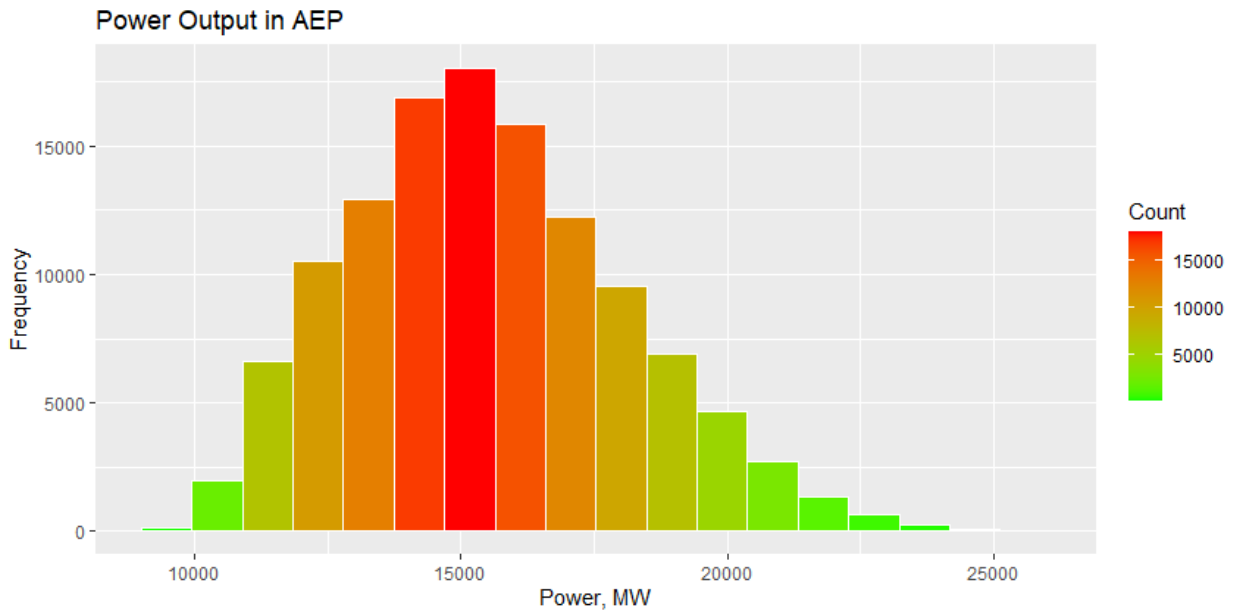


Figure 1: Using Sturges' formula to define number of bins

```
nclass.scott(energyData)
```

method is the number of bins, calculated as follows:

$$k = [\frac{\max(x) - \min(x)}{h}]$$

The braces indicate the ceiling function.

```
s=nclass.scott(energyData)
buildFreq(s)
```

Freedman-Diaconis' choice

Defined as:

$$h = 2\frac{IQR(x)}{n^{1/3}}$$

IQR denotes the interquartile range.

```
fd=nclass.FD(energyData)
buildFreq(fd)
```

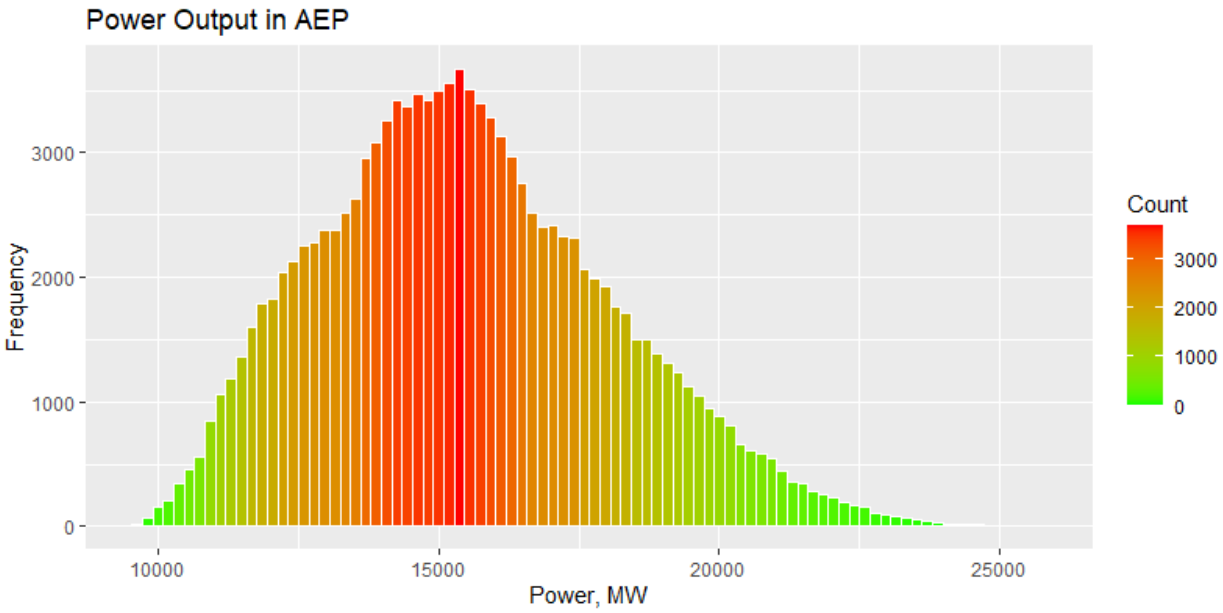


Figure 2: Using Scott's normal reference rule to define bin width

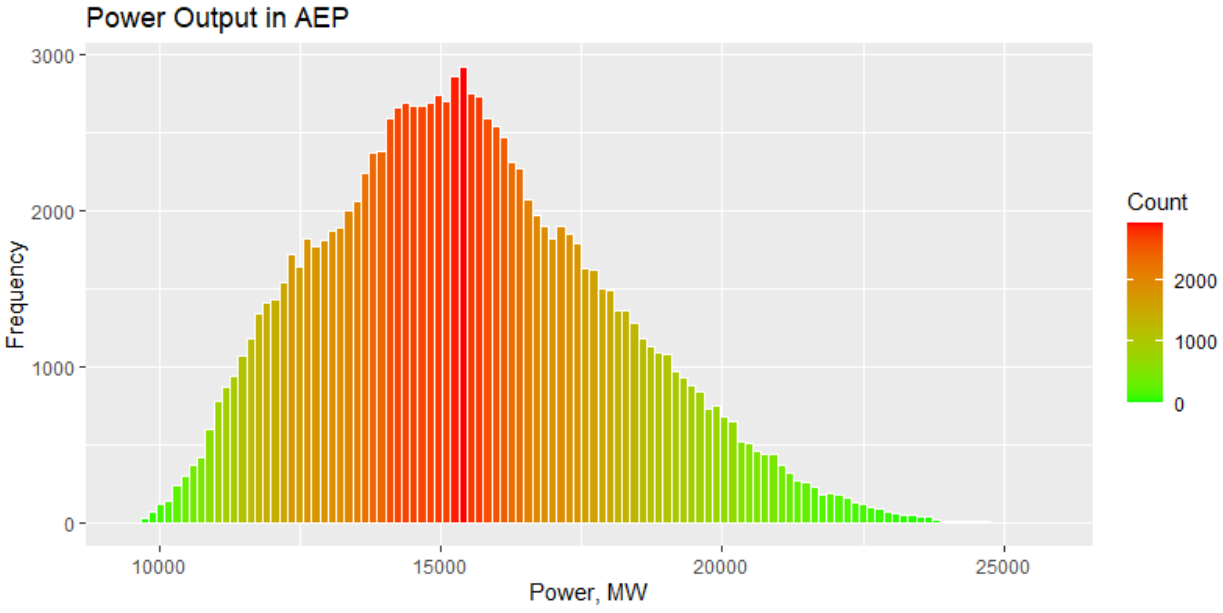


Figure 3: Using Freedman-Diaconis' choice to define bin width

Probability density function(PDF) and its kernel estimation

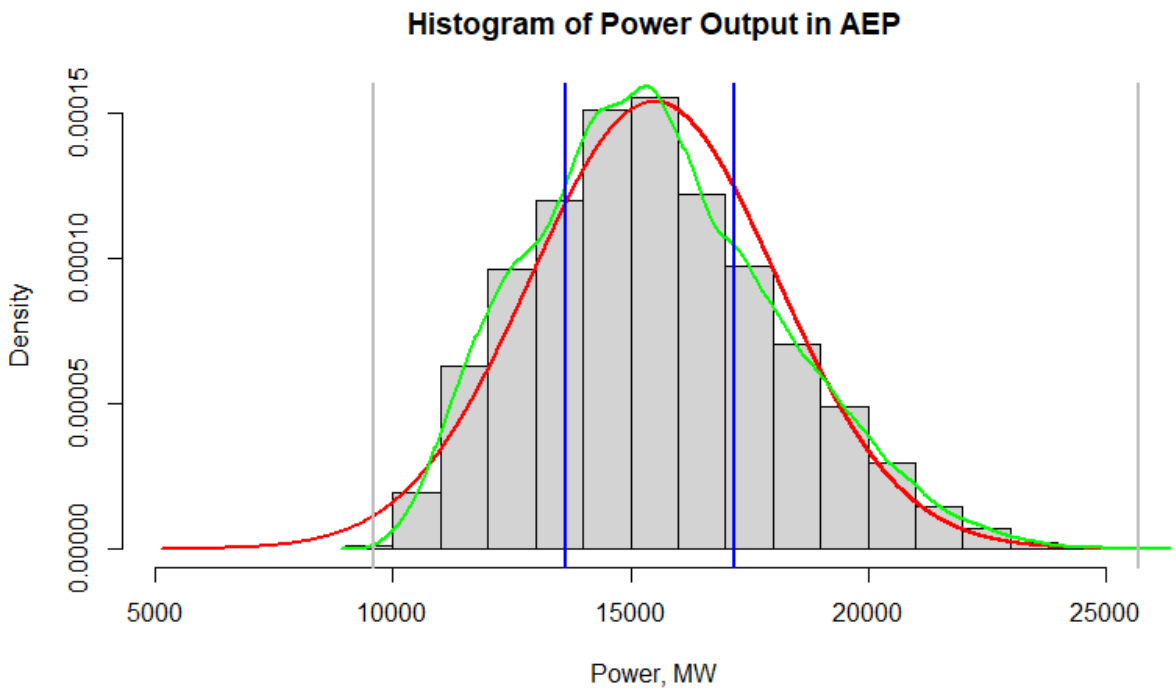
The kernel density estimation(KDE) is calculated as:

$$\hat{f}(x) = K\frac{\sum_{i=1}^{n}(x - x_i)}{h}$$

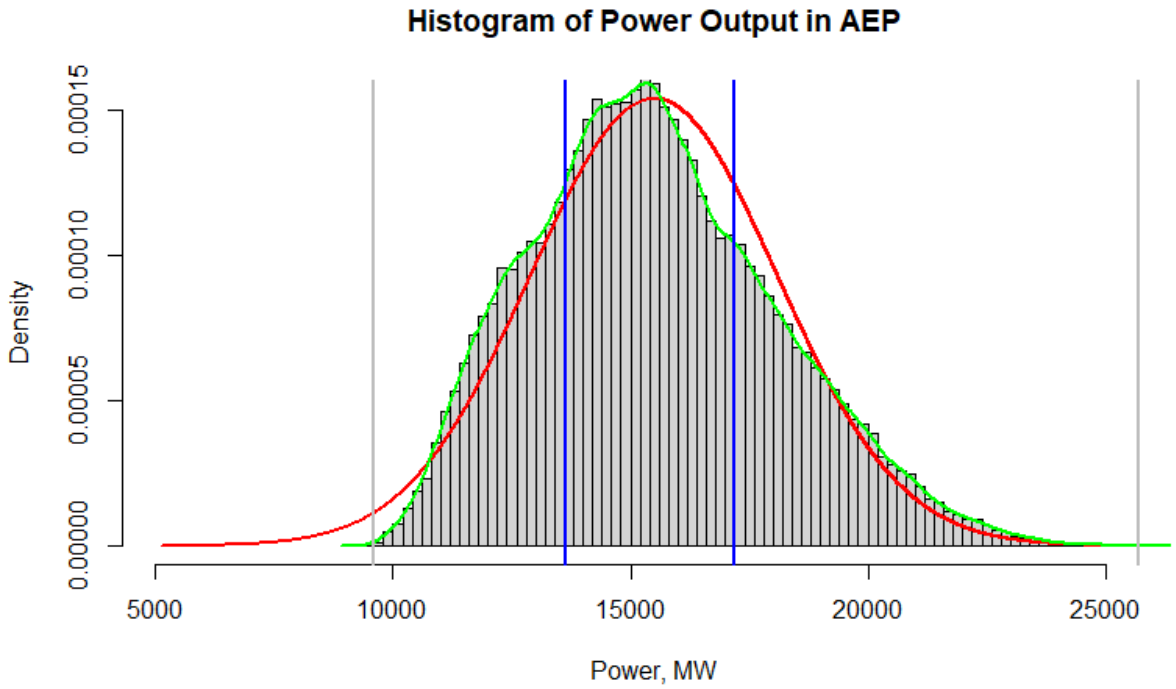
where $K > 0$ is a kernel function, and h is a bandwidth. KDE does not depend on the parameters of the RRV. Nevertheless the shape of the KDE curve is defined by width of bins h. The red curve in the following plots is the PDF of the random variable which is hypothetically distributed as $N \sim (\mu; \sigma^2)$, where μ is mean and σ^2 is variance. The green curve displays the KDE function. The blue lines are 1st and 3rd quartiles and the gray ones are minimum and maximum values of the observable range.

```
hist_and_norm <- function(energyData){
  hist_min <- mean(energyData) - 4*sd(energyData)
  hist_max <- mean(energyData) + 4*sd(energyData)
  normalx <- seq(hist_min, hist_max, by = 1)
  normaly <- dnorm(normalx, mean = mean(energyData), sd = sd(energyData))
  hist(data,
       main = "Histogram of Power Output in AEP",
       xlab = "Power, MW",
       xlim = c(hist_min, hist_max),
       ylim =c(0, max(normaly)),
       breaks = nclass.Sturges(energyData),
       #alternatively .scott(energyData)
       freq = FALSE)
  lines(normalx, normaly, col = "red", lwd="2")
  lines(density(energyData), col="green", lwd="2")
  abline(v=quantile(AEP_Data$AEP_MW, c(0.25, 0.75)), col="blue", lwd="2")
  abline(v=range(energyData), col="grey", lwd="2")
}
hist_and_norm(AEP_hourly$AEP_MW)
```

Sturges'numbers of bins



Scott's number of bins



Hypothesis proof

Two core visualization techniques should be taken into consideration before using a hypothesis proof criterion. P-P plots are used assess the agreement of the two cumulative distribution functions(CDFs) - empirical and theoretical. Q-Q plots allow us to compare CDFs by plotting their quantiles against each other.

P-P plot

Why use P-P plots?

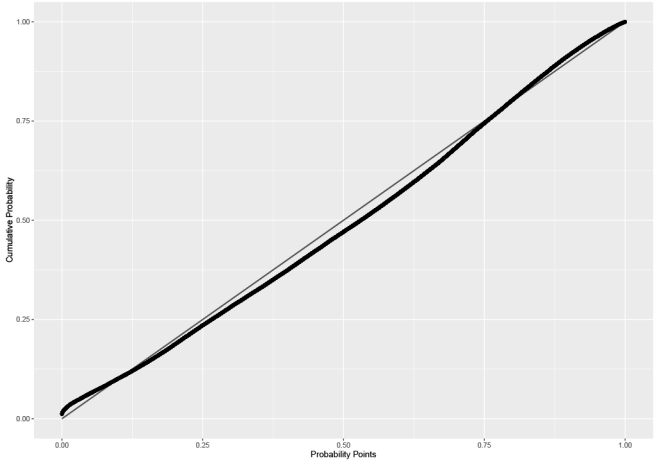
```
1. They are well suited to compare regions of high probability density (center of distribution) because
2. P-P plots can be used to visually evaluate the skewness of a distribution
3. The plot may result in weird patterns (e.g. following the axes of the chart) when the distributions a
```

The following listing uses extremely resource-demanding method

```
stat_pp_band()
```

for drawing P-P confidence bands. For this reason large data sets may overwhelm the computational machine's capability unless it is powerful enough to manage this. For instance 4GB of DDR was not enough to place more than 925 MB of data to complete the computation in decent time. That is why they they may not be seen in the plot.

```
buildPPplot<-function(){
  m <- mean(AEP_Data$AEP_MW)
  s <- sd(AEP_Data$AEP_MW)
  dp <- list(mean = m, sd = s)
  gg <- ggplot(data = AEP_Data, mapping = aes(sample = AEP_MW)) +
    stat_pp_band(distribution = "norm", dparams = dp) +
    qqplotr::stat_pp_line() +
    stat_pp_point(dparams = dp) +
    labs(x = "Probability Points", y = "Cumulative Probability")
  gg
}
```



Assume the variation range

$$x_{(0)} < x_{(1)} < ... < x_{(n)}$$

is given. Then the empirical CDF $\hat{F}_n$ will be defined as:

$$\hat{F}_n(x_i = i/n)$$

And the P-P plot is therefore built by points with coordinates

$$(F_{\xi_0}(x_i), i/n)$$

Hence the points on the y-axis are "fixed", whereas and the theoretical CDF of the RRV then takes the observable values from the variation range as an argument. Such a comparison gives insight into how different the two CDFs are by measuring "skewness" of the theoretical CDF with respect to the empirical one in every point of the variation range.

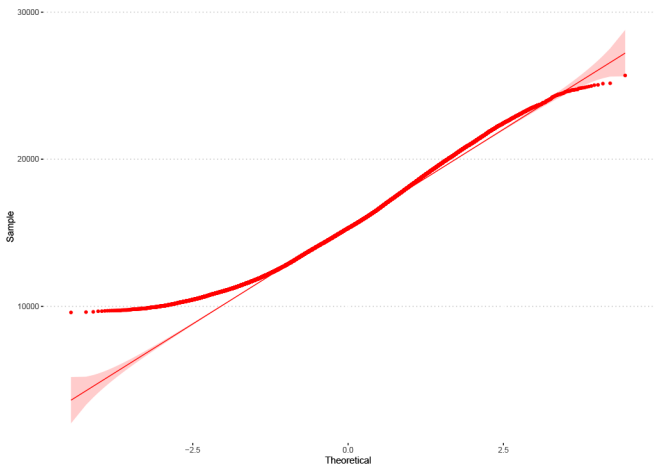
Q-Q plot

A few points on Q-Q plots:

```
1. Interpretation of the points on the plot: a point on the chart corresponds to a certain quantile com
2. Q-Q plots do not require specifying the location and scale parameters of the theoretical distributio
3. They can be used to visually evaluate the similarity of location, scale, and skewness of the two dis
```

This line of code is not as resource-violent as the previous one and yet it takes more time to build the Q-Q plot than a histogram.

```
ggqqplot(AEP_Data, x="AEP_MW", color = "red", ggtheme = theme_pubclean())
```



Expression

$$F_{\xi_0}^{-1}(\hat{F}_n(x_{(i)}))$$

defines the position of theoretical quantiles on the x-axis and takes empirical quantiles as argument, which

are situated on the y-axis respectively. The plot above implies that there are a lot more empirical quantiles than the theoretical ones in range $[9581; \sim 12000]$. For this reason the curve is somewhat flat in that range comparing to the

$$y = sx + a$$

line where a and s are unknown parameters. The rest of the variation range quantiles lie quite close to the theoretical CDF line, so the quantiles are almost "normal" in this case. The deviance residuals represent the contributions of individual samples to the deviance.

Pearson's chi-squared test

Person's criterion is reasonable for large data sets and checks exactly the correlation with the normal distribution.

```
library(nortest)
pearson.test(energyData)
```

```
##
##  Pearson chi-square normality test
##
## data:  energyData
## P = 4998.8, p-value < 2.2e-16
```

However it does not seem to be fair in case of the AEP dataset. The p-value is extremely low, which is because of:

```
1. The initial hypothesis is wrong.
OR
2. The p-value estimation failed due to large accumulated amount of minor elements which severely affec
```

Conclusion

Visual data implies that the RRV is normally distributed and moreover with high correlation rate. Nevertheless Pearson's chi-squared test does not provide definite answer on whether the initial distribution hypothesis is correct or not. Further analysis is therefore necessary to make any objective conclusions.