Describing Numerical Data

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

1 Introduction

2 Single Quantitative Variable Exploration
- Numerical Summaries
- Graphical Summaries: Analysis
- Graphical Summaries: How to Plot

3 Association Between Two Variables
- Two Quantitative Variables
- One Categorical and One Quantitative Variable
- Two Categorical Variables

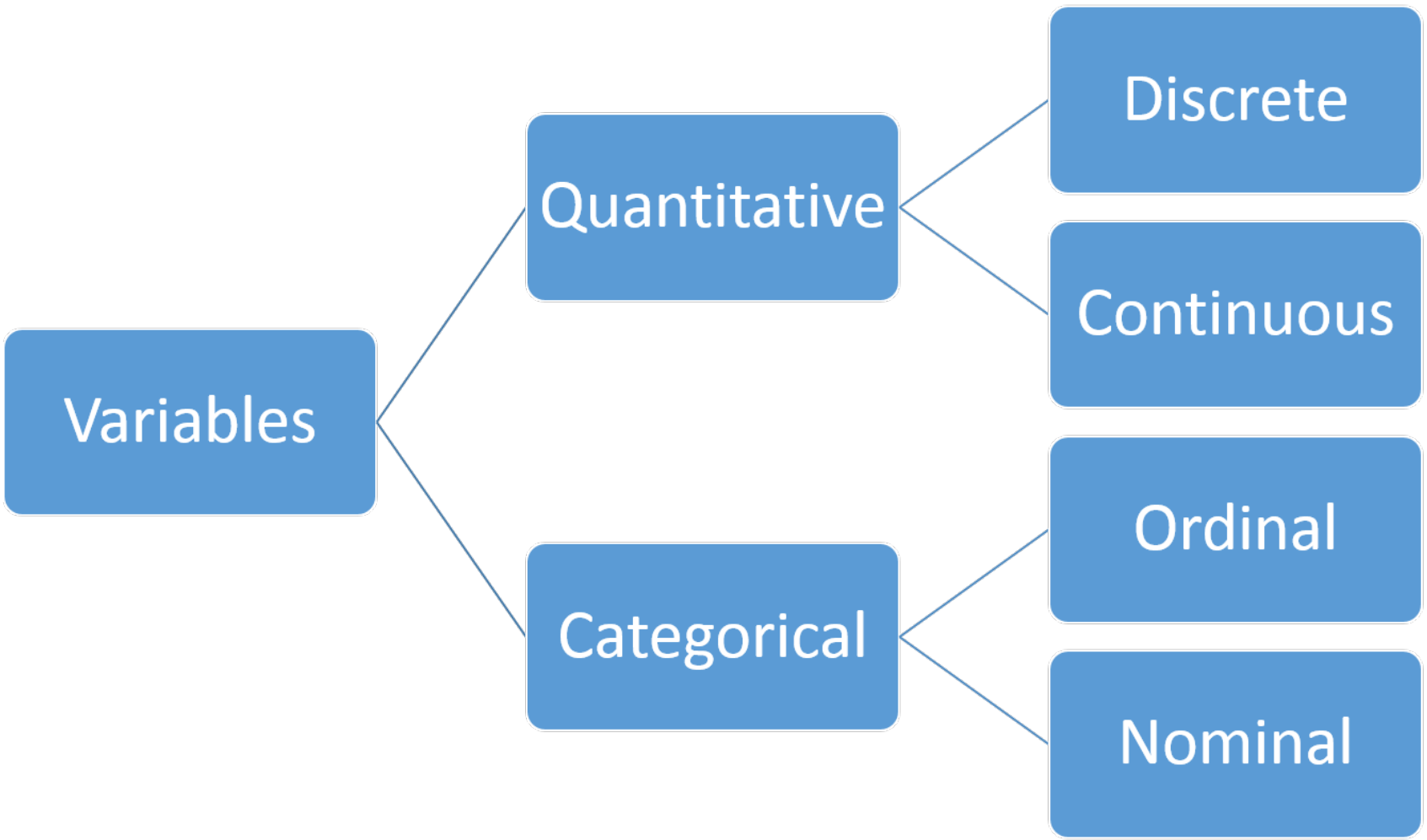
Introduction: Parameter and Statistic

Definition (Parameter and Statistic)

- A **parameter** is a numerical summary of the population. It is unknown.
- A **statistic** is a summary of a sample taken from the population. We compute it based on the data in our sample. There are two kinds of statistics:
  - Descriptive statistics
  - Inferential statistics

We need to use statistics, computed from the sample, to make inferences about a population parameter.

Types of Data



Introduction: Descriptive Statistics

- There are two major ways of describing data descriptively: numerical and graphical summaries.

- This topic will cover the descriptive statistics: **numerical and graphical summaries for single quantitative variable** and then summaries for the **association between two variables** (one categorical one quantitative or both quantitative variables) in the data.

- The inferential statistics (like hypothesis testing for one single sample or multiple samples will be introduced in later topics).

1 Introduction

2 Single Quantitative Variable Exploration
- Numerical Summaries
- Graphical Summaries: Analysis
- Graphical Summaries: How to Plot

3 Association Between Two Variables
- Two Quantitative Variables
- One Categorical and One Quantitative Variable
- Two Categorical Variables

Numerical and Graphical Summaries

- There are two major ways of describing numerical data:

  - **Numerical summaries**/descriptive measures: number of observations (sample size), location, variability and other measures.

  - **Graphical summaries**: histogram, boxplot, QQ plot (for checking normality of a dataset), scatter plot for bivariate data.

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

An Example: Midterm Mark

- Through out this section, we'll use a sample of midterm test score of 98 students (who took a ST21xx before).

- The sample is: 15.0, 18.0, 27.0, 20.0, 15.0, 19.5, 19.0, 24.5, 14.0, 24.5, 25.0, 24.0, 20.5, 13.5, 1.0, 27.0, 13.5, 21.0, 17.0, 15.5, 11.0, 19.5, 20.0, 13.5, 26.0, 7.0, 12.0, 25.0, 16.0, 28.0, 1.0, 26.0, 25.5, 6.0, 23.0, 23.5, 26.0, 14.5, 19.0, 27.5, 28.0, 26.5, 14.0, 1.0, 27.0, 17.0, 2.5, 18.0, 17.5, 8.0, 10.0, 12.5, 24.0, 15.0, 10.0, 10.5, 11.5, 16.5, 11.5, 11.5, 25.0, 14.0, 8.0, 27.5, 9.0, 18.5, 13.0, 20.0, 27.0, 26.0, 19.5, 24.5, 11.5, 23.0, 9.0, 9.0, 19.5, 13.5, 7.5, 24.0, 19.0, 11.0, 16.5, 19.5, 17.5, 21.0, 23.0, 27.0, 25.5, 27.5, 0.5, 13.0, 21.0, 7.0, 24.0, 6.5, 22.0, 27.5.

Summarizing the Center of Data

- Center of data should include the information on: mean, median and mode.

- About the midterm mark data, we can roughly describe:
  - all are positive (no zero mark);
  - range from 0.5 (min) to 28 (max);
  - there are few very low values: 0.5 or 1.

Creating/importing data in R

- Creating/importing data: you can key in every single value in R by:

```
> mark <- c(15.0, 18.0, 27.0, 20.0, 15.0, 19.5, 19.0, 24.5,
+           14.0, 24.5, 25.0, 24.0, 20.5, 13.5, 1.0, 27.0, 13.5,
+           21.0, 17.0, 15.5, 11.0, 19.5, 20.0, 13.5, 26.0, 7.0,
+           12.0, 25.0, 16.0, 28.0, 1.0, 26.0, 25.5, 6.0, 23.0,
+           23.5, 26.0, 14.5, 19.0, 27.5, 28.0, 26.5, 14.0, 1.0,
+           27.0, 17.0, 2.5, 18.0, 17.5, 8.0, 10.0, 12.5, 24.0,
+           15.0, 10.0, 10.5, 11.5, 16.5, 11.5, 11.5, 25.0,
+           14.0, 8.0, 27.5, 9.0, 18.5, 13.0, 20.0, 27.0, 26.0,
+           19.5, 24.5, 11.5, 23.0,  9.0, 9.0, 19.5, 13.5, 7.5,
+           24.0, 19.0, 11.0, 16.5, 19.5, 17.5, 21.0, 23.0,
+           27.0, 25.5, 27.5, 0.5, 13.0, 21.0, 7.0, 24.0, 6.5,
+           22.0, 27.5)
```

- Or, you can read the given data file `midterm_marks` by

```
> data<- read.csv("C:/Data/midterm_marks")
> mark<- data[,2]
> #data will have 2 columns: first is the index and
> #second column is the marks.
```

Descriptive statistics using R: Location

```
> length(mark)
[1] 98
> summary(mark)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.50   12.12   18.25   17.50   24.00   28.00
> mean(mark)
[1] 17.5
> median(mark)
[1] 18.25
> quantile(mark)
    0%    25%    50%    75%   100%
 0.500 12.125 18.250 24.000 28.000
```

Descriptive statistics using R: Variability

```
> range(mark)
[1]  0.5 28.0
> var(mark)
[1] 53.80412
> sd(mark)
[1] 7.335129
> IQR(mark)
[1] 11.875
> mark[order(mark)[1:5]] # The 5 smallest observations
[1] 0.5 1.0 1.0 1.0 2.5
> size<-length(mark)   #sample size = 98
> mark[order(mark)[(size-4):size]] #The 5 largest observations
[1] 27.5 27.5 27.5 28.0 28.0
```

Descriptive statistics: Skewness (1)

- If a distribution is unimodal (has just one peak), beside the location and variability, we would check whether it's symmetric or skewed to one side.
- If the bulk of the data is at the left and the right tail is longer, we say that the distribution is **skewed right or positively skewed**.
- If the peak is toward the right and the left tail is longer, we say that the distribution is **skewed left or negatively skewed**.
- Given a sample of size n, the sample skewness is

$$\frac{\sqrt{n(n-1)}}{n-2} \times \frac{m_3}{(m_2)^{3/2}},$$

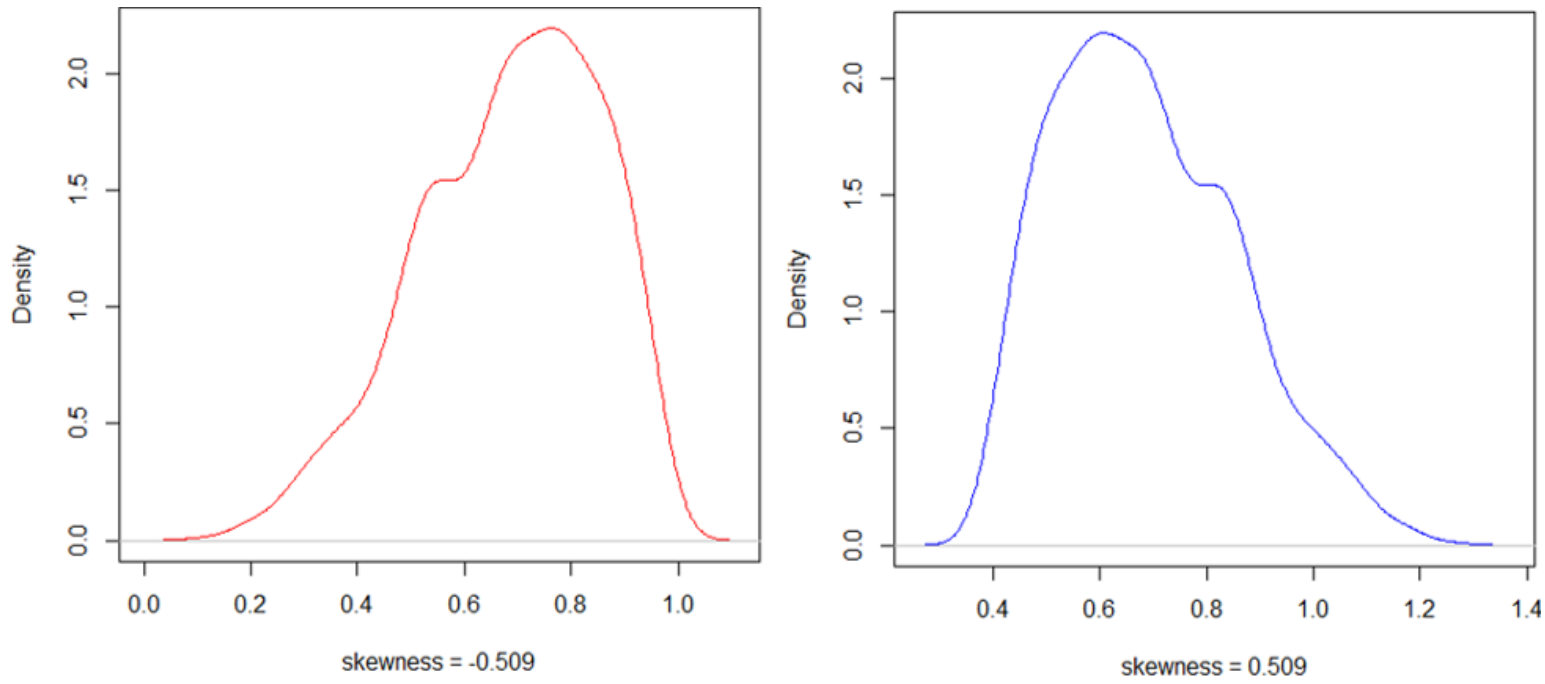
where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{and} \qquad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

- The skewness value represents the amount and direction of skew.

Descriptive statistics: Skewness (2)

- The two dataset below are having the same mean and same sd, however their shape are different.(x = Beta(4.5, 2) and y = mean(x) - x).



- If skewness = 0 then the data are perfectly symmetrical, but it's very rare!
- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -1/2 or between 1/2 and 1, the distribution is moderately skewed.
- If skewness is between -1/2 and 1/2, the distribution is approximately symmetric.

Descriptive statistics using R: Skewness

Create a function to calculate the sample skewness in R:

```
> skew <- function(x){
+ n <- length(x)
+ m3 <- mean((x-mean(x))^3)
+ m2 <- mean((x-mean(x))^2)
+ sk=m3/m2^(3/2)*sqrt(n*(n-1))/(n-2)
+ return(sk)
+ }
> skew(mark)
[1] -0.4205113
```

Descriptive statistics: Kurtosis (1)

- Beside skewness, **kurtosis** is another numerical measures of shape of a distribution.
- Higher values of kurtosis indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.
- Given a sample of size n, the sample kurtosis or actually excess kurtosis is:

$$\frac{n-1}{(n-2)(n-3)}\left[\frac{(n+1)m_4}{m_2^2} - 3(n-1)\right]$$

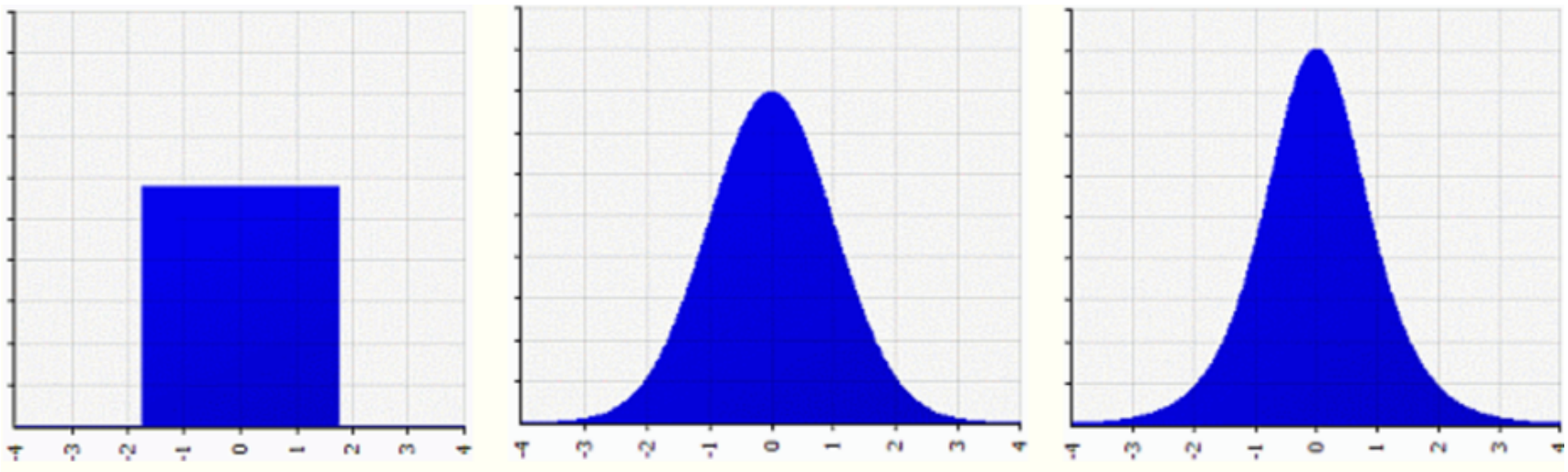
where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{and} \qquad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

- Kurtosis tells you how tall and sharp the central peak is, relative to a standard bell curve.

Descriptive statistics: Kurtosis (2)

- The three datasets below are having the same mean zero and same sd of 1 and same skewness of 0, however their kurtosis are different.



Descriptive statistics using R: Kurtosis

Create a function to calculate the sample kurtosis in R:

```
> kurt <- function(x){
+ n <- length(x)
+ m4 <- mean((x-mean(x))^4)
+ m2 <- mean((x-mean(x))^2)
+ kurt = (n-1)/((n-2)*(n-3))*((n+1)*m4/(m2^2)-3*(n-1))
+ return(kurt)
+ }
> kurt(mark)
[1] -0.6471977
```

Descriptive statistics using Python: Location

In [28]:
```
import numpy as np
import pandas as pd
from statistics import mean
from statistics import median
from statistics import variance
data = pd.read_csv (r"C:\Data\midterm_marks")
data.columns = ['Obs', 'mark'] # changing the columns' name

mean(data['mark'])
median(data['mark'])
variance(data['mark'])
np.quantile(data['mark'],0.25)
np.quantile(data['mark'],0.5)
np.quantile(data['mark'],0.75)
min(data['mark'])
max(data['mark'])
print('min: ',min(data['mark']) , "\n", 'Q1: ', np.quantile(data['mark'],0.25),
      "\n", 'mean: ',mean(data['mark']), "\n",  'Q2: ',median(data['mark']),
      "\n",'Q3: ',np.quantile(data['mark'],0.75),  "\n",'max: ', max(data['mark']),
     "\n", 'range: ',min(data['mark']), max(data['mark']) )
```

```
min:  0.5
 Q1:  12.125
 mean:  17.5
 Q2:  18.25
 Q3:  24.0
 max:  28.0
 range:  0.5 28.0
```

Descriptive statistics using Python: Variability

In [41]:
```
import statistics as st
from statistics import variance
variance(data['mark'])
st.stdev(data['mark']) #standard deviation
q75, q25 = np.percentile((data['mark']), [75 ,25])
iqr = q75 - q25

print('var: ', variance(data['mark']), '\n','sd: ', st.stdev(data['mark']), '\n','IQR: ', iqr)
```

```
var:  53.8041237113402
 sd:  7.335129427034005
 IQR:  11.875
```

Descriptive statistics using Python: Skewness

```
def skew(x):
    n = len(x)
    for i in range(n):
        y[i] = (x[i] - mean(x))**2
        z[i] = (x[i] - mean(x))**3
    m2 = mean(y)
    m3 = mean(z)
    sk = (m3/pow(m2,3/2))*pow(n*(n-1),1/2)/(n-2)
    return(sk)

print(skew(data['mark']))
```

```
-0.4205113167502236 3
```

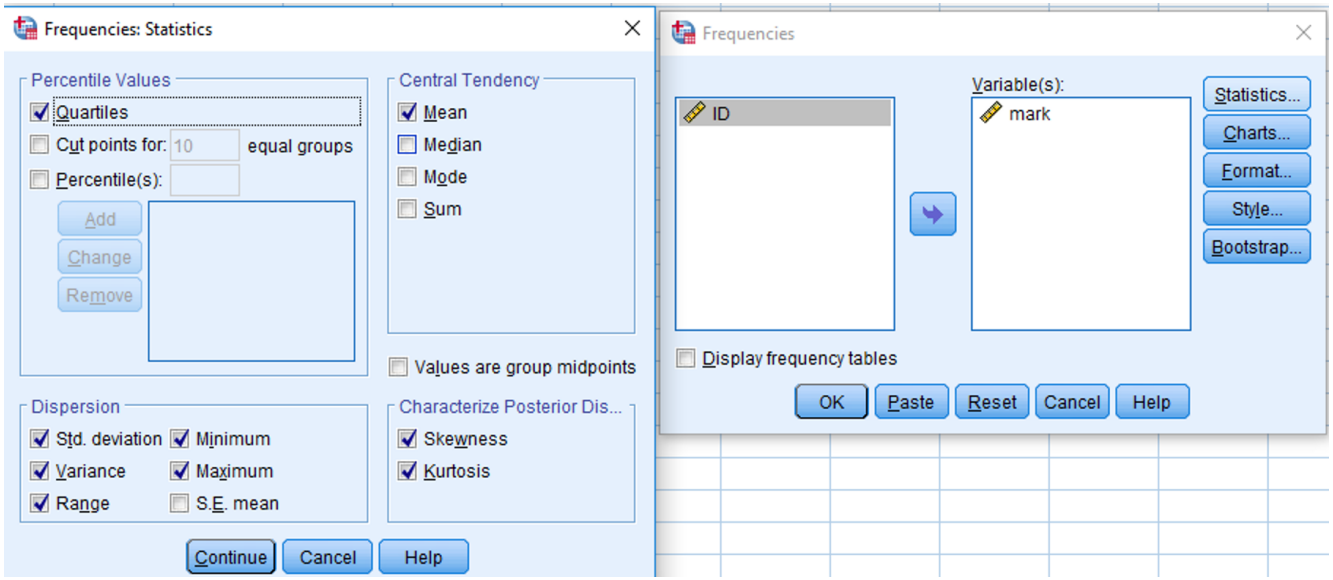
Descriptive statistics using Python: Kurtosis

```
def kurt(x):
    n = len(x)
    for i in range(n):
        y[i] = (x[i] - mean(x))**2
        z[i] = (x[i] - mean(x))**4
    m2 = mean(y)
    m4 = mean(z)
    kur =  (n-1)/((n-2)*(n-3))*((n+1)*m4/(m2**2) - 3*(n-1))
    return(kur)

print(kurt(data['mark']))
```

```
-0.6471976744377083
```

Descriptive statistics using SPSS (1)

- Open the data `midterm_marks_SPSS.sav`.
- "Analyze" $\rightarrow$ "Descriptive Statistics" $\rightarrow$ "Frequencies"...
- Move "mark" to the variable panel $\rightarrow$ click "Statistics" $\rightarrow$ choose statistics that you want (quantiles, mean,...) $\rightarrow$ "Continue" $\rightarrow$ "OK".


Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Characterize Posterior Dis...
Skewness
Kurtosis
Continue
Cancel
Help
Frequencies
ID
Variable(s):
mark
Statistics...
Charts...
Format...
Style...
Bootstrap...
Display frequency tables
OK
Paste
Reset
Cancel
Help

Descriptive statistics using SPSS (2)

The output is:

Statistics

mark

N	Valid	98
	Missing	0
Mean		17.500
Std. Deviation		7.3351
Variance		53.804
Skewness		-.421
Std. Error of Skewness		.244
Kurtosis		-.647
Std. Error of Kurtosis		.483
Range		27.5
Minimum		.5
Maximum		28.0
Percentiles	25	11.875
	50	18.250
	75	24.125

A Note on Numerical Summaries

- For a dataset, when mean is the same (or approximately the same) as median, then the data are close to symmetric.
- Mean is sensitive to the outlier while median is not.
- When mean is much larger than median, data are right (positively) skewed; while when mean is much smaller than median then data are left (negatively) skewed.

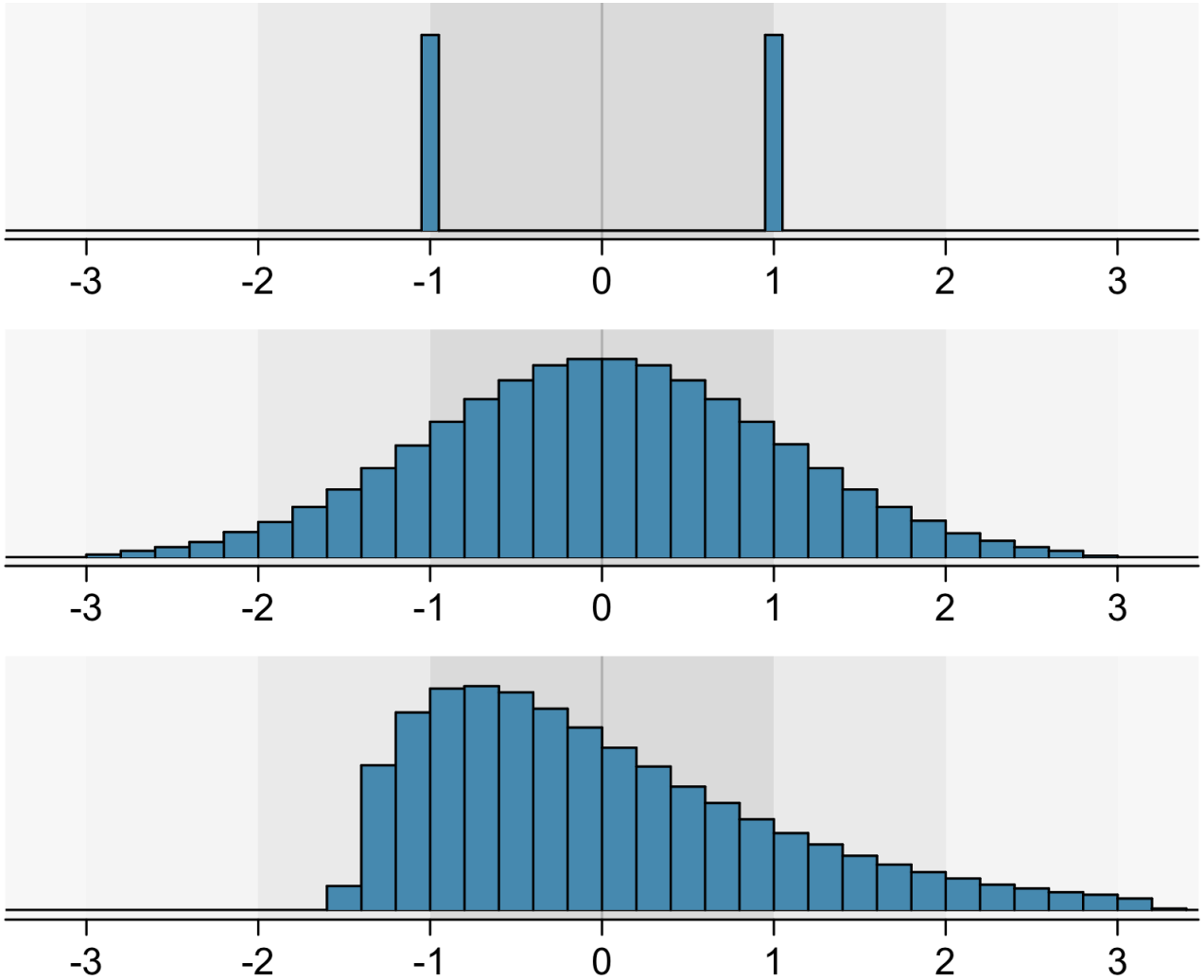
1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

Numerical Summaries Are Not Enough

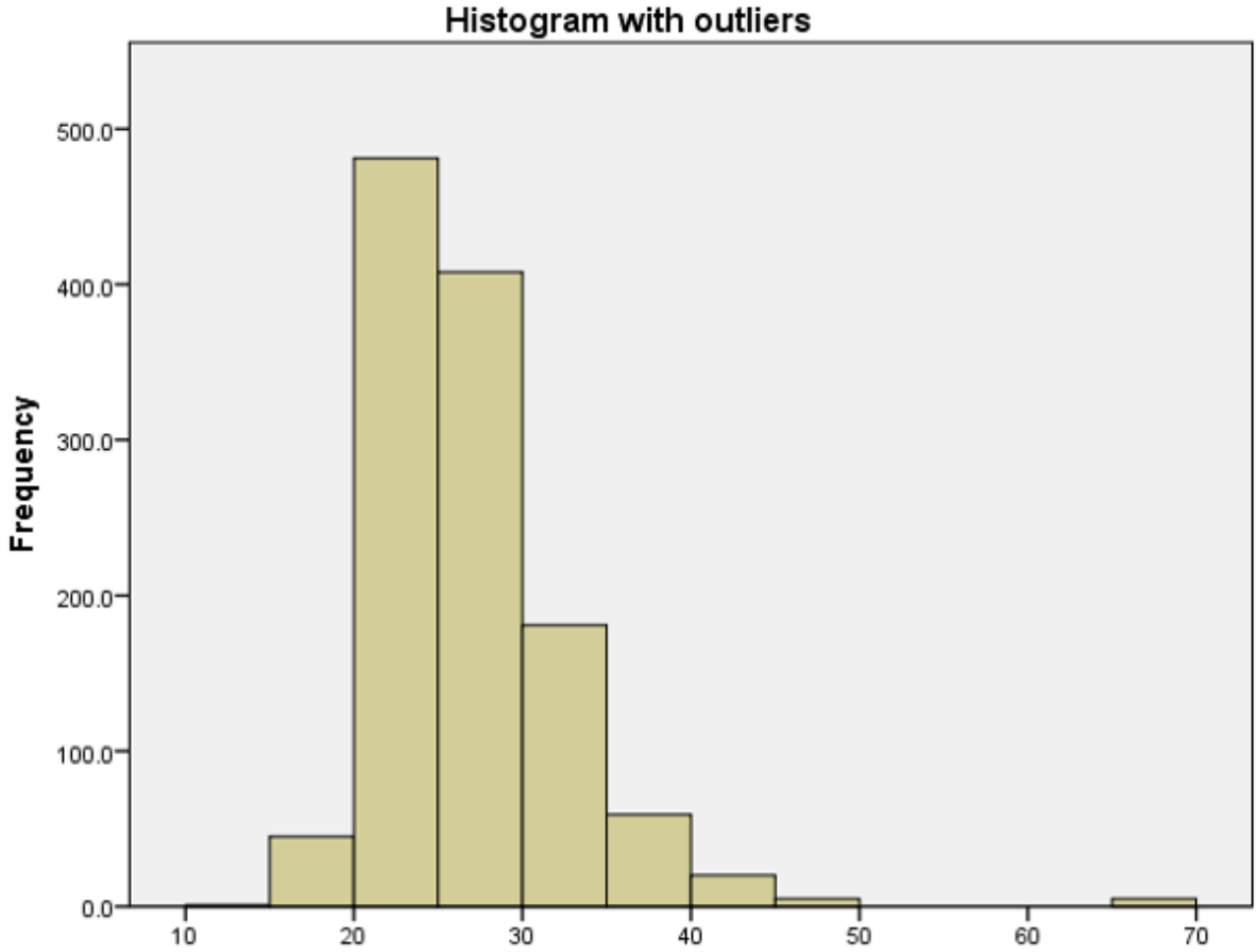
- No matter how many of the summary measures we report, nothing beats a picture.
- All 3 samples below had a sample mean of 0 and a sample variance of 1.



Histogram and Density Plot

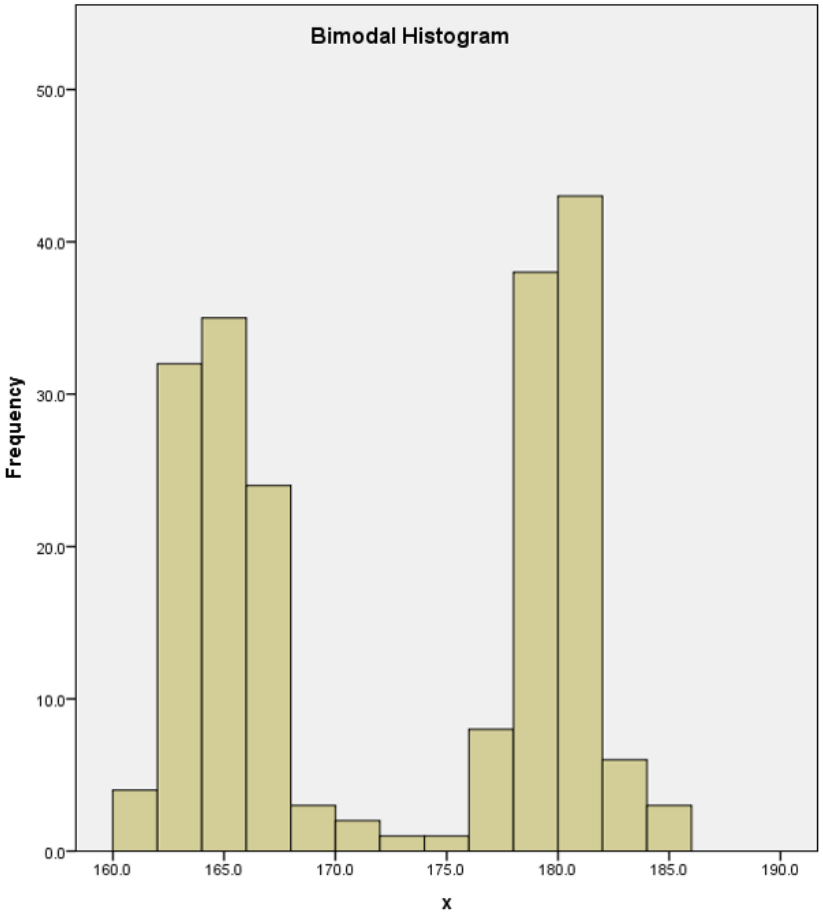
- A histogram is a graph that uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable.
- Density plots can be thought of as plots of smoothed histograms.
- What do we look for in a histogram?
  - The overall pattern. Do the data cluster together, or is there a gap such that one or more observations deviate from the rest?
  - Do the data have a single mound? This is known as a unimodal distribution. Data with two are known as bimodal, and data with many mounds are referred to as multimodal.
  - Is the distribution symmetric or skewed? Any suspected outliers?

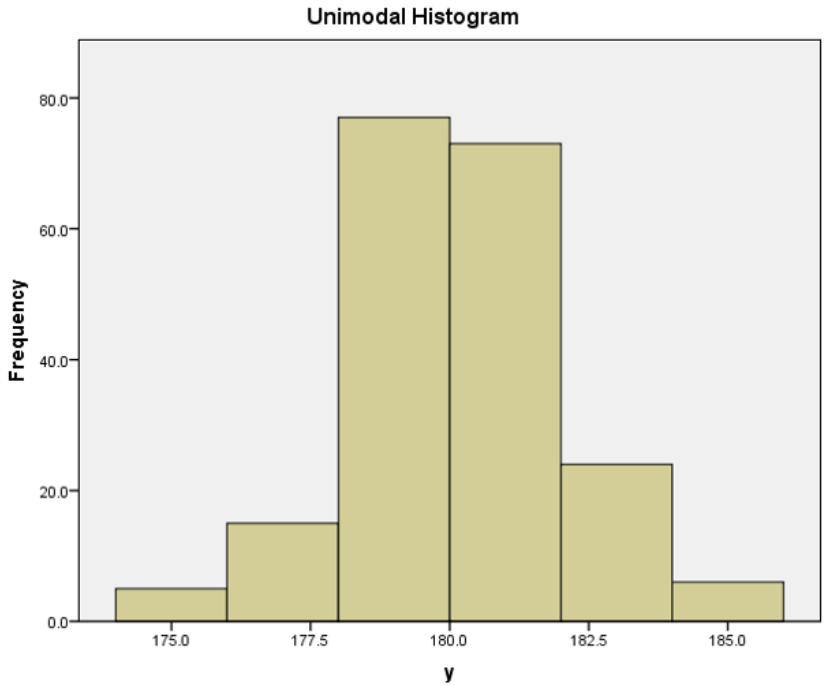
A Histogram With Outliers



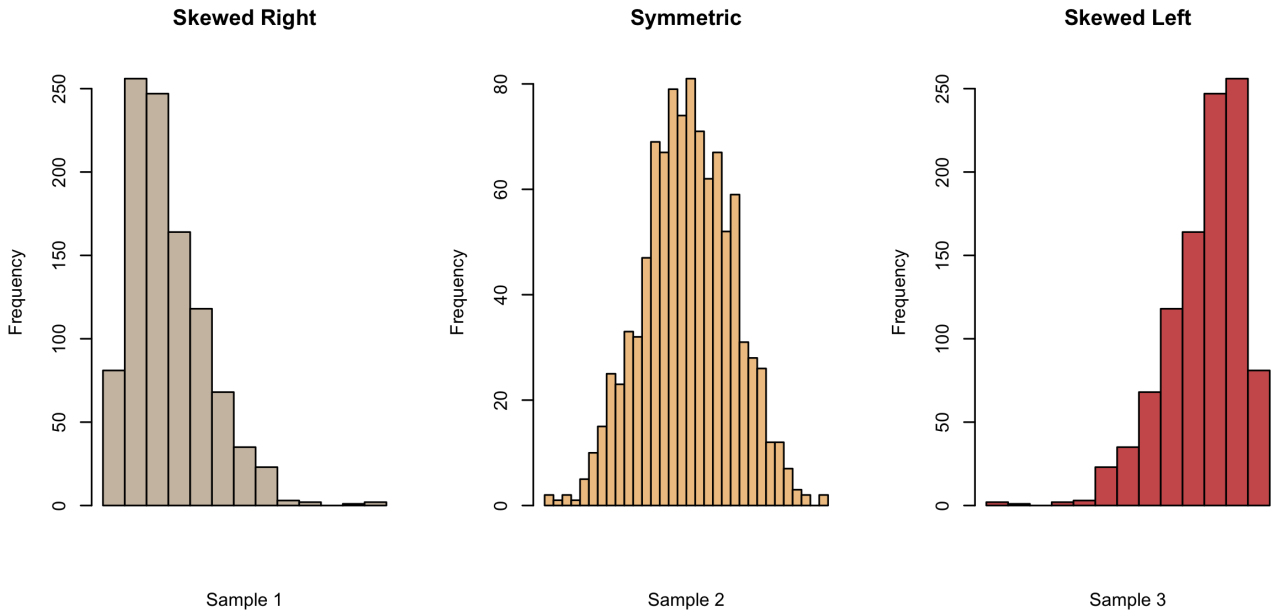
- This histogram is unimodal, but by visualization it has suspected outliers in the right.

Unimodal and Bimodal Histograms





Skewed Histograms



- Income is typically right-skewed.
- IQ is typically symmetric.
- Life-span is typically left-skewed.

Boxplots

- Boxplots provide a skeletal representation of a distribution, and they are very well suited for showing distributions for multiple variables.

Boxplots

- Boxplots provide a skeletal representation of a distribution, and they are very well suited for showing distributions for multiple variables.

- How the boxplot is drawn?
  - Determine Q_1, Q_2 and Q_3. The box is made of Q1 and Q3.
  - Determine the max-whisker reach by: $Q_3 + 1.5IQR$; the min-whisker reach by $Q_1 - 1.5IQR$.
  - Any data point that is out of the range from the min to max whisker reach is defined to be outlier.
  - Except the outliers, the maximum point determines the upper whisker and the minimum points determines the lower whisker of a boxplot.

Boxplots

- Boxplots provide a skeletal representation of a distribution, and they are very well suited for showing distributions for multiple variables.

- How the boxplot is drawn?
  - Determine Q_1, Q_2 and Q_3. The box is made of Q1 and Q3.
  - Determine the max-whisker reach by: $Q_3 + 1.5IQR$; the min-whisker reach by $Q_1 - 1.5IQR$.
  - Any data point that is out of the range from the min to max whisker reach is defined to be outlier.
  - Except the outliers, the maximum point determines the upper whisker and the minimum points determines the lower whisker of a boxplot.

- A boxplot might have mild outlier and extreme outlier. An outlier is defined to be extreme outlier if it's larger than $Q_3 + 3IQR$ or smaller than $Q_1 - 3IQR$.

Boxplots

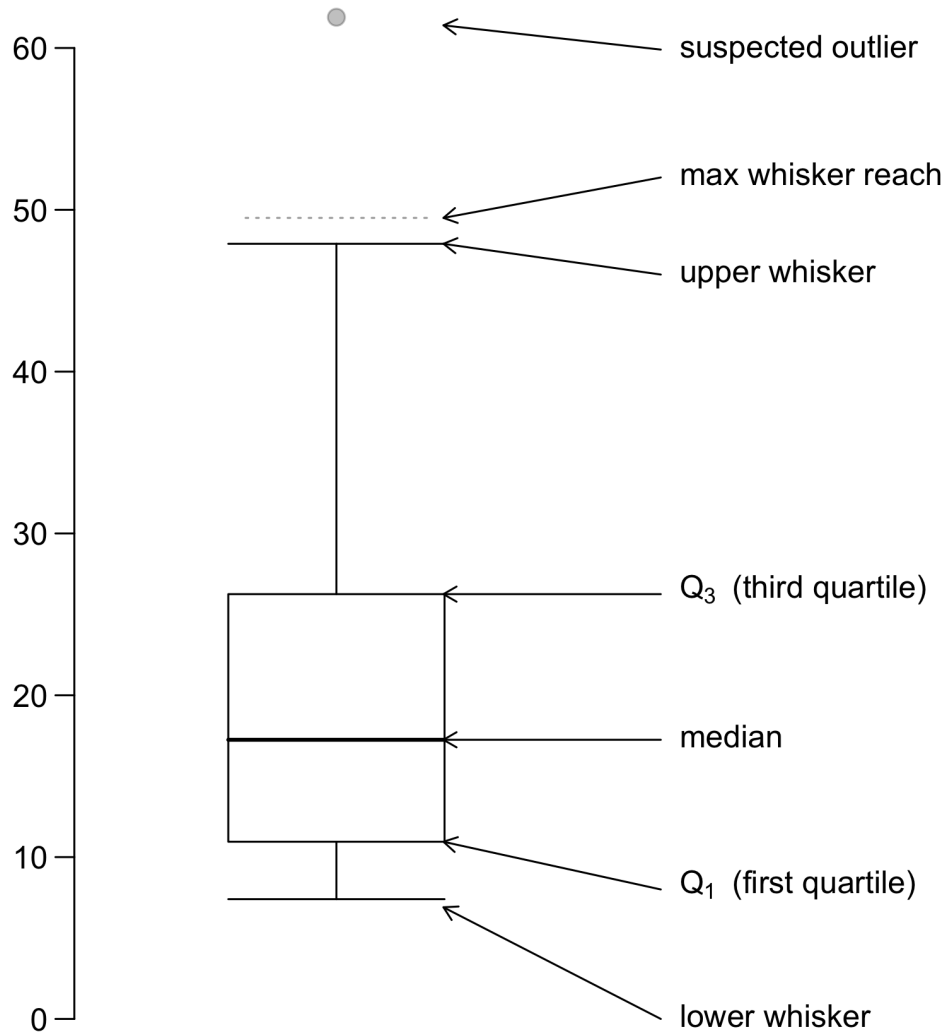
- Boxplots provide a skeletal representation of a distribution, and they are very well suited for showing distributions for multiple variables.

- How the boxplot is drawn?
  - Determine Q_1, Q_2 and Q_3. The box is made of Q1 and Q3.
  - Determine the max-whisker reach by: $Q_3 + 1.5IQR$; the min-whisker reach by $Q_1 - 1.5IQR$.
  - Any data point that is out of the range from the min to max whisker reach is defined to be outlier.
  - Except the outliers, the maximum point determines the upper whisker and the minimum points determines the lower whisker of a boxplot.

- A boxplot might have mild outlier and extreme outlier. An outlier is defined to be extreme outlier if it's larger than $Q_3 + 3IQR$ or smaller than $Q_1 - 3IQR$.

- A given boxplot helps us to identify median, lower and upper quantiles and outlier(s).

Ingredients of a Boxplot



QQ Plots

- The purpose of plotting a QQ plot of a dataset is to see if the data follow (approximately) a normal distribution or not.

QQ Plots

- The purpose of plotting a QQ plot of a dataset is to see if the data follow (approximately) a normal distribution or not.

- A QQ-plot plots the standardized sample quantiles against the theoretical quantiles of a N(0; 1) distribution. If they fall on a straight line, then we would say that there is evidence that the data came from a normal distribution.

QQ Plots

- The purpose of plotting a QQ plot of a dataset is to see if the data follow (approximately) a normal distribution or not.

- A QQ-plot plots the standardized sample quantiles against the theoretical quantiles of a N(0; 1) distribution. If they fall on a straight line, then we would say that there is evidence that the data came from a normal distribution.

- From the points on the plot, we can usually tell whether our sample data has longer or shorter tail than the normal, and on which side of the mean this occurs.

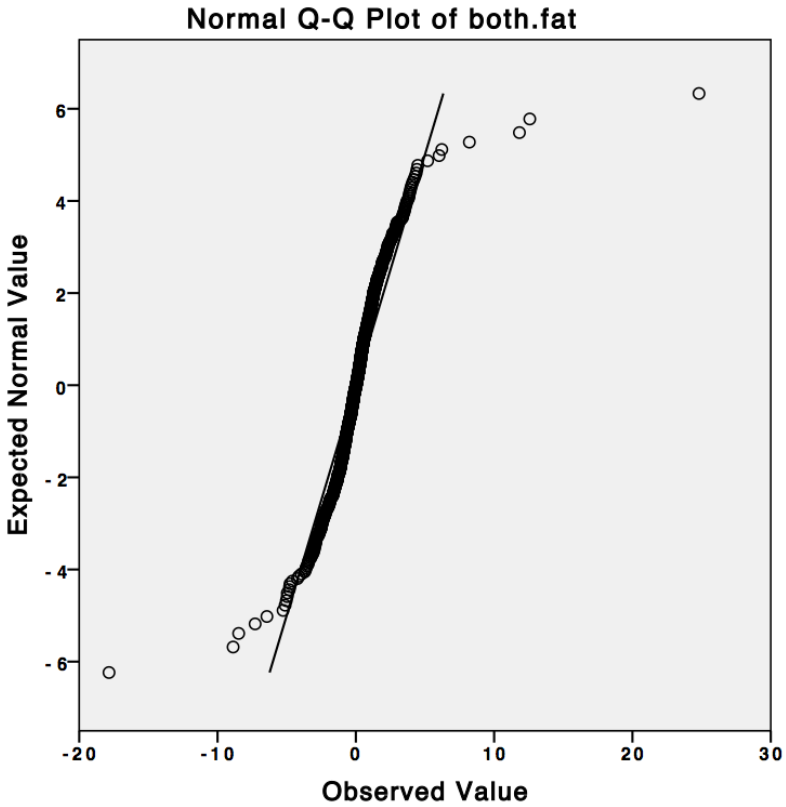
QQ plots and Normality Checking

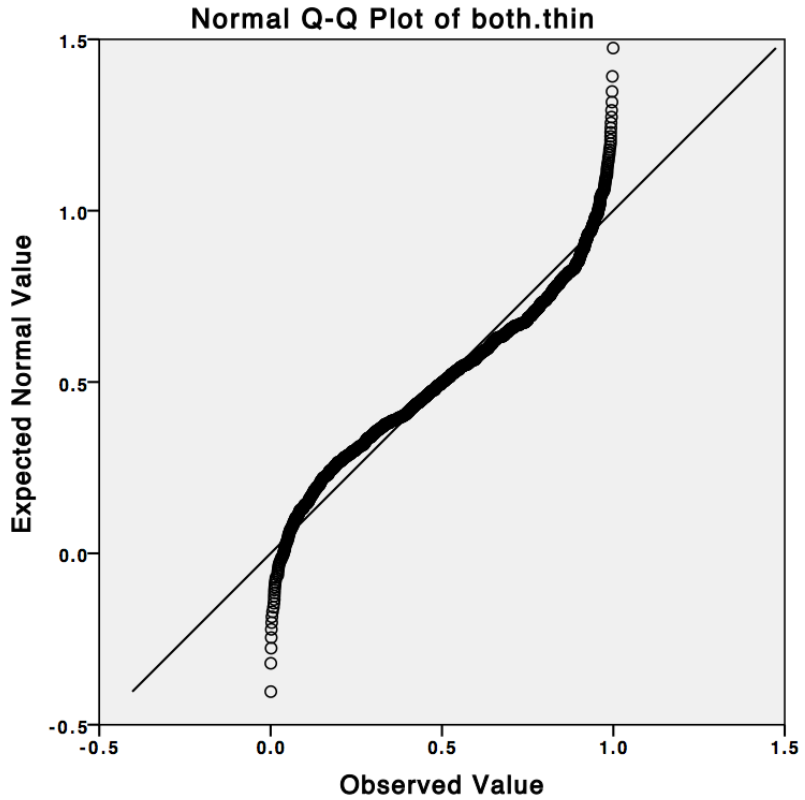
The comments below are for the QQ plot where the **sample quantiles are in the X-axis and theoretical quantiles are in Y-axis**.

- Right tail is below the straight line: longer tan Normal.
- Right tail is above the straight line: shorter than Normal.
- Left tail is below the straight line: shorter than Normal.
- Left tail is above the straight line: longer than Normal.

The comment should change accordingly if the sample quantiles are in the Y-axis and theoretical quantiles are in X-axis.
(For example, if right tail is below the straight line then it is shorter than Normal.)

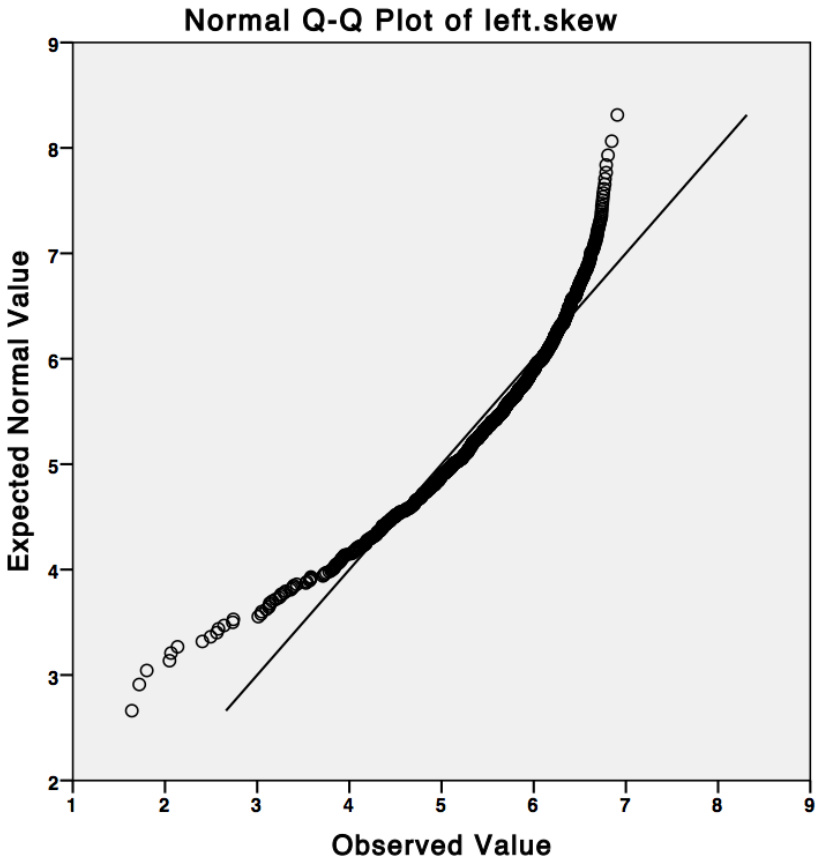
Boxplots: Few Examples (1)

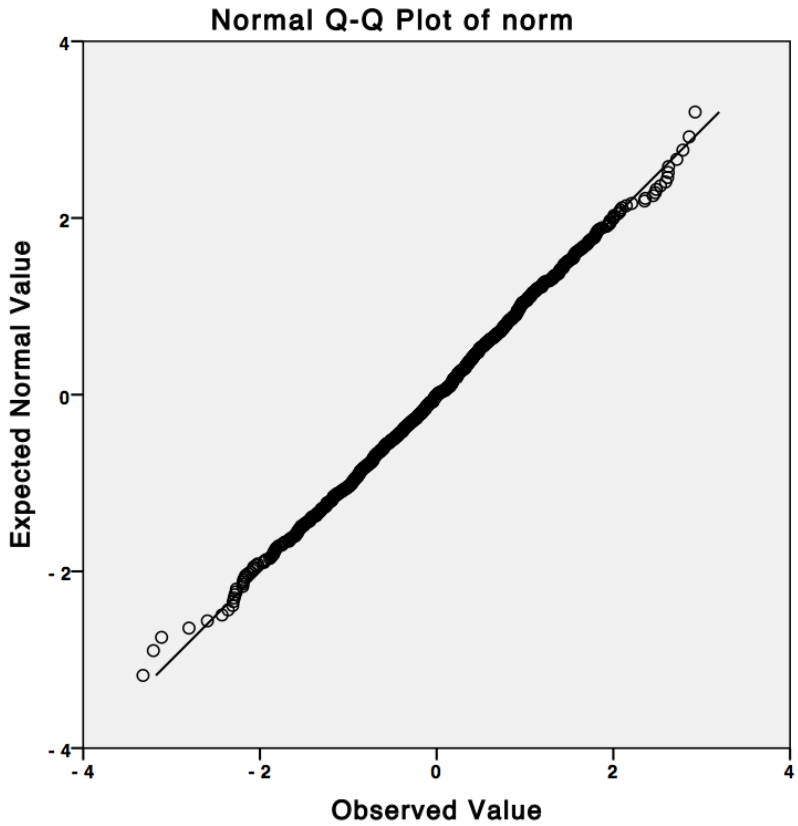




- Figure in the left is a data with both longer tails than normal.
- Figure in the right is a data with both shorter tails than normal.

Boxplots: Few Examples (2)





- Figure in the left is a data with left tail longer than normal but right tail is shorter than normal.
- Figure in the right is a data with both tails are normal.

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

Histogram and Density Plot in R

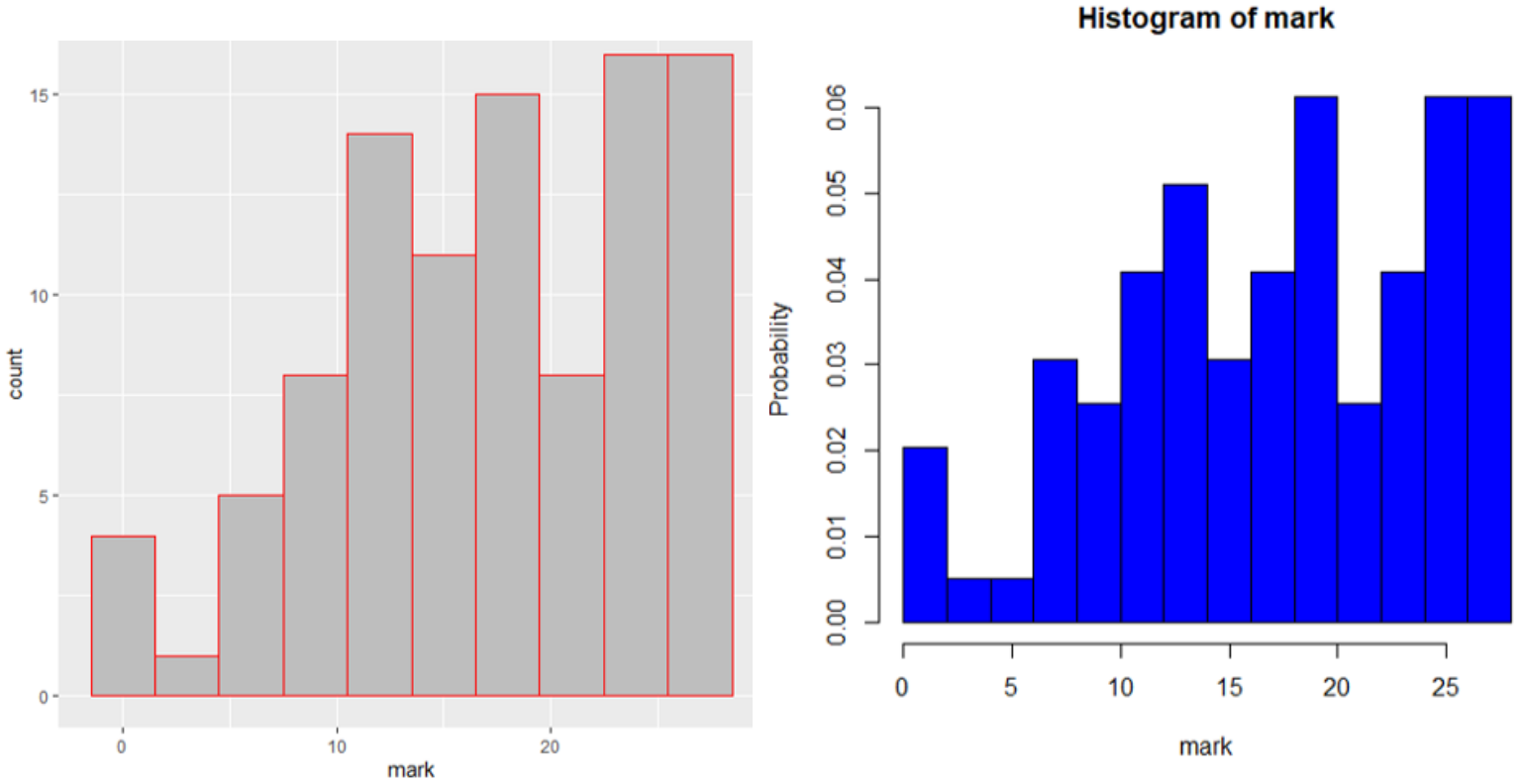
There are many ways to plot histograms in R:

- The `hist` function in the base `graphics` package;
- `truehist` in package `MASS`;
- `histogram` in package `lattice`;
- `geom_histogram` in package `ggplot2`.

```
## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = "lightgray", border = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, ...)
```

```
> hist(mark, freq=TRUE, main = paste("Histogram of mark"),
+ xlab = "mark", ylab="frequency", axes = TRUE, col = "blue")
```

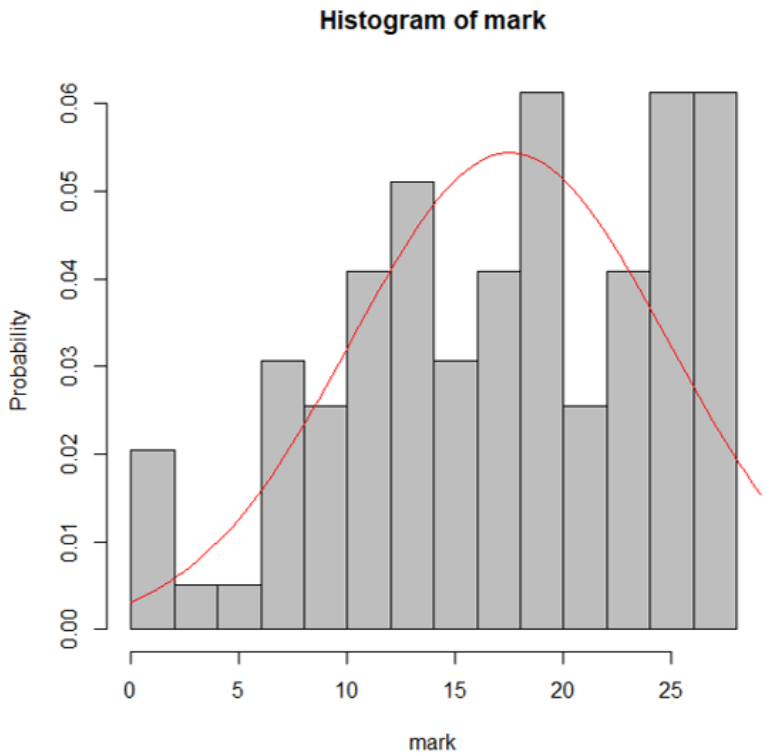
Histogram and Density Plot: in R (2)



- The histogram in the right is produced by `hist` with "nclass= 10"; while the histogram in the left is produced using `ggplot2` package.

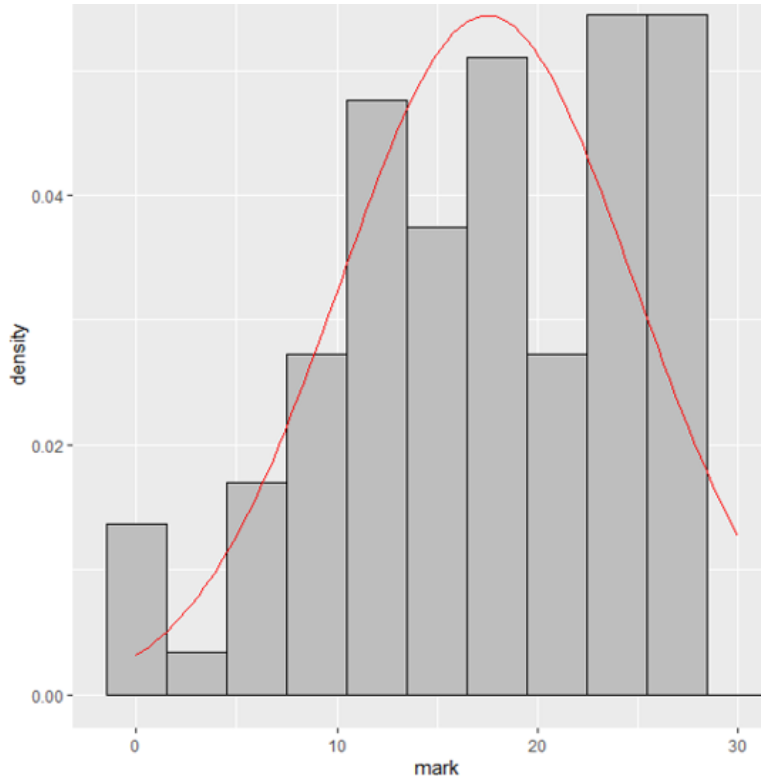
```
> library(ggplot2)
> data = data.frame(mark)
> ggplot(data) + geom_histogram(aes(x = mark),data = data,
+                  binwidth = 3, fill = "grey", color = "red")
```

Histogram and Density Plot in R (3)




```
> hist(mark, freq=FALSE, main = paste("Histogram of mark"),
+      xlab = "mark", ylab="frequency", axes = TRUE,
+      col = "grey",nclass = 10)
> x <- seq(0, 30, length.out=98)
> y <- dnorm(x, mean(mark), sd(mark))
> lines(x, y, col = "red")
```

Histogram and Normal Density Curve in R (4)




```
> p <- ggplot(data) + geom_histogram(aes(x = mark, y = ..density..),
+                    binwidth = 3, fill = "grey", color = "black")
> x <- seq(0, 30, length.out=98)
> y = dnorm(x, mean(mark), sd(mark))
> df <- data.frame(x = x, y = y)
> p + geom_line(data = data, aes(x = x, y = y), color = "red")
```

Boxplots in R

The code should be

```
> boxplot(mark, xlab = "mark")
```

Or

```
> ggplot(data) + geom_boxplot(aes(y = mark))
```

QQ Plots in R

The code should be

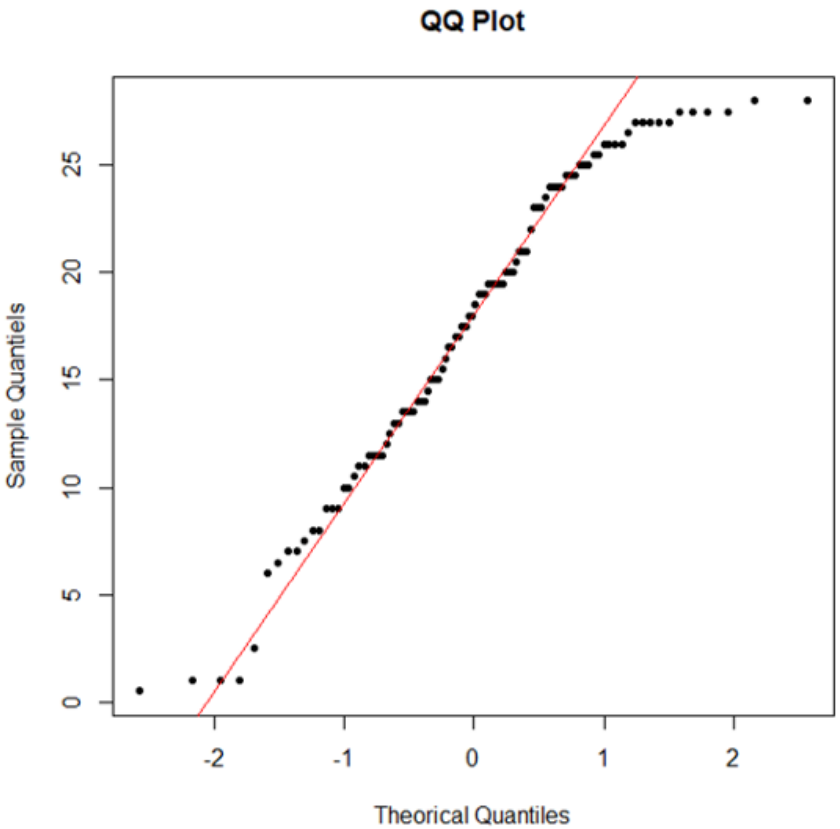
```
> qqnorm(mark, ylab = "Sample Quantiels", xlab = "Theorical
+         Quantiles", main = "QQ Plot", pch = 20)
> qqline(mark, col = "red")
```

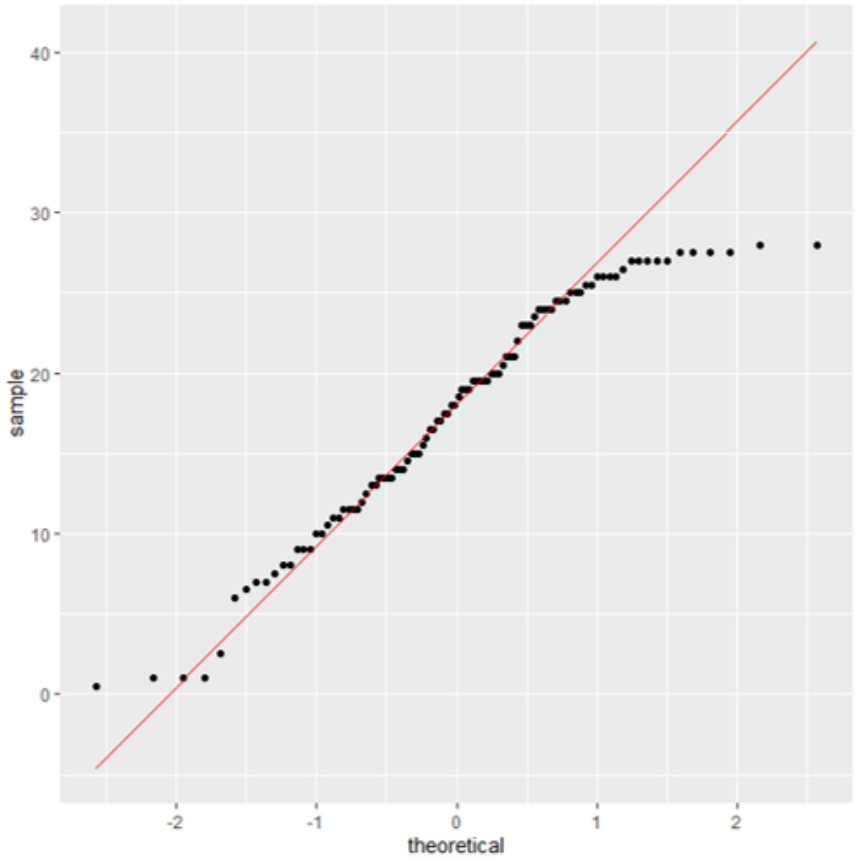
Or

```
> ggplot(data) + geom_qq(aes(sample = mark)) +
+   geom_qq_line(aes(sample = mark),color = "red")
```

QQ Plots in R

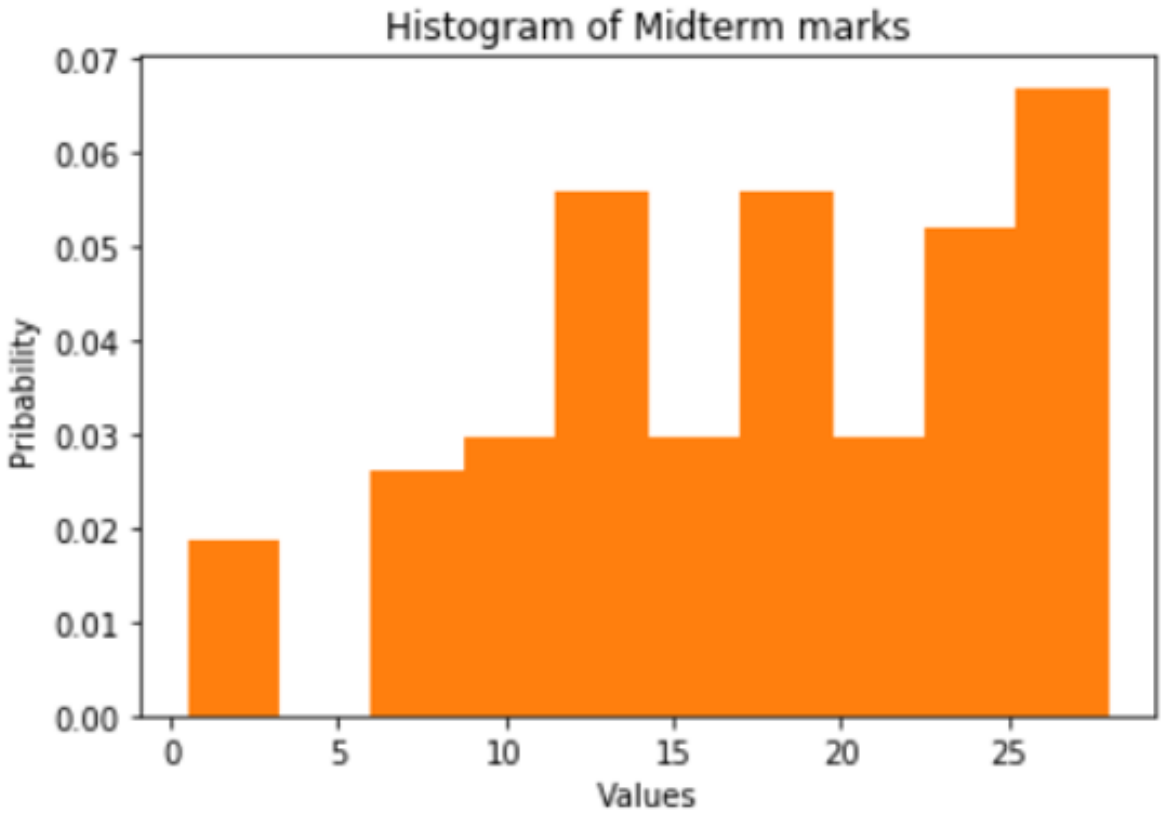
The output in the left is by `qqnorm` while the right one is by `ggplot2` package.





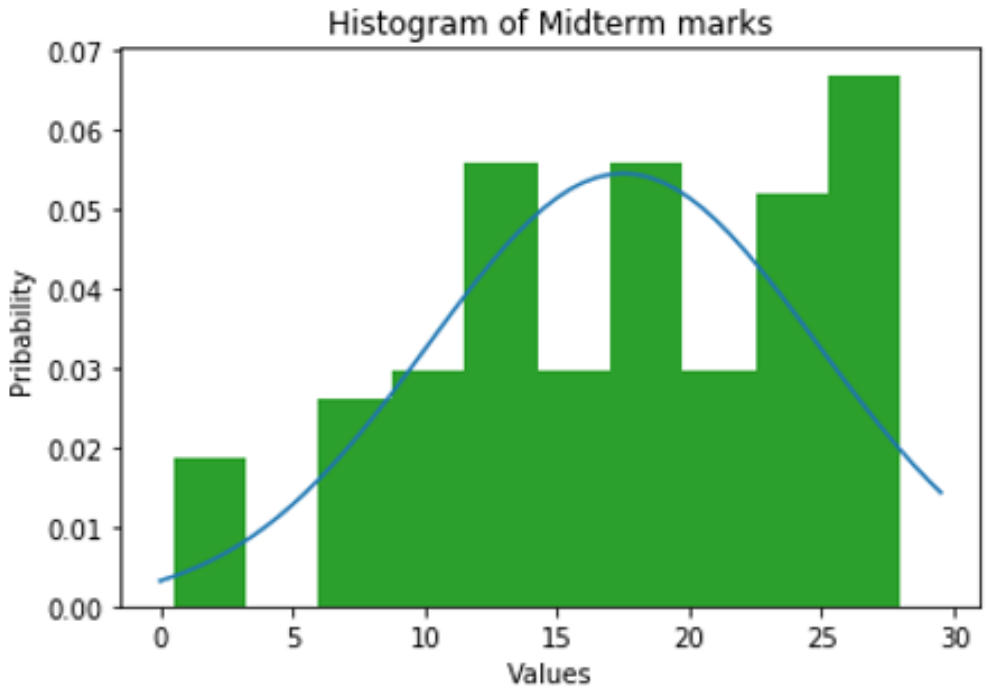
Histogram Plot in Python

```
In [23]: import matplotlib.pyplot as plt
         plt.hist(data['mark'], bins=None, range=None, density=True, color='C1')
         plt.title('Histogram of Midterm marks')
         plt.xlabel('Values')
         plt.ylabel('Pribability')
         plt.show()
```



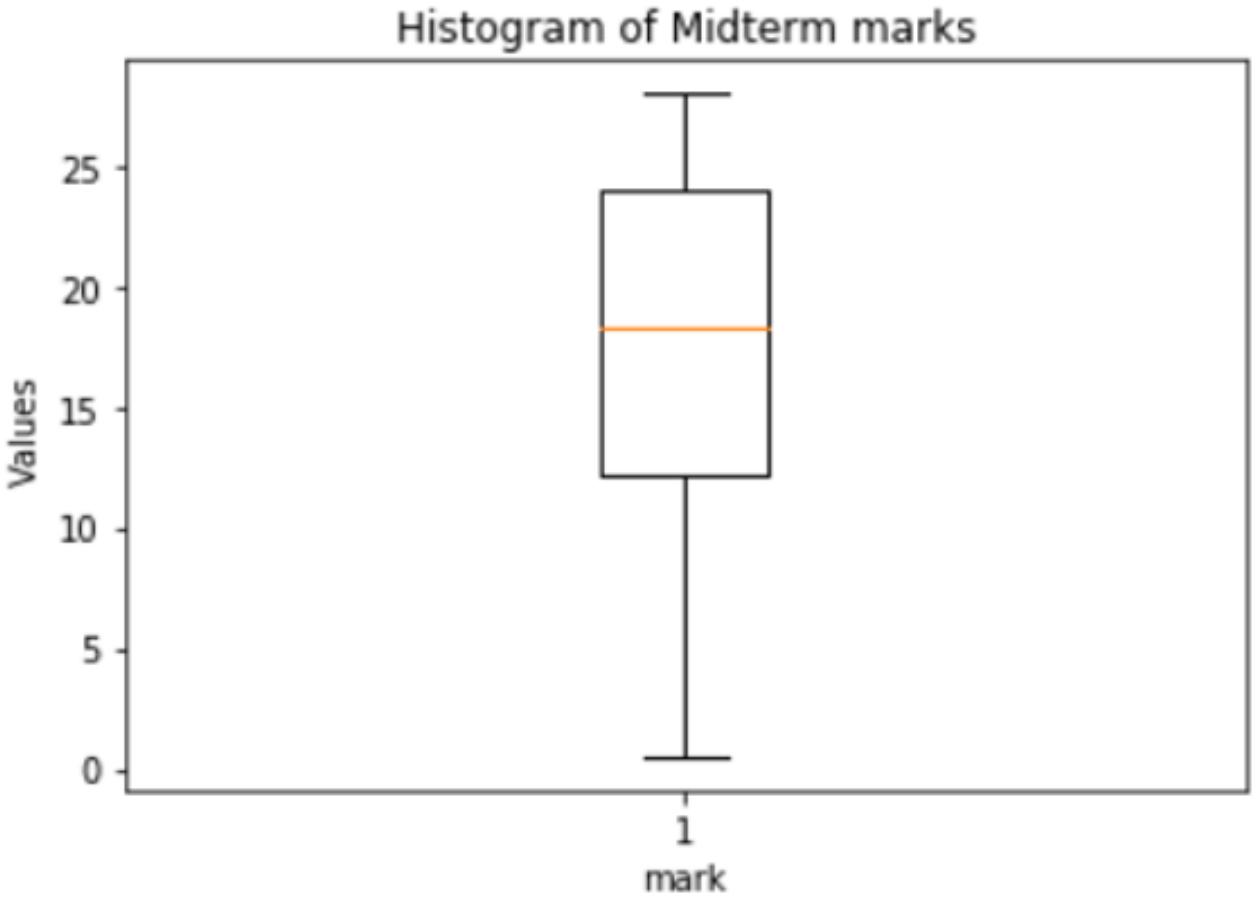
Histogram and Density Plots in Python

In [44]:
```
import scipy.stats as scst
import matplotlib.pyplot as plt
l = list(np.arange(0,30,0.5))
y = scst.norm.pdf(l,loc = mean(x),scale = st.stdev(x)) # this equivalent to qnorm in R
#print(y)
plt.plot(l, y)
plt.hist(data['mark'], bins=None, range=None, density=True, color='C2')
plt.title('Histogram of Midterm marks')
plt.xlabel('Values')
plt.ylabel('Pribability')
plt.show()
```



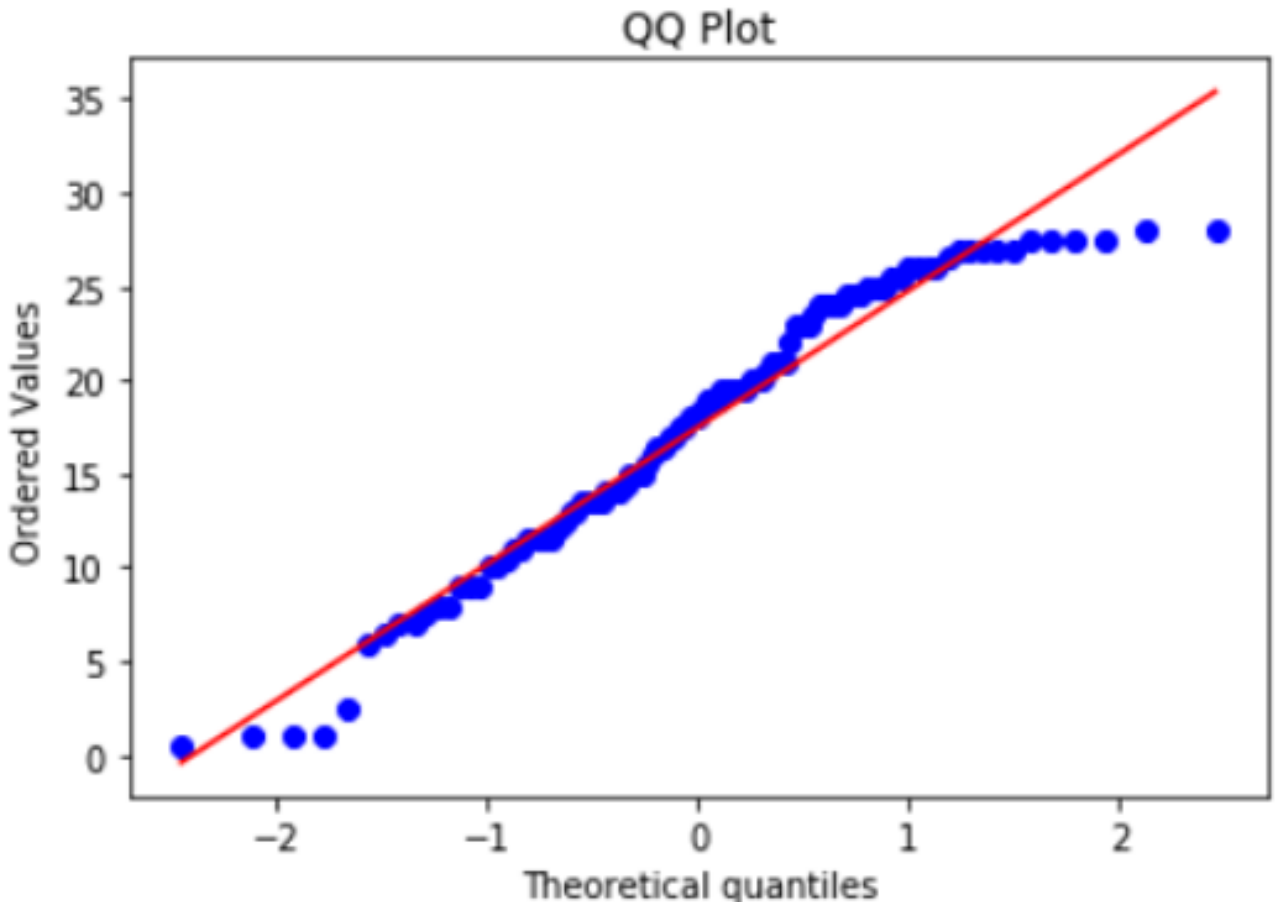
Boxplots in Python

```
import matplotlib.pyplot as plt
plt.boxplot(data['mark'])
plt.title('Histogram of Midterm marks')
plt.xlabel('mark')
plt.ylabel('Values')
plt.show()
```

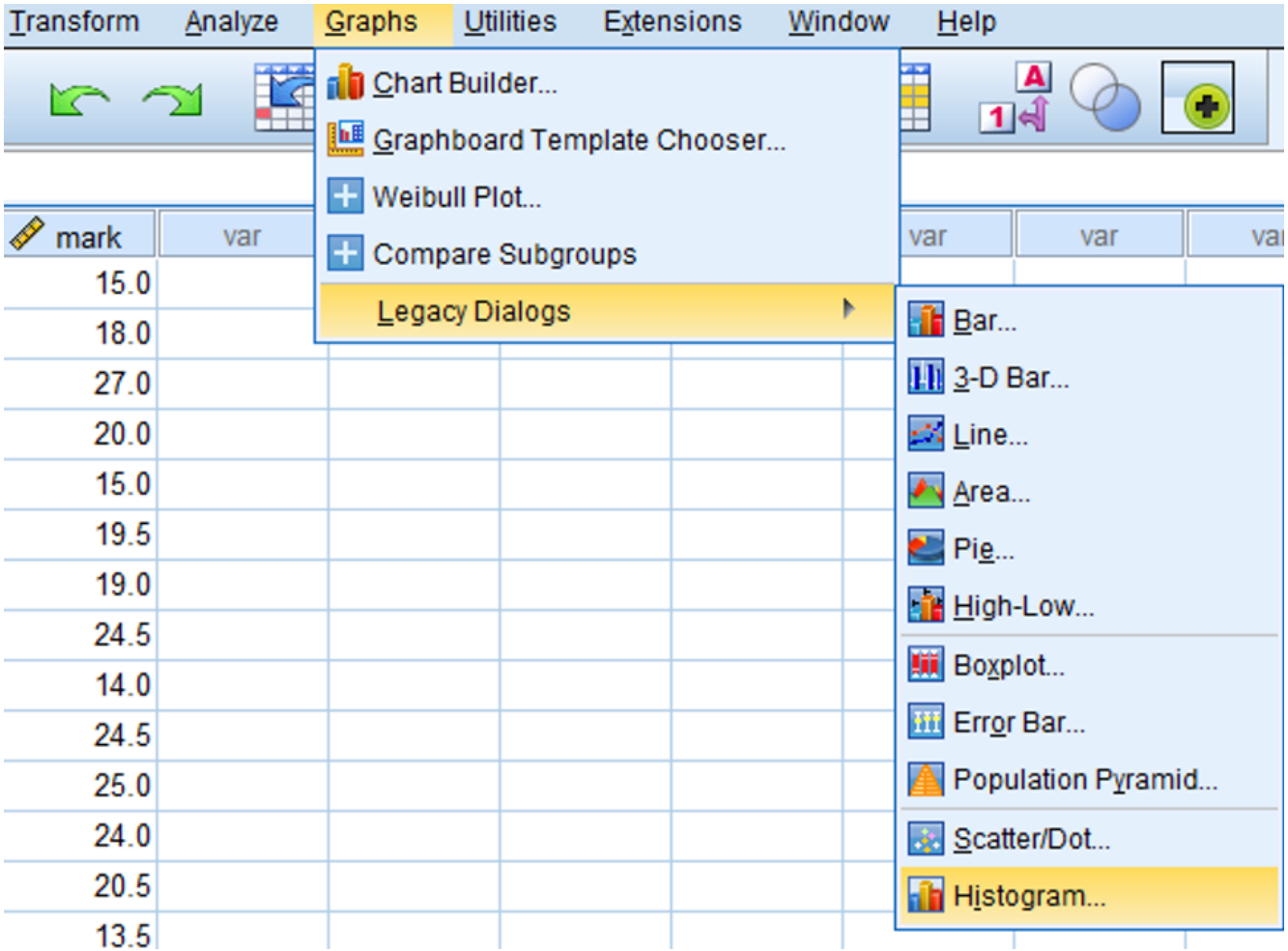


QQ Plots in Python

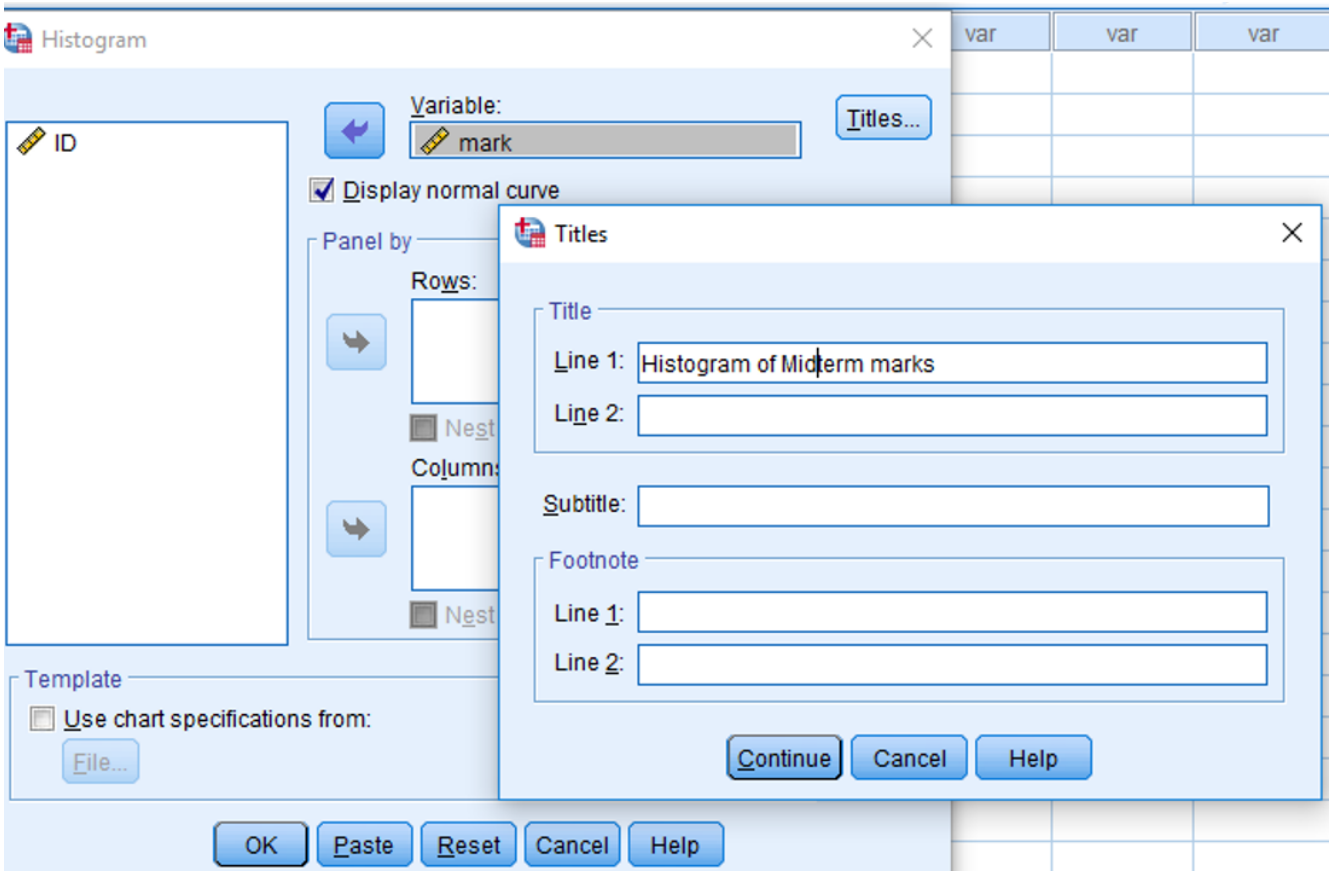
```
import pylab
import scipy.stats as scst

scst.probplot(x, dist="norm", plot=pylab)
pylab.title('QQ Plot')
pylab.show()
```



Plots in SPSS: Histograms (1)

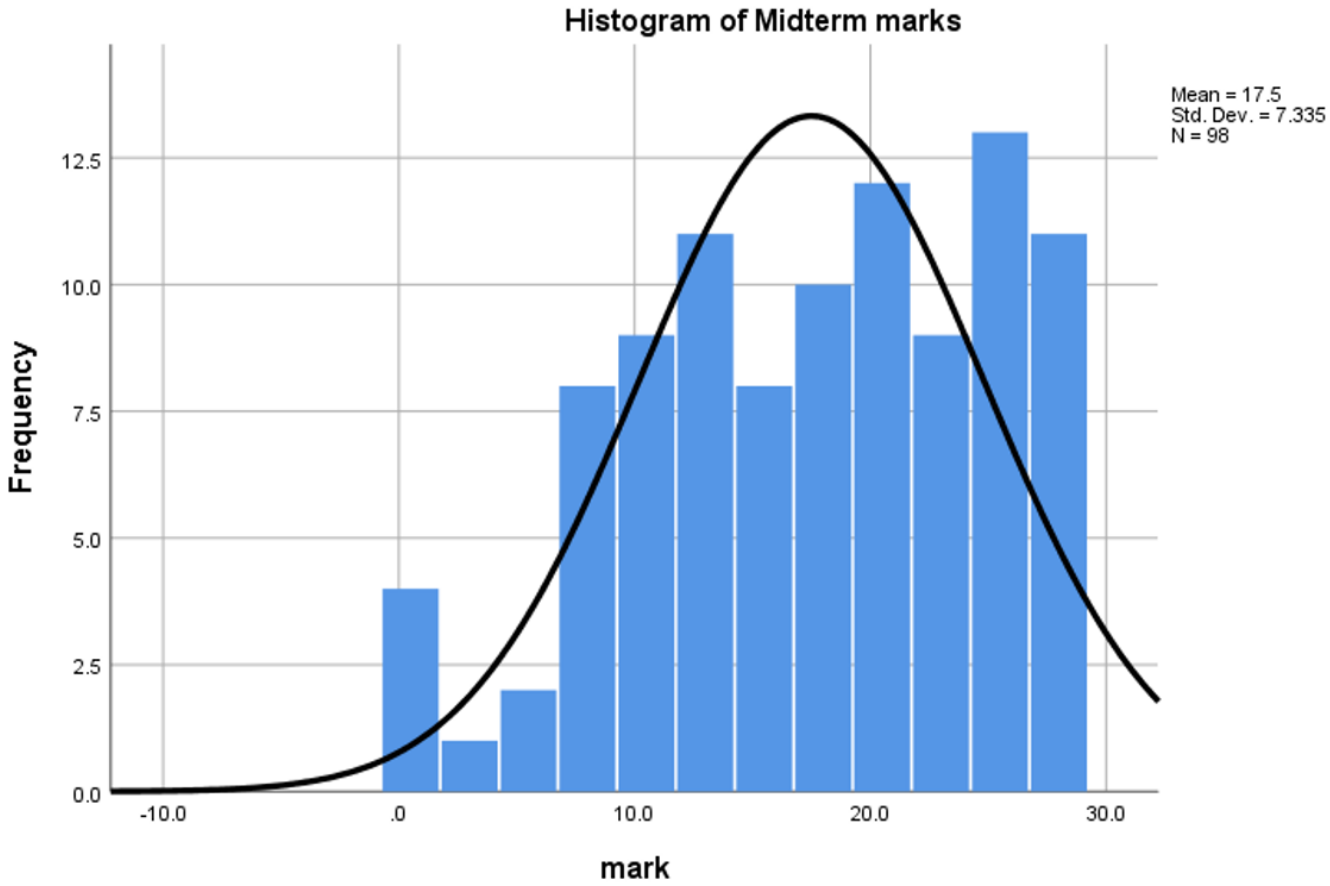


Plots in SPSS: Histograms (2)



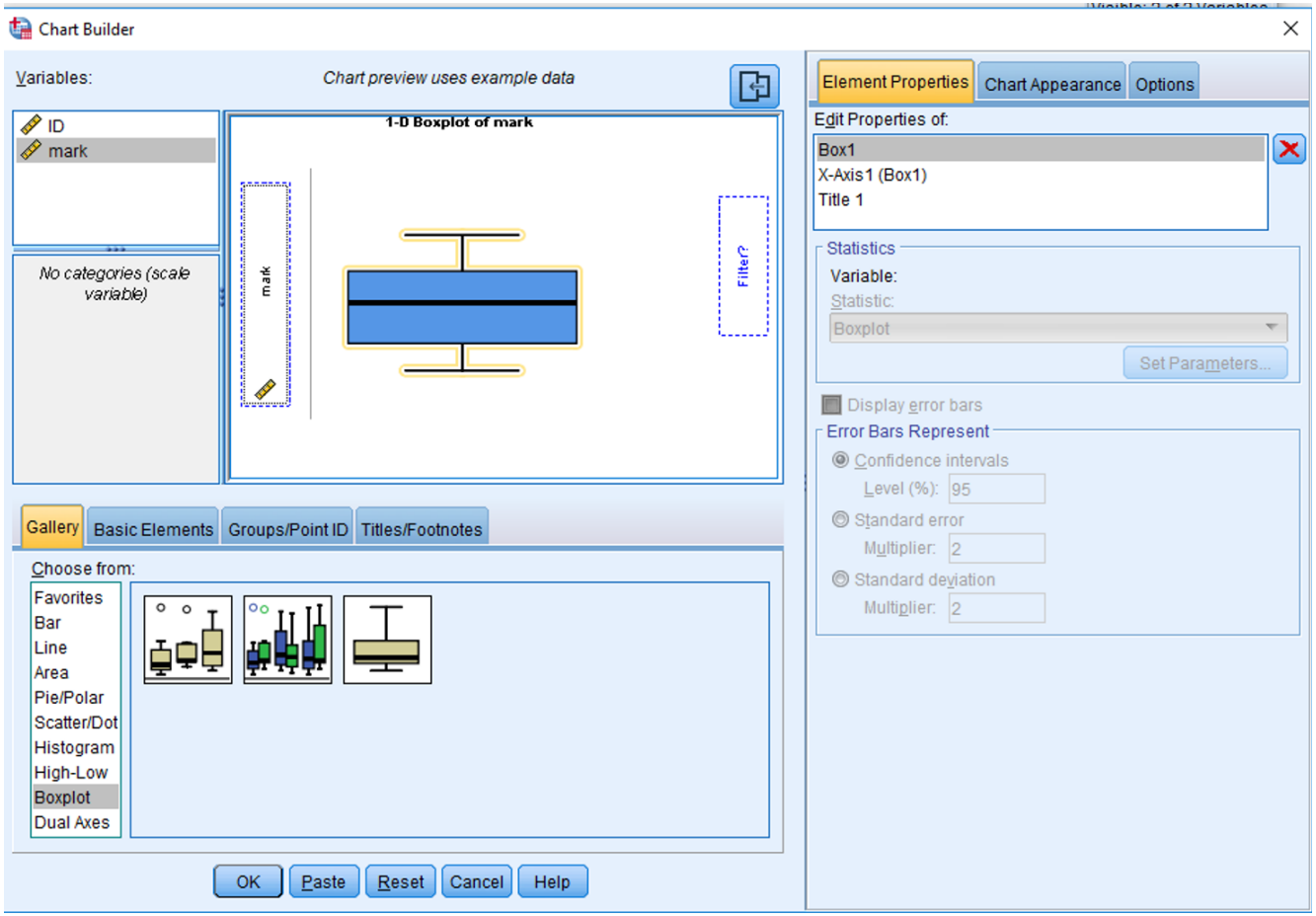
Plots in SPSS: Histograms (3)

The Output is:



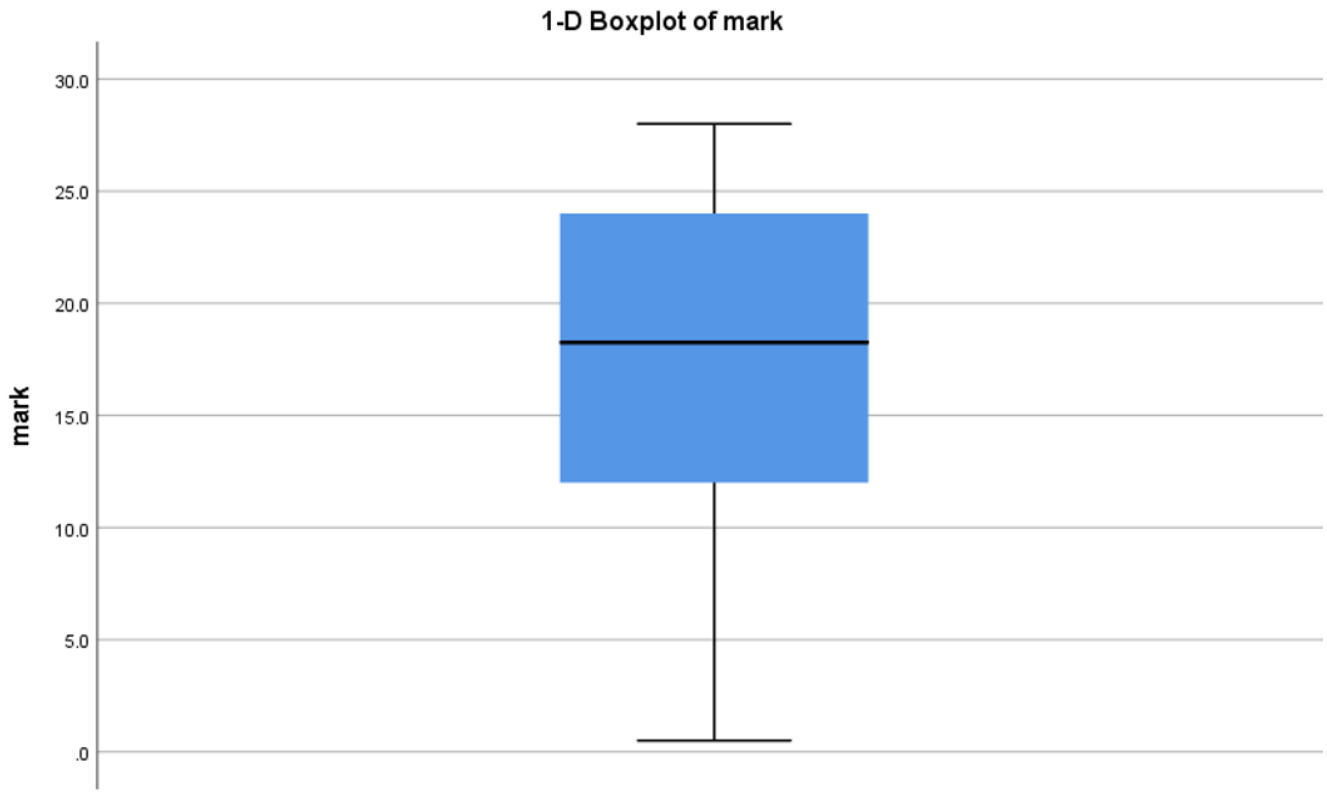
Plots in SPSS: Boxplots (1)

Under "Graphs" → "Chart Builder" → "Boxplot" (under "Gallery") → drag the type of boxplot into the "preview"...



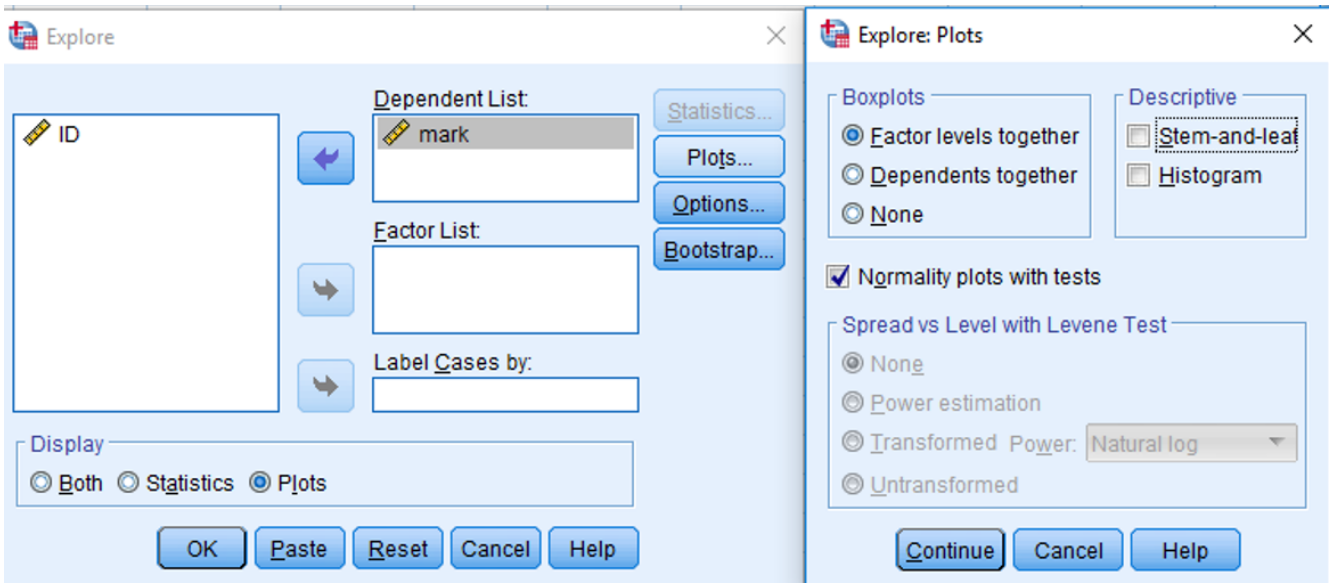
Plots in SPSS: Boxplots (2)

The output is:



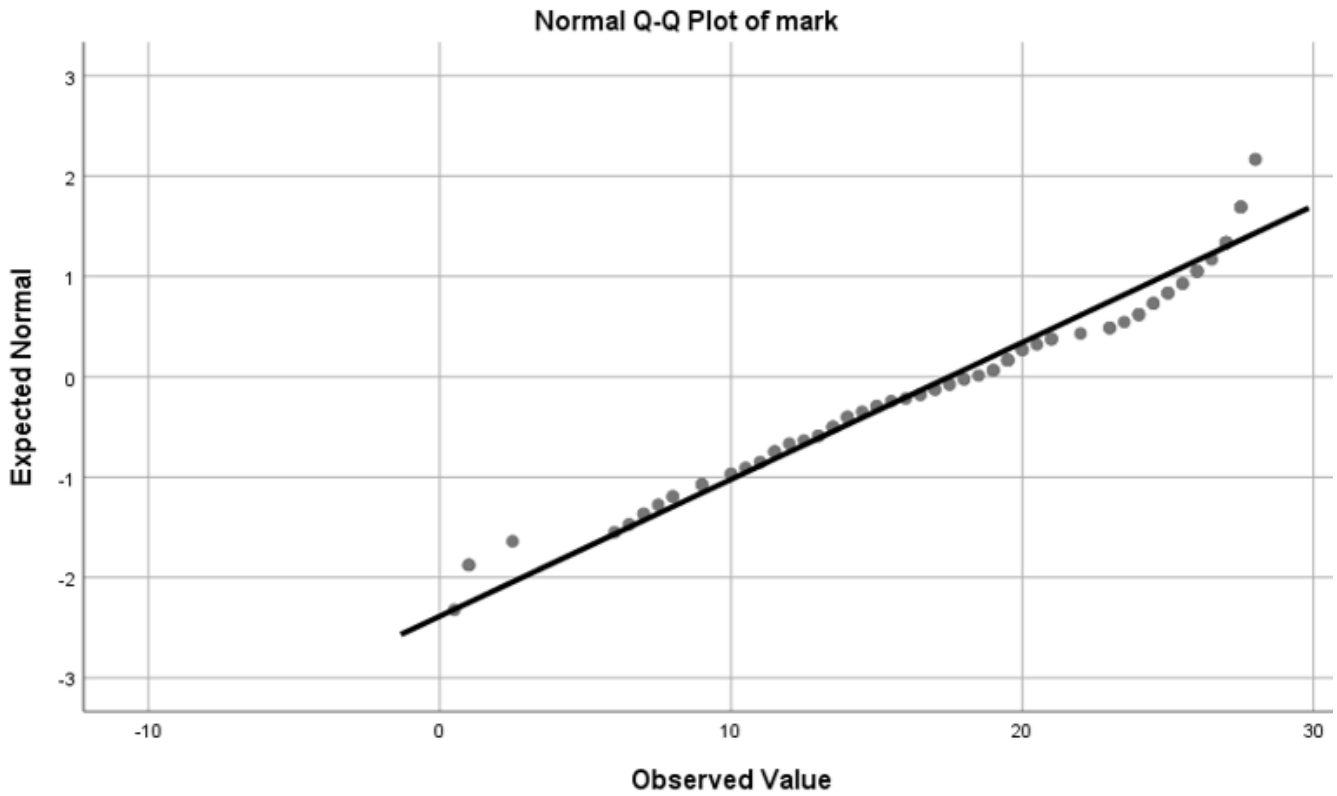
Plots in SPSS: QQ lots (1)

"Descriptive Statistics" → "Explore" → move "mark" into the dependent list panel... (see photo below).



Plots in SPSS: QQ lots (2)

The output is as below where the theoretical quantiles are in the y-axis.



1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

Association

- In the previous section, we covered exploratory techniques for summarizing a single variable at a time.

Association

- In the previous section, we covered exploratory techniques for summarizing a single variable at a time.
- However, many times variables are related or associated with others.

Association

- In the previous section, we covered exploratory techniques for summarizing a single variable at a time.
- However, many times variables are related or associated with others.
- As association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.

Association

- In the previous section, we covered exploratory techniques for summarizing a single variable at a time.
- However, many times variables are related or associated with others.
- As association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.
- We'll check the association for two cases: both are quantitative variables (like weight and height, CA1 and CA2,...) and one quantitative with one categorical variable (like weight and gender, or height and race,...).

Association

- In the previous section, we covered exploratory techniques for summarizing a single variable at a time.
- However, many times variables are related or associated with others.
- As association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.
- We'll check the association for two cases: both are quantitative variables (like weight and height, CA1 and CA2,...) and one quantitative with one categorical variable (like weight and gender, or height and race,...).
- The association between two variables can be explored by numerical summaries or by graphical summaries.

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

Quantifying the Association: Correlation Value

- For the case of two quantitative variables, the often numerical summary that can quantify the relationship between them is the correlation coefficient.

Quantifying the Association: Correlation Value

- For the case of two quantitative variables, the often numerical summary that can quantify the relationship between them is the correlation coefficient.
- Let X and Y are variables from a set of with n objects/people.

Quantifying the Association: Correlation Value

- For the case of two quantitative variables, the often numerical summary that can quantify the relationship between them is the correlation coefficient.
- Let X and Y are variables from a set of with n objects/people.
- The correlation between these two variables is defined as:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

where $\bar{X}, \bar{Y}$ are the sample means, s_X, s_Y are the sample standard deviations of the two variables.

Quantifying the Association: Correlation Value

- For the case of two quantitative variables, the often numerical summary that can quantify the relationship between them is the correlation coefficient.
- Let X and Y are variables from a set of with n objects/people.
- The correlation between these two variables is defined as:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

  where $\bar{X}, \bar{Y}$ are the sample means, s_X, s_Y are the sample standard deviations of the two variables.

- r is always between -1 and 1.

Quantifying the Association: Correlation Value

- For the case of two quantitative variables, the often numerical summary that can quantify the relationship between them is the correlation coefficient.
- Let X and Y are variables from a set of with n objects/people.
- The correlation between these two variables is defined as:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

  where $\bar{X}, \bar{Y}$ are the sample means, s_X, s_Y are the sample standard deviations of the two variables.

- r is always between -1 and 1.
- A positive value for r indicates a positive association and a negative value for r indicates a negative association.

Quantifying the Association: Correlation Value

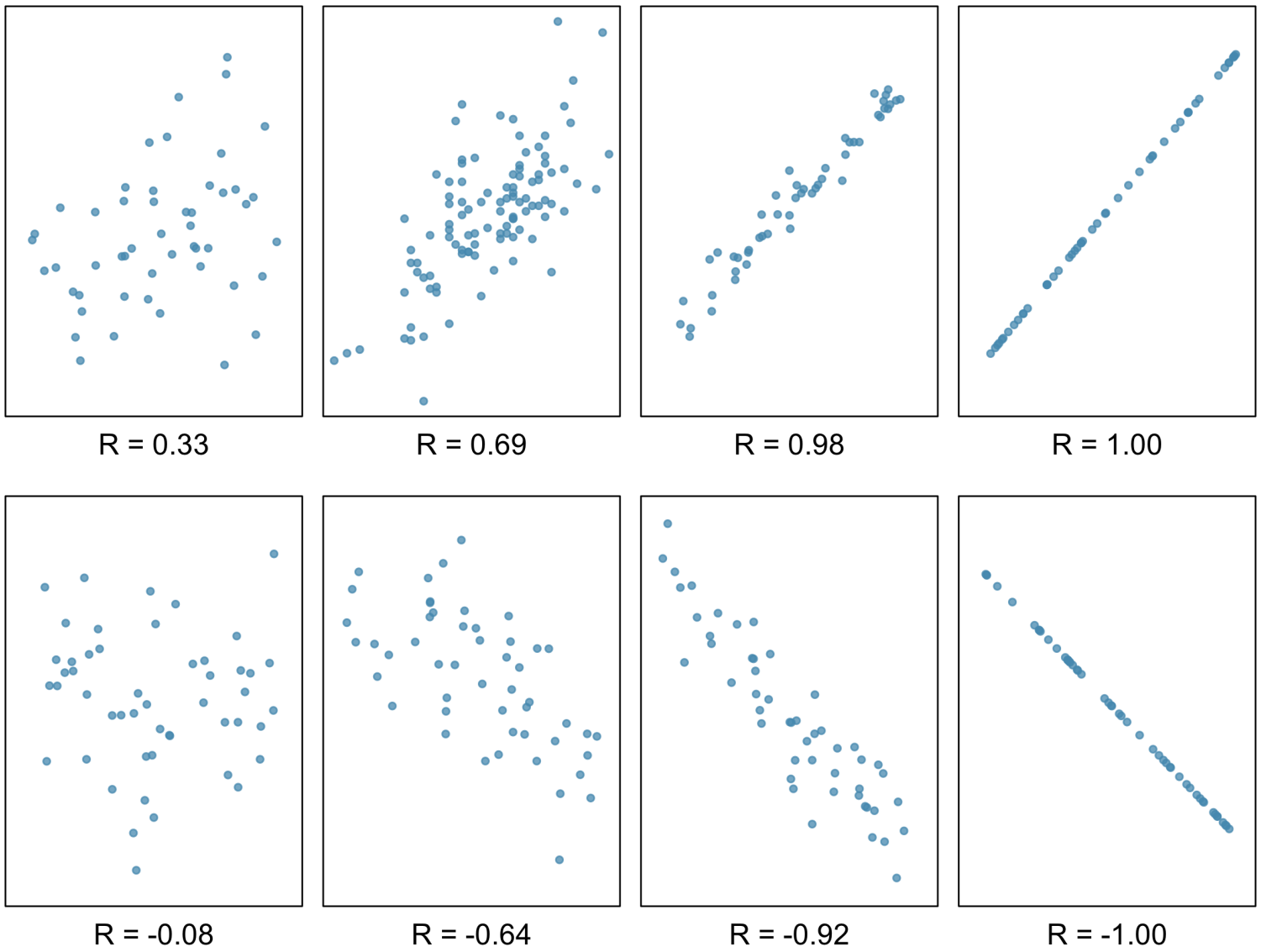
- For the case of two quantitative variables, the often numerical summary that can quantify the relationship between them is the correlation coefficient.
- Let X and Y are variables from a set of with n objects/people.
- The correlation between these two variables is defined as:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

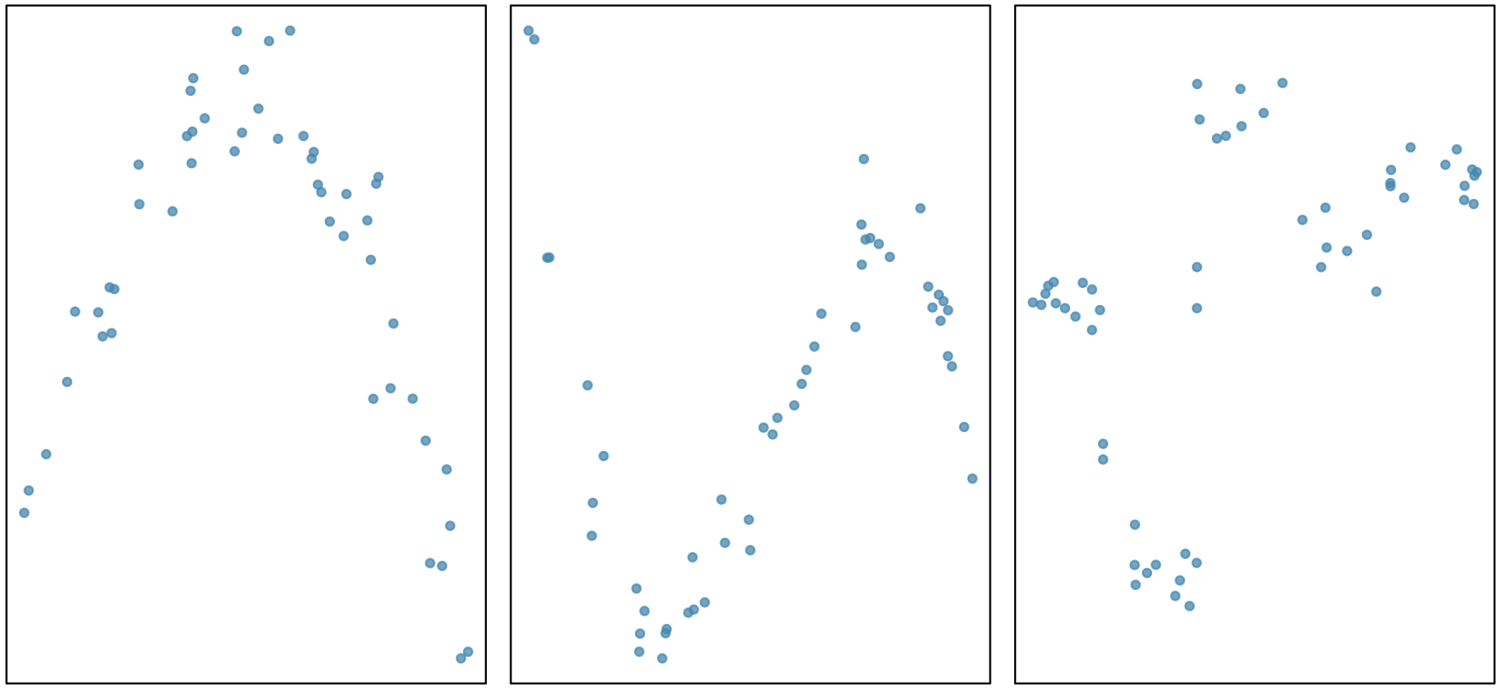
  where $\bar{X}, \bar{Y}$ are the sample means, s_X, s_Y are the sample standard deviations of the two variables.

- r is always between -1 and 1.
- A positive value for r indicates a positive association and a negative value for r indicates a negative association.
- Two variables have the same correlation, no matter which one is treated as the response and which is treated as the explanatory variable.

Examples of Correlation Values (Linear Relationship)



Examples of Correlation Values (Non-Linear Relationship)



A Dataset to Consider

- In this part, we continue to consider the `midterm_marks` dataset.

A Dataset to Consider

- In this part, we continue to consider the `midterm_marks` dataset.

- The score of the final test of the same batch of 98 students are released in the file `final_marks`.

A Dataset to Consider

- In this part, we continue to consider the `midterm_marks` dataset.
- The score of the final test of the same batch of 98 students are released in the file `final_marks`.
- We'll consider the relationship between the midterm from the dataset `midterm_marks` and final marks, to see if they are associated.

A Dataset to Consider

- In this part, we continue to consider the `midterm_marks` dataset.
- The score of the final test of the same batch of 98 students are released in the file `final_marks`.
- We'll consider the relationship between the midterm from the dataset `midterm_marks` and final marks, to see if they are associated.
- The name of the variables could be changed compared to the previous slides, where the midterm mark variable now is set as "midterm" or "M" and final mark variable now is set as "final" or "F".

Calculating Correlation in R

In R:

```
> final<-read.csv("C:/Data/final_marks")
> final <- final[,2]
> midterm<-read.csv("C:/Data/midterm_marks")
> midterm <- midterm[,2]
> cor(final,midterm)
[1] 0.7778648
```

In Python:

```
In [55]: midterm = pd.read_csv (r"C:\Data\midterm_marks")
         midterm.columns = ['Obs', 'M']
         final = pd.read_csv (r"C:\Data\final_marks")
         final.columns =  ['Obs', 'F']
         pearsonr(midterm['M'], final['F'])    #OR

         import numpy as np
         np.corrcoef(midterm['M'], final['F'])[0, 1]

Out[55]: 0.7778647904098349
```

Visualization the Association: Scatterplots

- Scatterplot can help to visualize the association between two quantitative variable well. **What to say given a scatterplot**:

Visualization the Association: Scatterplots

- Scatterplot can help to visualize the association between two quantitative variable well. **What to say given a scatterplot**:

- Is there any (possible) relationship between the 2 variables?

Visualization the Association: Scatterplots

- Scatterplot can help to visualize the association between two quantitative variable well. **What to say given a scatterplot**:
- Is there any (possible) relationship between the 2 variables?
- If yes, is the association positive or negative?

Visualization the Association: Scatterplots

- Scatterplot can help to visualize the association between two quantitative variable well. **What to say given a scatterplot**:
- Is there any (possible) relationship between the 2 variables?
- If yes, is the association positive or negative?
- If there is association, is it linear or non-linear type?

Visualization the Association: Scatterplots

- Scatterplot can help to visualize the association between two quantitative variable well. **What to say given a scatterplot**:
- Is there any (possible) relationship between the 2 variables?
- If yes, is the association positive or negative?
- If there is association, is it linear or non-linear type?
- Are some observations unusual, departing from the overall trend?

Visualization the Association: Scatterplots

- Scatterplot can help to visualize the association between two quantitative variable well. **What to say given a scatterplot**:
- Is there any (possible) relationship between the 2 variables?
- If yes, is the association positive or negative?
- If there is association, is it linear or non-linear type?
- Are some observations unusual, departing from the overall trend?
- Can answer this question also: Is the **variance** of the y-variable stable when the value of the x-variable changes?

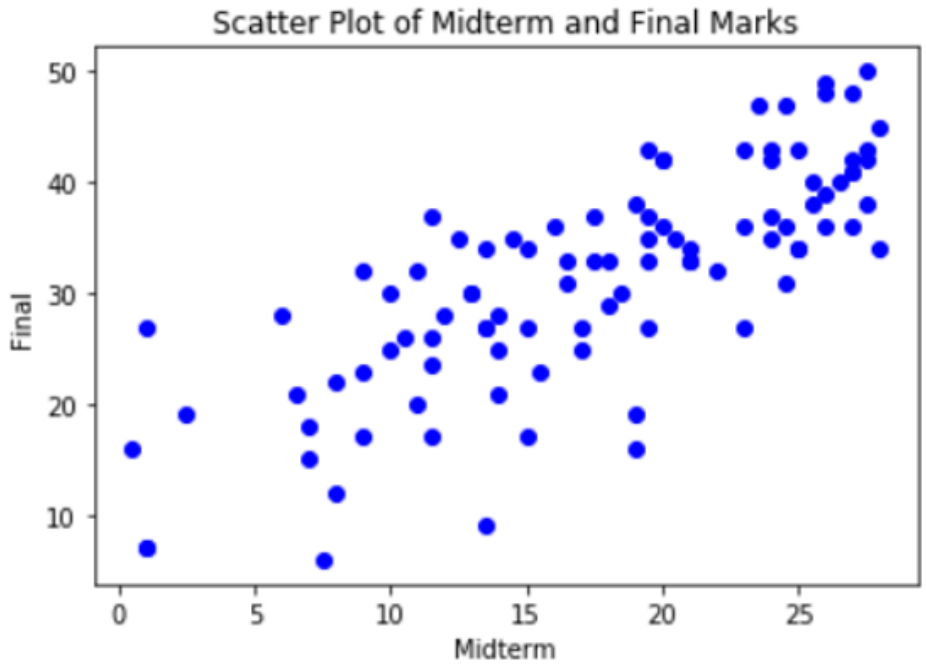
How to Produce Scatterplots

In R (this follows by the code in the previous slide):

```
> plot(midterm,final, pch = 20)
```

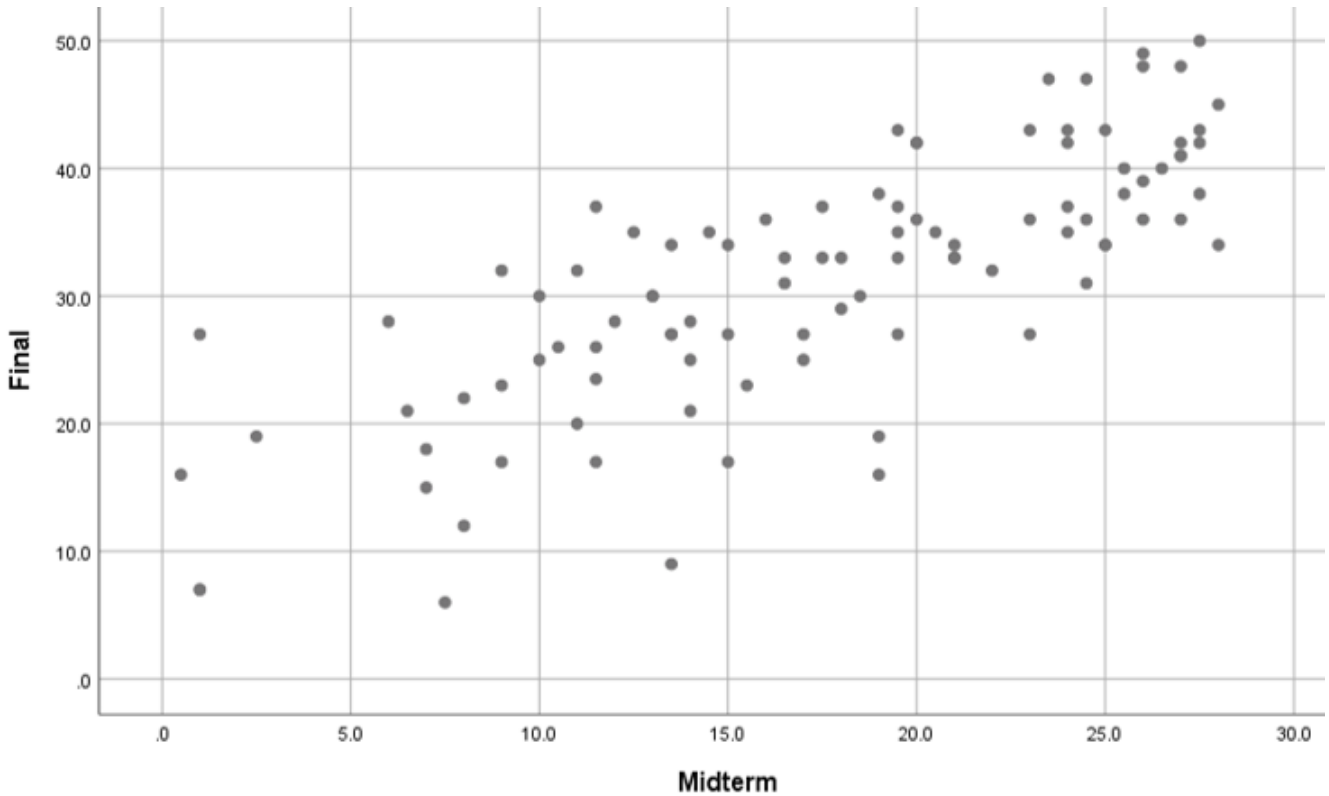
In Python (this follows by the code in the previous slide):

```
In [57]: import matplotlib.pyplot as plt

plt.scatter(midterm['M'], final['F'],label='Scatter Plot 1',color='b')
pyplot.xlabel('Midterm')
pyplot.ylabel('Final')
pyplot.title('Scatter Plot of Midterm and Final Marks')
pyplot.show()
```



Scatterplots in SPSS

- Firstly, open a data file (either the midterm or the final) then merge the other file by variable.
- "Graphs" → "Legacy Dialogs" → "Scatter/Dot..." → "Simple Scatter" → "Define" → move "mark" to X-axis and "final" to Y-axis → "Tittle" if you want → "OK".
- The output is:



1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

An Example

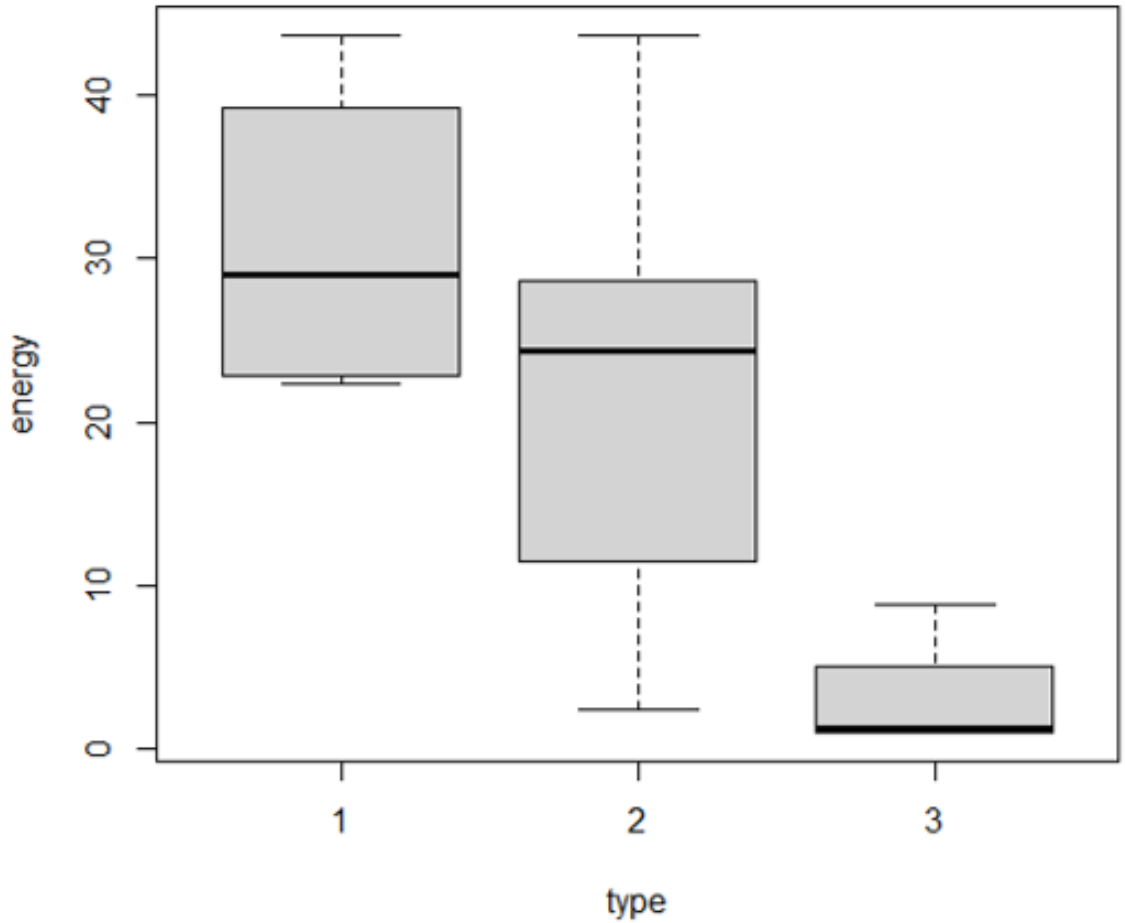
Example (Orientation of bats and birds)

To orientate themselves with respect to their surroundings, some bats use echolocations, i.e. they send out pulses (while flying) and read the echoes that are bounced back from surrounding objects.
Investigators are interested in explaining the total amount of energy expended by such bats when flying, after adjusting for body mass. The data is given in the file `bats.csv`, where the variable **type** has three different values:
Type 1: echolocating bats,
Type 2: non-echolocating bats and
Type 3: non-echolocating birds.

- This data have 3 variables: **mass** and **energy** are quantitative variables while **type** is a categorical variable with 3 categories. We would examine if **different types expends the energy differently**.
- In order to do so, we'll split the variable energy into 3 groups of type and observe the difference.
- Few often ways to visualize the difference is using boxplots or histograms of **energy by type**.

Boxplots of Multiple Groups in R

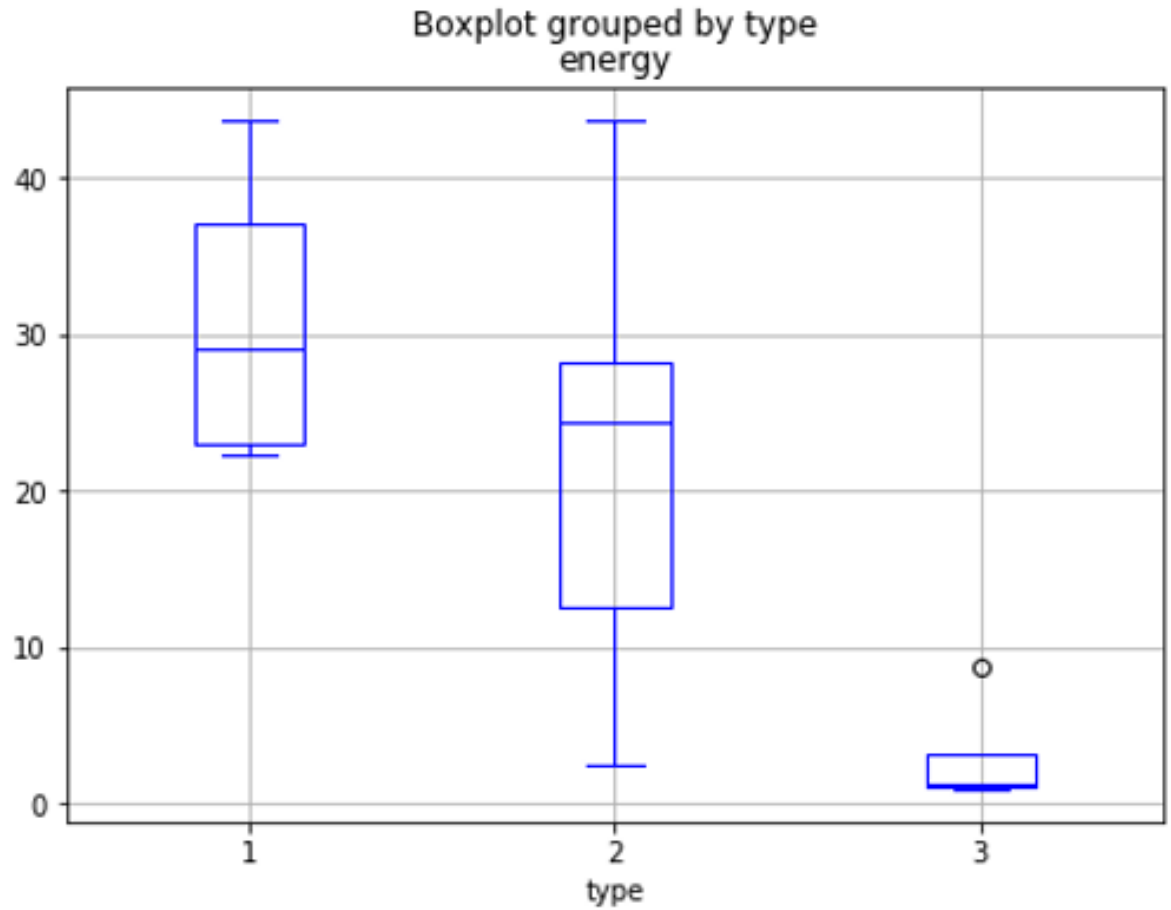
```
> bats<-final<-read.csv("C:/Data/bats.csv")
> bats <-data.frame(bats)
> attach(bats)
> boxplot(energy ~type)
```



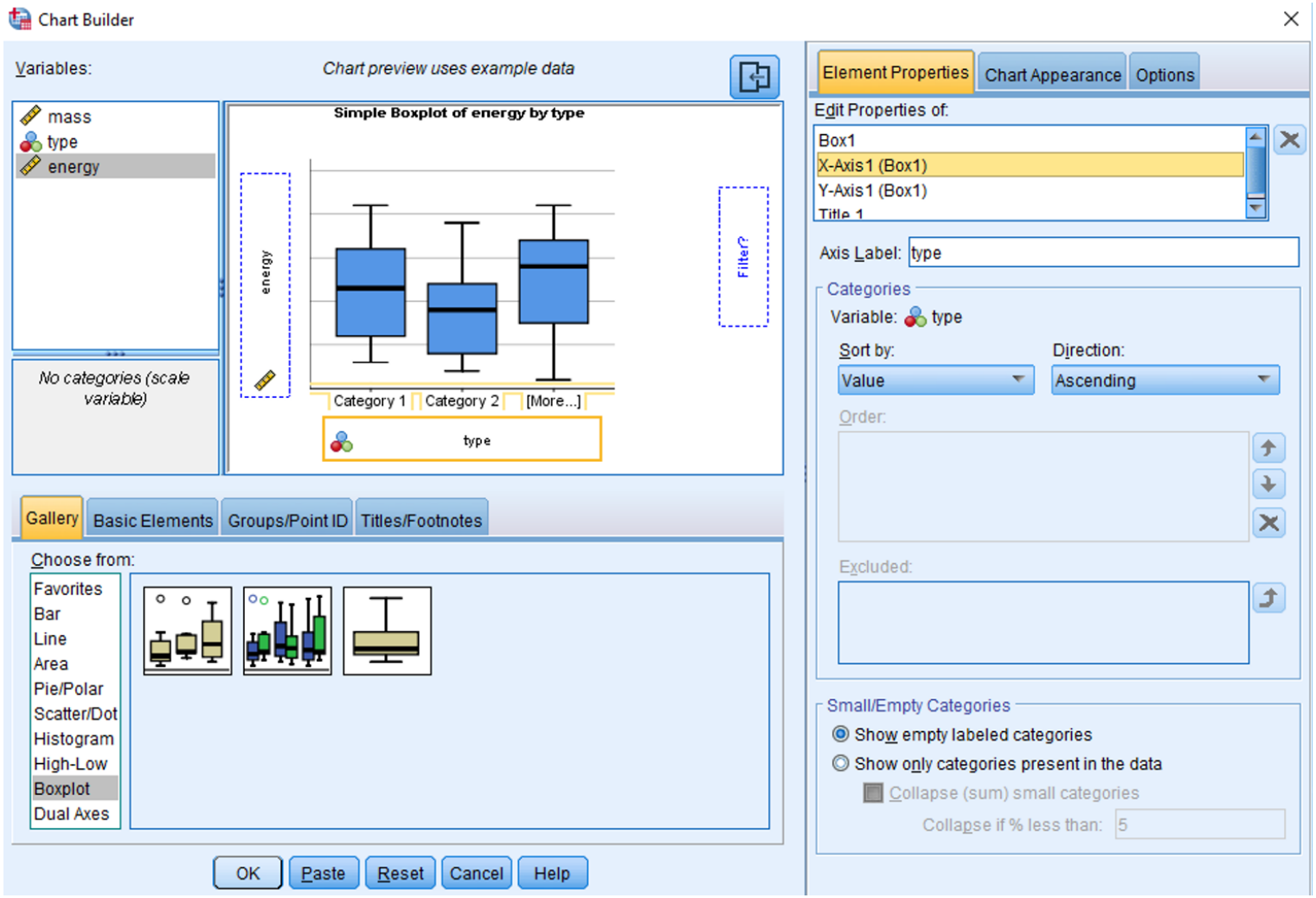
Boxplots of Multiple Groups in Python

```
import pandas as pd
bats = pd.read_csv('C:/Data/bats.csv')
fig, ax = plt.subplots(figsize=(7,5))
bats.boxplot(column=['energy'], by='type', ax=ax, color = 'b')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a523f9a048>
```

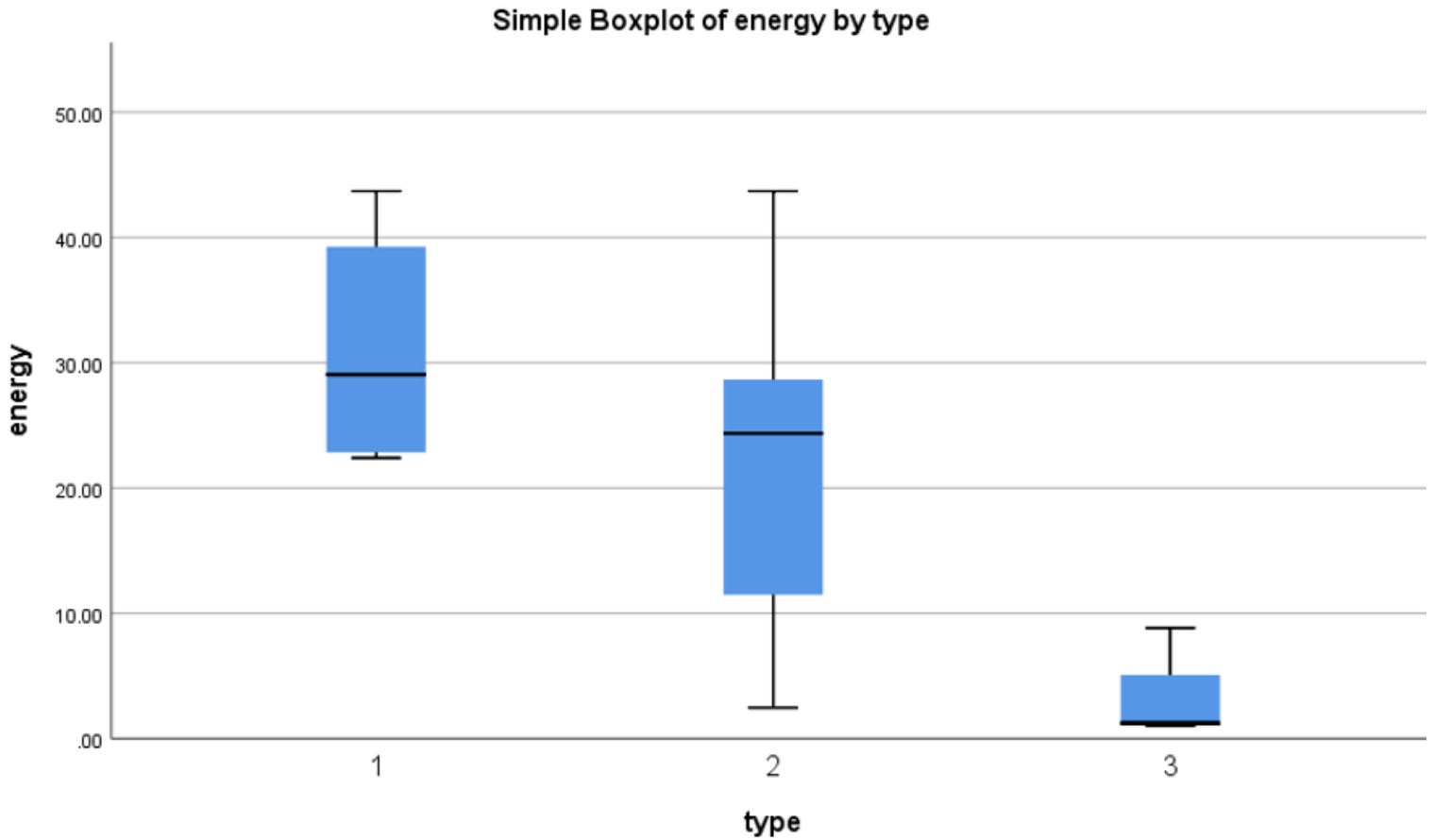


Boxplots of Multiple Groups in SPSS (1)



Boxplots of Multiple Groups in SPSS (2)

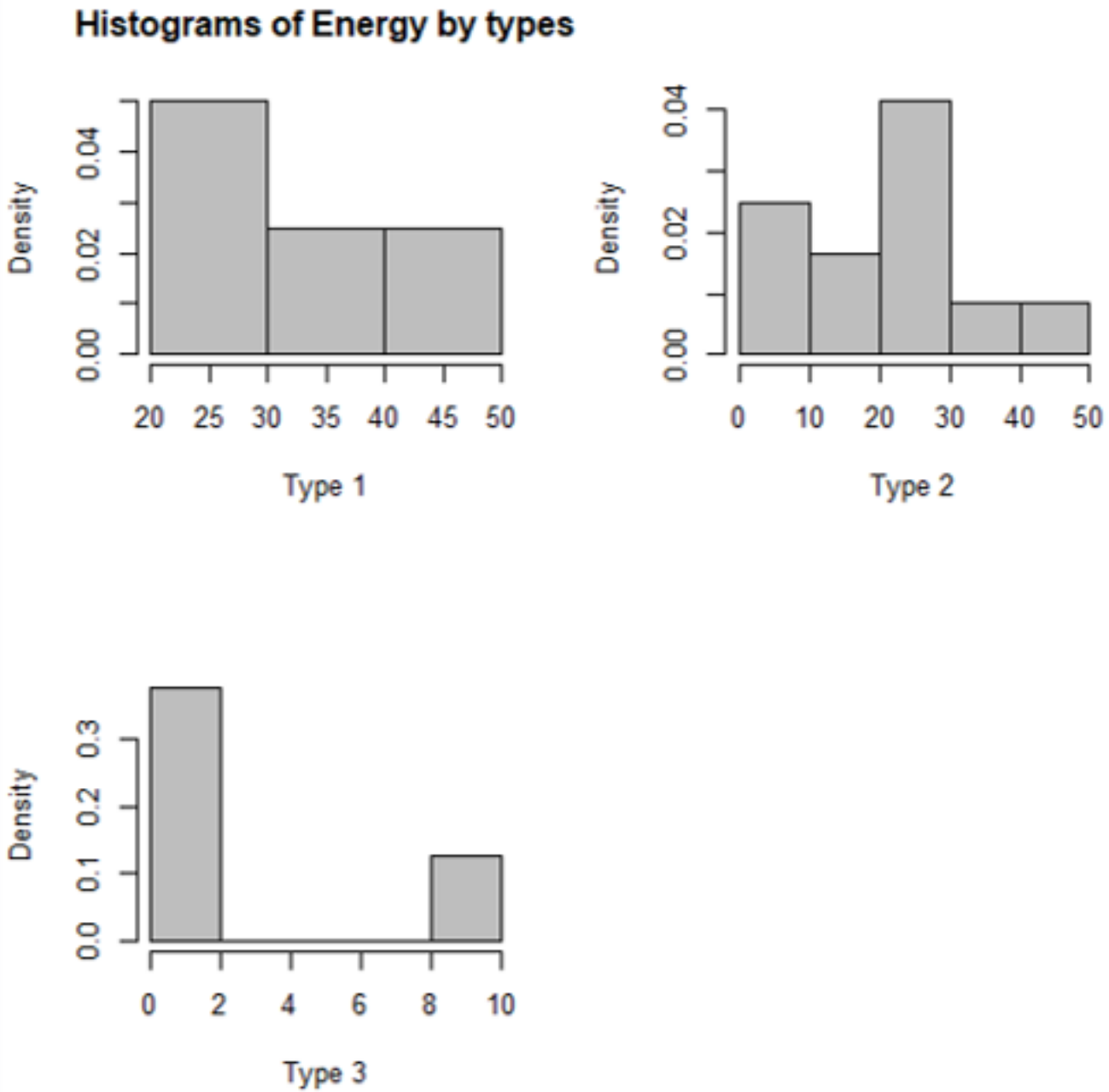
The output is:



Histograms of Multiple Groups in R (1)

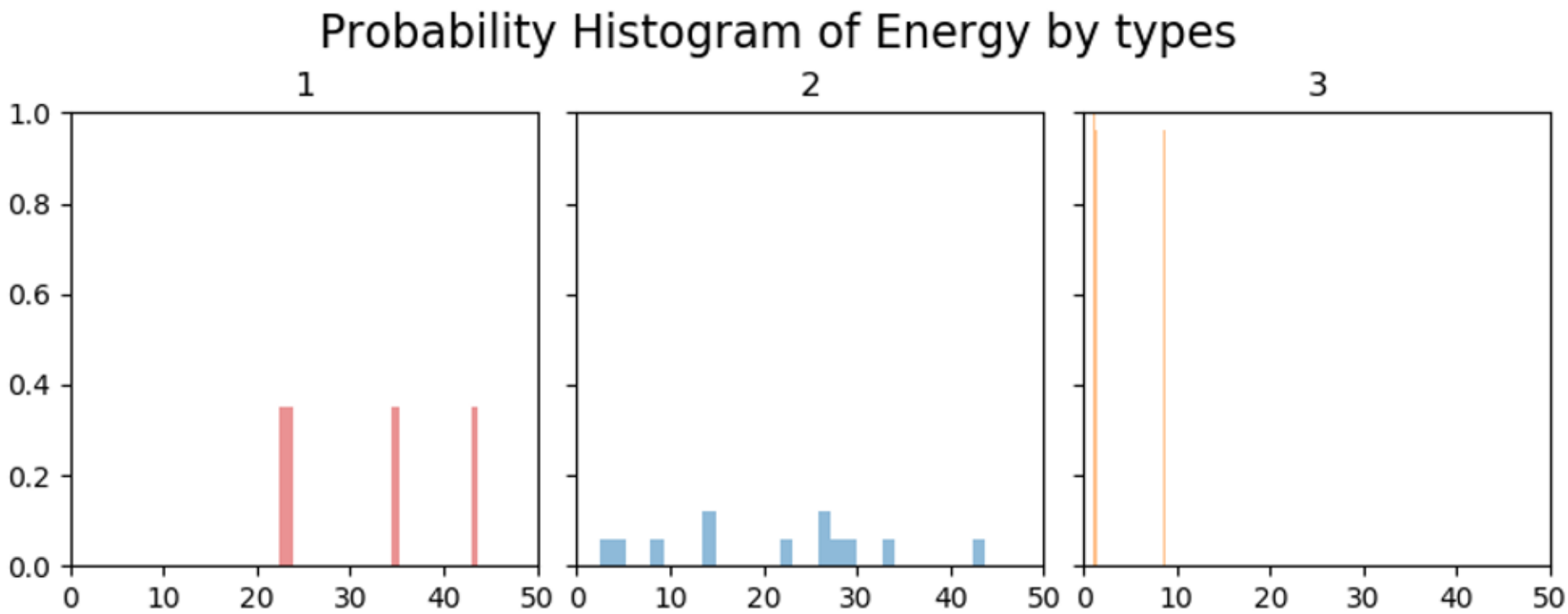
```
> # To specify that 3 graphs in one column in one page
> par(mfrow=c(2,2))
> #Histogram for the energy of type 1
> hist(energy[which(type ==1)], include.lowest = TRUE,freq=FALSE,
+   col="grey",xlab = "Type 1",main ="Histograms of Energy by types'
> hist(energy[which(type ==2)], include.lowest = TRUE, freq=FALSE,
+      col="grey", xlab = "Type 2",main = "")
> hist(energy[which(type ==3)], include.lowest = TRUE, freq=FALSE,
+      col="grey", xlab = "Type 3",main = "")
> #To get back to 1 graph in one page.
> par(mfrow=c(1,1))
```

Histograms of Multiple Groups in R (2)

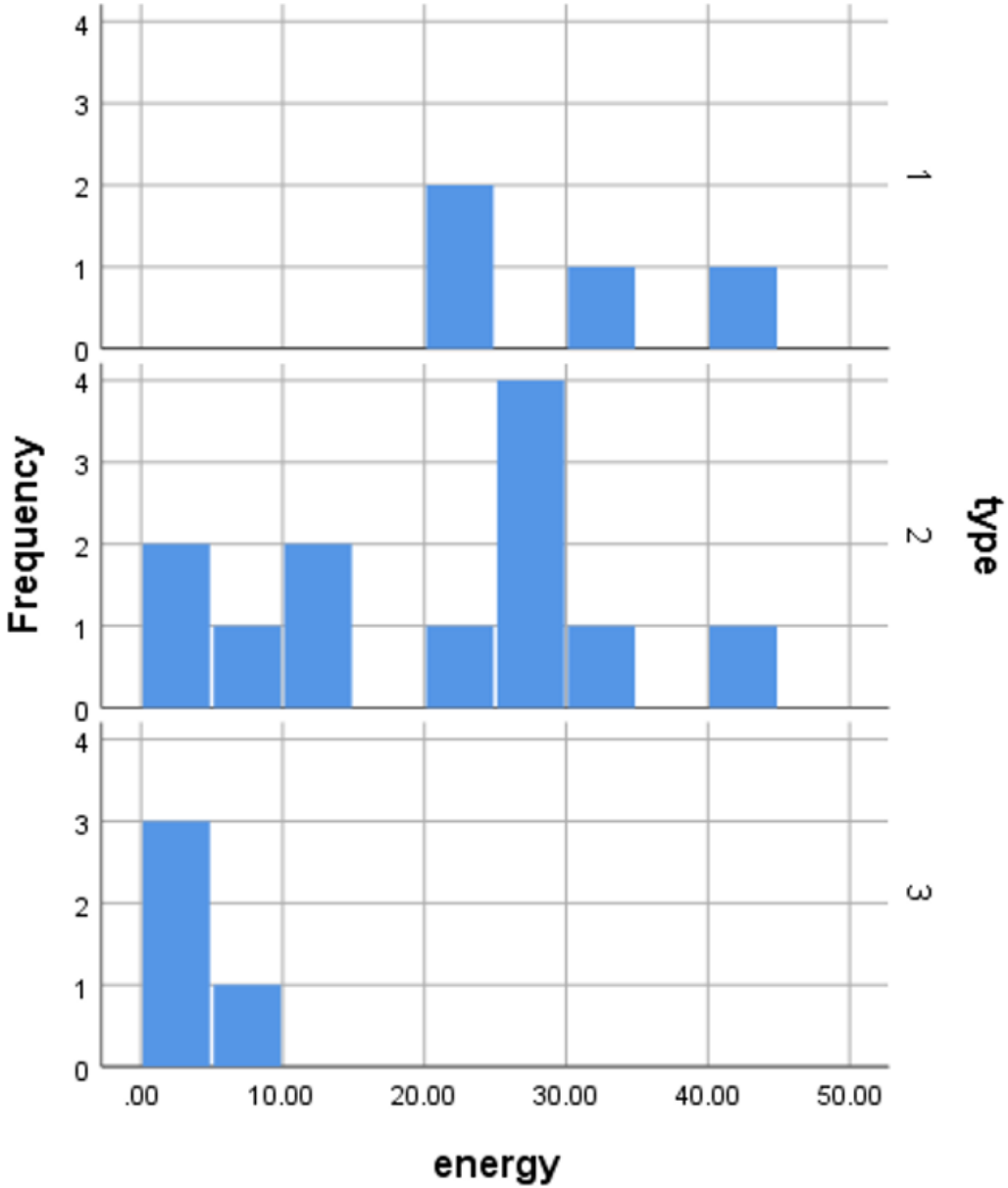


Histograms of Multiple Groups in Python

```
fig, axes = plt.subplots(1, 3, figsize=(8,3), dpi=100, sharex=True, sharey=True)
colors = ['tab:red', 'tab:blue', 'tab:orange']

for i, (ax, type) in enumerate(zip(axes.flatten(), bats.type.unique())):
    x = bats.loc[bats.type==type, 'energy']
    ax.hist(x, alpha=0.5, bins=30, density=True, stacked=True, label=str(type), color=colors[i])
    ax.set_title(type)

plt.suptitle('Probability Histogram of Energy by types', y=1.05, size=16)
ax.set_xlim(0, 50); ax.set_ylim(0, 1);
plt.tight_layout();
```



Histograms of Multiple Groups in SPSS



Comments on the Plots of Energy by Types

- Firstly, this dataset is not large, only 20 observations. Comments about the trend of data based on the figures might not be very helpful.

Comments on the Plots of Energy by Types

- Firstly, this dataset is not large, only 20 observations. Comments about the trend of data based on the figures might not be very helpful.

- From the boxplots, we can observe the difference of the energy expended by different type of bats/birds, especially between type 1 (echolocating bats) and type 3 (non-echolocating birds) and between type 2 (non-echolocating bats) and type 3, where type 3 expends lower amount of energy.

Comments on the Plots of Energy by Types

- Firstly, this dataset is not large, only 20 observations. Comments about the trend of data based on the figures might not be very helpful.

- From the boxplots, we can observe the difference of the energy expended by different type of bats/birds, especially between type 1 (echolocating bats) and type 3 (non-echolocating birds) and between type 2 (non-echolocating bats) and type 3, where type 3 expends lower amount of energy.

- From the histogram, we cannot say much about the distribution of the energy for each type, since the number of data points from different types is not large, especially type 3 with only 4 data points. However we can observe the range of energy values from the 3 types: The range of type 3 is much smaller (0 - 10).

1. Introduction

2. Single Quantitative Variable Exploration
   - Numerical Summaries
   - Graphical Summaries: Analysis
   - Graphical Summaries: How to Plot

3. Association Between Two Variables
   - Two Quantitative Variables
   - One Categorical and One Quantitative Variable
   - Two Categorical Variables

- The association between two categorical variables (by numerical and graphical summaries) will be presented in a later topic.