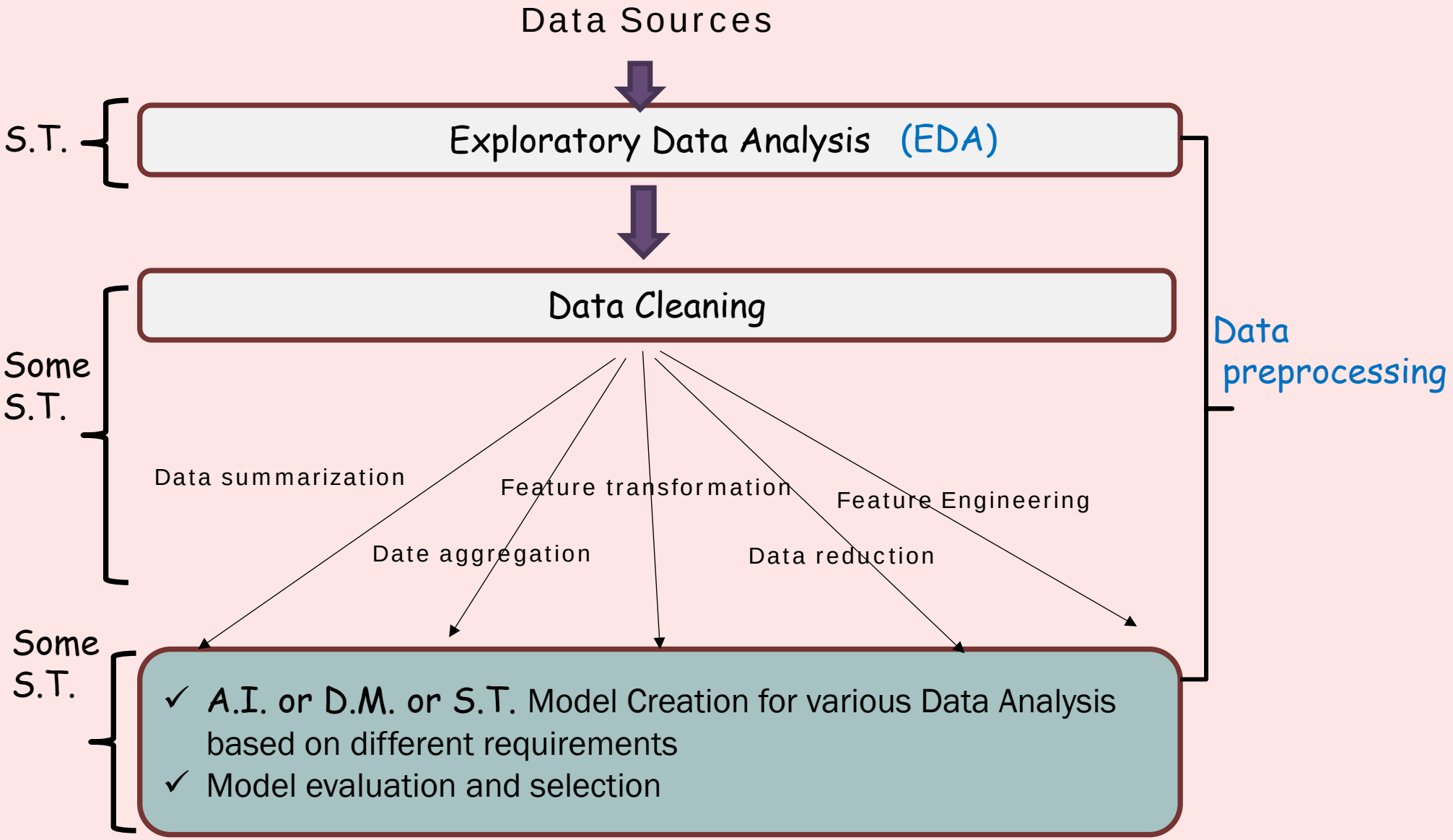
EXPLORATORY DATA ANALYSIS

Contents

Overview

Exploratory Data Analysis (EDA)

- In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.
- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
- Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling of your data. It provides the context needed to develop an appropriate model – and interpret the results correctly
- Exploratory Data Analysis (EDA) is, the most important part of Machine Learning Modeling in new datasets. If EDA is not executed correctly, it can cause us to start modeling with "unclean" data, and this is just as a snowball rolling downhill, it gets bigger and worse.
- EDA is presented in two ways: digital exploration and visual exploration.



* Nowdays data preprocessing contains data cleaning for getting bigger and complicated data.

Used Data Set for this Section

- The MASS package and Insurance data set are used for this Chapter.
- The relevant instructions for loading and viewing the data set are below.

```
install.packages("MASS")
library(MASS)
data(Insurance)
head(Insurance)
nrow(Insurance);ncol(Insurance)
dim(Insurance)
```

```
> head(Insurance)
  District  Group    Age Holders Claims
1        1    <1l    <25      197     38
2        1    <1l  25-29      264     35
3        1    <1l  30-35      246     20
4        1    <1l    >35     1680    156
5        1 1-1.5l    <25      284     63
6        1 1-1.5l  25-29      536     84
> nrow(Insurance);ncol(Insurance)
[1] 64
[1] 5
> dim(Insurance)
[1] 64  5
```

- The data set is structured for car policyholder below.

Format

This data frame contains the following columns:

`District`

factor: district of residence of policyholder (1 to 4): 4 is major cities.

`Group`

an ordered factor: group of car with levels <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.

`Age`

an ordered factor: the age of the insured in 4 groups labelled <25, 25–29, 30–35, >35.

`Holders`

numbers of policyholders.

`Claims`

numbers of claims

Digital Exploration

Digital Exploration

- The ideas are to retrieve relevant indexes for data exploration such as data structure(dataset shape)、variable types and situation 、 data distribution 、 missing and null values 、data correlation 、duplicated values, and descriptive statistics.
- EDA focuses on exploring data to understand the data's underlying structure and variables, to develop intuition about the data set, to consider how that data set came into existence, and to decide how it can be investigated with more formal statistical methods.

Understand Variables Structure –summary() 1/n

- The following are used to understand variables

```
names(Insurance)
attributes(Insurance)
str(Insurance)
summary(Insurance)
```

- The execution is as illustrated:
- Based on you have learned, what are the data types?
- They are both measures of central tendency (along with the mode and the midrange). The mean is the sum of the data point's values divided by the number of data points. The median is the geographic middle of the data when the list of data is put in ascending order.

```
> names(Insurance)
[1] "District" "Group"    "Age"      "Holders"  "Claims"
> attributes(Insurance)
$names
[1] "District" "Group"    "Age"      "Holders"  "Claims"

$class
[1] "data.frame"

$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34
[35] 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
60 61 62 63 64

> str(Insurance)
'data.frame':   64 obs. of  5 variables:
 $ District: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 $ Group   : Ord.factor w/ 4 levels "<1l"<"1-1.5l"<..: 1 1 1 1 2 2 2 2 3 3 ...
 $ Age     : Ord.factor w/ 4 levels "<25"<"25-29"<..: 1 2 3 4 1 2 3 4 1 2 ...
 $ Holders : int  197 264 246 1680 284 536 696 3582 133 286 ...
 $ Claims  : int  38 35 20 156 63 84 89 400 19 52 ...
> summary(Insurance)
 District    Group       Age        Holders           Claims
 1:16     <1l   :16   <25  :16   Min.   :   3.00   Min.   :  0.00
 2:16     1-1.5l:16   25-29:16   1st Qu.:  46.75   1st Qu.:  9.50
 3:16     1.5-2l:16   30-35:16   Median : 136.00   Median : 22.00
 4:16     >2l   :16   >35  :16   Mean   : 364.98   Mean   : 49.23
                                 3rd Qu.: 327.50   3rd Qu.: 55.50
                                 Max.   :3582.00   Max.   :400.00
```

Understand Variables Structure - summary() 2/n

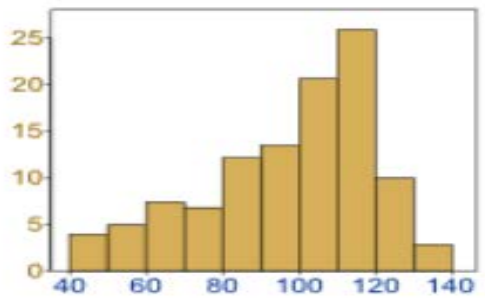
- You can see the variables' structures based on the following:

```
> attributes(Insurance)
$names
[1] "District" "Group"    "Age"      "Holders"  "Claims"

$class
[1] "data.frame"

$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
[34] 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
```

```
    Holders          Claims
Min.   :   3.00   Min.   :  0.00
1st Qu.:  46.75   1st Qu.:  9.50
Median : 136.00   Median : 22.00
Mean   : 364.98   Mean   : 49.23
3rd Qu.: 327.50   3rd Qu.: 55.50
Max.   :3582.00   Max.   :400.00
```

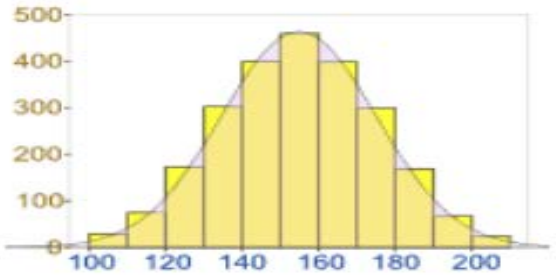
- Then, use **str()** to see the internal structure. As you can see, the number of observation is 64 in which the sample have 5 variables, and 4 levels in the first 3 variables: District, Group , and Age.
- And, use summary() to obtain the descriptive statistics. The first 3 variables have different data types from the last 2 variables.
- And, we can check the mean and median to see how skew the data is. If the value difference of the two statistics numbers is large, then left or right skewed distribution is obvious.
- For example on the variable **holders**, the data distribution presents **right skewed** distribution when its mean value doubles in median value. Further, the data of variable **holders likely** exists abnormal values resulting in the big difference between mean and median. ☺

More talks on skew distributions and abnormal values

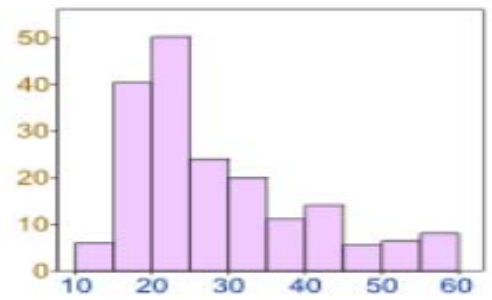


Negative Skew

(Left skew)

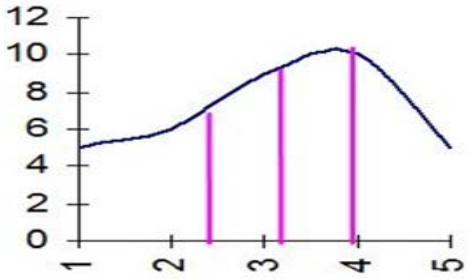


No Skew

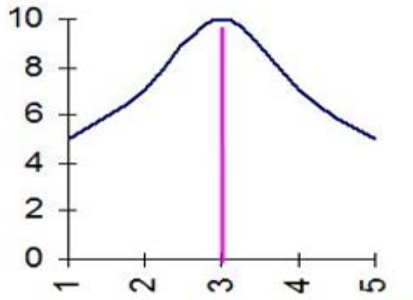


Positive Skew

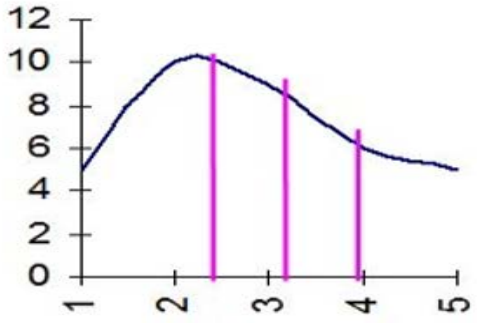
(Right skew)



Mean< Med<Mode



Mean = Median = Mode



Mode < Med< Mean

p.s.

1. Suggest to have your mode calculated to ensure your data-skew distribution. ☺
2. Find out your abnormal values to ensure the model creation is all right. ☺

Variables Description - describe() 1/n

- The Harrell Miscellaneous package contains many functions useful for data analysis, high-level graphics, utility operations, functions for computing sample size and power, importing and annotating datasets, imputing missing values, advanced table making, variable clustering, character string manipulation, conversion of R objects to LaTeX and html code, and recoding variables.

```
install.packages("Hmisc")
library(Hmisc)
```

- use describe() to view data distribution in more detailed.

Quantile the values to see the data distribution

Gmd: Gini's Mean Difference. Someone use it instead of variance.
Info: (關於變量的連續性強度) is a data distribution measure using the relative efficiency of a proportional odds/Wilcoxon test on the variable relative to the same test on a variable that has **no ties**. Info is related to how continuous the variable values are, and **ties are less harmful** while the more untied values there are. ☺

n: number of samples.
missing: number of missing samples
distinct: (unique) number of distinct values

- Use describe() based on different types

```
> describe(Insurance[,1:3])
Insurance[, 1:3]

 3  Variables      64  Observations
---------------------------------------
District
       n  missing distinct
      64        0        4

Value          1     2     3     4
Frequency     16    16    16    16
Proportion  0.25  0.25  0.25  0.25
---------------------------------------
Group
       n  missing distinct
      64        0        4

Value          <1l 1-1.5l 1.5-2l    >2l
Frequency       16     16     16     16
Proportion    0.25   0.25   0.25   0.25
---------------------------------------
Age
       n  missing distinct
      64        0        4

Value        <25 25-29 30-35   >35
Frequency     16    16    16    16
Proportion  0.25  0.25  0.25  0.25
---------------------------------------
```

```
> describe(Insurance[,4:5])
Insurance[, 4:5]

 2  Variables      64  Observations
---------------------------------------------------------------------------------------
Holders
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
      64        0       63        1      365    497.1    16.30    24.00    46.75   136.00
     .75      .90      .95
  327.50   868.90  1639.25

lowest :    3    7    9   16   18, highest: 1635 1640 1680 2443 3582
---------------------------------------------------------------------------------------
Claims
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
      64        0       46    0.999    49.23    60.66     3.15     4.30     9.50    22.00
     .75      .90      .95
   55.50   101.70   182.35

lowest :   0   2   3   4   5, highest: 156 187 233 290 400
---------------------------------------------------------------------------------------
```

Variables Description - describe() 2/n

- use describe() to view data distribution in more detailed.
- However, describe() justifies data type based on: (**constraints**)
  - if the number of distinct values for a variable with numerical type is not greater than 10, the variable is treated as **categorical** type. ☺
  - If the number of distinct values for a variable not greater than 20 and greater than 10, describe() shows frequency table.
  - If a variable has a number of samples greater than 20, the 5 lowest and 5 highest frequencies are listed.
  - Namely, describe() uses biased estimator for small samples while using unbiased estimator for non-small samples. (p.s. In statistics, **bias** (or bias function) of **an estimator** is the difference between this estimator's **expected value** and **the true value of the parameter being estimated.** An estimator or decision rule with zero bias is called unbiased. Otherwise the estimator is said to be biased).
  - The variable **claims** has more tied values as compared to Holders according to the **distinct** and **Info** in which Info is calculated through *a proportional odds/Wilcoxon test*.

```
Holders
       n  missing distinct      Info
      64        0       63         1
     .95
 1639.25

lowest :    3    7    9   16   18,
-----------------------------------
Claims
       n  missing distinct      Info
      64        0       46     0.999
     .95
  182.35
```

Variables Description - describe() 3/n

- Of all measures of variability, the **variance** is the most popular by far .
- However, Gini's Mean Difference (**Gmd**) is an alternative index of variability, shares many properties with the variance, but can be more informative about the properties of distributions that depart from normality.
- ***Gini's Mean Difference*** *can be used to check non-normal distribution based on left or right skewed distribution.*
- For example, the data distributions of the variables **Holders** and **Claims** presents **right skewed** distribution according to the mean and median, and **Gmd** in which the values are 4971. and 60.66 respectively. However, **Gmd** efficiency is sensitive to number of division, and degree of dispersion.
- The more division the quantile divides, the more completed data distribution you will get. However, sometimes, the situation may cause inaccurate **Gmd** value.

fBasics package – basicStats() 1/n

- The Rmetrics "fBasics" package is a collection of functions to explore and to investigate basic properties of financial returns and related quantities.
- The covered fields include techniques of explorative data analysis and the investigation of distributional properties, including parameter estimation and hypothesis testing.
- Even more there are several utility functions for data handling and management.
- BasicStatistics () computes basic financial time series statistics. For now, we check the data in the variable Holders as illustrated.

```
install.packages("fBasics")
library(fBasics)
basicStats(Insurance$Holders)
```

```
> basicStats(Insurance$Holders)
             X..Insurance.Holders
nobs                 6.400000e+01
NAs                  0.000000e+00
Minimum              3.000000e+00
Maximum              3.582000e+03
1. Quartile          4.675000e+01
3. Quartile          3.275000e+02
Mean                 3.649844e+02
Median               1.360000e+02
Sum                  2.335900e+04
SE Mean              7.784632e+01
LCL Mean             2.094209e+02
UCL Mean             5.205478e+02
Variance             3.878432e+05
Stdev                6.227706e+02
Skewness             3.127833e+00
Kurtosis             1.099961e+01
```

- The S.T. values in the red frame are different from summary()
- **nobs**: #observation

  NAs: #missing values

  Sum: summarize the values

  SE Mean: standard error mean

  LCL/ UCL mean: Lower/ Upper confidence interval for mean value

  variance: it measures how far a set of (random) numbers are spread out from their average value

  stdev: standard deviation

  skewness (偏度)

  kurtosis (峰度)

fBasics" package - basicStats() 2/n

- According to the sum value, there are totally 23,359 insurance holders in the data set.
- As listed, the average of insurance holders is 365 (e.g. 23,359/64) under considering District, Group, and Age.
- The mean value 365 is reliable while value is in the confidence interval [209, 521] (i.e. [LCL mean, UCL mean]). Namely, if a mean value is in a confidence interval, we say that the mean value has 95% confidence. It is credible.
- Skewness and kurtosis are discussed in data distribution.

  skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.
- Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable.
- For analyzing data, please you explore data according to the required information. Rather than exploring data as more as you can.

```
> basicStats(Insurance$Holders)
             X..Insurance.Holders
nobs                 6.400000e+01
NAs                  0.000000e+00
Minimum              3.000000e+00
Maximum              3.582000e+03
1. Quartile          4.675000e+01
3. Quartile          3.275000e+02
Mean                 3.649844e+02
Median               1.360000e+02
Sum                  2.335900e+04
SE Mean              7.784632e+01
LCL Mean             2.094209e+02
UCL Mean             5.205478e+02
Variance             3.878432e+05
Stdev                6.227706e+02
Skewness             3.127833e+00
Kurtosis             1.099961e+01
```

Distribution Index

- In Statistics, Binomial distribution(二項分佈), Poisson distribution(卜瓦松分佈) , and Geometric distribution(幾何分佈) are used for discrete variables.
- In Statistics, Uniform distribution (均勻分布), exponential distribution (指數分佈), Normal distribution (常態分佈) are used for continuous variables.
- However, a data distribution is a function or a listing which shows all the possible values (or intervals) of the data. It also (and this is important) tells you how often each value occurs.
- In this section, skewness and kurtosis are discussed and explored by using basicStats() in fBasics package ,and skewness()+kurtosis() in timeDate package for continuous variables.
- Kurtosis is the degree of peakedness of a distribution

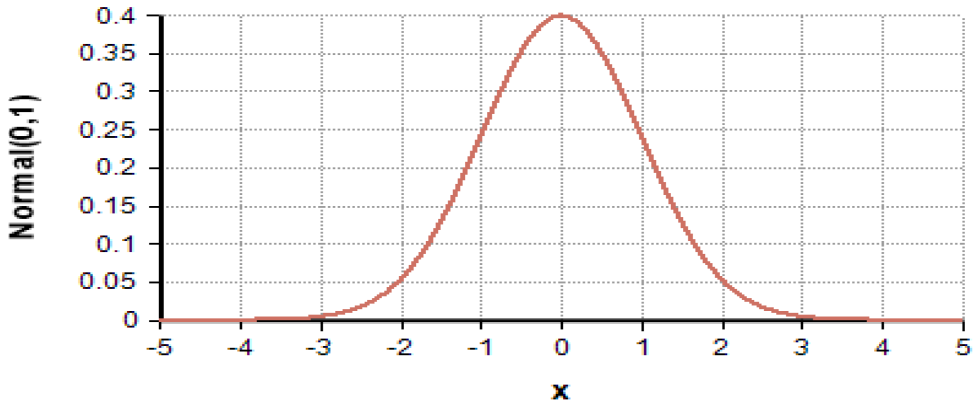
```
#Data-distribution index
install.packages("timeDate")
library(timeDate)
skewness(Insurance[,4:5])
kurtosis(Insurance[,4:5])
```

```
> library(timeDate)
> skewness(Insurance[,4:5])
 Holders   Claims
3.127833 2.877292
> kurtosis(Insurance[,4:5])
  Holders    Claims
10.999610  9.377258
```

p.s. As the numbers shown above, the skewness and kurtosis of the variable Holders are higher than the other one. Thus, suggest to do the data cleaning for the abnormal data in the variable Holder while considering skew distribution. By the way, they are right skewed because its skewness > 1.

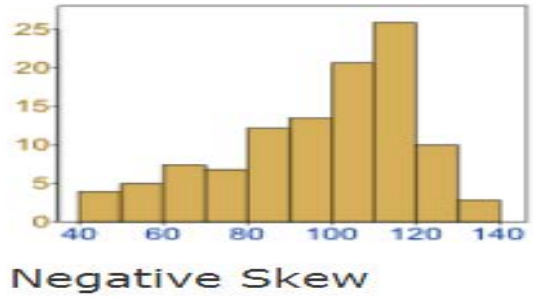
Distribution Index - skewness 1/n

- In probability theory and statistics, the skew normal distribution is a continuous probability distribution that generalizes the normal distribution to allow for **non-zero skewness**.
- If a normal distribution is limited to the interval [-1, 1] in axis x, that means the symmetry of the data distribution is stronger. And, its skewness is 0 as illustrated. While normal distribution is one of the most common forms of distribution, not all data sets follow this basic curve.
- The data distribution **is right skewed if its skewness > 1**, and is left skewed if its skewness < -1.
- Please pay more attention on that if your data distribution does not belong to a normal distribution, then does not mean you need to change your data distribution to a normal distribution.

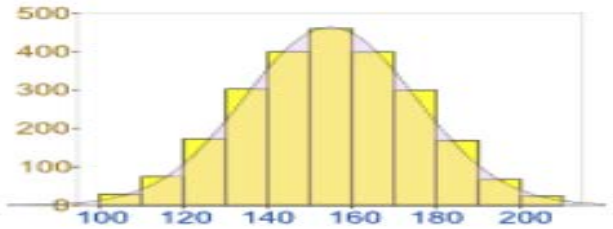


Distribution Index - skewness 2/n

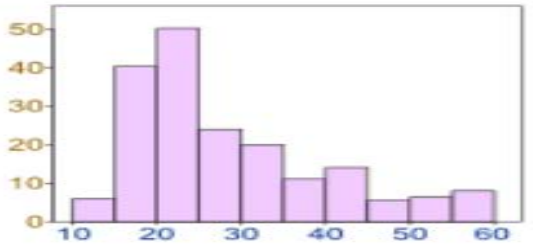
- Data can be "skewed", meaning it tends to have a **long tail** on one side or the other. For example,



Negative Skew

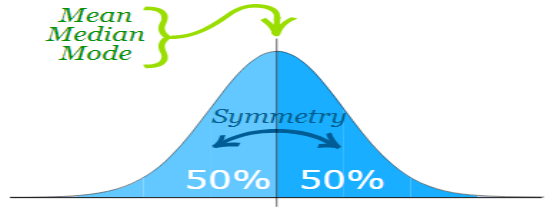


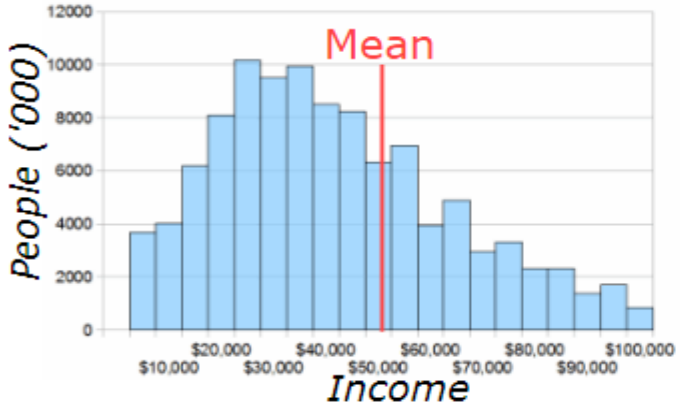
No Skew



Positive Skew

- Why is it called negative/ positive skew? Because the long "tail" is on the negative/ positive side of the peak.
- The Normal Distribution has No Skew
- For example,





Example: Income Distribution

Here is some data extracted from a recent Census.

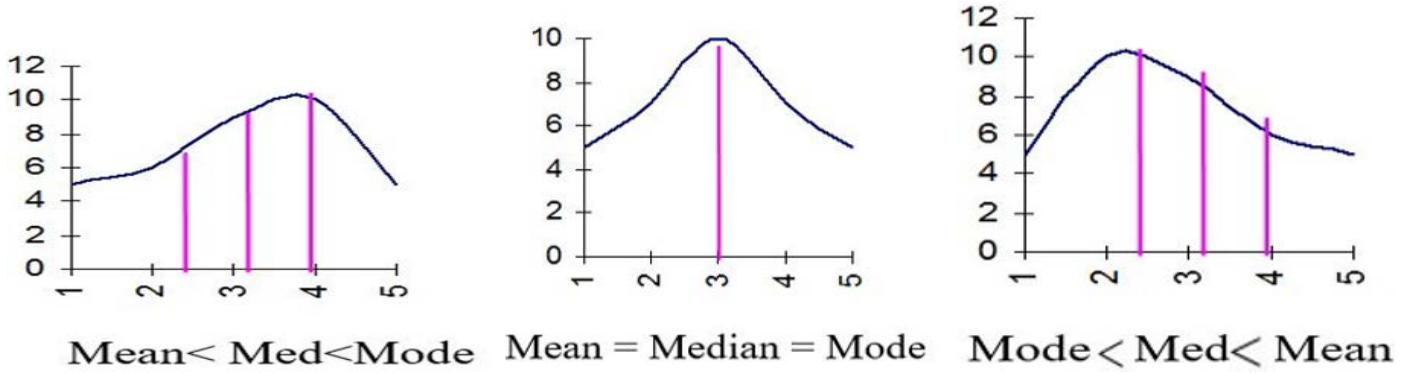
As you can see it is **positively skewed** ... in fact the tail continues way past $100,000

Distribution Index - skewness 3/n

- As executed on skewness(), the data distribution of variable **Holders** and **Claims** have have a **long tail** on right side since their sknewness are all greater than 1.
- Namely, the density of its data distribution have a long tail on right side.

```
> skewness(Insurance[,4:5])
 Holders   Claims
3.127833 2.877292
```

- Skewness can be quantified to define the extent to which a distribution differs from a normal distribution. In such a right-skewed distribution, usually (but not always) the mean is greater than the median, or equivalently, the mean is greater than the mode; in which case the skewness is greater than zero, and vice versa to left-skewed one.
- To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.



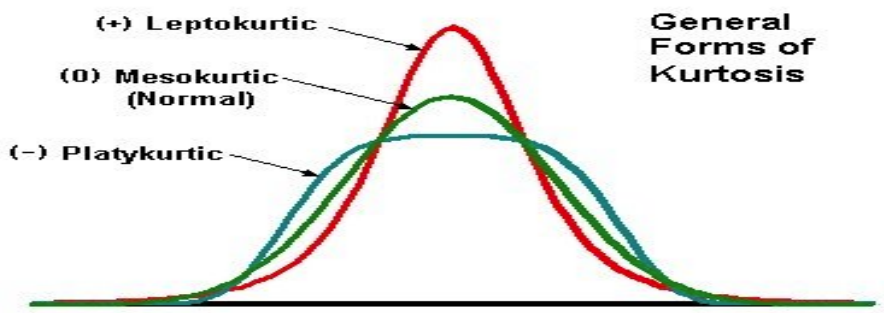
Distribution Index - skewness 4/n

- Skewness and symmetry become important when we discuss probability distributions.
- An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the **mean**; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.
- You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Distribution Index - Kurtosis 1/n

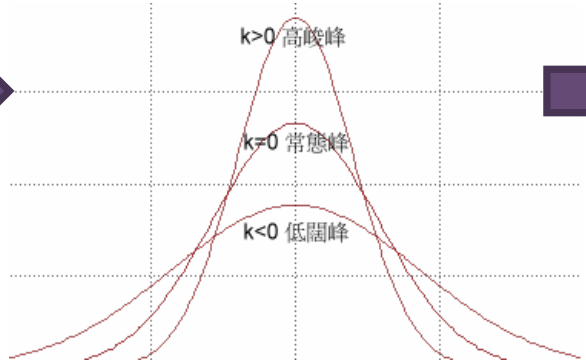
- Another one for data distribution index is kurtosis. It often works with skewness along.
- Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.
- The normal curve is called Mesokurtic curve (M. curve for short). If the curve of a distribution is peaked than a normal or Mesokurtic curve then it is referred to as a Leptokurtic curve (L. curve for short). If a curve is less peaked than a normal curve, it is called as a Platykurtic curve (P. curve for short). That's why kurtosis of normal distribution equal to 3.
- A Mesokurtic distribution is one in which the returns do not exhibit any behaviour that is different from one without kurtosis. This type of distribution has a coeffecient of kurtosis of 3 which is the same as that of a normal distribution. This distribution is zero kurtosis excess.
- $$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

這也被稱為超值峰度（excess kurtosis）。「減3」是為了讓正態分布的峰度為0。



$$K = \frac{\sum (x_i - \bar{x})^4}{ns^4} - 3$$

其中s是標準差，而n是樣本。



習題練習：

收集11位同學罰球投籃10次，投中次數分別為：

3 2 3 7 4 3 6 4 3 3 6

求其峰態係數？

解答：峰態係數為-1.198，圖形為低闊峰。

Distribution Index - Kurtosis 2/n

- Kurtosis is a measure of the combined sizes of the two tails. It measures the amount of probability in the tails. The value is often compared to the kurtosis of the normal distribution, which is equal to 3.
- Usually, It is to let the kurtosis of normal distribution to be 0, so -3 for being 0.
- The values for asymmetry and kurtosis between -2 and +2 are considered acceptable in order to prove normal univariate distribution(單變量分佈). However, it depends on mainly the sample size. Most software packages that compute the skewness and kurtosis, also compute their standard error.
- In the dataset we used, the kurtosis of variable Holders and Claims are greater 0, and even greater than +2. That means there are some abnormal data in the data set. For the abnormal data, you either find out the problem further, or find another algorithm instead not depends on data distribution.☺

```
> kurtosis(Insurance[,4:5])
  Holders    Claims
10.999610  9.377258
```

p.s. Based on the kurtosis calculation, suggest to find out the abnormal data in variable Holders first, then recreate the model. If the abnormal data is still existed, try to check out the abnormal data in variable Claims and rerun the model creation accordingly. If don't work via using the above methods, find another algorithm instead not depends on data distribution.

Distribution Index - Kurtosis 3/n

- A further characterization of the data includes skewness and kurtosis. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
- It measures the amount of probability in the tails. The value is often compared to the kurtosis of the normal distribution, which is equal to 3. If the kurtosis is greater than 3, then the dataset has heavier tails than a normal distribution (more in the tails).
- Like skewness, kurtosis is a statistical measure that is used to describe the distribution. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean)

Data Sparseness 1/n

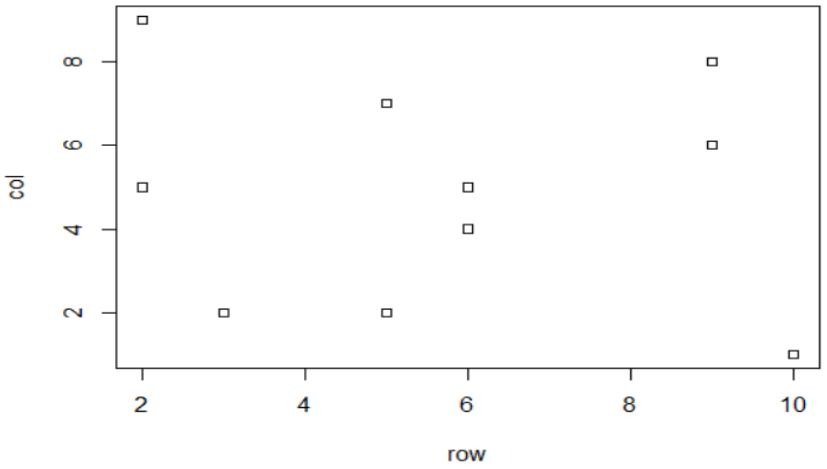
- In numerical analysis and scientific computing, a sparse matrix or sparse array is a matrix in which most of the elements are zero or null. By contrast, if most of the elements are nonzero and not null, then the matrix is considered dense.
- R uses the package Matrix to explore data sparseness, and provides functions to process dense matrix or sparse matrix.
- sample() takes a sample of the specified size from the elements of x using either with or without replacement. As illustrated, sample() function samples 10 times repeatedly and randomly (i.e. After sampling, it is replaced each time) in which the value is in 1 to 10 each time. For example, sample(1:6, 10, replace=TRUE). That means we roll a dice 10 time in which is 1 to 6 each time. And, the events we roll the dice are independent so the replace value is true. Namely, the number of population is the same after rolling the dice.

```
i=sample(1:10,10,replace=TRUE)
```

Data Sparseness

- sparseMatrix() builds a 10 x 10 matrix, and sets up 1 to the elements in which there are 10 elements filled in at most. This is a way to simulate and view your sparse data.
- which() filters out that the values in the matrix are 1. Then plot() function plots the sparse matrix.
- The argument **pch** is a symbol used in the plot() function as illustrated below. However, the symbols are dependent on the version of plot() you use. They are a little bit different.

```
install.packages("Matrix")
library(Matrix)
i=sample(1:10,10,replace=TRUE)
j=sample(1:10,10,replace=TRUE)
(A=sparseMatrix(i, j, x = 1))
loca=which(A==1, arr.ind=TRUE)
plot(loca,pch = 22)
```




```
 [1,] .  .  .  .  .  .  .  .  .
 [2,] .  .  .  .  1  .  .  .  1
 [3,] .  1  .  .  .  .  .  .  .
 [4,] .  .  .  .  .  .  .  .  .
 [5,] .  1  .  .  .  .  1  .  .
 [6,] .  .  .  1  1  .  .  .  .
 [7,] .  .  .  .  .  .  .  .  .
 [8,] .  .  .  .  .  .  .  .  .
 [9,] .  .  .  .  .  1  .  1  .
[10,] 1  .  .  .  .  .  .  .  .
```

Missing Values

- md.pattern() in the mice package is used to retrieve the missing-value patterns in a data set.

```
install.packages("mice")
library(mice)
md.pattern(Insurance)
```

- However, we will build a random-small data set (i.e. 64 x 5 elements) before using md.pattern(). We use the following loop to set up 10 NA (i.e. missing value) and store them on Insurance[ ]. Then, present and retrieve the missing values of Insurance[ ] by using md.pattern() .
- The “1” on the pattern table presents no missing value. “0” indicates a missing value.

```
# Missing values: build a random-small dataset
for(i in 1:10)
{ row=sample(1:64,1)
  col=sample(1:5,1)
  Insurance[row,col]=NA
}
```

```
> md.pattern(Insurance)
   Age Holders District Group Claims
54   1       1        1     1      1  0
5    1       1        1     1      0  1
2    1       1        1     0      1  1
2    1       1        0     1      1  1
1    1       0        1     1      1  1
     0       1        2     2      5 10
```

Visual Exploration

The End