Statistical Inference

An Introduction

Basic Terms as Prerequisite

- **Variable** (under study) – What you measure (ex. monthly salary of employees)
- **Population**- Set of all units in the study (all employees in the organization)
- **Sample**- Subset of units selected from population (ex. monthly salary of few selected employees in the organization)
- **Distribution**-How values of variable are distributed in the population (ex. normal distribution)
- **Factor**- Defines subgroups in the study.(ex. Gender, where gender wise salary distribution can be studied.)
- **Descriptive Statistics**- mean, median, standard deviation etc of the variable under study.. (ex. Average salary)

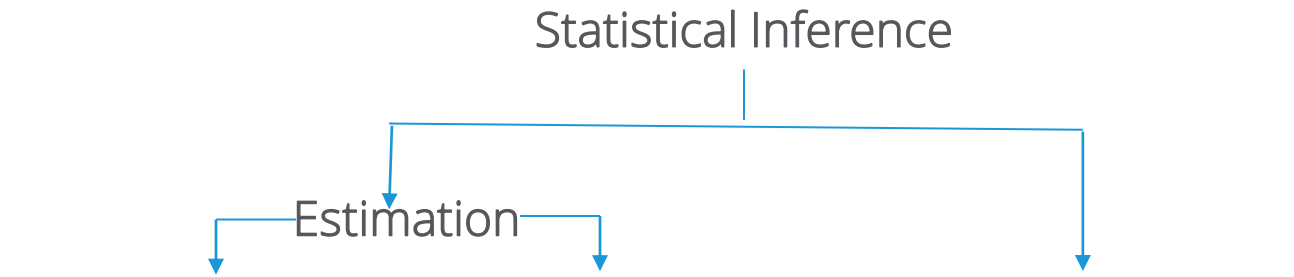
What is Statistical Inference ?

- Statistical inference is the process of drawing conclusion about unknown population properties, using a sample drawn from the population.

- These unknown population properties can be:
  - Mean
  - Proportion
  - Variance etc.

- Such unknown population properties are called as ‘Parameters’.

What is Statistical Inference ?



Point estimation

- Summarize the sample by a single value as an estimate of the population parameter.
- Ex. Average salary of junior data scientists is. 55,000 euros.

Interval estimation

- A range of values within which, we believe, the true population parameter lies with high probability.
- Ex. Average salary of junior data scientists is in the range of (52,000,55,000) With 95% confidence level.

Testing of Hypothesis

- To decide whether a statement regarding population parameter is true or false, based on sample data.
- Ex. Claim: Average salary of junior data scientists is greater than.50,000 euros annually.

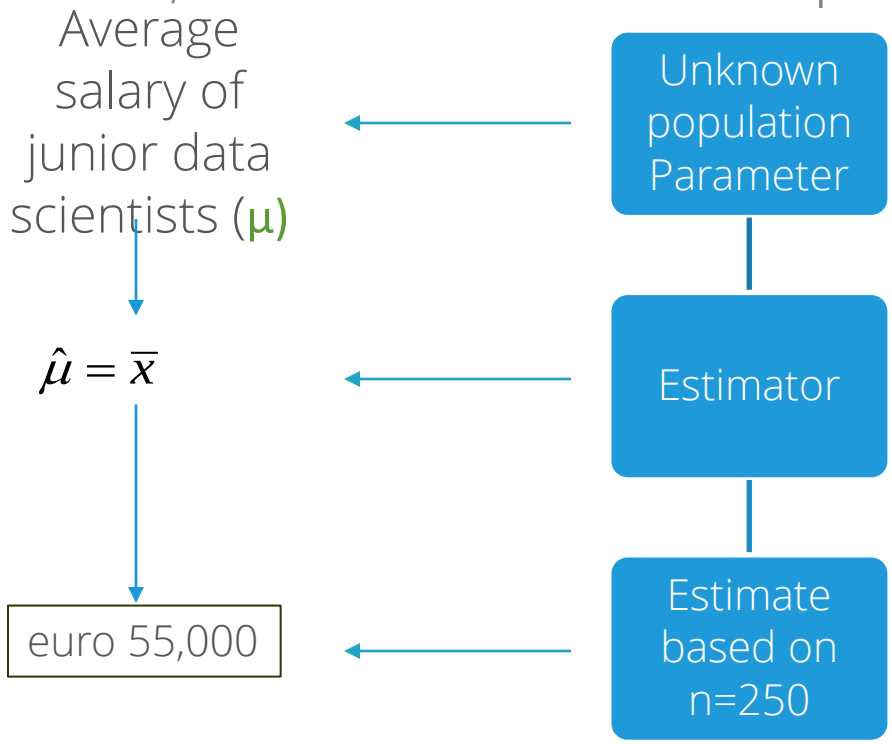
Parameter, Estimator, Estimate

- **Parameter**: Unknown property or characteristic of population
  - (population mean (μ), variance ($\boldsymbol{\sigma}^2$), proportion (P))
- **Estimator**: A rule or function based on sample observations which is used to estimate the parameter
  - (sample mean, sample variance, sample proportion)
- **Estimate:** A particular value computed by substituting the sample observations into an Estimator.

Parameter, Estimator, Estimate

- Research Question: What is the average salary of junior data scientists in Europe?

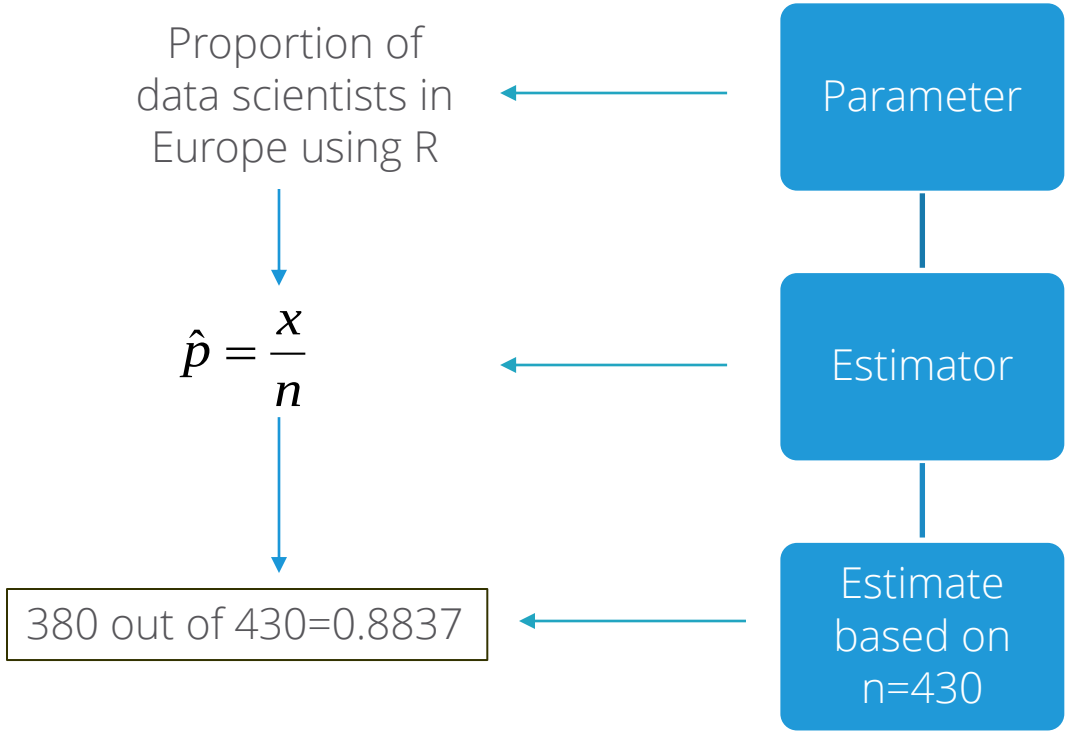
  - Average salary of junior data scientists in Europe is Population **Parameter.**
  - Sample of 250 junior data scientists is observed and Sample mean is computed.
  - Sample mean is used as **Estimator** of Population Mean.
  - Sample mean "55,000" which is calculated from sample of 250 is the **Estimate.**



Parameter, Estimator, Estimate

- Research Question: What is the proportion of data scientists in Europe who use R for data analysis?

- Proportion of data scientists in Europe who use R for data analysis is population parameter.
- Sample of 430 data scientists observed and proportion (or percentage) is calculated.
- Sample proportion is used as an estimator of population proportion.
- 380 out of 430 which is calculated from sample is **Estimate.**



Point Estimation vs. Interval Estimation

- In both the previous examples, (estimation of average salary of junior data scientists and proportion of data scientists using R) estimator is a single value estimating unknown population parameter.

- A confidence interval gives an estimated range of values which is likely to include an unknown population parameter with some probability, the estimated range being calculated from a given set of sample data.

- Generally, 95% or 90% Confidence Intervals are used.
- 95% confidence interval is a range estimate within which the true value of the parameter lies with probability 0.95.

Sampling distribution and Sampling error

- Research Question: What is the average salary of junior data scientists in Europe?

  - 50 samples, each of size 250 junior data scientists are observed and sample mean for each of theses 50 samples are computed. Here, sample mean will vary based on sample values.

- A probability distribution of all these means of the sample is called the **sampling distribution** of mean.

- **Standard error** is standard deviation of the these mean values.

Hypothesis Testing

- **Hypothesis**: An assertion about the distribution / parameter of the distribution of one or more random variables.
- **Null Hypothesis (H_0)**: An assertion which is generally believed to be true until researcher rejects it with evidence.
- **Alternative Hypothesis (H_1)**: A researcher's claim which contradicts null hypothesis.
- In simple words, testing of hypothesis is to decide whether a statement regarding population parameter is true or false, based on sample data.
- **Test Statistic**: The statistic on which decision rule of rejection of null hypothesis is defined.
- **Critical region or Rejection region**: the region, in which, if the value of test statistic falls, the null hypothesis is rejected.

Hypothesis Testing : Example

Objective	A consumer protection agency wants to test a Paint Manufacturer's claim, that average drying time of their new paint is less than 20 minutes.

- Sample: n=36 boards were painted from 36 different cans and the drying time was observed.
- Estimator of mean drying time is sample mean $\hat{\mu} = \bar{x}$

Null Hypothesis (H_0): $\mu = 20$

Alternate Hypothesis (H_1): $\mu < 20$

Test Statistic	In this case the test statistic is based or $\bar{x}$
Decision Criteria	Reject null hypothesis if test statistic based on sample mean is less than critical value.

Two types of error

- While testing the hypothesis using any decision rule, one of the following scenario might occur.

Decision	Reality	
	Ho is true	Ho is false
Reject Ho	Type I error	Correct
Do Not Reject Ho	Correct	Type II error

- **For example**, in legal system,
  H_0: person is not guilty H_1: person is guilty

Decision	Reality	
	Not Guilty	Guilty
Guilty	Type I Error -- Innocent person goes to jail	Correct
Not Guilty	Correct	Type II error Guilty person is set free

Two Types of error

- **Level of significance (LOS):** Probability of Type I error is called as 'Level of Significance (α)' generally set as 5% (los=0.05) and null hypothesis is rejected if observed risk(p value) is less than 0.05
- **α =** Probability [Type I Error] = Probability [Reject H_0 | H_0 is True]
- **β =** Probability [Type II Error] = Probability [Do not reject H_0 | H_0 is not True]
- **Power of the test** is given by: (1 – β)

One tailed and two tailed tests

- Hypothesis test where the alternative hypothesis is one-tailed (right-tailed or left-tailed), is called a **one-tailed test**.

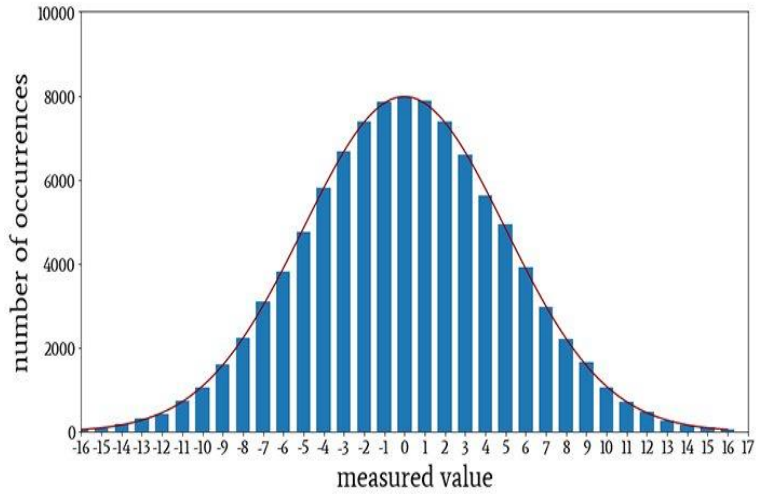
  $H_0: \mu = \mu_0$
  $H_1: \mu > \mu_0$ (Right-tailed) or $H_1: \mu < \mu_0$ (left-tailed)

- Hypothesis test where the alternative hypothesis is two-tailed is called **two-tailed test**.

  $H_0: \mu = \mu_0$
  $H_1: \mu \neq \mu_0$

Normal Distribution Assumption

- Is my data coming from Normal Distribution?
- This is a question you will ask yourself many times in data science project.
- If the answer is YES then many statistical methods/models become more reliable.



- If the answer is NO then you will ask next questic
  How do I transform it into Normal Distribution ?
  Which is alternative method if Normality can not be achieved ?

But how do you check or confirm the assumption of Normality?

Normality **Assessment**

- An assessment of the normality of data is a prerequisite for many statistical tests because normal distribution is an underlying assumption in parametric testing.

- Normality can be assessed using two approaches: graphical and numerical.
  - Graphical approach

    - Box-Whisker plot (It is used to asses symmetry rather than normality.)
    Quantile-Quantile plot (Q-Q plot).

  - Numerical (Statistical) approach
    Shapiro-Wilk test
    Kolmogorov-Smirnov test

* Box-Whisker plot is used to asses symmetry rather than normality. Hence, only Q-Q plot method is explained.

Case Study - 1

To assess normality of data in R, we shall consider the below case as an example.

Background

Data has 2 variables recorded for 80 guests in a large hotel.
Customer Satisfaction Index (csi) & Total Bill Amount. (billamt)

Objective

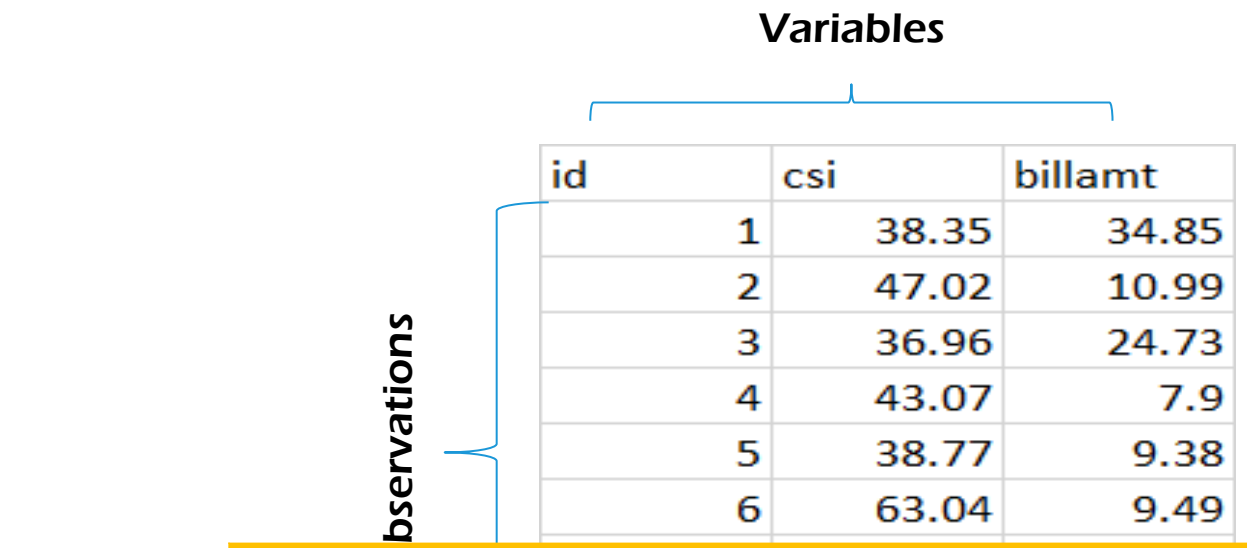
To check if variables follow normal distribution

Sample Size

Sample size: 80
Variables: id, csi, billamt

Data Snapshot

Normality Testing Data



Column	Description	Type	Measurement	Possible Values
id	Customer ID	Numeric		
csi	Customer Satisfaction Index	Numeric		Positive value
billamt	Total Bill Amount .	Numeric		Positive value

Quantile-Quantile plot

- Very powerful graphical method of assessing Normality.
- Quantiles are calculated using sample data and plotted against expected quantiles under Normal distribution.
- If Normality assumption is valid then high correlation is expected between sample quantiles and expected(theoretical quantiles under normal distribution) quantiles.
- The Y axis plots the actual quantiles values based on sample.
  The X axis plots theoretical values.
- If the data is truly sampled from a Normal distribution, the QQ plot will be linear.

Q-Q plot using R

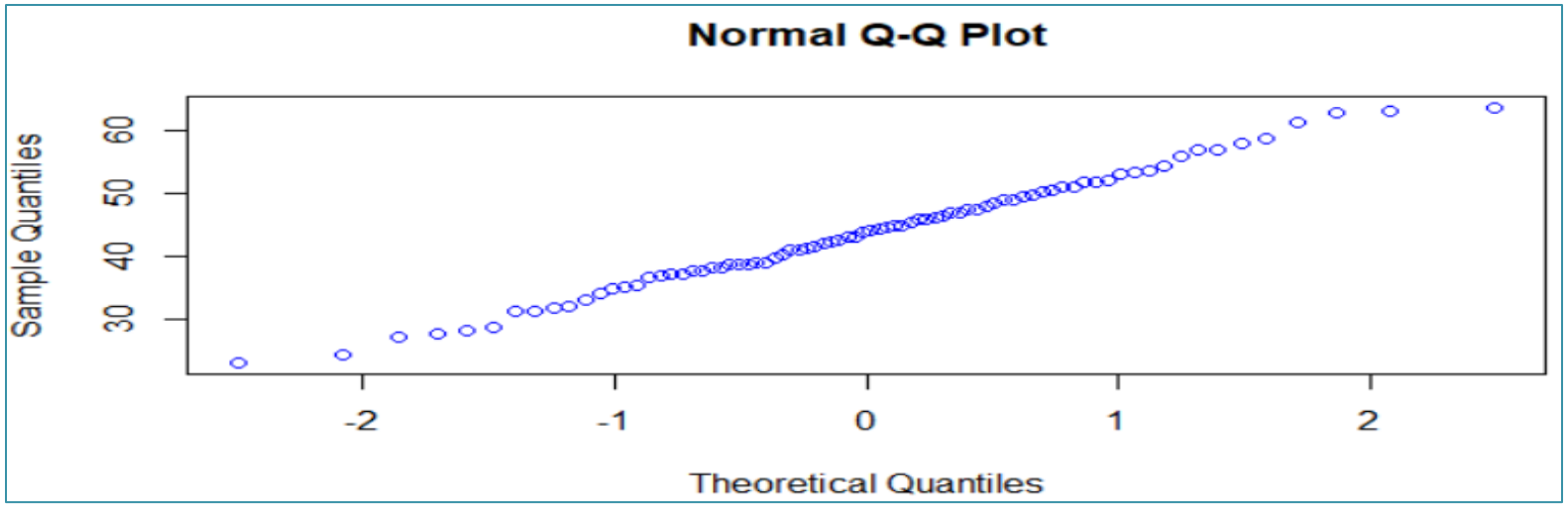
```
# Import data
data<-read.csv("Normality Testing Data.csv", header=TRUE)

# Q-Q plot for the variable csi
qqnorm(data$csi,col="blue")
```

- ***data$csi*** *is the variable for which normality is to be checked.*
- ***Col=blue*** *specifies the line color on graph.*

Q-Q plot using R

```
# Output:
```



Interpretation :

- *Q-Q plot is Linear. Distribution of 'csi' can be assumed to be normal.*

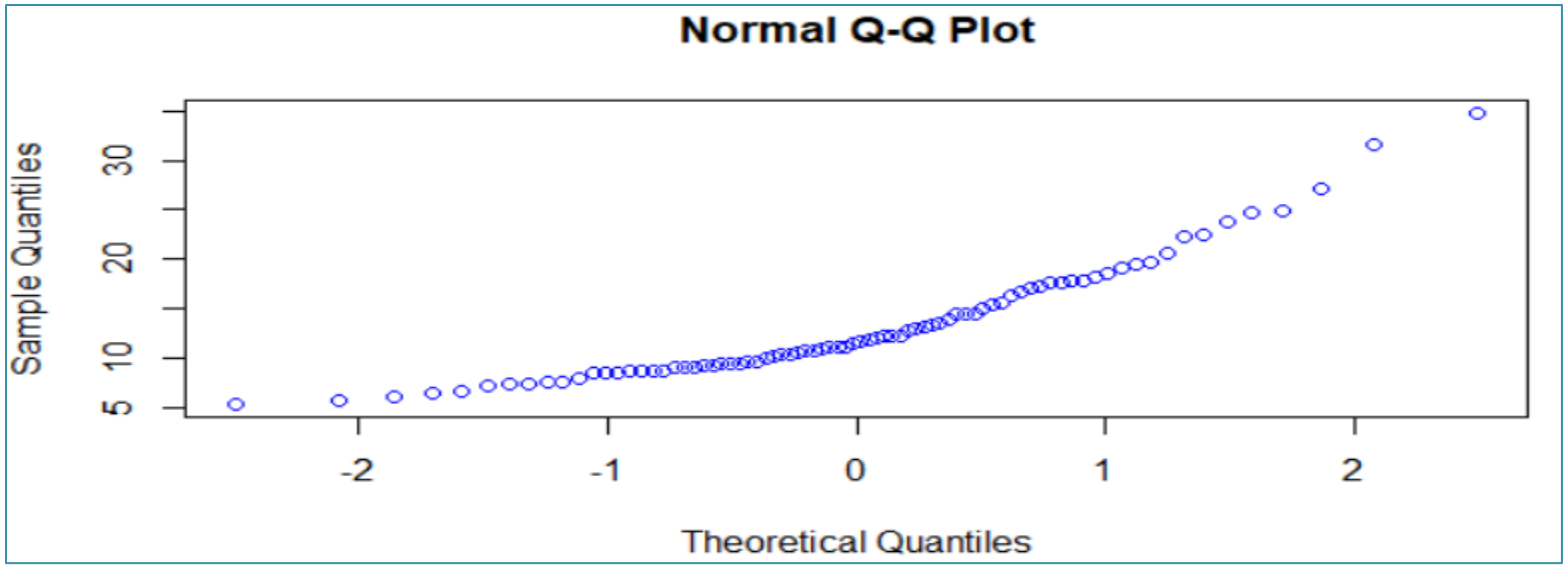
Q-Q plot using R

```
# Q-Q plot for the variable billamt
qqnorm(data$billamt,col="blue")
```

- ***data$billamt*** *is the variable for which normality is to be checked.*
- ***Col=blue*** *specifies the line color on graph.*

Q-Q plot using R

```
# Output:
```



Interpretation :

- *Q-Q plot is deviated from linearity. Distribution of 'billamt' appears to be non-normal.*

Shapiro-Wilk test

Shapiro-Wilk test is widely used statistical test for assessing **Normality**.

Objective	To test the **normality** of the data.

Null Hypothesis (H_0): **Sample is drawn from Normal Population**

Alternate Hypothesis (H_1): **Sample is drawn from Non-Normal Population**

The test is performed for the variables, 'csi' and 'billamt' separately.

Test Statistic	It correlates sample ordered values with expected Normal scores. (actual calculation is very complex so we will avoid details)
Decision Criteria	Reject the null hypothesis **if p-value < 0.05**

* Here we continue to use previous data of CSI and Bill Amount for Shapiro-Wilk test.

Shapiro-Wilk test in R

```
# Shapiro Wilk test for the variable csi
shapiro.test(data$csi)
```

- ***data$csi*** *is the variable for which normality is to be checked.*

```
# Output:
        Shapiro-Wilk normality test

data:  data$csi
W = 0.99196, p-value = 0.9038
```

Interpretation :

- *Since p-value is >0.05,do not reject H0. Distribution of 'csi' can be assumed to be normal.*

Shapiro-Wilk test in R

```
# Shapiro Wilk test for the variable billamt

shapiro.test(data$billamt)
```

- ***data$billamt*** *is the variable for which normality is to be checked.*

```
# Output:

        Shapiro-Wilk normality test

data:  data$billamt
W = 0.89031, p-value = 4.858e-06
```

Interpretation :

- *Since p-value is <0.05, reject H0. Distribution of 'billamt' appears to be non-normal.*

Standard Parametric Tests

- **One Sample t test**
- **Independent Sample t test**
- Paired t test
- F test for two variances
- Analysis of Variance- One way
- Analysis of Variance- Two way

NOTE: Above tests assume that the distribution of variable under study is “Normal”
As an alternative, there exists “Non-Parametric Tests”

More about this during next session

Thank You