Statistical Inference

Recap and One Sample t test

DATA SCIENCE
INSTITUTE

What is Statistical Inference ?

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

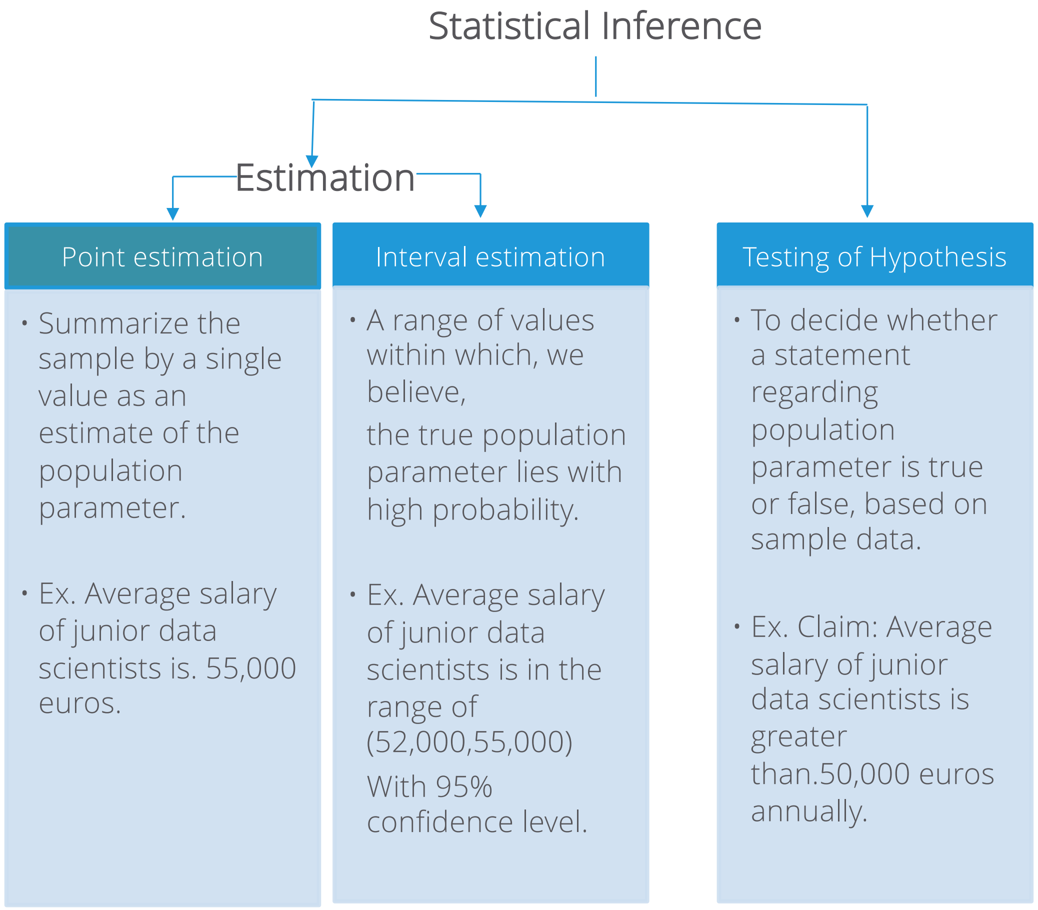
- These unknown population properties can be:
  - Mean
  - Proportion
  - Variance etc.

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

Sample Results
Inference regarding Population

DATA SCIENCE INSTITUTE

What is Statistical Inference ?



Statistical Inference

Estimation

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

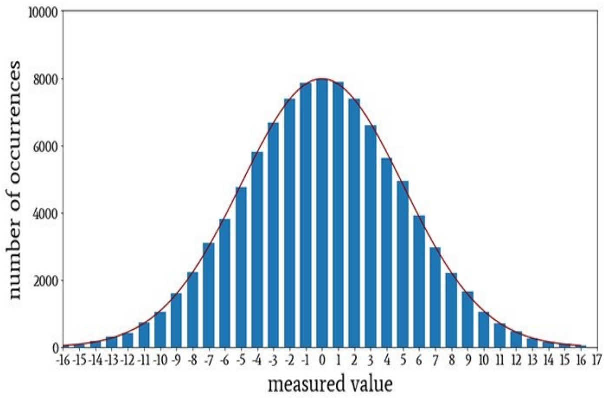
- To decide whether a statement regarding population parameter is true or false, based on sample data.
- Ex. Claim: Average salary of junior data scientists is greater than.50,000 euros annually.

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.

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.

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”

STATISTICAL INFERENCE
PARAMETRIC TESTS

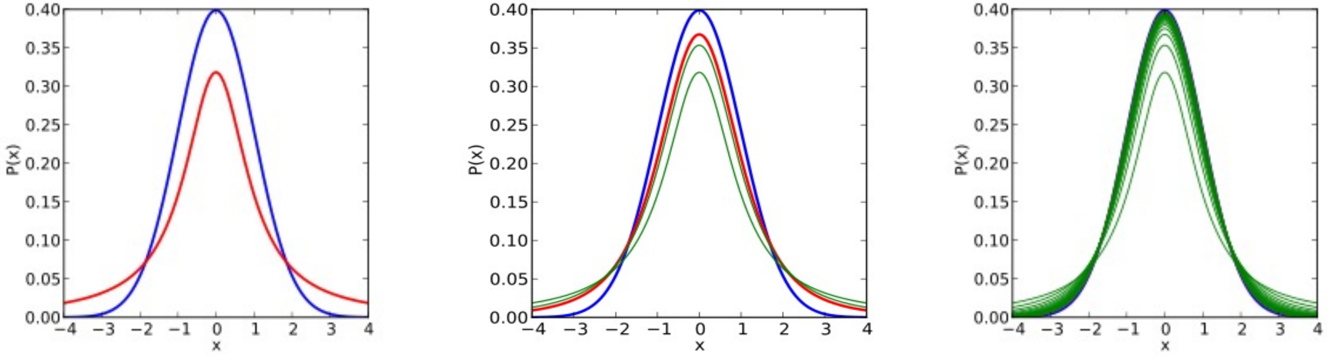
Tests of Significance based on t distribution

t distribution

The t distribution is symmetric and its overall shape resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider.

As the number of degrees of freedom grows, the *t*-distribution approaches the normal distribution with mean 0 and variance 1.

Blue- Normal



One Sample t test (for mean)

- One sample t test is used to test the hypothesis about a single population mean.

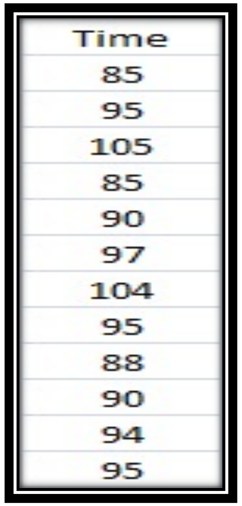
- We use one-sample *t*-test when we collect data on a single sample drawn from a defined population.

- In this design, we have one group of subjects, collect data on these subjects and compare our sample statistic to the hypothesized value of population parameter.

- Subjects in the study can be patients, customers, retail stores etc.

One Sample t test Example

- A large company is concerned about time taken by employees to complete weekly MIS report. The general manager claims that 'average time taken to complete the MIS report is more than 90 minutes'.
- Time taken to complete MIS report is observed for 12 randomly selected employees.



Time
85
95
105
85
90
97
104
95
88
90
94
95

Null Hypothesis H0: $\mu=\mu0$

Alternative Hypothesis H1: $\mu > \mu0$

Here μ**0=90 (test value)**

Assumptions for one sample t test

- The assumptions of the one-sample *t*-test are listed below:

1. Random sampling from a defined population
2. Population is normally distributed

The validity of the test is not seriously affected by moderate deviations from 'Normality' assumption.

Test Statistic

- The test statistic for one sample t test is given below:

$$\boxed{t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}}$$

- Where $\bar{x}$ is the sample mean, *s* is the sample standard deviation,n is the sample size and μ_0 is the test value.

- The quantity t follows a distribution called as 't distribution' with n-1 degrees of freedom.

Drawing Inference

	Notation	Value
Sample size	n	12
Mean	$\bar{x}$	93.5833
S.D.	S	6.4731
Standard Error	$S/\sqrt{n}$	1.8686
Difference	$\bar{x}-\mu_0$	3.5833
t	$(\bar{x}-\mu_0)/S.E.$	1.9176

Drawing Inference(continued)

- Compare p value with level of significance (typically 0.05 or 5%)

- Most of the softwares provide p value.

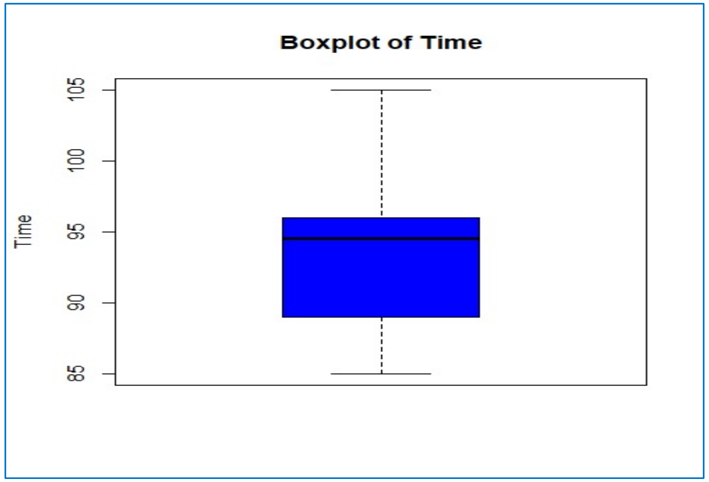
  If p value is less than 0.05 (Level of Significance) then Reject H0

  (Here p value is 0.0407)

Inference: Reject H0 and conclude that time taken to complete MIS report is significantly more than 90 minutes.

One Sample t test using R

```
#Import csv data set  ONE SAMPLE t TEST
mistime<-read.csv(file.choose(),header=T)
# Boxplot of  variable “Time”
boxplot(mistime$Time,col="blue",ylab = "Time", main = "Boxplot of Time")
```



One Sample t test using R

```r
# Use t.test by specifying variable, alternate hypothesis and μ0

t.test(mistime$Time, alternative="greater", mu=90)

#mistime$Time is the variable under study
#alternative=“greater” and mu=90 specify:
#Null Hypothesis:       H0: μ=90
#Alternative Hypothesis: H1: μ > 90,
```

R-Output for One Sample t test

```
One Sample t-test
    data:  mistime$Time
    t = 1.9176, df = 11, p-value = 0.04074
    alternative hypothesis: true mean is greater than 90
    95 percent confidence interval:
    90.22748      Inf
    sample estimates:
    mean of x
    93.58333
```

→ **Reject H_0**

t is the calculated value of the test statistic and p-value is used to make inference about acceptance or rejection of null hypothesis (if p-value < 0.05 reject H_0)

Inference: Reject H0 and conclude that time taken to complete MIS report is significantly more than 90 minutes.