Data Science with R

Day 3 - Statistics for Data Science – Basics and Advanced

TACT 360° Training

LEARN. ACHIEVE. STANDOUT

www. tactlearn.com

Today's Agenda

✓ Basics of Statistics

- Type of Random Variables - Based on Scale of Measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Variance
- Standard Deviation

✓ Advanced Statistics P1

- Normal Distribution
- Standard Normal Distribution and Z- Score
- Binomial Distribution
- Poisson Distribution

Basics of Statistics

Type of Random Variable - Based on Scale of Measurement

Based on scale of measurement RV can belong to one of the following types -

NOMINAL

- No Order
- No Comparison
- No Calculation
- No Interval

Ex.
Gender {M,F}

ORDINAL

- Order
- Comparison
- No Calculation
- No Interval

Ex.
Size {S<M<L}

INTERVAL

- Order
- Comparison
- Calculation
- Regular Interval
- No Absolute Zero
- Cannot calculate ratio

Ex.
Temp.{0 C= 32F}, IQ

RATIO

- Order
- Comparison
- Calculation
- Regular Interval
- Absolute Zero
- Can calculate Ratio

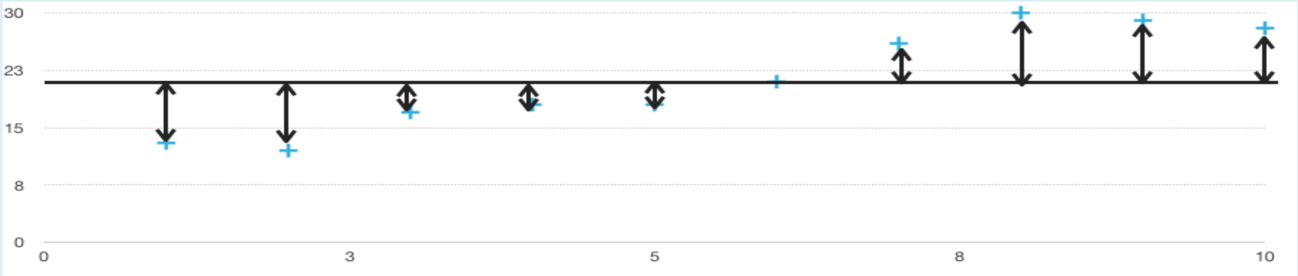
Ex.
Height, distance

Hemant Rathore

Basics of Statistics

Variance(σ^2) and Standard Deviation(σ) –

Variance (σ^2) - Squared deviation of value from Mean : Var(X) = 1/n [Sum $[X-Mu]^2$]
Standard Deviation (σ) - Square Root of Variance : √ 1/n [Sum $[X-Mu]^2$]



Data Point	Speed	X-Mu	$(X-Mu)^2$
1	13	-8	64
2	12	-9	81
3	17	-4	16
4	18	-3	9
5	18	-3	9
6	21	0	0
7	26	5	25
8	30	9	81
9	29	8	64
10	28	7	49
Average	21		398

Variance = 398/10 = 39.8

SD = 6.3

Hemant Rathore

Basics of Statistics

Mean(μ), Variance(σ²) and Standard Deviation(σ) -

Discrete random variable :

Mean (μ) -

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i.$$

Variance (σ²) -

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2,$$

Standard Deviation (σ) -

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$

$$\sigma = \sqrt{\sum_{i=1}^{N} p_i (x_i - \mu)^2},$$

https://en.wikipedia.org

Hemant Rathore

Basics of Statistics

Variance(σ^2) and Standard Deviation(σ) -

Continues random variable -

Mean (μ) -

$$\mu = \int x\, f(x)\, dx$$

Variance (σ^2) -

$$\operatorname{Var}(X) = \sigma^2 = \int (x-\mu)^2\, f(x)\, dx$$

Standard Deviation (σ) -

$$\sigma = \sqrt{\int_{\mathbf{X}} (x-\mu)^2\, p(x)\, \mathrm{d}x},$$

https://en.wikipedia.org/wiki/Mode_(statistics)

Hemant Rathore

Advanced Statistics

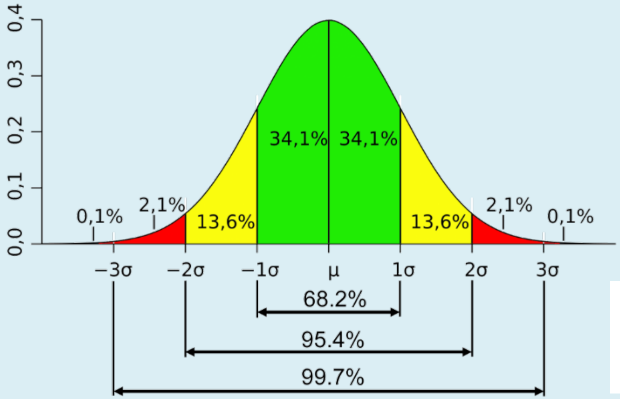
Normal Distribution

Normal or Gaussian or bell shaped curve distribution is a very common continuous probability distribution. Normal Distribution has bell shaped curve, it's a symmetric single model distribution with highest density at and around the mean :

Ex. Age, Marks

Some Important properties :

- Mean = Median = Mode
- Area within 1 Std. Dev around the mean ~ 68.3 %
- Area within 2 Std. Dev around the mean ~ 95.4 %
- Area within 3 Std. Dev around the mean ~ 99.7 %



$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

https://en.wikipedia.org/wiki/Normal_distribution

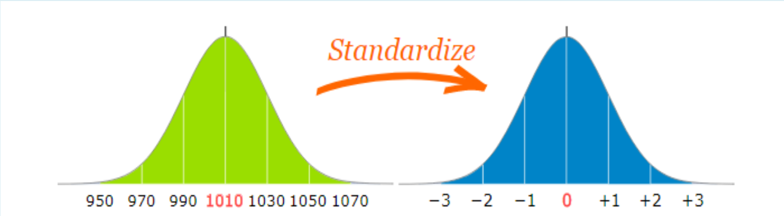
Advanced Statistics

Standard Normal Distribution and z Score/ z Statistic

Special case of Normal distribution with Mean = 0 and Variance = 1, Std. Deviation = 1. **it has total area under the curve = 1** which represents probability.

Any Normal distribution can be converted into Standard Normal Distribution by applying following transformation :

Z = (x - µ) / σ {This is called as **z Score**, it tells us how many SD far we are from mean in SND}



Why do we need Standard Normal Distribution?

To easily make the decision about Probability distribution by using some known properties of Standard Normal Distribution and z table.

https://www.mathsisfun.com/data/standard-normal-distribution.html

Advanced Statistics

Binomial Distribution

Bi→ 2, Nomial→ Nominal → Only 2 possible outcomes (Success or Failure)

When we perform any given experiment multiple times and we are interested in knowing #successes, this type of experiments are known as Binomial experiments, Ex. Flipping the coins multiple times. Using Binomial Distribution we can answer probability related questions for any Binomial experiments.

Probability of getting ‘x’ #Successes out of ‘n’ trials using Binomial Distribution –

$$P(x) = {}^{n}c_{x}\, P^{x}\ (1-P)^{n-x}$$; **P = Probability of Success in 1 trial**

Some Important properties :

- N Fixed Number of Trials
- Only 2 Possible Exclusive Outcomes
- Probability of success remain same during the experiment
- All the trials are independent

Hemant Rathore

Advanced Statistics

Poisson Distribution

When we analyze the probability of occurrence of any event during some specified interval of time or according to some other binding conditions.

Probability of ‘x’ occurrence using Poisson Distribution –

$$P(x) = (\lambda^x e^{-\lambda})/x! \; ; \; \lambda = \text{Mean/Expected \#Occurrence}$$

Some Important properties :

- All the occurrences are independent
- Expected #Occurrence doesn’t change over the period of time

Hemant Rathore

“Qs & As”

Data Science with R

Day 4 - Statistics for Data Science – Basics and Advanced

TACT 360⁰ Training

LEARN. ACHIEVE. STANDOUT

www. tactlearn.com

Today's Agenda

✓ Inferential Statistics

- Sampling
- Inferential Statistics
- Sampling Distribution
- Central Limit Theorem
- Central Limit Theorem Exercise

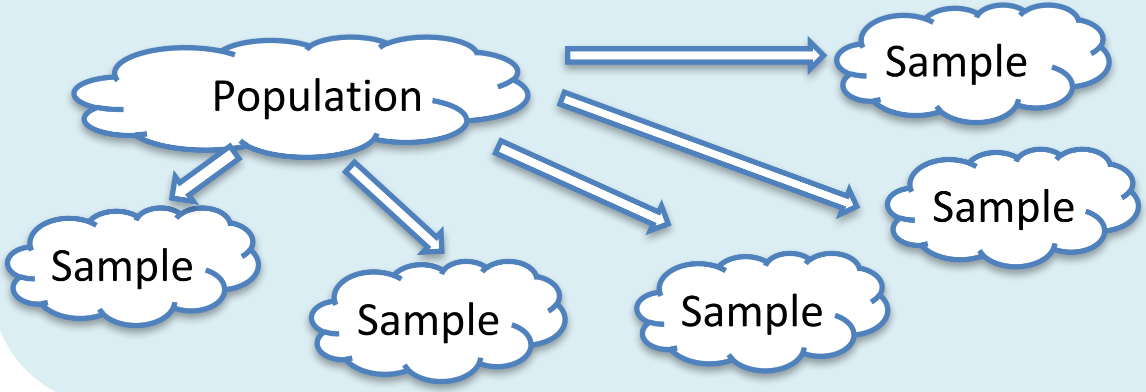
✓ Hypothesis Testing

- Hypothesis and hypothesis Testing
- One tail/Two tail test
- Type I and Type II Errors
- Hypothesis Testing using z test
- Hypothesis Testing t test

Advanced Statistics - Inferential Statistics

Sampling

Sampling is taking random samples from over all population, sampling is done in order to make some judgements about overall population because many a time it is not possible or practical to analyze the overall population and instead we can get approximately same results even using sampling with sufficient sample size

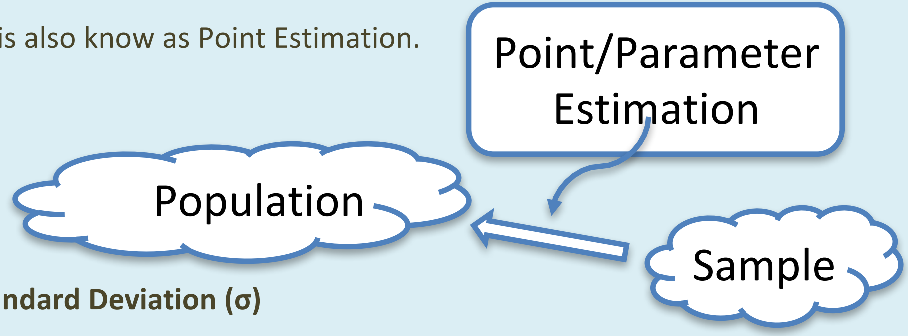


Hemant Rathore

Advanced Statistics - Inferential Statistics

Inferential Statistics

With inferential statistics, we try to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the overall population might think. Or, we use inferential statistics to make judgments of the probability of the overall population. This is also know as Point Estimation.



Point Estimators

- **Sampling Mean (μ XBar) -> Population Mean (μ)**
- **Sampling Standard Deviation (σ XBar) –> Population Standard Deviation (σ)**

Hemant Rathore

Advanced Statistics - Inferential Statistics

Sampling Distribution

When we use the distribution of samples taken randomly from population to make judgment about the overall population. Different Samples taken from same population can show different characteristics this is know as sampling variability. Larger the sample size – less the variability.

Expected Value E(x) or Sampling Mean (μ XBar)-

We take multiple samples from overall population and analyze the distribution of these samples to make the decision about overall population. The mean of these samples is known as expected value or sampling mean and it can be considered as Overall population mean (μ).

Sample Size (n)-> Very large than Expected Value E (x) -> μ

Standard Error of the Mean (σ XBar)-

Standard deviation of sampling distribution is know as Standard Error of the Mean.

Sample Size (n)-> Very large than Standard Error of the mean -> 0

Hemant Rathore

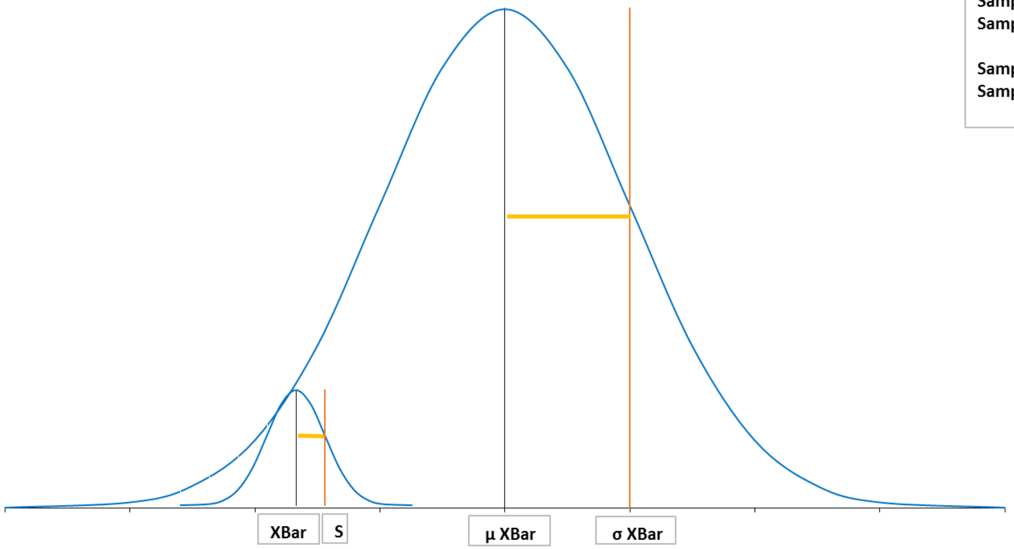
Advanced Statistics - Inferential Statistics

Sampling Distribution

Population Mean - µ
Population SD - σ

Sample Mean - XBar
Sample SD - S

Sampling Mean - µ XBar
Sampling SD - σ XBar



Hemant Rathore

Advanced Statistics - Inferential Statistics

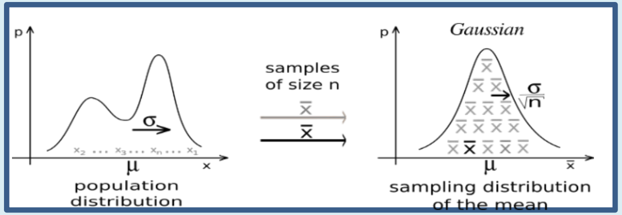
Central Limit Theorem (CLT)

"Sample Mean will be approximately normally distributed for larger sample size regardless of the original distribution from which we are taking samples."

With Mean = Population Mean (μ)

SD = σ /√n $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

{in case σ is not known then SD = s/ √n, s = Sample SD}



Application of CLT

So we can use standard normal distribution concepts for any non normal population by taking samples because as per CLT Samples will be normally distributed for large sample size.

From CLT we know, sampling SD **σ_x = σ /√n**

From Standard Normal distribution we know – **Z = (x - μ) / σ**

So for any sampling distribution we can say – **Z = (X - μ) / (σ /√n),** so now we can calculate the probability using SND for any Non normal Population.

Hemant Rathore

Advanced Statistics - Inferential Statistics

Exercise - Central Limit Theorem

A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean μ = 205 pounds and standard deviation σ = 15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

Solution – Given : μ = 205 , σ = 15 , n=49; Average total weight = 49*205 = 10045 > 9800

We know nothing about the original probability distribution weather its normal or not but from CLT we know sample mean will be normally distributed,

We are interested in the weight of 49 boxes not 1 so lets calculate : μ and σ for 49 boxes :

μ = 10045 , σ = 15*49= 735, now we need to know the probability that total weight would be <=9800 so X=9800

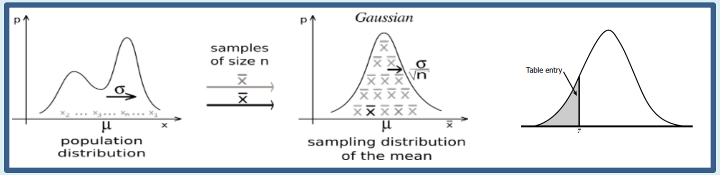
Z = (X - μ) / σ/ √n

(9800- 10045) / (735/7)

-245/105

Z = **-2.33**

let use z table to get the probability for z <= -2.33 → 0.0099



Advanced Statistics – Hypothesis Testing

Hypothesis

A hypothesis (plural hypotheses) is a proposed explanation for a phenomenon. In Statistics Hypothesis can be any theory about the data that we want to validate (generally accept or reject) – we will be mainly working of two type of hypotheses :

1. **Null Hypothesis (H0)** – Current Assumption or Theory which is currently assumed to be correct
2. **Alternative Hypothesis (H1)** – Claim or theory that we want to prove

Ex. H0: While flipping a coin the probability of getting head is 0.5; H1 : Probability of getting head is less than 0.5

Hemant Rathore

TACT 360° Training
LEARN. ACHIEVE. STANDOUT

Advanced Statistics – Hypothesis Testing

Hypothesis Testing

Validating the null hypothesis (H0) against some Alternative Hypothesis (H1) based on some given **sample data** , Steps involved in Hypothesis Testing using P values –

1. Define your Null and Alternative hypothesis, H0 & H1
2. Decide the type of test (One tail or two tail)
3. Define level of significance **α** , generally assumed to be 0.05 or 0.01
4. Find the Test Statistics TS (t Test or z test)
5. Find P Value
6. Reject the null hypothesis or you may accept the alternative hypothesis if **P < α**

- **P value – Probability of getting the given sample or even more extreme samples**
- **Significance level (α) – Minimum acceptable P Value/ border line**

Two tail test –

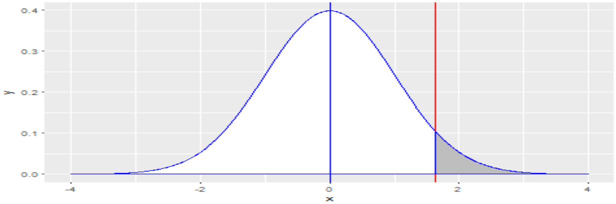
H0 : μ = 200

H1 : μ != 200

One tail Test –

H0 : μ <= 200

H1 : μ >200



One tail Test –

H0 : μ >= 200

H1 : μ <200

Hemant Rathore

Advanced Statistics – Hypothesis Testing

Type I Error or False Positive

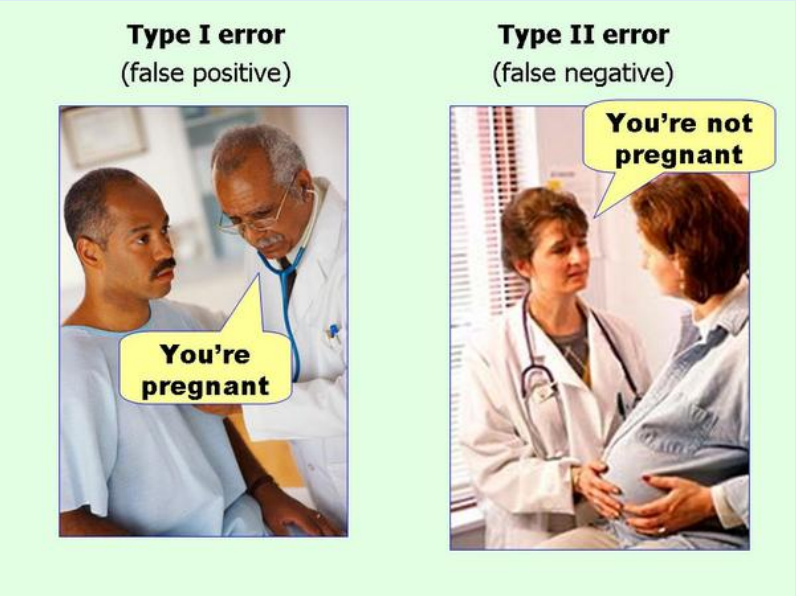
Getting Positive result when it should be Negative in reality.

Rejecting null hypothesis (H0) while H0 is correct and should not be rejected, Probability of Type I error is known as Alpha Risk.

Type II Error or False Negative

Getting Negative result when it should be Positive in reality.

Failing to Reject null hypothesis (H0) while H0 is not correct and should be rejected, Probability of Type II error is known as Beta Risk.



Hemant Rathore

TACT LEARN. ACHIEVE. STANDOUT

Advanced Statistics – Hypothesis Testing

Z-Test for Hypothesis Testing

Z-Test is used to perform hypothesis testing when we know population Standard Deviation (**σ**) or the sample size $n > 30$, Steps for 1 sample z test –

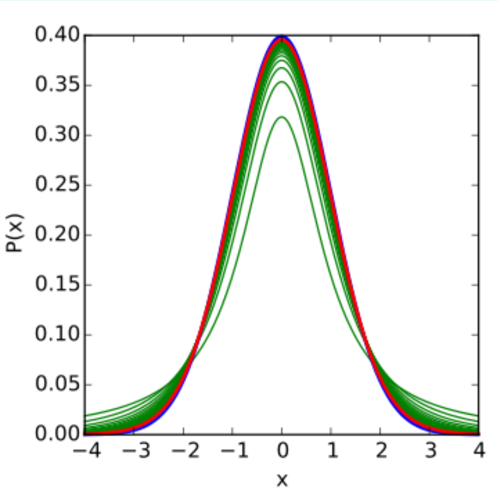
1. Define your Null and Alternative hypothesis, H0 & H1
2. Define level of significance **α** , generally assumed to be 0.05 or 0.01
3. Find the Test Statistics using **TS = (X- μ)/ (σ/ √n) or TS = (X- μ)/ (s/ √n)** {when σ is not known but n>=30}
4. Find P Value using Z Table and TS, if it's a two sided test then double P value
5. Reject the null hypothesis or you may accept the alternative hypothesis if **P < α**

Advanced Statistics – Hypothesis Testing

T-Test for Hypothesis Testing

T-Test is used to perform hypothesis testing when we don't know population Standard Deviation (**σ**) and sample size n < 30, Steps for 1 sample t test –

1. Define your Null and Alternative hypothesis, H0 & H1
2. Define level of significance **α** , generally assumed to be 0.05 or 0.01
3. Calculate Degree of Freedom DF = **n-1**
4. Find the Test Statistics using **TS = (X- μ)/ (s/ √n)**
5. Find P Value or P Range using T Table for calculated TS and DF
6. Reject the null hypothesis or you may accept the alternative hypothesis if **P < α**



Hemant Rathore

“Qs & As”

For more information or to set up an appointment, please contact us today.

jointact@tactlearn.com