Statistics in Data science

Data science involve the analysis and interpretation of complex datasets to extract valuable insights and support decision-making. Statistics plays a crucial role in data science, providing the foundation for various methods and techniques used in the field.

statistical concepts and techniques in data science

1. **Descriptive Statistics (normal distribution)**
   task 4

2. **Inferential Statistics** task 5

1- **Descriptive Statistics**

Descriptive Statistics -> Descriptive statistics are a set of techniques used to summarize and describe the main features of a dataset .

Descriptive statistics divided into :

1- central tendency -> **Mean, Median, Mode**

2- measures of dispersion" variability" -> **Range, Variance, Standard Deviation**

1. central tendency -> aim to identify a representative or central value around which the data points cluster. They provide a single value that summarizes the central location of the data.

    1. Mean -> the sum of all values divided by the number of observations. It represents the central point of a dataset.
    2. Median -> The middle value of a dataset when arranged in ascending or descending order. It is less sensitive to extreme values than the mean.
    3. Mode -> The value or values that appear most frequently in a dataset.

2. measures of dispersion” variability”=>quantify the spread, variability, or extent to which data points deviate from the central tendency. They provide information about how "spread out" the values are.

1. **Range ->** The difference between the maximum and minimum values in a dataset.

2. **Variance ->** A measure of how spread out the values in a dataset are from the mean “the unit is the square of original unit => cm^2 “ so the variance difficult to interpret .

3. **Standard Deviation ->** The square root of the variance. It provides a more … ic mean“ .

Population

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

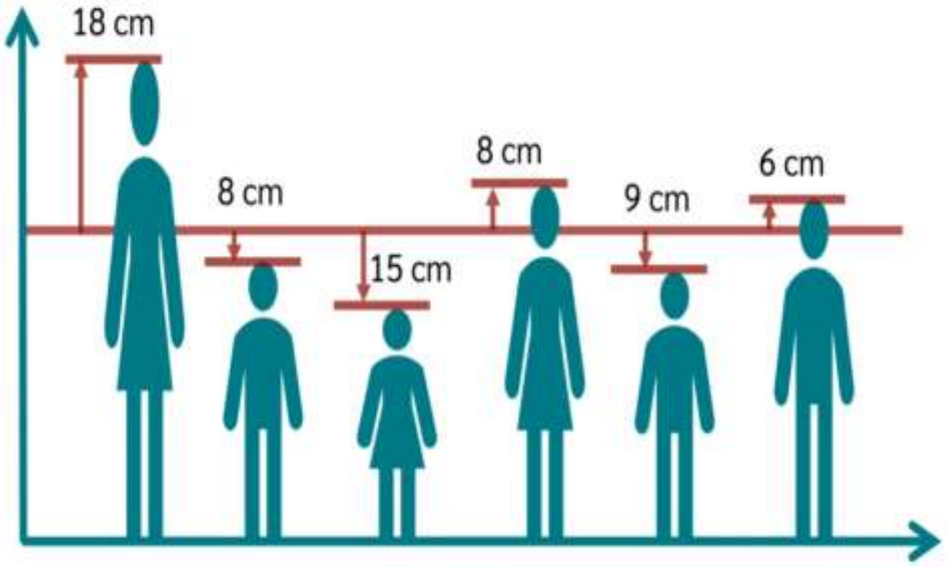
Sample

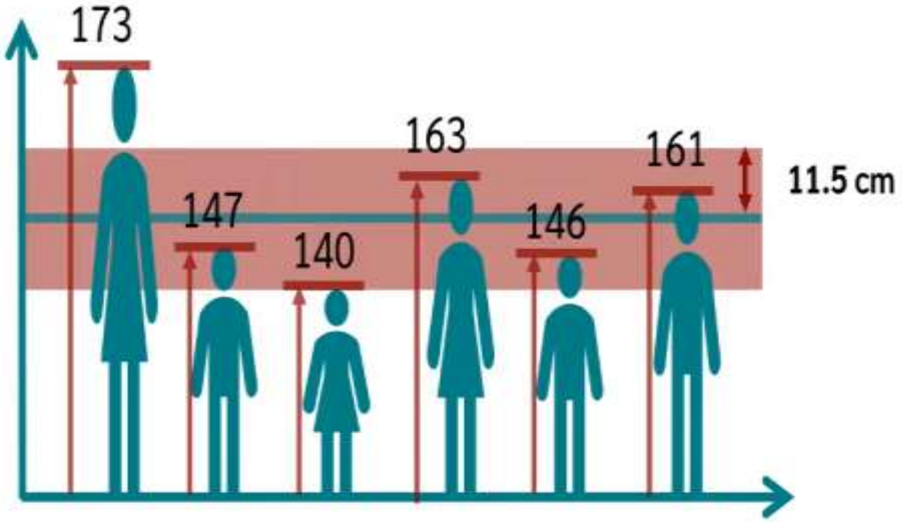
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Dispersion measures provide information about the variability or spread of values in a dataset.

Ex:

we want to know how much the **persons deviate** from the **mean value on average**.





The result is a standard deviation of **11.5 cm**.

$$= \sqrt{\frac{(173-155)^2+(147-155)^2+\cdots+(161-155)^2}{6}} = 11.5$$

central tendency vs measures of dispersion

central tendency => describe the center or average of a dataset,

dispersion => provide information about how the individual data points are spread around that center.

In data analysis => understanding not only the average income (mean) but also the spread of incomes (standard deviation) provides a more complete picture of the economic situation.

Type of data

task 5

Two type of data:

1. **Qualitative ➔** non-numerical data but words descriptive by observation.
   - Involve 5 sense like (seeing – feeling – test – hear – smell)

2. **Quantitative ➔** numerical data
   - **Numerical data :**
     1. **Discrete (counting) : integer number .**
     2. **Continues (measurement):** decimal number.

Scales of Measurement

Type of scaling ➔ **Nominal Scale, Ordinal Scale, Interval Scale, Ratio Scale.**

Scales of Measurement

Data	Nominal	Ordinal	Interval	Ratio
Labeled	✓	✓	✓	✓
Meaningful Order	✗	✓	✓	✓
Measurable Difference	✗	✗	✓	✓
True Zero Starting Point	✗	✗	✗	✓

2- inferential statistics

inferential statistics => is predictions about a population based on a sample of data drawn from that population.

techniques in inferential statistics

1. **Hypothesis Testing =>** **is an idea that can be tested.**
   - **Null Hypothesis (H0):** A statement that there is no effect or no difference.
   - **Alternative Hypothesis (H1 or Ha):** A statement expressing the presence of an effect or difference.
   - **P-value =>** The probability of obtaining results as extreme as the observed results, assuming the null hypothesis is true.

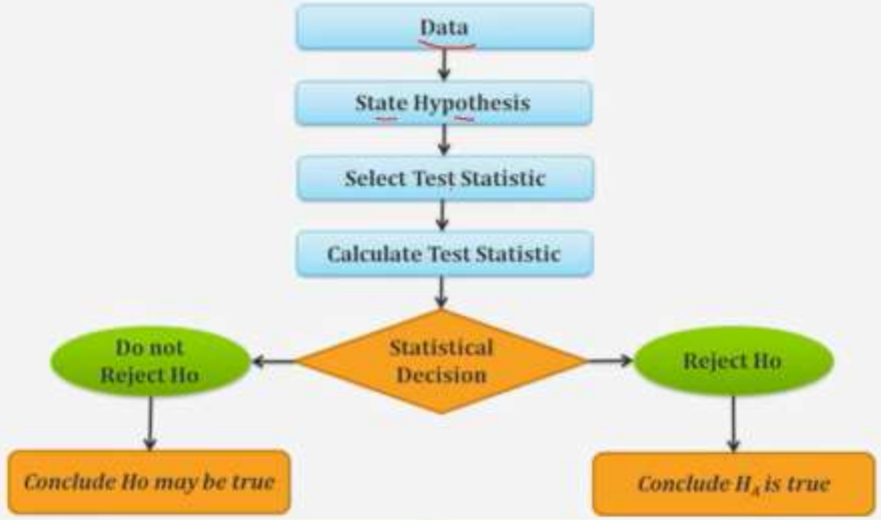
2. **Linear regression Analysis =>** used in data science to explore the relationship between a dependent variable and one or more independent variables.

   - Simple Linear regression => Involves one independent variable.
   - **Multiple Linear Regression =>** Involves more than one independent variable.

1. Hypothesis Testing

1

Hypotheses Testing Process



2

Test Statistic Calculation

- z-distribution

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- t-distribution

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Test Statistic Selection

Case	Data	Statistic
1	Normal population (σ Known)	Z
2	Not-Normal population ($n \geq 30$)	Z
3	Normal population (σ Unknown)	t

3

Statistical Decision (z)

Level α Rejection Regions for Testing $\mu = \mu_0$ (normal population and σ known)

Alternative hypothesis	Reject null hypothesis if:
$\mu < \mu_0$	$Z < -z_\alpha$
$\mu > \mu_0$	$Z > z_\alpha$
$\mu \neq \mu_0$	$Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

Example of hypothesis

Example 4

- A manufacturer of a pizza measures the amount of cheese used per run. Suppose that a consumer agency wishes to establish that the population mean is **less than** 71 pounds, the target amount established for this product. There are n = **80** observations and a computer calculation gives $\bar{x}$ = 68.45 and s = 9.583. What can it conclude if the probability of a Type I error is to be at most 0.01?

$\geq 30 \Rightarrow Z$

$\rightarrow H_a : \mu < 71 \rightarrow$ claim

$H_0 : \mu \geq 71$

α

Solution

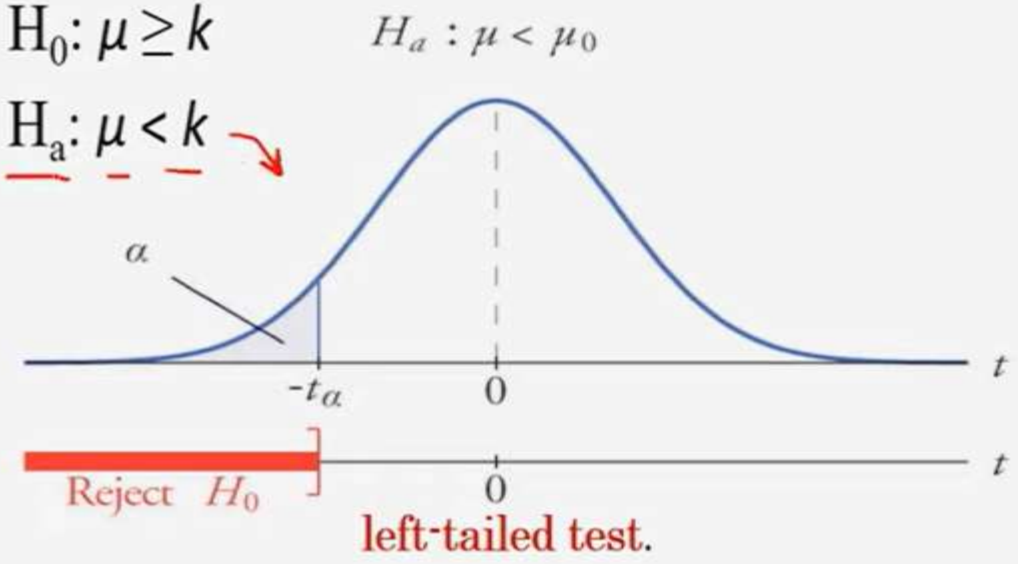
- Null hypothesis: μ ≥ 71 pounds
- Alternative hypothesis: μ < 71 pounds
- Level of significance: α ≤ 0.01 (Z = −2.33)

$$Z = \frac{68.45 - 71}{9.583/\sqrt{80}} = -2.38$$

- Decision: Since Z = −2.38 is less than −2.33, the null hypothesis must be rejected at level of significance 0.01. In other words, the suspicion that μ < 71 pounds is confirmed.

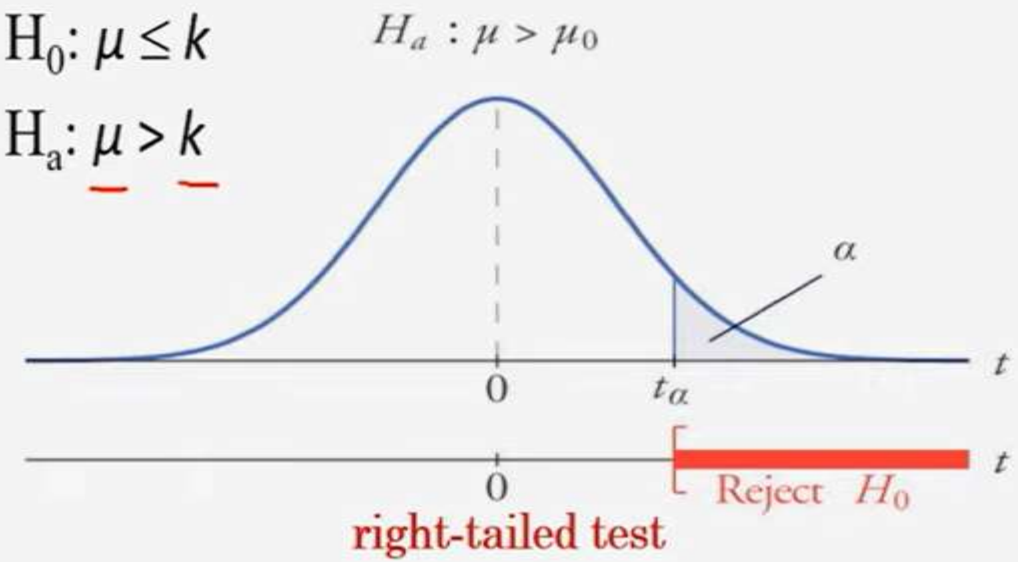
$H_0: \mu \geq k$

$H_a: \mu < k$



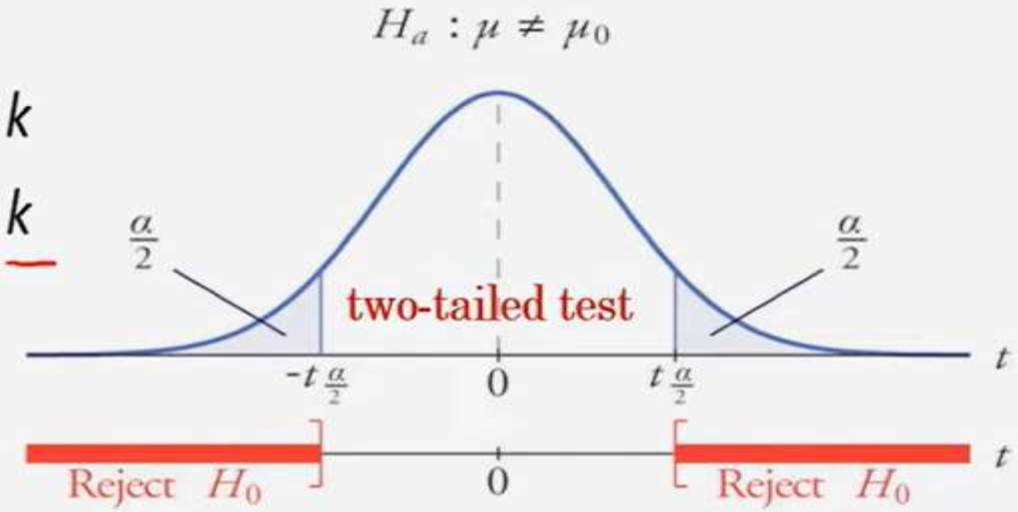
$H_0: \mu \leq k$

$H_a: \mu > k$



$H_0: \mu = k$

$H_a: \mu \neq k$

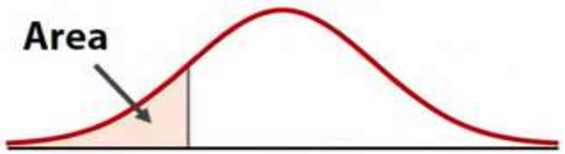


To calc Zα

Significance Level (α): This is the predetermined threshold used to determine statistical significance. Common choices are 0.05, 0.01, or 0.10. ” start of reject ”

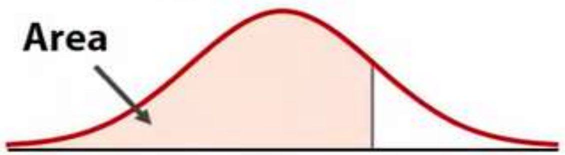
gave are
of left

Negative z



$z = -1.22$

Positive z



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

P-value

Z/t-Value vs P-value

Z or t-Value

- Level of significance (α).
- Sample (Z or t-Value).
- **Convert α to Zc or tc-Value**
- **Compare Z and Zc**
- Take the Decision

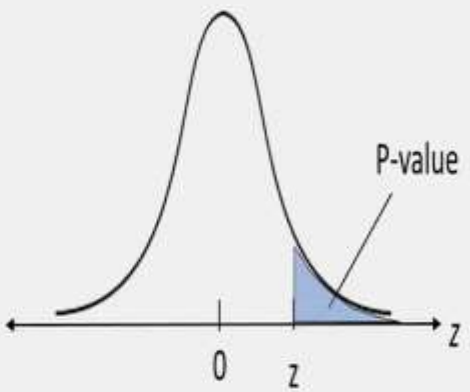
P-value

- Level of significance (α).
- Sample (Z or t-Value)
- **Convert Z or t-Value to P-Value**
- **Compare P-Value and α**
- Take the Decision

Decision Rule Based on P-value

- To use a P-value to make a conclusion in a hypothesis test, compare the P-value with α.

1. If P-value $\leq \alpha$, then **reject** Ho.
2. If P-value $> \alpha$, then **fail to reject** Ho.



Way to calc p-value

ex:

Finding the *P*-value

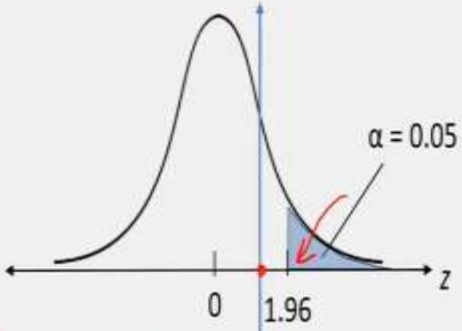
- After determining the hypothesis test's standardized test statistic and the test statistic's corresponding area, do one of the following to find the *P*-value.
  1) For a left-tailed test, P = (Area in left tail).
  2) For a right-tailed test, P = (Area in right tail).
  3) For a two-tailed test, $P = 2 *$ (Area in tail of test statistic).

Z-Value vs P-value

1. Z-Value Method

$\alpha = 0.05$ means $Z_\alpha = Z_c = 1.96$

Zsmaple = 1.56



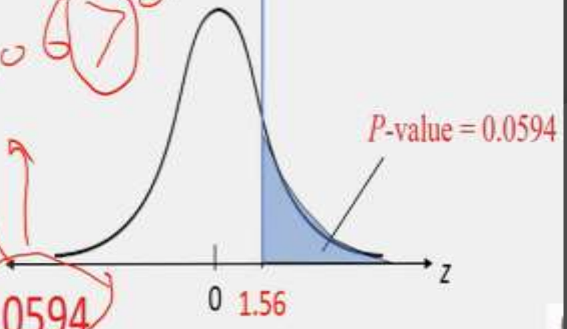
H_0 ✓ → H_0 ✓

$0.06 > 0.05$ α

2. P-Value Method

$\alpha = 0.05$

Zsmaple = 1.56 means P-value = 0.0594

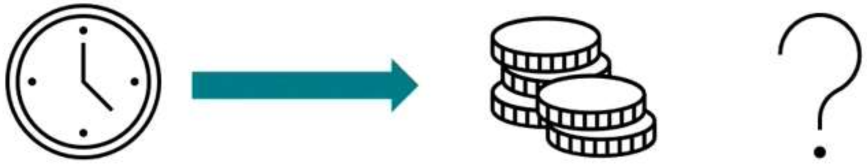


2- Linear regression ➔ is a statistical method used to model the relationship between a dependent variable (outcome) and one or more independent variables (predictor)

- **Have two type :**
  - **Simple linear regression:** Involves one independent variable.
    - Equ : $y = b.x + a + \epsilon$
  - **Multiple linear regression:** Involves more than independent variable.
    - Equ : $y = b_1.x_1 + b_2.x_2 + \ldots\ldots + b_k.x_k + a$

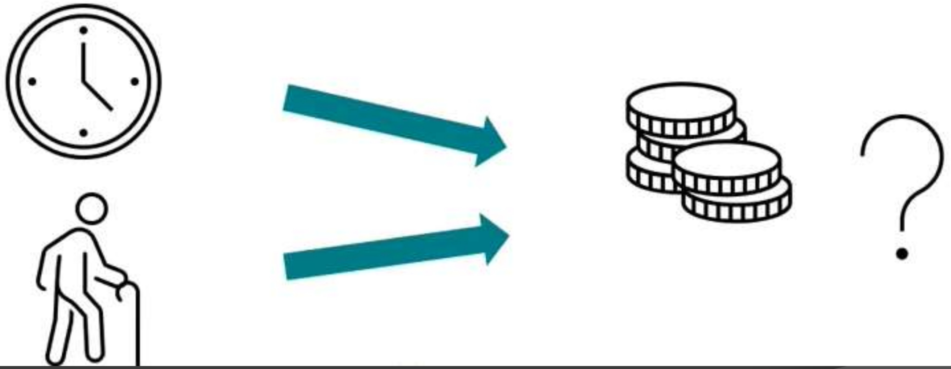
Simple linear regression

Does **the weekly working time** have an influence on the **hourly salary** of employees?

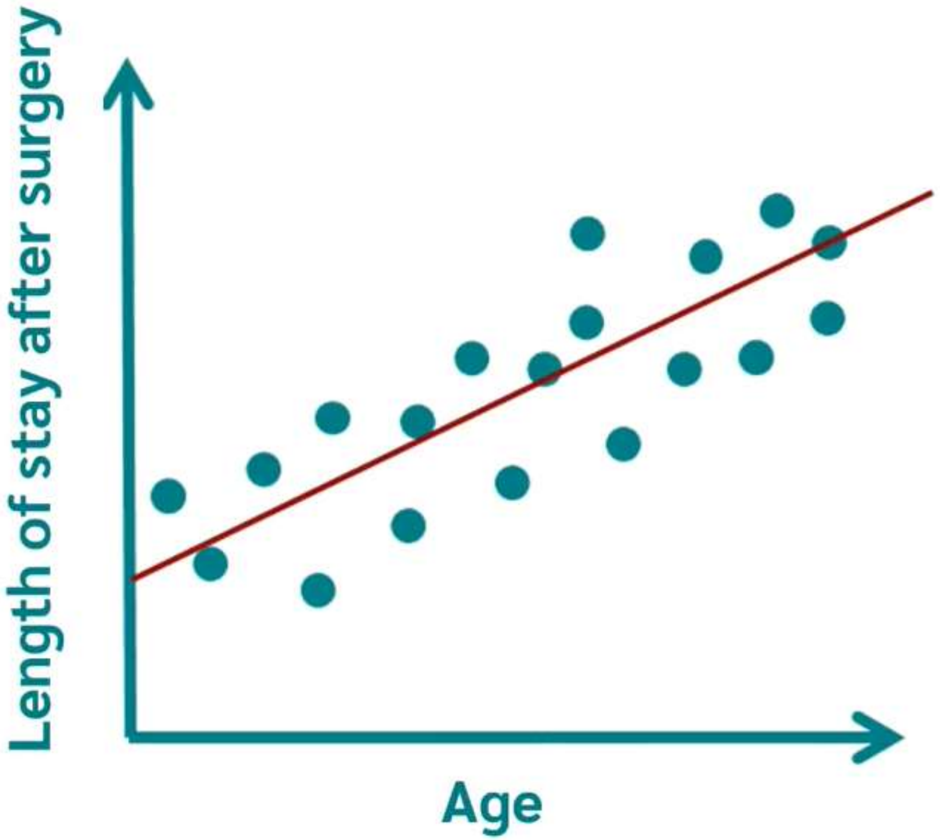


Multiple linear regression

Do the **weekly working hours** and the **age** of employees have an influence on their **hourly salary**?



Simple Linear Regression



Estimated length of stay

Age

$$\hat{y} = b \cdot x + a$$

$$\hat{y} = 0.14 \cdot x + 1.2$$

$$5.82 = 0.14 \cdot 33 + 1.2$$

Calculation of a and b

$$b = r\frac{s_y}{s_x} \qquad a = \bar{y} - b \cdot \bar{x}$$

End