Statistics:

Types of sampling methods:

1. **Simple Random Sampling** is a type of method which tells us about the creating of random number generation for the given population. Selection of the people is based on their number and it is done equally.

**FOR EXAMPLE:** In a company of 1-1000 employees. Everyone is given a random number equally either based on their joining or designation. We can call or select any number between 1-1000. So, it is based on the simple sampling method.

1. **Systematic Sampling (the selected nth person from the sample):** It is just like the simple random sampling but it works on the intervals not on the random number.

**FOR EXAMPLE:** In aproject of 100 people if we are selecting the 5th person then we will add the 10th of the interval of and we will create the interval (5,15,25,35… upto so on). So, we will select from the intervals but not individually.

1. **Stratified Sampling:** The population is divided into sub groups, called **strata**, based on (age, gender, income etc). We can also use Simple Random and Systematic way.

**FOR EXAMPLE:** In the office there are 600 employees. So, our subgroup will be based on gender which includes 500 females and 100 males now we are going to create subgroups out of them.

1. **Clustering Sampling:** We select an entire subgroup from the population. In this we do not select an individual person or interval.

**FOR EXAMPLE:**

If we want to buy some groceries from the market, we will list different groups like spices, vegetables or meat. We will select the whole subgroup.

**NON-PROBABILITY SAMPLING METHODS:**

* **Convenience Sampling (a particular survey or we can say field):** In this voluntarily an individual can be a part of the sampling from the population.

**FOR EXAMPLE:** In a survey we can ask questions individually related to the research work.

* **Voluntary Response Sampling:** As the name suggest it is based on the voluntary response of an individual or the groups (maybe subgroups).

**FOR EXAMPLE:** Online public surveys, it can be done individually or in a group.

* **Snowball Sample:** In this the sampling is biased. As it is inter-dependent. Let’s talk about the recruitment process. So, it can be through reference or we can say interview.

**FOR EXAMPLE:** A research is based on the homeless people, we can find homeless people from a homeless people or through the centre point of their base. There are no such details regarding homeless people whatsoever.

* **Purposive Sampling:** It is based on the purposefully selected samples. It is based on the expertise and judgement.

**FOR EXAMPLE:** Purposely selection of the cricket players in a training camp because of some approach.

**Descriptive** and **inferential** statistics are two major areas. **Descriptive** statistics methods are used to describe the characteristics of samples and population. **Inferential** statistics utilizes **Descriptive** statistics properties to test the hypothesis and draw inferences.

**Variable:**  A variable is a property that can take on any value.

**Height or weight are the best example.**

**VARIABLES**

**Quantitative variable Qualitative variable**

**(**measure numerical**) (**Categorical variable**)**

**Quantitative variable**

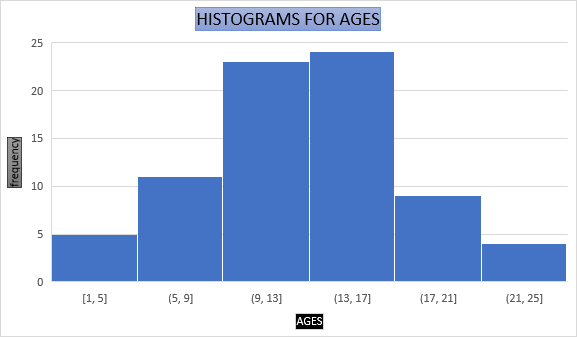
**Discrete variable continuous variable**

**(e.g. Whole number) (e.g. Height or weight)**

**VARIABLE MEASUREMENT SCALES:**

1. **NOMINAL (Categorical data)**
2. **ORDINAL (order of the data matters, value does not)**
3. **INTERVAL (order matters and the value also matters, natural zero is not there)**
4. **RATIO ()**

**BAR GRAPH** for frequency distribution. (for discrete)



**Histogram (continuous).**

**The line indicates the Kernal Destiny Estimator.**

**PDF (probability destiny function => smoothening of histography).**

**INTERMEDIATE PROBABILITY**

**Arithmetic mean of population and sample:**

**Population** (N) **Sample** (n)

S = {1,1,2,2,3,3,4,5,5,6,100}

**Median: Sort the given data**

**Odd number of counts= 11**

**median = 3**

**Even number of the counts = 12,**

S = {1,1,2,2,3,3,4,5,5,6,100,112}

**Avg. = 3+4/2= 3.5(Median)**

**{Median works well with the outlier}**

**Mode =** {1,2,2,3,4,5,6,6,6,7,8,100,100,100,100} [**for this kind of data we will use median]**

**Mode = most frequent element**

|  |
| --- |
| **Type of flower petal length petal width dataset** |
| Rose |
| Lily |
| Sunflower |
| - **{we use mode to fill up the null values for the categorical dataset}** |
| - |
| - |

**Measure of Dispersion(spread)**

1. **Variance**
2. **Standard deviation**

To identify two different samples and compare them we will use measure of dispersion.

**Population variance Sample variance**

**Here it is,**

**X x- *(*x- *)2***

**1 2.83 -1.83 3.34**

**2 2.83** **-0.83 0.6889**

**2 2.83 0.83 0.6889**

**3 2.83 0.17 0.03**

**4 2.83 1.17 1.37**

**5 2.83 2.17 4.71 *(*x- *)2* /6= 1.81**

**Sum= 10.84**

**this is the standard deviation = 1.345**

**And it will draw a bell curve.**



**Percentile and Quartiles {to find the outliers}**

**% of the number that are odd? In 1,2,3,4,5**

**% = # of numbers that are odd / total numbers = 3/5=0.6 = 60%**

**Percentile: a percentile is a value below which a certain percentage of observation lies.**

**Dataset: 2,2,3,4,5,5,5,6,7,8,8,8,8,8,9,9,10,11,11,12**

**What is the percentile ranking of the 10?**

**X = 10**

**(Number of values below x/n) 100 = 16/20 (100) = 80 percentile (80%)**

**80% of the entire distribution is less than 10**

**And For 11 = 85%**

**To find the value for the percentile we have**

**Value = percentile (n+1)/100= index position**

**{1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27}**

**Lower fence = Q1 – 1.5(IQR) = -3**

**Upper fence =Q3 +1.5(IQR) = 13**

**IQR (interquartile range) = Q3-Q1**

Q1 = 25%, VALUE = 5th index = 3

Q3 = 75%, VALUE = 15th index = 7

IQR = 7-3 = 4



**We have minimum = 1**

**Q1 = 3**

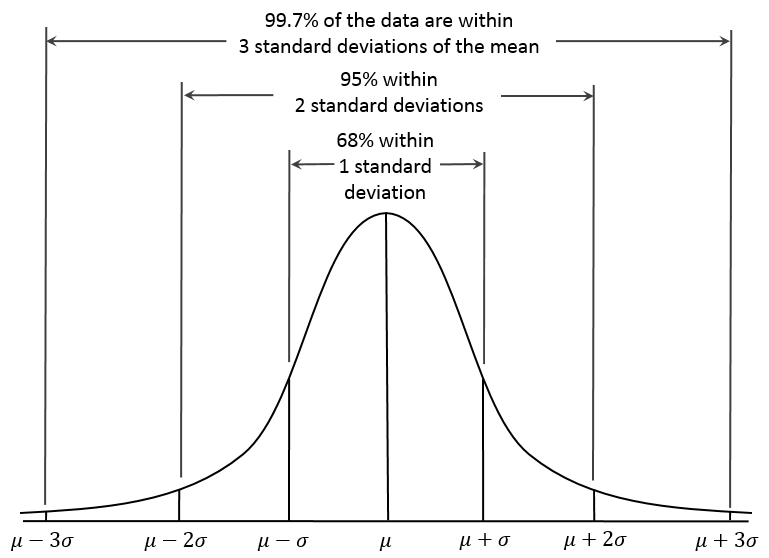
**Median = 5**

**Q3 = 7**

**Maximum = 9**

**Advance statistics**

**Gaussian / normal distribution:**



**Empirical rule:**

**68-95-99.7%**

**Let if and then we can see that there is a .5 right deviation of value from the mean.**

**Now, to calculate the deviation is it from the right side or the left side we use Z score.**

**Z score = = +0.5(which is to the right deviation)**

**{1,2,3,4,5,6,7}**

**If we apply Z score to this dataset we will get {-3, -2, -1, 0, 1, 2, 3}**

**By this we can see the 1st,2nd,3rd standard deviation by the use of Z score we can see the and call it standard normal distribution. We can compare previous table with the recent values with the help of Z score.**

**We convert normal distribution into standard normal distribution to get .**

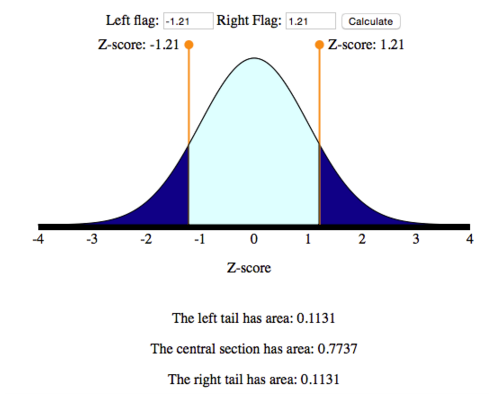
**If we have and a table, the use of Z score with each column It will give us standardization.**

|  |  |  |
| --- | --- | --- |
| **Age (years)** | **Salary ($)** | **Weight (kg)** |
| **23** | **120k** | **67** |
| **24** | **230k** | **68** |
| **25** | **340k** | **69** |

**Normalization**: We have an option to change any column from 0 to 1. The use of normalization in CNN, we know that there are 0 to 255 pixels. So, to use the normalization we divide each and every value of the column to get the value from 0-1 intervals.

**If we have Z score = -1.25 and Z score = -1, the area of the first Z score is lager than the Z score = -1. Now, the conclusion is that the Z score = -1 is better than the other Z score.**

**“NOW, WE HAVE THE VALUES OF THE Z SCORE AND TO FIND THE PERCENTAGE OF THE Z SCORE WE WILL GO THROUGH THE Z TABLE.”**



**PROBABILITY**

Sample {1,2,3,4,5,6}

P(x)= number of ways an event occur / total possibility of the outcome

**Additive rule (OR):**

Mutual exclusive events: Two events are mutual exclusive if they cannot occur at the same time.

Example: {H, T} tossing a coin.

Non mutual exclusive events: Multiple events can occur at the same time.

Example: Deck of the cards {Queen of hearts}

Q1. Of getting heads or tails.

P (H OR T) = p(H) +p(T) – p (H and T)

Or

and

P (H or T) = ½= 0.5

**Multiplication rule:**

Independent events (rolling a dice)

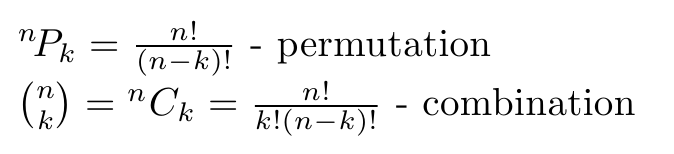
P (A and B) = P(A)\*P(B)

Dependent events (probability of getting out red marbles out of the bag.)

**Conditional probability: P (A and B) = P(A) P(B/A)**

**PERMUTATION AND COMBINATION**

We have permutation and combination. By permutation we can arrange or create the order for the problem. In combination we can select from the sample or any event.



**Statistics:**

1. **Probability distribution function -**

* **Probability destiny function**
* **Probability mass function**

1. **Cumulative distribution function**

**Probability distribution function**

**probability destiny probability m.f.**

**function (continuous random variable) (discrete random variable)**

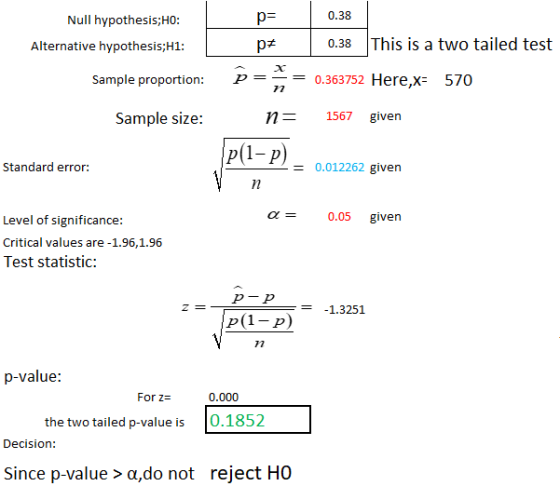
**Hypothesis Testing and Statistical Analysis:**

**Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.**

Inferential statistics sample data to population data. So, we use hypothesis testing.

Tossing a coin. Coin is fair or not.

1. Null hypothesis (Ho) = coin is fair (there is no significant difference or effect, or no change, in a specified population) Researchers assume that the null hypothesis is true until there is enough evidence to reject it.
2. Alternate hypothesis (H1) = coin is not fair (It contradict with the null hypothesis. It tells us that there is a significant difference, effect, or change in the population parameter.



**T test and Z test use?**

Do you know the population standard deviation?

What is the sample data? is it > 30?

If it is then we will use z test.

If it is not then in both case, we will use T test.

We have population mean normal average = 168 cm

Std = 3.9

Mean is different (total 36 individuals)

Average after what the doctor calculated ()= 169.5

*and (alternate hypothesis)*

confidence interval = 0.95= 95%

α = 1- C.I. = 1-0.95 = 0.05

1. Decision boundary

1 – 0.025 = 0.9750

Z = -1.96< null hypothesis< Z = +1.96 it will not be rejected.

If it does not lie in the range then it will reject the hypothesis.

Z test = where the value of n = 1.

Z test = 2.31

Therefore, the null hypothesis is rejected.

**Hypothesis test and statistical analysis using t test:**

In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on the intelligence or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence?

µ= 100

n = 30

= 140

S = 20

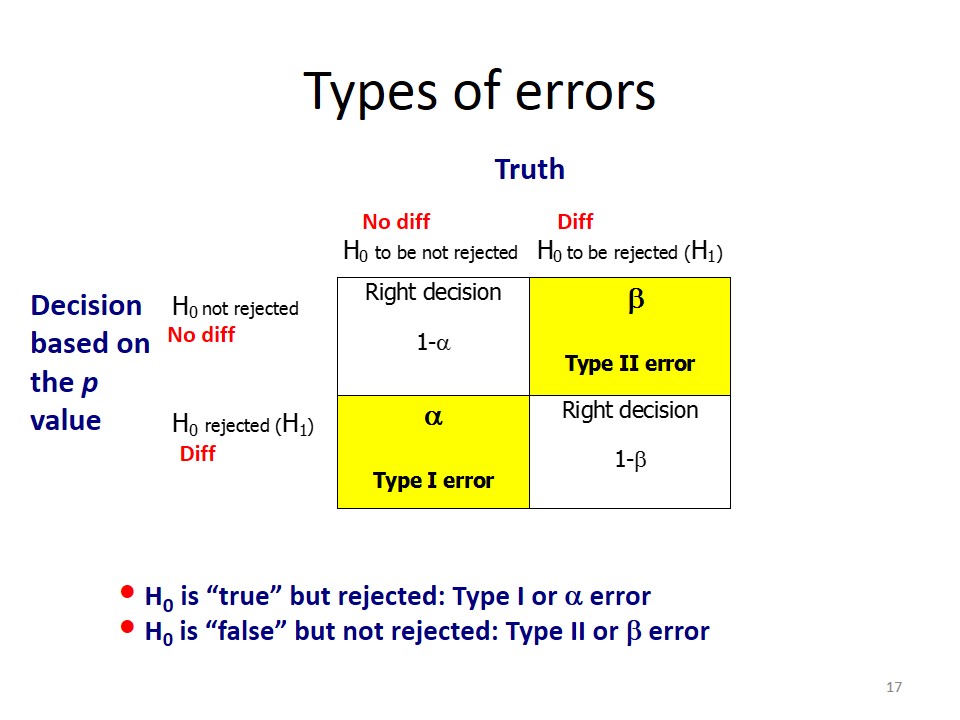
Null hypothesis = 100

Alternative hypothesis! = 100

Significance value α = 0.05

Degree of freedom = n-1 = 30-1 =29 (basically it is the choices for any sample)

t = = 10.96 (reject null hypothesis)



**Type 1 and type 2 error (M.L. concept)**