10/8/25 Data Science Journey Yasir Baig

{Statistics}

Basic Concepts
- Introduction
- Gathering data
- Describing data
- Making conclusions
- Prediction
- Populations
- Parameters
- Study Types
- Sample Types
- Data Types
- Measurement levels

Descriptive Statistics
- Descriptive statistics
- Frequency Tables
- Histograms
- Bar charts
- Box Plots
- Average
- Mean / Medium / Mode
- Variation
- Range
- Quartile / Percentile
- Interquartile range
- Standard deviation

Inferential Statistics
- Statistical Inference
- Normal Distribution

- Standard Normal
- T-Distribution
- Estimation
- Proportion Estimation
- Mean Estimation
- Hypothesis Testing
- Testing porportion
- Testing Mean.

"Statistics Introduction"

Statistics gives us methods of gaining knowledge from data.

→ What is statistics used for?

Statistics is used in all kinds of science & business application. Statistics gives us more acurte knowledge which helps us make better decisions. Statistics can focus on making predictions about what will happen in future. It can also focus on explaining how different things are connected.

Note: Good Statistical explanations are also useful for predictions.

Typical Steps of Statistical Methods

3 Step: or methods

- Gathering data
- Describing & visualizing data
- Making conclusions.

It is important to keep all 3 steps in mind for any questions we want more knowledge about. Knowing which types of data are available can tell you what kinds of questions you can answer with statistical methods. Knowing which questions you want to answer can help guide what sort of data you need. A lot of data might be available, and knowing want to focus on is important

How is Statistics Used?

Statistics can be used to explain things in a precise way. you can use it to understand & make conclusions about the group that you want to know more about. This group is called population.

A population could be many diff kinds of group:

- All of people in a country
- All the business in industry
- All the customers of business
- All people that play football who are older than 45

and so on - It just depends on what you want to know about.

Gathering data about population will give you a sample. This is part of whole population. Statistical methods are then used on that sample. The results of statistical methods from the sample is used to make conclusions about population.

Note: The word 'statistic' can also refer to specific bits of knowledge; like the average value of something

Statistics - Gathering Data

Gathering data is the first step in statistical analysis

- Say for example that you want to know something about all the people in France.

The population is then all of people in France.

→ It is too much effort to gather information about all members of a population (e.g all 67 million people in France). It is often much easier to collect a small group of that population & analyze that. This is called a sample.

A Representative Sample

The sample needs to be similar to the whole ~~to the~~ who population of France. It should have the same characteristics as the population. If you only include (easier to collect a smaller group of that population) people named jacques living in Paris who are 48 years old, the sample will not be similar to the whole population.

- So for a group sample (needs to be similar to the whole) you will need people from all over France, with different ages, profession & so on.
- If the members of the sample have similar characteristic (like age, profession etc) to the whole population of France, we say that the sample is representative of the population
- A good representative sample is crucial for statistical methods.

Statistics - Describing Data

Describing data is typically the second step of statistical analysis after gathering data.

Descriptive Statistics

The information (data) from your sample or population can be visualized with graphs or summerized by numbers. This will show key information in a sample way than just looking at raw data. It can help us understand how the data is disbributed.

Graphs can visually show the data distribution Example of graphs include:

- Histograms
- Pie charts
- Bar charts
- Box Plots

Some graphs have a close connection to numerical summary statistics. Calculating those give us the basis of these graphs. For example

- A box plot visually show the quartiles of data distribution.
- Quartiles are the data split into four equal size parts or quartiles. A quartile is one of summary statistics.

Summary statistics

Summary statistics condense large amounts to data into a few key numbers that describe it.

- Mean, Medium, Mode : Show the center of data
- Range, Interquartile range: show how spread out
- Quartile, percentiles : Show data position
- Standard Deviation, Variance: Show how much data v

Why it is useful.

- Gives a quick overview of dalsa
- Helps decide what to analyze further
- Guides deeper statistical analysis.

Statistics Making Conclusions

Statistical ~~inter~~ inference - Easy guide

Defination:

Using data from a sample to make conclusions about an entire population.

Probability Theory:

It helps estimate how certain we can about those conclusions.

They are some uncertainty when working with samples.

Key Tools:

- Confidence Intervals: Show the range where the true value for the population is likely to be.
- Hypothesis Testing: Check if a statement about a population is likely to be true.

Example:

- Are people in Netherlands taller than in Denmark
- Do people prefer pepsi or coke?
- Does a new medicine work?

Causal Inference:

- Find out if one thing causes another (not related)
- Example: Does rain makes plants grow?
- Requires good experiments or special statistical method
- Hard to prove because of ethical & pratical limits.

Statistic - Prediction & Explanation

Some types of statistical methods are focĕd on predicting what will happen.

Other types of statistical methods are focused on explaining how these are connected.

Prediction Vs Explanation

- Focus: Accuracy of results, not why they happen.
- Can work well with out explaining connection.
- Some machine learning models predict well but are hard to understand.
- Predictions about the future = forecasts but prediction can also be about unknow things in the present or past.
- Risk: May fail if condition change, since the how is unclear.

Explanation

Focus: Understanding how things are connected

May not give the best predictions

Often explains only part of situation

if it explains all important factors, it can also predict well.

Causal inference = finding out if not one thing cause anoter

Causality is tricky to prove, especially when many things connetcd

Conclusions about cause should be made carefully.

Statistics - Population & Samples

In statistics, the term population & sample are very important because most statistical studies deal with them.

Population

The entire group you want to study or learn about. Can be people, animals, products, events, anything you are studing

Example

- All people in Germany
- Every Netflix customer in world
- All car manufactorers globally.

Sample

A small part of the population that you actually collect data From.

Must be carefully chosen or it represents the whole population.

Examples:

- 500 Randomly choosen germans
- 300 Netflix Customers.
- Tesla, Toyota, BMW, all manufacturers

Why Sample Are Important?

- Studing an entire population is often impossible (Too big, expensive, or time consuming)
- A good sample can give accurate information about the population
- If the sample is not similar to population, Conclusions maybe wrong or useless.

Representative Samples

- A representative sample has the same key characteristics as the population (e.g; age, gender, location, habits).

Example: If 50% of the country is Female, a representative sample should also be about 50% Female.

Key Points:

- In everyday language, 'population' usually means people.
- In statistics 'population' means any group you want to study, people, animals, objects, events or measurements.

Statistics - Parameters and Statistics.

Parameter

A number that tell us something about the whole population.

Usually unknown because we can't study everyone or everything.

Statistics

A number that tells us something about a sample (part of the population).

We use it to guess the parameters

How They Work Together

Population - Parameter (true, value but unknown)

Sample - Statistic (known value, used to estimate parameter)

A good, representative sample gives a better estimate.

They always some uncertainty in the estimate.

Examples

Parameters (Population)	Statistics (Sample)
Mean (average)	Sample mean
Median (middle value)	Sample median
Mode (most common values)	Sample mode
Variance (spread of values)	Sample variance
Standard deviation (spread)	Sample standard deviation

What They Tell Us

Mean, Median, Mode 'Typical' value:

- Example: average age in country, average profits of a company, average range of electric car.

Variance & Standard Deviation: How spread out data is

- Example

one classroom = about same age → low spread
whole country = many different ages → high spread.

Note:

Parameter = describes the population (unknown)
Statistic = describes the sample (known)
Statistics help us guess parameters.

Statistics - Types of Studies

When doing statistic, we can collect data in different ways. The two main studies are:

Observational Study

- We watch and record data without changing anything.
- Example: Counting how many people wear seatbelts by observing traffic
- Good for finding patterns, not strong for proving cause-and-effect.

Experimental Study

- We change something to see its effect.

Usually compar 2 groups:

- Treatment group - gets the change (medicine)
- Control group - does not gets the change

Example:

Testing if a new drug works by giving it to one group and not to another.

Better for finding cause-and-effect, but harder

Note:

Observational = Record what happens naturally

Experimental = Change conditions & see results.

Experimental studies, if well-designed, give stronger proof about causes.

Statistics - Sample Types

When doing a study, we need participants How we choose them matters, some methods gives better results, might be harder.

Random Sampling

- Everyone in the population has equal chance of being chosen.
- Best Method (most accurate)
- Harder to make perfectly random in real life.

Tip: All other sampling methods are compared to this

Convenience Sampling

- Everything in population has equal chance of bei
- Pick the people easiest to reach.
- Easiest to do, but often not representative of the population.
- Can lead to misleading results.

Systematic Sampling

- Choose participants using a fixed rule of pattern
- Example:
  - → First 30 people in queue
  - → Every third person on list.
  - → First 10 & last 10 names.

Stratified Sampling

- Step 1: Split the population into groups (called strata) based on something like:

- Age groups
- Jobs/professions

- Step 2: Select participants from each group (often using random sampling).
- Good for making sure all important groups are represented.

Cluster Sampling

- Step 1: Split the population into clusters (natural groups e.g cities, schools).
- Step 2: Randomly pick some clusters.
- Step 3: Either use all people in those clusters or randomly choose some from them.

Quick Memory Trick:

Random = Fair Chance For Everyone
Convenience = Easy For Use/You
Systematic = Fixed Pattern
Stratified = Groups → sample from all
Cluster = Groups → pick some groups only

Statistics - Data Types

In statistics, data comes in different forms. The type of data decides how we can study & analyze it.

Two Main Types of Data

Qualitative Data (also called Categorical Data)
Quantitative Data (also called Numerical Data)

Qualitative (categorical) data

What it means:

Information that tell us what someth
is, but not how much or how many.
It describes categories or labels, not numbers.

Examples:

- Brand (Nike, Adidas, Puma)
- Nationality (Pakistani, Indian, American)
- Profession (Doctor, Engineer, Teacher)

What we can do with it:

- Count how many people are in each category
- Find percentages or proportions.
  - Example "30% of people prefer "Brand A"

Quantitative (Numerical) data

What it means:

Information we can measure & write in num
It tells us quantity, amount, or size.

Examples:

- Age (25 Years)
- Height (170 cm)
- Income ($ 50, 0000/Year)

What we can do with it:

- Find averages (mean, median)
- Find ranges, totals, or differences
  - Example: "The average height of Players 175

Note:

If data is names / Labels – its qualitative
if data is numbers we can measure or calculate with – its Quantitati

Statistics - Measurements Levels

Different kinds of data are measured in different ways. These measurement levels tells us:

- What types of statistics we can use
- How to present the data correctly

Nominal level

- What it is: Just categories or label, No order.
- Can we rank them? No
- Example: Brand Name, countries, color.

Ordinal level

- What it is: Categories that have an order, but the exact difference between them is unknown or not meaningful
- Can we rank them? Yes
- Examples:
  - Letter grades (F, D, C, B, A)
  - Military ranks
  - Satisfaction level (Poor, Fair, Good, Excellent)

Interval level

- What it is: Numbers with equal spacing between values, but no true zero.
- Can we rank them? Yes
- Are gaps equal? Yes
- True zero? No
- Example:
  - Year (2000, 2005, 2010),
  - Temperature in °C or °F.

Ratio level

- What it is: Numbers with equal spacing and a true zero.
- Can we rank them? Yes
- Are gaps equal? Yes
- True Zero? Yes
- Examples:
  - Money, age, time, height, weight

Quick Summary Table

Level	Types of Data	Order	Gaps	Zero	Examples
Nominal	Qualitative	No	No	No	Color
Ordinal	Qualitative	Yes	No	No	ranks
interval	Quantitative	Yes	Yes	No	Years
Ratio	Quantitative	Yes	Yes	Yes	Age, tall

Descriptive Statistics

Description / Descriptive statistics help us understand a dataset quickly without looking at every single number. We do this by:

1. Calculating key numbers that summarize the data
2. Drawing graphs to see the data visually

Three Main Things We Look AT

1. Center of the data: → Where is the most of data?
   - Shows the middle point or where values tend to gathe
   - Measured by:
     - Mean → average of all values
     - Median → middle value when data is sorted.
     - Mode → value that appears most often
   - These are called location parameters (they tell us where the data is located on a number line).

2. Variation of the Data → "How spread out is the data"?
   - Shows how much the data changes from the center
   - Measured by:
     - Range → biggest value → smallest value
     - Standard deviation → how far values are from mean
     - Quartiles → divide the data into 4 equal parts
   - These are called scale parameters

- Shape of the Data → "What does the distribution look like?
  - Shows wheather the data is symmetrical or skewed.
  - Measured by:
    - Skew:
      - Right skew → tail longer on right side.
      - Left skew → tail longer on left side.

- Shape tells us if values are bunched on one side more than the other.

How we present the Data?

1- Frequency Tables

- Count how many times each value occurs.
- Often grouped into intervals for continuous data.
- Useful for making graphs.

2- Visualizing the data

Different graphs for different data types:

- Pie charts → for qualitative (categories like colors, brands)
- Histograms → for quantitative (numbers like height, weight
- Scatter plots → for bivariate data (two variables at one
- Box plots → show quartiles, median, minimum, maximum an

Why it's useful

- Quick summary of large datasets
- Helps see patterns, trends, outliers, without checking valu
- Guides what type of statistical test / model we use late

Summary

Feature	Question & Answer	Example of Measur
center	where's the middle?	Mean, Median, Mod
Variation	How spread out?	Range, Std, Quarti
Shape	Symmetry or skew	Skewness

Statistics - Frequency Tables

A frequency table is a way to organize data so we can see patterns and distributions quickly.

What Frequency Means

- Frequency = How many times a value appears in dataset
- Example: if "age 50" appears 3 times, then frequency is 3

Why use a Frequency Table?

- It summarize large datasets into neat table.
- Helps us see how values are spread out.
- If there are too many unique values, we group them into intervals (also called bins).

Example - Noble prize Winners Age (up to 2020)

Age Interval	Frequency
10 - 19	1
20 - 29	2
30 - 39	48
40 - 49	158
50 - 59	236
60 - 69	262
70 - 79	174
80 - 89	50
90 - 99	3

We can See:

- Only 1 winner was 10 - 19 years old.
- Most winners were in their 60s (262 people).

1- Relative Frequency Table

- Relative Frequency = Frequency ÷ Total data points
- Shows percentage instead of counts.
- Helps compare across different-sized datasets.

Example – Nobel Price Winners Relative Frequency (934)

Age Interval	Relative Frequency
10 – 19	0.11 %
20 – 29	0.21 %
30 – 39	5.14 %
40 – 49	16.92 %
50 – 59	25.27 %
60 – 69	28.05 %
70 – 79	18.35 %
80 – 89	5.35 %
90 – 99	0.32 %

2- Cumulative Frequency Table

- Cumulative frequency = Adds up all frequencies up to pa
- Shows "how many values are less than equal to giv

Example - Nobel Prize Winners Cumulative Frequency

Age Group	Cumulative Frequency
younger than 20	1
younger than 30	3
younger than 40	51
younger than 50	209
younger than 60	445
younger than 70	707
younger than 80	881
Younger than 90	931
Younger than 100	934

we can also Cumaltive Relative Frequency (in percentages)

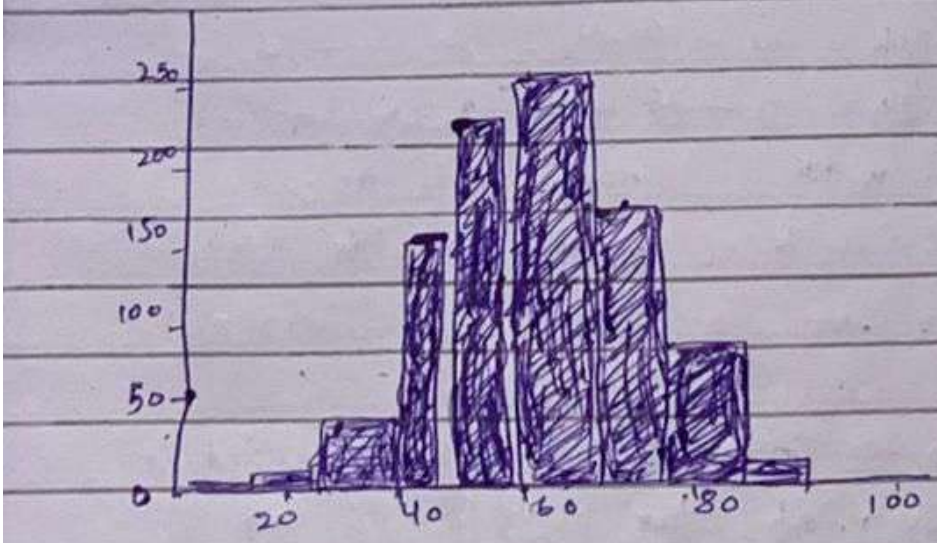
Key Terms Recap

Term	Meaning
Frequency	Number of times a value appears.
Relative Frequency	Frequency as a percentage of total.
Cumulative Frequency	Total count up to a certain point
Bin (interval)	Group of values combined into range

Statistics - Histograms

A Histogram is a graph that shows how numerical data is distributed.

Its only for quantitative (number) data → not categories.



How a Histogram Works

- We group data into intervals (called bins)
- Each bin is shown as a bar
- Height of bar = how many values (frequency) are in bi
- Bars are placed next to each other on a number line (no gaps, unlike bar charts, for categories).

Example Nobel Price Winner Ages

- Bin: 10-19, 20-29, 30-39, ... 90-99
- The tallest bar is for 60-69, meaning most winners were in their 60s.

Histogram Vs Bar Graphs

Histogram	Bar Graph
Show numerical data	Show categoral data
Bar touch	Bars have graph
order is by number range	order can be any

Bin Width

The length of interval is called bin width.

Example:

- Bin width = 10 → 10-19, 20-29, 30-39, ...
- Bin width = 5 → 15-19, 20-24, 25-29, ...

- Smaller bin width → more detail in graph
- Larger bin width → simpler, but less detail.

Quick Recap

- Histogram → Shows how numbers are spread out
- Bar → Touch each other, each bar is bin.
- Bin width → Controls the level of detail
- Best practice → choose a bin width that shows enough detail but isn't confusing.

Statistics – Bar Graphs

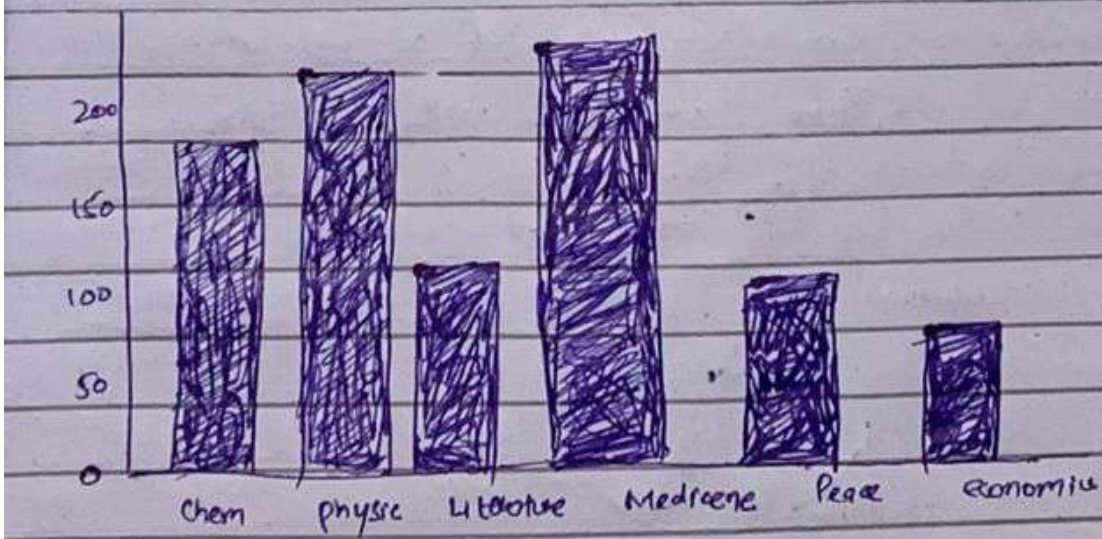
A bar graph is a chart that shows categorical (qualitative) data visually.

How a Bar Graph Works

- Data is split into categories (e.g physics, chemistry, liter)
- Each category is shown as bar
- Height of bar = how many time that category appears.
- Bars are usually separated by spaces (unlike histograms)

Example – Nobel Prize Categories (upto 2020)

- Each bar shows the number of winners in that Nobel prize category.
- Some categories have more winner because:
  - They have existed Longer
  - They sometimes have multiple winners in Year.



Quick Recap

- Bar Graph → for categories (qualitative data).
- Height of Bar → shows Frequency.
- Bar have space b/w them.
- Similar to histogram, but for different types of data.

Statistics- Pie Charts

A pie chart is a circle chart use to show categoral (qualitative) data

How a pie chart works

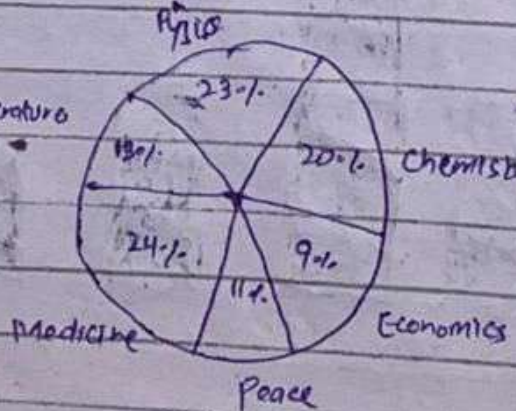
- The whole pie = 100% of data
- Each slice = one category
- Size of slice = frequency or relative frequency
- Bigger size/slice = category has more values

Frequency vs Relative Frequency

- Frequency → number of times a category appears
- Relative Frequency → precentage of total

Example:

- Each slice shows the percentage of winner
- Some slices are bigger because
  - They category existed for more year
  - Some years had multiple winners



Quick recap:

- Pie chart → For categories show proportions
- Each slice size = Frequency or percentage
- Whole pie = 100% of data
- Best when categories are few & distinct

Statistics - Box Plots

A box plot (box or whisker plot) is a graph that shows the main features of numerical data at a glance.

What a box plot shows:

1- Median → The middle value

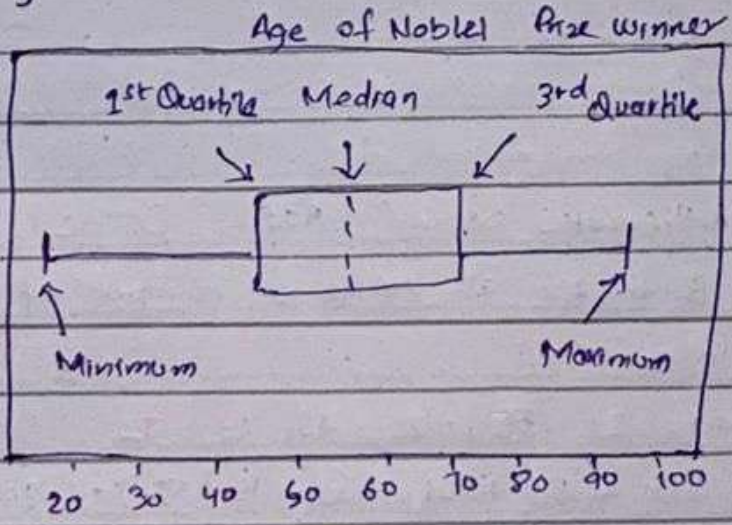
2- Quartile

- Q1 (1st Quartile) → value at 25% of data
- Q3 (3rd Quartile) → value at 75% of data.

3. Interquartile Range (IQR): Distance b/w Q1 & Q3 (50%)

4. Minimum & Maximum → Smallest & largest values

5- Range → Max - Min.



Median → 60 years (Red line inside box)

Q1 → ~~60~~ 51 years (25% of winners younger than this)

Q3 → ~~81~~ 69 years (75% of winners younger than this)

IQR - 69 - 51 = 18 years (middle half b/w 51 & 69)

Minimum = 17 years (youngest winners)

Maximum = 97 years (oldest winners)

Range = 97 - 17 = 80 years

Why box plots Are Useful

- Summarize a lot of information in one small g
- Show center, spread, & range at once.
- Make it easy to spot outliers (value for outside box)

Statistics - Averages.

An average is a number that shows where most of the values in a dataset are located.
It is called a measure of central tendency. because it describe the center of data.

The most common Averages.

Mean
Median
Mode

1- Mean (Arthematic Mean) Add & Divide,
Formula = Sum of all values ÷ number of value
Example:
Values → 40, 21, 55, 21, 48, 13, 72
Mean = (40 + 21 + 55 + 21 + 48 + 13 + 72) ÷ 7 = 38.57
Important: The mean is affected by extreme value (very large or small value)

2- Median "Middle value"
Arrange the number in order and pick the middle one.
Ordered → 13, 21, 21, 40, 48, 55, 77
Median = 40

If we change the last value from 72 to 1001
Median is still 40
Note: The median is not much affected by extreme values.

3- Mode "Most Frequent."
- The value that appears most often.

Example:
40, 21, 48, 21, 56, 45, 21, 72 = 21 is mode

A dataset have
one mode (unimodel).
More than one mode (bimodel / multimodel)
No mode (if all values are unique)

Statistics - Mean

What is the Mean?
- The mean is a type of average that shows the center of the datasets
- It is calculated by adding up all the values and dividing by the number of value.
- Mean is only calculated for numerical variables (like age, height, income).
- Here, "mean" refers to the arithmetic mean, the most common type.

Formular for the means

Population Mean (μ, mu)	Sample Mean ($\bar{x}$, x-bar)
$\mu = \frac{\Sigma X}{N}$	$\bar{x} = \frac{\Sigma X}{n}$

Where:

Σ = Sum of all values

X = the value in dataset

N = total number of values in population

n = total number of values in sample

Example Calculation:

4, 11, 7, 14 => 4+11+7+14 => 36%4=9

Mean = 9.

Mean in Programming (Python - Numpy)

```
import numpy as np
values = [4, 11, 7, 14]
mean_value = np.mean(values)
print(mean_value)
```

Symbols in Mean Calculations.

μ = population mean

$\bar{x}$ = Sample mean

Σ = Sum of value

x = Data value

i = Index num of data point

N = Num of value in population

n = Num of values in sample.

Key points to Remember

- Mean works best for numerical data.
- Sensitive to outliers
- Commonly used in statistics, ML & data analysis
- For large data, software tools are preferred over manual collection/calculations

Statistics - Median

The median is the middle value in a dataset when its ordered from smallest to largest.
It tells us where the center of data is
Can only be calculated for numerical variables

How to Find the Median

- Order the data from smallest to largest
- Count the number of observations (n)
- If
  - n is odd - The median is middle value
  - n is even - The median is aver of 2 middles values

Example:

13, 21, 40, 48, 55, 72

n = 7 (odd)

Middle value = 4th value = 40

Important Notes

- Always sort the data first before finding the median.
- The median is not affected much by extreme values (outliers), unlike the mean.

Median in programming (Python Numpy)

```
import numpy as np
values = [12, 13, 21, 40, 42, 48, 55, 72]
median_value = np.median(values)
print(median_value)
```

Key points to Remember
- Median - middle value of ordered data.
- Odd n → single middle value
- Even n → average of 2 middles values.
- Less affected by outliers than mean
- Works For numerical data only

Statistics - Mode

What is Mode?
- The mode is the value that appears most often in data set.
- Unlike the mean & median, The mode is used for numbers and categories.

Types of Modes
Unimodel → only one mode
Bimodel → two modes
Multimodel → more than 2 modes

Example 4, 7, 3, 11, 2, 4, 4, 12, 3
Mode = 4

Python Example
```
from statistics import multimode
values = [4, 7, 3, 2, 11, 4, 4, 13]
moder = multimode(values)
print(mode)
```

Key Points
- Use For both numerical & categorical
- most common occurance

Statistics - Variation

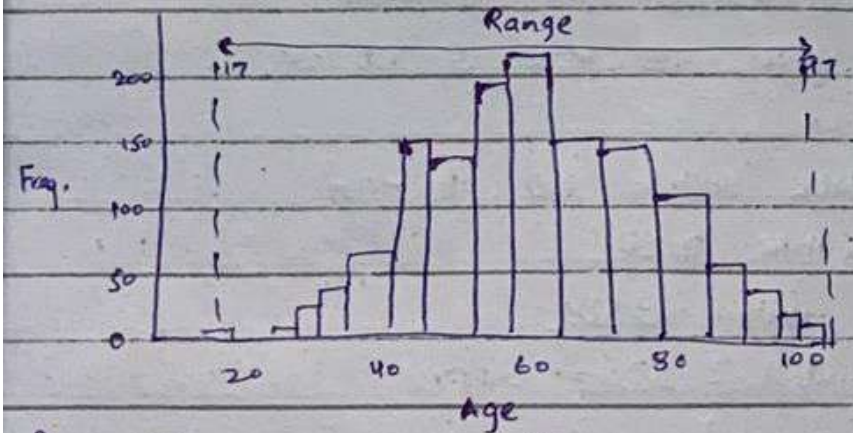
What is variation?

- Variation tells us how spread out the data is around the center (mean or median)
- It measures how far apart the values are from each other
- Can only be calculated for numerical data.

Range

The difference between the largest and smallest value

Formula: Range = Max Value − Min Value



Pros: very easy to calculate

Cons: Affected by extreme values (outliers)

Quartiles & percentiles

- Quartiles - split data into 4 equal parts
  - Q_0 Minimum value
  - Q_1 25%
  - Q_2 Median (50%)
  - Q_3 75%
  - Q_4 Max value
- Percentile - Split data into 100 equal parts.

  eg → 90th percentile = value below which 90% of data lies.

Interquartile Range (IQR)

- Middle 50% of data.
- Formula: IQR = Q3 − Q1
- IQR = 69 − 51 = 18 years.

Why Important?

- Removes the effect of extreme value
- Shows where central bulk of data lies.

Standard Deviation (σ or s)

- Measures the average distance of each data point from the mean.
- Small SD → values are close to mean.
- Large SD → values are more spread out.

Key points:

- 68% of data within 1SD of mean.
- 95% is within 2 SDs.
- 99.7% is within 3 SDs.
- Data beyond 3 SDs = outliers.

Statistics - Range

Range tells us how spread out the data is.
It is difference between the largest / smallest value.
Works only for numerical data.
Formula: Range = Max Val − Min Val.

Why Range is useful

Pros: Simple to calculate and understand.
Cons: Very sensitive to extreme values (outliers).

Programming Example

```
import numpy as np
values = [13, 21, 21, 40, 48, 55, 72]
x = np.ptp(values) # peak to peak difference
print(x)
```

Summary Table

Step 1	Find smallest value (min)
Step 2	Find Largest value (max)
Step 3	Substract min from max

Statistics - Quartiles and Percentiles

Why are they?

- Both are measure of variation (spread of data)
- They are quantiles - values that divide data into equal parts.

Quartiles

- Divide data into 4 parts

Q_0, Q_1, Q_2 Q_3 & Q_4

How to Interpret

- $Q_0 \rightarrow Q_1$: Lowest 25% of data
- $Q_1 \rightarrow Q_2$: Next 25% of data
- $Q_2 \rightarrow Q_3$: Next 25% of data
- $Q_3 \rightarrow Q_4$: Highest 25% of data

Python Example:

```
import numpy as np
values = [13, 12, 22, 25, 40, 55, 72]
quartiles = np.quartile(values, [0, 0.25, 0.5, 0.75, 1])
Print (quartiles)
```

Percentiles

Divide Data into 100 equals part.

P_{25} = 25 percentile → Same as Q1

P_{50} = 50th percentile → same as median (Q_2)

P_{75} = 75th percentile → same as Q3

Example

P95 = Value below which 95% of data lies.

Python Example

```
import numpy as np
values = [13, 21, 21, 40, 42, 48, 55, 72]
P65 = np.percentile(values, 65)
print(P65)
```

Quick Summary Table

Measure	Divides Data into	Example
Quartile	4 Parts	Q_1 Q_2 Q_3
Percentile	100 parts	P_{25}, P_{50}, P_{75}
Q_1 = P_{25}	first quartile	25% below
Q_2 = P_{50}	Median	50% below
Q_3 = P_{75}	Third quartile	75% below

Statistics - Interquartile Range

IQR is measure of variation - it tells us how spread out middle part of data is.

It is difference between the Q_3 & Q_1

$IQR = Q_3 - Q_1$

It focuses on middle 50% of data which reduces effect of extreme values (outliers).

Understanding Q_1 & Q_3

- Q_1 (First quartile) The value below which 25% of the data falls (separates the bottom 25% from rest)
- Q_3 (Third Quartile): The value below which 75% of data falls (Separate the top 25% from the rest)

Why use IQR?

- It ignores extreme values & focuses on central spread.
- Good For comparing variability b/w diff datasets

Programming Example

```
from scipy import stats
values = [13, 21, 21, 40, 42, 48, 55, 72]
x = stats.iqr(values)
print(x)
```

Statistics - Standard Deviation

Standard deviation is the most common measure of variation.

It shows how far the data values typically are from average (mean).

A small standard deviation → values are close to mean

A large standard deviation → values are more spread out

Standard deviation in a Normal Distribution.

If the data is normally distributed (bell-shaped curve)

- 68.3% of data is within 1 SDs of mean.
- 95.5% of data is within 2 SDs.
- 99.7% is within 3 SDs.

This is called empirical rule.

Population Vs Sample Standard Deviation

We calculate standard deviation differently for:

1- Population (σ) → uses the total no. of observations (N)

2- Sample (s) → uses $N-1$

Formulas

Population SD

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

Sample SD

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N-1}}$$

x_i is each data value

μ is population mean

$\bar{x}$ is sample mean

N is No. of observation

Python Example

Python population

```
import numpy as np
values = [4, 11, 7, 14]
print(np.std(values)
```

Python Sample

```
import numpy as np
values = [4, 11, 7, 14]
print(np.std(values, ddof=1)
```

Data: 4, 11, 7, 14

$$\mu = \frac{4+11+7+14}{4} = 9$$

Difference From mean

(4-9) = -5

(11-9) = 2

(7-9) = -2

(14-9) = 5 & add them

Square each diff ⅀ 58

Divide by N $58/4 = 14.5$

$\sigma = \sqrt{14.5} \approx 3.81$

Statistical Inference

Meaning: Using data from a sample to make conclusions about the whole population

Two main types: Estimation & Hypothesis Testing.

- Estimation
  - We use sample statistics (like average, proportion) to guess the real population value
  - The single best guess is called point estimate.
  - There always some uncertainty in estimate
  - We express this uncertainty with confidence interval or range that probably contain the true value

Example: Dutch people own b/w 3.5 & 6 bikes on avg.

Hypothesis Testing

- ~~(We use sample statistics)~~
- A way to check if a claim about a population is true or false using sample data
- Steps depend on:
  - Data type: Categorial (eg gender) or numerical.
  - Comparison Type:
    - One group only
    - Only / One group vs Other

Example:

- 90% of Australians are left-handed.
- Average dog weight is more than 40kg.
- Doctors earn more money than Lawyers

Probability Distributions

- All statistical inference is based on probability

calculations.

- These use probability distributions (patterns of h data is spread) to make decisions.

Statistics - Normal Distribution

The normal distribution is one of most important patte in statistics.

Many relal-life datasets follow this pattern, where

- Most values are close to average
- Very Few values are extremly high or low

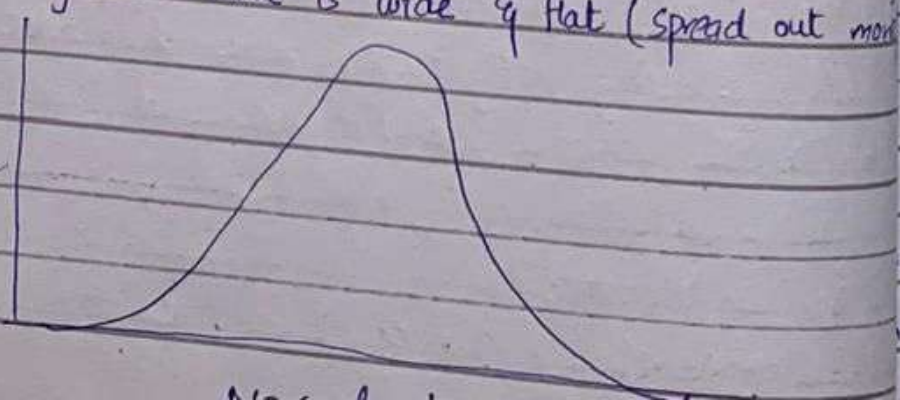
The curve look like a bell - highest in the middle tapering, off on both side.

Main Features:

- Mean, Median, & Mode are same. All in Center
- Symmetrical shape - left & right sides are mirror im
- Unimodal - Only one peak (most common Value)
- Area under curve = 1 (100%) All possible outcom
- Probabilities are shown by area b/w two point on cu

Mean (μ) & Standard Deviation (σ)

- Mean (μ) - center point of curve (average)
- Standard deviation (σ) - how spread out values ar
  - Small σ = curve is tall & narrow (close to mea
  - Large σ = curve is wide & flat (spread out mor



68-95-99.7 rule

In a normal distribution

- About 68.3% of data is within 1σ of the the mean ($\mu - \sigma$ to $\mu + \sigma$)
- About 95.5% of data is within 2σ
- Abaut 99.7% of data is within 3σ

This is useful for quickly estimating probabilities.

Effect of changing Mean or Standard Deviation

- Changing mean (μ) shifts the curve left or right
- Changing standard deviation (σ) changes the width and height of the curve, but the total area stays 1.

Real-Life Examples

Data that is often normally distributed.

- Human Heights
- Test scores in large groups
- Birth weights
- Age of Nobel Prize winners

Probability Distributions & the Normal curve

- A probability distribution decribes how likely each value is
- Simple examples:
  - Coin toss: 50% heads, 50% tails
  - Dice roll: Each faces has a 1/6 chances ($\approx$ 16.67%)
- When we add up many random result (like sums of dice dolls), the distribution starts to look like a normal curve.

- This happens because of central limit Theorem. Many small random effects combine into a bell shaped pattern.

Statistics - Standard Normal Distribution

A normal distribution where:

- Mean (μ) = 0
- Standard deviation (σ) = 1

We can convert (standarized) any normal dataset into this form.

This makes it easier to compare different datasets & calculate probabilities.

Why use standard Normal Distribution?

- Used in:
  - Confidence intervals
  - Hypothesis tests
- Standardizing makes probability calculates simpler, you can use Z-tables or software instedd of complex math.

Z-Values (Z-scores)

- Show how many standard deviations a value is from the mean:

$$Z = \frac{\text{value} - \mu}{\sigma}$$

e.g

Example:

- Mean height in Germany = 170cm
- Standard deviation = 10 cm
- Bob's height = 200 cm
- Bob's is 30 cm above avg = 30/10 = 3 → z=3
- Means bob is 3 standard deviations taller than the average German.

Finding Probabilities (P-Values)

- P-value: probability that a value is less than (or between) a certain z-score.
- Example with z = 3:
  - P(z < 3) = 0.9987 → Bob is taller than 99.87% of germans.
  - P(z > 3) = 1 − 0.9987 = 0.0013 → only 0.13% are taller than Bob.

Probability Between Two Values:

Example:

- Mean = 170cm, σ = 10cm
- Between 155cm & 165cm
  - z(155) = (155 − 170)/10 = −1.5 → P = 6.68%
  - z(165) = (165 − 170)/10 = −0.5 → P = 30.85%
  - Probability between = 30.85% − 6.68% = 24.17%

Finding z from Probability

Example:

- Taller than 90% of german → P = 0.9
- z from table/software = 1.281
- Convert to height

$$x = \mu + z\sigma = 170 + (1.281)(10) = 182.81\text{cm}$$

- You need to be 182.81 cm to be taller than 90% of Germans.

Key Takeaways

- Z-score put all normal data on the same scale (mean = 0, σ = 1)
- They help:
  - Compare across different datasets
  - Calculate probabilities easily
- Z-tables or software (python) are used to find probabilities and critical values.

Statistics - Student's T Distribution

- Look similar to the normal distribution but is used when there is extra uncertainty → usually with small sample size
- Main use: Estimation & hypothesis testing of a population mean

Why It's Different From Normal

- When the sample is small, we don't know the population standard deviation well → extra uncertainty
- The t-distribution
  - Has wider tails than the normal curve (more chance of extreme values)
  - Gets Narrower as sample size increases
- Large sample size → t-distribution becomes almost identical to the standard normal distribution.

Degrees of Freedom (df)

- A number that adjusts for sample size
  $df = n - 1$
- Example: If sample size $n = 30$, then $df = 29$.
- Smaller df → wider t-distribution

When to Use

- Estimating a mean with small sample.
- Hypothesis testing when population standard deviation is unknown.
- Finding critical values & p-values

Finding Probabilities (P-values)

- Use t-table or software
- Example in python

```
import scipy.stats as stats
stats.t.cdf(2.1, 29)
```

Finding T from a probability

```
stats.t.ppt(0.75, 29)
```

Key Points

- Small sample size → use t-distribution instedd of normal
- Degrees of freedom = sample size - 1
- As sample size grows, t-distribution → Normal distribution
- Use it to get t-values & p-values for tests about means

Statistics - Estimation

Estimation is about using a sample to guess a population value.

Point Estimate

- The most likely value of a population parameter
- Calculated from a sample

Depends on data type:

- Categorical data: no. of occurances ÷ Sample size
- Numerical data: mean (average) of sample.

Example:

The average height of people in Denmark (sample) = 180 cm → point estimate.

Note: Estimates are never ~~exist~~/exact they have uncertainty.

Confidence Interval (CI)

- A range of values around the point estimate where the true population value is likely to fall.
- Defined by a lower bound and an upper bound

Example:

Average height = 180cm → CI = 170 - 190 (lower 170, up

Confidence level

- Show how sure we are that the interval cont the true value.
- Common Values: 90% (0.90), 95% (0.95), 99% (0.9
- Higher confidence → wider interval.

Example: (average height CI)

- 90% CI →

95% CI - 170 - 190cm
99% CI - 160 - 200cm

Margin of error
- Distance from the point estimate to the lower and upper bounds.
- CI = point Estimate ± margin of error

Example:
- Point Estimate = 180cm
- MOE = 5 → CI = 175 - 180
- MoE = 10 → CI = 170 - 190
- MoE = 20 → CI = 160 - 200

Steps to Calculate Confidence Interval
- Check the conditions (sample must be random, large enough, and type of parameters.
- Find the point estimate
- Choose the confidence level
- Calculate margin of error
- Form the confidence interval: Point Est ± MOE

Common Parameters:
- Proportion: → for qualitative data (e.g % of voter)
- Mean Values → for numerical data (avg. height)

Statistics - Estimating Population Proportions

- What is a population proportion?
  - The sharp of population that belongs to a certain category.
  - Example: % of Nobel price winners born in USA

Point Esitomate:

The best guess of the population proportion from sample:

$$\hat{p} = \frac{\text{Number in Category}}{\text{Sample Size}}$$

Example: 6 out of 30 Nobel price winner born in USA $\rightarrow \hat{p} = 6/30 = (0.2)(20\%)$

Confidence Interval (CI)

- A range around the point estimate showing where the true proportion likely lies.
- Lower bound = Point estimate - MOE
- Upper bound = PE + MOE
- Example: Point estimate = 0.2, MOE = 0.143
  CI = 0.057 to 0.344
- Interpretation: We are 95% confident that 5.7% to 34.4% of all Nobel Prize winners were born in USA.

Confidence level:

- Shows how sure we are that CI contains the true value.
- Common levels 90%

Calculate CI (conditions):

- Sample is random
- Two possible outcomes: in category / not in category
- Each category has at least 5 samples (for standard)

Margin of Error +-

Margin of Error [MoE) = $Z_{critical}$ × Standard Error (SE)

Standard Error for proportion

$$SE = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Z-critical depends on confidence level (from Z-table)

Steps to calculate CI

- Check conditions
- Find estimate ($\hat{P}$)
- Decide confidence level (90%, 95%, 99%)
- Calculate Standard Error (SE)
- Find Z-critical Value
- Calculate MOE = Z × SE
- CI = $\hat{p} \pm$ ME

```
import scipy.stats as stats
import math
x = 6
n = 30
confidence_level = 0.95
p_hat = x/n
alpha = 1 - confidence_level
z = stats.norm.ppf(1 - alpha/2)
SE = math.sqrt(p_hat*(1-p_hat)/n)
```

ME = z*SE

lower = phat - ME

upper = phat + ME

print(f"CI: [{lower:.3f}, {upper:.3f}]")

Statistics - Estimating Population Means

Kllhat is population Mean?

- The avg. of numerical variable in population
- Example: Avg. age of Nobel prize winners

Point Estimate

The best guess of papulation mean from sample

$\bar{x}$ = sum of samples values / Sample Size

Example: Sample mean age = 62.1 → point estimate

Confidence level (CI)

- A range around the mean showing where the true population mean likely lies
- lower bound = $\bar{x}$ - Margin of Error
- Upper bound = $\bar{x}$ + Margin of Error
- Example: Mean = 62.1, Margin of Error = 5.04 → a = 57.06 to 67.14

Confidence level

- How sure we that CI contains the true mean
- Common levels 90%, 95%, 99%
- Higher confidence → wider interval
- Uses t-distribution for small samples
- Degrees of freedom (df) = n - 1

Conditions to calculate CI

- Sample is random
- Population is normally distributed or,
- Sample size is large enough ($n \geq 30$)

Margin of Error

$ME = t_{critical} \times SE$

Standard Error (SE): $SE = \frac{s}{\sqrt{n}}$

$t_{critical}$ depends on confidence level & df

Programming CI

```
import scipy.stats as stats
import math
x_bar = 62.1
s = 13.46
n = 30
confidence_level = 0.95
alpha = 1 - confidence_level
df = n - 1
SE = s/math.sqrt(n)
t_critical = stats.t.ppf(1 - alpha/2, df)
ME = t_crit * SE
lower = x_bar - ME
upper = x_bar + ME
print(f"CI: [{lower:.2f}, {upper:.2f}]")
```

Statistics - Hypothesis Testing

A formal way to check if a claim about a popu
is true. Example claims:

- Avg. height of people in Denmark > 170cm
- Share of left-handed people in Australia ≠ 10%
- Avg. income of dentists < Avg income of lawye

Null and Alternative Hypothesis

- Null hypothesis (H_0): Claim we assume true initially
- Alternative hypothesis (H_1 or H_a): We want to pr

Example: Claim Avg. height in Denmark > 170 cm

- $H_0 : \mu = 170cm$
- $H_1 = \mu > 170cm$

Only one of H_0 or H_1 can be true

Significance level (α)

- Probability of rejecting a true H_0
- Typical levels: 10% (0.10), 5% (0.05), 1% (0.01)
- Lower α = need stronger evidence to reject H_0

Example $\alpha = 0.05 \rightarrow$ 5 out of 100 times we may wrongly reject H_0.

Test Statistics

- A standardized value calculated from sample
- Converts sample data into probability distribution
- Depends on type of test.
  - Z (Standard Normal) for proportions
  - T (student's t) → for mean

Decision Approaches

a) Critical Value Approach.
- Compare test statistic to critical value.
- Critical value separates rejection region from rest of distribution.
- If test statistic is in the rejection region → reject H_0

b) P-Value Approach
- Compare p-value to α
- P-value = probability of observing the test statistic under H_0.
- If p-value $< \alpha$ → reject H_0
- If p-value $> \alpha$ → Keep H_0

Statistics – Hypothesis Testing a Proportion Mean

- A population mean (μ) is the avg of population
- Hypothesis tests check claims about that mean.
- Example: "The avg age of Nobel prize winners when they received the prize is more than 55

Steps For testing a Mean

1- Check the conditions
- Random sample
- Population is normally distributed or
- Sample size is large enough ($n > 30$)

2- Define the claims
- Null Hypothesis (H_0): mean = claimed value
- Alternative Hypothesis (H_1) = mean $\neq$?, or $<$ claimed value

3- Decide Significance level (α)
- Probability of rejecting a true H_0
- Common: 0.10, 0.05, 0.01
- Lower α → stronger evidence needed to reject H_0.

4. Calculate Test statistic

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$\bar{x}$ = sample mean

H_0 = claimed population (mean) (H_0)

s = sample standard deviation

n = sample size

Python

```
import scipy.stats as stats, math
x_bar, s, mu=null, n = 62.1, 13.46, 55, 30
test_stat = (x_bar, - mu_null) / (s/math.sqrt(n)
print (test_stat)
```

5. Conclude

a) Critical Value Approach

- Right-tailed: find critical T-value at α w
  df = n-1
- Compare TS with CV:
  - TS > CV → reject H_0
  - TS ≤ CV → keep H_0

b) P-value Approach

- P-value = probability of observing TS or
  extreme under H_0.
- Compare P-value with α:
  - $P < \alpha$ → reject H_0
  - $P \geq \alpha$ → keep H_0

6 - Interpretation

- Null hypothesis rejected - sample supports alter
  hypothesis
- P-value = smallest α at which H_0 can be rej
- Smaller P-value → stronger evidence against H_0

Hypothesis's Testing a Proportion

- A population proportion is the share of population in category
- Hypothesis tests check claims about that proportion
- Example: More than 20% of Nobel Prize winners were born in US.

Steps For Testing a Proportion

1- Check the conditions
- Random sample
- Only two elements/category: in category / Not in
- At least 5 in each category (for standard method)
- Example 10 US-born out of 40 - conditions.

2- Define the Claims
- Null Hypothesis (H_0) proportion = claimed value
- Alternative Hypothesis (H_i) proportion $\neq$, $>$ or $<$ claimed value

3- Decide Significance level (α)
- propability of rejecting true H_0
- Common: 0.10, 0.05, 0.01
- Lower α → stronger evidence needed to get H_0

4- Calculate Test Statistic

$$TS = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

$\hat{p}$ = sample proportion
p = claimed proportion (H_0)
n = sample size

Python

```
import scipy.stats as stats, math
n, p = 10, 40, 0.2
hat = x/n
est_stat = (p_hat - P) / math.sqrt( p*(1-P)/n)
int( best_stat)
```

5- Conclude

a) Critical Value Approach

- Right-tailed: Find critical Z-value at
- Compare TS with critical value:
  - TS > CV → reject H_0
  - TS < CV → Keep H_0

b) P-value Approach

- P-value = probability of observing TS or more extreme under H_0
- Compare P-value with α:
  - $P < \alpha$ – reject H_0
  - $P \geq \alpha$ – Keep H_0

6- Interpretation

- Sample does not support claim if H_0 is not rejected.
- True proportion might still be higher, but evidence is insufficient.

Statistics – Z – Table

Z-Distribution

- The z-distribution is a standard normal distribution with mean = 0 & standard deviation = 1
- It tell use the probability that a value is below than a given Z-score
- Negative Z-scores → values below the mean.
- Positive Z-scores → values above the mean

How to read the table:

1- Find the z-score by combining the row (first two digits) and coloum (second decimal)

2- The table value gives the probability (decimal

T - table

A T-table provide critical t-values for different degrees of freedom (df) and significance levels (α). It is used in hypothesis testing to determine whether to reject the null hypothesis, especially for small sample sizes. As df increases, the t-distribution approaches the normal distribution.