



Chapter No.	Chapter Title	Subtopics	Page No.
1	Introduction to Statistics (Statistics ka Taaruf) 📖	- Definition and Scope of Statistics - Role of Statistics in Decision Making	2
2	Importance of Statistics in Data Science	- Role of Statistics in Data Analysis - Applications in Machine Learning and AI	9
3	Measurement	- Types of Measurement Scales (Nominal, Ordinal, Interval, Ratio) - Reliability and Validity in Measurement	21
4	Data Collection	- Primary vs. Secondary Data - Data Collection Techniques (Surveys, Experiments, Observations)	30
5	Descriptive Statistics	- Measures of Central Tendency (Mean, Median, Mode) - Measures of Dispersion (Range, Variance, Standard Deviation) - Data Visualization (Graphs, Charts, Histograms)	49
6	Outliers and Missing Values	- Detection and Treatment of Outliers - Handling Missing Data (Imputation Methods)	92
7	Inferential Statistics	- Population vs. Sample - Hypothesis Testing (Null and Alternative Hypotheses) - Confidence Intervals and p-Values	—
8	Probability	- Basic Probability Concepts (Events, Sample Space, Probability Rules) - Probability Distributions (Normal, Binomial, Poisson)	—



CHAPTER 1

Introduction to Statistics



1.1 Important Definitions



1.1.1 Statistics (Shumariyat)

Statistics ya shumariyat wo science hai jo data ko collect karna, organize karna, analyze karna, aur interpret karna sikhati hai.  

- **Kirdar:** Ye data se patterns aur trends ko samajhne aur future predictions karne mein madad karta hai. Jaise, ek company apne customers ke behavior ko analyze kar ke future sales ka andaza lagati hai.
- **Ahmiyat:** Har field mein, chahe wo business ho, healthcare, education, ya government, statistics ka istemal hota hai decision-making process mein.  



1.1.2 Data Science (Data ka Ilm)

Data science ya data ilm ek interdisciplinary field hai jo statistics, computer science, aur domain expertise ko combine karta hai taake complex data se insights nikale jaa sakein.  


- **Kirdar:** Data science data ko deep level par samajhne aur us par based actionable decisions lene mein madad karta hai.
- **Misal:** Jaise, ek e-commerce website data science techniques ka istemal kar ke customer buying patterns ko samajh kar unhe behtar product recommendations deti hai.  



1.1.3 Machine Learning (Machine ki Seekh)

Machine learning ya machine ki seekh woh hissa hai computer science ka jo machines (computers) ko bina explicitly programmed kiye data se seekhne ki salahiyat deta hai.  


- **Kirdar:** Machine learning models data se patterns aur relationships ko samajh kar predictions ya decisions lene mein madad karte hain.
- **Misal:** Jaise, email services spam emails ko filter karna ya mobile apps jo user ke likhne ke style ko samajh kar text prediction karte hain.  

1.1.4 Artificial Intelligence (Masnoi Zahanat)



Artificial Intelligence ya masnoi zehan woh technology hai jo machines ko insani jaisi sochne aur problem-solve karne ki salahiyat deta hai.  

- **Kirdar:** AI systems complex tasks ko perform kar sakte hain, jaise human language ko samajhna, images ko pehchan'na, aur complex decisions lene mein madad karna.
- **Misal:** Jaise, autonomous driving cars jo traffic ko analyze karte hain aur decisions lete hain, ya virtual assistants jo voice commands ko samajh kar actions perform karte hain.  



1.2 Data Science aur ML mein Statistics ki Bunyadi Ahmiyat

Is chapter mein, hum statistics ke basic concepts ko explore karenge. Statistics ki samajh aapko na sirf data ko behtar tareeqe se analyze karne mein madad degi, balkay ye aapko data se chhupi patterns aur trends ko samajhne ki taqat bhi degi. Ye tamam skills data science aur ML ke field mein aapki buniyad banayengi. 

1.2.1 Data Ki Samajh aur Tahlil (Understanding and Analyzing Data)

Statistics ki madad se, data scientists aur ML engineers raw data ko samajh sakte hain. Aapko maloom hona chahiye ke data kya keh raha hai, us mein patterns kya hain, aur kya khaas baat hai us data mein. Jaise, ek e-commerce website par customers ki khareedari ke patterns ko samajhna, takay behtar marketing strategies banai ja saken.  

1.2.2 Faisla Sazi Mein Madadgar (Aiding in Decision Making)

Faisla sazi ke liye data pe mabni evidence bohot zaroori hai. Statistics aapko yeh sakhti deta hai ke aap bade data sets ko analyze kar ke, mawafiq decisions le sakein. Maslan, ek hospital mein, kis qisam ke ilaaj se zyada behtar nataij aaye, ye jaanne ke liye statistics bohot ahem hai.  

1.2.3 Peshgoiyan aur Andaza Lagana (Predictions and Forecasting)

ML models jo peshgoiyan karti hain, un ka asaas statistics par hota hai. Ye models historical data ko dekh kar future trends ka andaza lagati hain. Jaise, mausam ki peshgoi, stock market analysis, ya sales forecasts. Ye sab statistics ki madad se mumkin hota hai. 🌤️✅📈

1.2.4 Machine Learning Models ki Bunyad (Foundation for ML Models)

ML models, jaise ke linear regression, decision trees, ya neural networks, mein statistics ki concepts ka istemal hota hai. Ye models data ko samajhne aur us se seekhne ke liye statistics ke principles pe rely on karte hain. Aapko in models ko effectively train karne aur unki accuracy ko behtar banane ke liye statistics ki achi samajh honi chahiye. 🤖📚

1.2.5 Risks aur Anomalies ki Pehchan (Identifying Risks and Anomalies)

Kisi bhi data set mein risks ya anomalies ko pehchanne ke liye statistics bohot zaroori hai. Jaise, banking sector mein fraud detection ya manufacturing mein product failures ki pehchan. Is tarah, statistics se na sirf faide hasil hote hain balkay nuqsanat se bachne mein bhi madad milti hai. 🏢🔍

1.2.6 Data Integrity aur Quality Control (Ensuring Data Integrity and Quality)

Data ko sahi tareeqe se collect karna aur uski quality ko yaqeeni banane ke liye bhi statistics istemal hota hai. Ye ensure karta hai ke aap jo conclusions nikal rahe hain, wo dependable aur valid hain. Jaise, research studies ya quality assurance processes mein data ko verify karna. 📄🔍

1.2.7 Mushkil Masail ka Hal (Solving Complex Problems)

Aakhir mein, statistics aapko complex masail ko asaan banane mein madad karta hai. Data se deep insights nikalna aur un insights ko asaan fahmi se samjhana, statistics ke baghair mumkin nahi. Chahe wo healthcare, finance, education, ya technology ho, har jagah statistics ki ahmiyat apni jagah qaim hai. 🌐🔑

In summary, statistics plays a pivotal role in data science and ML. It is not only essential for understanding and analyzing data but also for making informed decisions, predicting future trends, building, and refining ML models, identifying risks and anomalies, ensuring data integrity, and solving complex

problems. This foundational knowledge is crucial across various domains, including e-commerce, healthcare, finance, and technology.

1.3 Basic Concepts: Mean, Median, Mode

1.3.1 Mean (Ausat):

Mean, yaani ausat, kisi data set ke tamam numbers ka total kar ke, unki tadad se taqseem karne ka amal hai. Ye aik basic lekin bohot ahem measure hai jo data set ki “central tendency” ko batata hai.

- **Istemaal:** Mean ko rozmarra ke kaamon jaise monthly ghar ke kharche, ya school ke class mein students ke average marks calculate karne mein istemal kiya jata hai.
- **Ehmiyat:** Data science mein, mean ka istemal kisi bhi data set ke general trend ko samajhne ke liye hota hai. Masalan, kisi website par average user spend time ya ek factory mein average production cost.
- **Limitation:** Agar data mein outliers (bohot zyada ya kam values) hain, to mean distort ho sakta hai. Is liye sometimes median ko zyada reliable samjha jata hai.

1.3.1.1 Median (Darmiyani Qeemat):

Median woh value hoti hai jo data set ko do hisson mein taqseem karti hai, yani ke aadhe numbers is se kam aur aadhe zyada hote hain. Agar data set mein numbers ki tadad odd hai, to beech ka number median hota hai; agar even hai, to darmiyani do numbers ka average median hota hai.

- **Istemaal:** Median ka istemal housing prices, salaries, aur isi tarah ke data sets ke liye kiya jata hai, jahan outliers ki mojudgi mean ko distort kar sakti hai.

- **Ehmiyat:** Median, data set mein mojud extremes (bohot zyada ya kam values) ki wajah se mean ke distortion se bachata hai, aur is liye kai dfa zyada accurate picture pesh karta hai.

1.3.2 Mode (Aam Tareen Qeemat):

Mode wo value hoti hai jo kisi data set mein sab se zyada baar aati hai. Ye bata sakta hai ke kis item, value ya category ko log zyada pasand karte hain ya zyada istemal karte hain.

- **Istemaal:** Mode ko fashion industry mein popular clothing sizes, education sector mein sab se common grades, ya marketing mein sab se zyada bikne wale products ko identify karne ke liye istemal kiya jata hai.
- **Ehmiyat:** Kuch cases mein, jaise customer preferences ya voting patterns, mode sab se zyada informative statistic ho sakta hai.

In tamaam concepts - mean, median, aur mode - ki samajh data science aur ML mein deeply important hai. Ye statistics ke basic tools hain jo data ko analyze karne, us mein patterns aur trends ko identify karne, aur data-driven decisions lene mein madad karte hain. Har concept ki apni jagah aur ahmiyat hai, aur kisi bhi data set ko samajhne ke liye inka istemal zaroori hota hai. Ye concepts na sirf data science ke professionals ke liye, balkay aam logon ke liye bhi, unke daily life decisions mein madadgar sabit ho sakte hain.

1.4 Practical Applications of Statistics in Daily Life

1.4.1 Khareedari aur Consumer Behavior (Shopping and Consumer Behavior)

- **Example:** Jab aap online shopping karte hain, to product ratings aur reviews mein statistics ka istemal hota hai. Average rating (mean) ye batata hai ke zyadatar khareedar product se kitne mutma'in hain. Isi tarah, sale items ki popularity ko mode ke zariye samjha ja sakta hai - kaunsa size ya color sab se zyada bik raha hai.
- **Ehmiyat:** Ye information consumers ko behtar decisions lene mein madad karta hai aur retailers ko customer preferences samajhne mein.

1.4.2 Sehat aur Tandruti (Health and Fitness)

- **Example:** Aapke smartphone ya fitness tracker mein daily steps, dil ki dhadkan, aur neend ke patterns ko track karne ke liye statistics ka istemal hota hai. Yahan median aur mean ye batate hain ke aapka average performance kya hai over time.
- **Ehmiyat:** Ye data aapko apne health goals ke mutabiq adjust karnay mein madad karta hai, masalan exercise badhana ya neend ke auqaat ko behtar banana.

1.4.3 Taleem (Education)

- **Example:** Schools aur colleges mein students ke grades aur test scores analyze karne ke liye statistics istemal hota hai. Teachers mean aur median ka istemal kar ke class ki overall performance ko samajhte hain, aur mode se ye dekhte hain ke zyadatar students kis grade range mein hain.
- **Ehmiyat:** Is se teachers ko ye samajhne mein madad milti hai ke kis subject ya topic par students ko zyada focus ki zaroorat hai.

1.4.4 Mausam ki Peshgoi (Weather Forecasting)

- **Example:** Mausam ki peshgoi ke liye meteorologists past weather data ka analysis karte hain. Yahan statistics ki madad se patterns aur trends ko samajhna possible hota hai, jaise average temperature, barish ki miqdar, ya hawa ki raftaar.
- **Ehmiyat:** Ye information logon ko apne rozmarra ke plans banane, agriculture ke decisions lene, ya emergency situations ke liye tyar hone mein madad karti hai.

1.4.5 Karobar aur Marketing (Business and Marketing)





- **Example:** Companies apne products ki sales, customer feedback, aur market trends ko samajhne ke liye statistics ka istemal karti hain. Yahan data analysis se wo samajh sakti hain ke kaunse products zyada popular hain (mode), average sales kya hai (mean), aur sales mein variation kitna hai (standard deviation).
- **Ehmiyat:** Ye insights businesses ko unki strategies ko behtar banane, naye products develop karne, aur customer satisfaction ko barhane mein madad karte hain.


In tamaam misalon se ye waziha hota hai ke statistics sirf kitabon tak mehdood nahi hai, balkay hamari rozmarra ki zindagi mein deep impact rakhta hai. Ye na sirf professionals balkay aam logon ko bhi unke decisions mein madad karta hai, chahe wo shopping ho, health, education, ya business se related ho. Statistics ke concepts ko samajh kar, hum apni life ke mukhtalif pehluon ko behtar tareeqe se manage kar sakte hain aur zyada informed decisions le sakte hain.






CHAPTER 2

IMPORTANCE OF STATISTIC

Is chapter mein hum dekhenge ke kaise statistics data science ke har pehlu mein bunyadi kirdar ada karta hai.  

Data science aaj ke daur ka ek intehai ahem shoba ban chuka hai. Ye technology, karobar, sehat, aur bahut se doosre shobajaat mein inqilabi tabdeeliyan la raha hai.   Har jagah data ka istemal hota hai, aur statistics woh zariya hai jis se hum is data ko samajhte hain, analyze karte hain, aur us se meaningful nateeja nikalte hain.  

Data science ka taalluq mukhtalif shobajaat se hai jisme statistics ek ahem hissa hai. 

Har field, chahe wo marketing ho, finance, ya healthcare, data se bharpoor hai. Statistics ke zariye, hum is data ko decode kar sakte hain aur is se useful maloomat hasil kar sakte hain.    Statistics ke baghair, ye data sirf numbers ka majmua lagta hai, lekin iski madad se, ye numbers kahaniyan sunane lagte hain. Aik effective data scientist ke liye statistics ki mazboot samajh bohot zaroori hai.  

Is chapter mein, hum statistics ke kuch basic tools aur techniques ko explore karenge, jo har data scientist ko maloom hone chahiye. Hum real-life examples ke zariye in concepts ko samjhenge, takay aap dekh saken ke yeh kaise aapke rozmarra ke kaamon mein madadgar sabit ho sakte hain.



Chapter Outline in English

This introduction sets the stage for the chapter, highlighting the fundamental role of statistics in data science across various fields. It emphasizes the transformative power of data when interpreted through statistical methods and prepares the reader to explore basic tools and techniques with practical, real-life examples.

2.1 Statistics to Reedh ki Haddi hy bhai!

Jab hum data science ki baat karte hain, to statistics uski bunyad hoti hai. Aik mazboot reedh ki haddi ke baghair jis tarah aik jism sambhal nahi sakta, isi tarah, statistics ke baghair data science ka koi wujood nahi hota. 📦🔍

Misal: Data sets ko safai aur tayyari ke liye statistical methods ka istemal, jaise ek business ki sales data ko analyze karna.

Sochein ke aap ek business run kar rahe hain aur aapko apni sales ki performance ko samajhna hai. Yahan statistics aapki madad karta hai. 📈🏢
Aap pehle data ko “clean” karte hain, yaani ke kisi bhi ghair zaroori ya galat maloomat ko hata dete hain. Is ke baad, aap statistical methods jaise mean, median, aur mode ka istemal kar ke apne sales ke data ko samajhte hain. 📊🔍

Is se aapko pata chalta hai ke aapki average sales kya hain, sales mein sab se zyada aur kam hone wale din kaun se hain (mode), aur sales data mein variation kitna hai (standard deviation). Ye maloomat aapko apne karobar ke liye behtar faislay karne mein madad deti hai, jaise ke stock management, marketing strategies, aur customer preferences ko samajhna. 📈📁

Statistics ke zariye, data jo pehle sirf numbers ka dhair lagta tha, ab aapke liye kahaniyan sunata hai, aur ye kahaniyan aapko apne karobar ko agay barhane ke liye zaroori raaste dikhate hain. 📖🚀

Outline

Is section mein, statistics ke data science mein kirdar ko ujagar kiya gaya hai, jisme humne dekha ke kaise statistics ek business ki sales performance ko samajhne mein madadgar sabit hota hai. Ye sirf ek misal hai; aisi hi kai aur misalon se statistics data science ke har pehlu mein apni ahmiyat sabit karta hai.

2.2 Descriptive Statistics (Data ko Smajhna) 📊🔍


Descriptive statistics (mean, median, mode, range, variance, standard deviation) ke zariye data sets ko samajhna. 📊🔍

Descriptive statistics wo basic tools hain jo humein kisi bhi data set ki pehli aur sab se zaroori jhalak faraham karti hain. Ye aalat humein data ke “general behavior” ko samajhne mein madad karti hain. 🧩🔍


- Mean (Ausat): Ye batata hai ke data set mein mojud tamam values ka average kya hai.

- Median (Darmiyani Qeeymat): Ye data set ko darmiyan se taqseem karta hai, takay hum samajh sakein ke aam tor par values kitni hain.
- Mode (Aam Tareen Qeeymat): Ye wo value hai jo data set mein sab se zyada baar aati hai.
- Range: Ye fark batata hai sab se kam aur zyada value ke darmiyan.
- Variance aur Standard Deviation: Ye dono measures data mein variation ya diversity ko show karte hain.

Misal k tor per: Kisi retail karobar mein customers ke data ko analyze karna takay khareedari ke patterns ko samjha ja sake.

Sochein ke aap ek retail store ke malik hain. Aapke paas har roz ke customers ke data hain: kya cheezein khareedi gayi hain, kitne paise kharch hue hain, aur customers ki tadad kya thi. 

- Mean: Aap rozana ki average sales calculate kar sakte hain.
- Median: Ye dekhte hain ke aam tor par ek customer kitna kharch karta hai.
- Mode: Aap dekh sakte hain ke kaunsi product sab se zyada bikti hai.
- Range aur Standard Deviation: Ye measures aapko batate hain ke aapki sales mein din ba din kitna variation hota hai.

Ye tamam measures aapko apne customers ki khareedari ke patterns ko samajhne mein madad karte hain. Is se aap apne stock ko manage kar sakte hain, marketing strategies tayyar kar sakte hain, aur customer satisfaction ko behtar bana sakte hain. 

Outline

Is section mein, humne dekha ke kaise descriptive statistics ke basic tools ki madad se aap apne karobar ki better understanding develop kar sakte hain, khas tor par retail sector mein. Ye tamam tools data science mein buniyadi aur zaroori hain aur kisi bhi data-driven decision making process ke liye nihayat ahem hain.

2.3 Inferential Statistics (Sample se Population ki pesh goi karna)

Inferential Statistics ke zariye sample data se poori population ke liye peshgoiyan aur generalizations nikalne ka tareeqa bayan karna.

Inferential statistics woh jadu hai jo humein chote namune (sample) se badi population ke bare mein qayasiyaat (inferences) nikalne ki ijaazat deta hai. Yeh data science mein anjaam diye jaane wale tajziyaat ka aik bohot ahem hissa hai.



Misal k tor pe: Siyasi Election mein Voting Behavior ki Peshgoi

Sochein ke aap ek political analyst hain aur aapko aane wale elections ke nataij ka andaza lagana hai. Aap ek chota sample lete hain - matlab chand so ya hazaar logon se unki raaye maloom karte hain. 📦PK

- **Namunay se Peshgoi (Predicting from a Sample):** Aap is chote group ke data ka tajzia kar ke, poori voting population ke rujhanat ka andaza lagate hain.
- **Confidence Interval:** Is se aap ye jaan sakte hain ke aapki peshgoi kitni reliable hai.
- **Margin of Error:** Ye aapko batata hai ke aapki peshgoi mein kitni Ghati ho sakti hai.

Ye sab kuch inferential statistics ke concepts ke zariye kiya jata hai. Aapke sample ka size aur diversity, aapki peshgoi ki accuracy ko behtar bana sakti hai.




Inferential statistics ki madad se aap nafaqat elections balkay market research, public opinion, aur kisi bhi field ke trends ko samajh sakte hain aur unke baare mein reliable predictions kar sakte hain. Ye data science mein decision making ko aik scientific aur quantifiable base faraham karta hai. 🧠💡

Outline


Is section mein, inferential statistics ke zariye kaise chote data samples se badi population ke liye meaningful conclusions aur predictions nikale ja sakte hain, is ki wazahat ki gayi hai. **Siyasi elections ka example is baat ko samajhne ka aik aasan aur dilchasp tareeqa hai, jo humein dikhaata hai ke data science practical zindagi ke decisions mein kaise madadgar sabit ho sakta hai.**

2.4 Faisla Sazi mein Probability ka Kirdar 📊


Probability ke zariye data science mein predictions aur decisions lene mein uncertainty ko manage karne ka tareeqa bayan karna.


Faisla sazi, khaas tor par jab uncertainty ka samna ho, mein probability aik aham kirdar ada karti hai. Ye humein batata hai ke kisi event ke hone ke imkaana kitne hain. Data science mein, is ka istemal kisi bhi anjaam ya nateeje ki likelihood ko samajhne ke liye hota hai. 

Misal k tor pe: Mausam ki Peshgoi aur Zaraat

Sochein ke aap ek kisan hain ya agricultural planning kar rahe hain. Aapko fasal lagane ka sahi waqt aur qisam ka faisla karna hai, jo mausam par mabni hota hai. 

- Probability Models: Mausam ki peshgoi ke liye meteorologists mukhtalif types ke data (temperature, barish, hawa ki raftar) ko analyze karte hain aur probability models ka istemal kar ke mausam ke trends ka andaza lagate hain.
- Faisla Sazi: Is information ki bunyad par, aap fasal lagane ya fasal bachane ke liye zaroori hifazati iqdamat kar sakte hain. Yeh aapko nuqsan se bacha sakta hai aur munafe ko maximise karne mein madad karta hai.

Is tarah, probability ki madad se, hum faisle lenay mein ziada informed aur scientific approach apna sakte hain. Ye sirf agriculture tak mehdood nahi, balkay karobar, sehat, engineering, aur mali maamlaat jese shobajaat mein bhi istemal hota hai. 



Probability humein sirf future predictions hi nahi deti, balkay ye humein risk management aur resource allocation mein bhi rehnumai faraham karti hai. Is ke istemal se, data science mein faisle lene ka amal ziada baasaroorat aur effective ban jata hai. 

Outline



Is section mein, probability ke data science mein faisle lene ke amal mein kirdar ko ujagar kiya gaya hai, aur ye dikhaya gaya hai ke kaise ye kisi bhi field, jaise ke agriculture, mein faisle lene ke liye scientific aur quantifiable base muhayya karta hai. Ye section humein batata hai ke kaise data science real-world problems ko solve karne mein madadgar hai aur kaise ye humare faisle ko behtar aur informed bana sakta hai.

2.5 Research mein Hypothesis Testing



Hypothesis testing ka tareeqa aur iski data science mein ahmiyat ko samjhana.

Hypothesis testing, research aur data analysis ka aik ahem juz hai. Is process mein hum aik hypothesis (qayaas) banate hain aur phir data ke zariye is ki janch karte hain ke ye sahi hai ya nahi. Ye method scientific research aur data-driven decision-making mein nihayat ahem hai.  



Misal k tor pe: Online Marketing Campaigns mein A/B Testing

Sochein ke aap ek digital marketer hain aur aapko ye dekhna hai ke aapke do different marketing campaigns mein se konsa zyada effective hai.  

- Hypothesis (Qayaas): Aapka qayaas ho sakta hai ke Campaign A, Campaign B se zyada customer engagement laega.
- A/B Testing: Aap kuch waqt ke liye dono campaigns ko chala ke dekhte hain aur unke results ko measure karte hain.
- Data Analysis: Phir aap statistical methods ka istemal kar ke ye dekhte hain ke kya aapka hypothesis sahi tha. Kya waqai Campaign A ne zyada engagement diya hai?

Is tarah ke tests se aapki marketing strategies zyada effective ho sakti hain, aur aap apne resources ko behtar tareeqe se allocate kar sakte hain.  

Hypothesis testing sirf marketing tak mehdood nahi, balkay scientific research, product development, aur health studies jese shobajaat mein bhi istemal hoti hai. Masalan, ek new medicine ki effectiveness ko test karne ke liye bhi hypothesis testing ka sahara liya jata hai.



Is process ki madad se, hum data ko behtar samajh sakte hain aur apne faislay zyada data-driven aur reliable bana sakte hain. Ye humein ye bhi batata hai ke kab hamare data mein significant changes aaye hain jo ke hamari soch ya strategy ko badalne ke liye kafi ho.  

Outline



Is section mein, hypothesis testing ke process aur uske applications ko detail mein samjhaya gaya hai, ye dikhate hue ke ye kaise data science aur research mein ahem kirdar ada karta hai. Ye method na sirf marketing campaigns ko behtar banane mein madad karta hai, balkay scientific research aur product development mein bhi nihayat zaroori hai. Ye humein bataata hai ke kaise data ke zariye informed decisions liye ja sakte hain.

2.6 Trends aur Peshgoi ke Liye Regression Analysis

Regression analysis ka istemal kar ke variables ke darmiyan relationships ko samajhna aur mustaqbil ke trends ya nataij ki peshgoi karna.



Regression analysis, ya wapasad ka tajzia, ek aham statistical tool hai jo data science mein variables ke darmiyan rishte ko samajhne aur future predictions karne ke liye istemal hota hai. Ye method humein batata hai ke kaise ek ya zyada variables, dusre variable(s) ko kaise mutassir karte hain.  



Misal k tor pe: Real Estate Prices ki Prediction

Sochein ke aap ek real estate analyst hain aur aapko property ke prices ka future trend samajhna hai.  

- Variables: Yahan pe variables ho sakte hain jese ke property ka size, location, qareebi sahuiliyat (schools, hospitals, etc.).
- Linear Regression: Aap linear regression model ka istemal kar ke in variables aur property ke prices ke darmiyan rishte ko samajh sakte hain.
- Prediction: Is model ki madad se, aap ye predict kar sakte hain ke mustaqbil mein prices kis tarah behave karenge, based on current trends.

Ye sirf ek misal hai. Regression analysis ko har qisam ke data sets par lagaya ja sakta hai, chahe wo business ho, sehat, education ya kisi aur field ka ho.

Masalan, ek company apni product ki sales ko predict karne ke liye, ya ek hospital patient ki recovery rate ko samajhne ke liye regression analysis ka istemal kar sakta hai.  

Regression analysis ki taqat yeh hai ke ye humein complex data ko simplify kar ke samajhne aur us par based decisions lene ki sahulat deta hai. Is se hum accurate aur reliable predictions kar sakte hain jo hamare karobar ya research ko behtar bana sakte hain.  

Outline

Is section mein, regression analysis ke istemal aur uske faide ko ujagar kiya gaya hai, ye dikhate hue ke kaise ye tool data science mein variables ke darmiyan rishto ko samajhne aur future trends ki peshgoi mein madadgar hai. Regression analysis ki versatility aur utility ko real estate ki misal ke zariye samjhaya gaya hai, jo ye batata hai ke ye kaise mukhtalif scenarios mein istemal kiya ja sakta hai.

2.7 Machine Learning aur Jadeed Statistics 🤖

Machine learning mein advanced statistical methods ka istemal aur unka kirdar.

Machine learning (ML), jo ke artificial intelligence (AI) ka aik hissa hai, modern statistics ke jadeed tareeqon par mabni hai. ML algorithms data se sikhne aur us par mabni predictions ya decisions lene ke liye statistics ki jadeed techniques ka istemal karte hain. 🧠💻

Streaming Services ke Liye Recommendation Systems

Ek aam misal hai streaming services jaise Netflix ya YouTube ke recommendation systems. 🎬🌐

- Data Collection: Ye services pehle aapke viewing history aur preferences collect karti hain.
- Statistical Analysis: Phir, advanced statistical models ka istemal kar ke ye analyze karte hain ke aap kis tarah ke content ko prefer karte hain.
- ML Algorithms: Inke zariye, system aapko woh movies ya videos suggest karta hai jo aapke interests se match karte hain.

Yeh sirf ek misal hai. ML aur statistics ka istemal aur bhi bohat se shobajaat mein hota hai, jaise ke facial recognition systems, autonomous vehicles, aur health diagnosis systems. 🚗👤

ML ki algorithms, jaise neural networks, decision trees, aur random forests, statistics ke complex models par mabni hain. Ye algorithms data se patterns ko samajhne aur mustaqbil ke liye accurate predictions karne mein madad karte hain. 🌲🌲

Is section ka maqsad yeh hai ke samjhaya jaye ke kaise modern ML techniques, traditional statistics ke concepts par mabni hain aur kaise ye dono mil kar data science ke field ko revolutionize kar rahe hain. ML aur statistics ki is combination se humein data ko deeper level par samajhne aur us par based smarter decisions lene ki taqat milti hai. 🚀💡

Outline

Is section mein, machine learning aur statistics ke darmiyan gehre rishte ko ujagar kiya gaya hai. Yeh section ye batata hai ke kaise ML ki advanced techniques, traditional statistical methods par mabni hain aur kaise ye dono mil kar data ko analyze karne aur us par mabni decisions lene ke tareeqon ko behtar

bana rahe hain. Streaming services ki recommendation systems ko misal ke taur par istemal kiya gaya hai, jo ye dikhata hai ke ye amalgamation kaise practical applications mein istemal hota hai.

Absolutely! Let's expand Section 8 in Roman Urdu, incorporating emojis for better engagement:

2.8 Real-life Data ki complexities aur Statistical Ahmiyat



Asal dunya ke data ki complexities aur statistical significance ki ahmiyat ko samjhana.

Asal duniya ka data aksar pechida aur unpredictable hota hai. Is mein noise, outliers, aur incomplete information hoti hai. Statistics ki madad se, hum is pechidagi ko samajh sakte hain aur accurate conclusions nikal sakte hain.



Medical Trials mein Nai Dawai ki Effectiveness ka Ta'een

Sochein ke ek nai dawai ki testing ki ja rahi hai. 🏥💊

- Data Collection: Medical trials mein mareezon par dawai ke asraat ko record kiya jata hai.
- Statistical Analysis: Is data ko analyze karne ke liye advanced statistical methods ka istemal hota hai.
- Statistical Significance: Ye determine karta hai ke kya dawai ka asar real hai ya phir chance ki wajah se.
- Conclusions: Agar results statistically significant hain, to scientists ye conclude kar sakte hain ke dawai effective hai.

Is tarah ke analysis se, hum nafaqat new treatments ko develop karne mein madad karte hain, balkay patient safety ko bhi yaqeeni banate hain. 🧪✅

Ye sirf ek misal hai. Statistical significance har qisam ke research aur data analysis mein zaroori hota hai, chahe wo business ho, environmental studies, ya social sciences. Ye humein batata hai ke hamare findings reliable hain aur in par based decisions lene mein madad karta hai. 📝📊



Is hisse mein, asal duniya ke data ki complexities aur statistical significance ki ahmiyat ko ujagar kiya gaya hai. Ye bataata hai ke kaise statistics humein complex data sets ko samajhne aur us par based reliable aur accurate decisions lene mein madad karta hai, khaas tor par jab baat medical trials jaise sensitive aur ahem mauzoon ki ho.

Outline







Is section mein asal duniya ke data ki complexities aur statistical analysis ki ahmiyat ko ujagar kiya gaya hai, khaas taur par medical trials ki misal ke zariye. Ye section ye samjhaata hai ke kaise statistics humein complex aur unpredictable real-world data ko samajhne aur us par mabni important decisions lene mein madad karta hai. Ye dikhata hai ke statistics ki ahmiyat sirf theoretical nahi, balkay practical aur real-world applications mein bhi bohot zyada hai.

2.9 Mustaqbil mein Data Science ke Andar Statistics ka role



Statistics ke mustaqbil ke role aur new techniques aur approaches ka jaaiza.





Jaise jaise technology tezi se evolve ho rahi hai, waise hi statistics ka role bhi data science mein badal raha hai. Aane wale waqt mein, hum expect kar sakte hain ke statistics aur bhi advanced aur sophisticated tools aur techniques ka istemal karega.  

Advanced Techniques aur Approaches



- Big Data Analysis: Data ki matra mein izafa ke sath, complex statistical models aur algorithms ki zarurat hogi.  
- Artificial Intelligence aur Machine Learning: In fields mein statistics ke advanced forms ki demand barhegi, jaise deep learning models aur predictive analytics.  
- Real-time Data Processing: Jaise 5G aur IoT devices zyada common hote ja rahe hain, real-time data analysis ke liye statistics ke tez aur dynamic models ki zarurat hogi.  

Data Science ke Mustaqbil ka Role

- Ethics aur Transparency: Data privacy aur ethical use ke sawalaat ke jawab dene mein statistics ka kirdar ahem hoga.  

- Customization aur Personalization: Businesses aur services ke liye customer ki zarurat ke mutabiq tailor kiye gaye solutions provide karne ke liye statistics bohot zaroori hoga.  
- Predictive Health Care: Medical field mein, personalized treatment plans aur disease prediction ke liye statistics ka istemal barhega.  

Continuous Learning ka Ahmiyat



Aakhri mein, data science ke students aur professionals ke liye statistics mein continuous learning bohot zaroori hogi. Jaise jaise field evolve karega, naye tools aur techniques seekhne ki zarurat hogi.  





Outline





Is hisse mein, mustaqbil mein data science aur statistics ke evolving role ko explore kiya gaya hai. Yeh section ye batata hai ke kaise statistics ke new approaches aur techniques data science ko aur bhi advanced bana rahe hain. Big data, AI, ML, aur real-time data processing jese topics ko cover karte hue, ye hissa ye samjhaata hai ke statistics ka role sirf badal nahi raha, balkay aur bhi zyada ahem ho raha hai, khas taur par ethics, customization, aur healthcare jese shobajaat mein. Ye section ye bhi emphasize karta hai ke data science mein continuous learning ki kitni ahmiyat hai.

2.10 Conclusion of the chapter



Is chapter ka ikhtitam karne se pehle, chaliye statistics aur data science ke darmiyan rishte ko yaad karein.

Humne dekha hai ke statistics data science ke har pehlu mein kaise zaroori hai. Yeh sirf numbers ko samajhne ka zariya nahi, balkay yeh humein data ki deeper understanding faraham karta hai.  

- Descriptive aur Inferential Statistics: Humne sikha ke kaise descriptive statistics humein data sets ko samajhne mein madad karta hai, aur inferential statistics humein chote namunay se badi population ke baare mein predictions karne mein madad karta hai.  
- Probability aur Hypothesis Testing: Probability se humne sikha ke kaise uncertainty ke sath smart decisions liye ja sakte hain, aur hypothesis testing se humne sikha ke kaise data ke zariye hamare qayas ko test kiya ja sakta hai.  

- Regression Analysis: Regression analysis ne humein bataya ke kaise different variables ke darmiyan relationships ko samjha ja sakta hai aur future predictions ki ja sakti hain.  
- Machine Learning: Aur aakhir mein, machine learning aur statistics ke combination ne humein bataya ke kaise data se sikhne aur us par mabni smarter decisions lene ka process aur bhi advanced aur effective ho sakta hai.  

Akhri Jumlay (is chapter k)

Is bab ko parhne ke baad, umeed hai ke aapko statistics aur data science ke beech ka gehra rishta samajh aaya hoga, aur aapko yeh bhi andaza ho gaya hoga ke kaise statistics aapke rozmarra ke decisions aur professional life mein madadgar sabit ho sakta hai. Data science ka safar sirf shuru hua hai, aur is mein statistics ka kirdar hamesha se ahem rahega.  

CHAPTER 3

MEASUREMENT

Jab hum data science ki baat karte hain, to pemaish ya measurement ka role bohot ahem hota hai. Is chapter mein, hum dekheinge ke kaise data science mein pemaish ke mukhtalif pehlu hote hain aur ye kyun zaroori hai. 📊📈

3.1 Importance of Measurement

Data science mein, measurement ka role bohot ahem hota hai. Agar hum kisi cheez ko measure nahi kar sakte, to hum us cheez ko analyze bhi nahi kar sakte. Is liye, measurement bohot zaroori hai.

Data ki Sifaat ka Taayun (Determining Data Quality):

Measurement se hum data ki accuracy, reliability, aur validity ka taayun karte hain. Yeh samajhna zaroori hai ke aap jo data use kar rahe hain wo qabile bharosa aur durust hai. 📊🔍

Misal ke taur par, agar aap ek research kar rahe hain jisme aap logon se unke khane ke adat ke bare mein sawalat pooch rahe hain. Yahan, aapko yeh dekhna hoga ke jawab kitne sahi hain, kya log sach bol rahe hain, ya unke jawab mein kuch bias to nahi hai. 🗣️📝

Measurement Scales aur Unka Istemal:

Har type ka data alag tarah se measure kiya jata hai aur iske liye alag scales hoti hain. Ye scales hain: nominal, ordinal, interval, aur ratio. Har ek ka apna unique faida aur istemal hai. 📏✂️

Jaise, nominal scale mein hum cheezon ko naam se pehchante hain, masalan, kisi survey mein mard ya aurat ke options. Ordinal scale mein, hum order ya darja bandi karte hain, jaise, hotel reviews mein stars. Interval scale mein koi fix zero point nahi hota, masalan, temperature. Aur ratio scale mein fixed zero point hota hai, jaise, kisi cheez ka wazan ya lambai. ⚖️🌡️

3.2 Scales/Levels of Measurement 📏📊

Pemayesh Ke Paimane ya Darje

Measurement scales, ya pemayesh ke paimane, data science mein data ko categorize aur analyze karne ka ek basic framework faraham karte hain. Har scale data ki mukhtalif kism ke properties ko measure karta hai aur iska apna unique use hota hai.

3.2.1 Nominal Scale (Nam Ka Paimana):

- **Definition:** Nominal scale sab se basic level ka measurement scale hai. Is mein data ko categories mein divide kiya jata hai, lekin in categories mein koi numeric order ya value nahi hoti.
- **Example:** Jaise, a survey mein logon ki nationality ya unka profession poocha jata hai. Pakistani, Indian, Teacher, Doctor, etc., are examples of nominal data.
- **Use in Data Science:** Data sorting aur categorization ke liye istemal hota hai, jaise customer segmentation ya demographic studies mein. 🌍 👤

3.2.2 Ordinal Scale (Tarteebi Paimana):

- **Definition:** Ordinal scale mein, data categories mein hota hai, lekin in categories mein ek specific order ya sequence hoti hai.
- **Example:** Jaise, ek survey mein logon se unki education level ke bare mein poocha jata hai: Matric, Intermediate, Bachelor's, Master's. Yahan, har category ka ek specific order hai.
- **Use in Data Science:** Data ko rank ya order mein rakhne ke liye istemal hota hai, masalan, customer satisfaction surveys mein. 😊 📊

3.2.3 Interval Scale (Waqt Ke Faslay Ka Paimana):

- **Definition:** Interval scale numeric values ke sath aata hai, aur is mein equal intervals ya differences hote hain, lekin iska koi true zero point nahi hota.
- **Example:** Temperature Celsius ya Fahrenheit mein. Yahan, 0 degrees ka matlab ye nahi ke koi temperature nahi hai; ye sirf ek point hai scale par.
- **Use in Data Science:** Data mein variations ko samajhne aur analyze karne ke liye, jaise climate change studies. 🌡️ 🌍

3.2.4 Ratio Scale (Tanasubi Paimana):

- **Definition:** Ratio scale interval scale ki tarah hota hai lekin is mein ek absolute zero point hota hai.
- **Example:** Distance (meters ya kilometers mein), weight (kilograms), ya age (saal mein). Yahan, zero ka matlab hai ke us cheez ka non-existence hai.
- **Use in Data Science:** Quantitative analysis aur scientific calculations ke liye, jaise physics ya engineering applications. 📊 ⚖️

Outline

Is section mein, measurement ke mukhtalif scales ya darje aur unke data science mein istemal ko tafseel se samjhaya gaya hai. Har scale ke unique features aur examples ko include kiya gaya hai, taake readers ko clear understanding ho ke kaise ye scales data ko samajhne aur analyze karne mein madadgar hain. Ye section data science practitioners ke liye important hai kyun ke ye unhe guide karta hai ke kis tarah ke data ko kaise handle kiya jaye aur kis tarah ke analysis ke liye konsa scale behtar hai.

Scale Type	Definition	Examples	Usage in Data Science
Nominal Scale	Categories without any numeric order. Differentiates by type, not quantity or order.	Gender, Nationality, Occupation	Used for categorizing and segmenting data, like in customer segmentation or demographic studies.
Ordinal Scale	Categories with a specific order or sequence, but the intervals are not necessarily equal.	Education Level, Satisfaction Ratings	Used for ranking or ordering data, like in customer satisfaction surveys or educational qualifications.
Interval Scale	Numeric scale with equal intervals between values, but no true zero point.	Temperature (Celsius/Fahrenheit), Calendar Years	Used for measuring differences and averages in data, like in climate studies or historical timelines.
Ratio Scale	Similar to interval scale but with a true zero point, allowing for statements of magnitude.	Weight, Height, Age, Distance	Used for comprehensive quantitative analysis and scientific calculations, like in physics or engineering applications.

A tabulated form to describe the scales of measurement in detail:

3.3 Data Collection and Measurement

Data Collection ke Process (The Process of Data Collection):

Data Collection Techniques: Data science mein data ikhatta karne ke mukhtalif tareeqe hote hain, jaise surveys, experiments, aur field studies. Har technique ka apna unique maqsad aur faida hota hai. 📋🔍

Misal ke taur par, agar aap market research kar rahe hain to aap online surveys ya focus groups ka istemal kar sakte hain. Ye aapko tezi se aur wasee range mein data faraham karta hai. Ya phir, agar aap environmental studies kar rahe hain, to field observations aur experiments zyada munasib ho sakte hain.



Measurement Errors ki Samajh (Understanding Measurement Errors):

Common Errors: Data collection process mein aane wale common errors mein shamil hain sampling error, bias, aur data entry mistakes. Ye errors aapke data ke results ko significantly affect kar sakte hain. ⚠️🚫

Jaise, agar aap ek survey mein sirf aik khas age group ke logon ko include karte hain, to ye sampling bias create kar sakta hai. Ya phir, data entry mein ghalti se galat information enter ho jaye, to ye bhi results ko distort kar sakta hai. 🖥️📉

Errors ko Kam Karna (Minimizing Errors):

Strategies to Reduce Errors: Kuch strategies jin se aap errors ko kam kar sakte hain, jaise careful planning, diverse sampling, aur data verification processes. Is se aapka data zyada reliable aur accurate banega. 📊✅

Misal ke taur par, aap pehle se hi decide kar lein ke aapki sample population kaisi hogi, taake aapke data mein diversity ho. Aur data collection ke baad, aap data verification aur cleaning process se guzar kar kisi bhi possible errors ko identify aur correct kar sakte hain. 🛠️🔧

3.4 Operationalization and Proxy Measures

3.4.1 Operationalization (Amliyat ka Tareeqa-kar) 📋🔧

Operationalization, ya amliyat ka tareeqa-kar, ek research process ka hissa hai jisme complex concepts ko measurable form mein tabdeel kiya jata hai. 📐🧠


- **Tafseel:** Jab hum research karte hain, to kai dafa humein abstract concepts (jaise khushi, ghurbat, ya sehat) ko quantify karna parta hai. Operationalization is process ko kehte hain jisme hum in concepts ko aise variables mein convert karte hain jo hum measure kar sakein. 🌐💬


Misal ke taur par, agar aap “khushi” ko measure karna chahte hain, to aap isay various indicators jaise life satisfaction, positive experiences, ya smile frequency ke through measure kar sakte hain. 😊📊


- **Application:** Operationalization research design mein crucial hai kyun ke yeh humein specific, measurable, aur quantifiable data provide karta hai jo humare conclusions aur analysis ko more credible banata hai. 📝✅

3.4.2 Proxy Measurement (Proxy Pemayesh)


Proxy measurement, ya proxy pemayesh, tab istemal hoti hai jab direct measurement mushkil ya na-mumkin ho. 


- **Tafseel:** Proxy measurement ek ‘stand-in’ ya alternate measurement hoti hai jo asal variable ki jagah use ki jati hai. Yeh tab kiya jata hai jab asal variable ko direct measure karna mushkil ho. 


Jaise, agar aap kisi mulk ki economic health measure karna chahte hain, to direct isay measure karna mushkil hai. Is ki jagah, aap GDP growth rate, unemployment rate, ya consumer spending jaise indicators ka istemal kar sakte hain as proxies. 

- **Application:** Proxy measurements research mein common hain, khaas tor par social sciences aur economics mein, jahan direct measurement ke liye resources ya access limited ho. Ye technique humein phir bhi important insights provide karti hai, albeit with some level of assumption or indirectness. 


3.5 Surrogate Endpoints


Surrogate Endpoints, ya mutaabadil anjaam ke nuqaat, medical research aur clinical trials mein istemal hone wale aise markers hain jo barah-e-raast bemari ke anjaam ko naapne ke bajaye uske effects ya risk factors ko measure karte hain. 

- **Tafseel:** Yeh aksar un halaton mein istemal hota hai jahan asal clinical endpoint (jaise, marz ki rok-thaam ya ilaj ki kamyabi) ko measure karna mushkil ho ya bohot waqt le. Surrogate endpoints se researchers ko jaldi aur aasaani se samajhne mein madad milti hai ke aik treatment ya dawa kitni effective hai. 

Misal ke taur par, agar ek nai dawai ka test kiya ja raha hai jo cholesterol ko kam karta hai, to researchers direct heart attacks ya strokes ki kami ko naapne ke bajaye cholesterol levels ko measure karte hain as a surrogate endpoint. Ye assumption yeh hota hai ke kam cholesterol level se heart attacks ka risk bhi kam ho jata hai. 



- **Application:** Surrogate endpoints zyada tar chronic diseases (jaise diabetes, hypertension) ke research mein istemal hote hain. Ye researchers

ko enable karta hai ke wo tezi se aur kam resources ke sath potential treatments ki efficacy ko samjhein aur evaluate karein. 

- **Ehmiyat aur Tanqeed:** Surrogate endpoints ka istemal time aur resources ki bachat to karta hai, lekin iska istemal kabhi-kabhi misleading bhi ho sakta hai. Agar surrogate endpoint aur asal health outcome ke darmiyan strong relationship na ho, to is se galat conclusions nikal sakte hain. Is liye, in endpoints ka chayan aur interpretation bohut soch-samajh ke aur scientific evidence ke sath karna chahiye. 



3.6 Quantitative and Qualitative Measurement

Quantitative Data ki Tafseel (Detailing Quantitative Data):

Definition and Examples: Quantitative data wo hota hai jo numbers mein measure kiya ja sakta hai. Is mein typically counts, percentages, ya numerical values شامل hain.  



Misal ke taur par, ek company ki monthly sales, ek website par rozana ke visitors, ya kisi school ke students ke exam scores. Ye data humein concrete aur measurable information deta hai, jaise kitna, kitni baar, aur kis darje mein.

Qualitative Data ka Analysis (Analyzing Qualitative Data):

Nature and Interpretation: Qualitative data non-numeric hota hai aur ismein text, images, ya observations شامل hote hain. Is data ko samajhna aur interpret karna often zyada complex hota hai.  

Jaise, customer reviews, interview transcripts, ya observational notes. Ye data humein deeper insights deta hai jaise log kya sochte hain, kyun kisi cheez ko pasand ya napasand karte hain, aur unke experiences kaise hote hain.

Combining Quantitative and Qualitative Data (Dono Types ke Data ko Milana):

Hybrid Approach: Behtareen insights often dono types ke data ko combine kar ke milte hain. Is approach se hum both measurable outcomes aur deeper human experiences ko samajh sakte hain.  



Ek retail store ka misal lein: Store quantitative data se sales trends aur popular items ko track karta hai, jabke customer interviews aur feedback se ye samajhne

ki koshish karta hai ke customers kyun kisi product ko prefer karte hain ya unke shopping experience mein kya behtar kiya ja sakta hai.

3.7 Data and Types of Data

What is Data?

Data is the raw material of data science. It is the information that we collect and analyze to gain insights and make decisions. Data can be quantitative or qualitative, and it can be collected through various methods, like surveys, experiments, or field studies. Data is the foundation of data science, and it is the basis of all data science processes and techniques.

Is liye, data science mein data ki ahmiyat bohot zyada hai. Is chapter mein, hum dekheinge ke data kya hota hai aur data ki mukhtalif types kya hain.  

Primary and Secondary data are two fundamental categories based on the source and nature of the data collection process.

3.7.1 Primary Data vs. Secondary Data

3.7.1.1 Primary Data

Definition: Primary data is data collected directly by the researcher for the specific purpose of their study. It is original and collected at the source.

Methods of Collection: Includes surveys, interviews, experiments, questionnaires, observations, and focus groups.

Examples:

1. A researcher conducting a survey to study consumer behavior.
2. Field experiments in environmental studies.

Uses:

1. Tailored to the specific needs and questions of the research.
2. Provides up-to-date and relevant data for the study.

Pros:

1. Specific to the researcher's requirements.
2. More control over the data quality.

Cons:

1. Can be time-consuming and costly to gather.
2. Risk of bias in data collection methods.

3.7.1.2 Secondary Data

Definition: Secondary data refers to data that was collected by someone else for a different purpose but is used by a researcher for their study.

Sources: Includes government publications, websites, books, journal articles, internal records of organizations, and previously conducted studies.

Examples:

1. Using census data for demographic studies.
2. Analyzing data from scientific journals for a literature review.

Uses:

1. Useful for obtaining a broad understanding of the topic.

CHAPTER 4

DATA COLLECTION

4.1 What is Data Collection?

Data Collection, woh process hai jis mein hum data ko collect karte hain.  




Data, Data Science or Data Analysis k liay RAW material ka kaam karta hy or agar raw material kharab ho ga tu hamari end product b kharab nikaly ge, is liay hamen data ko collect karne se pehlay b planning karni chahyeay.







Angrez kehta hy k:



“Data collection is a systematic process used to gather information from various sources to answer research questions, test hypotheses, or evaluate outcomes.”

4.2 Steps of Data Collection

Here’s an overview of the typical steps involved in this process:

1. **Define Objectives and Research Questions** : Clearly identify what you want to achieve and the questions you need answers to. This step sets the direction for your entire data collection process.
 - Wazeh karein ke aap kya hasil karna chahte hain aur kin sawalaat ke jawab chahiye. Yeh step pooray data collection process ko direction deta hai.
2. **Design the Data Collection Method/Tool**  : Choose the most appropriate method(s) for collecting data. This could be surveys, interviews, observations, experiments, or use of existing data sources. Design tools like questionnaires or interview guides accordingly.
 - Sahi tareeqa chunain data ikattha karne ke liye. Yeh ho sakta hai surveys, interviews, observations, experiments, ya existing data sources ka istemal. Munaasib tools design karein jaise questionnaires ya interview guides.

3. **Determine the Sample** : Decide on the sample size and the sampling method (random, stratified, convenience, etc.). This step is crucial to ensure the data is representative of the population being studied.
 - Sample size aur sampling method ka faisla karen (random, stratified, convenience, waghera). Yeh step yeh ensure karne ke liye crucial hai ke data mutal'aa karne wale population ki numainda ho.
4. **Collect Data** : Implement the data collection method. This could involve conducting surveys, performing experiments, observing behaviors, or compiling data from existing resources.
 - Data collection method ko implement karen. Yeh involve kar sakta hai surveys conduct karna, experiments perform karna, behaviors observe karna, ya existing resources se data compile karna.
5. **Ensure Data Quality** : Check the data as it's collected for accuracy and completeness. This may involve reviewing responses, checking for missing data, and ensuring the data collection method is being followed correctly.
 - Data ko check karein us ke ikatthe hone ke waqt accuracy aur completeness ke liye. Yeh review karna, missing data check karna, aur yeh ensure karna shamil hai ke data collection method sahi follow ho raha hai.
6. **Data Processing** : Once collected, the data needs to be processed. This might include data entry, coding, transcription, and cleaning to prepare it for analysis.
 - Ikattha kiye gaye data ko process karna. Yeh include kar sakta hai data entry, coding, transcription, aur cleaning tayyar karna ke liye analysis ke liye.
7. **Data Analysis** : Analyze the data using appropriate statistical methods or qualitative analysis techniques, depending on the nature of your data and research objectives.
 - Data ko appropriate statistical methods ya qualitative analysis techniques ke zariye analyze karen, data ki nature aur research objectives ke hisab se.
8. **Interpreting Results** : Draw conclusions from the data analysis. Interpret the results in the context of the research objectives and questions.

- Data analysis se nateejon ko nikalain. Results ko research objectives aur sawalaat ke context mein interpret karein.
9. **Report Writing and Presentation** : Prepare a report or presentation to communicate the findings. This should include the methodology, analysis, results, and conclusions.
- Findings ko communicate karne ke liye ek report ya presentation tayyar karen. Is mein methodology, analysis, results, aur conclusions shamil hone chahiye.
10. **Data Storage and Management** : Store the data securely, ensuring it is organized and accessible for future reference or further analysis.
- Data ko mehfooz aur munazzam tareeqe se store karein taake mustaqbil mein reference ya mazid analysis ke liye accessible ho.

Throughout this process, it's essential to consider ethical aspects, such as participant consent, confidentiality, and data privacy. Each step requires careful planning and execution to ensure the data collected is reliable, valid, and suitable for answering the research questions.

Population vs. Sample

Hamen Agay move karne se pehlay ye samajhna zaroori hy k akhir sample hta kia hy? or Population kia hti hy? or in dono main kia farq hy?

4.3 Population vs. Sample

Yahan main ne ek table banaya hai jo “Population” aur “Sample” ke darmiyaan farq ko wazeh karta hai:

Pehlu (Aspect)	Population	Sample
Tareef (Definition)	The entire group that is the subject of the study.	A subset of the population, selected for analysis.
Size	Usually large, encompassing all the individuals or items of interest.	Smaller, manageable in size, representing the population.
Selection	Not selected but defined as the total group of interest.	Selected using various sampling methods (random, stratified, etc.).
Usage	Used for census or when comprehensive data is needed.	Used for surveys or experiments due to time and resource constraints.
Example	All residents of Lahore for a study on city-wide traffic patterns.	A group of 1,000 residents from different areas of Lahore for the same study.

Is table se aap dekh sakte hain ke “Population” ka matlab hota hai poori group jis ke bare mein study ki ja rahi hai, jabke “Sample” is group ka ek chhota hissa hota hai jo analysis ke liye select kiya jata hai. Har aspect mein dono ki qualities ko samjha gaya hai.

4.4 Case Study for Data Collection

Kahani: Gulshan-e-Iqbal Park Mein Data Collection Ka Safar

Ek dhoop bhari subah Lahore mein, Raza aur Amna, dono researchers, ne decide kiya ke Gulshan-e-Iqbal Park mein logon ki recreational habits pe ek study karenge. Unka maqsad tha pata lagana ke log park mein kitna waqt guzarte hain aur kin activities mein involve hote hain. Unhone project ka naam rakha, “Green Spaces in Urban Lahore: A Study on Usage Patterns.”

Jab Raza aur Amna ne park mein jaake surveys shuru kiye, to unko jald hi ek masla ka samna karna pada. Unka sample size bahut chhota tha aur wo sirf

weekends pe data collect kar rahe the. Is se weekdays ke patterns miss ho rahe the.

Ek din, chai peete hue, Amna ne kaha, “Raza, kya tumhe nahi lagta humein weekdays aur alag alag waqt mein bhi data collect karna chahiye?” Raza ne jawab diya, “Bilkul, humein comprehensive data ke liye ye zaroori hai. Chalo, hum apne method mein tabdeeli laate hain.” 😞☕

Dono ne apna approach badla aur diverse timings aur weekdays ko include kiya. Unhone ek app ka istemal shuru kiya jo location aur time tracking mein madad karta tha. Is se unka data collection zyada effective ho gaya.

Jab unhone apni final report likhi, to us mein unhone apne initial mistakes aur unse seekhe gaye sabak ko bhi shamil kiya. Unki study ne local authorities ko valuable insights diye park ki betterment ke liye.

Is tajurbe se Raza aur Amna ne bohat kuch seekha. Unhone samjha ke data collection ek evolving process hai, jisme flexibility aur adaptability key hai. Unka yeh safar unhe aur bhi samajhdar aur experienced researchers banane mein madadgar sabit hua.

Aakhir mein, unki study naye researchers ke liye ek misaal ban gayi. Yeh dikhaya ke kaise mistakes se seekh kar data collection process ko behtar kiya ja sakta hai aur quality data se community ke liye positive changes laaye ja sakte hain. Raza aur Amna ne is se ek important lesson liya: har ghalti ek mauka hoti hai seekhne ka aur behtar researcher banne ka. 📖🌟🌳☀️

Raza aur Amna ne Gulshan-e-Iqbal Park mein apne data collection ke safar mein best practices ka istemal kar ke high-quality data ikattha kiya. Unki kahani mein yeh bhi shamil hai ke kaise unhone apni ghaltiyon se seekh kar behtar data collect kiya. ☀️📖

Jab Amna ne chai peete hue Raza se baat ki, “Raza, kya humein weekdays aur alag alag waqt mein bhi data collect nahi karna chahiye?” Raza ne jawab diya, “Bilkul, humein comprehensive data ke liye ye zaroori hai. Chalo, hum apne method mein tabdeeli laate hain.” ☕😞

Is ke baad, unhone kuch best practices ko apnaya:

1. **Diverse Sampling:** Unhone apna sample diverse banaya, different days aur times ko cover karte hue, takay data zyada accurately reflect kare park ko use karne walon ki variety.

2. **Reliable Tools:** Unhone ek mobile app ka istemal kiya jo location aur time tracking mein madadgar tha, is se data collection zyada accurate aur less time-consuming hua.
3. **Data Verification:** Raza aur Amna ne regular intervals pe data ki accuracy check ki, taake kisi bhi tarah ki ghaltiyan jaldi pakdi ja saken.
4. **Feedback and Improvement:** Unhone apne approach mein flexibility dikhayi aur feedback ke base par continuous improvements kiye.
5. **Ethical Considerations:** Unhone ensure kiya ke participants ki privacy aur consent ka khayal rakha jaaye, especially personal information collect karte waqt.
6. **Clear Objectives:** Unka maqsad clear tha - park ki usage patterns ko samajhna. Is ne unhe focused rakha aur relevant data collect karne mein madad ki.
7. **Organized Data Management:** Collect kiye gaye data ko unhone systematically organize kiya, is se analysis aur reporting mein asani hui.

Jab unhone apni final report likhi, to unka data comprehensive, accurate, aur reliable tha. Unki study ne park ki usage patterns ke bare mein deep insights provide kiye, jis se local authorities ko park ki planning aur improvements mein madad mili.

Is safar ne Raza aur Amna ko sikhaya ke kaise best practices ko follow kar ke high-quality data collect kiya ja sakta hai, aur yeh ke har ghalti se seekhne ka mauka milta hai. Unki mehnat ne unhe naye researchers ke liye ek misaal bana diya, yeh dikhate hue ke high-quality data collection se community mein positive impact kaise laya ja sakta hai. 🌳 ✨ 📊 🔍

4.5 Reliability and Validity 📊

Aaiye baat karte hain reliability aur validity ke concepts ke baare mein, statistics mein! 📊 ✨

4.5.1 Reliability - Bharosemandi 🔄

Jab Lahore ke galiyon mein aap ek dukaan se har roz achha kharidte hain, aur har dafa uska taste wahi hota hai, toh aap kehte hain, “Yeh dukaan toh bharosemand hai!” Isi tarah, statistics mein, reliability ka matlab hota hai consistency. Agar aap ek survey bar bar conduct karte hain, aur har dafa lagbhag same results aate hain, toh keh sakte hain ke aapka data reliable hai.

Kaise Check Karein Reliability? 😞 - **Test-Retest**: Ek hi test ko different waqt pe dobara karna aur dekhna ke results consistent hain ya nahi. - **Internal Consistency**: Survey ke andar different items ke responses ko compare karna aur dekhna ke kya woh aapas mein consistent hain.

4.5.2 Validity - Darusti 🔍

Ab sochiye, aap ek survey kar rahe hain ke log Lahore mein kitni chai peete hain. Agar aapka survey accurately chai peene ki aadat ko measure kar raha hai, toh keh sakte hain ke aapka data valid hai. Yani, aap jo measure karne ki koshish kar rahe hain, aapka tool wohi measure kar raha hai.

Kaise Check Karein Validity? 🧐 - **Content Validity**: Check kijiye ke aapke questions wohi cover kar rahe hain jo aap measure karna chahte hain. - **Criterion-Related Validity**: Aapke results ko kisi established standard ke results ke sath compare kijiye. - **Construct Validity**: Yeh check karta hai ke kya aapka measure woh theoretical concepts accurately capture kar raha hai jinko aap study kar rahe hain.

4.5.3 Reliability aur Validity Ka Importance 🌟

Sochiye aap ek research kar rahe hain ke Lahore mein traffic ki wajah se log kitna stressed hote hain. Agar aapka data reliable nahi hai, yani har baar alag results aate hain, toh aapke conclusions pe bharosa karna mushkil ho jayega. Aur agar data valid nahi hai, yani aap jo measure karne ki koshish kar rahe hain woh actually measure hi nahi ho raha, toh aapke nateeje ka koi maani nahi hoga.

Isliye, jab bhi statistics ka use karte hain, toh yeh ensure karna bohot zaroori hai ke aapka data reliable bhi ho aur valid bhi. Sirf is tarah se aapke research ke nateeje qabile bharosa aur mufeed hote hain. 📊🔍🌟

Reliability and Validity kaisay ensure karen

Reliability aur validity ko ensure karne ke liye, aapko apne data collection process mein kuch best practices ko follow karna chahiye. Yeh practices aapke data ko reliable aur valid banane mein madadgar sabit hongi. Triangulation is one of them 📊🌟

4.6 Triangulation 📊🌟

Aaiye baat karte hain “triangulation” ke concept ke baare mein! 🌐🔍

Triangulation - Teen Zavai Pemaish 📐 Imagine kariye, aap Lahore ke famous Food Street mein hain, aur aapko best biryani ki dukaan dhoondhni hai.

Aap teen alag-alag logon se poochte hain, aur agar teeno ek hi dukaan ki taraf ishara karte hain, toh aapko yakeen ho jata hai ke wohi behtareen jagah hogi. Isi tarah, research mein triangulation ka concept istemal hota hai. Yeh basically yeh hai ke aap ek hi research question ko alag-alag methods, sources, ya perspectives se dekhte hain, aur agar sab ek hi result ki taraf ishara karte hain, toh aapke nateeje zyada bharosemand hote hain.

Triangulation ke Types 🎲 1. **Data Triangulation:** Alag-alag sources se data ikattha karna. Jaise, surveys, interviews, aur observational data. 2. **Methodological Triangulation:** Alag-alag methods use karna, jaise quantitative (numbers pe focus) aur qualitative (words aur meanings pe focus) methods. 3. **Researcher Triangulation:** Alag-alag researchers ya analysts ka data ko analyze karna, taake bias kam se kam ho. 4. **Theoretical Triangulation:** Alag-alag theories ya nazariyaat ko istemal karna data ko samajhne ke liye.

Triangulation ka Faida ✨ Sochiye aap ek research kar rahe hain ke Lahore mein air pollution kis had tak health problems create karta hai. Agar aap sirf ek method (jaise sirf surveys) ya sirf ek source (jaise sirf hospital records) pe depend on karte hain, toh ho sakta hai aap kuch important aspects miss kar jaayein. Lekin agar aap triangulation ka istemal karte hain, yani surveys, environmental data, aur doctors ki opinions bhi lete hain, toh aapke nateeje zyada comprehensive aur accurate honge.

Triangulation ki Ahmiyat 📈 Triangulation research ko zyada mazboot banata hai. Ye aapko yeh confidence deta hai ke aapke nateeje sirf ek angle se nahi aaye hain, balke multiple sources aur perspectives ne usay support kiya hai. Yeh aapke research ko zyada qabile bharosa aur impactful bana sakta hai.

Toh, jab bhi aap kisi research project pe kaam kar rahe hoon, triangulation ko zaroor consider karein. Yeh aapke findings ko depth aur strength deta hai, aur aapke conclusions ko zyada convincing banata hai. 📖 ✨ 📊 🌐

4.7 True and Error Scores 📊 📄

4.7.1 True Score (Asal Score) 🎯

True Score, ya asal score, woh score hai jo kisi shakhs ki asal salahiyat ya performance ko accurately represent karta hai. ✨ 📄

- **Tafseel:** Ye maan'na hai ke jab koi test ya assessment liya jata hai, to jo score milta hai, woh ideally us shakhs ki asal ability ya knowledge ko

darust taur par bayan karta hai. Lekin, reality mein, kayi factors ki wajah se yeh hamesha mumkin nahi hota.

- **Application:** Educational assessments, psychological tests, ya kisi bhi qisam ke performance evaluations mein true score ko samajhna zaroori hota hai. Yeh humein batata hai ke agar tamam external factors ko control kiya jaaye, to shakhs ka actual performance kya hoga. 🏠📚
- **Ahmiyat:** True score ki understanding yeh ensure karti hai ke evaluations fair aur accurate hain, aur yeh bhi ke assessments ko improve kiya ja sake taake woh asal abilities ko behtar taur par measure kar sakein. 🎓⚖️

4.7.2 Error Score (Ghalti ka Score) ❌

Error Score, ya ghalti ka score, woh farq hai jo true score aur observed score ke darmiyan hota hai. 📊?

- **Tafseel:** Jab koi assessment ya test liya jata hai, to jo score hasil hota hai, us mein kuch inaccuracies ya ghaltiyan ho sakti hain. Ye ghaltiyan external factors jaise test-taker ki thakan, misunderstanding of questions, ya testing environment ki wajah se ho sakti hain.
- **Application:** Error score ko samajhna educational testing, psychological assessments, aur kisi bhi qisam ke performance evaluation mein ahem hota hai. Yeh batata hai ke assessment kitna reliable ya valid hai. 😊📊
- **Ahmiyat:** Error score ko kam se kam rakhne ki koshish ki jati hai taake assessment zyada accurate ho. Iske liye, test design aur administration ke methods ko behtar banaya jata hai, aur sometimes, statistical adjustments bhi kiye jaate hain. 📊🔧

The equation that relates True Score and Error Score is fundamental in psychometrics and assessment theory. It's expressed as:

Observed Score= True Score+Error Score

or we can also write this as:

$$X=T+E$$

In this equation:

- **Observed Score X (Mushahada Shuda Score):** This is the score actually obtained from the assessment or test. It's the measurable outcome you see.


- **True Score T (Asal Score):** The score that would theoretically be obtained if there were no errors in measurement. It represents the actual ability or characteristic being measured.
- **Error Score E (Ghalti ka Score):** This represents the portion of the score that deviates from the true score due to various factors like measurement errors, environmental factors, or test-taker's condition.





In an ideal situation, where there's no error, the error score would be zero, and the observed score would equal the true score. However, in real-world scenarios, there's almost always some degree of error, making the observed score slightly different from the true score.

Errors are important

Errors ko samjhna zaroori hy, yehi nahi k Data collection main sirf or sirf measurement main ghaltian ati hyn, balky data collection k har process main ghaltian ati hyn, jin ko samjhna or kam karna zaroori hy. Isi liay ap ko yeh janna zaroori hy k ghaltian kis tarah ki hoti hyn, or un ko kam karny k liay ap kis tarah ki strategies use kar sakti hyn.

4.8 Types of Errors (Ghalthiyon ke Iqsaam)

There are several Error Types that can affect the observed score. These include: Bilkul, aaiye dekhte hain data collection mein hone wale errors ko Roman Urdu mein, aur Pakistani style mein, thodi si emojis ke sath! 

1. **Sampling Error (Namunaati Ghalti)** : Jab sample poori population ka sahi se representation na kare. Masalan, agar Lahore ke sirf ek hisse ke logon se data liya jaye, magar pooray Lahore ke liye generalize kar diya jaye.
2. **Measurement Error (Pemaish Ki Ghalti)** : Jab jo cheez measure karni ho, woh sahi se na ho paaye. Jaise, kisi survey mein “aap kitni dafa chai peete hain?” ka sawaal ho, aur log ghalat jawab dein ya sawaal ka ghalat matlab le lein.
 1. **Random Error (Ittefaqi Ghalti)** : Ye woh ghalthiyan hain jo bas ho jaati hain, bina kisi wajah ke. Yeh har data collection mein hoti hain aur inhe poori tarah se khatam nahi kiya ja sakta.
 2. **Systematic Error (Nizaamati Ghalti/Bias)** : Jab koi consistent ghalti ho, jo bar-bar ho. Jaise, agar survey mein sawaal hi aise ho ke logon ko ek khaas tarah ka jawab dene par majboor kare.

3. **Processing Error (Processing Ki Ghalti)** 🖥️: Data ko process karte waqt ghaltiyan, jaise ghalat data entry ya code. Misal ke taur pe, naam “Ahmed” ko “Ahmad” likh dena.
4. **Coverage Error (Coverage Ki Ghalti)** 🌐: Jab sample mein kuch log ya areas miss ho jaayein. Maan lo, Karachi ke kisi survey mein Clifton ka area hi na cover ho.
5. **Nonresponse Error (Jawab Na Milne Ki Ghalti)** 🙋: Jab log survey respond hi na karein ya kuch sawaalat ka jawab na dein. Socho, aap ne 100 logon se poocha magar sirf 50 ne jawab diya.

Umeed hai ke aap ko Pakistani style mein yeh examples samajh aaye honge!

🌸 PK Remember, data collection mein ghaltiyan aam baat hain, magar inhen samajhna aur kam karna zaroori hai! 🤪 ✨

4.8.1 Sampling Error 📊

Sampling Error, ya namunaati ghalti, woh farq hota hai jo sample aur population ke darmiyan hota hai. 📊 ?

Aaiye baat karte hain “Sampling Error” ke baare mein! 🎲 📊

Sampling Error - Namunaati Ghalti 🔄

Sochiye, aap Lahore ke kisi bazaar mein jate hain aur wahan ke kuch logon se poochte hain ke unhe cricket pasand hai ya nahi. Phir, aap ye generalise kar dete hain ke pooray Lahore ko cricket pasand hai. Yeh ho sakta hai ke jo log aap se mile woh cricket ke fans ho, lekin pure Lahore ka yehi opinion nahi ho. Is tarah ke generalization mein jo error aata hai, usse hum “sampling error” kehte hain. Yani, jab aapka sample (namuna) poori population (aabaadi) ka sahi se representation na kare.

Sampling Error Kyun Hota Hai? 😞





1. **Chhota Sample Size:** Agar sample size bohot chhota ho, to yeh mushkil ho jata hai ke aap poori population ke trends ko accurately capture karein.
2. **Biased Selection:** Agar aapne sample ko biased tareeke se choose kiya ho, jaise sirf ek khaas area ke logon ko include kiya ho.
3. **Random Variability:** Kabhi-kabhi, pure ittefaq se, aapka sample aisi characteristics dikhata hai jo poori population mein nahi hoti.

Sampling Error Se Kaise Bacha Ja Sakta Hai? 🛡️



1. **Bigger and Diverse Sample:** Try karein ke aapka sample size bada aur diverse ho, taake woh better represent kare poori population ko.
2. **Random Sampling:** Use karein random sampling ka method, taake har individual ke selection ka chance barabar ho.
3. **Understanding the Population:** Pehle achi tarah samajh lein ke aapki population kya hai aur uske different aspects kya hain.



Sampling Error Ka Asar

Agar aapka data sampling error se affected hai, toh aapke research ke nateeje kamzor ho sakte hain. Lahore mein traffic patterns ke study mein agar sirf ek area se data liya gaya ho, toh aap pooray shehar ke traffic ke bare mein accurate conclusions nahi nikal sakte.

Isliye, jab bhi research kar rahe ho, toh sampling error ko samajhna aur usse bachne ki koshish karna bohot zaroori hai. Ye ensure karta hai ke aapke ikatthe kiye gaye data accurate aur qabile bharosa hain, aur aapke nateeje sahi mayene mein useful hote hain.    

4.8.2 Measurement Error (Pemayesh ki Ghalti)

Measurement Error, ya pemayesh ki ghalti, woh farq hota hai jo actual result (asal nateeja) aur measured result (pemayesh shuda nateeja) ke darmiyan hota hai.  

- **Tafseel:** Jab bhi hum kisi cheez ko measure karte hain, chahe woh educational test ho, scientific experiment, ya koi survey, to ghaltiyan hone ka imkaan hota hai. Yeh ghaltiyan asal aur measured outcomes ke darmiyan farq paida karti hain.
- **Ahmiyat:** In ghaltiyon ko samajhna aur unhen control karna zaroori hota hai taake hum jo nateeja nikal rahe hain woh zyada accurate aur reliable ho.  

4.8.2.1 Types of Measurement Errors (Pemayesh ki Ghaltiyon ke Iqsaam)





1. Random Errors (Ittefaqi Ghaltiyan):

- **Tafseel:** Ye wo ghaltiyan hoti hain jo unpredictably aur randomly hoti hain. In ka koi specific pattern, ya wajah nahi hoti.

- **Misal:** Jaise, ek survey mein respondent ka randomly koi sawaal galat samajh lena.
- **Application:** Random errors ko kam karne ke liye, data ko carefully analyze karna aur large sample sizes use karna hota hai. 🎲 🔍

2. Systematic Errors (Nizami Ghaltiyan/Bias):

- **Tafseel:** Ye ghaltiyan tab hoti hain jab measurement process mein koi constant bias ya error ho.
- **Misal:** Jaise, ek scientific instrument ka hamesha thoda sa zyada ya kam reading dena.
- **Application:** In errors ko identify kar ke unhen correct karna padta hai, jaise calibration of instruments ya testing procedures ko modify karna.  

Bias

Bias ko samajhna bohot zaroori hai taake aap apne data collection ko zyada accurate aur reliable bana sakein. Bias ke types aur us se bachne ke strategies ke baare mein aap yahan parh sakte hain: [Bias and Measurement Bias](#)

3. Human Errors (Insaani Ghaltiyan):

- **Tafseel:** Ye ghaltiyan insaan ki taraf se hoti hain, jaise ghalat data entry ya ghalat interpretation.
- **Misal:** Data ko galat tarike se enter karna ya kisi pattern ko ghalat samajhna.
- **Application:** Training aur careful review se in ghaltiyan ko kam kiya ja sakta hai. 🧠 📝 👁️

Measurement Error aur Reliability!

Is tafseeli wazahat mein, pemayesh ki ghaltiyan aur unke mukhtalif iqsaam ko Roman Urdu mein samjhaya gaya hai, unke asraat aur unhen kam karne ke tareeqon ke sath. Yeh understanding kisi bhi qisam ke research, assessment, ya data collection process ke liye crucial hai taake nateejaat zyada qabile bharosa aur durust ho sakein.

4.8.3 Processing Error (Processing Ki Ghalti) 🖥️

Len Janab-e-Aali, chaliye baat karte hain “Processing Error” ke baare mein!



Processing Error - Data Ko Process Karne Mein Ghalti

Imagine kariye, aap ek restaurant mein order dene ke baad dekhte hain ke aapka order galat aaya hai. Shayad waiter ne galat likha ho ya kitchen mein kuch mix up hua ho. Isi tarah, jab data collect karne ke baad usay process karte waqt ghaltiyan hoti hain, toh isey “processing error” kehte hain. Yeh tab hota hai jab data ko enter karte waqt, sort karte waqt, ya analyze karte waqt kisi tarah ka error ho jaye.

Processing Error Kyun Hota Hai?





1. **Data Entry Errors:** Jab data ko system mein enter kiya jata hai, toh typing mistakes ho sakti hain.
2. **Coding Mistakes:** Agar data ko code karte waqt galat codes use kiye jayein, jaise survey responses ko galat categories mein rakhna.
3. **Software Errors:** Kabhi-kabhi software mein glitches ya bugs ki wajah se bhi processing errors ho sakte hain.

Processing Error Se Kaise Bacha Ja Sakta Hai?

1. **Double-Check Data:** Data enter karne ke baad use dobara check karein.
2. **Automated Tools:** Jahan mumkin ho, automated tools ka use karein, jo errors ko kam karte hain.
3. **Training and Protocols:** Jo log data process kar rahe hain unko achi training dijiye aur clear protocols follow karne ko kahiye.

Processing Error Ka Asar

Agar aapke data mein processing error hai, toh aapke research ke nateeje galat ho sakte hain. Maan lijiye, aap Lahore mein air quality ka study kar rahe hain aur aapke data mein entry error hai. Is se aapka analysis galat ho sakta hai aur aapke nateeje bhi.

Isliye, data ko process karte waqt bohot dhyan rakhna zaroori hota hai. Achi quality control practices aur careful data handling se aap processing errors ko kam kar sakte hain, aur apne research ke nateeje ko zyada accurate aur qabile bharosa banate hain.    

4.8.4 Coverage Error (Coverage Ki Ghalti)

Bilkul, chaliye baat karte hain “Coverage Error” ke baare mein!  

Coverage Error - Data Collection Mein Kami

Imagine kariye, aap ek survey kar rahe hain ke Lahore ke log kis tarah ke mobile phones use karte hain. Agar aap sirf university students ko hi survey karte hain, toh aap baaki bohot se logon ko miss kar rahe hain jaise professionals, housewives, ya elderly log. Is tarah ke limitation ko “coverage error” kehte hain. Yeh tab hota hai jab aapka data collection ka method ya source poori population ko cover nahi karta.

Coverage Error Kyun Hota Hai?

1. **Limited Sources:** Agar aapka data collection ke sources limited hain, toh aap poori population ko represent nahi kar sakte.
2. **Outdated Lists:** Agar aap jo list use kar rahe hain woh outdated hai, toh new members ya recent changes miss ho sakte hain.
3. **Geographical Limitations:** Agar aap sirf ek specific geographical area ko cover karte hain, toh baaki areas miss ho jate hain.

Coverage Error Se Kaise Bacha Ja Sakta Hai?

1. **Comprehensive Sources:** Apne sources ko zyada comprehensive banayein, taake aap poori population ko represent kar sakein.
2. **Update Lists Regularly:** Apne data collection lists ko regularly update karein.
3. **Include Diverse Areas:** Different geographical areas ko include karein apne study mein.

Coverage Error Ka Asar

Agar aapke study mein coverage error hai, toh aapke nateeje poori population ke liye sahi nahi honge. Jaise agar aap Lahore ke mobile phone usage ka study kar rahe hain aur sirf university students ko include karte hain, toh aapke conclusions sirf unke liye valid honge, baaki population ke liye nahi.

Isliye, jab bhi aap kisi study ya research project pe kaam kar rahe hoon, toh coverage error ko samajhna aur usse bachne ki koshish karna bohot zaroori hai. Is se aapke data collection ko zyada accurate aur mufeed banaya ja sakta hai, aur aapke nateeje zyada qabile bharosa aur comprehensive hote hain.



4.8.5 Nonresponse Error (Jawab Na Milne Ki Ghalti)

Bilkul, chaliye baat karte hain “Nonresponse Error” ke baare mein! 📧🚫

Nonresponse Error - Jawab Na Milne Ki Ghalti 📊

Sochiye aap Karachi mein ek survey conduct kar rahe hain ke log kis tarah ke transport ko prefer karte hain. Aap 500 logon ko forms dete hain, lekin sirf 200 log hi jawab dete hain. Yeh jo missing data hai, usse “nonresponse error” kehte hain. Yeh tab hota hai jab logon se data collect karte waqt kuch log jawab hi nahi dete, ya phir kuch sawaalat chhod dete hain.

Nonresponse Error Kyun Hota Hai? 😞

1. **Long or Complex Surveys:** Agar survey bohot lamba ya mushkil ho, toh log jawab dene se gurez karte hain.
2. **Privacy Concerns:** Kuch log personal ya sensitive sawalaat pe jawab dene mein hesitant hote hain.
3. **Accessibility Issues:** Agar survey sirf English mein ho aur respondent Urdu ya local language mein comfortable hoon.

Nonresponse Error Se Kaise Bacha Ja Sakta Hai? 🛡️

1. **Short and Simple Surveys:** Surveys ko mukhtasar aur asaan banayein, taake zyada log jawab dein.
2. **Assure Privacy:** Logon ko yakeen dilayein ke unka data safe aur confidential rahega.
3. **Language Options:** Multiple languages mein surveys provide karein, especially local languages.

Nonresponse Error Ka Asar 📉

Agar aapke data mein nonresponse error hai, toh aapke study ke nateeje distorted ho sakte hain. Karachi ke transport preferences ke study mein, agar sirf aadhe log hi jawab dete hain, toh aapka data sirf unhi ke preferences ko represent karega, baaki population ka nahi.

Isliye, data collect karte waqt nonresponse error ko minimize karne ki koshish zaroori hoti hai. Is se aapke data collection ko zyada reliable aur accurate banaya ja sakta hai, aur aapke nateeje zyada comprehensive aur qabile bharosa hote hain. 📚📧🔍🌟

Bias

Bias ko samajhna bohot zaroori hai taake aap apne data collection ko zyada accurate aur reliable bana sakein. Bias ke types aur us se bachne ke strategies ke baare mein aap yahan parh sakte hain: [Bias and Measurement Bias](#)

4.9 Bias and Measurement Bias

Bilkul, chaliye detail se samajhte hain “bias” aur “measurement bias” ke concepts ko, Roman Urdu aur emojis ke saath! 📖🔍

4.9.1 General Bias

Bias research ya data collection mein systematic error ya haqeeqat se hat kar results ya inferencess ki taraf ishara karta hai. Research mein, bias data collection, analysis, interpretation, aur publication ke kisi bhi aspect mein aa sakta hai. Yeh biases mukhtalif sources se aate hain:

1. **Selection Bias (Intikhab Mein Jhukao):** Jab study mein شامل kiye gaye participants general population ka sahi representation na karein. Misal ke taur pe, agar health survey sirf urban hospitals mein kiya jaye, to rural areas ke trends miss ho sakte hain.
2. **Confirmation Bias (Tasdeeqi Jhukao):** Jab researchers ya data collectors jaan bujh kar ya anjaane mein data ko favor karte hain ya results ko interpret karte hain jo unke pehle se mojud beliefs ya hypotheses ko confirm karte hain.
3. **Publication Bias (Ishaa'at Mein Jhukao):** Sirf positive ya significant results ko publish karna, jabke negative ya non-significant findings ko nahi karna.
4. **Reporting Bias (Reporting Mein Jhukao):** Jab study ke sirf kuch outcomes ya aspects ko report kiya jata hai, aur dusre jo barabar important ho sakte hain, unhe chhoda jata hai.

4.9.2 Measurement Bias

Measurement Bias ek khaas tarah ka bias hai jo measurements ko collect, record, ya interpret karne mein systematic error se juda hota hai. Yeh study ke outcomes ki reliability aur validity par significantly asar daal sakta hai. Measurement bias ke types mein شامل hain:

1. **Instrument Bias (Aala Mein Jhukao):** Jab khud measurement instruments mein kharabi hoti hai. Jaise, ek scale jo hamesha kam wazan dikha raha ho.

2. **Observer Bias (Mushahida Karne Wale Ka Jhukao):** Jab measurement karne wala shakhs (observer) anjaane mein results ko influence karta hai, shayad apni expectations ya pehle se banaye gaye khayalat ki wajah se.
3. **Response Bias (Jawab Mein Jhukao):** Jab study mein शामिल log woh jawab dete hain jo unhe lagta hai ki expected hain ya socially acceptable hain, unke asal thoughts ya feelings ke bajaye.
4. **Sampling Bias (Namuna Mein Jhukao):** Measurement bias ka ek type jo namuna (sample) kaise chuna gaya hai us se juda hota hai. Agar sample population ka sahi representation na ho, to measurement biased ho sakta hai.

4.9.3 Bias Aur Measurement Bias Ko Kam Karna

Research ki integrity aur usefulness ke liye bias ko kam karna bohot zaroori hai. Yahan kuch strategies hain:

1. **Diverse Sampling (Mukhtalif Namunaat):** Yakeen karna ke sample jitna mumkin ho diverse aur representative ho.
2. **Blinding (Andha Banaye Rakhna):** Single ya double-blind study designs ka istemal karna jahan zaroori ho, taake observer aur participant biases ko roka ja sake.
3. **Calibration aur Maintenance (Tarteef aur Dekh-Bhaal):** Regularly instruments ko check aur calibrate karna unki accuracy ko yakeen mein lane ke liye.
4. **Training aur Standardization (Tarbiyat aur Ma'yari Amal):** Data collectors ko achi training dena aur measurement procedures ko standardize karna observer bias ko kam karne ke liye.
5. **Pilot Studies (Pehle Azmooda Mutala'a):** Pilot studies conduct karna taake main study se pehle biases ko pakda aur durust kiya ja sake.
6. **Triangulation (Teen Zavaii Pemaish):** Multiple methods, sources, ya theories ka istemal karna results ko cross-verify karne ke liye.
7. **Transparency in Reporting (Reporting Mein Shaffafiyat):** Sabhi findings aur methodologies ko wazeh aur mukammal tor par report karna taake reporting bias se bacha ja sake.
8. **Peer Review aur Replication (Sathi Jaiza aur Dohrao):** Research ko peer review ke liye pesh karna aur aise studies design karna jo replicate ki ja saken taake findings ki tasdeeq ho sake.

CHAPTER 5

DISSCRIPTIVE STATISTICS

5.1 What is Statistics?

Angrez kehta hai ka:

“Statistics is the science of collecting, organizing, presenting, analyzing and interpreting numerical data to assist in making more effective decisions.”

Asan alfaz main:

“Statistics is the science of data.”

Us se b asan alfaz main:

“Statistics data ka wo ilm hai jo data ko collect karna, usay organize karna, analyse karna, aur phir us se conclusions nikalne mein madad karta hai.”

Sochiye, aap Lahore ke traffic patterns ya Karachi ke weather trends ko samajhne ke liye data collect kar rahe hain; statistics yahan par aapko yeh samajhne mein madad karega ke data kya keh raha hai.

5.1.1 Statistics Ke Types




Broadly, statistics do main types mein divided hai:

1. **Descriptive Statistics (Tafseeli Shumariyat):** Ye data ko summarize karta hai, jaise mean (average), median, aur mode. Yeh aapko batata hai ke aapka data overall kaisa dikhta hai.
2. **Inferential Statistics (Istakhraji Shumariyat):** Ye larger populations ke bare mein conclusions draw karne ke liye sample data ka use karta hai. Iska istemal hypotheses testing, predictions, aur estimates banane ke liye kiya jata hai.



5.2 Descriptive Statistics

Descriptive Statistics data ko samajhne aur present karne ka aik asaan aur effective tareeqa hai. Is mein hum data ke basic features ko describe karte hain aur is se simple summaries about the sample and the measures banate hain. Yeh typically numbers ya graphs ke zariye kiya jata hai. Descriptive statistics mein shamil hain:

1. **Measures of Central Tendency (Markazi Rujhan Ke Pemaane):** Ye batata hai ke data ka central point kya hai. Is mein shamil hain:
 - **Mean (Ausat):** Tamam values ka average.
 - **Median (Wasti Qiymat):** Data set ki beech wali value jab values ko order mein rakha jaye.
 - **Mode (Aksar Aane Wali Qiymat):** Sab se zyada bar bar aane wali value.
2. **Measures of Variability (Tabdeeli Ke Pemaane):** Ye batata hai ke aapke data points kitne diverse hain. Is mein shamil hain:
 - **Range (Hudood):** Sab se kam aur zyada value ke beech ka farq.
 - **Interquartile Range (IQR):** Data set ki beech wali 50% values ka range.
 - **Variance (Ikhtilaaf):** Average se har value ke farq ka square.
 - **Standard Deviation (Mayaar Ki Hera-Phairi):** Ye batata hai ke data points mean se kitna door hain.
 - **Standard Error (Mayaar Ki Hera-Phairi Ki Khaata):** Ye batata hai ke sample mean population mean se kitna door hai.
3. **Graphs and Charts (Graphs aur Chart):** Data ko visually summarize karte hain, jaise bar charts, histograms, aur pie charts.

Descriptive statistics se aapko fori aur clear understanding milti hai ke aapka data kya keh raha hai, bina kisi complex analysis ke. Ye Lahore ke temperature patterns se le kar Karachi ke shopping trends tak, har tarah ke data ko samajhne mein pehla qadam hota hai.   

5.3 Population and Samples

Bilkul population and samples asan hy aik dam. chaliye iss concept ko detail se samajhte hain!  

Population vs. Sample: Ek Hi Data Set, Do Mukhtalif Tawajjuh

Ek hi data set ko kabhi population aur kabhi sample ke taur par dekha ja sakta hai, ye depend on karta hai ke aap data ko kyun aur kaise analyze kar rahe hain.

Misal: Farz karein, aap ke pass ek class ke final exam ke grades hain. Agar aapka maqsad sirf is class ke scores ka distribution describe karna hai, toh ye

grades aapke liye ek **population** ban jate hain. Lekin, agar aap in grades se koi inference nikalna chahte hain, maslan, dusre classes ya schools ke students ke scores ke bare mein, toh ye grades aapke liye ek **sample** ke tor par kaam aate hain.

5.3.1 Population Analysis 📊🌟

Jab aap **population** ko analyze karte hain, iska matlab hai ke aapki data set mein poori population of interest shamil hai. Aap is group ke tamam members par apne calculations kar rahe hote hain, aur aap seedhe taur par iss group ke characteristics ke bare mein statements kar sakte hain.

5.3.2 Sample Analysis 📊🔍

Is ke baraks, jab aap **sample** ko analyze karte hain, toh aap ek bade population se liye gaye chhote hisse par kaam kar rahe hote hain. Yahan jo statements aap bade group ke bare mein karte hain, woh probabilistic hote hain, yani ke yeh poore group par bilkul sahi na bhi ho sakte.

Descriptive vs. Inferential Statistics:

- **Parameters (Population ke Liye):** Population ko describe karne wale numbers ko *parameters* kehte hain, aur inhe Greek letters jaise μ (population mean ke liye) aur σ (population standard deviation ke liye) se signify kiya jata hai.
- **Statistics (Sample ke Liye):** Sample ko describe karne wale numbers ko *statistics* kehte hain, aur inhe Latin letters jaise \bar{x} (sample mean ke liye) aur s (sample standard deviation ke liye) se signify kiya jata hai.

Practical wajah se, aksar **population** ki bajaye **samples** ka analysis kiya jata hai, kyunki puri population ko direct study karna mumkin nahi ho sakta ya phir bohot mehnga pad sakta hai.

Is distinction ko samajhna statistics mein fundamental hai aur iske liye notational conventions aur terminology develop ki gayi hain. Har author ke yahan thoda farq ho sakta hai, lekin generally, yehi conventions follow kiye jate hain. 📖🌟📊


5.3.3 Notations for Descriptive Statistics

Notations

Statistics main mathematical notations ka bohot istemal hota hai. Is liye, aapko in notations ko samajhna bohot zaroori hai. Yeh rahi descriptive statistics mein

istemat hone wali kuch common notations ki list, unke ek line mein explanations ke saath:

- \bar{x} - **Sample Mean**: Sample data points ka average.
- μ - **Population Mean**: Puri population ke data points ka average.
- s - **Sample Standard Deviation**: Sample data ke variation ya spread ko measure karta hai.
- σ - **Population Standard Deviation**: Puri population ke data ke variation ya spread ko measure karta hai.
- s^2 - **Sample Variance**: Sample data points ke beech ke differences ka square.
- σ^2 - **Population Variance**: Population data points ke beech ke differences ka square.
- n - **Sample Size**: Sample mein total data points ki taadad.
- N - **Population Size**: Population mein total data points ki taadad.
- **min** - **Minimum Value**: Data set mein sab se kam value.
- **max** - **Maximum Value**: Data set mein sab se zyada value.
- **Q1, Q2, Q3** - **Quartiles**: Data set ko char hisson mein taqseem karne wale points, Q2 median bhi hai.
- **IQR** - **Interquartile Range**: Q3 aur Q1 ke beech ka farq, jo data spread ki central tendency ko batata hai.

Ye notations descriptive statistics mein data ko samajhne aur uski interpretation mein madadgar hoti hain. Har notation ka apna specific maqsad aur use hota hai. 

- **Yehi nahi abhi or suneay!**

Descriptive statistics mein kuch aur bhi notations hote hain jo data ko samajhne aur analyze karne mein madadgar hote hain. Yahan kuch aur examples hain:

- **Median** - **Median**: Data set ki beech wali value jab values ko order mein rakha jaye.
- **Mode** - **Mode**: Data set mein sab se zyada bar bar aane wali value.
- Σ - **Summation**: Kisi series ya sequence ke tamam elements ka total.

- **r - Correlation Coefficient:** Do variables ke beech ke relationship ki strength aur direction ko measure karta hai.
- **Skewness - Skewness:** Data distribution ki asymmetry ko measure karta hai, yani ke data kis taraf zyada jhuka hua hai.
- **Kurtosis - Kurtosis:** Data distribution ke “peakedness” ya “flatness” ko measure karta hai.
- **Range - Range:** Data set mein sab se zyada aur sab se kam values ke beech ka farq.
- **p - Proportion:** Kisi khaas category ya class mein falling observations ka proportion.
- **Frequency(f) - Frequency:** Kisi value ya class ki frequency, yani ke kitni dafa woh value ya class data set mein aati hai.

Ye additional notations aur bhi insight provide karte hain jab hum kisi data set ko descriptive statistics ke zariye analyze karte hain. Inka istemal data ke various aspects ko better samajhne aur interpret karne mein kiya jata hai.



5.4 Measure of Central Tendency

Angrez kehta hy:

Measures of central tendency, also known as **measures of location**, are descriptive statistics that describe the central position of a numeric data set.

- Measure of central tendency are typically among the first computation or calculation you do for a data set, specifically for the quantitative (continuous) data.

Bilkul, chaliye baat karte hain “Central Tendency” ke baare mein, Roman Urdu aur emojis ke saath! 🌟 📊

Central Tendency - Markazi Rujhan 🎯

“Central Tendency” ek statistical term hai jo describe karti hai ke ek data set ke andar values kis jagah concentrate hote hain ya jama hote hain. Yeh basically aapko batati hai ke aapke data ke “markaz” ya “centre” mein kya hota hai. Central Tendency ke measures data ke typical ya average behavior ko samajhne mein madad karte hain.

Central Tendency Ke Teen Main Measures 📊 📈 🎯

5.4.1 The Mean

Mean (Ausat): Yeh sab values ka average hota hai. Misal ke taur par, agar aap Lahore ke kisi college ke students ki umar ka average nikalna chahte hain, toh aap sab students ki age ka total le ke unki taadad se divide karenge.

Mean-Misal k tor per

Average Income 💰: Farz karein, aap Karachi ke ek mohalle mein rehne walon ki average monthly income janna chahte hain. Aap mohalle ke har ghar se income collect karte hain aur un sab ko jamaa kar ke total gharon ki taadad se divide karte hain. Jo number aata hai, woh is mohalle ki average monthly income hogi. Ye mean ya ausat kehlata hai.

5.4.1.1 Mathematical Equation

Mean (Ausat) 📊

- **Formula:** $\bar{x} = \frac{\sum x_i}{n}$
- **Jahan:**
 \bar{x} mean ya ausat hai, $\sum x_i$ tamam values ka sum hai, n values ki total taadad hai.
- **Example:** Agar aapke paas 5 students ki marks hain [60, 70, 80, 90, 100], to mean hoga: $\frac{60+70+80+90+100}{5} = 80$

Mean in python can be calculated as show in [Figure 5.2](#)

5.4.1.2 Mean (Ausat) Ki Types

1. Arithmetic Mean (Riyazi Ausat) 📊

- **Definition:** Tamam values ka total sum divided by values ki total taadad. Ye sab se common type hai.
- **Formula:** $\bar{x} = \frac{\sum x_i}{n}$
- **Jahan:** \bar{x} mean hai, $\sum x_i$ tamam values ka sum hai, n values ki total taadad hai.
- **Example:** Agar aap Lahore ke ek school ke har class ke students ki height ka average nikalna chahte hain, to aap sab heights ko jama karein aur phir unki taadad se taqseem karein. Maan lijiye heights hain [150cm, 160cm, 140cm], to mean hoga: $\frac{150+160+140}{3} = 150\text{cm}$




2. Geometric Mean (Jyamiti Ausat)

- **Definition:** Tamam values ke product ka nth root, jahan n values ki taadad hai. Ye rates of change ya percentages jaise data ke liye use hota hai.
- **Formula:** $G = (\prod_{i=1}^n x_i)^{1/n}$
- **Jahan:** G geometric mean hai, $\prod_{i=1}^n x_i$ tamam values ka product hai, n values ki taadad hai.
- **Example:** Agar aap Karachi mein property prices ke annual growth rate ka average nikalna chahte hain aur rates hain [10%, 15%, 20%], to geometric mean hoga: $1.10 \times 1.15 \times 1.203 \approx 1.148$ ya 14.8%.

3. Harmonic Mean (Moseeqi Ausat)

- **Definition:** Ye rates ya ratios ke reciprocal ke average ke reciprocal hota hai. Ye speed ya rates jaise data ke liye use hota hai.
- **Formula:** $H = n \sum 1/x_i$
- **Jahan:** H harmonic mean hai, n values ki taadad hai, $1/x_i$ har value ka reciprocal hai.
- **Example:** Agar aap Islamabad to Rawalpindi ke safar ke different speeds ko measure kar rahe hain aur speeds hain [40km/h, 60km/h], to harmonic mean hoga: $2/1/40 + 1/60 = 48$ km/h

5.4.1.3 Mean Ki Importance

Mean, ya arithmetic mean, data science aur everyday analysis mein sab se zyada istemal hone wala measure hai. Ye aapko data set ke general trend ko samajhne mein madad karta hai. Business decisions se le kar scientific research tak, mean ka istemal data ke average behaviour ko samajhne aur us par mabni decisions lene ke liye kiya jata hai. Ye simple hai, lekin powerful tool hai jo data ke central tendency ko capture karta hai.   

5.4.2 The Median

Median (Wasti Qeemat): Jab values ko order mein arrange kiya jaye, toh jo value beech mein aati hai, woh median hoti hai. Jaise, agar aap Islamabad mein ghar ki prices ko order mein lagaen, toh jo price beech mein hogi, woh median hogi.

Median-Misal k tor per

Property Rates 🏠: Lahore ke DHA mein property rates ki range bohot zyada hoti hai. Agar aap sab properties ki values ko order mein arrange karein aur phir beech ki value dekhein, toh woh median property rate hogi. Ye aapko batayega ke aam taur par DHA mein property ki qiymat kya hogi, is se high aur low extremes ka asar kam ho jata hai.

5.4.2.1 Mathematical Equation

Median (Wasti Qiymat) 📊 - **Odd Number of Values:** Median woh middle value hai jab values ko ascending order mein arrange kiya jata hai. - **Even Number of Values:** Middle do values ka average. - **Example:** Agar aapke paas values hain [10, 20, 30, 40, 50], to median 30 hoga (middle value). Agar values hain [10, 20, 30, 40], to median hoga: $20+30 \div 2 = 25$

Median in python can be calculated as show in [Figure 5.2](#)

5.4.2.2 Examples

1. Odd Number of Values

- **Situation:** Aap Karachi ke ek school ke class mein students ki heights record kar rahe hain aur unki taadad odd hai, jaise [150cm, 155cm, 160cm, 165cm, 170cm].
- **Median:** Heights ko order mein arrange karne ke baad, beech ki value, yani 160cm, median hogi.




2. Even Number of Values

- **Situation:** Islamabad ke ek hospital mein aane wale patients ki daily count record ki ja rahi hai aur ek haftay mein counts hain [30, 35, 40, 45].
- **Median:** Yahan do middle values hain, 35 aur 40, to median hoga ($= 37.5$).

5.4.2.3 Median Ka Importance 📌

Median, khaas taur par us waqt ahem ho jata hai jab data skewed ho ya outliers contain karta ho. Jaise, agar aap Lahore mein property prices dekhein aur kuch bahut hi high ya low prices hon, to mean distorted ho sakta hai, lekin median aapko zyada reliable picture deta hai.


Median data ke distribution ko beech se cut karta hai, jis se humein pata chalta hai ke data set ke aadhe values is value se kam hain aur aadhe zyada. Ye

especially helpful hota hai market analysis, income surveys, ya real estate prices jaise situations mein, jahan outliers mean ko affect kar sakte hain.   


5.4.3 The Mode

Mode (Aksar Aane Wali Qiymat): Wo value jo data set mein sab se zyada bar repeat hoti hai. Maan lijiye, Karachi ke ek area mein agar aap dekhein ke zyadatar log konsi car drive kar rahe hain, toh jo car model sab se zyada nazar aaye, woh mode hoga.

Mode-Misal k tor per

Most Common Car : Islamabad ke F-10 sector mein, agar aap notice karein ke zyadatar log Honda Civic drive kar rahe hain, to Honda Civic is sector ka mode hoga. Yani yeh woh car hai jo is area mein sab se zyada common hai.

5.4.3.1 Mathematical Equation

Mode (Aksar Aane Wali Qiymat)  - **Formula:** Mode woh value hai jo data set mein sab se zyada bar aati hai. - **Example:** Agar aapke paas values hain [1, 2, 2, 3, 4], to mode 2 hoga (kyunki 2 do dafa aaya hai aur baaki sab ek ek bar).

Mode in python can be calculated as follows:

Code

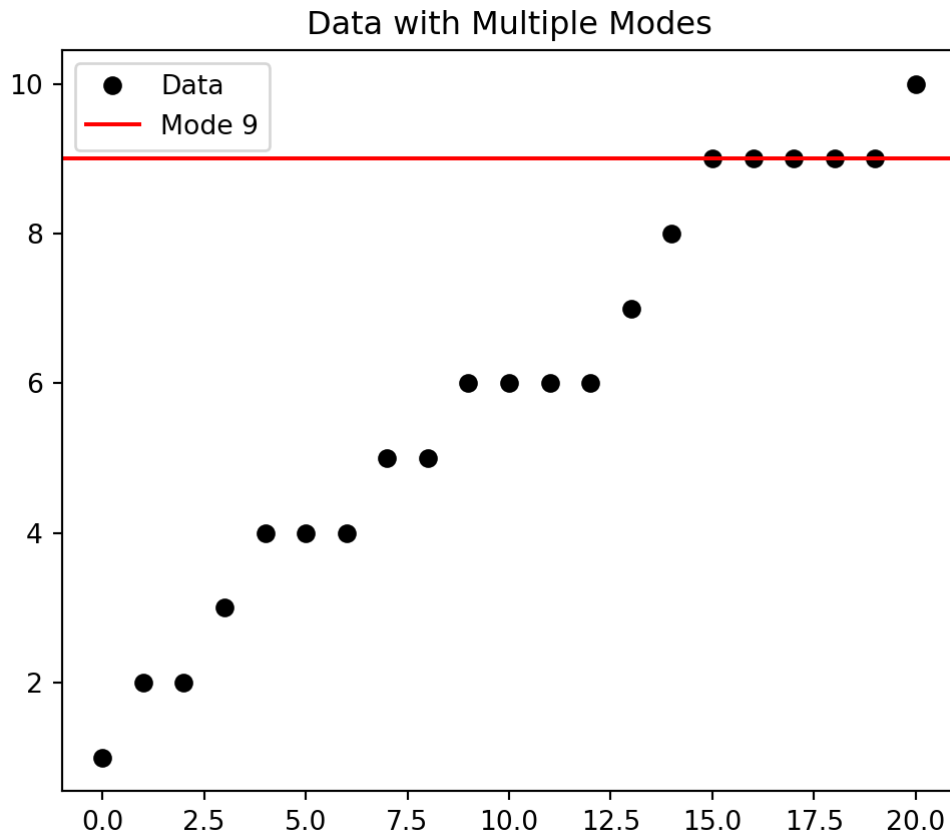


Figure 5.1: Mode of a data set

5.4.3.2 Mode Ke Types

1. Unimodal (Yak Mode)

- **Definition:** Jab ek hi data set mein sirf ek hi mode ho.
- **Example:** Agar aap Islamabad ke ek school mein students ki pasandida ice cream flavors ki list banate hain aur sab se zyada “Chocolate” flavor aata hai, toh yeh unimodal hai - yani Chocolate yahan ka mode hai.

2. Bimodal (Do Mode)

- **Definition:** Jab ek data set mein do alag modes hote hain.
- **Example:** Karachi ke ek shopping mall mein agar aap customers se unki age poochein aur sab se zyada do age groups, maan lijiye 25 aur 40, bar bar aate hain, to yeh bimodal hai - yani yahan ke do modes hain: 25 aur 40 years.

3. Multimodal (Kayi Modes)

- **Definition:** Jab ek data set mein do se zyada modes hote hain.
- **Example:** Lahore ke ek music festival mein alag-alag music genres ki popularity check karte waqt, agar aapko pata chale ke “Pop”, “Rock”, aur “Classical” teeno genres equally popular hain, toh yeh multimodal hai - yani teen modes hain: Pop, Rock, aur Classical.

4. No Mode (Koi Mode Nahin) 🚫

- **Definition:** Jab koi bhi value data set mein doosri values se zyada bar nahi aati.
- **Example:** Agar aap Peshawar ke ek mohalle mein sabhi gharo ki construction dates dekhein aur har ghar ki construction date alag ho, toh is situation mein koi mode nahi hoga.

5.4.3.3 Mode ki Importance

Mode ka use data ke distribution aur uske most common ya repeated elements ko samajhne ke liye hota hai. Ye aksar categorical data, jaise favorite items, categories, ya classes ko analyze karne mein madadgar hota hai. Mode se humein yeh insight milta hai ke kis cheez ko log sab se zyada pasand karte hain ya sab se zyada istemal karte hain, jo ke market research, public opinion surveys, aur social sciences mein khaas taur par useful hota hai. 📖 🔍 ✨ 🎯

5.4.4 Central Tendency Ki Importance 🔍

Central Tendency humein data sets ke baare mein aham insights deti hai. Yeh batata hai ke aam taur par data kis tarah distribute hua hai. Business decisions se le kar scientific research tak, har jagah Central Tendency ka analysis crucial hota hai. Is se hum data ke general pattern ko samajh sakte hain aur complex data sets ko asaan tareeqe se summarize kar sakte hain.

Central Tendency ka sahi istemal aapko aapke data ke markaz ya central point ki deep understanding provide karta hai, jo ke kisi bhi statistical analysis ya data science project ke liye foundation ka kaam karta hai. 📊 ✨ 📖

5.4.5 Median is Better than Mean

Median is Better than Mean: Jab data set mein outliers ya extreme values hain, toh median mean se zyada reliable hota hai. Is liye, jab aap data set ko analyze kar rahe hain aur aapko pata hai ke data mein extreme values hain, toh aap median ka istemal karen.

Let's see an example to understand this better. 📊 ✨ 📖

Code

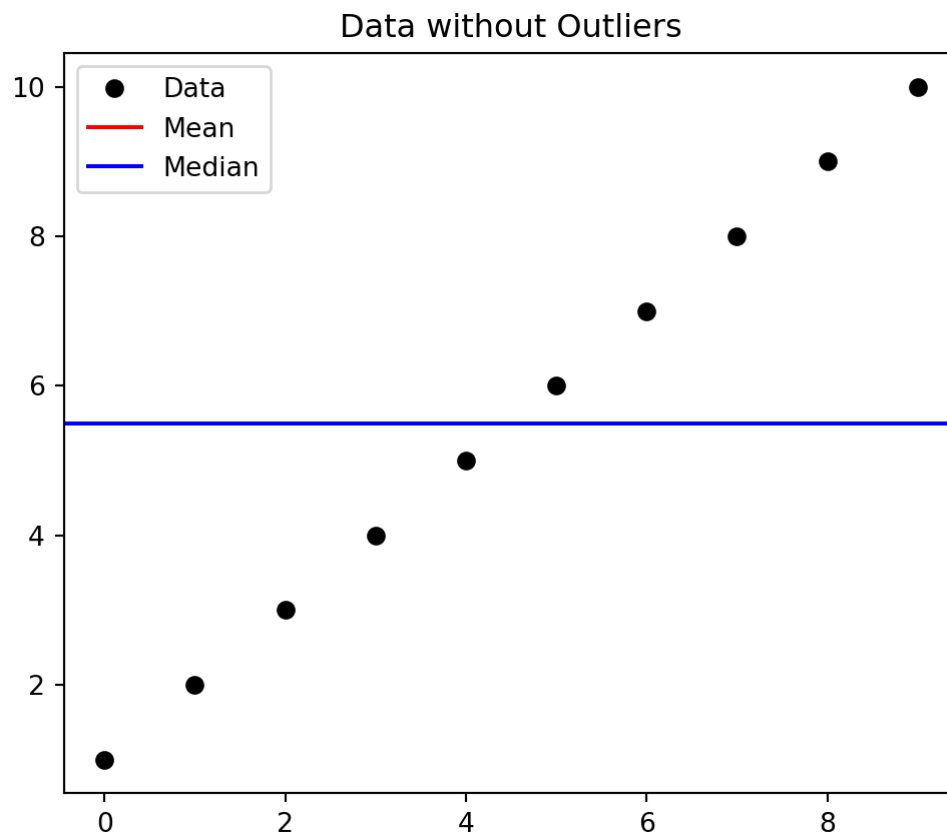


Figure 5.2: Figure without any Outlier (Mean and Median are same)

Code

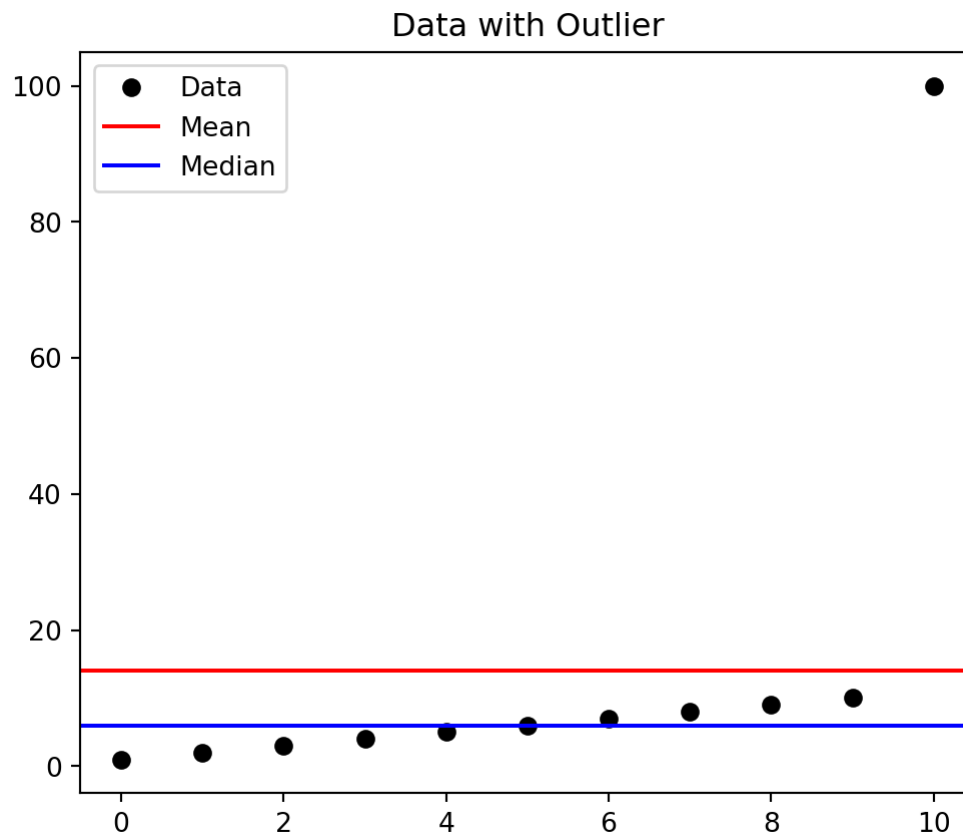


Figure 5.3: Figure with one Outlier (mean is affected more than median)

Code

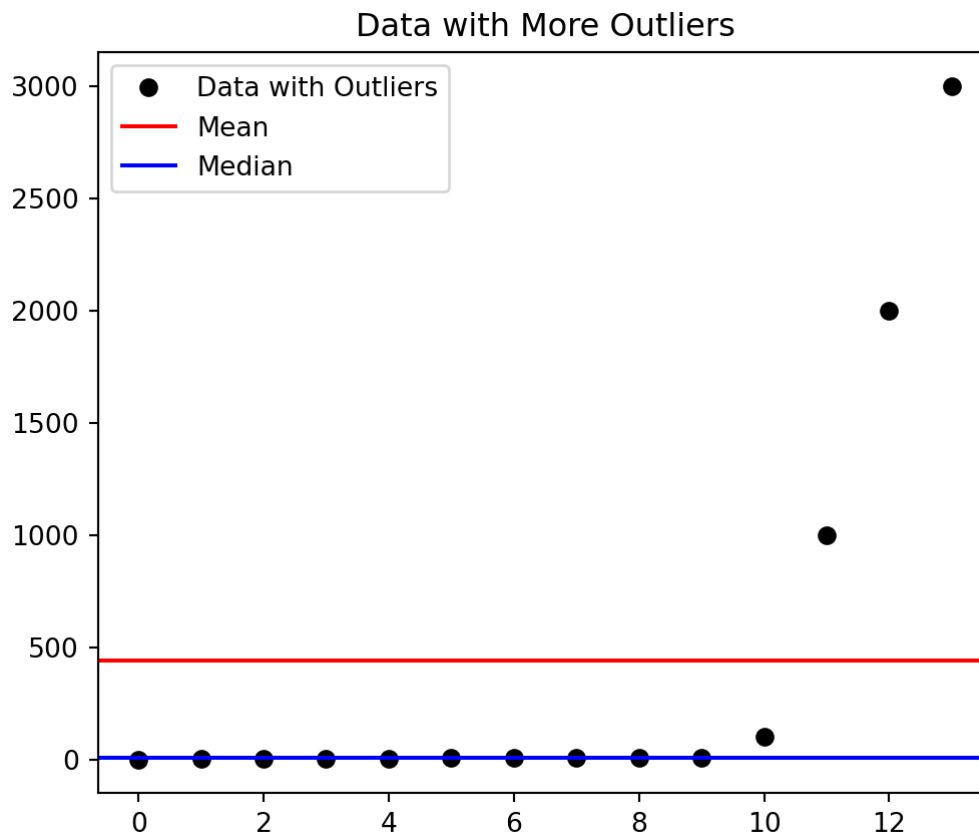


Figure 5.4: Figure with more outliers added this time the mean is even affected more than the median.

5.5 Measure of Variability

Variability ya **Tabdeeli**, statistics mein data points ke darmiyan hone wale ikhtilaaf ya farq ko bayan karta hai. Yeh aapko batata hai ke aapke data mein diversity ya inconsistency kitni hai. Iska matlab hai ke aapke data points kitne similar ya dissimilar hain ek dusre se.

Variability is also known as **dispersion** ya **spread**.

Variability Ke Main Measures 🖋️

5.5.1 Range 📐

- **Definition:** Ye simplest form hai variability ka, jo ke highest aur lowest values ke darmiyan ke farq ko show karta hai.
- **Formula:** $\text{Range} = \text{Maximum Value} - \text{Minimum Value}$
- **Example:** Agar Quetta mein alag-alag dukaanon par ek jaisi cheez ki alag-alag qiymaten hain, to range sab se kam aur sab se zyada price ke beech ka farq hoga. ### **Interquartile Range (IQR)** 📊

- **Definition:** Ye data set ki beech wali 50% values ka range hai. Ye data set ke extremes ko ignore karta hai.
- **Formula:** $IQR = Q3 - Q1$
- Where Q3 third quartile hai aur Q1 first quartile hai.
- **Example:** Agar aap Lahore ke ek school ke students ke test scores ko analyze kar rahe hain, to IQR आपको batayega ke middle 50% students ke scores kitne hain.

5.5.2 Variance

- **Definition:** Ye batata hai ke average se har data point kitna door hai, square kiya hua.
- **Formula:**
- Population ke liye: $\sigma^2 = \sum (x_i - \bar{x})^2 / n$
- Sample ke liye: $s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$
- Where σ^2 population variance hai, s^2 sample variance hai, $\sum (x_i - \bar{x})^2$ har value ke farq ka square hai, n values ki total taadad hai, aur \bar{x} mean hai.
- **Example:** Agar Faisalabad ke kisi college ke students ke test scores hain [65, 70, 75], to variance un scores aur unke average ke beech ke farq ka square hoga.

5.5.3 Standard Deviation

- **Definition:** Ye variance ka square root hai aur ye batata hai ke data mean se average kitna door hai.
- **Formula:**
- Population ke liye: $\sigma = \sqrt{\sigma^2}$
- Sample ke liye: $s = \sqrt{s^2}$
- Where σ population standard deviation hai, s sample standard deviation hai, σ^2 population variance hai, s^2 sample variance hai.
- **Example:** Agar Karachi mein alag-alag schools ke matric ke result scores hain, to standard deviation se humein pata chalega ke average score se har school kitna vary karta hai.

5.5.4 Standard Error

- **Definition:** Ye batata hai ke sample mean population mean se kitna door hai.
- **Formula:** $\text{Standard Error} = \frac{\sigma}{\sqrt{n}}$
- Where σ population standard deviation hai, n sample size hai.
- **Example:** Agar aap Lahore ke ek school mein students ke test scores ko analyze kar rahe hain, to standard error aapko batayega ke aapke sample mean population mean se kitna door hai.

Bilkul, yeh raha ek aam zindagi se related example jis se “variability” ka concept samajh mein aayega, Roman Urdu aur emojis ke saath! 📊🌍

5.5.5 Coefficient of Variation (CV)

Coefficient of Variation (CV), ya **Tabdeeli Ka Coefficient**, ek statistical measure hai jo data ke variability ko quantify karta hai, lekin isey standard deviation ke relative terms mein express kiya jata hai. Ye batata hai ke data ke standard deviation ka mean ke sath kya rishta hai.

5.5.5.1 CV Ka Formula 📐

CV ka formula hai: $CV = \frac{\sigma}{\bar{x}} \times 100\%$ Jahan σ standard deviation hai aur \bar{x} mean hai.

5.5.5.2 CV Ka Istemal Aur Ahmiyat ✨

1. Comparing Variability Between Different Datasets (Mukhtalif Data Sets Ke Variability Ka Mawazna):

- CV ko especially tab istemal kiya jata hai jab hum alag-alag datasets ya groups ke variability ko compare karna chahte hain, jin ke means alag ho sakte hain.
- **Example:** Agar aap Lahore aur Karachi ke schools ke students ke test scores ka mawazna karna chahte hain aur in dono cities ke average scores mein farq hai, to CV aapko batayega ke kis city mein variability zyada hai relative to their average.

2. Scaling Variability (Tabdeeli Ko Scale Karna):

- Kyunki CV mean ke relative terms mein hota hai, ye kisi bhi size ya scale ke data ke liye applicable hota hai, yeh scale-independent measure hai.

- **Example:** Agar aap different industries ke financial returns ko compare kar rahe hain, jahan revenues ka scale bohot alag ho sakta hai, CV aapko har industry ke returns ki relative variability ko samajhne mein madad karega.

3. Risk Assessment in Finance (Finance Mein Khatraat Ka Andaza):

- Investment aur financial analysis mein, CV ko often risk assessment ke liye use kiya jata hai. High CV ka matlab hota hai zyada risk.
- **Example:** Islamabad stock market mein alag-alag stocks ki investment risk ko measure karne ke liye, analysts CV ka use karte hain.

CV ek versatile tool hai jo data ke spread ya variability ko relative terms mein samajhne mein madad karta hai, aur ye kai fields mein, jaise finance, research, aur marketing mein, insights provide karne ke liye istemal hota hai.



5.5.6 Examples

School Ki Performance

Situation: Aap ek education board ke analyst hain aur aapko Lahore ke alag-alag schools ke matriculation ke exam results ka analysis karna hai. Aap dekh rahe hain ke har school ke students ke marks mein kitna farq hai.

1. **Data Collection:** Aap paanch different schools se students ke matric ke exam scores collect karte hain. Yeh scores kuch is tarah hain:

- School A: [70%, 75%, 80%, 85%, 90%]
- School B: [50%, 55%, 60%, 65%, 70%]
- School C: [65%, 65%, 65%, 65%, 65%]
- School D: [70%, 72%, 74%, 76%, 78%]
- School E: [60%, 80%, 60%, 80%, 60%]

2. **Analyzing Variability:**

- **Range (Hudood):** Aap pehle har school ke scores ka range dekhte hain. School A ka range hai 20% (90% - 70%), School B ka bhi 20%, School C ka 0% (sab scores same hain), School D ka 8%, aur School E ka 20%.

- **Standard Deviation (Mayaar Ki Hera-Phairi):** Phir aap standard deviation calculate karte hain taake zyada precise understanding mile. Aapko pata chalta hai ke School C ka standard deviation sab se kam hai, jo indicate karta hai ke uske students ke marks mein kam variability hai.

3. **Conclusion:** Is analysis se aapko yeh insight milti hai ke kuch schools mein students ke marks mein zyada variability hai (jaise School A, B, aur E), jabke School C mein students ke performance mein kam variability hai. Is se education board ko yeh samajhne mein madad milti hai ke kis school mein teaching methods zyada consistent results la rahe hain aur kahan par student performance mein zyada variation hai.

Is tarah ke analysis se stakeholders ko valuable insights milte hain jo unhe policies aur interventions design karne mein madad karte hain. Variability ka yeh analysis business, health, sports, aur bhi bohot se fields mein useful hota hai. 📈 🏠 🌟 📖

5.5.7 Examples in Python

Code

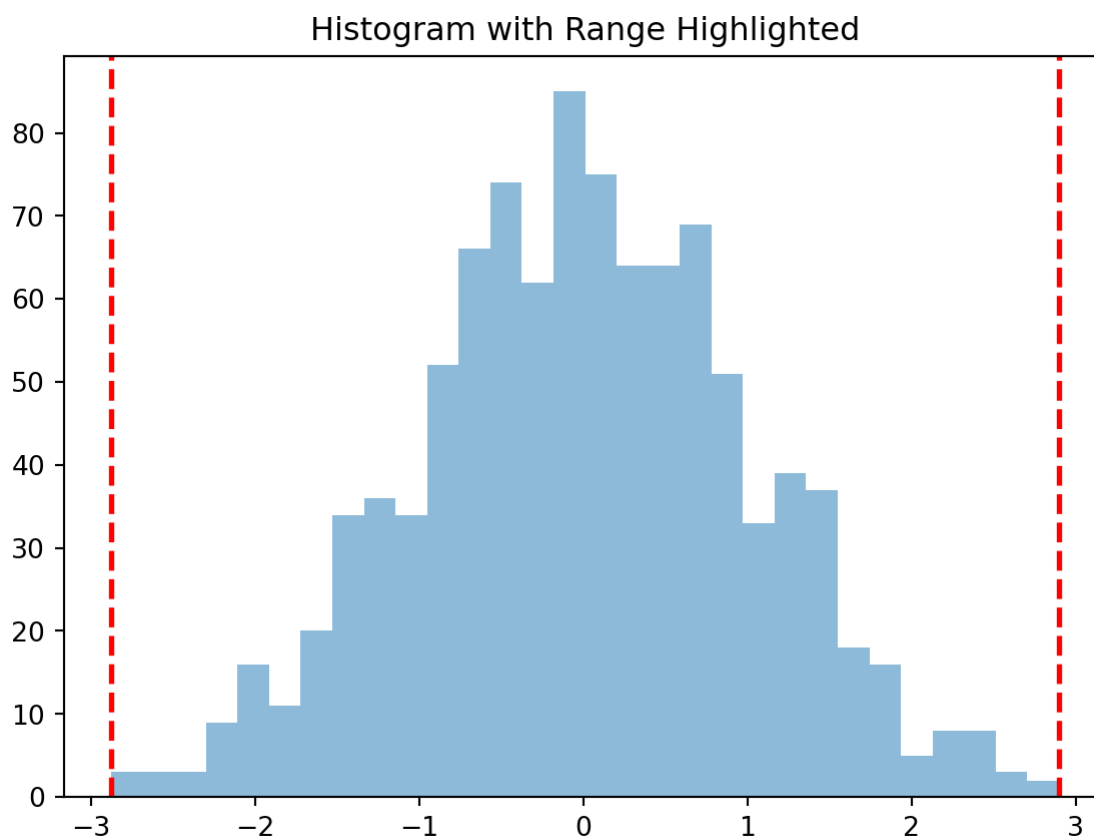


Figure 5.5: Figure showing the range of a data set

Code

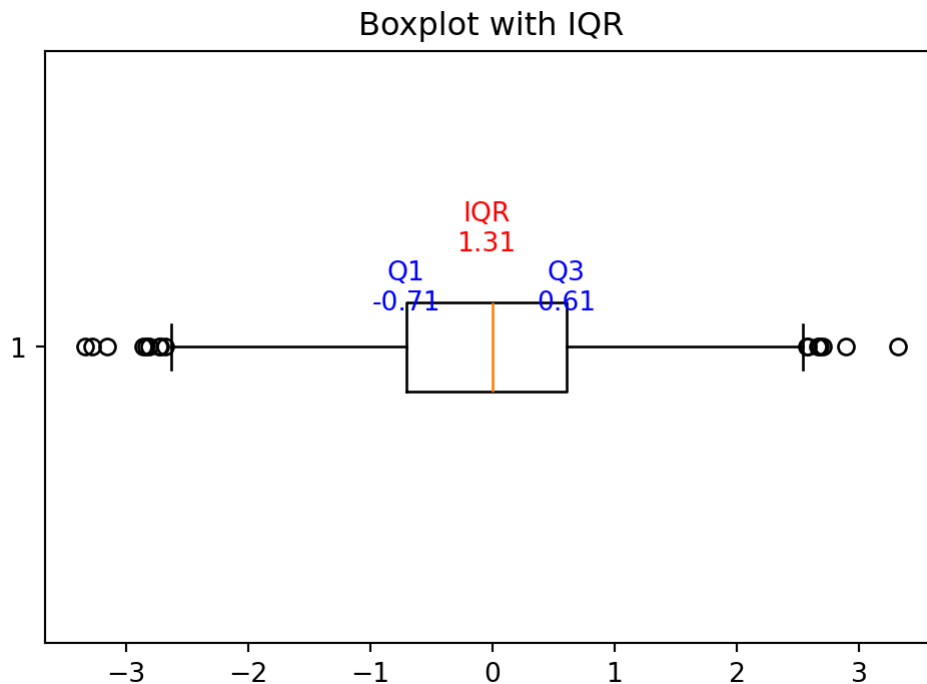


Figure 5.6: Figure showing the interquartile range of a data set (The box represents the IQR, the whiskers represent the range, and the red line represents the median).

The Interquartile Range (IQR) is: 1.311518777203355

Code

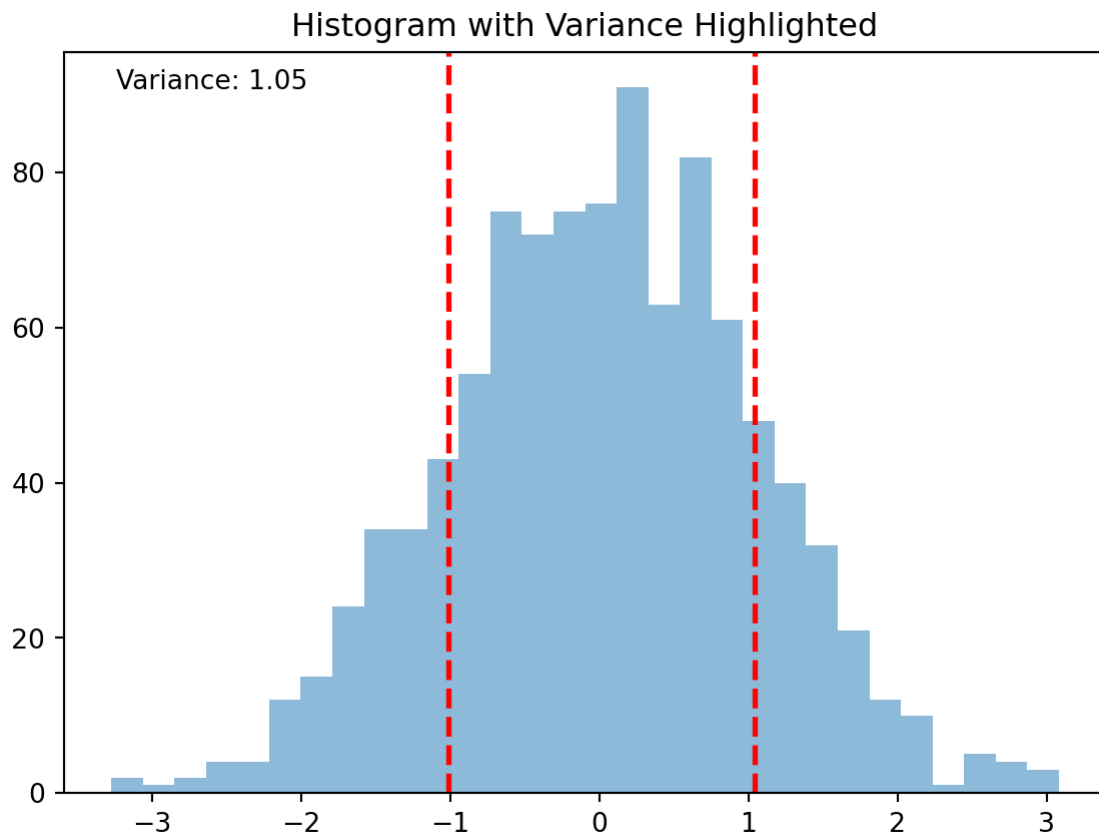


Figure 5.7: Figure showing the variance of a data set. (In the histogram, the dashed red lines represent one standard deviation away from the mean, which gives a visual representation of the variance.)

The variance is: 1.049007173333975

Code

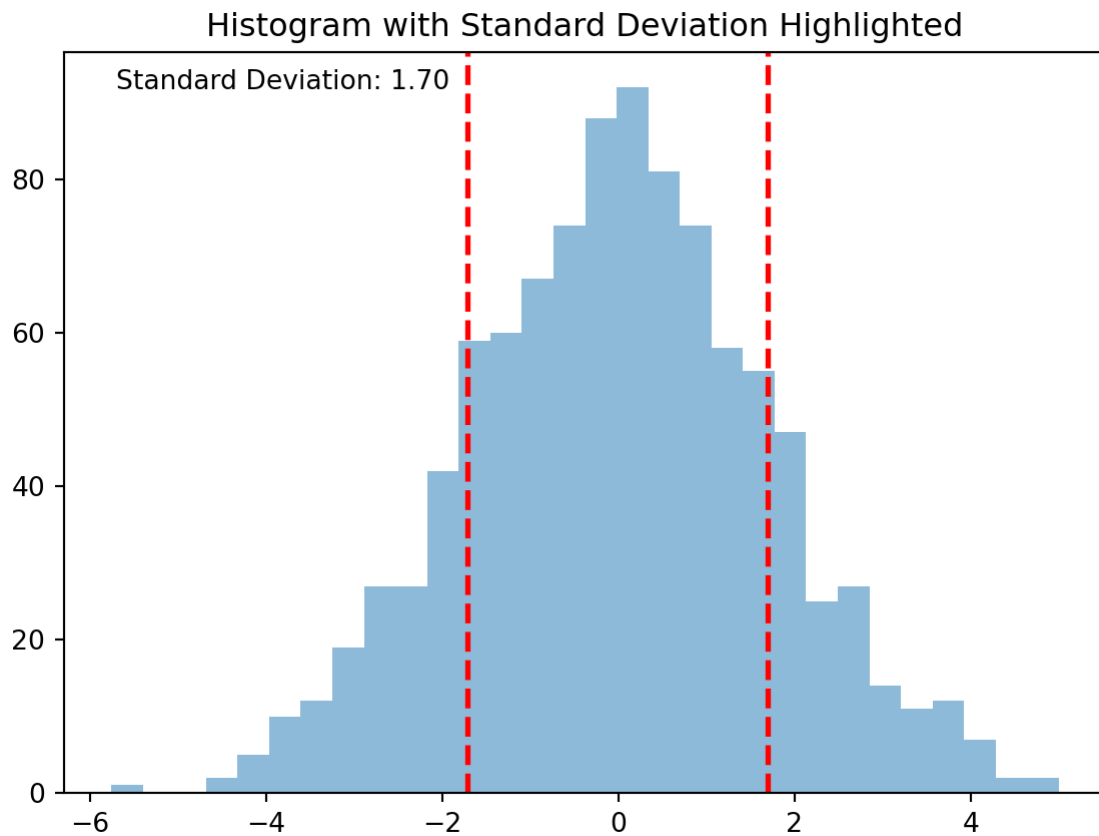


Figure 5.8: Figure showing the standard deviation of a data set. In the histogram, the dashed red lines represent one standard deviation away from the mean, which gives a visual representation of the standard deviation.

The standard deviation is: 1.7033920898240134

Code

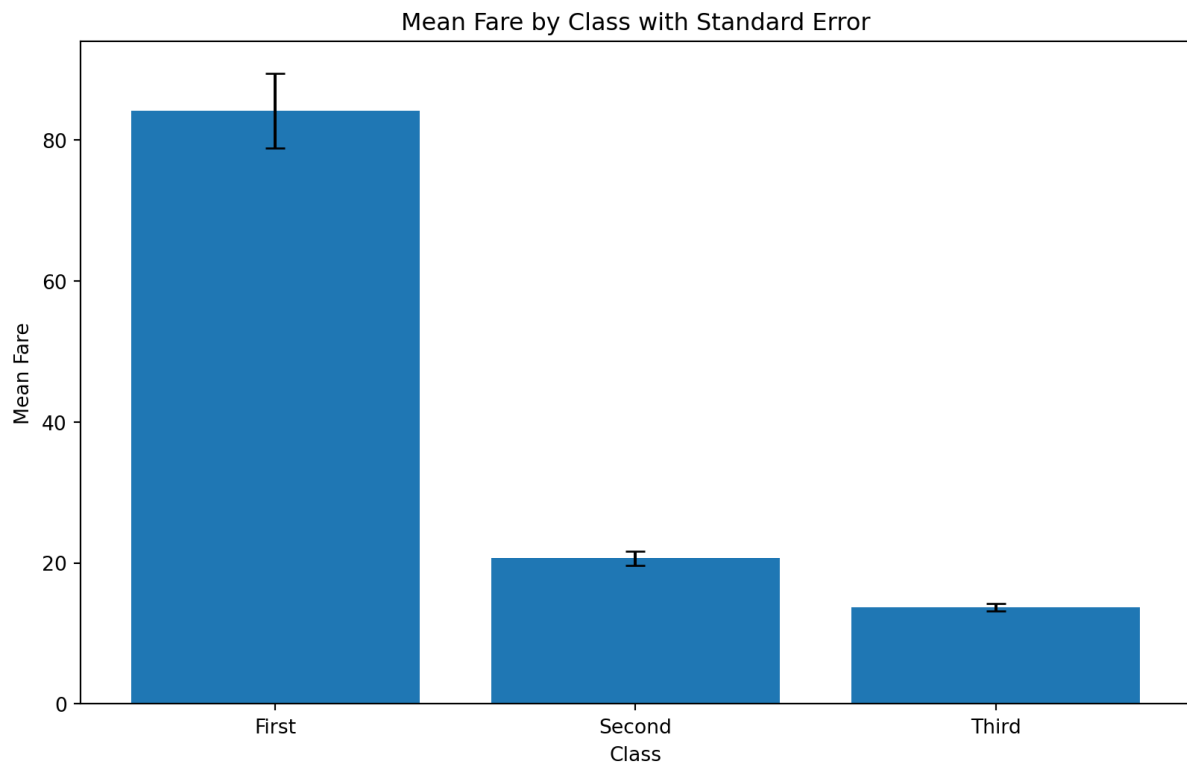


Figure 5.9: Figure showing the standard error in a bar plot of a titanic dataset.

5.5.8 Importance of Variability in Data Science and Machine Learning

Data Science mein, **variability** ya **tabdeeli** ka kirdar bohot ahem hota hai. Ye measure karta hai ke data points kitne alag hain ek dusre se aur ye insights provide karta hai ke data kis tarah distribute hua hai.

5.5.8.1 Variability Ka Importance 🌟

1. Data Understanding (Data Ki Samajh):

- Variability se data scientists ko data sets ke structure aur spread ki deep understanding milti hai. Yeh unhe batata hai ke data mein kis tarah ke patterns ya anomalies hain.
- **Example:** Karachi ke different areas mein air quality index (AQI) ki variability analyze kar ke, scientists pollution sources aur patterns ko better samajh sakte hain.

2. Model Accuracy (Model Ki Durusti):

- Machine learning models mein, high variability ka matlab ho sakta hai ke model ko train karne ke liye zyada complex ya diverse data ki zarurat hogi.

- **Example:** Lahore mein traffic flow predict karne wale model ke liye, road par different times mein hone wale traffic ki variability ko samajhna zaroori hai.

3. Risk Assessment (Khatraat Ka Andaza):


- Businesses aur financial analysts variability ko use karte hain risks ko assess karne ke liye. High variability ka matlab hai zyada risk.
- **Example:** Islamabad ke stock market mein investment ke decisions lene ke liye, different stocks ki price variability ka analysis karna crucial hota hai.

4. Quality Control (Mayaar Par Control):

- Manufacturing ya production processes mein, variability ka kam hona quality control ki achhi indication hoti hai.
- **Example:** Faisalabad ke textile mills mein cloth ki quality check karne ke liye, production line ke output mein variability ko monitor karna important hota hai.

5. Customer Insights (Grahak Ki Maloomat):

- Marketing aur customer behavior analysis mein, variability ko samajhna helps karta hai different customer segments aur unki preferences ko samajhne mein.
- **Example:** Multan mein ek retail store ke customer purchase patterns ki variability ko analyze kar ke, store apne products aur marketing strategies ko optimize kar sakta hai.

In sab examples se clear hota hai ke Data Science mein variability ko samajhna essential hai. Ye aapko data ke nature ko samajhne, risks ko manage karne, aur better decisions lene mein madad karta hai. 

5.6 Outliers

Aaiye baat karte hain “Outliers” ke baare mein Data Science aur statistics mein!



Outliers, ya **intehai qiymat**, woh data points hote hain jo baqi data se bohot alag hote hain. Ye aise values hoti hain jo ya to bohot zyada bari ya chhoti hoti hain baqi data ke comparison mein. In points ko outliers kehte hain kyunki ye “normal” ya expected range se bahar hote hain.

5.6.0.1 Outliers Ki Importance 🌟

1. Data Cleaning (Data Saaf Karna):

- Data science projects mein, outliers ko pehchanna aur unka proper handling zaroori hota hai. Kabhi-kabhi inhe remove karna better hota hai taake model ya analysis accurate ho.
- **Example:** Karachi ke traffic data mein, agar kisi khaas din (jaise kisi badi event ke din) traffic unusually high ho, toh ye outlier consider kiya jaa sakta hai.

2. Error Detection (Ghalti Ka Pata Lagana):

- Outliers kabhi-kabhi data collection ya processing ki ghaltiyon ki nishani bhi ho sakti hain. Inhe identify karna helps karta hai errors ko correct karne mein.
- **Example:** Lahore ke hospital mein patient ki age galat entry ki gayi ho jaise 200 years, ye ek obvious outlier hoga.

3. Insights and Discoveries (Maloomat aur Daryaft):

- Outliers se kabhi-kabhi new discoveries ya important insights mil sakte hain.
- **Example:** Islamabad ke market research data mein, agar kisi product ki sales unexpectedly zyada ya kam ho, toh ye outlier kisi trend ya market change ki nishani ho sakti hai.

4. Statistical Analysis (Shumariyati Tahlil):

- Outliers ka impact statistical measures jaise mean par hota hai, jo overall analysis ko affect kar sakte hain.
- **Example:** Peshawar ke school mein test scores ke analysis mein, agar ek ya do students ne unusually high ya low score kiya ho, toh ye mean score ko distort kar sakta hai.

5.6.1 Outliers Ka Handling 🛠️

Outliers ko handle karna carefully kiya jana chahiye. Kabhi-kabhi inhe data set se hata diya jata hai, lekin kabhi-kabhi inhe analyze karna bhi zaroori hota hai, khas taur par jab ye kisi real phenomenon ya important information ko represent karte hain. Outliers ko identify karne ke liye various methods jaise scatter plots,

box plots, aur statistical tests (e.g., Z-score, IQR) ka use kiya jata hai.



Detecting and removing outliers is a crucial step in data preprocessing, especially in data science and machine learning projects. Python, with its libraries like Pandas, NumPy, and SciPy, provides efficient tools to handle this task. Here's a general approach to detect and remove outliers in Python:

5.6.2 Detecting Outliers

1. Using Statistical Methods:

- **Standard Deviation and Z-Score:**

- Calculate the Z-score for each data point. Z-score indicates how many standard deviations an element is from the mean.
- Typically, data points with a Z-score greater than 3 or less than -3 are considered outliers.

The Z-score method is a statistical technique used to measure how far away a data point is from the mean, relative to the standard deviation of the dataset. The equation for calculating the Z-score of a data point is:

$$Z = \frac{(X - \mu)}{\sigma}$$

Where:

Z is the Z-score.

X is the value of the data point.

μ is the mean of the dataset.

σ is the standard deviation of the dataset.

Explanation of the Z-score Formula:

1. **(X- μ):** This part of the formula calculates the difference between the data point and the mean of the dataset. It shows how far the data point is from the mean.
2. **Division by σ :** This step normalizes the difference based on the standard deviation of the dataset. It essentially tells you how many standard deviations away from the mean your data point is.

Use of Z-score:

- A Z-score can be positive or negative, indicating whether the data point is above or below the mean, respectively.

- In most cases, a Z-score beyond +3 or below -3 is considered as an outlier, as it lies far from the mean (more than 3 standard deviations).

This method is widely used in statistics and data science to identify outliers and understand the distribution of data points within a dataset.

- Example: Here is the code to find outliers in python.

```
import numpy as np

data = np.array([10, 12, 12, 13, 12, 11, 14, 13, 15, 102, 12, 14, 14, 17, 19, 107,
10, 13, 12, 14, 12])

z_scores = np.abs((data - np.mean(data)) / np.std(data))

outliers = data[z_scores > 3]

print(outliers)

[107]
```

- **Interquartile Range (IQR):**
 - Calculate the IQR, which is the difference between the 75th and 25th percentile of the data.
 - Any data points that fall below 25th percentile - 1.5 * IQR or above 75th percentile + 1.5 * IQR are typically considered outliers.
- Example: Here is the code to find outliers in python.

```
import numpy as np

data = np.array([2, 10, 12, 12, 13, 12, 11, 14, 13, 15, 102, 12, 14, 14, 17, 19,
107, 10, 13, 12, 14, 12, 207]) # Your data array

Q1 = np.quantile(data, 0.25)

Q3 = np.quantile(data, 0.75)

IQR = Q3 - Q1

outliers = data[(data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR))]

print(outliers)

[ 2 102 19 107 207]
```

1. Using Visualization Tools:

- **Box Plots:**

- Box plots are a great way to visualize outliers.
- Data points that fall outside of the whiskers (1.5 times the IQR) are outliers.
- Example: Here is an example with code to find outliers in python using box plots.

Code

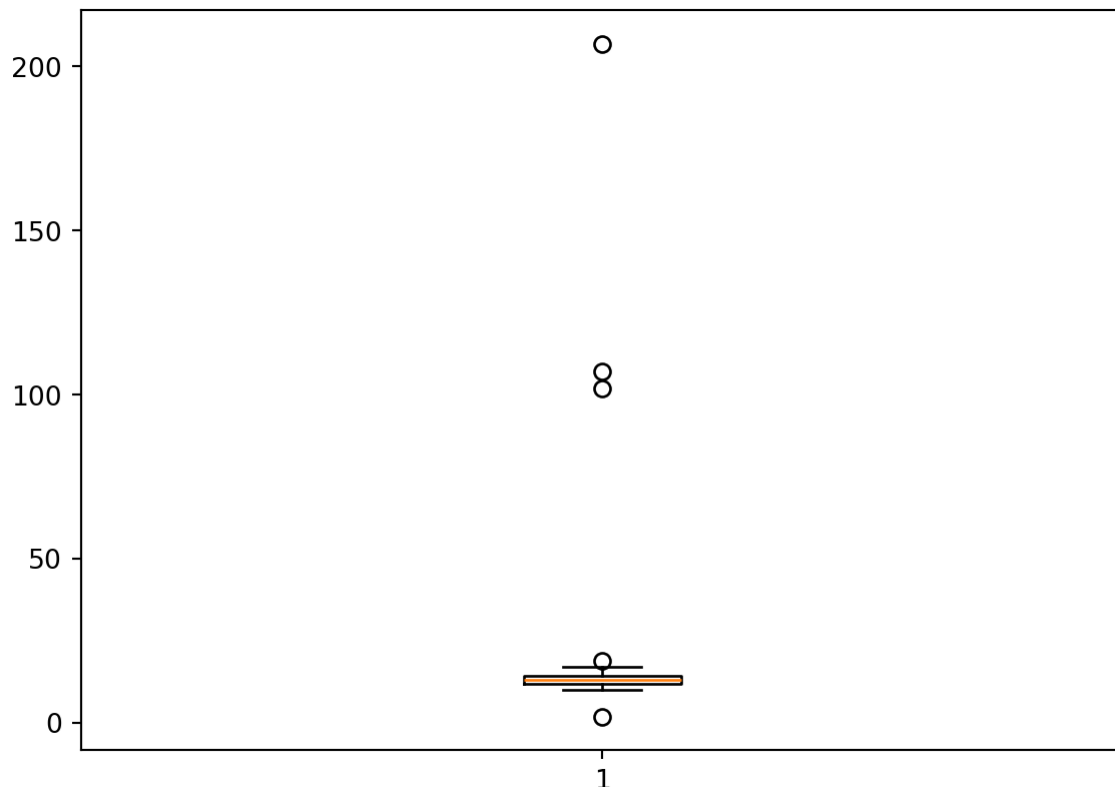


Figure 5.10: Figure showing box plot with outliers.

5.6.3 Removing Outliers

Once outliers are identified, you can choose to remove them to clean your dataset. This can be done by filtering the data.

- **Using Conditions:**

```
#import libraries  
import pandas as pd  
import numpy as np  
import seaborn as sns
```

```
data = sns.load_dataset('titanic')

# show the content of the 'age' column
print(f'the length of age column in Data is {len(data['age'])}')

# Remove outliers from the 'age' column
Q1 = data['age'].quantile(0.25)
Q3 = data['age'].quantile(0.75)
IQR = Q3 - Q1
filtered_data = data[~((data['age'] < (Q1 - 1.5 * IQR)) | (data['age'] > (Q3 + 1.5 * IQR)))]

# data without outliers
print(f'The length of age column in Data without outliers is {len(filtered_data['age'])}')

the length of age column in Data is 891
The length of age column in Data without outliers is 880

Here is an example before and after outliers are removed, you can also unfold the code to see the output.

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

data = np.array([2, 10, 12, 12, 13, 12, 11, 14, 13, 15, 102, 12, 14, 14, 17, 19, 107, 10, 13, 12, 14, 12, 207]) # Your data array

plt.figure(figsize=(4, 5))
fig1=sns.boxplot(data)

# Remove outliers
Q1 = np.quantile(data, 0.25)
Q3 = np.quantile(data, 0.75)
```

$IQR = Q3 - Q1$

```
filtered_data = data[~((data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR)))]
```

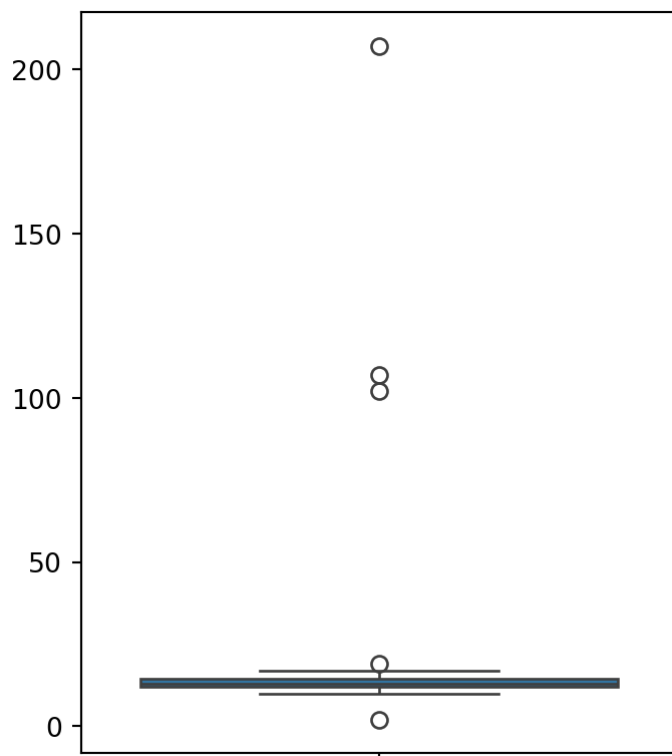
```
plt.figure(figsize=(4, 5))
```

```
fig2 = sns.boxplot(filtered_data)
```

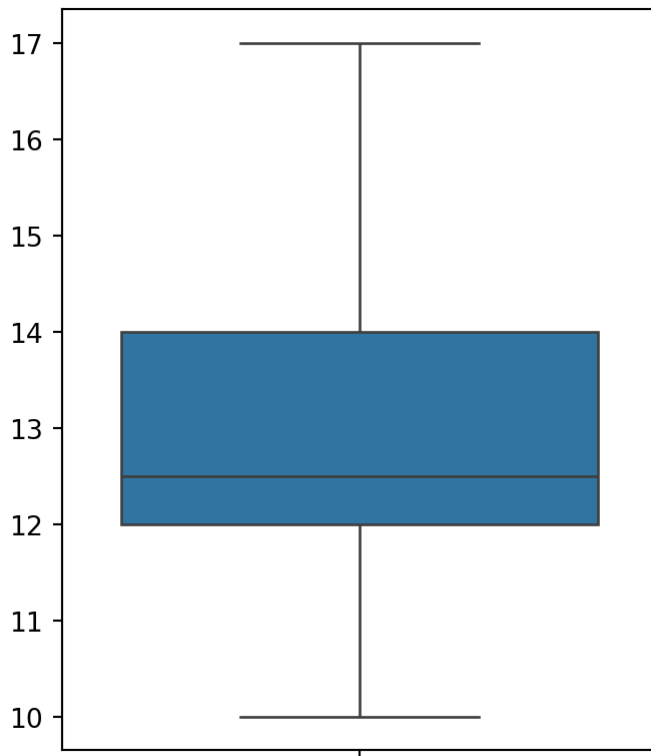
```
fig1
```

```
fig2
```

```
plt.show()
```



(a) Box plot with outliers



(b) Box plot without outliers

Figure 5.11: Figure showing box plot with and without outliers.

5.6.3.1 Important Considerations

- **Context Matters:** Before removing outliers, it's important to understand the context. Sometimes, outliers carry important information.
- **Impact on Dataset:** Removing outliers can significantly alter your results, especially in small datasets.

Using these methods in Python, you can effectively detect and handle outliers, ensuring that your data analysis or machine learning models are robust and reliable.

5.7 Graphical Methods

Graphical methods are a simple yet effective way to visualize the distribution of numerical variables. They show the number of observations in each category of a variable. Graphical methods are also known as **graphical displays**.

There are many different types of graphical methods, including: 1. **Frequency Tables** 2. **Bar Charts** 3. **Histograms** 4. **Box Plots** 5. **Scatter Plots** 6. **Line Plots** 7. **Pie Charts** 8. **Heat Maps** 9. **Venn Diagrams** 10. **Tree Maps** 11. **Word Clouds** 12. **Sankey Diagrams** 13. **Network Diagrams** 14. **Flow**





Charts 15. Cartograms 16. Choropleth Maps 17. Geographical Maps 18.
and many more!

I would suggest you to explore the [Andrew Abela's Chart Suggestions](#) to get a better understanding of which chart to use for which type of data.

5.7.1 Python libraries for data visualization

Python mein bohot se libraries hain jo data visualization ke liye use kiye jate hain. Kuch popular libraries hain:

1. [Matplotlib](#)
2. [Seaborn](#)
3. [Plotly](#)
4. [Bokeh](#)
5. [Altair](#)
6. [ggplot](#)
7. and many more...

Is section main ham kuch popular libraries ko use kar ke data visualization ke examples dekhenge.     or graphical methods se data ko understand karna seekhen gay.

5.7.2 Frequency Tables

Frequency tables are a simple way to visualize the distribution of categorical variables. They show the number of observations in each category of a variable. Frequency tables are also known as **contingency tables**.

5.7.2.1 Frequency Tables Ka Formula

Frequency tables ka formula hai:

$$\text{Frequency} = \frac{\text{Number of Observations in a Category}}{\text{Total Number of Observations}} \times 100\%$$

Frequency Tables, ya **Tadaad Ki Tables**, aik aham tool hain statistics aur data analysis mein. Ye tables data ko organize aur summarize karne ke liye use kiye jate hain, khaas taur par jab aapko data ke distribution ya patterns ko quickly samajhna ho.

5.7.2.2 Frequency Table Ki Structure

1. **Categories (Zumray):** Aapke data ke different groups ya classes.
2. **Frequency (Tadaad):** Har category mein kitni dafa woh value ya observation aayi hai.
3. **Relative Frequency (Nisbi Tadaad):** Ye batata hai ke har category ki frequency total observations ke hisse ke tor par kitni hai.
4. **Cumulative Frequency (Jammi Tadaad):** Ye batata hai ke kisi particular point tak total kitni frequencies accumulate ho chuki hain.

5.7.2.3 Frequency Tables Ki Misal 🌟

Example: Maan lijiye aap ek survey conduct kar rahe hain Islamabad ke ek school mein aur aapko yeh janna hai ke students rozana kitni dair TV dekhte hain. Aap categories bana sakte hain jaise “1 ghanta”, “2 ghante”, “3 ghante”, etc., aur phir count karte hain ke har category mein kitne students aate hain.

Frequency Tables in Python

```
import pandas as pd
```

```
# Example dataset: student survey on hours spent on social media daily
```

```
data = {  
    'Hours on Social Media': ['<1 hour', '1-2 hours', '2-3 hours', '3-4 hours', '>4  
hours'],  
    'Number of Students': [15, 30, 25, 10, 5]  
}
```

```
# Creating a DataFrame
```

```
df = pd.DataFrame(data)
```

```
# Displaying the DataFrame as a Frequency Table
```

```
Df
```

	Hours on social media	Number of Students
0	<1 hour	15
1	1-2 hours	30
2	2-3 hours	25
3	3-4 hours	10
4	>4 hours	5

Figure 5.12: Figure showing a frequency table in Python.

Let's have another example of titanic dataset.

Code

	Survived	Percentage
survived		
0	549	61.616162
1	342	38.383838

Figure 5.13: Figure showing a frequency table of Titanic dataset in Python.

Titanic Dataset results

In this table you can see that 61.6% of the passengers did not survive the Titanic disaster, while 38.4% survived.

5.7.2.4 Frequency Tables Ka Use 🛠️

1. Data Organization (Data Ko Tarteel Dena):

- Aap complex ya badi matra mein data ko asani se samajhne ke liye tarteel de sakte hain.
- **Example:** Karachi ke hospitals mein aane wale different types ke patients ka data organize karne ke liye.

2. Pattern Identification (Namoonay Ka Taayun):


- Data mein mojud patterns ya trends ko pehchanne mein madad milti hai.
- **Example:** Lahore mein kisi specific month mein hony wali traffic accidents ki frequency se traffic patterns ka analysis.

3. Decision Making (Faisla Sazi):

- Business ya policy decisions lene mein insights provide karta hai.
- **Example:** Peshawar ke ek retail store ke product sales data ko analyze kar ke inventory decisions lena.

4. Statistical Analysis (Shumariyati Tahlil):

- Kisi bhi further statistical analysis ya visualizations banane ke liye base provide karta hai.
- **Example:** Multan mein students ki examination performance analysis ke liye.

Frequency tables simple yet powerful tools hain jo data ko samajhne aur us par based decisions lene mein bohot madadgar sabit hote hain. Ye especially tab useful hote hain jab data sets bade hote hain ya jab aapko quick insights chahiye hote hain. 

5.7.3 Bar Charts

Bar charts are a simple yet effective way to visualize the distribution of categorical variables. They show the number of observations in each category of a variable. Bar charts are also known as **bar graphs**.

The bar chart is particularly appropriate for displaying discrete data with only a few categories. The bars can be plotted vertically or horizontally. The height or length of each bar is proportional to the number of observations in the category.

The [Figure 5.14](#) shows the bar chart of the Titanic dataset, you can also unfold the code to see the output. The barchart shows the number of passengers who survived and who did not survive the Titanic disaster.

Code

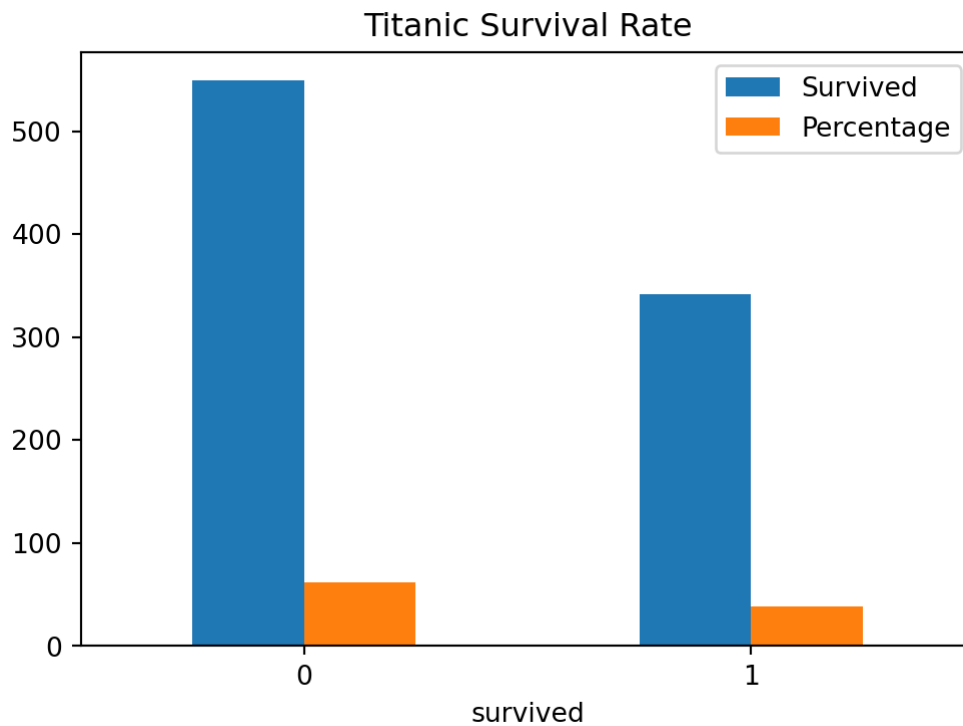


Figure 5.14: Figure showing a bar chart of Titanic dataset in Python.

Ap bar chart ko horizontal bhi bana sakte hain, jaisa ke [Figure 5.15](#) mein dikhaya gaya hai.

Code

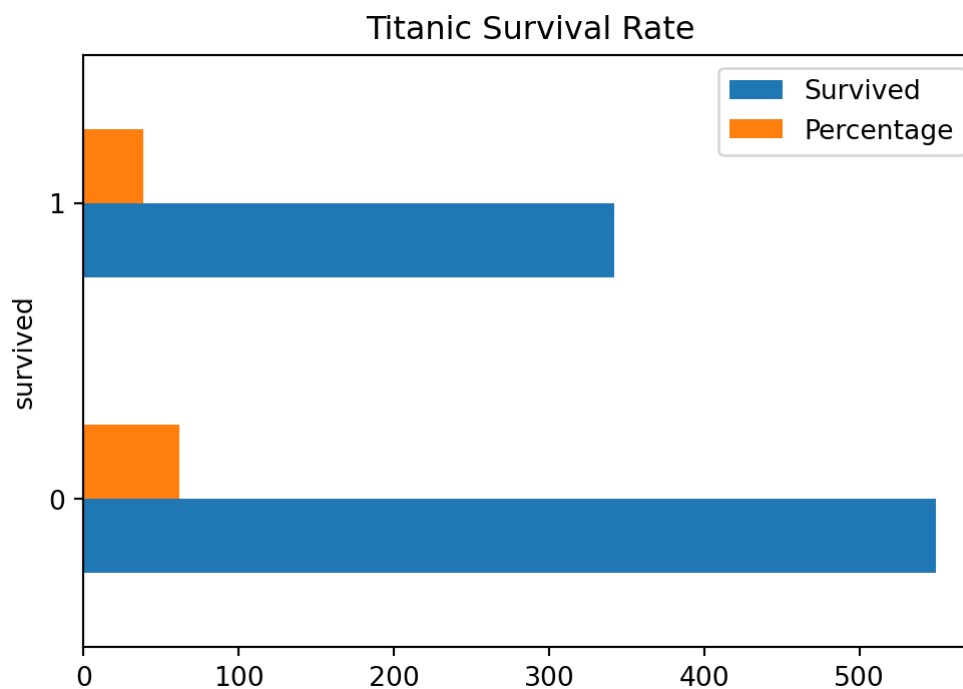


Figure 5.15: Figure showing a horizontal bar chart of Titanic dataset in Python.

5.7.3.1 Bar Charts Ka Formula 📊

Bar charts ka formula hai:

$$\text{Bar Height} = \frac{\text{Number of Observations in a Category}}{\text{Total Number of Observations}} \times 100\%$$

Bar Charts, ya **Bar Graphs**, aik aham tool hain statistics aur data analysis mein. Ye graphs data ko organize aur summarize karne ke liye use kiye jate hain, khaas taur par jab aapko data ke distribution ya patterns ko quickly samajhna ho.

You can also make stacked bar charts in Python. Stacked bar charts are used to show how a larger category is divided into smaller categories and what the relationship of each part has on the total amount. The [Figure 5.16](#) shows the stacked bar chart of the Titanic dataset, you can also unfold the code to see the output.

Code

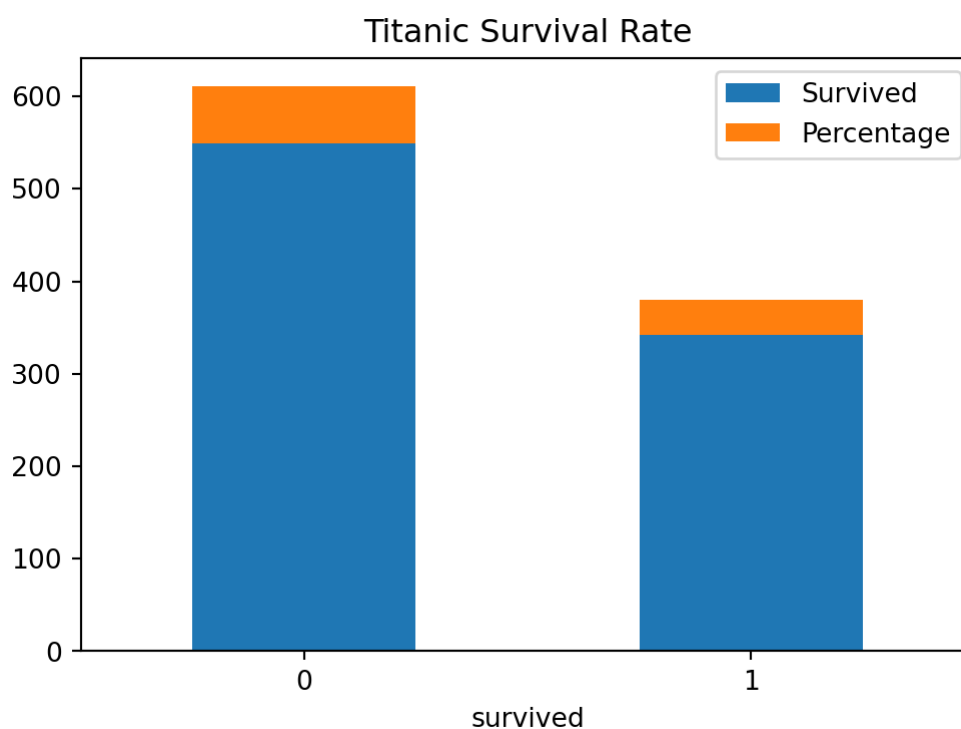


Figure 5.16: Figure showing a stacked bar chart of Titanic dataset in Python.

We can also draw bar charts using plotly library. The [Figure 5.17](#) shows the bar chart of the Titanic dataset using plotly library, you can also unfold the code to see the output.

Code

Figure 5.17: Figure showing a bar chart of Titanic dataset in Python using plotly library.

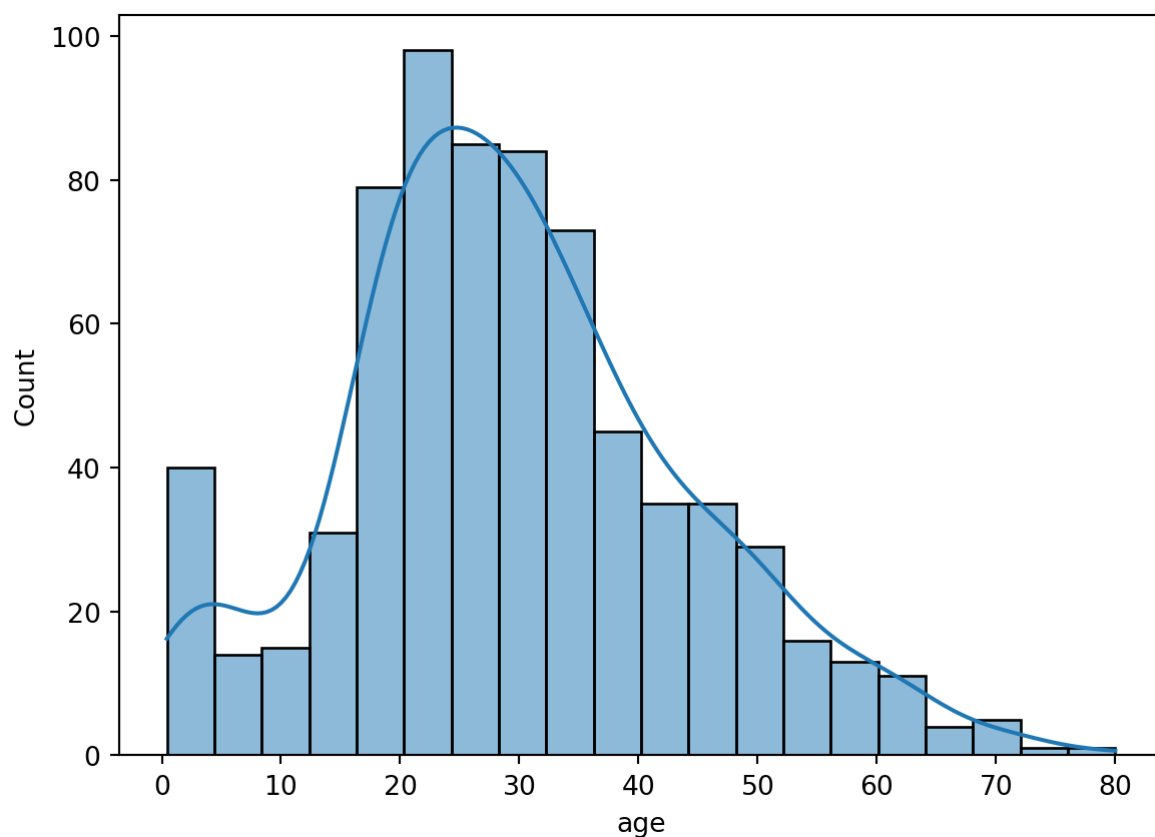
5.7.4 Histograms

Histograms are a simple yet effective way to visualize the distribution of numerical variables. They show the number of observations in each category of a variable. Histograms are also known as **frequency histograms**.

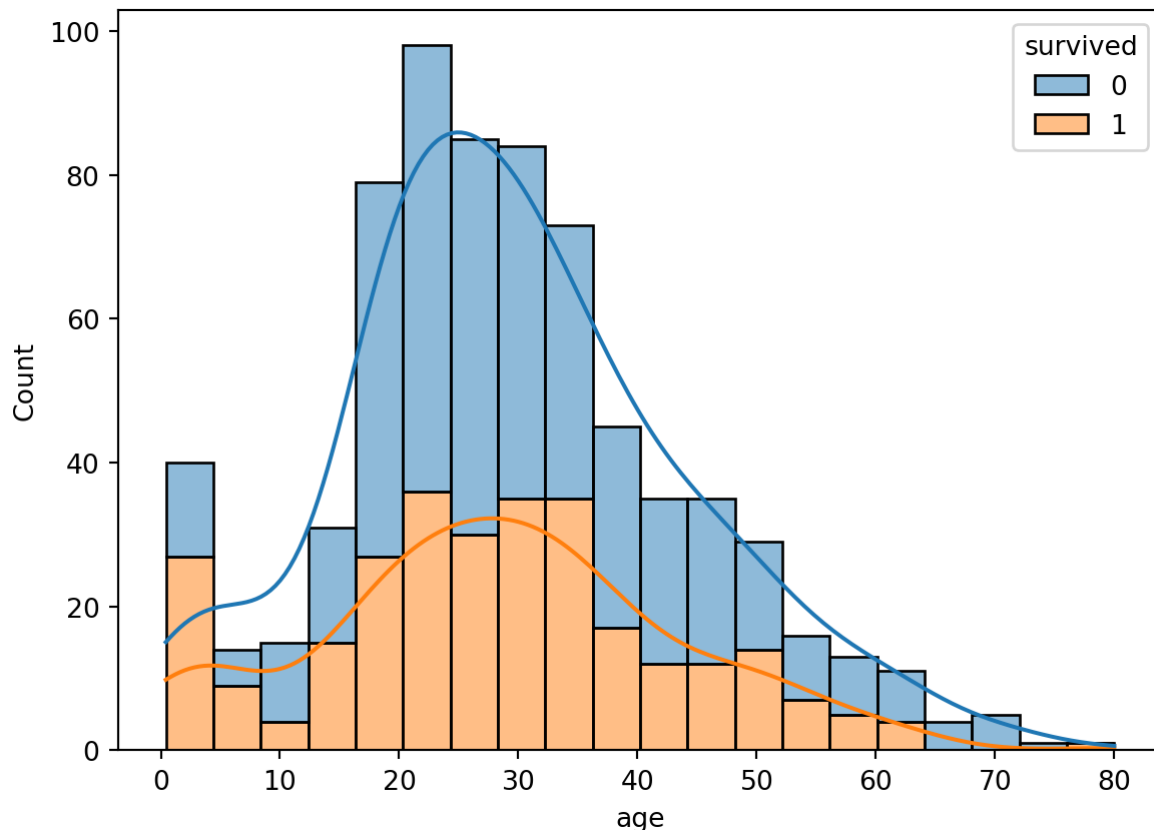
The histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. To construct a histogram, the first step is to “bin” (or “bucket”) the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent and are often (but are not required to be) of equal size.

The [Figure 5.18](#) shows the histogram of the Titanic dataset, you can also unfold the code to see the output. The histogram shows the number of passengers who survived and who did not survive the Titanic disaster based on their age.

Code



(a) Histogram of age column



(b) Histogram of Age column grouped by Survived column

Figure 5.18: Figure showing a histogram of Titanic dataset in Python.

5.7.5 Pie Charts

Pie charts are a simple yet effective way to visualize the distribution of categorical variables. They show the number of observations in each category of a variable. Pie charts are also known as **pie graphs**.

The pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented. The earliest known pie chart is generally credited to William Playfair's Statistical Breviary of 1801.

The [Figure 5.19](#) shows the pie chart of the Titanic dataset, you can also unfold the code to see the output. The pie chart shows the number of passengers who survived and who did not survive the Titanic disaster.

Code

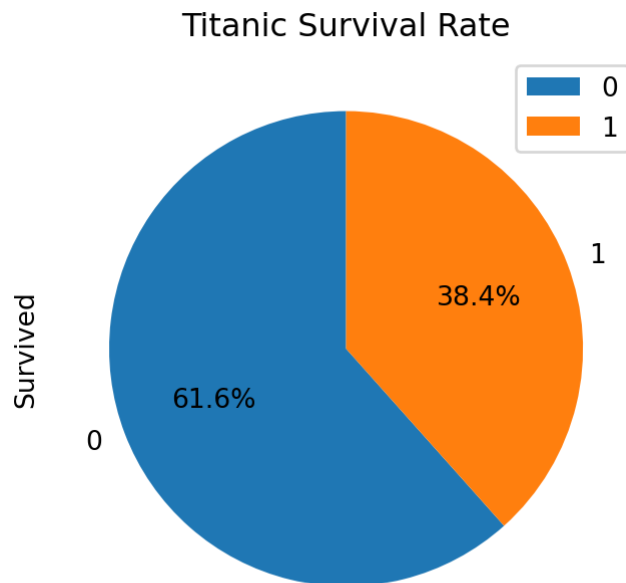


Figure 5.19: Figure showing a pie chart of Titanic dataset in Python.

We can also create pie chart using plotly to show the rate of survival of passengers in the Titanic disaster. The [Figure 5.20](#) shows the pie chart of the Titanic dataset using plotly library where data was grouped based on the class of travelling on titanic dataset. You can also unfold the code to see the output.

Code

Figure 5.20: Figure showing a pie chart of Titanic dataset in Python using plotly library.

5.7.6 Box Plots

Box plots are a simple yet effective way to visualize the distribution of numerical variables. They show the number of observations in each category of a variable. Box plots are also known as **box and whisker plots**.

The box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum. It is also known as a box and whisker plot. The box plot is compact and efficient, displaying only the most important summary statistics. It also allows for easy identification of any outliers and a visual representation of the data symmetry and skewness.

The [Figure 5.21](#) shows the box plot of the Titanic dataset, you can also unfold the code to see the output. The box plot shows the number of passengers who survived and who did not survive the Titanic disaster based on their age.

Code

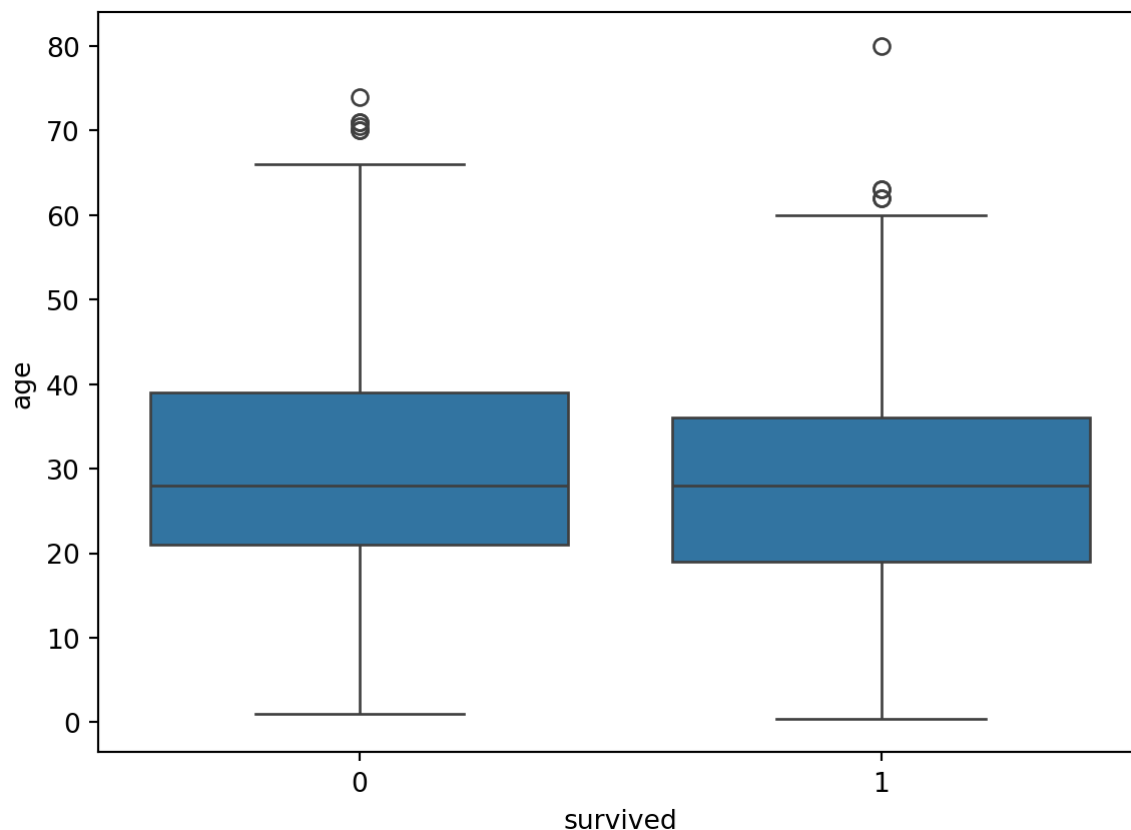


Figure 5.21: Figure showing a box plot of Titanic dataset in Python.

We can also show the box plot using plotly library. The [Figure 5.22](#) shows the box plot of the Titanic dataset using plotly library, you can also unfold the code to see the output. The box plot shows the number of passengers who survived and who did not survive the Titanic disaster based on their class.

Code

Figure 5.22: Figure showing a box plot of Titanic Survival Rate by Class and Age Python using plotly library.

5.7.7 Bi-variate Charts

Bi-variate charts are a simple yet effective way to visualize the relationship between two numerical variables. They show the number of observations in each category of a variable. Bi-variate charts are also known as **scatter plots**.

The scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are color-coded, one additional variable can be displayed. The data is displayed as a collection of points, each having the value of one variable

determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

The [Figure 5.23](#) shows the scatter plot of the Titanic dataset, you can also unfold the code to see the output. The scatter plot shows the relationship between the age and fare of the passengers who survived and who did not survive the Titanic disaster.

Code

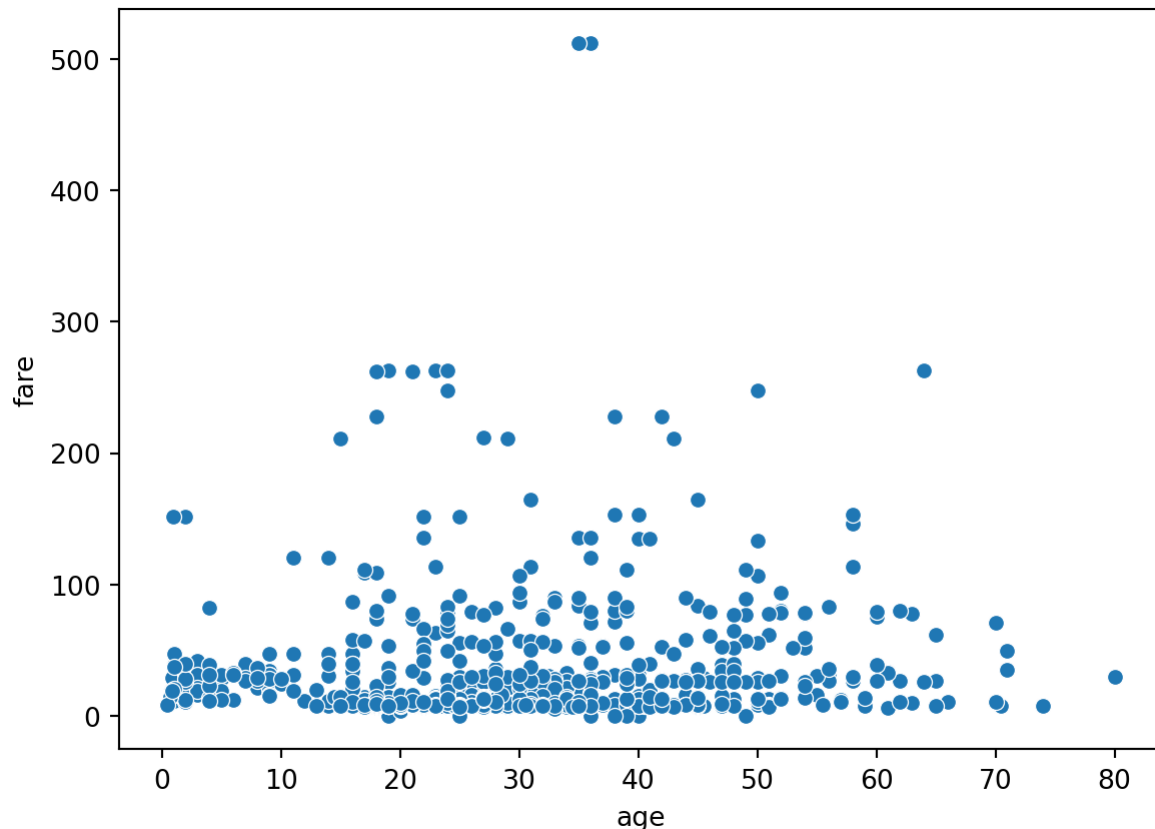


Figure 5.23: Figure showing a scatter plot of Titanic dataset in Python.

IN 2 variables ko ham further group kar sakte hain aur unke relationship ko visualize kar sakte hain. The [Figure 5.24](#) shows the scatter plot of the Titanic dataset, you can also unfold the code to see the output. The scatter plot shows the relationship between the age and fare of the passengers who survived and who did not survive the Titanic disaster based on their class.

Code

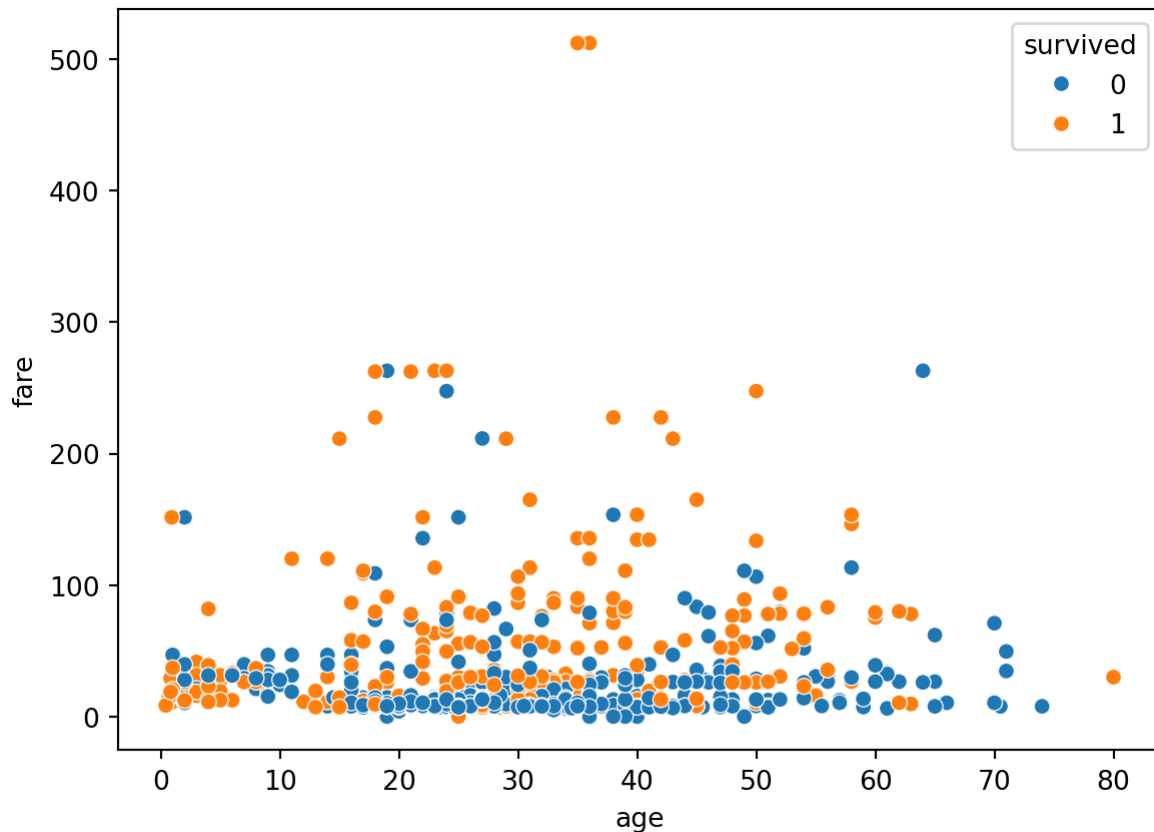


Figure 5.24: Figure showing a scatter plot of Titanic dataset in Python.

We can also draw scatter plots using plotly library. The [Figure 5.25](#) shows the scatter plot of the Titanic dataset using plotly library, you can also unfold the code to see the output. The scatter plot shows the relationship between the age and fare of the passengers who survived and who did not survive the Titanic disaster based on their class.

Code

Figure 5.25: Figure showing a scatter plot of Titanic dataset in Python using plotly library.

5.7.8 Line Plots

Line plots are a simple yet effective way to visualize the relationship between two numerical variables. They show the number of observations in each category of a variable. Line plots are also known as **line graphs**.

The line chart or line graph is a type of chart that displays information as a series of data points called ‘markers’ connected by straight line segments. It is a basic type of chart common in many fields. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and

joined with straight line segments. A line chart is often used to visualize a trend in data over intervals of time – a time series – thus the line is often drawn chronologically.

The [Figure 5.26](#) shows the line plot of the Titanic dataset, you can also unfold the code to see the output. The line plot shows the relationship between the age and fare of the passengers who survived and who did not survive the Titanic disaster.

Code

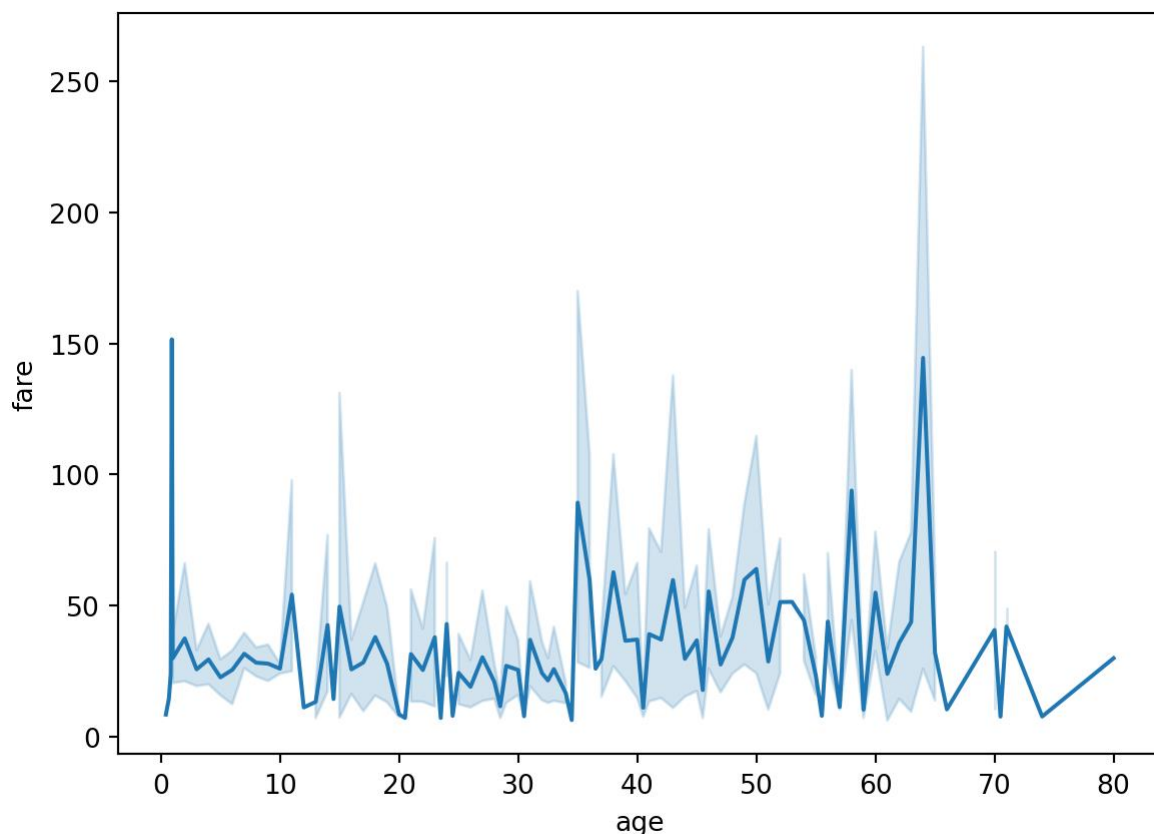


Figure 5.26: Figure showing a line plot of Titanic dataset in Python.

Ham line plot ko group kar ke bhi visualize kar sakte hain.

The [Figure 5.27](#) shows the line plot of the Titanic dataset, you can also unfold the code to see the output. The line plot shows the relationship between the age and fare of the passengers who survived and who did not survive the Titanic disaster based on their class.

Code

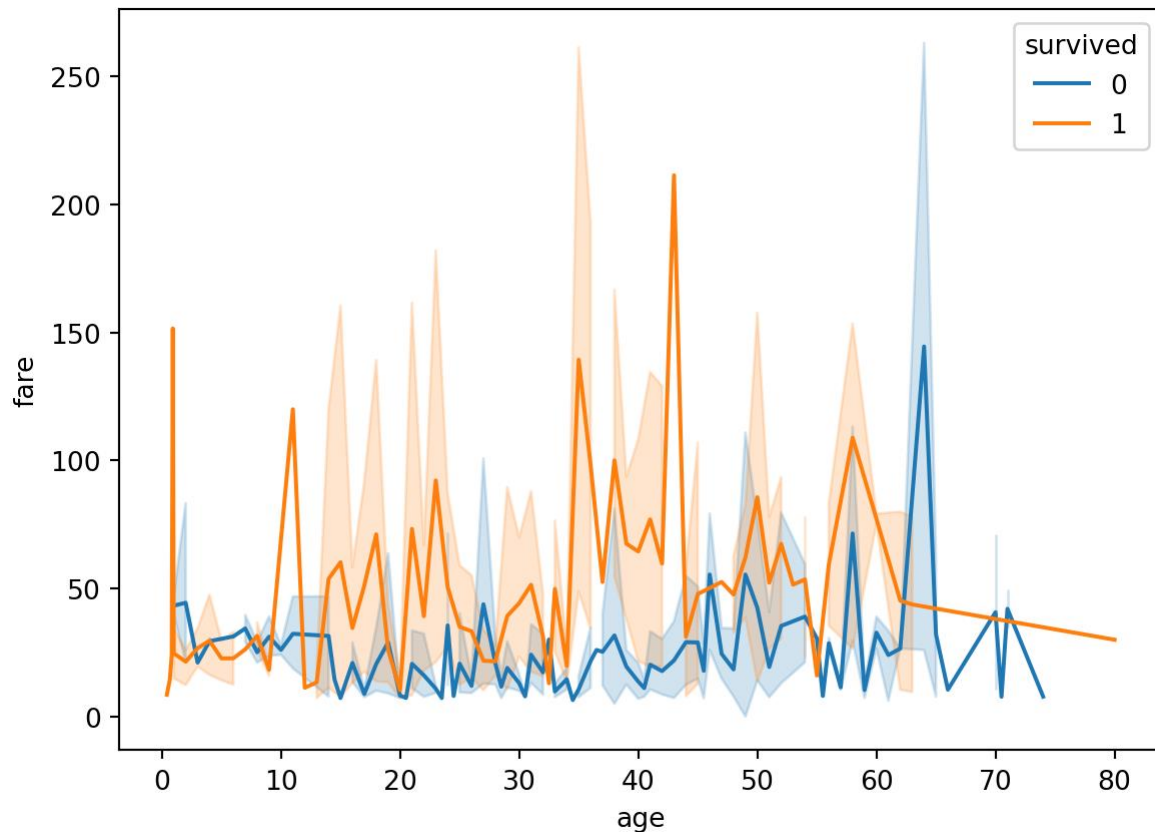


Figure 5.27: Figure showing a line plot of Titanic dataset in Python.

Data Composition Distribution Relationship Comparison Bar Chart Pie Chart
Tree Map Word Cloud Network Diagram Histogram Box Plot Scatter Plotline
Plot Scatter Plumbable Chart Bar Chart Line Plot Box Plot

Another Idea of Descriptive Statistics:

YesNoYesNoYesNoYesNoYesNoStart: Choose a Chart Do you have time series data? Use Line chart Do you need to compare parts to a whole? Use Pie Chart or Bar Chart Do you need to compare items? Use Bar Chart or Column Chart Do you have relational data? Use Scatter Plotted to show distribution? Use Histogram or Box Plot Consider other types of charts or revise data presentation needs

CHAPTER 6

Outliers and Missing Values

.1 Outliers

Data Mein Ghair Mamooli Cheezon ki Talash

Jab hum data ko samajhne aur us se insights nikalne ki baat karte hain, to kuch aise elements hote hain jo baqi data se hat ke hote hain. Inhe hum ‘anomalies’ ya ‘outliers’ kehte hain. Is chapter mein, hum inhi anomalies ko kaise pehchanein, unka kya asar hota hai, aur unhen kaise handle karein, is par baat karenge.

Outliers woh data points hote hain jo baqi data set se kafi alag hote hain.

- **Misal:** Aapke shehar mein, agar aksar temperature 20°C se 35°C ke darmiyan hota hai, to ek din ka temperature 50°C hona ek outlier hoga.
- **Ahmiyat:** Outliers ko identify karna zaroori hai kyun ke ye kabhi-kabhi data collection mein error ya kisi unusual event ki nishani ho sakte hain.

Outliers are also known as: 1. Abberant observations 2. Deviants 3. Outlying cases 4. Anomalous points 5. Abnormalities

6.1.1 Types of Outliers

Outliers nine types mein classify kiya ja sakta hai:

1. **Univariate:** Ye woh outliers hote hain jo sirf ek variable mein hote hain. For example, agar aapke data mein sirf age variable hai, to age ke outliers univariate outliers honge.
2. **Multivariate:** Ye woh outliers hote hain jo ek se zyada variables mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.
3. **Global:** Ye woh outliers hote hain jo poore data set mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.

4. **Local:** Ye woh outliers hote hain jo sirf ek cluster mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.
5. **Point:** Ye woh outliers hote hain jo sirf ek point mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.
6. **Contextual:** Ye woh outliers hote hain jo sirf ek cluster mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.
7. **Collective:** Ye woh outliers hote hain jo sirf ek cluster mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.
8. **Recurrent:** Ye woh outliers hote hain jo sirf ek cluster mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.
9. **Periodic:** Ye woh outliers hote hain jo sirf ek cluster mein hote hain. For example, agar aapke data mein age aur income dono variables hain, to age aur income ke outliers multivariate outliers honge.

6.1.2 Causes of Outliers

Outliers ki wajah kuch bhi ho sakti hai. Kuch common causes neeche diye gaye hain:

1. **Data Entry Errors:** Data ko enter karte waqt, kisi human error ki wajah se outliers ho sakte hain.
 - **Misal ke taur par,** agar aapke data mein age variable hai, aur kisi ne age ko 100 saal ki jagah 1000 saal enter kar diya, to ye ek outlier hoga.
2. **Measurement Errors:** Data ko measure karte waqt, kisi human error ki wajah se outliers ho sakte hain.
 - For example, agar aapke data mein height variable hai, aur kisi ne height ko 5 feet ki jagah 50 feet measure kar diya, to ye ek outlier hoga.
3. **Experimental Errors:** Data ko experiment karte waqt, kisi human error ki wajah se outliers ho sakte hain.

- For example, agar aapke data mein weight variable hai, aur kisi ne weight ko 50 kg ki jagah 500 kg measure kar diya, to ye ek outlier hoga.
- 4. **Intentional Outliers:** Kisi ne intentionally data mein outliers add kiye hon.
 - For example, agar aapke data mein age variable hai, aur kisi ne age ko 100 saal ki jagah 1000 saal enter kar diya, to ye ek outlier hoga.
- 5. **Data Processing Errors:** Data ko process karte waqt, kisi human error ki wajah se outliers ho sakte hain.
 - For example, agar aapke data mein age variable hai, aur kisi ne age ko 100 saal ki jagah 1000 saal enter kar diya, to ye ek outlier hoga.
- 6. **Sampling Errors:** Data ko sample karte waqt, kisi human error ki wajah se outliers ho sakte hain.
 - For example, agar aapke data mein age variable hai, aur kisi ne age ko 100 saal ki jagah 1000 saal enter kar diya, to ye ek outlier hoga.
- 7. **Natural Outliers:** Data mein outliers ki wajah natural events ho sakte hain.
 - For example, agar aapke data mein age variable hai, aur kisi ne age ko 100 saal ki jagah 1000 saal enter kar diya, to ye ek outlier hoga.

6.1.3 Why should we care about Outliers?

1. **Hidden Clues:** Outliers humein hidden clues dete hain. Inhe identify kar ke hum kisi hidden pattern ko discover kar sakte hain.
2. **Data Quality:** Outliers ki wajah se data quality kam ho jati hai. Inhe identify kar ke hum data quality ko improve kar sakte hain.
3. **Impact Analysis:** Outliers ki wajah se humari analysis mein error aa jata hai. Inhe identify kar ke hum analysis ko improve kar sakte hain.
4. **Better Decisions:** Outliers ki wajah se humari decisions par bhi asar padta hai. Inhe identify kar ke hum better decisions le sakte hain.
5. **Better Models:** Outliers ki wajah se humari models ki accuracy kam ho jati hai. Inhe identify kar ke hum better models bana sakte hain.
6. **Better Insights:** Outliers ki wajah se humari insights par bhi asar padta hai. Inhe identify kar ke hum better insights nikal sakte hain.

7. **Better Visualization:** Outliers ki wajah se humari visualizations ki quality kam ho jati hai. Inhe identify kar ke hum better visualizations bana sakte hain.
8. **Better Storytelling:** Outliers ki wajah se humari storytelling par bhi asar padta hai. Inhe identify kar ke hum better stories bana sakte hain.
9. **Better Data Products:** Outliers ki wajah se humari data products ki quality kam ho jati hai. Inhe identify kar ke hum better data products bana sakte hain.
10. **Better Data Science:** Outliers ki wajah se humari data science ki quality kam ho jati hai. Inhe identify kar ke hum better data science kar sakte hain.

6.1.4 Detect and remove Outliers

Outliers ko identify karne ke liye, hum kuch techniques use karte hain. In techniques ko hum 'Outlier Detection Techniques' kehte hain. In techniques mein se kuch neeche diye gaye hain:

1. **Z-Score**
2. **IQR**
3. **DBSCAN**
4. **Isolation Forest**
5. **Local Outlier Factor**
6. **Elliptic Envelope**
7. **One-Class SVM**
8. **Mahalanobis Distance**
9. **Robust Random Cut Forest**
10. **Histogram-based Outlier Score**
11. **K-Nearest Neighbors**
12. **K-Means Clustering**
13. **Local Correlation Integral**
14. and many more...

Ham sirf Z-Score, IQR or k-means clustering ko dekhenge.

6.1.5 Z-Score Method

Z-Score method mein, hum ye dekhte hain ke koi data point kitne standard deviations (SD) dur hai mean se.

Z-Score ki formula ye hai: $Z = \frac{x - \mu}{\sigma}$

Where:

Z: is the Z-Score

x: is the data point

μ : is the mean of the data

σ : is the standard deviation of the data

$x - \mu$: is the difference between the data point and the mean

$x - \mu\sigma$: is the difference between the data point and the mean in terms of standard deviations

Z-Score ki properties ye hain: 1. Z-Score ka mean 0 aur standard deviation 1 hota hai. 2. Z-Score ki value jitni zyada hogi, utna data point mean se zyada dur hoga. 3. Z-Score ki value jitni kam hogi, utna data point mean ke qareeb hoga. 4. Z-Score ki value 3 se zyada ya -3 se kam hogi, to data point outlier hoga.

Z-Score ki values ko interpret karne ke liye, neeche diye gaye table ko dekhein:

Z-Score	Data Point	Interpretation
-3	3 SDs below the mean	Outlier
-2	2 SDs below the mean	Outlier
-1	1 SD below the mean	Outlier
0	Mean	Not an outlier
1	1 SD above the mean	Not an outlier
2	2 SDs above the mean	Not an outlier
3	3 SDs above the mean	Not an outlier

6.1.5.1 Z-Score Method Example in Python

Z-Score method ko Python mein implement karne ke liye, neeche diye gaye steps follow karein:

6.1.5.1.1 Using numpy

Run the code below to see the steps.

Step 1: Import the required libraries

```
import pandas as pd
```

```
import numpy as np
```

Step 2: Create the data

```
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 50]})
```

Step 3: Calculate the mean and standard deviation

```
mean = np.mean(data['Age'])
```

```
std = np.std(data['Age'])
```

Step 4: Calculate the Z-Score

```
data['Z-Score'] = (data['Age'] - mean) / std
```

Step 5: Print the data

```
print("-----")
```

```
print(f'Here is the data with outliers:\n {data}')
```

```
print("-----")
```

Step 6: Print the outliers

```
print(f'Here are the outliers based on the z-score threshold, 3:\n {data[data['Z-Score'] > 3]}')
```

```
print("-----")
```

Step 7: Remove the outliers

```
data = data[data['Z-Score'] <= 3]
```

Step 8: Print the data without outliers

```
print(f'Here is the data without outliers:\n {data}')
```

Here is the data with outliers:

	Age	Z-Score
0	20	-0.938954
1	21	-0.806396
2	22	-0.673838
3	23	-0.541280
4	24	-0.408721
5	25	-0.276163
6	26	-0.143605
7	27	-0.011047
8	28	0.121512
9	29	0.254070
10	30	0.386628
11	50	3.037793

Here are the outliers based on the z-score threshold, 3:

	Age	Z-Score
11	50	3.037793

Here is the data without outliers:

	Age	Z-Score
0	20	-0.938954
1	21	-0.806396
2	22	-0.673838
3	23	-0.541280

```
4 24 -0.408721
5 25 -0.276163
6 26 -0.143605
7 27 -0.011047
8 28 0.121512
9 29 0.254070
10 30 0.386628
```

6.1.5.1.2 Using scipy library

You can also follow the steps below to implement the Z-Score method in Python, using scipy library:

Run the code below to see the steps.

```
# Import libraries
```

```
import numpy as np
```

```
from scipy import stats
```

```
# Sample data
```

```
data = [2.5, 2.7, 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0, 110.0]
```

```
# Calculate the Z-score for each data point
```

```
z_scores = np.abs(stats.zscore(data))
```

```
# Set a threshold for identifying outliers
```

```
threshold = 2.5
```

```
outliers = np.where(z_scores > threshold)[0]
```

```
# print the data
```

```
print("-----")
```

```
print("Data:", data)
print("-----")

print("Indices of Outliers:", outliers)
print("Outliers:", [data[i] for i in outliers])

# Remove outliers
data = [data[i] for i in range(len(data)) if i not in outliers]
print("-----")
print("Data without outliers:", data)
-----
Data: [2.5, 2.7, 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0, 110.0]
-----
Indices of Outliers: [9]
Outliers: [110.0]
-----
Data without outliers: [2.5, 2.7, 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0]
```

6.1.6 IQR Method

IQR method mein, hum ye dekhte hain ke koi data point kitne IQRs dur hai median se.

IQR ki formula ye hai:

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Where:

IQR: is the Interquartile Range

Q3: is the third quartile

Q1: is the first quartile

Q3–Q1: is the difference between the third quartile and the first quartile

IQR ki properties ye hain:

1. **IQR ka median 0 aur standard deviation 1 hota hai.**

2. **IQR ki value jitni zyada hogi, utna data point median se zyada dur hoga.**

3. **IQR ki value jitni kam hogi, utna data point median ke qareeb hoga.**

6.1.6.1 IQR Method Example in Python

IQR method ko Python mein implement karne ke liye, neeche diye gaye steps follow karein:

6.1.6.1.1 Using numpy

Run the code below to see the steps.

Step 1: Import the required libraries

```
import pandas as pd
```

```
import numpy as np
```

Step 2: Create the data

```
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 50]})
```

Step 3: Calculate the first and third quartile

```
Q1 = np.percentile(data['Age'], 25, interpolation = 'midpoint')
```

```
Q3 = np.percentile(data['Age'], 75, interpolation = 'midpoint')
```

Step 4: Calculate the IQR

```
IQR = Q3 - Q1
```

Step 5: Calculate the lower and upper bound

```
lower_bound = Q1 - (1.5 * IQR)
```

```
upper_bound = Q3 + (1.5 * IQR)
```

Step 6: Print the data

```
print("-----")
```

```
print(f'Here is the data with outliers:\n {data}')
print("-----")
# Step 7: Print the outliers
print(f'Here are the outliers based on the IQR threshold:\n {data[(data['Age'] <
lower_bound) | (data['Age'] > upper_bound)]}')
print("-----")
# Step 8: Remove the outliers
data = data[(data['Age'] >= lower_bound) & (data['Age'] <= upper_bound)]

# Step 9: Print the data without outliers
print(f'Here is the data without outliers:\n {data}')
-----

Here is the data with outliers:

    Age
0    20
1    21
2    22
3    23
4    24
5    25
6    26
7    27
8    28
9    29
10   30
11   50
-----
```

Here are the outliers based on the IQR threshold:

Age

11 50

Here is the data without outliers:

Age

0 20

1 21

2 22

3 23

4 24

5 25

6 26

7 27

8 28

9 29

10 30

6.1.7 Clustering Method (K-Means)

Clustering method mein, hum data points ko clusters mein divide karte hain. This can be done using the K-Means clustering algorithm. Where we specify the number of clusters we want to divide the data into. Then we assign each data point to a cluster. Then we calculate the distance of each data point from the centroid of the cluster it belongs to. Then we remove the data points that are farthest from the centroid of the cluster they belong to.

Use the code below to see the steps.

```
# Import library
```

```
from sklearn.cluster import KMeans
```

```
# Sample data
```

```
data = [[2, 2], [3, 3], [3, 4], [30, 30], [31, 31], [32, 32]]

# Create a K-means model with two clusters (normal and outlier)
kmeans = KMeans(n_clusters=2, n_init=10)
kmeans.fit(data)

# Predict cluster labels
labels = kmeans.predict(data)

# Identify outliers based on cluster labels
outliers = [data[i] for i, label in enumerate(labels) if label == 1]

# print data
print("Data:", data)
print("Outliers:", outliers)

# Remove outliers
data = [data[i] for i, label in enumerate(labels) if label == 0]
print("Data without outliers:", data)
Data: [[2, 2], [3, 3], [3, 4], [30, 30], [31, 31], [32, 32]]
Outliers: [[2, 2], [3, 3], [3, 4]]
Data without outliers: [[30, 30], [31, 31], [32, 32]]
```

6.1.8 Handling Outliers

Outliers ko handle karne ke liye, hum kuch techniques use karte hain. In techniques ko hum ‘Outlier Handling Techniques’ kehte hain. In techniques mein se kuch neeche diye gaye hain:

1. **Removing the outlier:** This is the most common method where all detected outliers are removed from the dataset.

2. **Transforming and binning values:** Outliers can be transformed to bring them within a range. Techniques like log transformation or square root transformation can be used.
3. **Imputation:** Outliers can also be replaced with mean, median, or mode values.
4. **Separate treatment:** In some use-cases, it's beneficial to treat outliers separately rather than removing or imputing them.
5. **Robus Statistical Methods:** Some of the statistical methods to analyze and model the data are less sensitive to outliers and provide more accurate results in the data.

I have explained some of these techniques in the [section above](#).. Where we remove the outliers using the Z-Score, IQR and K-Means clustering methods. You can also use the other techniques by yourself and practice them.

6.1.9 Conclusion

- Outliers in a dataset are observations that deviate dramatically from the rest of the data points. They might arise as a result of data gathering mistakes or abnormalities, or they can be real findings that are just infrequent or extraordinary.
- If outliers are not appropriately accounted for, they might produce misleading, inconsistent, and erroneous findings. As a result, identifying and dealing with outliers is critical in order to produce accurate and useful data analysis findings.
- Outliers may be detected using a variety of methods, including the percentile approach, IQR method, and z-score method. Outliers can be dealt with in a variety of methods, including removal, transformation, imputation, and so on.

6.2 Missing Values

Missing Values Ko Kaise Handle Kiya Jaye? Aur Inhe Handle Karna Kyun Zaroori Hai?" - Data Science Ki Dunia Mein Iska Role 🤔🔧

Missing values yaani ghaib data se guzarne wala har data scientist ya researcher ko iski ahmiyat aur isse judi mushkilaat ka andaza ho sakta hai. Data Science ki duniya mein, yeh missing values se guzarne ka tajurba aksar humein milta hai. Agar aap mein se kuch khush naseeb hain jo is masle se guzre nahi, toh woh waqai kismat wale hain! 😊 Lekin un logon ke liye jo is masle ka samna karte

hain, unko yeh samajhne mein mushkil nahi hoti ke missing values kitne masail paida kar sakti hain.

6.2.1 Naukri, Missing Values aur Aik Bari Ghalti 🥲

Lahore ki ek mashhoor company Codanics Solutions mein Ahmed ek talented data scientist tha. Woh apne projects ko hamesha top priority deta tha aur is wajah se us ki company mein bhi bohat izzat thi. ✨

Ek roz, **Ahmed** apne doston ke sath lunch kar raha tha. 🍲

Ali (ek aur data scientist): “Ahmed bhai! Suna hai aap ko naya project mila hai?”

Ahmed: “Ji haan, Ali. Mujhe customers ki buying habits analyze karni hai. Lekin data mein kuch missing values hain, mujhe lagta hai koi masla nahi hoga agar main unhein ignore kar doon.” 😞

Ali: “Bhai, kabhi bhi missing values ko ignore mat karo. Yeh choti si baat model ki performance ko kharab kar sakti hai.”

Lekin **Ahmed** ne **Ali** ki baat ko nazar andaaz kiya aur apne tareeqe se kaam karna shuru kar diya.

Jab model tayyar hua aur us ko real-world data par test kiya gaya, to us ki predictions bilkul bhi sahi nahi thi. 😬 Company ko is wajah se bohat bada nuqsan hua.

CEO, **Mr. Usman**, ne Ahmed ko apne office mein bulaya. 🏢

Mr. Usman: “Ahmed, humein bohat zyada nuqsan hua hai is project se. Kya masla hai?”

Ahmed: “Sir, maine socha tha ke kuch missing values se koi masla nahi hoga. Lekin mujhe ab samajh aaya hai ke maine ghalat socha.” 😞

Mr. Usman: “Ahmed, aap jante hain data science mein kitni bhi choti ghalti badi problem create kar sakti hai. Mujhe afsos hai, lekin humein aap ko company se nikalna parega.”

Ahmed ko bohat afsos hua. Us ne realize kiya ke kabhi bhi data ko lightly nahi lena chahiye. Woh ghar wapas laut kar Ali ko call kiya. 📞

Ahmed: “Ali, tum sahi keh rahe the. Mujhe company se nikal diya gaya hai.”

Ali: “Afsos hai sun kar. Lekin Ahmed, har galti se humein kuch na kuch seekhne ko milta hai. Aap ab better tareeqe se kaam karenge.”

Ahmed ne apni galti se seekha aur woh ab missing values aur data preprocessing par khaas tawajjo dene laga. Chand mahine baad, Ahmed ne ek aur company mein job shuru ki, aur wahan us ne prove kiya ke woh ek maahir data scientist hai. Lekin, us ek ghalti ka sabak us ne hamesha yaad rakha.

Ab agar ap b ahmad ki trah risk lena chahtay hyn tu missing values ko seekhnay se pehlay ap is blog ko ignore kar den, warna agar ap interested hyn tu yaqeen manen ye blog ap ki Data Science or AI ki journey ko bht kamal karne wala hy, I know ap soch rahay hun gay k aisa kia hy is main, Q fir Pola Payen kareay Start? Han Bholay phir tayyar ho?

I know ye nick names hyn magar isi trah or b bht se nick names hyn missing values k, By the way ap apna nick name likhen gay comments main?

6.2.2 Missing Values k ultay naam

Agar ap b aik desi culture ki paidawar hyn tu ap k bhi bht saray ultay naam gay. hai na? like Achoo, Billa, Bhola, Pola, Saji, kala, chitta, mota, chota, kaddu etc., ye main nahi keh raha ap kahin b nazar dorayen tu aisay naaam htay hyn, or kuch tu bht hi adab se pukaray jatay hyn, jaisa k, Paye Kalay. Ab isi trah missing values k bhi naam hyn kaafi jo agar ap ko na pata hun tu ap preshan hun gay. Chalein phir dekhtay hyn!

Missing values ko mukhtalif namon se pukara jata hai, depend karta hai ke context kya hai aur kis domain ya field mein baat ho rahi hai. Lekin, Data Science aur statistics mein commonly istemal hone wale names hain:

1. **NA** (Not Available)
2. **NaN** (Not a Number): Khaas taur par programming languages jaise ke Python mein pandas library mein istemal hota hai.
3. **Null**: Database management systems jaise SQL mein istemal hone wala term hai.
4. **Undefined**
5. **Blank ya Empty**
6. **Placeholder Values**: Kabhi-kabhi kuch default values set ki jati hain jinhein hum recognize kar sakte hain ke yeh actual data nahi hai. Masalan, kisi age field mein -1 ya 999 set karna.

7. **Sentinel Values:** Yeh bhi ek tarah ke placeholder values hoti hain jo specific conditions ko represent karte hain.
8. **Dummy Data:** Placeholder ya test purpose ke liye istemal hoti hai.
9. **Missing Data:** Aam taur se research papers mein istemal hone wala term.

In tamaam terms mein se kuch specific situations ya tools ke liye hote hain, jabke baaz aam istemal ke liye hote hain. Hamesha zaroori hai ke jab aap data ko analyze ya preprocess kar rahe hoon, toh aap in different types ke missing values ko pehchanein aur unhein sahi tareeqay se handle karein.

6.2.3 How to Identify Missing Values?

Missing values ko identify karne ke liye, hum kuch techniques use karte hain. In techniques ko hum ‘Missing Value Detection Techniques’ kehte hain. In techniques mein se kuch neech diye gaye hain:

1. **Visual Inspection:** Data ko visualize kar ke missing values ko identify kiya jata hai.
2. **Descriptive Statistics:** Data ki descriptive statistics ko calculate kar ke missing values ko identify kiya jata hai.
3. **Missingno Library:** Missingno library ko use kar ke missing values ko identify kiya jata hai.

6.2.3.1 Visual Inspection

Visual Inspection mein, hum data ko visualize kar ke missing values ko identify karte hain.

Use the code below to see the steps.

```
# Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load titanic dataset
data = sns.load_dataset('titanic')
```

```
# Visualize the data
plt.figure(figsize=(8, 5))
sns.heatmap(data.isnull(), cbar=False)
plt.show()
```

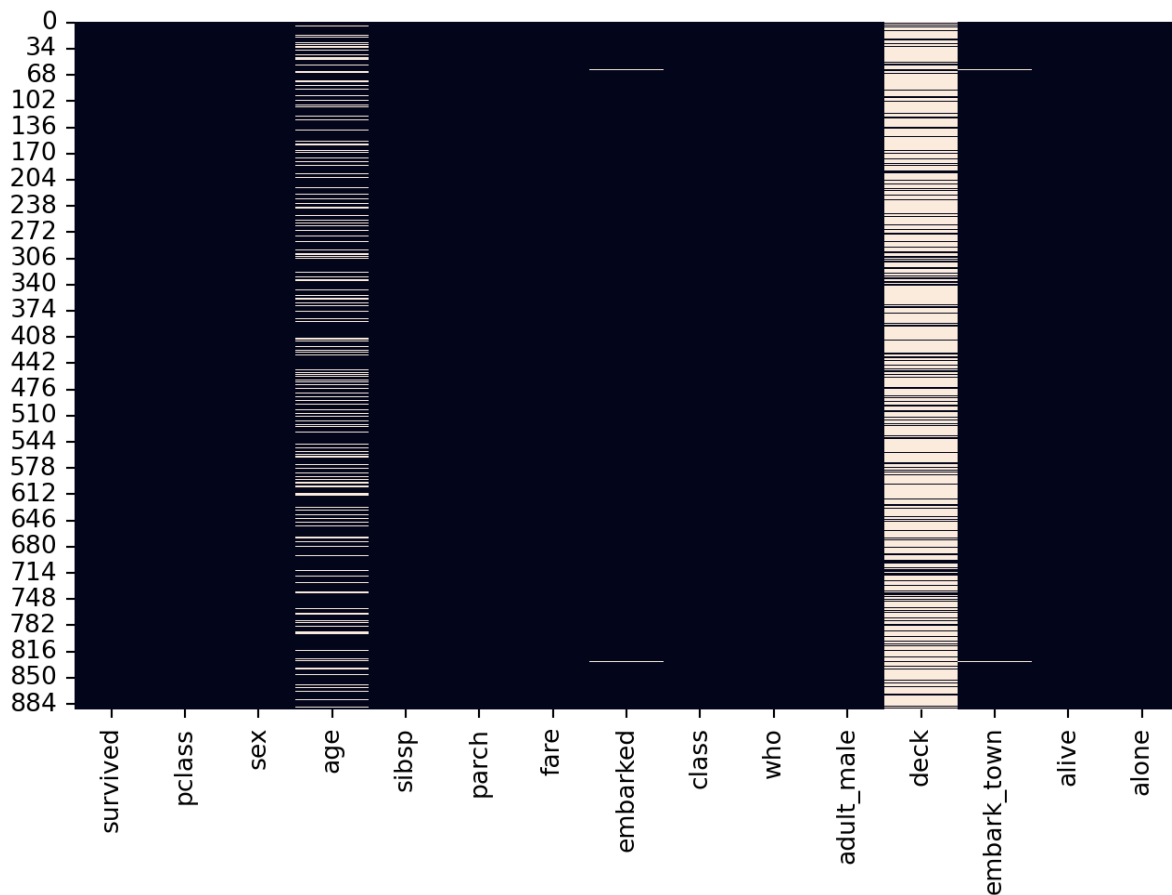


Figure 6.1: Visual Inspection of Missing Values of Titanic Dataset. The light colored lines represent missing values.

6.2.3.2 Descriptive Statistics

Descriptive Statistics mein, hum data ki descriptive statistics ko calculate kar ke missing values ko identify karte hain.

Use the code below to see the steps.

```
# Import libraries
import pandas as pd
import numpy as np
```

```
import seaborn as sns

# load titanic dataset

data = sns.load_dataset('titanic')


# calculate missing values

print("-----")

print(f'Missing values in each
column:\n{data.isnull().sum().sort_values(ascending=False)}')

print("-----")

print(f'Percentage of missing values in each
column:\n{round(data.isnull().sum() / len(data) * 100,
2).sort_values(ascending=False)}')
```

Missing values in each column:

deck	688
age	177
embarked	2
embark_town	2
survived	0
pclass	0
sex	0
sibsp	0
parch	0
fare	0
class	0
who	0
adult_male	0
alive	0

alone 0

dtype: int64

Percentage of missing values in each column:

deck 77.22

age 19.87

embarked 0.22

embark_town 0.22

survived 0.00

pclass 0.00

sex 0.00

sibsp 0.00

parch 0.00

fare 0.00

class 0.00

who 0.00

adult_male 0.00

alive 0.00

alone 0.00

dtype: float64

6.2.3.3 Missingno Library

Missingno library ko use kar ke bhi hum missing values ko identify kar sakte hain.

Use the code below to see the steps.

Code

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	N

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	N
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	N
...
886	0	2	male	27.0	0	0	13.0000	S	Second	man	True	N
887	1	1	female	19.0	0	0	30.0000	S	First	woman	False	B
888	0	3	female	NaN	1	2	23.4500	S	Third	woman	False	N
889	1	1	male	26.0	0	0	30.0000	C	First	man	True	C
890	0	3	male	32.0	0	0	7.7500	Q	Third	man	True	N

891 rows × 15 columns

Titanic Dataset

Import libraries

```
import pandas as pd
```

```
import numpy as np
```

```
import missingno as msno
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as snss
```

load titanic dataset

```
data = sns.load_dataset('titanic')
```

Visualize the data

```
msno.matrix(data, labels=True, fontsize=12, width_ratios=(2, 4), color=(0.2, 0.4, 0.6))
```

```
plt.show()
```

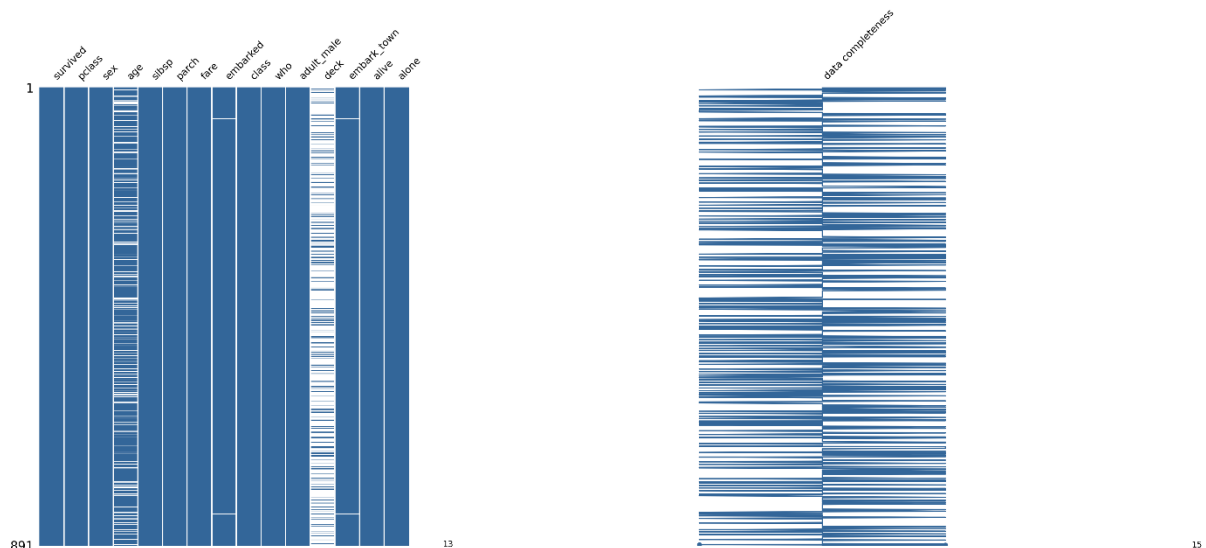



Figure 6.2: Missingno Library to Visualize Missing Values of Titanic Dataset. The white lines represent missing values.

6.2.4 Missing Values Handle Karna Kyun Itna Ahem Hai? 🧠

1. **Model Ki Accuracy Par Gehra Asar:** ❤️ Missing values ke honay se machine learning models ki accuracy mein kami aati hai, aur iski performance par bhi bura asar hota hai.
2. **Data Ki Mayari Par Sawal:** 📉 Missing values data ki mayari ko kamzor banate hain, jisse hamare analysis aur faislay mein bhi ghalat fehmiyan paida ho sakti hain.
3. **Model Training Ka Waqt Barh Jata Hai:** ⌚ Kabhi-kabhi, missing values ki wajah se model training ka waqt barh jata hai, jo ke resources aur waqt dono ka zaya hai.

6.2.5 Ruku Zara Sabr Karo

Missing values ka hona kisi bhi dataset mein aam baat hai, lekin jab hum decide karte hain ke kisi column ko remove karna chahiye ya nahi, to iska faisla humein kuch factors par depend karta hai:

Data Ki Quantity: Agar aapke paas bohat zyada data hai aur aik specific column mein missing values ki tadad bohat zyada hai (masalan, 70% ya 80%), toh us column ko remove kar dena behtar ho sakta hai, kyun ke us column se faida uthana mushkil ho sakta hai.

Column Ki Importance: Agar missing values wala column aapke analysis ya model ke liye bohat ahem hai, toh us column ko remove karna acha nahi hoga. Aise mein aap missing values ko impute karne ke tareeqe istemal kar sakte hain.

Nature of Data: Kabhi-kabhi, missing values ka hona bhi kuch indicate karta hai. Masalan, kisi survey mein, agar kisi sawal ka jawab nahi diya gaya, toh yeh indicate kar sakta hai ke participant us sawal se comfortable nahi tha. Aise mein, missing value ko hata dena ya replace karna sahi nahi hoga.

Model Ki Sensitivity: Kuch machine learning models missing values ko handle kar sakte hain, jabke kuch models sensitive hoti hain. Aise mein, agar model missing values ke sensitive hai, toh आपको missing values ko handle karna parega.

Type of Data: Numeric data mein missing values ko mean, median ya mode se replace kiya ja sakta hai. Categorical data mein, missing values ko mode ya kisi specific category se replace kiya ja sakta hai.

Aam taur par, agar aapke column mein 50% se zyada data missing hai, toh us column ko consider karna chahiye ke kya usse remove karna behtar rahega ya nahi. Lekin, yeh hard and fast rule nahi hai. Har dataset unique hota hai aur uski requirements bhi alag hoti hain. Is liye, आपको har dataset ke context mein decide karna hoga ke missing values ko kaise handle kiya jaye.

6.2.6 Missing Values Ko Handle Karne Ke Mufassal Tariqay 🧠

6.2.6.1 Maujooda Data Source Se Phir Se Data Hasil Karna: 🔄 Agar aap ke paas woh resource maujood hai jahan se aapne data liya tha, toh aap missing values ko wahan se dobara hasil kar sakte hain.

6.2.6.2 Mean, Median, Ya Mode Se Data Ko Impute Karna: 📊 Agar aapke paas numerical data hai, toh usmein missing values ko mean ya median se replace kiya jata hai. Wahi, categorical data ke liye mode ka istemal hota hai.

Use following code to see the steps to fill missing values with mean, median or mode in Python:

1. Mean

```
# Import libraries  
import pandas as pd  
import numpy as np
```

```
# Create the data
```

```
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, np.nan, 50]})
```

```
# Print the data with missing value
```

```
print("-----")
```

```
print(f'Here is the data with missing value:\n {data}')
```

```
# Calculate the mean
```

```
mean = data['Age'].mean()
```

```
# Replace the missing values with mean
```

```
data['Age'] = data['Age'].fillna(mean)
```

```
print("-----")
```

```
# Print the data without missing value
```

```
print(f'Here is the data without missing value:\n {data}')
```

```
-----
```

```
Here is the data with missing value:
```

```
    Age
```

```
0  20.0
```

```
1  21.0
```

```
2  22.0
```

```
3  23.0
```

```
4  24.0
```

```
5  25.0
```

```
6  26.0
```

```
7  27.0
```

```
8 28.0
9 29.0
10 NaN
11 50.0
```

Here is the data without missing value:

```
      Age
0 20.000000
1 21.000000
2 22.000000
3 23.000000
4 24.000000
5 25.000000
6 26.000000
7 27.000000
8 28.000000
9 29.000000
10 26.818182
11 50.000000
```

2. Median

```
# Import libraries
```

```
import pandas as pd
```

```
import numpy as np
```

```
# Create the data
```

```
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, np.nan, 50]})
```

```
# Print the data with missing value
print("-----")
print(f'Here is the data with missing value:\n {data}')

# Calculate the median
median = data['Age'].median()

# Replace the missing values with median
data['Age'] = data['Age'].fillna(median)
print("-----")
# Print the data without missing value
print(f'Here is the data without missing value:\n {data}')
-----
Here is the data with missing value:
    Age
0  20.0
1  21.0
2  22.0
3  23.0
4  24.0
5  25.0
6  26.0
7  27.0
8  28.0
9  29.0
10 NaN
11 50.0
```

Here is the data without missing value:

```
Age
0  20.0
1  21.0
2  22.0
3  23.0
4  24.0
5  25.0
6  26.0
7  27.0
8  28.0
9  29.0
10 25.0
11 50.0
```

3. Mode

```
# Import libraries
```

```
import pandas as pd
```

```
import numpy as np
```

```
# Create the data categorical data with mode and missing value
```

```
data = pd.DataFrame({'Fruit': ['Apple', 'Banana', 'Apple', 'Banana', 'Apple',  
                                'Banana', 'Apple', 'Banana', 'Apple', 'Banana', np.nan, 'Banana']})
```

```
# Print the data with missing value
```

```
print("-----")
```

```
print(f"Here is the data with missing value:\n {data}")
```

```
# Find the mode
```

```
mode = data['Fruit'].mode()[0]
```

```
# Replace the missing values with mode
```

```
data['Fruit'] = data['Fruit'].fillna(mode)
```

```
print("-----")
```

```
# Print the data without missing value
```

```
print(f'Here is the data without missing value:\n {data}')
```

```
-----
```

Here is the data with missing value:

Fruit

0 Apple

1 Banana

2 Apple

3 Banana

4 Apple

5 Banana

6 Apple

7 Banana

8 Apple

9 Banana

10 NaN

11 Banana

```
-----
```

Here is the data without missing value:

Fruit

0 Apple
1 Banana
2 Apple
3 Banana
4 Apple
5 Banana
6 Apple
7 Banana
8 Apple
9 Banana
10 Banana
11 Banana

6.2.6.3 Forward Ya Backward Fill Ka Istemal: 🧑🏻 🧑🏻 Kuch data sets mein waqt ya tarikh ka silsila hota hai. Aise data sets mein, aik row ke missing value ko pichli ya agli row ki value se replace kiya jata hai.

Use following code to see the steps to fill missing values with forward or backward fill in Python:

1. Forward Fill

```
# Import libraries
```

```
import pandas as pd
```

```
import numpy as np
```

Create the data

```
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, np.nan, 50]})
```

```
# Print the data with missing value
```

```
print("-----")
```



```
print(f'Here is the data with missing value:\n {data}')

# Replace the missing values with forward fill
data['Age'] = data['Age'].ffill()

print("-----")

# Print the data without missing value
print(f'Here is the data without missing value:\n {data}')
```

Here is the data with missing value:

	Age
0	20.0
1	21.0
2	22.0
3	23.0
4	24.0
5	25.0
6	26.0
7	27.0
8	28.0
9	29.0
10	NaN
11	50.0

Here is the data without missing value:

	Age
0	20.0
1	21.0

2 22.0

3 23.0

4 24.0

5 25.0

6 26.0

7 27.0

8 28.0

9 29.0

10 29.0

11 50.0

2. Backward Fill

Import libraries

import pandas as pd

import numpy as np

Create the data

```
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, np.nan, 50]})
```

Print the data with missing value

```
print("-----")
```

```
print(f'Here is the data with missing value:\n {data}')
```

Replace the missing values with backward fill

```
data['Age'] = data['Age'].bfill()
```

```
print("-----")
```

Print the data without missing value

```
print(f'Here is the data without missing value:\n {data}')
```

Here is the data with missing value:

	Age
0	20.0
1	21.0
2	22.0
3	23.0
4	24.0
5	25.0
6	26.0
7	27.0
8	28.0
9	29.0
10	NaN
11	50.0


Here is the data without missing value:

	Age
0	20.0
1	21.0
2	22.0
3	23.0
4	24.0
5	25.0
6	26.0
7	27.0
8	28.0

9 29.0

10 50.0

11 50.0

6.2.6.4 KNN Imputation Ka Istemal:  Yeh ek advanced technique hai jahan missing value ko uske aas-paas ke data points ke average value se replace kiya jata hai. Aise libraries jaise scikit-learn mein yeh method maujood hai.

Use following code to see the steps to fill missing values with KNN imputation in Python:

```
# Import libraries

import pandas as pd
import numpy as np
from sklearn.impute import KNNImputer

# Create the data

data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, np.nan,
50]})

# Print the data with missing value

print("-----")
print(f"Here is the data with missing value:\n {data}")

# Initialize the KNNImputer

imputer = KNNImputer(n_neighbors=2)

# Replace the missing values with KNN imputation

data['Age'] = imputer.fit_transform(data[['Age']])

print("-----")

# Print the data without missing value
```

```
print(f'Here is the data without missing value:\n {data}')
```

Here is the data with missing value:

Age

0	20.0
1	21.0
2	22.0
3	23.0
4	24.0
5	25.0
6	26.0
7	27.0
8	28.0
9	29.0
10	NaN
11	50.0

Here is the data without missing value:

Age

0	20.000000
1	21.000000
2	22.000000
3	23.000000
4	24.000000
5	25.000000
6	26.000000
7	27.000000

8 28.000000
9 29.000000
10 26.818182
11 50.000000

6.2.6.5 Deep Learning Techniques Ka Istemal:

🧠 Deep learning techniques jaise autoencoders bhi missing values ko handle karne mein madadgar sabit ho sakte hain.

Use following code to see the steps to fill missing values with deep learning techniques in Python:

```
# Import libraries
import pandas as pd
import numpy as np
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer

# Create the data
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, np.nan,
50]})

# Print the data with missing value
print("-----")
print(f'Here is the data with missing value:\n {data}')

# Initialize the IterativeImputer
imputer = IterativeImputer()

# Replace the missing values with deep learning techniques
data['Age'] = imputer.fit_transform(data[['Age']])
print("-----")
```

Print the data without missing value

```
print(f'Here is the data without missing value:\n {data}')
```

Here is the data with missing value:

Age

```
0  20.0
1  21.0
2  22.0
3  23.0
4  24.0
5  25.0
6  26.0
7  27.0
8  28.0
9  29.0
10 NaN
11 50.0
```

Here is the data without missing value:

Age

```
0  20.000000
1  21.000000
2  22.000000
3  23.000000
4  24.000000
5  25.000000
6  26.000000
```

```
7 27.000000
8 28.000000
9 29.000000
10 26.818182
11 50.000000
```

6.2.6.6 Simply Delete Kar Dena: ✗ Agar aapke data set mein missing values ki tadad bahut kam hai, toh aap us specific row ya column ko bhi delete kar sakte hain.

Use following code to see the steps to delete missing values in Python:

Import libraries

```
import pandas as pd
import numpy as np
```

Create the data

```
data = pd.DataFrame({'Age': [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, np.nan, 50]})
```

Print the data with missing value

```
print("-----")
print(f"Here is the data with missing value:\n {data}")
```

Delete the rows with missing values

```
data = data.dropna()
print("-----")
# Print the data without missing value
print(f"Here is the data without missing value:\n {data}")
```

Here is the data with missing value:

Age

0 20.0
1 21.0
2 22.0
3 23.0
4 24.0
5 25.0
6 26.0
7 27.0
8 28.0
9 29.0
10 NaN
11 50.0

Here is the data without missing value:

Age

0 20.0
1 21.0
2 22.0
3 23.0
4 24.0
5 25.0
6 26.0
7 27.0
8 28.0
9 29.0
11 50.0

6.2.6.7 Agar main na handle karun tu?

Bachoo Jee! phir tu hargiz model acha kaam nahi kare ga, yehi nahi abhi or suneay!

Agar hum missing values ko nazar andaaz kar dein toh humein kai masail ka samna karna par sakta hai. Yahan kuch masail hain jo arise ho sakti hain:



Model Accuracy Mein Kami: Machine learning models ki accuracy kam ho sakti hai, kyun ke model ko complete information nahi milti.



Ghalat Analysis: Data analysis mein ghalat nataij nikal sakte hain, jo ke decisions par negative asar dal sakta hai.



Model Confusion: Kuch models missing values handle nahi kar pate, jis se model train nahi ho pata ya phir ghalat predictions karta hai.



Bias in Model: Missing values ki wajah se model mein bias aane ka khatra barh jata hai.



Data ka Ghalat Interpretation: Missing values ki wajah se humare paas adhoori ya ghalat malumat ho sakti hai, jis ki wajah se hum data ko ghalat tareeqe se interpret kar sakte hain.



Storage Issues: Agar missing values ko replace nahi kiya jaye toh storage mein bhi masail ho sakti hain, kyun ke kuch systems missing values ko store nahi kar pate.



Data Integration Masail: Different sources se aane wale data mein agar missing values hain toh integration mein masail ho sakti hain.



Features ka Ghalat Selection: Missing values ki presence mein, kuch aham features ko ignore kiya ja sakta hai jin ka model par asar hona chahiye.



Ghalat Experimental Results: Science ya research projects mein, missing values ki wajah se ghalat experimental nataij aa sakte hain.



Stress aur Extra Kaam: Data scientists ko extra kaam karna par sakta hai tajziyat mein, kyun ke missing values ko identify aur handle karna parta hai.

Is liye, missing values ko handle karna bohat zaroori hota hai ta ke hum upar diye gaye masail se bach saken. 🛠️ 🔧 🔍

6.2.7 Conclusion

Missing Values - Ek Badi Challenge Lekin Ek Behtareen Mauqa Bhi

✳️ Missing values se guzarne ka tajurba har data scientist ke liye ek challenge toh hai hi, lekin isse humein yeh bhi seekhne ko milta hai ke hum kaise data ki mayari ko behtar bana sakte hain. Aakhir mein, behtar quality wale data se hi behtar aur zaheen insights aur models tayyar hoti hain.

CHAPTER 7

Inferential Statistics

Inferential Statistics, woh branch hai statistics ki jo data sets se **conclusions nikalne, predictions karne**, aur **hypotheses test karne** mein madad karti hai. Yeh method aksar chhote samples se liye gaye data par mabni hota hai, aur phir us data ke base par bade populations ke baare mein generalizations ya predictions kiye jaate hain.

7.1 Inferential Statistics Ke Main Components 🧩

7.1.1 Dependent vs. Independent Variable

7.1.1.1 Dependent Variables

Dependent Variables, ya **Inhsar Karne Wale Muttahidat**, woh variables hote hain jinki value ya status independent variable ke asar se badalti hai. Ye experiment ya study ke outcome ya result ko represent karte hain.

- **Example:** Usi school ki research mein, “class performance” dependent variable hai jo ke “neend ki miqdaar” se affect hota hai.

7.1.1.2 Independent Variables 🌟

Independent Variables, ya **Khud Mukhtar Muttahidat**, woh variables hote hain jin ki value ya status khud se set hoti hai aur jinka asar dosre variables par hota hai. Ye research ya experiment mein woh factors hote hain jinhe researcher change karta hai ya control karta hai taake dekha ja sake ke unka asar dependent variable par kya hota hai.

- **Example:** Maan lijiye aap Karachi ke kisi school mein research kar rahe hain ke kitni neend students ko mil rahi hai aur iska unki class performance par kya asar hota hai. Yahan, “neend ki miqdaar” (hours of sleep) ek independent variable hai.

7.1.1.3 Independent aur Dependent Variables Ka Taluq

In dono variables ka taluq research aur data analysis mein bohot ahem hota hai. Independent variable woh factor hai jise change kiya jata hai ya jo naturally vary hota hai, aur dependent variable woh outcome hai jise measure kiya jata

hai. Inka sahi identification aur use se researchers aur scientists ye samajh sakte hain ke ek factor dusre par kaise asar daalta hai.

Example: Agar Lahore mein air quality ka study kiya ja raha hai aur dekha jaa raha hai ke iska asar logon ki health par kya hai, to “air quality” independent variable hoga aur “logon ki health” dependent variable. Yahaan ye dekha jayega ke air quality ke behtar ya bura hone se logon ki health kaise affected hoti hai.




Here’s a table comparing Dependent and Independent Variables:

Aspect	Independent Variable	Dependent Variable
Definition	Variable that is changed or controlled in a scientific experiment.	Variable being tested and measured in scientific experiment.
Control in Experiments	Manipulated or selected by the researcher.	Observed and measured - changes in response to the independent variable.
Role in Analysis	Determines the conditions or groups in the study.	Outcome or result that is measured to the effect of the independent variable.
Desi Example	Temperature levels in a study to see its effect on crop growth.	Crop growth in response to different temperature levels.

- The **Independent Variable** is the “Temperature levels”. This is what the researcher would change or control to observe the effects.
- The **Dependent Variable** is the “Crop growth”. This is what the researcher is measuring, and it’s expected to change in response to the different temperature levels.

7.1.2 Hypothesis Testing (Hypothesis Ka Imtehan):

- Yeh statistical analysis ka ek method hai jisme aapke data ko test kiya jaata hai.
- Kisi hypothesis ya assumption ko test karne ke liye use kiya jata hai.
- **Example:** Farz karein, aap ye test karna chahte hain ke Islamabad aur Lahore mein students ki science understanding mein koi significant farq hai ya nahi.

Aaiye baat karte hain “Hypothesis Testing” (Hypothesis Ka Imtehan) ke baare mein! 

Hypothesis Testing, ek statistical method hai jo ke science aur data analysis mein use hota hai taake kisi specific assumption ya da'wa (hypothesis) ko test kiya ja sake. Is process mein, hum pehle ek hypothesis banate hain, phir data collect karte hain, aur uske baad statistical methods se ye dekhte hain ke kya humare collected data se wo hypothesis support hota hai ya nahi.

7.1.2.1 Hypothesis Testing Ke Steps 📐

1. **Null Hypothesis (Null Hypothesis, H_0):** Yeh wo basic assumption hota hai jo kehti hai ke koi khaas effect ya farq nahi hai.
2. **Alternative Hypothesis (Mutebadil Hypothesis, H_1):** Yeh hypothesis null hypothesis ke opposite hota hai, aur ye kehta hai ke koi effect ya farq hai.
3. **Data Collection (Data Ikattha Karna):** Relevant data collect karna jo hypothesis ko test karne ke liye zaroori ho.
4. **Statistical Test (Shumariyati Test):** Ek suitable statistical test perform karna taake data ko analyze kiya ja sake.
5. **Result (Nateeja):** Test ke results se conclusion nikalna ki kya null hypothesis ko reject karna chahiye ya nahi.

7.1.2.2 Hypothesis Testing Ki Misal ✨

Example: Maan lijiye, aap ye test karna chahte hain ke Karachi aur Lahore mein students ki math skills mein koi significant difference hai ya nahi. Yahan, aapka null hypothesis H_0 yeh hoga ke dono cities ke students ki skills mein koi farq nahi hai, aur alternative hypothesis H_1 yeh hoga ke ek significant difference hai. Phir aap dono cities ke schools se data collect karte hain aur statistical test jaise t-test ya ANOVA perform karte hain taake dekha ja sake ke kya aapka data H_0 ko reject karta hai ya nahi.

7.1.2.3 Hypothesis Testing Ki Importance 📈

Hypothesis testing research, science, aur business decisions mein bohot ahem hota hai. Yeh method humein yeh samajhne mein madad karta hai ke kya humare observations kisi real effect ki wajah se hain ya sirf chance ki wajah se. Is se hum informed decisions le sakte hain aur new theories ya products develop kar sakte hain.

Ye especially tab zaroori hota hai jab hume kisi assumption ko validate karna ho ya jab hum new discoveries ya insights ki talash mein hote hain. 📚🔍🌟📊

7.1.2.4 Case Study for Hypothesis Testing-Health Drink Ka Asar 🥤💡

Situation: Aap ek nutritionist hain aur aapne recently ek naya health drink develop kiya hai. Aapka da'wa (claim) hai ke yeh drink regular istemal karne se logon ki overall health mein significant improvement hota hai.

7.1.2.4.1 Hypothesis Formulation 📝

1. **Null Hypothesis H_0 :** Health drink ka koi significant asar nahi hota hai logon ki health par.
2. **Alternative Hypothesis H_1 :** Health drink regular istemal karne se logon ki health mein significant improvement hota hai.

7.1.2.4.2 Data Collection and Experiment Setup 📊

- Aap Lahore ke ek group of volunteers ko choose karte hain aur unhe randomly do groups mein divide karte hain: Ek group ko aap naya health drink dete hain aur dusre group ko aam drink (placebo).
- Phir aap unke health parameters jaise blood pressure, energy levels, aur immune response ko measure karte hain start se pehle aur phir kuch weeks ke regular use ke baad.

7.1.2.4.3 Statistical Test and Analysis 📈

- Aap statistical tests jaise t-test ya ANOVA ka istemal karte hain taake compare kiya ja sake ke dono groups mein health parameters mein kya significant changes aaye hain.
- Test ke results se pata chalta hai ke health drink wale group mein significant improvements hain compare karte hue placebo group se.

7.1.2.4.4 Result and Conclusion ✨

- Agar test ke results significant hote hain, toh aap H_0 ko reject kar sakte hain aur conclude kar sakte hain ke health drink ka asal mein significant positive asar hota hai health par.
- Agar results significant nahi hote, toh aap H_0 ko reject nahi kar sakte aur conclude karte hain ke drink ka koi special asar nahi hai.

7.1.2.5 Importance of This Test 📌

Yeh hypothesis testing ka example real-world mein product testing aur research mein kaise important role ada karta hai dikhata hai. Yeh method se aap product

ke claims ko scientifically validate kar sakte hain, jo ke consumers aur regulatory bodies ke liye confidence build karta hai. Aise tests se aapko valuable insights milte hain jo aapke product development aur marketing strategies ko guide karte hain. 🗑️ 🔍 ✨ 📚

7.2 Confidence Intervals:

- Yeh range batati hai jisme aapke true population parameter hone ki probability hoti hai.
- **Example:** Agar aap Karachi mein ek survey conduct karte hain aur pata lagate hain ke average monthly household income ki confidence interval kya hai. Bilkul, aaiye baat karte hain “Confidence Interval” (Yaqeen Ke Waqfay) ke baare mein, Roman Urdu aur emojis ke saath! 📊 📈

Confidence Interval (CI), ya **Yaqeen Ke Waqfay**, ek statistical term hai jo describe karta hai ke ek certain range mein, with a specific probability (confidence level), asal population parameter ki value kahaan gir sakti hai. Ye basically ek estimate hai ke aapke sample data se nikale gaye result kitne accurate hain jab unhe poore population par apply kiya jaye.

7.2.1 Confidence Interval Ka Structure 📐

- **Upper and Lower Bounds:** CI mein usually ek lower bound aur ek upper bound hota hai, jo ke yeh batate hain ke aapke estimated parameter ki asal value kis range mein ho sakti hai.
- **Confidence Level:** Commonly, 95% confidence level istemal kiya jata hai, lekin ye 90%, 99%, ya kisi aur level ka bhi ho sakta hai. Ye level batata hai ke kitni bar, agar hum bohot saare samples lein, to asal value is interval mein gir sakti hai.

Example: Maan lijiye aapne Lahore ke ek college mein students ka average test score calculate kiya hai. Aapka sample mean 70 hai aur aapne 95% confidence interval calculate kiya hai jo 68 se 72 ke beech hai. Iska matlab ye hai ke aap 95% sure hain ke poore college ke students ka average score is range mein hoga.

7.2.2 Confidence Interval Ka Importance 📌

1. **Data Ki Accuracy Ka Andaza:** Ye aapko batata hai ke aapke sample se nikale gaye estimates kitne reliable hain.

2. **Research Mein:** Scientific research mein, CI ka use often results ko present karne ke liye kiya jata hai taake readers ko yeh pata chale ke findings kitne sure hain.
3. **Business Decisions:** Business mein, CI ka istemal market research aur financial forecasting mein hota hai taake risk aur uncertainty ko quantify kiya ja sake.
4. **Policy Making:** Governments aur policymakers CI ka use kar sakte hain taake samajh sakein ke unke decisions kitne accurate hain based on the data available to them.

CI ek powerful statistical tool hai jo complex data ko samajhne aur us par based decisions lene mein key role ada karta hai. Ye especially tab zaroori hota hai jab hume data ke precision aur reliability ko quantify karna ho. 📊🔍🌟📈

The equation for calculating a Confidence Interval (CI) typically revolves around the standard error of the mean and a multiplier derived from the desired confidence level. For a simple CI around a sample mean, the equation is:

$$CI = \bar{x} \pm (Z \times SE)$$

Where: - \bar{x} is the sample mean. - Z is the Z-score corresponding to the desired confidence level (for example, 1.96 for a 95% confidence interval). - SE is the standard error of the sample mean.

Standard Error (SE) Equation:

The standard error of the sample mean is calculated as:

$$SE = \frac{s}{\sqrt{n}}$$

Where: - s is the sample standard deviation. - n is the sample size.

7.2.3 Putting It All Together:



For a 95% confidence interval, the Z-score is approximately 1.96. So, if your sample mean is 50, the sample standard deviation is 10, and your sample size is 100, the confidence interval would be calculated as:

1. Calculate the standard error: $SE = \frac{10}{\sqrt{100}} = 1$.
2. Multiply by the Z-score: $1.96 \times 1 = 1.96$.
3. Apply to the sample mean: 50 ± 1.96 , which gives you an interval of 48.04 to 51.96.

This means you can be 95% confident that the true population mean lies between 48.04 and 51.96. Remember, the confidence interval width can be influenced by both the variability in the data (as captured by the standard deviation) and the size of the sample, with larger samples typically yielding narrower intervals.

7.3 P-value (P-Value)





P Value, statistics mein ek ahem concept hai jo hypothesis testing mein istemal hota hai.

- **Tafseel:** P value woh probability hoti hai ke test ke results itne extreme ho jaise ke actual mein dekhe gaye, agar humara null hypothesis (basic assumption) sahi ho. Ye basically batata hai ke kisi given dataset par jab aik statistical model apply kiya jata hai, to observed results kitne unusual hain.
- **Ahmiyat:** P value ki madad se, researchers ye tay kar sakte hain ke unke results kisi real effect ki wajah se hain ya sirf ittefaq.  
- **Low vs High P Value:**
 - **Low P Value (Kam P Qiymat):** Iska matlab hai ke observed results shayad ittefaq nahi hain aur koi asal effect ho sakta hai. Aksar, P value 0.05 (5%) se kam hone par, hum null hypothesis ko reject karte hain.
 - **High P Value (Zyada P Qiymat):** Iska matlab hai ke observed results ittefaq se ho sakte hain aur koi significant effect nahi hai.



7.3.1 P-value threshold

P-value threshold, woh point hota hai jis par faisla kiya jata hai ke kya humare natije significant hain ya nahi.

- **Tafseel:** Jab aap koi hypothesis test karte hain, to aap pehle ye decide karte hain ke kis level par aap results ko significant samjhenge. Yehi level aapka P-value threshold hai.  
- **Common Threshold:** Aksar, scientists aur researchers 0.05 (ya 5%) ko as a standard threshold choose karte hain. Iska matlab hai ke agar P-value 0.05 se kam ho, to hum samajhte hain ke results statistically significant hain.  

- **Kaise Set Karein:**

- **Research Context ko Samjhein:** Threshold set karne se pehle, आपको अपने research ke context ko samajhna zaroori hai. Kuch studies mein zyada strict threshold (jaise 0.01) zaroori hota hai, khaas kar jahan results ka serious implications ho sakta hai.
- **Risk ko Madde Nazar Rakhein:** Lower threshold se aap Type I error (false positive) ka risk kam karte hain, lekin is se Type II error (false negative) ka risk badh sakta hai. Is liye, balance important hai. ⚖️ 📊
- **Field ke Standards:** Different fields ke different standards hote hain. Medical research mein shayad zyada strict standards hote hain as compared to social sciences. Is liye, apne field ke norms aur past studies ko bhi dekhein. 📖 📚

7.3.1.1 Ahmiyat (Importance) ✨

P-value threshold set karna research design ka aham hissa hota hai kyun ke yeh आपके results ki interpretation ko directly affect karta hai. Sahi threshold set kar ke, aap more reliable aur accurate conclusions tak pahunch sakte hain. 🎯 📊

p-value threshold

Setting the P-value threshold is a fundamental part of hypothesis testing, ensuring that the conclusions drawn from a study are based on sound statistical reasoning. The choice of threshold depends on the specific context of the research, the inherent risks of making errors, and the standards of the particular field of study.

7.3.2 Alpha α and P-value

Alpha α , statistics mein hypothesis testing ke context mein istemal hota hai aur isay significance level bhi kaha jata hai.

Alpha ko hum usually 0.05 (ya 5%) par set karte hain, lekin ye research ke context aur zarurat ke mutabiq vary ho sakta hai. Ye basically ek threshold hota hai jis par hum decide karte hain ke koi finding statistically significant hai ya nahi.

Alpha ko α se denote kiya jata hai aur ye probability hoti hai ke hum null hypothesis ko false positive taur par reject kar dein. Iska matlab hai ke agar aapka alpha 0.05 hai, to aap 5% tak acceptable samajhte hain ke aap galat taur par null hypothesis ko reject kar dein.

- **Tafseel:** Alpha woh probability hoti hai jis par hum decide karte hain ke koi finding statistically significant hai ya nahi. Ye basically Type I error (false positive) ki probability ko represent karta hai - yani ke, hum kitni probability tak acceptable samajhte hain ke hum galat taur par null hypothesis ko reject kar dein. 🎲🚫
- **Common Value:** Aksar, alpha ko 0.05 (ya 5%) par set kiya jata hai, lekin ye research ke context aur zarurat ke mutabiq vary ho sakta hai. 💡🔍

7.3.2.1 Alpha aur P-value ka Rishta (Alpha and P-value's Link) 🔗📈

Alpha aur P-value aapas mein closely linked hote hain aur hypothesis testing mein unka ek sath istemal hota hai.

- **Comparison:** Jab aap hypothesis test karte hain, to aap jo P-value calculate karte hain, usay apne set kiye gaye alpha se compare karte hain. Agar P-value alpha se kam hota hai, to hum null hypothesis ko reject karte hain aur conclude karte hain ke result statistically significant hai. 📊✅
- **Example:** Agar aapka alpha 0.05 hai aur aapka P-value 0.03 aata hai, to iska matlab hai ke aapke results ka statistical significance alpha level se zyada hai, aur aap null hypothesis ko reject kar sakte hain. 🔍📈
- **Balance in Decision Making:** Alpha ki value ka careful selection zaroori hai kyun ke ye Type I error (false positive) aur Type II error (false negative) ke risk ko balance karta hai. 😊⚖️

7.3.2.2 Ahmiyat (Importance) ✨

Alpha aur P-value ka sahi istemal aur unka aapas mein rishta samajhna research mein bohot ahem hota hai. Ye dono values mil kar help karti hain ke hum research ke results ko kis tarah interpret karein aur kitne confidence ke sath kisi conclusion tak pahunch sakte hain. 🎯📊

alpha and p-value

Alpha α aur P-value dono hypothesis testing ke critical elements hain. Alpha aapke research ke risk tolerance ko set karta hai, jabke P-value aapke data se milne wale evidence ki strength ko measure karta hai. In dono ka sahi use aur samajh research mein robust aur credible conclusions tak pahunchne mein madad karta hai.

7.4 Statistical Tests

Which statistical test to use depends on the type of data you have and the research question you want to answer. The following flowchart will help you choose the right statistical test for your data.

7.4.1 Types of Statistical Tests

Parametric vs. Non-Parametric Tests are the two main types of statistical tests. Parametric tests assume that the data is normally distributed, whereas non-parametric tests do not make this assumption. The following flowchart will help you choose the right statistical test for your data.

yesyesnoStatistical TestsCategorical DataNumerical DataChi-Square TestNormally Distributed DataData is HomogenousOne Sample Chi-Square TestTwo Samples Chi-Square TestParametric TestsNon-Parametric Tests

7.4.2 Z-test vs. t-test

Choosing between a Z-test and a T-test for hypothesis testing depends primarily on two factors: the sample size and whether the population standard deviation is known.

7.4.2.1 Z-test:

1. When to Use:

- The population standard deviation is known.
- The sample size is large (commonly, $n \geq 30$). With large samples, the sample standard deviation approximates the population standard deviation.
- For proportions (e.g., testing the proportion of success in a sample against a known population proportion).

2. Characteristics:

- Based on the Z-distribution, which is a normal distribution as n becomes large.
- More commonly used in quality control and standardization processes.

7.4.2.2 T-test:

1. When to Use:

- The population standard deviation is unknown.

- The sample size is small (typically, $n < 30$).
- Suitable for cases where the data is approximately normally distributed, especially in small samples.

2. Characteristics:

- Based on the T-distribution, which accounts for the additional uncertainty due to the estimation of the population standard deviation from the sample.
- T-distribution becomes closer to the normal distribution as the sample size increases.

7.4.2.3 General Guidelines:

- **Large Samples:** With large sample sizes, the T-test and Z-test will give similar results. This is because the T-distribution approaches the normal distribution as the sample size increases.
- **Small Samples:** When the sample size is small and the population standard deviation is unknown, the T-test is generally the appropriate choice due to its ability to account for the uncertainty in the standard deviation estimate.
- **Unknown Population Standard Deviation:** Even with large samples, if the population standard deviation is unknown and cannot be reliably estimated, a T-test is usually preferred.

7.4.2.4 Conclusion:

- Use the **Z-test** for large sample sizes or when the population standard deviation is known.
- Use the **T-test** for small sample sizes or when the population standard deviation is unknown.

In practice, the T-test is more commonly used in many research scenarios due to the rarity of knowing the population standard deviation and often dealing with smaller sample sizes.

7.4.3 Parametric Tests

Parametric tests are used when the data follows a normal distribution. The following flowchart will help you choose the right parametric test for your data.

Which test to choose from t-test and z-test? The following flowchart will help you.

YesNoYesNoParametric TestsOne VariableTwo VariablesIs sample size less than 30?t-testIs population standard deviation known?z-test

The following flowchart will help you choose the right parametric test for your data.

Parametric TestsOne VariableTwo VariablesThree or more VariablesTwo Continuous Dependent variablesCorrelationPearson's Correlation CoefficientOne Sample t-test/z-testIndependent SamplesOnly one Continuous Dependent variableTwo or more Continuous Dependent variablesOne FactorTwo FactorsThree or more FactorsOne-Way ANOVATwo-Way ANOVAThree or N-Way ANOVAMANOVA Multivariate Analysis of Variance

7.4.4 Non-parametric Tests

Non-parametric tests are used when the data does not follow a normal distribution. The following flowchart will help you choose the right non-parametric test for your data.

Non-Parametric TestsOne VariableTwo VariablesCompares one Continuous Dependent Variable to a fixed valueCompares Two Paired samplesWilcoxon Signed Rank TestCompares Two Independent samplesMann-Whitney U TestThree or more Independent samplesKruskal-Wallis TestCompares three or more median ranks, 1 variableKruskal-Wallis TestCompares three or more median ranks, 2 variablesFreidman Test

Now we will see how to perform these tests in Python.

7.4.5 Chi-Square Test (Chi-Square Test)

Chi-Square Test, categorical data ko analyze karne ke liye use hota hai. We can also call it Chi-squared test of independence. Is test mein, hum dekhte hain ke kya observed frequencies aur expected frequencies mein koi significant difference hai ya nahi.

7.4.5.1 Assumptions of Chi-Square Test

The assumptions of the Chi-Squared test are:

1. The variables under study are categorical (nominal or ordinal) variables.
2. The observations are independent of each other. This means that the occurrence of an outcome does not affect the other outcomes.

3. The data should be frequency counts of categories and not percentages or transformed data.
4. Each observed frequency, O_i , and expected frequency, E_i , should be greater than 5. If this assumption is violated, then the results might not be valid.
5. These assumptions need to be met for the Chi-Squared test to be valid.

7.4.5.2 Chi-square Test in Python

To perform a Chi-Squared test on the Titanic dataset in Python, we can use the `scipy.stats` library. The Chi-Squared test is often used to determine whether there is a significant association between two categorical variables. In this example, let's test whether there is a significant association between the 'Sex' (male or female) and 'Survived' (0 = No, 1 = Yes) variables in the Titanic dataset.

Null Hypothesis H_0 : There is no significant association between gender ('Sex') and survival ('Survived') on the Titanic. This means any observed differences in survival rates between genders in the dataset are due to chance and not due to an underlying relationship.

Alternative Hypothesis H_1 : There is a significant association between gender ('Sex') and survival ('Survived') on the Titanic. This implies that the differences in survival rates are not just due to chance but are influenced by the passengers' gender.

In hypothesis testing, the null hypothesis is what we attempt to disprove using our data. If the p-value from the Chi-Squared test is less than a certain threshold (commonly 0.05), we reject the null hypothesis in favor of the alternative hypothesis, concluding that there is indeed a statistically significant association between the two variables. If the p-value is greater than the threshold, we fail to reject the null hypothesis, meaning we do not have enough evidence to claim a significant association.

```
import pandas as pd
import numpy as np
import seaborn as sns
from scipy.stats import chi2_contingency
```



```
# Load the dataset
titanic = sns.load_dataset('titanic')

# Create a contingency table
contingency_table = pd.crosstab(titanic['sex'], titanic['survived'])

# Perform the Chi-Squared test
chi2, p, dof, expected = chi2_contingency(contingency_table)

# Display the results
print(f'Chi-Squared Value: {chi2}')
print(f'P-value: {p}')
print(f'Degrees of Freedom: {dof}')
print(f'Expected Frequencies:\n {expected}')
print("-----")
# print the results based on if else condition
if p < 0.05:
    print("Reject the null hypothesis, as there is a significant association between
the variables.")
else:
    print("Fail to reject the null hypothesis, as there is no significant association
between the variables.")
Chi-Squared Value: 260.71702016732104
P-value: 1.1973570627755645e-58
Degrees of Freedom: 1
Expected Frequencies:
[[193.47474747 120.52525253]
 [355.52525253 221.47474747]]
```

Reject the null hypothesis, as there is a significant association between the variables.

7.4.5.3 Explanation:

1. **Load Dataset:** We use Seaborn's built-in function to load the Titanic dataset.
2. **Create Contingency Table:** We create a contingency table (or cross-tabulation) between 'Sex' and 'Survived' using Pandas.
3. **Chi-Squared Test:** We use `chi2_contingency` from `scipy.stats` to perform the Chi-Squared test. This function returns the Chi-Squared value, the p-value, the degrees of freedom, and the expected frequencies if there was no association between the variables.
4. **Results:** The results are printed out. The p-value is used to determine the statistical significance. Typically, a p-value less than 0.05 indicates a statistically significant association between the variables.

As the p_value in this test is P-value: 1.1973570627755645e-58, we reject the null hypothesis.

7.4.5.4 Calculating Chi-Squared Manually

The Chi-Squared test statistic is calculated using the following formula:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

Where:

- χ^2 is the Chi-Squared statistic.
- O_i is the observed frequency.
- E_i is the expected frequency.
- The summation \sum is over all categories.

The expected frequency is calculated as:

$$E_i = (\text{row total}) \times (\text{column total}) / \text{grand total}$$

The Chi-Squared test compares the observed frequencies in each category of a contingency table with the frequencies expected under the null hypothesis, which is that the variables are independent.

7.4.6 t-test (t-test)

T-test, numerical data ko analyze karne ke liye use hota hai. Is test mein, hum dekhte hain ke kya observed means aur expected means mein koi significant difference hai ya nahi.

7.4.6.1 Assumptions of t-test

1. **Normality**: Data ka distribution normal hona chahiye.
2. **Independence**: Data points independent hona chahiye.
3. **Randomness**: Data points random hona chahiye.
4. **Sample Size**: Sample size chhota hona chahiye ($n < 30$).
5. **Scale**: Data ka scale interval ya ratio hona chahiye.
6. **Outliers**: Data mein outliers nahi hona chahiye.
7. **Linearity**: Data ka relationship linear hona chahiye.
8. **Homoscedasticity**: Data ka variance same hona chahiye.

7.4.6.2 Types of t-test

t-test One Sample t-test Two Samples t-test Independent Samples t-test Paired Samples t-test

1. **One Sample T-test**: Is test mein, hum dekhte hain ke kya aik sample ka mean kisi specific value se different hai ya nahi.
2. **Independent Samples T-test**: Is test mein, hum dekhte hain ke kya do samples ke means mein koi significant difference hai ya nahi.
3. **Paired Samples T-test**: Is test mein, hum dekhte hain ke kya do samples ke means mein koi significant difference hai ya nahi, jab ke wo samples related hain.

7.4.6.3 One Sample T-test

One Sample T-test, aik sample ka mean kisi specific value se different hone ki probability ko test karta hai. Is test mein, hum dekhte hain ke kya aik sample ka mean kisi specific value se different hai ya nahi.

```
# Import the required libraries
```

```
import numpy as np
```

```
from scipy.stats import ttest_1samp
```

```
# Create a sample data
ages = np.array([32, 34, 29, 29, 22, 39, 38, 37, 38, 36, 30, 26, 22, 22])

mu = 30 # mean of the population

# Perform the one-sample t-test
t_statistic, p_value = ttest_1samp(ages, mu)

# Print the results
print(f't-statistic: {t_statistic}')
print(f'p-value: {p_value}')
print("-----")
# print the results based on if else condition
if p_value < 0.05:
    print("Reject the null hypothesis,\n as the sample mean is significantly
different from the population mean.")
else:
    print("Fail to reject the null hypothesis,\n as the sample mean is not
significantly different from the population mean.")
t-statistic: 0.5973799001456603
p-value: 0.5605155888171379
-----
Fail to reject the null hypothesis,
as the sample mean is not significantly different from the population mean.
Try to change the value of mu and see how the results change. and if it becomes
significantly different or not.
```

7.4.6.4 Two Samples t-test

7.4.6.4.1 Independent Samples t-test

Independent Samples t-test, do independent samples ke means mein koi significant difference hone ki probability ko test karta hai. Is test mein, hum dekhte hain ke kya do samples ke means mein koi significant difference hai ya nahi.

```
# Import the required libraries
```

```
import numpy as np
```

```
from scipy.stats import ttest_ind
```

```
# Create two sample data
```

```
ages1 = np.array([32, 34, 29, 29, 22, 39, 38, 37, 38, 36, 30, 26, 22, 22])
```

```
ages2 = np.array([27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34])
```

```
# Perform the two-sample t-test
```

```
t_statistic, p_value = ttest_ind(ages1, ages2)
```

```
# Print the results
```

```
print(f't-statistic: {t_statistic}')
```

```
print(f'p-value: {p_value}')
```

```
# print the results based on if else condition
```

```
if p_value < 0.05:
```

```
    print("Reject the null hypothesis,\n as the sample means are significantly  
different.")
```

```
else:
```

```
    print("Fail to reject the null hypothesis,\n as the sample means are not  
significantly different.")
```

```
t-statistic: 2.0979439363492083
```

p-value: 0.04577767375684831

Reject the null hypothesis,

as the sample means are significantly different.

7.4.6.4.2 Paired Samples t-test

Paired Samples t-test, do related samples ke means mein koi significant difference hone ki probability ko test karta hai. Is test mein, hum dekhte hain ke kya do samples ke means mein koi significant difference hai ya nahi, jab ke wo samples related hain.

```
# Import the required libraries
```

```
import numpy as np
```

```
from scipy.stats import ttest_rel
```

```
# Create two sample data before and after
```

```
before = np.array([32, 34, 29, 29, 22, 39, 38, 37, 38, 36, 30, 26, 22, 22])
```

```
after = np.array([27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34])
```

```
# Perform the paired t-test
```

```
t_statistic, p_value = ttest_rel(before, after)
```

```
# Print the results
```

```
print(f't-statistic: {t_statistic}')
```

```
print(f'p-value: {p_value}')
```

```
# print the results based on if else condition
```

```
if p_value < 0.05:
```

```
    print("Reject the null hypothesis,\n as the before mean is significantly  
different to the after mean.")
```

```
else:
```

```
print("Fail to reject the null hypothesis,\n as the before mean is not  
significantly different to the after mean.")
```

t-statistic: 1.8116836198069448

p-value: 0.0931902262006994

Fail to reject the null hypothesis,

as the before mean is not significantly different to the after mean.

7.4.6.5 Role of Variance in t-test

Variance, data points ke spread ko measure karta hai.

- **Low Variance:** Data points close honge mean ke.
- **High Variance:** Data points far honge mean se.

Variance ka role t-test mein:

- **Independent Samples t-test:** Is test mein, hum dekhte hain ke do samples ke means mein koi significant difference hai ya nahi. Is test mein, agar dono samples ka variance same hoga, to hum ttest_ind use karenge, aur agar dono samples ka variance different hoga, to hum ttest_ind with unequal variance use karenge.

ttest_ind with unequal variance:

Step-1: Import libraries

```
import numpy as np
```

```
from scipy.stats import ttest_ind
```

Step 2: Create two datasets with unequal variance

```
np.random.seed(0) # for reproducibility
```

```
# Create a dataset 'ages1' with mean=30, standard deviation=3, size=100
```

```
ages1 = np.random.normal(30, 3, 100)
```

```
# Create a dataset 'ages2' with mean=30, standard deviation=10, size=100
```

```
ages2 = np.random.normal(30, 10, 100)
```

```
# Perform the two-sample t-test
t_statistic, p_value = ttest_ind(ages1, ages2, equal_var=False)

# Print the results
print(f't-statistic: {t_statistic}')
print(f'p-value: {p_value}')



# print the results based on if else condition
if p_value < 0.05:
    print("Reject the null hypothesis,\n as the sample means are significantly
different.")
else:
    print("Fail to reject the null hypothesis,\n as the sample means are not
significantly different.")
t-statistic: -0.5913989290785231
p-value: 0.5554059984405396
Fail to reject the null hypothesis,
as the sample means are not significantly different.
```

When to use which test from t-test of z-test?

YesNoYesNoIs sample size > 30?Z-testIs population standard deviation known?T-test

7.4.7 Z-test

Z Test, statistics mein istemal hone wala ek tareeqa hai jise hum population ke mean (ausat) ke bare mein hypotheses test karne ke liye use karte hain.

- **Tafseel:** Z Test tab istemal hota hai jab population ka standard deviation (maiyaar-i-inhiraaf) maloom ho aur sample size kafi bada ho (usually 30 se zyada). Is test mein, normal distribution ka istemal hota hai.  

- **Formula:** Z Test ka formula hai: $Z = \frac{\text{sample mean} - \text{population mean}}{\text{standard deviation} / \sqrt{\text{sample size}}}$
- **Application:** Ye test aksar business, psychology, aur medical research mein istemal hota hai jahan large data sets hoti hain. 📊📖

7.4.7.1 Z Test ke Iqsaam (Types of Z Test) 🎲🔍

Z Test ke mukhtalif iqsaam hote hain jo mukhtalif scenarios aur zaruraton ke mutabiq use kiye jate hain.

1. One-Sample Z Test (Yek Namuna Z Imtehaan):

- **Istemaal:** Jab ek sample ke mean ko kisi maloom population mean ke sath compare kiya jata hai.
- **Misal:** Company ye test kar sakti hai ke unka naya product kya average sale time se zyada ya kam time mein bik raha hai.

2. Two-Sample Z Test (Do Namunon ka Z Imtehaan):

- **Istemaal:** Do alag samples ke means ko aapas mein compare karna.
- **Misal:** Do different factories ke production times ko compare karna ke kon si factory zyada efficient hai.

3. Z Test for Proportions (Tanasub ke Liye Z Imtehaan):

- **Istemaal:** Population ke kisi hisse (proportion) ke bare mein hypotheses test karna.
- **Misal:** Kisi election mein aik political party ke votes ke tanasub ka analysis.

7.4.7.2 Ahmiyat (Importance) ✨

Z Test ki significance is mein hai ke ye large samples ke sath precise aur reliable results provide karta hai, khaas kar jahan population parameters maloom hon. Ye test researchers ko enable karta hai ke wo data ke patterns ko samajhne aur informed decisions lene mein madad le sakte hain. 📊💡

Z Test aur uske mukhtalif types, statistics aur data analysis mein widely istemal hote hain. Ye tests various scenarios mein data ko analyze karne ke liye aik powerful tool sabit hote hain, khaas taur par large datasets ke sath.

7.4.7.3 Z Test in Python

Step-1: Import libraries

```
import numpy as np
```

```
from statsmodels.stats import weightstats as stests
```

Step-2: Create a sample data

```
ages = np.array([32, 34, 29, 29, 22, 39, 38, 37, 38, 36, 30, 26, 22, 22, 27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34, 27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34])
```

Step-3: Perform the Z-test

```
z_statistic, p_value = stests.ztest(ages, value=30, alternative='two-sided')
```

Step-4: Print the results

```
print(f'z-statistic: {z_statistic}')
```

```
print(f'p-value: {p_value}')
```

print the results based on if else condition

```
if p_value < 0.05:
```

```
    print("Reject the null hypothesis,\n as the sample mean is significantly different from the population mean.")
```

```
else:
```

```
    print("Fail to reject the null hypothesis,\n as the sample mean is not significantly different from the population mean.")
```

```
z-statistic: -2.2355502631512842
```

```
p-value: 0.025381245676198847
```

```
Reject the null hypothesis,
```

```
as the sample mean is significantly different from the population mean.
```

```
Two Sample Z Test:
```

Step-1: Import libraries

```
import numpy as np
```

```
from statsmodels.stats import weightstats as stests
```

Step-2: Create two sample data with more than 30 samples and known population standard deviation

```
ages1 = np.array([32, 34, 29, 29, 22, 39, 38, 37, 38, 36, 30, 26, 22, 22, 27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34, 27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34])
```

```
ages2 = np.array([27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34, 27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34, 27, 29, 21, 20, 25, 24, 24, 26, 27, 28, 29, 30, 32, 34])
```

Step-3: Perform the Z-test

```
z_statistic, p_value = stests.ztest(ages1, ages2, value=0, alternative='two-sided')
```

Step-4: Print the results

```
print(f'z-statistic: {z_statistic}')
```

```
print(f'p-value: {p_value}')
```

print the results based on if else condition

```
if p_value < 0.05:
```

```
    print("Reject the null hypothesis,\n as the sample means are significantly different.")
```

```
else:
```

```
    print("Fail to reject the null hypothesis,\n as the sample means are not significantly different.")
```













```
z-statistic: 1.4028075294672642
```

```
p-value: 0.1606742411215759
```

Fail to reject the null hypothesis,
as the sample means are not significantly different.

7.4.8 ANOVA (Analysis of Variance)

ANOVA, statistics mein istemal hone wala aik test hai jo multiple groups ke means mein koi significant difference hone ki probability ko test karta hai.

- **Tafseel:** ANOVA, multiple groups ke means mein koi significant difference hone ki probability ko test karta hai. Is test mein, hum dekhte hain ke kya do ya zyada groups ke means mein koi significant difference hai ya nahi.  
- **Types of ANOVA:** ANOVA ke mukhtalif types hote hain jin mein se sab se common hai One-Way ANOVA. Is ke ilawa, Two-Way ANOVA, Three-Way ANOVA, aur MANOVA bhi istemal hote hain.  
- **One-Way ANOVA:** Is test mein, hum dekhte hain ke kya do ya zyada groups ke means mein koi significant difference hai ya nahi.  
- **Two-Way ANOVA:** Is test mein, hum dekhte hain ke kya do ya zyada groups ke means mein koi significant difference hai ya nahi, jab ke wo groups related hain.  
- **N-Way ANOVA:** Is test mein, hum dekhte hain ke kya do ya zyada groups ke means mein koi significant difference hai ya nahi, jab ke wo groups related hain aur un groups mein aik categorical variable aur do continuous variables hain.  
- **MANOVA:** Is test mein, hum dekhte hain ke kya do ya zyada groups ke means mein koi significant difference hai ya nahi, jab ke wo groups related hain aur un groups mein aik categorical variable aur do ya zyada continuous variables hain.  

7.4.8.1 ANOVA ke Assumptions

ANOVA ke assumptions ye hain: 1. **Normality:** Data ka distribution normal hona chahiye. 2. **Independence:** Data points independent hona chahiye. 3. **Homogeneity of Variance:** Data ka variance same hona chahiye. 4. **Randomness:** Data points random hona chahiye.

7.4.8.2 One-Way ANOVA in Python

One-way ANOVA is used to compare two or more groups of samples across one continuous independent variable.

For example, you could use a one-way ANOVA to compare the height of people living in different cities.

```
import scipy.stats as stats
```

```
# Sample data: Growth of plants with three types of fertilizers
```

```
fertilizer1 = [20, 22, 19, 24, 25]
```

```
fertilizer2 = [28, 30, 27, 26, 29]
```

```
fertilizer3 = [18, 20, 22, 19, 24]
```

```
# Perform the one-way ANOVA
```

```
f_stat, p_val = stats.f_oneway(fertilizer1, fertilizer2, fertilizer3)
```

```
print("F-statistic:", f_stat)
```

```
print("p-value:", p_val)
```

```
# print the results based on if the p-value is less than 0.05
```

```
if p_val < 0.05:
```

```
    print(f'Reject null hypothesis: The means are not equal, as the p-value:  
    {p_val} is less than 0.05')
```

```
else:
```

```
    print(f'Accept null hypothesis: The means are equal, as the p-value: {p_val}  
    is greater than 0.05')
```

```
F-statistic: 15.662162162162158
```

```
p-value: 0.0004515404760997283
```

```
Reject null hypothesis: The means are not equal, as the p-value:  
0.0004515404760997283 is less than 0.05
```

One-way ANOVA can also be done using StatsModels.

```
# One-way ANOVA using statsmodels
```

```
import pandas as pd
```

```
import statsmodels.api as sm
```

```
from statsmodels.formula.api import ols
```

```
# Create a dataframe
```

```
# Sample data: Growth of plants with three types of fertilizers
```

```
fertilizer1 = [20, 22, 19, 24, 25]
```

```
fertilizer2 = [28, 30, 27, 26, 29]
```

```
fertilizer3 = [18, 20, 22, 19, 24]
```

```
df = pd.DataFrame({"fertilizer": ["fertilizer1"] * 5 + ["fertilizer2"] * 5 +  
["fertilizer3"] * 5,  
                  "growth": fertilizer1 + fertilizer2 + fertilizer3}))
```

```
# Fit the model
```

```
model = ols("growth ~ fertilizer", data=df).fit()
```

```
# Perform ANOVA and print the summary table
```

```
anova_table = sm.stats.anova_lm(model, typ=2)
```

```
print(anova_table)
```

```
# print the results based on if the p-value is less than 0.05
```

```
if anova_table["PR(>F)"][0] < 0.05:
```

```
    print(f'Reject null hypothesis: The means are not equal, as the p-value is less  
    than 0.05")
```

```
else:
```

```
print(f'Accept null hypothesis: The means are equal, as the p-value is greater than 0.05')
```

```
sum_sq    df      F    PR(>F)
fertilizer 154.533333  2.0 15.662162 0.000452
Residual   59.200000 12.0    NaN    NaN
```

Reject null hypothesis: The means are not equal, as the p-value is less than 0.05

/var/folders/4q/h5d6slgx2rs_drdwcmgf_hm0000gp/T/ipykernel_49854/3008466380.py:23: FutureWarning:

Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent with DataFrame behavior). To access a value by position, use `ser.iloc[pos]`

Based on the p-value, we can conclude that the means are not equal. In other words, the growth of plants is significantly different for the three types of fertilizers. We need to perform a post-hoc test to determine which fertilizers are significantly different from each other.

7.4.8.2.1 Post-Hoc Test for One-Way ANOVA

We will perform a post-hoc test to determine which fertilizers are significantly different from each other.

Post-hoc test for one-way ANOVA

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
```

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

```
# Create a dataframe
```

```
# Sample data: Growth of plants with three types of fertilizers
```

```
fertilizer1 = [20, 22, 19, 24, 25]
```

```
fertilizer2 = [28, 30, 27, 26, 29]
```

```
fertilizer3 = [18, 20, 22, 19, 24]
```

```
df = pd.DataFrame({"fertilizer": ["fertilizer1"] * 5 + ["fertilizer2"] * 5 +  
    ["fertilizer3"] * 5,  
    "growth": fertilizer1 + fertilizer2 + fertilizer3})
```

```
# Perform Tukey's test
```

```
tukey = pairwise_tukeyhsd(endog=df["growth"], groups=df["fertilizer"],  
alpha=0.05)
```

```
# plot the results
```

```
tukey.plot_simultaneous()
```

```
# Print the results
```

```
print(tukey)
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
```

```
=====
```

```
group1    group2  meandiff p-adj  lower  upper reject
```

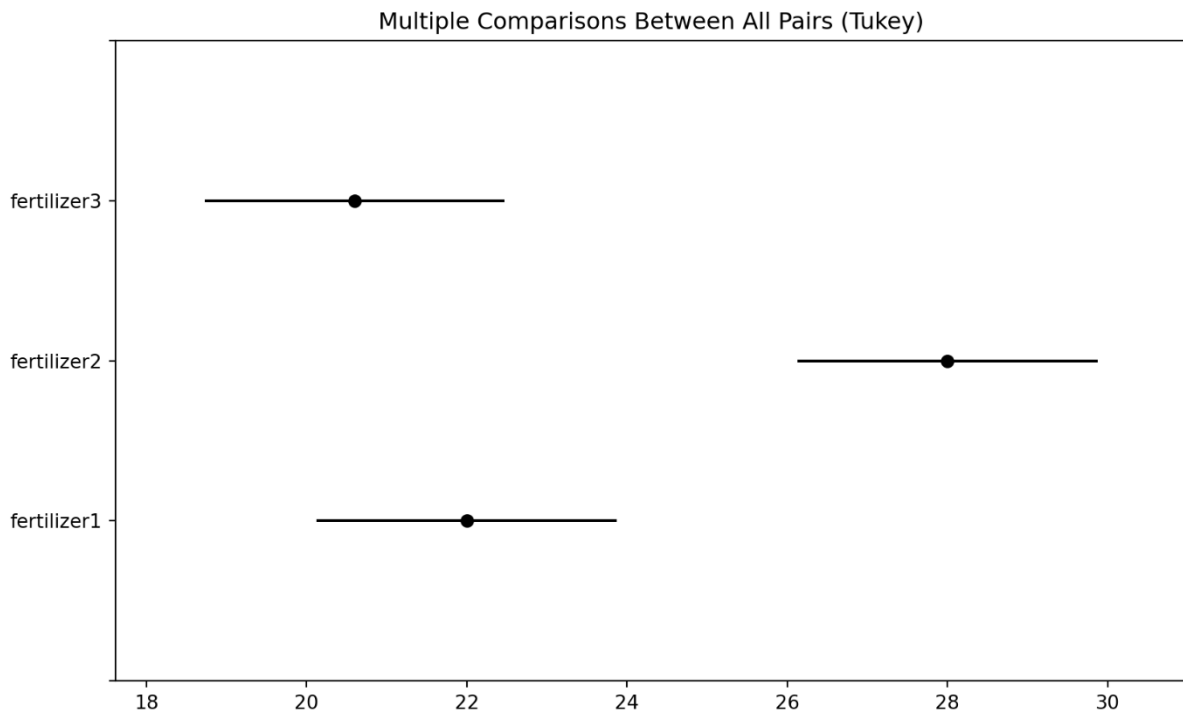
```
-----
```

```
fertilizer1 fertilizer2    6.0 0.0029  2.2523  9.7477  True
```

```
fertilizer1 fertilizer3   -1.4 0.5928 -5.1477  2.3477 False
```

```
fertilizer2 fertilizer3   -7.4 0.0005 -11.1477 -3.6523  True
```

```
-----
```

7.4.8.3 Two-Way ANOVA in Python

Two-way ANOVA is used to compare two or more groups of samples across two continuous independent variables.

For example, you could use a two-way ANOVA to compare the height of people living in different cities and different age groups.

Two-way ANOVA using statsmodels

```
import pandas as pd
```

```
import statsmodels.api as sm
```

```
from statsmodels.formula.api import ols
```

```
# Create a dataframe
```

```
# Sample data: Growth of plants with three types of fertilizers
```

```
fertilizer1 = [20, 22, 19, 24, 25]
```

```
fertilizer2 = [28, 30, 27, 26, 29]
```

```
fertilizer3 = [18, 20, 22, 19, 24]
```

```
df = pd.DataFrame({"fertilizer": ["fertilizer1"] * 5 + ["fertilizer2"] * 5 +
                                ["fertilizer3"] * 5,
                   "growth": fertilizer1 + fertilizer2 + fertilizer3,
                   "age": [20, 22, 19, 24, 25] * 3})

# Fit the model
model = ols("growth ~ fertilizer * age", data=df).fit()

# Perform ANOVA and print the summary table
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)

# print the results based on if the p-value is less than 0.05
if anova_table["PR(>F)"][0] < 0.05:
    print(f'Reject null hypothesis: The means are not equal, as the p-value is less
    than 0.05")
else:
    print(f'Accept null hypothesis: The means are equal, as the p-value is greater
    than 0.05")

sum_sq  df      F    PR(>F)
fertilizer  154.533333  2.0  22.736922  0.000303
age         16.615385  1.0   4.889336  0.054340
fertilizer:age  12.000000  2.0   1.765594  0.225490
Residual    30.584615  9.0     NaN     NaN

Reject null hypothesis: The means are not equal, as the p-value is less than 0.05
/var/folders/4q/h5d6slgx2rs_drdwcmgf_hm0000gp/T/ipykernel_49854/271202
5749.py:25: FutureWarning:
```

Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent with DataFrame behavior). To access a value by position, use `ser.iloc[pos]`

7.4.8.3.1 Post-Hoc Test for Two-Way ANOVA

We will perform a post-hoc test to determine which fertilizers are significantly different from each other.

```
# Post-hoc test for two-way ANOVA
```

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

```
import pandas as pd
```

```
import statsmodels.api as sm
```

```
from statsmodels.formula.api import ols
```

```
# Sample data
```

```
data = pd.DataFrame({  
    "Growth": [20, 22, 19, 24, 25, 28, 30, 27, 26, 29, 18, 20, 22, 19, 24, 21, 23, 20,  
25, 26, 29, 31, 28, 27, 30, 19, 21, 23, 20, 25],  
    "Fertilizer": ["F1", "F1", "F1", "F1", "F1", "F2", "F2", "F2", "F2", "F2", "F3",  
"F3", "F3", "F3", "F1", "F1", "F1", "F1", "F1", "F2", "F2", "F2", "F2",  
"F2", "F3", "F3", "F3", "F3", "F3"],  
    "Sunlight": ["High", "High", "High", "High", "High", "High", "High", "High",  
"High", "High", "High", "High", "High", "High", "Low",  
"Low", "Low", "Low", "Low", "Low", "Low", "Low", "Low", "Low", "Low",  
"Low", "Low", "Low", "Low"]  
})
```

```
tukey = pairwise_tukeyhsd(data['Growth'], data['Fertilizer'] + data['Sunlight'],  
alpha=0.05)
```

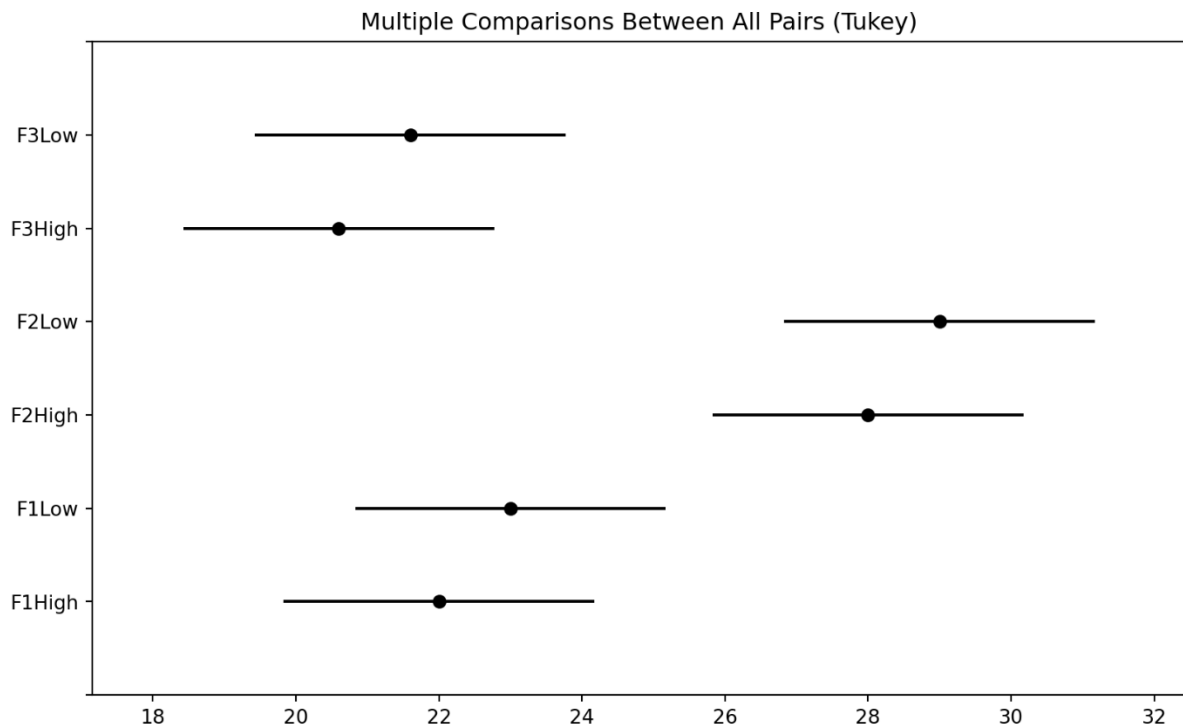
```
# plot the results
```

```
tukey.plot_simultaneous()
```

```
print(tukey)
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
F1High	F1Low	1.0	0.9786	-3.3434	5.3434	False
F1High	F2High	6.0	0.0032	1.6566	10.3434	True
F1High	F2Low	7.0	0.0006	2.6566	11.3434	True
F1High	F3High	-1.4	0.9145	-5.7434	2.9434	False
F1High	F3Low	-0.4	0.9997	-4.7434	3.9434	False
F1Low	F2High	5.0	0.0176	0.6566	9.3434	True
F1Low	F2Low	6.0	0.0032	1.6566	10.3434	True
F1Low	F3High	-2.4	0.5396	-6.7434	1.9434	False
F1Low	F3Low	-1.4	0.9145	-5.7434	2.9434	False
F2High	F2Low	1.0	0.9786	-3.3434	5.3434	False
F2High	F3High	-7.4	0.0003	-11.7434	-3.0566	True
F2High	F3Low	-6.4	0.0016	-10.7434	-2.0566	True
F2Low	F3High	-8.4	0.0	-12.7434	-4.0566	True
F2Low	F3Low	-7.4	0.0003	-11.7434	-3.0566	True
F3High	F3Low	1.0	0.9786	-3.3434	5.3434	False



7.4.8.4 N-Way ANOVA or factorial ANOVA in Python

N Way ANOVA is used to compare N groups of samples across one continuous independent variables. In this example we will choose only 3 groups.

```
import pandas as pd
```

```
import statsmodels.api as sm
```

```
from statsmodels.formula.api import ols
```

```
# Sample data
```

```
data = pd.DataFrame({
```

```
    "Growth": [20, 22, 19, 24, 25, 28, 30, 27, 26, 29, 18, 20, 22, 19, 24,
               21, 23, 20, 25, 26, 29, 31, 28, 27, 30, 19, 21, 23, 20, 25,
               20, 22, 21, 23, 24, 26, 28, 25, 27, 29, 17, 19, 21, 18, 20],
```

```
    "Fertilizer": ["F1", "F1", "F1", "F1", "F1", "F2", "F2", "F2", "F2", "F2", "F3",
                  "F3", "F3", "F3", "F1", "F1", "F1", "F1", "F1", "F2", "F2", "F2", "F2",
                  "F2", "F3", "F3", "F3", "F3", "F3"],
```

```
    "F1", "F1", "F1", "F1", "F1", "F2", "F2", "F2", "F2", "F2",
```

```
    "F3", "F3", "F3", "F3", "F3"],
```

```
"Sunlight": ["High", "High", "High", "High", "High", "High", "High",  
"High", "High", "High", "High", "High", "High", "High", "Low",  
"Low", "Low", "Low", "Low", "Low", "Low", "Low", "Low", "Low", "Low",  
"Low", "Low", "Low", "Low", "High", "High", "High", "High", "High", "High",  
"High", "High", "High", "High",  
"High", "High", "High", "High", "High"],  
"Watering": ["Regular", "Regular", "Regular", "Regular",  
"Regular", "Regular", "Regular", "Regular", "Regular", "Regular",  
"Regular", "Regular", "Regular", "Regular", "Regular", "Sparse",  
"Sparse", "Sparse", "Sparse", "Sparse",  
"Sparse", "Sparse", "Sparse", "Sparse", "Sparse",  
"Sparse", "Sparse", "Sparse", "Sparse", "Sparse",  
"Regular", "Regular", "Regular", "Regular", "Regular",  
"Regular", "Regular", "Regular", "Regular", "Regular",  
"Regular", "Regular", "Regular", "Regular", "Regular"]  
})
```

```
# Fit the model
```

```
model = ols('Growth ~ Fertilizer * Sunlight * Watering', data=data).fit()
```

```
# Perform three-way ANOVA
```

```
anova_results = sm.stats.anova_lm(model, typ=2)
```

```
print(anova_results)
```

```
# print the results based on if the p-value is less than 0.05
```

```
if anova_results["PR(>F)"][0] < 0.05:
```

```
print("Reject null hypothesis: The means are not equal, as the p-value is less
than 0.05")
```

else:

```
print("Fail to reject null hypothesis: The means are equal, as the p-value is
greater than 0.05")
```

	sum_sq	df	F	PR(>F)
Fertilizer	6.023247e+02	2.0	7.466835e+01	4.618789e-14
Sunlight	4.169869e-02	1.0	1.033852e-02	9.195328e-01
Watering	-1.975467e+02	1.0	-4.897852e+01	1.000000e+00
Fertilizer:Sunlight	2.489798e-14	2.0	3.086527e-15	1.000000e+00
Fertilizer:Watering	2.816616e-01	2.0	3.491673e-02	9.657160e-01
Sunlight:Watering	2.054444e+01	1.0	5.093664e+00	2.969139e-02
Fertilizer:Sunlight:Watering	1.088889e+00	2.0	1.349862e-01	8.741344e-01
Residual	1.573000e+02	39.0	NaN	NaN

Reject null hypothesis: The means are not equal, as the p-value is less than 0.05

/Users/aammar/Library/Python/3.9/lib/python/site-packages/statsmodels/base/model.py:1888: ValueWarning:

covariance of constraints does not have full rank. The number of constraints is 2, but rank is 1

/var/folders/4q/h5d6slgx2rs_drdwcmgf_htm0000gp/T/ipykernel_49854/1377226009.py:35: FutureWarning:

Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent with DataFrame behavior). To access a value by position, use `ser.iloc[pos]`

7.4.8.4.1 Post-Hoc Test for N-Way ANOVA

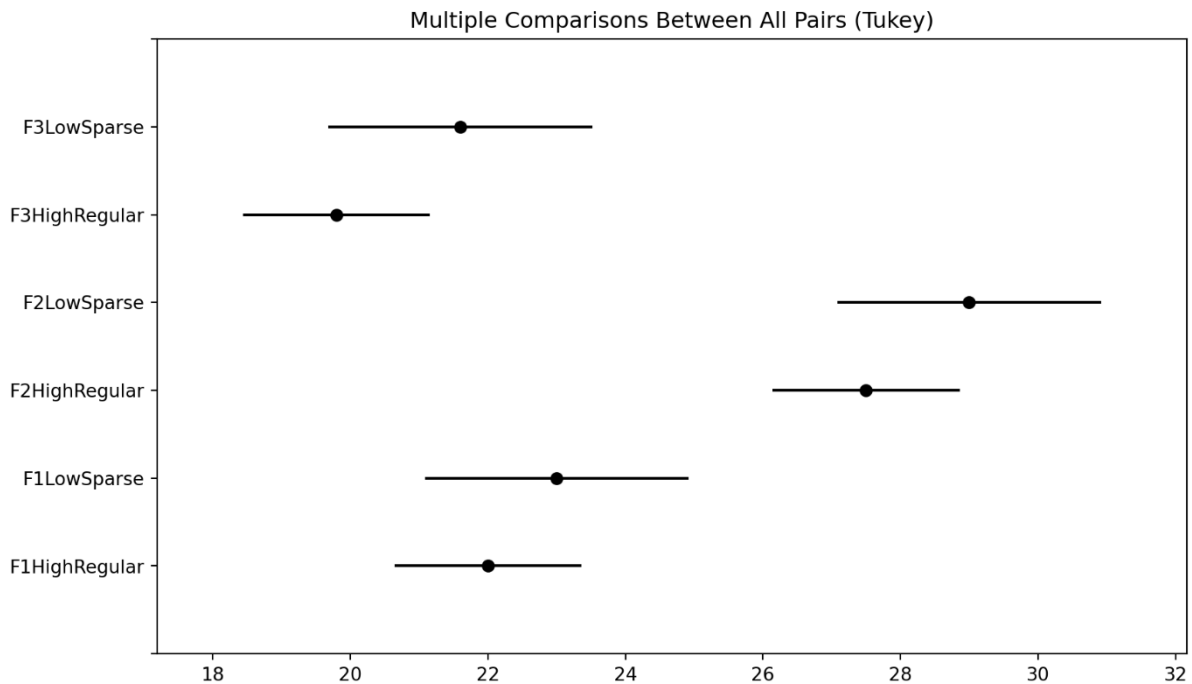
We will perform a post-hoc test to determine which Fertilizer * Sunlight * Watering interactions are significantly different from each other.

Code

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
F1HighRegular	F1LowSparse	1.0	0.9419	-2.2956	4.2956	False
F1HighRegular	F2HighRegular	5.5	0.0	2.8092	8.1908	True
F1HighRegular	F2LowSparse	7.0	0.0	3.7044	10.2956	True
F1HighRegular	F3HighRegular	-2.2	0.1647	-4.8908	0.4908	False
F1HighRegular	F3LowSparse	-0.4	0.9991	-3.6956	2.8956	False
F1LowSparse	F2HighRegular	4.5	0.0027	1.2044	7.7956	True
F1LowSparse	F2LowSparse	6.0	0.0004	2.1946	9.8054	True
F1LowSparse	F3HighRegular	-3.2	0.0613	-6.4956	0.0956	False
F1LowSparse	F3LowSparse	-1.4	0.8775	-5.2054	2.4054	False
F2HighRegular	F2LowSparse	1.5	0.7478	-1.7956	4.7956	False
F2HighRegular	F3HighRegular	-7.7	0.0	-10.3908	-5.0092	True
F2HighRegular	F3LowSparse	-5.9	0.0001	-9.1956	-2.6044	True
F2LowSparse	F3HighRegular	-9.2	0.0	-12.4956	-5.9044	True
F2LowSparse	F3LowSparse	-7.4	0.0	-11.2054	-3.5946	True
F3HighRegular	F3LowSparse	1.8	0.5804	-1.4956	5.0956	False

Pros	Cons	Ideal Usage
<ul style="list-style-type: none"> - Controls Type I error well - Comparatively robust 	<ul style="list-style-type: none"> - Can be conservative - Less powerful for unequal sample sizes 	When equal sample sizes and normally distributed data are assumed.
<ul style="list-style-type: none"> - Simple to compute - Very conservative 	<ul style="list-style-type: none"> - Increases Type II errors - Can be too stringent for many comparisons 	When few comparisons are made; useful in controlling Type I error in multiple testing.
<ul style="list-style-type: none"> - Flexible for any number of comparisons 	<ul style="list-style-type: none"> - More conservative than others - Can be less powerful 	When flexibility in hypothesis testing after ANOVA is needed; good for complex designs.
<ul style="list-style-type: none"> - Suitable for non-parametric data 	<ul style="list-style-type: none"> - Less powerful than parametric tests - Multiple comparison adjustments can be complex 	When data do not meet parametric assumptions, particularly with Kruskal-Wallis test.
<ul style="list-style-type: none"> - Less conservative than Bonferroni - Controls Type I error well 	<ul style="list-style-type: none"> - More complex calculation - Can still be conservative 	When a balance between Type I error control and power is needed, especially with multiple comparisons.
<ul style="list-style-type: none"> - More powerful (higher chance to detect real differences) 	<ul style="list-style-type: none"> - Higher risk of Type I error - Not recommended when there are many comparisons 	When comparisons are planned and limited, often used in exploratory data analysis.
<ul style="list-style-type: none"> - Controls the error rate well - Good for unequal sample sizes 	<ul style="list-style-type: none"> - Complex calculation - Can be conservative 	When there are unequal sample sizes, and control over Type I error is important.
<ul style="list-style-type: none"> - More powerful for detecting differences 	<ul style="list-style-type: none"> - Higher risk of Type I errors than other methods - Not recommended for many comparisons 	When sample sizes are equal and the data are normally distributed; less used due to higher error risks.



7.4.9 Other types of POST-HOC tests

There are several post hoc tests used in statistics, each with its own strengths and weaknesses. Here's a table summarizing some of the most common tests:

7.4.9.0.1 Notes:

1. **Choice of Test:** The choice of post hoc test largely depends on the nature of your data, the number of comparisons, and the balance you want to strike between the risks of Type I and Type II errors.
2. **Data Assumptions:** Some tests assume normally distributed data and equal variances, while others are non-parametric and do not make these assumptions.
3. **Type I and II Errors:** There's often a trade-off between the risk of Type I errors (false positives) and Type II errors (false negatives). More conservative tests (like Bonferroni) reduce the risk of Type I errors but increase the risk of Type II errors.

When conducting post hoc tests, it's essential to understand these pros and cons to choose the most appropriate test for your specific statistical analysis.

7.4.10 MANOVA (Multivariate Analysis of Variance)

Manova is a multivariate extension of ANOVA. It is used to model two or more dependent variables that are continuous with one or more categorical predictor variables. It is often used to assess for differences between two or more groups.

To perform a Multivariate Analysis of Variance (MANOVA) in Python, we typically use the statsmodels library. MANOVA is used when there are two or more dependent variables and one or more independent variables. It tests whether the mean differences among groups on a combination of dependent variables are likely to have occurred by chance.

Here's an example demonstrating how to create a MANOVA table in Python:

7.4.10.1 Example: MANOVA with StatsModels

Let's say we have a dataset with two dependent variables (e.g., test scores in mathematics and science) and one independent variable (e.g., teaching method). We want to know if there are statistically significant differences in the dependent variables across the levels of the independent variable.

7.4.10.2 Explanation:

- **Dataset Preparation:** The data dictionary and DataFrame (df) contain the sample data. Replace this with your actual data.
- **MANOVA Execution:** The MANOVA.from_formula method is used to perform the MANOVA. The formula 'MathScore + ScienceScore ~ Method' indicates that MathScore and ScienceScore are dependent variables, and Method is the independent variable.
- **Results:** The mv_test() method is used to get the MANOVA test results, which are printed to the console.

This script will output the MANOVA table, including Pillai's trace, Wilks' lambda, Hotelling-Lawley trace, and Roy's greatest root test statistics, along with their associated F-values, degrees of freedom, and p-values. These results will help you determine if there are statistically significant differences in the dependent variables across the levels of the independent variable.

```
# Import the required libraries
```

```
import pandas as pd
```

```
from statsmodels.multivariate.manova import MANOVA
```

```
# Create a sample dataset
```

```
data = {  
    'Method': ['A', 'A', 'A', 'B', 'B', 'B', 'C', 'C', 'C'],
```

```
'MathScore': [20, 22, 21, 19, 18, 20, 22, 23, 21],  
'ScienceScore': [30, 28, 29, 33, 32, 31, 29, 27, 28]  
}  
  
df = pd.DataFrame(data)  
  
# Perform the MANOVA  
maov = MANOVA.from_formula('MathScore + ScienceScore ~ Method',  
data=df)  
print(maov.mv_test())
```

Multivariate linear model

=====

Intercept	Value	Num DF	Den DF	F Value	Pr > F
-----------	-------	--------	--------	---------	--------

Wilks' lambda	0.0005	2.0000	5.0000	4711.5000	0.0000
Pillai's trace	0.9995	2.0000	5.0000	4711.5000	0.0000
Hotelling-Lawley trace	1884.6000	2.0000	5.0000	4711.5000	0.0000
Roy's greatest root	1884.6000	2.0000	5.0000	4711.5000	0.0000

Method	Value	Num DF	Den DF	F Value	Pr > F
--------	-------	--------	--------	---------	--------

Wilks' lambda	0.1802	4.0000	10.0000	3.3896	0.0534
Pillai's trace	0.8468	4.0000	12.0000	2.2031	0.1301
Hotelling-Lawley trace	4.4000	4.0000	5.1429	5.4000	0.0444

Roy's greatest root 4.3656 2.0000 6.0000 13.0969 0.0065

7.4.10.3 Interpretation of MANOVA Results

The MANOVA results provided contain two main parts: the test statistics associated with the intercept and the test statistics associated with the independent variable (Method). Each part includes four different test statistics: Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root. Let's interpret these results:

7.4.10.3.1 Intercept Part

1. **Wilks' Lambda:** A value close to 0 (0.0005) with a significant F-value (4711.5) and a p-value of 0.0000 indicates that the model with the intercept is significantly different from a model without the intercept.
2. **Pillai's Trace:** Similar to Wilks' lambda, a value close to 1 (0.9995) with a significant F-value and p-value indicates strong model significance.
3. **Hotelling-Lawley Trace:** A very high value (1884.6) with a significant F-value and p-value also suggests strong model significance.
4. **Roy's Greatest Root:** Like Hotelling-Lawley trace, a high value (1884.6) with a significant F-value and p-value indicates the model's significance.

7.4.10.3.2 Method Part

1. **Wilks' Lambda:** A value of 0.1802 with an F-value of 3.3896 and a p-value of 0.0534. This p-value is marginally above the typical alpha level of 0.05, suggesting that the differences in group means are not quite statistically significant at the 5% level.
2. **Pillai's Trace:** A value of 0.8468, F-value of 2.2031, and a p-value of 0.1301. This result further indicates that the group means are not significantly different, as the p-value is above 0.05.
3. **Hotelling-Lawley Trace:** A value of 4.4 with an F-value of 5.4 and a p-value of 0.0444. This p-value is below 0.05, indicating significant differences in the group means.
4. **Roy's Greatest Root:** A value of 4.3656, with an F-value of 13.0969 and a p-value of 0.0065. This result suggests significant differences in the group means, as indicated by this low p-value.

7.4.10.3.3 Overall Interpretation

- The significant intercept part indicates that the overall model is significant.
- For the Method part, different test statistics provide somewhat conflicting results. Wilks' Lambda and Pillai's Trace suggest that the means of different methods are not significantly different, while Hotelling-Lawley Trace and Roy's Greatest Root suggest significant differences.
- Such discrepancies can occur due to the sensitivity of each test to different assumptions and data characteristics. In practice, when results conflict, it's often advisable to further investigate the data, potentially considering other forms of analysis or looking into specific pairwise comparisons for more insights.

7.4.11 Correlation

Pearson correlation, statistics mein istemal hone wala aik measure hai jo do variables ke darmiyan linear relationship ko measure karta hai.

Types of Correlation tests:

- Pearson's correlation coefficient
- Spearman's rank correlation coefficient
- Kendall's rank correlation coefficient
- Point-Biserial correlation coefficient
- Biserial correlation coefficient
- Phi coefficient
- Cramer's V

7.4.11.1 Pearson's correlation coefficient

Pearson's correlation coefficient is a measure of the linear correlation between two variables X and Y. It has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

7.4.11.2 Spearman's rank correlation coefficient

Spearman's rank correlation coefficient is a nonparametric measure of the monotonicity of the relationship between two datasets. Unlike the Pearson correlation, the Spearman correlation does not assume that both datasets are normally distributed. Like other correlation coefficients, this one varies between +1 and -1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

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

7.4.11.3 Pearson's and Spearman's correlation in Python

There are many ways to calculate Pearson's correlation coefficient in Python. Here's are all the examples, one by one:

1. By defining function

You can unfold the codes by clicking on the arrow on the left side of the code.

Code

Pearson Correlation Coefficient: 0.7745966692414834

Highly Positive Correlation.

Code

Spearman Correlation Coefficient: 0.7378647873726218

Highly Positive Correlation

2. Using numpy

Code

Pearson Correlation Coefficient: 0.7745966692414834

3. Using Pandas on series

Code

Pearson Correlation Coefficient: 0.7745966692414834

4. Using Pandas on dataframe

Code

Pearson Correlation Coefficient:

Statistics for Data Science and Machine Learning

	x	y
x	1.000000	0.774597
y	0.774597	1.000000

=====

Spearman Correlation Coefficient:





	x	y
x	1.000000	0.737865
y	0.737865	1.000000

CHAPTER 8

Probability

Probability in Statistics (Imkanaat in Shumariyat)

Probability, ya Imkanaat, statistics mein ek bunyadi aur ahem concept hai jo ke kisi event ke hone ki likelihood ya chances ko measure karta hai.

- **Tafseel:** Probability ye batati hai ke koi khaas event ya natija kitna mumkin hai. Is ki range 0 se 1 tak hoti hai, jahan 0 ka matlab hai ke event bilkul bhi mumkin nahi aur 1 ka matlab hai ke event yaqeenan hone wala hai.  
- **Ahmiyat:** Probability statistics ke har shobe mein istemal hoti hai aur ye data analysis, predictions, aur decision-making mein madad karta hai. Ye humein uncertain situations mein rational decisions lene mein madad karta hai.  
- **Misal:** Jaise, ek sikke ko uchhalne par heads ya tails aane ka probability 0.5 (ya 50%) hota hai. Ya phir, mausam ki peshgoi karte waqt barish hone ka probability batana.

8.1 Probability ke Types (Iqsaam)

1. **Theoretical Probability (Nazari Imkanaat):** Ye woh probability hoti hai jo theory ya basic principles par mabni hoti hai. Jaise, sikke ki misal mein heads ya tails aane ka theoretical probability hamesha 0.5 hota hai.
2. **Experimental Probability (Tajurbaati Imkanaat):** Ye woh probability hoti hai jo actual experiments ya observations se calculate ki jati hai. Jaise, agar aap 100 dafa sikka uchhalte hain aur 60 dafa heads aata hai, to heads aane ki experimental probability 0.6 ho jayegi.
3. **Subjective Probability (Subjective Imkanaat):** Ye individual judgment ya anubhav par based hoti hai. Jaise, kisi doctor ka ye kehna ke kisi mareez ko kisi bimari hone ki kitni probability hai, based on unke past experiences.

Probability

Probability, statistics ka woh hissa hai jo humein uncertain events aur outcomes ko quantify karne mein madad karta hai. Is ke zariye, hum risk assessment karte hain, predictions karte hain, aur complex data sets ko better samajh sakte hain. Ye statistical analysis ka aham juz hai aur is ka istemal kai different fields mein hota hai, jaise finance, engineering, science, aur social sciences.

Let's explore a real-life example of probability and understand how to calculate it. I'll present this in an engaging and informative style:

8.1.1 Real-Life Example of Probability 🌍🎲

Example: Weather Forecasting (Mausam ki Peshgoi)

Imagine you're a meteorologist trying to predict the probability of rain tomorrow in your city. Based on historical weather data, you know that in your city, out of 365 days, it typically rains on 100 days.

8.1.1.1 Calculating Probability 📊

1. **Identify the Total Number of Outcomes (Mumkin Natijaat ki Tadaad):**
 - In this case, the total number of outcomes is 365 days (a year).
2. **Identify the Number of Favorable Outcomes (Sazgaar Natijaat ki Tadaad):**
 - The number of days it rains is 100.
3. **Use the Probability Formula (Imkanaat ka Formula):**
 - The formula for probability is: $P(\text{Event}) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Outcomes}}$
 - Applying this to our example: $P(\text{Rain}) = \frac{100}{365} \approx 0.274$

8.1.1.2 Interpretation (Tashreeh) 📈

- The probability of rain tomorrow, based on historical data, is approximately 0.274 or 27.4%.
- This means there's a 27.4% chance that it will rain tomorrow.

8.1.2 Real-Life Importance (Asal Zindagi Mein Ahmiyat) ✨

Probability calculations like this are used every day in weather forecasting. By analyzing historical weather data, meteorologists can predict future weather patterns and provide valuable information to the public and various industries.

Similarly, probability is used in many other fields like finance for risk assessment, in healthcare for disease prediction, and in sports for game strategy.

Calculating and understanding probabilities helps in making informed decisions based on statistical evidence rather than guesswork or intuition. This is crucial in fields where decisions can have significant consequences.

8.2 Basic Probability Concepts and Definitions

1. Probability:

- **Definition:** Probability kisi event ke hone ke chances ya likelihood ko measure karta hai.
- **Example:** Sikka uchalne par 'heads' aane ki Probability 0.5 hai.
- **Formula:** Probability ka formula $P(\text{Event}) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Outcomes}}$

2. Experiment (Tajurba):

- **Definition:** Koi bhi process ya activity jise perform kiya ja sakta hai aur jiska measurable outcome ho.
- **Example:** Sikka uchalna ek tajurba hai jiska outcome 'heads' ya 'tails' ho sakta hai.

3. Outcome (Natija):

- **Definition:** Kisi tajurbe ke natije mein se har ek mumkin result.
- **Example:** Sikka uchalne mein 'heads' ya 'tails' aana ek natija hai.

4. Sample Space:

- **Definition:** Kisi tajurbe ke sab mumkin natijaat ka majmua.
- **Example:** Sikka uchalne mein sample space hai {'heads', 'tails'}.

5. Event (Waqea):

- **Definition:** Sample space mein se kisi specific group of outcomes ka selection.
- **Example:** Ek card draw karne mein, 'king' card nikalna ek event hai.

6. Random Variable:

- **Definition:** Kisi tajurbe ke natijat ko numeric values se represent karta hai.
- **Example:** Sikka uchalne par, assign karen '1' for 'heads' aur '0' for 'tails'.

7. Independent Events (Azad Waqeaat):

- **Definition:** Jab ek event ke natije dusre event par asar nahi daalte.
- **Example:** Alag alag sikke uchalne ke natije ek dusre par asar nahi daalte.

8. Dependent Events (Mutassir Waqeaat):

- **Definition:** Jab ek event ka natija dusre event par asar daalta hai.
- **Example:** Agar ek bag se bina wapass daale hue ek card nikala jaye, to dusra card nikalne ke natije pehle card par depend karte hain.

9. Mutually Exclusive Events:

- **Definition:** Aise events jo ek sath nahi ho sakte.
- **Example:** Ek toss mein ek waqt mein 'heads' aur 'tails' dono nahi aa sakte.

10. Collectively Exhaustive Events:

- **Definition:** Aise events jo ek sath ho sakte hain aur jin ka union sample space ke barabar ho.
- **Example:** Ek toss mein ek waqt mein 'heads' aur 'tails' dono aa sakte hain.

11. Complement of an Event:

- **Definition:** Ek event ke sath us ke bilkul opposite event.
- **Example:** Ek card draw karne mein, 'king' card nikalne ka complement 'not king' card hai.
- **Formula:** Complement ka formula $P(\text{Event}) + P(\text{Complement of Event}) = 1$

12. Trial (Aazmaish):

- **Definition:** Imkanaat mein, aazmaish se murad woh individual execution ya performance hota hai jahan ek ya zyada outcomes mumkin hote hain.
- **Example:** Ek dice ko roll karna ek aazmaish hai, jahan outcomes 1 se 6 tak ho sakte hain.

13. Venn Diagram (Venn Diagram):

- **Definition:** Ye ek graphic organizer hai jo sets aur unke relationships ko visually show karta hai.
- **Example:** Do circles jo ek dusre ko overlap karte hain, un sets ke common elements ko show karte hain.

Code

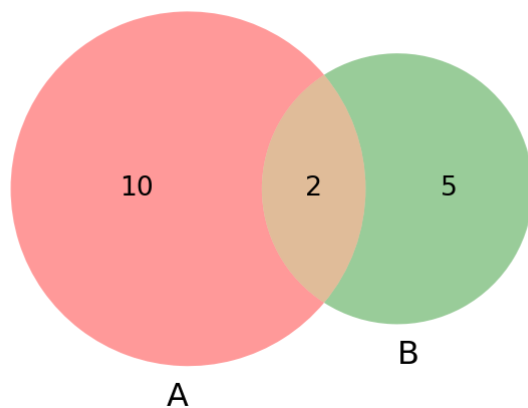


Figure 8.1: Ven Diagram

14. Union (Ittihad):

- **Definition:** Do ya zyada sets ka union woh set hota hai jo un sabhi sets ke elements ko include karta hai.
- **Example:** Agar Set A = {1, 2, 3} aur Set B = {3, 4, 5}, to A Union B = {1, 2, 3, 4, 5}.

15. Intersection (Goondh):

- **Definition:** Do ya zyada sets ka intersection woh set hota hai jo un sabhi sets ke common elements ko include karta hai.
- **Example:** Agar Set A = {1, 2, 3} aur Set B = {3, 4, 5}, to A Intersection B = {3}.

Code

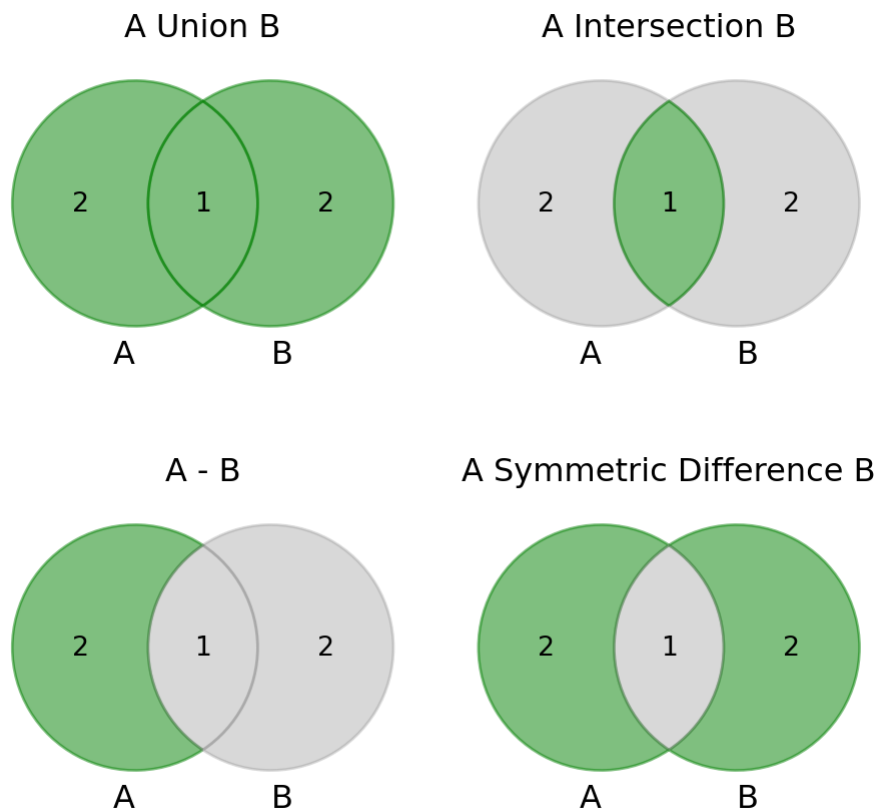


Figure 8.2: Set operations on Venn Diagram

16. Complement (Takmeel):

- **Definition:** Ek set ka complement woh set hota hai jo universal set ke elements ko include karta hai lekin pehle set ke elements ko nahi include karta.
- **Example:** Agar Set $A = \{1, 2\}$ aur universal set $= \{1, 2, 3, 4\}$, to A ka complement $= \{3, 4\}$.

Code

A Compliment in Universal Set U

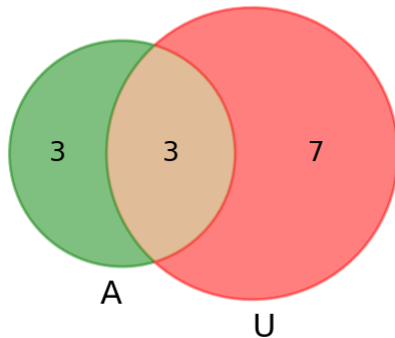


Figure 8.3: Set complement on Venn Diagram

17. Permutations (Tarteelat):

- **Definition:** Permutations se murad woh tamam mumkin ways hoti hain jin mein kisi set ke elements ko tarteeb di ja sakti hai.
- **Example:** Agar 3 alag numbers ko 2 ki sequence mein arrange karna hai, to $3P2 = 6$ ways hote hain.

18. Combinations (Imtizajat):

- **Definition:** Combinations woh tamam mumkin ways hoti hain jin mein kisi set ke elements ko group kiya ja sakta hai, lekin yahan tarteeb matter nahi karti.
- **Formula:** Combinations ka formula $nCr = \frac{n!}{r!(n-r)!}$
- Where n = total number of elements in the set, r = number of elements in the group and $!$ = factorial.
- **Example:** Agar 3 alag numbers mein se kisi 2 ko select karna hai, to $3C2 = 3$ ways hote hain.

8.3 Probability

There are several technical ways to define probability, but a definition useful for statistics is that **probability tells us how often something is likely to occur when an experiment is repeated.**

Probability humein batati hai ke koi cheez kitni baar ho sakti hai jab koi experiment repeat kiya jata hai. Masalan, ye kehna ke sikke ka toss karne par 'heads' aane ka probability kitna hai, isey hum sikke ko bohot baar uchaal kar dekh sakte hain aur note karte hain ke kitni dafa 'heads' aaya hai. Probability ke baare mein shayad sab se important baat yeh hai: Kisi bhi event ka probability

hamesha 0 aur 1 ke darmiyan hota hai. Agar kisi event ka probability 0 hai, toh iska matlab hai ke uska hone ka koi chance nahi hai, jabke agar kisi event ka probability 1 hai, toh yeh zaroor hone wala hai. Mathematics mein probability ko decimals mein specify karna conventional hai, isliye hum kehte hain ke kisi event ka probability 0 aur 1 ke beech hota hai, lekin percentages mein baat karna bhi acceptable hai (aur rozmarra ki baat-cheet mein zyada common hai), toh yeh bhi bilkul theek hai ke kisi event ka probability hamesha 0% aur 100% ke beech hota hai. Decimals se percent mein jaane ke liye, 100 se multiply karen (per cent = har 100 mein se), toh 0.4 probability bhi 40% probability hai ($0.4 \times 100 = 40$), aur 0.85 probability ko bhi 85% probability ke taur par keh sakte hain. Negative probability aur 100% se zyada ki probabilities logical impossibilities hain jo sirf baat-cheet ke taur par exist karti hain. Probability ka 0 aur 1 ke beech bounded hona mathematical implications rakhta hai jo hum is chapter mein baad mein explore karenge.

Tip

This fact also provides a useful check on your calculations. If you come up with a probability lower than 0 or greater than 1, you have certainly made a mistake somewhere along the way. Furthermore, if someone tells you there is a 200% chance that you will make a killing in the stock market if you follow his system, you should probably look for a new investment advisor.

Important Facts about Probability

1. The probability of an event is always between 0 and 1.
2. The probability of the sample space is always 1.
3. The probability of an event and its complement is always 1.
4. The probability of an event that cannot occur is 0.
5. The probability of an event that must occur is 1.

Probability Formula

Probability ka formula ye hai:

$$P(\text{Event}) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Outcomes}}$$

8.3.1 Expressing the Probability of an Event

Probability ko express karne ke liye, hum is ka formula use karte hain. Is formula mein, humein total number of outcomes aur number of favorable outcomes ko calculate karna hota hai. Total number of outcomes, ya sample

space, humein experiment ko perform karne se pehle hi pata hota hai. Lekin number of favorable outcomes ko calculate karne ke liye, humein experiment ko perform karna hota hai. Is liye, hum probability ko calculate karne ke liye experiment ko perform karte hain. Is ke baad, hum probability ko express karne ke liye is ka formula use karte hain.

8.3.1.1 Example: Probability of Rolling a Die

Question: Ek fair die ko roll karne par 3 aane ka probability kya hai?

Solution:

1. **Identify the Total Number of Outcomes (Mumkin Natijaat ki Tadaad):**
 - In this case, the total number of outcomes is 6 (1, 2, 3, 4, 5, 6).
 - Is ko hum sample space kehte hain.
 - Is ke baad, humein number of favorable outcomes ko calculate karna hota hai.
2. **Identify the Number of Favorable Outcomes (Sazgaar Natijaat ki Tadaad):**
3. **Use the Probability Formula (Imkanaat ka Formula):**
 - The formula for probability is: $P(\text{Event}) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Outcomes}}$
 - Applying this to our example: $P(\text{Rolling a 3}) = \frac{1}{6} \approx 0.167$

8.3.1.2 Example: Probability of Drawing a Card

Question: Ek standard deck of cards mein se ek card draw karne par 'king' card aane ka probability kya hai?

Solution:

1. **Identify the Total Number of Outcomes (Mumkin Natijaat ki Tadaad):**
 - In this case, the total number of outcomes is 52 (4 suits \times 13 ranks).
 - Is ko hum sample space kehte hain.
 - Is ke baad, humein number of favorable outcomes ko calculate karna hota hai.

2. **Identify the Number of Favorable Outcomes (Sazgaar Natijaat ki Tadaad):**

3. **Use the Probability Formula (Imkanaat ka Formula):**

- The formula for probability is: $P(\text{Event}) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Outcomes}}$
- Applying this to our example: $P(\text{Drawing a King}) = \frac{4}{52} \approx 0.077$

8.3.2 Conditional Probability

Conditional Probability, wo probability hoti hai jisme ek event ke hone ki likelihood ko dusre event ke hone ya na hone ke context mein dekha jata hai.

- **Tafseel:** Yeh basically batata hai ke agar humein pata ho ke koi specific condition ya event pehle hi occur ho chuka hai, to is ke baad kisi dusre event ke hone ki kya probability hogi.
- **Formula:** $P(A|B) = \frac{P(A \cap B)}{P(B)}$ Yahan, $P(A|B)$ ka matlab hai probability of A given B.
- **Misal:** Maan lijiye, ek bag mein 5 red aur 5 blue balls hain. Agar ek ball randomly nikali jati hai aur wo red hoti hai (Event B), to dobara red ball nikalne ki Imkanaat (Event A) kya hogi?

8.3.2.1 Conditional Probability ka Ahmiyat 🌟

1. **Complex Situations Mein Samajh:** Conditional probability complex situations ko samajhne mein madad karti hai, jahan ek event dusre se kis tarah linked hota hai.
2. **Decision Making:** Ye concept especially decision-making mein useful hota hai, jaise medical diagnosis ya financial forecasting mein. 🏠💰
3. **Statistics aur Data Analysis:** Conditional probability statistics aur data analysis ke advanced concepts jaise Bayes' Theorem mein istemal hoti hai.

8.3.2.2 Real-Life Application (Asal Zindagi Mein Istemal) 🌐💡

Conditional probability ka istemal rozmarra ki zindagi ke kayi decisions mein hota hai. Jaise, doctors ye jan'ne ke liye istemal karte hain ke agar koi patient kisi symptom ko show karta hai, to usay koi specific bimari hone ki kitni

imkanaat hai. Isi tarah, business analysts market trends aur consumer behavior ko samajhne ke liye conditional probability ka istemal karte hain.

For example, we have a bag with 5 red and 5 blue balls, and we want to find the probability of drawing a red ball again (Event A) given that a red ball has already been drawn (Event B), the calculation is as follows:

After drawing the first red ball, there are now 4 red balls left in a total of 9 balls. So, the conditional probability is calculated as:

$$P(A \text{ given } B) = \frac{4}{9} \approx 0.444$$

This means there's approximately a 44.4% chance of drawing another red ball given that one red ball has already been drawn. Jaisay jaisay zada red baal nikaltay ayen gay probability kam hoti jayegi, kyun ke red balls ki tadaad kam hoti jayegi.

8.3.3 Calculating the probability of multiple events

Multiple events ka probability calculate karne ke liye, hamen pata hna chahyeay k event independent hain ya dependent.

- **Independent Events:** Independent events woh events hoti hain jo ek dusre par asar nahi daalte. Jaise, ek coin ko uchhalne par 'heads' ya 'tails' aane ka probability 0.5 hai. Agar hum ek coin ko 10 dafa uchhalte hain aur 5 dafa 'heads' aur 5 dafa 'tails' aata hai, to ye independent events hain.
- **Dependent Events:** Dependent events woh events hoti hain jo ek dusre par asar daalte hain. Jaise, agar hum ek bag mein jis mein 5 red and 5 blue balls hain se ek ball nikalte hain aur wo red hoti hai, to dobara se ball nikalne par red ball aane ka probability kam ho jata hai. Kyun ke ab bag mein red balls ki tadaad kam ho jati hai. Is liye, ye dependent events hain.

8.3.3.1 1. Union of Mutually Exclusive Events

Example: Tossing a Coin - Events: Event A = Getting heads, Event B = Getting tails - **Mutually Exclusive:** These events are mutually exclusive because you can't get both heads and tails on a single coin toss. - **Union:** $P(A \text{ or } B) = P(A) + P(B)$. Since the probability of heads ($P(A)$) and tails ($P(B)$) is each 0.5, the union is $0.5 + 0.5 = 1$. This means either heads or tails is certain to occur.

8.3.3.2 2. Union of Events That Are Not Mutually Exclusive

Example: Drawing a Card from a Deck - Events: Event A = Drawing a heart, Event B = Drawing a queen - **Not Mutually Exclusive:** These events are not mutually exclusive as one card (the Queen of Hearts) satisfies both events.

- **Union:** $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. The probability of drawing a heart ($P(A)$) is $1/4$, a queen ($P(B)$) is $1/13$, and the Queen of Hearts ($P(A \text{ and } B)$) is $1/52$. So, $P(A \text{ or } B) = 1/4 + 1/13 - 1/52$.

8.3.3.3 3. Intersection of Independent Events

Example: Rolling Two Dice - Events: Event A = First die shows a 6, Event B = Second die shows a 6 - **Independent Events:** The result of the first die doesn't affect the second die. - **Intersection:** $P(A \text{ and } B) = P(A) \times P(B)$. The probability of each die showing a 6 is $1/6$, so $P(A \text{ and } B) = 1/6 \times 1/6$.

8.3.3.4 4. Intersection of Nonindependent Events

Example: Drawing Two Cards Successively Without

Replacement - Events: Event A = First card is an ace, Event B = Second card is an ace - **Nonindependent Events:** The outcome of the first draw influences the second. - **Intersection:** For $P(A \text{ and } B)$, first calculate $P(A)$, which is $4/52$. Then, given A has occurred, there are now 3 aces left in a deck of 51 cards, so $P(B \text{ given } A) = 3/51$. Thus, $P(A \text{ and } B) = P(A) \times P(B \text{ given } A) = 4/52 \times 3/51$.

8.4 Bayes' Theorem

Bayes' Theorem, ya Bayes Ka Usool, probability theory mein ek powerful formula hai jo conditional probabilities (shartia imkaaniyat) ko calculate karta hai.

Bayes' Theorem was named after Thomas Bayes, an English statistician who lived in the 18th century. It is also known as Bayes' Rule, Bayes' Law, or Bayes' Formula.

- **Tafseel:** Ye theorem humein yeh samajhne mein madad karta hai ke agar humein kisi event ke hone ki prior knowledge ho, to kaise hum us information ko use kar ke kisi dusre related event ki updated probability ko calculate kar sakte hain.
- **Formula:** $P(A|B) = P(B|A) \times P(A) / P(B)$ Yahan, $P(A|B)$ ka matlab hai probability of A given B.
- **Components:**
 - **$P(A|B)$:** Event A ke hone ki probability given ke event B ho chuka hai.

- **$P(B|A)$** : Event B ke hone ki probability given ke event A ho chuka hai.
- **$P(A)$ aur $P(B)$** : Event A aur B ke hone ki independent probabilities.

8.4.1 Bayes' Theorem Ki Ahmiyat 🌟

1. **Decision Making**: Bayes' Theorem ko decision making, especially under uncertainty mein use kiya jata hai. Ye especially medical field, finance, aur machine learning mein bohot useful hai.
2. **Updating Beliefs**: Ye formula humein nayi information milne par apne beliefs ko update karne mein madad karta hai, jaise diagnosis mein new test results ki basis par disease ki likelihood ko update karna. 🏠📈
3. **Predictive Modelling**: Machine learning aur data science mein, Bayes' Theorem ko predictive models banane aur unhe fine-tune karne ke liye istemal kiya jata hai.

8.4.2 Real-Life Application (Asal Zindagi Mein Istemal) 🧑🔬

Example: Medical Diagnosis - Maan lijiye, aap ek doctor hain aur aapko ye maloom hai ke kisi bimari ka overall prevalence ($P(B)$) bohot kam hai, lekin aapke patient mein kuch specific symptoms (A) hain. Bayes' Theorem ki madad se, aap calculate kar sakte hain ke given yeh symptoms, is patient ko woh bimari hone ki kitni probability hai ($P(A|B)$).

Bayes' Theorem

Bayes' Theorem, probability ke theory mein aik jadeed aur mufeed tool hai. Ye na sirf humein complex data sets ko samajhne mein madad karta hai, balki uncertain situations mein informed decisions lene mein bhi hamari rehnumai karta hai. Ye approach humein prior knowledge aur new evidence dono ko effectively combine karne ki taqat deta hai.

8.5 Sensitivity (Hassasiyat) 🔍

Sensitivity, ya hassasiyat, medical testing mein istemal hone wala ek metric hai jo ye measure karta hai ke kisi test ka disease ko detect karne ki kitni capability hai, especially jab disease present ho.

- **Formula:** $\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$
- **Real-Life Example:** Maan lijiye, aap ek medical researcher hain aur aap ek naye cancer detection test ka study kar rahe hain. Agar aapka test 95%

cancerous cases ko sahi taur par identify karta hai, to iska sensitivity 95% hai. Iska matlab hai ke ye test cancer patients ko identify karne mein high degree of accuracy rakhta hai.

8.5.1 Specificity (Tafseelat)

Specificity, ya tafseelat, ek aur important metric hai jo measure karta hai ke kisi test ka disease na hone par kitni accurately negative result dena ka capability hai.

- **Formula:** $\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$
- **Real-Life Example:** Agar wahi cancer test, non-cancerous cases ko 90% cases mein sahi taur par identify kar raha hai ke un mein cancer nahi hai, to iska specificity 90% hai. Iska matlab hai ke is test ka galat taur par cancer ka indication dene ka imkaan kam hai.

8.5.2 Probability of Disease in the Population (Aabadi Mein Bimari ki Imkaaniyat)

Ye metric population level par kisi specific bimari ke hone ki overall probability ko represent karta hai.

- **Importance:** Ye understanding epidemiologists aur public health officials ko help karti hai disease patterns ko samajhne aur health resources ko allocate karne mein.
- **Real-Life Example:** Maan lijiye, aapke shehar mein aik lakh log rehte hain, aur us mein se 1,000 logon ko diabetes hai. To, diabetes ki prevalence ya aabadi mein bimari ki imkaaniyat 1% hai (1,000 divided by 100,000). Ye data health planning aur resource allocation decisions ke liye crucial hai.