NPHYD

What is stats
Central tendency measure
Dispersion measure
Data distribution

Design of experiments
Sampling
Descriptive statistics
    Location of central tendency
        Mean
        Median
        Mode
    Dispersion of the data
        Range
        variance /standard deviation(smaller standard deviation, large standard deviation)
        Kurtosis
            Leptokurtic
            Mesokurtic
            Platykurtic
        Mean absolute deviation
        Interquartile range,quartile,percentile, decile
        sample stats vs population stats

Probability Theory

Shape or Distribution of the data

Measure of shape(symmetric or skewed)

Locating outlier with z score

The Empirical Rule

Types of distribution(Probability Distribution)

Probability Mass Distribution

Discrete Probability Distribution

Uniform

bernoulli

binomial distribution

Negative binomial

possion

Probability Density Distribution

Continuous Probability Distribution

Uniform

Normal

Standard normal distribution

Lognormal

Gamma

Chi square

T distribution

F distribution

Exponential

Power law

Pareto

Triangular

Geometric

Hypergeometric

Weibull

Cumulative probability distribution

Population/Sample

Methods or estimation of sample size

- Probability sampling
- Types of random sampling
    - Simple random sample
    - Stratified sampling
    - Systematic sampling
    - Cluster sampling

- Non probability sampling
    - Accidental sampling
    - Quota sampling
    - Judgmental sampling
    - Snowball sampling

- Errors in data collection
- Sampling error
- Non sampling error

Central limit theorem

Inferential Statistics

Point estimate
Confidence interval estimate
Bias
Efficiency
Standard error
Significance level(alpha)
Beta value
One tail test
Two tail test
Acceptance region

Rejection region
Critical value
Degree of freedom
P-value
Type 1 error
Type two error
Confusion matrix with alternative hypothesis and null hypothesis
Evidence collection
Z-table, t-table, chi-square-table, f-table

Hypothesis testing
Null hypothesis
Alternative hypothesis
Z-TEST
T-TEST(one sample t test, two sample t test, paired test, analysis of variance(ANOVA))
CHI-SQUARE TEST
F-TEST
ANOVA-TEST(one way, two way ….n-way anova)

Test for continuous data
Mean test
Variance test
    Homogeneity of variance(HOV)
Non parametric test
Parametric test
test
Data normality check
    Plot graph
    Shapiro wilk w test
    Anderson test
    Kolmogorov smirnov (KS) test
Non normal data
Reason for non normal data
Extreme values
Overlap of two or more process
Insufficient data discrimination

Sorted data
Values close to zero or a natural limit
Data follows a different distribution

Correlation
  - **Spearman rank**
  - **Pearson**
  - **kendall**

Covariance
==

Ineuron

Simple and composite hypothesis
Critical value and critical region
Hypothesis testing for large sample size
Compare two population sample means using z test
Hypothesis testing for small sample size

Testing of Hypothesis for population Variance Using Chi-Squared test
chi-square test for categorical variable
Chi-square test for the goodness of fit
Anova

A/B testing
Types of the data
Level of measurement
Mean median mode for the group data
Skewness effect and use of central tendency
Coefficient of variation(CV)

Discrete probability distribution

Binomial
Poisson
Bernoulli

Continuous probability distribution

PDF
PMS
CDF

Normal
Student T
Chi squared

Random variable
DISCRETE RANDOM VARIABLE
CONTINUES RANDOM VARIABLE

PARAMETRIC TEST

T-test
Z-test
F-test
Bayes theorem
Conditional probability
One way anova
Two way anova
K-way or n-way anova

Calculating using Python

1) The sum of squares between (SSbetween)

2) The sum of squares within (SSwithin)

3) The sum of squares total (SSTotal)

F-test(variance ratio test)

—---

Data distribution then relate this thing with prob distribution
Left and right skewed
Types of probability distributions

PMF
Uniform
Binomial and Multinomial
Possion

PDF
Normal
Log normal
Power Law
Pareto

CDF concept also tell them

Many things closely follow a Normal Distribution:

- heights of people
- size of things produced by machines
- errors in measurements
- blood pressure

- marks on a test

Plot names based on univariate bivariate and multivariate analysis

Common chart for the EDA.

1. Area Chart
2. Boxen Plot
3. Bubble Chart
4. Candlestick Chart
5. Donut Chart
6. Error Bar Chart
7. Funnel Chart
8. Hexbin Plot
9. Kde Plot (Kernel Density Estimate)
10. Lollipop Chart
11. Marimekko Chart
12. Network Graph
13. Parallel Coordinates Plot
14. Pareto Chart
15. Radial Plot
16. Sankey Diagram
17. Step Chart
18. Stacked Bar Chart
19. Sunburst Chart
20. Tree Map
21. Venn Diagram
22. Waterfall Chart
23. Waffle Chart
24. Line Plot
25. Scatter Plot
26. Histogram
27. Bar Plot
28. Bar Chart
29. Stacked Bar Chart
30. Grouped Bar Chart
31. Box Plot
32. Violin Plot
33. Kernel Density Plot
34. Area Plot
35. Stacked Area Plot

36. Stream Plot
37. Density Plot
38. Hexbin Plot
39. Pair Plot
40. Joint Plot
41. Heatmap
42. Clustermap
43. Correlation Plot
44. Network Graph
45. Tree Plot
46. Radial Plot
47. Circular Plot
48. Sankey Diagram
49. Treemap
50. Pie Chart
51. Donut Chart
52. Polar Plot
53. Parallel Coordinates Plot
54. Parallel Categories Plot
55. 3D Scatter Plot
56. 3D Line Plot
57. 3D Surface Plot
58. Contour Plot
59. Contourf Plot
60. Streamline Plot
61. Quiver Plot
62. Candlestick Plot
63. Waterfall Chart
64. Gantt Chart
65. Funnel Plot
66. Lollipop Plot
67. Waffle Chart
68. Word Cloud
69. Horizon Plot
70. Ridgeline Plot
71. Swarm Plot
72. Strip Plot
73. Categorical Dot Plot
74. Categorical Aberration Plot
75. Categorical Violin Plot
76. Categorical Box Plot
77. Categorical Point Plot

78. Categorical Bar Plot
79. Categorical Count Plot
80. Categorical Percentage Plot
81. Categorical Swarm Plot
82. Categorical Strip Plot
83. Categorical Stacked Bar Plot
84. Categorical Grouped Bar Plot
85. Categorical Heatmap
86. Categorical Clustermap
87. Categorical Correlation Plot
88. Candlestick Bar Chart
89. Errorbar Plot
90. Stem Plot
91. Step Plot
92. Filled Step Plot
93. Fan Chart
94. Polar Area Chart
95. Polar Bar Chart
96. Word Frequency Plot
97. Network Diagram
98. Chord Diagram
99. Alluvial Diagram
100. Spider Plot
101. Population Pyramid
102. Streamgraph
103. Kiviat Diagram
104. Marimekko Chart
105. Area Range Plot
106. Calendar Plot
107. Calendar Heatmap
108. Circular Heatmap
109. Violin Swarm Plot
110. Strip Box Plot
111. Marginal Histogram
112. Marginal Box Plot
113. Marginal Violin Plot
114. Marginal KDE Plot
115. Pairwise Density Plot
116. QQ Plot (Quantile-Quantile Plot)
117. ECDF Plot (Empirical Cumulative Distribution Function)
118. Rug Plot
119. Circular Bar Plot

120. Polar Scatter Plot
121. Ternary Plot
122. 3D Wireframe Plot
123. 3D Contour Plot
124. 3D Ribbon Plot
125. Radar Spider Plot
126. Packed Bubble Chart
127. Slope Chart
128. Circular Tree Map
129. Circular Dendrogram
130. Word Tree
131. Network Heatmap
132. Arc Diagram
133. Radial Stacked Bar Chart
134. Polar Histogram
135. Word Scatter Plot
136. Hive Plot
137. Rose Diagram
138. Choropleth Map
139. Cartogram
140. Hex Map
141. Streamgraph
142. Circular Network Diagram
143. Flowchart
144. Circos Plot
145. Radial Box Plot
146. Streamtube Plot
147. Joy Plot
148. Word Cloud Matrix
149. Circular Barbell Plot
150. Fan Chart
151. Word Arc Diagram
152. Spiral Plot
153. Spiral Heatmap
154. Spiral Line Plot
155. Spiral Scatter Plot
156. Circular Bubble Plot
157. Circle Packing
158. Hierarchical Edge Bundling
159. Alluvial Diagram
160. Sankey Diagram with Flows
161. Sunburst Plot

162. Area Tree Map
163. Mosaic Plot
164. Word Pyramid
165. Radial Line Chart
166. Chord Diagram with Groups
167. Circle Timeline
168. Circumplex Chart
169. Radial Heatmap
170. Streamgraph with Stream Ribbons
171. Stepped Area Chart
172. Parallel Sets Plot
173. Spiral Calendar Plot
174. Polar Wordcloud
175. Circular Heatmap with Connections
176. Donut Heatmap
177. Polar Scatter Plot with Connections
178. Streamgraph with Stream Bands
179. Ternary Heatmap
180. Chord Diagram with Arcs
181. Circular Voronoi Diagram
182. Sunflower Plot
183. Circular Word Cloud

Univariate Analysis:

1. Histogram
2. Kernel Density Plot
3. Area Plot
4. Density Plot
5. Bar Plot
6. Box Plot
7. Violin Plot
8. Swarm Plot
9. Strip Plot
10. Ridgeline Plot
11. Line Plot
12. Scatter Plot
13. Stem Plot
14. Step Plot
15. Filled Step Plot

16. Fan Chart
17. Polar Area Chart
18. Polar Bar Chart
19. Word Frequency Plot
20. Rug Plot

Bivariate Analysis:

1. Scatter Plot
2. Line Plot
3. Bar Plot
4. Box Plot
5. Violin Plot
6. Joint Plot
7. Pair Plot
8. Hexbin Plot
9. Heatmap
10. Clustermap
11. Correlation Plot
12. Network Graph
13. Streamplot
14. Quiver Plot
15. Contour Plot
16. Contourf Plot
17. Candlestick Plot
18. Waterfall Chart
19. Gantt Chart
20. Funnel Plot
21. Lollipop Plot
22. Waffle Chart

Multivariate Analysis:

1. Clustermap
2. Correlation Plot
3. Network Graph
4. Parallel Coordinates Plot
5. Parallel Categories Plot
6. 3D Scatter Plot
7. 3D Line Plot
8. 3D Surface Plot
9. Scatter Plot Matrix

10. Ternary Plot
11. Radial Plot
12. Circular Plot
13. Sankey Diagram
14. Treemap
15. Word Cloud
16. Alluvial Diagram
17. Chord Diagram
18. Radial Stacked Bar Chart
19. Hive Plot
20. Word Scatter Plot

can you give me all the graph names which we use in the EDA

Certainly! Here are some common types of graphs used in Exploratory Data Analysis (EDA):

1. Histogram: A histogram displays the distribution of a continuous variable by dividing it into bins and showing the frequency or count of data points in each bin.
2. Bar Chart: A bar chart represents categorical data using rectangular bars, where the length of each bar corresponds to the frequency or count of data points in that category.
3. Pie Chart: A pie chart displays the proportion of different categories as sectors of a circular pie. Each category is represented by a slice, and the size of the slice represents its proportion of the whole.
4. Line Plot: A line plot shows the relationship between two variables by connecting data points with straight lines. It is often used to visualize trends and patterns over time or any continuous variable.
5. Scatter Plot: A scatter plot displays the relationship between two continuous variables as a collection of points on a two-dimensional plane. It helps to identify patterns, clusters, and correlations between variables.
6. Box Plot: A box plot, also known as a box-and-whisker plot, provides a summary of the distribution of a continuous variable. It displays the median, quartiles, and potential outliers of the data.
7. Violin Plot: A violin plot combines aspects of a box plot and a kernel density plot. It displays the distribution of a variable as a rotated kernel density plot on each side, with a box plot in the middle.
8. Heatmap: A heatmap represents the correlation or relationship between multiple variables as a grid of colored cells. It is commonly used to visualize the patterns and strengths of relationships in a dataset.

9. Area Plot: An area plot, also known as a stacked area chart, displays the cumulative contribution of different variables over time or any continuous dimension. It helps to visualize the overall composition and changes in composition.
10. Pair Plot: A pair plot, also known as a scatter plot matrix, displays pairwise relationships between multiple variables in a dataset. It shows scatter plots for each combination of variables, allowing for a comprehensive view of their relationships.
11. Radar Chart: A radar chart, also known as a spider chart or star plot, displays multivariate data on a two-dimensional plane with multiple axes. It is useful for comparing and visualizing multiple variables simultaneously.

Complete guid of data sampling

Population/Sample

Methods or estimation of sample size

Probability sampling

Types of random sampling

- Simple random sample
- Stratified sampling
- Systematic sampling
- Cluster sampling

Non probability sampling

- Accidental sampling
- Quota sampling
- Judgmental sampling
- Snowball sampling

Errors in data collection

Sampling error

Non sampling error

Data sampling refers to the process of selecting a subset of data points from a larger dataset. It is commonly used in machine learning for various purposes, such as creating training and validation sets, addressing class imbalance, or reducing the computational requirements of the model. Here are some popular data sampling methods:

1. Random Sampling: Randomly selects data points from the dataset without any specific criteria or bias. It is a simple and straightforward sampling technique.
2. Stratified Sampling: Ensures that the sample maintains the same class distribution as the original dataset. It is useful when dealing with imbalanced datasets or when preserving the representation of different classes is important.
3. Oversampling: Increases the number of instances in the minority class to balance the class distribution. Techniques such as Random Oversampling, SMOTE (Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling) are commonly used for oversampling.
4. Undersampling: Reduces the number of instances in the majority class to balance the class distribution. Techniques like Random Undersampling and Tomek Links are commonly used for undersampling.
5. Synthetic Sampling: Generates synthetic data points in the minority class to balance the class distribution. SMOTE and ADASYN are examples of synthetic sampling methods.
6. Cluster-based Sampling: Identifies clusters or groups of similar instances in the dataset and selects representative samples from each cluster. This method helps to reduce the redundancy of data points.
7. Near-Miss Sampling: Selects instances from the majority class that are close to instances in the minority class, aiming to improve classification performance in imbalanced datasets.
8. Stratified Sampling with Oversampling: Combines stratified sampling with oversampling to create a balanced sample while preserving the distribution of different classes.
9. Ensemble-based Sampling: Involves creating multiple samples using different sampling techniques and combining them to form an ensemble dataset. This approach can help improve the model's robustness and generalisation ability.
10. Bootstrap Sampling: Randomly samples data points with replacement from the original dataset. It is commonly used for estimating uncertainty and constructing confidence intervals.
11. Adversarial Sampling: Selects data points that are challenging for the model to classify correctly. It helps in creating a more robust and generalizable model by exposing it to difficult instances.
12. Importance Sampling: Assigns different sampling probabilities to data points based on their importance or relevance. It is often used in scenarios where certain instances are more critical or representative than others.

These are some of the commonly used data sampling methods in machine learning. The choice of sampling technique depends on the specific problem, the nature of the data, and the goals of the analysis or modelling task.

Certainly! Here is a comprehensive list of data sampling methods commonly used in machine learning:

1. Random Sampling
2. Stratified Sampling
3. Oversampling
    - Random Oversampling
    - SMOTE (Synthetic Minority Over-sampling Technique)
    - ADASYN (Adaptive Synthetic Sampling)
    - Borderline-SMOTE
    - SVM SMOTE (Support Vector Machine Synthetic Minority Over-sampling Technique)
    - SMOTE-ENN (SMOTE Edited Nearest Neighbors)
    - SMOTE-Tomek
4. Undersampling
    - Random Undersampling
    - Tomek Links
    - Edited Nearest Neighbors (ENN)
    - Instance Hardness Threshold (IHT)
    - NearMiss
5. Synthetic Sampling
    - SMOTE (Synthetic Minority Over-sampling Technique)
    - ADASYN (Adaptive Synthetic Sampling)
    - Borderline-SMOTE
    - SVMSMOTE (Support Vector Machine Synthetic Minority Over-sampling Technique)
    - SMOTE-ENN (SMOTE Edited Nearest Neighbors)
    - SMOTE-Tomek
6. Cluster-based Sampling
    - K-Means Clustering
    - Mini-Batch K-Means
    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
    - Hierarchical Clustering
7. Near-Miss Sampling
    - NearMiss-1
    - NearMiss-2
    - NearMiss-3

8. Stratified Sampling with Oversampling
9. Ensemble-based Sampling
10. Bootstrap Sampling
11. Adversarial Sampling
12. Importance Sampling
13. Systematic Sampling
14. Cluster Sampling
15. Stratified Cluster Sampling
16. Multi-stage Sampling
17. Time-based Sampling
18. Density-based Sampling
19. Active Learning Sampling
20. Cost-sensitive Sampling
21. SMOTEBoost
22. RUSBoost
23. EasyEnsemble
24. BalanceCascade
25. Condensed Nearest Neighbors (CNN)
26. One-Sided Selection (OSS)
27. Neighborhood Cleaning Rule (NCL)
28. Random Subspace Sampling

These methods cover a range of sampling techniques used for various purposes, including class imbalance handling, dataset reduction, uncertainty estimation, and model robustness. Keep in mind that the suitability of each method depends on the specific dataset and the machine learning problem at hand.

Certainly! Probability-based sampling and non-probability-based sampling are two broad categories of sampling methods used in research and statistics.

1. Probability-Based Sampling: Probability-based sampling involves selecting samples from a population in such a way that each element has a known and non-zero chance of being included in the sample. The main characteristic of probability-based sampling is that it allows for the calculation of sampling probabilities and sampling errors, enabling statistical inference and generalisation. Here are some common probability-based sampling techniques:
    - Simple Random Sampling: Each element in the population has an equal chance of being selected. Random selection can be done with or without replacement.
    - Systematic Sampling: Elements in the population are selected at fixed intervals after an initial random start.

    - Stratified Sampling: The population is divided into homogeneous subgroups called strata, and samples are independently selected from each stratum. This method ensures representation from each subgroup.
    - Cluster Sampling: The population is divided into clusters (e.g., geographical regions), and a subset of clusters is randomly selected. All elements within the selected clusters are included in the sample.
    - Multistage Sampling: A combination of cluster and random sampling, where clusters are successively sampled in multiple stages until individual elements are selected.

Probability-based sampling methods provide a solid foundation for statistical inference, as they allow researchers to estimate population characteristics and make generalisations with known degrees of uncertainty.

2. Non-Probability-Based Sampling: Non-probability-based sampling methods do not involve random selection and do not provide a basis for calculating sampling probabilities or measuring sampling errors. These methods are often used when probability-based sampling is impractical or not feasible. Non-probability sampling methods are commonly employed in qualitative research, exploratory studies, or when the focus is on specific subgroups rather than generalisation to a larger population. Here are a few examples:
    - Convenience Sampling: Selecting participants based on their convenient availability or accessibility.
    - Purposive Sampling: Handpicking participants who meet specific criteria or possess desired characteristics.
    - Snowball Sampling: Identifying initial participants and then asking them to refer others who meet the study criteria.
    - Quota Sampling: Selecting individuals to match predefined quotas based on certain characteristics (e.g., age, gender, occupation) to ensure representation of subgroups.

Non-probability-based sampling methods have limitations regarding generalizability and potential bias, as the sample may not be representative of the population. However, they can still be useful in certain research contexts, such as exploratory studies, qualitative research, or when specific subgroups are of interest.

It's important to carefully consider the research goals, available resources, and characteristics of the population when selecting a sampling method, as each method has its own advantages and limitations.