Statistics and Probability - Detailed Notes

---

1. Introduction to Statistics

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make decisions. It helps in understanding patterns and trends in data.

---

2. Sampling

Definition: Sampling is the process of selecting a subset of individuals from a population to estimate characteristics of the whole population.

Types of Sampling:

- **Random Sampling**: Each individual has an equal chance of being selected.
- **Stratified Sampling**: Population divided into groups (strata), then samples taken from each.
- **Systematic Sampling**: Every kth item is selected from a list.

Example:
If we have a population of 1000 students, and we randomly pick 100 students to analyze their performance, it is called sampling.

---

3. Central Tendencies

Central tendency refers to the middle or typical value in a dataset.

- **Mean (Average)** = Sum of all values / Total number of values
- **Median** = Middle value after sorting data
- **Mode** = Most frequent value

Example:
Given data: [2, 3, 5, 7, 7, 10]

- Mean = (2+3+5+7+7+10)/6 = 5.67
- Median = (5+7)/2 = 6
- Mode = 7

4. Null Values

Definition: Null values represent missing or undefined data in a dataset.

Handling Techniques:

- Removing null rows/columns
- Imputing values using mean/median/mode

Example:

Name	Age
A	23
B	NaN

Impute with mean: Age = 23 (if only one value exists)

5. Duplicates

Definition: Duplicate records are repeated entries in a dataset.

Handling:

- Use `drop_duplicates()` in Python (Pandas) to remove them.

Example:

Name	Age
A	23
A	23

After dropping duplicates, only one record remains.

---

6. Range

Definition: Range is the difference between the maximum and minimum values in a dataset.

Formula: Range = Max - Min

Example:
Data: [3, 7, 2, 9, 5] → Range = 9 - 2 = 7

---

7. Variance

Definition: Variance measures the spread of data points around the mean.

Formula:
Variance=1n∑i=1n(xi−xˉ)2\text{Variance} = \frac{1}{n} \sum_{i=1}^{n}(x_i - \bar{x})^2

Example:
Data: [2, 4, 4, 4, 5, 5, 7, 9]
Mean = 5, Variance = 4

---

8. Standard Deviation (SD)

Definition: Standard deviation is the square root of variance. It indicates how data values spread around the mean.

Example:
If variance = 4, then SD = √4 = 2

---

9. Percentile

Definition: A percentile indicates the value below which a given percentage of observations fall.

Example:
If you are in the 90th percentile in a test, you scored better than 90% of test takers.

---

10. Quantile

Definition: Quantiles divide the dataset into equal-sized intervals.

- Quartiles (4 parts), Deciles (10 parts), Percentiles (100 parts)

Example:
25th percentile = Q1 (first quartile), 50th percentile = median (Q2)

---

11. Outliers

Definition: Outliers are extreme values that differ significantly from other observations.

Detection Methods:

- IQR method:
  - IQR = Q3 - Q1
  - Lower bound = Q1 - 1.5 * IQR
  - Upper bound = Q3 + 1.5 * IQR

Example:
Data: [2, 3, 4, 5, 6, 100] → 100 is an outlier

---

12. Correlation

Definition: Correlation measures the relationship between two variables.

Range: -1 to 1

- +1: Perfect positive correlation
- -1: Perfect negative correlation
- 0: No correlation

Example:
Height and weight often show positive correlation.

---

13. Plots

Useful for visualizing data.

- **Histogram**: Distribution of numerical data
- **Boxplot**: Detects outliers and spread
- **Scatter plot**: Correlation between two variables
- **Bar chart**: Categorical data comparison

Example:
Use matplotlib/seaborn in Python:

```
import seaborn as sns
sns.boxplot(data=[2,4,4,4,5,5,7,9,100])
```

---

14. Probability

a. Joint Probability
Probability of two events happening together.
P(A∩B)=P(A)⋅P(B∣A)P(A \cap B) = P(A) \cdot P(B|A)

b. Marginal Probability
Probability of a single event irrespective of others.

c. Conditional Probability
Probability of A given B has occurred.
P(A∣B)=P(A∩B)P(B)P(A|B) = \frac{P(A \cap B)}{P(B)}

Example:
If 60% students play cricket and 30% of those also play football:

- Joint = 0.6 * 0.3 = 0.18

---

16. Skewness

Definition: Skewness measures the asymmetry of the distribution.

- Positive skew: Tail on right
- Negative skew: Tail on left

Example:
Income data often shows positive skew due to few very high incomes.

---

17. Kurtosis

Definition: Kurtosis measures the "tailedness" of a distribution.

- High kurtosis: heavy tails (outliers)
- Low kurtosis: light tails

Example:
Normal distribution has kurtosis = 3 (mesokurtic)

---

18. Label Encoding

Definition: Converts categorical labels into numerical form.

Example:
Colors: [Red, Green, Blue] → [0, 1, 2]

In Python:

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
le.fit_transform(['Red', 'Green', 'Blue'])
```

End of Notes