BENIN REPUBLIC

****

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

****

UNIVERSITY OF PARAKOU

NATIONAL SCHOOL OF PLANNING, STATISTICS AND DEMOGRAPHY

****

Level : Master I

Groupe 3 practical work

Theme :

Terms used in Statistics

Produced and presented by :

1- ADENIYI Zaya-Dine

2- BORORI SANNI Eve

3- CODJOKINKIN Y. Alain

4- VODA Eustache

5- OLOUKOU Fabrice

6- ISSIFOU Abidardaye

7- SABI K. B. Kpénté

Under the supervision of :

Dr. Béatrice N. K. M'PO

January 2025

Summary

Plan

Introduction

Statistics is a fundamental discipline that involves the collection, analysis, interpretation, presentation and organization of data. Understanding key statistical terms is essential to effectively communicating and interpreting data-driven insights.

Indeed, our daily lives are punctuated by data, and statistical terminologies are increasingly present in our conversations and activities.

This work aims to present some terms relating to the main areas of statistics which are : descriptive statistics, inferential Statistics, probability, regression Analysis, data Analysis, sampling Methods and experimental design.

I- Descriptive statistics

Statistic : A statistic is a value that has been produced from a data collection, such as a summary measure, an estimate or projection. Statistical information is data that has been organised to serve a useful purpose.

Data : They are measurements or observations that are collected as a source of information.

Data item (or variable) : A data item is a characteristic (or attribute) of a data unit which is measured or counted, such as height, country of birth, or income.

Administrative data : They are collected as part of the day to day processes and record keeping of organisations.

Mean (Average) : The sum of values divided by the number of observations.

Median : The middle value when data is ordered.

Mode : The most frequently occurring value.

Standard Deviation (SD) : A measure of the spread or dispersion of data.

Variance : The square of the standard deviation.

Range : The difference between the highest and lowest values.

Quartiles : Values dividing data into four equal parts.

Interquartile Range (IQR): The difference between the third and first quartiles.

Skewness : A measure of data symmetry.

Kurtosis : A measure of the ”tailedness” of the data distribution.

Weighted Mean : A mean where some values contribute more than others.

Geometric Mean : The *n*th root of the product of *n* values, used for growth rates.

Harmonic Mean : The reciprocal of the average of reciprocals, often used for rates.

Coefficient of Variation (CV) : A standardized measure of dispersion.

Empirical Rule : In a normal distribution, about 68%, 95%, and 99.7% of values fall within 1, 2, and 3 standard deviations from the mean.

Z-Scores : Standardized values that indicate the number of standard deviations a data point is from the mean.

Boxplot (Five-Number Summary): Minimum, Q1, Median, Q3, and Maximum.

II- Inferential Statistics

Population : The entire group being studied.

Sample : A subset of the population.

Parameter : A numerical characteristic of a population.

Statistic : A numerical characteristic of a sample.

Confidence Interval (CI) : A range of values likely to contain the population parameter.

Hypothesis Testing : A method to test assumptions about a population.

P-value : The probability of observing results as extreme as those in your data, assuming the null hypothesis is true.

Significance Level (α) : The threshold for rejecting the null hypothesis.

Null Hypothesis ($H0$) : The default assumption, e.g., no effect or no difference.

Alternative Hypothesis (*H*1) : The hypothesis contrary to the null hypothesis.

Type I Error : Rejecting a true null hypothesis.

Type II Error : Failing to reject a false null hypothesis.

Likelihood Ratio Test : A hypothesis test comparing the goodness of fit of two models.

Bayesian Inference: A method using Bayes' theorem to update probabilities based on evidence.

Bootstrap Methods: Resampling techniques for estimating the sampling distribution of a statistic.

Monte Carlo Simulation: Using random sampling to model and analyze the behavior of a system.

Degrees of Freedom: The number of independent pieces of information in a calculation.

Effect Size : A measure of the strength of a phenomenon, e.g., Cohen's *d* or eta-squared (η 2).

Nonparametric Tests : Tests that do not assume a specific distribution (e.g., Mann-Whitney U, Kruskal-Wallis)

III- Probability

Probability Distribution : A function defining the likelihood of outcomes.

Normal Distribution : A bell-shaped probability distribution.

Binomial Distribution : For experiments with two possible outcomes (success/failure).

Poisson Distribution : For the probability of a given number of events in a fixed interval.

Random Variable : A variable whose values depend on random phenomena.

Cumulative Distribution Function (CDF) : The probability that a variable takes on a value less than or equal to X.

Probability Mass Function (PMF) : For discrete variables, the probability of each possible outcome.

Probability Density Function (PDF) : For continuous variables, the like lihood of a variable falling within a certain range.

Markov Chains : A stochastic process with memoryless properties.

Bayes' Theorem : A formula describing the probability of an event based on prior knowledge.

Central Limit Theorem (CLT) : The principle that sample means approximate a normal distribution as the sample size grows.

Joint Probability Distribution : The probability of two events occurring together.

Conditional Probability : The probability of an event given another event.

IV- Regression Analysis

Linear Regression : A method to model the relationship between a dependent and independent variable.

Multiple Regression : Regression with more than one independent variable.

Coefficient of Determination (*R*2) : Measures how well the regression fits the data.

Residual : The difference between observed and predicted values.

Logistic Regression : Used for binary or categorical dependent variables.

Ridge Regression (L2 Regularization) : A technique to prevent multicollinearity in linear regression by adding a penalty for large coefficients.

Lasso Regression (L1 Regularization) : A regression method that performs variable selection and regularization.

Interaction Effects : How the effect of one variable changes depending on another.

Heteroscedasticity : Non-constant variance of residuals across levels of an independent variable.

Autocorrelation : Correlation of a variable with itself over successive time intervals.

Generalized Linear Models (GLM) : An extension of linear regression for non-normal response distributions.

Variance Inflation Factor (VIF) : A measure of multicollinearity in regression models.

V- Data Analysis

Outliers : Data points significantly different from others.

Correlation : A measure of the relationship between two variables.

Covariance : A measure of how two variables vary together.

Time Series Analysis : Analyzing data points collected over time.

Data Transformation : Modifying data for analysis (e.g., logarithmic transformation).

Principal Component Analysis (PCA): A method for reducing dimensionality while retaining variance.

Cluster Analysis : Grouping observations into clusters (e.g., k-means, hierarchical clustering).

Factor Analysis : Identifying underlying relationships between variables.

Discriminant Analysis : A technique for classifying observations into predefined groups.

Entropy : A measure of randomness or uncertainty in data.

Cross-Validation : Splitting data into training and testing sets to evaluate model performance.

Hidden Markov Models (HMM) : Statistical models for time-series data where the system state is partially observable.

VI- Sampling Methods

Random Sampling : Each member of the population has an equal chance of being selected.

Stratified Sampling : Dividing the population into strata, then sampling within each.

Cluster Sampling : Sampling entire clusters or groups.

Systematic Sampling : Selecting every *n*th observation.

Snowball Sampling : Recruiting subjects through referrals from initial participants.

Sequential Sampling : Sampling in stages based on previous results.

Multistage Sampling : A combination of multiple sampling methods.

Adaptive Sampling : Modifying the sampling process based on initial findings.

Jackknife Sampling : A resampling technique for estimating the bias and variance of a statistic.

Importance Sampling : A variance reduction technique used in Monte Carlo simulations.

VII- Experimental Design

Control Group : The group not exposed to the experimental treatment.

Treatment Group : The group exposed to the treatment.

Randomization : Assigning subjects randomly to groups.

Replication : Repeating experiments to increase reliability.

Latin Square Design : A design ensuring each treatment appears exactly once in each row and column.

Factorial Design : Testing all possible combinations of factors and levels.

Randomized Block Design : Grouping similar experimental units into blocks before random assignment.

Cross-Over Design : Subjects receive different treatments in different periods.

Split-Plot Design: A design for experiments with two levels of randomization.

Confounding Variables : Variables that distort the relationship between the independent and dependent variables.

Interaction Effects : When the effect of one factor depends on another.

Conclusion

Understanding these statistical terms enhances comprehension of data analysis processes and improves communication among researchers and practitioners in various fields. Mastery of these concepts allows for better interpretation of research findings and facilitates informed decision-making based on statistical evidence. This will also allow ordinary people to better understand statistical concepts and limit errors in their uses.

Bibliographic references

Holosko, M. J., & Thyer, B. A. (2011). Pocket glossary for commonly used research terms. SAGE Publications.

Australian Bureau of Statistics : https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/statistical-terms-and-concepts-glossary

Table of contents