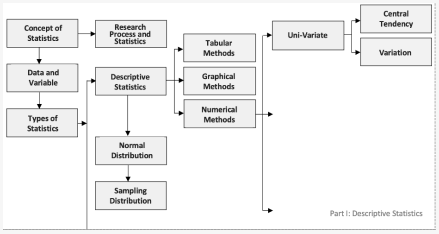
Descriptive Statistics Pathway



Types of Statistical Methods

Descriptive Statistics

> Methods that help us to describe sample or population **Example:** Standard deviation

Process

> 1- Identify population or sample 2- Identify variable of interest 3- Collect data 4- Describe data

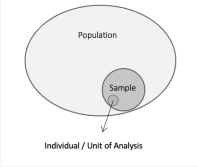
Inferential Statistics

> Methods that help us to use sample information in order to draw conclusions regarding the population **Example:** Confidence Interval

Process

> 1- Identify population 2- Identify variable of interest 3- Collect sample data 4- Inference about population based on sample data (Make statements or predictions about the population)

Population /Sample



Ch1. Review Questions

Walter Wallace's wheel of science best illustrates which of the following?

> The relationship between theory and data

A characteristic of the individual to be measured or observed is:

> Variable

You are studying relationship between activity level and body fat percentage of females in New Jersey. What is the independent variable and what is the control variable of your study?

> Activity level: Independent variable. Gender:Control variable

Identify: 1) the level of measurement 2) Whether the variable is discrete or continuous

What is the main mode of transportation you use to commute to work? 1. by foot 2. by bicycle 3. by private vehicle, including car, truck, van, taxicab, and motorcycle 4. by public transportation, including bus, rail, and ferry

> **1)**Level of Measurement: Nominal.
> **2)**Neither Continuous or Discrete.

People convicted of first-degree murder should be executed. Strongly Agree Agree Neither Agree nor Disagree Disagree Strongly Disagree

> **1)**Level of Measurement:Ordinal.
> **2)**Neither Continuous or Discrete.

Methods

Data Collection

Data Organization

Data Presentation

Data Analysis and Interpretation

Data Collection Critical Issues

SOP / MOP

> A Standard Operating Procedure or Manual Operating Procedure is a set of step-by-step instructions compiled by a research team to guide individuals carry out the research project operations.

Blind Group

> *Single Blind:* Participant is blind to the collected data. *Double Blind:* Participant and researcher are blind to the collected data. *Triple Blind:* Participant, researcher, and project manager are blind to the collected data.

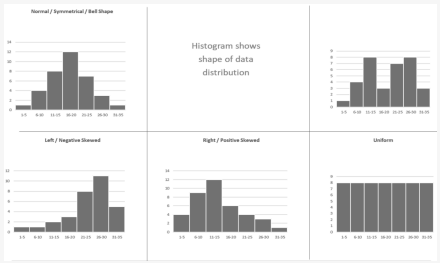
Placebo

> A Substance that's designed to have no therapeutic value.

QA / QC

> Quality Assurance is process orientated and focuses on problem prevention. QAs are set of approaches, methods, processes and actions formulated to assure quality of the project and its focus is on problem identification and correction.

Shape of Data Distribution



Variation Analysis: Coefficient of Skewness

$$\text{Coefficient of Skewness (CS)} = \frac{3 * (\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Coefficient of Skewness

Concept:

What is shape of data distribution using numerical form

Shows

direction of skewness (sign). Strength of skewness (value). Comparison of several distributions

Variation Analysis: Coefficient of Variation

$$\text{Coefficient of Variation (CV)} = \frac{\text{Standard Deviation}}{\text{Mean}} * 100$$

Relative Standard Deviation

Coefficient of Variation

Concept:

What is shape of data distribution using numerical form

Potentials

Shows degree of skewness within the data set. Easy to compare several distributions

Chebyshev Theorem

$$1 - \frac{1}{k^2}$$

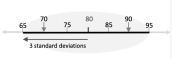
k ← number of standard deviations on either side of the mean

Concept: How much of data is located close to mean and how much is far from the mean.

Chebyshev Theorem Example

Ex: Average grade of students is 80 and its standard deviation is 5. What percent of students received between 65 and 95?

$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 0.89 = 89\%$$



Interpretation: **At least** 88.9% of students received between 65 and 95.

Variation Analysis: Z Score

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score
x = observed value
μ = mean of the sample
σ = standard deviation of the sample

Concept: How far one data is from the mean, in units of standard deviation
Positive Z indicates that the data is above average, Negative Z indicates that it is below average
If data distribution is normal, then Z is Standard Z Score

Quantitative Vs Qualitative Methods

Quantitative Research

Test Theories/hypothesis

Seeks to Measure/Quantify

Methodology

o Numerical Values

o Closed Questions

Required Data

Requires Many Respondents

Outcomes

o May be generalized to the population.

o Guidance for decision making.

Key Terms

Testing, Measurement, Objectivity, replicability

Quantitative Vs Qualitative Methods

Qualitative Research

Seeks to explain and understand.

Exploring ideas and formulating theory and hypothesis

Methodology

o Text/Narratives

o Open-ended Question

Required Data

Requires Few respondents.

Outcome

Quantitative Vs Qualitative Methods (cont)

o Not projectable to total populations

o Understandings for new ideas

Key Terms

Understanding, context, complexity, subjectivity

Population/ Sample

Population

all individuals of objects under study.

Sample

a portion of population.

Individual/Unit of Analysis

an indivudals or object under study .

Parameter

a numerical measure that describes an aspect of a population.

Statistic

a numerical mesure that describes an aspect of a sample.

Sampling Frame

a list of individuals from which a sample is taken.

Under-coverage

the outcome of omitting some individuals from the sample frame.

Random Sample

a sample selected in a manner that every individual has equal chance to be selected as a member of the sample.

Population/ Sample (cont)

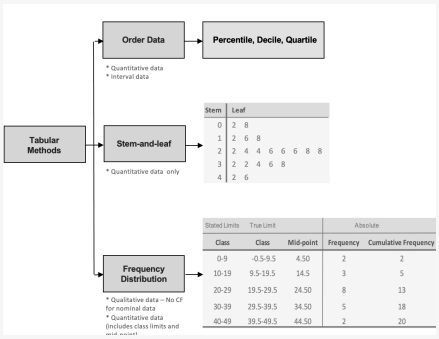
Sampling Error

the difference between a sample statistics and corresponding population parameter.*Sampling error can not be determined.Sampling error most probably decreases as the sample size is increased.*

Non-sampling Error

Mainly the result of: Poor sample design, problematic data collection, biases involved in data collection

Data Organization: Tabular Methods



Data Organization: Frequency Distribution

Applications

Distribution of data within categories.

Guidelines

Must be mutually exclusive (Non-Overlap classes) and Exhaustive (Classes for every observation) .

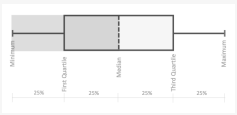
If possible:

5-10 classes. Equal class width (Size, Range, Interval). Avoid open-ended classes (More than $100,000). List all classes including those which have zero frequency.

Data Presentation

Types:	Static, Interactive, Animated
Nature of Data	
Univariate Charts	Line Chart, Bar Chart, Pie Chart, Histogram, Bi-Histogram, Polygon, Box-Plot, Q-Q Plot, P-P Plot
Bivariate Charts	Scatter Plot
Multivariate Charts	Star Chart, Radar Chart, Web Chart, Area Chart

Variation Analysis: Box Plot



Variation Analysis: Range

Range

Gap between 2 data *Gap between minimum and maximum*

Range = R= Maximum - Minimum

Mid-Range

Half of maximum and minimum data together

Mid-range = MR = Maximum + Minimum/2

Inter-Quartile Range

IQR=Q3-Q1

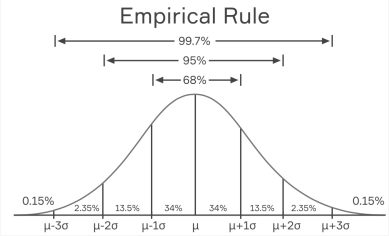
Inter-Quartile Deviation

IQD = QD = (Q3-Q1) / 2

Index of Qualitative Variation

Concept:	The index of qualitative variation (IQV) is a measure of variability for nominal data.
IQV =	Number of observed differences / Maximum possible differences

Empirical Rule (Normal)



Concept: If data is normally distributed, what percentages of data is located close to mean and how much is far from the mean.

About 68% of the data fall in the interval of μ - 1 to μ + 1

About 95% of the data fall in the interval of μ - 2 to μ + 2

About 99% of the data fall in the interval of μ - 3 to μ + 3

Standard deviation is about one forth of the range (R/4)

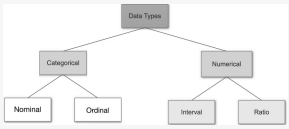
Classical Probability

- Probability of an event = $\frac{f}{n}$
- Probability of event A = P(A) = $\frac{frequency\ of\ A}{Total\ number\ of\ possibilities}$

Ex: When we roll a fair die, what is probability of having 2?

Probability of 2 = P(2) = 1/6 = 16.6%

Types of Data



Data & Variable

Data

> Facts and values of variables collected for analysis/presentation

Qualitative. Data

> Values have qualitative format. **Example:** Mode of transportation

Quantitative. Data

> Values of a meaningful quantitative format. **Example:** Age of Customers

Data & Variable

Variable any characteristics of an individual that can change from individual to individual.

Variable Types

1. Independent Variable (explanatory)

the variable that you purposely change or control in order to see what effect it has.

Variable that is changed or controlled in a study / scientific experiment

2. Dependent Variable (response)

the variable that responds to the change in the independent variable.

3. Lurking Variable

the variable that can have effect on the variables being studied but is not included in the study.

4. Control Variable

the variable that you keep it controlled in order to reduce effects on known and unknown variables.

5. Confounded Variable

a variable that influences both the dependent variable and independent variable.

Variable Type Example

Example A researcher wants to assess the effectiveness of drug X on recovery duration.

Independent Variable (explanatory):

> Amount of drug X intake (mg)

Dependent Variable (response):

> Recovery duration (days)

Lurking Variables:

> Gender, Age, Weight, Ethnicity, Other illnesses, etc

Control Variable:

> Researcher will collect data from 25-40 years old patients to keep age constant.

Confounded Variable:

> Other drugs the patient is taking

Frequency Distribution: Relative Analysis

Proportion

> $Frequency/(Sample\ size)$

Percentage

> $Frequency/(Sample\ size)*100$

Percent Change

> $(F2-F1)/F1*100$

> F1 is the earlier frequency. F2 is the later frequency. % change can be positive or negative

Ratio

> $F1/F2$

> F1 is the number of cases in 1st category. F2 is the number of cases in 2nd category.

Frequency Distribution: Relative Analysis (cont)

Rate

> (# *of actual occurance*)/(# ☐ ☐*f total possibilities*)

> 122 death / 7000 population = 0.0174.
> Crude Death Rate = 0.0174 * 1000.
> Crude Death Rate = 17.4 per 1000.

Density

> $Frequency/(Class\ Width)$

> Density is the only measure that requires quantitative data.

Data Category/ Level of Measurement/- Nature of Data

Qualitative (Categorical)

> Codes: Religious background, Political Affiliation

Nominal :

> Objects fall into unordered categories. Example: Race, Religious background

Ordinal:

> Objects fall into naturally ordered categories. Example: Performance, Economic class

Binary

Non-Binary

> Objects fall into sevral catergories. Example: Level of community participation

Quantitative (Numerical)

> Numbers: Income, Family Size

Mean

Median/Quartiles

Position of Median in Data Set	$(n+1)/2$
Quartiles	Position of Q1: $(n+1)* 0.25$
	Position of Q3: $(n+1)* 0.75$

Mean Deviation & Standard Deviation

$$MD = \frac{\sum |X - \bar{X}|}{n}$$

Sample Variance

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

Population Variance (Lowercase sigma)

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Variation Analysis: Percentile, Decile, Quartile

Concept: where one data stands in comparison to the rest of data within the data set
Percentile: Percent of data which is equal or less than a given data
Decile: Which decile a data falls in
Quartile: Which quartile a data falls in