

Research Methods Workshops

Data Analysis: Statistical Analysis
Principles and where to use them

Readings

Essential Readings:

- Berenson et al. Chapter 1, 3 and 4
- Groebner et al. Chapter 4
- *Research Methods for Business Students*, Mark Saunders, Philip Lewis, Adrian Thornhill, Pearson Education, 2009

Learning Outcomes

By the end of this session you should be able to:

- Understand what statistics is and what it is used for.
- Explain what is meant by descriptive statistics and inferential statistics.
- Calculate the arithmetic mean, the weighted mean (not all data points contribute equally to the mean), the geometric mean (measure of central tendency), the median, and the mode.
- Compute and interpret the range, the variance, the standard deviation, covariance and correlation coefficient.
- Match the appropriate statistical test to the types of data you have collected.

What is statistical analysis?

- **Statistics** is the science of collecting, organizing, presenting, analyzing, and interpreting numerical/non-numerical data for the purpose of assisting in making a more effective decision or forming a judgment as to the likelihood of a hypothesis being true.
- It is also a branch of mathematics which transform numbers (especially from very large data sets) into useable information for researchers, decision makers and interpreters of those data sets.
- It also provides several means of validating whether our analysis of data justifies the inferences we make, based on that data and the confidence we can place in those inferences.

Types of Statistical Analysis

- The branch of mathematics that transforms data into useful information to be presented as evidence.



Descriptive Statistics

Collecting, summarizing, and describing data - in order to present evidence for findings.

Converting numerical data sets into graphical representations.

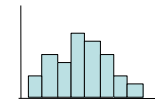


Inferential Statistics

Drawing conclusions and/or establishing relationships amongst two or more variables, concerning a population based only on sample data.

Descriptive Statistics

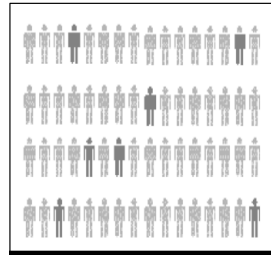
- Collect data
 - e.g., Survey
- Present data
 - e.g., Tables and graphs



- Characterize data
 - e.g., Sample mean = $\frac{\sum X_i}{n}$ (arithmetic mean)

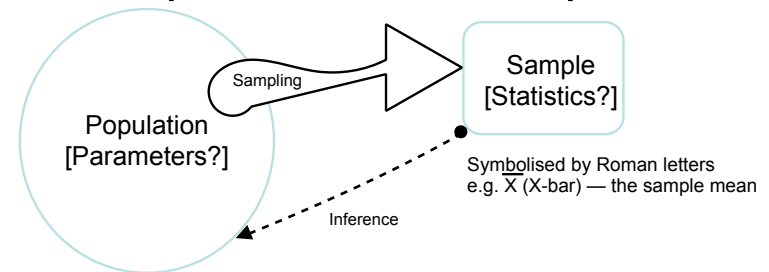
Inferential Statistics

- Estimation
 - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
 - e.g., Test the claim that the population mean weight is 55 Kg.
- Testing a possible relationship between two or more variables



Drawing a conclusion about a population based on data sampled from a subset of the population

Population and samples



Symbolised by Greek letters
e.g. μ (Mu) — the population mean

Symbolised by Roman letters
e.g. $\bar{X}$ (X-bar) — the sample mean

- Both descriptive and inferential statistics require sampling a subset of a population (defined by characteristic parameters - e.g. consumers spending £20 or more on clothing (per week)).
- Statistical data analysis is applied to the sample in the hope that inferences can be drawn which will hold true of the population as a whole.
- There is no guarantee that the sample mean and the population mean will coincide - depends on whether the sample is representative.

Measures of Central Tendency

(and when you should use them!)

Measure of Central Tendency (“the mean (several types)”, “average”, “median”, “mode”, “mid-inter quartile range”): A single value that summarises a set of data. It locates the centre of the values -- and indicates (in a normally distributed population) the most likely value – one that is most typical (or representative).

The **central tendency** is the extent to which all the data values group around a typical or central value.

Note:

There is not just one measure of central tendency; in fact, there are many. i.e. the arithmetic mean, the weighted mean, the median, the mode, the geometric mean and the harmonic mean (and sometimes the mid-point interquartile range).

[See Exercise 2].

Types of variable (and “levels of measurement”)

- Most inferential statistics is designed to test whether a correlation exists between two or more variables (and how strong the correlation is).
- Typically, we attempt to establish a correlation between ‘independent variables (IV)’ and a ‘dependent variable (DV)’:

IV = one or more variables we either manipulate or observe changes in numerical value which we believe will influence changes in DV.

DV = a variable we predict is influenced by the changes in IV.

- **Examples:**
 - i. Test whether ‘footfall’ (number of people entering a shop) - DV - is influenced by whether it is raining outside or not - IV.
 - ii. Attempt to correlate level of satisfaction of employees at work (DV) with their level of pay (IV).
 - iii. Evaluate whether there is a significant correlation between change in the ‘industrial production index (IPI in the USA) - IV - and national economic growth - DV.

Types of Data (2)

- Despite some debate amongst academic researchers, it is customary to classify data collected from a sample into one of four types (from work by Stanley Stevens in the early 1950's):
 1. **Nominal** (also called "categorical"): names or categories of parameters within the population - no ordering of responses.
 2. **Ordinal**: data can be ordered on a scale: high/low, most/least etc.
 3. **Interval-Based**: Interval scales are numerical scales in which intervals have the same interpretation throughout
 4. **Ratio**: An interval scale with the additional property that its zero position indicates the absence of the item being measured
- The kinds of statistical test and measure of central tendency you can use depend upon which of these four types your data fits.

Applying a statistical test to the wrong type of data makes your calculation *statistically meaningless*!

- We will consider each of these four in turn, with examples, and discuss which "measures of central tendency" are appropriate to each type of data.

11

Nominal Data

- When measuring using a nominal scale, one simply names or categorises possible responses. 'Gender', 'job title', 'favourite colour', and 'religion' are examples of variables measured on a nominal scale.
- A good nominal scale must ensure that categories occupy the same level within the scale (and be mutually exclusive): e.g.

[Clothing purchases made by customers]:

hats, scarves, shirts, tops, dresses, trousers, socks — is **OK**.

hats, shirts, women's clothing, men's clothing, accessories - is **NOT**.

- There is no ordering implied by the responses, therefore it is **meaningless** to calculate the 'mean response (average)' across all responses.
- When evaluating nominal data, better to use 'frequency' measures and the 'mode' (as a measure of central tendency) - **NOT** the average (mean)!
- Nominal data is sometimes called 'non-parametric' (within the population) because there is no measure of rank nor order of response.

12

Ordinal Data

- The items in this scale are ordered, ranging from low to high. This is what distinguishes ordinal from nominal scales.
- Likert scaled responses (to a questionnaire) are typically regarded as ordinal (e.g. 'Strongly Disagree, Disagree, Neither agree nor disagree, Agree, Strongly Agree') because there is an implied ordering.
- Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the desired characteristic. [For example, our 'agreement' ordering makes it meaningful to assert that one person agrees more than another with (say) the quality of service].
- The difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels.

Owing to this (though some statisticians disagree), it is usually **meaningless** to calculate the mean ('average') across all responses to ordinal data. The more appropriate 'measure of central tendency' is the **median** score (because its value is on the scale - whereas an 'average' may not be).

13

Interval Data

- Interval scales are numerical scales in which intervals have the same interpretation throughout.

As an example, the Fahrenheit scale of temperature: the difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees (unlike an ordinal scale). Consider also income scales, age scales.

- Interval scales are not perfect. They do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature.
- Since an interval scale has no true zero point, it makes **no sense** to calculate ratios of temperatures.

For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting property is preserved across the two ratios. Is 40 degrees so much hotter than 20 degrees than 100 degrees is so much hotter than 50 degrees? (both measures are in the same ratio (one half)).

- The mode, median, and arithmetic mean are allowed to measure central tendency of interval variables, while measures of statistical dispersion include range and standard deviation.

14

Ratio Data

- The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured.
- The ratio scale represents the three earlier scales rolled up in one:
Like a nominal scale, it provides a name or category for each object (the numbers serve as labels).
Like an ordinal scale, the objects are ordered (in terms of an ordering from low to high).
Like an interval scale, the same difference at two places on the scale has the same meaning.
In addition, the same ratio at two places on the scale also carry the same meaning.
- An example of a ratio scale is the amount of money you have in your pocket (25p, 55p, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have 0 money, this implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50p has twice as much money as someone with 25p.
- All measures of central tendency and dispersion are applicable to ratio data: mean, median, mode (for comparison purposes) as well as range, inter-quartile range, standard deviation.

Arithmetic Mean

- The arithmetic mean (often (lazily) called the “average”) is the most common measure of central tendency
 - For a sample of size n :

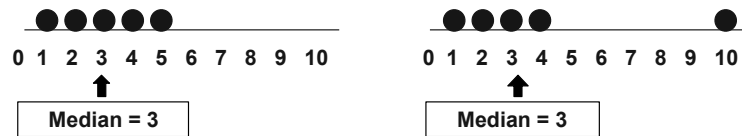
The diagram shows the formula for the arithmetic mean with labels and arrows pointing to specific parts:

- Pronounced x-bar**: Points to the $\bar{X}$ symbol.
- Sample size**: Points to the n in the denominator of the first fraction.
- The i^{th} value**: Points to the x_i in the numerator of the first fraction.
- Observed values**: Points to the $x_1, x_2, \dots, x_n$ in the expanded numerator.

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

The Median

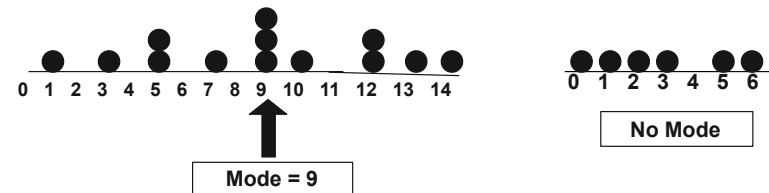
- In an ordered array, the median is the “middle” number (50% above, 50% below).



- It is not affected by extremely large or small values and is therefore a valuable *measure of central tendency* when such values occur in a data sample.
- It can be misleading in non-normal (multi-modal) distributions.

The Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical (nominal) data
- There may may be no mode if data is evenly distributed.
- There may be several modes, if there are peaks of distribution



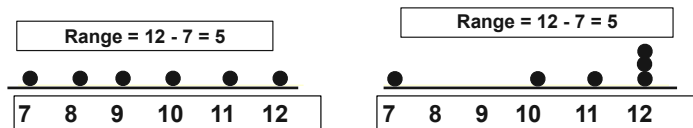
The Range

Range is the simplest measure of variation

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

The Range can be misleading

- Ignores the way in which data are distributed



- Sensitive to outliers

1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 5

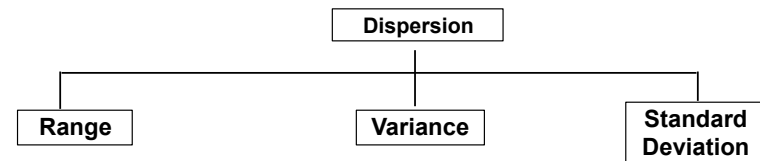
$$\text{Range} = 5 - 1 = 4$$

1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 120

$$\text{Range} = 120 - 1 = 119$$

Measures of Dispersion

- We will consider several measures of dispersion.



- Measures of variation give information on the spread or variability or dispersion of the data values.

The Variance

- Average of squared deviations of values from the mean
 - Sample variance:

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

Where $\bar{X}$ = arithmetic mean
 n = sample size
 X_i = i^{th} value of the variable X

- Population variance (p):

$$p^2 = \frac{\sum_{i=1}^N (X_i - m)^2}{N}$$

Where m = population mean
 N = population size
 X_i = i^{th} value of the variable X

The Standard Deviation (STDEV)

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the variance
- Has the same units as the original data

- Sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

- Population standard deviation:
 where p = STDEV of population and
 m = population mean (arithmetic),
 N = population size.

$$p = \sqrt{\frac{\sum_{i=1}^N (X_i - m)^2}{N}}$$

The Covariance

- The covariance measures the strength of the linear relationship between **two numerical variables** (X & Y).
- The sample covariance:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

- Only concerned with the strength of the relationship.
- No causal effect is implied.

The Coefficient of Correlation

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

where
$$r = \frac{\text{cov}(X, Y)}{S_X S_Y}$$

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

Discrete Probability Distribution

- A probability distribution for a discrete random variable is a mutually exclusive listing of all possible numerical outcomes for that variable and a probability of occurrence associated with each outcome.

Number of purchases made	Probability
2	0.2
3	0.4
4	0.24
5	0.16

Continuous Probability Distributions

- A continuous random variable is a variable that can assume any value on a continuum (can assume an uncountable number of values)
 - thickness of an item.
 - time required to complete a task.
 - temperature of a solution.
 - height, in centimeters.
- These can potentially take on any value depending only on the ability to measure precisely and accurately.

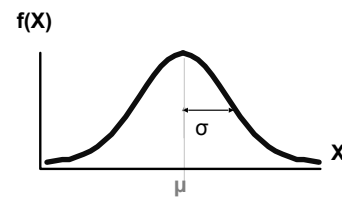
The Normal Distribution

- 'Bell Shaped'
- Symmetrical
- Mean, Median and Mode are equal
- Location is determined by the mean, μ

Spread is determined by the standard deviation, σ

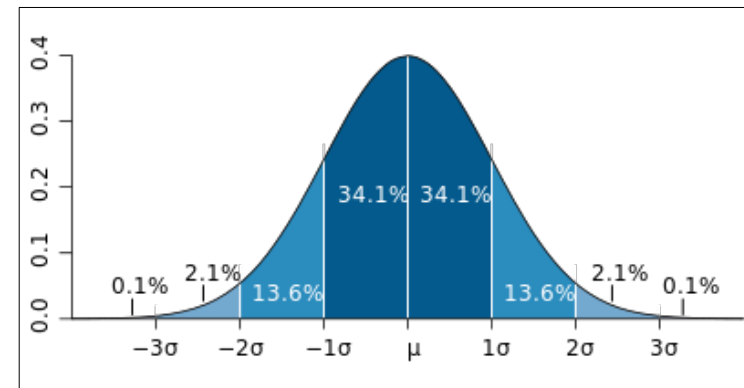
The random variable has an infinite theoretical range:

$+\infty$ to $-\infty$



μ = Mean
= Median
= Mode

Properties of the Normal (Gaussian) Distribution



C. F. Gauss (1777-1855) proved the very useful theorem that, if a sample distribution is 'normal', then

•68.2% of the sample lie within 1 SD (1σ) of the mean

•95.4% of the sample lie within 2 SD (2σ) of the mean

•99.6% of the sample lie within 3 SD (3σ) of the mean