STATISTICS - EXAM NOTES

Weekly Notes
Preview of Week One

Week One – Introduction to Stats and SPSS

Branches of Statistics

- **Descriptive Statistics:** Summarise and describe numbers
  - Describe characteristics of a group
  - Gives summary information that uses numbers to reflect characteristics of the group (describes the sample)
- **Inferential Statistics:** Used to estimate a number that is unknown in a population
  - Used as representative sample of population to make statements about the broader population from which that sample is drawn
  - Draw conclusions and inferences to the population

Types of Data

- **Data:** Variables that have been organised for analysis
  - Variables are the columns – observations are the rows
- **Variable:** Something that varies and is measured
  - **1. Categorical -** Arbitrary values/labels represent categories (no inherent numerical meaning)
    - e.g. Gender – can have numbers assigned but arbitrary
    - e.g. Types of Crime - 1. Robbery, 2. Larceny, 3. Burglary (can switch numbers wouldn't matter)
  - **2. Discrete –** Can only have certain values within a range – whole numbers
    - e.g. number of crimes reported 218, number of symptoms, number of goals
  - **3. Continuous –** Can take on any value (usually fractional)
    - e.g. GPA
    - e.g. Last years crime rate 8.23456 per 10,000. Temp, speed, intelligence
- **Dichotomising** - (dividing into two categories) of continuous and discrete variables is quite common as it enables us to find out if there are differences between groups who may be at the extremes of the continuous or discrete variables
  - Loose effect power, spurious variables, loss of information

Levels of Measurement – NOIR

- **Qualitative Measures** (numbers assigned but no real meaning)
  - **Nominal –** Categories associated with the variable are **different**
    - E.g. Y/N, gender, cities etc.
    - No numeric value assigned to them/no order
    - Simplest level of data/lowest level of measurement

    - **Ordinal –** Categories associated with variables are qualitatively **'different'** AND **'rankable'**
        - E.g. Likert Scale (strongly agree, agree..), letter grades
        - Order but no equal numeric distance between the categories – no ranked order
- **Quantitative Measures** (Truly numeric values)
    - **Interval –** Categories associated with the variable are "**different**" AND the categories are "**rankable**" AND the intervals are **"equidistant**"
        - No absolute zero
        - Equal distance across the gaps used in this scale
    - **Ratio –** Categories associated with the variable are "**different**" AND the categories are "**rankable"** AND the intervals are "**equidistant**" AND there is a "**true zero**"
        - Zero value can be observed

Differences vs. Relationships

- **Differences between two or more groups – Example:** Do male (independent) college students experience greater college student victimization (dependant) than do female college students?
- **Whether two variables are related** – **Example**: Is there a relationship between risky behaviours and an increased likelihood of college student victimization? (looking for correlation)

Types of Variables – Dependant and Independent

- **Variables –** assigned an actual value and more concrete then a concept (crime = concept, assaults/thefts = corresponding variables)
    - **Extraneous variables** – factors that the researcher might not have accounted for but which may have influenced study
    - **Confounding variable** – A specific type of extraneous variable is one that is correlated with both of the main variables that we are interested in
- **Independent Variables (IV) –** value changes independently of a change in another variable.
    - It is controlled or manipulated by the researcher
    - CAUSE (predictor/antecedent)
    - Considered the cause in cause-effect relationship (causes the outcome (DV))
- **Dependant Variables (DV) –** value changes as a result of the change in another variable
    - Not controlled or manipulated by the researcher
    - OUTCOME
    - Effect in cause-effect relationship (outcome/effect dependant on the IV change)

Types of Research Design

- **Experimental –** Can infer causation if researcher **manipulates IV** and true experiment involves the **random allocation** to minimise confounding variables (e.g. ensure outcome is not caused by other factors)
    - Journal of experimental criminology..
    - Statistical Tests – T-tests, ANOVA, Mann-Whitney U Test
- **Quasi-experimental –** Compare outcomes for an IV but **IV not manipulated** or participants have **not been randomly allocated**.
    - Can infer causation but harder to infer
    - Statistical Tests – T-tests, ANOVA, Mann-Whitney U Test, Wilcoxon
- **Correlational –** Tests relationships but **cannot infer causation**
    - Correlation doesn't equal causation
    - Statistical Tests – Pearson's and Spearman's Rho

Between Participants and Within Participants Designs

- **Between –** independent/unrelated design has different groups of participants in each condition of the IV
    - Less likely to get bored, tired, frustrated and more likely to perform at optimum level
    - Less susceptible to practice/fatigue effects and less likely to work out the rationale for the study
    - Reduce order and demand effects and eliminate these factors as extraneous variables from the study
    - Negative – need more participants and loose certain degree of control over inter-participant confounding variables (people bring different characteristics)
- **Within –** repeated measures or related designs
    - Greater control of inter-individual confounding variables
    - Fewer participants needed
    - Increased likelihood of order effects – fatigue, practice, boredom
        - Can counter balance by halving participants and going in different order
    - More likely to realise the experiment
    - Can not really be used in Quasi-experimental designs

Week Two – Measures of Central Tendency

Samples and Populations

- **Population –** entire set of events or group of people
    - Computed summary values are **parameters**
- **Sample –** smaller selection of events or people from a population (must be as representative of target population as possible)
    - Computed summary values are **statistics**
- **Inferences –** Inferential statistics draw conclusions about parameters based on sample statistics about populations
    - Estimate/make guess about what the population parameter truly is)

Sampling Error

- Difference between population parameter and the sample statistic

Reducing Sampling Error

- Need random selection of participants
- Equality of the draw to ensure randomness
- Independence of the draw – selecting one element will not affect the sampling of another
- Random Number Tables – help random selection (now computers select random sample)

Measures of Central Tendency – How we describe the typical case

- **Mean –** Average (total of all values/number of values)
    - **Needs interval or ratio data**
        - Doesn't have extreme scores
    - **Advantages:** Can be manipulated algebraically and stable across samples (better estimate of population mean)
    - **Disadvantages:** Influenced by extreme score (**outliers**), value may not actually exist (interval/ratio data can but may not exist in population),
- **Median –** Midpoint – has 50% of ranked data fall above and below
    - **Need at least ordinal data**
        - Use if extreme scores
    - Rank order values
    - Median position (e.g. 6^{th} position) or median value (actual number)
    - **Advantages:** Unaffected by extreme scores and skewed distributions, does not require assumptions about interval properties of the scale
    - **Disadvantages:** does not readily enter equations and not stable from sample to sample
- **Mode –** Most common/frequently observed value
    - **Nominal/Categorical data**
    - Can be bi-modal (two appear the most)
    - **Advantages:** Must occur, represents largest number of scores, highest probability of being chosen, applicable to categorical data
    - **Disadvantages:** Changes with different data grouping, may not be particularly representative

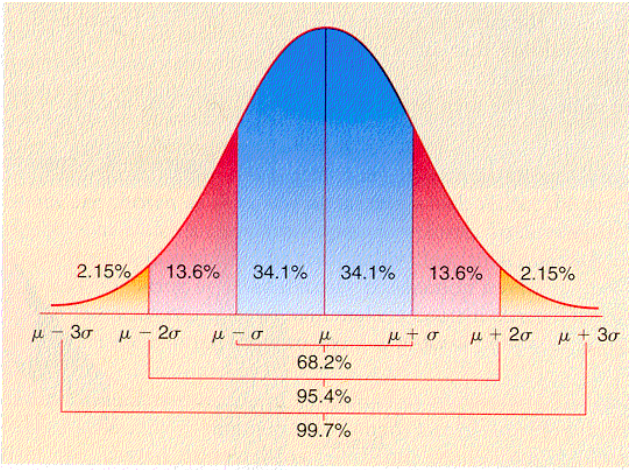
Displaying Data

- **Exploratory Data Analysis (EDA) –** generates insight about data and under covers underlying structure of data
    - Identified important variables (which to use in research) and detects outliers and abnormalities
    - Tests underlying assumptions
    - EDA is an approach or style not prescriptive
- **Frequency Distribution –**
    - **Tables:** number of individuals in each category/percentage
    - **Histogram:** Continuous data (rectangles used to represent frequencies)
        - Can identify problems in data (e.g., error inputting data)
        - Can quickly identify mode
        - Can see distribution of scores

- **Boxplot –** helps to clearly identify extreme scores
    - Visual display of distribution that combines the 25%, 50% (median) and 75% percentile scores and the range (min/max)
        - Inter-quartile range
    - Elongated if wide range numbers
    - Outliers can also be displayed
    - Limited by type of data can be used - ordinal level data or above but not categorical or nominal level data
- **Pie Charts -** Start at 12 and go around starting with the biggest
    - The single valuable feature of a pie chart is that it shows how the parts contribute to the whole
        - Don't use 3D - this is lost when a 3-D effect is used.
- **When to display data**
    - If it aids in interpretation of the data – keep it simple and label x-axis and y-axis appropriately and concise and accurate title

The Normal (Curve) Distribution - Distributions tell us how our data look

- Statistical tests make certain assumptions about distributions – so need to look at distributions to make sure they meet the assumptions of statistical tests
- **Normal Curve Characteristics –**
    - Symmetrical (no skewness)
        - Sides are mirror images of one another
    - Tails never touch the x-axis
        - Ends go on to infinity
    - Can be kurtosis
    - Uni-modal (only one peak)
    - Mean, mode and median – same value under the normal curve
        - Located at the centre
    - Bell shaped (function of variability)



- **Skewness –** Concerned with symmetry of distribution
    - Symmetrical distribution very rare in the real world
    - Skewed distributions have one side of the distribution that is different to the other
    - **Positive skew:** the tail of the graph points towards the positive end of the scale
        - **Extended tail to the right**
    - **Negative skew:** the tail of the graph points towards the negative end of the scale