BUSINESS STATISTICS STAT150

Statistics Introduction

- Statistics is the science of learning from data
- Involves collection, presenting summaries, analyzing and interpreting data

Objective and Scope of Statistics

- Aim: obtain information about a target population (all subjects relevant) using a sample (measurable subset)

Study Design

- Formulate question
- Specify target population
- Variables (measurements collected)
- Define method of data collection
- Ideally observations should be independent of each other

Why take a sample: understand population, Representative, Census

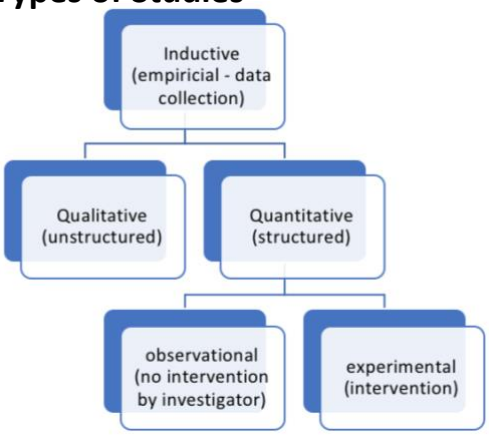
Selecting a Sample

- Representative – make inferences about target population, unbiased, large
- Random sample: each member of target population has an equal chance of being chosen

The Data We Collect: Variables

- Measurements are taken according to variables of interest
- Types of variables: Predictors (determinant), Influence, Outcomes

Types of Studies



Bias: Sample Size

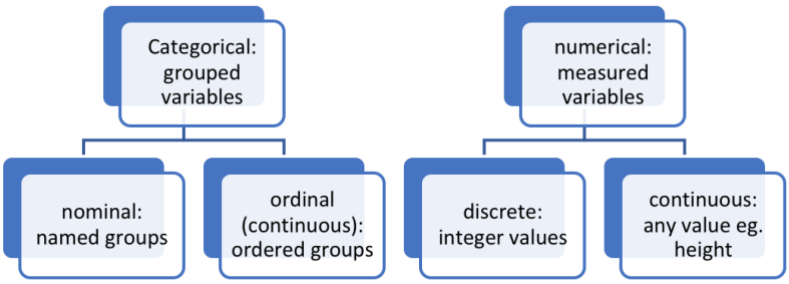
- Bias: any systematic error
- Types of bias
  - Selection bias: way subjects selected
  - measurement bias: measurement of variables
  - response bias: response rate too low (>75%), those who respond have different characteristic (often strongest opinions)
  - confounding: distorting of the apparent effect of one variable (predictor) on another (outcome). Another explanation for correlation

Accuracy

- sample size (n)
- variability (spread): more variable = less accurate

Data

Data Classification



DATA	categorical	numerical	
categorical	clustered bar charts	comparative box plots	bar charts or pie charts
numerical	comparative box plots	scatter plots	histograms
	bar charts or pie charts	histograms	

Categorical Data

Data Included: count, proportion, percentage, mode

→ ONE CATEGORICAL VARIABLE

- Frequency table: variable name, category, count/proportion, percentage of observations, total sample size
- Bar Charts/Pie Charts

→ TWO CATEGORICAL VARIABLES

- Contingency table: whether one variable is contingent (associated) with another.
- Clustered Bar charts: easy to make comparisons

Numerical Data: Describing distributions

(1) shape: symmetry, mode, outliers (left tail = skewed left)

(2) centre: median, mean

a. median is more robust (less influenced by outliers and best to use with skewed distributions)

(3) spread/variation: range, inter-quartile range (most robust), standard deviation

Inter- Quartile Range

- upper quartile – lower quartile.
- In excel: Quartile.exc(range, quartile number)

the p^{th} percentile is the value such that p% of the values in the distribution are lower than it

Variance

- Variance: work out mean. For each number subtract mean and square the difference. Average differences.

Standard Deviation

- excel: Stdev.S(range)

Numerical Data Graphs and Summaries

- Histogram: visual representation
- Box plot: sort data into quarters
  - Fences determine whether data has any outliers.
    - (1) IQR = UQ - LQ
    - LQ – (1.5)(IQR) = low fence
    - UQ – (1.5)(IQR) = high fence
  - Comparative Box Plots: several samples simultaneous, concise graph

- Scatter Plots: determine relation between two numerical variables
  - X axis: predictor, independent
  - Y axis: response, outcome, dependent
  - Positive/negative. Linear/Non-Linear. Strong/Weak (points close to line of best fit = strong).
- Table of Descriptive Statistics

Distributions

Population Distributions

- Small samples may not "fit" as well as larger samples
- Histograms of large samples should represent parent population distribution
- Population distribution: curve off histogram
- Sample statistics are used to estimate population parameters

Discrete and Continuous Random Variables (X)

- Probability distribution: set of all outcomes and associated probability
- Uniform distribution: same probability of occurring

Discrete Distributions: The Binomial Distribution

- Probability of success of an event = p
- Probability of failure = q = 1-p
- The distribution of successes, x, will be symmetric and bell-shaped when: $np \geq 10$ and $nq \geq 10$ (success and failure condition)
- When n is small or p is close to 0 or 1, the binomial distribution is skewed
- When n is large or p is close to 0.5, the binomial distribution is symmetric and bell shaped

Continuous Distributions: the Normal distribution

- Symmetric "bell shaped", unlimited range
- Mean and standard deviation

Areas Under a Normal Curve

- Standard deviation 1 68%
- Standard deviation 2 95%
- Standard deviation 3 99.7%

The Standard Normal Distribution

- Normal distribution can be converted to a standard normal (z)
- Mean = 0, sd = 1
- $z = \frac{y-u}{\sigma}$
- z value indicates how many standard deviations a y value is away from the mean
- can be positive and negative

z-scores and probability

- $z = \frac{observation(y) - mean}{sd}$
- rearrange to find y = u + sd(z)
- area under the curve = probability = proportion of all values greater than y
- Probability formula in excel: =Norm.Dist(observation, mean, standard deviation, true)
  - If trying to find area above line do 1-norm.dist(..)
- The p^{th} percentile is the value such that p% of the values in the distribution are lower than it.
  - Eg. looking for 10^{th} percentile means probability = 0.1 (norm.dist = 1)
  - Y = Norm.Inv(probability,mean,sd)