1) Ordinal scale: classifies values into distinct categories in which ranking is implied (adv: poss more of a property, scale ordered/weak: does not account for the amount of the difference between categories) – numbers allow ranking but no distance. (differences between rankings are not equal) Eg. Preferences included for each scenario/rank from lowest to highest level.

- Numerical variables (quantitative): have values that represent actual number quantities and have meaning as measurement. (identified as either discrete or continuous variables)
    1) Interval scale: ordered scale in which the difference between consecutive measurement is a meaningful quantity – distances are equal (no true zero point: does not mean absence of phenomenon) – measure the magnitude of the differences. Eg. Temperature, specific time
    2) Ratio scale: ordered scale in which the difference between the measurements involves a true zero-point (represents the absence of phenomenon/fixed point) – provide proportion of difference (exact figures on objective) Eg. Amounts of time
    - Discrete variables: have numerical values that arises from a counting process (representing the countably infinite case [true zero point]/variable that assumes a finite number of exact number.) Eg. Number of children: 13, defects per hour: 32.
    - Continuous variable: produce numerical response that arises from a measuring process. (possible values cannot be counted and can only be described using intervals on the real number line from range) Eg. Cost: 108.56, financial return: 23.1

DCOVA: 5 steps for business task

1. **Define** the variables that you want to study in order to solve a business problem or meet business objective.
2. **Collect** the data from appropriate sources. (represent sample properly)
3. **Organize** the data collected by developing tables.
4. **Visualize** the data by developing charts. (graphic – understand)
5. **Analyse** the data by examining the appropriate charts and tables to reach conclusions.

Sampling

Smaller research group is used to represent the wider data set (cost is lower and more quickly)

- Sample must be representative of the entire population in order to create a result reflective in the population.
- Sampling must be done partially, impartially and objectively.
- Only trustable sampling scheme is random sampling (individuals are selected
- totally at random)

Collecting data

- Primary data source: collect your own data for analysis.
- Secondary data source: your analysis has been collected by someone else. Eg. Survey, organization
    - ✓ Organisation: secondary sources that obtained from primary sources. Eg. financial data, industry/market data, weather conditions…
    - ✓ Designed experiments: design of any task that aims to describe or explain the variation of information under conditions that are hypothesized to reflect the variation. It is subject because often involves sophisticated statistical procedures. (possible effect of a treatment on subjects – assigned to a control group) Eg. Consumer testing, quality testing, market testing.
    - ✓ Survey (draw sample): Ask questions about people's beliefs, attitudes, behaviours and other characteristics. Eg. Satisfaction about product, political poll
    - ✓ Observational studies: collecting data directly observing a behaviour. (measuring) Eg. Measure volume of the traffic, time take for customers to be served…
    - ✓ Automated and streaming data: Big data -> volume, velocity, variety, veracity (Eg. Mobile phone, data usage, GPS data)
    - ✓ Cleaning data: Cleaning process remove errors, flag strange…

SAMPLING AND DISTRIBUTION DATA

Organizing data

Categorical data

*Discrete random variable: the set of all possible outcomes is finite (integer value, no gap between values – counting process) represented by histogram. Note: gap between &2.01 and $2.02 is still finite and not differentiable => discrete

$$\boxed{f(x) = \lambda e^{-\lambda x} \quad ; \quad x > 0}$$

Pr(x < X) = EXPONDIST(X value, mean, true)

SAMPLING DISTRIBUTION

+ Usefulness of sampling: less time consuming & less costly & more practical (however, inaccurate or biased results can result if parts of the population are excluded)

Nonprobability sampling (select without knowing the probabilities of selection)

+ Convenience sample: selected based only on being easy, inexpensive, quick to sample (Adv: accelerated data collection & ready availability & cost effectiveness/Dis: potential bias due to unrepresented of interest & insufficient power to identify differences)

+ Judgement sample: perceived experts or most appropriate items are selected by convenience [Adv: convenience, low cost, speed/Dis: lack of accuracy due to selection bias (subject to some degree of bias due to not identical frame and population & not a form of random selection and have inherit bias)]

+ Quota sample (similar to stratified sampling): pre-set quotas of groups chosen by convenience: segmented into mutually exclusive sub-groups -> select based on specific proportion

Non-probability sample

> + Sampling method where some elements have no chance of selection because the selection is based on assumptions regarding the population of interest. (Dis: not accurately determined & non-random, nonprobability does not allow estimations of sampling errors that give rise to exclusive bias and place limits)
>
> + Include convenience sampling, quota sampling and purposive sampling

Probability sample

Sample in which every unit of population has a chance of being selected (probability can be accurately determined – combinations produce unbiased estimates)

+ Simple random sample: each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. (Adv: simple, cheap to use/Dis: vulnerable to sampling error because randomness does not reflect the makeup of the population & also be cumbersome and tedious) Excel: =RANDBETWEEN (x, x)

+ Systematic sampling (interval sampling): relies on arranging the study population according to some order scheme and select at regular interval through its ordered list (Adv: ensure the spread out of sample randomness & easy to implement and stratification induced create efficiency & effective against many types of bias/Dis: vulnerable to periodicities in the list which make the scheme less accurate – unrepresentative of overall population if periodicity & sampling errors due to variation make it difficult to quantify that accuracy) k = N/n

+ Stratified sampling: frame can be organized by distinct categories into separate strata according to important characteristics (enable drawing inferences about specific subgroups – use the best approach suited) which lead to more efficient statistical measurement. -sample size proportional to size of each strata (Adv: focus more on important representation population & improve efficiency and accuracy against bias & greater balancing of statistical power of differences between data/Dis: increase cost and complexity of sample selection (reduce utility of strata) & requires the selection of relevant stratification variables which can be difficult & expensive to implement)

+ Cluster sampling: sampling is often clustered by geography or time periods which is more cost-effective and increase the variability of sample estimates - population is divided into several clusters, each representative of the population (cluster level frame: multistage sampling – two of more levels of units are embedded one in the other and could substantially reduce sampling costs) Clusters are naturally occurring designations and should be mutually exclusive and collectively exhaustive

Types of survey errors

- Coverage error or selection bias (excluded from frame) – certain groups of items are excluded from the frame (selection bias & inability to obtain information about all sample cases) gaps between the sampling frame and the total population lead to biased results and can affect the variance of results. Coverage error is a kind of non-sampling error
- Non-response error or selection bias – failure to collect data on all items in the sample and result in a non-response bias some respondents included in the sample do not respond - key difference here is that the error comes from an absence of respondents instead of the collection of erroneous data.