Fundamental of Statistics:

Two areas of statistics:

1. Descriptive Statistics (DS): concisely summarise characteristics of data, using both graphical and numerical techniques
2. Statistical inferences (SI): process of making an estimate, forecast, or decision about a population based on a sample drawn from that population
    - ⇒ Constructing estimates of the characteristics of data
    - ⇒ Testing hypothesis about the world (population), based on data

- Modelling and analysis combines descriptive and inferential statistics, to build models that represent relationships and trends we are interested in, in a systematic way

Population vs. Samples:

- Population: full set of all items of interest to us, characteristic of population is called a parameter
- Sample: subset of data drawn from a population, characteristic of a sample is called a statistic

Variable Types:

- Variable: a characteristic of a population or sample
- Three main types of variables:
    1. Numerical: quantitative, "interval"; e.g. income, age, height
    2. Nominal: categorical, quantitative, discrete; e.g. gender, marital status, nationality
    3. Ordinal: ranked; e.g. excellent ⇒ good ⇒ fair ⇒ poor ⇒ very poor

Data Types:

- Data: observed values of a variable
- Cross – section data: many observations at one period in time
- Time Series data: measurements at successive points in time
- Panel data: cross – section data over serval time periods, on the same members

Biased and Unbiased Samples:

- Unbiased sample: a genuine representation of the population, yielding sample statistics that are close to the relevant population parameters
- Some bias situations include:
    - ⇒ Error in data acquisition: incorrect responses are recoded perhaps due to misinterpretation of questions or resistance to truth revelation
    - ⇒ Non-response bias: do not get responses from all members of a selected sample
    - ⇒ Selection bias: the selection methods means that some subjects are more likely to be surveyed than others

Sampling Plans:

- Simple Random Sampling: all population members have an equal chance of being sampled
    - ⇒ Sample Size involves trade-off between accuracy and cost; sample size increases:
        - The cost increase

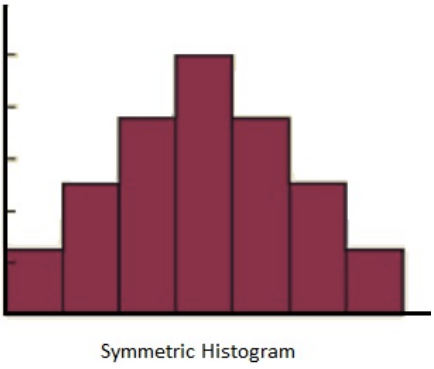
        - The sample statistics become closer to the population parameters
- Stratified Random Sampling: population is split into mutually exclusive groups, from which random samples are drawn
- Cluster Sampling: involves choosing a number of clusters (of groups) at random

Why sample?

- Analysing population data to determine a specific population parameter with 100% certainty may be impractical or impossible
- Collecting and analysing a small sample of observations taken from our population of interest is faster and cheaper
- Use sample statistic as an estimate of the population parameter of interest, but cannot be 100% confident our estimate is correct
- A good sample should be unbiased and representative, with a small sampling error
    - $\Rightarrow$ Sampling Error: the difference between our estimate based on our sample (statistic) and the true population parameter

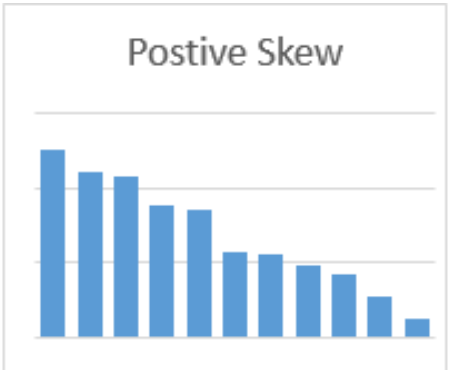
Descriptive Statistics:

Shapes of Histograms:

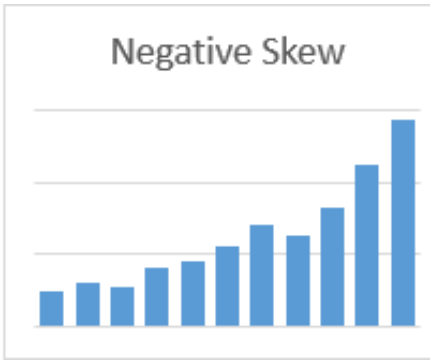


Symmetric Histogram

- Mean ≈ Median ≈



- Mean > Median >



- Mean < Median < Mode

- Number of modal classes: a modal class is the one with the largest number of observations (a peak)
  - ⇒ Unimodal Histogram: one peak
  - ⇒ Bimodal Histogram: two peaks

Numerical Descriptive Methods:

- Population: parameters, denote in Greek letters
- Sample: statistics, denote in Normal letters

1. Measure of Central Location: calculate where the centre of a set of data is
   - ⇒ Mean:
     - Sample Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
     - Population Mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
     - The $\bar{x}$ is an estimator of the μ
     - Sampling Error = $\bar{x} - \mu$
     - As sample size n increases, the sample becomes more representative of the population and $\bar{x}$ will converge to μ
     - Easy to calculate, often used for making inferences, but sensitive to extreme observations
   - ⇒ Median:
     - Middle value in ordered data, if n is odd
     - Average of the two-middle value, if n is even
     - Nice interpretation and not sensitive to extreme observations, so sometimes used when the mean might provide a distorted measure of central location
   - ⇒ Mode:
     - The most frequently occurring value in the set of data
     - Used for less often, but position of mode relative to mean and median may tell us something about the shape of the distribution
   - ⇒ The normal distribution is bell-shaped, and many statistical techniques require that the population be approximately normally distributed
2. Measure of Dispersion

⇒ Population Variance: average of all the squared deviations between each observation and the population mean

- $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- Standard Deviation: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

⇒ Sample Variance:

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Standard Deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

⇒ An important difference between s^2 and σ^2:

- To calculate σ^2, divisor is population size N
- To calculate s^2, divisor is sample size less 1, $n - 1$

⇒ The difference is important for ensuring that our sample statistic s^2 is an unbiased estimator of the population parameter σ^2

- Coefficients of Variation: depends on the magnitude of observations in the data to be whether big/small
  - Population coefficient of variation: $CV = \frac{\sigma}{\mu}$
  - Sample coefficient of variation: $cv = \frac{s}{\bar{x}}$
- Mean Absolute Deviations (MADs):
  - Population MAD = $\frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$
  - Sample MAD = $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$

Amenable to Statistical inference

- Range: Largest Value (Max.) – Smallest Value (Min.)

3. Measure of Relative Standing:

⇒ Percentile: The p-th percentile is the value for which at most p% of observations lie below that value, and at most (100-p)% of observations lie above it

- Location of any percentile p: $L_p = (n+1)\frac{p}{100}$
- First (lower) decile = 10th percentile
  First (lower) quartile (Q_1) = 25th percentile
  Second (middle) quartile (Q_2) = 50th percentile, median value
  Third (upper) quartile (Q_3) = 75th percentile
  Ninth (upper) decile = 90th percentile

⇒ Inter-Quartile Range (IQR): $IQR = Q_3 - Q_1$

Nominal Data:

- Nominal Data denote qualitative or categorical characteristics of individuals, firms, etc
- Present in both Graphical and Tabular techniques:

1. Pie and Bar Charts:

⇒ Pie Chart: size of pie slice reflects proportion (percentage) of total

⇒ Bar Chart: height of bar reflects proportion (percentage) of total, compare two separate breakdowns of nominal data

⇒ Which Chart type is best?

- Pie Chart, if we want to see components of whole entity in a manner that indicates the relative sizes of components
- Bar Chart, if we want to compare size or frequency of various categories