Statistics Descriptive statistics, graphs and statistical studies

Measures of the centre

- Sample mean – average value
- Median – number in the centre
- Mode – the most typical value

Measures of spread

- Range

$$range(x) = max(x) - min(x)$$

- Interquartile range

$$IQR(x) = med(\text{upper half}) - med(\text{lower half}) = Q3 - Q1$$

- Population standard deviation $\sigma = \sqrt{\sigma^2}$

$$\boldsymbol{\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Convenient formula (n is the size of the *population*)

$$\sigma^2 = \mu(x^2) - \big(\mu(x)\big)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

- Sample standard deviation $s = \sqrt{s^2}$

$$\boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Clarification. There is $n - 1$ in the denominator to make s^2 an unbiased estimate of σ^2.

Convenient formula (n is the size of the *sample*)

$$s^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right) = \frac{n}{n-1}\left(\left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right)$$

Relative location of elements

- z-scores

$$z\text{-score of } x = \frac{x - \bar{x}}{s}$$

- Percentiles and quartiles
- Outliers

$$outlier < Q1 - 1.5 \times IQR \quad \text{or} \quad outlier > Q3 + 1.5 \times IQR$$

Shape of the sample

- Symmetry

$$med \approx mean \Rightarrow \text{symmetric}$$

- Right skewed (positively skewed) – the tail is on the right

$$med < mean \Rightarrow \text{right skewed}$$

- Left skewed (negatively skewed) – the tail is on the left

$$med > mean \Rightarrow \text{left skewed}$$

- Number of peaks: unimodal (1 peak), bimodal (2 peaks), etc.
- Clusters – subgroups of the sample
- Gaps – holes between clusters

Types of graphs

- Box-and-whisker plot
- Dot plot

- Stem-and-leaf plot
- Histogram (for quantitative data)
- Bar chart (for categorical data)
- Cumulative frequencies plot (ogive)

Comparison of the samples

- Centre: mean, median
- Spread: range, IQR, standard deviation
- Shape: symmetry/skewness, number of peaks
- Special features: outliers, clusters, gaps

Graphs for comparison of the samples

- Parallel box-and-whisker plots
- Back-to-back stem and leaf plots
- Parallel histograms
- Back-to-back histograms

Statistical studies

- *Observational* studies (without intervention)
- *Experimental* studies (with intervention)

Observational studies

- *Census*: a study of the entire population
  Clarification. All other methods are sample studies.

Non-probability sampling (some population units have zero probability of being selected)

- *Convenience sampling*: only convenient units are selected
- *Judgement sampling*: researchers rely on an expert's opinion in forming sample
- *Quota sampling*: sampling frame is unavailable, i.e. population is not known fully
  Clarification. Quota sampling is a non-random equivalent of the stratified random sampling.

Probability sampling (each population unit has some positive probability of being selected)

- *Simple random sampling* (SRS): each unit has equal positive probability of being selected; is obtained through randomization; provides i.i.d. r.v.
- *Systematic sampling*: k-th and then each n-th unit is selected
- *Stratified sampling* (proportionate or disproportionate stratification): some elements from each stratum are included in the sample
- *One-stage cluster sampling*: all units from some clusters are included in the sample
- *Two-stage cluster sampling*: some units from some clusters are included in the sample
- *Multistage sampling*: techniques can be mixed but not necessarily

Errors in the observational studies

Sampling error (since sample study is conducted instead of census)

Non-sampling error (occurs both in sample study and in census)

- *Selection* bias: *undercoverage* bias, *non-response* bias, *voluntary response* bias
- *Response* bias: *non-anonymous survey*, *non-neutral questions*, *fear of consequences*

Variables distorting conclusions of the study

- *Lurking variables* (affect both dependent and independent variables indicating a false relationship)
- *Confounding variables* (affect dependent variable and makes it unclear which of the independent and confounding variables affects the dependent variable)

Parts of an experiment

- *Experimental units/participants* divided into *treatment groups* and *control groups*
- *Explanatory variables/independent variables/factors*
- *Response variable/dependent variable*
- *Treatment levels* (choices for each factor)
- *Treatments* (combinations of the assigned treatment levels)

Methods to verify causal relationship

- *Control groups*
- *Time order* (*longitudinal* and *panel* surveys)

Problems with experiments and their elimination

- *Placebo effect – blinding* or *double blinding*
- *Confounding factors – randomization* (completely randomized design of the experiment); *blocking*/randomized block design; *matched pairs design*