Descriptive Statistics

Anjali Bhavan

19-12-2020

1 Measures of central tendency

1. Arithmetic mean = $\frac{1}{N}\sum_{i=1}^{n} f_i x_i$ where $N = \sum_{i=1}^{n} f_i$
   Shifting of origin: $x = a + hu$, where a is the shifted coordinate and h is the new scale.
2. Geometric mean
3. Harmonic mean
4. Median = $l + \frac{h}{f}(\frac{N}{2} - C)$
   where l is the lower limit of the median class (class with cumulative frequency just more than $N/2$ (or whatever is the first value in the bracket)), h and f the width and frequency of median class, C is cumulative frequency of pre-median class.
5. Mode = $l + \frac{h(f_m - f_1)}{2f_m - f_1 - f_2}$
   where l is the lower limit of the modal class (class with highest frequency), $f1$ and $f2$ are frequencies of pre and post modal class.
6. Partition values:
   Quartiles: $Q_i = l + \frac{h}{f}(\frac{iN}{4} - C)$
   Reminder: l, h, f and C are all to be figured out based on the first value in the bracket. That has to be used for determining which is median class.

2 Measures of dispersion

1. Range: IQR = $Q3 - Q1$
2. Std deviation: **Has two different formulae for population and sample.** Also, stddev is used rather than variance because square numbers get big, and variance has same unit as mean, hence.
   With shifting of origin: both variance and stddev are unaffected by origin shift, but are affected by scale shift. Variance becomes h^2 times and stddev becomes h times.
3. Karl Pearson's coefficient of variation: used to measure variability in data (mostly for comparison). Given by $\frac{\sigma}{\bar{x}} * 100$
4. coefficient of dispersion based on quartile deviation: $\frac{Q3-Q1}{Q3+Q1}$

Definition 1 (Skewness)**.** *Skewness is the measure of lack of symmetry in the distribution. Positively skewed means the longer tail of the distribution is towards the right, negatively skewed means vice-versa. In positive skew,* ***mode*** $<$ ***median*** $<$ ***mean****. In negative,* ***mean*** $<$ ***median*** $<$ ***mode****.*

Definition 2 (Kurtosis)**.** *Kurtosis is the measure of lack of peakedness of the distribution. Leptokurtic distribution means more peaked than normal, mesokurtic means normal symmetric curve, platykurtic means flatter than normal curve. The value of β_2 (described below) gives an estimation of kurtosis.*

3 Skewness, Kurtosis, Moments

1. rth moment of a distribution: given by

$$\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a)^r \tag{1}$$

   First moment is mean, second variance, third skewness, fourth kurtosis.

2. Pearson's coefficients:
   $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\gamma_1 = \sqrt{\beta_1}$
   $\beta_2 = \frac{\mu_4}{\mu_2^2}$, $\gamma_2 = \beta_2 - 3$

3. Karl Pearson's coefficient of skewness: $\frac{mean-mode}{\sigma}$

4. Bowley's coefficient of skewness: $\frac{Q3+Q1-2Q2}{Q3-Q1}$

Theorem 1. *Chebyshev's Inequality: Given a number $k \geq 1$ and a dataset of n observations, atleast $1 - \frac{1}{k^2}$ of the observations will lie within k standard devs of the mean.*