Basic Statistics

Sudipta Das

Assistant Professor,
Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

1. Data Summary Measures
   - Central Tendency
   - Measures of Dispersion
   - Skewness
   - Kurtosis

Chapter 4: Data Summary Measures

Introduction

Summary Measures

- There are several types of summary measures which are useful for describing different aspects of the data (more specifically, different aspects of the distribution of data over the possible values).

Central Tendency I

- The first type is called measures of central tendency, which represent a central value for the collection of observations in the sample.
- Common measures of central tendency
  1. Mean,
     1. Arithmetic,
     2. Geometric,
     3. Harmonic
  2. Median
     - Quartile, Decile, Percentile, Quantile
  3. Mode

(Arithmetic) Mean is the simple average of all the observations

- For un-grouped or simple series data: n observations $(X_1, X_2, \ldots, X_n)$
    - Mean is computed by dividing the sum of all the observations by the total number of observations.
        - So, the mean is $\frac{x_1+x_2+\cdots+x_n}{n} (= \frac{1}{n}\sum_{i=1}^{n} x_i)$, denoted by $\bar{x}$.

Central Tendency III

- For grouped data
  - Discrete variable:

Values(x)	x_1	x_2	$\ldots$	x_i	$\ldots$	x_k	Total
Freq(f)	f_1	f_2	$\ldots$	f_i	$\ldots$	f_k	N

  - Continuous variable:

Class Interval	$x_1' - x_1''$	$x_2' - x_2''$		$x_k' - x_k''$	Total
Freq(f)	f_1	f_2	$\ldots$	f_k	N
Class Marks(x)	x_1	x_2	$\ldots$	x_k	-

,

where $x_i = \frac{x_i' + x_i''}{2}$.

  - Mean is $\bar{x} = \frac{1}{N} \sum_{i=1}^{k} x_i f_i$

Central Tendency IV

- Some properties of A.M.
  1. Sum of deviations from AM is zero
  2. If the variable X has mean $\bar{x}$ then the variable $Y = \frac{X-a}{b}$ has mean $\bar{y} = \frac{\bar{x}-a}{b}$
  3. If two groups of observations $\{x_{1,1}, \ldots, x_{1,n_1}\}$ and $\{x_{2,1}, \ldots, x_{2,n_2}\}$ have AMs $\bar{x}_1$ and $\bar{x}_2$, respectively then
     - the combined mean of all the observations is
     $$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}.$$
     - if $\bar{x}_1 < \bar{x}_2$,
     $$\bar{x}_1 < \bar{\bar{x}} < \bar{x}_2$$

Central Tendency V

Geometric and Harmonic mean for *only* positive observations

- Geometric mean

$$\bar{x}_g = \begin{cases} \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} & \text{, ungrouped} \\ \left(\prod_{i=1}^{k} x_i^{f_i}\right)^{\frac{1}{N}} & \text{, grouped,} \end{cases}$$

- Harmonic mean

$$\bar{x}_h = \begin{cases} \dfrac{n}{\sum_{i=1}^{n} \dfrac{1}{x_i}} & \text{, ungrouped} \\ \dfrac{N}{\sum_{i=1}^{k} \dfrac{f_i}{x_i}} & \text{, grouped,} \end{cases}$$

Central Tendency VI

- Some relations between A.M., G.M and H.M.
  1. The logarithm of the G.M. is equal to the A.M. of the logarithmic values.
  2. The reciprocal of the H.M is the A.M of the reciprocal values
  3. For any set of strictly positive real numbers, $A.M \geq G.M. \geq H.M.$
     - Equality hold when all numbers are equal

Central Tendency VII

Median is the middle-most value among the ordered observations

- To compute median for series data
  - Arrange the observations $x_1, x_2, \ldots, x_n$ from the smallest to the largest.
  - Then, if n is odd, the middle value (value of the $((n+1)/2)^{th}$ observational unit) is the median.
  - If n is even, the average of the two middlemost values (average of the values of $(n/2)^{th}$and $(n/2+1)^{th}$ observational units) is the median.

Central Tendency VIII

- To compute median for grouped data:- discrete variable

Values(x)	x_1	x_2	$\ldots$	x_i	$\ldots$	x_k	Total
Freq(f)	f_1	f_2	$\ldots$	f_i	$\ldots$	f_k	N
Cum Freq($\leq$)	f'_1	f'_2	$\ldots$	f'_i	$\ldots$	f'_k	-

  - Find the *first* class, whose cumulative frequency is greater than equal to $N/2$
    - If it is the i^{th} class then $\frac{N}{2} \leq f'_i$ and $f'_{i-1} < \frac{N}{2}$
  - Then the median is $\tilde{x} = x_i$.

Central Tendency IX

- To compute median for grouped data:- continuous variable

Class Interval	$x_1'-x_1''$	$x_2'-x_2''$		$x_i'-x_i'$		$x_k'-x_k''$	Total
Freq(f)	f_1	f_2	$\ldots$	f_i	$\ldots$	f_k	N
Cum Freq($\leq$)	f_1'	f_2'	$\ldots$	f_i'	$\ldots$	f_k'	-

  - Find the median class, i.e., *first* class, whose cumulative frequency is greater than equal to $N/2$
    - If it is the i^{th} class then $\frac{N}{2} \leq f_i'$ and $f_{i-1}' < \frac{N}{2}$
  - Then the median is

$$\tilde{x} = x_i' + \frac{\frac{N}{2} - f_{i-1}'}{f_i} \times c.$$

A more general measure is Quantile

- The p^{th} quantile $(0 < p < 1)$, of a frequency distribution Z_p is a value which divides the distribution in the ratio $p : (1 - p)$

Central Tendency XI

- Different choices for p
  - $p = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$
    - $Z_{i/4} = Q_i$ is called i^{th} **quartile**
    - $Q_1 < Q_2 < Q_3$
  - $p = \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \ldots, \frac{9}{10}$
    - $Z_{i/10} = d_i$ is called i^{th} **decile**
    - $d_1 < d_2 < d_3 < \ldots < d_9$
  - $p = \frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \ldots, \frac{99}{100}$
    - $Z_{i/100} = p_i$ is called i^{th} **percentile**
    - $p_1 < p_2 < p_3 < \ldots < p_{99}$

Central Tendency XII

- Common examples of quartiles, percentiles, deciles, etc..
    - In view of the brief discussion on percentiles before, the first quartile, denoted by Q_1, is the 25th percentile.
    - Similarly, the second quartile, denoted by Q_2, is the 50th percentile which is also the median.
    - The third quartile Q_3 is the 75th percentile.
    - The deciles are similarly defined as the 10th percentile, 20th percentile and so on.
    - In particular, the 5th decile is the 50th percentile, or the second quartile, or the median.

Central Tendency XIII

- To compute p^{th} quantile for series data
  - Consider the proportion and cumulative proportion for the ordered observations

Ordered Obs	$x_{(1)}$	$x_{(2)}$		$x_{(i)}$		$x_{(n)}$
Freq(f)	$1/n$	$1/n$	$\dots$	$1/n$	$\dots$	$1/n$
Cum Freq($\leq$) (proportion set)	$1/n$	$2/n$	$\dots$	i/n	$\dots$	1

  - Look for the target proportion p in the proportion set.
    - if p is exactly one of the values in proportion set: $z_p =$("that value"+"next value")/2
    - if p is crossed over the proportion set for the first time: $z_p =$"that value"

Central Tendency XIV

- To compute p^{th} quantile for grouped data
  - Find the class containing z_p , i.e., *first* class, whose cumulative frequency is greater than equal to Np
    - If it is the i^{th} class then $Np \leq f'_i$ and $f'_{i-1} < Np$
  - Then

$$z_p = x'_i + \frac{Np - f'_{i-1}}{f_i} \times c.$$

Central Tendency XV

Mode is the value that occurs with highest frequency in the data

- Simple series: No mode
- Discrete grouped data

Values(x)	x_1	x_2	$\dots$	x_i	$\dots$	x_k	Total
Freq(f)	f_1	f_2	$\dots$	f_i	$\dots$	f_k	N

- If f_i is the unique maximum of $(f_1, f_2, \dots, f_k)$, then the mode is

$$\check{x} = x_i$$

- Note: There can be more than one mode.

Central Tendency XVI

- Continuous grouped data

Values(x)	$x_1'-x_1''$	$x_2'-x_2''$	$\dots$	$x_i'-x_i''$	$\dots$	$x_k'-x_k''$	Total
Freq(f)	f_1	f_2	$\dots$	f_i	$\dots$	f_k	N

  - If f_i is the unique maximum of $(f_1, f_2, \dots, f_k)$, then $x_i'-x_i''$ is the modal class
  - The mode is

$$\check{x} = x_i' + \frac{f_i - f_{i-1}}{2f_i - f_{i-1} - f_{i+1}} \times c$$

Measures of Dispersion I

- Measures of Dispersion describe the spread of the observations in the sample. More the spread, larger are these measures.
- Common measures of dispersion are range, interquartile range, variance, standard deviation, coefficient of variation, etc..

Measures of Dispersion II

Dispersion Measures based on extreme values

1. Range
   - Range is defined as the difference between the maximum value and the minimum value of the observations.
   - For series data: $x_{(n)} - x_{(1)}$
   - For grouped data: Discrete-$x_k - x_1$ and Continuous$x''_k - x'_1$
2. Interquartile range (IQR)
   - $Q_3 - Q_1$, representing the range of values in which the middle 50% of the observations lie.
   - Quartile deviation/Semi IQR: $(Q_3 - Q_1)/2$

Dispersion Measures based on all values

1. Mean deviation about A
   - For series data: $MD_A = \frac{1}{n}\sum_{i=1}^{n} |x_i - A|$
   - For grouped data: $MD_A = \frac{1}{N}\sum_{i=1}^{k} f_i|x_i - A|$
   - Note:
     - If $A = \bar{x}$, it is called Mean deviation about mean and denoted by $MD_{\bar{x}}$
     - If $A = \tilde{x}$, it is called Mean deviation about median and denoted by $MD_{\tilde{x}}$

2. Root mean square deviation about A
   - For series data: $RMSD_A = +\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2}$
   - For grouped data: $RMSD_A = +\sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i\,(x_i - A)^2}$
   - Note:
     - If $A = \bar{x}$, it is called Standard deviation and denoted by $RMSD_{\bar{x}} = s$
     - The square of the standard deviation is called variance and denoted by s^2

- Simplified formula for variance

$$s^2 = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 & \text{, ungrouped} \\ \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2 & \text{, grouped,} \end{cases}$$

Measures of Dispersion VI

- Some results on dispersion
  1. Mean deviation about A is smallest when measured about median:
  $$MD_{\tilde{x}} \leq MD_A$$
  Hint: $n\, MD_A = \sum_{i=1}^{n} |x_i - A| = |x_{(1)} - A| + |x_{(2)} - A| + \cdots + |x_{(n-1)} - A| + |x_{(n)} - A|$
  2. Root Mean Square deviation about A is smallest when measured about mean:
  $$RMSD_{\bar{x}} \leq RMSD_A$$
  Hint: $n\, RMSD_A^2 = \sum_{i=1}^{n} (x_i - A)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - A)^2$
  3. $s = 0 \Leftrightarrow$ all observations are equal

Measures of Dispersion VII

4. If the variable X having n values $(x_1, \ldots, x_n)$ has s.d. s_x, then the variable $Y = a + bX$ has s.d. $|b|s_x$.
5. If two groups of observations $\{x_{1,1}, \ldots, x_{1,n_1}\}$ and $\{x_{2,1}, \ldots, x_{2,n_2}\}$ have AMs $\bar{x}_1$ and $\bar{x}_2$, respectively and have standard deviations s_1 and s_2, respectively then the combined mean and sd of all the observations are
$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}.$$
$$s = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2}{n_1 + n_2}}$$
6. $MD_{\bar{x}} \leq RMSD_{\bar{x}}$

   Hint: Take $a_i = |x_i - \bar{x}|$ and $b_i = 1 \forall i$ in the following CS inequality
   $(a_1 b_1 + \cdots + a_n b_n)^2 \leq (a_1^2 + \cdots + a_n^2)(b_1^2 + \cdots + b_n^2)$

Measures of Dispersion VIII

Dispersion Measures based on mutual differences

1. Gini's Coefficients: a single number aimed at measuring the degree of inequality among values of a frequency distribution

- Gini's Coefficients = half of the relative mean absolute difference,

$$G = \frac{1}{2} \frac{\left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \right)}{\bar{x}}$$

Measures of Dispersion IX

- Note:
  - $G \geq 0$
    - G=0 *iff* all observations have same values
  - $G \leq 1$
    Hint : $|x_i - x_j| \leq |x_i| + |x_j|$
    - $G \approx 1$ when all but one observations are zero

Measures of Dispersion X

Relative Measures of dispersion

1. Coefficient of variation (CV)
   - It is SD/Mean (multiplied by 100 when expressed as a percentage)
   - that is, $CV = \frac{s}{\bar{x}} \times 100$
2. Coefficient of mean-deviation (CMD)
   - $CMD = \frac{MD_{\bar{x}}}{\bar{x}}$
   - Can also be calculated w.r.t. $\tilde{x}$ and $\check{x}$
3. Coefficient of quartile deviation (CQD)
   - $CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$

- Note: All of the above are unit-less

Skewness I

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- A frequency distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
- For the n observations $x_1, x_2, \ldots, x_n$, some measures of skewness are
  1. Fisher-Pearson coefficient of skewness: $$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}}$$
  2. Galton/Bowley's skewness: $(Q_3 - 2Q_2 + Q_1)/(Q_3 - Q_1)$

Skewness II

- Zero skewness: Symmetric
  - A frequency distribution, or data set, is symmetric if it looks the same to the left and right of the center point.
- Positive skewness: right skewed
  - Frequency distribution, or data set has longer right tail than left tail
- Negative skewness: left skewed
  - Frequency distribution, or data set has longer left tail than right tail

Kurtosis I

- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
- For the n observations $x_1, x_2, \ldots, x_n$, the formula for kurtosis is

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}.$$

Kurtosis II

- Mesokurtic:
    - For observations from normal data kurtosis is 3
- Leptokurtic:
    - a “positive" or thin distribution (fatter/heave tails), kurtosis is more than 3
        - Its does not imply the distribution is "tall" as sometimes reported.
        - It produces more outliers than the normal distribution.
- Platykurtic:
    - a “negative" or wide distribution (thin/lighter tails), kurtosis is lesser than 3
        - Its does not imply the distribution is "flat-topped" as sometimes reported.
        - It produces fewer and less extreme outliers than does the normal distribution