Descriptive Statistics

1. The following data represent the hospitalization stay in days for a random sample from a flue epidemic source: 27, 33, 28, 27, 25, 31, 32, 34, 38, 41, 37, 22, 23, 27, 35, 25, 41, and 30.

   - Which is the sample size?

     Sample size = 18

   - Compute for the hospitalization stay the following statistics: mean, median, mode, variance, standard deviation and coefficient of variance.

     Mean: (27+33+28+27+25+31+32+34+38+41+37+22+23+27+35+25+41+30)/18 = 30.89

     Median:

     ÷ Sort the data ascending: 22 (X_1), 23 (X_2), 25 (X_3), 25 (X_4), 27 (X_5), 27 (X_6), 27 (X_7), 28 (X_8), 30 (X_9), 31 (X_{10}), 32 (X_{11}), 33 (X_{12}), 34 (X_{13}), 35 (X_{14}), 37 (X_{15}), 38 (X_{16}), 41 (X_{17}), 41 (X_{18})

     ÷ Choose the formula: $Me = (X_{18/2} + X_{18/2+1})/2 = (X_9 + X_{10})/2 = (30+31)/2 = 30.5$

     Mode: 27

     Variance: $s^2 = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

     s2 = $[(27-30.89)^2+(33-30.89)^2+(28-30.89)^2+(27-30.89)^2+(25-30.89)^2+(31-30.89)^2+(32-30.89)^2+(34-30.89)^2+(38-30.89)^2+(41-30.89)^2+(37-30.89)^2+(22-30.89)^2+(23-30.89)^2+(27-30.89)^2+(35-30.89)^2+(25-30.89)^2+(41-30.89)^2+(30-30.89)^2]/(18-1) = 34.69$

     Standard deviation: $s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$

     S = sqrt(34.69) = 6 (where sqrt = √)

     Coefficient of variance: $CV = \frac{s}{\bar{X}}$

     CV = 6/30.89 = 0.19

   - Based on the value of coefficient of variation specify the homogeneity of the series.

     Since the 10 ≤ CV < 20 the series is relative homogenous

   - Compute the quartiles for this series. What can be saying about the symmetry of the data?

     Q1 = 27

     Q2 = 30.5

     Q3 = 34.75

     Q2-Q1 = 3.5

     Q3-Q2 = 4.25

Since Q3-Q2 ~ Q2-Q1 the data could be symmetrical.

2. The following data represent the age (in years) at which the infection with HIV was diagnosis on a sample of 27 randomly selected cases:

39, 50, 26, 45, 71, 51, 33, 40, 40, 51, 66, 63, 55, 36, 57, 41, 61, 47, 44, 48, 59, 42, 54, 47, 53, 54, 47

- Compute with a precision of two decimals the following statistics:(i) median; (ii) mode; (iii) mean; (iv) central value; (v) amplitude; (vi) variation; (viii) coefficient of variation; (vii) standard deviation.

(i) Median

- Arrange data ascending

26 (X_1), 33 (X_2), 36 (X_3), 39 (X_4), 40 (X_5), 40 (X_6), 41 (X_7), 42 (X_8), 44 (X_9), 45 (X_{10}), 47 (X_{11}), 47 (X_{12}), 47 (X_{13}), 48 (X_{14}), 50 (X_{15}), 51 (X_{16}), 51 (X_{17}), 53 (X_{18}), 54 (X_{19}), 54 (X_{20}), 55 (X_{21}), 57 (X_{22}), 59 (X_{23}), 61 (X_{24}), 63 (X_{25}), 66 (X_{26}), 71 (X_{27})

- N is odd (27)
- Me: $X_{n+1}/2 \rightarrow X_{27+1}/2 \rightarrow X_{28}/2 \rightarrow X_{14} \rightarrow 48$
- Me = 48

(ii) Mode = 47

26, 33, 36, 39, 40, 40, 41, 42, 44, 45, 47, 47, 47, 48, 50, 51, 51, 53, 54, 54, 55, 57, 59, 61, 63, 66, 71

(iii) Mean = (26+33+36+39+40+40+41+42+44+45+47+47+47+48+50+51+51+53+54+54+55+57+59+61+63+66+71)/27 = 48.89

(iv) central value = $(X_{min} + X_{max})/2$ = (26+71)/2 = 48.5

(v) amplitude = $(X_{max}-X_{min})/2$ = (71-26)/2 = 22.5

(vi) variation = $((26-48.88)^2 + (33-48.88)^2 + (36-48.88)^2 + (39-48.88)^2 + (40-48.88)^2 + (40-48.88)^2 + (41-48.88)^2 + (42-48.88)^2 + (44-48.88)^2 + (45-48.88)^2 + (47-48.88)^2 + (47-48.88)^2 + (47-48.88)^2 + (48-48.88)^2 + (50-48.88)^2 + (51-48.88)^2 + (51-48.88)^2 + (53-48.88)^2 + (54-48.88)^2 + (54-48.88)^2 + (55-48.88)^2 + (57-48.88)^2 + (59-48.88)^2 + (61-48.88)^2 + (63-48.88)^2 + (66-48.88)^2 + (71-48.88)^2)/(27-1)$ = 106.56

(vii) coefficient of variation = sqrt(variance)/m = sqrt(106.56)/48.89 = 0.2111

(viii) standard deviation = sqrt(variance) = 10.32

- How many observation will be contain in the following ranges:
  - i. $\overline{X} \pm 1 \cdot s$: (48.89-10.32) (48.89+10.32) → range of (38.57 – 59.21) → 20 observations → 74% of the observations
  - ii. $\overline{X} \pm 2 \cdot s$: (48.89-2*10.32) (48.89+2*10.32) → range of (28.25 – 69.53) → 25 observations → 92.59% of the observations

iii. $\overline{X} \pm 3 \cdot s$: (48.89-3*10.32)-(48.89+3*10.32) → range of (17.93 – 79.85) → 27 observations → 100% of the observations

- Specify the level of homogeneity (or heterogeneity) of the sample.

Since the CV = 0.2111 the series is relative hererogenous.

- Assess the symmetry of distribution of data using quartiles.

Q1 = 41.5

Q2 = 48

Q3 = 54.4

(Q2 – Q1) = 6.5

(Q3 – Q2) = 6.5

Since (Q2 – Q1) = (Q3 – Q2) the series is symmetrical

3. Compute the following statistics for the sample of days of incubation:

7, 3, 5, 7, 10, 6, 8, 4, 5, 3, 7, 6, 5, 4, 8, 8, 7, 10, 12, 3, 2, 5, 6, 7, 8.

- Mean = (7+3+5+7+10+6+8+4+5+3+7+6+5+4+8+8+7+10+12+3+2+5+6+7+8)/25 = 6.24
- Median

2 (X_1), 3 (X_2), 3 (X_3), 3 (X_4), 4 (X_5), 4 (X_6), 5 (X_7), 5 (X_8), 5 (X_9), 5 (X_{10}), 6 (X_{11}), 6 (X_{12}), 6 (X_{13}), 7 (X_{14}), 7 (X_{15}), 7 (X_{16}), 7 (X_{17}), 7 (X_{18}), 8 (X_{19}), 8 (X_{20}), 8 (X_{21}), 8 (X_{22}), 10 (X_{23}), 10 (X_{24}), 12 (X_{25})

- N is odd (25)
- Me: $X_{n+1}/2 \rightarrow X_{25+1}/2 \rightarrow X_{26}/2 \rightarrow X_{13} \rightarrow 6$
- Me = 6
- Mode = 7
- Amplitude = $(X_{max}-X_{min})/2 = (12-2)/2 = 5$
- Standard deviation

Variance = $((7-6.24)^2 + (3-6.24)^2 + (5-6.24)^2 + (7-6.24)^2 + (10-6.24)^2 + (6-6.24)^2 + (8-6.24)^2 + (4-6.24)^2 + (5-6.24)^2 + (3-6.24)^2 + (7-6.24)^2 + (6-6.24)^2 + (5-6.24)^2 + (4-6.24)^2 + (8-6.24)^2 + (8-6.24)^2 + (7-6.24)^2 + (10-6.24)^2 + (12-6.24)^2 + (3-6.24)^2 + (2-6.24)^2 + (5-6.24)^2 + (6-6.24)^2 + (7-6.24)^2 + (8-6.24)^2)/(25-1) = 5.94$

Standard deviation = sqrt(variance) = 2.44

- Standard error = (standard deviation)/(sqrt(n)) = 2.44/(sqrt(25)) = 0.49
- Coefficient of variation. Give the interpretation of the obtained value.

CV = (standard deviation)/(arithmetic mean) = 2.44/6.24 = 0.3910

The series is heterogeneous.

- Q_1 (25), Q_2 (50), Q_3 (75).

$Q_1 = 5$

$Q_2 = 6$

$Q_3 = 8$

$Q_2 – Q_1 = 6-5 = 1$

$Q_3 – Q_2 = 8-6 = 2$

Since $Q_3 – Q_2$ is different by $Q_2 – Q_1$ the series is not symmetrical.