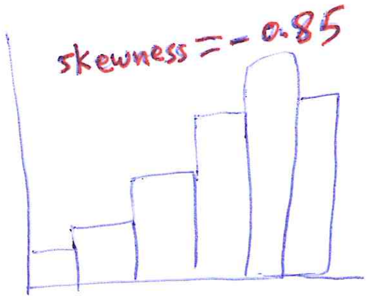
3.3 Measures of Distribution shape

An important numerical measure of the shape of a distribution is called skewness.

- For data skewed to the left, the skewness is negative (see A)
- " " " " " right, " " " positive (see B)
- If the data are symmetric, then the skewness is zero (see C)

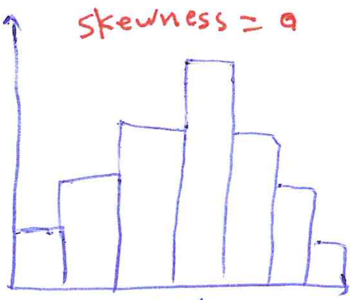


(A)

Moderately skewed left
(Mean < Median)

(B)

Moderately skewed right
Mean > Median



(C)

Symmetric
(Mean = Median)

(D)

Highley skewed Right

- Note that the median is the best measure of location when the data are highly skewed. Because the median is not affected by very high or very lower values.

Z - Scores

Z-score measures the relative location that determine how far a particular value from the mean.

- Suppose that a sample of n observations with values $x_1, x_2, \cdots, x_n$. Then a Z-score is defined by

$$z_i = \frac{x_i - \bar{x}}{s}$$

where
x_i is the i^{th} observation
$\bar{x}$ is the sample mean
s is the sample standard dev.
z_i is the Z-score for x_i

- Note that z-score is also called the standardized value.
  Z-score is the number of standard deviations x_i from $\bar{x}$.

- $Z_1 = 1.2$ means that x_1 is 1.2 standard deviation $> \bar{x}$
  $Z_2 = -0.5$ means that x_i is 0.5 standard deviation $< \bar{x}$

- If $Z_i > 0$, then $x_i > \bar{x}$
  If $Z_i < 0$, then $x_i < \bar{x}$
  If $Z_i = 0$, then $x_i = \bar{x}$

Example (Q 26 page 102) Consider a sample with a mean of 500 and a standard deviation of 100. What are the z-scores for the data values 520, 650, 500, 450, 280.

$\bar{x} = 500$
$s = 100$

$x_1 = 520$, $x_2 = 650$, $x_3 = 500$, $x_4 = 450$, $x_5 = 280$

$$Z_1 = \frac{x_1 - \bar{x}}{s} = \frac{520 - 500}{100} = 0.2$$

$$Z_2 = \frac{x_2 - \bar{x}}{s} = \frac{650 - 500}{100} = 1.5$$

$$Z_3 = \frac{x_3 - \bar{x}}{s} = \frac{500 - 500}{100} = 0$$

$$Z_4 = \frac{x_4 - \bar{x}}{s} = \frac{450 - 500}{100} = -0.5$$

$$Z_5 = \frac{x_5 - \bar{x}}{s} = \frac{280 - 500}{100} = -2.2$$

Chebyshev's Theorem

Chebyshev's Theorem: At least $\left(1-\frac{1}{z^2}\right)$ of the data values must be within z standard deviations of the mean, where $z > 1$.

Chebyshev's Theorem enables us to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean.

- At least 75% of the data values must be within $z = 2$ standard deviations of the mean.
- At least 89% of the data values must be within $z = 3$ standard deviations of the mean
- At least 94% of the data values must be within $z = 4$ standard deviations of the mean

---

Example: Suppose for a given sample with mean = 20 and standard deviation of 2. Use Chebyshev's theorem to determine the percentage of the data within each of the following ranges;

$\bar{x} = 20$
$s = 2$

[a] 16 — 24 [b] 14 — 26 [c] 10 — 30 [d] 15 - 25

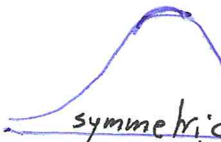
[a] $z = \frac{24-20}{2} = \frac{4}{2} = 2 \Rightarrow \left(1-\frac{1}{(2)^2}\right) = 1 - \frac{1}{4} = 0.75$ or 75% (at least)

[b] $z = \frac{26-20}{2} = \frac{6}{2} = 3 \Rightarrow \left(1-\frac{1}{(3)^2}\right) = 1 - \frac{1}{9} \approx 89\%$ (at least)

[c] $z = \frac{30-20}{2} = \frac{10}{2} = 5 \Rightarrow \left(1-\frac{1}{(5)^2}\right) = 1 - \frac{1}{25} = 96\%$ (at least)

[d] $z = \frac{25-20}{2} = \frac{5}{2} = 2.5 \Rightarrow \left(1-\frac{1}{(2.5)^2}\right) = 1 - \frac{1}{6.25} = 84\%$ (at least)

Empirical Rule



For data having a bell-shaped distribution — symmetric mound or bell-shaped or normal dist.

- Approximately 68 % of the data values will be within 1 standard deviation of the mean.
- Approximately 95 % of the data values will be within 2 standard deviations of the mean.
- Almost all of the data values will be within 3 standard deviations of the mean.

Example (Q28 page 102) Suppose the data have a bell-shaped distribution with mean 30 and standard deviation 5. Use the empirical rule to determine the percentage (or range) of the data within each of the following ranges:

$\bar{x} = 30$
$s = 5$

$\boxed{a}$ 20 to 40 $\boxed{b}$ 15 to 45 $\boxed{c}$ 25 - 35

$\boxed{a}$ $z = \frac{40-30}{5} = \frac{10}{5} = 2 \Rightarrow 95\%$ "approximately"

$\boxed{b}$ $z = \frac{45-30}{5} = \frac{15}{5} = 3 \Rightarrow$ Almost all data values

$\boxed{c}$ $z = \frac{35-30}{5} = \frac{5}{5} = 1 \Rightarrow 68\%$ "approximately"

Detecting Outliers:

Outliers : are observations that are unusually large or unusually small (extreme values)

- They may be included incorrectly in the data set.
- We use z-score to detect outliers.
- Since within $z = 3$, we almost have all the data values in a bell-shaped distribution. Then any observation with $z < -3$ or $z > 3$ is an outlier.