OUTLIERS

Descriptive Statistics



Quick Overview

An outlier is an observation unusually far from the other values in a data set.

Outliers may result from errors or may be genuine but unusual observations.

They can affect the mean, range, variance, and standard deviation.

1. What Is an Outlier?

An outlier is a value that differs substantially from the general pattern of the remaining observations.

Example: 12, 13, 14, 15, 16, 17, 65

The value 65 is much larger than the other observations and may be a potential outlier.

Whether it is an outlier should be assessed using the data context and an appropriate statistical method.

2. Why Do Outliers Occur?

Measurement error: an instrument may be inaccurate or used incorrectly.

Data-entry error: a value may have been typed or recorded incorrectly.

Sampling variation: an individual or item may genuinely have an unusual measurement.

Natural variation: some populations contain valid extreme observations.

3. Identifying Potential Outliers: IQR Method

The interquartile range measures the spread of the middle 50% of ordered observations.

$IQR = Q_3 - Q_1$

Q_1 is the first quartile (25th percentile); Q_3 is the third quartile (75th percentile).

Lower bound = $Q_1 - 1.5(IQR)$

Upper bound = $Q_3 + 1.5(IQR)$

Values below the lower bound or above the upper bound are flagged as potential outliers.

4. Identifying Potential Outliers: Z-Scores

A z-score indicates how many standard deviations an observation lies from the sample mean.

$z = (x - \bar{x}) / s$

x = observation; $\bar{x}$ = sample mean; s = sample standard deviation.

A large absolute z-score—often around 3 or more—may indicate a potential outlier.

This guideline may be misleading for highly skewed data and is not a universal rule.

5. Effects on Descriptive Statistics

Mean: may be pulled toward an unusually large or small observation.

Median: usually less affected because it depends on ordered position.

Range: may increase substantially because it uses the minimum and maximum.

Variance and standard deviation: may increase because extreme values can have large squared deviations.

IQR: generally more resistant to extreme observations than the range and standard deviation.

Key Terms

Outlier: an observation unusually distant from the general pattern.

Quartiles: values dividing ordered data into four parts.

IQR: the difference between the third and first quartiles.

Z-score: the number of standard deviations an observation lies from the mean.

Quick Review Questions

1. What is an outlier?
2. State two possible causes of outliers.
3. Write the formula for the interquartile range.
4. What are the lower and upper bounds in the 1.5 × IQR method?
5. Which is usually more resistant to outliers: the mean or the median?

Given the lower quartile (Q_1) of 120 mg/dl and the upper quartile (Q_3)of 180 mg/dl for blood glucose level in diabetes patients, which of the following values would be considered an outlier

A) 150
B) 200
C) 35
D) 275
E) 230