

Chapter 2

Methods for Describing Sets of Data

Numerical Measures of Variability

The Range

- Largest measurement minus the smallest measurement
- Loses sensitivity when data sets are large

These 2 distributions have the same range. How much does the range tell you about the data variability?

4

Objectives

Describe Data using Graphs

Describe Data using Numerical Measures

2

Numerical Measures of Variability

The Sample Variance (s²)

- The sum of the squared deviations from the mean divided by (n-1). Expressed as units squared

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- Why square the deviations? The sum of the deviations from the mean is zero

5

Numerical Measures of Variability

- Variability – the spread of the data across possible values
- 3 commonly used measures of Variability
 - Range
 - Variance
 - Standard Deviation

3

Numerical Measures of Variability

The Sample Standard Deviation (s)

- The positive square root of the sample variance

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{s^2}$$

- Expressed in the original units of measurement

6

Numerical Measures of Variability

Samples and Populations - Notation

	Sample	Population
Variance	s^2	σ^2
Standard Deviation	s	σ

7

Numerical Measures of Relative Standing

Descriptive measures of relationship of a measurement to the rest of the data

Common measures:

- percentile ranking or percentile score
- z-score

10

Interpreting the Standard Deviation

How many observations fit within $\pm n$ s of the mean?

	Chebyshev's Rule	Empirical Rule
$\pm 1s$ or $\pm 1\sigma$	No useful info	Approximately 68%
$\pm 2s$ or $\pm 2\sigma$	At least 75%	Approximately 95%
$\pm 3s$ or $\pm 3\sigma$	At least 8/9	Approximately 99.7%

8

Numerical Measures of Relative Standing

Percentile rankings make use of the p th percentile

The median is an example of percentiles.

Median is the 50th percentile – 50 % of observations lie above it, and 50% lie below it

For any p , the p th percentile has $p\%$ of the measures lying below it, and $(100-p)\%$ above it

11

Interpreting the Standard Deviation

You have purchased compact fluorescent light bulbs for your home. Average life length is 500 hours, standard deviation is 24, and frequency distribution for the life length is mound shaped. One of your bulbs burns out at 450 hours. Would you send the bulb back for a refund?

Interval	Range	% of observations included	% of observations excluded
$\pm 1s$	476 - 524	Approximately 68%	Approximately 32%
$\pm 2s$	452 - 548	Approximately 95%	Approximately 5%
$\pm 3s$	428 - 572	Approximately 99.7%	Approximately 0.3%

9

Numerical Measures of Relative Standing

z-score – the distance between a measurement x and the mean, expressed in standard units

Use of standard units allows comparison across data sets

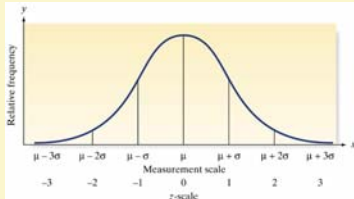
$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{x - \bar{x}}{s}$$

12

Numerical Measures of Relative Standing

More on z-scores

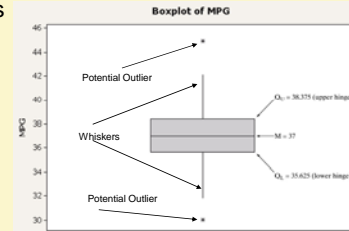
Z-scores follow the empirical rule for mounded distributions



13

Methods for Detecting Outliers

Box Plots



Not on plot – inner and outer fences, which determine potential outliers

16

Methods for Detecting Outliers

Outlier – an observation that is unusually large or small relative to the data values being described

Causes

- Invalid measurement
- Misclassified measurement
- A rare (chance) event

2 detection methods

- Box Plots
- z-scores

14

Methods for Detecting Outliers

Rules of thumb

•Box Plots

- measurements between inner and outer fences are suspect
- measurements beyond outer fences are highly suspect

•Z-scores

- Scores of ± 3 in mounded distributions (± 2 in highly skewed distributions) are considered outliers

17

Methods for Detecting Outliers

Box Plots

- based on **quartiles**, values that divide the dataset into 4 groups
- Lower Quartile Q_L – 25th percentile
- Middle Quartile - median
- Upper Quartile Q_U – 75th percentile
- Interquartile Range (IQR) = $Q_U - Q_L$

15

Summary

Numerical measures of central tendency

- Mean
- Median
- Mode

•Numerical measures of variation

- Range
- Variance
- Standard Deviation

18

Summary

Distribution Rules

- Chebyshev's Rule
- Empirical Rule

•Measures of relative standing

- Percentile scores
- z-scores

•Methods for detecting Outliers

- Box plots
- z-scores

19