Chapter 3
Statistics for Describing, Exploring, and Comparing Data

3-1 Review and Preview

3-2 Measures of Center

3-3 Measures of Variation

3-4 Measures of Relative Standing and Boxplots

Definition

The **standard deviation** of a set of sample values, denoted by *s*, is a measure of variation of values about the mean.

1

Sample Standard Deviation Formula

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

Sample Standard Deviation (Shortcut Formula)

$$s = \sqrt{\frac{n\Sigma(x^2) - (\Sigma x)^2}{n\,(n-1)}}$$

Sample Standard Deviation (Shortcut Formula)

$$s = \sqrt{\frac{\sum\left(x^2\right) - n\left(\overline{x}\right)^2}{n-1}}$$

2

Standard Deviation - Important Properties

- The standard deviation is a measure of variation of all values from the mean.
- The value of the standard deviation s is usually positive.
- The value of the standard deviation s can increase dramatically with the inclusion of one or more outliers (data values far away from all others).
- The units of the standard deviation s are the same as the units of the original data values.

Old Faithful Geyser
Intervals between eruptions (min)

98 92 95 87 96 90 65 92 95 93 98 94

Find:

A. Sample standard deviation

$$\overline{x} = 91.25$$

$$\sum_{i=1}^{n} X_i^2 = 100781$$

B. Sample standard deviation using Range Rule of Thumb

Comparing Variation in Different Samples

It's a good practice to compare two sample standard deviations only when the sample means are approximately the same.

When comparing variation in samples with very different means, it is better to use the coefficient of variation, which is defined later in this section.

3

Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma\,(x-\mu)^2}{N}}$$

This formula is similar to the previous formula, but instead, the population mean and population size are used.

Variance

- The **variance** of a set of values is a measure of variation equal to the square of the standard deviation.
- Sample variance: s^2 - Square of the sample standard deviation s
- Population variance: σ^2 - Square of the population standard deviation σ

Unbiased Estimator

The sample variance s^2 is an **unbiased estimator** of the population variance σ^2, which means values of s^2 tend to target the value of σ^2 instead of systematically tending to overestimate or underestimate σ^2.

Variance - Notation

s = *sample* standard deviation

s^2 = *sample* variance

σ = *population* standard deviation

σ^2 = *population* variance

Part 2

Beyond the Basics of Measures of Variation

Range Rule of Thumb

is based on the principle that for many data sets, the vast majority (such as 95%) of sample values lie within two standard deviations of the mean.

Range Rule of Thumb for Interpreting a Known Value of the Standard Deviation

Informally define *usual* values in a data set to be those that are typical and not too extreme. Find rough estimates of the minimum and maximum "usual" sample values as follows:

Minimum "usual" value = (mean) – 2 × (standard deviation)

Maximum "usual" value = (mean) + 2 × (standard deviation)

Range Rule of Thumb for Estimating a Value of the Standard Deviation *s*

To roughly estimate the standard deviation from a collection of known sample data use

$$s \approx \frac{\text{range}}{4}$$

where

range = (maximum value) – (minimum value)

Properties of the Standard Deviation

- Measures the variation among data values
- Values close together have a small standard deviation, but values with much more variation have a larger standard deviation
- Has the same units of measurement as the original data

Properties of the Standard Deviation

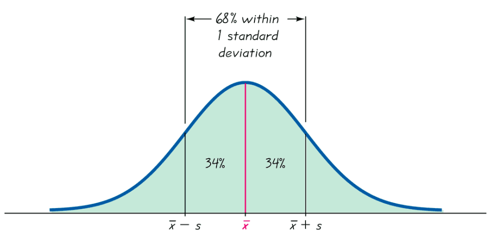
- For many data sets, a value is *unusual* if it differs from the mean by more than two standard deviations
- Compare standard deviations of two different data sets only if the they use the same scale and units, and they have means that are approximately the same

Empirical (or 68-95-99.7) Rule

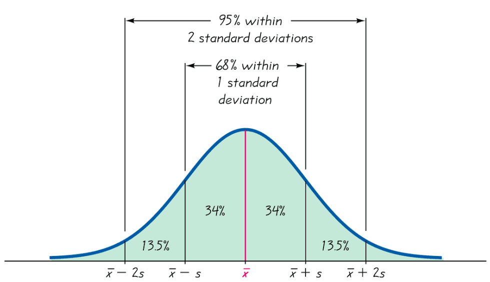
For data sets having a distribution that is approximately bell shaped, the following properties apply:

- **About 68% of all values fall within 1 standard deviation of the mean.**
- **About 95% of all values fall within 2 standard deviations of the mean.**
- **About 99.7% of all values fall within 3 standard deviations of the mean.**

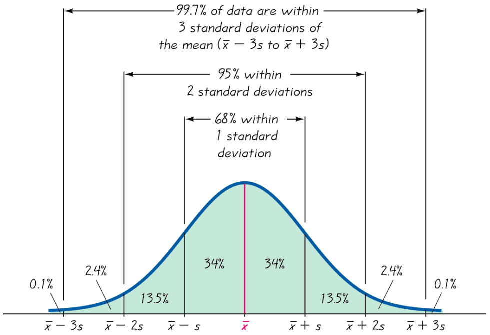
The Empirical Rule



The Empirical Rule



The Empirical Rule



Chebyshev's Theorem

The proportion (or fraction) of any set of data lying within K standard deviations of the mean is always at least $1–1/K^2$, where K is any positive number greater than 1.

- **For $K = 2$, at least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean.**
- **For $K = 3$, at least 8/9 (or 89%) of all values lie within 3 standard deviations of the mean.**

Rationale for using $n – 1$ versus n

There are only $n – 1$ independent values. With a given mean, only $n – 1$ values can be freely assigned any number before the last value is determined.

Dividing by $n – 1$ yields better results than dividing by n. It causes s^2 to target σ^2 whereas division by n causes s^2 to underestimate σ^2.

Coefficient of Variation

The **coefficient of variation** (or **CV**) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

Sample

$$CV = \frac{s}{\bar{x}} \bullet 100\%$$

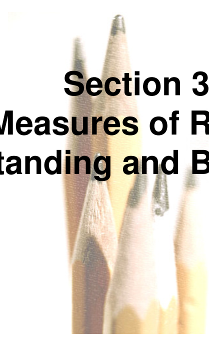
Population

$$CV = \frac{\sigma}{\mu} \bullet 100\%$$

Recap

In this section we have looked at:

- Range
- Standard deviation of a sample and population
- Variance of a sample and population
- Range rule of thumb
- Empirical distribution
- Chebyshev's theorem
- Coefficient of variation (CV)



Section 3-4
Measures of Relative Standing and Boxplots

Key Concept

This section introduces measures of relative standing, which are numbers showing the location of data values relative to the other values within a data set. They can be used to compare values from different data sets, or to compare values within the same data set. The most important concept is the *z* score. We will also discuss percentiles and quartiles, as well as a new statistical graph called the boxplot.

Part 1

Basics of *z* Scores, Percentiles, Quartiles, and Boxplots

Z score

- ***z* Score** (or standardized value)

  the number of standard deviations that a given value *x* is above or below the mean

Measures of Position *z* Score

Sample

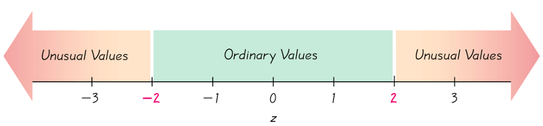
$$z = \frac{x - \bar{x}}{s}$$

Population

$$z = \frac{x - \mu}{\sigma}$$

Round *z* scores to 2 decimal places

Interpreting *Z* Scores



Whenever a value is less than the mean, its corresponding *z* score is negative

Ordinary values: $-2 \le z \text{ score} \le 2$

Unusual Values: $z \text{ score} < -2$ or $z \text{ score} > 2$

Percentiles

are measures of location. There are 99 percentiles denoted P_1, P_2, . . . P_{99}, which divide a set of data into 100 groups with about 1% of the values in each group.

Finding the Percentile of a Data Value

$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Converting from the *k*th Percentile to the Corresponding Data Value

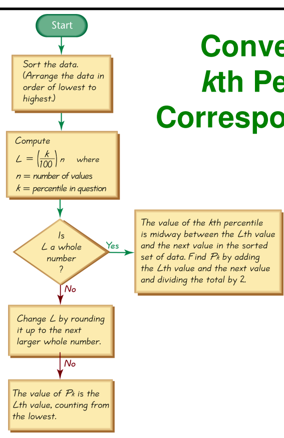
Notation

$$L = \frac{k}{100} \cdot n$$

- n total number of values in the data set
- k percentile being used
- L locator that gives the position of a value
- P_k *k*th percentile

12

Converting from the *k*th Percentile to the Corresponding Data Value



Quartiles

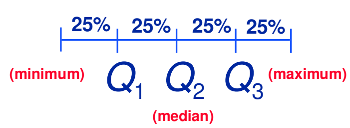
Are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with about 25% of the values in each group.

- Q_1 **(First Quartile)** separates the bottom 25% of sorted values from the top 75%.
- Q_2 **(Second Quartile)** same as the median; separates the bottom 50% of sorted values from the top 50%.
- Q_3 **(Third Quartile)** separates the bottom 75% of sorted values from the top 25%.

Quartiles

$$Q_1,\ Q_2,\ Q_3$$

divide **ranked** scores into four equal parts



Some Other Statistics

- **Interquartile Range (or IQR):** $Q_3 - Q_1$
- **Semi-interquartile Range:** $\dfrac{Q_3 - Q_1}{2}$
- **Midquartile:** $\dfrac{Q_3 + Q_1}{2}$
- **10 - 90 Percentile Range:** $P_{90} - P_{10}$

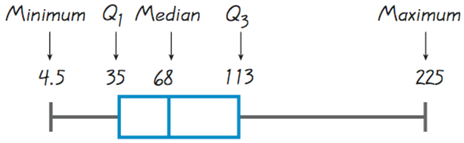
5-Number Summary

- For a set of data, the **5-number summary** consists of the minimum value; the first quartile Q_1; the median (or second quartile Q_2); the third quartile, Q_3; and the maximum value.

Boxplot

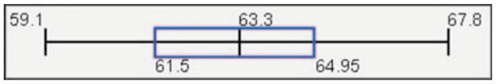
- A **boxplot** (or **box-and-whisker-diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, Q_1; the median; and the third quartile, Q_3.

Boxplots



Boxplot of Movie Budget Amounts

Boxplots - Normal Distribution



Normal Distribution:
Heights from a Simple Random Sample of Women

Boxplots - Skewed Distribution

Skewed Distribution:
Salaries (in thousands of dollars) of NCAA Football Coaches

Part 2

Outliers and Modified Boxplots

Outliers

- An outlier is a value that lies very far away from the vast majority of the other values in a data set.

Important Principles

- An outlier can have a dramatic effect on the mean.
- An outlier can have a dramatic effect on the standard deviation.
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured.

Outliers for Modified Boxplots

For purposes of constructing *modified boxplots*, we can consider outliers to be data values meeting specific criteria.

In modified boxplots, a data value is an outlier if it is . . .

above Q_3 by an amount greater than $1.5 \times$ IQR

or

below Q_1 by an amount greater than $1.5 \times$ IQR

Modified Boxplots

Boxplots described earlier are called **skeletal** (or **regular**) boxplots.

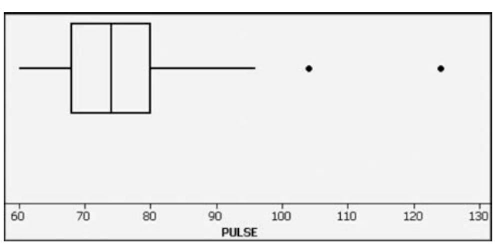
Some statistical packages provide **modified boxplots** which represent outliers as special points.

Modified Boxplot Construction

A modified boxplot is constructed with these specifications:

- A special symbol (such as an asterisk) is used to identify outliers.
- The solid horizontal line extends only as far as the minimum data value that is not an outlier and the maximum data value that is not an outlier.

Modified Boxplots - Example



Pulse rates of females listed in Data Set 1 in Appendix B.

Recap

In this section we have discussed:

- z Scores
- z Scores and unusual values
- Percentiles
- Quartiles
- Converting a percentile to corresponding data values
- Other statistics
- 5-number summary
- Boxplots and modified boxplots
- Effects of outliers

Putting It All Together

Always consider certain key factors:

- Context of the data
- Source of the data
- Sampling Method
- Measures of Center
- Measures of Variation
- Distribution
- Outliers
- Changing patterns over time
- Conclusions
- Practical Implications