**Chapter 2b: Methods for Describing Sets of Data (cont.)**

Empirical Rule

If a Histogram is Bell Shaped:

1. Approximately **68%** of all observations fall within **one** standard deviation of the mean.
2. Approximately **95%** of all observations fall within **two** standard deviations of the mean
3. Approximately **99.7%** of all observations fall within **three** standard deviations of the mean



**Example - The height of male students at Clemson is approximately normally distributed (bell shaped) with a mean of 71 inches and standard deviation of 2.5 inches.**

1. **What percent of the male students are shorter than 66 inches?**
2. **Taller than 73.5 inches?**
3. **Between 66 & 73.5 inches?**

**Example - Given a data set comprised of 4037 measurements that is bell-shaped with a mean of 256. If 99.7% of the data lies between -26 and 538 then what is the standard deviation?**

Chebyshev’s Rule

* Applies to data of ANY shape (including bell shaped and symmetric)
* No useful information is provided on the fraction of measurements that fall within 1 standard deviation of the mean
* Generally, for any number of k (>1), at least
 of the measurements will fall within k standard deviations of the of the mean

**Example - For any data set, what percentage of the data should lie within 2 standard deviations of the mean?**

**Example - Consider the sample data:**

**41 44 45 47 48 51 53 58 66**

**What exact percentage of this data lies within 2 standard deviations of the mean?**

**Example - Consider the sample data:**

**20 37 48 48 49 50 53 61 64 70**

**What exact percentage of this data lies within 2 standard deviations of the mean?**

Approximation of Sample Standard Deviation:

We know from both the Empirical and Chebyshev’s rules that most of the measurements from a data set will be within 2 standard deviations from the mean and that most all of the measurements from a data set will be within 3 standard deviations from the mean. Consequently, we would expect the range of the measurements to be between 4 and 6 standard deviations in length.

(sometimes in exercises and examples your book uses range/4 as a crude approximation for s.

**Measures of Position**

* **Percentiles**
* **Quartiles**
* **Z-scores**

Percentiles

* The pth percentile is the value such that *p* percent of the observations fall below or at that value



Quartiles

* Splits data into quarters (fourths)
* First quartile = 25th percentile
* Second quartile = 50th percentile
* Third quartile = 75th percentile

Finding Quartiles

1. Arrange data in order
2. Consider the median (the midpoint). That is the **second quartile**, Q2
3. Consider the lower half of the observations. The median of these observations is the **first quartile**, Q1.
4. Consider the upper half of observations. Their median is the **third quartile**, Q3

Note: When using your calculator or any other technology, different devices use different methods for finding quartiles.

**Simple Example - Suppose a personnel manager has hired 10 new employees. The ages of each of these employees sorted from low to high is listed as follows:**

**23 25 25 34 35 45 46 47 52 54**

Interquartile Range

**IQR = Q3 – Q1**

* Measure of spread
* Resistant statistic

Five Number Summary

* Min
* Q1
* Median
* Q3
* Max

**Example - A report from the U.S. Department of Justice gave the following percent increase in federal prison populations in 20 northeastern & mid-western states in 1999.**

**5.9 1.3 5.0 5.9 4.5 5.6 4.1 6.3 4.8 6.9**

**4.5 3.5 7.2 6.4 5.5 5.3 8.0 4.4 7.2 3.2**

**Find the Five Number Summary.**

Boxplots

Why use them?

* ease of construction
* convenient handling of outliers
* construction is not subjective (like histograms)
* Used with medium or large size data sets (n > 10)
* useful for comparative displays

Disadvantages of Boxplots

* does not retain the individual observations
* should **not** be used with small data sets (n < 10)

How to Construct Boxplot

1. find five-number summary

Min Q1 Med Q3 Max

1. draw box from Q1 to Q3
2. draw median as center line in the box
3. Calculate Fences
4. whiskers extend to largest (smallest) data value **inside** the fence
5. Mark outliers with an asterisk

Fences (Or Bounds)



Shapes of Boxplots



**Example - A report from the U.S. Department of Justice gave the following percent increase in federal prison populations in 20 northeastern & mid-western states in 1999.**

**5.9 1.3 5.0 5.9 4.5 5.6 4.1 6.3 4.8 6.9**

**4.5 3.5 7.2 6.4 5.5 5.3 8.0 4.4 7.2 3.2**

**Create a boxplot. Describe the distribution.**

RECALL: When describing graphs or charts describe Center, Shape, and Spread

**Example –**

Cancer

No Cancer

**The median radon concentration for the no cancer group is lower than the median for the cancer group. The range of the cancer group is larger than the range for the no cancer group. Both distributions are skewed right. The cancer group has outliers at 39, 45, 57, and 210. The no cancer group has outliers at 55 and 85.**

Z-scores

* In a bell shaped distribution a potential outlier is more than 3 standard deviations from the mean.
* Z-score measures the number of standard deviations from the mean



**Example - If  and inches, what is the z-score for the observation of 19.4?**

**Example - Ellen is taking Math and English. On her first test in English she scored a 78. The mean of her English class was 74 and the standard deviation was 5. On her first test in Math she scored a 91. The mean of her Math class was 89 and the standard deviation was 6.**

**What is the z-score for her English score?**

**What is the z-score for her Math score?**

**Which class is she doing better in relative to the rest of her class?**

**Scatterplot**

* When a dataset has multiple variables and we want to see a relationship between two of the quantitative variables we can plot the data using a scatterplot.
* A scatterplot is created using one quantitative variable on the x-axis and the other on the y-axis.
* We will look at bivariate data and scatterplots further in Chapter 10.

Graphical Misuses

Guidelines for Constructing Effective Graphs

* Label both axes and provide a heading
* The vertical axis should start at 0.
* Be cautious in using figures
* Be cautious of comparing more than one group on a single graph

Tricky Graphs

 

 



Time Plots

* Data collected over time is called a **time series**
* Use a **time plot** to display time series data
* Look for **trends** (rise, fall, etc.)

**Example – DOW Jones Industrial Average (1900-Present Monthly)**

****

[**http://stockcharts.com/freecharts/historical/djia1900.html**](http://stockcharts.com/freecharts/historical/djia1900.html)