Chapter 4

Numerical Methods for Describing Data

1

4.1: Describing the Center of a Data Set

The **sample mean** of a numerical sample, $x_1, x_2, x_3, \ldots, x_n$, denoted $\bar{x}$, is

$$\bar{x} = \frac{\text{sum of all observations in the sample}}{\text{number of observations in the sample}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

2

Describing the Center of a Data Set with the arithmetic mean

The **population mean** is denoted by μ, is the average of all x values in the entire population.

3

Example calculations

During a two week period 10 houses were sold in Fancytown.

House Price in Fancytown x
231,000
313,000
299,000
312,000
285,000
317,000
294,000
297,000
315,000
287,000
$\sum x =$ 2,950,000

$$\bar{x} = \frac{\sum x}{n} = \frac{2,950,000}{10} = 295,000$$

The "average" or mean price for this sample of 10 houses in Fancytown is $295,000

4

Example calculations

During a two week period 10 houses were sold in Lowtown.

House Price in Lowtown x
97,000
93,000
110,000
121,000
113,000
95,000
100,000
122,000
99,000
2,000,000
$\sum x =$ 2,950,000

Outlier

$$\bar{x} = \frac{\sum x}{n} = \frac{2,950,000}{10} = 295,000$$

The "average" or mean price for this sample of 10 houses in Lowtown is $295,000

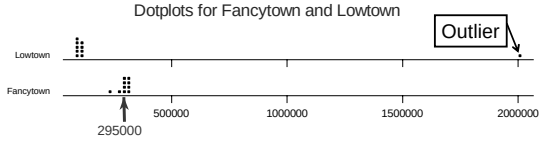
5

Reflections on the Sample calculations

Looking at the dotplots of the samples for Fancytown and Lowtown we can see that the mean, $295,000 appears to accurately represent the "center" of the data for Fancytown, but it is not representative of the Lowtown data.

Clearly, the mean can be greatly affected by the presence of even a single *outlier*.

Dotplots for Fancytown and Lowtown



6

Comments

1. In the previous example of the house prices in the sample of 10 houses from Lowtown, the mean was affected very strongly by the one house with the extremely high price.
2. The other 9 houses had selling prices around $100,000.
3. This illustrates that the mean can be very sensitive to a few extreme values.

7

Describing the Center of a Data Set with the median

The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included, so that every sample observation appears in the ordered list). Then

$$\text{sample median} = \begin{cases} \text{the single middle value if n is odd} \\ \text{the mean of the middle two values if n is even} \end{cases}$$

8

Example of Median Calculation

Consider the Fancytown data. First, we put the data in numerical increasing order to get

231,000 285,000 287,000 294,000
297,000 299,000 312,000 313,000
315,000 317,000

Since there are 10 (even) data values, the median is the mean of the two values in the middle.

$$\text{median} = \frac{297000 + 299000}{2} = \$298,000$$

9

Example of Median Calculation

Consider the Lowtown data. We put the data in numerical increasing order to get

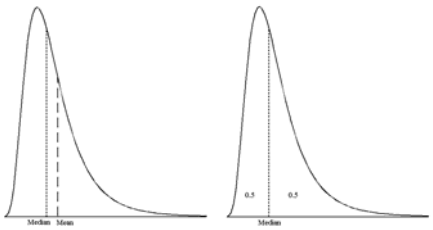
93,000 95,000 97,000 99,000
100,000 110,000 113,000 121,000
122,000 2,000,000

Since there are 10 (even) data values, the median is the mean of the two values in the middle.

$$\text{Median} = \frac{100000 + 110000}{2} = 105{,}000$$

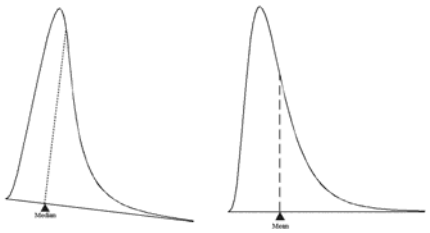
10

Comparing the Sample Mean & Sample Median



11

Comparing the Sample Mean & Sample Median



12

Comparing the Sample Mean & Sample Median

Notice from the preceding pictures that the median splits the area in the distribution in half and the mean is the point of balance.

Typically,

1. when a distribution is skewed positively, the mean is larger than the median,
2. when a distribution is skewed negatively, the mean is smaller then the median, and
3. when a distribution is symmetric, the mean and the median are equal.

13

The Trimmed Mean

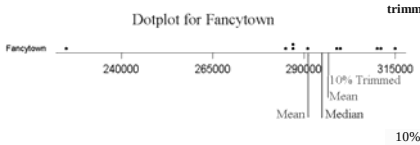
A **trimmed mean** is computed by first ordering the data values from smallest to largest, deleting a selected number of values from each end of the ordered list, and finally computing the mean of the remaining values.

The **trimming percentage** is the percentage of values deleted from each end of the ordered list.

14

Example of Trimmed Mean

Med=298K



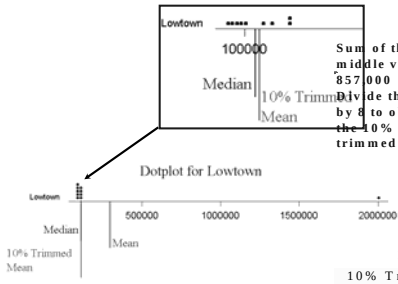
Sum of the eight middle values is 2,402,000

Divide this value by 8 to obtain the 10% trimmed mean.

House Price in Fancytown
231,000
285,000
287,000
294,000
297,000
299,000
312,000
313,000
315,000
317,000
$\sum x =$ 2,950,000
$\bar{x} =$ 295,000
median = 298,000
10% Trim Mean = 300,250

15

Example of Trimmed Mean



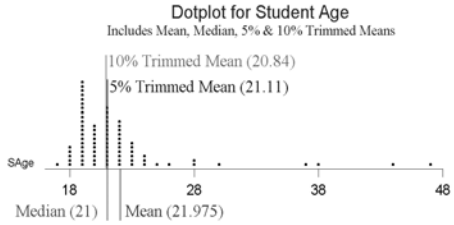
Sum of the eight middle values is 857,000
Divide this value by 8 to obtain the 10% trimmed mean.

House Price in Lowtown
93,000
95,000
97,000
99,000
100,000
110,000
113,000
121,000
122,000
2,000,000
$\sum x = 2{,}950{,}000$
$\bar{x} = 295{,}000$
median = 105,000
10% Trim Mean = 107,125

16

Another Example

Here's an example of what happens if you compute the mean, median, and 5% & 10% trimmed means for the Ages for the 79 students taking Data Analysis



17

Categorical Data - Sample Proportion

The **sample proportion of successes**, denoted by p, is

$$p = \text{sample proportion of successes} = \frac{\text{number of S's in the sample}}{n}$$

Where S is the label used for the response designated as success. The population proportion of successes is denoted by π.

18

Categorical Data - Sample Proportion

If we look at the student data sample, consider the variable gender and treat being female as a success, we have 25 of the sample of 79 students are female, so the sample proportion (of females) is

$$p = \frac{25}{79} = 0.316$$

19

4.2: Describing Variability

The simplest numerical measure of the variability of a numerical data set is the **range**, which is defined to be the difference between the largest and smallest data values.

range = maximum - minimum

20

Describing Variability

The n **deviations from the sample mean** are the differences:

$$x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, \ldots, x_n - \bar{x},$$

Note: The sum of all of the deviations from the sample mean will be equal to 0, except possibly for the effects of rounding the numbers. This means that the average deviation from the mean is always 0 and cannot be used as a measure of variability.

21

Sample Variance

The sample **variance**, denoted $\mathbf{s^2}$ is the sum of the squared deviations from the mean divided by n-1.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

Note: $S_{xx} = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$

$$= \sum (x - \bar{x})^2$$

22

Sample Standard Deviation

The sample **standard deviation**, denoted **s** is the positive square root of the sample variance.

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

The **population standard deviation** is denoted by σ.

23

Example calculations

10 Macintosh Apples were randomly selected and weighed (in ounces).

x	$x - \bar{x}$	$(x - \bar{x})^2$
7.52	0.096	0.0092
8.48	1.056	1.1151
7.36	-0.064	0.0041
6.24	-1.184	1.4019
7.68	0.256	0.0655
6.56	-0.864	0.7465
6.40	-1.024	1.0486
8.16	0.736	0.5417
7.68	0.256	0.0655
8.16	0.736	0.5417
74.24	0.000	5.5398

Range $= 8.48 - 6.24 = 2.24$

$$\bar{x} = \frac{\sum x}{n} = \frac{74.24}{10} = 7.424$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{5.5398}{10-1}$$

$$= \frac{5.5398}{9} = 0.61554$$

$$s = \sqrt{0.61554} = 0.78456$$

24

Calculator Formula for s^2 and s

A little algebra can establish the sum of the square deviations,

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \Big(\sum x\Big)^2 \Big/ n$$

A **computational formula** for the sample variance is given by

$$s^2 = \frac{S_{xx}}{n-1} = \frac{\sum x^2 - \Big(\sum x\Big)^2 \Big/ n}{n-1}$$

25

Calculations Revisited

x	x^2
7.52	56.5504
8.48	71.9104
7.36	54.1696
6.24	38.9376
7.68	58.9824
6.56	43.0336
6.40	40.9600
8.16	66.5856
7.68	58.9824
8.16	66.5856
74.24	556.6976

$$n = 10,\ \sum x = 74.24,\ \sum x^2 = 556.6976$$

$$S_{xx} = \sum x^2 - \Big(\sum x\Big)^2 \Big/ n$$

$$= 556.6976 - \frac{(74.24)^2}{10} = 5.53984$$

$$s^2 = \frac{5.53984}{9} = 0.615538$$

$$s = \sqrt{0.615538} = 0.78456$$

The values for s^2 and s are exactly the same as were obtained earlier.

26

Quartiles and the Interquartile Range

Lower quartile (Q_1) = median of the lower half of the data set.

Upper Quartile (Q_3) = median of the upper half of the data set.

The **interquartile range** (iqr), a resistant measure of variability is given by

iqr = upper quartile – lower quartile

$= Q_3 - Q_1$

Note: If n is odd, the median is excluded from both the lower and upper halves of the data.

27

Quartiles and IQR Example

15 students with part time jobs were randomly selected and the number of hours worked last week was recorded.

19, 12, 14, 10, 12, 10, 25, 9, 8, 4, 2, 10, 7, 11, 15

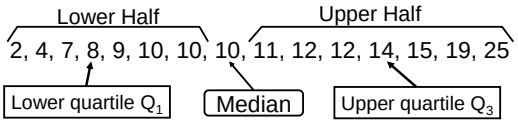
The data is put in increasing order to get

2, 4, 7, 8, 9, 10, 10, 10, 11, 12, 12, 14, 15, 19, 25

28

Quartiles and IQR Example

With 15 data values, the median is the 8th value. Specifically, the median is 10.



Lower quartile = 8 **Upper quartile = 14**

Iqr = 14 - 8 = 6

29

4.3: Boxplots

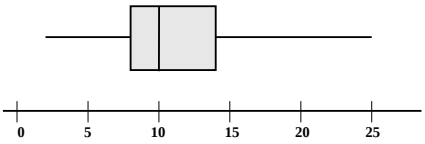
Constructing a Skeletal Boxplot

1. Draw a horizontal (or vertical) scale.
2. Construct a rectangular box whose left (or lower) edge is at the lower quartile and whose right (or upper) edge is at the upper quartile (the box width = iqr). Draw a vertical (or horizontal) line segment inside the box at the location of the median.
3. Extend horizontal (or vertical) line segments from each end of the box to the smallest and largest observations in the data set. (These lines are called whiskers.)

30

Skeletal Boxplot Example

Using the student work hours data we have



31

Outliers

An observations is an **outlier** if it is more than 1.5 iqr away from the closest end of the box (less than the lower quartile minus 1.5 iqr or more than the upper quartile plus 1.5 iqr.

An outlier is **extreme** if it is more than 3 iqr from the closest end of the box, and it is **mild** otherwise.

32

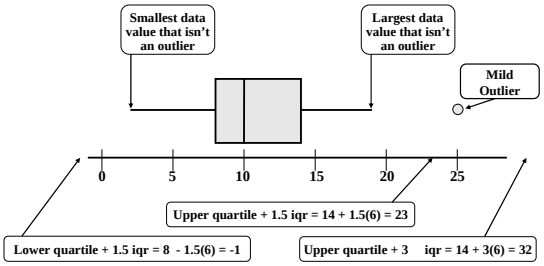
Modified Boxplots

A **modified boxplot** represents mild outliers by shaded circles and extreme outliers by open circles. Whiskers extend on each end to the most extreme observations that are *not* outliers.

33

Modified Boxplot Example

Using the student work hours data we have



34

Modified Boxplot Example

Consider the ages of the 79 students from the classroom data set from the slideshow Chapter 3.

Iqr = 22 – 19 = 3

Lower quartile – 3 iqr = 10 **Lower quartile – 1.5 iqr =14.5**
Upper quartile + 3 iqr = 31 **Upper quartile + 1.5 iqr = 26.5**

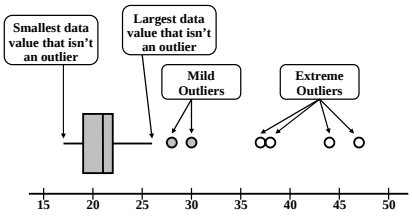
17	18	18	18	18	18	19	19	19	19	
19	19	19	19	19	19	19	19	19	19	Lower Quartile
19	19	19	19	19	19	20	20	20	20	
20	20	20	20	20	20	21	21	21	21	Median
21	21	21	21	21	21	21	21	21	21	
22	22	22	22	22	22	22	22	22	22	Upper Quartile
22	23	23	23	23	23	23	24	24	24	
25	26	28	28	30	37	38	44	47		

Moderate Outliers Extreme Outliers

35

Modified Boxplot Example

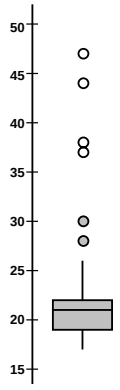
Here is the modified boxplot for the student age data.



36

Modified Boxplot Example

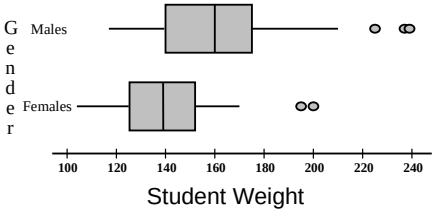
Here is the same boxplot reproduced with a vertical orientation.



37

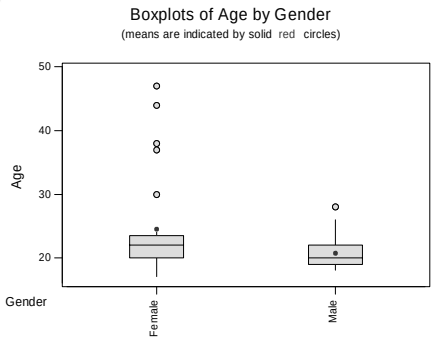
Comparative Boxplot Example

By putting boxplots of two separate groups or subgroups we can compare their distributional behaviors.
Notice that the distributional pattern of female and male student weights have similar shapes, although the females are roughly 20 lbs lighter (as a group).



38

Comparative Boxplot Example



39

4.4: Interpreting Variability
Chebyshev's Rule

Does not need to be a normal distribution

Chebyshev's Rule

Consider any number k, where $k \geq 1$. Then the percentage of observations that are within k standard deviations of the mean is at least

.

$$100\left(1-\frac{1}{k^2}\right)\%$$

40

Interpreting Variability
Chebyshev's Rule

For specific values of k Chebyshev's Rule reads

- At least 75% of the observations are within 2 standard deviations of the mean.
- At least 89% of the observations are within 3 standard deviations of the mean.
- At least 90% of the observations are within 3.16 standard deviations of the mean.
- At least 94% of the observations are within 4 standard deviations of the mean.
- At least 96% of the observations are within 5 standard deviations of the mean.
- At least 99% of the observations are with 10 standard deviations of the mean.

41

Example - Chebyshev's Rule

Consider the student age data

17	**18**	**18**	**18**	**18**	**18**	**19**	**19**	**19**	**19**
19	**19**	**19**	**19**	**19**	**19**	**19**	**19**	**19**	**19**
19	**19**	**19**	**19**	**19**	**19**	**20**	**20**	**20**	**20**
20	**20**	**20**	**20**	**20**	**20**	**21**	**21**	**21**	**21**
21	**21**	**21**	**21**	**21**	**21**	**21**	**21**	**21**	**21**
22	**22**	**22**	**22**	**22**	**22**	**22**	**22**	**22**	**22**
22	**23**	**23**	**23**	**23**	**23**	**23**	**24**	**24**	**24**
25	**26**	**28**	**28**	**30**	**37**	**38**	**44**	**47**	

Color code: within 1 standard deviation of the mean
within 2 standard deviations of the mean
within 3 standard deviations of the mean
within 4 standard deviations of the mean
within 5 standard deviations of the mean

42

Example - Chebyshev's Rule

Summarizing the student age data

Interval	Chebyshev's	Actual
within 1 standard deviation of the mean	≥ 0%	72/79 = 91.1%
within 2 standard deviations of the mean	≥ 75%	75/79 = 94.9%
within 3 standard deviations of the mean	≥ 88.8%	76/79 = 96.2%
within 4 standard deviations of the mean	≥ 93.8%	77/79 = 97.5%
within 5 standard deviations of the mean	≥ 96.0%	79/79 = 100%

Notice that Chebyshev gives very conservative lower bounds and the values aren't very close to the actual percentages.

43

Empirical Rule

If the histogram of values in a data set is reasonably symmetric and unimodal (specifically, is reasonably approximated by a normal curve), then

1. Approximately 68% of the observations are within 1 standard deviation of the mean.
2. Approximately 95% of the observations are within 2 standard deviation of the mean.
3. Approximately 99.7% of the observations are within 3 standard deviation of the mean.

44

Z Scores

The **z score** corresponding to a particular observation in a data set is

$$\text{zscore} = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}$$

The z score is how many standard deviations the observation is from the mean.

A positive z score indicates the observation is above the mean and a negative z score indicates the observation is below the mean.

45

Z Scores

Computing the z score is often referred to as **standardization** and the z score is called a **standardized score**.

The formula used with sample data is

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

46

Example

A sample of GPAs of 38 statistics students appear below (sorted in increasing order)

2.00	2.25	2.36	2.37	2.50	2.50	2.60
2.67	2.70	2.70	2.75	2.78	2.80	2.80
2.82	2.90	2.90	3.00	3.02	3.07	3.15
3.20	3.20	3.20	3.23	3.29	3.30	3.30
3.42	3.46	3.48	3.50	3.50	3.58	3.75
3.80	3.83	3.97				

$$\bar{x} = 3.0434 \text{ and } s = 0.4720$$

47

Example

The following stem and leaf indicates that the GPA data is reasonably symmetric and unimodal.

2	0
2	233
2	55
2	667777
2	88899
3	0001
3	2222233
3	444555
3	7
3	889

Stem: Units digit
Leaf: Tenths digit

48

Example

Using the formula $z\ score = \frac{x - \bar{x}}{s}$ we compute the z scores and color code the values as we did in an earlier example.

-2.21	-1.68	-1.45	-1.43	-1.15	-1.15
-0.94	-0.79	-0.73	-0.73	-0.62	-0.56
-0.52	-0.52	-0.47	-0.30	-0.30	-0.09
-0.05	0.06	0.23	0.33	0.33	0.33
0.40	0.52	0.54	0.54	0.80	0.88
0.93	0.97	0.97	1.14	1.50	1.60
1.67	1.96				

49

Example

Interval	Empirical Rule	Actual
within 1 standard deviation of the mean	≈ 68%	27/38 = 71%
within 2 standard deviations of the mean	≈ 95%	37/38 = 97%
within 3 standard deviations of the mean	≈99.7%	38/38 = 100%

Notice that the empirical rule gives reasonably good estimates for this example.

50

Comparison of Chebyshev's Rule and the Empirical Rule

The following refers to the weights in the sample of 79 students. Notice that the stem and leaf diagram suggest the data distribution is unimodal but is positively skewed because of the outliers on the high side. Nevertheless, the results for the Empirical Rule are good.

10	3
11	37
12	011444555
13	000000455589
14	000000000555
15	000000555567
16	000005558
17	0000005555
18	0358
19	5
20	00
21	0
22	55
23	79

Stem: Hundreds & tens digits
Leaf: Units digit

51

Comparison of Chebyshev' Rule and the Empirical Rule

Interval	Chebyshev's Rule	Empirical Rule	Actual
within 1 standard deviation of the mean	≥ 0%	≈ 68%	56/79 = 70.9%
within 2 standard deviations of the mean	≥ 75%	≈ 95%	75/79 = 94.9%
within 3 standard deviations of the mean	≥ 88.8%	≈99.7%	79/79 = 100%

Notice that even with moderate positive skewing of the data, the Empirical Rule gave a much more usable and meaningful result.

52

4.5: Wrap-up

- Measures of Central Tendency don't tell the whole story.
- Mean & Median just give a number related to the middle of the data
- Need to know something about variability
- Need to also look @ the data to understand shape of data
- Both mean & SD are sensitive to extreme values
  - If data set has outliers, median & interquartile range may be better choice for describing the center & spread

53

Wrap-up

- Be careful of small data set, everything can be misleading
- Not all distributions are normal, or even remotely normal (You can't apply the empirical rule to non normal data sets)
- Outliers can mess up your data & influence you interpretations. Datasets with outliers should be analyzed with caution.

54