Chapter 2
Location, Variability and Process

Section 2.1
Numerical Summaries Based on Deviations

David W. Gerbing

The School of Business
Portland State University

- Numerical Summaries Based on Deviations
  - Summary Statistics
  - The Center as the Mean
  - Variability About the Mean

2.1a
Summary Statistics

Summaries of a Distribution

Numerical summaries of all of the values aid understanding

- The midterm for 58 students has been administered and scored, so how well did the students perform?
- To understand a data set of interest clearly does not involve memorizing all the data values
- Instead, a distribution of the data values of a variable, such as midterm scores, is largely understood through graphs and numerical summary indices, including
  - **Location**: Position of a value relative to the remaining values of the distribution, such as the center, or a value that cuts off a specified percentage of values, such as the lowest 25%
  - **Variability**: How spread out, how different the values in the distribution are from each other

General Summary Statistics

Efficient summary of one or all variables in the data table

- Basic summary statistics are provided in addition to the graphical output for functions that display the shape of the distribution, such as the Histogram() function from Chapter 1
- There is also a dedicated `lessR` function that provides just the numerical summaries of a distribution, `SummaryStats()`, or `ss()`
- The brief version of SummaryStats, referenced with the option `brief=TRUE`, or as `ss_brief()`, provides the following basic summary statistics, explained in this chapter

n: number of data values, miss: number of missing data values

mean: arithmetic average (m), sd: standard deviation (s)

min: minimum value, median: median, max: maximum value

General Summary Statistics II

Efficient summary of one or all variables in the data table

```
> d <-
    Read("https://web.pdx.edu/~gerbing/data/
outlier.csv")
```

```
> ss_brief(Income)
```

```
 n miss      mean       sd      min      mdn      max
10    0  101076.0  77362.6  29750.0  69660.0 273800.0
```

- The long form of SummaryStats() provides more summary statistics, discussed later in this chapter
- The `SummaryStats()` function can also summarize *all* the variables in a data frame, defaulting to `d`

  `> SummaryStats()` or `ss()`

Missing Values

Real data sets often have missing values

- **Missing value**: A cell in the data table for which no data value is present
  - Excel data file: the corresponding cell is blank
  - csv data file: two commas with nothing in between
- R represents a missing value with NA for "not available"
- Ex: If the 5th of 7 values of the variable Age is missing

```
> values(Age)
[1] 48 35 36 59 NA 61 25
```

- The SummaryStats() function accounts for any missing data

```
> ss_brief(Age)

 n miss    mean     sd     min   median     max
 6    1    44.0   14.4    25.0     42.0    61.0
```

2.1b
The Center as the Mean

Summation Notation

To obtain data summaries need to count and sum

- **Sample**: The set of data values for one or more variables to be analyzed, such as for generic variable Y
- A common operation in statistical analysis is to sum a list of numbers, such as the data values of a variable
- Summation symbol is the upper-case Greek letter sigma, $\sum$
- For variable Y, where Y_i indicates the $\mathrm{i^{th}}$ data value, refer to the sum of these data values with $\sum Y_i$
- **Sample Size**: n, number of data values for the variable in the sample
- Ex: Three test scores: $Y_1 = 87, Y_2 = 79, Y_3 = 97$
  - Number of observations: $n = 3$
  - Sum: $\sum Y_i = Y_1 + Y_2 + Y_3 \ = \ 87 + 79 + 97 \ = \ 263$
- These two data summaries are sample size, n, and the sum, Σ

The Arithmetic Mean

Most common indicator of the center of a distribution

- **Sample Mean**: m, sum of the numerical data values for a variable divided by the number of values[1]
- $m = \dfrac{\sum Y_i}{n} = \dfrac{87 + 79 + 97}{3} = \dfrac{263}{3} = 87.67$
- To explicitly indicate the variable to which the sample mean refers, subscript the m with the variable name, such as m_Y
- More formally, the mean defined here is the arithmetic mean, as there are other types of means such as the harmonic mean encountered in Chapter 6

[1] An older, less elegant symbol for the sample mean, developed before computers were invented, is the name of the variable with a bar on top, such as $\bar{Y}$. This symbol is more difficult to produce in a word processor, is inconsistent with other statistical symbols, and is incompatible with both the input and output of computer software for data analysis.

Excel: Arithmetic Mean

- To illustrate the calculation of the sample mean, *m*, manually enter its formula into a worksheet

	B	C
1		Cost
2		Y
3	1	$16.00
4	2	$8.63
5	3	$22.50
6	4	$24.50
7	5	$49.75
8	6	$34.13
9	7	$17.25
10	8	$6.50
11	9	$41.38
12	10	$13.00
13	Sum	$233.64
14	Sample size	10
15	Mean	$23.36

=AVERAGE(C3:C12)

	C	
13	=SUM(C3:C12)	← ΣY_i
14	=COUNT(C3:C12)	← n
15	=C13/C14	

$$m = \frac{\sum Y_i}{n} = \frac{\$16.00 + \$8.63 + \ldots + \$13.00}{10} = \frac{233.64}{10} = \$23.36$$

Weighted Mean

Generalization of the usual arithmetic mean

- **Weighted mean** of Y: Sum of each value Y_i multiplied by its associated weight, w_i, divided by sum of the weights

$$m.wt = \frac{\sum w_i Y_i}{\sum w_i}$$

- Ex: Joe received an 87 and 79 on two midterms and a 97 on the Final, weighted twice as much as either midterm

$$m.wt = \frac{(1)87 + (1)79 + (2)97}{1 + 1 + 2} = \frac{360}{4} = 90$$

- R, for this example `> weighted.mean(Y, c(1,1,2))`
- Arithmetic mean is a weighted mean with weights of 1
  - Denominator: Sum of weights is just sample size n
  - Numerator: Each $\sum w_i Y_i$ term is just $\sum (1) Y_i = \sum Y_i$

Meaning of the Mean

Deviation from the mean

- What is the meaning, the motivation, of summing a list of data values and then dividing by the number values?
- The answer follows from a foundational concept of statistics, the mean deviation
- **Mean deviation** of i^{th} data value for a variable: Distance of the i^{th} data value from the mean,

$$\text{deviation}_i = Y_i - m$$

- **Key Concept**: Statistics is the study of variability, which, for numeric variables, is expressed in terms of deviation scores
- The concept of deviation from the mean is the basis of the assessment of variability for numeric variables, those of ratio or interval quality

Meaning of the Mean

Mean as balance point

- To reveal an important property of the mean, sum the mean deviations for all the data values
- Consider two different distributions of assembly times for a sample of 5 assemblies, for each of two employees

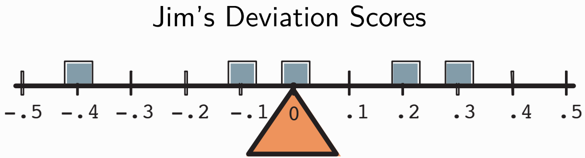
	Jim			Bob		
	Y	mean	dev	Y	mean	dev
1	5.6	6.0	-0.4	4.0	6.0	-2.0
2	5.9	6.0	-0.1	4.0	6.0	-2.0
3	6.0	6.0	0.0	5.0	6.0	-1.0
4	6.2	6.0	0.2	7.0	6.0	1.0
5	6.3	6.0	0.3	10.0	6.0	4.0
Sum			0.0			0.0

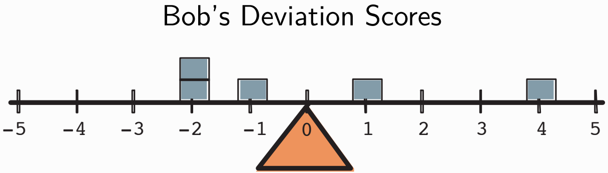
- Each distribution has the same mean, but different variability
- Regardless, the mean deviations always sum to zero

Mean as Balance Point

What being in the middle means

- The mean is in the center of a distribution in the sense that the sum of deviations about the mean is *always* zero





2.1c
Variability About the Mean

Introducing the Standard Deviation

The primary index of variability for numerical data

- Statistics is the tool to formally analyze the naturally occurring variation in the world around us
- To analyze variability we need a statistic that assesses the variability of the values of a variable
- **Key Concept**: The standard deviation is the primary statistic to assess the variability of the values of a continuous variable
- The larger the standard deviation of the values of a variable, the larger their variability
  - 5.6 5.9 6.0 6.2 6.3 : less variability
  - 4.0 4.0 5.0 7.0 10.0 : more variability
  - The second distribution has the larger standard deviation
- The standard deviation of a variable is based on the mean deviations for all of its data values

The Conceptual Basis of the Standard Deviation

Squared Deviation Scores

- Sum of mean deviations cannot summarize variability because the sum is always zero
- To remove the negative signs, square each deviation because the squared deviation is part of equation for the normal curve

	Jim				Bob			
	Y	mean	dev	dev^2	Y	mean	dev	dev^2
1	5.6	6.0	-0.4	.16	4.0	6.0	-2.0	4.00
2	5.9	6.0	-0.1	.01	4.0	6.0	-2.0	4.00
3	6.0	6.0	0.0	.00	5.0	6.0	-1.0	1.00
4	6.2	6.0	0.2	.04	7.0	6.0	1.0	1.00
5	6.3	6.0	0.3	.09	10.0	6.0	4.0	16.00
Sum			0.0	0.30			0.0	26.00

SSY: Sum of squared deviations of Y

or "sum of squares" of Y

From the Sum to the Mean of the Squared Deviations

Remove the confound of sample size

- The sum of squared deviations, SSY, confounds variability with sample size
- That is, typically, the larger the sample the larger is SSY because there are more squared deviations to sum
- A better index of variability than the sum of squared deviations is the corresponding *mean* of the squared deviations
- The statistic of interest here, the standard deviation, is ultimately based directly on this mean of the squared deviations
- However, there is one issue that must be addressed before calculating this mean, the concept of data dependency

Data Dependency

Need sample mean before sample standard deviation

- To calculate the standard deviation requires calculating the deviation scores
  - First calculate the value of one statistical estimate from the data, the sample mean, m
  - Next, from the *same* data, calculate the deviations, $Y_i - m$, with the same m obtained from the first pass of the data
- The second pass through the *same* data introduces a data dependency that uses a value calculated from the first pass
- **Data Dependency**: A data value constrained to be dependent on the remaining data values and any statistical estimates previously computed
- The calculation of the sample standard deviation depends on the prior calculation of the sample mean, m

Illustration of a Data Dependency

Re-cycling through the same data

- Refer back to the previous sum of the three test scores:

  Sum: $\sum Y_i = Y_1 + Y_2 + Y_3 \;=\; 87 + 79 + 97 \;=\; 263$

- After the first pass through the data to calculate the deviation scores, the sum (or the mean) is already known
- If any two data values and the sum are known, the third or remaining data value is fixed, no longer free to vary in the 2^{nd} pass through the same data to calculate the deviations
- In this example, the first two values and the sum are known, so the third value Y_3 is fixed

  $87 + 79 + ?? \;=\; 263$

- Because of this data dependency, the value of the fixed data value is determined and is no longer free to vary

  $Y_3 = \sum Y_i - (Y_1 + Y_2) = 263 - (87 + 79) = 97$

Degrees of Freedom

Correct the sample standard deviation for bias

- **Degrees of freedom** (df) of a statistic: Number of data values *not* constrained by other statistical estimates previously calculated from the *same* data
- df **for the standard deviation**: To account for the data dependency of using the mean from the same data to calculate the mean deviations, $df = n - 1$
- The df can be considered to be the effective sample size after resolving the data dependency
- Now base the mean of the squared deviations, and ultimately the sample standard deviation, on this df
- **Variance**: Mean of the squared deviations based on the degrees of freedom, SSY/df
- This sample variance is denoted s^2, or, to explicitly indicate the variable of interest, such as Y, s_Y^2

Example of Variance

Variance is an average

- To calculate the variance: Square all the mean deviations, sum the squared mean deviations to get SSY, and then divide by the degrees of freedom, $n - 1$

	Jim				Bob			
	Y	mean	dev	dev^2	Y	mean	dev	dev^2
1	5.6	6.0	-0.4	.16	4.0	6.0	-2.0	4.00
2	5.9	6.0	-0.1	.01	4.0	6.0	-2.0	4.00
3	6.0	6.0	0.0	.00	5.0	6.0	-1.0	1.00
4	6.2	6.0	0.2	.04	7.0	6.0	1.0	1.00
5	6.3	6.0	0.3	.09	10.0	6.0	4.0	16.00
Sum			0.0	0.30			0.0	26.00
df				4				4
Mean				0.075				6.500

Variance

Notation and Formulas

Variance and standard deviation

- **Sample Variance**: $s^2 = \dfrac{SSY}{df} = \dfrac{\sum(Y_i - m)^2}{n-1}$
- By definition, the variance is expressed in squared units of the original variable Y, so if Y is measured in inches, then s^2 is in squared inches
- To derive an index of variability that remains in the original units of the measured variable, move to the square root
- **Standard deviation** of Y: Square root of the variance
- Denote the standard deviation by s when computed from data, or, to explicitly indicate the variable, s_Y for variable Y
- Ex: The mean squared deviations, the variance or s^2, for Jim and Bob are 0.075 and 6.500, respectively
  - Jim's standard deviation: $\sqrt{s^2} = \sqrt{0.075} = 0.274$
  - Bob's standard deviation: $\sqrt{s^2} = \sqrt{6.500} = 2.550$

Expression for the Sample Standard Deviation

Understanding and computing

- **Sample standard deviation**: Square root of the average squared deviation score based on degrees of freedom $n - 1$

$$s = \sqrt{\frac{\sum(Y_i - m)^2}{n - 1}}$$

data value
deviation from mean
squared deviation from mean
sum of squared deviations from mean
average of squared deviations from mean based on *df*
square root of average of squared deviations based on *df*

Meaning of the Standard Deviation: Part I

How large are the deviation scores?

- The standard deviation, the amount of variability inherent in the data, directly reflects the sizes of the mean deviations
- **Key Concept**: The standard deviation provides the size of the "typical" deviation, that is, how far the data values tend to be spread out from their mean
- A smaller standard deviation indicates that the data values tend to cluster around the mean
- For the extreme case of no variability, that is, all the data values equal each other, the standard deviation is zero
- A larger standard deviation indicates the data values are more dispersed about the mean, in which case the mean is a *less* effective summary of the distribution of data values

Illustration: Standard Deviation and Mean Deviations

Example of two different standard deviations

- Consider two different distributions of test score percentages that share the *same* mean of 86.1%, 10 scores per distribution
- For Distribution #1, the scores only vary from 84% to 89%
  - Scores: 84 85 85 85 86 86 86 87 88 89
  - Deviations: -2.1 -1.1 -1.1 -1.1 -0.1 -0.1 -0.1 0.9 1.9 2.9
  - Standard Deviation: $s = 1.52$ for $m = 86.1$
- For Distribution #2, the scores vary more, from 77% to 96%
  - Scores: 77 81 81 82 86 87 90 91 92 94
  - Deviations: -9.1 -5.1 -5.1 -4.1 -0.1 0.9 3.9 4.9 5.9 7.9
  - Standard Deviation: $s = 5.67$ for $m = 86.1$
- The standard deviation, here $s_{Test1} = 1.52$ vs $s_{Test2} = 5.67$, summarizes the extent of the variability, as indicated by the size of the corresponding deviation scores

Meaning of the Standard Deviation: Part II

Relation to the normal curve

- As shown, the standard deviation indicates the size of the "typical" deviation from the mean
- The standard deviation provides additional information for the analysis of normally distributed data, the bell-shaped distribution that describes many, many distributions of data across a wide range of topics and applications
- The standard deviation is intimately linked to the normal distribution probabilities
    - For example, approximately 95% of normally distributed data values are within two standard deviations of the mean
- **Key Concept**: The normal distribution/curve and its relation to the standard deviation are fundamental concepts across much of statistical analysis
- These topics are discussed further in the next chapter

Index Subtract 2 from each listed value to get the Slide

- The End