Minitab

LESSON: DESCRIBING DATA *NUMERICALLY*
This lesson includes an overview of the subject, instructor notes, and example exercises using Minitab.

Describing Data *Numerically*

Lesson Overview

Statistics is the discipline concerned with the **optimal acquisition** (where garbage in equals garbage out) and **analysis of data** in order to **model** a population or process.

We can begin to analyze a data set by describing it both numerically and graphically. This lesson considers important **numerical summaries of data**. In this lesson, we will use sample data taken from a large population, and we are only considering quantitative (numeric) data, not qualitative (categorical) data. For the data sets of interest, we will select only one variable of interest; that is, we will be working with univariate data, not bivariate or multivariate data.

Prerequisites

This lesson requires knowledge of basic arithmetic. Symbolic notation will be introduced and used to simplify the formulas for the computation of numerical measurements. In Minitab, computations will be made on single columns of data.

Learning Targets

This lesson teaches students how to:

- Calculate basic numerical measures of **center** for a sample set of data, including its mean, median, and mode
- Determine which measure of center may be more appropriate for a given data set
- Calculate basic numerical measures of **spread** for a sample set of data, including its range, variance, and standard deviation

Time Required

It will take the instructor 30-45 minutes in class to introduce the descriptive statistics formulas. We recommend starting the activity sheet in class so that students can ask the instructor

questions while working on it. The exercises on the activity sheet will take an additional 30-45 minutes, and they can be used as homework or quiz problems.

Materials Required

- Minitab desktop (20 or higher) or Minitab web app
- Minitab worksheet of sample data, entitled ***DescribingDataNumerically_Lesson.mtw***
- Internet access (optional example)

Assessment

The activity sheet contains exercises for students to assess their understanding of the learning targets for this lesson.

Possible Extensions

This lesson provides good introductory examples for students new to statistics. The instructor may want to do the ***Sampling*** lesson first so that students know how data is being selected from the population. The recommended follow-up lesson is ***Describing Data Graphically***.

References

Tranquilizing Sheep – Reaction Time Online Game:
http://www.freeonlinegames.com/game/sheep-reaction

Instructor Notes with Examples

Sample Data

Since we are calculating numerical values on sample data, below is the definition of a sample. There is another lesson devoted entirely to sampling.

Definition: A **sample** is a **subset of subjects** from the population for which observations are actually made.

The numerical values that are calculated on a sample are called **statistics**.

Definition: A **sample statistic** is a **numerical value** characterizing the sample (e.g. center, range, spread, shape). Statistics are typically "English" letters: $\bar{x}$, *s*, or *m*.

There are two branches of statistics that are discussed in introductory statistics courses – **descriptive statistics** and **inferential statistics**. Later lessons will be devoted to inferential statistics.

Definition: Descriptive statistics (also called summary statistics) uses **graphical and/or numerical summaries** for **describing** or **summarizing data** from a **sample**.

- The most common descriptive statistics provide information about a sample's central tendency (mean, median, mode) and variability (variance, standard deviation, range).
- Some graphical methods for displaying and describing data include: dotplot, stem-and-leaf plot, histogram, boxplot, and time series plot (time ordered data). Additional lessons describe these graphs.

Notation: When discussing samples throughout this lesson, we need to have notation for a generic sample of size n. We'll use:

$$x_1, x_2, \ldots, x_i, \ldots, x_n,$$

where

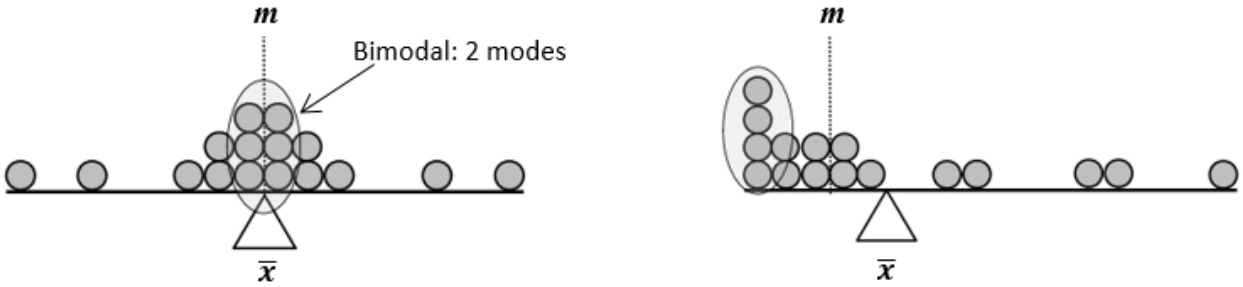
x_1 denotes the numeric value of the first item in the sample,
x_2 denotes the numeric value of the second item in the sample,
⋮
x_i denotes the numeric value of the i^{th} item in the sample,
⋮
x_n denotes the numeric value of the n^{th} item in the sample.

Sample Mean

Definition: The ***sample mean***, denoted by $\bar{x}$, is the arithmetic **average** of the n **data values** in the sample.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

By picture, we can think of the sample mean as the fulcrum point that keeps a weightless ruler, in which each observation is represented by the same weight, in perfect balance.



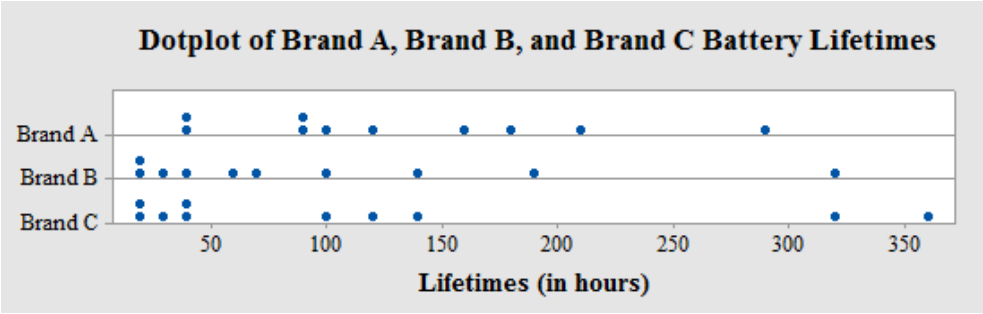
Also noted in each picture are the modes (circled) and location of the medians (*m*). The definitions of these statistics are contained in the following pages.

Example 1

Ten batteries from brands A, B, and C were tested to determine their lifetimes (in hours).

Brand A:	41	289	214	102	38	94	179	87	116	155
Brand B:	39	65	22	64	22	191	99	32	142	317
Brand C:	24	95	139	122	41	360	318	34	43	18

Here are the lifetimes plotted as comparison dotplots in Minitab:



The sample mean lifetimes of battery brands A, B, and C are:

$$\bar{x}_A = \frac{41 + 289 + \cdots + 155}{10} = 131.5 \text{ hours}$$

$$\bar{x}_B = \frac{39 + 65 + \cdots + 317}{10} = 99.3 \text{ hours}$$

$$\bar{x}_C = \frac{24 + 95 + \cdots + 18}{10} = 119.4 \text{ hours}$$

Sample Median

Definition: The ***sample median*** is the **middle** ordered data value if the sample size ***n*** **is odd** and the **average of the middle two** ordered data values if the sample size ***n*** **is even**.

- 50% of the data is less than or equal to the median.
- 50% of the data is greater than or equal to the median.
- The **median** provides a measure which is **less affected by extreme scores than the mean is.**

Example 2

The sample median lifetimes of batteries from brands A, B, and C are:

- Battery brand A **ordered** lifetimes: 38, 41, 87, 94, 102, 116, 155, 179, 214, 289. Since there is an even number of data points, the sample median is: $\frac{102+116}{2} = \mathbf{109}$ **hours**.
- Battery brand B **ordered** lifetimes: 22, 22, 32, 39, 64, 65, 99, 142, 191, 317. The sample median is 64.5 hours.
- Battery brand C **ordered** lifetimes: 18, 24, 34, 41, 43, 95, 122, 139, 318, 360. The sample median is $\frac{43+95}{2} = \mathbf{69}$ **hours**.
- As an additional example, suppose we have battery brand D with ordered lifetimes: 20, 32, 45, 67, 69, 142, 150. Since there are an odd number of data points, the sample median is **67**, the middle ordered data value.

Sample Mode

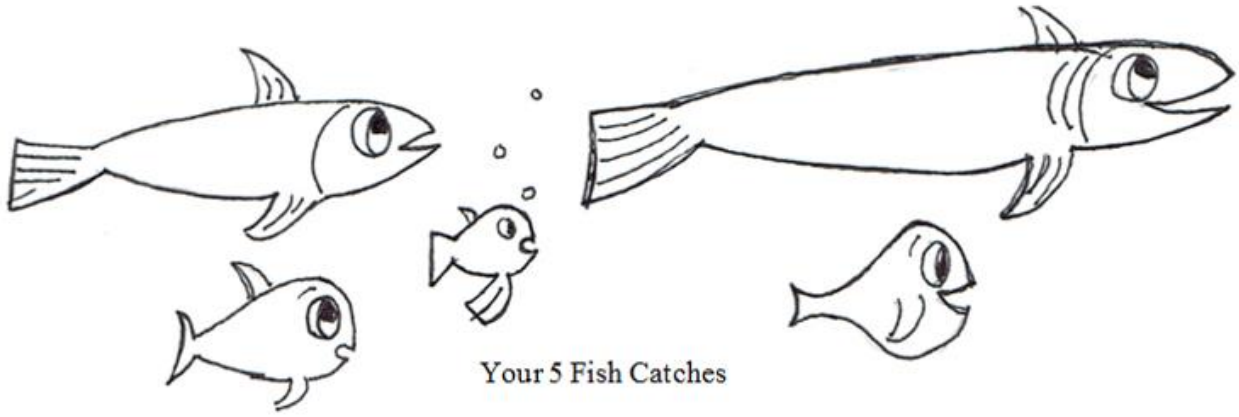
Definition: The **most frequently occurring** sample data value is the ***mode***. There can be more than one mode.

Example 3

Battery brands A and C do not have modes. Battery brand B has mode **22 hours.**

Example 4

You decide to participate in a fishing contest at a local pond. Each contestant must catch 5 fish, and the winner will be determined by the contestant with the "longest" catches overall. Given you caught the following 5 fish below, would you rather the judges use the mean or median to determine longest catches?



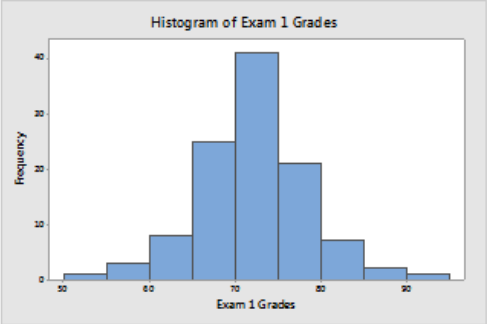
Answer: You want to win the contest! So, hopefully the judges will determine the longest catches using the mean of the five catches. The length of the median catch definitely won't win you the top prize!

Skewed Data

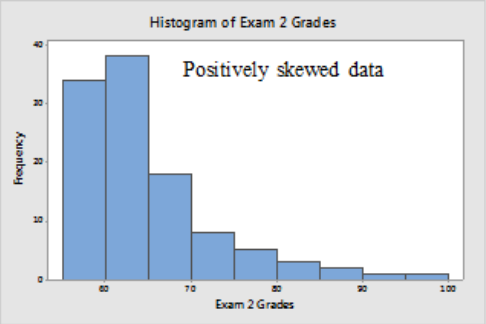
A data set is said to be **skewed** if it is asymmetric, either positively or negatively, as denoted in the figures below.

- For **positively skewed** data, *generally* the mean is greater than the median.
- For **negatively skewed** data, *generally* the mean is less than the median.
- For symmetric data, the mean and median tend to be close to the same value.

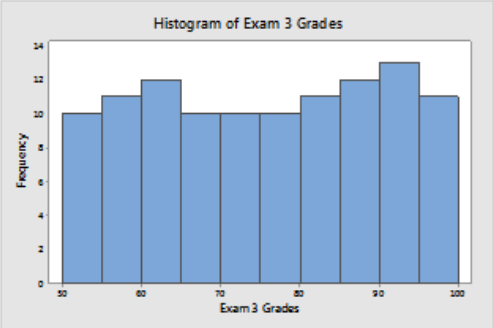
Below are histograms of exam scores for 110 students. Note: All histogram bins contain their left endpoints.



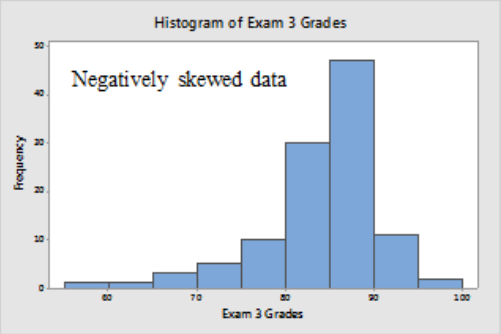
Mean (~71.86) and median (73) are about the same



Mean (~66.32) is greater than the median (62)



Mean (~75.04) and median (75) are about the same



Mean (~83.50) is less than the median (89)

Measures of Spread

We can observe three measures of spread for a sample: the sample range, sample variance, and sample standard deviation.

Sample Range

Definition: The ***sample range*** for a data set is the difference between the largest (maximum) and smallest (minimum) data values in the sample.

Returning to **Example 1** (data below), we can calculate sample ranges for battery brands A, B, and C.

Brand A:	41	289	214	102	38	94	179	87	116	155
Brand B:	39	65	22	64	22	191	99	32	142	317
Brand C:	24	95	139	122	41	360	318	34	43	18

- Sample range for battery brand A lifetimes: 289 – 38 = **251 hours**
- Sample range for battery brand B lifetimes: 317 – 22 = **295 hours**
- Sample range for battery brand C lifetimes: 360 – 18 = **342 hours**

Sample Variance and Sample Standard Deviation

Definition: The ***sample variance*** is the most common estimate of data spread, and we use it in conjunction with the sample mean. It is a measure of deviation from the sample mean $\bar{x}$. For instance, the difference $(x_1 - \bar{x})$ is the **deviation** of the first data point from the sample mean. Hence, we have the ***n* deviations**:

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_i - \bar{x}), \dots, (x_n - \bar{x})$$

Some deviations are negative, while others are positive, and summing the deviations yields 0. In order to make all deviations positive, we **square each deviation**. The sample variance is the **sum of the squared deviations divided by** (***n* – 1)** and is denoted by the symbol $\boldsymbol{s^2}$.

$$\boldsymbol{s^2} = \frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] = \boldsymbol{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Comments regarding the sample variance, s^2:

- The sample variance (s^2) measures the average scatter of the data values about the sample mean. It is the **average** of the squared deviations.
- Why do we divide by $n - 1$ instead of n? Because dividing by $n - 1$ gives us a BETTER ESTIMATE of the true population variance σ^2.
- The **units of s^2 are squared units**. For example, if our data consists of peoples' weights in pounds, s^2 has units pounds squared. To return to the same units as the sample mean, we take the **square root of the sample variance**; it is called the ***sample standard deviation***, and it is denoted by ***s***.

The sample variance and sample standard deviation for battery brand A lifetimes from **Example 1** are computed as follows.

We already computed the sample mean of battery brand A as $\bar{x} = 131.5$ hours. So, the sum of the squared deviations is:

$$(41 - 131.5)^2 + (289 - 131.5)^2 + (214 - 131.5)^2 + \cdots + (155 - 131.5)^2 = 55850.5 \text{ hours}^2$$

Thus,

$$\boldsymbol{s^2} = \frac{55850.5}{10-1} \cong \mathbf{6205.61\ hrs^2}, \text{ and } \boldsymbol{s} = \sqrt{\frac{55850.5}{9}} \cong \mathbf{78.78\ hrs}.$$

Minitab Calculations

All computations we just did by hand in previous examples can be easily calculated in Minitab.

Example 5

Ten batteries from brands A, B, and C were tested to determine their lifetimes (in hours).

Brand A:	41	289	214	102	38	94	179	87	116	155
Brand B:	39	65	22	64	22	191	99	32	142	317
Brand C:	24	95	139	122	41	360	318	34	43	18

Open the Minitab worksheet ***DescribingDataNumerically_Lesson.mtw***. Data for battery brand A, B, and C lifetimes are in columns C1, C2, and C3, respectively.

How to compute descriptive statistics in Minitab:

1. Choose **Stat > Basic Statistics > Display Descriptive Statistics**.
2. In **Variables**, enter '*Brand A' 'Brand B' 'Brand C'*.
3. Click **Statistics** and check **Mean, Standard deviation, Variance, Median, Mode, Minimum, Maximum, Range**, and **N total**.
4. Click **OK** in each dialog box.

The Minitab output is:

Statistics

Variable	Total Count	Mean	StDev	Variance	Minimum	Median	Maximum	Range	Mode	N for Mode
Brand A	10	131.5	78.8	6205.6	38.0	109.0	289.0	251.0	*	0
Brand B	10	99.3	94.4	8907.1	22.0	64.5	317.0	295.0	22	2
Brand C	10	119.4	123.4	15219.6	18.0	69.0	360.0	342.0	*	0

Before beginning the activity sheet, here's a fun riddle for remembering the mean, median, mode, and range.

Hey diddle diddle,
The median's in the middle;
You add and divide for the mean.
The mode is the one that appears the most,
And the range is the difference between.