Lecture 4:
Lesson and Activity Packet

MATH 232: Introduction to Statistics

September 14, 2016

Last time, we discussed:

- Grouped data
  - Class frequencies
  - Class limits
  - Class boundaries
  - Class mark
- Measures of Central Tendency and Dispersion for Grouped Data
  - Mean for Grouped Data
  - Median for Grouped Data
  - Modal class for Grouped Data
  - Range for Grouped Data
  - Variance for Grouped Data
  - Standard Deviation for Grouped Data
- Chebyshev's Theorem
- Empirical rule
- z-scores
- Coefficients of Variation

Questions?

Recall the defining property of the *median* of an ungrouped data set:

Definition 1 (*Median*)

The median *is a value that is exceeded by as many values as it exceeds. (There are as many values greater than the median as there are values less than the median.)*

Actually, the median is just one of several *fractiles*, or (as the Schaum's text calls them) *measures of position*, that divide data into parts as nearly equal as possible. The most commonly used ones are the...

- *Percentiles*, which divide the data into 100 parts;
- *Deciles*, which divide the data into 10 parts;
- *Quartiles*, which divide the data into 4 parts;
- *Median*, which divides the data into 2 parts.

Question 1

Why do we say "as nearly equal as possible"? [Hint: Think about whether there is a way to divide a set of data into for equal parts when the sample size is, say, $n = 27$ or $n = 33$.]

Group Exercise 1 (*2 minutes*)

- *The following data represents temperatures in Fahrenheit taken at 2 p.m. on 12 days of July this year:* $\{90, 75, 86, 77, 85, 72, 78, 79, 94, 82, 74, 93\}$. *Arrange the data in an increasing list.*
- *Draw three vertical lines on your list that separate the data into four parts, each with the same number of data points in it.*

What you did in Exercise 1 was to find the *quartiles* of the data. Just as Definition 1 gave the defining property of the median, the defining properties of the three quartiles, which are called Q_1, Q_2, and Q_3, are:

Definition 2 (*Quartiles*)

- *Q_1 is exceeded by three times as many values as it exceeds, and Q_3 exceeds three times as many values as are exceeded by it.*
- *There are as many values less than Q_1 as there are between Q_1 and Q_2, between Q_2 and Q_3, and greater than Q_3.*
- *Half the data fall between Q_1 and Q_3.*

In a perfect world, all data sets would have quartiles that exactly satisfy this definition. But remember the answer to Question 1: not all data sets have n divisible by 4; also, some data sets have repeating values near quartile positions. **The simple exercise you just did worked because $n = 12$ is evenly divisible by 4, and because it had no repeated data.**

In the general case (even for not-so-perfect data sets), the procedure for finding quartiles is:

Formula 1 (*Finding Quartiles*)

- *Arrange the data in increasing order;*
- *Find the median M;*
- *Q_2 is the same as M.*
- *Write down the set of all values to the left of the median position of the whole set. Q_1 is the median of those.*
- *Write down the set of all values to the right of the median position of the whole set. Q_3 is the median of those.*

Group Exercise 2

Suppose that the following data represents the times, in minutes, of twenty power outages in Berkshire County following last year's big snowstorm in March:

$$\{18, 125, 44, 96, 31, 26, 80, 49, 125, 63, 45, 33, 89, 12, 103, 75, 40, 80, 61, 28\}.$$

Find the quartiles of this data.

Definition 3 (*Interquartile range*)

The interquartile range *is denoted IQR, and is defined as $IQR := Q_3 - Q_1$.*

Group Exercise 3

- *Compute the interquartile range of the data in Exercise 2.*
- *Compute the interquartile range of the data in Exercise 1.*

The information given by the median and the two other quartiles is useful both as a measure of central tendency and as a measure of dispersion; sometimes, this data is presented in graphical form using what's called a *box-and-whisker plot*, often called simply a *box plot*.

Formula 2 (*Constructing a Box Plot*)

- *Find the quartiles of the data;*
- *Draw an axis with a sensible range of values (it doesn't have to be the range of the data; for example, if the data were test scores, I might draw the axis from 0 to 100). This axis can be horizontal or vertical: your choice (Triola's textbook uses horizontal, and Schaum's uses vertical, for example).*
- *Draw a rectangular box extending from Q_1 to Q_3;*
- *Split the box perpendicular to the axis at the position of the median (Q_2);*
- *Put your pencil point on the middle of the edge of the box at Q_1; starting from that point, draw a line parallel to the axis, extending down to the smallest value in the data set.*
- *Put your pencil point on the middle of the edge of the box at Q_3; starting from that point, draw a line parallel to the axis, extending up to the largest value in the data set.*

Example 1

The following are the scores of nine students on a statistics test: $\{66, 73, 74, 79, 82, 86, 88, 90, 94\}$. *The median of this data is 82, the lower quartile is* $Q_1 = 73.5$, *and the upper quartile is* $Q_3 = 89$ *(verify this, if you want!).*

The box plot drawn horizontally looks like:

The box plot drawn vertically looks like:

.

Group Exercise 4

Draw a box plot for the data in Exercise 2 or for the data in Exercise 1.

Another *fractile* that is commonly used and important is the **percentile**. Just like the quartile divides the data into four roughly equal parts, the percentile divides the data into 100 roughly equal parts. It is typically most useful when examining very large data sets.

Formula 3 (*Computing the percentiles P_1, P_2, . . . , P_{100} of a data set*)

- *Arrange the data into a sorted list from lowest to highest;*
- *Let p be the percentile you are trying to compute (if you want to find P_{61}, the 61^{st} percentile, then $p = 61$).*
- *Compute $\frac{p}{100} \cdot n$, where n is the number of values in the data set (the sample size). Call this number you just computed as "i".*
- *If i is an integer (a whole counting number), then the p^{th} percentile is the mean of the data in the i^{th} and $(i+1)^{th}$ **positions** in the sorted list you created from your data set.*
- *If i is not an integer, then the p^{th} percentile is the value of the data point in the in the next highest position of the sorted list you created from your data set.*

Example 2

The following are 40 recorded speeds of cars on $I-91$ south in miles per hour: {68, 68, 72, 73, 65, 74, 73, 72, 68, 65, 65, 73, 66, 71, 68, 74, 66, 71, 65, 73, 59, 75, 70, 56, 66, 75, 68, 75, 62, 72, 60, 73, 61, 75, 58, 74, 60, 73, 58, 75}. To compute the 85^{th} percentile of this data, we find $i = \frac{85}{100} \cdot 40 = 34$. Since $i = 34$ is an integer, the 85^{th} percentile of the data is the mean of the values in the 34^{th} and 35^{th} positions in the sorted list; those two values happen to both be 74 miles per hour, so the 85^{th} percentile is $\frac{74+74}{2} = 74$ miles per hour.

Group Exercise 5

Compute the 65^{th} percentile of the data in Exercise 1 or Exercise 2.

Group Exercise 6

Compute P_{25}, P_{50}, and P_{75} for the data in Exercise 1. How do these values relate to Q_1, Q_3, and the median?

The last fractile to cover is the *decile*. Deciles divide the data into 10 roughly equal parts. They can be computed using their own formulas, but most of the time, they are computed using the formulas that relate them to the corresponding percentiles:

$$D_1 = P_{10} \quad D_2 = P_{20}, D_3 = P_{30}, \ldots, D_9 = P_{90}.$$

Recap

- Fractiles (measures of position)
  - Median (this is not new)
  - Quartiles
  - Deciles
  - Percentiles
- Interquartile range
- Box-and-Whisker Plot

Homework

- Homework 2 due Friday on everything through the end of this packet.