INFO 1301

Prof. Michael Paul
Prof. William Aspray

Monday, September 12, 2016

Topics for Today

- Describing Data
  - Mean, median, and mode
  - Dot plots
  - Histograms

Measures of Central Tendency - Mean

Want a single value that identifies the central position within a data set. Three common choices:

- (Arithmetic) Mean [aka the average] – sum of all the data points divided by the number of data points
- Example 1: 1,2, 4, 4, 5, 9 is a data set. The mean is (1+2+4+4+5+8)/6 = 25/6 or approximately 4.17
- Note that the mean is not one of the numbers that appeared in the data set
- When not to use the mean – when there are outliers
- Example 2: staff salary (thousands of dollars): 15, 18, 16, 14, 15, 15, 12, 17, 90, 95 The mean is $30.7 thousand. But most of the salaries are in the 12-18K range. What happened?

Measures of Central Tendency - Median

- Median = the middle score
- Procedure: order the data in ascending order; if an odd number of data points, pick the middle one; if an even number of data points, average the two middle ones
- Example 3: Data set is 65, 55, 89, 56, 35, 14, 56, 55, 87, 45, 92

Reorder as: 14, 35, 45, 55, 55, **56,** 56, 65, 87, 89, 92

Median is 56.

- Example 4: 11, 19, 26, 24, 17, 3

Reorder as 3,11**,17**,**19**,24,26

Median = (17+19)/2 = 18

Measure of Central Tendency - Mode

- Mode = most common value in the data set
- Often used for categorical data when we want to know the most common category
- Example 5: Transportation to campus: car (10), **bus (23),** walk (8), bicycle (13), skateboard (9)
- Example 6: Favorite baseball team of sample of citizens of New Haven, CT: Yankees (11), Red Sox (11), Cubs (3), Phillies (2), Rangers (1), Dodgers (3), Giants (3) Braves (3) – Which mode is best?
- Example 7: Weight of students in this class to the tenth of a kilogram – 20 different values (even though some close) so no mode – Mode rarely used with continuous data.

3 Measures of Central Tendency

- What is best measure of central tendency?
  - Depends entirely on the application
  - Mode used with (non-ordinal, i.e. nominal) categorical data but not often with numerical data; it is the least used measure.
  - Continuous numerical data seldom uses a mode because there are typically few repeats of the same value in the data set.
  - Debated question whether there can be a mean for ordinal categorical data. Some people report means on Likert scales, but other statisticians question its meaningfulness.
  - Mean, median, and mode – or two of them – sometimes the same. These are cases of high regularity of the data set.

Frequency of Categorical Data - Dotplots

- Dotplots – each dot represents an instance of a particular value of a categorical variable [regular, stacked]
- Dotplot itself provides a visual representation of the frequency of values of a categorical variable
- Favorite color of students in our class (ROYGBIV) – make the dotplot
- Dotplots good with small data sets but not with large data sets – just a big blur with large data sets

Density of Data - Histograms

- For bigger data sets (both categorical and numerical), we can put the individual data points into bins and count the number of items in each bin
- Gives us a sense of the density of data – higher bars where the data are more common.
- Remember the email50 data set discussed in the book in section 1.2.1 (p. 10)? We had data on the number of characters in each of these 50 emails from 2012.
- Place them in bins that are 5000 characters wide.
- You get: 19 (0-5), 12 (5-10), 6 (10-15), 2 (15-20), 3 (20-25), 5 (25-30), …, 0 (55-60), 1 (60-65) [convention on what to do at boundaries]
- See histogram in Fig 1.22 (p. 31) in textbook

Shape of Data

- Histograms give visual clues to skewing
- Right skewed – smaller and longer to right
- Left skewed
- Symmetric – trail off similarly in both directions
- Unimodal, bimodal, multimodal – number of peaks

Class Exercise

- We will calculate the mean, median, and mode of the salaries of Scholars in Residence on the CU Boulder campus last academic year.
- Go to https://www.cu.edu/budgetpolicy/cu-data
- Click on Faculty and Staff
- Click on online database under Faculty and Staff salaries
- At the bottom of the page enter 599 in the box and hit the return key, which will take you to page 599 of the database
- Write the data about Scholars in Residence from p. 599 and p. 600 into Minitab Express (get to p. 600 from p. 599 by using the single arrow at the bottom of the page of the spreadsheet)
- Use Minitab Express to calculate the mean, median, and mode, and make a dot plot and histogram of this data.