Describing Data
Part 1: Centrality and Variability

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

February 6, 2017

Prof. Michael Paul

Descriptive Statistics

- Statistics that summarize a dataset
- Provide information about samples, but not necessarily populations

Two main categories:

- Central tendency
- Variability

Measures of Central Tendency

Three Ms:

- Mean
- Median
- Mode

When to use one over another?

Mean

- Also called the **average**
- Sum of all the data points divided by the number of data points
  - Example: 1,2, 4, 4, 5, 9

  - The mean is (1+2+4+4+5+9)/6 = 25/6, or approximately 4.17
- Note that the mean is not necessarily one of the numbers that appeared in the data set

Mean

- Example: 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?

- When not to use the mean: when there are **outliers**

Median

- The middle value
- 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: 65, 55, 89, 56, 35, 14, 56, 55, 87, 45, 92
  - Reorder: 14, 35, 45, 55, 55, **56,** 56, 65, 87, 89, 92
  - Median is 56.
- Example: 11, 19, 26, 24, 17, 3
  - Reorder: 3,11**,17**,**19**,24,26
  - Median = (17+19)/2 = 18

Median

- Median is less sensitive to outliers
- Salary example:
  15, 18, 16, 14, 15, 15, 12, 17, 90, 95
  - Reorder: 12, 14, 15, 15, **15**, **16**, 17, 18, 90, 95
  - Median is 15.5

Mode

- Most common value in the dataset
- Often used for categorical data
- Example:
  Transportation to campus:
  car (10), **bus (23),** walk (8), bicycle (13), skateboard (9)
- Example:
  Height of students in this class in millimeters:
  20 different values (even though some close) so no mode
  - Mode rarely used with continuous data

Measures of Central Tendency

Which to use depends on the application

- Mode is least common for numerical data, but most common for categorical data
- Median is better when there are outliers
- Mean is better for a small amount of data

Measures of Variability

How much do points deviate from the average?

- Consider two data sets:
  - A is 4,5,6,7,8
  - B is 2,4,6,8,10
- The mean is the same in both cases.
- But you can intuitively say that data set B varies more from its mean that data set A.

Measures of Variability

How much do points deviate from the average?

- Variance
- Standard deviation

Variance

Dataset A: 4,5,6,7,8

- Calculate difference between each point and the mean
  - 4-6 = -2
  - 5-6 = -1
  - 6-6 = 0
  - 7-6 = 1
  - 8-6 = 2
- Square each difference: 4, 1, 0, 1, 4
- Then average the squares: (4+1+0+1+4)/5 = 2

Variance

Dataset A: 4,5,6,7,8

A more accurate way to calculate variance is to divide by one less than the actual number of observations

- Average the squares: (4+1+0+1+4)/4 = 2.5

The math behind this is beyond the scope of this class. We'll allow either method in this class.

Standard Deviation

Standard deviation is the square root of variance

Standard deviation is usually denoted with σ

- and σ^2 is variance

Standard deviation is a more common statistic than variance (but you can get one from the other)

Standard Deviation

Standard deviation can tell you the distribution of values in your dataset. In a typical dataset:

- 60% of data will be within 1 σ of the mean
- 95% of data will be within 2 σ of the mean

Dataset A: 4,5,6,7,8

Standard deviation is 1.6

5, 6, 7 (60%) within 1.6 of the mean

4,5,6,7,8 (100%) within 3.2 of the mean

Percentiles

- Sometimes you want to know how a particular value compares to the entire dataset.
- For ordinal variables, we can create **percentiles**
- Example: SAT scores are given in percentiles. Thus if you are in the 90^{th} percentile on a given variable (e.g. you SAT general math aptitude test), that means your value is higher than 90% of the people who took the test at the same time.

Percentiles

- 25^{th} Percentile: First Quartile (Q1)
- 75^{th} Percentile: Third Quartile (Q3)
- 50^{th} Percentile: Second Quartile (Q2)
  - Also called the median!

- Inter-quartile range (IQR) = Q3 – Q1
  - Tells you how spread out the middle 50% are
  - Another way of measuring variability/spread

Percentiles

7	
21	
31	Q_1
47	
75	
87	Q_2 (median)
115	
116	
119	Q_3
119	
155	
177	

IQR = 119 – 31 = 88

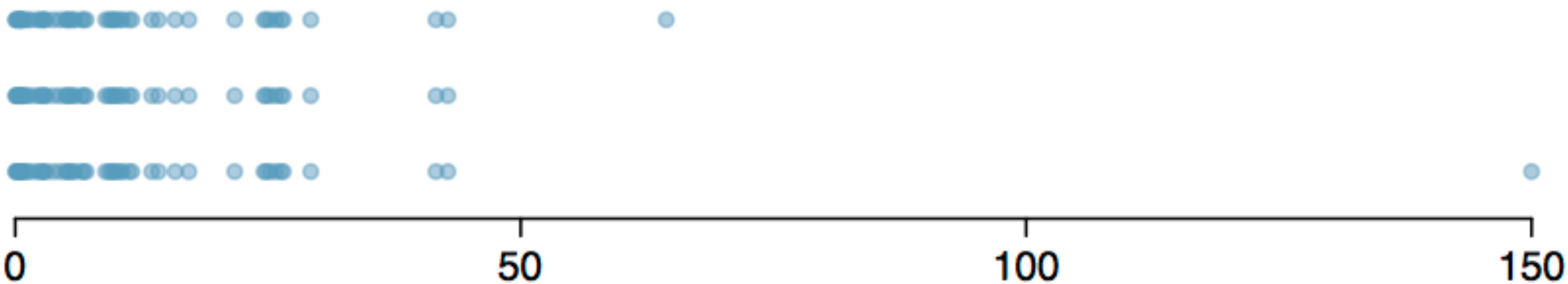
Outliers

Loose definition: values that are unusually far away from other values in a dataset

No single definition, but common definitions:

- More than 2 standard deviations away from the mean (in either direction)
- More than 1.5*IQR below Q1 or more than 1.5*IQR above Q3

Robustness



robust		not robust	
median	IQR	$\bar{x}$	s
6,890	12,875	11,600	13,130
6,768	11,702	10,521	10,798
6,890	12,875	13,310	22,434

Practice

cu-salaries dataset in D2L

5 job categories:

- Regular faculty (professors)
- Research faculty (researchers, postdocs)
- Other faculty (lecturers)
- Classified staff (office staff, laborers)
- Officer/Professional (directors, deans)

Practice

1. In which job categories is the mean larger than the median? In which is it smaller? In which is it about the same? Why?
2. How many outliers are there for each job category? Are the outliers high or low?
3. Which department has the highest median salary? Which has the lowest?
4. Look at each department’s salary histogram and say whether it appears to be left-skewed, right-skewed, or symmetric.