HUMBEHV 3ST3
Measures of Central Tendency & Variability
Week 2

Prof. Patrick Bennett

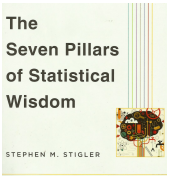
Describing Data Distributions

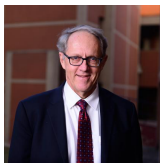
- Often we wish to summarize data distributions, rather than simply illustrating them in graphs histograms
- Two basic descriptions of a distribution include its "middle" (central tendency) and "how spread out it is" (variability)

Part 1 - Central Tendency

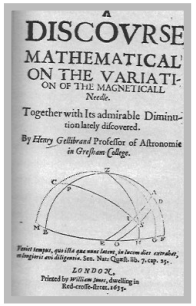
- Measures of central tendency attempt to identify a "typical" score
- How should we define a "typical" score?
- Common measures:
  - mode (most common)
  - median (middle score; 50th percentile)
  - mean (average)

Does statistical "aggregation" make sense?



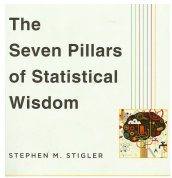


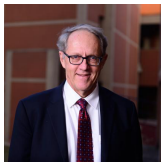
The taking of a mean of any sort is a rather radical step in an analysis. In doing this, the statistician is discarding information… the individuality of each observation is lost… The strong temptation is, and has always been, to select one observation thought to be the best rather than corrupt it by averaging with others of suspected lesser value.



Gillebrand (1635) One of the first texts to use the arithmetic mean to summarize data.

Statistical “aggregation” in antiquity



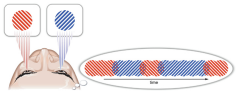


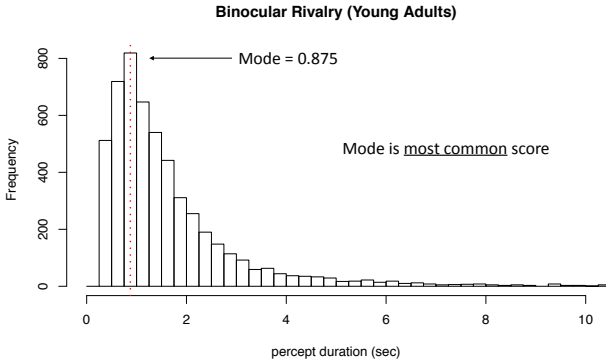
Thucydides

“Ladders were made to match the height of the enemy’s wall, which they measured by the layers of bricks… These were counted by many persons at once; and though some might miss the right calculation, **most** would hit upon it… The length required for the ladders was thus obtained…” (p 30) From Thucydides’ *History of the Peloponnesian War*

5

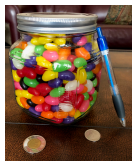
Example of Mode – Binocular Rivalry Times

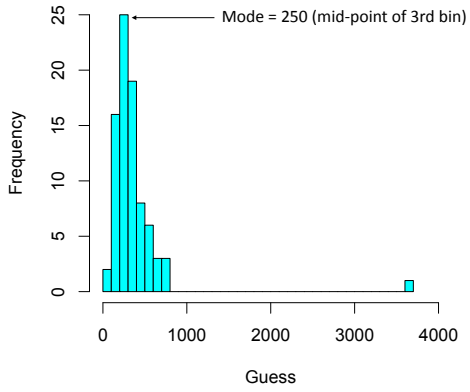




6

Example of Mode – Jelly Bean Estimates

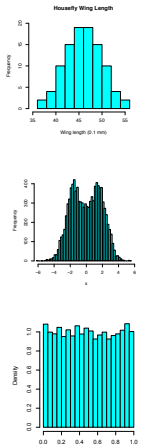




7

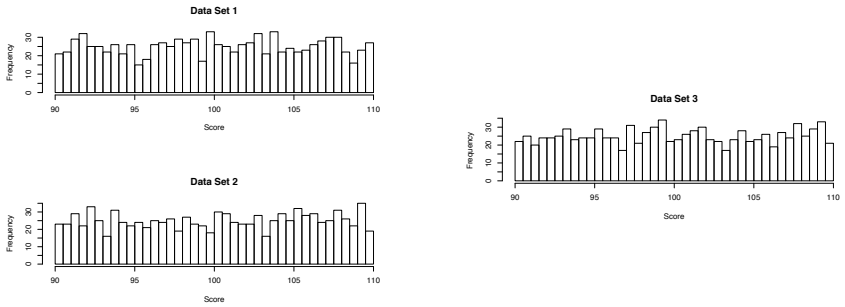
Mode - Most Common Score

- The mode is the most common/frequent score
- If 2 adjacent scores occur with equal frequency:
  - mode = average of those two scores
- If 2 non-adjacent scores occur with equal frequency:
  - distribution is bi-modal
  - report both numbers
- If range of number occur with nearly equal frequency:
  - mode is ill-defined
  - report as: “the mode fell within the range of X to Y”



8

Mode may be poorly defined



In some cases, small fluctuations in frequencies can produce BIG changes in mode

Median = Middle Score

Odd number of scores (N=11)

> scores:
92 95 103 108 79 111 106 99 100 108 85

> sorted scores:
79 85 92 95 99 100 103 106 108 108 111

> median(scores):
100

Median Location = (N+1)/2 = (11+1)/2 = 6

Median = Middle Score

Even number of scores (N=10)

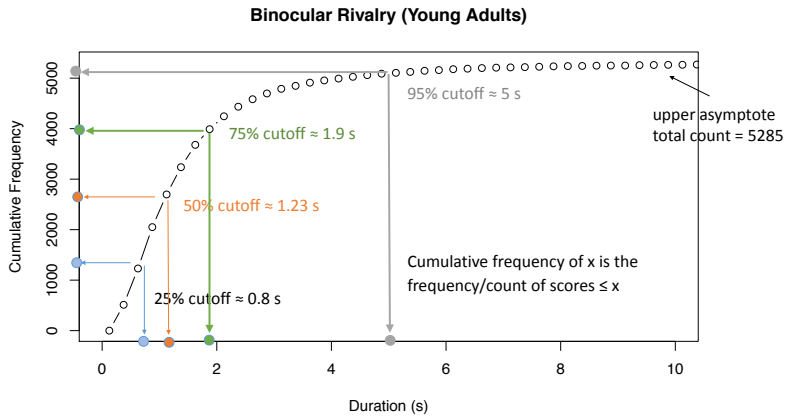
> scores:
92 95 108 79 111 106 99 100 108 85

> sorted scores:
79 85 92 95 99 100 106 108 108 111

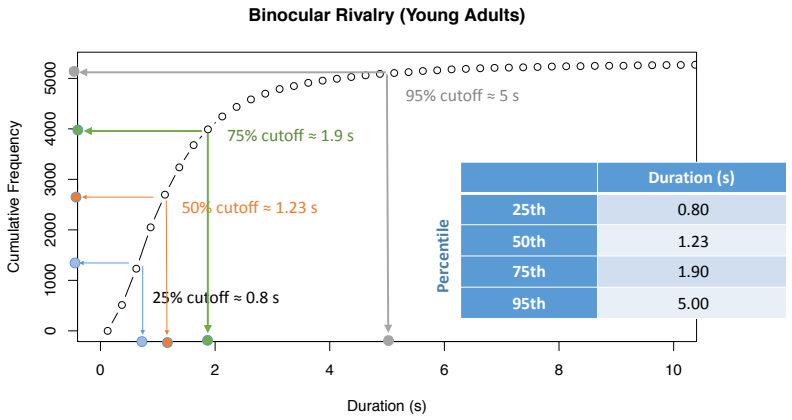
> median(scores):
(99+100)/2 = 99.5

Median Location = (N+1)/2 = (10+1)/2 = 5.5

Cumulative Frequency Plot & Percentiles



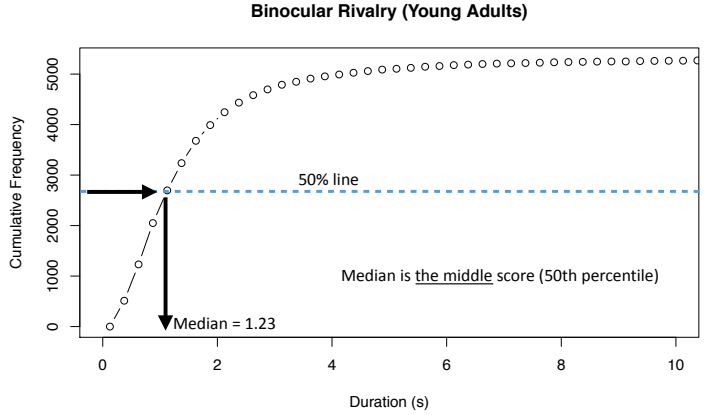
Cumulative Frequency Plot & Percentiles



Percentile	Duration (s)
25th	0.80
50th	1.23
75th	1.90
95th	5.00

13

Cumulative Frequency Plot (Median = 50th Percentile)



14

Median is Robust to Outliers

Odd number of scores (N=11)

replace 108 with 99,999

> scores:

92 95 103 108 79 111 106 99 100 99,999 85

> sorted scores:

79 85 92 95 99 100 103 106 108 111 99,999

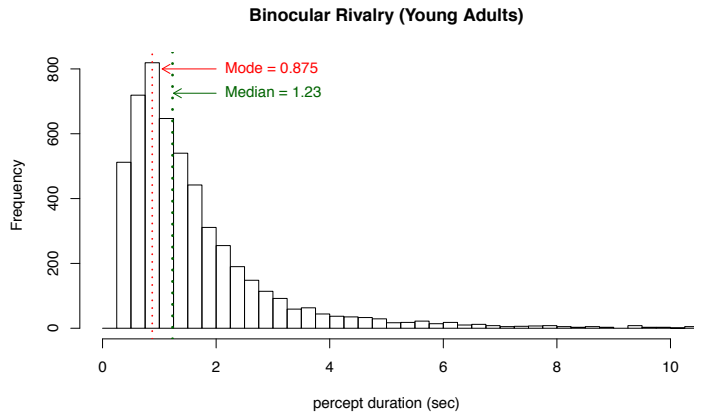
> median(scores):

100

Median Location = (N+1)/2 = (11+1)/2 = 6

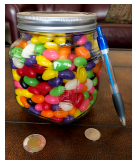
15

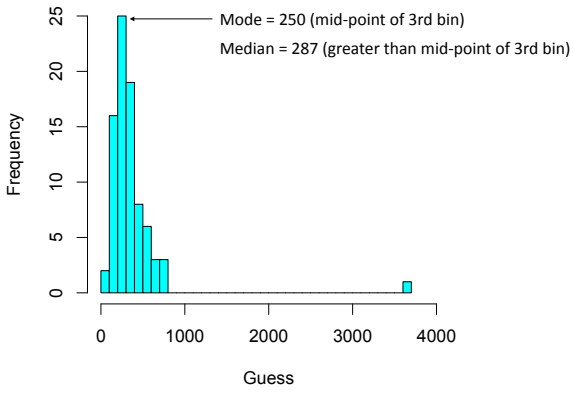
Mode vs. Median



16

Mode vs. Median – Jelly Bean Estimates





Interlude: Summation Notation

- Suppose you have a set of n scores, X:
  - $X_1, X_2, X_3, \ldots, X_n$
  - represent an arbitrary score as X_i
- Following notation represents the summation of all n scores:

Sigma → $\sum_{i=1}^{n} X_i = (X_1 + X_2 + X_3 + \cdots + X_n)$

Summation (continued)

- We often use a compact form to represent summing over all scores:

$$\sum X$$

equivalent!

$$\sum_{i=1}^{n} X_i = (X_1 + X_2 + X_3 + \cdots + X_n)$$

Summation (continued)

- Operations to the right of Sigma are performed on individual Y's before summation

$$\sum X^2 = (X_1^2 + X_2^2 + X_3^2 + \cdots + X_n^2)$$

Summation (continued)

• Brackets indicate operations performed <u>after</u> summation

$$\left(\sum X\right)^2 = (X_1 + X_2 + X_3 + \cdots + X_n)^2$$

Mean

• most commonly used measure of central tendency
• it is the <u>average</u> score:
  - (sum of all scores) divided by (number of scores)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{(X_1 + X_2 + X_3 + \cdots + X_n)}{n}$$

early illustration of calculating a mean

1.7 Köbel's depiction of the determination of the lawful rod. (*Köbel 1522*)

The Seven Pillars of Statistical Wisdom, Stephen M Stigler

Why use the mean?

• consider the following game:
  - we both inspect/study a set of numbers (i.e, the data)
  - on each trial, I randomly pick a number from the set
  - you guess the value
  - repeat for many trials
  - define "error" as difference between guess & number
  - your goal: minimize the total error (summed across trials)
• what is the best strategy?
  - Answer: guess the mean on every trial
  - the mean minimizes the sum of errors & sum of squared errors
  - mean is "least squares" estimate of a "typical" score

Mean is NOT Robust to Outliers

Replacing score of 108 with 99,999 had no effect on median:

> scores:
92 95 103 108 79 111 106 99 100 99,999 85 ~~108~~

> sorted scores:
79 85 92 95 99 100 103 106 108 111 99,999

> median(scores):
100

but has a BIG effect on the mean:

with 108, mean = 98.7; with 99,999, mean = 9179.7

Trimmed Means

- To make the mean less sensitive to extreme values, we can trim a certain percentage of values off of the tails
  - Example: scores = {2, 2, 3, 5, 6, 7, 8, 8, 9, 501}
    - mean = 55.1
  - Now trim 10% off tails at <u>both</u> ends of sorted scores:
  - trimmed scores = {~~2~~, 2, 3, 5, 6, 7, 8, 8, 9, ~~501~~}
    - 10% trimmed mean = 6
- Amount of trimming varies across applications/situations:
  - e.g., 20% trimmed mean removes upper & lower 20% of scores

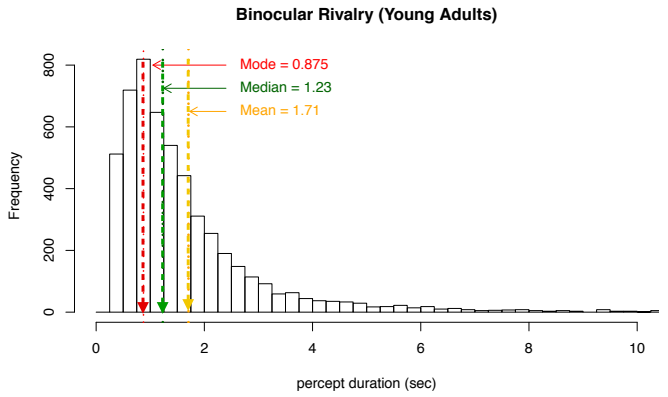
	sorted data:	10% trimming	20% trimming	40% trimming
1	87.40	~~87.40~~	~~87.40~~	~~87.40~~
2	87.64	~~87.64~~	~~87.64~~	~~87.64~~
3	89.17	89.17	~~89.17~~	~~89.17~~
4	89.56	89.56	~~89.56~~	~~89.56~~
5	91.92	91.92	91.92	~~91.92~~
6	92.23	92.23	92.23	~~92.23~~
7	93.74	93.74	93.74	~~93.74~~
8	94.42	94.42	94.42	~~94.42~~
9	94.83	94.83	94.83	94.83
10	96.69	96.69	96.69	96.69
11	97.04	97.04	97.04	97.04
12	97.28	97.28	97.28	97.28
13	99.25	99.25	99.25	~~99.25~~
14	104.31	104.31	104.31	~~104.31~~
15	107.57	107.57	107.57	~~107.57~~
16	109.55	109.55	109.55	~~109.55~~
17	110.38	110.38	~~110.38~~	~~110.38~~
18	112.68	112.68	~~112.68~~	~~112.68~~
19	113.52	~~113.52~~	~~113.52~~	~~113.52~~
20	115.85	~~115.85~~	~~115.85~~	~~115.85~~

10% trimmed mean → 98.79

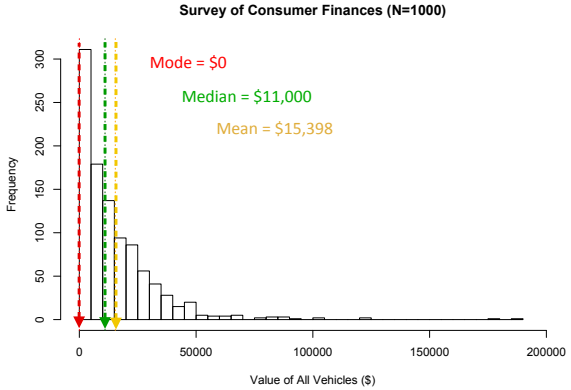
20% trimmed mean → 98.24

40% trimmed mean → 96.46

Mode, Median, & Mean

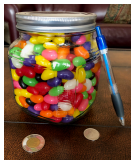


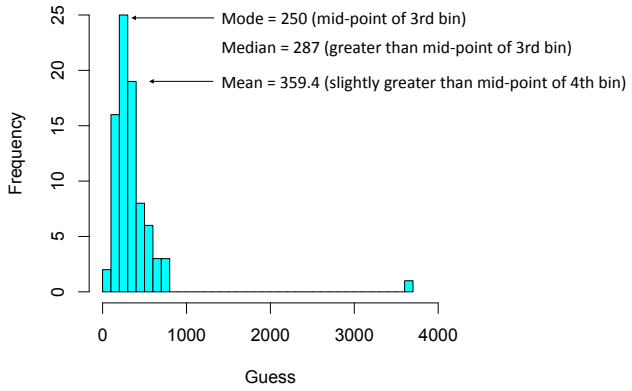
Mode, Median, & Mean



29

Mode, Median, Mean – Jelly Bean Estimates





30

Mode – Advantages & Disadvantages

- Advantages
  - robust to outliers (extreme scores)
  - value actually appears in the data
  - represents the greatest probability of subjects having a score
  - can be found for nominal data
    - e.g., mode of household pet type "dog"; no analogous mean or median
- Disadvantages
  - depends on how we bin scores
  - can be poorly defined & unstable for flat distributions

31

Median – Advantages & Disadvantages

- Advantages
  - robust to outliers (extreme scores)
  - can be calculated even with flat distributions
  - good index of "typical" score in skewed distributions
- Disadvantages
  - no mathematical formula for the median
    - difficult to use median in mathematical derivations/equations
  - in some situations, between-sample variability is greater for median than mean (i.e., median less stable than mean)

32

Mean – Advantages & Disadvantages

- Advantages:
  - in some situations, between-sample variation is less for sample means than sample modes & medians
    - i.e., means are more stable across samples
  - mean is easy to use in statistical formulas
- Disadvantages:
  - value may not actually exist in the data
  - less robust than median to extreme values
    - use trimmed means instead?

Part 1 - Central Tendency (summary)

- Mode, Median, Mean
  - methods of calculation
  - advantages & disadvantages
- Summation Notation
- Trimmed Means