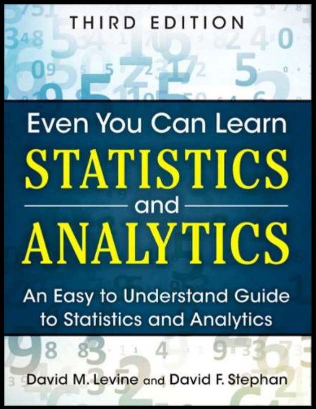
IMGD 2905

Descriptive Statistics

Chapter 3



THIRD EDITION

Even You Can Learn

STATISTICS

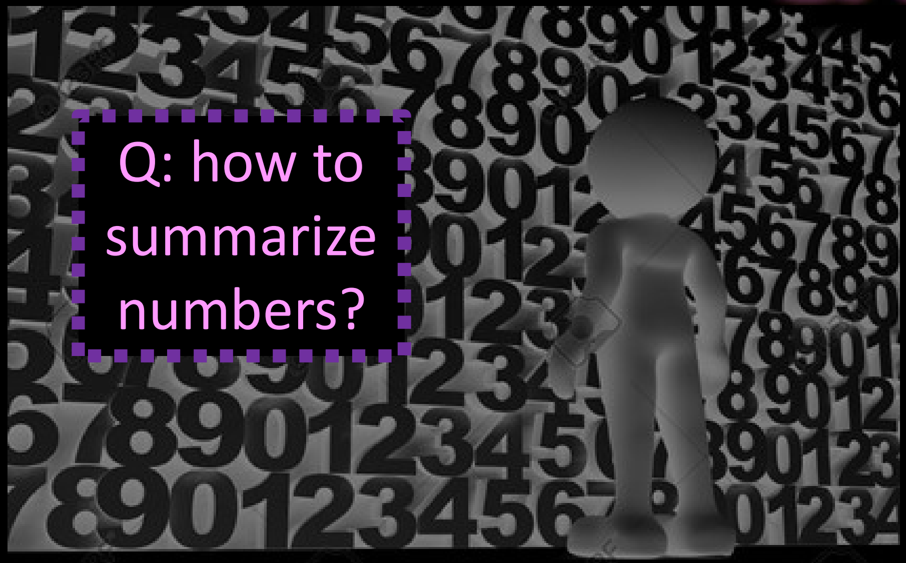
and

ANALYTICS

An Easy to Understand Guide to Statistics and Analytics

David M. Levine and David F. Stephan

Summarizing Data



Q: how to summarize numbers?

- With lots of playtesting, there is a lot of data (good!)
- But raw data is often just a pile of numbers
  - Rarely of interest, or even sensible

Summarizing Data GA

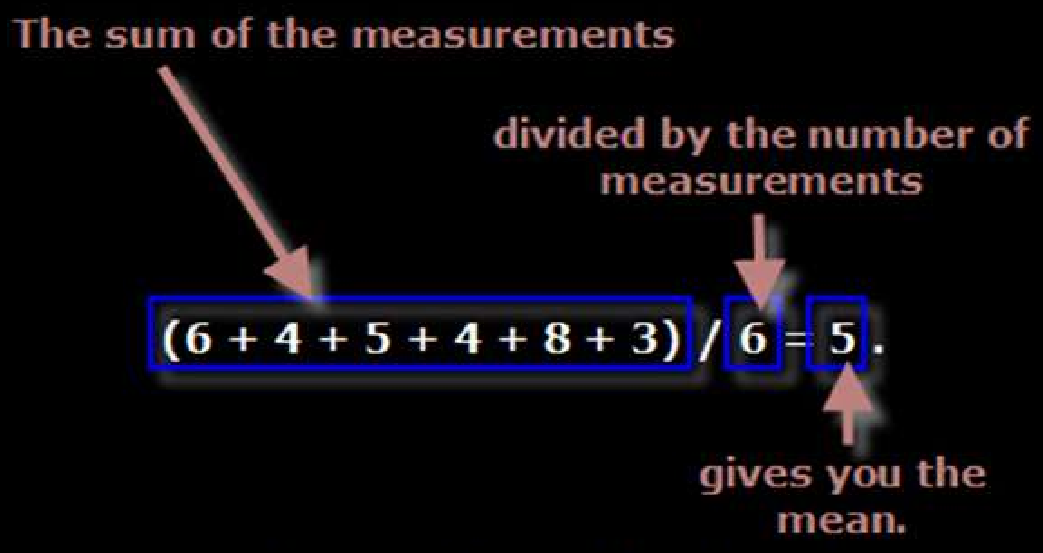
Measures of central tendency

Groupwork

4 3 7 8 3 4 22 3 5 3 2 3

- Indicate *central tendency* with **one** number?
- What are *pros* and *cons* of each?

Measure of Central Tendency: Mean



The sum of the measurements

divided by the number of measurements

$(6 + 4 + 5 + 4 + 8 + 3) / 6 = 5$.

gives you the mean.

http://www.cdn.sciencebuddies.org/Files/463/9/MeanEquation.jpg

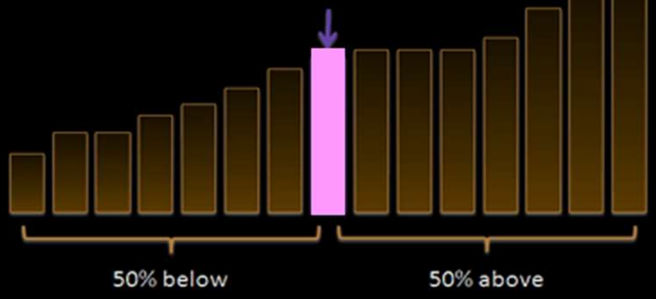
- Aka: "arithmetic mean" or "average"

=AVERAGE(range)

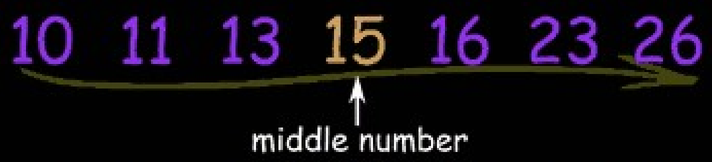
=AVERAGEIF() – averages if numbers meet certain condition

Measure of Central Tendency: Median

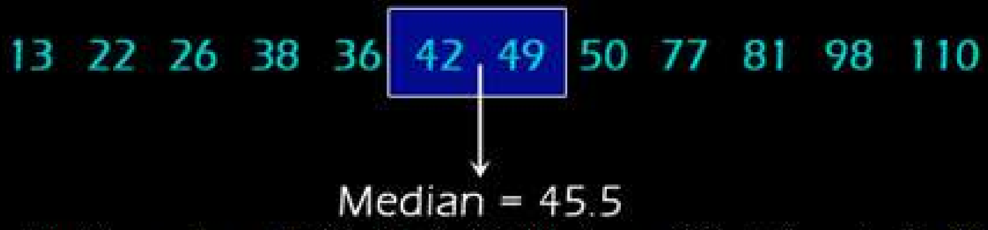
- Sort values low to high and take middle value



https://betterexplained.com/wp-content/uploads/average/median.png



https://www.mathsisfun.com/definitions/images/median.gif



http://www.nedarc.org/statisticalHelp/basicStatistics/measuresOfCenter/images/median.gif

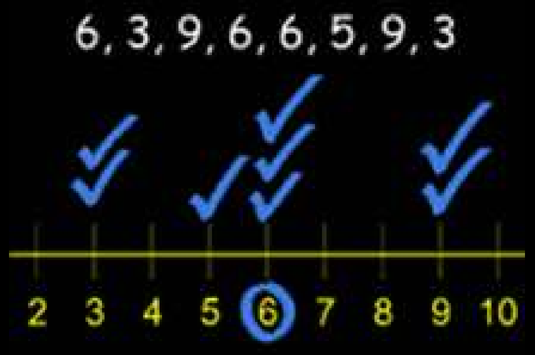
=MEDIAN(range)

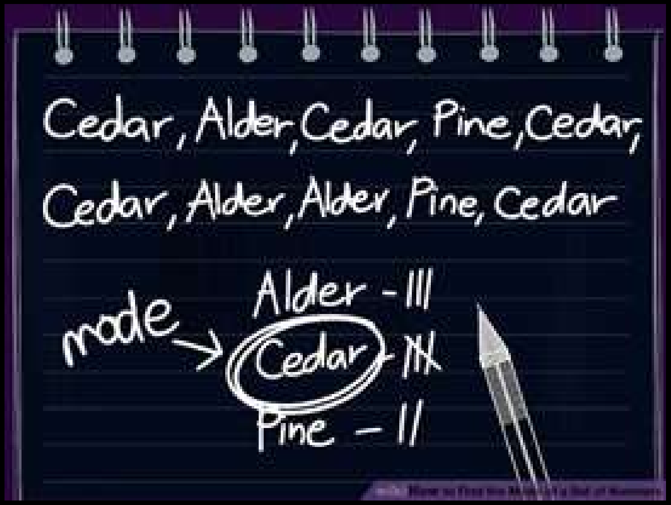
Measure of Central Tendency: Mode

- Number which occurs most frequently
- Not too useful in many cases

→ Best use for categorical data

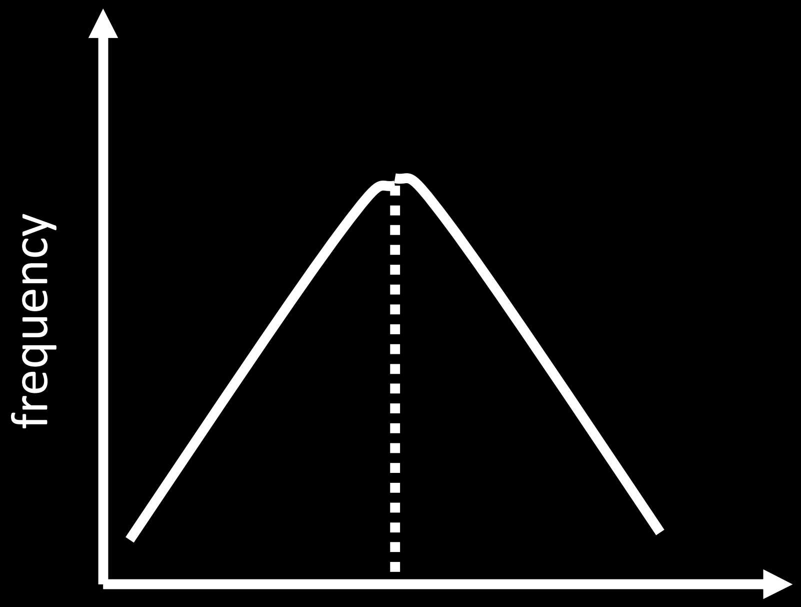
  - e.g., most popular Hero group in Heroes of the Storm



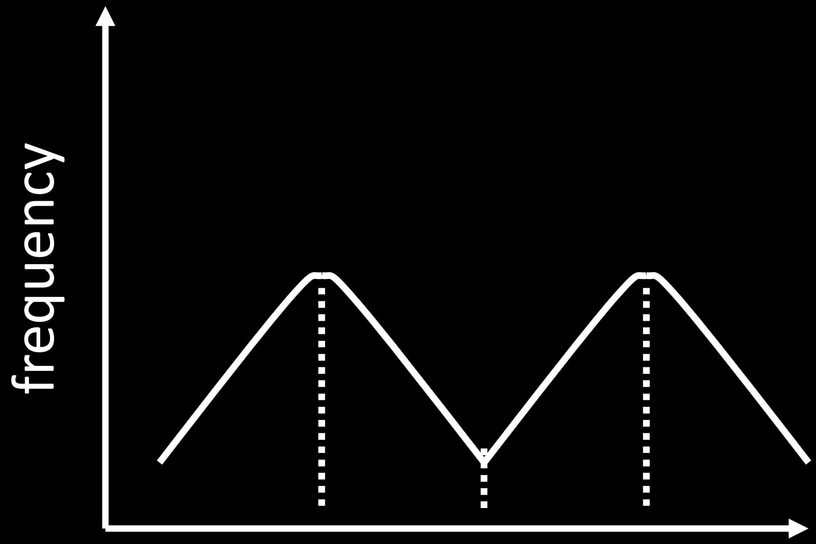

Cedar, Alder, Cedar, Pine, Cedar,
Cedar, Alder, Alder, Pine, Cedar
mode
Alder - III
Cedar - ##
Pine - II

=MODE( )

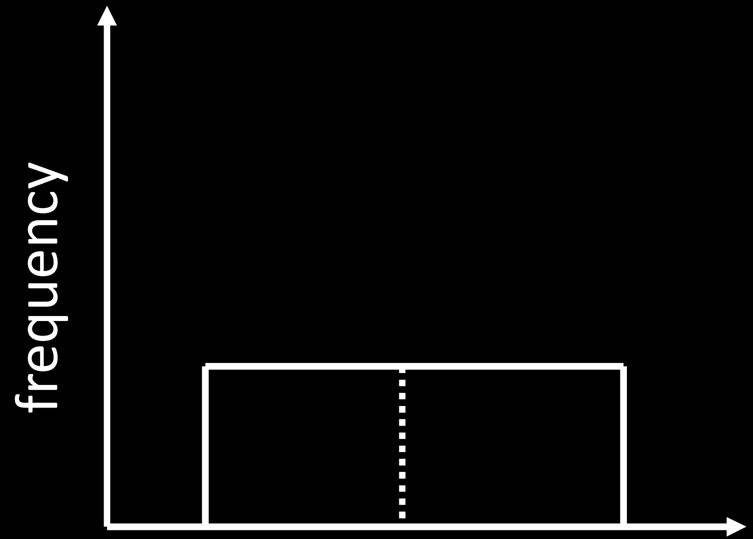
Mean, Median, Mode?



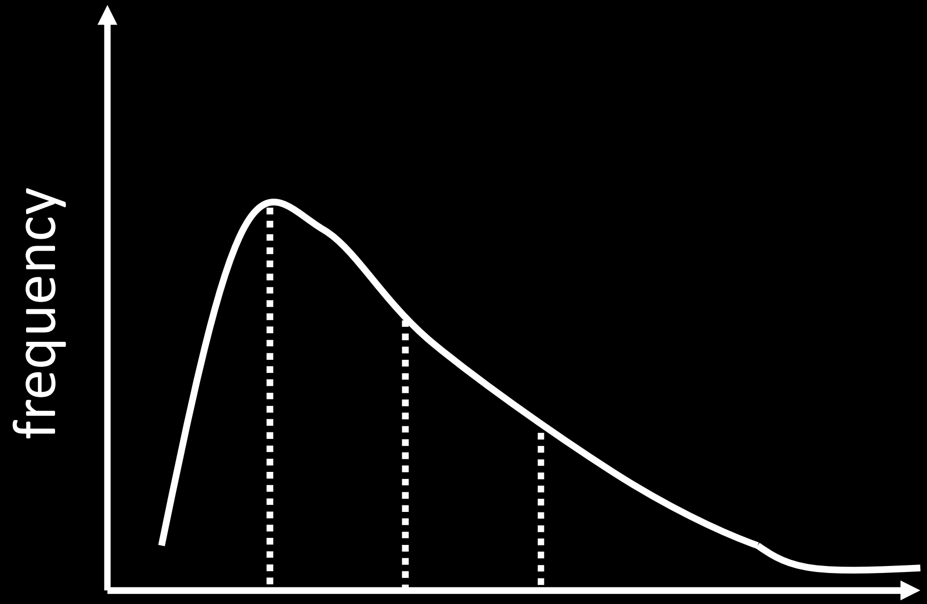
Mean, Median, Mode?



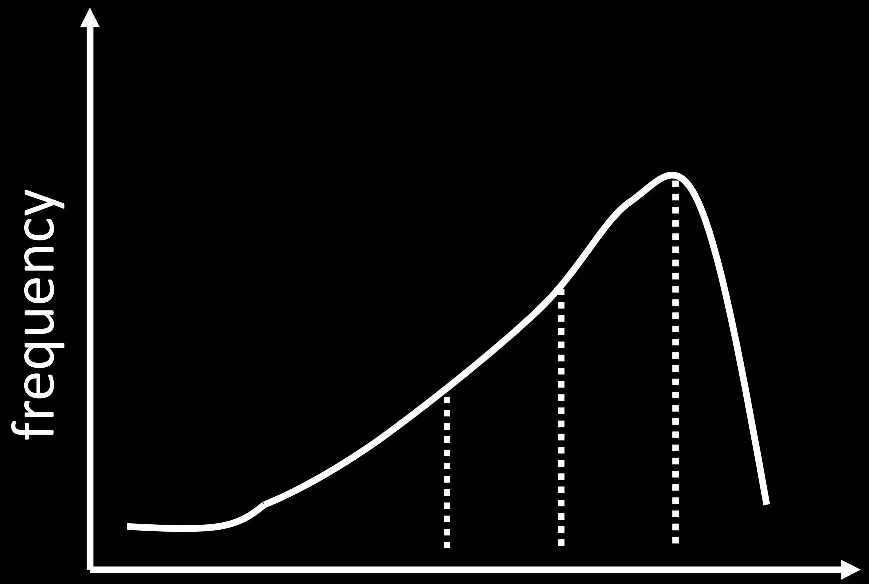
Mean, Median, Mode?



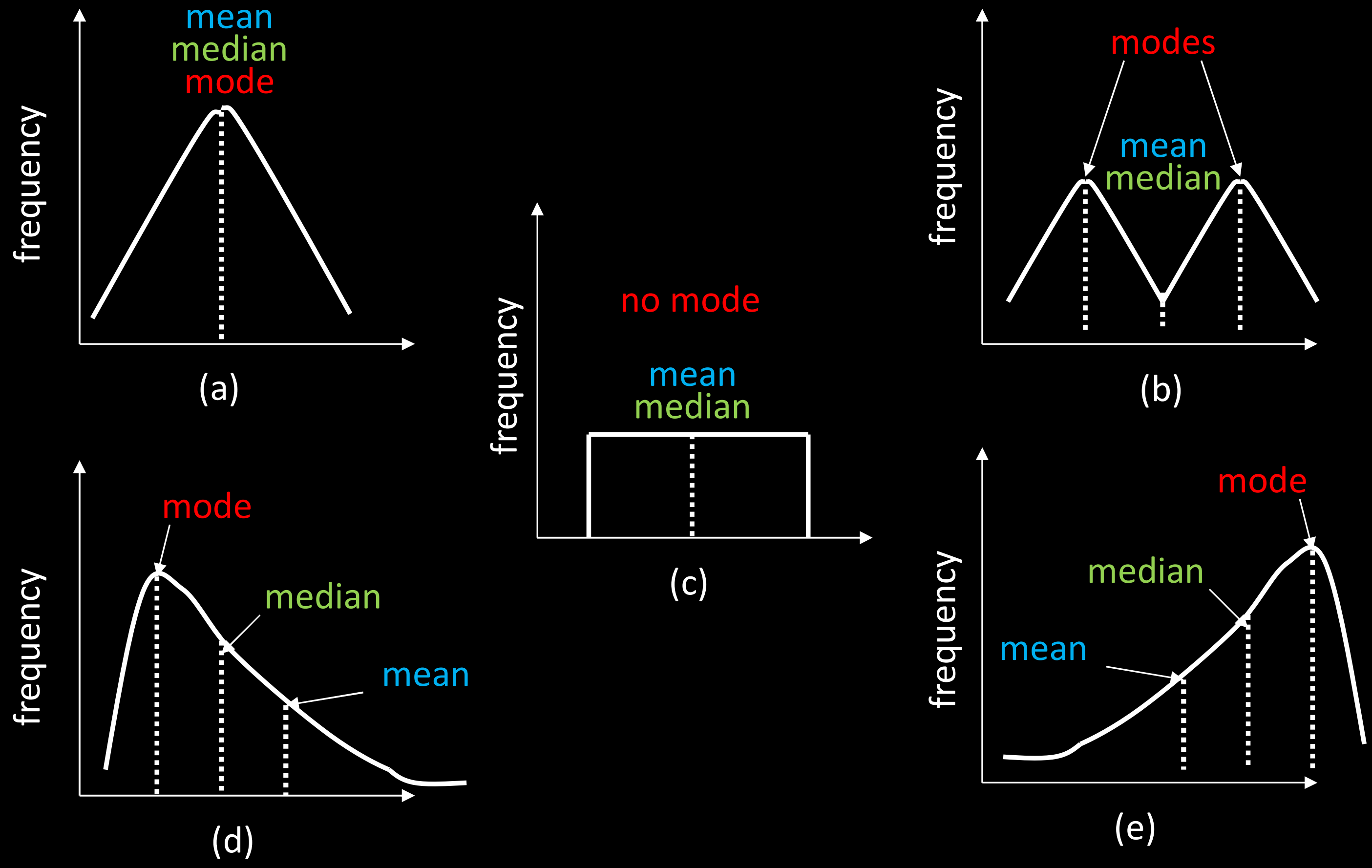
Mean, Median, Mode?



Mean, Median, Mode?



Mean, Median, Mode?



Which to Use:
Mean, Median, Mode?

Which to Use:
Mean, Median, Mode?

- Mean many statistical tests that use sample
  - Estimator of population mean
  - Uses all data

Which to Use: Mean, Median, Mode?

- Median is useful for skewed data
  - e.g., income data (US Census) or housing prices (Zillo)
  - e.g., *Overwatch* team (6 players): 5 people level 5, 1 person level 275
    - Mean is 50 - not so useful since no one at this level
    - Median is 5 – perhaps more representative
  - Does not use all data. “Resistant” to extremes (e.g., 275)
  - But what if were project scores? Hard to “bring up” grade

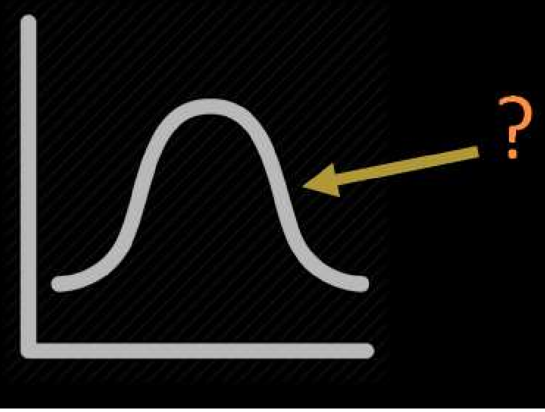
Which to Use:
Mean, Median, Mode?

- Mode is useful primarily for categorical data only
  - Most played League champion, most popular maze, …

Other Measures of Position

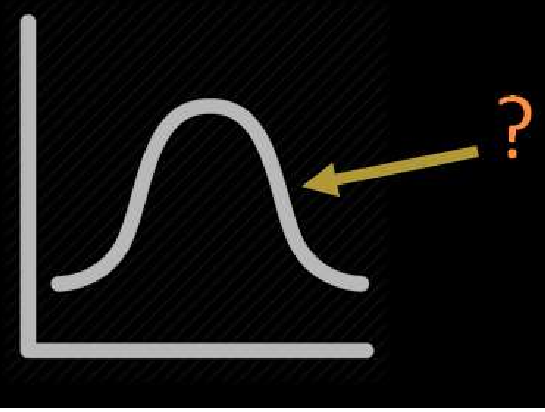
- May not always want center
  - e.g., want to know best LoL Champions

- What other positions may be desired?

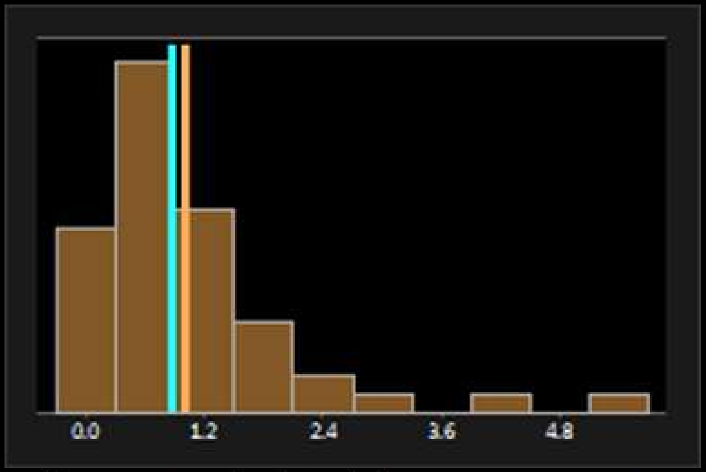


Other Measures of Position

- Maximum / Minimum
  - Not discussed more
- Trimmed Mean
- Quartiles
- Percentiles



Trimmed Mean



http://support.minitab.com/en-us/minitab/17/histogram_mean_vs_trimmed_mean.png

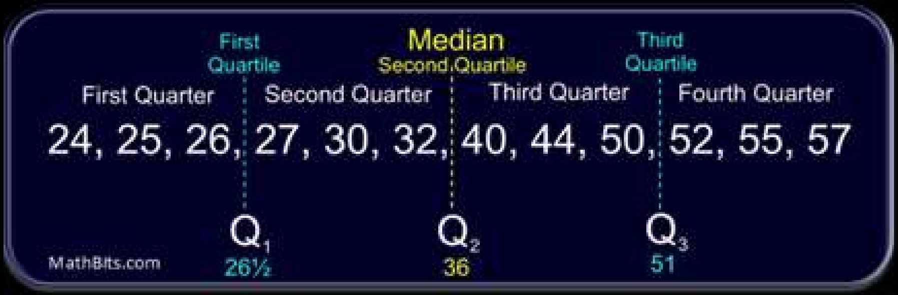
- Take “trimming” off top and bottom (typically 5% or 10%)
  - Reduces effects of extreme values, like median

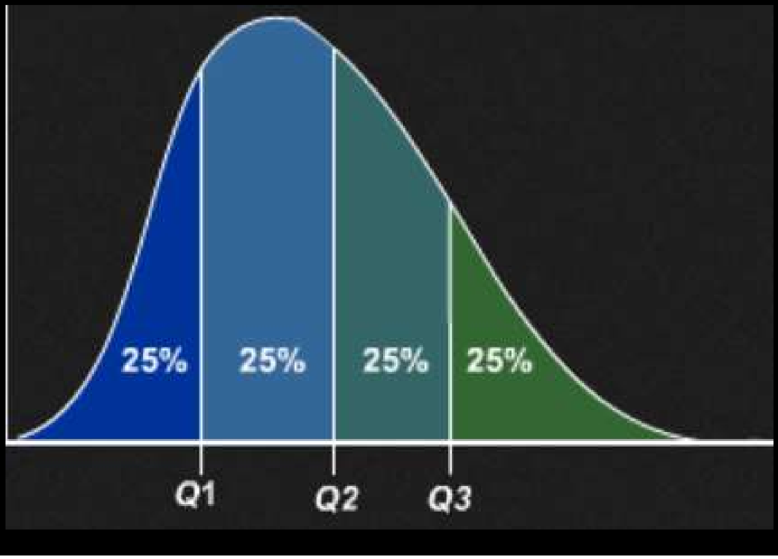
`=TRIMMEAN(array,percent)`

Quartiles

- Sort values
- First quartile (Q1) is 25% from bottom
- Third quartile (Q3) is 75% from bottom
- (What is second quartile?)

```
=QUARTILE(array,n)
```



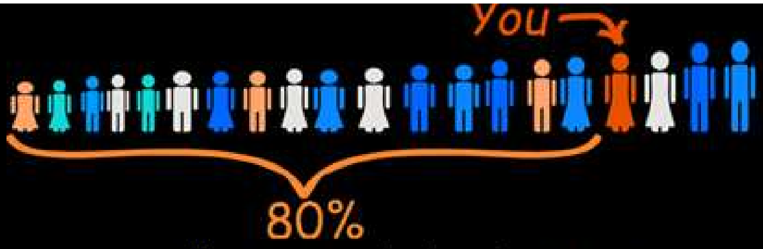


Percentiles

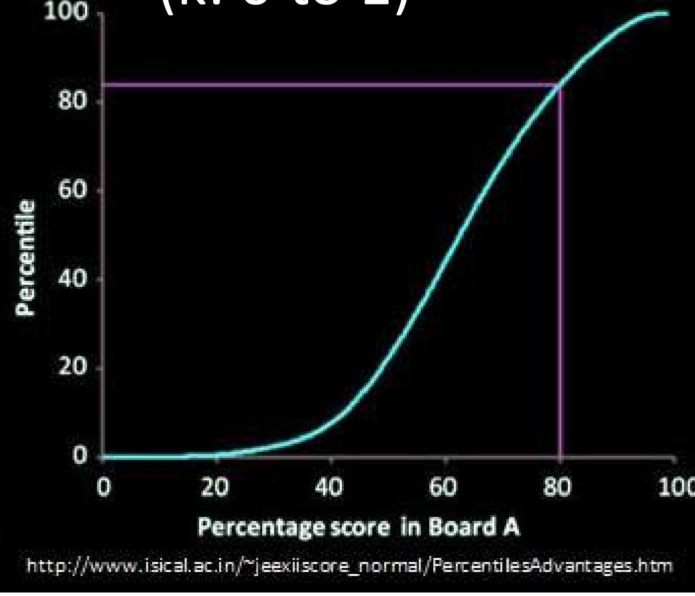
- Generalization of quartiles
- *N*th percentile is data point *n*% from bottom of data
- Interpolate as for first quartile

=PERCENTILE(array,k)

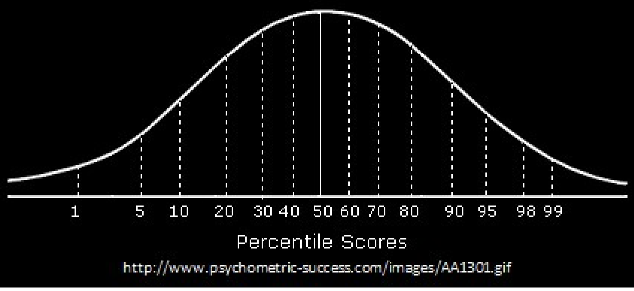
(k: 0 to 1)



https://www.mathsisfun.com/data/images/percentile-80.svg

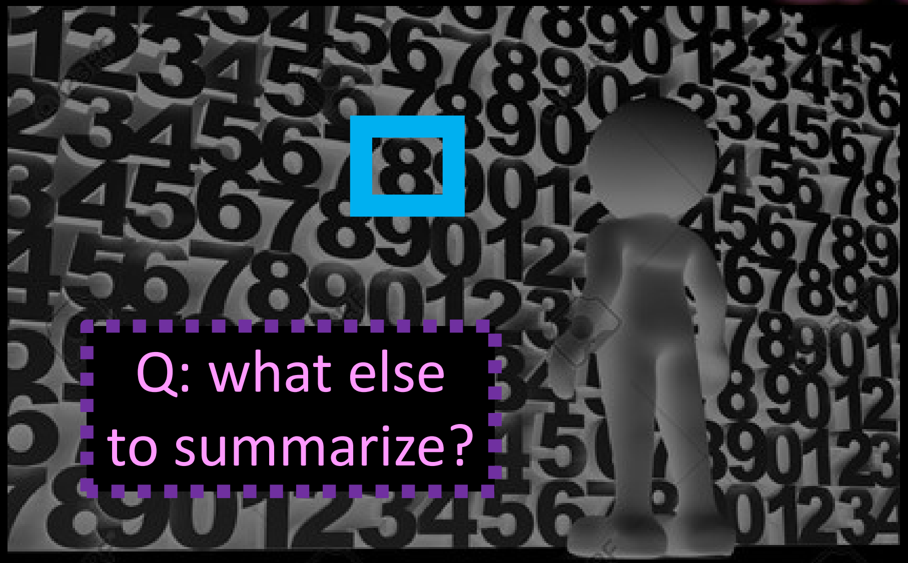


http://www.isical.ac.in/~jeexiiscore_normal/PercentilesAdvantages.htm



http://www.psychometric-success.com/images/AA1301.gif

Summarizing Data, Part 2



Q: what else to summarize?

- Ok, pile of numbers can now be summarized as *one* number
  - Mean, median, mode
- But is that enough?

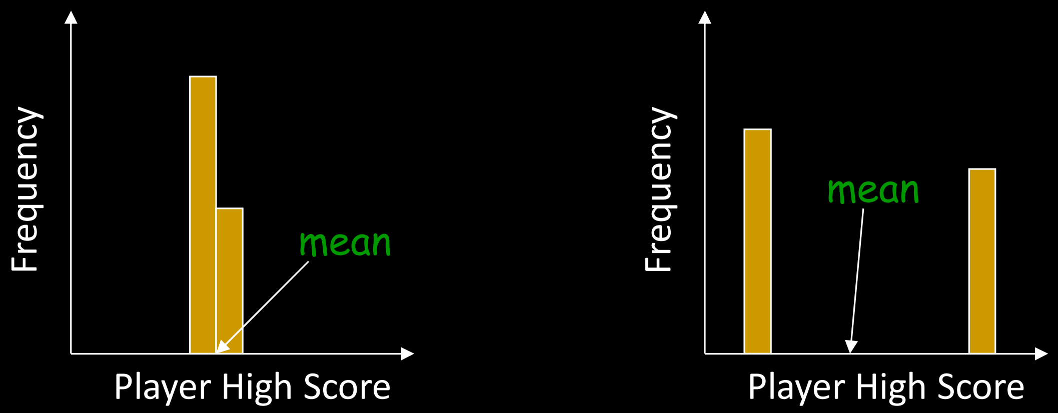
Summarizing Data, Part 2

- Measures of variation
- (*aka* measures of *dispersion*, or measures of *spread*)

Summarizing Data, Part 2 GA

“Then there is the man who drowned crossing a stream with an average depth of six inches.” – W.I.E. Gates

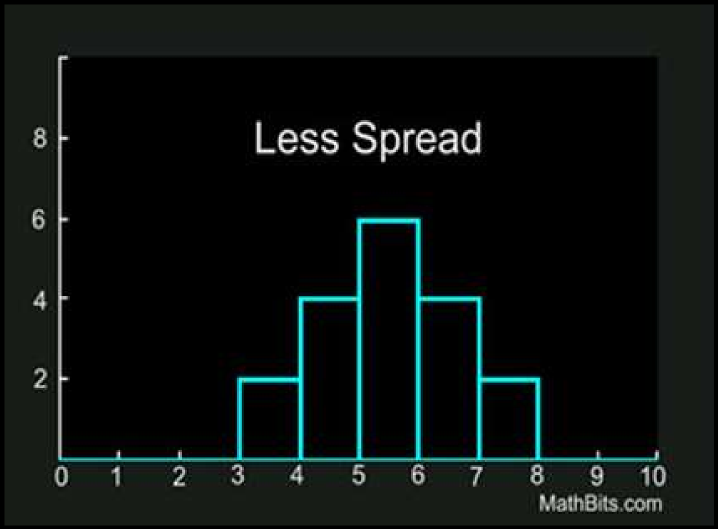
- Summarizing by single number rarely enough → need statement about dispersion (aka variation)

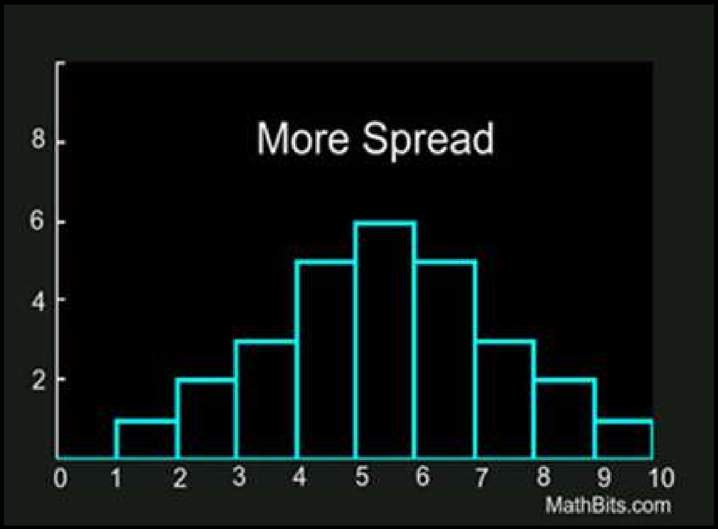


Above: does single number (mean) tell you enough about data?

Dispersion Overview (1 of 3)

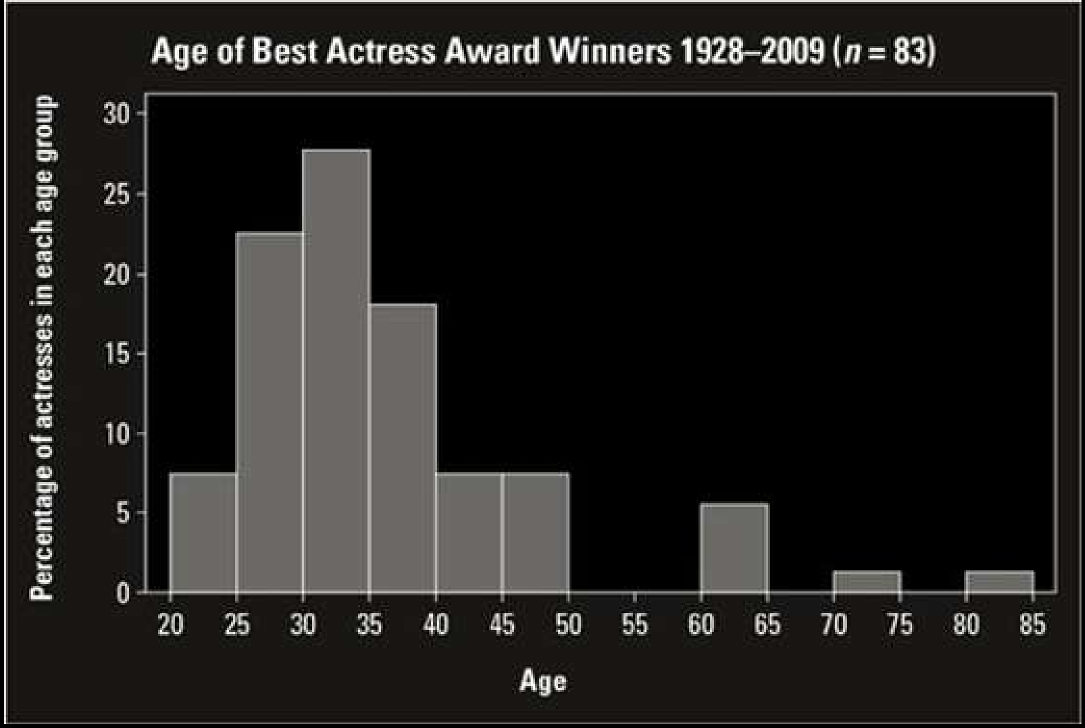
- Is data clumped or spread out?





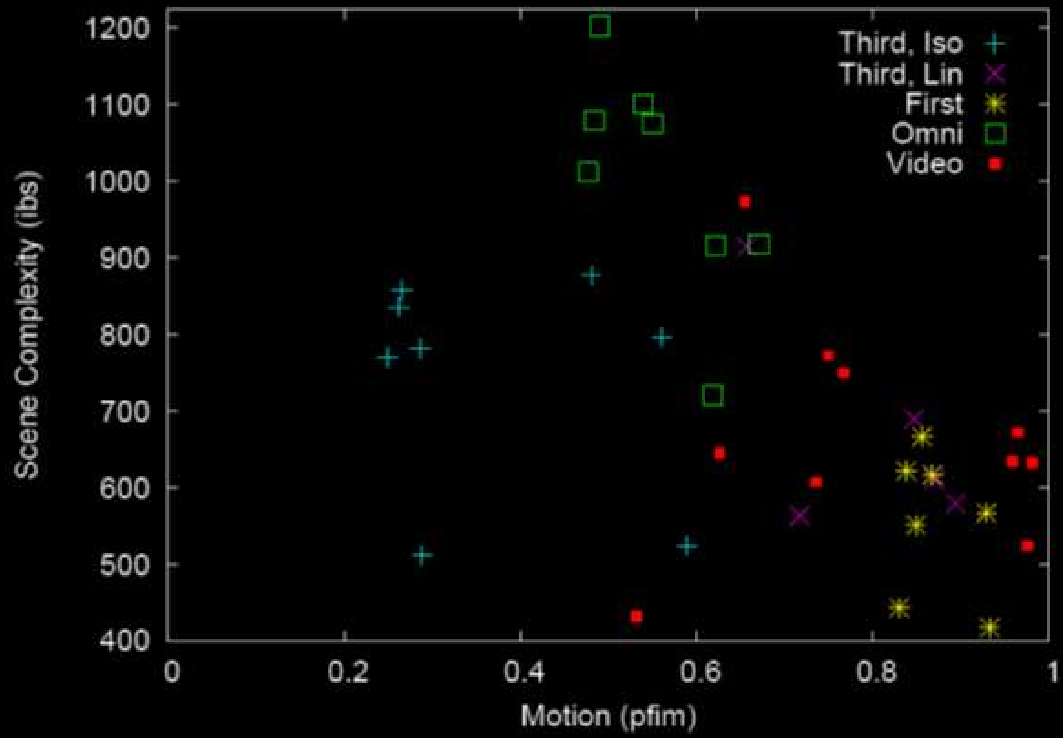
Dispersion Overview (2 of 3) GA

- Is data clumped or spread out?



Dispersion Overview (3 of 3)

- Is data clumped or spread out?



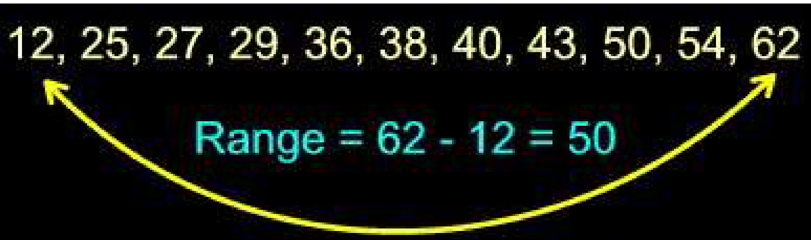
"Motion and Scene Complexity for Streaming Video Games"

Measures of Dispersion?

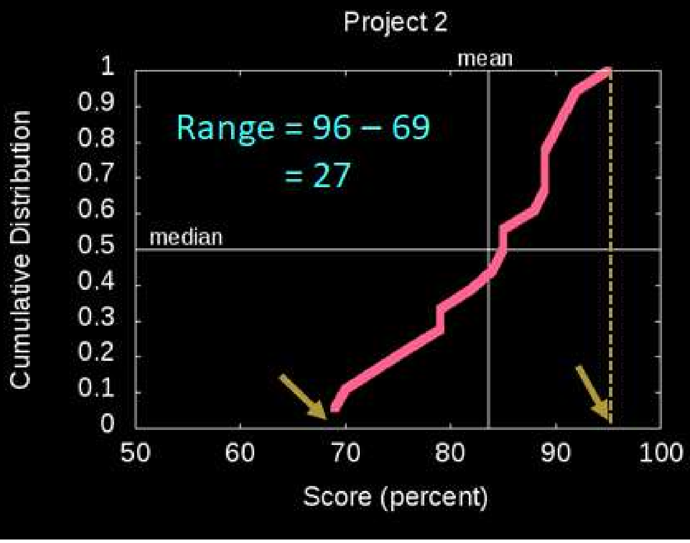
Range GA

- Difference between smallest and largest value
- Somewhat obvious, but doesn’t tell you much about “clumping”
  - Minimum may be zero
  - Maximum can be from outlier
    - Event not related to phenomena studied (e.g., 0 on project)
  - Maximum gets larger with # samples, so no “stable” point

=MAX(array)-MIN(array)



http://idolosol.com/images/range-3.jpg



Variance

- Compute mean of sample
- Compute how far each value in sample is from mean
  - Some can be less than mean, some greater
  - → So square this difference (why square?)
- Divide by number of sample values – 1
  - The “-1” corrects “bias” when trying to estimate *population variance* using *sample variance*

$$\text{Sample Variance} = s^2 = \frac{\overset{\text{“sum up all”}}{\Sigma}(X - \overset{\text{“mean”}}{\bar{X}})^2}{n-1}$$

Variance Example

- Sample kills in *LoL* match
  - 12, 20, 16, 18, 19
  - What is sample variance?

- First, mean = 85 / 5 = 17

Kills	X – mean	(X – mean)^2
12	-5	25
20	3	9
16	-1	1
18	1	1
19	2	4

s^2 = (25 + 9 + 1 + 1 + 4) / (5 – 1)

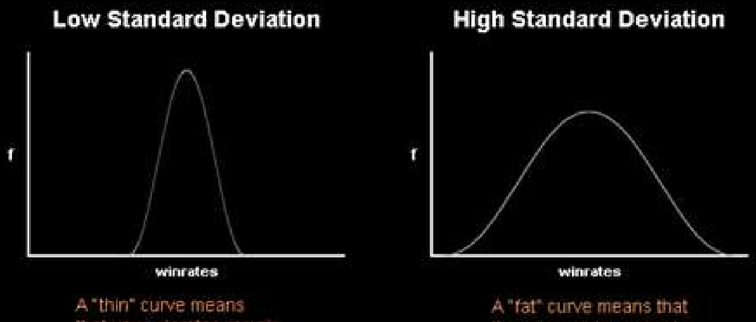
= 40 / 4 = 10 kills squared

"Larger" means "more spread" … but units odd

=VAR(array)

Standard Deviation

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$



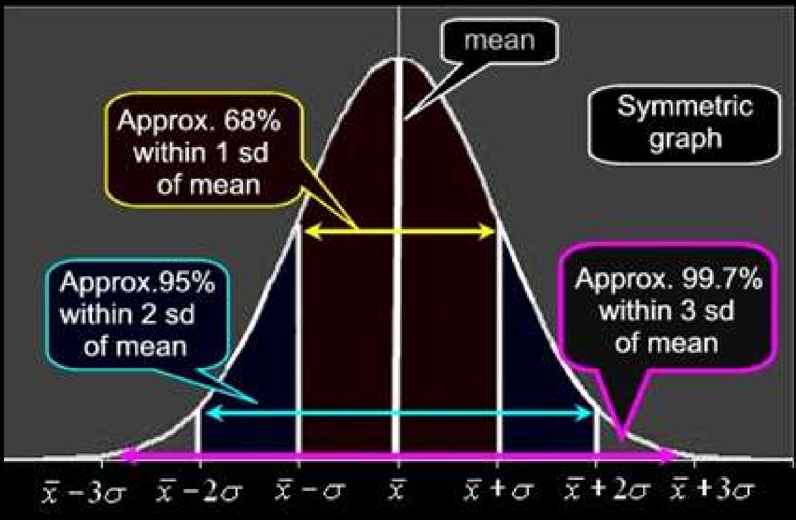
A "thin" curve means that your winrates remain close to the mean average.

A "fat" curve means that there is a wider spread of your winrates.

- Square-root of variance
- Usually, use standard deviation instead of variance
  - Why? → Same *units* as data (e.g., "kills" in previous example)
- Can compare standard deviation to mean (*coefficient of variation*, next)
- But first:
  - Mendenhall's Empirical Rule
  - Z-score

Mendenhall’s Empirical Rule

1. About 68% data within one standard deviation of mean
   - interval between mean s and mean+s contains about 68% of data
2. About 95% within 2 standard deviations of mean
3. Almost all data within 3 standard deviations of mean



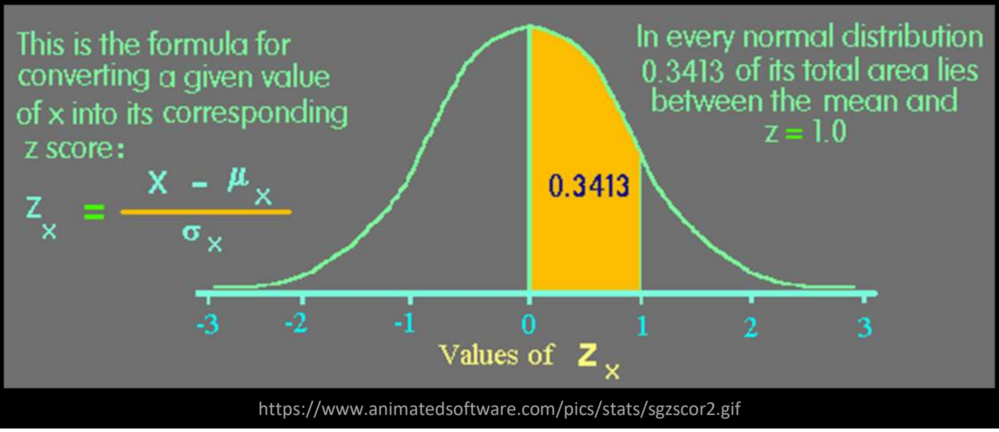
https://mathbitsnotebook.com/Algebra1/StatisticsData/normalgrapha.jpg

Rule assumes normal (“Bell curve”) distribution

Z-Score

- Measure of how "far" from center (mean) single data point is
  - *Not* measure of dispersion for whole data set

$$z = \frac{X - \bar{X}}{s}$$



This is the formula for converting a given value of x into its corresponding z score:

$$Z_x = \frac{X - \mu_x}{\sigma_x}$$

In every normal distribution 0.3413 of its total area lies between the mean and z = 1.0

https://www.animatedsoftware.com/pics/stats/sgzscor2.gif

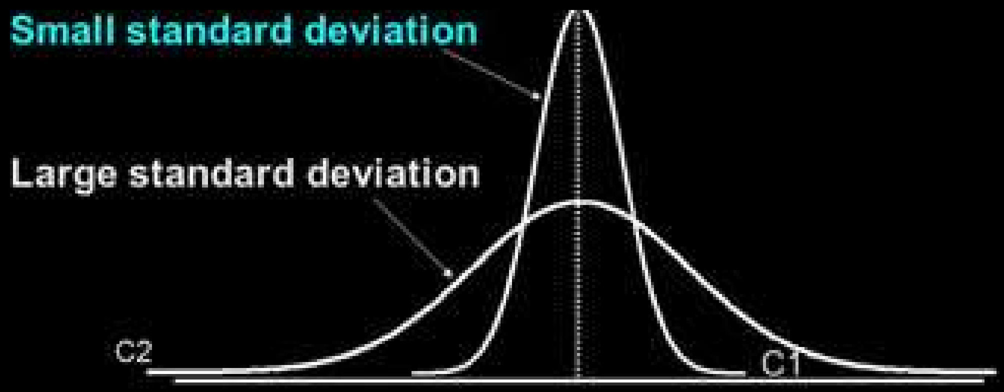
Example

Mean	469
Std dev	119
X	650

Z-score for X?

(650 – 469)/119 → 1.52

Coefficient of Variation (CV)



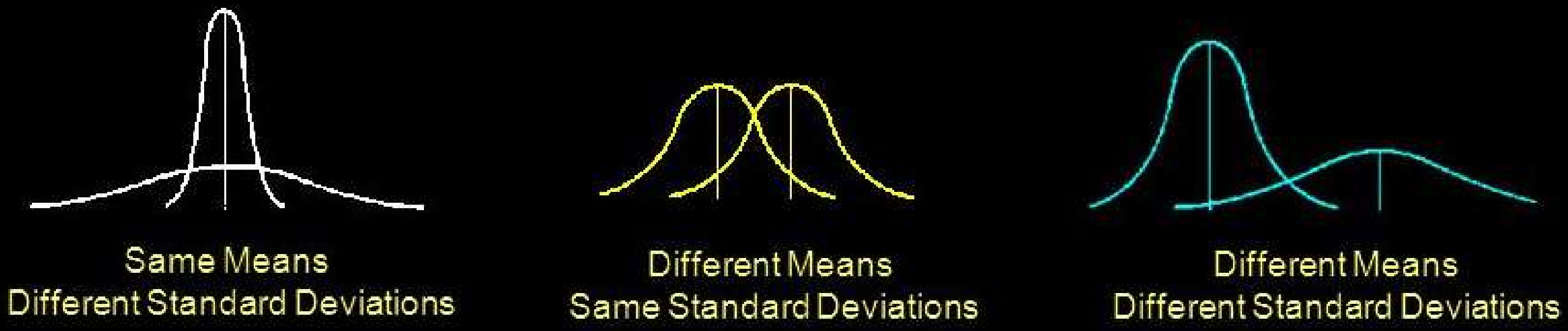
- Size of standard deviation relative to mean
  - e.g., large sd & large mean, not so spread
  - but large sd & small mean, more spread
- Standard deviation divided by mean
  - Can do this since same units!

$$CV = \frac{s}{\bar{x}} \times 100$$

percent

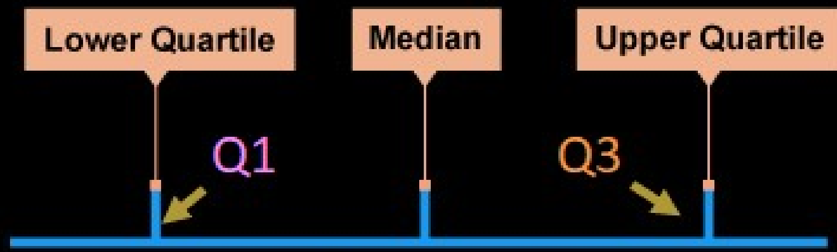
Coefficient of Variation (CV)

- What is the relative CV for each curve?



Semi-Interquartile Range

- ½ distance between Q3 (75th percentile) and Q1 (25th percentile)



http://www.bbc.co.uk/staticarchive/9629000486ef4b1a40efa565c162cb779e0bd82c.png

$$\frac{Q3 - Q1}{2}$$

- Guideline: use semi-interquartile (SIQR) for index of dispersion when using median as index of central tendency

Index of Dispersion Example

(sorted) Lap Times
1.9
2.7
3.9
4.1
4.2
4.2
4.4
4.5
4.5
4.8
4.9
5.1
5.1
5.3
5.6
5.9

- First, sort. Then, compute:
  - Mean = 4.4
  - Min = 1.9, Max = 5.9
  - Median = [16 / 2] = 8th = 4.5
  - Q1 = 16 / 4 = 8th = 4.1
  - Q3 = 3 * 16 / 4 = 12th = 5.1

SIQR = (Q3 - Q1) / 2 = 0.5

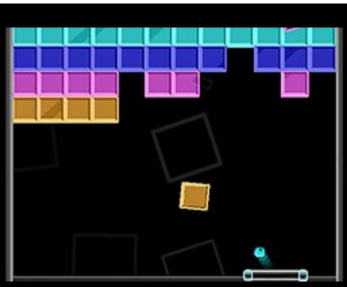
Variance = 0.96

Stddev = 0.98

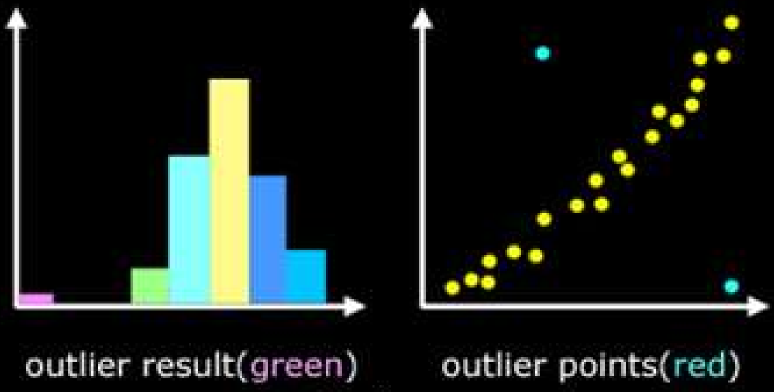
CV = stddev/mean = 0.22

Range = max – min = 4

Breakout 3



- Rank *measures of dispersion* by sensitivity to outliers
  - Standard Deviation
  - Coefficient of Variation
  - Semi-interquartile Range
  - Variance
  - Range



http://www.a-levelmathstutor.com/images/statistics/outliers-graph01.jpg

Ranking of Affect by Outliers?

Measure of Dispersion

Susceptibility

- Variance
- Range
- Standard Deviation
- Coefficient of Variation
- Semi-interquartile Range

??

Ranking of Affect by Outliers?

Measure of Dispersion

- Variance
- Range
- Standard Deviation
- Coefficient of Variation
- Semi-interquartile Range

Susceptibility

- Range
  susceptible
- Variance
  - Standard Deviation
  - Coefficient of Variation
- SIQR
  resistant

Measures of Dispersion – Categorical Data

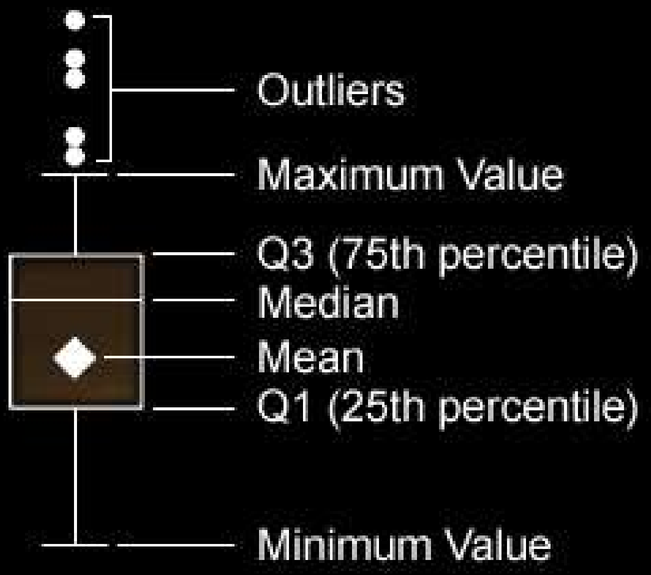
- Only for quantitative data!
  - categorical can’t quantify spread since no ‘distance’
- Instead, give categories for given percentile of sample
  - e.g., “90% of samples are in 3 categories”

Depicting Dispersion in Charts

- Histogram
- Cumulative distribution
- Box-and-Whiskers
- Error Bars

Box-and-Whiskers Chart

- Way of showing variation
- Highlight middle 50% (interquartile range, IQR)
  - "Box"
- Lines go to smallest non-outlier
  - "Whiskers"
- Points indicate outliers
- Middle line shows median
- Sometimes with mean
- Outlier? → Data value "way out there", "far" from the rest
  - Formally, 1.5+ IQRs away from quartile

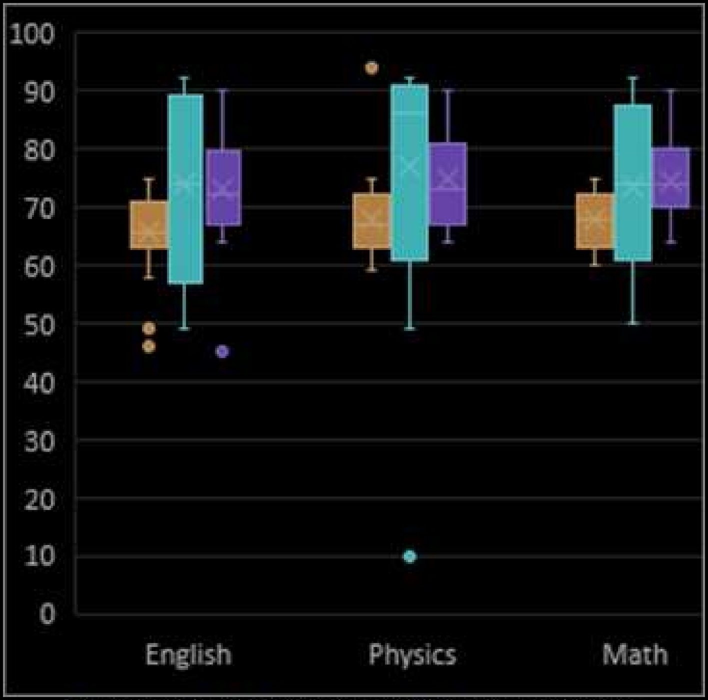


http://support.sas.com/documentation/cdl/en/vaug/65747/HTML/default/images/boxplot.png

Also called "boxplot"

Box-and-Whiskers Chart GA

- Way of showing variation
- Highlight middle 50% (interquartile range, IQR)
  - "Box"
- Lines go to smallest non-outlier
  - "Whiskers"
- Points indicate outliers
- Middle line shows median
- Sometimes with mean
- Outlier? → Data value "way out there", "far" from the rest
  - Formally, 1.5+ IQRs away from quartile



https://support.office.com/en-us/article/Create-a-box-and-whisker-chart-62f4219f-db4b-4754-aca8-4743f6190f0d

Also called "boxplot"

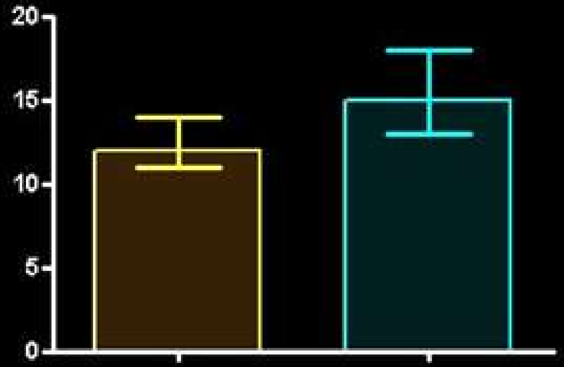
Error Bars for Columns and Points

- Line through graph point parallel to axis with “caps”
- Denotes uncertainty (variation) in value

- Often:
  - 1 standard deviation
- Can be (discuss later):
  - 1 standard error
  - 1 confidence interval

State clearly!

click “+” → “Error Bars” → “type”



https://s3.amazonaws.com/cdn.graphpad.com/faq/804/images/804b.jpg

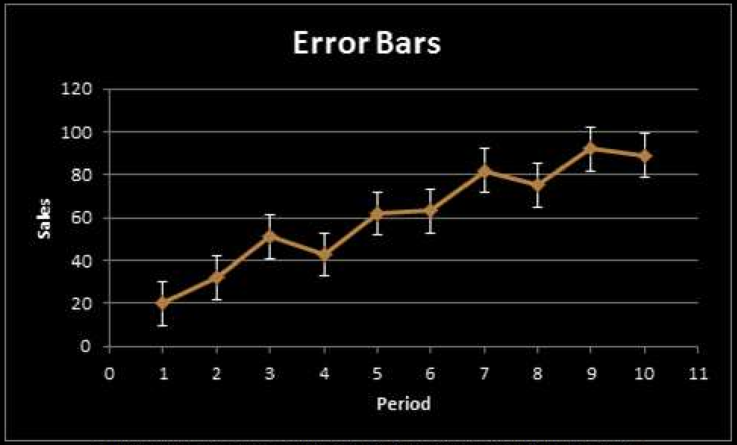
Error Bars for Columns and Points

- Line through graph point parallel to axis with “caps”
- Denotes uncertainty (variation) in value

- Often:
  - 1 standard deviation
- Can be (discuss later):
  - 1 standard error
  - 1 confidence interval

State clearly!

click “+” → “Error Bars” → “type”



http://www.excel-easy.com/examples/images/error-bars/error-bars.png