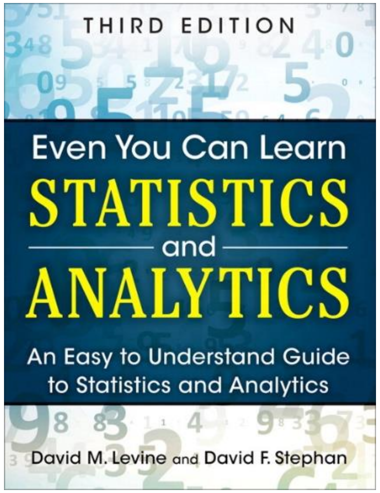
IMGD 2905

Descriptive Statistics

Chapter 3



THIRD EDITION

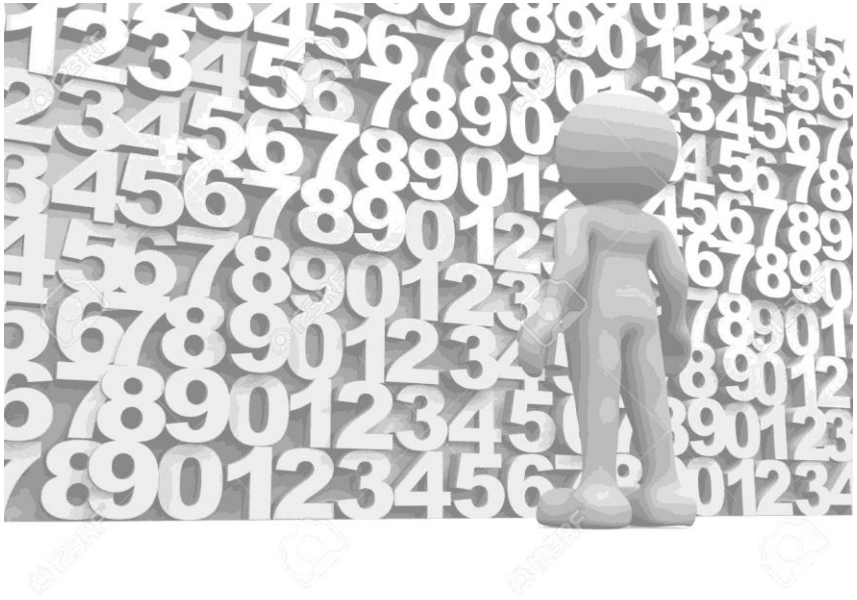
Even You Can Learn STATISTICS and ANALYTICS

An Easy to Understand Guide to Statistics and Analytics

David M. Levine and David F. Stephan

Summarizing Data

- With lots of playtesting, there is a lot of data
  - This is a good thing!
- But raw data is often just a pile of numbers
  - Rarely of interest
  - Or even sensible
- Q: How to summarize all this information?



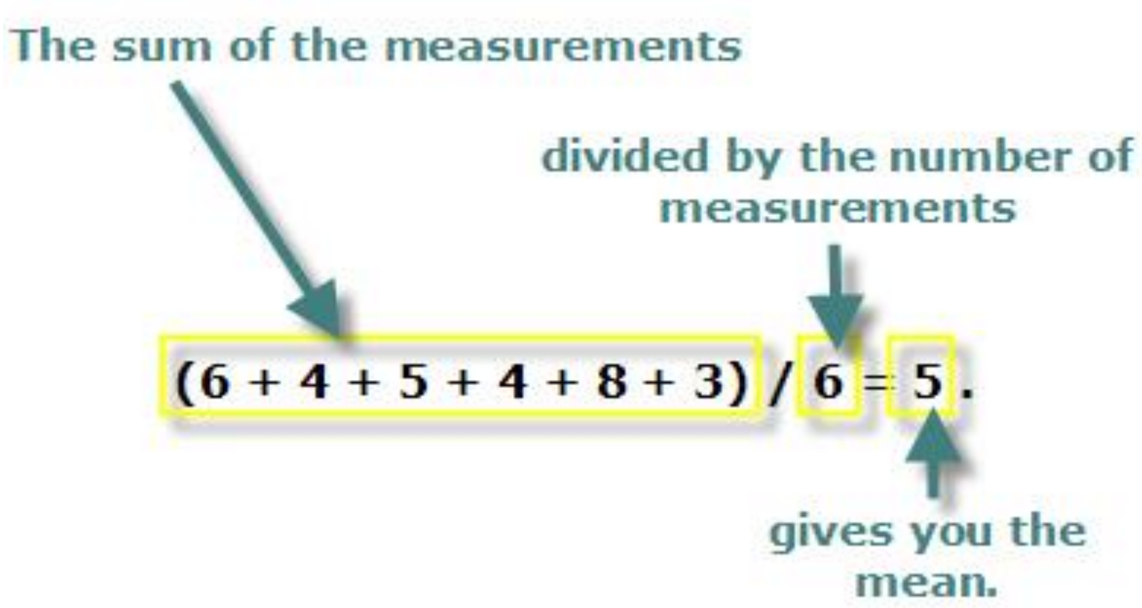
Summarizing Data

- With lots of playtesting, there is a lot of data
  - This is a good thing!
- But raw data is often just a pile of numbers
  - Rarely of interest
  - Or even sensible
- Q: How to summarize all this information?

Measures of central tendency

Examples? Pros and Cons?

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.

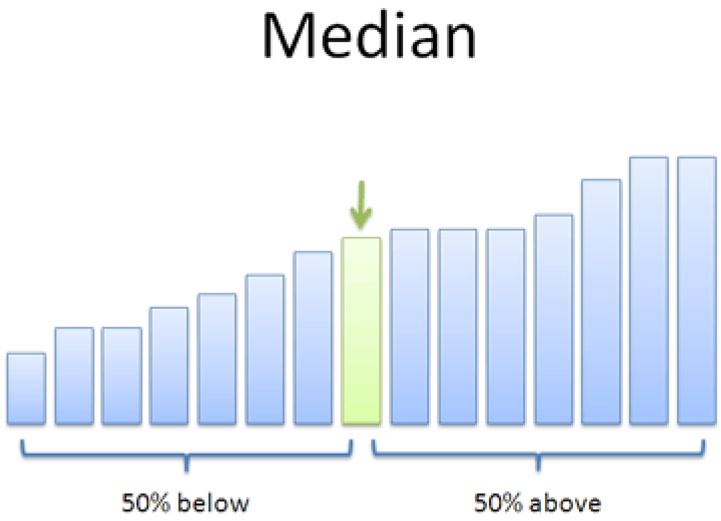
http://www.cdn.sciencebuddies.org/Files/463/9/MeanEquation.jpg

- Also called the "arithmetic mean" or "average"
- In Excel, `=AVERAGE(range)`

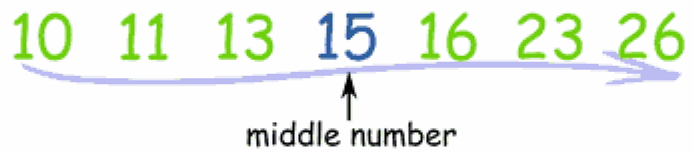
  `=AVERAGEIF()` – averages if numbers meet certain condition

Measure of Central Tendency: Median

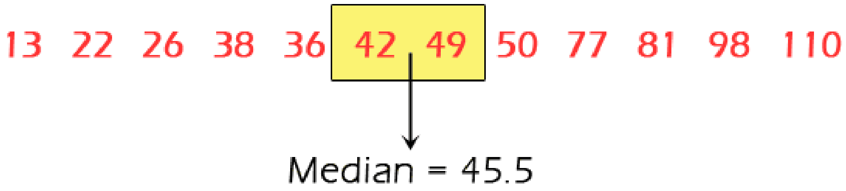
- Sort values low to high and take middle value



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

- In Excel, =MEDIAN(range)

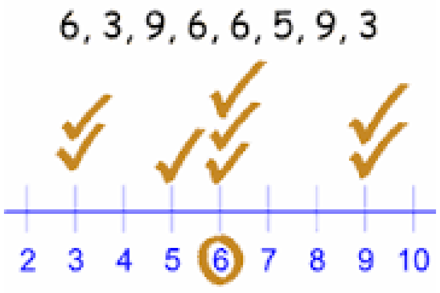
Measure of Central Tendency: Mode

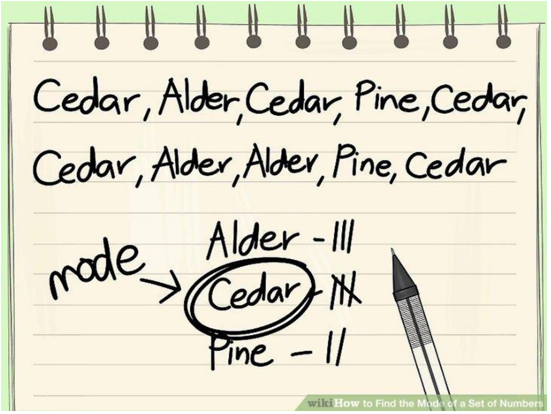
- Number which occurs most frequently
- Not so useful in many cases

→ Best use for categorical data

  - e.g., most popular Champion group in League of Legends

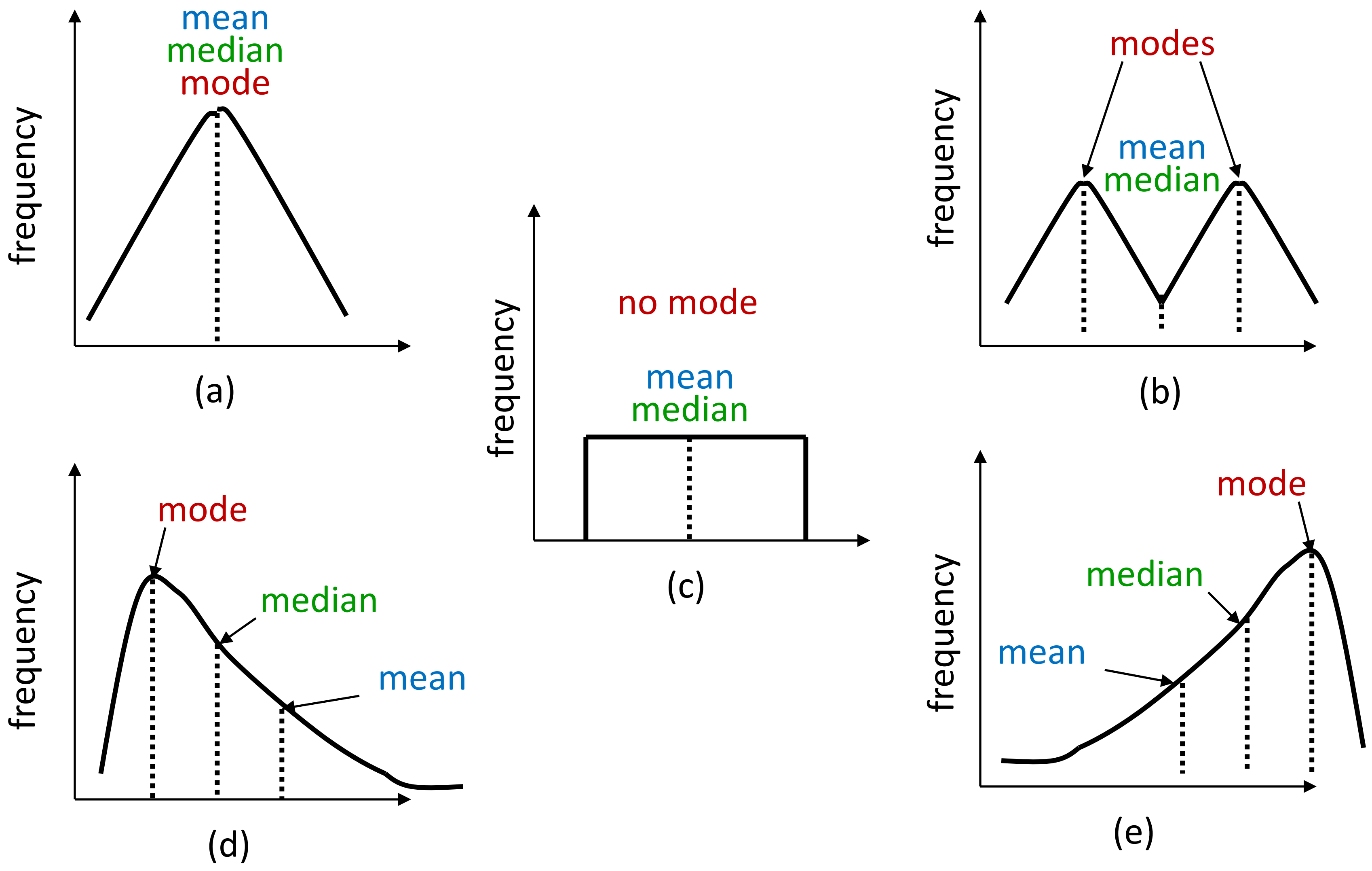
- In Excel, =MODE( )




Cedar, Alder, Cedar, Pine, Cedar,
Cedar, Alder, Alder, Pine, Cedar
Alder - III
mode
Cedar - ⅢⅠ
Pine - II
wikiHow to Find the Mode of a Set of Numbers

http://pad3.whstatic.com/images/thumb/c/cd/Find-the-Mode-of-a-Set-of-Numbers-Step-7.jpg/aid130521-v4-728px-Find-the-Mode-of-a-Set-of-Numbers-Step-7.jpg

Depiction: Mean, Median, Mode?

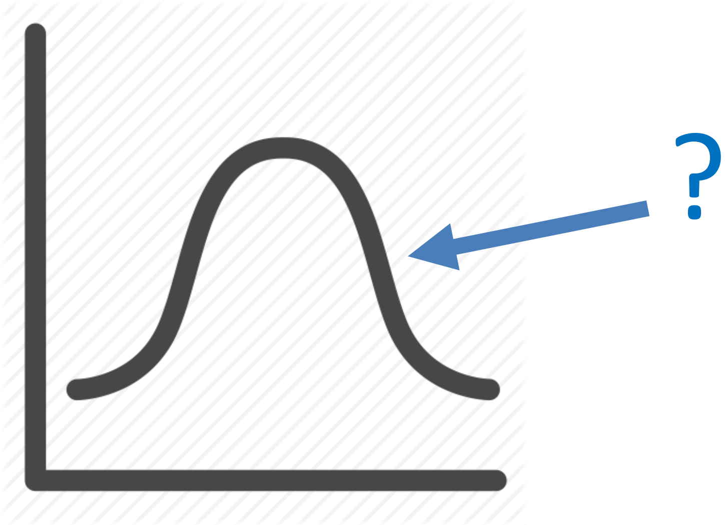


Which to Use, Mean, Median, Mode?

- Mean many statistical tests with sample
  - Estimator of population mean
  - Uses all data
- Median can be 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 - more representative
  - Does not use all data. "Resistant" to extremes (e.g., 275)
  - But what if were exam scores? Hard to "bring up" grade
- Mode can be useful primarily for categorical data
  - Most played League champion, most popular maze, …

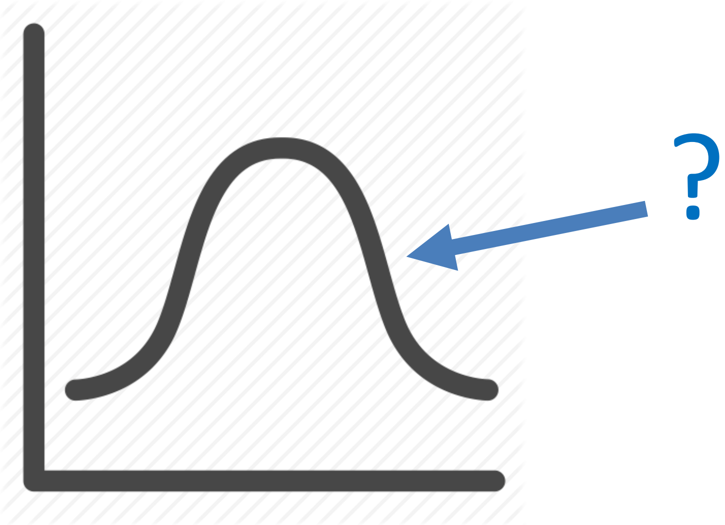
Other Measures of Position?

- May not always want center
  - e.g., what weapon that gets most kills in PUBG
- What other positions may be desired?



Other Measures of Position

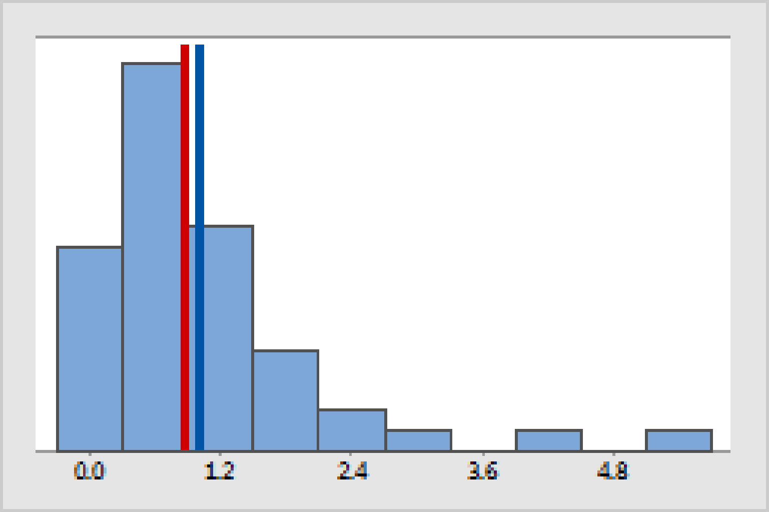
- May not always want center
  - e.g., want to know *best* LoL Champions



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

Trimmed Mean

- Take "trimming" off top and bottom (typically 5% or 10%)
  - Reduces effects of extreme values, like median
- In Excel, `=TRIMMEAN(array,percent)`

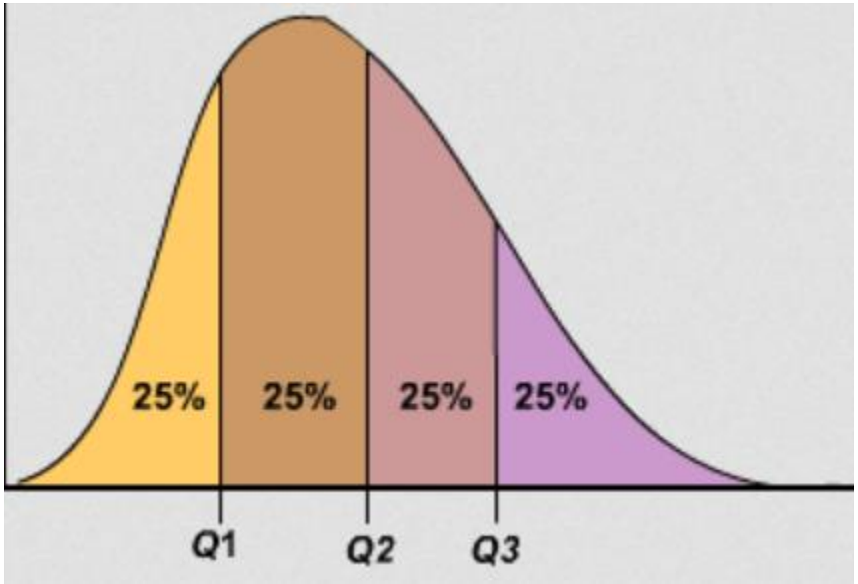


http://support.minitab.com/en-us/minitab/17/histogram_mean_vs_trimmed_mean.png

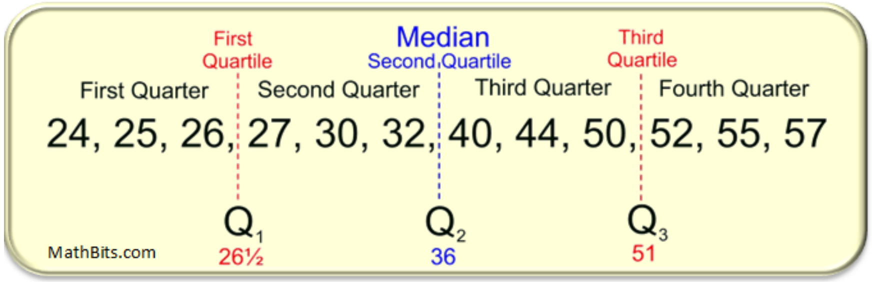
Blue – original mean
Red – trimmed mean

Quartiles

- Sort values
- First quartile (Q1) is 25% from bottom
- Third quartile (Q3) is 75% from bottom
- (What is second quartile?)
- In Excel, =`QUARTILE(array,n)`



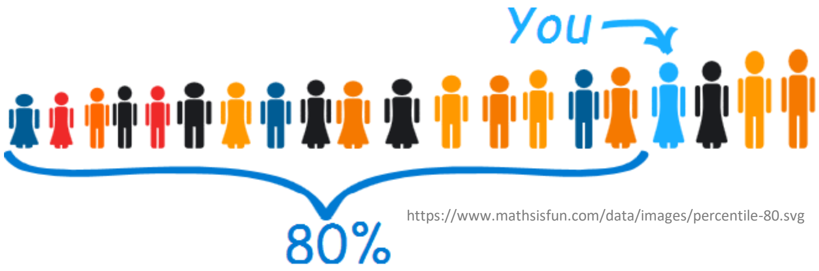
https://www.hackmath.net/images/quartiles.png



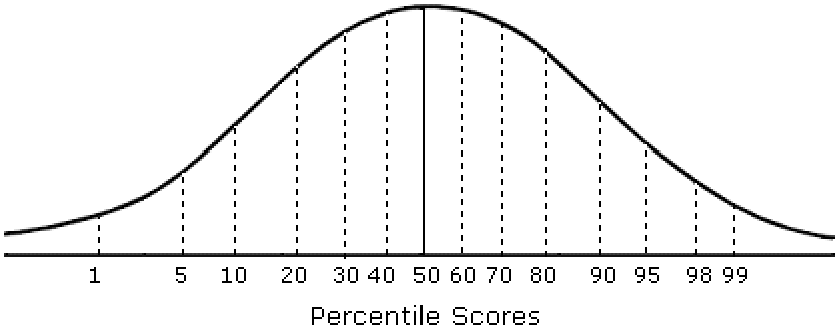
https://mathbitsnotebook.com/Algebra1/StatisticsData/quartileboxview2.png

Percentiles

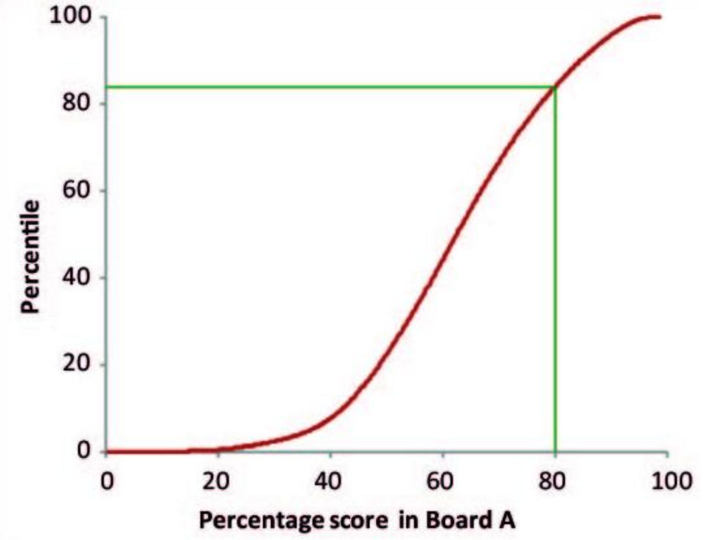
- Generalization of quartiles
- *N*th percentile is data point *n*% from bottom of data
- Interpolate as for first quartile
- In Excel, `=PERCENTILE(array,k)` (k: 0 to 1)



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



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



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

Summarizing Data, Part 2

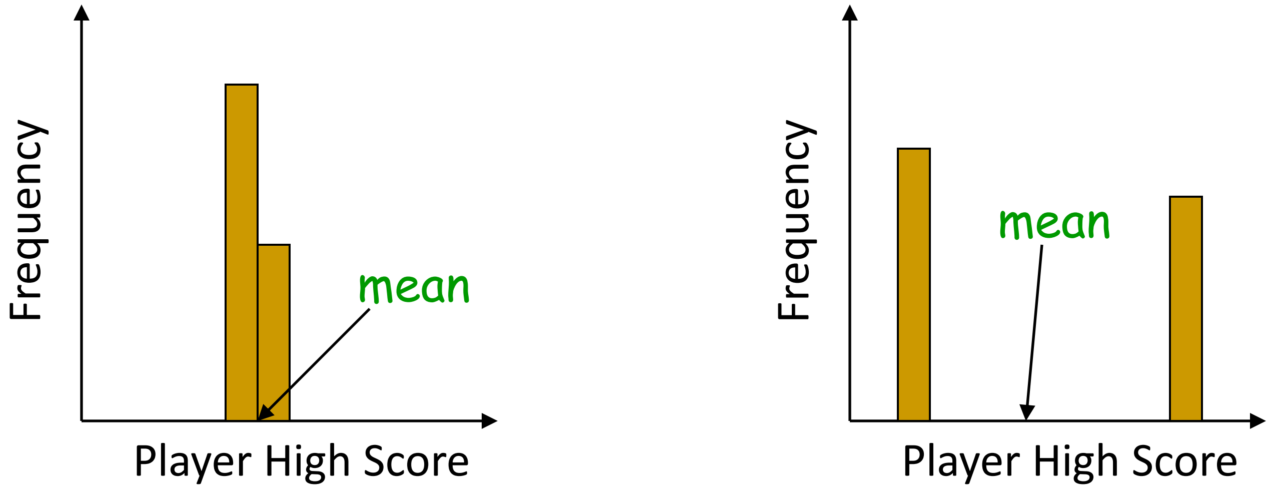
- Ok, pile of numbers can now be summarized as *one* number
  - Mean, median, mode
- But is that enough?
- Q: What other major aspect of numbers haven't we summarized?

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

Summarizing Data, Part 2

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

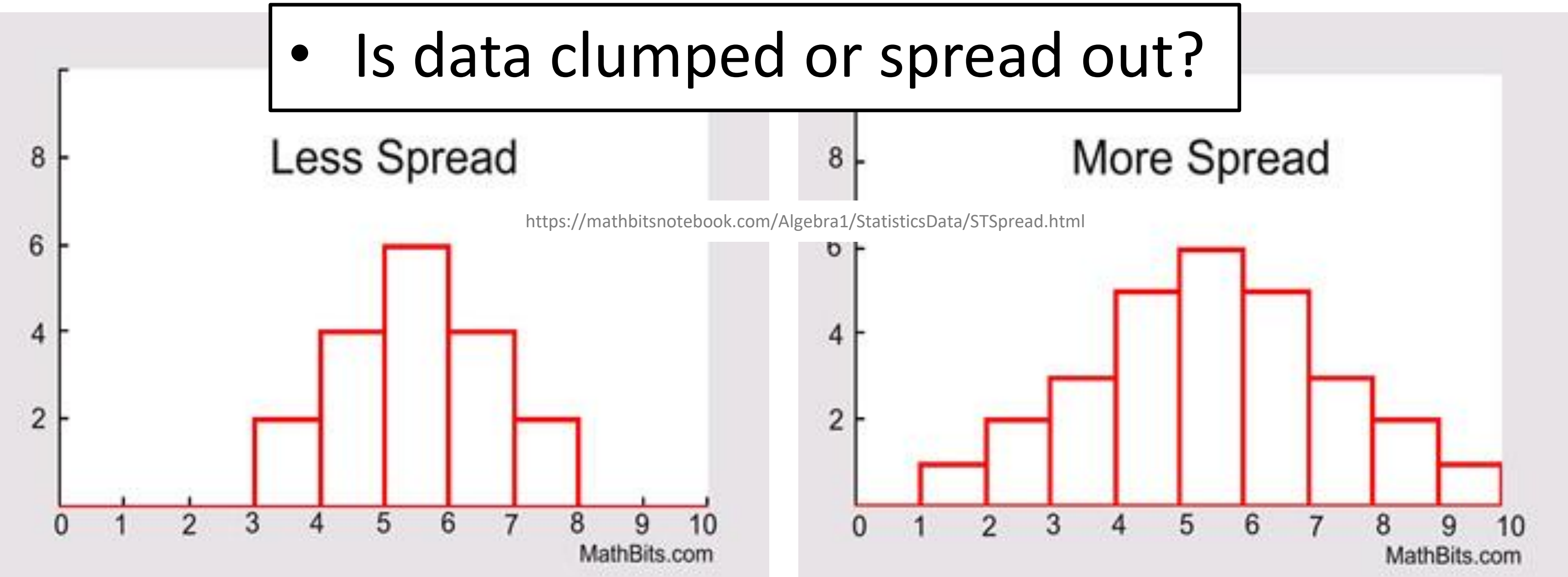
- Summarizing by single number rarely enough → need statement about dispersion (aka variation)



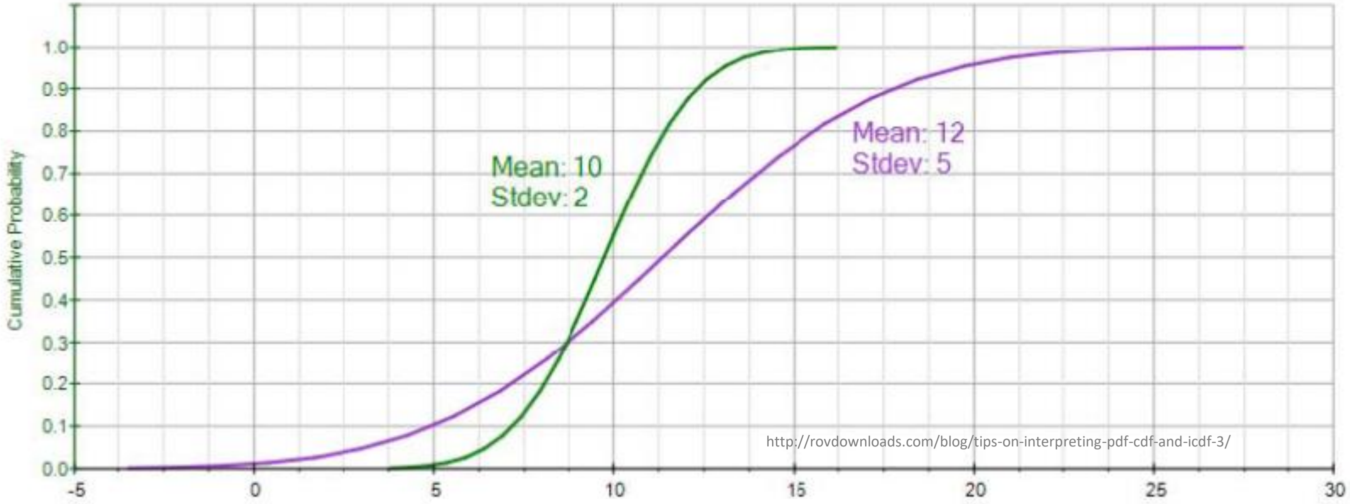
Above: does single number (mean) tell you enough about data?

Dispersion Overview (1 of 3)

- Is data clumped or spread out?



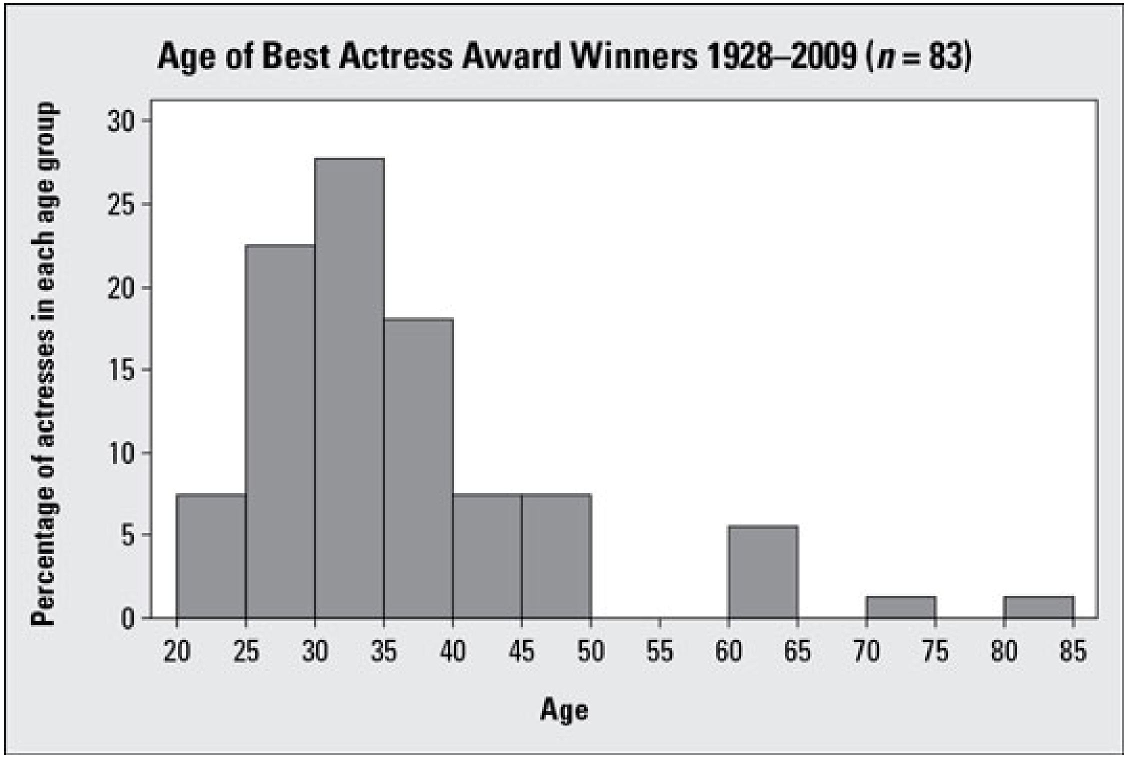
https://mathbitsnotebook.com/Algebra1/StatisticsData/STSpread.html



http://rovdownloads.com/blog/tips-on-interpreting-pdf-cdf-and-icdf-3/

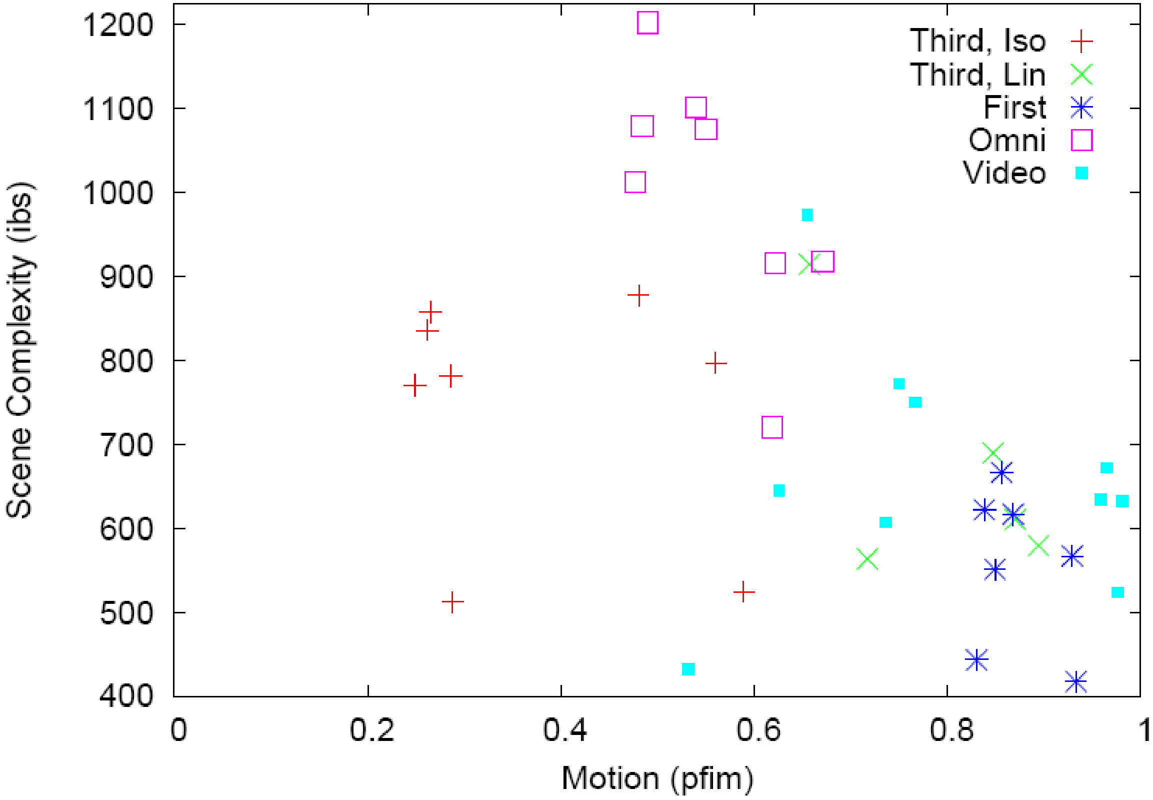
Dispersion Overview (2 of 3)

Is data clumped or spread out?



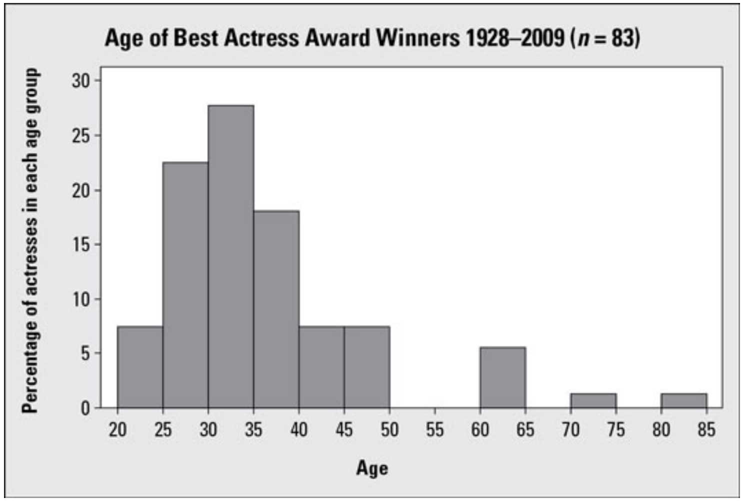
Dispersion Overview (3 of 3)

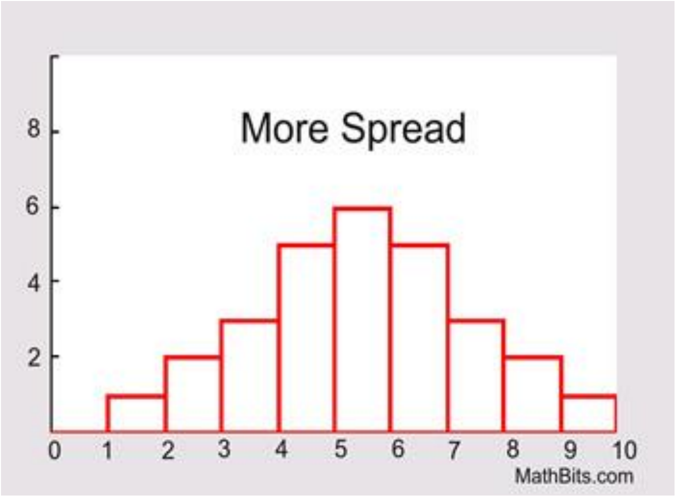
Is data clumped or spread out?

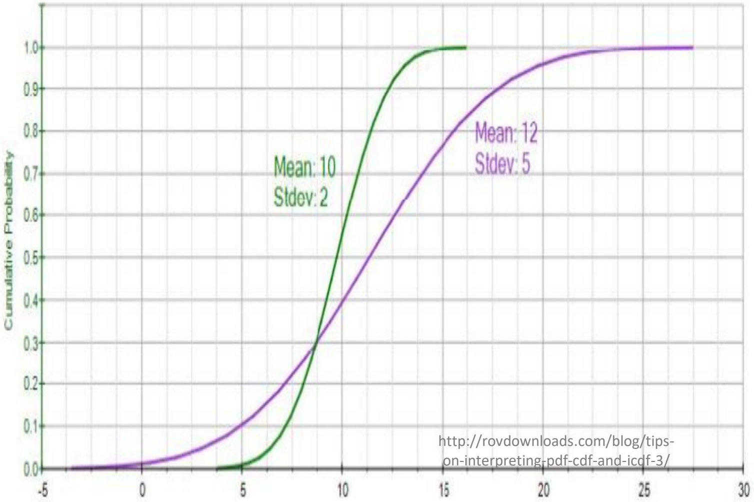


"Motion and Scene Complexity for Streaming Video Games"

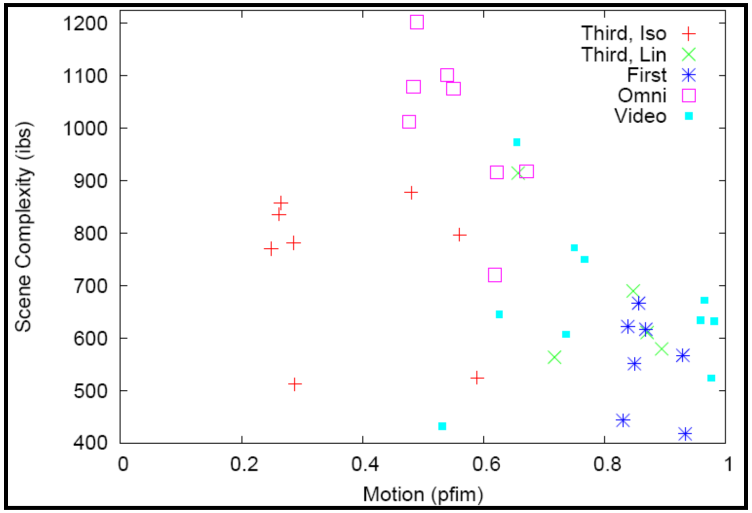
What are Some Measures of Dispersion? → Groupwork







http://rovdownloads.com/blog/tips-on-interpreting-pdf-cdf-and-icdf-3/



Groupwork

Group A: **0 6 12 18 26**

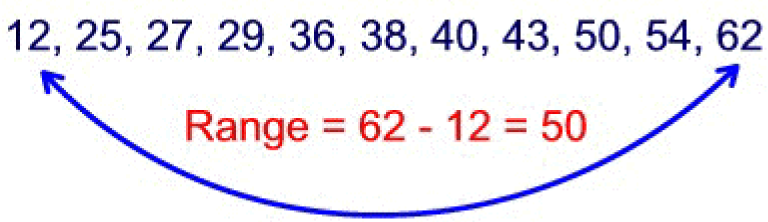
Group B: **0 18 20 22 26**

- Different ways to report *dispersion* with **one** number?
- What are pros and cons of each?
- Icebreaker, Groupwork, Questions

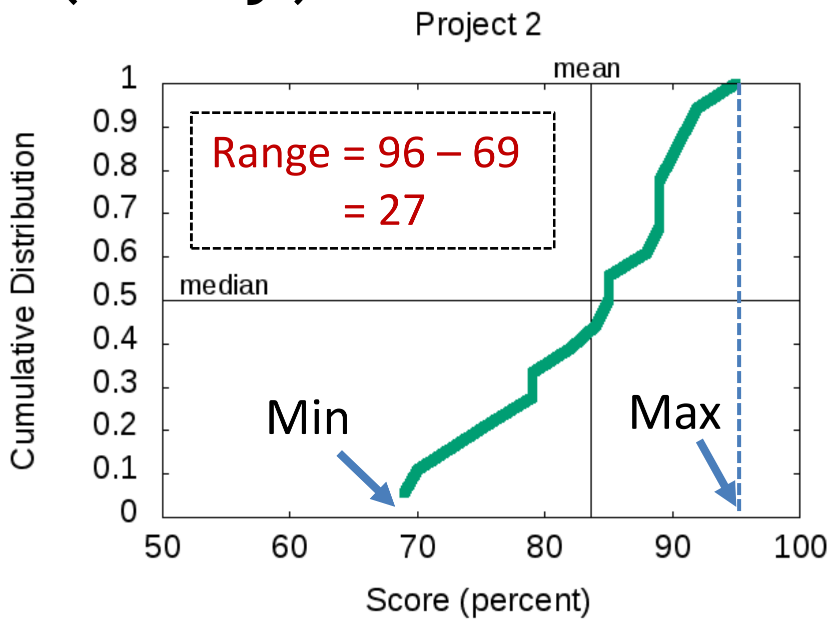
https://web.cs.wpi.edu/~imgd2905/d23/groupwork/3-dispersion/handout.html

Range

- 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
- In Excel, `=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 (what does squaring do?)
- Divide by number of sample values – 1
  - The “-1” corrects “bias” when trying to estimate *population variance* using *sample variance*

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

Variance Example

- Sample kills in *PUBG* matches
  - 12, 20, 16, 18, 19
  - What is sample variance?
- First, mean = 85 / 5 = 17

Kills	X – mean	$(X - \text{mean})^2$
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

In Excel, =VAR(array)

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

Standard Deviation

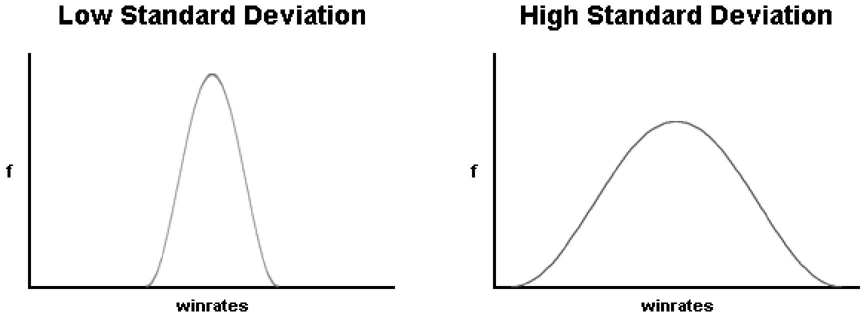
- Square-root of variance

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

- 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

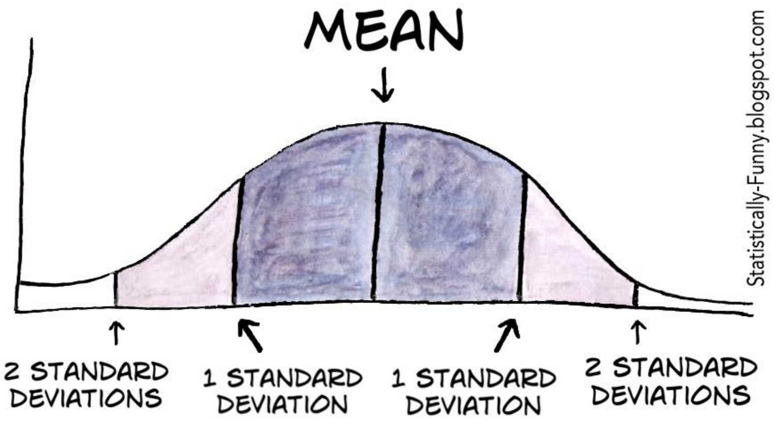
Average “distance” of points from mean

$$c = \sqrt{(a^2 + b^2)}$$



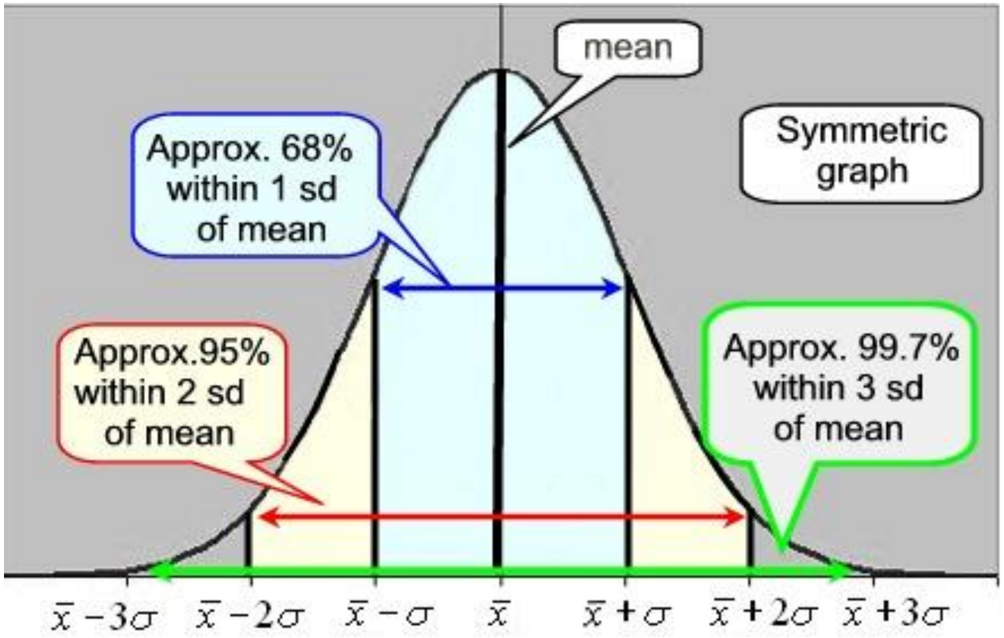
A "thin" curve means that your winrates remain close to the mean average.

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



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



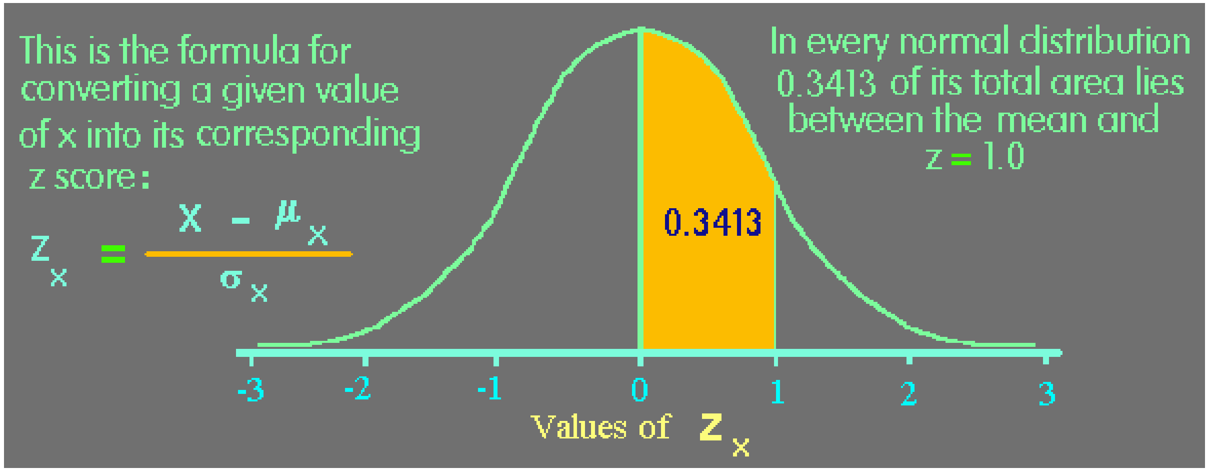
https://mathbitsnotebook.com/Algebra1/StatisticsData/normalgrapha.jpg

Rule holds for 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?

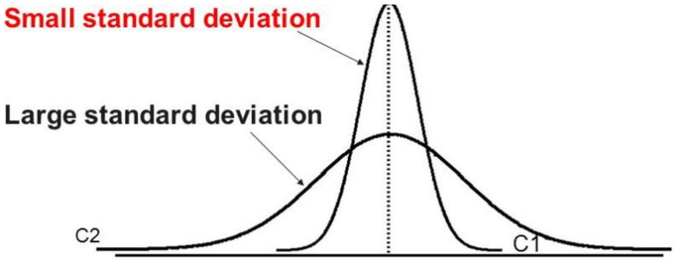
(650 – 469)/119 1.52

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 is “unit-less”, so measure of spread independent of quantity
  - E.g. seconds, clicks, spaces

Shown as percent (multiply by 100)

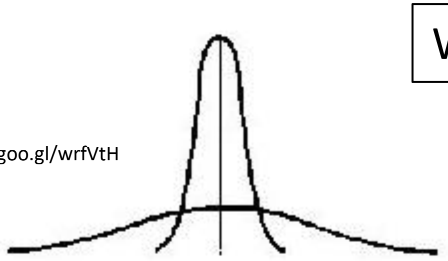
$$CV = \frac{s}{\bar{x}} \times 100$$



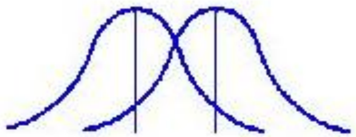
http://images.slideplayer.com/35/10391754/slides/slide_59.jpg

What is the relative CV for each curve?

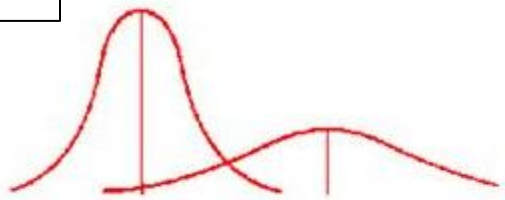
http://goo.gl/wrfVtH



Same Means
Different Standard Deviations



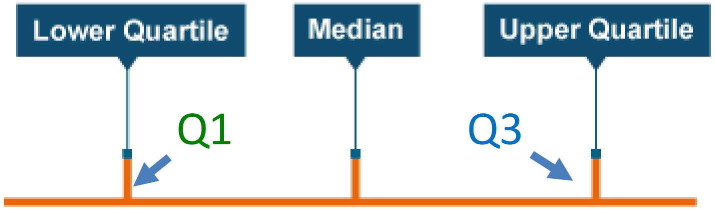
Different Means
Same Standard Deviations



Different Means
Different Standard Deviations

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 whenever using median as index of central tendency

Index of Dispersion Example

(sorted)
Lap Times

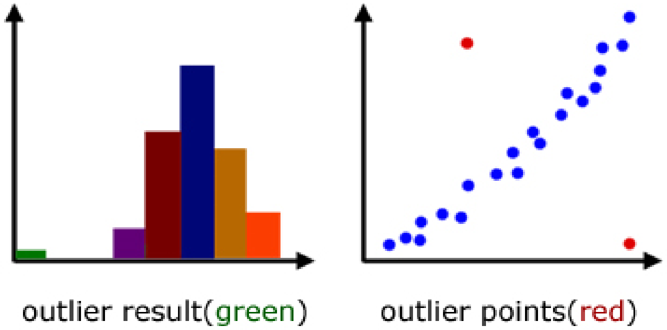
1.9
2.7
3.9
4.1
4.2
4.2
4.4
4.5
4.5
4.8
4.9
5.1
5.1
5.3
5.6
5.9

- 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

Groupwork

- Rank *measures of dispersion* by sensitivity to outliers
  - – CoV
  - – Range
  - – Std Dev
  - – Semi-interquartile Range



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

https://web.cs.wpi.edu/~imgd2905/d23/groupwork/4-outlier-effect/handout.html

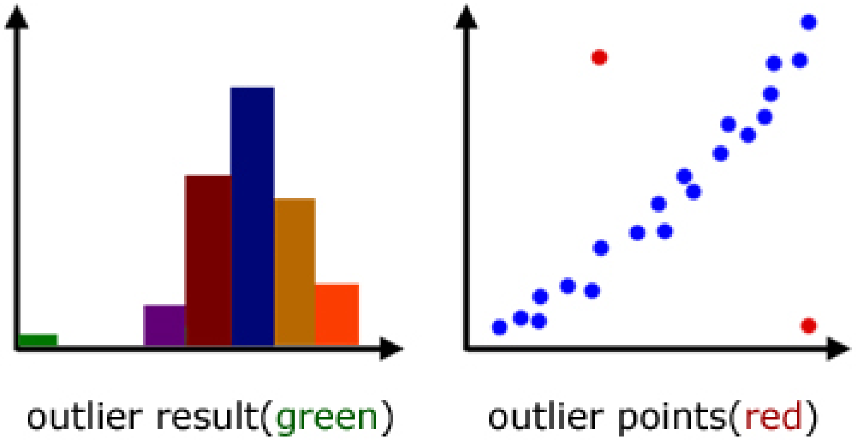
Ranking of Affect by Outliers?

Measure of Dispersion

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

Most to Least

?



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

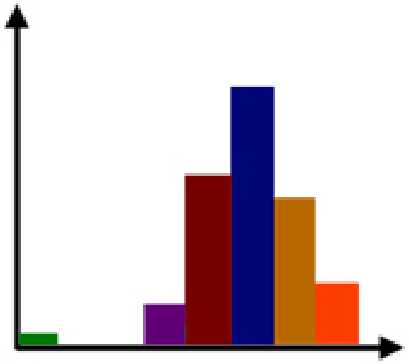
Ranking of Affect by Outliers?

Measure of Dispersion

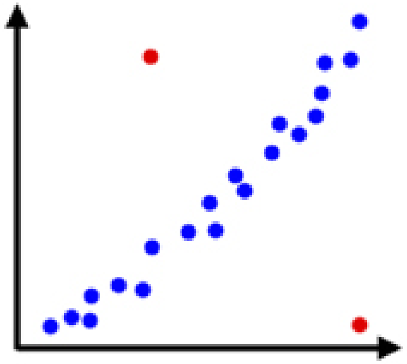
- Range
- Standard Deviation
- Coefficient of Variation
- Semi-interquartile Range

Most to Least

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



outlier result(green)



outlier points(red)

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

Only for quantitative data!

categorical can't quantify spread since no 'distance'

Instead, give categories for given percentile of samples

e.g., "90% of samples are in 3 categories" (Pareto chart)

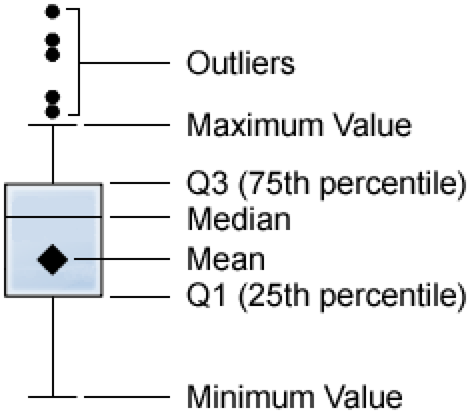
Depicting Dispersion in Charts

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

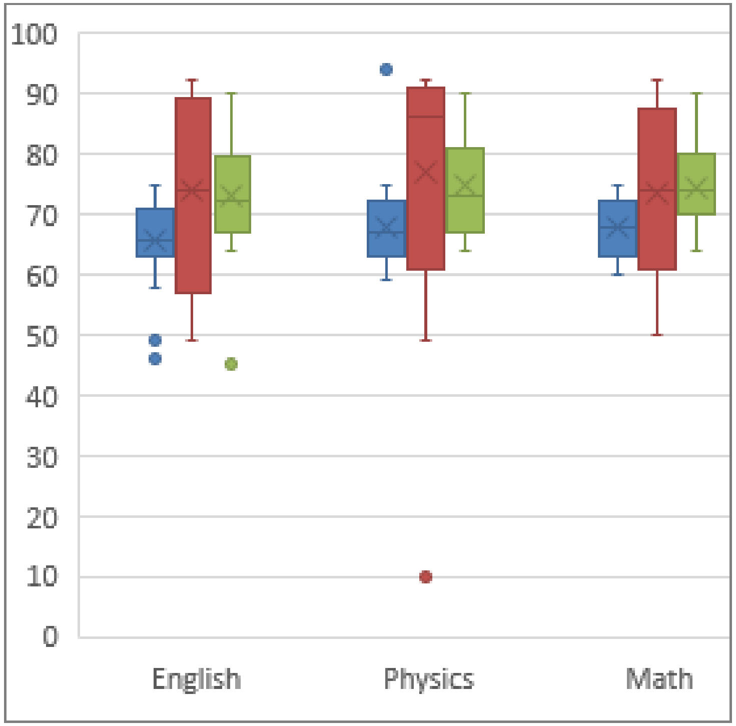
Box-and-Whiskers Chart

- 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
- Available in Excel

Also called "boxplot"



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



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

Cumulative Distribution

- Cumulative amount of data with value or less
- Easy to see min, max, median
- Compare shapes of distributions

Demo: lol-patches.xlsx

Select column R (Bug Fixes)

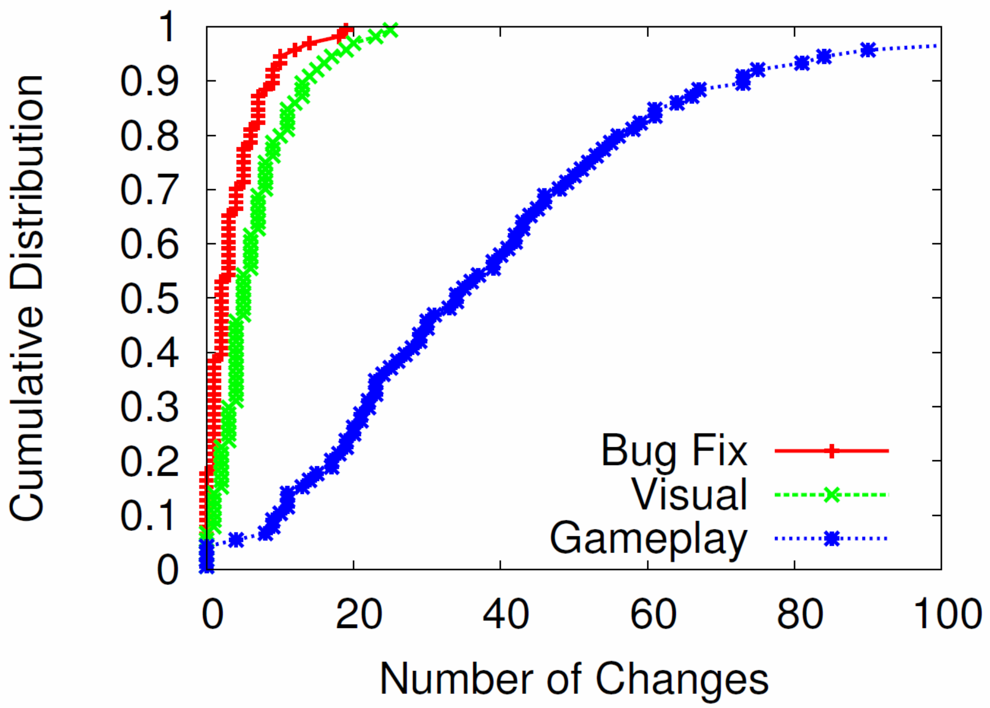
Sort low to high

New column S for percent [=ROW()/164]

Select column → paste down all

Select both column R and S

Insert → Scatter plot with lines



“Nerfs, Buffs and Bugs - Analysis of the Impact of Patching on League of Legends”

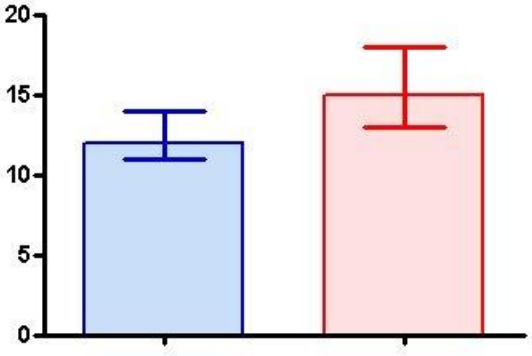
http://www.cs.wpi.edu/~claypool/papers/lol-crawler/

Error Bars for Columns and Points

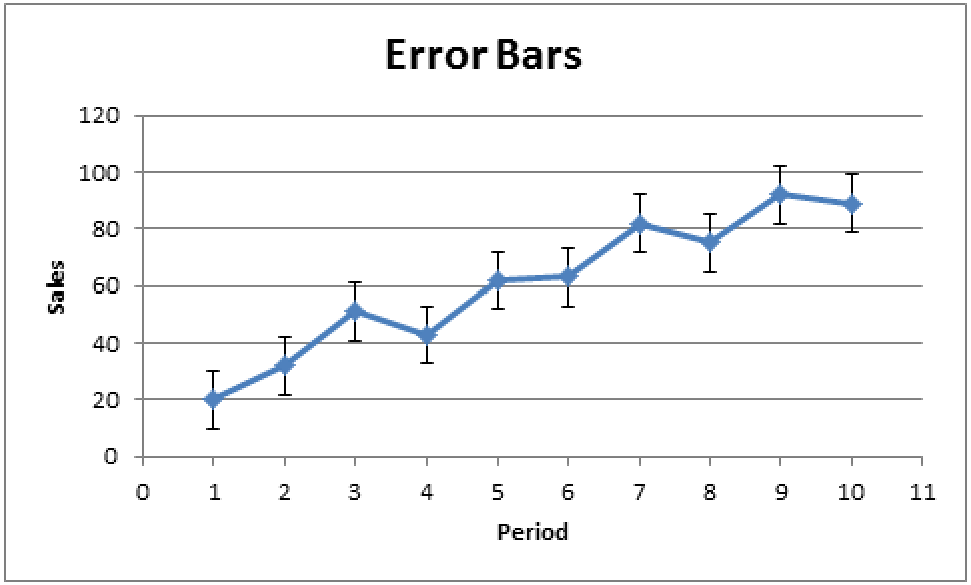
- Line through graph point parallel to axis with "caps"
- Denotes uncertainty (variation) in value
- Excel: click "+" → "Error Bars" → "type"

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

State clearly!



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



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