INM373
Research Methods and Professional Issues

Quantitative Research Part 1

Nikos Komninos, Ph.D.

nikos.komninos.1@city.ac.uk

CITY
UNIVERSITY OF LONDON
EST 1894

Overview

- Quantitative Research Methods
  - Statistics
    - Descriptive first half of today
    - Inferential – second half and next week

CITY
UNIVERSITY OF LONDON
EST 1894

Acknowledgements

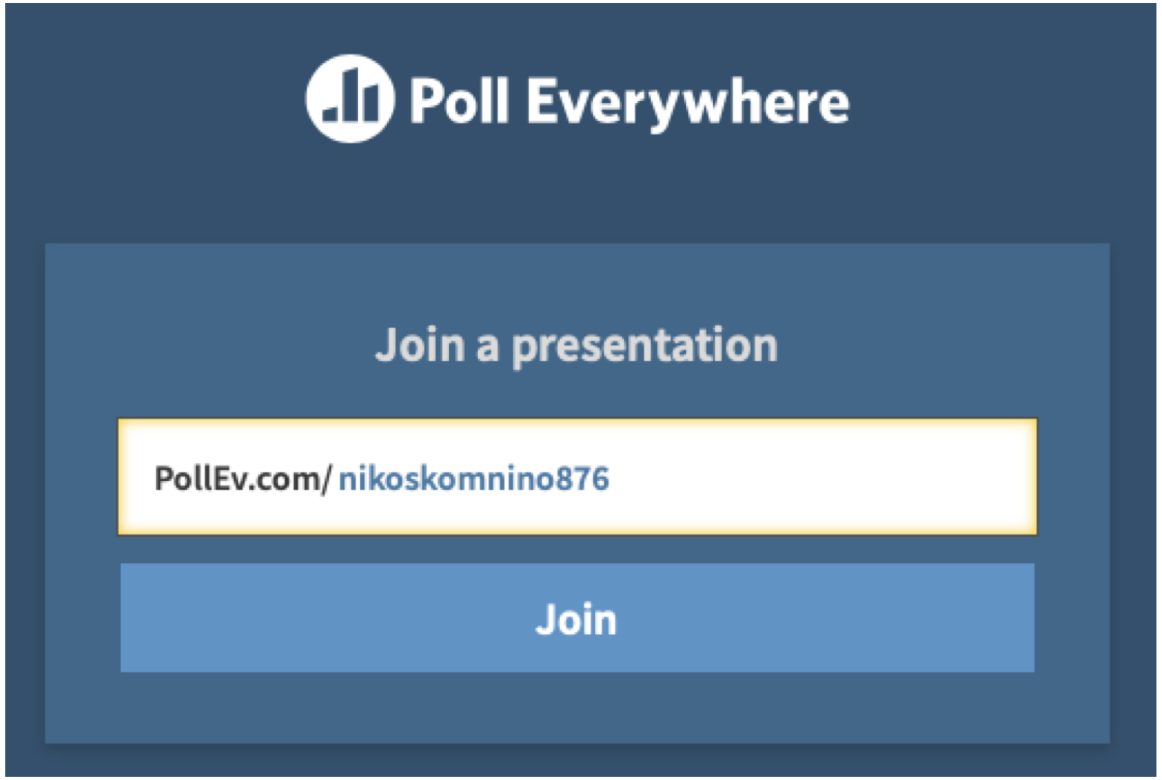
- To Dr Eugenio Alberdi who prepared most of the lecture slides in the previous years.

Join Poll Everywhere

Respond at **PollEv.com/nikoskomnino876**

Text a **CODE** to **07480 781235**

SMS charges vary per mobile carrier



What represents you best?

I have studied statistics and enjoyed it	**328439**
I have studied statistics and I did not enjoy it	**328454**
I have studied statistics and used it more than once	**328455**
I have never studied statistics	**328463**

Last Week (1)

- **Research questions and aims**
  - All research should have one or more clear aims – *what* the research seeks to accomplish (not how)
  - Research **questions** are aims, stated in a form of questions

- **Research Objectives**
  - Research objectives are ***how* you will answer** your **research question** (or achieve your research aims)

CITY
UNIVERSITY OF LONDON
EST 1894

Last Week (2)

- Use **divergent thinking** (e.g., brainstorming) to generate a broad range of research questions (4+ recommended)
  - These can be questions you might want to follow up on for your project.
  - Or, they might just be questions about topics you are interested in

CITY
UNIVERSITY OF LONDON
EST 1894

Numerical Methods: Statistics

- What is “statistics”?
  - an academic discipline
  - a set of methods to process data
  - collections of data gathered through these methods
  - specially calculated (summary) figures

“The statistics department use statistics to gather statistics that allow them to quote statistics.”

(Rowntree, 2000)

CITY
UNIVERSITY OF LONDON
EST 1894

Statistics

- "Describing things with numbers and assessing the odds" (Rugg & Petre, 2006)

- Descriptive – first half of today
  - "what you've got" (Rugg & Petre, 2006)
  - "summarizing essential features" (Rowntree,2000)
- Inferential – second half and next week
  - "what are the odds against your findings being due to random chance?" (Rugg & Petre, 2006)
  - "whether you've found something remarkable" (Rugg & Petre, 2006)

Numbers

Quantitative

- Research often involves collecting quantitative data, numerical information:
  - Quantities
  - measurements
  - Counts
  - Scores
  - Ranks
  - Timing
  - Intervals
  - Frequencies
  - Percentages

Statistics: methods to make sense of those numbers

CITY
UNIVERSITY OF LONDON
EST 1894

Numbers: example of research questions

- Do these people behave differently to those people?
  - frequency of behaviours (how often they do this or that)
- Do people perform better with this or that piece of software?
  - number of errors; how long it takes;
- Does this system work better than that system?
  - quicker, more accurate, preferable, etc.
- Do we achieve better detection with SVM or k-means?
  - more accurate, preferable, etc.
- Do people agree or disagree with Trump's politics on Twitter?
  - numbers or proportions of tweets

Measurements

CITY
UNIVERSITY OF LONDON
EST 1894

Measurement

- Levels of measurement
  - **CATEGORIES** – arithmetic operations are not applicable
    - Nominal (unordered categorical data)
    - Ordinal (ordered categorical data)
  - **NUMBERS** – arithmetic operations are applicable
    - Ratio (discrete vs. continuous)
    - Interval (arbitrary zero)

(OTHER: ranks, rates, percentages,...)

What type of measurement is this: "reliable" vs "unreliable"

Ratio	**328480**
Nominal	**328485**
Interval	**328487**
Ordinal	**328500**

What type of measurement is: Stage cancer I/II/III/IV

Ratio	**328501**
Ordinal	**328530**
Nominal	**328588**
Interval	**328817**

What sort of measurement is this: Blood Pressure: Low- Ideal- High

Ratio	**328818**
Nominal	**328819**
Ordinal	**328820**
Interval	**328821**

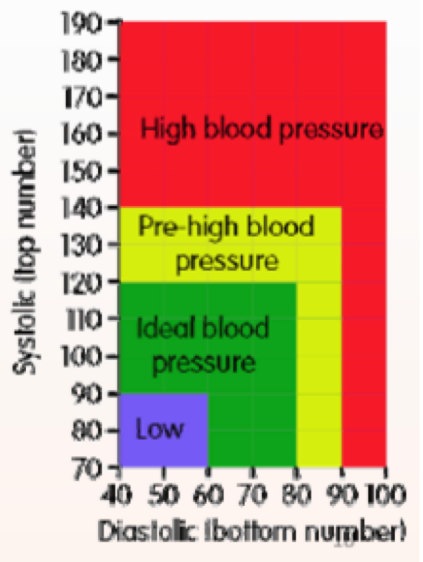
What type of measurement is: HEIGHT in metres?

Ordinal	**328822**
Ratio	**328823**
Interval	**328839**
Nominal	**328847**

CITY
UNIVERSITY OF LONDON
EST 1894

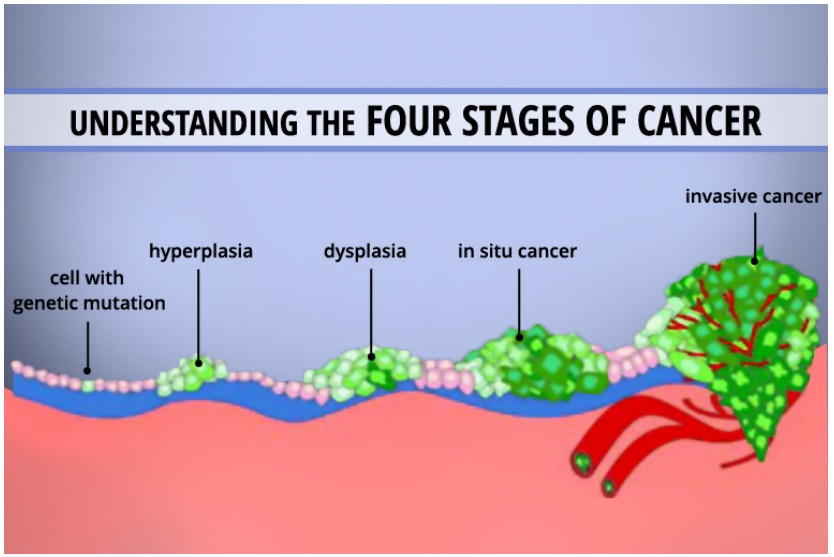
Categorical Examples

- **Nominal** (category membership without ordering)
  - Binary:
    - male/female, attack/no attack, diabetic/non-diabetic, married/single, smoker/non-smoker
  - N-ary:
    - Blood group: A/B/AB/O
    - DDoS attacks: Application / Zero Day/ / Ping Flood / SNMP Flood /...
    - Marital status: married/single/divorced/ /separated/widowed



Categorical Examples (cont…)

- **Ordinal** (ordered categories)
  - penetration testing:
    gathering target information, scanning, gaining access, maintaining access, covering tracks
  - degrees of pain: minimal / moderate /severe/ unbearable
  - stage of cancer: I/ II/ III/ IV
  - non-/ex-/light/heavy smoker



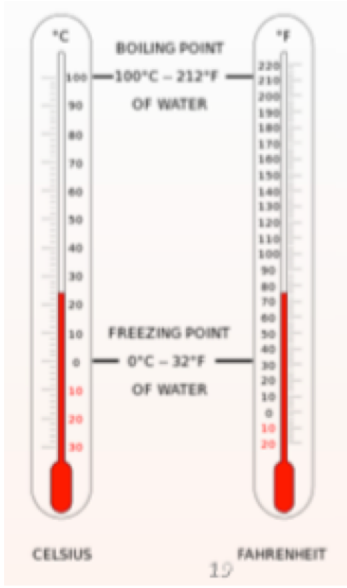
Numerical Examples

- **Ratio** (most common measurement)
  - Discrete (like “ordinal” but can only take numerical values)
    - No. of children, No of devices in a network, No. of visits to GP in a year, 2 parts flour to 1 part fat,...
      - these can take only certain numerical values
  - Continuous
    - body weight, height, blood pressure, age (can be),...
      - you can have fractions

Numerical Examples (cont…)

- **Interval** (arbitrary zero)
  - degrees in measuring temperature (Celsius, Fahrenheit), time of day (on a 12-hour clock), score on standardized scale of political orientation...
    - and other scales constructed so as to possess equal intervals
    - numerical values expressed on a scale

    with fixed units so that the interval between them is meaningful but ratios are not

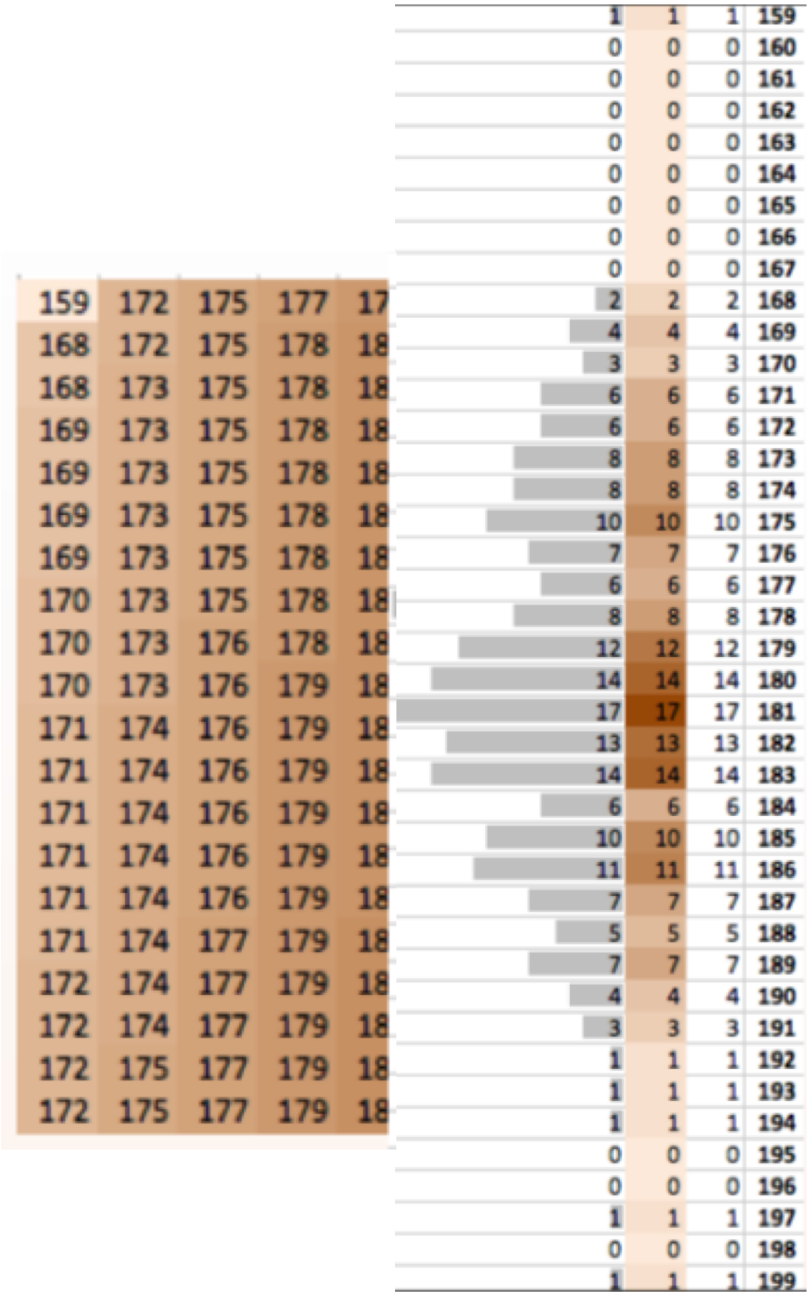


CITY
UNIVERSITY OF LONDON
EST 1894

Useful Representation?

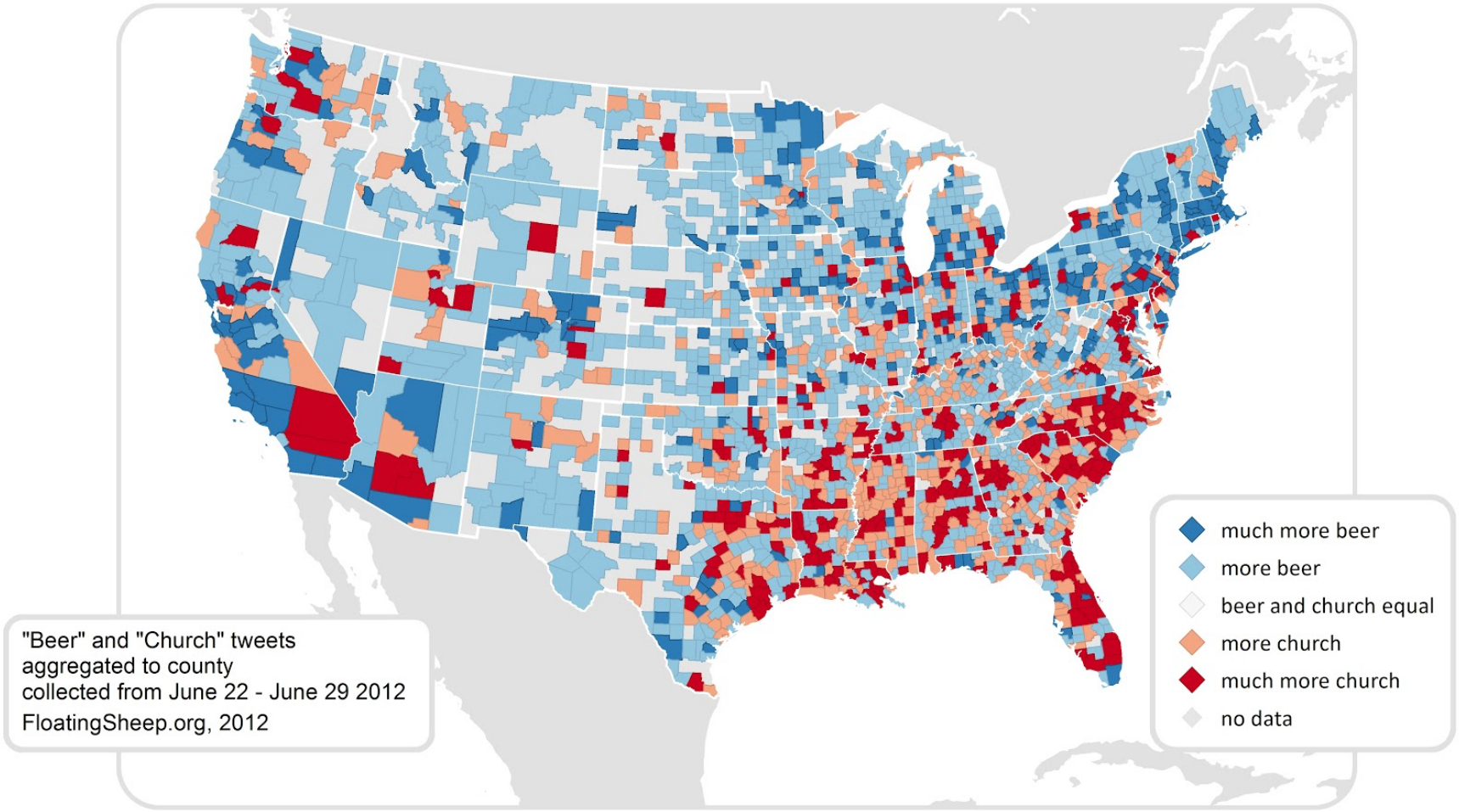
- Summarises data
- Reveals underlying patterns

- for example:
  - table (ordered / aligned)
  - ordered groups
  - graphics
    - frequency histograms
    - bar charts
    - pie charts



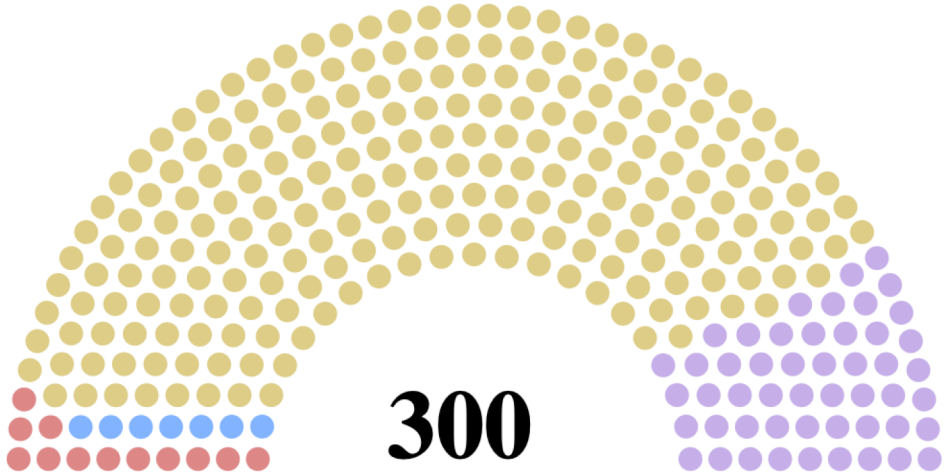
CITY
UNIVERSITY OF LONDON
EST 1894

Example: Beer or Church on Twitter?



CITY
UNIVERSITY OF LONDON
EST 1894

Example: Elections in Greece 2019



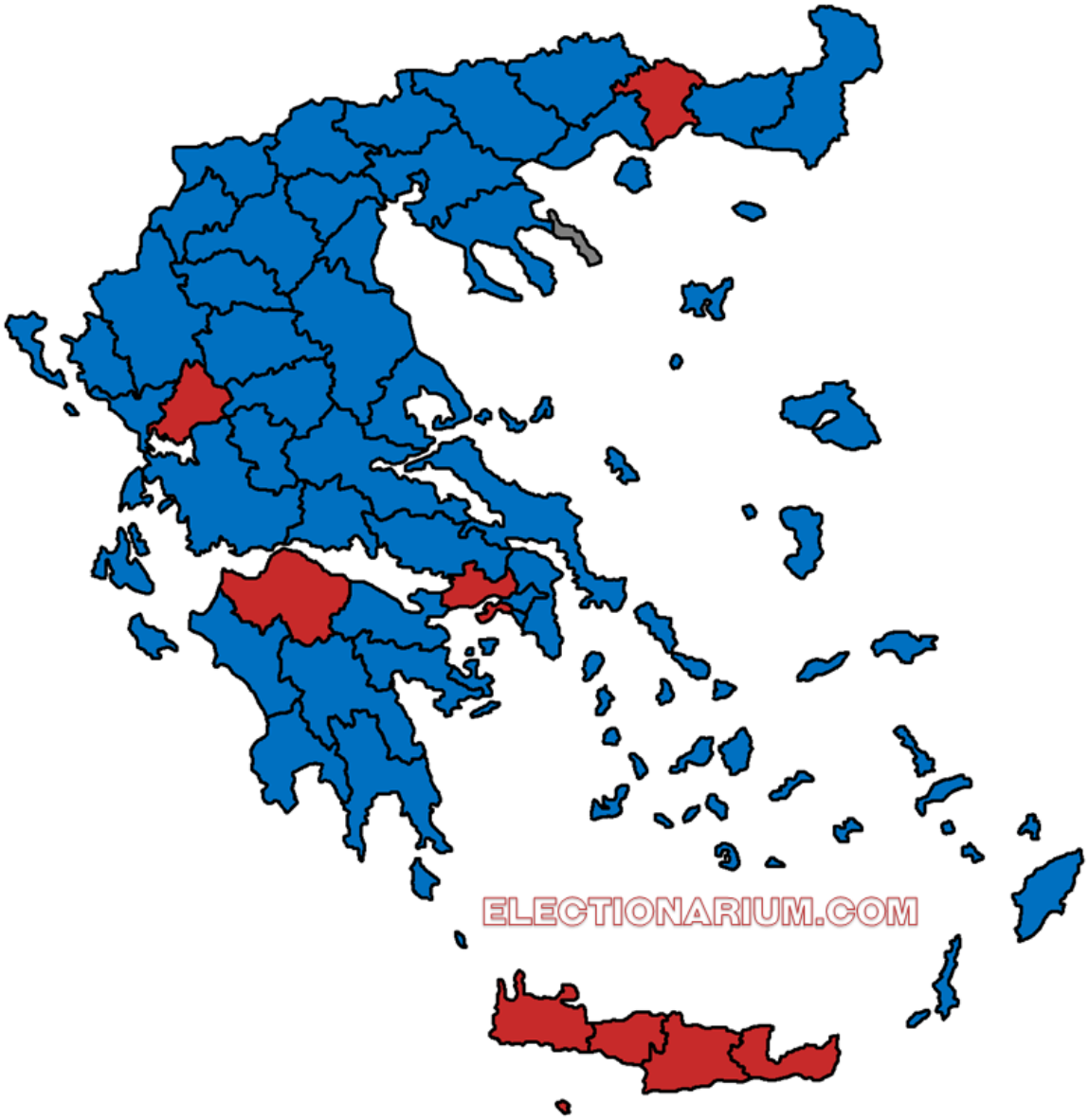
231	Seats elected in multi-seat constituencies proportionally
50	Seats awarded to the largest party or coalition as a bonus
12	Seats elected as a single national multi-seat constituency
7	Seats elected in single-seat constituencies (effectively First Past the Post)

CITY
UNIVERSITY OF LONDON
EST 1894

GREECE ELECTION 2019

SUBDIVISIONS

ELECTIONARIUM.COM



CITY
UNIVERSITY OF LONDON
EST 1894

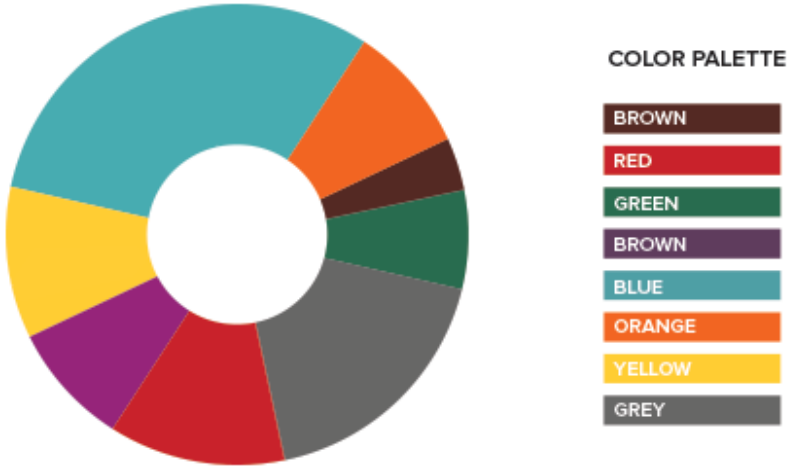
What makes a visualisation memorable?

Fig. 1. **Left:** The top twelve overall most memorable visualizations from our experiment (most to least memorable from top left to bottom right). **Middle:** The top twelve most memorable visualizations from our experiment when visualizations containing human recognizable cartoons or images are removed (most to least memorable from top left to bottom right). **Right:** The twelve least memorable visualizations from our experiment (most to least memorable from top left to bottom right).

http://cvcl.mit.edu/papers/Borkin_etal_MemorableVisualization_TVCG2013.pdf

Borkin et al. (2013): some conclusions

- **colour** and human **recognizable objects** enhance memorability
  - "we are best at remembering "natural" looking visualizations, as they are similar to scenes, objects, and people, and that pictorial and rounded features help memorability"



CITY
UNIVERSITY OF LONDON
EST 1894

Borkin et al. (2013): some conclusions (cont…)

- **unique visualization** types are more memorable than common graphs
  - "pictoral, grid/matrix, trees and networks, and diagrams" vs. "circles, area, points, bars, and lines"
  - "novel and unexpected visualizations can be better remembered than the visualizations with limited variability that we are exposed to since elementary school"

- **FUNCTIONAL ART:** "something that achieves beauty not through the subjective, freely wandering self-expression of the painter or sculptor, but through the careful and restrained tinkering of the engineer."

CITY
UNIVERSITY OF LONDON
EST 1894

Is ”memorable” the same as “effective”?

Using good graphs is important

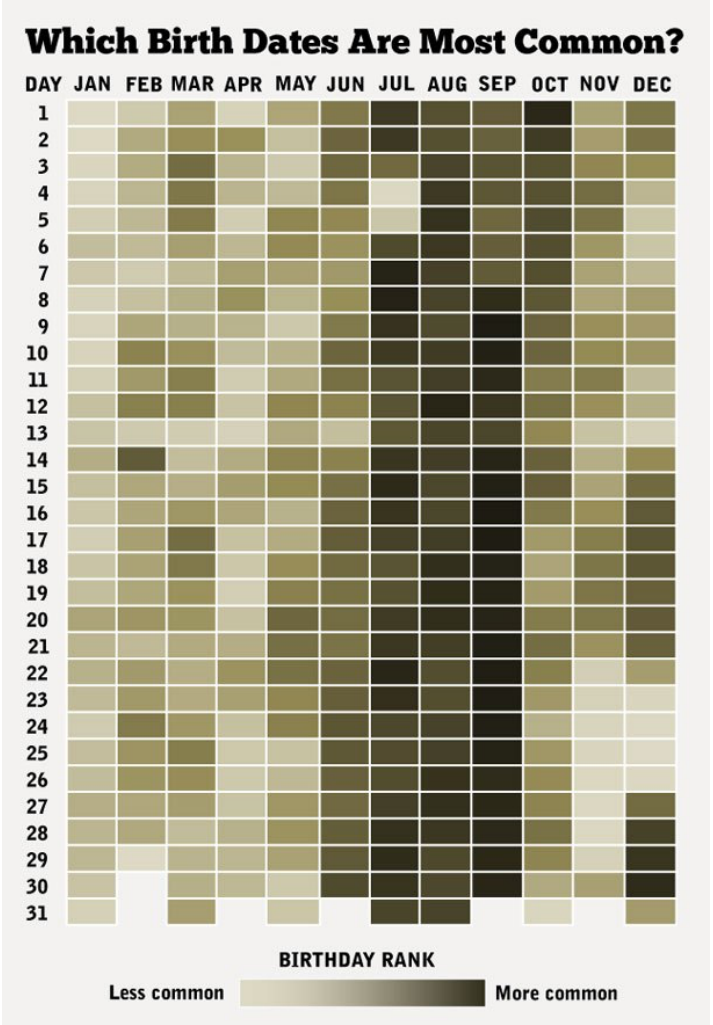
Graphs can highlight useful patterns in data

BUT

statistical analysis often is more than just graphs

DON'T STOP THERE

Graphical information normally complements the outcomes of **statistical tests** that allow us to estimate whether the data patterns are **meaningful**



CITY
UNIVERSITY OF LONDON
EST 1894

Summary Statistics

The ..... is the sum of all measurements divided by the number of measurements

Median	**329199**
Mean	**329238**
Mode	**329239**
Standard Deviation	**329253**

The ..... divides the distribution in half

Range	**334035**
Mean	**334046**
Mode	**334047**
Median	**334062**

The mode is .....

The arithmetic average	**334063**
The most frequent score	**335208**
The standard score	**335209**
A measure of dispersion	**335210**

Which of the following is a measure of central tendency?

The mean	**335303**
The standard deviation	**335304**
Both	**335306**
Neither	**335880**

The standard deviation is ……..

The square of the variance	**335885**
Smaller than the mean	**338364**
A measure of dispersion	**339018**
All of the above	**341176**

Central Tendency

the most representative value

Central Tendency – 'average'

- ARITHMETIC MEAN
  - Numeric scale

- Sum of all the measurements divided by the number of measurements

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

DATA: { 16, 17, 10, 13, 20, 18, 13, 14, 18 }

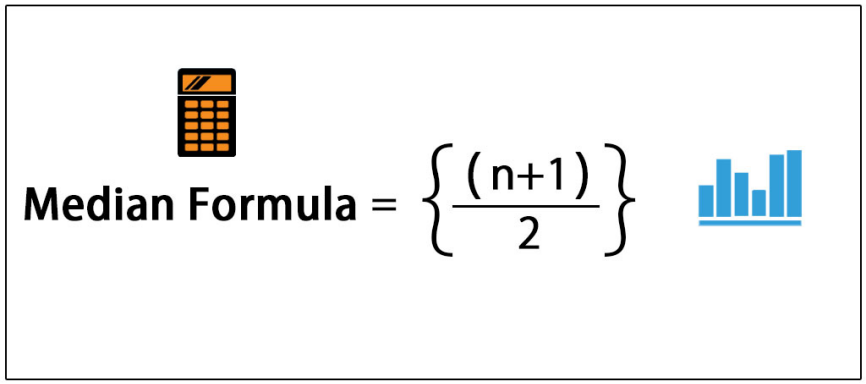
$$\bar{x} = \frac{16 + 17 + 10 + 13 + 20 + 18 + 13 + 14 + 18}{9}$$

$$\bar{x} = \frac{139}{9}$$

$\bar{x} = 15.444$ (rounded to three decimal places)

Central Tendency – 'average'

- MEDIAN
  - Ordinal
  - ratio



The middle value when data are arranged from smallest to larges in an array

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

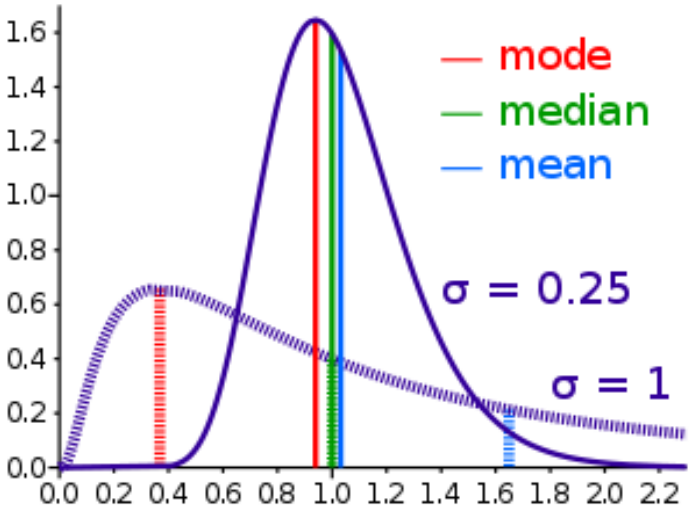
Median = (4 + 5) ÷ 2

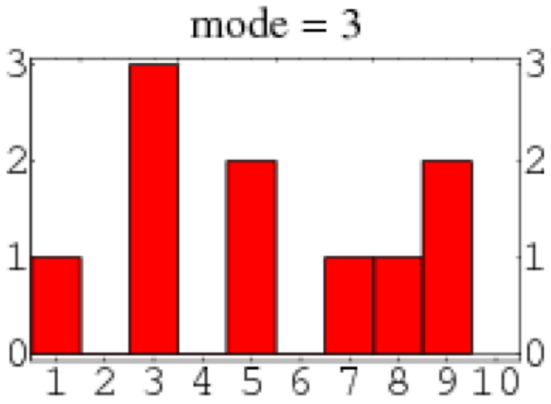
= 4.5

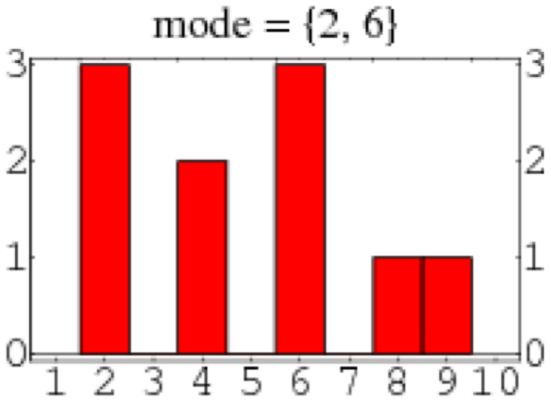
Central Tendency – ‘average’

- MODE
  - Nominal

The score with the highest frequency and the most frequent score







CITY
UNIVERSITY OF LONDON
EST 1894

Dispersion

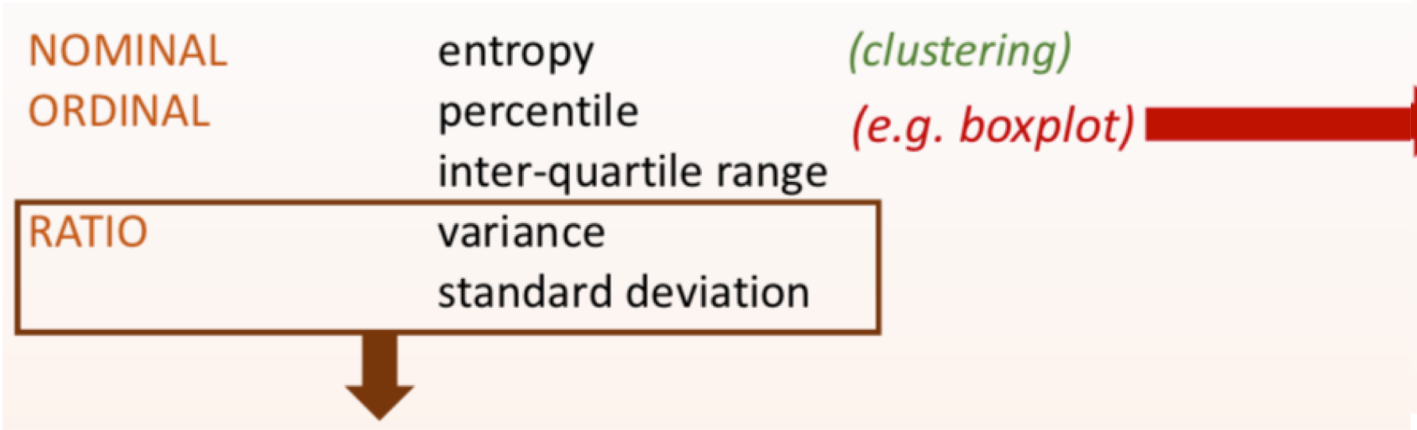
amount of variation around the most representative value

Is the average always a good summary?

Is the average always a good summary?

Dispersion – 'width'

- **how far are things from the centre?**
  - **if** our **measurements are very similar** to one another, then the **mean is** a **good representation** of the distribution as a whole
  - **if** they **are very different**, then the mean will be a less good summary



Dispersion – ‘width’

- Which ratio distribution has greatest dispersion?

CITY
UNIVERSITY OF LONDON
EST 1894

Dispersion – 'width'

- STANDARD DEVIATION
  - the square-root of the variance

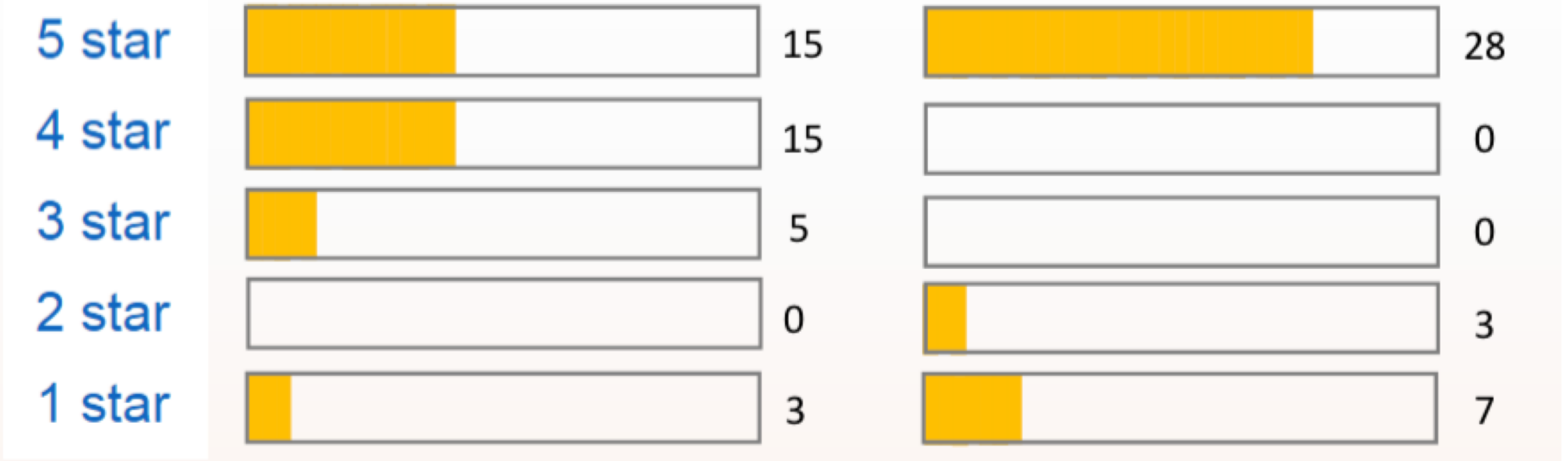
- It uses the same units as the original measurements themselves

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

CITY
UNIVERSITY OF LONDON
EST 1894

standard deviation

4.0 out of 5 stars



SD: 1.126747

SD: 1.668254

CITY
UNIVERSITY OF LONDON
EST 1894

Dispersion – ‘width’

- VARIANCE
  - the **difference** between each ***measurement*** and the ***mean***
    - squaring those differences
    - and dividing by the number of measurements

equivalent to the 'average squared deviation from the mean’

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

Dispersion – ‘width’

- Which **ratio** distribution has greatest dispersion?

CITY
UNIVERSITY OF LONDON
EST 1894

Standardising

allows us to compare dispersion between distributions

CITY
UNIVERSITY OF LONDON
EST 1894

How do we compare centimetres and kilograms?

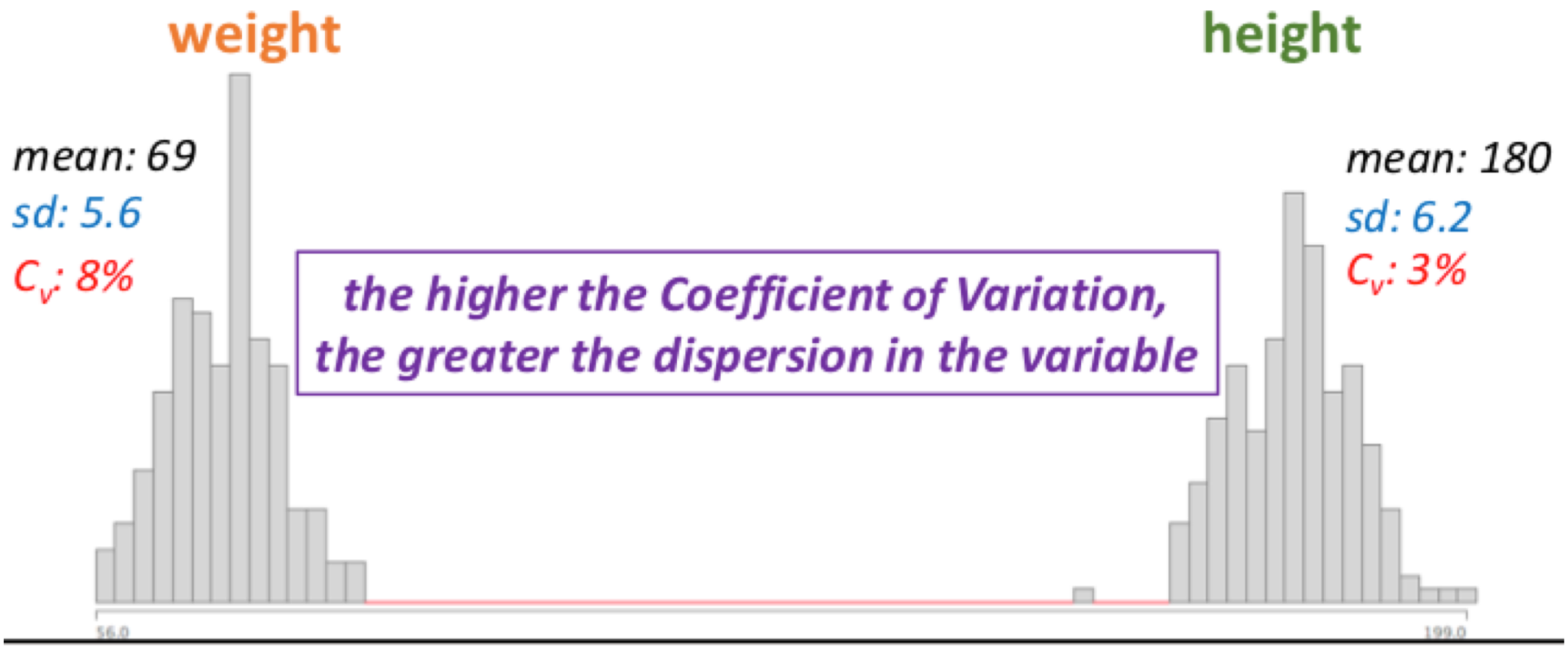
- COEFFICIENT OF VARIATION (CV)
  - measure of **overall dispersion** relative to the **mean** (μ)

$$CV = \frac{\sigma}{\mu}$$

$$CV\ (\%) = \left(\frac{Standard\ deviation}{Mean}\right) \times 100$$

Dispersion – ‘width’

- Which **ratio** distribution has greatest dispersion?



CITY
UNIVERSITY OF LONDON
EST 1894

How do we compare centimetres and kilograms?

- Z-SCORE
  - how **extreme** a **value** is in a given distribution?
    - difference between any individual value (x) and the mean (μ)
    - in terms of standard deviation (σ)

the z-score has the same meaning in one unit or another, facilitating comparisons

$$z = \frac{x - \mu}{\sigma}$$

Sampling

getting the right things to measure

What is μ?

The standard error of the mean	**341177**
The mean of the sample	**341178**
The mean of the population	**341182**
The mean of a normal distribution	**341258**

CITY
UNIVERSITY OF LONDON
EST 1894

Sampling

- POPULATION
  - everything we are interested in

- STATISTICAL SAMPLE
  - the things you get to measure **measurements** made in your **study**

making the research process manageable (efficient)

CITY
UNIVERSITY OF LONDON
EST 1894

Sampling: bias

- BIAS
  - systematic **differences between** the **sample** and the **population**

- SAMPLING FRAME
  - the **subset** of the **population** from which we can actually sample

In order for us to generalise from the sample to the population, it is important that the sample is :

Random	**341259**
Large	**341518**
Representative	**341519**
All of the above	**341523**

CITY
UNIVERSITY OF LONDON
EST 1894

Sampling: Methods

- RANDOM SAMPLING (equally likely any to be selected)
  - http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/samplinghirev2.shtml
- SYSTEMATIC SAMPLING
  - e.g. “every 10th or 15th customer”
- STRATIFIED SAMPLING (representative of different groups)
  - http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/samplinghirev3.shtml
- IMPORTANT:
  - each item in the sampling frame should have an **equal chance of selection** (be mechanical to avoid subjective influence)

Statistical Conventions

POPULATION	SAMPLE
Mean: μ (“mu”)	Mean: ”x-bar”
Standard deviation: σ (”sigma”)	Standard deviation: **s** or $\mathbf{s_x}$

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

standard deviation for sample

“statistic”

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

“parameter”

standard deviation for population

CITY
UNIVERSITY OF LONDON
EST 1894

Sampling: Reliability

- Is our sample representative?

- Is the sample mean a good indication of population mean?
  - $\bar{x}$ vs. μ

- How do we know if our sample is any good?
- Can we estimate the reliability of our sample?

Statistics Without Tears: a Primer for Non-Mathematicians
Rowntree, Derek 1991

Which is likely to be larger, a population mean or a sample mean?

Both means should be very similar or identical	**341524**
The sample mean	**341525**
The population mean	**341536**
No generalisation can be made since it will always depend on the sample that is taken	**341541**

Sampling: Reliability

STANDARD ERROR OF THE MEAN (SE)

- Is the amount of likely variation between
  - the **mean of a sample**
  - and the **mean of the population**

- Quantifies reliability of sample

Larger samples have lower standard error and are therefore more reliable

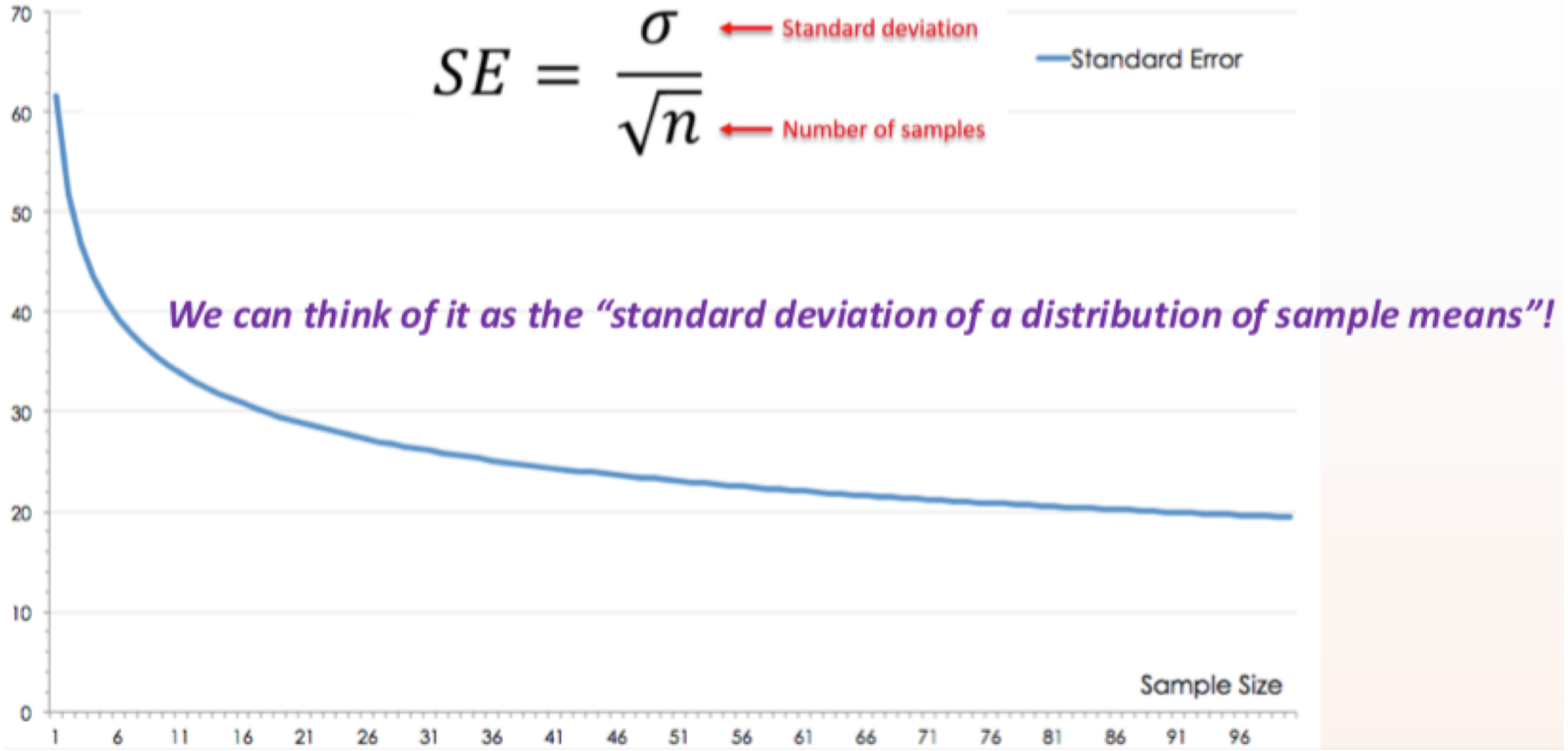
– **with diminishing/making less returns**

Standard Error of the Mean

$$SE = \frac{\sigma}{\sqrt{n}}$$

Standard deviation

Number of samples



We can think of it as the "standard deviation of a distribution of sample means"!

CITY
UNIVERSITY OF LONDON
EST 1894

Sampling can be crucial

- based on past experience in RMPI:
  - the selection of **inappropriate sample sizes and** inappropriate **sampling strategies results** in **more wasted work and inconclusive results** than any other factor **in MSc dissertations**

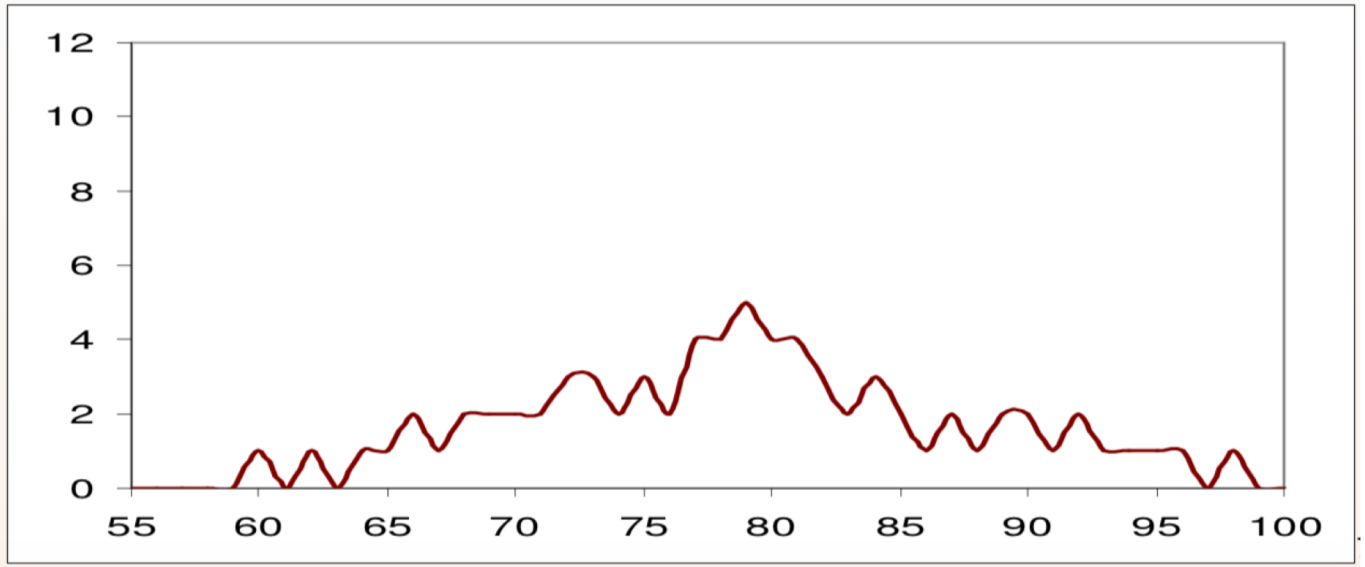
CITY
UNIVERSITY OF LONDON
EST 1894

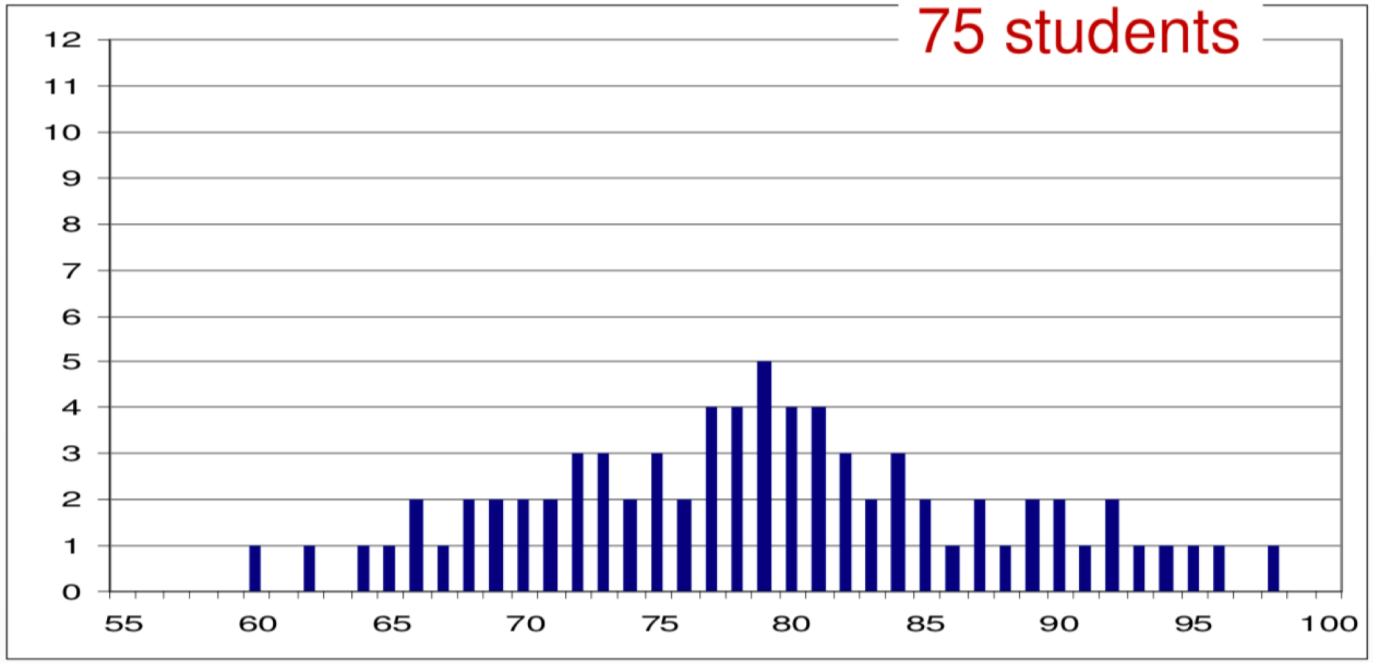
Introducing the Normal Distribution

All statements are true of the NORMAL distribution EXCEPT

Mean, median and mode coincide	**341554**
The most common scores lie near the mean	**341586**
Its mean value is always 0	**341675**
The right half of the curve is a mirror image of the left half	**341676**

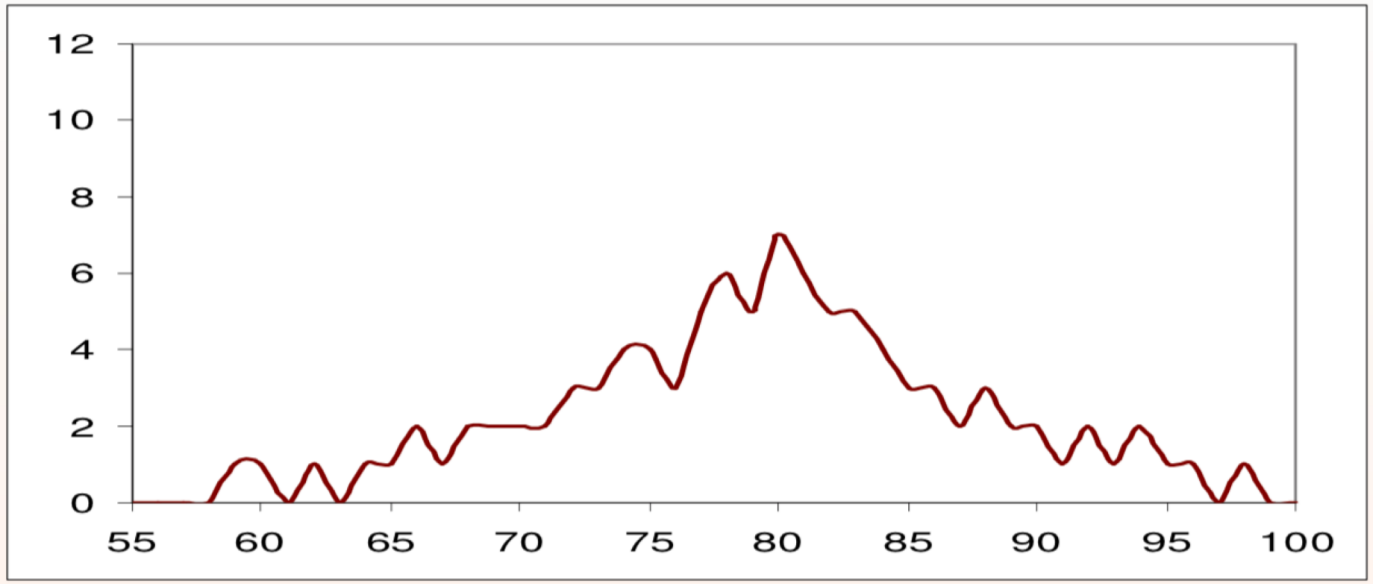
CITY
UNIVERSITY OF LONDON

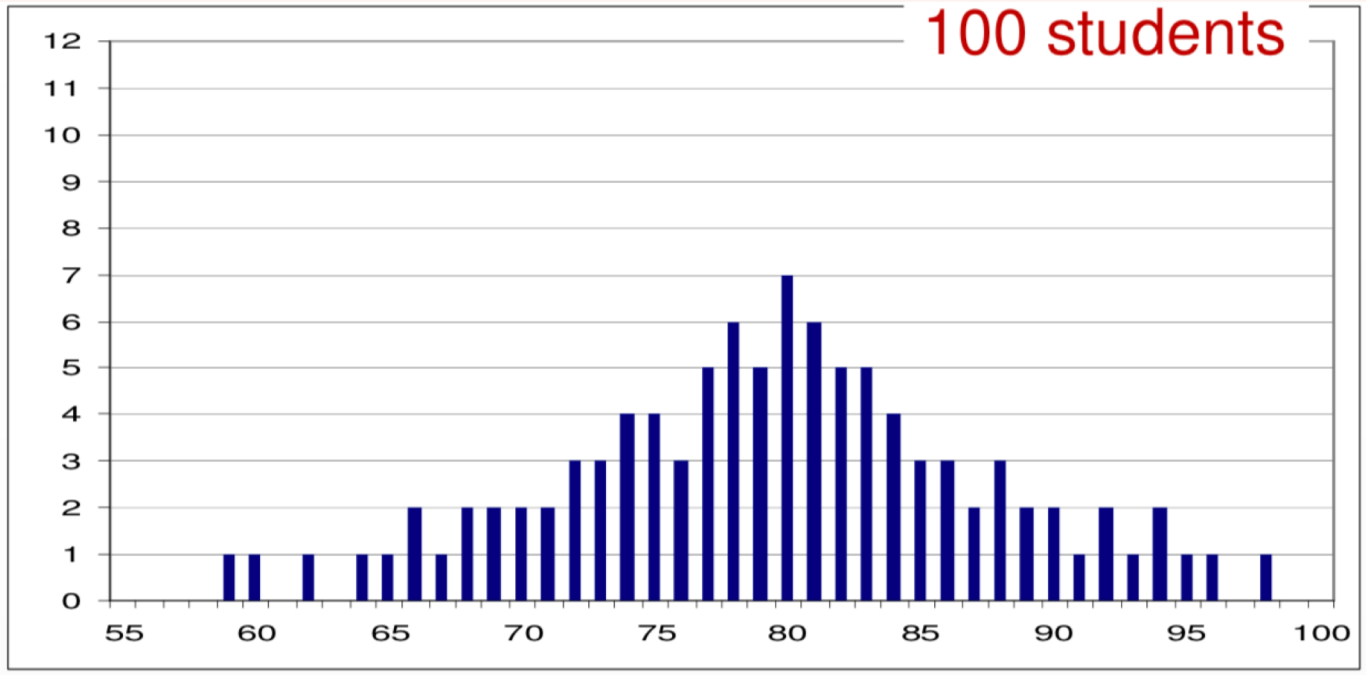




creativity score	number of students
55	0
56	0
57	0
58	0
59	0
60	1
61	0
62	1
63	0
64	1
65	1
66	2
67	1
68	2
69	2
70	2
71	2
72	3
73	3
74	2
75	3
76	2
77	4
78	4
79	5
80	4
81	4
82	3
83	2
84	3
85	2
86	1
87	2
88	1
89	2
90	2
91	1
92	2
93	1
94	1
95	1
96	1
97	0
98	1
99	0
100	0

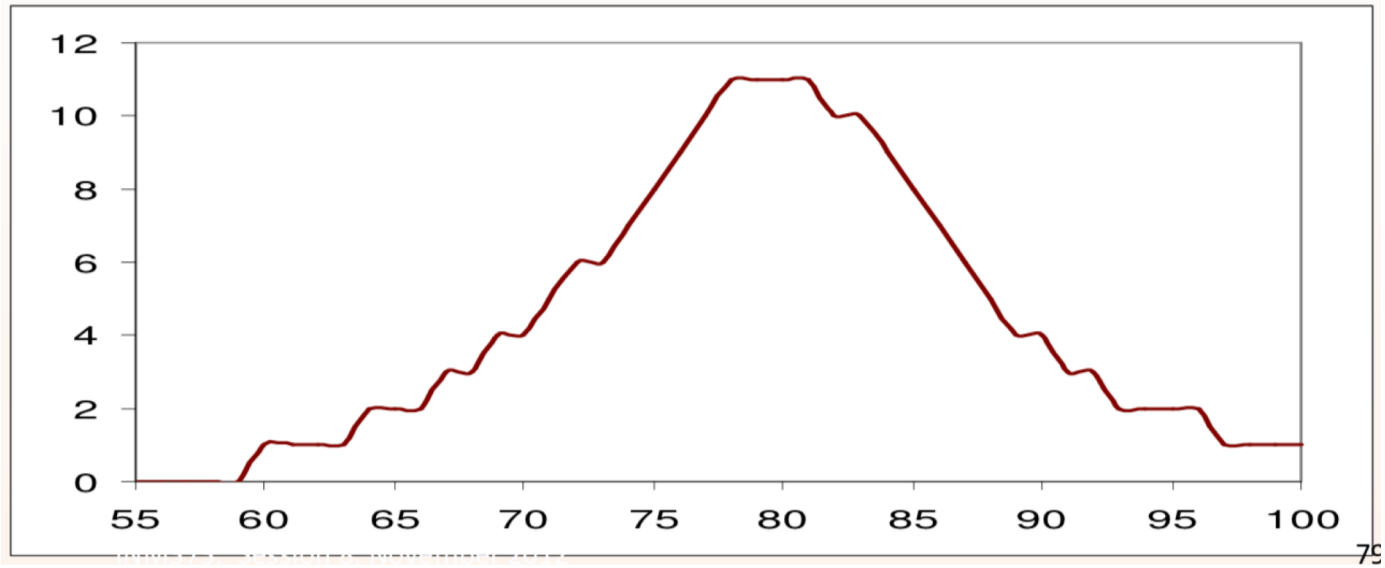
CITY

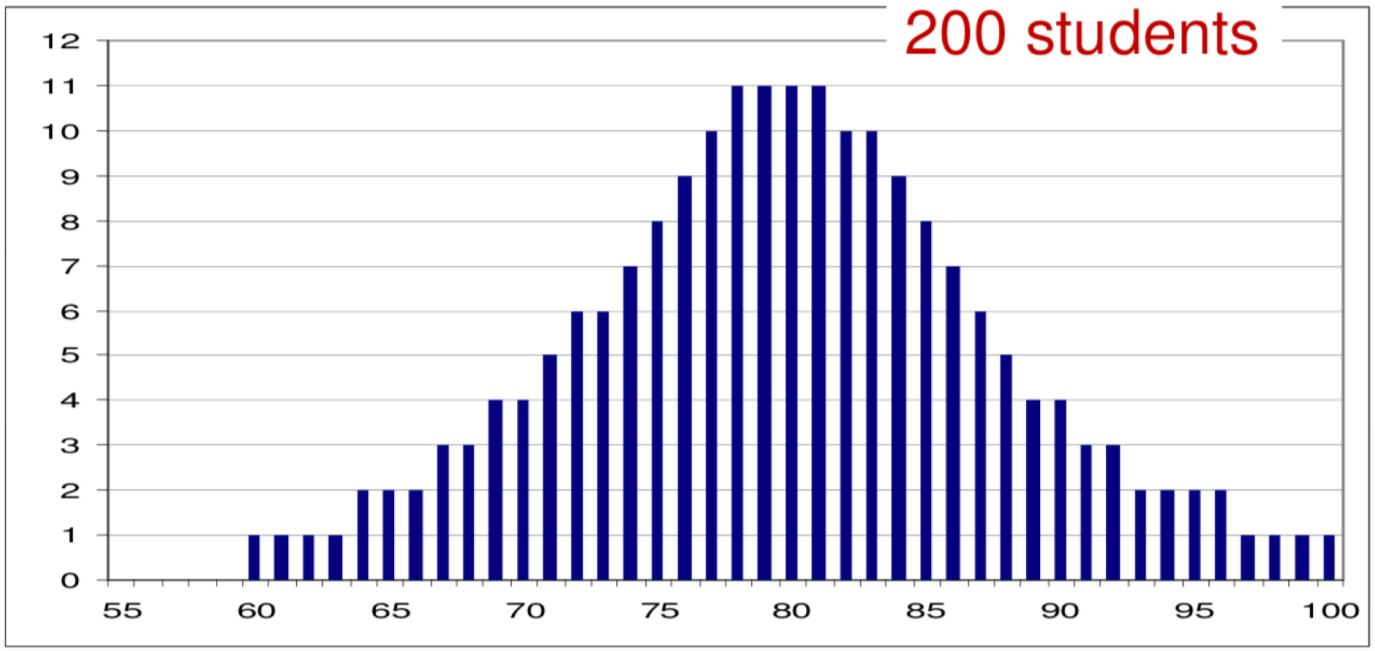




creativity score	number of students
55	0
56	0
57	0
58	0
59	1
60	1
61	0
62	1
63	0
64	1
65	1
66	2
67	1
68	2
69	2
70	2
71	2
72	3
73	3
74	4
75	4
76	3
77	5
78	6
79	5
80	7
81	6
82	5
83	5
84	4
85	3
86	3
87	2
88	3
89	2
90	2
91	1
92	2
93	1
94	2
95	1
96	1
97	0
98	1
99	0
100	0

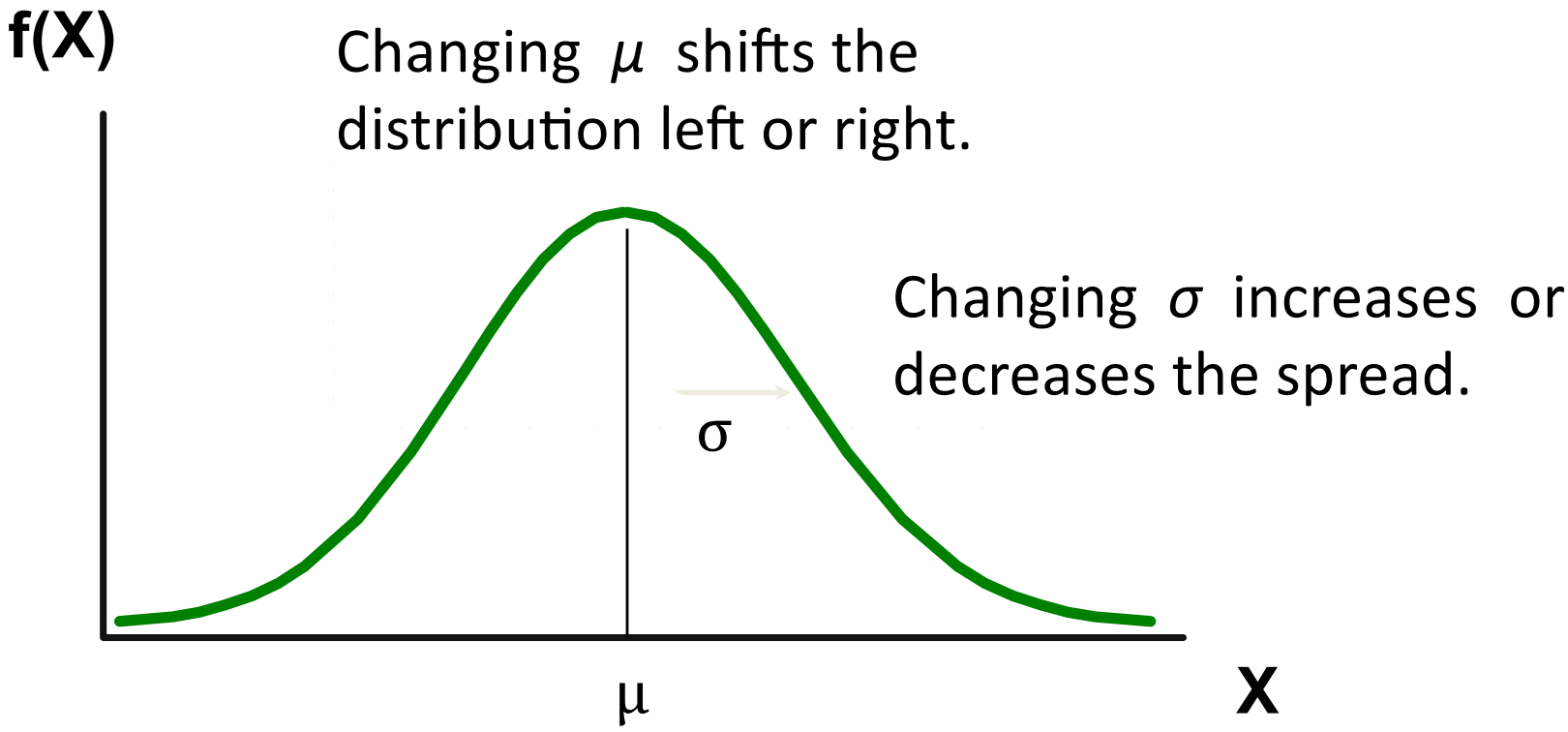
CITY
UNIVERSITY OF LONDON
EST 1894





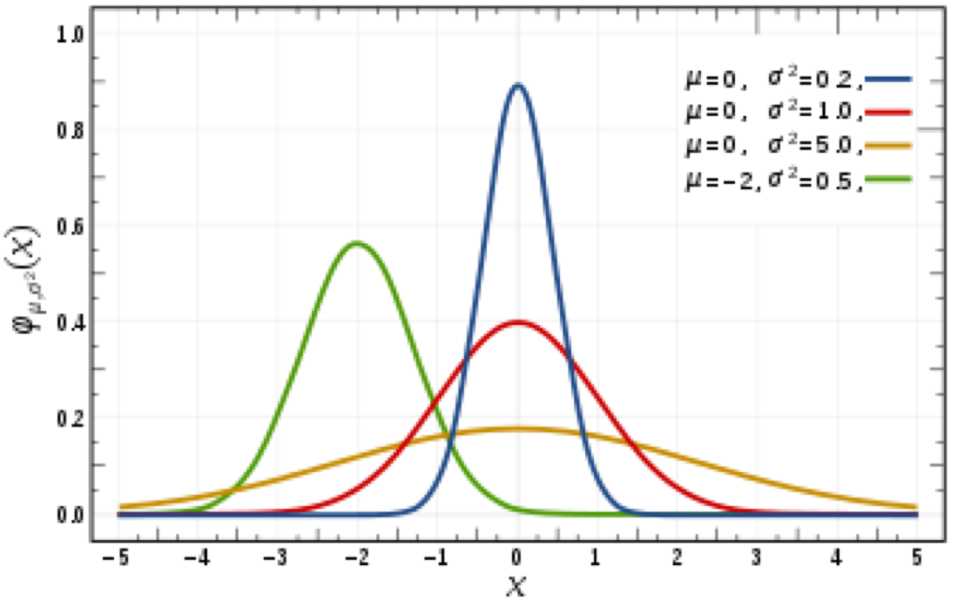
creativity score	number of students
55	0
56	0
57	0
58	0
59	0
60	1
61	1
62	1
63	1
64	2
65	2
66	2
67	3
68	3
69	4
70	4
71	5
72	6
73	6
74	7
75	8
76	9
77	10
78	11
79	11
80	11
81	11
82	10
83	10
84	9
85	8
86	7
87	6
88	5
89	4
90	4
91	3
92	3
93	2
94	2
95	2
96	2
97	1
98	1
99	1
100	1

The Normal Distribution



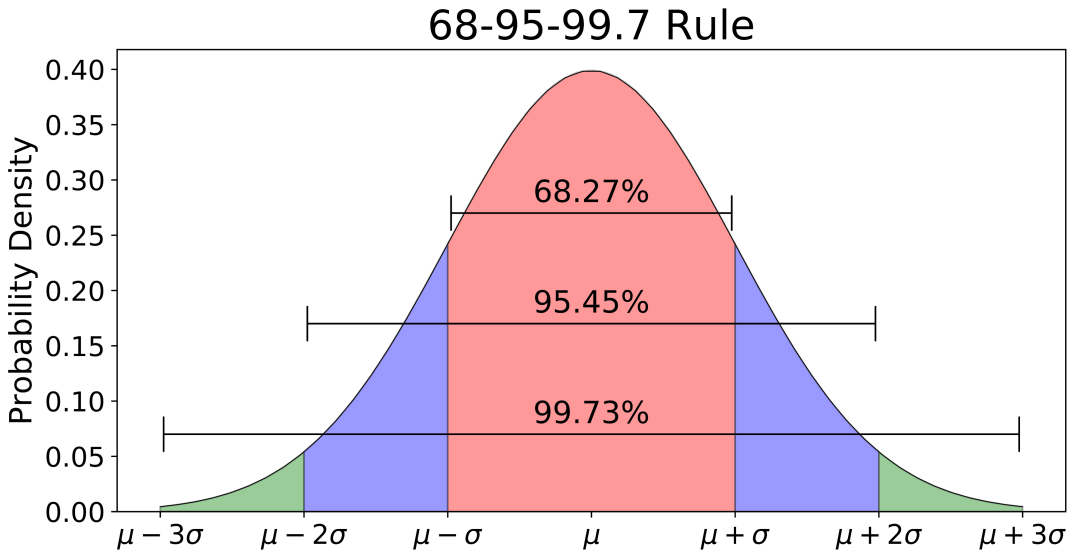
The Normal Distribution (cont…)

- Gaussian curve bell shaped
  - the most frequently occurring values are at the centre of the distribution
- The ***mean***, ***median*** and ***mode*** all coincide at this point



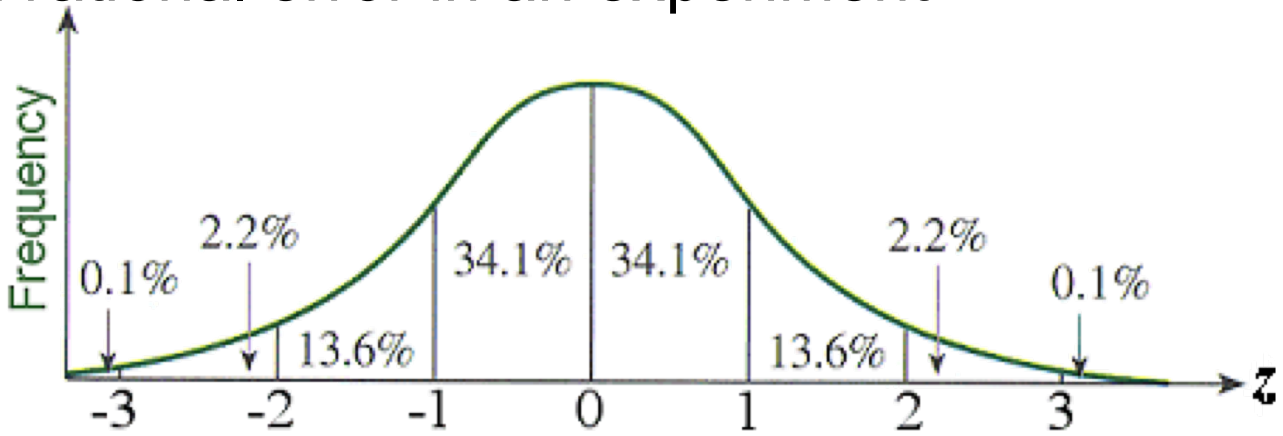
The Normal Distribution (cont…)

- Values which occur **further away from the mean are less and less frequent**
- Quickly **falls off towards plus/minus** infinity
- **Symmetrical**: there are as many values above the mean as there are below an idealisation



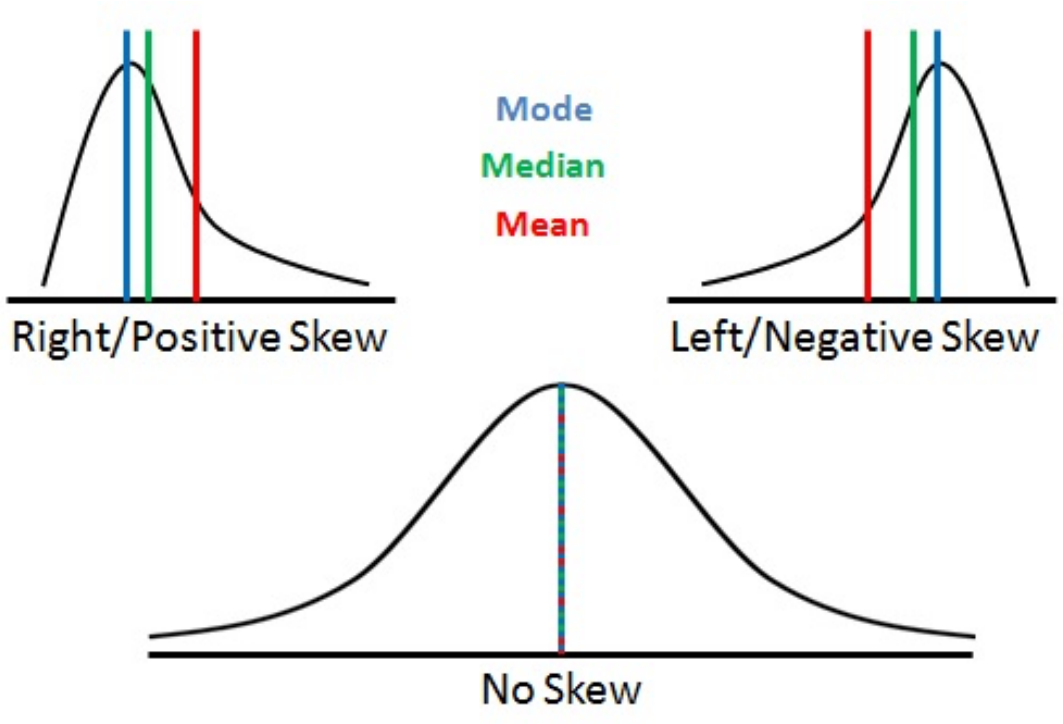
The Normal Distribution (cont…)

- **Many naturally occurring phenomena:**
- IQ
- height, weight, shoe size
- body temperature
- exam marks?
- observational error in an experiment



Skewed Distributions

- Examples:
  - salaries
    (positive skew)
  - exam marks
    (negative skew)



- We can **quantify** '**skewness**' (kurtosis)
- We can **transform skewed distributions** so that they **exhibit** an **approximately normal distribution**

CITY
UNIVERSITY OF LONDON