Introduction

> The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.
>
> — Nate Silver, *The Signal and the Noise*[1]

Why We Need Statistics

Harold Shipman was Britain's most prolific convicted murderer, though he does not fit the archetypal profile of a serial killer. A mild-mannered family doctor working in a suburb of Manchester, between 1975 and 1998 he injected at least 215 of his mostly elderly patients with a massive opiate overdose. He finally made the mistake of forging the will of one of his victims so as to leave him some money: her daughter was a solicitor, suspicions were aroused, and forensic analysis of his

A.S

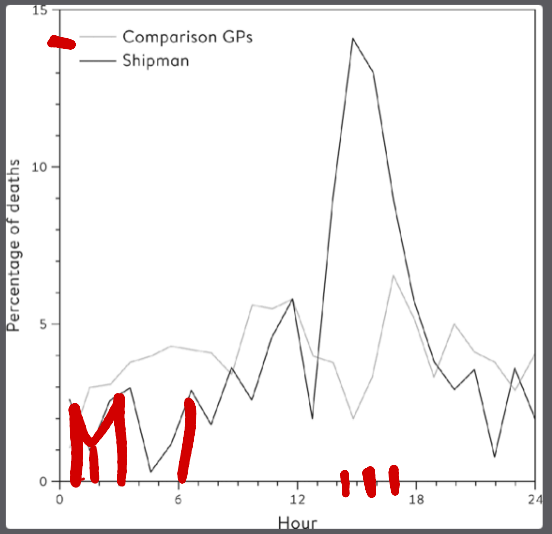


Figure 0.2
The time at which Harold Shipman's patients died, compared to the times at which patients of other local general practitioners died. The pattern does not require sophisticated statistical analysis.

BARPLOT : VIZ TOOL FOR ADDING AESTHETICS TO SUMMARY OF CATEGORICAL DATA
↳ PERC / FREQ

X - AXIS : DIFFERENT CATEGORIES

Y - AXIS : FREQ / PERC ASSOCIATED TO EACH CATEGORY

NUMERICAL DATA

ACTUAL VALUES ARE INTERPRETABLE, CAN BE ORDERED, AMENABLE TO NUMERICAL CALCULATIONS SUCH AS SUMMING, MULTIPLYING BY A NUMBER ETC

- COUNT DATA (INTEGERS)
  - "HOW MANY"

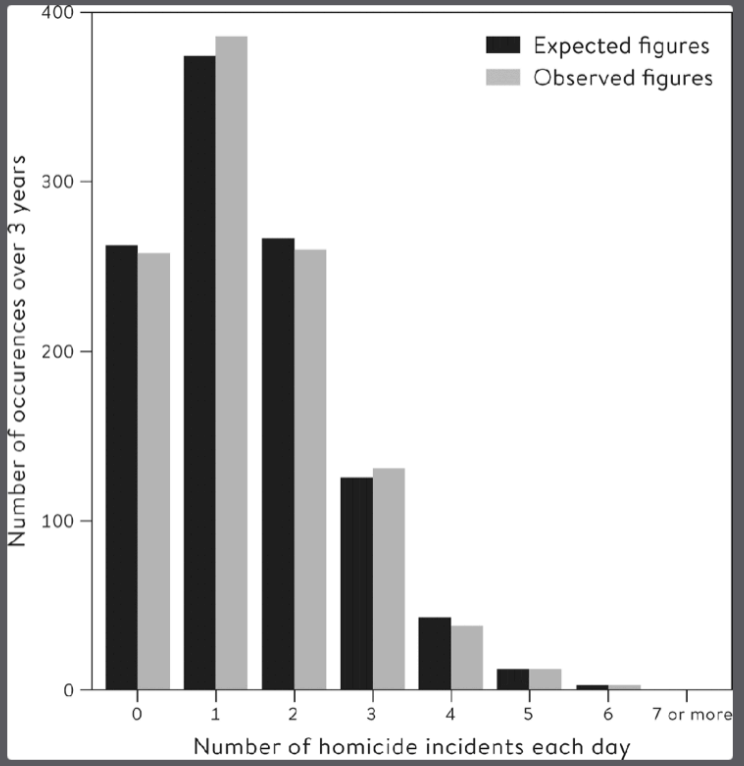


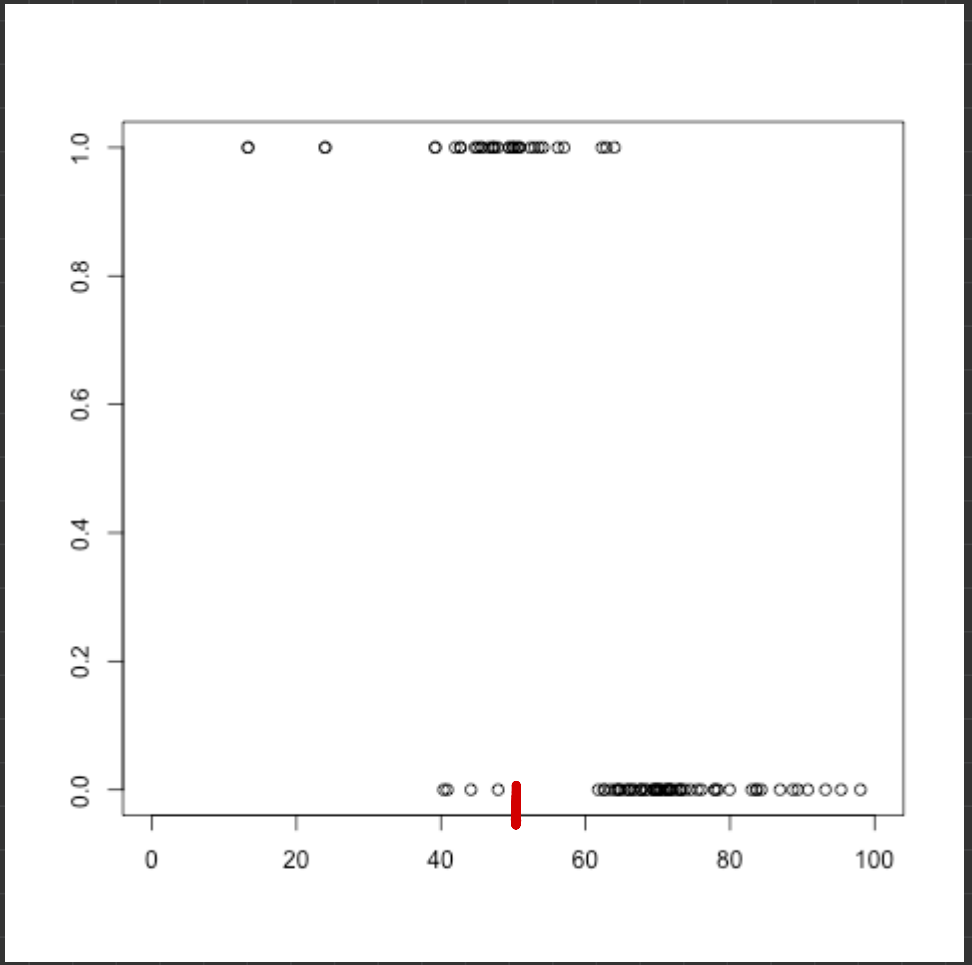
Figure 8.5
Observed and expected (assuming a Poisson distribution) daily number of recorded homicide incidents, 2014 to 2016, England and Wales.[3]

Focus on grey bars

A.S

CONTINUOUS DATA

- DATA FOR WHICH MANY DIFFERENT VALUES ARE POSSIBLE



i : A STUDENT

X_i : FINAL GRADE

Y_i : ATTENDED (0) OR NOT (1)

CHAPTER 2

Summarizing and Communicating Numbers. Lots of Numbers

Can we trust the wisdom of crowds?

In 1907 Francis Galton, cousin of Charles Darwin and polymath originator of identification using fingerprints, weather forecasts and eugenics,[fn1] wrote a letter to the prestigious science journal *Nature* about his visit to the Fat Stock and Poultry Exhibition in the port city of Plymouth. There he saw a large ox displayed and contestants paying sixpence to guess the 'dressed' weight of the resulting meat after the poor beast had been slaughtered. He got hold of 787 of the tickets that had been filled out and chose the middle value of 1,207 lb (547 kg) as the democratic choice, 'every other estimate being condemned as too high or too low by the majority of voters'. The dressed weight turned out to be 1,198 lb (543 kg), which was remarkably close to his choice based on the 787 votes.[1] Galton titled his letter '*Vox Populi*' (voice of the people), but this process of decision-making is now better known as the **wisdom of crowds**.

Galton carried out what we might now call a data summary: he took a mass of numbers written on tickets and reduced them to a single estimated weight of 1,207 lb. In this chapter we look at the

→ THE MEDIAN

WEAK LEARNER
(ENSEMBLE METHOD)

- SUMMARIZATION OF NUMERICAL DATA
  (LOTS OF NUMBER)
  - MEASURES / SUMMARIES OF LOCATION
    - SAMPLE MEDIAN

      (MIDDLE VALUE)

      EXAMPLE : $-100, -0.5, 2, 100, 1$

SAMPLE MEAN / AVERAGE

x_i: VALUE OF NUMERICAL VAR x ON THE i'TH "INDIVIDUAL"

$i = 1, \ldots, n$

$$\text{mean} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n$$

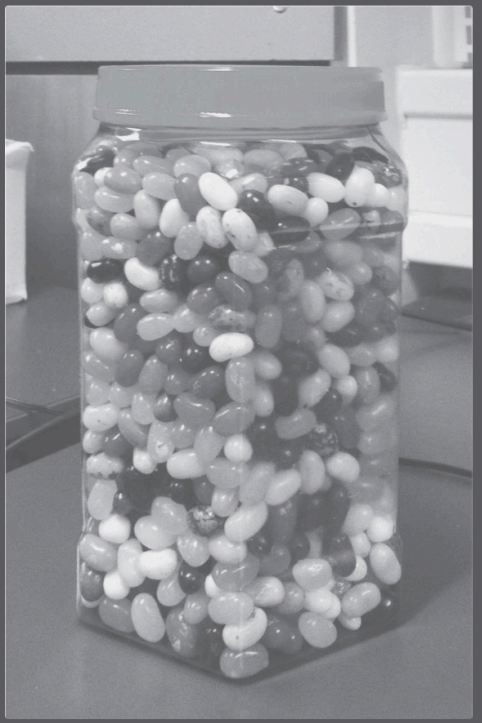


Figure 2.1
How many jelly beans are in this jar? We asked this on a YouTube video and got 915 responses. The answer will be given later.

A.S

A.S

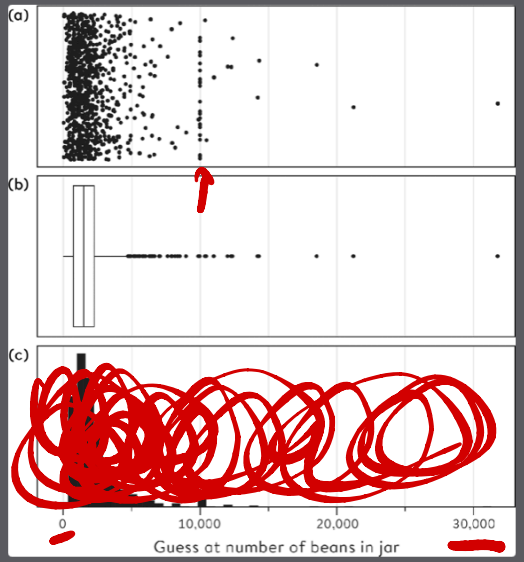


Figure 2.2
Different ways of showing the pattern of 915 guesses of the number of jelly beans in the jar. (a) A strip-chart or dot-diagram, with a jitter to prevent points lying on top of each other; (b) a box-and-whisker plot; (c) a histogram

← PLOT IF DATA

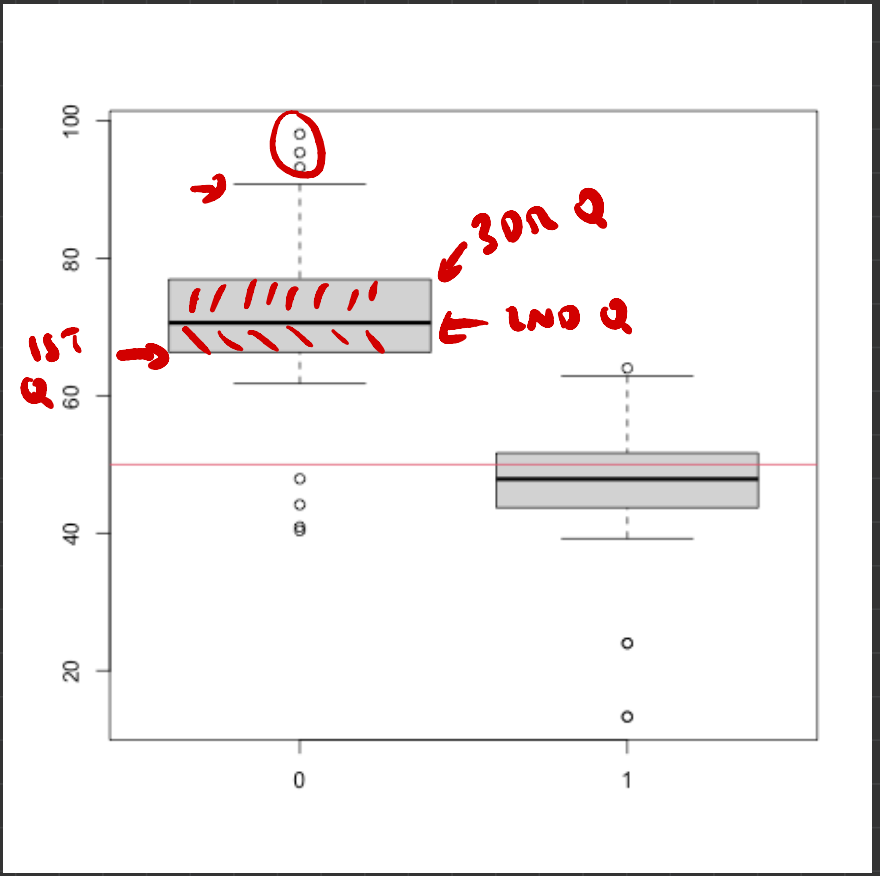
i: EACH DOT IS A PERSON'S GUESS

2:

QUARTILES

- 2ND QUARTILE IS MEDIAN
- 1ST Q
  - 25% OF DATA ON ITS LEFT
  - 75% " " " RIGHT
- 3RD Q
  - 25% ON THE RIGHT
  - 75% " " LEFT

GRADES DATASET



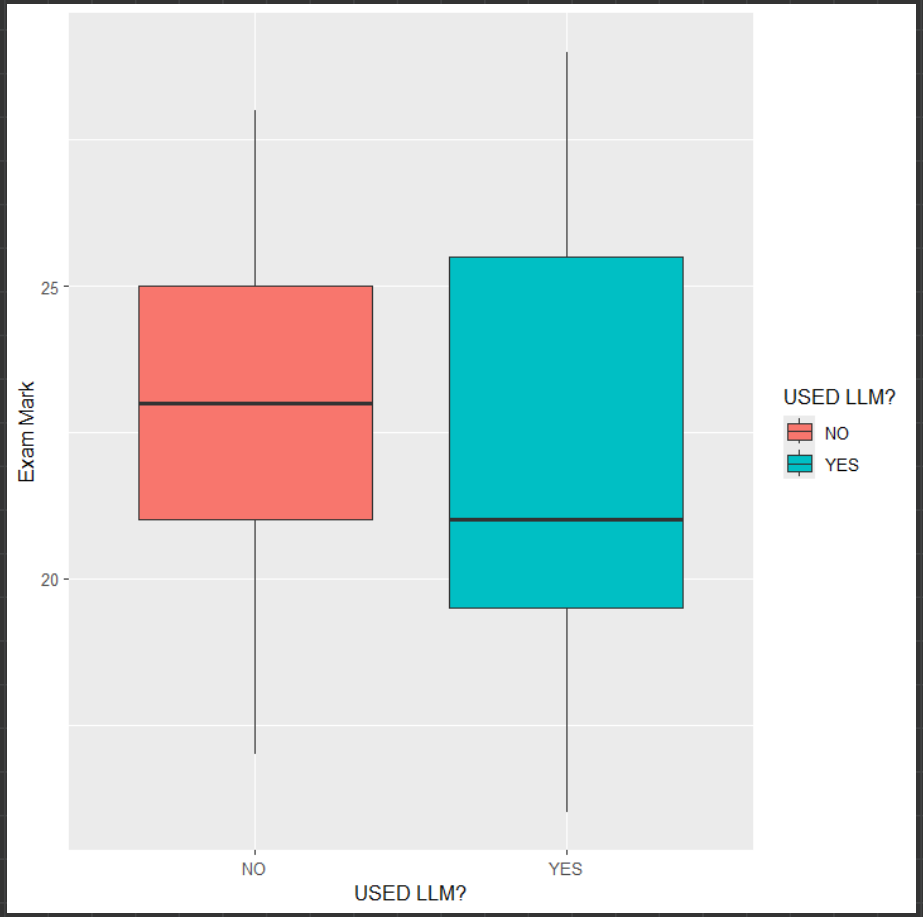
VIZ OF SUMMARY

↓

QUANTILES

↓

BOXPLOT


Exam Mark
25
20
NO
YES
USED LLM?
USED LLM?
NO
YES

ASYMMETRY VS SYMMETRY

(ESPECIALLY FOR POSITIVE DATA)

→ INFORMALLY: WHETHER 1ST & 3RD Q'S ARE SYMMETRIC AROUND MEDIAN

↳ WHEN NOT: SKEWNESS

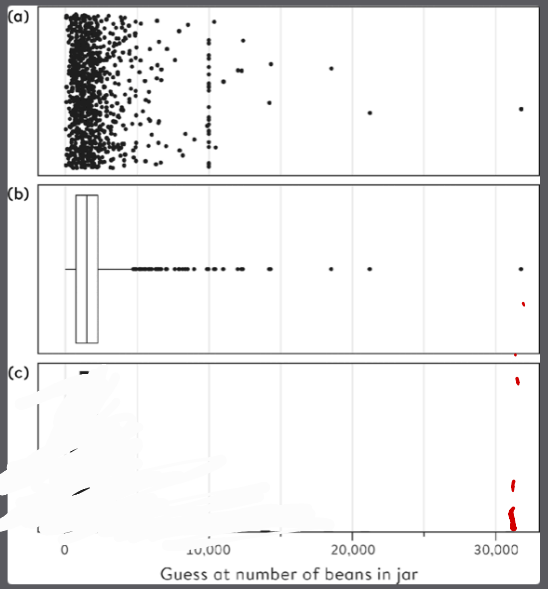


Figure 2.2
Different ways of showing the pattern of 915 guesses of the number of jelly beans in the jar. (a) A strip-chart or dot-diagram, with a jitter to prevent points lying on top of each other; (b) a box-and-whisker plot; (c) a histogram

← MANIFESTATION OF SKEWNESS

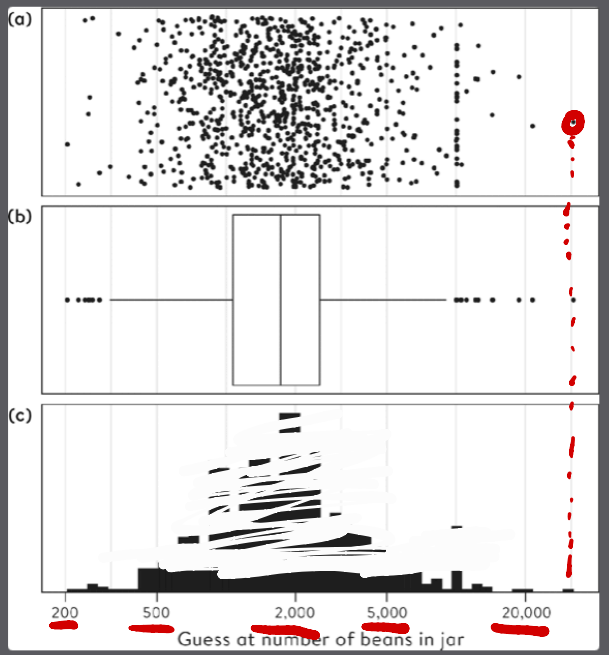
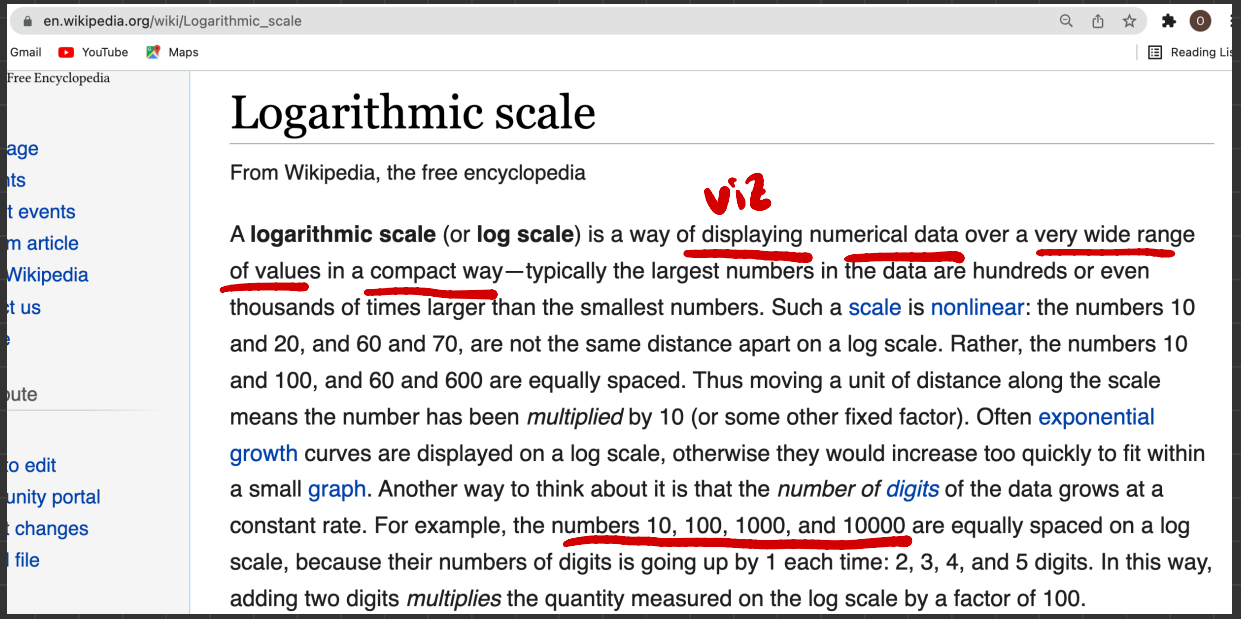


Figure 2.3
Graphical displays of the jelly-bean guesses plotted on a logarithmic scale. (a) Strip-chart; (b) box-and-whisker plot; (c) histogram all show a fairly symmetric pattern.

SAME DATA BUT ON LOGARITHMIC SCALE (LOG-SCALE)


en.wikipedia.org/wiki/Logarithmic_scale
Gmail
YouTube
Maps
Reading Lis
Free Encyclopedia
age
nts
t events
m article
Wikipedia
t us
oute
o edit
unity portal
t changes
l file
Logarithmic scale
From Wikipedia, the free encyclopedia
viz
A **logarithmic scale** (or **log scale**) is a way of displaying numerical data over a very wide range of values in a compact way—typically the largest numbers in the data are hundreds or even thousands of times larger than the smallest numbers. Such a scale is nonlinear: the numbers 10 and 20, and 60 and 70, are not the same distance apart on a log scale. Rather, the numbers 10 and 100, and 60 and 600 are equally spaced. Thus moving a unit of distance along the scale means the number has been *multiplied* by 10 (or some other fixed factor). Often exponential growth curves are displayed on a log scale, otherwise they would increase too quickly to fit within a small graph. Another way to think about it is that the *number of digits* of the data grows at a constant rate. For example, the numbers 10, 100, 1000, and 10000 are equally spaced on a log scale, because their numbers of digits is going up by 1 each time: 2, 3, 4, and 5 digits. In this way, adding two digits *multiplies* the quantity measured on the log scale by a factor of 100.

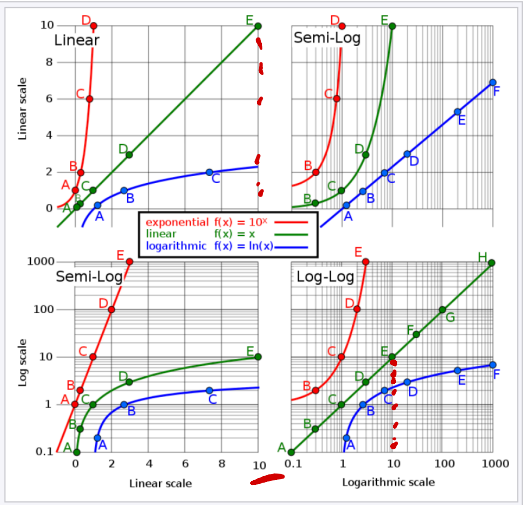
Graphic representation [edit]

The top left graph is linear in the X and Y axes, and the Y-axis ranges from 0 to 10. A base-10 log scale is used for the Y axis of the bottom left graph, and the Y axis ranges from 0.1 to 1,000.

The top right graph uses a log-10 scale for just the X axis, and the bottom right graph uses a log-10 scale for both the X axis and the Y axis.

Presentation of data on a logarithmic scale can be helpful when the data:

- covers a large range of values, since the use of the logarithms of the values rather than the actual values reduces a wide range to a more manageable size;
- may contain exponential laws or power laws, since these will show up as straight lines.



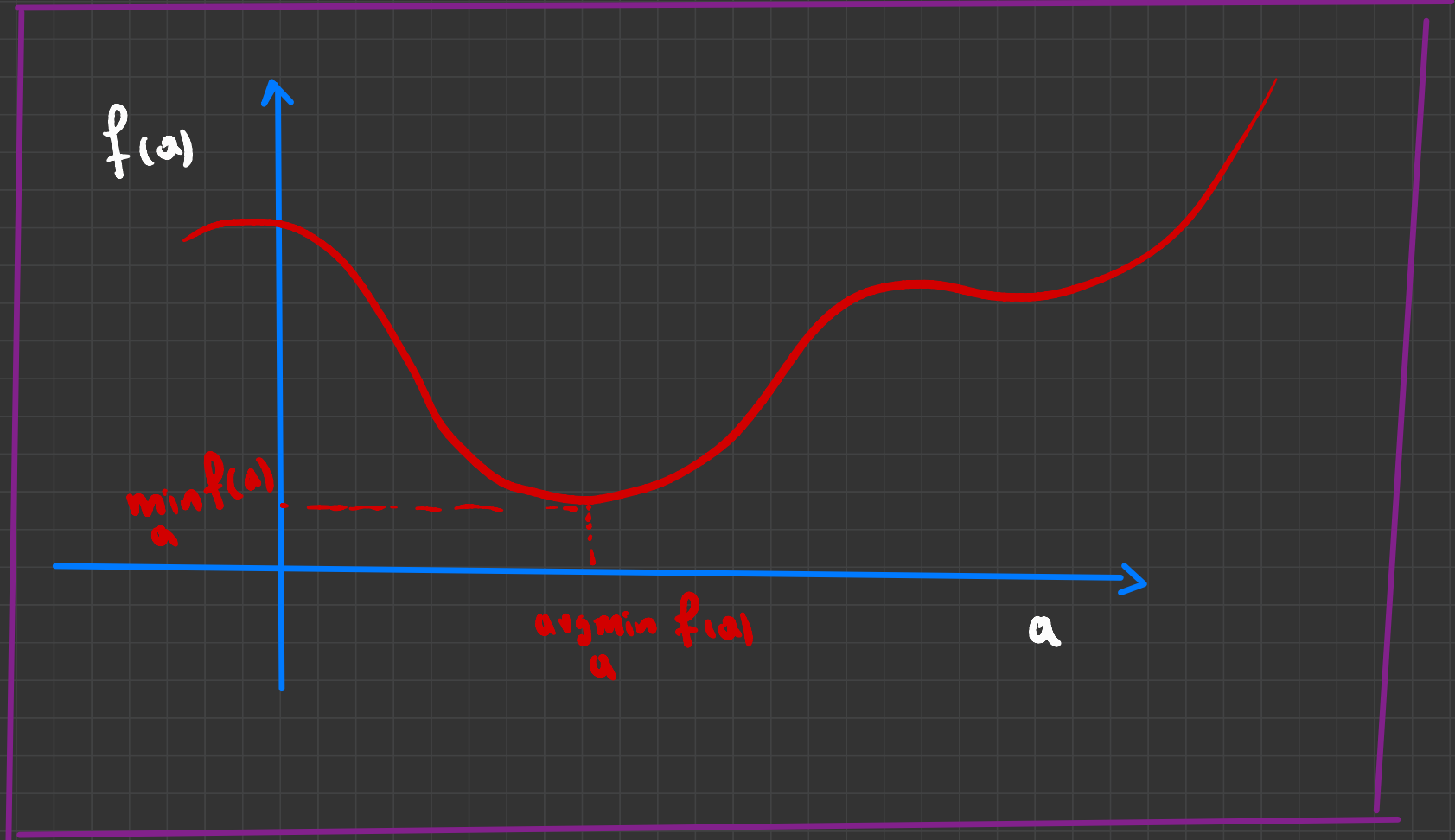
Various scales: lin–lin, lin–log, log–lin, and log–log. Plotted graphs are: $y = 10^x$ (red), $y = x$ (green), $y = \log_e(x)$ (blue).

VARIATIONAL FORMULATION OF MEAN & MEDIAN

$$\text{mean} = \arg\min_a \; g(a)$$

$$\text{median} = \arg\min_a \; f(a)$$

$$g(a) = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2 \qquad f(a) = \frac{1}{n} \sum_{i=1}^{n} |x_i - a|$$


f(a)
min f(a)
a
argmin f(a)
a
a