VARIATIONAL FORMULATION OF MEAN & MEDIAN

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

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

$$g(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 \quad \text{(L2 NORM)}$$

$$f(a) = \frac{1}{n}\sum_{i=1}^{n}|x_i - a| \quad \text{(L1 NORM)}$$

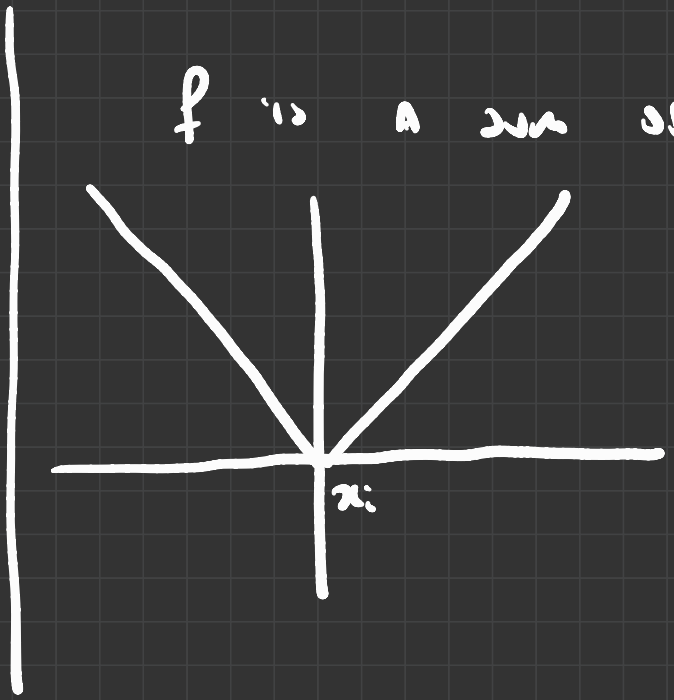
$$g(a) = \frac{1}{n} \sum_i (x_i - a)^2$$

$$g'(a) = \frac{2}{n} \sum_{i=1}^{n} (-1)(x_i - a) = 0$$

$$\Rightarrow \sum_i x_i = n a \Rightarrow \underline{a = \frac{1}{n} \sum_i x_i}$$

WHAT ABOUT

$$f(a) = \frac{1}{n} \sum_i |x_i - a|$$




a →

MEAN vs MEDIAN ('INFLUENCE VS ROBUSTNESS')

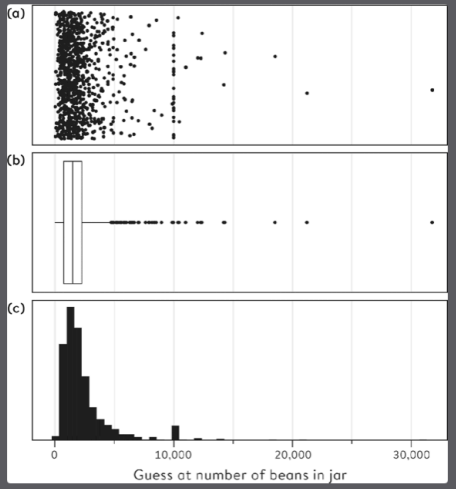


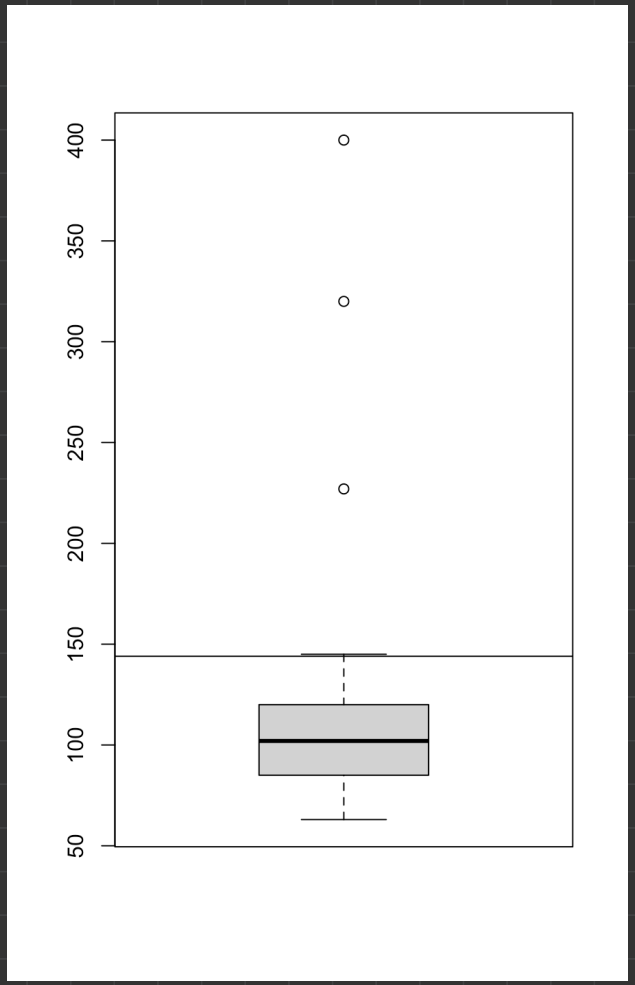
Figure 2.2
Different ways of showing the pattern of 915 guesses of the number of jelly beans in the jar. (a) A strip-chart or dot-diagram, with a jitter to prevent points lying on top of each other; (b) a box-and-whisker plot; (c) a histogram

MEAN : 2408

MEDIAN : 1775

TRUTH : 1616

Ping-Pong Results



MEASURES / SUMMARIES OF SPREAD / DISPERSION / VARIABILITY

- RANGE : MAX - MIN
  - VERY UNINFORMATIVE
- INTERQUANTILE RANGE : 3RD Q - 1ST Q

- STANDARD DEVIATION

$$sd = \sqrt{var}$$

$$var = \frac{1}{n} \sum_{i=1}^{n} (x_i - mean)^2 = \min_a \frac{1}{n} \sum_i (x_i - a)^2$$

- MEAN ABSOLUTE DEVIATION

$$MAD = \frac{1}{n} \sum_i |x_i - median| = \min_a \frac{1}{n} \sum_i |x_i - a|$$

HISTOGRAMS

THIS IS NOT A SUMMARY OR VISUALIZATION OF THE DATA.

↳ IT IS AN INFERENCE ABOUT DATA

CREATING A HISTOGRAM

1ST STEP : DEFINE DATA INTERVALS
↳ BINS

2ND STEP : FREQ OF NUMBER OF DATA IN EACH BIN

3RD STEP : BARPLOT

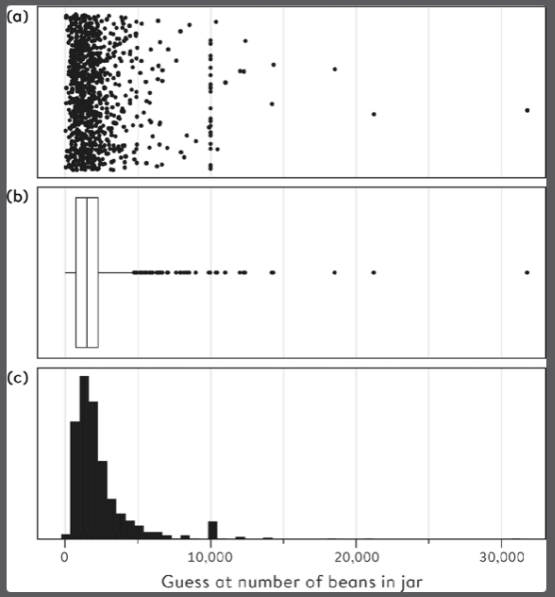
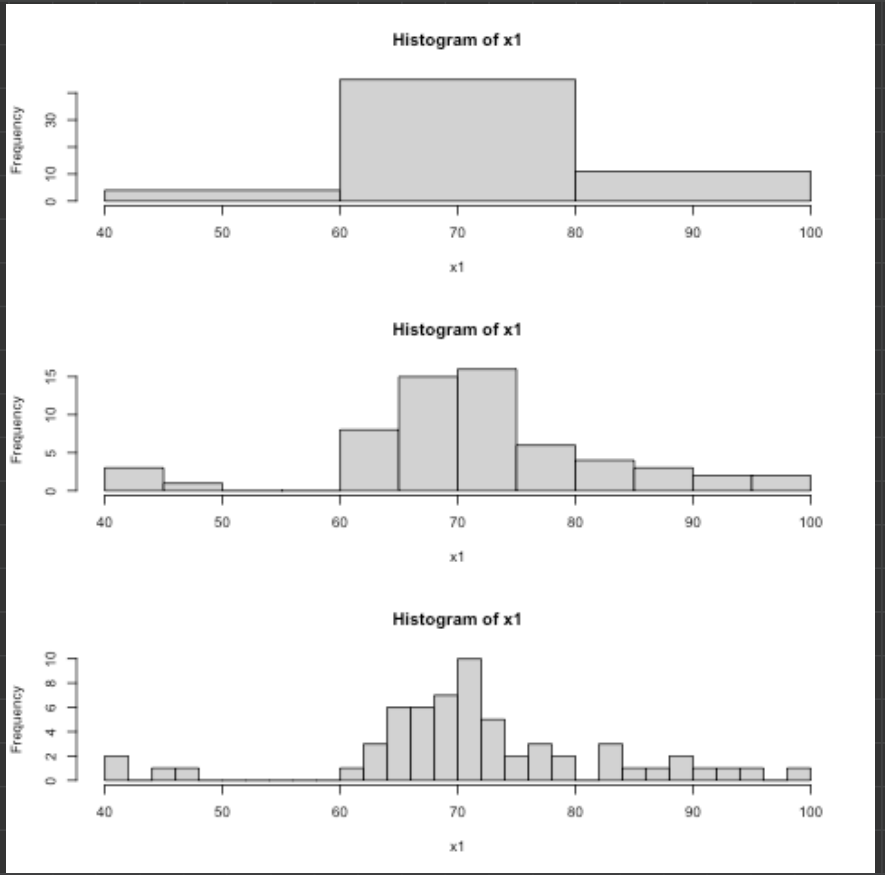


Figure 2.2
Different ways of showing the pattern of 915 guesses of the number of jelly beans in the jar. (a) A strip-chart or dot-diagram, with a jitter to prevent points lying on top of each other; (b) a box-and-whisker plot; (c) a histogram



GRADES DATASET (ATTENDING GROUP)

BIVARIATE DATA

(x_i, y_i) $\quad i = 1, \dots, n$

x, y NUMERICAL

A.S

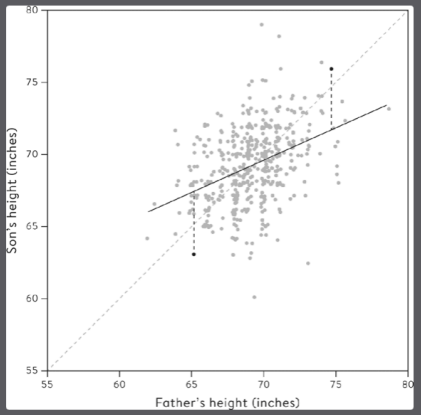


Figure 5.1
Scatter of heights of 465 fathers and sons from Galton's data (many fathers are repeated since they have multiple sons). A jitter has been added to separate the points, and the diagonal dashed line represents exact equality between son and father's heights. The solid line is the standard 'best-fit' line. Each point gives rise to a 'residual' (dashed line), which is the size of the error were we to use the line to predict a son's height from his father's.

SCATTER PLOT

IGNORE SOLID BLACK LINE

"REGRESSION TO THE MEAN"

CORRELATION COEFFICIENT

MEASURE OF HOW 'TIGHTLY' THE SCATTER PLOT IS AROUND A STRAIGHT LINE

$(x_i, y_i) \qquad i = 1, \ldots, n$

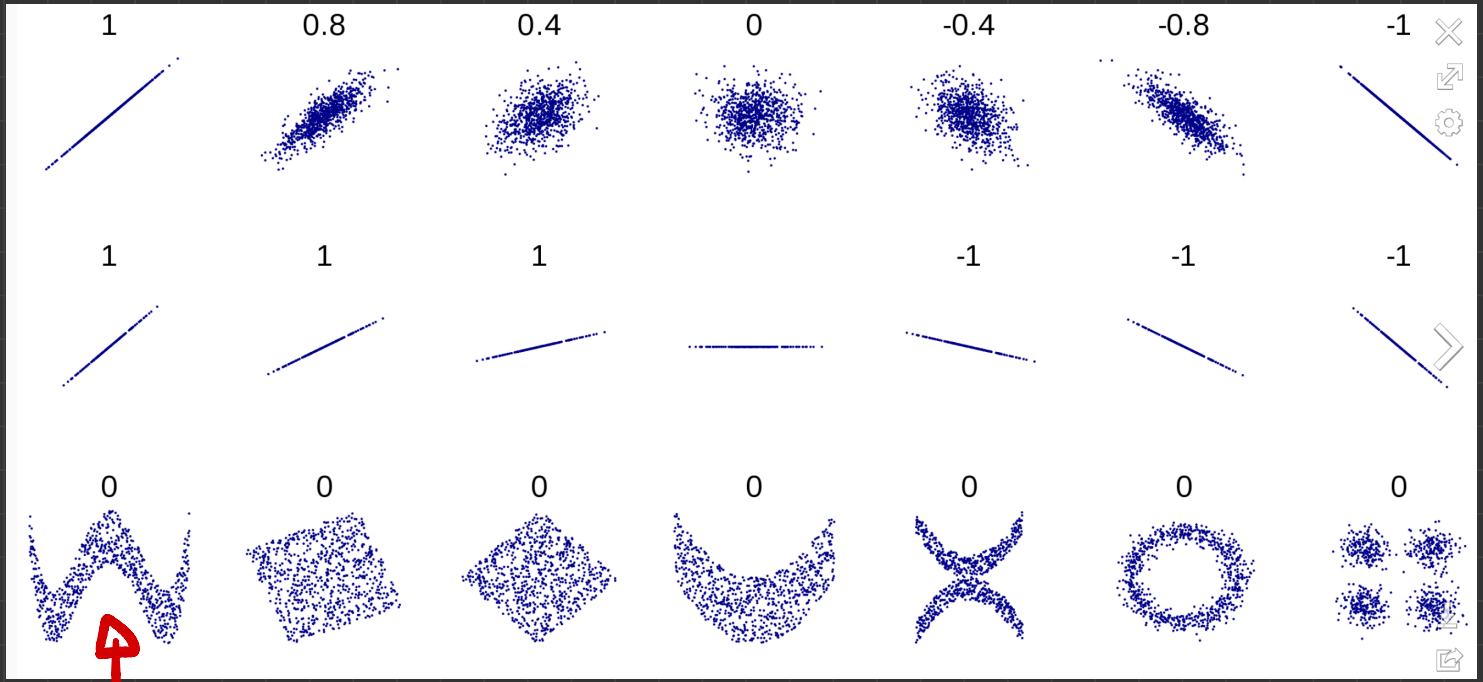
$$\text{mean}(x) = \frac{1}{n} \sum_i x_i \qquad \text{mean}(y)$$

$$\text{sd}(x) = \frac{1}{n} \sum_i (x_i - \text{mean}(x))^2 \qquad \text{sd}(y)$$

$$corr(x,y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - mean(x)}{sd(x)} \times \frac{y_i - mean(y)}{sd(y)} \right)$$

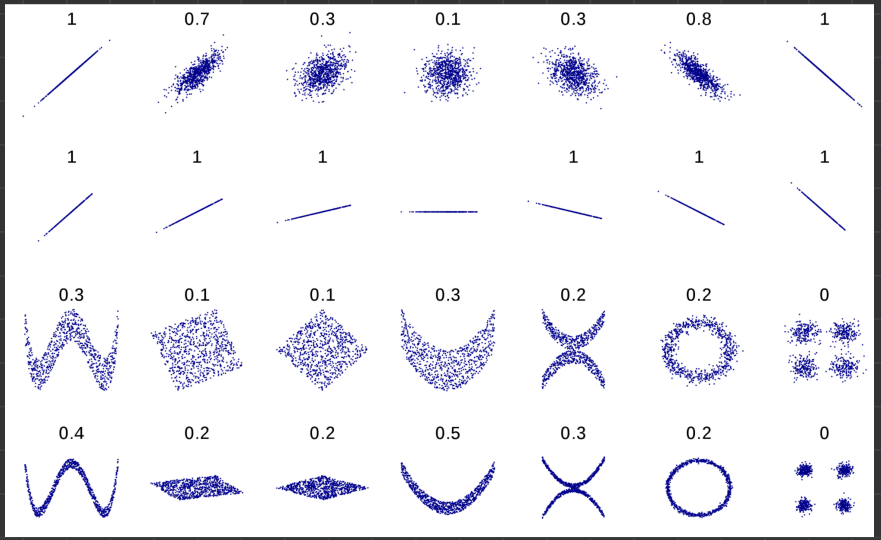
- $\frac{x_i - mean(x)}{sd(x)}$: Z-SCORE OR THE STANDARDIZATION OF DATA POINTS

- DEFINITION IMPLIES : $-1 \leq corr(x,y) \leq 1$
  - $corr(x,y)$ IS A STANDARDIZED MEASURE OF DEPENDENCE


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0
Y
X

DIFFERENT MEASURES OF (NON-LINEAR) DEPENDENCE

DISTANCE CORRELATION



AoS

Corr = 0