Statistics and probability

Topic 4 · SL

The biggest poll that failed

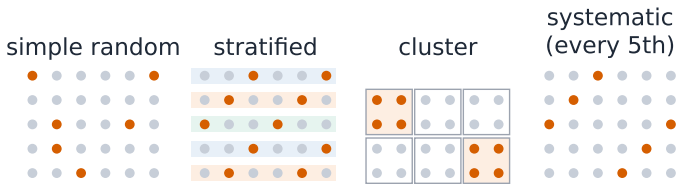
> A huge biased **sample** is worse than a small fair one.

- 2.4 million replies predicted the wrong winner.
- Gallup's 50 000 got it right.
- This unit: data $\rightarrow$ regression $\rightarrow$ probability $\rightarrow$ distributions.

Sampling techniques

Technique	Weakness
Simple random	Needs a full list
Systematic	Biased by patterns in the list
Stratified	Needs group sizes
Quota	Interviewer chooses
Convenience	Rarely representative

Random sampling designs



SRS: equal chance · stratified: sample within groups · cluster: take whole clusters · systematic: every kth

> Stratified sampling keeps the groups in proportion

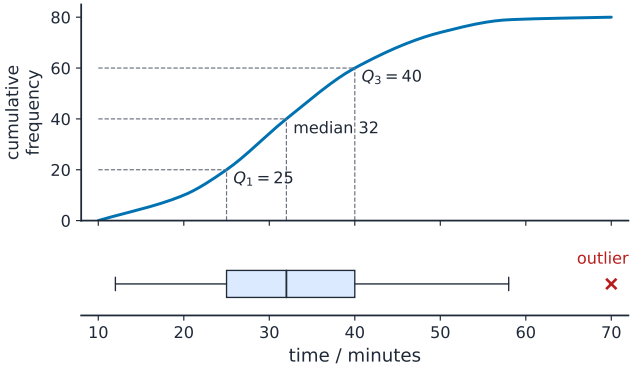
Centre and spread

1. **Centre.** **Mean** 9, **median** 7.5, **mode** 7
2. **Quartiles.** $Q_1 = 6$, $Q_3 = 11$, IQR 5
3. **Spread.** $\sigma^2 = 24$, $\sigma \approx 4.90$

Decision

The outlier pulls the mean above the median

Quartiles from a curve



- Read across at $\frac{n}{4}$, $\frac{n}{2}$, $\frac{3n}{4}$; compare centre and spread in context.

Mean, median, mode, range



$$\text{median} = 3$$

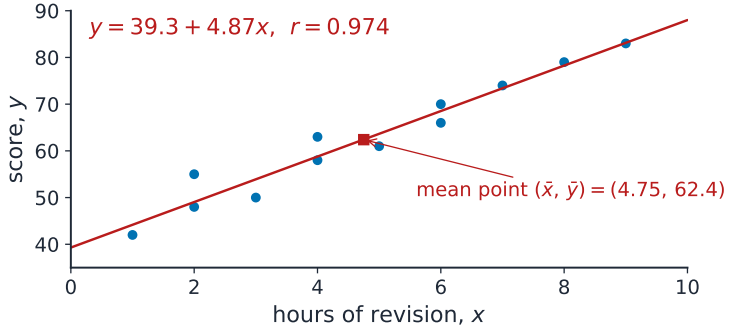
$$\text{mean} = \frac{2+3+3+5+7}{5} = 4$$

$$\text{mode} = 3 \text{ (most common)}$$

$$\text{range} = 7 - 2 = 5$$

- Data 2, 3, 3, 5, 7: **mean** 4, **median** 3, **mode** 3.
- Range $7 - 2 = 5$.

Regression



- Through the **mean** point $(\overline{x}, \overline{y})$.
- Correlation is not causation.

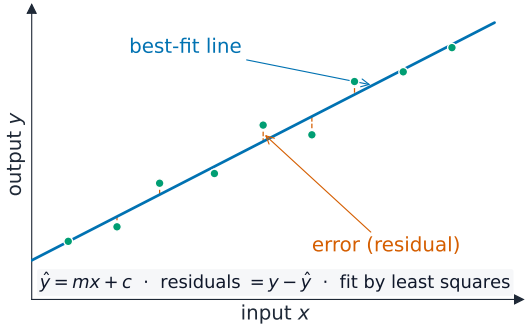
Using the line

1. **5.5 hours.** About 66 marks, **interpolation**
2. **Gradient.** About 4.9 marks per hour
3. **15 hours.** 112 marks, impossible **extrapolation**

Decision

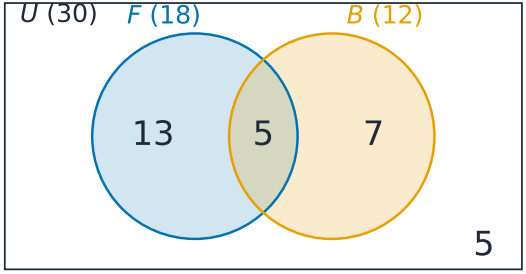
Predict only within the data

Least squares



- The line minimises the squared vertical errors.
- Predict y from x, inside the data range.

Venn diagrams



$P(F \cup B) = \frac{25}{30}$

$P(F \mid B) = \frac{5}{12}$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- Fill from the overlap outwards.

Conditional and independent

1. **Given** B. $P(F \mid B) = \frac{5}{12}$
2. **Product.** $P(F)P(B) = 0.24$
3. **Compare.** $P(F \cap B) \approx 0.167 \neq 0.24$

Decision

Not **independent**

Without replacement

1. **Both red.** $\frac{3}{5} \times \frac{2}{4} = 0.3$
2. **Same colour.** $0.3 + 0.1 = 0.4$
3. **Second red.** $0.3 + 0.3 = 0.6$

Decision

Second-draw probabilities change

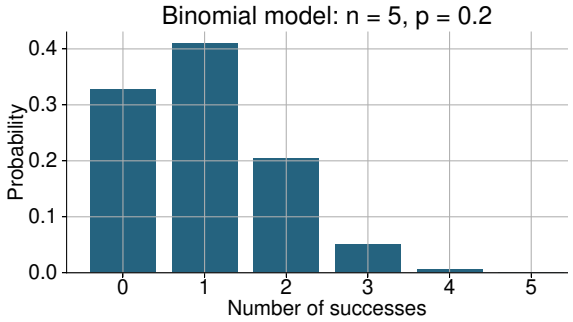
Expected value

1. **Expectation.** $E(X) = 0.6 + 2 = 2.6$
2. **Gain.** $2.6 - 3 = -0.4$ yuan
3. **Fair?** No, expected gain is not 0

Decision

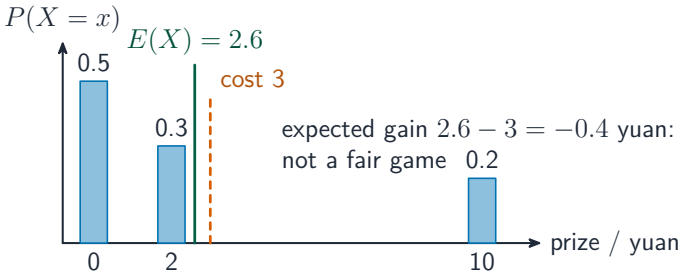
$E(X) = \sum x P(X = x)$

The binomial distribution



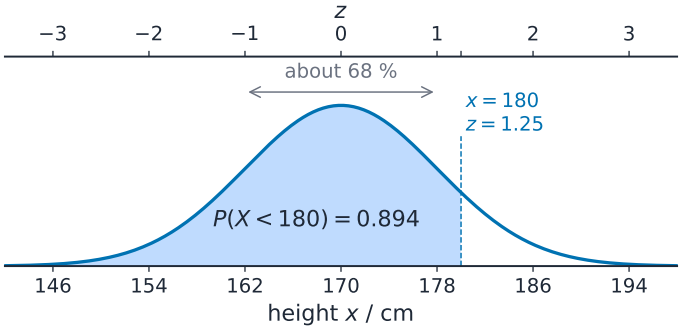
- Fixed n, two outcomes, **independent**, constant p.
- $E(X) = np$, $\text{Var}(X) = np(1-p)$.

Is the game fair?



- $E(\text{prize}) = 0(0.5) + 2(0.3) + 10(0.2) = 2.6.$
- Cost 3, so the expected gain is -0.4.

The normal distribution



- $z = \frac{x-\mu}{\sigma}$.
- 68 % within 1σ, 95 % within 2σ.

An unknown mean

1. **z.** $z = 1.645$
2. **Equation.** $\frac{50-\mu}{4} = 1.645$
3. **Solve.** $\mu \approx 43.4$

Decision

Standardise, then solve

Check: which distribution?

Your turn

The number of faulty bulbs in a box of 12, each faulty with probability 0.05, follows which model?

A. $B(12, 0.05)$
B. $N(12, 0.05^2)$
C. $B(0.05, 12)$
D. A uniform distribution

Exam tips

> Compare centre and spread, in context, every time.

- State the distribution and parameters.
- $P(X \geq r) = 1 - P(X \leq r - 1)$.
- Sketch the normal curve first.

Check yourself

Try these

For $4, 7, 7, 9, 12, 15, 18, 40$, find the quartiles and test 40 as an outlier.

$P(A) = 0.5$, $P(B) = 0.4$, $P(A \cup B) = 0.7$. Are A and B independent?

$X \sim N(\mu, 4^2)$ and $P(X < 30) = 0.2$. Find μ.