Statistics and probability

Topic 4 · SL and HL

The lady tasting tea

> Could the result have happened by chance alone?

- Eight cups, four of each kind, in random order.
- Guessing all four right has probability $\frac{1}{70}$.
- This unit: data $\to$ probability $\to$ distributions $\to$ tests.

Sampling techniques

Technique	Risk
Simple random	needs a full list
Systematic	biased if the list has a pattern
Stratified	needs group sizes
Quota, convenience	not random, so bias

Guessing the tea cups

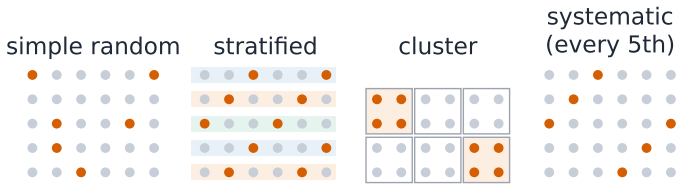


M: milk first, T: tea first; 4 of each in random order

choosing the right 4 by guessing: $1/\binom{8}{4} = 1/70 \approx 0.014$

So correct guesses are evidence of skill

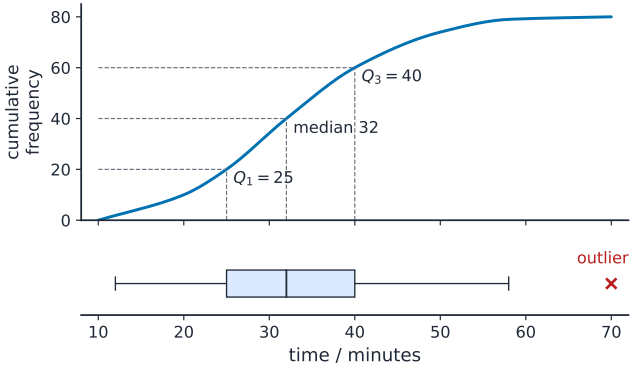
Four sampling designs



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

Choose the one that fits the **population**

Box plots



- **Outlier** beyond $Q_3 + 1.5 \times \text{IQR}$; compare a centre and a spread, in context.

Mean, median and mode



$$\text{median} = 3$$

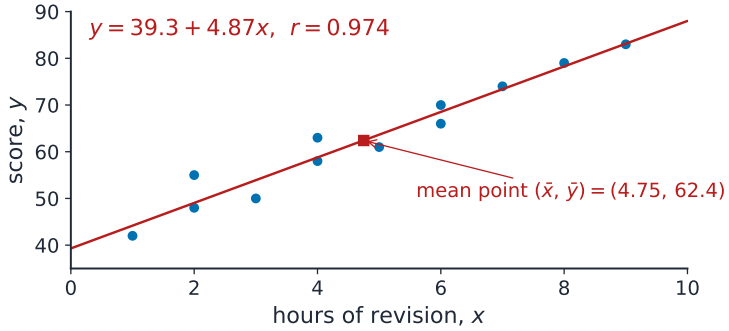
$$\text{mean} = \frac{2+3+3+5+7}{5} = 4$$

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

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

Choose the average that suits the data

Regression



- $r = 0.974$: strong positive linear correlation.
- Interpolate, do not extrapolate.

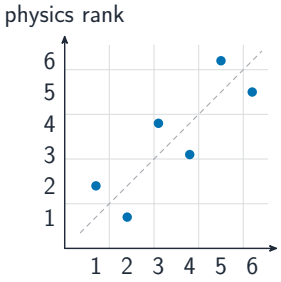
Spearman's rank

1. **Ranks.** maths $1, \ldots, 6$; physics $2, 1, 4, 3, 6, 5$
2. **GDC.** $r_s \approx 0.829$
3. **Interpret.** strong agreement between the rankings

Decision

Use it for ranks, outliers or curves that keep rising

Spearman's rank correlation

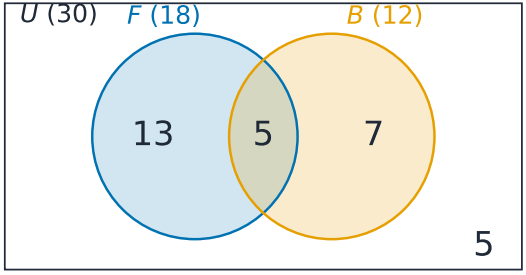


$r_s =$ Pearson's r on ranks

$r_s \approx 0.829$: strong positive agreement

A strong positive agreement in rank

Venn diagrams



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

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

- $P(F \mid B) = \frac{5}{12}$.
- $0.6 \times 0.4 \neq \frac{5}{30}$, so not **independent**.

Check: which formula?

Your turn

Which shows that A and B are independent?

A. $P(A)P(B) = 0.2 = P(A \cap B)$
B. $P(A) + P(B) < 1$
C. $P(A \cap B) \neq 0$
D. $P(A) > P(B)$

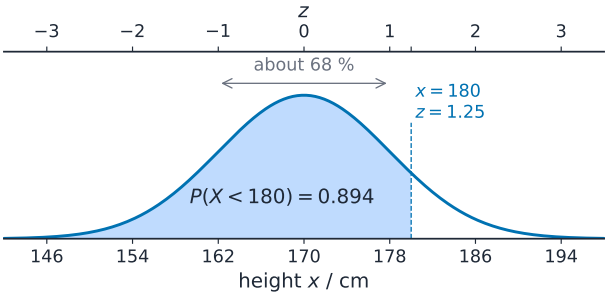
The binomial distribution

1. **Model.** $X \sim B(5, 0.2)$
2. **One.** $P(X = 1) = 0.4096$
3. **At least one.** $1 - 0.8^5 \approx 0.672$

Decision

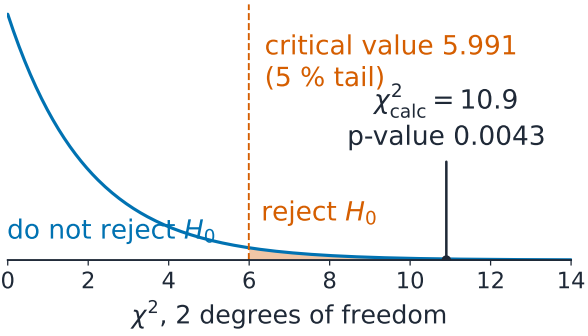
pdf for $=$, cdf for $\leq$

The normal distribution



- $z = \frac{x-\mu}{\sigma}$.
- Inverse normal finds a value from a probability.

Chi-squared for independence



- $E = \frac{\text{row} \times \text{column}}{\text{total}}$, all at least 5.
- **Degrees of freedom** $(r-1)(c-1)$.

The t-test

1. **Hypotheses.** $H_0 : \mu_A = \mu_B$, $H_1 : \mu_A < \mu_B$
2. **GDC.** $t \approx -3.29$, p-value 0.0041
3. **Conclude.** $0.0041 < 0.05$, reject H_0

Decision

State the conclusion in context

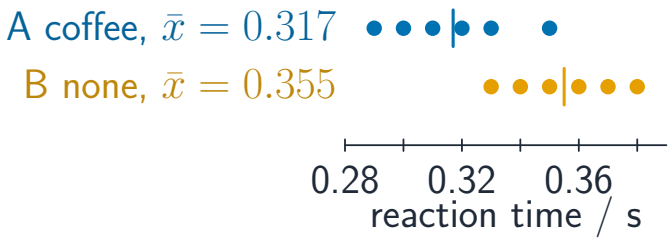
Check: what does p mean?

Your turn

What is the correct conclusion?

A. Not enough evidence that the die is biased
B. The die is proved fair
C. The die is biased
D. 63.9 % of rolls are fair

A two-sample t-test



pooled $t \approx -3.29$

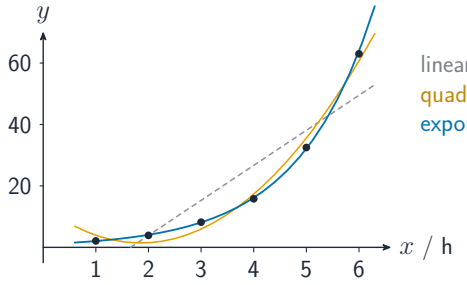
$p \approx 0.0041 < 0.05$

Below 0.05: reject H_0

Choosing a non-linear model

Model	R^2
linear	0.826
quadratic	0.989
exponential 1.03×1.99^x	0.9998

Comparing regression models



linear: $R^2 = 0.826$
quadratic: $R^2 = 0.989$
exponential: $R^2 = 0.9998$

Choose by fit and by context

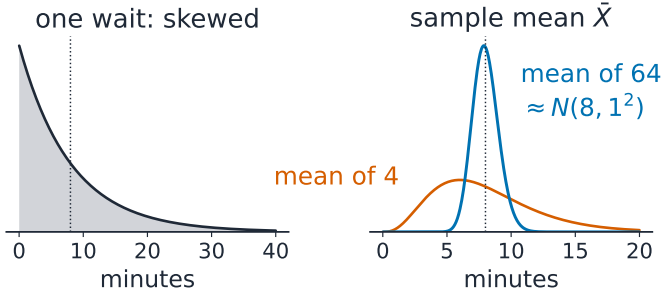
Combinations of variables

1. **Total.** $C + J \sim N(750, 10^2)$
2. **Probability.** $P(C + J > 765) \approx 0.0668$
3. **Estimate.** $s_{n-1}^2 = \frac{n}{n-1} s_n^2$

Decision

Square the coefficients, then add

The central limit theorem



- $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ for large n.
- $n = 64$ waits gives $N(8, 1^2)$.

Adding independent normals

jam C
$N(500,\, 6^2)$

$+$

jar J
$N(250,\, 8^2)$

$=$

total T
$N(750,\, 10^2)$

variances add: $36 + 64 = 100$ (standard deviations do not: $6 + 8 \neq 10$)

$$P(T > 765) \approx 0.0668$$

> Variances add, not standard deviations

What 95\% confidence means



CI form: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ · level = long-run capture rate of the method

> Confidence is in the method, not one interval

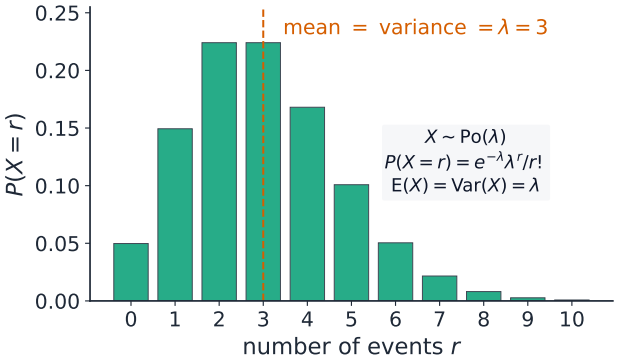
The Poisson distribution

1. **None.** $P(X = 0) = e^{-3} \approx 0.0498$
2. **At most 2.** $P(X \leq 2) \approx 0.423$
3. **Two hours.** $Po(6)$, $P(Y \geq 10) \approx 0.0839$

Decision

Scale the **mean** to the interval

Reading a Poisson bar chart



Events at a constant average rate

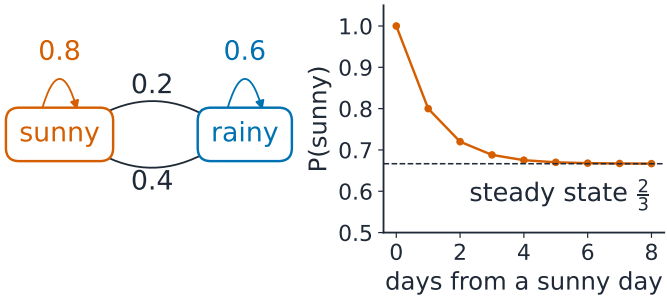
Critical regions and errors

1. **Boundary.** $P(X \geq 14) \approx 0.0577$, $P(X \geq 15) \approx 0.0207$
2. **Region.** $X \geq 15$; Type I 0.0207
3. **If** $p = 0.7$**.** Type II $P(X \leq 14) \approx 0.584$

Decision

List the probabilities either side

Markov chains



- Columns of T add to 1; $\mathbf{s}_n = T^n \mathbf{s}_0$.
- **Steady state** $T\mathbf{s} = \mathbf{s}$: $\left(\frac{2}{3}, \frac{1}{3}\right)$.

Exam tips

> A test never proves H_0; it finds evidence against it, or not.

- Write H_0 and H_1 in context.
- Compare the p-value with the level, then conclude in words.
- Expected frequencies at least 5.

Check yourself

Try these

Find $P(160 < X < 180)$ for heights $N(170, 8^2)$.

Find the expected frequencies for the drinks table.

Test whether a die with 8, 9, 12, 7, 10, 14 is fair.