

Basics & scales

Descriptive: summarise the data you have.

Inductive: sample → population (needs probability). Explore → Summarize → Model → Estimate → Test.

Population (defined by content, time, space) · **unit** (case/object/subject) · **attribute** (variable) · **value** · **raw data list**. **Parameter** μ, σ, π (population) vs **statistic** $\bar{x}, s, p$ (sample).

scale	meaning	allowed
nominal	categories, no order	mode
ordinal	natural order, distances meaningless	+ median, quantiles, Spearman
metric discrete	countable values	+ mean, SD, Pearson
metric continuous	any value, arbitrary precision	+ classes / histogram

Nested: metric $\supset$ ordinal $\supset$ nominal properties (never the other way).

Frequencies & charts

f_i absolute, $p_i = f_i / N$ relative, $F_i = \sum_{j \leq i} p_j$ cumulative. Discrete → **bar chart** (gaps). Continuous → classes $[e_{i-1}, e_i)$, width $d_i = e_i - e_{i-1}$, mark $x_i = (e_{i-1} + e_i) / 2$, $k \approx \sqrt{N}$ classes → **histogram** (no gaps).

Unequal widths: bar height = **density** p_i / d_i (area = share).

ECDF $F(x)$ = share $\leq x$: step function (raw/discrete), polygon (grouped).

$P(a < X \leq b) = F(b) - F(a)$, above x : $1 - F(x)$.

Location

$$\bar{x} = \frac{1}{N} \sum x_i \quad \bar{x} = \frac{1}{N} \sum_{i=1}^k f_i x_i$$

(raw · grouped with class marks) · weighted

$$\bar{x}_w = \sum w_i x_i / \sum w_i$$

Median (sort first): N odd $x_{((N+1)/2)}$; N even $\frac{1}{2}(x_{(N/2)} + x_{(N/2+1)})$. Robust; mean is efficient but outlier-sensitive.

Tukey's hinges (raw quartiles): split sorted data into halves — N odd: median belongs to **both** halves. Q_1 = median of lower half, Q_3 = median of upper half. e.g. 4 5 6 7 8 9 9 → $Q_1 = 5.5$, $Q_2 = 7$, $Q_3 = 8.5$

Grouped quantile (class with $F_{i-1} \leq Np \leq F_i$):

$$\tilde{x}_p = e_{i-1} + \frac{d_i}{Np - F_{i-1}}(Np - F_{i-1})$$

Mode: most frequent (nominal OK), may be multimodal; modal class = highest density.

Geometric (growth): $\bar{r} = \sqrt[n]{x_n / x_0} - 1$ ·

Harmonic (rates over equal amounts):

$$\bar{v} = \frac{n}{\sum 1/v_i} \rightarrow 60 \text{ \& } 100 \text{ km/h} \rightarrow 75.$$

Right-skewed: mode < median < mean · left-skewed: mean < median < mode. One value changes by Δ → mean changes by Δ/N . Mixture:

$$\bar{x} = w\bar{x}_1 + (1-w)\bar{x}_2.$$

Spread

Range = max - min · **IQR** = $Q_3 - Q_1$ (robust)

$$\sigma^2 = \frac{1}{N} \sum (x_i - \bar{x})^2 = \frac{1}{N} \sum x_i^2 - \bar{x}^2$$

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

Grouped: $\sigma^2 = \frac{1}{N} \sum f_i x_i^2 - \bar{x}^2$. $\sigma = \sqrt{\sigma^2}$.

Describe a population → $\div N$; **estimate** from a sample → $\div (n-1)$ (unbiased; R/Excel VAR.S).

Transform $y = a + bx$: $\bar{y} = a + b\bar{x}$,

$$\sigma_y^2 = b^2 \sigma_x^2, \sigma_y = |b| \sigma_x \text{ (shift-invariant).}$$

$CV = \sigma / \bar{x}$ (unit-free) · skewness

$$\gamma = \frac{1}{N} \sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^3 \text{ (<0 left, >0 right) · excess}$$

$$\text{kurtosis } w = \frac{1}{N} \sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 - 3$$

Sigma rule (normal): $\mu \pm 1\sigma$ 68.3% · $\pm 2\sigma$ 95.4% · $\pm 3\sigma$ 99.7%.

Boxplot

Five numbers: min, Q_1 , median, Q_3 , max.

Modified: fences $Q_1 - 1.5 IQR$,

$Q_3 + 1.5 IQR$; whiskers end at the last values **inside**; beyond = outlier O; beyond $3 \cdot IQR$ = extreme *.

Ages 31 35 42 45 47 49 50 51 78 90: Q_1 42, $\bar{x}$ 48, Q_3 51, IQR 9, upper fence 64.5 → 78 O, 90 * (> 78).

Two variables

$$\text{Cov} = \frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N} \sum x_i y_i - \bar{x}\bar{y}$$

$$r = \frac{\text{Cov}(x, y)}{\text{SD}(x) \text{SD}(y)}$$

Cov sign = direction only. **Pearson r** $\in [-1, 1]$: strength of a **linear** link; unit-free (km/h → mph: same r); $r = 0 \neq$ "no relation"; outlier-sensitive.

Spearman ρ_S : Pearson on **ranks** (ties → mean rank, e.g. places 2,3 → 2.5) — ordinal data / outliers, measures **monotonic** links.

Correlation $\neq$ causation: $X \rightarrow Y$, $Y \rightarrow X$, confounder C, or chance.

Exam traps

- Raw-data quartiles → **Tukey**; grouped → **interpolate**; mean of grouped data uses **class marks**.
- Whole population given → $\div N$; "estimate", "sample" → $\div (n-1)$.
- Unequal class widths → plot **density**, not counts.
- Speed over equal distances → harmonic; growth → geometric; prices $\times$ quantities → weighted.
- Change of units: Var $\times b^2$, SD $\times |b|$, Cov $\times b_x b_y$, **r unchanged**.
- Ordinal → median / Spearman; nominal → mode only.
- Outliers / skew → report median & IQR (boxplot).
- Regression from summaries: $b = r s_y / s_x$; predict only inside the data range.

Linear regression

$$\hat{y} = a + bx, \quad a = \bar{y} - b\bar{x}$$

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = r \frac{\text{SD}(y)}{\text{SD}(x)}$$

Least squares: minimise $\sum e_i^2$, $e_i = y_i - \hat{y}_i$; line passes through $(\bar{x}, \bar{y})$. b = change of y per unit x ; $a = y$ at $x = 0$.

$SQT = SQE + SQR$ with $\sum (y_i - \bar{y})^2$, $\sum (\hat{y}_i - \bar{y})^2$, $\sum (y_i - \hat{y}_i)^2$; $R^2 = SQE / SQT = r^2 \in [0, 1]$; acceptable if > 0.6 . Beware extrapolation.

Summary data: mean 68/75, Var 81/100, Cov 40.5 → $b = 0.5$, $a = 41$, $\hat{y}(51) = 66.5$.

Probability basics

Experiment, sample space Ω , event $A \subseteq \Omega$. **Laplace:** $P(A) = |A| / |\Omega|$ (equally likely). **Empirical:** relative frequency; **law of large numbers:** it converges to $P(A)$.

Kolmogorov: $0 \leq P(A) \leq 1$, $P(\Omega) = 1$, disjoint: $P(A \cup B) = P(A) + P(B)$. $\Rightarrow P(\bar{A}) = 1 - P(A)$. Frequentist (long-run frequency) vs Bayesian (degree of belief).

Counting

choose k of n	no repetition	repetition
order matters	$\frac{n!}{(n-k)!}$	n^k
order irrelevant	$\binom{n}{k}$	$\binom{n+k-1}{k}$

All n arranged: $n!$ · identical groups:

$\frac{n!}{k_1! k_2! \dots}$ (ASSESSMENT $\frac{10!}{4! 2!} = 75\,600$) · stages multiply ($4 \cdot 7 \cdot 5 = 140$).

Tricks: "at least one" = all - none · fixed members: choose the rest · never together: all - both in · only together: both in + both out · sit together: glue to one block ($\cdot \text{block!}$) · $\binom{n}{k} = \binom{n}{n-k}$.

Rules

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\begin{aligned} 3 \text{ events: } P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(AB) - P(AC) - P(BC) + P(ABC) \end{aligned}$$

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

$$\text{Multiplication: } P(A \cap B) = P(A | B)P(B)$$

$$\begin{aligned} \text{Independent} &\Leftrightarrow P(A \cap B) = P(A)P(B) \\ &\Leftrightarrow P(A | B) = P(A). \text{ Disjoint } (P(A \cap B) = 0) \text{ events are } \textit{dependent} \text{ (if } P > 0). \end{aligned}$$

Total probability (partition $A_1 \dots A_k$):

$$\begin{aligned} P(B) &= \sum P(B | A_i)P(A_i) \\ P(A_j | B) &= \frac{P(B | A_j)P(A_j)}{\sum_i P(B | A_i)P(A_i)} \end{aligned}$$

Test: sens = $P(+|ill)$, spec = $P(-|healthy)$, PPV = $P(ill|+)$. Think 1000 people: 3% ill, sens 96%, spec 93% $\rightarrow$ 28.8 TP, 67.9 FP $\rightarrow$ PPV 29.8%. Tree: multiply along, add across.

Random variables

Discrete: $f(x) = P(X = x)$, $F(x) = P(X \leq x)$ (steps). **Continuous:** density $f \geq 0$, $\int f = 1$, $P(a \leq X \leq b) = \int_a^b f = F(b) - F(a)$, $P(X = x) = 0$, $F' = f$.

$$E(X) = \sum x f(x) \text{ or } \int x f(x) dx$$

$$\text{Var}(X) = E(X^2) - E(X)^2$$

$$E(aX + b) = aE(X) + b,$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X),$$

$$E(X + Y) = EX + EY; \text{ independent:}$$

$$\text{Var}(X \pm Y) = \text{Var}X + \text{Var}Y.$$

Joint $f(x, y)$; marginals = row/column

$$\text{sums; } f(y | x) = f(x, y) / f_X(x);$$

$$\text{independent} \Leftrightarrow f(x, y) = f_X(x)f_Y(y);$$

$$\text{Cov} = E(XY) - E(X)E(Y), \rho = \frac{\text{Cov}}{\text{SD}_X \text{SD}_Y}.$$

Worked examples

- Bayes** — fire alarm: $P(F) = 0.0002$, $P(A|F) = 0.99$, $P(A|\bar{F}) = 4/365 \rightarrow P(F | A) = \frac{0.99 \cdot 0.0002}{0.99 \cdot 0.0002 + \frac{4}{365} \cdot 0.9998} = 0.0178$
- Binomial** — 10 free throws, $p = 0.8$: $P(X > 2) = 1 - P(0) - P(1) - P(2) = 0.99992$
- Poisson** — 1 customer / 2 min, window 5 min $\rightarrow \lambda = 2.5$: $P(X = 2) = 0.2565$
- Hypergeom.** — 4 "Johns" in 20, choose 6: $P(\text{exactly 1}) = \frac{\binom{4}{1} \binom{16}{5}}{\binom{20}{6}} = 0.4508$; none 0.2066
- Normal** — $N(50, 4)$: $P(X < 51) = \Phi(0.5) = 0.6915$; $P(48 < X < 52) = 2\Phi(1) - 1 = 0.6827$
- Counting** — 12 candidates, 4 chosen: 495; two refuse each other: $495 - \binom{10}{2} = 450$; only together: $45 + 210 = 255$
- Monty Hall** — switching wins with 2/3, staying with 1/3.

Distributions

	model	$P(X = k) / f(x)$	E	Var
Uniform {1..n}	equally likely	$1 / n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Bernoulli	one trial	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial	n indep. trials, successes	$\binom{n}{k} p^k(1-p)^{n-k}$	np	$np(1-p)$
Hypergeom.	n of N without replacement, M successes	$\frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$	$n \frac{M}{N}$	$n \frac{M}{N} (1 - \frac{M}{N}) \frac{N-n}{N-1}$
Poisson	events per interval, rate λ	$\frac{\lambda^k}{k!} e^{-\lambda}$	λ	λ
Geometric	trials until 1st success	$(1-p)^{k-1} p$	$1 / p$	$\frac{1-p}{p^2}$
Neg. binomial	failures before r-th success	$\binom{k+r-1}{k} p^r (1-p)^k$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Normal	$N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$\chi^2(\text{df})$	sum of df squared $N(0,1)$	right-skewed, ≥ 0	df	2df
t(df)	$N(0,1)$ -like, heavy tails	$\rightarrow N(0,1)$ as $\text{df} \rightarrow \infty$	0	$\frac{df}{df-2}$
F(df₁, df₂)	ratio of variances	≥ 0	$\frac{df_1}{df_2-2}$	

Approximations: Hyp $\approx$ Bin(n, M/N) if $n / N \leq 0.05$ · Bin $\approx$ Poiss(np) if $n \geq 50$, $p \leq 0.05$ · Bin $\approx$ N(np, np(1-p)) if $np(1-p) \geq 9$, continuity correction: $P(X \leq k) \approx \Phi(\frac{k+0.5-np}{\sqrt{np(1-p)}})$, lower limit -0.5.

Which one? fixed n, constant $p \rightarrow$ Bin · small urn, no replacement $\rightarrow$ Hyp · counts per time/area $\rightarrow$ Poisson (scale λ to the window!) · waiting for 1st $\rightarrow$ Geo.

Normal & CLT

$$Z = \frac{X - \mu}{\sigma}, \quad P(X \leq x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad x_p = \mu + \sigma u_p$$

$\Phi(-z) = 1 - \Phi(z)$; $P(X = x) = 0$. Sums (indep.):
 $\sum X_i \sim N(n\mu, n\sigma^2)$; $X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$ — variances add.

CLT: $\bar{X} \approx N(\mu, \sigma^2 / n)$ for any distribution, $n > 30$. Standard error $\sigma / \sqrt{n}$.

Estimation & CIs

$\hat{\mu} = \bar{x}, \hat{\sigma}^2 = s^2 (\div n-1), \hat{\pi} = k / n$. Unbiased $E(\hat{\theta}) = \theta$ · consistent (Var $\rightarrow 0$) · efficient (smallest Var).

σ known	$\bar{x} \pm u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$
σ unknown	$\bar{x} \pm t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}$
proportion	$p \pm u_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$
one-sided	$\mu \leq \bar{x} + t_{n-1, 1-\alpha} \frac{s}{\sqrt{n}}$
sample size	$n \geq \left(\frac{2u_{1-\alpha/2}\sigma}{L}\right)^2$ (round up)
level for L	$u = \frac{L\sqrt{n}}{2\sigma}, S = 2\Phi(u) - 1$

"95% of such intervals cover μ " — not "95% of values lie inside". More confidence $\rightarrow$ wider; $4 \times n \rightarrow$ half width.

Testing logic

Claim to prove $\rightarrow H_1$; H_0 = opposite (with "="). Choose α (5%) **before**. Compute T, compare with critical value, or $p < \alpha \Rightarrow$ reject. Not rejected $\neq H_0$ proven.

	H₀ true	H₁ true
reject H₀	type I α (false +)	power $1 - \beta$
keep H₀	✓	type II β (false -)

p-value: P(result at least as extreme | H_0). One-sided (only if the other direction is impossible): $t_{1-\alpha}$, p halves. Test $\Leftrightarrow$ CI: reject iff $\mu_0 \notin$ CI. Significant $\neq$ relevant (check CI). **Multiple tests:** $\alpha_{glob} = 1 - (1 - \alpha)^k$; Bonferroni α / k .

$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$ · CI for $\mu_x - \mu_y$:
 $(\bar{x} - \bar{y}) \pm t_{n+m-2, 1-\alpha/2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$

χ^2 assumptions: $n > 30$, all $e \geq 1$, $\geq 80\%$ of $e \geq 5$ (else Fisher). Cramér $V = \sqrt{\frac{\chi^2}{n(l-1)}}$, $l = \min(\text{rows}, \text{cols})$: ≤ 1 none · ≤ 3 low · ≤ 5 moderate · > 5 high.

Which test?

test	use when	statistic	reject H ₀ if
Gauss (z)	μ vs μ_0 , σ known	$\frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$	$ T > u_{1-\alpha/2}$
one-sample t	μ vs μ_0 , σ unknown	$\frac{\bar{x} - \mu_0}{s} \sqrt{n}$	$ T > t_{n-1, 1-\alpha/2}$
pooled t	2 indep. groups, $\sigma_x^2 = \sigma_y^2$	$\frac{\bar{x} - \bar{y}}{s_p} \sqrt{\frac{nm}{n+m}}$	$ T > t_{n+m-2, 1-\alpha/2}$
Welch t	2 indep. groups, variances differ	$\frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}}$	t with Welch df (round down)
paired t	same objects twice	$\frac{\bar{d}}{s_d} \sqrt{n}, \quad d_i = x_{i1} - x_{i2}$	$ T > t_{n-1, 1-\alpha/2}$
F-test	equal variances?	$F = s_x^2 / s_y^2$ (larger on top)	$F > F_{n_x-1, n_y-1, 1-\alpha}$
binomial	π vs π_0 , $np(1-p) \geq 9$	$\frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$	$ T > u_{1-\alpha/2}$
χ^2 fit	1 variable vs p_i , $e_i = np_i$	$\sum \frac{(o_i - e_i)^2}{e_i}$	$> \chi^2_{k-1, 1-\alpha}$
χ^2 indep.	m×k table, $e_{ij} = \frac{r_i c_j}{n}$	$\sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$	$> \chi^2_{(m-1)(k-1), 1-\alpha}$
Fisher	small 2×2, exact	sum P of tables $\leq$ P(obs), same margins	$p < \alpha$
McNemar	paired yes/no; b, c changers	$\frac{(b-c)^2}{b+c}; b+c \leq 30$ $\frac{(b-c -1)^2}{b+c}$	> 3.841

Tables

$\Phi(z) = P(Z \leq z)$

z	+0.00	+0.05	+0.10	+0.15
0.0	0.5000	0.5199	0.5398	0.5596
0.2	0.5793	0.5987	0.6179	0.6368
0.4	0.6554	0.6736	0.6915	0.7088
0.6	0.7257	0.7422	0.7580	0.7734
0.8	0.7881	0.8023	0.8159	0.8289
1.0	0.8413	0.8531	0.8643	0.8749
1.2	0.8849	0.8944	0.9032	0.9115
1.4	0.9192	0.9265	0.9332	0.9394
1.6	0.9452	0.9505	0.9554	0.9599
1.8	0.9641	0.9678	0.9713	0.9744
2.0	0.9772	0.9798	0.9821	0.9842
2.2	0.9861	0.9878	0.9893	0.9906
2.4	0.9918	0.9929	0.9938	0.9946
2.6	0.9953	0.9960	0.9965	0.9970
2.8	0.9974	0.9978	0.9981	0.9984
3.0	0.9987	0.9989	0.9990	0.9992

t- and χ^2 -quantiles

	df	t _{0.95}	t _{0.975}	t _{0.995}	$\chi^2_{0.95}$
1	6.314	12.706	63.657	3.84	
2	2.920	4.303	9.925	5.99	
3	2.353	3.182	5.841	7.81	
4	2.132	2.776	4.604	9.49	
5	2.015	2.571	4.032	11.07	
6	1.943	2.447	3.707	12.59	
7	1.895	2.365	3.499	14.07	
8	1.860	2.306	3.355	15.51	
9	1.833	2.262	3.250	16.92	
10	1.812	2.228	3.169	18.31	
11	1.796	2.201	3.106	19.68	
12	1.782	2.179	3.055	21.03	
13	1.771	2.160	3.012	22.36	
14	1.761	2.145	2.977	23.68	
15	1.753	2.131	2.947	25.00	
16	1.746	2.120	2.921	26.30	
17	1.740	2.110	2.898	27.59	
18	1.734	2.101	2.878	28.87	
19	1.729	2.093	2.861	30.14	
20	1.725	2.086	2.845	31.41	
22	1.717	2.076	2.819	33.92	
25	1.708	2.060	2.787	37.65	
30	1.697	2.042	2.750	43.77	
40	1.684	2.021	2.704	55.8	
60	1.671	2.000	2.660	79.1	
120	1.658	1.980	2.617	146.6	
∞	1.645	1.960	2.576		

F_{d1, d2, 0.95}

d ₂ \ d ₁	1	2	3	4	5	6	8	10	12	15
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.68	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	4.00	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.57	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.28	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	3.07	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.91	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.69	2.62
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.54	2.48	2.40
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.28	2.20
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	2.09	2.01

Statistics & data

- **Statistics** = science of collecting, organising, presenting and interpreting data ("learning from data").
- **Descriptive**: summarise observed data, no generalisation. **Inductive / inferential**: draw conclusions from a sample about a population; rests on **probability theory**.
- **Empirical population**: all objects of interest, defined by *content*, *time* and *space* — never arbitrary. **Sample**: subset actually observed.
- **Representative sample** mirrors the population; **simple random sample**: every object has the same chance. **Random error** can be controlled by inductive statistics, **systematic error (bias)** cannot.
- **Observational unit** (case, object, subject) · **attribute** (variable, characteristic) · **attribute value** · **raw data list** (uncompressed values).
- **Parameter** describes a population (μ , σ , π); **statistic** is computed from a sample ($\bar{x}$, s , p) and estimates a parameter.

Levels of measurement

- **Nominal**: categories, only = / $\neq$ (blood type). **Ordinal**: natural order, distances not interpretable (grades, Likert, pain). **Metric**: differences interpretable.
- **Metric discrete**: countable values (children, visits). **Metric continuous**: any value in an interval, arbitrary precision (time, length, volume).
- Hierarchy: metric $\Rightarrow$ has ordinal $\Rightarrow$ has nominal properties; **never** the reverse. "Categorical" = nominal or ordinal.
- Ordinal values need not be integers; numbers as codes (postal code) can still be nominal.
- Allowed: nominal $\rightarrow$ mode, frequencies, χ^2 ; ordinal $\rightarrow$ + median, quantiles, Spearman; metric $\rightarrow$ + mean, SD, Pearson, t-tests.

Describing data — concepts

- Bar chart (gaps) for discrete/categorical; **histogram** (no gaps) for continuous classes; area $\propto$ share, so unequal widths need density.
- ECDF: non-decreasing, from 0 to 1, step function for raw data, polygon for grouped data.
- **Mean**: uses all information (efficient) but sensitive to outliers; only metric. **Median**: robust, ordinal OK, less efficient. **Mode**: any scale, may be multimodal.
- Similar mean and median $\Rightarrow$ roughly symmetric; large gap $\Rightarrow$ skewed $\rightarrow$ report median. Right skew: mode < median < mean.
- Deviations from the mean always sum to 0 $\Rightarrow$ only $n - 1$ are free (**degrees of freedom**) $\Rightarrow s^2$ divides by $n - 1$ and is **unbiased**.
- Variance: ≥ 0 , = 0 only for constant data, squared units, **location-invariant but not scale-invariant** ($\times b^2$). SD in original units. CV unit-free (compare different units).
- Range sensitive to outliers; IQR robust (middle 50%). Boxplot = five-number summary; outlier > 1.5·IQR, extreme > 3·IQR beyond the box. Skewness < 0 left, > 0 right; excess kurtosis compares tails with the normal.

Correlation & regression — concepts

- Covariance: sign = direction of a monotonic link; size not interpretable (unit-dependent).
- **Pearson r**: strength & direction of a **linear** link; $-1 \leq r \leq 1$; dimensionless; symmetric in X, Y; unchanged by linear unit changes (up to sign); outlier-sensitive; $r = 0 \Rightarrow$ no **linear** link (a curve may exist); undefined if a variance is 0.
- **Spearman p**: Pearson on ranks; for ordinal data/outliers; measures **monotonic** links.
- **Correlation $\neq$ causation**: $X \rightarrow Y$, $Y \rightarrow X$, confounder (spurious correlation, e.g. storks & births), or chance.
- Regression: $X =$ explanatory/independent/predictor, $Y =$ response/dependent. Least squares minimises the sum of squared **vertical** residuals; line passes through $(\bar{x}, \bar{y})$.
- **R²** = share of the variation of Y explained by the model (0...1), = r^2 only in simple linear regression; > 0.6 acceptable. Extrapolation outside the data is unreliable.

Probability — concepts

- Random experiment $\rightarrow$ sample space Ω (all outcomes); event = subset; elementary event = one outcome; Ω certain, $\emptyset$ impossible.
- **Laplace** needs finitely many *equally likely* outcomes. **Empirical** = relative frequency; **law of large numbers**: it converges to P as $n \rightarrow \infty$.
- **Frequentist**: probability = long-run frequency, parameters fixed. **Bayesian**: probability = degree of belief, updated by data (prior $\rightarrow$ posterior).
- Kolmogorov axioms: non-negativity, $P(\Omega) = 1$, additivity for disjoint events.
- **Disjoint $\neq$ independent**: disjoint events with $P > 0$ are dependent. Independent: knowing B does not change $P(A)$.
- $P(A \mid B) \neq P(B \mid A)$. Sensitivity = $P(+ \mid \text{ill})$, specificity = $P(- \mid \text{healthy})$, PPV = $P(\text{ill} \mid +)$ depends strongly on the **base rate** (prevalence).

Estimation — concepts

- **Point estimate**: very precise, little confidence. **Interval estimate**: less precise, chosen confidence $S = 1 - \alpha$. Precision and confidence compete.
- **Unbiased**: $E(\theta) = \theta$. **Consistent**: $\text{Var}(\theta) \rightarrow 0$ as n grows. **Efficient**: smallest variance among unbiased estimators.
- **SD** describes the spread of single values; **standard error** $\sigma/\sqrt{n}$ the spread of the mean.
- CI meaning: the *procedure* covers the true parameter in $(1 - \alpha)$ of repeated samples. **Not**: "95% of the values lie inside"; a single computed interval either contains μ or not.
- Wider CI with: higher confidence, larger σ , smaller n . Larger $n \Rightarrow$ more precision and/or confidence; t-interval wider than z-interval for small n .

Random variables & distributions — concepts

- A **random variable** maps outcomes to numbers. Discrete: pmf $P(X = x)$. Continuous: density f — **not** a probability, may exceed 1; $P(X = x) = 0$; probability = area.
- CDF F : non-decreasing, right-continuous, $F(-\infty) = 0$, $F(\infty) = 1$. Quantile x_p : $F(x_p) = p$.
- $E(X)$ = long-run average / balance point; $\text{Var}(X)$ = expected squared deviation. Independence $\Rightarrow \text{Cov} = 0$ (not conversely).
- **Binomial**: fixed n , independent trials, two outcomes, constant p . **Hypergeometric**: without replacement (p changes). **Poisson**: rare independent events per interval, $E = \text{Var} = \lambda$. **Geometric**: waiting time to 1st success.
- **Normal**: symmetric bell, mean = median = mode, determined by μ and σ , inflection points $\mu \pm \sigma$, 68–95–99.7 rule; sums of independent normals are normal.
- **t**: symmetric, heavier tails (accounts for estimating σ), $\rightarrow N(0,1)$ as $df \rightarrow \infty$. $\chi^2 \geq 0$, right-skewed. $F \geq 0$, ratio of variances.
- **CLT**: the mean (sum) of many independent variables is approx. normal whatever their distribution ($n > 30$). De Moivre–Laplace is its binomial case.

Testing — concepts

- The statement to be **proven goes into H₁**; H₀ is its opposite and contains the equality. Proving " $\mu = \mu_0$ " exactly is impossible (needs an equivalence test).
- A significance test controls only α (type I, false positive). β (type II, false negative) is not controlled; power = $1 - \beta$. $\alpha \downarrow \Rightarrow \beta \uparrow$; they do not add to 1.
- Not rejecting H₀ $\neq$ H₀ true ("H₀ cannot be rejected", never "accepted").
- **p-value** = $P(\text{result at least as extreme} \mid H_0)$. It is **not** $P(H_0 \text{ true})$, not the effect size, not $1 - P(H_1)$. Small $p \Rightarrow$ significant.
- **Significance $\neq$ relevance**: huge n makes tiny effects significant — judge with the CI and a relevance limit.
- Two-sided unless the opposite direction is impossible; one-sided p = half the two-sided p (choose before seeing data).
- Test $\Leftrightarrow$ CI: reject at level α exactly when μ_0 lies outside the $(1 - \alpha)$ CI.
- **Parametric** tests (assume e.g. normality) have more power when assumptions hold; **non-parametric** = distribution-free.
- Assumptions: independence, normality (or $n > 30$), equal variances for the pooled t-test (check with F-test, else Welch), paired data $\rightarrow$ paired test.
- Multiple testing inflates the global α ; Bonferroni α/k is simple but conservative (less power).
- χ^2 tests are **asymptotic** ($n > 30$, $e \geq 1$, 80% of $e \geq 5$); Fisher is **exact** (small samples, expensive). McNemar for paired categorical data. Cramér's V: strength only, no direction.

True or false? — classic traps

- | | |
|--|--|
| ✓ A metric attribute always has the properties of an ordinal one. | ✗ A density value can never be larger than 1. |
| ✗ A nominal attribute always has the properties of an ordinal one. | ✗ A 95% CI contains 95% of all observations. |
| ✗ Estimating a party's vote share from a poll is descriptive statistics. | ✗ If H ₀ is not rejected, H ₀ has been proven. |
| ✗ Adding a constant to all values changes the variance. | ✗ The p-value is the probability that H ₀ is true. |
| ✗ $r = 0$ means X and Y are independent. | ✓ Lowering α increases the risk of a type II error. |
| ✓ Converting km/h to mph leaves Pearson's r unchanged. | ✓ A statistically significant result need not be practically relevant. |
| ✗ Two disjoint events with positive probabilities are independent. | ✗ The χ^2 test is exact for any sample size. |