Statistics 1

Pearson Edexcel · International A-Level · Further Mathematics

Statistics 1: data, probability, regression and distributions

Statistics 1 (S1, paper WST01/01, 1 hour 30 minutes, 75 marks, first assessment June 2019) is the statistics partner of M1 in International AS Mathematics; P1 and P2 are assumed. The formula booklet carries the statistical tables; the formulas below are yours to know. This reference covers data representation and summary, probability, correlation and regression, discrete random variables and the Normal distribution.

Must-know for S1 (not in the booklet):

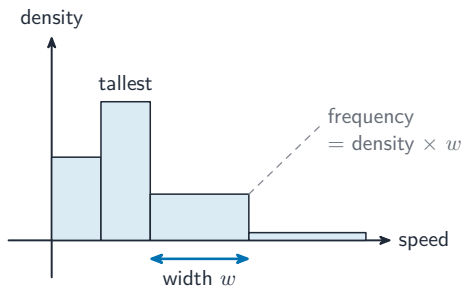
Family	Formulae
means	$\bar{x} = \dfrac{\sum x}{n}$ or $\dfrac{\sum fx}{\sum f}$
probability	$P(A') = 1 - P(A)$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, $P(A \cap B) = P(A)P(B \mid A)$
coding	$E(aX + b) = aE(X) + b$, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
Normal	$Z = \dfrac{X - \mu}{\sigma}$ for $X \sim N(\mu, \sigma^2)$

Mathematical models in probability and statistics

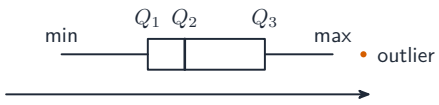
A statistical model is a simplified description of a real situation used to represent, analyse and predict. Reasons to use one: it is quicker and cheaper than the real experiment, it can be repeated and refined, and it exposes relationships. Know the standard reasons – "to make predictions" alone is not enough for the mark.

Representing data: histograms, stem and leaf, box plots

In a **histogram** the AREA is frequency: bar height is frequency density $\dfrac{\text{frequency}}{\text{class width}}$, so unequal classes get unequal widths, and comparing bars is comparing densities. A **stem and leaf** keeps the original values (back-to-back compares two groups). A **box plot** shows minimum, lower quartile, median, upper quartile, maximum, with **outliers** marked as separate crosses under whatever rule the question states.



area is frequency: height
= f/w for unequal classes



box plot: quartiles and
whiskers, outliers as crosses

Worked example – histogram heights by ratio (WST01 June 2025 Q5a). Speeds of 225 cars in classes of widths 5, 5, 10, 15, 15 with frequencies 54, 90, 60, 15, 6: the densities are 10.8, 18, 6, 1, 0.4. The tallest bar (25–30, density 18) is drawn 10 cm tall, so the second-tallest (20–25, density 10.8) measures $\frac{10.8}{18} \times 10 = 6$ cm. Beware: the 30–40 bar ALSO has density 6 – only the ratio argument earns the marks, not "60 ÷ 10".

Measures of location and dispersion

For grouped data, estimate the **median** or quartiles by linear interpolation inside the containing class: $Q_2 = a + w \times \dfrac{\frac{n}{2} - F}{f}$. The **interquartile range** is $Q_3 - Q_1$; variance is $\sigma^2 = \dfrac{\sum x^2}{n} - \bar{x}^2$ (or $\dfrac{\sum fx^2}{\sum f} - \bar{x}^2$). Combining groups: total $\sum x = \sum n_i \bar{x}_i$, total $\sum x^2$ adds directly, then reapply the formulas. The **standard deviation** is the square root of the variance. Coding reverses: if $w = 11.5d$, then $S_{td} = \dfrac{S_{tw}}{11.5}$ and $S_{dd} = \dfrac{S_{ww}}{11.5^2}$.

Worked example – combining two classes (WST01 June 2025 Q2). Class A: $n = 24$, $\bar{x} = 47$, $\sum x^2 = 66876$; class B: $n = 30$, $\bar{x} = b$, $\sum x^2 = 73826$; the 54 combined students average 45. Then $\sum x = 54(45) = 2430$, so $\sum x_B = 2430 - 24(47) = 1302$ and $b = 43.4$. The combined standard deviation: $\sigma = \sqrt{\frac{66876+73826}{54} - 45^2} = \sqrt{24.095\ldots} = 24.1$. If every B mark rises by 2: B's variance is unchanged (shifting moves nothing), the combined mean increases, and the combined SD decreases – B now sits closer to A, so the whole group is less spread.

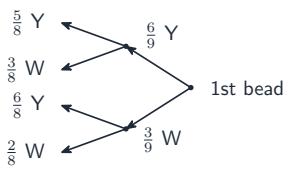
Worked example – interpolation quartiles (WST01 June 2025 Q5c–d). With cumulative frequencies 54, 144, 204, 219, 225: the median is the 112.5th value, inside 25–30: $Q_2 = 25 + 5 \times \frac{112.5-54}{90} = 28.25$. Q_3 is the 168.75th, inside 30–40: $Q_3 = 30 + 10 \times \frac{168.75-144}{60} = 34.125$, so with $Q_1 = 25.14$ given, $IQR = 8.985$.

Skewness

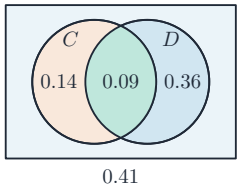
Compare distances: $(Q_3 - Q_2)$ versus $(Q_2 - Q_1)$, or mean versus median. Above, $5.875 > 3.11$ (and mean $30.6 >$ median 28.25), so the speeds are **positively skewed** (right-skewed) – which is also why a Normal model would be refused.

Probability

Work from the **sample space** : listing outcomes beats formula-grabbing. For any two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ and $P(A \cap B) = P(A)P(B \mid A)$; **independence** means $P(B \mid A) = P(B)$, giving the product rule. **Conditional probability** reverses the tree: $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$. Without replacement, the second-layer fractions drop by one in numerator and denominator.



without replacement: second-layer fractions drop by one top and bottom



$P(C \cup D) = P(C) + P(D) - P(C \cap D)$ fills the regions

Worked example – Venn to conditional (WST01 June 2025 Q4). $P(C \cup D) = 0.59$, $P(D) = 0.45$, $P(C \mid D) = 0.2$. Then $P(C \cap D) = 0.2 \times 0.45 = 0.09$ and $P(C) = 0.59 + 0.09 - 0.45 = 0.23$ – fill the Venn diagram and both answers fall out.

Worked example – a tree with reversal (WST01 June 2025 Q7). Bag: 3 white, 6 yellow; two drawn without replacement. Tree: $\frac{6}{9}, \frac{3}{9}$, then $\frac{5}{8}, \frac{3}{8}$ after yellow and $\frac{6}{8}, \frac{2}{8}$ after white. Exactly one yellow: $\frac{6}{9} \cdot \frac{3}{8} + \frac{3}{9} \cdot \frac{6}{8} = \frac{1}{2}$. At least one white: $1 - \frac{6}{9} \cdot \frac{5}{8} = \frac{7}{12}$. Given at least one white, both white: $\dfrac{\frac{3}{9} \cdot \frac{2}{8}}{\frac{7}{12}} = \dfrac{1}{12} \cdot \dfrac{12}{7} = \dfrac{1}{7}$.

Correlation and regression

The **product moment correlation coefficient** $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$ measures LINEAR association ($-1 \le r \le 1$); interpret in context. The least-squares **regression line** of y on x is $y = a + bx$ with $b = \dfrac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$ – only usable to predict y FROM x, and only within the data range (extrapolation is the flagged danger). Swap x and y and the line is different.

Worked example – coded regression (WST01 June 2025 Q3). With $w = 11.5d$, $S_{tt} = 26.43$, $S_{tw} = -91.55$, $\sum w = 339.25$, $\sum t = 9.1$, $\sum w^2 = 10036.45$ over 12 days: $S_{ww} = 10036.45 - \frac{339.25^2}{12} = 445.57$; the pmcc $r = \dfrac{-91.55}{\sqrt{26.43 \times 445.57}} = -0.844$ – warmer days, less gas. The regression of w on t: $b = \frac{-91.55}{26.43} = -3.46$, $a = \frac{339.25}{12} - (-3.46)\frac{9.1}{12} = 30.9$, so $w = -3.46t + 30.9$; divide through by 11.5 for d on t: $d = -0.301t + 2.69$. At $t = 2$: $d = 2.09$ m^3, and each extra degree means about 0.3 m^3 less gas.

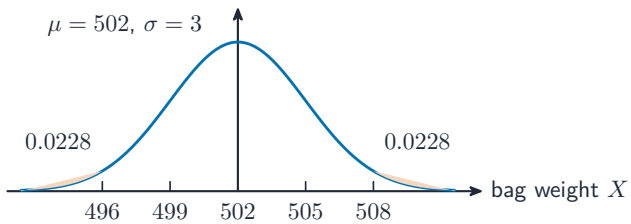
Discrete random variables

A **discrete random variable** takes isolated values with probabilities $p(x) = P(X = x)$ summing to 1; its **probability function** table is the object. **Expectation** $E(X) = \sum xp(x)$, $E(X^2) = \sum x^2 p(x)$, **variance** $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$. The cumulative function $F(x_0) = P(X \le x_0)$ adds probabilities up to x_0. The discrete UNIFORM distribution on $1, \ldots, n$ has $E(X) = \frac{n+1}{2}$ and $\mathrm{Var}(X) = \frac{n^2-1}{12}$.

Worked example – the full routine (WST01 June 2025 Q1). X with probabilities 0.3, 0.2, 0.35, 0.15 on 1–4. $F(3.5) = 0.3 + 0.2 + 0.35 = 0.85$. The inequality $3X - 3 < X + 2$ solves to $X < 2.5$, so $P = 0.5$ – solve for X FIRST, then add. $E(X) = 2.35$, $E(X^2) = 6.65$, $\mathrm{Var}(X) = 6.65 - 2.35^2 = 1.1275$; and $\mathrm{Var}(5 - 2X) = 4(1.1275) = 4.51$ by the coding rule.

The Normal distribution

For the **Normal distribution** $X \sim N(\mu, \sigma^2)$, standardise: $Z = \dfrac{X - \mu}{\sigma}$ then read the table. Symmetry does half the work: $P(X > \mu + 2\sigma) = P(Z > 2) = 0.0228$ and $P(\mu - 2\sigma < X < \mu + 2\sigma) = 1 - 2(0.0228) = 0.9544$. Reverse problems give z from the table, then $x = \mu + z\sigma$; two conditions on the same k give simultaneous equations.



standardise $Z = \dfrac{X - \mu}{\sigma}$: tails at $z = \pm 2$ hold 0.0228 each

Worked example – bags of flour (WST01 June 2025 Q6). Small bags $X \sim N(502, 3^2)$: $P(X > 508) = P(Z > 2) = 0.0228$. Of 4 bags, exactly 2 over 508:

$6(0.0228)^2(0.9772)^2 = 0.00298$. Large bags $Y \sim N(1024, \sigma^2)$ with 85th percentile 1038.51: $z_{0.85} = 1.0364$, so $\sigma = \frac{14.51}{1.0364} = 14$. If $P(X > k) = P(Y < 2k)$, then $\frac{k-502}{3} = \frac{1024-2k}{14}$, so $14k - 7028 = 3072 - 6k$, $k = 505$, and $p = P(Z > 1) = 0.159$.

Checklist before you call this unit done

- Give several reasons why statistical models are used; keep "cheap and quick" plus one more.
- Compute frequency densities for unequal classes; compare histogram bars by ratio; interpolate quartiles inside the containing class; mark outliers under the stated rule.
- Combine data sets through $\sum x$ and $\sum x^2$; reverse coding on S statistics; predict the effect of shifting or scaling on mean, variance and SD.
- Solve probability questions from Venn diagrams and trees, with and without replacement; reverse conditionals with $P(A \cap B)/P(A)$.
- Compute r from S statistics and interpret it; fit $y = a + bx$; predict only within range; pass coded data back to the original variable.
- Run the discrete DRV routine: probabilities sum to 1, solve inequalities for X first, $E(X)$, $E(X^2)$, variance, linear coding; quote the uniform results.
- Standardise both directions for Normal questions; use symmetry; finish two-condition problems with simultaneous equations.