Statistics 3

Pearson Edexcel · International A-Level · Further Mathematics

Statistics 3: samples, estimation, chi-squared and correlation

Statistics 3 (S3, paper WST03/01, 1 hour 30 minutes, 75 marks, first assessment June 2020) is the final statistics unit: S1–S2 are assumed. The booklet carries the Normal and chi-squared tables plus the Spearman critical values. This reference covers combinations of Normal variables, sampling methods, estimators and the sample mean's distribution, confidence intervals and **hypothesis tests** for means (one and two samples), goodness of fit and contingency tables, and rank correlation with its test.

Must-know for S3 (not in the booklet):

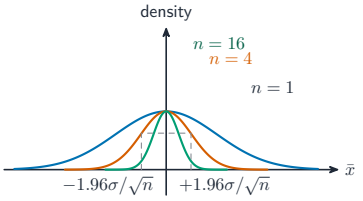
Family	Formulae
combinations	$X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ independent $\Longrightarrow aX \pm bY \sim N(a\mu_x \pm b\mu_y,\ a^2\sigma_x^2 + b^2\sigma_y^2)$
sample mean	$\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$; SE $= \dfrac{\sigma}{\sqrt{n}}$
unbiased variance	$s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1} = \dfrac{\sum x_i^2 - n\bar{x}^2}{n-1}$
test statistics	$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$; two samples $z = \dfrac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}$
chi-squared	$X^2 = \sum \dfrac{(O-E)^2}{E}$, $E = \dfrac{\text{row total} \times \text{col total}}{\text{grand total}}$
Spearman	$r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$

Combinations of random variables

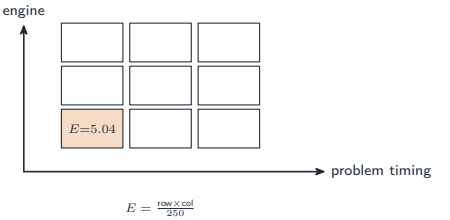
Independent Normal variables combine to a Normal: add means, add **squared** coefficients times variances. For $3S$ where $S \sim N(18, 0.03^2)$: $3S \sim N(54, 9 \times 0.0009)$ – the mean triples but the variance multiplies by nine. Differences work the same way: $C_1 - C_2 \sim N(0, 2\sigma^2)$, and $P(|C_1 - C_2| > c)$ is twice the one-tail probability by symmetry.

Worked example (WST03 June 2025 Q8). $S \sim N(18, 0.03^2)$, $C \sim N(25, 0.03^2)$, $P \sim N(15, 0.2^2)$ independently (kg). (a) $S_1 + S_2 + S_3 \sim N(54, 0.0027)$ (sd 0.05196): $P(> 54.1) = 1 - \Phi(1.924) = 0.0274$. (b) $C_1 - C_2 \sim N(0, 0.0018)$: $P(|C_1 - C_2| > 0.02) = 2(1 - \Phi(0.4714)) = 2 \times 0.3192 = 0.638$. (c) With $T = \sum_1^{28} S_i + \sum_1^5 C_i + P_1$, the event $T < 30P_1 + 190$ rearranges to $\sum S_i + \sum C_i - 29P_1 < 190$ – a Normal with mean $504 + 125 - 435 = 194$ and variance $28(0.0009) + 5(0.0009) + 29^2(0.04) = 33.6697$, so $P = \Phi\left(\frac{190-194}{5.804}\right) = 1 - 0.7549 = 0.245$.

Sampling methods



SE $= \sigma/\sqrt{n}$: the interval shrinks like $1/\sqrt{n}$



$X^2 = \sum \frac{(O-E)^2}{E} = 12.3$ on $(3-1)(3-1) = 4$ df

The **simple random sample** gives every member (and every sample of size n) an equal chance; use random numbers on a sampling frame. **Systematic sampling** takes every xth member from a random start – for 20 from 360, $x = 18$; it works well on ordered lists and removes selector bias, but members in the same stretch can never appear together (spacings are locked). **Stratified sampling** splits by strata and samples each proportionally – best when groups differ internally. **Quota sampling** fills interviewer-set quotas without a frame – cheap but vulnerable to interviewer bias. Know one advantage and one disadvantage of each.

Estimators, bias and the sample mean

A **statistic** contains no unknown population parameters. An estimator is **unbiased** if its expectation equals the parameter; its **bias** is the gap: $\bar{X}$ for μ, $s^2 = \dfrac{\sum x_i^2 - n\bar{x}^2}{n-1}$ for σ^2 (the $n-1$ is the correction). Bias is $\mathrm{E}(\hat{\theta}) - \theta$ in sign and size. The sample mean satisfies $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$: bigger samples shrink the **standard error** $\sigma/\sqrt{n}$.

Worked example (WST03 June 2025 Q6). $R = \frac{1}{2}X_1 + \frac{3}{4}X_{20}$: $\mathrm{E}(R) = (\frac{1}{2} + \frac{3}{4})\mu = \frac{5}{4}\mu \neq \mu$, so R is biased with bias $\frac{\mu}{4}$. Given $S = \frac{aX_1 - bX_2}{2}$ and $T = \frac{a\sum_1^{14} X_i + b\sum_{15}^{20} X_i}{20}$ both unbiased: the expectations force $a - b = 2$ and $14a + 6b = 20$, so $a = 1.6$, $b = -0.4$.

Worked example – recovering parameters (WST03 June 2025 Q7). Given $P(\bar{X}_n < 45.2) = 0.3446$ and $P(\bar{X}_n > 75.2) = 0.0179$: the z-values are -0.4 and 2.1, so $\mu - 0.4\frac{\sigma}{\sqrt{n}} = 45.2$ and $\mu + 2.1\frac{\sigma}{\sqrt{n}} = 75.2$. Subtracting: $2.5\frac{\sigma}{\sqrt{n}} = 30$, giving $\sigma = 12\sqrt{n}$ (SE 12) and $\mu = 45.2 + 4.8 = 50$; then $P(\bar{X}_n > 59) = 1 - \Phi(0.75) = 0.2266$.

Confidence intervals and tests for a mean

A **confidence interval** for μ (variance known) is $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$ with $z = 1.96$ (95%), 1.6449 (one-tailed 95% critical), 2.5758 (99%). Interpretation: the procedure captures μ in 95% of samples – not "the probability that μ lies inside is 0.95". The test

statistic $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$; when the variance is unknown and n is large, substitute s (the **Central Limit Theorem** also licenses non-Normal populations when n is large). Conclusions state the evidence in context.

Worked example – two samples with a gap (WST03 June 2025 Q4). Grey squirrels ($n = 80$, $\bar{x} = 500$, $s_x^2 = 1156$) and red ($n = 80$; $\sum y = 23200$, $\sum y^2 = 6741351$, giving $\bar{y} = 290$ and $s_y^2 = \frac{6741351 - 80(290)^2}{79} = 169$). Test $\mathrm{H}_0\colon \mu_x - \mu_y = 200$ against > 200 at 5%:

$$z = \frac{(500 - 290) - 200}{\sqrt{1156/80 + 169/80}} = \frac{10}{4.0698} = 2.46 > 1.6449,$$

so reject H_0 – the grey mean is more than 200 g above the red. The variance-substitution assumption ($\sigma^2 = s^2$ for each population) is what licenses the formula.

Goodness of fit and contingency tables

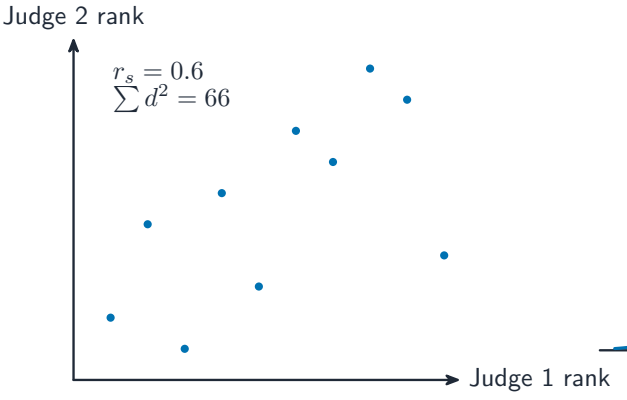
Goodness of fit compares observed counts with a model's expected ones: $X^2 = \sum \frac{(O-E)^2}{E}$. Degrees of freedom = (cells − 1) − (parameters estimated from the data): a binomial with $\hat{p}$ estimated loses one more. Combine tail cells whose **expected frequency** is below 5 before anything else. **Contingency tables** test association between two classifications with $E = \frac{\text{row} \times \text{column}}{\text{total}}$ and $\mathrm{df} = (r-1)(c-1)$.

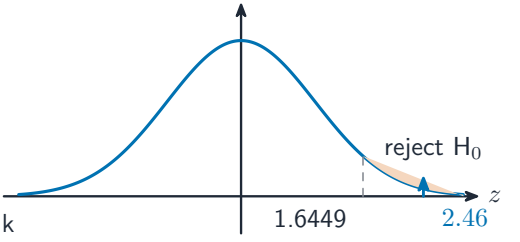
Worked example – binomial fit (WST03 June 2025 Q5). Males in 250 litters of 5 ($O = 2, 40, 90, 85, 30, 3$): $\hat{p} = \frac{610}{1250} = 0.488$, giving expected frequencies $8.79, 41.92, 79.91, 76.16, 36.30, 6.92$. The r-class contributes $\frac{(2-8.79)^2}{8.79} = 5.25$, so with the given $\sum = 5.70$ for the rest, $X^2 = 10.95$ on **degrees of freedom** $6 - 1 - 1 = 4$; the 5% critical value 9.488 is exceeded – reject: the binomial is not a suitable model.

Worked example – association (WST03 June 2025 Q3). Engine type × first-problem timing (250 cars, 3×3): the diesel/month 1–6 cell has $E = \frac{35 \times 36}{250} = 5.04$ and contributes $\frac{(10-5.04)^2}{5.04} = 4.88$; with 7.444 for the other eight cells, $X^2 = 12.325 > 9.488$ on 4 df – evidence of an association between engine type and when problems start.

Rank correlation and its test

Spearman's rank correlation coefficient measures monotonic agreement between two rankings; test $\mathrm{H}_0\colon \rho = 0$ against the table and conclude in words about the rankings.





positive but short of the 1% critical value 0.7455

one-tailed 5%: $z = 2.46 > 1.6449$ — the gap exceeds 200 g

Spearman's rank correlation coefficient $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$ measures monotonic agreement between two rankings (no ties examined). Test $H_0: \rho = 0$ against one- or two-tailed alternatives using the table of critical values; conclude in context ("insufficient evidence of positive agreement between the judges").

Worked example (WST03 June 2025 Q2). Two judges rank 10 performances: the rank differences give $\sum d^2 = 66$, so $r_s = 1 - \frac{6(66)}{990} = 0.6$. At the 1% one-tailed level the **critical value** is 0.7455: $0.6 < 0.7455$, so do not reject H_0 – insufficient evidence of positive correlation between the judges' rankings.

Checklist before you call this unit done

- Combine independent Normals (squared coefficients on variances); handle differences and |differences| by symmetry.
- Name and justify the four sampling methods; compute systematic spacing and its impossibility properties.
- Check statistics for parameters, compute expectations for bias, solve paired unbiased-estimator equations.
- Use SE $\sigma/\sqrt{n}$; recover (μ, σ) from two sample-mean probabilities; compute probabilities for X.
- Build and interpret z-based confidence intervals; run one- and two-sample tests with variance substitution and CLT stated.
- Compute expected frequencies (model or contingency), combine low cells, subtract estimated parameters from the df, compare with the χ^2 table.
- Compute Spearman's r_s from ranks and test against the table with hypotheses stated in symbols and the conclusion in words.