Statistics 2

Pearson Edexcel · International A-Level · Mathematics

Statistics 2: distributions, continuous variables and hypothesis tests

Statistics 2 (S2, paper WST02/01, 1 hour 30 minutes, 75 marks, first assessment June 2020) follows S1 in the applied route: P1–P2 and S1 are assumed. The booklet carries the binomial, Poisson and Normal tables. This reference covers the binomial and Poisson distributions with their approximations, continuous random variables, the uniform distribution, sampling and the full hypothesis-testing routine.

Must-know for S2 (not in the booklet):

Family	Formulae
binomial	$X \sim B(n,p)$: $E(X) = np$, $\mathrm{Var}(X) = np(1-p)$, $P(X=r) = \binom{n}{r} p^r (1-p)^{n-r}$
Poisson	$X \sim \mathrm{Po}(\lambda)$: $E(X) = \mathrm{Var}(X) = \lambda$; rates scale with time
continuous	$P(a < X < b) = \int_a^b f(x)\,dx$, $f(x) = F'(x)$, $E(X) = \int x f(x)\,dx$, $E(X^2) = \int x^2 f(x)\,dx$
uniform	on $[a,b]$: mean $\frac{a+b}{2}$, variance $\frac{(b-a)^2}{12}$, density $\frac{1}{b-a}$

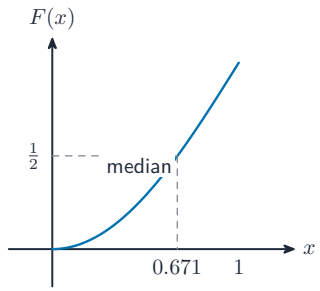
The binomial and Poisson distributions

Binomial distribution : n independent trials, constant probability p – the two assumptions to state in context. **Poisson distribution** : events occur singly, independently, at a constant average rate – the rate scales ($\mathrm{Po}(\lambda)$ per minute gives $\mathrm{Po}(5\lambda)$ in 5 minutes). Moments come as formulas: np and $np(1-p)$; λ and λ. Given the mean and variance of a binomial, solve for n and p simultaneously. The **Poisson approximation** to the binomial needs n large and p small (use $\lambda = np$).

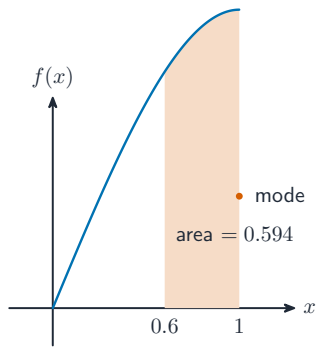
> **Worked example – moments then approximations (WST02 June 2025 Q3).** Blue sweets per small bag have mean 3 and variance 2.55: $np = 3$ and $np(1-p) = 2.55$ give $1-p = 0.85$, so $p = 0.15$ and $n = 20$. A medium bag of 40 with 20% green: $X \sim B(40, 0.2)$, $E(X) = 8$, and $P(X > 8) = 1 - P(X \le 8) = 0.407$. A large bag of 80 with 10% yellow: with $n = 80$ large and $p = 0.1$ small, $X \approx \mathrm{Po}(8)$ and $P(X \ge 12) = 1 - P(X \le 11) = 0.112$.

Continuous random variables

A **continuous random variable** has a **probability density function** f with total area 1 and $P(X = x) = 0$ at every point; the **cumulative distribution function** $F(x) = P(X \le x)$ satisfies $f = F'$ and $F(a) = \int_{-\infty}^{a} f$. Piecewise densities: normalise $\int f = 1$ (which pins the unknown constant), then read off the **median** ($F = \frac{1}{2}$), quartiles, and the **mode** from the peak of f (set $f' = 0$, or inspect corners). $E(X) = \int x f\,dx$, $E(X^2) = \int x^2 f\,dx$, $\mathrm{Var} = E(X^2) - [E(X)]^2$; skewness compares mean, median and mode.



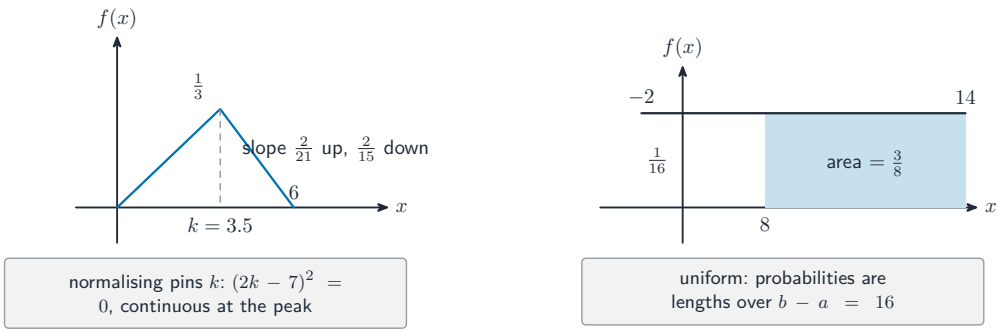
median solves $F(m) = \frac{1}{2}$;
mode differentiates $f = F'$



$P(X > 0.6) = 1 - F(0.6)$;
peak of f at the boundary

Left: the CDF F(x) = (6x^2 - x^4)/5 on [0, 1] rising from 0 to 1 with the median marked at F = 1/2 (x = 0.671). Right: the density f(x) = (12x - 4x^3)/5 with the mode at x = 1 and the area beyond x = 0.6 shaded (0.594).

> **Worked example – a CDF worked through (WST02 June 2025 Q1).** $F(x) = k(6x^2 - x^4)$ on $[0,1]$: $F(1) = 1$ gives $5k = 1$. $P(X > 0.6) = 1 - \frac{6(0.36) - 0.1296}{5} = 0.594$. The median: $\frac{6m^2 - m^4}{5} = \frac{1}{2}$, i.e. $2m^4 - 12m^2 + 5 = 0$, so $m = 0.671$. The mode: $f(x) = \frac{12x - 4x^3}{5}$ and $f'(x) = \frac{12 - 12x^2}{5} = 0$ gives $x = 1$ (reject -1, outside the support). Skewness: mean $0.64 <$ median $0.671 <$ mode 1 – negative skew.



Left: a triangular pdf rising along 2x/21 from 0 to k = 3.5 then falling along 2(6-x)/15 to x = 6, peak value 1/3 marked. Right: the uniform density on [-2, 14] at height 1/16 with the interval [8, 14] shaded (area 3/8).

> **Worked example – a piecewise-linear pdf, no calculator (WST02 June 2025 Q5).** With $f(x) = \frac{2x}{21}$ on $[0, k]$ and $f(x) = \frac{2(6-x)}{15}$ on $(k, 6]$: normalising,
>
> $$\frac{k^2}{21} + \frac{2}{15}\left(18 - 6k + \frac{k^2}{2}\right) = 1 \implies 4k^2 - 28k + 49 = 0 \implies (2k-7)^2 = 0,$$
>
> so $k = 3.5$ (the pdf is continuous there, both pieces equalling $\frac{1}{3}$). Then $E(X) = \frac{2}{21} \cdot \frac{k^3}{3} + \frac{2}{15}\int_{3.5}^{6}(6x - x^2)dx = \frac{19}{6}$, and with $E(X^2) = \frac{277}{24}$ given, $\mathrm{Var}(X) = \frac{277}{24} - \frac{361}{36} = \frac{109}{72}$.

The continuous uniform distribution

The **continuous uniform distribution** on $[a, b]$ Uniform on $[a, b]$: density $\frac{1}{b-a}$, straight CDF, mean $\frac{a+b}{2}$, variance $\frac{(b-a)^2}{12}$. Two facts pin a and b (an interval probability gives $b - a$; the mean gives $a + b$). Coding scales the mean; probabilities are lengths over $b - a$.

> **Worked example – uniform on a wide interval (WST02 June 2025 Q6).** Assume $a \le 2$ and $b \ge 5$. Then $P(2 < X < 5) = \frac{3}{16}$ gives $b - a = 16$; $E(X) = 6$ gives $a + b = 12$: so $a = -2$, $b = 14$ Without that support assumption, $a = 4.4$, $b = 7.6$ also satisfies the two given facts; the following calculations use the stated support assumption. $E(cX + 1) = 3$ gives $c = \frac{1}{3}$; $\mathrm{Var} = \frac{256}{12} = \frac{64}{3}$; $E(X^2) = \frac{64}{3} + 36 = \frac{172}{3}$; and $P(\frac{3}{2}X - b > a) = P(X > \frac{2(a+b)}{3}) = P(X > 8) = \frac{14-8}{16} = \frac{3}{8}$.

Populations, samples and statistics

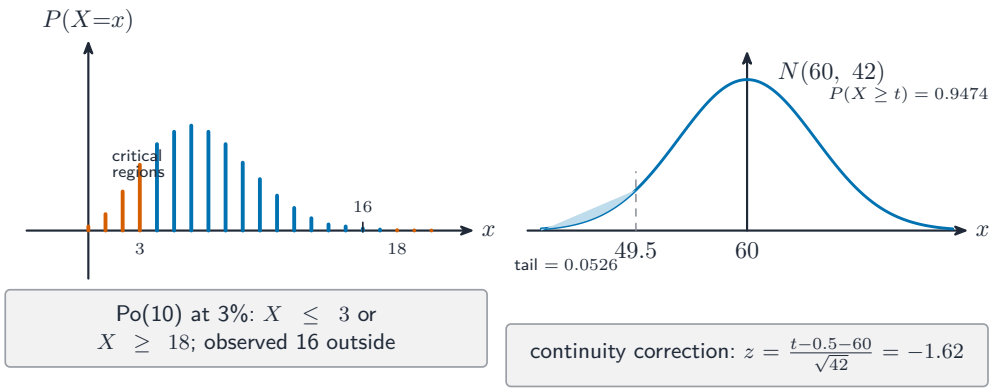
A **census** aims to measure everybody (complete but costly, slow, possibly destructive). To **sample** is to measure a selected part of the population; the selected units are also called a sample (cheaper and faster, but with sampling error). The **sampling frame** lists the sampling units. A **statistic** is a function of the sample values only (no unknown parameters) – $\bar{X}$ is, $\bar{X} - \mu$ is not; its distribution over repeated samples is the **sampling distribution** .

> **Worked example – a sampling distribution (WST02 June 2025 Q4).** Three beads from a 2:5 black:white bag, scoring 4 per black and 1 per white: with $P(\text{black}) = \frac{2}{7}$, the score X takes 3, 6, 9, 12 with probabilities $\frac{125}{343}, \frac{150}{343}, \frac{60}{343}, \frac{8}{343}$ (binomial weights $\binom{3}{0}(\frac{5}{7})^3$ etc.). If Y counts 9-point sets among n independent plays, $Y \sim B(n, \frac{60}{343})$ and $1 - (\frac{283}{343})^n > 0.95$ needs $n > 15.6$, so $n = 16$.

Hypothesis tests

A **hypothesis test** compares data against a null model.

State the **null hypothesis** H_0 (no change, the quoted parameter) and H_1 (one- or two-tailed as asked). Find the **critical region** at the stated nominal level by choosing tails whose total null probability does not exceed α; report the actual **significance level** as the total tail probability. One-tailed tests put the whole level in one tail; two-tailed split it. Conclude IN CONTEXT: "insufficient evidence that … has changed". Normal approximations to the binomial: mean np, variance $np(1-p)$, with the **continuity correction** (at least t means $t - 0.5$).



Left: the Po(10) distribution as vertical bars with the critical regions X <= 3 and X >= 18 in a contrasting colour and the observed 16 marked between them. Right: the N(60, 42) curve with the area below 49.5 shaded and the continuity-corrected z = -1.62 marked.

Worked example – a Poisson rate test (WST02 June 2025 Q2). Customers average 20/hour; after alterations one 30-minute interval is watched: $X \sim \text{Po}(10)$, $H_0\colon \lambda = 10$, $H_1\colon \lambda \neq 10$. At 3%: $P(X \leq 3) = 0.0103$ but $P(X \leq 4) = 0.0293$ (too big); $P(X \geq 17) = 0.0270$ but $P(X \geq 18) = 0.0143$ – critical region $X \leq 3$ or $X \geq 18$, actual level $0.0103 + 0.0143 = 2.46\%$. The observed 16 lies outside $\rightarrow$ do not reject H_0.

Worked example – binomial test with a Normal middle (WST02 June 2025 Q7). 30% of customers take insurance; a sample of 40 gives $X \sim B(40, 0.3)$. Largest r with $P(X < r) < 0.05$ (fewer than r): $P(X \leq 6) = 0.0238 < 0.05$ but $P(X \leq 7) = 0.0553 > 0.05$, so $r = 7$. For 200 customers, $B(200, 0.3) \approx N(60, 42)$: $P(X \geq t) = 0.9474$ means $\frac{t-0.5-60}{\sqrt{42}} = -1.62$, so $t = 50$. After a special offer, 11 of 25 buy: $H_0\colon p = 0.3$, $H_1\colon p > 0.3$; $P(X \geq 11) = 0.0978 < 0.10$, so reject H_0 – evidence the proportion has increased.

Checklist before you call this unit done

- State binomial/Poisson assumptions in context; solve mean/variance pairs for n, p; scale Poisson rates with time.
- Normalise piecewise densities (constants, perfect-square k); differentiate F for f; find median/quartiles from F, mode from f' or corners; run E(X), E(X^2), Var; read skewness from mean–median–mode.
- Quote and derive the uniform facts; pin a and b from an interval plus the mean; keep probabilities as lengths.
- Distinguish census vs sample with pros and cons; recognise statistics; build small sampling distributions by enumeration; wrap counts in B(n, p) for follow-ups.
- Run one- and two-tailed tests: hypotheses in symbols and words, critical region by accumulation, actual significance level, in-context conclusion; Normal approximation with continuity correction when n is large.