Statistics 2

Edexcel IAL S2 (WST02/01); the statistics unit after S1 in IAL Mathematics

Statistics 2: distributions, continuous variables and hypothesis tests

- use the binomial and Poisson distributions, and approximations to them;
- work with continuous random variables and their density and distribution functions;
- use the **continuous uniform distribution**;

Statistics 2: distributions, continuous variables and hypothesis tests (2)

- understand populations, samples and sampling distributions;
- carry out hypothesis tests for a binomial proportion and a Poisson mean.

The binomial and Poisson distributions

- there is a fixed number n of trials;
- each trial has two outcomes, success or failure;
- the trials are independent;

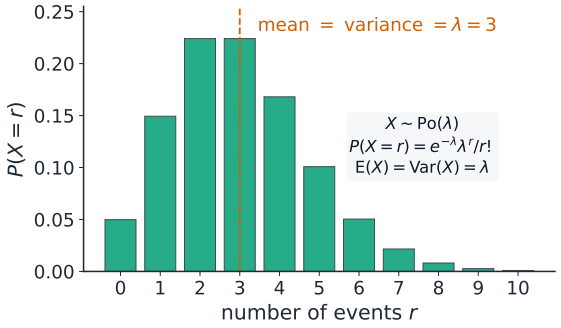
The binomial and Poisson distributions (2)

- the probability of success p is the same for every trial.
- if p is small, $B(n, p)$ is approximately $\text{Po}(np)$;
- if p is close to 0.5, $B(n, p)$ is approximately $N(np, np(1-p))$;

The binomial and Poisson distributions (3)

- if λ is large, Po(λ) is approximately $N(\lambda, \lambda)$.
- For a binomial model: $np = 3$ and $np(1-p) = 2.55$. Divide: $1 - p = 0.85$, so $p = 0.15$ and $n = 20$.
- A medium bag holds 40 sweets, of which 20% are green on average: $X \sim B(40, 0.2)$, so $E(X) = 8$ and $P(X > 8) = 1 - P(X \leq 8) = 1 - 0.5931 = 0.407$.
- A large bag holds 80 sweets, 10% yellow. n is large and p is small, so $X \approx$ Po(8), and $P(X \geq 12) = 1 - P(X \leq 11) = 1 - 0.8881 = 0.112$.

Figure: The binomial and Poisson distributions



- The **Poisson distribution** $\mathrm{Po}(3)$: for a Poisson variable the mean and variance both equal λ.
- Read the axes and units first.

Continuous random variables

- $f(x) \geq 0$, and the total area under f is 1;
- probability is area: $P(a < X < b) = \int_a^b f(x)\,dx$, and $P(X = a) = 0$ for any single value.
- The **median** m satisfies $F(m) = \frac{1}{2}$; the quartiles satisfy $F = \frac{1}{4}$ and $F = \frac{3}{4}$.

Continuous random variables (2)

- The **mode** is the value of x where $f(x)$ is greatest. Find it with $f'(x) = 0$, or from a sketch: it may be at an end of the interval.
- $\text{Var}(X) = E(X^2) - [E(X)]^2$.
- Skewness: if mean $<$ **median** $<$ **mode**, the distribution has negative skew; if mean $>$ median $>$ mode, positive skew.

Continuous random variables (3)

- $F(1) = 1$, so $5k = 1$ and $k = \frac{1}{5}$.
- $P(X > 0.6) = 1 - \dfrac{6(0.36) - 0.1296}{5} = 0.594.$
- **Median**: $\dfrac{6m^2 - m^4}{5} = \frac{1}{2}$, so $2m^4 - 12m^2 + 5 = 0$. Then $m^2 = \dfrac{12 - \sqrt{104}}{4} = 0.4505$ (the other root is bigger than 1), and $m = 0.671$.

Continuous random variables (4)

- pdf: $f(x) = \dfrac{12x - 4x^3}{5}$. Its derivative $\dfrac{12 - 12x^2}{5}$ is positive for $0 < x < 1$, so f is increasing and the **mode** is 1.
- Mean: $\int_0^1 \dfrac{12x^2 - 4x^4}{5}\, dx = \dfrac{4 - 0.8}{5} = 0.64.$
- Mean $<$ **median** $<$ mode, so the distribution has negative skew.

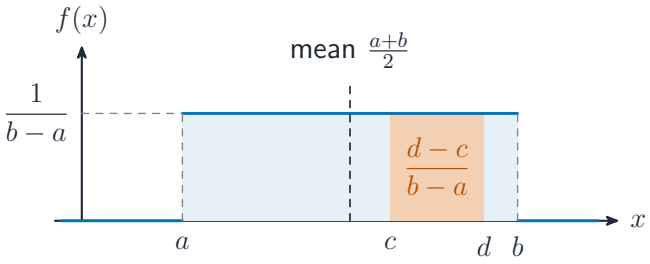
Continuous random variables (5)

- Total area: $\frac{k^2}{21} + \frac{2}{15}\left(18 - 6k + \frac{k^2}{2}\right) = 1.$
- Multiply by 105: $5k^2 + 252 - 84k + 7k^2 = 105$, so $12k^2 - 84k + 147 = 0$, that is $4k^2 - 28k + 49 = 0$, or $(2k-7)^2 = 0$. So $k = 3.5$, and both parts equal $\frac{1}{3}$ there.
- $E(X) = \int_0^{3.5} \frac{2x^2}{21}\,dx + \int_{3.5}^{6} \frac{2(6x - x^2)}{15}\,dx = 1.361 + 1.806 = \frac{19}{6}.$
- With $E(X^2) = \frac{277}{24}$, $\mathrm{Var}(X) = \frac{277}{24} - \left(\frac{19}{6}\right)^2 = \frac{109}{72}.$

The continuous uniform distribution

- $\frac{3}{b-a} = \frac{3}{16}$, so $b-a = 16$; and $\frac{a+b}{2} = 6$, so $a+b = 12$. So $a = -2$ and $b = 14$.
- $\text{Var}(X) = \frac{16^2}{12} = \frac{64}{3}$, so $E(X^2) = \frac{64}{3} + 6^2 = \frac{172}{3}$.
- If $E(cX+1) = 3$, then $6c+1 = 3$ and $c = \frac{1}{3}$.
- $P(\frac{3}{2}X - b > a) = P(X > \frac{2}{3}(a+b)) = P(X > 8) = \frac{14-8}{16} = \frac{3}{8}$.

Figure: The continuous uniform distribution



probability = length $\div$ $(b-a)$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

- Uniform on [a, b]: probabilities are lengths divided by b − a
- Match each part of the diagram to the text.

Populations, samples and statistics

- A **census** measures every member of the population. It gives complete information, but it is slow and expensive, and impossible if testing destroys the item.
- To **sample** is to measure only part of the population. A **sample** is quicker and cheaper, but the results vary from sample to sample.
- A **sampling frame** is a list of the members of the population that the sample is chosen from.

Populations, samples and statistics (2)

- $P(\text{black}) = \frac{2}{7}$. The number of black beads is $B(3, \frac{2}{7})$, and 0, 1, 2 or 3 black beads give a total of 3, 6, 9 or 12.
- The game is played n times. For the probability of at least one score of 9 to be over 0.95: $1 - \left(\frac{283}{343}\right)^n > 0.95$, so $n > \dfrac{\ln 0.05}{\ln(283/343)} = 15.6$, and the least n is 16.

Hypothesis tests

- H_0: $\lambda = 10$ (per 30 minutes), H_1: $\lambda \neq 10$. Under H_0, $X \sim \text{Po}(10)$. Put 1.5% in each tail.
- Lower tail: $P(X \leq 3) = 0.0103$, but $P(X \leq 4) = 0.0293$, so the lower **critical region** is $X \leq 3$.
- Upper tail: $P(X \geq 18) = 1 - 0.9857 = 0.0143$, but $P(X \geq 17) = 0.0270$, so the upper critical region is $X \geq 18$.

Hypothesis tests (2)

- The actual **significance level** is $0.0103 + 0.0143 = 0.0246$, or 2.46%.
- 16 is not in the **critical region**, so do not reject H_0: there is not enough evidence that the rate of customers has changed.
- For a **sample** of 40, $X \sim B(40, 0.3)$. The largest r with $P(X < r) < 0.05$: $P(X \leq 6) = 0.0238$ but $P(X \leq 7) = 0.0553$, so $r = 7$.

Hypothesis tests (3)

- For a **sample** of 200, $X \approx Y \sim N(60, 42)$. If $P(X \geq t) = 0.9474$, then with the **continuity correction** $\dfrac{t - 0.5 - 60}{\sqrt{42}} = -1.62$, so $t = 50$.
- After a special offer, 11 of 25 customers buy insurance. Test at the 10% level: H_0: $p = 0.3$, H_1: $p > 0.3$, with $X \sim B(25, 0.3)$.
  $P(X \geq 11) = 1 - 0.9022 = 0.0978 < 0.10$, so reject H_0: there is evidence that the proportion has increased.

Sources for this lesson

- Pearson Edexcel, *International Advanced Level Mathematics, Further Mathematics and Pure Mathematics specification* (2018), unit S2.
- L. von Bortkiewicz, *Das Gesetz der kleinen Zahlen* (Leipzig, 1898).
- Pearson Edexcel, WST02/01 June 2025, questions 1 to 7 (adapted).

Common errors

- Stating the conditions for a binomial or Poisson model without the context.
- Forgetting to scale the Poisson mean to the length of the interval.
- Treating $f(x)$ as a probability; for a continuous variable, probability is an area.

Common errors (2)

- Forgetting that the **mode** can be at an end of the interval.
- Leaving out the **continuity correction** when a Normal distribution approximates a discrete one.
- Writing hypotheses in terms of X or $\overline{x}$ instead of the parameter p or λ.

Common errors (3)

- For a two-tailed test, using the whole **significance level** in each tail.
- Writing "H_0 is true" or "this proves ..."; a test gives evidence, not proof.

Key terms

- **Poisson distribution**: the distribution of the number of random events in a fixed interval at a constant average rate.
- **Probability density function**: a function whose area over an interval gives the probability that X is in that interval.
- **Cumulative distribution function**: $F(x) = P(X \leq x)$.

Key terms (2)

- **Sampling distribution**: the distribution of a statistic over all possible samples.
- **Critical region**: the set of values of the test statistic for which H_0 is rejected.
- **Significance level**: the probability of rejecting H_0 when it is true.

Check yourself

- 1.
- Calls arrive at a rate of 3 per hour.
- What is the distribution of the number of calls in 2 hours? 2.
- $X \sim B(10, 0.4)$.