Week 45

A-Level Further Mathematics

Homework bundle for this week

本周作业合集

exercises.lol

Contents 目录

4. Further Probability & Statistics

Starter Quiz · 课前小测

A-Level Further Mathematics Name: _____ Score: _____

1. For f(x) = ½x on [0,2], F(x) = ¼x^2. The median solves ¼m^2 = 0.5. What is m? (2 dp)

   _____

2. A sample n = 10 has s = 4. With t = 2.262, what is the margin t × s/√n? (2 dp)

   _____

3. Observed counts 20, 30, 25, 25, each expected 25. What is $\chi^2 = \Sigma(O-E)^2/E$?

   _____

4. The key advantage of a non-parametric test is that it makes no assumption that the data is:

   ☐ normally distributed

   ☐ positive

   ☐ discrete

   ☐ paired

5. X has G(t) = 0.5 + 0.3t + 0.2t^2. Since E(X) = G′(1) and G′(t) = 0.3 + 0.4t, what is E(X)?

   _____

6. What must the total area under a probability density function equal?

   _____

7. You use the t-distribution (rather than the normal) when:

   ☐ the sample is small and the population variance is unknown

   ☐ the sample is very large

   ☐ the population variance is known

   ☐ the data is categorical

8. The chi-squared test statistic is:

   ☐ $\Sigma\ (O - E)^2/E$

   ☐ $\Sigma\ (O - E)$

   ☐ $\Sigma\ O/E$

   ☐ $(O - E)^2$

9. Under the null hypothesis, the sign test uses which distribution for the counts above/below the median?

- ☐ binomial with p = ½
- ☐ normal
- ☐ Poisson
- ☐ chi-squared

10. The probability generating function of X is defined as:
    - ☐ $E(t^x) = \Sigma\ P(X=x)\ t^x$
    - ☐ E(X)
    - ☐ $\Sigma\ P(X=x)$
    - ☐ e^x

A-Level Further Mathematics

1. **1.41**

   ¼$m^2 = 0.5 \rightarrow m^2 = 2 \rightarrow m = \sqrt{2} \approx 1.41$.

2. **2.86**

   $2.262 \times 4/\sqrt{10} = 2.262 \times 1.265 = 2.86$.

3. **2**

   $((-5)^2 + 5^2 + 0 + 0)/25 = 50/25 = 2$.

4. **normally distributed**

   Non-parametric tests work without assuming an underlying normal distribution.

5. **0.7**

   G'(1) = 0.3 + 0.4(1) = 0.7.

6. **1**

   A pdf integrates to 1 over its whole range.

7. **the sample is small and the population variance is unknown**

   Small n with unknown $\sigma \rightarrow$ estimate with s and use t.

8. **$\Sigma\ (O - E)^2/E$**

   χ^2 sums (observed − expected)^2 ÷ expected over all categories.

9. **binomial with p = ½**

   Each value is equally likely above or below the median, so the count is binomial with p = ½.

10. **$E(t^x) = \Sigma\ P(X{=}x)\ t^x$**

    $G(t) = E(t^x) = \Sigma\ P(X{=}x)\ t^x$.

A-LEVEL FURTHER MATHEMATICS • EXERCISE SHEET

4.1 Continuous random variables

Name: _____ Class: _____ Date: _____ **Total: 14 marks**

KEY WORDS probability density function 概率密度函数 • cumulative distribution function 累积分布函数 • median 中位数 • percentiles 百分位数 • continuous random variables 连续型随机变量

Objective

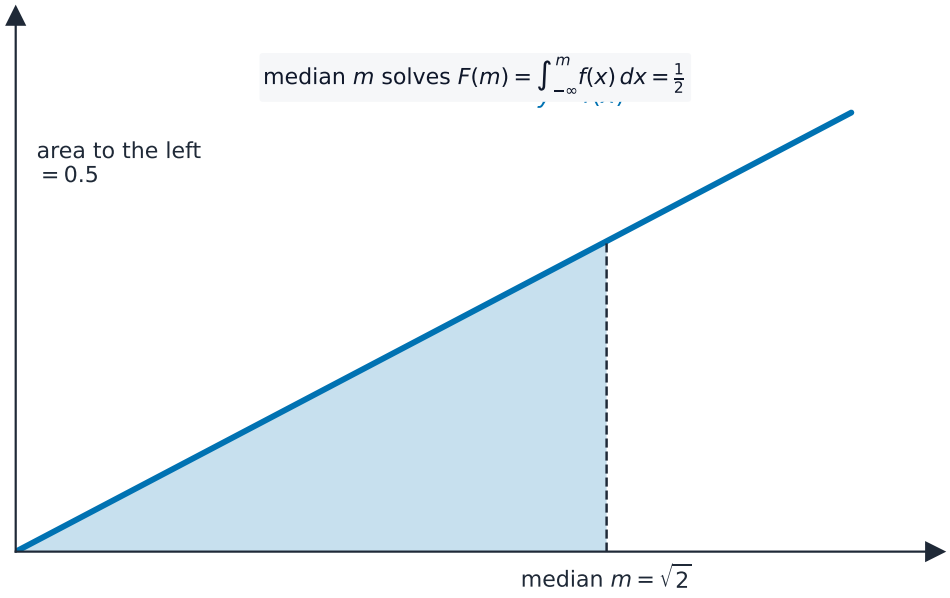
Build the skills to answer exam questions on **continuous random variables** 连续型随机变量 — the **probability density function** 概率密度函数, $E(g(X))$, and the **cumulative distribution function** 累积分布函数.

You must be able to:

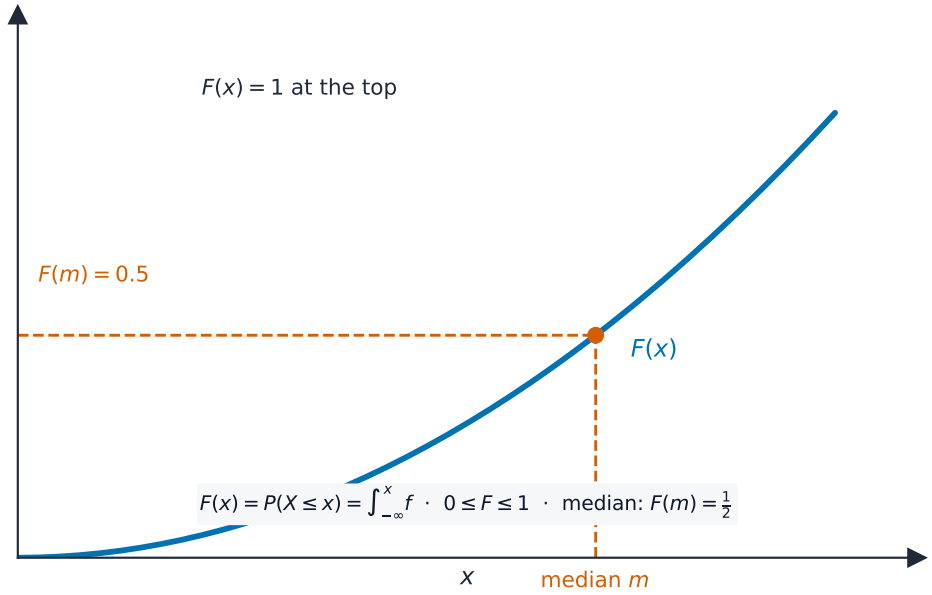
- find a constant / probability from a pdf (possibly piecewise)
- find $E(X)$, $\mathrm{Var}(X)$ and $E(g(X))$ by integration
- find $F(x)$ and use it for the median and percentiles

1 Worked examples

■ pdf, cdf and median



The median is where the area under $f(x)$ to its left is 0.5



$F(x)$ rises from 0 to 1; the median is where $F(x) = 0.5$

$f(x) = \frac{1}{2}x$ **on** $[0, 2]$**:** $F(x) = \frac{1}{4}x^2$; median $\frac{1}{4}m^2 = 0.5 \Rightarrow m = \sqrt{2}$.

2 Practice

2.1 For $f(x) = \frac{1}{2}x$ on $[0, 2]$, find $E(X)$. [2]

2.2 Write the relationship between $f(x)$ and $F(x)$. [1]

3 Exam-style questions

3.1 The median of a continuous variable solves: [1]

- **A** $f(m) = 0.5$
- **B** $F(m) = 0.5$
- **C** $E(X) = m$
- **D** $F(m) = 1$

3.2 A continuous variable has $f(x) = kx^2$ for $0 \leqslant x \leqslant 3$ (and 0 elsewhere).

(a) Find k. [3]

(b) Find $E(X)$. [3]

3.3 A variable has cdf $F(x) = \dfrac{x^2}{16}$ for $0 \leqslant x \leqslant 4$.

(a) Find the pdf $f(x)$. [2]

(b) Find the median. [2]

Solutions

2.1 $E(X) = \displaystyle\int_0^2 x \cdot \tfrac{1}{2}x\,dx = \left[\tfrac{x^3}{6}\right]_0^2 = \frac{4}{3}$.

2.2 $f(x) = F'(x)$.

3.1 B. The median solves $F(m) = 0.5$.

3.2 (a) $\displaystyle\int_0^3 kx^2\,dx = \left[\tfrac{kx^3}{3}\right]_0^3 = 9k = 1 \Rightarrow k = \tfrac{1}{9}$.

(b) $E(X) = \displaystyle\int_0^3 x \cdot \tfrac{1}{9}x^2\,dx = \tfrac{1}{9}\left[\tfrac{x^4}{4}\right]_0^3 = \tfrac{1}{9} \cdot \tfrac{81}{4} = \frac{9}{4}$.

3.3 (a) $f(x) = F'(x) = \dfrac{x}{8}$ for $0 \leqslant x \leqslant 4$.

(b) $\dfrac{m^2}{16} = 0.5 \Rightarrow m^2 = 8 \Rightarrow m = 2\sqrt{2} = 2.83$.

5 A continuous random variable X has probability density function f given by

$$\mathrm{f}(x) = \begin{cases} \frac{1}{16}\sqrt{x} & 0 \leqslant x < 4, \\ \frac{1}{k\sqrt{x}} & 4 \leqslant x \leqslant 9, \\ 0 & \text{otherwise,} \end{cases}$$

where k is a constant.

(a) Show that $k = 3$. [2]

.....

.....

.....

.....

.....

.....

.....

.....

(b) Find the median value of X. [3]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

The random variable Y is defined by $Y = \sqrt{X}$.

(c) Find the probability density function of Y. [5]

4 A continuous random variable X has cumulative distribution function F given by

$$\mathrm{F}(x) = \begin{cases} 0 & x < 1, \\ \frac{1}{5}x + a & 1 \leqslant x < 4, \\ \frac{1}{50}x^2 + b & 4 \leqslant x \leqslant 6, \\ 1 & x > 6, \end{cases}$$

where a and b are constants.

(a) Find the value of a and the value of b. [2]

.....

.....

.....

.....

.....

.....

.....

.....

(b) Find the probability density function of X. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(c) Given that $\mathrm{E}(X) = \frac{529}{150}$, find $\mathrm{Var}(X)$. [3]

(d) Find the 10th and 90th percentiles of X. [3]

2 The continuous random variable X has probability density function f given by

$$\mathrm{f}(x) = \begin{cases} \frac{4}{9}(x+1) & 0 \leqslant x \leqslant 1, \\ (x-2)^2 & 1 < x \leqslant 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the cumulative distribution function of X. [3]

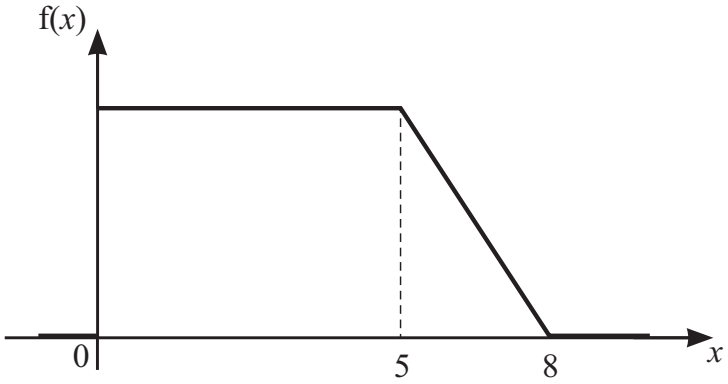
(b) Find the exact value of the median of X. [2]

The random variable Y is defined by $Y = \sqrt{X}$.

(c) Find the cumulative distribution function of Y. [3]

(d) Determine whether the median of Y is greater than, or less than, the median of X. [2]

2



As shown in the diagram, the continuous random variable X has probability density function f given by

$$f(x) = \begin{cases} a & 0 \leqslant x \leqslant 5, \\ b - cx & 5 \leqslant x \leqslant 8, \\ 0 & \text{otherwise,} \end{cases}$$

where a, b and c are constants.

(a) Show that $a = \frac{2}{13}$ and find the values of b and c. [3]

(b) Find the mean of X. [3]

(c) Find the median of X. [2]

The random variable Y is defined by $Y = X^2$.

(d) Find the cumulative distribution function for Y. [4]

3 A continuous random variable X has probability density function f given by

$$\mathrm{f}(x) = \begin{cases} kx & 0 \leqslant x < 1, \\ k(8-x) & 1 \leqslant x \leqslant 8, \\ 0 & \text{otherwise,} \end{cases}$$

where k is a constant.

(a) Show that $k = \frac{1}{25}$. [2]

.....

.....

.....

.....

.....

.....

(b) Find the median value of X. [3]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

The random variable Y is defined by $Y = \sqrt[3]{X}$.

(c) Find the probability density function of Y. [5]

4.2 Inference using normal and t-distributio

Name: _____ Class: _____ Date: _____ **Total: 12 marks**

KEY WORDS confidence interval 置信区间 • degrees of freedom 自由度

Objective

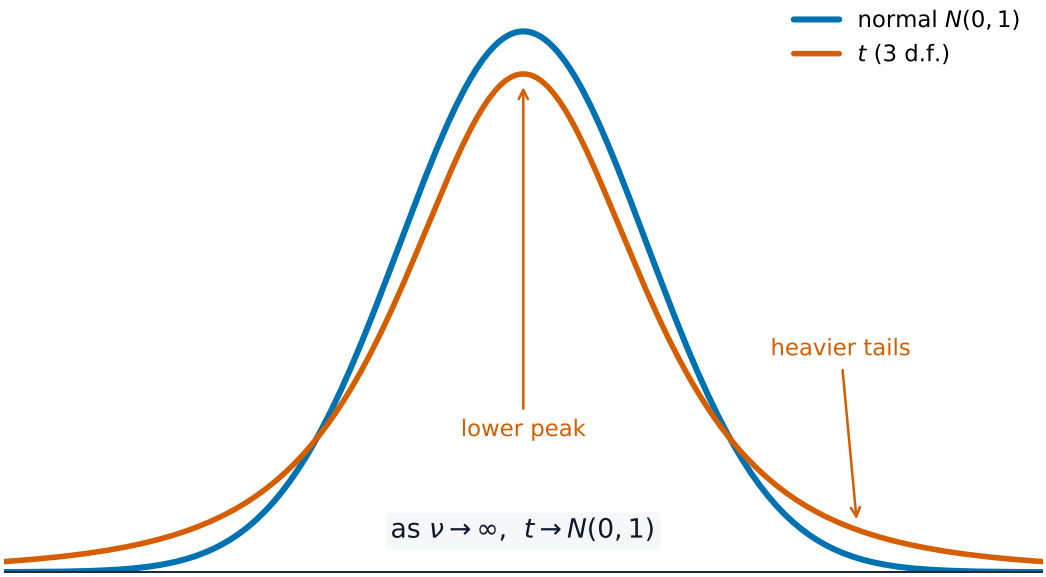
Build the skills to answer exam questions on **inference using the normal and t-distributions** 正态与 t 分布推断 — small-sample **confidence intervals** 置信区间 and **t-tests** t 检验.

You must be able to:

- find a confidence interval $\bar{x} \pm t\dfrac{s}{\sqrt{n}}$ (with $n-1$ d.f.)
- carry out a one-sample t-test
- carry out a two-sample / paired t-test (pooled variance where needed)

1 Worked examples

Small-sample interval



For a small sample the t-distribution is flatter with heavier tails, so its critical values are larger

$n = 10,\ \bar{x} = 50,\ s = 4,\ 95\%$ $(t = 2.262,\ 9$ **d.f.**$)$: $50 \pm 2.262\dfrac{4}{\sqrt{10}} = 50 \pm 2.86 =$

(47.1, 52.9).

2 Practice

2.1 State the number of degrees of freedom for a one-sample t-test with $n = 12$. [1]

2.2 State when a t-distribution is used instead of the normal. [1]

3 Exam-style questions

3.1 A 95% confidence interval based on a small sample uses a t value that is, compared with 1.96: [1]

- **A** smaller
- **B** larger
- **C** equal
- **D** zero

3.2 A sample of 8 measurements has $\bar{x} = 20.5$ and $s = 1.2$.

(a) Find a 95% confidence interval for the population mean (use $t = 2.365$, 7 d.f.). [3]

(b) State one assumption needed for this interval to be valid. [1]

3.3 A machine should produce rods of mean length 25 cm. A sample of 6 rods has $\bar{x} = 25.4$ cm and $s = 0.5$ cm. Test at the 5% level (two-tailed, $t_{\text{crit}} = 2.571$) whether the mean differs from 25. [5]

Solutions

2.1 $n - 1 = 11$.

2.2 when the sample is small and the population variance is unknown.

3.1 B. The small-sample t value is larger than 1.96.

3.2 (a) $20.5 \pm 2.365\dfrac{1.2}{\sqrt{8}} = 20.5 \pm 2.365(0.4243) = 20.5 \pm 1.00$; (19.5, 21.5).
(b) the population is (approximately) normally distributed.

3.3 $H_0 : \mu = 25$, $H_1 : \mu \neq 25$; $t = \dfrac{25.4 - 25}{0.5/\sqrt{6}} = \dfrac{0.4}{0.2041} = 1.96$; compare with 2.571; since $1.96 < 2.571$, do not reject H_0: no evidence the mean differs from 25.

1 A group of 10 school children are asked to estimate the size of an angle $\theta°$ in a given acute angled triangle. These estimates, in degrees, are as follows.

84 85 77 85 84 87 86 88 83 85

(a) Stating any assumptions you make, calculate a 95% confidence interval for θ. [5]

(b) Give a reason why the assumptions made in part **(a)** may not be appropriate in this case. [1]

3 A random sample of 10 newborn baby boys is taken and their masses in kg are recorded. From this sample, the population standard deviation of all newborn baby boys is estimated as 0.6 kg. A random sample of 5 newborn baby girls is taken and their masses in kg are recorded as follows.

3.9 3.1 2.9 3.1 3.6

It is assumed that the masses of newborn baby boys and girls have the same population standard deviation, σ kg.

By pooling the two samples, calculate an estimate of σ. [4]

6 Nine athletes in a club have a new coach. The coach adopts a new training programme which he believes will reduce the race times of these athletes. Each athlete completes a 1500 m time trial before and after completing the new training programme. Their times, in seconds (s), are recorded.

Athlete	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Time before training (s)	250	251	252	267	276	291	310	320	335
Time after training (s)	245	251	253	261	275	293	302	313	320

(a) Carry out a paired *t*-test at the 5% significance level to test the coach's belief. [7]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Further research suggests that the effects of the training programme tend to reduce the times of the slower athletes by more than those of the faster athletes.

(b) Suggest a reason why the paired *t*-test used in part **(a)** may not have been an appropriate test in this case. [1]

.....

.....

.....

(c) Suggest a suitable alternative test that could have been used instead of a paired *t*-test. [1]

.....

.....

.....

2 A factory produces packets of biscuits. The total mass of biscuits in a packet has a normal distribution with mean μ. A random sample of 12 packets is taken and the mass of the contents of each packet, x g, is recorded. The results are summarised as follows.

$$\sum x = 2390 \qquad \sum x^2 = 476117$$

(a) Find a 99% confidence interval for μ. [4]

A test of the null hypothesis $\mu = k$ is carried out on this sample using a 5% significance level. The test does not support the alternative hypothesis $\mu < k$.

(b) Find the greatest possible value of k. [3]

5 An engineer is comparing the tensile strengths of steel rods made from two machines, A and B. The engineer randomly selects 8 rods from machine A and 6 rods from machine B. The tensile strengths, in appropriate units, are given in the following table.

Machine A	402	403	415	412	409	407	406	410
Machine B	401	398	395	397	410	405		

You should assume that the two distributions are normal and have the same population variance.

Use a t-test at the 5% significance level to test whether there is any difference in the mean tensile strengths of steel rods from the two machines. [9]

1 Local residents are concerned about the speed of cars travelling through their village. They record the speed, in km h^{-1}, of a random sample of 13 cars travelling through their village. The recorded speeds are as follows.

40 53 59 42 43 48 62 67 46 82 66 45 70

(a) Construct a 90% confidence interval for the population mean speed of cars passing through the village. [5]

(b) State an assumption that is necessary for the confidence interval found in **1(a)** to be valid. [1]

3 Two different types of juice extractor, machine X and machine Y, are being compared. The manufacturer claims that machine Y extracts more juice per orange on average than machine X.

A random sample of 20 oranges is selected and each carefully measured. Oranges of a similar size and mass are then paired and numbered from 1 to 10. One orange from each pair is randomly allocated to machine X with the other allocated to machine Y. The amount of juice, in ml, extracted from each orange is recorded in the following table.

	pair									
	1	2	3	4	5	6	7	8	9	10
machine X	65	73	58	61	72	79	64	65	69	71
machine Y	68	72	64	63	75	82	63	63	72	74

(a) Use a t-test to test the manufacturer's claim at the 1% significance level. [7]

(b) State an assumption required for the test in **3(a)** to be valid. [1]

.....

.....

.....

.....

The manufacturer notices that the amount of juice extracted from the oranges in pair 4 were recorded incorrectly. In fact, the amount of juice extracted from machine X was 60 ml and from machine Y was 62 ml.

(c) Explain, with justification, whether the conclusion of the test in **3(a)** remains the same. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

1 The manager of a hardware store is interested in whether there is a difference in the amount spent per customer on weekdays ($\$x$) compared to weekends ($\$y$). Random samples of 120 customers on weekdays and 80 customers on weekends are taken and the amount spent by each customer is recorded. The results are summarised as follows.

$$\sum x = 10\,470 \qquad \sum (x-\overline{x})^2 = 12\,283 \qquad \sum y = 6560 \qquad \sum (y-\overline{y})^2 = 13\,520$$

Test at the 1% significance level whether there is a difference in the mean amount spent per customer on weekdays compared to weekends. You should not assume that the population variances of the amounts spent on weekdays and weekends are equal. [7]

4 A researcher is interested in whether there is a difference between two schools in students' aptitude for English. She randomly chooses 10 students from school X and 8 students of a similar age from school Y to take a written English test. The scores for the students from school X (x) and school Y (y) are summarised as follows.

$$\sum x = 612 \qquad \sum x^2 = 40\,104 \qquad \sum y = 444 \qquad \sum y^2 = 27\,460$$

You should assume that the two distributions are normal and have the same population variance.

(a) Find a 95% confidence interval for the difference in the mean scores for students from school X and students from school Y in the written English test. [6]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(b) Use the confidence interval you found in part **(a)** to explain why there is insufficient evidence at a 5% significance level to suggest that the English scores of students from school *X* and students from school *Y* are different. [1]

.....

.....

2 A farmer is investigating whether using a new fertiliser will increase the yield of tomato plants. The farmer selects 40 tomato plants at random and gives them the new fertiliser. The crop mass, x kg, of each of these 40 plants is recorded. The farmer selects a further 60 tomato plants at random and gives them a standard fertiliser. The crop mass, y kg, of each of these 60 plants is recorded. The results are summarised as follows.

$$\Sigma x = 168 \quad \Sigma x^2 = 720 \quad \Sigma y = 228 \quad \Sigma y^2 = 900$$

Find a 90% confidence interval for the difference in mean crop mass associated with each type of fertiliser. [7]

5 A doctor is investigating the concentration of blood glucose in patients at risk of developing type 2 diabetes, where blood glucose is measured in appropriate units. The doctor claims that a particular intervention reduces the concentration by more than k units on average. A group of 8 at risk patients is selected at random and each patient follows the intervention for six months. The blood glucose concentrations before and after the intervention are given in the following table.

Patient	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Before	183	165	172	165	143	176	161	153
After	164	148	164	149	134	153	155	148

(a) Use a t-test at the 5% significance level to find the range of values of k for which the result of the test is to reject the null hypothesis. [7]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(b) State an assumption necessary for the test in part **(a)** to be valid. [1]

.....

.....

A-LEVEL FURTHER MATHEMATICS • EXERCISE SHEET

4.3 Chi-squared tests

Name: _____ Class: _____ Date: _____ **Total: 11 marks**

KEY WORDS goodness of fit 拟合优度 • contingency table 列联表 • independence 独立性 • chi-squared test 卡方检验 • degrees of freedom 自由度

Objective

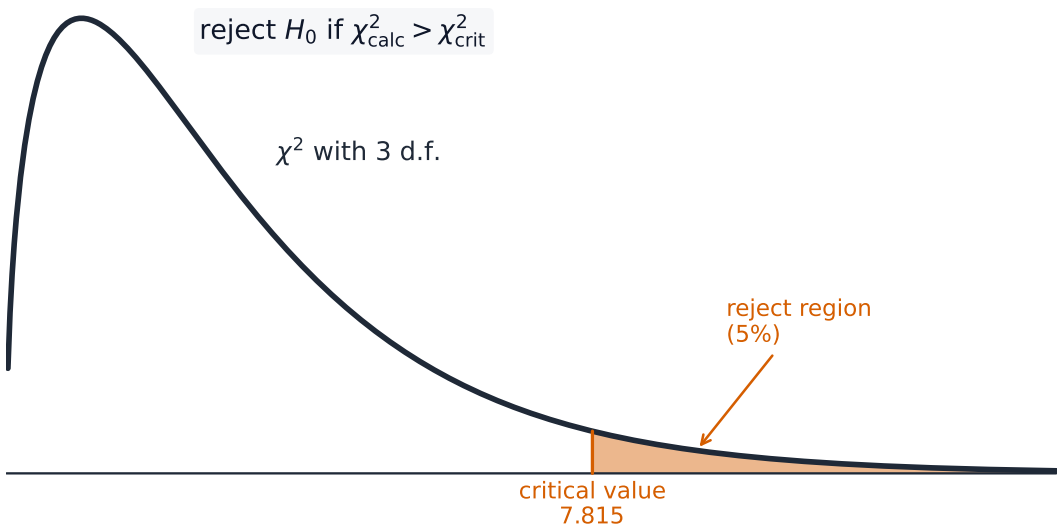
Build the skills to answer exam questions on **chi-squared tests** 卡方检验 — **goodness of fit** 拟合优度 and **independence** 独立性 in a **contingency table** 列联表.

You must be able to:

- compute expected counts and $\chi^2 = \sum \dfrac{(O-E)^2}{E}$
- find the degrees of freedom and compare with the critical value
- carry out a goodness-of-fit and an independence test

1 Worked examples

■ Goodness of fit



Reject the model when χ^2 exceeds the critical value (in the shaded tail)

Four equal categories, observed $20, 30, 25, 25$ **(each** $E = 25$**):** $\chi^2 = \dfrac{25+25+0+0}{25} = 2$. With 3 d.f. the 5% value is 7.815; $2 < 7.815 \rightarrow$ do not re-

ject.

2 Practice

2.1 Write down the χ^2 test statistic formula. [1]

2.2 A goodness-of-fit test has 5 categories and no estimated parameters. State the degrees of freedom. [1]

3 Exam-style questions

3.1 In a χ^2 test, a **large** value of χ^2 suggests: [1]

- **A** observed and expected agree well
- **B** observed and expected differ a lot
- **C** the sample is small
- **D** the test is invalid

3.2 A die is rolled 120 times, giving the counts:

Face	**1**	**2**	**3**	**4**	**5**	**6**
Observed	15	22	23	19	18	23

Test at the 5% level whether the die is fair ($\chi^2_{\text{crit}} = 11.07$, 5 d.f.). [6]

3.3 For a 3×4 contingency table, state the number of degrees of freedom for a test of

independence. [2]

Solutions

2.1 $\chi^2 = \sum \dfrac{(O-E)^2}{E}$.

2.2 $5 - 1 = 4$.

3.1 **B**. A large χ^2 means observed counts differ greatly from expected.

3.2 each $E = \dfrac{120}{6} = 20$; $\chi^2 = \dfrac{(15-20)^2 + (22-20)^2 + (23-20)^2 + (19-20)^2 + (18-20)^2 + (23-20}{20}$ $= \dfrac{25+4+9+1+4+9}{20} = \dfrac{52}{20} = 2.6$; $2.6 < 11.07$; do not reject — no evidence the die is unfair.

3.3 degrees of freedom $= (3-1)(4-1) = 6$.

2 The manager of a car park claims that the number of cars entering the car park follows a Poisson distribution with mean 2.8. The numbers of cars entering the car park are recorded on a working day during successive 5-minute periods. The following table contains the observed frequencies, together with most of the expected frequencies and their contributions to the χ^2-test statistic.

Number of cars	0	1	2	3	4	5	$\geqslant 6$
Observed frequency	2	15	31	29	13	3	7
Expected frequency	6.081	17.03	23.84	p	15.57	8.721	6.511
χ^2-test statistic	2.739	0.241	2.152	q	0.425	3.753	0.037

(a) Find the value of p and the value of q. [2]

.....

.....

.....

.....

.....

(b) Carry out a goodness of fit test at the 5% significance level to investigate the manager's claim. [4]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

3 A traffic expert claims that the number of breakdowns occurring each day on a busy section of a motorway follows a Poisson distribution with mean 0.7. The number of breakdowns each day over a 200-day period was recorded. The following table contains the observed frequencies together with some of the expected frequencies using the expert's distribution.

Number of breakdowns per day	0	1	2	3	4	$\geqslant 5$
Observed frequency	88	73	26	7	3	3
Expected frequency	99.317	m	24.333	5.678	0.994	n

(a) Find the value of m and the value of n, correct to 3 decimal places. [2]

.....

.....

.....

(b) Carry out a goodness of fit test at the 5% significance level to investigate the expert's claim. [6]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

5 A driving instructor believes that the performance (pass or fail) of students when taking a driving test is associated with their age. The following table summarises the number of students who pass and who fail, and the ages in years of the students taking the test, over a period of three years.

	age of student			
	under 20	20–40	over 40	total
pass	34	41	6	81
fail	16	39	9	64
total	50	80	15	145

Test, at the 10% significance level, whether performance is independent of age of student. [7]

3 Eggs in a supermarket are sold in boxes of six. A supermarket manager wishes to model the number of broken eggs in the boxes sold in the store. A random sample of 2000 boxes is taken and the number of broken eggs recorded. The observed frequencies are shown in the table below.

Number of broken eggs	0	1	2	3	4	5	6
Observed frequency	1844	143	11	0	1	0	1

(a) Use the data to estimate the probability that an egg is broken. Give your answer correct to 4 significant figures. [1]

.....

.....

.....

.....

.....

.....

It is decided to carry out a goodness of fit test at the 0.5% significance level to determine whether a binomial distribution fits the data.

The observed frequencies and the expected frequencies are given in the following table.

Number of broken eggs	0	1	2	3	4	5	6
Observed frequency	1844	143	11	0	1	0	1
Expected frequency	1831.3	a	6.016	0.119	0.001	0.000	0.000

(b) Show that $a = 162.6$ correct to 1 decimal place. [1]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(c) Carry out a goodness of fit test at the 0.5% level of significance to test whether a binomial distribution is a satisfactory model for the data. [5]

(d) Give a reason why a binomial distribution may not be a suitable model in this situation. [1]

1 A person's eye colour may be categorised as "brown", "blue" or "other". A scientist claims that these eye colours are uniformly distributed and hence are equally likely to occur in the population. A survey of 120 people from this population found that 38 people had brown eyes, 52 people had blue eyes and 30 people had eyes which were neither brown nor blue.

Use the data to carry out a goodness of fit test at the 5% significance level to test the scientist's claim. [6]

A-LEVEL FURTHER MATHEMATICS • EXERCISE SHEET

4.4 Non-parametric tests

Name: _____ Class: _____ Date: _____ **Total: 11 marks**

KEY WORDS rank-sum 秩和 • sign test 符号检验 • wilcoxon signed-rank 符号秩 • wilcoxon signed-rank test 威尔科 • non-parametric tests 非参数检验

Objective

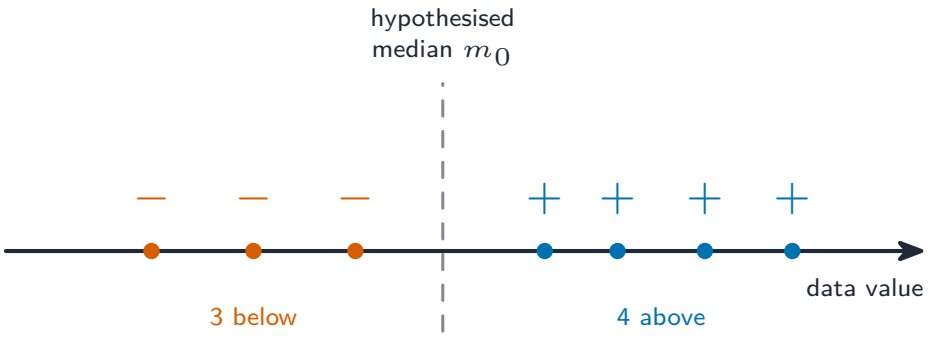
Build the skills to answer exam questions on **non-parametric tests** 非参数检验 — the **sign test** 符号检验, the **Wilcoxon signed-rank** 符号秩 (matched-pairs) test, and the **rank-sum** 秩和 test.

You must be able to:

- carry out a sign test using a binomial model
- carry out a Wilcoxon signed-rank test on paired data
- choose the right non-parametric test and state when to use one

1 Worked examples

■ The sign test



The sign test counts values above/below a proposed median, then tests the counts with a binomial model

Sign test for median $= m_0$: count + and − signs; under H_0 the number of + is $\sim B(n, 0.5)$; a very lopsided count is significant.

2 Practice

2.1 State when a non-parametric test is preferred over a t-test. [1]

2.2 In a sign test, state the distribution of the number of positive signs under H_0. [1]

3 Exam-style questions

3.1 The Wilcoxon signed-rank test uses, in addition to the signs of the differences: [1]

- **A** nothing else
- **B** the sizes (ranks) of the differences
- **C** the sample mean
- **D** the population variance

3.2 Ten people rate a product before and after a change. Eight rate it higher (+) and two lower (−).

(a) State H_0 and H_1 for a sign test of "no change" vs "an improvement". [2]

(b) Using $X \sim B(10, 0.5)$, find $P(X \geqslant 8)$ and state the conclusion at the 5% level. [4]

3.3 Explain why a non-parametric test might be chosen instead of a t-test for a small sample. [2]

Solutions

2.1 when the data cannot be assumed to be normally distributed.

2.2 $\sim B(n, 0.5)$.

3.1 B. The signed-rank test also uses the ranks (sizes) of the differences.

3.2 (a) H_0: median difference $= 0$ (no change); H_1: median difference > 0 (improvement).
(b) $P(X \geqslant 8) = P(8) + P(9) + P(10) = \dfrac{45 + 10 + 1}{1024} = \dfrac{56}{1024} = 0.0547$; $0.0547 > 0.05$, so do not reject H_0 — insufficient evidence of improvement.

3.3 a t-test assumes the population is normal; for a small sample that assumption cannot be checked reliably; a non-parametric test makes no such assumption, so it is safer when normality is doubtful.

4 A researcher believes that the median m of a population has changed from its known previous value m_0. The researcher collects a random sample of size 28. She ranks the data and calculates a test statistic T using the Wilcoxon signed-rank test. The conclusion of the test carried out at a 1% significance level is that there is not sufficient evidence to support her belief.

Using a normal approximation, find the least possible value of T. [5]

1 A large company claims that the median salary of its employees is \$32 500. The salaries (\$) of 15 randomly selected employees are listed below.

18 750	30 500	125 000	42 500	25 000
26 000	52 500	23 000	27 500	19 500
25 500	33 000	30 000	21 500	29 000

(a) Explain why a Wilcoxon signed-rank test may not be appropriate to test the company's claim in this case. [1]

.....

.....

.....

(b) Carry out a sign test at the 10% significance level to investigate the company's claim. [5]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

6 Students of the same age from two schools, school *A* and school *B*, take a large number of quizzes throughout the year and are each awarded a mark out of 1000. The marks of 123 students in school *A* and 147 students in school *B* are ranked from lowest (rank 1) to highest (rank 270). The sum of the ranks of the students from school *A* is 15 355.

(a) Carry out a Wilcoxon rank-sum test at the 5% significance level to investigate whether there is a difference in average marks between the students in school *A* and school *B*. [6]

(b) State an assumption that is required for the Wilcoxon rank-sum test to be valid. [1]

Additional page

If you use the following page to complete the answer to any question, the question number must be clearly shown.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

5 Two jigsaw puzzles have the same number of pieces with identical shapes but have different pictures printed on them. One puzzle has a seaside picture and the other has a cartoon picture. A researcher believes that children will complete the cartoon puzzle more quickly. To test this belief, 10 children are randomly selected. The time taken in seconds for each child to complete each puzzle is recorded below.

Child	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Seaside	182	130	193	181	192	204	184	192	180	189
Cartoon	161	111	195	159	202	200	168	165	145	160

(a) Carry out a Wilcoxon matched-pairs signed-rank test at the 5% significance level to test the researcher's belief. [6]

(b) Show that using a paired-sample sign test at the 5% significance level would result in the opposite conclusion to that found in part **(a)**. [3]

It was later discovered that the experiment had been conducted such that each child completed the seaside puzzle first followed by the cartoon puzzle.

(c) Comment on the validity of using this experiment to test the researcher's belief. [1]

4 A researcher claims that older people take longer to react to a sudden loud noise than younger people. To investigate this, the researcher randomly selects 6 people over 50 years old and 8 people under 25 years old and records their reaction times, in milliseconds, to a sudden loud noise. The reaction times are as follows.

Over 50	198	212	217	229	235	242		
Under 25	178	181	183	192	203	209	223	231

Carry out a Wilcoxon rank-sum test at the 5% significance level to test the researcher's claim. [8]

A-LEVEL FURTHER MATHEMATICS • EXERCISE SHEET

4.5 Probability generating functions

Name: _____ Class: _____ Date: _____ **Total: 14 marks**

KEY WORDS probability generating functions 概率母函数

Objective

Build the skills to answer exam questions on **probability generating functions** 概率母函数 — building $G(t)$, getting the mean and variance from its derivatives, and combining independent variables.

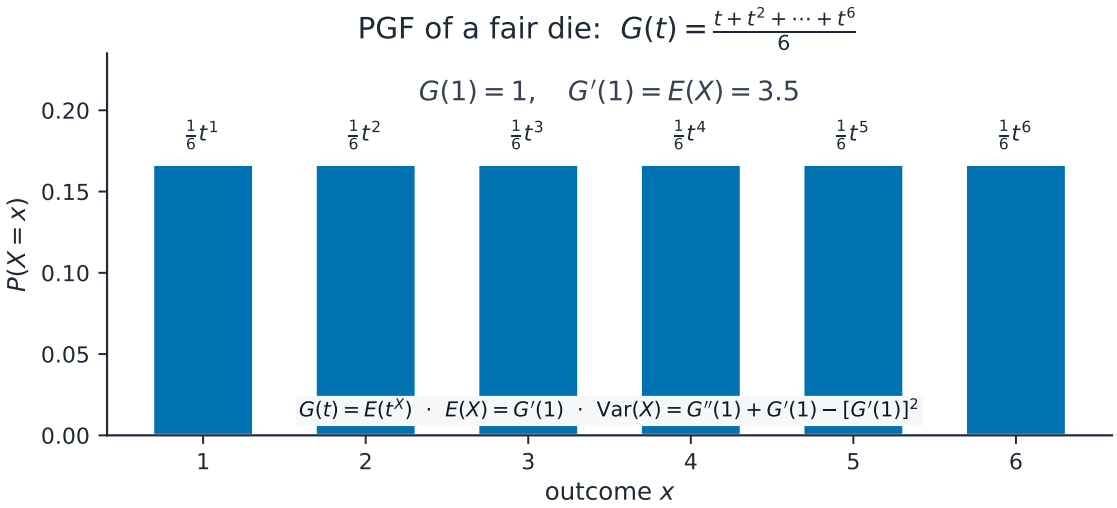
You must be able to:

- write $G(t) = E(t^X) = \sum P(X = x)t^x$
- use $E(X) = G'(1)$ and $\text{Var}(X) = G''(1) + G'(1) - [G'(1)]^2$
- multiply PGFs for a sum of independent variables

1 Worked examples

■ Mean from the PGF

A PGF packs a whole discrete distribution into one function $G(t)$



The PGF of a fair die: $P(X = x)$ is the coefficient of t^x, and $G'(1) = E(X) = 3.5$

$P(X = 0) = 0.5$, $P(X = 1) = 0.3$, $P(X = 2) = 0.2$: $G(t) = 0.5 + 0.3t + 0.2t^2$; $G'(t) = 0.3 + 0.4t$, so $E(X) = G'(1) = 0.7$.

2 Practice

2.1 Write down $G(t)$ for a variable with $P(X = 1) = P(X = 2) = P(X = 3) = \frac{1}{3}$. [2]

2.2 State the formula for $E(X)$ in terms of G. [1]

3 Exam-style questions

3.1 For any PGF, $G(1)$ equals: [1]

- **A** 0
- **B** $E(X)$
- **C** 1
- **D** $\text{Var}(X)$

3.2 A variable has $G(t) = \frac{1}{4}(1+t)^2$.

(a) Find the probability distribution of X. [3]

(b) Find $E(X)$ and $\text{Var}(X)$. [4]

3.3 X and Y are independent with PGFs $G_X(t) = \frac{1}{2}(1+t)$ and $G_Y(t) = \frac{1}{3}(1+t+t^2)$. Find the PGF of $X+Y$ and hence $P(X+Y=0)$. [3]

Solutions

2.1 $G(t) = \frac{1}{3}(t + t^2 + t^3)$.

2.2 $E(X) = G'(1)$.

3.1 C. $G(1) = \sum P(X = x) = 1$.

3.2 (a) $G(t) = \frac{1}{4}(1 + 2t + t^2) = \frac{1}{4} + \frac{1}{2}t + \frac{1}{4}t^2$, so $P(X = 0) = \frac{1}{4}$, $P(X = 1) = \frac{1}{2}$, $P(X = 2) = \frac{1}{4}$.
(b) $G'(t) = \frac{1}{2} + \frac{1}{2}t \Rightarrow E(X) = G'(1) = 1$; $G''(t) = \frac{1}{2} \Rightarrow \text{Var}(X) = \frac{1}{2} + 1 - 1^2 = \frac{1}{2}$.

3.3 $G_{X+Y}(t) = G_X(t)G_Y(t) = \frac{1}{2}(1+t) \cdot \frac{1}{3}(1+t+t^2) = \frac{1}{6}(1+t)(1+t+t^2)$; $P(X+Y=0)$ is the constant term $= \frac{1}{6}$.

7 A discrete random variable X takes values $r = 0, 1, 2$ with probabilities $\mathrm{P}(X = r)$ as given in the following table.

r	0	1	2
$\mathrm{P}(X = r)$	a	$2a$	b

(a) Write down the probability generating function of X, and use it to find an expression for $\mathrm{E}(X)$ in terms of a and b. [2]

(b) Show that $\mathrm{Var}(X) = 2b + 2(a+b)(1-2a-2b)$. [3]

The random variable Y is defined by $Y = X_1 + X_2 + X_3 + \cdots + X_{10}$ where $X_1, X_2, X_3, \ldots, X_{10}$ are ten independent observations of X.

(c) Using the probability generating function of Y, and your answer to part **(a)**, show that $\mathrm{E}(Y) = 10\mathrm{E}(X)$. [3]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(d) For the case $b = 0$, define fully the distribution of Y. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Additional page

If you use the following page to complete the answer to any question, the question number must be clearly shown.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

6 The discrete random variable X has probability generating function $G_X(t)$ given by

$$G_X(t) = \frac{t}{(3-2t)^2} .$$

(a) Find $E(X)$ and $Var(X)$. [5]

The discrete random variable Y has probability generating function $G_Y(t)$ given by

$$G_Y(t) = \frac{t^2}{(3-2t)^2}.$$

The random variable Z is the sum of the random variables X and Y.

(b) Assuming X and Y are independent, find $P(Z > 4)$. [5]

Additional page

If you use the following page to complete the answer to any question, the question number must be clearly shown.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

4 The random variable X takes values 1 and 2 with probabilities $\frac{2}{5}$ and $\frac{3}{5}$ respectively.

(a) Write down the probability generating function $G_X(t)$ of X. [1]

.....

.....

.....

The random variable Y is the sum of four independent observations of X.

(b) Find the probability generating function $G_Y(t)$ of Y. Give your answer in the form $G_Y(t) = at^m(b+ct)^n$, where a, b, c, m and n are constants to be determined. [2]

.....

.....

.....

(c) Use $G_Y(t)$ to find $P(Y = 6)$. [2]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(d) Find Var(Y). [5]

6 Y is a discrete random variable which takes the values 0, 2, 4, … . The probability generating function of Y is given by

$$G_Y(t) = \frac{k}{1-at^2}.$$

(a) Find k in terms of a. [1]

(b) Show that $P(Y > 2) = a^2$. [3]

It is now given that $a = 0.2$.

(c) Find the value of $\mathrm{E}(Y)$. [2]

Additional page

If you use the following page to complete the answer to any question, the question number must be clearly shown.

6 A bag contains 7 red balls and 3 blue balls. Kieran selects 2 balls at random, without replacement. The number of red balls selected by Kieran is denoted by X, and the number of different colours present in Kieran's selection is denoted by Y.

(a) Find the probability generating functions, $G_X(t)$ of X and $G_Y(t)$ of Y. [4]

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

The random variable Z is the sum of the number of red balls and the number of different colours present in Kieran's selection. Kieran claims that the probability generating function of Z is equal to $G_X(t) \times G_Y(t)$.

(b) Explain why Kieran is incorrect. [1]

.....

.....

.....

(c) Find the probability generating function of Z, expressing your answer as a polynomial in t. [4]

(d) Use the probability generating function of Z to find E(Z). [2]

Additional page

If you use the following page to complete the answer to any question, the question number must be clearly shown.