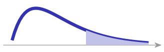
Further Probability & Statistics

A-Level Further Mathematics ▪ Topic 4

A-Level Further Mathematics



What we will cover

1 Continuous random variables

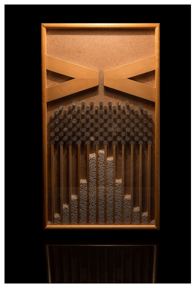
2 t-distribution inference

3 Chi-squared tests

4 Non-parametric tests

5 Generating functions

6 Recap



1 Continuous and inference

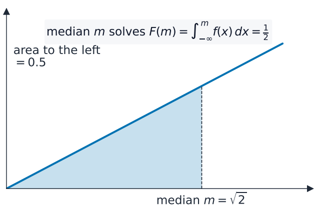
Continuous random variables

$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f$, and $f = F'$.

$f(x) = \frac{1}{2}x$, $0 \leqslant x \leqslant 2$

$F = \frac{1}{4}x^2$; median $m = \sqrt{2} = 1.41$

Continuous random variables, in detail



median: area 0.5 to its left

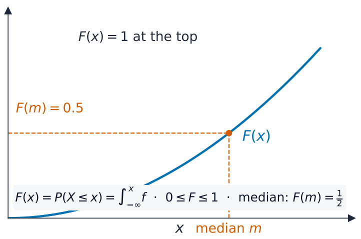
$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f$, and $f = F'$.

The cumulative distribution function

- **cdf 累积分布函数** $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$ – the running total of the density
- **going back** $f(x) = \frac{dF}{dx}$: differentiate the cdf to recover the density
- **its properties** non-decreasing, $F(-\infty) = 0$, $F(\infty) = 1$
- **the median** solve $F(m) = 0.5$; **percentiles 百分位数** solve $F(x) = p$
- **why it matters** any probability is a difference of two values: $P(a < X \leq b) = F(b) - F(a)$

For a probability, reach for F; for a mode or an expectation, reach for f. Knowing which one the question needs saves the whole calculation.

The cumulative graph $F(x)$ rises from 0



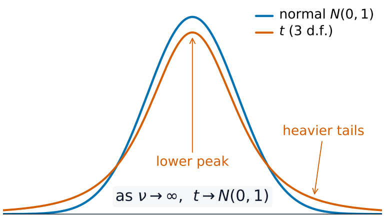
> The cumulative graph $F(x)$ rises from 0 to 1; the height 0.5 is reached at the median.

The distribution of a related variable

the setup	X has a known distribution and $Y = g(X)$; find Y's
the method	work through the cumulative function, never the density directly
for $Y = X^2$	$F_Y(y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$
then	differentiate to get $f_Y(y)$
the caution	mind the range of Y, and whether g is one-to-one over X's range

> Never substitute into the density and hope. The cdf route is the only one that handles a many-to-one g correctly.

t-distribution inference



Small sample, unknown variance
$\Rightarrow$ use the t-**distribution** t 分布:

$$\bar{x} \pm t\frac{s}{\sqrt{n}} \quad (n-1 \text{ d.f.}).$$

$n = 10,\ \bar{x} = 50,\ s = 4,\ t = 2.262$

$50 \pm 2.86 \Rightarrow (47.1,\ 52.9)$

When to use t rather than the normal

the condition	the sample is small and the population variance is unknown
the estimate	use the sample's own **unbiased estimate** 无偏估计 $s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2$
degrees of freedom	$\nu = n - 1$ for a single sample; the t-curve is wider than the normal and approaches it as ν grows
two samples	a **pooled estimate** 合并估计 of the common variance, with $\nu = n_1 + n_2 - 2$
confidence interval 置信区间	$\bar{x} \pm t_\nu \frac{s}{\sqrt{n}}$ – same shape as the normal one, with t in place of z

> The extra width of the t-curve is the price of estimating the variance from the same small sample. That is what makes the interval honest.

2 Tests

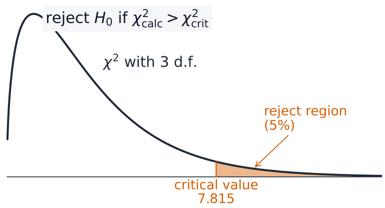
Chi-squared tests

Compare observed counts O with expected E:

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

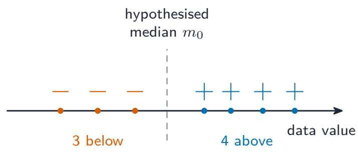
> Large $\chi^2 \Rightarrow$ reject H_0. Degrees of freedom $=$ categories -1, minus any estimated parameters.

Chi-squared tests, in detail



> The test rejects the model when χ^2 exceeds the critical value, landing in the shaded 5%

Non-parametric tests



No assumption that the data is normal:

- **sign test** 符号检验 – counts above/below a median
- **Wilcoxon** 威尔科克森 signed-rank & rank-sum – use the sizes too

Worked example – a sign test

Ten values are tested against H_0: median $= 5$. Nine fall above it. Is that evidence against H_0?

- under H_0, each value is above with probability 0.5, so the count $X \sim B(10, 0.5)$
- $P(X \geq 9) = \binom{10}{9}(0.5)^{10} + \binom{10}{10}(0.5)^{10}$
- $= \dfrac{10+1}{1024} = 0.0107$
- $0.0107 < 0.05$, so reject H_0 at the 5% level

> The whole test is a binomial tail. Note it uses *only* the signs – discard how far above the median each value was, and a great deal of information with it.

The three non-parametric tests, and what each uses

sign test 符号检验	uses only which side of the hypothesised median each value falls – the weakest, and the most assumption-free
Wilcoxon signed-rank 威尔科克森符号秩检验	the matched-pairs test: uses the sizes of the differences as well as their signs, so it is more powerful
Wilcoxon rank-sum 秩和检验	for two independent samples – ranks all the data together and compares the rank totals
when to use them	when normality cannot be assumed, or the data are ordinal
goodness of fit 拟合优度	χ^2 compares observed with expected frequencies; a **contingency table** 列联表 tests independence

> Matched pairs → signed-rank. Two independent samples → rank-sum. Choosing the wrong one loses the whole question, so read for paired or not first.

Which test, and on what assumption

Hypothesis test	Every **hypothesis test** 假设检验 states H_0, a test statistic, a significance level and a conclusion in context.
Parametric tests	t and χ^2 assume a **theoretical distribution** 理论分布 – normality, or expected counts of at least 5.
Wilcoxon signed-rank	The **Wilcoxon signed-rank test** 威尔科克森符号秩检验: paired data, no normality assumed. Rank the differences and compare the rank sums.
Wilcoxon rank-sum	The **Wilcoxon rank-sum test** 威尔科克森秩和检验: two *independent* samples, again distribution-free.

> A **probability density function** 概率密度函数 is what the parametric tests assume and the non-parametric ones avoid – which is why the latter are used on small or skewed samples.

Probability generating functions

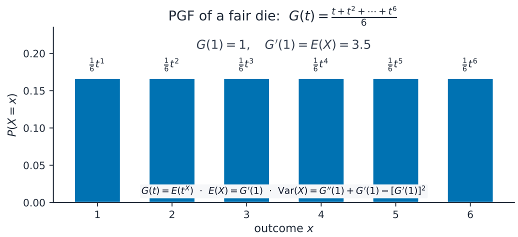
$$G(t) = E(t^X) = \sum_x P(X{=}x)\, t^x.$$

$E(X) = G'(1)$; the PGF of a sum of independent variables is the product of their PGFs.

$P = 0.5, 0.3, 0.2$ for $0, 1, 2$

$G = 0.5 + 0.3t + 0.2t^2$
$E(X) = G'(1) = 0.7$

PGF of a fair die packs equal



> PGF of a fair die packs equal masses at 1 through 6 into G(t)

Probability generating functions, continued

> Dice: a starting point for probability and discrete random variables.

Exam tips

- For a **continuous random variable**, the pdf integrates to 1 over its range, and $E(X) = \int xf(x)\,dx$.
- Use the t-**distribution** when the sample is small and the population variance is unknown; state the degrees of freedom.
- For a **chi-squared** test compute $\sum(O-E)^2/E$, compare with the critical value at the right degrees of freedom, and **combine classes with** $E < 5$.
- State H_0 and H_1 and give the conclusion **in context** for every test.

3 Recap

Topic 4 – key takeaways

1 Continuous
median solves $F(m) = 0.5$

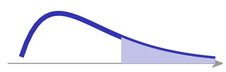
2 t-inference
small samples use the t-distribution

3 χ^2
$\sum(O-E)^2/E$; fit & independence

4 PGF
$G(t) = E(t^X)$; $E(X) = G'(1)$

Check yourself

1. Which distribution for a small-sample mean test?
2. Write the chi-squared statistic.
3. What does the sign test compare?
4. How do you get $E(X)$ from a PGF?



> **Answers** **1.** the t-distribution **2.** $\sum(O-E)^2/E$ **3.** counts above/below a median **4.** $E(X) = G'(1)$