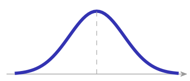
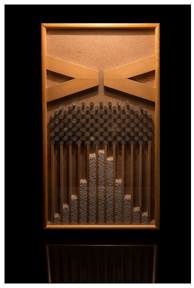
Probability & Statistics 1

A-Level Mathematics ▪ Topic 5

A-Level Mathematics

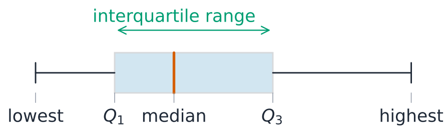


What we will cover



1 Data

Box plot: quartiles



box: Q_1–Q_3 · median line · whiskers: min/max

box plot: quartiles

> box plot: quartiles

Choose the diagram that suits the data

Diagram	What it is for
stem-and-leaf 茎叶图	keeps the original values *and* shows the shape; a **back-to-back** 背靠背 version compares two sets round a shared stem
histogram 直方图	grouped data, where the area of each bar gives the frequency
box-and-whisker	the five-figure summary: min, Q_1, median, Q_3, max
cumulative frequency	read the median across from half the total, quartiles from a quarter and three quarters

> In a histogram it is area, not height, that means frequency – which is why unequal class widths need frequency *density*.

The diagrams, and what each shows

Stem-and-leaf diagram — A **stem-and-leaf diagram** 茎叶图 keeps every original value while showing the shape – so the median and quartiles can be read straight off it.

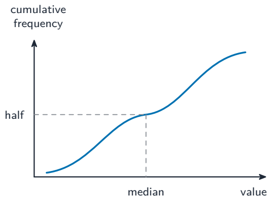
Box-and-whisker plot — A **box-and-whisker plot** 箱线图 shows the five-number summary and makes two distributions easy to compare side by side.

Histogram — For continuous data in unequal classes: it is the *area* that represents frequency, so the vertical axis is frequency density.

Cumulative frequency — Plot at the upper class boundary and read **percentiles** 百分位数 off the curve.

> A **measure of central tendency** 集中趋势度量 says where the data sits; the **range (of data)** 全距, interquartile range and standard deviation say how spread out it is.

Representation of data



A cumulative frequency curve is S-shaped; read the median across from half the total frequency, and the quartiles from one and three quarters

Averages and spread

- **mean** 平均数 $\bar{x} = \frac{\sum x}{n}$, **median** 中位数, **mode** 众数
- **standard deviation** 标准差 $\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$
- **interquartile range** 四分位距 $Q_3 - Q_1$

The mean, median and mode are **measures of central tendency** 集中趋势; the standard deviation and IQR are measures of **variation** 离散程度. **Coding** 编码 makes big numbers easy: put $t = x - a$ for an **assumed mean** 假定平均数 a, then $\bar{x} = a + \bar{t}$ and the standard deviation is unchanged.

$n = 10$, $\sum x = 50$, $\sum x^2 = 300$

$\bar{x} = 5$, $\sigma = \sqrt{30 - 25} = 2.24$

The measures of spread

the range	largest minus smallest – and wholly determined by two extreme values
the interquartile range 四分位距	upper quartile minus lower quartile – the middle 50%, so outliers cannot distort it
the standard deviation	the root-mean-square distance from the mean – uses every value
the variance	its square, which is what the formulas add and subtract
which to quote	the IQR with a median, the standard deviation with a mean
an outlier test	more than $1.5 \times$ IQR beyond a quartile

Pair the measures correctly. A mean with an IQR, or a median with a standard deviation, is a mismatch the examiner notices.

2 Counting and probability

Permutations and combinations

Permutation 排列 (order matters): $^nP_r = \frac{n!}{(n-r)!}$.

Combination 组合: $^nC_r = \binom{n}{r}$.

NEEDLESS (E$\times 3$, S$\times 2$)

$\frac{8!}{3!\,2!} = 3360$

permutation (order matters) **combination (order does not)**

AB BA $\{A, B\}$

NPr = n!/(n−r)!; nCr = nPr/r!

Arranging 3 of 4 objects ($^4P_3$)

$\boxed{4} \times \boxed{3} \times \boxed{2} = 24$

1st pick 2nd pick 3rd pick

Choosing 3 of 4 (order ignored): $^4C_3 = \frac{^4P_3}{3!} = 4$

$^nP_r = n!/(n-r)!$ · $^nC_r = {^nP_r}/r!$

nPr = n!/(n−r)!; nCr = nPr/r!

Worked example – arrangements with repeated letters

How many different arrangements are there of the letters of NEEDLESS?

- 8 letters, with E three times and S twice
- divide by the factorial of each repeat count
- $\dfrac{8!}{3!\,2!} = \dfrac{40320}{12} = 3360$

Repeated letters are indistinguishable, so every arrangement has been counted $3! \times 2!$ times. Dividing by the repeats is the whole correction.

Probability

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

Independent: $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. Conditional: $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$.

Mutually exclusive $\Rightarrow$ intersection is empty, so you just add. Independent is not the same as exclusive.

3 Distributions

Discrete random variables

$E(X) = \sum x\,P(X{=}x)$. **Binomial** 二项分布 $B(n, p)$:

$$P(X{=}r) = \binom{n}{r} p^r (1-p)^{n-r}, \quad E(X) = np.$$

$B(10, 0.3)$

$P(X{=}2) = 0.233$, $E(X) = 3$

The named distributions, and the probability rules

binomial 二项分布	n fixed independent trials, constant p: $P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$
geometric 几何分布	the trial on which the first success happens: $P(X = r) = (1-p)^{r-1} p$
the difference	binomial counts successes in a fixed number of trials; geometric counts trials until the first success
addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, and the last term is zero for mutually exclusive events
multiplication	for “and” with independent events, $P(A \cap B) = P(A)P(B)$
conditional 条件概率	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ – and independence means $P(A \mid B) = P(A)$

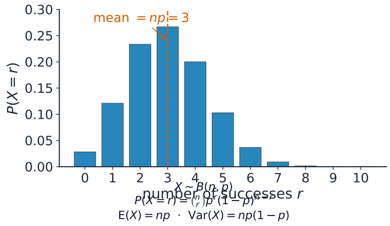
Check independence before multiplying. Drawing without replacement makes the second event conditional, and that is the trap these questions set.

The named discrete distributions

Probability distribution table	Any discrete random variable can be written as a **probability distribution table** 概率分布表: every value, with its probability, summing to 1.
Expectation	The **expectation** 期望 $\mathrm{E}(X) = \sum x\mathrm{P}(X = x)$ – the long-run mean, not a value the variable need ever take.
Binomial distribution	The **binomial distribution** 二项分布 $\mathrm{B}(n, p)$: a fixed number of independent trials, each a success or failure with the same p.
Geometric distribution	The **geometric distribution** 几何分布 counts trials *up to and including* the first success, so $\mathrm{P}(X = r) = p(1-p)^{r-1}$.

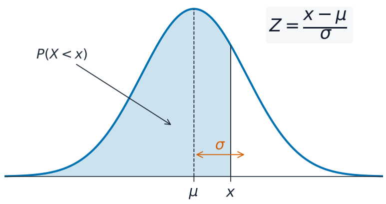
Choose by the question asked: how many successes in n trials, or how many trials until the first success.

Discrete random variables, in detail



The distribution $B(10, 0.3)$: each bar is $P(X = r)$, clustered around the mean $np = 3$.

The normal distribution



Normal 正态分布 $X \sim N(\mu, \sigma^2)$.
Standardise 标准化:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Read $P(Z < z)$ from the Φ table; a large-n binomial $\approx$ normal (with a **continuity correction** 连续性校正).

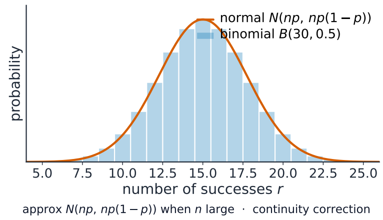
Worked example – finding an unknown mean

Bags of rice have mass $X \sim N(\mu, 0.14^2)$, and $P(X < 1.48) = 0.22$. Find μ.

- standardise: $Z = \frac{X - \mu}{\sigma}$
- from the table, $P(Z < z) = 0.22$ gives $z = -0.772$
- $\frac{1.48 - \mu}{0.14} = -0.772$
- $\mu = 1.48 + 0.772(0.14) = 1.59$ kg

The probability is below 0.5, so z must be negative – and reading the table's positive value and forgetting the sign is the standard error here.

When n is large the binomial bars



When n is large the binomial bars follow a normal curve of the same mean and variance.

Worked example – mean and standard deviation

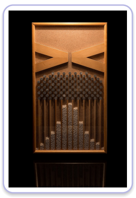
For 10 values, $\sum x = 50$ and $\sum x^2 = 300$. Find the mean and standard deviation.

$$\bar{x} = \frac{50}{10} = 5$$

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{30 - 25} = 2.24$$

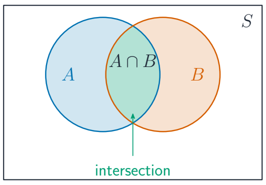
Variance $= \frac{\sum x^2}{n} - \bar{x}^2$ – the mean of the squares minus the square of the mean. Subtract in that order.

The normal distribution, continued



A Galton board: balls falling through pins pile up into the bell-shaped normal distribution.

Probability, continued



The overlap of the two circles is $A \cap B$; the addition rule subtracts it once so it is not counted twice.

Independence, conditioning, and the normal approximation

Independent events	**Independent events** 独立事件: $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. One happening tells you nothing about the other.
Conditional probability	**Conditional probability** 条件概率: $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$ – restrict to the times B happened, then take A's share.
Continuous random variable	A **continuous random variable** 连续随机变量 takes any value in a range, so probability is an area and $\mathrm{P}(X = x) = 0$.
Standardisation	**Standardisation** 标准化: $Z = \dfrac{X - \mu}{\sigma}$ turns any normal variable into the standard one, so a single table serves them all.

A binomial with large n is well approximated by a normal – an **approximation** 近似 that needs a continuity correction of ± 0.5.

Exam tips

- Decide whether order matters: **permutations (^nP_r) when it does, combinations (^nC_r)** when it does not.
- For a **discrete random variable**, check the probabilities sum to 1 and use $E(X) = \sum x\,P(X = x)$.
- For the **normal distribution**, standardise with $z = (x - \mu)/\sigma$, **sketch and shade**, then read the table.
- Apply a **continuity correction** when approximating a discrete variable by the normal.

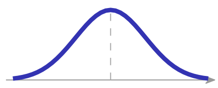
4 Recap

Topic 5 – key takeaways

1. **Data** — mean, median, mode; standard deviation
2. **Counting** — ^nP_r order matters; ^nC_r not
3. **Probability** — add for "or", multiply for independent "and"
4. **Distributions** — binomial $B(n, p)$; normal $N(\mu, \sigma^2)$

Check yourself

1. Does order matter for a combination?
2. The test for independence of A and B?
3. $E(X)$ for $X \sim B(20, 0.5)$?
4. How do you standardise a normal variable?



Answers **1.** no **2.** $P(A \cap B) = P(A)P(B)$ **3.** $np = 10$ **4.** $Z = (X - \mu)/\sigma$