# Probability & Statistics 2

A-Level Mathematics

This handout covers Topic 6: **Probability & Statistics** 概率统计 2. It adds the Poisson model, combining random variables, continuous distributions, and the ideas of estimation and testing.

## The Poisson distribution

*People arriving at random in a queue follow a Poisson distribution.*

Image: W.carter, CC0 (commons.wikimedia.org)

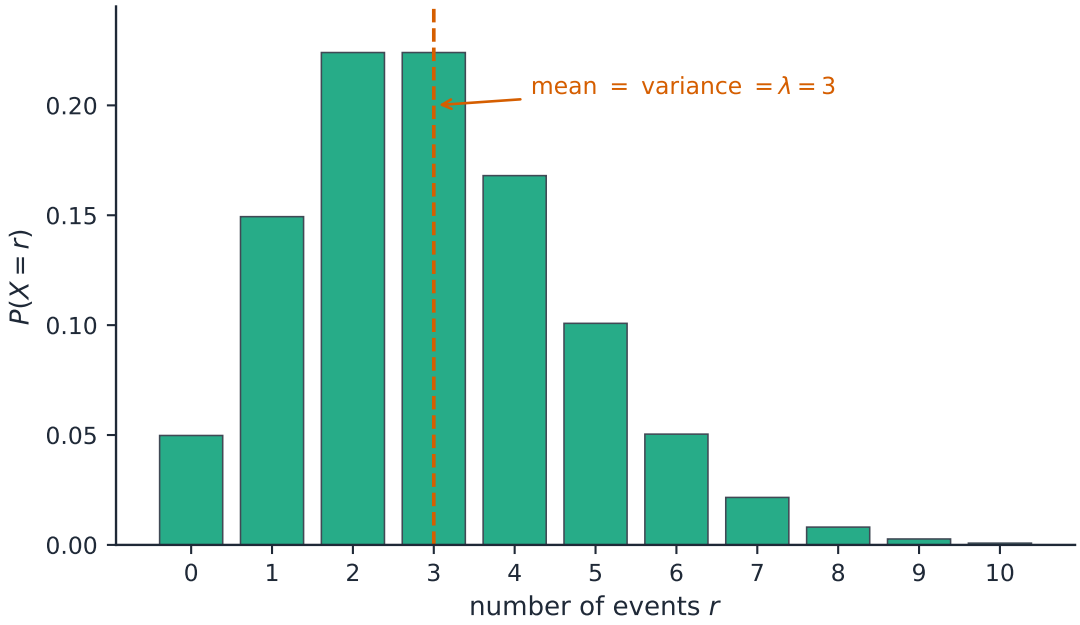
The **Poisson distribution** 泊松分布 $X \sim \text{Po}(\lambda)$ models the number of random events in a fixed interval, when events happen at a steady average rate $\lambda$:

$$P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}.$$

For a Poisson variable the mean and the variance are both equal to $\lambda$. The Poisson distribution is a good **approximation** 近似 to the binomial when $n$ is large and $p$ is small. The normal distribution (with continuity correction) approximates the Poisson when $\lambda$ is large.

**Worked example.** $X \sim \text{Po}(3)$. Find $P(X = 2)$.

$$P(X = 2) = e^{-3}\frac{3^2}{2!} = e^{-3} \times 4.5 = 0.224.$$




*The Poisson distribution* Po(3)*: for a Poisson variable the mean and variance both equal* $\lambda$*.*

## Linear combinations of random variables

When you change a variable by a linear rule, the **expectation** 期望 (mean) and **variance** 方差 follow these rules:

$$E(aX+b) = aE(X)+b, \qquad \operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X).$$

For two **independent** variables $X$ and $Y$:

$$E(aX+bY) = aE(X)+bE(Y), \qquad \operatorname{Var}(aX+bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y).$$

Two useful facts: if $X$ has a **normal distribution** 正态分布 then so does $aX+b$; and the sum of independent Poisson variables is again Poisson.

**Worked example.** $X$ has mean 5 and variance 4. Find $E(3X-1)$ and $\operatorname{Var}(3X-1)$.

$$E(3X-1) = 3(5)-1 = 14, \qquad \operatorname{Var}(3X-1) = 3^2 \times 4 = 36.$$

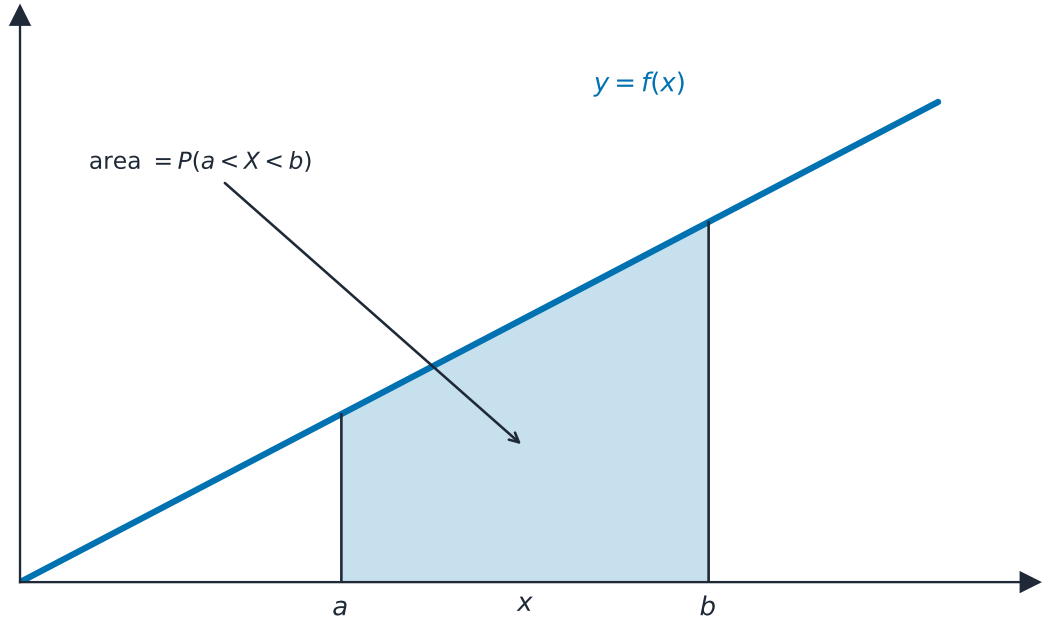
## Continuous random variables

A **continuous random variable** 连续型随机变量 can take any value in a range. Its probabilities come from a **probability density function** 概率密度函数 $f(x)$, with two key properties:

$$f(x) \geqslant 0, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

A probability is the area under $f$, and the mean is found by integration:

$$P(a < X < b) = \int_a^b f(x)\,dx, \qquad E(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx.$$




*For a continuous variable, the probability $P(a < X < b)$ is the area under $f(x)$ between $a$ and $b$.*

**Worked example.** A continuous variable has $f(x) = \frac{1}{2}x$ for $0 \leqslant x \leqslant 2$ (and 0 elsewhere). Find $E(X)$.

$$E(X) = \int_0^2 x \cdot \tfrac{1}{2}x \, dx = \int_0^2 \tfrac{1}{2}x^2 \, dx = \left[\tfrac{x^3}{6}\right]_0^2 = \frac{8}{6} = \frac{4}{3}.$$

# Sampling and estimation

*Statistics studies a sample to learn about a whole population.*

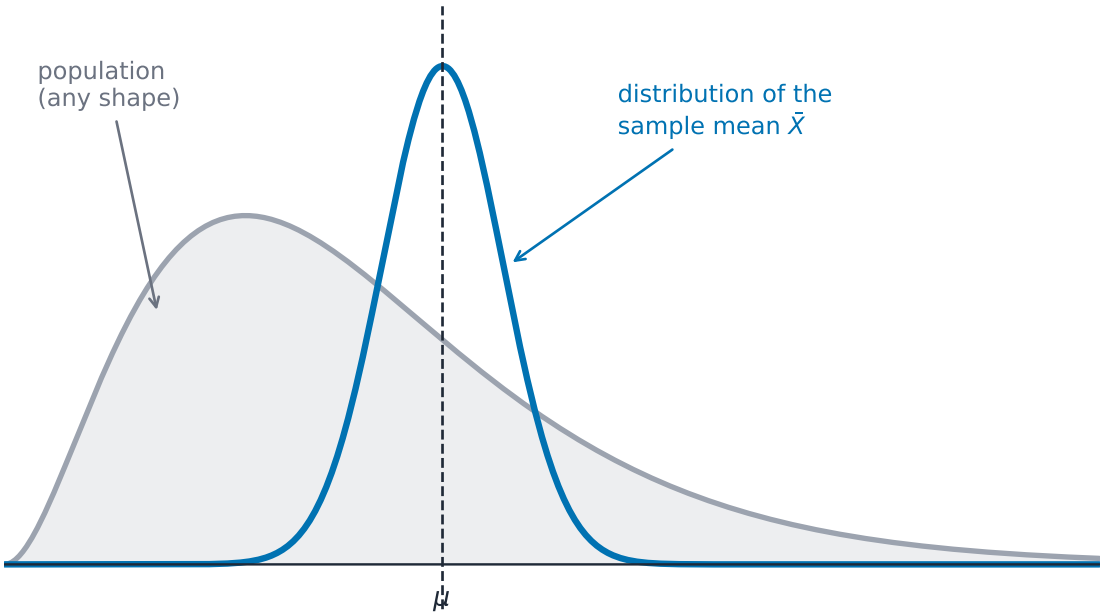
Image: Yann Kemper, CC0 (commons.wikimedia.org)

A **sample** 样本 is a small group chosen from the whole **population** 总体. Good sampling needs **randomness** 随机性, so that every member has a fair chance of being chosen.

The sample mean $\bar{X}$ is itself a random variable, with

$$E(\bar{X}) = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

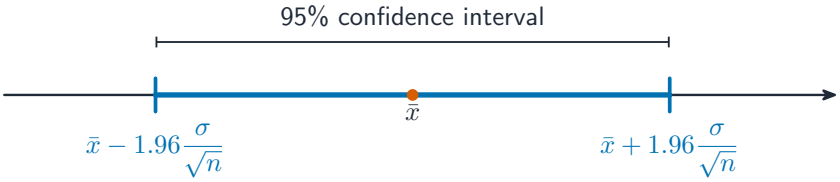
By the **Central Limit Theorem** 中心极限定理, for a large sample $\bar{X}$ is approximately normal, whatever the shape of the population.




*Whatever the population's shape, the sample mean $\bar{X}$ has a narrow, near-normal distribution centred on $\mu$.*

From a sample you can find **unbiased estimates** 无偏估计 of the population mean and variance. A **confidence interval** 置信区间 gives a range that probably contains the true mean. When the population is normal with known $\sigma$ (or the sample is large), a 95% interval is

$$\bar{x} \pm 1.96\,\frac{\sigma}{\sqrt{n}}.$$




*A 95% confidence interval stretches 1.96 standard errors each side of the sample mean.*

You can also find a confidence interval for a **population proportion** 总体比例 from a large sample.

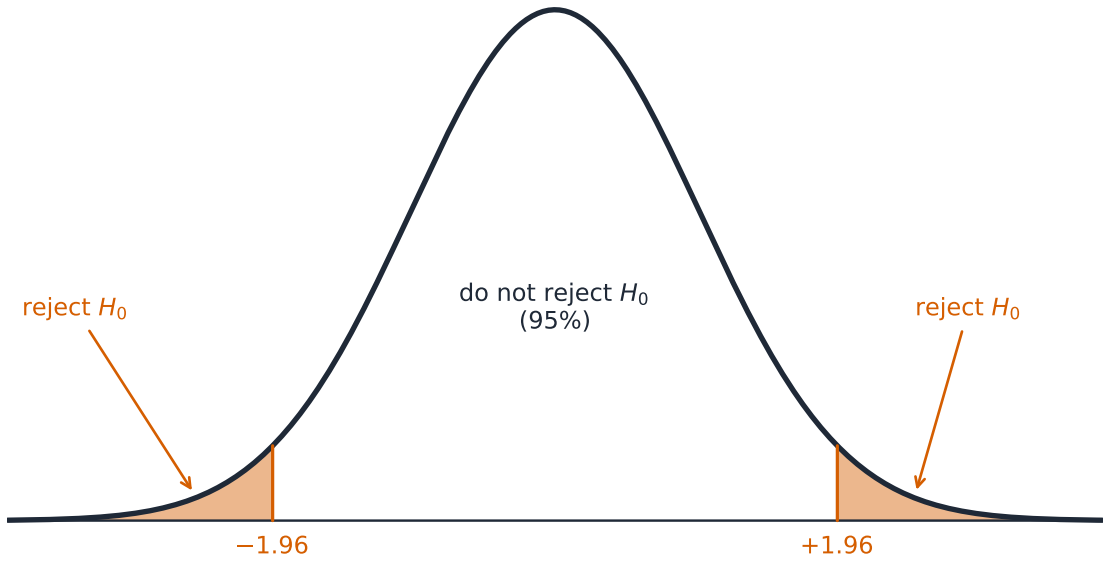
**Worked example.** A sample of $n = 64$ has mean $\bar{x} = 50$, from a population with $\sigma = 8$. Find a 95% confidence interval for the population mean.

$$50 \pm 1.96 \times \frac{8}{\sqrt{64}} = 50 \pm 1.96 \;\Rightarrow\; (48.0,\ 52.0).$$

# Hypothesis tests

A **hypothesis test** 假设检验 uses sample data to judge a claim. You set up two statements: the **null hypothesis** 原假设 $H_0$ (the claim being tested, usually "no change") and the **alternative hypothesis** 备择假设 $H_1$ (what you suspect instead). The test is **one-tailed** 单尾 if $H_1$ points one way (e.g. $\mu > 50$) and **two-tailed** 双尾 if it allows both ways ($\mu \neq 50$).

You fix a **significance level** 显著性水平 (often 5%), work out a **test statistic** 检验统计量 from the data, and see whether it lands in the **rejection region** 拒绝域. If it does, you reject $H_0$.




*A two-tailed test at 5% rejects $H_0$ only if the test statistic falls in a shaded tail beyond $\pm 1.96$.*

Two mistakes are possible: a **Type I error** 第一类错误 is rejecting $H_0$ when it is actually true; a **Type II error** 第二类错误 is accepting $H_0$ when it is actually false.

**Worked example.** A population is claimed to have mean 50, with $\sigma = 8$. A sample of $n = 64$ gives $\bar{x} = 52$. Test at the 5% level whether the mean has changed.

$H_0 : \mu = 50$ and $H_1 : \mu \neq 50$ (two-tailed). The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{52 - 50}{8/8} = 2.$$

The critical value at 5% (two-tailed) is 1.96. Since $2 > 1.96$, you reject $H_0$: there is evidence the mean has changed.