

Answers to Questions

Probability and Statistics: Detailed Explanations

1. Random variables

A **random variable** assigns a real number to each outcome of a random experiment.

$$X : \Omega \longrightarrow \mathbb{R}, \quad \omega \longmapsto X(\omega).$$

- ▶ Ω is the sample space of possible outcomes.
- ▶ On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, X must be measurable:

$$\{\omega : X(\omega) \leq x\} \in \mathcal{F} \quad \text{for every } x \in \mathbb{R}.$$

This ensures that probabilities such as $\mathbb{P}(X \leq x)$ are defined.

Example: toss a coin and set $X(H) = 1$, $X(T) = 0$.

1. Outcomes, events and numerical values

Consider two coin tosses. The sample space is

$$\Omega = \{HH, HT, TH, TT\}.$$

Let X count the number of heads:

Outcome ω	HH	HT	TH	TT
Value $X(\omega)$	2	1	1	0

- ▶ An **outcome** is one complete result, such as HT .
- ▶ An **event** is a set of outcomes, such as $\{HT, TH\}$.
- ▶ $X = 1$ describes this event: $\{\omega : X(\omega) = 1\} = \{HT, TH\}$.

For independent fair tosses, $\mathbb{P}(X = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

1. The probability space

The three parts of $(\Omega, \mathcal{F}, \mathbb{P})$ have different roles:

- ▶ Ω : all possible outcomes.
- ▶ $\mathcal{F}$: the collection of events to which we assign probabilities.
- ▶ $\mathbb{P}$: the rule assigning a number in $[0, 1]$ to each event.

For one fair coin toss,

$$\Omega = \{H, T\}, \quad \mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}.$$

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}, \quad \mathbb{P}(\Omega) = 1.$$

A sigma-algebra contains Ω and is closed under complements and countable unions. These properties let us combine events consistently.

1. What measurability means

The condition

$$\{\omega : X(\omega) \leq x\} \in \mathcal{F} \quad \text{for every } x \in \mathbb{R}$$

means that every threshold question about X corresponds to an allowed event.

For $X = \text{number of heads in two tosses}$,

$$\{X \leq 1\} = \{HT, TH, TT\}, \quad \mathbb{P}(X \leq 1) = \frac{3}{4}.$$

- ▶ X is a fixed rule. Randomness comes from the outcome supplied to it.
- ▶ In a finite sample space with all subsets in $\mathcal{F}$, every real-valued function is measurable.
- ▶ Checking all thresholds also ensures that inverse images of all Borel sets are events.

2. Probability distributions: discrete variables

The **distribution** of X specifies the probabilities of its possible values. For a discrete variable, its **probability mass function (PMF)** is

$$p(x) = \mathbb{P}(X = x), \quad p(x) \geq 0, \quad \sum_x p(x) = 1.$$

Example: for the score X on a fair six-sided die,

$$p(x) = \frac{1}{6} \quad (x = 1, \dots, 6), \quad p(x) = 0 \quad \text{otherwise.}$$

Hence

$$\mathbb{P}(X \leq 2) = p(1) + p(2) = \frac{1}{3}.$$

2. Probability distributions: densities

For a variable with a **probability density function (PDF)** p ,

$$p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x) dx = 1, \quad \mathbb{P}(a \leq X \leq b) = \int_a^b p(x) dx.$$

- ▶ Probability is the **area under the density** over an interval.
- ▶ A single point has probability $\mathbb{P}(X = x) = 0$.

Example: if $X \sim \text{Uniform}(0, 1)$, then

$$\mathbb{P}(0.2 \leq X \leq 0.5) = \int_{0.2}^{0.5} 1 dx = 0.3.$$

A general distribution can be described by its cumulative distribution function $F_X(x) = \mathbb{P}(X \leq x)$, whether or not it has a density.

2. A random variable and its distribution

For two independent fair coin tosses, let X count heads. Different outcomes may give the same value of X .

x	0	1	2
Outcomes	TT	HT, TH	HH
$p(x)$	$1/4$	$1/2$	$1/4$

$$\mathbb{P}(X \geq 1) = p(1) + p(2) = \frac{3}{4}.$$

General definition: the distribution of X assigns to each Borel set B the probability

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega : X(\omega) \in B\}).$$

The PMF records this information for a discrete variable.

2. Density height and probability

Let X be uniform on $[0, \frac{1}{2}]$. Its density is

$$p(x) = \begin{cases} 2, & 0 \leq x \leq \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

The total area is $2 \times \frac{1}{2} = 1$, so this is a valid density.

$$\mathbb{P}(0.1 \leq X \leq 0.3) = \int_{0.1}^{0.3} 2 \, dx = 0.4.$$

- ▶ A density value may exceed 1. Its integral gives probability.
- ▶ For a short interval near a point where p is continuous, $\mathbb{P}(x \leq X \leq x + h) \approx p(x)h$.
- ▶ An interval of zero width has zero area, hence $\mathbb{P}(X = x) = 0$.

2. The cumulative distribution function

The CDF gives the probability of being at or below a threshold:

$$F_X(x) = \mathbb{P}(X \leq x).$$

For the number of heads in two independent fair tosses,

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1/4, & 0 \leq x < 1, \\ 3/4, & 1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

For any distribution, $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$. If a density exists, $F_X(x) = \int_{-\infty}^x p(t) dt$.

3. Confidence intervals

A **confidence interval** uses sample data D to estimate an unknown parameter θ :

$$[L(D), U(D)].$$

An exact $100(1 - \alpha)\%$ confidence procedure satisfies

$$\mathbb{P}_{\theta}(L(D) \leq \theta \leq U(D)) = 1 - \alpha \quad \text{for every } \theta.$$

- ▶ The parameter θ is fixed. The endpoints vary across samples.
- ▶ A 95% procedure covers θ in 95% of repeated samples in the long run.
- ▶ Conservative procedures have coverage at least $1 - \alpha$.

3. Confidence intervals: a numerical example

Suppose $X_1, \dots, X_n$ are independent $N(\mu, \sigma^2)$ observations, with known standard deviation σ .

A 95% confidence interval for μ is approximately

$$\left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \quad \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right].$$

Example: $\bar{x} = 10$, $\sigma = 2$, $n = 100$ gives

$$10 \pm 1.96 \frac{2}{10} = 10 \pm 0.392, \quad [9.608, 10.392].$$

The 95% refers to the coverage of the **procedure** across repeated samples. Once we observe the data, this particular interval either contains μ or does not.

3. Why the interval contains a standard error

For independent observations with mean μ and variance σ^2 ,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \mathbb{E}[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

The **standard error** of $\bar{X}$ is its standard deviation:

$$\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

- ▶ σ describes variability of individual observations.
- ▶ $\sigma/\sqrt{n}$ describes variability of sample means across repeated samples.
- ▶ Multiplying the sample size by four halves the standard error.

3. Deriving the normal confidence interval

Under independent normal sampling with known σ ,

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

Approximately 95% of standard normal probability lies between -1.96 and 1.96 :

$$\mathbb{P}\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95.$$

Rearranging the inequalities gives

$$\mathbb{P}\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95.$$

The interval follows directly from the sampling distribution of $\bar{X}$.

3. Confidence level, width and interpretation

- ▶ **More confidence gives a wider interval.** A 99% normal interval uses approximately 2.576 instead of 1.96.
- ▶ **More data gives a narrower interval.** The width scales as $1/\sqrt{n}$ when σ is fixed.
- ▶ **Unknown variability must be estimated.** For independent normal data with unknown σ , use

$$\bar{X} \pm t_{n-1, 0.975} \frac{S}{\sqrt{n}}, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The confidence statement concerns repeated sampling under the model. It does not assign a 95% posterior probability to a fixed parameter lying in the observed interval.

4. Expectation

The **expectation** is a probability-weighted average of possible values.

For a variable with density p ,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x) dx.$$

For a discrete variable with PMF p ,

$$\mathbb{E}[X] = \sum_x xp(x).$$

Example: for a fair die,

$$\mathbb{E}[X] = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5.$$

These formulas give a finite mean when $\mathbb{E}[|X|] < \infty$. The mean need not be a possible observed value.

4. Expectation as a long-run average

Suppose a payoff X is 0 with probability 0.8 and 10 with probability 0.2.

$$\mathbb{E}[X] = 0(0.8) + 10(0.2) = 2.$$

Across many independent repetitions, the average payoff tends towards 2. An individual payoff is still either 0 or 10.

For independent identically distributed variables with $\mathbb{E}[|X_1|] < \infty$, the law of large numbers gives

$$\frac{1}{n} \sum_{i=1}^n X_i \longrightarrow \mathbb{E}[X_1] \quad \text{almost surely.}$$

This is the foundation of estimating expectations by simulation.

4. Expectations of functions

To average a function $g(X)$, weight its values by the distribution of X :

$$\mathbb{E}[g(X)] = \sum_x g(x)p(x) \quad \text{or} \quad \mathbb{E}[g(X)] = \int g(x)p(x) dx.$$

If $X \sim \text{Uniform}(0, 1)$,

$$\mathbb{E}[X] = \int_0^1 x dx = \frac{1}{2}, \quad \mathbb{E}[X^2] = \int_0^1 x^2 dx = \frac{1}{3}.$$

Thus $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2 = 1/4$.

Linearity: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ for integrable variables. Independence is unnecessary for this identity.

5. Variance

The **variance** measures the average squared distance from the mean $\mu = \mathbb{E}[X]$:

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2.$$

For a variable with density p ,

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx.$$

- ▶ Variance is nonnegative and has squared units.
- ▶ The standard deviation $\sigma = \sqrt{\text{Var}(X)}$ has the same units as X .

Example: if X equals 0 or 2 with equal probabilities, then $\mu = 1$ and $\text{Var}(X) = \frac{1}{2}(0 - 1)^2 + \frac{1}{2}(2 - 1)^2 = 1$.

5. Why we square deviations

Writing $\mu = \mathbb{E}[X]$, signed deviations always average to zero:

$$\mathbb{E}[X - \mu] = \mathbb{E}[X] - \mu = 0.$$

Squaring prevents positive and negative deviations from cancelling.

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mu)^2] \\ &= \mathbb{E}[X^2 - 2\mu X + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2.\end{aligned}$$

For $X \sim \text{Uniform}(0, 1)$, $\text{Var}(X) = 1/3 - 1/4 = 1/12$.

5. How transformations change variance

For constants a, b and finite variance,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b, \quad \text{Var}(aX + b) = a^2 \text{Var}(X).$$

- ▶ Adding b shifts every value equally and preserves spread.
- ▶ Multiplying by a multiplies distances by $|a|$ and squared distances by a^2 .

If $\mathbb{E}[X] = 1$ and $\text{Var}(X) = 4$, then $Y = 3X + 5$ satisfies

$$\mathbb{E}[Y] = 8, \quad \text{Var}(Y) = 36, \quad \text{SD}(Y) = 6.$$

6. Normal distribution with mean 1 and variance 4

For $X \sim N(\mu, \sigma^2)$, the density is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Substituting $\mu = 1$ and $\sigma^2 = 4$ gives

$$p(x) = \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{(x - 1)^2}{8}\right).$$

- ▶ The mean is $\mathbb{E}[X] = 1$.
- ▶ The variance is $\text{Var}(X) = 4$, so the standard deviation is $\sigma = 2$.
- ▶ Standardising gives $Z = (X - 1)/2 \sim N(0, 1)$.

6. Normal probabilities by standardisation

For $X \sim N(1, 4)$, subtract the mean and divide by the standard deviation:

$$Z = \frac{X - 1}{2} \sim N(0, 1).$$

For example,

$$\begin{aligned}\mathbb{P}(-1 \leq X \leq 3) &= \mathbb{P}\left(\frac{-1 - 1}{2} \leq Z \leq \frac{3 - 1}{2}\right) \\ &= \mathbb{P}(-1 \leq Z \leq 1) \\ &= \Phi(1) - \Phi(-1) \approx 0.6827.\end{aligned}$$

Here $\Phi(z) = \mathbb{P}(Z \leq z)$ is the standard normal CDF. The interval $[-1, 3]$ extends one standard deviation either side of the mean.

7. Independent events

Events A and B are **independent** if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

When $\mathbb{P}(B) > 0$, this is equivalent to

$$\mathbb{P}(A \mid B) = \mathbb{P}(A).$$

Knowing that B occurred leaves the probability of A unchanged.

Example: toss two independent fair coins. Let A be “the first coin is heads” and B be “the second coin is heads”.

$$\mathbb{P}(A \cap B) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2}.$$

7. Independence and conditional probability

Conditional probability restricts attention to outcomes in B :

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \mathbb{P}(B) > 0.$$

For a fair die, let $A = \{2, 4, 6\}$ and $B = \{4, 5, 6\}$.

$$\mathbb{P}(A) = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{1}{2}, \quad \mathbb{P}(A \cap B) = \frac{2}{6} = \frac{1}{3}.$$

Therefore

$$\mathbb{P}(A \mid B) = \frac{1/3}{1/2} = \frac{2}{3} \neq \frac{1}{2}.$$

These events are dependent: knowing the score exceeds 3 changes the chance of an even score.

7. Disjoint events and independent variables

Disjoint events cannot happen together: $A \cap B = \emptyset$. If both have positive probability, then

$$\mathbb{P}(A \cap B) = 0 < \mathbb{P}(A)\mathbb{P}(B),$$

so they are dependent. Knowing one occurred rules out the other.

Independent random variables satisfy

$$\mathbb{P}(X \in C, Y \in D) = \mathbb{P}(X \in C)\mathbb{P}(Y \in D)$$

for every pair of Borel sets C, D .

For discrete variables, this is equivalent to $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for every x, y .

8. Pearson correlation coefficient

For variables with finite, strictly positive variances,

$$\boxed{\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}}, \quad \sigma_X = \sqrt{\text{Var}(X)}, \quad \sigma_Y = \sqrt{\text{Var}(Y)}.$$

- ▶ $-1 \leq \rho_{X,Y} \leq 1$. Correlation has no units.
- ▶ A positive value indicates a positive linear association.
- ▶ A negative value indicates a negative linear association.
- ▶ A zero value indicates no linear association, but dependence may remain.

Example: if $Y = 2X + 3$ and $0 < \text{Var}(X) < \infty$, then $\rho_{X,Y} = 1$.

8. Why correlation divides by standard deviations

Standardise both variables:

$$Z_X = \frac{X - \mu_X}{\sigma_X}, \quad Z_Y = \frac{Y - \mu_Y}{\sigma_Y}.$$

Their average product is

$$\mathbb{E}[Z_X Z_Y] = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \rho_{X,Y}.$$

- ▶ Standardisation removes the original measurement units.
- ▶ Both standardised variables have mean 0 and variance 1.
- ▶ Cauchy–Schwarz gives $|\mathbb{E}[Z_X Z_Y]| \leq \sqrt{\mathbb{E}[Z_X^2] \mathbb{E}[Z_Y^2]} = 1$.

8. Zero correlation can coexist with dependence

Let X take values $-1, 0, 1$ with equal probabilities and set $Y = X^2$.

X	-1	0	1
Y	1	0	1
XY	-1	0	1

$$\mathbb{E}[X] = 0, \quad \mathbb{E}[Y] = \frac{2}{3}, \quad \mathbb{E}[XY] = 0.$$

Thus $\text{Cov}(X, Y) = 0$ and $\rho_{X,Y} = 0$ since both variances are positive.

Yet Y is completely determined by X . For example, $\mathbb{P}(Y = 0 \mid X = 0) = 1$, whereas $\mathbb{P}(Y = 0) = 1/3$.

9. Covariance

The **covariance** measures how two variables vary together:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Equivalently, when the second moments are finite,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- ▶ Positive covariance: deviations from the means tend to have the same sign.
- ▶ Negative covariance: deviations tend to have opposite signs.
- ▶ Independence implies zero covariance when the second moments are finite.

Example: if $Y = 2X + 3$, then $\text{Cov}(X, Y) = 2 \text{Var}(X)$.

9. A covariance calculation

Suppose (X, Y) equals $(0, 1)$ or $(2, 5)$ with equal probabilities. Then $\mu_X = 1$ and $\mu_Y = 3$.

X	Y	$X - 1$	$Y - 3$	$(X - 1)(Y - 3)$
0	1	-1	-2	2
2	5	1	2	2

$$\text{Cov}(X, Y) = \frac{1}{2}(2) + \frac{1}{2}(2) = 2.$$

Also $\text{Var}(X) = 1$ and $\text{Var}(Y) = 4$, so

$$\rho_{X,Y} = \frac{2}{1 \cdot 2} = 1.$$

This agrees with the exact linear relationship $Y = 2X + 1$.

9. Covariance in the variance of a sum

Expanding the squared deviation of $X + Y$ gives

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y).$$

- ▶ Positive covariance increases the variance of the sum.
- ▶ Negative covariance reduces it.
- ▶ Independent variables have zero covariance when their variances are finite.

For independent variables X_i with common variance σ^2 ,

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}.$$

This explains the standard error used in the confidence interval.

10. Bayes' formula

For events A and B with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Here the displayed conditional $\mathbb{P}(B \mid A)$ also assumes $\mathbb{P}(A) > 0$.

Example: choose a fair coin or a double-headed coin, each with probability $1/2$. After one toss gives heads, what is the probability of having chosen the double-headed coin?

$$\mathbb{P}(H) = 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}, \quad \mathbb{P}(D \mid H) = \frac{1 \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}.$$

The observed head increases the probability of D from $1/2$ to $2/3$.

10. Deriving Bayes' formula

The same joint probability can be written in two ways:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A).$$

Dividing by $\mathbb{P}(B) > 0$ yields Bayes' formula.

The denominator follows from splitting the sample space into A and A^c :

$$\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\mathbb{P}(A^c).$$

- ▶ $\mathbb{P}(A)$: probability before observing B (prior).
- ▶ $\mathbb{P}(B \mid A)$: probability of the observation under A .
- ▶ $\mathbb{P}(A \mid B)$: updated probability after observing B (posterior).

Assume the conditioning events have positive probability.

10. The coin example through expected counts

Imagine 400 repetitions of choosing a coin and tossing it once. The expected counts are:

Chosen coin	Heads	Tails	Total
Double-headed	200	0	200
Fair	100	100	200
Total	300	100	400

Among the 300 expected heads, 200 come from the double-headed coin:

$$\mathbb{P}(D \mid H) = \frac{200}{300} = \frac{2}{3}.$$

These are expected counts under the model. Actual counts vary across repetitions.

10. Updating after two heads

Choose the coin once, then toss it twice. Tosses are independent conditional on the chosen coin.

$$\mathbb{P}(HH \mid D) = 1, \quad \mathbb{P}(HH \mid F) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

With equal prior probabilities,

$$\mathbb{P}(D \mid HH) = \frac{1 \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2}} = \frac{4}{5}.$$

The posterior probability rises from $1/2$ before any toss to $2/3$ after one head and $4/5$ after two heads.

Both models allow heads, so even two heads leave some probability on the fair coin.