Introduction to Probability

Computer Science Tripos – Part IA, Paper 1

Comprehensive Tripos Revision Guide

University of Cambridge – Department of Computer Science and Technology
Based on Lectures 1–12 (Jamnik & Sauerwald) and Tripos Papers 2020–2025

Block	Lectures	Key Skills
Foundations	1: Bayes & Independence	Axioms, conditioning, Bayes' theorem, total probability
Random Variables	2–5: RVs, Variance, Named Distributions	PMF/PDF/CDF, $\mathbb{E}[X]$, $\mathrm{Var}(X)$, discrete & continuous laws
Multivariate	6–7: Joints, Marginals, Covariance	Joint/marginal laws, independence, Cov, Corr
Limit Theorems	8–9: Weak Law, CLT	Markov/Chebyshev, convergence, normal approximation
Statistics	10–11: Estimators	Bias, MSE, unbiased estimators, Jensen's inequality
Applications	12: Online Algorithms	Optimal stopping, secretary problem

0 Contents

1 Foundations: Axioms, Conditioning & Bayes' Theorem

1.1 Probability Axioms

Kolmogorov's Axioms

Let S be the sample space and $\mathcal{F}$ a collection of events. A probability function P satisfies:
1. $P(S) = 1$.
2. $0 \le P(A) \le 1$ for all events A.
3. **Countable additivity**: if $A_1, A_2, \ldots$ are pairwise disjoint, then $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$.

Immediate consequences: $P(\emptyset) = 0$; $P(A^c) = 1 - P(A)$; if $A \subseteq B$ then $P(A) \le P(B)$.

Inclusion–Exclusion

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

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Boole's inequality (union bound): $P\left(\bigcup_i A_i\right) \le \sum_i P(A_i)$.

1.2 Conditional Probability & Independence

Conditional Probability

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

Multiplication rule: $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$.
Chain rule: $P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \cdots P(A_n \mid A_1, \ldots, A_{n-1})$.

Independence

Events A, B are **independent** iff

$$P(A \cap B) = P(A)P(B) \quad \Longleftrightarrow \quad P(A \mid B) = P(A) \text{ (for } P(B) > 0).$$

$A_1, \ldots, A_n$ are **mutually independent** iff for *every* subset $I \subseteq \{1, \ldots, n\}$,

$$P\Big(\bigcap_{i \in I} A_i\Big) = \prod_{i \in I} P(A_i).$$

Trap: Pairwise $\neq$ Mutual Independence

Pairwise independence (every pair satisfies $P(A_i \cap A_j) = P(A_i)P(A_j)$) does *not* imply mutual independence. Always check the full product condition when asked to justify mutual independence, and always state *which* definition you are verifying.

1.3 Law of Total Probability & Bayes' Theorem

Law of Total Probability

If $\{F_1, \ldots, F_n\}$ is a **partition** of S (i.e. disjoint, exhaustive, $P(F_i) > 0$), then for any event E:

$$P(E) = \sum_{i=1}^{n} P(E \mid F_i)\, P(F_i).$$

Bayes' Theorem

$$P(F_j \mid E) = \frac{P(E \mid F_j)\, P(F_j)}{P(E)} = \frac{P(E \mid F_j)\, P(F_j)}{\sum_i P(E \mid F_i)\, P(F_i)}.$$

In words: posterior $\propto$ likelihood $\times$ prior.

Worked Example: Diagnostic Testing

A disease has prevalence $P(D) = 0.01$. A test has sensitivity $P(+ \mid D) = 0.99$ and false-positive rate $P(+ \mid D^c) = 0.05$. Find $P(D \mid +)$.

Solution. By total probability,

$$P(+) = P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c) = 0.99(0.01) + 0.05(0.99) = 0.0099 + 0.0495 = 0.0594.$$

By Bayes,

$$P(D \mid +) = \frac{0.0099}{0.0594} \approx 0.1667.$$

Insight: even with a 99% sensitive test, a low base rate keeps the posterior low – this is the classic base-rate trap examiners love to test.

Exam Technique for Bayes Questions

- Always draw (or imagine) a tree diagram first: it makes the partition $\{F_i\}$ explicit.
- Compute $P(E)$ via total probability *before* substituting into Bayes – do not try to do it in one line.
- When a question gives conditionals "the wrong way round" (e.g. $P(E \mid F)$ given, $P(F \mid E)$ wanted), that is your cue: use Bayes.

1.4 Useful Combinatorics Recap

Counting

Permutations of n distinct objects: $n!$. k **from** n**, ordered**: $\dfrac{n!}{(n-k)!}$.

Combinations (k from n, unordered): $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$.

Binomial theorem: $(x+y)^n = \displaystyle\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$.

2 Random Variables & Expectation

2.1 Discrete Random Variables

PMF and CDF

A discrete random variable X takes values in a countable set. Its **probability mass function** (PMF) is

$$p_X(a) = P[X = a], \qquad \sum_a p_X(a) = 1.$$

Its **cumulative distribution function** (CDF) is $F_X(a) = P[X \le a] = \sum_{x \le a} p_X(x)$, which is right-continuous, non-decreasing, $F_X(-\infty) = 0$, $F_X(\infty) = 1$.

Expectation

$$\mathbb{E}[X] = \sum_a a\, p_X(a) \qquad \text{(if the sum converges absolutely).}$$

Law of the Unconscious Statistician (LOTUS): for any function g,

$$\mathbb{E}[g(X)] = \sum_a g(a)\, p_X(a).$$

Linearity of Expectation

For any random variables X, Y (*not* necessarily independent) and constants $a, b \in \mathbb{R}$:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \qquad \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Indicator Variables – the Single Most Useful Trick

For an event A, define $\mathbb{1}_A = 1$ if A occurs, 0 otherwise. Then $\mathbb{E}[\mathbb{1}_A] = P(A)$.
Method: to find $\mathbb{E}$[count of something], write the count as a sum of indicators $\sum_i \mathbb{1}_{A_i}$ and apply linearity – this works *even when the A_i are dependent.*
Classic example (fixed points of a random permutation): let X = number of fixed points of a uniformly random permutation of $\{1, \ldots, n\}$. Let $\mathbb{1}_i = \mathbb{1}[\pi(i) = i]$, so $P(\mathbb{1}_i = 1) = 1/n$ and

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}_i] = n \cdot \frac{1}{n} = 1,$$

regardless of n – despite the $\mathbb{1}_i$ being dependent.

2.2 Functions of Random Variables

Trap: $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$ in general

Only true when g is *linear*. For convex g, Jensen's inequality gives $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$ (see Section 10). E.g. $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ always (equivalent to $\mathrm{Var}(X) \geq 0$).

Worked Example: Coin Tosses – First Head

Toss a fair coin repeatedly. Let X be the toss number of the first head. Then $P[X = k] = (1/2)^k$ for $k \geq 1$ (this is Geom(1/2)), and

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \left(\frac{1}{2}\right)^k = \frac{1/2}{(1 - 1/2)^2} = 2.$$

Alternative (first-step) method: condition on the first toss, $\mathbb{E}[X] = 1 + \frac{1}{2}\mathbb{E}[X] \Rightarrow \mathbb{E}[X] = 2$. This first-step decomposition generalises to many recursive expectation problems.

2.3 Independence of Random Variables

Independent Random Variables

X, Y are independent iff $P[X \leq a, Y \leq b] = P[X \leq a]P[Y \leq b]$ for all a, b; equivalently (discrete case) $p_{X,Y}(a, b) = p_X(a)p_Y(b)$ for all a, b.
If X, Y independent: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, and $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.

3 Variance & Discrete Distributions

3.1 Variance

Variance and Standard Deviation

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \qquad \sigma_X = \sqrt{\mathrm{Var}(X)}.$$

Scaling: $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$ (shifts do not affect spread; scaling by a scales variance by a^2).
Sums: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$; if X, Y independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
More generally, for independent $X_1, \ldots, X_n$: $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i)$.

Trap #1: Variance does not add for dependent variables

$\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$ unless $\mathrm{Cov}(X,Y) = 0$ (e.g. X, Y independent). Always check independence before dropping the covariance term.

3.2 Key Discrete Distributions

Distribution	PMF $P(X=k)$	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	Typical use
$\mathrm{Bern}(p)$	$p^k(1-p)^{1-k}$, $k \in \{0,1\}$	p	$p(1-p)$	Single yes/no trial
$\mathrm{Bin}(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$, $0 \leq k \leq n$	np	$np(1-p)$	# successes in n i.i.d. Bernoulli trials
Discrete $\mathrm{Unif}(a,b)$	$\frac{1}{b-a+1}$, $k = a, \ldots, b$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	Equally likely integers
$\mathrm{Hyp}(N,n,m)$	$\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$	$n\frac{m}{N}$	$n\frac{m}{N}\left(1-\frac{m}{N}\right)\frac{N-n}{N-1}$	# successes drawing n *without replacement* from N (m "successes")

Sum of Independent Binomials

If $X \sim \mathrm{Bin}(n,p)$, $Y \sim \mathrm{Bin}(m,p)$ independent (*same* p), then $X+Y \sim \mathrm{Bin}(n+m,p)$.

Binomial → Sum of Indicators

$X \sim \mathrm{Bin}(n,p)$ can always be written $X = \sum_{i=1}^{n} X_i$ with $X_i \overset{\text{iid}}{\sim} \mathrm{Bern}(p)$. This immediately gives $\mathbb{E}[X] = np$ and, by independence, $\mathrm{Var}(X) = np(1-p)$ without touching the PMF sum.

Worked Example: Sampling Without Replacement (Hypergeometric via Indicators)

An urn has N balls, m red. Draw n without replacement; let $X_i = \mathbb{1}[i\text{-th ball red}]$ and $X = \sum X_i$.
By symmetry $P(X_i = 1) = m/N$ for every i (each position is equally likely to be any ball), so $\mathbb{E}[X] = n\,m/N$ – *identical* to the with-replacement (Binomial) mean, even though the X_i are *not* independent (drawing a red ball lowers the chance of the next being red). The dependence only shows up in $\mathrm{Var}(X)$, which is smaller than the binomial variance $np(1-p)$ by the finite-population correction factor $\frac{N-n}{N-1}$.

Trap #2: With vs. Without Replacement

"Without replacement, small sample relative to population" ⇒ Hypergeometric, *not* Binomial. If $n \ll N$, Hypergeometric is well-approximated by Binomial$(n, m/N)$ – state this approximation explicitly if asked to justify it.

4 Poisson & Geometric Random Variables

4.1 Poisson Distribution

Poisson Distribution

$X \sim \mathrm{Pois}(\lambda)$ models the number of events in a fixed interval when events occur independently at a constant average rate λ.

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0,1,2,\ldots, \qquad \mathbb{E}[X] = \mathrm{Var}(X) = \lambda.$$

Sum: if $X \sim \mathrm{Pois}(\lambda_1)$, $Y \sim \mathrm{Pois}(\lambda_2)$ independent, then $X+Y \sim \mathrm{Pois}(\lambda_1+\lambda_2)$.

Poisson as a Limit of Binomial (Poisson Approximation)

If $X_n \sim \text{Bin}(n, p_n)$ with $np_n \to \lambda$ as $n \to \infty$ (i.e. n large, p small, $np = \lambda$ moderate), then

$$P(X_n = k) \to \frac{e^{-\lambda}\lambda^k}{k!}.$$

Rule of thumb for exam use: $n \geq 20$ and $p \leq 0.05$ (or $n \geq 100$, $np \leq 10$) – use $\text{Pois}(np)$ to approximate $\text{Bin}(n, p)$.

4.2 Geometric Distribution

Geometric Distribution

$X \sim \text{Geom}(p)$ = number of *trials up to and including* the first success in i.i.d. $\text{Bern}(p)$ trials.

$$P(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots, \qquad \mathbb{E}[X] = \frac{1}{p}, \qquad \text{Var}(X) = \frac{1-p}{p^2}.$$

Tail: $P(X > k) = (1-p)^k$.

Memoryless Property

Geometric (and, in continuous time, Exponential) is the *only* memoryless discrete (continuous) distribution:

$$P(X > s + t \mid X > s) = P(X > t), \qquad \forall s, t \geq 0.$$

Interpretation: past waiting time gives no information about future waiting time – "the process has no memory".

Worked Example: Memoryless Property

Waiting times between buses are Geom(1/2) on $\{1, 2, \ldots\}$. Given the waiting time is at least 2, find $P[\text{waiting time} \geq 7]$.

Solution. By memorylessness, $P(X \geq 7 \mid X \geq 2) = P(X \geq 7 - 1 \mid X \geq 1)$. More directly: $P(X \geq s + t \mid X \geq s + 1) = P(X \geq t + 1)$ using the discrete shift; concretely here

$$P(X \geq 7 \mid X \geq 2) = P(X \geq 6) = (1-p)^5 = (1/2)^5 = \tfrac{1}{32}.$$

(Always re-derive directly from $P(X \geq k) = (1-p)^{k-1}$ rather than quoting memorylessness blindly, to avoid off-by-one slips.)

4.3 Negative Binomial Distribution

Negative Binomial

$X \sim \text{NegBin}(r, p)$ = number of trials up to and including the r-th success.

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \ldots, \qquad \mathbb{E}[X] = \frac{r}{p}, \qquad \text{Var}(X) = \frac{r(1-p)}{p^2}.$$

Sum of Geometrics: $X = \sum_{i=1}^{r} G_i$ where $G_i \overset{\text{iid}}{\sim} \text{Geom}(p)$ are the gaps between consecutive successes – immediately gives $\mathbb{E}[X] = r/p$, $\text{Var}(X) = r(1-p)/p^2$.

Choosing the Right Discrete Distribution – Quick Diagnostic

- **Fixed number of trials, count successes** → Binomial.
- **Trials until 1st success** → Geometric. **Trials until r-th success** → Negative Binomial.
- **Rate per unit time/space, count of rare events** → Poisson.
- **Fixed sample *without replacement* from a finite population** → Hypergeometric.

5 Continuous Distributions

5.1 Density Functions

PDF, CDF, Expectation

X is continuous with density $f_X \geq 0$ if $P[a \leq X \leq b] = \int_a^b f_X(x)\,\mathrm{d}x$ and $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\,\mathrm{d}x, \qquad f_X(x) = F_X'(x) \text{ (where differentiable).}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x, \qquad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x \quad \text{(LOTUS)}.$$

Trap: $f_X(a) \neq P[X = a]$

For continuous X, $P[X = a] = 0$ for every single point a. The density $f_X(a)$ is *not* a probability – it can even exceed 1. Only $\int f_X$ over an interval gives a probability.

Recovering f_X from a piecewise CDF

Given F_X defined piecewise, differentiate *each piece* to get f_X on the corresponding interval; check continuity of F_X at the boundaries (density need not be continuous, but F_X must be) and that f_X integrates to 1 as a sanity check.

5.2 Key Continuous Distributions

Distribution	**PDF** $f_X(x)$	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	**Notes**
$\mathrm{Unif}(a,b)$	$\frac{1}{b-a}$, $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	Constant density
$\mathrm{Exp}(\lambda)$	$\lambda e^{-\lambda x}$, $x \geq 0$	$1/\lambda$	$1/\lambda^2$	Memoryless; CDF $1 - e^{-\lambda x}$
$N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2	Bell curve; symmetric about μ

Exponential: Memoryless & Poisson-Process Link

$P[X > t] = e^{-\lambda t}$, and $P[X > s + t \mid X > s] = P[X > t]$ (memoryless).
If events occur as a Poisson process of rate λ, the *gaps between consecutive events* are i.i.d. $\mathrm{Exp}(\lambda)$, and the number of events in an interval of length t is $\mathrm{Pois}(\lambda t)$. This is the continuous analogue of Geometric $\leftrightarrow$ Poisson.

Standard Normal & Standardisation

$Z = \dfrac{X - \mu}{\sigma} \sim N(0,1)$ whenever $X \sim N(\mu, \sigma^2)$. Write $\Phi(z) = P[Z \leq z]$ for the standard normal CDF (tabulated).

$$P[X \leq a] = \Phi\left(\frac{a - \mu}{\sigma}\right), \qquad \Phi(-z) = 1 - \Phi(z) \text{ (symmetry)}.$$

Sums of independent normals: if $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ independent, $aX + bY \sim N(a\mu_1 + b\mu_2,\ a^2\sigma_1^2 + b^2\sigma_2^2)$.

5.3 Transformations of Continuous Random Variables

Change of Variables (Monotone g)

If $Y = g(X)$ with g strictly monotone and differentiable, and g^{-1} exists:

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|.$$

Method (safer for exams): find $F_Y(y) = P[Y \le y]$ directly by manipulating the inequality (careful with direction flips if g is decreasing), then differentiate.

Worked Example: Order Statistics – Min of Uniforms

$X, Y \overset{\text{iid}}{\sim} \text{Unif}(0, 1/2)$, $Z = \min(X, Y)$. Find F_Z and $\mathbb{E}[Z]$.

Solution. $P[Z > z] = P[X > z]P[Y > z]$ (independence) $= \left(\dfrac{1/2 - z}{1/2}\right)^2$ for $0 \le z \le 1/2$, so

$$F_Z(z) = 1 - (1 - 2z)^2, \qquad f_Z(z) = 4(1 - 2z), \quad 0 \le z \le 1/2.$$

$$\mathbb{E}[Z] = \int_0^{1/2} z \cdot 4(1 - 2z)\,\mathrm{d}z = \frac{1}{6}.$$

General pattern: for n i.i.d. continuous X_i, $P[\min > z] = (P[X_1 > z])^n$ and $P[\max \le z] = (P[X_1 \le z])^n$ – memorise this trick for *any* min/max question.

Markov's & Chebyshev's Inequalities (Continuous or Discrete)

See Section 8 for full statements – these apply to any non-negative / any random variable with finite mean and variance, and are frequently combined with continuous distributions in exam questions (e.g. bounding $P[X \le c]$ for a Normal without computing Φ exactly).

Trap: Forgetting the Support

Always state the support of $f_{X,Y}$ or f_X explicitly (e.g. $0 < y < x < 1$) and set the density to 0 outside it – integration limits in joint-density questions depend entirely on getting this region right.

6 Joint Distributions & Marginals

6.1 Discrete Joint Distributions

Joint PMF, CDF and Marginals

$$p_{X,Y}(a, b) = P[X = a, Y = b], \qquad F_{X,Y}(a, b) = P[X \le a, Y \le b].$$

Marginal PMF of X: sum out Y,

$$p_X(a) = \sum_b p_{X,Y}(a, b), \qquad F_X(a) = \lim_{b\to\infty} F_{X,Y}(a, b).$$

Trap: The Joint Distribution Contains *More* Information Than Both Marginals

Two different joint distributions can have identical marginals (classic example: sampling without replacement gives X_i's with the same marginal as with replacement, but a totally different joint distribution – the X_i are dependent). **You cannot reconstruct the joint from the marginals alone unless you are told the variables are independent.**

6.2 Continuous Joint Distributions

Joint Density

X, Y jointly continuous with density $f_{X,Y} \geq 0$ if

$$P[a_1 \leq X \leq b_1,\ a_2 \leq Y \leq b_2] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x,y)\,\mathrm{d}y\,\mathrm{d}x, \qquad \iint_{\mathbb{R}^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1.$$

Marginal density: $f_X(x) = \displaystyle\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y.$

Relation to CDF: $f(x,y) = \dfrac{\partial^2}{\partial x\,\partial y} F(x,y).$

Independence via Factorisation

X, Y independent $\iff$ $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all x, y (equivalently $p_{X,Y}(a,b) = p_X(a) p_Y(b)$ discretely) $\iff$ the joint density/PMF **factorises** *and* the support is a rectangle (product set) – both conditions matter!

Trap: Factorisation Alone Is Not Enough

If the support of $f_{X,Y}$ is *not* a rectangle (e.g. $0 < y < x < 1$), then X, Y are *automatically dependent*, even if the formula for $f(x,y)$ looks like it factorises as a product of functions of x and y separately – because the range of Y depends on X. Always check the support first.

Worked Example: Non-Rectangular Support

$f(x,y) = cx$ for $0 < y < x < 1$. Find c, the marginals, and determine independence.

Solution. Normalise: $\displaystyle\int_0^1 \int_0^x cx\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 cx^2\,\mathrm{d}x = c/3 = 1 \Rightarrow c = 3.$

Marginals: $f_X(x) = \displaystyle\int_0^x 3x\,\mathrm{d}y = 3x^2,\ 0 < x < 1; \quad f_Y(y) = \int_y^1 3x\,\mathrm{d}x = \tfrac{3}{2}(1 - y^2),\ 0 < y < 1.$

Independence: $f_X(x) f_Y(y) \neq f(x,y) = 3x$ (and the support is triangular, not rectangular) $\Rightarrow$ **not independent**.

6.3 Conditional Distributions

Conditional PMF / Density

Discrete: $p_{X|Y}(a \mid b) = \dfrac{p_{X,Y}(a,b)}{p_Y(b)}$, provided $p_Y(b) > 0$.

Continuous: $f_{X|Y}(x \mid y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$, provided $f_Y(y) > 0$.

Conditional expectation: $\mathbb{E}[X \mid Y = y] = \displaystyle\sum_x x\, p_{X|Y}(x \mid y)$ or $\displaystyle\int x\, f_{X|Y}(x \mid y)\,\mathrm{d}x.$

Law of Total Expectation (Tower Property)

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \sum_y \mathbb{E}[X \mid Y = y]\, P(Y = y).$$

This is the multivariate analogue of the Law of Total Probability, and often the fastest route to $\mathbb{E}[X]$ when a natural conditioning variable exists (e.g. first-step analysis).

6.4 Sums of Random Variables (Convolution)

Distribution of a Sum

If X, Y independent discrete, $P[X+Y=n] = \sum_k p_X(k)\, p_Y(n-k)$.

If X, Y independent continuous, $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x$ (convolution).

7 Covariance & Correlation

Covariance

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

Properties:

- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ (symmetric).
- **Bilinearity:** $\mathrm{Cov}(aX+b,\, cY+d) = ac\,\mathrm{Cov}(X,Y)$.
- $\mathrm{Cov}\left(\sum_i X_i, \sum_j Y_j\right) = \sum_i \sum_j \mathrm{Cov}(X_i, Y_j)$.
- If X, Y independent $\Rightarrow \mathrm{Cov}(X,Y) = 0$ (**converse is false** – see trap below).

Trap: Zero Covariance Does Not Imply Independence

$\mathrm{Cov}(X,Y) = 0$ only rules out *linear* dependence. A classic counterexample: $X \sim \mathrm{Unif}(-1,1)$, $Y = X^2$. Then $\mathrm{Cov}(X,Y) = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0 - 0 = 0$, but Y is a deterministic (hence maximally dependent) function of X. Always state "uncorrelated" rather than "independent" unless you have verified factorisation of the joint law.

Variance of a Sum (General)

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y), \qquad \mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y).$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Correlation Coefficient

$$\rho(X,Y) = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \in [-1,1].$$

$|\rho| = 1 \iff Y = aX + b$ exactly (perfect linear relationship), sign of a matches sign of ρ. This bound follows from the **Cauchy–Schwarz inequality**: $(\mathrm{Cov}(X,Y))^2 \le \mathrm{Var}(X)\mathrm{Var}(Y)$.

Worked Example: Covariance from a Joint Table

$X, Y \in \{-1, 0, 1\}$ with joint table (partially given, filled using marginals summing to 1 across rows/columns). Suppose the completed table gives $P(X=-1, Y=-1) = \frac{1}{4}$, $P(X=1, Y=1) = \frac{1}{4}$, $P(X=0, Y=0) = \frac{1}{2}$, all other cells 0.
$\mathbb{E}[X] = \mathbb{E}[Y] = 0$ by symmetry. $\mathbb{E}[XY] = (-1)(-1)\frac{1}{4} + (1)(1)\frac{1}{4} + 0 = \frac{1}{2}$.

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \tfrac{1}{2} - 0 = \tfrac{1}{2}.$$

Since $P(X=1, Y=-1) = 0 \neq P(X=1)P(Y=-1)$, X and Y are **not independent** – consistent with the positive covariance.

Worked Example: Sampling Without Replacement

Urn with balls labelled $0, \ldots, n-1$; draw all n without replacement, X_i = label drawn at step i. Find $\mathrm{Cov}(X_1, X_2)$.

Solution. By symmetry each X_i is $\mathrm{Unif}\{0, \ldots, n-1\}$, so $\mathbb{E}[X_i] = \frac{n-1}{2}$, $\mathrm{Var}(X_i) = \frac{n^2-1}{12}$.
Since $X_1 + \cdots + X_n = 0 + 1 + \cdots + (n-1)$ is a *constant*, $\mathrm{Var}(\sum X_i) = 0$. Expanding,

$$0 = \sum_i \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) = n\mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1, X_2)$$

(using symmetry, all pairwise covariances equal), so

$$\mathrm{Cov}(X_1, X_2) = -\frac{\mathrm{Var}(X_1)}{n-1} = -\frac{n+1}{12}.$$

Key takeaway: whenever a sum of random variables is a known *constant*, set $\mathrm{Var}(\text{sum}) = 0$ and expand – a very efficient way to extract covariances by symmetry.

8 Tail Bounds & The Weak Law of Large Numbers

8.1 Markov's Inequality

Markov's Inequality

If $X \geq 0$ and $a > 0$, then

$$P[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

Requires only that X is non-negative and $\mathbb{E}[X]$ exists – no variance needed. Usually a *weak* but universally applicable bound.

8.2 Chebyshev's Inequality

Chebyshev's Inequality

For any random variable X with finite mean μ and variance σ^2, and any $k > 0$:

$$P[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}.$$

Derivation: apply Markov's inequality to the non-negative variable $(X-\mu)^2$ with $a = k^2$:

$$P[(X-\mu)^2 \geq k^2] \leq \frac{\mathbb{E}[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}.$$

Two-sided & distribution-free: unlike Markov, it needs no non-negativity assumption, and applies to *any* distribution with finite variance – generally gives a tighter bound than Markov when both are applicable.

Choosing Which Bound to Use

- Only $\mathbb{E}[X]$ known, $X \geq 0 \to$ Markov.
- $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ both known $\to$ Chebyshev (generally tighter).
- Sum of many i.i.d. variables and an *exact* (asymptotic) probability wanted $\to$ CLT (Section 9) – gives the tightest answer but only asymptotically valid.

A question that gives you a variance and asks you to "bound" (not "approximate") a probability is almost always asking for Chebyshev.

Worked Example: Chebyshev Bound

$F \sim N(10, 2)$ models a female elephant's weight (tonnes). Bound $P[F \leq 5]$ using Chebyshev.

Solution. $P[F \leq 5] = P[F - 10 \leq -5] \leq P[|F-10| \geq 5] \leq \frac{\sigma^2}{k^2} = \frac{2}{25} = 0.08$.

(Note Chebyshev only needs mean/variance, not normality – so this bound would hold for *any* distribution with this mean and variance, not just Normal.)

8.3 Weak Law of Large Numbers

Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots$ be i.i.d. with finite mean μ and finite variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then for every $\varepsilon > 0$:

$$P\big[\,|\overline{X}_n - \mu| \geq \varepsilon\,\big] \to 0 \quad \text{as } n \to \infty.$$

We say $\overline{X}_n \xrightarrow{P} \mu$ (**convergence in probability**).

Proof Sketch (via Chebyshev)

$\mathbb{E}[\overline{X}_n] = \mu$ and $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ (independence). By Chebyshev,

$$P\big[|\overline{X}_n - \mu| \geq \varepsilon\big] \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty.$$

This is the standard *examinable* proof – know it, it is a favourite short-answer / "prove that" question.

Trap: WLLN Gives No Rate / Says Nothing About a Single Trial

WLLN only tells you $\overline{X}_n$ concentrates around μ *as* $n \to \infty$; it gives no information about how large n needs to be without also using Chebyshev/CLT to get an explicit bound. Do not confuse "the sample mean converges" with "every individual outcome converges" (that would be a Strong Law statement, not required here).

Worked Example: Container Loading (Markov vs Chebyshev vs CLT)

$X_1, \ldots, X_5 \overset{\text{iid}}{\sim} \mathrm{Exp}(1/2)$ (packet weights), $X = \sum X_i$. Find w such that $P[X \leq w] \geq 0.95$.
Setup: $\mathbb{E}[X] = 5 \cdot 2 = 10$, $\mathrm{Var}(X) = 5 \cdot 4 = 20$.
Markov (on $X \geq 0$): want $P[X > w] \leq \mathbb{E}[X]/w \leq 0.05 \Rightarrow w \geq 10/0.05 = 200$.
Chebyshev: want $P[|X - 10| \geq w - 10] \leq \sigma^2/(w-10)^2 \leq 0.05 \Rightarrow (w-10)^2 \geq 20/0.05 = 400 \Rightarrow w \geq 30$.
CLT (Section 9): $X \approx N(10, 20)$, so want $\Phi\left(\dfrac{w-10}{\sqrt{20}}\right) \geq 0.95 \Rightarrow \dfrac{w-10}{\sqrt{20}} \geq 1.645 \Rightarrow w \approx 17.4$.
Takeaway: the three methods give increasingly tight (and increasingly assumption-heavy) bounds: Markov $\geq$ Chebyshev $\geq$ CLT-approximation. Tripos questions often ask for all three back-to-back precisely to test this ordering.

9 The Central Limit Theorem

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots$ be i.i.d. with finite mean μ and finite variance $\sigma^2 > 0$. Then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty,$$

equivalently, writing $S_n = \sum_{i=1}^n X_i$,

$$S_n \mathrel{\dot\sim} N(n\mu,\ n\sigma^2) \quad \text{for large } n.$$

Remarkable feature: holds regardless of the shape of the X_i's distribution (as long as mean/variance are finite) – this is why the Normal distribution is ubiquitous.

Standard CLT Exam Recipe

1. Identify S_n (or $\overline{X}_n$) as a sum/average of i.i.d. terms; state $\mu = \mathbb{E}[X_i]$, $\sigma^2 = \mathrm{Var}(X_i)$.
2. Approximate $S_n \sim N(n\mu, n\sigma^2)$ (state you are invoking the CLT).
3. Standardise: $Z = \dfrac{S_n - n\mu}{\sqrt{n}\sigma}$, and express the target probability using Φ.
4. If S_n is a *count* of a discrete variable (e.g. Binomial), apply a **continuity correction** (± 0.5) for a more accurate approximation, if requested.

Normal Approximation to Binomial & Poisson

$X \sim \mathrm{Bin}(n,p)$, for n large (rule of thumb $np \geq 5$ and $n(1-p) \geq 5$):

$$X \mathrel{\dot\sim} N\big(np,\ np(1-p)\big).$$

$X \sim \mathrm{Pois}(\lambda)$, for λ large: $X \mathrel{\dot\sim} N(\lambda, \lambda)$ (since Poisson is itself a sum of i.i.d. contributions in disjoint sub-intervals).

Worked Example: Normal Approximation with Continuity Correction

$H \sim \mathrm{Bin}(300, 0.8)$. Approximate $P[220 < H < 260]$.
Solution. $\mathbb{E}[H] = 240$, $\mathrm{Var}(H) = 300(0.8)(0.2) = 48$, $\sigma = \sqrt{48} \approx 6.93$.
With continuity correction, $P[220 < H < 260] \approx P[220.5 \leq H \leq 259.5]$

$$= \Phi\left(\frac{259.5 - 240}{6.93}\right) - \Phi\left(\frac{220.5 - 240}{6.93}\right) = \Phi(2.81) - \Phi(-2.81) = 2\Phi(2.81) - 1.$$

(Continuity correction is optional unless explicitly requested, but always mention you *could* apply it – shows examiners you know the subtlety of approximating a discrete distribution with a continuous one.)

Trap: CLT vs. Exact Distribution vs. Poisson Approximation

- np moderate, n large, p small $\rightarrow$ **Poisson** approximation is usually preferred (better in the tails).
- np and $n(1-p)$ both reasonably large (≥ 5) $\rightarrow$ **Normal** approximation via CLT.
- Exam questions often explicitly ask you to justify *why* an approximation is valid – always cite the relevant condition (n large & p small for Poisson; $np, n(1-p)$ large for Normal) rather than just stating the approximating distribution.

10 Estimators: Bias, Variance & Mean Squared Error

10.1 Basic Definitions

Estimator, Bias, Unbiasedness

Given i.i.d. samples $X_1, \ldots, X_n$ from a distribution with unknown parameter θ, an **estimator** $\hat\theta = \hat\theta(X_1, \ldots, X_n)$ is any function of the data used to estimate θ (itself a random variable).

$$\mathrm{Bias}(\hat\theta) = \mathbb{E}[\hat\theta] - \theta.$$

$\hat\theta$ is **unbiased** if $\mathrm{Bias}(\hat\theta) = 0$, i.e. $\mathbb{E}[\hat\theta] = \theta$ for *every* value of θ.

Sample Mean & Sample Variance

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad E[\overline{X}_n] = \mu, \quad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \quad (\overline{X}_n \text{ unbiased for } \mu).$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{is unbiased for } \sigma^2 : \ \mathbb{E}[S^2] = \sigma^2.$$

Trap: The "Obvious" Variance Estimator Is Biased

$\frac{1}{n}\sum(X_i - \overline{X}_n)^2$ (dividing by n, not $n-1$) is *biased* – it underestimates σ^2 on average, because $\overline{X}_n$ is itself estimated from the same data ("uses up a degree of freedom"). Always use $n-1$ for an unbiased variance estimator, and be ready to prove this via $\mathbb{E}[(X_i - \overline{X}_n)^2]$ expansion if asked.

10.2 Mean Squared Error

Mean Squared Error (MSE) – Bias–Variance Decomposition

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = \mathrm{Var}(\hat{\theta}) + \big(\mathrm{Bias}(\hat{\theta})\big)^2.$$

Derivation: write $\hat{\theta} - \theta = (\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta}] - \theta)$, square, and note the cross term has expectation 0.
MSE is the standard criterion for comparing estimators: a *biased* estimator with small variance can beat an unbiased estimator with large variance.

Standard MSE Question Structure

1. Propose/identify $\hat{\theta}$.
2. Compute $\mathbb{E}[\hat{\theta}] \to$ Bias.
3. Compute $\mathrm{Var}(\hat{\theta})$.
4. Combine: $\mathrm{MSE} = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2$.
5. If comparing two estimators, whichever has smaller MSE (possibly only asymptotically, i.e. as $n \to \infty$) is preferred.

10.3 Jensen's Inequality

Jensen's Inequality

If g is convex, $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. If g is concave, $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$.
Key consequence for estimators: if $\hat{\theta}$ is unbiased for θ (i.e. $\mathbb{E}[\hat{\theta}] = \theta$) and g is strictly convex (e.g. $g(x) = \sqrt{x}$ is *concave*, $g(x) = x^2$ is *convex*), then $g(\hat{\theta})$ is **generally biased** for $g(\theta)$:

$$\mathbb{E}[g(\hat{\theta})] \neq g(\mathbb{E}[\hat{\theta}]) = g(\theta) \quad \text{in general.}$$

Worked Example: $\sqrt{\hat{\theta}}$ Is Not Unbiased Even if $\hat{\theta}$ Is

Suppose T is an unbiased estimator of θ^2 (e.g. $T = \frac{3}{k}\sum Y_i^2$ for $Y_i \sim \mathrm{Unif}(-\theta, \theta)$, a standard Tripos setup). Is $\sqrt{T}$ unbiased for θ?
Solution. $g(x) = \sqrt{x}$ is strictly *concave*, so by Jensen $\mathbb{E}[\sqrt{T}] \leq \sqrt{\mathbb{E}[T]} = \sqrt{\theta^2} = \theta$, with equality *only if T is a.s. constant*. Since T has positive variance (it is built from random samples), the inequality is strict:

$$\mathbb{E}[\sqrt{T}] < \theta.$$

So $\sqrt{T}$ is **biased** (it *underestimates* θ) – this is a very common Tripos trap: *unbiasedness does not transfer through nonlinear functions.*

10.4 Case Study: Estimating a Population Size ("Tank Problem")

Setup

Balls/tanks labelled $1, \ldots, N$ (N unknown); observe $X_1, \ldots, X_n$ i.i.d. Unif$\{1, \ldots, N\}$ (sampling with replacement) or without replacement. Two classical unbiased estimators of N:

Model 1: Estimator from the Sample Mean

Since $\mathbb{E}[X_i] = \frac{N+1}{2}$, an unbiased estimator is

$$\hat{N}_1 = 2\overline{X}_n - 1, \qquad \mathbb{E}[\hat{N}_1] = 2 \cdot \frac{N+1}{2} - 1 = N.$$

$$\mathrm{Var}(\hat{N}_1) = 4\,\mathrm{Var}(\overline{X}_n) = \frac{4}{n} \cdot \mathrm{Var}(X_1) = \frac{4}{n} \cdot \frac{N^2-1}{12} = \frac{N^2-1}{3n}.$$

Model 2: Estimator from the Sample Maximum

Let $M = \max(X_1, \ldots, X_n)$. One can show $\mathbb{E}[M] = \frac{n}{n+1}(N+1)$, so the **bias-corrected**, unbiased estimator is

$$\hat{N}_2 = \frac{n+1}{n} M - 1, \qquad \mathbb{E}[\hat{N}_2] = N.$$

Why correct at all? M itself *underestimates* N on average ($\mathbb{E}[M] < N$ for finite n) since the maximum of a sample rarely reaches the true upper limit – the factor $\frac{n+1}{n}$ exactly corrects this bias.
$\mathrm{Var}(\hat{N}_2) = \frac{(N-n)(N+1)}{n(n+2)}$ – crucially, for large n this is $O(N^2/n^2)$, dramatically **smaller** than $\mathrm{Var}(\hat{N}_1) = O(N^2/n)$: **Model 2 (max-based) is far more efficient than Model 1 (mean-based)** for large N.

Why This Matters (German Tank Problem)

This is the famous *German tank problem* (WWII estimation of enemy tank production from serial numbers). Tripos loves testing: (a) deriving $\mathbb{E}[M]$, (b) constructing the bias-corrected estimator, (c) comparing $\mathrm{Var}(\hat{N}_1)$ vs. $\mathrm{Var}(\hat{N}_2)$ via MSE to argue which is preferable.

10.5 Existence of Unbiased Estimators

Trap: An Unbiased Estimator May Not Exist

Not every target quantity has an unbiased estimator based on a given sample. E.g. estimating $1/p$ for Geom(p) is easy ($\overline{X}_n$ works, $\mathbb{E}[\overline{X}_n] = 1/p$), but estimating p itself *unbiasedly* from geometric samples is subtle – direct plug-in estimators like $1/\overline{X}_n$ are typically *biased* (again by Jensen, since $x \mapsto 1/x$ is convex on $x > 0$, so $\mathbb{E}[1/\overline{X}_n] \geq 1/\mathbb{E}[\overline{X}_n] = p$). Constructing an exactly unbiased estimator for p requires more care (e.g. using the indicator $\mathbb{1}[X_1 = 1]$, which has $\mathbb{E}[\mathbb{1}[X_1 = 1]] = p$ exactly, though with much higher variance than a biased plug-in).

Worked Example: Unbiased Estimator for p (Geometric)

$X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Geom(p). Construct an unbiased estimator of $1/p$, then of p.

For $1/p$: $\mathbb{E}[\overline{X}_n] = 1/p$ directly, so $\widehat{1/p} = \overline{X}_n$ is unbiased, with MSE $= \mathrm{Var}(\overline{X}_n) = \frac{(1-p)/p^2}{n}$.

For p: use $\hat{p} = \mathbb{1}[X_1 = 1]$ (Bernoulli indicator that the *first* sample was an immediate success). Since $P(X_1 = 1) = p$, $\mathbb{E}[\hat{p}] = p$ exactly – unbiased, though wasteful (ignores $X_2, \ldots, X_n$) and has variance $p(1-p)$, generally much larger than a biased alternative like $1/\overline{X}_n$. This illustrates the classic bias–variance trade-off central to estimator comparison via MSE.

11 Online Algorithms & Optimal Stopping

11.1 Stopping Problem 1: The Dice Game (Last Success Problem)

Setup

Roll a fair die n times. After each roll you may STOP or CONTINUE. You **win** iff you stop on the *last* 6 to appear among the n rolls. Find the optimal (threshold) strategy.

Key Computation

Probability that a 6 appearing at a given position is the *last* 6 among the final k rolls, i.e. exactly one 6 occurs in the last k rolls:

$$P[\text{exactly one 6 in last } k \text{ rolls}] = \binom{k}{1}\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{k-1} = \frac{k}{6}\left(\frac{5}{6}\right)^{k-1}.$$

This probability, as a function of k, is **unimodal** (increases then decreases) – the optimal strategy is a **threshold rule**: continue until only k^* rolls remain (where k^* maximises the above), then stop at the next 6.

11.2 Stopping Problem 2: The Secretary Problem

Setup

n candidates arrive in uniformly random order; you observe them one at a time and can only compare each to those seen so far (relative ranks). You must accept or reject immediately (no recall), and want to maximise the probability of selecting the *overall best* candidate.

Optimal Strategy & Success Probability

Threshold rule: reject (observe only) the first $r-1$ candidates, then accept the first subsequent candidate who is better than *all* candidates seen so far.
Optimal threshold: $r^* \approx n/e$ (reject roughly the first n/e candidates).
Optimal success probability:

$$P[\text{select the best}] \to \frac{1}{e} \approx 0.368 \quad \text{as } n \to \infty.$$

Sketch of the Argument

Given threshold r, the probability of success is

$$P_n(r) = \sum_{i=r}^{n} P[\text{candidate } i \text{ is best overall, and is selected}] = \frac{r-1}{n}\sum_{i=r}^{n}\frac{1}{i-1}.$$

As $n \to \infty$ with $r/n \to x$, this converges to $-x \ln x$, which is maximised at $x = 1/e$, giving optimal value $1/e$.

Exam Takeaways for the Secretary Problem

- Memorise: **reject first** $\approx n/e$, then **accept the first record-breaker**; success probability $\to 1/e \approx 0.37$.
- Understand *why* $1/e$ is remarkable: this optimal probability does *not* decay to 0 as $n \to \infty$, despite there being n candidates and only one "best".
- Be ready to justify the threshold structure of the optimal policy from first principles (a candidate should only ever be accepted if they are the best seen so far – accepting a non-record is never optimal).

11.3 Generalisation: The Odds Algorithm (Non-Examinable Background)

Odds Algorithm (Bruss, 2000) – context only

Generalises the secretary problem to sequences of independent "success" indicators with known probabilities p_i. Defines odds $r_i = p_i/(1-p_i)$; the optimal stopping rule is to stop at the first success once the sum of remaining odds $\sum_{j \geq i} r_j$ drops below 1. The secretary problem is the special case $p_i = 1/i$. Included in lectures for context – **flagged non-examinable**, but useful for intuition about why $1/e$ emerges.

11 CONSOLIDATED FORMULA SHEET

11.3 Discrete Distributions

Distribution	**PMF**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
$\mathrm{Bern}(p)$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
$\mathrm{Bin}(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
$\mathrm{Pois}(\lambda)$	$e^{-\lambda}\lambda^k/k!$	λ	λ
$\mathrm{Geom}(p)$	$(1-p)^{k-1}p$	$1/p$	$(1-p)/p^2$
$\mathrm{NegBin}(r,p)$	$\binom{k-1}{r-1}p^r(1-p)^{k-r}$	r/p	$r(1-p)/p^2$
$\mathrm{Hyp}(N,n,m)$	$\binom{m}{k}\binom{N-m}{n-k}/\binom{N}{n}$	$n\frac{m}{N}$	$n\frac{m}{N}(1-\frac{m}{N})\frac{N-n}{N-1}$
Discrete $\mathrm{Unif}(a,b)$	$\frac{1}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$

11.3 Continuous Distributions

Distribution	**PDF**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
$\mathrm{Unif}(a,b)$	$1/(b-a)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathrm{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
$N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2

11.3 Core Identities

Identity	**Statement**
Bayes' Theorem	$P(F \mid E) = P(E \mid F)P(F)/P(E)$
Total Probability	$P(E) = \sum_i P(E \mid F_i)P(F_i)$
Linearity of Expectation	$\mathbb{E}[aX+bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ (always, no independence needed)
Variance	$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$; $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$
Sum Variance	$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$
Covariance	$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
Correlation	$\rho(X,Y) = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y) \in [-1,1]$
Tower Property	$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$
Markov	$P[X \geq a] \leq \mathbb{E}[X]/a \quad (X \geq 0)$
Chebyshev	$P[\lvert X-\mu\rvert \geq k] \leq \sigma^2/k^2$
WLLN	$\overline{X}_n \xrightarrow{P} \mu$
CLT	$(\overline{X}_n - \mu)/(\sigma/\sqrt{n}) \xrightarrow{d} N(0,1)$
MSE	$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2$
Jensen (convex g)	$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$
Secretary problem	reject first n/e, success prob. $\to 1/e$

11 TRIPOS EXAM STRATEGY

11.3 Paper Structure & High-Value Topics (based on 2020–2025 Paper 1, Q5–Q6)

Topic	Typical Question Style
Named distributions	Identify the correct distribution (Binomial/Poisson/Geometric/Hypergeometric/Normal/Exponential/Uniform) from a word problem and state PMF/PDF, mean, variance
Approximations	Justify *and* apply a Poisson or Normal approximation to a Binomial; sometimes chained with Markov/Chebyshev/CLT bounds on the same random variable
Conditional probability & independence	Build a joint table from partial information (using row/column sums); test independence *and* compute covariance from the same table
Joint continuous densities	Find normalising constant c; derive marginal densities; test independence via support *and* factorisation
Order statistics	min/max of i.i.d. uniforms or exponentials; CDF-first method
Tail bounds	Apply Markov, then Chebyshev, then CLT to the *same* scenario – know the hierarchy
Estimators & MSE	Propose an estimator, check unbiasedness, compute variance/MSE, sometimes with a Jensen's-inequality twist (e.g. $\sqrt{\hat{\theta}}$)
Indicator-variable expectations	Number of matches/fixed points/successes – always decompose into indicators first
Secretary / optimal stopping	State the $1/e$ rule and threshold structure

11.3 General Exam Technique

Exam Technique

- **State definitions/theorems explicitly** before using them (e.g. "By Bayes' theorem…", "By the CLT…") – method marks are awarded for correctly *naming* the tool used.
- **Always define your random variables and their distributions** at the start of each part – ambiguity costs marks.
- When a question says "you do not need to compute a numerical value", **leave the answer in terms of $\Phi(\cdot)$ or unevaluated sums/binomials** – do not waste time simplifying.
- For joint-table questions, **use the marginal constraint (rows/columns sum to the given marginal)** to fill in unknown cells before answering later parts.
- Cross-check variance/covariance answers using symmetry or a variance-of-a-constant-sum trick where applicable (Section 7).
- For estimator questions, always compute **bias first, then variance, then combine into MSE** – do not attempt to compute MSE directly from the definition unless asked to derive the decomposition itself.

11.3 Top 10 Traps – Commit These to Memory

The 10 Most Common Tripos Errors

1. **Independence**: confusing $\mathrm{Cov}(X,Y)=0$ with independence – it only rules out linear dependence.
2. **Joint density support**: forgetting that a non-rectangular support (e.g. $0<y<x<1$) forces dependence even if the formula looks separable.
3. **Variance of a sum**: dropping the $2\mathrm{Cov}(X,Y)$ term without checking independence first.

4. **Sample variance**: using n instead of $n-1$ for an unbiased variance estimator.
5. **Jensen's inequality**: assuming a nonlinear function of an unbiased estimator is still unbiased (it generally is not).
6. **With vs. without replacement**: using Binomial when Hypergeometric is required (or vice-versa when $n \ll N$).
7. **Continuous densities**: treating $f_X(a)$ as a probability rather than a density.
8. **Tail-bound hierarchy**: applying Chebyshev when only $\mathbb{E}[X] \geq 0$ is known (no variance given) – use Markov instead.
9. **Approximation justification**: choosing Normal vs. Poisson approximation without checking/citing the correct regime ($np, n(1-p)$ both large vs. n large, p small).
10. **Off-by-one errors**: in Geometric/Negative-Binomial tail probabilities and in the memoryless property – always re-derive from $P(X \geq k) = (1-p)^{k-1}$ rather than quoting the shortcut blindly.

11.3 Final Checklist

Before You Walk Into the Exam

- Can you write down, from memory, the PMF/PDF, mean and variance of every distribution in the tables above?
- Can you state Bayes' theorem, the Law of Total Probability/Expectation, Markov's and Chebyshev's inequalities, the WLLN and the CLT precisely (with all conditions)?
- Can you derive $\mathbb{E}[M]$ for the maximum of n i.i.d. discrete/continuous uniforms, and build a bias-corrected estimator from it?
- Can you compute a joint PMF/PDF, its marginals, and test independence *and* covariance from a single joint table or density?
- Can you state the secretary problem's optimal policy and its $1/e$ success probability?