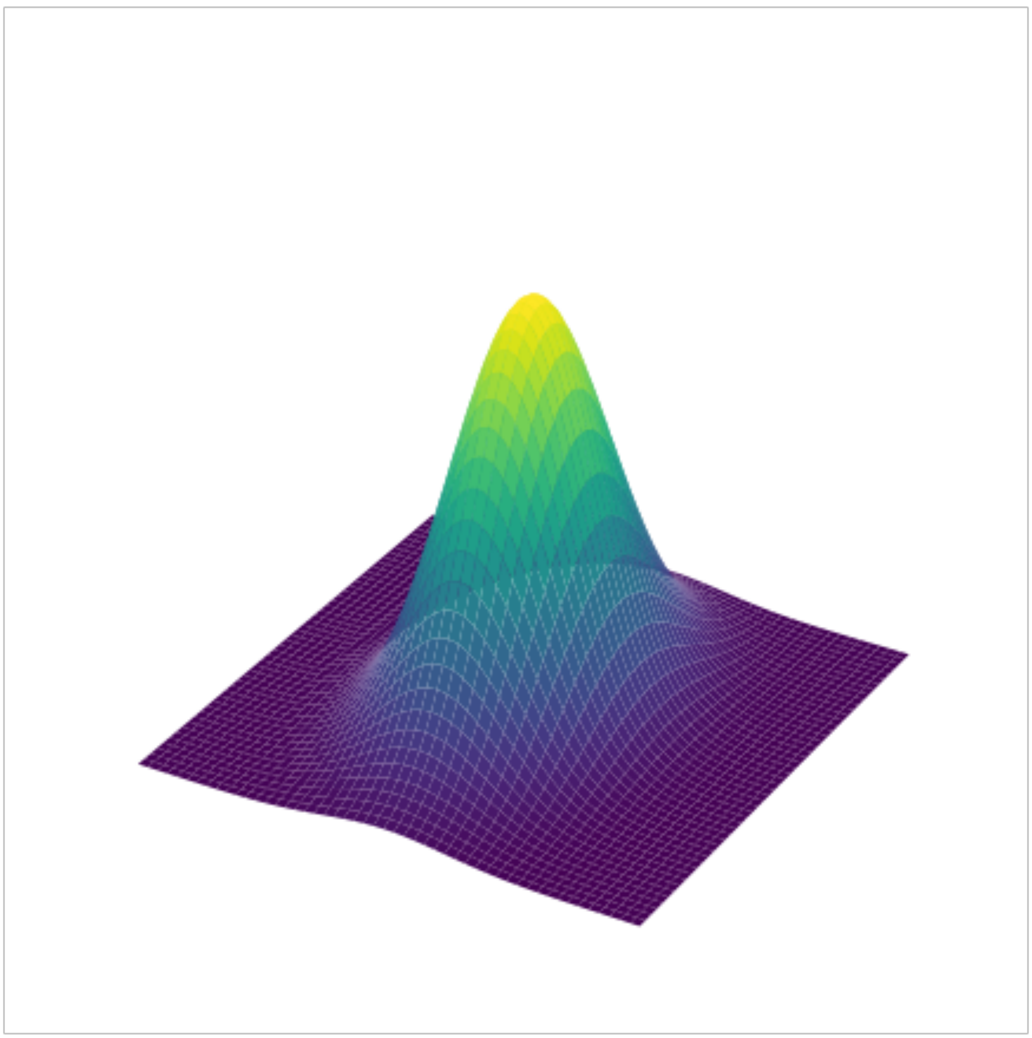
The University of Melbourne

Semester 2 2020

Course Notes

MAST20004 Probability



MAST20004 Probability

1 Axioms of Probability

Terminology and Definitions

Terminology and Definitions	
Random experiment	A process leading to a number (which may be infinite) of possible *outcomes* and the actual outcome that occurs depends on influences that cannot be predicted beforehand.
Sample space	Ω is the set of all possible outcomes of a random experiment.
Events	An event is a **set** of possible outcomes, a subset of the sample space $A \subseteq \Omega$
Outcomes	An outcome is a member of the sample space, $\omega \in \Omega$ (a possible result of a random experiment; different outcomes are mutually exclusive)
Impossible event	The *empty set* is an example of an impossible event
Certain event	The sample space Ω is a certain event, as one of the outcomes in Ω must occur

Warning: Note that outcomes *belong* to a set (the sample space), and events are *subsets* of the sample space. Do not confuse the two.
Note: There can be sets other than Ω that have probability 1, and there can be sets other than $\varnothing$ that have probability 0.

Event Relations

Event Relations	
Union	$A \cup B = \{x : x \in A \vee x \in B\}$
Intersection	$A \cap B = \{x : x \in A \wedge x \in B\}$
Complement	$A^C = \Omega \backslash A = \{x \in \Omega \vert x \notin A\}$
Exclusion	$A \setminus B = \{x \in A \vert x \notin B\} = B \cap A^c$
Disjoint events	$A \cap B = \emptyset$
Exhaustive events	$A_1 \cup A_2 = \Omega$
Collectively exhaustive events	$\bigcup_{i=1}^{n} A_i = \Omega.$
Mutually disjoint events	$A_1, A_2, \ldots$ are disjoint if no two events have outcomes in common: $A_i \cap A_j = \varnothing, \quad \forall i \neq j$

Note: Mutually disjoint events $A_i \cap A_j = \varnothing, \quad \forall i \neq j$ does <u>not</u> imply that $\bigcap_{i=1}^{n} A_i = \varnothing$

Algebra of Set Theory

Algebra of Set Theory	
Commutative law	$A \cup B = B \cup A$ $A \cap B = B \cap A$
Associative law	$(A \cup B) \cup C = A \cup (B \cup C)$ $(A \cap B) \cap C = A \cap (B \cap C)$
Distributive law	$A \cap (B \cup C) = (A \cap B) \cup (B \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (B \cup C)$
De Morgan's Laws	$(A \cup B)^C = A^C \cap B^C$ $(A \cap B)^C = A^C \cup B^C$ $\overline{\bigcap A_i} \equiv \bigcup \overline{A_i}$ $\overline{\bigcup A_i} \equiv \bigcap \overline{A_i}$
Idempotent laws	$A \cup A = A$ $A \cap A = A$
Domination laws	$A \cup \Omega = \Omega$ $A \cap \varnothing = \varnothing$
Absorption laws	$A \cup (A \cap B) = A$ $A \cap (A \cup B) = A$
Uniqueness of complements	If $A \cup B = \Omega$, and $A \cap B = \varnothing$, then $B = A^C$

Note: The event A and its complement are mutually exclusive and exhaustive.
Note: Exhaustive events need not also be disjoint.

Common set decompositions

- $B = A \cup (A^c \cap B) = (A \cap B) \cup (A^c \cap B)$
- $A \cup B = A \cup (A^c \cap B)$
- $A \cap B = A \backslash (A \backslash B)$
- $(B \backslash A)^C = A^C \cap B$

Element-wise set equality

Element-wise approach to set equality

1. Pick an arbitrary element w in the LHS, and show that it belongs to the RHS. $(A \subseteq B)$
2. Pick an arbitrary element w in the RHS, and show that it belongs to the LHS. $(B \subseteq A)$

Ex: Prove $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

1. Let $w \in LHS$, then $(w \in A)$ and $(w \in B$ or $w \in C)$
   Case 1: If $w \in B$, then $w \in A \cap B$, but $A \cap B$ is a subset of RHS. Thus, $w \in RHS$
   Case 2: If $w \in C$, then $w \in A \cap C$, but $A \cap C$ is a subset of RHS. Thus, $w \in RHS$
2. Let $w \in RHS$, then $(w \in A \cap B)$ or $(w \in A \cap C)$.
   Case 1: If $w \in A \cap B$, then ($w \in A$ and $w \in B$). And $B \subseteq (B \cup C)$, thus $w \in LHS$.
   Case 2: If $w \in A \cap C$, then ($w \in A$ and $w \in C$). And $C \subseteq (B \cup C)$, thus $w \in LHS$.

20]

Cardinality relations for finite sets (counting outcomes)

Cardinality of Finite Sets	
For disjoint sets A, B	$\lvert A \cup B \rvert = \lvert A \rvert + \lvert B \rvert$
Unions and Intersections	$\lvert C \cup D \rvert + \lvert C \cap D \rvert = \lvert C \rvert + \lvert D \rvert$
Countable sets:	Sets that are finite or can be enumerated by a 1-1 correspondence (Bijection) with the Natural numbers. Examples: $\mathbb{N}, \mathbb{Z}, \mathbb{N} \times \mathbb{N}, \mathbb{Z} \times \mathbb{Z}$, Counterexamples: $[0, 1], \mathbb{R}$

Probability Axioms

Probability Axioms	$\mathbb{P}(\cdot) : A \mapsto [0,1]$
1. Non-negativity	$\mathbb{P}(A) \geq 0, \quad \forall A \subseteq \Omega$
2. Unitarity	$\mathbb{P}(\Omega) = 1$
3. Countable additivity	$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, where $\{A_1, A_2, A_3, \ldots\}$ is a sequence of **mutually disjoint events**

Probability Properties		**Proof**
Probability of empty set	$\mathbb{P}(\varnothing) = 0$	$\varnothing \cup \varnothing \cup \cdots = \varnothing$ is mutually disjoint set, so by (3), $\mathbb{P}(\varnothing) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing) + \cdots + \mathbb{P}(\varnothing)$, which only holds if $\mathbb{P}(\varnothing) = 0$
Finite additivity	$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$, where $\{A_1, A_2, A_3, \ldots, A_n\}$ is a *finite* set of **mutually disjoint events**	Let $A_1, \ldots, A_n$ be mutually disjoint, construct $A_1, \ldots, A_n, \varnothing, \varnothing, \ldots$. Since $\mathbb{P}(\varnothing) = 0$ and by countable additivity, $\mathbb{P}(A_1 \cup \cdots \cup A_n \cup \varnothing) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n) + 0 + 0 \cdots$
Complement rule	$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$	$\mathbb{P}\left(A \cup A^C\right) = \mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$
Monotonicity	If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$	If $A \subseteq B$, then $B = A \cup (B \setminus A)$, which is disjoint, thus $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$ since probabilities are non-negative
Probability lies in unit interval	$0 \leq \mathbb{P}(A) \leq 1$	$\varnothing \subseteq A \subseteq \Omega$, thus $\mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) \implies 0 \leq \mathbb{P}(A) \leq 1$
Addition theorem	$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)}$	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B))$ And $\mathbb{P}(B) = \mathbb{P}(B \setminus (A \cap B)) + \mathbb{P}(A \cap B)$, eliminate $\mathbb{P}(B \setminus (A \cap B))$
Continuity	(i) Let $\{A_n : n \geq 1\}$ be an increasing sequence of events, namely, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, and define $A = \bigcup_{n=1}^{\infty} A_n$. Then $\mathbb{P}(A) = \lim_{n\to\infty} A_n$ (ii) Let$\{A_n : n \geq 1\}$ be a decreasing sequence of events, namely, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, and define $A = \bigcap_{n=1}^{\infty} A_n$. Then $\mathbb{P}(A) = \lim_{n\to\infty} A_n$	

Classical Probability Model	
Criteria	(i) The sample space Ω is a *finite* set, consisting of N outcomes (ii) The probability of each outcome in Ω occur equally likely. $\mathbb{P}(\{\omega\}) = \dfrac{1}{N}, \forall \omega \in \Omega$
For any event A	$\mathbb{P}(A) = \dfrac{\#A}{\#\Omega}$

2 Conditional Probability

Independence of events

Independence of events	
Independence of two events	$A \perp B \equiv \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$
Pairwise independent of $\{A_i\}$	The events $\{A_i\}$ are *pairwise* independent if $A_i \perp A_j$ for any pair of $i \neq j$
Mutual independence of $\{A_i\}$	The events $\{A_i\}$ are *mutually* independent if for <u>any sub-collection</u> of $\{A_i\}$, $\mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_n}\right) = \mathbb{P}\left(A_{i_1}\right) \times \cdots \times \mathbb{P}\left(a_{i_n}\right)$
Property of mutually independent events	Let $\{A_i : 1 \leq i \leq n\}$ be mutually independent events. We can form new subsets of mutually independent events out of them by selecting <u>non-overlapping</u> events and performing arbitrary operations on them.

Notes

- In general, to prove mutual independence of n events, you need to show $2^n - 1 - n$ equations
- If A and B are disjoint, then there are certainly *not* independent; in fact, they are *negatively* related.
- A *mutually* independent set of events is (by definition) pairwise independent; but the converse is <u>not</u> necessarily true. If A_i are mutually independent then $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)\ldots\mathbb{P}(A_n)$ but the converse need not be true.
- An event is independent of itself iff $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$ (the empty set is independent of itself).

Conditional Probability

Conditional probability	
Conditional probability	$\boxed{\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}$
Multiplication theorem	$\mathbb{P}(A \cap B) = \mathbb{P}(B)\,\mathbb{P}(A \mid B)$
Positive and negative relations	$\begin{cases} \text{Positive relation : } \mathbb{P}(A \mid B) > \mathbb{P}(A) \\ \text{Negative relation : } \mathbb{P}(A \mid B) < \mathbb{P}(A) \end{cases}$

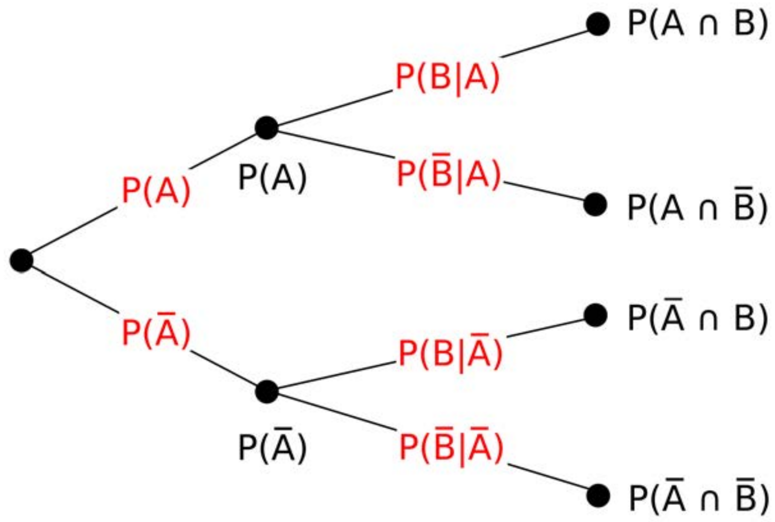
Law of Total Probability and Bayes' theorem

> Intuition for the Law of Total Probability: We think of an event as the effect /result due to one of several distinct causes/reasons. In this way, we compute the probability of the event by *conditioning* on each of the possible causes and adding up all these possibilities. We interpret A as an effect and interpret a partition $\{B_i\}$ as the possible distinct causes leading to the effect A.

Law of Total Probability & Bayes' Theorem	
Law of Total Probability	$\boxed{\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(B_i)\,\mathbb{P}(A \mid B_i)}$, where $\{B_i\}$ is a *partition* of Ω
Bayes' Theorem	$\boxed{\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\sum_j \mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j)}}$, for a given partition $\{A_j\}$
Bayes' theorem for the partition $\Omega = A \cup B$	$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B \mid A)\,\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\,\mathbb{P}(A^c)}$
Partition of Ω	The set of events $\{A_i\}$ is a partition of Ω if they are mutually disjoint and collectively exhaustive : $(A_i \cap A_j = \varnothing) \cap (\bigcup_i A_i = \Omega)$

Tree Diagrams

- A tree diagram can be used to represent a conditional probability space
- Each node represents an event, and each edge is weighted by the corresponding probability (the root node is the certain event, with probability of one)
- Each set of sibling branches is a partition of the parent event, and sums to one
- The Law of Total Probability is equivalent to reaching the 'effect' (leaf node) though any of the possible 'causes' (intermediary branches). Multiply along the branches and add the result.


P(A ∩ B)
P(B|A)
P(A)
P(A)
P(B̄|A)
P(A ∩ B̄)
P(Ā ∩ B)
P(Ā)
P(B|Ā)
P(Ā)
P(B̄|Ā)
P(Ā ∩ B̄)

3 Random Variables

Random Variables	
Random Variables	A random variable maps <u>*outcomes*</u> of the sample space to a *real number.* $X : \Omega \to \mathbb{R}, \omega \mapsto X(\omega)$
State space	The state space of a random variable is the range of X, the set of all possible values of $X(\omega)$. $S_X \subseteq \mathbb{R} = \{X(\omega) : \omega \in \Omega\}$
Events of a random variable	*Events* of a random variable for a given $x \in S_X$, are the events in the sample space that are mapped by the random variable to the value x. $A_x = \{\omega \in \Omega : X(\omega) = x\}$
Probability of an event	$\mathbb{P}\left(X = x\right) = \mathbb{P}\left(A_x\right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = x\}\right)$
Distribution of a random variable	$(X, \mathbb{P}\left(X = x\right))$

Discrete Random Variables

> Where the state space S_X is *countable*

Probability Mass Functions (pmf) p_X	
Discrete random variables	A discrete r.v. is one for which the state space S_X is *countable*
Probability Mass Functions	The probability mass function of X maps the outcomes of X to a probability. $p_X : S_X \to [0, 1], p_X(x) = \mathbb{P}\left(X = x\right)$
Probability of a given event in the sample space	$\mathbb{P}\left(x \in A\right) = \sum_{x \in A} \mathbb{P}\left(X = x\right) = \sum_{x \in A} p_X(x)$
Properties of a PMF	1. $p_X(x) \geq 0, \quad \forall x \in S_X$ 2. $\sum_{x \in S_X} p_X(x) = 1$
PMF from CDF	$\mathbb{P}\left(X = k\right) = F_X(k) - F_X(k-1)$, for $S_X \in \mathbb{Z}$

Cumulative Distribution Function (cdf)

Cumulative Distribution Function (cdf)	$F_X : \mathbb{R} \to [0,1], \quad F_X(x) = \mathbb{P}(X \leq x) = \sum_{y \in S_x,\, y \leq x} p_X(y)$
CDF lies in unit interval	$0 \leq F_X(x) \leq 1$
Probability of $a < X \leq b$	$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$
Limits to infinity	$\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$

Continuous Random Variables

> Where the state space S_X is *uncountable*

Probability density function (pdf) f_X	
Probability density function	$f_X : \mathbb{R} \to [0,\infty), \quad f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$, the function that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$
Interpretation as a 'probability density' around the point x	$\mathbb{P}(X \approx x) \approx f_X(x)\delta$
Non-negativity	$f_X(x) \geq 0$ (since $F_X(x)$ is increasing function, so $F_X' \geq 0$; the derivative of an increasing function is non-negative)
Probability as area under f_x	$\mathbb{P}(a \leq x \leq b) = \int_a^b f_X(t)\, dt = F_X(b) - F_X(a)$
Area under f_x is 1	$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

Notes:

- Any function that satisfies $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$ and $f_X(x) \geq 0$ is the pdf of some random variable.
- The **range** of $f_X(x)$ need not be ≤ 1, probability densities represent *areas*, not probabilities themselves

Expectation

Expectation	
Discrete random variables	$\mathbb{E}[X] = \sum_{x \in S_X} x \cdot p_X(x) = \sum_{\omega \in \Omega} X(\omega)\,\mathbb{P}(\omega)$
Continuous random variables	$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$
Expectation of *functions* of random variables $\psi(X) : \mathbb{R} \to \mathbb{R}$	$\mathbb{E}[\psi(X)] = \begin{cases} \sum_{x \in S_X} \psi(x)\,p_X(x) \\ \int_{-\infty}^{\infty} \psi(x)\,f_X(x)\,dx \end{cases}$
Linearity property of expectation	$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
Expectation via tail probabilities (for continuous **non-negative** rvs)	$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X > x)\;dx = \int_0^{\infty} 1 - F_X(x)\,dx$
Expectation via tail probabilities (for discrete rvs where $S_X \subseteq \mathbb{N}$)	$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n)\,, \quad n \in \mathbb{N}$

Geometric interpretation of Expectation: We can interpret the average value or expectation as the 'centre of mass' of the probability distribution.

Variance

Variance	
Variance	$\boxed{\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2}$
Variance of linear function	$\mathrm{Var}[aX + b] = a^2\,\mathrm{Var}[X]$
Standard deviation	$\mathrm{sd}(X) = \sqrt{\mathrm{Var}[X]}$
Non-negativity	$\mathrm{Var}[X] \geq 0$
Constant X value	$\mathrm{Var}[X] = 0 \iff \mathbb{P}(X = \mu_X) = 1$

Moments

Moments	$n \in \mathbb{N}$
Moment	$\boxed{\mu_n = \mathbb{E}\left[X^n\right]}$
Central moment	$\boxed{\nu_n = \mathbb{E}\left[(X - \mu_X)^n\right]}$
Expectation as first moment	$\mathbb{E}\left[X\right] = \mu_1$
Variance as second central moment	$\mathrm{Var}[X] = \nu_2$
Moments via tail probabilities for $n \geq 1$ and **non-negative** X	$\mathbb{E}\left[X^n\right] = n \int_0^\infty x^{n-1}(1 - F_X(x))\, dx = n \int_0^\infty x^{n-1}\, \mathbb{P}\left(X > x\right)\, dx$
Discrete moments	$\mathbb{E}\left[X^n\right] = \sum_{x \in S_X} x^n p_X(x)$

Note:

- If the n-th moment about any point exists, so does the $(n-1)$-th moment (and thus, all lower-order moments) about every point.
- The zeroth moment of any pdf is 1, since the area under any pdf must be equal to one.

Taylor Approximations of Expectation and Variance

Taylor Approximations of Expectation and Variance	
Taylor Approximation for Expectation	$\boxed{\mathbb{E}\left[\psi(X)\right] \approx \psi(\mu) + \frac{1}{2}\psi''(\mu)\, \sigma_X^2}$
Taylor Approximation for Variance	$\boxed{\mathrm{Var}[X] \approx \left[\psi'(\mu)\right]^2\, \sigma_X^2}$

Note:

- The second order Taylor expansion of $\psi(X)$ about $x = \mu$ is $\psi(X) \approx \psi(\mu) + \psi'(\mu)(X - \mu) + \frac{1}{2}\psi''(\mu)(X - \mu)^2$
- If we want the approximation to be more accurate, we need to consider higher order Taylor approximations of $\psi(X)$.

Chebyshev and Markov inequalities

Inequalities	$\lambda > 0$
Chebyshev's inequality	$\boxed{\mathbb{P}\left(\lvert X-\mu\rvert \geq \lambda\right) \leq \dfrac{\sigma^2}{\lambda^2}}$ or $\mathbb{P}\left(\left\lvert \dfrac{X-\mu}{\sigma}\right\rvert \geq \lambda\right) \leq \dfrac{1}{\lambda^2}$
Markov's Inequality	$\boxed{\mathbb{P}\left(\lvert X\rvert \geq \lambda\right) \leq \dfrac{\mathbb{E}[\lvert X\rvert^k]}{\lambda^k}}$

Corollaries to Chebyshev's inequality:

- $\mathbb{P}\left(\lvert X-\mu\rvert < \lambda\right) > 1 - \dfrac{\sigma^2}{\lambda^2}$
- $\mathbb{P}\left(a < X < b\right) = \mathbb{P}\left(\left\lvert X - \dfrac{a+b}{2}\right\rvert \leq \dfrac{b-a}{2}\right) \geq 1 - \dfrac{\sigma^2 + \left(\mu - \dfrac{a+b}{2}\right)^2}{\left(\dfrac{b-a}{2}\right)^2}$
- $\mathbb{P}\left(X \geq \mu + a\right) \leq \dfrac{\sigma^2}{\sigma^2 + a^2}$

Transformations of Random Variables

General principle: Calculate the CDF of $Y = \psi(X)$ by inverting $\psi(X)$ in $F_Y(y) = \mathbb{P}\left(y \leq y\right) = \mathbb{P}\left(\psi(X) \leq y\right)$ and then differentiate to obtain the pdf.

Warning: Be careful of domains of the transformed random variable!

Example Transformations of Random Variables	**CDF of Y**	**PDF of Y**
Linear transformations $Y = aX + b$	$F_Y(y) = \mathbb{P}\left(X \leq \dfrac{y-b}{a}\right) = F_X\left(\dfrac{y-b}{a}\right)$	$f_Y(y) = \dfrac{1}{\lvert a\rvert} f_X\left(\dfrac{y-b}{a}\right)$
Square transformations $Y = X^2$	$\begin{aligned} F_Y(y) &= \mathbb{P}\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= \mathbb{P}\left(X \leq \sqrt{y}\right) - \mathbb{P}\left(X < -\sqrt{y}\right) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$	$f_Y(y) = \begin{cases} \dfrac{1}{2\sqrt{y}}\left(F_X(\sqrt{y}) + f_X(-\sqrt{y})\right), & y > 0 \\ 0, & \text{Else} \end{cases}$
Minimums $Y = \min(X, M)$, where $M \in \mathbb{R}$	$F_Y(y) = \begin{cases} F_X(y), & y < M \\ 1, & y \geq M \end{cases}$	

Note: For $Y = \min(X, M)$ where $M \in \mathbb{R}$, the cdf of Y is often discontinuous at the threshold value M. Thus $Y = \min(X, M)$ is often an example of a random variable that has *neither* a discrete nor a continuous distribution.

General Transformations and applications

Standardised random variable	The standardised random variable $\boxed{X_s = \dfrac{X-\mu}{\sigma}}$ has mean 0 and variance 1.
General Monotonic (strictly increasing or decreasing) Transformations	$\boxed{f_Y(y) = \left\lvert \dfrac{1}{\psi'(\psi^{-1}(y))} f_X(\psi^{-1}(y)) \right\rvert}$
Linear transformation of normal distribution	$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \implies aX + b \sim \mathcal{N}\left(a\mu + b, a^2\sigma^2\right)$
Square of standard normal distribution	$X \sim \mathcal{N}(0,1) \implies X^2 \sim \gamma(1/2, 1/2)$

Monotonic transformations

Let $Y = \psi(X)$, where ψ is an *increasing* function, then

$$\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \le y) \\
&= \mathbb{P}(\psi(X) \le y) \\
&= \mathbb{P}\left(X \le \psi^{-1}(y)\right) \\
&= F_X(\psi^{-1}(y)) \\
\therefore f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_X(\psi^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} \psi^{-1}(y) \\
&= \frac{1}{\psi'(\psi^{-1}(y))} f_X(\psi^{-1}(y))
\end{aligned}$$

Simulating random variables with a given distribution function F

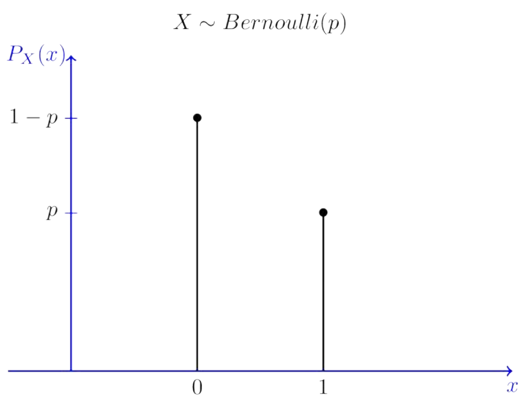
1. Let $U \sim R(0,1)$ be the uniform distribution on the unit interval, and let F be a given arbitrary cumulative distribution function. Then the cdf of $F^{-1}(U) = F$.
2. Thus, to simulate random variates sampled from F, we sample from U and apply $F^{-1}(U)$.

4 Discrete Distributions

Bernoulli Distribution

> If the random experiment has a binary set of outcomes (Success & Failure, Heads & Tails, True & False…).

Bernoulli distribution	$X \sim \text{Bi}(1, p)$
Sample space and random variable state space	$\Omega = \{S, F\}$ where $X(\omega) = \begin{cases} 1, & \omega = S \\ 0, & \omega = F \end{cases}$
PMF	$p(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$
CDF	$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$
Expectation	$\mathbb{E}[X] = p$
Variance	$\text{Var}[X] = p(1-p) = pq$



Binomial Distribution

> The number of **successes** (x) in a *sequence* of n independent Bernoulli trials with probability of success p.

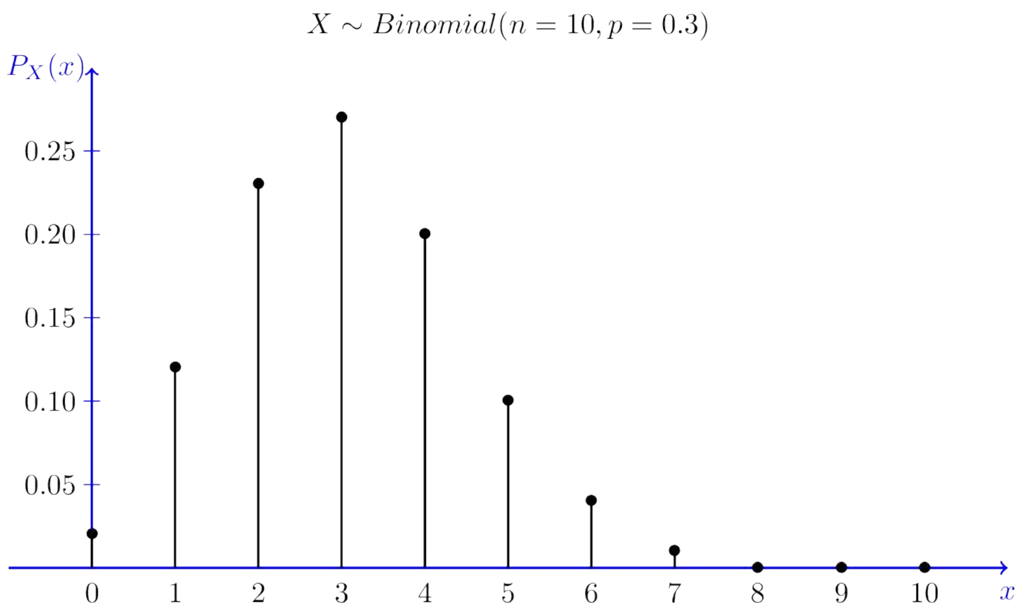
> **Sampling with replacement**: Binomial models the number of successes in a sample of size n drawn **with replacement** from a population of size N.

Binomial distribution	$X \sim \mathrm{Bi}(n, p)$
Sample space and random variable state space	The set of all sequences of n letters consisting either of S or F. $\Omega = \{(x_1, x_2, \ldots, x_n) : x_i \in \{S, F\}\}$. $S_X = \{0, 1, 2, 3 \ldots\}$
PMF	$\boxed{\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}}$, $k \in \{0, 1, 2, \ldots, n\}$
Expectation	$\mathbb{E}[X] = np$
Variance	$\mathrm{Var}[X] = np(1-p) = npq$
Normal approximation to the binomial	$X \dot{\sim} \mathcal{N}(np, np(1-p))$
As sum of Bernoulli	A Binomial $\mathrm{B}(n, p)$ random variable can be thought of as the sum of n independent $\mathrm{B}(1, p)$ Bernoulli random variables. $X = B_1 + B_2 + \cdots + B_n$.

Shape of distribution (Recursive formula): Since the ratio of successive probabilities $\dfrac{\mathbb{P}(X = k)}{\mathbb{P}(X = k-1)} = 1 + \dfrac{(n+1)p - k}{k(1-p)}$, if $x < p(n+1)$ then the pmf is increasing, and if $x > p(n+1)$ the pmf is decreasing.

Number of peaks: Where $(n+1)p \notin \mathbb{Z}$ the distribution has a single peak, and if $(n+1)p \in \mathbb{Z}$, the distribution has two equal neighbouring peaks - at $(n+1)p$ and $(n+1)p - 1$.

Derivation: The formula can be understood as follows: k successes occur with probability p^k and $n-k$ failures occur with probability $(1-p)^{n-k}$. Together, there are $\binom{n}{k}$ ways to distribute the k successes amongst the n trials,



Geometric distribution

> The number of **failures** (k or x) before the *first* success in an infinite sequence of independent Bernoulli trials with probability of success p.

Geometric distribution	$X \sim G(p)$
Sample space and random variable state space	$S_X = \{0, 1, 2, 3, \ldots\}$
PMF	$\boxed{\mathbb{P}(X = k) = (1 - p)^k p}$, where $k \in \{0, 1, 2, 3, \ldots\}$
CDF	$1 - (1 - p)^{k+1}$
Expectation	$\mathbb{E}[X] = \dfrac{1 - p}{p}$
Variance	$\text{Var}[X] = \dfrac{1 - p}{p^2}$
Memoryless property	$\mathbb{P}(X \geq a + b \mid X \geq b) = \mathbb{P}(X \geq a)$

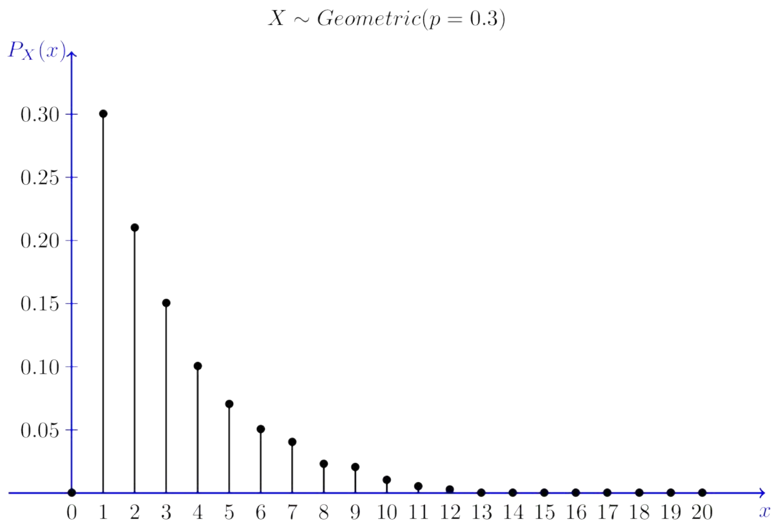
Derivation: $\mathbb{P}(X = 0) = \mathbb{P}(S) = p$, $\mathbb{P}(X = 1) = \mathbb{P}(FS) = (1 - p)p$, $\mathbb{P}(X = 2) = \mathbb{P}(FFS) = (1 - p)^2 p$ and so on.

Warning: Careful of the parametrisation of the geometric distribution! *MAST20004 Probability* uses the number of **failures** (X) before the first success, but a common alternate parametrisation is the number of **trials** before the first success ($Y = X + 1$).

Memorylessness: If you intend to repeat an experiment until the first success, then, given that the first success has not yet occurred, the conditional probability distribution of the number of additional trials does not depend on how many failures have been observed. The die one throws or the coin one tosses does not have a "memory" of these failures.

Hence, given that the first a trials were all failures, the *residual* time $T-a$ until the first success has the same distribution as the original waiting time T.

Warning: Memorylessness means that $\mathbb{P}(X > 40 \mid X \geq 30) = \mathbb{P}(X > 10)$, **not** that $\mathbb{P}(X > 40 \mid X \geq 30) = \mathbb{P}(X > 40)$



Negative Binomial distribution

> The **total number of failures** (x or k) *before the r-th* success in an **infinite** sequence of independent Bernoulli trials.

Negative Binomial distribution	$X \sim \mathrm{NB}(r, p), \quad r > 0$
Sample space and random variable state space	$S_X = \{0, 1, 2, 3, \ldots\}$
PMF	$\boxed{p_X(k) = \binom{-r}{k}(p-1)^k p^r}$, where $k \in \{0, 1, 2, 3, \ldots\}$
Expectation	$\mathbb{E}[X] = r\dfrac{(1-p)}{p}$
Variance	$\mathrm{Var}[X] = r\dfrac{(1-p)}{p^2}$
Relation with Geometric	$X \sim \mathrm{NB}(r, p) = X_1 + \cdots + X_r$, where $X_i \sim \mathrm{G}(p)$ and $X_1, \ldots, X_r$ are independent.
Relation with Binomial	$Z \sim \mathrm{NB}(r, p),\ X \sim \mathrm{B}(n, p)$ $n \geqslant r \Longrightarrow \mathbb{P}(Z \leqslant n - r) = \mathbb{P}(X \geqslant r)$

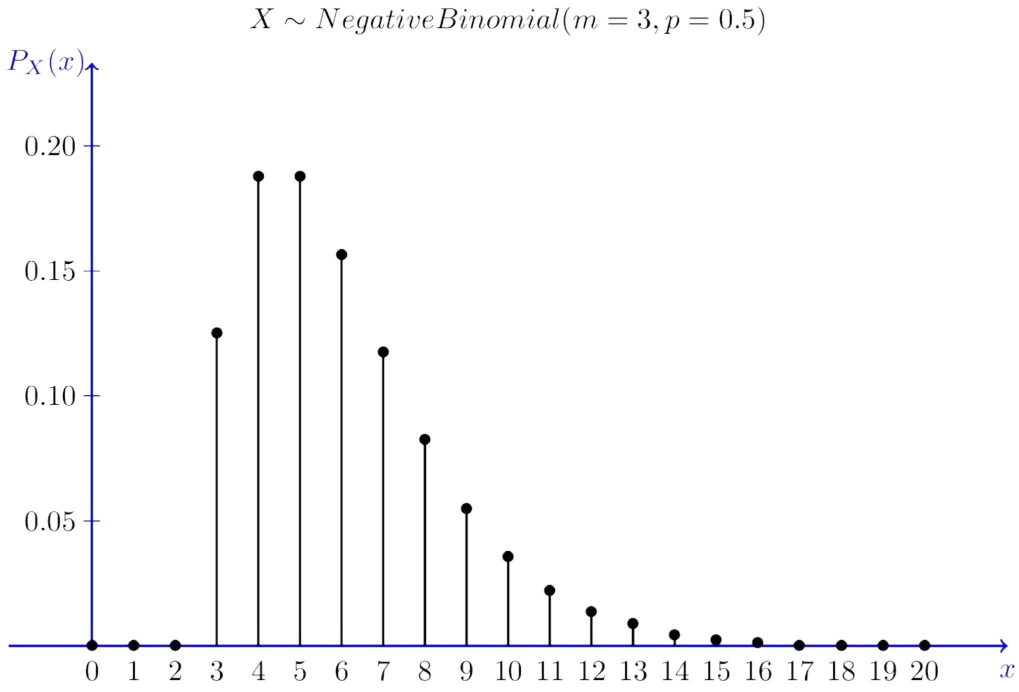
Note: The geometric distribution is a special case of the negative binomial distribution, with $r = 1$.

Negative Binomial relation with Binomial

Let $n \geq r$ be given positive integers. Let Z be given by the total number of failures before the r-th success. Let X be the total number of successes among the first n trials. Then $Z \sim \mathrm{NB}(r, p)$, $X \sim \mathrm{B}(n, p)$ and $\{Z \leq n - r\} = \{X \geq r\}$

Shape of negative binomial distributions

Computing ratio of successive probabilities, we have $\frac{p_X(k)}{p_X(k-1)} = \left(\frac{r-1}{k} + 1\right)(1-p)$, and $\frac{p_X(k)}{p_X(k-1)} \geq 1 \implies k \leq \frac{(r-1)(1-p)}{p}$. Thus, as a result, if $r \leqslant 1$, the *pmf is strictly decreasing for all time*. If $r > 1$, the pmf is strictly increasing before the threshold $\frac{(r-1)(1-p)}{p}$ and strictly decreasing after $\frac{(r-1)(1-p)}{p}$
.



Hypergeometric Distribution

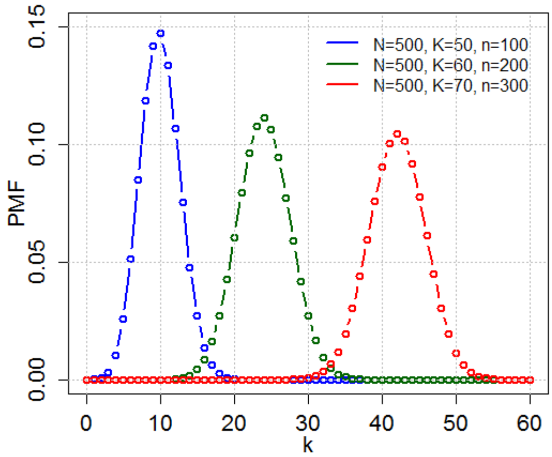
> Sampling **without** replacement. The probability of selecting n defective items (with replacement) from a population of N, where D items in the population are defective.

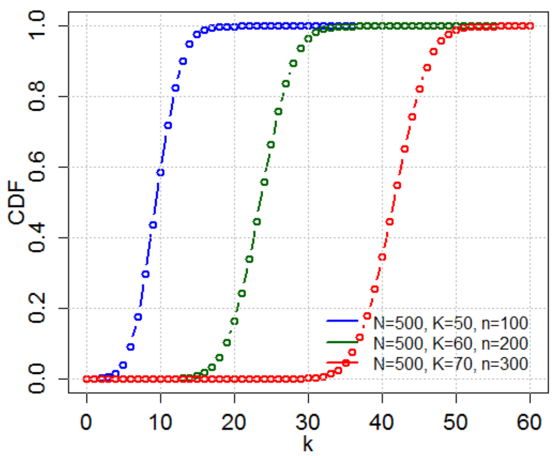
Hypergeometric distribution	$X \sim \mathrm{Hg}(n, D, N), \quad n \leq N$
Sample space and random variable state space	$S_X = \{0, 1, 2, \ldots, n\}$
PMF	$\boxed{p_X(k) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}}$, where $k \in \{0, 1, 2, \ldots, n\}$
Expectation	$\mathbb{E}[X] = \frac{nD}{N}$
Variance	$\mathrm{Var}[X] = \frac{nD(N-D)}{N^2}\left(1 - \frac{n-1}{N-1}\right)$
Binomial approximation to the Hypergeometric	As $n \to \infty$, we have $X \sim \mathrm{Hg}(n, D, N) \approx \mathrm{Bi}(n, D/N)$

Binomial approximation to the Hypergeometric

Suppose that N and D are both very large but the ratio $p = N/D$ is fixed. We also assume that n is fixed. Since the population N is very large comparing with n, it makes little difference about whether we are sampling with or without replacement (the proportion of defective items in the box is almost unchanged). Thus, when N is large, $X \sim \mathrm{Hg}(n, D, N) \approx \mathrm{Bi}(n, D/N)$. Precisely, we have

$\lim_{N\to\infty} \frac{\binom{D}{k}\cdot\binom{N-D}{n-k}}{\binom{N}{n}} = \binom{n}{k} p^k (1-p)^{n-k}.$





Poisson distribution

> Performing independent Bernoulli trials **continuously** over a unit time interval with given success "rate" (or arrivals rate/density) λ - the **expected number of arrivals per unit time**. The Poisson random variable counts the number of successes or arrivals in a given period with a given arrival rate λ.

Poisson distribution	$X \sim \text{Pn}(\lambda), \quad \lambda > 0$
State space	$S_X = \{0, 1, 2, 3, \ldots\}$
PMF	$\boxed{p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}}$, where $k \in \{0, 1, 2, 3, \ldots\}$
Time-rate of events	If $r =$ time rate of events $[1/\text{time}]$, then $\mathbb{P}\left(\text{k events in interval } t)\right) = \frac{(rt)^k e^{-rt}}{k!}$
Expectation	$\mathbb{E}\left[X\right] = \lambda$
Variance	$\text{Var}[X] = \lambda$
General <u>Poisson approximations</u>	1. $X = X_1 + \cdots + X_m, \quad X_i \sim \text{Bi}(1, p_i)$, where m large, p_i small and need not be equal. 2. $X_1, \ldots, X_m$ are 'weakly dependent' Then, $X \sim \text{Pn}(\lambda = p_1 + \cdots + p_m)$
Poisson approximation to the Binomial	For a binomial $X \sim \text{Bi}(n, p) \approx \text{Pn(np)}$, for p small.
Convolution	$X \sim \text{Pn}(\lambda), Y \sim \text{Pn}(\mu), X \perp Y \implies X + Y \sim \text{Pn}(\lambda + \mu)$

Derivation: We can consider taking the Binomial pmf with $p = \lambda/n$ and shrinking the time period to 0 by taking $n \to \infty$: $\mathbb{P}\left(X = k\right) = \binom{n}{k}(\lambda/n)^x(1 - \lambda/n)^{n-k} \to \frac{\lambda^k}{k!}e^{-\lambda}$

Poisson Modelling:

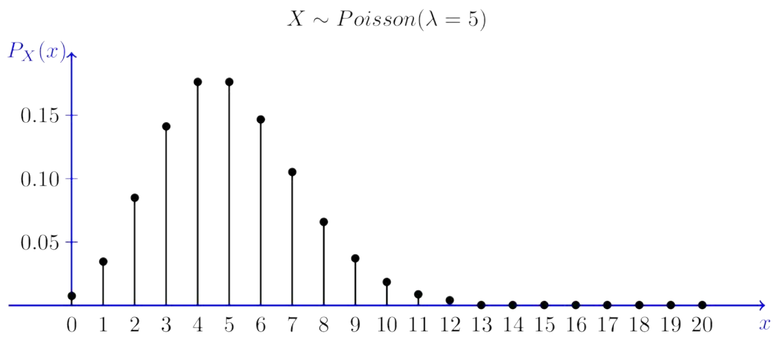
- The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
- The average rate at which events occur is independent of any occurrences. For simplicity, this is usually assumed to be constant, but may in practice vary with time.
- Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur.

Poisson Approximations

Conditions:

1. The random variable X can be written as the sum of m Bernoulli variables with parameter p_i: $X = X_1 + \cdots + X_m, \quad X_i \sim \text{Bi}(1, p_i)$. Note that the Bernoulli trials need not be fully independent, nor do they need to share the same success probability p.
2. The number of Bernoulli variables (m) is very large, each p_i is very small, and $P_1 + \cdots + P_m$ is of 'normal' scale.
3. The dependence among the random variables $X_1, \ldots, X_m$is weak, in the sense that for each X_i, the number of random variables that are dependent on (related to) X_i is much smaller than m.

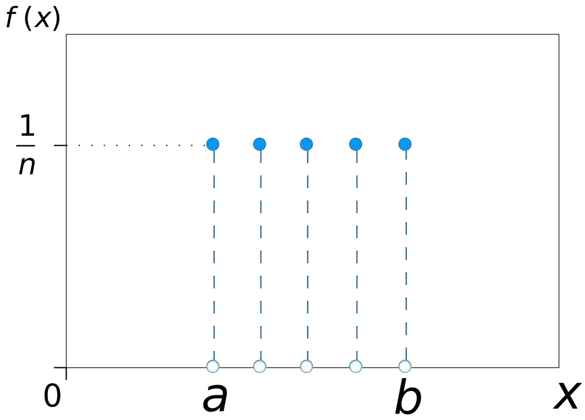
Then, we can approximate the pmf of X by the Poisson distribution with parameter $\lambda = \sum_i^m p_i$.

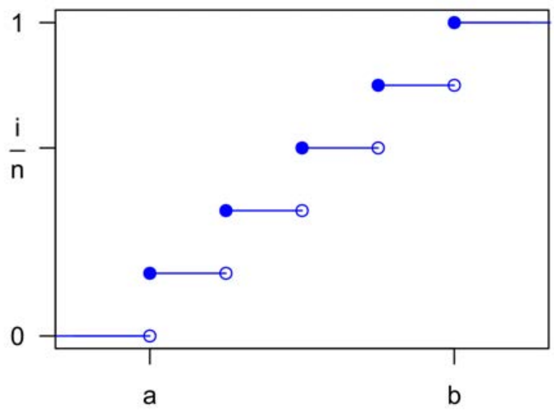


Discrete Uniform Distribution

> A discrete distribution that takes on integer values between a and b **uniformly**.

Discrete Uniform Distribution	$X \sim \mathrm{U}(a, b), \quad a < b$
State space	$S_X = \{a, a+1, \ldots, b\}$ (set of integers between a and b)
PMF	$\boxed{p_X(k) = \frac{1}{b-a+1}}$, where $k \in \{a, a+1, \ldots, b\}$
Expectation	$\mathbb{E}[X] = \frac{a+b}{2}$
Variance	$\mathrm{Var}[X] = \frac{(b-a+1)^2 - 1}{12}$


f (x)
1/n
0
a
b
x

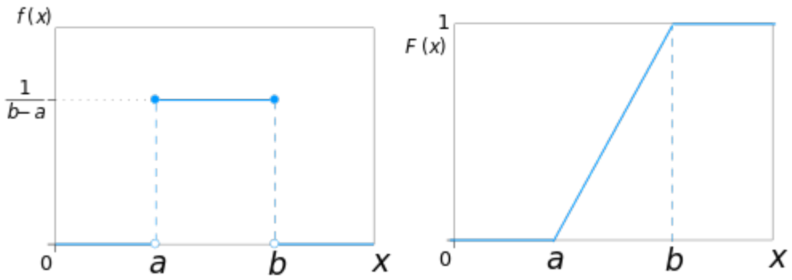

1
i/n
0
a
b

5 Continuous Distributions

Continuous Uniform Distribution

> A continuous distribution that takes on any real value between a and b **uniformly**.

Continuous Uniform Distribution	$X \sim \mathrm{R}(a, b)$
PDF	$\boxed{f_X(x) = \frac{1}{b-a}, \quad a < x < b}$
CDF	$\boxed{F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}}$
Probability of $c < x < d$	$\mathbb{P}(c < x < d) = \frac{d-c}{b-a}$
Expectation	$\mathbb{E}[X] = \frac{a+b}{2}$
Variance	$\mathrm{Var}[X] = \frac{(b-a)^2}{12}$



Exponential distribution

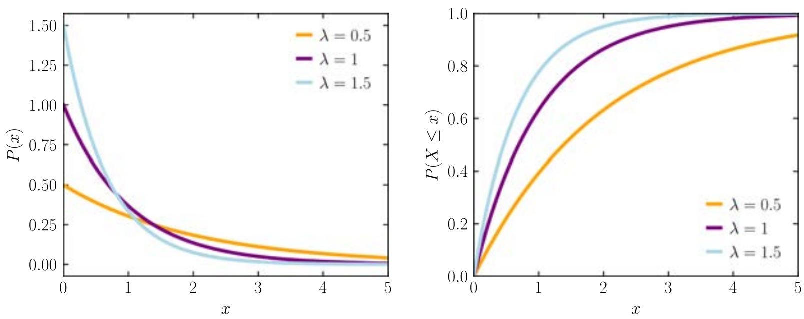
> Consider **infinite** independent Bernoulli trials in **continuous** time with rate α, then X is the **time until the first success**. Exponential distributions are often used to model the waiting time *between* the occurrence of events.
>
> The parameter λ is naturally interpreted as the average number of arrivals per unit interval. More generally, we can always regard α as the rate at which the underlying event occurs per unit time.
>
> The Exponential distribution is the **continuous** time analogue of the **Geometric** (measures time until first success)

Exponential distribution	$X \sim \exp(\lambda), \quad \lambda > 0$
PDF	$\boxed{f_X(t) = \lambda e^{-\lambda t}, \quad t \geqslant 0}$
CDF	$\boxed{F_X(t) = \mathbb{P}(X \leqslant t) = \begin{cases} 1 - e^{-\lambda t}, & t \geqslant 0 \\ 0, & t < 0 \end{cases}}$
Expectation	$\mathbb{E}[X] = \frac{1}{\lambda}$
Variance	$\text{Var}[X] = \frac{1}{\lambda^2}$
Memoryless property	$\mathbb{P}(X \geq s + t \mid X \geq s) = \mathbb{P}(X \geq t)$
As special case of Gamma Distribution	$\exp(\lambda) \sim \gamma(1, \lambda)$

Note: The geometric distribution is a particular case of the gamma distribution.

Intuition for rate parameter: If you receive phone calls at an average rate of 2 per hour, then you can expect to wait *half an hour* for every call.

Intuition for memoryless property: When X is interpreted as the waiting time for an event to occur relative to some initial time, this relation implies that, if X is conditioned on a failure to observe the event over some initial period of time s, the distribution of the remaining waiting time is the same as the original unconditional distribution. For example, if an event has not occurred after 30 seconds, the conditional probability that occurrence will take at least 10 more seconds is equal to the unconditional probability of observing the event more than 10 seconds after the initial time.



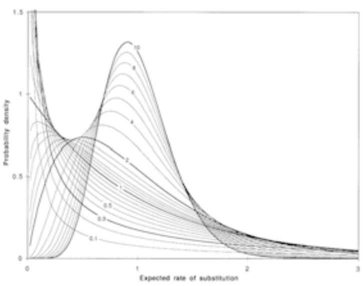
Gamma Distribution

> The Gamma distribution is the continuous analogue fo the Negative Binomial distribution. The Gamma distribution models the waiting time until the r-th occurrence of an event, in **continuous** time.

Gamma distribution	$X \sim \gamma(r, \alpha), \quad r, \alpha > 0$
PDF	$\boxed{f_X(t) = \dfrac{\alpha^r}{\Gamma(r)} e^{-\alpha t} t^{r-1}, \quad t \geqslant 0}$
Expectation	$\mathbb{E}[X] = \dfrac{r}{\alpha}$
Variance	$\mathrm{Var}[X] = \dfrac{r}{\alpha^2}$
Moment	$\mathbb{E}\left[X^k\right] = \dfrac{\Gamma(r+k)}{\Gamma(r)\alpha^k}$
Relation to Exponential Distribution	Let $X_1, X_2, \ldots, X_n$ be n iid exponential variables $X_i \sim \exp(\lambda)$, then $\sum X_i \sim \gamma(n, \frac{1}{\lambda})$.

Gamma function

Gamma function	
Gamma function	$\boxed{\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx, \quad r > 0}$
Recursive definition	$\boxed{\Gamma(r) = (r-1)\Gamma(r-1)}, \quad r > 0$
Gamma and Factorial	$\Gamma(n) = (n-1)!, \quad \forall n \in \mathbb{N}$
Base cases	$\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$

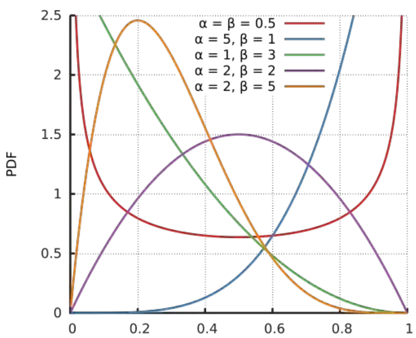


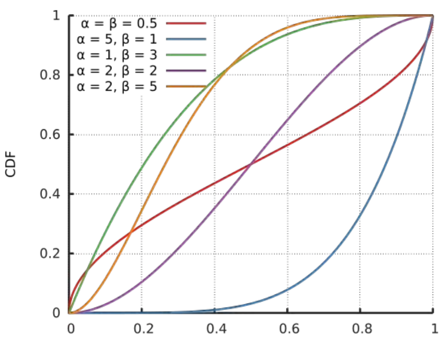
Beta distribution

Beta distribution	$X \sim \text{Beta}(\alpha, \beta), \quad \alpha, \beta > 0$
PDF	$\boxed{f_X(x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}}, \quad 0 \le x \le 1$
Mean	$\mathbb{E}\left[X\right] = \dfrac{\alpha}{\alpha+\beta}$
Variance	$\text{Var}[X] = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Moments	$\mathbb{E}\left[X^k\right] = \dfrac{B(\alpha+k,\beta)}{B(\alpha,\beta)} = \dfrac{\Gamma(a+k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}$

Beta function

Beta function	
Beta function	$\boxed{B(\alpha,\beta) \equiv \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx, \quad \alpha, \beta > 0}$
Beta and Gamma function	$B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

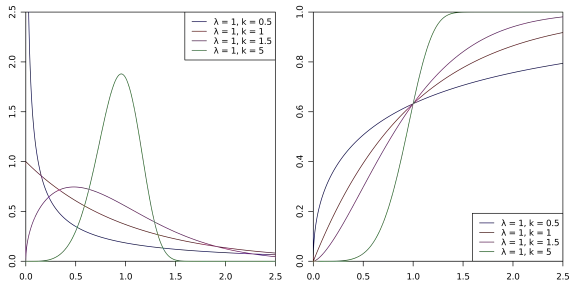




Weibull Distribution

Weibull distribution	$X \sim \text{Weibull}(\beta, \gamma), \quad \beta, \gamma > 0$
PDF	$\boxed{\frac{\gamma}{\beta^\gamma} x^{\gamma-1} e^{-(x/\beta)^\gamma}, \quad x \geq 0}$
CDF	$1 - e^{-1(x/\beta)^\gamma}, \quad x \geq 0$
Mean	$\mathbb{E}\left[X\right] = \beta\Gamma\left(\frac{\gamma+1}{\gamma}\right)$
Variance	$\text{Var}[X] = \beta^2 \left(\Gamma\left(\frac{\gamma+2}{\gamma}\right) - \left[\Gamma\left(\frac{\gamma+1}{\gamma}\right)\right]^2\right)$

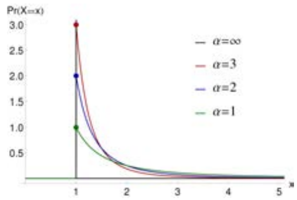
Special cases: When $\gamma = 1$, the Weibull distribution becomes an exponential distribution. When $\gamma = 2$, the tail behaviour of the Weibull distribution is the same as the *normal* distribution.



Pareto distribution

> Tail decay is a power *function* $\mathbb{P}\left(X > x\right) = \left(\frac{\alpha}{x}\right)^\gamma$

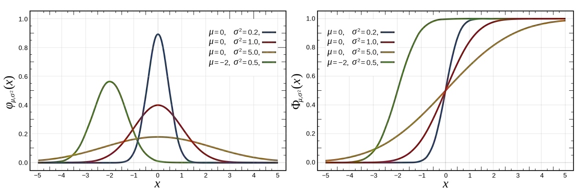
Weibull distribution	$X \sim \text{Pareto}(\alpha, \gamma), \quad \alpha, \gamma > 0$
PDF	$\boxed{\frac{\gamma\alpha^\gamma}{x^{\gamma+1}}, \quad x \geq \alpha}$
CDF	$1 - \left(\frac{\alpha}{x}\right)^\gamma, \quad x \geq \alpha$
Mean	$\mathbb{E}\left[X\right] = \frac{\gamma\alpha}{\gamma - 1}, \quad \gamma > 1$
Variance	$\text{Var}[X] = \frac{\gamma\alpha^2}{(\gamma-1)^2(\gamma-2)}, \quad \gamma > 2$



Normal distribution

Normal distribution	$X \sim \mathcal{N}(\mu, \sigma^2)$, $Z \sim \mathcal{N}(0,1)$
PDF	$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$
CDF	$\boxed{\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt}$
Standardisation	If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ If $Z \sim \mathcal{N}(0,1)$, then $X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$
Moments of Normal Distribution	$\boxed{\mathbb{E}\left[Z^n\right] = (n-1)\mathbb{E}\left[Z^{n-2}\right]} = \begin{cases} 0, & n \text{ odd} \\ \frac{(2k)!}{2^k (k)!}, & n = 2k \text{ is even} \end{cases}$
Closure under linear transformations	$X \sim \mathcal{N}(\mu, \sigma^2) \implies \boxed{aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)}$
Symmetry property of the Standard Normal CDF	$\Phi(-z) = 1 - \Phi(z)$
68–95–99.7 rule of thumb	$\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 0.6827$ $\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545$ $\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$

Symmetry properties of the Normal distribution: The Normal distribution is symmetric about its mean, and when centred about 0, an even function. Points of inflection occur at $\mu \pm \sigma$.



Log-normal distribution

> If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = e^X$, then Y is Lognormal (its logarithm is normal)

Lognormal distribution	$X \sim \mathrm{LN}(\mu, \sigma^2)$
PDF	$\frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2}, \quad x > 0$

6 Bivariate Random Variables

Bivariate Random Variables

Bivariate Random Variables	
Bivariate Random Variable (X, Y)	$(X, Y) : \Omega \to \mathbb{R}^2, \quad \omega \mapsto (X(\omega), Y(\omega))$ is a random *vector* from the sample space to points in $\mathbb{R}^2$
Range of (X, Y)	$S_{X,Y} = \{(X(\omega), Y(\omega)) : \omega \in \Omega)\} \subseteq \mathbb{R}^2$
Joint CDF	$F_{XY} : \mathbb{R}^2 \to [0, 1]$, $F_{XY}(x, y) = \mathbb{P}(\{X \le x\} \cap \{Y \le y\}) = \mathbb{P}(X \le x, Y \le y)$

Note:

- The joint probability is the probability that both X and Y take on a value simultaneously, whereas the marginal probability is the probability that X takes a given value regardless of what Y is (ranging over all possible values of Y for that given value of X)
- The joint cdf is a function defined on the entire $\mathbb{R}^2$ (i.e. for all $(x, y) \in \mathbb{R}^2$).

Discrete Bivariate Random Variables

> Discrete bivariate random variables are those for which $S_{X,Y}$ is countable. Equivalently, discrete bivariate random variables occur when X, Y are both discrete.

Joint PMFs	
Joint Probability Mass Function	$p_{XY} : S_{XY} \to [0,1], \quad p_{XY}(x,y) = \mathbb{P}(X = x, Y = y)$
Probability of a given event in the sample space	$\boxed{\mathbb{P}((X,Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x,y)}$
Non-negativity	$p_{X,Y}(x,y) \geq 0, \quad \forall (x,y) \in S_{X,Y}$
Unitarity	$\sum_{(x,y)\in S_{X,Y}} p_{X,Y}(x,y) = 1$
Joint CDF to joint PDF	$\boxed{\begin{aligned} p_{X,Y}(x,y) &= F_{X,Y}(x,y) - F_{X,Y}(x-1,y) \\ &- F_{X,Y}(x,y-1) + F_{X,Y}(x-1,y-1), \quad (x,y) \in \mathbb{Z}^2 \end{aligned}}$
Joint PMFs to *Marginal* pmf	$\boxed{p_X(x) = \sum_{y\in S_y} p_{X,Y}(x,y) \quad (x \in S_X), \quad p_Y(y) = \sum_{x \in S_X} p_{X,Y}(x,y) \quad (y \in S_Y)}$

Joint CDFs	$F_{X,Y} : \mathbb{R}^2 \to [0,1]$
Cumulative Distribution Function (cdf)	$F_{X,Y}(a,b) = \sum_{(x,y)\in S_{XY},\, x\leq a, y \leq b} p_{XY}(x,y)$
Probability of rectangular region	$\mathbb{P}(a < X \leq b, c < Y \leq d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c)$
Joint CDF to Marginal CDF	$\boxed{F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y), \quad F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y)}$

Note:

- You *cannot* (generally) recover the Joint pmf from the Marginal pmfs, (unless the random variables are *independent*)
- We often use a table to represent the joint pmf and the marginal pmf's. Add the last row and the last column to record the marginal pmf's of X and Y, respectively.

Y \ X	0	1	2	$P_Y(y)$
0	1/8	1/8	0	1/4
1	1/8	1/4	1/8	1/2
2	0	1/8	1/8	1/4
$P_X(x)$	1/4	1/2	1/4	

Continuous Bivariate Random Variables

Bivariate Joint PDFs	$f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$
Bivariate Joint PDF	$\boxed{\mathbb{P}\left((X,Y) \in D\right) = \iint_D f_{X,Y}(x,y)\,dx\,dy, \quad D \subseteq \mathbb{R}^2}$
Probability of a rectangular region	$\mathbb{P}(a \leqslant X \leqslant b, c \leqslant Y \leqslant d) = \int_a^b \int_c^d f_{X,Y}(x,y)\,dy\,dx = \int_c^d \int_a^b f_{X,Y}(x,y)\,dx\,dy$
Unitarity	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$
Joint CDF to joint PDF	$\boxed{f_{X,Y} = \frac{\partial^2}{\partial y \partial x} F_{X,Y} = \frac{\partial^2}{\partial x \partial y} F_{X,Y}}$
Joint PDFs to Marginal PDFs	$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,d\color{red}{y}, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,d\color{red}{x}}$

Warning:

- The joint pdf uniquely determines *the* marginal pdf's, but the converse is <u>not</u> true. Marginal pdf's do <u>not</u> determine the joint distribution.
- To find the marginal PDFs from the Joint PDFs, we integrate the *other* variable out, meaning that our bounds on our integral will be in terms of the *other* variable, and not in terms of the marginal variable.

Recovering joint pdf from joint cdf by partial differentiation

$$F_{X,Y} = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du$$

Let the inner integral $\varphi(u;y) = \int_{-\infty}^{y} f_{X,Y}(u,v)\,dv$, then $F_{X,Y} = \int_{-\infty}^{x} \varphi(u;y)\,du$. Thus, $\frac{\partial}{\partial x} F_{X,Y}(x,y) = \varphi(x;y) = \int_{-\infty}^{y} f_{X,Y}(x,v)\,dv$, and thus $\frac{\partial^2}{\partial x \partial y} F_{X,Y} = f_{X,Y}$.

Bivariate Joint CDFs	$F_{X,Y} : \mathbb{R}^2 \to [0,1]$
Bivariate Joint CDF	$F_{X,Y}(x,y) = \mathbb{P}\left(X \leq x, Y \leq y\right) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\,du\,dv$
Non-negativity	$F_{X,Y}(x,y) \geq 0, \quad \forall (x,y) \in \mathbb{R}^2$
Limit to infinity	$\lim_{x,y \to \infty} F_{X,Y}(x,y) = 1$

Conditional Bivariate distributions

Hint: Treat the given conditions as generating *another* random variable with its own distribution functions.

Discrete Conditional PMFs	$p_{X\|Y}(\cdot \mid y) : S_X \to [0,1]$ **and** $p_{Y\|X}(\cdot \mid x) : S_Y \to [0,1]$
Conditional PMF	$\boxed{p_{X\|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)},\ x \in S_X}$ and $\boxed{p_{Y\|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)},\ y \in S_Y}$
Conditional probability of event	$\mathbb{P}(X \in A \mid Y = y) = \sum_{x \in A} p_{X\|Y}(x \mid y)$

Note: In the conditional formula we think of x as a variable and of y as <u>fixed</u> with $y \in S_Y$ (as previously defined) and $p_Y(y) \neq 0$.

Continuous Conditional PDFs	$f_{X\|Y}, f_{Y\|X} : \mathbb{R} \to [0, \infty)$
Conditional PDF	$\boxed{f_{X\|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, x \in \mathbb{R}}$ and $\boxed{f_{Y\|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, y \in \mathbb{R}}$
Conditional probability of event	$\mathbb{P}(X \in A \| Y = k) = \int_A f_{X\|Y}(x \mid k)\, dx$

Independence of Random Variables

Independence of Random Variables	
Theoretical definition of Independent random variables	$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(X \in B)$, for <u>all</u> choices of subsets $A, B \subseteq \mathbb{R}$
Joint and Marginal CDFs	$F_{X,Y}(x,y) = F_X(x)F_Y(y), \quad \forall x, y \in \mathbb{R}$
Discrete (Joint) PMF	$\boxed{\begin{array}{l} p_{X,Y}(x,y) = p_X(x)p_Y(y) \\ \equiv p_{X\mid Y}(x \mid y) = p_X(x) \\ \equiv p_{Y\mid X}(y \mid x) = p_Y(y) \end{array}}$
Continuous (Joint) PDF	$\boxed{\begin{array}{l} f_{X,Y}(x,y) = f_X(x)f_Y(y) \\ \equiv f_{X\mid Y}(x \mid y) = f_X(x) \\ \equiv f_{Y\mid X}(y \mid x) = f_Y(y) \end{array}}$
Independence of n Random Variables	$\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n)$, for <u>all</u> choices of subsets $A_, \ldots, A_n \subseteq \mathbb{R}$
Independence of a *sequence* of Random Variables	A sequence $\{X_1, X_2, X_3, \ldots\}$ of random variables are said to be independent if $X_1, \ldots, X_n$ are independent <u>for each</u> $n \geq 1$.

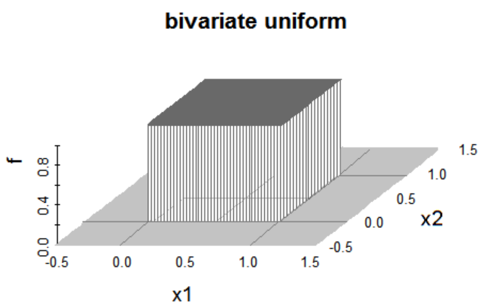
Intuition: The distribution of X should remain unaffected regardless of how much information we know about the values of Y (that is, the *conditional* probability is the same as *marginal* probability).

<u>Warning:</u> You need to check for <u>ALL</u> choices of x, y, including where the cdf/pdf/pmf takes on a *zero* value. Use piecewise functions carefully!

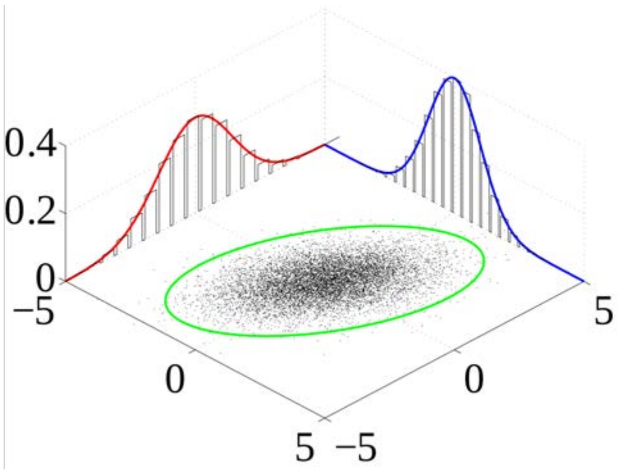
Bivariate distributions

Bivariate Uniform distributions

Geometric intuition: Let (X, Y) be a "uniform random point" within the unit square, then $\mathbb{P}((X,Y) \in A) = \dfrac{\text{area of A}}{\text{area of D}}$



Bivariate Normal distributions



***Standard* bivariate normal distribution**	
Standard bivariate normal	$(X,Y) \sim \mathcal{N}_2(\rho), \quad \rho(-1,1)$
Joint PDF	$f_{X,Y}(x,y) = \dfrac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\dfrac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$
Marginal distributions of $\mathcal{N}_2(\rho)$	$\boxed{X, Y \sim \mathcal{N}(0,1)}$ $\boxed{f_X(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}, \text{ and } f_Y(y) = \dfrac{1}{\sqrt{2\pi}} e^{-y^2/2}}$
Conditional distributions of $\mathcal{N}_2(\rho)$	$\boxed{X \mid_{Y=y} \sim \mathcal{N}\left(\rho y, 1-\rho^2\right), \text{ and } Y \mid_{X=x} \sim \mathcal{N}\left(\rho x, 1-\rho^2\right)}$
Conditional expectations	$\mathbb{E}[X \mid Y = y] = \rho y$
Standardisations	$X_s \mid Y = y = \left(\dfrac{X - \mu_X}{\sigma_X} \mid Y = y\right) \sim \mathcal{N}(0,1)$

Warning: $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ does <u>not</u> imply that $(X,Y) \sim \mathcal{N}_2(\rho)$ is jointly normal

***General* bivariate normal distributions**	
General bivariate normal distributions	$(X,Y) \sim \mathcal{N}_2\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho \in [-1,1]\right)$ where $\left(\dfrac{X-\mu_X}{\sigma_X}, \dfrac{Y-\mu_Y}{\sigma_Y}\right) \sim \mathcal{N}_2(\rho)$
Joint pdf	$f_{X,Y}(x,y) = \dfrac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\dfrac{1}{2(1-\rho^2)}\left(\dfrac{(x-\mu_X)^2}{\sigma_X^2} - \dfrac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \dfrac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right)$
Marginal distributions of general $\mathcal{N}_2$	$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$
Conditional distributions of general $\mathcal{N}_2$	$X\mid_{Y=y} \sim N\left(\rho\sigma_X y_s + \mu_X, \sigma_X^2\left(1-\rho^2\right)\right)$, where $x_s = \dfrac{x-\mu_X}{\sigma_X}, y_s = \dfrac{y-\mu_Y}{\sigma_Y}$ $Y\mid_{X=x} \sim N\left(\rho\sigma_Y x_s + \mu_Y, \sigma_Y^2\left(1-\rho^2\right)\right)$
Linear transformation	If (X,Y) is a bivariate normal random variable and $a,b,c,d \in \mathbb{R}$, then $(aX+bY, cX+dY)$ is also a bivariate normal random variable.
Decomposition of the bivariate normal	$\boxed{X = \mu_X + \dfrac{\sigma_X\rho}{\sigma_Y}(Y-\mu_Y) + \sigma_X\sqrt{1-\rho^2}Z}$
Decomposition of the univariate normal	$\boxed{X = \sigma_X Z + \mu_X}$
Constructing bivariate normals from standard normals Z_1, Z_2	$(X_1 = Z_1,\ X_2 = \rho Z_1 + \sqrt{1-\rho^2}Z_2) \sim \mathcal{N}_2(\rho)$

Geometry of *Standard* Bivariate normal distributions

- The peak of the graph is at the origin $(0,0)$
- The value of $f_{X,Y}(x,y)$ decays to zero in all directions as (x,y) approaches infinity.
- positive relationship if $\rho > 0$, negative relationship if $\rho < 0$
- As $|\rho| \to 1$, the relationship between X and Y becomes stronger.

Marginal distributions from Standard bivariate normal distribution

- It is in general not possible to reconstruct the joint distribution from the marginal distributions.
- We know that the marginal distributions of $N_2(\rho)$ are both standard normal distribution. However, the converse is not true in general – the *joint distribution* of two standard normal random variables needs not be bivariate normal.
- In particular, note that $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ does not imply that $(X,Y) \sim \mathcal{N}_2(\rho)$.

Sufficient condition to determine whether (X,Y) is a bivariate normal random variable:

- If $aX + bY$ is a normal random variable for <u>any</u> choice of $a, b \in \mathbb{R}$, then (X, Y) is a bivariate normal random variable.

Note:

- We can include the cases $\rho = \pm 1$ in the definition of standard bivariate normal distributions. When $\rho = 1$, we have $X \sim N(0,1), Y = X$. When $\rho = -1$, we have $X \sim N(0,1), Y = -X$.
- *Normally*, $\text{Cov(X, Y)} = 0 \not\Longrightarrow \text{ X} \perp \text{Y}$, but in the special case where $(X, Y) \sim \mathcal{N}_2(\rho)$, then if the covariance is zero, the random variables are independent (Can be seen since covariance is zero means the correlation coefficient ρ is zero, and thus we can factorise the joint pdf into the marginal pdfs for X, Y and thus $X \perp Y$).

Expectation of functions of Bivariate random variables

Expectation of functions of Bivariate random variables	$\psi : \mathbb{R}^2 \to \mathbb{R}$
Expectation of functions of Bivariate random variables	$\boxed{\mathbb{E}[\psi(X,Y)] = \begin{cases} \displaystyle\sum_{(x,y) \in S_{X,Y}} \psi(x,y)\, p_{X,Y}(x,y) \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(x,y)\, f_{X,Y}(x,y)\, dx\, dy \end{cases}}$
Linearity property of expectation	$\boxed{\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]}$
Expectation of products (<u>independent</u>)	$\boxed{X \perp Y \implies \mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]}$ $X_1, \ldots, X_n$ independent then $\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n]$

Warning: The converse does not imply, i.e. if $\mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[XY]$, this does not imply that X, Y are independent.

Conditional Expectation and Variance

<u>Warning</u>: There are different *types* of conditional expectation, be extremely careful not to confuse them.

- The conditional expectation $\mathbb{E}[X \mid Y]$ conditioning on a random variable is <u>itself a random variable</u>
- The conditional expectation $\mathbb{E}[X \mid Y = y]$ conditioning on a *fixed value of a random variable* is a fixed value in terms of the value of y.
- The conditional expectation $\mathbb{E}[X \mid A]$ conditioning on an event A, using the distribution and probability functions generated by the even

Conditional Expectation

Conditioning on a random variate	$\boxed{\mathbb{E}[X \mid Y = y] = \begin{cases} \displaystyle\sum_{x \in S_X} x\, p_{X\mid Y}(x \mid y) \\ \displaystyle\int_{-\infty}^{\infty} x\, f_{X\mid Y}(x \mid y)\, dx \end{cases}}$
Conditioning on a random variable	Define $\eta(y) = \mathbb{E}[X \mid Y = y]$ (as above), a number in terms of y. Then $\mathbb{E}[X \mid Y] = \eta(Y)$, the function composed with the random variable Y, itself a random variable
Law of Total Expectation	$\boxed{\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]}$
Law of Total Expectation (applied)	Multivariate case: $\boxed{\mathbb{E}[\psi(X,Y)] = \mathbb{E}[\mathbb{E}[\psi(X,Y) \mid Y]]}$ Products: $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY \mid Y]] = \mathbb{E}[Y\,\mathbb{E}[X \mid Y]]$ In terms of probability functions: $\boxed{\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \sum_y \mathbb{E}[X \mid Y = y]\,\mathbb{P}(Y = y)}$
Conditional expectation of an event A	$\boxed{\mathbb{E}[X \mid A] = \begin{cases} \displaystyle\sum_{x \in S_X} x p_{X\mid A}(x) \\ \displaystyle\int_{-\infty}^{\infty} x\, f_{X\mid A}(x)\, dx \end{cases}}$

Conditional Variance

Law of Total Variance	$\boxed{\mathrm{Var}[X] = \mathrm{Var}[\mathbb{E}[X \mid Y]] + \mathbb{E}[\mathrm{Var}[X \mid Y]]}$
Conditional variance of an event A	$\boxed{\mathrm{Var}[X \mid A] = \mathbb{E}\left[X^2 \mid A\right] - (\mathbb{E}[X \mid A])^2}$

Convolutions of independent random variables

Convolutions of *independent* random variables	$\psi(X,Y) = X + Y$
Discrete convolution formula	$\boxed{p_{X+Y}(k) = \sum_{i=0}^{k} p_X(i) p_Y(k-i), \quad k \in \{0, 1, 2\ldots\}}$ (Where X, Y are independent and non-negative)
Continuous convolution formula	$\boxed{f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx = \int_{-\infty}^{\infty} f_Y(y) f_X(z-y)\, dy}$

Warning: Careful with convolution limits or terminals in the continuous case.

Applications of convolutions	$\psi(X,Y) = X+Y$
Poisson rvs	$X \sim \text{Pn}(\lambda), Y \sim \text{Pn}(\mu) \implies X+Y \sim \text{Pn}(\mu+\lambda)$
Binomial rvs	$X \sim \text{Bi}(n,p), Y \sim \text{Bi}(m,p) \implies X+Y \sim \text{Bi}(n+m,p)$
Negative binomial rvs	$X \sim \text{NB}(r,p), Y \sim \text{NB}(s,p) \implies X+Y \sim \text{NB}(r+s,p)$
Exponential rvs	$X, Y \sim \exp(\alpha) \implies X+Y \sim \gamma(2,\alpha)$
Gamma rvs	$X \sim \gamma(r,\alpha), Y \sim \gamma(s,\alpha) \implies X+Y \sim \gamma(r+s,\alpha)$

Derivation for discrete convolution formula: Consider the decomposition $\{X+Y=k\} = \bigcup_{i=0}^{k}\{X=i, Y=k-i\}$, then by finite additivity and independence,

$$\begin{aligned}\mathbb{P}(X+Y=k) &= \mathbb{P}\left(\bigcup_{i=0}^{k}\{X=i, Y=k-i\}\right) \\ &= \sum_{i=0}^{k}\mathbb{P}(X=i, Y=k-i) \\ &= \sum_{i=0}^{k}\mathbb{P}(X=i)\,\mathbb{P}(Y=k-i)\end{aligned}$$

Decomposition of Binomial and Negative Binomial as sum of independent Bernoulli and Geometric random variables, respectively:

- Intuitively, $Bi(n,p)$ is the sum of n independent $B(1,p)$-random variables.
- Similarly, $N(r,p)$ is the sum of r independent $G(p)$-random variables.

Covariance

Covariance

Covariance	$\boxed{\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]}$
Variance of a sum	$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}[Y]$
Variance in terms of covariance	$\boxed{\mathrm{Var}[W] = \mathrm{Cov}[W,W]}$
Variance of a sum of *independent* rvs	$X \perp Y \implies \mathrm{Cov}(X,Y) = 0 \implies \mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$
Variance of linear transformation	$\mathrm{Var}[aX+bY] = a^2\,\mathrm{Var}[X] + 2ab\,\mathrm{Cov}(X,Y) + b^2\,\mathrm{Var}[Y]$
Positive/Negative Relationships	$\begin{cases} \mathrm{Cov}(X,Y) > 0 : \text{X,Y positively related} \\ \mathrm{Cov}(X,Y) < 0 : \text{X,Y negatively related} \\ \mathrm{Cov}(X,Y) = 0 : \text{X,Y uncorrelated} \end{cases}$
Bilinearity property	$\mathrm{Cov}(aX+bY, cX+dY) = ac\,\mathrm{Var}[X] + (ad+bc)\,\mathrm{Cov}(X,Y) + bd\,\mathrm{Var}[Y]$ $\mathrm{Cov}(X+Y, Z+W) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(X,W)\,\mathrm{Cov}(Y,Z) + \mathrm{Cov}(Y,W)$
Commutativity property	$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
Independence implies uncorrelated	$X \perp Y \implies \mathrm{Cov}(X,Y) = 0$
Uncorrelated does <u>not</u> imply independence	$\mathrm{Cov}(X,Y) = 0 \not\Longrightarrow X \perp Y$
Covariance in terms of correlation coefficien	$\mathrm{Cov}(X,Y) = \rho(X,Y)\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}$

Correlation coefficient

Correlation coefficient	$\boxed{\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} \in [-1,1]}$
Rescaling the correlation coefficient	$\rho(aX,bY) = \rho(X,Y)$

Variance of a sum:

$$\begin{aligned} V[X+Y] &= \mathbb{E}\left[\left((X+Y) - (\mu_X+\mu_Y)\right)^2\right] \\ &= \mathbb{E}\left[\left((X-\mu_X) + (Y-\mu_Y)\right)^2\right] \\ &= \mathbb{E}\left[(X-\mu_X)^2\right] + 2\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] + \mathbb{E}\left[(Y-\mu_Y)^2\right] \\ &= V[X] + 2\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] + V[Y] \end{aligned}$$

Warning:

- X, Y are independent implies X, Y are uncorrelated. However, the converse is <u>not</u> true, if $\mathrm{Cov}(X, Y) = 0$, then X, Y need not be independent.

7 Generating Functions

Probability generating function (pgf)

> Power series whose coefficients are given by the pmf of a **discrete non-negative** random variable

Probability generating function (pgf)	$S_X = \{0, 1, 2 \ldots\}$
Probability generating function	$\boxed{P_X(z) := \mathbb{E}\left[z^X\right] = \sum_{k=0}^{\infty} p_X(k)\, z^k}$
pmf from pgf (Inversion formula)	$\boxed{p_X(k) = \frac{P_X^{(k)}(0)}{k!}}$
Expectation	$\mathbb{E}[X] = P'_X(1)$
Variance	$\mathrm{Var}[X] = P''_X(1) + P'_X(1) - [P'_X(1)]^2$
Convolution theorem for PGFs	$X \perp Y \implies \boxed{P_{X+Y}(z) = P_X(z) P_Y(z)}$
Linear combination	$P_{aX+b}(z) = P_X(z^a) z^b$
Relationship between pgf and mgf	$P_X(z) = M_X(\ln z)$
Uniqueness theorem	If X, Y have an identical PGF (in some interval containing 0), then X, Y are identically distributed.

Properties

- The k-th coefficient of $P_X(k)$ is *exactly* the probability mass function $p_X(k)$.
- $P_X(1) = 1$ (since probabilities sum to 1; useful for finding scale coefficients for the pgf)
- $P_X(z)$ is well defined when $|z| \leq 1$, for any z with $|z| \leq 1$, $|P_X(z)| \leq 1$
- The Taylor expansion of PGF around $z = 0$ is $P_X(z) = \sum_{k=0}^{\infty} \frac{P_X^{(k)}(0)}{k!} z^k$

Random Sums using the pgf

Let $\{X_i : 1 \leq i \leq N\}$ be a sequence of N independent identically distributed random variables, where N is an independent non-negative discrete random variable. Let $S_N = \sum_{I=1}^{N} X_i$ be the random sum. Then the generating function of S_N is $G_{S_N}(z = G_N(G_X(z)))$; since
$G_{S_N}(z) = \mathbb{E}\left[z^{S_N}\right] = \mathbb{E}\left[z^{\sum_{i=1}^N X_i}\right] = \mathbb{E}\left[\mathbb{E}\left[z^{\sum_{i=1}^N X_i} \mid N\right]\right] = \mathbb{E}\left[(G_X(z))^N\right] = G_N(G_X(z))$, by definition of G_{S_N}, definition of S_N, law of total expectation, using the convolution theorem, and definition of G_X, G_N, respectively.

Moment generating function (mgf)

> *Arbitrary* random variable. The k-th coefficient of the mgf is the k-th moment of X

Moment generating function (mgf)	dom $M_X = T = \{t \in \mathbb{R} : \mathbb{E}\left[e^{tX}\right] < \infty\}$
Moment generating function (mgf)	$\boxed{M_X(t) := \mathbb{E}\left[e^{tX}\right] = \begin{cases} \sum_{x \in S_X} e^{tx}\, p_X(x) \\ \int_{-\infty}^{\infty} e^{tx}\, f_X(x)\, dx \end{cases}}$
Uniqueness theorem	Let X, Y be two random variables. Suppose that $M_X(t), M_Y(t)$ are both well defined and equal *in some neighbourhood of the origin* $t = 0$. Then X and Y have the same distribution.
Relation between the pgf and the mgf	$M_X(t) = P_X(e^t)$
Convolution theorem for mgfs	$X \perp Y \implies M_{X+Y}(t) = M_X(t)M_Y(t)$
Linear transformation	$M_{aX+b}(t) = M_X(at)e^{bt}$
Moments via Taylor expansion	$M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}\left[X^n\right] \frac{t^n}{n!}$
Computing *moments*	$\boxed{\mathbb{E}\left[X^n\right] = M_X^{(n)}(0)}$
Computing *central moments*	$\mathbb{E}\left[(X-\mu)^n\right] = \frac{\mathrm{d}^n}{\mathrm{d}t^n}(e^{-\mu t} M_X(t))\,\Big\vert_{t=0}$
Expectation	$\mathbb{E}\left[X\right] = M_X'(0)$
Variance	$\mathrm{Var}[X] = M_X''(0) - [M_X'(0)]^2$

Notes on the Domain

- By definition, the mgf is *always positive*.
- In addition, we have $\boxed{M_X(0) = 1}$ (zeroth moment is unitary) and thus $0 \in T$.
- It is typical that T is an interval containing the origin $t = 0$.
- However, there are examples where the mgf is *only* well defined for $t = 0$.
- Remember to specify the domain of the mgf carefully! Including where $t = 0$.

Computing moments from mgf

- We can compute moments of X by differentiating the mgf
- We can also achieve this by writing down the Taylor expansion of $M_X(t)$ directly.

Cumulant generating function (cgf)

> *Arbitrary* random variable. The k-th coefficient of the mgf is the k-th moment of X. The first cumulant is the mean, the second cumulant is the variance, and the third cumulant is the same as the third central moment. But fourth and higher-order cumulants are neither moments nor central moments, but rather more complicated polynomial functions of the moments.

Cumulant generating function (cgf)	$\operatorname{dom} K_X = T = \{t \in \mathbb{R} : \mathbb{E}\left[e^{tX}\right] < \infty\}$
Cumulant generating function	$\boxed{K_X(t) := \ln \mathbb{E}\left[e^{tX}\right] = \ln M_X(t)}$
Relation between the mgf and the cgf	$M_X(t) = e^{K_X(t)}$
Convolution theorem for cgfs	$X \perp Y \implies K_{X+Y}(t) = K_X(t) + K_Y(t)$
Linear transformation	$M_{aX+b}(t) = M_X(at)e^{bt}$
Cumulants via Taylor expansion	$K_X(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}$
Cumulants	$\boxed{\kappa_n = K_X^{(n)}(0)}$
Expectation	$\mathbb{E}[X] = M'_X(0)$
Variance	$\operatorname{Var}[X] = M''_X(0) - [M'_X(0)]^2$

Properties of cumulants

- **Invariance**: $\kappa_1(X + c) = \kappa_1(X) + c$
- **Equivariance**: $\kappa_n(X + c) = \kappa_n(X), n \geq 2$
- **Homogeneity**: $\kappa_n(aX) = a^n \kappa_n(X)$

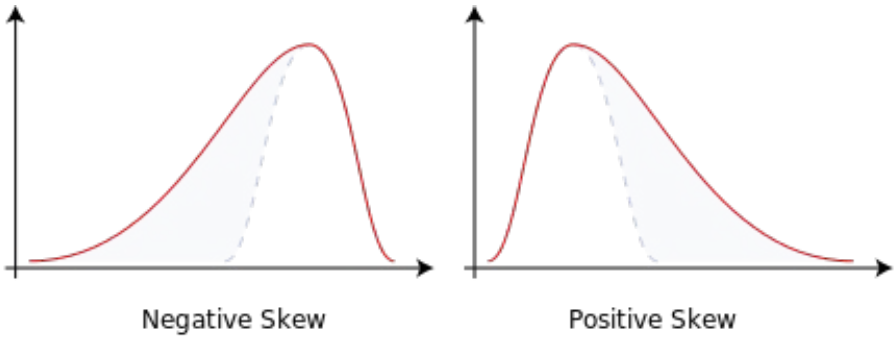
First few cumulants

1. $\kappa_1 = \mu_1 = \mathbb{E}[X]$
2. $\kappa_2 = \text{Var}[X] = \sigma^2 = \mu_2 - \mu_1^2$
3. $\kappa_3 = \mathbb{E}\left[(X - \mathbb{E}[X])^3\right] = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3$
4. $\kappa_4 = \mathbb{E}\left[(X - \mathbb{E}[X])^4\right] - 3\sigma^4 = \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4$

Cumulants of Normal Distribution $X \sim \mathcal{N}(\mu, \sigma^2)$:

- $\kappa_1 = \mu, \kappa_2 = \sigma^2, \kappa_r = 0, r \geq 3$

Skewness and Kurtosis		
Coefficient of Skewness	$\text{Skew}(X) = \frac{\kappa_3}{\sigma^3} = \mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sigma}\right)^3\right]$	Skewness reflects the level of symmetry of the distribution around its mean. If the pdf of X is symmetric about its mean, then $\text{Skew}(X)$ is identically zero.
Coefficient of Kurtosis	$\text{Kurt}(X) = \frac{\kappa_4}{\sigma^4} = \mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sigma}\right)^4\right] - 3$	For a flatter and shorter tailed distribution than $\mathcal{N}(0,1)$, the kurtosis is negative; while for a more peaked and longer tailed distribution than $\mathcal{N}(0,1)$, the kurtosis is positive. Kurtosis reflects the 'heaviness' of the tail.



The Generating functions of common distributions

Warning: Careful with parameterisations!!

Distribution	PGF	MGF	CGF
Bernoulli $X \sim \mathrm{Bi}(1,p)$	$1-p+pz, \quad z \in \mathbb{R}$	$1-p+pe^t$	$\ln(1-p+pe^t)$
Binomial $X \sim \mathrm{Bi}(n,p)$	$(1-p+pz)^n, \quad z \in \mathbb{R}$	$(1-p+pe^t)^n, \quad t \in \mathbb{R}$	$n\ln(1-p+pe^t)$
Geometric $X \sim \mathrm{G}(p)$	$\dfrac{p}{1+(p-1)z}, \quad \lvert z\rvert < \dfrac{1}{1-p}$	$\dfrac{p}{1+(p-1)e^t}$	
Negative Binomial $X \sim \mathrm{NB}(r,p)$	$p^r(1+(p-1)z)^{-r}, \quad \lvert z\rvert < \dfrac{1}{1-p}$	$\left(\dfrac{p}{1-(1-p)e^t}\right)^r, \quad t \in \left(-\infty, \ln\frac{1}{1-p}\right)$	
Poisson $X \sim \mathrm{Pn}(\lambda)$	$e^{\lambda(z-1)}, \quad z \in \mathbb{R}$	$e^{\lambda(e^t-1)}, \quad t \in \mathbb{R}$	$\lambda(e^t-1)$
Exponential $X \sim \exp(\alpha)$		$\dfrac{\alpha}{\alpha-t}, \quad t \in (-\infty, \alpha)$	
Gamma $X \sim \gamma(r,\alpha)$		$\left(\dfrac{\alpha}{\alpha-t}\right)^r, \quad t \in (-\infty, \alpha)$	
Standard Normal $Z \sim \mathcal{N}(0,1)$		$e^{t^2/2}, \quad t \in \mathbb{R}$	
Normal $X \sim \mathcal{N}(\mu,\sigma^2)$		$e^{\mu t+\sigma^2 t^2/2}, t \in \mathbb{R}$	$\mu t+\sigma^2 t^2/2$

8 Limit Theorems

Convergence in distribution

Convergence in distribution	$\{X_n : n \geq 1\}$
By distribution function	$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \implies X_n \xrightarrow{d} X$ (for all x at which the function $F_Y(x)$ is continuous)
By expectation property	$\lim_{n\to\infty} \mathbb{E}[\varphi(X_n)] = \mathbb{E}[\varphi(X)] \iff X_n \xrightarrow{d} X$ (for all bounded continuous functions $\varphi : \mathbb{R} \to \mathbb{R}$)
By the **mgf**	$\lim_{n\to\infty} M_{X_n}(t) = M_X(t), \quad \forall t \in (-\delta, \delta) \implies X_n \xrightarrow{d} X$ (where the mgf's of X_n and X are well defined in a common neighbourhood of the origin, say $(-\delta, \delta)$.)

Note: The converse of the above property is *not* true.

Law of Large Numbers

Let $\{X_n : n \geqslant 1\}$ be a sequence of independent and identically distributed random variables. Let $\mu = \mathbb{E}[X_1]$, which is the same for all X_n by assumption.

Let $S_n = X_1 + \cdots + X_n$ be the *partial sum*, and let $\bar{S}_n = \dfrac{S}{n}$ be the *sample average*. Then

$$\boxed{\bar{S}_n = \frac{S_n}{n} \xrightarrow{d} \mu \quad \text{as } n \to \infty}$$

Here the constant μ is viewed as the deterministic random variable $Y = \mu$

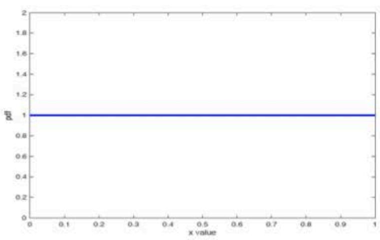
Note: The law of large numbers holds in more general contexts. For instance, the result holds if we only assume that the sequence of random variables are independent, and have the same mean μ and same variance σ^2, but not necessarily identically distributed. The convergence indeed holds in a much stronger sense than convergence in *distribution*.

Central Limit Theorem

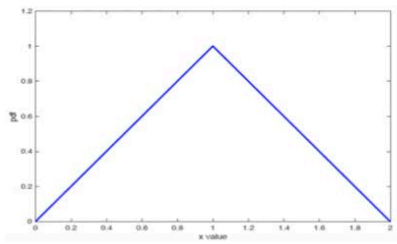
Central Limit Theorem	**Let $\{X_n : n \geqslant 1\}$ be a i.i.d sequence, and $S_n = X_1 + \cdots + X_n$ be the partial sum**
Central Limit Theorem	$$Z_n = \frac{S_n - \mathbb{E}\left[S_n\right]}{\sqrt{V\left[S_n\right]}} = \frac{\bar{S}_n - \mathbb{E}\left[\bar{S}_n\right]}{\sqrt{V\left[\bar{S}_n\right]}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as } n \to \infty$$
Distribution function	$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leqslant x\right) = \Phi(x), \quad \forall x \in \mathbb{R}$$

Intuitively, the central limit theorem asserts that when n is very large, the partial sum $S_n = X_1 + \cdots + X_n$ is approximated distributed like $\mathcal{N}(n\mu, n\sigma^2)$ regardless of what the individual distribution of X_n is.

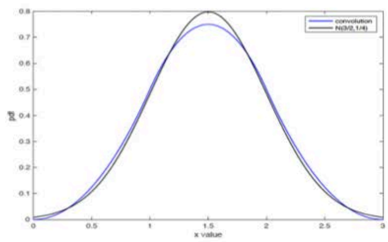
Example (Normal approximation to S_N, where $X_i \sim R(0,1)$)



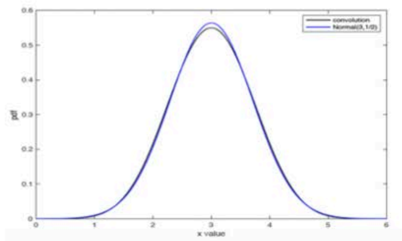
f_{X_1}



f_{S_2}



f_{S_3}



f_{S_6}

Applications of the CLT	
Normal approximation to general partial sum	$\boxed{S_n = X_1 + \cdots + X_n \approx \mathcal{N}(n\mu, n\sigma^2)}$
Normal approximation to Binomial	$\text{Bi}(n,p) \approx \mathcal{N}(np, np(1-p))$
Normal approximation to Gamma	$\gamma(n,\alpha) \approx \mathcal{N}(n/\alpha, n/\alpha^2)$ (X_n is iid $X_i \sim \exp(\alpha) \implies S_n \sim \gamma(n,\alpha)$)
Stirling's Formula	$\boxed{n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n}$

9 Stochastic Processes

Stochastic Processes	**A stochastic process is a family of random variables, $X(t), t \in T$.**
State space	For each $t \in T$, $X(t)$ takes values in a set S, called the state space of the stochastic process.
Index set	The set T is called the index set, with t usually denoting time.
Discrete-time stochastic process	If the index set T is discrete (e.g. $T = \{0, 1, 2, \ldots\}$). It is common practice to denote a discrete-time stochastic process by $\{X_n, n = 0, 1, 2, \ldots\}$.
Continous-time stochastic process	If the index set T is continuous (e.g. $T = [0, \infty]$)

Discrete-time Markov Chains

A **Discrete-Time Markov Chain (DSMC)** is a Stochastic Process such that $\{X_n : n \geq 0\}$, where

1. **Discreteness**: The state-space S is countable; and
2. **Markov Property:** For *any* $i, j, i_0 \ldots i_n \in S$,
   $\mathbb{P}(X_{n+1} = j \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$, that is, the *probability of the next state ($X_{n+1} = j$) is dependent only on the current state* $X_n = I$, and not on the past (
   $X_0 = i_0, X_1 = i_1 \ldots X_{n-1}$).
   - $\mathbb{P}(\text{ future } \mid \text{ past \& present }) = \mathbb{P}(\text{ future } \mid \text{ present })$

Time-homogenous Discrete-time Markov Chains

- $\forall i, j, m, n \quad \mathbb{P}(X_{n+m} = j \mid X_n = I) = \mathbb{P}(X_m = j \mid X_0 = I)$. That is, the m step transition probability depends on the *number* of steps and not on the starting point n.

Discrete-time Markov Chains	
One-step transition probability	$P_{ij} := \mathbb{P}\left(X_{n+1} = j \mid X_n = i\right)$
m-step transition probability	$P_{ij}^{(m)} = \mathbb{P}\left(X_{n+m} = j \mid X_n = i\right)$
Transition Matrix	$P_{\lvert S\rvert \times \lvert S\rvert} = [P_{ij}]_{i,j\in S} = [\mathbb{P}\left(X_{n+1} = j \mid X_n = i\right)]_{i,j\in S}$
Stochastic matrix	1. $P_{ij} \geq 0$ 2. $\sum_{j\in S} P_{ij} = 1, \quad \forall i \in S$

Note: The state space S can be infinite, in which case P has infinitely many rows and columns.

Equilibrium Distributions	
Initial distribution	$\boldsymbol{\pi}_0 = \left(\mathbb{P}\left(X_0 = 0\right), \mathbb{P}\left(X_0 = 1\right) \ldots\right)$
n-step distribution	$\boxed{\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 P^n}$
Law of Total Probability	$\boxed{\mathbb{P}\left(X_n = i\right) = \sum_k \mathbb{P}\left(X_0 = k\right) \mathbb{P}\left(X_n = i \mid X_0 = k\right) = \boldsymbol{\pi}_0 [P^n]_{\_,i}}$
Equilibrium distribution	$\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$, where $\boxed{\boldsymbol{\pi} = \boldsymbol{\pi} P}$ (The equilibrium distribution is the left-eigenvector of P with eigenvalue 1)
Limiting interpretation	$\boxed{\pi_j = \lim_{n\to\infty} P_{ij}^{(n)}}$ (The limiting probability of the stochastic process being in state j)
Stationary interpretation	$\boxed{\boldsymbol{\pi} = \boldsymbol{\pi} P}$ (If the process starts with probability distribution it will persist with that distribution *forever*.)
Ergodic interpretation	$\boldsymbol{\pi}$ Is the long term *relative frequencies* of being in state

Eigenvalue method for finding stationary distribution

1. Find an eigenvector $\boldsymbol{x} = (x_1 \ldots x_N)$ of $P^\intercal$ corresponding to eigenvalue $\lambda = 1$ (need not be unique)
2. Normalise $\mathbf{x}$ to make it a pmf, $\boldsymbol{\pi} = \left(\dfrac{x_1}{x_1 + \cdots + x_N}, \ldots, \dfrac{x_N}{x_1 + \cdots + x_N}\right)$

Limiting method for finding stationary distribution

- Decompose the transition matrix $P = KDK^{-1}$ where K is an invertible matrix of eigenvectors and D is a diagonal matrix of corresponding eigenvalues
- Find the stationary distribution by computing limit of matrix powers: $\lim_{n\to\infty} P^n = K(\lim_{n\to\infty} D^n)K^{-1}$, where the power of a diagonal matrix is simply the entries of the diagonal raised to the corresponding power

Linear equations method for finding stationary distribution

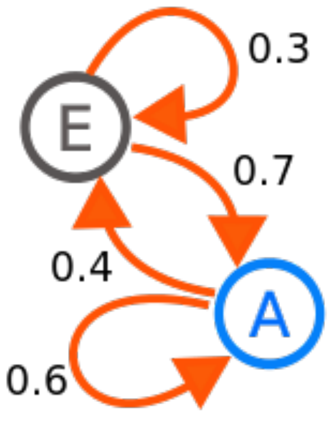
- Solve the linear system and normalising constraint $\begin{cases} \boldsymbol{\pi} = \boldsymbol{\pi} P \\ \pi_1 + \cdots + \pi_N = 1 \end{cases}$ (where 1 equation is linearly dependent)

State-space Diagrams

1. Each node in the graph represents a state in S
2. Each directed weighted edge is the transition between state i and j (loops are allowed)

<u>**Example:**</u>

The stat-space diagram generated by the square transition matrix $P = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$



1 Axioms of Probability

Probability Axioms	$\mathbb{P}(\cdot) : A \mapsto [0,1]$
1. Non-negativity	$\mathbb{P}(A) \geq 0, \quad \forall A \subseteq \Omega$
2. Unitarity	$\mathbb{P}(\Omega) = 1$
3. Countable additivity	$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, where $\{A_1, A_2, A_3, \ldots\}$ is a sequence of **mutually disjoint events**

Probability Properties		Proof
Probability of empty set	$\mathbb{P}(\varnothing) = 0$	$\varnothing \cup \varnothing \cup \cdots = \varnothing$ is mutually disjoint set, so by (3), $\mathbb{P}(\varnothing) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing) + \cdots + \mathbb{P}(\varnothing)$, which only holds if $\mathbb{P}(\varnothing) = 0$
Finite additivity	$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$, where $\{A_1, A_2, A_3, \ldots, A_n\}$ is a *finite* set of **mutually disjoint events**	Let $A_1, \ldots, A_n$ be mutually disjoint, construct $A_1, \ldots, A_n, \varnothing, \varnothing, \ldots$. Since $\mathbb{P}(\varnothing) = 0$ and by countable additivity, $\mathbb{P}(A_1 \cup \cdots \cup A_n \cup \varnothing) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n) + 0 + 0 \cdots$
Complement rule	$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$	$\mathbb{P}\left(A \cup A^C\right) = \mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$
Monotonicity	If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$	If $A \subseteq B$, then $B = A \cup (B \setminus A)$, which is disjoint, thus $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$ since probabilities are non-negative
Probability lies in unit interval	$0 \leq \mathbb{P}(A) \leq 1$	$\varnothing \subseteq A \subseteq \Omega$, thus $\mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) \implies 0 \leq \mathbb{P}(A) \leq 1$
Addition theorem	$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)}$	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B))$ And $\mathbb{P}(B) = \mathbb{P}(B \setminus (A \cap B)) + \mathbb{P}(A \cap B)$, eliminate $\mathbb{P}(B \setminus (A \cap B))$
Continuity	(i) Let $\{A_n : n \geq 1\}$ be an increasing sequence of events, namely, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, and define $A = \bigcup_{n=1}^{\infty} A_n$. Then $\mathbb{P}(A) = \lim_{n\to\infty} A_n$	(ii) Let $\{A_n : n \geq 1\}$ be a decreasing sequence of events, namely, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, and define $A = \bigcap_{n=1}^{\infty} A_n$. Then $\mathbb{P}(A) = \lim_{n\to\infty} A_n$

2 Conditional Probability

Independence of events	
Independence of two events	$A \perp B \equiv \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$
Pairwise independent of $\{A_i\}$	The events $\{A_i\}$ are *pairwise* independent if $A_i \perp A_j$ for any pair of $i \neq j$
Mutual independence of $\{A_i\}$	The events $\{A_i\}$ are *mutually* independent if for <u>any sub-collection</u> of $\{A_i\}$, $\mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_n}\right) = \mathbb{P}\left(A_{i_1}\right) \times \cdots \times \mathbb{P}\left(a_{i_n}\right)$
Property of mutually independent events	Let $\{A_i : 1 \leq i \leq n\}$ be mutually independent events. We can form new subsets of mutually independent events out of them by selecting <u>non-overlapping</u> events and performing arbitrary operations on them.

Conditional probability	
Conditional probability	$\boxed{\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}$
Multiplication theorem	$\mathbb{P}(A \cap B) = \mathbb{P}(B)\,\mathbb{P}(A \mid B)$
Positive and negative relations	$\begin{cases} \text{Positive relation} : \mathbb{P}(A \mid B) > \mathbb{P}(A) \\ \text{Negative relation} : \mathbb{P}(A \mid B) < \mathbb{P}(A) \end{cases}$

Law of Total Probability & Bayes' Theorem	
<u>Law of Total Probability</u>	$\boxed{\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(B_i)\,\mathbb{P}(A \mid B_i)}$, where $\{B_i\}$ is a *partition* of Ω
<u>Bayes' Theorem</u>	$\boxed{\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\sum_j \mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j)}}$, for a given partition $\{A_j\}$
Bayes' theorem for the partition $\Omega = A \cup B$	$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B \mid A)\,\mathbb{P}(A) + \mathbb{P}(B \mid A^c)\,\mathbb{P}(A^c)}$

3 Random Variables

Random Variables	
Probability Mass Functions	The probability mass function of X maps the outcomes of X to a probability. $p_X : S_X \to [0,1], p_X(x) = \mathbb{P}(X = x)$
Probability of a given event in the sample space	$\mathbb{P}(x \in A) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} p_X(x)$
Properties of a PMF	1. $p_X(x) \geq 0, \quad \forall x \in S_X$ 2. $\sum_{x \in S_X} p_X(x) = 1$
PMF from CDF	$\mathbb{P}(X = k) = F_X(k) - F_X(k-1)$, for $S_X \in \mathbb{Z}$
Cumulative Distribution Function (cdf)	$F_X : \mathbb{R} \to [0,1], \quad F_X(x) = \mathbb{P}(X \leq x) = \sum_{y \in S_x,\, y \leq x} p_X(y)$
CDF lies in unit interval	$0 \leq F_X(x) \leq 1$
Probability of $a < X \leq b$	$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$
Limits to infinity	$\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$
Probability density function	$f_X : \mathbb{R} \to [0, \infty), \quad f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$, the function that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$
Interpretation as a 'probability density' around the point x	$\mathbb{P}(X \approx x) \approx f_X(x)\delta$
Non-negativity	$f_X(x) \geq 0$ (since $F_X(x)$ is increasing function, so $F_X' \geq 0$; the derivative of an increasing function is non-negative)
Probability as area under f_x	$\mathbb{P}(a \leq x \leq b) = \int_a^b f_X(t)\,dt = F_X(b) - F_X(a)$
Area under f_x is 1	$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

Expectation and Variance	
Discrete random variables	$\mathbb{E}[X] = \sum_{x \in S_X} x \cdot p_X(x) = \sum_{\omega \in \Omega} X(\omega)\,\mathbb{P}(\omega)$
Continuous random variables	$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$
Expectation of *functions* of random variables $\psi(X) : \mathbb{R} \to \mathbb{R}$	$\mathbb{E}[\psi(X)] = \begin{cases} \sum_{x \in S_X} \psi(x)\,p_X(x) \\ \int_{-\infty}^{\infty} \psi(x)\,f_X(x)\,dx \end{cases}$
Linearity property of expectation	$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
Expectation via tail probabilities (for continuous **non-negative** rvs)	$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X > x)\;dx = \int_0^{\infty} 1 - F_X(x)\,dx$
Expectation via tail probabilities (for discrete rvs where $S_X \subseteq \mathbb{N}$)	$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n), \quad n \in \mathbb{N}$
Variance	$\boxed{\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2}$
Variance of linear function	$\mathrm{Var}[aX + b] = a^2\,\mathrm{Var}[X]$
Standard deviation	$\mathrm{sd}(X) = \sqrt{\mathrm{Var}[X]}$

Moments	$n \in \mathbb{N}$
Moment	$\boxed{\mu_n = \mathbb{E}[X^n]}$
Central moment	$\boxed{\nu_n = \mathbb{E}[(X - \mu_X)^n]}$
Moments via tail probabilities for $n \geq 1$ and **non-negative** X	$\mathbb{E}[X^n] = n \int_0^{\infty} x^{n-1}(1 - F_X(x))\,dx = n \int_0^{\infty} x^{n-1}\,\mathbb{P}(X > x)\;dx$
Discrete moments	$\mathbb{E}[X^n] = \sum_{x \in S_X} x^n p_X(x)$

Taylor Approximations and Inequalities	
Taylor Approximation for Expectation	$\boxed{\mathbb{E}[\psi(X)] \approx \psi(\mu) + \frac{1}{2}\psi''(\mu)\,\sigma_X^2}$
Taylor Approximation for Variance	$\boxed{\mathrm{Var}[X] \approx [\psi'(\mu)]^2\,\sigma_X^2}$
Chebyshev's inequality	$\boxed{\mathbb{P}(\lvert X - \mu\rvert \geq \lambda) \leq \frac{\sigma^2}{\lambda^2}}$ or $\mathbb{P}\left(\left\lvert\frac{X - \mu}{\sigma}\right\rvert \geq \lambda\right) \leq \frac{1}{\lambda^2}$
Markov's Inequality	$\boxed{\mathbb{P}(\lvert X\rvert \geq \lambda) \leq \frac{\mathbb{E}[\lvert X\rvert^k]}{\lambda^k}}$

4 Discrete Distributions

Distribution	PMF	State Space	Expectation	Variance
Bernoulli $X \sim \mathrm{Bi}(1,p)$	$p(x) = \begin{cases} p, & x=1 \\ 1-p, & x=0 \end{cases}$	$\Omega = \{S,F\}$ where $X(\omega) = \begin{cases} 1, & \omega = S \\ 0, & \omega = F \end{cases}$	$\mathbb{E}[X] = p$	$\mathrm{Var}[X] = p(1-p) = pq$
Binomial $X \sim \mathrm{Bi}(n,p)$	$\boxed{\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}}$ $, k \in \{0,1,2,\dots,n\}$	$\Omega = \{(x_1, x_2, \dots, x_n) : x_i \in \{S,F\}\}$ $S_X = \{0,1,2,3\dots\}$	$\mathbb{E}[X] = np$	$\mathrm{Var}[X] = np(1-p) = npq$
Geometric $X \sim G(p)$	$\boxed{\mathbb{P}(X=k) = (1-p)^k p}$, where $k \in \{0,1,2,3,\dots\}$	$S_X = \{0,1,2,3,\dots\}$	$\mathbb{E}[X] = \frac{1-p}{p}$	$\mathrm{Var}[X] = \frac{1-p}{p^2}$
Negative Binomial $X \sim \mathrm{NB}(r,p), \quad r>0$	$\boxed{p_X(k) = \binom{-r}{k} (p-1)^k p^r}$, where $k \in \{0,1,2,3,\dots\}$	$S_X = \{0,1,2,3,\dots\}$	$\mathbb{E}[X] = r\frac{(1-p)}{p}$	$\mathrm{Var}[X] = r\frac{(1-p)}{p^2}$
Hypergeometric $X \sim \mathrm{Hg}(n,D,N)$	$\boxed{p_X(k) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}}$, where $k \in \{0,1,2,\dots,n\}$	$S_X = \{0,1,2,\dots,n\}, n \le N$	$\mathbb{E}[X] = \frac{nD}{N}$	$\mathrm{Var}[X] = \frac{nD(N-D)}{N^2}\left(1 - \frac{n-1}{N-1}\right)$
Poisson $X \sim \mathrm{Pn}(\lambda), \quad \lambda > 0$	$\boxed{p_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}}$, where $k \in \{0,1,2,3,\dots\}$	$S_X = \{0,1,2,3,\dots\}$	$\mathbb{E}[X] = \lambda$	$\mathrm{Var}[X] = \lambda$
Discrete Uniform $X \sim \mathrm{U}(a,b), \quad a<b$	$\boxed{p_X(k) = \frac{1}{b-a+1}}$, where $k \in \{a, a+1, \dots, b\}$	$S_X = \{a, a+1, \dots, b\}$ (set of integers between a and b)	$\mathbb{E}[X] = \frac{a+b}{2}$	$\mathrm{Var}[X] = \frac{(b-a+1)^2 - 1}{12}$

5 Continuous Distributions

Gamma and Beta functions	
Gamma function	$\boxed{\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx, \quad r > 0}$
Recursive definition	$\boxed{\Gamma(r) = (r-1)\Gamma(r-1)}, \quad r > 0$ $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$ $\Gamma(n) = (n-1)!, \quad \forall n \in \mathbb{N}$
Beta function	$\boxed{B(\alpha,\beta) \equiv \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx, \quad \alpha, \beta > 0}$
Beta and Gamma function	$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Distribution	PDF	CDF	Expectation/Moments
Continuous Uniform $X \sim \mathrm{R}(a,b)$	$\boxed{f_X(x) = \frac{1}{b-a}, \quad a < x < b}$	$\boxed{F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}}$	$\mathbb{E}[X] = \frac{a+b}{2}$
Exponential $X \sim \exp(\lambda), \quad \lambda > 0$	$\boxed{f_X(t) = \lambda e^{-\lambda t}, \quad t \geqslant 0}$	$\boxed{F_X(t) = \mathbb{P}(X \leqslant t) = \begin{cases} 1 - e^{-\lambda t}, & t \geqslant 0 \\ 0, & t < 0 \end{cases}}$	$\mathbb{E}[X] = \frac{1}{\lambda}$
Gamma $X \sim \gamma(r,\alpha), \quad r, \alpha > 0$	$\boxed{f_X(t) = \frac{\alpha^r}{\Gamma(r)} e^{-\alpha t} t^{r-1}, \quad t \geqslant 0}$		$\mathbb{E}[X] = \frac{r}{\alpha}$ $\mathbb{E}[X^k] = \frac{\Gamma(r+k)}{\Gamma(r)\alpha^k}$
Beta $X \sim \mathrm{Beta}(\alpha,\beta), \quad \alpha,\beta > 0$	$\boxed{f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}}, \quad 0 \le x \le 1$		$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ $\mathbb{E}[X^k] = \frac{B(\alpha+k,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(a+k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}$ \$
Weibull $X \sim \mathrm{Weibull}(\beta,\gamma), \quad \beta,\gamma > 0$	$\boxed{\frac{\gamma}{\beta^\gamma} x^{\gamma-1} e^{-(x/\beta)^\gamma}, \quad x \ge 0}$	$1 - e^{-1(x/\beta)^\gamma}, \quad x \ge 0$	$\mathbb{E}[X] = \beta\Gamma\left(\frac{\gamma+1}{\gamma}\right)$
Pareto $X \sim \mathrm{Pareto}(\alpha,\gamma), \quad \alpha,\gamma > 0$	$\boxed{\frac{\gamma\alpha^\gamma}{x^{\gamma+1}}, \quad x \ge \alpha}$	$1 - \left(\frac{\alpha}{x}\right)^\gamma, \quad x \ge \alpha$	$\mathbb{E}[X] = \frac{\gamma\alpha}{\gamma-1}, \quad \gamma > 1$
Normal $X \sim \mathcal{N}(\mu,\sigma^2)$, $Z \sim \mathcal{N}(0,1)$	$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$	$\boxed{\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2}\, dt}$	$\boxed{\mathbb{E}[Z^n] = (n-1)\mathbb{E}[Z^{n-2}]} = \begin{cases} 0, & n \text{ odd} \\ \frac{(2k)!}{2^k (k)!}, & n = 2k \text{ is eve} \end{cases}$
Lognormal $X \sim \mathrm{LN}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2}, \quad x > 0$		

6 Bivariate Random Variables

Bivariate Random Variables	
Bivariate Random Variable (X,Y)	$(X,Y) : \Omega \to \mathbb{R}^2, \quad \omega \mapsto (X(\omega), Y(\omega))$ is a random *vector* from the sample space to points in $\mathbb{R}^2$
Range of (X,Y)	$S_{X,Y} = \{(X(\omega), Y(\omega)) : \omega \in \Omega)\} \subseteq \mathbb{R}^2$
Joint CDF	$F_{XY} : \mathbb{R}^2 \to [0,1]$, $F_{XY}(x,y) = \mathbb{P}(\{X \le x\} \cap \{Y \le y\}) = \mathbb{P}(X \le x, Y \le y)$
Joint Probability Mass Function	$p_{XY} : S_{XY} \to [0,1], \quad p_{XY}(x,y) = \mathbb{P}(X = x, Y = y)$
Probability of a given event in the sample space	$\boxed{\mathbb{P}((X,Y) \in A) = \sum_{(x,y) \in A} p_{X,Y}(x,y)}$
Non-negativity	$p_{X,Y}(x,y) \ge 0, \quad \forall (x,y) \in S_{X,Y}$
Unitarity	$\sum_{(x,y) \in S_{X,Y}} p_{X,Y}(x,y) = 1$
Joint CDF to joint PDF	$\boxed{\begin{aligned} p_{X,Y}(x,y) &= F_{X,Y}(x,y) - F_{X,Y}(x-1,y) \\ &\quad - F_{X,Y}(x,y-1) + F_{X,Y}(x-1,y-1), \quad (x,y) \in \mathbb{Z}^2 \end{aligned}}$
Joint PMFs to *Marginal* pmf	$\boxed{p_X(x) = \sum_{y \in S_y} p_{X,Y}(x,y) \quad (x \in S_X), \quad p_Y(y) = \sum_{x \in S_X} p_{X,Y}(x,y) \quad (y \in S_Y)}$
Cumulative Distribution Function (cdf)	$F_{X,Y}(a,b) = \sum_{(x,y) \in S_{XY},\, x \le a, y \le b} p_{XY}(x,y)$
Probability of rectangular region	$\mathbb{P}(a < X \le b, c < Y \le d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c)$
Joint CDF to Marginal CDF	$\boxed{F_X(x) = \lim_{y \to \infty} F_{X,Y}(x,y), \quad F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x,y)}$

Bivariate Joint PDFs and CDFs	$f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}, F_{X,Y} : \mathbb{R}^2 \to [0,1]$
Bivariate Joint PDF	$\boxed{\mathbb{P}((X,Y) \in D) = \iint_D f_{X,Y}(x,y)\, dx\, dy, \quad D \subseteq \mathbb{R}^2}$
Probability of a rectangular region	$\mathbb{P}(a \leqslant X \leqslant b, c \leqslant Y \leqslant d) = \int_a^b \int_c^d f_{X,Y}(x,y)\, dy\, dx = \int_c^d \int_a^b f_{X,Y}(x,y)\, dx\, dy$
Unitarity	$\int_{-\infty}^\infty \int_{-\infty}^\infty f_{X,Y}(x,y)\, dx\, dy = 1$
Joint CDF to joint PDF	$\boxed{f_{X,Y} = \frac{\partial^2}{\partial y \partial x} F_{X,Y} = \frac{\partial^2}{\partial x \partial y} F_{X,Y}}$
Joint PDFs to Marginal PDFs	$\boxed{f_X(x) = \int_{-\infty}^\infty f_{X,Y}(x,y)\, dy, \quad f_Y(y) = \int_{-\infty}^\infty f_{X,Y}(x,y)\, dx}$
Bivariate Joint CDF	$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u,v)\, du\, dv$
Non-negativity	$F_{X,Y}(x,y) \ge 0, \quad \forall (x,y) \in \mathbb{R}^2$
Limit to infinity	$\lim_{x,y \to \infty} F_{X,Y}(x,y) = 1$

Discrete/Continuous Conditional PMFs	$p_{X\mid Y}(\cdot \mid y) : S_X \to [0,1]$ and $p_{Y\mid X}(\cdot \mid x) : S_Y \to [0,1], f_{X\mid Y}, f_{Y\mid X} : \mathbb{R} \to [0,\infty)$
Conditional PMF	$\boxed{p_{X\mid Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \; x \in S_X}$ and $\boxed{p_{Y\mid X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}, \; y \in S_Y}$
Conditional probability of event	$\mathbb{P}(X \in A \mid Y = y) = \sum_{x \in A} p_{X\mid Y}(x \mid y)$
Conditional PDF	$\boxed{f_{X\mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, x \in \mathbb{R}}$ and $\boxed{f_{Y\mid X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, y \in \mathbb{R}}$
Conditional probability of event	$\mathbb{P}(X \in A \mid Y = k) = \int_A f_{X\mid Y}(x \mid k)\, dx$

Independence of Random Variables	
Theoretical definition of Independent random variables	$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(X \in B)$, for all choices of subsets $A, B \subseteq \mathbb{R}$
Joint and Marginal CDFs	$F_{X,Y}(x,y) = F_X(x)F_Y(y), \quad \forall x,y \in \mathbb{R}$
Discrete (Joint) PMF	$\boxed{\begin{aligned} &p_{X,Y}(x,y) = p_X(x)p_Y(y) \\ &\equiv p_{X\mid Y}(x \mid y) = p_X(x) \\ &\equiv p_{Y\mid X}(y \mid x) = p_Y(y) \end{aligned}}$
Continuous (Joint) PDF	$\boxed{\begin{aligned} &f_{X,Y}(x,y) = f_X(x)f_Y(y) \\ &\equiv f_{X\mid Y}(x \mid y) = f_X(x) \\ &\equiv f_{Y\mid X}(y \mid x) = f_Y(y) \end{aligned}}$
Independence of n Random Variables	$\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n)$, for all choices of subsets $A_, \dots, A_n \subseteq \mathbb{R}$
Independence of a *sequence* of Random Variables	A sequence $\{X_1, X_2, X_3, \dots\}$ of random variables are said to be independent if $X_1, \dots, X_n$ are independent for each $n \ge 1$.

Standard bivariate normal distribution	
Standard bivariate normal	$(X,Y) \sim \mathcal{N}_2(\rho), \quad \rho(-1,1)$
Joint PDF	$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$
Marginal distributions of $\mathcal{N}_2(\rho)$	$\boxed{X, Y \sim \mathcal{N}(0,1)}$ $\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \text{ and } f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}}$
Conditional distributions of $\mathcal{N}_2(\rho)$	$\boxed{X \mid_{Y=y} \sim \mathcal{N}(\rho y, 1-\rho^2), \text{ and } Y \mid_{X=x} \sim \mathcal{N}(\rho x, 1-\rho^2)}$
Conditional expectations	$\mathbb{E}[X \mid Y = y] = \rho y$
Standardisations	$X_s \mid Y = y = \left(\frac{X - \mu_X}{\sigma_X} \mid Y = y\right) \sim \mathcal{N}(0,1)$

General bivariate normal distributions	
General bivariate normal distributions	$(X,Y)\sim\mathcal{N}_2\left(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho\in[-1,1]\right)$ where $\left(\frac{X-\mu_X}{\sigma_X},\frac{Y-\mu_Y}{\sigma_Y}\right)\sim\mathcal{N}_2(\rho)$
Joint pdf	$f_{X,Y}(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2}-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right)$
Marginal distributions of general $\mathcal{N}_2$	$X\sim\mathcal{N}(\mu_X,\sigma_X^2)$ and $Y\sim\mathcal{N}(\mu_Y,\sigma_Y^2)$
Conditional distributions of general $\mathcal{N}_2$	$X\vert_{Y=y}\sim N\left(\rho\sigma_X y_s+\mu_X,\sigma_X^2\left(1-\rho^2\right)\right)$, where $x_s=\frac{x-\mu_X}{\sigma_X}, y_s=\frac{y-\mu_Y}{\sigma_Y}$ $Y\vert_{X=x}\sim N\left(\rho\sigma_Y x_s+\mu_Y,\sigma_Y^2\left(1-\rho^2\right)\right)$
Linear transformation	If (X,Y) is a bivariate normal random variable and $a,b,c,d\in\mathbb{R}$, then $(aX+bY,cX+dY)$ is also a bivariate normal random variable.
Decomposition of the bivariate normal	$\boxed{X=\mu_X+\frac{\sigma_X\rho}{\sigma_Y}(Y-\mu_Y)+\sigma_X\sqrt{1-\rho^2}Z}$
Decomposition of the univariate normal	$\boxed{X=\sigma_X Z+\mu_X}$
Constructing bivariate normals from standard normals Z_1,Z_2	$(X_1=Z_1,\ X_2=\rho Z_1+\sqrt{1-\rho^2}Z_2)\sim\mathcal{N}_2(\rho)$

Expectation of functions of Bivariate random variables	$\psi:\mathbb{R}^2\to\mathbb{R}$
Expectation of functions of Bivariate random variables	$\boxed{\mathbb{E}[\psi(X,Y)]=\begin{cases}\sum_{(x,y)\in S_{X,Y}}\psi(x,y)\,p_{X,Y}(x,y)\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(x,y)\,f_{X,Y}(x,y)\,dx\,dy\end{cases}}$
Linearity property of expectation	$\boxed{\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]}$
Expectation of products (<u>independent</u>)	$\boxed{X\perp Y\implies\mathbb{E}[XY]=\mathbb{E}[X]\,\mathbb{E}[Y]}$ $X_1,\dots,X_n$ independent then $\mathbb{E}[X_1\cdots X_n]=\mathbb{E}[X_1]\cdots\mathbb{E}[X_n]$

Conditional Expectation and Variance	
Conditioning on a random variate	$\boxed{\mathbb{E}[X\mid Y=y]=\begin{cases}\sum_{x\in S_X}x\,p_{X\mid Y}(x\mid y)\\ \int_{-\infty}^{\infty}x\,f_{X\mid Y}(x\mid y)\,dx\end{cases}}$
Conditioning on a random variable	Define $\eta(y)=\mathbb{E}[X\mid Y=y]$ (as above), a number in terms of y. Then $\mathbb{E}[X\mid Y]=\eta(Y)$, the function composed with the random variable Y, itself a random variable
<u>Law of Total Expectation</u>	$\boxed{\mathbb{E}[X]=\mathbb{E}[\mathbb{E}[X\mid Y]]}$
Law of Total Expectation (applied)	Multivariate case: $\boxed{\mathbb{E}[\psi(X,Y)]=\mathbb{E}[\mathbb{E}[\psi(X,Y)\mid Y]]}$ Products: $\mathbb{E}[XY]=\mathbb{E}[\mathbb{E}[XY\mid Y]]=\mathbb{E}[Y\,\mathbb{E}[X\mid Y]]$ In terms of probability functions: $\boxed{\mathbb{E}[X]=\mathbb{E}[\mathbb{E}[X\mid Y]]=\sum_y\mathbb{E}[X\mid Y=y]\,\mathbb{P}(Y=y)}$
Conditional expectation of an <u>event</u> A	$\boxed{\mathbb{E}[X\mid A]=\begin{cases}\sum_{x\in S_X}x\,p_{X\mid A}(x)\\ \int_{-\infty}^{\infty}x\,f_{X\mid A}(x)\,dx\end{cases}}$
<u>Law of Total Variance</u>	$\boxed{\operatorname{Var}[X]=\operatorname{Var}[\mathbb{E}[X\mid Y]]+\mathbb{E}[\operatorname{Var}[X\mid Y]]}$
Conditional variance of an <u>event</u> A	$\boxed{\operatorname{Var}[X\mid A]=\mathbb{E}\left[X^2\mid A\right]-(\mathbb{E}[X\mid A])^2}$

Convolutions of *independent* random variables	$\psi(X,Y)=X+Y$
Discrete convolution formula	$\boxed{p_{X+Y}(k)=\sum_{i=0}^{k}p_X(i)p_Y(k-i),\quad k\in\{0,1,2\dots\}}$ (Where X,Y are independent and non-negative)
Continuous convolution formula	$\boxed{f_{X+Y}(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\,dx=\int_{-\infty}^{\infty}f_Y(y)f_X(z-y)\,dy}$

Covariance	
Covariance	$\boxed{\operatorname{Cov}(X,Y)=\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]=\mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y]}$
Variance of a sum	$\operatorname{Var}[X+Y]=\operatorname{Var}[X]+2\operatorname{Cov}(X,Y)+\operatorname{Var}[Y]$
Variance in terms of covariance	$\boxed{\operatorname{Var}[W]=\operatorname{Cov}[W,W]}$
Variance of a sum of *independent* rvs	$X\perp Y\implies\operatorname{Cov}(X,Y)=0\implies\operatorname{Var}[X+Y]=\operatorname{Var}[X]+\operatorname{Var}[Y]$
Variance of linear transformation	$\operatorname{Var}[aX+bY]=a^2\operatorname{Var}[X]+2ab\operatorname{Cov}(X,Y)+b^2\operatorname{Var}[Y]$
Bilinearity property	$\operatorname{Cov}(aX+bY,cX+dY)=ac\operatorname{Var}[X]+(ad+bc)\operatorname{Cov}(X,Y)+bd\operatorname{Var}[Y]$ $\operatorname{Cov}(X+Y,Z+W)=\operatorname{Cov}(X,Z)+\operatorname{Cov}(X,W)\operatorname{Cov}(Y,Z)+\operatorname{Cov}(Y,W)$
Independence implies uncorrelated	$X\perp Y\implies\operatorname{Cov}(X,Y)=0$
Correlation coefficient	$\boxed{\rho(X,Y)=\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}}\in[-1,1]}$
Rescaling the correlation coefficient	$\rho(aX,bY)=\rho(X,Y)$

7 Generating Functions

Probability generating function (pgf)	$S_X=\{0,1,2\dots\}$
Probability generating function	$\boxed{P_X(z):=\mathbb{E}\left[z^X\right]=\sum_{k=0}^{\infty}p_X(k)\,z^k}$
pmf from pgf (Inversion formula)	$\boxed{p_X(k)=\frac{P_X^{(k)}(0)}{k!}}$
Expectation	$\mathbb{E}[X]=P_X'(1)$
Variance	$\operatorname{Var}[X]=P_X''(1)+P_X'(1)-[P_X'(1)]^2$
<u>Convolution theorem for PGFs</u>	$X\perp Y\implies\boxed{P_{X+Y}(z)=P_X(z)P_Y(z)}$
Linear combination	$P_{aX+b}(z)=P_X(z^a)z^b$
Relationship between pgf and mgf	$P_X(z)=M_X(\ln z)$

Moment generating function (mgf)	$\operatorname{dom}M_X=T=\{t\in\mathbb{R}:\mathbb{E}\left[e^{tX}\right]<\infty\}$
Moment generating function (mgf)	$\boxed{M_X(t):=\mathbb{E}\left[e^{tX}\right]=\begin{cases}\sum_{x\in S_X}e^{tx}\,p_X(x)\\ \int_{-\infty}^{\infty}e^{tx}\,f_X(x)\,dx\end{cases}}$
Relation between the pgf and the mgf	$M_X(t)=P_X(e^t)$
Convolution theorem for mgfs	$X\perp Y\implies M_{X+Y}(t)=M_X(t)M_Y(t)$
Linear transformation	$M_{aX+b}(t)=M_X(at)e^{bt}$
Moments via Taylor expansion	$M_X(t)=\sum_{n=0}^{\infty}\mathbb{E}[X^n]\,\frac{t^n}{n!}$
Computing *moments*	$\boxed{\mathbb{E}[X^n]=M_X^{(n)}(0)}$
Computing *central moments*	$\mathbb{E}[(X-\mu)^n]=\frac{\mathrm{d}^n}{\mathrm{d}t^n}(e^{-\mu t}M_X(t))\Big\vert_{t=0}$
Expectation	$\mathbb{E}[X]=M_X'(0)$
Variance	$\operatorname{Var}[X]=M_X''(0)-[M_X'(0)]^2$

Cumulant generating function (cgf)	$\operatorname{dom}K_X=T=\{t\in\mathbb{R}:\mathbb{E}\left[e^{tX}\right]<\infty\}$
Cumulant generating function	$\boxed{K_X(t):=\ln\mathbb{E}\left[e^{tX}\right]=\ln M_X(t)}$
Relation between the mgf and the cgf	$M_X(t)=e^{K_X(t)}$
Convolution theorem for cgfs	$X\perp Y\implies K_{X+Y}(t)=K_X(t)+K_Y(t)$
Linear transformation	$M_{aX+b}(t)=M_X(at)e^{bt}$
Cumulants via Taylor expansion	$K_X(t)=\sum_{n=1}^{\infty}\kappa_n\frac{t^n}{n!}$
Cumulants	$\boxed{\kappa_n=K_X^{(n)}(0)}$
Expectation	$\mathbb{E}[X]=M_X'(0)$
Variance	$\operatorname{Var}[X]=M_X''(0)-[M_X'(0)]^2$
Coefficient of Skewness	$\operatorname{Skew}(X)=\frac{\kappa_3}{\sigma^3}=\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sigma}\right)^3\right]$
Coefficient of Kurtosis	$\operatorname{Kurt}(X)=\frac{\kappa_4}{\sigma^4}=\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sigma}\right)^4\right]-3$

Distribution	PGF	MGF	CGF
Bernoulli $X\sim\operatorname{Bi}(1,p)$	$1-p+pz,\quad z\in\mathbb{R}$	$1-p+pe^t$	$\ln(1-p+pe^t)$
Binomial $X\sim\operatorname{Bi}(n,p)$	$(1-p+pz)^n,\quad z\in\mathbb{R}$	$(1-p+pe^t)^n,\quad t\in\mathbb{R}$	$n\ln(1-p+pe^t)$
Geometric $X\sim\operatorname{G}(p)$	$\frac{p}{1+(p-1)z},\quad \lvert z\rvert<\frac{1}{1-p}$	$\frac{p}{1+(p-1)e^t}$	
Negative Binomial $X\sim\operatorname{NB}(r,p)$	$p^r(1+(p-1)z)^{-r},\quad \lvert z\rvert<\frac{1}{1-p}$	$\left(\frac{p}{1-(1-p)e^t}\right)^r,\quad t\in\left(-\infty,\ln\frac{1}{1-p}\right)$	
Poisson $X\sim\operatorname{Pn}(\lambda)$	$e^{\lambda(z-1)},\quad z\in\mathbb{R}$	$e^{\lambda(e^t-1)},\quad t\in\mathbb{R}$	$\lambda(e^t-1)$
Exponential $X\sim\exp(\alpha)$		$\frac{\alpha}{\alpha-t},\quad t\in(-\infty,\alpha)$	
Gamma $X\sim\gamma(r,\alpha)$		$\left(\frac{\alpha}{\alpha-t}\right)^r,\quad t\in(-\infty,\alpha)$	
Standard Normal $Z\sim\mathcal{N}(0,1)$		$e^{t^2/2},\quad t\in\mathbb{R}$	
Normal $X\sim\mathcal{N}(\mu,\sigma^2)$		$e^{\mu t+\sigma^2t^2/2}, t\in\mathbb{R}$	$\mu t+\sigma^2t^2/2$

8 Limit Theorems

Convergence in distribution	$\{X_n:n\geq 1\}$
By distribution function	$\lim_{n\to\infty}F_{X_n}(x)=F_X(x)\implies X_n\xrightarrow{d}X$ (for all x at which the function $F_Y(x)$ is continuous)
By expectation property	$\lim_{n\to\infty}\mathbb{E}[\varphi(X_n)]=\mathbb{E}[\varphi(X)]\iff X_n\xrightarrow{d}X$ (for all bounded continuous functions $\varphi:\mathbb{R}\to\mathbb{R}$)
By the **mgf**	$\lim_{n\to\infty}M_{X_n}(t)=M_X(t),\quad\forall t\in(-\delta,\delta)\implies X_n\xrightarrow{d}X$ (where the mgf's of X_n and X are well defined in a common neighbourhood of the origin, say $(-\delta,\delta)$.)

Central Limit Theorem	Let $\{X_n : n \geqslant 1\}$ be a i.i.d sequence, and $S_n = X_1 + \cdots + X_n$ be the partial sum
Central Limit Theorem	$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{V[S_n]}} = \frac{\bar{S}_n - \mathbb{E}[\bar{S}_n]}{\sqrt{V[\bar{S}_n]}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as } n \to \infty$
Distribution function	$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leqslant x\right) = \Phi(x), \quad \forall x \in \mathbb{R}$

9 Stochastic Processes

Discrete-time Markov Chains	
One-step transition probability	$P_{ij} := \mathbb{P}(X_{n+1} = j \mid X_n = i)$
m-step transition probability	$P_{ij}^{(m)} = \mathbb{P}(X_{n+m} = j \mid X_n = i)$
Transition Matrix	$P_{\lvert S\rvert \times \lvert S\rvert} = [P_{ij}]_{i,j\in S} = [\mathbb{P}(X_{n+1} = j \mid X_n = i)]_{i,j\in S}$
Stochastic matrix	1. $P_{ij} \geq 0$ 2. $\sum_{j\in S} P_{ij} = 1, \quad \forall i \in S$

Equilibrium Distributions	
Initial distribution	$\boldsymbol{\pi}_0 = (\mathbb{P}(X_0 = 0), \mathbb{P}(X_0 = 1)\ldots)$
n-step distribution	$\boxed{\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 P^n}$
Law of Total Probability	$\boxed{\mathbb{P}(X_n = i) = \sum_k \mathbb{P}(X_0 = k)\,\mathbb{P}(X_n = i \mid X_0 = k) = \boldsymbol{\pi}_0 [P^n]_{\_,i}}$
Equilibrium distribution	$\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots)$, where $\boxed{\boldsymbol{\pi} = \boldsymbol{\pi} P}$ (The equilibrium distribution is the left-eigenvector of P with eigenvalue 1)
Limiting interpretation	$\boxed{\pi_j = \lim_{n\to\infty} P_{ij}^{(n)}}$ (The limiting probability of the stochastic process being in state j)
Stationary interpretation	$\boxed{\boldsymbol{\pi} = \boldsymbol{\pi} P}$ (If the process starts with probability distribution it will persist with that distribution *forever*.)
Ergodic interpretation	$\boldsymbol{\pi}$ Is the long term *relative frequencies* of being in state