MATH0057 – Probability and Statistics

Based on lectures by Dr Kayvan Sadeghi
Notes taken by Robert Moye

Term 2 of 2021 – 2022

Probability is the quantification of uncertainty. Statistics is the interpretation of information under uncertainty.

The sections for interval until occurrence distributions are

Random variable	Distribution			
	Discrete		Continuous	
Number of incidents	Binomial	(1.3.3)	Poisson	(1.3.7)
Interval until 1^{st} incident	Geometric	(1.3.4)	Exponential	(1.5.2)
Interval until k^{th} incident	Neg. binomial	(1.3.5)	Gamma	(1.5.3)

Contents

1 Probability

1.1 Basic probability theory

1.1.1 Definitions

Definition (Sample space)**.** Probability statements, made in the context of an experiment, may result in a number of possible outcomes, each of which may be denoted ω. The set of all possible outcomes is a sample space, denoted Ω.

Definition (Event)**.** A collection of outcomes is an event. An event E occurs if the outcome of the experiment of E. $E \subseteq \Omega$.

1.1.2 Sets

Notation. For events E, F, recall that

- E^c is the complement of E, the set of outcomes that are not in E.
- $E \cup F$ is the union of E and F, the set of outcomes in E or in F (or both).
- $E \cap F$ is the intersection of E and F, the set of outcomes in both E and F.
- $E \subseteq F$ denotes that E is a subset of F, so every element of E is in F.
- $E \subset F$ denotes that E is a subset of F, but E and F are not identical.
- $E \backslash F = E \cap F^c$ is the set of elements of E that are not in F.
- $\emptyset = \{\} = \Omega^c$ is the empty set, that which contains no elements.
- If $E \cap F = \emptyset$ then no outcomes are common to both E and F. These events are called disjoint, or mutually exclusive (m.e.)
- If all possible pairs of events $E_1, \dots E_n$ are disjoint, they are called mutually disjoint or pairwise disjoint.

Law 1.1 (De Morgan's)**.** For events E, F,

$$(E \cap F)^c = E^c \cup F^c \tag{1.1}$$
$$(E \cup F)^c = E^c \cap F^c \tag{1.2}$$

Law 1.2 (Distribution)**.** For events E, F, G,

$$E \cap (F \cup G) = (E \cap F) \cup (E \cap G) \tag{1.3}$$
$$E \cup (F \cap G) = (E \cup F) \cap (E \cup G) \tag{1.4}$$

1.1.3 Axioms

Definition (σ-algebra event space)**.** For sample space Ω and events $E_i \in \Omega$ in an event space $\mathcal{F}$, $\mathcal{F}$ is a σ-algebra if

$$\Omega \in \mathcal{F} \tag{1.5}$$
$$E \in \mathcal{F} \implies E^c \in \mathcal{F} \tag{1.6}$$
$$E_1 \in \mathcal{F}, \dots, E_n \in \mathcal{F} \implies \bigcup_{i=1}^{n} E_i \in \mathcal{F} \tag{1.7}$$

Example 1.1. Consider an experiment with fair coin tossed once, which lands on heads or tails, H or T. The sample space is $\Omega = \{H, T\}$ whilst the event space is $\mathcal{F} = \big\{\emptyset, \{H\}, \{T\}, \Omega\big\}$.

Axiom 1.3 (Kolmogrov's)**.** A probability function P is a mapping $P : \mathcal{F} \to \mathbb{R}$ such that $\forall E, F \in \mathcal{F}$,

$$P(E) \geq 0 \tag{1.8}$$
$$P(\Omega) = 1 \tag{1.9}$$
$$E \cap F = \emptyset \implies P(E \cup F) = P(E) + P(F) \tag{1.10}$$

Remark. For countably infinite sequences $E_1, E_2, \dots$ of mutually disjoint sets in $\mathcal{F}$, equation (1.10) may need to be adapted to

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

1.1.4 Constructing probability functions

Remark. The axioms capture how probabilities behave intuitively, based on relative frequency (how many times an event occurs over many repetitions of an experiment), but allocation of useful probabilities rely on mathematical modelling. All results will be deduced from the axioms.

Method. For a countable sample space, specify probabilities for each individual $\omega \in \Omega$. For $\Omega = \{w_1, w_2, \dots\}$, let $\{p_1, p_2\}$ be a set of 'weights' satisfying

$$p_k \geq 0, \quad \sum_{k=1}^{\infty} p_k = 1$$

Then for any event $E \subseteq \Omega$,

$$P(E) = \sum_{k : w_k \in E} p_k$$

Clearly this satisfies equations (1.8) and (1.9). For equation (1.10), see that

$$P(E \cup F) = \sum_{i : \omega_i \in E \cup F} p_i = \sum_{i : w_i \in E} p_i + \sum_{j : w_j \in F} p_j = P(E) + P(F)$$

Remark. For an uncountable sample space, construction of a probability function is more difficult, but the axioms still work.

1.1.5 Deductions

Corollary 1.4. $\forall E, F, E_1, \dots E_n \in \mathcal{F}$,

$$P(E^c) = 1 - P(E) \tag{1.11}$$
$$P(E) \leq 1 \tag{1.12}$$
$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{1.13}$$
$$E \subseteq F \implies P(F \backslash E) = P(F) - P(E) \quad \text{and} \quad P(E) \leq P(F) \tag{1.14}$$
$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i) \tag{1.15}$$

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) = &\sum_{i=1}^{n} P(E_i) \\ &- \sum_{i<j} P(E_i \cap E_j) \\ &+ \sum_{i<j<k} P(E_i \cap E_j \cap E_k) \\ &\quad\vdots \\ &+ (-1)^{n-1} P(E_1 \cap \cdots \cap E_n) \end{aligned} \tag{1.16}$$

Proof.

(1.11). By equations (1.9) and (1.10),

$$P(E) + P(E^c) = P(E \cup E^c) = P(\Omega) = 1$$

(1.12). By equations (1.8) and (1.11),

$$P(E) \leq P(E) + P(E^c) = 1$$

(1.13). By equations (1.8) and (1.11),

$$\begin{aligned} P(E \cup F) &= P\big(E \cup (F \cap E^c)\big) \\ &= P(E) + P(F \cap E^c) \\ &= P(E) + P(F^{cc} \cap E^c) \\ &= P(E) + P(F^c \cup E)^c \\ &= P(E) + 1 - P(F^c \cup E) \\ &= P(E) + 1 - P\big(F^c \cup (E \cap F)\big) \\ &= P(E) + 1 - P(F^c) - P(E \cap F) \\ &= P(E) + P(F) - P(E \cap F) \end{aligned}$$

(1.14). Since $E \subseteq F$, $F^c \subseteq E^c$, so $F \cup E^c = \Omega$. So by equations (1.11) and (1.13),

$$\begin{aligned} P(F \backslash E) &= P(F \cap E^c) \\ &= P(F) + P(E^c) - P(F \cup E^c) \\ &= P(F) + 1 - P(E^c) - P(\Omega) \\ &= P(F) - P(E) \end{aligned}$$

By equation (1.8),

$$\begin{aligned} P(E) + P(F \backslash E) &= P(F) \\ P(E) &\leq P(F) \end{aligned}$$

(1.15). (by induction) The case $n = 1$ is obvious (and $n = 2$ is given by equation (1.13)). Let

$$U_r = \bigcup_{i=1}^{r} E_i$$

Suppose true for $n = k$. Then

$$P\left(U_k\right) \leq \sum_{i=1}^{k} P(E_i)$$

By equation (1.13),

$$\begin{aligned} P(U_{k+1}) &= P(U_k \cup E_{r+1}) \\ &= P(U_k) + P(E_{k+1}) - P(U_k \cap E^k + 1) \end{aligned}$$

By equation (1.8),

$$\begin{aligned} &\leq P(U_k) + P(E_{k+1}) \\ &= \sum_{i=1}^{k} P(E_i) + P(E_{k+1}) \\ P(U_{k+1}) &\leq \sum_{i=1}^{k+1} P(E_i) \end{aligned}$$

So true for $n = k + 1$.

True for $n = 1$, and true for $n = k \implies$ true for $n = k + 1$, so by induction, true for $n \in \mathbb{N}$.

(1.16). (by induction) The case $n = 2$ is given by equation (1.13). As before, let

$$U_r = \bigcup_{i=1}^{r} E_i$$

Suppose true for $n = k$. Then by equation (1.13),

$$P(U_{k+1}) = P(U_r \cup E_{k+1}) = P(U_k) + P(E_{k+1}) - P(U_k \cap E_{k+1})$$

By equation (1.3),

$$\begin{aligned} U_k \cap E_{k+1} &= (E_1 \cup \cdots \cup E_k) \cap E_{k+1} \\ &= (E_1 \cap E_{k+1}) \cup \cdots \cup (E_k \cup E_{k+1}) \\ &= \sum_{i=1}^{k} E_i \cap E_{k+1} =: \sum_{i=1}^{k} I_i \end{aligned}$$

By the assumption,

$$\begin{aligned} P(U_k) &= \sum_{i=1}^{k} P(E_i) - \cdots + (-1)^{r-1} P(E_1 \cap \cdots \cap E_{k+1}) \\ P\left(\bigcup_{i=1}^{k} I_i\right) &= \sum_{i=1}^{k} P(I_i) - \cdots + (-1)^{r-1} P(I_1 \cap \cdots \cap I_{k+1}) \end{aligned}$$

So

$$
\begin{aligned}
P(U_{k+1}) &= P(U_k) + P(E_{k+1}) - P(I_k) \\
&= \sum_{i=1}^{k} P(E_i) - \dots + (-1)^{r-1} P(E_1 \cap \dots \cap E_{k+1}) + P(E_{k+1}) \\
&\quad - \left(\sum_{i=1}^{k} P(I_i) - \dots + (-1)^{r-1} P(I_1 \cap \dots \cap I_{k+1}) \right) \\
&= \sum_{i=1}^{k+1} P(E_{k+1}) \\
&\quad - \sum_{1 \le i < j \le k} P(E_i \cap E_j) + \dots + (-1)^{r-1} P(E_1 \cap \dots \cap E_{k+1}) \\
&\quad - \left(\sum_{i=1}^{k} P(I_i) - \dots + (-1)^{r-1} P(I_1 \cap \dots \cap I_{k+1}) \right)
\end{aligned}
$$

Substitute back $I_i = E_i \cap E_{k+1}$, and pair up and combine terms from the two lines.

$$
\begin{aligned}
&= \sum_{i=1}^{k+1} P(E_{k+1}) \\
&\quad - \left(\sum_{1 \le i < j \le k} P(E_i \cap E_j) + \sum_{i=1}^{k} P(E_i \cap E_{k+1}) \right) \\
&\quad + \left(\sum_{1 \le i < j < l \le k} P(E_i \cap E_j \cap E_l) + \sum_{1 \le i < j \le k} P(I_i \cap I_j) \right) \\
&\quad \vdots
\end{aligned}
$$

$$
\begin{aligned}
P(U_{k+1}) = \quad & \sum_{i=1}^{k+1} P(E_i) - \sum_{1 \le i < j \le k+1} P(E_i \cap E_j) + \dots \\
& \dots + (-1)^k P(E_1 \cap \dots \cap E_{k+1})
\end{aligned}
$$

Hence true for $n = k + 1$.

True for $n = 2$, and true for $n = k \implies$ true for $n = k + 1$, so by induction, true for $n \in \mathbb{Z},\ n \ge 2$.

$\square$

1.1.6 Independence

Definition (Independent)**.** Events E and F are independent if

$$P(E \cap F) = P(E)P(F)$$

Events $E_1, \dots, E_n$ are mutually independent if, for any collection $E_{i_1}, \dots, E_{i_k}$,

$$P(E_{i_1} \cap \dots E_{i_k}) = \prod_{j=1}^{k} P(E_{i_j})$$

Remark. Independence captures the idea that events are unrelated, that one does not affect the other.

Remark. For independent events E, F,

$$\begin{aligned} \text{Mutually exclusive} &\iff P(E \cap F) = 0 \iff P(E)P(F) = 0 \\ &\iff P(E) = 0 \text{ or } P(F) = 0 \end{aligned}$$

Remark. Mutual independence is a stronger condition than pairwise independence, which only requires pairs of events E_i, E_j to be independent.

1.1.7 Conditional probability

Definition (Conditional probability)**.** The conditional probability of F given E is the probability that F occurs when E is known to have occurred. For $P(E) > 0$, it is given by

$$P(F|E) = \frac{P(E \cap F)}{P(E)}$$

Remark. The definition of conditional probability may be interpreted as changing the sample space from Ω to E, that is $P(F|E)$ is the probability of F given that the outcome of the experiment is known to be in E. Note that $P(E|E) = 1$.

Remark. If E and F are independent, and $P(E) \neq 0$, then

$$P(F|E) = \frac{P(E \cap F)}{P(E)} = \frac{P(E)P(F)}{P(E)} = P(F)$$

So E occurring does not change the chance of F occurring.

Law 1.5 (Total probability)**.** Let $\{E_i\}$ be a partition of Ω (the E_is are disjoint, and their union gives Ω). Then for any event F,

$$P(F) = \sum_i P(F|E_i)P(E_i)$$

Proof. By induction on equation (1.3),

$$F = F \cap \Omega = F \cap \left(\bigcup_i E_i \right) = \bigcup_i E_i \cap F$$

So

$$P(F) = \sum_i P(F \cap E_i) = \sum_i P(F|E_i)P(E_i) \qquad \square$$

Theorem 1.6 (Bayes')**.** Let $\{E_i\}$ be a partition of Ω and F be any event with $P(F) > 0$. Then

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{P(F)} = \frac{P(F|E_i)P(E_i)}{\sum_i P(F|E_i)P(E_i)}$$

Proof. By definition of conditional probability

$$P(E_i|F) = \frac{P(E_i \cap F)}{P(F)} \quad \text{and} \quad P(E_i \cap F) = P(F|E_i)P(E_i)$$

The denominator comes from the Law of Total Probability (1.5). $\square$

Example 1.2. An urn contains five red balls and one white ball. A ball is drawn and then it and another ball of the same colour are placed back in the urn. Finally a second ball is drawn.

1. What is the probability that the second ball is white?
2. If the second ball is white, what is the probability that the first was red?

Let W_i be the probability the i^{th} ball is white, and R_i the corresponding probability for red.

1. By the Law of Total Probability (1.5),

$$P(W_2) = P(W_2|R_1)P(R) + P(W_2|W_1)P(W_1) = \frac{1}{7}\frac{5}{6} + \frac{2}{7}\frac{1}{6} = \frac{1}{6}$$

2. By Bayes' Theorem (1.6),

$$P(R_1|W_2) = \frac{P(W_2|R_1)P(R_1)}{P(W_2)} = \frac{\frac{1}{7}\frac{5}{6}}{\frac{1}{6}} = \frac{5}{7}$$

Law 1.7 (Generalised Multiplication)**.** For events $E_1, \ldots, E_n$ that satisfy

$$P(E_1 \cap \cdots \cap E_{n-1}) > 0,$$

$$P\left(\bigcap_{i=1}^{n} E_i\right) = \prod_{i=1}^{n} P(E_i|E_1 \cap \cdots \cap E_{i-1})$$
$$P(E_1 \cap \cdots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2) \ldots P(E_n|E_1 \cap \cdots \cap E_{n-1})$$

Proof. All conditional probabilities are defined, since equation (1.13) gives

$$P(E_1) \geq P(E_1 \cap E_2) \geq \cdots \geq P(E_1 \cap \cdots \cap E_{n-1}) > 0$$

By the definition of conditional probability,

$$\begin{aligned}
\prod_{i=1}^{n} P(E_i|E_1 \cap \cdots \cap E_{i-1}) &= \prod_{i=1}^{n} \frac{P(E_1 \cap \cdots \cap E_i)}{P(E_1 \cap \cdots \cap E_{i-1})} \\
&= \prod_{i=1}^{n} P(E_1 \cap \cdots \cap E_i) \prod_{i=1}^{n-1} \frac{1}{P(E_1 \cap \cdots \cap E_i)} \\
&= \prod_{i=1}^{n-1} \frac{P(E_1 \cap \cdots \cap E_i)}{P(E_1 \cap \cdots \cap E_i)} P\left(\bigcap_{i=1}^{n} E_i\right) \\
&= P\left(\bigcap_{i=1}^{n} E_i\right)
\end{aligned}$$

or written out,

$$\begin{aligned}
P(E_1 \cap \cdots \cap E_n) &= P(E_1)\frac{P(E_1 \cap E_2)}{P(E_1)}\frac{P(E_1 \cap E_2 \cap E_3)}{P(E_1 \cap E_2)} \cdots \frac{P(E_1 \cap E_2 \cap E_n)}{P(E_1 \cap E_2 \cap E_{n-1})} \\
&= P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2) \ldots P(E_n|E_1 \cap \cdots \cap E_{n-1})
\end{aligned}$$

□

Remark. This may be useful for problems where an event of interest arises from a sequence of contributing events.

1.2 Discrete random variables

1.2.1 Distributions

Definition (Discrete r.v.)**.** For a countable sample space Ω, a function

$$X : \Omega \to \mathbb{R}$$

is called a discrete random variable (discrete r.v.). This maps outcomes of an experiment to numbers. For brevity, its usage may be notated

$$P\big(\{\omega : X(\omega) = x\}\big) = P\big(X(\omega) = x\big) = P(X = x)$$

Example 1.3. When tossing a coin twice, the number of heads obtained may be given by the function

$$X = \big\{(HH, 2), (HT, 1), (TH, 1), (TT, 0)\big\}$$

Definition (Probability mass function)**.** For a discrete random variable X taking values in the set $\{x_i\}$, the function $p_X(x) = P(X = x)$ is the probability mass function (pmf) of X.

If unambiguous, this may be written $p_X(x) = p(x)$. For integers, this may be written $p(X = n) = p_n$.

Remark. For a collection $\{p(x_i)\}$ to be a pmf, it must satisfy the axioms of probability (1.3).

(1.8). Since the ps are probabilities of events, $\forall i\ p(x_i) \geq 0$.

(1.9). Since the sample space may be written as a union of disjoint events

$$\Omega = \bigcup_i \{\omega : X(\omega) = x_i\}$$

by equation (1.10),

$$\begin{aligned}\sum_i p(x_i) &= P\left(\bigcup_i \{\omega : X(\omega) = x_i\}\right)\\ &= \sum_i P\big(\{\omega : X(\omega) = x_i\}\big) = P(\Omega) = 1\end{aligned}$$

(1.10). Satisfied by the selection of mutually independent events $\{x_i\}$.

Remark. The collection $\{x_i\}$ may be regarded as a sample space in its own right, with the probability function $P\big(\{x_i\}\big) = p(x_i.)$. This may be appropriate to use when only the values taken by X are of interest.

Notation. It may be useful to consider the probability that X lies within an interval.

$$P(a \leq X \leq b) = P\big(\{\omega : a \leq X(\omega) \leq b\}\big) = \sum_{i : a \leq x_i \leq b} p_x(x_i)$$

Definition (CDF)**.** For any random variable X, the (Cumulative) distribution function (cdf) of X is

$$F_X(x) = P(X \leq x) = \sum_{i:x_i \leq x} p_X(x_i)$$

Rule 1.8.

$$F(x) \to 0 \quad \text{as} \quad x \to -\infty \tag{1.17}$$
$$F(x) \to 1 \quad \text{as} \quad x \to \infty \tag{1.18}$$
$$F(x) \text{ is monotonic increasing} \tag{1.19}$$
$$F(x) \text{ is a step function, with discontinuities at } \{x_i\} \text{ of height } p(x_i) \tag{1.20}$$

Proof. Obvious from the axioms of probability (1.3). □

Remark. The pmf may be derived from the cdf as

$$p(x) = P(X = x) = P(X \leq x) - P(X < x) = \lim_{\delta \to 0^+} \big(F(x) - F(x - \delta)\big)$$

A general version of this result follows in rule 1.9

Rule 1.9.

$$a < b \implies P(a < X \leq b) = F(b) - F(a)$$

Proof. By equation (1.10)

$$\begin{aligned} F(b) &= P(X \leq b) \\ &= P\big((X \leq a) \cup (a < X \leq b)\big) \\ &= P(X \leq a) + P(a < X \leq b) \\ &= F(a) + P(a < X \leq b) \end{aligned}$$

□

1.2.2 Expectation

Definition (Median)**.** The median of a distribution is

$$x : F(x) = \frac{1}{2}$$

If repeating an experiment, a random variable would be expected to lie in equal proportions above and below the median.

Definition (Mode)**.** The mode is the value with highest probability.

Definition (Expectation)**.** The expectation, or expected value, or (population) mean of a discrete random variable X is

$$E(X) := \sum_i x_i P(X = x_i) = \sum_i x_i p(x_i)$$

for probability mass function p, provided the sum is well-defined.

Remark. In the countably-infinite case, the sum is not guaranteed to converge. $E(X)$ exists if the sum converges absolutely, that is, if

$$\sum_i |x_i| \, P(X = x_i) < \infty$$

Remark. $E(X)$ represents an idealised long-run average for the values of X.

Remark. Expected value may be calculated using the values taken by X, or directly on the sample space.

$$E(x) = \sum_i x_i P\{\omega : X(\omega) = x_i\} = \sum_{\omega\in\Omega} X(\omega)P(\omega)$$

Proposition 1.10 (Expectation of a function)**.** For a random variable X and $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(X)$ is a random variable $\phi(X) : \Omega \to \mathbb{R}$, so its expectation may be calculated

$$\begin{aligned} E\big(\phi(X)\big) &= \sum_{\omega\in\Omega} \phi\big(X(\omega)\big)P(\{\omega\}) \\ &= \sum_i \phi(x_i)P(\{\omega : X(w) = x_i\}) \\ &= \sum_i \phi(x_i)p_X(x_i) \end{aligned}$$

Proposition 1.11. For all constants a, c,

$$P(X = c) = 1 \implies E(X) = c \tag{1.21}$$

$$Y = aX + c \implies E(Y) = aE(X) + c \tag{1.22}$$

Proof. By definition,

(1.21)

$$E(X) = \sum_i x_i p(x_i) = cP(X = c) = c$$

(1.22)

$$\begin{aligned} E(Y) &= \sum_i (ax_i + b)p(x_i) \\ &= a\sum_i x_i p(x_i) + b\sum_i p(x_i) \\ &= aE(X) + b \end{aligned}$$

□

Proposition 1.12. If a random variable X has a symmetric probability mass function and $E(X)$ exists, then $E(X)$ is the central value.

Proof. Suppose the pmf is symmetric about zero, so $X = \{0, \pm x_1, \pm x_2, \dots\}$ with respective probabilities $p_0, p_1, \dots$. Then

$$E(X) = 0p_0 + (x_1p_1 - x_1p_1) + (x_2p_2 - x_2p_2) + \cdots = 0$$

If X is symmetric about some other value μ, then define $Y = X - \mu$ to get a pmf symmetric about zero, so $E(Y) = 0$. Then by equation (1.21),

$$E(X) = E(Y) + \mu = \mu$$

□

1.2.3 Variance

Definition (Variance)**.** When it exists, the r^{th} moment of X about α is

$$E\big((X-\alpha)^r\big)$$

Definition. For a random variable X with mean $E(X) = \mu$, the variance of X is

$$\operatorname{Var}(X) = E\big((X-\alpha)^2\big)$$

Remark. $E(X)$ is the first moment about zero, and $\operatorname{Var}(X)$ is the second moment about $\mu = E(X)$.

Proposition 1.13. For a random variable X with $\mu = E(X)$,

$$\operatorname{Var}(X) = E(X^2) - \mu^2 \tag{1.23}$$
$$\operatorname{Var}(X) = E\big(X(X-1)\big) - \mu(\mu-1) \tag{1.24}$$

Proof. Using equation (1.21) gives

(1.23)

$$\begin{aligned} E\big((X-\alpha)^2\big) &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu\mu + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

(1.24)

$$\begin{aligned} E\big((X-\alpha)^2\big) &= E(X^2 - X + X(1-2\mu) + \mu^2) \\ &= E(X^2 - X) + (1-2\mu)E(X) + \mu^2 \\ &= E\big(X(X-1)\big) + (1-2\mu)\mu + \mu^2 \\ &= E\big(X(X-1)\big) - \mu(\mu-1) \end{aligned}$$

□

Proposition 1.14. For a random variable X and all constants a, c,

Where it exists,
$$\operatorname{Var}(X) \geq 0 \tag{1.25}$$
$$\operatorname{Var}(X) = 0 \iff P(X=c) = 1 \tag{1.26}$$
$$Y = aX + c \implies \operatorname{Var}(Y) = a^2 \operatorname{Var}(X) \tag{1.27}$$

Proof.

(1.25) Variance is a sum of non-negative terms

$$\operatorname{Var}(X) = E(X-\mu)^2 = \sum_i (x_i - \mu)p(x_i)$$

(1.26) By equation (1.21),

$$P(X=c) = 1 \implies E(X) = c$$

so by the definition of varience,

$$\implies \operatorname{Var}(X) = 0$$

Now suppose X is not constant. Then X takes at least two values with non-zero probability. Therefore at least one term in

$$\operatorname{Var}(X) = E(X-\mu)^2 = \sum_i (x_i - \mu)p(x_i)$$

must not be zero, so $\operatorname{Var}(X) > 1$. Hence in this case, $\operatorname{Var}(X) \neq 0$.

(1.27) By equation (1.22),

$$\begin{aligned}\operatorname{Var}(Y) &= E\big(Y - E(Y)\big)^2 \\ &= E\big(aX + c - aE(X) - c\big)^2 \\ &= a^2 E\big(X - E(X)\big)^2 \\ &= a^2 \operatorname{Var}(X)\end{aligned}$$

□

Definition (Standard definition)**.** Where it exists, the standard deviation of X is

$$\operatorname{sd}(X) = \sqrt{\operatorname{Var}(X)}$$

Remark. Unlike variance, standard deviation is measured in the same units of X, so has a direct interpretation.

Notation. Expectation, variance and standard deviation may be notated as μ, σ^2, σ respectively, or if necessary, $\mu_X, \mu_Y, \sigma_X^2, \ldots$.

Remark. Expectation may be interpreted as a measure of location, and varience and standard deviation as measures of spread or dispersion.

Definition (Coefficient of variation)**.** The coefficient of variation of a random variable X is the dimensionless ratio

$$\frac{\sigma}{\mu}$$

1.2.4 Probability generating function

Definition (PGF)**.** For a random variable X taking values $0, 1, 2, \ldots$, the probability generation function (pgf) of X is the power series

$$\Pi_X(z) = E(z^X) = \sum_{k=0}^{\infty} z^k p(k)$$

for all values of z where the expectation is defined.

Proposition 1.15. For a random variable X with pgf $\Pi(z)$, where the expectations exist, the n^{th} factorial moment of X is

$$\Pi(1) = 1 \tag{1.28}$$
$$\Pi'(1) = E(X) \tag{1.29}$$
$$\Pi''(1) = E\big(X(X-1)\big) \tag{1.30}$$
$$\Pi^{(n)}(1) = E\left(\frac{X!}{(X-n)!}\right) \tag{1.31}$$

Proof by induction.

(1.28)

$$\Pi(1) = E(1) = 1$$

(1.31) Obvious, by induction on the previous method.

$$\begin{aligned}
\Pi^{(n+1)}(z) &= \frac{\mathrm{d}\Pi^{(n)}(z)}{\mathrm{d}z} \\
&= \frac{\mathrm{d}}{\mathrm{d}z}\left(\sum_{k=n}^{\infty} \frac{k!}{(k-n)!} z^{k-n} p(k)\right) \\
&= \sum_{k=n}^{\infty} \frac{k!}{(k-n)!}(k-n) z^{k-(n+1)} p(k) \\
&= \sum_{k=n+1}^{\infty} \frac{k!}{\big(k-(n+1)\big)!} z^{k-(n+1)} p(k)
\end{aligned}$$

so

$$\Pi^{(n)}(1) = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} p(k) = \sum_{k=0}^{\infty} \frac{k!}{(k-n)!} p(k) = E\left(\frac{X!}{(X-n)!}\right) \quad \square$$

1.3 Standard discrete distributions

1.3.1 Choosing a distribution

Bernoulli An experiment with two outcomes that form a partition of the sample space

$$X : \Omega \to \{0, 1\}$$

$$\begin{aligned}
\text{“success”} &= \{\omega \in \Omega : X(\omega) = 1\} \\
\text{“faliure”} &= \{\omega \in \Omega : X(\omega) = 0\}
\end{aligned}$$

may be called a binary experiment, or Bernoulli trial. X has a Bernoulli distribution

Binomial The number of successes of $n \in \mathbb{N}1$ independent Bernoulli trials, each with the same probability of success.

Geometric The number of independent Bernoulli trials, each with the same probability of success, until a success is observed.

Negative Binomial The number of independent Bernoulli trials, each with the same probability of success, until a $r \in \mathbb{N}1$ successes are observed.

Hypergeometric The number of objects with an attribute, when n objects are drawn from a population of N, with M having that attribute.

Poisson The number of instantaneous occurrences within an interval.

1.3.2 Bernoulli

Rule 1.16. For $P(\text{success}) = p$ and,

$$X = \begin{cases} 1 & \text{if success} \\ 0 & \text{if faliure} \end{cases}$$

$$\begin{aligned} E(X) &= 1p + 0(1-p) = p \\ E(X^2) &= 1^2 p + 0^2(1-p) = p \\ \text{Var}(X) &= E(X^2) - \big(E(X)\big)^2 = p - p^2 = p(1-p) \end{aligned}$$

1.3.3 Binomial

Notation. The number of successes X of n independent Bernoulli trials, each with the same probability of success p, follows a binomial distribution with parameters n and p, written

$$X \sim \text{Bin}(n, p)$$

Proposition 1.17. $X \sim \text{Bin}(n, p)$ has pmf

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Proof omitted.

Proposition 1.18. For $X \sim \text{Bin}(n, p)$,

$$\begin{aligned} E(X) &= np \\ \text{Var}(X) &= np(1-p) \end{aligned}$$

Proof omitted. See chapter "sum of independent variables" when written.

1.3.4 Geometric

Notation. The number X of independent Bernoulli trials, each with the same probability of success p, until a success is observed, follows a geometric distribution with parameters p, written

$$X \sim \text{Geo}(p)$$

Proposition 1.19. $X \sim \text{Geo}(p)$ has pmf

$$P(X = k) = (1-p)^{k-1} p$$

Proof. The derivation comes from that $k - 1$ trials must fail before the k^{th} succeeds. Now check that this is a pmf. Clearly it is always non-negative. By the sum of a geometric series,

$$\sum_{k=1}^{\infty} P(X = k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{k=1}^{\infty} (1-p)^{k-1} = p \frac{1}{1-(1-p)} = 1 \qquad \square$$

Proposition 1.20. For $X \sim \mathrm{Geo}(p)$,

$$E(X) = \frac{1}{p}$$
$$\mathrm{Var}(X) = \frac{1-p}{p^2}$$

Proof. The pgf for X is

$$\Pi(z) = E(z^X) = \sum_{k=1}^{\infty} z^k p(1-p)^{k-1} = pz \sum_{k=1}^{\infty} \big(z(1-p)\big)^{k-1} = \frac{pz}{1-(1-p)z}$$

for $|(1-p)z| \leq 1$. Then

$$\begin{aligned}
\Pi'(z) &= \frac{p}{1-(1-p)z} + (1-p)\frac{pz}{(1-(1-p)z)^2} \\
&= \frac{p(1-(1-p)z) + (1-p)pz}{(1-(1-p)z)^2} \\
&= \frac{p}{(1-(1-p)z)^2} \\
\Pi''(z) &= \frac{2p(1-p)}{(1-(1-p)z)^3}
\end{aligned}$$

Hence

$$\begin{aligned}
E(X) &= \Pi'(1) = \frac{1}{p} \\
\mathrm{Var}(X) &= E\big(X(X-1)\big) - E(X)\big(E(X)-1\big) \\
&= \Pi''(1) - \frac{1}{p}\left(\frac{1}{p} - 1\right) \\
&= \frac{2(1-p)}{p^2} - \frac{1}{p}\frac{1-p}{p} = \frac{1-p}{p^2}
\end{aligned}$$

□

Remark. If asked to calculate the probability $P(X > c)$, don't use the geometric distribution, just consider that there are c fails first, so the probability is $(1-p)^c$.

1.3.5 Negative Binomial

Notation. The number X of independent Bernoulli trials, each with the same probability of success p, up to and including the r^{th} success, follows a negative binomial distribution with parameters r and p, written

$$X \sim \mathrm{NB}(r, p)$$

Proposition 1.21. $X \sim \mathrm{NB}(r, p)$ has pmf

$$P(X = k) = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r} \times p = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

Proof. This formula may be derived from that there must be $r-1$ successes in the first $k-1$ throws, then one success. Now check that this is a pmf. Clearly it is always non-negative. By negative binomial expansion,

$$\begin{aligned}\sum_{k=r}^{\infty} P(X=k) &= \sum_{k=r}^{\infty} \binom{k-1}{r-1} p^r (1-p)^{k-r} \\ &= p^r \sum_{k=r}^{\infty} \binom{k-1}{r-1} (1-p)^{k-r} \\ &= p^r \sum_{k=0}^{\infty} \binom{r+k}{r} (1-p)^k \\ &= \frac{p^r}{\big(1-(1-p)\big)^r} = 1\end{aligned}$$

□

Proposition 1.22. For $X \sim \mathrm{NB}(r,p)$,

$$\begin{aligned} E(X) &= \frac{r}{p} \\ \mathrm{Var}(X) &= \frac{r(1-p)}{p^2}\end{aligned}$$

Proof omitted. See chapter "sum of independent variables" when written.

Remark. As with the geometric distribution, if asked to calculate the probability $P(X > c)$, don't use the negative binomial distribution, use the binomial distribution on the first c trials.

1.3.6 Hypergeometric

Notation. Consider a population of finite size N, with M individuals of interest, and n objects are sampled at random. If objects are sampled with replacement, then the number of objects X out of n of interest is

$$X \sim \mathrm{Bin}(n, \tfrac{M}{N})$$

Otherwise, if objects are sampled without replacement,

$$X \sim \mathrm{H}(n, M, N)$$

Proposition 1.23. $X \sim \mathrm{H}(n, M, N)$ has pmf

$$P(X=k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

Proof. For $X = k$, k objects of interest out of M must be selected and $n-k$ objects out of $N-M$ must be selected. Hence the chance of this happening when n objects are chosen from N is given by the formula. Now check that this is a pmf. Clearly it is always non-negative.

Use the binomial theorem.

$$\begin{aligned}(1+x)^N &= (1+x)^M (1+x)^{N-M} \\ &= \left(\sum_{r=0}^{M} \binom{M}{r} x^r\right)\left(\sum_{r=0}^{N-M} \binom{N-M}{r} x^r\right)\end{aligned}$$

The x_n coefficient of this is

$$\binom{N}{n} = \sum_{k=0}^{n} \binom{M}{k}\binom{N-M}{n-k}$$

If $n > N - M$, the first few terms will be zero, and if $M < n$, then the terms for $k > M$ will be zero. Hence

$$\binom{N}{n} = \sum_{k=0}^{\min(n,M)} \binom{M}{k}\binom{N-M}{n-k}$$

and so the sum of probabilities is 1. $\square$

Proposition 1.24. For $X \sim \mathrm{H}(n, M, n)$,

$$\begin{aligned} E(X) &= \frac{nM}{N} \\ \mathrm{Var}(X) &= \frac{nM(N-M)(N-n)}{N^2(N-1)} \end{aligned}$$

Method (Binomial approximation)**.** For $X \sim \mathrm{H}(n, M, N)$ which gives values $\{0, 1, \dots, \min(n, M)\}$,

$$\begin{aligned} P(X = k) &= \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \\ &= \frac{\frac{M!}{(M-k)!k!}\frac{(N-M)!}{(N+k-M-n)!(n-k)!}}{\frac{N!}{(N-n)!n!}} \\ &= \frac{M!(N-M)!(N-n)!n!}{(M-k)!k!(N+k-M-n)!(n-k)!N!} \\ &= \left(\frac{n!}{k!(n-k)!}\right)\left(\frac{M!}{(M-k)!}\frac{(N-M)!}{(N+k-M-n)!}\right)\left(\frac{N!}{(N-n)!}\right)^{-1} \\ &= \binom{n}{k}\frac{M(M-1)\cdots(M-k+1)\times(N-M)(N-M-1)\cdots(N-M-n+k+1)}{N(N-1)\cdots(N-n+1)} \end{aligned}$$

Numerator and denominator both contain k terms, so divide both by N^k.

$$= \binom{n}{k}\frac{\frac{M}{N}\cdots\left(\frac{M}{N}-\frac{k+1}{N}\right)\times\frac{N-M}{N}\cdots\left(\frac{N-M}{N}-\frac{n+k+1}{N}\right)}{1\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)}$$

Take the limit as $N \to \infty$, with fixed $\frac{M}{N} = p, n, k$.

$$\xrightarrow{N\to\infty} \binom{n}{k}\left(\frac{M}{N}\right)^k\left(\frac{N-M}{N}\right)^{n-k} = \binom{n}{k}p^k(1-p)^{n-k}$$

This is the binomial distribution. Clearly the limiting behaviour of $E(X)$ and $\mathrm{Var}(X)$ agree with the corresponding binomial expressions.

1.3.7 Poisson

Remark. The lecturer uses language of accidents in time intervals, but I generalised this to incidents in measures.

Definition (Poisson process)**.** For some constant rate of process λ, a Poisson process satisfies

1. The probability of one incident during some measure δ is $\delta\lambda + o(\lambda)$.
2. The probability of no incidents during the same measure is $1 - \lambda\delta + o(\lambda)$.
3. If they don't overlap, the number of incidents in any measure is independent of any other.

The first two assumption ensure that for small enough δ, the probability of an incident during a measure is approximately zero, thus two coincident occurrences are prohibited.

Notation. Notate the distribution of the number of accidents in a Poisson process X, during an measure, as the Poisson distribution

$$X \sim \mathrm{Poi}(\lambda)$$

Proposition 1.25. For $\lambda = \mu$, $X \sim \mathrm{Poi}(\lambda)$ has pmf

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

and

$$\begin{aligned} E(X) &= \mu \\ \mathrm{Var}(X) &= mu \end{aligned}$$

Remark. Split the measure $[0, t]$ into n intervals of length $\delta = \frac{t}{n}$, and set up a Bernoulli trial in each interval, where success means an incident occurred. Then taking the limit as n gets bigger, and λ gets smaller,

$$X(t) \sim Po(\lambda) \approx \mathrm{Bin}(n, \lambda\delta) = \mathrm{Bin}(n, \tfrac{\lambda t}{n})$$

Hence

$$\begin{aligned} E\big(X(t)\big) &= n\frac{\lambda t}{n} = \lambda t \\ E\big(X(1)\big) &= \lambda \end{aligned}$$

Thus $\lambda = \mu$ is the mean number of incidents in a measure of 1 unit.

1.4 Continuous random variables

1.4.1 Probability density functions

Definition (Continuous random variable)**.** A random variable is a function from a sample space Ω to $\mathbb{R}$. For associated event space $\mathcal{F}$, probability function P and random variable $X : \Omega \to \mathbb{R}$, for any real $a, b \in \mathbb{R}$,

$$a \leq X \leq b = \{\omega : a \leq X(\omega) \leq b\} \in \mathcal{F}$$

In particular, the cumulative distribution function may be written

$$F_X(x) = P(X \leq x)$$

Definition (Probability density function)**.** If F_X is continuous and differentiable, with derivative f_X, then X is a continuous random variable. f_X is the probability density function of X. If unambiguous, the subscript X may be omitted.

Differentiation and integration are inverse operations, and $F(-\infty) = 0$, so

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$$
$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)\,\mathrm{d}x \quad \text{for} \quad a \leq b$$

This gives the interpretation that the probability that X lies in an interval $(a, b]$ is given by the area under the graph of the pdf between a and b.

Remark.

$$F(\infty) = P(X \leq \infty) = \int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$$

Proposition 1.26. The conditions required for a function f to be a pdf are

$$\forall x \in \mathbb{R} \quad f(x) \geq 0 \tag{1.32}$$
$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \tag{1.33}$$

Remark. A continuous random variable may be described completely by knowing either F or f, since one may be obtained from the other.

Remark. If X is continuous, then F is continuous, so

$$P(X = x) = \lim_{\varepsilon \to 0} P(x - \varepsilon < X \leq x) = \lim_{\varepsilon \to 0} F(x) - F(x - \varepsilon) = 0$$

Remark. F may be neither discontinuous on a countable set nor continuous and non-differentiable. This may be harder to deal with, but these situation are unlikely to arise. As an exception, F may be a mix of the two, which is straightforward to handle.

Definition (Median)**.** The median satisfies $F(x) = \frac{1}{2}$.

Definition (Mode)**.** The mode of X is the value of x for which $f(x)$ is a maximum.

1.4.2 Expectation and varience

Remark. From the definition of differentiation,

$$f(x) = \lim_{\delta x \to 0} \frac{F(x + \delta x) - F(x)}{\delta x} = \lim_{\delta x \to 0} \frac{P(x < X < x + \delta x)}{\delta x}$$

Hence $f(x)$ must be non-negative everywhere. Providing δx is small,

$$P(x < X < x + \delta x) \approx f(x)\,\delta x$$

so a continuous random variable may be approximated by a discrete one which takes values on a finely spaced grid of points, separated by intervals of length

δx.Then the probability associated with an interval starting at point x would be $f(x)\,\delta x$.

The expectation of such a discrete random variable is the sum over every x value in the grid of

$$\sum x f(x)\,\delta x$$

As $\delta x \to 0$, this sum tends to an integral, motivating the following definitions. Generally, continuous random variable formulae look similarly to discrete random variable formulae.

Definition (Expectation)**.** For continuous random variable X with density f, the expected value of X is

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

if the integral converges absolutely

$$\int_{-\infty}^{\infty} |x|\, f(x)\,\mathrm{d}x < \infty$$

Definition (Variance)**.** For continuous random variable X with density f and expectation $\mu = E(X)$,

$$\mathrm{Var}(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$$

Remark. Expectation and variance for continues random variable inherit the same properties as for the discrete case.

1.4.3 Moment generating function

Definition (Moment generating function)**.** The moment generating function (mgf) of a continuous random variable X is

$$M_X(t) = E(e^{tX})$$

for those values of t where the expectation exists.

Remark. Note that

$$M(t) = E(e^{tX}) = \left(1 + tX + \frac{(tX)^2}{2!} + \cdots\right)$$

At $t = 0$,

$$M(0) = E(e^{0X}) = E(1) = 1$$

Differentiating with respect to t gives

$$\begin{aligned} M'(t) &= E\left(X + tX^2 + \frac{t^2X^3}{2!} + \cdots\right) \\ M'(0) &= E(X + 0 + \cdots) = E(X) \\ M''(t) &= E\left(X^2 + tX^3 + \frac{t^2X^4}{2!} + \cdots\right) \\ M''(0) &= E(X^2 + 0 + 0 + \cdots) = E(X^2) \end{aligned}$$

The k^{th} moment of X is

$$M^{(k)}(0) = E(X^k)$$

1.4.4 Functions of random variables

Proposition 1.27. For $Y = g(X)$ for a strictly monotone function g, where X has pdf f_X, the pdf of Y is

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

Proof. Suppose g is strictly increasing, that is,

$$x_1 < x_2 \implies g(x_1) < g(x_2)$$

Then the cdf of Y is

$$F_Y = P(Y \leq y) = P\big(g(X) \leq y\big) = P\big(X \leq g^{-1}(y)\big) = F_X\big(g^{-1}(y)\big)$$

Differentiating gives

$$f_Y(y) = \frac{\mathrm{d}F_Y}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}F_X\big(g^{-1}(y)\big) = f_X\big(g^{-1}(y)\big)\frac{\mathrm{d}}{\mathrm{d}y}\big(g^{-1}(y)\big) = f_X\big(g^{-1}(y)\big)\frac{\mathrm{d}x}{\mathrm{d}y}$$

and similarly for a strictly decreasing sequence. □

Remark. If g is not monotonic, then proceed from first principles.

Example 1.4. For $Z \sim \mathcal{N}(0,1)$ and $Y = Z^2$, what is the distribution of Y?
$g(z) = z^2$ is not monotonic in $\mathbb{R}$. For $y > 0$,

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) = P(Z^2 < y) \\
&= P\big(-y^{-\frac{1}{2}} \leq Z \leq y^{\frac{1}{2}}\big) \\
&= \Phi(y^{\frac{1}{2}}) - \Phi(-y^{\frac{1}{2}}) \\
&= \Phi(y^{\frac{1}{2}}) - \big(1 - \Phi(y^2\frac{1}{2})\big) \\
f_Y(y) &= \frac{\mathrm{d}F_Y}{\mathrm{d}y} \\
&= 2\phi(y^{\frac{1}{2}})\frac{1}{2}y^{-\frac{1}{2}} \\
&= 2\frac{1}{\sqrt{2\pi}}e^{\frac{-y}{2}}\frac{y^{-\frac{1}{2}}}{2} \\
&= \left(\tfrac{1}{2}\right)^{\frac{1}{2}}y^{-\frac{1}{2}}e^{-\frac{y}{2}}\frac{1}{\sqrt{\pi}} \\
&= \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}y^{-\frac{1}{2}}e^{-\frac{y}{2}}}{\Gamma(\frac{1}{2})}
\end{aligned}$$

This is exactly the pdf for $\Gamma(\frac{1}{2}, \frac{1}{2})$, hence $Y \sim \Gamma(\frac{1}{2}, \frac{1}{2})$.

1.5 Standard continuous distributions

1.5.1 Uniform

Definition (Uniform distribution)**.** A continuous random variable X has uniform (or rectangular) distribution, written $X \sim \mathrm{U}(a, b)$, if it has

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Proposition 1.28. This is a valid pdf

Proof.

(1.32) $f(x) \geq 0$ everywhere is trivial.

(1.33) The integral is just the area of the rectangle, $\frac{1}{b-a}(b-a) = 1$. □

Remark. For $(c, c+d) \in (a, b)$,

$$P(c < X \leq c+d) = \int_c^{c+d} \frac{\mathrm{d}x}{b-a} = \frac{1}{b-a} x\Big|_c^{c+d} = \frac{d}{b-a}$$

Hence the probability X lies in an interval lies only in its length, not its location.

Remark. The cdf of X is

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u = \begin{cases} 0 & x < a \\ \int_a^x \frac{\mathrm{d}u}{b-a} = \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

Proposition 1.29. The expectation and variance of $X \sim U(a, b)$ are

$$E(X) = \frac{a+b}{2} \tag{1.34}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12} \tag{1.35}$$

Proof.

(1.34)

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \\ &= \int_a^b \frac{x}{b-a}\,\mathrm{d}x \\ &= \frac{x^2}{2(b-a)}\Big|_a^b \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2} \end{aligned}$$

(1.35)

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \\
&= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{\mathrm{d}x}{b-a} \\
&= \frac{1}{3(b-a)} \left(x - \frac{a+b}{2}\right)^3 \Big|_a^b \\
&= \frac{1}{24(b-a)} \big((b-a)^3 - (a-b)^3\big) \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

□

Remark. The uniform distribution is often too simple to be useful, but may be helpful for the simulation of random numbers, since $U(0,1)$ may be transformed into almost any other distribution.

1.5.2 Exponential

Definition (Exponential distribution)**.** For incidents which occur at a Poisson rate λ, the length of the interval X until the first incident occurs is distributed with (negative) exponential distribution

$$X \sim \mathrm{Exp}(\lambda)$$

Proposition 1.30. $X \sim \mathrm{Exp}(\lambda)$ has pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Proof. The number of incidents $N(t)$ occurring in an interval $(0,t)$ is distributed $N(t) \sim \mathrm{Poi}(\lambda t)$, so cdf of X at x is

$$\begin{aligned}
F(x) &= P(X \le x) \\
&= P\big(\text{incident in } (0,x]\big) \\
&= P\big(N(x) \ge 1\big) \\
&= 1 - P\big(N(x) = 0\big) \\
&= 1 - \frac{(\lambda x)^0 e^{-\lambda x}}{0!} \\
&= 1 - e^{-\lambda x}
\end{aligned}$$

Differentiating then gives the pdf. Check that this is a pdf.

(1.32) $f(x) \ge 0$ everywhere is trivial.

(1.33)

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = \int_0^{\infty} \lambda e^{\lambda x}\,\mathrm{d}x = -e^{-\lambda x}\big|_0^{\infty} = 0 - (-1) = 1$$

□

Proposition 1.31. The expectation and variance of $X \sim \text{Exp}(\lambda)$ are

$$E(X) = \frac{1}{\lambda} \tag{1.36}$$

$$\text{Var}(X) = \frac{1}{\lambda^2} \tag{1.37}$$

Proof. Find the moment generating function for $X \sim \text{Exp}(\lambda)$.

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= \int_{\mathbb{R}} e^{tx} f(x) \, \mathrm{d}x \\ &= \lambda \int_0^\infty e^{tx} e^{-\lambda x} \, \mathrm{d}x \\ &= \lambda \frac{e^{(t-\lambda)x}}{t-\lambda} \Big|_0^\infty \end{aligned}$$

This is only finite for $t < \lambda$. Since λ is fixed and positive, $t > 0$ can always be picked such that this is the case.

$$M(t) = \lambda \left(0 - \frac{1}{t-\lambda} \right) = \frac{\lambda}{\lambda - t}$$

The moments can be found by differentiating with respect to t.

$$\begin{aligned} M'(t) &= \frac{\lambda}{(\lambda - t)^2} \\ M''(t) &= \frac{2\lambda}{(\lambda - t)^3} \end{aligned}$$

Only $n = 1, 2$ are needed for this proof, but the general case is included for completion.

$$M^{(n)}(t) = \frac{n!\lambda}{(\lambda - t)^{r+1}}$$

Then

$$\begin{aligned} E(X) &= M'(0) = \frac{1}{\lambda} \\ E(X^2) &= M''(0) = \frac{2}{\lambda^2} \\ E(X^n) &= M^{(n)}(0) = \frac{n!}{\lambda^n} \end{aligned}$$

So

$$E(X) = \frac{1}{\lambda} \tag{1.36}$$

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2} \tag{1.37}$$

□

Proposition 1.32. The median of $X \sim \text{Exp}(\lambda)$ is $\frac{\ln 2}{\lambda}$, so such a distribution cannot have the same median as mean.

Proof. For the median m,

$$\begin{aligned} F(m) &= \frac{1}{2} \\ 1 - e^{-\lambda m} &= \frac{1}{2} \\ m &= \frac{\ln 2}{\lambda} \end{aligned}$$

Since the mean is $\frac{1}{\lambda}$, this cannot be equal to the median. □

Remark. The mean time between incidents is $\frac{1}{\lambda}$ for $X \sim \text{Exp}(\lambda)$. If the last incident was L units prior, what is the distribution of the interval until the next emission. In other words, what is the distribution of $X - L$ given that $X > L$?

$$\begin{aligned} P(X - L > x \mid X > L) &= \frac{P\big((X - L > x) \cap (X > L)\big)}{P(X > L)} \\ &= \frac{P\big((X > L + x) \cap (X > L)\big)}{P(X > L)} \\ &= \frac{P(X > L + x)}{P(X > L)} \\ &= \frac{e^{-\lambda(L+x)}}{e^{-\lambda L}} = e^{-\lambda x} \end{aligned}$$

Hence $X - L \mid X > L \sim \text{Exp}(\lambda)$. The distribution of length of remaining interval until the next incident does not depend on the previous incident. This is called the lack of memory property, which is a result of independence between non-overlapping intervals in Poisson process.

Remark. Just as the Poisson is the limit of a binomial distribution as $n \to \infty$, for fixed np, the exponential distribution is the limit of a geometric.

Remark. The exponential distribution is often used to model things like the times to failure of electrical components, or the lifespans of individuals.

1.5.3 Gamma

Remark. Just as the exponential distribution may be seen as the limit of a geometric distribution, since it is the distribution of the interval until the first incident in a Poisson process, the gamma distribution may be considered as the limit of the negative binomial distribution, and used to model the distribution of interval until the k^{th} incident.

The Poisson process and negative binomial both involve factorials, so clearly a continuous extension to this will be needed for the gamma distribution.

Definition (Gamma function)**.** The gamma function, for any $\alpha > 0$, is

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, \mathrm{d}x$$

Remark. The Gamma function usually cannot be evaluated analytically, but it has several useful properties. It is called the generalised factorial, since it agrees with the factorial for $\alpha \in \mathbb{N}$, but is also defined for non-integers.

Proposition 1.33. It may be useful to know that

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

Proof is in section 1.5.4.

Proposition 1.34.

$$\forall \alpha > 0 \quad \Gamma(\alpha+1) = \alpha\Gamma(\alpha) \tag{1.38}$$
$$\forall \alpha \in \mathbb{N} \quad \Gamma(\alpha) = (\alpha-1)! \tag{1.39}$$

Proof.

(1.38) Integrate by parts. For $\alpha > 1$.

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x \\ &= -x^{\alpha-1}e^{-x}\Big|_0^\infty + \int_0^\infty (\alpha-1)x^{\alpha-2}e^{-x}\,\mathrm{d}x \\ &= 0 + (\alpha-1)\int_0^\infty x^{\alpha-2}e^{-x}\,\mathrm{d}x \\ &= (\alpha-1)\Gamma(\alpha-1)\end{aligned}$$

Hence substituting $\alpha+1$ with $\alpha > 0$ in place of α gives the required result.

(1.39) Proof by induction. For $\alpha = 1$,

$$\Gamma(1) = \int_0^\infty e^{-x}\,\mathrm{d}x = \frac{e^{-x}}{-1}\Big|_0^\infty = 0 - (-1) = 1$$

The recurrence relation (1.38) gives the result by induction. □

Remark. Let X be the interval until the k^{th} accident in a Poisson process. Then X takes values in $\mathbb{R}^+$, and for $x > 0$ has cdf

$$\begin{aligned}F(x) &= 1 - P(X > x) \\ &= 1 - P\big(\text{at most } k-1 \text{ incidents in } (0,x]\big) \\ &= 1 - \sum_{n=0}^{k-1}\frac{e^{\lambda x}(\lambda x)^n}{n!}\end{aligned}$$

Integrate to get the pdf.

$$\begin{aligned}f(x) = F'(x) &= -\sum_{n=0}^{k-1}\frac{1}{n!}\frac{\mathrm{d}}{\mathrm{d}x}\big(e^{-\lambda x}(\lambda x)^n\big) \\ &= -\sum_{n=0}^{k-1}\frac{1}{n!}\left((\lambda x)^n\frac{\mathrm{d}}{\mathrm{d}x}(e^{-\lambda x}) + e^{-\lambda x}\frac{\mathrm{d}}{\mathrm{d}x}(\lambda x)^n\right) \\ &= \sum_{n=0}^{k-1}\frac{1}{n!}\lambda(\lambda x)^n e^{-x} - \sum_{n=1}^{k-1}\frac{1}{n!}\lambda n e^{-\lambda x}(\lambda x)^{n-1}\end{aligned}$$

Substitute $\ell = n-1$ into the second sum.

$$= \sum_{n=0}^{k-1} \frac{1}{n!}\lambda(\lambda x)^n e^{-x} - \sum_{\ell=0}^{k-2} \frac{1}{\ell!}\lambda n e^{-\lambda x}(\lambda x)^\ell$$
$$= \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$$

This motivates the following definition.

Definition (Gamma distribution)**.** A continuous random variable X has a gamma distribution with parameters $a, \lambda > 0$, written $X \sim \Gamma(\alpha, \lambda)$, if it has pdf

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Proposition 1.35. This is a valid pdf.

Proof. This is a valid pdf since it is non-negative everywhere, and

$$\int_{\mathbb{R}} f(x)\ \mathrm{d}x = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\lambda x}\ \mathrm{d}x$$

Substitute $u = \lambda x$.

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{u}{\lambda}\right)^{\alpha-1} e^{-u} \frac{\mathrm{d}u}{\lambda}$$
$$= \frac{1}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} e^{-u}\ \mathrm{d}u$$
$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1$$

□

Remark. As should be expected, $\Gamma(1, \lambda)$ is the same distribution as $\mathrm{Exp}(\lambda)$.

Remark. The gamma distribution is useful in many more situations than that which motivated it, due to the flexible range of shapes it offers. α is called the index or shape parameter, and λ is called the scale parameter, or sometimes just the parameter.

Proposition 1.36. The expectation and variance of $X \sim \Gamma(\alpha, \lambda)$ are

$$E(X) = \frac{\alpha}{\lambda} \tag{1.40}$$
$$\mathrm{Var}(X) = \frac{\alpha}{\lambda^2} \tag{1.41}$$

Proof. Find the moment generating function for $X \sim \Gamma(\alpha, \lambda)$.

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= \int_{\mathbb{R}} e^{tx} f(x)\ \mathrm{d}x \\ &= \int_0^\infty e^{tx} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}\ \mathrm{d}x \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x}\ \mathrm{d}x \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \left(\frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} \mathrm{d}x \right) \end{aligned}$$

For $t < \lambda$, the bracketed expression is the integral of the pdf of the distribution $X \sim \Gamma(\alpha, \lambda - t)$, so evaluates to 0. Since $\lambda > 0$, such a t can always be found, hence

$$M(t) = \frac{\lambda^\alpha}{(\lambda - t)^\alpha}$$

The moments can be found by differentiating with respect to t.

$$\begin{aligned} M'(t) &= \frac{\alpha\lambda^\alpha}{(\lambda - t)^{\alpha+1}} \\ M''(t) &= \frac{\alpha(\alpha+1)\lambda^\alpha}{(\lambda - t)^{\alpha+2}} \end{aligned}$$

Then

$$\begin{aligned} E(X) = M'(0) &= \frac{\alpha}{\lambda} \\ E(X^2) = M''(0) &= \frac{\alpha(\alpha+1)}{\lambda^2} \end{aligned}$$

So

$$E(X) = \frac{\alpha}{\lambda} \tag{1.40}$$

$$\operatorname{Var}(X) = E(X^2) - E(X)^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2} \tag{1.41}$$

□

1.5.4 Beta

Definition. The beta function, for $\alpha, \beta > 0$, is

$$\mathrm{B}(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x$$

Proposition 1.37.

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{1.42}$$

$$\forall \alpha, \beta \in \mathbb{N},\ a, b > 1 \quad \mathrm{B}(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} \tag{1.43}$$

Proof.

$$\begin{aligned} \Gamma(\alpha)\Gamma(\beta) &= \int_{u=0}^\infty e^{-u}u^{\alpha-1}\,\mathrm{d}u \cdot \int_{v=0}^\infty e^{-v}v^{\beta-1}\,\mathrm{d}v \\ &= \int_{v=0}^\infty \int_{u=0}^\infty e^{-u-v}u^{\alpha-1}v^{\beta-1}\,\mathrm{d}u\,\mathrm{d}v \end{aligned}$$

Let $u = zt,\ v = z(1-t)$. This has Jacobian

$$\begin{pmatrix} u_z & u_t \\ v_z & v_t \end{pmatrix} = \begin{pmatrix} t & z \\ 1-t & -z \end{pmatrix} = -z$$

So $\mathrm{d}u\ \mathrm{d}v = z\ \mathrm{d}z\ \mathrm{d}t$. $z = u + v$ and $t = \frac{u}{u+v}$, so the limits of integration for z are 0 to ∞, and for t are 0 to 1. So

$$\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) &= \int_{z=0}^{\infty}\int_{t=0}^{1} e^{-z}(zt)^{\alpha-1}\big(z(1-t)\big)^{\beta-1} z\ \mathrm{d}t\ \mathrm{d}z \\
&= \int_{z=0}^{\infty} e^{-z} z^{\alpha+\beta-1}\ \mathrm{d}z \cdot \int_{t=0}^{1} t^{\alpha-1}(1-t)^{\beta-1}\ \mathrm{d}t \\
&= \Gamma(\alpha+\beta)\,\mathrm{B}(\alpha,\beta)
\end{aligned}$$

Hence equation (1.42) is true. equation (1.43) is immediate from equations (1.39) and (1.42). ☐

Definition (Beta distribution)**.** A continuous random variable X has a beta distribution with parameters $\alpha, \beta > 0$, written $X \sim \mathrm{B}(\alpha, \beta)$, if it has pdf

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

This is a valid pdf since it is non-negative everywhere, and integrating gives 1 by the definition of the beta function

Proposition 1.38. The expectation and variance of $X \sim \mathrm{B}(\alpha, \beta)$ are

$$E(X) = \frac{\alpha}{\alpha+\beta} \tag{1.44}$$

$$\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{1.45}$$

Proof. The r^{th} moment of X about 0 is

$$\begin{aligned}
E(X^r) &= \frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 x^r x^{\alpha-1}(1-x)^{\beta-1}\ \mathrm{d}x \\
&= \frac{\mathrm{B}(\alpha+r,\beta)}{\mathrm{B}(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+r)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+r)} \\
&= \frac{\alpha(\alpha+1)\cdots(\alpha+r-1)\Gamma(\alpha)\Gamma(\alpha+\beta)}{\Gamma(\alpha)(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+r-1)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha(\alpha+1)\cdots(\alpha+r-1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+r-1)} \\
&= \prod_{i=0}^{r-1}\frac{\alpha+i}{\alpha+\beta+i}
\end{aligned}$$

The expectation follows immediately, and

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\ &= \frac{\alpha(\alpha+1)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\end{aligned}$$

□

Remark. The beta family is extremely flexible. It might be expected to provide a good model for the distribution of any random variable which arises as a proportion (the integrand looks like a binomial).

Proof of Proposition 1.33.

$$\begin{aligned}B(\tfrac{1}{2}, \tfrac{1}{2}) &= \int_0^1 x^{\frac{1}{2}-1}(1-x)^{\frac{1}{2}-1}\,\mathrm{d}x \\ &= \int_0^1 \frac{\mathrm{d}x}{\sqrt{x}\sqrt{1-x}}\end{aligned}$$

Substitute $u = \sqrt{x}$. Then $\mathrm{d}u = \frac{1}{2u}\,\mathrm{d}x$, so

$$\begin{aligned}&= \int_0^1 \frac{2u}{\sqrt{1-xu}}\,\mathrm{d}u \\ &= 2\int_0^1 \frac{\mathrm{d}u}{\sqrt{1-u^2}} \\ &= 2 \lim_{\substack{b\to 1^- \\ a\to 0^+}} \big[\arcsin(u)\big]_a^b \\ &= \pi\end{aligned}$$

So

$$\begin{aligned}B(\tfrac{1}{2}, \tfrac{1}{2}) &= \frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2}+\frac{1}{2})} \\ \Gamma(\tfrac{1}{2})^2 &= 1\pi \\ \Gamma(\tfrac{1}{2}) &= \sqrt{\pi}\end{aligned}$$

□

1.5.5 Normal

Definition (Normal distribution)**.** A continuous random variable X has a normal (or Gaussian)distribution with parameters μ, σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if it has pdf

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Proposition 1.39. This is a valid pdf.

Proof. Clearly it is non-negative.

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{\sigma^2}}\,\mathrm{d}x$$

Substitute $z = \frac{x-\mu}{\sigma}$.

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \cdot 2 \int_{0}^{\infty} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\
&= \sqrt{\frac{2}{\pi}} \left(\int_0^\infty e^{-\frac{1}{2}z^2}\,\mathrm{d}z \int_0^\infty e^{-\frac{1}{2}y^2}\,\mathrm{d}y \right)^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{\pi}} \left(\int_0^\infty \int_0^\infty e^{-\frac{1}{2}(y^2+z^2)}\,\mathrm{d}z\,\mathrm{d}y \right)^{\frac{1}{2}}
\end{aligned}$$

Substitute $y = sz$.

$$\begin{aligned}
&= \sqrt{\frac{2}{\pi}} \left(\int_0^\infty \int_0^\infty e^{-\frac{1}{2}(s^2z^2+z^2)} z\,\mathrm{d}z\,\mathrm{d}s \right)^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{\pi}} \left(\int_0^\infty \int_0^\infty e^{-\frac{1}{2}z^2(1+s^2)} z\,\mathrm{d}z\,\mathrm{d}s \right)^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{\pi}} \left(\int_0^\infty \left[-\frac{1}{1+s^2} e^{-\frac{1}{2}z^2(1+s^2)} \right]_0^\infty \mathrm{d}s \right)^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{\pi}} \left(\int_0^\infty \frac{1}{1+s^2}\,\mathrm{d}s \right)^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{\pi}} \left(\big[\arctan(s) \big]_0^\infty \right)^{\frac{1}{2}} \\
&= \sqrt{\frac{2}{\pi}} \left(\frac{\pi}{2} \right)^{\frac{1}{2}} = 1
\end{aligned}$$

□

Remark. A normal distribution is symmetric about its mean, and is unimodal. It is often described as bell-shaped.

Proposition 1.40. The expectation and variance of $X \sim \mathcal{N}(\mu, \sigma^2)$ are

$$E(X) = \mu \tag{1.46}$$

$$\mathrm{Var}(X) = \sigma^2 \tag{1.47}$$

Proof.

(1.46)

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} (x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x + \mu\int_{-\infty}^{\infty}\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}(-\sigma^2)e^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big|_{-\infty}^{\infty} + \mu \\
&= \mu
\end{aligned}$$

(1.47)

$$\begin{aligned}
\mathrm{Var}(X) &= E\left((X-\mu)^2\right) \\
&= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}(x-\mu)^2e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x
\end{aligned}$$

Substitute $t = \frac{x-\mu}{\sigma}$.

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma^2t^2e^{-\frac{t^2}{2}}\,\mathrm{d}t$$

Use integration by parts.

$$\begin{aligned}
&= \frac{\sigma^2}{\sqrt{2\pi}}\left[-te^{\frac{-t^2}{2}}\right]_{-\infty}^{\infty} + \sigma^2\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\,\mathrm{d}t \\
&= 0 + \sigma^2 = \sigma^2
\end{aligned}$$

□

Definition (Standard normal)**.** The standard normal distribution $\mathcal{N}(\mu,\sigma^2)$ has pdf $\phi(x)$ and cdf $\Phi(x)$ of

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}, \quad \Phi(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}u^2}\,\mathrm{d}u$$

Proposition 1.41. Any normal distribution $\mathcal{N} \sim (0,1)$ and the standard normal distribution $Z \sim \mathcal{N}(0,1)$ are related by

$$Z = \frac{X-\mu}{\sigma}$$

Proof. For $Z = g(X) = \frac{X-\mu}{\sigma}$, $X = g^{-1}(Z) = \sigma Z + \mu$. σ is positive, so $g(X)$ is strictly increasing. So

$$\begin{aligned}
f_Z(z) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(g^{-1}(z)-\mu)^2}{2\sigma^2}}\frac{\mathrm{d}g^{-1}(z)}{\mathrm{d}z} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\sigma z+\mu-\mu)^2}{2\sigma^2}}\frac{\mathrm{d}(\sigma z+\mu)}{\mathrm{d}z} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{z^2}{2}}\sigma \\
&= \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} = \phi(z)
\end{aligned}$$

Hence Z has a standard normal distribution. □

Method. For $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

These values of Φ can then be found in tables to calculate the probabilities. Also, use the median $\Phi(0) = 0.5$ and the symmetry

$$\Phi(-x) = 1 - \Phi(x)$$

Remark. The normal distribution is important for several reasons.

1. Many natural phenomena are approximately normally distributed, and even more can be transformed to fit, such as taking logs or cubes.

2. Normal distribution provides a good model for random measurement errors in experiments.

3. Many useful normal approximations to other distributions exist, such as binomial, Poisson and gamma distributions.

4. Linear combinations of normal distributions are also normal.

5. The distribution of the sum of a large number of independent random variable from any distribution, non of which predominate, tends in the limit to a normal distribution (see **??**).

2 Example questions

Sheet 0

Question 1. *Explain the difference between the sample space and the event space. Write down the elements of the two spaces for the experiment where a coin is tossed once (use the notation H, T to denote heads and tails).* ***3 Marks***

The event space is the collection of outcomes; a subset of the sample space, possible individual outcomes. The sample space is $\Omega = \{H, T\}$ whilst the event space is $\mathcal{F} = \big\{\emptyset, \{H\}, \{T\}, \Omega\big\}$.

Question 2. *Consider an experiment with a countable (but possibly infinite) sample space ω in which the individual outcomes are denoted by $\omega_1, \omega_2, \ldots$. Let $p_1, p_2 \ldots$, be a corresponding set of non-negative 'weights' with $\sum_i p_i = 1$. For any event $E \subseteq \omega$ define*

$$P(E) = \sum_{i:\omega_i \in E} p_i$$

Show that the function $P()$ is a probability function (i.e. that it satisfies Kolmogorov's three axioms). ***4 Marks***

1) $P(E) \geq 0$ is obvious ✓
2) $P(\Omega) = \sum_i p_i = 1$ ✓
3) E, F disjoint.

$$P(E \cup F) = \sum_{i:\omega_i \in E \cup F} p_i = \sum_{i:w_i \in E} p_i + \sum_{j:w_j \in F} p_j = P(E) + P(F)$$

Question 3. *Use the axioms of probability to prove that*

$$P(E^c) = 1 - P(E)$$

3 Marks

Clearly $\Omega = E \cup E^c$ and E, E^c are disjoint. Then by equations (1.6) and (1.7),

$$1 = P(\Omega) = P(E) + P(E^c)$$

Question 4. *Use the axioms of probability to prove that for any two events E and F,*

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

4 Marks

The trick is to find something that is disjoint. By repeated use of equation (1.7),

$$
\begin{aligned}
E \cup F &= E \cup \big\{F \cap (E \cap F)^c\big\} \\
P(E \cup F) &= P(E) + P\big(F \cap (E \cap F)^c\big) \\
F &= (E \cap F) \cup \big(F \cap (E \cap F)^c\big) \\
P(F) - P(E \cap F) &= P\big(F \cap (E \cap F)^c\big)
\end{aligned}
$$

Question 5. *Is it possible for two events to be mutually exclusive and also independent?* ***3 Marks***

$$
\begin{aligned}
\text{Mutually exclusive} &\iff P(E \cap F) = 0 \\
&\iff P(E)P(F) = 0 \\
&\iff P(E) = 0 \text{ or } P(F) = 0
\end{aligned}
$$

Question 6. *Suppose that E and F are two events with $P(F) > 0$. Show that if $P(E|F) > P(E)$ then. $P(F|E) > P(F)$. Interpret this result in words.* ***4 Marks***

$$
P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(F|E)P(E)}{P(F)}
$$

so

$$
\begin{aligned}
&\frac{P(F|E)P(E)}{P(F)} > P(E) \\
&\implies P(F|E) > P(F)
\end{aligned}
$$

E and F tend to occur together more often than if they were independent.

Question 7. *A proportion of 1% of a population are known to have a particular disease. A screening test has a 95% chance of detecting the disease (i.e., of giving a positive result) in an individual who has the disease. There is also a 6% chance that a disease-free individual will give a positive result and be thought to have the disease. What is the probability that a positive result is obtained on an individual chosen at random from the population? If a positive result is obtained, what is the probability that the individual does have the disease?* ***2 Marks***

Let D be the probability individual has a disease and E be the probability of a positive test result. Then

$$
P(D) = 0.01, \quad P(E|D) = 0.95, \quad P(E|D^c) = 0.06
$$

So the probability of a positive result is

$$
P(E) = P(E|D)P(D) + P(E|D^c)P(D^) = 0.95 \times 0.01 + 0.06 \times 0.99 = 0.0689
$$

The probability of having the disease given a positive test is, by Bayes

$$P(D|E) = \frac{P(E|D)P(D)}{P(E)} = \frac{0.95 \times 0.01}{0.0689} = 0.138$$

Question 8. *An urn contains five red balls and one white ball. A ball is drawn and then it and another ball of the same colour are placed back in the urn. Finally a second ball is drawn.*

(a) What is the probability that the second ball is white?

(b) If the second ball is white, what is the probability that the first was red?

2 Marks

Let W_i be the probability the i^{th} ball is white, and R_i the corresponding probability for red.

(a) $P(W_2) = P(W_2|R_1)P(R) + P(W_2|W_1)P(W_1) = \frac{1}{7}\frac{5}{6} + \frac{2}{7}\frac{1}{6} = \frac{1}{6}$

(b) $P(R_1|W_2) = \frac{P(W_2|R_1)P(R_1)}{P(W_2)} = \frac{\frac{1}{7}\frac{5}{6}}{\frac{1}{6}} = \frac{5}{7}$

Sheet 1

Question 1. *For X the number of heads on two throws of a coin, find $E(X)$, $E(\frac{1}{1+X})$. Verify*

$$E\left(\frac{1}{1+X}\right) \neq \frac{1}{1+E(X)}$$

Find Var$(\frac{1}{1+X})$. ***6 marks***

By symmetry, $E(X) = 1$.

$$E\left(\frac{1}{1+X}\right) = \sum_{x=0}^{2} \frac{1}{1+x}P(x) = 1 \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{4} = \frac{7}{12}$$

Indeed

$$\frac{1}{1+E(X)} = \frac{1}{2} \neq \frac{7}{12} = E\left(\frac{1}{1+X}\right)$$

To find variance,

$$E\left(\left(\frac{1}{1+x}\right)^2\right) = 1 \times \frac{1}{4} + \frac{1}{4} \times \frac{1}{2} + \frac{1}{9} \times \frac{1}{4} = \cdots$$

So

$$\text{Var}\left(\frac{1}{1+x}\right) = E\left(\left(\frac{1}{1+x}\right)^2\right) - \left(E\left(\frac{1}{1+X}\right)\right)^2 = \cdots$$

Question 2. *A sample of two balls is selected at random and without replacement from an urn containing five red, three blue and two white balls. You win* *two pounds for each blue ball chosen and lose one pound for each red ball. Let X be your winnings. Write down the possible values of X and the corresponding probability mass function. Find your expected profit. What is the probability that you lose more than a pound, given that you do make a loss?* **5 marks**

X takes values $-2, -1, 0, 1, 2, 4$.

$$\begin{aligned}
P(X = -2) &= P(\{RR\}) = \frac{5}{10} \times \frac{4}{9} = \frac{2}{9} \\
P(X = -1) &= P(\{RW, WR\}) = 2 \times \frac{5}{10} \times \frac{2}{9} = \frac{2}{9} \\
P(X = 0) &= P(\{WW\}) = \frac{2}{10} \times \frac{1}{9} = \frac{1}{45} \\
P(X = 1) &= P(\{RB, BR\}) = 2 \times \frac{5}{10} \times \frac{3}{9} = \frac{1}{3} \\
P(X = 2) &= P(\{BW, WB\}) = 2 \times \frac{2}{10} \times \frac{3}{9} = \frac{2}{15} \\
P(X = 3) &= 0 \\
P(X = 2) &= P(\{BB\}) = \frac{3}{10} \times \frac{2}{9} = \frac{1}{15}
\end{aligned}$$

Adding these up should give 1.

$$E(X) = -2 \times \frac{2}{9} - 1 \times \frac{2}{9} + \frac{1}{3} + 2 \times 215 + 4 \times 415 = £0.20$$

Finally,

$$P(X < -1 \mid X < 0) = P(X = -2 \mid X < 0) = \frac{P(X = -2)}{P(X < 0)} = \frac{\frac{2}{9}}{\frac{2}{9} + \frac{2}{9}} = \frac{1}{2}$$

Question 3. *In a biochemical experiment, n organisms are placed in a nutrient medium and X, the number of organisms that survive for a given period, is recorded. If*

$$P(X = r) = \frac{2(r+1)}{(n+1)(n+2)}$$

for $r = 0, \ldots, n$ and is zero otherwise, calculate the probability that at most a proportion $\alpha = \frac{k}{n}$ of the organisms survive. Deduce that for large n, this probability is approximately α^2. Find the smallest value of n for which the probability of there being at least one survivor among the n organisms is at least 0.95. **4 marks**

The probability that at most a proportion $\alpha = \frac{k}{n}$ survive is

$$\begin{aligned}
P(X \leq k) &= \sum_{r=0}^{k} P(X = r) \\
&= \frac{2}{(n+1)(n+2)} \sum_{r=0}^{k} (r+1) \\
&= \frac{(k+1)(k+2)}{(n+1)(n+2)} \\
&= \frac{(n\alpha+1)(n\alpha+2)}{(n+1)(n+2)} \to \alpha^2 \quad \text{as} \quad n \to \infty
\end{aligned}$$

The smallest value of n for which the probability of there being at least one survivor among the n organisms is at least 0.95 is

$$\begin{aligned}
0.95 &\leq P(X \geq 1) \\
&\leq 1 - P(X = 0) \\
P(X = 0) &\leq 0.05 \\
\frac{2}{(n+1)(n+2)} &\leq \frac{1}{20} \\
(n+1)(n+2) &\geq 40 \\
n &\geq 5
\end{aligned}$$

Question 4. *Meteorologists are required to issue daily forecasts of the probability of rainfall in a certain location. In an attempt to make them more accountable, or to save money, it is proposed to link their salary to their forecast performance. Performance is measured by scoring each forecast probability* $\hat{p}$ *against the subsequent outcome*

$$X = \begin{cases} 1 & \textit{rain } p \\ 0 & \textit{no rain } (1-p) \end{cases}$$

It is proposed to define a the forecast score as

$$S = X \ln \hat{p} + (1 - X) \ln(1 - \hat{p})$$

(a) *If the true probability of rain is p, find the forecast value* $\hat{p}$ *that maximises the expected score. If forecasters receive salary bonuses when they score highly according to this rule, how should they act so as to maximise their salaries in the long run?*

(b) *In the West of Ireland, rain has occurred on 69% of all days in the last 50 years. In New South Wales, Australia, rain has occurred on 24% of all days. Would it be appropriate to use this scoring rule to compare the performance of Irish and Australian forecasters?*

5 marks

(a)

$$\begin{aligned} E(S) &= \ln \hat{p}\, E(X) + \ln(1-\hat{p})\big(1 - E(X)\big) \\ \frac{\partial E(S)}{\partial \hat{p}} &= \frac{p}{\hat{p}} - \frac{1-p}{1-\hat{p}} \end{aligned}$$

When this is zero,

$$\hat{p} = p$$

(b) The score S is symmetric for $\hat{p}$ and $1-\hat{p}$. But $1 - 0.24 \neq 0.69$, so it still isn't quite fair.