Required for: MATH40005 Probability and Statistics

Based on the lectures of Almut Veraart and Dean Bodenham, Imperial College London

1 The Sample Space

1.0.1 Definition

The sample space Ω is the set of all possible outcomes of an experiment. $\omega \in \Omega$ is a sample point.

1.0.2 Definition

A subset of Ω is an event.

1.0.3 Definition

Suppose A and B are events.
The union $A \cup B$ denotes the event that at least one of A or B occur.
The intersection $A \cap B$ denotes the event that both A and B occur.
$A^c = \Omega \setminus A$ is the complement of A in Ω.

2 Interpretations of Probability

2.0.1 Definition: Naive Definition of Probability

Suppose Ω is finite (or of finite area, for instance). The naive probability of the event A is

$$P_{\text{Naive}}(A) = \frac{|A|}{|\Omega|}.$$

This definition assumes each sample point has the same weight.

2.0.2 Definition: Limiting Frequency Definition of Probability

Let an experiment be replicated N times and let the event A occur n of those times. Another definition of the probability of the event A is

$$P_{\text{Limiting}}(A) = \lim_{N \to \infty} \frac{n}{N}.$$

2.0.3 Definition: Subjective Definition of Probability

For an event A, $P(A)$ may be assigned based on personal belief, or according to past information, potentially differing between individuals. While difficult to quantify or apply, subjectivity is a widely accepted interpretation of probability.

3 Counting

3.1 The Multiplication Principle

3.1.1 Theorem

For two experiments, A which has a possible outcomes and B which has b possible outcomes, performing A and B once each in any order has ab possible outcomes.

3.2 Power Sets

3.2.1 Definition

The power set of a set S, $\mathcal{P}(S)$ is the set of all subsets of S.

3.2.2 Theorem

For a finite sample space Ω, $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$.

3.3 Sampling With and Without Replacement

3.3.1 Theorem: Sampling Without Replacement - Ordered

For an ordered sample of k items without replacement from n items, where Ω is the set of possible samples, $|\Omega| = \frac{n!}{(n-k)!}$ (sometimes written ${}^nP_k$).

3.3.2 Theorem: Sampling Without Replacement - Unordered

For an unordered sample of k items without replacement from n items, where Ω is the set of possible samples, $|\Omega| = \frac{n!}{k!(n-k)!}$ (sometimes written ${}^nC_k$ or as the binomial $\binom{n}{k}$).

3.3.3 Theorem: Sampling With Replacement - Ordered

For an ordered sample of k items with replacement from n items, where Ω is the set of possible samples, $|\Omega| = n^k$.

3.3.4 Theorem: Sampling With Replacement - Unordered

For an unordered sample of k items with replacement from n items, where Ω is the set of possible samples, $|\Omega| = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$.

3.3.5 Table of 3.3.1-4

The cardinality of the sample space for each type of sampling above is given in the following table:

	Ordered	Unordered
Without replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$
With replacement	n^k	$\binom{n+k-1}{k}$

4 Axiomatic Definition of Probability

4.1 The Event Space ($\mathcal{F}$)

4.1.1 Definition

A set of subsets of Ω (events), $\mathcal{F}$, is an algebra if

i $\emptyset \in \mathcal{F}$.

ii $\mathcal{F}$ is closed under complement: $A \in \mathcal{F} \implies A^c \in \mathcal{F}$.

iiia $\mathcal{F}$ is closed under pairwise union: $A_1, A_2 \in \mathcal{F} \implies A_1 \cup A_2 \in \mathcal{F}$.

and is a σ-algebra if additionally

iiib $\mathcal{F}$ is closed under countable union: $A_1, A_2, A_3, ... \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

4.1.2 Corollaries

1. All algebras contain Ω.
2. All algebras are closed under both finite union and finite intersection.
3. All σ-algebras are also closed under countable intersection.

4.1.3 Definition

The trivial σ-algebra is $\{\emptyset, \Omega\}$ and the total σ-algebra is $\mathcal{P}(\Omega)$.

4.2 Probability Measure and Probability Space

4.2.1 Definition: The Axiomatic Definition of Probability

A probability measure on $(\Omega, \mathcal{F})$ is a map $P : \mathcal{F} \longrightarrow \mathbb{R}$ which satisfies three conditions:

i $P(A) \geq 0 \; \forall \, A \in \mathcal{F}$.

ii $P(\Omega) = 1$.

iii For any disjoint set ($i \neq j \implies A_i \cap A_j = \emptyset$) of events $A_1, A_2, A_3, ... \in \mathcal{F}$,

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

4.2.2 Definition

The triplet $(\Omega, \mathcal{F}, P)$ is a probability space.

4.2.3 Theorem

In a probability space $(\Omega, \mathcal{F}, P)$, $\forall \, A, B \in \mathcal{F}$,

1. $P(A^c) = 1 - P(A)$.
2. $A \subseteq B \implies P(A) \leq P(B)$.
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

5 Conditional Probability

5.0.1 Definition

In a probability space $(\Omega, \mathcal{F}, P)$, for events $A, B \in \mathcal{F}$ $(P(B) > 0)$, the conditional probability of A given B is

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

5.0.2 Theorem

In a probability space $(\Omega, \mathcal{F}, P)$, let $A, B \in \mathcal{F}$, let $P(B) > 0$ and let $Q(A) = P(A|B)$. $(\Omega, \mathcal{F}, Q)$ is a probability space.

5.0.3 Lemma

$P(A \cap B) = P(A|B)P(B)$.

5.0.4 Theorem

For any events $A_1, ..., A_n$ where $P(A_2 \cap ... \cap A_n) > 0$,

$$\begin{aligned} P(A_1 \cap ... \cap A_n) =& P(A_1|A_2 \cap ... \cap A_n)P(A_2|A_3 \cap ... \cap A_n)... \\ & P(A_{n-2}|A_{n-1} \cap A_n)P(A_{n-1}|A_n)P(A_n). \end{aligned}$$

5.1 Bayes' Rule and the Law of Total Probability

5.1.1 Theorem: Bayes' Rule

Let $A, B \in \mathcal{F}$ and let $P(A), P(B) > 0$.

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

5.1.2 Definition

A partition of Ω is a set of disjoint events $\{A_i \mid i \in \mathcal{I}\}$ ($i \neq j \in \mathcal{I} \implies A_i \cap A_j = \emptyset$, where $\mathcal{I}$ is a countable index set) such that $\bigcup_{i\in\mathcal{I}} A_i = \Omega$.

5.1.3 Theorem: Law of Total Probability

Let $\{B_i \mid i \in \mathcal{I}\}$ be a partition of Ω where $P(B_i) > 0 \; \forall \, i \in \mathcal{I}$.

$$P(A) = \sum_{i\in\mathcal{I}} P(A \cap B_i) = \sum_{i\in\mathcal{I}} P(A|B_i)P(B_i).$$

5.1.4 Theorem

Let $\{B_i \mid i \in \mathcal{I}\}$ be a partition of Ω where $P(B_i) > 0 \; \forall \, i \in \mathcal{I}$. $\forall \, A \in \mathcal{F}$, $P(A) > 0$,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum\limits_j P(A|B_j)P(B_j)}.$$

6 Independence

6.0.1 Definition

A and B are independent $\iff$

$$P(A \cap B) = P(A)P(B) \; \Big(\iff P(A|B) = P(A) \text{ and } P(B|A) = P(B) \; \Big).$$

6.0.2 Theorem

$$\begin{aligned} A \text{ and } B \text{ are independent } \implies & A \text{ and } B^c \text{ are independent} \\ \wedge \; & A^c \text{ and } B \text{ are independent} \\ \wedge \; & A^c \text{ and } B^c \text{ are independent.} \end{aligned}$$

6.0.3 Definition

A finite set of events $A_1, ..., A_n$ is independent $\iff$

$$P(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap ... \cap A_{i_k}) = P(A_{i_1})P(A_{i_2})P(A_{i_3})...P(A_{i_k})$$

$\forall$ subsets $\{A_{i_1}, ..., A_{i_k}\}$ where $k = 1, ..., n$.
A countable or uncountable set of events is independent $\iff$ every finite subset is independent.

6.1 Continuity of the Probability Measure and The Product Rule

6.1.1 Lemma

Let $A_1, A_2, A_3, ... \in \mathcal{F}$, let $D_i = A_i \setminus \left(\cup_{i=1}^{i-1} A_i\right)$ (and $D_1 = A_1$). $\{D_i\}$ are all disjoint and $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} D_i$ (any countable union can be written as a countable union of disjoint events).

6.1.2 Definition

A sequence of events $(A_i)_{i\geq 1}$ increases [Resp. decreases] to A ($A_i \uparrow A$ [Resp. $A_i \downarrow A$]) $\iff$ $A_1 \subset A_2 \subset A_3 \subset ...$ and $\cup_{i=1}^{\infty} A_i = A$ [Resp. $A_1 \supset A_2 \supset A_3 \supset ...$ and $\cap_{i=1}^{\infty} A_i = A$].

6.1.3 Theorem

Let $A_1, A_2, A_3, ... \in \mathcal{F}$. $A_i \uparrow A \; \vee \; A_i \downarrow A \implies \lim_{i\to\infty} P(A_i) = P(A)$.

6.1.4 Theorem: Product Rule

Let $\{A_1, A_2, A_3, ...\}$ be a countable and independent set of events.

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \prod_{i=1}^{\infty} P(A_i).$$

7 Discrete Random Variables

7.0.1 Definition

A discrete random variable on $(\Omega, \mathcal{F}, P)$ is a map $X : \Omega \longrightarrow \mathbb{R}$ such that

i $\{X(\omega) \mid \omega \in \Omega\} := \operatorname{Im} X$ is a countable subset of $\mathbb{R}$.

ii $\{\omega \in \Omega \mid X(\omega) = x\}\big(:= X^{-1}(x)\big) \in \mathcal{F}\ \forall\, x \in \mathbb{R}$.

Note: the second condition, which states that each set of ω which are mapped to a given x (the preimage of x) is an event (and they are all disjoint sets), is so that probabilities can be assigned to them. The first condition makes X discrete.

7.0.2 Definition

The probability mass function of X is

$$\begin{aligned} p_X : \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto P(\{\omega \in \Omega \mid X(\omega) = x\}). \end{aligned}$$

$p_X(x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$ is often written $P(X = x)$.

7.0.3 Corollary

$$\sum_{x \in \mathbb{R}} p_X(x) = 1.$$

7.0.4 Theorem

Let $S = \{s_i \in \mathbb{R} \mid i \in \mathcal{I}\}$ be countable and distinct and $\{\pi_i \in \mathbb{R} \mid i \in \mathcal{I},\ \pi_i \geq 0\ \forall\, i\}$ be such that $\sum_{i \in \mathcal{I}} \pi_i = 1$. $\exists$ a probability space $(\Omega, \mathcal{F}, P)$ and a discrete random variable X with probability mass function

$$p_X(s_i) = \begin{cases} \pi_i, & i \in \mathcal{I} \\ 0, & s_i \notin S. \end{cases}$$

7.1 Some Discrete Distributions

7.1.1 Definition: Bernoulli Distribution

A discrete random variable X has Bernoulli distribution with parameter $p \in (0,1) \iff \operatorname{Im} X = \{0,1\}$ and

$$p_X(1) = p, \quad p_X(0) = 1 - p \quad (\text{and } x \notin \operatorname{Im} X \implies p_X(x) = 0)$$

$\iff : \boxed{X \sim \text{Bern}(p)}$.

7.1.2 Definition

Let $A \in \mathcal{F}$. The indicator variable of A is

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A. \end{cases}$$

Note: $I_A \sim \text{Bern}(P(A))$.

7.1.3 Definition: Binomial Distribution

A discrete random variable X has binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1) \iff \operatorname{Im} X = \{0, 1, ..., n\}$ and

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (\text{and } x \notin \operatorname{Im} X \implies p_X(x) = 0)$$

$\iff : \boxed{X \sim \text{Bin}(n, p)}$.

(X represents the number of successes in a sequence of n identical Bernoulli trials with success parameter p, or the number of successes observed in n draws with replacement from a population where a proportion p are to be successful).

7.1.4 Definition: Hypergeometric Distribution

A discrete random variable X has hypergeometric distribution with parameters $N \in \mathbb{N}$, $K \in \{0, 1, ..., N\}$ and $n \in \{0, 1, ..., N\} \iff \operatorname{Im} X = \{0, 1, ..., \min\{n, K\}\}$ and

$$p_X(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad (\text{and } x \notin \operatorname{Im} X \implies p_X(x) = 0)$$

$\iff : \boxed{X \sim \text{HGeom}(N, K, n)}$.

(X represents the number of successes observed in n draws without replacement from a population of size N where K are to be successful).

7.1.5 Definition: Discrete Uniform Distribution

Let $C \neq \emptyset$ be a finite set of numbers.
A discrete random variable X follows the discrete uniform distribution on $C \iff \operatorname{Im} X = C$ and

$$p_X(x) = \frac{1}{|C|} \quad (\text{and } x \notin \operatorname{Im} X \implies p_X(x) = 0)$$

$\iff : \boxed{X \sim \text{DU}(C)}$.

7.1.6 Definition: Poisson Distribution

A discrete random variable X has Poisson distribution with parameter $\lambda > 0 \iff \operatorname{Im} X = \mathbb{N}$ and

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad (\text{and } x \notin \operatorname{Im} X \implies p_X(x) = 0)$$

$\iff : \boxed{X \sim \text{Poi}(\lambda)}$.

(X represents the number of successes (random occurrences) in a unit of time where λ is the rate or the expected successes per unit of time).

7.1.7 Definition: Geometric Distribution

A discrete random variable X has geometric distribution with parameter $p \in (0,1) \iff \operatorname{Im} X = \mathbb{N} \setminus \{0\}$ and

$$p_X(x) = (1-p)^{x-1} p \quad (\text{and } x \notin \operatorname{Im} X \implies p_X(x) = 0)$$

$\iff : \boxed{X \sim \text{Geom}(p)}$.

(X represents the number of successive identical Bernoulli trials to obtain the first success. Note that if X is to represent the number of failures before the first success, then $p_X(x) = (1-p)^x p$ and the image is $\mathbb{N}$, including 0).

7.1.8 Definition: Negative Binomial Distribution

A discrete random variable X has negative binomial distribution with parameters $r \in \mathbb{N}$ and $p \in (0,1) \iff \operatorname{Im} X = \mathbb{N}$ and

$$p_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x \quad (\text{and } x \notin \operatorname{Im} X \implies p_X(x) = 0)$$

$\iff : \boxed{X \sim \text{NBin}(r,p)}$.

(X represents the number of failures observed in a sequence of identical Bernoulli trials with success parameter p before r successes have occurred).

8 Continuous Random Variables

8.0.1 Definition: Random Variable (General)

A random variable on $(\Omega, \mathcal{F}, P)$ is a map $X : \Omega \longrightarrow \mathbb{R}$ such that

$$\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F} \; \forall \, x \in \mathbb{R}.$$

Note that this condition is similar to the second condition in 7.0.1 and that the definition of a discrete random variable also satisfies this definition, since $\mathcal{F}$ is closed under countable union

$$\implies \bigcup_{x \in \operatorname{Im} X, \; x \leq x^*} \{\omega \mid X(\omega) = x\} \in \mathcal{F} \iff \{\omega \in \Omega \mid X(\omega) \leq x^*\} \in \mathcal{F}.$$

8.0.2 Definition

The cumulative distribution function of X is

$$\begin{aligned} F_X : \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto P(\{\omega \in \Omega \mid X(\omega) \leq x\}). \end{aligned}$$

$F_X(x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\})$ is often written $P(X \leq x)$.

8.0.3 Theorem

1. F_X is monotonically non-decreasing.
2. F_X is right-continuous.
3. $\forall$ random variables X, $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

8.0.4 Theorem

$P(a < X \leq b) = F_X(b) - F_X(a)$.

8.0.5 Definition

A random variable X is continuous if, for a function $f_X : \mathbb{R} \longrightarrow \mathbb{R}$ (probability density function) for which $f_X(x) \geq 0 \; \forall \, x \in \mathbb{R}$ and $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$,

$$F_X(x) = \int_{-\infty}^{x} f_X(u) \, du \quad \forall \, x \in \mathbb{R}.$$

8.0.6 Theorem

For a continuous random variable X with p.d.f f_X,

1. $P(X = x) = 0 \; \forall \, x \in \mathbb{R}$.
2. $P(a \leq x \leq b) = \int_a^b f_X(u) \, dx$.

8.0.7 Summary Table: Properties of Discrete and Continuous Random Variables

Let $p_X : \mathbb{R} \longrightarrow [0,1]$ be the p.m.f of a discrete X and $f_X : \mathbb{R} \longrightarrow \mathbb{R}$ be the p.d.f of a continuous X.

Discrete random variable	Continuous random variable
Both satisfy the general definition in 8.0.1	
$p_X(x) \geq 0 \ \forall \, x \in \mathbb{R}$	$f_X(x) \geq 0 \ \forall \, x \in \mathbb{R}$
$\sum_{x \in \operatorname{Im} X} p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x) \ dx = 1$
$F_X(x) = \sum_{u \in \operatorname{Im} X, \ u \leq x} p_X(u)$	$F_X(x) = \int_{-\infty}^{x} f_X(u) \ du$

8.1 Some Continuous Distributions

8.1.1 Definition: Uniform Distribution

A continuous random variable X has uniform distribution on the interval $(a,b) \iff$

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

$$\left(\text{and} \quad F_X(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \geq b \end{cases} \right)$$

$\iff : \boxed{X \sim \mathrm{U}(a,b)}$.

8.1.2 Definition: Exponential Distribution

A continuous random variable X has exponential distribution with parameter $\lambda > 0 \iff$

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

$$\left(\text{and} \quad F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0 \end{cases} \right)$$

$\iff : \boxed{X \sim \mathrm{Exp}(\lambda)}$.

8.1.3 Definition: Gamma Function

$$\Gamma(t) := \int_0^{\infty} x^{t-1} e^{-x} \ dx \quad (t > 0).$$

Note: for $\forall \, t > 1, \ \Gamma(t) = (t-1)\Gamma(t-1)$ and $\forall \, t \in \mathbb{N}, \ \Gamma(t) = (t-1)!$.

8.1.4 Definition: Gamma Distribution

A continuous random variable X has gamma distribution with parameters $\alpha > 0$ (shape) and $\beta > 0$ (rate) $\iff$

$$f_X(x) = \begin{cases} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$\iff : \boxed{X \sim \mathrm{Gamma}(\alpha, \beta)}$.

In the special case $\alpha = n \in \mathbb{N} \setminus \{0\}$, the Erlang distribution is given by

$$f_X(x) = \begin{cases} \frac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

There is no closed form c.d.f for the gamma distribution.

8.1.5 Definition: χ-Squared Distribution

A continuous random variable X has χ-squared distribution with $\nu \in \mathbb{N}$ degrees of freedom $\iff$

$$f_X(x) = \begin{cases} \frac{1}{2\Gamma(\frac{\nu}{2})} \left(\frac{x}{2}\right)^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$\iff : \boxed{X \sim \chi^2(\nu)}$ (also written $X \sim \chi^2_\nu$).

Note that $X \sim \chi^2(\nu) \iff X \sim \text{Gamma}(\frac{\nu}{2}, \frac{1}{2})$.
There is no closed form c.d.f for the χ-squared distribution.

8.1.6 Definition: F-Distribution

A continuous random variable X has F-distribution with $d_1, d_2 > 0$ degrees of freedom $\iff$

$$f_X(x) = \begin{cases} \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)\left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\left(1+\frac{d_1}{d_2}x\right)^{\frac{d_1+d_2}{2}}}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$\iff : \boxed{X \sim \text{F}(d_1, d_2)}$.

Note that $X_n \sim \chi^2_n \wedge X_m \sim \chi^2_m \iff \frac{X_n/n}{X_m/m} \sim \text{F}(n, m)$.
There is no closed form c.d.f for the F-distribution.

8.1.7 Definition: Beta Function

$$B(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx \quad (\alpha, \beta > 0).$$

8.1.8 Definition: Beta Distribution

A continuous random variable X has beta distribution with parameters $\alpha, \beta > 0 \iff$

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

$\iff : \boxed{X \sim \text{Beta}(\alpha, \beta)}$.

There is no closed form c.d.f for the beta distribution.

8.1.9 Definition: Normal Distribution

A continuous random variable X has normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0 \iff$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in \mathbb{R}$$

$\iff : \boxed{X \sim \text{N}(\mu, \sigma^2)}$.

In the special case $\mu = 0,\ \sigma = 1$, the probability density function for the standard normal distribution is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{for } x \in \mathbb{R}$$

and the cumulative distribution function by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt \quad \text{for } x \in \mathbb{R}.$$

There is no simpler closed form of the c.d.f for the standard normal distribution beyond this integral.

8.1.10 Corollary

1. $\phi(x) = \phi(-x) \ \forall \, x.$
2. $\Phi(x) = 1 - \Phi(x) \ \forall \, x.$

8.1.11 Definition: Standard Cauchy Distribution

A continuous random variable X has standard Cauchy distribution $\iff$

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } x \in \mathbb{R}$$

$$\left(\text{and} \quad F_X(x) = \frac{1}{\pi} \arctan x + \frac{1}{2} \quad \text{for } x \in \mathbb{R} \right).$$

$\iff : \boxed{X \sim \text{Cauchy}(1,0)}$.

Note that $X \sim \text{N}(0,1) \wedge Y \sim \text{N}(0,1) \iff \frac{X}{Y} \sim \text{Cauchy}(1,0)$.

8.1.12 Definition: Student's t-Distribution

A continuous random variable X has (student's) t-distribution with $\nu > 0$ degrees of freedom $\iff$

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } x \in \mathbb{R}.$$

$\iff : \boxed{X \sim t_\nu}$.

There is no closed form c.d.f for the student's t-distribution.

9 Transformations of Random Variables

9.1 Discrete Case

9.1.1 Theorem

Let X be a discrete random variable and Y be such that $Y(\omega) = g(X(\omega))$, then
$\{\omega \in \Omega \mid Y(\omega) = y\} \in \mathcal{F} \ \forall \, y \in \mathbb{R}$ and

$$p_Y(y) = \sum_{x \in \text{Im } X, \ g(x) = y} p_X(x) \quad \forall \, y \in \text{Im } Y.$$

9.1.2 Theorem

Let X be a discrete random variable and Y be such that $Y(\omega) = g(X(\omega))$ as above. If additionally g is a bijection, then $\exists$ a unique x for which $y = g(x)$, and so $p_Y(y) = p_X(g^{-1}(y)) \ \forall \, y \in \text{Im } Y$.

9.2 Continuous Case

9.2.1 Theorem

Let X be a continuous random variable and Y be such that $Y(\omega) = g(X(\omega))$, then

$$F_Y(y) = \int_{x \in \text{Im } X, \ g(x) \leq y} f_X(x) \ dx.$$

The integral in this result is analogous to the sum in 9.1.1 and is a necessary definition for cases where g may not be strictly increasing or decreasing (analgous to the fact that g in 9.1.1 is not necessarily bijective), but it is not particularly useful on its own. Two special bijective cases are given on the next page in which, by differentiating the cumulative distribution function, a probability denisty function can be obtained.

9.2.2 Theorem

Let X be a continuous random variable and Y be such that $Y(\omega) = g(X(\omega))$ as above, but where additionally g is strictly increasing, differentiable and has inverse g^{-1}. Let $x = g^{-1}(y)$, then

$$F_Y(y) = F_X(g^{-1}(y)) = F_X(x)$$

$$\text{and so} \quad f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}[g^{-1}(y)] = f_X(x)\frac{dx}{dy} \quad \forall\, y \in \mathbb{R}.$$

In the case of strictly increasing g therefore, this second equation can be written $f_Y(y)dy = f_X(x)dx$.

9.2.3 Theorem

If instead g is strictly decreasing (and differentiable), then

$$F_Y(y) = 1 - F_X(g^{-1}(y)) = 1 - F_X(x)$$

$$\text{and so} \quad f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dy}[g^{-1}(y)] = f_X(x)\left|\frac{dx}{dy}\right| \quad \forall\, y \in \mathbb{R},$$

since the derivative of the inverse of g will be negative; so note that by the absolute value, this theorem in fact holds when g is either strictly increasing or decreasing. $\left|\frac{dx}{dy}\right|$ is called the Jacobian of the transformation g^{-1}.

10 Expectation of Random Variables

10.0.1 Definition

The expectation (or mean) of a discrete random variable X with p.m.f p_X is

$$\mathrm{E}(X) = \sum_{x \in \mathrm{Im}\, X} x p_X(x)$$

(provided $\sum_{x\,\mathrm{Im}\,X} |x| p_X(x)$ is convergent).

10.0.2 Definition

The expectation (or mean) of a continuous random variable X with p.d.f f_X is

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\ dx$$

(provided $\int_{-\infty}^{\infty} |x| f_X(x)\ dx$ is convergent).

10.1 Law of the Unconscious Statistician

10.1.1 Theorem

Let X be a discrete random variable with p.m.f p_X and let $g : \mathbb{R} \longrightarrow \mathbb{R}$.

$$\mathrm{E}(g(X)) = \sum_{x \in \mathrm{Im}\, X} g(x) p_X(x)$$

(provided $\sum_{x\,\mathrm{Im}\,X} |g(x)| p_X(x)$ is convergent).

10.1.2 Theorem

Let X be a continuous random variable with p.d.f f_X and let $g : \mathbb{R} \longrightarrow \mathbb{R}$.

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\ dx$$

(provided $\int_{-\infty}^{\infty} |g(x)| f_X(x)\ dx$ is convergent).

10.1.3 Definition

Let $k \in \mathbb{N}$ and $g(x) = x^k$. $\mathrm{E}(g(X))$ is the k^{th} moment of the random variable X.

10.1.4 Theorem

1. If $X(\omega) \geq 0 \; \forall \omega, \; \mathrm{E}(X) \geq 0$.
2. $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b \; \forall a, b \in \mathbb{R}$.

Note that this result is not sufficient for linearity - see theorem 11.6.3.

10.2 Variance

10.2.1 Definition

Let X be a discrete or continuous random variable. The variance of X is

$$\mathrm{Var}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right).$$

10.2.2 Theorem

$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$.

10.2.3 Theorem

$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X) \; \forall a, b \in \mathbb{R}$.

11 Multivariate Random Variables

11.1 Multivariate Distributions and Independence

11.1.1 Definition

For two arbitrary random variables X and Y on the same probability space, the joint distribution function of the random vector (X, Y) is

$$\begin{aligned} F_{X,Y} : \mathbb{R}^2 &\longrightarrow [0, 1] \\ (x, y) &\longmapsto P(\{\omega \in \Omega \mid X(\omega) \leq x \; \wedge \; Y(\omega) \leq y\}). \end{aligned}$$

$F_{X,Y}(x, y) = P(\{\omega \in \Omega \mid X(\omega) \leq x \wedge Y(\omega) \leq y\})$ is often written $P(X \leq x, Y \leq y)$.

11.1.2 Theorem

1. $F_{X,Y}$ is monotonically non-decreasing:

$$x_1 < x_2 \; \wedge \; y_1 < y_2 \implies F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$$

2. $\forall$ random variables X, Y,

$$\lim_{x \to -\infty, y \to -\infty} F_{X,Y}(x, y) = 0 \quad \text{and} \quad \lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1.$$

Note that the limits above determine the marginal distributions uniquely:

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y) \quad \text{and} \quad F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x, y).$$

11.1.3 Definition

The random variables X and Y are independent $\iff$

$$\begin{aligned} &\text{the events } \{\omega \mid X(\omega) \leq x\} \text{ and } \{\omega \mid Y(\omega) \leq y\} \text{ are independent} \quad \forall x, y \in \mathbb{R} \\ \iff &P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \quad \forall x, y \in \mathbb{R} \\ \iff &F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \forall x, y \in \mathbb{R}. \end{aligned}$$

11.2 Extension to n Dimensions

11.2.1 Definition

For arbitrary random variables $X_1, ..., X_n$ on the same probability space, the joint distribution function of the random vector $\mathbf{X} := (X_1, ..., X_n)$ is

$$\begin{aligned} F_{\mathbf{X}} : \mathbb{R}^n &\longrightarrow [0,1] \\ (\mathbf{x}) &\longmapsto P(\{\omega \in \Omega \mid X_1(\omega) \leq x_1 \ \wedge \ ... \ \wedge \ X_n(\omega) \leq x_n\}) \\ &= P(X_1 \leq x_1, ..., X_n \leq x_n) \end{aligned}$$

(where $(\mathbf{x}) = (x_1, ..., x_n)$) and $X_1, ..., X_n$ are (group) independent $\iff$

$$\begin{aligned} & P(X_1 \leq x_1, ..., X_n \leq x_n) = P(X_1 \leq x_1)...P(X_n \leq x_n) \quad \forall \mathbf{x} \in \mathbb{R}^n \\ \iff & F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1)...F_{X_n}(x_n) \quad \forall \mathbf{x} \in \mathbb{R}^n. \end{aligned}$$

11.2.2 Definition

The random variables $X_1, ..., X_n$ are pairwise independent $\iff$

$$\begin{aligned} & P(X_i \leq x_i, X_j \leq x_j) = P(X_i \leq x_i)P(X_j \leq x_j) \quad \forall x_i, x_j \in \mathbb{R} \ \forall i \neq j \\ \iff & F_{X_i, X_j}(x_i, x_j) = F_{X_i}(x_i)F_{X_j}(x_j) \quad \forall x_i, x_j \in \mathbb{R} \ \forall i \neq j. \end{aligned}$$

11.2.3 Definition

An infinite family of random variables $\{X_i \mid i \in \mathcal{I}\}$ is independent $\iff$ all finite subsets are group independent $\iff$

$$P\left(\bigwedge_{j \in \mathcal{J}} X_j \leq x_j\right) = \prod_{j \in \mathcal{J}} P(X_j \leq x_j) \quad \forall x_j \in \mathbb{R}, \quad \forall \text{ finite } \mathcal{J} \subset \mathcal{I}.$$

11.3 Multivariate Discrete Distributions and Independence

11.3.1 Definition

For two discrete random variables X and Y on the same probability space, the joint probability mass function of the random vector (X, Y) is

$$\begin{aligned} p_{X,Y} : \mathbb{R}^2 &\longrightarrow [0,1] \\ (x, y) &\longmapsto P(\{\omega \in \Omega \mid X(\omega) = x \ \wedge \ Y(\omega) = y\}). \end{aligned}$$

$p_{X,Y}(x, y) = P(\{\omega \in \Omega \mid X(\omega) = x \ \wedge \ Y(\omega) = y\})$ is often written $P(X = x, Y = y)$.

11.3.2 Corollary

$$\sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} p_{X,Y}(x, y) = 1.$$

Note that the marginal probability mass functions are then given by

$$p_X(x) = \sum_{y \in \mathbb{R}} p_{X,Y}(x, y) \quad \text{and} \quad p_Y(y) = \sum_{x \in \mathbb{R}} p_{X,Y}(x, y).$$

Generally, $P((X, Y) \in A \subseteq \mathbb{R}^2) = \sum \sum_{(x,y) \in A} p_{X,Y}(x, y)$.

11.3.3 Definition

The discrete random variables X and Y are independent $\iff$

$$\begin{aligned} & \text{the events } \{\omega \mid X(\omega) = x\} \text{ and } \{\omega \mid Y(\omega) = y\} \text{ are independent} \quad \forall x, y \in \mathbb{R} \\ \iff & P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y \in \mathbb{R} \\ \iff & p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \forall x, y \in \mathbb{R}. \end{aligned}$$

11.4 Multivariate Continuous Distributions and Independence

11.4.1 Definition

The random vector (X, Y) is jointly continuous if, for a function $f_{X,Y} : \mathbb{R}^2 \longrightarrow \mathbb{R}$ (joint probability density function) for which $f_{X,Y}(x, y) \geq 0 \; \forall \, x, y \in \mathbb{R}$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \; dx \; dy = 1$,

$$F_{X,Y}(x, y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u, v) \; dv \; du \quad \forall \, x, y \in \mathbb{R}.$$

Note that the marginal distributions are then given, as in 11.1.2, by

$$F_X(x) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f_{X,Y}(u, v) \; dv \; du \quad \text{and} \quad F_Y(y) = \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{y} f_{X,Y}(u, v) \; dv \; du$$

and, differentiating, the marginal probability density functions by

$$f_X(x) = \int_{\infty}^{\infty} f_{X,Y}(x, v) \; dv \quad \text{and} \quad f_Y(y) = \int_{\infty}^{\infty} f_{X,Y}(u, y) \; du.$$

Generally, $P((X, Y) \in A \subseteq \mathbb{R}^2) = \int \int_{(x,y)\in A} f_{X,Y}(x, y) \; dx \; dy$.

11.4.2 Note

As with the univariate case, the joint density can be obtained from the joint distribution.

$$f_{X,Y}(x, y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y), & \text{where the derivative exists} \\ 0, & \text{otherwise.} \end{cases}$$

11.4.3 Definition

By differentiating in 11.1.3, the jointly continuous random variables X and Y are independent $\iff$

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

11.5 Transformations of Random Variables - Bivariate Case

Given the jointly continuous random vector (X, Y) with j.d.f $f_{X,Y}$, let $u, v : \mathbb{R}^2 \longrightarrow \mathbb{R}$ and let the random variable $U = u(X, Y)$ and $V = v(X, Y)$ while

$$\begin{aligned} T &: \mathbb{R}^2 \longrightarrow \mathbb{R}^2 \\ (x, y) &\longmapsto (u(x, y), v(x, y)) \end{aligned}$$

is assumed to be a bijection from $D = \{(x, y) \mid f_{X,Y}(x, y) > 0\} \subseteq \mathbb{R}^2 \longrightarrow S \subseteq \mathbb{R}^2$. Let the inverse of T be denoted by $T^{-1} : S \longrightarrow D$ where $T^{-1}(u, v) = (x(u, v), y(u, v))$. While the Jacobian of g^{-1} in 9.2.3 was $\left|\frac{dx}{dy}\right|$, the Jacobian of T^{-1} in this case is

$$|J(u, v)| = \left|\frac{\partial(x, y)}{\partial(u, v)}\right| := \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$$

and the j.d.f of (U, V) is given by

$$f_{U,V}(u, v) = f_{X,Y}(T^{-1}(u, v))|J(u, v)| = f_{X,Y}(x(u, v), y(u, v))|J(u, v)| \quad \forall \, (u, v) \in S$$

$$(\text{and } (u, v) \notin S \implies f_{U,V}(u, v) = 0).$$

11.6 Law of the Unconscious Statistician - 2 dimensions

11.6.1 Theorem

Let X, Y be discrete random variables with j.p.m.f $p_{X,Y}$ and let $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$. $g(X, Y)$ is also a discrete random variable and

$$\mathrm{E}(g(X, Y)) = \sum_{x \in \operatorname{Im} X} \sum_{y \in \operatorname{Im} Y} g(x, y) p_{X,Y}(x, y).$$

11.6.2 Theorem

Let X, Y be jointly continuous random variables with j.p.d.f $f_{X,Y}$ and let $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$.

$$\mathrm{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\; dx\; dy.$$

11.6.3 Theorem: Linearity of the Expectation

Using these new results, it can be proved that

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y) \quad \forall\, a, b \in \mathbb{R}$$

and, more generally, that

$$\mathrm{E}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i\,\mathrm{E}(X_i) \quad \forall\, \{c_i\} \subset \mathbb{R}.$$

11.7 Covariance

11.7.1 Definition

The covariance of two random variables X and Y is

$$\mathrm{Cov}(X,Y) = \mathrm{E}\Big((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\Big).$$

Note that $\mathrm{Cov}(X,X) = \mathrm{E}\Big((X - \mathrm{E}(X))^2\Big) = \mathrm{Var}(X)$.

11.7.2 Theorem

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

11.7.3 Theorem

For two discrete or jointly continuous random variables X and Y,

$$X \text{ and } Y \text{ are independent } \implies \mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y) \iff \mathrm{Cov}(X,Y) = 0$$

and, more generally,

$$X_1, ..., X_n \text{ are independent } \implies \mathrm{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathrm{E}(X_i).$$

11.7.4 Theorem

The equation in 11.7.3 is not sufficient to imply independence. In fact,

$$X \text{ and } Y \text{ are independent } \iff \mathrm{E}(g(X)h(Y)) = \mathrm{E}(g(X))\,\mathrm{E}(h(Y)) \;\forall\, g, h : \mathbb{R} \longrightarrow \mathbb{R}.$$

11.7.5 Theorem: Variance of the Sum

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

12 Generating Functions

12.0.1 Definition

Let X be a non-negative integer valued discrete random variable. Let

$$\mathcal{S}_X = \left\{ s \in \mathbb{R} \mid \sum_{x=0}^{\infty} |s|^x p_X(x) \text{ is finite} \right\}.$$

The probability generating function of X is $G_X : \mathcal{S}_X \longrightarrow \mathbb{R}$ such that

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x p_X(x).$$

Note that $G_X(0) = p_X(0)$ and $G_X(1) = 1$.

12.0.2 Theorem

The probability generating function of a discrete random variable uniquely determines its probability mass function.

12.0.3 Theorem

Let X and Y be non-negative integer valued discrete random variables.

$$G_X(s) = G_Y(s) \ \forall\, s \in \mathcal{S}_X \cap \mathcal{S}_Y \iff p_X(x) = p_Y(x) \ \forall\, x \in \operatorname{Im} X.$$

12.0.4 Theorem

Let X and Y be non-negative integer valued discrete random variables. Let X and Y be independent.

$$G_{X+Y}(s) = G_X(s)G_Y(s) \quad \forall\, s \in \mathcal{S}_X \cap \mathcal{S}_Y.$$

12.0.5 Corollary

Let $X_1, ..., X_n$ be non-negative integer valued discrete random variables. Let $X_1, ..., X_n$ be independent.

$$G_{\sum_{i=1}^n X_i}(s) = \prod_{i=1}^n G_{X_i}(s) \quad \forall\, s \in \bigcap_{i=1}^n \mathcal{S}_{X_i}.$$

12.0.6 Theorem

Let X be a non-negative integer valued discrete random variable.

$$G_X^{(k)}(1) = \mathrm{E}\Big(X(X-1)...(X-k+1)\Big).$$

12.1 Moments

12.1.1 Definition

The moment generating function of a random variable X is

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

12.1.2 Theorem

The k^{th} moment of X can be found by $\mathrm{E}(X^k) = M_X^{(k)}(0)$.

12.1.3 Theorem

$M_{aX+b}(t) = e^{bt}M_X(at) \ \forall\, a, b \in \mathbb{R}$.

12.1.4 Theorem

Let $X_1, ..., X_n$ be independent random variables.

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t).$$

12.1.5 Theorem: Characterisation Theorem

$$M_X(t) = M_Y(t) \text{ in a neighbourhood of } 0 \implies F_X(u) = F_Y(u) \quad \forall\, u$$

(the moment generating function characterises the distribution of a random variable uniquely).

12.1.6 Note: Characteristic Functions and the Laplace Transform

While the moment generating function does not exist for all distributions, the characteristic function

$$\phi_X(t) = \mathrm{E}(e^{itX})$$

does (see 21.0.2). Another useful function is the Laplace transform:

$$\mathcal{L}_X(t) = M_X(-t) = \mathrm{E}(e^{-tX}).$$

13 Conditional Distribution and Conditional Expectation

13.0.1 Definition

Let X be a discrete random variable on $(\Omega, \mathcal{F}, P)$ and let $B \in \mathcal{F}$. The conditional distribution of X given B is given by

$$P(X = x|B) = \frac{P(\{X = x\} \cap B)}{P(B)}$$

and the conditional expectation of X given B is

$$\mathrm{E}(X|B) = \sum_{x \in \mathrm{Im}\, X} xP(X = x|B)$$

(provided $\sum_{x\, \mathrm{Im}\, X} |x|P(X = x|B)$ is convergent).

13.1 Conditioning on a Discrete Random Variable

13.1.1 Definition

Consider the definitions above in the case where $B = \{Y = y\}$. For two discrete random variables X and Y, the conditional probability mass function of X given $Y = y$ (for which $p_Y(y) > 0$) is given by

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

$p_{X|Y}(x, y)$ is often written $P(X = x|Y = y)$.
The conditional expectation of X given $Y = y$ is

$$\mathrm{E}(X|Y = y) = \sum_{x \in \mathrm{Im}\, X} xp_{X|Y}(x|y) = \sum_{x \in \mathrm{Im}\, X} x\frac{p_{X,Y}(x, y)}{p_Y(y)}$$

(provided $\sum_{x \in \mathrm{Im}\, X} |x|p_{X|Y}(x|y)$ is convergent).

13.1.2 Theorem: Conditional Law of the Unconscious Statistician

Let $g : \mathbb{R} \longrightarrow \mathbb{R}$.

$$\mathrm{E}(g(X)|Y = y) = \sum_{x \in \mathrm{Im}\, X} g(x)p_{X|Y}(x|y) = \sum_{x \in \mathrm{Im}\, X} g(x)\frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

13.1.3 Theorem: Law of Total Expectation (Discrete Random Variables)

Let $\{B_i \mid i \in \mathcal{I}\}$ be a partition of Ω where $P(B_i) > 0\ \forall\, i \in \mathcal{I}$. Let X be a discrete random variable.

$$\mathrm{E}(X) = \sum_{i \in \mathcal{I}} \mathrm{E}(X|B_i)P(B_i)$$

(provided $\sum_{i \in \mathcal{I}} |\,\mathrm{E}(X|B_i)|P(B_i)$ is convergent). Note that the set of events $\{Y = y\}$ ranging over all $y \in \mathrm{Im}\, Y$ is an example of a partition of Ω, so a form of the theorem is $\mathrm{E}(X) = \sum_{y,\ p_Y(y)>0} \mathrm{E}(X|Y = y)p_Y(y)$.

13.2 Conditioning on a Continuous Random Variable

13.2.1 Definition

For two jointly continuous random variables X and Y, the conditional probability density function of X given $Y = y$ (for which $f_Y(y) > 0$) is given by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

and the conditional distribution function of X given $Y = y$ by

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} \frac{f_{X,Y}(u, y)\ du}{f_Y(y)}.$$

$F_{X|Y}(x|y)$ is often written $P(X \leq x|Y = y)$.
The conditional expectation of X given $Y = y$ is

$$\mathrm{E}(X|Y = y) = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)\ dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x, y)}{f_Y(y)}\ dx$$

(provided $\int_{-\infty}^{\infty} |x|f_{X|Y}(x|y)\ dx$ is convergent).

13.2.2 Theorem: Conditional Law of the Unconscious Statistician

Let $g : \mathbb{R} \longrightarrow \mathbb{R}$.

$$\mathrm{E}(g(X)|Y = y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\ dx = \int_{-\infty}^{\infty} g(x) \frac{f_{X,Y}(x,y)}{f_Y(y)}\ dx.$$

13.2.3 Theorem: Law of Total Expectation (Jointly Continuous Random Vector)

For two jointly continuous random variables X and Y,

$$\mathrm{E}(X) = \int_{y,\ f_Y(y)>0} \mathrm{E}(X|Y = y) f_Y(y)\ dx$$

(provided $\int_{-\infty}^{\infty} |\,\mathrm{E}(X|Y = y)| f_Y(y)\ dx$ is convergent). This theorem is analogous to 13.1.3 since again, the set of events $\{Y = y\}$ ranging over all $y \in \operatorname{Im} Y$ is effectively a partition of Ω.

13.3 General Theorems of Conditional Expectation

In these results, X and Y may be continuous or discrete.

13.3.1 Theorem

$\mathrm{E}\Big(\,\mathrm{E}(X|Y)\Big) = \mathrm{E}(X)$.

13.3.2 Theorem

$\mathrm{E}\Big(g(X)h(Y)|Y\Big) = h(Y)\,\mathrm{E}\Big(g(X)|Y\Big)$.

14 Central Tendency and Dispersion

14.1 Mean, Variance and Higher Order Moments

14.1.1 Theorem

$\mathrm{E}\Big((X-a)^2\Big) \geq \mathrm{E}\Big((X-\mathrm{E}(X))^2\Big)\ \forall\, a \in \mathbb{R}$.

14.1.2 Theorem

$\mathrm{E}\Big((X-g(Y))^2\Big) \geq \mathrm{E}\Big((X-\mathrm{E}(X|Y))^2\Big)\ \forall\, g : \mathbb{R} \longrightarrow \mathbb{R}$.

14.1.3 Theorem

$\mathrm{E}\Big(\,\mathrm{E}(X|Y)\Big) = \mathrm{E}(X)$.

14.1.4 Definition

$\mu'_k := \mathrm{E}(X^k)$ is the k^{th} (raw) moment of the random variable X (Definition 10.1.3).
$\mu_k := \mathrm{E}\Big((X-\mu)^k\Big)$ is the k^{th} central moment of the random variable X (where $\mu = \mu'_1 = \mathrm{E}(X)$).

14.1.5 Corollary

Let the mean and variance of X be μ and σ^2 respectively. $\mu'_2 = \mu^2 + \sigma^2$.

14.2 Sample Mean and Variance

14.2.1 Definition

The sample mean of the random variables $X_1, ..., X_n$ is their arithmetic mean:

$$\overline{X} = \frac{X_1 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Their sample variance is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

14.2.2 Theorem

Let $X_1, ..., X_n$ be sampled independently from a distribution with mean and variance μ and σ^2 respectively.

1. $\mathrm{E}(\overline{X}) = \mu$.
2. $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.
3. $\mathrm{E}(S^2) = \sigma^2$.

(4. If sampling from a normal distribution, $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$ - see 14.6.4.)

14.3 The Markov and Chebyshev Inequalities

14.3.1 Theorem: Markov's Inequality

Let X be a random variable taking only non-negative values.

$$P(X \geq a) \leq \frac{\mathrm{E}(X)}{a} \quad \forall\, a > 0.$$

14.3.2 Theorem: Chebyshev's Inequality

Let the mean and variance of X be μ and σ^2 respectively.

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} \quad \forall\, c > 0.$$

14.3.3 Lemma

Let X take non-zero values only in the closed interval $[a, b]$. $\mathrm{Var}(X) \leq \frac{(b-a)^2}{4}$.

14.3.4 Theorem (Application)

Suppose a sample of size n is taken from a distribution $X \sim \mathrm{Bern}(p)$ for an unknown probability p, and the sample proportion recorded as $\widehat{p}$. We can say that $\widehat{p}$ differs from p by at most ε with a confidence of at least $1 - \frac{1}{4n\varepsilon^2}$ (see section 14.5).

14.4 Other Measures

14.4.1 Definition

The mode of a random variable X with probability density function f_X is

$$\mathrm{mode}(X) = \arg\max_x f_X(x) \quad \Big(\text{or} \quad \arg\max_x p_X(x)\Big).$$

14.4.2 Definition

A median of a random variable X is m such that

$$P(X \leq m) \geq \tfrac{1}{2} \quad \text{and} \quad P(X \geq m) \geq \tfrac{1}{2}.$$

m may be written median(X). The median is not unique in this definition - in cases where any value in the interval $[a, b]$ satisfies the definition, the median of X may be taken as $\frac{a+b}{2}$.

14.4.3 Theorem

$\mathrm{E}\Big(|X - a|\Big) \geq \mathrm{E}\Big(|X - \mathrm{median}(X)|\Big)\ \forall\, a \in \mathbb{R}$.

14.4.4 Definition

The sample median m of an ordered sample of observations $x_1, ..., x_n$ is

$$m = \begin{cases} x_{((n+1)/2)}, & n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{((n/2)+1)}), & n \text{ is even.} \end{cases}$$

Note: when n is even, any value in the open interval $(x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)})$ is a median.

14.4.5 Definition

Given a random variable X with cumulative distribution F_X, $F_X^{-1} : [0,1] \longrightarrow \mathbb{R}$, provided it exists, is called the quantile function.

14.4.6 Definition

The existence of a quantile function for the random variable X implies the uniqueness of the following statistics.
The median of X is $m = F_X^{-1}(0.5)$.
The lower quartile of X is $q_{0.25} = F_X^{-1}(0.25)$.
The upper quartile of X is $q_{0.75} = F_X^{-1}(0.75)$.
The interquartile range of X is $\text{IQR} = q_{0.75} - q_{0.25}$.

14.5 Parameter Estimation

A parameter is a characteristic that determines the distribution of a random variable or joint distribution of random variables. A statistic is a function of a sample of observed random variables. The frequentist inference of statistics draws conclusions about a parameter unknown to the statistician from observations of a statistic:

14.5.1 Definition

A point estimator is a function $\widehat{\Theta}(X_1, ..., X_n)$ on a sample of random variables $X_1, ..., X_n$.

14.5.2 Note

Since a point estimator is a function of random variables, it is also a random variable. Any statistic is a point estimator, for example note that the sample mean has its own expectation and variance. As in section 11.2, provided there is no ambiguity, I will refer to a vector of random variables by $\mathbf{X}$ and a vector of observations by $\mathbf{x}$. A realisation of the point estimator $\widehat{\Theta}$ will be denoted by $\widehat{\theta}$, representing an estimate of the parameter θ.

14.5.3 Definition

An interval estimate (of a parameter θ) is a pair of functions L, U on a sample which satisfy $L(\mathbf{x}) \leq U(\mathbf{x})\ \forall\, \mathbf{x}$. The random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is an interval estimator.

14.5.4 Definition

The coverage probability of $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that θ lies in the random interval $[L(\mathbf{X}), U(\mathbf{X})]$. Coverage probability only exists given an unknown value of θ, so is written $P(\theta \in [L(\mathbf{X}), U(\mathbf{X})] \,|\, \theta)$, or commonly $P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$.

14.5.5 Definition

Assuming L, U are chosen satisfying 14.5.3, suppose

$$P_\theta(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) \geq 1 - \alpha$$

for $\alpha \in (0,1)$. $[L(\mathbf{X}), U(\mathbf{X})]$ is then called a $1-\alpha$ confidence interval (or a $[100(1-\alpha)]\%$ confidence interval).

14.5.6 Definition

The estimation error of the point estimator $\widehat{\Theta}$ to the parameter θ is $\widehat{\Theta} - \theta$.

14.5.7 Definition

The mean error or bias of $\widehat{\Theta}$ to θ is the expectation of the estimation error:

$$b_\theta(\widehat{\Theta}) = \text{E}(\widehat{\Theta} - \theta) = \text{E}(\widehat{\Theta}) - \theta.$$

14.5.8 Definition

$\widehat{\Theta}$ is unbiased $\iff \text{E}(\widehat{\Theta}) = \theta\ (\forall\, \theta) \iff b_\theta(\widehat{\Theta}) = 0\ (\forall\, \theta)$.

14.5.9 Definition

The mean squared error of $\widehat{\Theta}$ to θ is the expectation of the square of the estimation error:

$$\mathrm{E}\left((\widehat{\Theta}-\theta)^2\right).$$

14.5.10 Theorem

$$\mathrm{E}\left((\widehat{\Theta}-\theta)^2\right) = \left(b_\theta(\widehat{\Theta})\right)^2 + \mathrm{Var}(\widehat{\Theta}).$$

14.6 Special Case: Normal Random Variables

This short section states some theorems about sample mean and sample variance (see 14.2) in the special case where $X_1, ..., X_n$ are normally distributed. Some of the proofs require ideas from other year one courses.

14.6.1 Theorem

Let $X_1, ..., X_n$ be independent random variables $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i \in \{1, ..., n\}$. Let $Y = \sum_{i=1}^n X_i$.

$$Y \sim \mathrm{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

14.6.2 Corollary: Distribution of the Sample Mean

Let $X_1, ..., X_n$ be independent and identically distributed random variables $X_i \sim \mathrm{N}(\mu, \sigma^2)\ \forall\, i$. The sample mean is then distributed $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ (with parameters exactly as seen in 14.2.2).

14.6.3 Theorem

Let $Z_1, ..., Z_n \sim \mathrm{N}(0, 1)$ be independent and let $\mathbf{Z}$ be the column vector $(Z_1, ..., Z_n)^T$. Let A be an orthogonal $n \times n$ matrix and let $\mathbf{Y} = (Y_1, ..., Y_n)^T = A\mathbf{Z}$. The Y_i are also independent and each distributed $\sim \mathrm{N}(0, 1)$, and $\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Z_i^2$.

14.6.4 Corollary: Distribution of the Sample Variance

Let $X_1, ..., X_n$ be independent and identically distributed random variables $X_i \sim \mathrm{N}(\mu, \sigma^2)\ \forall\, i$. The sample variance S^2 satisfies $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. A useful further corollary is that $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$.

15 Statistical Models

15.1 Probability Models

Recall definitions 1.0.1 (sample space) and 4.2.1 (probability measure).

15.1.1 Definition: Probability Model

A probability model consists of

i A sample space $\Omega \neq \emptyset$;

ii A set of subsets of Ω (events);

iii A probability measure P which assigns a probability to each event.

15.2 Inference Using a Probability Model

Suppose we know the probability model for a random variable X, and we would like to make an inference about a future observation x. We could find a plausible value (e.g. $\mathrm{E}(x)$), but we may instead prefer to find a subset of $\mathrm{Im}\, X$ which has a high probability of containing x (by integration with unknown limits, for example).

15.3 Statistical Models

In reality, knowing the probability model and predicting outcomes is rare; we are normally drawing conclusions about the probability model based on observations. This requires statistical modelling. Suppose instead that we have made observations $\mathbf{x}$, and we don't know about the probability model of $\mathbf{X}$. We might consider a statistical model for the data $\mathbf{x}$ as a set of probability measures $\{P_\theta \mid \theta \in \Theta\}$, one of which is the true probability measure (along with the true value of the parameter θ) that gave rise to $\mathbf{X} = \mathbf{x}$.

15.3.1 Definition

Θ (above) is the space of all possible values of θ, called the parameter space. Note: In both statistical modelling and parameter estimation (14.5), θ refers to a parameter of interest, but be aware of the difference between $\widehat{\Theta}, \widehat{\theta}$, denoting an estimator and an observed estimation, and Θ, which is simply a set containing the true value of θ.

15.3.2 Definition: Statistical Model

A statistical model consists of

i Identifying random variables of interest (which are hypothetically observable);

ii Specifying a family of possible distributions for the random variables;

iii Specifying unknown parameters of the distributions which may be hypothetically observable;

iv Potentially specifying distributions for the unknown parameters.

If the parameters are thought of as random (as in item iv), then the distributions of the random variables corresponding to θ are conditional distributions given θ.

16 Likelihood

16.1 The Likelihood Function

16.1.1 Definition

Suppose $\{P_\theta \mid \theta \in \Theta\}$ outlines a statistical model for $\mathbf{X}$, and let each P_θ correspond to a probability density function f_θ. Having made an observation $\mathbf{x}$, the likelihood function is defined

$$\begin{aligned} L(\cdot|\mathbf{x}) : \Theta &\longrightarrow \mathbb{R} \\ \theta &\longmapsto f_\theta(\mathbf{x}). \end{aligned}$$

16.1.2 Definition

For a $\theta \in \Theta$, $L(\theta|\mathbf{x})$ is called the likelihood of θ given the observation $\mathbf{x}$.

16.1.3 Note

Since $f_\theta(\mathbf{x})$ can be thought of as $f(\mathbf{x}|\theta)$, the joint probability density (or mass) function of the random variables $\mathbf{X}$ given θ, we then have that

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta).$$

16.1.4 Theorem: Discrete Case

If the P_θ in the statistical model are discrete, then $f(\mathbf{x}|\theta)$ is simply the probability of observing $\mathbf{x}$ given that the parameter's true value is θ, and by the relationship above, $L(\theta|\mathbf{x})$ is equal to this probability. It is <u>not</u> the probability that the true value is θ given $\mathbf{x}$ is observed: suppose $L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x})$, we say that θ_1 is more plausible (or likely), not probable, since despite θ being unknown, it is fixed.

16.1.5 Theorem: Continuous Case

If X is continuous, $P_\theta(\mathbf{X} = \mathbf{x}) = 0 \; \forall \, \mathbf{x}$, so instead, we can consider a likelihood ratio, which for discrete random variables is simply

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})} = \frac{P_{\theta_1}(\mathbf{X} = \mathbf{x})}{P_{\theta_2}(\mathbf{X} = \mathbf{x})}, \quad \theta_1, \theta_2 \in \Theta.$$

For continuous random variables, consider the fact that $P_\theta(x - \delta < X < x + \delta) \approx 2\delta f(x|\theta)$ and that $2\delta f(x|\theta) = 2\delta L(\theta|x)$, giving

$$\frac{L(\theta_1|x)}{L(\theta_2|x)} \approx \frac{P_{\theta_1}(x - \delta < X < x + \delta)}{P_{\theta_2}(x - \delta < X < x + \delta)}, \quad \theta_1, \theta_2 \in \Theta,$$

so we can find an approximation of the ratio of probabilities of observing $\mathbf{x}$ for two values of θ.

16.1.6 Definition

Any likelihood function $L'(\theta|\mathbf{x}) = cL(\theta|\mathbf{x})$ where $c > 0$ is an equivalent likelihood function to $L(\theta|\mathbf{x})$, motivated by the fact that likelihood ratios are unchanged. Note that likelihood equivalence is an equivalence relation.

16.2 The Likelihood Principle

16.2.1 Principle

All evidence provided by a sample relevant to the parameters in a statistical model arises from the likelihood function. In other words, suppose different samples $\mathbf{x}_1$ and $\mathbf{x}_2$ give equivalent likelihood functions for θ, i.e.

$$L(\theta|\mathbf{x}_1) = C(\mathbf{x}_1, \mathbf{x}_2)L(\theta|\mathbf{x}_2) \quad \forall \, \theta,$$

the sets of conclusions to be made about θ from $\mathbf{x}_1$ and $\mathbf{x}_2$ should be the equivalent.

16.2.2 Note

The section above is stated as a principle because some believe it to be inconsistent with various statistical methods, and that additional information such as sampling procedure has an effect on the inferences made.

16.3 Maximum Likelihood Estimation

16.3.1 Definition

The maximum likelihood estimate of θ with likelihood function $L(\theta|\mathbf{x})$ is $\widehat{\theta}(\mathbf{x})$, the value at which $L(\theta|\mathbf{x})$ is at its maximum ($\widehat{\theta}(\mathbf{x}) = \arg\max L(\theta|\mathbf{x})$ over θ). Note that $\widehat{\theta}(\mathbf{x})$ is not a realisation of any point estimator $\widehat{\Theta}$, but the notation is used because it too represents an estimate for θ.

16.3.2 Definition

If the maximum likelihood estimate of θ with likelihood function $L(\theta|\mathbf{x})$ is $\widehat{\theta}(\mathbf{x})$, $\widehat{\theta}(\mathbf{X})$ is the maximum likelihood estimator of θ based on the random vector $\mathbf{X}$.

16.3.3 Definition

The log-likelihood, $\log L(\theta|\mathbf{x})$, is an example of a monotonic transformation that may instead be maximised if it is difficult to find the maximum of $L(\theta|\mathbf{x})$; finding the argument of its maximum finds the argument of the maximum of $L(\theta|\mathbf{x})$.

17 Correlation

17.0.1 Lemma

$Z \geq 0 \implies \mathrm{E}(Z) \geq 0$ (as in 10.1.4). The function $f(t) := \mathrm{E}[((X-\mu_X)t+(Y-\mu_Y))^2]$ ($\mu_X = \mathrm{E}(X)$, $\mu_Y = \mathrm{E}(Y)$), along with this lemma, is useful for some of the following proofs.

17.0.2 Theorem

1. $\mathrm{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \,\mathrm{Cov}(X_i, Y_j)$ (bilinearity of the covariance).
   We can then effectively combine theorems 10.2.3 and 11.7.5: for any constants $a, b \in \mathbb{R}$,
   $$\mathrm{Var}(aX + bY) = a^2 \,\mathrm{Var}(X) + b^2 \,\mathrm{Var}(Y) + 2ab \,\mathrm{Cov}(X, Y).$$
2. For two random variables X and Y, with variances σ_X^2 and σ_Y^2,
   $$|\,\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y.$$

17.0.3 Definition

Having seen the definitions of sample mean and variance, we now also define the sample covariance of the random variables $X_1, ..., X_n$ and $Y_1, ..., Y_n$, as

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

17.1 Correlation

17.1.1 Definition

The correlation (specifically the product-moment correlation coefficient or Pearson correlation) of the two random variables X and Y is

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the the square roots of the variances σ_X^2 and σ_Y^2 (known as standard deviation).

17.1.2 Corollary

$-1 \leq \rho_{XY} \leq 1$ $\forall$ random variables X, Y (this is a corollary to 17.0.2).

17.1.3 Lemma

Let Z be a random variable taking only non-negative values. $\mathrm{E}(Z) = 0 \iff p_Z(0) = 1$.

17.1.4 Corollary

$|\rho_{XY}| = 1 \iff \exists\, a, b \in \mathbb{R}$ such that $P(Y = aX + b) = 1$.
$(a > 0 \iff \rho_{XY} = 1$ and $a < 0 \iff \rho_{XY} = -1)$.

17.1.5 Definition

The sample correlation for a set of pairs of observations $(x_1, y_1), ..., (x_n, y_n)$ is the same as the Pearson correlation, but with observed sample covariance and square roots of observed sample variances (S_{xy}, S_{xx}, S_{yy} are defined as these quantities without the coefficient $\frac{1}{n-1}$):

$$r_{XY} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} := \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}.$$

17.1.6 Theorem

Suppose we have observed $(x_1, y_1), ..., (x_n, y_n)$ and define $u_i = ax_i + b$ and $v_i = cy_i + d$ ($a, b, c, d \in \mathbb{R}$, $a, c > 0$ or $a, c < 0$), then

$$r_{XY} = r_{UV}$$

and if a and c have different signs, then $r_{XY} = -r_{UV}$, (i.e. $r_{UV} = \left(\frac{a}{|a|}\right)\left(\frac{c}{|c|}\right) r_{XY}$).

18 Simple Linear Regression

While correlation gives a numerical interpretation of the strength of the relationship between random variables X and Y, simple linear regression attempts to establish whether that relationship can be written in the form $Y = \beta_0 + \beta_1 X$, which would facilitate the prediction of a realisation y_{n+1} from a new realisation x_{n+1}.

18.0.1 Definition

Suppose the relationship between X and Y can be specified by $Y = f(X)$. X is the predictor and Y is the response.

18.1 Finding the Parameters for Simple Linear Regression

Suppose now that we have n pairs of observations (x_i, y_i) (which are not necessarily perfectly correlated). We can obviously find n numbers e_i such that

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \forall\, i$$

where the r_i are the errors associated with the approximation of each y_i by $\beta_0 + \beta_1 x_i$. Since β_0, β_1 are undecided, the values of e_i are unknown, so suppose $\widehat{\beta_0}, \widehat{\beta_1}$ are decided upon, then $\widehat{e_i} := y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$ and the best choices for $\widehat{\beta_0}, \widehat{\beta_1}$ are therefore those that minimise $\sum_{i=1}^n (\widehat{e_i})^2 = \sum_{i=1}^n (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2$. Defining the residual sum of squares

$$\mathrm{RSS}(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

the problem is now to find $\widehat{\beta_0}, \widehat{\beta_1}$ such that $\mathrm{RSS}(\widehat{\beta_0}, \widehat{\beta_1}) \leq \mathrm{RSS}(\beta_0, \beta_1)\ \forall\, \beta_0, \beta_1 \in \mathbb{R}$.

$\widehat{\beta_0}$: $\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \equiv \sum_{i=1}^n ((y_i - \beta_1 x_i) - \beta_0)^2$ so let $z_i = y_i - \beta_1 x_i$, and we know that $\sum_{i=1}^n (z_i - \beta_0)^2$ is minimised by $\beta_0 = \overline{z} = \overline{y} - \beta_1 \overline{x}$.

$\widehat{\beta_1}$: Now, $\mathrm{RSS}(\widehat{\beta_0}, \beta_1) = \sum_{i=1}^n ((y_i - (\overline{y} - \beta_1 \overline{x}) - \beta_1 x_i)^2 \equiv \sum_{i=1}^n ((y_i - \overline{y}) - \beta_1 (x_i - \overline{x}))^2 \equiv \sum_{i=1}^n \big((y_i - \overline{y})^2 - 2\beta_1 (y_i - \overline{y})(x_i - \overline{x}) + \beta_1^2 (x_i - \overline{x})^2\big) = S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx}$. Completing the square for β_1 we obtain

$$\mathrm{RSS}(\widehat{\beta_0}, \beta_1) = S_{xx} \left(\beta_1 - \tfrac{S_{xy}}{S_{xx}}\right) + S_{yy} - \tfrac{{S_{xy}}^2}{S_{xx}}$$

and so the value of β_1 which minimises $\mathrm{RSS}(\widehat{\beta_0}, \beta_1)$ is $\widehat{\beta_1} := \frac{S_{xy}}{S_{xx}}$, and we have altogether:

$$\widehat{\beta_1} = \tfrac{S_{xy}}{S_{xx}} \qquad \widehat{\beta_0} = \overline{y} - \left(\tfrac{S_{xy}}{S_{xx}}\right) \overline{x}.$$

18.1.1 Theorem

The values of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are found to be the same if they are obtained instead by assuming $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ in the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and then maximising $L(\beta_0, \beta_1, \sigma^2 | (\mathbf{x}, \mathbf{y})) = \prod_{i=1}^n f_{\beta_0, \beta_1, \sigma^2}(\varepsilon_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}$ for fixed σ^2 (finding the maximum likelihood estimate for β_0, β_1).

18.1.2 Theorem

Using the same assumptions as in 18.1.1, once $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are found, the maximum likelihood estimate $\widehat{\sigma}^2$ is $\frac{1}{n} \sum_{i=1}^n (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2$.

18.1.3 Theorem

Now let $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i, ε_i are treated as independent, unobservable random variables. They are observable for the fixed values $\widehat{\beta_0}, \widehat{\beta_1}$ of β_0, β_1 and, letting σ^2 be $\mathrm{Var}(\varepsilon_i)$ as before,

1. $\mathrm{Cov}(\widehat{\beta_0}, \widehat{\beta_1}) = -\left(\frac{\overline{x}}{S_{xx}}\right) \sigma^2$.
2. $\mathrm{Cov}(Y_i, \widehat{\beta_0}) = \left(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}}\right) \sigma^2$.
3. $\mathrm{Cov}(Y_i, \widehat{\beta_1}) = \left(\frac{x_i - \overline{x}}{S_{xx}}\right) \sigma^2$.

19 Hypothesis Testing

19.0.1 Definition

The null hypothesis for an experiment (often written H_0) is the hypothesis that there is no significant difference between characteristics of a population - that any observed difference is due to chance or error.

19.0.2 Definition

The significance threshold (often α) is the probability of rejecting H_0, given H_0 is assumed to be true,

$$\alpha = P(H_0 \text{ is rejected} \mid H_0 \text{ is true}).$$

The p-value for an observation (p) is the probability of an observation being at least as extreme as this observation given H_0 were true; H_0 is rejected when the observation is statistically significant: $p < \alpha$ (so the choice of alpha and the decision to reject H_0 under this condition is in fact what defines it as above).

19.0.3 Definition

A one-tailed test is a test in which the significance threshold α is taken to represent a critical region at only one end of the distribution, decided upon beforehand by the nature of the alternative hypothesis.
A two-tailed test splits the critical region into two regions of size $\frac{\alpha}{2}$, at either end of the distribution.

19.1 Prerequisite: Using the t-Distribution

19.1.1 Lemma

Suppose $U \sim \mathrm{N}(0,1)$ and $V \sim \chi^2_\nu$,

$$\frac{U}{\sqrt{V/\nu}} \sim t_\nu$$

(not examinable in this module, but required for this section).

19.1.2 Theorem

(Corollary of 14.6.2): let $\overline{X}$ be the sample mean of n independently distributed random variables $X_i \sim \mathrm{N}(\mu, \sigma^2)$.

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

19.1.3 Theorem

Let $X_1, ..., X_n$ be as above, now distributed with unknown variance σ^2 but sample variance S^2.

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

(The proof follows from 19.1.1: $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ can be written in the form $\frac{U}{\sqrt{\frac{V}{n-1}}}$, with U as in 19.1.2 and V as in 14.6.4).

19.2 Confidence Intervals

19.2.1 Definition

A confidence interval is a range of plausible values for an unknown parameter, constructed such that the probability of the parameter falling inside the interval is equal to a given confidence level. They can be found by two-tailed tests using the assumed distribution of the unknown parameter (see below).

19.2.2 Theorem

If the sample mean ($\overline{X}$) and sample variance (S^2) of a vector of observations are evaluated, the equation $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ can be rearranged for μ, and thus a confidence interval for μ is $(\overline{X} - \frac{S}{\sqrt{n}}T_{(1-\frac{\alpha}{2})}, \overline{X} + \frac{S}{\sqrt{n}}T_{(1-\frac{\alpha}{2})})$, where $P(T < T_{(1-\frac{\alpha}{2})}) = 1 - \frac{\alpha}{2}$ and $1 - \alpha$ is the required confidence level.

19.2.3 Theorem

If instead the actual variance σ^2 is known, a confidence interval for μ is the same as above, but with $T_{(1-\frac{\alpha}{2})}$ found using the standard normal distribution.

20 Bayesian Inference

20.1 Prior and Posterior Distributions

20.1.1 Definition

Recall the definitions of joint distribution, marginal distribution and likelihood and note that, supposing $\mathbf{X}$ and θ have a joint distribution $f(\mathbf{x}, \theta)$ (and that the support of θ is the set Θ), the marginal joint density of $\mathbf{X}$ is denoted by $m(\mathbf{x})$, and derived as

$$m(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}, \theta)\ d\theta.$$

20.1.2 Definition

Suppose we treat the unknown parameter θ as a random variable, the distribution that θ is thought to follow (before any observations) is called the prior distribution and its density is denoted by $\pi(\theta)$.

20.1.3 Definition

Suppose that the random variables $\mathbf{X}$ are observed as $\mathbf{x}$. The conditional distribution of θ given $\mathbf{X} = \mathbf{x}$ is called the posterior distribution and its conditional density is denoted by $\pi(\theta|\mathbf{x})$.

20.1.4 Theorem

Suppose that the random variables $\mathbf{X}$ have joint density function $f(\mathbf{x}|\theta)$ and that $\mathbf{X}$ and θ have joint density $f(\mathbf{x}, \theta)$ with $\mathbf{X}$ having marginal joint density $m(\mathbf{x})$, but also that the value of θ is unknown and has prior distribution $\pi(\theta)$.

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}.$$

20.1.5 Note

Recall that the density funtion $f(\mathbf{x}|\theta)$ is equal to the likelihood function $L(\theta|\mathbf{x})$ and may be referred to as such.

20.1.6 Definition

Suppose Ψ is the family of distributions from which the prior distribution is chosen. If these prior distributions have their own parameters, they are known as hyperparameters; for example suppose that the density $f(x|\theta)$ is normally distributed with known variance and unknown mean θ, which is in turn normally distributed with mean μ and variance σ^2, μ and σ^2 are hyperparameters.

20.2 Conjugate Prior Distributions

20.2.1 Definition

Let $\mathbf{X}$ be conditionally distributed given θ and let $\mathcal{F}$ be the family of conditional distributions $f(\mathbf{x}|\theta)$. Let Ψ be the family of distributions from which the prior distribution $\pi(\theta)$ is chosen. Suppose, for any $\pi(\theta) \in \Psi$ and any observation $\mathbf{x} \subset \Omega$ (the sample space of $\mathbf{X}$) that $\pi(\theta|\mathbf{x}) \in \Psi$ also, then Ψ is a conjugate family to samples that follow distributions in $\mathcal{F}$.

20.2.2 Theorem

The beta distribution is conjugate to the Bernoulli distribution and the binomial distribution.

20.2.3 Corollary

If the prior distribution is uniform on the interval $[0, 1]$ and the likelihood is Bernoulli or binomial, then the posterior distribution is a beta distribution (the $\mathrm{U}(0, 1)$ distribution is equivalent to the $\mathrm{Beta}(1, 1)$ distribution).

20.2.4 Theorem

The normal distribution is conjugate to itself.

20.2.5 Theorem

The gamma distribution is conjugate to the exponential distribution.

21 The Central Limit Theorem

21.0.1 Definition

Let the sequence of random variables $X_1, X_2, X_3...$ have cumulative distribution functions $F_1, F_2, F_3, ...$ respectively. The sequence (X_n) converges in distribution to $X \iff F_n(x) \to F(x)\ \forall\, x$ at which F, the cumulative distribution function of X, is continuous, written $X_n \xrightarrow{\mathcal{D}} X$.

21.0.2 Definition

The characteristic function of the random variable X is

$$\phi_X(t) = \mathrm{E}(e^{itX}) = \int_{-\infty}^{\infty} e^{itx}\ F(dx)$$

where F is the cumulative distribution function of X. Note that the integration here is with respect to $F(dx)$. When the probability density function f_X is known, the characteristic function is of course $\int_{-\infty}^{\infty} e^{itx} f_X(x)\ dx$.

21.0.3 Note

Since $|e^{itX}| = 1$, $\phi_X(t) = \mathrm{E}(e^{itX})$ always exists.

21.0.4 Theorem

For any random variable X,

1. ϕ_X is continuous.
2. $\phi_X(0) = 1$
3. $|\phi_X(t)| \leq 1\ \forall\, t$
4. $\phi_{aX+b}(t) = e^{ibt}\phi_X(at)\ \forall\, a, b \in \mathbb{R}$.

21.0.5 Theorem

Let $X_1, ..., X_n$ be independent random variables.

$$\phi_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^{n} \phi_{X_i}(t).$$

21.0.6 Theorem

Random variables X and Y follow the same distribution $\iff \phi_X(t) = \phi_Y(t)\ \forall\, t$.

21.0.7 Definition

Let the cumulative distribution function of X be F. Recall the definition of the n^{th} raw moment: $\mu'_n := \mathrm{E}(X^n)$ (14.1.4). In this section it will be denoted by m_n, and we also define the n^{th} absolute moment M_n.

$$m_n = \mathrm{E}(X^n) = \int_{-\infty}^{\infty} x^n\ F(dx), \qquad M_n = \mathrm{E}(|X|^n) = \int_{-\infty}^{\infty} |x|^n\ F(dx).$$

21.0.8 Lemma

M_n (of X) is convergent $\implies$ the n^{th} derivative of ϕ_X exists and is equal to

$$\phi_X^{(n)}(t) = i^n \int_{-\infty}^{\infty} e^{itx} x^n\ F(dx).$$

21.0.9 Corollary

m_2 (of X) is convergent $\implies$

$$\phi'_X(0) = im_1 \quad \text{and} \quad \phi''_X(0) = -m_2.$$

21.1 The Continuity Theorem

21.1.1 Theorem: Lévy's Continuity Theorem

The sequence of distributions (F_n) converges pointwise to a distribution $F \iff$ the corresponding sequence of characteristic functions (ϕ_n) converges pointwise to a characteristic function ϕ, which is continuous at 0. Note that a characteristic function is defined entirely by a distribution, so the random variable subscript is not included in this theorem. Additionally, ϕ is the characteristic function of F and hence is continuous everywhere, and the convergence $\phi_n \to \phi$ is uniform.

21.1.2 Corollary

If it exists, the pointwise limit of a sequence of characteristic functions is also a characteristic function.

21.2 The Central Limit Theorem

21.2.1 Definition

The n^{th} normalised sum of the sequence of random variables $X_1, X_2, X_3, ...$ is

$$S_n = \frac{X_1 + ... + X_n}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

Note that this is not equal to the sample mean.

21.2.2 Lemma

If the mean and variance of a distribution are finite, so is the second raw moment.

21.2.3 Lemma

The characteristic function of the standard normal distribution is $e^{-\frac{t^2}{2}}$.

21.2.4 Theorem: The Central Limit Theorem

Let $X_1, X_2, X_3, ...$ be a sequence of identically and independently distributed random variables, for which we have (or can assume that) $\mathrm{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 1 \; \forall \, i$. Let S_n be the n^{th} normalised sum of the sequence.

$$S_n \xrightarrow{\mathcal{D}} Z \sim \mathrm{N}(0, 1).$$

21.2.5 Corollary

The central limit theorem can be generalised to the case where the sequence of identically and independently distributed random variables $X_1, X_2, X_3, ...$ have common mean and variance μ and σ^2 respectively. Instead, let S_n be the n^{th} normalised sum of the sequence $Z_1, Z_2, Z_3, ...$ where $Z_i = \frac{X_i - \mu}{\sigma}$.

$$S_n \xrightarrow{\mathcal{D}} Z \sim \mathrm{N}(0, 1).$$