1 Introduction

1.1 Motivation

1.2 Preliminaries

Sets

A **set** A is an unordered collection of objects, $x_1, \dots, x_n, \dots$:

$$A = \{x_1, \dots, x_n, \dots\}$$

We write $x \in A$ to say that 'x is an element of A', or 'x belongs to A'. The collection of all possible objects in a given context is called the **universe** Ω.
An **ordered set** is written $A = (1, 2, \dots)$.

Subsets

A set A is a **subset** of a set B if $x \in A$ implies that $x \in B$: we write $A \subset B$.
Note that $\emptyset \subset A$ for every set A. Thus,

$$\emptyset \subset \mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}, \qquad \emptyset \subset \mathbb{I} \subset \mathbb{C}$$

Cardinal of a set

A finite set A has a finite number of elements, and this number is called its **cardinal** :

$$\text{card } A, \quad \#A, \quad |A|$$

Boolean operations

Let $A, B \subset \Omega$. Then we can define

the **union** and the **intersection** of A and B to be

$$A \cup B = \{x \in \Omega : x \in A \text{ or } x \in B\}, \quad A \cap B = \{x \in \Omega : x \in A \text{ and } x \in B\}$$

the **complement** of A in Ω to be $A^c = \{x \in \Omega : x \notin A\}$

the **difference between** A **and** B to be

$$A \setminus B = A \cap B^c = \{x \in \Omega : x \in A \text{ and } x \notin B\}, \qquad A \setminus B \neq B \setminus A$$

the **symmetric difference** to be

$$A \bigtriangleup B = (A \setminus B) \cup (B \setminus A)$$

If $\{A_j\}_{j=1}^{\infty}$ is an infinite set of the subsets of Ω, then

$$\bigcup_{j=1}^{\infty} A_j = A_1 \cup A_2 \cup \cdots : \text{ those } x \in \Omega \text{ that belong to } \textbf{at least} \text{ one } A_j$$

$$\bigcap_{j=1}^{\infty} A_j = A_1 \cap A_2 \cap \cdots : \text{ those } x \in \Omega \text{ that belong to } \textbf{every} \ A_j$$

Partition

A **partition** of Ω is a collection of non-empty subsets $A_1, \dots, A_n$ in Ω such that

the A_j are **exhaustive**, i.e., $A_1 \cup \cdots \cup A_n = \Omega$

the A_j are **disjoint**, i.e., $A_i \cap A_j = \emptyset$, for $i \neq j$

A partition can also be composed of an infinity of sets $\{A_j\}_{j=1}^{\infty}$

Cartesian product

The **Cartesian product** of two sets A, B is the set of **ordered** pairs

$$A \times B = \{(a,b) : a \in A, \ b \in B\}$$

In the same way

$$A_1 \times \cdots \times A_n = \{(a_1, \dots, a_n\} : a_1 \in A_1, \dots, a_n \in A_n$$

If $A_1 = \cdots = A_n = A$, then we write $A_1 \times \cdots \times A_n = A^n$.
As the pairs are ordered, $A \times B \neq B \times A$.
If $A_1, \dots, A_n$ are all finite, then

$$|A_1 \times \cdots \times A_n| = |A_1| \times \cdots \times |A_n|$$

1.3 Combinatorics

Reminders

If $A_1, \dots, A_k$ are sets, then

multiplication :

$$|A_1 \times \cdots \times A_k| = |A_1| \times \cdots \times |A_k|$$

addition : if the A_j are disjoint, then

$$|A_1 \cup \cdots \cup A_k| = |A_1| + \cdots + |A_k|$$

Ordered selection

A **permutation** of n distinct objects is an ordered set of those objects.
Given n distinct objects, the number of different permutations (without repetition) of length $r \leq n$ is

$$n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

Given $n = \sum_{i=1}^{r} n_i$ objects of r different types, where n_i is the number of objects of type i that are indistinguishable from one another, the number of permutations (without repetition) of the n objects is

$$\frac{n!}{n_1!n_2!\cdots n_r!}$$

Multinomial and binomial coefficients

Let $n_1, \dots, n_r$ be integers in $0, 1, \dots, n$, having total $n_1 + \cdots + n_r = n$. Then

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1!n_2!\cdots n_r!}$$

is called the **multinomial coefficient**.
The most common case arises when $r = 2$:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}(= C_n^k)$$

is called the **binomial coefficient**.

Non-ordered selection

The number of ways of choosing a set of r objects from a set of n distinct objects without repetition is

$$\frac{n!}{r!(n-r)!} = \binom{n}{r}$$

The number of ways of distributing n distinct objects into r distinct groups of size $n_1, \dots, n_r$, where $n_1 + \cdots + n_r = n$, is

$$\frac{n!}{n_1!n_2!\cdots n_r!}$$

Binomial coefficients

If $n, m \in \{1, 2, 3, \dots\}$ and $r \in \{0, \dots, n\}$, then :

$\binom{n}{r} = \binom{n}{n-r}$

$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$ (Pascal's triangle)

$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$ (Newton's binomial theorem)

$\lim_{n\to\infty} n^{-r}\binom{n}{r} = \frac{1}{r!}, \quad r \in \mathbb{N}$

Partitions of integers

The number of distinct vectors $(n_1, \dots, n_r)$ of positive integers, $n_1, \dots, n_r > 0$, satisfying $n_1 + \cdots + n_r = n$, is

$$\binom{n-1}{r-1}$$

The number of distinct vectors $(n_1, \ldots, n_r)$ of positive integers, $n_1, \ldots, n_r \geq 0$, satisfying $n_1 + \cdots + n_r = n$, is

$$\binom{n+r-1}{n}$$

Some series

A **geometric series** is of the form $a, a\theta, a\theta^2, \ldots$; we have

$$\sum_{i=0}^{n} a\theta^i = \begin{cases} a\frac{1-\theta^{n+1}}{1-\theta} & \theta \neq 1 \\ a(n+1) & \theta = 1 \end{cases}$$

If $|\theta| < 1$, then $\sum\limits_{i=0}^{\infty} \theta^i = \frac{1}{1-\theta}$, and

$$\sum_{i=0}^{\infty} \frac{i!}{r!(i-r)!}\theta^{i-r} = \frac{1}{(1-\theta)^{r+1}}, \qquad r = 1, 2, \ldots$$

The **exponential series**

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

converges absolutely for all $x \in \mathbb{C}$.

2 Probability

2.1 Probability Spaces

Calculation of probabilities

$$\text{probability of an event} = \frac{\text{\# of times event takes place}}{\text{\# experiments carried out}}$$

Probability space

A random experiment is modelled by a **probability space**.

A **probability space** $(\Omega, \mathcal{F}, \mathrm{P})$ is a mathematical object associated with a random experiment, comprising :

a set Ω, the **sample space (universe)**, which contains all the possible **results (outcomes, elementary events)** ω of the experiment

a collection $\mathcal{F}$ of subsets of Ω. These subsets are called **events**, and $\mathcal{F}$ is called the **event space**

a function $\mathrm{P} : \mathcal{F} \mapsto [0,1]$ called a **probability distribution**, which associates a probability $\mathrm{P}(A) \in [0,1]$ to each $A \in \mathcal{F}$

Sample space

The sample space Ω is the space composed of elements representing all the possible results of a random experiment. Each element $\omega \in \Omega$ is associated with a different result.

Event space

$\mathcal{F}$ is a set of subsets of Ω which represents the events of interest. An **event space** $\mathcal{F}$ is a set of the subsets of Ω such that if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. $\mathcal{F}$ is also called a **sigma-algebra**. $\Omega \in \mathcal{F}$, $\emptyset \in \mathcal{F}$.

Probability function

A **probability distribution** P assigns a probability to each element of the event space $\mathcal{F}$, with the following properties :

(P1) if $A \in \mathcal{F}$, then $0 \leq \mathrm{P}(A) \leq 1$

(P2) $\mathrm{P}(\Omega) = 1$

(P3) if $\{A_i\}_{i=1}^{\infty}$ are pairwise disjoint (i.e., $A_i \cap A_j = \emptyset, i \neq j$), then

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

Let $A, B, \{A_i\}_{i=1}^{\infty}$ be events of the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Then

(a) $\mathrm{P}(\emptyset) = 0$

(b) $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$

(Note) $\mathrm{P}(\Omega) = 1$

Continuity

A function f is continuous at x if for every sequence $\{x_n\}$ such that

$$\lim_{n\to\infty} x_n = x, \text{ we have } \lim_{n\to\infty} f(x_n) = f(x)$$

For all sequences of sets for which

$$\lim_{n\to\infty} A_n = A, \text{ we have } \lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A)$$

Hence P is called a **continuous set function**.

Inclusion-exclusion formulae

If $A_1, \dots, A_n$ are events of $(\Omega, \mathcal{F}, \mathrm{P})$, then

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{r=1}^{n}(-1)^{r+1} \sum_{1\le i_1<\cdots<i_r\le n} \mathrm{P}(A_{i_1} \cap \cdots \cap A_{i_r})$$

The number of terms is

$$\binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n - 1$$

2.2 Conditional Probability

Let A, B be events of the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, such that $\mathrm{P}(B) > 0$. Then the **conditional probability of A given B** is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

Bayes' theorem

(Law of total probability) Let $\{B_i\}_{i=1}^{\infty}$ be pairwise disjoint events (i.e. $B_i \cap B_j = \emptyset,\ i \neq j$) of the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, and let A be an event satisfying $A \subset \bigcup_{i=1}^{\infty} B_i$. Then

$$\mathrm{P}(A) = \sum_{i=1}^{\infty} \mathrm{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\mathrm{P}(B_i)$$

(Bayes) Suppose that the conditions above are satisfied, and that $\mathrm{P}(A) > 0$. Then

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\mathrm{P}(B_j)}{\sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\mathrm{P}(B_i)}, \qquad j \in \mathbb{N}$$

Multiple conditioning

('Prediction decomposition') Let $A_1, \dots, A_n$ be events in a probability space. Then

$$\mathrm{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=2}^{n} \mathrm{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1}) \times \mathrm{P}(A_1)$$

2.3 Independence

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are **independent** (we write $A \perp\!\!\!\perp B$) iff

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$$

This implies that

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A)\mathrm{P}(B)}{\mathrm{P}(B)} = \mathrm{P}(A)$$

and by symmetry $\mathrm{P}(B \mid A) = \mathrm{P}(B)$.

Types of independence

The events $A_1, \ldots, A_n$ are **(mutually) independent** if for all sets of indices $F \subset \{1, \ldots, n\}$,

$$\mathrm{P}\left(\bigcap_{i \in F} A_i\right) = \prod_{i \in F} \mathrm{P}(A_i)$$

The events $A_1, \ldots, A_n$ are **pairwise independent** if

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i)\mathrm{P}(A_j), \qquad 1 \leq i < j \leq n$$

The events $A_1, \ldots, A_n$ are **conditionally independent given** B if for all sets of indices $F \subset \{1, \ldots, n\}$,

$$\mathrm{P}\left(\bigcap_{i \in F} A_i \mid B\right) = \prod_{i \in F} \mathrm{P}(A_i \mid B)$$

2.4 Edifying Examples

3 Random Variables

3.1 Basic Ideas

Random variables

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A **random variable (rv)** $X : \Omega \mapsto \mathbb{R}$ is a function from the sample space Ω taking values in the real numbers $\mathbb{R}$.
The set of values taken by X,

$$D_X = \{x \in \mathbb{R} : \exists \omega \in \Omega \text{ such that } X(\omega) = x\}$$

is called the **support** of X. If D_X is countable, then X is a **discrete random variable**.
The random variable X associates probabilities to subsets S included in $\mathbb{R}$, given by

$$\mathrm{P}(X \in S) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \in S\})$$

In particular, we set $A_x = \{\omega \in \Omega : X(\omega) = x\}$.

Bernoulli random variables

A random variable that takes only the values 0 and 1 is called an **indicator variable**, or a **Bernoulli random variable**, or a **Bernoulli trial**.

Mass functions

The **probability mass function (PMF)** of a discrete random variable X is

$$f_X(x) = \mathrm{P}(X = x) = \mathrm{P}(A_x), \qquad x \in \mathbb{R}$$

It has two key properties :

(i) $f_X(x) \geq 0$, and it is only positive for $x \in D_X$, where D_X is the image of the function X, i.e., the **support** of f_X

(ii) the total probability $\sum_{\{i : x_i \in D_X\}} f_X(x_i) = 1$

When there is no risk of confusion, we write $f_X \equiv f$ and $D_X \equiv D$.

Binomial random variable

A **binomial** random variable X has PMF

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n, \qquad n \in \mathbb{N}, \qquad 0 \leq p \leq 1 \qquad \heartsuit$$

We write $X \sim \mathrm{B}(n, p)$, and call n the **denominator** and p the **probability of success**.
We use $\sim$ to mean 'has the distribution'.
The binomial model is used when we are considering the number of 'successes' of a trial which is independently repeated a fixed number of times, and where each trial has the same probability of success.

Geometric distribution

A **geometric** random variable X has PMF

$$f_X(x) = p(1-p)^{x-1}, \qquad x = 1, 2, \ldots, \qquad 0 \leq p \leq 1 \qquad \heartsuit$$

We write $X \sim \mathrm{Geom}(p)$, and we call p the **success probability**.
This models the waiting time until a first event, in a series of independent trials having the same success probability.
(Lack of memory) If $X \sim \mathrm{Geom}(p)$, then

$$\mathrm{P}(X > n + m \mid X > m) = \mathrm{P}(X > n)$$

This is also sometimes called **memorylessness**.

Negative binomial distribution

A **negative binomial** random variable X with parameters n and p has PMF

$$f_X(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}, \qquad x = n, n+1, n+2, \ldots, \qquad 0 \leq p \leq 1$$

We write $X \sim \mathrm{NegBin}(n, p)$.
It models the waiting time until the nth success in a series of independent trials having the same success probability.

Negative binomial distribution : alternative version

We sometimes write the geometric and negative binomial variables in a more general form, setting $Y = X - n$, and then the probability mass function is

$$f_Y(y) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha) y!} p^\alpha (1-p)^y, \qquad y = 0, 1, 2, \ldots, \qquad 0 \leq p \leq 1, \qquad \alpha > 0$$

where

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du, \qquad \alpha > 0$$

is the **Gamma function**. The principal properties of $\Gamma(\alpha)$ are :

$$\Gamma(1) = 1$$

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \qquad \alpha > 0$$

$$\Gamma(n) = (n-1)!, \qquad n = 1, 2, 3, \ldots$$

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

Hypergeometric distribution

We draw a sample of m balls without replacement from an urn containing w white balls and b black balls. Let X be the number of white balls drawn. Then

$$\mathrm{P}(X = x) = \frac{\binom{w}{x}\binom{b}{m-x}}{\binom{w+b}{m}}, \qquad x = \max(0, m-b), \ldots, \min(w, m)$$

and the distribution of X is **hypergeometric**. We write $X \sim \mathrm{HyperGeom}(w, b; m)$.

Discrete uniform distribution

A **discrete uniform** random variable X has PMF

$$f_X(x) = \frac{1}{b-a+1}, \qquad x = a, a+1, \ldots, b, \qquad a < b, \qquad a, b \in \mathbb{Z}$$

We write $U \sim \mathrm{DU}(a,b)$.
This definition generalizes the outcome of a die throw, which corresponds to the $\mathrm{DU}(1,6)$ distribution.

Poisson distribution

A **Poisson** random variable X has the PMF

$$f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \qquad x = 0, 1, \ldots, \qquad \lambda > 0$$

We write $X \sim \mathrm{Pois}(\lambda)$.
The Poisson distribution appears everywhere in probability and statistics, often as a model for counts, or for a number of rare events.
It also provides approximations to probabilities, for example for random permutations or the binomial distribution.

Cumulative distribution function

The **cumulative distribution function (CDF)** of a random variable X is

$$F_X(x) = \mathrm{P}(X \le x), \qquad x \in \mathbb{R}$$

If X is discrete, we can write

$$F_X(x) = \sum_{\{x_i \in D_X : x_i \le x\}} \mathrm{P}(X = x_i)$$

which is a step function with jumps at the points of the support D_X of $f_X(x)$.
When there is no risk of confusion, we write $F \equiv F_X$.
Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \mapsto \mathbb{R}$ a random variable. Its cumulative distribution function F_X satisfies :

(a) $\lim_{x \to -\infty} F_X(x) = 0$

(b) $\lim_{x \to \infty} F_X(x) = 1$

(c) F_X is non-decreasing, so $F_X(x) \le F_X(y)$ for $x \le y$

(d) F_X is continuous on the right, thus

$$\lim_{t \to 0} F_X(x+t) = F_X(x), \qquad x \in \mathbb{R}$$

(e) $\mathrm{P}(X > x) = 1 - F_X(x)$

(f) if $x < y$, then $\mathrm{P}(x < X \le y) = F_X(y) - F_X(x)$

Transformations of discrete random variables

If X is a random variable and $Y = g(X)$, then

$$f_Y(y) = \sum_{x : g(x) = y} f_X(x)$$

3.2 Expectation

Let X be a discrete random variable for which $\sum_{x \in D_X} |x| f_X(x) < \infty$, where D_X is the support of f_X. The **expectation** (or **expected value** or **mean**) of X is

$$\mathrm{E}(X) = \sum_{x \in D_X} x \mathrm{P}(X = x) = \sum_{x \in D_X} x f_X(x)$$

$\mathrm{E}(X)$ is also sometimes called the "average of X".

Expected value of a function

Let X be a random variable with mass function f, and let g be a real-valued function of X. Then

$$\mathrm{E}\{g(X)\} = \sum_{x \in D_X} g(x) f(x)$$

when $\sum\limits_{x \in D_X} |g(x)| f(x) < \infty$.

Let X be a random variable with a finite expected value $\mathrm{E}(X)$, and let $a, b \in \mathbb{R}$ be constants. Then

(a) $\mathrm{E}(\cdot)$ is a linear operator, i.e., $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$

(b) if $g(X)$ and $h(X)$ have finite expected values, then

$$\mathrm{E}\{g(X) + h(X)\} = \mathrm{E}\{g(X)\} + \mathrm{E}\{h(X)\}$$

(c) if $\mathrm{P}(X = b) = 1$, then $\mathrm{E}(X) = b$

(d) if $\mathrm{P}(a < X \leq b) = 1$, then $a < \mathrm{E}(X) \leq b$

(e) $\{\mathrm{E}(X)\}^2 \leq \mathrm{E}(X^2)$

Moments of a distribution

If X has a PMF $f(x)$ such that $\sum\limits_{x} |x|^r f(x) < \infty$, then

(a) the rth **moment** of X is $\mathrm{E}(X^r)$

(b) the rth **central moment** of X is $\mathrm{E}[\{X - \mathrm{E}(X)\}^r]$

(c) the **variance** of X is $\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \mathrm{E}(X^2) - \mathrm{E}(X)^2$ (the second central moment)

(d) the **standard deviation** of X is defined as $\sqrt{\mathrm{var}(X)}$ (non-negative)

(e) the rth **factorial moment** of X is $\mathrm{E}\{X(X-1)\cdots(X-r+1)\}$

$\mathrm{E}(X)$ and $\mathrm{var}(X)$ are the most important moment : they represent the 'average value' $\mathrm{E}(X)$ of X, and the 'average squared distance' of X from its mean, $\mathrm{E}(X)$.

Let X be a random variable whose variance exists, and let a, b be constants. Then $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X)$.

If X takes its values in $\{0, 1, \dots\}$, $r \geq 2$, and $\mathrm{E}(X) < \infty$, then

$$\mathrm{E}(X) = \sum_{x=1}^{\infty} \mathrm{P}(X \geq x) \qquad \heartsuit$$

$$\mathrm{E}\{X(X-1)\cdots(X-r+1)\} = r \sum_{x=r}^{\infty} (x-1)\cdots(x-r+1)\mathrm{P}(X \geq x) \qquad \xcancel{\heartsuit}$$

3.3 Conditional Probability Distributions

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, on which we define a discrete random variable X, and let be $B \in \mathcal{F}$ with $\mathrm{P}(B) > 0$. Then the **conditional probability mass function** of X given B is

$$f_X(x \mid B) = \mathrm{P}(X = x \mid B) = \frac{\mathrm{P}(A_x \cap B)}{\mathrm{P}(B)}$$

where $A_x = \{\omega \in \Omega : X(\omega) = x\}$.
Often B is an event of form $X \in \mathcal{B}$, for some $\mathcal{B} \subset \mathbb{R}$, and then

$$f_X(x \mid B) = \frac{\mathrm{P}(X = x, X \in \mathcal{B})}{\mathrm{P}(X \in \mathcal{B})} = \frac{\mathrm{P}(X \in \mathcal{B} \mid X = x)\mathrm{P}(X = x)}{\mathrm{P}(X \in \mathcal{B})} = \frac{I(x \in \mathcal{B})}{\mathrm{P}(X \in \mathcal{B})} f_X(x)$$

so $f_X(x \mid B) = 0 \ (x \notin \mathcal{B})$

Conditional expected value

Suppose that $\sum_x |g(x)| f_X(x \mid B) < \infty$. Then the conditional expected value of $g(X)$ given B is

$$\mathrm{E}\{g(X) \mid B\} = \sum_x g(x) f_X(x \mid B)$$

Let X be a random variable with expected value $\mathrm{E}(X)$, let $\{B_i\}_{i=1}^{\infty}$ be a partition of Ω, $\mathrm{P}(B_i) > 0$ for all i, and let the sum be absolutely convergent. Then

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} \mathrm{E}(X \mid B_i)\mathrm{P}(B_i)$$

3.4 Notions of Convergence

Convergence of distributions

Let $\{X_n\}, X$ be random variables whose cumulative distribution functions are $\{F_n\}, F$. Then we say that the random variables $\{X_n\}$ **converge in distribution** (or **converge in law**) to X, if, for all $x \in \mathbb{R}$ where F is continuous,

$$F_n(x) \to F(x), \qquad n \to \infty$$

We write $X_n \xrightarrow{D} X$
If $D_X \subset \mathbb{Z}$, then $F_n(x) \to F(x)$ if $f_n(x) \to f(x)$ for all $x, n \to \infty$

Law of small numbers

Let $X_n \sim B(n, p_n)$, and suppose that $np_n \to \lambda > 0$ when $n \to \infty$. Then $X_n \xrightarrow{D} X$, where $X \sim \mathrm{Pois}(\lambda)$.
It can be used to approximate binomial probabilities for large n and small p by Poisson probabilities.

Which distribution ?

Is X based on independent trials (0/1) with a same probability p, or on draws from a finite population, with replacement ?

If **Yes**, is the total number of trials n fixed, so $X \in \{0, \ldots, n\}$?

If **Yes** : use the **binomial** distribution, $X \sim \mathrm{B}(n, p)$ (and thus the **Bernoulli** distribution if $n = 1$)

If $n \approx \infty$ or $n \gg np$, we can use the **Poisson** distribution, $X \sim \mathrm{Pois}(np)$

If **No**, then $X \in \{n, n+1, \dots\}$, and we use the **geometric** (if X is the number of trials until one success) or **negative binomial** (if X is the number of trials until the last of several successes) distributions

If **No**, then if the draw is independent but without replacement from a finite population, then $X \sim$ **hypergeometric** distribution

4 Continuous Random Variables

4.1 Basic Ideas

Continuous random variables

Until now we supposed that the support

$$D_X = \{x \in \mathbb{R} : X(\omega) = x, \omega \in \Omega\}$$

of X is countable, so X is a discrete random variable. We suppose now that D_X is not countable, which implies also that Ω itself is not countable.
Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. The **cumulative distribution function** of a rv X defined on $(\Omega, \mathcal{F}, \mathrm{P})$ is

$$F(x) = \mathrm{P}(X \leq x) = \mathrm{P}(\mathcal{B}_x), \qquad x \in \mathbb{R}$$

where $\mathcal{B}_x = \{\omega : X(\omega) \leq x\} \subset \Omega$.

Probability density functions

A random variable X is **continuous** if there exists a function $f(x)$, called the **probability density function (or density) (PDF)** of X, such that

$$\mathrm{P}(X \leq x) = F(X) = \int_{-\infty}^{x} f(u)du, \qquad x \in \mathbb{R}$$

The properties of F imply that

(i) $f(x) \geq 0$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

(Remark) $f(x) = \frac{dF(x)}{dx}$

(Remark) Since $\mathrm{P}(x < X \leq y) = \int_x^y f(u)du$ for $x < y$, for all $x \in \mathbb{R}$,

$$\mathrm{P}(X = x) = \lim_{y \to x} \mathrm{P}(x < X \leq y) = \lim_{y \to x} \int_x^y f(u)du = \int_x^x f(u)du = 0$$

(Remark) If X is discrete, then its PMF $f(x)$ is often also called its density function

Basic distributions

(Uniform distribution) The random variable U having density

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

is called a **uniform random variable**. We write $U \sim \mathrm{U}(a, b)$.
(Exponential distribution) The random variable X having density

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

is called an **exponential random variable** with parameter $\lambda > 0$. We write $X \sim \exp(\lambda)$.

Gamma distribution

The random variable X having density

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

is called a **gamma random variable** with parameters $\alpha, \lambda > 0$. We write $X \sim \text{Gamma}(\alpha, \lambda)$. Here α is called the **shape parameter** and λ is called the **rate**, with λ^{-1} the **scale parameter**. By letting $\alpha = 1$ we get the exponential density, and when $\alpha = 2, 3, \ldots$ we get the **Erlang** density.

Laplace distribution

The random variable X having density

$$f(x) = \frac{\lambda}{2} e^{-\lambda |x-\eta|}, \qquad x \in \mathbb{R}, \qquad \eta \in \mathbb{R}, \qquad \lambda > 0$$

is called a **Laplace random variable** (or sometimes a **double exponential** random variable).

Pareto distribution

The random variable X with cumulative distribution function

$$F(x) = \begin{cases} 0 & x < \beta \\ 1 - \left(\frac{\beta}{x}\right)^\alpha & x \geq \beta \end{cases} \qquad \alpha, \beta > 0$$

is called a **Pareto random variable**.

Moments

Let $g(x)$ be a real-valued function, and X a continuous random variable of density $f(x)$. Then if $\text{E}\{|g(X)|\} < \infty$, we define the **expectation** of $g(X)$ to be

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

In particular the **expectation** and the **variance** of X are

$$\text{E}(X) = \textstyle\int_{-\infty}^{\infty} x f(x) dx$$

$$\text{var}(X) = \textstyle\int_{-\infty}^{\infty} \{x - \text{E}(X)\}^2 f(x) dx = \text{E}(X^2) - \text{E}(X)^2$$

Conditional densities

For reasonable subsets $\mathcal{A} \subset \mathbb{R}$ we have

$$F_X(x \mid X \in \mathcal{A}) = \text{P}(X \leq x \mid X \in \mathcal{A}) = \frac{\text{P}(X \leq x \cap X \in \mathcal{A})}{\text{P}(X \in \mathcal{A})} = \frac{\int_{\mathcal{A}_x} f(y) dy}{\text{P}(X \in \mathcal{A})}$$

where $\mathcal{A}_x = \{y : y \le x, y \in \mathcal{A}\}$, and

$$f_X(x \mid X \in \mathcal{A}) = \begin{cases} \frac{f_X(x)}{\mathrm{P}(X \in \mathcal{A})} & x \in \mathcal{A} \\ 0 & otherwise \end{cases}$$

With $I(X \in \mathcal{A})$ the indicator variable of the event $X \in \mathcal{A}$, we can write

$$\mathrm{E}\{g(X) \mid X \in \mathcal{A}\} = \frac{\mathrm{E}\{g(X)\ I(X \in \mathcal{A})\}}{\mathrm{P}(X \in \mathcal{A})}$$

X discrete or continuous ?

	Discrete	Continuous
Support D_X	countable	contains an interval $(x_-, x_+) \subset \mathbb{R}$
f_X	mass function dimensionless $0 \le f_X(x) \le 1$ $\sum_{x \in \mathbb{R}} f_X(x) = 1$	density function units $[x]^{-1}$ $0 \le f_X(x)$ $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$
$F_X(a) = \mathrm{P}(X \le a)$	$\sum_{x \le a} f_X(x)$	$\int_{-\infty}^{a} f_X(x)\ dx$
$\mathrm{P}(X \in \mathcal{A})$	$\sum_{x \in \mathcal{A}} f_X(x)$	$\int_{\mathcal{A}} f_X(x)\ dx$
$\mathrm{P}(a < X \le b)$	$\sum_{\{x: a < x \le b\}} f_X(x)$	$\int_a^b f_X(x)\ dx$
$\mathrm{P}(X = a)$	$f_X(a) \ge 0$	$\int_a^a f_X(x)\ dx = \mathbf{0}$
$\mathrm{E}\{g(X)\}$ (if well defined)	$\sum_{x \in \mathbb{R}} g(x) f_X(x)$	$\int_{-\infty}^{\infty} g(x) f_X(x)\ dx$

4.2 Further Ideas

Quantiles

Let $0 < p < 1$. We define the p **quantile** of the cumulative distribution function $F(x)$ to be

$$x_p = \inf\{x : F(x) \ge p\}$$

For most continuous random variables, x_p is unique and equals $x_p = F^{-1}(p)$, where F^{-1} is the inverse function F ; then x_p is the value for which $\mathrm{P}(X \le x_p) = p$. In particular, we call the 0.5 quantile the **median** of F.

Transformations

We often consider $Y = g(X)$, where g is a known function, and we want to calculate F_Y and f_Y given F_X and f_X.

General transformation

Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a function and $\mathcal{B} \subset \mathbb{R}$ any subset of $\mathbb{R}$. Then $g^{-1}(\mathcal{B}) \subset \mathbb{R}$ is the set for which $g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$. Let $Y = g(X)$ be a random variable and $\mathcal{B}_y = (-\infty, y]$. Then

$$F_Y(y) = \mathrm{P}(Y \le y) = \begin{cases} \int_{g^{-1}(\mathcal{B}_y)} f_X(x)\ dx & X \text{ continuous} \\ \sum_{x \in g^{-1}(\mathcal{B}_y)} f_X(x) & X \text{ discrete} \end{cases}$$

where $g^{-1}(\mathcal{B}_y) = \{x \in \mathbb{R} : g(x) \le y\}$. When g is monotone increasing or decreasing and has differentiable inverse g^{-1}, then

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X\{g^{-1}(y)\}, \qquad y \in \mathbb{R}$$

4.3 Normal Distribution

A random variable X having density

$$f(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \qquad x \in \mathbb{R}, \qquad \mu \in \mathbb{R}, \qquad \sigma > 0 \qquad\qquad \heartsuit$$

is a **normal random variable** with expectation μ and variance σ^2 : we write $X \sim \mathcal{N}(\mu, \sigma^2)$. (The standard deviation of X is $\sqrt{\sigma^2} = \sigma > 0$.)
When $\mu = 0$, $\sigma^2 = 1$, the corresponding random variable Z is **standard normal**, $Z \sim \mathcal{N}(0,1)$, with density

$$\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}, \; z \in \mathbb{R}$$

Then

$$F_Z(x) = \mathrm{P}(Z \le x) = \Phi(x) = \int_{-\infty}^{x} \phi(z)dz = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} e^{-z^2/2} dz$$

The normal distribution is often called the **Gaussian distribution**.

- The density function is centred at μ, which is the most likely value and also the median.
- The standard deviation σ is a measure of the spread of the values around μ :
  - 68% of the probability lies in the interval $\mu \pm \sigma$
  - 95% of the probability lies in the interval $\mu \pm 2\sigma$
  - 99.7% of the probability lies in the interval $\mu \pm 3\sigma$

The density $\phi(z)$, the cumulative distribution function $\Phi(z)$, and the quantiles z_p of $Z \sim \mathcal{N}(0,1)$ satisfy, for all $z \in \mathbb{R}$:

(a) the density is symmetric with respect to $z = 0$, i.e., $\phi(z) = \phi(-z)$

(b) $\mathrm{P}(Z \le z) = \Phi(z) = 1 - \Phi(-z) = 1 - \mathrm{P}(Z \ge z)$

(c) the standard normal quantiles z_p satisfy $z_p = -z_{1-p}$, for all $0 < p < 1$

(d) $z^r \phi(z) \to 0$ when $z \to \pm\infty$, for all $r > 0$. This implies that the moments $\mathrm{E}(Z^r)$ exist for all $r \in \mathbb{N}$

(e) we have

$$\phi'(z) = -z\phi(z), \qquad \phi''(z) = (z^2 - 1)\phi(z), \qquad \phi'''(z) = -(z^3 - 3z)\phi(z), \qquad \ldots$$

This implies that $\mathrm{E}(Z) = 0$, $\mathrm{var}(Z) = 1$, $\mathrm{E}(Z^3) = 0$, etc.

(f) If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

Note that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then we can write $X = \mu + \sigma Z$, where $Z \sim \mathcal{N}(0,1)$.

Normal approximation to the binomial distribution

(de Moivre-Laplace) Let $X_n \sim \mathrm{B}(n,p)$, where $0 < p < 1$, let

$$\mu_n = \mathrm{E}(X_n) = np, \qquad \sigma_n^2 = \mathrm{var}(X_n) = np(1-p)$$

and let $Z \sim \mathcal{N}(0,1)$. When $n \to \infty$,

$$\mathrm{P}\left(\frac{X_n - \mu_n}{\sigma_n} \le z\right) \to \Phi(z),\ z \in \mathbb{R}\ ; \qquad \text{i.e., } \frac{X_n - \mu_n}{\sigma_n} \xrightarrow{D} Z$$

This gives us an approximation of the probability that $X_n \le r$:

$$\mathrm{P}(X_n \le r) = \mathrm{P}\left(\frac{X_n - \mu_n}{\sigma_n} \le \frac{r - \mu_n}{\sigma_n}\right) \doteq \Phi\left(\frac{r - \mu_n}{\sigma_n}\right)$$

which corresponds to $X_n \,\dot{\sim}\, \mathcal{N}\{np, np(1-p)\}$.
In practice the approximation is bad when $\min\{np, np(1-p)\} < 5$.
The Poisson approximation to the binomial distribution is valid for large n and small p. The normal approximation is valid for large n and $\min\{np, n(1-p)\} \ge 5$.

Continuity correction

A better approximation to $\mathrm{P}(X_n \le r)$ is given by replacing r by $r+\frac{1}{2}$; the $\frac{1}{2}$ is called the **continuity correction**. This gives

$$\mathrm{P}(X_n \le r) \doteq \Phi\left(\frac{r + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \qquad \xcancel{\heartsuit}$$

4.4 Q-Q Plots

Which density ?

Uniform variables lie in a finite interval, and give equal probability to each part of the interval

Exponential and **gamma** variables lie in $(0, \infty)$, and are often used to model waiting times and other positive quantities

> the gamma has two parameters and is more flexible, but the exponential is simpler and has some elegant properties

Pareto variables lie in the interval (β, ∞), so are not appropriate for arbitrary positive quantities (which could be smaller than β), but are often used to model financial losses over some threshold β

Normal variables lie in $\mathbb{R}$ and are used to model quantities that arise (or might arise) through averaging of many small effects (e.g., height and weight, which are influenced by many genetic factors), or where measurements are subject to error

Laplace variables lie in $\mathbb{R}$; the Laplace distribution can be used in place of the normal in situations where outliers might be present

5 Several Random Variables

5.1 Basic Notions

Discrete random variables

Let (X, Y) be a discrete random variable : the set

$$D = \{(x, y) \in \mathbb{R}^2 : \mathrm{P}\{(X, Y) = (x, y)\} > 0\}$$

is countable. The **(joint) probability mass function** of (X, Y) is

$$f_{X,Y}(x, y) = \mathrm{P}\{(X, Y) = (x, y)\}, \qquad (x, y) \in \mathbb{R}^2$$

and the **(joint) cumulative distribution function** of (X, Y) is

$$F_{X,Y}(x, y) = \mathrm{P}(X \leq x, Y \leq y), \qquad (x, y) \in \mathbb{R}^2$$

Continuous random variables

The random variable (X, Y) is said to be **(jointly) continuous** if there exists a function $f_{X,Y}(x, y)$, called the **(joint) density** of (X, Y), such that

$$\mathrm{P}\{(X, Y) \in A\} = \int\int_{(u,v)\in A} f_{X,Y}(u, v)\; du\; dv, \qquad \mathcal{A} \subset \mathbb{R}^2$$

By letting $\mathcal{A} = \{(u, v) : u \leq x, v \leq y\}$, we see that the **(joint) cumulative distribution function** of (X, Y) can be written

$$F_{X,Y}(x, y) = \mathrm{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)\; du\; dv, \qquad (x, y) \in \mathbb{R}^2$$

and this implies that

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$$

Exponential families

Let $(X_1, \ldots, X_n)$ be a discrete or continuous random variable with mass/density function of the form

$$f(x_1, \ldots, x_n) = \exp\left\{\sum_{i=1}^{p} s_i(x)\theta_i - \kappa(\theta_1, \ldots, \theta_p) + c(x_1, \ldots, x_n)\right\}, \qquad (x_1, \ldots, x_n) \in D \subset \mathbb{R}^n \qquad \cancel{\heartsuit}$$

where $(\theta_1, \ldots, \theta_p) \in \Theta \subset \mathbb{R}^p$. This is called an **exponential family** distribution - not to be confused with the exponential distribution.

Marginal and conditional distributions

The **marginal probability mass/density function** of X is

$$f_X(x) = \begin{cases} \sum_y f_{X,Y}(x, y) & \text{discrete case} \\ \int_{-\infty}^{\infty} f_{X,Y}(x, y)\; dy & \text{continuous case} \end{cases} \qquad x \in \mathbb{R}$$

The **conditional probability mass/density function** of Y given X is

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \qquad y \in \mathbb{R}$$

provided $f_X(x) > 0$. If (X,Y) is discrete,

$$f_X(x) = \mathrm{P}(X = x), \qquad f_{Y|X}(y \mid x) = \mathrm{P}(Y = y \mid X = x)$$

The conditional density $f_{Y|X}(y \mid x)$ is undefined if $f_X(x) = 0$.
Analogous definitions exist for $f_Y(y)$, $f_{X|Y}(x \mid y)$, and for the conditional cumulative distribution functions $F_{X|Y}(x \mid y)$, $F_{Y|X}(y \mid x)$.

Multivariate random variables

Let $X_1, \dots, X_n$ be rvs defined on the same probability space. Their **joint cumulative distribution function** is

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathrm{P}(X_1 \le x_1, \dots, X_n \le x_n)$$

and their **joint density/mass probability function** is

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \begin{cases} \mathrm{P}(X_1 = x_1, \dots, X_n = x_n) & \text{discrete case} \\ \frac{\partial^n F_{X_1,\dots,X_n}(x_1,\dots,x_n)}{\partial x_1 \cdots \partial x_n} & \text{continuous case} \end{cases}$$

We analogously define the conditional and marginal densities, the cumulative distribution functions, etc., by replacing (X,Y) by $X = X_{\mathcal{A}}$, $Y = X_{\mathcal{B}}$, where $\mathcal{A}, \mathcal{B} \subset \{1,\dots,n\}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$.
Subsequently everything can be generalised to n variables, but for ease of notation we will mostly limit ourselves to the bivariate case.

Multinomial distribution

The random variable $(X_1,\dots,X_k)$ has the **multinomial distribution of denominator m and probabilities $(p_1,\dots,p_k)$** if its mass function is

$$f(x_1,\dots,x_k) = \frac{m!}{x_1! \times \cdots \times x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \qquad x_1,\dots,x_k \in \{0,\dots,m\}, \qquad \sum_{j=1}^k x_j = m$$

where $m \in \mathbb{N}$ and $p_1,\dots,p_k \in [0,1]$, with $p_1 + \cdots + p_k = 1$.
This distribution appears as the distribution of the number of individuals in the categories $\{1,\dots,k\}$ when m independent individuals fall into the classes with probabilities $\{p_1,\dots,p_k\}$. It generalises the binomial distribution to $k > 2$ categories.

Independence

Random variables X, Y defined on the same probability space are **independent** if

$$\mathrm{P}(X \in \mathcal{A}, Y \in \mathcal{B}) = \mathrm{P}(X \in \mathcal{A})\mathrm{P}(Y \in \mathcal{B}), \qquad \mathcal{A}, \mathcal{B} \subset \mathbb{R}$$

By letting $\mathcal{A} = (-\infty, x]$ and $\mathcal{B} = (-\infty, y]$, we find that

$$F_{X,Y}(x,y) = \cdots = F_X(x)F_Y(y), \qquad x, y \in \mathbb{R}$$

implying the equivalent condition

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad x,y \in \mathbb{R} \qquad (*)$$

which will be our criterion of independence. This condition concerns the **functions** $f_{X,Y}(x,y)$, $f_X(x)$, $f_Y(y)$: X,Y are independent iff $(*)$ remains true **for all** $x,y \in \mathbb{R}$.
If X,Y are independent, then for all x such that $f_X(x) > 0$,

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y), \qquad y \in \mathbb{R}$$

Thus knowing that $X = x$ does not affect the density of Y : this is an obvious meaning of "independence". By symmetry $f_{X|Y}(x \mid y) = f_X(x)$ for all y such that $f_Y(y) > 0$.

A **random sample of size** n from a distribution F of density f is a set of n independent random variables which all have a distribution F. Equivalently we say that $X_1, \ldots, X_n$ are **independent and identically distributed (iid)** with distribution F, or with density f, and write $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ or $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$.
By independence, the joint density of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_X(x_j)$$

Mixed distributions

We sometimes encounter distributions with X discrete and Y continuous, or vice versa.

5.2 Dependence

Joint moments

Let X,Y be random variables of density $f_{X,Y}(x,y)$. Then if $\mathrm{E}\{|g(X,Y)|\} < \infty$, we can define the **expectation** of $g(X,Y)$ to be

$$\mathrm{E}\{g(X,Y)\} = \begin{cases} \sum_{x,y} g(x,y) f_{X,Y}(x,y) & \text{discrete case} \\ \iint g(x,y) f_{X,Y}(x,y)\; dx\; dy & \text{continuous case} \end{cases}$$

In particular we define the **joint moments** and the **joint central moments** by

$$\mathrm{E}(X^r Y^s), \qquad \mathrm{E}[\{X - \mathrm{E}(X)\}^r \{Y - \mathrm{E}(Y)\}^s], \qquad r,s \in \mathbb{N}$$

The most important of these is the **covariance** of X and Y,

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

Let X,Y,Z be random variables and $a,b,c,d \in \mathbb{R}$ constants. The covariance satisfies :

$\mathrm{cov}(X,X) = \mathrm{var}(X)$

$\mathrm{cov}(a,X) = 0$

$\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$ **(symmetry)**

$\mathrm{cov}(a + bX + cY, Z) = b\ \mathrm{cov}(X,Z) + c\ \mathrm{cov}(Y,Z)$ **(bilinearity)**

$\operatorname{cov}(a+bX, c+dY) = bd \operatorname{cov}(X,Y)$

$\operatorname{cov}(X,Y)^2 \le \operatorname{var}(X)\operatorname{var}(Y)$ **(Cauchy-Schwarz inequality)**

Independence and covariance

If X and Y are independent and $g(X), h(X)$ are functions whose expectations exist, then

$$\mathrm{E}\{g(X)h(Y)\} = \cdots = \mathrm{E}\{g(X)\}\mathrm{E}\{h(Y)\}$$

By letting $g(X) = X - \mathrm{E}(X)$ and $h(Y) = Y - \mathrm{E}(Y)$, we can see that if X and Y are independent, then

$$\operatorname{cov}(X,Y) = \cdots = 0$$

Thus X, Y indep $\implies \operatorname{cov}(X,Y) = 0$. However, the converse is false.

Linear combinations of random variables

The **average** of random variables $X_1, \dots, X_n$ is $\overline{X} = n^{-1} \sum_{j=1}^{n} X_j$.

Let $X_1, \dots, X_n$ be random variables and $a, b_1, \dots, b_n$ be constants. Then

(a)

$$\mathrm{E}(a + b_1X_1 + \cdots + b_nX_n) = a + \sum_{j=1}^{n} b_j \mathrm{E}(X_j)$$

$$\operatorname{var}(a + b_1X_1 + \cdots + b_nX_n) = \sum_{j=1}^{n} b_j^2 \operatorname{var}(X_j) + \sum_{j \ne k} b_j b_k \operatorname{cov}(X_j, X_k)$$

(b) If $X_1, \dots, X_n$ are independent, then $\operatorname{cov}(X_j, X_k) = 0$, $j \ne k$, so

$$\operatorname{var}(a + b_1X_1 + \cdots + b_nX_n) = \sum_{j=1}^{n} b_j^2 \operatorname{var}(X_j)$$

(c) If $X_1, \dots, X_n$ are independent and all have mean μ and variance σ^2, then

$$\mathrm{E}(\overline{X}) = \mu, \qquad \operatorname{var}(\overline{X}) = \frac{\sigma^2}{n}$$

Correlation

The covariance depends on the units of measurement, so we often use the following dimensionless measure of dependence.

The **correlation** of X, Y is

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\{\operatorname{var}(X)\operatorname{var}(Y)\}^{1/2}}$$

This measures the linear dependence between X and Y.

Let X, Y be random variables having correlation $\rho = \operatorname{corr}(X,Y)$. Then :

(a) $-1 \le \rho \le 1$

(b) if $\rho = \pm 1$, then there exist $a, b, c \in \mathbb{R}$ such that

$$aX + bY + c = 0$$

with probability 1 (X and Y are then linearly dependent)

(c) if X, Y are independent, then $\mathrm{corr}(X, Y) = 0$

(d) the effect of the transformation

$$(X, Y) \mapsto (a + bX, c + dY)$$

is

$$\mathrm{corr}(X, Y) \mapsto \mathrm{sign}(bd)\mathrm{corr}(X, Y)$$

Note that :

correlation measures **linear** dependence

we can have strong nonlinear dependence, but correlation zero

correlation can be strong but **specious**

Correlation $\neq$ causation

Two variables can be very correlated without one **causing** changes in the other.

Conditional expectation

Let $g(X, Y)$ be a function of a random vector (X, Y). Its **conditional expectation** given $X = x$ is

$$\mathrm{E}\{g(X, Y) \mid X = x\} = \begin{cases} \sum_y g(x, y) f_{Y|X}(y \mid x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} g(x, y) f_{Y|X}(y \mid x)\, dy & \text{in the continuous case} \end{cases}$$

on the condition that $f_X(x) > 0$ and $\mathrm{E}\{|g(X, Y)| \mid X = x\} < \infty$. Note that the conditional expectation $\mathrm{E}\{g(X, Y) \mid X = x\}$ is a function of x.

Expectation and conditioning

Sometimes it is easier to calculate $\mathrm{E}\{g(X, Y)\}$ in stages.
If the required expectations exist, then

$$\mathrm{E}\{g(X, Y)\} = \mathrm{E}_X[\mathrm{E}_Y\{g(X, Y) \mid X = x\}]$$

$$\mathrm{var}\{g(X, Y)\} = \mathrm{E}_X[\mathrm{var}_Y\{g(X, Y) \mid X = x\}] + \mathrm{var}_X[\mathrm{E}_Y\{g(X, Y) \mid X = x\}]$$

where E_X and var_X represent expectation and variance according to the distribution of X.

5.3 Generating Functions

We define the **moment-generating function** of a random variable X by

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

for $t \in \mathbb{R}$ such that $M_X(t) < \infty$.

- $M_X(t)$ is also called the **Laplace transform** of $f_X(x)$
- The MGF is useful as a summary of all the properties of X, we can write

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}\left(\sum_{r=0}^{\infty} \frac{t^r X^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mathrm{E}(X^r)$$

  from which we can obtain all the moments $\mathrm{E}(X^r)$ by differentiation

Important theorems I

(1) If $M(t)$ is the MGF of a random variable X, then

$M_X(0) = 1$

$M_{a+bX}(t) = e^{at} M_X(bt)$

$\mathrm{E}(X^r) = \frac{\partial^r M_X(t)}{\partial t^r}\Big|_{t=0}$

$\mathrm{E}(X) = M'_X(0)$

$\mathrm{var}(X) = M''_X(0) - M'_X(0)^2$

(2) There exists an injection between the cumulative distribution functions $F_X(x)$ and the moment-generating functions $M_X(t)$.
This theorem is very useful, as it says that if we recognise a MGF, we know to which distribution it corresponds.

Linear combinations

Let $a, b_1, \dots, b_n \in \mathbb{R}$ and $X_1, \dots, X_n$ be independent rv's whose MGFs exist. Then $Y = a + b_1 X_1 + \cdots + b_n X_n$ has MGF

$$M_Y(t) = \cdots = e^{ta} \prod_{j=1}^{n} M_{X_j}(tb_j)$$

In particular, if $X_1, \dots, X_n$ is a random sample, then $S = X_1 + \cdots + X_n$ has

$$M_S(t) = M_X(t)^n$$

Let $X_1, \dots, X_n$ be independent with $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$.

$$Y = a + b_1 X_1 + \cdots + b_n X_n \sim \mathcal{N}(a + b_1\mu_1 + \cdots + b_n\mu_n, b_1^2\sigma_1^2 + \cdots + b_n^2\sigma_n^2)$$

thus a linear combination of normal rv's is normal.

Important theorems II

($\xrightarrow{D}$) Let $\{X_n\}, X$ be random variables whose cumulative distribution functions are $\{F_n\}, F$. Then we say that the random variables $\{X_n\}$ **converge in distribution** to X, if, for all $x \in \mathbb{R}$ where F is continuous,

$$F_n(x) \to F(x), \qquad n \to \infty$$

We then write $X_n \xrightarrow{D} X$.
(3, Continuity) Let $\{X_n\}, X$ be random variables with distribution functions $\{F_n\}, F$, whose MGFs $M_n(t), M(t)$ exist for $0 \leq |t| < b$. Then if $M_n(t) \to M(t)$ for $|t| \leq a < b$ when $n \to \infty$, then $X_n \xrightarrow{D} X$, i.e., $F_n(x) \to F(x)$ at

each $x \in \mathbb{R}$ where F is continuous.

Mean vector and covariance matrix

Let $X = (X_1, \dots, X_p)^\mathrm{T}$ be a $p \times 1$ vector of random variables. Then

$$\mathrm{E}(X)_{p\times 1} = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{pmatrix}$$

$$\mathrm{var}(X)_{p\times p} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_p) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_1, X_p) & \mathrm{cov}(X_2, X_p) & \cdots & \mathrm{var}(X_p) \end{pmatrix}$$

are called the **expectation (mean vector)** and the **(co)-variance matrix** of X.
The matrix $\mathrm{var}(X)$ is positive semi-definite, since

$$\mathrm{var}\left(\sum_{j=1}^{p} a_j X_j\right) = a^\mathrm{T} \mathrm{var}(X) a \geq 0$$

for all vectors $a = (a_1, \dots, a_p)^\mathrm{T} \in \mathbb{R}^p$.

Multivariate case

The **moment-generating function (MGF)** of a random vector $X_{p\times 1} = (X_1, \dots, X_p)^\mathrm{T}$ is

$$M_X(t) = \mathrm{E}(e^{t^\mathrm{T} X}) = \mathrm{E}(e^{\sum_{r=1}^{p} t_r X_r}), \qquad t = (t_1, \dots, t_p) \in \mathcal{T}$$

where $\mathcal{T} = \{t \in \mathbb{R}^p : M_X(t) < \infty\}$. Let the rth and (r, s)th elements of the **mean vector** $\mathrm{E}(X)_{p\times 1}$ and of the **covariance matrix** $\mathrm{var}(X)_{p\times p}$ be the quantities $\mathrm{E}(X_r)$ and $\mathrm{cov}(X_r, X_s)$.
The MGF has the following properties :

- $0 \in \mathcal{T}$, thus $M_X(0) = 1$
- we have

$$\mathrm{E}(X)_{p\times 1} = M'_X(0) = \left.\frac{\partial M_X(t)}{\partial t}\right|_{t=0}, \qquad \mathrm{var}(X)_{p\times p} = \left.\frac{\partial^2 M_X(t)}{\partial t \partial t^\mathrm{T}}\right|_{t=0} - M'_X(0) M'_X(0)^\mathrm{T}$$

- if $\mathcal{A}, \mathcal{B} \subset \{1, \dots, p\}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$, and we write $X_\mathcal{A}$ for the subvector of X containing $\{X_j : j \in \mathcal{A}\}$, etc., then $X_\mathcal{A}$ and $X_\mathcal{B}$ are independent iff

$$M_X(t) = \mathrm{E}(e^{t_\mathcal{A}^\mathrm{T} X_\mathcal{A} + t_\mathcal{B}^\mathrm{T} X_\mathcal{B}}), \qquad t \in \mathcal{T}$$

- there is an injective mapping between MGFs and probability distributions

5.4 Multivariate Normal Distribution

The random vector $X = (X_1, \dots, X_p)^{\mathrm{T}}$ has a **multivariate normal distribution** if there exist a $p \times 1$ vector $\mu = (\mu_1, \dots, \mu_p)^{\mathrm{T}} \in \mathbb{R}^p$ and a $p \times p$ symmetric matrix Ω with elements ω_{jk} such that

$$u^{\mathrm{T}}X \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u), \qquad u \in \mathbb{R}^p$$

then we write $X \sim \mathcal{N}_p(\mu, \Omega)$.

- Since $\mathrm{var}(u^{\mathrm{T}}X) = u^{\mathrm{T}}\Omega u \geq 0$ for any $u \in \mathbb{R}^p$, Ω must be positive semi-definite
- This definition allows degenerate distributions, for which there exists a u such that $\mathrm{var}(u^{\mathrm{T}}X) = 0$. This gives mathematically clean results but can be avoided in applications by reformulating the problem to avoid degeneracy, effectively working in a space of dimension $m < p$

(a) We have

$$\mathrm{E}(X_j) = \mu_j, \qquad \mathrm{var}(X_j) = \omega_{jj}, \qquad \mathrm{cov}(X_j, X_k) = \omega_{jk}, \qquad j \neq k$$

so μ and Ω are called the **mean vector** and **covariance matrix** of X

(b) The moment-generating function of X is $M_X(u) = \exp\left(u^{\mathrm{T}}\mu + \frac{1}{2}u^{\mathrm{T}}\Omega u\right)$, for $u \in \mathbb{R}^p$ $\xcancel{\heartsuit}$

(c) If $\mathcal{A}, \mathcal{B} \subset \{1, \dots, p\}$, and $\mathcal{A} \cap \mathcal{B} = \emptyset$ then

$$X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \iff \Omega_{\mathcal{A},\mathcal{B}} = 0 \qquad \heartsuit$$

(d) If $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $X_{n\times 1} = (X_1, \dots, X_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$

(e) An affine transform of a normal variable is normal :

$$a_{r\times 1} + B_{r\times p}X \sim \mathcal{N}_r(a + B\mu, B\Omega B^{\mathrm{T}}) \qquad \xcancel{\heartsuit}$$

The random vector $X \sim \mathcal{N}_p(\mu, \Omega)$ has a density function on $\mathbb{R}^p$ if and only if Ω is positive definite, i.e., Ω has rank p. If so, the density function is

$$f(x; \mu, \Omega) = \frac{1}{(2\pi)^{p/2}|\Omega|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)\right\}, \qquad x \in \mathbb{R}^p \qquad \xcancel{\heartsuit}$$

If not, X is a linear combination of variables that have a density function on $\mathbb{R}^m$, where $m < p$ is the rank of Ω.

Marginal and conditional distributions

Let $X \sim \mathcal{N}_p(\mu_{p\times 1}, \Omega_{p\times p})$, where $|\Omega| > 0$, and let $\mathcal{A}, \mathcal{B} \subset \{1, \dots, p\}$ with $|\mathcal{A}| = q < p$, $|\mathcal{B}| = r < p$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$. Let $\mu_{\mathcal{A}}$, $\Omega_{\mathcal{A}}$ and $\Omega_{\mathcal{AB}}$ be respectively the $q \times 1$ subvector of μ, $q \times q$ and $q \times r$ submatrices of Ω conformable with $\mathcal{A}$, $\mathcal{A} \times \mathcal{A}$ and $\mathcal{A} \times \mathcal{B}$. Then :

(a) the marginal distribution of $X_{\mathcal{A}}$ is normal

$$X_{\mathcal{A}} \sim \mathcal{N}_q(\mu_{\mathcal{A}}, \Omega_{\mathcal{A}})$$

(b) the conditional distribution of $X_{\mathcal{A}}$ given $X_{\mathcal{B}} = x_{\mathcal{B}}$ is normal

$$X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}} \sim \mathcal{N}_q\{\mu_{\mathcal{A}} + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}), \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}}\} \qquad \xcancel{\heartsuit}$$

This has two important implications :

- (a) implies that any subvector of X also has a multivariate normal distribution
- (b) implies that two components of $X_{\mathcal{A}}$ are conditionally independent given $X_{\mathcal{B}}$ if and only if the corresponding off-diagonal element of $\Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}}$ equals zero

Regression to the mean

The slope of the line < 1 :

- the children of tall parents are smaller than them, on average, and the children of small parents are larger than them, on average
- someone with an above-average mark on a midterm test will tend to do worse in the final, on average

5.5 Transformations

Transformation of random variables

We often want to calculate the distributions of random variables based on other random variables.

- Let $Y = g(X)$, where the function g is known. We want to obtain F_Y and f_Y from F_X and f_X
- Let $g : \mathbb{R} \mapsto \mathbb{R}$, $\mathcal{B} \subset \mathbb{R}$, and $g^{-1}(\mathcal{B}) \subset \mathbb{R}$ be the set for which $g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$. Then

  $$\mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\}$$

  since $X \in g^{-1}(\mathcal{B})$ iff $g(X) = Y \in g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$.
- To find $F_Y(y)$, we take $\mathcal{B}_y = -(\infty, y]$, giving

  $$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}\{g(X) \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\}$$
- If the function g is monotonic increasing with (monotonic increasing) inverse g^{-1}, and if X and Y are continuous RV, then

  $$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X\{g^{-1}(y)\}}{dy} = f_X\{g^{-1}(y)\} \times \left|\frac{dg^{-1}(y)}{dy}\right|$$

  where the $|\cdot|$ ensures that the same formula holds with monotonic decreasing g

X bivariate

We calculate $\mathrm{P}(Y \in \mathcal{B})$, with $Y \in \mathbb{R}^d$ a function of $X \in \mathbb{R}^2$ and

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix} = \begin{pmatrix} g_1(X_1, X_2) \\ \vdots \\ g_d(X_1, X_2) \end{pmatrix} = g(X)$$

Let $g : \mathbb{R}^2 \mapsto \mathbb{R}^d$ be a known function, $\mathcal{B} \subset \mathbb{R}^d$, and $g^{-1}(\mathcal{B}) \subset \mathbb{R}^2$ be the set for which $g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$. Then

$$\mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\}$$

Transformations of joint continuous densities

Let $X = (X_1, X_2) \in \mathbb{R}^2$ be a continuous random variable, and let $Y = (Y_1, Y_2)$ with $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$, where :

(a) the system of equations $y_1 = g_1(x_1, x_2), y_2 = g_2(x_1, x_2)$ can be solved for all (y_1, y_2), giving the solutions $x_1 = h_1(y_1, y_2), x_2 = h_2(y_1, y_2)$

(b) g_1 and g_2 are continuously differentiable and have Jacobian

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} = \left(\frac{\partial g_1}{\partial x_1} \frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2} \frac{\partial g_2}{\partial x_1} \right)$$

which is positive if $f_{X_1,X_2}(x_1, x_2) > 0$. Then

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2) \times |J(x_1, x_2)|^{-1} \Big|_{x_1=h_1(y_1,y_2), x_2=h_2(y_1,y_2)}$$

Sums of independent variables

If X, Y are independent random variables, then the PDF or PMF of their sum $S = X + Y$ is the convolution $f_X * f_Y$ of the PDFs or PMFs f_X, f_Y :

$$f_S(s) = f_X * f_Y(s) = \begin{cases} \int_{-\infty}^{\infty} f_X(x) f_Y(s - x)\ dx & X, Y \text{ continuous} \\ \sum_x f_X(x) f_Y(s - x) & X, Y \text{ discrete} \end{cases}$$

Multivariate case

The transformations of joint continuous densities extend to random vectors with continuous density, $Y = g(X) \in \mathbb{R}^n$, where $X \in \mathbb{R}^n$ is a continuous variable :

$$(X_1, \ldots, X_n) \mapsto (Y_1 = g_1(X_1, \ldots, X_n), \ldots, Y_n = g_n(X_1, \ldots, X_n))$$

If the inverse transformation h exists, and has Jacobian

$$J(x_1, \ldots, x_n) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix}$$

then

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\ |J(x_1, \ldots, x_n)|^{-1}$$

evaluated at $x_1 = h_1(y_1, \ldots, y_n), \ldots, x_n = h_n(y_1, \ldots, y_n)$.

Convolution and sums of random variables

If $X_1, \ldots, X_n$ are independent random variables, then the PDF or PMF of $S = X_1 + \cdots + X_n$ is the convolution

$$f_S(s) = f_{X_1} * \cdots * f_{X_n}(s)$$

In fact it is easier to use the MGFs for convolutions, if possible.

5.6 Order Statistics

The **order statistics** of the rv's $X_1, \ldots, X_n$ are the ordered values

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$$

If the $X_1, \ldots, X_n$ are continuous and independent, then no two of the X_j can be equal, i.e.

$$X_{(1)} < X_{(2)} < \cdots < X_{(n-1)} < X_{(n)}$$

In particular, the **minimum** is $X_{(1)}$, the **maximum** is $X_{(n)}$, and the **median** is

$$X_{(m+1)} \qquad (n = 2m+1,\ \text{odd}), \qquad \frac{1}{2}(X_{(m)} + X_{(m+1)})(n = 2m,\ \text{even})$$

The median is the central value of $X_1, \ldots X_n$.
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, then :

$F_{X_{(n)}}(x) = \mathrm{P}(X_{(n)} \leq x) = F(x)^n$

$\mathrm{P}(X_{(1)} \leq x) = 1 - \{1 - F(x)\}^n$

If F corresponds to a density f, then $f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} F(x)^{r-1} f(x) \{1 - F(x)\}^{n-r}, \qquad r = 1, \ldots, n$

6 Approximation and Convergence

6.1 Inequalities

If X is a random variable, $a > 0$ a constant, h a non-negative function and g a convex function, then

$\mathrm{P}\{h(X) \geq a\} \leq \mathrm{E}\{h(X)\}/a$ (basic inequality) ♡

$\mathrm{P}(|X| \geq a) \leq \mathrm{E}(|X|)/a$ (Markov's inequality) ♡

$\mathrm{P}(|X| \geq a) \leq \mathrm{E}(X^2)/a^2$ (Chebyshov's inequality) ♡

$\mathrm{E}\{g(X)\} \geq g\{\mathrm{E}(X)\}$ (Jensen's inequality) ♡

On replacing X by $X - \mathrm{E}(X)$, Chebyshov's inequality gives

$$\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\} \leq \mathrm{var}(X)/a^2 \qquad \heartsuit$$

$Y_1, \ldots, Y_n$ i.i.d., define $\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$

$$\mathrm{var}(\overline{Y}) = \mathrm{cov}(\overline{Y}, \overline{Y}) = \mathrm{cov}\left(\frac{Y_1 + \cdots + Y_n}{n}, \frac{Y_1 + \cdots + Y_n}{n}\right)$$

$$= \frac{1}{n^2} \sum_{i,j=1}^{n} \mathrm{cov}(Y_i, Y_j) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{cov}(Y_i, Y_i) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}(Y_i) = \frac{\mathrm{var}(Y_1)}{n}$$

$$\Rightarrow \mathrm{var}(\overline{Y}) = \frac{\mathrm{var}(Y_1)}{n}$$

Hoeffding's inequality

Let $Z_1, \ldots, Z_n$ be independent random variables such that $\mathrm{E}(Z_i) = 0$ and $a_i \leq Z_i \leq b_i$ for constants $a_i < b_i$. If $\varepsilon > 0$, then for all $t > 0$

$$\mathrm{P}\left(\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq e^{-t\varepsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8} \qquad \xcancel{\heartsuit}$$

This inequality is much more useful than the others for finding powerful bounds in practical situations.

6.2 Convergence

(Deterministic convergence) If $x_1, x_2, \ldots, x$ are real numbers, then $x_n \to x$ iff for all $\varepsilon > 0$, there exists N_ε such that $|x_n - x| < \varepsilon$ for all $n > N_\varepsilon$.
Probabilistic convergence is more complicated... We could hope that (for example) $X_n \to X$ if either

$$\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x), \qquad x \in \mathbb{R}$$

or

$$\mathrm{E}(X_n) \to \mathrm{E}(X)$$

when $n \to \infty$.

Modes of convergence of random variables

Let $X, X_1, X_2, \ldots$ be random variables with cumulative distribution function $F, F_1, F_2, \ldots$. Then

(a) X_n converges to X **almost surely**, $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathrm{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1 \qquad \heartsuit$$

(b) X_n converges to X **in mean square**, $X_n \xrightarrow{2} X$, if

$$\lim_{n\to\infty} \mathrm{E}\{(X_n - X)^2\} = 0, \qquad \text{where } \mathrm{E}(X_n^2), \mathrm{E}(X^2) < \infty \qquad \heartsuit$$

(c) X_n converges to X **in probability**, $X_n \xrightarrow{P} X$, if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \varepsilon) = 0 \qquad \heartsuit$$

(d) X_n converges to X **in distribution**, $X_n \xrightarrow{D} X$, if

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ at each point } x \text{ where } F(x) \text{ is continuous} \qquad \heartsuit$$

To understand $X_n \xrightarrow{\text{a.s.}} X$ better :

- All the variables $\{X_n\}, X$ must be defined on the same probability space, $(\Omega, \mathcal{F}, \mathrm{P})$. It is not trivial to construct this space (we need 'Kolmogorov's extension theorem')
- Then to each $\omega \in \Omega$ corresponds a sequence

  $$X_1(\omega), X_2(\omega), \ldots, X_n(\omega), \ldots$$

  which will converge, or not, as a sequence of real numbers
- If $X_n \xrightarrow{\text{a.s.}} X$, then

  $$\mathrm{P}\Big(\Big\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1$$

  the set of values of ω for which $X_n(\omega) \not\to X(\omega)$ has probability 0

Relations between modes of convergence

- If $X_n \xrightarrow{\text{a.s.}} X$, $X_n \xrightarrow{2} X$ or $X_n \xrightarrow{P} X$, then $X_1, X_2, \ldots, X$ must all be defined with respect to only one probability space. This is not the case for $X_n \xrightarrow{D} X$, which only concerns the probabilities. This last is thus weaker that the others
- These modes of convergence are related to one another in the following way :

  $$\begin{array}{ll} X_n \xrightarrow{\text{a.s.}} X \Rightarrow & \\ & X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X \\ X_n \xrightarrow{2} X \Rightarrow & \end{array}$$

  All other implications are in general false
- The most important modes of convergence in this course are $\xrightarrow{P}$ and $\xrightarrow{D}$, since we often wish to approximate probabilities, and $\xrightarrow{D}$ gives us a way to do so

Continuity theorem

Let $\{X_n\}, X$ be random variables with cumulative distribution functions $\{F_n\}, F$, whose MGFs $M_n(t), M(t)$ exist for $0 \leq |t| < b$. If there exists a $0 < a < b$ such that $M_n(t) \to M(t)$ for $|t| \leq a$ when $n \to \infty$, then $X_n \xrightarrow{D} X$, that is to say, $F_n(x) \to F(x)$ at each $x \in \mathbb{R}$ where F is continuous.

- We could replace $M_n(t)$ and $M(t)$ by the cumulant-generating functions $K_n(t) = \log M_n(t)$ and $K(t) = \log M(t)$
- We established the law of small numbers by using this result

Combinations of convergent sequences

Let x_0, y_0 be constants, $X, Y, \{X_n\}, \{Y_n\}$ random variables, and h a function continuous at x_0. Then

$$X_n \xrightarrow{D} x_0 \Rightarrow X_n \xrightarrow{P} x_0$$

$$X_n \xrightarrow{P} x_0 \Rightarrow h(X_n) \xrightarrow{P} h(x_0)$$

$$X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{P} y_0 \Rightarrow X_n + Y_n \xrightarrow{D} X + y_0,\ X_n Y_n \xrightarrow{D} X y_0$$

The third line is known as **Slutsky's lemma**. It is very useful in statistical applications.

Limits for maxima

- In applications, we often have to take into account the greatest or the smallest random variables considered
- A system of n composants can break down when any composant of the system becomes faulty. What is the distribution of the failure time ?
- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, and $M_n = \max\{X_1, \ldots, X_n\}$. Then

$$\mathrm{P}(M_n \leq x) = \mathrm{P}(X_1 \leq x, \ldots, X_n \leq x) = F(x)^n \to \begin{cases} 0 & F(x) < 1 \\ 1 & F(x) = 1 \end{cases}$$

- Hence M_n must be renormalised to get a non-degenerate limit distribution. Let $\{a_n\} > 0$ and $\{b_n\}$ be sequences of constants, and consider the convergence in distribution of

$$Y_n = \frac{M_n - b_n}{a_n}$$

  where a_n, b_n are chosen so that a non-degenerate limit distribution for Y_n exists

Fisher-Tippett theorem

Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, where F is a continuous cumulative distribution function. Let $M_n = \max\{X_1, \ldots, X_n\}$, and suppose that the sequences of constants $\{a_n\} > 0$ and $\{b_n\}$ can be chosen so that $Y_n = (M_n - b_n)/a_n \xrightarrow{D} Y$, where Y has a non-degenerate limit distribution $H(y)$ when $n \to \infty$. Then H must be the **generalised extreme-value (GEV) distribution**

$$H(y) = \begin{cases} \exp\left[-\{1 + \xi(y - \eta)/\tau\}_+^{-1/\xi} \right] & \xi \neq 0 \\ \exp\left[-\exp\{-(y - \eta)/\tau\} \right] & \xi = 0 \end{cases}$$

where $u_+ = \max(u, 0)$, and $\eta, \xi \in \mathbb{R}$, $\tau > 0$.

6.3 Laws of Large Numbers

The first part of our limit results concern the behaviour of averages of independent random variables.

Weak law of large numbers

Let $X_1, X_2, \dots$ be a sequence of independent identically distributed random variables with finite expectation μ, and write their average as

$$\overline{X} = n^{-1}(X_1 + \cdots + X_n)$$

Then $\overline{X} \xrightarrow{P} \mu$; i.e., for all $\varepsilon > 0$

$$\mathrm{P}(|\overline{X} - \mu| > \varepsilon) \to 0, \qquad n \to \infty$$

- Thus, under mild conditions, the averages of samples of important size converge towards the expectation of the distribution from which the sample is taken
- If the X_i are independent Bernoulli trials, we return to our primitive notion of probability as a limit of relative frequencies. The circle is complete
- When $\mathrm{E}(X_i)$ does not exist, the possibility of huge values of X_i implies that $\overline{X}$ cannot converge

Strong law of large numbers

Under the conditions of the weak law of large numbers, $\overline{X} \xrightarrow{\text{a.s.}} \mu$:

$$\mathrm{P}\Big(\lim_{n\to\infty} \overline{X} = \mu\Big) = 1$$

- This is stronger in the sense that for all $\varepsilon > 0$, the weak law allows the event $|\overline{X} - \mu| > \varepsilon$ to occur an infinite number of times, though with smaller and smaller probabilities. The strong law excludes this possibility : it implies that the event $|\overline{X} - \mu| > \varepsilon$ can only occur a finite number of times

6.4 Central Limit Theorem

Standardisation of an average

The law of large numbers shows us that the average $\overline{X}$ approaches μ when $n \to \infty$ ($\overline{X} - \mu \xrightarrow{\text{a.s.}} 0$). If $\mathrm{var}(X_j) < \infty$, then the linear combinations of random variables tells us that

$$\mathrm{E}(\overline{X}) = \mu, \qquad \mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}$$

so, for all n, the difference between $\overline{X}$ and its expectation relative to its standard deviation

$$Z_n = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\mathrm{var}(\overline{X})^{1/2}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma}$$

has expected value zero and unit variance.

Central limit theorem (CLT)

Let $X_1, X_2, \dots$ be independent random variables with expectation μ and variance $0 < \sigma^2 < \infty$. Then

$$Z_n = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} Z, \qquad n \to \infty$$

where $Z \sim \mathcal{N}(0,1)$. ("$\overline{X} \underset{n\to\infty}{\simeq} \mu + \frac{\sigma}{\sqrt{n}} Z,\ Z \sim \mathcal{N}(0,1)$")

Thus

$$\mathrm{P}\left\{\frac{n^{1/2}(\overline{X}-\mu)}{\sigma} \le z\right\} \doteq \mathrm{P}(Z \le z) = \Phi(z)$$

for large n.

The CLT is used to approximate probabilities involving the sums of independent random variables. Under the previous conditions, we have

$$\mathrm{E}\left(\sum_{j=1}^{n} X_j\right) = n\mu, \qquad \mathrm{var}\left(\sum_{j=1}^{n} X_j\right) = n\sigma^2$$

so

$$\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\overline{X}-\mu)}{\sqrt{n\sigma^2}} = \frac{n^{1/2}(\overline{X}-\mu)}{\sigma} = Z_n$$

can be approximated using a normal variable :

$$\mathrm{P}\left(\sum_{j=1}^{n} X_j \le x\right) = \mathrm{P}\left\{\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} \le \frac{x-n\mu}{(n\sigma^2)^{1/2}}\right\} \doteq \Phi\left\{\frac{x-n\mu}{(n\sigma^2)^{1/2}}\right\}$$

The accuracy of the approximation depends on the underlying variables : it is (of course) exact for normal X_j, works better if the X_j are symmetrically distributed (e.g., uniform), and typically is adequate if $n > 25$ or so.

6.5 Delta Method

Sample quantiles

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, and $0 < p < 1$. Then the p **sample quantile** of $X_1, \ldots, X_n$ is the rth order statistic $X_{(r)}$, where $r = \lceil np \rceil$.

(Asymptotic distribution of order statistics) Let $0 < p < 1$, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, and $x_p = F^{-1}(p)$. Then if $f(x_p) > 0$

$$\frac{X_{(\lceil np \rceil)} - x_p}{[p(1-p)/\{nf(x_p)^2\}]^{1/2}} \xrightarrow{D} \mathcal{N}(0,1), \qquad n \to \infty$$

This implies that

$$X_{(\lceil np \rceil)} \mathrel{\dot\sim} \mathcal{N}\left\{x_p, \frac{p(1-p)}{nf(x_p)^2}\right\}$$

To get an intuition, note that $X_{(r)} \le x$ iff $T = \sum I(X_j \le x) \ge r$, and apply the CLT to T.

7 Exploratory Statistics

7.1 Introduction

7.2 Data

7.3 Graphs

7.4 Numerical Summaries

7.5 Boxplot

7.6 Choice of a Model

8 Statistical Inference

8.1 Introduction

Definitions

(Notation) We will use y and Y to represent the data $y_1, \dots, y_n \in \mathbb{R}$ and $Y_1, \dots, Y_n$.
A **statistical model** is a probability distribution $f(y)$ chosen or constructed to learn from observed data y or from potential data Y.

- If $f(y) = f(y; \theta)$ is determined by a parameter θ of finite dimension, it is a **parametric model**, and otherwise it is a **nonparametric model**

- A perfectly known model is called **simple**, otherwise it is **composite**

A **statistic** $T = t(Y)$ is a known function of the data Y.
The **sampling distribution** of a statistic $T = t(Y)$ is its distribution when $Y \sim f(y)$.
A **random sample** is a set of independent and identically distributed random variables $Y_1, \dots, Y_n$, or their realisations $y_1, \dots, y_n$.

8.2 Point Estimation

Statistical models

Note that :

- a statistic T is a function of the random variables $Y_1, \dots, Y_n$, so T is itself a random variable

- the **sampling distribution of** T depends on the distribution of the Y_j

- if we cannot deduce the exact distribution of T from that of the Y_j, we must sometimes make do with knowing $\mathrm{E}(T)$ and $\mathrm{var}(T)$, which give partial information on the distribution of T, and thus may allow us to approximate the distribution of T (often using the central limit theorem)

Estimation methods

Examples of common methods for estimating the parameters of models are :

- **method of moments** (simple, can be inefficient)

- **maximum likelihood estimation** (general, optimal in many parametric models)

- **M-estimation** (even more general, can be robust, but loses efficiency compared to maximum likelihood)

Method of moments

- The **method of moments estimate** of a parameter θ is the value $\tilde{\theta}$ that matches the theoretical and empirical moments

- For a model with p unknown parameters, we set the theoretical moments of the population equal to the empirical moments of the sample $y_1, \dots, y_n$, and solve the resulting equations, i.e.,

$$\mathrm{E}(Y^r) = \int y^r f(y; \theta)\, dy = \frac{1}{n} \sum_{j=1}^{n} y_j^r, \qquad r = 1, \dots, p$$

- We thus need as many (finite !) moments of the underlying model as there are unknown parameters

- We may have more than one choice of moments to use, so in principle the estimate is not unique, but in practice we usually use the first r moments, because they give the most stable estimates

Maximum likelihood estimation

If $y_1, \dots, y_n$ is a random sample from the density or mass $f(y;\theta)$, then the **likelihood** for θ is

$$L(\theta) = f(y_1, \dots, y_n; \theta) = f(y_1;\theta) \times f(y_2;\theta) \times \cdots \times f(y_n;\theta)$$

The data are treated as fixed, and the likelihood $L(\theta)$ is regarded as a function of θ.
The **maximum likelihood estimate (MLE)** $\widehat{\theta}$ of a parameter θ is the value that gives the observed data the highest likelihood. Thus

$$L(\widehat{\theta}) \geq L(\theta) \text{ for each } \theta$$

We simplify the calculations by maximising $\ell(\theta) = \log L(\theta)$ rather than $L(\theta)$. The approach is :

- calculate the log-likelihood $\ell(\theta)$ (and plot it if possible)

- find the value $\widehat{\theta}$ maximising $\ell(\theta)$, which often satisfies $\frac{d\ell(\widehat{\theta})}{d\theta} = 0$

- check that $\widehat{\theta}$ gives a maximum, often by checking that $\frac{d^2\ell(\widehat{\theta})}{d\theta^2} < 0$

M-estimation

- This generalises maximum likelihood estimation. We maximise a function of the form

$$\rho(\theta; Y) = \sum_{j=1}^{n} \rho(\theta; Y_j)$$

  where $\rho(\theta; y)$ is (if possible) concave as a function of θ for all y. Equivalently we minimise $-\rho(\theta; Y)$

- We choose the function ρ to give estimators with suitable properties, such as small variance or robustness to outliers

- Taking $\rho(\theta; y) = \log f(y;\theta)$ gives the maximum likelihood estimator

Bias

The **bias** of the estimator $\widehat{\theta}$ of θ is

$$b(\widehat{\theta}) = \mathrm{E}(\widehat{\theta}) - \theta$$

- Interpretation of the bias :

  - if $b(\widehat{\theta}) < 0$ for all θ, then on average $\widehat{\theta}$ underestimates θ
  - if $b(\widehat{\theta}) > 0$ for all θ, then on average $\widehat{\theta}$ overestimates θ
  - if $b(\widehat{\theta}) = 0$ for all θ, then $\widehat{\theta}$ is said to be **unbiased**

- If $b(\widehat{\theta}) \approx 0$, then $\widehat{\theta}$ is 'in the right place' on average

Mean square error

The **mean square error (MSE)** of the estimator $\widehat{\theta}$ of θ is

$$\text{MSE}(\widehat{\theta}) = \text{E}\{(\widehat{\theta} - \theta)^2\} = \cdots = \text{var}(\widehat{\theta}) + b(\widehat{\theta})^2$$

This is the average squared distance between $\widehat{\theta}$ and its target value θ.
Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators of the same parameter θ. Then

$$\text{MSE}(\widehat{\theta}_1) = \text{var}(\widehat{\theta}_1) + b_1(\widehat{\theta})^2 = \text{var}(\widehat{\theta}_1)$$

$$\text{MSE}(\widehat{\theta}_2) = \text{var}(\widehat{\theta}_2) + b_2(\widehat{\theta})^2 = \text{var}(\widehat{\theta}_2)$$

and we say that $\widehat{\theta}_1$ is **more efficient** than $\widehat{\theta}_2$ if

$$\text{var}(\widehat{\theta}_1) \leq \text{var}(\widehat{\theta}_2)$$

If so, then we prefer $\widehat{\theta}_1$ to $\widehat{\theta}_2$.

8.3 Interval Estimation

Pivots

A key element of statistical thinking is to assess uncertainty of results and conclusions.
Let $t = 1$ be an estimate of an unknown parameter θ based on a sample of size n :

- if $n = 10^5$ we are much more sure that $\theta \approx t$ than if $n = 10$
- as well as t we would thus like to give an interval which will be wider when $n = 10$ than when $n = 10^5$, to make the uncertainty of t explicit

We suppose that we have

- **data** $y_1, \dots, y_n$, which are regarded as a realisation of a
- **random sample** $Y_1, \dots, Y_n$ drawn from a
- **statistical model** $f(y; \theta)$ whose unknown
- **parameter** θ is estimated by the
- **estimate** $t = t(y_1, \dots, y_n)$, which is regarded as a realisation of the
- **estimator** $T = t(Y_1, \dots, Y_n)$

Let $Y = (Y_1, \dots, Y_n)$ be sampled from a distribution F with parameter θ. Then a **pivot** is a function $Q = q(Y, \theta)$ of the data and the parameter θ, where the distribution of Q is known and does not depend on θ. We say that Q is **pivotal**.

Confidence intervals

Let $Y = (Y_1, \dots, Y_n)$ be data from a parametric statistical model with scalar parameter θ. A **confidence interval (CI)** (L, U) **for** θ with lower confidence bound L and upper confidence bound U is a random interval that contains θ with a specified probability, called the **(confidence) level** of the interval.

- $L = l(Y)$ and $U = u(Y)$ are statistics that can be computed from the data $Y_1, \dots, Y_n$. They do not depend on θ

- In a continuous setting (so $<$ gives the same probabilities as $\leq$), and if we write the probabilities that θ lies below and above the interval as

$$\mathrm{P}(\theta < L) = \alpha_L, \quad \mathrm{P}(U < \theta) = \alpha_U$$

  then (L, U) has confidence level

$$\mathrm{P}(L \leq \theta \leq U) = 1 - \mathrm{P}(\theta < L) - \mathrm{P}(U < \theta) = 1 - \alpha_L - \alpha_U$$

- Often we seek an interval with equal probabilities of not containing θ at each end, with $\alpha_L = \alpha_U = \frac{\alpha}{2}$, giving an **equi-tailed** $(1 - \alpha) \times 100\%$ **confidence interval**
- We usually take standard values of α, such that $1 - \alpha = 0.9, 0.95, 0.99, \dots$

Construction of a CI :

- We use pivots to construct CIs :
  - we find a pivot $Q = q(Y, \theta)$ involving θ
  - we obtain the quantiles $q_{\alpha_U}, q_{1-\alpha_L}$ of Q, which do not depend on θ
  - then we transform the equation

$$\mathrm{P}\{q_{\alpha_U} \leq q(Y, \theta) \leq q_{1-\alpha_L}\} = (1 - \alpha_L) - \alpha_U$$

    into the form

$$\mathrm{P}(L \leq \theta \leq U) = 1 - \alpha_L - \alpha_U$$

    where the bounds L, U depend on Y, $q_{1-\alpha_L}$ and q_{α_U}, but not on θ
- In many cases, the bounds are of a standard form

Interpretation of a CI :

- (L, U) is a random interval that contains θ with probability $1 - \alpha$
- We imagine an infinite sequence of repetitions of the experiment that gave (L, U)
- In that case, the CI that we calculated is one of an infinity of possible CIs, and we can consider that our CI was chosen at random from among them
- Although we do not know whether our particular CI contains θ, the event $\theta \in (L, U)$ has probability $1 - \alpha$, matching the confidence level of the CI

One- and two-sided intervals

- A **two-sided confidence interval** (L, U) is generally used, but **one-sided confidence intervals**, of the form $(-\infty, U)$ or (L, ∞), are also sometimes required
- For one-sided CIs, we take $\alpha_U = 0$ or $\alpha_L = 0$, giving respective intervals (L, ∞) or $(-\infty, U)$
- To get a one-sided $(1 - \alpha) \times 100\%$ interval, we can compute a two-sided interval with $\alpha_L = \alpha_U = \alpha$, and then replace the unwanted limit by $\pm\infty$ (or another value if required in the context)

Approximate normal confidence intervals

- We can often construct approximate CIs using the CLT, since many statistics that are based on averages of $Y = (Y_1, \ldots, Y_n)$ have approximate normal distributions for large n. If $T = t(Y)$ is an estimator of θ with standard error $\sqrt{V}$, and the theorem of standard errors applies, then

  $$T \dot{\sim} \mathcal{N}(\theta, V)$$

  and so $\frac{T-\theta}{\sqrt{V}} \dot{\sim} \mathcal{N}(0,1)$. Thus

  $$\mathrm{P}\left\{ z_{\alpha_U} < \frac{T-\theta}{\sqrt{V}} \leq z_{1-\alpha_L} \right\} \doteq \Phi(z_{1-\alpha_L}) - \Phi(z_{\alpha_U}) = 1 - \alpha_L - \alpha_U$$

  implying that an approximate $(1 - \alpha_L - \alpha_U) \times 100\%$ CI for θ is

  $$(L, U) = (T - \sqrt{V} z_{1-\alpha_L}, T - \sqrt{V} z_{\alpha_U})$$

  Recall that if $\alpha_L, \alpha_U < \frac{1}{2}$, then $z_{1-\alpha_L} > 0$ and $z_{\alpha_U} < 0$, so $L < U$

- Often we take $\alpha_L = \alpha_U = 0.025$, and then $z_{1-\alpha_L} = -z_{\alpha_U} = 1.96$, giving the 'rule of thumb' $(L, U) \approx T \pm 2\sqrt{V}$ for a two-sided 95% confidence interval

(Chi-square distribution) Let $Z = (Z_1, \ldots, Z_n) \sim \mathcal{N}_n(0, I_n)$. Then the distribution of $||Z||^2 = Z_1^2 + \cdots + Z_n^2$ is called the **chi-square distribution** with n degrees of freedom and denoted $||Z||^2 \sim \chi_n^2$.

$$\mathrm{E}(||Z||^2) = \mathrm{E}(Z_1^2) + \cdots + \mathrm{E}(Z_n^2) = n$$

If $Z \sim \mathcal{N}(0, \sigma^2 I_n)$, then $||Z||^2 \sim \sigma^2 \chi_n^2$ meaning that $||Z||^2 = \sigma^2 S^2$ with $S^2 \sim \chi_n^2$
(Student's t-distribution) Student's t-distribution with n degrees of freedom is the distribution of

$$T = \frac{Z}{\sqrt{V/n}} \text{ where } \begin{cases} Z \sim \mathcal{N}(0,1) \\ V \sim \chi_n^2 \\ Z \text{ and } V \text{ are independent} \end{cases}$$

We denote $T \sim t_n$.

Normal random sample

If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$\left.\begin{array}{l} \overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n) \\ (n-1)S^2 = \sum_{j=1}^n (Y_j - \overline{Y})^2 \sim \sigma^2 \chi_{n-1}^2 \end{array}\right\} \text{independent}$$

where χ_ν^2 represents the **chi-square distribution with ν degrees of freedom**.
The first result here implies that if σ^2 is known, then

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$$

is a pivot that provides an exact $(1-\alpha_L-\alpha_U)$ confidence interval for μ, of the form

$$(L,U)=\left(\overline{Y}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha_L},\overline{Y}-\frac{\sigma}{\sqrt{n}}z_{\alpha_U}\right)$$

where z_p denotes the p quantile of the standard normal distribution.

Unknown variance

- In applications σ^2 is usually unknown. If so, the theorem of normal random sample implies that

  $$\frac{\overline{Y}-\mu}{\sqrt{S^2/n}}\sim t_{n-1},\qquad \frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$$

  are pivots that provide confidence intervals for μ and σ^2, respectively, i.e.,

  $$(L,U)=\left(\overline{Y}-\frac{S}{\sqrt{n}}t_{n-1}(1-\alpha_L),\overline{Y}-\frac{S}{\sqrt{n}}t_{n-1}(\alpha_U)\right)$$

  $$(L,U)=\left(\frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha_L)},\frac{(n-1)S^2}{\chi^2_{n-1}(\alpha_U)}\right)$$

  where :

  - $t_\nu(p)$ is the p quantile of the **Student t distribution with ν degrees of freedom**
  - χ^2_ν is the p quantile of the **chi-square distribution with ν degrees of freedom**

- For symmetric densities such as the normal and the Student t, the quantiles satisfy

  $$z_p=-z_{1-p},\qquad t_\nu(p)=-t_\nu(1-p)$$

  so equi-tailed $(1-\alpha)\times 100\%$ CIs have the forms

  $$\overline{Y}\pm n^{-1/2}\sigma z_{1-\alpha/2},\qquad \overline{Y}\pm n^{-1/2}St_{n-1}(1-\alpha/2)$$

Comments

- The construction of confidence intervals is based on pivots, often using the central limit theorem to approximate the distribution of an estimator, and thus giving approximate intervals
- A confidence interval (L,U) not only suggests where an unknown parameter is situated, but its width $U-L$ gives an idea of the precision of the estimate
- In most cases

  $$U-L\propto\sqrt{V}\propto n^{-1/2}$$

  so multiplying the sample size by 100 increases precision only by a factor of 10
- Having to estimate the variance using V decreases precision, and thus increases the width
- To get a one sided $(1-\alpha)\times 100\%$ interval, we can compute a two-sided interval with $\alpha_L=\alpha_U=\alpha$, and then replace the unwanted limit by $\pm\infty$ (or another suitable limit)

- In some cases, especially normal models, exact CIs are available

8.4 Hypothesis Tests

Confidence intervals and tests

- We can use confidence intervals (CIs) to assess the plausibility of a value θ^0 of θ :
  - If θ^0 lies inside a $(1-\alpha)\times 100\%$ CI, then we **cannot reject** the hypothesis that $\theta = \theta^0$, at **significance level** α
  - If θ^0 lies outside a $(1-\alpha)\times 100\%$ CI, then we **reject** the hypothesis that $\theta = \theta^0$, at **significance level** α
- The discussion of the scientific method at the start of §7 tells us that data cannot prove correctness of a theory (hypothesis), because we can always imagine that future data or a new experiment might undermine it, but data can falsify theory. Hence we can **reject** or **not reject (provisionally accept)** a hypothesis, but we cannot **prove** it
- The decision to reject or not depends on the chosen significance level α : we will reject less often if α is small, since then the CI will be wider
- If α is small and we do reject, this gives stronger evidence against θ^0
- Use of a two-sided CI (L, U) implies that seeing either $\theta^0 < L$ or $\theta^0 > U$ would be evidence against the theory. In general we should consider whether to use $(-\infty, U)$ or (L, ∞) instead

Null and alternative hypotheses

In a general testing problem we aim to use the data to decide between two hypotheses.

- The **null hypothesis** H_0, which represents the theory/model we want to test
  - For the coin tosses, H_0 is that the coin is fair, i.e., $\mathrm{P(heads)} = \theta = \theta^0 = \frac{1}{2}$
- The **alternative hypothesis** H_1, which represents what happens if H_0 is false
  - For the coin tosses, H_1 is that the coin is not fair, i.e., $\mathrm{P(heads)} \neq \theta^0$
- When we decide between the hypotheses, we can make two sorts of error :

  Type I error (false positive) : H_0 is true, but we wrongly reject it (and choose H_1)

  Type II error (false negative) : H_1 is true, but we wrongly accept H_0

		Decision	
		Accept H_0	Reject H_0
State of Nature	H_0 true	Correct choice (True negative)	Type I Error (False positive)
	H_1 true	Type II Error (False negative)	Correct choice (True positive)

Taxonomy of hypotheses

A **simple hypothesis** entirely fixes the distribution of the data Y, whereas a **composite hypothesis** does not fix the distribution of Y

ROC curve

The **receiver operating characteristic (ROC) curve** of a test plots $\beta(t)$ against $\alpha(t)$ as the cut-off t varies, i.e., it shows $(\mathrm{P}_0(T \geq t), \mathrm{P}_1(T > t))$, when $t \in \mathbb{R}$.

Size and power

- As μ increases, it becomes easier to detect when H_0 is false, because the densities under H_0 and H_1 become more separated, and the ROC curve moves 'further north-west'
- When H_0 and H_1 are the same, i.e., $\mu = 0$, then the curve lies on the diagonal. Then the hypotheses cannot be distinguished
- In applications, μ is usually unknown, so we fix α (often at some conventional value, e.g., $0.05, 0.01$) and then accept the resulting $\beta(\alpha)$
- We also call
  - the **false positive probability** the **size** α of the test
  - the **true positive probability** the **power** β of the test

Let $\mathrm{P}_0(\cdot)$ and $\mathrm{P}_1(\cdot)$ denote probabilities computed under null and alternative hypotheses H_0 and H_1 respectively. Then the **size** and **power** of a statistical test of H_0 against H_1 are

$$\text{size } \alpha = \mathrm{P}_0(\text{reject } H_0), \qquad \text{power } \beta = \mathrm{P}_1(\text{reject } H_0)$$

Power and confidence intervals

- If the test is based on a $(1-\alpha) \times 100\%$ CI, the size is the probability that the true value of the parameter lies outside the CI, so it is α
- Taking a smaller value of α gives a wider interval, so it must decrease the power
- Usually the width of the interval (L, U) satisfies

  $$U - L \propto n^{1/2}$$

  so increasing n gives a narrower interval and will increase the power of the test. This makes sense, because having more data should allow us to be more certain in our conclusions
- Unfortunately, not all tests correspond to confidence intervals, so we need a more general approach
- For example, checking the fit of a model is not usually possible using a confidence interval...

Testing goodness of a fit

We may want to assess whether a statistical model fits data appropriately.

Pearson statistic

Let $O_1, \ldots, O_k$ be the number of observations of a random sample of size $n = n_1 + \cdots + n_k$ falling into the categories $1, \ldots, k$, whose expected numbers are $E_1, \ldots, E_k$, where $E_i > 0$. Then the **Pearson statistic (or chi-square statistic)** is

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Let $Z_1, \ldots, Z_\nu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then $W = Z_1^2 + \cdots + Z_\nu^2$ follows the **chi-square distribution with ν degrees of freedom**, whose density function is

$$f_W(w) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} w^{\nu/2-1} e^{-w/2}, \qquad w > 0, \qquad \nu = 1, 2, \ldots$$

where $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\, du$, $a > 0$, is the gamma function.

- If the joint distribution of $O_1, \ldots, O_k$ is multinomial with denominator n and probabilities $p_1 = E_1/n, \ldots, p_k = E_k/n$, then $T \dot\sim \chi^2_{k-1}$, the approximation being good if $k^{-1} \sum E_i \geq 5$
- We can use T to check the agreement between the data $O_1, \ldots, O_k$ and the theoretical probabilities $p_1, \ldots, p_k$

Rationale :

- If $O_i \approx E_i$ for all i, then T will be small, otherwise it will tend to be bigger
- If the joint distribution of $O_1, \ldots, O_k$ is multinomial with denominator n and probabilities $p_i = E_i/n$, then each $O_i \sim \mathrm{B}(n, p_i)$, and thus

  $$\mathrm{E}(O_i) = np_i = E_i, \qquad \mathrm{var}(O_i) = np_i(1-p_i) = E_i(1 - E_i/n) \approx E_i$$

  thus $Z_i = \frac{O_i - E_i}{\sqrt{E_i}} \dot\sim \mathcal{N}(0,1)$ for large n, and we would imagine that

  $$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} Z_i^2 \dot\sim \chi^2_k$$

  but the constraint $\sum_i O_i = n$ means that only $k-1$ of the Z_i can vary independently, thus reducing the degrees of freedom to $k-1$

Evidence and P-values

A statistical hypothesis test has the following elements :

- a **null hypothesis** H_0, to be tested against an **alternative hypothesis** H_1
- **data**, from which we compute a **test statistic** T, chosen such that large values of T provide evidence against H_0
- the observed value of T is t_{obs}, which we compare with the **null distribution** of T, i.e., the sampling distribution of T under H_0
- we measure the evidence against H_0 using the **P-value**

  $$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}})$$

  where small values of p_{obs} suggest that either

  - H_0 is true but something unlikely has occurred
  - H_0 is false

- If $p_{\text{obs}} < \alpha$, then we say that the test is **significant at level** α or **significant at the** $\alpha \times 100\%$ **level**

- If we must make a decision, then we **reject** H_0 if $p_{\text{obs}} < \alpha$, where α is the significance level of the test, and we (provisionally) **accept** H_0 if $p_{\text{obs}} \geq \alpha$

Decision procedures and measures of evidence

We can use a test of H_0 in two related ways :

- as a **decision procedure**, where we
  - choose a level α at which we want to test H_0
  - reject H_0 (i.e., choose H_1) if the P-value is less than α
  - do not reject H_0 if the P-value is greater than α

- as a **measure of evidence** against H_0, with
  - small values of p_{obs} suggesting stronger evidence against H_0
  - H_1 need not be explicit, though the type of departure from H_0 that we seek is implicit in the choice of T

- Knowing the exact value of p_{obs} is more useful than knowing that H_0 has been rejected, so the measure of evidence is more informative

- The strength of the evidence contained in a P-value can be summarised as follows :

α	Evidence against H_0
0.05	Weak
0.01	Positive
0.001	Strong
0.0001	Very strong

Choice of α

- As with CIs, conventional values are often used, such as $\alpha = 0.05, 0.01, 0.001$

- The most common value is $\alpha = 0.05$, which corresponds to a Type I error probability of 5%, i.e., H_0 will be rejected once in every 20 tests, even when it is true

- When many tests are performed, using large α can give many **false positives**, i.e., significant tests for which in fact H_0 is true

- Consider a microarray experiment, where we test 1000 genes at significance level α, to see which genes influence some disease. If only 100 genes have effects, we can write

$$\mathrm{P}(H_0) = 900/1000, \quad \mathrm{P}(H_1) = 100/1000, \quad \mathrm{P}(S \mid H_0) = \alpha, \quad \mathrm{P}(S \mid H_1) = \beta$$

where α is the size of the test, $\beta > \alpha$ is its power, and S denotes the event that the test is significant at level α. Bayes' theorem gives

$$\mathrm{P}(H_0 \mid S) = \frac{\mathrm{P}(H_0)\mathrm{P}(S \mid H_0)}{\mathrm{P}(H_0)\mathrm{P}(S \mid H_0) + \mathrm{P}(H_1)\mathrm{P}(S \mid H_1)} = \frac{0.9\alpha}{0.9\alpha + 0.1\beta}$$

Hence with $\alpha = 0.05$, $\beta = 0.8$, say, $\mathrm{P}(H_0 \mid S) \doteq 0.36$, so over one-third of significant tests will not be intersting. If instead we set $\alpha = 0.005$, we have $\mathrm{P}(H_0 \mid S) \doteq 0.05$, which is more reasonable.

8.5 Comparison of Tests

Types of test

There are many different tests for different hypotheses. Two important classes of tests are :

- **parametric tests**, which are based on a parametric statistical model, such as $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and $H_0 : \mu = 0$
- **nonparametric tests**, which are based on a more general statistical model, such as $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f$, and $H_0 : \mathrm{P}(Y > 0) = \mathrm{P}(Y < 0) = \frac{1}{2}$, i.e., the median of f is at $y = 0$

The main advantage of a parametric test is the possibility of finding a (nearly-)optimal test, if the underlying assumptions are correct, though such a test could perform badly in the presence of outliers.
A nonparametric test is often more robust, but it will suffer a loss of power compared to a parametric test, used appropriately.

ROC curve II

- We previously met the ROC curve as a summary of the properties of a test
- A good test will have a ROC curve lying as close to the upper left corner as possible
- A useless test has a ROC curve lying on (or close to) the diagonal
- This suggests that if we have a choice of tests, we should choose one whose ROC curve is as close to the north-west as possible, i.e., we should choose the test that maximises the power for a given size
- This leads us to the **Neyman-Pearson lemma**, which says how to do this (in ideal circumstances)

Most powerful tests

- We aim to choose our test statistic T to maximise the power of the test for a given size
- A decision procedure corresponds to partitioning the sample space Ω containing the data Y into a **rejection region**, $\mathcal{Y}$, and its complement, $\overline{\mathcal{Y}}$, with

$$Y \in \mathcal{Y} \implies \text{Reject } H_0, \qquad Y \in \overline{\mathcal{Y}} \implies \text{Accept } H_0$$

- We aim to choose $\mathcal{Y}$ such that $\mathrm{P}_1(Y \in \mathcal{Y})$ is the largest possible such that $\mathrm{P}_0(Y \in \mathcal{Y}) = \alpha$

(Neyman-Pearson) Let $f_0(y), f_1(y)$ be the densities of Y under simple null and alternative hypotheses. Then if it exists, the set

$$\mathcal{Y}_\alpha = \{y \in \Omega : \frac{f_1(y)}{f_0(y)} > t\}$$

such that $\mathrm{P}_0(Y \in \mathcal{Y}_\alpha) = \alpha$ maximises $\mathrm{P}_1(Y \in \mathcal{Y}_\alpha)$, amongst all the $\mathcal{Y}'$ such that $\mathrm{P}_0(Y \in \mathcal{Y}') \leq \alpha$. Thus to maximise the power of a given threshold, we must base the decision on $\mathcal{Y}_\alpha$.

Power and distance

- A canonical example is where $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu = \mu_1$$

- If σ^2 is known, then the Neyman-Pearson lemma can be applied, and we find that the most powerful test is based on $\overline{Y}$ and its power is $\Phi(z_\alpha + \delta)$, where $\Phi(z_\alpha) = \alpha$, and

$$\delta = n^{1/2} \frac{|\mu_1 - \mu_0|}{\sigma}$$

  is the **standardized distance** between the models

- We see that
  - the power increases if n increases, or if $|\mu_1 - \mu_0|$ increases, since in either case the difference between the hypotheses is easier to detect
  - the power decreases if σ increases, since then the data become noisier
  - if $\delta = 0$, then the power equals the size, because the two hypotheses are the same, and therefore $\mathrm{P}_0(\cdot) = \mathrm{P}_1(\cdot)$
- Many other situations are analogous to this, with power depending on generalised versions of δ

Summary

- We have considered the situation where we have to make a binary choice between
  - the null hypothesis, H_0, against which we want to test
  - the alternative hypothesis, H_1

  using a test statistic T whose observed value is t_{obs}, computing the P-value

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}})$$

  which is computed assuming that H_0 is true

- We can consider p_{obs} as a measure of the evidence in the data against H_0
- For a test with significance level α, we reject H_0 and choose H_1 if $p_{\text{obs}} < \alpha$
- We must accept that we can make mistakes :

		Decision	
		Accept H_0	Reject H_0
State of Nature	H_0 true	Good choice	Type I Error
	H_1 true	Type II Error	Good choice

- If we try to minimise the probability of Type II error (i.e., maximise power) for a given probability of Type I error (fixed size), we can construct an optimal test, but this is only possible in simple cases. Otherwise we usually have to compare tests numerically

9 Likelihood

9.1 Motivation

Illustration

- When we toss a coin, small asymmetries influence the probability of obtaining heads, which is not necessarily $\frac{1}{2}$. If $Y_1, \ldots, Y_n$ denote the results of independent Bernoulli trials, then we can write

$$\mathrm{P}(Y_j = 1) = \theta, \quad \mathrm{P}(Y_j = 0) = 1 - \theta, \quad 0 \leq \theta \leq 1, \quad j = 1, \ldots, n$$

- Below is such a sequence for a 5Fr coin with $n = 10$

$$1\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 1$$

Basic Idea

For a value of θ which is not very credible, the density of the data will be smaller : the higher the density, the more credible the corresponding θ. Since the $y_1, \ldots, y_{10}$ result from independent trials, we have

$$f(y_1, \ldots, y_{10}; \theta) = \prod_{j=1}^{10} f(y_j; \theta) = f(y_1; \theta) \times \cdots \times f(y_{10}; \theta) = \theta^5 \times (1 - \theta) \times \theta^4 = \theta^9 (1 - \theta)$$

which we will consider as a function of θ for $0 \leq \theta \leq 1$, called the **likelihood** $L(\theta)$.

Relative likelihood

- To compare values of θ, we only need to consider the ratio of the corresponding values of $L(\theta)$:

$$\frac{L(\theta_1)}{L(\theta_2)} = \frac{f(y_1, \ldots, y_{10}; \theta_1)}{f(y_1, \ldots, y_{10}; \theta_2)} = \frac{\theta_1^9 (1 - \theta_1)}{\theta_2^9 (1 - \theta_2)} = c$$

  implies that θ_1 is c times more plausible than θ_2

- The most plausible value is $\widehat{\theta}$, which satisfies

$$L(\widehat{\theta}) \geq L(\theta), \quad 0 \leq \theta \leq 1$$

  $\widehat{\theta}$ is called the **maximum likelihood estimate**

- To find $\widehat{\theta}$, we can equivalently maximise the **log likelihood**

$$\ell(\theta) = \log L(\theta)$$

- The **relative likelihood** $RL(\theta) = L(\theta)/L(\widehat{\theta})$, $0 \leq RL(\theta) \leq 1$ gives the plausibility of θ with respect to $\widehat{\theta}$

9.2 Scalar Parameter

Likelihood

Let y be a set of data, whose joint probability density $f(y;\theta)$ depends on a parameter θ, then the **likelihood** and the **log likelihood** are

$$L(\theta) = f(y;\theta), \qquad \ell(\theta) = \log L(\theta)$$

considered a function of θ.
If $y = (y_1, \ldots, y_n)$ is a realisation of the independent random variables of $Y_1, \ldots, Y_n$, then

$$L(\theta) = f(y;\theta) = \prod_{j=1}^{n} f(y_j;\theta), \qquad \ell(\theta) = \sum_{j=1}^{n} \log f(y_j;\theta)$$

where $f(y_j;\theta)$ represents the density of one of the y_j.

Maximum likelihood estimation

The **maximum likelihood estimate** $\widehat{\theta}$ satisfies

$$L(\widehat{\theta}) \geq L(\theta) \text{ for all } \theta$$

which is equivalent to $\ell(\widehat{\theta}) \geq \ell(\theta)$, since $L(\theta)$ and $\ell(\theta)$ have their maxima at the same value of θ. The corresponding random variable is called the **maximum likelihood estimator (MLE)**.

- Often $\widehat{\theta}$ satisfies

  $$\frac{d\ell(\widehat{\theta})}{d\theta} = 0, \qquad \frac{d^2\ell(\widehat{\theta})}{d\theta^2} < 0$$

  In this course we will suppose that the first of these equations has only one solution (not always the case in reality)

Information

Assume $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$. The **observed information** $J(\theta)$ (depends on the observation $y = (y_1, \ldots, y_n)$) and the **expected information (or Fisher information)** $I(\theta)$ (does not depend on the observation, expectation under $y_1, \ldots, y_n \overset{\text{iid}}{\sim} f(\cdot;\theta)$) are

$$J(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2}, \qquad I(\theta) = \mathrm{E}\{J(\theta)\} = \mathrm{E}\left\{-\frac{d^2\ell(\theta)}{d\theta^2}\right\}$$

They measure the curvature of $-\ell(\theta)$: the larger $J(\theta)$ and $I(\theta)$, the more concentrated $\ell(\theta)$ and $L(\theta)$ are.

Limit distribution of the MLE

Let $Y_1, \ldots, Y_n$ be a random sample from a parametric density $f(y;\theta)$, and let $\widehat{\theta}$ be the MLE of θ. If f satisfies **regularity conditions** $\xcancel{\heartsuit}$, then

$$J(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} \mathcal{N}(0,1), \qquad n \to \infty \qquad\qquad \heartsuit$$

Thus for large n,

$$\widehat{\theta} \mathrel{\dot\sim} \mathcal{N}\left\{\theta, J(\widehat{\theta})^{-1}\right\} \qquad\qquad \heartsuit$$

and a two-sided equi-tailed CI for θ with approximate level $(1-\alpha)$ is

$$\mathcal{I}^{\widehat{\theta}}_{1-\alpha} = (L, U) = (\widehat{\theta} - J(\widehat{\theta})^{1/2} z_{1-\alpha/2}, \widehat{\theta} + J(\widehat{\theta})^{1/2} z_{1-\alpha/2})$$

We can show that for large n (and a regular model) no estimator has a smaller variance than $\widehat{\theta}$, which implies that the CIs $\mathcal{I}^{\widehat{\theta}}_{1-\alpha}$ are as narrow as possible.

Likelihood ratio statistic

Let $\ell(\theta)$ be the log likelihood for a scalar parameter θ, whose MLE is $\widehat{\theta}$. Then the **likelihood ratio statistic** is

$$W(\theta) = 2\Big\{\ell(\widehat{\theta}) - \ell(\theta)\Big\} = 2\Big\{\log L(\widehat{\theta}) - \log L(\theta)\Big\} = 2\log\frac{L(\widehat{\theta})}{L(\theta)} = -2\log RL(\theta)$$

If θ^0 is the value of θ that generated the data, then under the regularity conditions giving $\widehat{\theta}$ a normal limit distribution

$$W(\theta^0) \xrightarrow{D} \chi^2_1, \qquad n \to \infty$$

Hence $W(\theta^0) \dot{\sim} \chi^2_1$ for large n.
If k is a positive integer, the χ^2_k distribution with k degrees of freedom is the distribution of

$$Z_1^2 + \cdots + Z_k^2, \text{ where } Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \qquad \heartsuit$$

$$\theta_0, \qquad Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta_0)$$

- Suppose we want to test the hypothesis $H_0 : \theta = \theta^0$, where θ^0 is fixed. If H_0 is true, the theorem implies that $W(\theta^0) \dot{\sim} \chi^2_1$. The larger $W(\theta^0)$ is, the more we doubt H_0. Thus we can take $W(\theta^0)$ as a test statistic, whose observed value is w_{obs}, and with

$$p_{\text{obs}} = \mathrm{P}\{W(\theta^0) \geq w_{\text{obs}}\} \doteq \mathrm{P}\{\chi^2_1 \geq w_{\text{obs}}\}$$

  as significance level. The smaller p_{obs} is, the more we doubt H_0

- Let $\chi^2_\nu(1-\alpha)$ be the $(1-\alpha)$ quantile of the χ^2_ν distribution. The theorem of likelihood ratio statistic implies that a CI for θ^0 at the $(1-\alpha)$ level is the set

$$\mathcal{I}^W_{1-\alpha} = \{\theta : W(\theta) \leq \chi^2_1(1-\alpha)\} = \Big\{\theta : 2\Big\{\ell(\widehat{\theta}) - \ell(\theta)\Big\} \leq \chi^2_1(1-\alpha)\Big\} = \Big\{\theta : \ell(\theta) \geq \ell(\widehat{\theta}) - \frac{1}{2}\chi^2_1(1-\alpha)\Big\}$$

- With $1-\alpha = 0.95$ we have $x^2_1(0.95) = 3.84$. Thus the 95% CI for a scalar θ contains all θ such that $\ell(\theta) \geq \ell(\widehat{\theta}) - 1.92$. In this case we have

$$RL(\theta) = \frac{L(\theta)}{L(\widehat{\theta})} = \exp\{\ell(\theta) - \ell(\widehat{\theta})\} \geq \exp(-1.92) \approx 0.15$$

CIs based on the likelihood ratio statistic

- When n increases, the CI becomes narrower and more symmetric about $\widehat{\theta}$
- When $1-\alpha$ increases (i.e., α decreases), the CI becomes wider

Regularity

The regularity conditions are complicated. Situations where they are false are often cases where

- one of the parameters is discrete

- the support of $f(y;\theta)$ depends on θ
- the true θ is on the limit of its possible values

The conditions are satisfied in the majority of cases met in practice.

9.3 Vector Parameter

9.4 Statistical Modelling

9.5 Linear Regression

10 Bayesian Inference

10.1 Basic Ideas

10.2 Bayesian Modelling

Summaries

Which distribution ?

Is X based on independent trials (0/1) with a same probability p, or on draws from a finite population, with replacement ?

If **Yes**, is the total number of trials n fixed, so $X \in \{0, \dots, n\}$?

If **Yes** : use the **binomial** distribution, $X \sim \mathrm{B}(n, p)$ (and thus the **Bernoulli** distribution if $n = 1$)

If $n \approx \infty$ or $n \gg np$, we can use the **Poisson** distribution, $X \sim \mathrm{Pois}(np)$

If **No**, then $X \in \{n, n+1, \dots\}$, and we use the **geometric** (if X is the number of trials until one success) or **negative binomial** (if X is the number of trials until the last of several successes) distributions

If **No**, then if the draw is independent but without replacement from a finite population, then $X \sim$ **hypergeometric** distribution

X discrete or continuous ?

	Discrete	Continuous
Support D_X	countable	contains an interval $(x_-, x_+) \subset \mathbb{R}$
f_X	mass function dimensionless $0 \le f_X(x) \le 1$ $\sum_{x \in \mathbb{R}} f_X(x) = 1$	density function units $[x]^{-1}$ $0 \le f_X(x)$ $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$
$F_X(a) = \mathrm{P}(X \le a)$	$\sum_{x \le a} f_X(x)$	$\int_{-\infty}^{a} f_X(x)\ dx$
$\mathrm{P}(X \in \mathcal{A})$	$\sum_{x \in \mathcal{A}} f_X(x)$	$\int_{\mathcal{A}} f_X(x)\ dx$
$\mathrm{P}(a < X \le b)$	$\sum_{\{x: a < x \le b\}} f_X(x)$	$\int_a^b f_X(x)\ dx$
$\mathrm{P}(X = a)$	$f_X(a) \ge 0$	$\int_a^a f_X(x)\ dx = \mathbf{0}$
$\mathrm{E}\{g(X)\}$ (if well defined)	$\sum_{x \in \mathbb{R}} g(x) f_X(x)$	$\int_{-\infty}^{\infty} g(x) f_X(x)\ dx$

Which density ?

Uniform variables lie in a finite interval, and give equal probability to each part of the interval

Exponential and **gamma** variables lie in $(0, \infty)$, and are often used to model waiting times and other positive quantities

the gamma has two parameters and is more flexible, but the exponential is simpler and has some elegant properties

Pareto variables lie in the interval (β, ∞), so are not appropriate for arbitrary positive quantities (which could be smaller than β), but are often used to model financial losses over some threshold β

Normal variables lie in $\mathbb{R}$ and are used to model quantities that arise (or might arise) through averaging of many small effects (e.g., height and weight, which are influenced by many genetic factors), or where measurements are subject to error

Laplace variables lie in $\mathbb{R}$; the Laplace distribution can be used in place of the normal in situations where outliers might be present

The following pages contain a table that summarizes the Random Variables seen in this course. The first is a printable version, split on 2 pages for readability. The second is a computer version, that fits on 1 page for practicality.

Random Variable		Type	Support	PMF or PDF
	$X \sim$		D_X	$f_X(x)$
Bernoulli / Indicator	$\mathrm{Ber}(p)$ / $\mathbb{I}$	Discrete	$0, 1$	$\begin{cases} 1-p & \text{if } x=0 \\ p & \text{if } x=1 \end{cases}$
Binomial	$\mathrm{B}(n,p)$	Discrete	$0, 1, \ldots, n$	$\binom{n}{x} p^x (1-p)^{n-x}$
Geometric	$\mathrm{Geom}(p)$	Discrete	$1, 2, \ldots$	$p(1-p)^{x-1}$
Negative binomial	$\mathrm{NegBin}(n,p)$	Discrete	$n, n+1, n+2, \ldots$	$\binom{x-1}{n-1} p^n (1-p)^{x-n}$
Negative binomial (alternate)		Discrete	$0, 1, 2, \ldots$	$\frac{\Gamma(y+\alpha)}{\Gamma(\alpha) y!} p^\alpha (1-p)^y$
Hypergeometric	$\mathrm{HyperGeom}(w, b; m)$	Discrete	$\max(0, m-b), \ldots, \min(w, m)$	$\frac{\binom{w}{x}\binom{b}{m-x}}{\binom{w+b}{m}}$
Discrete uniform	$\mathrm{DU}(a,b)$	Discrete	$a, a+1, \ldots, b$	$\frac{1}{b-a+1}$
Poisson	$\mathrm{Pois}(\lambda)$	Discrete	$0, 1, \ldots$	$\frac{\lambda^x}{x!} e^{-\lambda}$
Uniform	$\mathrm{U}(a,b)$	Continuous	$\mathbb{R}$	$\begin{cases} \frac{1}{b-a} & a \leq u \leq b \\ 0 & \text{otherwise} \end{cases}$
Exponential	$\exp(\lambda)$	Continuous	$\mathbb{R}$	$\begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$
Gamma	$\mathrm{Gamma}(\alpha, \lambda)$	Continuous	$\mathbb{R}$	$\begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$
Laplace / Double exponential		Continuous	$\mathbb{R}$	$\frac{\lambda}{2} e^{-\lambda \lvert x-\eta \rvert}$
Pareto		Continuous	$\mathbb{R}$	
Normal / Gaussian	$\mathcal{N}(\mu, \sigma^2)$	Continuous	$\mathbb{R}$	$\frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$
Standard normal	$\mathcal{N}(0,1)$	Continuous	$\mathbb{R}$	$(2\pi)^{-1/2} e^{-z^2/2}$

Random Variable	CDF	Expectation	Variance	MGF
	$F_X(x)$	$\mathrm{E}(X)$	$\mathrm{var}(X)$	$M_X(t)$
Bernoulli / Indicator		p	$p(1-p)$	$(1-p)+pe^t$
Binomial		np	$np(1-p)$	$((1-p)+pe^t)^n$
Geometric		$\frac{1}{p}$	$\frac{1-p}{p^2}$	$pe^t\frac{1}{1-e^t(1-p)},\ e^t(1-p)<1$
Negative binomial				
Negative binomial (alternate)				
Hypergeometric				
Discrete uniform				
Poisson		λ	λ	$e^{\lambda(e^t-1)}$
Uniform	$\begin{cases} 0 & u<a \\ \frac{u-a}{b-a} & a\le u\le b \\ 1 & u>b \end{cases}$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)} & t\neq 0 \\ 1 & t=0 \end{cases}$
Exponential	$\begin{cases} 1-e^{-\lambda x} & x>0 \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t},\ t<\lambda$
Gamma		$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\frac{1}{(1-\frac{t}{\lambda})^\alpha},\ t<\lambda$
Laplace / Double exponential				
Pareto	$\begin{cases} 0 & x<\beta \\ 1-\left(\frac{\beta}{x}\right)^\alpha & x\ge\beta \end{cases}$			
Normal / Gaussian		μ	σ^2	
Standard normal		0	1	

Random Variable		Type	Support	PMF or PDF	CDF	Expectation	Variance	MGF
	$X \sim$		D_X	$f_X(x)$	$F_X(x)$	$\mathrm{E}(X)$	$\mathrm{var}(X)$	$M_X(t)$
Bernoulli / Indicator	$\mathrm{Ber}(p)$ / $\mathbb{I}$	Discrete	$0, 1$	$\begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$		p	$p(1-p)$	$(1-p) + pe^t$
Binomial	$\mathrm{B}(n, p)$	Discrete	$0, 1, \dots, n$	$\binom{n}{x} p^x (1-p)^{n-x}$		np	$np(1-p)$	$((1-p) + pe^t)^n$
Geometric	$\mathrm{Geom}(p)$	Discrete	$1, 2, \dots$	$p(1-p)^{x-1}$		$\frac{1}{p}$	$\frac{1-p}{p^2}$	$pe^t \frac{1}{1-e^t(1-p)}, \; e^t(1-p) < 1$
Negative binomial	$\mathrm{NegBin}(n, p)$	Discrete	$n, n+1, n+2, \dots$	$\binom{x-1}{n-1} p^n (1-p)^{x-n}$				
Negative binomial (alternate)		Discrete	$0, 1, 2, \dots$	$\frac{\Gamma(y+\alpha)}{\Gamma(\alpha) y!} p^\alpha (1-p)^y$				
Hypergeometric	$\mathrm{HyperGeom}(w, b; m)$	Discrete	$\max(0, m-b), \dots, \min(w, m)$	$\frac{\binom{w}{x}\binom{b}{m-x}}{\binom{w+b}{m}}$				
Discrete uniform	$\mathrm{DU}(a, b)$	Discrete	$a, a+1, \dots, b$	$\frac{1}{b-a+1}$				
Poisson	$\mathrm{Pois}(\lambda)$	Discrete	$0, 1, \dots$	$\frac{\lambda^x}{x!} e^{-\lambda}$		λ	λ	$e^{\lambda(e^t - 1)}$
Uniform	$\mathrm{U}(a, b)$	Continuous	$\mathbb{R}$	$\begin{cases} \frac{1}{b-a} & a \leq u \leq b \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 0 & u < a \\ \frac{u-a}{b-a} & a \leq u \leq b \\ 1 & u > b \end{cases}$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & t \neq 0 \\ 1 & t = 0 \end{cases}$
Exponential	$\exp(\lambda)$	Continuous	$\mathbb{R}$	$\begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - t}, \; t < \lambda$
Gamma	$\mathrm{Gamma}(\alpha, \lambda)$	Continuous	$\mathbb{R}$	$\begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$		$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\frac{1}{(1-\frac{t}{\lambda})^\alpha}, \; t < \lambda$
Laplace / Double exponential		Continuous	$\mathbb{R}$	$\frac{\lambda}{2} e^{-\lambda \lvert x - \eta \rvert}$				
Pareto		Continuous	$\mathbb{R}$		$\begin{cases} 0 & x < \beta \\ 1 - \left(\frac{\beta}{x}\right)^\alpha & x \geq \beta \end{cases}$			
Normal / Gaussian	$\mathcal{N}(\mu, \sigma^2)$	Continuous	$\mathbb{R}$	$\frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$		μ	σ^2	
Standard normal	$\mathcal{N}(0, 1)$	Continuous	$\mathbb{R}$	$(2\pi)^{-1/2} e^{-z^2/2}$		0	1	