# I. Introduction

## Example

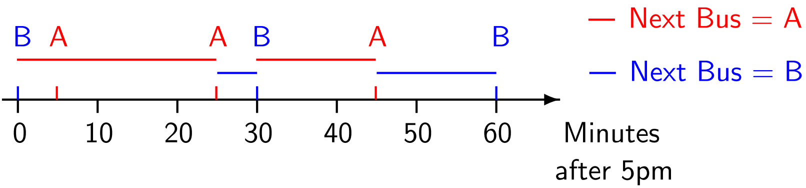
**Catching a bus**

Each day you arrive at the bus stop at a random time, evenly distributed between 5pm and 6pm.

> Bus A runs at 5:05, 5:25 and 5:45.
>
> Bus B runs at 5:00, 5:30 and 6:00.

You observe that you take bus A $2/3$ of the time.

Why is this so?



# Mathematical model of a probability space

## Definition

The **sample space** $\Omega$ is the set of all possible outcomes.

In the Bus example, let elements of $\Omega$ be the outcomes

$$\omega = \text{Time of arrival measured in minutes after 5pm.}$$

Then $\Omega = [0, 60]$.

## Definition

**Events** are subsets of $\Omega$.

Some events in the Bus example are

$$\text{Arrive at a time to catch bus B} = \{0\} \cup (25, 30] \cup (45, 60],$$

$$\text{Wait at least 8 minutes} = (5, 17] \cup (30, 37] \cup (45, 52].$$

# Mathematical model of a probability space

### Definition

Denote the **set of events** by $\mathcal{F}$.

### Definition

**Probability** is a function $\mathsf{P} : \mathcal{F} \to [0, 1]$.

How should we define P in our example?

If $E = (a, b) \subset \Omega$, set

$$\mathsf{P}(E) = \frac{b - a}{60}.$$

If $E = (a_1, b_1) \cup \ldots \cup (a_k, b_k) \subset \Omega$, where the intervals $(a_j, b_j)$ are disjoint, set

$$\mathsf{P}(E) \ = \ \sum_{j=1}^{k} \frac{b_j - a_j}{60} \ = \ \frac{\text{Total length of } E}{60}.$$

# Mathematical model of a probability space

**The probability of a single outcome**

For every $\omega \in \Omega$,

$$\mathsf{P}\{\omega\} \;=\; \frac{\text{Length of } (\omega,\omega)}{60} \;=\; 0.$$

Each individual outcome $\omega$ has probability zero — but $\mathsf{P}(\Omega) = 1$.

Why does this not contradict the axioms of probability?

The set $\Omega$ is uncountable, so it *makes no sense* to write

$$\mathsf{P}(\Omega) = \sum_{\omega} \mathsf{P}(\omega) = \ldots$$

### Corollary

*In the bus example,*

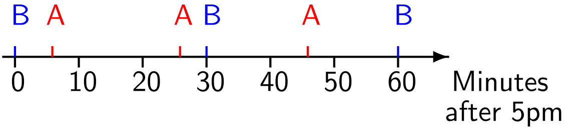
$$\mathsf{P}([a,b)) \;=\; \mathsf{P}((a,b)) \;=\; \mathsf{P}((a,b]).$$

# Random variables

**A random variable (RV)** is a real-valued function defined on the sample space,

$$X : \Omega \to \mathrm{R}.$$

Outcome $\omega$ gives the value $X(\omega)$.



Examples of RVs in the bus example are

a) $X =$ Time until bus B arrives

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = 0, \\ 30 - \omega & \text{if } 0 < \omega \leq 30, \\ 60 - \omega & \text{if } 30 < \omega \leq 60. \end{cases}$$

# Random variables

b) $Y$ = Arrival time of next bus B

$$Y(\omega) \;=\; \begin{cases} 0 & \text{if } \omega = 0, \\ 30 & \text{if } 0 < \omega \leq 30, \\ 60 & \text{if } 30 < \omega \leq 60. \end{cases}$$

c) $Z$ = Your arrival time at the bus stop

$$Z(\omega) \;=\; \omega.$$

d) Let $B$ be the event that Bus B is the next bus to arrive, so

$$B \;=\; \{0\} \,\cup\, (25, 30] \,\cup (45, 60]$$

and define the **indicator variable**

$$I_B(\omega) \;=\; \begin{cases} 1 & \text{if } \omega \in B, \\ 0 & \text{if } \omega \notin B. \end{cases}$$

# Random variables

Note that $Y$ and $I_B$ are **discrete** RVs:

$$\mathsf{P}(Y = 30) \ = \ \frac{1}{2}, \quad \mathsf{P}(Y = 60) \ = \ \frac{1}{2}$$

and

$$\mathsf{P}(I_B = 0) \ = \ \frac{2}{3}, \quad \mathsf{P}(I_B = 1) \ = \ \frac{1}{3}.$$

In contrast, $X$ and $Z$ are **continuous** RVs.

This course unit is concerned with developing a way to describe continuous RVs and to calculate their properties.

# II Random variables and cumulative distribution functions

## II.a Probability density functions (PDFs)

Recall that a random variable is a function $X : \Omega \to \mathsf{R}$.

A **discrete** RV takes values in a set which is finite or countable,

$$X(\omega) \in \{x_1, x_2, \ldots\} \quad \text{for } \omega \in \Omega,$$

and

$$\mathsf{P}(a \leq X \leq b) = \sum_{x_i :\, a \leq x_i \leq b} \mathsf{P}(X = x_i).$$

### Definition

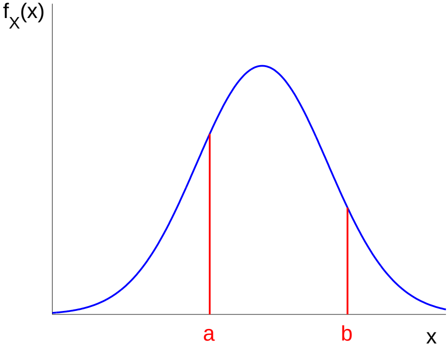
We say $X$ is a **continuous** RV if there exists a piecewise continuous function $f_X : \mathsf{R} \to [0, \infty)$ such that for all $a \leq b$

$$\mathsf{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

Then, $f_X(x)$ is the **probability density function** (PDF) of $X$.

# Probability density functions

The probability that $X$ lies in an interval is given by the area under the curve $f_X(x)$ over that interval.



$$\mathsf{P}(a \le X \le b) \; = \; \int_a^b f_X(x)\, dx.$$

# Probability density functions

Since

$$\mathsf{P}(a \le X \le b) \ = \ \int_a^b f_X(x)\,dx,$$

it is necessary that

(i) $f_X(x) \geq 0$ for all $x \in \mathsf{R}$

(ii) $\int_{-\infty}^{\infty} f_X(x)\,dx \ = \ 1.$

Note that, in general, $f_X(x) \neq \mathsf{P}(X = x)$.

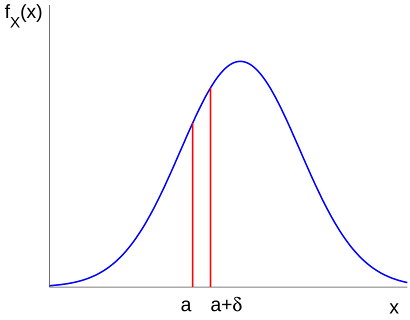
In fact,

$$\mathsf{P}(X = a) \ = \ \int_a^a f_X(x)\,dx \ = \ 0$$

and

$$\mathsf{P}(a < X < b) \ = \ \mathsf{P}(a \le X \le b).$$

# Probability density functions



Consider the event that $X$ lies in the interval $(a,\, a+\delta)$.

If $f_X(x)$ is continuous at $a$, then for small $\delta$

$$\mathsf{P}(a < X < a+\delta) \;=\; \int_a^{a+\delta} f_X(x)\,dx \;\approx\; \delta\, f_X(a).$$

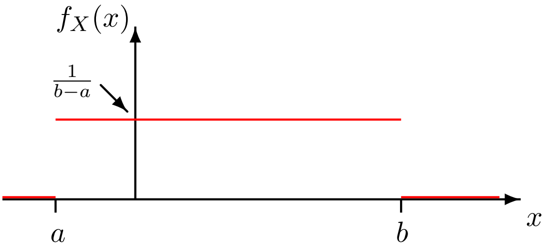
# Probability density functions

The **uniform distribution**

## Definition

The random variable $X$ has a uniform distribution on $(a, b)$, written as $X \sim \mathrm{Unif}(a, b)$ or $X \sim U(a, b)$, if it has PDF

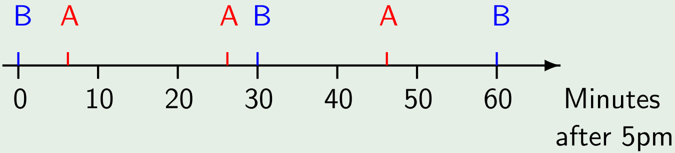
$$f_X(x) \;=\; \begin{cases} \frac{1}{b-a} & \text{for } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$



# The uniform distribution

## Example

In the bus example



Your arrival time $Z$ is a $\mathrm{Unif}(0, 60)$ RV.

## Example

Suppose $X \sim \mathrm{Unif}(0, 60)$.

Find $\mathrm{P}(2 < X < 15)$.

*See calculations on board*

# II.b Expectation and variance

Recall that if $X$ is a **discrete** RV, its expectation is

$$\mathsf{E}(X) \;=\; \sum_{x_i} x_i \, P(X = x_i).$$

### Definition

The expectation of a **continuous** random variable $X$ is

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

as long as

$$\int_{-\infty}^{\infty} |x| \, f_X(x) \, dx \;<\; \infty.$$

Question: How should we define $\mathsf{E}(X^2)$ ?

# Properties of $\mathsf{E}(X)$

## Proposition

**The law of the unconscious statistician**

*For a function* $g : \mathsf{R} \to \mathsf{R}$,

$$\mathsf{E}[g(X)] \;=\; \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

*whenever the integral exists.*

## Proof

*Treat* $Y = g(X)$ *as a random variable. Find* $f_Y(y)$, *the PDF of* $Y$, *then use the original definition of expectation:*

$$\mathsf{E}(Y) \;=\; \int_{-\infty}^{\infty} y\, f_Y(y)\, dy,$$

*and show this equals* $\int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$ *— see Ch. IV for details.*

# Properties of $\mathsf{E}(X)$

## Proposition

*Whenever the integrals exist:*

$$\text{(i)}\ \ \mathsf{E}(a\,X+b) \;=\; a\,\mathsf{E}(X)+b$$

$$\text{(ii)}\ \ \mathsf{E}[g(X)+h(X)] \;=\; E[g(X)]+\mathsf{E}[h(X)].$$

## Proof

*Apply the law of the unconscious statistician.*

*See calculations on board*

# Variance

The **variance** of $X$ is defined to be

$$\mathsf{Var}(X) \;=\; \mathsf{E}[\,(X - \mathsf{E}(X))^2\,]$$

— just as for a discrete RV.

The **standard deviation** of $X$ is

$$SD(X) \;=\; \sqrt{\mathsf{Var}(X)}.$$

# Variance

## Lemma

$$\mathsf{Var}(X) \; = \; \mathsf{E}(X^2) - [\mathsf{E}(X)]^2.$$

## Proof

*Using previous propositions (i) and (ii),*

$$\begin{aligned}
\mathsf{Var}(X) &= \mathsf{E}[\,(X - \mathsf{E}(X))^2\,] \\
&= \mathsf{E}[\,X^2 - 2\,X\,\mathsf{E}(X) + [\mathsf{E}(X)]^2\,] \\
&= \mathsf{E}(X^2) - 2\mathsf{E}(X)\mathsf{E}(X) + [\mathsf{E}(X)]^2 \\
&= \mathsf{E}(X^2) - [\mathsf{E}(X)]^2.
\end{aligned}$$

# Variance

SELF STUDY SLIDE

## Lemma

$$\mathsf{Var}(a + b\,X) \;=\; b^2\,\mathsf{Var}(X)$$

## Proof

*Check in your own time:*

$$\begin{aligned}
\mathsf{Var}(a + b\,X) &= \mathsf{E}[\,(a + b\,X)^2\,] - [\mathsf{E}(a + b\,X)]^2 \\
&= \mathsf{E}[a^2 + 2ab\,X + b^2X^2] - [a + b\,\mathsf{E}(X)]^2 \\
&= a^2 + 2ab\,\mathsf{E}(X) + b^2\,\mathsf{E}(X^2) \\
&\quad -a^2 - 2ab\,\mathsf{E}(X) - b^2[\mathsf{E}(X)]^2 \\
&= b^2(\,\mathsf{E}(X^2) - [\mathsf{E}(X)]^2\,) \\
&= b^2\,\mathsf{Var}(X).
\end{aligned}$$

# Variance

SELF STUDY SLIDE

## Example

Suppose $X$ has PDF

$$f_X(x) = \begin{cases} \frac{3}{4}(1-x^2) & -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

In your own time, show that

$$\mathsf{E}(X) = 0 \quad \text{and} \quad \mathsf{E}(X^2) = \frac{1}{5}.$$

Hence, conclude

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 = \frac{1}{5}.$$

# Where we are Lecture 3

**I** Introduction – An example of a non-discrete probability space

**II** Random variables and cumulative distribution functions

**II.a** Probability density functions

**II.b** Expectation and variance

**II.c** Independence of random variables

**II.d** Cumulative distribution functions

In this lecture, we shall cover sections **II.c** and **II.d**.

# II.c Independence of random variables

## Definition

Events $A$ and $B$ are independent if

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B).$$

## Definition

The random variables $X$ and $Y$ are independent if

$$\mathsf{P}(X \le x, Y \le y) = \mathsf{P}(X \le x)\,\mathsf{P}(Y \le y) \quad \text{for all } x \text{ and } y. \quad (1)$$

Note: This definition applies to both discrete and continuous RVs.

However, for continuous RVs, the property

$$\mathsf{P}(X = x, Y = y) = \mathsf{P}(X = x)\,\mathsf{P}(Y = y) \quad \text{for all } x \text{ and } y$$

does **not** imply independence — both sides are automatically zero.

# Independence of random variables

We might instead have defined RVs $X$ and $Y$ to be independent if

$$\mathsf{P}(X \in (x_1, x_2], Y \in (y_1, y_2]) = \mathsf{P}(X \in (x_1, x_2])\,\mathsf{P}(Y \in (y_1, y_2]) \quad \text{for all } x_1,\ x_2,\ y_1 \text{ and } y_2 \in \mathsf{R}. \tag{2}$$

In fact, condition (1) $\Leftrightarrow$ condition (2).

So we could use either definition.

(Proof: In the next Problems Class).

*For continuous RVs, independence can be stated in terms of PDFs — but this involves the joint PDF of two RVs $X$ and $Y$, which we have not yet introduced.*

*We shall return to this later in the course.*

# II.d Cumulative distribution functions

**Definition**

The cumulative distribution function (CDF) of the random variable $X$ is the function $F_X : \mathsf{R} \to [0, 1]$ defined by

$$F_X(x) \ = \ \mathsf{P}(X \leq x).$$

Note: This definition applies to both discrete and continuous RVs.

We may sometimes omit the subscript $X$ and write $F(x)$ if it is clear from the context that we are referring to the RV $X$.

# Cumulative distribution functions

## Theorem

*The CDF of the random variable $X$ has the following properties*

*(i)* $F_X$ *is increasing, i.e., if* $x \leq y$*, then* $F_X(x) \leq F_X(y)$*,*

*(ii)* $\lim_{x\to-\infty} F_X(x) = 0$*,*

*(iii)* $\lim_{x\to\infty} F_X(x) = 1$*,*

*(iv)* $F_X$ *is right-continuous, i.e., if* $x_n \downarrow x$*, then* $F_X(x_n) \downarrow F_X(x)$*.*

Notation: Here, $a_n \downarrow a$ means that $\{a_n\}$ is a decreasing sequence with $a_n > a$ for all $n$ and $\lim_{n\to\infty} a_n = a$.

*See calculations on board*

# Cumulative distribution functions

Proof of (i): $F_X(x) \ \le \ F_X(y)$ for $x < y$

We have

$$F_X(x) \ = \ \mathsf{P}(X \le x) \ = \ \mathsf{P}\{\omega : X(\omega) \le x\},$$

$$F_X(y) \ = \ \mathsf{P}(X \le y) \ = \ \mathsf{P}\{\omega : X(\omega) \le y\}.$$

Now, $x < y \ \Rightarrow \ \{\omega : X(\omega) \le x\} \ \subseteq \ \{\omega : X(\omega) \le y\}.$

So

$$\mathsf{P}\{\omega : X(\omega) \le x\} \ \le \ \mathsf{P}\{\omega : X(\omega) \le y\},$$

i.e.,

$$F_X(x) \ \le \ F_X(y).$$

□

# Cumulative distribution functions

Before proving the rest of the Theorem, we prove a Lemma.

## Lemma

*(i) If $A_1 \subset A_2 \subset \dots$ are events, then*

$$\mathsf{P}(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathsf{P}(A_n),$$

*(ii) If $B_1 \supset B_2 \supset \dots$ are events, then*

$$\mathsf{P}(\cap_{n=1}^{\infty} B_n) = \lim_{n\to\infty} \mathsf{P}(B_n).$$

# Cumulative distribution functions

## Proof of Lemma

*The proof uses the axioms of probability:*

$\mathsf{P}(E) \in [0, 1]$ *for any event* $E$.

$\mathsf{P}(\Omega) = 1, \quad \mathsf{P}(\emptyset) = 0.$

*If* $E_1$ *and* $E_2$ *are disjoint,*

$$\mathsf{P}(E_1 \cup E_2) = \mathsf{P}(E_1) + \mathsf{P}(E_2).$$

*If* $E_1, E_2, \ldots$ *are disjoint,*

$$\mathsf{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathsf{P}(E_i).$$

# Cumulative distribution functions

## Proof of Lemma

*Suppose we write* $E_1 = A_1$
*and, for* $n \geq 2$, $E_n = A_n \backslash A_{n-1} = A_n \backslash (A_1 \cup \cdots \cup A_{n-1})$
*Note that the* $E_i$ *are disjoint*
*Note also* $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} E_i$
*Last axiom tells us that*

$$\mathsf{P}(A_n) = \mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathsf{P}(E_i)$$

*And hence using last axiom again*

$$\lim_{n\to\infty} \mathsf{P}(A_n) = \sum_{i=1}^{\infty} \mathsf{P}(E_i) = \mathsf{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right)$$

# Cumulative distribution functions

## Proof of Lemma

*Note that* $1 = \mathsf{P}(A + A^c) = \mathsf{P}(A) + \mathsf{P}(A^c)$ *so*

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

*Now note that*

$$\omega \in \bigcup_{i=1}^{\infty} B_i^c \iff \omega \in B_j^c \text{ for some } j$$

*Hence*

$$\omega \in \left(\bigcup_{i=1}^{\infty} B_i^c\right)^c \iff \omega \notin B_j^c \text{ for all } j \iff \omega \in \bigcap_{i=1}^{\infty} B_i$$

# Cumulative distribution functions

## Proof of Lemma

*Hence*

$$\mathsf{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = \mathsf{P}\left(\left(\bigcup_{i=1}^{\infty} B_i^c\right)^c\right) = 1 - \mathsf{P}\left(\bigcup_{i=1}^{\infty} B_i^c\right)$$

*As* $B_1^c \subset B_2^c \subset B_3^c$ *because* $B_1 \supset B_2 \supset B_3 \cdots$
*we can use part (i) to deduce*

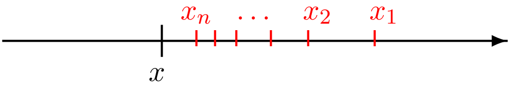
$$\mathsf{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = 1 - \lim_{n\to\infty} \mathsf{P}(B_n^c) = \lim_{n\to\infty} 1 - \mathsf{P}(B_n^c) = \lim_{n\to\infty} \mathsf{P}(B_n).$$

# Cumulative distribution functions

We are now ready to prove the rest of the Theorem.

Proof of (iv): $F_X(x)$ is right-continuous at $x$

Let $x_n \downarrow x$



Define

$$
\begin{aligned}
B_n &= \{\omega : X(\omega) \le x_n\}, \quad n = 1, 2, \dots, \\
B &= \{\omega : X(\omega) \le x\}.
\end{aligned}
$$

Then $B_1 \supset B_2 \supset \dots$ and, I claim,

$$\cap_{n=1}^{\infty} B_n = B.$$

# Cumulative distribution functions

SELF STUDY SLIDE

Proof of the claim that $\cap_{n=1}^{\infty} B_n = B$.

(a) Suppose $\omega \in B$,

then $X(\omega) \leq x \leq x_n$ for all $n$

so $\omega \in B_n$ for all $n$

and $\omega \in \cap_{n=1}^{\infty} B_n$.

(b) Suppose $\omega \in \cap_{n=1}^{\infty} B_n$,

then $X(\omega) \leq x_n$ for all $n$

therefore $X(\omega) \leq \lim_{n\to\infty} x_n = x$

so $\omega \in B$.

Thus, the claim is proved.

# Cumulative distribution functions

SELF STUDY SLIDE

Now we use Lemma (ii).

Since $B_1 \supset B_2 \supset \ldots$,

$$\mathsf{P}\{\cap_{n=1}^{\infty} B_n\} \;=\; \lim_{n\to\infty} \mathsf{P}(B_n) \tag{3}$$

The RHS of (3) is

$$\lim_{n\to\infty} \mathsf{P}(X \le x_n) \;=\; \lim_{n\to\infty} F_X(x_n).$$

The LHS of (3) is

$$\mathsf{P}(B) \;=\; \mathsf{P}(X \le x) \;=\; F_X(x).$$

So we have $\lim_{n\to\infty} F_X(x_n) \;=\; F_X(x)$, as required.

□

Proofs of (ii) and (iii): *see Problem Sheet 2.*

# Where we are

**I** Introduction – An example of a non-discrete probability space

**II** Random variables and cumulative distribution functions

**II.a** Probability density functions

**II.b** Expectation and variance

**II.c** Independence of random variables

**II.d** Cumulative distribution functions

In this lecture, we continue with section **II.d**.

# Cumulative distribution functions

## Example

**A discrete random variable**

Let $X$ be the number of Hs in 4 tosses of a fair coin.

| $x$ | $\mathsf{P}(X = x)$ |
|---|---|
| 0 | 1/16 |
| 1 | 4/16 |
| 2 | 6/16 |
| 3 | 4/16 |
| 4 | 1/16 |

The CDF is defined as

$$F_X(x) = \mathsf{P}(X \leq x), \quad x \in \mathsf{R}.$$

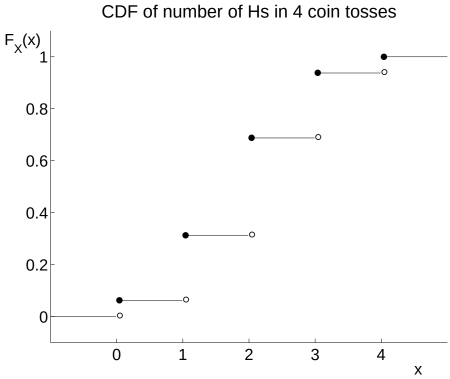
# Coin tossing example

The CDF is $F_X(x) = \mathsf{P}(X \leq x), \quad x \in \mathsf{R}$.

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1/16 & \text{for } 0 \leq x < 1 \\ 5/16 & \text{for } 1 \leq x < 2 \\ 11/16 & \text{for } 2 \leq x < 3 \\ 15/16 & \text{for } 3 \leq x < 4 \\ 1 & \text{for } x \geq 4 \end{cases}$$

# Coin tossing example

The CDF is $F_X(x) = \mathrm{P}(X \leq x), \quad x \in \mathrm{R}$.



The size of the jump at $x$ is $\mathrm{P}(X = x)$, for $x = 0$, 1, 2, 3 and 4.

# Cumulative distribution functions

## Example

**A continuous random variable**

Scrat the squirrel buries an acorn in a 1 metre square patch of earth.

Scrat chooses the location **uniformly** over the square:

Define co-ordinates in the range 0 to 1.

Then, for any set $A \subset [0,1] \times [0,1]$,

$$\mathrm{P}(\text{Acorn is buried in area } A) \ = \ \text{Area}(A).$$

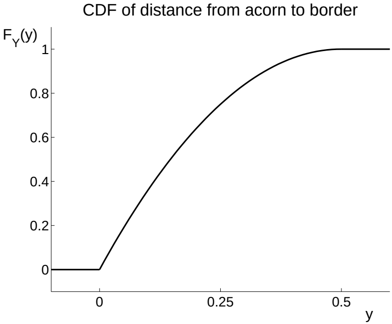
Let $Y$ be the distance from the acorn to the border of the square.

*Question:* What is the CDF of the random variable $Y$ ?

# Buried acorn example

*Answer:* The CDF of $Y$ is *See calculations on board*

$$F_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ 4y(1-y) & \text{for } 0 \le y \le 1/2 \\ 1 & \text{for } y > 1/2 \end{cases}$$



# Relating the PDF and CDF

We restrict attention here to the case of **continuous** RVs.

Recalling the definition of a PDF, the following relationship holds between the CDF $F_X(x)$ and PDF $f_X(x)$

$$F_X(x) \;=\; \mathsf{P}(X \leq x) \;=\; \int_{-\infty}^{x} f_X(u)\,du.$$

So, by the fundamental theorem of calculus,

$$f_X(x) \;=\; \frac{d}{dx} F_X(x). \tag{4}$$

Note: To be precise, property (4) holds where $f_X(x)$ is continuous.

*See calculations on board*

# Buried acorn example

With $Y =$ The distance from the acorn to the border of the square, we found the CDF

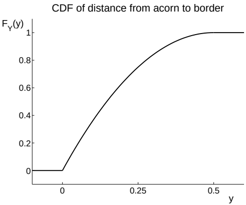
$$F_Y(y) \;=\; \begin{cases} 0 & \text{for } y < 0 \\ 4y - 4y^2 & \text{for } 0 \le y \le 1/2 \\ 1 & \text{for } y > 1/2. \end{cases}$$

Differentiating with respect to $y$, we find the PDF is

$$f_Y(y) \;=\; \begin{cases} 0 & \text{for } y < 0 \\ 4 - 8y & \text{for } 0 < y < 1/2 \\ 0 & \text{for } y > 1/2. \end{cases}$$

# Buried acorn example

We did not define the PDF at $y = 0$ or $y = 1/2$.



We cannot differentiate $F_Y(y)$ at $y = 0$.

Although the form of $F$ changes at $y = 1/2$, it does have a derivative there — of zero.

However, we can define $f_Y(0)$ and $f_Y(1/2)$ arbitrarily — since this will not affect any probabilities

$$\mathsf{P}(a \le Y \le b) \; = \; \int_a^b f_Y(y)\, dy.$$

# Buried acorn example

Let us find $\mathrm{P}(Y \geq 1/4)$ in two ways:

(a) Using the CDF

*See calculations on board*

(b) Using the PDF

*See calculations on board*

# Expressing P$(X < a)$

## Definition

If $F$ is a CDF and $a \in$ R, we define

$$F(a-) \ = \ \lim_{n\to\infty} F(a_n)$$

where $\{a_n\}$ is a sequence such that $a_n \uparrow a$.

Here, $a_n \uparrow a$ means that $\{a_n\}$ is an increasing sequence with $a_n < a$ for all $n$ and $\lim_{n\to\infty} a_n = a$.

# Expressing $\mathsf{P}(X < a)$

SELF STUDY SLIDE

## Lemma

*For any random variable $X$ and $a \in \mathsf{R}$,*

$$\mathsf{P}(X < a) = F_X(a-).$$

## Proof

*Check in your own time:*

*Let $a_n \uparrow a$.*

*Then*

$$\begin{aligned} \mathsf{P}(X < a) &= \mathsf{P}\left[\cup_{n=1}^{\infty}\{X \leq a_n\}\right] \\ &= \lim_{n\to\infty} P(X \leq a_n) \\ &= F_X(a-). \end{aligned}$$

□

# Where we are

**I** Introduction – An example of a non-discrete probability space

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**III.a** The uniform distribution

**III.b** The normal distribution

**III.c** The exponential distribution

**III.d** Some other families of continuous random variables

In the next four lectures, we shall explore examples of distributions.

Today, we shall consider the uniform distribution and make a start on the normal distribution.

# Terminology

By the **distribution** of a random variable, we mean

*for a discrete RV:*

its probability mass function,

*for a continuous RV:*

its PDF.

Alternatively, the CDF specifies a RV's distribution in both the discrete and continuous cases.

# III.a The uniform distribution

## Definition

The RV $X$ has a uniform distribution on $(a, b)$, denoted

$$X \sim \mathrm{Unif}(a, b) \quad \text{or} \quad X \sim U(a, b),$$

if it has PDF

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

The uniform distribution provides a model for an observation which must lie between $a$ and $b$, and for which all parts of this range are equally likely.

It is the archetypal "random" distribution.

# CDF of the uniform distribution

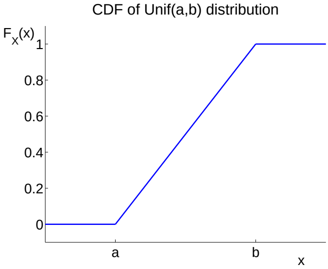
The CDF of a $\mathrm{Unif}(a,b)$ RV is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \int_{a}^{x} \frac{1}{b-a}\,du$$

$$= \frac{x-a}{b-a} \qquad \text{for } a \le x \le b.$$

So, for the whole range of $x$,

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \le x \le b, \\ 1 & \text{if } x > b. \end{cases}$$



# The uniform distribution

## Lemma

*Let* $U \sim \mathrm{Unif}(0,1)$ *and*

$$X = a + (b-a)U,$$

*where* $a \in \mathrm{R}$, $b \in \mathrm{R}$ *and* $a < b$.

*Then*

$$X \sim \mathrm{Unif}(a,b).$$

## Proof

*See calculations on board*

# Mean and variance of a $\mathrm{Unif}(0,1)$ RV

Let $U \sim \mathrm{Unif}(0,1)$. Then,

$$\mathsf{E}(U) \;=\; \int_{-\infty}^{\infty} u\, f_U(u) du \;=\; \int_0^1 u\, 1\, du \;=\; \frac{1}{2}.$$

Also,

$$\mathsf{E}(U^2) \;=\; \int_{-\infty}^{\infty} u^2\, f_U(u) du \;=\; \int_0^1 u^2\, 1\, du \;=\; \frac{1}{3},$$

so

$$\mathsf{Var}(U) \;=\; \mathsf{E}(U^2) - [\mathsf{E}(U)]^2 \;=\; \frac{1}{3} - \left(\frac{1}{2}\right)^2 \;=\; \frac{1}{12}.$$

# Mean and variance of a $\mathrm{Unif}(a,b)$ RV

If $X \sim \mathrm{Unif}(a,b)$, we can write this RV as

$$X \;=\; a + (b-a)U,$$

where $U \sim \mathrm{Unif}(0,1)$.

Hence,

$$\mathsf{E}(X) \;=\; \frac{1}{2}\,(a+b),$$

$$\mathsf{Var}(X) \;=\; \frac{(b-a)^2}{12}.$$

*See calculations on board*

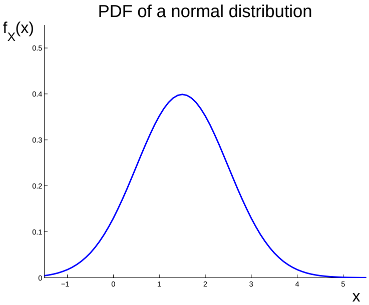
# Example of a uniform RV

A stick of length 50cm is broken in two at a random point, uniformly distributed along the stick.

Find the distribution of the longer piece of the stick.

*See calculations on board*

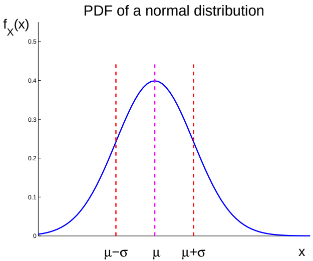
# III.b The normal distribution



The normal or "Gaussian" distribution is a common choice for modelling experimental data.

The normal distribution arises in theory as the limiting distribution of the sum of a large number of independent RVs.

# PDF of the normal distribution



About 68% of the $N(\mu, \sigma^2)$ distribution lies between $\mu - \sigma$ and $\mu + \sigma$,

about 95% lies between $\mu - 2\sigma$ and $\mu + 2\sigma$.

## Definition

The RV $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$, denoted $X \sim N(\mu,\, \sigma^2)$, if it has PDF

$$f_X(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left\{ \frac{-(x-\mu)^2}{2\,\sigma^2} \right\}, \quad x \in \mathsf{R}.$$

We shall see this does indeed imply mean $\mu$ and variance $\sigma^2$.

# PDF of the normal distribution

## Definition

The random variable $Z$ is said to be **standard normal** if it follows a $N(0,\,1)$ distribution.

Thus, if $Z$ is a standard normal RV, it has PDF

$$f_Z(z) \;=\; \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-z^2}{2}\right\}, \quad z \in \mathsf{R}.$$

It is implicit in this definition of the PDF of a $N(0,\,1)$ RV that

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{\frac{-z^2}{2}\right\}dz \;=\; 1.$$

This is true, although we shall not prove it here.

# Where we are

**I** Introduction – An example of a non-discrete probability space

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**III.a** The uniform distribution

**III.b** The normal distribution

**III.c** The exponential distribution

**III.d** Some other families of continuous random variables

Today, we continue to look at the normal distribution.

# The normal distribution

## Proposition

*Suppose* $X \sim N(\mu, \sigma^2)$ *and* $Y = a\,X + b$*, where* $a, b \in \mathrm{R}$*,* $a \neq 0$*.*

*Then*

$$Y \;\sim\; N(a\,\mu + b,\, a^2\,\sigma^2).$$

## Corollary

*If* $X \sim N(\mu, \sigma^2)$*, then* $X - \mu \sim N(0, \sigma^2)$ *and*

$$Z \;=\; \frac{X-\mu}{\sigma} \;\sim\; N(0,\, 1).$$

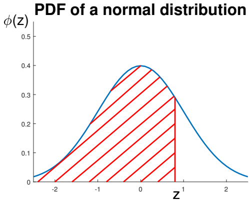
# CDF of the normal distribution

## Notation

*The CDF of the standard normal distribution is denoted by $\Phi(z)$. It is equal to*

$$\Phi(z) \;=\; \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du \;=\; \int_{-\infty}^{z} \phi(u)\, du,$$

*where $\phi$ is used to denote the standard normal PDF.*



The area under the curve to the left of $z$ is the CDF $\Phi(z)$.

By symmetry, $\Phi(-z) = 1 - \Phi(z)$ for $z \in$ R.

# Mean and variance of a $N(0,\,1)$ RV

We use the fact that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-z^2}{2}\right\} dz \;=\; 1.$$

*See calculations on board*

Let $Z \sim N(0,\,1)$.

Then,

$$\begin{aligned} \mathsf{E}(Z) \;&=\; \int_{-\infty}^{\infty} z\,\frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-z^2}{2}\right\} dz \\ &=\; \frac{1}{\sqrt{2\pi}} \left[-\exp\left\{\frac{-z^2}{2}\right\}\right]_{-\infty}^{\infty} \;=\; 0. \end{aligned}$$

# Mean and variance of a $N(0,\,1)$ RV

With $Z \sim N(0,\,1)$,

$$\begin{aligned} \mathsf{E}(Z^2) &= \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-z^2}{2}\right\} dz \\ &= \left[\frac{-1}{\sqrt{2\pi}}\, z \exp\left\{\frac{-z^2}{2}\right\}\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \left(\frac{-1}{\sqrt{2\pi}}\right) \exp\left\{\frac{-z^2}{2}\right\} dz \\ &= 0 + 1 = 1. \end{aligned}$$

Here, we have used integration by parts:

$$\int_a^b u(z)\, v'(z)\, dz = [u(z)\, v(z)]_a^b - \int_a^b u'(z)\, v(z)\, dz$$

with $u(z) = z$ and $v(z) = -\exp(-z^2/2)$, so $v'(z) = z\exp(-z^2/2)$.

# Mean and variance of normal RVs

For $Z \sim N(0,\, 1)$,

$$\mathsf{Var}(Z) \;=\; \mathsf{E}(Z^2) - (E(Z))^2 = 1 - 0^2 \;=\; 1.$$

Now consider $X \sim N(\mu,\, \sigma^2)$.

We can write $X = \mu + \sigma\, Z$, where $Z \sim N(0,\, 1)$.

Thus,

$$\mathsf{E}(X) \;=\; \mu + \sigma\, \mathsf{E}(Z) = \mu$$

and

$$\mathsf{Var}(X) \;=\; \sigma^2\, \mathsf{Var}(Z) \;=\; \sigma^2.$$

So we do have mean $\mu$ and variance $\sigma^2$.

# The deMoivre-Laplace Theorem

We state this theorem without proof.

### Theorem

*Let* $0 < p < 1$ *and* $S_n \sim \mathrm{Binom}(n, p)$.

*Then, for any* $a \in \mathsf{R}$,

$$\mathsf{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le a\right) \to \Phi(a) \quad \textit{as } n \to \infty.$$

Note that $(S_n - np)/\sqrt{np(1-p)}$ has mean zero and variance one.

# Application of the deMoivre-Laplace Theorem

## Example

A new diet is designed to reduce cholesterol levels.

A group of 200 subjects with high cholesterol are put into pairs and one in each pair is randomly chosen to receive the new diet.

In 65 of the 100 pairs, the patient on the new diet shows the greater reduction in cholesterol level.

Let $X$ be the RV denoting the number of pairs in which the patient on the new diet has the greater improvement.

If the new diet has no benefit, then $X \sim \mathrm{Binom}\,(100, 0.5)$ and, according to the deMoivre-Laplace Theorem, we can treat

$$\frac{X - 100 \times 0.5}{\sqrt{100 \times 0.5 \times 0.5}}$$

as approximately $N(0,\ 1)$.

# Application of the deMoivre-Laplace Theorem

Calculating probabilities for the case $X \sim \mathrm{Binom}\,(100, 0.5)$,

$$\begin{aligned} \mathsf{P}(X \geq 65) &= \mathsf{P}\left(\frac{X - 100 \times 0.5}{\sqrt{100 \times 0.5 \times 0.5}} \geq \frac{65 - 100 \times 0.5}{\sqrt{100 \times 0.5 \times 0.5}}\right) \\ &\approx \mathsf{P}(Z > 3.0) \;=\; 1 - \Phi(3.0) \;=\; 0.00135, \end{aligned}$$

where $Z$ denotes a $N(0,\, 1)$ random variable.

The value of $\Phi(3.0)$ is found by the R command **pnorm()**.

Since there is only a small probability of such a high value of $X$ if the new diet offers no advantage, we may conclude that the new diet has at least some beneficial effect.

# Continuity correction

Using the deMoivre-Laplace theorem, we approximate the distribution of $X \sim \mathrm{Binom}\,(n, p)$ by that of a normal RV

$$Y \sim N(np,\, np(1-p)).$$

Since $X$ takes integer values, it it is tricky to match its distribution to that of the continuous RV $Y$.

We can think of $X = x$ for the discrete $X$ as corresponding to $Y \in (x - 0.5, x + 0.5)$ for the continuous $Y$.

*See calculations on board*

Consequently, we match the events

$$X \leq x \quad \text{and} \quad Y \leq x + 0.5$$

and we match

$$X \geq x \quad \text{and} \quad Y \geq x - 0.5.$$

# Continuity correction

Applying this idea in making a normal approximation to a binomial probability is known as making a "continuity correction".

Using the continuity correction in our example we obtain:

$$\begin{aligned}
\mathsf{P}(X \geq 65) &= \mathsf{P}(X > 64.5) \\
&= \mathsf{P}\left(\frac{X - 100 \times 0.5}{\sqrt{100 \times 0.5 \times 0.5}} > \frac{64.5 - 100 \times 0.5}{\sqrt{100 \times 0.5 \times 0.5}}\right) \\
&\approx \mathsf{P}(Z > 2.9) \;=\; 1 - \Phi(2.9) \;=\; 0.00187.
\end{aligned}$$

Without the continuity correction, we obtained the answer 0.00135.

The true probability, using the full binomial calculation, is 0.00176 — so the continuity correction has helped.

# Where we are

**I** Introduction – An example of a non-discrete probability space

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**III.a** The uniform distribution

**III.b** The normal distribution

**III.c** The exponential distribution

**III.d** Some other families of continuous random variables

Today, we shall learn about the exponential distribution.

# III.c The exponential distribution

## Definition

The RV $X$ has an exponential distribution with rate parameter $\lambda\ (> 0)$, denoted $\mathrm{Exp}(\lambda)$, if it has PDF

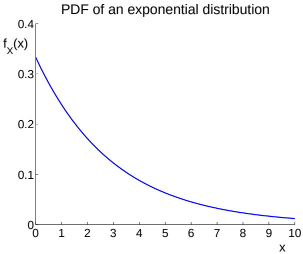
$$f_X(x) \;=\; \begin{cases} \lambda \exp\{-\lambda\, x\} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

**Uses of the exponential distribution**

Survival times (medical)

Failure times (industrial)

Waiting times



# CDF of the $\mathrm{Exp}(\lambda)$ distribution

For $x < 0$, the CDF is $F_X(x) = 0$ .

For $x \geq 0$, the CDF is

$$\begin{aligned} F_X(x) &= \int_0^x f_X(u)\,du \\ &= \int_0^x \lambda\,\exp\{-\lambda\,u\}\,du \\ &= [\,-\exp\{-\lambda\,u\}\,]_0^x \\ &= 1 - \exp\{-\lambda\,x\}. \end{aligned}$$

Note this implies $F_X(x) \to 1$ as $x \to \infty$.

So, we have checked that the PDF does integrate to one.

# Mean and variance of an $\mathrm{Exp}(\lambda)$ RV

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

$$\mathsf{E}(X) = \frac{1}{\lambda},$$

$$\mathsf{E}(X^2) = \frac{2}{\lambda^2}.$$

Hence,

$$\mathsf{Var}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 = \frac{1}{\lambda^2}.$$

*See calculations on board*

# The memoryless property of the exponential distribution

Suppose $X \sim \mathrm{Exp}(\lambda)$ and consider the conditional probability that $X > t + s$ given that $X > t$, where $t > 0$ and $s > 0$.

Recall the definition of conditional probability,

$$\mathsf{P}(A \,|\, B) \;=\; \frac{\mathsf{P}(A \text{ and } B)}{\mathsf{P}(B)}.$$

So,

$$\begin{aligned}
\mathsf{P}\{X > t + s \,|\, X > t\} &= \frac{\mathsf{P}(X > t + s \;\text{ and }\; X > t)}{\mathsf{P}(X > t)} \\
&= \frac{\mathsf{P}(X > t + s)}{\mathsf{P}(X > t)} = \frac{1 - F_X(t + s)}{1 - F_X(t)} \\
&= \frac{\exp\{-\lambda\,(t + s)\}}{\exp\{-\lambda\, t\}} = \exp\{-\lambda\, s\}.
\end{aligned}$$

# The memoryless property of the exponential distribution

Since

$$\mathsf{P}\{X > t + s \,|\, X > t\} \;=\; \exp\{-\lambda\, s\}$$

does not depend on $t$, we say the exponential distribution is memoryless.

## Example

An angler knows that the waiting time (in minutes) before he catches a fish follows an $\mathrm{Exp}(0.02)$ distribution.

How long does he have to wait to have a probability of $0.5$ of catching a fish?

Suppose he has been waiting 30 minutes and not yet caught a fish, how much longer does he need to wait to have a $0.5$ probability of catching a fish?

*See calculations on board*

# The “hazard rate” of the exponential distribution

## Definition

The **hazard rate** at time $t$ of a survival time distribution is

$$h(t) \;=\; \lim_{\delta t \downarrow 0} \frac{1}{\delta t}\, \mathsf{P}\{X \in (t, t+\delta t] \,|\, X > t\},$$

where $X$ is a RV following the specified distribution.

This can be viewed as the instantaneous rate of failure at time $t$, given survival up to time $t$.

The hazard rate is a very natural property of a lifetime distribution.

Risks from particular hazards are often expressed in terms of a hazard rate: for example, in the statement

> “The rate of incidence of lung cancer is higher by a factor $n$ for smokers than for non-smokers”.

# The hazard rate of a survival distribution

If $f_X$ is continuous at $t$,

$$\mathsf{P}\{X \in (t, t+\delta t] \text{ and } X > t\} \;=\; \mathsf{P}\{X \in (t, t+\delta t]\} \approx f_X(t)\,\delta t.$$

Also,

$$\mathsf{P}(X > t) \;=\; 1 - F_X(t).$$

Thus, we have

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{1}{\delta t}\, \mathsf{P}\{X \in (t, t+\delta t] \,|\, X > t\}, \\ &= \lim_{\delta t \to 0} \frac{1}{\delta t}\, \frac{\mathsf{P}\{X \in (t, t+\delta t]\}}{1 - F_X(t)}, \\ &= \lim_{\delta t \to 0} \frac{\mathsf{P}\{X \in (t, t+\delta t]\}}{\delta t}\, \frac{1}{1 - F_X(t)} \;=\; \frac{f_X(t)}{1 - F_X(t)}. \end{aligned}$$

# The "hazard rate" of the exponential distribution

For an exponential RV, $X \sim \mathrm{Exp}(\lambda)$,

$$h(t) \;=\; \frac{f_X(t)}{1 - F_X(t)} \;=\; \frac{\lambda \exp\{-\lambda\, t\}}{\exp\{-\lambda\, t\}} \;=\; \lambda.$$

The constant hazard rate is in keeping with the memoryless property.

Working in the other direction, suppose we know a positive RV follows a continuous, "memoryless" distribution, and so has a constant hazard rate.

What distribution does this RV follow?

# The “hazard rate” of the exponential distribution

Let $X$ be a positive and continuous RV with hazard rate $h(t) = k$. Then,

$$\frac{f_X(t)}{1 - F_X(t)} = k,$$

$$\int_0^x \frac{f_X(t)}{1 - F_X(t)}\,dt = \int_0^x k\,dt,$$

and

$$[\,-\log\{1 - F_X(t)\}\,]_0^x = k\,x.$$

It follows that

$$-\log\{1 - F_X(x)\} = k\,x$$

and

$$F_X(x) = 1 - \exp(-k\,x),$$

so $X$ is an $\mathrm{Exp}(k)$ random variable — which is why I referred to $k$ as the “rate parameter”.

# Where we are

**I** Introduction – An example of a non-discrete probability space

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

- **III.a** The uniform distribution
- **III.b** The normal distribution
- **III.c** The exponential distribution
- **III.d** Some other families of continuous random variables

Today, we shall learn about the Gamma and Weibull distributions.

# III.d The Gamma distribution

First, we need to define the **Gamma function**.

## Definition

The Gamma function is defined for $t > 0$ as

$$\Gamma(t) \;=\; \int_0^\infty x^{t-1}\, e^{-x}\, dx.$$

Note that

$$\begin{aligned}
\Gamma(t+1) \;&=\; \int_0^\infty x^t\, e^{-x}\, dx \\
&=\; \left[ -x^t\, e^{-x} \right]_0^\infty + \int_0^\infty t\, x^{t-1}\, e^{-x}\, dx \\
&=\; t\,\Gamma(t) \;=\; t\,(t-1)\,\Gamma(t-1) \;=\; \ldots\; .
\end{aligned}$$

# The Gamma function

For $t = 1$,

$$\Gamma(1) \;=\; \int_0^\infty e^{-x}\,dx \;=\; 1.$$

For integer values of $t$, $\Gamma(t+1) = t\,\Gamma(t)$ and $\Gamma(1) = 1$ imply

$$\Gamma(t) \;=\; (t-1)!$$

For $t = 1/2$,

$$\Gamma(1/2) \;=\; \int_0^\infty x^{-1/2} e^{-x}\,dx \;=\; \sqrt{\pi}$$

— to see this, make a change of variable to $y = \sqrt{2x}$ and use

$$\int_0^\infty \frac{1}{\sqrt{2\pi}}\,\exp(-y^2/2)\,dy \;=\; \frac{1}{2}.$$

# The Gamma distribution

## Definition

The RV $X$ has a Gamma distribution with parameters $\lambda$ and $k$ ($\lambda > 0$, $k > 0$), denoted $\mathrm{Gamma}\,(\lambda, k)$, if it has PDF

$$f_X(x) \;=\; \begin{cases} \frac{1}{\Gamma(k)}\,\lambda^k\,x^{k-1}\,\exp\{-\lambda\,x\} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We can check:

$$\begin{aligned} \int_0^\infty f_X(x)\,dx \;&=\; \int_0^\infty \frac{1}{\Gamma(k)}\,\lambda^k\,x^{k-1}\,\exp\{-\lambda\,x\}\,dx \\ &=\; \int_0^\infty \frac{1}{\Gamma(k)}\,u^{k-1}\,\exp\{-u\}\,du \;=\; 1 \end{aligned}$$

— substituting $u = \lambda\,x$, with "$du = \lambda\,dx$".

# Mean and variance of the $\mathrm{Gamma}\,(\lambda, k)$ distribution

If $X \sim \mathrm{Gamma}\,(\lambda, k)$,

$$\mathsf{E}(X) \;=\; \frac{k}{\lambda}$$

$$\mathsf{E}(X^2) \;=\; \frac{(k+1)\,k}{\lambda^2}$$

Hence,

$$\mathsf{Var}(X) \;=\; \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 \;=\; \frac{(k+1)\,k}{\lambda^2} - \left\{\frac{k}{\lambda}\right\}^2 \;=\; \frac{k}{\lambda^2}$$

*See calculations on board*

# Properties of the Gamma distribution

**Scaling**

$$X \sim \mathrm{Gamma}\,(\lambda, k) \;\Rightarrow\; Y = c\,X \sim \mathrm{Gamma}\,(\lambda/c, k).$$

Proof.

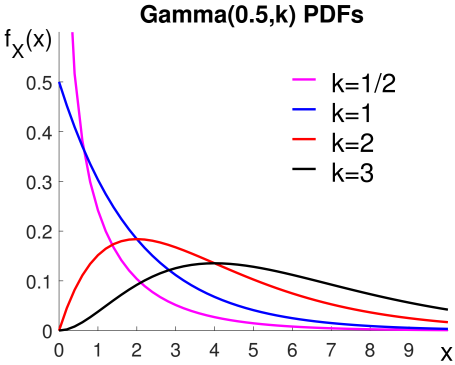
*See calculations on board*

□

The parameter $\lambda$ serves to scale the Gamma distribution, but the mean is *inversely* proportional to $\lambda$.

So, the role of $\lambda$ is similar to that of the rate parameter in the exponential distribution — and we shall see more of a connection in due course.

# Shape of the Gamma distribution



The parameter $k$ determines the shape of the Gamma distribution.

# Relation between the Gamma and exponential distributions

Note that the $\mathrm{Gamma}\,(\lambda, 1)$ distribution has density

$$f_X(x) \;=\; \lambda\, e^{-\lambda\, x} \quad \text{for } x \geq 0,$$

and so is an $\mathrm{Exp}(\lambda)$ distribution.

*We state but do not prove here that:*

If $X_1, \ldots, X_k$ are independent $\mathrm{Exp}(\lambda)$ RVs, then

$$X_1 + \ldots + X_k \;\sim\; \mathrm{Gamma}\,(\lambda, k).$$

*Hence*, for integers $k_1$ and $k_2$, if $Y_1 \sim \mathrm{Gamma}\,(\lambda, k_1)$ and $Y_2 \sim \mathrm{Gamma}\,(\lambda, k_2)$ are independent RVs, then

$$Y_1 + Y_2 \;\sim\; \mathrm{Gamma}\,(\lambda, k_1 + k_2).$$

(Express $Y_1$ and $Y_2$ as sums of independent $\mathrm{Gamma}\,(\lambda, 1)$ RVs.)

# Relation between the Gamma and normal distributions

*We state but do not prove here that:*

1. If $X \sim N(0,1)$, then $X^2 \sim \mathrm{Gamma}\,(\frac{1}{2},\frac{1}{2})$.

2. If $X_1,\ldots,X_k$ are independent $N(0,1)$ RVs, then

$$X_1^2+\ldots+X_k^2 \sim \mathrm{Gamma}\left(\frac{1}{2},\frac{k}{2}\right)$$

— this is also known as the $\chi_k^2$ distribution, or $\chi^2$ (chi squared) distribution on $k$ degrees of freedom.

Thus, if $X_1$ and $X_2$ are independent $N(0,1)$ RVs, $X_1^2+X_2^2$ is the sum of two independent $\mathrm{Gamma}\,(\frac{1}{2},\frac{1}{2})$ RVs, so

$$X_1^2+X_2^2 \sim \mathrm{Gamma}\left(\frac{1}{2},1\right) \sim \mathrm{Exp}\left(\frac{1}{2}\right).$$

# The Weibull distribution

## Definition

The RV $X$ has a Weibull distribution with parameters $\lambda$ and $\beta$, denoted $\mathrm{Weib}(\lambda, \beta)$, if it has PDF

$$f_X(x) \;=\; \begin{cases} \lambda\,\beta\,x^{\beta-1}\,\exp\{-\lambda\,x^{\beta}\} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The CDF of the $\mathrm{Weib}(\lambda, \beta)$ distribution is

$$F_X(x) \;=\; \begin{cases} 0 & x < 0, \\ 1 - \exp\{-\lambda\,x^{\beta}\} & x \geq 0. \end{cases}$$

# The Weibull distribution

The hazard rate of the Weibull distribution at time $t$ is

$$h(t) \;=\; \frac{f_X(t)}{1 - F_X(t)} \;=\; \frac{\lambda\,\beta\,t^{\beta-1}\;\exp\{-\lambda\,t^{\beta}\}}{\exp\{-\lambda\,t^{\beta}\}} \;=\; \lambda\,\beta\,t^{\beta-1}.$$

The value of $\beta$ shapes the hazard rate function $h(t)$.

For $\beta = 1$, $h(t)$ is constant and the $\mathrm{Weib}(\lambda, \beta)$ distribution is also the $\mathrm{Exp}(\lambda)$ distribution,

For $\beta > 1$, $h(t)$ increases with $t$ — old items are more likely to fail than new ones,

For $\beta < 1$, $h(t)$ decreases with $t$ — old items are less likely to fail than new ones.

# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**IV.a** Joint PDFs, marginal distributions and independence

**IV.b** Conditional PDFs

**IV.c** More on expectation

**IV.d** Covariance

Today, we start on joint distributions and marginal distributions.

# IV.a Joint PDFs

The joint distribution for **discrete** RVs is easily defined.

As an example, suppose we roll a die and set

$$Y = \text{Score shown on the die.}$$

Then we toss a coin once if $Y$ is odd and twice if $Y$ is even, and set

$$X = \text{Number of Heads obtained.}$$

The joint distribution of $(X, Y)$ is given by the table of probabilities P$(X = x$ and $Y = y)$:

| | | $y$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | 0 | 1/12 | 1/24 | 1/12 | 1/24 | 1/12 | 1/24 |
| $x$ | 1 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 |
| | 2 | 0 | 1/24 | 0 | 1/24 | 0 | 1/24 |

# Joint PDFs

With continuous RVs, we need a multi-dimensional version of the probability density function that we have seen for a single RV.

We start by considering just two RVs, $X$ and $Y$ say.

The **joint PDF** needs to be able to capture connections between the RVs $X$ and $Y$, as seen in the previous discrete example.

Two RVs may vary together in a systematic way. E.g.,

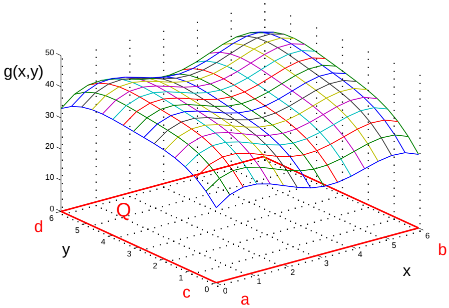
$$X = \textit{Height}, \;\; Y = \textit{Weight} \quad \textit{of the same individual.}$$

Possible values of one RV may depend on the value of the other. E.g., $Y \leq X$ if

$$\begin{aligned} X &= \textit{Time spent working on Example Sheets,} \\ Y &= \textit{Time spent working on MA10212 Example Sheets.} \end{aligned}$$

As a preliminary, we need to discuss **double integrals**.

# Double integrals

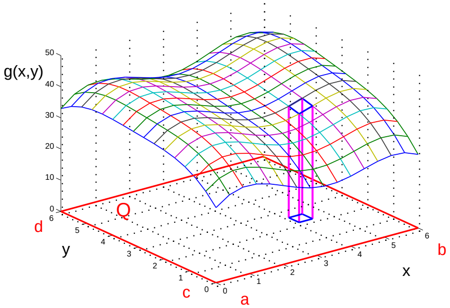


Consider a function $g$: $Q \to$ R, where $Q = (a,b) \times (c,d) \subset$ R$^2$.

We define

$$\int\!\!\int_Q g(x,y)\,dy\,dx = \int_a^b \left[\int_c^d g(x,y)dy\right] dx = \int_c^d \left[\int_a^b g(x,y)dx\right] dy.$$

# Double integrals



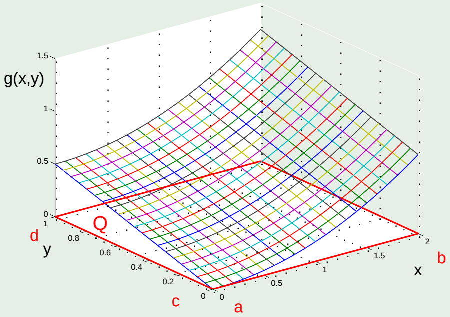
The double integral represents the limit of a sum of volumes under the surface $g(x, y)$ — as the grid becomes finer and finer.

The order of integration — over $x$ first or over $y$ first — corresponds to the order of summation of these volumes.

For a well-behaved function $g$, the answer is the same either way.

# Double integrals

## Example



$Q = (0,2) \times (0,1)$ and

$$g(x,y) \;=\; \frac{3}{16}\,x^2 + \frac{1}{2}\,y.$$

# Double integrals

With $Q = (0,2) \times (0,1)$ and $g(x,y) \;=\; \frac{3}{16}\,x^2 + \frac{1}{2}\,y$,

$$\int\int_Q g(x,y)\,dy\,dx \;=\; \int_0^2 \left\{\int_0^1 \left(\frac{3}{16}\,x^2 + \frac{1}{2}\,y\right)\,dy\right\}\,dx \;=\; 1.$$

*See calculations on board*

**Exercise:** Check you get the same answer by writing this double integral as

$$\int\int_Q g(x,y)\,dx\,dy \;=\; \int_0^1 \left\{\int_0^2 \left(\frac{3}{16}\,x^2 + \frac{1}{2}\,y\right)\,dx\right\}\,dy$$

and integrating with respect to $x$ first, then with respect to $y$.

# Double integrals

We can allow the region of integration to be something other than a rectangle $(a, b) \times (c, d)$.

## Example

Consider

$$T \ = \ \{(x, y)\colon\ 0 \le y \le x \le 1\}$$

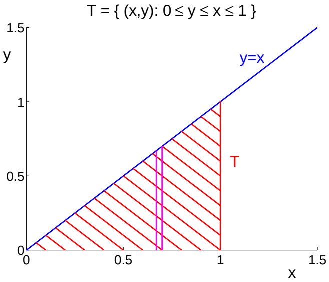
and

$$g(x, y) \ = \ 2\, e^{-x}\, e^{-2y}.$$

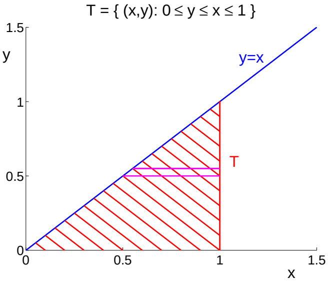
Find

$$\int\!\!\int_T g(x, y)\, dx\, dy.$$

# Double integrals



or



**Exercise:** Evaluate the integral of $g(x,y)$ over $T$ by writing

$$\int\int_T g(x,y)\,dx\,dy \;=\; \int_0^1 \left\{\int_0^x g(x,y)\,dy\right\} dx \;=\; \ldots$$

and also by

$$\int\int_T g(x,y)\,dx\,dy \;=\; \int_0^1 \left\{\int_y^1 g(x,y)\,dx\right\} dy \;=\; \ldots\,.$$

# Joint distribution for continuous RVs

## Definition

We say $X$ and $Y$ are jointly continuous RVs if there exists a function $f_{X,Y}(x,y)$: $\mathsf{R}^2 \to [0,\infty)$ such that for any region $A \subset \mathsf{R}^2$,

$$\mathsf{P}\{(X,Y) \in A\} \; = \int\int_A f_{X,Y}(x,y)\,dy\,dx.$$

The function $f_{X,Y}(x,y)$ is called the joint PDF of $X$ and $Y$.

## Example

$$f_{X,Y}(x,y) \; = \begin{cases} \frac{3}{16}\,x^2 + \frac{1}{2}\,y & \text{if } 0 < x < 2 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The function $f_{X,Y}(x,y)$ is positive everywhere and we have seen that it integrates to 1.

# Marginal distributions

## Proposition

*If $X$ and $Y$ have joint PDF $f_{X,Y}(x,y)$, then $X$ and $Y$ have PDFs*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \quad \textit{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

*Here, $f_X(x)$ and $f_Y(y)$ are called the* **marginal PDFs** *of $X$ and $Y$.*

## Proof.

$$\begin{aligned} \mathsf{P}(a \le X \le b) &= \mathsf{P}(a \le X \le b \text{ and } -\infty < Y < \infty) \\ &= \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = \int_a^b f_X(x)\,dx \end{aligned}$$

for $f_X(x)$ as defined above. So $f_X(x)$ is the PDF of $X$.

The proof that $f_Y(y)$ is the PDF of $Y$ follows similarly. □

# Marginal distributions

## Example

Consider the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{3}{16}\,x^2 + \frac{1}{2}\,y & \text{if } 0 < x < 2 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The marginal PDF of $X$ is as follows:

$$\begin{aligned} f_X(x) &= 0 \quad \text{for } x \le 0 \text{ or } x \ge 2, \\ f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^1 \left(\frac{3}{16}\,x^2 + \frac{1}{2}\,y\right) dy \\ &= \left[\frac{3}{16}\,x^2\,y + \frac{1}{4}\,y^2\right]_0^1 = \frac{3}{16}\,x^2 + \frac{1}{4} \quad \text{for } 0 < x < 2. \end{aligned}$$

# Marginal distributions

For the same joint PDF,

$$f_{X,Y}(x,y) \;=\; \begin{cases} \frac{3}{16}\,x^2 + \frac{1}{2}\,y & \text{if } 0 < x < 2 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise,} \end{cases}$$

the marginal PDF of $Y$ is:

$$\begin{aligned} f_Y(y) &= 0 \quad \text{for } y \le 0 \text{ or } y \ge 1, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \;=\; \int_0^2 \left(\frac{3}{16}\,x^2 + \frac{1}{2}\,y\right)\,dx \\ &= \left[\frac{x^3}{16} + \frac{1}{2}\,x\,y\right]_0^2 \;=\; \frac{1}{2} + y \quad \text{for } 0 < y < 1. \end{aligned}$$

# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**IV.a** Joint PDFs, marginal distributions and independence

**IV.b** Conditional PDFs

**IV.c** More on expectation

**IV.d** Covariance

Today, we consider independence of two or more RVs in terms of their joint PDF and marginal PDFs.

# IV.a Independence

## Definition

The joint CDF of RVs $X$ and $Y$ is the function $F_{X,Y}\colon \mathsf{R}^2 \to [0,1]$ defined by

$$F_{X,Y}(x,y) \;=\; \mathsf{P}(X \leq x \text{ and } Y \leq y).$$

Recall that the marginal CDFs are

$$F_X(x) \;=\; \mathsf{P}(X \leq x) \quad \text{and} \quad F_Y(y) \;=\; \mathsf{P}(Y \leq y).$$

## Definition (from Lecture 3)

*The RVs $X$ and $Y$ are independent if*

$$\mathsf{P}(X \leq x,\ Y \leq y) \;=\; \mathsf{P}(X \leq x)\,\mathsf{P}(Y \leq y) \;\; \textit{for all } x \in \mathsf{R},\ y \in \mathsf{R},$$

*i.e., if*

$$F_{X,Y}(x,y) \;=\; F_X(x)\,F_Y(y) \quad \textit{for all } x \in \mathsf{R},\ y \in \mathsf{R}.$$

# Independence and joint PDFs

## Theorem

*The jointly continuous RVs $X$ and $Y$ are independent if and only if*

$$f_{X,Y}(x,y) \;=\; f_X(x)\, f_Y(y) \quad \textit{for all } x \in \mathrm{R},\ y \in \mathrm{R}. \tag{5}$$

Note: Strictly speaking, we should allow $f_{X,Y}$, $f_X$ and $f_Y$ to fail to satisfy (5) on a finite set of points — since changing the value of a density at a point does not affect the probability of any event.

## Proof.

(i) We need to show that (5) implies independence, according to the definition in terms of $F_{X,Y}$, $F_X$ and $F_Y$.

(ii) We also need to show that the definition of independence, in terms of $F_{X,Y}$, $F_X$ and $F_Y$, implies (5).

□

*See calculations on board*

# Independence of RVs

## Example

Consider the joint PDF

$$f_{X,Y}(x,y) \; = \begin{cases} 2\,e^{-x}\,e^{-2y} & \text{if } x > 0 \text{ and } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Integrating over $y$ gives

$$f_X(x) \; = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

and integrating over $x$ gives

$$f_Y(y) \; = \begin{cases} 2\,e^{-2y} & y > 0 \\ 0 & y \leq 0. \end{cases}$$

# Independence of RVs

## Example, continued

*Since*

$$f_{X,Y}(x,y) \;=\; f_X(x)\,f_Y(y) \quad \textit{for all } x \in \mathsf{R},\ y \in \mathsf{R},$$

*the Theorem tells us that $X$ and $Y$ are independent.*

By inspection, the marginal distribution of $X$ is $\mathrm{Exp}(1)$ and the marginal distribution of $Y$ is $\mathrm{Exp}(2)$.

Thus, the pair $(X,Y)$ is made up of two independent exponential RVs, with different rate parameters.

In fact, we can see that $f_{X,Y}(x,y)$ factorises into one term involving $x$ only and another term involving $y$ only — and this is enough to imply independence of $X$ and $Y$.

# Independence of RVs

## Example

Consider the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{3}{16}\,x^2 + \frac{1}{2}\,y & \text{if } 0 < x < 2 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Integrating over $y$ gives

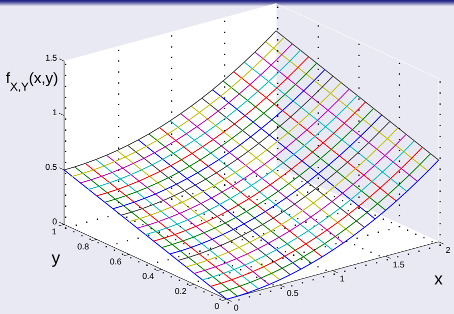
$$f_X(x) = \begin{cases} \frac{3}{16}\,x^2 + \frac{1}{4} & 0 < x < 2 \\ 0 & x \leq 0 \ \text{ or } \ x \geq 2 \end{cases}$$

and integrating over $x$ gives

$$f_Y(y) = \begin{cases} \frac{1}{2} + y & 0 < y < 1 \\ 0 & y \leq 0 \ \text{ or } \ y \geq 1. \end{cases}$$

# Independence of RVs

## Example, continued



*The Theorem tells us that $X$ and $Y$ are not independent, since*

$$f_{X,Y}(x,y) \neq f_X(x)\, f_Y(y) \quad \textit{for all } x \in \mathrm{R},\ y \in \mathrm{R}.$$

*This is also evident from the form of $f_{X,Y}(x,y)$ and from the plot.*

# More than two RVs

## Definition

We say $X_1, \ldots, X_n$ are jointly continuous RVs if there exists a function $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$: $\mathrm{R}^n \to [0,\infty)$ such that for any region $A \subset \mathrm{R}^n$,

$$\mathrm{P}\{(X_1, \ldots, X_n) \in A\} = \int \cdots \int_A f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_1 \ldots dx_n.$$

The function $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ is called the joint PDF of $X_1, \ldots, X_n$.

If $X_1, \ldots, X_n$ are independent,

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

# More than two RVs

## Example

Suppose $X_1, \ldots, X_n$ are independent RVs, each following an $\mathrm{Exp}(\lambda)$ distribution.

Then

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

$$= \begin{cases} \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i) & \text{if } x_1 > 0, \ldots, x_n > 0, \\ 0 & \text{otherwise.} \end{cases}$$

# Properties of joint CDFs

## Theorem

*If $X$ and $Y$ are two RVs with joint CDF $F_{X,Y}(x,y)$, then*

*(i) For all $x \in$ R,*

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

*(ii) For all $y \in$ R,*

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

*(iii)* $\lim_{x,\,y\to\infty} F_{X,Y}(x,y) = 1.$

*(iv) If $x_n \downarrow x$ and $y_n \downarrow y$,* $\lim_{n\to\infty} F_{X,Y}(x_n,y_n) = F_{X,Y}(x,y)$.

*(v) For every $a < b$ and $c < d$,*

$$F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c) \geq 0.$$

# Properties of joint CDFs

SELF STUDY SLIDE

## Proof

*The methods used here are similar to those used in Lecture 3 to prove properties of the CDF of a single RV.*

*In doing this, we also call on the Lemma from that lecture.*

*You should check the details fully in your own time.*

*(i) Pick a sequence $y_n \uparrow \infty$, then by part (i) of the Lemma*

$$\begin{aligned}\lim_{n\to\infty} \mathsf{P}(X \le x, Y \le y_n) &= \mathsf{P}\left(\cup_{n=1}^{\infty} \{X \le x, Y \le y_n\}\right)\\ &= \mathsf{P}(X \le x) = F_X(x).\end{aligned}$$

*Now pick a sequence $y_n \downarrow -\infty$, then by part (ii) of the Lemma*

$$\begin{aligned}\lim_{n\to\infty} \mathsf{P}(X \le x, Y \le y_n) &= \mathsf{P}\left(\cap_{n=1}^{\infty} \{X \le x, Y \le y_n\}\right)\\ &= \mathsf{P}(\emptyset) = 0.\end{aligned}$$

# Properties of joint CDFs

SELF STUDY SLIDE

## Proof, continued

*The proof of (ii) is similar to that of (i) with the roles of $X$ and $Y$ interchanged.*

*The proof of (iii) is similar to that of (i) and (ii) but we let both $x_n \uparrow \infty$ and $y_n \uparrow \infty$ and note that*

$$\cup_{n=1}^{\infty} \{X \leq x_n, Y \leq y_n\} = \Omega.$$

*(iv) We use the fact that*

$$\cap_{n=1}^{\infty} \{X \leq x_n, Y \leq y_n\} = \{X \leq x, Y \leq y\}.$$

*(v) The expression is equal to* $\mathsf{P}(a < X \leq b, c < Y \leq d)$, *which is greater than or equal to zero.*

# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**IV.a** Joint PDFs, marginal distributions and independence

**IV.b** Conditional PDFs

**IV.c** More on expectation

**IV.d** Covariance

Today, we consider the conditional distribution of one continuous RV given the value of another.

# IV.b Conditional PDFs

Recall the definition of the conditional probability of event A given event B

$$\mathsf{P}(A \,|\, B) \;=\; \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} \quad \text{if } \mathsf{P}(B) > 0.$$

So, for discrete RVs

$$\mathsf{P}(Y = y \,|\, X = x) \;=\; \frac{\mathsf{P}(Y = y,\, X = x)}{\mathsf{P}(X = x)} \quad \text{if } P_X(x) > 0.$$

We want a similar function for continuous RVs, which captures how the value of one RV, $X$, is affected by that of another RV, $Y$.

# Conditional PDFs

## Definition

Let $X$ and $Y$ be continuous RVs with joint PDF $f_{X,Y}(x,y)$ and marginal PDFs $f_X(x)$ and $f_Y(y)$.

For a value $x$ with $f_X(x) > 0$, the **conditional PDF** of $Y$ given $X = x$, written as $f_{Y|X}(y|x)$ or $f_{Y|X}(y \,|\, X = x)$, is

$$f_{Y|X}(y|x) \;=\; f_{Y|X}(y \,|\, X = x) \;=\; \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{for } y \in \mathrm{R}.$$

Similarly, for values $y$ with $f_Y(y) > 0$, **the conditional PDF** of $X$ given $Y = y$ is

$$f_{X|Y}(x|y) \;=\; f_{X|Y}(x \,|\, Y = y) \;=\; \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{for } x \in \mathrm{R}.$$

# Conditional PDFs

Note that $f_{Y|X}(y|x)$ is a PDF for $Y$ since it is positive and

$$\begin{aligned}
\int_{-\infty}^{\infty} f_{Y|X}(y|x)\,dy &= \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_X(x)}\,dy \\
&= \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\
&= \frac{1}{f_X(x)}\, f_X(x) \;=\; 1.
\end{aligned}$$

Similarly, $f_{X|Y}(x|y)$ is a PDF for $X$.

Motivation of this definition:

*See calculations on board*

# Conditional PDFs

For the distribution of $Y$ given $X = x$, we can define conditional probabilities, the conditional CDF, and conditional expectation — based on the **conditional PDF** $f_{Y|X}(y|x)$.

**Conditional probabilities given** $X = x$ are of the form

$$\mathsf{P}(a \leq Y \leq b \,|\, X = x) \;=\; \int_a^b f_{Y|X}(y|x)\,dy.$$

The **conditional CDF of** $Y$ **given** $X = x$ is

$$F_{Y|X}(y|x) \;=\; \mathsf{P}(Y \leq y \,|\, X = x) \;=\; \int_{-\infty}^{y} f_{Y|X}(u|x)\,du.$$

The **conditional expectation of** $Y$ **given** $X = x$ is

$$\mathsf{E}(Y \,|\, X = x) \;=\; \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\,dy,$$

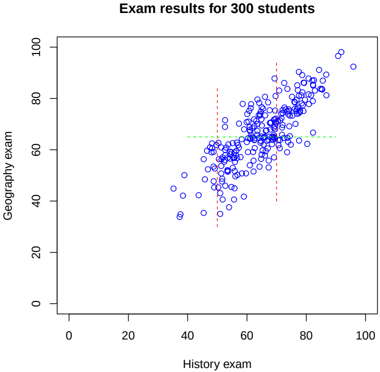
when this integral exists.

# Conditional PDFs

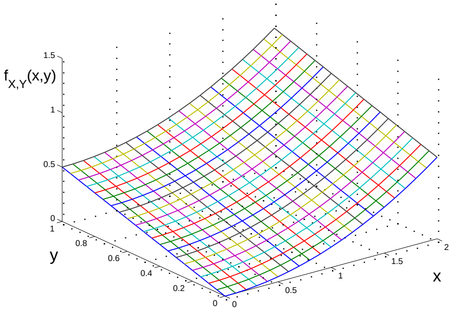
## Example

In a large college, 300 students sit exams in History and Geography.

Those scoring highly in one exam tend to do well in the other.



# Conditional PDFs



## Example

Consider the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{3}{16}\,x^2 + \frac{1}{2}\,y & \text{if } 0 < x < 2 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

# Conditional PDFs

## Example, continued

*In lecture 10, we derived the marginal PDFs for this* $f_{X,Y}(x,y)$*,*

$$f_X(x) \;=\; \begin{cases} \frac{3}{16}x^2+\frac{1}{4} & 0<x<2 \\ 0 & x\le 0 \;\textit{ or }\; x\ge 2 \end{cases}$$

*and*

$$f_Y(y) \;=\; \begin{cases} \frac{1}{2}+y & 0<y<1 \\ 0 & y\le 0 \;\textit{ or }\; y\ge 1. \end{cases}$$

*Hence, the conditional PDF for* $Y$ *when* $X=x\in(0,2)$ *is*

$$f_{Y|X}(y|x) \;=\; \frac{f_{X,Y}(x,y)}{f_X(x)} \;=\; \frac{3x^2/16+y/2}{3x^2/16+1/4} \;=\; \frac{3x^2+8y}{3x^2+4}$$

*for* $0<y<1$*, and 0 otherwise.*

# Conditional PDFs

### Example, continued

*Similarly, the conditional PDF for $X$ when $Y = y \in (0,1)$ is*

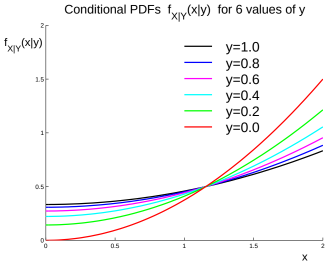
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{3x^2/16 + y/2}{1/2 + y} = \frac{3x^2 + 8y}{8 + 16y}$$

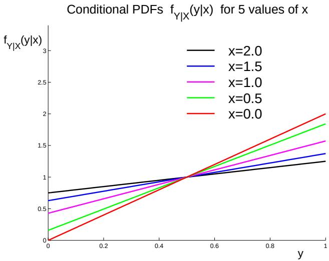
*for $0 < x < 2$, and 0 otherwise.*

It is convenient to define $f_{Y|X}(y|x) = 0$ when $f_X(x) = 0$, i.e., for $x < 0$ or $x > 2$.

Likewise, we set $f_{X|Y}(x|y) = 0$ for $y < 0$ or $y > 1$.

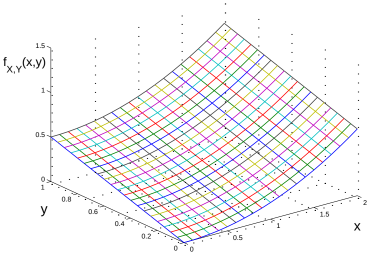
# Conditional PDFs


Conditional PDFs $f_{X|Y}(x|y)$ for 6 values of y


Conditional PDFs $f_{Y|X}(y|x)$ for 5 values of x

*The shapes of $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ can be seen in the 2D plot of $f_{X,Y}(x,y)$.*

*Dividing by $f_Y(y)$ or $f_X(x)$ scales the conditional PDFs, so that $\int f_{X|Y}(x|y)dx = 1$ and $\int f_{Y|X}(y|x)dy = 1$.*



# Conditional PDFs

Suppose we wish to find the probability that $Y > 0.5$ given that we know $X = 1.5$.

We simply calculate

$$\int_{1/2}^{1} f_{Y|X}\left(y \,|\, X = \frac{3}{2}\right) dy \;=\; \int_{1/2}^{1} \frac{3(3/2)^2 + 8y}{3(3/2)^2 + 4}\, dy$$

$$= \;\frac{1}{43} \int_{1/2}^{1} 27 + 32y\, dy \;\;=\;\; \frac{51}{86}.$$

# The multiplication rule

It follows from the definitions of $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$ that

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)\, f_{Y|X}(y|x) \\ &= f_Y(y)\, f_{X|Y}(x|y) \end{aligned}$$

whenever $f_X(x) > 0$ and $f_Y(y) > 0$.

Recall that $X$ and $Y$ are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)\, f_Y(y) \quad \text{for all } x \in \mathsf{R},\ y \in \mathsf{R}.$$

So the property of independence is equivalent to

$$f_{Y|X}(y|x) = f_Y(y) \quad \text{for all } y \in \mathsf{R}, \text{ for all } x \text{ such that } f_X(x) > 0$$

and to

$$f_{X|Y}(x|y) = f_X(x) \quad \text{for all } x \in \mathsf{R}, \text{ for all } y \text{ such that } f_Y(y) > 0.$$

# Multiplying conditional probabilities

## Example

As part of a arbitration process in the ikh khural the following division of wealth is agreed:

The Bogd Khan must take a random number $X \sim \mathrm{Unif}(0,1)$ and give a fraction $X$ of his gold to Queen Mandukhai.

Queen Mandukhai must take $Y \sim \mathrm{Unif}(0,1)$ and give a fraction $Y$ of the gold she received to Chinggis Khan.

Let $Z$ be the fraction of the Bogd Khan's gold that eventually passes to Chinggis Khan (so $Z = X\,Y$). Find the PDF of $Z$.

We have

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases} \qquad f_{Z|X}(z|x) = \begin{cases} 1/x & 0 < z < x \\ 0 & \text{otherwise.} \end{cases}$$

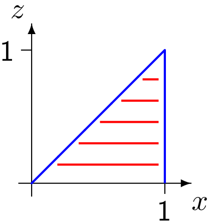
*See calculations on board*

# Multiplying conditional probabilities

From $f_X(x)$ and $f_{Z|X}(z|x)$, non-zero values of the joint PDF are

$$\begin{aligned} f_{X,Z}(x,z) &= f_X(x)\, f_{Z|X}(z|x) \\ &= 1 \cdot \frac{1}{x} \qquad \text{for } 0 < x < 1 \text{ and } 0 < z < x. \end{aligned}$$

*Note the triangular shape of the region where $f_{X,Z}(x,z) > 0$.*



Hence, for $0 < z < 1$,

$$f_Z(z) = \int_z^1 f_{X,Z}(x,z)\, dx = \int_z^1 \frac{1}{x}\, dx = [\log(x)]_z^1 = -\log(z).$$

# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**IV.a** Joint PDFs, marginal distributions and independence

**IV.b** Conditional PDFs

**IV.c** More on expectation

**IV.d** Covariance

Today, we return to the Law of the Unconscious Statistician, we consider expectation of a function of two continuous RVs, and we tackle the topic of covariance.

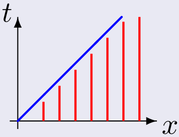
# IV.c More on Expectation

## Lemma

*Suppose $X$ is a continuous RV with $P(X \geq 0) = 1$ and* $\mathsf{E}(X) < \infty$. *Then*

$$\mathsf{E}(X) = \int_0^\infty \mathsf{P}(X \geq t)\, dt.$$

## Proof

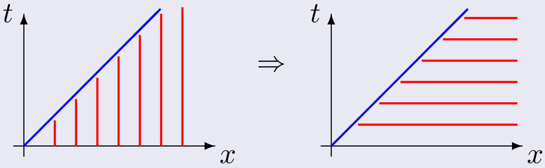


$$\mathsf{E}(X) = \int_0^\infty x\, f_X(x)\, dx = \int_0^\infty \left\{ \int_0^x dt \right\} f_X(x)\, dx.$$

# Proof of the lemma

## Proof, continued

*Rearrange the integral and change the order of integration:*



$$\begin{aligned}\mathsf{E}(X) &= \int_0^\infty x\, f_X(x)\, dx \;=\; \int_0^\infty \int_0^x f_X(x)\, dt\, dx \\ &= \int_0^\infty \int_t^\infty f_X(x)\, dx\, dt \;=\; \int_0^\infty \mathsf{P}(X \geq t)\, dt. \qquad \square\end{aligned}$$

# Law of the Unconscious Statistician

## Theorem

*Suppose $X$ is a continuous RV with PDF $f_X(x)$, $g\colon \mathsf{R} \to \mathsf{R}$ is a positive function, and $g(X)$ is a continuous RV.*

*Then*

$$\mathsf{E}[\,g(X)\,] \;=\; \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx. \tag{6}$$

## Proof

*Since $g(X)$ is a positive, continuous RV, by the lemma,*

$$\mathsf{E}[\,g(X)\,] \;=\int_0^\infty \mathsf{P}[\,g(X) \geq t\,]\, dt \;=\int_0^\infty \left\{ \int_{\{x\colon g(x) \geq t\}} f_X(x)\, dx \right\} dt.$$

# Law of the Unconscious Statistician

## Proof, continued

$$
\begin{aligned}
\mathsf{E}[g(X)] &= \int_{t=0}^{\infty} \int_{\{x:\, g(x)\geq t\}} f_X(x)\; dx\; dt \\
&= \int_{x=-\infty}^{\infty} \int_{\{t:\, 0\leq t\leq g(x)\}} f_X(x)\; dt\; dx \\
&= \int_{x=-\infty}^{\infty} \int_{0}^{g(x)} f_X(x)\; dt\; dx \\
&= \int_{x=-\infty}^{\infty} \left\{ \int_{0}^{g(x)} dt \right\} f_X(x)\, dx = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx.
\end{aligned}
$$

□

# Law of the Unconscious Statistician

SELF STUDY SLIDE

The theorem can be extended to the case of a function $g$ which takes positive and negative values, as long as $\mathsf{E}(g(X))$ is defined.

*You should check the details below in your own time.*

First, generalise the lemma to a general, continuous RV $X$.

## Lemma

*If* $\mathsf{E}(|X|) < \infty$, *then*

$$\mathsf{E}(X) \;=\; -\int_{-\infty}^{0} \mathsf{P}(X \le t)\,dt + \int_{0}^{\infty} \mathsf{P}(X \ge t)\,dt.$$

## Proof

*Write*

$$\mathsf{E}(X) \;=\; \int_{-\infty}^{\infty} x\, f_X(x)\,dx \;=\; \int_{-\infty}^{0} x\, f_X(x)\,dx + \int_{0}^{\infty} x\, f_X(x)\,dx.$$

# Law of the Unconscious Statistician

SELF STUDY SLIDE

## Proof, continued

*Apply the argument used to prove the original lemma to show that*

$$\int_{-\infty}^{0} x\, f_X(x)\, dx \;=\; -\int_{-\infty}^{0} \mathsf{P}(X \le t)dt.$$

*We already have*

$$\int_{0}^{\infty} x\, f_X(x)\, dx \;=\; \int_{0}^{\infty} \mathsf{P}(X \ge t)dt,$$

*so the result of the new lemma follows.*

□

# Law of the Unconscious Statistician

SELF STUDY SLIDE

Applying the new lemma to the continuous RV $g(X)$ gives

$$\mathsf{E}(g(X)) \;=\; -\int_{-\infty}^{0} \mathsf{P}(g(X) \le t)\, dt + \int_{0}^{\infty} \mathsf{P}(g(X) \ge t)\, dt.$$

Similar arguments to those used to prove (6) imply the above expression is equal to

$$-\int_{-\infty}^{\infty} I[g(x) \le 0]\; (-g(x))\, f_X(x)\, dx$$

$$+\int_{-\infty}^{\infty} I[g(x) \ge 0]\; g(x)\, f_X(x)\, dx$$

and the result follows.

# Expectation with two random variables

We now state, without proof, a 2-dimensional version of the Law of the Unconscious Statistician.

A proof can be constructed following the same steps as in the 1-dimensional case.

### Theorem

*Suppose $X$ and $Y$ are continuous RVs with joint PDF $f_{X,Y}(x,y)$ and $h$: $\mathsf{R}^2 \to \mathsf{R}$ is a continuous function.*

*Then*

$$\mathsf{E}[\,h(X,Y)\,] \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\, f_{X,Y}(x,y)\, dy\, dx$$

*as long as $\mathsf{E}(|h(X,Y)|)$ is defined, i.e., finite.*

# Expectation with two random variables

## Proposition

*If $X$ and $Y$ are jointly continuous RVs, $a \in \mathsf{R}$ and $b \in \mathsf{R}$, then*

$$\mathsf{E}[\,a\,X + b\,Y\,] \;=\; a\,\mathsf{E}(X) + b\,\mathsf{E}(Y).$$

**Proof** *See calculations on board*

## Proposition

*Suppose $X$ and $Y$ are* **independent***, jointly continuous RVs, and $g$ and $h$ are functions from $\mathsf{R} \to \mathsf{R}$. If the relevant integrals exist,*

$$(i) \qquad \mathsf{E}[\,XY\,] \;=\; \mathsf{E}(X)\,\mathsf{E}(Y),$$

$$(ii) \qquad \mathsf{E}[\,g(X)\,h(Y)\,] \;=\; \mathsf{E}[g(X)]\,\mathsf{E}[h(Y)].$$

**Proof** *See calculations on board*

# IV.d Covariance

Suppose $X$ and $Y$ are random variables and $\mathsf{Var}(X)$ and $\mathsf{Var}(Y)$ both exist.

## Definition

The **covariance** of $X$ and $Y$ is

$$\mathsf{Cov}(X,Y) \;=\; \mathsf{E}\{\,[X-\mathsf{E}(X)]\,[Y-\mathsf{E}(Y)]\,\}.$$

Note that

$$\begin{aligned}\mathsf{Cov}(X,Y) &= \mathsf{E}\{\,[X-\mathsf{E}(X)]\,[Y-\mathsf{E}(Y)]\,\}\\ &= \mathsf{E}(X\,Y)-\mathsf{E}(X)\,\mathsf{E}(Y)-\mathsf{E}(X)\,\mathsf{E}(Y)+\mathsf{E}(X)\,\mathsf{E}(Y)\\ &= \mathsf{E}(X\,Y)-\mathsf{E}(X)\,\mathsf{E}(Y).\end{aligned}$$

# Correlation

## Definition

The **correlation** between $X$ and $Y$ is

$$\mathsf{Corr}(X,Y) \;=\; \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}}\,.$$

# Properties of covariance and correlation SELF STUDY SLIDE

You should **know** these properties (proofs are on the Moodle page).

**Proposition**

*(i)* $\mathsf{Cov}(X, X) \;=\; \mathsf{Var}(X)$

*(ii)* $\mathsf{Cov}(X, Y) \;=\; \mathsf{Cov}(Y, X)$

*(iii)* $\mathsf{Cov}(a\,X + b\,Y,\; Z) \;=\; a\,Cov(X, Z) + b\,\mathsf{Cov}(Y, Z)$

*(iv)* $\mathsf{Var}(a\,X + b\,Y) \;=\; a^2\,Var(X) + b^2\,\mathsf{Var}(Y) + 2\,a\,b\,\mathsf{Cov}(X, Y)$

*(v)* $-1 \;\le\; \mathsf{Corr}(X, Y) \;\le\; 1$

*(vi)* $\mathsf{Corr}(X, Y) = 1 \;\Leftrightarrow\; Y = a\,X + b$ *for some* $a > 0$ *and* $b \in \mathsf{R}$

*(vii)* $\mathsf{Corr}(X, Y) = -1 \Leftrightarrow Y = a\,X + b$ *for some* $a < 0$ *and* $b \in \mathsf{R}$

*(viii)* *If* $X$ *and* $Y$ *are independent, then* $\mathsf{Cov}(X, Y) = 0$ — *but* $\mathsf{Cov}(X, Y) = 0$ *does* ***not*** *imply* $X$ *and* $Y$ *are independent.*

# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

**V.a** Transforming random variables

**V.b** Applications to simulation

Today, we shall learn about the distribution of a transformation of a random variable.

We shall then start to look at how transformations can be used in simulating from tricky looking distributions.

# IV.a Transformations of random variables

Suppose $X$ is a RV and $g\colon$ R $\to$ R is a function. We know the distribution of $X$ and we wish to find the distribution of $Y = g(X)$.

If P$(a < X < b) = 1$, we may consider $g\colon\ (a, b) \to$ R.

## Example

If $X \sim N(0, 1)$ and $Y = X^2$, what is the distribution of $Y$?

Note that P$(0 < Y < \infty) = 1$.

Take $y > 0$, then

$$\begin{aligned} \mathrm{P}(Y \le y) &= P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = \Phi(\sqrt{y}) - \{1 - \Phi(\sqrt{y})\} \\ &= 2\,\Phi(\sqrt{y}) - 1. \end{aligned}$$

# Example: Transformation of a random variable

## Example, continued

*So, if* $X \sim N(0,1)$ *and* $Y = X^2$*, the CDF of* $Y$ *is*

$$F_Y(y) \; = \; \begin{cases} 2\,\Phi(\sqrt{y}) - 1 & y > 0, \\ 0 & y \leq 0. \end{cases}$$

*Now differentiate* $F_Y(y)$ *to find the PDF,* $f_Y(y)$*. Recall that*

$$\Phi(x) \; = \; \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\, du \; = \; \int_{-\infty}^{x} \phi(u)\, du.$$

*The fundamental theorem of calculus implies that*

$$\frac{d}{dx}\,\Phi(x) \; = \; \phi(x) \; = \; \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

# Example: Transformation of a random variable

## Example, continued

*Hence, the PDF of* $Y$ *at* $y > 0$ *is*

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\{2\,\Phi(\sqrt{y}) - 1\} = 2\,\Phi'(\sqrt{y})\,\frac{1}{2\sqrt{y}} \\ &= \phi(\sqrt{y})\,\frac{1}{\sqrt{y}} = \frac{1}{\sqrt{2\pi}}\,e^{-y/2}\,y^{-1/2}, \end{aligned}$$

*and* $f_Y(y) = 0$ *for* $y < 0$.

*By inspection* $Y \sim \mathrm{Gamma}\,(1/2, 1/2)$, *as* $\Gamma(1/2) = \sqrt{\pi}$ *implies*

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(1/2)}\,\left(\frac{1}{2}\right)^{1/2}\,y^{-1/2}\,e^{-\frac{1}{2}y} & y > 0, \\ 0 & y \le 0. \end{cases}$$

*The random variable* $Y$ *is also said to have a* $\chi^2_1$ *distribution.*

# The CDF method for finding the distribution of $g(X)$

The preceding example illustrates a general approach to finding the CDF and PDF of $g(X)$ from the distribution of $X$.

(i) Express the event $\{Y \leq y\}$ in terms of the RV $X$.

(ii) Find $F_Y(y)$, the probability of $\{Y \leq y\}$, from the CDF of $X$.

(iii) Differentiate $F_Y(y)$ with respect to $y$ to find the PDF $f_Y(y)$.

# The transformation formula

## Theorem

*Let $X$ be a continuous RV with PDF $f_X(x)$ and* $\mathsf{P}(a < X < b) = 1$.

*Suppose $g : (a, b) \to$* R *is continuous, strictly increasing and differentiable.*

*Then $Y = g(X)$ is a continuous RV with PDF*

$$f_Y(y) \;=\; f_X(g^{-1}(y))\,\frac{d}{dy}\,(g^{-1}(y)).$$

*If we set $x = g^{-1}(y)$, so $y = g(x)$, then we can write*

$$f_Y(y) \;=\; f_X(x)\,\frac{dx}{dy} \;=\; f_X(x)\,\left(\frac{dy}{dx}\right)^{-1}.$$

## Proof

*See calculations on board*

# The transformation formula

## Example

Let $X \sim \mathrm{Exp}(1)$ and $Y = \sqrt{X}$.

What is the distribution of $Y$?

Define the function $g(x) = \sqrt{x}$ for $x \in (0, \infty)$.

Then $g(x)\colon (0,\infty) \to (0,\infty)$ is continuous, strictly increasing and differentiable, and we can apply the theorem.

With $x = g^{-1}(y) = y^2$,

$$\frac{dx}{dy} \;=\; 2\,y$$

and

$$f_Y(y) \;=\; f_X(x)\,\frac{dx}{dy} \;=\; e^{-x}\,2\,y \;=\; 2\,y\,e^{-y^2} \quad \text{for } y > 0.$$

We recognise this distribution as $Y \sim \mathrm{Weib}(1,2)$.

# Independence of transformed RVs

We state the following theorem without proof.

**Theorem**

*Suppose $X$ and $Y$ are independent RVs and we have functions $g$ : R $\to$ R and $h$ : R $\to$ R.*

*Then $g(X)$ and $h(Y)$ are also independent.*

# IV.b Applications of transformed RVs to simulation

There are good methods to simulate independent $\mathrm{Unif}(0,1)$ RVs.

These "pseudo random number" generators lie at the heart of methods for simulating from more complex distributions.

We can take

$$U \;\sim\; \mathrm{Unif}(0,1)$$

then produce a new RV

$$X \;=\; h(U).$$

**Example**

Using the function $h(u) = -\log(1-u)$ gives values $X \sim \mathrm{Exp}(1)$.

*See calculations on board*

The next challenge is how to choose the function $h$ to obtain $X$ following a specified distribution.

# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

- **V.a** Transforming random variables
- **V.b** Applications to simulation

We continue to study how transformations can be used in simulating from a variety of distributions.

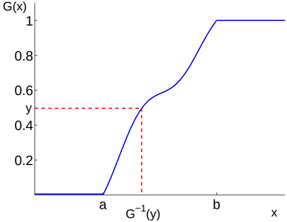
# The inverse CDF method for simulation

Suppose we have a way of generating $\mathrm{Unif}(0,1)$ RVs but we wish to generate a RV from a distribution with CDF $G(x)$.

Consider the case where $G(x)$ is a continuous, strictly increasing function $G\colon [a,b] \to [0,1]$ with $G(a) = 0$ and $G(b) = 1$.

For $b = \infty$, we write the function $G$ as $G\colon [a,\infty) \to [0,1)$.
If also $a = -\infty$, we write $G\colon (-\infty,\infty) \to (0,1)$.

The function $G^{-1}\colon (0,1) \to (a,b)$ is well-defined.



Note that

$$G(G^{-1}(y)) = y.$$

# The inverse CDF method for simulation

**A key fact:**

Suppose the continuous random variable $X$ has CDF $G(x)$, where $G(x)$ is a continuous, strictly increasing function on $[a, b]$.

Define the random variable $Y = G(X)$ — so $X = G^{-1}(Y)$.

Then, $Y \sim \mathrm{Unif}(0, 1)$.

**Proof**

By definition, $Y$ only takes values in the range $[0, 1]$.

For $0 \leq y \leq 1$,

$$\begin{aligned} \mathsf{P}(Y \leq y) &= \mathsf{P}\{G(X) \leq y\} \\ &= \mathsf{P}\{X \leq G^{-1}(y)\} \\ &= G(G^{-1}(y)) = y. \qquad \square \end{aligned}$$

# The inverse CDF method for simulation

## Theorem

*Suppose $G(x)$ is a continuous, strictly increasing function from $[a,b] \to [0,1]$ with $G(a) = 0$ and $G(b) = 1$.*

*Cases $a = -\infty$ and $b = \infty$ are also allowed, as noted above.*

*Let $U \sim \mathrm{Unif}(0,1)$ and set $X = G^{-1}(U)$.*

*Then $X$ is a RV with CDF $G(x)$.*

## Proof

*Since $\mathsf{P}(0 < U < 1) = 1$, $X$ is well-defined with probability 1.*

*For $x \in (a,b)$,*

$$\begin{aligned} F_X(x) &= \mathsf{P}(X \le x) = \mathsf{P}(G^{-1}(U) \le x) \\ &= \mathsf{P}\{G(G^{-1}(U)) \le G(x)\} = \mathsf{P}(U \le G(x)) = G(x). \end{aligned}$$

# Example of the inverse CDF method

## Example

Generate a RV with CDF

$$G(x) \;=\; \begin{cases} 1 - \exp\{-\lambda\, x^2\} & x > 0 \\ 0 & x \leq 0, \end{cases}$$

where $\lambda > 0$.

*See calculations on board*

We find

$$G^{-1}(u) \;=\; \sqrt{\frac{-\log(1-u)}{\lambda}},$$

so take $U \sim \mathrm{Unif}(0,1)$ and

$$X \;=\; G^{-1}(U) \;=\; \sqrt{\frac{-\log(1-U)}{\lambda}}.$$

# Another example of the inverse CDF method

## Example

Generate a RV with PDF

$$f_X(x) \;=\; \begin{cases} \frac{1}{2\sqrt{x}} & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Integrate $f_X$ to obtain

$$F_X(x) \;=\; \begin{cases} 0 & x \leq 0 \\ \sqrt{x} & 0 < x < 1 \\ 1 & x \geq 1. \end{cases}$$

We have $F_X(x) \colon (0,1) \to (0,1)$ and we see this is continuous and strictly increasing. We need to find the inverse of $F_X$.

# Another example of the inverse CDF method

## Example, continued

*We have*

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \sqrt{x} & 0 < x < 1 \\ 1 & x \geq 1. \end{cases}$$

If

$$u = F_X(x) = \sqrt{x},$$

then

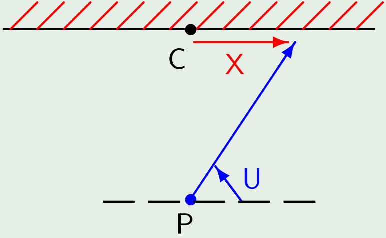
$$x = u^2 = F_X^{-1}(u).$$

So, by the theorem, we take $U \sim \mathrm{Unif}(0,1)$ and set $X = U^2$.

# Creating a Cauchy RV: The novice tennis player

## Example

A tennis player stands at a point P, $10m$ from a wall and hits balls in random directions. The closest point on the wall is $C$.



Denote by $U$ the angle between the ball's trajectory and a line parallel to the wall.

Consider cases where $0 < U < \pi$ and let $X$ be the distance (in $m$) to the right of $C$ that a ball hits the wall.

If $U \sim \mathrm{Unif}(0, \pi)$, what is the distribution of $X$?

# The novice tennis player

Express $X$ in terms of $U$ and, hence, find an expression for $\mathsf{P}(X \le x)$.

*See calculations on board*

We obtain

$$F_X(x) \; = \; \frac{1}{2} + \frac{1}{\pi} \, \tan^{-1}\left(\frac{x}{10}\right) \quad \text{for } x \in \mathsf{R}.$$

Differentiating gives

$$f_X(x) \; = \; \frac{1}{\pi} \frac{1}{10\,(1 + x^2/10^2)} \quad \text{for } x \in \mathsf{R}.$$

The RV $X$ follows a *Cauchy* distribution.

# The Cauchy distribution

## Definition

The RV $Y$ follows a central Cauchy distribution with parameter 1 if

$$f_Y(y) \ = \ \frac{1}{\pi} \frac{1}{1+y^2} \quad \text{for } y \in \mathrm{R}.$$

The RV $Y$ follows a central Cauchy distribution with parameter $\theta$ if

$$f_Y(y) \ = \ \frac{1}{\pi} \frac{1}{\theta\,(1+y^2/\theta^2)} \quad \text{for } y \in \mathrm{R}.$$

We have seen a method to generate a Cauchy random variable:

Take $U \sim \mathrm{Unif}(0,\pi)$ — e.g., let $U$ be $\pi$ times a $\mathrm{Unif}(0,1)$ RV.

Then

$$X \ = \ \theta \tan\left(\frac{\pi}{2} - U\right) \ \sim \ \mathrm{Cauchy}\,(\theta).$$

# The Cauchy distribution

What is the mean of a Cauchy RV with parameter $\theta$ ?

*See calculations on board*

Calculation shows that

$$\int_{-\infty}^{\infty} \frac{|y|}{\pi\,(1+y^2)}\,dy \;=\; \infty.$$

We can (with care) write $\mathsf{E}(|Y|) = \infty$.

But $\mathsf{E}(Y)$ is **undefined**.

Have we seen a previous example of a RV with infinite expectation?

# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

**VI** Sums of random variables

- **VI.a** Mean and variance of sums of RVs
- **VI.b** The Central Limit Theorem
- **VI.c** Distribution of a sum of RVs
- **VI.d** Distribution of a sum of normal RVs

Our next topic is the distribution of a sum of random variables.

We start with the mean and variance of a sum of RVs.

# VI.a Mean and variance of a sum of RVs

Recall that

$$\mathsf{E}(X_1 + X_2) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$$

and

$$\mathsf{Var}(X_1 + X_2) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2) + 2\,\mathsf{Cov}(X_1, X_2).$$

### Proposition

$$\mathsf{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{E}(X_i).$$

### Proposition

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

*See calculations on board*

# A sum of independent random variables

## Proposition

*Suppose* $X_1, \ldots, X_n$ *are independent RVs. Then*

$$\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{Var}(X_i).$$

## Proof.

Since $X_1, \ldots, X_n$ are independent, we have

$$\mathsf{Cov}(X_i, X_j) = 0 \quad \text{for each pair } i \text{ and } j.$$

Substituting this in the previous result gives the desired answer.

□

# Random samples

## Definition

Let $X_1, \ldots, X_n$ be independent RVs from the same distribution. Then, $X_1, \ldots, X_n$ are said to be **independent and identically distributed,** or **i.i.d.** for short.

The values taken by these RVs, $x_1, \ldots, x_n$, are said to form a **random sample** from the distribution in question.

## Example

An experiment is performed $n$ times, independently. We define

$$X_i = \begin{cases} 1 & \text{if a certain event occurs in experiment } i \\ 0 & \text{if the event does not occur in experiment } i. \end{cases}$$

The set of values $\{x_1, \ldots, x_n\}$ is a random sample from the $\mathrm{Bernoulli}(p)$ distribution, where $p$ is the probability the event occurs in a particular experiment.

# VI.b The Central Limit Theorem

In the preceding example, the sum of the Bernoulli RVs is

$$S_n \ = \ X_1 + \ldots + X_n \ \sim \ \mathrm{Binom}(n, p).$$

It is easy to check that, for a single Bernoulli RV,

$$\mathsf{E}(X_i) \ = \ p \quad \text{and} \quad \mathsf{Var}(X_i) \ = \ p\,(1-p).$$

Hence, by our earlier results

$$\mathsf{E}(S_n) \ = \ n\,p \quad \text{and} \quad \mathsf{Var}(S_n) \ = \ n\,p\,(1-p).$$

The deMoivre-Laplace theorem tells us that, for large $n$, approximately

$$\frac{S_n - n\,p}{\sqrt{n\,p\,(1-p)}} \ \sim \ N(0,1).$$

The **Central Limit Theorem** is a more general form of this result.

# The Central Limit Theorem

We state without proof:

### Theorem

*Suppose* $X_1, X_2, \ldots$ *are i.i.d. RVs with finite expectation and variance*

$$\mathsf{E}(X_i) \;=\; \mu \;\textit{ and }\; \mathsf{Var}(X_i) \;=\; \sigma^2 > 0 \;\textit{ for each } i = 1, 2, \ldots\,.$$

*Let* $S_n \;=\; X_1 + \ldots + X_n$*. Then, for any* $a \in \mathsf{R}$*,*

$$\mathsf{P}\left(\frac{S_n - n\,\mu}{\sqrt{n\,\sigma^2}} < a\right) \;\to\; \Phi(a) \quad \textit{as } n \to \infty,$$

*where* $\Phi(x)$ *is the CDF at* $x$ *of a standard normal distribution.*

Previous results imply $(S_n - n\,\mu)/\sqrt{n\,\sigma^2}$ has mean zero and variance 1. The Central Limit Theorem gives us the full information about the shape of this distribution — for large $n$.

# An illustration of the Central Limit Theorem

## Example

Suppose
$$X_i \sim \mathrm{Exp}(0.2), \quad i = 1,\ 2,\ \dots\ .$$

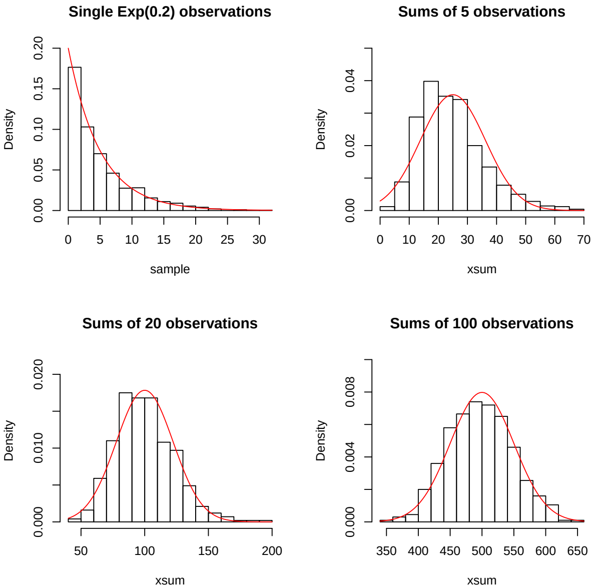
We can look at the distribution of one observation $X_i$ by taking 1000 realisations and plotting the histogram.

Suppose, instead, we draw a sample of $5$ RVs $X_1, \dots, X_5$ and take their sum. Now repeat this 1000 times and plot the histogram.

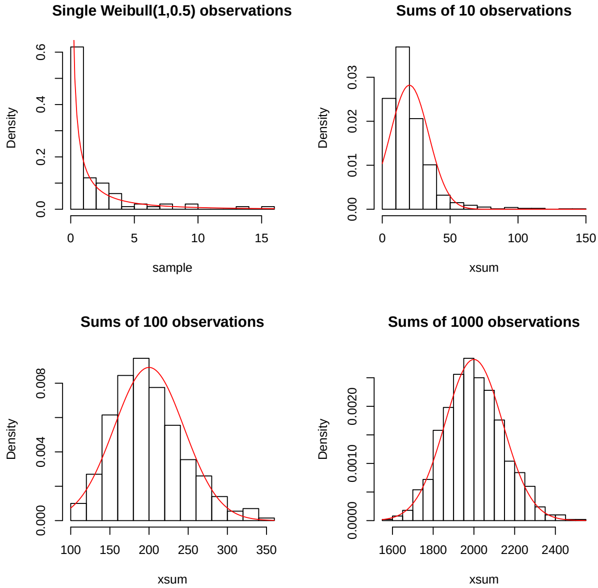
We see a distribution whose *shape* is somewhat like that of a normal distribution. However, the left hand tail is rather short and the right hand tail is rather long.

Repeating with sums of 20 or 100 $\mathrm{Exp}(0.2)$ RVs, the histograms look to follow the shape of a normal density closely.

# Histograms of sums of $\text{Exp}(0.2)$ RVs



# Histograms of sums of $\mathrm{Weib}(1, 0.5)$ RVs



# Where we are

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

**VI** Sums of random variables

- **VI.a** Mean and variance of sums of RVs
- **VI.b** The Central Limit Theorem
- **VI.c** Distribution of a sum of RVs
- **VI.d** Distribution of a sum of normal RVs

We now consider the distribution of the sum of two RVs, finding the precise PDF of the sum from the RVs' joint distribution.

# VI.c Distribution of the sum of two RVs

Suppose $X$ and $Y$ are continuous RVs with joint PDF $f_{X,Y}(x, y)$.

Let $W = X + Y$.

What is the PDF, $f_W(w)$, of $W$?

First consider the CDF,

$$\mathsf{P}(W \leq w) \;=\; \mathsf{P}(X + Y \leq w).$$

If we can write this in the form

$$\mathsf{P}(W \leq w) \;=\; \int_{-\infty}^{w} h(v)\, dv,$$

then we can deduce $f_W(w) = h(w)$.

# Distribution of the sum of two RVs

Let $A_w = \{\, (x,y) : \ x + y \leq w \,\}$.

Then,

$$
\begin{aligned}
\mathsf{P}(W \leq w) &= \int_{A_w} f_{X,Y}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{w-x} f_{X,Y}(x,y)\, dy \right\} dx \\
&= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{w} f_{X,Y}(x, v-x)\, dv \right\} dx,
\end{aligned}
$$

using the substitution $v = y + x$ in the inner integral.

*See calculations on board*

# Distribution of the sum of two RVs

Re-arrange this equation as

$$\begin{aligned}\mathsf{P}(W \le w) &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{w} f_{X,Y}(x, v-x)\, dv \right\} dx \\ &= \int_{-\infty}^{w} \left\{ \int_{-\infty}^{\infty} f_{X,Y}(x, v-x)\, dx \right\} dv.\end{aligned}$$

## Convolution formula, general case

*Comparing the above formula with*

$$\mathsf{P}(W \le w) = \int_{-\infty}^{w} f_W(v)\, dv,$$

*we can deduce that $W = X + Y$ has PDF*

$$f_W(w) = \int_{-\infty}^{\infty} f_{X,Y}(x, w-x)\, dx.$$

# The convolution formula for two independent RVs

## Convolution formula, independent RVs

*If $X$ and $Y$ are independent, continuous RVs, then their sum $W = X + Y$ has PDF*

$$\begin{aligned} f_W(w) &= \int_{-\infty}^{\infty} f_{X,Y}(x, w-x)\, dx \\ &= \int_{-\infty}^{\infty} f_X(x)\, f_Y(w-x)\, dx. \end{aligned}$$

*This is known as the "convolution formula".*

# VI.d The sum of two normal random variables

## Proposition

*If $X$ and $Y$ are independent and*

$$X \sim N(\mu_X, \sigma_X^2) \quad \textit{and} \quad Y \sim N(\mu_Y, \sigma_Y^2),$$

*then*

$$W = X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

**Proof:** We first consider the case where $\mu_X = \mu_Y = 0$.

Using the convolution formula,

$$\begin{aligned} f_W(w) &= \int_{-\infty}^{\infty} f_X(x)\, f_Y(w-x)\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(\frac{-x^2}{2\sigma_X^2}\right) \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(\frac{-(w-x)^2}{2\sigma_Y^2}\right) dx. \end{aligned}$$

# The sum of two normal random variables

Hence,

$$f_W(w) \;=\; \frac{1}{2\pi}\,\frac{1}{\sqrt{\sigma_X^2\,\sigma_Y^2}}\,\int_{-\infty}^{\infty} \exp\left(\frac{-1}{2}\left[\frac{x^2}{\sigma_X^2} + \frac{w^2 - 2\,w\,x + x^2}{\sigma_Y^2}\right]\right) dx.$$

We can re-arrange the terms in the exponential as a quadratic in $x$ (which involves $w$) plus a term in $w^2$.

This gives an expression of the form

$$\begin{aligned} f_W(w) &= k \int_{-\infty}^{\infty} \exp\{-a(x - b\,w)^2 - c\,w^2\}\,dx, \\ &= k\,e^{-c\,w^2} \int_{-\infty}^{\infty} \exp\{-a(x - b\,w)^2\}\,dx, \end{aligned}$$

where $k$, $a$, $b$ and $c$ are constants (not involving $x$ or $w$).

# The sum of two normal random variables

We have

$$f_W(w) \; = \; k \, e^{-c\,w^2} \int_{-\infty}^{\infty} \exp\{-a(x - b\,w)^2\}\, dx.$$

Note that in

$$\int_{-\infty}^{\infty} \exp\{-a(x - b\,w)^2\}\, dx$$

the integrand is a multiple of the PDF of a normal RV $X$ with mean $bw$, so the answer is a constant which does not involve $w$.

It follows that $f_W(w)$ has the form

$$f_W(w) \; = \; h \, e^{-c\,w^2}$$

for a constant $h$ — so $W$ is normally distributed (with mean 0).

# The sum of two normal random variables

Given that $W$ is normally distributed, all that remains is to find $\mathsf{E}(W)$ and $\mathsf{Var}(W)$.

Using standard results for independent RVs $X$ and $Y$,

$$\mathsf{E}(W) \;=\; \mathsf{E}(X) + \mathsf{E}(Y) \;=\; 0$$

and

$$\mathsf{Var}(W) \;=\; \mathsf{Var}(X) + \mathsf{Var}(Y) \;=\; \sigma_X^2 + \sigma_Y^2.$$

Thus, we have

$$W \;\sim\; N(0, \sigma_X^2 + \sigma_Y^2) \;=\; N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2),$$

which proves the result for the case $\mu_X = \mu_Y = 0$.

# The sum of two normal random variables SELF STUDY SLIDE

*An algebraic derivation* giving $\mathsf{E}(W)$ and $\mathsf{Var}(W)$ is possible — *check this in your own time.*

We have

$$f_W(w) = \frac{1}{2\pi} \frac{1}{\sqrt{\sigma_X^2 \sigma_Y^2}} \int_{-\infty}^{\infty} \exp\left( \frac{-1}{2} \left[ \frac{x^2}{\sigma_X^2} + \frac{(w-x)^2}{\sigma_Y^2} \right] \right) dx. \quad (7)$$

First, note that

$$\begin{aligned} \frac{x^2}{\sigma_X^2} + \frac{(w-x)^2}{\sigma_Y^2} &= \sigma_X^{-2}\, x^2 + \sigma_Y^{-2}\, (x^2 - 2\, w\, x + w^2) \\ &= (\sigma_X^{-2} + \sigma_Y^{-2})\, x^2 - 2\, \sigma_Y^{-2}\, w\, x + \sigma_Y^{-2}\, w^2 \\ &= (\sigma_X^{-2} + \sigma_Y^{-2}) \left\{ x^2 - \frac{2\, \sigma_Y^{-2}\, w}{\sigma_X^{-2} + \sigma_Y^{-2}}\, x \right\} + \sigma_Y^{-2}\, w^2. \end{aligned}$$

# The sum of two normal random variables SELF STUDY SLIDE

**The algebra, continued:**

We have

$$\frac{x^2}{\sigma_X^2} + \frac{(w-x)^2}{\sigma_Y^2} = (\sigma_X^{-2} + \sigma_Y^{-2}) \left\{ x^2 - \frac{2\,\sigma_Y^{-2}\,w}{\sigma_X^{-2} + \sigma_Y^{-2}}\, x \right\} + \sigma_Y^{-2}\, w^2$$

$$= (\sigma_X^{-2} + \sigma_Y^{-2}) \left\{ x - \frac{\sigma_Y^{-2}\,w}{\sigma_X^{-2} + \sigma_Y^{-2}} \right\}^2 - \frac{\sigma_Y^{-4}\,w^2}{\sigma_X^{-2} + \sigma_Y^{-2}} + \sigma_Y^{-2}\, w^2$$

$$= \frac{(\sigma_X^2 + \sigma_Y^2)}{\sigma_X^2\,\sigma_Y^2} \left\{ x - \frac{\sigma_Y^{-2}\,w}{\sigma_X^{-2} + \sigma_Y^{-2}} \right\}^2 + \frac{w^2}{(\sigma_X^2 + \sigma_Y^2)}.$$

# The sum of two normal random variables SELF STUDY SLIDE

**Algebra continued:**

Substituting into (7), we get

$$f_W(w) \;=\; \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}}\exp\left(\frac{-1}{2}\frac{w^2}{(\sigma_X^2+\sigma_Y^2)}\right)\int_{-\infty}^{\infty} g_X(x)\,dx,$$

where

$$g_X(x) \;=\; \sqrt{\frac{(\sigma_X^2+\sigma_Y^2)}{2\pi(\sigma_X^2\,\sigma_Y^2)}}\exp\left(\frac{-(\sigma_X^2+\sigma_Y^2)}{2\,\sigma_X^2\,\sigma_Y^2}\left\{x-\frac{\sigma_Y^{-2}\,w}{\sigma_X^{-2}+\sigma_Y^{-2}}\right\}^2\right)$$

is the PDF of a random variable

$$X \;\sim\; N\left(\frac{\sigma_Y^{-2}}{\sigma_X^{-2}+\sigma_Y^{-2}}\,w,\;\frac{\sigma_X^2\,\sigma_Y^2}{(\sigma_X^2+\sigma_Y^2)}\right).$$

# The sum of two normal random variables SELF STUDY SLIDE

**Algebra continued:**

Since

$$\int g_X(x)\,dx = 1,$$

we have

$$f_W(w) \;=\; \frac{1}{\sqrt{2\pi(\sigma_X^2+\sigma_Y^2)}}\exp\left(\frac{-1}{2}\frac{w^2}{(\sigma_X^2+\sigma_Y^2)}\right)$$

and we see that

$$W \;\sim\; N(0,\,\sigma_X^2+\sigma_Y^2).$$

Thus, if $X \sim N(0,\,\sigma_X^2)$ and $Y \sim N(0,\,\sigma_Y^2)$ are independent,

$$X+Y \;\sim\; N(0,\,\sigma_X^2+\sigma_Y^2).$$

□

# The sum of two normal random variables

**Proof of proposition continued:** Let $X$ and $Y$ be independent,

$$X \sim N(\mu_X, \sigma_X^2) \text{ and } Y \sim N(\mu_Y, \sigma_Y^2).$$

Recall that if $Z \sim N(\mu, \sigma^2)$, then $Z + a \sim N(\mu + a, \sigma^2)$.

Let $X' = X - \mu_X \sim N(0, \sigma_X^2)$ and $Y' = Y - \mu_Y \sim N(0, \sigma_Y^2)$.

These RVs are independent with mean zero, so we have proved that

$$X' + Y' \sim N(0, \sigma_X^2 + \sigma_Y^2).$$

Hence

$$\begin{aligned} X + Y &= X' + \mu_X + Y' + \mu_Y \\ &\sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \end{aligned}$$

□

# The sum of $n$ i.i.d. normal random variables

## Proposition

*Suppose* $X_1, \ldots, X_n$ *are independent, identically distributed with* $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$. *Then*

$$\sum_{i=1}^{n} X_i \ \sim \ N(n\,\mu, \, n\,\sigma^2).$$

## Proof.

By induction.

The previous result covers the case $n = 2$.

Given

$$\sum_{i=1}^{n-1} X_i \ \sim \ N((n-1)\,\mu, \, (n-1)\,\sigma^2),$$

apply the previous result to the two RVs $\sum_{i=1}^{n-1} X_i$ and $X_n$. □

# The mean of $n$ i.i.d. normal (or other) random variables

Suppose $X_1, \ldots, X_n$ are independent, identically distributed with

$$X_i \sim N(\mu,\, \sigma^2), \quad i = 1, \ldots, n.$$

Then

$$\bar{X} \;=\; \frac{1}{n} \sum_{i=1}^{n} X_i \;\sim\; N(\mu,\, \frac{\sigma^2}{n}).$$

More generally, for any i.i.d. RVs $X_1, \ldots, X_n$, with $\mathsf{E}(X_i) = \mu$ and $\mathsf{Var}(X_i) = \sigma^2$, $i = 1, \ldots, n$,

$$\mathsf{E}(\bar{X}) \;=\; \mu \quad \text{and} \quad \mathsf{Var}(\bar{X}) \;=\; \frac{\sigma^2}{n}$$

and the Central Limit Theorem implies the distribution of $\bar{X}$ is *approximately* normal for large $n$.

# Where we are — Lecture 17

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

**VI** Sums of random variables

**VII** Estimation

- **VII.a** Introduction to model fitting
- **VII.b** Estimation by the method of moments
- **VII.c** Estimates and estimators
- **VII.d** Maximum likelihood estimation
- **VII.e** Sampling distributions, bias and mean square error
- **VII.f** Assessment of goodness of fit

# VII Estimation

We have seen how a specified model leads to random samples.

$$\text{Model} \quad \rightarrow \quad \text{Data}$$

We now consider the situation where we have observed data and wish to learn about the model that produced them.

This is the transition from Probability to Statistics.

Note that the word "data" is plural.

**Definition**

| | |
|---|---|
| Datum | One piece of information |
| Data | Several or many pieces of information |

So, write "the data are", not "the data is".

# Learning about the model that generated the data

The problem is most interesting when data concern several variables.

We could consider the relation between the variables

$$\begin{aligned} H &= \text{Height} \\ W &= \text{Weight} \\ S &= \text{Systolic blood pressure} \end{aligned}$$

measured on a set of individuals.

We can add the variable

$$X = \text{Subject has a heart attack in the next 5 years}$$

and consider the dependence of $X$ on $H$, $W$ and $S$.

# Learning about the model that generated the data

Suppose the analysis of these data shows that the likelihood of a heart attack is increased if an individual has a high value of

$$S = \text{Systolic Blood Pressure}$$

or a high value of

$$\text{Body Mass Index} = \frac{W}{H^2},$$

where $W$ is measured in $kg$ and $H$ in $m$.

If interventions (diet or medication) are made to reduce $S$ or $BMI$, will this reduce the risk of a heart attack?

In a study of such interventions, care should be taken to allow for other risk factors, e.g., age and family history of heart disease.

Cholesterol level is another risk factor — this could be treated too.

For proper control of variation between the individuals in a study, a **Randomised Clinical Trial** is the best approach.

# VII.a Introduction to model fitting

We start with the simple case of modelling the distribution of a single random variable.

Here, we have to choose a suitable distribution and estimate parameters of this distribution.

## Example

Two fish are randomly positioned in a $30cm \times 50cm \times 100cm$ aquarium, and their positions are independent.

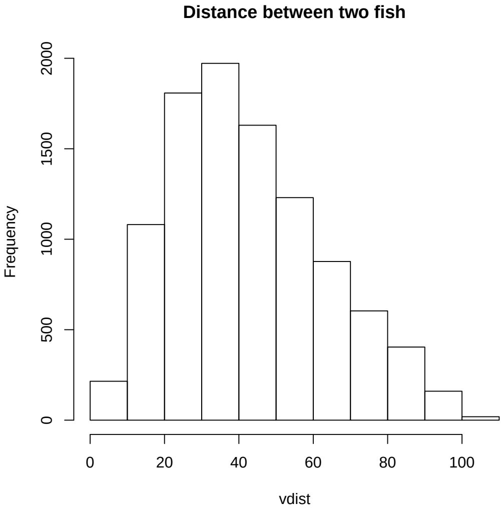
The random variable $Y$ is the distance (in $cm$) between the fish.

This is a similar example to the distance between two flies on a sphere (Computer Lab Sheet 3, Problem 1) but the distribution has a rather different shape.

Can we model the distribution of $Y$, at least approximately, as some standard distribution?

# The distance between two fish

A histogram of 1000 realisations of the distance variable $Y$.



# The distance between two fish

In these data, the minimum observed value of $Y$ was 1.6 and the maximum was 108.8.

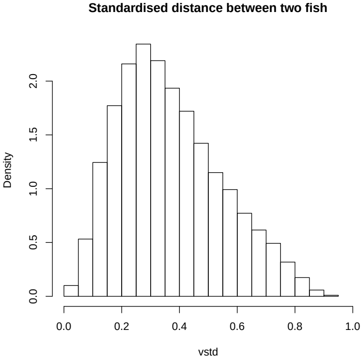
The range of possible values for $Y$ is from 0 to

$$\sqrt{\{30^2 + 50^2 + 100^2\}} \ = \ 115.8.$$

We define the standardised distance between the fish to be

$$X \ = \ \frac{Y}{115.8}$$

and re-draw the histogram.



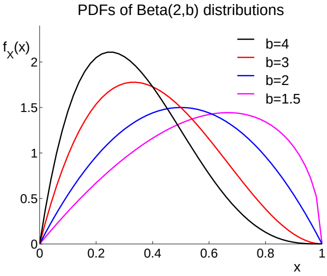
# The distance between two fish

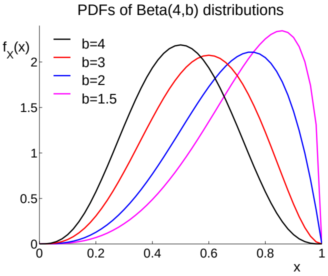
One standard distribution for a random variable taking values in the interval $(0,1)$ is the Beta distribution.

The $\mathrm{Beta}\,(a,b)$ distribution has PDF

$$f_X(x) \;=\; \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,x^{a-1}\,(1-x)^{b-1} \qquad \text{for } x \in (0,1),$$

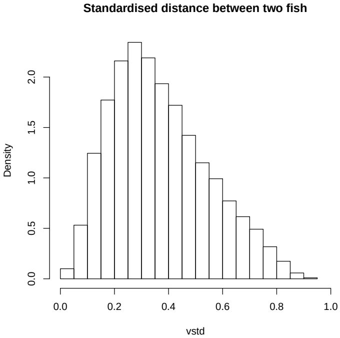
and $f_X(x) = 0$ otherwise.


PDFs of Beta(2,b) distributions


PDFs of Beta(4,b) distributions

# Fitting a Beta distribution

We want to find values for the parameters $a$ and $b$ such that the $\mathrm{Beta}\,(a, b)$ distribution matches the histogram for our data.



We could use trial and error to find values for $a$ and $b$ for which the PDF $f_X(x)$ follows the histogram (plotted as a density by giving the `prob=TRUE` command).

We shall take a more systematic approach.

# Fitting a Beta distribution

We shall match properties of the $\mathrm{Beta}\,(a, b)$ distribution to properties of our sample.

This is known as estimation by "The Method of Moments".

If $X \sim \mathrm{Beta}\,(a, b)$, it can be shown that

$$\mathsf{E}(X) \;=\; \frac{a}{a+b} \quad \text{and} \quad \mathsf{Var}(X) \;=\; \frac{ab}{(a+b+1)\,(a+b)^2}\,.$$

Hence

$$\mathsf{E}(X^2) \;=\; \mathsf{Var}(X) + [\mathsf{E}(X)]^2 \;=\; \frac{ab}{(a+b+1)\,(a+b)^2} + \frac{a^2}{(a+b)^2}\,.$$

The average of the observed values $x_1, \ldots, x_{1000}$ is 0.368.

The average value of $x_i^2$ for our sample is 0.167.

# Fitting a Beta distribution

Equating sample averages of $x_i$ and $x_i^2$ to $\mathsf{E}(X)$ and $\mathsf{E}(X^2)$ gives

$$\frac{a}{a+b} = 0.368 \tag{8}$$

and

$$\frac{ab}{(a+b+1)\,(a+b)^2} + \frac{a^2}{(a+b)^2} = 0.167. \tag{9}$$

Substituting (8) into (9), gives

$$\frac{ab}{(a+b+1)\,(a+b)^2} = 0.167 - 0.368^2 = 0.032.$$

Hence

$$a+b+1 = \left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)\frac{1}{0.032} = \frac{0.368 \times (1-0.368)}{0.032},$$

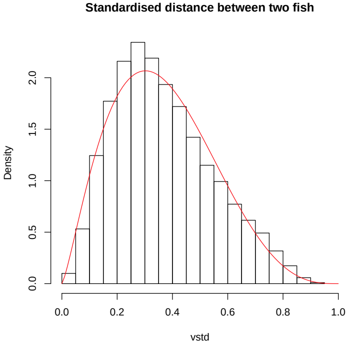
which implies $a+b = 6.3$.

# Fitting a Beta distribution

Combining $a/(a+b) = 0.368$ and $a+b = 6.3$, we get

$$a = 2.3 \quad \text{and} \quad b = 4.0.$$

Superimposing the $\mathrm{Beta}\,(2.3, 4.0)$ distribution on the histogram of $X$ values gives a reasonably good fit.



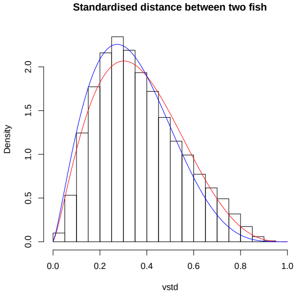
# Fitting a Beta distribution

The R command `help(rbeta)` provides information about a "Non-central Beta distribution".

Experimenting with this, we find $a = 2.3$ and $b = 5.0$ along with a non-centrality parameter of 0.5 fits the mode of the sample better.

The same is true for further sets of simulated data.

Superimposing the $\mathrm{Beta}\,(2.3, 5.0, 0.5)$ distribution for our example produces the blue curve in the figure below.



# Fitting a Beta distribution

We have modelled $X = Y/115.8$ as $X \sim \mathrm{Beta}\,(a, b)$.

A model for the original distance $Y$ follows since

$$Y \ \sim \ 115.8\,\mathrm{Beta}\,(a, b).$$

Even using a Non-central Beta distribution, it is difficult to match both the mode and the upper tail of the histogram.

We conclude that the data do not follow a Beta distribution exactly — but this might be a useful approximation for some purposes.

Suppose you are asked to estimate $\mathsf{E}(Y)$:

Having a particular distribution in mind does not necessarily help — in fitting the Beta distribution we effectively estimated $\mathsf{E}(X)$ by $(x_1 + \ldots + x_{1000})/1000$ in order to fit values for $a$ and $b$.

However, in other cases, there may be better ways to estimate $\mathsf{E}(X)$ than by the sample mean.

# Mean and variance of a Beta distribution SELF STUDY SLIDE

The $\mathrm{Beta}\,(a,b)$ distribution has PDF $f_X(x)=0$ for $x<0$ and $x>1$, and

$$f_X(x) \;=\; \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,x^{a-1}\,(1-x)^{b-1} \qquad \text{for } x\in(0,1).$$

Hence

$$\begin{aligned}\mathsf{E}(X) \;&=\; \int_0^1 x\,\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,x^{a-1}\,(1-x)^{b-1}\,dx \\ &=\; \ldots \;=\; \frac{a}{a+b}\end{aligned}$$

and

$$\begin{aligned}\mathsf{E}(X^2) \;&=\; \int_0^1 x^2\,\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,x^{a-1}\,(1-x)^{b-1}\,dx \\ &=\; \ldots \;=\; \frac{a(a+1)}{(a+b)(a+b+1)}.\end{aligned}$$

*See calculations on board*

# Where we are

Lecture 18

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

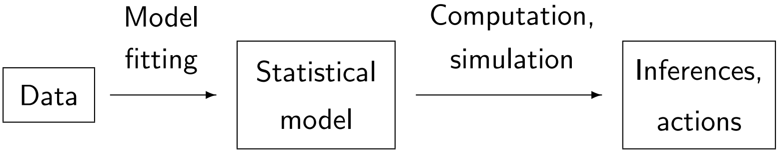
**VI** Sums of random variables

**VII** Estimation

- **VII.a** Introduction to model fitting
- **VII.b** Estimation by the method of moments
- **VII.c** Estimates and estimators
- **VII.d** Maximum likelihood estimation
- **VII.e** Sampling distributions, bias and mean square error
- **VII.f** Assessment of goodness of fit

# VII.a Introduction to model fitting

Given observed data, we want to set up a statistical model that describes the data, then use this model to make inferences or guide future actions.



With i.i.d. observations of a single random variable, we need to

- Choose a suitable distribution:

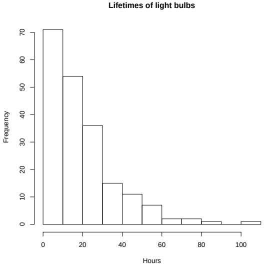
  Discrete or continuous

  Real valued, positive, or taking values in an interval

- Estimate parameters of this distribution

# Estimating parameter values

Suppose a graphical display suggests that the data follow an exponential distribution.



How should we estimate the rate parameter $\lambda$?

We shall look at two methods of estimation:

The Method of Moments,

Maximum Likelihood Estimation.

# Method of Moments: "Distance between fish" example

We assumed observations, $X_1, \ldots, X_n$, to be i.i.d. RVs following a $\mathrm{Beta}\,(a, b)$ distribution. We needed two pieces of information to estimate two parameters, $a$ and $b$.

We found formulae for $\mathsf{E}(X)$ and $\mathsf{E}(X^2)$ when $X \sim \mathrm{Beta}\,(a, b)$,

$$\mathsf{E}(X) \;=\; \frac{a}{a+b}\,, \quad \mathsf{E}(X^2) \;=\; \frac{ab}{(a+b+1)\,(a+b)^2} + \left(\frac{a}{a+b}\right)^2 .$$

With observed values $x_1, \ldots, x_n$, we set up equations matching the sample means of $x_i$ and $x_i^2$ to expected values $\mathsf{E}(X)$ and $\mathsf{E}(X^2)$,

$$\frac{a}{a+b} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \frac{ab}{(a+b+1)\,(a+b)^2} + \left(\frac{a}{a+b}\right)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Denoting the solutions of this pair of equations by $\hat{a}$ and $\hat{b}$, we concluded that, approximately, each $X_i \sim \mathrm{Beta}\,(\hat{a}, \hat{b})$.

# VII.b Estimation: the Method of Moments

### Definition

For $k > 0$, the $k$th moment of a RV $X$ is

$$\mu_k \;=\; \mathsf{E}(X^k) \quad \text{(if this exists).}$$

So $\mu_0 = 1$, $\mu_1 = \mathsf{E}(X)$, $\mu_2 = \mathsf{E}(X^2)$, etc.

Suppose we observe random variables $X_1, \ldots, X_n$.

Denote the observed values of these variables by $x_1, \ldots, x_n$.

We assume $X_1, \ldots, X_n$ are i.i.d. and follow a specific form of distribution which involves unknown parameters $\theta_1, \ldots, \theta_p$.

We estimate the values of $\mu_1, \ldots, \mu_p$ from the observed data.

Then, we find $\theta_1, \ldots, \theta_p$ that give these values for $\mu_1, \ldots, \mu_p$.

# Estimating $\mu_1$

Assume $\mathsf{E}(X_1) = \mu_1$ and $\mathsf{Var}(X_1) = \sigma^2$ both exist.

Each $X_i$ has mean $\mu_1$ and variance $\sigma^2$, so

$$\mathsf{E}\left(\sum_{i=1}^{n} X_i\right) \;=\; n\,\mu_1 \quad \text{and} \quad \mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) \;=\; n\,\sigma^2.$$

Thus,

$$\mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \mu_1 \quad \text{and} \quad \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \sigma^2/n.$$

By the Central Limit Theorem,

$$\overline{\mathsf{X}} \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i$$

behaves like a $N(\mu_1, \sigma^2/n)$ RV, for large $n$.

# Estimating $\mu_1$

We have, approximately, for large $n$,

$$\mathsf{\overline{X}} - \mu_1 \sim N(0, \frac{\sigma^2}{n}) \quad \text{and} \quad (\mathsf{\overline{X}} - \mu_1)\,\frac{\sqrt{n}}{\sigma} \sim N(0,1).$$

Thus, for large $n$,

$$\begin{aligned} \mathsf{P}\left(\left|\mathsf{\overline{X}} - \mu_1\right| > \epsilon\right) &= \mathsf{P}\left(\left|\mathsf{\overline{X}} - \mu_1\right| \frac{\sqrt{n}}{\sigma} > \epsilon\,\frac{\sqrt{n}}{\sigma}\right) \\ &\approx \mathsf{P}\left(|Z| > \epsilon\,\frac{\sqrt{n}}{\sigma}\right), \end{aligned}$$

where $Z \sim N(0,1)$, and this probability $\to 0$ as $n \to \infty$.

Hence, for large $n$, with high probability $\mathsf{\overline{X}} \approx \mu_1$.

Thus, it is reasonable to approximate $\mu_1$ by $\bar{\mathsf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where $x_1, \ldots, x_n$ are the observed values of $X_1, \ldots, X_n$.

# Estimating $\mu_k$ for $k > 1$

We can apply the same reasoning to RVs $X_i^k$, $i = \ldots, n$, assuming both $\mathsf{E}(X_i^k) = \mu_k$ and $\mathsf{Var}(X_i^k)$ exist

We conclude that, with high probability,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \approx \mu_k$$

and it is reasonable to approximate $\mu_k$ by $\frac{1}{n}\sum_{i=1}^{n} x_i^k$.

# Method of Moments estimation

Suppose a probability distribution has unknown parameters $\theta_1, \ldots, \theta_p$ and we observe i.i.d. observations $X_1, \ldots, X_n$.

## Definition

Method of Moments Estimates of $\theta_1, \ldots, \theta_p$ are the solutions to

$$\frac{1}{n} \sum_{i=1}^{n} x_i^k \;=\; \mu_k \;=\; \mu_k(\theta_1, \ldots, \theta_p) \quad \text{for } k = 1, \ldots, p,$$

where $x_1, \ldots, x_n$ denote the observed values of $X_1, \ldots, X_n$.

For large $n$, there is a high probability that each $\frac{1}{n} \sum_{i=1}^{n} X_i^k$ is close to the true value of $\mu_k$, $k = 1, \ldots, p$.

Hence, it is also likely that the estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_p$ are close to the true values $\theta_1, \ldots, \theta_p$.

# Fitting an exponential distribution

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathrm{Exp}(\lambda)$ RVs and we wish to estimate the parameter $\lambda$.

The first moment of each $X_i$, $i = 1, \ldots, n$, is

$$\mathsf{E}(X_i) \;=\; \frac{1}{\lambda}.$$

With observed values $x_1, \ldots, x_n$, we solve

$$\frac{1}{n} \sum_{i=1}^{n} x_i \;=\; \frac{1}{\lambda}\,. \tag{10}$$

It is standard practice to denote an estimate by placing a “hat” over the name of the parameter.

So, from (10), we obtain the estimate

$$\widehat{\lambda} \;=\; \frac{n}{x_1 + \ldots + x_n}\,.$$

# Fitting a Uniform distribution

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathrm{Unif}(0, \theta)$ RVs and we wish to estimate the parameter $\theta$.

The first moment of each $X_i$, $i = 1, \ldots, n$, is

$$\mathsf{E}(X_i) \;=\; \frac{\theta}{2}.$$

With observed values $x_1, \ldots, x_n$, we solve

$$\frac{1}{n} \sum_{i=1}^{n} x_i \;=\; \frac{\theta}{2}$$

to obtain the estimate

$$\widehat{\theta} \;=\; \frac{2\,(x_1 + \ldots + x_n)}{n}\,.$$

# Fitting a normal distribution

Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ RVs and we wish to estimate the parameters $\mu$ and $\sigma^2$.

The first moment of each $X_i$, $i = 1, \ldots, n$, is

$$\mathsf{E}(X_i) \;=\; \mu.$$

The second moment of each $X_i$ is

$$\mathsf{E}(X_i^2) \;=\; \mathsf{Var}(X_i) + [\mathsf{E}(X_i)]^2 \;=\; \sigma^2 + \mu^2.$$

So, we need to solve

$$\frac{1}{n} \sum_{i=1}^{n} x_i \;=\; \mu \tag{11}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \;=\; \sigma^2 + \mu^2. \tag{12}$$

# Fitting a normal distribution

Defining the random variable

$$\overline{\mathsf{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the observed value of this RV,

$$\overline{\mathsf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

From (11), we have

$$\widehat{\mu} = \overline{\mathsf{x}}$$

and substituting this into (12) gives

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{\mathsf{x}}^2.$$

# VII.c Estimates and estimators

We have used upper and lower case letters to distinguish between:

the name of a random variable, $X$, and

the value $x$ that this variable takes.

Similarly, we wish to distinguish between:

a parameter estimate viewed as a random variable and

the value (i.e., a number) this estimate takes for given data.

The first of these is defined in terms of the RVs $X_1, \ldots, X_n$.

The second is defined in terms of observed values $x_1, \ldots, x_n$.

# Estimates and estimators

## Definition

An estimate is a real number computed from the data.

An estimate of the parameter $\theta$ based on observations $x_1, \ldots, x_n$ can be written as

$$\widehat{\theta} \;=\; h(x_1, \ldots, x_n)$$

for the appropriate function $h$.

## Definition

An estimator is a random variable, a function of the random variables $X_1, \ldots, X_n$ that comprise the data.

If the estimate of $\theta$ based on observations $x_1, \ldots, x_n$ is $\widehat{\theta} = h(x_1, \ldots, x_n)$, the estimator is the random variable

$$h(X_1, \ldots, X_n).$$

# A notational quandary

What name should we give to the estimator $h(X_1, \ldots, X_n)$?

For consistency, we ought to use the upper case version of the estimate $\widehat{\theta}$ — which would be $\widehat{\Theta}$.

However, it is not usual to use upper case Greek letters in this way.

We can introduce a new name altogether, e.g.,

$$T = h(X_1, \ldots, X_n).$$

This enables to talk about the estimator as a random variable and write down its "sampling distribution", e.g.,

$$T \sim N(\theta, \frac{\sigma^2}{n}). \tag{13}$$

It is tempting to use $\widehat{\theta}$ in place of $T$ in (13) — but then we would be using the same symbol both for the name of a random variable and for the value it takes.

# Where we are — Lecture 19

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

**VI** Sums of random variables

**VII** Estimation

- **VII.a** Introduction to model fitting
- **VII.b** Estimation by the method of moments
- **VII.c** Estimates and estimators
- **VII.d** Maximum likelihood estimation
- **VII.e** Sampling distributions, bias and mean square error
- **VII.f** Assessment of goodness of fit

# VII.d Maximum likelihood estimation

Suppose the random variables $X_1, \ldots, X_n$ follow a certain type of distribution but the parameters of that distribution are unknown.

As an example, if we assume $X_i \sim \mathrm{Exp}(\lambda)$, $i = 1, \ldots, n$, there is one unknown parameter, $\lambda$, to estimate.

### Definition

Let $X_1, \ldots, X_n$ be i.i.d. *continuous* RVs with PDF $f_X(\theta, x)$, where $\theta$ denotes a parameter or a vector of parameters.

Suppose values $X_i = x_i$, $i = 1, \ldots, n$, are observed, and define $\underset{\sim}{x} = (x_1, \ldots, x_n)$.

The **likelihood** function for these continuous RVs is

$$L(\theta, \underset{\sim}{x}) \;=\; \prod_{i=1}^{n} f_X(\theta, x_i).$$

# Maximum likelihood estimation

## Definition

Let $X_1, \ldots, X_n$ be i.i.d. *discrete* RVs for which

$$\mathsf{P}(X = x) = p_X(\theta, x), \quad x \in \Omega,$$

where $\theta$ denotes a parameter or a vector of parameters.

Suppose values $X_i = x_i$, $i = 1, \ldots, n$, are observed, and let $\underset{\sim}{x} = (x_1, \ldots, x_n)$.

The **likelihood** function for these discrete RVs is

$$L(\theta, \underset{\sim}{x}) \;=\; \prod_{i=1}^{n} p_X(\theta, x_i).$$

# Maximum likelihood estimation

### Definition

In both the continuous and discrete cases, the log-likelihood function is

$$\mathcal{L}(\theta, \underset{\sim}{x}) \; = \; \log\{L(\theta, \underset{\sim}{x})\}.$$

### Definition

The **Maximum Likelihood Estimate** (MLE) of a parameter or vector of parameters, $\theta$, is the value of $\theta$ that maximises the likelihood function $L(\theta, \underset{\sim}{x})$ for the observed data.

Equivalently, the MLE is the value of $\theta$ that maximises the log-likelihood, $\mathcal{L}(\theta, \underset{\sim}{x})$.

# Maximum likelihood estimation

Why is the Maximum Likelihood Estimate (MLE) a good choice?

A parameter value under which the observed data are “likely” seems plausible.

Case by case, one can usually show this estimate has good properties.

Theory for large sample sizes shows that maximum likelihood estimation is an efficient, all purpose method (see Statistics 2a and later courses where maximum likelihood estimation is used in more complex models).

There are exceptional situations where other methods may be preferable, even with large sample sizes: this tends to be the case when the set of values for which $f_X(\theta, x) > 0$ depends on $\theta$.

# Maximum likelihood estimation: Bernoulli observations

## Example

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathrm{Bernoulli}(p)$ random variables, so

$$\mathsf{P}(X_i = x) \; = \; \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Let

$$S \; = \; \sum_{i=1}^{n} X_i \quad \text{and} \quad s \; = \; \sum_{i=1}^{n} x_i.$$

Calculations show the maximum likelihood estimate is

$$\widehat{p} \; = \; \frac{s}{n} \; = \; \frac{1}{n} \sum_{i=1}^{n} x_i.$$

*See calculations on board*

# Maximising $L(\theta, \underset{\sim}{x})$ over possible values of $\theta$

In many cases (but not all), we find the MLE of $\theta$ by solving

$$\frac{d}{d\theta} L(\theta, \underset{\sim}{x}) \; = \; 0.$$

We still need to prove this gives a maximum of $L(\theta, \underset{\sim}{x})$.

Possible arguments are:

(i) $d^2 L(\theta, \underset{\sim}{x})/d\theta^2 < 0$ for all $\theta$;

(ii) $d\, L(\theta, \underset{\sim}{x})/d\theta = 0$ at $\theta = \tilde{\theta}$,

$d\, L(\theta, \underset{\sim}{x})/d\theta > 0$ for $\theta < \tilde{\theta}$, and

$d\, L(\theta, \underset{\sim}{x})/d\theta < 0$ for $\theta > \tilde{\theta}$;

(iii) $L(\theta, \underset{\sim}{x}) > 0$ for all $\theta$, $L(\theta, \underset{\sim}{x}) \to 0$ as $\theta \to -\infty$ and $\theta \to \infty$,

# Maximising $L(\theta, \underset{\sim}{x})$ over possible values of $\theta$

Sometimes it is simpler to work with $\mathcal{L}(\theta, \underset{\sim}{x}) \; = \; \log\{L(\theta, \underset{\sim}{x})\}$.

Consider the Bernoulli example, where

$$\mathcal{L}(p, \underset{\sim}{x}) \; = \; s \log(p) + (n-s) \log(1-p).$$

For $1 \leq s \leq n-1$,

$$\frac{d}{dp} \mathcal{L}(p, \underset{\sim}{x}) \; = \; \frac{s}{p} - \frac{n-s}{1-p}$$

and

$$\frac{d^2}{dp^2} \mathcal{L}(p, \underset{\sim}{x}) \;\; = \;\; \frac{-s}{p^2} - \frac{n-s}{(1-p)^2} \;\; < \;\; 0.$$

The solution to $d\mathcal{L}(p, \underset{\sim}{x})/dp = 0$ is $\widehat{p} = s/n$.

Since $d^2\mathcal{L}(p, \underset{\sim}{x})/dp^2 < 0$, $\widehat{p} = s/n$ gives the maximum of $\mathcal{L}(\theta, \underset{\sim}{x})$

— and therefore of $L(\theta, \underset{\sim}{x})$.

# Maximum likelihood estimation: Normal observations

## Example

Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables.

Find the maximum likelihood estimate of $\theta = (\mu, \sigma^2)$.

We have

$$\begin{aligned} L(\theta, \underset{\sim}{x}) &= \prod_{i=1}^{n} f_X(\theta, x_i) \\ &= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\}, \end{aligned}$$

and

$$\mathcal{L}(\theta, \underset{\sim}{x}) = \text{constant} - n\log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

# Maximum likelihood estimation: Normal observations

We shall find the value of $\theta = (\mu, \sigma^2)$ that maximises $\mathcal{L}(\theta, \underset{\sim}{x})$.

First, we consider the problem of maximising $\mathcal{L}(\theta, \underset{\sim}{x})$ with respect to $\mu$ for a fixed value of $\sigma^2$.

Then we shall maximise over $\sigma^2$.

**Step 1**

For given $\sigma^2$, we maximise

$$\mathcal{L}(\theta, \underset{\sim}{x}) \;=\; \text{constant} \;-\; n\log(\sigma) \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

over $\mu$. To do this, we need to minimise

$$\sum_{i=1}^{n}(x_i - \mu)^2.$$

# Maximum likelihood estimation: Normal observations

Note that

$$\frac{d}{d\mu}\sum_{i=1}^{n}(x_i-\mu)^2 \;=\; -\sum_{i=1}^{n}2(x_i-\mu)$$

and

$$\frac{d^2}{d\mu^2}\sum_{i=1}^{n}(x_i-\mu)^2 \;=\; -\sum_{i=1}^{n}(-2) \;\;>\; 0.$$

So, the minimum of $\sum_{i=1}^{n}(x_i-\mu)^2$ is the solution to

$$\frac{d}{d\mu}\sum_{i=1}^{n}(x_i-\mu)^2 \;=\; 0,$$

which has $n\mu \;=\; \sum_{i=1}^{n}x_i$, i.e., $\mu \;=\; \frac{1}{n}\sum_{i=1}^{n}x_i \;=\; \bar{\mathsf{x}}$.

# Maximum likelihood estimation: Normal observations

**Step 2**

Remember, our aim is to maximise the log likelihood over values of $\theta = (\mu,\, \sigma^2)$.

We have seen that, for a given value of $\sigma^2$, the log likelihood is maximised by taking $\mu = \bar{\mathsf{x}}$.

Since we get the same answer for any $\sigma^2$, the MLE of $\mu$ is $\widehat{\mu} = \bar{\mathsf{x}}$.

We now maximise the log likelihood over $\sigma^2$ when $\mu = \bar{\mathsf{x}}$, so

$$\begin{aligned}\mathcal{L}(\theta, \underset{\sim}{x}) &= \text{constant} - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{\mathsf{x}})^2 \\ &= c - n\log(\sigma) - A/(2\sigma^2),\end{aligned}$$

where $A = \sum_{i=1}^{n}(x_i - \bar{\mathsf{x}})^2$.

# Maximum likelihood estimation: Normal observations

SELF STUDY SLIDE

Consider minimising

$$h(\sigma) \;=\; n\log(\sigma) + A/(2\sigma^2),$$

where $A$ is positive.

The minimum occurs when

$$\sigma^2 \;=\; \frac{A}{n}.$$

*See calculations on board*

So, we have

$$\widehat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathsf{x}})^2$$

to go with $\widehat{\mu} = \bar{\mathsf{x}}$.

# Maximum likelihood estimation: Normal observations

We have derived the maximum likelihood estimates of $\mu$ and $\sigma^2$,

$$\widehat{\mu} = \bar{\mathsf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2.$$

Recall that the method of moments gave

$$\widehat{\mu} = \bar{\mathsf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \widehat{\mu}^2.$$

The two versions of $\widehat{\sigma}^2$ are, in fact, the same since

$$\sum_{i=1}^{n}(x_i - \bar{\mathsf{x}})^2 = \sum_{i=1}^{n}(x_i^2 - 2\,x_i\,\bar{\mathsf{x}} + \bar{\mathsf{x}}^2) = \left(\sum_{i=1}^{n} x_i^2\right) - n\,\bar{\mathsf{x}}^2.$$

In general, these two methods **can** lead to different estimates.

# Where we are

- **I** Introduction
- **II** Random variables and cumulative distribution functions
- **III** Important families of continuous random variables
- **IV** Joint distributions, independence and expectation
- **V** Transformations of random variables and simulation
- **VI** Sums of random variables
- **VII** Estimation
  - **VII.a** Introduction to model fitting
  - **VII.b** Estimation by the method of moments
  - **VII.c** Estimates and estimators
  - **VII.d** Maximum likelihood estimation
  - **VII.e** Sampling distributions, bias and mean square error
  - **VII.f** Assessment of goodness of fit

# VII.e Sampling distributions, bias and mean square error

Example

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathrm{Unif}(0, \theta)$ RVs, where the parameter $\theta > 0$ is unknown.

**Method of moments estimation**

We saw in Lecture 19, the method of moments estimator for $\theta$ is

$$T_1(X_1, \ldots, X_n) \;=\; \frac{2}{n} \sum_{i=1}^{n} X_i.$$

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

**Maximum likelihood estimation**

We can write

$$f_X(x) = \begin{cases} \frac{1}{\theta}\,\mathsf{I}(x \le \theta) & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$$

where $\mathsf{I}(x \le \theta) = 1$ if $x \le \theta$ and 0 otherwise.

Let $\underset{\sim}{x} = (x_1, \ldots, x_n)$, where each $x_i \ge 0$.

The likelihood function is

$$\begin{aligned} L(\theta, \underset{\sim}{x}) &= \prod_{i=1}^{n} \frac{1}{\theta}\,\mathsf{I}(x_i \le \theta) \\ &= \frac{1}{\theta^n}\,\mathsf{I}(\theta \ge \max{}_{i=1,\ldots,n}\{x_i\}). \end{aligned}$$

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

We have

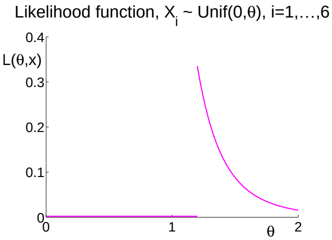
$$L(\theta, \underset{\sim}{x}) \ = \ \frac{1}{\theta^n}\, \mathrm{I}(\theta \geq \max_{i=1,\ldots,n}\{x_i\}).$$

With observed data

$$\underset{\sim}{x} = (0.4, 0.12, 0.8, 1.05, 0.23, 1.2),$$

the maximum of $L(\theta, \underset{\sim}{x})$ is at

$$\widehat{\theta} \ = \ \max_{i=1,\ldots,6}\{x_i\} \ = \ 1.2.$$



In general, for data $\underset{\sim}{x} = (x_1, \ldots, x_n)$, the maximum likelihood **estimate** is $\max_{i=1,\ldots,n}\{x_i\}$.

Thus, the maximum likelihood **estimator** of $\theta$ is

$$T_2(X_1, \ldots, X_n) \ = \ \max_{i=1,\ldots,n}\{X_i\}.$$

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

So, we now have two possible estimators:

**Method of moments**

$$T_1(X_1, \ldots, X_n) \ = \ \frac{2}{n} \sum_{i=1}^{n} X_i$$

**Maximum likelihood**

$$T_2(X_1, \ldots, X_n) \ = \ \mathsf{max}_{i=1,\ldots,n} \{X_i\}$$

What should we consider when choosing between these estimators?

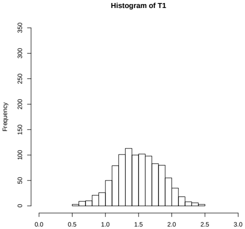
# Sampling distribution of an estimator

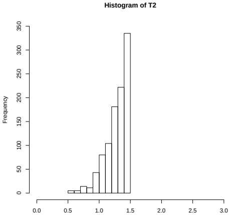
## Definition

The sampling distribution of an estimator $T(X_1, \ldots, X_n)$ is the distribution of the random variable $T$ — which follows from the definition of $T$ and the joint distribution of $X_1, \ldots, X_n$.

## Example

Simulations of $X_1, \ldots, X_6 \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$, with $\theta = 1.5$, gave 1000 values of the method of moments estimator $T_1(X_1, \ldots, X_6)$ and the maximum likelihood estimator $T_2(X_1, \ldots, X_6)$:





Note:

We always have $T_2 \leq \theta$.

On average, $T_2$ is closer to the true $\theta$ than $T_1$.

# Bias and precision

### Definition

An estimator $T(X_1, \ldots, X_n)$ of $\theta$ is unbiased if

$$\mathsf{E}(T) \;=\; \theta \quad \text{for all } \theta.$$

The bias of an estimator $T(X_1, \ldots, X_n)$ of $\theta$ is

$$\mathsf{Bias}(T) \;=\; \mathsf{E}(T) - \theta.$$

Note: Since $\mathsf{E}(T)$ and $\mathsf{Bias}(T)$ depend on $\theta$, we could indicate this by writing $\mathsf{E}_\theta(T)$ and $\mathsf{Bias}_\theta(T)$.

### Definition

The precision of an estimator $T(X_1, \ldots, X_n)$ of $\theta$ is

$$\mathsf{Precision}(T) \;=\; \frac{1}{\mathsf{Var}(T)}\,.$$

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

In our example, the **method of moments** estimator is

$$T_1 \;=\; \frac{2}{n} \sum_{i=1}^{n} X_i.$$

This has expected value

$$\mathsf{E}(T_1) \;=\; \frac{2}{n}\, n\, \frac{\theta}{2} \;=\; \theta,$$

so $T_1$ is an unbiased estimator of $\theta$.

Since

$$\mathsf{Var}(T_1) \;=\; \frac{4}{n^2}\, n\, \frac{\theta^2}{12} \;=\; \frac{\theta^2}{3\,n}\,,$$

the precision of $T_1$ is

$$\mathsf{Precision}(T_1) \;=\; \frac{3\,n}{\theta^2}\,.$$

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

In the same example, the **maximum likelihood** estimator is

$$T_2 \;=\; \mathsf{max}_{i=1,\ldots,n}\{X_i\}.$$

This estimator has CDF

$$F_{T_2}(t_2) \;=\; \mathsf{P}(T_2 \le t_2) \;=\; \left(\frac{t_2}{\theta}\right)^n,$$

and its PDF is

$$f_{T_2}(t_2) \;=\; \frac{d\,F_{T_2}(t_2)}{d\,t_2} \;=\; \frac{n\,t_2^{n-1}}{\theta^n} \quad \text{for } 0 \le t_2 \le \theta.$$

Hence, we can calculate

$$\mathsf{E}(T_2) \;=\; \frac{n}{n+1}\,\theta \quad \text{and} \quad \mathsf{Var}(T_2) \;=\; \frac{n}{(n+1)^2\,(n+2)}\,\theta^2.$$

*See calculations on board*

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

The **maximum likelihood** estimator $T_2$ has expectation

$$\mathsf{E}(T_2) \;=\; \frac{n}{n+1}\,\theta$$

so this estimator has bias

$$\mathsf{Bias}(T_2) \;=\; \mathsf{E}(T_2) - \theta \;=\; \frac{-\theta}{n+1}\,.$$

The variance of $T_2$ is

$$\mathsf{Var}(T_2) \;=\; \frac{n\,\theta^2}{(n+1)^2\,(n+2)} \;<\; \frac{\theta^2}{3\,n} \;=\; \mathsf{Var}(T_1).$$

However, low variance (and high precision) is not so helpful if an estimator's distribution is not centred on the true parameter value.

# Mean square error

## Definition

The mean square error of an estimator $T(X_1, \ldots, X_n)$ of $\theta$ is

$$\mathsf{MSE}(T) \;=\; \mathsf{E}\{(T-\theta)^2\}.$$

Now

$$\begin{aligned}
\mathsf{MSE}(T) \;&=\; \mathsf{E}\{\,([T-\mathsf{E}(T)] + [\mathsf{E}(T)-\theta])^2\,\} \\
&=\; \mathsf{E}\{[T-\mathsf{E}(T)]^2\} + 2\,\mathsf{E}\{T-\mathsf{E}(T)\}\,[\mathsf{E}(T)-\theta] \\
&\quad + [\mathsf{E}(T)-\theta]^2 \\
&=\; \mathsf{Var}(T) + 0 + (\mathsf{Bias}(T))^2.
\end{aligned}$$

So, the MSE combines bias and variance as

$$\mathsf{MSE}(T) \;=\; \mathsf{Var}(T) + (\mathsf{Bias}(T))^2.$$

# Mean square error

In our example, the **method of moments** estimator is unbiased and so has mean square error

$$\mathsf{MSE}(T_1) \;=\; \mathsf{Var}(T_1) \;=\; \frac{\theta^2}{3\,n}\,.$$

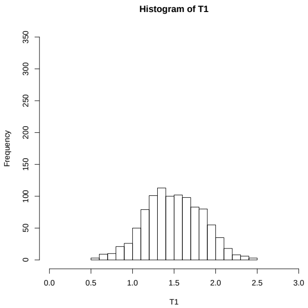
The **maximum likelihood** estimator has mean square error

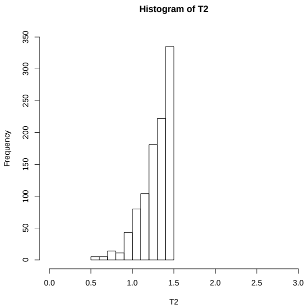
$$\begin{aligned}\mathsf{MSE}(T_2) &= \mathsf{Var}(T_2) + (\mathsf{Bias}(T_2))^2 \\ &= \frac{n\,\theta^2}{(n+1)^2\,(n+2)} + \frac{\theta^2}{(n+1)^2} \\ &= \frac{2\,\theta^2}{(n+1)\,(n+2)}\,.\end{aligned}$$

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

In our example, simulations were conducted with $n = 6$, for which

$$\mathsf{MSE}(T_1)/\mathsf{MSE}(T_2) = 14/9.$$





This ratio increases with $n$, so maximum likelihood estimation appears to be superior to the method of moments for this problem.

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

There is a simple way to modify the maximum likelihood estimate to make it unbiased. We know that

$$\mathsf{E}(T_2) \;=\; \frac{n}{n+1}\,\theta.$$

Thus, defining

$$T_3 \;=\; \frac{n+1}{n}\,T_2 \;=\; \frac{n+1}{n}\,\mathsf{max}_{i=1,\ldots,n}\,\{X_i\}$$

gives an unbiased estimator with variance

$$\mathsf{Var}(T_3) \;=\; \left(\frac{n+1}{n}\right)^2 \mathsf{Var}(T_2) \;=\; \frac{\theta^2}{n\,(n+2)}\,.$$

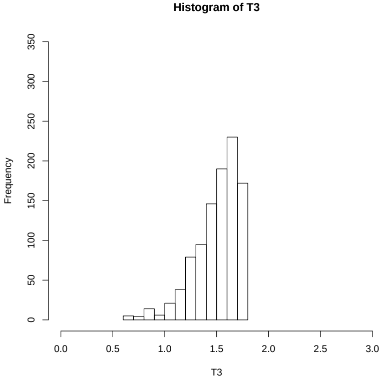
Since $T_3$ is unbiased, it has mean square error

$$\mathsf{MSE}(T_3) \;=\; \mathsf{Var}(T_3) \;=\; \frac{\theta^2}{n\,(n+2)},$$

which can be seen to be lower that that of $T_1$ and $T_2$.

# Example: $X_1, \ldots, X_n \sim$ i.i.d. $\mathrm{Unif}(0, \theta)$

The histogram shows result of 1000 simulated values of the modified maximum likelihood estimator, $T_3$.



Note that the true value $\theta = 1.5$ is now at the "centre of mass" of the distribution.

# Finding the best estimator

Later statistics courses will address the issues:

Finding the best possible estimator in terms of bias or MSE,

General theory of maximum likelihood estimation,

Proof that MLE gives best the possible estimators (for most problems) when the sample size is large,

Estimation in complex models involving several parameters.

# Where we are

Lecture 21

**I** Introduction

**II** Random variables and cumulative distribution functions

**III** Important families of continuous random variables

**IV** Joint distributions, independence and expectation

**V** Transformations of random variables and simulation

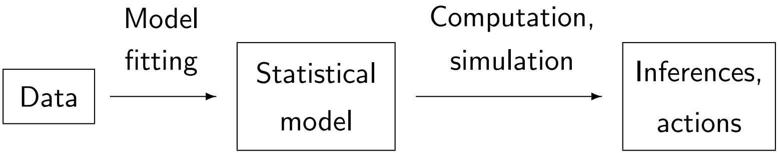
**VI** Sums of random variables

**VII** Estimation

- **VII.a** Introduction to model fitting
- **VII.b** Estimation by the method of moments
- **VII.c** Estimates and estimators
- **VII.d** Maximum likelihood estimation
- **VII.e** Sampling distributions, bias and mean square error
- **VII.f** Assessment of goodness of fit

# VII.f Assessment of goodness of fit

Recall the scheme for modelling data and drawing inferences:



Data

Model fitting

Statistical model

Computation, simulation

Inferences, actions

Within "Model fitting" we have taken a two step approach to model a random variable:

1. Select a family of distributions, e.g., exponential or normal
2. Estimate the parameter or parameters of this distribution.

We can extend this approach to allow validation of the choice of distribution to check that the model agrees with the observed data.

# Goodness of fit

An iterative approach to model fitting:

1. Select a family of distributions, e.g., exponential or normal
2. Estimate the parameter or parameters for this distribution.
3. Check whether the data are typical for the fitted distribution:

   If the data follow the fitted distribution well, STOP HERE.

   If not, go back to (1) and try another distribution.

One option on returning to step (1) is to consider a larger family:

The Weibull distribution contains the exponential as a special case but allows other possibilities too,

The Gamma family contains cases close to the normal distribution — plus a variety of asymmetric distributions.

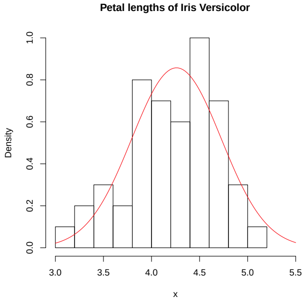
# Goodness of fit

In Problems Class 9, we looked at data comprising the petal length (in $cm$) for 50 specimens of Iris versicolor.

Supposing these lengths follow a $N(\mu, \sigma^2)$ distribution, the method of moments (or maximum likelihood) estimates are

$$\hat{\mu} = 4.26 \quad \text{and} \quad \hat{\sigma}^2 = 0.216.$$

Superimposing a $N(4.26, 0.216)$ PDF on the data histogram gives



Is this model a good enough fit to the data?

# Quantile-quantile plots

Quantile-quantile plots (q-q plots) avoid the “blocking” effect of histograms and so present data in more detail.

Suppose we observe i.i.d. RVs $X_1, \ldots, X_n$.

Denote the ordered values of the observed data $x_1, \ldots, x_n$ by

$$x_{(1)} \leq \ldots \leq x_{(n)}.$$

Let $F_X(x)$ be the CDF of a distribution proposed for the RVs $X_i$.

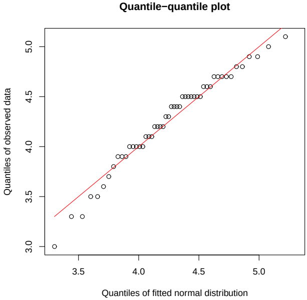
A q-q plot is a graph of

$$x_{(i)} \text{ against } F_X^{-1}\left(\frac{i}{n+1}\right).$$

If the RVs do indeed come from the distribution with CDF $F_X(x)$, then the q-q plot should be, approximately, a straight line through the origin with slope 1.

# Quantile-quantile plots

The q-q plot for the petal lengths of Iris versicolor is shown below.



Questions to consider:

1. Why do we expect a line, $y = x$, if the model is correct?
2. How close should a q-q plot be to a straight line?
3. What do departures from a straight line tell us?

# 1. Why we expect to see $x_{(i)} \approx F_X^{-1}\{i/(n+1)\}$

For given $x$, the number of observations out of $X_1, \dots, X_n$ taking values less than or equal to $x$ is

$$\mathrm{Binom}(n, F_X(x)).$$

The expected value of a $\mathrm{Binom}(n, F_X(x))$ RV is $n\,F_X(x)$, so we "expect" this many observations to be $\leq x$ — and the binomial variance indicates the likely variation in this number.

Applying this argument with $x = x_{(i)}$, we might expect $n\,F_X(x_{(i)})$ observations $\leq x_{(i)}$.

Since $i$ observations are less than or equal to $x_{(i)}$, this implies

$$\frac{i}{n} \;\approx\; F_X(x_{(i)}) \quad \text{and so} \quad x_{(i)} \;\approx\; F_X^{-1}\left(\frac{i}{n}\right).$$

# Why we expect to see $x_{(i)} \approx F_X^{-1}\{i/(n+1)\}$

Usually, we can ignore the probability that a continuous RV is exactly equal to a particular value.

However, we have just considered $x = x_{(i)}$ where, by definition, one of the RVs $X_1, \ldots, X_n$ has this as an observed value.

If we do not "count" the observation at $x_{(i)}$ we get

$$x_{(i)} \;\approx\; F_X^{-1}\left(\frac{i-1}{n}\right).$$

In fact, it can be shown (see Example Sheet 6, Question 5) that

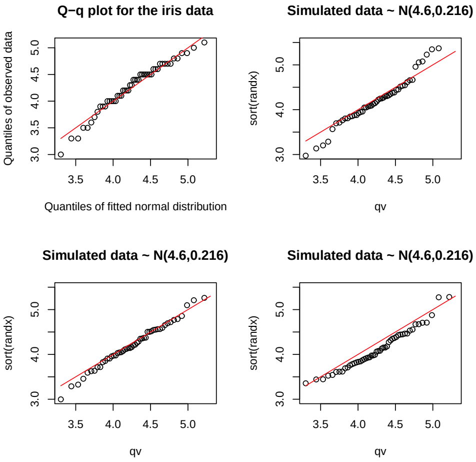
$$\mathsf{E}\{\, F_X(X_{(i)})\,\} \;=\; \frac{i}{n+1},$$

suggesting we might "expect" to see

$$x_{(i)} \;\approx\; F_X^{-1}\left(\frac{i}{n+1}\right).$$

# 2. How close should a q-q plot be to a straight line?

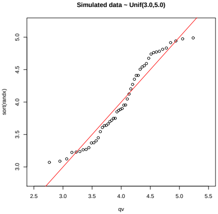
We can use simulation to see how a typical q-q plot might appear when a fitted model is actually correct. The following q-q plots are for the real iris data and three sets of fifty $N(4.26, 0.216)$ RVs:

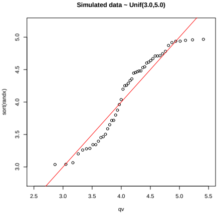


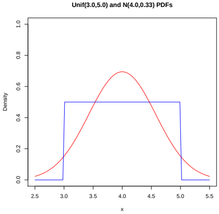
# 3. What a q-q plot can tell us

If the points in a q-q plot do not follow a straight line with slope 1, their pattern should indicate the failings of the assumed model.

Suppose observations from a $\mathrm{Unif}(3.0, 5.0)$ distribution are modelled by a normal distribution:







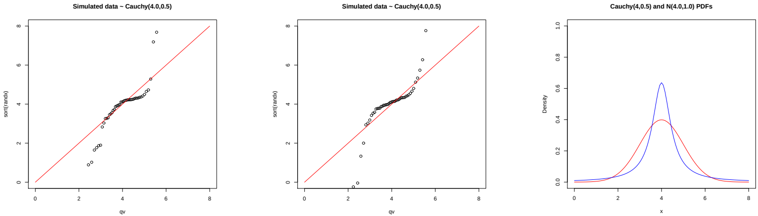
The “S” shaped q-q plots reflect the fact that the $\mathrm{Unif}(3.0, 5.0)$ distribution does not extend as far outwards as the normal.

Thus, the ordered values, $x_{(1)}, \ldots, x_{(50)}$, flatten off just above 3.0 and just below 5.0.

# What a q-q plot can tell us

Now consider observations from a Cauchy distribution centred on 4, with shape parameter 0.5.

The q-q plot tests whether these data can be fitted by a $N(4, 1)$ distribution.



The $N(4, 1)$ distribution matches the $\mathrm{Cauchy}\,(4, 0.5)$ distribution reasonably well at its centre, but the Cauchy PDF has heavier tails.

In fact, the tails are so long that in both simulated data sets, a few data points were off the vertical scale — by a long way!

# The role of simulation

Simulation plays a large role in modern statistics.

It is nice to derive exact formulae for probabilities and other properties of distributions — but that can be just about impossible for complex models.

Simulation provides a very powerful way to carry out calculations — basically, high dimensional integrals.

For some problems, even simulation is difficult — there is much current research on methods for simulating from complex, high dimensional distributions.

# What next?

**Statistics**

Modelling a response distribution in terms of other variables

Time series and forecasting — sequences of random variables

Multivariate data — multiple observations on each subject

Formal statistical inference

Applications of statistics:

Medicine — drug development, clinical trials

The environment, agriculture, biostatistics, social science, psychology, assessing risk . . .