ISE 2: PROBABILITY and STATISTICS

Lecturer

Dr Emma McCoy,

Room 522, Department of Mathematics, Huxley Building.

Phone: X 48553

email: e.mccoy@ic.ac.uk WWW: http://www.ma.ic.ac.uk/∼ejm/ISE.2.6/

Course Outline

1. Random experiments
2. Probability
   - axioms of probability
   - elementary probability results
   - conditional probability
   - the law of total probability
   - Bayes theorem
3. Systems reliability
4. Discrete random variables
   - Binomial distribution
   - Poisson distribution
5. Continuous random variables
   - uniform distribution

- Exponential distribution
- Normal distribution

6. Statistical Analysis
   - Point and interval estimation
7. Reliability and hazard rates

Textbooks

Statistics for the engineering and computer sciences, second edition. William Mendenhall and Terry Sincich, Dellen Publishing Company.

Probability concepts in engineering planning and design. Alfredo Ang and Wison Tang, Wiley.

1 Introduction – Random experiments

In general a *statistical analysis* may consist of:

- Summarization of data.
- Prediction.
- Decision making.
- Answering a specific question of interest.

We may view a statistical analysis as attempting to separate the *signal* in the data from the *noise*.

The accompanying handout of datasets gives two examples of types of data that we may be interested in.

1.1 Methods of summarizing data

Dataset 1

We need to find a simple *model* to describe the variability that we see.

This will allow us to answer questions of interest such as:

- What is the probability that 3 or more syntax errors will be present in another module?

Here the number of syntax errors is an example of a *discrete random variable* – it can only take one of a countable number of values.

Other examples of discrete random variables include:

- The toss of a coin where the outcome is Head/Tail.

- No of defective items found when 10 components are selected and tested.

Can we summarize dataset 1 in a more informative manner ?

Frequency Table:

NUMBER OF ERRORS	FREQUENCY	PROPORTION
0	14	0.467
1	10	0.333
2	4	0.133
3	2	0.067
> 3	0	0.0
TOTAL	30	1.0

So referring to the question of interest described above the *observed* proportion of 3 or more syntax errors is 0.067. This is an *estimate* of the 'true' probability. We envisage an infinite number of modules – this is the *population*. We have obtained a *sample* of size 30 from this population. We are really interested in the population probability of 3 or greater syntax errors.

We are interested in finding those situations in which we can find 'good' estimates.

Frequency diagram (bar chart)

A frequency diagram provides a graphical representation of the data. If we plot the proportions against number of syntax errors then we have an *estimate* of the probability distribution from which the data have been generated.

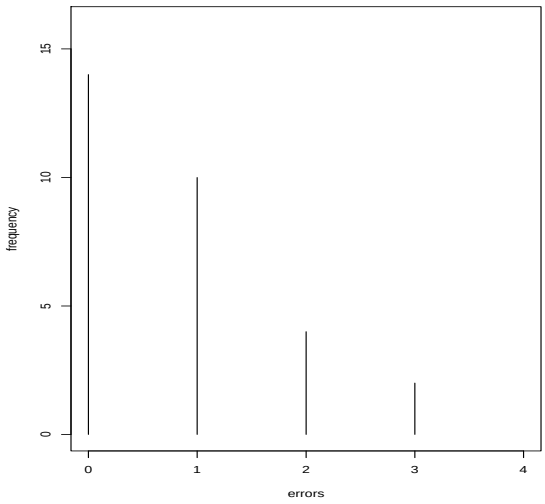


Figure 1: Frequency diagram for dataset 1

Dataset 2

Each value is a realization of what is known as a *continuous random variable*. One *assumption* we may make is that there is no trend with time. The values change from day to day in a way which we cannot predict – again we need to find a *model* for the downloading time. For example we assume that the downloading times are in general close to some value μ but there is some unpredictable variability about this value which produces the day-to-day values. The way such variability is described is via a *probability distribution*. For example we may assume that the day-to-day values can be modelled by a probability distribution which is *symmetric* about μ. We then may wish to estimate the value of μ.

Questions of interest here may include:

- What is our best guess (or estimate) of the average downloading time ?

- What sort of spread around the average value would we expect ?

We can summarize the data more informatively using a *histogram*.

A histogram is simply a grouping of the data into a number of boxes and is an *estimate* of the probability distribution from which the data arise.

This allows us to understand the variability more easily – for example are the data symmetric or skewed? Are there any outlying (or extreme) observations?

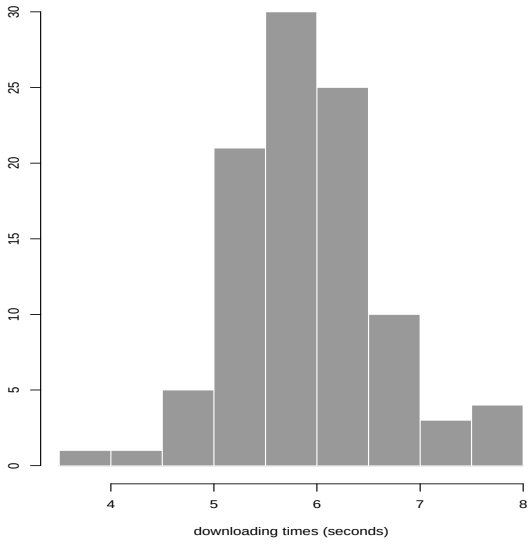


Figure 2: Histogram of Dataset 2

1.2 Summary statistics

Any quantity which is calculated from the data is known as a *statistic*. Statistics can be used to summarize a given dataset. Most simple statistics are measures of either the *location* or the *spread* of the data.

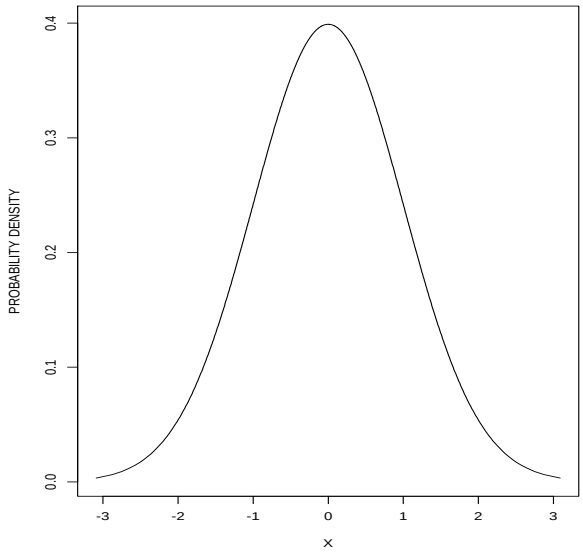


Figure 3: Normal distribution

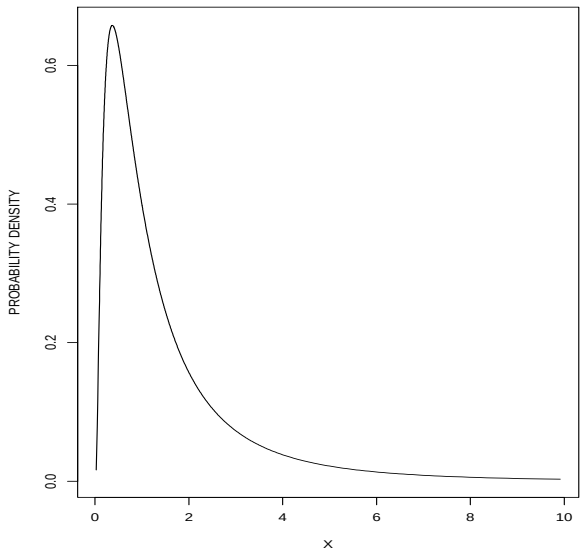


Figure 4: A right-skewed distribution

Suppose that n measurements have been taken on the random variable under consideration (eg number of syntax errors, downloading time).

Denote these measurements by $x_1, x_2, \ldots, x_n$. So x_1 is the first observation, x_2 is the second observation, etc.

Definition: the *sample mean* (or sample expected value) of the observations is given by:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

For dataset 1 we have

$$x_1 = 1, x_2 = 0, \cdots, x_{30} = 0 \text{ and } \overline{x} = \frac{24}{30} = 0.8 \tag{2}$$

For dataset 2 we have

$$x_1 = 6.36, x_2 = 6.73, \cdots, x_{100} = 5.18 \text{ and } \overline{x} = 5.95$$

Definition: the *sample median* of a sample of n measurements is the middle number when the measurements are arranged in ascending order.

If n is odd then the sample median is given by measurement $x_{(n+1)/2}$.

If n is even then the sample median is given by $(x_{n/2} + x_{(n/2)+1})/2$.

Examples

For dataset 1 we have the value 0 occurring 14 times and the value 1 occurring 10 times and so the median is the value 1. – for data of these type the sample median is not such a good summary of a 'typical' value of the data.

For dataset 2 the sample median is given by the sample average of the 50th and 51st observations. In this case both of these values are 5.88 and so the median is 5.88 also.

When the distribution from which the data has arisen is *heavily skewed* (that is, not symmetric) then the median may be thought of as 'a more typical value':

Example

Yearly earnings of maths graduates (in thousands of pounds):

$$24\ 19\ 18\ 23\ 22\ 19\ 16\ 55$$

For these data the sample mean is 24.5 thousand pounds whilst the sample median is 20.5 thousand pounds. Hence the latter is a more representative value.

Definition: the *sample mode* is that value of x which occurs with the greatest frequency.

Examples

For dataset 1 the sample mode is the value 0.

For the yearly earnings data the sample mode is 19.

All of these sample quantities we have defined are *estimates* of the corresponding population characteristics.

Note that for a symmetric probability distribution the mean, the median and the mode are the same.

As well as a measure of the location we might also want a measure of the spread of the distribution.

Definition: the *sample variance* s^2 is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

It is easier to use the formula:

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right]$$

The *sample standard deviation* s is given by the square root of the variance. The standard deviation is easier to interpret as it is on the same scale as the original measurements.

As a measure of spread it is most appropriate to think about the standard deviation for a symmetric distribution (such as the normal).

2 Probability

As we saw in Section 1 we are required to propose probabilistic models for data. In this chapter we will look at the theory of probability. Probabilities are defined upon events and so we first look at set theory and describe various operations that can be carried out on events.

2.1 Set theory

2.1.1 The sample space

The collection of all possible outcomes of an experiment is called the *sample space* of the experiment. We will denote the sample space by S.

Examples:

Dataset 1: $S = \{0, 1, 2, ...\}$

Dataset 2: $S = \{x : x \geq 0\}$

Toss of a coin: $S = \{H, T\}$

Roll of a six-sided die: $S = \{1, 2, 3, 4, 5, 6\}$.

2.1.2 Relations of set theory

The statement that a possible outcome of an experiment s is a member of S is denoted symbolically by the relation $s \in S$.

When an experiment has been performed and we say that some *event* has occurred then this is shorthand for saying that the outcome of the experiment satisfied certain conditions which specified that event. Any event can be regarded as a certain subset of possible outcomes in the sample space S.

Example: roll of a die. Denote by A the event that an even number is obtained. Then A is represented by the subset $A = \{2, 4, 6\}$.

It is said that an event A is contained in another event B if every outcome that belongs to the subset defining the event A also belongs to the subset B. We write $A \subset B$ and say that A is a subset of B. Equivalently, if $A \subset B$, we may say that B contains A and may write $B \supset A$.

Note that $A \subset S$ for any event A.

Example: roll of a die, suppose A is the event of an even number being obtained and C is the event that a number greater than 1 is obtained. Then $A = \{2, 4, 6\}$ and $C = \{2, 3, 4, 5, 6\}$ and $A \subset C$.

Example 2.1

2.1.3 The empty set

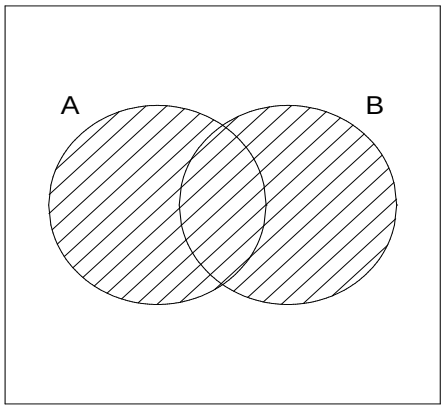
Some events are impossible. For example when a die is rolled it is impossible to obtain a negative number. Hence the event that a negative number will be obtained is defined by the subset of S that contains no outcomes. This subset of S is called the *empty set* and is denoted by the symbol ϕ.

Note that every set contains the empty set and so $\phi \subset A \subset S$.

2.1.4 Operations of set theory

Unions

If A and B are any 2 events then the *union* of A and B is defined to be the event containing all outcomes that belong to A alone, to B alone or to both A and B. We denote the union of A and B by $A \cup B$.



For any events A and B the union has the following properties:

$A \cup \phi = A$

$A \cup A = A$

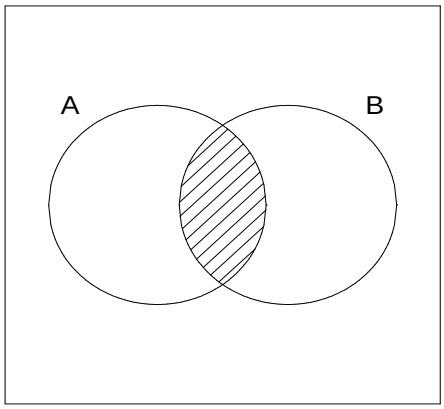
$A \cup S = S$

$A \cup B = B \cup A$

The union of n events $A_1, A_2, \cdots, A_n$ is defined to be the event that contains all outcomes which belong to at least one of the n events.
Notation: $A_1 \cup A_2 \cup \cdots \cup A_n$ or $\cup_{i=1}^{n} A_i$.

Intersections

If A and B are any 2 events then the *intersection* of A and B is defined to be the event containing all outcomes that belong both to A and to B. We denote the intersection of A and B by $A \cap B$.



For any events A and B the intersection has the following properties:

$A \cap \phi = \phi$

$A \cap A = A$

$A \cap S = A$

$A \cap B = B \cap A$

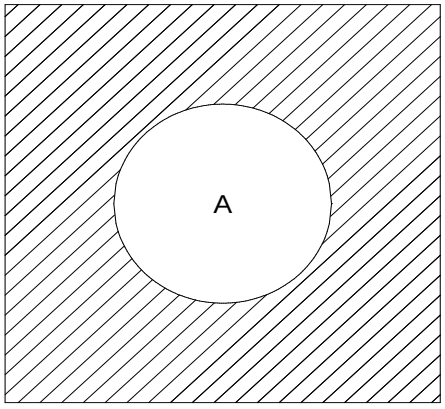
The intersection of n events $A_1, A_2, \cdots, A_n$ is defined to be the event that contains all outcomes which belong to all of the n events.

Notation: $A_1 \cap A_2 \cap \cdots \cap A_n$ or $\cap_{i=1}^{n} A_i$.

Example 2.2

2.1.5 Compliments

The complement of an event A is defined to be event that contains all outcomes in the sample space S which *do not* belong to A.
Notation: the compliment of A is written as A'.



The compliment has the following properties:

$(A')' = A$

$\phi' = S$

$A \cup A' = S$

$A \cap A' = \phi$

Examples:

Dataset 1: if $A = \{\geq 3\}$ errors then $A' = \{0, 1, 2\}$.

Dataset 2: if

$$A = \{x : 3 \leq x \leq 5\}$$

then

$$A' = \{x : 0 \leq x < 3, x > 5\}.$$

Example 2.3

2.1.6 Disjoint events

It is said that 2 events A and B are *disjoint* or *mutually exclusive* if A and B have no outcomes in common. It follows that A and B are disjoint if and only if $A \cap B = \phi$.

Example: roll of a die: suppose A is the event of an odd number and B is the event of an even number then $A = \{1, 3, 5\}, B = \{2, 4, 6\}$ and A and B are disjoint

Example: roll of a die: suppose A is the event of an odd number and C is the event of a number greater than 1 then $A = \{1, 3, 5\}, C = \{2, 3, 4, 5, 6\}$ and A and C are not disjoint since they have the outcomes 3 and 5 in common.

Example 2.4

2.1.7 Further results

a) DeMorgan's Laws: for any 2 events A and B we have

$$(A \cup B)' = A' \cap B'$$

$$(A \cap B)' = A' \cup B'$$

b) For any 3 events A, B and C we have:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Examples 2.5–2.7

2.2 The definition of probability

Probabilities can be defined in different ways:

Physical.

Relative frequency.

Subjective.

Probabilities, however assigned, must satisfy three specific axioms:

Axiom 1: for any event A, $\mathrm{P}(A) \geq 0$.

Axiom 2: $\mathrm{P}(S) = 1$.

Axiom 3: For any sequence of *disjoint* events $A_1, A_2, A_3, \cdots$

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

The mathematical definition of probability can now be given as follows:

Definition: A *probability distribution*, or simply a *probability* on a sample space S is a specification of numbers $\mathrm{P}(A)$ which satisfy Axioms 1–3.

Properties of probability:

(i) $\mathrm{P}(\phi) = 0$.

(ii) For any event A, $\mathrm{P}(A') = 1 - \mathrm{P}(A)$.

(iii) For any event A, $0 \leq \mathrm{P}(A) \leq 1$.

(iv) *The addition law of probability*

For any 2 events A and B:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).$$

So if A and B are *disjoint*

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).$$

For any 3 events A, B and C:

$$\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C)$$

$$-\mathrm{P}(A \cap B) - \mathrm{P}(A \cap C) - \mathrm{P}(B \cap C)$$

$$+\mathrm{P}(A \cap B \cap C).$$

Example 2.8

2.2.1 Conditional probability

We now look at the way in which the probability of an event A changes after it has been learned that some other event B has occurred. This is known as the *conditional probability* of the event A *given* that the event B has occurred. We write this as $\mathrm{P}(A \mid B)$.

Definition: if A and B are any 2 events with $\mathrm{P}(B) > 0$, then

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

Note that now we can derive the *multiplication law* of probability:

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

2.2.2 Independence

Two events A and B are said to be *independent* if

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

that is, knowledge of B does not change the probability of A.

So if A and B are independent we have

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

Examples 2.9–2.10

2.2.3 Bayes Theorem

Let S denote the sample space of some experiment and consider k events $A_1, \cdots, A_k$ in S such that $A_1, \cdots, A_k$ are *disjoint* and $\cup_{i=1}^{n} A_i = S$. Such a set of events is said to form a *partition* of S.

Consider any other event B, then

$$A_1 \cap B, A_2 \cap B, \cdots, A_k \cap B$$

form a partition of B. Note that $A_i \cap B$ may equal ϕ for some A_i.

So

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_k \cap B)$$

Since the k events on the right are disjoint we have

$$\mathrm{P}(B) = \textstyle\sum_{j=1}^{k} \mathrm{P}(A_j \cap B) \text{ (Addition Law of Probability)}$$

Now $\mathrm{P}(A_j \cap B) = \mathrm{P}(A_j)\mathrm{P}(B|A_j)$ so

$$\mathrm{P}(B) = \sum_{j=1}^{k} \mathrm{P}(A_j)\mathrm{P}(B|A_j)$$

Recall that

$$\mathrm{P}(A_i|B) = \mathrm{P}(B|A_i)\mathrm{P}(A_i)/\mathrm{P}(B)$$

using the above we can therefore obtain **Bayes Theorem**:

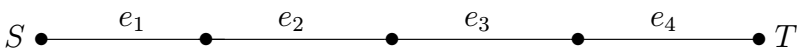
$$\mathrm{P}(A_i|B) = \mathrm{P}(B|A_i)\mathrm{P}(A_i)/\sum_{j=1}^{k} \mathrm{P}(A_j)\mathrm{P}(B|A_j)$$

Example 2.11

3 Systems reliability

Consider a system of components put together so that the whole system works only if certain combinations of the components work. We can represent these components as a NETWORK.

3.1 Series system



If any component fails $\Rightarrow$ system fails. Suppose we have n components each operating independently:

Let C_i be the event that component e_i *fails*, $i = 1, ..., n$.

$$\mathrm{P}(C_i) = \theta,\ i = 1, ..., n,\ \mathrm{P}(C_i') = 1 - \theta$$

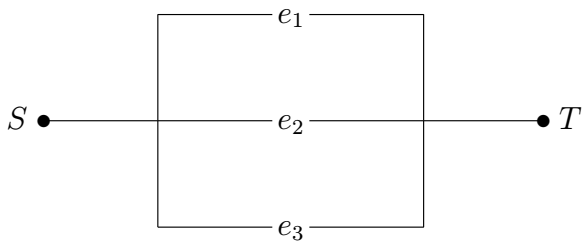
$$\mathrm{P}(\text{system functions}) = \mathrm{P}(C_1' \cap C_2' \cap \cdots \cap C_n')$$

$$= \mathrm{P}(C_1') \times \mathrm{P}(C_2') \times \cdots \times \mathrm{P}(C_n') \text{ since the events are independent}$$

$$= (1 - \theta)^n$$

Example: $n = 3$, $\theta = 0.1$, $\mathrm{P}(\text{system functions}) = 0.9^3 = 0.729$

3.2 Parallel system



Again suppose we have n components each operating independently:

Again let $\mathrm{P}(C_i) = \theta$.

The system fails if all n components fail and

$$\mathrm{P}(\text{system functions}) = 1 - \mathrm{P}(\text{system fails})$$

$$\mathrm{P}(\text{system fails}) = \mathrm{P}(C_1 \cap C_2 \cap \cdots \cap C_n)$$

$$= \mathrm{P}(C_1) \times \mathrm{P}(C_2) \times \cdots \times \mathrm{P}(C_n) \text{ since independent}$$

So

$$\mathrm{P}(\text{system functions}) = 1 - \theta^n$$

Example: $n = 3$, $\theta = 0.1$, $\mathrm{P}(\text{system functions}) = 1 - 0.1^3 = 0.999$.

3.3 Mixed system

– in lectures.

4 Random Variables and their distributions

4.1 Definition and notation

Recall:

Dataset 1: number of errors X, $S = \{0, 1, 2, ...\}$.

Dataset 2: time Y, $S = \{x : x \geq 0\}$.

Other example: Toss of a coin, outcome Z: $S = \{H, T\}$.

In the above X, Y and Z are examples of *random variables*.

Important note: capital letters will denote random variables, lower case letters will denote particular values (realizations).

When the outcomes can be listed we have a *discrete random variable*, otherwise we have a *continuous random variable*.

Let $p_i = \mathrm{P}(X = x_i),\ i = 1, 2, 3, ...$ Then any set of p_i's such that

i) $p_i \geq 0$, and

ii) $\sum_{i=1}^{\infty} p_i = \mathrm{P}(X \in S) = 1$

forms a *probability distribution* over $x_1, x_2, x_3, ...$

The *distribution function* $F(x)$ of a *discrete random variable* is given by

$$F(x_j) = \mathrm{P}(X \leq x_j) = \sum_{i=1}^{j} p_i = p_1 + p_2 + ... + p_j.$$

We now give some examples of distributions which can be used as models for discrete random variables.

4.2 The uniform distribution

Suppose that the value of a random variable X is *equally likely* to be any one of the k integers $1, 2, ..., k$.

Then the probability distribution of X is given by:

$$\mathrm{P}(X = x) = \begin{cases} \frac{1}{k} & \text{for } x = 1, 2, ..., k \\ 0 & \text{otherwise} \end{cases}$$

The *distribution function* is given by

$$F(j) = \mathrm{P}(X \leq j) = \sum_{i=1}^{j} p_i = p_1 + p_2 + ... + p_j = \frac{j}{k} \quad \text{for } j = 1, 2, ..., k.$$

Examples:

Outcome when we roll a fair die.

Outcome when a fair coin is tossed.

4.3 The Binomial Distribution

Suppose we have n *independent* trials with the outcome of each being either a *success* denoted 1 or a *failure* denoted 0. Let Y_i, $i = 1, ..., n$ be random variables representing the outcomes of each of the n trials.

Suppose also that the probability of success on each trial is constant and is given by p. That is,

$$\mathrm{P}(Y_i = 1) = p$$

and

$$\mathrm{P}(Y_i = 0) = 1 - p.$$

for $i = 1, ..., n$.

Let X denote the number of successes we obtain in the n trials, that is

$$X = \sum_{i=1}^{n} Y_i.$$

The sample space of X is given by $S = 0, 1, 2, ..., n$, that is, we can obtain between 0 and n successes. We now derive the probabilities of obtaining each of these outcomes.

Suppose we observe x successes and $n - x$ failures.

Then the probability distribution of X is given by:

$$\mathrm{P}(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, ..., n \\ 0 & \text{otherwise} \end{cases}$$

where

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

Derivation in lectures.

Examples 4.1–4.2

4.4 The Poisson Distribution

Many physical problems are concerned with events occurring independently of one another in time or space.

Examples

(i) Counts measured by a Geiger counter in 8-minute intervals

(ii) Traffic accidents on a particular road per day.

(iii) Number of imperfections along a fixed length of cable.

(iv) Telephone calls arriving at an exchange in 10 second periods.

A *Poisson process* is a simple model for such examples. Such a model assumes that in a unit time interval (or a unit length) the event of interest occurs randomly (so there is no clustering) at a rate $\lambda > 0$.

Let X denote the number of events occurring in time intervals of length t. Since we have a rate of λ in unit time, the rate in an interval of length t is λt.

The random variable X then has the Poisson distribution given by

$$\mathrm{P}(X = x) = \begin{cases} \frac{e^{-\lambda t}(\lambda t)^x}{x!} & \text{for } x = 0, 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

Note: the sample here space is $S = \{0, 1, 2, ...\}$.

We can simplify the above by putting $\mu = \lambda t$. We then obtain

$$\mathrm{P}(X = x) = \begin{cases} \frac{e^{-\mu}\mu^x}{x!} & \text{for } x = 0, 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

The *distribution function* is given by

$$F(j) = \mathrm{P}(X \leq j) = \sum_{i=0}^{j} \mathrm{P}(X = i) \quad \text{for } \ j = 0, 1, 2, ...$$

Example 4.3

4.5 Continuous Distributions

For a continuous random variable X we have a function f, called the *probability density function* (pdf). Every probability density function must satisfy:

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

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

For any interval A we have

$$\mathrm{P}(X \in A) = \int_A f(x)dx.$$

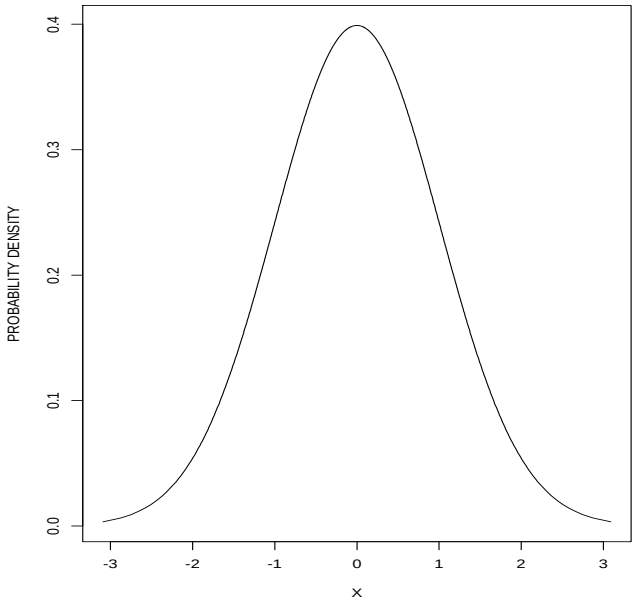


Figure 5: Probability density function

The *distribution function* is given by

$$F(x_0) = \mathrm{P}(X \leq x_0) = \int_{-\infty}^{x_0} f(x)dx.$$

Notes:

(i) $\mathrm{P}(X = x) = 0$ for a continuous random variable X.

(ii) $\frac{d}{dx}F(x) = f(x)$.

(iii)

$$\mathrm{P}(a \leq X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx$$

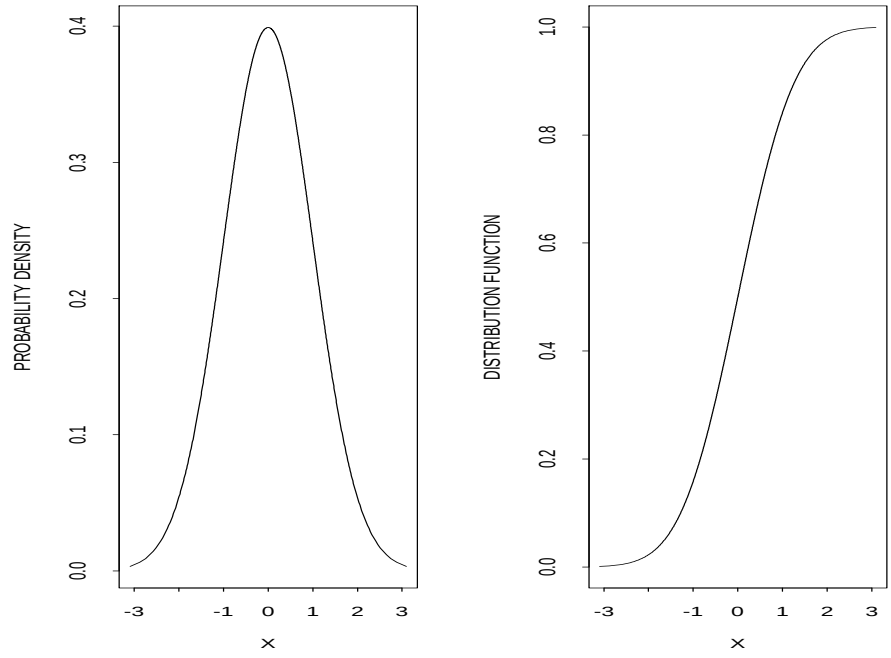


Figure 6: Probability density function and distribution function

4.6 The Uniform Distribution

Suppose we have a continuous random variable that is equally likely to occur in any interval of length c within the range $[a, b]$. Then the probability density function of this random variable is

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

Notes: show that this is a pdf and derive distribution function in lectures.

4.7 The exponential distribution

Suppose we have events occurring in a Poisson process of rate λ but we are interested in the random variable describing the time between events T. It can be shown that this random variable has an *exponential distribution* with parameter λ. The probability density function of an exponential random variable is given by:

$$f(t) = \begin{cases} \lambda \exp(-\lambda t) & \text{for } t > 0. \\ 0 & \text{otherwise, that is for } t \leq 0 \end{cases}$$

This probability density function is shown with various values of λ in Figure 7.

Notes: show that this is a probability density function and derive in lectures the distribution function, which is given by

$$F(t_0) = 1 - \exp(-\lambda t_0) \quad \text{for } t_0 > 0.$$

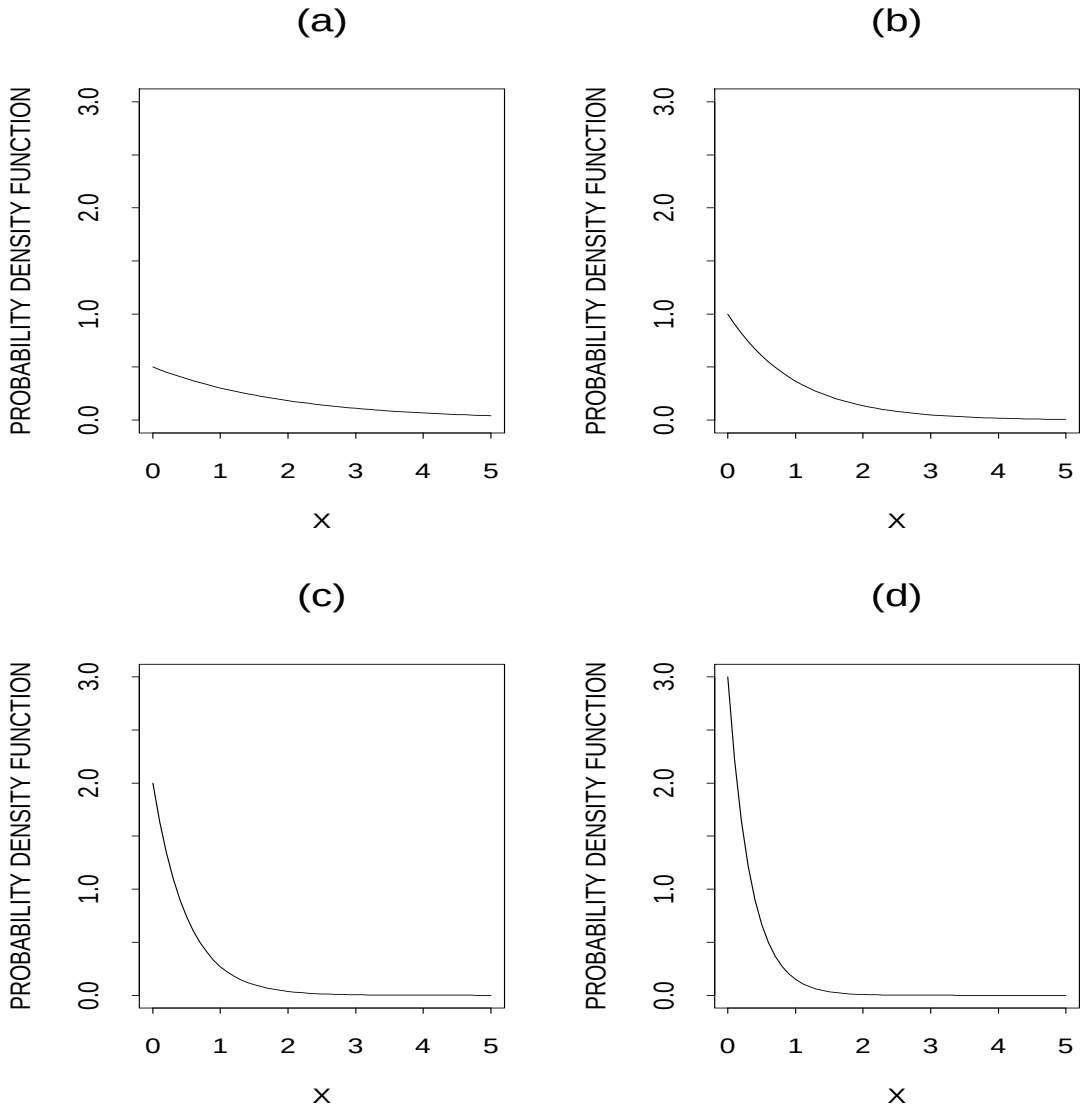


Figure 7: Exponential probability density functions with (a) $\lambda = 0.5$, (b) $\lambda = 1.0$, (c) $\lambda = 2.0$, (d) $\lambda = 3.0$

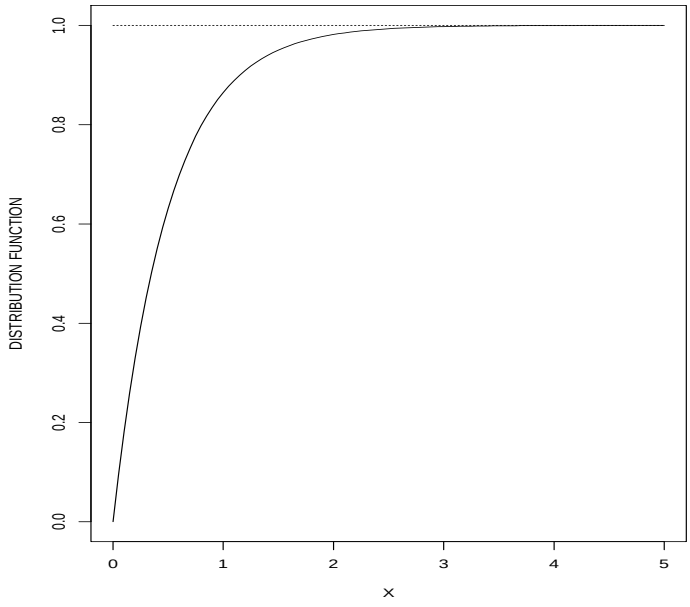


Figure 8: Exponential distribution function for $\lambda = 2.0$

As well as arising naturally as the time between events in a Poisson process the exponential distribution is also used to model the lifetimes of components which do not fail – the reason for this is the *lack of memory property* of the exponential distribution which will be discussed in lectures. This property states that:

$$\mathrm{P}(T > s + t | T > s) = \mathrm{P}(T > t).$$

4.8 The normal distribution

The normal distribution is of great importance for the following reasons:

1. It is often suitable as a probability model for measurements of weight, length, strength, etc.
2. Non-normal data can often be transformed to normality.
3. The central limit theorem states that when we take a sample of size n from any distribution with a mean μ and a variance σ^2, the sample mean $\bar{X}$ will have a distribution which gets closer and closer to normality as n increases.
4. Can be used as an approximation to the binomial or the Poisson distributions when we have large n or λ respectively (though this is less useful now that computers can be used to evaluate binomial/Poisson probabilities).
5. Many standard statistical techniques are based on the normal distribution.

If the continuous random variable X has probability density function

$$\phi(x) = f(x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-x^2/2\right) \quad -\infty < x < \infty$$

then we say that X has the (standard) normal distribution with mean 0 and variance 1.

We write $X \sim N(0,1)$.

The standard normal distribution is symmetric about 0 and is shown in Figure 1.

The distribution function is given by:

$$F(x_0) = \mathrm{P}(X \leq x_0) = \Phi(x_0) = \int_{-\infty}^{x_0} \frac{1}{(2\pi)^{1/2}} \exp\left(-x^2/2\right) dx.$$

This integral cannot be evaluated analytically hence its values are calculated numerically and are widely-tabulated. The density and distribution function of the standard normal distribution is shown in Figure 6.

Example 4.4: how to use standard normal tables.

4.9 Mean and variance

Recall in Section 1 we informally discussed measures of location and spread for sets of data. Here we define the mean and the variance as summaries of a distribution. Note that for some distributions the mean is not a good measure of the location of the distribution and the variance is not a good measure of the spread of a distribution. For normal data the mean and the standard deviation (which is the square root of the variance) are good summaries of location and spread.

For a discrete random variable X the expected value (or mean) is defined as

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

For a continuous random variable with probability density function $f(x)$ the expected value (or mean) is defined as

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

Example 4.5:

Properties:

(i) For constants a and b we have

$$E(aX + b) = aE(X) + b.$$

(ii) For random variables X and Y we have

$$E(X + Y) = E(X) + E(Y).$$

(iii) For constants $a_1, ..., a_k$ and b and random variables $X_1, ..., X_k$ we have

$$E(a_1X_1 + ... + a_kX_k + b) = a_1E(X_1) + ... + a_kE(X_k) + b.$$

The *variance* of a random variable X is defined as

$$\text{var}(X) = E[(X - \mu)^2]$$

where

$$\mu = E(X).$$

So the variance is the mean of the squared distances from the mean.

The *standard deviation* is defined as

$$\text{sd}(X) = \text{var}(X)^{1/2}.$$

Note that the standard deviation is on the same scale as the data.

Properties:

(i) For constants a and b we have

$$\text{var}(aX + b) = a^2\text{var}(X).$$

(ii) For *independent* random variables X and Y we have

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

(iii) For constants $a_1, ..., a_k$ and b and *independent* random variables $X_1, ..., X_k$ we have

$$\text{var}(a_1X_1 + ... + a_kX_k + b) = a_1^2\text{var}(X_1) + +... + a_k^2\text{var}(X_k).$$

Examples:

Binomial distribution: $E(X) = np$, $\text{var}(X) = np(1 - p)$.

Poisson distribution: $E(X) = \mu$, $\text{var}(X) = \mu$.

Uniform distribution: $E(X) = (b + a)/2$, $\text{var}(X) = (b - a)^2/12$.

Exponential distribution: $E(X) = \lambda^{-1}$, $\text{var}(X) = \lambda^{-2}$.

Proofs: in class.

4.10 The normal distribution revisited

If the continuous random variable X has probability density function

$$f(x) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty$$

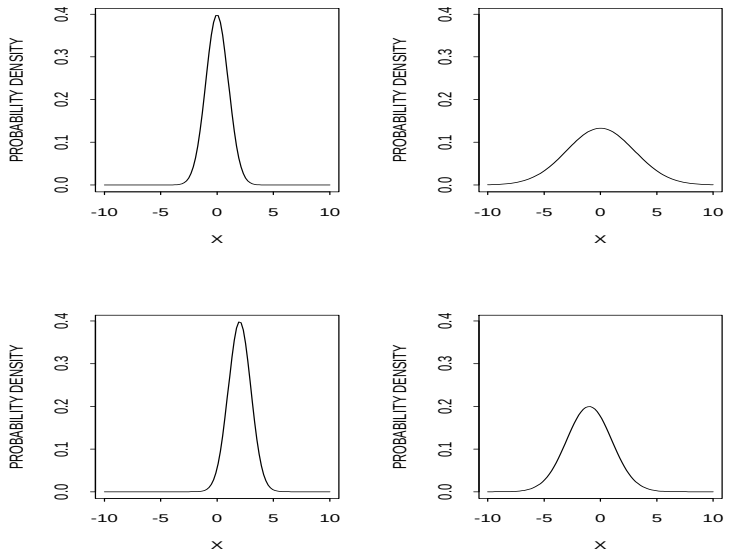


Figure 9: Normal distributions with: (a) $\mu = 0, \sigma^2 = 1$, (b) $\mu = 0, \sigma^2 = 9$, (c) $\mu = 2, \sigma^2 = 1$,(d) $\mu = -1, \sigma^2 = 4$

then X has a normal distribution with mean μ and standard deviation σ. We write $X \sim N(\mu, \sigma^2)$. Figure 4 shows the effect of changing the values of μ and σ.

Important result: if $X \sim N(\mu, \sigma^2)$ then the random variable Z defined by

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

is distributed as $Z \sim N(0, 1)$. This is useful for looking up in tables the values of the distribution function for arbitrary normal distributions, that is, we don't need a separate set of tables for each value of μ and σ.

Example 4.6: how to use the normal tables for arbitrary normal distributions.

Further important results:

1. If X_1 and X_2 are independent random variables with distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, then the random variable defined by

$$Z = X_1 + X_2$$

is also normal with mean

$$E(Z) = E(X_1) + E(X_2) = \mu_1 + \mu_2$$

and variance

$$\text{var}(Z) = \text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) = \sigma_1^2 + \sigma_2^2.$$

2. If X_1 and X_2 are independent random variables with distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, then the random variable defined by

$$Z = X_1 - X_2$$

is also normal with mean

$$E(Z) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

and variance

$$\text{var}(Z) = \text{var}(X_1 - X_2) = \text{var}(X_1) + \text{var}(X_2) = \sigma_1^2 + \sigma_2^2.$$

3. **General result** If $X_1, ..., X_n$ are independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$ and $a_1, ..., a_n$ are constants then the random variable defined by

$$Z = \sum_{i=1}^{n} a_i X_i$$

has a normal distribution with mean

$$E(Z) = E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mu_i$$

and variance

$$\text{var}(Z) = \text{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

Very important result

Definition: a *random sample* of size n from a distribution $f(x)$ is a set of independent and identically distributed random variables $X_1, ..., X_n$ each with the distribution $f(x)$.

If $X_1, ..., X_n$ are a *random sample* from the normal distribution $N(\mu, \sigma^2)$ then the random variable

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

has a normal distribution with mean

$$E(\bar{X}) = \mu$$

and variance

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Note here that $\bar{X}$ is a random variable. Suppose we carry out an experiment where we draw random samples of size n from a normal distribution – if we do this repeatedly and evaluate the mean then each time we will get a different answer. If we were to repeat the experiment an infinite number of times then the distribution of the $\bar{X}$'s that we obtain is $N(\mu, \sigma^2/n)$.

5 Statistical Analysis

5.1 Introduction

Probability theory provides us with models for random experiments. Such experiments provide realizations x of random variables X. The observation $X = x$ is only one of many possible values of X. We wish to learn about the probability distribution or probability density function which produced x – this will us allow us to make inference about the process under study.
For example, we will be able to predict future outcomes/summarize the underlying system.

The notation we shall now use to represent our probability model is $p(x|\theta)$ where θ represents the parameter of the model – this notation makes explicit that to evaluate the probability of seeing a particular observation we need to know the value of θ.

In Section 4 the probability models which we looked at contained *parameters*, for example
(i) a binomial random variable with probability of success p so $\theta = p$,
(ii) a Poisson random variable with rate λ so $\theta = \lambda$,
(iii) a normal random variable with mean μ and variance σ^2 so $\theta = (\mu, \sigma^2)$.

In Section 3 these parameters were assumed known and we could predict the outcomes we were likely to see, for example we could evaluate probabilities such as $\mathrm{P}(X > 3)$.
In statistical applications we are in the opposite situation – we obtain realizations (that is, observations) from some unknown probability model and we wish to guess the values of the parameters of the model.

After observing outcomes x_1, x_2, ...,x_n we may be interested in:

Point estimation: guessing the value of θ.

Interval estimation: giving an interval in which we are 'confident' that θ lies.

5.2 Point estimation

Suppose we obtain a random sample of size n from some probability model $f(x|\theta)$. That is, we obtain measurements $\underline{x}= (x_1, ..., x_n)$ on n independent and identically distributed random variables $\underline{X} = (X_1, ..., X_n)$ each with the same distribution $f(x|\theta)$.

In this case we have the joint distribution

$$f(\underline{x}|\theta) = f(x_1, ..., x_n|\theta) = f(x_1|\theta)...f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Viewing the joint distribution as a function of θ gives what is known as the *likelihood function.*

Example: dataset 2, downloading times.

We could assume that x_1, ...,x_{100} are a random sample from $f(x_i|\mu, \sigma^2)$ where $f(.|.)$ is a normal distribution, that is:

$$X_i \sim N(\mu, \sigma^2).$$

Definition: any real-valued function $T = g(x_1, ..., x_n)$ of the observations in the random sample is called a *statistic.*

T is known as an *estimator* if we use it as a guess of the value of some unknown parameter.

Example 5.1: A random sample of size 10 of response times were obtained (in seconds) for an online holiday booking query:

5.3, 4.9, 6.2, 5.7, 4.8

5.4, 6.1, 6.3, 5.6, 5.5.

Suppose it were assumed that these measurements formed a random sample of size 10 from a normal distribution with mean μ and known variance $\sigma^2 = 0.5^2$. How could we guess the value of μ ?

Various possibilities are open to us: we could try the mean, the median, the average of the smallest and the largest ,...

How do we judge an estimator ?

Important idea: our estimator $T = g(\underline{X})$ is a function of the random variables X_1, ...,X_n and so the estimator itself has a distribution which is known as the *sampling distribution.*

We would like the distribution of our estimator to be 'close to the unknown parameter θ' with high probability.

If we have a distribution for our estimator then how could we summarize this distribution?

(i) **The average value**: the estimator $T = g(\underline{X})$ is said to be *unbiased* if

$$E[T] = \theta$$

– on average we obtain the correct answer. If we repeated the experiment an infinite number of times then our average answer would be the correct answer.

(ii) **The variance**: if we have two estimators that are unbiased then we would rather use the one which has the smaller variance.

Recall the result from Section 4 that if we have a random sample of size n from the normal distribution $N(\mu, \sigma^2)$ then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

has a normal distribution with $E[\bar{X}] = \mu$ and $var(\bar{X}) = \frac{\sigma^2}{n}$.

So the sample mean is an unbiased estimator of μ.

Example 5.1 revisited

Result: if we have a random sample of size n from the normal distribution $N(\mu, \sigma^2)$ with μ and σ^2 both unknown then the estimator

$$T = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of σ^2.

5.3 Maximum Likelihood Estimation

One way of estimating the parameter θ, is by choosing the value of θ that maximises the likelihood:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

In other words, we choose θ to give the highest probability to the observed data. This value of θ is known as the Maximum Likelihood Estimator (MLE).

Example 5.2:

5.4 Interval estimation

We have seen that associated with any point estimator of an unknown parameter θ there is a measure of the uncertainty of that estimator.

Example: Consider a random sample of size n $X_1, ..., X_n$ from $N(\mu, \sigma^2)$ with σ known but μ unknown. The estimator $\hat{\mu} = \bar{X}$ has

$$\mathrm{E}[\bar{X}] = \mu$$

and

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

An alternative method of describing the uncertainty associated with an estimate is to give an interval within which we think that θ lies. One way of doing this is via a **confidence interval**.

In lectures derive the confidence interval for the mean μ of a normal distribution when the variance σ^2 is known. The $100(1-\alpha)\%$ confidence interval is given by:

$$\left(\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is that value such that

$$1 - \alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2})$$

where $Z \sim N(0,1)$. So for example for $\alpha = 0.10$ we have $z = 1.6449$, for $\alpha = 0.05$ we have $z = 1.9600$ and for $\alpha = 0.01$ we have $z = 2.5758$. These 3 cases are appropriate for confidence intervals of, respectively, 90 % , 95 % and 99 %.

Example 5.3:

Suppose now that $X_1, ..., X_n$ are a random sample from the normal distribution $N(\mu, \sigma^2)$ where now **both** μ and σ^2 are unknown.

In the case where σ^2 was known we used the fact that

$$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

has the standard normal distribution $N(0,1)$. We cannot do this here since σ is unknown.

We can, however, replace σ by the estimator

$$s = \left[\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)\right]^{1/2}$$

consider the new random variable

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

This random variable does not have a normal distribution but what is known as the **(student's) t-distribution** with $n-1$ degrees of freedom. The t-distribution is widely tabulated for different degrees of freedom. Like the normal distribution it is symmetrical about zero but has 'fatter tails', that is, there is more probability in the tails. As the degrees of freedom n tends to infinity the t-distribution tends to the normal distribution. The intuition here is that as we obtain more and more data our estimator for σ, namely s, becomes better and better in the sense that with high probability we will be close to the true value. Hence it is 'as if' we knew σ and can use the normal distribution. Figure 1 shows a selection of t-distributions.

The confidence interval for μ is given by:

$$\left(\bar{x} - \frac{t_{\alpha/2}s}{\sqrt{n}}, \bar{x} + \frac{t_{\alpha/2}s}{\sqrt{n}}\right)$$

where $t_{\alpha/2}$ is that value such that

$$1 - \alpha = P(-t_{\alpha/2} < T < t_{\alpha/2})$$

where T has a t-distribution with $n-1$ degrees of freedom.

Note: this interval is wider than when the variance σ^2 is known, reflecting the extra uncertainty in the estimate of σ^2.

Example 5.4:

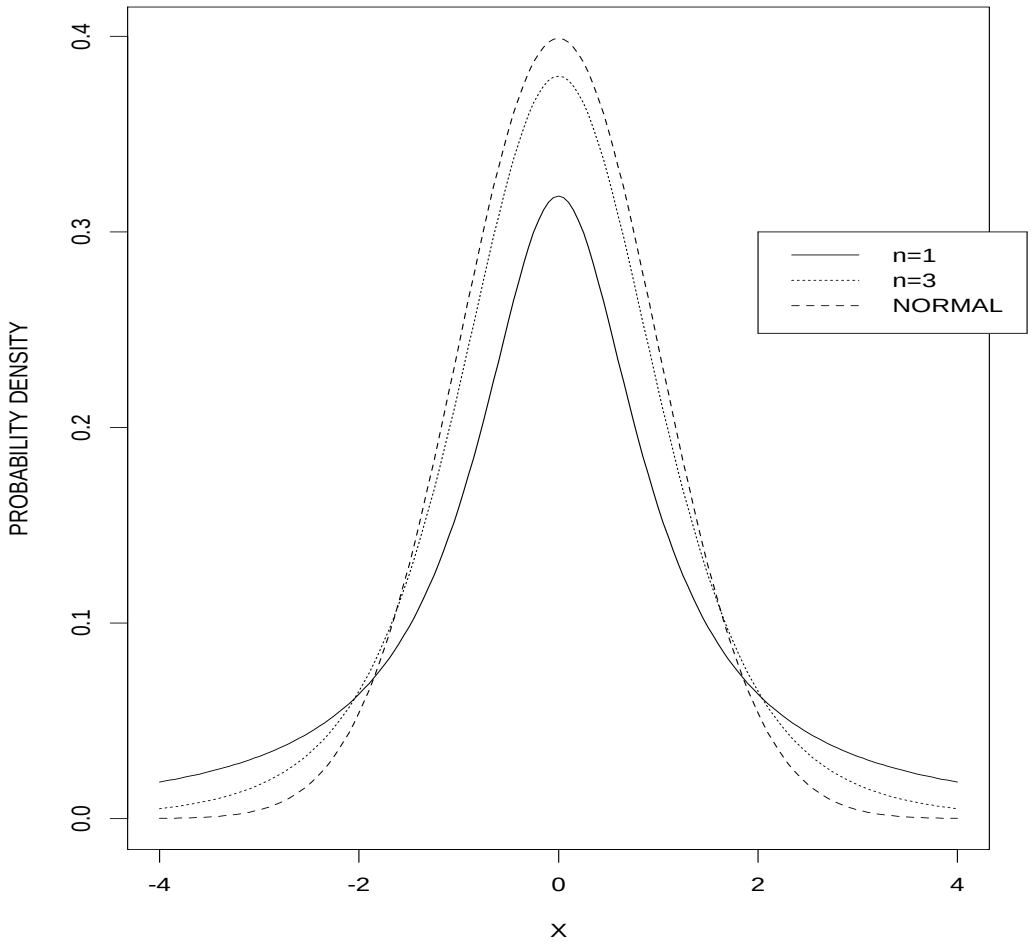


Figure 10: A selection of t distributions

6 Product Reliability

Let

$$T = \begin{cases} \text{failure – time} \\ \text{life – time} \end{cases}$$

of a product e.g. an automobile, stereo, computing equipment.

The random variable, T is non-negative, by its nature and has distribution function $F(t)$, the failure-time distribution, and probability density function $f(t)$, the failure-time density.

The probability that the product will fail before time t_0:

$$\mathrm{P}(T \leq t_0) = F(t_0) = \int_0^{t_0} f(t)\,\mathrm{d}t.$$

The probability that the product will survive until time t_0:

$$\mathrm{P}(T > t_0) = 1 - F(t_0) = R(t_0), \text{ say.}$$

$R(t_0)$ is the *probability of reliability.*

We note that,

$f(t)\mathrm{d}t \approx \mathrm{P}(t < T < t + \mathrm{d}t) =$ probability that a product will fail in $(t, t + \mathrm{d}t)$.

6.1 Alternative ways to describe life characteristics of a product

Here we use a measure of the probability of failure as the product gets older – the probability that the product will fail in $(t, t + \mathrm{d}t)$ given that it has survived to time t.

Let A be event 'Product fails in $(t, t + \mathrm{d}t)$'

B be event 'Product survives until t'

Probability of failure in $(t, t + \mathrm{d}t)$, given survival until t, is

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A)}{\mathrm{P}(B)} \approx \frac{f(t)\,\mathrm{d}t}{1 - F(t)} = \frac{f(t)\,\mathrm{d}t}{R(t)}.$$

The quantity

$$z(t) = \frac{f(t)}{R(t)} \propto \mathrm{P}(A|B)$$

and is called the hazard rate of the product.

Knowledge about a product's hazard rate often helps us to select the appropriate failure time density function for the product.

6.2 Examples of Hazards

Figure 11 shows some examples of hazard functions.

6.3 Examples of commonly used failure-time densities

6.3.1 Exponential Distribution

A positive random variable T has an exponential distribution with parameter λ ($\lambda > 0$) if its pdf is

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

The exponential distribution is used as the model for the failure time distribution of systems under 'chance' failure conditions, e.g.

1. Lifetime of an electron tube.

2. Time intervals between successive breakdowns of an electronics system.

Figure 12 shows the reliability and hazard function for an exponential distribution with $\lambda = 1$.

Example 6.1:

Show

$$R(t) = e^{-\lambda t} \quad t > 0$$
$$z(t) = \lambda \qquad t > 0 \textit{ constant hazard rate.}$$

6.3.2 Two parameter Weibull Distribution

A positive random variable T has a Weibull distribution with parameters λ, β $(\lambda, \beta > 0)$ if its pdf is

$$f(t) = \begin{cases} \lambda\beta(\lambda t)^{\beta-1} e^{-(\lambda t)^{\beta}} & t \geq 0 \\ 0 & t < 0. \end{cases}$$

β is a *shape* parameter, λ is a *scale* parameter. Figure 13 shows the reliability and hazard functions for a Weibull distribution with $\lambda = 1$, and $\beta = 0.5$ and 3.

Example 6.2:

Show

$$R(t) = e^{-(\lambda t)^{\beta}} \qquad t > 0.$$
$$z(t) = \lambda\beta(\lambda t)^{\beta-1} \quad t > 0 \textit{ not memoryless.}$$

Example 6.3:

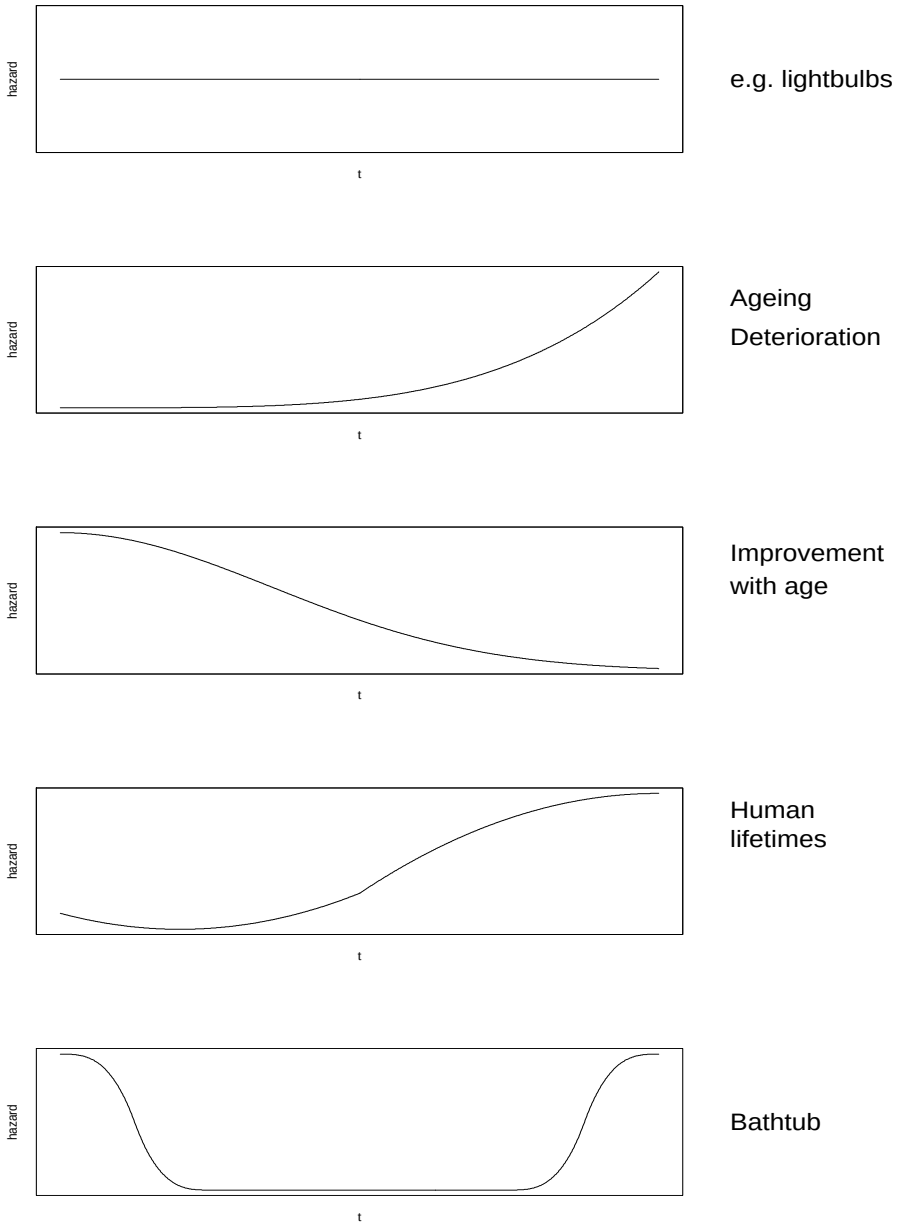


Figure 11: Examples of Hazards

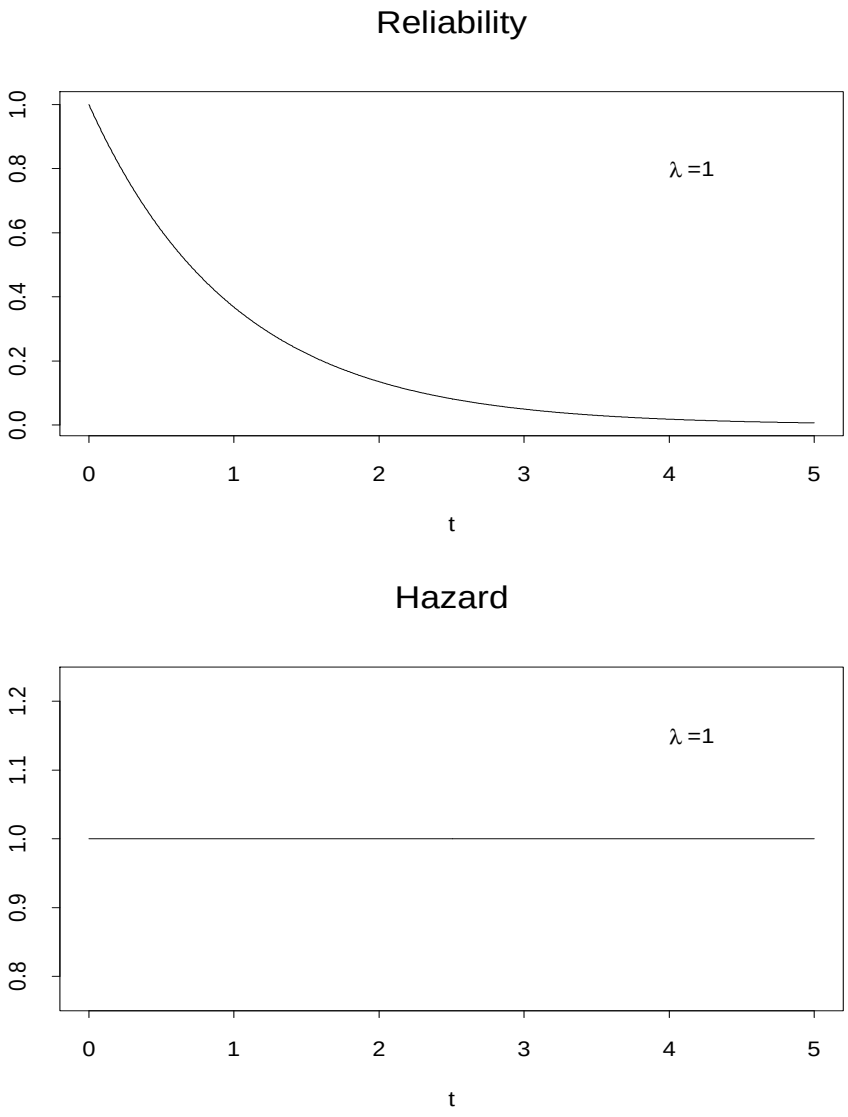


Figure 12: Exponential Distribution

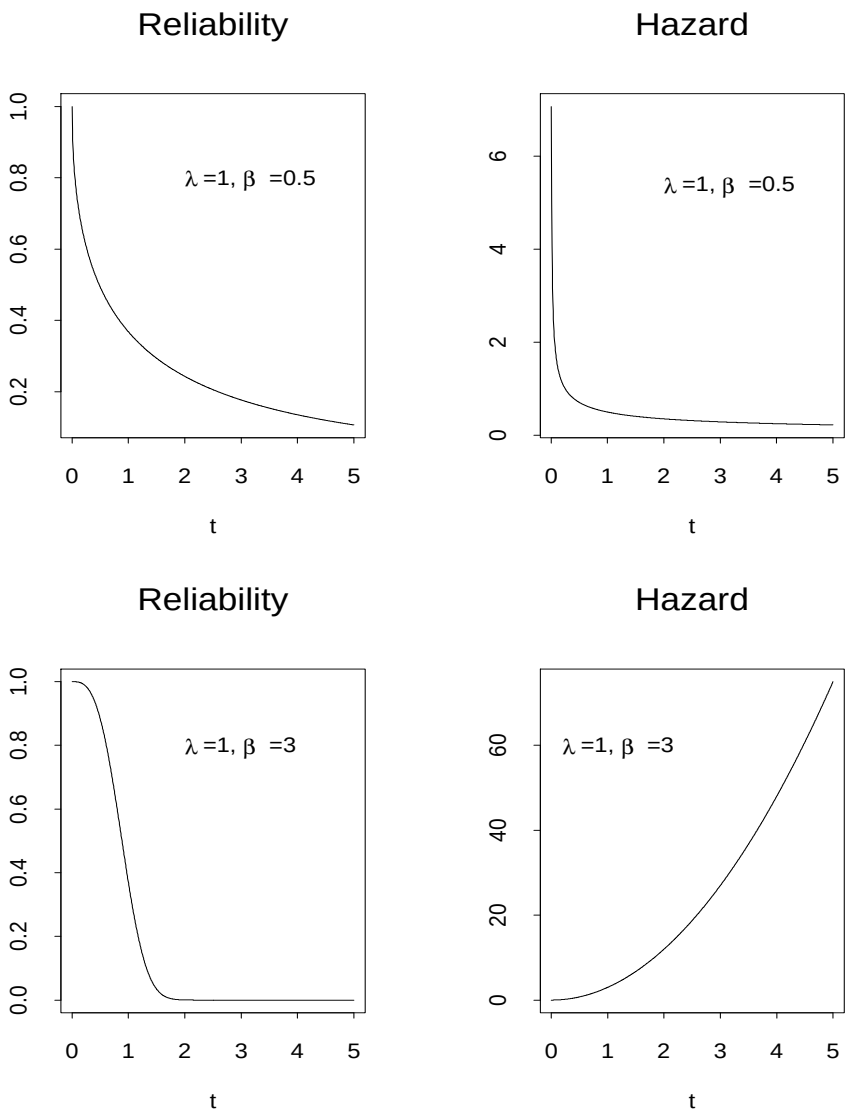


Figure 13: Weibull Distribution