



## CHAPTER 2: PROBABILITY AND PROBABILITY DISTRIBUTIONS

### *MOTIVATION*

This chapter explores the underpinnings of inferential statistics. The practice of *statistical inference* – generalizing conclusions based on limited data – depends on probabilities.

At first it may seem curious that decisions concerning, for example, treatment options for patients, or public policy on disease screening, or nominations for a characteristic or behaviour to be labelled as a “risk factor” for cancer (with consequent further decisions regarding health policy), will to some extent be based on the rules of chance. But remember that we are dealing not with a predictable, constant world, but the real world where variation is endemic. No individual is “typical” of the population, so we must observe and quantify characteristics of groups – which we hope will reflect the underlying population that is our real interest. We need strategies (a) to select the groups – these are our *samples*, (b) to construct *models* of the underlying population, and (c) to relate the observed data of the sample back to the models.

Probability theory provides models for the population against which we can compare whatever information is at hand in our sample.

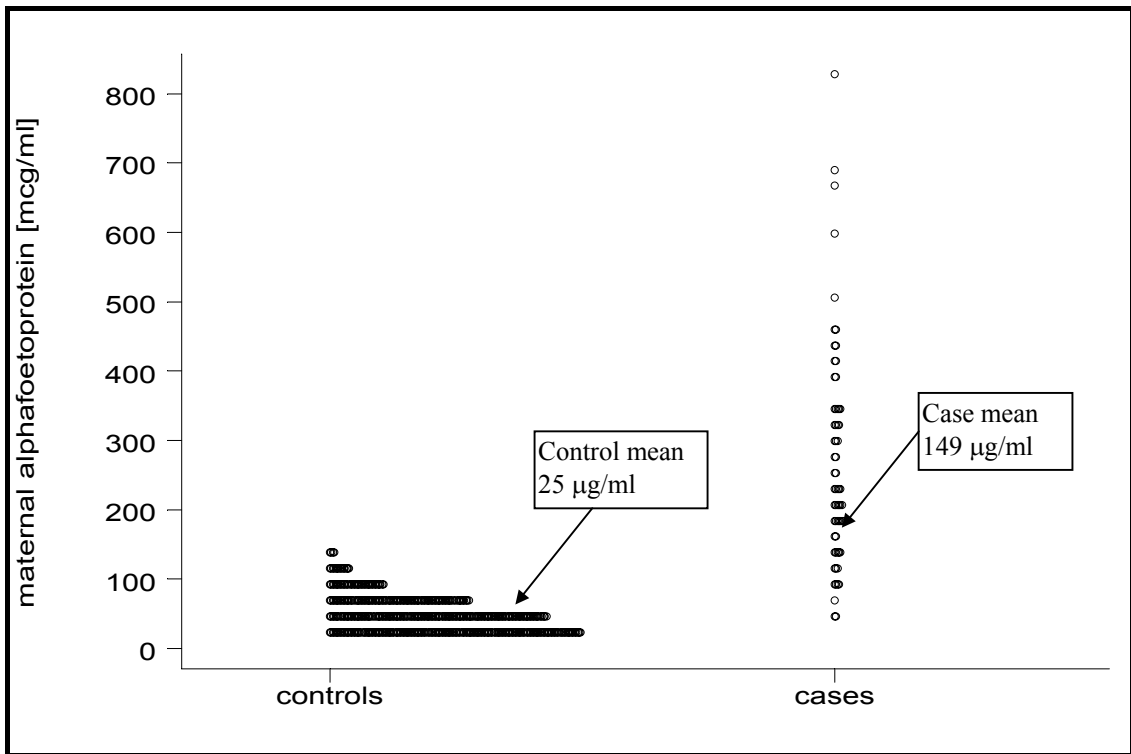
Look at it this way: the data of our sample, if collected validly, provide a small window onto the population of interest. On the other hand, the probability model is like an artist’s rendition of what the entire population is expected to look like (assuming that the artist has no surreal tendencies). We then try to see if the view through the window resembles the painting. If the painting (our model) accords with the evidence of our eyes (our sample data) then the painting is held to be a true reflection of the nature of the population, at least until something better is painted. If the painting does not resemble the view through the window then the original painting is discarded and another may be commissioned. That is, a new model for the characteristics of the population will need to be proposed.

Does your group of patients with cancer survive longer than anybody else’s group? Do infants from socio-economically deprived households fail to gain weight compared with all children? Is the incidence of sport or work related injury related to alcohol consumption? If we have sample data and probability models describing cancer survival or weight gain or the relationship of alcohol and injury in the population, then these problems will have solutions. Just as important, probability models enable us to quantify how *incorrect* our solution may be.

## § 2.1 INTRODUCTION

Researchers do not just describe data; often they seek to use the knowledge gained from a relatively small sample to make statements of general applicability to the whole population. This is the practice of *inferential statistics*. Before tackling this subject we need to understand something of Probability Theory and Probability Distributions which provide the justification for our inferential processes.

Consider the following situation. A researcher collects blood from a representative sample of mothers, 464 of whom have given birth to a normal infant, and 64 of whom have given birth to an infant with a neural tube defect (NTD). Interest centres on the maternal level of alpha-foetoprotein ( $\alpha$ -fp), a known marker for certain congenital malformations. The mean  $\alpha$ -fp level in the 64 mothers with affected children (“cases”) is 159  $\mu\text{g/ml}$ , and among the 464 controls is 25  $\mu\text{g/ml}$ . Fig 2.1 shows a “dotplot” representation of the distribution of  $\alpha$ -fp levels by case-control status.



**Fig 2.1** Distribution of maternal  $\alpha$ -foetoprotein levels

Can the researcher *draw the inference* from her samples that, in the general population, mothers of babies with a NTD have a higher mean blood  $\alpha$ -fp than mothers of non-affected babies? After all, the means of the two samples are different. On the other hand, you could say that, with the biological variability of alpha-foetoprotein levels and the (relatively) small number of people tested, some apparent difference in means was virtually inevitable. The questions really are: is the apparent difference found in these samples reliable? Will the magnitude and direction of the difference remain pretty constant from sample to sample? Or is the difference found

on this occasion the result of non-systematic (chance) factors which will vary from study to study so that no consistent difference will be found? If the latter is the case, the researcher would be on shaky ground in claiming a difference in mean  $\alpha$ -fp levels between mothers of NTD and normal babies in the population.

***Probability theory provides the logical basis for deciding between alternative interpretations of research findings.***

Probability theory is a vast and complicated branch of mathematics, but luckily for us, just a few basic concepts are needed to appreciate the practical aspects of inferential statistics, that is, ***parameter estimation*** and ***hypothesis testing*** (see Chapter 3).

Along the way, we'll look at a practical use of some basic probability laws – ***Bayes' Theorem***.

## § 2.2      **PROBABILITY**

### § 2.2.1    **Definitions**

We will define the ***probability of an event*** as the *long-run relative frequency of the event in repeated trials under similar conditions*. Every word of this “frequentist” definition has been chosen carefully, so read it again.

#### **Example 2.1**

We know that the probability of getting a head on a coin toss is  $\frac{1}{2}$  or 0.5.

The notation is:  $P(\text{head})$  or  $Pr(\text{head}) = 0.5$

We say this even though we know the actual result for the single toss will be either a head or a tail – what we are really saying is that if enough coins are tossed (that is, in the *long run*), the relative frequency of heads will approach 1 in 2.

Next, in a statistical world – a world which is characterised by *variability* – an experiment or ***trial*** can lead to different results or ***events***. The collection of every possible event is called the ***sample space*** for the experiment in question.

If two events cannot possibly exist simultaneously then they are termed ***mutually exclusive***. For example, consider the possible results of a life-saving operation, say, an emergency tracheotomy on a person with acute laryngeal obstruction. The resultant event of the single “trial” would either be success (patient lives) or failure (patient dies).

Together, these two events make up the sample space. The two events cannot co-exist.

Two or more events are said to be *independent* if the occurrence or non-occurrence of one event is unaffected by the occurrence or non-occurrence of the other event(s). For example, in 100 successive tosses of a “fair” coin, even if the first 99 trials lead to a heads result (an unlikely but still possible state of affairs), the 100<sup>th</sup> toss will still lead to a heads result, with probability 0.5, or a tails result with probability 0.5 because each trial’s result is independent of all the others.

On the other hand, consider the example of the allocation of successive patients to wards in a hospital on a busy day. It may be that if one patient was sent to a North Wing ward, the next might be admitted to an East Wing ward in an effort to balance the workload. The second event is *dependent* or *conditional* on the first.

### § 2.2.2 Some Properties and Rules of Probability

The great French mathematician, Laplace, maintained that probability was just “common sense with numbers”. The following rules bear this out.

The probability of an event, denoted as  $P(E)$ , cannot be less than zero or greater than one. This follows directly from the definition of probability as a *relative* frequency.

$$0 \leq P(E) \leq 1 \quad \text{E2.1}$$

Next, it is clear that the sum of the probabilities of all the *mutually exclusive* events in the sample space must equal one. For example, if the only possible human blood types were A, B, AB and O, the probabilities of these four types when summed would equal one. For  $n$  possible results of a trial we have:

$$P(E_1) + P(E_2) + \dots + P(E_{n-1}) + P(E_n) = 1 \quad \text{E2.2}$$

The *complement* of a specified event is all the *other* mutually exclusive events in the sample space. The complement of an event  $E$  is denoted by  $\bar{E}$ . The unknown probability of an event can be easily calculated if the probability of the complement is known:

$$P(\bar{E}) = P(\text{not } E) = 1 - P(E) \quad \text{E2.3}$$

#### **Example 2.2**

A community nurse knows that a type of infant skin rash can only mean one of three conditions: A, B or C. That is, A, B and C “exhaust” the sample space of possible diagnoses. She knows (from the literature) that the probability of the rash being caused by condition A is 0.1 and by condition B, 0.65. So, it is immediately apparent that there is a 0.25 chance of C being the underlying condition.

If events are mutually exclusive, then the probability of *either* event occurring is equal to the sum of the probabilities of each event. So if we have two mutually exclusive events, A and B, then:

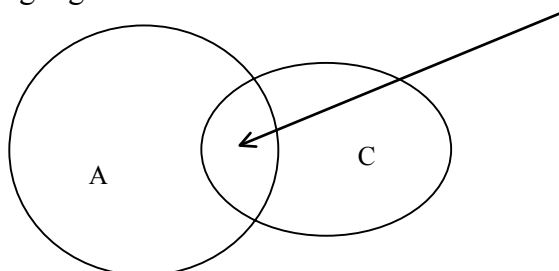
$$P(A \text{ or } B) = P(A) + P(B)$$

E2.4

E2.4 is called the *Additive Rule of Probability*.

### Example 2.2 continued

If diagnoses A, B and C are mutually exclusive, we can say the probability of the rash being due to either A or C is:  $0.10 + 0.25 = 0.35$ . (Note that if A and C are *not* mutually exclusive, then we would need to subtract from 0.35 the joint probability of A and C occurring together – otherwise we would be “double-counting”.)



Note the use of the word “or” and its connection with the addition of probabilities. In probability, the word “and” is not used in connection with the simple addition of probabilities of mutually exclusive events. Sometimes this seems to be at variance with common English usage, but that’s life.

In fact the word “and” is used in connection with another rule, the *Multiplicative Rule*. This rule has two forms, one for independent events and another for events where the probability of one of the events is dependent or conditional on the other.

For *independent* events, say A and B, we have by the multiplicative rule :

$$P(A \text{ and } B) = P(A) \times P(B)$$

E2.5

where the expression on the left side of the equation means the *joint probability* of both events A and B occurring together.

### Example 2.2 continued

Let’s say that two babies are presented with the same sort of rash. The (joint) probability that the first will have condition A and the second will have condition B is:  $0.10 \times 0.65 = 0.065$ . This is true only if the events are independent.

For *dependent* events, say when the outcome of the second trial is conditional on the outcome of the first trial, the multiplicative rule has the following form :

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

E2.6

where the symbol “ $P(B|A)$ ” is read as “the probability of event B occurring *given* that event A occurs or has occurred”. In no sense does the vertical slash “|” imply a division of B by A. Note that if A and B are independent, then the occurrence of B is not affected by the occurrence or non-occurrence of A, so that  $P(B|A) = P(B)$ , and the two forms of the multiplicative rule E2.5 and E2.6 coincide.

### Example 2.2 continued

Let’s say that our two babies are siblings and that condition A is a contagious disorder. If it is known that the first baby has condition A, then we cannot assume that the probability of the second baby having condition B is  $P(B) = 0.65$ , since the occurrence of the condition in the second baby must surely reflect the contagious condition of the first. That is, an assumption of independence is unwarranted.

To make a sensible calculation of  $P(A \text{ and } B)$  we would first need to know the conditional probability of the second child having condition B, given that the first baby had condition A. That is, we would need to know, or estimate,  $P(B|A)$ .

The Law of Complements applied to conditional probabilities is:

$$P(\bar{A}|B) = P[\textit{not A} | B] = 1 - P(A|B) \quad \mathbf{E2.7}$$

### Example 2.3

The probability of breast cancer recurring after treatment within 5 years, *given* that it has spread to the lymph glands at the time of initial diagnosis (called *Stage II cancer*), is 0.6. What is the probability that a Stage II breast cancer will not recur within 5 years?

Solution:

$$\begin{aligned} P(\textit{No recurrence}|\textit{Stage II}) &= 1 - P(\textit{recurrence}|\textit{Stage II}) \\ &= 0.4 \end{aligned}$$

## § 2.3 BAYES’ THEOREM

Let us now take a bit of a detour from the development of probability theory and examine a practical use of some of the rules we have discussed.

### § 2.3.1 Introduction

The first payoff from knowing a little basic probability is *Bayes’ Theorem*, named for the eighteenth century English cleric and mathematician, Thomas Bayes. This famous result was read to the Royal Society in 1763, after Bayes’ death. Briefly, (see §2.3.3 for details) Bayes’ Theorem gives a method for manipulating conditional probabilities – and these crop up frequently in medicine and the health sciences. As far as we are concerned, Bayes’ Theorem has two applications:

- [1] It has proved very useful in approximating mathematically the way diagnoses are made. This has been put to use in programming computers to help diagnose illnesses. To arrive at a diagnosis, we use:
- information from the patient – symptoms, signs, test results [*a computer program for medical diagnosis would also require these data*];
  - knowledge of how common these clinical findings are in *each* of the diseases being considered as alternative diagnoses [*a computer program would need access to the relevant conditional probabilities, such as:  $P(\text{Test}^+|\text{Disease A})$  and  $P(\text{Test}^+|\text{Disease B})$ ]; and*
  - how common the possible diseases are in the population – the disease prevalences [*a computer program would need the unconditional probabilities of the diseases being considered:  $P(\text{Disease A})$ ,  $P(\text{Disease B})$  etc].*

**Example 2.4**

If a 30 year old woman presents to her general practitioner with marked weight loss, the doctor (with limited insight) might entertain a differential diagnosis of either anorexia nervosa (so-called “slimming disease”) or thyrotoxicosis (overactive thyroid gland). On the one hand, he thinks that weight loss always occurs in anorexia nervosa and not quite as frequently in thyrotoxicosis – this might sway the doctor into diagnosing anorexia nervosa. On the other hand, thyrotoxicosis might be a more common disease in 30 year old women than anorexia, leading to the alternative diagnosis.

A computer faced with this problem would utilize Bayes’ Theorem to arrive at a quantitative solution based on probabilities.

The other application of Bayes’ Theorem is:

- [2] The evaluation of *population-based screening programmes*. A screening programme aims at detecting disease at an *early* stage in *asymptomatic* persons so that effective treatment (or prevention for others) can be instituted. Examples of screening programmes are the use of Pap smears in cervical cancer, chest X-rays for tuberculosis, mammography for breast cancer, faecal blood tests for bowel cancer, biochemical tests for inborn errors of metabolism in newborns, and various immunological tests for HIV infection (“AIDS” virus).

You’ll learn more about such matters in epidemiology, but Bayes’ Theorem (hence probability theory) is crucial to this controversial field of public health.

**§ 2.3.2 The Formula for Bayes’ Theorem**

Here is the formula when *only two* possible diagnoses or conditions are possible. The formula is easily generalised to handle three or more possibilities, but two will suffice to illustrate the principles.

$$P(D_1|T) = \frac{P(T|D_1) \times P(D_1)}{[P(T|D_1) \times P(D_1)] + [P(T|D_2) \times P(D_2)]} \quad \mathbf{E2.8}$$

In E2.8,  $D_1$  and  $D_2$  refer to the two mutually exclusive and exhaustive diagnostic possibilities. That is, for this simplified version of Bayes' Theorem, the patient can only have one of the two diseases and there is no possibility of another disease. Note that these conditions are fulfilled when  $D_1$  corresponds to the *presence* of a particular disease (often denoted:  $D^+$ ) and  $D_2$  corresponds to that disease's *absence* (often denoted:  $D^-$ ). This is equivalent to saying that  $D_2$  could refer collectively to *all the other* possible conditions (including good health!). Or,  $D_1$  and  $D_2$  could be two different diseases that exhaust the diagnostic possibilities, as in the weight loss example mentioned in Example 2.4. For the two-state situation we are considering, it is obvious that  $P(D_1) = 1 - P(D_2)$ ; that is, the events are complements (see E2.3).

$T$ , the **conditioning event**, usually refers to some event of interest which describes the environment in which the diagnosis needs to be made, for example,  $T$  could be the presence of a symptom or the result of a diagnostic test. A positive test result would be denoted:  $T^+$ , a negative test by:  $T^-$ .

### § 2.3.3 What does Bayes' Theorem do?

Bayes' Theorem provides a method for "reversing a conditional probability". If you plug in the  $T|D$  and  $D$  information on the right hand side of E2.8, the  $D|T$  information pops out on the left hand side. This is just the way information often needs to be dealt with in medical or public health studies. Bayes' rule gives the probability of the particular diagnosis conditional on (or given) the presence of the symptom or a test result: what you need to do is plug in the unconditional probabilities of the diseases, the  $P(D)$ , and the conditional probabilities of the symptoms or test results given the diseases, the  $P(T|D)$ . Of course, the formula is symmetrical for  $D_1$  and  $D_2$ , so it's just as easy to find  $P(D_2|T)$ . Also, by using the rule of complements in probability, you can redefine  $T$  to get the probability of a diagnosis given the *absence* of a symptom or given a *negative* test result.

Example 2.5 looks at a practical calculation using Bayes' Theorem, and also provides a running tutorial on the general concepts of screening.

**Example 2.5**

In a certain population, a diagnostic test for the Human Immuno-Deficiency Virus will detect the disease in 90% of those who actually are afflicted. Also, if a person is *not* infected, the test will be negative for HIV with probability 0.95. It is estimated that 0.1% (or 0.001) of this population is infected with HIV.

A person is chosen at random from the population and is tested for HIV. The test is positive for the disease. What is the probability that the person is in fact HIV infected (in other words, that the test got it right)?

*Solution:*

The question asks for  $P(\text{HIV infected}|\text{Test positive})$ . The notation is:  $P(D^+|T^+)$ . We define our events first, and assign probabilities to them. The assumption is that the proportions provided by the data (the prevalence, sensitivity and specificity) are valid estimates of the required probabilities.

$D^+$  is the event “HIV infected” and  $P(D^+) = 0.001$

$D^-$  is the event “HIV free” and  $P(D^-) = 0.999$

$P(D^+)$  would be called the *prevalence* of HIV infection by epidemiologists; the prevalence in this population is 0.001, or 1 per 1000. The prevalence of non-infection,  $P(D^-)$ , is 0.999, or 999 per 1000.

Note that the unconditional probabilities of the mutually exclusive and exhaustive diagnostic possibilities,  $P(D^+)$  and  $P(D^-)$ , sum to 1, as they must always do.

Continuing with definitions:

$T^+|D^+$  is the event “test positive *given* HIV infected”

so:  $P(T^+|D^+) = 0.90$  (*straight from the data*)

$T^-|D^+$  is the event “test negative *given* HIV infected”

so:  $P(T^-|D^+) = 0.10$  (*using complements and E2.7*)

Epidemiologists would call  $P(T^+|D^+)$  the *sensitivity* of the test – it is the probability that a test will register positive in an infected person. Its complement,  $P(T^-|D^+)$ , is the probability of having a *false negative* test – the test incorrectly registers negative in an infected person.

$T^-|D^-$  is the event “test negative *given* HIV free”

so:  $P(T^-|D^-) = 0.95$  (*straight from the data*)

$T^+|D^-$  is the event “test positive *given* HIV free”

so:  $P(T^+|D^-) = 0.05$  (*using complements and E2.7*)

**Example 2.5 continued**

Epidemiologists would call  $P(T^-|D^-)$  the *specificity* of the test – it is the probability that a test will register negative in a non-infected person. Its complement,  $P(T^+|D^-)$ , is the probability of having a *false positive* test – the test incorrectly registers positive in a non-infected person.

Sensitivity and specificity are in-built attributes of the particular testing procedure and equipment used, but also depend on the existence of diagnostic “gold-standards” and the nature of the sub-population being screened. False positives and false negatives bedevil *all* screening tests and *all* laboratory diagnostic tests.

Back to solving our problem. We just plug the relevant probabilities into the formula E2.8:

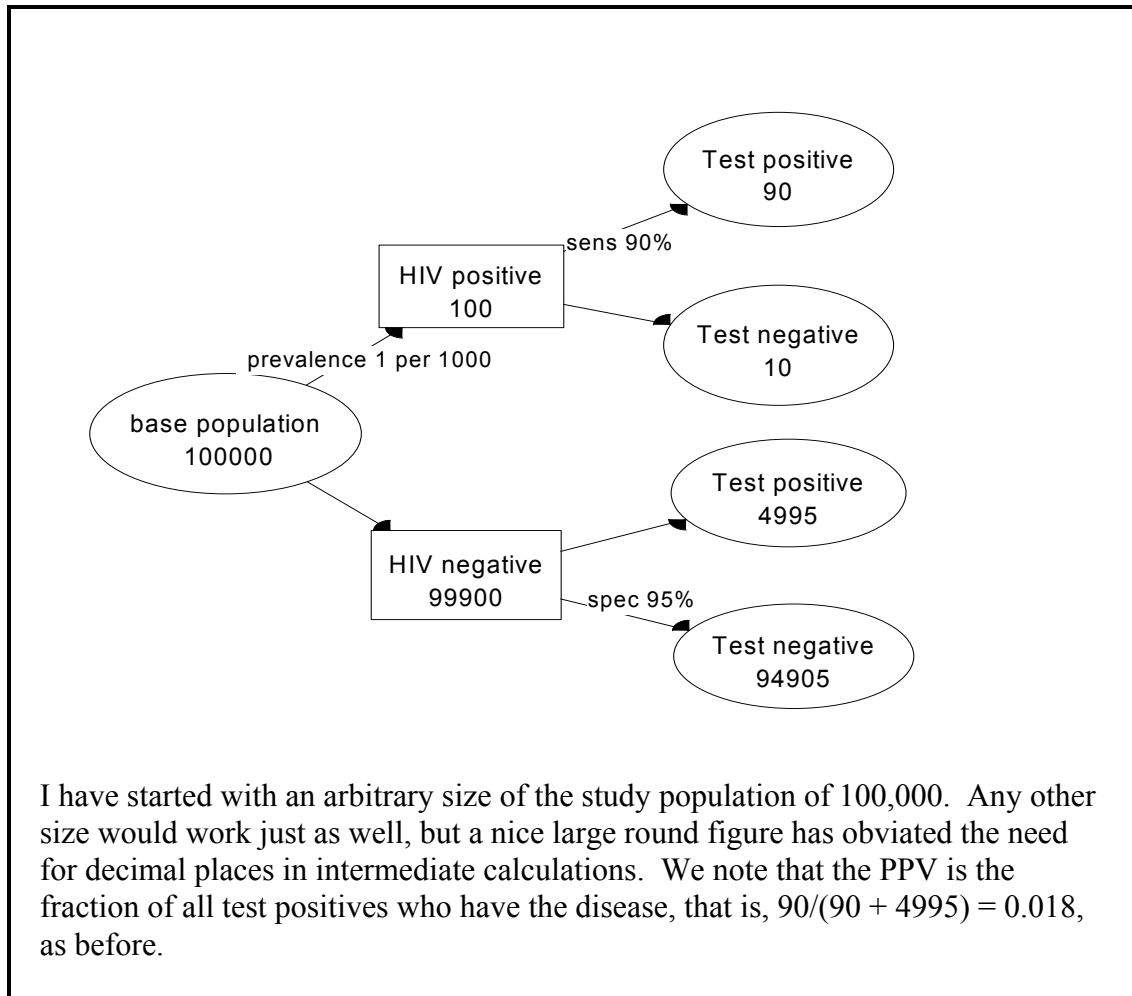
$$\begin{aligned}
 P(D^+|T^+) &= \frac{P(T^+|D^+) \times P(D^+)}{[P(T^+|D^+) \times P(D^+)] + [P(T^+|D^-) \times P(D^-)]} \\
 &= \frac{0.90 \times 0.001}{(0.90 \times 0.001) + (0.05 \times 0.999)} = 0.018 \quad (\text{to 3 decimal places})
 \end{aligned}$$

In other words, for every 1000 persons whose test is positive, only 18 will actually be infected with HIV. The value we have calculated is termed the *yield* or the *positive predictive value (PPV)* of the screening test. A yield of 1.8% means that 98.2% of those whose tests are positive for HIV will *not* be infected. You may care to speculate on the personal and economic costs associated with such a state of affairs.

*Just when you thought you'd got it right department:* the definitions I have given for false positives,  $P(T^+|D^-)$ , and false negatives,  $P(T^-|D^+)$ , while the commonest used, are not universally agreed upon. Some authors define false positives as  $P(D^+|T^-)$ , which in our terms would be the complement of the negative predictive value; they would define false negatives as  $P(D^-|T^+)$ , which we would call the complement of the positive predictive value. The lesson here is to pay close attention to definitions when reading papers on screening, and to be doubly careful in your interpretation and assessment if the author does not explicitly define his or her terms.

### § 2.3.4 An Alternative Approach – the Tree Diagram

Note that a simple problem such as that posed in Example 2.5 can also be readily solved by drawing up a tree diagram (Fig 2.2). If you find such an approach simpler, then use it! (but always try and relate the branches of your tree to the conditional probabilities that are operating to produce them). You will probably come to the conclusion that using a tree is easier for simple problems, but that a more complex problem, say, one involving multiple diagnostic categories, would require a computer program utilising a generalisation of Bayes' Theorem.



**Fig 2.2** Tree diagram for the data of Example 2.5

### § 2.3.5 Lessons from Example 2.5

In Chapter 1, the general concepts of error and variability were introduced. Problems associated with screening tests and diagnostic laboratory tests illustrate the world of variability that health workers and researchers inhabit. A test result cannot be interpreted sensibly without considering these points:

- *Tests are imperfect*; no test has 100% sensitivity and 100% specificity. Usually a trade-off needs to be made. A highly sensitive test will also mean an increase in false positives; a highly specific test will mean an increase in false negatives.
- *Whenever you are faced with a test result* remember: A positive test does not necessarily imply a diseased person (problem of poor Positive Predictive Value); a negative test does not necessarily imply absence of the disease (problem of poor Negative Predictive Value).
- *Whenever you are faced with a patient* remember: An ill patient may not have a positive test (problem of false negatives); a healthy patient may have a positive test (problem of false positives).

- Even a highly sensitive and highly specific test (one that correctly identifies a high proportion of both diseased and non-diseased persons) when applied to a population in which the disease prevalence *is low*, will deliver a low Positive Predictive Value. Such a screening programme will lead to the mislabelling of healthy persons as being diseased

That's the end of our detour. Those of you studying epidemiology will learn more about screening. Now we'll rejoin our discussion of probability.

## § 2.4 PROBABILITY DISTRIBUTIONS

### § 2.4.1 Introduction

We have met the term *distribution* before in the context of a table or graph representing the frequencies or relative frequencies of events. A ***probability distribution*** is a representation of the allocation of probabilities to values of a random variable.

Using the definition of probability as the *long run* relative frequency, we could use the relative frequencies of our “short-run” sample data to estimate the probabilities that we might find if we collected a much larger sample or, in the end, collected data on the entire population. The more data we have, hopefully the better our empirical relative frequency distribution approximates the probability distribution.

#### **Example 2.6:**

We all “know” that the long run relative frequency (the probability) of heads or tails on coin tossing is 0.5. This most simple of probability distributions is given as:

<u>value</u>	<u>probability</u>
head	0.5
tail	0.5

To be specific, the random variable here is “the result of a coin toss”. It takes two values: head or tail. You could toss a coin 100 times and perhaps get a 55:45 heads:tails result. If you had the patience, you might toss it 1000 times – it would not surprise you that the ratio more closely approaches 1:1, say 515:485. After 1,000,000 tosses the result might be 500,010:499,090 – and so on, getting closer and closer to the theoretical ratio of 1:1.

It turns out that, in many situations, we can do a lot better than this and don't even need to perform the trials at all. In fact, we can sometimes calculate the exact relative frequency (for each value of the variable in question) that would occur given an infinite number of trials, that is, we can calculate the distribution of probabilities. These theoretical probability distributions are given by mathematical formulae called ***probability functions*** – you plug in the value of the random variable and the formula gives you the probability associated with that value in the population. ***A probability function is a rule for the allocation of probabilities to values of a random variable.***

Theoretical probability distributions can describe certain variables which take discrete values or certain variables which are continuous. In the continuous case, the probability function is often called a **density function** (remember, in the context of a relative frequency distribution, the histogram gave you an idea of how “dense” the data were in a sub-interval?).

The general formula for a probability function is:

$f(x) = \{\text{complicated expression}\}$	[for a valid domain of x]	<b>E2.9</b>
--	---------------------------	-------------

Let’s look at E2.9 from a general point of view, before moving on to examine some examples (see §2.4.2 and §2.4.3).

The notation  $f(x)$  is the standard mathematical notation for a function. It says “this is a *rule* for operating on, or manipulating, the values of the variable, x”. One of the simplest mathematical functions is:  $f(x) = x$ , which just says that the rule is to leave x alone. Another is  $f(x) = \sqrt{x}$ , which tells us to take the square root of x. Unfortunately, probability functions are more complicated than this!

The *{complicated expression}* on the right hand side of E2.9 defines the rule for operating on x. The expression will differ for different probability functions. The *form* of this expression identifies the **type** (or **class**) of probability distribution. Examples of classes of probability distributions are the **Binomial** distribution, the **Normal** distribution and the **Chi-Square** distribution. We will meet these later on.

The expression involves the values of the random variable, x, and perhaps one or more **parameters** which are (hopefully) known constants identifying the population being considered. For example, the arithmetic mean and the standard deviation, which we called *statistics* when they were calculated from sample data, are examples of *parameters* when used in the context of populations. Parameters can be spotted easily since they are usually represented by Greek letters, as opposed to the Roman letters used for sample statistics. The *population* mean is **mu** ( $\mu$ ), and the *population* standard deviation is **sigma** ( $\sigma$ ); [this is small sigma; we’ve already met capital Sigma,  $\Sigma$ , the summation operator]. The particular values of the parameters identify a *unique* distribution within its *class*. I’ll give some examples below.

The comment *[for a valid domain of x]* simply reminds us that some random variables may not necessarily take all possible values on the real number line.

### **Example 2.7**

One cannot have a value of less than zero for the variable: *weights of children*. The valid domain here would be all values greater than zero, even though some small and some large numbers would never be realised. The notation here is:  $x: x \geq 0$ , which is read: “x, such that x takes all non-negative values”.

### § 2.4.2. The Binomial Distribution

The Binomial Distribution is one of a number of classes of distributions used with *discrete* data. (Other classes of discrete distributions are the *Poisson*, *Geometric*, and *Hypergeometric* distributions; we won't consider them further in this short course.) Remember that a discrete probability distribution describes probabilities for a random variable which can take only a finite (or, for the mathematicians, “countably” infinite) number of values.

Consider the variable of interest to be the *number of heads* that arise from tossing a coin 5 times. (Admittedly, coin tossing gets a bit boring, but it does provide a simple *model* for many real-life situations; see Example 2.8). It is known that such a variable has a Binomial distribution. The class of Binomial distributions is described by the function:

$$f(x) = {}^n C_x \cdot \pi^x \cdot (1 - \pi)^{n-x} \quad [\text{for } x=0,1,2,3,\dots,n] \quad \mathbf{E2.10}$$

where:

- $x$  takes on the possible values of the random variable; in our coin tossing example,  $x$  could take values 0,1,2,3,4,5. In general,  $x$  ranges from 0 to  $n$ .
- $C$  is the combinatoric operator, which you may have met in high school. It involves factorials, denoted by the symbol “!” . You will remember that  $k! = 1 \times 2 \times 3 \times \dots \times (k-2) \times (k-1) \times k$ . The symbol  ${}^n C_x$  equals  $n!/[x!(n-x)!]$ . So, for example,  $4! = 4 \times 3 \times 2 = 24$  and  $6! = 6 \times 5 \times 4! = 720$ . If you pose the question: how many ways can I choose 4 objects from 6 (omitting different orderings of the same chosen 4), the answer is  ${}^6 C_4 = 6!/[4!(6-4)!] = 6!/(4! \times 2!) = 15$ .
- $n$  is the number of trials; it is one of two parameters (but not a Greek letter this time!) needed to specify which *particular* binomial distribution we are dealing with. In our example  $n = 5$ .
- $\pi$  is the Greek small letter “pi”; the probability of “success” (getting a head) in a *single* trial. This has got *nothing* to do with the mathematical constant relating the circumference of a circle to its diameter ( $\pi \sim 3.1415$ ). To avoid confusion, we often designate this parameter as “p” – even though, to be consistent, p should really stand for our *sample* proportion of successes. Anyway, success is defined arbitrarily – in a coin tossing example we could call a head a success (and a tail a failure) or a tail a success (and the head a failure) – just as long as we state our definition from the start and are consistent.  $\pi$  is the other parameter of the binomial distribution. The complement of  $\pi$ ,  $1-\pi$ , is the probability of “failure”, in our case getting a tail.

Note that E2.10 uses an alternative notation for multiplying factors: “.”. This is to avoid confusion with “x” also being the notation for the random variable (and probably add to confusion if any decimal points are around).

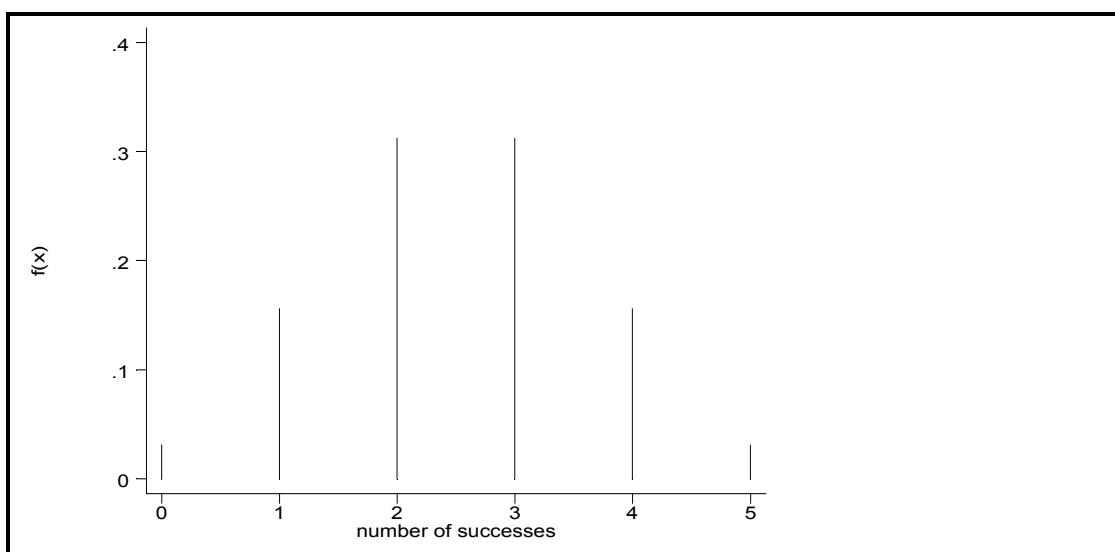
In our example we have specified the values of the two required parameters,  $n = 5$  and  $\pi = 0.5$ , so we know our *particular* binomial distribution has the function:

$$f(x) = {}^5C_x \cdot 0.5^x \cdot (1 - 0.5)^{5-x}$$

and we can go about calculating probabilities associated with each of the possible outcomes:  $x = 0, 1, 2, 3, 4$  or  $5$ . For example, using the binomial formula one could easily find the probability of getting  $x = 2$  heads (and therefore 3 tails) in the  $n = 5$  coin toss trials.

$$\text{Prob}(2 \text{ heads}) = f(2) = {}^5C_2 \cdot 0.5^2 \cdot (1 - 0.5)^3 = 5!/[2!(5 - 2)!] \cdot 0.5^2 \cdot 0.5^3 = 0.3125$$

Calculate each probability for  $x = 0, \dots, 5$  and add them up; the total should, of course, be 1 (see E2.2).



**Fig 2.3** Binomial distribution with parameters  $n=5$ ;  $\pi=0.5$

Fig 2.3 is a graphical representation of the binomial probability distribution with parameters  $n = 5$  and  $\pi = 0.5$ . For example it tells us that over a very large number of 5-toss experiments (that is, *in the long run*), we expect that 31.25% of the trials will yield 2 heads and 3 tails. In a discrete distribution such as this, the probabilities are drawn as vertical lines with heights proportional to the probabilities. Sum the heights – they should come to 1.

*To summarise so far:* if you know that you are dealing with a variable following a theoretical binomial distribution and you know the specific values of the parameters of the distribution ( $n$  and  $\pi$ ), then you can calculate the expected probabilities for any value of the variable. (And you *do not* have to repeat experiments endlessly.)

It is obvious that there is an infinite number of possible binomial distributions depending on the parameters supplied. For example, a different binomial distribution describes the number of heads on 6 tosses of a coin; another describes the number of times a die face of 4 occurs on 23 rolls of a die – here the parameters are  $n = 23$  and  $\pi = 1/6$ , since the probability of a 4 on a single trial is  $1/6$ . Published tables give probability results for many combinations of  $n$  and  $\pi$  or you can use a calculator and E2.10.

How do you know that the random variable you are dealing with follows a binomial distribution? If the trials that generate the data are ***Bernoulli trials*** (named after an early 18<sup>th</sup> century Swiss mathematician) then the variable has a binomial distribution (in shorthand, it is a “binomial variable”). A Bernoulli trial is a trial where:

- there are only *two* possible mutually exclusive outcomes;
- the probability of success,  $\pi$ , is constant from trial to trial; and
- each trial’s result is unaffected by any other trial’s result, that is, the trials are *independent*.

So a coin toss fulfills these criteria. What has all this got to do with medicine and the health sciences? Well, if you can convince yourself that your data could have been generated by repetitions of a Bernoulli process (and in the context of proposing a model to simplify life, convincing oneself of such things is not too hard) then you can use the binomial distribution to calculate probabilities and provide a theoretical justification for any claims you make. Here is an example from an occupational health setting.

### **Example 2.8**

Let’s say we are interested in how often a “needlestick” injury, from a needle used to collect blood from a known Hepatitis B patient, leads to the appearance of core antigen (evidence of infection) in unvaccinated, non-immune nurses working in a drug treatment clinic.

Assume it is known from large surveys in such clinics that the probability of infection (“success”) in the population of nurses after such an incident (a trial) when there is no protective vaccination is:  $\pi = 0.2$ . At the clinic in question, 10 vaccinated nurses suffer a needlestick injury during 1996. What is the probability that more than the expected 2 nurses will become infected?

Let’s look at the Binomial probability distribution with parameters  $\pi = 0.2$  and  $n = 10$ . You could (and should!) calculate this table of expected probabilities using E2.10.

number of “successes” (hepatitis cases)	probability (to 3 decimal places)
0	0.107
1	0.268
2	0.302
3	0.201
<i>etc</i>	<i>etc</i>
10	0.000

We note that the probability of 0, 1 or 2 nurses becoming infected is  $0.107 + 0.268 + 0.302 = 0.677$ . So by the Law of Complements E2.3, the probability of more than 2 becoming infected is 0.323.

### § 2.4.3 The Normal Distribution

We now turn to probability distributions for continuous data. Some of the commonly encountered distributions are the *Normal* distribution, the *Student t* distribution and the *Chi-square* distribution. Each of these is a class of distributions – once again, within each class, values of the associated parameters define a unique distribution. We will consider the t-distribution and the Chi-square distribution in later chapters.

The most important of the continuous probability distributions is the *Normal* or *Gaussian distribution*, named for the contributions in this area of Carl Friedrich Gauss, a German mathematician (1777-1865). He is considered by many to be the greatest mathematician who ever lived. The mathematical formula for the Normal distribution was first published by De Moivre in 1733. *This distribution will haunt you for the rest of your life, so you should become absolutely familiar with it.*

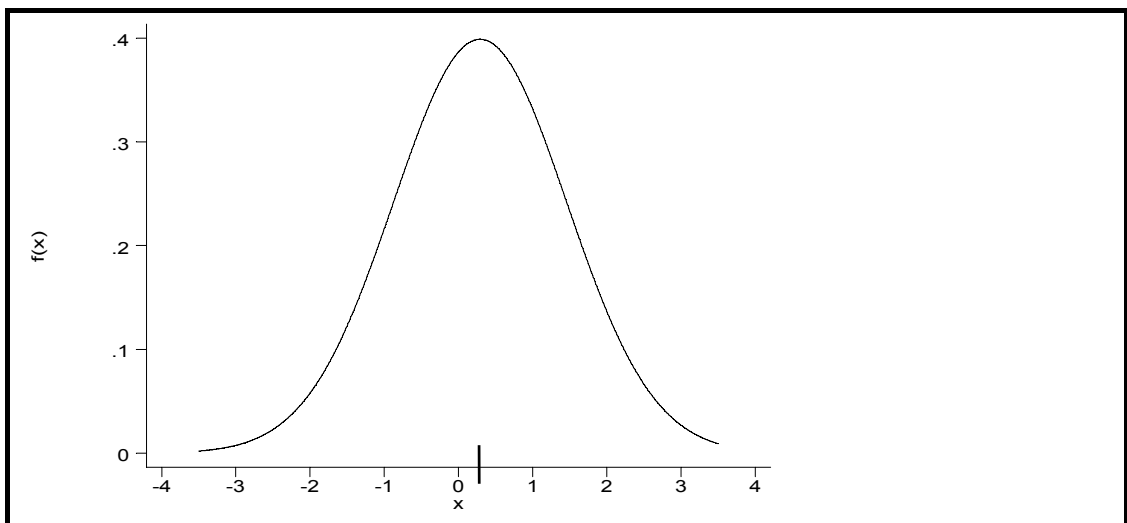
The Normal probability function is defined by E2.11:

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot e^{-\left\{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}} \quad [\text{for all real } x] \quad \text{E2.11}$$

In this function,  $\pi$  and  $e$  are the usual mathematical constants ( $\pi=3.14159$  and  $e$ , the base of natural logarithms, = 2.71828 approximately).

To define a *unique* Normal distribution, we need only specify the values of the *two* required parameters, the population mean,  $\mu$ , and the population standard deviation,  $\sigma$ . Theoretically,  $x$ , the values of the random variable in question, can range from -infinity to +infinity ( $-\infty$  to  $+\infty$ ).

Note that the distribution is symmetrical about the population mean, and that the standard deviation describes its spread. It is often referred to as a *bell-shaped curve*. See Fig 2.4 for a curve centred on (that is, the mean equals) 0.25.



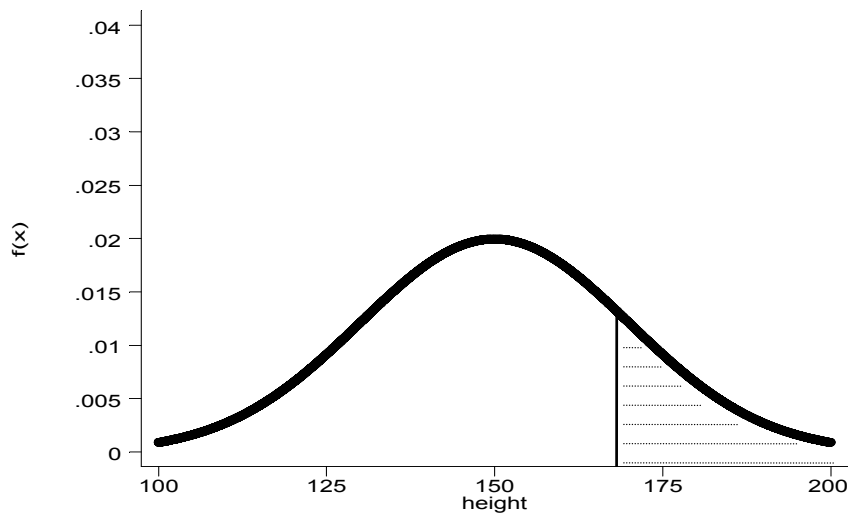
**Fig 2.4** Normal distribution (mean = 0.25)

In accordance with our basic rules of probability the total *area* under the curve (from  $x = -\infty$  to  $x = +\infty$ ) equals 1. (This can be proven by integration of E2.11 after transforming to polar coordinates, for those with an interest in such esoterica.)

One does not usually speak of the probability of getting a *particular* value in a continuous distribution (for example, the probability of the height of an adult male being 170 cm), but rather the probability of getting a value as large (implying “as large or larger”) or as small (implying “as small or smaller”) as a specified value.

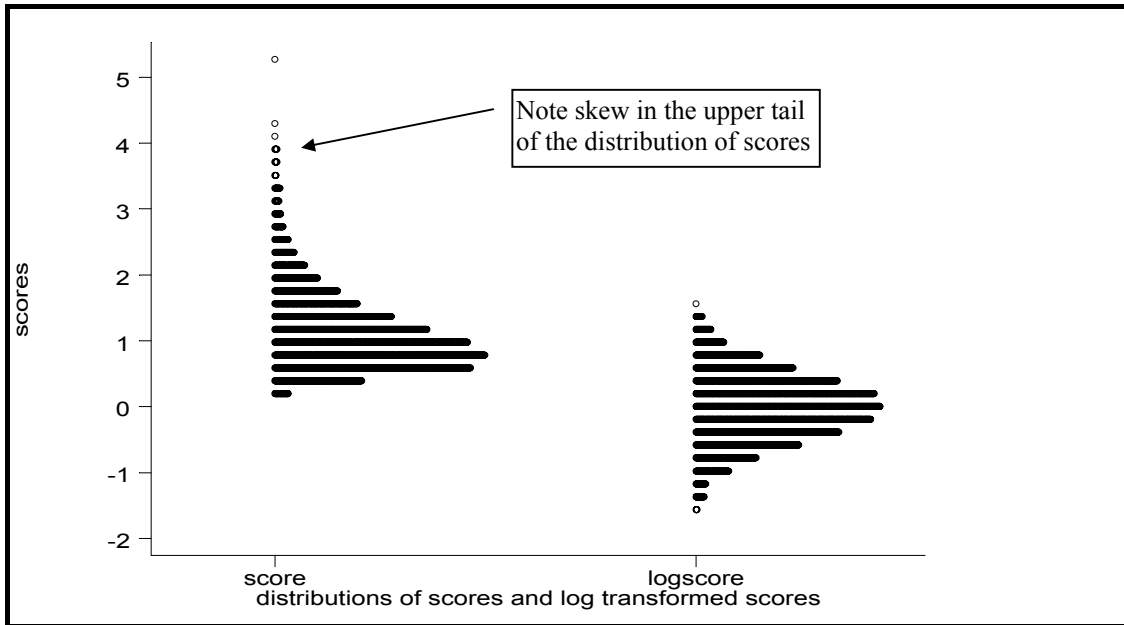
### Example 2.9

The probability of an adult male’s height being 170 cm or more would be given by the area under the Normal distribution curve for adult male heights between values on the abscissa, or height axis, of 170 and  $+\infty$ . This is the shaded area in Fig 2.5. To calculate the value of this area, we would need to know (i) the mean of the distribution and (ii) the measure of “spread” – the standard deviation.



**Fig 2.5** Normal curve of adult male heights in a population (mean 150 cm)

The Normal distribution curve has some interesting and very useful features. It describes the probabilities of values of many biological variables, for example, weight, height, systolic blood pressure and the crown-rump length of new-borns. Even if a variable does not follow the Normal distribution, sometimes a simple **transformation** of the data will bring it much closer to Normal – and this may be useful when attempting to use certain statistical testing procedures. For example, a random variable may have a highly positively skewed distribution, but if the logarithm of each raw score is taken, then these transformed scores may become distributed in a very near-Normal fashion. This may facilitate, for example, testing if mean scores differ between two defined groups. Fig 2.6, a dot-plot, shows the effect of using the log transform on scores whose original distribution was markedly positively skewed.



**Fig 2.6** Effect of the log transform on a positively skewed distribution

We note in passing that other types of distributions may require different transformations depending on the circumstances and objectives: examples include the *arcsine transformation* for proportions, *square root transformations*, also useful for positively skewed data, and the *logistic transform* to enable modelling of binary (0/1) data, as might be the case if an epidemiologist wishes to determine which pre-existing factors predict *whether or not* (coded as 1 or 0) a subject develops a malignant disease.

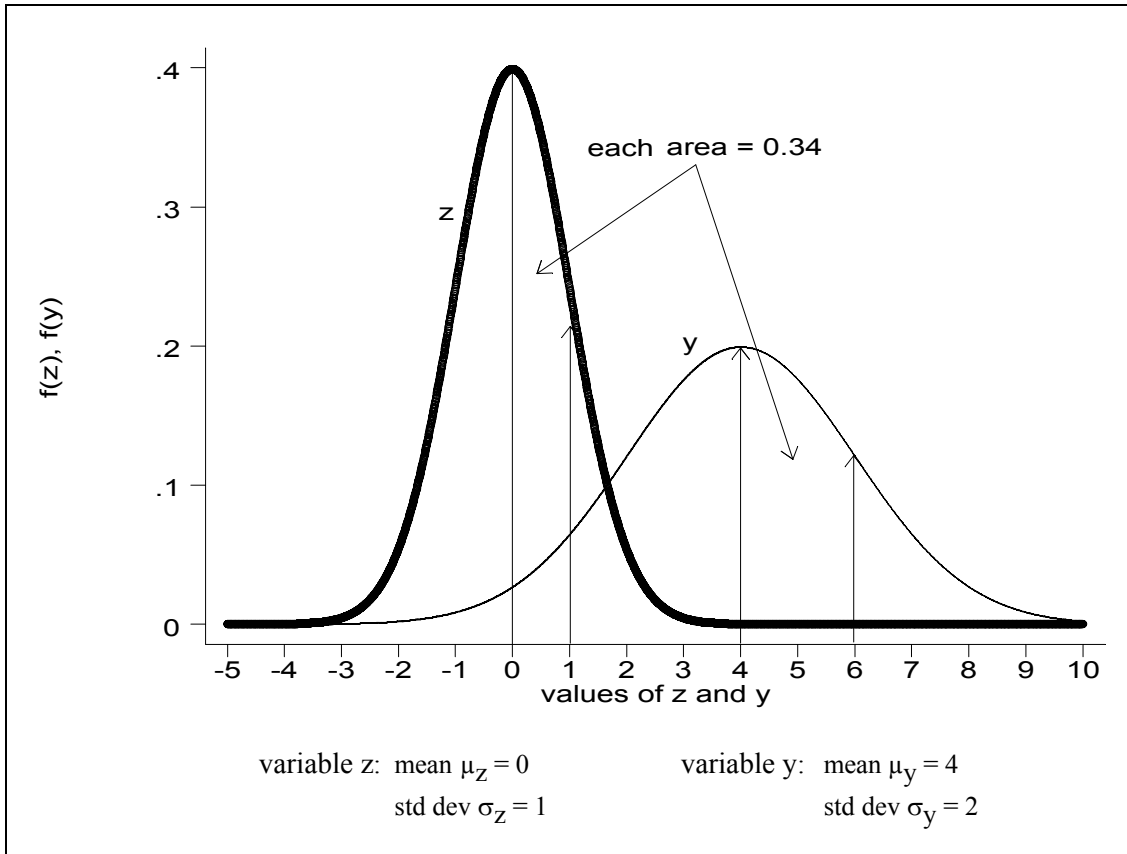
Why should the Normal distribution be so good at describing the probabilities of values of biological variables? The *central limit theorem* states that the addition of many variables, which themselves may not be Normal, tends to produce a resultant variable which is Normally distributed. You can think of a biological variable, say blood pressure, as being the resultant of the contributions of many component variables, for example, genetic endowment, age, stress level, condition of the arteries and the heart etc, so you can see that, by the central limit theorem, many quantities of interest to us should have a near-Normal distribution. This is one of life's few great freebies.

Since there is an infinite number of possible combinations of different means and standard deviations there is also an infinite number of Normal probability curves. Refer back to Fig 1.4 which shows just three of them.

#### § 2.4.4 The Standard Normal Distribution

A problem arises from the multitude of Normal distributions. If we wished to look up probabilities associated with a particular curve, for example the distribution of birthweights in Tasmanian babies, we would need tables of probabilities specific for that curve. It is obviously impractical to construct statistical tables for every Normal distribution.

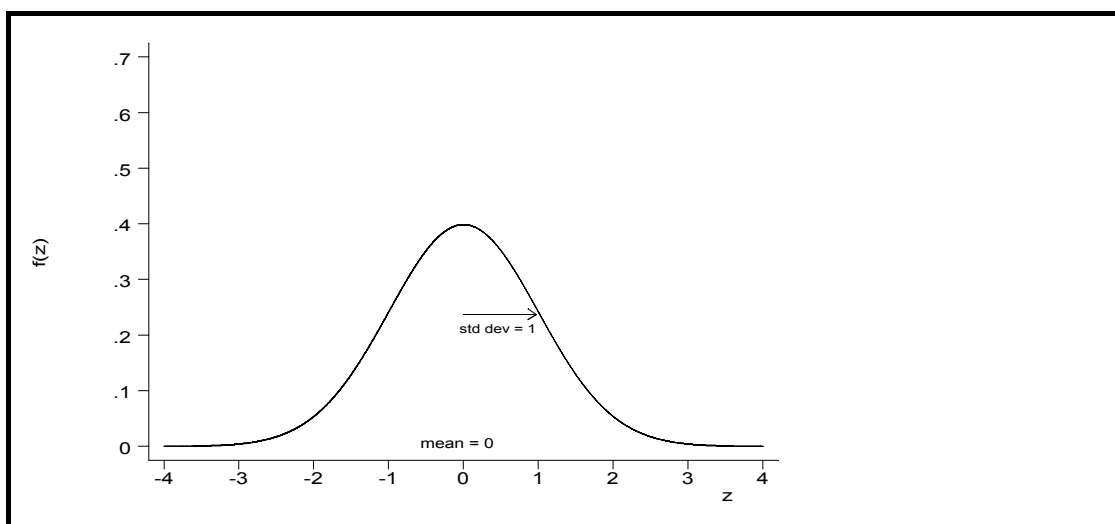
However, a very convenient feature of the Normal curve is that if one measures out a distance from the mean value of the variable, along the x axis, in units of the standard deviation of that variable, then the area swept out under the curve (that is, the *probability*) will be the same for any of the family of Normal curves, regardless of the particular mean and standard deviation.



**Fig 2.7** Equal areas (probabilities) of different Normal curves

For example, 0.68 or 68% of the total probability is between  $-1$  and  $+1$  standard deviations from the mean, and 95% within  $\pm 1.96$  standard deviations from the mean, *regardless of which particular Normal curve we are dealing with*. So, referring to Fig 2.7, about 34% of the area under the curve lies between 0 and  $+1$  for the variable  $z$ , which has a standard deviation of 1, and, similarly, 34% lies between 4 and 6 for the variable  $y$ , which has a standard deviation of 2. This suggests a way we can calculate areas under curves for any Normal curve we come across.

For simplicity we agree to call the particular Normal distribution which has a mean of 0 and a standard deviation of 1 the **Standard Normal Distribution** or **z-distribution** (since values of standard Normal variables are usually designated by a “z” rather than an “x”). The z values are often called **standard Normal deviates**, not because of any inherent personality defect, but because they represent *deviations from the mean in units of the standard deviation*.



**Fig 2.8** Standard Normal distribution (z) curve

For this special Normal curve (Fig 2.8), probabilities relating to the area under the curve for values between  $-\infty$  and  $+\infty$ , have been tabulated (see the Appendix). When you are dealing with a variable whose distribution is Normal but not *Standard* Normal then one can still use the Standard Normal tables to calculate probabilities since *every x-value of a non-standard Normal variable has a corresponding Standard Normal z-value*. The correspondence is given by the *standard Normal transformation* (also called the *z-transformation*):

$$z = (x - \mu) / \sigma$$

**E2.12**

where  $x$  is the original value of the variable,  $\mu$  is the mean of the original distribution,  $\sigma$  is the standard deviation of the original distribution and  $z$  is the new value of the standard Normal variable whose corresponding probability is in the tables.

***The probability associated with the z-value will be identical to the probability associated with the original x-value.***

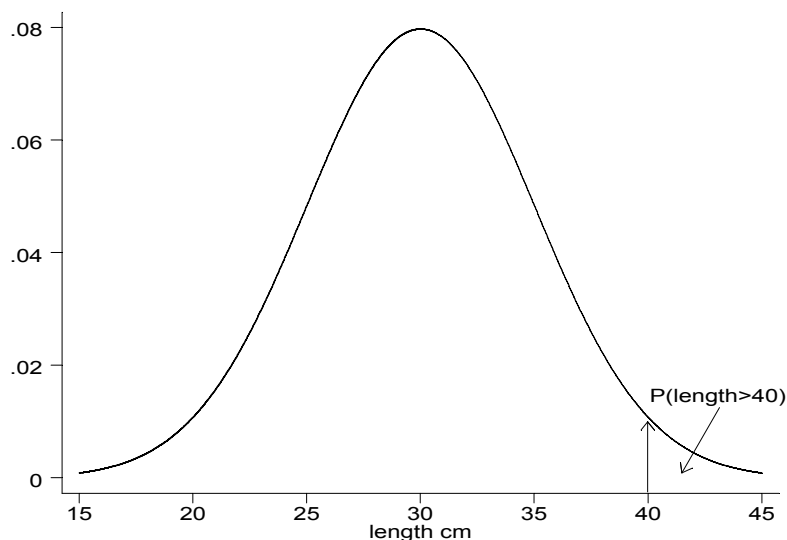
Since the standard Normal curve is symmetric about the mean (zero), tables often just give results for positive values of  $z$  and you have to do a mirror image trick in your mind to deal with negative values and/or use the property of complements E2.3.

So to find the probability of getting a value as large or larger than  $x$ , say, you transform your  $x$ -value to a standard  $z$ -value and look up a standard Normal table.

**Example 2.10**

It is known that the mean and standard deviation of the Normal curve describing the distribution of babies' crown-rump length at birth are 30 cm and 5 cm respectively. What is the probability of a baby having a crown-rump length of at least 40 cm?

Always (that means *always*) draw a little diagram. This step will often save you from coming up with nonsensical results. The question implies finding  $P(\text{length} \geq 40 \text{ cm})$ . We are after the area under the probability curve indicated in Fig 2.9.



**Fig 2.9** Normal distribution for data in Example 2.10

Now we can't look up a table of probabilities for this curve (since none exists) so we'll transform the original length measurement into a z-score using E2.12. A length of 40 cm becomes a z-score of 2:

$$z = (40 - 30)/5 = 2 \quad (\text{note: no units})$$

If you check Table [1], the area under the standard Normal curve for  $z \geq 2$  is 0.023 or 2.3%. From our discussion above, this is the *same area* as that under the original Normal curve (with mean = 30 and standard deviation = 5) above a length of 40 cm.

*Conclusion:* the chance of getting a baby as big or bigger than 40 cm in this population is only 0.023 (2.3%) – which is pretty low.

You can immediately see two other things. First, by the law of complements, the probability of getting a baby whose length is *less* than 40 cm is  $1 - .023 = 0.977$  or 97.7%. Second, since the curve is symmetric about the mean, the probability of getting a baby with length 20 cm or less is also 0.023. You will note that the z-score in this case is  $(20 - 30)/5 = -2$ . Put still another way,  $1 - (2 \times 0.023) = 0.954$  or 95.4% of babies are born with crown-rump lengths between 20 and 40 cm. So, as we expect, about 95% of babies lie between  $\pm 2$  standard deviations of the mean. (A little more accurately, 95% of the population of babies lie within  $\pm 1.96$  SD of the mean.)

You should remember the following z-scores and the probabilities they cut off.

- $z = \pm 1.96$  cuts off 0.025 in *each* tail of the standard curve (0.05 or 5% in total)
- $z = \pm 1.645$  cuts off 0.05 in *each* tail of the curve (0.1 or 10% in total).

These z-scores are called **critical values** because they are commonly used as criteria for deciding if an observed value is – given what is known or assumed about the population – *usual or unusual*. In Example 2.10 above, we might be justified in saying that a baby of 40 cm or more would be unusual, given the reported mean and variability of the population birth lengths. You might go further and wonder (if such a birth length was the result of a valid random selection from the population) whether the true mean of the population really was 30 cm. In other words, in the face of the observed evidence, is the model claimed for the population wrong?

Actually, any z-score can be defined as “critical”: it just depends on how low the probability gets before you wish to consider that something unusual is happening. But the two quoted above are by far the commonest in use.

## § 2.5 SAMPLING DISTRIBUTIONS

### § 2.5.1 Introduction and Definition

So far we have been looking at the probability distribution of all the possible individual values of a variable in a population, for example, weight, height or blood sugar in a population of adult males. But we could also look at the distribution of possible values of one or other of the *statistics* which can be calculated on *repeated samples* taken from a population.

#### **Example 2.11**

The *minimum value* of a set of observations is one of a number of different statistics that can be calculated on a sample. (Other possibilities are the mean, the range, the maximum, the median etc.) We could construct a distribution of the minimum values of titres of antibody to rubella virus observed in random samples of schoolgirls aged 13, each sample containing 12 girls, one such sample taken from every school in the state. If there are 200 schools in our sampling frame, then we will be able to form a relative frequency distribution of minimum rubella titres (for samples of size 12) using the 200 minimum values we have collected.

Now take this a (theoretical) step further, and select an infinite number of samples – a *long run* process. We would be in a position to calculate the *long run relative frequencies* of values of the statistic.

*The allocation of probabilities to each possible value of a sample statistic is called the **sampling distribution** of that statistic. As we will now discuss, *sampling**

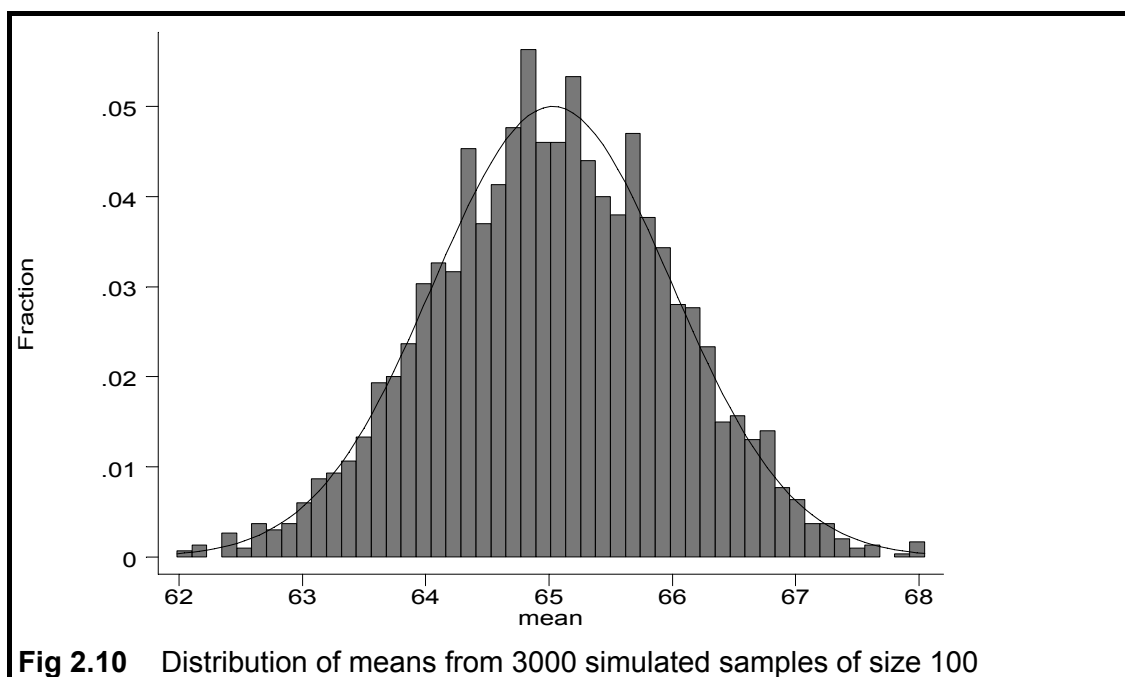
*distributions provide models which are the bases of inferential procedures.* The concept of sampling distributions is the most important concept in any development of the inferential statistics.

### § 2.5.2 Sampling Distribution of the Mean

A commonly used sampling distribution is that of sample means. It is referred to as the *sampling distribution of the mean* or, alternatively, as *the distribution of sample means*. For example, we could choose a random sample of  $n = 100$  people from a population and calculate the mean weight of the sample. Then we choose another sample of 100 people from the same population and calculate the mean weight of this second sample. Repeat the process until all possible samples of size 100 have been drawn. If the original population is infinite, as is usually assumed, then you will end up with an infinite number of sample means. These sample means are just numbers, so they will have their own *long run relative frequency distribution* with a mean (that is, the mean of the means) and a standard deviation (termed the ***standard error*** of the mean). This distribution can be plotted, if you have the patience or a computer.

#### **Example 2.12**

I most certainly did *not* have the patience to take repeated samples of size 100 *endlessly* from the population and calculate the mean weight of each sample. But I still want to demonstrate aspects of such a sampling distribution, so I programmed a statistical package to perform a ***simulation***. A simulation uses the computer's capacity for endless and mindless iterations of a task to achieve empirical results that would be impossible "by hand". I constructed a notional population of adult weights with a Normal distribution, mean = 65 kg, standard deviation = 10 kg. The program then chose a sample of  $n = 100$  weights at random from this population, calculated the mean of the sample, stored it away, then repeated the process. 3000 samples (and a hot chocolate in Rundle Street) later, I called a halt to the process, and plotted the "nearly long-run" relative frequency distribution of the 3000 mean weights. Here is the result, with a superimposed theoretical Normal curve:



**Fig 2.10** Distribution of means from 3000 simulated samples of size 100

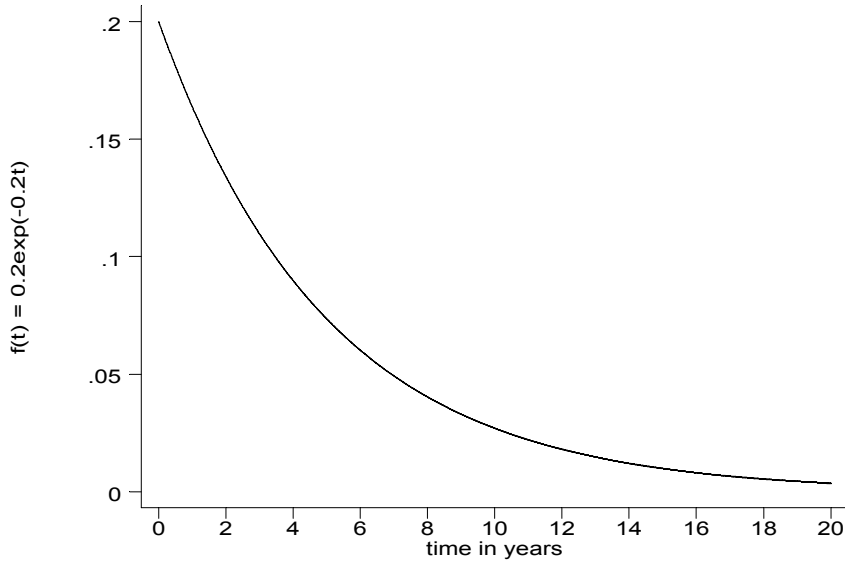
The *remarkable thing* is that this distribution of the sample means will be very close to Normal.

Possibly, you are not yet sufficiently impressed? After all, in Example 2.12, the underlying population variable, people's weight, does happen to be approximately Normally distributed (and, in my computer-based simulation, *was* Normal by construction), so you might feel it is natural that the sampling distribution of mean weights based on this population distribution will also be Normal. The following example, however, cannot fail to impress.

### Example 2.13

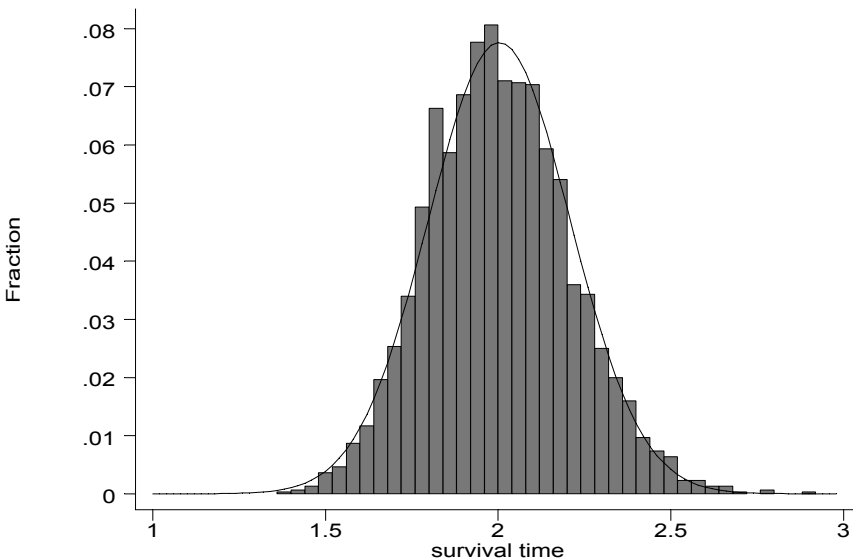
Let us consider a theoretical distribution of survival times for subjects with a particular malignant disease. The *exponential probability distribution* is a commonly used model for such a situation. Here is a plot of an exponential distribution with mean survival time of 5 years (equivalent to a death rate of 0.2 per year):

**Example 2.13 continued**



**Fig 2.11** Exponential probability distribution with hazard  $\lambda = 0.2$

For now, the details of the exponential distribution are not important – just as long as we agree that, even by a quick graphical inspection, it is *not* Normal. Next we repeat our simulation, taking 3000 random samples each consisting of 100 observed survival times (in years) from this population distribution. We calculate the mean survival time for each sample and finally construct a relative frequency distribution of the 3000 means.



**Fig 2.12** Relative frequency distribution of 3000 mean survival times from samples of size 100 drawn from an exponential distribution of survival times in the population

**Example 2.13 continued**

Only the most world-weary of individuals, completely given to a cynical and dissolute existence, would fail to be impressed, startled even, by this truly amazing result. We see that *even when the distribution of the variable of interest in the population is far from Normal, as long as the sample size is large enough, the distribution of sample means will be close to Normal.* (Had I taken a sample size larger than 100 and/or allowed the simulation to run longer and collected more than 3000 samples, Fig 2.12 would have been even closer to Normal.)

The result demonstrated in Examples 2.12 and 2.13 (and which can be proven analytically using the central limit theorem) provides us with an underlying model of immense utility. Here is the rationale:

In medical and health-related research, due to limitations of time, money, patience and patients, we usually just collect data on *one* sample of the population and calculate that sample's mean height, respiratory function, or whatever. The construction of a sampling distribution by endless repeated sampling is *purely notional* (and certainly was worked out long before we could cheat using a computer simulation). The essential things to grasp are:

- the mean you calculate on your single study sample is just one possible value of the mean (since your sample is but one of the innumerable potential samples in the entire population); and
- your mean fits somewhere on the abscissa (the horizontal axis) of a Normal probability curve describing the distribution of all the possible means of samples derived from the underlying population.

If you know, or can somehow estimate, the parameters of this Normal sampling curve, that is, its mean and standard error, then by consulting the Normal curve probability tables you can tell how likely it is that your sample was drawn from the assumed underlying population. Put another way: how usual or unusual would it be to obtain a sample such as yours given the known or assumed characteristics of the population from which the sample allegedly came? The sampling distribution tells you how, given the assumed value of the mean of the population, you might *expect* sample means to “behave”, so you have a basis on which to judge your particular sample. If you come to understand the above, then much of the rest of inferential statistics will fall into place.

It turns out to be quite easy to get the mean and standard deviation (*standard error*) of the sampling distribution of the mean.

Let  $\mu$  and  $\sigma$  be the mean and standard deviation respectively of the underlying population from which you *notionally* took repeated random samples of size  $n$  to form the theoretical sampling distribution of the mean. Then:

$$\text{the mean of the sampling distribution of means} = \mu \qquad \text{E2.13}$$

$$\text{the standard error of the sampling distribution of means} = \sigma/\sqrt{n} \qquad \text{E2.14}$$

The latter is most often referred to by the shorter term “standard error of the mean”. You might note from E2.14 that as the sample size increases, the standard error decreases. As in most of life’s endeavours, so the general wisdom has it, if you put more effort in (take a larger sample), you will reap the reward (less error, that is, greater precision).

**Example 2.12 continued**

Given the parameters of the underlying distribution of weights ( $\mu = 65$  kg,  $\sigma = 10$  kg) and the sample size  $n = 100$ , E2.12 and E2.13 predict that the mean of the sampling distribution of mean weights would be 65 kg (in fact, the simulation delivered a mean of 65.03 kg) and the standard error would be  $10/\sqrt{100} = 1$  kg (the simulation delivered standard error = 0.97 kg).

Finally, you won’t be surprised to learn that the version of the standard Normal transformation used for mapping sample means into z-scores is given by (compare E2.12 & E2.14):

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\text{standard error}} \quad \text{E2.15}$$

Note that the  $\bar{x}$  value in E2.15 is not an individual raw datum but an observed value of the sample mean. E2.15 says: how many standard errors does the distance from the mean of this sample to the mean of the sampling distribution represent?

**§ 2.5.3 Sampling Distribution of a Proportion**

We have seen how the sample mean is distributed. Another common summary statistic is the sample proportion. The distribution of proportions calculated on repeated samples selected from a population wherein the proportion of subjects *with* a characteristic is  $\pi$ , and the proportion *without* is  $1-\pi$  will be developed by way of an example.

First, I state without proof that the sampling distribution is approximately Normal (best for  $0.3 < \pi < 0.7$ , a consequence of trying to approximate a *discrete* Binomial distribution of the number of successes in trials by a *continuous* Normal model) and that the mean and standard error of the sampling distribution are given by E2.16 and E2.17 respectively:

$$\text{the mean of the sampling distribution of proportions} = \pi \quad \text{E2.16}$$

$$\text{the standard error of the sampling distribution} = \sqrt{\frac{\pi(1-\pi)}{n}} \quad \text{E2.17}$$

So, the mapping of a sample proportion,  $p$ , into a z-score is given by:

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

E2.18

**Example 2.14**

You will recall from Example 2.8 that the population proportion of *unvaccinated* nurses who become infected with Hep B after a contaminated needlestick injury was quoted as  $\pi = 0.2$  (20%). Let us suppose that a sample of  $n = 10$  injured nurses is selected from a population of nurses who have been previously *vaccinated* against Hep B. Only one (10%) becomes infected. Is this evidence of a decrease in the rate of infection, that is, evidence of vaccine efficacy? Or is the decrease in infection just a chance finding on this occasion? Let us look at the behaviour of very many samples of size 10 drawn from the *unvaccinated* population, which will be the basis for construction of a model. If it turns out to be unusual to get samples with only 0 or 1 infections from the unvaccinated population, then maybe the vaccination is showing its worth. On the other hand, if it is quite usual for an unvaccinated population to give rise to samples with infection rates of 0.1 or lower, then we would be foolish to make any claims of efficacy for vaccination.

Another simulation is in order. Remember we are constructing a model based on sampling from an *unvaccinated* population where  $\pi = 0.2$ .

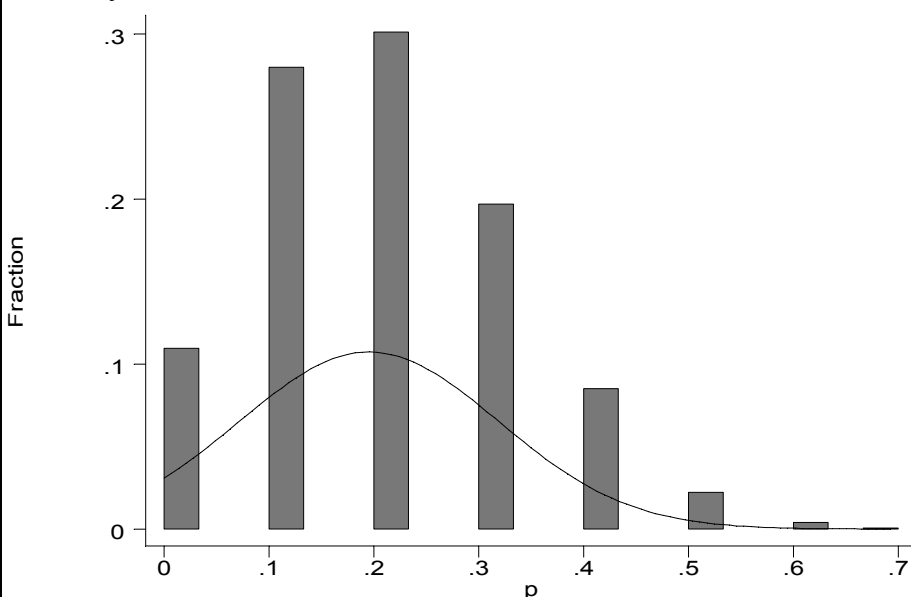
First assume that each needlestick injury represents a Bernoulli trial (you could and should dispute this, but for the moment we will maintain this simplifying assumption). Then the distribution of the number of infections in 10 such trials is Binomial with parameters  $n = 10$  and  $\pi = 0.2$ . I have drawn at random 3000 samples from a simulated population with such a distribution. For each sample the number of “successes” (Hep B infections) was recorded. Dividing each by 10 gave the sample proportion of success. Here is a listing of the results of the first 7 of the 3000 samples just to get you orientated:

sample number	number of infections	proportion infected
1	2	0.2
2	4	0.4
3	0	0.0
4	3	0.3
5	5	0.5
6	2	0.2
7	2	0.2
etc	etc	etc

**Example 2.14 continued**

The mean of these 3000 proportions (see E1.2) was 0.1954, the standard deviation – *standard error* since we are dealing with a sampling distribution – was 0.124 (see E1.3 and E1.4). These values are quite close to the mean and standard error expected from the theoretical Normal sampling distribution: using E2.16 and E2.17, the expected mean is 0.2 and the standard deviation is  $\sqrt{[0.2 \times 0.8 / 10]}$  or 0.126. So far, so good.

But the graph of the distribution of sample proportions from my simulation (Fig 2.13) does not give me confidence that the use of a Normal model will be valid. In fact, since theory suggests that a Normal model is a good approximation for the sampling distribution for  $0.3 < \pi < 0.7$ , my problem involving  $\pi = 0.2$  was always going to be a bit tricky.



**Fig 2.13** Distribution of proportions from samples  $n = 10$  from a population with  $\pi = 0.2$  (with superimposed Normal curve)

What can I do? To solve my problem, a Normal model does not appear useful. From Fig 2.13, I see that any probabilities I calculate under the Normal curve would not correspond very well with those from the empirical distribution of the simulation. I have several alternatives.

The first is to abandon a theoretical (Normal) model and just use the empirical relative frequency distribution of the simulation. How often did my simulation of an *unvaccinated* population generate a sample with less than the 2 expected infections? In fact, 39% of the samples had 0 or 1 infections. You can see this by adding up the heights of the two leftmost columns in Fig 2.13. An event that occurs on 39% of occasions is hardly an unusual event. On this basis, we would doubt that the low infection rate in the observed sample of vaccinated nurses had anything to do with their vaccination status.

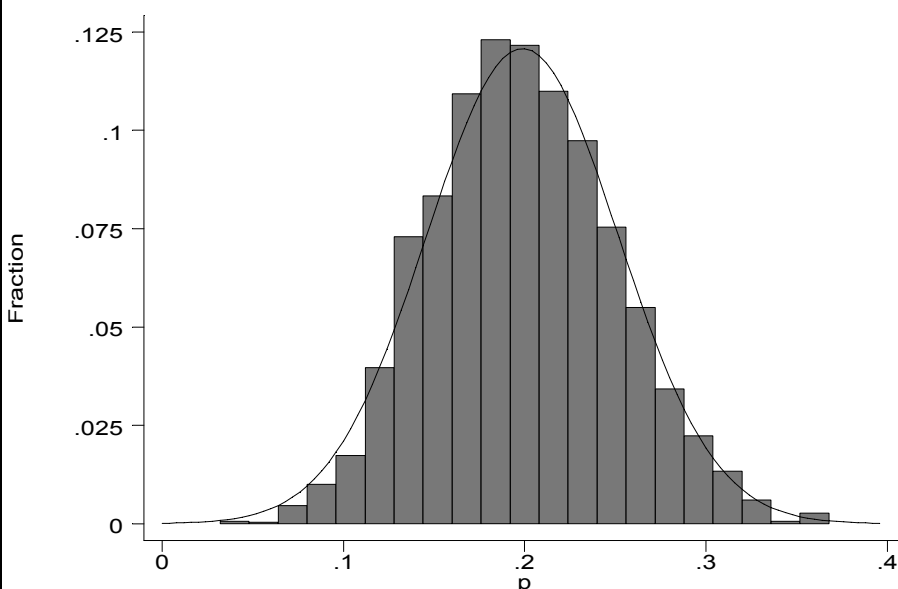
**Example 2.14 continued**

(Also, if you were really keen you could use E2.18 with  $n = 10$ ,  $p = 0.1$  and  $\pi = 0.2$  and then the Tables of the Standard Normal Distribution, to show that the Normal model would give a lower tail probability of about 0.21, quite different from 0.39, confirming the inadequacy of the Normal model.)

There are at least two major problems remaining if we just use the simulated empirical data. One is that we do not have a measure of the standard error, which eventually (see Chapter 3) we may want to make statements regarding precision. To get this we still need a nice model. The other is that we may not always be in a position to use a simulation. A well-behaved theoretical model can be quite handy.

One approach is to try a *transformation* of the sample proportion (see more advanced texts) in the hope that such a manoeuvre will yield a more valid Normal approximation.

Another is to (notionally) go back to the population and obtain repeated *samples of a larger size*. Theory (central limit theorem again) suggests that, even for values of  $\pi$  more extreme than 0.3 or 0.7, this will help. To demonstrate this, another simulation! Let's try a sample size of, say,  $n = 60$ . (I tried  $n = 25$  and again with  $n = 40$  which were both improvements over  $n = 10$ , but still not good enough.) Fig 2.14 shows how a Normal model now appears to be reasonable for describing the distribution of proportions in samples of size 60 taken from a population with infection rate of 0.2. I note that my simulation gave a mean sample proportion of 0.199 (*cf* 0.2 theoretically from E2.16) and a standard error of 0.053 (*cf* 0.052 from E2.17).



**Fig 2.14** Distribution of proportions from samples  $n = 60$  from a population with  $\pi = 0.2$  (with superimposed Normal curve)

**Example 2.14 continued**

So, let's say we did have the luxury of being able to redesign the study and choose a sample of 60 vaccinated nurses. Say 6 (10%, as before) became infected, whereas we would have expected 12 (20%) to have been infected in the unvaccinated population. To see if there is a decrease in the proportion – other than just a “chance” finding on our particular sample – we need to see how often (in repeated sampling) a sample proportion of 0.1 or less might arise when the true proportion in the unvaccinated population is 0.2. If this is a frequent occurrence – associated with a large probability – then evidently we are not to think it is unusual that even an unvaccinated population gives rise to samples that have only half the infection rate expected. In other words, vaccination may be doing nothing. On the other hand, if such an occurrence is infrequent – small probability – then we have evidence that it would be unusual for such a sample to be generated from an unvaccinated population, and that maybe vaccination is affording protection.

Since we have an easy-to-deal-with Normal model describing samples of size 60 from the population, it is simple to calculate the probability. We seek the area in the tail of the curve below  $p = 0.1$ . Given the mean and standard error of the Normal model, we calculate that:

$$z = (0.1 - 0.2)/0.052 = -1.923 \qquad \text{from E2.18}$$

The probability (see Appendix Table 1) associated with this z-score is 0.027. My simulation gave a probability of 0.032, which is quite close. That is to say, 27 times in 1000, in samples of size 60 from an *unvaccinated* population of clinic nurses we would get a proportion infected with Hep B after needlestick injury of 0.1 or less. It is now up to you to judge if the observed infected proportion of 0.1 in the *vaccinated* nurses convinces you that vaccination was worthwhile. I think that .027 or 2.7% represents an unusual event, so we would rarely expect an unvaccinated population to yield such a low rate. Vaccination is starting to look good.

Note that the final result in Example 2.14 is quite different from that obtained in the first simulation using sample sizes of  $n = 10$ . The larger sample size of  $n = 60$  gives us *more information* upon which to base a *better decision*. This is a fundamental concept in sampling. The down side, of course, is that it costs more, both financially and logistically, to take a larger sample for a study. These factors must be weighed up when a study is being designed.

### § 2.5.4 Further examples

Here are some more examples using the results of §2.5.2.

#### Example 2.15

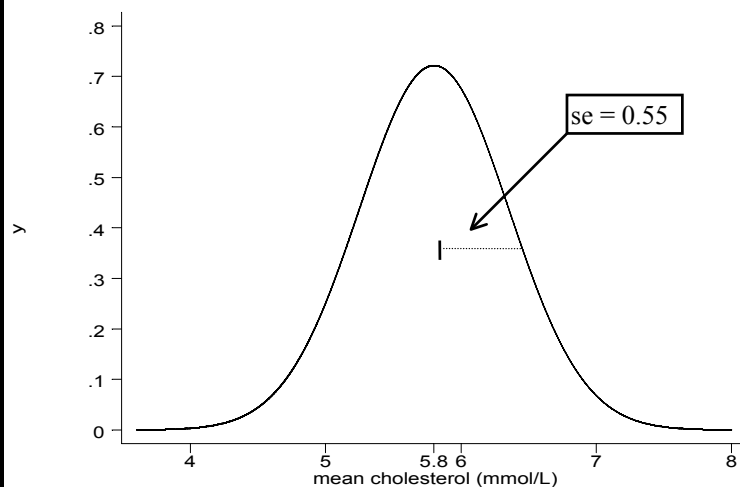
The overall population plasma cholesterol in Australian adult males aged between 45 and 50 is Normally distributed with known population mean  $\mu = 5.8$  mmol/litre and standard deviation  $\sigma = 2.65$  mmol/litre. A cardiologist looks at the medical records of 23 adult males who have each suffered a heart attack within the preceding 12 months. The mean plasma cholesterol in this sample (measured prior to their heart attacks), was found to be 7.0 mmol/litre.

What is the probability of getting such a result?

*Solution:*

Draw a diagram of the notional sampling distribution of the mean. From E2.13 and E2.14 we know this distribution has a mean of 5.8 mmol/litre and a standard error of  $2.65/\sqrt{23} = 0.55$  mmol/litre.

Mark in on the x-axis the sample mean that the doctor found, 7.0 mmol/L.



**Fig 2.15** Sampling distribution of the mean for  $\mu=5.8$ ,  $\sigma=2.65$ ,  $n=23$

Now the z-score corresponding to mean cholesterol of 7.0 mmol/L is:

$$z = (7.0 - 5.8)/(2.65/\sqrt{23}) = 2.17$$

That is, the distance of 7 from 5.8 is 2.17 standard errors. And the probability of getting such a score or greater – a so called upper **one-tailed** probability – is  $P(z > 2.17)$ .

**Example 2.15 continued**

From Table 1 in the Appendix this is 0.015 or 1.5%. So, the chances of getting a mean of 7.0 mmol/L or greater in a sample of size 23 from a population with mean 5.8 mmol/L is only 0.015 or 1.5%. This is a pretty small chance. It makes you wonder if the doctor's sample actually comes, not from the overall Oz male population, but from another differently defined population – one with different parameters – perhaps one with a higher mean cholesterol.

In Example 2.15, there is no completely right or wrong interpretation. It is still *possible* that the doctor's patients were drawn from the general population, but, *on the basis of probabilities*, it would be unwise to claim this. As the previous discussions on errors and variability suggest, decisions are made using *available evidence*, in the presence of *ambiguity*, and sometimes will be *wrong*.

Sometimes problems are phrased such that the probability required is a *two-tailed* probability. In this case we are interested in the chance of getting a result as *extreme* as the one our sample provides, in *either* tail of the distribution.

**Example 2.16**

A researcher is interested in calibrating a device to measure blood glucose prior to initiating a screening programme for diabetes in an at-risk section of the population. The manufacturer says that, using a standard solution of glucose, the mean reading on repeated tests should be 5.0 mmol/L with standard deviation 1.8 mmol/L. The researcher gets a mean reading of 7.0 mmol/L after testing 3 times.

What is the probability of getting a result as extreme as this: 2 mmol/L off the alleged true mean?

*Solution:*

Now the researcher would be just as, if not more, worried about getting a mean reading that was 2.0 mmol/L *below* the claimed population mean of 5.0 mmol/L (at 3.0 mmol/L) as she would about getting a reading 2.0 mmol/L *above* (at 7.0 mmol/L). The former situation would lead to an increase in false negatives on the screen (falsely low blood glucose recorded in a subject with diabetes), the latter would lead to an increase in false positives (falsely high glucose in subjects without disease). To take account of the possibilities of both extremes she would need to calculate the two-tailed probability – the sum of the shaded regions in the diagram.

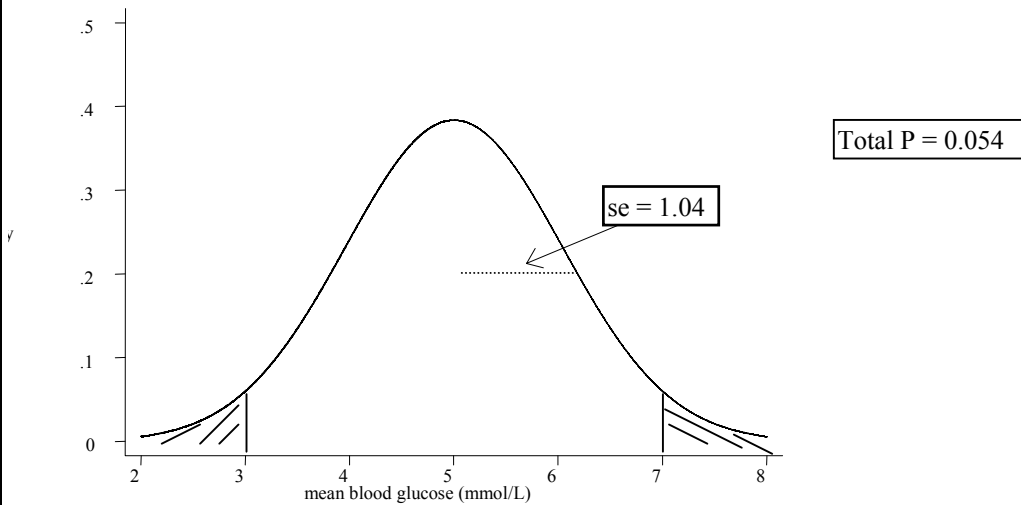
Some tables give two-tailed probabilities straight off (see Appendix Table 2), or, since the curve is symmetrical, we can just *double* either of the one-tailed probabilities. Doing this (arbitrarily choosing the low tail), and using Appendix Table 1, we seek:

$$\begin{aligned}
 & 2 \times P(\bar{x} \leq 3) \\
 = & 2 \times P(z \leq [3-5]/[1.8/\sqrt{3}]) \quad \text{(using E2.15)}
 \end{aligned}$$

**Example 2.16 continued**

$$= 2 \times P(z \leq -1.925)$$

$$= 2 \times 0.027 = 0.054 \text{ or } 5.4\% \quad (\text{rather small})$$



So, if you consider 0.054 to be a rather low probability, something *may* be amiss. Either the manufacturer's standard solution did not meet its stated specifications (that is, the population parameters were not as stated), or the researcher's machine, or her technique, is a bit suspect. *If the mean were truly 5.0 mmol/L then chance sampling variation leading to a 2.0 mmol/L or greater discrepancy should only occur on about 5% of occasions.* We would proceed to check out each possibility.

**§ 2.6 SOME HANDY HINTS**

- Always draw diagrams.
- z-scores and x-scores can be less than zero, probabilities can't be. If you come up with a probability less than zero or greater than one, you've done something wrong.
- Think about a problem carefully to see if it involves a sampling distribution (as in Example 2.13) or a population distribution problem (compare Example 2.10).
- If in doubt, calculate two-tailed probabilities.
- Always remember that the probabilities upon which you base your decisions are themselves based on assumptions regarding the true nature of the underlying distribution (for example, is it Normal?) and the validity of the sampling procedure that generated the data.

## § 2.7 SUMMARY

Probability theory provides models of reference for our research endeavours. To this end we have developed the notion of sampling distributions, in particular, that of the sample mean and the sample proportion. It is important to understand that we could (almost) as easily have chosen to use other sample statistics as the basis of this discussion; perhaps the difference in proportions between two groups, or, say, one of the common measures of disease risk such as the relative risk or the odds ratio. All these statistics have sampling distributions (not necessarily Normal) which enable us to make judgements about the *particular* sample we may be dealing with. As we shall see in more detail in Chapter 3, decisions based on these models are subject to error, but at least these errors are quantifiable.

There are no statistical free lunches. We only create a *model* because we deem the *reality* it represents to be too complex for our purposes. As the statistician G.E.P Box remarked: “*All models are wrong, some are useful*”. A model is often used to provide the framework for answering a question in scientific research. An appropriate model can only be chosen, and the answer it provides only assessed, if full consideration is given to the design and conduct of the study, and to the myriad of ways in which bias may have been introduced.

