The Concept of a Statistic

Study guide and advice

To understand the concepts explained in this chapter you require an understanding of the distinction between a permutation (an ordered set of items in a list) and a combination (an unordered set of items in a list). It is suggested that the (optional) chapter entitled *Permutations and Combinations* is studied in advance of this one. However, examination style problems at this level can be solved by means of a probability tree or equivalent, thus not actually requiring drill in the use of permutations and combinations. So you can, if you wish, read the examples here as worked examples, which contain full explanations of calculations of permutations where required, and tackle the exam-style questions using probability trees.

Prerequisites

You should be familiar with discrete probability distributions and how to calculate these using probability trees.

Example (1)

Two bags contain a mixture of balls. In the first bag there are three balls having the numbers 0, 1 and 2; in the second bag there are five balls, numbered 1, 1, 1, 2 and 3. One ball is drawn from the first bag, and another ball from the second.

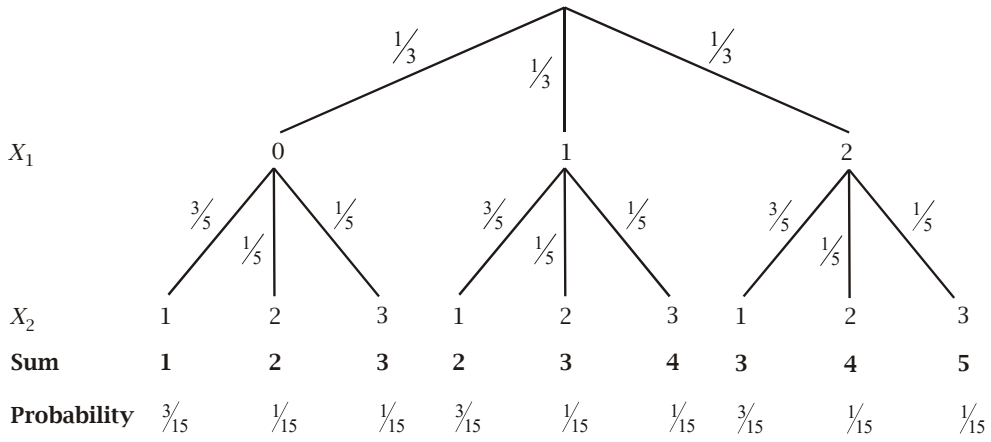
(*a*) Write down the discrete probability distribution of the sum Y of the two values.

(*b*) Find the expectation and variance of Y.

(*c*) List all the possible permutations of the two balls together with their sum.

Solution

Let X_1 represent the discrete random variable for the result of drawing a ball from the first bag. Let X_2 represent the discrete random variable for the result of drawing a ball from the second bag. Let $Z = X_1 + X_2$ represent the discrete random variable for the sum of the two numbers of the balls. Then the entire probability tree for the experiment is as follows.



(*a*) From which we can derive the following probability distribution for X

z	1	2	3	4	5
$P(Z=z)$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{5}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

(*b*) $E(Z)=\text{Sum}\left(z\times P(Z=z)\right)$ $\left[E(Z)=\sum z\times P(Z=z)\right]$

$$=\left(1\times\frac{3}{15}\right)+\left(2\times\frac{4}{15}\right)+\left(3\times\frac{5}{15}\right)+\left(4\times\frac{2}{15}\right)+\left(5\times\frac{1}{15}\right)$$

$$=\frac{39}{15}$$

$E\left(Z^2\right)=\text{Sum}\left(z^2\times P(Z=z)\right)$ $\left[E\left(Z^2\right)=\sum z^2\times P(Z=z)\right]$

$$=\left(1^2\times\frac{3}{15}\right)+\left(2^2\times\frac{4}{15}\right)+\left(3^2\times\frac{5}{15}\right)+\left(4^2\times\frac{2}{15}\right)+\left(5^2\times\frac{1}{15}\right)$$

$$=\frac{121}{15}$$

$$\text{var}(Z)=E\left(Z^2\right)-\left[E(Z)\right]^2$$

$$=\frac{121}{15}-\left(\frac{39}{15}\right)^2$$

$$=\frac{294}{225}$$

(*c*) To answer this question we observe first that in the second bag there are three balls carrying the same number. These are "identical" in the sense that they all carry the same number, but "not identical" in that they are actually different balls. To show this we shall mark each of these balls to show that they are different balls but bearing the same number.

1 1* 1**

Then every possible permutation in the experiment is given in the following table.

X_1	X_2	Z
0	1	1
0	1*	1
0	1**	1
0	2	2
0	3	3

X_1	X_2	Z
1	1	2
1	1*	2
1	1**	2
1	2	3
1	3	4

X_1	X_2	Z
2	1	3
2	1*	3
2	1**	3
2	2	4
2	3	5

There are 15 possible permutations, each one being a possible outcome of the experiment, each having a probability of $\frac{1}{15}$. This table comprises the entire sample space (possibility space) for the experiment. From this we could also deduce the probability distribution of $Z = X_1 + X_2$ given in part (*a*). For example

$$P(Z=1) = \frac{\text{number of outcomes (permutations) in which } Z=1 \text{ occurs}}{\text{total number of outcomes (permutations)}} = \frac{3}{15} = \frac{1}{5}$$

Sampling distributions

In example (1) a single observation is drawn from two different bags. The discrete random variables X_1 and X_2 represent distinct probability distributions. The random variable $Z = X_1 + X_2$ is a linear combination of both of these. Suppose on the other hand there was only one bag, and two balls were taken from this same bag, with or without replacement. In that case we would be *sampling* the same bag twice. We would have a *sample* made of two *observations* of the same *population*. Denote by X_1 the result of the first observation, and by X_2 the result of the second observation and by $Z = X_1 + X_2$ the sum of the two observations. We will assume also that X_1 and X_2 are random observations and that at in each of the two trials the probability of any single outcome is equally likely.

Example (2)

One bag contains a mixture of five balls bearing the numbers 1, 1, 1, 2 and 3. Two balls are drawn from the same bag (*i*) with replacement, (*ii*) without replacement. Let Z denote the sum of the two numbers on the balls. For both cases

(*a*) List all the possible permutations.

(*b*) Define a *sample* to be any combination of the same two numbers. Find the number of permutations corresponding to each sample, and determine the probability of each sample.

(*c*) Draw the probability tree corresponding to the sum Z.

(*d*) Determine the distribution of Z.

(*e*) Are the two observations that comprise the sum Z independent?

Solution

Let X_1 represent the discrete random variable for the first observation of the bag. Let X_2 represent the discrete random variable for the second observation. Let $Z = X_1 + X_2$ represent the discrete random variable for the sum of the two numbers of the balls.

(*i*) With replacement

(*a*) As before we shall label the different balls bearing the same number.

1 1* 1**

In the problem with replacement any ball may be chosen twice. The 25 permutations (outcomes) are as follows.

X_1	X_2	Z
1	1	2
1	1*	2
1	1**	2
1	2	3
1	3	4

X_1	X_2	Z
1*	1	2
1*	1*	2
1*	1**	2
1*	2	3
1*	3	4

X_1	X_2	Z
1**	1	2
1**	1*	2
1**	1**	2
1**	2	3
1**	3	4

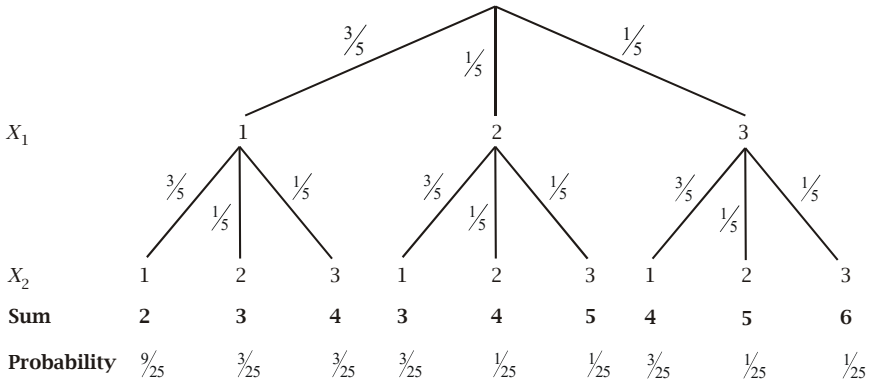
X_1	X_2	Z
2	1	3
2	1*	3
2	1**	3
2	2	4
2	3	5

X_1	X_2	Z
3	1	4
3	1*	4
3	1**	4
3	2	5
3	3	6

(*b*)

Sample	{1,1}	{1,2}	{1,3}	{2,2}	{2,3}	{3,3}
Event	$Z=2$	$Z=3$	$Z=4$	$Z=4$	$Z=5$	$Z=6$
No. of permutations	9	6	6	1	2	1
Probability	$\frac{9}{25}$	$\frac{6}{25}$	$\frac{6}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

(*c*) The probability tree is



(*d*) The distribution of Z is

Event, $Z = z$	$Z = 2$	$Z = 3$	$Z = 4$	$Z = 5$	$Z = 6$
$P(Z = z)$	$\frac{9}{25}$	$\frac{6}{25}$	$\frac{7}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

The approaches based on (1) finding all the samples (permutations) and (2) the probability tree are equivalent, as we would expect.

(*e*) The two observations are taken *with replacement* and are independent. As the probability tree illustrates the probabilities for the second branches are the same as those for the first and the probability that X_2 takes a value is not affected by the outcome of X_1 and vice-versa.

(*ii*) Without replacement

(*a*) The 20 permutations (outcomes) are as follows.

X_1	X_2	Z
1	1*	2
1	1**	2
1	2	3
1	3	4

X_1	X_2	Z
1*	1	2
1*	1**	2
1*	2	3
1*	3	4

X_1	X_2	Z
1**	1	2
1**	1*	2
1**	2	3
1**	3	4

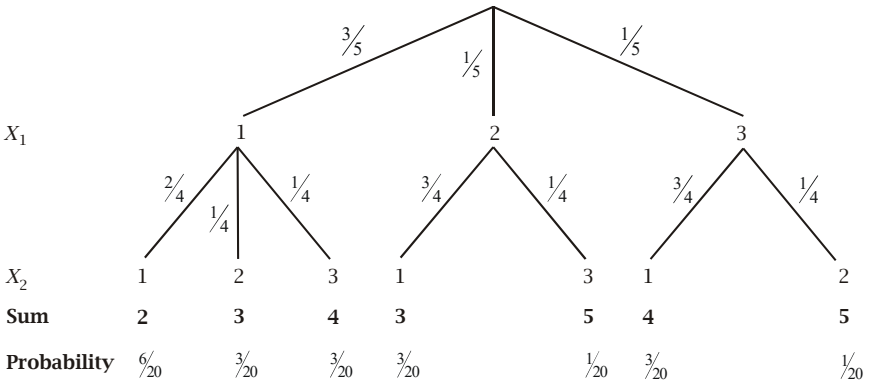
X_1	X_2	Z
2	1	3
2	1*	3
2	1**	3
2	3	5

X_1	X_2	Z
3	1	4
3	1*	4
3	1**	4
3	2	5

(*b*)

Sample	$\{1,1\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
Event	$Z=2$	$Z=3$	$Z=4$	$Z=5$
No. of permutations	6	6	6	2
Probability	$\frac{6}{20}$	$\frac{6}{20}$	$\frac{6}{20}$	$\frac{2}{20}$

(*c*) The probability tree is



(*c*) The distribution of Z is

Event, $Z=z$	$Z=2$	$Z=3$	$Z=4$	$Z=5$
$P(Z=z)$	$\frac{6}{20}$	$\frac{6}{20}$	$\frac{6}{20}$	$\frac{2}{20}$

(*d*) X_2 is **not** independent of X_1. Recall that for independent events

$P(A \textbf{ and } B)=P(A)\times P(B)$ or $P(A\cap B)=P(A)\times P(B)$

For example, let A denote the event that X_1 takes the value 1 and let B denote the event that X_2 takes the value 1. Then $P(A)=\frac{3}{5}$, $P(B)=\frac{3}{5}$ and $P(A\cap B)=\frac{3}{10}$ but $P(A)\times P(B)=\frac{3}{5}\times\frac{3}{5}=\frac{9}{25}$ so $P(A\cap B)\neq P(A)\times P(B)$. Alternatively, using the approach based on conditional probability, if A and B are independent events, then $P(B|A)=P(B)$. However, $P(B|A)=\frac{1}{2}\neq P(B)=\frac{3}{5}$.

Let us summarise what is taking place in this example. Instead of drawing single observations from two or more populations and combining them in some way, we are making one or more observations from a **single** population. The result of these observations is called a sample. A single observation of one population is a sample of size 1. If we draw 2 observations from a single population then the sample size is 2, and so forth. We remind you of the definition of a discrete random variable.

Formal definition of a discrete random variable

Let X be a variable such that

(*a*) It is discrete, meaning it can only take n exact values $x_1, x_2, \ldots, x_n$. When X takes the value x_i we write $X = x_i$.

(*b*) It is random, meaning that with each value that the variable takes, there is associated a probability p. We write this $P(X = x) = p$, which is read, "The probability that the random variable X takes the value x is p".

(*c*) Because it is random it obeys the law of total probability. The sum of all the probabilities for all n values is equal to 1.

By this definition a single observation drawn from the same population is a random variable. When we take a sample of size greater than 1, we must then distinguish between several concepts that we have already denoted by different symbols, where appropriate, in our worked example (2).

X Let this denote the random variable for a population as a whole (or denotes the *population* for short).

X_i Let this denote the *i*th observation of X. This is also a random variable. If the samples are taken *with replacement*, each X_i has the same distribution as the *parent population* X. In problems *without replacement* the distribution of X_2 will not be the same as that of X_1 and so forth.

x [Lower case letter.] Let this represent a particular value that the population X or observation X_i can take.

Y Let this represent a *sample*, which comprises n observations of a random variable X..

$Y =$ a combination of values of $X_1, X_2, \ldots, X_n = \{X_1, X_2, \ldots, X_n\}$

Y follows a *sample distribution* that is derived from the probability distribution of each observation. By the definition above it is also a discrete random variable.

y [Lower case letter.] Let this represent a particular value that the sample Y can take. Each value of y is a particular unordered set of observations.

$y = \{x_1, x_2, \ldots, x_n\}$ where x_i is the particular value of X_i.

Each value y is a distinct combination of the values that each X_i can take and is itself a set of values.

Z Let this represent a *function* of the n observations of the random variable X. Any function satisfying certain rules (that we specify below) is called a *statistic*. For example the sum of the n observations

$$Z = X_1 + X_2 + ... + X_n$$

is a *statistic*. Z is also a discrete random variable and like Y (to which it is closely related) it has a probability distribution based on the probability distributions of each observation.

z [Lower case letter.] Let this be a particular value that the *statistic* Z can take. If, for example, Z is the statistic

$$Z = X_1 + X_2 + ... + X_n$$

then z is the value

$$z = x_1 + x_2 + ... + x_n$$

where x_i is the particular value of the ith observation of X.

Remarks

(1) In some contexts either the *sample* or the *statistic* may also be denoted by X where $X_1, X_2, ..., X_n$ are n observations of a single population. Here we have tried to avoid confusing (1) the population X from (2) the sample Y drawn from n observations of X and both of these from (3) the statistic Z that is a function of the n observations in the sample Y. We do this by using different symbols for each concept. However, in context any symbol may be used for any of these concepts and you should be aware of this.

(2) In example (2) we considered the case where a sample Y is drawn from just 2 observations and the statistic Z is the sum of the two observations. Our definitions and symbols above apply to samples of any size n. Below we will consider an example where the sample size is greater than 2.

(3) In the case where the population X is sampled *without replacement* the probability distributions of each $X_1, X_2, ..., X_n$ are different. That is to say X_1 is **not** independent of X_2 and so forth. However, in many contexts it is a requirement that each observation $X_1, X_2, ..., X_n$ in the sample **is independent**. In that case the sample

$$Y = \{X_1, X_2, ..., X_n\}$$

is called a *simple random sample*. Thus we use the term *random sample* for a variable that may be drawn from n observations of a population X, whether independent or not, and *simple random sample* where it is specific that each observation in the sample is independent of each other.

(4) Throughout the above we have specified that the population is a *discrete* random variable. This restriction is unnecessary. The ideas introduced in this chapter apply equally to discrete and to continuous random variables making appropriate adjustments where necessary (i.e. probabilities defined for intervals rather than for discrete values). However, in this chapter our examples shall all be drawn from discrete probability distributions and we shall reserve the full discussion of the continuous case to a later chapter.

Example (3)

Five cards are numbered 1, 1, 1, 2, 3 respectively.

(*a*) Let X represent the discrete variable for the number on the card when any card is drawn at random Determine the probability distribution of X.

(*b*) This population is to be observed at random three times. Let X_1, X_2, X_3 denote the three observations in the sample. If the observations are made with replacement determine the probability distribution of each of X_1, X_2, X_3.

(*c*) This population is observed at random three times without replacement. Let Y denote the resultant sample. List all the possible samples of Y and identify the number of permutations of the observations X_1, X_2, X_3 that correspond to each sample.

(*d*) Let Z be the statistic given by

$Z = X_1 + X_2 + X_3$

Find the probability distribution of Z.

(*e*) Let $\bar{X}$ be the statistic representing the sample mean. Find the probability distribution of $\bar{X}$.

(*f*) Let M be the statistic representing the sample median. Find the probability distribution of M.

Solution

(*a*)

Event, $X = x$	$X = 1$	$X = 2$	$X = 3$
$P(X = x)$	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

(*b*) The probability distribution of each observation X_1, X_2, X_3 is the same as that of the *parent population X*.. Hence for $i = 1,2,3$

Event, $X_i = x$	$X_i = 1$	$X_i = 2$	$X_i = 3$
$P(X_i = x)$	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

(*c*) This question requires one to determine the value of certain permutations. Let us explain how this is done first.

1. Denote by ${}^3P_2$ the permutations of 2 objects taken from a choice of 3. We have three cards with the number 1 on them. Label these 1, 1* and 1**. Suppose we have to choose 2 of these. Then by the above this is the number ${}^3P_2$. To find this number, note that there are 3 ways to choose the first card; when that card has been chosen, there are 2 ways to choose the second card. Hence ${}^3P_2 = 3 \times 2 = 6$.

2. There are 3 permutations of 1, 1, 2. These are (1, 1, 2), (1, 2, 1) and (2, 1, 1).

3. There are 6 permutations of 1, 2, 3. The first number can be chosen in 3 ways; the second number can be chosen in 2 ways; then the third number can be chosen in only 1 way. So there are $3 \times 2 \times 1 = 6$ permutations. The permutations are specifically (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2) and (3, 2, 1).

4. The number of permutations n corresponding to 1, 1, 2 in any order is

$$\begin{aligned} n &= \text{permutations of 1, 1, 2} \times \text{ways of choosing two 1s from three} \\ &= 3 \times {}^3P_2 \\ &= 3 \times 6 \\ &= 18 \end{aligned}$$

This should help you to understand the following table. In this experiment there are 60 permutations (outcomes) in all.

Sample	No of permutations	Event
$\{1,1,1\}$	Permutations of $1,1^*,1^{**} = 3! = 6$	$Z = 3$
$\{1,1,2\}$	${}^3P_2 \times$ permuations of $1,1,2 = 6 \times 3 = 18$	$Z = 4$
$\{1,1,3\}$	${}^3P_2 \times$ permuations of $1,1,3 = 6 \times 3 = 18$	$Z = 5$
$\{1,2,3\}$	${}^3P_1 \times$ permuations of $1,2,3 = 3 \times 3! = 18$	$Z = 6$

(*d*) $Z = X_1 + X_2 + X_3$

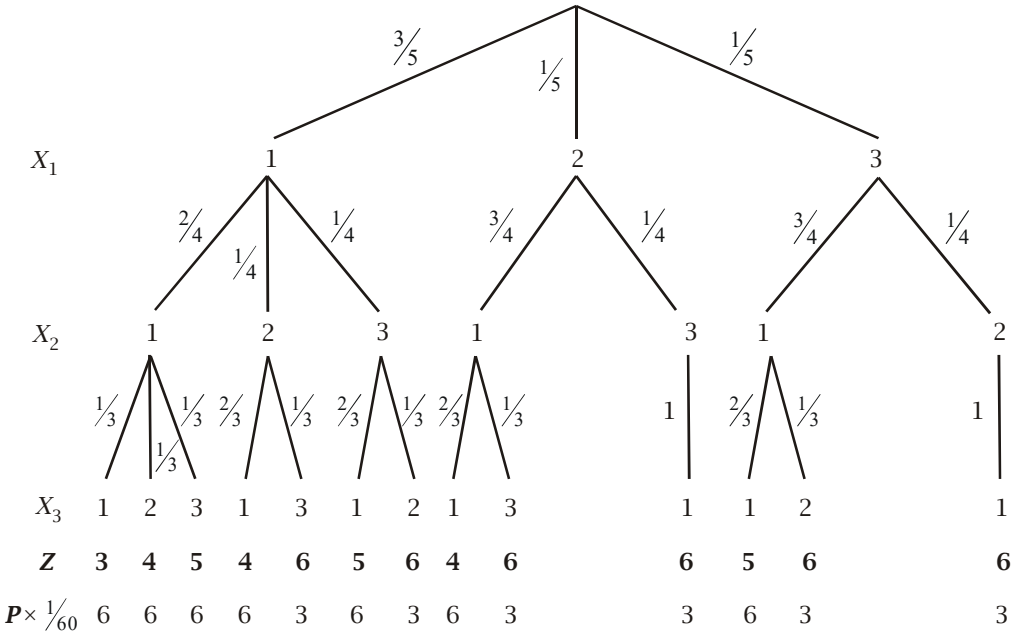
Using the definition of probability as

$$P(A) = \frac{\text{number of outcomes (permutations) in which } A \text{ occurs}}{\text{total number of outcomes (permutations)}}$$

and the table of the numbers of permutations given in part (*c*) we have

Event, $Z = z$	$Z = 3$	$Z = 4$	$Z = 5$	$Z = 6$
$P(Z = z)$	$\frac{6}{60} = 0.1$	$\frac{18}{60} = 0.3$	$\frac{18}{60} = 0.3$	$\frac{18}{60} = 0.3$

If the approach using permutations were not followed, then the same table could be deduced from the probability tree for the experiment.



(*e*) $\bar{X}$ represents the sample mean. The probability distribution of $\bar{X}$ is

Sample	$\{1,1,1\}$	$\{1,1,2\}$	$\{1,1,3\}$	$\{1,2,3\}$
Event, $\bar{X} = \bar{x}$	$\bar{X} = 1$	$\bar{X} = \frac{4}{3}$	$\bar{X} = \frac{5}{3}$	$\bar{X} = 2$
$P(\bar{X} = \bar{x})$	0.1	0.3	0.3	0.3

(*f*) *M* represents the sample median. We have

Sample	$\{1,1,1\}$	$\{1,1,2\}$	$\{1,1,3\}$	$\{1,2,3\}$
Median $M = m$	$m = 1$	$m = 1$	$m = 1$	$m = 2$
$P(M = m)$	0.1	0.3	0.3	0.3

Hence the probability distribution for M is

Median $M = m$	$m = 1$	$m = 2$
$P(M = m)$	0.7	0.3

Statistics and parameters

The conceptual progress of this chapter may be summarised by the following diagram.

population (X) $\longrightarrow$ observation (X_i) $\longrightarrow$ sample (Y) $\longrightarrow$ statistic (Z)

We can now formally define a statistic

Definition of a statistic

If $X_1, X_2, X_3, \ldots X_n$ is a random sample of size n from some population then a statistic Z is a random variable consisting of any function of the X_i that involves no other quantities.

Thus, for example $Z = X_1 + X_2$ is a statistic. It will take the values defined by $y = y_{i,j} = x_i + x_j$. The values are all the possible values of the two observations, X_1 and X_2, in every possible combination.

When Z is also based on a *simple random sample* so that each observation is independent of the other, we can prove directly that Z is a random variable. Since the observations X_1 and X_2 are independent of each other, then

$$P(Z = z) = P(X = x_1 \text{ and } X_2 = x_2) = P(X = x_1) \times P(X = x_2)$$

To find the sum of all the probabilities of the values that Y can take

$$\begin{aligned}\sum P(Z = z) &= \sum P(X_1 = x_i) \times \sum P(X_2 = x_j) \\ &= 1 \times 1 \\ &= 1\end{aligned}$$

This shows that Z is a random variable.

Recognising what is a statistic

At this point we could benefit from some "rules" for recognising which functions are statistics. Firstly, we will assume that X_1 and X_2 are observations of a random variable X. Then

(1) A linear combination of X_1 and X_2 is also a random variable. For example $X_1 + X_2$ and $X_1 - X_2$ are random variables.

(2) Translation. The addition or subtraction of a scalar quantity will not prevent the variable from being a random variable. If X is a random variable then $X + a$ is also a random variable, where a is a scalar.

(3) Scaling. The multiplication of random variables by scalar quantities will not prevent the resultant variable from being a random variable. If X is a random variable then aX is a random variable.

We can summarise all of this by saying that statistics are *linear combinations* of observations of a random variable.

Parameters

Examine again the following diagram of concepts.

$$\text{population } (X) \longrightarrow \text{observation } (X_i) \longrightarrow \text{sample } (Y) \longrightarrow \text{statistic } (Z)$$

We see that a statistic is a property of a sample computed according to some rule. Let us use the term *parameter* to represent a property of a population. Examples of parameters are the true mean, median and variance of the population. We use the symbols μ and σ^2 to represent the true population mean and variance respectively. We see automatically from the diagram above that a statistic cannot be a function of any population parameter, for that would be circular. The statistic must be a function of the sample and not of the population.

Rule for what is definitely not a statistic

A statistic cannot include terms representing **unknown** population parameters, such as the true population mean, μ, and the true population variance, σ^2.

Thus, for example, the sample mean and sample standard deviation **are statistics**

$$\bar{x} = \frac{\Sigma x}{n}$$

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n}$$

Note, these are statistics computed from the sample. They are **not** the true population mean and variance. Let the true population mean, μ, the true population variance, σ^2, be unknown. Then the following are not statistics.

$$\sum X - \mu$$

$$\sum \frac{x - \mu}{\sigma}$$

These are expressions involving unknown population parameters, which are not drawn from the sample.

Estimation

The fundamental reason for distinguishing between parameters (properties of the population) and statistics (properties of a sample) is that statistics are used to *estimate* population parameters. In all the examples discussed in this chapter the probability distribution of the population X has been given. Since it is given we can determine precisely such population parameters as the mean and variance of X. However, the population distributions here were based on theoretical distributions, and in particular on the assumption of classical probability that all outcomes are equally likely. The practical use of the theory of statistics is in order to *estimate* population parameters where the population parameters are **unknown**. For example, pollsters wish to forecast which party will win an election. The proportion of people who will vote for a certain party is a population parameter and until the election is actually conducted it is unknown. In advance of the election pollsters predict the true value of this parameter based on samples of what voters say they will do on the day of the election. A statistic based on the sample is used to estimate the true population proportion, which is a parameter. When a statistic is used to estimate a parameter it is said to be an *estimator*. This serves as an introduction to the theory of estimation, which is taken up in subsequent chapters.

Remark

We take up the theory of estimation in subsequent chapters and advise you at this stage not to make any assumptions about which statistics may be used to estimate particular population parameters. In particular, do not use the sample variance to estimate the population variance. This is an introduction only to the theory of estimation.

Summary

A *statistic* is drawn from a *sample*.
A *parameter* is a property of a *population*.
Statistics are used to estimate *parameters*.