4 Probability and Sampling Distributions

The reasoning of statistical inference is based on asking "How often would this method give me the right answer?" Inference is most secure when we produce data by random sampling or randomized comparative designs. When we use chance, the laws of probability answer the above question.

4.1 Randomness

Since we collect data from a sample but we would like to make inferences about the population we would like to differentiate between values that represent the sample and values that represent the population. A **parameter** is a value that represents the population. It is usually unknown but is the value we are interested in. The **statistic** is a number computed from the sample without any unknown parameters and usually is used to estimate or make a guess about the paramter.

Parameters are usually represented by the greek letters. The mean of the population is represented by μ and is different from the mean of the sample which we refer to by $\bar{x}$. As an example, if the mean content of 6 cans of a soda is 298.5 ml as opposed to being what we expect (which is 300 ml), the number 298.5 corresponds to $\bar{x}$, the mean of our sample and the number 300 corresponds to μ, the overall or population mean.

Another parameter is σ, which is the standard deviation of the population and its corresponding statistic is s, the standard deviation of the sample.

4.1.1 The idea of probability

So how does the value of $\bar{x}$ help us find the value of μ? If we take another sample, the sample mean $\bar{x}$ is going to be different from the sample mean of the first sample. This difference in sample means due to different samples is refered to as **sampling variability.**

Why is sampling variability not fatal? In the short run a chance occurence is unpredictable, but in the long run, there will be a predictable pattern. We may not be able to predict the outcome of the next toss of a fair coin, but in the long run we would expect the coin to land on heads 50% of the time.

A phenomenon is **random** if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

The **probability** of an event is the proportion of times the event occurs in a very long series of repetitions.

4.1.2 Thinking about randomness

To understand randomness, observe random behavior (both long-run and short-run).

- You must have a long series of **independent** trials.
- The idea of probability is empirical. Computer simulations can imitate random behavior but cannot estimate probabilities.
- Computer simulations are useful since they give us a long run of trials.

The following graphs are two simulations of tossing a coin.

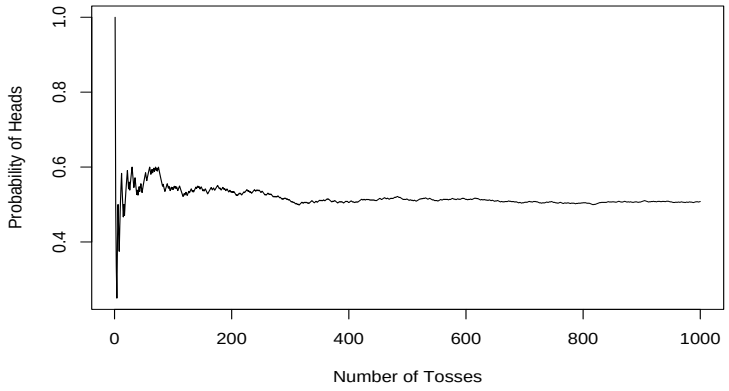

Figure 1: Proportion of Heads obtained - Simulaton 1

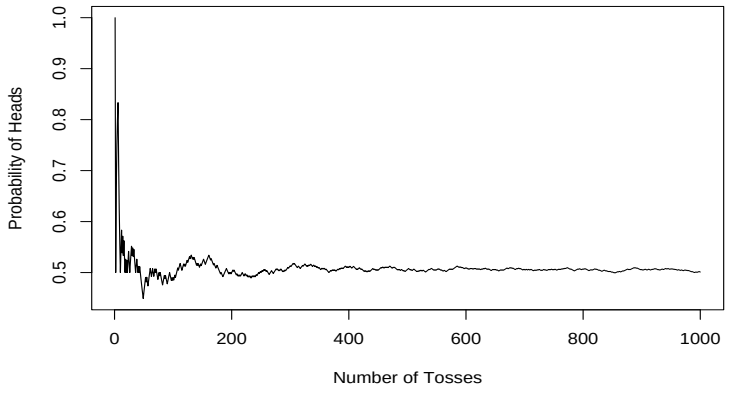

Figure 2: Proportion of Heads obtained - Simulation 2

4.2 Probability Models

For a random event, we cannot predict the outcome of the next trial, and we cannot predict correctly the outcome of the trial after that. What we can do is to list out all possible outcomes and the probability of getting such outcomes.

The **sample space** S of a random phenomenon is the set of all possible outcomes.

An **event** is any outcome or a set of outcomes of a random phenomenon.

A **probability model** is a mathematical description of a random phenomenon consisting of two parts: a sample space S and a way of assigning probabilities to events.

When we toss two coins we could get the following sample space.

$$\text{S} = \{\text{HH, HT, TH, TT}\}$$

where HT denotes getting a head on the first toss and a tail on the second toss.

Similarly when we toss two dice we can get 36 possible outcomes. In this case S=

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Suppose we were interested in just the sum of the faces of the two dice that were rolled. In that case the sample space is

$$\text{S} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

4.2.1 Probability rules

So what is the probability of obtaining a head in tosses of a coin or the probability of obtaining two sixes when rolling two dice? To obtain these probabilities we need to repeat the tosses (of coins or dice) many times and record the number of times we obtain a head or two sixes, giving the long run proportion of observed successes as the probability. However there are a few rules that need to be observed in any asssignment of probabilities.

1. **Any probability is a number between 0 and 1.**Any proportion, and so any probability is a number between 0 and 1. An event with probability 0 never occurs, an event with probability 1 always occurs and an event with probability 0.5 occurs in half the trials in the long run.

2. **All possible outcomes together must have probability 1.**Since an outcome has to occur, the occurence of any one outcome in the sample space is a sure event and hence has probability 1.

3. **The probability that an event does not occur is 1 minus the probability that the event does occur.**At any given instant an event occurs or doesnt. Since that is a sure event, it has probability 1 and so the sum of the probabilities of an event occurring and not occuring has to add upto 1.

4. **If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual properties.**Two events are said to be disjoint if they

cannot happen at the same time. In such a case the probability of either event is just the sum of their individual properties.

These rules can be summarized as follows.

1. $0 \leq P(A) \leq 1$.
2. $P(S) = 1$.
3. $P(A^c) = 1 - P(A)$
4. $P(A \text{ or } B) = P(A) + P(B)$ where A and B are disjoint.

Consider the example of tossing two dice and summing the faces that turn up. Our sample space is

$$\text{S} \quad = \quad \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Suppose we were interested in the following events.

A - get an even sum, i.e. 2, 4, 6, 8, 10, 12.

B - get 2, 3, 4, 5, 6.

C - get 7, 8, 9, 10, 11, 12.

D - get 3, 5.

How can we calculate the probabilities of each of these events? To do this we need to know the probabilities of each one of the individiual outcomes in the sample space S. Suppose the dice were not loaded and hence had an equal chance of turning up one of 6 faces. In that case we have 36 possible face combinations and each combination has an equal chance of occuring.

S		Combinations	Probability
2	-	(1,1)	$\frac{1}{36}$
3	-	(1,2), (2,1)	$\frac{2}{36}$
4	-	(1,3), (2,2), (3,1)	$\frac{3}{36}$
5	-	(1,4), (2,3), (3,2), (4,1)	$\frac{4}{36}$
6	-	(1,5), (2,4), (3,3), (4,2), (5,1)	$\frac{5}{36}$
7	-	(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)	$\frac{6}{36}$
8	-	(2,6), (3,5), (4,4), (5,3), (6,2)	$\frac{5}{36}$
9	-	(3,6), (4,5), (5,4), (6,3)	$\frac{4}{36}$
10	-	(4,6), (5,5), (6,4)	$\frac{3}{36}$
11	-	(5,6), (6,5)	$\frac{2}{36}$
12	-	(6,6)	$\frac{1}{36}$

We can now calculate the probabilities of the events A, B, C and D.

1. $P(A) = P(2, 4, 6, 8, 10, 12) = P(2) + P(4) + P(6) + P(8) + P(10) + P(12) = \frac{1+3+5+5+3+1}{36} = \frac{18}{36}$

2. $P(B) =$
3. $P(C) =$
4. $P(D) =$
5. $P(A \text{ or } D) =$
6. $P(B \text{ and } D) =$
7. $P(B \text{ and } C) =$
8. $P(C \text{ or } D) =$

Two events are said to be **independent** when the occurence of one event has no affect on the probability of the other event occurring. If A and B are independent events then the $P(A \text{ and } B) = P(A)P(B)$. Note that disjoint events cannot be independent.

If we roll two dice and define the event A as the event of getting a 5 on the first die and B as the event of getting a 2 or a 4 on the second die. In this case, if the dice are unbiased, then $P(A) = \frac{1}{6}$ and $P(B) = \frac{2}{6}$, and hence the probability of observing a 5 on the first roll and a 2 or a 4 on the second roll is the product of the two probabilities $P(A).P(B) = \frac{1}{6}.\frac{2}{6} = \frac{2}{36}$.

4.2.2 Assigning probabilities: finite number of outcomes

From the above problem we now formally state how to assign probabilities to events when we have a finite number of outcomes.

- Assign probabilities to each individual outcome. These probabilities must satisfy probability rules 1 and 2.
- The probability of any event is the sum of the probabilities of the outcomes that make up the event.

4.2.3 Assigning probabilities: interval of outcomes

An interval of outcomes usually means that you have an infinite number of outcomes. In this case, the probability of an individidual outcome is zero. However the probability of a set of outcomes is the area under the curve that describes the probability.

Consider the following example where the x-axis represents the waiting time in a clinic.

1. What is the probability that you wait less than 5 minutes?
   If X is the amount of time you wait, then
   $P(X < 5) =$ Area under the triangle to the left of 5$= \frac{1}{2}(5)(0.2) = 0.5$
2. What is the probability that you wait more than 8 minutes?
   (Hint: The height of the line above the point 8 is 0.04)

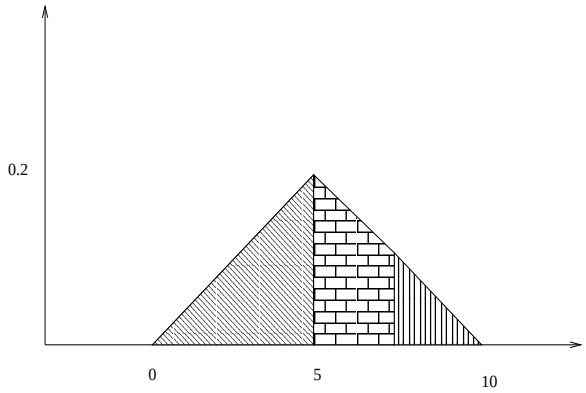

Figure 3: Distribution of waiting times at a Clinic

3. What is the probability that you wait between 5 and 8 minutes?
   (Hint: Use 1 and 2 to answer this one!)

4.2.4 Normal probability distributions

One of the most used probability distribution is the normal probability distribution. It is a good approximation to many other distributions and is many times an idealized description for data. Consider the height of all young women. The heights follow a normal distribution closely with $\mu = 64.5$ inches and $\sigma = 2.5$ inches. Choose one woman at random and measure her height. Call her height X. If we repeat this many times the heights we would get would resemble a normal distribution.

What is the probability that a randomly selected women is between 68 and 70 inches?

In the above example we used X to describe the height of a randomly selected woman. So the value that X takes changes when we choose another woman. Since X changes values randomly (depending on the random selection of the woman), X is referred to a random variable.

$\Rightarrow$ A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

$\Rightarrow$ The **probability distribution** of a random variable X tells us what values X can take and how to assign probabilities to those values.

4.3 Sampling Distributions

A random variable is random because its value is dependant on a random phenomenon. A statistic is random because we collect samples in a random manner. Hence we can describe the behavior of the statistic we are interested in by asking us the question "What would happen if we did this many times?"

4.3.1 Statistical estimation and the law of large numbers

The population mean μ is a fixed number but in practise it is usually unknown. The sample mean $\bar{x}$ is a random quantity (if obtained as the mean of a random sample), but since an SRS is supposed to represent the population, we can use $\bar{x}$ to estimate μ. However if we took another sample we would get a different sample mean and hence a different estimate for the populaton mean. So how is this a valid method? Suppose we increase our sample size. In this case the sample mean is going to coincide with the population mean when n gets very large. This is the **Law of Large Numbers.**

Draw SRS of different sizes from the population. As the sample size increases, the mean $\bar{x}$ of the observed values gets closer to the population mean.

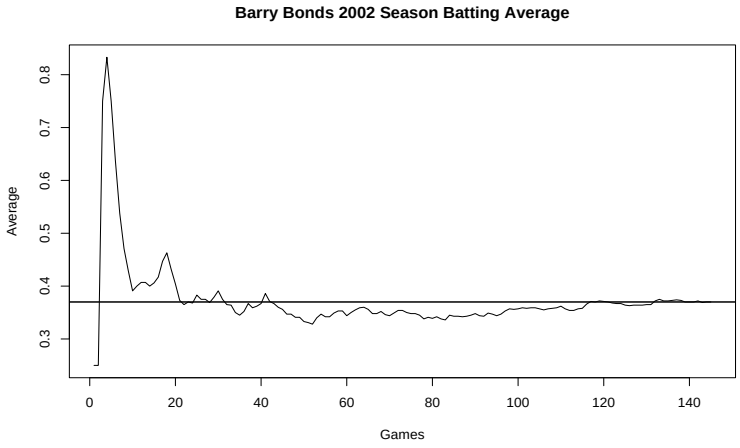

Figure 4: Law of Large Numbers in "action"

4.3.2 Sampling distributions

Since $\bar{x}$ is a random variable, we will get different values for it when we take a different samples. However there is a long-run regularity to all the values that we get.

1. Suppose our population is {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}.
2. Collect many samples of size 2.
3. Calculate $\bar{x}$ for each sample.

4. Make a histogram of the values of $\bar{x}$
5. Examine the distribution, i.e. shape, center and spread.
6. Repeat this for samples of size 4.

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

4.3.3 The mean and the standard deviatiation of $\bar{x}$

Usually we are interested in estimating the population mean μ and the statistic used to estimate it is the sample mean $\bar{x}$. The sample mean has some interesting properties with relation to the population mean.

Suppose that $\bar{x}$ is the mean of an SRS drawn from a large population with mean μ and standard deviation σ, then the mean of all sample means (sample means from all possible samples) is the population mean μ and the standard deviation of all possible sample means is $\sigma/\sqrt{n}$.

The sample mean is an **unbiased estimator** of the population mean. An unbiased estimator is one that has no tendency to consistently over-estimate or under-estimate the population parameter.

Also note that the standard deviation of $\bar{x}$ depends on the sample size n. The larger your sample size (n), the smaller the standard deviation, implying that your results are more accurate.

4.3.4 The central limit theorem

Sampling Distribution of $\bar{x}$.
If a population is normally distributed with mean μ and standard deviation σ, then the sample mean from a sample of size n is also normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

The normal distribution is very useful in the sense that it describes many of the variables we might be interested in. However there are times when a population distribution may not be normally distributed. Note that the mean of all sample means is the same as the population mean, regardless of whether the population is normally distributed. The shape of the distribution of $\bar{x}$ depends on the shape of the original population. However as n increases, the shape tends to resemble that of a normal distribution, which is the basis of the central limit theorom.

Central Limit Theorem
Draw an SRS from any population with mean μ and standard deviation σ. If n is large, the sampling distribution of $\bar{x}$ is approximately normal.

A study of rush hour traffic counts the number of people in each of 700 cars entering a busy freeway at rush hour. The count has mean μ=1.5 and standard deviation σ=0.75.

1. Could the distribution of the count be normal? Why or why not?
2. What is the distribution of the sample mean from a sample of size 700?
3. What is the probability that 700 cars have more than 1075 people?