
1. Random Sample and Sampling Distributions

1.1	Random sample	1-1
1.2	Statistics and their sampling distributions	1-2
1.3	Exercises	1-5

1.1 Random sample

Statistical science is concerned with data collection and using them to draw conclusions. We then define the **population** as the set of individuals that we want to draw conclusions about while the **sample** is defined as the portion of the population that we actually examine. The number of individuals in the sample corresponds to the **sample size**.

The measured characteristic from each individual in the sample is a random variable and the collection of characteristics from all individuals in the sample is called a **random sample**. Each element in the random sample is an observation from the same population. The set of all possible values of these random variables is called the **sample space**.


Often we wish to measure some unknown characteristic of the population. A characteristic of the population is called a **parameter**. The set of all possible values of the parameters is called the **parameter space**. We use the sample to infer the value of the parameter. Any quantity calculated from the sample is called a **statistic**. A statistic is therefore a random variable and its distribution is called the **sampling distribution**.

Let us consider the following example.

Example 1.1. The office for national statistics (ONS) wishes to measure the unemployment rate in the UK. To that end, it chooses people of working age within the UK and asks them whether they are employed or seeking employment. The proportion among those asked who are seeking employment can be used to measure the unemployment rate.

In this example the population consists of all individuals able to work in the UK. The parameter we wish to estimate is the unemployment rate p which is a proportion so the parameter space is the set $[0, 1]$.

Because the ONS cannot ask every individual, it asks a subset of the population. The individuals asked consist of the sample. The proportion in the sample seeking employment is a statistic because it is calculated from the sample and not the whole population.

Suppose n individuals were asked and let X_i denote the response of the i th individual, $i = 1, \dots, n$. We let $X_i = 1$ if the i th individual is seeking employment and 0 if not so in this case the sample space is the set $\{0, 1\}$. The random sample is the set $\{X_1, \dots, X_n\}$. The proportion in the sample is also the mean of the X_i 's, denoted by \bar{X} . Each X_i is distributed as $X_i \sim \text{Bernoulli}(p)$, so the sampling distribution of \bar{X} is the distribution of the sample proportion, $\text{Bin}(n, p)/n$. 

Formally, we define a random sample as follows.

Definition 1.1 (Random sample).

The random variables X_1, \dots, X_n are called a **random sample** of size n from the population $f(x|\theta)$ depending on a parameter θ if X_1, \dots, X_n are mutually independent random variables and the pdf/pmf of each X_i is the same function $f(x|\theta)$. The variables X_1, \dots, X_n are also called **independent and identically distributed (iid) random variables**. We write $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x|\theta)$.

Often we are interested in the joint distribution of our sample. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x|\theta)$. Then the joint pdf/pmf of X_1, \dots, X_n is

$$f(x_1, \dots, x_n|\theta) = f(x_1|\theta) \times \dots \times f(x_n|\theta) = \prod_{i=1}^n f(x_i|\theta),$$

where the first equality is true because the random variables are mutually independent.

Example 1.2. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\mu)$. For example X_1, \dots, X_n may correspond to the failure times (measured in years) for n identical circuit boards that are put to test and used until they fail. Each X_i has pdf

$$f(x|\mu) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right),$$

so the joint pdf of the sample is

$$\begin{aligned} f(x_1, \dots, x_n|\mu) &= \prod_{i=1}^n f(x_i|\mu) \\ &= \frac{1}{\mu} \exp\left(-\frac{x_1}{\mu}\right) \times \frac{1}{\mu} \exp\left(-\frac{x_2}{\mu}\right) \times \dots \times \frac{1}{\mu} \exp\left(-\frac{x_n}{\mu}\right) \\ &= \frac{1}{\mu^n} \exp\left(-\frac{1}{\mu} \sum_{i=1}^n x_i\right). \end{aligned}$$

►

1.2 Statistics and their sampling distributions

In statistical inference, we are interested in describing the distribution of the population. In most cases, a suitable calculation using the sampled values can help.

Definition 1.2 (Statistic and its sampling distribution).

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x|\theta)$. A function $T = T(X_1, \dots, X_n)$ of the variables X_1, \dots, X_n , which does not depend on θ , is called a **statistic**. The statistic is itself a random variable. The probability distribution of T is called its **sampling distribution**.

Another way to think of the sampling distribution is as the distribution of all possible values of T for all possible random samples of size n from the population $f(x|\theta)$.

Example 1.3. Let X_1, \dots, X_n be a random sample of size n . Two of the most frequently used statistics are the sample mean, \bar{X} , and the sample variance S^2 defined by

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

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

Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then, the sampling distribution of \bar{X} is

$$\bar{X} \sim N(\mu, \sigma^2/n),$$

i.e., the normal distribution with mean μ and variance σ^2/n , and the sampling distribution of S^2 is

$$(n-1) \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2,$$

i.e., the chi-squared distribution with $n-1$ degrees of freedom times the constant $\sigma^2/(n-1)$. Moreover, as stated in the next theorem, \bar{X} and S^2 are independent in the case of normal populations. ►

Theorem 1.1.

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then the statistics \bar{X} and S^2 are independent.

Proof. Let $Y_i = X_i - \bar{X}$ for $i = 1, \dots, n$. We will show that \bar{X} is independent of each of the Y_i 's. Because S^2 is only a function of the Y_i 's, it follows that S^2 is independent of \bar{X} .

To prove independence between \bar{X} and Y_i , it is enough to show that $\text{Cov}(Y_i, \bar{X}) = 0$ because both Y_i and \bar{X} are normally distributed. Indeed,
 $\text{Cov}(Y_i, \bar{X}) = \text{Cov}(X_i - \bar{X}, \bar{X}) = \text{Cov}(X_i, \bar{X}) - \text{Cov}(\bar{X}, \bar{X}) = \text{Cov}(X_i, \frac{1}{n} \sum_j X_j) - \text{Var } \bar{X} = \frac{1}{n} \sum_j \text{Cov}(X_i, X_j) - \text{Var } \bar{X} = \frac{1}{n} \sigma^2 - \frac{1}{n} \sigma^2 = 0.$ □

The sampling distribution is not always easy to derive, either because the distribution of the population is unknown or because the statistic does not have a straightforward expression. Sometimes we can state asymptotic results as the sample size increases.

Theorem 1.2 (Law of large numbers).

Let X_1, \dots, X_n be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}(|\bar{X} - \mu| \geq \varepsilon) \rightarrow 0, \text{ as } n \rightarrow \infty.$$

Proof. This is easily proved by Chebychev's inequality: $\mathbb{P}(g(X) \geq r) \leq \mathbb{E} g(X)/r \ \forall r > 0$.

$$\mathbb{P}(|\bar{X} - \mu| \geq \varepsilon) = \mathbb{P}((\bar{X} - \mu)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}(\bar{X} - \mu)^2}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad \square$$

The law of large numbers simply states that the probability of small deviations of the sample mean from the population mean can be made very small if we choose a large enough sample size.

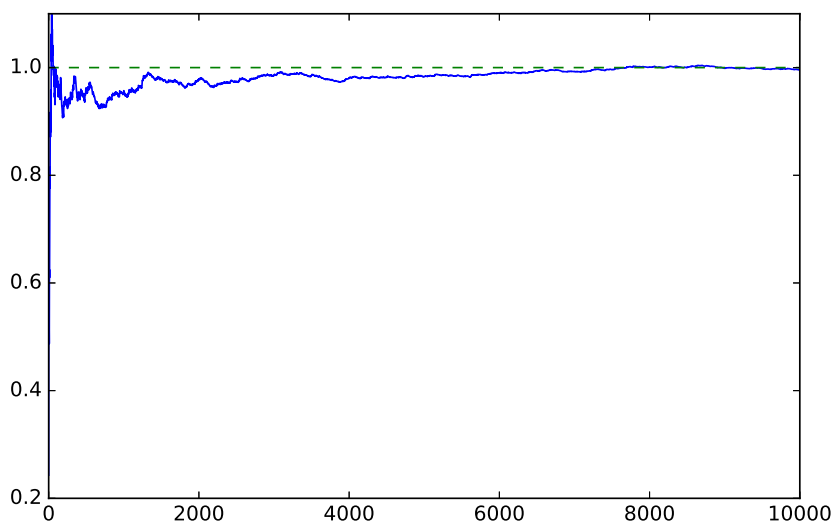
Example 1.4. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\mu)$. Then $\mathbb{E} X_i = \mu$, therefore $\mathbb{E} \bar{X} = \mu$. The law of large numbers says that the probability that $|\bar{X} - \mu|$ exceeding a small number ε can become arbitrarily small by increasing the sample size n . This is illustrated by the following python code.

```
import matplotlib.pyplot as plt
import scipy
import scipy.stats
```

```

N = 10000 # Max sample size
mu = 1    # The mean (scale) parameter
x = scipy.stats.expon.rvs(size=N, scale=mu)
xbar = scipy.cumsum(x)/scipy.arange(1,N+1)
plt.plot(xbar)                # xbar at n = 1,2,...,N
plt.plot([0,N],[mu,mu], '--') # Horizontal line
plt.show()

```



The sample mean is ubiquitous in statistics and is important to know its sampling distribution. The next theorem summarises the large-sample behaviour of the sample mean.

Theorem 1.3 (Central limit theorem).

Let X_1, \dots, X_n be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Define $Z_n = \sqrt{n}(\bar{X} - \mu)/\sigma$. Then, for any $z \in \mathbb{R}$,

$$\mathbb{P}(Z_n < z) \rightarrow \Phi(z), \text{ as } n \rightarrow \infty,$$

where $\Phi(z)$ denotes the CDF of the $N(0, 1)$ distribution.

In other words, the central limit theorem says that the CDF of \bar{X} and the CDF of $N(\mu, \sigma^2/n)$ are visually indistinguishable for large sample size. Since in many cases we cannot come up with the sampling distribution of the sample mean, the approximate normal distribution can be used assuming that the sample size is large.

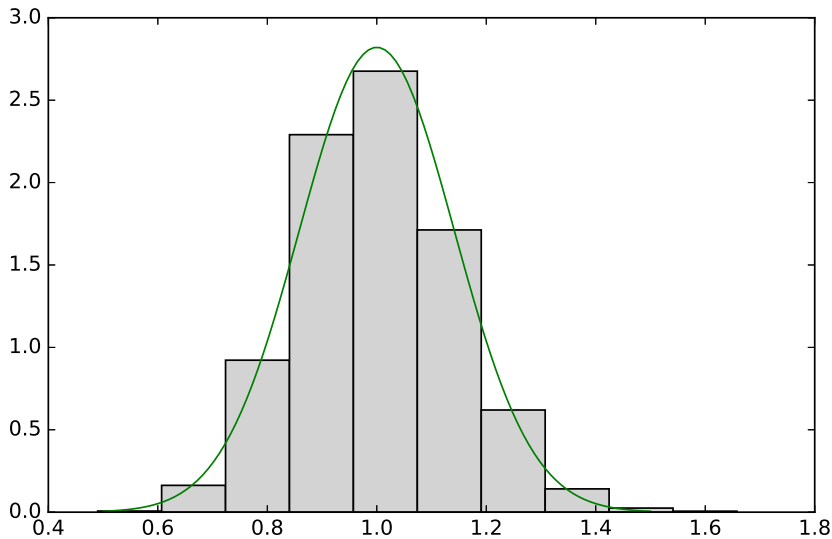
Example 1.5. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\mu)$. Then $\mathbb{E} X_i = \mu$ and $\text{Var} X_i = \mu^2$. By the central limit theorem, the distribution of \bar{X} is approximately $N(\mu, \mu^2/n)$ for large n . This is illustrated by the following python code.

```

import matplotlib.pyplot as plt
import scipy
import scipy.stats
N = 10000 # Number of repetitions
n = 50    # Sample size for each repetition
mu = 1    # The mean (scale) parameter
x = scipy.stats.expon.rvs(size=(N,n), scale = mu)

```

```
xbar = scipy.mean(x,1)          # Sample mean across rows
xx = scipy.linspace(.5,1.5,100)
plt.hist(xbar, normed=True, alpha=0.5, facecolor='lightgray')
plt.plot(xx, scipy.stats.norm.pdf(xx, mu, mu/scipy.sqrt(n))) # Normal pdf
plt.show()
```



1.3 Exercises

1. A coffee shop buys roasted coffee from a supplier. In order to assess the quality of the supplied coffee, the manager of the shop conducts a tasting experiment where she selects a small portion of coffee beans from different batches and tastes the coffee from each portion. For each portion she gives a score in the scale $1, 2, \dots, 10$ with 10 corresponding to coffee of the best taste and uses the results to assess the quality of the coffee.

Identify the population, parameter, and statistic.

2. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$. Derive the sampling distribution of \bar{X} given in Example 1.3.
3. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p)$.
 - a) Derive the sampling distribution of \bar{X} .
 - b) Derive the asymptotic distribution of \bar{X} from the central limit theorem.
 - c) Draw a graph of the exact and approximate CDFs when $n = 20$ and $p = 0.4$.