

## Behavior of the sample mean

We observe  $n$  independent and identically distributed (iid) draws from a random variable  $X$ . Denote the observed values by  $X_1, X_2, \dots, X_n$ .

Assume the  $X_i$  come from a population with expected value  $\mu$  and variance  $\sigma^2$ .

$$EX_i = \mu$$

$$\text{var}X_i = \sigma^2$$

The sample mean

$$\bar{X} = (X_1 + \dots + X_n)/n = \sum_i X_i/n$$

is a random variable. The “population” for  $\bar{X}$  is different from, but related to, the population of the individual observations. The distribution of  $\bar{X}$  is called the “sampling distribution.”

**Fundamental fact:** The expected value of  $\bar{X}$  is

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

and the variance of  $\bar{X}$  is

$$\text{var}\bar{X} = \sigma^2/n$$

**Mathematical definition of the expected value:** If a random variable  $X$  has a density function  $q(x)$ , then the expected value of  $X$  is

$$\int xq(x)dx$$

and the population variance is

$$\int (x - EX)^2 q(x) dx$$

**Example:** What are the expected value and variance of a uniform random variable on  $(0, 1)$ ?

**Solution:** The density is  $q(x) = 1$  on  $(0, 1)$ , and  $q(x) = 0$  elsewhere. The expected value is

$$\int_0^1 x dx = 1/2.$$

The variance is

$$\int_0^1 (x - 1/2)^2 dx = 1/12.$$

**Example:** What are the expected value and variance of a standard exponential random variable, which has density  $q(x) = \exp(-x)$  on  $(0, \infty)$ .

**Solution:** Using integration by parts, the expected value is

$$\int x \exp(-x) dx = - \int_0^\infty x \exp(-x) + \int \exp(-x) dx = 1.$$

The variance is

$$\begin{aligned} \int (x - 1)^2 \exp(-x) dx &= \int x^2 \exp(-x) dx - 2 \int x \exp(-x) dx + \int \exp(-x) dx \\ &= \int x^2 \exp(-x) dx - 1 \\ &= - \int_0^\infty x^2 \exp(-x) + 2 \int x \exp(-x) dx - 1 \\ &= 1. \end{aligned}$$

### Simulations:

We can study the sampling behavior of  $\bar{X}$  by simulating data sets under different conditions and calculating the  $\bar{X}$  values of each set.

To learn about  $E\bar{X}$ , we sample many data sets from a fixed population, and calculate  $\bar{X}$  from each of them. Then we average the  $\bar{X}$  values together. To learn about  $\text{var}\bar{X}$ , instead of averaging we would calculate the sample standard deviation of the  $\bar{X}$  values.

**R syntax for random arrays:** To generate a vector of  $m$  iid standard normal values in R, use:

```
Z <- rnorm(m)
```

To generate a  $m \times n$  array of standard normal values, use

```
Z <- rnorm(m*n)
A <- array(Z, c(m,n))
```

or

```
A <- array(rnorm(m*n), c(m,n))
```

**Exercises:** What do you expect the value of  $V$  to be after executing each of the following programs?

- Program 1

```
## Generate a 50x1000 array of standard normal draws.
X <- array(rnorm(50000), c(50,1000))

## Get the mean of each column of X.
Y <- apply(X, 2, mean)

## Calculate the sample variance of Y.
V <- var(Y)
```

- Program 2

```
## Generate a 50x1000 array of uniform draws on (0,1).
X <- array(runif(50000), c(50,1000))

## Get the mean of each column of X.
Y <- apply(X, 2, mean)

## Calculate the sample variance of Y.
V <- var(Y)
```

- Program 3

```
## Generate a 50x1000 array of uniform draws on (0,1).  
X <- array(runif(50000), c(50,1000))  
  
## Get the mean of each column of X.  
Y <- apply(X, 2, mean)  
  
## Calculate the sample variance of Y.  
V <- var(Y)
```

- Program 4

```
## Generate a 50x1000 array of standard exponential draws.  
X <- array(rexp(50000), c(50,1000))  
  
## Get the mean of each column of X.  
Y <- apply(X, 2, mean)  
  
## Calculate the sample variance of Y.  
V <- var(Y)
```

We can also look at how the variance of  $\bar{X}$  is affected by changes in the sample size.

```

## Initialize V so we can append onto it.
V <- NULL

## The value of r loops through the values 10,20,30,40,50.
for (r in c(10,20,30,40,50))
{
  ## Generate a rx1000 array of standard exponential draws.
  X <- array(rexp(r*1000), c(r,1000))

  ## Get the mean of each column of X.
  Y <- apply(X, 2, mean)

  ## Calculate the sample variance of Y.
  V <- c(V, var(Y))
}

```

### More properties of the sample mean:

If  $c$  is a constant and  $X$  and  $Y$  are random variables, the expected value has the following properties:

$$E(X + c) = (EX) + c$$

$$E(c \cdot X) = c \cdot EX$$

$$E(X + Y) = EX + EY$$

If  $X$  and  $Y$  are *uncorrelated* then

$$E(X \cdot Y) = EX \cdot EY.$$

The sample mean has similar properties. If  $Y_i = X_i + c$  and  $Z_i = cX_i$ , then

$$\bar{Y} = \bar{X} + c$$

$$\bar{Z} = c\bar{X}$$

If  $U_i$  and  $V_i$  are sequences of numbers and  $W_i = U_i + V_i$ , then

$$\bar{W} = \bar{U} + \bar{V}.$$

**Simulations:** We can explore some of these properties using simulations. In the following example,  $X$  and  $Y$  are uncorrelated, while  $U$  and  $V$  are correlated. You can check how well the identity  $E(X \cdot Y) = EX \cdot EY$  holds for  $X$  and  $Y$ , and for  $U$  and  $V$ .

```
X <- rexp(1e4)
Y <- rexp(1e4)
Z <- X*Y
```

```
A <- rexp(1e4)
U <- A + rexp(1e4)
V <- A + rexp(1e4)
W <- U*V
```

### Functions of random variables:

If  $X$  is a random variable and  $f(x)$  is a mathematical function, then  $Y = f(X)$  is a random variable. It has its own population, distribution, expected value, etc. that generally differ from those of  $X$ .

If  $f(x)$  is a concave function (e.g. log or square-root), then

$$Ef(X) \leq f(EX).$$

What will be the sign of E1-E2 after executing the following program?

```
## 1000 standard exponential draws
X <- rexp(1000)
```

```
## Estimate E log(X)
E1 <- mean(log(X))
```

```
## Estimate log(EX)
E2 <- log(mean(X))
```

If  $f(x)$  is a convex function (e.g.  $f(x) = x^2$ ) then

$$Ef(X) \geq f(EX).$$

What will be the sign of E1-E2 after executing the following program?

```
## 1000 standard normal draws
X <- rnorm(1000)
```

```
## Estimate E X^2
E1 <- mean(X^2)
```

```
## Estimate (EX)^2
E2 <- mean(X)^2
```

Mathematically, a smooth function  $f(x)$  is concave if  $f''$  is always non-positive, and is convex if  $f''$  is always non-negative. Note that many functions are neither (e.g.  $f(x) = x^3$ ).

### Exceptional cases:

The *Cauchy distribution* has no mean and infinite variance. It is defined as the tangent of a uniform random angle. That is,  $X = \tan(U\pi)$  where  $U$  is uniform on  $(0, 1)$  has a Cauchy distribution.

What do you expect the value of V to be after executing the following program?

```

## Initialize V so we can append onto it.
V <- NULL

## The value of r loops through the values 10,20,30,40,50.
for (r in c(10,20,30,40,50))
{
  ## Generate a 50x1000 array of standard exponential draws.
  X <- array(rcauchy(r*1000), c(r,1000))

  ## Get the mean of each column of X.
  Y <- apply(X, 2, mean)

  ## Calculate the sample variance of Y.
  V <- c(V, var(Y))
}

```

## Estimating the variance

The variance is not a linear function of the data, so formulating an estimator for the variance is not as simple as in the case of the expected value. Since the population variance is defined by

$$E(X - EX)^2$$

it might make sense to estimate the population variance using

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

How does this estimate perform? Let's do a simulation.

```

## Create a list to append onto.
V <- NULL

for (i in 1:1000)
{
  ## Generate a list of 20 standard normal draws.
  X <- rnorm(20)

  ## Center X.
  X <- X - mean(X)

  ## The second moment estimate.
  C2_hat = sum(X^2)/20

  ## Stick it onto the end of V.
  V[i] <- C2_hat
}

Y <- mean(V)

```

If you run this several times, you will see that the value of Y is consistently below the true value of 1. However if we use the usual sample variance estimate

$$\sum_{j=1}^n (X_j - \bar{X})^2 / (n - 1)$$

we get better results (check this by replacing the 20 in the denominator of C2\_hat with 19).