5. Random Variables and Probability Distributions

Learning Objectives

1. Determine probabilities from probability mass functions and cumulative distribution functions.
2. Understand the assumptions for probability distributions.
3. Select an appropriate probability distribution to calculate probabilities.
4. Calculate probabilities, means and variances for probability distributions.

Definitions

A **random variable** is a function that associates a real number with each element in the sample space.

The probability distribution of a random variable *X* gives the probability for each value of *X*.

Random variables

A random variable takes a **numeric** value based on the outcome of a random event.

Denote by capital letter — X, Y, Z, etc.

A particular value of a random variable will be denoted with a lower case letter — x, y, z

There are two types of random variables:

- **Discrete** random variables: can take one of a finite number of distinct outcomes.
- **Continuous** random variables: can take any numeric value within a range of values.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.

The possible outcomes, together with their probabilities are:-

Sample	prob	X
NNN		
DNN		
NDN		
NND		
DDN		
DND		
NDD		
DDD		

Random variable X is the number of defectives in the sample.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.
The possible outcomes, together with their probabilities are:-

Sample	Prob	X
NNN	$(0.9)^3$	
DNN	$(0.1)(0.9)^2$	
NDN	$(0.9)(0.1))0.9)$	
NND	$(0.9)^2(0.1)$	
DDN	$(0.1)^2(0.9)$	
DND	$(0.1)(0.9)(0.1)$	
NDD	$(0.9)(0.1)^2$	
DDD	$(0.1)^3$	

Random variable X is the number of defectives in the sample.

Example: Discrete Random Variable

Computer chips may be classed as defective (D) or non-defective (N).
A large batch contains a proportion 0.1 of defectives, and 3 are sampled at random.
The possible outcomes, together with their probabilities are:-

Sample	Prob	X
NNN	$(0.9)^3$	0
DNN	$(0.1)(0.9)^2$	1
NDN	$(0.9)(0.1))0.9)$	1
NND	$(0.9)^2(0.1)$	1
DDN	$(0.1)^2(0.9)$	2
DND	$(0.1)(0.9)(0.1)$	2
NDD	$(0.9)(0.1)^2$	2
DDD	$(0.1)^3$	3

Random variable X is the number of defectives in the sample.

Probability model (discrete)

The collection of all possible values of a random variable together with associated probabilities is called the **probability model**

In the example, $\Pr(X = 1)$ can be determined by adding up the probabilities of the 3 sample points associated with the event $X = 1$, etc

x	$\Pr(X = x)$
0	
1	
2	
3	

Probability model (discrete)

The collection of all possible values of a random variable together with associated probabilities is called the **probability model**

In the example, $\Pr(X = 1)$ can be determined by adding up the probabilities of the 3 sample points associated with the event $X = 1$, etc

x	$\Pr(X = x)$
0	$(0.9)^3$
1	$3(0.1)(0.9)^2$
2	$3(0.1)^2(0.9)$
3	$(0.1)^3$

Example

A couple having children will stop when they have a child of each sex or three children.

outcome	GGG	GGB	GB	BG	BBG	BBB
Probability						

Let the random variable X be the number of girls in the family

x	0	1	2	3
$P(X = x)$				

Probability function for a discrete random variable represented pictorially by a **bar graph**.

Example

A couple having children will stop when they have a child of each sex or three children.

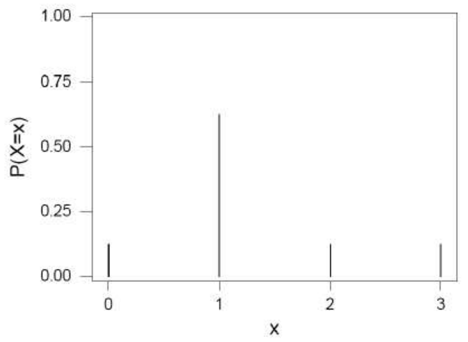
outcome	GGG	GGB	GB	BG	BBG	BBB
Probability	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Let the random variable X be the number of girls in the family

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Probability function for a discrete random variable represented pictorially by a **bar graph**.

Probability Function: Bar Graph



Class survey: choosing a number at random

Frequency

````
```{r}
survey.data %>% select (number) %>% table
```
````

````
number
 1  2  3  4  5  6  7  8  9 10 
 4 19 17 21 15 15 38 22 17  5 
````


Choosing a number at random

Probability

````
```{r}
survey.data %>% select (number) %>% table %>% prop.table %>% round(digits = 2)
```
````

````
number
   1    2    3    4    5    6    7    8    9   10 
0.02 0.11 0.10 0.12 0.09 0.09 0.22 0.13 0.10 0.03 
````


Choosing a number at random

Probability

````
```{r}
survey.data %>% select (number) %>% table %>% prop.table %>% round(digits = 2)
```
````

````
       number
X         1    2    3    4    5    6    7    8    9   10 
P(X=x) 0.02 0.11 0.10 0.12 0.09 0.09 0.22 0.13 0.10 0.03 
````

Note that capital X denotes the random variable while small x denotes one of its value

***P*(X=7) = ??**

Discrete Probability Distributions

The set of ordered pairs $(x, f(x))$ is a **probability function**, **probability mass function**, or **probability distribution** of the discrete random variable X if, for each possible outcome x,

1. $f(x) \geq 0$,
2. $\sum_x f(x) = 1$,
3. $P(X = x) = f(x)$.

Capital letters for random variables, small letter for one of its values.

Definitions

The **cumulative distribution function** $F(x)$ of a discrete random variable X with probability distribution $f(x)$ is

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{for } -\infty < x < \infty.$$

The cumulative distribution function, is the probability that a random variable *X* with a given probability distribution will be found at a value less than or equal to *x*.

Cumulative Distribution Functions

Consider the probability distribution for the 'choose a number' example. Find the probability of choosing a 3 or less

- The event ($X \leq 3$) is the total of the events:
  ($X = 0$), ($X = 1$), ($X = 2$), and ($X = 3$).
- From the table:

1	2	3	4	5	6	7	8	9	10
0.02	0.11	0.10	0.12	0.09	0.09	0.22	0.13	0.10	0.03

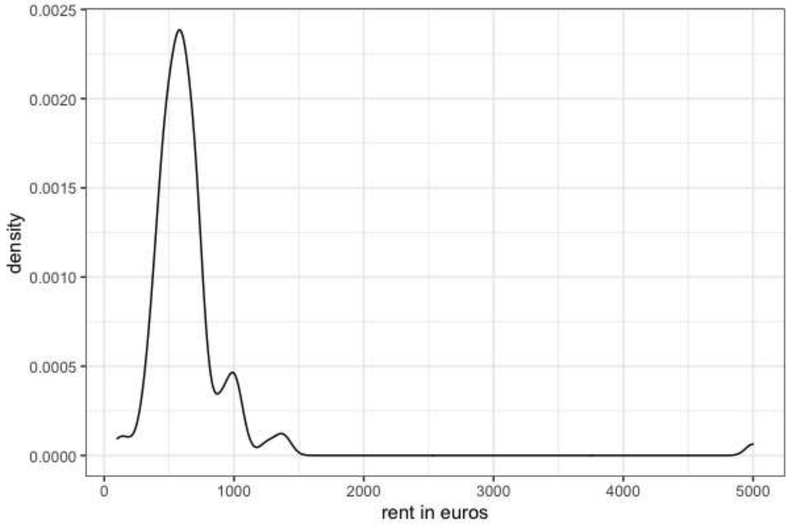
$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0.23$

Continuous Probability Distributions

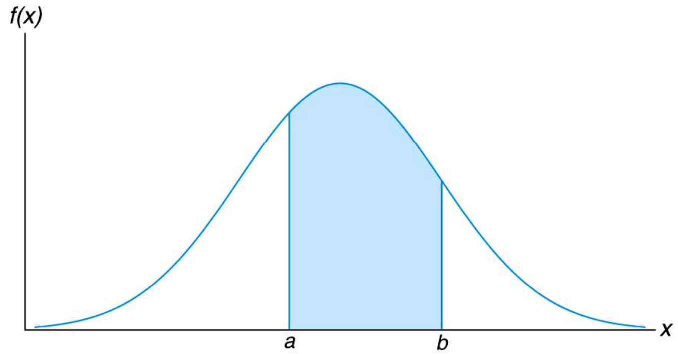
The function $f(x)$ is a **probability density function** (pdf) for the continuous random variable X, defined over the set of real numbers, if

1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\, dx$.

Note $P(X = x) = 0$ i.e. there is no area exactly at x !



$P(a < X < b)$



Popular discrete and continuous distributions

- Discrete:
  - Binomial
  - Poisson
  - Hypergeometric
- Continuous:
  - Uniform
  - Normal
  - Exponential
- What do they look like ?
- When are they used ?

Expected Value — Location

A useful summary of interest is the average, or **expected value**, of a random variable — denoted by $E[X]$ and μ.

Expected Value — Location

A useful summary of interest is the average, or **expected value**, of a random variable — denoted by $E[X]$ and μ.

The expected value of a random variable can be found by summing the products of each possible value by the probability that it occurs:

$$\mu = E[X] = \sum_x xP(X = x)$$

Expected Value — Location

A useful summary of interest is the average, or **expected value**, of a random variable — denoted by $E[X]$ and μ.

The expected value of a random variable can be found by summing the products of each possible value by the probability that it occurs:

$$\mu = E[X] = \sum_x xP(X = x)$$

Example
:

$$E[No.defectives] = 0\times(0.9)^3+1\times3(0.1)(0.9)^2+2\times3(0.1)^2(0.9)+3\times(0.1)^3 =$$

Variance, Standard Deviation — Spread

The **variance** of a random variable measures the squared deviation from the mean:

$$\sigma^2 = \mathrm{Var}(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(X = x)$$

Variance, Standard Deviation — Spread

The **variance** of a random variable measures the squared deviation from the mean:

$$\sigma^2 = \mathrm{Var}(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(X = x)$$

Or more usefully the **standard deviation** is:

$$\sigma = \mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$$

this has the advantage of being in the *same units* as X (and μ).

Variance, Standard Deviation — Spread

The **variance** of a random variable measures the squared deviation from the mean:

$$\sigma^2 = \text{Var}(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(X=x)$$

Or more usefully the **standard deviation** is:

$$\sigma = \text{sd}(X) = \sqrt{\text{Var}(X)}$$

this has the advantage of being in the *same units* as X (and μ).

Example
:

$\text{Var}(No.defectives) = ((0-0.3)^2 \times 0.9^3) + ((1-0.3)^2 \times 3 \times 0.1 \times 0.9^2) +$

$((2\text{-}0.3)^2 \times 3 \times 0.1^2 \times 0.9) + ((3-0.3)^2 \times 0.1^3) = 0.27$

Variance of a Random Variable

$$\text{Var}(X)=E(X^2)-E^2(X)$$

Where

$$E(X^2) = \sum x^2 P(X=x)$$


```{r}
survey.data %>% select (number) %>% summarise (mean = mean(number, na.rm=TRUE), variance = var(number, na.r=TRUE), sd = sd(number, na.rm=TRUE), nas= sum(is.na(number)))
```

Description: df [1 × 4]

mean <dbl>	variance <dbl>	sd <dbl>	nas <int>
5.705202	5.813752	2.411172	4

More on Means and Variances

Adding or subtracting a constant from data shifts the mean but does not change the variance or standard deviation:

$E[X+c] = E[X]+c \qquad \text{Var}(X+c) = \text{Var}(X) \qquad \text{sd}(X+c) = \text{sd}(X)$

$E[X-c] = E[X]-c \qquad \text{Var}(X-c) = \text{Var}(X) \qquad \text{sd}(X-c) = \text{sd}(X)$

Multiplying a random variable by a constant multiplies the mean by that constant and the variance by the square of the constant:

$E[aX] = aE[X] \qquad \text{Var}(aX) = a^2\text{Var}(X) \qquad \text{sd}(aX) = |a|\,\text{sd}(X)$