8. Random variables and distributions

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

We did:

- Probability
  - Definitions
  - Rules of Probability: Addition, complement, multiplication
  - Conditional Probability
  - Mutually exclusive events
  - Independent events

Next we will see...

- Random Variables: Discrete or Continuous
- Discrete Random variables (By hand and using R)
  - P.m.f
  - Expectation
  - Variance
  - C.d.f

Mutually exclusive and independent events

- Two events, A and B, are independent if the occurrence of one event does not change the probability of the occurrence of the other event
  - $P(A|B) = P(A)$
- Two events are mutually exclusive if they cannot occur together. ($P(A \cap B) = 0$)
  - $P(A|B) = 0$

For events A and B

- Addition rule, OR rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - For mutually exclusive events : $P(A \cup B) = P(A) + P(B)$

- Multiplication rule, AND rule: $P(B \cap A) = P(A)P(B|A)$
  - For independent events: $P(A \cap B) = P(A)P(B)$

- Law of total probability: $P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$

- Law of complement: $P(A^c) = 1 - P(A)$

Random variable

A random variable is a variable whose numeric value is based on the outcome of a random experiment.

- We use a capital letter, like X, to denote a random variable
- The values of a random variable are denoted with a lowercase letter, in this case x
- For example, $P(X = x)$

Example Toss a fair coin and count number of "H"(or 1).

Outcome	H	T
Values: $\mathrm{X} = \mathrm{x}$	1	0
Probability: $\mathrm{P}(\mathrm{X} = \mathrm{x})$	$1/2$	$1/2$

Two types of random variables

- **Discrete RV**, where X can take only a finite (or countably infinite) number of values
  - 'things you count'
  - ex.: number of heads in 4 flips, cars that enter in a parking lot in a given period of time, etc.

- **Continuous RV**, where X can take any value on the real line in a bounded or unbounded interval.
  - 'things you measure'
  - ex.: height of PSTAT 10 students, time till the next bus arrives

Discrete Random variable

Example Flip a fair coin once

$$S = \{H, T\}$$

$$X(H) = 1, X(T) = 0$$

,

$$X = \begin{cases} 1 & \text{if coin lands heads} \\ 0 & \text{if coin lands tails} \end{cases}$$

In words, $X =$ number of heads in one coin flip

Example: Flipping two coins

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

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0,$$

$$X = \text{number of heads in flipping two coins}$$

$$X = \begin{cases} 2 & \text{if HH} \\ 1 & \text{if HT or TH} \\ 0 & \text{if TT} \end{cases}$$

In words, X = number of heads in two independent coin flips

Why RV's?

- describe events succinctly
- "Flipping two coins and getting at most one head" or "flipping two coins and getting either no or one head" vs "$X \leq 1$"
- "Flipping two coins" and getting exactly one head " vs "$X = 1$"

Discrete Probability Distribution or p.m.f

A discrete probability distribution, also known as a **probability mass function** or p.m.f, consists of all of the values a random variable can take, along with the corresponding probabilities of taking those values.

Example Flip a fair coin once

Outcome	H	T
Values: $X = x$	1	0
Probability: $P(X = x)$	1/2	1/2

- Note: The sum of these probabilities must be equal to 1.

Toss a coin twice and record the number of heads.

Outcome	TT	HT	HT	HH
# of Heads	0	1	1	2
Probability	0.25	0.25	0.25	0.25

The resulting pmf is the table

x	0	1	2
$P(X = x)$	0.25	0.5	0.25

$$\sum_{\text{all } x} P(X = x) = P(X = 0) + P(X = 1) + P(X = 2) = 1$$

Your turn : Theory

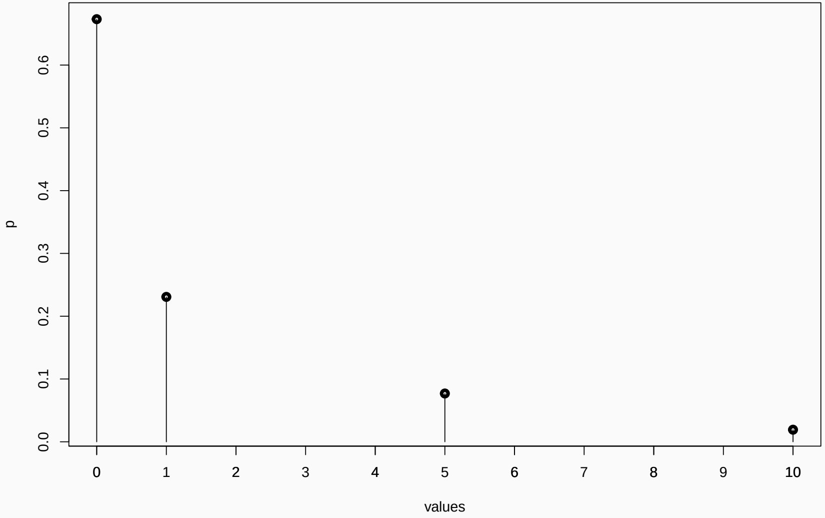
In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability mass function for your winnings.

- (What is the experiment?)
- What are the outcomes, sample space?
- What are you interested in counting?
- What is the random variable? it's values and probabilities?

The pmf for card game

Event	X	$P(X)$
Heart (not ace)	1	$\frac{12}{52} \approx 0.23$
Ace	5	$\frac{4}{52} \approx 0.08$
King of spades	10	$\frac{1}{52} \approx 0.02$
All else	0	$\frac{35}{52} \approx 0.67$
Total		1

The pmf for card game




```
## [1] 0.67 0.23 0.08 0.02
```

Expected Value of a RV or Expectation

Given a random variable X with probability mass function (p.m.f.) $p(x) = P(X = x)$,

Expected Value of X is

$$E(X) = \sum_{i=1}^{k} x_i p(x_i) = \sum_{i=1}^{k} x_i P(X = x_i)$$

Why?

- interested in what we might expect to see (the average outcome)
- We call this the **expected value** (mean, average value or expectation),
- It's the average of all possible values of X, weighted by their probabilities.

Your turn : Theory

Calculate your expected winning in the card game.

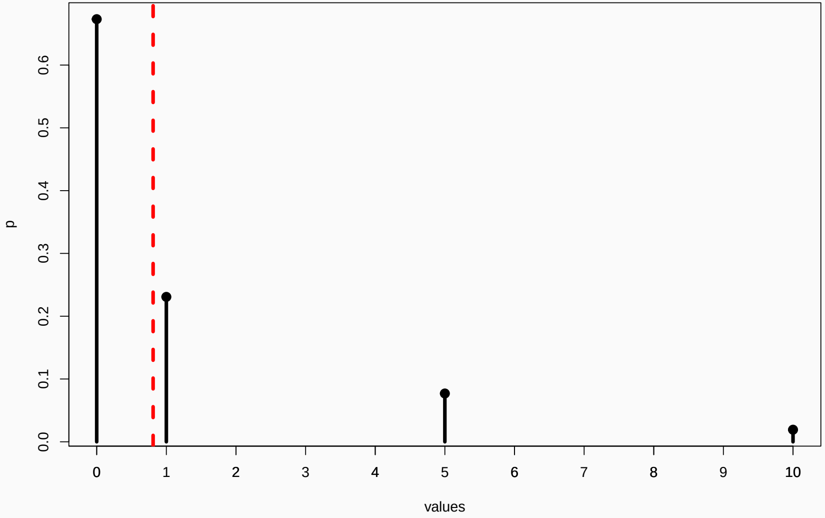
EV of card game from theory

Event	X	$P(X)$	$X\ P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

On average, you expect to make 0.81 in this game.

Note Expected value doesn't need to be one of the values that the variable can take.

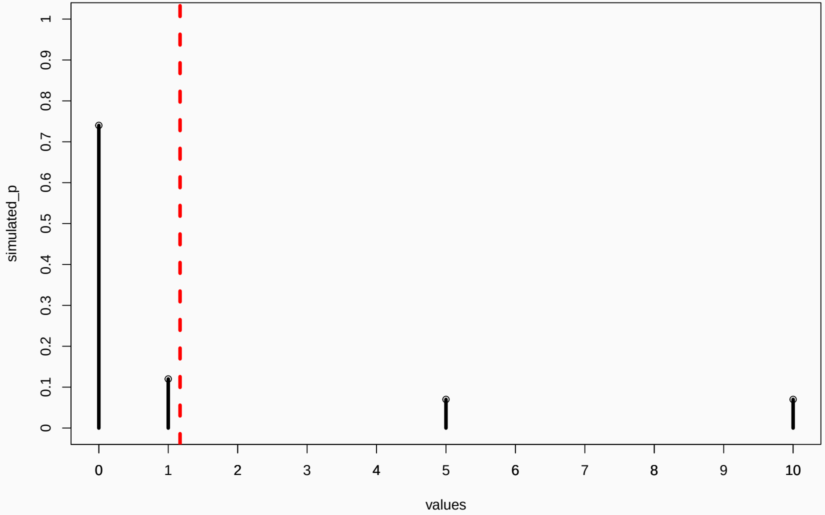
EV of card game from theory




```
## [1] "ev from theory is  0.81 probabilities are  0.67 0.23 0.08 0.02"
```

r

From simulation

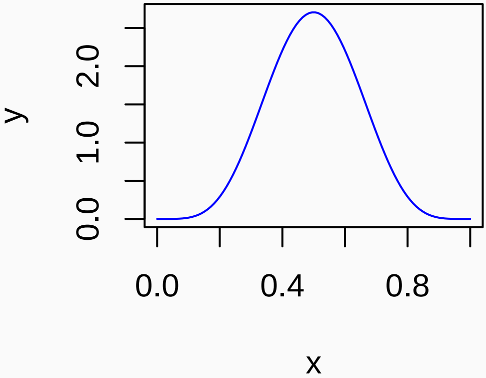



```
## [1] "ev from simulation is  1.17 , simulated proportions are  0.74 0.12 0.07 0.07"
```

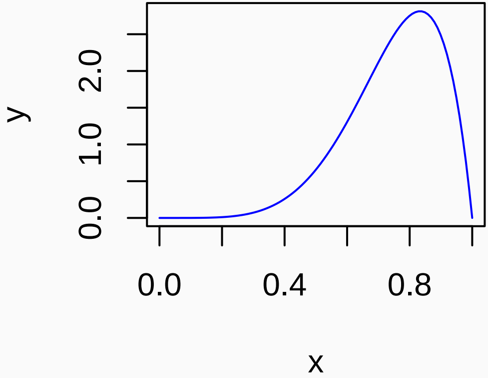
- This is a right skewed distribution since it has a long tail to the right.

Common Distribution Shapes

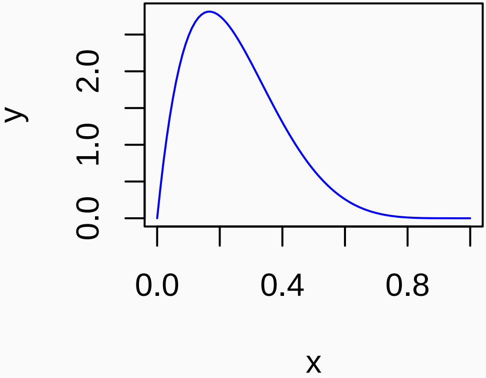
- **Symmetric**:



- **Left-Skewed**:



- **Right-Skewed**:



Variability

We are also often interested in the variability in the values of a random variable (around the mean value of the rv).

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

Variability of a discrete random variable

For the previous card game example, how much is the variability in the winnings?

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

X	$P(X)$	$X\ P(X)$	$(X - E(X))^2$	$(X - E(X))^2 P(X)$
1	$\frac{12}{52}$	$1 \cdot \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \cdot 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \cdot \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \cdot 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \cdot \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \cdot 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \cdot \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \cdot 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$ $SD(X) = \sqrt{3.4246}$ $SD(X) = 1.85$

Interpretation

```
## [1] "EV from theory is  0.81"

## [1] "Variance from theory is  3.42"

## [1] "SD from theory is  1.85"
```

Your typical winnings are somewhere between 0.81 ± 1.85 ie between -1.04 to 2.66

Average of winnings from the 100 simulated games is 1.17

Cumulative distribution function (c.d.f)

For a discrete random variable X, cdf is given by,

$$F(k) = P(X \leq k) = \sum_{x \leq k} P(X = x)$$

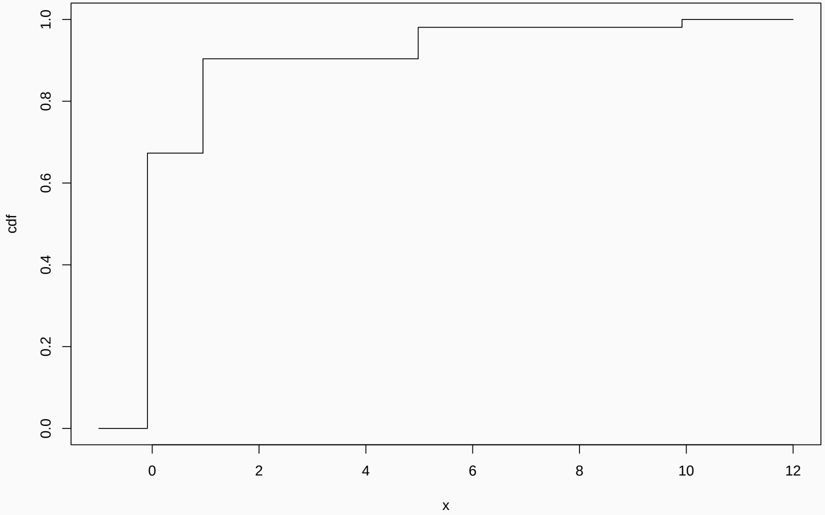
The cumulative distribution function (or CDF) ,$F(k)$ is the probability that the random variable X is at most some particular value k, or no bigger than k ie $P(X \leq k)$

- `cumsum()` function in R executes a cumulative summation element by element

For the previous card game example,

x	0	. 1	5.	. 10.
$P(X = x)$	$35/52 = 0.67$	$12/52 = 0.23$	$4/52 = 0.08$	$1/52 = 0.02$
$P(X \leq x)$	$35/52 = 0.67$	$35/52 + 12/52 = 47/52 = 0.9$	$47/52 + 4/52 = 51/52 = 0.98$	$51/52 + 1/52 = 52/52 = 1$

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 35/52 & \text{if } 0 \leq x < 1 \\ 47/52 & \text{if } 1 \leq x < 5 \\ 51/52 & \text{if } 5 \leq x < 10 \\ 1 & \text{if } x \geq 10 \end{cases}$$



- What is $P(X \leq 7)$? F(7)

Use R

What is $P(X \leq 7)$? ie F(7) using R?

```
values <- c(0,1,5,10)
p <- c(35/52, 12/52, 4/52, 1/52 )
p
```

```
## [1] 0.67307692 0.23076923 0.07692308 0.01923077
```

```
cp <- cumsum(p)
cp
```

```
## [1] 0.6730769 0.9038462 0.9807692 1.0000000
```

```
cp[3]
```

```
## [1] 0.9807692
```

we did:

- Random Variables: Discrete or Continuous
- Discrete Random variables (By hand and using R)
  - P.m.f
  - E(X)
  - V(X)
  - C.d.f