5. Random Variables

Transfer exploration seminar: Statistics and Data Science

Dr. Uma Ravat

PSTAT 194TR

Summary: Basic Probability Theory

- Experiment, Sample space, Events
- Probability - Classical(Frequentist) Definition and Simulation based approach
- Probability Properties and Rules:

1. The probability of an event A, denoted by $P(A)$, is a number between 0 and 1. $0 \leq P(A) \leq 1$
2. For the sample space $S, P(S) = 1$
3. $P(\emptyset) = 0$; $\emptyset$ is the null/empty set containing no elements.
4. Complement: $P(A^c) = 1 - P(A)$
5. Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - **Special case:** If A and B are mutually exclusive events, that is, $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$
6. Multiplication rule: $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$, where $P(B|A)$ is the probability of event B given that event A happened.
    - **Special case:**If A and B are independent, ie $P(B|A) = P(B)$, then $P(A \cap B) = P(A) \times P(B)$

Next we will see...

Random Variables - distribution function and probability calculation - Summarizing random variables - Expected value - Variance

Become familiar with the notation and concepts, the algebra will follow much more easily

Where do random varaibles come from?

Recall, for data(a sample) we said a variable can be

- Numerical - discrete or continuous
- Categorical - ordinal or nominal

When generalizing from a sample to the population

- There's always some uncertainty about the true distributions and relationships in the population
- Probability is the mathematical tool used to measure and express this uncertainty. (PSTAT 120A)
- **random variables** are the mathematical tool that allow us to incorporate uncertainty in the variables in our data.

What is a random variable?

A random variable X assigns a numerical value to each possible outcome (and event) of a random experiment.

Notation/Convention: Capital letters towards the end of the alphabet such as X, Y, Z are used to denote random variables

Corresponding lower case letters x, y, z are used to denote the observed outcomes(observed values/sample values).

Become familiar with the notation and concepts, the algebra will follow much more easily

Example

Experiment: Toss a fair coin

Outcome	H	T
Values: X = x	1	0

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

Probability associated with values of a random variables

Instead of discussing probability of outcomes and events, we can focus on probability of values that a random variable takes.

Example Toss a fair coin

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

Then, we can say that $P(X = 1) = 0.5$, i.e., X is equal to 1 with probability of 0.5

Probability distribution of a random variable

The probability distribution of a random variable specifies its possible values (i.e., its range) and their corresponding probabilities.

$X =$ number of heads in a fair coin toss

As a table

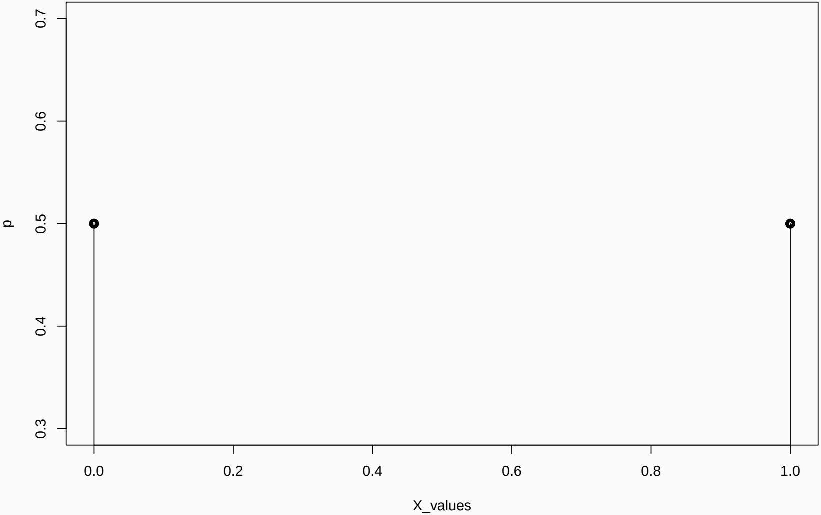
Values: X = x	0	1
Probability: $P(X = x)$	0.5	0.5

As a mathematical function

$$P(X = x) = \begin{cases} 1/2, & \mathrm{X} = 1 \\ 1/2, & \mathrm{X} = 0 \end{cases}$$

Probability distribution of a random variable

As a picture/graph/visualization



$$P(X = x) = \begin{cases} 1/2, & \mathrm{X} = 1 \\ 1/2, & \mathrm{X} = 0 \end{cases}$$

The total probability for the random variable is still 1 and all the probability rules we discussed still hold

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

Your turn : Theory : Getting ready for Las Vegas!

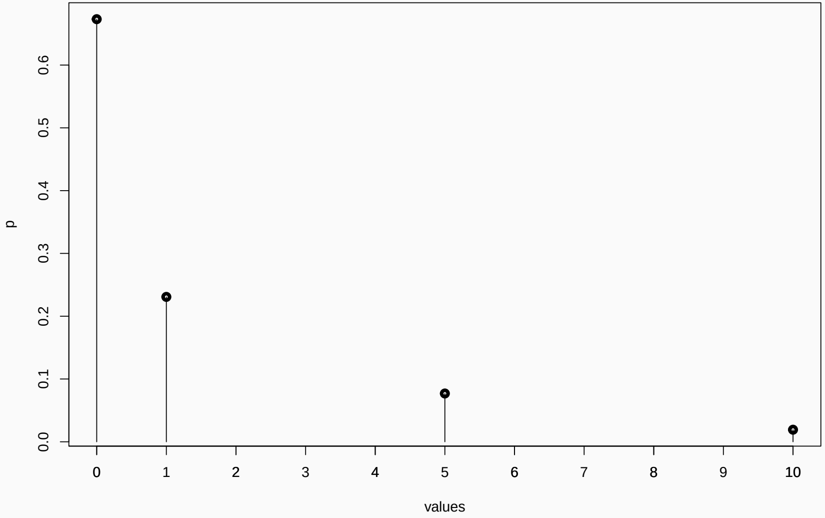
In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability mass function for your winnings.

- (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
```

Would you bet on winning \$5 or \$10 at Las Vegas?

What is the probability that you win either \$5 or \$10?

$X = \$$ you win

$$
\begin{aligned}
P(X = 5 \text{ or } X = 10) &=_{\text{mutually exclusive}} P(X = 5) + P(X = 10) \quad \text{(addition rule)} \\
&= 0.08 + 0.02 \\
&= 0.1
\end{aligned}
$$

You have a 10% chance of winning either \$5 or \$10. Would you do it?

Study Skills: For practice Make up another event wrt this experiment and calculate its probability. Be sure to write your solution out with explanations - Definitions, probability rules you use and how their usage is justified.

Types of random variables

Recall: Types of variables in Statistical Data

For data(a sample) we said a variable can be

- Numerical - discrete or continuous
- Categorical - ordinal or nominal

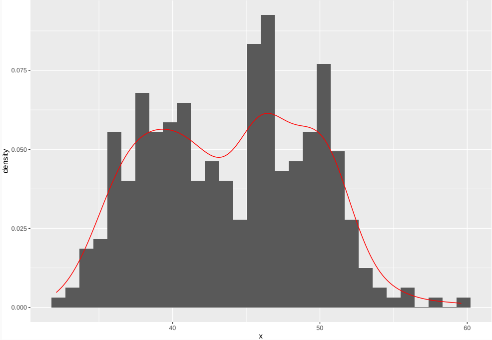
Types of random variables

A **discrete random variable** X has a finite or countably infinite number of possible outcomes. ie we can *count* the number of outcomes

- The distribution function $P(X = x)$ is called a **probability mass function** (pmf)

A **continuous random variable** X takes all values in an interval of real numbers. i.e we can *measure* the outcomes.

From histograms to continuous distributions



If we use histograms to estimate continuous functions that describe all possible outcomes, we have created a probability density function.

Thus the probability for a continuous random variable can now be estimated by the **area under the probability density function/curve** $P(X < 40), P(40 < X < 50)$ etc

Distribution of a continuous RV

- is specified by its **p**robability **d**ensity **f**unction (p.d.f.)
  - pdf gives the relative likelihood of the continuous random variable within the sample space.
  - can be represented by
    - a function $f(x)$, the density function or
    - its graph, the density curve
- the probabilities are given by the area under the graph between specified values.
  - If X is a continuous r.v., then $\mathrm{P}(\mathrm{X} = \mathrm{x}) = 0$ for all values x.
- The total area under a density curve is always equal to 1.

Get used to the notation and the algebra/calculus will follow more easily.

Properties of the Probability Density Function (pdf) - $f(x)$

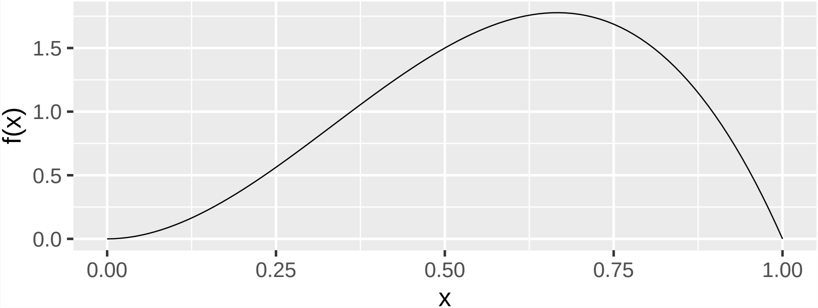
$$f(x) \geq 0 \text{ for all } x \epsilon S$$

$$\int_{x \ \epsilon \ S} f(x)dx = 1$$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

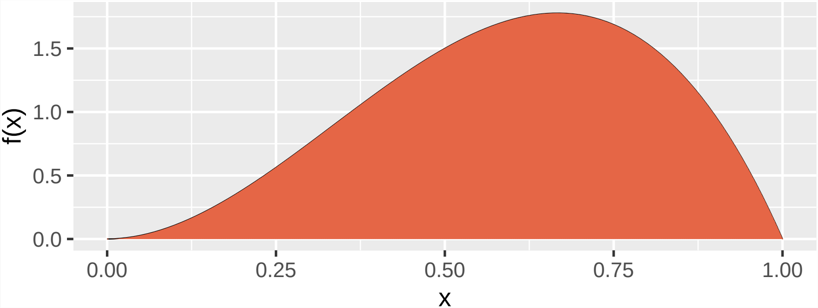
Example of continuous random variable

$$f(x) = \begin{cases} 12\, x^2\, (1 - x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



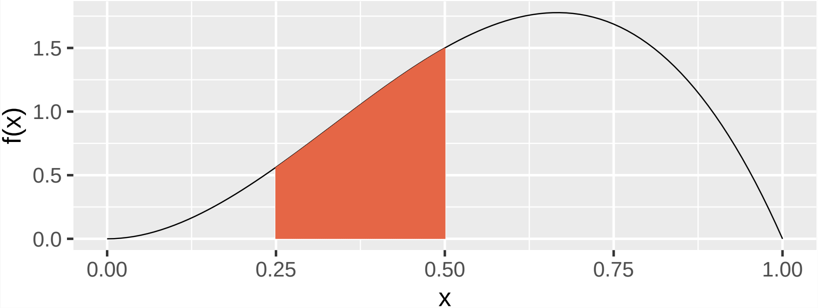
Area under the curve = 1

$f(x) = 12\ (x^2)\ (1-x),\ 0 \leq x \leq 1$



$$\int_{x\ \epsilon\ S} f(x)dx = \int_0^1 12(x^2)(1-x)dx$$
$$= 12\int_0^1 (x^2 - x^3)dx = 12\Big[\frac{x^3}{3} - \frac{x^4}{4}\Big|_0^1\Big] = 1$$

Probability is Area Under the Curve



$$P(0.25 < X < 0.50) = \int_{0.25}^{0.50} 12(x^2)(1-x)dx = 12\int_{0.25}^{0.50}(x^2 - x^3)dx$$

$$= 12\Big[\frac{x^3}{3} - \frac{x^4}{4}\Big|_{0.25}^{0.50}\Big] = 0.2617188$$

Descriptive summary for random variables

From sample measures to analogus population measures

For numerical data in a sample, we calculated **sample mean** $\bar{x}$ and **sample variance** s^2

We calculate analogous measures of center and spread for random variables (discrete, continuous).

The mathematical tools we use are **sums and series** for discrete random variables and **integrals** for continuous random variables.

Measure of center : Mean or average value

	Notation	Formula
Sample	$\bar{x}$	$\frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i \frac{1}{n}$
Discrete RV, pmf $P(X = x)$	$\bar{X} = E(X)$	$\sum_{x_i \ \epsilon \ S} x_i \ P(X = x_i)$
Continuous RV, pdf $f(x)$	$\bar{X} = E(X)$	$\int_{x \ \epsilon \ S} x \ f(x) \ dx$

Measure of spread: variance

	Notation	Formula
Sample	s^2	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Discrete Random Variable	$\sigma^2 = Var(X)$	$\sum_{x_i \ \epsilon \ S}(x_i - \bar{X})^2 \ P(X = x_i)$
Continuous Random Variable	$\sigma^2 = Var(X)$	$\int_{x \ \epsilon \ S}(x - \bar{X})^2 \ f(x) \ dx$

Summary:

- Random Variables: Discrete , Continuous
  - distribution function: pmf $P(X = x)$, pdf $f(x)$
  - probabilities
  - Expected value: $\bar{X}, E(X)$
  - Variance: $\sigma^2, Var(X)$

Become familiar with the notation and concepts, the algebra will follow much more easily

Next we will see...

- Continue working with Random Variables
- Wrap up with a review of Intro to R and Intro to Probability.
- Hear your feedback about these modules.
- Discuss anything else you want.

Pre-reading: Math Review:

- Sum and series
- Integrals