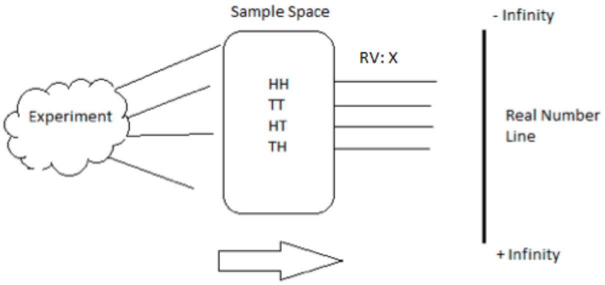
STAT 350 Lecture 3: Random Variables and Probability Modeling

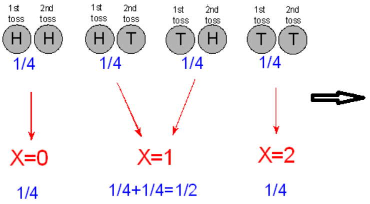
Random Variables and Expectation

(Chapter 3, 4 of WMMY)

Outline

Introduction





List of possible values

Probability of each value

x	0	1	2
P(X=x)	1/4	1/2	1/4

Introduction

- Suppose we perform a random experiment, very often it is not the *sample outcome* that is of interest, but rather some (numerical) characteristic of the sample outcome.
  - E.g. Toss a coin 3 times. $\mathcal{S} = \{\text{HHH, THH, HTH, } \cdots\}$.
  - Not interested in such sequences. Interested in "# of heads" based on sequences, a numerical characteristic of sample outcomes.
- This characteristic is called a *random variable*, random because its value depends on the outcome of the random experiment.
- More precisely, **a random variable is a function of the outcome of the random experiment**.
- An important part of probability theory is the study of random variables.
- **Random variable is a variable which can take potential values.**

Outline

Random variables (RVs)

- Start with a sample space $\mathcal{S}$.
- Let s denote a generic outcome in $\mathcal{S}$.
- A *random variable* (RV) X is a function $X : \mathcal{S} \to \mathbb{R}$ that takes an outcome $s \in \mathcal{S}$ to a real number $X(s) \in \mathbb{R}$.
- Example:
  - A coin is tossed 3 times and the outcome (TTH) is observed. Interested in "# of heads"
  - $s =$ TTH. $X(\text{TTH}) =$ "# of heads of TTH" $= 1$ is just a characteristic of outcome $s =$ TTH.
- We will typically use upper-case letters U, V, W, X, Y, and Z to denote RVs.

Two examples

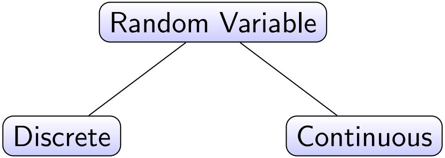
A. Randomly choose a STAT 350 student and record how many siblings he/she has.

- $\mathcal{S} = \{\text{Stat 350 class}\}$
- $X =$ number of siblings
- Possible values of X: 0,1,2,3,...
- e.g., $X(\text{Honglang}) = 1$, since I have one litter brother.

B. Randomly choose a battery from a production line and record its lifetime.

- $\mathcal{S} = \{\text{batteries on the production line}\}$
- $Y =$ lifetime (in hours)
- Possible values of Y: $[0, \infty)$

Discrete vs. continuous RVs



- Like when discussing data, we actually have two types of RVs
- A RV is *discrete* if it takes values in a finite or countable set (e.g., integers)
- A RV is *continuous* if it takes values in an uncountable set (e.g., an interval)
- It is important to make a distinction between these two types of RVs, because the tools used to analyze discrete and continuous RVs are different.
- Example A is discrete and Example B is continuous.

Exercise

Consider a population of college athletes, and choose one at random. *Which of the RVs below are discrete and which are continuous?*

- $V =$ height
- $W =$ weight
- $X =$ number of college credits earned
- $Y =$ grade-point average

Study of Random Variables—Distribution

For understanding of a given random variable X, we need to figure out the **distribution** of this random variable

- Potential variables X can take, i.e. range of X.
- Corresponding probabilities for X to take values in the above range.

Example – coin flipping

Let X = number of Heads when a fair coin is tossed thrice.

- $\mathcal{S} = \{$HHH, HHT, HTH, THH, HTT, THT, TTH, TTT$\}$
- Construct a table of s vs. $X(s)$:

s	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(s)$	3	2	2	2	1	1	1	0

- $\Pr(X = 2) = \Pr(\text{HHT or HTH or THH}) = 3/8$.
- $\Pr(X \leq 1) = \Pr(\text{HTT or THT or TTH or TTT}) = 4/8$.

Outline

Cumulative Distribution Function (CDF)

- The cumulative distribution function (CDF) of a random variable X is

$$F(x) = \Pr(X \leq x), \quad x \in (-\infty, \infty).$$

- Properties of $F(x)$ – why?
  - $F(x)$ is non-decreasing in x.
  - $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
  - $\Pr(a < X \leq b) = F(b) - F(a)$. (More on this later!)
- It turns out that if we know the CDF of X, then we know all there is to know about X!
- But there are related functions that are sometimes easier to work with, and these are specific to the two kinds of RVs.

Probability Mass Function (PMF): Discrete

- For a discrete RV X, the PMF is

$$p(x) = \Pr(X = x),$$

  where x ranges over some countable set, such as the integers.
- By Kolmogorov's 3rd axiom, if A is a countable set, then

$$\Pr(X \in A) = \sum_{x \in A} p(x).$$

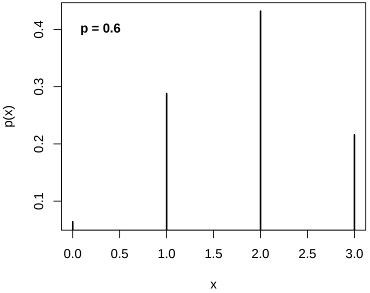
- The PMF can be described by a formula or, **if X takes only finitely many values, by a table**.

Example – coin tossing

Let X denote the number of Heads in three tosses of a coin that lands on Heads with probability p.

$$p(x) = \Pr(X = x) = \binom{3}{x} p^x (1-p)^{3-x}, \quad x = 0, 1, 2, 3.$$

x	0	1	2	3
$p(x)$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3



Example – coin tossing 2

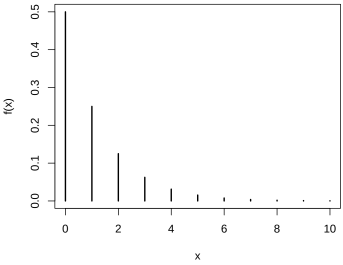
Toss a fair coin until a Heads appears; let X be the number of Tails before the first Heads appears.

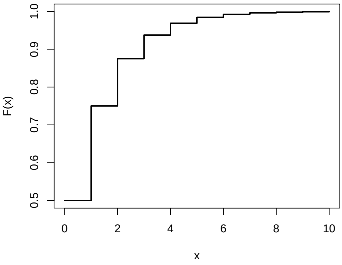
$$p(x) = (1/2)^{x+1} \quad F(x) = 1 - (1/2)^{x+1}, \quad x = 0, 1, 2, \ldots$$

x	Outcome	$p(x)$	$F(x)$
0	H	0.5	0.5
1	TH	0.25	0.75
2	TTH	0.125	0.875
3	TTTH	0.0625	0.9375
4	TTTTH	0.03125	0.96875
$\vdots$			

Question: Can you prove that $\sum_{x=0}^{\infty} p(x) = 1$?

Example – coin tossing 2 (cont.)





Exercises

1. Suppose X is a random variable with CDF

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.25 & \text{for } 0 \leq x < 1 \\ 0.75 & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2 \end{cases}$$

   1. Calculate $\Pr(X \leq 1.2)$.
   2. Calculate $\Pr(X > 0.5)$.
   3. Find the PMF $p(x)$ and plot it.

Exercise

Consider a box containing 10 balls in the following configuration:

$$\Big[2 \text{ Red}, \quad 3 \text{ Blue}, \quad 5 \text{ Green}\Big].$$

Draw two balls from the box at random without replacement. You win \$5 per Red, \$1 per Blue, and \$0 per Green. Let X denote the total winnings.

1. Find the PMF of X.
2. Calculate $\Pr(X \geq 4)$.
3. Find $\Pr(X \geq 4 \mid X > 0)$.

Exercise

Consider a box containing 10 balls in the following configuration:

$$\Big[2 \text{ Red}, \quad 3 \text{ Blue}, \quad 5 \text{ Green}\Big].$$

Draw two balls from the box at random without replacement. You win \$5 per Red, \$1 per Blue, and \$0 per Green. Let X denote the total winnings.

1. Find the PMF of X.
2. Calculate $\Pr(X \geq 4)$.
3. Find $\Pr(X \geq 4 \mid X > 0)$.

x	0	1	2	5	6	10
$p(x)$	$\frac{5}{10}\frac{4}{9}$	$\frac{5}{10}\frac{3}{9}+\frac{3}{10}\frac{5}{9}$	$\frac{3}{10}\frac{2}{9}$	$\frac{2}{10}\frac{5}{9}+\frac{5}{10}\frac{2}{9}$	$\frac{2}{10}\frac{3}{9}+\frac{3}{10}\frac{2}{9}$	$\frac{2}{10}\frac{1}{9}$

Exercise

Consider a box containing 10 balls in the following configuration:

$$\Big[2 \text{ Red}, \quad 3 \text{ Blue}, \quad 5 \text{ Green}\Big].$$

Draw two balls from the box at random without replacement. You win \$5 per Red, \$1 per Blue, and \$0 per Green. Let X denote the total winnings.

1. Find the PMF of X.
2. Calculate $\Pr(X \geq 4)$.$= p(5) + p(6) + p(10) = 34/90$
3. Find $\Pr(X \geq 4 \mid X > 0)$.

x	0	1	2	5	6	10
$p(x)$	$\frac{5}{10}\frac{4}{9}$	$\frac{5}{10}\frac{3}{9} + \frac{3}{10}\frac{5}{9}$	$\frac{3}{10}\frac{2}{9}$	$\frac{2}{10}\frac{5}{9} + \frac{5}{10}\frac{2}{9}$	$\frac{2}{10}\frac{3}{9} + \frac{3}{10}\frac{2}{9}$	$\frac{2}{10}\frac{1}{9}$

Exercise

Consider a box containing 10 balls in the following configuration:

$$\Big[2 \text{ Red}, \quad 3 \text{ Blue}, \quad 5 \text{ Green}\Big].$$

Draw two balls from the box at random without replacement. You win \$5 per Red, \$1 per Blue, and \$0 per Green. Let X denote the total winnings.

1. Find the PMF of X.
2. Calculate $\Pr(X \geq 4)$. $= p(5) + p(6) + p(10) = 34/90$
3. Find $\Pr(X \geq 4 \mid X > 0)$. $= \frac{\Pr(\{X \geq 4\} \cap \{X > 0\})}{\Pr(X > 0)} = \frac{\Pr(X \geq 4)}{1 - \Pr(X = 0)} = \frac{34/90}{1 - 20/90} = 34/70$

x	0	1	2	5	6	10
$p(x)$	$\frac{5}{10}\frac{4}{9}$	$\frac{5}{10}\frac{3}{9} + \frac{3}{10}\frac{5}{9}$	$\frac{3}{10}\frac{2}{9}$	$\frac{2}{10}\frac{5}{9} + \frac{5}{10}\frac{2}{9}$	$\frac{2}{10}\frac{3}{9} + \frac{3}{10}\frac{2}{9}$	$\frac{2}{10}\frac{1}{9}$

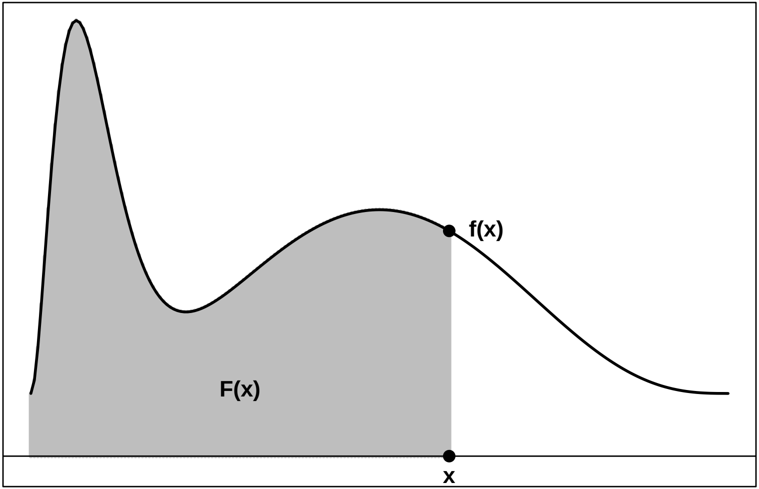
Probability Density Function (PDF): Continuous

- If X is a continuous RV, then there exists a non-negative function $f(x)$, called the PDF, such that

$$\Pr(a < X \leq b) = \int_a^b f(x)\,dx.$$

- Important properties:
  - $\int_{-\infty}^{+\infty} f(x)\,dx = 1$
  - $F(x) = \int_{-\infty}^{x} f(u)\,du$ [See illustration below!]
  - $f(x) = \frac{d}{dx}F(x)$
- NOTE: $f(x) \neq \Pr(X = x)$!!!

Relation between PDF and CDF



Exercises

1. Suppose $F(x) = 1 - e^{-x}$ for $x \geq 0$; $F(x) = 0$ otherwise.
    1. Find $\Pr(X = 2)$.
    2. Calculate $f(x)$.
    3. Compare $f(2)$ with the answer to part (1).
2. Find the value of c such that $f(x)$ is a valid PDF.
    1. $f(x) = c$ for $a \leq x \leq b$; $f(x) = 0$ otherwise.
    2. $f(x) = cx^2$ for $-1 \leq x \leq 1$; $f(x) = 0$ otherwise.
    3. $f(x) = c|x|$ for $-2 \leq x \leq 2$; $f(x) = 0$ otherwise.
3. Suppose X is a continous RV with PDF

$$f(x) = \begin{cases} \frac{2}{5}(x+1) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

    1. Find the CDF $F(x)$.
    2. Find $\Pr(X > 1.75 \text{ or } X < 1.25)$.

Functions of a RV

- Given a RV X it is possible to define new random variables by applying functions to X; that is, if $g(x)$ is some function, then $Y = g(X)$ is a new RV.
- For example, $g(x) = ax + b$ or $g(x) = x^2$ or $g(x) = \cos x$.
- If we know the PMF/PDF of X then, at least in principle, we can find the PMF/PDF of $Y = g(X)$.
- **You are only required to know the case when X is discrete.**

Exercise

Let X be the number showing on the roll of a fair six-sided die. Define a new RV $Y = (X - 3)^2$ and calculate its PMF $p_Y(y)$.

Outline

Expected value

- Intuition: Expected value is average value of the RV over infinitely many repetitions of the experiment.
- Probabilistic interpretation: if we were to guess the value of X *before* the experiment, then $\mathrm{E}(X)$ would be our guess.
- Formal definition:

$$\mathrm{E}(X) = \begin{cases} \sum_x x p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} x f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

- So if you know the PMF/PDF of X, then we can get $\mathrm{E}(X)$.

Exercises

1. Suppose X is a discrete RV with the following PMF

x	0	1	2	3
$p(x)$	0.1	0.2	0.4	0.3

   Find the expected value of X.

2. Suppose X is a continous RV with PDF

$$f(x) = \begin{cases} \frac{2}{5}(x+1) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

   Find $\mathrm{E}(X)$.

More general "expected values"

- Take any function $g(x)$.
- If X is a RV then so is $g(X)$, so it makes sense to consider $\mathrm{E}[g(X)]$.
- The definition is straightforward:

$$\mathrm{E}[g(X)] = \begin{cases} \sum_x g(x)p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} g(x)f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

- If $g(x) = x$ then we get back the previous definition.
- Important fact: **Expected value is linear**; i.e.,

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b.$$

Exercise

Suppose X is a continous RV with PDF

$$f(x) = \begin{cases} \frac{2}{5}(x+1) & \text{if } 1 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

1. Find $\mathrm{E}(X^2)$.
2. Find the value m such that $F(m) = 0.5$. This value m is called the *median*.

Variance

- Mean describes the center of the distribution.
- The variance (or SD) describes the spread.
- Formal definition: If $\mathsf{E}(X) = \mu$, then

$$\mathsf{V}(X) = \mathsf{E}[(X - \mu)^2],$$

  and $\mathsf{SD}(X) = \sqrt{\mathsf{V}(X)}$.
- That is, $\mathsf{V}(X)$ is just $\mathsf{E}[g(X)]$ for $g(x) = (x - \mu)^2$.
- Interpretation: RVs with larger variance have distribution that's more spread out.

Properties of Variance

- $\mathrm{V}(X) \geq 0$, and $\mathrm{V}(X) = 0$ iff $\Pr(X = \mathrm{E}(X)) = 1$
- $\mathrm{V}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.
- Properties 1 and 2 imply that $\mathrm{E}(X^2) \geq [\mathrm{E}(X)]^2$.
- If a and b are two numbers, then

$$\mathrm{V}(aX + b) = a^2\mathrm{V}(X) \quad \text{and} \quad \mathrm{SD}(aX + b) = |a|\mathrm{SD}(X).$$

Exercises

1. Suppose X is a discrete RV with PMF

x	0	1	2	3
$f(x)$	0.1	0.2	0.4	0.3

   Find $\mathrm{V}(X)$, and $\mathrm{V}(2X - 3)$.
2. Suppose $\mathrm{E}(Y) = 2$ and $\mathrm{Var}(Y) = 6$. Find $\mathrm{E}(Y^2)$ and $\mathrm{E}[(Y - 1)^2]$.

Outline

Definition

- Intuitively, a set of RVs $X_1, \dots, X_n$ are independent if the observed value of X_i has nothing to do with the observed value of X_j ($i \neq j$).
- For example, we expect that the weather tomorrow in the UK is independent of the height of a randomly chosen student on the IUPUI campus.
- Formal defintion: A set of RVs $X_1, \dots, X_n$ are independent if

$$\Pr(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^{n} \Pr(X_i \in A_i),$$

  for any sets $A_1, \dots, A_n$. We say $X_1, \dots, X_n$ are independent and identically distributed (I.I.D.) if they're independent and they all have the same PMF/PDF.

Properties

Suppose $X_1, \dots, X_n$ are indepenent RVs (not necessarily I.I.D.).

- $\mathrm{E}(X_1 + \cdots + X_n) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)$. [Independence is not necessary here!]
- $\mathrm{V}(X_1 + \cdots + X_n) = \mathrm{V}(X_1) + \cdots + \mathrm{V}(X_n)$. (Give an example of *dependent* RVs X and Y for which $\mathrm{V}(X + Y) \neq \mathrm{V}(X) + \mathrm{V}(Y)$.)

Proofs are not hard, but require a notion of joint PMF/PDF which we will not cover in STAT 350.

Important exercise

Suppose $X_1, \ldots, X_n$ are I.I.D. RVs with common mean μ and common variance σ^2. Define the sample mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Find the expected value and variance of $\overline{X}$.