Lecture 02: Random Variables and Distributions

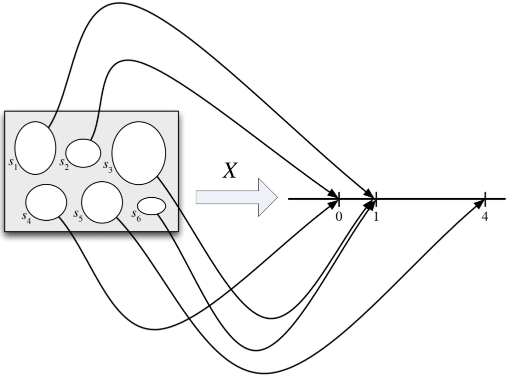
Mathematical Statistics I, MATH 60061/70061

Thursday September 2, 2021

Reference: Casella & Berger, 1.4-1.6

Random variable

Definition: Given an experiment with sample space $S = \{s_1, \ldots, s_n\}$ with a probability function P, a **random variable** (r.v.) is a function from the sample space S to the real numbers $\mathbb{R}$.



Random variable

Definition: Given an experiment with sample space $S = \{s_1, \ldots, s_n\}$ with a probability function P, a **random variable** (r.v.) is a function from the sample space S to the real numbers $\mathbb{R}$.

- A random variable X assigns a numerical value $X(s)$ to each possible outcome $s \in S$.
- The randomness comes from the fact that we have a *random experiment*; the mapping itself is *deterministic*.

Probability function with random variable

Given a random variable X and a subset A of the real line, define $X^{-1}(A) = \{s \in S : X(s) \in A\}$, the probability function with the random variable X, P_X is given by

$$P_X(X \in A) = P(X^{-1}(A)) = P(\{s \in S : X(s) \in A\})$$
$$P_X(X = x) = P(X^{-1}(x)) = P(\{s \in S : X(s) = x\})$$

- Verify that P_X satisfies the Axioms of Probability
- Because of the equivalence, we will simply write $P(X = x)$ rather than $P_X(X = x)$

Example: coin tosses

Consider an experiment where we toss a fair coin twice. The sample space is $S = \{HH, HT, TH, TT\}$. Some random variables:

- Let X be the number of Heads

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

- Let Y be the number of Tails (note that Y and $2 - X$ are the same r.v.)

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

- Let I be 1 if the first toss lands Heads and 0 otherwise.

$$I(HH) = 1, I(HT) = 1, I(TH) = 0, I(TT) = 0$$

Discrete random variable

Two main types of random variables used in practice:

- Discrete random variables
- Continuous random variables

A random variable X is said to be *discrete* if there is a finite list of values $a_1, a_2, \ldots, a_n$ or an infinite list of values $a_1, a_2, \ldots$ such that $\sum_j P(X = a_j) = 1$.

If X is a discrete r.v., then the finite or countably infinite set of values x such that $P(X = x) > 0$ is called the *support* of X.

Distributions

The **distribution** of a random variable specifies the probabilities of *all events* associated with the r.v.

- $P(X = 10)$
- $P(X > 100)$
- $\cdots$

For a discrete r.v., the most natural way to describe its distribution is using the probability mass function.

Probability mass functions

The **probability mass function** (PMF) of a discrete random variable X is the function f_X given by

$$f_X(x) = P(X = x) \quad \text{for all } x.$$

The PMF is positive if x is in the support of X (e.g., $\{x_1, x_2, \ldots\}$), and 0 otherwise.

A valid PMF f_X must satisfy two criteria:

- Nonnegativity: $f_X(x) \geq 0$ if $x = x_j$ for some j, and $f_X(x) = 0$ otherwise;
- Sums to 1: $\sum_{j=1}^{\infty} f_X(x_j) = 1$.

Example: two fair coin tosses, continued

- Let X be the number of Heads

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

- The PMF of X:

$$\begin{aligned} f_X(0) &= P(X = 0) = P(\{TT\}) = 1/4 \\ f_X(1) &= P(X = 1) = P(\{HT, TH\}) = 1/2 \\ f_X(2) &= P(X = 2) = P(\{HH\}) = 1/4 \end{aligned}$$

and $f_X(x) = 0$ for all other values of x.

Example: two fair coin tosses, continued

- Let Y be the number of Tails

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

- The PMF of Y:

$$\begin{aligned} f_Y(0) &= P(Y = 0) = P(\{HH\}) = 1/4 \\ f_Y(1) &= P(Y = 1) = P(\{HT, TH\}) = 1/2 \\ f_Y(2) &= P(Y = 2) = P(\{TT\}) = 1/4 \end{aligned}$$

and $f_Y(y) = 0$ for all other values of y.

Example: two fair coin tosses, continued

- Let I be 1 if the first toss lands Heads and 0 otherwise.

$$I(HH) = 1, I(HT) = 1, I(TH) = 0, I(TT) = 0$$

- The PMF of I:

$$f_I(0) = P(I = 0) = P(\{TH, TT\}) = 1/2$$
$$f_I(1) = P(I = 1) = P(\{HH, HT\}) = 1/2$$

and $f_I(i) = 0$ for all other values of i.

Cumulative distribution functions

Besides the PMF, another function that describes the distribution of a random variable is the cumulative distribution function.

Definition: The **cumulative distribution function** (CDF) of a random variable X is the function F_X given by $F_X(x) = P(X \leq x)$, for all x.

Cumulative distribution functions

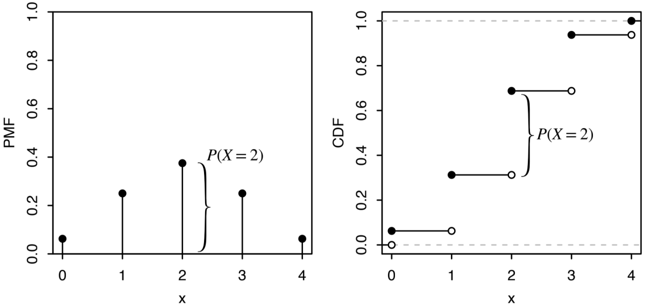
Besides the PMF, another function that describes the distribution of a random variable is the cumulative distribution function.

Definition: The **cumulative distribution function** (CDF) of a random variable X is the function F_X given by $F_X(x) = P(X \leq x)$, for all x.

- We sometimes drop the subscript and just write F for a CDF, when there is no risk of ambiguity.
- The PMF is generally easier to work with for discrete r.v.s, since evaluating the CDF requires a summation.
- The CDF is defined for all r.v.s; on the other hand, only discrete r.v.s possess the PMF.

Conversion from PMF to CDF

Flip a fair coin four times. Let X be the number of Heads.

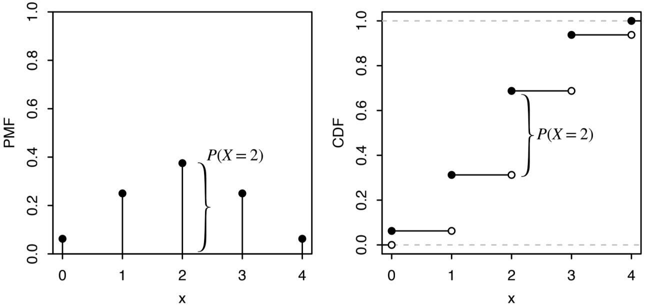


To find $F_X(1.5)$:

$$P(X \leq 1.5) = P(X = 0) + P(X = 1) = \left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^4 = \frac{5}{16}$$

Conversion from CDF to PMF

Flip a fair coin four times. Let X be the number of Heads.



- The CDF of a discrete r.v. consists of jumps and flat regions.
- The height of a jump in the CDF at x is equal to the value of the PMF at x.

Valid CDFs

A function F mapping the real line to $[0, 1]$ is a CDF for some probability P if and only if F satisfies the following three conditions:

1. F is non-decreasing: $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.
2. F is right-continuous: $F(x) = F(x^+)$ for all x, where

$$F(x^+) = \lim_{\substack{y \to x \\ y > x}} F(y).$$

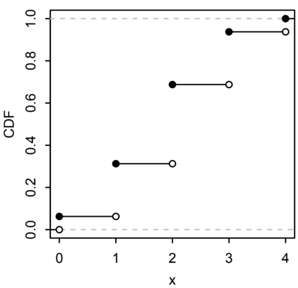
3. F converges to 0 and 1 in the limits:

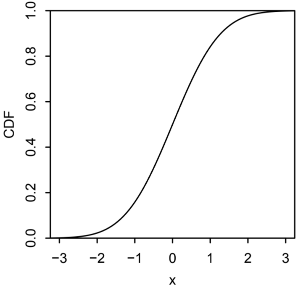
$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1.$$

Continuous random variables

A random variable has a **continuous distribution** if its CDF is *differentiable*. We allow there to be endpoints (or finitely many points) where the CDF is continuous but not differentiable, as long as the CDF is differentiable everywhere else.

A **continuous random variable** is a random variable with a continuous distribution, where $P(X = x) = 0$ for all x.





Probability density function

For a continuous random variable X with CDF F, the **probability density function** (PDF) of X is the derivative of the CDF, given by $f(x) = F'(x)$. The *support* of X, and of its distribution, is the set of all x where $f(x) > 0$.

A valid PDF f must satisfy the following two criteria:

- Nonnegative: $f(x) \geq 0$
- Integrates to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

PDF, PMF, and probability

PDF vs. PMF

- For a PDF f, the quantity $f(x)$ is not a probability.
- It is possible to have $f(x) > 1$ for some values of x.

For a continuous R.V. X, $P(X = a) = 0$. The probability of X being *very close* to a

$$P(a - \epsilon/2 < X < a + \epsilon/2) = \int_{a-\epsilon/2}^{a+\epsilon/2} f(x)dx \approx f(a)\epsilon$$

To obtain a probability, we need to *integrate* the PDF (*density*).

PDF to CDF

Let X be a continuous R.V. with PDF f. Then the CDF of X is given by

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

The results is immediate from the fundamental theorem of calculus. By definition of PDF, F is an antiderivative of f, so

$$\int_{-\infty}^{x} f(t)dt = F(x) - F(\infty) = F(x).$$

The CDF is the *accumulated area* under the PDF. In the discrete case, we obtain the value of a discrete CDF at x by summing the PMF over all values less than or equal to x.

Probability of a continuous R.V. within a range

By definition of CDF and the fundamental theorem of calculus,

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

For continuous R.V.s, $P(X = a) = P(X = b) = 0$, so

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$$

To get a desired probability, integrate the PDF over the appropriate range.

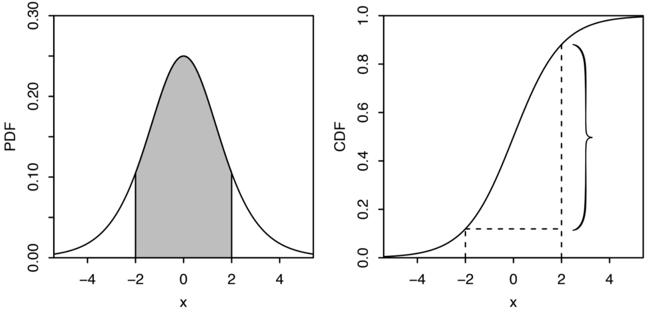
Logistic distribution

The Logistic distribution has CDF

$$F(x) = \frac{e^x}{1+e^x}, \quad x \in \mathbb{R}.$$

Differentiating the CDF gives the PDF

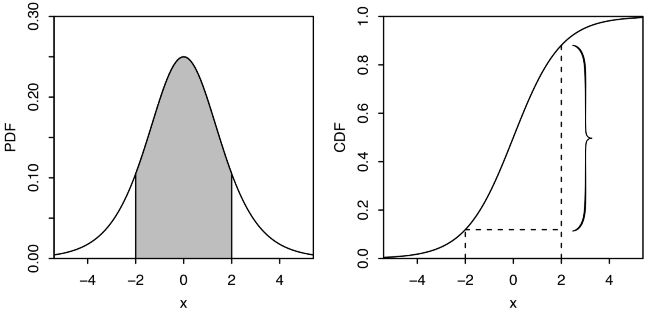
$$f(x) = \frac{e^x}{(1+e^x)^2}, \quad x \in \mathbb{R}.$$



Logistic distribution

Let $X \sim \text{Logistic}$,

$$P(-2 < X < 2) = \int_{-2}^{2} \frac{e^x}{(1+e^x)^2} dx = F(2) - F(-2) \approx 0.76.$$



CDFs and random variables

- A random variable X is **continuous** if $F_X(x)$ is a **continuous function** of x.
- A random variable X is **discrete** if $F_X(x)$ is a **step function** of x.

Identically distributed random variables

F_X completely determines the probability distribution of a random variable X.

The following two statements are equivalent:

1. The random variables X and Y are **identically distributed**.
2. $F_X(x) = F_Y(x)$ for every x.

Two random variables that are identically distributed are *not* necessarily equal.

Consider the experiment of tossing a fair coin twice, and let X be the number of Heads and Y be the number of Tails.

$$F_X = F_Y, \quad X \neq Y$$