Econ 1620: Introduction to Econometrics

Week 3, Lecture 3/Week 4, Lecture 1

Random variables and probability distributions -- Introduction

[LM 3.1, 3.3, 3.4 & SW 2.1]

Dimitra Politi -- Brown University -- Spring 2015

Random variables

- So far, we have described a random process in terms of a collection of possible outcomes, and we have discussed ways to assign probabilities to outcomes and events.

- The methodology we developed is particularly useful when each basic outcome of the sample space is equally likely, and there is a finite (or countably infinite) number of possible outcomes. This is the case for many games of chance.

- Starting today, we will look at other, useful ways of assigning probabilities to outcomes of random processes. This will involve "redefining" the sample space using **functions** known as random variables.

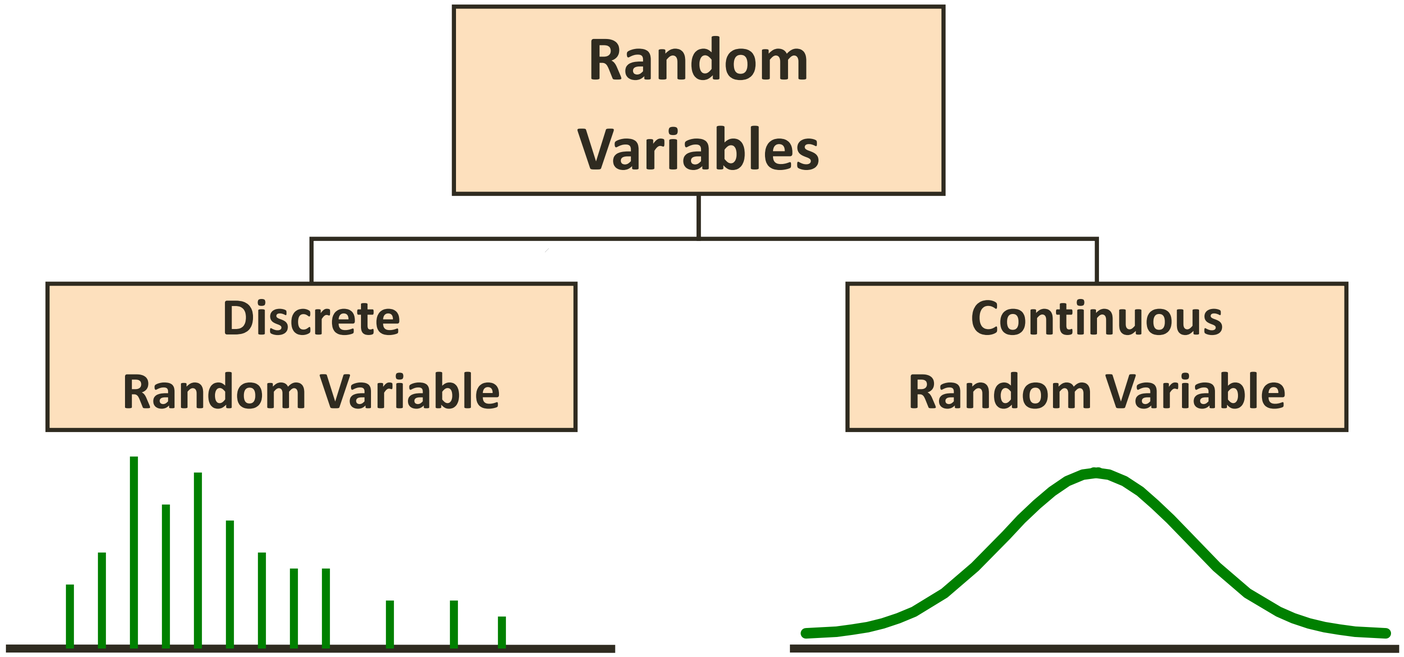
- **Random variable (or stochastic variable):** variable whose numerical value is determined by the outcome of a random process.

Random variables (continued)

- Note that we can describe outcomes of random processes verbally (e.g. girl or boy, heads or tails), but it's easier to describe them numerically.
  - Example #1: Girl: outcome 1, Boy: outcome 0.
  - Example #2: Tails: outcome 1, Heads: outcome 0

- A random variable is a <u>real-valued function defined on a sample space</u>.
  - Example #1: Number of allergic reactions to an experimental drug tested on 10 volunteers.
  - Example #2: Height among sophomores.
  - Example #3: The number of times your computer crashes while you write a paper.

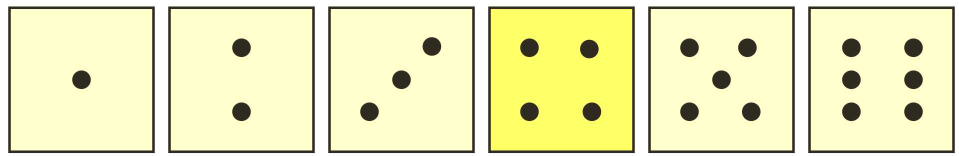
Discrete and continuous random variables



The biggest distinction among random variables is whether their values are countable (discrete random variables), or whether they occur in an interval (continuous random variables).

Discrete random variables

Discrete random variables can only take on **a countable number of values**.



Examples:

- Roll a die twice, and let random variable X be the number of times that 4 comes up. $X = \{0,1,2\}$
- Let Y be the number of female students who get an A in Econ 1620.
  $Y\{0,1,2, \dots, number\ of\ \ female\ students\ in\ Econ\ 1620\}$
- Toss a coin 5 times, and let Z be the number of heads. $Z = \{0,1,2,3,4,5\}$

Probability mass function (PMF)

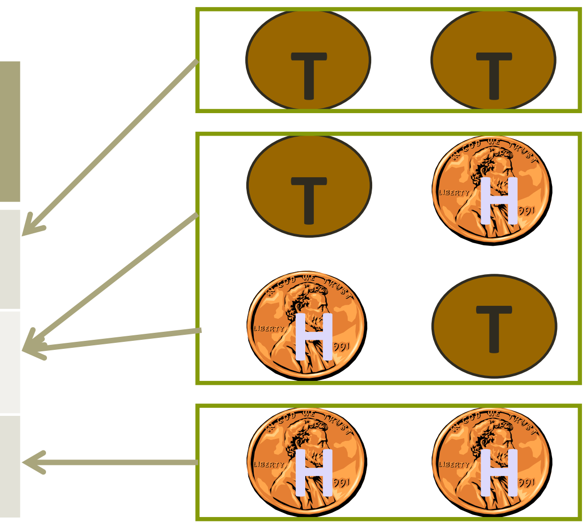
- The best way to describe a discrete random variable is to use its probability mass function (PMF), or probability distribution.
- The PMF lists <u>all possible values</u> of the random variable, and the <u>probability</u> that each value will occur. We often symbolize it by $f(x) = \Pr(X = x)$. That is, $f(x)$ is the probability that our random variable X takes on the value x.

- <u>Example</u>: let's toss 2 coins and let X be the number of heads. What is the PMF of X? X can take on 3 possible values: 0, 1, and 2.

There are 4 possible outcomes for this random process:
$\{(T,T),(T,H),(H,T),(H,H)\}$

The PMF of X is:

Value of X	Probability $f(x)$
0	$1/4 = 0.25$
1	$2/4 = 0.50$
2	$1/4 = 0.25$

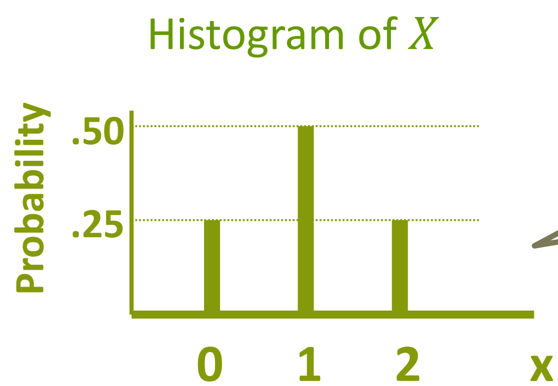


Probability Distribution

- A probability distribution needs to satisfy some **properties**:

1. $f(x) = P(X = x) \geq 0$ for all x. This is not a surprise, given that $f(x)$ is a probability!
2. $\sum_x f(x) = 1$: the sum across all possible values of X is equal to 1.

- The probability distribution is very useful, because using it we can calculate probabilities of events: $P(A) = \sum_{x \in A} f(x)$.
  - Using the previous example: $P(X \geq 1) = f(1) + f(2) = 0.75$
- Graphically, we show a probability distribution using a histogram.

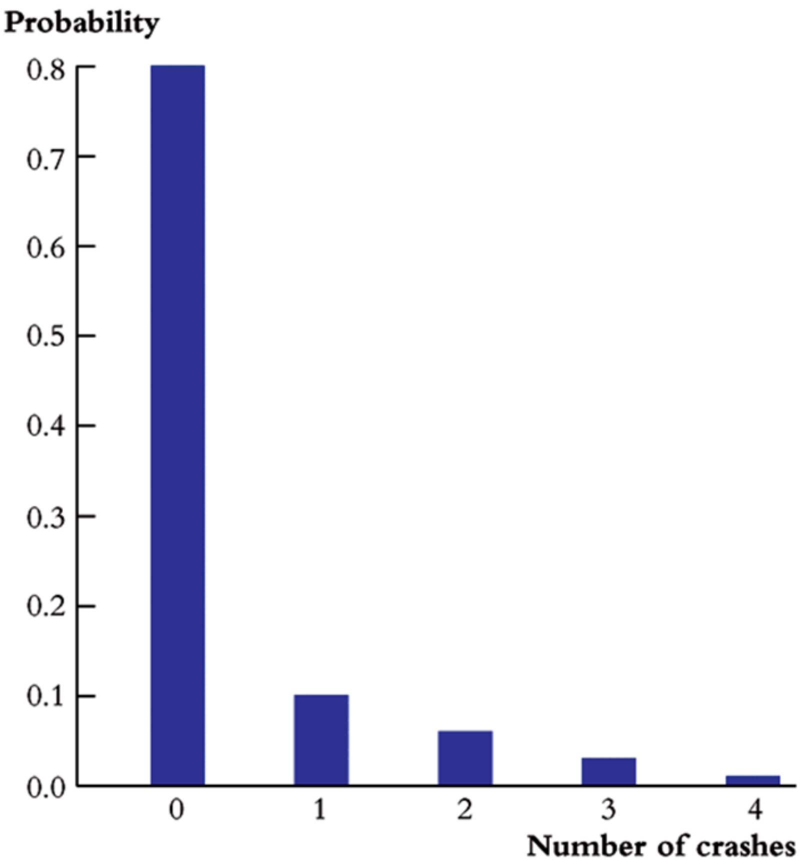


Histogram from previous example

A simple histogram

FIGURE 2.1 **Probability Distribution of the Number of Computer Crashes**

The height of each bar is the probability that the computer crashes the indicated number of times. The height of the first bar is 0.8, so the probability of 0 computer crashes is 80%. The height of the second bar is 0.1, so the probability of 1 computer crash is 10%, and so forth for the other bars.



Cumulative Distribution Function

- Another way to describe a (discrete or continuous) random variable X is to use its cumulative distribution function (cdf), which we usually symbolize by $F(x)$. The cdf expresses the probability that X does not exceed a certain value x:

$$\boxed{F(x) = P(X \leq x) \; for \; -\infty < x < \infty}$$

- Using the cdf, we can determine the probability that X will lie in a specified interval:

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

- Also note: $P(X > x) = 1 - F(x)$

TABLE 2.1 Probability of Your Computer Crashing *M* Times

	Outcome (number of crashes)				
	0	1	2	3	4
Probability distribution	0.80	0.10	0.06	0.03	0.01
Cumulative probability distribution	0.80	0.90	0.96	0.99	1.00

Properties of the cdf

Any cdf (for both discrete and continuous random variables) always satisfies the following conditions:

- $0 \leq F(x) \leq 1$
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- $F(x)$ is a non-decreasing function of x: if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$
- $F(x)$ is continuous from the right

CDF and PMF for discrete random variables

- The cdf and pmf of a discrete random variable X can be generated from each other:
  - If we have the pmf of X, we can generate the cdf by recursion.
  - If we have the cdf of X, we can generate the pmf by subtraction.

- For example: suppose that random variable X can take on the values $x_1 < x_2 < \cdots < x_K$. Then:
  - $f(x_1) = F(x_1)$
  - For $i = 2, \dots, K$:

$$F(x_i) = F(x_{i-1}) + f(x_i) = \sum_{j=1}^{i} f(x_j)$$

$$f(x_i) = F(x_i) - F(x_{i-1})$$

Examples of discrete probability distributions

- **Bernoulli**: a one-time "yes/no" trial, $X = 1$ if "yes", $X = 0$ if "no"
  - $\Pr(X = 1) = p, \Pr(X = 0) = 1 - p.$
  - p is the probability of "success" ("yes", $X = 1$).
  - Example: Getting tails if you toss a coin, or getting 6 if you roll a die once.

- **Uniform**: K possible outcomes/values, each equally likely (random)
  - $\Pr(x) = \frac{1}{K}, x = 1,2,3, \dots, K.$
  - Example: the winning number in a daily lottery.
  - The uniform distribution has both a discrete and a continuous version.

- **Binomial**: X is the number of "successes" after n independent trials of a Bernoulli process with p probability of success.
  - $\Pr(x) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x}$
  - The Bernoulli distribution is the Binomial distribution with $n = 1$.
  - Examples: Factory line with p probability of producing non-defective item, clinical trial

Continuous random variables

- A continuous random variable is a variable that can assume **any value in an interval on the real line, or in a collection of intervals.**
  - time required to complete a task
  - temperature of a solution
  - height

- These variables can potentially take on any value within the interval, depending only on the ability to measure accurately.

- That's why it is not possible to talk about the probability that our random variable takes a particular value: $P(X = a) = 0$ by definition. This doesn't mean that $X = a$ is impossible! It means that the probability is distributed so thinly that you can only see it on sets like intervals.

- Instead, we talk about the probability of the random variable taking values within an interval.

Probabilities for a continuous random variable

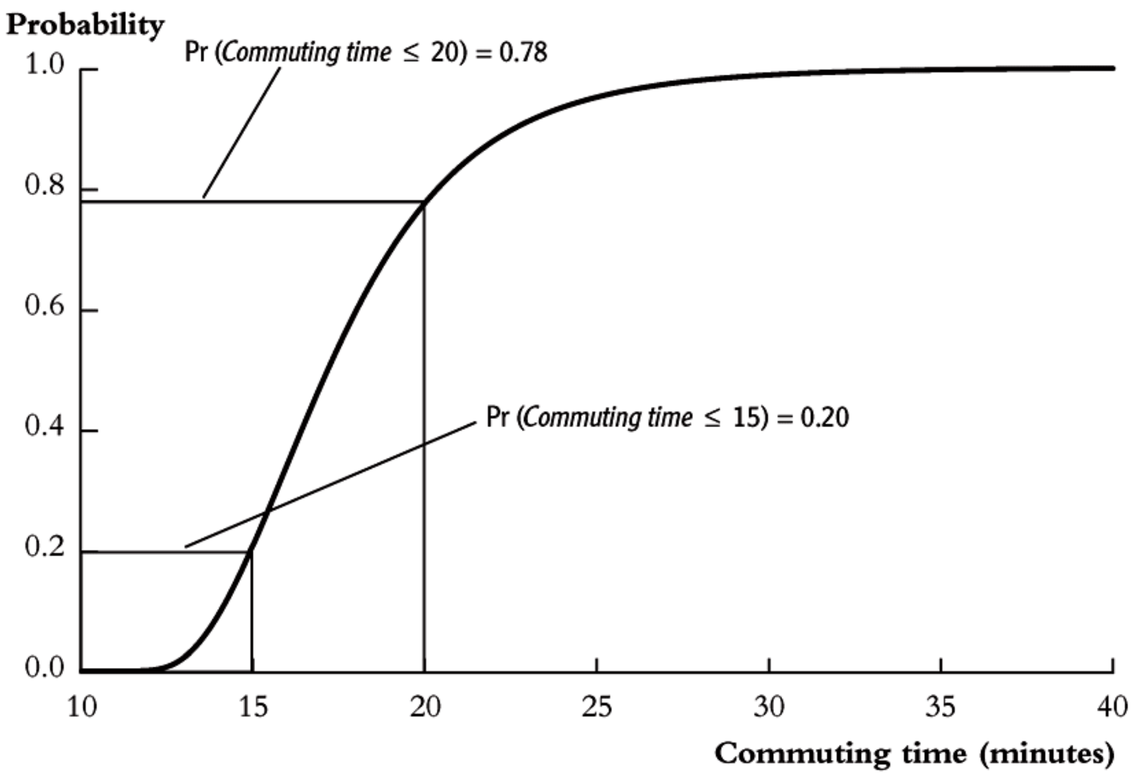
- Because continuous random variables do not have probabilities attached to specific values, but rather to intervals of values, they do not have probability mass functions like discrete random variables.

- Instead, to describe a continuous random variable, we use its cumulative distribution function $F(x)$, and its first derivative, the probability density function (pdf), which we symbolize by $f(x)$. Note that $f(x)$, in the case of a continuous random variable X, is **not** the probability that $X = x$!

- The cdf has the same meaning and utility here that it does with discrete random variables: $F(x) = P(X \leq x)$

CDF for a continuous random variable

FIGURE 2.2 Cumulative Distribution and Probability Density Functions of Commuting Time



(a) Cumulative distribution function of commuting time

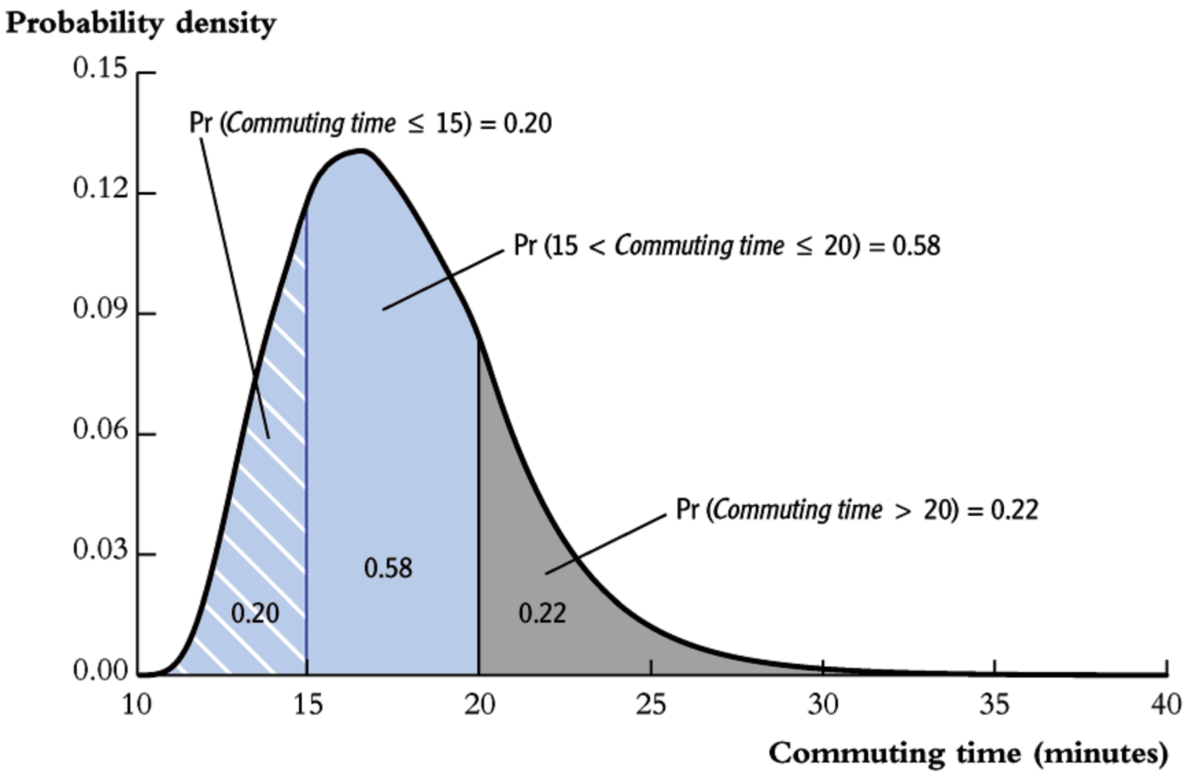
Probability Density Function

- Let X be a continuous random variable with a cdf given by $F(x)$. The probability density function $f(x)$ is the first derivative of the cdf:

$$f(x) = \frac{dF(x)}{dx}$$

- The pdf $f(x)$ of random variable X has the following properties:
  - $f(x) \geq 0 \; for \; all \; x$: $f(x)$ is an non-negative function
  - $\int_{-\infty}^{\infty} f(x)dx = 1$: the area under the probability density function over all values of x is equal to 1.
  - $P(a < X < b) = \int_{a}^{b} f(x)dx$: The probability that X takes values within the interval (a, b) is the area under the density function graph between a and b. [Remember that integrals are "continuous sums".]

PDF for a continuous random variable

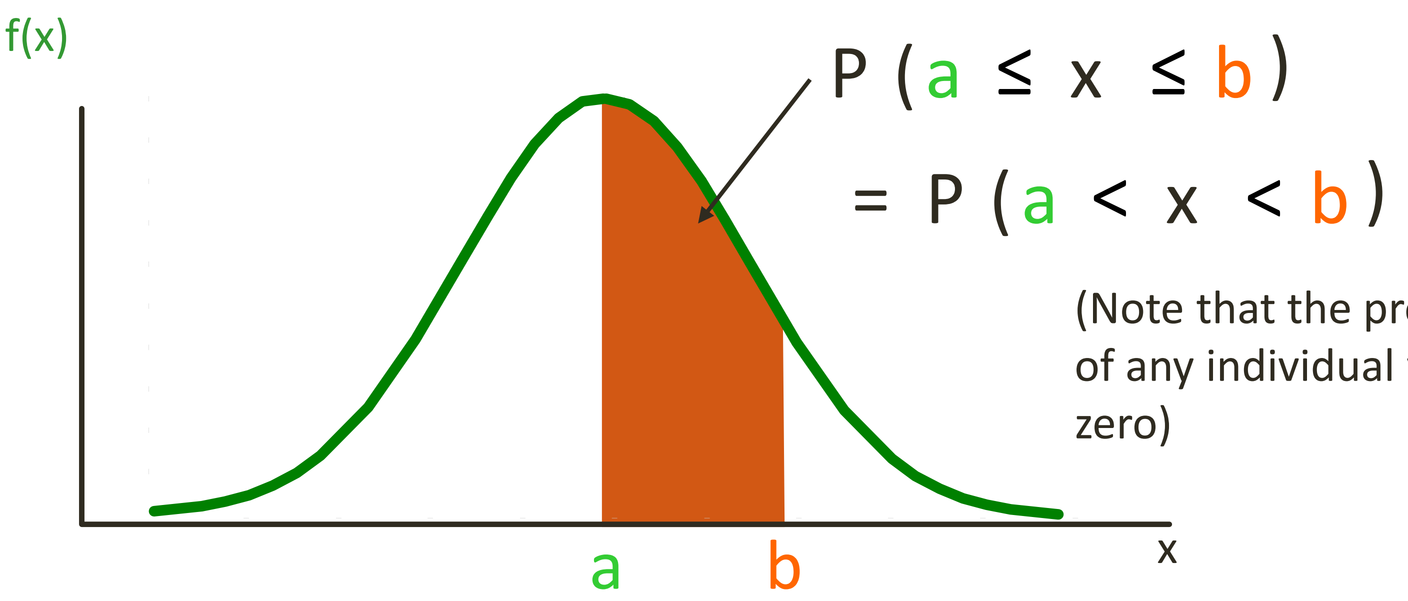


(b) Probability density function of commuting time

Probability as an area

The shaded area under the curve is the probability that X is between a and b.

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



(Note that the probability of any individual value is zero)

CDF and PDF for continuous random variables

- The probability density function $f(x)$ is the first derivative of the cdf:

$$f(x) = \frac{dF(x)}{dx}$$

- Alternatively, we can express the cumulative distribution function $F(x_0)$ of a continuous random variable as the area under the probability density function from the minimum x value up to x_0:

$$F(x_0) = \int_{-\infty}^{x_0} f(x)dx$$

Examples of continuous probability distributions

- **Uniform distribution on an interval**:
  - A point X must be selected from an interval S such that the probability that X belongs to a sub-interval of S is proportional to the length of that sub-interval.
  - We can say that a point is chosen at random from the interval (a, b).
  - $f(x) = \frac{1}{b-a}, for\ a \leq x \leq b, and\ 0\ otherwise$

- **Normal (or Gaussian) distribution**: we will talk in length about this family of distributions in later lectures! Its pdf is a bell-shaped curve, and we use it to describe many random processes (e.g. IQ in the population).