ISI-BUDS
PROBABILITY - PART III

Veronica J. Berrocal

July 11, 2022

PART III PLAN:

- Continuous random variables
- Probability and Cumulative Density Functions (PDFs and CDFs)
- Some common distributions for continuous random variables

CONTINUOUS RANDOM VARIABLES

- We have seen that a continuous random variable is a variable for which the outcome can be any value in an interval or collection of intervals.
- By definition, the probability that a continuous random variable is equal to any specified value is 0. For a continuous random variable X, we are only able to find the probability that X falls between two values.

PROBABILITY DENSITY FUNCTION (P.D.F.)

- For a continuous random variable, we do not have a probability distribution function, but we have a probability density function (p.d.f.), which is used to find the probability that the random variable falls into a specified interval of values.
- The p.d.f. for a random variable X is a curve such that the area under the curve over an interval equals the probability that X is in that interval.
- The area under the curve for the entire range of possible values is equal to 1.

NOTATIONS FOR A CONTINUOUS RANDOM VARIABLE

- When calculating probabilities for a continuous random variable we will use the following notation:
  - X denotes the random variable
  - The two endpoints of an interval are denoted by a and b and the interval of values of X that fall between a and b is indicated with $a \leq X \leq b$.
  - The probability that X has a value between a and b is written as $P(a \leq X \leq b)$.

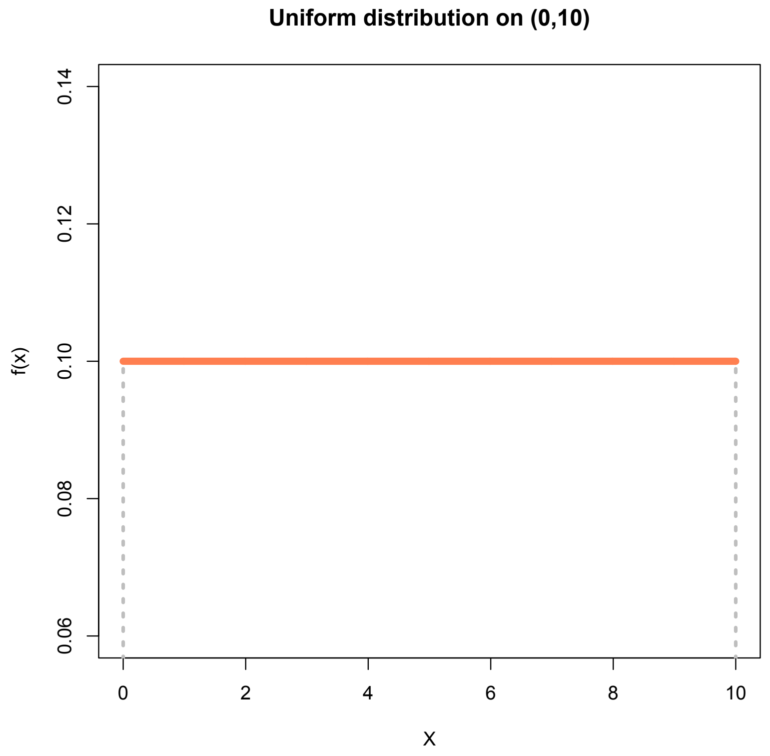
EXAMPLE: BUS WAITING TIME

A bus arrives at a bus stop every 10 minutes. If a person arrives at the bus stop at a random time, how long will he or she have to wait for the next bus?

If we call X the random variable "waiting time until the next bus arrives", the value of X could be any number between 0 and 10 minutes, hence X is a continuous random variable.

EXAMPLE: BUS WAITING TIME

- The p.d.f. for the random variable X, "bus waiting time", is shown on the right.
- The p.d.f. is a flat line that covers the interval between 0 and 10.
- A p.d.f. that is flat assigns the same probability to all intervals that have the same width. Because of this, such p.d.f. are called uniform distribution, and a random variable with a uniform p.d.f. is called a Uniform random variable.



UNIFORM P.D.F.

- If a Uniform random variable can assume values between a and b, then, since the Uniform p.d.f. must have an area under the curve of 1, this means that the Uniform p.d.f. has:
  - height equal to ..... between a and b; and
  - height equal to ...... elsewhere.

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

- Hence, a uniform p.d.f. has expression:

EXPECTED VALUE AND VARIANCE

- If X is a uniform random variable, then the expected value and the variance of X are given by:

$$\boxed{E(X) = \frac{b-a}{2}}$$

$$\boxed{Var(X) = \frac{(b-a)^2}{12}}$$

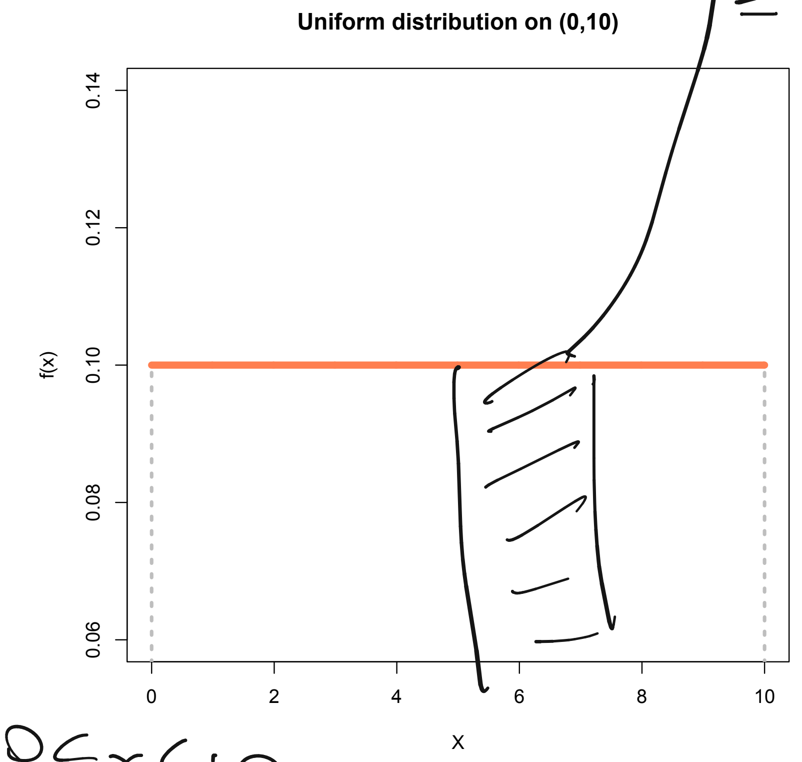
DERIVING PROBABILITIES FOR ANY INTERVAL

$P(5 \leq X \leq 7)$

$= \text{area}$

$= \int_5^7 f(x)\,dx$

- Going back to the random variable X or "bus waiting time", what is the probability that a person waits at bus stop between 5 and 7 minutes?
- To find this area, we need to find the area under the curve between 5 and 7.
- What is the value of this probability?



$$f(x) = \begin{cases} 1/10 & 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

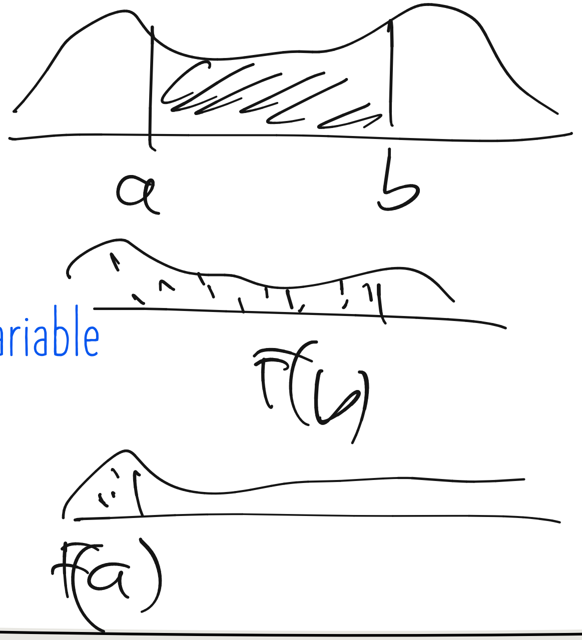
CUMULATIVE DISTRIBUTION FUNCTION

$P(a \le X \le b)$
$= F(b) - F(a)$

- The cumulative distribution function $F(x)$ of a random variable X is a function that provides the probability that X is less or equal than any specific value:

non-decreasing

$$\boxed{F(x) = P(X \le x) = P(-\infty \le X \le x)}$$

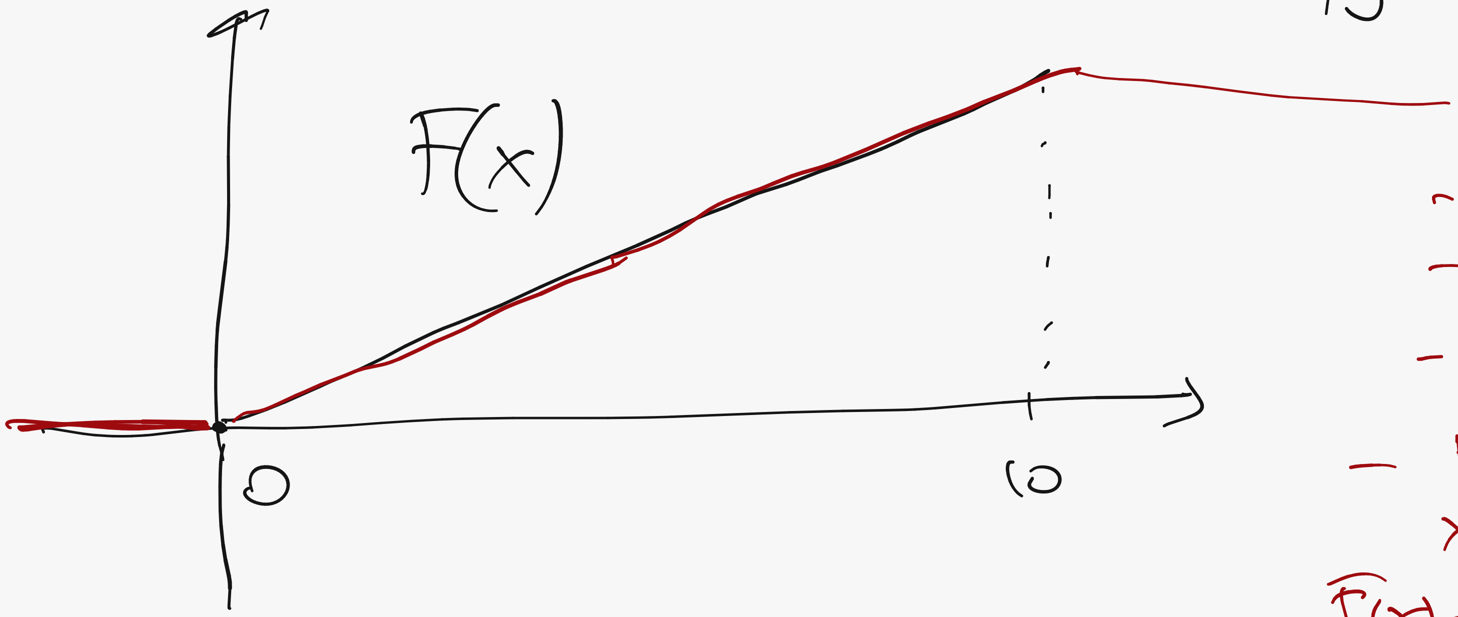


- In other words, the cumulative distribution function $F(x)$ of a random variable X reports the area under the curve to the left of the specified value.

pdf of $X \sim Unif((0,10))$

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{x} \frac{1}{10}\,dt = \left[\frac{t}{10}\right]_0^x$$

$$= \frac{x}{10}$$



- non-decreasing
- linear
- continuous
- piecewise
- $F(x) = 0$ when $x = 0$

$F(x) = 1$ when $x = 10$

By definition:

$$\text{cdf} = F(x) = \int_{-\infty}^{x} \underbrace{f(t)}_{\text{pdf}}\, dt$$

$$\text{P.d.f} \downarrow \text{c.d.f}$$

What if I have c.d.f. can I derive p.d.f ? $\boxed{\text{Yes}}$

Just take the derivative.

$$f(x) = \frac{dF(t)}{dt}\Big|_{t=x} = F'(x).$$

DERIVING PROBABILITIES USING A C.D.F.

- Going back to the example of the random variable X, "bus waiting time", and our interest in calculating the probability that an individual will wait for the bus between 5 and 7 minutes.
- Can you calculate this probability using the C.D.F. of X?

NORMAL RANDOM VARIABLES

- The most commonly encountered type of continuous random variables are normal random variables.
- The p.d.f. of a continuous normal random variable has a bell shape, called the normal curve, which is symmetric and is completely characterized once values for the mean and variance of the normal curve are specified.

NORMAL P.D.F.

$f(x \mid \mu, \sigma^2)$

- The expression of the normal p.d.f. of a normal random variable X with expected value, $\mathrm{E}(X) = \mu$ and variance $Var(X) = \sigma^2$, is:

parameters

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \underbrace{exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}_{\text{kernel}}$$

data

$x \in \mathbb{R}$

EXPECTED VALUE AND VARIANCE OF A NORMAL R.V.

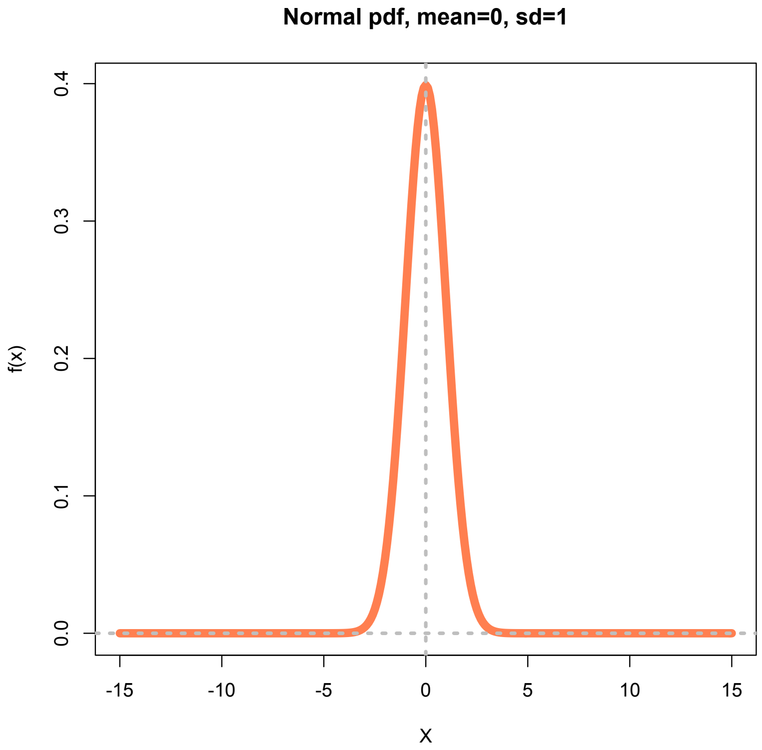
$X \sim N(\mu, \sigma^2)$

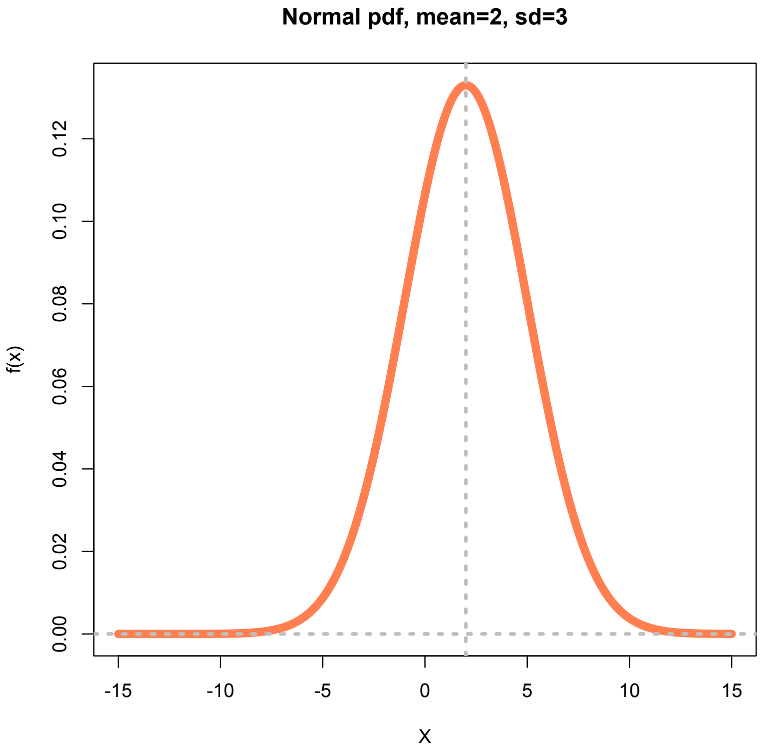
- As indicated, the expected value and the variance of a normal random variable are given respectively by: μ and $\boldsymbol{\sigma^2}$.

$E(X) = \mu \qquad Var(X) = \sigma^2$

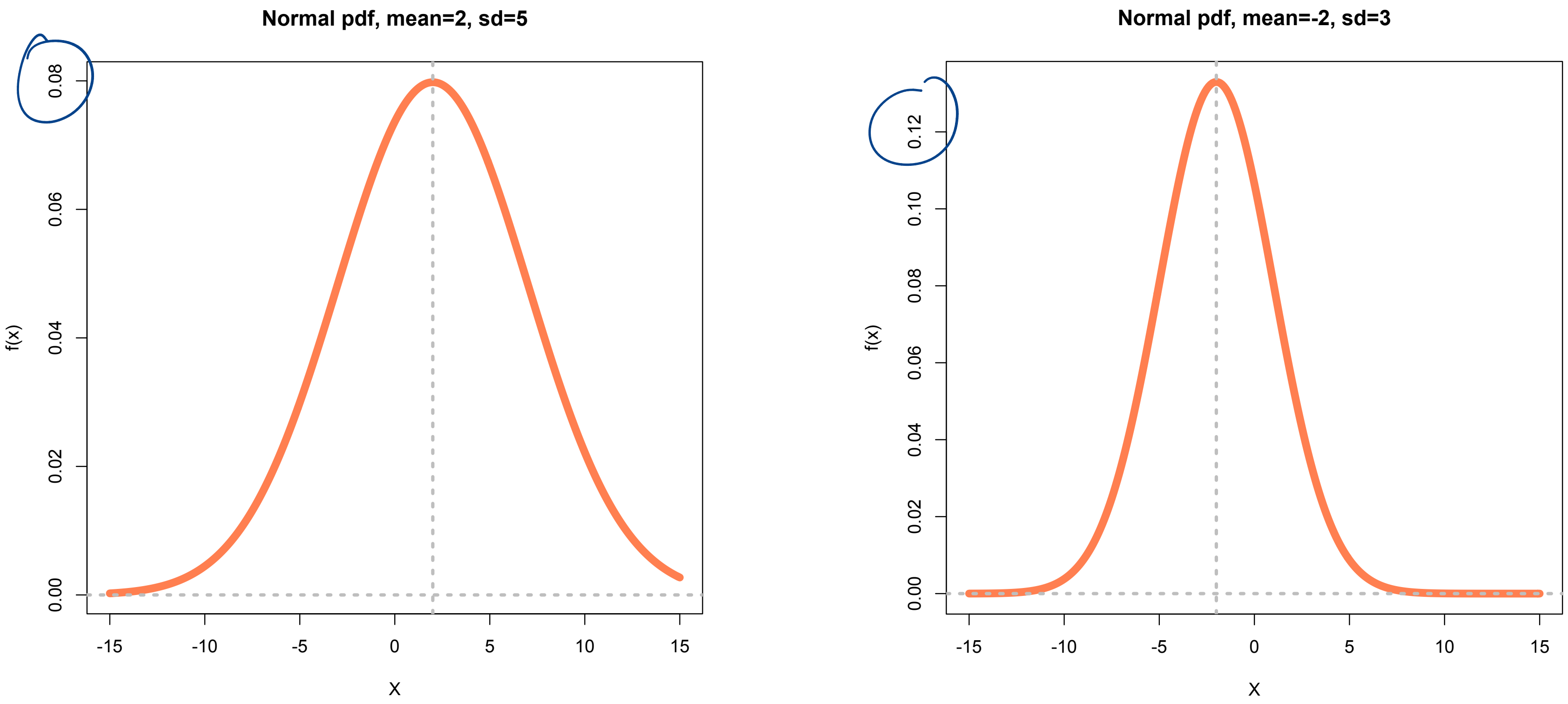
- A normal random variable whose expected value, μ, is equal to 0 and whose variance, $\boldsymbol{\sigma^2}$, are equal to 1, is called a standard normal random variable.

NORMAL P.D.F.


Normal pdf, mean=0, sd=1
f(x)
0.0
0.1
0.2
0.3
0.4
-15
-10
-5
0
5
10
15
X


Normal pdf, mean=2, sd=3
f(x)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
-15
-10
-5
0
5
10
15
X

NORMAL P.D.F.


Normal pdf, mean=2, sd=5
f(x)
0.00
0.02
0.04
0.06
0.08
-15
-10
-5
0
5
10
15
X
Normal pdf, mean=-2, sd=3
f(x)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
-15
-10
-5
0
5
10
15
X

SOME CHARACTERISTICS OF THE NORMAL DISTRIBUTION (I)

Some features shared by all normal curves and normal random variables are:

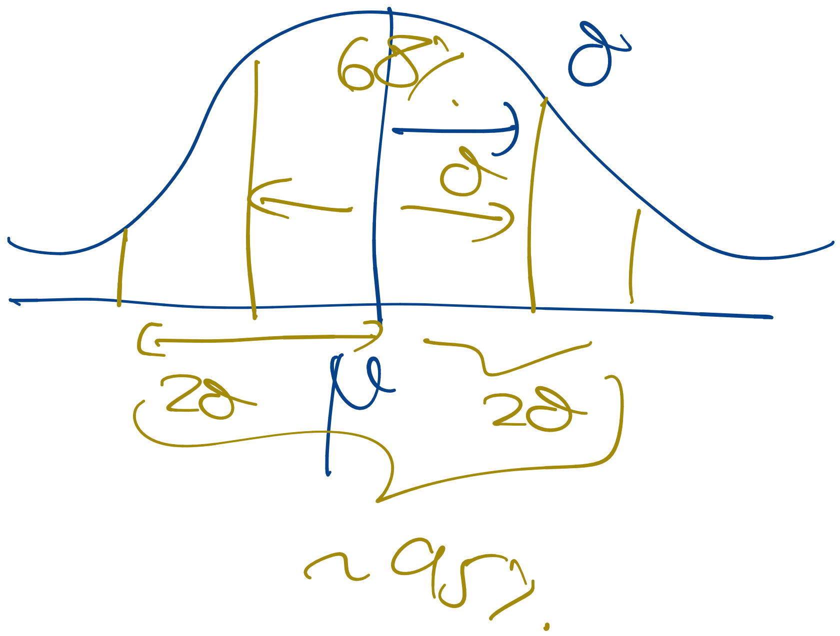
- The normal curve is symmetric and bell shaped (but not all symmetric and bell-shaped density curves are normal curves).

- $P(X \leq \mu) = P(X \geq \mu) = 0.5$
- $P(X \leq \mu - d) = P(X \geq \mu + d)$ for any positive number d.

SOME CHARACTERISTICS OF THE NORMAL DISTRIBUTION (I)

Some features shared by all normal curves and normal random variables are:

- $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$



CALCULATING A STANDARDIZED OR Z SCORE

- It is possible to transform any value assumed x assumed by a normal random variable X, with expected value or mean μ and variance, σ^2 (and thus, standard deviation σ), into the corresponding z-score, or value z, of a standard normal variable:

$$z = \frac{x-\mu}{\sigma}.$$

- A z-score measures the number of standard deviation that a values falls from the mean.

FINDING PERCENTILES AND PROBABILITIES IN R:

- If X is a normal random variable with expected value or mean μ and variance, σ^2 (and thus, standard deviation σ), it is possible to find percentiles and probabilities that X falls in any interval (a,b), using the following functions in R:

  - `qnorm(p,mean=…, sd=…)` where p is the probability corresponding to the percentile sought
  - `pnorm(x, mean=.., sd=..)` which reports the area to the left of x or $P(X \leq x)$.
  - dnorm (x, mean=.., sd=..)

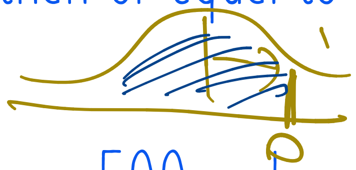
EXAMPLE: MATH SAT SCORE

Suppose that the scores on the math section of the SAT test are normally distributed with mean μ =515 and standard deviation σ=100.

a. What is the probability that a randomly selected test-taker has a score less than or equal to 600?

→ find z-score $z = \frac{600 - 515}{100}$

→ pnorm(600, 515, 100)



b. What is the probability that a randomly selected test-taker has a score between 500 and 600?

pnorm(600, 515, 100) − pnorm(500, 515, 100)

c. What score must a test-taker have to be in 90th percentile?

qnorm(.9, 515, 100)

GAMMA RANDOM VARIABLES

- A continuous random variable X is said to be a Gamma random variable and follow a Gamma distribution if:

  - X can only take positive values
  - Its probability density function has a shape with a potentially long tail to the right.

  - Gamma random variables are used to model environmental phenomena, such as accumulated rainfall amount in an interval of time, time in between sequences of earthquakes, etc.

GAMMA P.D.F.

- If the random variable X follows a Gamma distribution, its probability density function has the following expression:

normalization constant

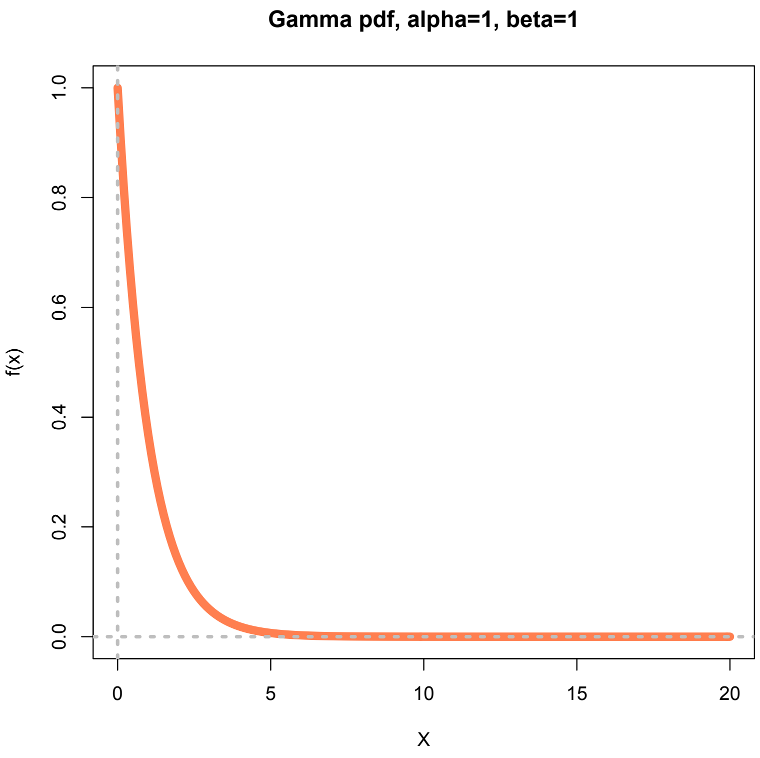
Kernel

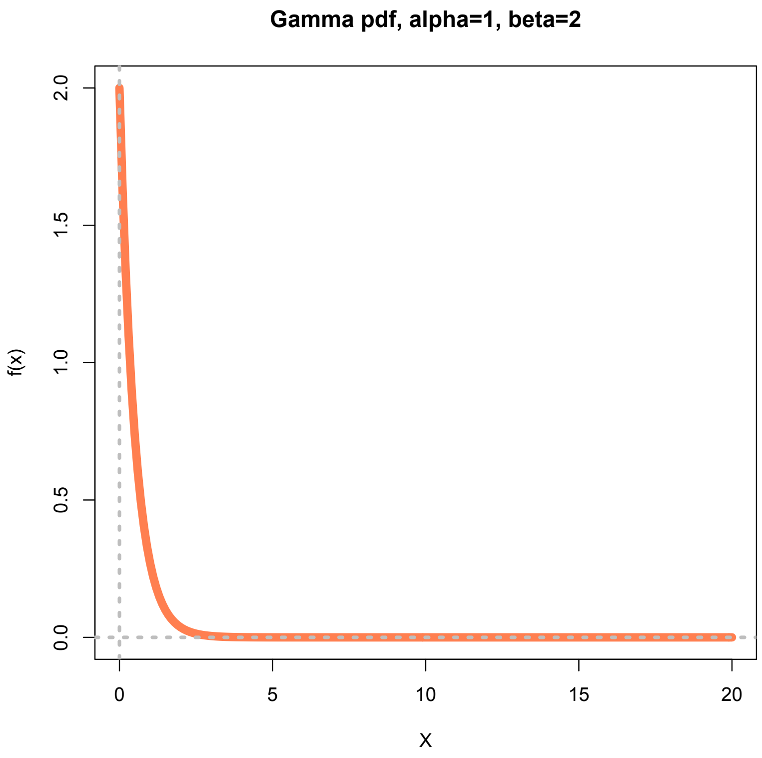
$$\boxed{f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \boxed{x^{\alpha-1} \cdot e^{-\beta x}}}$$

The shape of the p.d.f. curve is determined by α and β, respectively, called the shape and rate parameter of the distribution. These two parameters are both positive.

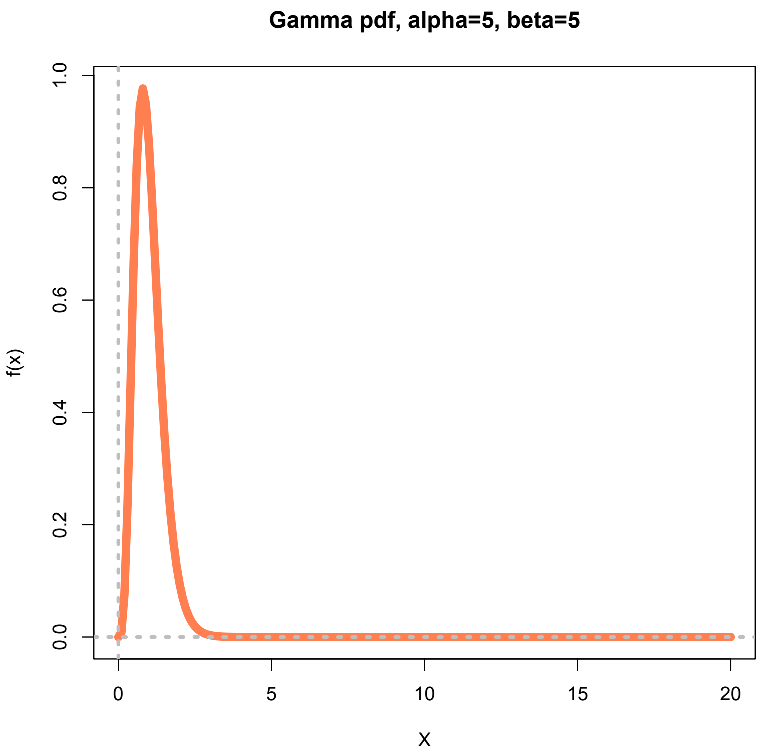
The function at the denominator, $\Gamma(\cdot)$, is called the Gamma function and it is a function whose definition involve an integral.

GAMMA P.D.F.

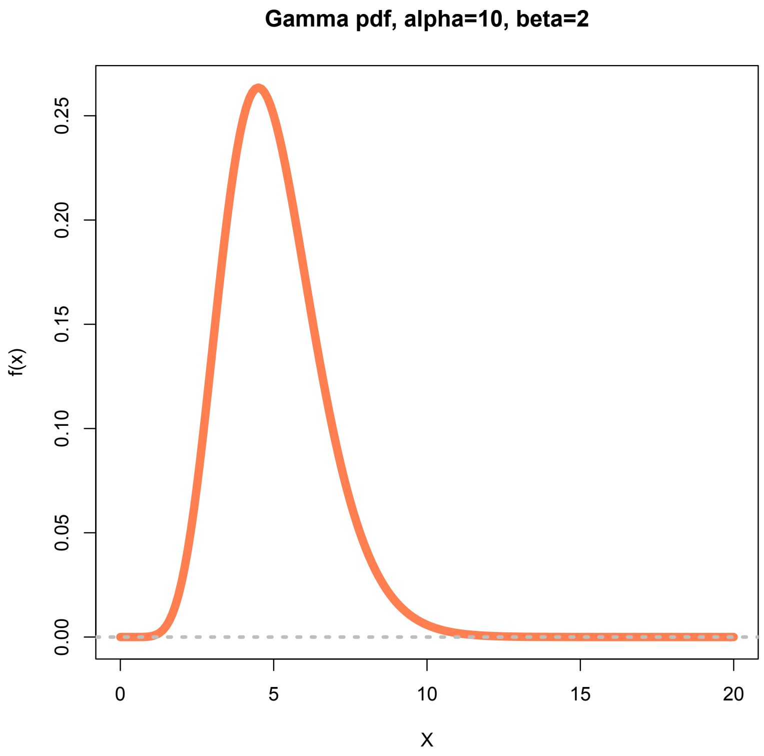

Gamma pdf, alpha=1, beta=1
f(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


Gamma pdf, alpha=1, beta=2
f(x)
x
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20

GAMMA P.D.F.



Gamma pdf, alpha=5, beta=5
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
X



Gamma pdf, alpha=10, beta=2
f(x)
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
X

EXPECTED VALUE AND VARIANCE

- If X is a random variable following a Gamma distribution with shape parameter α and rate parameter β, then:

$\frac{\text{shape}}{\text{rate}}$ but 1/rate = scale

= shape * scale

$$\boxed{E(X) = \frac{\alpha}{\beta}}$$

$$\boxed{Var(X) = \frac{\alpha}{\beta^2}}$$

- The Gamma distribution is often used in Bayesian statistics.

BETA RANDOM VARIABLES

- Another distribution that is often used in Bayesian statistics is the Beta distribution.
- A random variable X is said to be a Beta random variable, and follow a Beta distribution if:
  - X can only take values between 0 and 1.

BETA P.D.F

$$Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}$$

- If the random variable X follows a Beta distribution, its probability density function has the following expression:

$$\boxed{f(x;\alpha,\beta) = \frac{x^{\alpha-1}\cdot(1-x)^{\beta-1}}{B(\alpha,\beta)}}$$

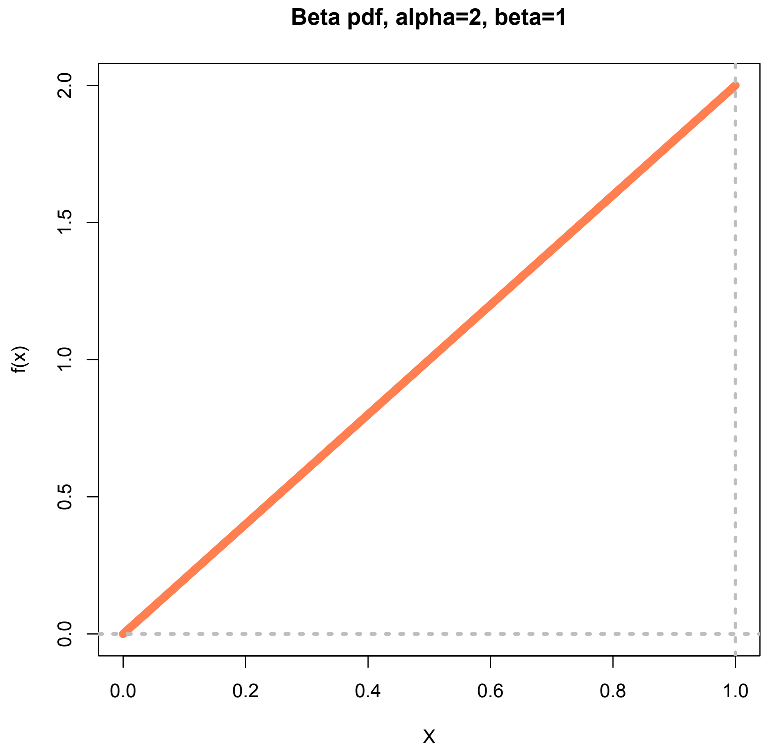
kernel

beta function

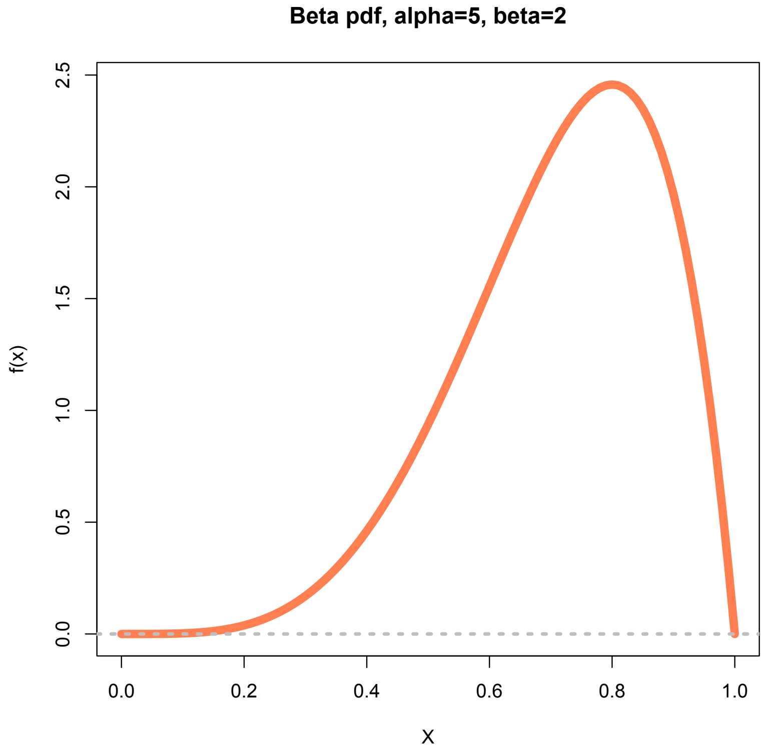
The shape of the p.d.f. curve is determined by α and β, both called the shape parameters of the distribution. These two parameters are both positive.

The function at the denominator, $\mathrm{B}(\cdot)$, is called the Beta function and it is a function whose definition involves the Gamma function $\Gamma(\cdot)$.

BETA P.D.F.

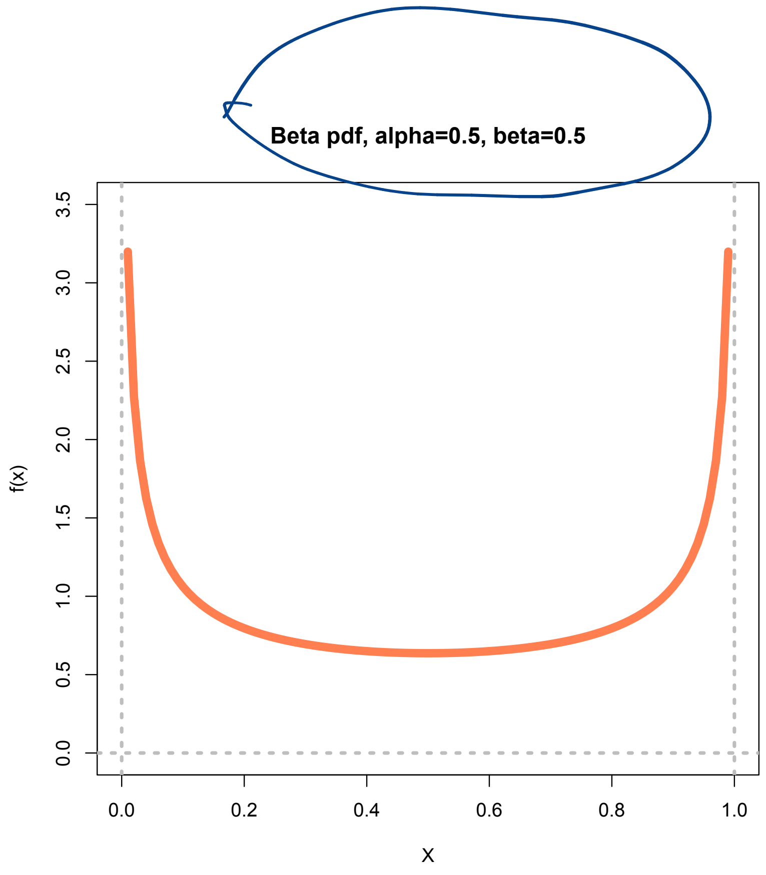


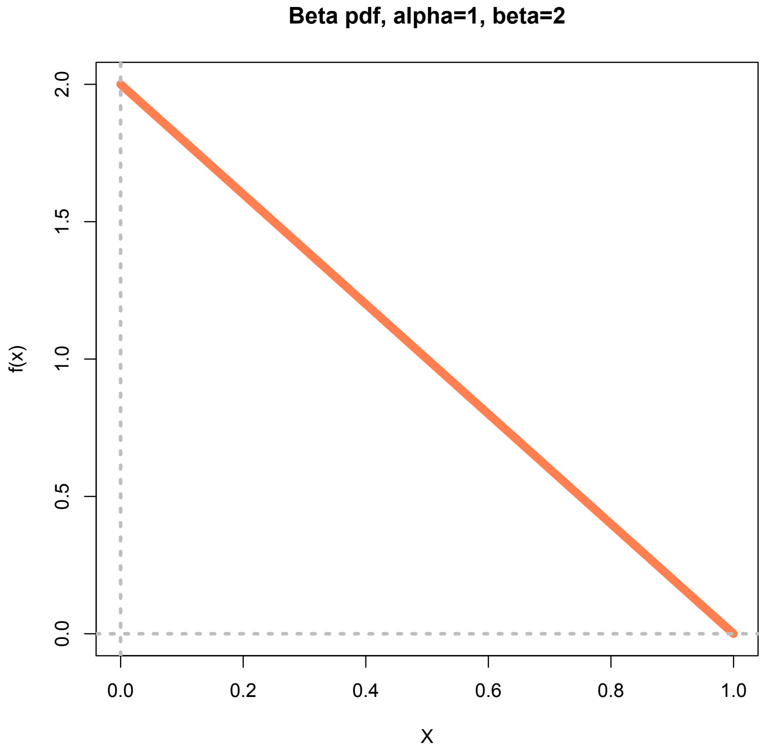
Beta pdf, alpha=2, beta=1
f(x)
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
X



Beta pdf, alpha=5, beta=2
f(x)
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
X

BETA P.D.F.


Beta pdf, alpha=0.5, beta=0.5
f(x)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
X


Beta pdf, alpha=1, beta=2
f(x)
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
X

EXPECTED VALUE AND VARIANCE

- If X is a random variable following a Beta distribution with shape parameters α and β, then:

$$\boxed{E(X) = \frac{\alpha}{\alpha+\beta}}$$

$$\boxed{Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2 \cdot (\alpha+\beta+1)}}$$

EXERCISES:

1. A line segment of length 1 is cut once at random. What is the probability that the longer piece is more than twice ~~the~~ that of the shorter piece?
2. Suppose that X has as p.d.f. $f(x) = cx^2$ for $0 \leq x \leq 1$ and $f(x) = 0$ otherwise. Find c. Additionally what is $P(0.1 \leq X \leq 0.5)$?