Probability Distributions in R

Contents

- Random Variable
- Probability Mass function
- Cumulative Distribution Function

Random Variable

- A random variable *x* takes on a defined set of values with different probabilities.
    - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
- Probability is how frequently we expect different outcomes to occur if we repeat the experiment over and over

Random Variables

There are two types of Random Variables

–Discrete random variables

–Continuous random variables.

Discrete Data Vs Continuous Data

A discrete space is one in which all possible outcomes can be clearly identified and counted. Data that has a clearly finite number of values

- Die Roll: S = {1,2,3,4,5,6}
- Number of Boys among 4 children: S = {0, 1, 2, 3, 4}
- Number of baskets made on two free throws: S = {0, 1, 2}

A non-discrete space, which is called a <u>continuous</u> interval/ space, is one in which the outcomes are too numerous to identify every possible outcome:

Height of Human Beings: S = [0.00 inches, 100.00 inches] → It is impractical to specify every possible height from 0.00 to 100.00.

Examples

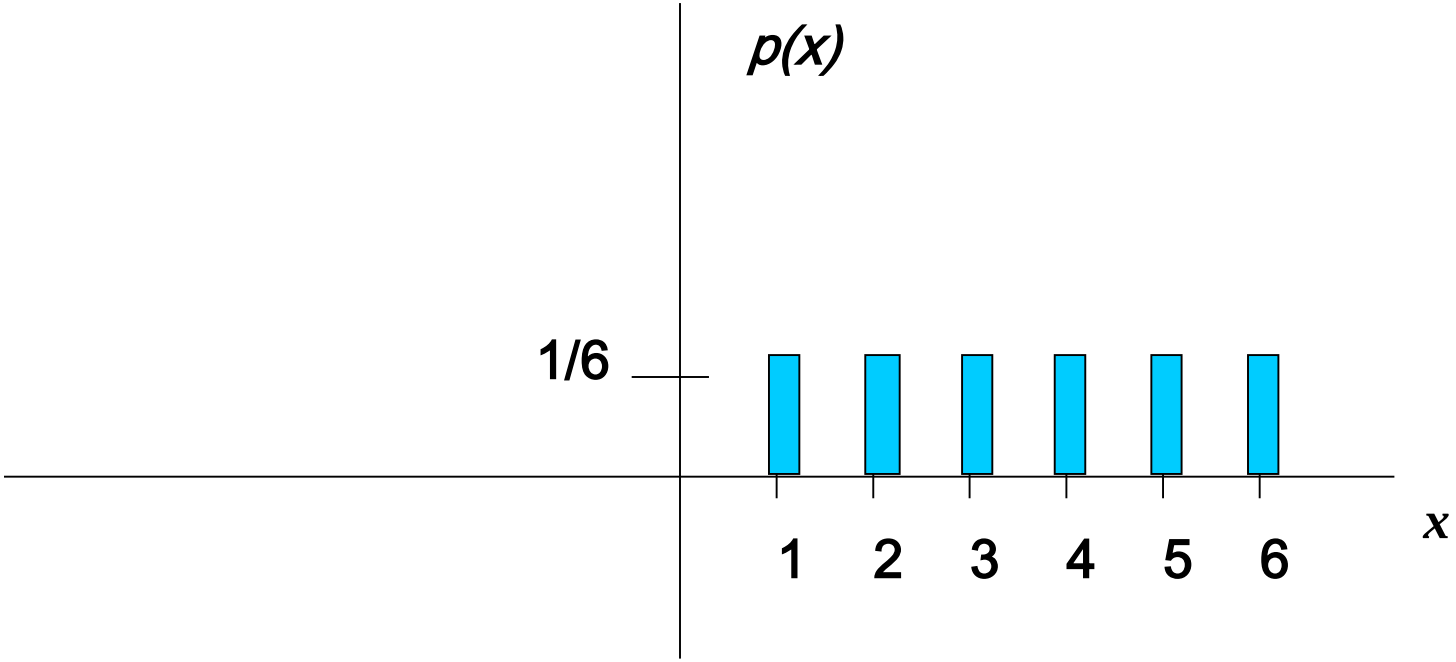
- **Discrete random variables**
  - Number of sales
  - Number of calls
  - Shares of stock
  - People in line
  - Mistakes per page

- **Continuous random variables**
  - Length
  - Depth
  - Volume
  - Time
  - Weight

Probability Mass functions

- A probability mass function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)*
- *p(x)* is a number from 0 to 1.0.
- The area under a probability function is always 1.

Discrete example: roll of a die



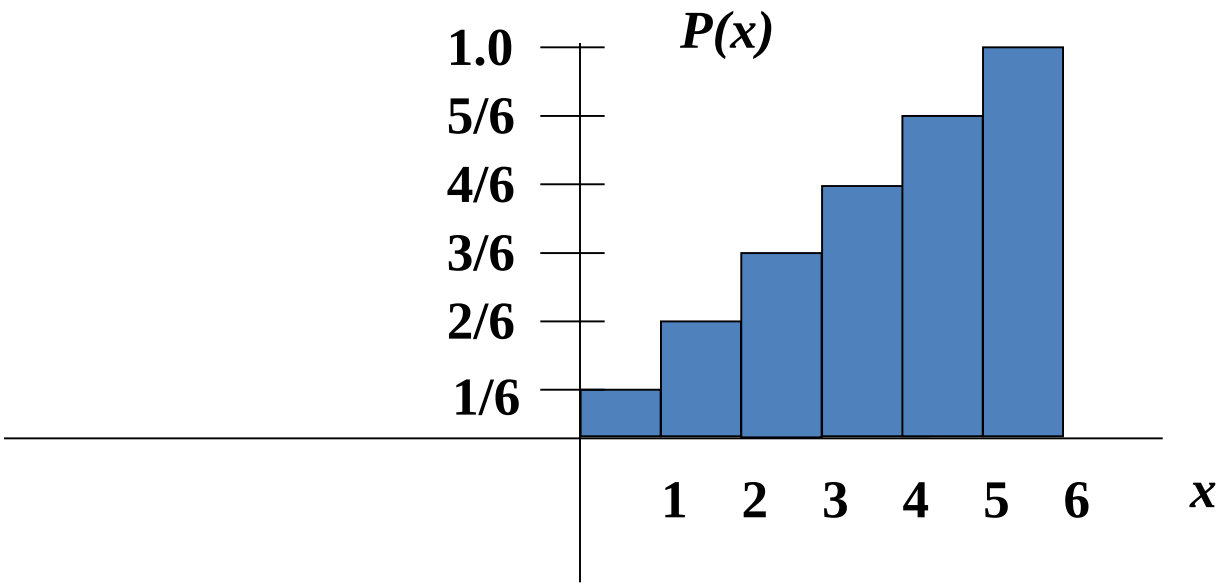
$$\sum_{\text{all } x} P(x) = 1$$

Probability mass function (pmf)

x	*p(x)*
1	*p(x=1)*=1/6
2	*p(x=2)*=1/6
3	*p(x=3)*=1/6
4	*p(x=4)*=1/6
5	*p(x=5)*=1/6
6	*p(x=6)*=1/6

1.0

Cumulative distribution function (CDF)



Cumulative distribution function

x	$P(x \leq A)$
1	$P(x \leq 1) = 1/6$
2	$P(x \leq 2) = 2/6$
3	$P(x \leq 3) = 3/6$
4	$P(x \leq 4) = 4/6$
5	$P(x \leq 5) = 5/6$
6	$P(x \leq 6) = 6/6$

Practice Problem:

- The number of patients seen in the ER in any given hour is a random variable represented by *x*. The probability distribution for *x* is:

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

Find the probability that in a given hour:

a. exactly 14 patients arrive $p(x=14) = .1$

b. At least 12 patients arrive $p(x \geq 12) = (.2 + .1 + .1) = .4$

c. At most 11 patients arrive $p(x \leq 11) = (.4 + .2) = .6$

Probability Distributions for Discrete Random Variables

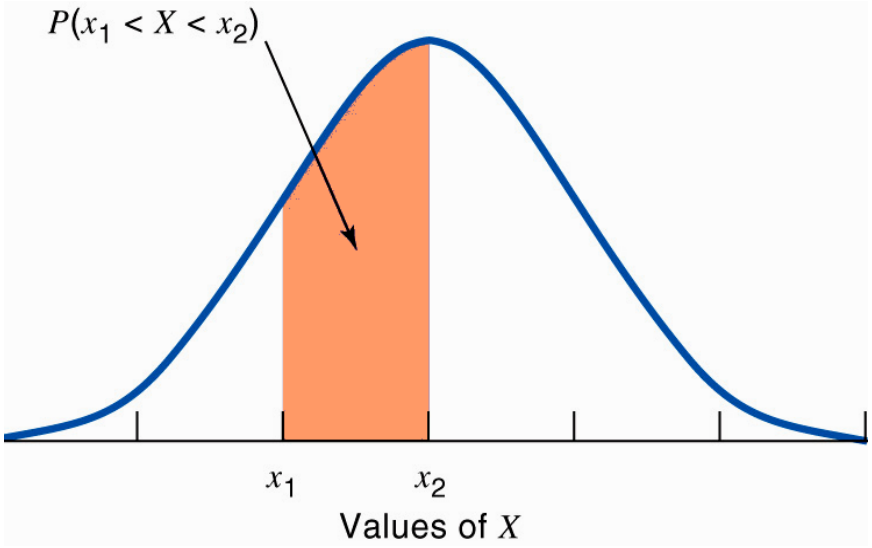
- Say a random variable *x* is rolling a die.
- This table gives the probabilities

$$P(x) = \begin{cases} \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise.} \end{cases}$$

x	*P(x)*
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Probability Distributions for Continuous Random Variables

The probability of any event in which the outcomes are continuous is the area under the density curve for the values of X that make up the event.



Only intervals can have a non-zero probability

The probability of a single event is ***meaningless*** for a continuous random variable. In fact, the calculated value is 0.

Example: The height of a sample of women has a distribution of approximately N(64.5, 2.5). What is the probability, if we pick one woman at random, that her height will be between 68 and 70 inches. I.e. ***P*(68 < *X* < 70)**?

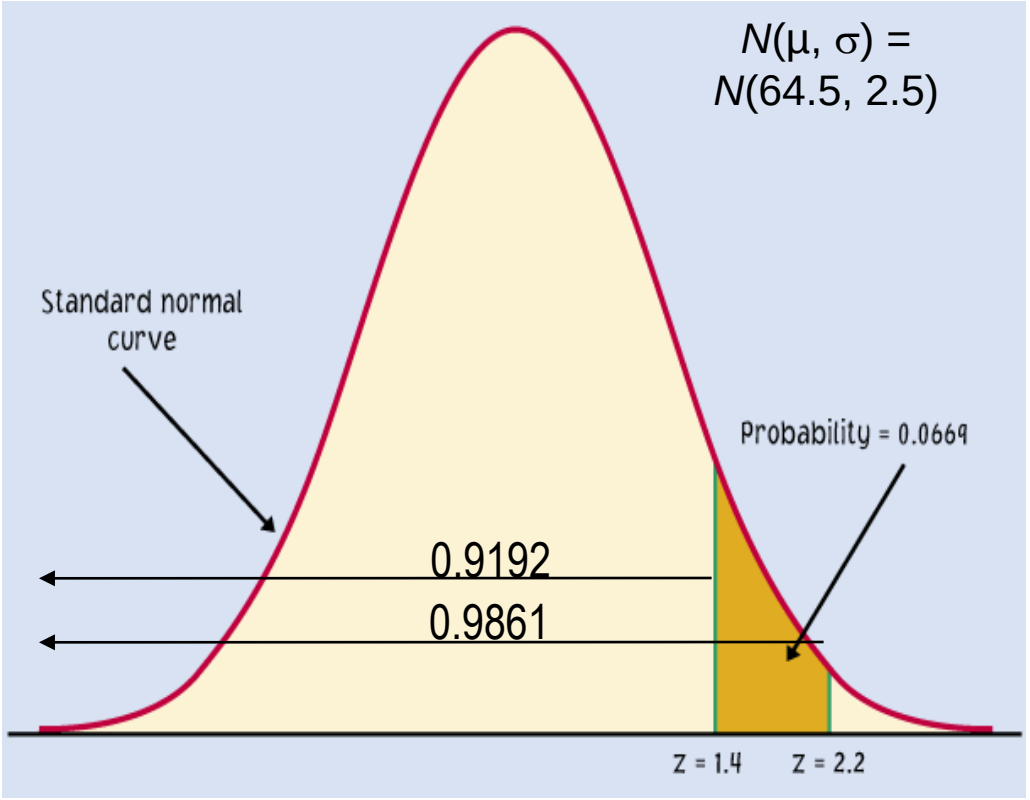
$$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$$

$$z = \frac{(x-\mu)}{\sigma}$$

we calculate the *z*-scores for 68 and 70 and determine the area between them.

For x = 68", $z = \frac{(68-64.5)}{2.5} = 1.4$

For x = 70", $z = \frac{(70-64.5)}{2.5} = 2.2$



The area under the curve for the interval [68" to 70"] is 0.9861 − 0.9192 = 0.0669.

I.e. The probability that a randomly chosen woman falls into this range is 0.0669.

Mean of a discrete random variable

For a discrete random variable *X* with the probability distribution shown here,

Value of X	x_1	x_2	x_3	$\cdots$	x_k
Probability	p_1	p_2	p_3	$\cdots$	p_k

the mean μ of *X* is found by multiplying each possible value of *X* by its probability, and then adding the products.

$$\mu_X = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$
$$= \sum x_i p_i$$

A basketball player shoots three free throws. The random variable *X* is the number of baskets successfully made.

Value of *X*	**0**	**1**	**2**	**3**
Probability	**0.064**	**0.288**	**0.432**	**0.216**

The mean μ of *X* is

μ = (0*0.064) + (1*0.288) + (2*0.432) + (3*0.216)

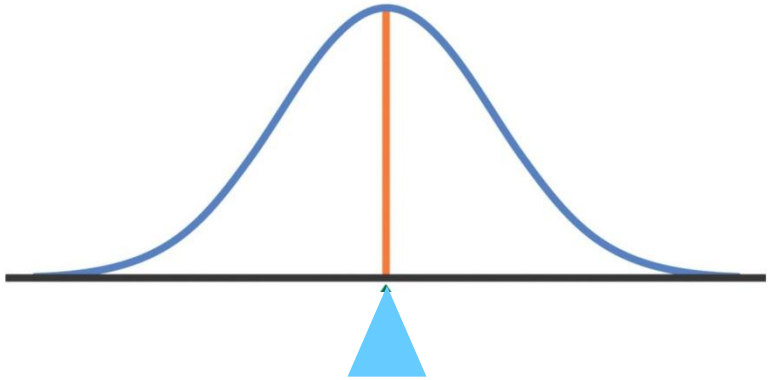
= 1.8

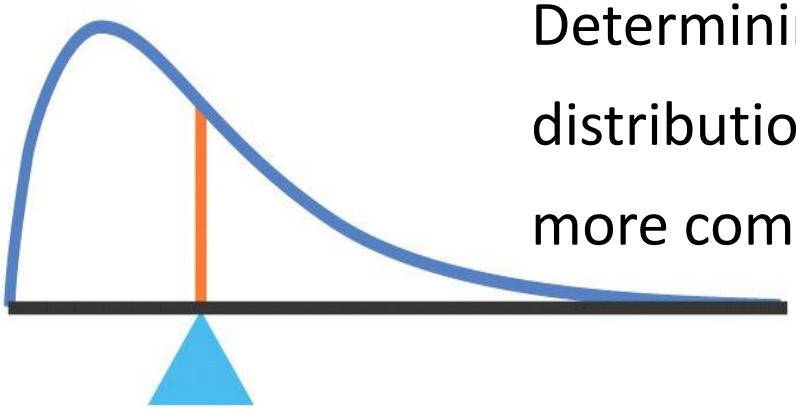
In other words, out of 3 throws, in the long run, this player would make 1.8 baskets.

Mean of a continuous random variable

As we did with probabilities, when working with continuous data, we rely on the density curve to make various determinations.

With symmetric curves, such as the Normal curve, the mean lies at the center.





Determining the exact mean of a distribution with a skewed density curve is more complex.

Probability Distributions

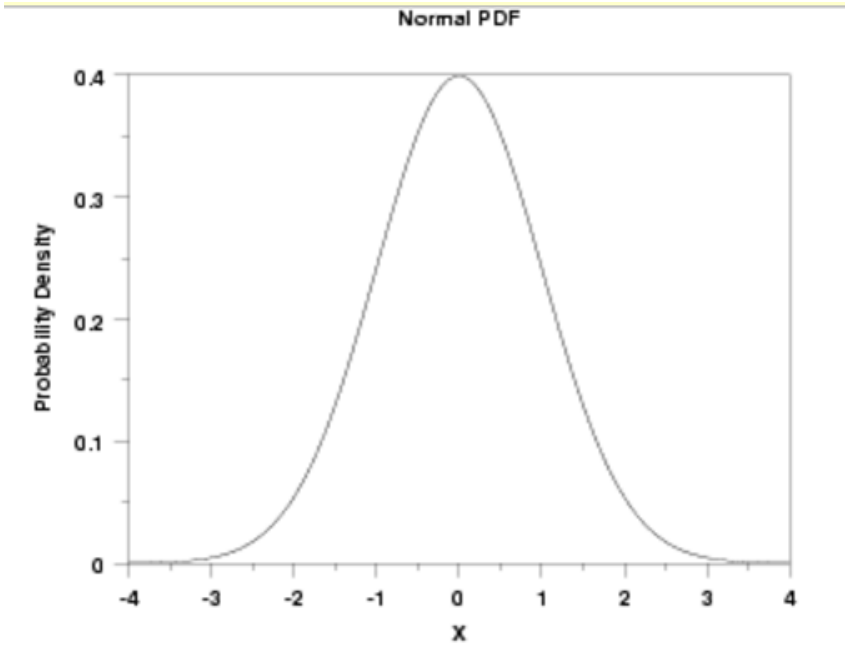
- Normal
- Binomial
- Poisson
- Uniform

Normal Distribution- probability density function

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

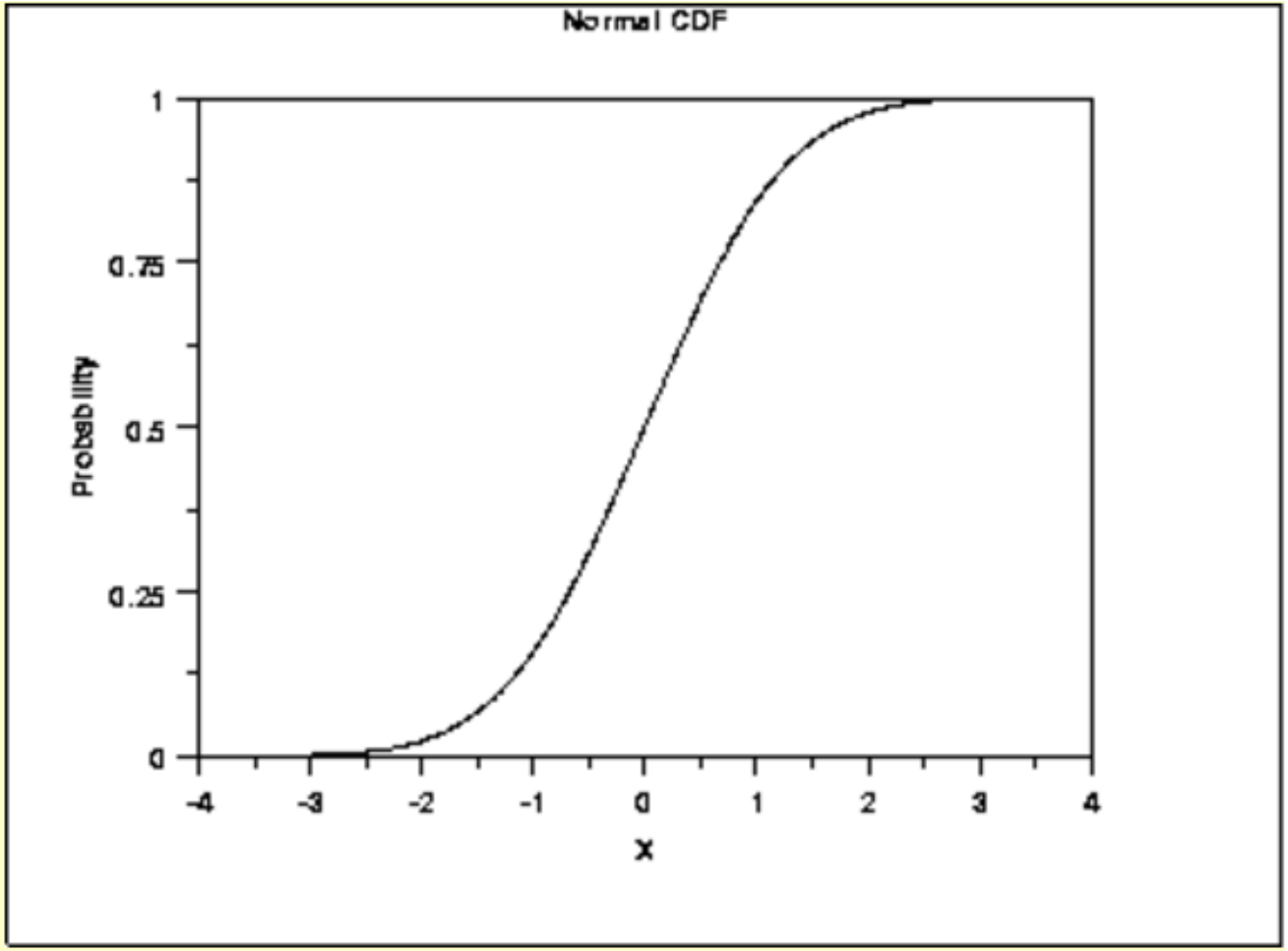
$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

where $\mu = 0$ and $\sigma = 1$



Normal Distribution- Cumulative Distribution Function

The formula for cdf is $$F(x) = \int_{-\infty}^{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$



R function for Normal Distribution

R has four in built functions to generate normal distribution.

They are described below.

rnorm(n, mean, sd)
dnorm(x, mean, sd)
pnorm(x, mean, sd)
qnorm(p, mean, sd)

Following is the description of the parameters used in above functions −

- **x** is a vector of numbers.
- **p** is a vector of probabilities.
- **n** is number of observations(sample size).
- **mean** is the mean value of the sample data. It's default value is 0.
- **sd** is the standard deviation. It's default value is 1.

rnorm

- This function is used to generate random numbers whose distribution is normal.
- It takes the sample size as input and generates that many random numbers.
- Eg 1: `y <- rnorm(50)`

mean = 0 and sd = 1 are the default arguments for the dnorm function.

Eg 2: rnorm(20, mean = 50, sd = 3)

dnorm()

- This function gives height of the probability distribution at each point for a given mean and standard deviation.
- As we all know the probability density for the normal distribution is: $f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- The function dnorm returns the value of the probability density function for the normal distribution given parameters for X, μ, and σ
- dnorm(0, mean = 0, sd =1) 0.3989
- dnorm(2, mean = 5, sd = 3) 0.0865

pnorm

- This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function"

qnorm

- This function takes the probability value and gives a number whose cumulative value matches the probability value.