Probability

Intro to Data Science

Anton Kalén

University of Skövde

Sep 16, 2022

> Mathematics is the logic of certainty;
> probability is the logic of uncertainty.

Certainty

Constants

$$\pi = 3.1415927$$

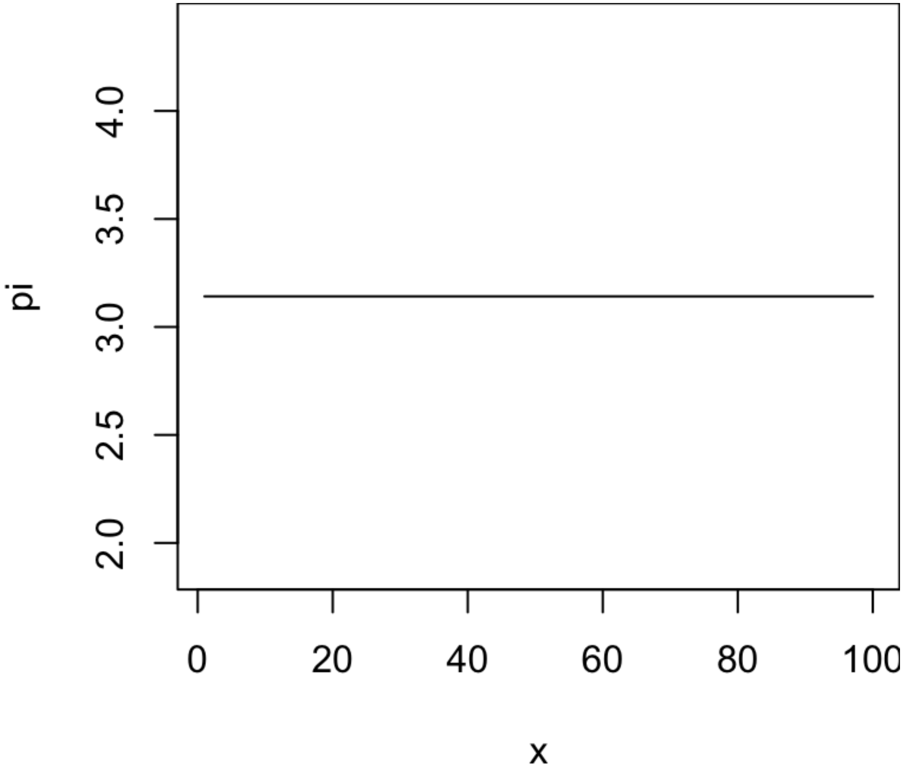
```r
x <- seq(
  from = 1,
  to = 100,
  length.out = 100
)

pi <- rep(
  3.1415927,
  times = 100
)
```

Certainty

Constants

$\pi = 3.1415927$

```r
x <- seq(
  from = 1,
  to = 100,
  length.out = 100
)

pi <- rep(
  3.1415927,
  times = 100
)

par(mar = c(5.1, 4.1, 0.01, 0.01))
plot(x, pi, type = "l")
```



Certainty

Deterministic variables:

$$VAT = 0.25 \times price$$

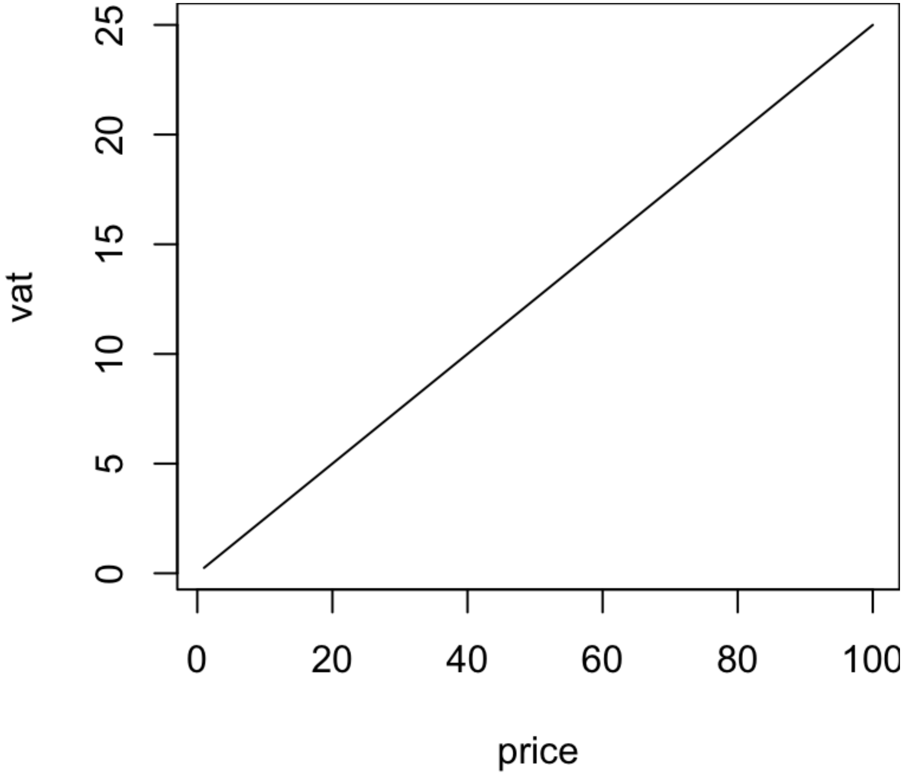
```r
price <- seq(
  from = 1,
  to = 100,
  by = 1
)

vat = 0.25 * price
```

Certainty

Deterministic variables

$$VAT = 0.25 \times price$$

```r
price <- seq(
  from = 1,
  to = 100,
  by = 1
)

vat = 0.25 * price

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(price, vat, type = "l")
```



Uncertainty

Will this person buy or not?

Number of visitors tomorrow

Height of the next person coming

Total weight of the next transport

Exercise

> You work in a online retailer and are tasked with predicting if a visitor will buy the product sold or not.

In groups of 2–4 persons, discuss what sources of uncertainty exist that makes it hard to be sure about if the visitor buys or not. 5 min.

Sources of Uncertainty

- Lack of knowledge of what determines if a visitor buys.
- Lack of knowledge about the visitor.
- Measurement error of potential predictors.
- Measurement error of previous sales.
- Sampling uncertainty of past visitors.
- Randomness.

Probability of buys

> The retailer has 3 visitors per day. How many of them will buy a product?

Probability of buys

1. What are all possible outcomes?

Sample space

$\Omega = \{\text{nnn, bnn, nbn, nnb, bbn, bnb, nbb, bbb}\}$

$\text{n} = \text{no buy}$

$\text{b} = \text{buy}$

Probability of buys

2. What are we interested in knowing?

Random Variable

Let X be the number of buys in a day.

$X : \Omega \to \mathbb{R}$

A random variable is a function mapping the sample space, Ω, to the real line, $\mathbb{R}$ (i.e., it assigns a number to each possible event).

Probability of buys

2. What are we interested in knowing?

Random Variable

$X(\text{nnn}) = 0$

$X(\text{bnn}) = 1$

$X(\text{nbn}) = 1$

$X(\text{nnb}) = 1$

$X(\text{bbn}) = 2$

$X(\text{bnb}) = 2$

$X(\text{nbb}) = 2$

$X(\text{bbb}) = 3$

Probability of buys

3. What is the probability of the different outcomes?

Probability function

$P(X)$

$P : X \rightarrow [0, 1]$

The probability function, P, assigns a probability between 0 and 1 to each possible outcome of the random variable X

Probability of buys

3. What is the probability of the different outcomes?

Probability function

$$P(X = 0) = \frac{1}{8} = 0.125$$

$$P(X = 1) = \frac{3}{8} = 0.375$$

$$P(X = 2) = \frac{3}{8} = 0.375$$

$$P(X = 3) = \frac{1}{8} = 0.125$$

If probability of each visitor buying = 0.5

Probability notation

X — Random variable (r.v.)

$\Omega_X = \{x_1, \dots, x_n\}$ — Sample space of r.v. X

$P(X)$ — Probability function on r.v. X

$P(X = x_1)$ or $P(x_1)$ — Probability of X taking value x_1

Probability distribution

We can represent random variables using probability distributions.

Distribution of r.v.

$X \sim \text{Bin}(n, p)$

$n = \text{number of trials.}$

$p = \text{probability of success.}$

The random variable X is binomial distributed, with n trials and p probability of success.

Binomial distribution

> The retailer has 3 visitors per day. The probability of each visitor buying a product is 0.5. How many of them will buy a product?

$n = 3$

$p = 0.5$

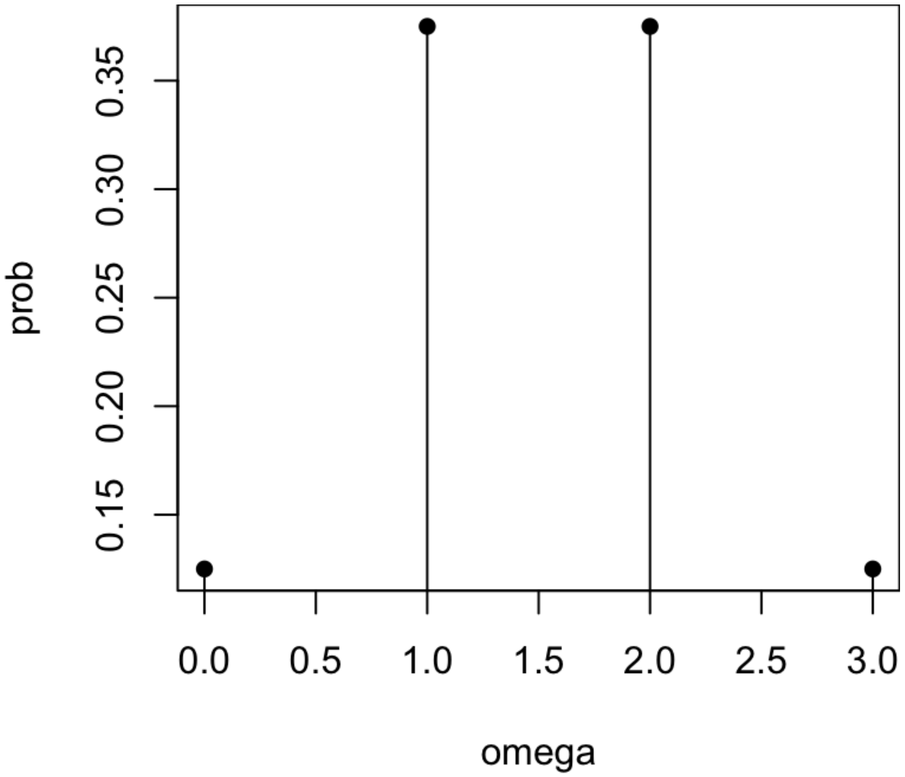
$X \sim \text{Bin}(3, 0.5)$

Binomial distribution

> The retailer has **3** visitors per day. The probability of each visitor buying a product is **0.5**. How many of them will buy a product?

```r
n <- 3
p <- 0.5

omega <- 0:n

prob <- dbinom(
  omega,
  size = n,
  prob = p
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

```

```r
plot(omega, prob, type = "h")
points(omega, prob, pch = 16)
```

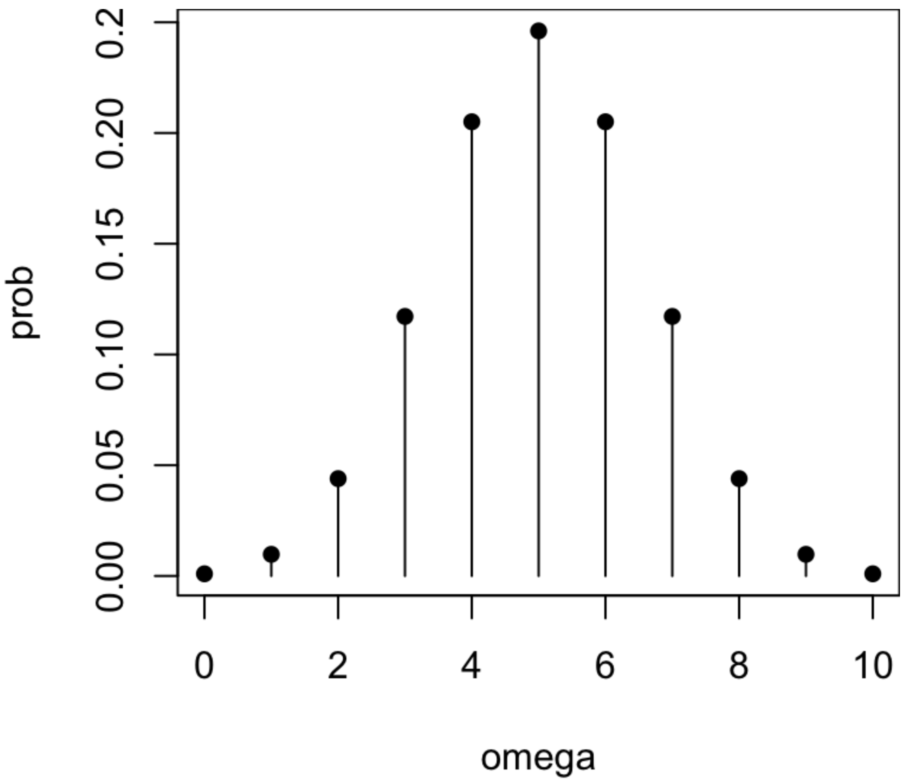


Binomial distribution

> The retailer has **10** visitors per day. The probability of each visitor buying a product is **0.5**. How many of them will buy a product?

```r
n <- 10
p <- 0.5

omega <- 0:n

prob <- dbinom(
  omega,
  size = n,
  prob = p
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

```

```r
plot(omega, prob, type = "h")
points(omega, prob, pch = 16)
```

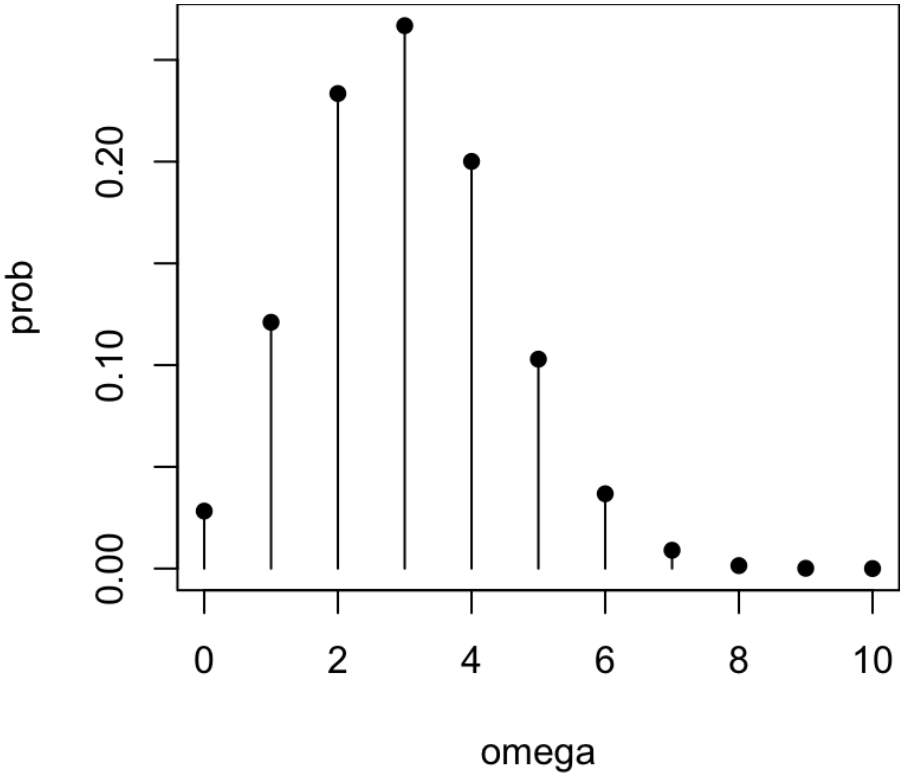


Binomial distribution

> The retailer has **10** visitors per day. The probability of each visitor buying a product is **0.3**. How many of them will buy a product?

```r
n <- 10
p <- 0.3

omega <- 0:n

prob <- dbinom(
  omega,
  size = n,
  prob = p
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

```

```
plot(omega, prob, type = "h")
points(omega, prob, pch = 16)
```



Probability mass function (PMS)

$$p_X(x_i) = P(X = x_i)$$

```
dbinom()
```

The probability that the discrete r.v. X is equal to x_i.

Probability mass function (PMS)

$$p_X(x_i) = P(X = x_i)$$

`dbinom()`

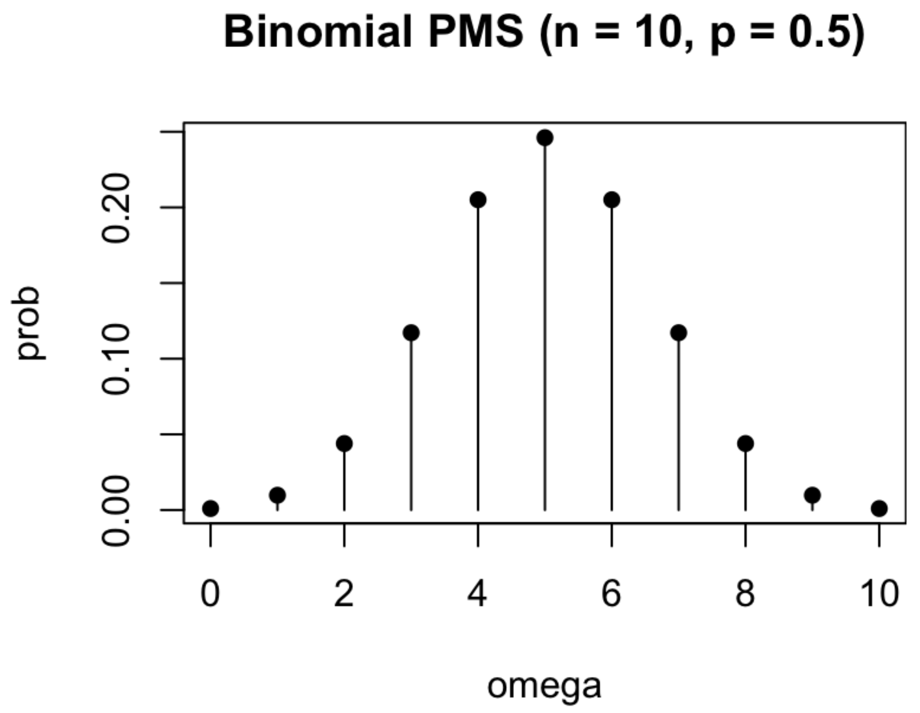
Cumulative distribution function (CDF)

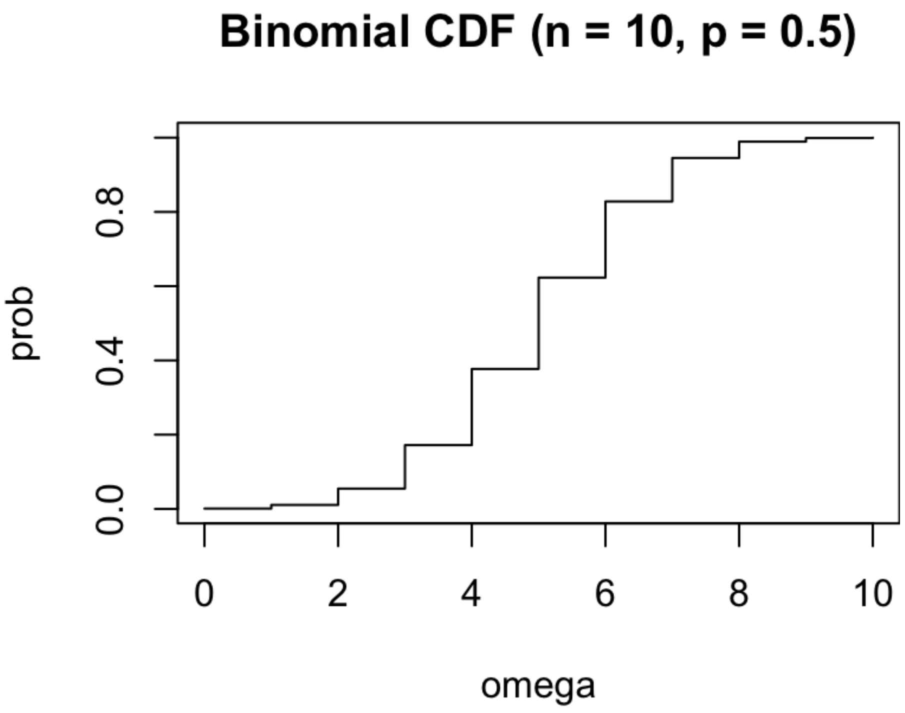
$$F_X(x_i) = P(X \leq x_i)$$

`pbinom()`

The probability that the discrete r.v. X is smaller or equal to x_i.

Binomial function





Poisson distribution

> Discrete duistribution for counting success in a specific time, where there is a large number of events and low probability.

For example, the online retailer has thousands of visitors each day and only a small proportion ends up buying.

Poisson distribution

$X \sim \mathrm{Pois}(\lambda)$

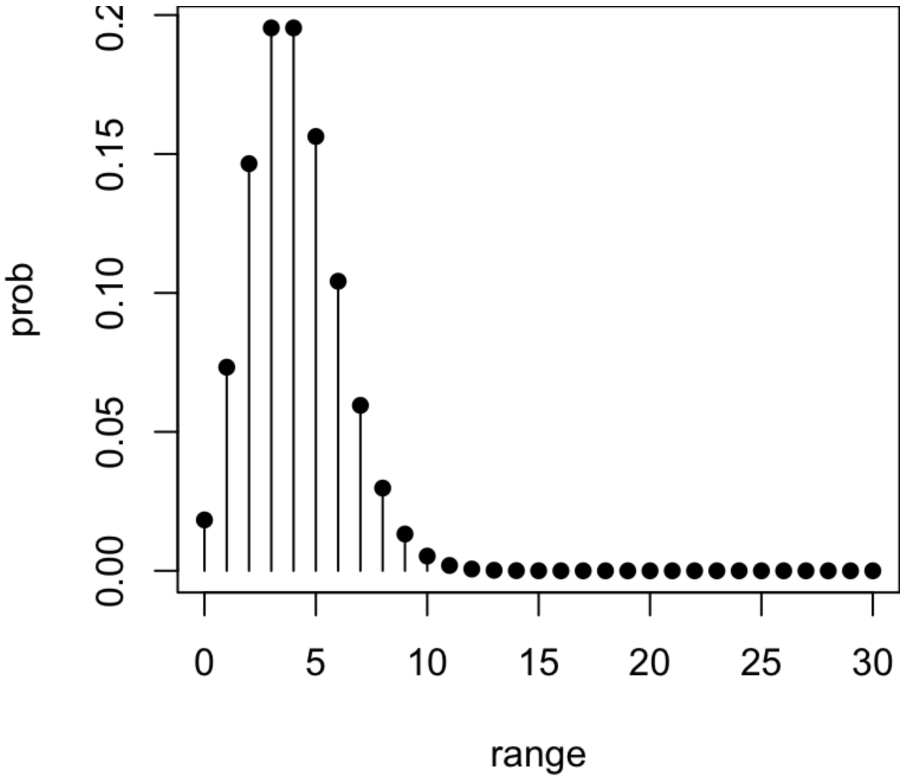
$\Omega_X = \mathbb{I}^+$ (Possible values are positive integers)

λ — Mean and variance of X

Poisson distribution (PMF)

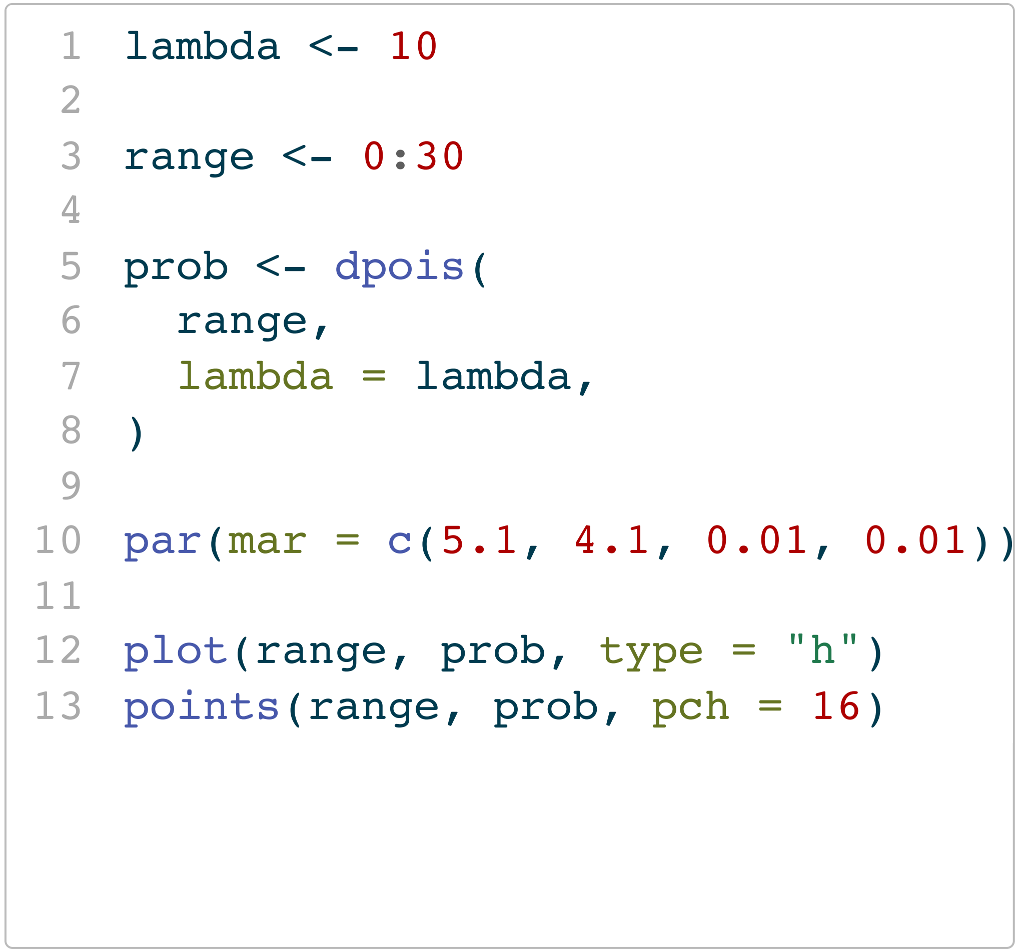
> The average amount of daily sales are **4**.

```r
lambda <- 4

range <- 0:30

prob <- dpois(
  range,
  lambda = lambda,
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(range, prob, type = "h")
points(range, prob, pch = 16)
```

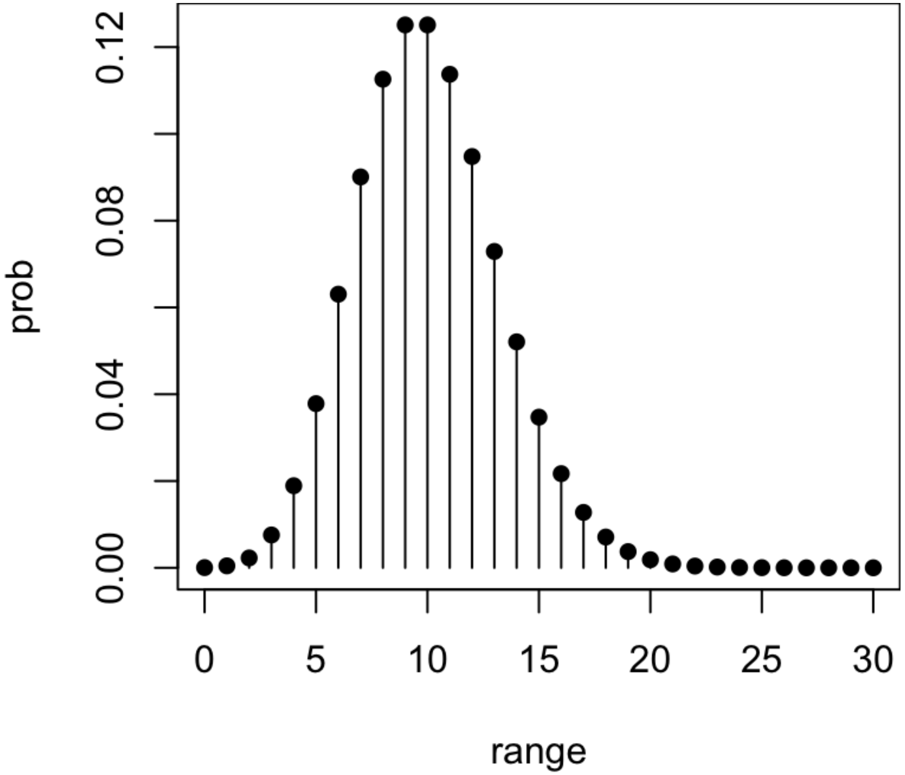


Poisson distribution (PMF)

> The average amount of daily sales are **10**.



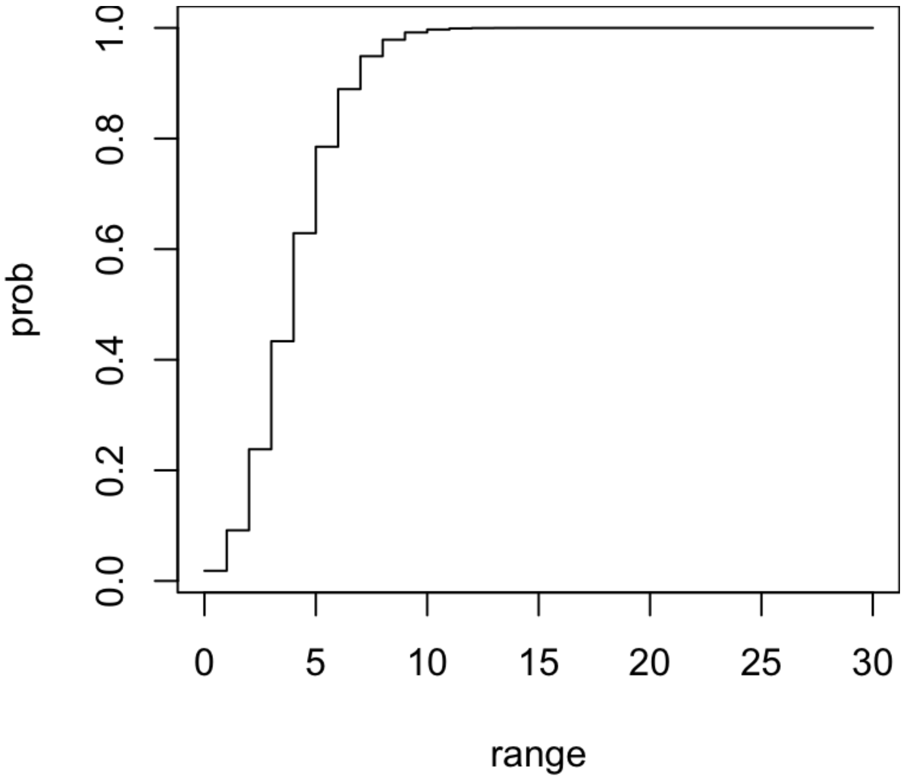
```r
lambda <- 10

range <- 0:30

prob <- dpois(
  range,
  lambda = lambda,
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(range, prob, type = "h")
points(range, prob, pch = 16)
```



Poisson distribution (CDF)

> The average amount of daily sales are **4**.

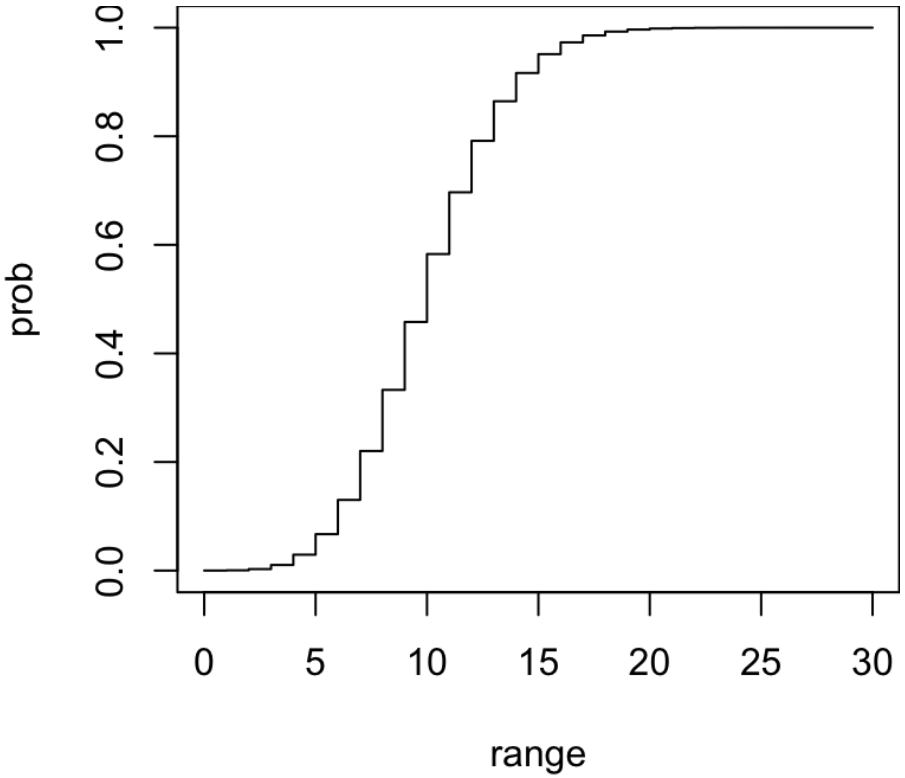
```r
lambda <- 4

range <- 0:30

prob <- ppois(
  range,
  lambda = lambda,
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(range, prob, type = "s")
```



Poisson distribution (CDF)

> The average amount of daily sales are **4**.

```r
lambda <- 10

range <- 0:30

prob <- ppois(
  range,
  lambda = lambda,
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(range, prob, type = "s")
```



Conditional probability

> What is the probability that a visitor buys a product, given that it is a raining day?

	Buy	No Buy
Sun	13	39
Rain	26	104

Random variables & sample space

Finish visit: $X, \Omega_X = \{\text{Buy}, \text{No Buy}\}$

Wheather: $Y, \Omega_Y = \{\text{Sun}, \text{Rain}\}$

Conditional probability

> What is the probability that a visitor buys a product, given that it is a raining day?

	Buy	No Buy
Sun	13	39
Rain	26	104

Marginal probability

Marginal probabilities X: $P(X = \text{Buy}), P(X = \text{No buy})$

Marginal probabilities of Y: $P(Y = \text{Sun}), P(Y = \text{Rain})$

Conditional probability

> What is the probability that a visitor buys a product, given that it is a raining day?

	Buy	No Buy
Sun	13	39
Rain	26	104

Conditional probability $P(X|Y)$

Probability of X given Sun: $P(X|Y = \text{Sun})$, or $P(X|\text{Sun})$

Probability of Buy, given Rain: $P(X = \text{Buy}|Y = \text{Rain})$, or $P(\text{Buy}|\text{Rain})$

Conditional probability

> What is the probability that a visitor buys a product, given that it is a raining day?

	Buy	No Buy
Sun	13	39
Rain	26	104

Joint probability $P(X, Y)$

Probability of Buy when it rains: $P(X = \text{Buy}, Y = Rain)$, or $P(X = \text{Buy}, Y = Rain)$

Conditional probability

```r
# Create the matrix

sun <- c(13, 39)
rain <- c(26, 104)

sales <- rbind(sun, rain)

colnames(sales) <- c("buy", "no buy")

sales
```

```
     buy no buy
sun   13     39
rain  26    104
```

Conditional probability

Marginal probability $P(X = \text{Buy})$

```r
# P(X = Buy)
sum(sales[,"buy"]) / sum(sales)
```

```
[1] 0.2142857
```

Marginal probability $P(Y = \text{Sun})$

```r
# P(Y = Sun)
sum(sales["sun",]) / sum(sales)
```

```
[1] 0.2857143
```

Conditional probability

Conditional probability $P(X = \text{Buy}|Y = \text{Rain})$

```r
# P(X = Buy|Y = Rain)
sales["rain","buy"] / sum(sales["rain", ])
```

```
[1] 0.2
```

Conditional probability $P(X = \text{Buy}|Y = \text{Sun})$

```r
# P(X = Buy|Y = Sun)
sales["sun","buy"] / sum(sales["sun", ])
```

```
[1] 0.25
```

Conditional probability $P(X|Y)$

```r
sales / rowSums(sales)
```

```
      buy no buy
sun  0.25   0.75
rain 0.20   0.80
```

Conditional probability

Joint probability $P(X = \text{Buy}, Y = \text{Rain})$

```r
sales["rain", "buy"] / sum(sales)
```

```
[1] 0.1428571
```

Joint probability $P(X, Y)$

```r
sales / sum(sales)
```

```
            buy    no buy
sun  0.07142857 0.2142857
rain 0.14285714 0.5714286
```

Probabilities sums to 1

```r
sum(sales / sum(sales))
```

```
[1] 1
```

Continuous random variables

> Random variables that can take any real value in an interval

- Infinite amount of real numbers → any specific real number has zero probability.

- Probability can only exist over measurable areas/volumes.

- Instead of probability mass functions (PMF), we work with probability densities functions (PDF)

Normal distribution

$X \sim \mathcal{N}(\mu, \sigma)$

mu — Mean of X

σ — Standard deviation of X

$\Omega = \mathbb{R} = (-\infty, \infty)$

Normal distribution

Probability density function

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\int_{-\infty}^{\infty} f(X) dX = 1$$

The probability density function $f(X)$ over all real numbers integrates to one .

Normal distribution

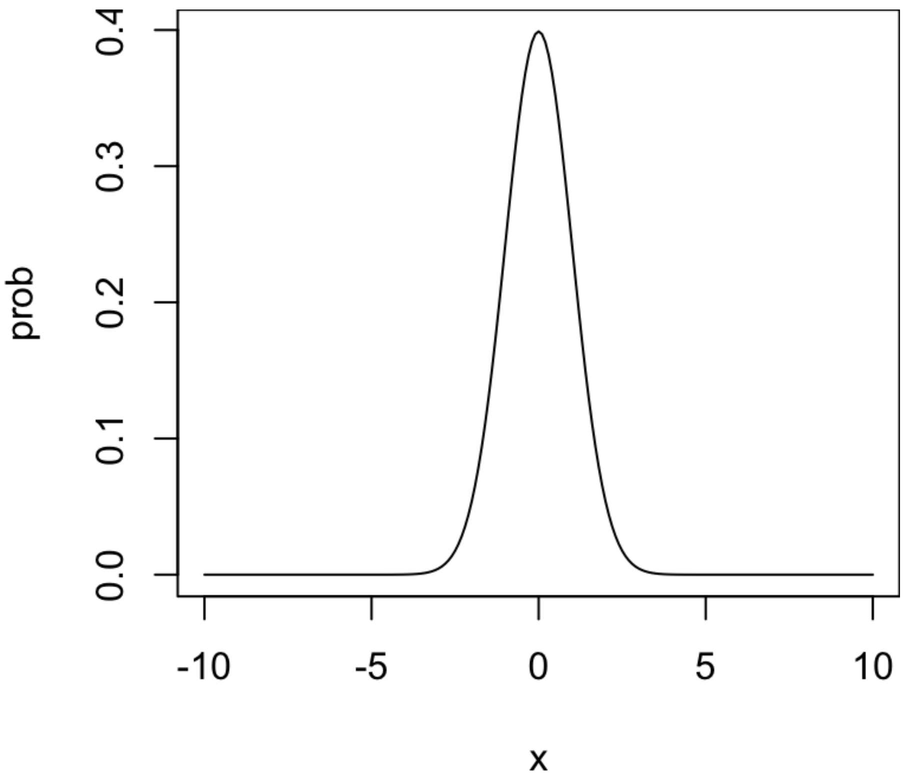
Cumulative distribution function (CDF)

$$F(X) = \int_{-\infty}^{X} f(X)dX$$

Normal distribution

Probability density function of $\mathcal{N}(0, 1)$.

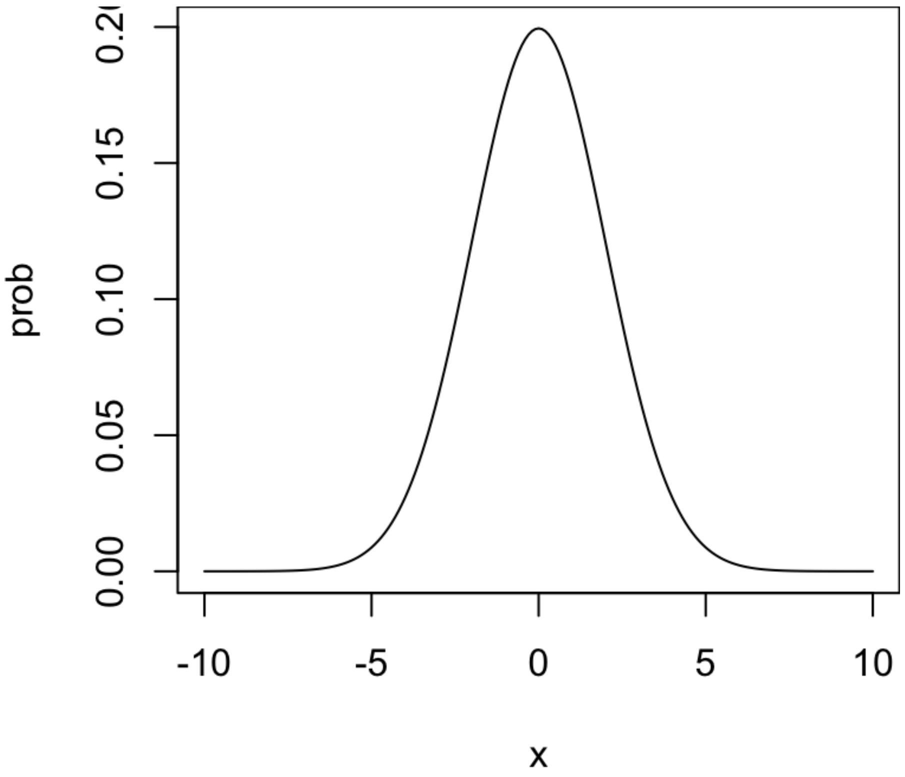
```r
mu <- 0
sigma <- 1

x <- seq(-10, 10, by = .1)

prob <- dnorm(
  x,
  mean = mu,
  sd = sigma
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(x, prob, type = "l")
```



Normal distribution

Probability density function of $\mathcal{N}(0, 2)$.

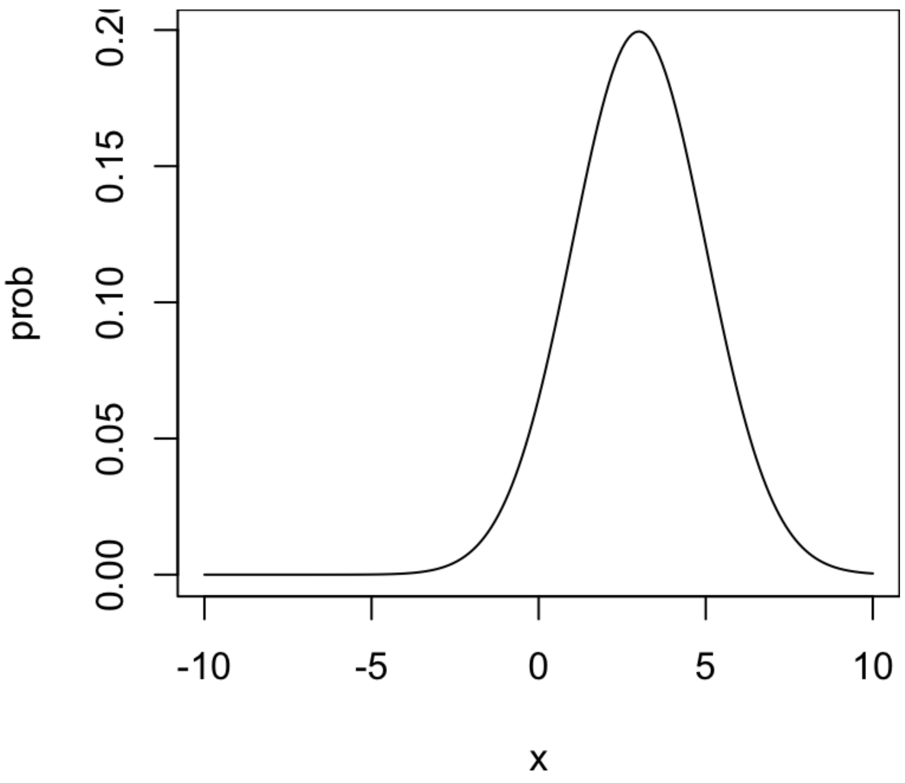
```r
mu <- 0
sigma <- 2

x <- seq(-10, 10, by = .1)

prob <- dnorm(
  x,
  mean = mu,
  sd = sigma
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(x, prob, type = "l")
```



Normal distribution

Probability density function of $\mathcal{N}(3, 2)$.

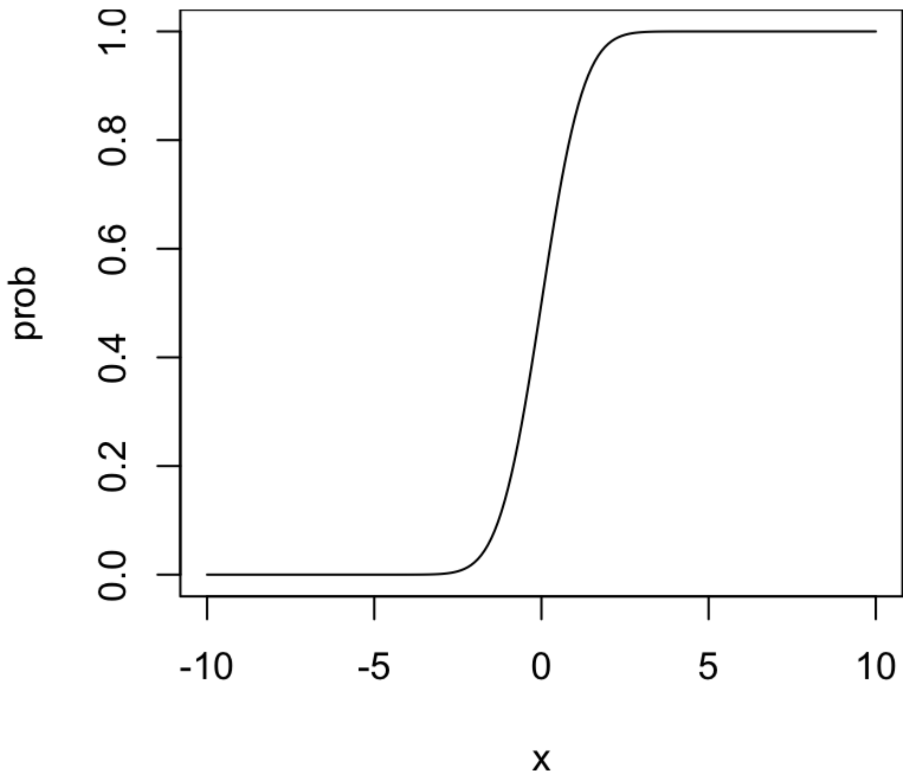
```r
mu <- 3
sigma <- 2

x <- seq(-10, 10, by = .1)

prob <- dnorm(
  x,
  mean = mu,
  sd = sigma
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(x, prob, type = "l")
```



Normal distribution

Cumulative distribution function of $\mathcal{N}(0, 1)$.

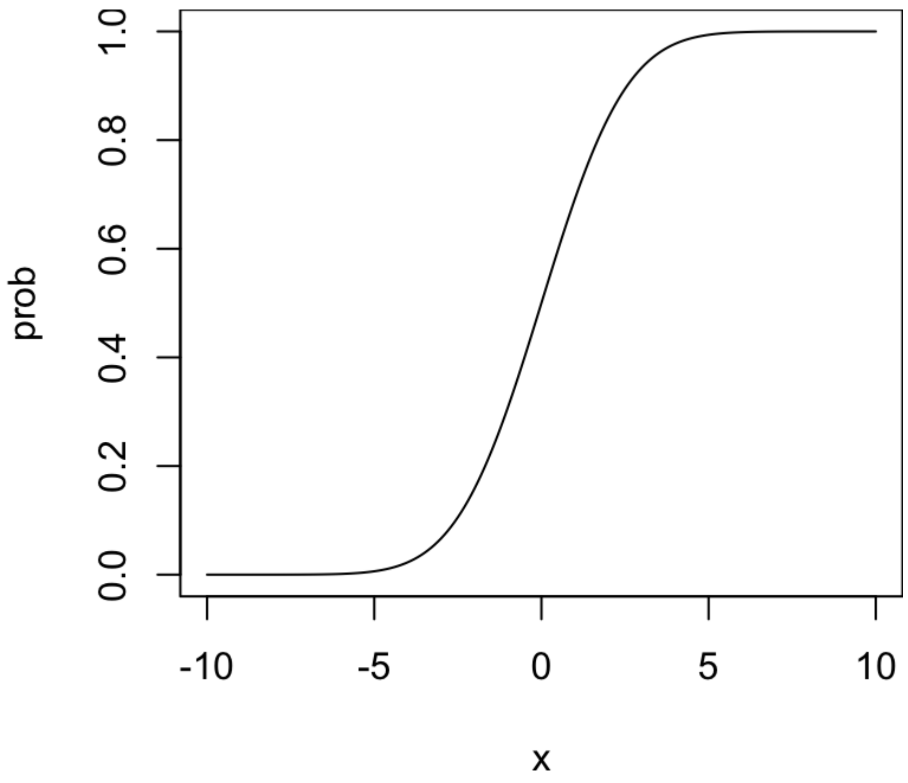
```r
mu <- 0
sigma <- 1

x <- seq(-10, 10, by = .1)

prob <- pnorm(
  x,
  mean = mu,
  sd = sigma
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(x, prob, type = "l")
```



Normal distribution

Cumulative distribution function of $\mathcal{N}(0, 2)$.

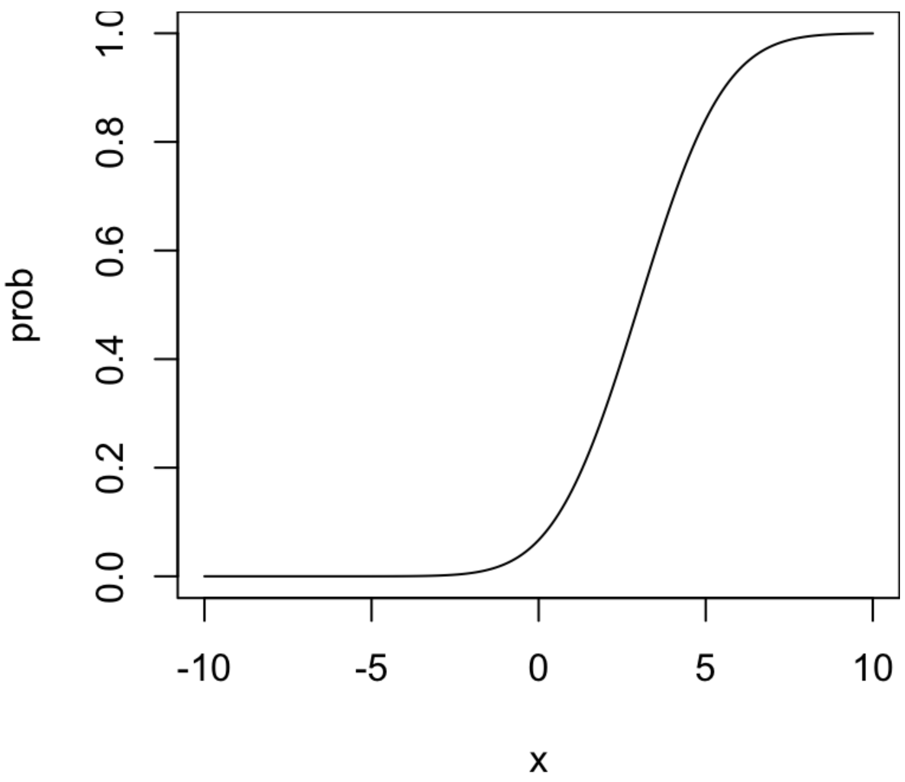
```r
mu <- 0
sigma <- 2

x <- seq(-10, 10, by = .1)

prob <- pnorm(
  x,
  mean = mu,
  sd = sigma
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(x, prob, type = "l")
```



Normal distribution

Cumulative distribution function of $\mathcal{N}(3, 2)$.

```r
mu <- 3
sigma <- 2

x <- seq(-10, 10, by = .1)

prob <- pnorm(
  x,
  mean = mu,
  sd = sigma
)

par(mar = c(5.1, 4.1, 0.01, 0.01))

plot(x, prob, type = "l")
```



Normal distribution

Given $X \sim \mathcal{N}(1, 0)$, the $P(X \leq 1) = F_X(1)$.

```r
mu <- 0
sigma <- 1
x <- 1

pnorm(
  x,
  mean = mu,
  sd = sigma
)
```

```
[1] 0.8413447
```

Normal distribution

Given $X \sim \mathcal{N}(1, 0)$, calculate the probability that $x \in [-1, 1]$.

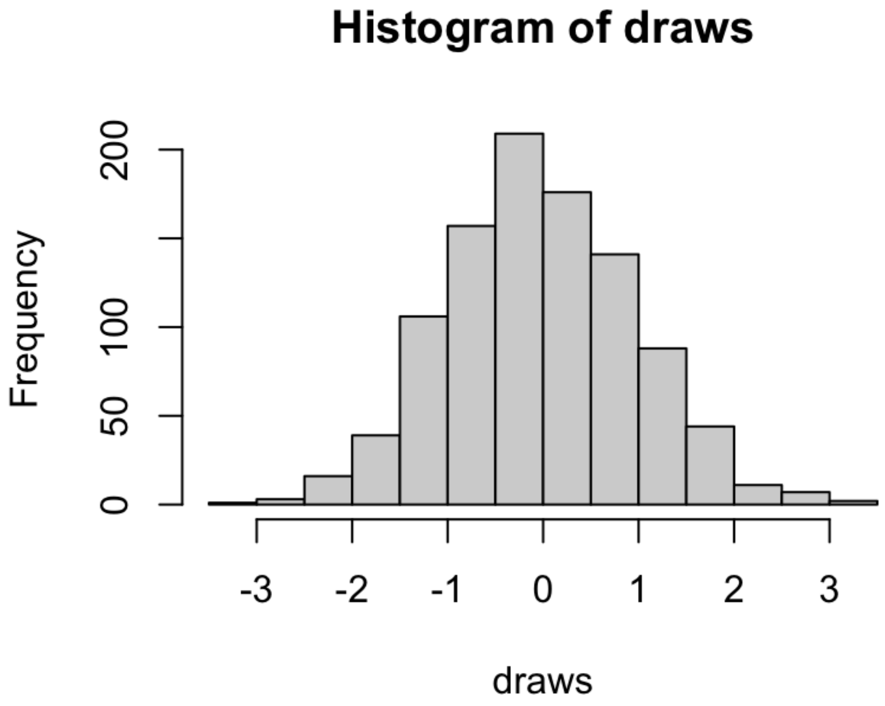
```r
mu <- 0
sigma <- 1
x_1 <- -1
x_2 <- 1

p_1 <- pnorm(x_1, mu, sigma)

p_2 <- pnorm(x_2, mu, sigma)

p_2 - p_1
```

```
[1] 0.6826895
```

Normal distribution

Simulate 1000 draws from $X \sim \mathcal{N}(1, 0)$

```r
mu <- 0
sigma <- 1
n <- 1000

draws <- rnorm(n, mu, sigma)

par(mar = c(5.1, 4.1, 4.1, 0.01))

hist(draws)
```



Lab assignment

Lab assignment

- Similar structure to Lab 1: R programming
- You need to hand in correct solution with working code for **all** excercises.
- Work individual:
  - One report per student.
  - No duplicate reports.
- You can only use base R, no external libraries.
- No copy paste from internet or other sources.

Lab assignment

- You upload a PDF report in Canvas.
- PDF should contain:
  - Text explanation of your solution (choices etc.)
  - R code (carefully commented)
  - Any plots/figures should be included

Lab assignment

- Make sure code is easy to test the code by copy-pasting it into an R-session.
- Reports will be checked for plagarism.
- Two opportunities to submit.

Resources

- Introduction to Probability (free online book): http://probabilitybook.net
- Harvard's "Statistics 110: Probability"-lectures: https://projects.iq.harvard.edu/stat110/youtube
- Base R cheat sheet: https://github.com/rstudio/cheatsheets/blob/main/base-r.pdf
- R help pages: `?distributions`, `?plot` …