

# Statistical programming

Overview of probability. Simulation

Marc Comas-Cufí



Statistical Programming Marc  
Comas and Karina Gibert Statistical  
Programming Marc Comas  
and Karina Gibert Statistical Pro-  
gramming Marc Comas and Karina  
Gibert Statistical Programming  
Marc Comas and Karina Gibert

# Today's session

- Overview of probability
  - Probability
  - Random variables
  - The Central limit theorem
  - Simulation

# A brief review of probability

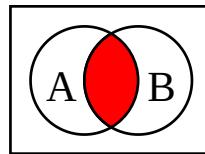
# Event

- An *event* is a set of possible results in a particular experiment.
- The event containing all possible events/results is called the *sample space*.
- For completeness, the *impossible event* is defined and it is denoted as  $\emptyset$ .
- Example:
  - “Face 6” is a possible event when rolling a six-sided dice.
- Events are nicely represented with Euler diagrams:

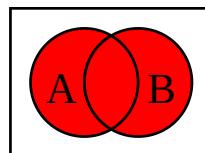
# Basic event operations

Given two events  $A$  and  $B$ .

- $A \cap B$ . Intersection of events  $A$  and  $B$ . Event defined as  $A$  and  $B$  occur.



- $A \cup B$ . Union of events  $A$  and  $B$ . Event defined as either  $A$  or  $B$  occur.



- $A^c$ . Complement of an event  $A$ . Event defined as  $A$  does not occur.

# Probability (of an event)

- The *probability of an event* measures “how likely” the event is or “how big” with respect the sample space the event is.
- The probability of an event  $A$  is denoted as  $P(A)$ .
- Probabilities are quantities between 0 and 1, i.e. for any event  $A$ ,  $0 \leq P(A) \leq 1$ .
- $P(\emptyset) = 0$  and  $P(\Omega) = 1$ , where  $\Omega$  denotes the sample space.

What are probabilities?

- **Frequentist.** Probabilities are long run relative frequencies of events.
- **Bayesian.** Probabilities are used to quantify our uncertainty about events.
- Example:
  - “Face 6” event has probability  $1/6$  when rolling a six-sided dice.

# Conditional probability

The conditional probability of one event  $A$  given another event  $B$  is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- $P(A|B)$  measures how likely is to happen  $A$  once we know  $B$  has happened.
- We can think conditioning as if we were reducing our sample space with events where  $B$  occurs.
- Example: “Face 6” event has probability  $1/3$  once we are told the result is even.

## Bayes' rule

For two events  $A$  and  $B$ . Then,

$$\begin{aligned} P(A|B) &= \frac{P(A)P(B|A)}{P(B)} \\ &= \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A^c)P(B|A^c)}. \end{aligned}$$

# Simulating events with R (1)

- We can randomize a set of possible events with function `sample()`.

```
1 cards = c('2', '3', '4', '5', '6', '7', '8', '9', '10', 'J', 'Q', 'K', 'A')
2 sample(cards)
3 #> [1] "Q"   "10"  "J"   "8"   "9"   "7"   "A"   "4"   "3"   "K"   "6"   "5"   "2"
```

- We can pick a certain number of events:

```
1 sample(cards, 5)
2 #> [1] "A" "9" "8" "7" "Q"
```

- By default, `sample()` picks elements without replacement. To force replacement:

```
1 sample(cards, 5, replace = TRUE)
2 #> [1] "Q" "7" "3" "A" "A"
```

# Simulating events with R (2)

- We can assign a probability to each event:

```
1 breast_cancer = c('yes', 'no')
2 sample(breast_cancer, 200, replace = TRUE, prob = c(0.004, 0.996))
3 #> [1] "no"  "no"
4 #> [13] "no"  "no"
5 #> [25] "no"  "no"
6 #> [37] "no"  "no"
7 #> [49] "no"  "no"
8 #> [61] "no"  "no"
9 #> [73] "no"  "no"
10 #> [85] "no"  "no"
11 #> [97] "no"  "no"
12 #> [109] "no"  "no"
13 #> [121] "no"  "no"
14 #> [133] "no"  "no"
15 #> [145] "no"  "no"
16 #> [157] "no"  "no"
17 #> [169] "no"  "no"  "no"  "no"  "no"  "no"  "no"  "no"  "yes" "no"  "no"  "no"
18 #> [181] "no"  "no"
19 #> [193] "no"  "no"
```

# Activity: Medical diagnosis

Suppose you are a women in your 40s, and you decide to have a medical test for breast cancer called a mammogram.

- If the test is positive, what is the probability you have cancer?

Information:

- The probability of having breast cancer at 40s is 0.004.
- The test has a **sensitivity** of 80%. In other words, if you have breast cancer, the test will be positive with probability 0.8.
- The test has a **specificity** 90%. In other words, if you don't have breast cancer, the test will be negative with probability 0.9.

# Activity: Medical diagnosis

```
1 library(tidyverse)
2 N = 100000
3 test_result_sampling = function(breast_cancer){
4   if(breast_cancer == 'yes'){
5     sample(c('+', '-'), 1, prob = c(0.8, 0.2))
6   }else{
7     sample(c('+', '-'), 1, prob = c(0.1, 0.9))
8   }
9 }
10 women40s = tibble(
11   breast_cancer = sample(c('yes', 'no'), N, replace = TRUE, prob = c(0.004, 0.996)),
12   test_result = map_chr(breast_cancer, ~test_result_sampling(.x)))
13 )
```

- What is the proportion of women with breast cancer in the population having a positive test?

# Activity: Medical diagnosis

Use probability theory to calculate the exact probability of a women having breast cancer once we know she got a positive test.

**Hint.** Consider the following events:

- $A$  = “To have cancer”,  $A^c$  = “Not to have cancer” and
- $B$  = “Test is positive”.

# Independence between events

- Two events  $A$  and  $B$  are said to be independent, denoted  $A \perp B$ , if

$$P(A) = P(A|B).$$

- A practical equivalent definition is  $A$  and  $B$  are independent if

$$P(A \cap B) = P(A)P(B).$$

# Conditional independence

- Two events  $A$  and  $B$  are said to be conditionally independent given  $C$ , denoted  $A \perp B|C$ , if

$$P(A|C) = P(A|B \cap C),$$

or

$$P(A \cap B|C) = P(A|C)P(B|C).$$

# Activity

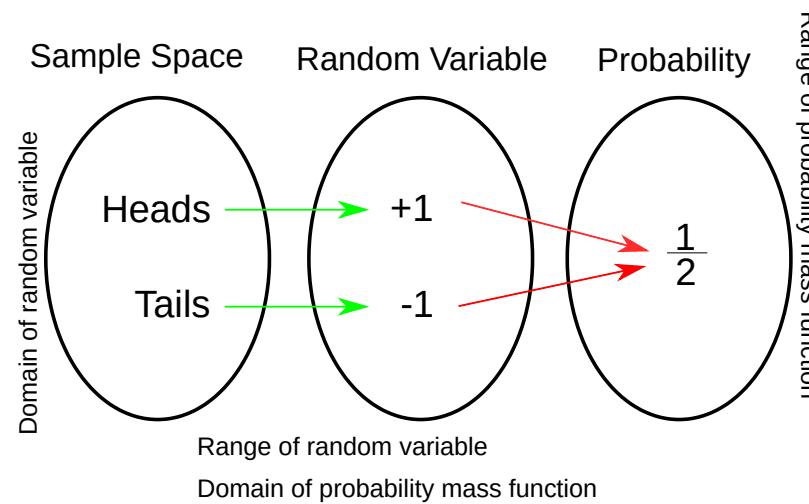
- Check if event A = “face is even” and event B = “the face is smaller or equal than 4” are independent when rolling a 6-faced dice.
- Check if event A = “face is 6 in the blue dice” and B = “face is 6 in the red dice” are independent when rolling a blue and red 6-faced dice.
- Check if event A = “face is 6 in the blue dice” and B = “face is 6 in the red dice” are conditionally independent when rolling a blue and red 6-faced dice when we know about event C = “the sum is even”.

# Activity: Monty Hall problem

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

# Random variables

- A random variable (r.v.) is a function from a sample space to a measurable space, like the real numbers.
- We can think of a random variable as a variable taking values randomly.
- If values are discrete, the function assigning a probability to its values is called *probability mass function* (pmf).

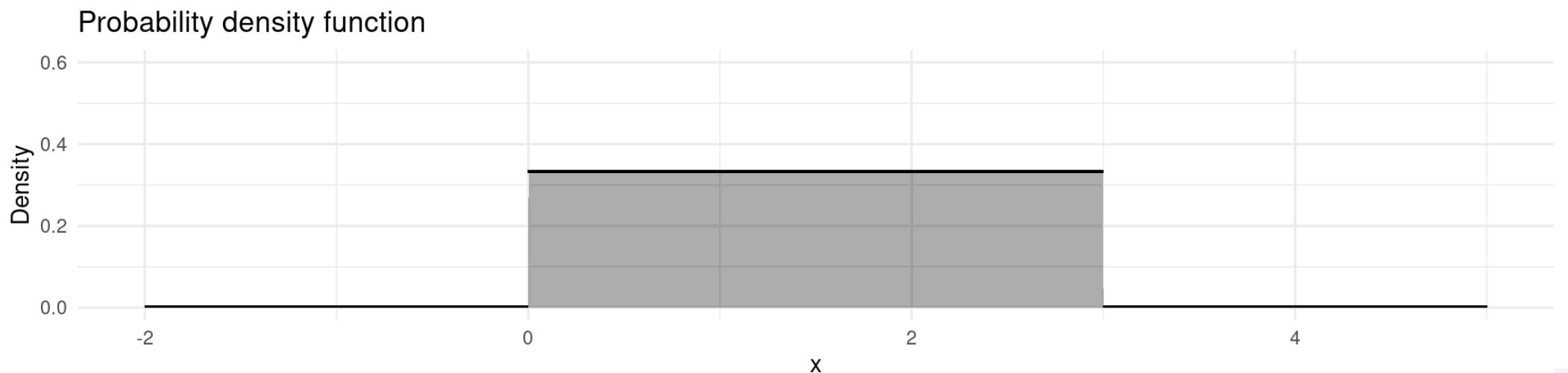


# Discrete random variables

- Discrete r.v. take values in a countable set of elements. For example a finite set of numbers or the integers.
- Discrete r.v. can be characterised with giving  $P(X = x)$  for each possible value  $x$ . Function  $p(x) = P(X = x)$  is called the *probability mass function* (pmf).
- **Example:** Discrete r.v.  $X$  is uniquely determined by
  - $X \in \{1, 2, 3\}$  and
  - the pmf:  $p(1) = 0.25, p(2) = 0.5$  and  $p(3) = 0.25$ .

# Continuous random variables

- Continuous random variables take values in the real line.
- Possible events for continuous r.v. are combinations of subsets the real line.
- Continuous r.v. are characterised by providing  $P(X \in E)$  for any event  $E$ .
- The *probability density function* (pdf) of a continuous r.v. is a function  $f(x)$  such that  $P(X \in E) = \int_{E} f(x) dx$
- **Example:** Continuous r.v.  $X$  is uniquely determined with the pdf  $f(x)$  defined as  $f(x) = \frac{1}{3}$  if  $x \in (0, 3)$  and  $f(x) = 0$  otherwise.



# The expected value or mean of a r.v.

- The *expected value* of a r.v.  $X$ , denoted  $\mathbb{E}[X]$ , is the mean value we expect after an infinite number of runs.
  - For a discrete r.v.  $X$ , the expected value is

$$\mathbb{E}[X] = \sum_x x p(x) \quad [ \sum_x x P(X = x) ]$$

\* For a continuous r.v.  $X$ , the expected value is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$

- Example: the expected value of the r.v. with the value of rolling a 6-faced dice is

$$1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

# More expected values

- The variance of a r.v., denoted  $\text{var}[X]$ , is defined as

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

- The covariance between two r.v.'s  $X$  and  $Y$ , denoted  $\text{cov}[X, Y]$ , is defined as

$$\text{cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

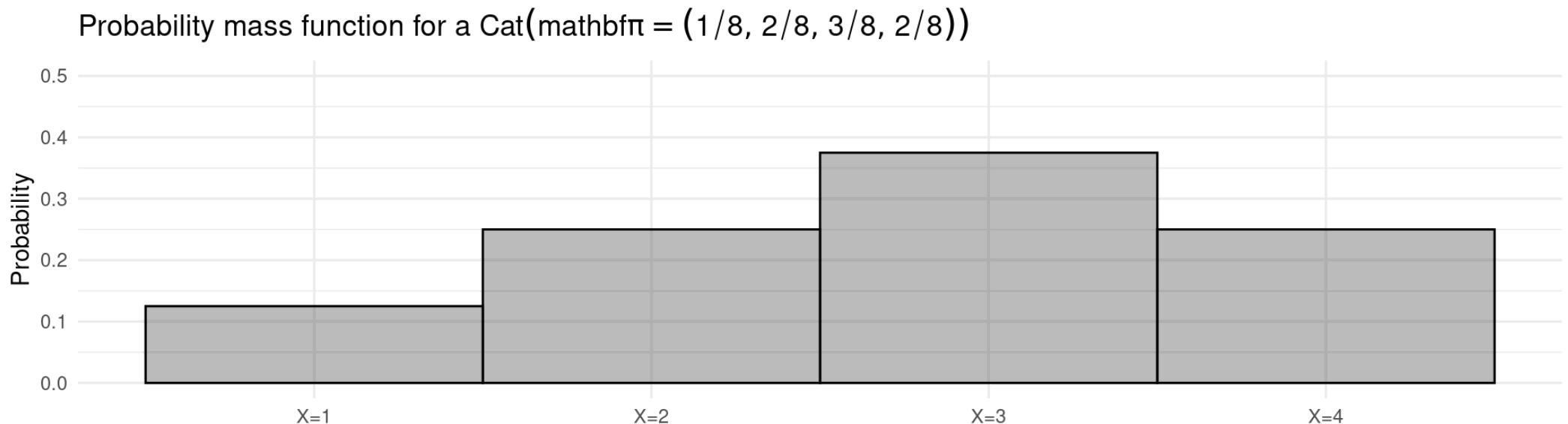
- See visualisation at <https://seeing-theory.brown.edu/basic-probability/index.html>
- For any function  $g(x)$ , we can define the expected value of  $g(X)$  as
  - $\mathbb{E}[g(X)] = \sum_x g(x) p(x)$  when  $X$  is discrete or
  - $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$  when continuous.

# Some common probability distributions

# Categorical distribution

The categorical distribution is the distribution obtained after associating natural number 1 to  $k$  to a set of events with a certain probability  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$ .

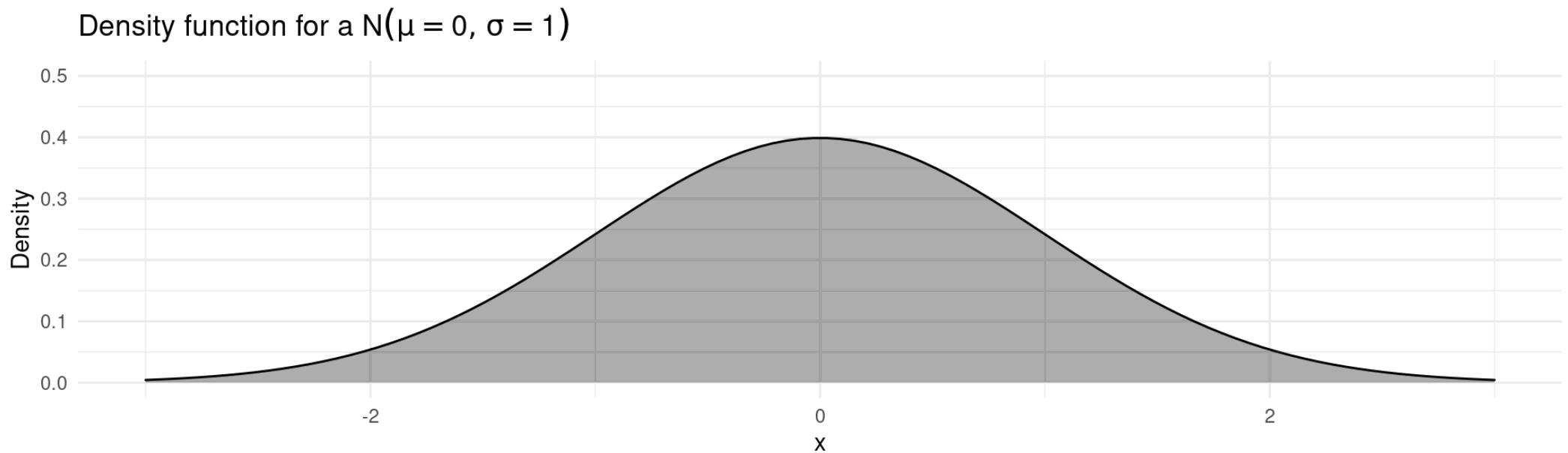
- For a categorical r.v.  $X$ , denoted  $X \sim \text{Cat}(\boldsymbol{\pi} = (\pi_1, \dots, \pi_k))$ ,
  - $\mu = \mathbb{E}[X] = \sum_{i \leq k} i \pi_k$  and
  - $\text{var}[X] = \sum_{i \leq k} (i - \mu)^2 \pi_k$ .



# Gaussian (normal) distribution

The gaussian (normal) distribution is the distribution that reduces the probability exponentially (with velocity  $1/\sigma^2$ ) for events far from a certain center ( $\mu$ ).

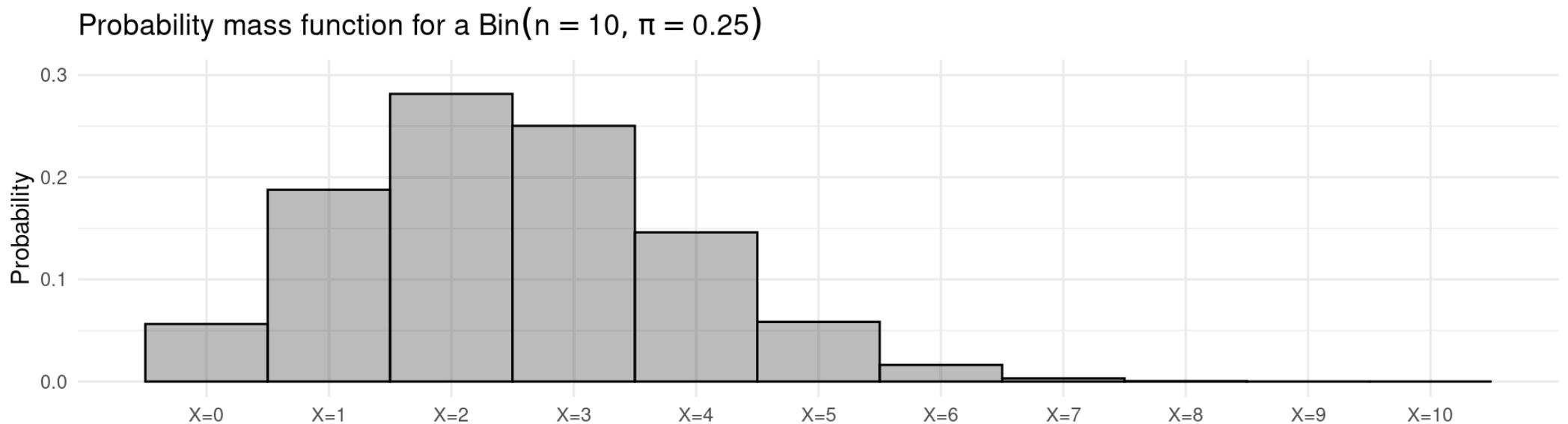
- For a gaussian r.v.  $X, X \sim N(\mu, \sigma)$ ,
  - $\mathbb{E}[X] = \mu$  and
  - $\text{var}[X] = \sigma^2$ .



# Binomial distribution

The binomial distribution counts the number of success after repeating a certain experiments  $n$  times when the probability of success is  $\pi$ .

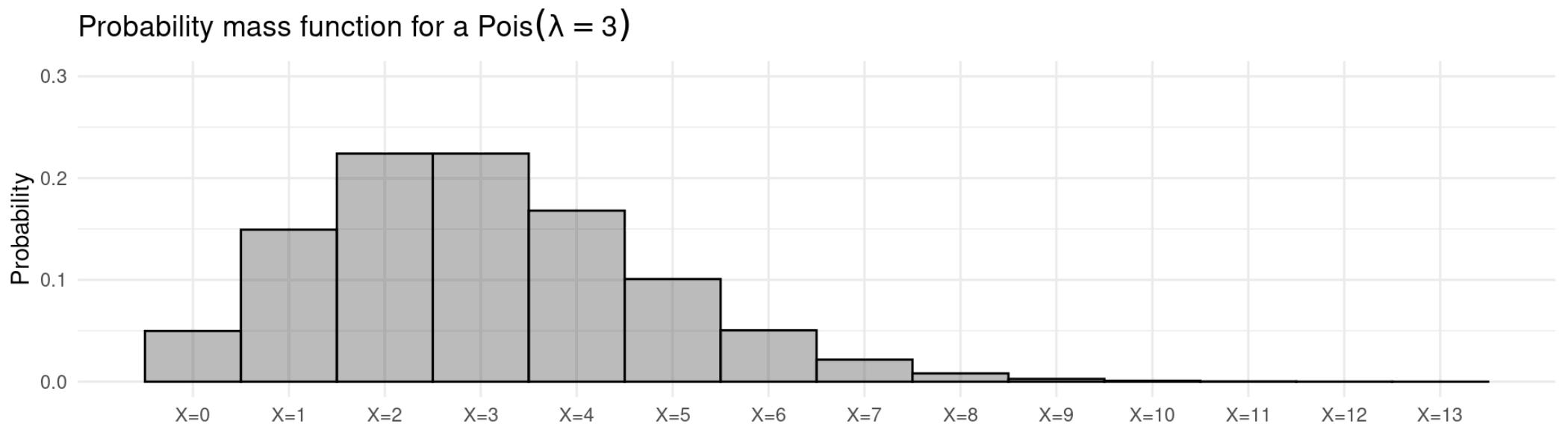
- For a binomial r.v.  $X, X \sim Bin(n, \pi)$ ,
  - $\mathbb{E}[X] = n\pi$  and
  - $\text{var}[X] = n\pi(1 - \pi)$ .



# Poisson distribution

The Poisson distribution counts the number of success detected in an interval of time or space when the expected number of successes is  $\lambda$ .

- For a Poisson r.v.  $X, X \sim Poiss(\lambda)$ ,
  - $\mathbb{E}[X] = \lambda$  and
  - $\text{var}[X] = \lambda$ .



# Generating r.v. with R

- Categorical r.v.

```
1 sample(1:4, 6, prob = c(1/8,2/8,3/8,2/8), replace=TRUE)
2 #> [1] 2 4 4 4 2 2
```

- Gaussian r.v.

```
1 rnorm(6, mean = 0, sd = 1)
2 #> [1] 1.6741280 -1.5617505 1.0604928 1.3124341 1.5795088 -0.5374763
```

- Binomial r.v.

```
1 rbinom(6, size = 10, prob = 0.25)
2 #> [1] 5 3 3 1 3 2
```

- Poisson r.v.

```
1 rpois(6, lambda = 3)
2 #> [1] 8 2 2 2 2 5
```

# Activity

Thing about possible variables that can be modelled with one of the four seen distributions:

- Categorical distribution
- Gaussian distribution
- Binomial distribution
- Poisson distribution

Think about possible parameters to use.

# Joint probability distribution

The *joint probability distribution* is a multivariate model for random variables. The joint probability distribution of r.v.  $X_1, \dots, X_k$  is completely determined by providing

- the probability of all possible events for discrete r.v.'s

$$p(x_1, \dots, x_k) = P(X_1 = x_1 \cap \dots \cap X_k = x_k),$$

- or the multivariate density for continuous r.v's

$$f(x_1, \dots, x_k).$$

# Transformations of random variables

- If we transform a r.v. with a function  $g(x)$  we get another r.v.
- Linear transformations. If we linearly transform a r.v.  $X = (X_1, \dots, X_k)$  with  $f(x) = Ax + b$  then
  - $\mathbb{E}[f(X)] = A\mu + b$  where  $\mu = \mathbb{E}[X]$ , and
  - $\text{cov}[f(X)] = A\Sigma A^T$  where  $\Sigma = \text{cov}[X]$ .
- For non-linear transformations we can simulate the r.v. to obtain these expected values.

**Example:** The log-normal distribution is defined as the distribution obtained after exponentiating a gaussian r.v.,  $f(x) = e^x$ ,  $Y = e^X$ , where  $X \sim N(\mu, \sigma)$ . Check using simulating gaussian r.v. that

- $\mathbb{E}[Y] = e^{\mu+\sigma^2/2}$  and
- $\text{var}[Y] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$ .

# The central limit theorem

If  $X_1, \dots, X_n$  are r.v. equally distributed with expected value  $\mathbb{E}[X]$  and variance  $\text{var}[X]$ . Then, when  $n$  is big, we have

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \sim N(\mathbb{E}[X], \sqrt{\text{var}[X]/n})$$

---

Next plot shows the distribution of 1000 realisations of a r.v. obtained as the average of 30 binomial distributions with parameters  $n = 10$  and  $\pi = 0.25$ . Blue line represents the gaussian distribution with  $\mu = n\pi$  and  $\sigma^2 = \sqrt{n\pi(1 - \pi)/30}$

# The central limit theorem

If  $X_1, \dots, X_n$  are r.v. equally distributed with expected value  $\mathbb{E}[X]$  and variance  $\text{var}[X]$ . Then, when  $n$  is big, we have

$$\frac{\bar{X} - \mathbb{E}[X]}{\sqrt{\text{var}[X]/n}} \sim N(0, 1)$$

---

Next plot shows the distribution of 1000 standardised realisations of a r.v. obtained as the average of 30 binomial distributions with parameters  $n = 10$  and  $\pi = 0.25$ . Blue line represents the gaussian distribution with  $\mu = 0$  and  $\sigma^2 = 1$

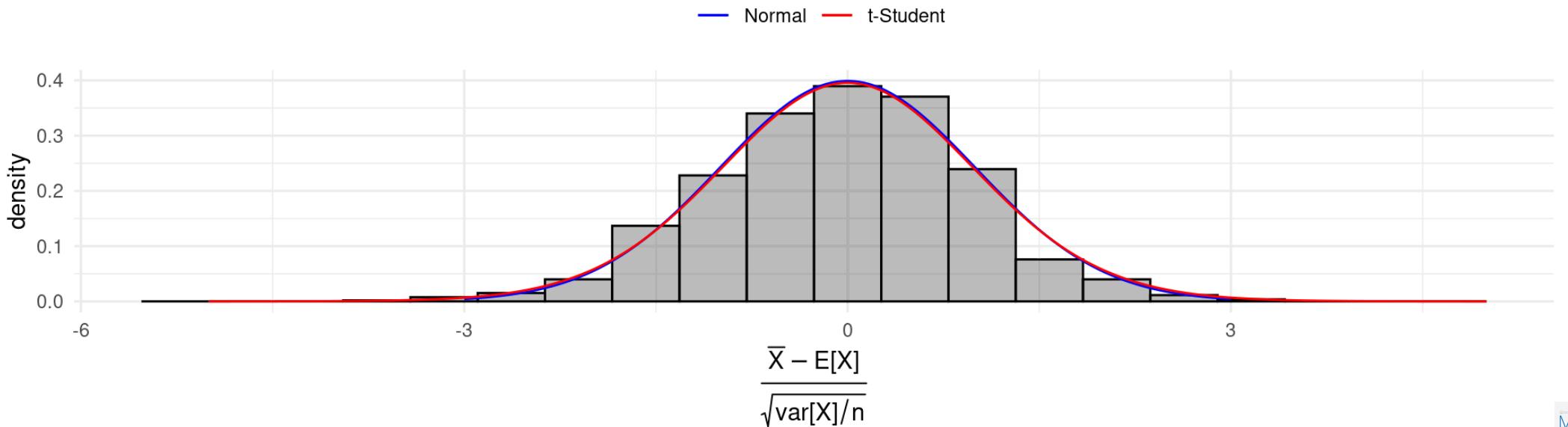
# Sampling distribution

If  $x_1, \dots, x_n$  is a sample i.i.d with expected value  $\mathbb{E}[X]$ . Then, when  $n$  is big, we have

$$\frac{\bar{x} - \mathbb{E}[X]}{\sqrt{s_x/n}} \sim t_{n-1}$$

---

Next plot shows the distribution of 1000 standardised realisations of a r.v. obtained as the average of 30 binomial distributions with parameters  $n = 10$  and  $\pi = 0.25$ . Normal and Student distributionss are compared.



# Final activity

- Fix some r.v.  $Y$  after some transformation  $f(x)$ .
  - For example by transforming certain binomial distribution with function  $f(x) = \sin(x)$ .
- Take a sample of size 50 and calculate the mean of this sample. We we call it:  $x_1$ .
- Repeat the process 1000 times obtaining a sample of means:  $X_m = \{x_1, \dots, x_{1000}\}$ .
- Visualise the resulting distribution of sample  $X_m$ .
- Think about how we can use  $x_i$  to obtain information of  $\mathbb{E}[Y]$ ?

**That's all for today**

# Next week session

- Overview of probability and statistics
  - Chi-squared test: goodness of fit, independence
  - Student's t-test: one sample, paired samples, independent two-sample
  - One-way ANOVA test
  - Normality test
  - Homocedasticity test