Probability Distributions

Giovanni Saraceno

Contents

Probability Distributions

The concepts of probability and randomness are central to statistics. In particular, to understand statistical methods, it is important to view data as samples derived from distributions.

This section outlines the basic ideas of probability and functions in `R` for random sampling and handling theoretical distributions.

Random Sampling

Early examples in probability theory primarily dealt with gambling and games where the core concept was random sampling, such as shuffling a deck of cards or drawing numbered balls. In `R`, we can simulate such situations using the `sample` function. For example, to draw 5 numbers (randomly!) from the set `1:90`, we can use:

```
sample(1:90, 5)
```

```
## [1]  4 28 72 39 46
```

The first argument, `x`, is a vector indicating the set to sample from, while `size` specifies the sample size. By default, the sample function samples **without replacement** (i.e., the sample cannot contain duplicate values), so the sample size cannot exceed the length of the set. To sample **with replacement**, use the option `replace = TRUE`. For example, to simulate 10 coin tosses:

```
sample(c("H", "T"), 10, replace = TRUE)
```

```
##  [1] "H" "H" "H" "T" "H" "H" "T" "H" "T" "H"
```

In a fair coin toss, the events "Heads" and "Tails" are equally likely (i.e., each has a probability of $\frac{1}{2}$). In R, we can also consider cases where the events are not equally probable using the `prob` option:

```
sample(c("H", "T"), 10, replace = TRUE, prob = c(0.9, 0.1))
```

```
##  [1] "H" "H" "H" "H" "H" "H" "H" "H" "H" "H"
```

Note: The sum of the values in the prob vector must equal 1.

Combinatorics

Consider the example of sampling 5 numbers without replacement. The probability of a specific number being drawn first is $\frac{1}{90}$, for the second $\frac{1}{89}$, and so on. Thus, the probability of a specific sample is:

```
1 / prod(90:86)
```

```
## [1] 1.896126e-10
```

This is the probability of drawing specific numbers. If this scenario corresponds to a lottery, we are interested in the probability of guessing a specific set of 5 numbers. In this case, we must account for all possible orders of the 5 numbers, which is 5! or $5 \times 4 \times 3 \times 2 \times 1$. The probability of winning the lottery is:

```
factorial(5) / prod(90:86)
```

```
## [1] 2.275351e-08
```

Alternatively, we can calculate the total number of ways to choose 5 elements from 90 using the binomial coefficient:

$$\binom{90}{5} = \frac{90!}{5!85!} = 43949268$$

In R, we use the choose function:

```
1 / choose(90, 5)
```

```
## [1] 2.275351e-08
```

Distributions in R

Consider independent replications of a given experiment. From a probabilistic perspective, we are often less interested in individual outcomes and more focused on the total number of outcomes. This result is random and thus described by a random variable.
Remark: A *random variable* is a real-valued function from an outcome space into the real line more generally into R^n . The probability law defined on the outcome space induces (defines explicitly) a probability model for the random variable. It is the basic quantity used in probability theory to characterize a probability process. Random variables are categorized as discrete or continuous.

A discrete random variable X takes values in a finite set and is characterized by its probability mass function $f(x) = P(X = x)$ or its cumulative distribution function $F(x) = P(X \leq x)$.

A (uni-variate) continuous random variable can take values in the real line and is characterized by its density function $f(x)$ and distribution function (or cumulative distribution function):

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

`R` includes implementations of major probability distributions, both discrete and continuous, as these are central to statistical modeling and hypothesis testing (discussed later), replacing traditional statistical tables. Examples include:

Distribution	**R Name**
Binomial	`binomial`
Chi-squared	`chisq`
Exponential	`exp`
Geometric	`geom`
Poisson	`pois`
Normal	`norm`
t-Student	`t`

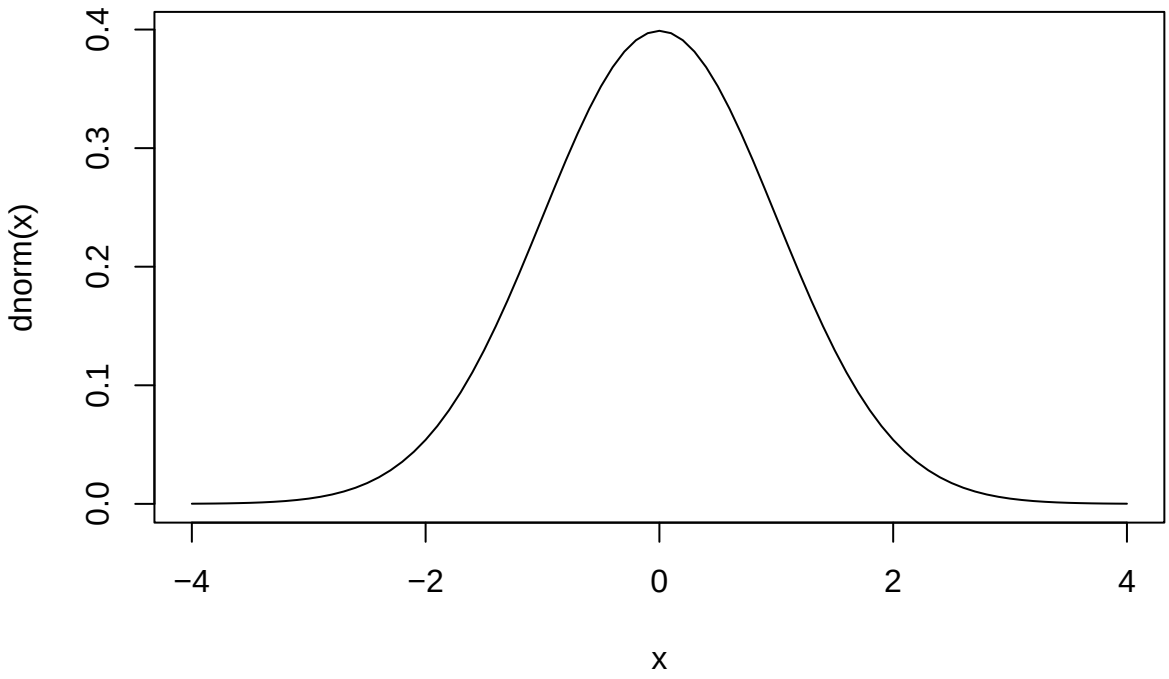
Each distribution allows for four fundamental operations:Probability density/mass, Cumulative distribution, Quantiles, Random number generation. By adding a prefix to the distribution name in R, we can compute these quantities:

prefix	function
d	Density
p	Cumulative distribution
q	Quantile
r	Random Generation

Density

The density function is rarely used directly but is helpful for plotting:

```
x <- seq(-4, 4, 0.1)
plot(x, dnorm(x), type = "l")
```



or using gggplo2

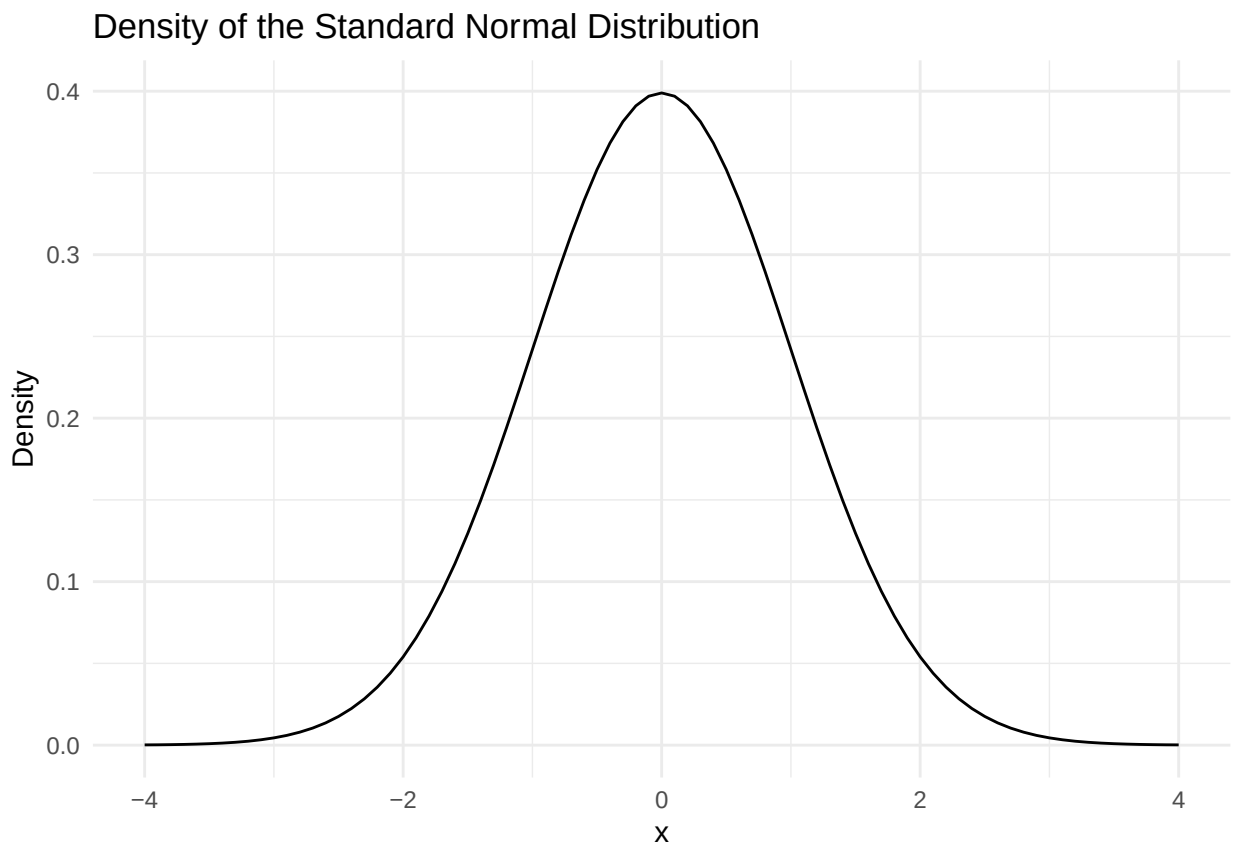
```
library(ggplot2)
```

```
## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.3.3
```

```
x <- seq(-4, 4, 0.1)
df <- data.frame(x = x, density = dnorm(x))

ggplot(df, aes(x = x, y = density)) +
  geom_line() +
```

```
  labs(title = "Density of the Standard Normal Distribution",
       x = "x", y = "Density") +
  theme_minimal()
```

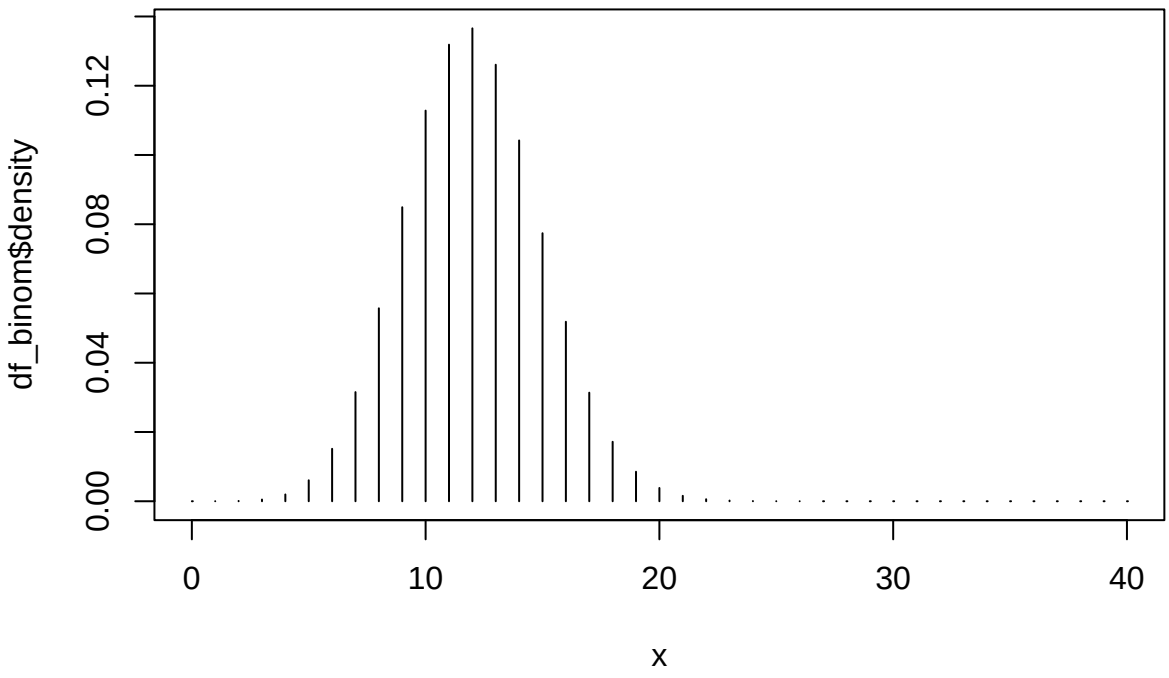


Alternatively, we can use:

```
curve(dnorm(x), from = -4, to = 4)
```

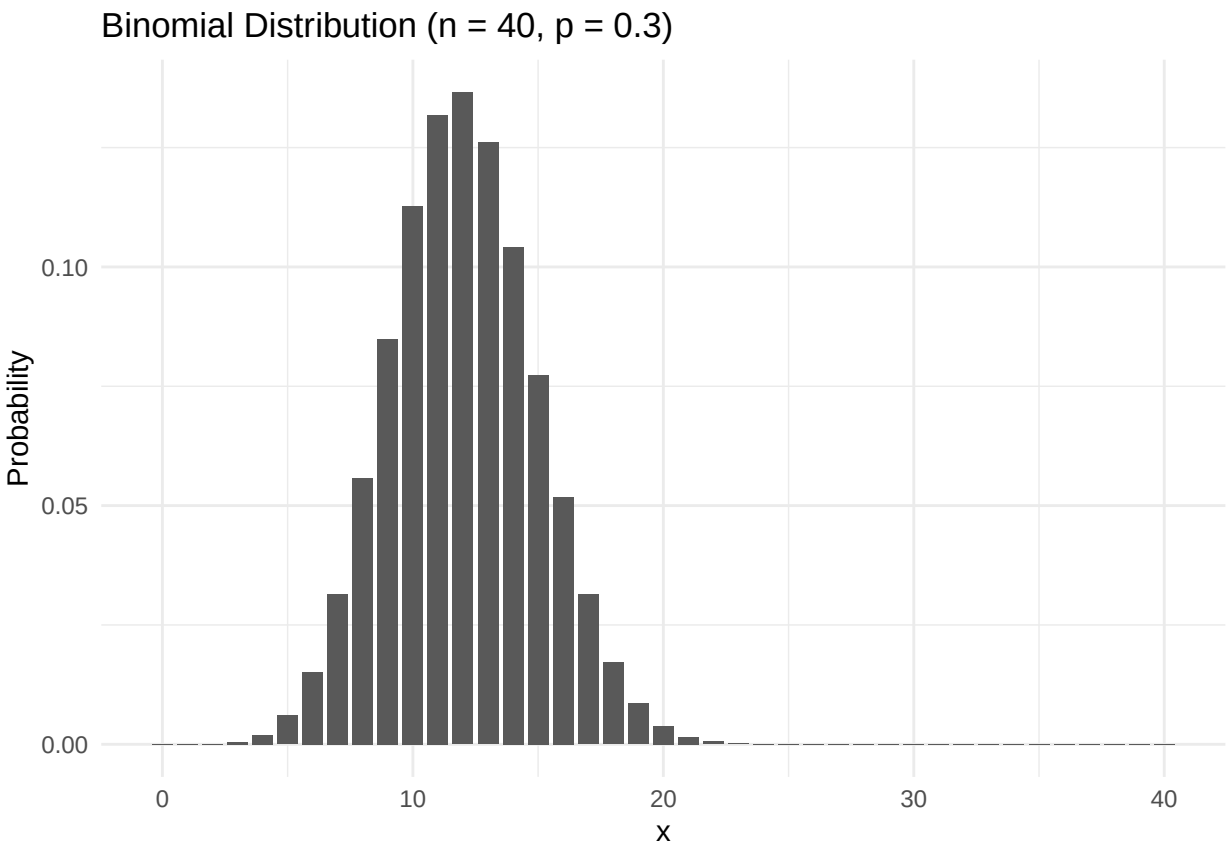
For discrete distributions, a "stick diagram" is often used. For example, the binomial distribution with $n = 40$,$p = 0.3$:

```
x <- 0:40
df_binom <- data.frame(x = x, density = dbinom(x, size = 40, prob = 0.3))
plot(x, df_binom$density, type = "h")
```



or

```
ggplot(df_binom, aes(x = x, y = density)) +
  geom_col(width = 0.8) +
  labs(title = "Binomial Distribution (n = 40, p = 0.3)",
       x = "x", y = "Probability") +
  theme_minimal()
```



Cumulative Distribution

These functions are used to calculate probabilities. For example, suppose a biochemical measure in healthy individuals is normally distributed with a mean of 132 and a standard deviation of 13. For a patient with a value of 160:

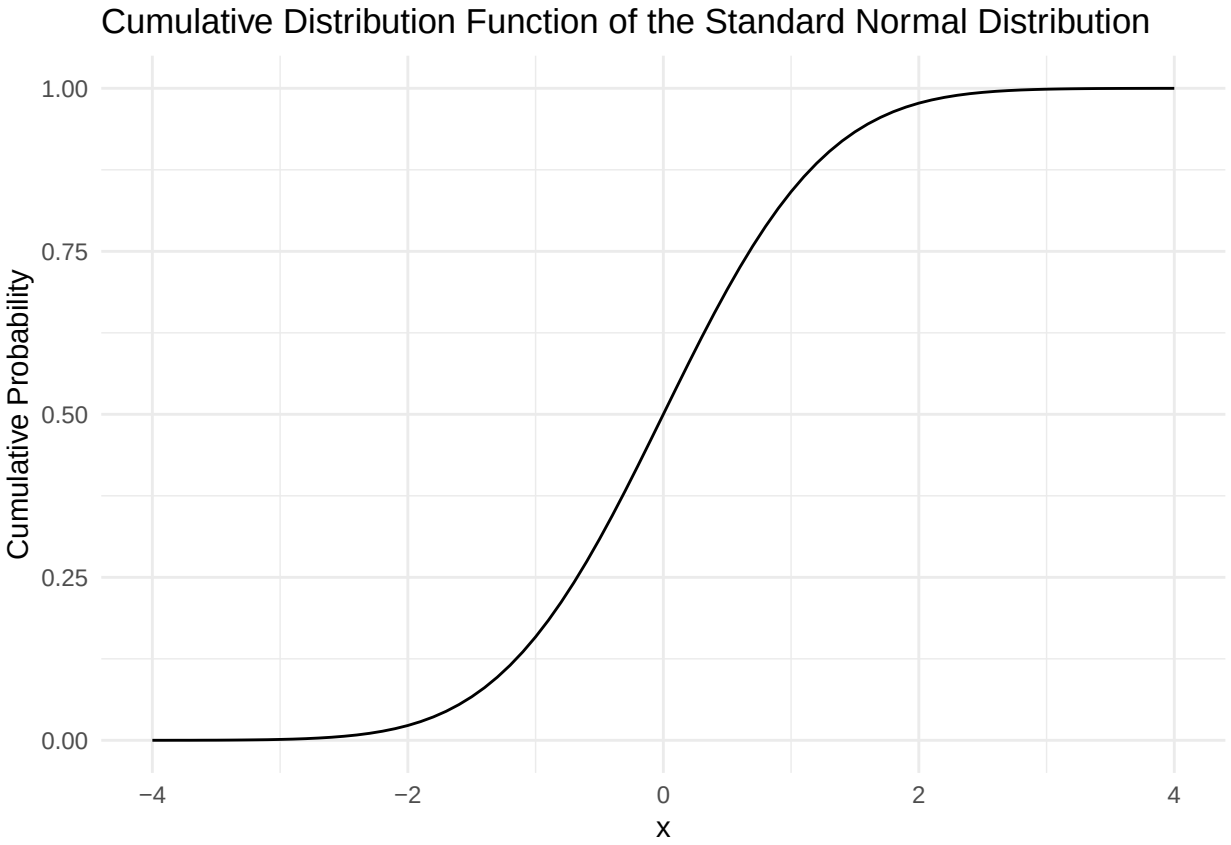
```
1 - pnorm(160, mean = 132, sd = 13)
```

```
## [1] 0.01562612
```

This indicates that approximately 1.5% of the general population exceeds this value. We can visualize the cumulative distribution function of a normal distribution:

```
x <- seq(-4, 4, 0.1)
df_cdf <- data.frame(x = x, cdf = pnorm(x))

ggplot(df_cdf, aes(x = x, y = cdf)) +
  geom_line() +
  labs(title = "Cumulative Distribution Function of the Standard Normal Distribution",
       x = "x", y = "Cumulative Probability") +
  theme_minimal()
```



Quantiles

Quantile functions are the inverse of cumulative distribution functions. For instance, to find z such that $\Phi(z) = 0.9$:

```
qnorm(0.9)
```

```
## [1] 1.281552
```

Random Numbers

We can generate random numbers as realizations of random variables:

```
rnorm(10, mean = 3, sd = 2)
```

```
##  [1] 2.048430 3.479713 3.509233 2.177056 3.245878 3.049808 4.332770 2.742566
##  [9] 1.539767 2.364050
```

```
rbinom(15, size = 50, prob = 0.2)
```

```
##  [1] 13 10  9  6 15  6  8  8 10 12  8 12 14  8 10
```

Random data is useful for studying mathematical approximations via simulation study.

Examples

Bernoulli probability model

Learning Experiment (Jog et al. 1999): To help establish neural correlates of procedural learning, neurons in the striatum of rats are recorded over several days as they executed a procedure learning task.

In this task, the rat used auditory cues to learn which one of two arms of a T-maze to enter in order to receive a reward. On each trial, the rat was placed in a T-maze. A tone was played. If it was a low tone the animal had to go left to receive a reward, whereas if it was a high tone it had to go right to obtain its reward. Suppose that on the previous day, the animal executed this task 40 times and, in so doing, made 22 correct choices and 18 incorrect choices. Before, the start of the 40 trials today, what is the probability that the animal will give a correct response on a given trial?

In this problem, there are only two possible outcomes: a correct response or an incorrect response. The outcomes are mutually exclusive. That is, when one outcome occurs, the other cannot occur. Let p be the probability of a correct response.

The Bernoulli probability model would be a good model for this problem. It defines the probability of a correct or incorrect response on each trial. Either the quantity

```
p_hat = 22/40
```

or $p = 0.5$ would be a reasonable estimate (guess) of p for a trial. Use of the parameter `p_hat` would suggest that we expect the performance on the trials today to be like the performance on trials yesterday. Use of $p = 0.5$ as the guess of p for today would indicate a belief that performance will be indistinguishable from chance on today's trial.

Exercise: Plot the Bernoulli probability mass function and the cumulative distribution.

Binomial probability model

If today we test the rat on the same task and give it 40 trials, and we assume that the probability of a correct choice is p, what are the possible outcomes and what is the probability of each outcome?

On each trial, there can be either a correct or an incorrect response. Across the 40 trials, there can be any combination of correct and incorrect responses such that the sum of correct and incorrect responses equals 40. That is, there are k correct responses and $40 - k$ incorrect responses for $k = 0, \ldots, 40$. If we assume that the trials are independent, and that on each trial the probability of a correct response is p and the probability of an incorrect response is $1 - p$ then the probability of this event is

$$Pr(k \text{ successes}|40 \text{ trials}) = p^k 1 - p)^{40-k}$$

.

Exercise: Plot the Bernoulli probability mass function and the cumulative distribution.

Poisson probability model

The Quantal Release Hypothesis for the release of acetylcholine at the frog motor neuromuscular junction states that in response to stimulation, acetylcholine is released from the motor nerve terminal in discrete "packets" or quanta. Normal endplate potentials (EPPs) are the result of several hundred quanta. Miniature EPP's are the result of spontaneous release of single quanta. An important corollary of the quantal release hypothesis is that there is most likely a large population of quanta in the nerve terminal, each one of which has a small probability of being released by a nerve impulse. We now know that these quanta are packaged in vesicles. For a fixed small time interval (fraction of a millisecond) can we compute the probability that a given number of quanta or vesicles will be released?

To study this problem we can formulate a binomial probability model in which N is the number of quanta or release sites and p is the probability of release in a given small time interval. Let us assume that the release sites behave independently. This then leads to the binomial probability model and hence the probability of observing exactly k quanta released in the specified small time interval. As a practical matter, as N becomes large, we can approximate this calculation by assuming that the probability of release decreases so that $N \times p \to \lambda$. Hence, for N sufficiently large, we have $Np \approx \lambda$ or $p = \frac{\lambda}{N}$.

Exercise: Plot the Poisson probability mass function and the cumulative distribution.

Exercises

Exercise 1 Calculate the probability of each of the following events:

1. $(X > 3)$, where $X \sim N(0, 1)$:

   ```
   ## [1] 0.001349898
   ```

2. $(X > 42)$, where $X \sim N(35, 36)$:

   ```
   ## [1] 0.1216725
   ```

3. $(X = 10)$, where $X \sim Bin(10, 0.8)$:

   ```
   ## [1] 0.1073742
   ```

4. $(X < 0.9)$, where $X \sim N(0, 1)$:

   ```
   ## [1] 0.8159399
   ```

5. $(X > 6.5)$, where $X \sim \chi^2_2$:

   ```
   ## [1] 0.03877421
   ```

Exercise 2 It is known that 5% of the normal distribution lies outside the interval $(-2s, 2s)$, centered at the mean. What are the corresponding intervals for 1%, 0.5%, and 0.1%? What is the position of the quantiles expressed in terms of the standard deviation s?

Exercise 3 Consider a disease where the probability of post-operative complications is 20%. A surgeon suggests a new procedure and tests it on 10 patients, none of whom have complications. What is the probability of operating on 10 patients successfully (without complications) using the traditional procedure?

Exercise 4 The toss of a coin can be simulated in R using rbinom instead of sample. How exactly can this be done?
Hint: Simulate 10 tosses of a fair coin using rbinom.

Exercise 5 Suppose that we have collected the birthdays from 16 students in a class:

January	29
February	17 21 24
March	11 13
April	10 27
May	11
June	11 28
July	6 25
August	17
November	5 15

Plot the probability mass function using the given sample.

References

Jog, Mandar S, Yasuo Kubota, Christopher I Connolly, Viveka Hillegaart, and Ann M Graybiel. 1999. "Building Neural Representations of Habits." *Science* 286 (5445): 1745–49.