

Statistics and Data Analysis - Lab-01

Barbara Żogała-Siudem

2026/2027

1 Probability distributions

Let us start with basic probability distributions:

- $\mathcal{U}(a, b)$ - uniform distribution,
- $\mathcal{N}(\mu, \sigma^2)$ - normal distribution,
- $\Gamma(k, \theta)$ - gamma distribution,
- $\chi^2(n)$ - chi-squared distribution,
- $\text{Exp}(\lambda)$ - exponential distribution,
- $\text{Bin}(n, p)$ - binomial distribution,
- $\mathcal{P}(\lambda)$ - Poisson distribution,
- $\text{geom}(p)$ - geometric distribution,
- Cauchy - Cauchy distribution.

1.1 How to find a model? What distribution can we use?

Process $\longrightarrow$ Values $\longrightarrow$ Assumptions $\longrightarrow$ Distribution

Example: Number of heads while tossing a coin 20 times.

Assumptions:

- fixed number of tosses: $n = 20$,
- we obtain 20 values: heads, tails, tails, ...
- if coin is fair, then probability in one toss $p = \frac{1}{2}$, so $\mathbb{P}(X = \text{heads}) = \frac{1}{2}$,
- independent tosses.

Therefore:

$$X \sim \text{Binomial}\left(20, \frac{1}{2}\right).$$

Choose from the process and assumptions — not only the histogram.

1.2 Discrete or continuous?

1.2.1 Discrete variable

Takes separate, countable values.

Examples:

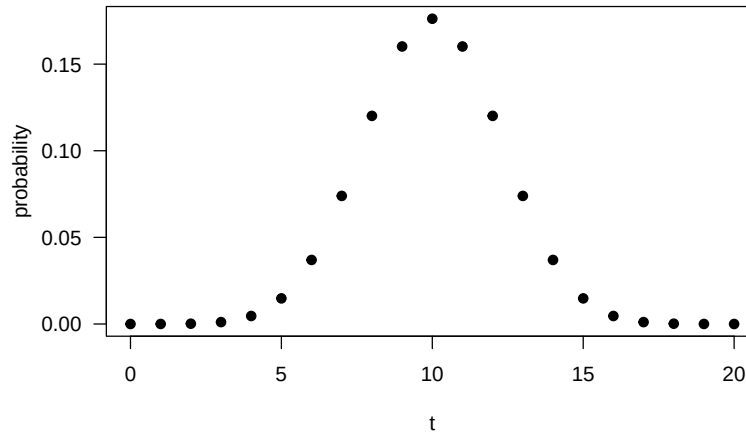
- number of events,
- number of successes,
- number of trials until success.

Probability mass function:

$$p_k = P(X = k)$$

A PMF value **is** a probability.

```
t = seq(0, 20, 1)
plot(t, dbinom(t, 20, 0.5), pch = 19, ylab="probability", las = 1)
```



1.2.2 Continuous variable

Takes values on an interval.

Examples:

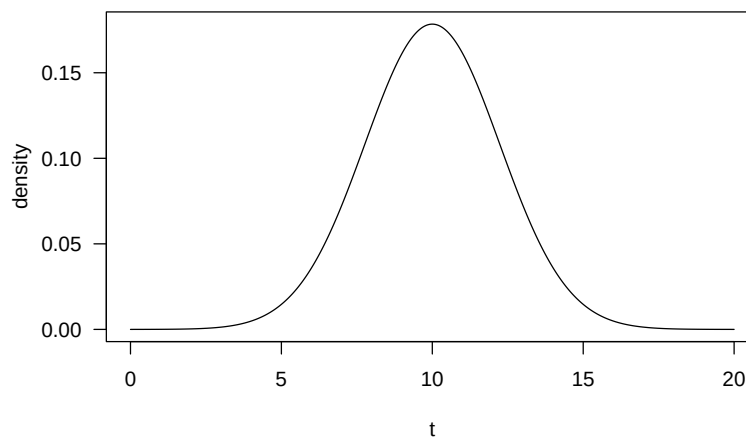
- waiting time,
- measurement error,
- lifetime or concentration

Probability density function:

$$f_X(x) = \text{pdf}(x).$$

A PDF value is a **density**, not $P(X = x)$.

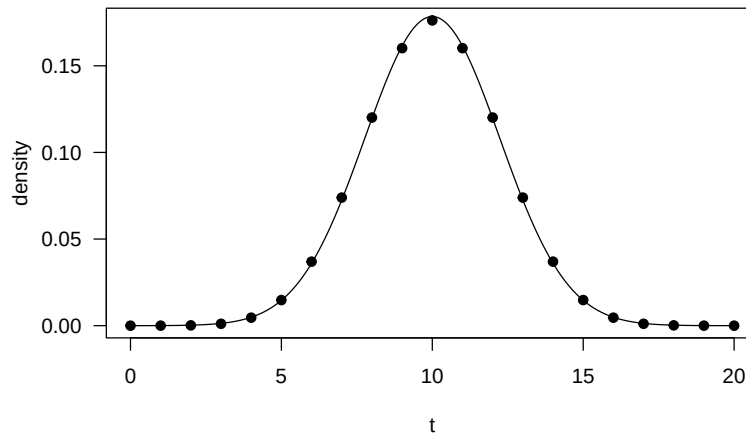
```
t = seq(0, 20, 0.01)
plot(t, dnorm(t, 10, sqrt(5)), type = "l", ylab="density", las = 1)
```



For a continuous variable, probabilities are areas under the density curve.

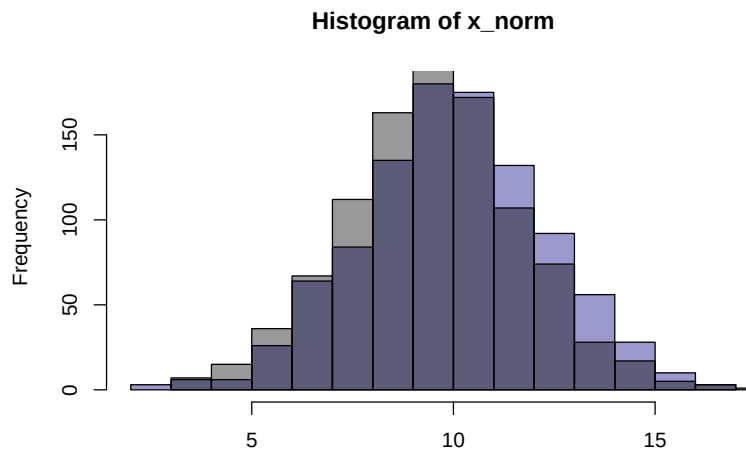
1.3 Compare normal PDF and binomial PMF:

```
t = seq(0, 20, 1)
plot(t, dbinom(t, 20, 0.5), pch = 19, ylab="density", las = 1)
t = seq(0, 20, 0.01)
lines(t, dnorm(t, 10, sqrt(5)))
```



1.4 Compare normal and binomial histograms:

```
x_bin = rbinom(1000, 20, 0.5)
x_norm = rnorm(1000, 10, sqrt(5))
hist(x_norm, col = rgb(0, 0, 0.5, 0.4), xlab = "", breaks = 20)
hist(x_bin, col = rgb(0, 0, 0, 0.4), add = TRUE, breaks=20)
```



1.5 One consistent scipy.stats interface

```
from scipy import stats

rv = stats.norm(loc=0, scale=1)  # continuous distribution

rv.rvs(size=100)                # random sample
rv.pdf(x)                       # density function: f_X(x)
```

```
rv.cdf(x)          #  $P(X \leq x)$ 
rv.ppf(p)          # quantile (percentile)
rv.mean(), rv.var(), rv.std() # some statistics
```

For a discrete distribution:

```
rv = stats.binom(n=20, p=0.25)
rv.pmf(k)          # probability mass function:  $P(X = k)$ 
rv.cdf(k)          #  $P(X \leq k)$ 
rv.ppf(p)          # quantile (percentile)
rv.mean(), rv.var(), rv.std() # some statistics
```

2 Exercises

2.1 One consistent `scipy.stats` interface

```
from scipy import stats

rv = stats.norm(loc=0, scale=1) # continuous distribution

rv.rvs(size=100)               # random sample
rv.pdf(x)                      # density function:  $f_X(x)$ 
rv.cdf(x)                      #  $P(X \leq x)$ 
rv.ppf(p)                      # quantile (percentile)
rv.mean(), rv.var(), rv.std()  # some statistics
```

For a discrete distribution:

```
rv = stats.binom(n=20, p=0.25)
rv.pmf(k)          # probability mass function:  $P(X = k)$ 
rv.cdf(k)          #  $P(X \leq k)$ 
rv.ppf(p)          # quantile (percentile)
rv.mean(), rv.var(), rv.std() # some statistics
```

Remember: `rvs` simulates; `pmf/pdf` describe mass or density; `cdf` accumulates probability; `ppf` inverts the CDF.

Always check the documentation for parameters.

2.2 Exercise

See how specific distributions are represented in `scipy.stats`. How can you find the density ($f_X(t)$) or the probability mass function ($\mathbb{P}(X = k)$) for specific distribution?

Plot (on one plot) densities of the given probability distributions:

- standard normal distribution and standard Cauchy distribution. Can you see a difference. Is it a ‘big’ difference? Try different parameters for normal distribution, can you pick such ones, so that those densities look alike?
- gamma distribution with shape parameter equal to 30 and scale parameter equal to 0.5 along with normal distribution which resembles the given gamma distribution (what parameters would you choose and why?)

2.3 Exercise (GS):

Let us assume that a student came totally unprepared for an exam. The exam is in a form of a quiz that consists of 20 questions, and each question is given 4 answers from which exactly one is correct.

- What is the probability of obtaining 0/20, 5/20, 10/20, 15/20, 20/20 points?
- What distribution can be used to model this situation?

Plot the probability mass function for this distribution with appropriate parameters.

2.4 Excercise (KO):

Which probability distributions can appropriately model the following processes:

- (a) The number of lorries driving through a state road section between 3 and 4 a.m.?
- (b) The number of cars driving out of Warsaw by A2 motorway between 3 and 4 p.m.?
- (c) The time of waiting for the first speeding vehicle after speed measurements begin on a given road section?
- (d) The time of waiting for the first 3 speeding vehicles?
- (e) The time of waiting for the first 100 speeding vehicles?
- (f) The number of speeding vehicles in a quarter?
- (g) The number of speeding vehicles in a day?
- (h) The number of heads obtained during 5 flips of a loaded coin?