Statistics

L2: Random Variables - revisited

Katarzyna Filipiak

Institute of Mathematics
Poznań University of Technology

2025/2026

Random variable

Random variable X – real values function $X = X(\omega)$ defined on the set Ω, that is $X : \Omega \to \mathbb{R}$

Discrete random variable X – if its possible values constitute a sequence of separated points on the number line, i.e., $x_1, x_2, \ldots$

Continuous random variable – if the set of its all possible values is an interval

Since the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to its possible values

Discrete random variables

Probability distribution

Probability distribution of the discrete random variable X (with n possible values labeled by $x_1, \ldots, x_n$) is the collection of probabilities

$$P(X = x_i) = p_i, \qquad i = 1, 2, \ldots, n.$$

Example 1

In the experiment three automatic cameras have been used to save its performance. In a given conditions the probability of taking correct photo equal 0.6 and it is the same for each camera. Give the probability distribution of random variable counting the cameras that took correct photo. Compute probability that:
(a) none of the cameras takes correct photo;
(b) at least two cameras take correct photo.

X – random variable counting the cameras that took correct photo

probability distribution:

x_i					
p_i					

$$P(X = 0) =$$
$$P(X \geq 2) =$$

Binomial distribution $\mathrm{bin}(n,\ p)$

- for a single "object" there are only two possibilities: success or failure – at probability tree there are just two branches at each level
- probability of success is the same for every "object" – at probability tree we have the same pair of probabilities at each level
- the set of values of random variable X that counts successes is $\{0, 1, 2, \dots, n\}$

$X \sim \mathrm{bin}(n, p)$ $\qquad$ $\square\texttt{binom}(\square, n, p)$

Probability distribution (density): $\qquad$ $\texttt{dbinom}(x, n, p)$

Example 1

In the experiment three automatic cameras have been used to save its performance. In a given conditions the probability of taking correct photo equal 0.6 and it is the same for each camera. Give the probability distribution of X denoting the number of cameras that takes correct photo in this experiment.

$$X \sim \mathrm{bin}(3,\ 0.6)$$

$$x = 0 : 3$$

$$\mathsf{prob} = \mathtt{dbinom}(x, 3, 0.6)$$

Expectation, variance and standard deviation

Expectation (mean) for discrete random variable:

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_i$$

Variance: $\mathbb{D}^2(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$

Standard deviation: $\mathbb{D}(X) = \sqrt{\mathbb{D}^2(X)}$

Example 1 - cont.

In the experiment three automatic cameras have been used to save its performance. In a given conditions the probability of taking correct photo equal 0.6 and it is the same for each camera. How many correct photos can be expected? What is the standard deviation?

$$X \sim \mathrm{bin}(3,\ 0.6)$$

$$x = 0:3, \qquad \mathrm{prob} = \mathtt{dbinom}(x, 3, 0.6)$$

$$\begin{aligned}
\mathrm{expect} &= \mathtt{sum}(x * \mathrm{prob}) \\
\mathrm{variance} &= \mathtt{sum}(x * x * \mathrm{prob}) - \mathrm{expect} * \mathrm{expect} \\
\mathrm{sd} &= \mathtt{sqrt}(\mathrm{variance})
\end{aligned}$$

Discrete distributions in R

name = name of the distribution

param = parameters of the distribution

Probability/density:	d (density)	+	*name*	=	d*name*(x, *param*)
CDF:	p (probability)	+	*name*	=	p*name*(x, *param*)
Quantile:	q (quantile)	+	*name*	=	q*name*(x, *param*)
Random number:	r (random)	+	*name*	=	r*name*(x, *param*)

Discrete distribution name:

- binomial: `binom`
- Poisson: `pois`

Discrete distribution histogram (line graph):

plot(x, d*name*(x, *param*), type = "h")

Continuous random variables

Probability density function

Probability density function (pdf) of the continuous random variable X, that can be used to obtain probabilities, is the following function

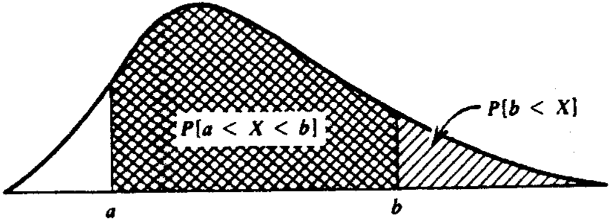
a) $f : \mathbb{R} \longrightarrow \mathbb{R}^{+} \cup \{0\}$,

b) $P(a \leq X \leq b) =$ the area under the curve $f(x)$ between a and b.

The properties:

- $f(x) \geq 0$
- $P(a \leq X \leq b) = P(a < X < b) = \int_a^b f(x)\,\mathrm{d}x$
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
- $P(X > b) = 1 - P(X \leq b)$

Density function



$$
\begin{aligned}
P(a < X \leq b) &= P(a < X < b) \\
&= P(a \leq X \leq b) \\
&= P(a \leq X < b)
\end{aligned}
$$

$$P(X = a) = 0$$

Example 2

A radar telemetry tracking station requires a vast quantity of high-quality magnetic tape. It has been established that the distance (X) (in cm) between tape-surface flaws has the following pdf:

$$f(x) = \begin{cases} 0.01\mathrm{e}^{-0.01x} & \text{for} \quad x \geq 0, \\ 0 & \text{for} \quad x < 0. \end{cases}$$

Suppose one flaw has been identified. What is the probability that an additional flaw is found within the next 50 cm of tape?

X - random variable counting the distance between flaws

$$P(X \leq 50) = \int_0^{50} 0.01\mathrm{e}^{-0.01x}\,\mathrm{d}x = \mathtt{integrate}(f, 0, 50)$$

$$f = \mathtt{function}(x)\{0.01 * \mathtt{exp}(-0.01 * x)\}$$

$$= 0.3934693$$

Cumulative distribution function

Cumulative distribution function (cdf) $F : \mathbb{R} \to [0,\ 1]$ is defined as

$$F(x) = P(X \leq x)$$

$$\texttt{pname}(x, \text{parameters})$$

Cdf of continuous random variable:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

Properties of cdf

- $0 \leq F(x) \leq 1$
- $F(x)$ is nondecreasing
- $F(x)$ right-hand side continuous, that is:

$$\lim_{x \to x_0^+} F(x) = F(x_0)$$

For arbitrary $a, b \in \mathbb{R}$:

$$P(X \leq a) = F(a)$$

$$P(X > a) = 1 - F(a)$$

$$P(a < X \leq b) = F(b) - F(a)$$

Example 2 - cont.

A radar telemetry tracking station requires a vast quantity of high-quality magnetic tape. It has been established that the distance (X) (in cm) between tape-surface flaws has the following cdf:

$$F(x) = \begin{cases} 0 & \text{for} \quad x < 0, \\ 1 - \mathrm{e}^{-0.01x} & \text{for} \quad x \geq 0, \end{cases}$$

Suppose one flaw has been identified. Using cdf find the probability that an additional flaw is found within the next 50 cm of tape.

$$P(X \leq 50) = F(50) = 1 - \mathrm{e}^{-0.01 \cdot 50} = 0.3934693$$

$$X \sim \mathrm{EXP}(\lambda), \qquad \lambda = 0.01$$

$$P(X \leq 50) = F(50) = \mathtt{pexp}(50,\ 0.01)$$

Expectation, variance and standard deviation

Expectation (mean) for continuous random variable:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

Variance: $\mathbb{D}^2(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$

Standard deviation: $\mathbb{D}(X) = \sqrt{\mathbb{D}^2(X)}$

Example 2 - cont.

A radar telemetry tracking station requires a vast quantity of high-quality magnetic tape. It has been established that the distance (X) (in cm) between tape-surface flaws has the following pdf:

$$f(x) = \begin{cases} 0.01\mathrm{e}^{-0.01x} & \text{for} \quad x \geq 0, \\ 0 & \text{for} \quad x < 0. \end{cases}$$

What is the expected distance between flaws?

$$\mathbb{E}(X) = \int_0^\infty x \cdot 0.01\mathrm{e}^{-0.01x}\,\mathrm{d}x = \texttt{integrate}(f, 0, \texttt{Inf})$$

$$f = \texttt{function}(x)\{x * 0.01 * \texttt{exp}(-0.01 * x)\}$$

$$= 100$$

$$X \sim \mathrm{EXP}(\lambda) \quad \Longrightarrow \quad \mathbb{E}(X) = \frac{1}{\lambda} = \frac{1}{0.01} = 100$$

Normal distribution $N(\mu,\sigma)$

Pdf ($\mu \in \mathbb{R}, \quad \sigma \in \mathbb{R}^+$):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \texttt{dnorm}(x,\mu,\sigma)$$

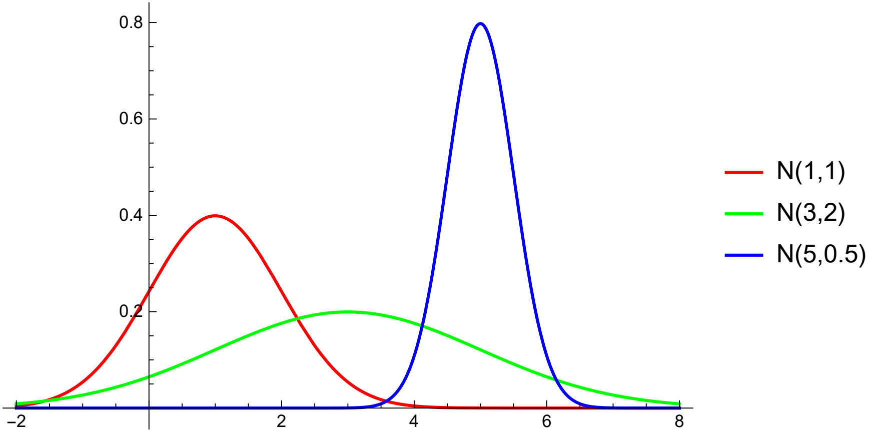
Cdf:

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \qquad \texttt{pnorm}(x,\mu,\sigma)$$

Expectation and variance:

$$\mathbb{E}(X) = \mu, \qquad \mathbb{D}^2(X) = \sigma^2$$

Normal distribution



Standard normal distribution $N(0,1)$

$$X \sim N(\mu,\sigma) \quad \Rightarrow \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

$\Phi(z)$ – the cdf of $N(0,1)$ $\diamondsuit$

$$P(X \leq b) = P\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right)$$

$$P(X \leq b) = F(b) = \texttt{pnorm}(b,\mu,\sigma)$$

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$P(a \leq X \leq b) = F(b) - F(a) = \texttt{pnorm}(b,\mu,\sigma) - \texttt{pnorm}(a,\mu,\sigma)$$

Example 3

The average active-ingredient yield per liter of raw material for samples of vials may be approximated by a normal distribution with mean $\mu = 30$ grams and standard deviation $\sigma = 0.2$ gram. Find the probability that the average yield of a sample is

(a) less than 29.55 grams;

(b) greater than 30.45 grams;

(c) between 29.5 and 30.25 grams.

$$X \sim N(30,\ 0.2)$$

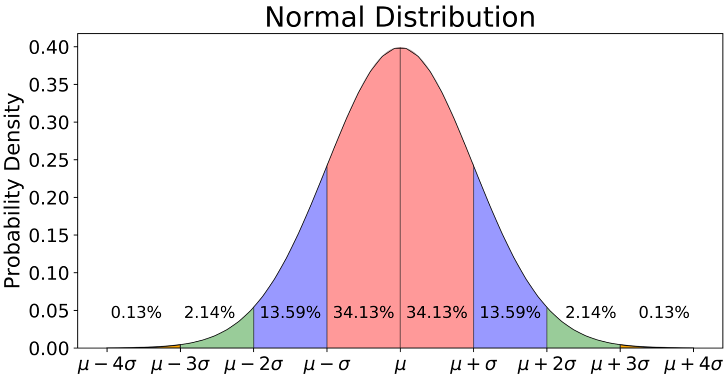
$$P(X < 29.55) = F(29.55) = \texttt{pnorm(29.55, 30, 0.2)}$$

$$P(X > 30.45) = 1 - F(30.45)$$

$$=$$

$$P(29.5 < X < 30.25) =$$

3-sigma rule


Normal Distribution
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
$\mu-4\sigma$
$\mu-3\sigma$
$\mu-2\sigma$
$\mu-\sigma$
μ
$\mu+\sigma$
$\mu+2\sigma$
$\mu+3\sigma$
$\mu+4\sigma$

3-sigma rule

Let $X \sim N(\mu, \sigma)$. Then

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

Approximation of binomial distribution

Let $X \sim \mathrm{bin}(n, p)$ with large n and $p \in (0,\ 1)$.
It is known that

$$\mathbb{E}(X) = n \cdot p, \qquad \mathbb{D}(X) = \sqrt{n \cdot p \cdot q}.$$

Then

$$X \underset{\text{app}}{\sim} N(n \cdot p,\ \sqrt{n \cdot p \cdot q})$$

Example 3

The *U.S. Statistical Abstract* reports that the median family outcome in the U.S. for 1981 was 22400\$. Among 180 randomly selected families find the probability, that at most 70 has an income below 22400\$.

B - random variable counting the number of families with the income below 22400\$

$$B \sim \text{bin}(180,\, 0.5)$$

$$B \underset{\text{app}}{\sim} N(180 \cdot 0.5,\, \sqrt{180 \cdot 0.5 \cdot 0.5})$$

$$P(B \leq 70) = F_{\text{bin}}(70) = \texttt{pbinom}(70, 180, 0.5) = 0.001766602$$

$$P(B \leq 70) \approx F_N(70) = \texttt{pnorm}(70,\ 90,\ \sqrt{45}) = 0.001434556$$

Continuous distributions in R

name = name of the distribution

param = parameters of the distribution

Probability/density:	d (density)	+	*name*	=	d*name*(x, *param*)
CDF:	p (probability)	+	*name*	=	p*name*(x, *param*)
Quantile:	q (quantile)	+	*name*	=	q*name*(x, *param*)
Random number:	r (random)	+	*name*	=	r*name*(x, *param*)

Continuous distribution name:

exponential:	`exp`	t-Student:	`t`
normal:	`norm`	chi-square:	`chisq`
		F-Snedecor:	`f`

Continuous distribution function plot: curve(d*name*(x, *param*))

Short summary

Discrete distributions:

- binomial bin(n, p):
  - probability: `dbinom(`*point*`,`*n*`,`*p*`)`
  - graph: `plot(x, dbinom(x, `*n*`, `*p*`), type = "h")`
  - $\mathbb{E}(X) = n \cdot p$, $\mathbb{D}(X) = \sqrt{n \cdot p \cdot (1 - p)}$

Continuous distributions:

- exponential EXP(λ):
  - probability (red formula): `pexp(`*point*`,`λ`)`
  - graph: `curve(dexp(x, `λ`))`
  - $\mathbb{E}(X) = 1/\lambda$, $\mathbb{D}(X) = 1/\lambda$
- normal N(μ, σ):
  - probability (red formula): `pnorm(`*point*`,`μ`,`σ`)`
  - graph: `curve(dnorm(x, `μ`, `σ`))`
  - $\mathbb{E}(X) = \mu$, $\mathbb{D}(X) = \sigma$