Applied Statistics

for Computer Science BSc, Exam

Probability Theory and Mathematical Statistics

for Computer Science Engineering BSc, Term grade

István Fazekas
University of Debrecen

2020/21 fall

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

Main topics

1. Probability theory
2. Statistics

Mathematical tools: combinatorics, calculus

Computer tool: Matlab

Book:

Yates, Goodman:

Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers

Lecture 8

The normal distribution

The standard normal distribution

Definition. The standard normal probability density function is

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}. \tag{1}$$

Notation: $X \sim \mathcal{N}(0,1)$.
The standard normal distribution is the most important distribution in statistics.

The standard normal PDF

f is symmetric.
Proposition. The function f in (1) is a PDF.
Proof. $f(x) > 0$
f is measurable because it is continuous.

$$\int_{-\infty}^{\infty} f(y)dy = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} e^{-y^2/2} dy = \sqrt{\frac{2}{\pi}} I. \quad (2)$$

The standard normal PDF...

Calculate I^2. Using double integral

$$I^2 = \int_0^\infty e^{-x^2/2}dx \int_0^\infty e^{-y^2/2}dx = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)/2}dxdy.$$

Applying substitution $x = r\cos\varphi$, $y = r\sin\varphi$ that is using polar coordinates, we see that the Jacobi determinant of the transformation is r. Then

$$I^2 = \int_0^{\pi/2} \int_0^\infty re^{-r^2/2}drd\varphi = \frac{\pi}{2}\left[-e^{-r^2/2}\right]_0^\infty = \frac{\pi}{2}.$$

By (2), $\int_{-\infty}^{\infty} f(x)dx = 1$, so f is a PDF.

The standard normal PDF...

The standard normal PDF is a symmetric bell-shaped curve.

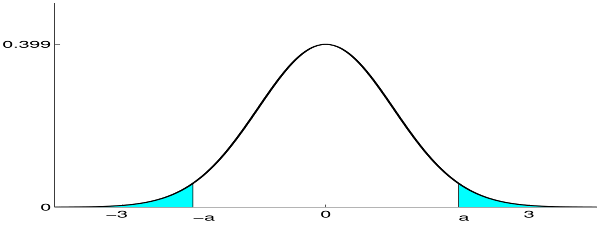

Figure: The PDF of the standard normal distribution

The 0.025 quantile of the standard normal distribution is $-a = -1.96$ and the 0.975 quantile is $a = 1.96$.
It means that both blue domains below the curve of the standard normal PDF have area 0.025.

The standard normal CDF

The standard normal CDF is a strictly increasing continuous curve.

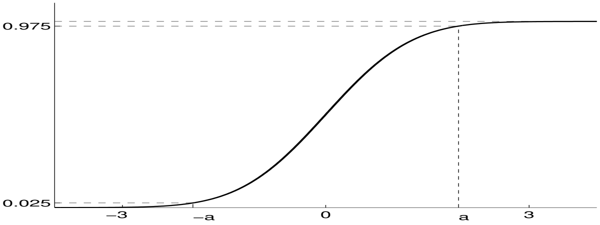

Figure: The CDF of the standard normal distribution

On the figure we see that the 0.025 quantile of the standard normal distribution is $-a = -1.96$ and the 0.975 quantile is $a = 1.96$.
The values of the standard normal CDF are contained in tables and they are calculated by statistical programs.

The moments of the standard normal distribution

$\mathbb{E}X^k = 0$, if k is odd,
$\mathbb{E}X^k = (k-1)!!$, if k is even.

Here $(k-1)!! = (k-1)(k-3)\cdots 1$, that is $(k-1)$ semi-factorial.

In particular

$$\mathbb{E}X = 0,\ \mathbb{E}X^2 = 1,\ \mathbb{E}X^3 = 0,\ \mathbb{E}X^4 = 3$$

So

$$\mathrm{Var}X = \mathbb{E}X^2 - \mathbb{E}^2X = 1 - 0 = 1$$

The moments of the standard normal distribution...

Proof.

$$\mathbb{E}X^k = \int_{-\infty}^{\infty} x^k \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

So it is 0 for odd k.
Now for even k, using partial integration,

$$\mathbb{E}X^k = \int_{-\infty}^{\infty} x^{k-1} \frac{1}{\sqrt{2\pi}} x\, e^{-x^2/2} dx = -\left[x^{k-1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right]_{-\infty}^{+\infty} +$$

$$+ \int_{-\infty}^{\infty} (k-1) x^{k-2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = (k-1)\mathbb{E}X^{k-2}.$$

As $\mathbb{E}X^0 = 1$, therefore, by mathematical induction,
$\mathbb{E}X^k = (k-1)!!$

The normal distribution

Definition. The distribution of Y is called normal distribution if its PDF is

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)}, \tag{3}$$

where $m \in \mathbb{R}$, $\sigma > 0$.
Notation $Y \sim \mathcal{N}(m, \sigma^2)$.

The normal distribution...

Proposition. Let X have standard normal distribution, $\sigma \neq 0$, then $Y = \sigma X + m$ has normal distribution.
Proof. Let $X \sim \mathcal{N}(0,1)$.
Then for $Y = \sigma X + m$ (if $\sigma > 0$)

$$F_Y(x) = P(Y < x) = P(\sigma X + m < x) = P\left(X < \frac{x-m}{\sigma}\right) = F_X\left(\frac{x-m}{\sigma}\right)$$

So the PDF of Y is

$$f_Y(x) = f_X\left(\frac{x-m}{\sigma}\right) \cdot \frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2}.$$

The normal distribution...

Exercise. Let $Y \sim \mathcal{N}(m, \sigma^2)$.
Show that $\mathbb{E}Y = m$ and $\mathrm{Var}(Y) = \sigma^2$.
Hint. Let $X \sim \mathcal{N}(0,1)$. Then for $Y = \sigma X + m$ we have $Y \sim \mathcal{N}(m, \sigma^2)$.

Remark. If Y is normally distributed, then $aY + b$ is also normally distributed for any numbers $a \neq 0$ and b.
Remark. If $Y \sim \mathcal{N}(m, \sigma^2)$, then $X = (Y - m)/\sigma$ has standard normal distribution.
Remark. The PDF of $\mathcal{N}(m, \sigma^2)$ is a bell-shaped curve which is symmetric around m.

The normal PDF

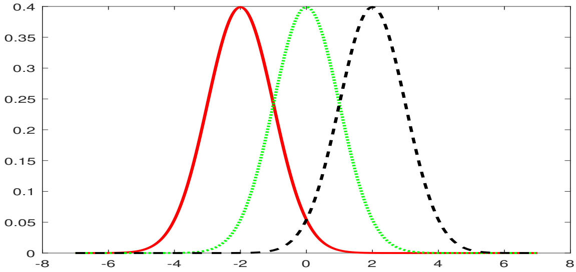

Figure: PDF of normal distributions

solid line: PDF of $\mathcal{N}(-2, 1)$
dotted line: PDF of $\mathcal{N}(0, 1)$
dashed line: PDF of $\mathcal{N}(2, 1)$

The normal PDF

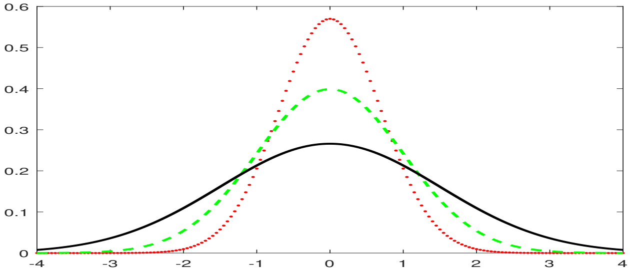

Figure: PDF of normal distributions

solid line: PDF of $\mathcal{N}(0, 1.5^2)$
dashed line: PDF of $\mathcal{N}(0, 1^2)$
dotted line: PDF of $\mathcal{N}(0, 0.7^2)$

Normal PDF and histogram

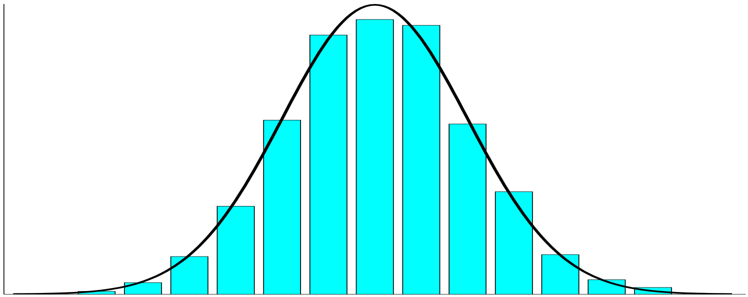

Figure: A histogram and the fitted normal PDF

Approximation of the standard normal CDF

There is no closed formula for the standard normal CDF.
But there are good approximations.
Let Φ be the standard normal CDF. Then

$$\Phi(x) \approx 1 - 0.5(1 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4)^{-4},$$

if $x \geq 0$, where $a_1 = 0.196854$, $a_2 = 0.115194$, $a_3 = 0.000344$, $a_4 = 0.019527$.
The error of the approximation is less than 2.5×10^{-4}.

The three-sigma rule

If $X \sim \mathcal{N}(m, \sigma^2)$, then

$$P(m - 3\sigma < X < m + 3\sigma) \approx 0.9972$$

A normal random variable falls outside the $\pm 3\sigma$ interval around the expectation with negligible (around 0.0028) probability.
This $\boldsymbol{3\sigma}$-rule is important in quality control.