Introduction to Probability

Lecture 5+: Continuous random variables

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Outline

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

Normal (Gaussian) random variable

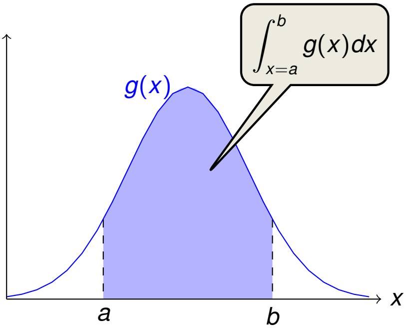
From discrete to continuous RV

- So far, all RV were discrete: can only take on integer values.
- If RV need to take on values in the real number domain ($\mathbb{R}$), then continuous random variable.
- Examples of continuous RV: Uniform RV, Exponential RV, Normal RV.
- Continuous RV are just like discrete RV, except that every **sum** becomes an **integral**.
- Example of possible values of continuous RV X:

$$\begin{aligned}
(0,1) &= \{x \in \mathbb{R}; 0 < x < 1\} \\
[0,1] &= \{x \in \mathbb{R}; 0 \leq x \leq 1\} \\
[0,1) &= \{x \in \mathbb{R}; 0 \leq x < 1\} \\
(-\infty,\infty) &= \text{all real numbers}
\end{aligned}$$

- Examples:
  - X: price of a stock
  - X: time that a machine works before breakdown
  - X: error in an experimental measurement

Integrals revision



Integral = area under a curve = $\int_{x=a}^{b} g(x)dx = G(x)\Big|_a^b = G(b) - G(a)$

where $G(x)$ is the antiderivative for g(x).

Some examples:

$$\int_a^b x^2 dx = \frac{x^3}{3}\Big|_a^b = \frac{b^3 - a^3}{3} \qquad \int a\,dx = ax + C$$

$$\int \frac{1}{x} dx = ln|x| + C \qquad \int e^x dx = e^x + C$$

Continuous paradigm

- The most important property of discrete RV was probability mass function (PMF) denoting the probability of the RV taking on a certain value.
- But in the continuous world this is impossible:
  What is the probability that a newborn child weighs exactly 3.215438765432532 kg? **NONE**
- Real values are defined with infinite precision, thus the probability that a RV takes on a specific value is not meaningful when the RV is continuous.
- We need a function that says how likely is it that a RV takes on a particular value relative to other values that it could take on: probability density function.

Definition of continuous RV

Continuous random variable

A random variable X is continuous if there is a probability density function (PDF), $f(x) \geq 0$ such that for $-\infty < x < \infty$:

$$\mathbf{P}[a \leq X \leq b] = \int_a^b f(x)dx$$

To preserve the axioms that guarantee that $\mathbf{P}[a \leq X \leq b]$ is a probability, the following properties must hold:

$$0 \leq \mathbf{P}[a \leq X \leq b] \leq 1$$

$$\mathbf{P}[-\infty < X < \infty] = 1 \qquad \left(= \int_{-\infty}^{\infty} f(x)dx\right)$$

- Note: we also write $f(x)$ as $f_X(x)$.
- In continuous world, every RV has a PDF: its relative value wrt to other possible values.
- Integrate $f(x)$ to get probabilities.

Comparing PMF and PDF

Discrete random variable X	Continuous random variable X
Probability mass function (PMF):	Probability density function (PDF):
$p(x)$	$f(x)$
Compute probability:	Compute probability:
$\mathbf{P}[X = x] = p(x)$	
$\mathbf{P}[a \leq X \leq b] = \sum_{x=a}^{b} p(x)$	$\mathbf{P}[a \leq X \leq b] = \int_{x=a}^{b} f(x)dx$

Both are measures of how likely is X to take on a value.

Computing probability example

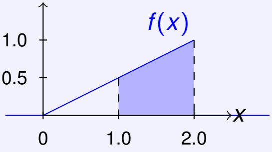
Example

Let X be a continuous RV with PDF:

$$f(x) = \begin{cases} \frac{1}{2}x & \text{if } 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

What is $\mathbf{P}[X \ge 1]$?

Answer



Method 1: integrate

$$\mathbf{P}[1 \le X < \infty] = \int_1^\infty f(x)dx = \int_1^2 \frac{1}{2}xdx$$

$$= \frac{1}{2}\left(\frac{1}{2}x^2\right)\Big|_1^2 = \frac{1}{2}\left(2 - \frac{1}{2}\right) = \frac{3}{4}$$

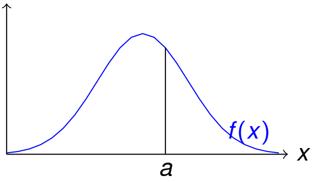
Method 2: areas of triangles

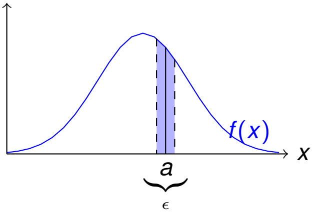
area$=\frac{(2)(1)}{2} - \frac{(1)(0.5)}{2} = 1 - \frac{1}{4} = \frac{3}{4}$

PDF properties

- $f(x)$ is NOT a probability, it is probability density:

$$\mathbf{P}[X = a] = \int_a^a f(x)dx = 0 \neq f(a)$$





$$\mathbf{P}\left[a - \frac{\epsilon}{2} \leq X \leq a + \frac{\epsilon}{2}\right] = \int_{a-\frac{\epsilon}{2}}^{a+\frac{\epsilon}{2}} f(x)dx \approx \text{width } \times \text{ height} = \epsilon f(a)$$

Thus, $\mathbf{P}[X = a] = \lim_{\epsilon \to 0} \epsilon f(a) = 0$.

- $\mathbf{P}[a \leq X \leq b] = \mathbf{P}[a < X \leq b] = \mathbf{P}[a \leq X < b] = \mathbf{P}[a < X < b]$

PDF and probability example

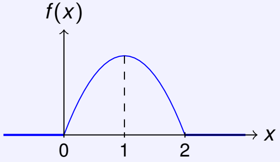
Example

Let X be a continuous RV with PDF:

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{when } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the value of the constant C? What is $\mathbf{P}[X > 1]$?

Answer



C is a normalisation constant. We know that PDF must sum to 1:

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} f(x)dx + \int_{0}^{2} f(x)dx + \int_{2}^{\infty} f(x)dx$$

$$= \int_{-\infty}^{0} 0dx + \int_{0}^{2} C(4x - 2x^2)dx + \int_{2}^{\infty} 0dx$$

$$= \int_{0}^{2} C(4x - 2x^2)dx = C\left(2x^2 - \frac{2x^3}{3}\right)\Bigg|_0^2 = C\left(8 - \frac{16}{3}\right) = C\frac{8}{3}$$

Thus $C = \frac{3}{8}$

PDF and probability example cont.

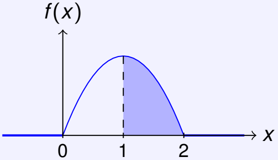
Example

Let X be a continuous RV with PDF:

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{when } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the value of the constant C? What is $\mathbf{P}[X > 1]$?

Answer



$$\mathbf{P}[X > 1] = \int_1^\infty f(x)dx = \int_1^2 f(x)dx + \int_2^\infty 0dx$$

$$= \int_1^2 \frac{3}{8}(4x - 2x^2)dx = \frac{3}{8}\left(2x^2 - \frac{2x^3}{3}\right)\Big|_1^2 =$$

$$= \frac{3}{8}\left(\left(8 - \frac{16}{3}\right) - \left(2 - \frac{2}{3}\right)\right) = \frac{1}{2}$$

Outline

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

Normal (Gaussian) random variable

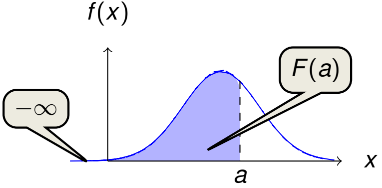
Cumulative distribution function

- Since PDF is not a probability, we need to solve an integral every single time we want to calculate a probability.
- To save effort, cumulative distribution function (CDF) computes this: $F(a) = F_X(a) = \mathbf{P}[X \leq a]$ where $-\infty < a < \infty$.
- Recall: CDF for *discrete* RV is $F(a) = \sum_{\text{all } x \leq a} p(x)$

Cumulative distribution function for a continuous RV

For a continuous random variable X with PDF $f(x)$, the cumulative distribution function (CDF) is:

$$F_X(a) = \mathbf{P}[X \leq a] = \int_{-\infty}^{a} f(x)dx$$

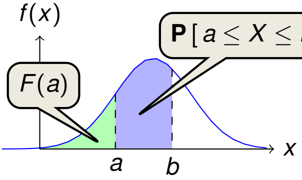


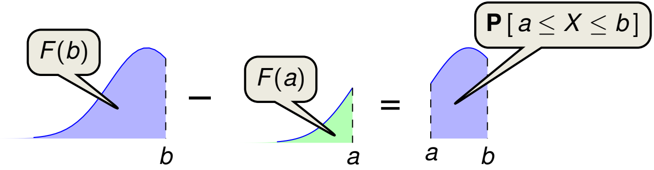
CDF properties

- While PDF is not a probability, CDF is.
- If you learn to use CDFs, you can avoid integrating the PDF.
- It is a matter of convention that CDF is probability that a RV takes on a value less than (or equal to) the input value as opposed to greater than.
- Useful examples of using CDF:

Probability question	**Solution**	**Explanation**
$\mathbf{P}[X \leq a]$	$F(a)$	Definition of CDF
$\mathbf{P}[X < a]$	$F(a)$	Note that $\mathbf{P}[X = a] = 0$
$\mathbf{P}[X > a]$	$1 - F(a)$	$\mathbf{P}[X \leq a] + \mathbf{P}[X > a] = 1$
$\mathbf{P}[a < X < b]$	$F(b) - F(a)$	$F(a) + \mathbf{P}[a < X < b] = F(b)$

Computing CDF





$$
\begin{aligned}
F(b) - F(a) &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx \\
&= \left(\int_{-\infty}^{a} f(x)dx + \int_{a}^{b} f(x)dx \right) - \int_{-\infty}^{a} f(x)dx \\
&= \int_{a}^{b} f(x)dx = \mathbf{P}[a < X < b] = \mathbf{P}[a \le X \le b]
\end{aligned}
$$

Expectation and variance for continuous RV

Discrete RV X

$$\mathbf{E}[X] = \sum_x xp(x)$$

$$\mathbf{E}[g(X)] = \sum_x g(x)p(x)$$

Continuous RV X

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

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

Both continuous and discrete RVs

$$\mathbf{E}[aX + b] = a\mathbf{E}[X] + b \qquad \text{Linearity of expectation}$$

$$\mathbf{V}[X] = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X]^2) \qquad \text{Properties of}$$

$$\mathbf{V}[aX + b] = a^2\mathbf{V}[X] \qquad \text{variance}$$

Outline

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

Normal (Gaussian) random variable

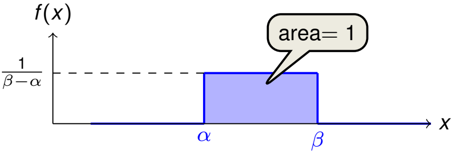
Uniform continuous RV

Uniform continuous random variable

A uniform continuous random variable X is defined as follows:

$$\mathbf{X} \sim \mathbf{Uni}(\alpha, \beta)$$

$$\begin{aligned}
\text{Range:} \quad & [\alpha, \beta], \text{ sometimes } (\alpha, \beta) \\
\text{PDF:} \quad & f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{when } \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases} \\
\text{Expectation:} \quad & \mathbf{E}[X] = \frac{\alpha+\beta}{2} \\
\text{Variance:} \quad & \mathbf{V}[X] = \frac{(\beta-\alpha)^2}{12}
\end{aligned}$$



- Notice that the density $\frac{1}{\beta-\alpha}$ is exactly the same regardless of the value of x. This makes it uniform.
- The PDF is $\frac{1}{\beta-\alpha}$ since it is a constant such that the integral over all possible inputs evaluates to 1.

Public transport example

Example

The University bus arrives at the Computer Lab bus stop at 7:00, 7:15 and so on at 15 minute intervals. You arrive at the bus stop a time uniformly distributed in the interval between 1pm and 1:30pm. What is the probability that you wait less than 5 minutes for the bus?

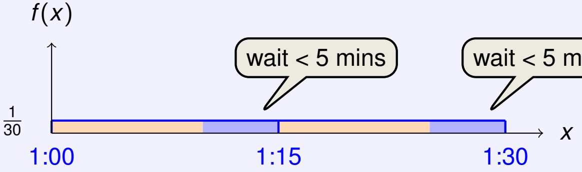
Answer

Let X be a RV for the time you arrive after 1pm to the bus stop.

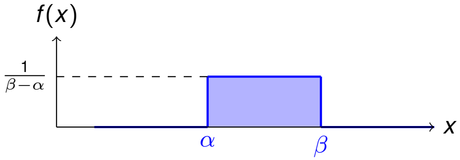
Define RVs: $X \sim Uni(0, 30)$

Solve:

$$\mathbf{P}[10 < X < 15] + \mathbf{P}[25 < X < 30] = \int_{10}^{15} \frac{1}{30} dx + \int_{25}^{30} \frac{1}{30} dx = \frac{5}{30} + \frac{5}{30} = \frac{1}{3}$$



Expectation for Uniform RV



$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_{\alpha}^{\beta} x \cdot \frac{1}{\beta - \alpha}\, dx$$

$$= \frac{1}{\beta - \alpha} \frac{1}{2} x^2 \Big|_{\alpha}^{\beta} = \frac{1}{\beta - \alpha} \frac{1}{2} (\beta^2 - \alpha^2)$$

$$= \frac{1}{2} \frac{(\beta + \alpha)(\beta - \alpha)}{\beta - \alpha} = \frac{\alpha + \beta}{2}$$

Outline

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

Normal (Gaussian) random variable

Exponential continuous RV

Exponential continuous random variable

An exponential random variable X represents the time until an event (first success) occurs. It is parametrised by $\lambda > 0$, the constant rate at which the event occurs.

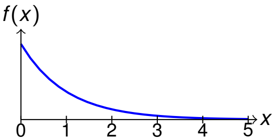
$$\mathbf{X} \sim \mathbf{Exp}(\lambda)$$

Range: $[0, \infty)$

PDF: $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{when } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Expectation: $\mathbf{E}[X] = \dfrac{1}{\lambda}$ (time)

Variance: $\mathbf{V}[X] = \dfrac{1}{\lambda^2}$



- Examples: time until next earthquake, time for request to reach web server, time until end of mobile phone contract.
- Note that λ is the same as the one in the Poisson RV.
- Poisson RV counts # of events that occur in a fixed interval, exponential RV measures the amount of time until the next event occurs.

Pandemic example

Example

Major pandemics occur once every 100 years. What is the probability of a major pandemic in the next 5 years? What is the standard deviation of years until the next pandemic?

Answer

Let X be a RV for the time when the next pandemic happens.
Let a unit of time be 1 year.

Define RVs: $X \sim Exp(\lambda)$, $\mathbf{E}[X] = \frac{1}{\lambda} = 100$, thus $\lambda = \frac{1}{100} = 0.01$
$X \sim Exp(\lambda = 0.01)$.

Solve: Compute $\mathbf{P}[X < 5]$, $\mathbf{SD}[X]$.

$$\mathbf{P}[X < 5] = \int_0^5 0.01 e^{-0.01x} dx \qquad \text{(remember that } \int e^{cx} dx = \frac{1}{c} e^{cx}\text{)}$$

$$= 0.01 \frac{1}{-0.01} e^{-0.01x} \Big|_0^5$$

$$= -(e^{-0.05} - e^0) \approx 0.049$$

$$\mathbf{SD}[X] = \sqrt{\mathbf{V}[X]} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda} = 100 \text{ years}$$

CDF of Exponential RV

CDF for Exponential RV

If X is an exponential continuous random variable, $X \sim Exp(\lambda)$, then its cumulative distribution function CDF (where $x \geq 0$) is

$$F(x) = 1 - e^{-\lambda x}$$

Proof:

$$\begin{aligned} F(x) = \mathbf{P}[X \leq x] &= \int_0^x \lambda e^{-\lambda x} dx \\ &= \lambda \frac{1}{-\lambda} e^{-\lambda x} \Big|_0^x \\ &= -1(e^{-\lambda x} - e^{-\lambda 0}) \\ &= 1 - e^{-\lambda x} \end{aligned}$$

Outline

Continuous random variables

Cumulative distribution function, expectation, variance

Uniform random variable

Exponential random variable

Normal (Gaussian) random variable

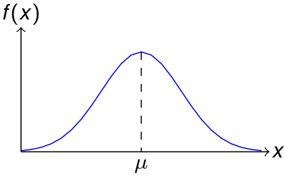
Normal continuous RV

Normal continuous random variable

A normal random variable X, parametrised over mean μ and variance σ^2 is defined as

$$\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$$

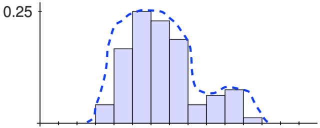
$$\begin{aligned} \text{Range:} \quad & (-\infty, \infty) \\ \text{PDF:} \quad & f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \\ \text{Expectation:} \quad & \mathbf{E}[X] = \mu \\ \text{Variance:} \quad & \mathbf{V}[X] = \sigma^2 \end{aligned}$$



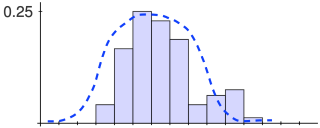
- The most important random variable type, AKA Gaussian RV and Bell curve.
- Generated from summing independent RV, thus occurs often in nature (cf. Central Limit Theorem in Lecture 8).
- Used to model entropic (conservative) distribution of data with mean and variance.

Normal RV paradigm

Goal: translate problem statement into a RV – model real life situation with probability distributions (e.g., height distribution in a class).



Perfect fit!
But what about another class?
Overfit?



Same mean and variance!
Generalises well.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. PDF of X:

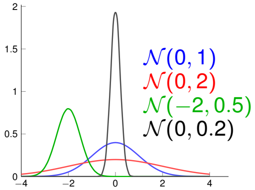
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

exponential tail

symmetric around μ

normalising constant

variance σ^2 manages spread



Walking example

Example

You spent X minutes walking to the department every day. The average time you spend is $\mu = 10$ minutes. The variance from day to day of the time spent to get to the department is $\sigma^2 = 2$ minutes^2. Suppose X is normally distributed. What is the probability you spend ≥ 12 minutes travelling to the department?

Answer

$$X \sim \mathcal{N}(\mu = 10, \sigma^2 = 2)$$

$$\mathbf{P}[X \geq 12] = \int_{12}^{\infty} f(x)dx = \int_{12}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Cannot be solved analytically!

That is, no closed form for the integral of the Normal PDF. (But...)

Properties for Normal RV

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with CDF $\mathbf{P}[X \le x] = F(x)$.

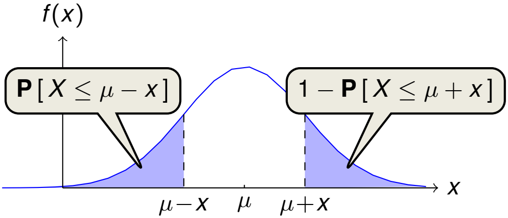
- Linear tranformations of Normal RVs are also Normal RVs.

$$\text{If } Y = aX + b, \text{ then } Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

  Proof outline:
  - $\mathbf{E}[Y] = \mathbf{E}[aX + b] = a\mathbf{E}[X] + b = a\mu + b$ (linearity of expectation)
  - $\mathbf{V}[Y] = \mathbf{V}[aX + b] = a^2\mathbf{V}[X] = a^2\sigma^2$
  - Y is also Normal.

- The PDF of a Normal RV is symmetric about the mean μ.

$$F(\mu - x) = 1 - F(\mu + x)$$



Computing probabilities with Normal RV

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. How do we compute CDF, $\mathbf{P}[X \leq x] = F(x)$?

- We cannot analytically solve the integral (it has no closed form).
- But we can solve numerically using a function Φ, which is a precomputed function:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

CDF of the Standard Normal, Z

Z: Standard Normal RV

Standard Normal random variable Z

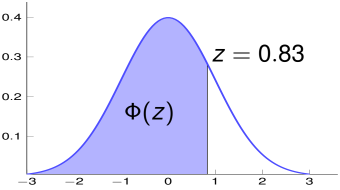
The Standard Normal continuous random variable Z is defined as

$$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

Expectation: $\mathbf{E}[Z] = \mu = 0$ (zero mean)

Variance: $\mathbf{V}[Z] = \sigma^2 = 1$ (unit variance)

- Not a new distribution: a special case of the Normal ($\mathcal{N}(\mu, \sigma^2) = \mu + \sigma \mathcal{N}(0, 1)$).
- CDF of Z defined as $\mathbf{P}[Z \leq z] = \Phi(z)$.



$\mathbf{P}[Z \leq 0.83] = \Phi(0.83) = 0.7967$

Table A.3 Standard Normal Curve Areas (*cont.*) $\Phi(z) = P(Z \leq z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633

Walking example revisited

Example

You spent X minutes walking to the department every day. The average time you spend is $\mu = 10$ minutes. The variance from day to day of the time spent to get to the department is $\sigma^2 = 2$ minutes^2. Suppose X is normally distributed. What is the probability you spend ≥ 12 minutes travelling to the department?

Answer

$X \sim \mathcal{N}(\mu = 10, \sigma^2 = 2)$

(But $\mathbf{P}[X \geq 12] = \int_{12}^{\infty} f(x)dx$ has no analytic solution.)

1. Compute $z = \frac{(x-\mu)}{\sigma}$:

$$\mathbf{P}[X \geq 12] = 1 - F_x(12) = 1 - \Phi\left(\frac{12-10}{\sqrt{2}}\right) \approx 1 - \Phi(1.41)$$

2. Look up $\Phi(z)$ in table:

$$1 - \Phi(1.41) \approx 1 - 0.9207 = 0.0793$$