Course 02323 Introduction to Statistics & 02402 Statistics (Polytechnical Foundation)

Lecture 3: Random variables and continuous distributions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

1. Summary
2. Continuous Distributions
   - Density and Distribution Functions
   - Mean, variance, and covariance
3. Specific continuous distributions
   - The Uniform distribution
   - The Normal distribution
   - The log-normal distribution
   - The Exponential distribution
4. Rules for stochastic variables
5. Extra: Multidimensional Random Variables

Overview

Summary: Week 2, Discrete Distributions

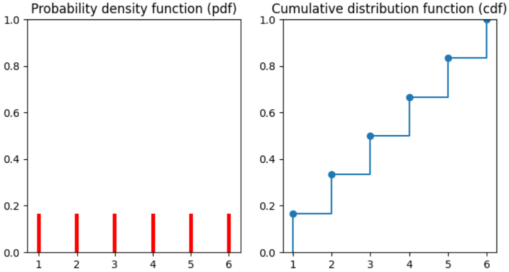
A stochastic variable: X

Probability density function (*pdf* / *pmf*):

$$f(x) = \mathrm{P}(X = x)$$

Cumulative distribution function (*cdf*):

$$F(x) = \mathrm{P}(X \leq x)$$



Expectation value:

$$\mathrm{E}[X] = \sum_{x \in A} x \, \mathrm{P}(X = x) = \mu$$

Variance:

$$\mathrm{V}[X] = \sum_{x \in A} (x - \mu)^2 \, \mathrm{P}(X = x) = \sigma^2$$

Alternatively:

$$\mathrm{V}[X] = \mathrm{E}[(X - \mu)^2]$$

Overview

Continuous Random Variables

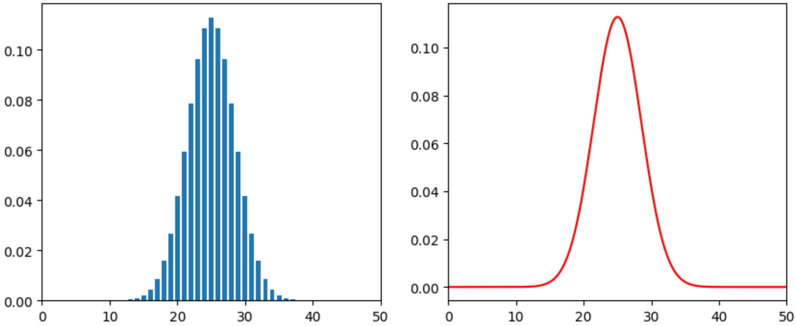
Random variable X.

The sample space S is now continuous.

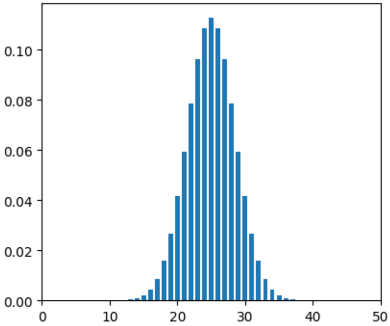
Examples:

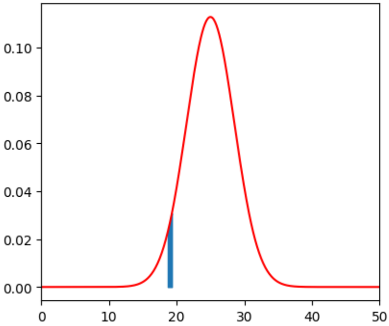
- Height of students
- Measurement of wind speed
- Time to cycle to DTU
- Measurement of blood sugar in patients
- ...

Continous probability distribution

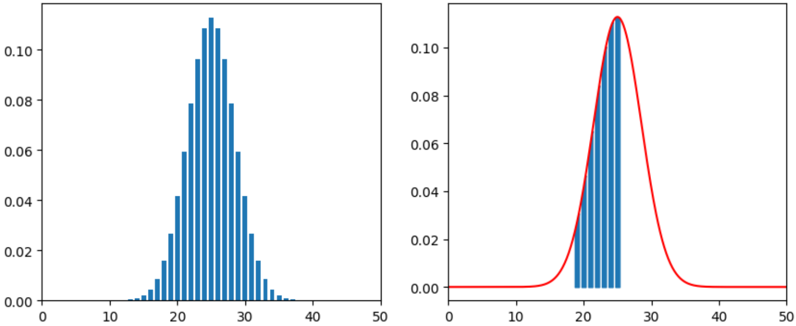


Continous probability distribution





Continous probability distribution



The density function, Definition 2.32

- The probability density function (pdf) for a random variable is denoted by $f(x)$.

The density function, Definition 2.32

- The probability density function (pdf) for a random variable is denoted by $f(x)$.
- The density function says something about the frequency of the outcome x for the random variable X.

The density function, Definition 2.32

- The probability density function (pdf) for a random variable is denoted by $f(x)$.
- The density function says something about the frequency of the outcome x for the random variable X.
- The probability that X takes a value in the interval $[a;b]$ is given by the integral (the area under the curve):

$$P(a < X \leq b) = \int_a^b f(x)\,dx.$$

The density function, Definition 2.32

- The probability density function (pdf) for a random variable is denoted by $f(x)$.
- The density function says something about the frequency of the outcome x for the random variable X.
- The probability that X takes a value in the interval $[a;b]$ is given by the integral (the area under the curve):

$$P(a < X \leq b) = \int_a^b f(x)\,dx\,.$$

- *No* direct probability for pdf. In fact, $P(X = x) = 0$ for all x.

The density function, Definition 2.32

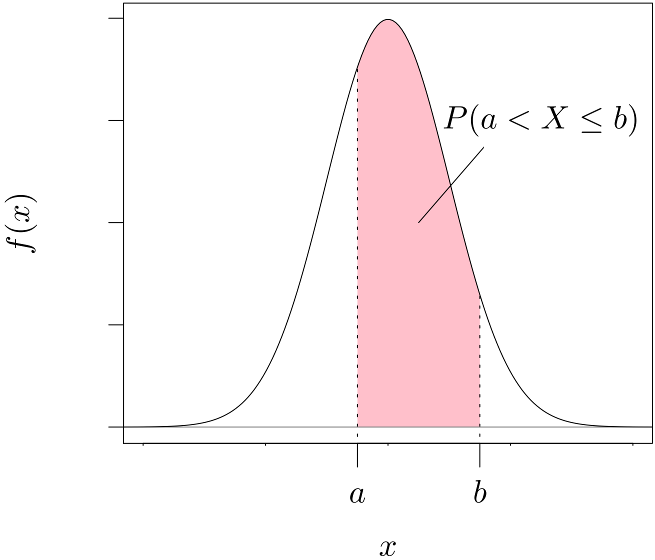
- The probability density function (pdf) for a random variable is denoted by $f(x)$.
- The density function says something about the frequency of the outcome x for the random variable X.
- The probability that X takes a value in the interval $[a;b]$ is given by the integral (the area under the curve):

$$P(a < X \leq b) = \int_a^b f(x)\,dx\,.$$

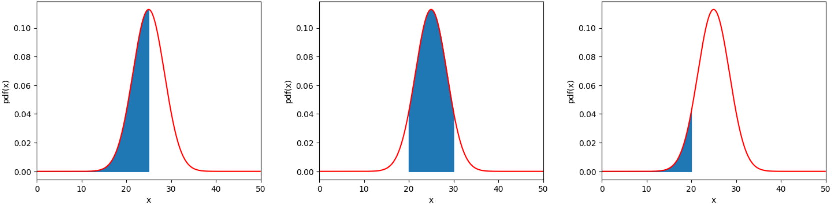
- *No* direct probability for pdf. In fact, $P(X = x) = 0$ for all x.
- The density function $f(x)$ for the distribution of a continuous random variable satisfies that

$$f(x) \geq 0 \text{ for all } x \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1\,.$$

The density function (Continuous)

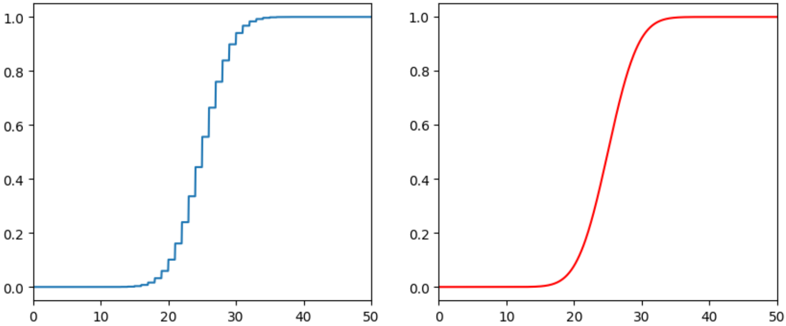


Kahoot



(x4)

Distribution Function for Continuous Variables



The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.

The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.
- The distribution function is defined by

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

The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.
- The distribution function is defined by

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

- Note that as a consequence of this definition,

$$f(x) = F^{'}(x)\,.$$

The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.
- The distribution function is defined by

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

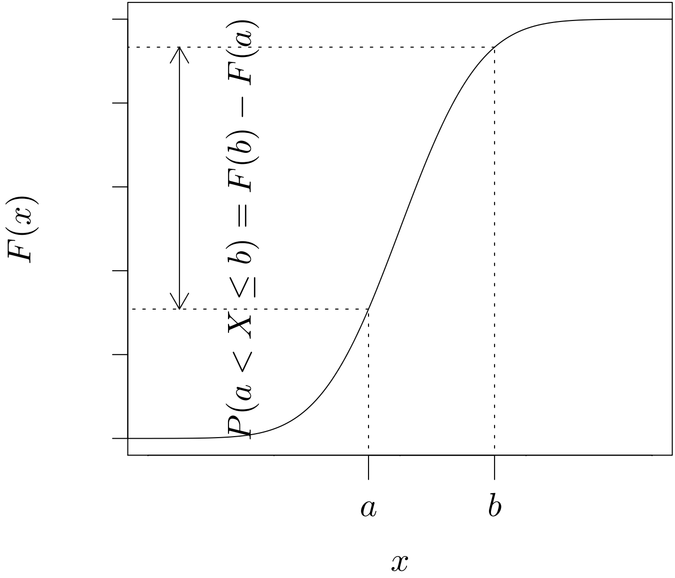
- Note that as a consequence of this definition,

$$f(x) = F^{'}(x)\,.$$

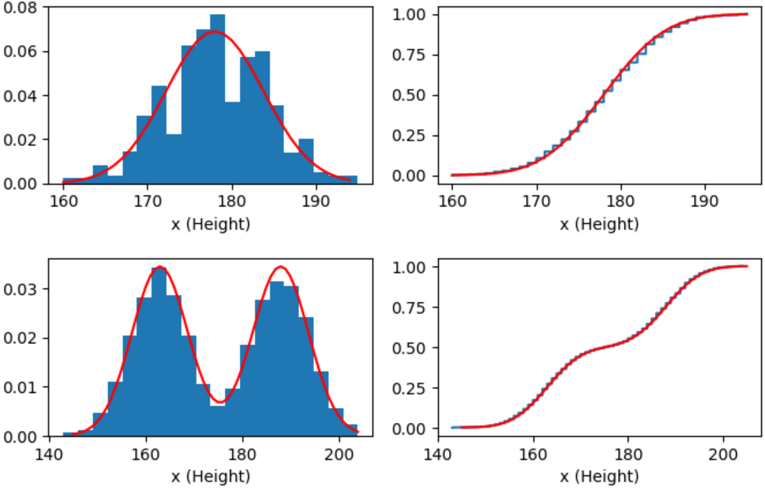
- It's particularly useful to note that

$$P(a < X \leq b) = \int_{a}^{b} f(x)\, dx \; = F(b) - F(a).$$

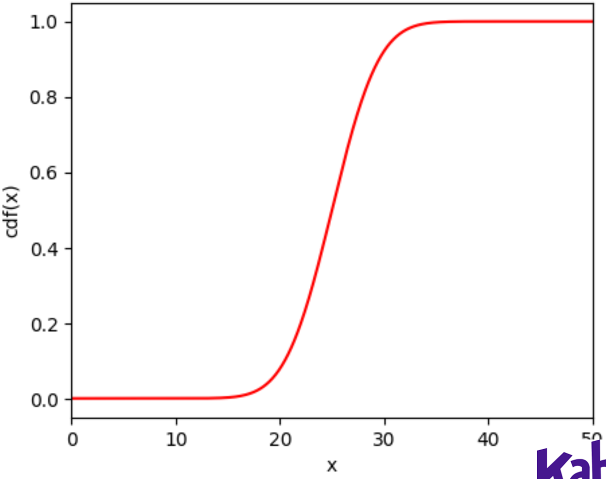
Continuous distribution function



More Examples



Kahoot



Kahoot!

(x2)

Mean, continuous random variable, Definition 2.34

The mean/expected value of a continuous random variable:

$$\mu = \mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

Mean, continuous random variable, Definition 2.34

The mean/expected value of a continuous random variable:

$$\mu = \mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

Compare with the mean of a discrete random variable:

$$\mu = \mathrm{E}[X] = \sum_{\text{all } x} x f(x)$$

Variance, continuous random variable, Definition 2.34

The variance of a continuous random variable:

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

Variance, continuous random variable, Definition 2.34

The variance of a continuous random variable:

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

Compare with the variance of a discrete random variable:

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \sum_{\text{all } x} (x-\mu)^2 f(x)$$

Covariance, Definition 2.58

The covariance between two random variables:

Let X and Y be two random variables. Then, the covariance between X and Y is

$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Relationship between covariance and independence:

If two random variables are *independent* their covariance is 0. *The reverse is not necessarily true!*

Overview

Specific continuous distributions

A number of statistical distributions exist (both continuous and discrete) that can be used to describe and analyze different types of problems.

Today, we'll take a closer look at the following continuous distributions:

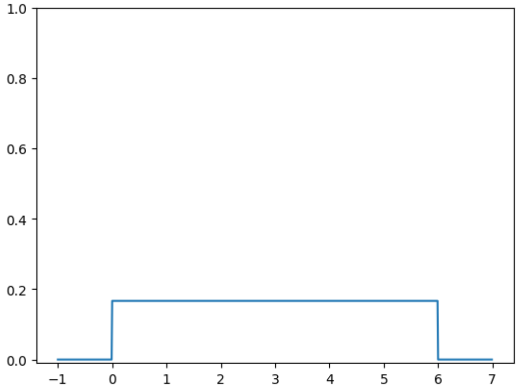
- The uniform distribution
- The normal distribution
- The log-normal distribution
- The exponential distribution

As we did with the discrete distributions, will we use `Scipy.stats` for the continuous distributions (see documentation online).

General 'methods' for different distributions are:

`scipy.stats`	`.uniform/.norm/.lognorm/.expon`
`.rvs`	'random variates' (simulate random numbers)
`.pdf`	'probability density function' (pdf/density function)
`.cdf`	'cumulative distribution function' (distribution function)
`.ppf`	'percent point function' (inverse cdf / quantile function)
`.mean /.var /.std`	'mean'/'variance'/'standard deviation'

Density for a Uniform Distribution (Example)



The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$X \sim U(\alpha, \beta)$

The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$X \sim U(\alpha, \beta)$

Density function:

$f(x) = \frac{1}{\beta - \alpha}$ for $\alpha \leq x \leq \beta$

The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$X \sim U(\alpha, \beta)$

Density function:

$f(x) = \frac{1}{\beta - \alpha}$ for $\alpha \leq x \leq \beta$

Mean:

$\mu = \frac{\alpha + \beta}{2}$

The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$X \sim U(\alpha, \beta)$

Density function:

$f(x) = \frac{1}{\beta - \alpha}$ for $\alpha \leq x \leq \beta$

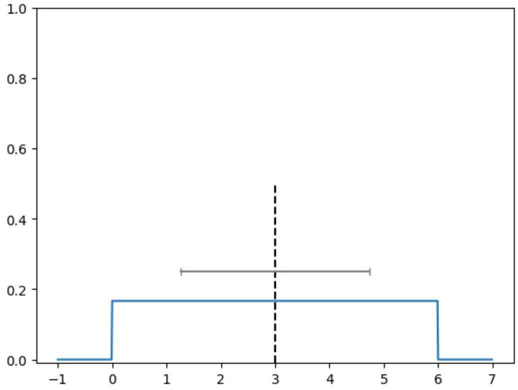
Mean:

$\mu = \frac{\alpha + \beta}{2}$

Variance:

$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$

Density for a Uniform Distribution (Example)



Example 1

Students in a statistics course arrive at a lecture between 8:00 and 8:30. It is assumed that the arrival time can be described by a uniform distribution.

Let $X \sim U(0, 30)$ represent the "arrival time" for a randomly selected student.

Question:

What is the probability that a randomly selected student arrives between 8:20 and 8:30?

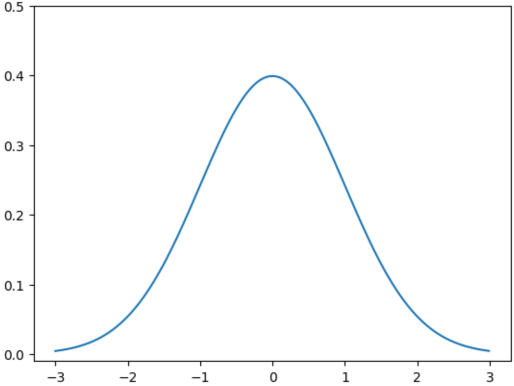
Question:

What is the probability that a randomly selected student arrives after 8:30?

Kahoot! (x2)

Visual Studio Code

Density of a normal distribution (example)



The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$X \sim N(\mu, \sigma^2)$

The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$$X \sim N(\mu, \sigma^2)$$

Density function:

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

The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$X \sim N(\mu, \sigma^2)$

Density function:

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

Mean:

$\mu = \mu$

The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$$X \sim N(\mu, \sigma^2)$$

Density function:

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

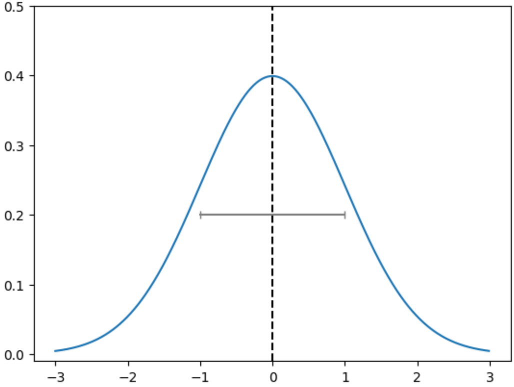
Mean:

$$\mu = \mu$$

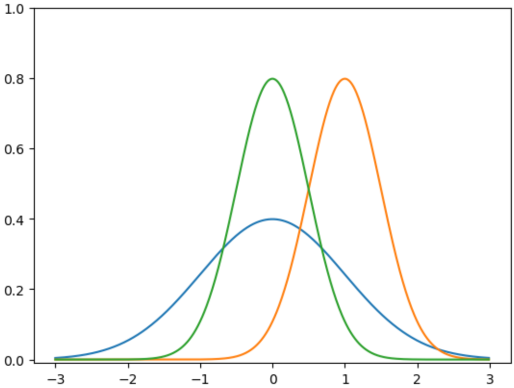
Variance:

$$\sigma^2 = \sigma^2$$

Density for a normal distribution (example)



More examples



The standard normal distribution

The standard normal distribution:

$$Z \sim N(0, 1^2)$$

The normal distribution with mean 0 and variance 1.

The standard normal distribution

The standard normal distribution:

$$Z \sim N(0, 1^2)$$

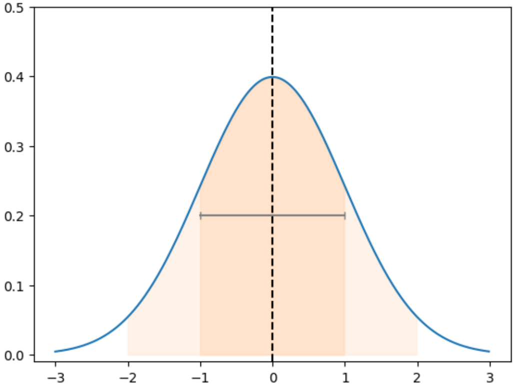
The normal distribution with mean 0 and variance 1.

Standardization:

An arbitrary normal distributed variable $X \sim N(\mu, \sigma^2)$ can be *standardized* by

$$Z = \frac{X - \mu}{\sigma}$$

The Standard Normal Distribution



Within $\mu \pm \sigma$, about 68.3% of the probability mass lies
Within $\mu \pm 2\sigma$, about 95.4% of the probability mass lies

Example 2

Measurement Error:

A given scale has a measurement error (measured in grams), Z, which can be described by a standard normal distribution, $Z \sim N(0, 1^2)$.

This means that the average measurement error is $\mu = 0$ grams and the standard deviation is $\sigma = 1$ gram.

Question:

a) What is the probability that the scale gives a result that is at least 2 grams less than the true weight of the product?
b) What is the probability that the scale gives a result that is at least 2 grams more than the true weight of the product?
c) What is the probability that the scale has a deviation of at most ± 1 gram?

Visual Studio Code

Example 3

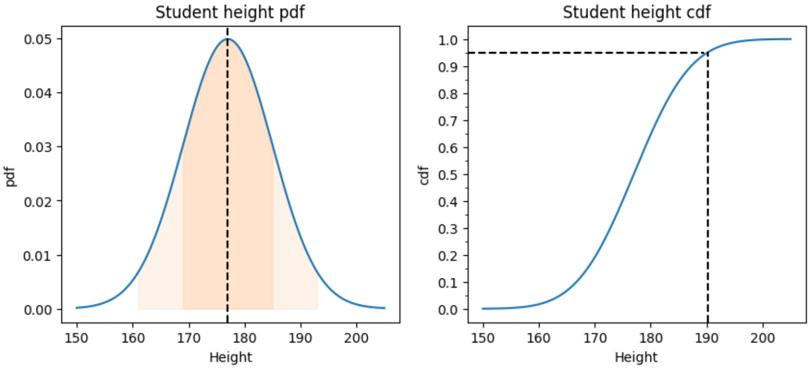
Income Distribution:

It is assumed that the salary of primary school teachers can be described by a normal distribution with a mean value of $\mu = 290$ (in 1000 DKK) and a standard deviation of $\sigma = 4$ (1000 DKK).

Question:

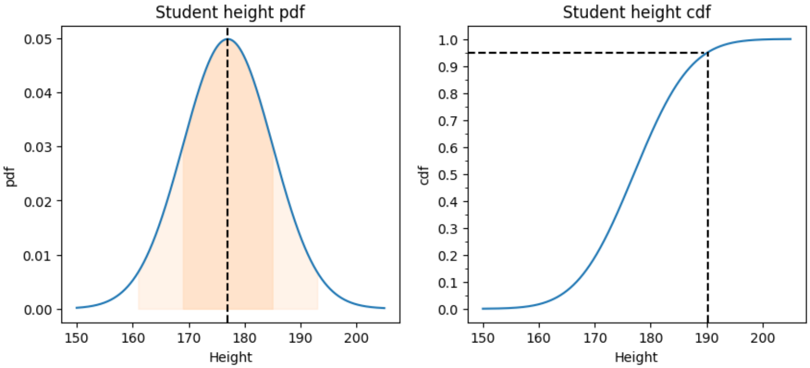
a) What is the probability that a randomly selected teacher earns more than 300,000 DKK?
b) (Inverse question) Specify a salary range (which is symmetric around the mean) that covers 95% of teachers' salaries.

Connection between distribution and quantiles



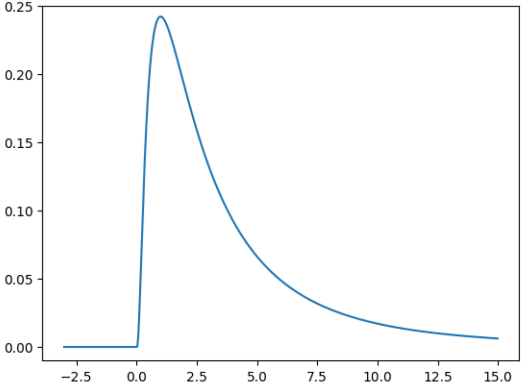
- What is the median height? (and how do you read it from the two plots?)
- What are Q1 (the first quartile) and Q3 (the third quartile)?
- How tall should a door be if 95% of students should pass through without bending?

Connection between distribution and percentiles



- Quantiles (percentiles)3 = "Averaged inverted cdf"
- In Python, you use ".ppf" - e.g. `stats.norm.ppf(q=0.95, loc=`μ`, scale=`σ`)`

The log-normal distribution



The log-normal distribution, Def. 2.46 & Thm. 2.47

Notation:

$X \sim LN(\alpha, \beta^2)$ (where $\beta > 0$)

The log-normal distribution, Def. 2.46 & Thm. 2.47

Notation:

$X \sim LN(\alpha, \beta^2)$ (where $\beta > 0$)

Density function:

$$f(x) = \begin{cases} \frac{1}{\beta\sqrt{2\pi}} x^{-1} e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The log-normal distribution, Def. 2.46 & Thm. 2.47

Notation:

$X \sim LN(\alpha, \beta^2)$ (where $\beta > 0$)

Density function:

$$f(x) = \begin{cases} \frac{1}{\beta\sqrt{2\pi}} x^{-1} e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean:

$\mu = e^{\alpha+\beta^2/2}$

The log-normal distribution, Def. 2.46 & Thm. 2.47

Notation:

$X \sim LN(\alpha, \beta^2)$ (where $\beta > 0$)

Density function:

$$f(x) = \begin{cases} \frac{1}{\beta\sqrt{2\pi}} x^{-1} e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean:

$$\mu = e^{\alpha+\beta^2/2}$$

Variance:

$$\sigma^2 = e^{2\alpha+\beta^2}(e^{\beta^2} - 1)$$

The log-normal distribution

Log-normal and normal distribution:

A log-normally distributed variable $Y \sim LN(\alpha, \beta^2)$ can be transformed into a normally distributed variable X by:

$$X = \ln(Y).$$

Here X is normally distributed with mean α and variance β^2, i.e. $X \sim N(\alpha, \beta^2)$.

The log-normal distribution

Log-normal and normal distribution:

A log-normally distributed variable $Y \sim LN(\alpha, \beta^2)$ can be transformed into a normally distributed variable X by:

$$X = \ln(Y).$$

Here X is normally distributed with mean α and variance β^2, i.e. $X \sim N(\alpha, \beta^2)$.

By standardizing X through

$$Z = \frac{X - \mathrm{E}[X]}{\sqrt{\mathrm{V}[X]}} = \frac{\ln(Y) - \alpha}{\beta}$$

you get a standard normally distributed variable $Z \sim N(0, 1)$.

The log-normal distribution

Log-normal and normal distribution:

A log-normally distributed variable $Y \sim LN(\alpha, \beta^2)$ can be transformed into a normally distributed variable X by:

$$X = \ln(Y).$$

Here X is normally distributed with mean α and variance β^2, i.e. $X \sim N(\alpha, \beta^2)$.

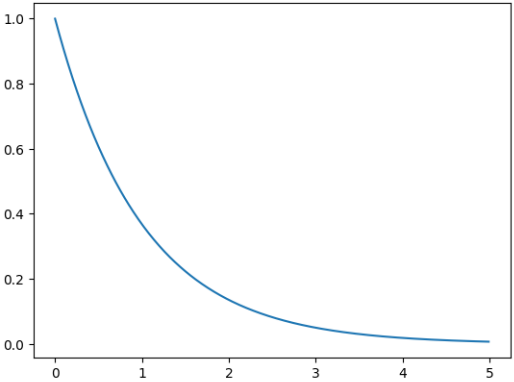
By standardizing X through

$$Z = \frac{X - \mathrm{E}[X]}{\sqrt{\mathrm{V}[X]}} = \frac{\ln(Y) - \alpha}{\beta}$$

you get a standard normally distributed variable $Z \sim N(0, 1)$.

Note: In Python, Y (the log-normally distributed variable) is parameterized with loc $= 0$, $s = \beta$ and scale $= e^{\alpha}$

The exponential distribution



The exponential distribution, Def. 2.48 & Thm. 2.49

Notation:

$X \sim \text{Exp}(\lambda)$, where $\lambda > 0$.

Density function:

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

Mean:

$\mu = \frac{1}{\lambda}$

Variance:

$\sigma^2 = \frac{1}{\lambda^2}$

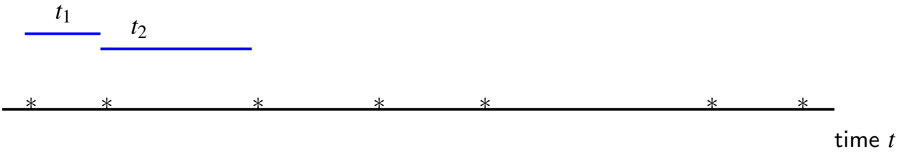
The exponential distribution

- The exponential distribution is often used to model lifetimes and waiting times.
- The exponential distribution can describe the (waiting) time between events in a Poisson process.
- The exponential distribution is a special case of the *gamma distribution*.

Connection between the exponential and Poisson distribution

Poisson: Number of events per (time) unit

Exponential: Continuous distance between events



Example 4

Queue model: Poisson process

The time between customer arrivals at a post office follows an exponential distribution with a mean of $\mu = 2$ minutes.

Question:

a) A customer has just arrived. What is the probability that no more customers will arrive within a period of 2 minutes?
b) Now use the **Poisson** *distribution to calculate the probability that no more customers will arrive within the next two minutes.*

Visual Studio Code

Overview

Rules for stochastic variables

These rules apply to both continuous and discrete random variables!

Let X be a random variable, while a and b are constants.

Mean rule:

$$\mathrm{E}[aX+b] = a\mathrm{E}[X] + b$$

Variance rule:

$$\mathrm{V}[aX+b] = a^2\mathrm{V}[X]$$

Example 5

Let X be a random variable with mean $\mathrm{E}[X] = 4$ and variance $\mathrm{V}[X] = 6$.

Question:

Calculate the mean and variance of $Y = -3X + 2$.

Visual Studio Code

Example 5

Let X be a random variable with mean $\mathrm{E}[X] = 4$ and variance $\mathrm{V}[X] = 6$.

Question:

Calculate the mean and variance of $Y = -3X + 2$.

Answer:

$$\mathrm{E}[Y] = -3\mathrm{E}[X] + 2 = -3 \cdot 4 + 2 = -10$$
$$\mathrm{V}[Y] = (-3)^2\mathrm{V}[X] = 9 \cdot 6 = 54$$

Visual Studio Code

Rules for Random Variables

Let $X_1, \ldots, X_n$ be *independent* random variables.

Rules for Random Variables

Let $X_1, \ldots, X_n$ be *independent* random variables.

Mean Rule:

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n]$$
$$= a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n]$$

Rules for Random Variables

Let $X_1, \ldots, X_n$ be *independent* random variables.

Mean Rule:

$$\mathrm{E}[a_1X_1 + a_2X_2 + \cdots + a_nX_n]$$
$$= a_1\mathrm{E}[X_1] + a_2\mathrm{E}[X_2] + \cdots + a_n\mathrm{E}[X_n]$$

Variance Rule:

$$\mathrm{V}[a_1X_1 + a_2X_2 + \cdots + a_nX_n]$$
$$= a_1^2\mathrm{V}[X_1] + \cdots + a_n^2\mathrm{V}[X_n]$$

Example 6

Planning for Airline

The individual weight of passengers on a flight, X, is assumed to be normally distributed as $X \sim N(70, 10^2)$.

A plane that can take 55 passengers can be loaded with a maximum of 4000 kg (only the passengers' weight is considered here as the load).

Example 6

Planning for Airline

The individual weight of passengers on a flight, X, is assumed to be normally distributed as $X \sim N(70, 10^2)$.

A plane that can take 55 passengers can be loaded with a maximum of 4000 kg (only the passengers' weight is considered here as the load).

Question:

Calculate the probability that the plane will be overloaded.

Example 6

Planning for Airline

The individual weight of passengers on a flight, X, is assumed to be normally distributed as $X \sim N(70, 10^2)$.

A plane that can take 55 passengers can be loaded with a maximum of 4000 kg (only the passengers' weight is considered here as the load).

Question:

Calculate the probability that the plane will be overloaded.

What is the total weight Y of 55 passengers on a flight?

How do we mathematically describe the random variable Y?

Example 6

Planning for Airline

The individual weight of passengers on a flight, X, is assumed to be normally distributed as $X \sim N(70, 10^2)$.

A plane that can take 55 passengers can be loaded with a maximum of 4000 kg (only the passengers' weight is considered here as the load).

Question:

Calculate the probability that the plane will be overloaded.

What is the total weight Y of 55 passengers on a flight?

How do we mathematically describe the random variable Y?

...

NOT: $Y = 55 \cdot X$

Example 8

What is the total weight Y of 55 passengers on a flight?

$Y = \sum_{i=1}^{55} X_i$, where $X_i \sim N(70, 10^2)$ (assumed to be independent)

Example 8

What is the total weight Y of 55 passengers on a flight?

$Y = \sum_{i=1}^{55} X_i$, where $X_i \sim N(70, 10^2)$ (assumed to be independent)

Mean and Variance of Y:

Example 8

What is the total weight Y of 55 passengers on a flight?

$Y = \sum_{i=1}^{55} X_i$, where $X_i \sim N(70, 10^2)$ (assumed to be independent)

Mean and Variance of Y:

$$\mathrm{E}[Y] = \sum_{i=1}^{55} \mathrm{E}[X_i] = \sum_{i=1}^{55} 70 = 55 \cdot 70 = 3850$$

$$\mathrm{V}[Y] = \sum_{i=1}^{55} \mathrm{V}[X_i] = \sum_{i=1}^{55} 100 = 55 \cdot 100 = 5500$$

Example 8

What is the total weight Y of 55 passengers on a flight?

$Y = \sum_{i=1}^{55} X_i$, where $X_i \sim N(70, 10^2)$ (assumed to be independent)

Mean and Variance of Y:

$$\mathrm{E}[Y] = \sum_{i=1}^{55} \mathrm{E}[X_i] = \sum_{i=1}^{55} 70 = 55 \cdot 70 = 3850$$

$$\mathrm{V}[Y] = \sum_{i=1}^{55} \mathrm{V}[X_i] = \sum_{i=1}^{55} 100 = 55 \cdot 100 = 5500$$

Y is normally distributed, so we can find $P(Y > 4000)$ by:

$1 - F(y = 4000; \mu = 3850, \sigma^2 = 5500) = 0.02156$ (try it yourself in Python)

Example 8 - WRONG Analysis

What is Y?

NOT: $Y = 55 \cdot X$

Example 8 - WRONG Analysis

What is Y?

NOT: $Y = 55 \cdot X$

Mean and Variance of WRONG Y:

$$E[Y] = 55 \cdot 70 = 3850$$
$$V[Y] = 55^2 V[X] = 55^2 \cdot 100 = 302500$$

Example 8 - WRONG Analysis

What is Y?

NOT: $Y = 55 \cdot X$

Mean and Variance of WRONG Y:

$$\mathrm{E}[Y] = 55 \cdot 70 = 3850$$

$$\mathrm{V}[Y] = 55^2 \mathrm{V}[X] = 55^2 \cdot 100 = 302500$$

The WRONG Y is also normally distributed. Here we find $P(Y > 4000)$ with WRONG Y:

$1 - F(y = 4000; \mu = 3850, \sigma^2 = 550^2) = 0.3925$ (try it yourself in Python)

Example 8 - WRONG Analysis

What is Y?

NOT: $Y = 55 \cdot X$

Mean and Variance of WRONG Y:

$$\mathrm{E}[Y] = 55 \cdot 70 = 3850$$
$$\mathrm{V}[Y] = 55^2 \mathrm{V}[X] = 55^2 \cdot 100 = 302500$$

The WRONG Y is also normally distributed. Here we find $P(Y > 4000)$ with WRONG Y:

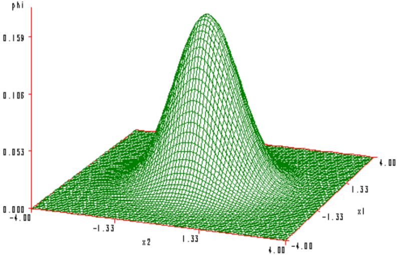
$1 - F(y = 4000; \mu = 3850, \sigma^2 = 550^2) = 0.3925$ (try it yourself in Python)

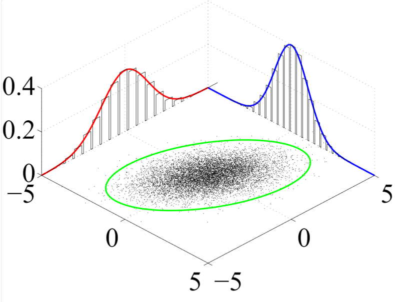
Consequence of incorrect calculation:

MANY wasted money for the airline!!!

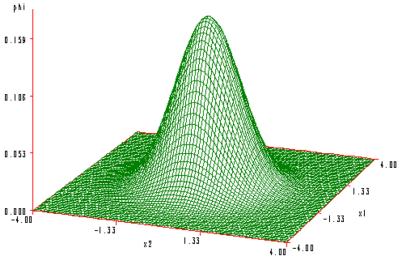
Overview

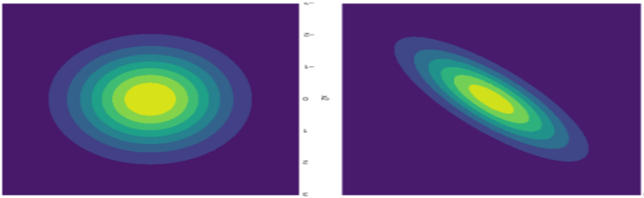
Extra: Multidimensional Random Variables



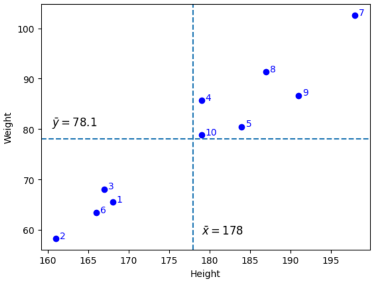


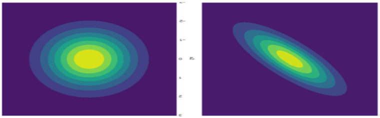
Extra: Multidimensional Random Variables





Extra: Multidimensional Random Variables





Sample Covariance:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Covariance:

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$$

Agenda