Introduction to Statistics

Week 3: Continuous distributions

Peder Bacher

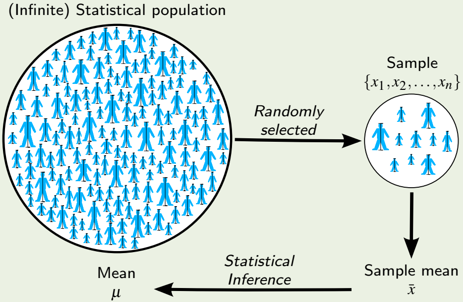
DTU Compute, Dynamical Systems
Building 303B, Room 009
Technical University of Denmark
2800 Lyngby – Denmark
e-mail: pbac@dtu.dk

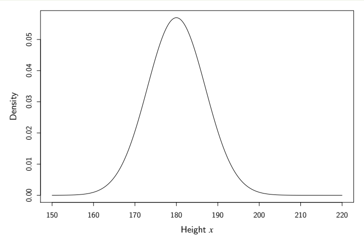
Spring 2017

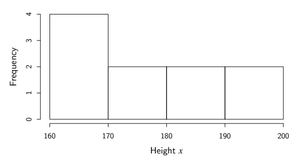
Agenda

1. Continuous random variables and distributions
   - The density function
   - Distribution function
   - The mean of a continuous random variable
   - The variance of a continuous random variable
2. Specific statistical distributions
   - Continuous distributions in R
   - Uniform distribution
   - Normal distribution
   - Log-normal distribution
   - Exponential distribution
3. Identities for the mean and variance

Example: Population and distribution



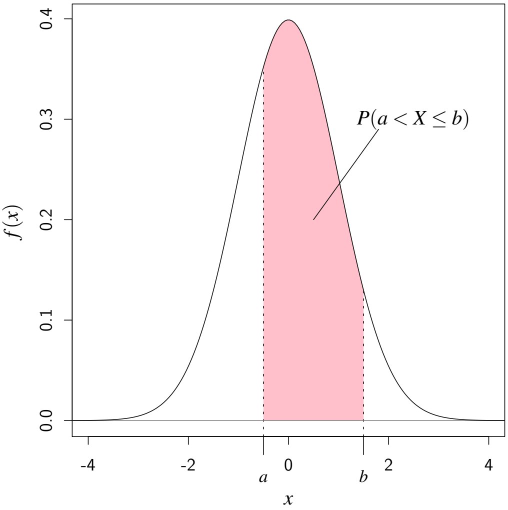




The density function (pdf)

- The density function for a stochastic variable is denoted by $f(x)$
- $f(x)$ says something about the frequency of the outcome x for the stochastic variable X
- The density function for continuous variables does not correspond to the probability, that is $f(x) \neq P(X = x)$
- A nice plot of $f(x)$ is a histogram (continuous)

The density function for continuous variables



The density function for continuous variables

- The probability density function (pdf) for a continuous variable is written by

$$f(x)$$

- The following is valid:
  - No negative values

$$f(x) \geq 0 \quad \text{for all possible values of } x$$

  - The area under the curve is one

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

Distribution function or cumulative density function (cdf))

- The distribution function for a continuous stochastic variable is denoted by

$$F(x)$$

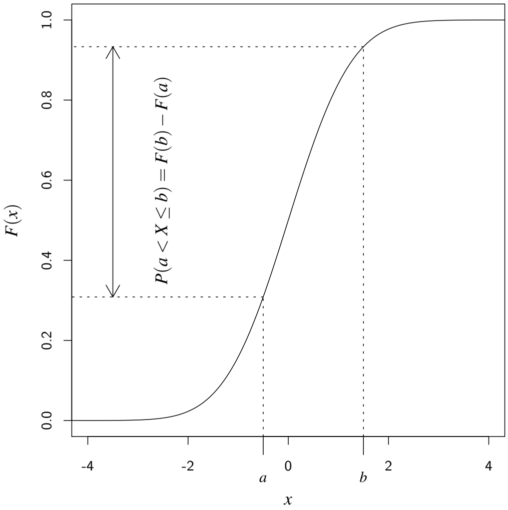
- The distribution function corresponds to the cumulative density function:

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

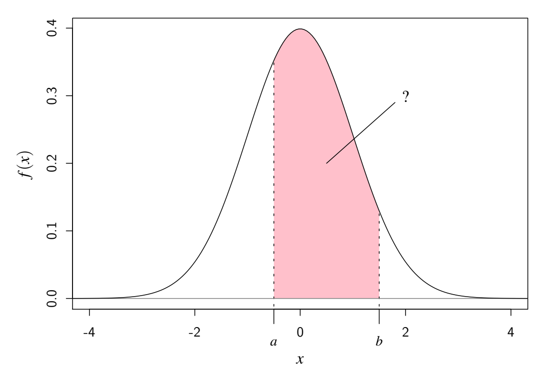
$$F(x) = \int_{t=-\infty}^{x} f(t)dt$$

$$f(x) = F^{'}(x)$$

The distribution function (cdf)



Question about probabilities

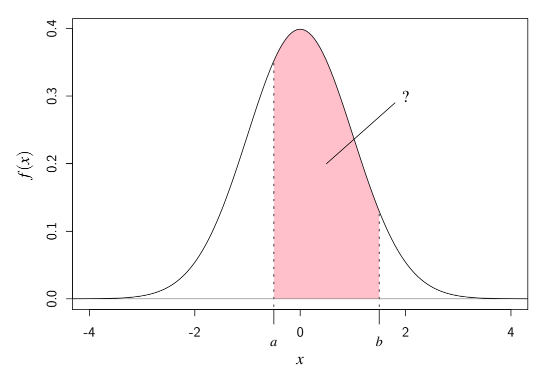


Which area (probability) is marked?

A: $\int_{-\infty}^{b} f(x)dx$ B: $1 - \int_{a}^{b} f(x)dx$ C: $\int_{a}^{b} f(x)dx$ D: $1 - \int_{a}^{\infty} f(x)dx$

Answer C: $\int_{a}^{b} f(x)dx$

Question about probabilities



How can we easiest calculate the marked area?

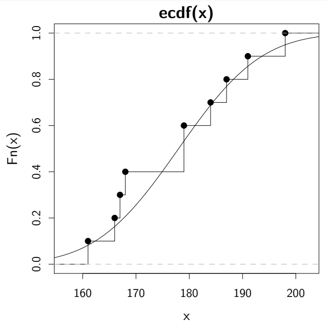
A: $\int_a^b f(x)dx$ B: $\int_a^b F(x)dx$ C: $f(b) - f(a)$ D: $F(b) - F(a)$

Answer D: $F(b) - F(a)$ *(we do it in R by (normal distributed):* `pnorm(b) - pnorm(a)`)

Example: ecdf vs. cdf

Student height example from Chapter 1:

```
## Plot the empirical cdf (ecdf) and estimated cdf
## Heights sample
x <- c(168,161,167,179,184,166,198,187,191,179)
## Plot the empirical cdf
plot(ecdf(x), verticals = TRUE)
## An x sequence
xp <- 150:210
## The estimated cdf
lines(xp, pnorm(xp, mean(x), sd(x)))
```



The mean of a continuous random variable

The mean of a continuous random variable

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Compare with the discrete definition: $\mu = \sum_{\text{all } x} x \cdot f(x)$

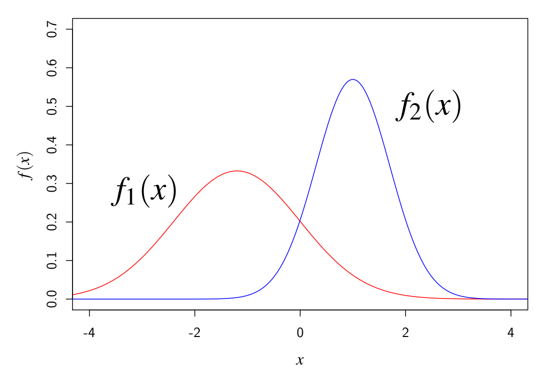
The variance of a continuous random variable

The variance of a continuous random variable:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x) dx$$

Compare with the discrete definition: $\sigma^2 = \sum_{\text{all } x} (x-\mu)^2 \cdot f(x)$

Question about mean

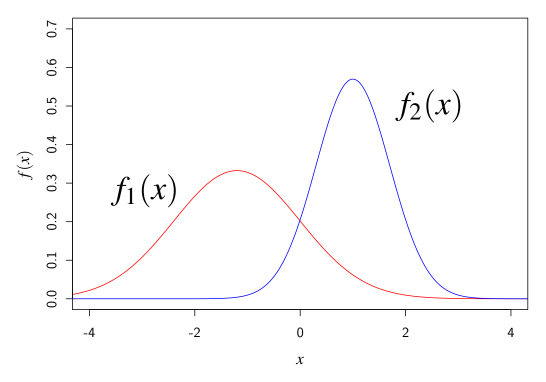


Which pdf have the highest mean (both are symmetric)?

A: $f_1(x)$ B: $f_2(x)$ C: None, $\mu_1 = \mu_2$ D: Cannot be answered

Answer B: $f_2(x)$ i.e. $\mu_1 < \mu_2$

Spørgsmål om spread



Which pdf has the highest variance (both are symmetric)?

A: $f_1(x)$ B: $f_2(x)$ C: None, $\sigma_1^2 = \sigma_2^2$ D: Cannot be answered

Answer A: $f_1(x)$ i.e. $\sigma_1^2 > \sigma_2^2$

Specific Statistical Distributions

A number of statistical distributions exist that can be used to describe and analyze different kind of problems

- Now we consider <u>continuous</u> distributions:
  - The uniform distribution
  - The normal distribution
  - The log-normal distribution
  - The exponential distribution

Continuous distributions in R

R	Name
`norm`	Normal distribution
`unif`	Uniform distribution
`lnorm`	Log-normal distribution
`exp`	Exponential distribution

- `d` Density function $f(x)$ (probability density function, pdf)
- `p` Distribution function $F(x)$ (cumulative distribution function, cdf)
- `q` Quantile in distribution
- `r` Random numbers from the distribution

Uniform distribution

Syntax:

$X \sim U(\alpha, \beta)$ (Read: X follows a uniform distribution with parameters α and β)

Density function (pdf):

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

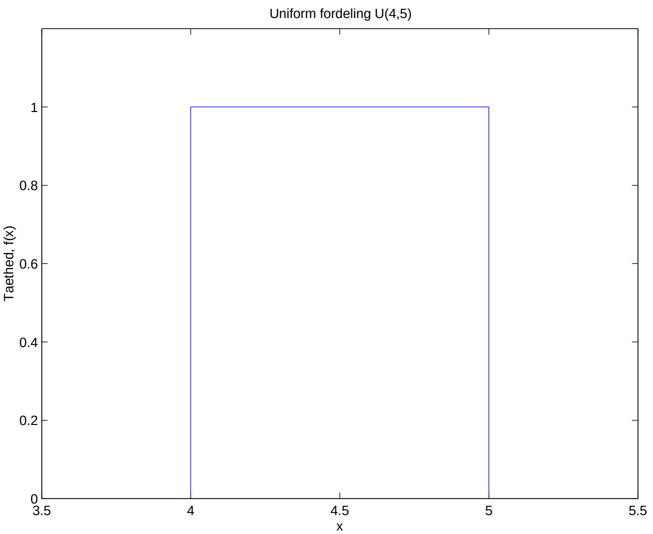
Mean:

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

Variance:

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

Example: Uniform distribution



Question: Uniform distribution 1

Employees arrives between 8:00 and 8:30. It is assumed that the arival times can be described by a uniform distribution.

What is the probability that a randomly selected employee arrives between 8:20 og 8:30?

A: $1/2$ B: $1/6$ C: $1/3$ D: 0

Answer C: $10/30 = 1/3$

```
punif(q=30,min=0,max=30) - punif(q=20,min=0,max=30)
```

[1] 0.33

Question: Uniform distribution 2

Employees arrives between 8:00 and 8:30. It is assumed that the arival times can be described by a uniform distribution.

What is the probability that a randomly selected employee arrives later than 8:30?

A: $1/2$ B: $1/6$ C: $1/3$ D: 0

Answer D: $P(X > 30) = 0$

```
1 - punif(q=30,min=0,max=30)
```

[1] 0

Normal distribution

Syntax:

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

Density function (pdf):

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

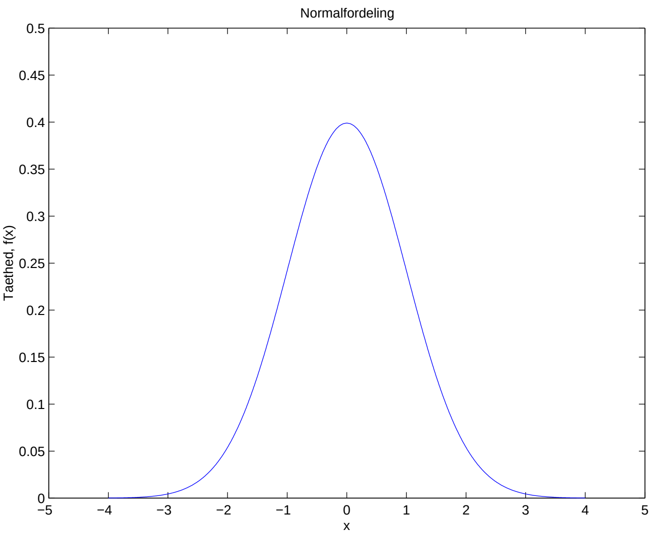
Mean:

$\mu = \mu$

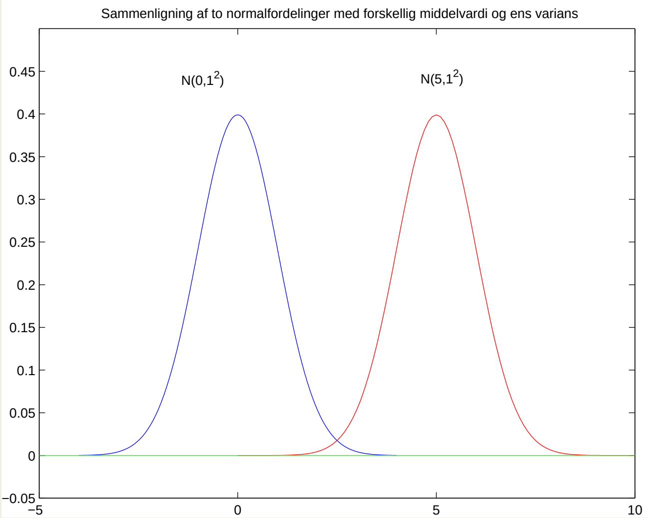
Variance:

$\sigma^2 = \sigma^2$

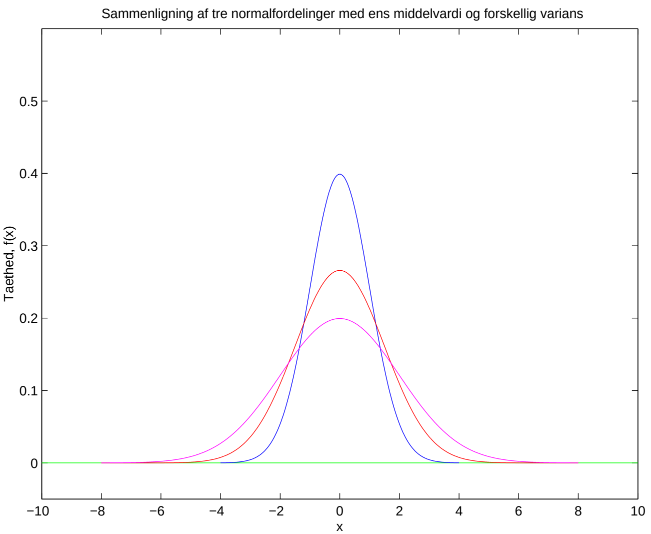
Example: Normal distributionen



Example: Normal distributionen



Example: Normal distributionen



Example: Normal distribution, probabilities

Distribution of weights of rye bread:

Assume that the weight of a rye bread from a production can be described with the normal distribution

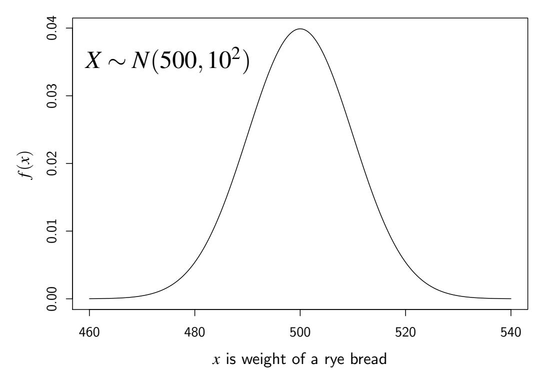
$$X \sim N(500, 10^2)$$

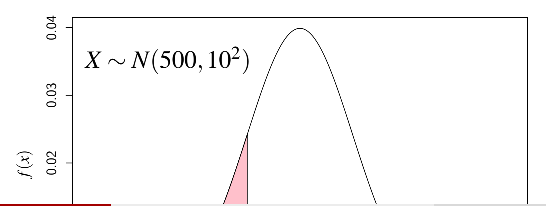
Hence, mean is $\mu = 500$ gram and standard deviation is $\sigma = 10$ gram.
We plan to measure the weight of a randomly chosen bread from the production.

Question:

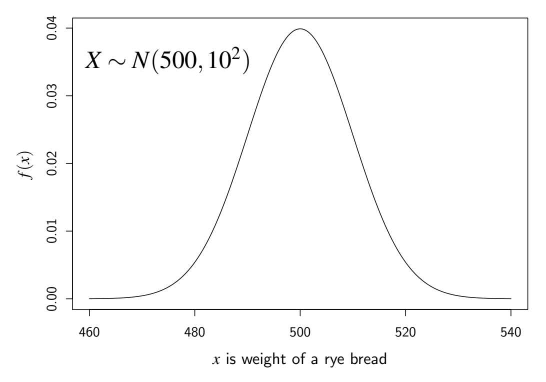
1: What is the probability that the bread weights less than 490 g?
2: What is the probability that the bread weights more than ± 20 g away from 500 g?

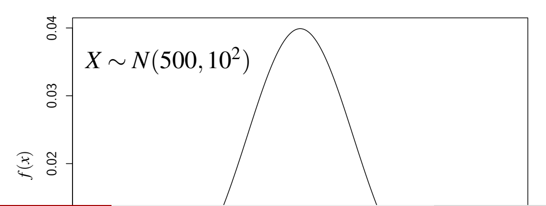
Example: Normal distribution, Question 1



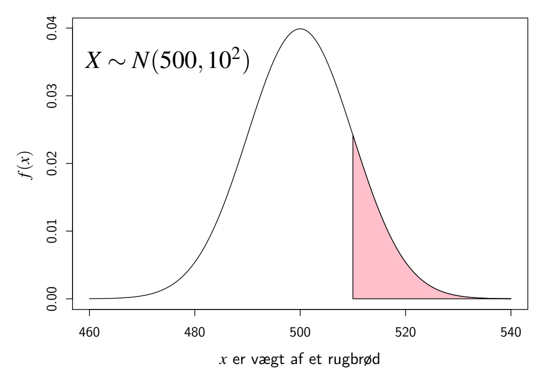


Example: Normaldistribution, Question 2





Question: Probability in the normal distribution

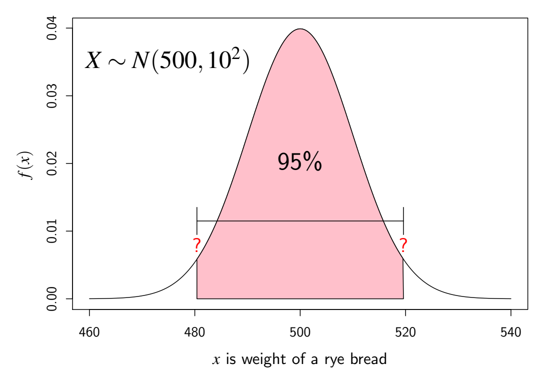


What is the probability that the bread weights more that 510 g equal to?

A: $F(510)$ B: $1 - F(490)$ C: $1 - F(520)$ D: $1 - F(510)$

Answer D: $P(X > 510) = 1 - P(X \leq 510) = 1 - F(510) = 0.16$

Example: Normal distribution quantiles



"Inverse question": *Which interval covers 95% of the rye breads?*

```
qnorm(c(0.025,0.975), mean=500, sd=10)
```

```
[1] 480.4 519.6
```

Standard normal distribution

A standard normal distribution

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

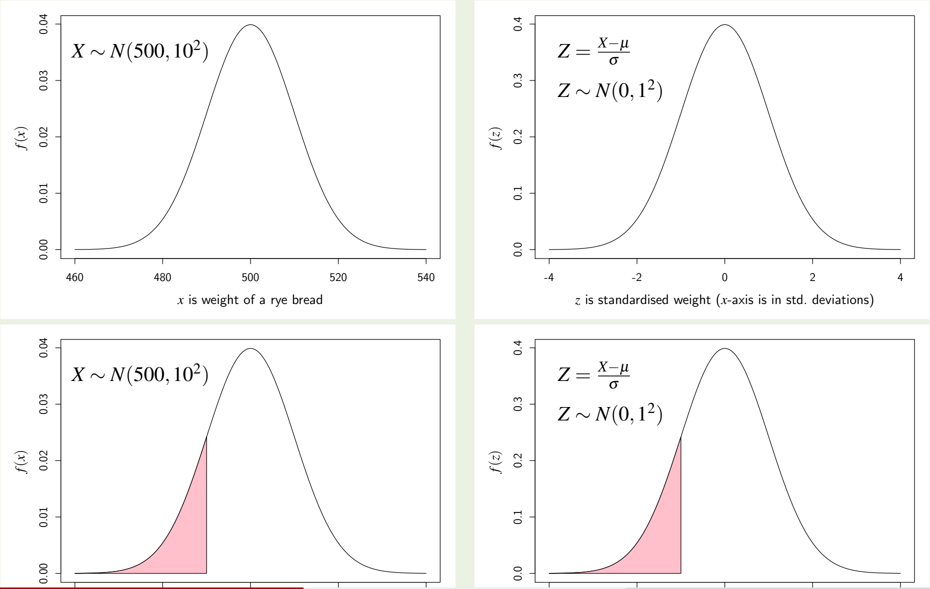
A normal distribution with mean 0 and variance 1.

Standardizing

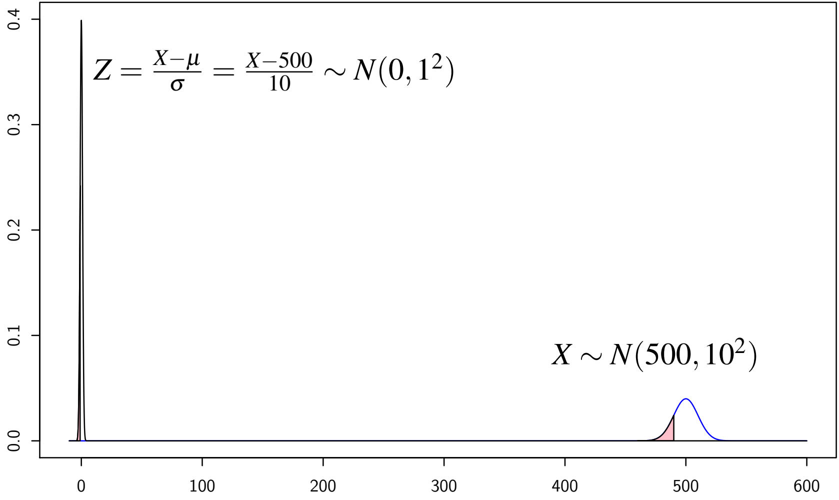
An arbitrary normally distributed variable $X \sim N(\mu, \sigma^2)$ can be standardized by

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

Example: Standard Normal distribution



Example: Transformation to standard normal distribution



Log-Normal distribution

Syntax:

$X \sim LN(\alpha, \beta^2)$ (If X follows the log-normal then $\ln(X)$ follows the normal distribution)

Density function, (pdf):

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

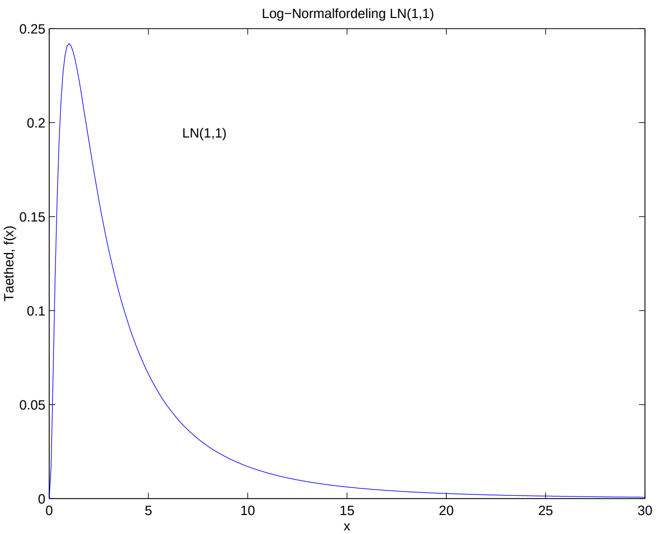
Mean:

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

Variance:

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

Example: Log-normal distribution



Log-normal distribution

Lognormal and normal distribution:

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

$$X = \frac{\ln(Y) - \alpha}{\beta}$$

hence

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

Exponential distribution

Syntax:

$X \sim Exp(\lambda)$

Density function (pdf):

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

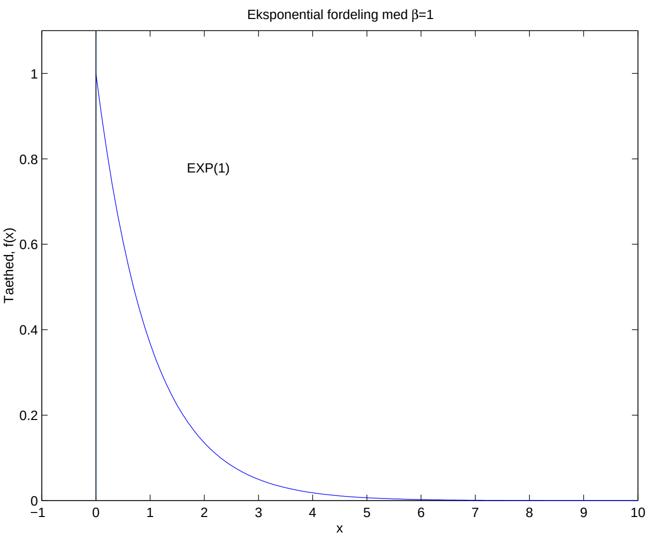
Mean:

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

Variance:

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

Example: Exponential distribution



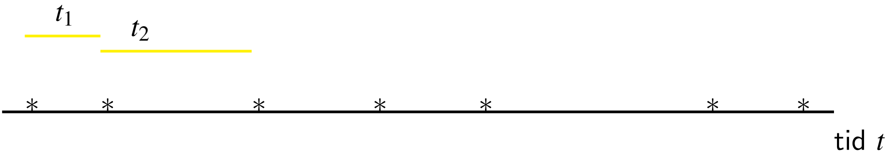
Exponential distribution

- The exponential distribution is a special case of the gamma distribution
- The exponential distribution is used to describe lifespan and waiting times
- The exponential distribution can be used to describe (waiting) time between Poisson events

Relation between exponential- og poisson distribution

Poisson: Discrete events pr./ unit

Exponential: Continuous distance between events



Example: Exponentiel distribution

Qeuing model - poisson proces

The time between customer arrivals at a post office is exponentially distributed with mean $\mu = 2$ minutes.

Question:

One customer is just arrived. What is the probability that no other costumers will arrive in the next period of 2 minutes?

Answer:

Use exponential: $X_{\text{exp}} \sim Exp(\lambda)$ with $\lambda = \frac{1}{\mu} = \frac{1}{2}$, find $P(X_{\text{exp}} > 2)$.

Use Poisson: $\lambda_{\text{2min}} = 1$, find $P(X_{\text{pois}} = 0) = \frac{e^{-1}}{1!} 1^0 = e^{-1}$.

```
1-pexp(q=2, rate=1/2); dpois(x=0, lambda=1); exp(x=-1)
```

```
[1] 0.37 [1] 0.37 [1] 0.37
```

Identities for the mean and variance

(Holds for AS WELL continuous as discrete variables)

- X is a random variable
- We assume that a and b are constants

Then:

Mean rule:

$$\mathrm{E}(aX+b) = a\,\mathrm{E}(X)+b$$

Variance rule:

$$\mathrm{V}(aX+b) = a^2\,\mathrm{V}(X)$$

Example: Calculation rules 1

X is a random variable

A random variable X has mean 4 and variance 6.

Question:

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

Answer:

$$\mathrm{E}(Y) = -3\,\mathrm{E}(X) + 2 = -3 \cdot 4 + 2 = -10$$

$$\mathrm{V}(Y) = (-3)^2\,\mathrm{V}(X) = 9 \cdot 6 = 54$$

Calculation rules for random variable

(Holds for AS WELL continuous as discrete variables)

- $X_1,\ldots,X_n$ are random variables
- $X_1,\ldots,X_n$ are independent

Then:

Mean rule:

$$\mathrm{E}(a_1X_1+a_2X_2+..+a_nX_n)=a_1\,\mathrm{E}(X_1)+a_2\,\mathrm{E}(X_2)+..+a_n\,\mathrm{E}(X_n)$$

Variance rule::

$$\mathrm{V}(a_1X_1+a_2X_2+..+a_nX_n)=a_1^2\,\mathrm{V}(X_1)+..+a_n^2\,\mathrm{V}(X_n)$$

Example: Calculation rules 2

Airline Planning

The weight of the passengers on a flight is assumed normal distributed $X \sim N(70, 10^2)$.

A plane, which can take 55 passengers, must not have a load exceeding more than 4000 kg (only the weight of the passengers is considered as load).

Question:

Calculate the probability that the plain is overloaded.

What is the total passenger weight Y?

A: $Y = 55 \cdot X$ B: $Y = \sum_{i=1}^{55} X_i$ C: $Y = 55 + X$ D: Not A,B or C

Answer B: $Y = \sum_{i=1}^{55} X_i$, it is the sum of 55 different passengers.

Example: Calculation rules 2

What is Y="Total passenger weight"?

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

Mean and variance of Y:

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

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

We use a normal distribution for Y:

```
1-pnorm(q=4000, mean = 3850, sd = sqrt(5500))
```

```
[1] 0.022
```

Example: Calculation rules 3 - WRONG ANALYSIS

What is Y?

Definitely NOT: $Y = 55 \cdot X$!!!!!!

Mean and variance of Y:

$$\mathrm{E}(Y) = 55 \cdot 70 = 3850$$

$$\mathrm{V}(Y) = 55^2\,\mathrm{V}(X) = 55^2 \cdot 100 = 550^2$$

We use a normal distribution for Y:

```
1-pnorm(q=4000, mean = 3850, sd = 550)
```

[1] 0.39

Consequence of wrong calculation:

A LOT of wasted money for the airline company!!!

Agenda

1. Continuous random variables and distributions
   - The density function
   - Distribution function
   - The mean of a continuous random variable
   - The variance of a continuous random variable
2. Specific statistical distributions
   - Continuous distributions in R
   - Uniform distribution
   - Normal distribution
   - Log-normal distribution
   - Exponential distribution
3. Identities for the mean and variance