Statistics

Lecture 7

Continuous probability distributions

SILESIAN UNIVERSITY
SCHOOL OF BUSINESS ADMINISTRATION IN KARVINA

David Bartl
Statistics
INM/BASTA

Outline of the lecture

- Continuous probability distributions
- Continuous uniform distribution
- Gaussian normal distribution
- Exponential distribution
- Some other continuous probability distributions

Experiment — Trial — Random variable

An **experiment** is any physical procedure which can end up with a result from a set of possible outcomes, and the experiment can be repeated (up to) infinitely many times. Each individual repetition of the experiment is called a **trial.**

The final result ω of (each trial of) the experiment is called an **outcome.**

The set Ω of all the outcomes is called the **sample space.** An event E is a set of outcomes (i.e. $E \subseteq \Omega$). The **event space** $\mathcal{F}$ is the collection of all events.

The event space is a σ-algebra, the **probability** $P\colon \mathcal{F} \to \mathbb{R}$ is a non-negative and σ-additive function. The triple $(\Omega, \mathcal{F}, P)$ is a **probability space** and a **random variable** is any measurable function $X\colon \Omega \to \mathbb{R}$.

Random variable — Dataset

To conclude, **the random variable assigns a numerical value to each outcome** of the random experiment.

Now, a **dataset** is a collection of measurements and observations, i.e. it is a collection of data. A **data unit** is an entity of the population under study, and a <u>**data item**</u> or a <u>**variable**</u> is a characteristics of each data unit.

We are considering numerical (quantitative) variables now.

<u>**We assume the hypothesis**</u> that the data items in the dataset are realizations of the random variable, i.e. **the random variable** (via the trials of the random experiment) **generates the data**.

Examples of continuous random variables

- The result of measurement of the dimension or weight of some object is a random variable X which (in theory) can attain any positive value $X \in \mathbb{R}^+$.

- The waiting time spent by a customer in the queue at the counter is a random variable X which (in theory) can attain any non-negative value $X \in \mathbb{R}_0^+$ $(X \geq 0)$.

- The stock price is a random variable X which (in theory) can attain any non-negative value $X \in \mathbb{R}_0^+$ $(X \geq 0)$.

Random variable

Given the probability space $(\Omega, \mathcal{F}, P)$,

the random variable X is any measurable function $X\colon \Omega \to \mathbb{R}$.

(Recall that the function $X\colon \Omega \to \mathbb{R}$ is <u>measurable</u> if and only if

$$\{\omega \in \Omega : a < X(\omega) < b\} \in \mathcal{F} \qquad \text{for every} \quad a, b \in \mathbb{R}$$

That is, the preimage $X^{-1}\big((a,b)\big)$ of any open interval is an event.)

Assumptions to simplify the matters

We assume for simplicity that

— either the probability space $(\Omega, \mathcal{F}, P)$ is **discrete**, i.e. the sample space is finite ($\Omega = \{1, \dots, N\}$) or countably infinite ($\Omega = \{1, 2, 3, \dots\}$) and the event space $\mathcal{F} = 2^{\Omega} = \{E : E \subseteq \Omega\}$ is the collection of all subsets of the sample space Ω

— or the probability space $(\Omega, \mathcal{F}, P)$ is **continuous**, i.e. the sample space is the set of the real numbers ($\Omega = \mathbb{R}$), the event space $\mathcal{F} \subset 2^{\mathbb{R}}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$, and the random variable $X: \mathbb{R} \to \mathbb{R}$ is assumed to be the identity function $(X(x) = x)$ only

Probability mass function

The discrete case (cases I & II) was the topic of the previous lecture. The discrete case is characterized by that there exists the probability mass function $p : \Omega \to \mathbb{R}$ of the probability $P : \mathcal{F} \to \mathbb{R}$ so that

$$P(E) = \sum_{\omega \in E} p(\omega) \qquad \text{for every event} \quad E \in \mathcal{F}$$

Notice that the probability mass function can be seen as a special case of the probability density function (see below).

Notice also that the random variable X can be any function $X : \Omega \to \mathbb{R}$.

Assumptions to simplify the matters

The continuous case (case III) shall be the topic of this lecture.

For simplicity, we shall assume that

— the sample space $\Omega = \mathbb{R}$ is the set of the real numbers

— the event space $\mathcal{F}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$

— the random variable X is the identity function

$$X\colon \mathbb{R} \to \mathbb{R} \qquad X(x) = x \qquad \text{for every} \quad x \in \mathbb{R}$$

— there exists a probability density function

of the given probability P with respect to the Lebesgue measure λ on $(\mathbb{R}, \mathcal{F})$

Probability density function

Given the probability space $(\Omega, \mathcal{F}, P)$, the **probability density function** of the probability measure P with respect to a reference measure λ on $(\Omega, \mathcal{F})$ is a (measurable) function

$$f : \Omega \to \mathbb{R}$$

such that it holds

$$P(E) = \int_E f(\omega)\, \mathrm{d}\lambda \qquad \text{for every event} \quad E \in \mathcal{F}$$

(The integral on the right-hand side is the Lebesgue integral.)

Assumptions to simplify the matters

To sum up, we shall assume for simplicity in this lecture that

- the sample space $\Omega = \mathbb{R}$ is the set of the real numbers
- the event space $\mathcal{F}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$
- the random variable X is the identity function

$$X: \mathbb{R} \to \mathbb{R}$$

$$X: x \mapsto x \qquad \text{for every} \quad x \in \mathbb{R}$$

- and there exists a continuous probability density function $f: \mathbb{R} \to \mathbb{R}$ such that

$$P(E) = \int_E f(x)\,\mathrm{d}x \qquad \text{for every event} \quad E \in \mathcal{F}$$

Cumulative distribution function

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable.

Then the **cumulative distribution function** of the random variable X is the function

$$F : \mathbb{R} \to \mathbb{R}$$

defined by

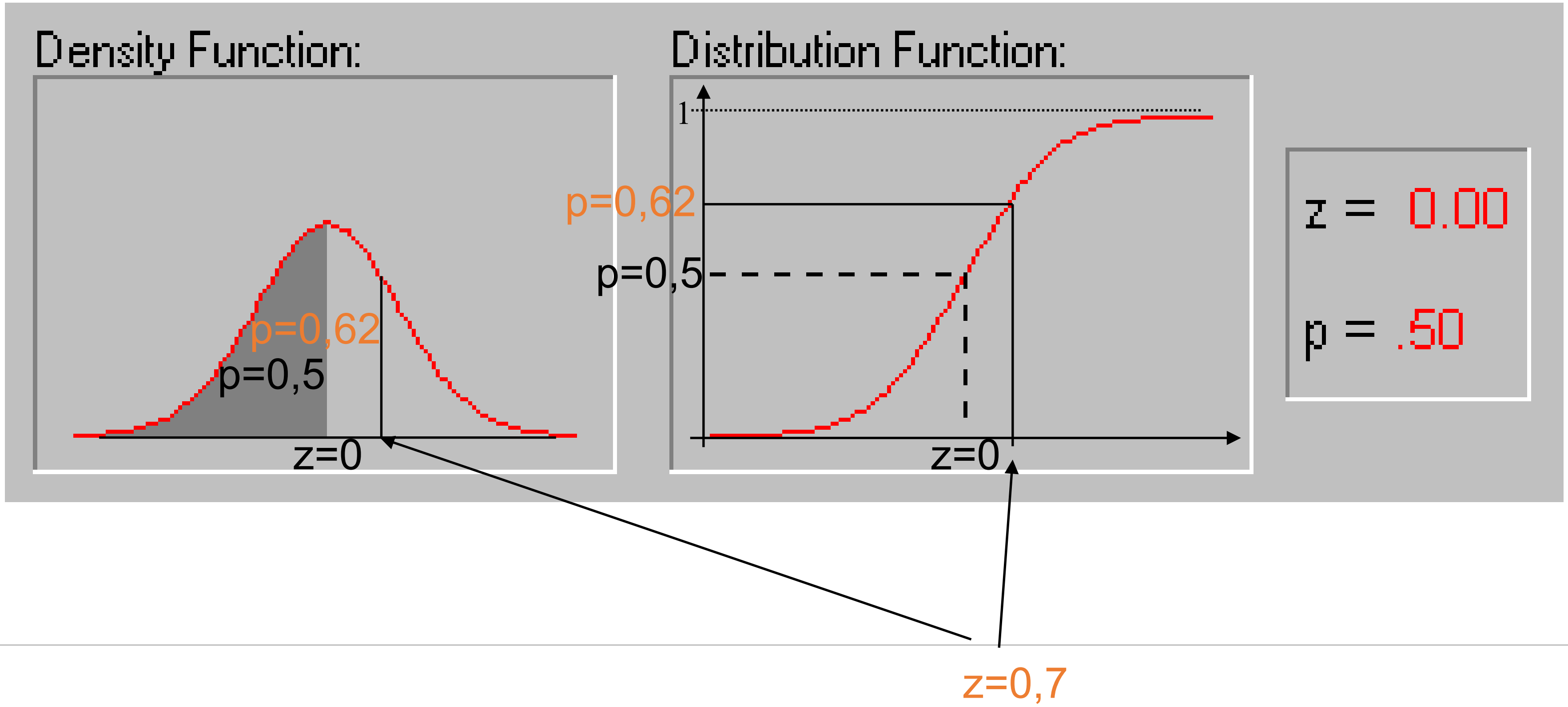
$$F(x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

Notice that the expression "$P(\{\omega \in \Omega : X(\omega) \leq x\})$" is often written as "$P(X \leq x)$" for short.

The density & the cumulative distribution function

It holds:

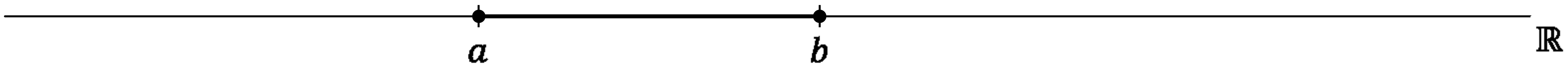
$$P(a < X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$



Continuous uniform distribution

Uniform distribution (continuous)

The continuous uniform distribution is another one which relates closely to the classical definition of probability: Considering a bounded interval $[a, b]$ to be the set of the outcomes of a random experiment, each point $x \in [a, b]$ of the interval is equally likely to occur.



Uniform distribution (continuous)

Consider a probability space $(\Omega, \mathcal{F}, P)$ where the sample space $\Omega = \mathbb{R}$, the event space $\mathcal{F} \subset 2^{\mathbb{R}}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$, and the probability P is given by its probability density function f which is

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \in \mathbb{R} \setminus [a,b] \end{cases}$$

where $a, b \in \mathbb{R}$, such that $a < b$, are given numbers.

Uniform distribution (continuous)

Consider the above probability space $(\Omega, \mathcal{F}, P)$ with the sample space $\Omega = \mathbb{R}$ and the probability P is given by its probability density function

$$f(x) = \frac{1}{b-a} \quad \text{for} \quad x \in [a,b] \qquad \text{and} \qquad f(x) = 0 \quad \text{otherwise}$$

Then the identity random variable $X: \mathbb{R} \to \mathbb{R}$

$$X(x) = x \qquad \text{for} \quad x \in \mathbb{R}$$

follows the continuous uniform distribution.

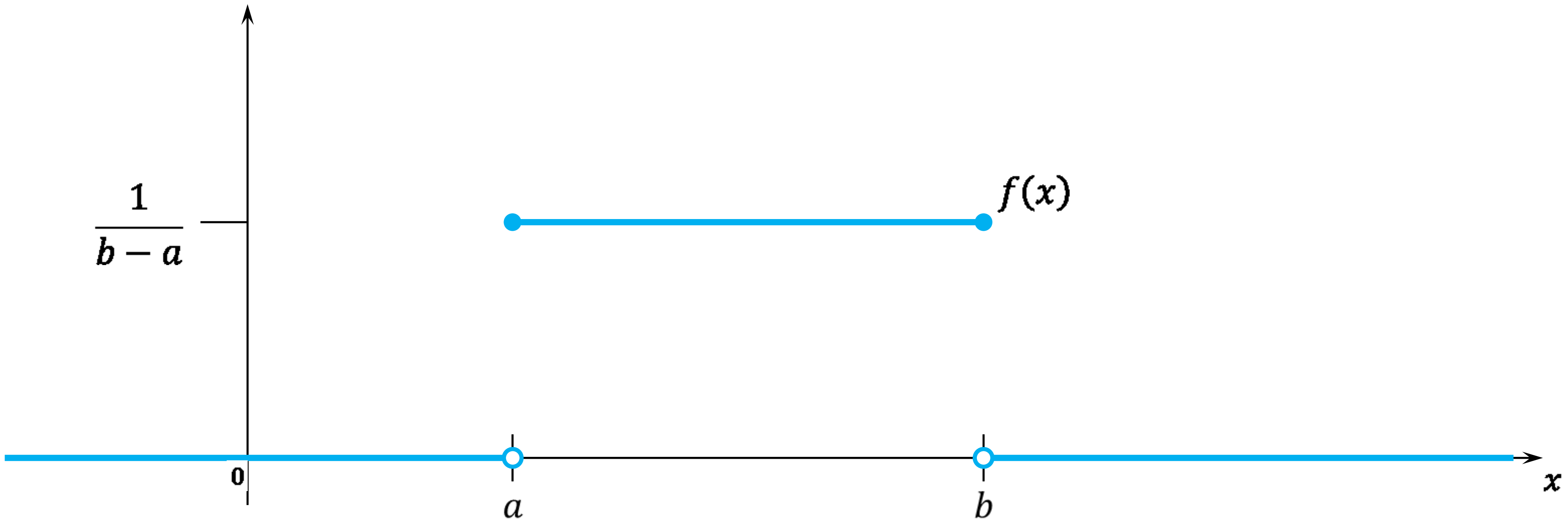
We then say that X is a **continuous uniform random variable** and write

$$X \sim \mathcal{U}(a,b)$$

Uniform distribution (continuous)

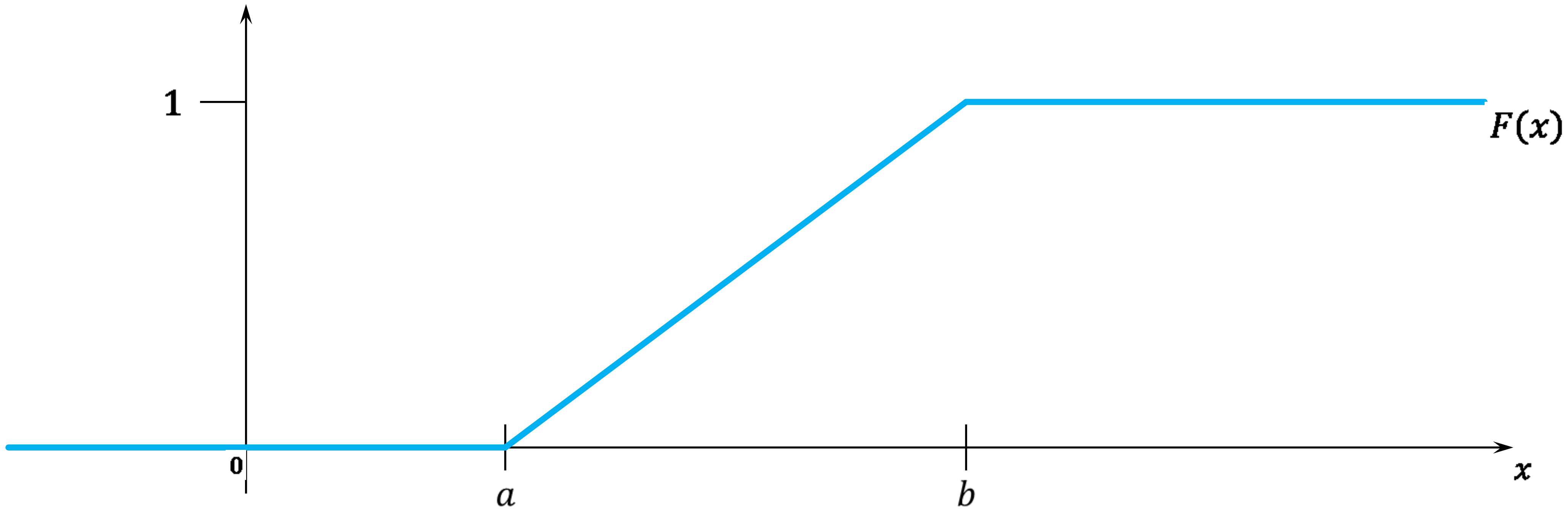
The **graph** of the **probability density function**

of a continuous uniform random variable $X \sim \mathcal{U}(a, b)$:



Uniform distribution (continuous)

The **graph** of the **cumulative distribution function** of a continuous uniform random variable $X \sim \mathcal{U}(a, b)$:



Uniform distribution (continuous)

Let the random variable $X \sim \mathcal{U}(a, b)$ follow the continuous uniform distribution.

<u>Calculate as an exercise:</u>

- Mean value: $\mu = \mathrm{E}[X] = \frac{a+b}{2}$
- Variance: $\sigma^2 = \mathrm{Var}(X) = \frac{(b-a)^2}{12}$
- Mode: $\hat{X}$ is any element $x \in [a, b]$
- Median: $\tilde{X} = \frac{a+b}{2}$

Uniform distribution (continuous) in Excel

In Excel, use the functions:

=RAND()	to get a uniformly distributed random number x $(0 \leq x < 1)$
=RANDBETWEEN()	to get a uniformly distributed <u>integer</u> random number n $(a \leq n \leq b)$

Remarks:

=RANDBETWEEN$(a;b) \in \{a, a+1, a+2, \dots, b-1, b\}$ where a, b are integer

=RAND$()*(b-a)+a$ yields a random number $x \in [a, b)$

=RAND$() \in [0,1)$

=INT(RAND$()*(b-a)+a+1)$ the same as =RANDBETWEEN$(a;b)$

Uniform distribution (continuous): Example

- The bus operates every 15 minutes. A passenger (who does not know the timetable of the bus) comes to the bus stop at a random time.

- The operator comes to the telephone set every 15 minutes. A customer makes a telephone call at a random time and waits as long as the operator comes.

Notice the difference:

— we have just one request here
— when applying the discrete Poisson distribution or the continuous exponential distribution, we have a series of requests coming at a constant rate

Gaussian normal distribution

- Normal distribution
- Normalized normal distribution
- Central Limit Theorem

Normal distribution

It has been observed in practice that the distribution of many real phenomena yields the typical "bell curve", i.e. the phenomena can be **approximated** by the classical **Gaussian normal distribution**. These phenomena include:

- the results of repeated measurements of lengths, distances, weights, etc.
- the results of various tests and examinations (exams)

We explain these practical observations by hypothesizing that many random quantities present in the experiment sum up together, cancel one another, and (by the Central Limit Theorem) yield the normal distribution approximately.

Normal distribution

Consider a probability space $(\Omega, \mathcal{F}, P)$ where the sample space $\Omega = \mathbb{R}$, the event space $\mathcal{F} \subset 2^{\mathbb{R}}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$, and the probability P is given by its probability density function f which is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for} \quad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$, so that $\sigma^2 > 0$, are given.

Normal distribution

Consider the above probability space $(\Omega, \mathcal{F}, P)$ with the sample space $\Omega = \mathbb{R}$ and the probability P is given by its probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for} \quad x \in \mathbb{R}$$

Then the identity random variable $X: \mathbb{R} \to \mathbb{R}$

$$X(x) = x \qquad \text{for} \quad x \in \mathbb{R}$$

follows the Gaussian normal distribution.

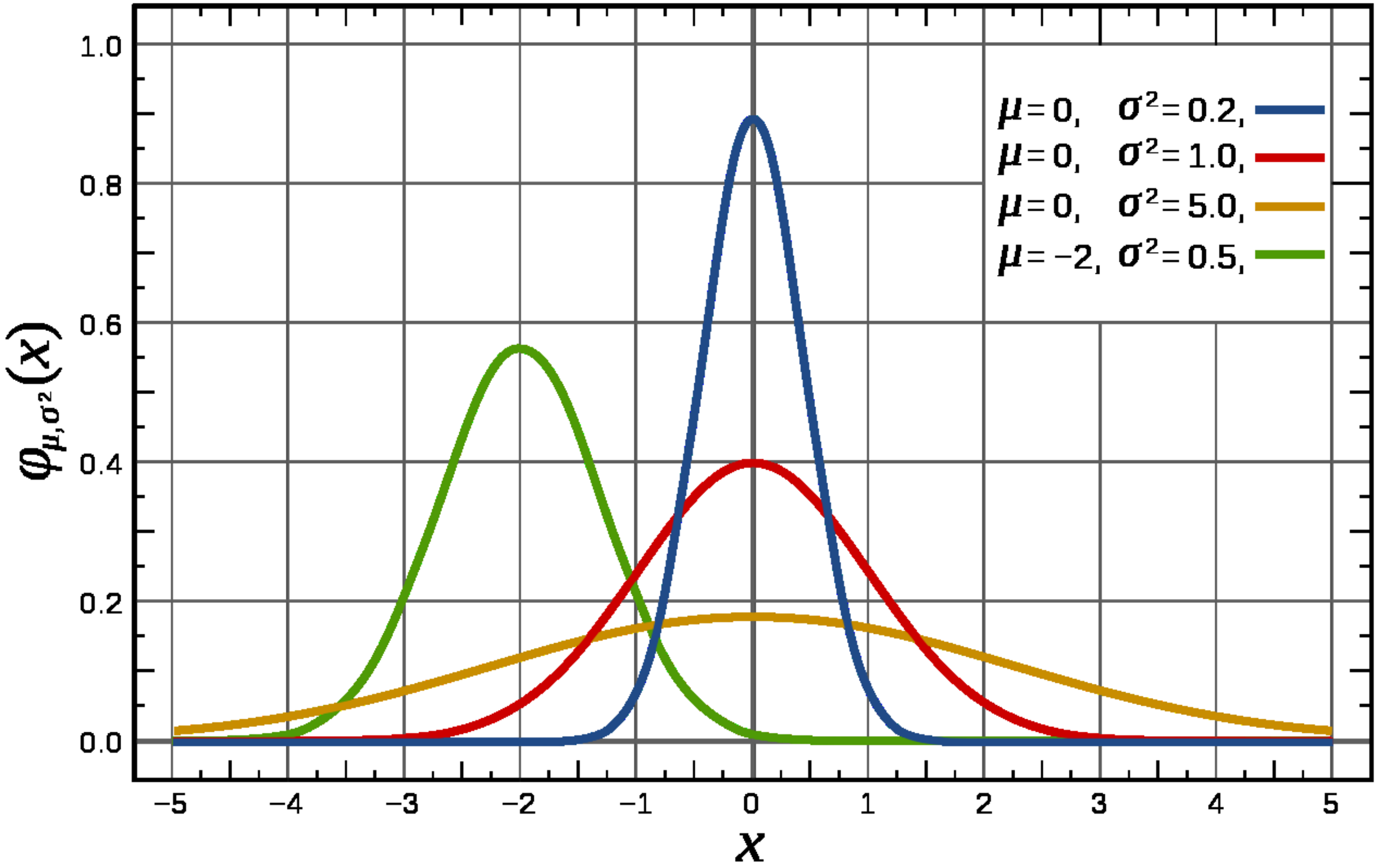
We then say that X is a **Gaussian normal random variable** and write

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

Normal distribution

The **graph** of the **probability density function**

of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ for various parameters:

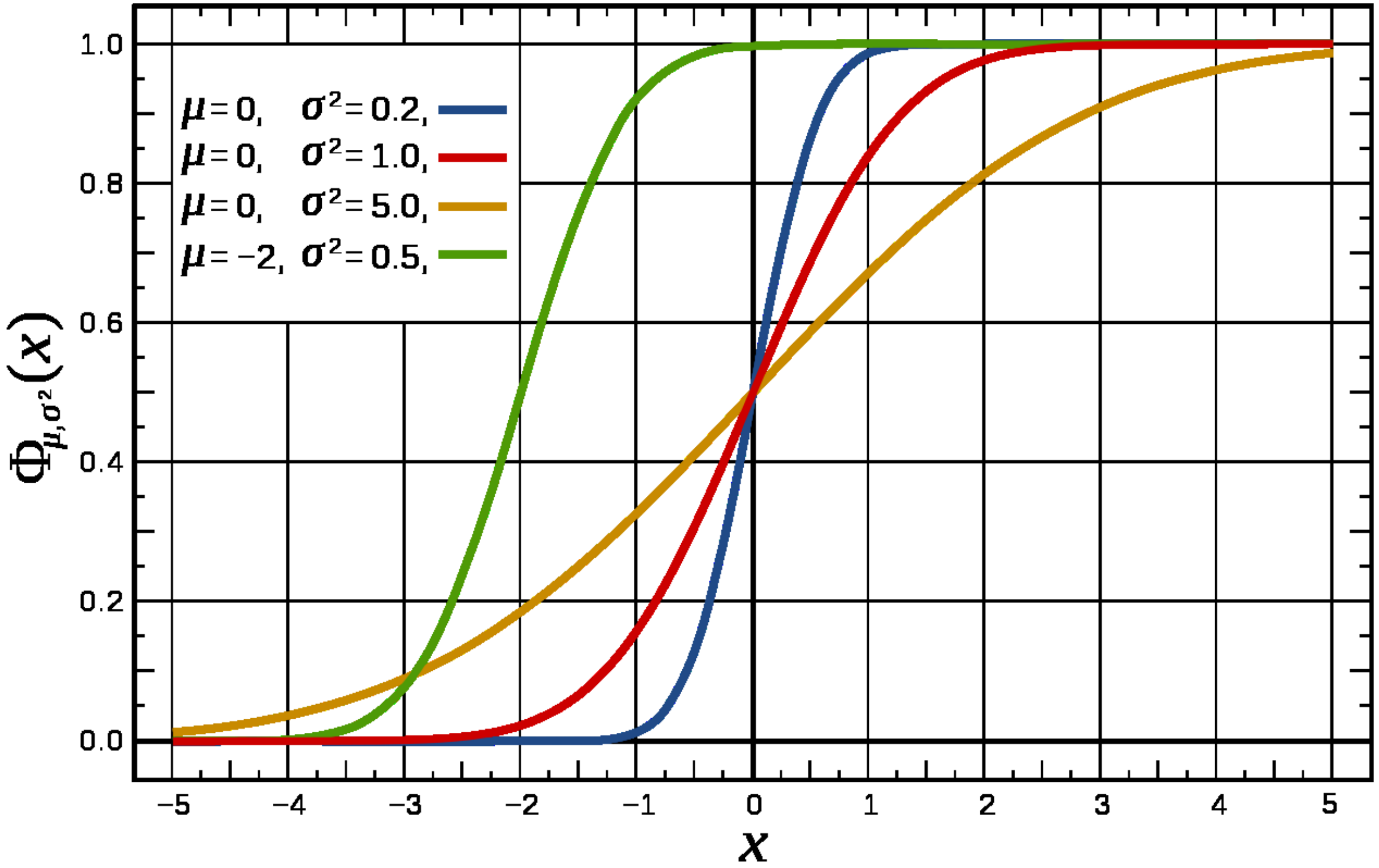


source: Wikipedia

Normal distribution

The **graph** of the **cumulative distribution function**

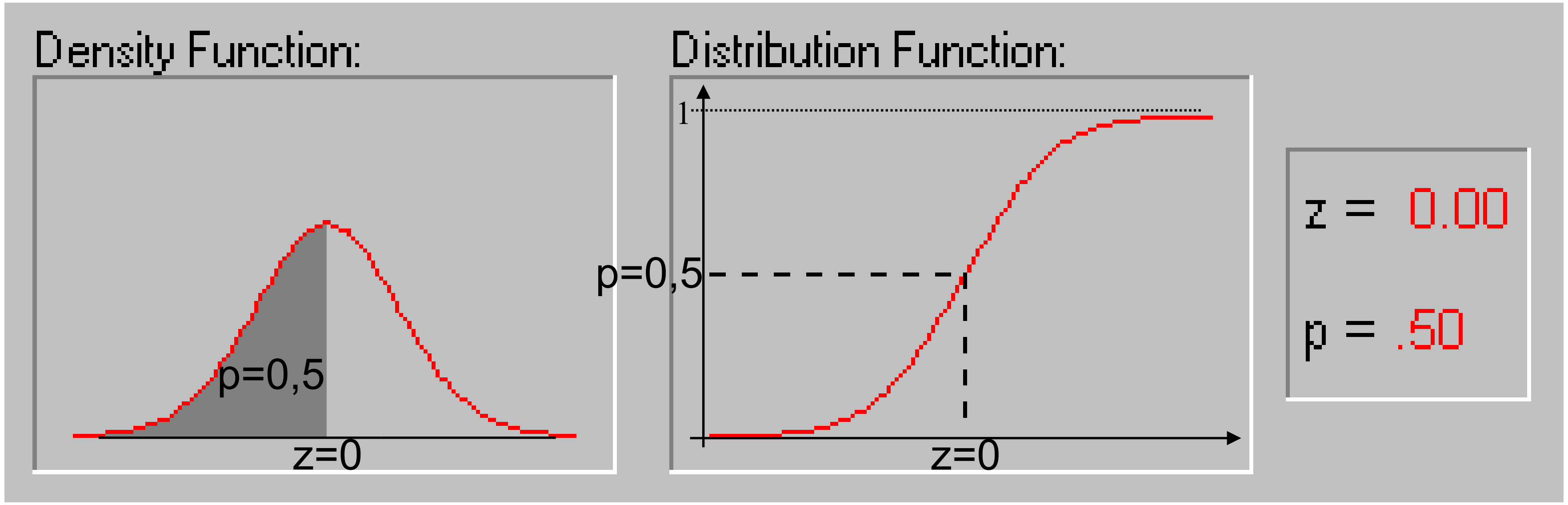
of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ for various parameters:



source: Wikipedia

Normal distribution

The graph of the probability density function & cumulative distribution function of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$:



Normal distribution

Let the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ follow the Gaussian normal distribution.

<u>Calculate as an exercise:</u>

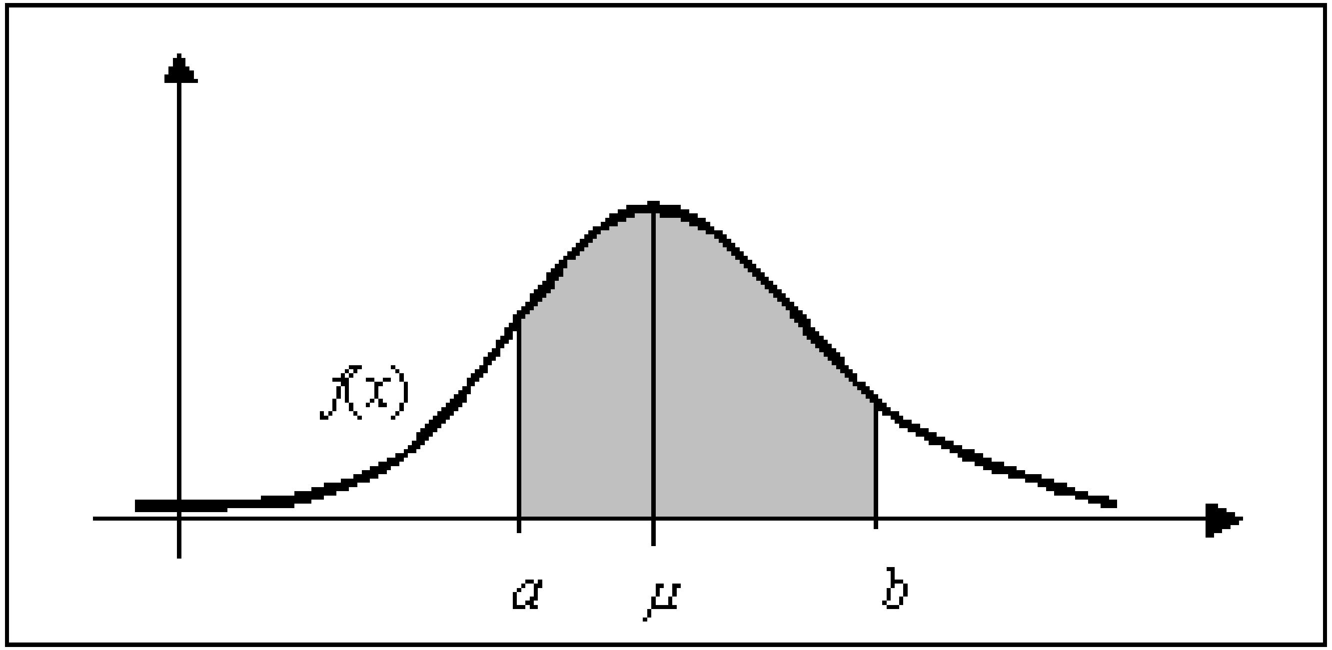
- Mean value: $\mu = \mathrm{E}[X] = \mu$
- Variance: $\sigma^2 = \mathrm{Var}(X) = \sigma^2$
- Mode: $\hat{X} = \mu$
- Median: $\tilde{X} = \mu$

Normal distribution

<u>**Recall:**</u> Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

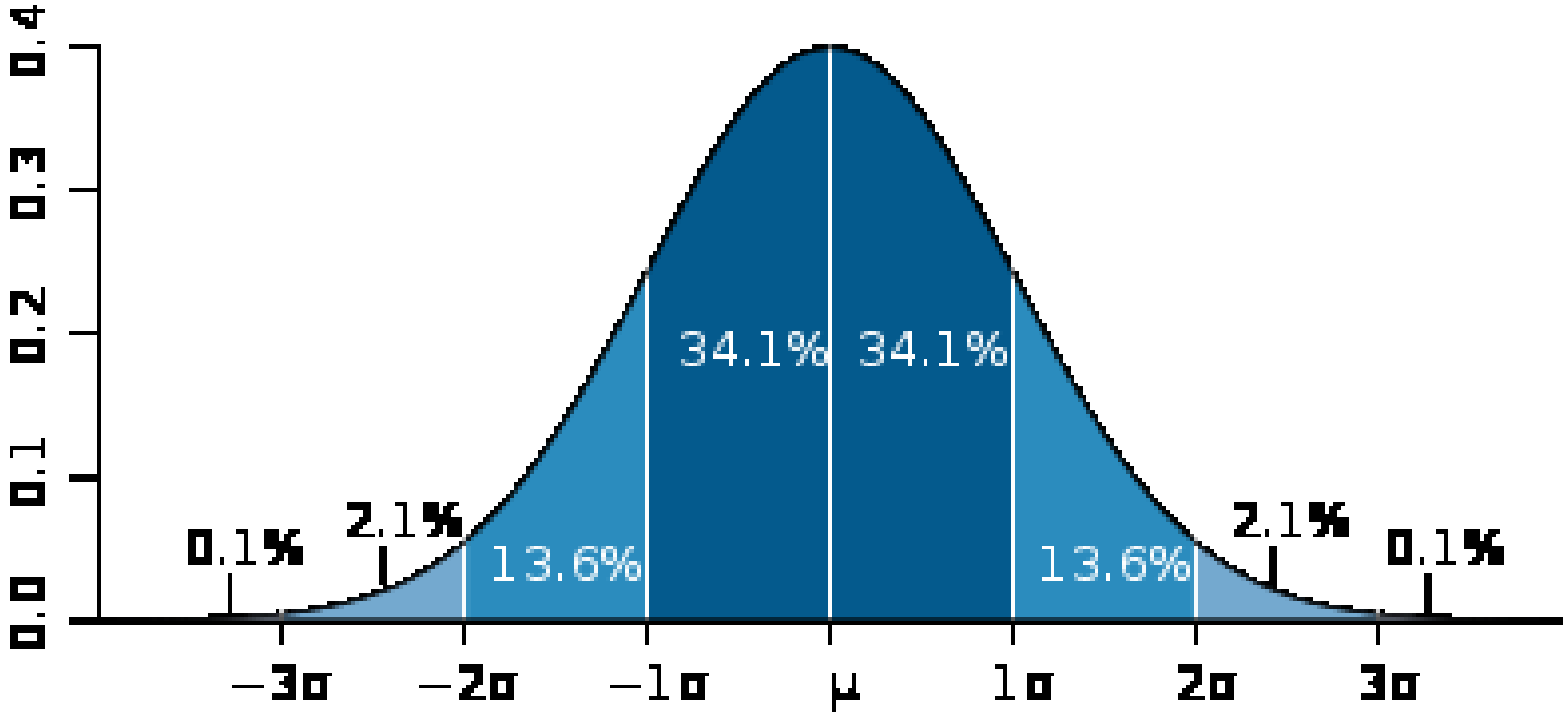
$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < x < b) =$$

$$= \int_a^b f(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x$$



Normal distribution: The 68–95–99.7 rule

The 68–95–99.7 rule / The three-sigma (3σ) rule



source: Wikipedia

Normalized normal distribution

The **normalized normal distribution** is

$$\mathcal{N}(0,1)$$

i.e. the normal distribution $\mathcal{N}(\mu,\sigma^2)$ with the parameters $\mu = 0$ and $\sigma^2 = 1$.

The **density** of the normalized normal distribution is denoted by $\varphi(x)$:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \qquad \text{for} \quad x \in \mathbb{R}$$

Normalized normal distribution

The **cumulative distribution function** of the normalized normal distribution is called the *Laplace function* and is denoted by $\Phi(x)$:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,\mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t \qquad \text{for} \quad x \in \mathbb{R}$$

Normalized normal distribution

Notice that, if X is a normally distributed random variable $(X \sim \mathcal{N}(\mu, \sigma^2))$, then

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

i.e. Z follows the normalized normal distribution. (Where $\sigma = \sqrt{\sigma^2}$.)

<u>Remark:</u> To properly understand, recall that the random variable is a measurable function $X: \Omega \to \mathbb{R}$. We consider $\Omega = \mathbb{R}$ here for simplicity. Hence, we have

$$Z(\omega) = \frac{X(\omega) - \mu}{\sigma} \qquad \text{for every} \quad \omega \in \Omega = \mathbb{R}$$

where $\omega \in \Omega = \mathbb{R}$ is the result of the random experiment.

Normalized normal distribution

Notice that, if X is a normally distributed random variable $(X \sim \mathcal{N}(\mu, \sigma^2))$, then

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

i.e. Z follows the normalized normal distribution. (Recall that $\sigma = \sqrt{\sigma^2}$.)

Notice that

$$P(-x < Z < x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} \mathrm{e}^{-\frac{t^2}{2}} \,\mathrm{d}t = \frac{2}{\sqrt{2\pi}} \int_{0}^{x} \mathrm{e}^{-\frac{t^2}{2}} \,\mathrm{d}t = 2\Phi^*(x) \qquad \text{for} \quad x \geq 0$$

(see below)

Normalized normal distribution

Recall the **cumulative distribution function** of the normalized normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,\mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t \qquad \text{for} \quad x \in \mathbb{R}$$

Defining the function

$$\Phi^*(x) = \int_{0}^{x} \varphi(t)\,\mathrm{d}t = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t \qquad \text{for} \quad x \geq 0$$

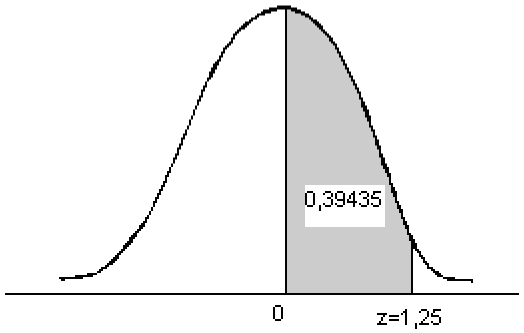
notice that

$$\Phi(x) - \Phi(-x) = 2\Phi^*(x) = 2\Phi(x) - 1 \qquad \text{for every} \quad x \geq 0$$

The values of the function $\Phi^*(x)$ can be found in statistical tables.

(Because the integral needs to be evaluated numerically.)

TABLE 1
The area under the curve of the normalized normal distribution $\mathcal{N}(0,1)$



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.10260	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.18824	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.36460	0.33891
1.0	0.34134	0.34375	0.34614	0.34850	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	**0.39435**	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42786	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408

Normal distribution in Excel

In Excel, use the functions:

=NORM.S.DIST$(x; \mathrm{TRUE})$	to get the value of the cumulative distribution function $\Phi(x)$ of the normalized normal distribution $(-\infty < x < +\infty)$
=NORM.S.DIST$(x; \mathrm{FALSE})$	to get the value of the probability density function $\varphi(x)$ of the normalized normal distribution
=NORM.S.INV(p)	to get the quantile $\Phi^{-1}(p)$ of the normalized normal distribution $(0 < p < 1)$

Normal distribution in Excel

In Excel, use the functions:

=NORM.DIST$(x; \mu; \sigma; \mathrm{TRUE})$ to get the value of the cumulative distribution function of the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ $(-\infty < x < +\infty)$

=NORM.DIST$(x; \mu; \sigma; \mathrm{FALSE})$ to get the value of the probability density function of the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ $(-\infty < x < +\infty)$

=NORM.INV$(p; \mu; \sigma)$ to get the quantile of the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ $(0 < p < 1)$

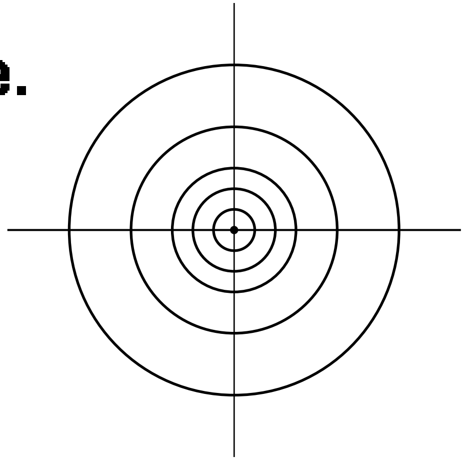
Normal distribution in Excel

In Excel, use the functions:

=NORMSDIST(x)	the same as =NORM.S.DIST(x; TRUE), deprecated
=NORMSINV()	the same as =NORM.S.INV(), deprecated
=NORMDIST()	the same as =NORM.DIST(), deprecated
=NORMINV()	the same as =NORM.INV(), deprecated

Normal distribution: Example

- An archer shoots an arrow against the plane with the intention to hit the origin. The sample space $\Omega = \mathbb{R}^2 = \{[x, y] : x, y \in \mathbb{R}\}$ is the set of all the points of the plane. The random variable X is the x-coordinate of the hit, i.e. $X(\omega) = X([x, y]) = x$.



- If we considered the random variable as both coordinates of the hit, i.e. $X(\omega) = X([x, y]) = \omega = [x, y]$, then the resulting distribution would be a bivariate normal distribution.

Central Limit Theorem (CLT)

There are several versions or variants of the Central Limit Theorem.

Its earlies version is now known as the de Moivre-Laplace Theorem. It states that the normal distribution is an approximation of the discrete binomial distribution.

We shall then mention the Lindeberg-Lévy Theorem, which is a comprehensible variant of the Central Limit Theorem.

CLT: de Moivre-Laplace Theorem (local form)

Considering the binomial distribution, choose the probability $p \in (0,1)$ and put $q = 1 - p$. (So that $p + q = 1$ and $p, q > 0$.) Then,

for any natural number $n \in \mathbb{N}$ and for any natural number $k \in \mathbb{N}$,

if

$$k \approx np$$

then

$$\binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi\left(\sqrt{npq}\right)^2}} \mathrm{e}^{-\frac{(k-np)^2}{2(\sqrt{npq})^2}}$$

as $n \to \infty$.

CLT: de Moivre-Laplace Theorem (integral form)

Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables, each of which represents the result of a Bernoulli Trial (independent of the other trials), attaining the value $X_i = 1$ (success) with a fixed probability $p \in (0,1)$ and attaining the value $X_i = 0$ (failure) with the fixed probability $q = 1 - p$. (So that $p + q = 1$ and $p, q > 0$.) It then holds for every $y \in \mathbb{R}$

$$\frac{\sum_{k=0}^{\lfloor x_n \rfloor} \binom{n}{k} p^k q^{n-k} - np}{\sqrt{npq}} \longrightarrow \underbrace{\int_{-\infty}^{x_n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \, \mathrm{d}t}_{\Phi(x_n)} \qquad \text{as} \quad n \to \infty$$

where $x_n = np + y\sqrt{npq}$.

CLT: Lindeberg-Lévy Theorem

Let $X_1, X_2, X_3, \ldots$ be a sequence of **independent & identically distributed** random variables with finite expected value $\mathrm{E}[X_i] = \mu$ and with finite variance $\mathrm{Var}(X_i) = \sigma^2$. Then

$$P\left(\left\{\omega \in \Omega^n : \sqrt{n}\left(\frac{X_1(\omega_1) + \cdots + X_n(\omega_n)}{n} - \mu\right) \leq x\right\}\right) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{t^2}{2\sigma^2}} \, \mathrm{d}t$$

as $n \to \infty$

for every $x \in \mathbb{R}$.

Exponential distribution

Exponential distribution

There are some events, such as

— customers coming to a shop during one hour (between 10:00 and 11:00, say)

— telephone calls incoming during one hour (between 10:00 and 11:00, say)

— requests incoming to a server during one minute (between 10:00 and 10:01)

— meteorites of diameter ≥ 1 meter hitting the Earth during a year

— decay events from a radioactive source

that (as we suppose) have some properties in common.

Exponential distribution

Suppose that a random event occurs repeatedly and satisfies the following assumptions:

- the event can occur at any time
- the average number of occurrences of the event during an interval of time of a fixed length is constant; the number does not depend on the beginning of the interval, and does not depend on the number of occurrences of the event before the beginning of the time interval
- the average number of occurrences of the event during an interval of time is proportional to the length of the interval
- …

Exponential distribution

Suppose that a random event occurs repeatedly and satisfies the following assumptions:

- …
- if the length of the interval is very small, then there is no more than one occurrence of the event in the interval;

  in other words, denoting by $p_t^{\geq 2}$ the probability that the event occurs at least two times during a time interval of length $t > 0$, it holds $p_t^{\geq 2} \times t \longrightarrow 0$ as $t \to 0$

Exponential distribution

We know already that, under the above assumptions, the probability that the event occurs exactly k times during the time interval of a given length $t > 0$ is

$$\frac{\lambda^k}{k!} e^{-\lambda}$$

where λ is the expected number of events during an interval of the given length t.

¿ What is the **probability distribution of the times** between the events ?

Exponential distribution

Consider a probability space $(\Omega, \mathcal{F}, P)$ where the sample space $\Omega = \mathbb{R}$, the event space $\mathcal{F} \subset 2^{\mathbb{R}}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$, and the probability P is given by its probability density function f which is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \in [0, +\infty) \\ 0, & x \in (-\infty, 0) \end{cases}$$

where $\lambda \in \mathbb{R}^{+}$, so that $\lambda > 0$, is a given number.

Exponential distribution

Consider the above probability space $(\Omega, \mathcal{F}, P)$ with the sample space $\Omega = \mathbb{R}$ and the probability P is given by its probability density function

$$f(x) = \lambda \mathrm{e}^{-\lambda x} \quad \text{for} \quad x \in [0, +\infty) \qquad \text{and} \qquad f(x) = 0 \quad \text{otherwise}$$

Then the identity random variable $X: \mathbb{R} \to \mathbb{R}$

$$X(x) = x \qquad \text{for} \quad x \in \mathbb{R}$$

follows the exponential distribution.

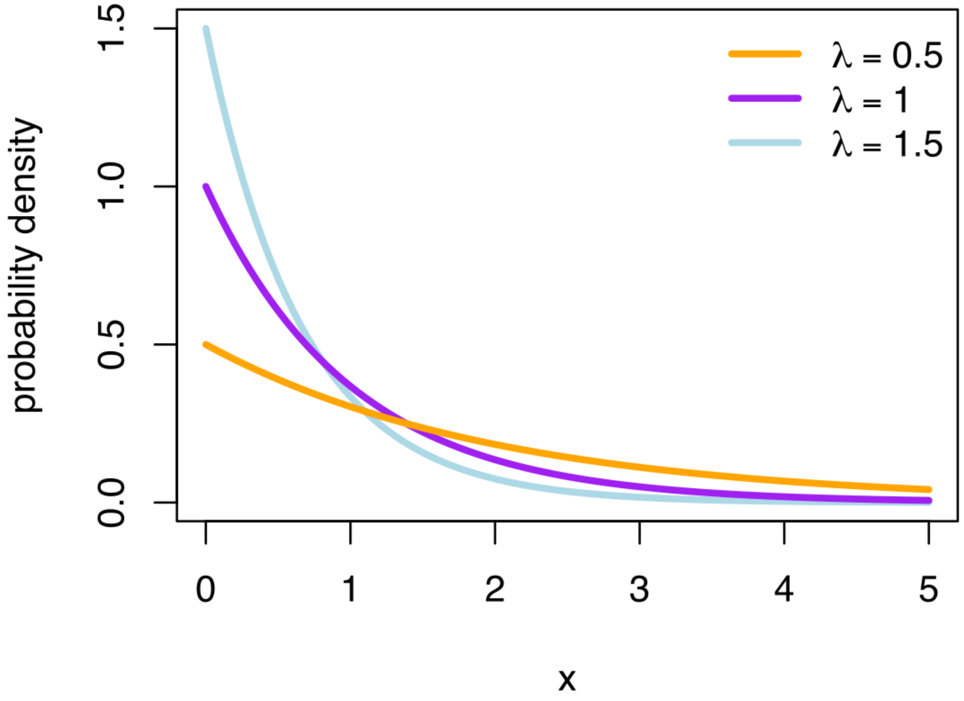
We then say that X is an **exponential random variable** and write

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

Exponential distribution

The **graph** of the **probability density function** of

an exponential random variable $X \sim \mathrm{Exp}(\lambda)$ for various values of the parameter:

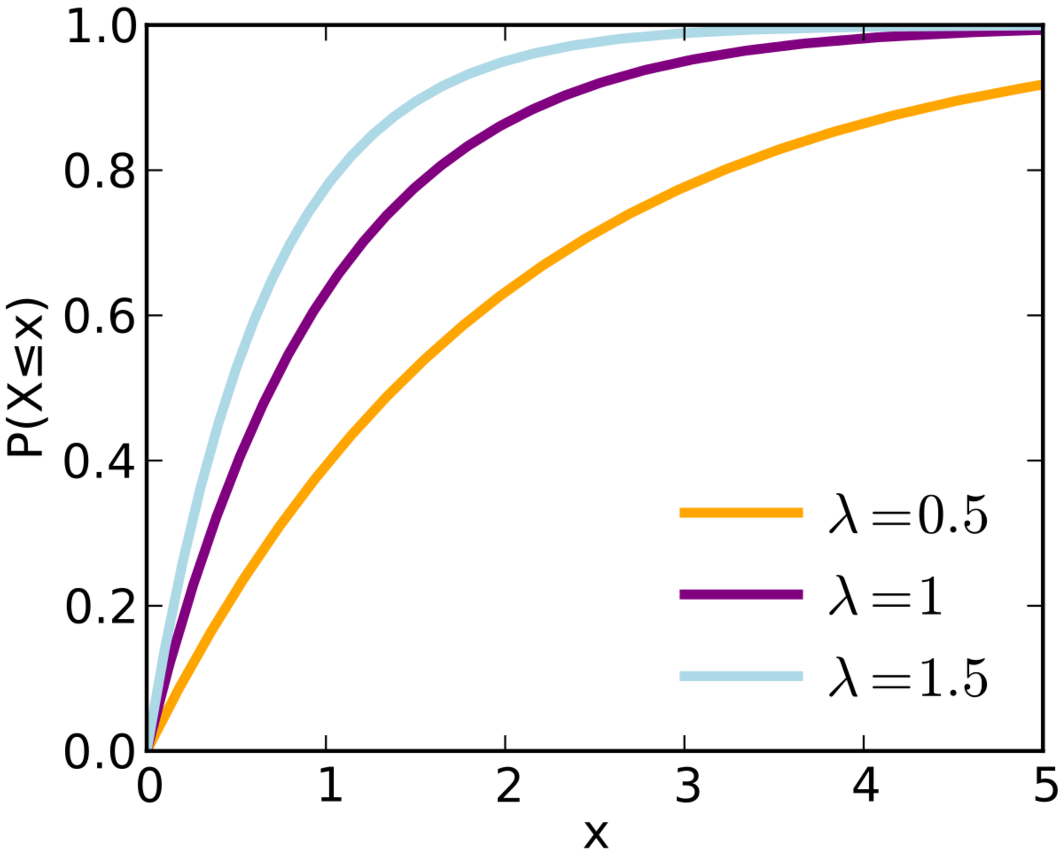


source: Wikipedia

Exponential distribution

The **graph** of the **cumulative distribution function** of

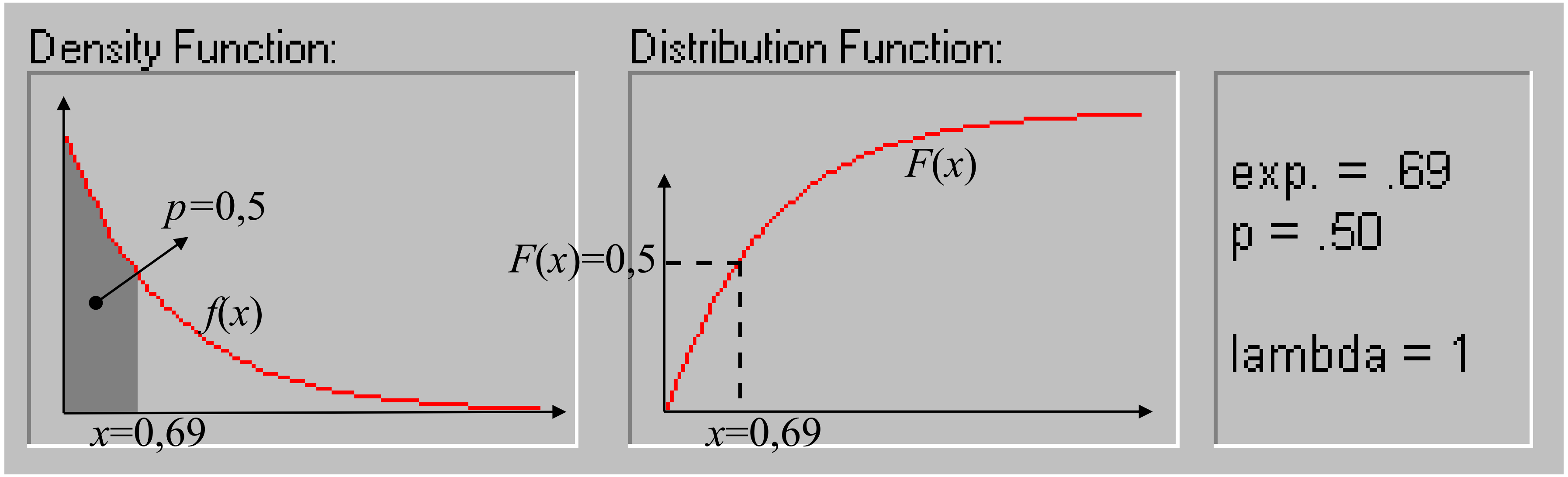
an exponential random variable $X \sim \mathrm{Exp}(\lambda)$ for various values of the parameter:



source: Wikipedia

Exponential distribution

The graph of the probability density function & cumulative distribution function of an exponential random variable $X \sim \text{Exp}(\lambda)$ with $\lambda = 1$:



Exponential distribution

Let the random variable $X \sim \mathrm{Exp}(\lambda)$ follow the exponential distribution.

<u>Calculate as an exercise:</u>

- Mean value: $\mu = \mathrm{E}[X] = \frac{1}{\lambda}$
- Variance: $\sigma^2 = \mathrm{Var}(X) = \frac{1}{\lambda^2}$
- Mode: $\hat{X} = 0$
- Median: $\tilde{X} = \frac{\ln 2}{\lambda}$

Exponential distribution: Examples

- The time till the next telephone call.
- The time until a radioactive particle decays.
- The time between clicks of a Geiger-Müller counter.
- The time until the next default in risk modelling.
- The time till the next failure / accident / …

Some continuous probability distributions derived from the normal distribution

- Pearson's χ^2 distribution
- Student's t distribution
- Fisher-Snedecor F distribution

Pearson's χ^2 distribution

Let

$$Z_1, Z_2, \ldots, Z_k \sim \mathcal{N}(0,1)$$

be <u>independent</u> & <u>normalized normal</u> random variables. Then the random variable

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

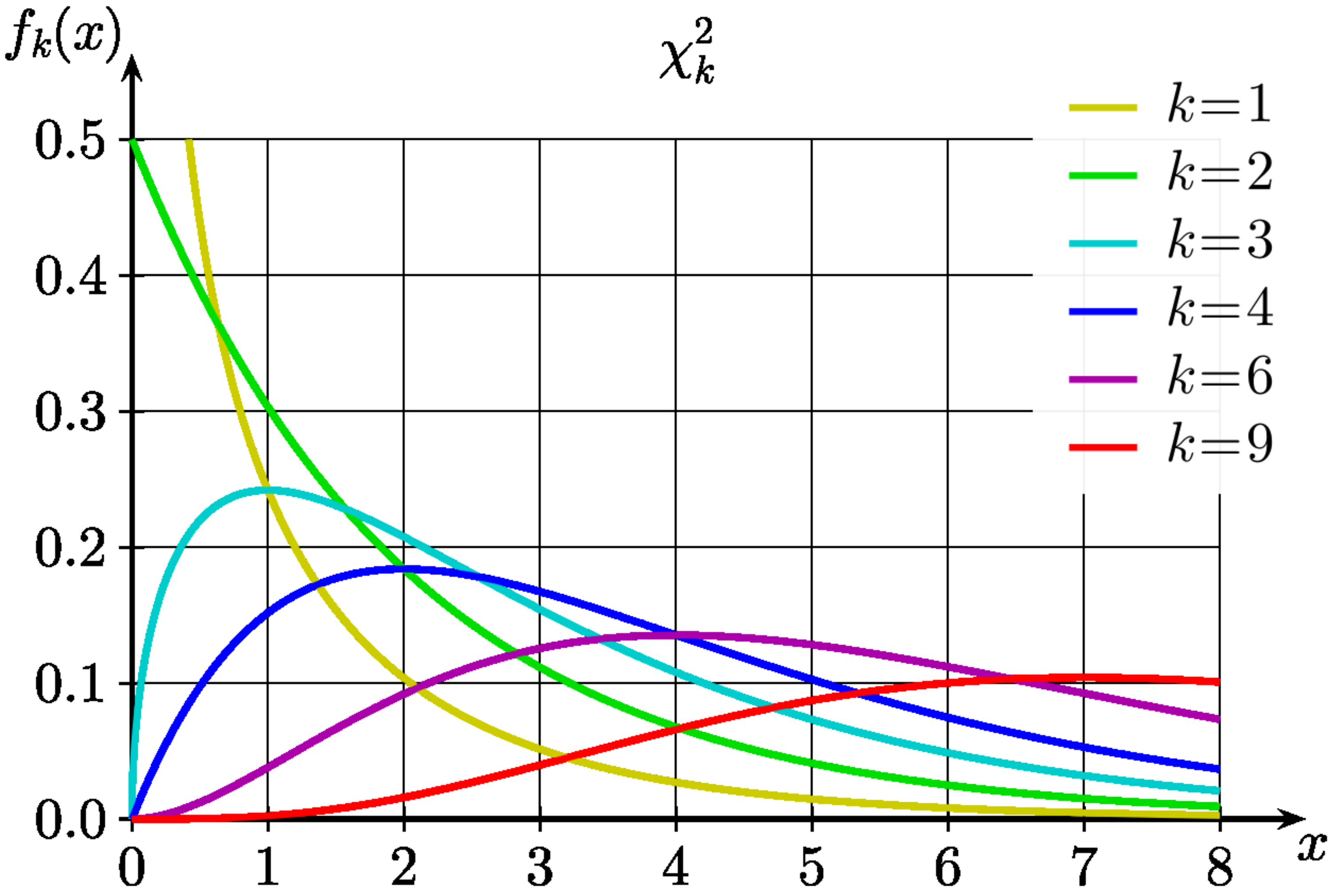
i.e. the sum of the squares of the random variables,

follows the **chi-squared distribution** with k degrees of freedom and we write

$$X \sim \chi^2(k) \qquad \text{or} \qquad X \sim \chi_k^2$$

chi-squared distribution

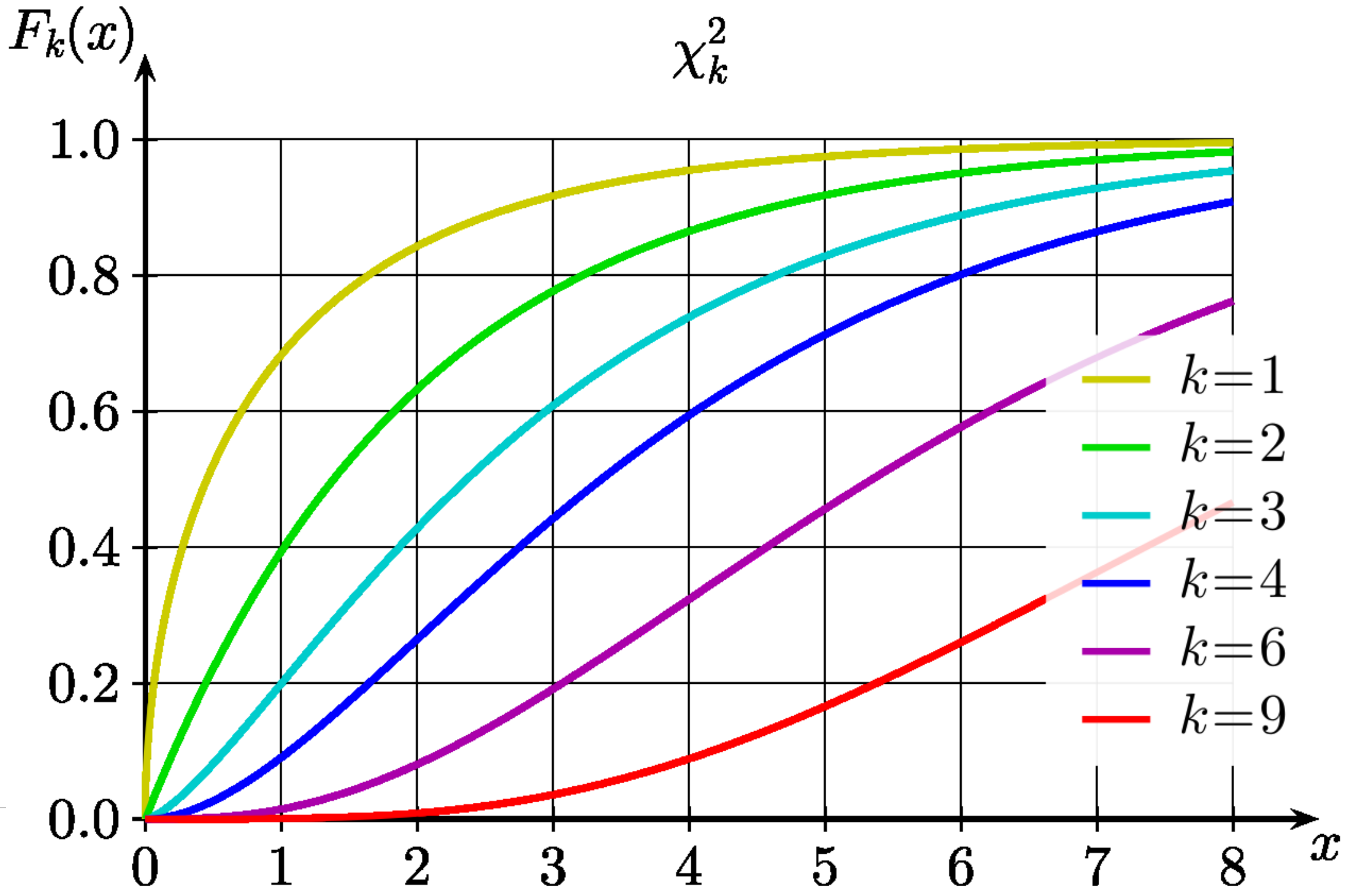
The graph of the probability density function of an $X \sim \chi_k^2$ for various k:



source: Wikipedia

chi-squared distribution

The **graph** of the **cumulative distribution function** of an $X \sim \chi^2_k$ for various k:



source: Wikipedia

chi-squared distribution

The probability density function:

$$f(x) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{\Gamma\left(\frac{k}{2}\right) 2^{\frac{k}{2}}}, & x \in (0, +\infty) \\ 0, & x \in (-\infty, 0] \end{cases}$$

Mean value: $\mu = \mathrm{E}[X] = k$

Mode: $\hat{X} = \max(k-2, 0)$

Variance: $\sigma^2 = \mathrm{Var}(x) = 2k$

The gamma function

$$\Gamma(z) = \int_0^{+\infty} x^z \mathrm{e}^{-x}\, \mathrm{d}x \qquad \text{for} \quad z \in \mathbb{C} \;\text{ such that }\; \mathrm{Re}(z) > 0$$

It is easy to calculate:

$$\Gamma(1) = 1$$

$$\Gamma(z+1) = z\Gamma(z)$$

Therefore:

$$\Gamma(n+1) = n! \qquad \text{for} \quad n = 0, 1, 2, 3, \ldots$$

The gamma function – another definition (due to Euler)

$$\Gamma(z) = \frac{1}{z}\prod_{n=1}^{\infty}\frac{\left(1+\frac{1}{n}\right)^{z}}{1+\frac{z}{n}} \qquad \text{for} \qquad z \in \mathbb{C} \setminus \{0, -1, -2, -3, \dots\}$$

chi-squared distribution in Excel

In Excel, use the functions:

=CHISQ.DIST$(x; k; \text{TRUE})$	to get the value of the cumulative distribution function of the random variable $X \sim \chi_k^2$ $(-\infty < x < +\infty)$
=CHISQ.DIST$(x; k; \text{FALSE})$	to get the value of the probability density function of the random variable $X \sim \chi_k^2$ $(-\infty < x < +\infty)$
=CHISQ.INV$(p; k)$	to get the quantile of the random variable $X \sim \chi_k^2$ $(0 < p < 1)$

chi-squared distribution in Excel

In Excel, use the functions:

=CHISQ.DIST.RT$(x; k)$ the same as =1-CHISQ.DIST$(x; k; \mathrm{TRUE})$

=CHISQ.INV.RT$(p; k)$ the same as =CHISQ.INV$(1 - p; k)$

That is, these functions calculate the cumulative distribution function from the right.

This is useful when one needs to calculate the <u>critical value</u> of some χ^2-test.

chi-squared distribution in Excel

In Excel, use the functions:

=CHIDIST()	the same as =CHISQ.DIST.RT(), deprecated
=CHIINV()	the same as =CHISQ.INV.RT, deprecated

Student's t distribution

Let

$$Z, Z_1, Z_2, \ldots, Z_k \sim \mathcal{N}(0,1)$$

be independent & normalized normal random variables. Then the random variable

$$X = \frac{Z}{\sqrt{\frac{Z_1^2 + Z_2^2 + \cdots + Z_k^2}{k}}} = \frac{Z}{\sqrt{Z_1^2 + Z_2^2 + \cdots + Z_k^2}} \sqrt{k}$$

follows the ***t*-distribution** with k degrees of freedom and we write

$$X \sim t(k) \qquad \text{or} \qquad X \sim t_k$$

Student's t distribution

Equivalently, let

$$Z \sim \mathcal{N}(0,1) \qquad \text{and} \qquad Y \sim \chi_k^2$$

be <u>independent</u> random variables. Then the random variable

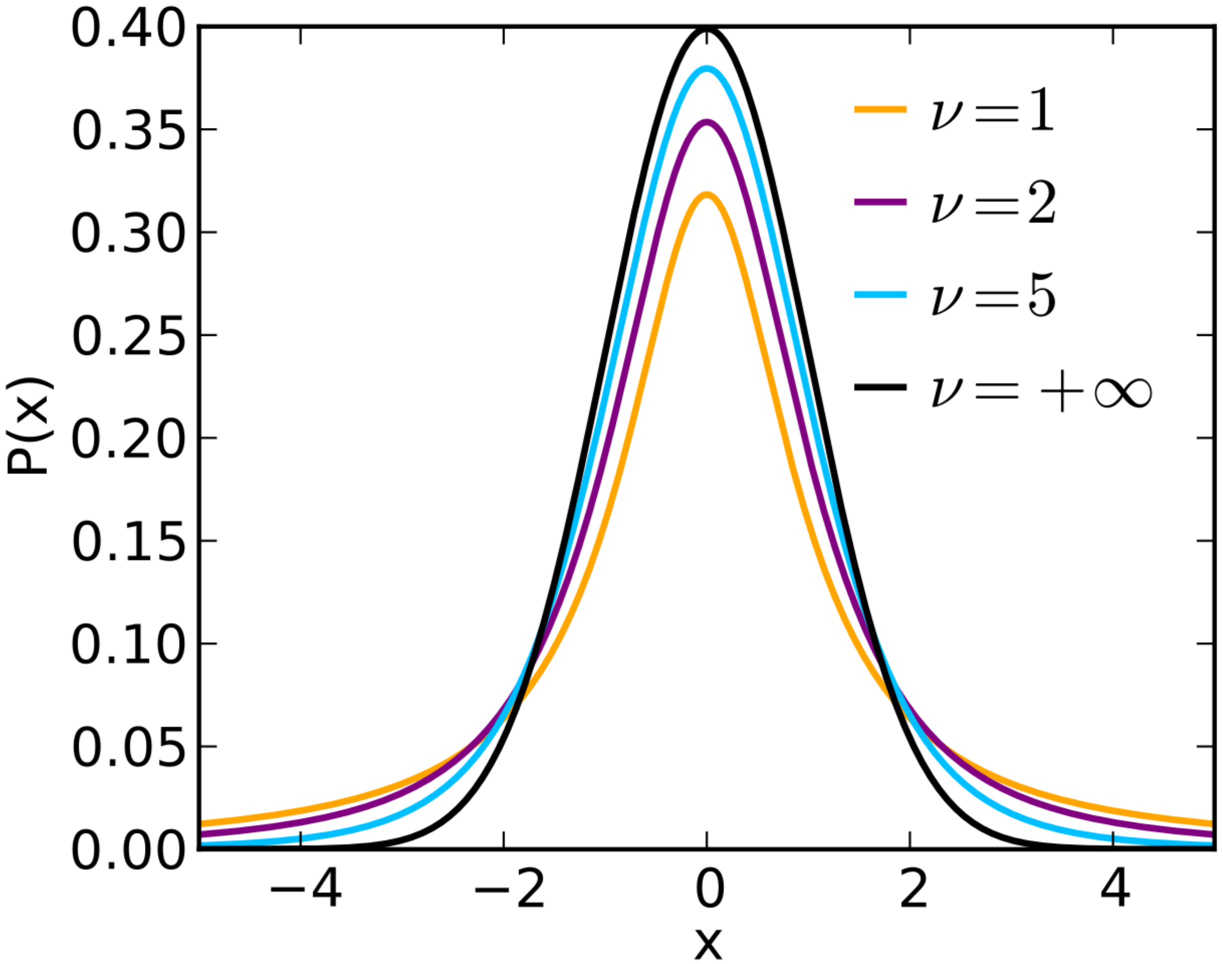
$$X = \frac{Z}{\sqrt{Y/k}} = \frac{Z}{\sqrt{Y}}\sqrt{k}$$

follows the ***t*-distribution** with k degrees of freedom and we write

$$X \sim t(k) \qquad \text{or} \qquad X \sim t_k$$

t-distribution

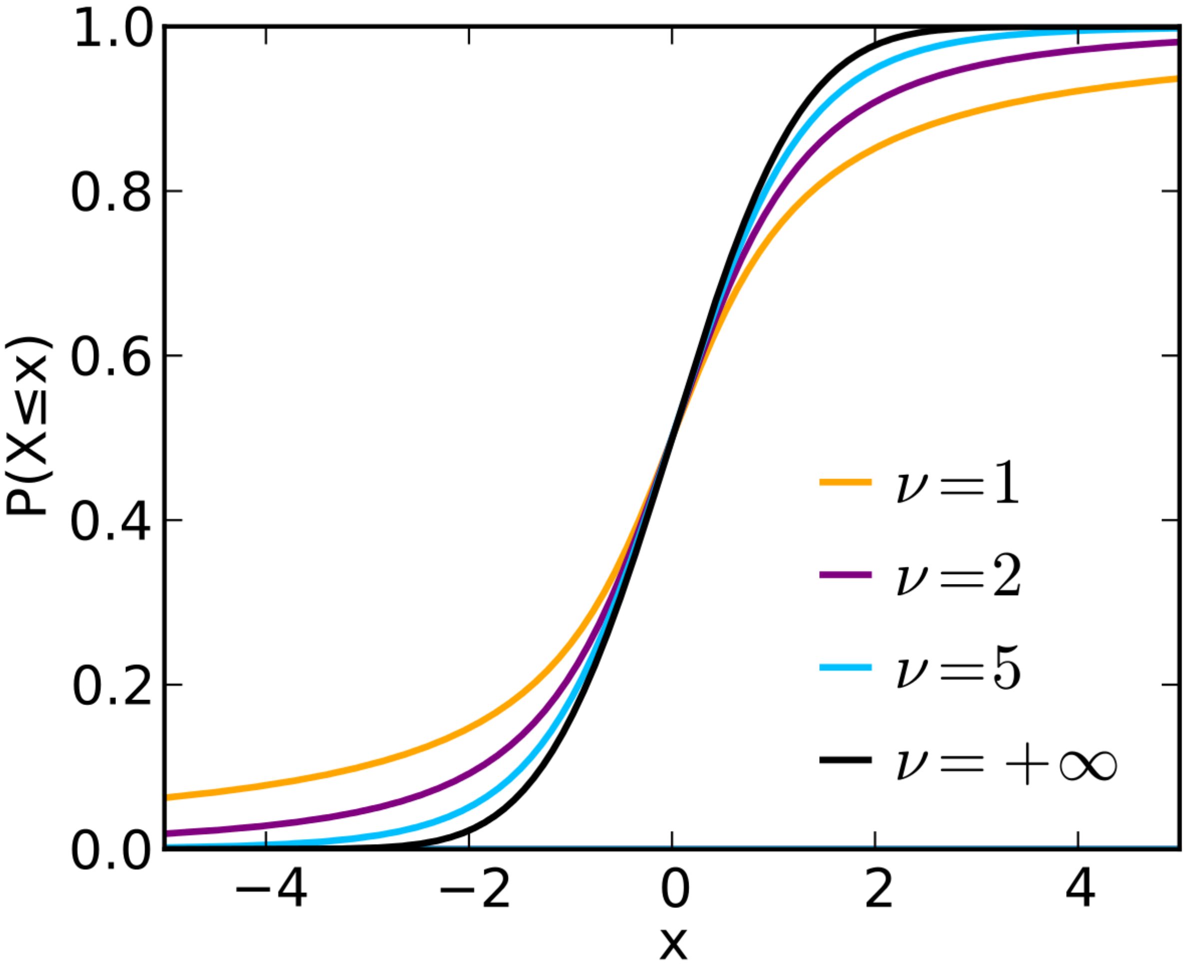
The **graph** of the **probability density function** of an $X \sim t_k$ for various k:



source: Wikipedia

t-distribution

The graph of the cumulative distribution function of an $X \sim t_k$ for various k:



source: Wikipedia

t-distribution

The probability density function:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}} \qquad \text{for} \quad x \in \mathbb{R}$$

Mean value: $\mu = \mathrm{E}[X] = \begin{cases} 0, & k = 2, 3, 4, \dots \\ \text{not exists}, & k = 1 \end{cases}$

Mode: $\hat{X} = 0$

Variance: $\sigma^2 = \mathrm{Var}(x) = \begin{cases} \frac{k}{k-2}, & k = 3, 4, 5, \dots \\ +\infty, & k = 2 \\ \text{not exists}, & k = 1 \end{cases}$

Median: $\tilde{X} = 0$

t-distribution in Excel

In Excel, use the functions:

=T.DIST$(x; k; \mathrm{TRUE})$	to get the value of the cumulative distribution function of the random variable $X \sim t_k$ $(-\infty < x < +\infty)$
=T.DIST$(x; k; \mathrm{FALSE})$	to get the value of the probability density function of the random variable $X \sim t_k$ $(-\infty < x < +\infty)$
=T.INV$(p; k)$	to get the quantile of the random variable $X \sim t_k$ $(0 < p < 1)$

t-distribution in Excel

In Excel, use the functions:

=T.DIST.RT$(x; k)$	the same as =1-T.DIST$(x; k; \text{TRUE})$
=T.DIST.2T$(x; k)$	the same as =2*T.DIST.RT$(x; k)$
=T.INV.2T$(p; k)$	the same as =T.INV$(1 - p/2\,; k)$

That is, the functions calculate the cumulative distribution function from the right.

This is useful when one needs to calculate the <u>critical value</u> of some t-test.

t-distribution in Excel

In Excel, use the functions:

=TDIST($x; k;$ tails) tails $= 1$:

the same as =T.DIST.RT($x; k$), deprecated

tails $= 2$:

the same as =T.DIST.2T($x; k$), deprecated

=TINV() the same as =T.INV.2T, deprecated

Fisher-Snedecor F distribution

Let

$$Z_1', Z_2', \dots, Z_{k_1}' \sim \mathcal{N}(0,1) \qquad \text{and} \qquad Z_1'', Z_2'', \dots, Z_{k_2}'' \sim \mathcal{N}(0,1)$$

be independent & normalized normal random variables. Then the random variable

$$X = \frac{\dfrac{{Z_1'}^2 + {Z_2'}^2 + \dots + {Z_{k_1}'}^2}{k_1}}{\dfrac{{Z_1''}^2 + {Z_2''}^2 + \dots + {Z_{k_2}''}^2}{k_2}} = \frac{{Z_1'}^2 + {Z_2'}^2 + \dots + {Z_{k_1}'}^2}{{Z_1''}^2 + {Z_2''}^2 + \dots + {Z_{k_2}''}^2} \Big/ \frac{k_1}{k_2}$$

follows the ***F*-distribution** with k_1 and k_2 degrees of freedom and we write

$$X \sim F(k_1, k_2) \qquad \text{or} \qquad X \sim F_{k_1,k_2}$$

Fisher-Snedecor F distribution

Equivalently, let

$$Y_1 \sim \chi^2_{k_1} \qquad \text{and} \qquad Y_2 \sim \chi^2_{k_2}$$

be independent random variables. Then the random variable

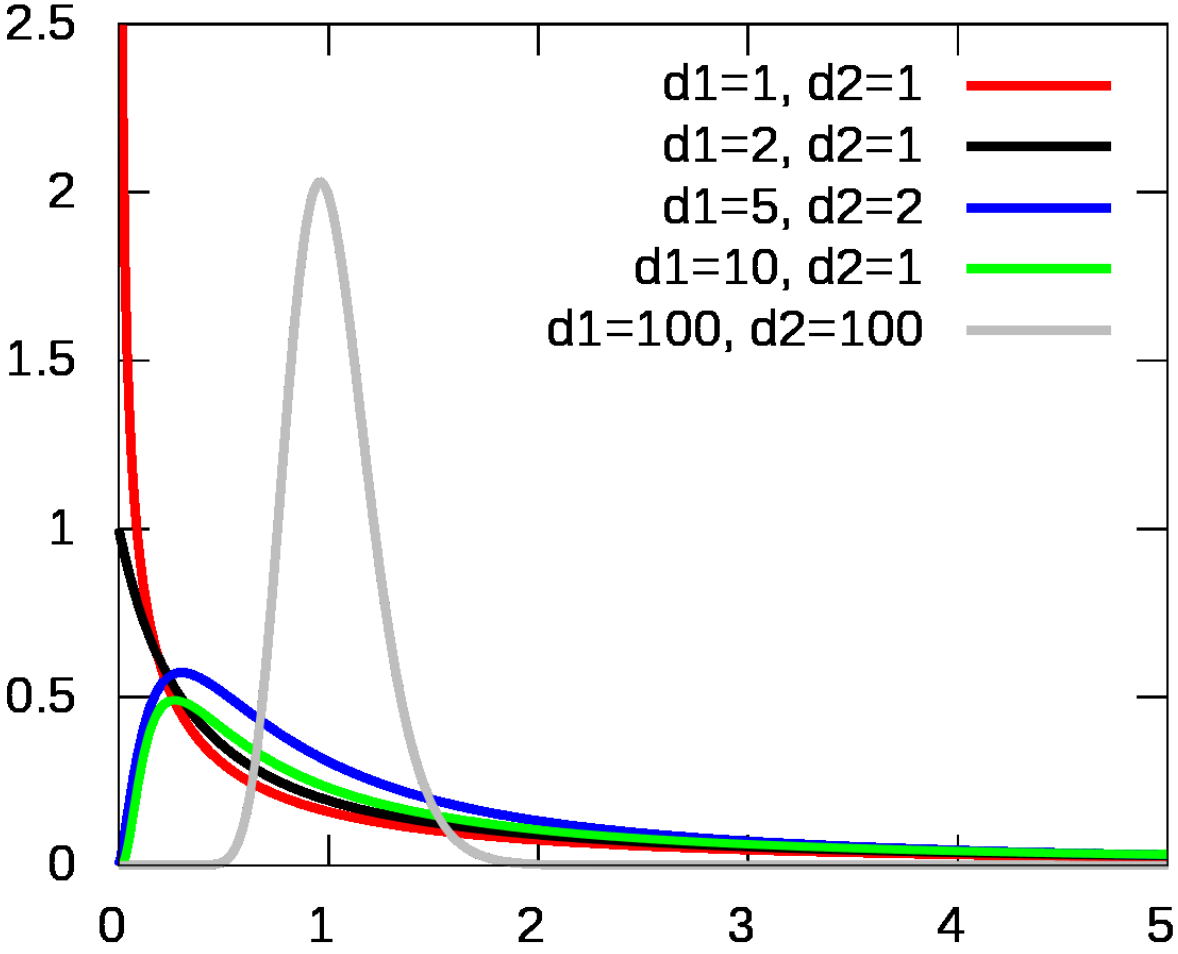
$$X = \frac{\frac{Y_1}{k_1}}{\frac{Y_2}{k_2}} = \frac{Y_1}{Y_2} \bigg/ \frac{k_1}{k_2}$$

follows the ***F*-distribution** with k_1 and k_2 degrees of freedom and we write

$$X \sim F(k_1, k_2) \qquad \text{or} \qquad X \sim F_{k_1, k_2}$$

F-distribution

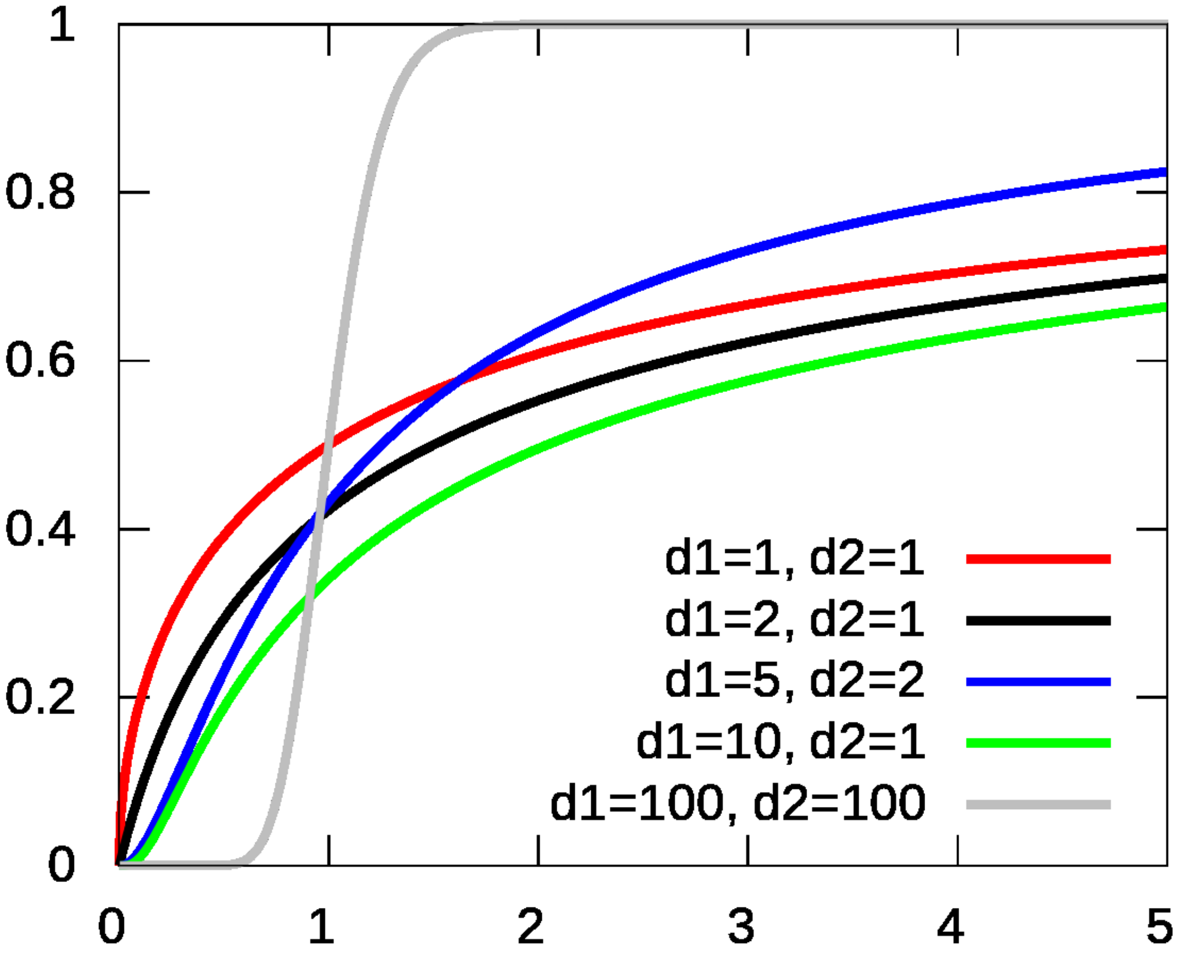
The graph of the probability density function of an $X \sim F_{k_1,k_2}$ for various k_1, k_2:



source: Wikipedia

F-distribution

The graph of the cumulative distribution function of an $X \sim F_{k_1,k_2}$ **for various** k_1, k_2**:**



source: Wikipedia

F-distribution

The probability density function:

$$f(x)=\begin{cases}\dfrac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}\left(\dfrac{k_1}{k_2}\right)^{\frac{k_1}{2}}x^{\frac{k_1}{2}-1}\left(1+\dfrac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}}, & x\in(0,+\infty)\\ 0, & x\in(-\infty,0]\end{cases}$$

F-distribution

Mean value (for $k_2 = 3, 4, 5, \ldots$):

$$\mu = \mathrm{E}[X] = \frac{k_2}{k_2 - 2}$$

Variance (for $k_2 = 5, 6, 7, \ldots$):

$$\sigma^2 = \mathrm{Var}(X) = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$$

Mode (for $k_1 = 3, 4, 5, \ldots$):

$$\hat{X} = \frac{k_1 - 2}{k_1}\frac{k_2}{k_2 + 2}$$

F-distribution in Excel

In Excel, use the functions:

=F.DIST$(x; k_1; k_2; \text{TRUE})$	to get the value of the cumulative distribution function of the random variable $X \sim F_{k_1 k_2}$ $(-\infty < x < +\infty)$
=F.DIST$(x; k_1; k_2; \text{FALSE})$	to get the value of the probability density function of the random variable $X \sim F_{k_1,k_2}$ $(-\infty < x < +\infty)$
=F.INV$(p; k_1; k_2)$	to get the quantile of the random variable $X \sim F_{k_1,k_2}$ $(0 < p < 1)$

F-distribution in Excel

In Excel, use the functions:

=F.DIST.RT$(x; k_1; k_2)$ the same as =1-F.DIST$(x; k_1; k_2; \text{TRUE})$

=F.INV.RT$(p; k_1; k_2)$ the same as =F.INV$(1 - p; k_1; k_2)$

That is, these functions calculate the cumulative distribution function from the right.

This is useful when one needs to calculate the critical value of some *F*-test.

F-distribution in Excel

In Excel, use the functions:

=FDIST()	the same as =F.DIST.RT(), deprecated
=FINV()	the same as =F.INV.RT, deprecated