Mathematics 6F – Statistics

http://cmp.felk.cvut.cz/~navara/m6f/eindex.htm

Mirko Navara, navara@math.feld.cvut.cz

Center for Machine Perception, Department of Cybernetics, FEE CTU

Karlovo náměstí, building G, room 104a

June 29, 2006

1 Basics of probability theory

1.1 Motivation example

A lot in a lottery is sold for the **price** of 2 EUR.

1 lot from 1000 wins 1000 EUR, others nothing. (This determines the **value** of the lot after the drawing of lots.)

What is the value of the lot **before** the drawing of lots?

Not 2 EUR, but $\frac{1}{1000}1000 = 1$ EUR $=$ average value after the drawing of lots.

This is a topic of **probability theory**.

Question "Lottery": Why do people participate in lotteries?

Here the rules were clearly described in the same units.

Modification: A lot of a tombola may win various prices whose value is individual and not exactly determined.

A good evaluation of such a lot requires handling incomplete information and more advanced arguments. The value of the lot is not exactly determined even after the drawing of lots.

This is a topic of **fuzzy sets theory and fuzzy logic**.

1.2 The role of statistics

So far we assumed that the parameters of the probabilistic model are known. This is rarely the case.

Example: In some games a winning strategy might be: Bet something else than the others. It is necessary to know the strategy of other players.

Example: In roulette, both sides want to know whether all numbers have the same probability. How to verify/deny this assumption? What is the risk of a wrong conclusion?

This is a topic of **statistics**.

Statistics gives us more: It is a tool for finding those laws of the world which are not apparent.

1.3 Probability (probability measure)

is a function P which maps events to numbers from $[0, 1]$ and satisfies

(P1) $P[true] = 1$,

(P2) $P\left[\bigvee_{n\in\mathbb{N}} A_n\right] = \sum_{n\in\mathbb{N}} P[A_n]$ if events $A_n,\ n \in \mathbb{N}$ are mutually exclusive.

Consequences:

$P[false] = 0, \qquad P[\neg A] = 1 - P[A]$,

if $A \implies B$, then $P[A] \leq P[B]$.

(For correctness, we need the collection of events to satisfy some conditions ...)

1.4 Random variable

is a measurable mapping of elementary events into $\mathbb{R}$*, i.e.,*

an object X described by probabilities $P[X \in I] = \omega_X(I)$ defined for all intervals $I \subseteq \mathbb{R}$ (and for any union of countably many intervals);

ω_X is a **probability measure** determining the distribution of random variable X. *(We restrict attention to so-called* perfect *measures, but others are not encountered in practice.)* It satisfies:

$\omega_X(\mathbb{R}) = 1$,

$\omega_X\left(\bigcup_{n\in\mathbb{N}} I_n\right) = \sum_{n\in\mathbb{N}} \omega_X(I_n)$ if intervals $I_n,\ n \in \mathbb{N}$ are mutually disjoint.

Consequences:

$\omega_X(\emptyset) = 0, \qquad \omega_X(\mathbb{R} \setminus I) = 1 - \omega_X(I),$

if $I \subseteq J$, then $\omega_X(I) \leq \omega_X(J)$ and $\omega_X(J \setminus I) = \omega_X(J) - \omega_X(I)$.

More efficient representation: we restrict attention to intervals of the form $I = (-\infty, t)$, $t \in \mathbb{R}$,

$$P[X \in (-\infty, t)] = P[X < t] = \omega_X((-\infty, t)) = F_X(t).$$

$F_X\colon \mathbb{R} \to [0, 1]$ is the **cumulative distribution function (cdf)** of a random variable X. This suffices because

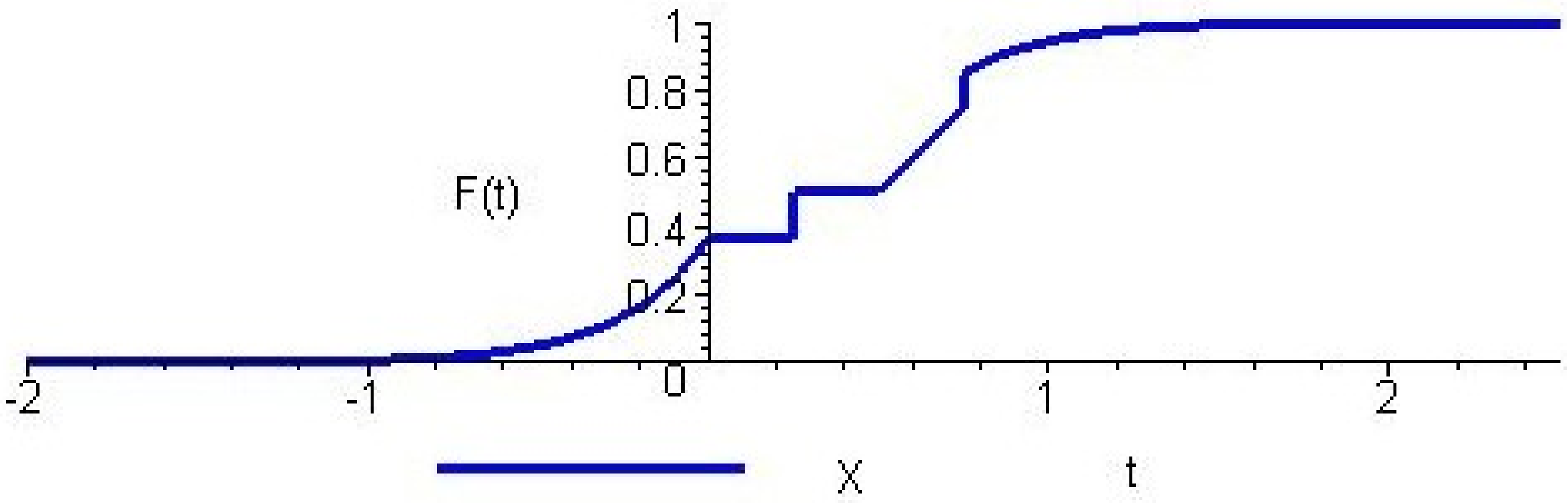
$$\begin{aligned}
[a, b) &= (-\infty, b) \setminus (-\infty, a),\\
[a, \infty) &= \mathbb{R} \setminus (-\infty, a),\\
(-\infty, a] &= \bigcap_{b\colon b>a} (-\infty, b),\\
\{a\} &= (-\infty, a] \setminus (-\infty, a),\\
&\dots
\end{aligned}$$

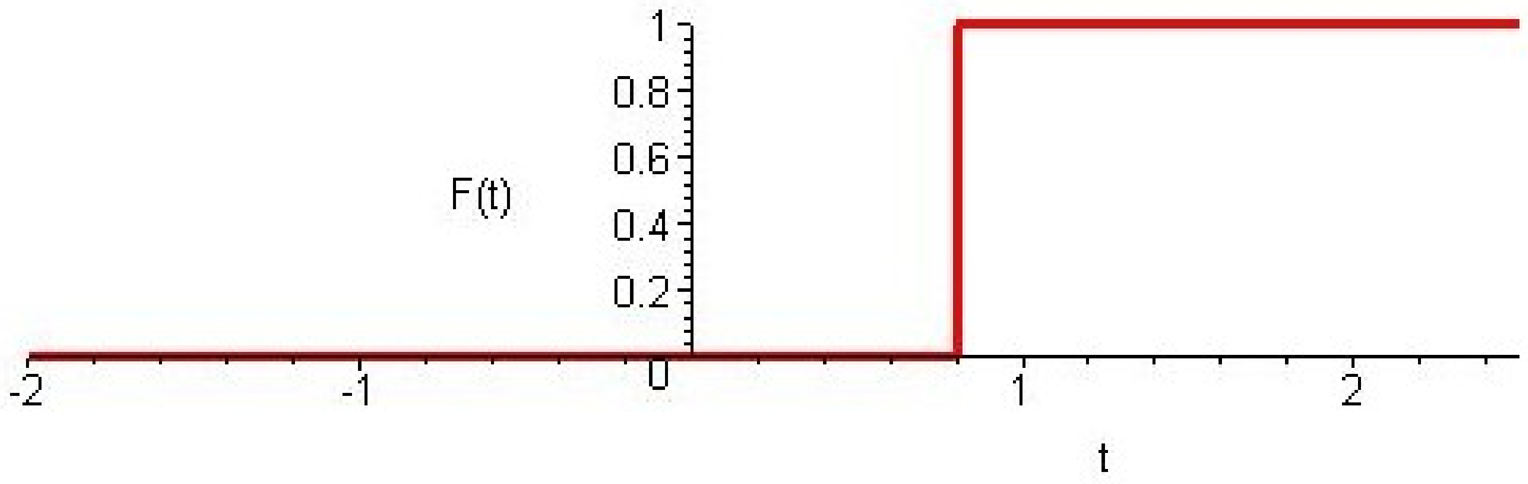
$$\begin{aligned}
\omega_X([a, b)) &= P[a \leq X < b] = F_X(b) - F_X(a),\\
\omega_X([a, \infty)) &= 1 - F_X(a),\\
\omega_X((-\infty, a]) &= P[X \leq a] = \lim_{b \to a+} F_X(b) = F_X(a+),\\
\omega_X(\{a\}) &= P[X = a] = F_X(a+) - F_X(a),\\
&\dots
\end{aligned}$$

A cumulative distribution function is:

- non-decreasing,
- left continuous,
- satisfies $\lim_{t \to -\infty} F_X(t) = 0, \quad \lim_{t \to \infty} F_X(t) = 1.$

Theorem: These above three conditions are **necessary** and **sufficient** for a function F_X to be a cdf of some random variable.


F(t)
1
0.8
0.6
0.4
0.2
0
-2
-1
1
2
X
t



Example: A real number r corresponds to a random variable (denoted also r) with the **Dirac** distribution concentrated in r:

$$\omega_r(I) = \begin{cases} 0 & \text{if} \quad r \notin I, \\ 1 & \text{if} \quad r \in I, \end{cases} \qquad F_r(t) = \begin{cases} 0 & \text{if} \quad t < r, \\ 1 & \text{if} \quad r \le t. \end{cases}$$

(F_r is a shifted Heaviside function.)

1.5 Random vector (vector random variable)

is a vector of random variables $\vec{X} = (X_1, \dots, X_n)$ determined by probabilities $P[X_1 \in I_1 \wedge \dots \wedge X_n \in I_n] = \omega_{\vec{X}}(I_1 \times \dots \times I_n)$, where $I_1, \dots, I_n$ are intervals in $\mathbb{R}$.

It suffices to restrict attention to intervals $I_k = (-\infty, t_k)$, $t_k \in \mathbb{R}$,

$$\begin{aligned} P[X_1 \in (-\infty, t_1) \wedge \dots \wedge X_n \in (-\infty, t_n)] &= P[X_1 < t_1 \wedge \dots \wedge X_n < t_n] \\ &= \omega_{\vec{X}}((-\infty, t_1) \times \dots \times (-\infty, t_n)) \\ &= F_{\vec{X}}(t_1, \dots, t_n). \end{aligned}$$

$F_{\vec{X}} \colon \mathbb{R}^n \to [0, 1]$ is the **cumulative distribution function (cdf)** of a random vector $\vec{X}$. It is

- non-decreasing (in all variables),

- left continuous (in all variables),

- $\lim_{t_1 \to -\infty, \dots, t_n \to -\infty} F_{\vec{X}}(t_1, \dots, t_n) = 0, \quad \lim_{t_1 \to \infty, \dots, t_n \to \infty} F_{\vec{X}}(t_1, \dots, t_n) = 1.$

Theorem: These are **necessary** and **sufficient** conditions.

Is is not sufficient to know the **marginal** distributions of random variables $X_1, \dots, X_n$, because they do not bear information about dependence. **Independent** variables $X_1, \dots, X_n$ allow a simplification to:

$$F_{\vec{X}}(t_1, \dots, t_n) = P\left[X_1 < t_1 \wedge \dots \wedge X_n < t_n\right] = \prod_{i=1}^{n} P\left[X_i < t_i\right]$$

1.6 More general random variables

A **complex random variable** is a random vector with two entries interpreted as a real and imaginary part.

Sometimes we admit “random variables” achieving non-numeric values which do not admit any arithmetic or ordering. We can choose a numbering of these values, but there is no canonical or natural way how to do it, so numerical computations with these values are meaningless.

1.7 Mixture of random variables

U, V with coefficients $c, 1-c \in [0,1]$ is a random variable $X = \text{Mix}(c, U;\ 1-c, V)$ with cdf

$$F_X = cF_U + (1-c)F_V,$$

$$F_X(t) = cF_U(t) + (1-c)F_V(t).$$

The corresponding probability measure is $\omega_X = c\omega_U + (1-c)\omega_V$.

A correct intruduction of these notions requires to define the corresponding probability space. Nevertheless, the distribution is obtained easily and it is usually sufficient for subsequent procedures.

More generally, the **mixture of random variables** $U_1, \dots, U_n$ **with coefficients** $c_1, \dots, c_n \in [0,1]$, $\sum_{i=1}^{n} c_i = 1$, is a random variable $\text{Mix}(c_1, U_1;\ \dots;\ c_n, U_n)$ with cdf $\sum_{i=1}^{n} c_i\, F_{U_i}$. The corresponding probability measure is $\sum_{i=1}^{n} c_i\, \omega_{U_i}$.

This can be generalized to countably many random variables.

Example: Mixture of reals $r_1, \ldots, r_n$ with coefficients $c_1, \ldots, c_n$ is a random variable $X = \mathrm{Mix}(c_1, r_1; \ldots; c_n, r_n)$,

$$\omega_X(I) = P[X \in I] = \sum_{i : r_i \in I} c_i, \qquad F_X(t) = \sum_{i : r_i \leq t} c_i.$$

It can be described by the **probability function** $p_X : \mathbb{R} \to [0, 1]$,

$$p_X(t) = \omega_X(\{t\}) = P[X = t] = \begin{cases} c_i & \text{if } t = r_i, \\ 0 & \text{otherwise.} \end{cases}$$

This can be generalized to countably many reals.

1.8 Types of random variables

1. **Discrete**: (from the previous example) There is a countable set O_X such that $\omega_X(\mathbb{R} \setminus O_X) = P\,[X \notin O_X] = 0$. The least such set (if it exists) is $\Omega_X = \{t \in \mathbb{R} : \omega_X(\{t\}) \neq 0\} = \{t \in \mathbb{R} : P[X = t] \neq 0\}$.

   A discrete random variable can be described by the **probability function** $p_X(t) = \omega_X(\{t\}) = P[X = t]$.

   It satisfies $\sum\limits_{t \in \mathbb{R}} p_X(t) = 1$.

2. **(Absolute) continuous**:

$$F_X(t) = \int_{-\infty}^{t} f_X(u)\, du$$

   for a non-negative function $f_X : \mathbb{R} \to [0, \infty)$ called a **probability density function (pdf)** of the random variable X.

   It satisfies $\int\limits_{-\infty}^{\infty} f_X(u)\, du = 1$.

It is not determined uniquely, but two pdfs f_X, g_X of the same random variable have a difference such that

$$\int_I (f_X(x) - g_X(x))\, dx = 0$$

for all intervals I.

We may choose $f_X(t) = \frac{dF_X(t)}{dt}$ if the derivative exists.

$\omega_X(\{t\}) = 0$ for all t.

3. **Mixed**: A mixture of the preceding cases;

   $\Omega_X \neq \emptyset$, $\omega_X(\mathbb{R} \setminus \Omega_X) = P\left[X \notin O_X\right] \neq 0$.

   It has neither a probability function, nor a pdf.

4. Other cases: E.g., a random variable with a continuous cdf which cannot be expressed as an integral. We exclude such cases in the sequel.

1.9 Description of a mixed random variable

A mixed random variable X can be **uniquely** expressed in the form $X = \mathsf{Mix}(c, U;\ 1-c, V)$, where U is discrete, V is continuous, and $c \in (0,1)$:

$$\begin{aligned}
c &= \omega_X(\Omega_X) = \omega_X(\{t \in \mathbb{R} : \omega_X(\{t\}) \neq 0\}),\\
c\,\omega_U(\{t\}) + (1-c)\,\underbrace{\omega_V(\{t\})}_{0} &= c\,\omega_U(\{t\}) = \omega_X(\{t\}),\\
p_U(t) = \omega_U(\{t\}) &= \frac{\omega_X(\{t\})}{c},\\
\Omega_U &= \Omega_X,\\
c\,\omega_U(I) + (1-c)\,\omega_V(I) &= \omega_X(I),\\
\omega_V(I) &= \frac{\omega_X(I) - c\,\omega_U(I)}{1-c},\\
F_V(t) &= \frac{F_X(t) - c\,F_U(t)}{1-c}.
\end{aligned}$$

Alternatively, without the use of a probability measure:

$$
\begin{aligned}
c &= \sum_{t\in\mathbb{R}} P[X=t], \\
c\,P[U=t] &= P[X=t], \\
p_U(t) = P[U=t] &= \frac{P[X=t]}{c}, \\
c\,P[U\in I] + (1-c)\,P[V\in I] &= P[X\in I], \\
P[V\in I] &= \frac{P[X\in I] - c\,P[U\in I]}{1-c}, \\
F_V(t) &= \frac{F_X(t) - c\,F_U(t)}{1-c}.
\end{aligned}
$$

(We may continue by a decomposition of the discrete part to a mixture of Dirac distributions.)

1.10 Quantile function of a random variable

$$\forall\alpha\in(0,1)\ \exists t\in\mathbb{R} : P[X<t]\le\alpha\le P[X\le t].$$

If there are more such numbers, they form an interval from which *(usually)* the center is taken; more precisely

$$Q_X(\alpha) = \frac{1}{2}\left(\sup\{t \in \mathbb{R} \mid P[X < t] \leq \alpha\} + \inf\{t \in \mathbb{R} \mid \alpha \leq P[X \leq t]\}\right).$$

The number $Q_X(\alpha)$ is called an **α-quantile** of the random variable X; the function $Q_X\colon (0,1) \to \mathbb{R}$ is the **quantile function** of the random variable X. In particular $Q_X(\frac{1}{2})$ is the **median**, also further quantiles have their names. Properties of the quantile function:

- it is nondecreasing,
- $Q_X(\alpha) = \frac{1}{2}(Q_X(\alpha-) + Q_X(\alpha+))$.

Theorem: These conditions are **necessary** and **sufficient**.

We may speak of a **vertical representation** of a random variable by a cdf $F_X\colon \mathbb{R} \to [0,1]$ and a **horizontal representation** by a quantile function $Q_X\colon (0,1) \to \mathbb{R}$.

The inverse transformation:

$$F_X(t) = \inf\{\alpha \in (0,1) \mid Q_X(\alpha) > t\}.$$

Functions F_X, Q_X are mutually inverse whenever they are continuous and increasing (it suffices to check this for one of them).

1.11 How to represent a random variable in a computer

1. **Discrete**: If it attains only finitely many values $t_k,\ k = 1, \ldots, n$, we need only these values and their probabilities $p_X(t_k) = \omega_X(\{t_k\}) = P[X = t_k]$; these fully describe the probability function by $2n$ numbers (up to the imprecise representation of reals in a computer).

   If a discrete random variable attains (countably) infinitely many values, we have to ignore some of them, e.g., those with small probabilities. For each $\varepsilon > 0$ we may choose finitely many $t_k,\ k = 1, \ldots, n$, so that $\omega_X(\mathbb{R} \setminus \{t_1,\ \ldots, t_n\}) = P[X \notin \{t_1,\ \ldots, t_n\}] \le \varepsilon$. A problem remains which value should be assigned to the remaining (less probable) cases.

2. **(Absolute) continuous**: The pdf can be approximated by values $f(t_k)$ in "sufficiently many" points $t_k,\ k = 1, \ldots, n$, provided that f is "sufficiently smooth". We are rather interested in the integrals

$$F_X(t_{k+1}) - F_X(t_k) = \int_{t_k}^{t_{k+1}} f_X(u)\, du,$$

from which the cdf can be approximated. We may use directly the values $F_X(t_k)$ for the representation. We need a "dense" set of points in places where the pdf is high.

The points $t_k,\ k = 1, \ldots, n$, may be chosen so that the increments $F_X(t_{k+1}) - F_X(t_k)$ have given magnitudes. We choose $\alpha_k \in (0, 1),\ k = 1, \ldots, n$, and find $t_k = Q_X(\alpha_k)$.

Memory requirements are high, they are dependent on the scale of values of the random variable and the cdf.

Very often the type of distribution is known and a few parameters suffice to determine it completely.

Many general cases are treated as mixtures of random variables with known types of distributions; then we manage with finitely many parameters.

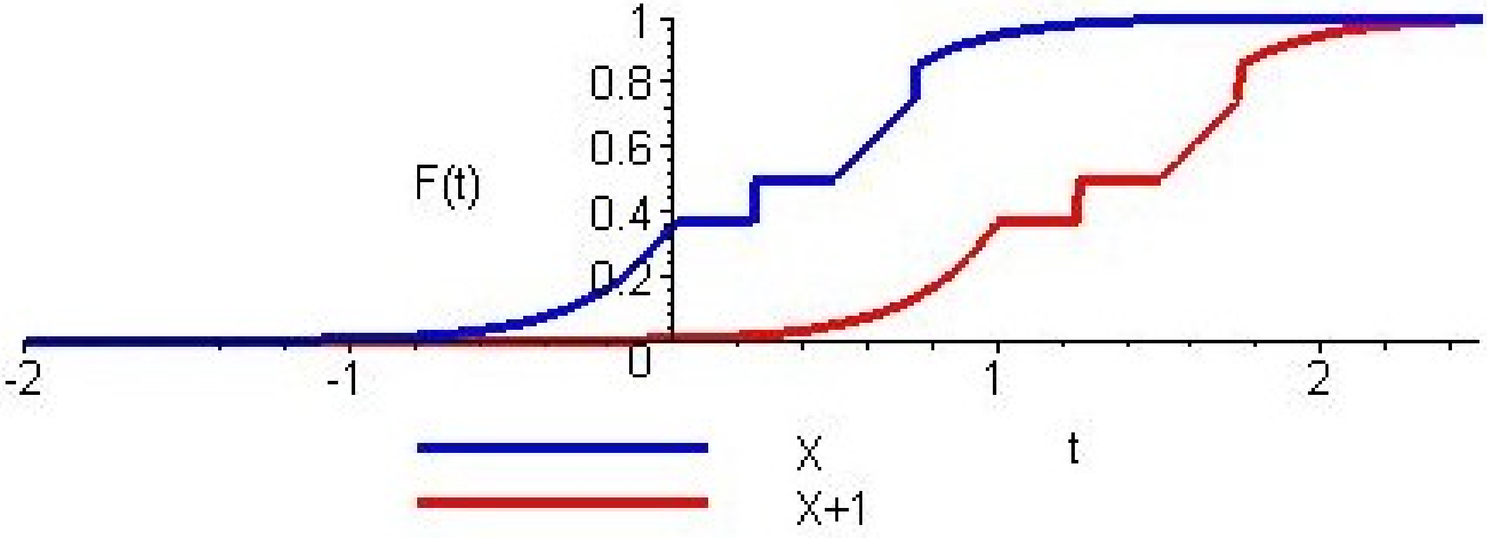
3. **Mixed**: The same as for a continuous random variable. However, this description for the discrete part is unnecessarily inaccurate.

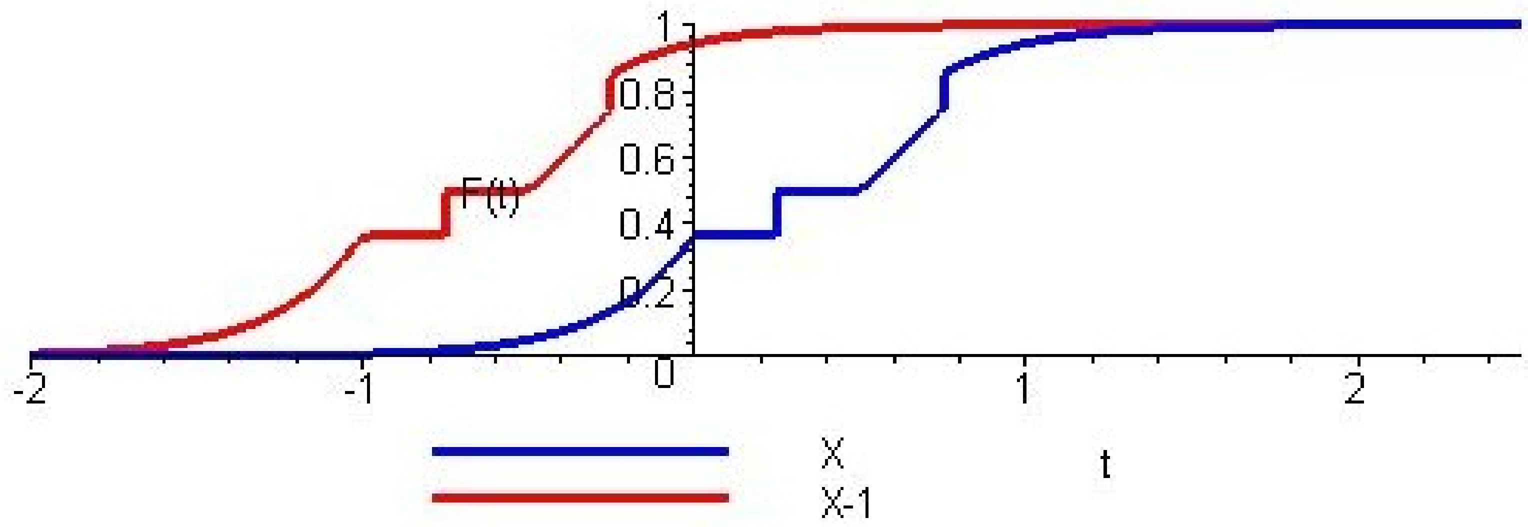
1.12 Operations with random variables

Here $I, J \subseteq \mathbb{R}$ are intervals of countable unions of intervals.

Addition of a constant r corresponds to a horizontal shift:

$$\begin{aligned} \omega_{X+r}(I+r) &= \omega_X(I), & \omega_{X+r}(J) &= \omega_X(J-r), \\ F_{X+r}(t+r) &= F_X(t), & F_{X+r}(u) &= F_X(u-r), \\ Q_{X+r}(\alpha) &= Q_X(\alpha)+r. \end{aligned}$$


1
0.8
0.6
F(t)
0.4
0.2
-2
-1
0
1
2
t
X
X+1

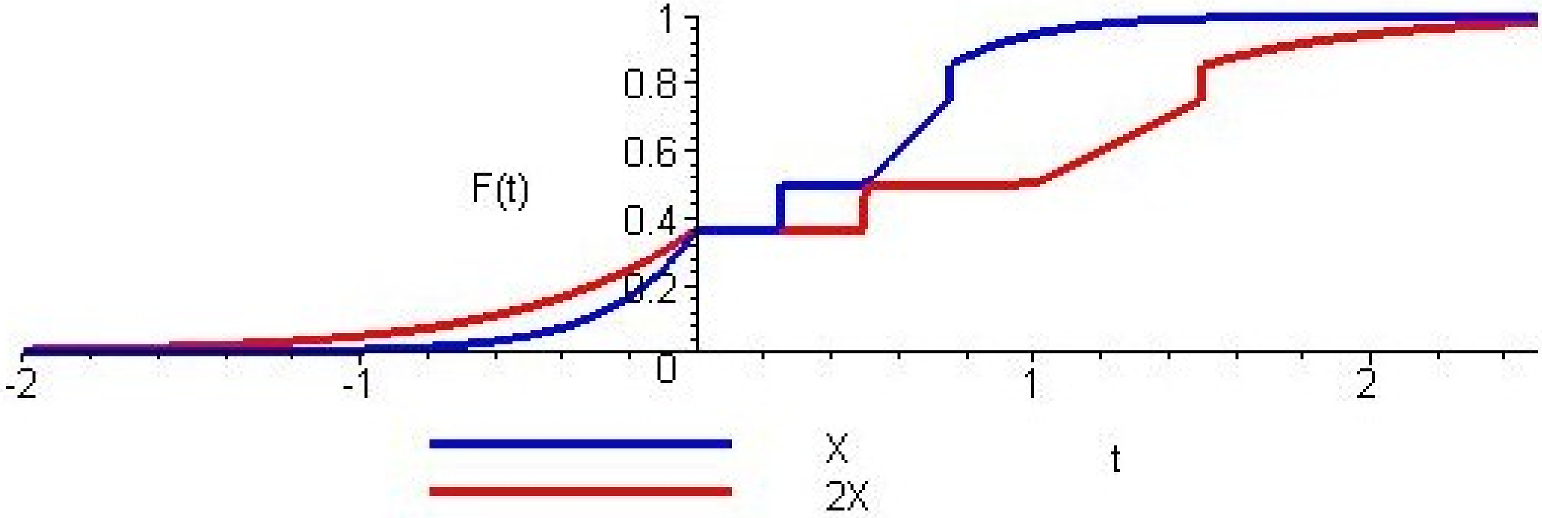


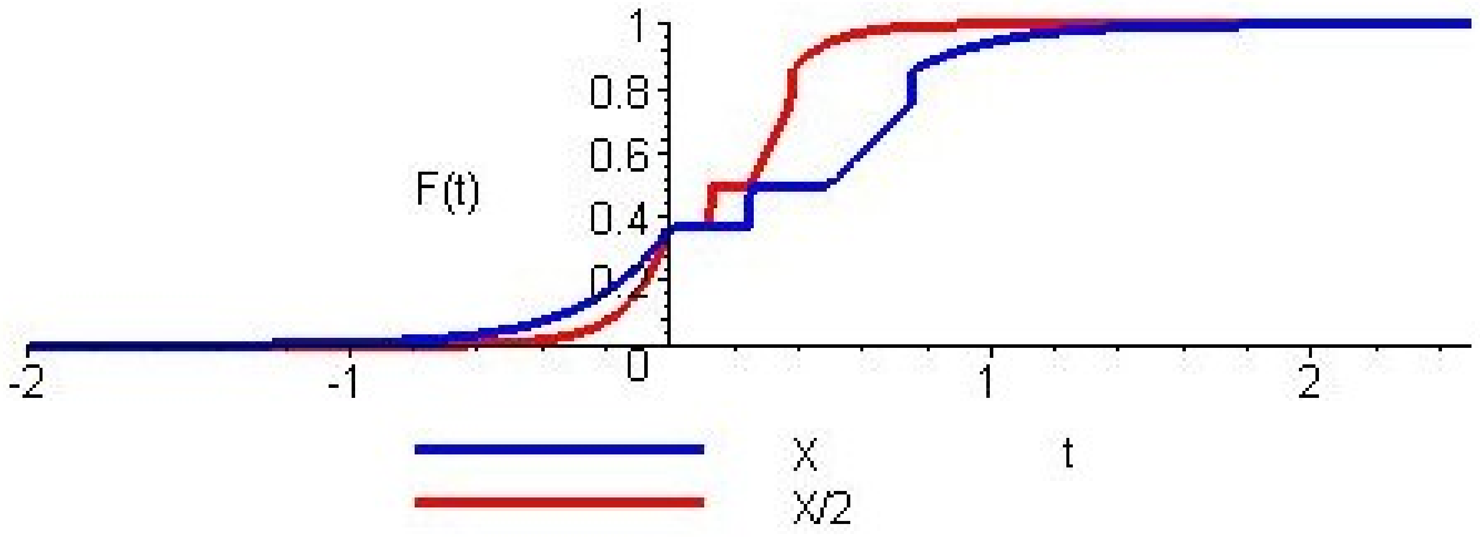
Multiplication by a non-zero constant r corresponds to a similarity transformation:

$$\omega_{rX}(rI) = \omega_X(I), \qquad \omega_{rX}(J) = \omega_X\left(\frac{J}{r}\right).$$

For cdf, we have to distinguish several cases:

- $r > 0$: $\qquad F_{rX}(rt) = F_X(t), \qquad F_{rX}(u) = F_X\left(\frac{u}{r}\right), \qquad Q_{rX}(\alpha) = r\,Q_X(\alpha),$


1
0.8
0.6
F(t)
0.4
0.2
-2
-1
0
1
2
t
X
2X

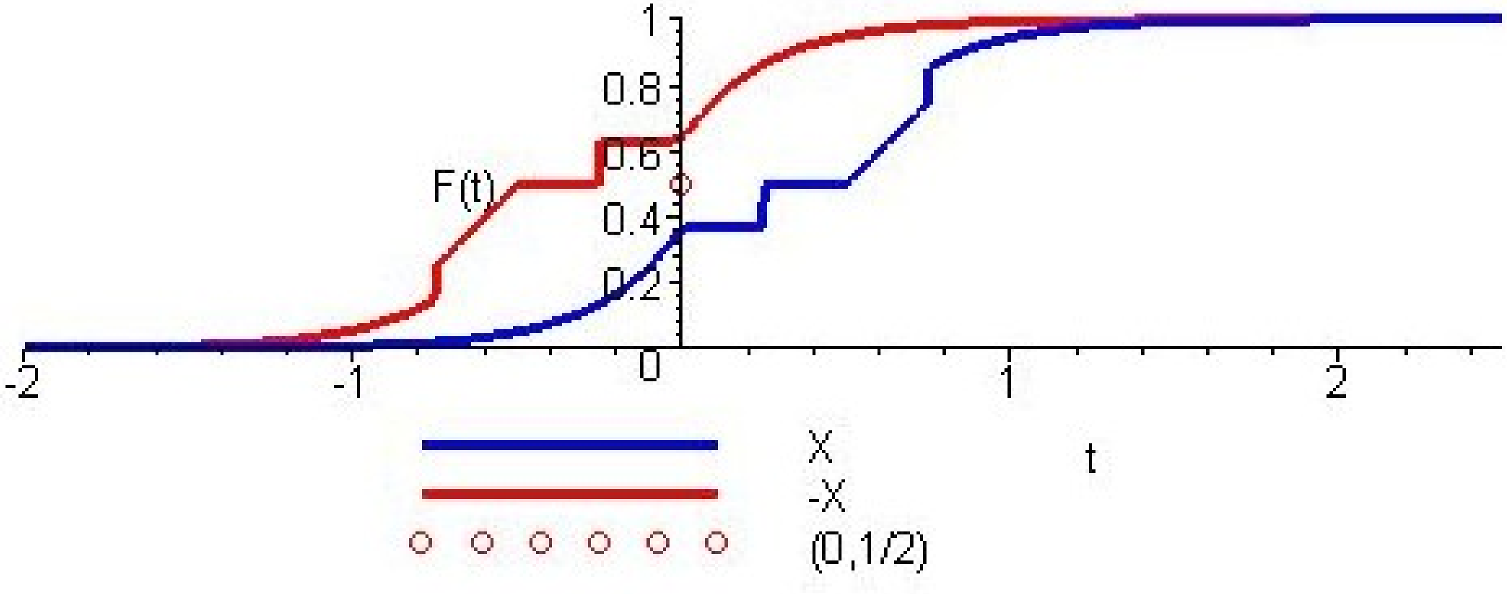


- $r = -1$: $\quad F_{-X}(-t) = \omega_{-X}((-\infty,-t)) = \omega_X((t,\infty)) = 1 - \omega_X((-\infty,t])$, **at continuity points** of the cdf

  $F_{-X}(-t) = 1 - \omega_X((-\infty,t]) = 1 - P[X \leq t] = 1 - P[X < t] = 1 - \omega_X((-\infty,t)) = 1 - F_X(t)$,

  $F_{-X}(u) = 1 - F_X(-u)$, the left limit in discontinuity points (symmetry w.r.t. point $\left(0, \frac{1}{2}\right)$ with a correction to left continuity),

$$Q_{-X}(\alpha) = -Q_X(1-\alpha),$$



- $r < 0$: combination of previous cases.

Mapping by a continuous increasing function h: $\omega_{h(X)}(h(I)) = \omega_X(I)$, $F_{h(X)}(h(t)) = F_X(t)$,
$Q_{h(X)}(\alpha) = h(Q_X(\alpha))$ **in points in which the quantile function is continuous**.

Mapping by a right continuous non-decreasing function h: $F_{h(X)}(u) = \sup\{F_X(t) \mid h(t) < u\}$.

Mapping by a piecewise monotonic function h**:**
We may express it as $h = h_+ - h_-$, where h_+, h_- are non-decreasing.
We express X as a mixture of two random variables; the range of the first, resp. the second, contains only points in which h is non-decreasing, resp. non-increasing.
The result is obtained as a mixture of two random variables resulting from mappings by function h "componentwise", i.e., $h(\mathrm{Mix}(c, U; 1 - c, V)) = \mathrm{Mix}(c, h(U); 1 - c, h(V))$.

Sum of random variables is not uniquely determined unless we assume their **independence**. Even in this case it can be highly non-trivial.

Mixture of random variables – see above. *In contrast to the sum, it is uniquely determined by the (marginal) distributions of input random variables and coefficients of the mixture.*

1.13 Computer implementation of a random variable

1. We create a generator of a random (or pseudorandom) variable X with the uniform distribution on $[0, 1]$.

2. The random variable $Q_Y(X)$ has the same distribution as Y. (It suffices to apply Q_Y to each realization of the random variable X.)

All **continuous** distributions are equivalent up to a (non-linear) change of scale.

1.14 Expectation (expected value, mean, mean value)

is defined separately

- for a **discrete** random variable U:

$$\mu_U = \sum_{t \in \mathbb{R}} t \cdot p_U(t) = \sum_{t \in \Omega_U} t \cdot p_U(t),$$

- for a **continuous** random variable V:

$$\mu_V = \int\limits_{-\infty}^{\infty} t \cdot f_V(t)\, dt,$$

- for a **mixture** of random variables $X = \mathsf{Mix}(c, U;\ 1-c, V)$, where U is discrete, V continuous:

$$\mu_X = c\mu_U + (1-c)\mu_V.$$

*(This **is not** a linearity of expectation!)*

All three cases can be covered by a single formula for the quantile function

$$\mu_X = \int\limits_0^1 Q_X(\alpha)\, d\alpha.$$

This can be easily generalized to the expectation of a function of a random variable:

$$\mu_{h(X)} = \int\limits_0^1 h\left(Q_X(\alpha)\right)\, d\alpha.$$

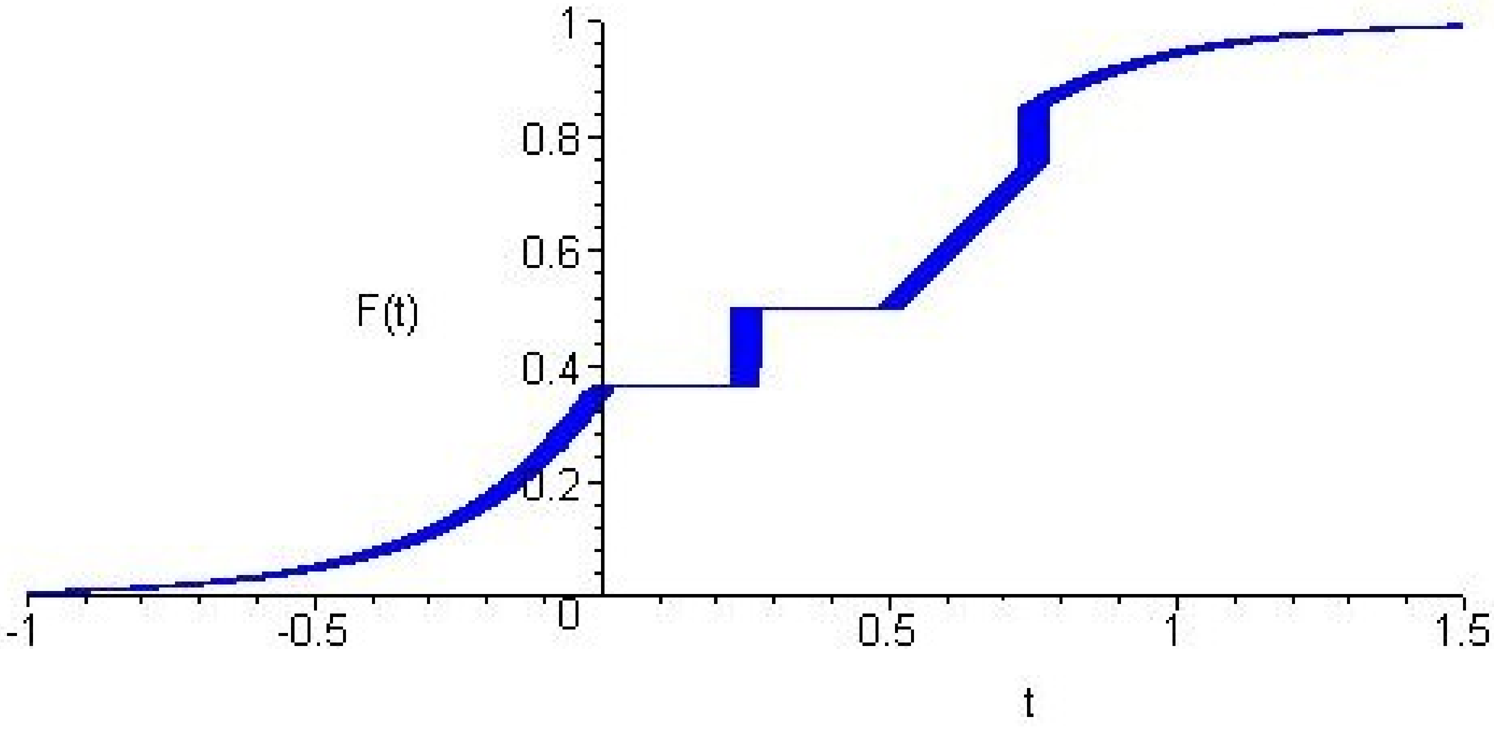
In particular, for a **discrete** random variable

$$\mu_{h(U)} = \sum_{t \in \Omega_U} h\left(t\right) \cdot p_U(t).$$

An analogy for a continuous random variable requires additional assumptions because the continuity of a random variable need not be preserved by the application of a function.

Alternative notation of expectation: EX.

It is the horizontal coordinate of the center of gravity of the graph of the cumulative distribution function, where elements are weigted by the increase of cdf:



Working with expectation, we assume its existence (which is not obvious).

Expectation

- of a vector random variable $\vec{X} = (X_1, \dots, X_N)$: $\quad \mu_{\vec{X}} = (\mu_{X_1}, \dots, \mu_{X_N})$

- of a complex random variable $X = \Re(X) + i\,\Im(X)$: $\mu_X = \Re(\mu_X) + i\,\Im(\mu_X)$

1.14.1 Properties of expectation

$$\begin{aligned}
\mu_r &= r, \\
\mu_{X+Y} &= \mu_X + \mu_Y, \qquad \text{in particular,} \qquad \mu_{X+r} = \mu_X + r, \\
\mu_{X-Y} &= \mu_X - \mu_Y, \\
\mu_{rX} &= r\,\mu_X, \qquad \text{more generally,} \qquad \mu_{rX+sY} = r\,\mu_X + s\,\mu_Y.
\end{aligned}$$

(This **is** *a linearity of expectation.)*

$$\mu_{\mathsf{Mix}(c,U;\,1-c,V)} = c\mu_U + (1-c)\mu_V.$$

(This **is not** *a linearity of expectation!)*

Only for **independent** random variables

$$\mu_{X\cdot Y} = \mu_X \cdot \mu_Y.$$

1.15 Variance (dispersion)

$$\sigma_X^2 = \mu_{(X-\mu_X)^2} = \mu_{X^2} - \mu_X^2$$

$$\mu_{X^2} = \mu_X^2 + \sigma_X^2$$

Alternative notations: DX, var X.

Properties:

$$\sigma_X^2 = \int_0^1 (Q_X(\alpha) - \mu_X)^2 \, d\alpha.$$

$$\sigma_X^2 \geq 0,$$
$$\sigma_r^2 = 0,$$
$$\sigma_{X+r}^2 = \sigma_X^2,$$
$$\sigma_{rX}^2 = r^2 \, \sigma_X^2.$$

$$
\begin{aligned}
\sigma^2_{\mathsf{Mix}(c,U;\,1-c,V)} &= \mu_{\mathsf{Mix}(c,U;\,1-c,V)^2} - \mu^2_{\mathsf{Mix}(c,U;\,1-c,V)} \\
&= c\mu_{U^2} + (1-c)\mu_{V^2} - (c\mu_U + (1-c)\mu_V)^2 \\
&= c\left(\sigma_U^2 + \mu_U^2\right) + (1-c)\left(\sigma_V^2 + \mu_V^2\right) - (c\mu_U + (1-c)\mu_V)^2 \\
&= c\sigma_U^2 + (1-c)\sigma_V^2 + c\mu_U^2 + (1-c)\mu_V^2 - (c\mu_U + (1-c)\mu_V)^2 \\
&= c\sigma_U^2 + (1-c)\sigma_V^2 + c(1-c)\left(\mu_U - \mu_V\right)^2
\end{aligned}
$$

Only for **independent** random variables

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y, \qquad \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y.$$

1.16 Standard deviation

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\mu_{(X-\mu_X)^2}}$$

Properties:

$$\sigma_X = \sqrt{\int_0^1 (Q_X(\alpha) - \mu_X)^2 \; d\alpha.}$$

$$\begin{aligned} \sigma_X &\geq 0, \\ \sigma_r &= 0, \\ \sigma_{X+r} &= \sigma_X, \\ \sigma_{rX} &= |r| \; \sigma_X. \end{aligned}$$

Only for **independent** random variables

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2} \qquad \text{(quadratic mean).}$$

1.17 General moments

$k \in \mathbb{N}$

kth **general moment** *(no special notation introduced here)*: μ_{X^k}, in particular:

for $k = 1$: $\mu_{X^1} = \mu_X$,

for $k = 2$: $\mu_{X^2} = \mu_X^2 + \sigma_X^2$.

Alternative notation: m_k.

kth **centred moment** (a kth **moment about** μ_X; *no notation introduced here*): $\mu_{(X-\mu_X)^k}$, in particular:

for $k = 1$: 0,

for $k = 2$: σ_X^2.

Alternative notation: μ_k.

$$\mu_{X^k} = \int_0^1 (Q_X(\alpha))^k \ d\alpha$$

$$\mu_{(X-\mu_X)^k} = \int_0^1 (Q_X(\alpha) - \mu_X)^k \ d\alpha$$

1.18 Normalized random variable

has a zero mean and a unit variance.

Normalization is made by the formula

$$\text{norm}\, X = \frac{X - \mu_X}{\sigma_X}$$

if it is meaningful.

Do not confuse it with (Gaussian) **normal distribution** $N(\mu, \sigma^2)$ which has the density

$$f_{N(\mu,\sigma^2)}(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(t-\mu)^2}{2\,\sigma^2}\right).$$

Its cdf is the Gauss integral. In particular, the pdf of the **normalized normal distribution** $N(0,1)$ (abreviated notation: N) is

$$f_{N(0,1)}(t) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right).$$

1.19 Chebyshev inequality

Theorem:

$$\forall\delta > 0 : P\left[|\text{norm}\, X| < \delta\right] \geq 1 - \frac{1}{\delta^2}\,,$$

where $\text{norm}\, X = \frac{X - \mu_X}{\sigma_X}$ (whenever all these expressions are meaningful).

Proof using the quantile function:

$$\underbrace{\sigma^2_{\text{norm}\, X}}_{1} = \mu_{(\text{norm}\, X)^2} - \underbrace{\mu^2_{\text{norm}\, X}}_{0}\,,$$

$$1 = \mu_{(\text{norm}\, X)^2} = \mu_Y\,,$$

where $Y = (\text{norm}\, X)^2$. An estimate of the probability $\beta = P\left[|\text{norm}\, X| < \delta\right] =$

$P[Y < \delta^2] = F_Y(\delta^2)$:

$$1 = \mu_Y = \int_0^1 Q_Y(\alpha)\, d\alpha = \int_0^\beta \underbrace{Q_Y(\alpha)}_{\geq 0} d\alpha + \int_\beta^1 \underbrace{Q_Y(\alpha)}_{\geq \delta^2}\, d\alpha \geq (1-\beta)\delta^2\,,$$

$$\beta \geq 1 - \frac{1}{\delta^2}\,.$$

Proof using a mixture: We may express Y as a mixture $Y = (\operatorname{norm} X)^2 = \operatorname{Mix}(\beta, L;\, 1-\beta, U)$, where

- L attains only values from $[0, \delta^2)$,
- U attains only values from $[\delta^2, \infty)$, hence $\mu_U \geq \delta^2$,
- $\beta = F_Y(\delta^2)$.

$$1 = \mu_Y = \beta\ \underbrace{\mu_L}_{\geq 0} + (1-\beta)\ \underbrace{\mu_U}_{\geq \delta^2} \geq (1-\beta)\ \delta^2\,.$$

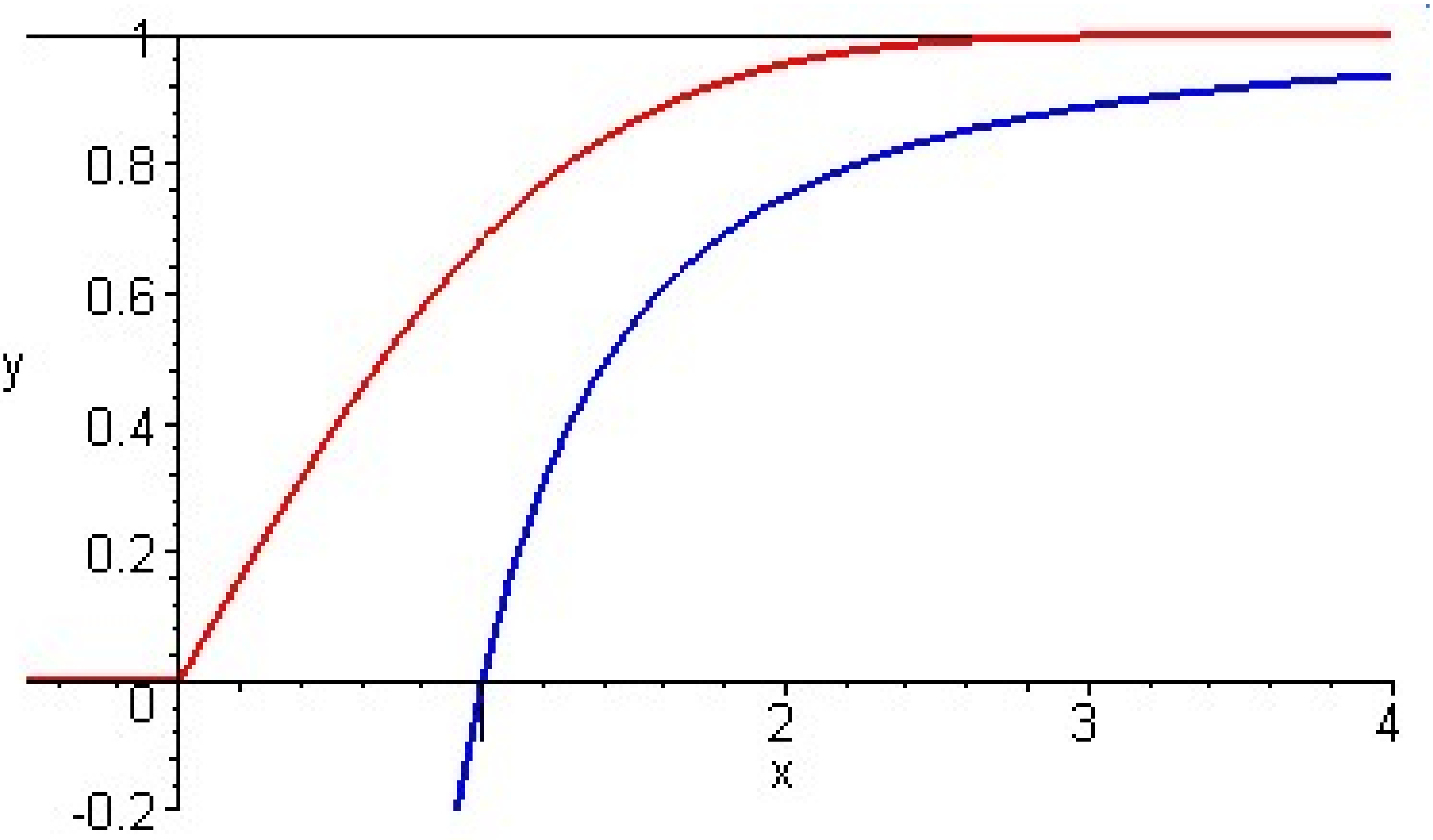
The equality occurs iff $U = \delta^2$, $L = 0$, i.e., for the discrete distribution

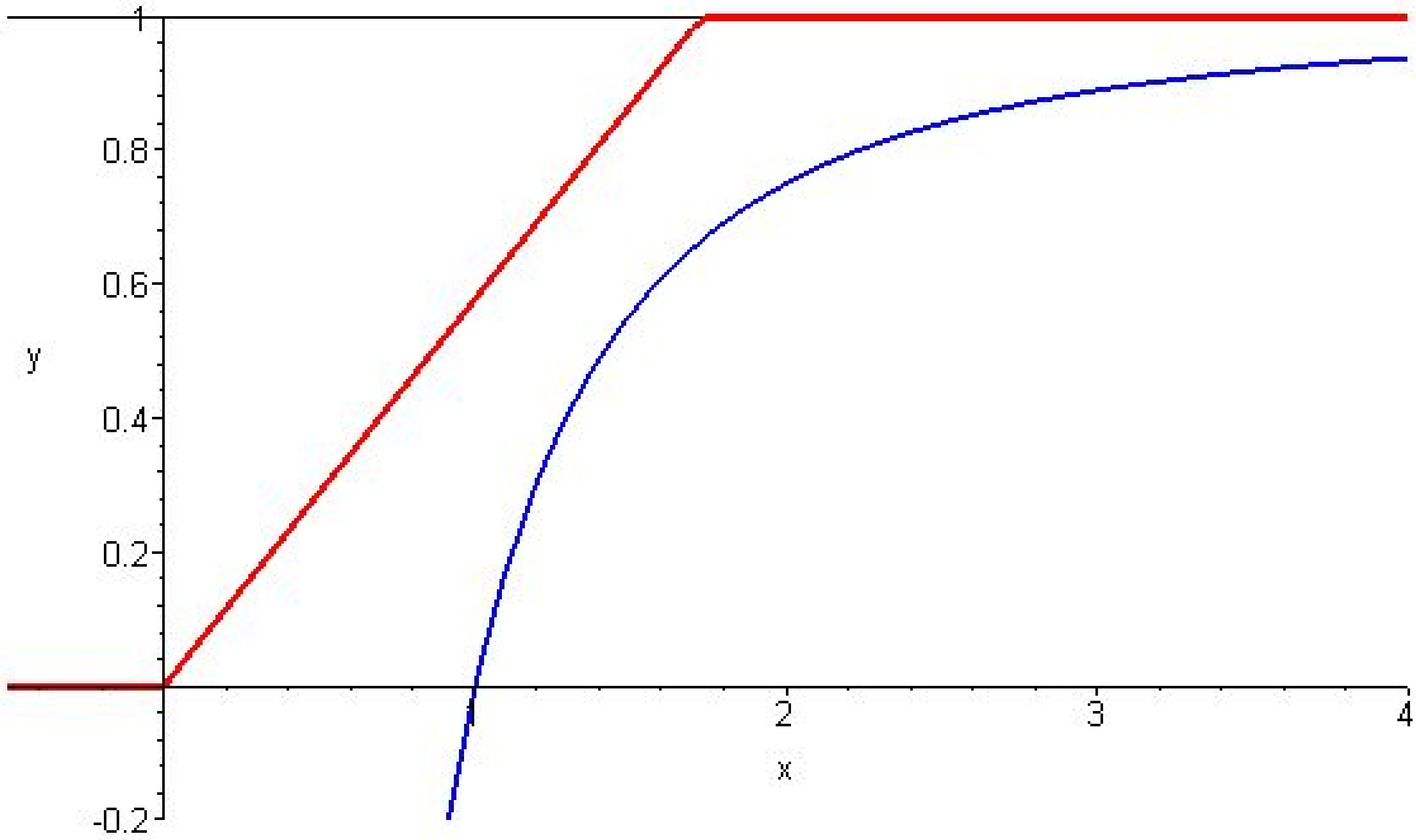
$\{(\mu_X - \delta\,\sigma_X, \frac{1-\beta}{2}), (\mu_X, \beta), (\mu_X + \delta\,\sigma_X, \frac{1-\beta}{2})\}$.

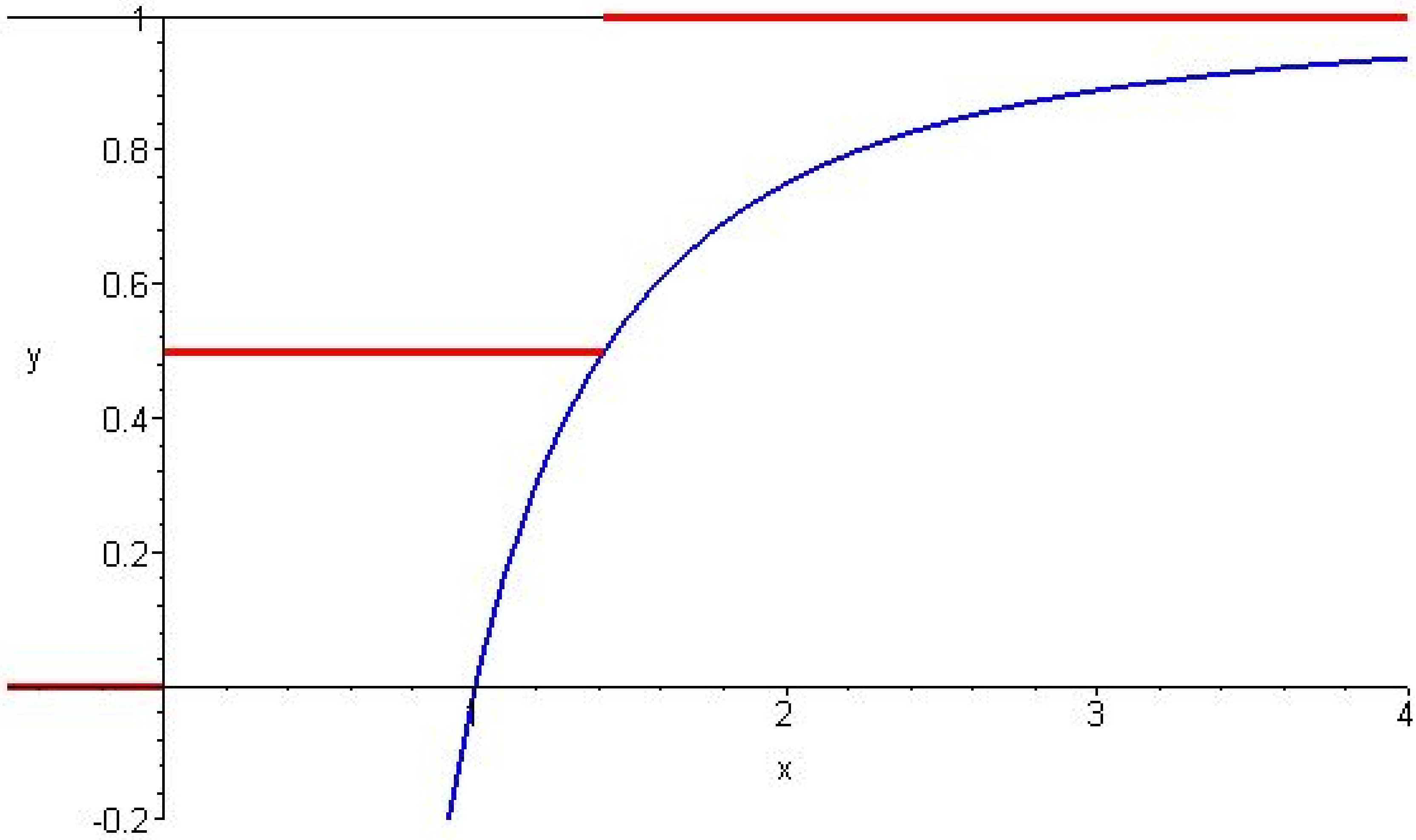
Equivalent forms ($\varepsilon = \delta\,\sigma_X$):

$$\forall\delta > 0 : P\left[\left|\frac{X - \mu_X}{\sigma_X}\right| \geq \delta\right] \leq \frac{1}{\delta^2},$$

$$\forall\varepsilon > 0 : P\left[|X - \mu_X| \geq \varepsilon\right] \leq \frac{\sigma_X^2}{\varepsilon^2}\,,$$


1
0.8
0.6
y
0.4
0.2
0
-0.2
1
2
3
4
x


1
0.8
0.6
y
0.4
0.2
-0.2
1
2
3
4
x


1
0.8
0.6
y
0.4
0.2
-0.2
1
2
3
4
x

2 Basic notions of statistics

2.1 Why do we need statistics?

We investigate **common** properties of many events.

Instead of studying all of them, we take only a sample (because of price, destructive tests, etc.).

- Estimation of parameters of probability models

- Hypotheses tests

2.2 Random sample, estimates

- **population**

- **sample**

The precision of a statistical estimate is determined by the sample size, not by the size of the population.

Random sample $\vec{X} = (X_1, \dots, X_n)$ is a vector of random variables which are **independent** and **equally distributed**.

(We omit indices, e.g., we write F_X instead of F_{X_k}.)

By an experiment we obtain the **realization of a random sample**, $\vec{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$.

n is the **sample size**.

Statistic is (any) measurable function G defined on a random sample of arbitrary size. (We compute it from the random variables, not from the parameters of the distribution which are unavailable.)

"**Measurability**" means that the probability

$$P[G(X_1, \ldots, X_n) < t] = F_{G(X_1,\ldots,X_n)}(t)$$

is defined for all $t \in \mathbb{R}$. Statistic – as a function of random variables – is itself also a random variable.

It is usually used as an **estimate of parameters of a distribution** (which remain hidden).

Notation:
θ ... actual parameter (real number),
$\hat{\theta}$ ... its estimate based on a random sample (random variable)

Desirable properties of estimates:

- $\mu_{\hat{\theta}} = \theta$ **unbiased**

- $\lim\limits_{n\to\infty} \mu_{\hat{\theta}} = \theta$ **asymptotically unbiased**

- **efficient** $=$ with a low variance; this is evaluated by $\mu_{(\hat{\theta}-\theta)^2}$, which reduces to $\sigma^2_{\hat{\theta}}$ for an unbiased estimate

- **the best unbiased** estimate is the most efficient among all unbiased estimates (nevertheless, there may exist more efficient *biased* estimates)

- $\lim\limits_{n\to\infty} \mu_{\hat{\theta}} = \theta$, $\lim\limits_{n\to\infty} \sigma_{\hat{\theta}} = 0$ **consistent**

- **robust**, i.e. resistant to noise (and outliers), “even noisy data lead to a good estimate”; here an exact criterion is missing, but it is of much practical use

2.3 Sample mean (sample average)

of a random sample $\vec{X} = (X_1, \ldots, X_n)$ is

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

Alternative notation: $\bar{X}_n$ *(when the sample size is important)*

Its realization is denoted by a small case letter:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j.$$

Theorem:

$$\mu_{\bar{X}_n} = \frac{1}{n}\sum_{j=1}^{n}\mu_X = \mu_X,$$

$$\sigma^2_{\bar{X}_n} = \frac{1}{n^2}\sum_{j=1}^{n}\sigma^2_X = \frac{1}{n}\sigma^2_X,$$

$$\sigma_{\bar{X}_n} = \sqrt{\frac{1}{n}\sigma^2_X} = \frac{1}{\sqrt{n}}\sigma_X,$$

if they exist. (Here $\mu_X = \mu_{X_j}$ etc.)

Corollary: The sample mean is a unbiased consistent estimate of the mean. (Independently on the type of distribution.)

Chebyshev inequality for $\bar{X}_n$ implies

$$P\left[\left|\bar{X}_n - \mu_X\right| \geq \varepsilon\right] \leq \frac{\sigma^2_{\bar{X}_n}}{\varepsilon^2} = \frac{\sigma^2_X}{n\,\varepsilon^2} \to 0 \qquad \text{for } n \to \infty.$$

This holds under more general assumptions (X_j need not have the same distribution) – the **weak law of large numbers**.

This is often called an “precise sum of imprecise numbers”. This is not correct because the sum $\sum_{j=1}^{n} X_j$ has variance $n\,\sigma_X^2 \to \infty$. The **relative** error of the sum **decreases**, the **absolute** error **increases**.

The distribution of the sample mean may be much more complex than the original one; an easy description exists only in special cases.

Theorem: The sample mean of the **normal** distribution $N(\mu_X, \sigma_X^2)$ has the normal distribution $N\left(\mu_X, \frac{1}{n}\sigma_X^2\right)$; it is the best unbiased estimate of the mean.

(Later on, we shall see a more efficient biased estimate.)

An analogous theorem holds for other distributions at least asymptotically:

Central Limit Theorem: Let $X_j,\ j \in \mathbb{N}$, be independent equally distributed random variables with mean μ_X and standard deviation $\sigma_X \neq 0$. Then the normalized random variables

$$Y_n = \frac{\sqrt{n}}{\sigma_X}(\bar{X}_n - \mu_X)$$

converge to the normalized normal distribution in the following sense:

$$\forall t \in \mathbb{R} : \lim_{n \to \infty} F_{Y_n}(t) = F_{N(0,1)}(t).$$

2.4 Sample variance (sample dispersion)

of a random sample $\vec{X} = (X_1, \ldots, X_n)$ is the statistic

$$S_X^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2.$$

Alternative notation: S_X^2 *(The upper index* 2 *does not mean a square!)*

Its realization is denoted by a small case letter:

$$s_X^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x}_n)^2.$$

A single-pass formula is more practical:

$$S_X^2 = \frac{1}{n-1} \sum_{j=1}^{n} X_j^2 - \frac{n}{n-1} \bar{X}_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} X_j^2 - \frac{1}{n(n-1)} \left(\sum_{j=1}^{n} X_j \right)^2 .$$

Theorem:

$$\mu_{S_X^2} = \sigma_X^2.$$

Proof: From the single-pass formula for S_X^2:

$$\begin{aligned} \mu_{S_X^2} = \frac{n}{n-1} \mu_{X^2} - \frac{n}{n-1} \mu_{\bar{X}_n^2} &= \frac{n}{n-1} \left(\sigma_X^2 + \mu_X^2 - \sigma_{\bar{X}_n}^2 - \mu_{\bar{X}_n}^2 \right) \\ &= \frac{n}{n-1} \left(\sigma_X^2 + \mu_X^2 - \frac{1}{n} \sigma_X^2 - \mu_X^2 \right) = \sigma_X^2. \end{aligned}$$

Theorem: The sample variance is an unbiased consistent estimate of variance (provided that the original distribution has a variance and the fourth centered moment).

The distribution of the sample variance may be much more complex than the original one.

In particular, for the distribution $N(0,1)$ and $n = 2$:

$\bar{X} = \dfrac{X_1 + X_2}{2}$, $\qquad X_1 - \bar{X} = -(X_2 - \bar{X}) = \dfrac{X_1 - X_2}{2}$ has the distribution $N\left(0, \frac{1}{2}\right)$,

$$S_X^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = 2\left(\frac{X_1 - X_2}{2}\right)^2 = \left(\frac{X_1 - X_2}{\sqrt{2}}\right)^2 = U^2,$$

where $U = \frac{X_1 - X_2}{\sqrt{2}}$ has the distribution $N(0,1)$.

This is called the χ^2-distribution with 1 degree of freedom.

2.4.1 Distribution χ^2

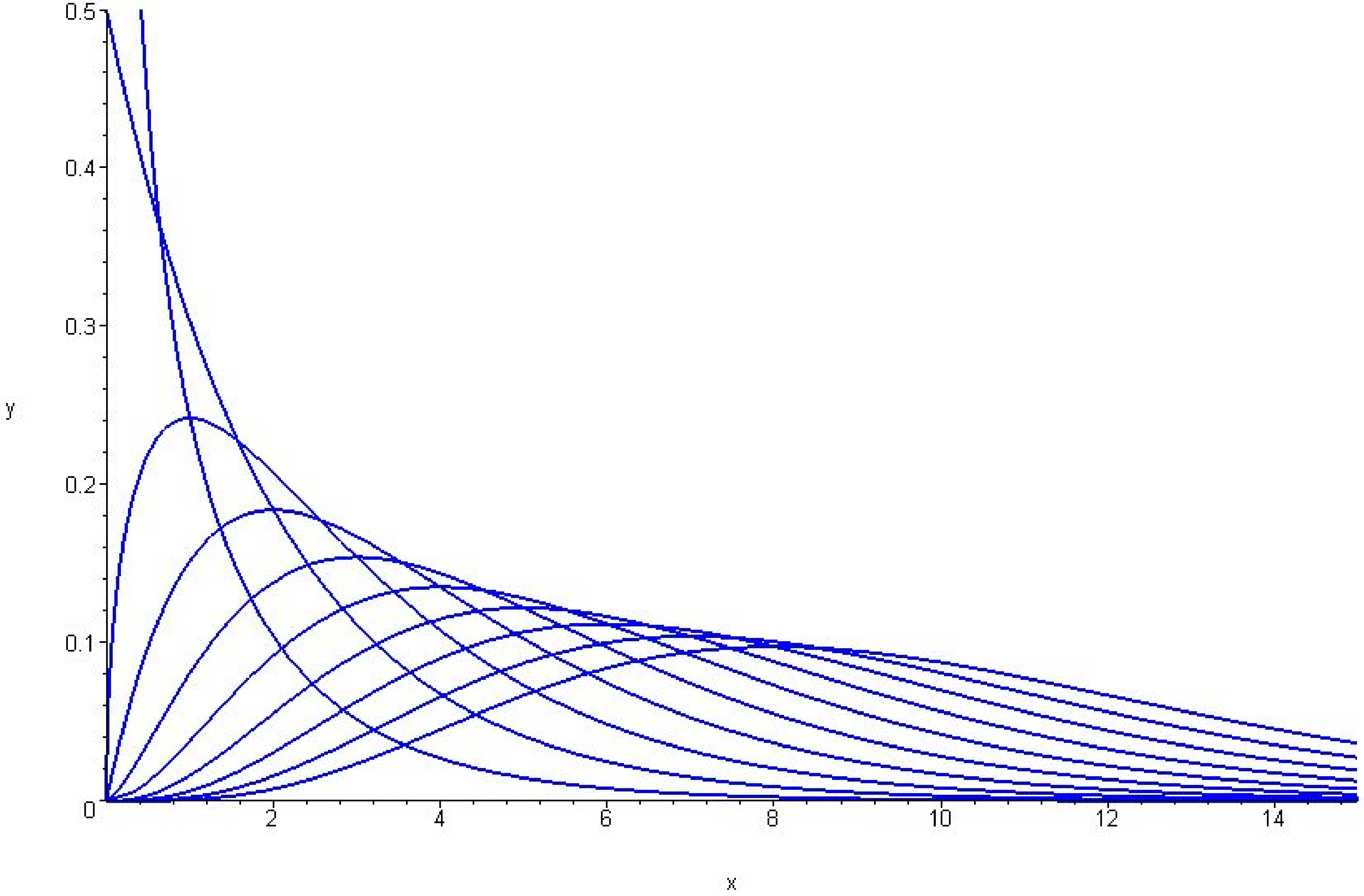
with η degrees of freedom is the distribution of the random variable $Y = \sum\limits_{j=1}^{\eta} U_j^2$, where U_j are **independent** random variables with the **normalized normal** distribution $N(0,1)$.

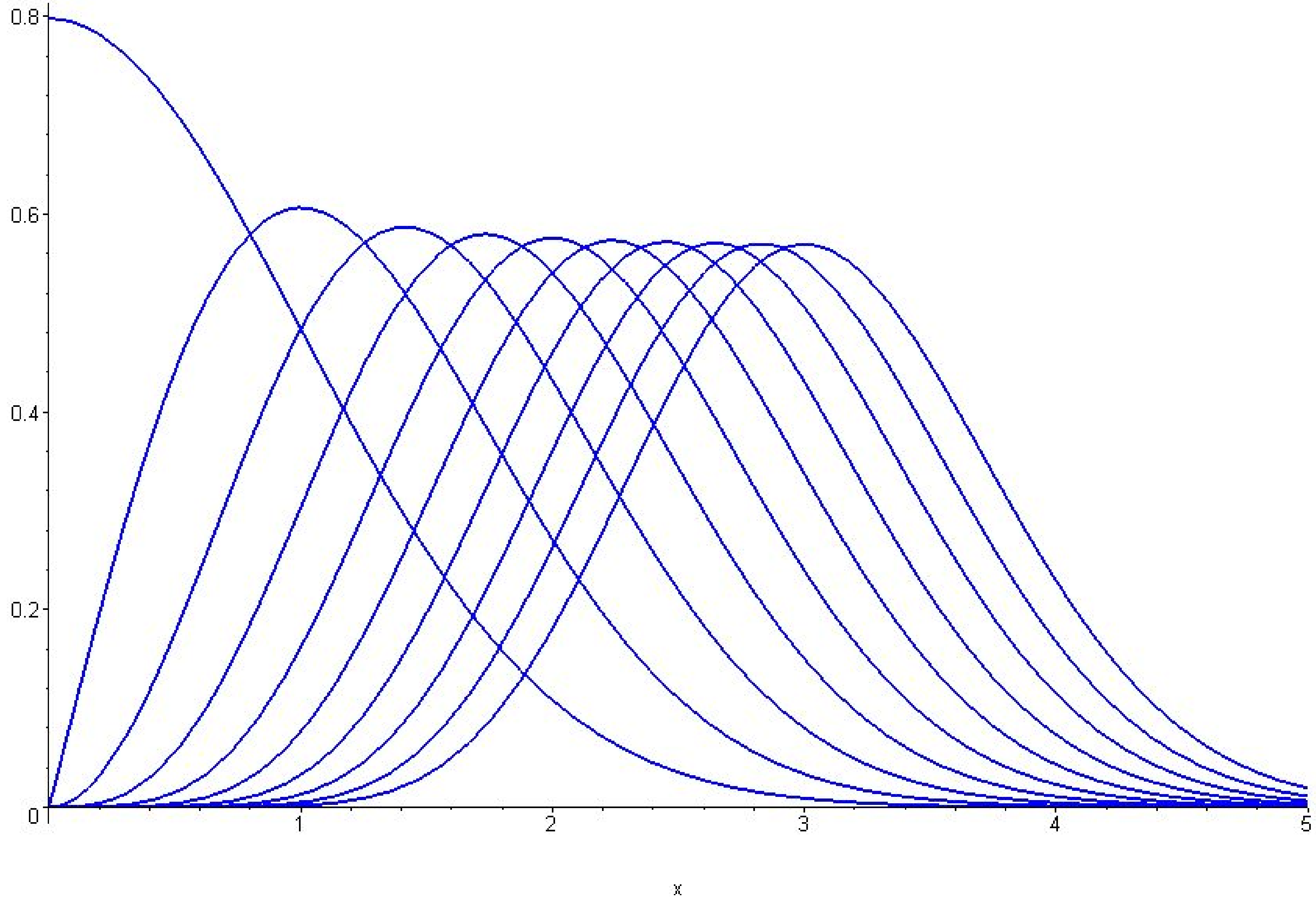
Notation: $\chi^2(\eta)$.

Its pdf is

$$f_Y(y) = \begin{cases} \dfrac{y^{\frac{\eta}{2}-1}\, e^{\frac{-y}{2}}}{2^{\frac{\eta}{2}}\, \Gamma\left(\frac{\eta}{2}\right)} & \text{for } y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(z) = \int\limits_0^{\infty} t^{z-1} e^{-t}\, dt$, in particular $\Gamma(n+1) = n!$ for all $n \in \mathbb{N}$.


0.5
0.4
0.3
0.2
0.1
0
y
2
4
6
8
10
12
14
x



Pdf's of χ^2 distributions with $1, 2, \ldots, 10$ degrees of freedom and their square roots ("distances from the target").

Theorem: Let X, Y be **independent** random variables with distributions $\chi^2(\xi), \chi^2(\eta)$, respectively. Then $X + Y$ has the distribution $\chi^2(\xi + \eta)$.

Theorem: A random variable Y with distribution χ^2 with η degrees of freedom satisfies

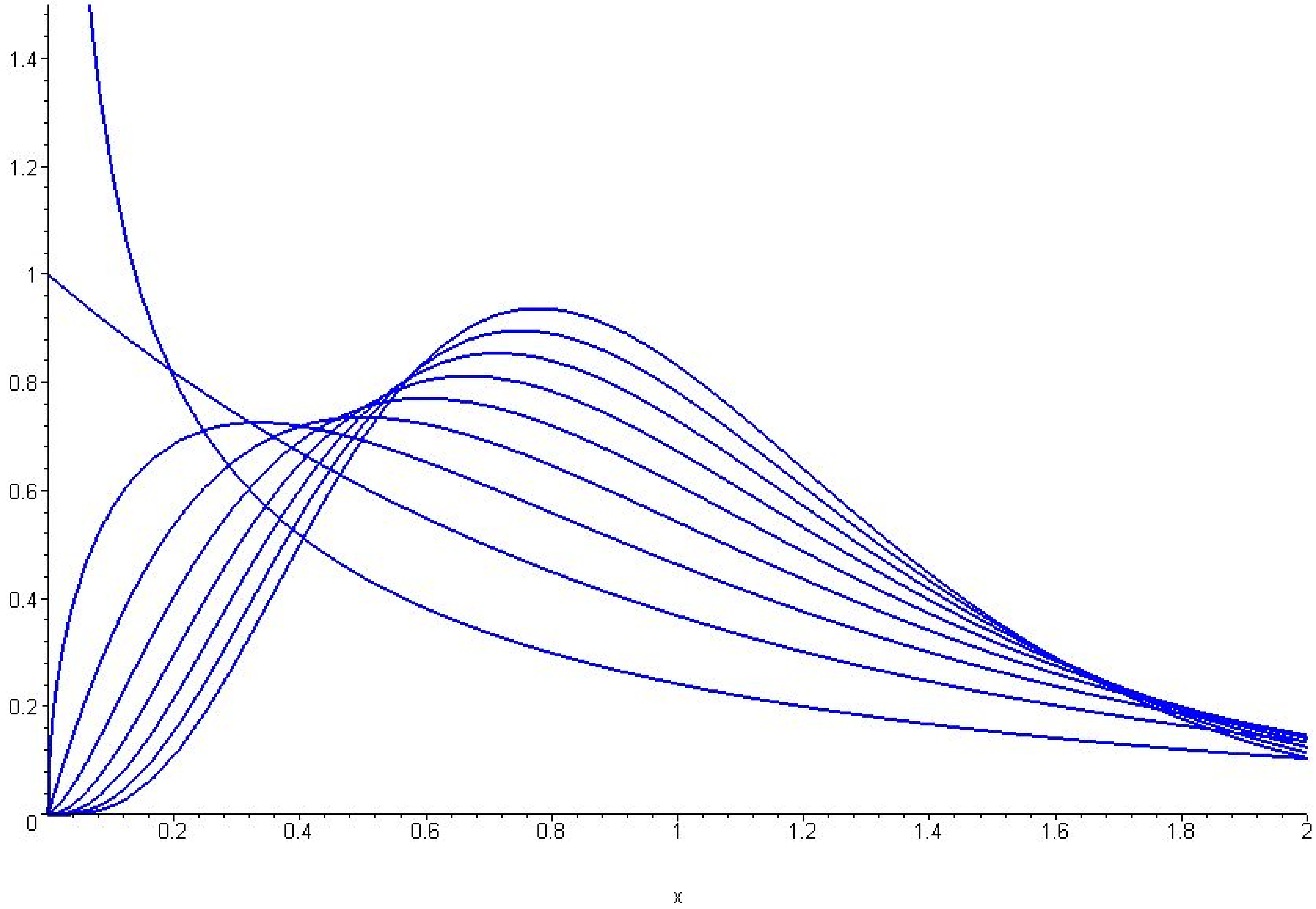
$$\mu_Y = \eta, \qquad \sigma_Y^2 = 2\eta.$$

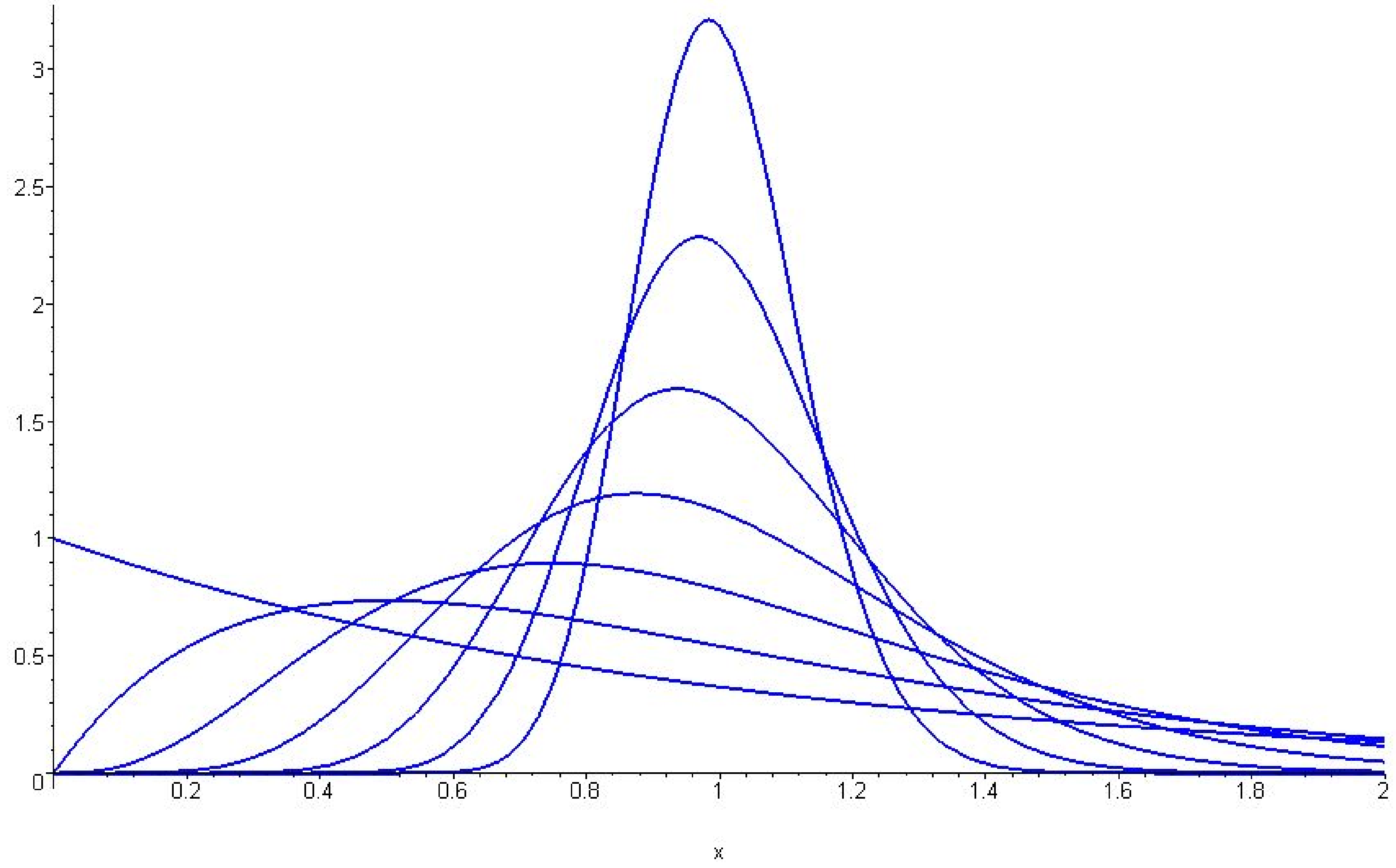
(We do not normalize this distribution.)

2.4.2 Sample variance

of a **normal** distribution $N(\mu_X, \sigma_X^2)$ satisfies:

$$\frac{(n-1)\ S_X^2}{\sigma_X^2} \text{ has distribution } \chi^2(n-1).$$


1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x



Pdf's of the distributions of sample variances from the normalized normal distribution with sample sizes $2, 3, \ldots, 10$ and $3 = 2^1 + 1, 2^2 + 1, \ldots, 2^7 + 1 = 129$.

Corollary: The variance of the sample variance is

$$\sigma^2_{S^2_X} = \frac{2}{n-1}\sigma^4_X,$$

where $\sigma^4_X = \left(\sigma^2_X\right)^2$.

Theorem: For a random sample $\vec{X} = (X_1, \ldots, X_n)$ from a **normal** distribution $\bar{X}$ is the best unbiased estimate of the mean, S^2_X is the best unbiased estimate of the variance, and statistics $\bar{X}, S^2_X$ are consistent and **independent**.

However, there is a biased estimate of variance which is more efficient:

2.4.3 Alternative estimate of variance

$$\widehat{\sigma^2_X} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2 = \frac{n-1}{n}S^2_X$$

Theorem: $\widehat{\sigma_X^2}$ is a biased consistent estimate of the variance.

Proof:

$$\mu_{\widehat{\sigma_X^2}} = \frac{n-1}{n}\sigma_X^2 \to \sigma_X^2,$$

$\widehat{\sigma_X^2}$ has a variance less than that of S_X^2, their proportion is $\left(\frac{n-1}{n}\right)^2$.

Efficiency cannot be compared in general; for a **normal** distribution:

1. efficiency of the estimate S_X^2:

$$\sigma^2_{S_X^2} = \frac{2}{n-1}\sigma_X^4$$

2. efficiency of the estimate $\widehat{\sigma_X^2}$:

$$\begin{aligned}\mu_{(\widehat{\sigma_X^2}-\sigma_X^2)^2} &= \sigma^2_{\widehat{\sigma_X^2}} + \left(\frac{1}{n}\sigma_X^2\right)^2 \\ &= \left(\frac{n-1}{n}\right)^2 \frac{2}{n-1}\sigma_X^4 + \frac{1}{n^2}\sigma_X^4 = \frac{2n-1}{n^2}\sigma_X^4.\end{aligned}$$

As

$$\frac{2n-1}{n^2} < \frac{2}{n} < \frac{2}{n-1},$$

the estimate $\widehat{\sigma_X^2}$ is more efficient than S_X^2 (which is the best unbiased one!).

2.5 Sample standard deviation

of a random sample $\vec{X} = (X_1, \ldots, X_n)$ is the statistic

$$S_X = \sqrt{S_X^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2}.$$

Alternative notation: S

Its realization is denoted by a small case letter:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2}.$$

Theorem:

$$\mu_{S_X} \leq \sigma_X .$$

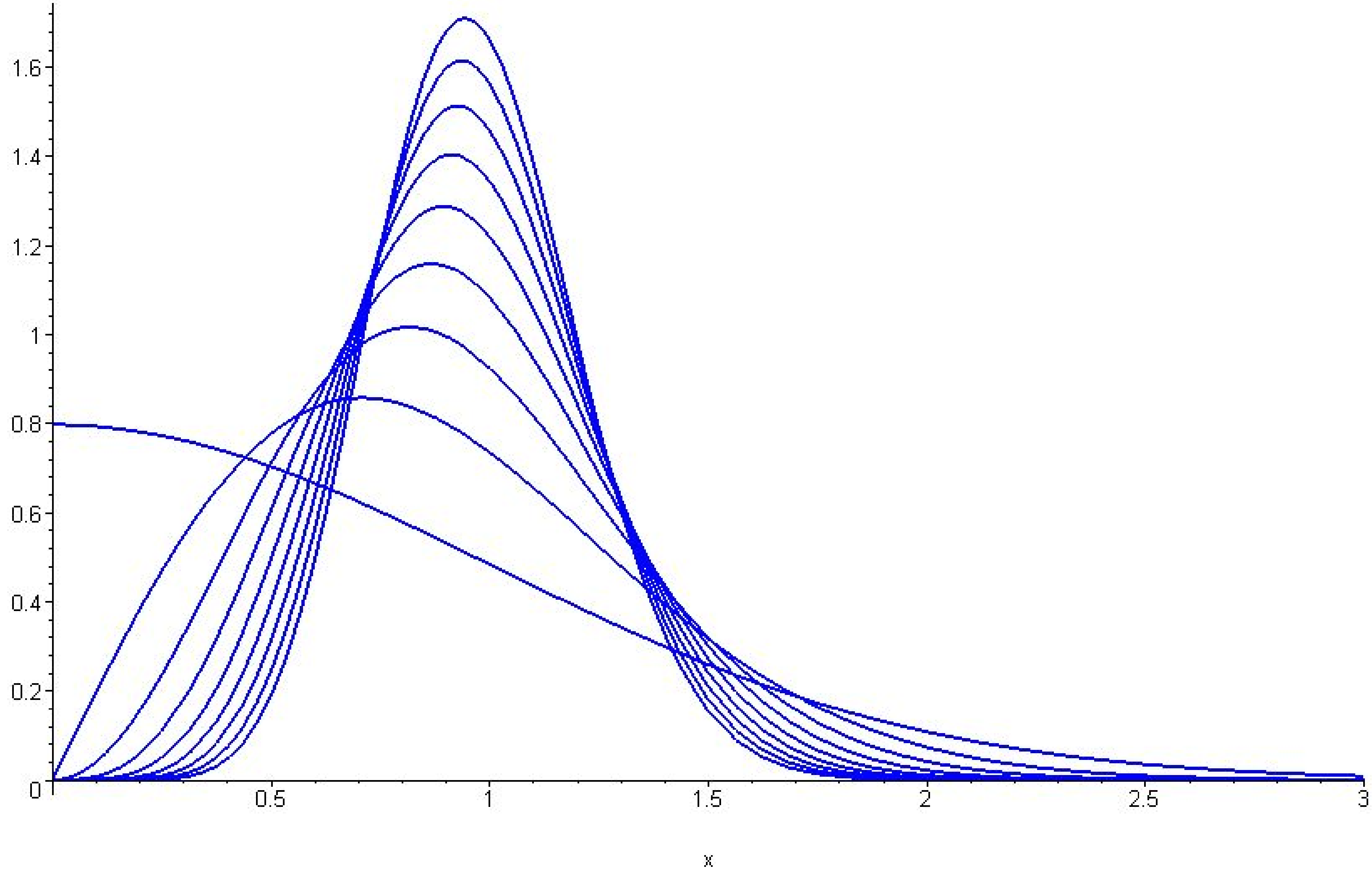
Equality does not hold if $\sigma_X > 0$, thus it is a **biased** estimate of the standard deviation!

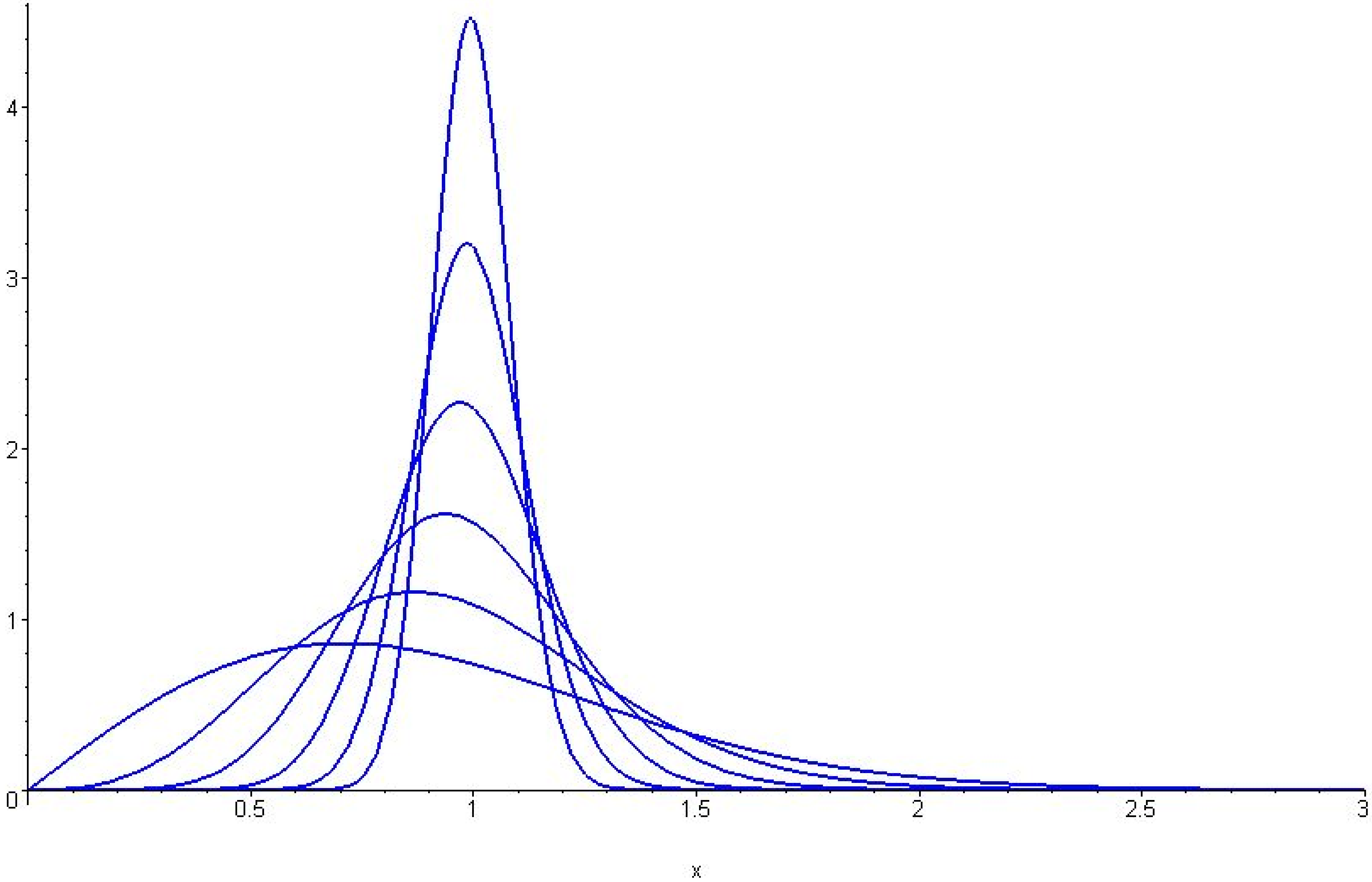
Proof:

$$\sigma_X^2 = \mu_{S_X^2} = \mu_{(S_X)^2} = \mu_{S_X}^2 + \underbrace{\sigma_{S_X}^2}_{\geq 0} \geq \mu_{S_X}^2$$

$$\sigma_X \geq \mu_{S_X}$$

Theorem: The sample standard deviation is a consistent estimate of the standard deviation (provided that the original distribution has a variance a 4th central moment).


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.5
1
1.5
2
2.5
3
x



Pdf's of the distributions of sample standard deviation from the normalized normal distribution with sample sizes $2, 3, \ldots, 10$ and $3 = 2^1 + 1, 2^2 + 1, \ldots, 2^6 + 1 = 65$.

2.6 Sample kth general moment

of a random sample $\vec{X} = (X_1, \ldots, X_n)$ is the statistic

$$M_{X^k} = \frac{1}{n} \sum_{j=1}^{n} {X_j}^k.$$

Alternative notation: M_k

Its realization is denoted by a small case letter:

$$m_{X^k} = \frac{1}{n} \sum_{j=1}^{n} {x_j}^k.$$

Theorem:

$$\mu_{M_{X^k}} = \mu_{X^k}.$$

(I.e., it is an **unbiased** estimate of the kth general moment.)

Theorem: The sample kth general moment is a consistent estimate of the kth general moment (provided that X has a kth and a $2k$th general moment).

Proof:

$$\sigma^2_{M_{X^k}} = \frac{1}{n^2} n \sigma^2_{X^k} = \frac{1}{n} \sigma^2_{X^k} = \frac{1}{n} \left(\mu_{(X^k)^2} - \mu^2_{X^k} \right) = \frac{1}{n} \left(\mu_{X^{2k}} - \mu^2_{X^k} \right).$$

2.7 Histogram and empirical distribution

Let us take a (non-random) vector $\vec{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ (obtained usually as a realization of a random sample). The order of entries is irrelevant, but their repetition is important. More economical representation of large vectors with a low number of different entries can be obtained by recording only the range $H = \{x_1, \dots, x_n\}$ and the **frequencies** n_t, $t \in H$. These data are usually represented by the **frequency table** or a graph called a **histogram**.

Normalization results in **relative frequencies** $r_t = \frac{n_t}{n}$, $t \in H$. As $\sum_{t \in H} r_t = 1$, they define a probability function $p_{Emp(\vec{x})}(t) = r_t$ of the so-called **empirical distribution** $Emp(\vec{x})$. It is a discrete distribution with at most n values which describes vector $\vec{x}$.

2.7.1 Properties of the empirical distribution

(Index $Emp(\vec{x})$ denotes parameters of any random variable with this distribution..)

$$\mu_{Emp(\vec{x})} = \sum_{t \in H} t\, r_t = \frac{1}{n} \sum_{t \in H} t\, n_t = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

$$\mu_{Emp(\vec{x})^k} = \sum_{t \in H} t^k\, r_t = \frac{1}{n} \sum_{t \in H} t^k\, n_t = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

$$\sigma^2_{Emp(\vec{x})} = \sum_{t \in H} \left(t - \mu_{Emp(\vec{x})}\right)^2 r_t = \frac{1}{n} \sum_{t \in H} (t - \bar{x})^2\, n_t = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n-1}{n} s^2_X$$

General moments of the empirical distribution equal the sample general moments of the original distribution.
Their computation from the histogram may be simpler than from the original realization of a random sample (if there are repeated values).

The variance of the empirical distribution is the alternative estimate of variance $\widehat{\sigma^2_X} = \frac{n-1}{n} S^2_X$, not the sample variance S^2_X.

2.8 Sample median

is the median of the empirical distribution, $Q_{Emp(\vec{x})}(\frac{1}{2})$. It gives us information different from the sample mean, often more useful (among others, it is more **robust** - resistent to noise and outliers). Besides, we know how it is mapped by a monotonic function.

Why it is used less than the sample mean:

- The computational complexity is higher; ordering of values requires time proportional to $n \ln n$, the sample mean only n.

- Memory requirements are higher – we need to remember all data, 2 registers suffice for the sample mean.

- The possibility of decentralization and parallel computation are very limited.

2.9 Interval estimates

So far, the actual value of a parameter, θ, has been estimated by a **point estimate** $\hat{\theta}$ (which is a random variable). Instead of it we search now for an **interval estimate**, so-called **confidence interval** I which is a minimal interval such that

$$P[\theta \in I] \geq 1 - \alpha,$$

where $\alpha \in (0, \frac{1}{2})$ is the probability of exceeding the bounds of the interval I; $1-\alpha$ is the **confidence coefficient.** Usually we search for an **upper**, resp. **lower one-sided** estimate,

$$I = (-\infty, Q_{\hat{\theta}}(1-\alpha)], \text{ resp. } I = [Q_{\hat{\theta}}(\alpha), \infty),$$

or the (**symmetric**) **two-sided** estimate,

$$I = \left[Q_{\hat{\theta}}\left(\frac{\alpha}{2}\right), Q_{\hat{\theta}}\left(1 - \frac{\alpha}{2}\right)\right].$$

For this, we need to know the distribution of the estimate $\hat{\theta}$.

2.10 Interval estimates of parameters of a normal distribution $N(\mu, \sigma^2)$

2.10.1 Estimate of the mean for a known variance σ^2

We estimate μ by the sample mean $\bar{X}$ with the distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$.

The normalized random variable $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)$ has the distribution $N(0,1) = N$;

$$
\begin{aligned}
&P\left[\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \in (-\infty, Q_N(1-\alpha)]\right] \\
&= 1 - \alpha \\
&= P\left[\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \le Q_N(1-\alpha)\right] \\
&= P\left[\mu \le \bar{X} + \frac{\sigma}{\sqrt{n}} Q_N(1-\alpha)\right] \\
&= P\left[\mu \in \left(-\infty, \bar{X} + \frac{\sigma}{\sqrt{n}} Q_N(1-\alpha)\right]\right].
\end{aligned}
$$

Similarly, other interval estimates are derived:

$$\left(-\infty, \bar{X} + \frac{\sigma}{\sqrt{n}} Q_N(1-\alpha)\right],$$

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} Q_N(1-\alpha), \infty\right),$$

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} Q_N\left(1-\frac{\alpha}{2}\right), \bar{X} + \frac{\sigma}{\sqrt{n}} Q_N\left(1-\frac{\alpha}{2}\right)\right],$$

where $\bar{X} - \frac{\sigma}{\sqrt{n}} Q_N(1-\alpha) = \bar{X} + \frac{\sigma}{\sqrt{n}} Q_N(\alpha)$
$(Q_N(\alpha) = -Q_N(1-\alpha)$ usually cannot be found in tables).

In applications, we replace the sample mean $\bar{X}$ by its realization $\bar{x}$.

2.10.2 Estimate of the mean for an unknown variance

We estimate μ by the sample mean $\bar{X}$ with the distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$,

σ^2 by the sample variance S^2; $\frac{(n-1)\,S^2}{\sigma^2}$ with the distribution $\chi^2(n-1)$.

We test the random variable $\frac{\sqrt{n}}{S}(\bar{X} - \mu)$ analogously; however, its distribution is not normal, although $\bar{X}, S$ are independent.

2.10.3 Student t-distribution [Gossett]

with η degrees of freedom is the distribution of the random variable

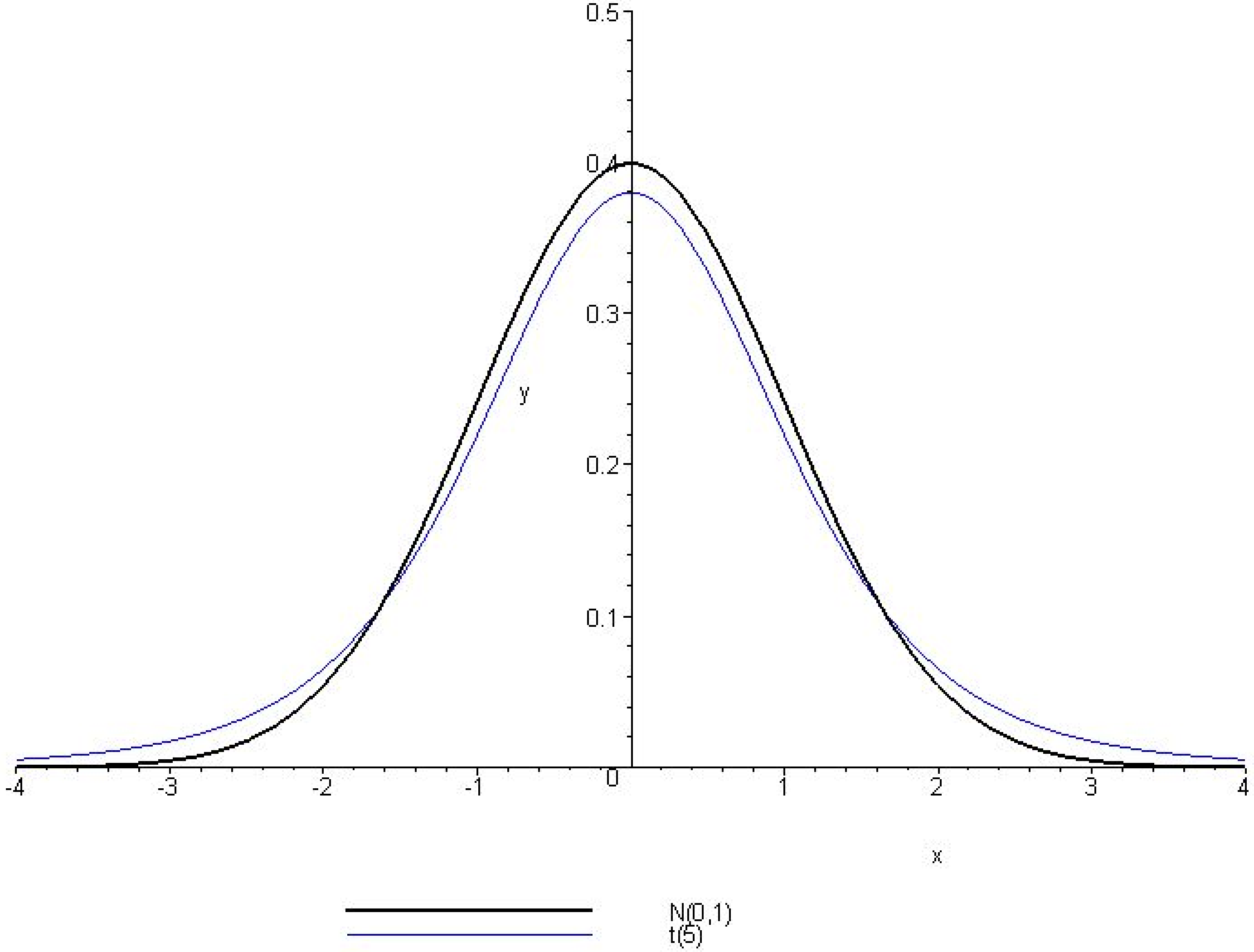
$$\frac{U}{\sqrt{\frac{V}{\eta}}},$$

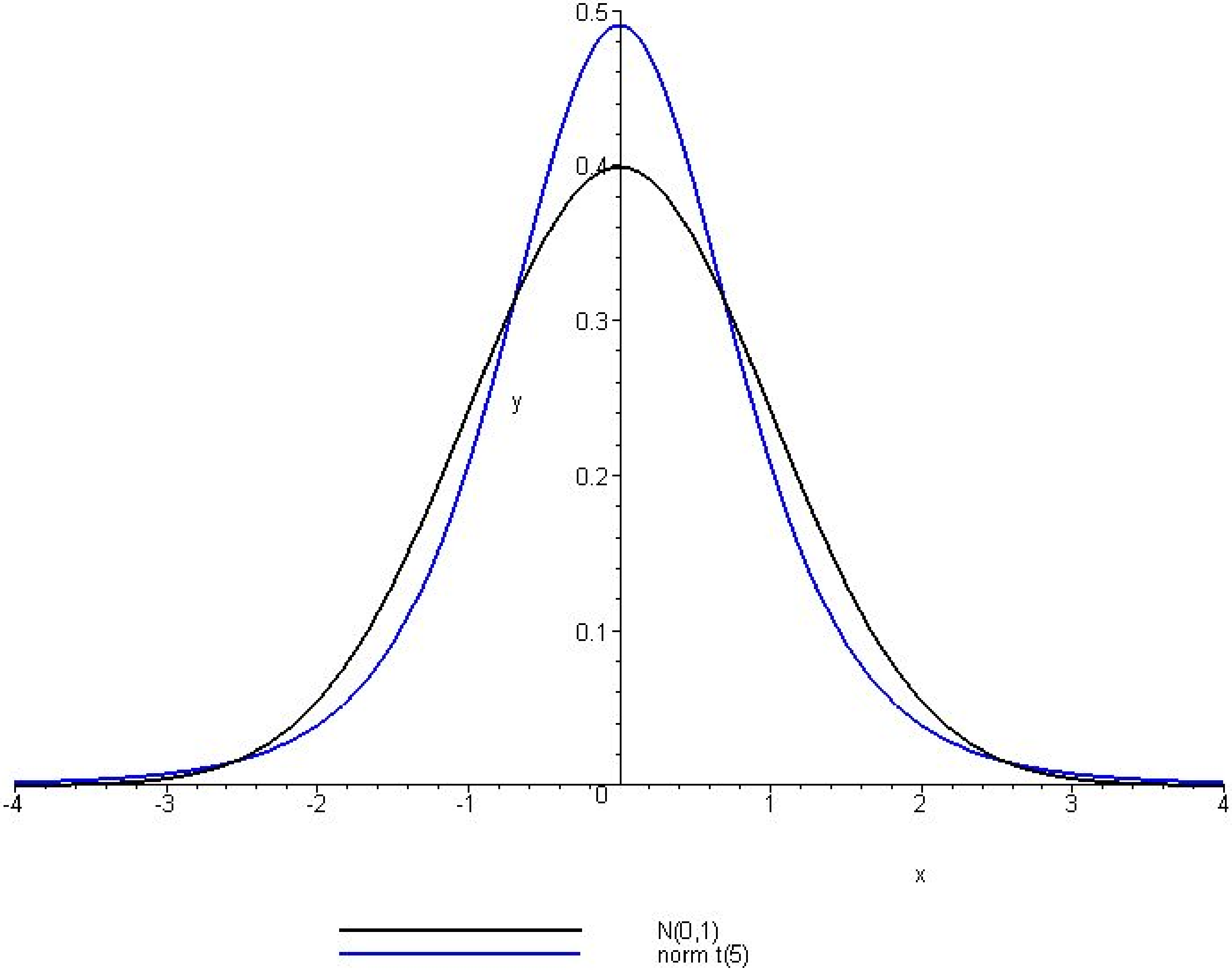
where U has the distribution $N(0, 1)$,
V has the distribution $\chi^2(\eta)$,
U, V are independent.

Notation: $t(\eta)$.

Symmetry w.r.t. zero $\implies Q_{t(\eta)}(1 - \alpha) = -Q_{t(\eta)}(\alpha)$

For a high number of degrees of freedom we replace it by a normal distribution.


0.5
0.4
0.3
0.2
0.1
y
-4
-3
-2
-1
0
1
2
3
4
x
N(0,1)
t(5)



Pdf's of the normalized normal distribution and the Student distribution with 5 degrees of freedom (original and normalized).

2.10.4 Estimate of the mean for an unknown variance II

In our case:

$$\begin{aligned} U &= \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) \text{ has } N(0,1), \\ V &= \frac{(n-1)\,S^2}{\sigma^2} \text{ has } \chi^2(n-1),\ \ \eta = n-1, \\ \frac{U}{\sqrt{\frac{V}{\eta}}} = \frac{\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu)}{\sqrt{\frac{S^2}{\sigma^2}}} &= \frac{\sqrt{n}}{S}(\bar{X}-\mu) \text{ has } t(n-1). \end{aligned}$$

This induces interval estimates

$$\left(-\infty, \bar{X} + \frac{S}{\sqrt{n}} Q_{t(n-1)}(1-\alpha)\right],$$

$$\left[\bar{X} - \frac{S}{\sqrt{n}} Q_{t(n-1)}(1-\alpha), \infty\right),$$

$$\left[\bar{X} - \frac{S}{\sqrt{n}} Q_{t(n-1)}(1-\frac{\alpha}{2}), \bar{X} + \frac{S}{\sqrt{n}} Q_{t(n-1)}(1-\frac{\alpha}{2})\right].$$

In applications, we replace the sample mean $\bar{X}$ by its realization $\bar{x}$ and the sample standard deviation S_X by its realization s_X.

2.10.5 Estimate of a variance

We estimate σ^2 by the sample variance S^2; $\frac{(n-1)S^2}{\sigma^2}$ has the distribution $\chi^2(n-1)$;

$$\begin{aligned}
&P\left[\frac{(n-1)\,S_X^2}{\sigma^2} \in (-\infty, Q_{\chi^2(n-1)}(1-\alpha)]\right] \\
&= 1-\alpha \\
&= P\left[\frac{(n-1)\,S_X^2}{\sigma^2} \leq Q_{\chi^2(n-1)}(1-\alpha)\right] \\
&= P\left[\frac{(n-1)\,S_X^2}{Q_{\chi^2(n-1)}(1-\alpha)} \leq \sigma^2\right] \\
&= P\left[\sigma^2 \in \left[\frac{(n-1)\,S_X^2}{Q_{\chi^2(n-1)}(1-\alpha)}, \infty\right)\right].
\end{aligned}$$

We obtained a lower estimate.

Similarly, other interval estimates are derived:

$$\left(-\infty, \frac{(n-1)\,S^2}{Q_{\chi^2(n-1)}(\alpha)}\right],$$

$$\left[\frac{(n-1)\,S^2}{Q_{\chi^2(n-1)}(1-\alpha)}, \infty\right),$$

$$\left[\frac{(n-1)\,S^2}{Q_{\chi^2(n-1)}\left(1-\frac{\alpha}{2}\right)}, \frac{(n-1)\,S^2}{Q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right)}\right].$$

In applications, we replace the sample variance S_X^2 by its realization s_X^2.

2.10.6 Interval estimates of continuous distributions which are not normal

can be converted to a normal distribution by a non-linear transformation

$$t \mapsto F_N^{-1}(F_X(t)) = Q_N(F_X(t))$$

($F_X(X)$ has the uniform distribution on $[0,1]$).

2.11 Parameters estimation

The distribution of X depends on a vector of parameters, $\mathbf{\Theta} = (\theta_1, \ldots, \theta_i) \in \mathbf{\Omega}$, where $\mathbf{\Omega} \subseteq \mathbb{R}^i$ is the **parameter space**, i.e., the set of all possible values of parameters; we denote the probability function by $p_X(t|\Theta)$, etc.

The task is to find an estimate $\hat{\Theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_i)$ using a realization $\vec{x} = (x_1, \ldots, x_n)$.

2.11.1 Method of moments

For $k = 1, 2, \ldots$, the kth general moment is a function of Θ,

$$\mu_{X^k}(\Theta) = \mu_{X^k}(\theta_1, \ldots, \theta_i)$$

(which can be computed from the probability model).
It can be estimated by the realization of the sample general moment, m_{X^k}.

The method of moments recommends the estimate $(\hat{\theta}_1, \ldots, \hat{\theta}_i)$ such that

$$\mu_{X^k}(\hat{\theta}_1, \ldots, \hat{\theta}_i) = m_{X^k} = \frac{1}{n} \sum_{j=1}^{n} {x_j}^k, \qquad k = 1, 2, \ldots.$$

To achieve a unique solution for i variables, we usually need (the first) i equations for $k = 1, 2, \ldots, i$.

Applicability of method of moments

Possible problems:

1. There is no solution $\Longrightarrow$ try a reduction of the number of equations.

2. There are infinitely many solutions $\Longrightarrow$ try to include more equations.

3. There is more than one solution (e.g., of a system of quadratic equations).

4. There is a unique solution, but it is hard to find it.

5. The system is ill-conditioned (typically for a large number of parameters).

6. We have found a unique solution, but it **does not satisfy the assumptions** of the model, $\boldsymbol{\Theta} \notin \boldsymbol{\Omega}$ (e.g., the parameters cannot be arbitrary numbers) $\Longrightarrow$ NO HOPE! **Always check the solution!**

7. All equations are considered equally important; this is sometimes strange (typically for a large number of parameters).

8. Non-numerical data are excluded (except for those which can be reasonably numbered).

Advantage:

1. Applicable to discrete, continuous, and **mixed** distributions without any change.

2.11.2 Method of maximum likelihood

For a discrete distribution

The probability of the realization,

$$\begin{aligned} p_{\vec{X}}(\vec{x}|\Theta) &= P\left[X_1 = x_1 \wedge \ldots \wedge X_n = x_n|\Theta\right] \\ &= \prod_{j=1}^{n} P\left[X_j = x_j|\Theta\right] = \prod_{j=1}^{n} p_X(x_j|\Theta) = \ell(\Theta), \end{aligned}$$

is a function $\ell\colon \mathbf{\Omega} \to [0,1]$, $\mathbf{\Omega} \subseteq \mathbb{R}^i$, of $\Theta = (\theta_1, \ldots, \theta_i)$ called the **likelihood of a discrete distribution**. We maximize it or rather the **log-likelihood**,

$$L(\Theta) = \ln \ell(\Theta) = \sum_{j=1}^{n} \ln p_X(x_j|\Theta).$$

(The case $p_X(x_j|\Theta) = 0$ has to be excluded, but it does not correspond to the maximum likelihood.)

Example: The empirical distribution is the maximum likelihood estimate of discrete distribution (if it is not restricted by additional conditions).

For a continuous distribution

Here each realization has a zero probability; instead of it, we use the pdf, but we obtain a completely **different notion,**

$$f_{\vec{X}}(\vec{x}|\Theta) = \prod_{j=1}^{n} f_X(x_j|\Theta) = \ell(\Theta).$$

Nevertheless, also this function $\ell\colon \mathbf{\Omega} \to [0, \infty)$, $\mathbf{\Omega} \subseteq \mathbb{R}^i$, is called the **likelihood of a continuous distribution** and

$$L(\Theta) = \ln \ell(\Theta) = \sum_{j=1}^{n} \ln f_X(x_j|\Theta)$$

the **log-likelihood**.

(The case $f_X(x_j|\Theta) = 0$ has to be excluded, but it does not correspond to the maximum likelihood.)

For a mixed distribution

Undefined!

Applicability of method of maximum likelihood

Possible problems:

1. There is more than one solution. (Possibly different values of parameters lead to the same distribution – does this mind?)

2. The solution need not exist. (This may happen only if the likelihood is not continuous or the parameter space is not closed.)

3. There is a unique solution, but it is hard to find it (local maxima need not be global).

4. The task is ill-conditioned.

5. The values of likelihood may be very small.

6. **It cannot be used to mixed distributions!**

Advantages:

1. It is easier “not to get lost” when looking for and optimum rather than a solution of a system of equations.

2. It respects the meaning of data of different nature.

3. It can be used also to non-numerical data.

3 Tests of hypotheses

3.1 Basic notions and principles

We test a hypothesis about the value of a parameter θ of a distribution (using a **criterion** or **test statistic** T).

Example: Parameter θ achieves only 2 values, 0 for “normal” population, 1 for “anomal” elements. Both classes have known distributions with different means μ_0, μ_1, where $\mu_0 < \mu_1$. Random sample $\vec{X}$ is made from one of the classes, we have to guess which one. For that, we use (not necessarily) $T = \bar{X}$ as an estimate of the mean. We choose $c \in (\mu_0, \mu_1)$ and classify the sample by 0 for $T \leq c$, 1 for $T > c$. Two types of errors are possible:

1. class 0 is classified by 1, with probability $\alpha(c)$ (non-increasing function of c),

2. class 1 is classified by 0, with probability $\beta(c)$, (non-decreasing function of c).

Possible criteria for the choice of the bound c:

- $\alpha(c) = \beta(c)$,
- $\min\limits_{c}(\alpha(c) + \beta(c))$,
- $\min\limits_{c} e(\alpha(c), \beta(c))$, e.g., $\min\limits_{c}(a\alpha(c) + b\beta(c))$, i.e., minimization of a **payment function**,
- $\alpha(c) =$ a small value fixed in advance.

Usually the latter option is taken because of

- technical reasons (easier task),
- we do not need the distribution of the anomal class,

- usually the task is complicated by allowing more than two values of the parameter.

Example: Should we stop the distribution of a medicine because of suspected undesirable side-effects?

Null hypothesis H_0: The producer is innocent, the risk does not increase.

Alternative hypothesis H_1: The producer is guilty, the risk increases.

Beside good decisions we risk:

Type I error: We reject a valid null hypothesis (we accuse an innocent).

Type II error: We do not reject an invalid null hypothesis (we dismiss a culprit).

By the choice of the bound we decrease the risk of one error and increase the other.

Accepted solution: A **critical value** c of a test is taken so that the risk of type I error is (less than or equal to) a given small probability $\alpha \in \mathbb{R}$ called a **significance level**.

Usually 1% or 5% is used (always $\alpha \ll \frac{1}{2}$).

Values of the criterion exceeding the critical value (which correspond to results not much probable under the assumption of the null hypothesis) are considered **statistically significant** and then we **reject the null hypothesis**.

In the opposite case, we **accept (=do not reject) the null hypothesis**, but also we **do not confirm it** becuase this could cause a type II error with an unknown probability β.

The **power of a test** is evaluated according to $1 - \beta$, i.e., the risk of type II error for a given risk of type I error.

The following notions are distinguished in the literature:

- a **simple hypothesis**: the null hypothesis corresponds to a single value of the parameter,

- a **compound hypothesis**: the null hypothesis corresponds to more values of the parameter,

and also

- a **simple alternative**: the alternative hypothesis corresponds to a single value of the parameter,

- a **compound alternative**: the alternative hypothesis corresponds to more values of the parameter.

Often the null and alternative hypotheses are formulated so that they do not cover all possible cases. It is safer to avoid this case by taking for the null hypothesis the negation of the alternative hypothesis.

E.g., if $H_1 : \theta > c$, we do not take $H_0 : \theta = c$ but $H_0 : \theta \le c$. (The highest risk of type I error usually occurs for $\theta = c$, hence the procedure is the same.)

For a compound hypothesis we require that the risk of type I error is at most α for all values of the parameter satisfying the null hypothesis.

(Statistical significance does not imply practical significance.)

Solution: The null hypothesis is rejected if and only if the value of the criterion does not belong to the $(1-\alpha)$-confidence interval. Thus the critical value is the bound of the confidence interval.

Conversely, one may ask about the **achieved significance (P)**, i.e., the significance level for which the critical value equals the observed value of the criterion. **(The lower value, the more significant result.)** This is the usual output of a program; it can be compared to any desired significance level and, moreover, it gives additional information about the distance from the critical value.

Typical form of a test: Test statistics T with a known distribution (more exactly, its realization, t) is compared to the quantile of the respective distribution and the null hypothesis is rejected for extreme values (low probable when the null hypothesis holds):

H_0	H_1	rejected for	significance
$\theta \leq c$	$\theta > c$	$t > Q_T(1-\alpha)$	$1 - F_T(t)$
$\theta \geq c$	$\theta < c$	$t < Q_T(\alpha)$	$F_T(t)$
$\theta = c$	$\theta \neq c$	$t > Q_T(1-\frac{\alpha}{2})$ or $t < Q_T(\frac{\alpha}{2})$	$2\min\left(F_T(t), 1 - F_T(t)\right)$

The following combinations of a null and an alternative hypotheses is encountered in the literature:

H_0	H_1
$\theta = c$	$\theta > c$
$\theta = c$	$\theta < c$

They are solved the same way as the first two cases mentioned above.

3.2 Tests of mean of normal distribution

3.2.1 For known variance σ^2

$$T := \frac{\bar{X} - c}{\sigma}\sqrt{n}$$

is compared to quantiles of the normalized normal distribution:

H_0	is rejected for	achieved significance
$\mu \leq c$	$t > Q_N(1-\alpha)$	$1 - F_N(t)$
$\mu \geq c$	$t < -Q_N(1-\alpha) = Q_N(\alpha)$	$F_N(t)$
$\mu = c$	$\lvert t\rvert > Q_N(1-\frac{\alpha}{2})$	$2\min\left(F_N(t), 1 - F_N(t)\right)$

3.2.2 For unknown variance

$$T := \frac{\bar{X} - c}{S_X}\sqrt{n}$$

is compared to quantiles of the **Student distribution** with $n-1$ degrees of freedom:

H_0	is rejected for	achieved significance
$\mu \leq c$	$t > Q_{t(n-1)}(1-\alpha)$	$1 - F_{t(n-1)}(t)$
$\mu \geq c$	$t < -Q_{t(n-1)}(1-\alpha)$	$F_{t(n-1)}(t)$
$\mu = c$	$\lvert t\rvert > Q_{t(n-1)}(1-\frac{\alpha}{2})$	$2\min\Big(F_{t(n-1)}(t), 1 - F_{t(n-1)}(t)\Big)$

3.3 Tests of variance of normal distribution

$$T := \frac{(n-1)\, S_X^2}{c}$$

is compared to quantiles of the χ^2**-distribution** with $n-1$ degrees of freedom:

H_0	is rejected for	achieved significance
$\sigma^2 \leq c$	$t > Q_{\chi^2(n-1)}(1-\alpha)$	$1 - F_{\chi^2(n-1)}(t)$
$\sigma^2 \geq c$	$t < Q_{\chi^2(n-1)}(\alpha)$	$F_{\chi^2(n-1)}(t)$
$\sigma^2 = c$	$t < Q_{\chi^2(n-1)}(\frac{\alpha}{2})$ or $t > Q_{\chi^2(n-1)}(1-\frac{\alpha}{2})$	$2 \min \left(F_{\chi^2(n-1)}(t), 1 - F_{\chi^2(n-1)}(t)\right)$

3.4 Comparison of two normal distributions

Assumption: **Independent** samples

$$(X_1, \ldots, X_m) \text{ from } N(\mu_X, \sigma_X^2),$$
$$(Y_1, \ldots, Y_n) \text{ from } N(\mu_Y, \sigma_Y^2).$$

3.4.1 Tests of variances of two normal distributions [Fisher]

If $\sigma_X^2 = \sigma_Y^2$, then $S_X^2 \approx S_Y^2$. The test statistic is chosen as

$$T := \frac{S_X^2}{S_Y^2}.$$

The **F-distribution (Fisher–Snedecor distribution) with ξ and η degrees of freedom** is the distribution of a random variable

$$F = \frac{\frac{U}{\xi}}{\frac{V}{\eta}},$$

where U, V are **independent** random variables with distributions $\chi^2(\xi), \chi^2(\eta)$, respectively.

Notation: $F(\xi, \eta)$

If $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, then we put

$$U := \frac{(m-1)S_X^2}{\sigma^2} \text{ has } \chi^2(m-1),$$
$$V := \frac{(n-1)S_Y^2}{\sigma^2} \text{ has } \chi^2(n-1),$$
$$\xi := m-1,\ \eta := n-1,$$
$$F = \frac{\frac{U}{\xi}}{\frac{V}{\eta}} = \frac{\frac{(m-1)S_X^2}{(m-1)\sigma^2}}{\frac{(n-1)S_Y^2}{(n-1)\sigma^2}} = \frac{S_X^2}{S_Y^2} = T.$$

We test T for the distribution $F(m-1,n-1)$:

H_0	is rejected for	achieved significance
$\sigma_X^2 \le \sigma_Y^2$	$t > Q_{F(m-1,n-1)}(1-\alpha)$	$1 - F_{F(m-1,n-1)}(t)$
$\sigma_X^2 \ge \sigma_Y^2$	$t < Q_{F(m-1,n-1)}(\alpha)$	$F_{F(m-1,n-1)}(t)$
$\sigma_X^2 = \sigma_Y^2$	$t < Q_{F(m-1,n-1)}(\frac{\alpha}{2})$ or $t > Q_{F(m-1,n-1)}(1-\frac{\alpha}{2})$	$2\min(F_{F(m-1,n-1)}(t),$ $1 - F_{F(m-1,n-1)}(t))$

For each significance level, we need a two-dimensional table of quantiles indexed by ξ, η; usually only one half is tabulated, for the second half we need the formula

$$Q_{F(\xi,\eta)}(\beta) = \frac{1}{Q_{F(\eta,\xi)}(1-\beta)}$$

(notice the reverse order of the degrees of freedom!) or we have to consider $\frac{S_Y^2}{S_X^2}$ instead of $\frac{S_X^2}{S_Y^2}$.

3.4.2 Tests of means of two normal distributions with known variance σ^2

$$
\begin{aligned}
\bar{X}_m \text{ has } N\left(\mu_X, \frac{\sigma^2}{m}\right), \\
\bar{Y}_n \text{ has } N\left(\mu_Y, \frac{\sigma^2}{n}\right), \\
\bar{X}_m - \bar{Y}_n \text{ has } N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right).
\end{aligned}
$$

Under the assumption $\mu_X = \mu_Y$,

$$
T := \frac{\bar{X}_m - \bar{Y}_n}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ has } N\left(0, 1\right).
$$

We test T for $N(0,1)$ (see Section 3.2.1).

3.4.3 Tests of means of two normal distributions with (the same) unknown variance

Assumption: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

As the first step, we have to verify this assumption (see Section 3.4.1).

(In fact, we cannot verify it; we try to reject it and if this fails, we continue. Otherwise, the test becomes more complicated, see [Mood et al.].)

We have two estimates S_X^2, S_Y^2 of the same value σ^2; we take their average weighted by the sample sizes (minus 1 for one degree of freedom lost by computing the sample mean):

$$\begin{aligned} \frac{(m-1)S_X^2}{\sigma^2} & \text{ has } \chi^2(m-1), \\ \frac{(n-1)S_Y^2}{\sigma^2} & \text{ has } \chi^2(n-1), \\ \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} & \text{ has } \chi^2(m+n-2), \end{aligned}$$

its mean is $m+n-2$, the mean of

$$\frac{(m-1)S_X^2+(n-1)S_Y^2}{(m+n-2)\,\sigma^2}=\frac{S^2}{\sigma^2}$$

is 1 and

$$S^2 := \frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}$$

is an unbiased estimate of σ^2,

$$S := \sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}}.$$

$$\begin{aligned}
\bar{X}_m \text{ has } N\left(\mu_X, \frac{\sigma^2}{m}\right),\\
\bar{Y}_n \text{ has } N\left(\mu_Y, \frac{\sigma^2}{n}\right),\\
\bar{X}_m - \bar{Y}_n \text{ has } N\left(\mu_X-\mu_Y, \sigma^2\left(\frac{1}{m}+\frac{1}{n}\right)\right).
\end{aligned}$$

Under the hypothesis $\mu_X = \mu_Y$,

$$\frac{\bar{X}_m - \bar{Y}_n}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}} \text{ has } N\left(0,1\right),$$

$$\frac{(m+n-2)S^2}{\sigma^2} = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \text{ has } \chi^2(m+n-2),$$

$$T := \frac{\bar{X}_m - \bar{Y}_n}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{\frac{\bar{X}_m - \bar{Y}_n}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}} \text{ has } t(m+n-2).$$

We test T for $t(m+n-2)$ (see Section 3.2.2).

3.5 Tests of means of two normal distributions for paired samples

(according to M.I. Schlesinger)

Example: Compare the mean temperatures at two places.
The standard test of means of two normal distributions is very weak because of high variance; however, the changes are almost synchronized. Thus the two samples **are not mutually independent**. We always measure both variables at the same time.

Assumption: Random variables X_j, Y_j $(j = 1, \ldots, n)$ have normal distributions $N(\mu_j, \sigma^2)$ with a constant variance σ^2 and variable means $\mu_j = \mu_{X_j} = \mu_{Y_j}$.

We may use variables $U_j := X_j - \mu_j, V_j := Y_j - \mu_j$ $(j = 1, \ldots, n)$ which **are independent** and have the distribution $N(0, \sigma^2)$.

Random variables $\Delta_j := X_j - Y_j = U_j - V_j$ $(j = 1, \ldots, n)$ are independent and have the distribution $N(0, 2\sigma^2)$.

The sample mean $\bar{\Delta}$ has $N\left(0, \frac{2\sigma^2}{n}\right)$.

3.5.1 For known variance σ^2

The unknown parameters of the joint distribution are $\mu_1, \ldots, \mu_n$, but we do not need them.

Following Section 3.2.1 (with $c = 0$), we test

$$T := \frac{\bar{\Delta}}{\sigma}\sqrt{\frac{n}{2}} = \frac{\bar{X} - \bar{Y}}{\sigma}\sqrt{\frac{n}{2}}$$

for $N(0, 1)$.

3.5.2 For unknown variance

The unknown parameters of the joint distribution are $\Theta = (\sigma^2, \mu_1, \ldots, \mu_n)$, but we need to estimate only $\sigma^2 = \sigma_X^2$.

Instead of it, we can work directly with the sample $(\Delta_1, \dots, \Delta_n)$ from a normal distribution.

Following Section 3.2.2 (with $c = 0$), we test

$$T := \frac{\bar{\Delta}}{S_\Delta}\sqrt{n}$$

for $t(n-1)$.

Exercise: Maximum likelihood estimate of the parameters:

$$\begin{aligned}
\ell(\Theta) &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_j-\mu_j)^2}{2\sigma^2}\right) \cdot \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_j-\mu_j)^2}{2\sigma^2}\right), \\
L(\Theta) &= -\sum_{j=1}^{n} \frac{(x_j-\mu_j)^2}{2\sigma^2} - \sum_{j=1}^{n} \frac{(y_j-\mu_j)^2}{2\sigma^2} - 2n \ln \sigma - 2n \ln \sqrt{2\pi}, \\
0 = \frac{\partial L(\hat{\Theta})}{\partial \hat{\mu}_j} &= \frac{\partial}{\partial \hat{\mu}_j}\left(-\frac{(x_j-\hat{\mu}_j)^2}{2\hat{\sigma}^2} - \frac{(y_j-\hat{\mu}_j)^2}{2\hat{\sigma}^2}\right) \\
&= \frac{1}{\hat{\sigma}^2}\Big((x_j-\hat{\mu}_j) + (y_j-\hat{\mu}_j)\Big) = \frac{1}{\hat{\sigma}^2}\left(x_j + y_j - 2\hat{\mu}_j\right), \\
\hat{\mu}_j &= \frac{x_j+y_j}{2}, \qquad j = 1, \dots, n.
\end{aligned}$$

The estimates $\hat{\mu}_j$ $(j = 1, \ldots, n)$ *are* **not consistent**.

We substitute them:

$$L(\hat{\Theta}) = -\sum_{j=1}^{n} \frac{(x_j - y_j)^2}{4\hat{\sigma}^2} - n \ln \hat{\sigma}^2 - 2n \ln \sqrt{2\pi},$$

$$0 = \frac{\partial L(\hat{\Theta})}{\partial \left(\hat{\sigma}^2\right)} = \sum_{j=1}^{n} \frac{(x_j - y_j)^2}{4\hat{\sigma}^4} - \frac{2n}{\hat{\sigma}^2},$$

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{j=1}^{n} (x_j - y_j)^2 = \frac{1}{2n} \sum_{j=1}^{n} \delta_j{}^2,$$

where δ_j is the realization of Δ_j. *The estimate* $\hat{\sigma}^2$ **is consistent**.

3.6 χ^2 goodness-of-fit test

We test a hypothesis that a random variable has some distribution. As we can only reject hypotheses, we shall never confirm that it is really so.

We test a **discrete distribution** (which could be a discretization of a continuous distribution).

H_0 : The random variable has a discrete distribution to k classes with probabilities $p_1, \ldots, p_k$.

Our test is based on a random sample of size n. We do not need the order of results, only their **frequencies** n_i, resp. **relative frequencies** $\frac{n_i}{n}$ $(i = 1, \ldots, k)$. We compare the frequency n_i to the **theoretical frequency** np_i. The test statistic is

$$T := \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

For $n \to \infty$, its distribution converges to $\chi^2(k-1)$. The null hypothesis is rejected for $T \geq Q_{\chi^2(k-1)}(1-\alpha)$.

An **empirical distribution** *is the (discrete) distribution whose probabilities of results equal the observed relative frequencies $\frac{n_i}{n}$. It is the maximum likelihood estimate.*

3.6.1 Modifications

Problem: We test for a distribution which is only an estimate of the actual one. This causes an unspecified additional error. The theoretical frequencies of the classes must not be too small (say, below 5) to justify our conclusion.

Modification: If the theoretical frequency of some classes is too small, we join them with others (possibly “close” ones).

Problem: The distribution may depend on unknown parameters.

Modification 1: We estimate the parameters using a **different** sample.

Modification 2: We estimate the parameters using **the same** sample. This reduces the degrees of freedom; we have to test for $\chi^2(k-1-q)$, where q is the number of estimated parameters.

Problem: We want to test goodness-of-fit to a **continuous** or **mixed** distribution.

Modification: We first discretize the distribution, i.e., we split all possible results to k disjoint classes. Elements from one class should be "close" in order to achieve an acceptable power of the test. All theoretical frequencies have to be sufficiently large and – preferably – approximately equal.

3.6.2 χ^2 goodness-of-fit test of equality of two distributions

(see [Mood et al.])

H_0 : Two random variables have the same discrete distribution.

Sample sizes are m, n, frequencies are m_i, n_i $(i = 1, \ldots, k)$. We assume a distribution with unknown theoretical probabilities p_i $(i = 1, \ldots, k)$.

$$\sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} \text{ converges to } \chi^2(k-1),$$

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \text{ converges to } \chi^2(k-1),$$

$$T := \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} + \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \text{ converges to } \chi^2(2(k-1)).$$

Unknown parameters p_i may be estimated by maximum likelihood,

$$p_i = \frac{m_i + n_i}{m + n};$$

only $k - 1$ of them are independent (because $\sum\limits_{i=1}^{k} p_i = 1$), hence the resulting number of degrees of freedom is $2(k-1) - (k-1) = k - 1$ and we test T for $\chi^2(k-1)$. The null hypothesis is rejected for $T \geq Q_{\chi^2(k-1)}(1 - \alpha)$.

3.6.3 χ^2 goodness-of-fit test of independence of two distributions

(see [Likeš, Machek])

H_0 : Two discrete random variables (with unknown distributions) are independent.

X attains k values with probabilities $p_1, \ldots, p_k$,
Y attains m values with probabilities $q_1, \ldots, q_m$.
A realization of a two-dimensional sample $((x_1, y_1), \ldots, (x_n, y_n))$ consists of couples of realizations of random variables X, Y; we again need only the frequencies n_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, m)$. These are usually organized into a so-called **contingence table**. The number of classes is $k\,m$.

Under the independence assumption, the probabilities of results are $p_i q_j$ $(i = 1, \ldots, k;\ j = 1, \ldots, m)$,

$$T := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n p_i q_j)^2}{n p_i q_j} \text{ converges to } \chi^2(km - 1).$$

Unknown parameters p_i, q_j can be estimated by maximum likelihood,

$$p_i = \frac{\sum\limits_{j=1}^{m} n_{ij}}{n}, \qquad q_j = \frac{\sum\limits_{i=1}^{k} n_{ij}}{n};$$

only $(k-1)+(m-1)$ are independent (because $\sum\limits_{i=1}^{k} p_i = 1$, $\sum\limits_{j=1}^{m} q_j = 1$), hence the resulting number of degrees of freedom is $k\,m - 1 - (k-1) - (m-1) = (k-1)(m-1)$ and we test T for $\chi^2((k-1)(m-1))$. The null hypothesis is rejected for $T \geq Q_{\chi^2((k-1)(m-1))}(1-\alpha)$.

3.7 Correlation, its estimate and testing

(see [Likeš, Machek])

Correlation $\varrho_{X,Y}$ of random variables X, Y (with non-zero variances) is the mean of the product of normalized random variables $\frac{X-\mu_X}{\sigma_X} \cdot \frac{Y-\mu_Y}{\sigma_Y}$,

$$\varrho_{X,Y} = \frac{\mu_{(X-\mu_X)(Y-\mu_Y)}}{\sigma_X \sigma_Y} \in [-1, 1].$$

It vanishes for random variables which are independent, but also for some others; then we call them **uncorrelated**.
The extreme values ±1 correspond to a linear dependence between X and Y.

Using a two-dimensional sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ we may estimate the correlation by the **sample correlation**

$$R_{X,Y} = \frac{\sum\limits_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{\left(\sum\limits_{j=1}^{n}(X_j - \bar{X})^2\right)\left(\sum\limits_{j=1}^{n}(Y_j - \bar{Y})^2\right)}}$$

$$= \frac{n\sum\limits_{j=1}^{n} X_j Y_j - \left(\sum\limits_{j=1}^{n} X_j\right)\left(\sum\limits_{j=1}^{n} Y_j\right)}{\sqrt{\left(n\sum\limits_{j=1}^{n} X_j^2 - \left(\sum\limits_{j=1}^{n} X_j\right)^2\right)\left(n\sum\limits_{j=1}^{n} Y_j^2 - \left(\sum\limits_{j=1}^{n} Y_j\right)^2\right)}}.$$

(The former is the formula for the cosine of the angle between vectors $X_j - \bar{X}, Y_j - \bar{Y} \in \mathbb{R}^n$. The latter is a single-pass formula.)

3.7.1 Test of correlation of two normal distributions

Assumption: A two-dimensional random variable (X, Y) has a (two-dimensional) normal distribution, $n \geq 3$.

$H_0 : \varrho_{X,Y} = 0$ (X, Y are uncorrelated).

The test statistic is

$$T := \frac{R_{X,Y}\sqrt{n-2}}{\sqrt{1-R_{X,Y}^2}};$$

if the variables X, Y are uncorrelated, the test statistic T has the distribution $t(n-2)$ and we proceed as in Section 3.2.2.

3.8 Nonparametrical tests

They do not require any knowledge about the distribution, but they are less powerful.

3.8.1 Sign test

We distinguish only the sign of the difference from a fixed value c. Thus we lose a quantitative information and the possibility to test, e.g., the mean. Instead of that, we test the median $Q_X(\frac{1}{2})$.

$H_0 : Q_X(\frac{1}{2}) = c$

Under the null hypothesis, differences of both signs should be equally probable. Zero differences are excluded from the sample. The test statistic T is the number of positive differences. It is tested for the binomial distribution $\mathsf{Bin}\left(n, \frac{1}{2}\right)$. The null hypothesis is rejected for

$$T < Q_{\mathsf{Bin}\left(n,\frac{1}{2}\right)}\left(\frac{\alpha}{2}\right) \qquad \text{or} \qquad T > Q_{\mathsf{Bin}\left(n,\frac{1}{2}\right)}\left(1 - \frac{\alpha}{2}\right).$$

(Similarly for one-sided tests.) The computation of quantiles is complex, but they are tabulated (depending on n and the significance level).

It is easier to compute the achieved significance.

For large n, we apply the central limit theorem and we test

$$T_0 := \frac{2T - n}{\sqrt{n}}$$

for $N(0, 1)$.

The sign test can be used also to comparison of two medians of a paired sample.

In contrast to the mean, the median always exists (however, it is difficult to define it uniquely).

3.8.2 Wilcoxon test (of one sample)

$H_0 : X$ has a distribution symmetric around c

(Then c is both the median and the mean.)

From a realization $(x_1, \ldots, x_n)$ we compute the sequence $(z_1, \ldots, z_n)$, where $z_j = x_j - c$. We order it according to increasing absolute values $\left|z_j\right| = \left|x_j - c\right|$. Thus we determine the order r_j of the j^{th} element. If more differences are equal, we assign to them the same order equal to the arithmetic mean. The test statistic is

$$T_1 := \sum_{j: z_j > 0} r_j$$

or

$$T_2 := \min\left(\sum_{j: z_j > 0} r_j, \sum_{j: z_j < 0} r_j\right);$$

we compare it to the tabulated critical value for this test.

4 What is missing here

4.1 More about mappings and sums of random variables

4.2 Characteristic function of a random variable

Problems of statistical research – see Rogalewicz

4.3 Proof of the Central Limit Theorem

References

[Chatfield] Chatfield, C.: Statistics for Technology. 3rd ed., Chapman & Hall, London, 1992.

[Likeš, Machek] Likeš, J., Machek, J.: Matematická statistika. 2nd ed., SNTL, Praha, 1988.

[Mood et al.] Mood, A.M., Graybill, F.A., Boes, D.C.: Introduction to the Theory of Statistics. 3rd ed., McGraw-Hill, 1974.

[Papoulis] Papoulis, A.: Probability and Statistics, Prentice-Hall, 1990.