Introduction to some topics in applied probability

Artem Kovalevskii

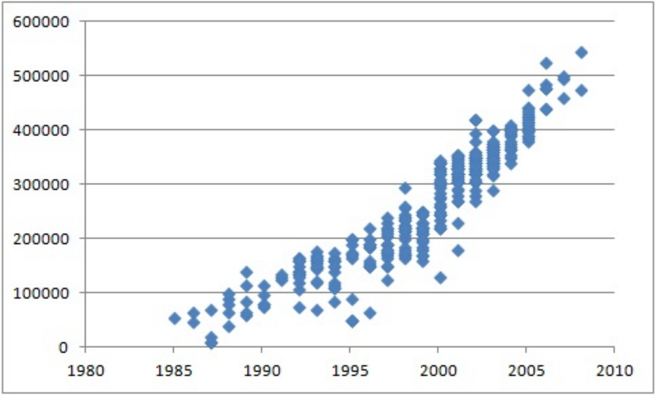
Novosibirsk State Technical University
Novosibirsk State University

pandorra@ngs.ru

Nankai, 2018

1. Example

We analyse ads about sales of Toyota Corolla cars at www.ngs.ru on 02.06.2012. We investigate dependence of price Y_i against production year X_i. Ads are ordered by the year. The order is random for cars of a same year.



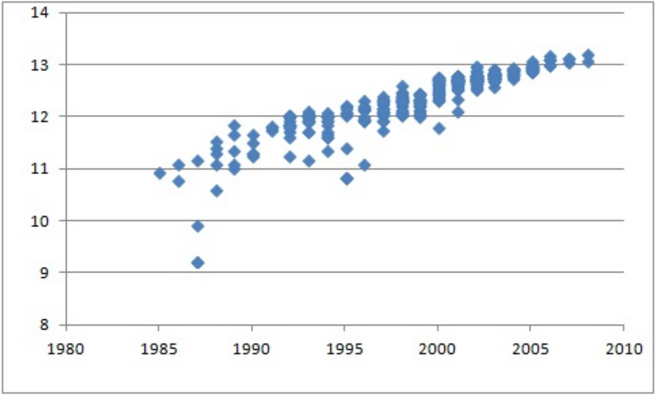
Prices of right-wheeled cars (in roubles, 382 ads)

The model is

$$\ln Y_i = aX_i + b + \varepsilon_i, \ i = 1, \ldots, n. \tag{1}$$

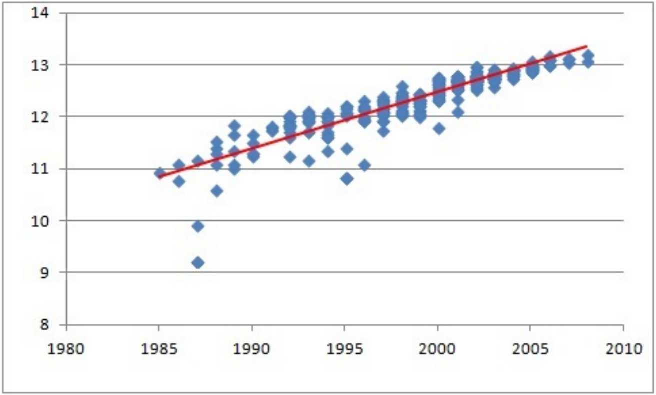
Here ε_i are independent and identically distributed, have zero mean and non-zero finite variance. Estimates of a and b are approximately

$$\widehat{a} = 0.1089, \ \widehat{b} = -205.3.$$

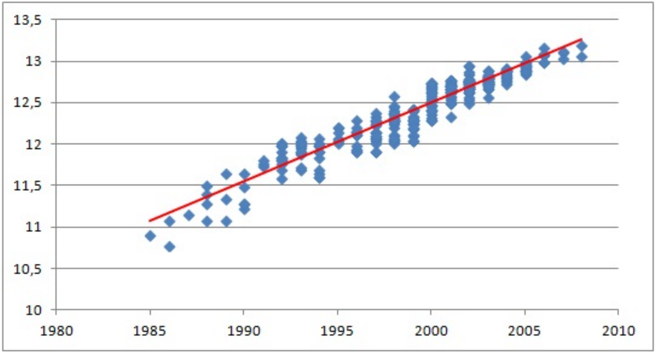


Logarifms of prices

We estimate Y_i and calculate regression residuals. The sample standard deviation of regression residuals is $S = 0.2469$.



Logarifms of prices with a trend line



Logarifms of prices after 3-sigma procedure

The linear model is not appropriate

2. Basic concepts and theorems of probability

Probability theory studies mathematical models of random experiments, that is, experiments with an unpredictable result but with convergent frequency.

A random experiment is associated with a space of elementary outcomes, an arbitrary non-empty set Ω. Measurable subsets of Ω are called events.

Definition *Probability space* is a space of elementary outcomes with a probability measure (probability) $\mathbf{P}$ on its subsets such that:

1) $\mathbf{P}(A) \geq 0$ for any event A (non-negativity);

2) $\mathbf{P}(A_1 \cup A_2 \cup \ldots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \ldots$ for any finite or countable set of disjoint events $A_1,\ A_2,\ \ldots$ (countable additivity);

3) $\mathbf{P}(\Omega) = 1$ (normalization).

Conditional probability of the event A provided that the event B has occurred is defined as $\mathbf{P}(A|B) = \mathbf{P}(AB)/\mathbf{P}(B)$.

Definition *Random vector* $\mathbf{X} = (X_1, \dots, X_n)$ is a mapping from the space Ω to n-dimentional arifmetic space $\mathbf{R}^n$ such that set $\{\omega \in \Omega : \ \mathbf{X}(\omega) < \mathbf{t}\}$ is an event for any $\mathbf{t} = (t_1, \dots, t_n) \in \mathbf{R}^n$, that is, probability of the set is defined.
Multidimensional cumulative distribution function (cdf) of random vector $\mathbf{X}$ is the probability as a function of the vector variable $\mathbf{t}$:

$$F_{\mathbf{X}}(\,\mathbf{t}\,) = \mathbf{P}\{\,\omega \in \Omega : \ \mathbf{X}(\omega) \le \mathbf{t}\,\} = \mathbf{P}\{\,\mathbf{X} \le \mathbf{t}\,\}.$$

Inequality $\mathbf{X} \leq \mathbf{t}$ is understood coordinate-wise, that is, it means the system of inequalities $X_1 \leq t_1, \ldots, X_n \leq t_n$.
For $n = 1$ we obtain definitions of a random variable and a cumulative distribution function.
A random vector $\mathbf{X}$ is called *discrete* if it takes a finite or a countable number of values $\mathbf{t}_1, \mathbf{t}_2, \ldots$. One can determine its distribution by its distribution table, that is, by a collection of probabilities $\mathbf{P}\{\mathbf{X} = \mathbf{t}_j\}$. In 2-dimensional case (if $n = 2$) one can write the distribution table of discrete vector $\mathbf{X} = (X, Y)$ as

$X \backslash Y$	b_1	b_2	$\ldots$
a_1	p_{11}	p_{12}	$\ldots$
a_2	p_{21}	p_{22}	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Here $p_{ij} = \mathbf{P}\{X = a_i,\ Y = b_j\}$.
Properties of the table of distribution of a two-dimensional discrete vector:
1) all a_i are distinct;
2) all b_j are distinct;
3) all p_{ij} are nonnegative;
4) the sum of all p_{ij} is 1.

Exercise

Let first player wins with probability $1/4$, second player wins with probability $1/4$, too. Nobody win with probability $1/2$. Let X and Y be indicators of wins of first and second player, respectively. Write a distribution table of random vector (X, Y).

In one-dimensional case, distribution of a discrete random variable is written in the form of a table, containing its values $a_1, \; a_2, \; \ldots$ and probabilities $p_i = \mathbf{P}\{X = a_i\}$. The table of one-dimensional distribution of random variable X is obtained from two-dimensional distribution table by formula $\mathbf{P}\{X = a_i\} = \sum_j p_{ij}$.

Example Poisson distribution
$\mathbf{P}\{\xi = k\} = \frac{\lambda^k}{k!}e^{-\lambda},\ k = 0,\ 1,\ \ldots.$
Here $\lambda > 0$ is a parameter of Poisson distribution.

Random vector $\mathbf{X}$ is said to have *multidimensional absolutely continuous distribution* if there exists an *multidimensional probability density function (pdf)* $f_{\mathbf{X}}(\mathbf{t})$ such that for any Borel set $B \subseteq \mathbf{R}^n$

$$\mathbf{P}\{\mathbf{X} \in B\} = \int\limits_B f_{\mathbf{X}}(\mathbf{t})d\mathbf{t}$$

(hereinafter, we use the notation $d\mathbf{t} = dt_1 \ldots dt_n$).

So if $\mathbf{X}$ has multidimensional absolutely continuous distribution then for any $\mathbf{t} \in \mathbf{R}^n$

$$F_{\mathbf{X}}(\mathbf{t}) = \int\limits_{\mathbf{u} \leq \mathbf{t}} f_{\mathbf{X}}(\mathbf{u}) d\mathbf{u}.$$

Properties of multidimensional pdf: $f_{\mathbf{X}}(\,\mathbf{t}\,) \geq 0$; $\int\limits_{\mathbf{R}^n} f_{\mathbf{X}}(\,\mathbf{t}\,) d\mathbf{t} = 1.$

One-dimensional pdfs of components of a random vector are calculated by integrating of the multidimensional pdf over all values of all other components.

Components of a random vector are called *independent* if for any subsets $B_1, \ \dots, \ B_n$ of the real line

$$\mathbf{P}\{X_1 \in B_1, \ \dots, \ X_n \in B_n\} = \mathbf{P}\{X_1 \in B_1\} \cdot \ldots \cdot \mathbf{P}\{X_n \in B_n\}.$$

The distribution of a vector with independent components is determined by the distributions of the components.

If a discrete random vector takes values $\mathbf{t}_1$, $\mathbf{t}_2$, ..., then its expectation is a vector

$$\mathbf{EX} = \sum_j \mathbf{t}_j \mathbf{P}\{\mathbf{X} = \mathbf{t}_j\}.$$

If series

$$\sum_j |\mathbf{t}_j| \mathbf{P}\{\mathbf{X} = \mathbf{t}_j\}$$

is divergent then one said that the mathematical expectation does not exist. Here, $|\cdot|$ denotes the Euclidean norm of the vector.

If $\mathbf{g}: \ \mathbf{R}^n \to \mathbf{R}^m$ be a vector function, $m \geq 1$, then

$$\mathbf{E}\mathbf{g}(\mathbf{X}) = \sum_j \mathbf{g}(\mathbf{t}_j)\mathbf{P}\{\mathbf{X} = \mathbf{t}_j\}.$$

In the absolutely continuous case, expectation is defined by formula

$$\mathbf{E}\mathbf{X} = \int\limits_{\mathbf{R}^n} \mathbf{t} f_{\mathbf{X}}(\,\mathbf{t}\,)d\mathbf{t}.$$

Expectation does not exist if

$$\int\limits_{\mathbf{R}^n} |\,\mathbf{t}\,| f_{\mathbf{X}}(\,\mathbf{t}\,)d\mathbf{t}$$

is divergent.

Calculation of expectation:

$$\mathbf{E}\mathbf{g}(\mathbf{X}) = \int\limits_{\mathbf{R}^n} \mathbf{g}(\,\mathbf{t}\,) f_{\mathbf{X}}(\,\mathbf{t}\,) d\mathbf{t}.$$

If components of vector **g** are written in the form of a matrix, the corresponding formulas give mathematical expectation of the random matrix.

The covariance matrix of random vector column **X** is the matrix

$$C(\mathbf{X}) = \mathbf{E}(\mathbf{X} - \mathbf{E}\mathbf{X})(\mathbf{X} - \mathbf{E}\mathbf{X})^T.$$

Matrix $C(\mathbf{X})$ is symmetric and nonnegatively definite. Its diagonal elements are variances of components of the random vector, and the off-diagonal elements are covariances of the corresponding pairs of components. The root-mean-square (standard) deviation of a component is the root of its variance. The correlation coefficient of two components is their covariance divided by the product of standard deviations.

Let random vector $\mathbf{X}$ have a multidimensional standard normal distribution, that is, its multidimensional pdf is

$$f_{\mathbf{X}}(\mathbf{t}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{t}^{\ T}\mathbf{t}\right),$$

$\mathbf{t} = (t_1,\ \ldots,\ t_n)^T$, $\mathbf{t}^{\ T}\mathbf{t} = t_1^2 + \ldots + t_n^2$.

Let column vector $\mathbf{Y}$ be expressed in terms of column vector $\mathbf{X}$ linearly:

$$\mathbf{Y} = \mathbf{a} + B\mathbf{X},$$

$\mathbf{a}$ is a non-random column vector, B is a square matrix. Then $\mathbf{Y}$ is said to have a multidimensional normal distribution.
Normal vector $\mathbf{Y}$ has an expectation vector $\mathbf{EY} = \mathbf{a}$ and a covariance matrix $C = C(\mathbf{Y}) = BB^T$. The distribution of the normal vector $\mathbf{Y}$ is completely determined by the mathematical expectation and the covariance matrix. The matrix B is defined by a given multidimensional normal distribution up to an orthogonal matrix.

A multidimensional normal distribution is said to be nondegenerate if the matrix B is non-degenerate, that is, $\det B \neq 0$, or, equivalently, $\det C > 0$.
In this case, there is a multidimensional pdf

$$f_{\mathbf{Y}}(\mathbf{t}) = \frac{1}{(2\pi)^{n/2}(\det C)^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{t}-\mathbf{a})^T C^{-1}(\mathbf{t}-\mathbf{a})\right).$$

Definition A sequence of random variables $\{Y_n\}$ is said to converge almost surely (a.s.) to a random variable Y if

$$\mathbf{P}\{\omega : \; Y_n(\omega) \to Y(\omega)\} = \mathbf{P}\{Y_n \to Y\} = 1.$$

Designation $Y_n \stackrel{a.s.}{\to} Y$.
Theorem (strong law of large numbers, SLLN) *Let random variables* $X_1, X_2, \ldots$ *be independent and identically distributed, and* $\mathbf{E}|X_1| < \infty$. *Let* $a = \mathbf{E}X_1$, $S_n = \sum_{i=1}^{n} X_i$. *Then*

$$\frac{S_n}{n} \stackrel{a.s.}{\to} a$$

as $n \to \infty$
Definition A sequence of random vectors $\{\mathbf{Y}_n\}$ is said to converge in distribution (weakly) to a random vector Y if

$$F_{\mathbf{Y}_n}(\mathbf{x}) \to F_{\mathbf{Y}}(\mathbf{x})$$

for all points $\mathbf{x}$ in that $F_{\mathbf{Y}}$ is continious.
Designation $Y_n \Rightarrow Y$.

Central limit theorem (CLT) *Let $X_1, X_2 \ldots$ be independent and identically distributed random variables. Let $S_n = X_1 + \ldots + X_n$, $a = \mathbf{E}X_1$, $\sigma^2 = \mathbf{Var}X_1$, and $0 < \sigma^2 < \infty$. Then $(S_n - na)/\sigma\sqrt{n} \Rightarrow Z$, Z has standard normal distribution. That is, for any $x \in \mathbf{R}$*

$$\mathbf{P}\left\{\frac{S_n - na}{\sigma\sqrt{n}} \le x\right\} = F_{\frac{S_n - na}{\sigma\sqrt{n}}}(x) \to \Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2}\, dt$$

as $n \to \infty$.

Multidimensional central limit theorem *Let* $\mathbf{X}_1, \mathbf{X}_2 \ldots$ *be independent and identically distributed random m-dimensional vectors with expectation vector* $\mathbf{a}$ *and non-degenerate covariance matrix* C*. Let* $\mathbf{S}_n = \mathbf{X}_1 + \ldots + \mathbf{X}_n$*. Then* $\frac{\mathbf{S}_n - n\mathbf{a}}{\sqrt{n}} \Rightarrow Z$*, that is, for any* $\mathbf{x} \in \mathbf{R}^m$

$$\mathbf{P}\left\{\frac{\mathbf{S}_n - n\mathbf{a}}{\sqrt{n}} \le \mathbf{x}\right\} \to \mathbf{P}\{\mathbf{Z} \le \mathbf{x}\}$$

as $n \to \infty$*,* $\mathbf{Z}$ *is a normal random vector with expectation* $\mathbf{0}$ *a covariance matrix* C*.*

3. Elements of the theory of stochastic processes

A stochastic process is a set of random variables indexed by time: $\xi = \{\xi(t),\ t \in G\}$. Here $G \subseteq \mathbf{R}$.

Poisson process $\{\Pi(t),\ t \geq 0\}$:

$$\mathbf{P}\{\Pi(t) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t},\ t > 0,\ k = 0;\ 1;\ \ldots .$$

$\lambda > 0$ is an intensity of the Poisson process.
Thus the Poisson process at time $t > 0$ has Poisson distribution with parameter λt.
The Poisson process has many amazing properties. We need the splitting property.

Theorem (on a Poisson infinite urn scheme) Let the total number of balls ξ is a Poisson random variable with parameter λ. Each of the balls independently of the others is placed in an urn with number i with probability $p_i > 0$, $i = 1,\ 2,\ \ldots$, $\sum_{i=1}^{\infty} p_i = 1$. Let ξ_i be a number of balls in i-th urn. Then ξ_i, $i = 1,\ 2,\ \ldots$, are independent Poisson random variables with parameters λp_i.

Proof

For any $m < \infty$ let $q_m = p_m + p_{m+1} + \ldots$, $\xi_{\geq m} = \xi_m + \xi_{m+1} + \ldots$. Then

$$\mathbf{P}\{\xi_i = k_i,\ i = 1,\ 2,\ \ldots,\ m-1,\ \xi_{\geq m} = k_m\}$$

$$= \mathbf{P}\{\xi = k_1 + \ldots + k_m\} \frac{(k_1 + \ldots + k_m)!}{k_1! \ldots k_m!} p_1^{k_1} \ldots p_{m-1}^{k_{m-1}} q_m^{k_m}$$

$$= \frac{\lambda^{k_1 + \ldots + k_m} p_1^{k_1} \ldots p_{m-1}^{k_{m-1}} q_m^{k_m}}{k_1! \ldots k_m!} e^{-\lambda}$$

$$= \frac{(\lambda p_1)^{k_1}}{k_1!} e^{-\lambda p_1} \ldots \frac{(\lambda p_{m-1})^{k_{m-1}}}{k_{m-1}!} e^{-\lambda p_{m-1}} \ldots \frac{(\lambda q_m)^{k_m}}{k_m!} e^{-\lambda q_m}.$$

So $\xi_1, \ldots, \xi_{m-1}$ are Poisson and independent for any m. So all $\xi_1, \xi_2, \ldots$ are Poisson and independent.
The proof is complete.

This result is valid for a finite Poisson urn scheme, too.
Corollary (on the splitting of a Poisson flow) Let $\Pi = \{\Pi(t),\ t \geq 0\}$ be a Poisson process with intensity λ, $p_i > 0$, $i = 1,\ 2,\ \ldots$, $\sum_{i=1}^{\infty} p_i = 1$. Let form stochastic processes Π_i, $i = 1,\ 2,\ \ldots$, according to the following rule: every moment of time, when $\Pi(t)$ grows by one, we assign with probability p_i to process Π_i (independently of the others moments of growth). Then processes Π_i, $i = 1,\ 2,\ \ldots$, are independent Poisson processes with intensities λp_i.

Exercise
Prove it for splitting on 2 processes.

Gaussian process $X = \{X(t),\ t \in G\}$ is a stochastic process with normal finite-dimensional distributions, that is, for any $t_1, \dots, t_m \in G$ vector $(X(t_1), \dots, X(t_m))$ have m-dimensional normal distribution.

Distribution of a Gaussian process is defined by its expectation $a(t) = \mathbf{E}X(t)$ and its covariance function $K(s,t) = \mathbf{E}X(s)X(t) - \mathbf{E}X(s)\mathbf{E}X(t)$. A stochastic process is called *centered* if its expectation equals to 0, that is, $a(t) \equiv 0$. We have $K(s,t) = \mathbf{E}X(s)X(t)$ for a centered process.

One can define m-dimensional Gaussian process with vector function of expectation $\mathbf{a}(t)$ and covariance matrix function $K(s,t) = (K_{ij}(s,t))$.

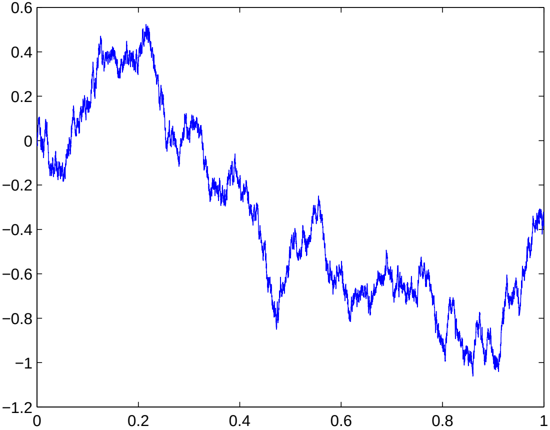
Let us describe the conditions under which a Gaussian process has continuous a.s. sample paths. Continuity of its correlation function does not guarantee the continuity of sample paths. A correct description of the continuity conditions uses an entropy approach. We give only a sufficient condition in terms of the correlation function.

Theorem (sufficient condition for the continuity a.s. of a Gaussian process)
If there are $0 < C < \infty$, $\alpha > 0$, $\eta > 0$ such that

$$\mathbf{E}(X(s) - X(t))^2 = K(s,s) + K(t,t) - 2K(s,t) \le \frac{C}{|\ln|s-t||^{1+\alpha}}$$

for all s, t from a compact set $G \subset \mathbf{R}$ such that $|s-t| < \eta$, then Gaussian process $X = \{X(t),\ t \in G\}$ has continious a.s. sample paths.

Standard Wiener process $W = \{W(t),\ t \in G\}$ is a centered Gaussian process with covariance function $K_W(s,t) = \min(s,t)$. We let $G = [0,1]$.


0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
0
0.2
0.4
0.6
0.8
1

So W is a Gaussian process with independent increments, zero expectation and variance $\mathbf{Var}\, W(t) = t$. The standard Wiener process with probability 1 has everywhere continuous but nowhere differentiable sample paths. Continuity a.s. of sample paths follows from the above theorem:
$\mathbf{E}(W(s) - W(t))^2 = |s - t|$, and $\eta \ln^{1+\alpha} \eta \to 0$ as $\eta \to 0$ for any $\alpha > -1$.

4. Brownian bridge

Standard Wiener process $W = \{W(t),\ t \geq 0\}$ played an important role in the preceding chapter. We recall that it is a Gaussian process with independent increments, for which the expectation is zero, and the variance at any time $t \geq 0$ is equal to t. Its covariance function

$$K_W(u, v) = \mathbf{E}W(u)W(v) = \min(u, v), \quad u, v \geq 0.$$

Brownian bridge $W^0 = \{W^0(t),\ 0 \leq t \leq 1\}$ is defined by $W^0(t) = W(t) - tW(1)$. From definition, $W^0(1) = 0$. Brownian bridge is also a Gaussian process with zero expectation but with dependent increments. From the definition, using properties of expectation, we obtain covariance function of Brownian bridge

$$K_W^0(u, v) = \mathbf{E}W^0(u)W^0(v) = \min(u, v) - uv, \ \ 0 \leq u, v \leq 1.$$

So $\mathbf{Var} W^0(t) = t(1-t)$, $t \in [0,\, 1]$.
Equivalent definition of the Brownian bridge: this is a Wiener process, provided that $W(1) = 0$, that is,

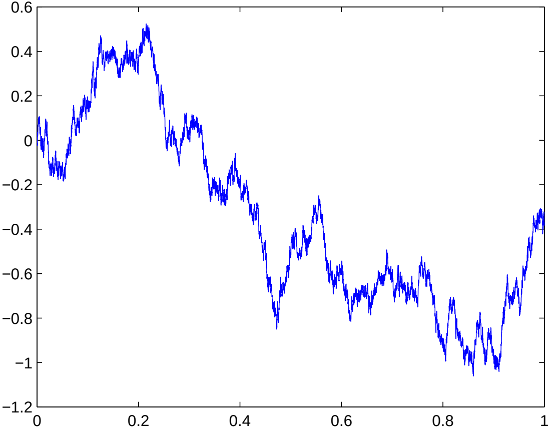
$$\mathbf{P}\{J(W^0) \le x\} = \lim_{\varepsilon\to 0} \mathbf{P}\{J(W) \le x | W(1) \in (-\varepsilon,\ \varepsilon)\}$$

$$= \lim_{\varepsilon\to 0} \frac{\mathbf{P}\{J(W) \le x,\ W(1) \in (-\varepsilon,\ \varepsilon)\}}{\mathbf{P}\{W(1) \in (-\varepsilon,\ \varepsilon)\}}$$

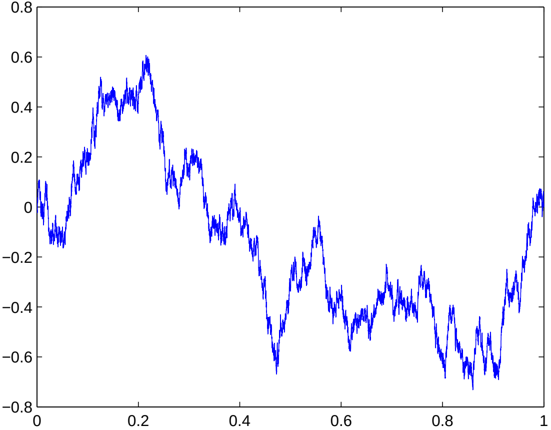
$$= \sqrt{2\pi} \lim_{\varepsilon\to 0} \frac{\mathbf{P}\{J(W) \le x,\ W(1) \in (-\varepsilon,\ \varepsilon)\}}{2\varepsilon}.$$

Here J is continuous in the uniform metric functional in $C(0,1)$, and x is any real number.
The last equality is true because $W(1) \sim \mathcal{N}_{0,1}$, and the pdf of the standard normal law at the point 0 is $1/\sqrt{2\pi}$.

A sample path of $W(t)$



A sample path of $W^0(t)$



5. Integral functionals

For any $u \geq 0$, $du \geq 0$ let designate $dW(u) = W(u + du) - W(u)$. Due to independence of increments $dW(u) \sim \mathcal{N}_{0,du}$.
Let function $g(t)$ be Riemann integrable on [0, 1]. Let integral $\int_0^1 g(t)dW(t)$ be a Gaussian random variable with expectation and variance equal to limits of expectations and variances of the corresponding integral sums.
The expectation of integral $\int_0^1 g(t)dW(t)$ is 0.
Theorem If g, h are Riemann integrable on [0, 1] then

$$\mathbf{cov}\left(\int_0^1 g(t)dW(t),\ \int_0^1 h(t)dW(t)\right) = \int_0^1 g(t)h(t)dt.$$

Proof
As

$$\mathbf{E}dW(u)dW(v) = du$$

for $u = v$, $du = dv$, and

$$\mathbf{E}dW(u)dW(v) = 0,$$

if $(u, u + du)$ and $(v, v + dv)$ don't intersect, we have

$$\mathbf{cov}\left(\int_0^1 g(t)dW(t),\ \int_0^1 h(t)dW(t)\right)$$

$$= \mathbf{E}\int_0^1 g(t)dW(t)\int_0^1 h(t)dW(t)$$

$$= \int_0^1\int_0^1 g(u)h(v)\mathbf{E}dW(u)dW(v) = \int_0^1 g(t)h(t)dt.$$

So

$$\mathbf{Var}\int_0^1 g(t)dW(t) = \int_0^1 g^2(t)dt.$$

The proof is complete.

We now consider functional $\int_0^1 W(t)dt$, it also has a zero mathematical expectation. Integrating by parts, we obtain

$$\int_0^1 W(t)dt = tW(t)|_0^1 - \int_0^1 tdW(t)$$

$$= W(1) - \int_0^1 tdW(t) = \int_0^1 (1-t)dW(t).$$

So

$$\mathbf{Var} \int_0^1 W(t)dt = \int_0^1 (1-t)^2 dt = \frac{1}{3}.$$

Thus $\int_0^1 W(t)dt \sim \mathcal{N}_{0,1/3}$.
Exercise Prove that $\int_0^1 W^0(t)dt \sim \mathcal{N}_{0,1/12}$.

6. Exrtemal functionals

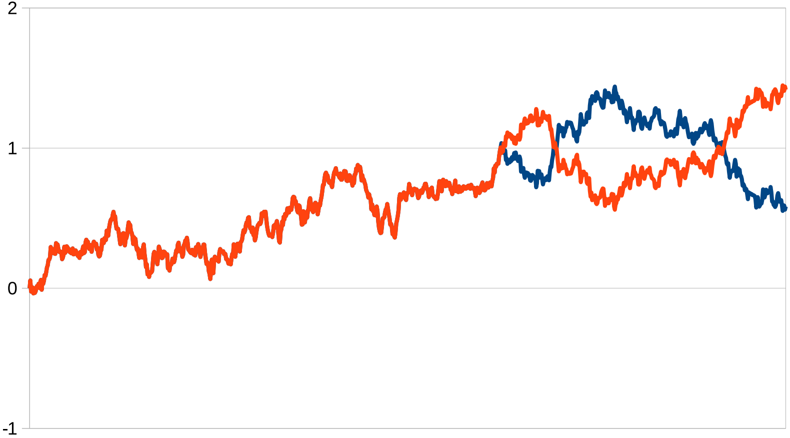
In this section we shall discuss distributions of random variables

$$\max_{0\le t\le 1} W(t), \qquad \max_{0\le t\le 1} W^0(t),$$

$$\max_{0\le t\le 1} |W(t)|, \qquad \max_{0\le t\le 1} |W^0(t)|.$$

These are functionals of the standard Wiener process and the Brownian bridge. But their distributions differ from normal. Formulas for their distribution are derived using a reflection principle.

Reflection of W on level 1



Recall that $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$ is the cdf of the standard normal law, $\overline{\Phi}(x) = 1 - \Phi(x)$ is the tail of the standard normal distribution.
Let $\widetilde{W}$ be reflected process. From symmetry, distributions of W and $\widetilde{W}$ are the same.
$W(1) \sim \mathcal{N}_{0,1}$, $\widetilde{W}(1) \sim \mathcal{N}_{0,1}$.

$$\mathbf{P}\{\max_{0 \le t \le 1} W(t) > x\} = \mathbf{P}\{W(1) > x\} + \mathbf{P}\{\widetilde{W}(1) > x\}$$

$$= 2\mathbf{P}\{W(1) > x\} = 2\overline{\Phi}(x).$$

So

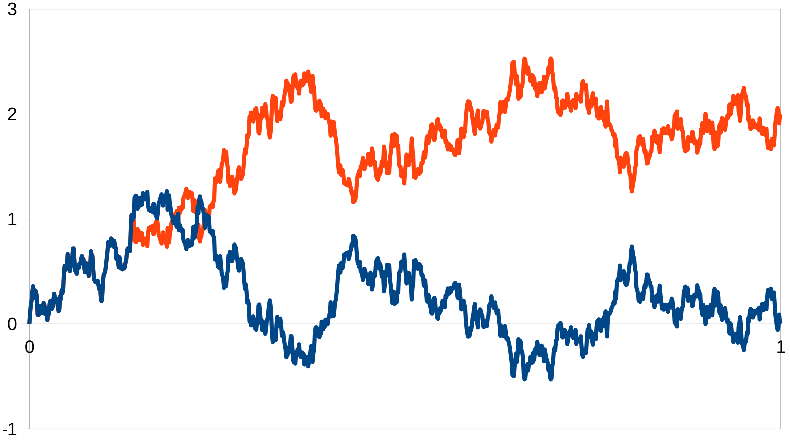
$$\mathbf{P}\{\max_{0 \le t \le 1} W(t) \le x\} = 1 - 2\overline{\Phi}(x).$$

Exercise

Calculate

$$\mathbf{P}\{\min_{0 \le t \le 1} W(t) \le -1.96\}$$

Reflection of W^0 on level 1



Remember that $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is a pdf for standard normal distribution.

$$\mathbf{P}\{\max_{0 \le t \le 1} W^0(t) > x\} = \lim_{\varepsilon \to 0} \frac{\mathbf{P}\{\widetilde{W}(1) \in (2x - \varepsilon,\, 2x + \varepsilon)\}}{W(1) \in (-\varepsilon,\, \varepsilon)\}}$$

$$= \frac{\varphi(2x)}{\varphi(0)} = e^{-2x^2},$$

so

$$\mathbf{P}\{\max_{0 \le t \le 1} W^0(t) \le x\} = 1 - e^{-2x^2}.$$

Exercise

Calculate

$$\mathbf{P}\{\min_{0 \le t \le 1} W^0(t) \le -2\}.$$

If we calculate

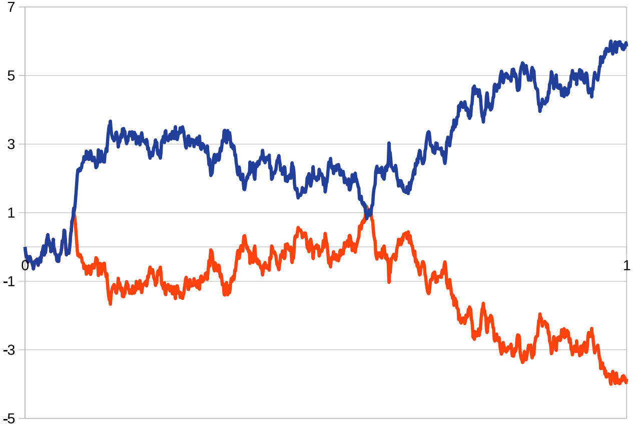
$$\mathbf{P}\{\max_{0\le t\le 1}|W(t)|>x\},$$

then we have two borders: x and $-x$.

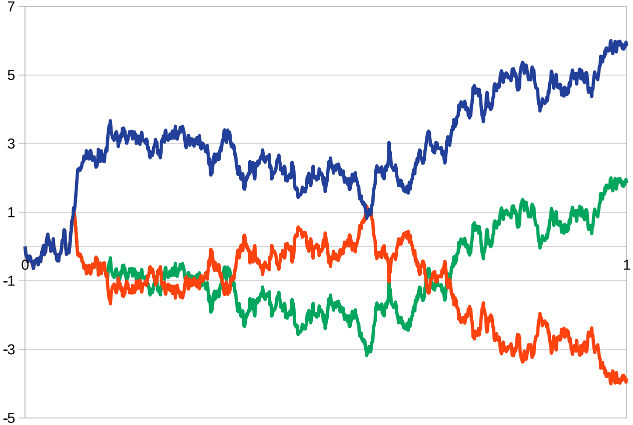
$$\mathbf{P}\{\max_{0\le t\le 1}|W(t)|>x\}$$

$$=\mathbf{P}\{W(1)>x \text{ or } \widetilde{W}(1)>x \text{ or } W(1)<-x \text{ or } \widetilde{W}(1)<-x\}.$$

But events
$\{W(1)>x\},\ \ \{\widetilde{W}(1)>x\},\ \ \{W(1)<-x\}\ \ \{\widetilde{W}(1)<-x\}$
intersect, that is, can take place simultaneusly. So we use multiple reflections.



$$\mathbf{P}\{\max_{0\le t\le 1}|W(t)|>x\}=4\mathbf{P}\{W(1)>x\}-4\mathbf{P}\{W(1)>3x\}\dots$$



$$\mathbf{P}\{\max_{0\le t\le 1}|W(t)|>x\}=4\mathbf{P}\{W(1)>x\}-4\mathbf{P}\{W(1)>3x\}$$

$$+4\mathbf{P}\{W(1)>5x\}-\dots$$

$$\mathbf{P}\{\max_{0\le t\le 1} |W(t)| \le x\} = 1 - 4\sum_{k=0}^{\infty}(-1)^k \mathbf{P}\{W(1) > (2k+1)x\}.$$

So

$$\mathbf{P}\{\max_{0\le t\le 1} |W(t)| \le x\} = 1 - 4\sum_{k=0}^{\infty}(-1)^k\,\overline{\Phi}((2k+1)x).$$

Analogously,

$$\mathbf{P}\{\max_{0\leq t\leq 1}|W^0(t)|>x\}=\frac{2\varphi(2x)-2\varphi(4x)+2\varphi(6x)-\dots}{\varphi(0)}$$

$$=2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2x^2}.$$

Thus

$$\mathbf{P}\{\max_{0\leq t\leq 1}|W^0(t)|\leq x\}=\sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2x^2}.$$

All these formulas are valid for $x>0$. The latter distribution is called the Kolmogorov distribution.

Exercise

Calculate

$$\mathbf{P}\{\max_{0\leq t\leq 1}|W^0(t)|\geq 2\}.$$

7. Omega squared functional

Let omega squared functional be $\omega^2 = \int\limits_0^1 \left(W^0(t)\right)^2 dt$.

Here W^0 is a standard Brownian bridge.

Then random variable ω^2 is called to have omega squared distribution.

To calculate its cdf, we introduce random Fourrier series.

Later we will use this approach for other Gaussian processes.

We use Smirnov's formula (Smirnov, 1937).

Theorem If $J = \sum_{k=1}^{\infty} \frac{\eta_k^2}{\lambda_k}$, $\eta_1, \eta_2, \ldots$ are independent and have standard normal distribution, $0 < \lambda_1 < \lambda_2 < \ldots$, then

$$F_J(x) = 1 + \frac{1}{\pi} \sum_{k=1}^{\infty} (-1)^k \int_{\lambda_{2k-1}}^{\lambda_{2k}} \frac{e^{-\lambda x/2}}{\sqrt{-D(\lambda)}} \cdot \frac{d\lambda}{\lambda},\ x > 0.$$

Here

$$D(\lambda) = \prod_{k=1}^{\infty} \left(1 - \frac{\lambda}{\lambda_k}\right),$$

integrals under summation must go to zero.

So, our task is to calculate λ_k, $k \geq 1$.

We will find a very compact formula for $D(\lambda)$ in our case.

Integrals are calculated numerically.

Remember

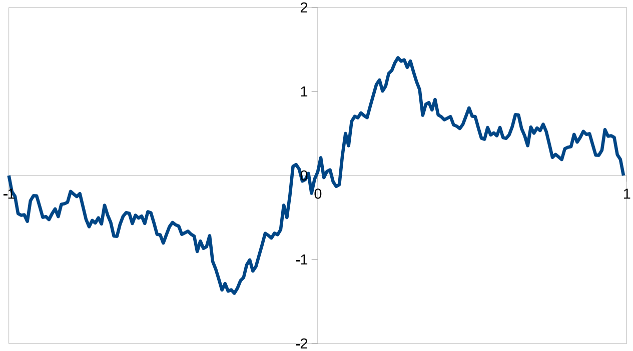
$$K_W^0(u,v) = \mathbf{E}W^0(u)W^0(v) = \min(u,v) - uv, \quad 0 \le u,v \le 1.$$

Note that for any $k \ge 1$

$$\int_0^1 K_W^0(u,v)\sin \pi kv \, dv = \frac{\sin \pi ku}{\pi^2 k^2}.$$

So $\sin \pi kv$ are eigenfunctions of kernel $K_W^0(u,v)$ with eigenvalues $\pi^{-2}k^{-2}$.

We let $W^0(-t) = -W^0(t)$ by definition, $0 \leq t \leq 1$.



So we have Fourrier series on $[-1, 1]$

$$W^0(t) = \frac{\alpha_0}{2} + \sum_{k=1}^{\infty} \left(\alpha_k \cos \pi kt + \beta_k \sin \pi kt\right).$$

From symmetry, $\alpha_k = 0$ for $k \geq 0$.
Remember that $\sin \pi kv$ are eigenfunctions and therefore β_k for $k \geq 1$ are independent random coefficients.
$\beta_k = \int_{-1}^{1} W^0(t) \sin \pi kt \, dt = 2 \int_0^1 W^0(t) \sin \pi kt \, dt$

As W^0 is a centered Gaussian process, β_k are independent normal random variables with zero expectations, $k \geq 1$.
To correspond Smirnov's theorem, we let

$$\beta_k = \frac{\eta_k \sqrt{2}}{\sqrt{\lambda_k}},$$

η_k are independent standard normal random variables, λ_k are positive constants, $k \geq 1$.

Really, we have

$$W^0(t) = \sum_{k=1}^{\infty} \frac{\eta_k \sqrt{2}}{\sqrt{\lambda_k}} \sin \pi k t,$$

and

$$\omega^2 = \int\limits_0^1 \left(W^0(t)\right)^2 dt = \sum_{k=1}^{\infty} \frac{2\eta_k^2}{\lambda_k} \int_0^1 \sin^2 \pi k t \, dt = \sum_{k=1}^{\infty} \frac{\eta_k^2}{\lambda_k}$$

due to ortogonality of functions $\sin \pi k t$ and $\sin \pi m t$ on [0, 1] for $m \neq k$.

As

$$\frac{\eta_k\sqrt{2}}{\sqrt{\lambda_k}} = 2\int_0^1 W^0(t)\sin \pi kt\,dt,$$

we have (taking expectations of squares and using linearity of expectation)

$$\mathbf{E}\left(\frac{\eta_k\sqrt{2}}{\sqrt{\lambda_k}}\right)^2 = 4\int_0^1\int_0^1 \mathbf{E}W^0(s)W^0(t)\sin \pi ks \sin \pi kt\,ds\,dt.$$

As $\mathbf{E}\eta_k^2 = 1$, LHS is $2/\lambda_k$. Remember that $\mathbf{E}W^0(s)W^0(t) = K_W^0(s,t)$.

Thus

$$\lambda_k = \frac{1}{2\int_0^1 \int_0^1 K_W^0(s,t) \sin \pi ks \sin \pi kt \, ds \, dt}.$$

W^0 is a standard Brownian bridge, and

$$K_W^0(s,t) = \min(s,t) - st.$$

Exercise Calculate λ_k.

$\lambda_k = \pi^2 k^2$.
Using formula

$$\prod_{k=1}^{\infty}\left(1-\frac{z^2}{k^2}\right)=\frac{1}{z\Gamma(z)\Gamma(1-z)}=\frac{\sin\pi z}{\pi z},$$

we have

$$D(\lambda)=\prod_{k=1}^{\infty}\left(1-\frac{\lambda}{\lambda_k}\right)=\frac{\sin\sqrt{\lambda}}{\sqrt{\lambda}}.$$

So we have Smirnov's formula for ω^2 cdf

$$F_{\omega^2}(x) = 1 + \frac{1}{\pi} \sum_{k=1}^{\infty} (-1)^k \int_{\pi^2(2k-1)^2}^{\pi^2(2k)^2} \frac{e^{-\lambda x/2}}{\sqrt{-\frac{\sin\sqrt{\lambda}}{\sqrt{\lambda}}}} \cdot \frac{d\lambda}{\lambda},\ x > 0.$$

Substituting $\lambda = \mu^2$, we have

$$F_{\omega^2}(x) = 1 + \frac{2}{\pi} \sum_{k=1}^{\infty} (-1)^k \int_{(2k-1)\pi}^{2k\pi} \frac{e^{-\mu^2 x/2}\, d\mu}{\sqrt{-\mu \sin \mu}},\ x > 0.$$

Integrals are calculated numerically.

8. Functional central limit theorem

Theorem *Let* $X_1, X_2 \ldots$ *be independent and identically distributed random variables,* $0 < \sigma^2 = \mathbf{Var} X_1 < \infty$. *Then*

$$W_n \Rightarrow W \quad \text{in} \quad C(0,1).$$

Let us explain each letter in the statement of the theorem. Let's start from the end.

Space $C(0,1)$ is a set of functions $x = \{x(t),\ 0 \leq t \leq 1\}$ that are continious on [0, 1]. Distance in uniform metric in space $C(0,1)$ between two functions is the maximum of their absolete difference:

$$dist(x,y) = \max_{t \in [0,1]} |x(t) - y(t)|.$$

This distance (metrics) defines a topology (a collection of neighborhoods) in $C(0,1)$ as follows: ε-neighborhood of function x is the set of functions that differ from it in the uniform metric by less than ε.

Exercise Shut your eyes and imagine a function that is continuous on [0, 1] by a point in some space. Then open your eyes and draw ε-neighborhood of this function in $C(0,1)$. Functions in this neighborhood are super-numerous, aren't it?

If we take some set of functions $B \subset C(0,1)$ then its boundary ∂B is a function, any neighborhood of which lies partly in B and partly not. What's this for? Of course, for the definition of weak convergence $\Rightarrow$.
Before formulating what it means in the space $C(0,1)$, let describe it on a real axis, that is, for random variables.
Weak convergence of a sequence of random variables $\{\xi_n\}$ to a random variable ξ is convergence of the distribution functions at all points of continuity of the limit function: if $F_\xi(t+0) = F_\xi(t)$ then $F_{\xi_n}(t) \to F_\xi(t)$ as $n \to \infty$.

Let formulate the equivalent condition for weak convergence. As already mentioned, boundary ∂B of the set B in some topological space is the set of such points that any its neighborhood contains both points of B and points not of B. For example, the boundary of a segment (as well as an open interval) on a real line is the set of two points, that is, its ends.

This equivalent definition of weak convergence holds for separable full topological spaces without changes. We decipher, however, what it means in the case of $C(0,1)$.
Sequence of stochastic processes $Z_n = \{Z_n(t),\ 0 \leq t \leq 1\}$ which are continious on [0, 1] is called weakly convergent to a continious stochastic process $Z = \{Z(t),\ 0 \leq t \leq 1\}$ if for any measurable set of functions $B \in C(0,1)$ with $\mathbf{P}(Z \in \partial B) = 0$ there is convergence $\mathbf{P}(Z_n \in B) \to \mathbf{P}(Z \in B)$.

Let define now the stochastic process $W_n = \{W_n(t),\ 0 \leq t \leq 1\}$. Let $Y_1,\ Y_2,\ \ldots$ be independent and identically distributed random variables, $\mathbf{E}\,Y_1 = a$, $0 < \mathbf{Var}\,Y_1 = \sigma^2 < \infty$. Let $S_0 = 0$, $S_k = Y_1 + \ldots + Y_k$, $k \geq 1$.
Stochastic process W_n is a broken line that is built by points

$$\left(\frac{k}{n},\ \frac{S_k - ka}{\sigma\sqrt{n}}\right), \quad k = 0;\ 1;\ \ldots;\ n.$$

Thus, the substance of the functional central limit theorem: a random polygonal line W_n constructed by centered and normed sums of independent random variables with finite nonzero variance, weakly converges in the space $C(0,1)$ to the Wiener process.
A functional central limit theorem is also called the invariance principle, since the limiting process is the same for any distribution of random variables Y_i with finite nonzero variance.

The usual central limit theorem is a corollary of the functional theorem. We obtain it by considering the processes at the right end of the segment, that is, at point 1:

$$W_n(1) = \frac{S_n - na}{\sigma\sqrt{n}} \Rightarrow W(1) \sim \mathcal{N}_{0,1}.$$

It is rather inconvenient to operate with sets in function spaces. Applications of the functional central limit theorem is based on the weak convergence of continuous functionals.

Functional $g : C(0,1) \to \mathbf{R}$, which assigns real numbers to functions, is said to be continuous (in the uniform metric), if from the convergence $dist(x_n, y) \to 0$ follows $g(x_n) \to g(y)$.

Exercise Prove that $dist(x, 0)$, $\int_0^1 x(t)dt$, $\int_0^1 x^2(t)dt$ are continuous functionals.

Theorem If $Z_n \Rightarrow Z$ in $C(0,1)$ and g is a continuous functional then $g(Z_n) \Rightarrow g(Z)$.

9. Empirical bridge

Let construct a process that is based only on empirical data and converges weakly to standard Brownian bridge W^0 in the case that the data satisfy the conditions of the Functional Central Limit Theorem. This process is called as *empirical bridge*.

We have data $\mathbf{X} = (X_1, \ldots, X_n)$.

Let $S_0 = 0,\ S_k = \sum_{i=1}^{k} X_i,\ k = 1,\ 2,\ \ldots,\ n$.

We calculate $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = S_n/n$, $\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, $s_X^2 = \overline{X^2} - (\overline{X})^2$.

Empirical bridge $Z_n^0 = \{Z_n^0(t), 0 \le t \le 1\}$ is a random broken line with nodes

$$\left(\frac{k}{n};\ \frac{S_k - kS_n/n}{s_X\sqrt{n}}\right), k = 0, \ldots, n.$$

From definition,

$$Z_n^0(t) = \frac{S_k - kS_n/n}{s_X\sqrt{n}} + \frac{nX_{k+1} - S_n}{s_X\sqrt{n}}\left(t - \frac{k}{n}\right),$$

$$\frac{k}{n} \le t < \frac{k+1}{n}, \quad k = 0, \ldots, n-1.$$

Exercise 1 Draw the empirical bridge for $\mathbf{x} = (1,\ 1,\ -1,\ -1)$.
Exercise 2 Draw the empirical bridge for
$\mathbf{x} = (-3,\ 1,\ -2,\ 6,\ 4,\ 1,\ -4,\ 5)$.

Theorem *Let X_1, X_2 ... be independent and identically distributed random variables,* $0 < \sigma^2 = \mathbf{Var} X_1 < \infty$. *Then*

$$Z_n^0 \Rightarrow W^0 \ \text{ in } \ C(0,1).$$

Proof

$$Z_n^0(t) = \frac{\sigma}{s_X}(W_n(t) - tW_n(1))$$

Note that $\frac{\sigma}{s_X} \to 1$ a.s., and $W_n \Rightarrow W$ due to Functional Central Limit Theorem.
As $W^0(t) = W(t) - tW(1)$, and the map $W \to W^0$ is continuous in the uniform metric in $C(0,1)$, we have $Z_n^0 \Rightarrow W^0$.
The proof is complete.

Corollary If assumptions of the Theorem hold then

$$\int_0^1 Z_n^0(t)\,dt \Rightarrow \mathcal{N}_{0,1/12},$$

$$\int_0^1 (Z_n^0(t))^2\,dt \Rightarrow \omega^2,$$

$$J_n^\infty \overset{def}{=} \max_{0\le t\le 1} |Z_n(t)| \Rightarrow K \quad \text{(Kolmogorov distribution)}.$$

So we calculate p-value as

$$\alpha^* = 1 - K(J_n^\infty),$$

$$K(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2t^2},\ t > 0.$$

Exercise Calculate p-value if $J_n^\infty = 4$.

10. Statistical tests

Recall that mathematical statistics constructs such probabilistic models of phenomena, in which the distributions of random variables are unknown or not known completely. *Statistical hypothesis* is a statement about the distribution of random variables, participating in the description of the model.
The statistical hypothesis can either fix a single distribution (such a hypothesis is called *simple*), or allocate a class of distributions consisting of more than one distribution (such a hypothesis is called *complex*).
Most often, as a complex hypothesis, we propose a statement about the belonging of a distribution to a certain parametric family.

Hypotheses, as a rule, are denoted by the Latin letter H (from the word *hypothesis*) with lower indices. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random sample, $\mathbf{X} \sim \mathbf{P}$, and $\mathbf{P}$ is a distribution of any one random variable X_i. This distribution is unknown (completely or partually).

Statistical test is a rule based on which a sample is associated with one of the hypotheses. More rigorously, the statistical criterion for sample size n is a function from a sample space (of the entire space $\mathbf{R}^n$ or of the set G^n, where $G \subset \mathbf{R}$ is the set on which the sample values are concentrated according to a priori assumptions) into the set of hypotheses. The number of hypotheses is always greater than one and can be finite or infinite (countable or uncountable).

We will consider a situation where there are only two hypotheses. One of them is called *basic*, and the other is called *alternative*, denoting H and $\overline{\mathrm{H}}$, respectively. In this situation, a statistical criterion is any rule that allows based on the observable sample vector $\mathbf{X}$ choose one of the hypotheses: basic or alternative. It is convenient to represent the statistical test as a function $\delta(\mathbf{X})$ of the sample vector, taking two values: H and $\overline{\mathrm{H}}$. The most common approach for construction of statistical tests is as follows.

$$\delta(\mathbf{X}) = \overline{H},\ \text{if}\ J(\mathbf{X}) \in J_\alpha,$$
$$\delta(\mathbf{X}) = H,\ \text{if}\ J(\mathbf{X}) \notin J_\alpha.$$

This rule is based on common sense: it prescribes to reject the hypothesis H (that is, accept $\overline{H}$), if event $\{J(\mathbf{X}) \in J_\alpha\}$ occures, which should not happen, be the hypothesis H true. The number $\alpha > 0$ is called *level of significance*, the statistic $J(\mathbf{X})$ is called by *test statistics*, and the set J_α by *a critical set*.

When applying a statistical test, errors of two kinds can arise. The error of the first kind is that the true basic hypothesis is rejected. An error of the second kind is that the true alternative hypothesis is rejected.

accepted hypothesis	hypothesis H is true	hypothesis $\overline{H}$ is true
H	no error	error of 2nd kind
$\overline{H}$	error of 1st kind	no error

A criterion is characterized by error probabilities:

$$\alpha_1 = \mathbf{P}_H(H \text{ is rejected}); \qquad \alpha_2 = \mathbf{P}_{\overline{H}}(\overline{H} \text{ is rejected}).$$

Here, the subscript of the probability symbol indicates hypothesis under which the probability is counted.

Statistics $J_n = J_n(\mathbf{X})$ should have the following properties:
1) if hypothesis H holds, statistic J_n has a known distribution or, at least, converges weakly to some random variable J with known distribution;
2) if hypothesis $\overline{H}$ holds, statistic J_n converges almost surely to infinity with an increase of the sample size.

The convergence of statistics J_n almost surely to infinity under the basic hypothesis guarantees the *consistency* of the test, that is, convergence of the error probability of the second kind α_2 to zero with increasing sample size.
For each sample realization **x** one can find the limit level $\alpha^* = \alpha^*(\mathbf{x})$ under which the hypothesis H can still be accepted. This value is called *p-value*. P-value α^* has the sense of probability to get worse agreement with the hypothesis being tested than real obtained if the hypothesis H is true. Therefore, the less α^*, the more it speaks against the hypothesis H.

P-value is calculated using the distribution of J:

$$\alpha^* = \mathbf{P}\{J \geq J(\mathbf{X})\} = 1 - F_J(J(\mathbf{X})).$$

In terms of p-value, the critical area has the following form

$$\mathrm{J}_\alpha = \{\alpha^* \leq \alpha\},$$

that is, the basic hypothesis is rejected at level α in the case $\alpha^* \leq \alpha$.

Example

We test the model of independent identically distributed random variables:
1) for air temperatures $y_1, \ldots, y_{30}$ in Moscow in November 2011 (data taken from the site
http://academic.udayton.edu/kissock/http/Weather/default.htm
and because the temperature is in Fahrenheit),
2) for temperature increments $x_i = y_{i+1} - y_i$, $i = 1, \ldots, 29$.

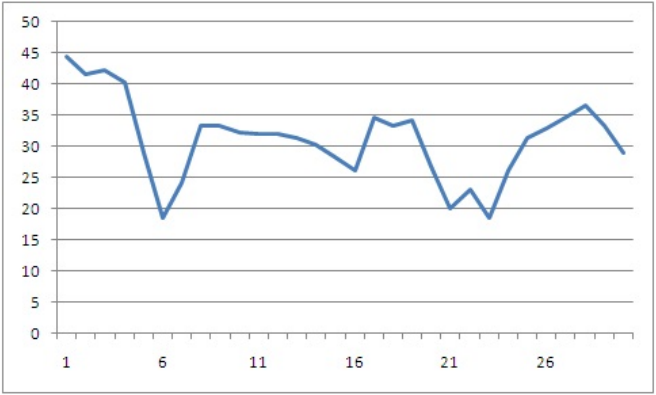


Figure: Temperature in Moscow in November 2011 in Fahrenheit

We use tests based on statistics

$J_n^\infty = \max_{t\in[0,1]} |Z_n(t)|$,

$J_n^{(1)} = 2\sqrt{3}\left|\int_0^1 Z_n(t)dt\right|$.

If the basic hypothesis is true then distribution of J_n^∞ converges to the Kolmogorov distribution, and distribution of statistics $J_n^{(1)}$ converges to distribution of absolete value of a standard normal random variable.

1)

$$J_n^\infty = \max_{1\le k\le n-1} \left| \frac{n\sum_{i=1}^k y_i - k\sum_{i=1}^n y_i}{s_y n\sqrt{n}} \right|,$$

$$J_n^{(1)} = \frac{\sqrt{3}}{s_y n\sqrt{n}} \left| \sum_{k=1}^n \left(\sum_{i=1}^{k-1} y_i + \sum_{i=1}^k y_i - \frac{2k-1}{n} \sum_{i=1}^n y_i \right) \right|.$$

Here $n = 30$, s_y is a sample standard deviation.

We have $J_n^\infty \approx 1.24$, p-value $\alpha^* = 1 - K(1.24) \approx 0.0929$.
$J_n^{(1)} \approx 1.77$, p-value $\alpha^* = 2(1 - \Phi_{0,1}(1.77)) \approx 0.0763$.
Both criteria are rejected the basic hypothesis at level 0.1, but accept it at level 0.05.
A rather poor correspondence of the random sample model to the data is explained by the fact that the temperatures in sequential days are significantly dependent.

2) We calculate the same statistics for temperature increments $x_i = y_{i+1} - y_i$, $i = 1, \ldots, 29$.

$$J_n^{\infty} = \max_{1 \le k \le n-1} \left| \frac{n \sum_{i=1}^{k} x_i - k \sum_{i=1}^{n} x_i}{s_x n \sqrt{n}} \right|,$$

$$J_n^{(1)} = \frac{\sqrt{3}}{s_x n \sqrt{n}} \left| \sum_{k=1}^{n} \left(\sum_{i=1}^{k-1} x_i + \sum_{i=1}^{k} x_i - \frac{2k-1}{n} \sum_{i=1}^{n} x_i \right) \right|.$$

Here $n = 29$, s_x is a sample standard deviation.

$J_n^\infty \approx 0,865$, p-value $\alpha^* = 1 - K(0,865) \approx 0,442$;
$J_n^{(1)} \approx 0,738$, p-value $\alpha^* = 2(1 - \Phi_{0,1}(0,738)) \approx 0,461$.
There is no reason to reject the hypothesis of homogeneity of increments.

Thus, the hypothesis of homogeneity of increments corresponds better to the data. According to this hypothesis, the average temperature change in November 2011 in Moscow is $\overline{x} \approx -0.397$ F per day with a sample standard deviation $s_x \approx 2.79$ F per day.

Analysis of text homogeneity

For analysis of the homogeneity of a text, one must define a map of text to a sequence of numbers.

We have developed a program that maps a text to
the sequence of indicators of occurrence of words in
dictionary of service words (prepositions, conjunctions, particles).

This is so-called dictionary of author's invariant.

In this example we analyse 25 fiction texts.

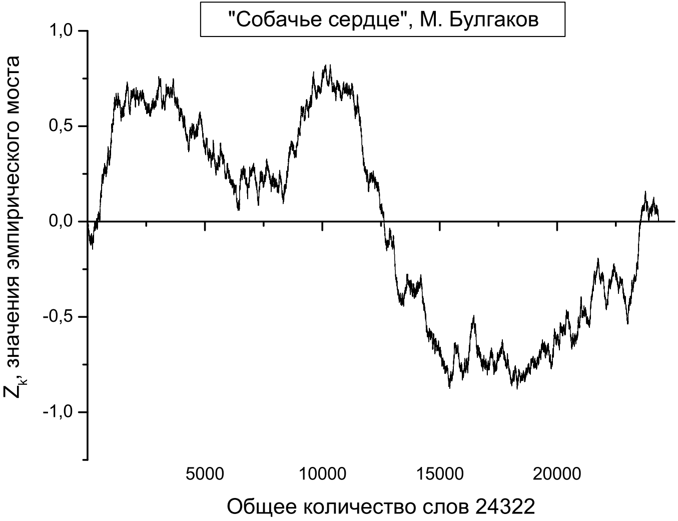
The texts were studied singly and in contaminations.

There are $25 \times 24 = 600$ pairwise contaminations of texts, including $3! + 9 \times 2 = 24$ contaminations of texts by one author and 576 by different authors.

For these texts, empirical bridges Z_n^0 were calculated.

We calculated $J_n^\infty = \max_{t \in [0;\ 1]} |Z_n^0(t)|$ and p-values

$$\alpha^* = 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 |J_n^\infty|^2}.$$


"Собачье сердце", М. Булгаков
1,0
0,5
0,0
-0,5
-1,0
Z_k, значения эмпирического моста
5000
10000
15000
20000
Общее количество слов 24322

Two texts by the same author

	ν_n	T_n	M_n	n	ε^*
aelit+giper	0.1680	35709	2.2519	112342	7.87431E-05
be-god+grad	0.1877	47346	2.7534	156524	5.19998E-07
chapaev+insec	0.2007	96633	1.7162	135134	0.005531114
chapaev+pel-g	0.1966	106004	1.6702	157081	0.007553569
dogheart+master	0.2117	33497	4.1050	136976	4.61903E-15
giper+aelit	0.1680	62746	2.5383	112342	5.06749E-06
grad+be-god	0.1877	35301	2.9651	156524	4.61764E-08
gramota+zhiwago	0.2019	82610	3.6979	178858	2.65059E-12
insec+chapaev	0.2007	48889	1.3334	135134	0.057121241
insec+pel-g	0.1979	59715	2.5414	110792	4.90759E-06

Novels *Virgin soil upturned* and *The Quiet Don* by Mikhail Sholokhov

novel	n	ν_n	T_n	M_n	ε^*
Virgin soil upturned	204955	0.21	97523	6.7	9.4E-40
The Quiet Don	421854	0.18	210643	9.0	5.61E-71

Here each of the novels gives very large deviation values of empirical bridge that not characteristic for any other texts never for a couple of novels of one author nor for single. The change point on the graphs turns out to be extremely clear.

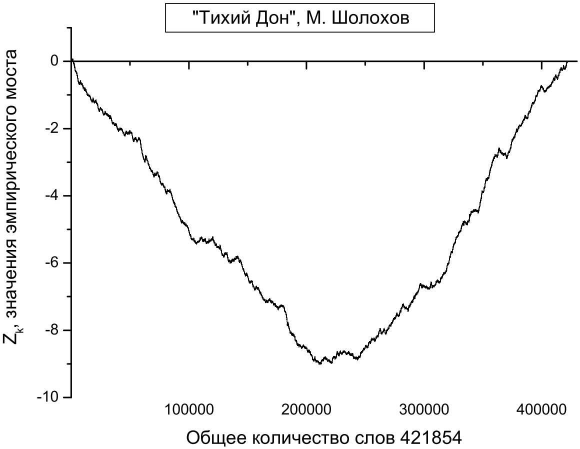


Figure: *The Quiet Don* by Mikhail Sholokhov

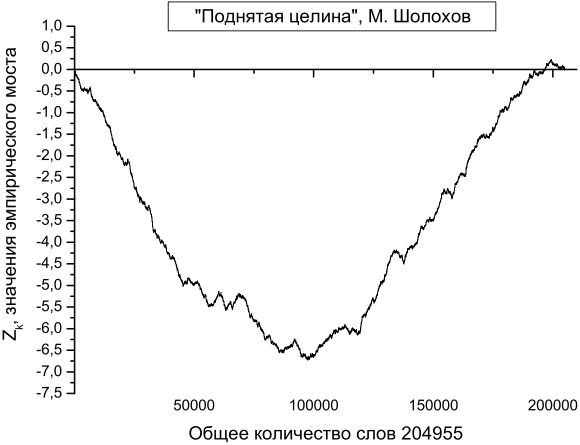


Figure: *Virgin soil upturned* by Mikhail Sholokhov

11. Order statistics

If random variables $X_1, \ldots, X_n$ are sorted in ascending order, then we get a new random variables, called *order statistics*:

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n-1)} \leq X_{(n)}.$$

So $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Empirical distribution function $F_n^*(t)$ is called the related frequency of elements that smaller than the given t. Empirical distribution function corresponding to sample $\mathbf{X} = (X_1, X_2, ..., X_n)$, can be constructed from this sample using any of the following formulas:

$$F_n^*(t) = \frac{\{\text{number of } X_i : X_i \leq t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i \leq t).$$

Here

$$\mathbf{I}(X_i \leq t) = \begin{cases} 1, & \text{if } X_i \leq t; \\ 0 & \text{else}; \end{cases}$$

is an indicator of event $\{X_i \leq t\}$.

Let $GL_F(t) = \int\limits_0^t F^{-1}(s)\, ds$
be the *theoretical general Lorenz curve*
(Gastwirth, 1971; Davydov and Zitikis, 2004)
where $F^{-1}(s) = \sup\{x : \ F(x) < s\}$, $0 < s < 1$,
is the inverse of distribution function $F(x)$.
Let $GL_F^0(t) = GL_F(t) - tGL_F(1)$ be its centered version.
Exercise 1 Find F^{-1}, GL_F and GL_F^0 for pdf $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.
Exercise 2 Find F^{-1}, GL_F and GL_F^0 if a random variable takes values 0 and 1 with probabilities $1/2$.

Similarly, let $GL_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_{(n)}$
be the *empirical Lorenz curve*.
$GL_n^0(t) = GL_n(t) - tGL_n(1)$ is its centered version.
Exercise Find F_n^{-1}, GL_n and GL_n^0 for $\mathbf{x} = (3,\ 1,\ 2,\ 2)$.
Goldie (1977) showed that, as $n \to \infty$, the empirical Lorenz curve converges a.s. to the theoretical curve in the uniform metric, i.e. $\sup_{t\in\mathbf{R}} |GL_n(t) - GL_F(t)| \to 0$ a.s.

12. Regression on order statistics

Brown et al. (1975) proposed a test for change of regression at unknown time. Their approach is based on computation of recursive residuals. MacNeill (1978) studied a linear regression against values of continuously differentiable functions. He obtained limit processes for sequences of partial sums of regression residuals. Later Bischoff (1997) showed that the MacNeill's theorem holds in more general setting, namely for continuous regressor functions. Aue et al. (2008) introduced a new test for polynomial regression functions which is analogous to the classical likelihood test. Stute (1997) proposed a class of tests that are based on regression residuals for one-parametric case.

We consider a model of a simple linear regression on order statistics. The need of this model comes from applications. Kovalevskii (2013) analysed dependence of logarithm of a car price on a production year basing on a list of ads. We have Y_i (a logarithm of price in this example) that is assumed to depend linearly on production year X_i and noise ε_i with zero mean. Then we reorder the data to correspond to acsending order of X_i. We need in a statistical test to verify the model.

To define the model, we introduce 2 mutually independent families of random variables:

1) $\{\varepsilon_i, i \geqslant 1\}$, a family of independent identically distributed random variables, $\mathbf{E}\,\varepsilon_1 = 0$, $\mathbf{Var}\,\varepsilon_1 = \sigma^2 > 0$;

2) $\{X_i\}_{i=1}^{\infty}$, a sequence of i.i.d. random variables with distribution function F and finite positive variance $\mathbf{Var}\,X_1$.

A regression model before ordering:

$$Y_i = a + bX_i + \varepsilon_i, \; i = 1, \ldots, n.$$

So we have a three-dimensional vector $(Y_i, X_i, \varepsilon_i)$. Then we order it on the second component (X_i) and obtain vector $(Y_{ni}, X_{ni}, \varepsilon_{ni})$. Here $X_{ni} = X_{i:n} = X_{(i)}$ is the i-th order statistic of the first n random variables $X_1, \ldots, X_n$. In particular, $X_{n1} = \min_{1 \leqslant i \leqslant n} X_i$ and $X_{nn} = \max_{1 \leqslant i \leqslant n} X_i$. Values Y_{ni}, ε_{ni} are values of Y and ε corresponding to X_{ni} (that is, induced order statistics, concomitants).

Exercise

Order by X_i's

X_i	Y_i
3	1
2	3
1	0
2	2

We have the following regression model after ordering:

$$Y_{ni} = a + bX_{ni} + \varepsilon_{ni}, \ i = 1, \ldots, n.$$

For this model, we introduce an *empirical bridge* and show its weak convergence to a centered Gaussian process.
Let

$$\widehat{b}_n = \frac{\overline{XY} - \overline{X}\ \overline{Y}}{\overline{X^2} - \overline{X}^2}, \ \widehat{a}_n = \overline{Y} - \widehat{b}_n \overline{X}$$

be the classical Gauss-Markov estimators for a and b. Here $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_{ni} = \frac{1}{n}\sum_{i=1}^n X_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_{ni} = \frac{1}{n}\sum_{i=1}^n Y_i$ etc.
Note that a sum on all i does not depend on order, therefore estimators coincide for models before and after ordering.

Exercise

Calculate $\widehat{b}_n$ and $\widehat{a}_n$

X_{ni}	Y_{ni}
1	0
2	3
2	2
3	1

Define *fitted values* $\widehat{Y}_{ni}$,
regression residuals $\widehat{\varepsilon}_{ni}$
and their *partial sums* $\widehat{\Delta}_{ni}$, by
$\widehat{Y}_{ni} = \widehat{a}_n + \widehat{b}_n X_{ni}$,
$\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$ and
$\widehat{\Delta}_{ni} = \ \widehat{\varepsilon}_{n1} + \ \ \ldots + \ \widehat{\varepsilon}_{ni}$
for $1 \leqslant i \leqslant n$, $\ \widehat{\Delta}_{n0} = 0$.
Note that $\widehat{\Delta}_{nn} = 0$.

Exercise

Calculate fitted values $\widehat{Y}_{ni}$, regression residuals $\widehat{\varepsilon}_{ni}$ and their partial sums $\widehat{\Delta}_{ni}$

X_{ni}	Y_{ni}
1	0
2	3
2	2
3	1

An *empirical bridge* is a random polygon $\widehat{Z}_n^0$ with nodes

$$\left(k/n, \quad \widehat{\Delta}_{nk}/\sqrt{ns_\varepsilon^2}\right), \; k = 0, \ldots, n,$$

where $s_\varepsilon^2 = \overline{\widehat{\varepsilon}^2}$ is an estimator of variance σ^2.

Exercise

Draw an empirical bridge of regression residuals

X_{ni}	Y_{ni}
1	0
2	3
2	2
3	1

Theorem 1 *The empirical bridge* $\widehat{Z}_n^0$ *converge weakly, as* $n \to \infty$*, to the centered Gaussian process* Z_F *with covariance function,* $K_F^0(t,s)$*, given by*

$$K_F^0(t,s) = \min\{t,s\} - ts - \frac{GL_F^0(t)GL_F^0(s)}{\mathbf{Var}X_1}, \; t,s \in [0,1].$$

Here weak convergence holds in the space $C(0,1)$ *of continuous functions on [0,1] endowed by the uniform metric.*

Exercise 1 Find $K_F^0(t,s)$ if $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$.
Exercise 2 Find $K_F^0(t,s)$ if X takes values 0 and 1 with probabilities $1/2$.

Proof of Theorem 1

Let $X_{ni}^0 = X_{ni} - \overline{X}$, $\varepsilon_{ni}^0 = \varepsilon_{ni} - \overline{\varepsilon}$ where $\overline{\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ni} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i$ because the sum on all i does not depend on order.

The proof includes four steps. In the first step, we show that, in the formulae under consideration, the sum $\sum_{i=1}^{n} \frac{\varepsilon_{ni}^{0} X_{ni}^{0}}{\sqrt{n}}$ may be replaced by the sum $\sum_{i=1}^{n} \frac{\varepsilon_{ni}^{0} \mathbf{E} X_{ni}^{0}}{\sqrt{n}}$.

Secondly, we prove weak convergence of a normalized vector with coordinates $(\widehat{\Delta}_{nk_1}, \ldots, \widehat{\Delta}_{nk_m})$ to a normalized vector with coordinates $(\Delta_{nk_1}, \ldots, \Delta_{nk_m})$ where Δ_{nk_i} are defined below. Then we prove weak convergence of finite-dimensional distributions.

The third step contains a proof of relative compactness of the family $\{\widehat{Z}_n(t), 0 \leqslant \ t \leqslant 1\}$. We complete with a proof of convergence of sample variance s_{ε}^2 to variance σ^2.
In what follows, notation $\overset{\mathbf{p}}{\to}$ states for convergence in probability.

Step 1

Note that

$$\widehat{\Delta}_{nk} = \sum_{i=1}^{k} \left(\varepsilon_{ni}^{0} - \frac{\overline{X^0 \varepsilon^0}}{\overline{(X^0)^2}} X_{ni}^{0} \right).$$

We show that

$$\frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \varepsilon_{ni}^{0} X_{ni}^{0} - \sum_{i=1}^{n} \varepsilon_{ni}^{0} \mathbf{E} X_{ni}^{0} \right) \xrightarrow{\mathbf{p}} 0.$$

We use Theorem 1 of Höeffding (1953) that implies
$\frac{1}{n}\sum\limits_{i=1}^{n} \mathbf{Var} X_{ni} \to 0$ as $n \to \infty$.
Note that $\mathbf{Var}\overline{X} = \mathbf{Var}X_1/n$,
$\frac{1}{n}\sum_{i,j=1}^{n} \mathbf{cov}(X_{ni}, X_{nj}) = \frac{1}{n}\mathbf{Var}\sum_{i=1}^{n} X_{ni} = \mathbf{Var}X_1$.

As $\sum\limits_{i=1}^{n}(X_{ni}^{0} - \mathbf{E}X_{ni}^{0}) = 0$ we have

$$\sum_{i=1}^{n} \varepsilon_{ni}^{0}(X_{ni}^{0} - \mathbf{E}X_{ni}^{0}) = \sum_{i=1}^{n} \varepsilon_{ni}(X_{ni}^{0} - \mathbf{E}X_{ni}^{0}).$$

Due to Chebyshev's inequality,

$$\mathbf{P}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_{ni}(X_{ni}^{0}-\mathbf{E}X_{ni}^{0})\right|\geqslant\delta\right\}\leqslant\frac{\mathbf{Var}\sum_{i=1}^{n}\varepsilon_{ni}(X_{ni}^{0}-\mathbf{E}X_{ni}^{0})}{n\delta^{2}}.$$

As $\{\varepsilon_{ni}\}$ are independent and do not depend on $\{X_{ni}\}$ we have

$$\mathbf{Var}\sum_{i=1}^{n}\varepsilon_{ni}(X_{ni}^{0}-\mathbf{E}X_{ni}^{0})=\sum_{i=1}^{n}\mathbf{Var}\varepsilon_{ni}\mathbf{Var}\left(X_{ni}^{0}-\mathbf{E}X_{ni}^{0}\right)$$

$$=\sum_{i=1}^{n}\mathbf{Var}\varepsilon_{ni}\mathbf{Var}X_{ni}^{0}.$$

$$\sum_{i=1}^{n} \mathbf{Var} X_{ni}^{0} = \sum_{i=1}^{n} \mathbf{Var} X_{ni} - 2\sum_{i=1}^{n} \mathbf{cov}(X_{ni}, \overline{X}) + n\mathbf{Var}\overline{X}$$

$$= \sum_{i=1}^{n} \mathbf{Var} X_{ni} - \frac{2}{n} \sum_{i,j=1}^{n} \mathbf{cov}(X_{ni}, X_{nj}) + \mathbf{Var} X_1$$

$$= \sum_{i=1}^{n} \mathbf{Var} X_{ni} - \mathbf{Var} X_1 = o(n).$$

Thus

$$\frac{1}{n}\mathbf{Var}\sum_{i=1}^{n}\varepsilon_{ni}(X_{ni}^{0}-\mathbf{E}X_{ni}^{0})\to 0.$$

So (2.1) is proved.

Step 2 Let $\lfloor t \rfloor$ be the integer part of t. For any fixed m and for $0 \leqslant s_1 < \cdots < s_m \leqslant 1$, $k_i = \lfloor n s_i \rfloor$, we establish weak convergence, as $n \to \infty$, of vector $\vec{\eta} = \frac{1}{\sigma\sqrt{n}}(\widehat{\Delta}_{nk_1}, \ldots, \widehat{\Delta}_{nk_m})$ to vector $\vec{Z}_F = (Z_F(s_1), \ldots, Z_F(s_m))$.

From (2.1) and from convergences $\overline{(X^0)^2} \to \mathbf{Var}X_1$ a.s.,

$$\frac{1}{n}\sum_{i=1}^{k_i} X_{ni}^0 \to \ GL_F^0(s_i)$$

a.s. (Goldie, 1975), it is enough to prove $\vec{\zeta} \Longrightarrow \vec{Z}_F$ where $\vec{\zeta} = \ \frac{1}{\sigma\sqrt{n}}(\Delta_{nk_1}, \ldots, \Delta_{nk_m})$,

$$\Delta_{nk_j} = \sum_{i=1}^{k_j} \varepsilon_{ni}^0 - \frac{GL_F^0(s_j)}{\mathbf{Var}X_1}\sum_{i=1}^{n} \varepsilon_{ni}^0 \mathbf{E}X_{ni}^0 = \sum_{i=1}^{k_j} \varepsilon_{ni}^0 - \frac{GL_F^0(s_j)}{\mathbf{Var}X_1}\sum_{i=1}^{n} \varepsilon_{ni} \mathbf{E}X_{ni}^0.$$

We prove weak convergence $\vec{\zeta} \Longrightarrow \vec{Z}_F^0$ using characteristic function

$$\varphi_{\vec{\zeta}}(\mathbf{t}\ ) = \mathbf{E} \prod_{j=1}^{m} \exp\left(\mathbf{i} \frac{t_j \widehat{\Delta}_{nk_j}}{\sigma\sqrt{n}} \right).$$

Notice that

$$\sum_{j=1}^{m} t_j \left(\sum_{i=1}^{k_j} (\varepsilon_{ni} - \overline{\varepsilon}) - \frac{GL_F^0(s_j)}{\mathbf{Var} X_1} \sum_{i=1}^{n} \varepsilon_{ni} \mathbf{E} X_{ni}^0 \right)$$

$$= \sum_{i=1}^{n} \varepsilon_{ni} \sum_{j=1}^{m} t_j \left(\mathbf{I}\{i \leqslant k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\mathbf{Var} X_1} \mathbf{E} X_{ni}^0 \right).$$

It is well known that the finiteness of $\mathbf{E}\psi_1$ implies convergence $\frac{\psi_{n:n}}{n} \to 0$ a.s. and in mean for a sequence of i.i.d random variables $\psi_1, \ldots, \psi_n$, and, more generally, for a stationary ergodic sequence as a consequence of the subadditive ergodic theorem (Kingman, 1968).

Applying this fact and using Hőlder's inequality we have $\mathbf{E}X_{ni}^0 = o(\sqrt{n})$ uniformly in $1 \leqslant i \leqslant n$.

Let $\beta_{ni} = \sum_{j=1}^{m} t_j \left(\mathbf{I}\{i \leqslant k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\mathbf{Var} X_1} \mathbf{E} X_{ni}^0 \right)$. Then $\beta_{ni}/\sqrt{n} \to 0$,

$$\sum_{i=1}^{n} \frac{\beta_{ni}^2}{n} \to C_F := \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} t_{j_1} t_{j_2} K_F(s_{j_1}, s_{j_2}).$$

As for any $t \to 0$

$$\mathbf{E}e^{\mathbf{i}t\varepsilon^v_{ni}} = -\frac{1}{2}t^2\mathbf{Var}\varepsilon^v_{ni}\left(1+o(1)\right),$$

and $\mathbf{Var}\varepsilon_{ni} \to \sigma^2$ as $i, n \to \infty$, then we have by integration on Markov chain distribution (as in Step 1)

$$\varphi_{\vec{\zeta}}(\mathbf{t}\ ) = \mathbf{E}\prod_{j=1}^{m}\exp\left(\mathbf{i}\frac{\varepsilon_{ni}\beta_{ni}}{\sigma\sqrt{n}}\right)$$

$$= \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{\beta^2_{nj}\mathbf{Var}\varepsilon_{ni}}{n\sigma^2}\right)(1+o(1)) \to \exp(-C_F/2).$$

So we have $\varphi_{\vec{\zeta}}(\mathbf{t}\ ) \to \exp(-C_F/2)$. Thus convergence of finite-dimensional distributions is proved.

Step 3 We show that the family of distributions $\{\widehat{Z}_n(t), 0 \leqslant t \leqslant 1\}$ is relatively compact.

Let $S_{nk} = \sum\limits_{i=1}^{k} X_{ni},\ k = 1, \ldots, n,\ S_{n0} = 0.$

By Prokhorov's theorem (section 1 §6 in Billingsley, 1968) it suffices to show that the family of distributions of random processes $\left\{\frac{\widehat{\Delta}_{n,\lfloor nt\rfloor}}{\sigma\sqrt{n}},\ 0 \leqslant t \leqslant 1\right\}$, $n = 1, 2, \ldots,$ is tight. Put $k = \lfloor nt \rfloor$ and let

$$\widehat{\Delta}^0_{nk} = \sum_{i=1}^{k}\left(\varepsilon_{ni} - \frac{\overline{X^0\varepsilon^0}}{\overline{(X^0)^2}} X_{ni}\right).$$

Then $\widehat{\Delta}_{nk} = \widehat{\Delta}^0_{nk} - \frac{k}{n}\widehat{\Delta}^0_{nn}$.
As $\{\varepsilon^v_{ni}\}$ are i.i.d. for any v, the invariance principle implies tightness of the family $\left\{\frac{\sum_{i=1}^{\lfloor nt\rfloor}\varepsilon^v_{ni}}{\sigma\sqrt{n}}, 0 \leqslant t \leqslant 1\right\}$ for any $v \in \{1, \ldots, M\}$. Thus $\left\{\frac{\sum_{i=1}^{\lfloor nt\rfloor}\varepsilon_{ni}}{\sigma\sqrt{n}}, 0 \leqslant t \leqslant 1\right\}$ is tight. Invariance principle for this Markov-modulated sequence goes from Corollary 3.8 (McLeish, 1975).

So, it is enough to establish tightness of

$$\left\{ \frac{\overline{X^0 \varepsilon^0} \sqrt{n}}{\sigma \overline{(X^0)^2}} \frac{S_{n,\lfloor nt \rfloor}}{n}, 0 \leqslant t \leqslant 1 \right\}.$$

In turn, by Theorem 8.3 (Billingsley, 1968), it suffices to prove that, for any $\varepsilon > 0,\ \alpha > 0$, there are $0 < \ \delta < \ 1,\ n_0 \in \mathbf{N}$ such that

$$\frac{1}{\delta}\mathbf{P}\left\{\sup_{t\leqslant s\leqslant t+\delta}\left|\frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}}\frac{S_{n,\lfloor ns\rfloor}-S_{n,\lfloor nt\rfloor}}{n}\right|\geqslant\varepsilon\right\}\leqslant\alpha, \qquad (2)$$

for all $n > n_0,\ 0 \leqslant t \leqslant 1$.

Notice that $\frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}} \Longrightarrow \frac{\zeta}{\sqrt{\mathbf{Var}X_1}}$, and (Goldie, 1977)

$$\sup_{t\leqslant s\leqslant t+\delta}\left|\frac{S_{n,\lfloor ns\rfloor}-S_{n,\lfloor nt\rfloor}}{n}\right|\to\sup_{t\leqslant s\leqslant t+\delta}|GL_F(s)-GL_F(t)| \text{ a.s..}$$

Here ζ is a standard normal random variable and $GL_F(x)$ is the general Lorenz curve.
By Cauchy-Bunyakowsky inequality,

$$\sup_{t\leqslant s\leqslant t+\delta}|GL_F(s)-GL_F(t)|\leqslant\sup_{t\leqslant s\leqslant t+\delta}\int_t^s|F^{-1}(x)|dx\leqslant\sqrt{\delta\mathbf{E}X_1^2}.$$

So one may choose a positive δ that satisfies (2.2).

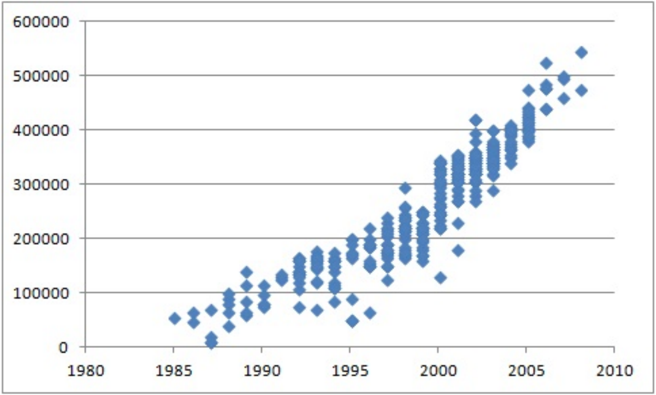
Step 4 It remains to prove $s_\varepsilon^2 \xrightarrow{\mathbf{p}} \sigma^2$. Indeed,

$$\overline{\widehat{\varepsilon}^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_{ni} - \overline{\varepsilon} - \frac{\overline{X^0\varepsilon^0}}{\overline{(X^0)^2}}(X_{ni} - \overline{X})\right)^2 = \overline{(\varepsilon^0)^2} - \frac{(\overline{X^0\varepsilon^0})^2}{\overline{(X^0)^2}} \xrightarrow{\mathbf{p}} \sigma^2.$$

This completes the proof of Theorem 1.

Example 1. A Regression Model for Prices of Second-Hand Cars

We analyse ads about sales of Toyota Corolla cars at www.ngs.ru on 02.06.2012. There are 525 ads. We explore a regression of logarithm of a sale price against a date of the ad, a steering wheel position (left or right), a year of production, an engine volume, gearboxes type, milage. Standard regression analysis gives p-values lesser than 0.01 only for a steering wheel position and a year of production. The number of cars with a left wheel is relatively small, so we choose right-wheeled cars (382 ads). We investigate dependence of price Y_i against production year X_i. Ads are ordered by the year. The order is random for cars of a same year.



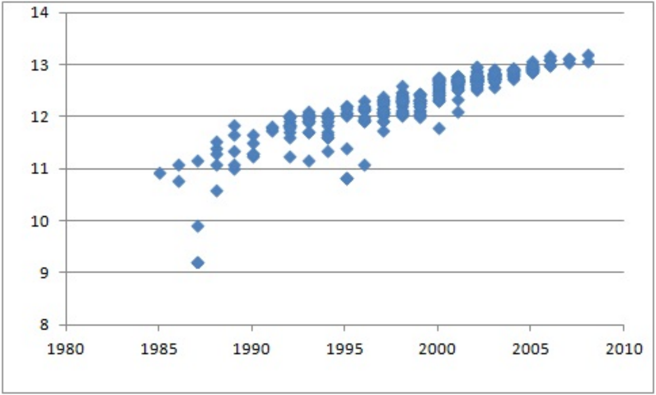
Prices of right-wheeled cars (in roubles, 382 ads)

The model is

$$\ln Y_i = aX_i + b + \varepsilon_i,\ \ i = 1, \ldots, n. \tag{3}$$

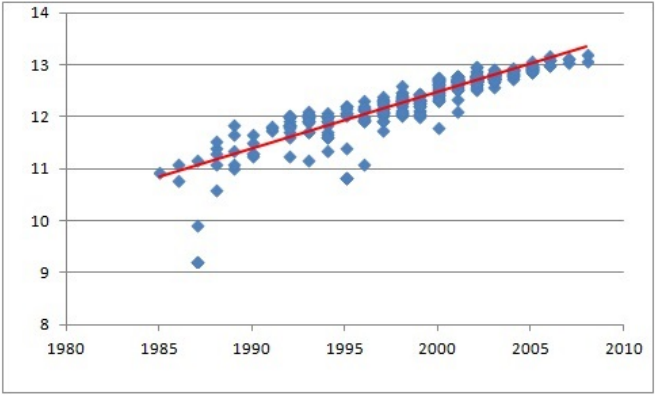
Here ε_i are independent and identically distributed, have zero mean and non-zero finite variance. Estimates of a and b are approximately

$$\widehat{a} = 0.1089,\ \ \widehat{b} = -205.3.$$



Logarifms of prices

We estimate Y_i and calculate regression residuals. The sample standard deviation of regression residuals is $S = 0.2469$.

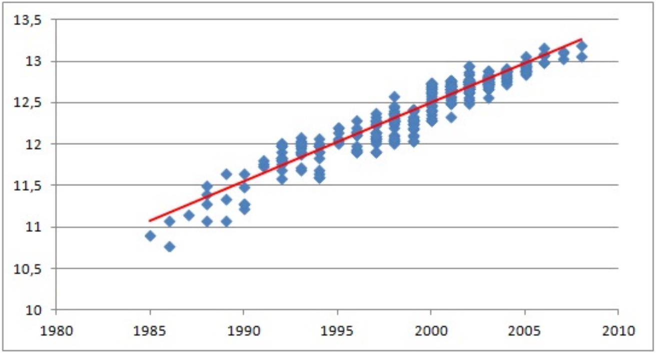


Logarifms of prices with a trend line

Then we delete consequently ads with regression residuals that are absolutely greater then 3-multiplied sample standard deviation (which is recalculated after each ad deletion). 364 ads remains after deletion, parameters estimations for it

$$\widehat{a} = 0.09558, \ \widehat{b} = -178.7, \ S = 0.1291.$$

We have decreased S almost twofold.

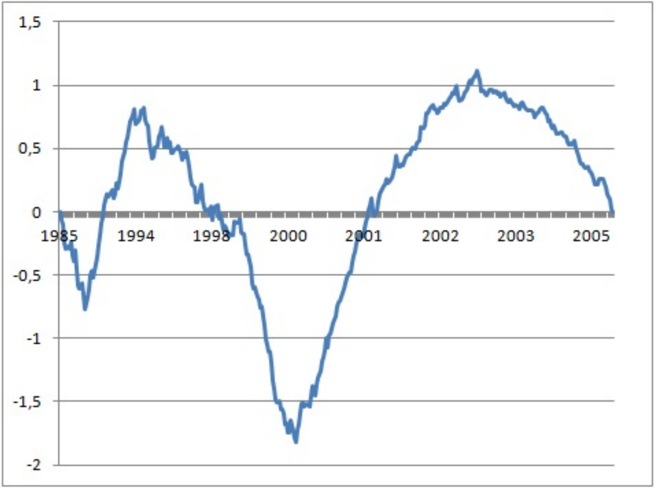


Logarifms of prices after 3-sigma procedure

We calculate an empirical bridge of regression residuals.
Let $\widehat{Y}_i = \widehat{a} + \widehat{b}X_i$, $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$, $\widehat{\Delta}_i^0 = \widehat{\varepsilon}_1 + \ \ldots + \widehat{\varepsilon}_i$.
The empirical bridge is a random polygon $\widehat{Z_n}$ with nodes

$$\left(\frac{k}{n}, \ \frac{\widehat{\Delta_k^0} - \frac{k}{n}\widehat{\Delta_n^0}}{\sqrt{s_\varepsilon^2 n}} \right) = \left(\frac{k}{n}, \ \frac{\widehat{\Delta_k^0}}{\sqrt{s_\varepsilon^2 n}} \right)$$

where $s_\varepsilon^2 = \overline{\widehat{\varepsilon}^2} - (\overline{\widehat{\varepsilon}})^2 = \overline{\widehat{\varepsilon}^2}$



The empirical bridge

The basic hypothesis is model (1):

$$\ln Y_i = aX_i + b + \varepsilon_i, \ i = 1, \ldots, n.$$

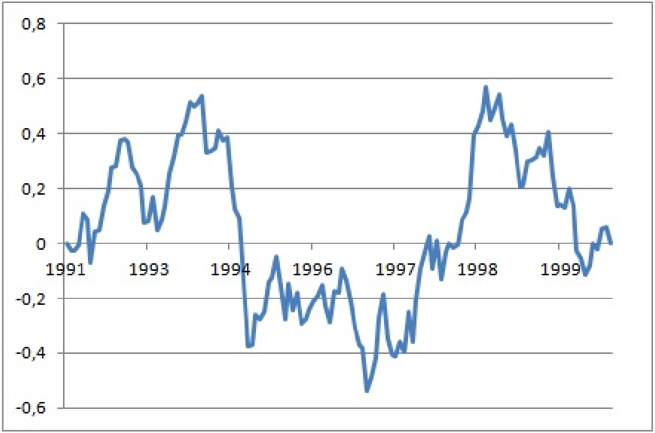
As distribution of an empirical bridge is close to distribution of a standard brownian bridge under this hypothesis, we neglect the basic hypothesis on level 0.01 on a base on maximal deviation of the empirical bridge from an absciss axe. Therefore we propose a new hypothesis: model (1) for each of intervals between sharp peaks of the empirical bridge, that is, for intervals from 1 to 16, from 17 to 154, from 155 to 364. We don't analyse points from 1 to 16 due to a small number of points. We slightly correct intervals to correspond production years. Interval from 17 to 148 corresponds to years from 1991 to 1999, and interval from 149 to 364 corresponds to years from 2000 to 2008.

We analyse each of these two intervals. For the first one

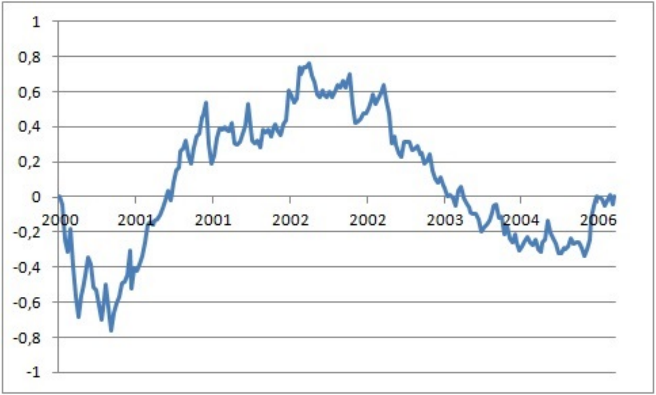
$$\widehat{a} = 0.06484, \ \widehat{b} = -117.3, \ S = 0.1356.$$

For the second one

$$\widehat{a} = 0.07144, \ \widehat{b} = -130.3, \ S = 0.08699.$$



The empirical bridge for 1991–1999



The empirical bridge for 2000–2008

The empirical bridge is a random polygon $\widehat{Z_n^0}$ with nodes

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta_k^0} - \frac{k}{n}\widehat{\Delta_n^0}}{\sqrt{s_\varepsilon^2 n}} \right) = \left(\frac{k}{n},\ \frac{\widehat{\Delta_k^0}}{\sqrt{s_\varepsilon^2 n}} \right)$$

where $s_\varepsilon^2 = \overline{\widehat{\varepsilon}^2} - (\overline{\widehat{\varepsilon}})^2 = \overline{\widehat{\varepsilon}^2}$.
Denote by

$$GL_F(t) = \int\limits_0^t F^{-1}(s)\,ds$$

a theoretical general Lorenz curve where

$$F^{-1}(s) = \sup\{x : F(x) < s\}$$

be a quantile function (a generalized inverse function) of a distribution function $F(x)$.

Denote by

$$GL_n(t) = \frac{1}{n} \sum_{i=1}^{[nt]} \xi_{i:n}$$

an empirical Lorenz curve.

Goldie (1977) proved a fundamental fact: an empirical Lorenz curve converges to a theoretical one in a uniform metric almost surely.

Let $GL_F^0(t) = GL_F(t) - tGL_F(1)$ be a centered theoretical general Lorenz curve.

We use the next theorem from [*A. Kovalevskii, E. Shatalin* (2016)]:

Theorem 1 *Let* $X_i = \xi_{i:n}$ *be order statistics generated by sample* $(\xi_1, \ldots, \xi_n)$ *with distribution function* F*, sequences* $\{\varepsilon_i\}$ *and* $\{\xi_i\}$ *are independent. If* $0 < \mathbf{Var}\xi_1 < \infty$ *then*
$\widehat{Z_n^0} \Longrightarrow Z_F^0$ *where* Z_F^0 *is a centered Gaussian process with a covariance kernel* $K_F^0(t, u)$*, given by*

$$K_F^0(t, u) = \min\{t, u\} - tu - \frac{GL_F^0(t) GL_F^0(u)}{\mathbf{Var}\xi_1}, \ t, u \in [0, 1].$$

We change $GL_F^0(t)$ by its estimation $GL_n^0(t) = GL_n(t) - tGL_n(1)$. We substitute sample variance s_X^2 for variance $\mathbf{Var}\xi_1$. Let

$$K_n^0(t,u) = \min\{t,u\} - tu - \frac{GL_n^0(t)GL_n^0(u)}{S^2},\ t,u \in [0,1].$$

Then $K_n^0(t,u) \to K_F^0(t,u)$ uniformly on $t,u \in [0,1]$ as $n \to \infty$.

Our statistical test use values of the empirical bridge in d points: let

$$\mathbf{a} = (a_1, \ldots, a_d) = \left(\frac{1}{d+1}, \ldots, \frac{d}{d+1}\right),$$

$$G = \big(K_F^0(a_i, a_j)\big)_{i,j=1}^{d}, \quad G_n = \big(K_n^0(a_i, a_j)\big)_{i,j=1}^{d},$$

$$q = (\widehat{Z_n}(a_1), \ldots, \widehat{Z_n}(a_d))^T.$$

If G^{-1} exists then $q^T G^{-1} q$ is a continuous functional of $\widehat{Z_n}$. Therefore we have

Corollary 1 *Let conditions of Theorem 1 be satisfied. If G^{-1} exists then $q^T G_n^{-1} q$ converge weakly to χ^2-distribution with parameter d.*

P-value for the test is $\alpha^* = 1 - F_{\chi^2_d}(q^T G_n^{-1} q)$.

We choose $d = [n^{1/3}] + 1$.

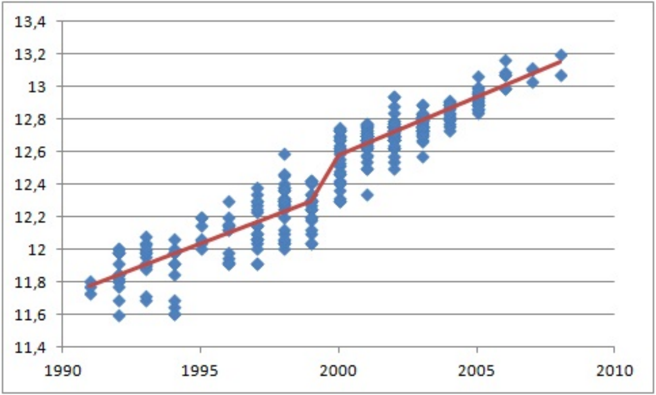
We have $n = 364$, $d = 8$ for all the sample,
$n_1 = 132$, $d_1 = 6$, $n_2 = 216$, $d_2 = 7$ for its 1st and 2nd parts (corresponding years 1991–1999 and 2000–2008).

Caculations give $\alpha^* << 10^{-4}$ for all the sample, $\alpha_1^* = 0.1677$ for its 1st part and $\alpha_2^* = 0.07505$ for its 2nd part. Therefore the test rejects the basic hypothesis in all the time interval at the 10^{-4} level and accepts it in intervals 1991–1999 and 2000–2008 at the 0.07 level.

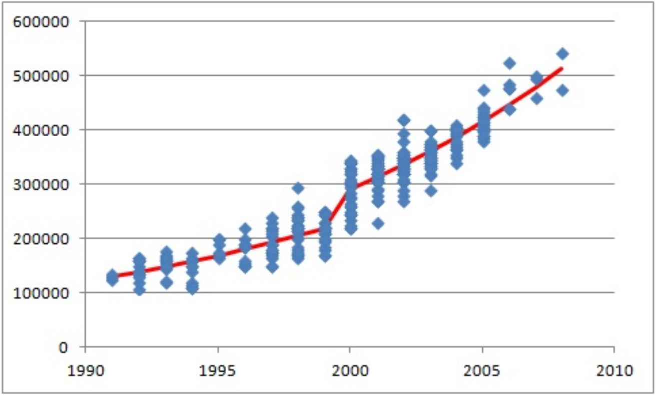
So one can calculate estimated prices of cars in these intervals using models with corresponding coefficients (see Table 1). There is a gap between 1999 and 2000. The price difference is about 6% per year in 1991–1999 and about 7% per year in 2000–2008. Cars of 1999 cheaper than cars of 2000 approximately 1.32 times.

Table: Estimated prices in thousands of roubles

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Price	130	139	148	158	169	180	192	205	219	290	312	335	360	386	415	446	479	514



Logs of prices with different trend lines for 1991–1999 and 2000–2008



Prices with different trend lines for 1991–1999 and 2000–2008 (in roubles, 348 ads)

Example 2. Dependence of weight on height

The initial data the is the information about heights (in cm) and body weights (in kg) of female students of the first course of the Volga state Medical University (2-dimensional sample of 750 items) \http://www.volgmed.ru/ru/

Models

$$W_i = \theta + H_i + \varepsilon_i$$

$$\ln W_i = a + \ln H_i + \varepsilon_i$$

$$\ln W_i = a + 1.5 \ln H_i + \varepsilon_i$$

$$\ln W_i = a + 2 \ln H_i + \varepsilon_i$$

$$\ln W_i = a + 2.5 \ln H_i + \varepsilon_i$$

$$\ln W_i = a + 3 \ln H_i + \varepsilon_i$$

$$\ln W_i = a + b \ln H_i + \varepsilon_i$$

$$W_i = a + bH_i + \varepsilon_i$$

$$W_i = a + bH_i^{1.5} + \varepsilon_i$$

$$W_i = a + bH_i^{2} + \varepsilon_i$$

$$W_i = a + bH_i^{2.5} + \varepsilon_i$$

$$W_i = a + bH_i^{3} + \varepsilon_i$$

$\omega_n^2 = \int\limits_0^1 (Z_n^0(t))^2 \, dt$ is calculated as

$$\omega_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{3} \left(Z_n^0 \left(\frac{i}{n} \right) - Z_n^0 \left(\frac{i-1}{n} \right) \right)^2 + Z_n^0 \left(\frac{i}{n} \right) Z_n^0 \left(\frac{i-1}{n} \right) \right)$$

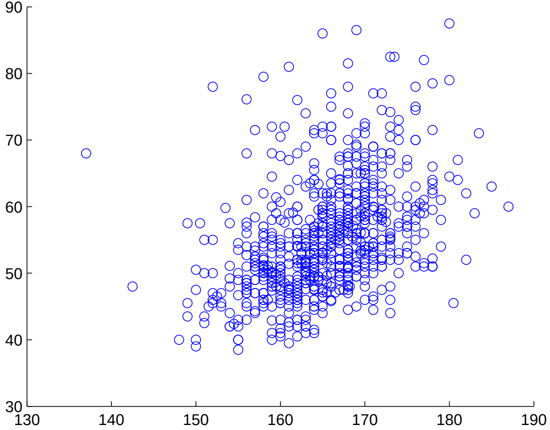
Table 1

Model	$\widehat{\sigma^2}$	ω_n^2	α^*
1	66,82	7,96	$< 10^{-5}$
2	0,0176	3,316	$< 10^{-5}$
3	0,0171	0,4503	0,053
4	0,0174	0,5928	0,024
5	0,0185	3,591	$< 10^{-5}$
6	0,0203	8,643	$< 10^{-5}$
7	0,0171	0,2697	0,0052
8	57,22	0,2273	0,013
9	57,17	0,2172	0,016
10	57,14	0,2101	0,018

Table 1 allows to compare models with each other. In particular, we can conclude that Model 3 is the best. Note the interesting effect. Model 3 is better than Model 7, in which the model parameter b is accurately estimated. It turns out that it is better to guess the model parameter than to estimate it. Of course, this effect is related to the fact that limit distribution of statistics $omega^2$ significantly different for one- and two-parametrical models : estimation of the second parameter should lead to much smaller deviations, but this does not occur in this example.

None of the considered models shows the high p-value, i.e. good compliance with the data being investigated. Therefore, the next phase of the study is analysis of outliers.

Dependence of body mass (in kg) on height (in cm)



The graph shows outliers (abnormally large deviations from any of the proposed regression dependencies), which can lead to the distortion of results. To address this shortcoming, we repeatedly performed the procedure of cleaning the sample using the rule of 3 sigma. We recalculated estimates of the parameters and sample variances of residuals each time after deleting items. The normality of the sample was checked and the procedure was repeated as long as no value was deleted at the next step. As a result, a new 2-dimensional sample was obtained for each model. All calculations were performed again.

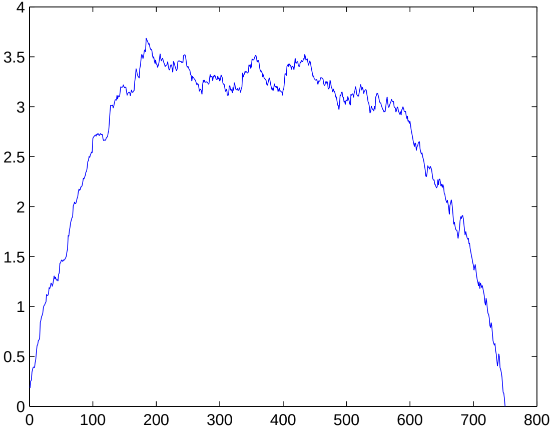
The results of calculations are given in Table 2 (Model 10 is excluded, because the sample have not passed the normality test at one of steps).

Table 2

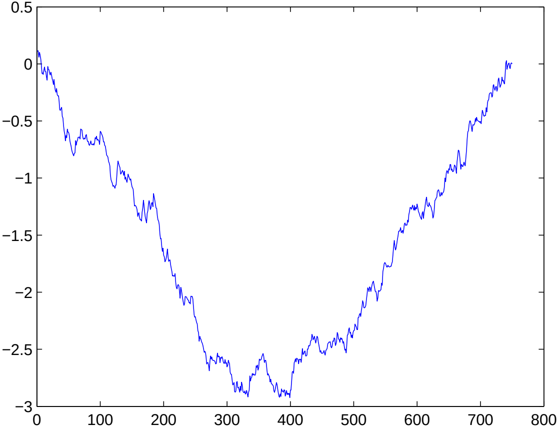
Model	Iterations	Deleted	$\widehat{\sigma^2}$	ω_n^2	α^*
1	4	14	53,07	6,87	$< 10^{-5}$
2	2	9	0,0158	4,35	$< 10^{-5}$
3	2	8	0,0158	0,8133	$0,0068$
4	4	10	0,0149	0,2718	0,165
5	2	9	0,0164	2,84	$< 10^{-5}$
6	2	8	0,0177	7,64	$< 10^{-5}$
7	3	9	0,0151	0,1741	0,0412
8	1	11	47,14	0,1605	0,0563
9	4	20	42,27	0,1674	0,047

Model 4 is the best. So Model 4 should be used to analyze body mass deviations from the norm.
$\widehat{a} \approx -6,2171.$
In this way, our study allows to determine the significance of body mass deviations from the norm on the basis of lognormal law with parameters
$\mu = -6,2171 + 2\ln H,$
$\sigma^2 = 0,0149.$
H is a first-year femail student height in cm.

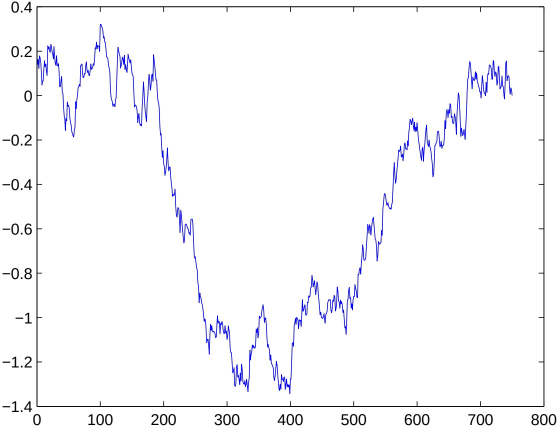
$W_i = \theta + H_i + \varepsilon_i$



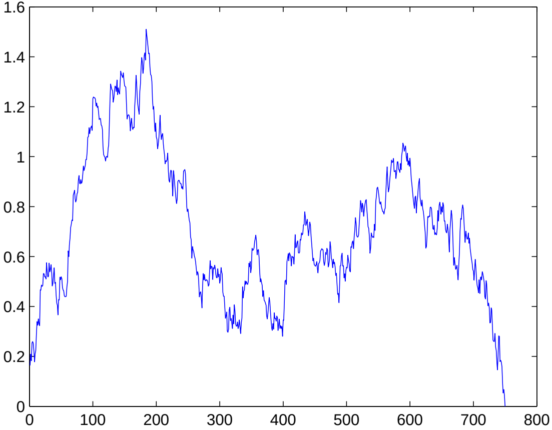
$\ln W_i = a + \ln H_i + \varepsilon_i$



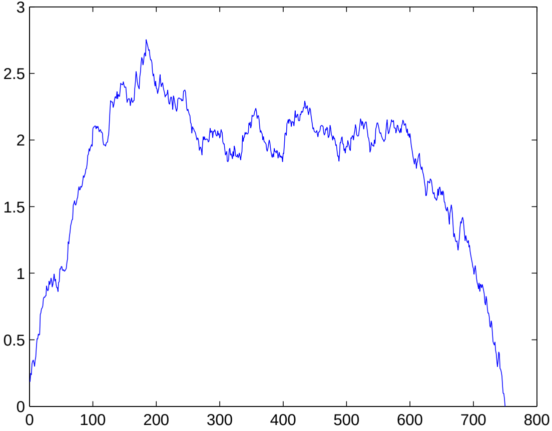
$\ln W_i = a + 1.5 \ln H_i + \varepsilon_i$



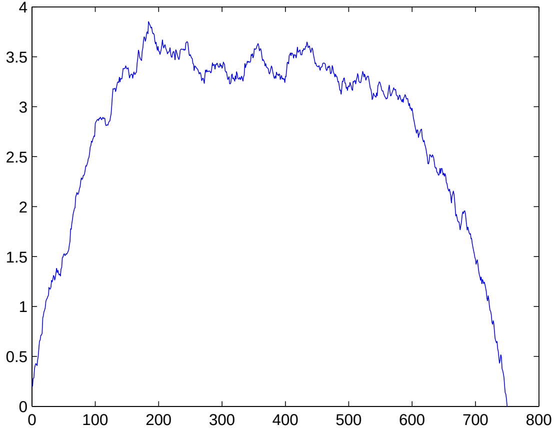
$\ln W_i = a + 2 \ln H_i + \varepsilon_i$



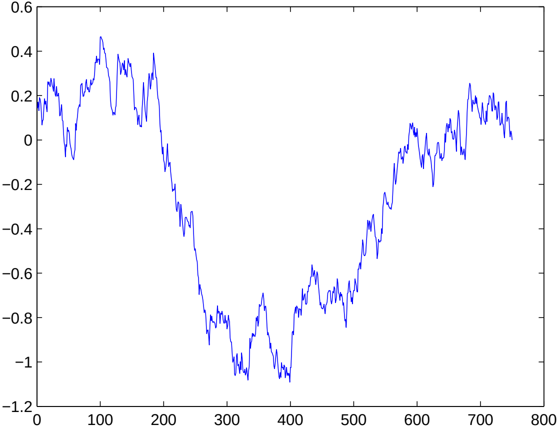
$\ln W_i = a + 2.5 \ln H_i + \varepsilon_i$



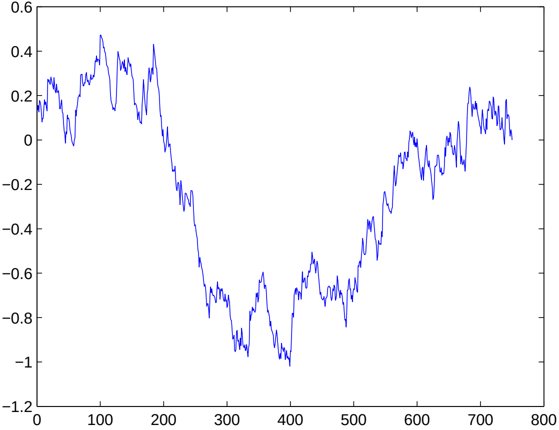
$\ln W_i = a + 3 \ln H_i + \varepsilon_i$



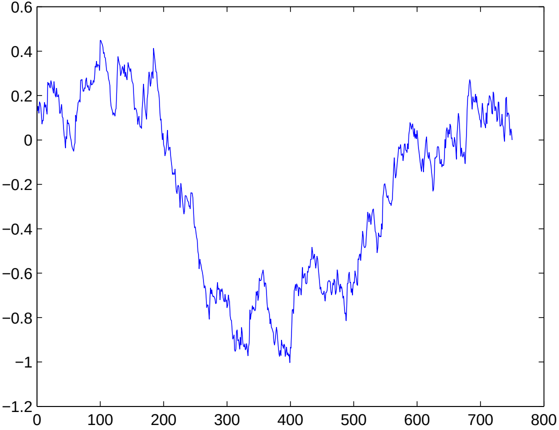
$$\ln W_i = a + b \ln H_i + \varepsilon_i$$



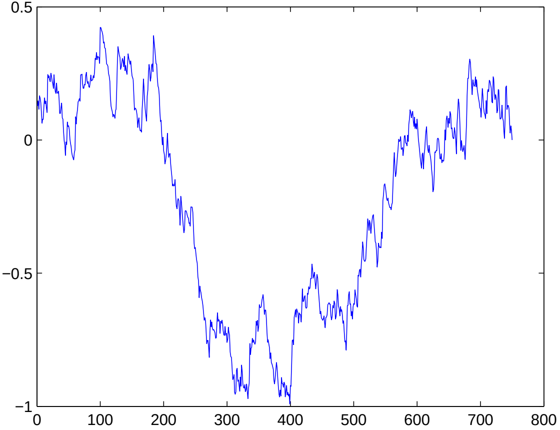
$$W_i = a + bH_i + \varepsilon_i$$



$$W_i = a + bH_i^{1.5} + \varepsilon_i$$



$W_i = a + bH_i^2 + \varepsilon_i$


0.5
0
−0.5
−1
0
100
200
300
400
500
600
700
800