Goodness-of-fit tests for exponentiality based on Yanev-Chakraborty characterization and their efficiencies

Volkova Ksenia

Saint-Petersburg State University, Russia

September, 2015

Introduction

Let $X_1, X_2, \ldots$ be i.i.d. observations with the continuous df F. We are interested in testing the hypothesis

$$H_0 : F \text{ is exponential with the density } \lambda e^{-\lambda x},\ x \geq 0,\ \lambda > 0,$$
$$H_1 : F \text{ is non-exponential df},$$

assuming that the alternative df is also concentrated on $[0, \infty)$.

Introduction

Definition

A df F will be said to belong to class $\mathcal{F}$, if the corresponding density f has derivatives of all orders in the neighbourhood of zero.

Arnold and Villasenor (2013) conjectured, and Yanev and Chakraborty (2013) proved that the following characterized the exponential law within the class $\mathcal{F}$:

Theorem

Let $X_1, \ldots, X_n$ be non-negative i.i.d. rv's with df F from class $\mathcal{F}$. Then the statistics $\max(X_1, X_2, X_3)$ and $\max(X_1, X_2) + \frac{X_3}{3}$ are identically distributed if and only if the df F is exponential.

Introduction

Let $F_n(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{X_i < t\}, t \in R^1$, be the usual empirical df based on the sample $X_1, \ldots, X_n$. According to the characterization we construct for $t \geq 0$ the U-empirical df's by the formulae

$$H_n(t) = \binom{n}{3}^{-1} \sum_{1 \leq i_1 < i_2 < i_3 \leq n} \mathbf{1}\{\max(X_{i_1}, X_{i_2}, X_{i_3}) < t\}, \quad t \geq 0,$$

$$G_n(t) = \frac{1}{3}\binom{n}{3}^{-1} \sum_{1 \leq i_1 < i_2 < i_3 \leq n} \Big[\mathbf{1}\{\max(X_{i_1}, X_{i_2}) + \frac{X_{i_3}}{3} < t\} +$$
$$+ \mathbf{1}\{\max(X_{i_2}, X_{i_3}) + \frac{X_{i_1}}{3} < t\} + \mathbf{1}\{\max(X_{i_3}, X_{i_1}) + \frac{X_{i_2}}{3} < t\}\Big], t \geq 0.$$

Consider two statistics for testing of H_0 against H_1 :

$$I_n = \int_0^\infty (H_n(t) - G_n(t))\, dF_n(t),$$
$$D_n = \sup_{t \geq 0} \mid H_n(t) - G_n(t) \mid .$$

Introduction

In this talk we will

- describe limiting distributions of statistics I_n and D_n under H_0.
- find logarithmic asymptotics of the large deviations under H_0.
- calculate the local Bahadur efficiency of statistics under some parametric alternatives.
- discuss conditions of the local asymptotic optimality of our tests and describe "most favorable" alternatives for them.

Limiting distribution of the statistic I_n

Note that we may take $\lambda = 1$.

It is well-known that non-degenerate U-statistics are asymptotically normal (Hoeffding, 1948). Let show that I_n belongs to this class.

The statistic I_n is asymptotically equivalent to the U-statistic of degree 4 with the centered kernel

$$\Psi(X_1, X_2, X_3, X_4) = \frac{1}{4} \sum_{\pi(i_1,\dots,i_4)} \mathbf{1}\{\max(X_{i_1}, X_{i_2}, X_{i_3}) < X_{i_4}\} -$$

$$- \frac{1}{24} \sum_{\pi(i_1,\dots,i_4)} \mathbf{1}\{\max(X_{i_1}, X_{i_2}) + \frac{X_{i_3}}{3} < X_{i_4}\}.$$

Limiting distribution of the statistic I_n

Let us calculate the projection of the kernel $\Psi(X_1, X_2, X_3, X_4)$:

$$\psi(s) = E(\Psi(X_1, X_2, X_3, X_4) \mid X_4 = s) = \frac{3}{8}e^{-s} - \frac{9}{16}e^{-2s} + \frac{1}{4}e^{-3s} - \frac{1}{12}e^{-s/3},$$

with the variance Δ^2 under H_0

$$\Delta^2 = E\psi^2(X_1) = \frac{23}{174720} \approx 0.0001316.$$

Hence the kernel Ψ is non-degenerate. By Hoeffding's theorem as $n \to \infty$

$$\sqrt{n}I_n \xrightarrow{d} \mathcal{N}(0, \frac{23}{10920}).$$

Limiting distribution of the statistic D_n

The rv $H_n(t) - G_n(t)$ for fixed $t \geq 0$ is asymptotically equivalent to a family of U-statistics with the kernels depending on $t \geq 0$:

$$\Xi(X,Y,Z;t) = \mathbf{1}\{\max(X,Y,Z) < t\} - \frac{1}{3}\mathbf{1}\{\max(X,Y) + \frac{Z}{3} < t\} -$$
$$-\frac{1}{3}\mathbf{1}\{\max(Y,Z) + \frac{X}{3} < t\} - \frac{1}{3}\mathbf{1}\{\max(X,Z) + \frac{Y}{3} < t\}.$$

The projection of this kernel for fixed t is equal to

$$\xi(s;t) = \mathbf{1}\{s < t\}\left[\frac{1}{3} - e^{-t} + e^{-2t} - e^{2s-3t} + \frac{2}{3}e^{3s-3t}\right] - \frac{1}{3}\mathbf{1}\{s < 3t\}(1 - e^{t-s/3})^2.$$

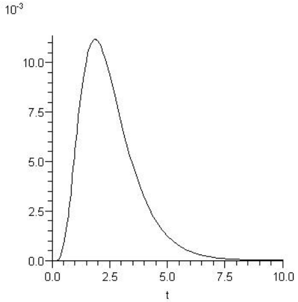
Limiting distribution of the statistic D_n

After some calculations we get that the variance of this projection under H_0 is

$$\delta^2(t) = \frac{1}{5}e^{-t} - \frac{2}{3}e^{-2t} + (\frac{26}{9} - \frac{4}{3}t)e^{-3t} - \frac{43}{21}e^{-4t} + \\ + \frac{1}{10}e^{-5t} - \frac{4}{45}e^{-6t} - \frac{2}{7}e^{-5t/3} + \frac{1}{2}e^{-7t/3} + e^{-8t/3} - \\ - \frac{8}{5}e^{-10t/3} - 2e^{-11t/3} + 2e^{-13t/3}.$$

Hence our family of kernels $\Xi(X, Y, Z; t)$ is non-degenerate and

$$\delta^2 = \sup_{t \geq 0} \delta^2(t) \approx 0.01119.$$



Limiting distribution of the statistic D_n

The limiting distribution of the statistic D_n is unknown. Using methods of Silverman (1983), one can show that the U-empirical process

$$\eta_n(t) = \sqrt{n}\left(H_n(t) - G_n(t)\right), \ t \geq 0,$$

weakly converges as $n \to \infty$ to the certain centered Gaussian process $\eta(t)$ with the calculable covariance. Then the sequence of statistics $\sqrt{n}D_n$ converges in distribution to the rv $\sup_{t\geq 0}|\eta(t)|$ with very complicated distribution (currently unknown). But the critical values for statistics D_n can be found via simulating their sample distribution.

n	0.1	0.05	0.01
10	0.41	0.46	0.57
20	0.25	0.29	0.36
30	0.20	0.22	0.28
40	0.17	0.19	0.23
50	0.15	0.17	0.20
100	0.10	0.11	0.13

Large deviations of the statistic I_n

The kernel Ψ is centered, bounded and non-degenerate. Therefore from the theorem of Nikitin and Ponikarov (1999) describing the large deviations of non-degenerate U-statistics we have

Theorem

For $a > 0$

$$\lim_{n\to\infty} n^{-1} \ln P(I_n > a) \sim \frac{5460}{23} a^2, \text{ as } a \to 0.$$

Some notions from the Bahadur theory

In the Bahadur theory the measure of the efficiency of the sequence of statistics $\{T_n\}$ is the exact slope $c_T(\theta)$.
According to the Bahadur theory exact slopes may be found by using the following Bahadur theorem:

Theorem

Suppose that the following two conditions hold:

$$\text{a)} \qquad T_n \xrightarrow{P_\theta} b(\theta), \qquad \theta > 0,$$

where $-\infty < b(\theta) < \infty$*, and* $\xrightarrow{P_\theta}$ *denotes the convergence in probability under* $G(\cdot\,;\theta)$*.*

$$\text{b)} \qquad \lim_{n\to\infty} n^{-1} \ln \, P_{H_0}(T_n \geq t\,) \;=\; -r(t)$$

for any t *in an open interval* I*, where* r *is continuous and* $\{b(\theta),\, \theta > 0\} \subset I$*. Then*

$$c_T(\theta) \;=\; 2\, r(b(\theta)).$$

Some notions from Bahadur theory

It is well-known according to the Bahadur-Raghavachari inequality, (Bahadur, 1971), that always

$$c_T(\theta) \leq 2K(\theta),\ \theta > 0,$$

where $K(\theta)$ is the Kullback-Leibler "distance" between the null-hypothesis and the alternative indexed by the real parameter θ. Therefore we may define the local Bahadur efficiency as

$$e^B(T) = \lim_{\theta \to 0} \frac{c_T(\theta)}{2K(\theta)}.$$

Local efficiency of the statistic I_n

Let us calculate the local Bahadur exact slope and the local efficiency of the sequence of statistics I_n for the alternative df $G(x,\theta)$ and the density $g(x,\theta)$ assuming their regularity and the possibility of differentiating under the integral sign. Denote also $h(x) = g'_\theta(x,0)$.

According to the Law of Large Numbers for U-statistics (Korolyuk and Borovskikh, 1994), the limit in probability of the sequence I_n under any such alternative is equal as $\theta \to 0$ to

$$b_I(\theta) = P_\theta(\max(X,Y,Z) < W) - P_\theta(\max(X,Y) + \frac{Z}{3} < W).$$

After some computation we get

$$b_I(\theta) \sim 4\theta \int_0^\infty \psi(s)h(s)ds,\ \theta \to 0.$$

Local efficiency of the statistic I_n

We will need in the sequel the expressions of the Kullback-Leibler "distance" between the null-hypothesis and the considered alternatives as $\theta \to 0$. Note that the null-hypothesis is the composite one. Put

$$K(\theta) = \inf_{\lambda>0} \int_0^\infty \ln[g_j(x,\theta)/\lambda \exp(-\lambda x)] g_j(x,\theta)\, dx.$$

Then

$$2K(\theta) \sim \left[\int_0^\infty h^2(x) e^x dx - \left(\int_0^\infty x h(x) dx \right)^2 \right] \theta^2, \text{ as } \theta \to 0.$$

Local efficiency of the statistic I_n

We present the following alternatives:

i) Makeham alternative with the density

$$g_1(x,\theta) = \big(1+\theta(1-e^{-x})\big)\exp\big(-x-\theta(e^{-x}-1+x)\big), \theta > 0;$$

ii) Weibull alternative with the density $g_2(x,\theta) = (1+\theta)x^{\theta}\exp(-x^{1+\theta}), \theta > 0;$

iii) Gamma alternative with the density $g_3(x,\theta) = \frac{x^{\theta}}{\Gamma(\theta+1)}e^{-x}, \theta > 0;$

iv) Exponential mixture with negative weights (EMNW(β)) with the density

$$g_4(x) = (1+\theta)e^{-x} - \theta\beta e^{-\beta x}, \theta \in \Big[0, \frac{1}{\beta-1}\Big], \beta > 1.$$

Alternative	Makeham	Weibull	Gamma	EMNV(4)
Efficiency	0.654	0.649	0.638	0.863

Large deviations of the statistic D_n

The family of kernels $\{\Xi(X, Y, Z; t), t \geq 0\}$ is not only centered but bounded. Hence using the result on large deviations of the supremum of a family of U-statistics (Nikitin, 2010), we obtain

Theorem

For $a > 0$

$$\lim_{n\to\infty} n^{-1} \ln P(D_n > a) \sim 4.966a^2, \text{ as } a \to 0.$$

Local efficiency of the statistic D_n

Let us calculate the local Bahadur slope and local efficiency of the statistic D_n for the alternative df $G(x, \theta)$. By Glivenko-Cantelli theorem for the U-empirical df's (Janssen, 1988) the limit of D_n almost surely under any alternative is equal as $\theta \to 0$ to

$$b_D(\theta) := \sup_{t \geq 0} |P_\theta(\max(X, Y, Z) < t) - P_\theta(\max(X, Y) + \frac{Z}{3} < t)|.$$

Assuming the regularity of the alternative df, we can deduce

$$b_D(\theta) \sim \sup_{t \geq 0} 3\theta| \int_0^\infty \xi(s; t) h(s) ds|, \ \theta \to 0.$$

Alternative	Makeham	Weibull	Gamma	EMNV(4)
Efficiency	0.123	0.079	0.066	0.107

Conditions of local asymptotic optimality of statistic I_n.

Let us derive conditions of the local asymptotic optimality (LAO) in the Bahadur sense. This means to describe the local structure of alternatives when the relation holds:

$$c_T(\theta) \sim 2K(\theta),\ \theta \to 0.$$

Consider functions

$$H(x) = G_\theta^{'}(x,\theta) \mid_{\theta=0}, \quad h(x) = g_\theta^{'}(x,\theta) \mid_{\theta=0} .$$

We will assume that the following regularity conditions are true:

$$h(x) = H'(x),\ x \geq 0, \quad \int_0^\infty h^2(x)e^x dx < \infty, \tag{1}$$

$$\frac{\partial}{\partial\theta}\int_0^\infty xg(x,\theta)dx \mid_{\theta=0} = \int_0^\infty xh(x)dx. \tag{2}$$

Denote by $\mathcal{G}$ the class of densities $g(x,\theta)$ with df's $G(x,\theta)$, satisfying the regularity conditions (1) - (2).

Conditions of local asymptotic optimality of statistic I_n.

First consider the integral statistic I_n with the kernel $\Psi(X_1, X_2, X_3, X_4)$ and its projection $\psi(x)$.
We recall that by the definition we have

$$\Delta^2 = \int_0^\infty \psi^2(x)e^{-x}dx,$$

$$b_I(\theta) \sim 4\theta \int_0^\infty \psi(x)h(x)dx,$$

$$2K(\theta) \sim \left[\int_0^\infty h^2(x)e^x dx - \Big(\int_0^\infty xh(x)dx\Big)^2\right]\theta^2, \; \theta \to 0.$$

Consequently the local BE takes the form

$$e^B(I) = \lim_{\theta\to 0} b_I^2(\theta)/\left(32\Delta^2 K(\theta)\right).$$

Conditions of local asymptotic optimality of statistic I_n.

From Cauchy-Schwarz inequality follows that $e^B(I) = 1$ holds iff

$$h(x) = e^{-x}(C_1\psi(x) + C_2(x-1))$$

for some constants $C_1 > 0$ and C_2. Such distributions constitute the LAO class in the class $\mathcal{G}$.
The simplest example of such alternative density $g(x,\theta)$ for small $\theta > 0$ satisfies the formula

$$g(x,\theta) = e^{-x}\left(1 + \theta\left(\frac{3}{8}e^{-x} - \frac{9}{16}e^{-2x} + \frac{1}{4}e^{-3x} - \frac{1}{12}e^{-x/3}\right)\right), x \geq 0.$$

Conditions of local asymptotic optimality of statistic D_n.

Now consider the Kolmogorov-type statistic D_n with the family of kernels $\Xi(X, Y, Z; t)$ and their projections $\xi(x; t)$. In this case, the following asymptotics is valid

$$\delta^2(t) = \int_0^\infty \xi^2(x) e^{-x} dx,$$

$$b_D(t, \theta) \sim 3\theta \int_0^\infty \xi(x; t) h(x) dx,$$

$$2K(\theta) \sim \Big[\int_0^\infty h^2(x) e^x dx - \Big(\int_0^\infty x h(x) dx \Big)^2 \Big] \theta^2, \; \theta \to 0.$$

Hence, the local asymptotic efficiency takes the form

$$e^B(D) = \lim_{\theta \to 0} \left[b_D^2(\theta) / \sup_{t \geq 0} \left(18 \delta^2(t) \right) \cdot K(\theta) \right].$$

Conditions of local asymptotic optimality of statistic D_n.

It follows that the sequence of statistics D_n is locally optimal iff

$$h(x) = e^{-x}(C_1\xi(x;t_0) + C_2(x-1)) \text{ for } t_0 = \arg\max_{t\geq 0} \delta^2(t)$$

and some constants $C_3 > 0$ and C_4. The distributions having such function $h(x)$ form the domain of LAO in the corresponding class.
The simplest example of such alternative density $g(x,\theta)$ for small $\theta > 0$ is given by the formula

$$g(x,\theta) = e^{-x}\Big(1 + \theta\mathbf{1}\{x < t_0\}\left[\frac{1}{3} - e^{-t_0} + e^{-2t_0} - e^{2x-3t_0} + \frac{2}{3}e^{3x-3t_0}\right] - \\ -\frac{\theta}{3}\mathbf{1}\{x < 3t_0\}(1 - e^{t_0 - x/3})^2\Big), x \geq 0,$$

where

$$t_0 = \arg\max_{t\geq 0} \delta^2(t) \approx 1.854.$$

Thank you for your attention!