Anderson-Darling Type Goodness-of-fit Statistic Based on a Multifold Integrated Empirical Distribution Function

S. Kuriki (Inst. Stat. Math., Tokyo) and H.-K. Hwang (Academia Sinica)

Bernoulli Society Satellite Meeting to ISI2013

Wed 4 Sept 2013, Univ Tokyo

Outline

I. Anderson-Darling statistic and its extension

Goodness-of-fit tests

- $X_1, \ldots, X_n$: i.i.d. sequence from cdf F
- Goodness-of-fit test:

$$H_0 : F = G \quad \text{vs.} \quad H_1 : F \neq G$$

  (G is a given cdf)
- When G is continuous, we can assume $G(x) = x$ (i.e., Unif(0,1)) WLOG.
- Empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$$

- Test statistic is defined as a measure of discrepancy between $F_n(x)$ and $G(x) = x$.

Goodness-of-fit tests (cont)

- We focus on two integral-type test statistics.
- Anderson-Darling (1952) statistic:

$$A_n = n\int_0^1 \frac{1}{x(1-x)}(F_n(x)-x)^2 dx$$

- Watson (1961) statistic
(for testing uniformity on the unit sphere in $\mathbb{R}^2$):

$$U_n = n\int_0^1 (F_n(x)-x)^2 dx - n\left\{\int_0^1 (F_n(x)-x)dx\right\}^2$$

- Limiting null distributions:
Let $\xi_1, \xi_2, \ldots$ be i.i.d. sequence from $N(0,1)$. As $n \to \infty$,

$$A_n \xrightarrow{d} \sum_{k=1}^{\infty} \frac{1}{k(k+1)}\xi_k^2, \qquad U_n \xrightarrow{d} \sum_{k=1}^{\infty} \frac{1}{2\pi^2 k^2}(\xi_{2k-1}^2 + \xi_{2k}^2)$$

Closely looking at Anderson-Darling

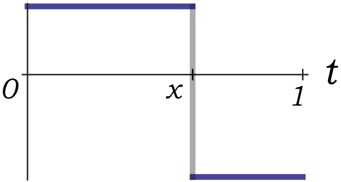
- Anderson-Darling statistic

$$A_n = \int_0^1 \frac{B_n(x)^2}{x(1-x)} dx \quad \text{where} \quad B_n(x) = \sqrt{n}(F_n(x) - x)$$

Here,

$$B_n(x) = \sqrt{n} \int_0^1 h^{[0]}(t; x) dF_n(t), \qquad h^{[0]}(t; x) = \mathbb{1}(t \leq x) - x$$

- $h^{[0]}(t; x)$



An extension to Anderson-Darling

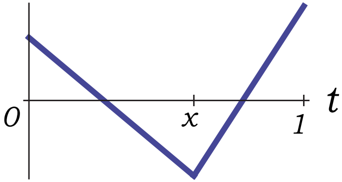
- To propose a new class of test statistics, instead of $h^{[0]}(\cdot; x)$, we prepare different type $h^{[m]}(\cdot; x)$.
- Note first that $h^{[0]}(\cdot; x)$ is piecewise constant s.t.

$$\int_0^1 h^{[0]}(t; x) \cdot 1 \, dt = 0$$

- Define $h^{[1]}(\cdot; x)$ to be continuous and piecewise linear s.t.

$$\int_0^1 h^{[1]}(t; x) \cdot (at + b) \, dt = 0, \quad \forall a, b$$

- $h^{[1]}(\cdot; x)$



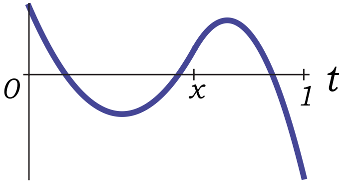
An extension to Anderson-Darling (cont)

- General from of $h^{[m]}(t;x)$:

$$h^{[m]}(t;x) = \frac{1}{m!}(x-t)^m \mathbb{1}(t \leq x) - \sum_{k=0}^{m} \int_0^x \frac{1}{m!}(x-u)^m L_k(u)du \times L_k(t)$$

where $L_k(\cdot)$ is the Legendre polynomial of degree k

- $h^{[2]}(\cdot;x)$



An extension to Anderson-Darling (cont)

- We propose an extension of Anderson-Darling:

$$A_n^{[m]} = \int_0^1 \frac{B_n^{[m]}(x)^2}{\{x(1-x)\}^{m+1}} dx$$

where

$$\begin{aligned} B_n^{[m]}(x) =& \sqrt{n} \int_0^1 h^{[m]}(t;x) dF_n(t) \\ =& \sqrt{n} \Big\{ \int\cdots\int_{x>x_m>\cdots>x_1>0} F_n(x_1)dx_1\cdots dx_m \\ & - \sum_{k=0}^{m} \int\cdots\int_{x>x_m>\cdots>x_1>0} L_k(x_1)dx_1\cdots dx_m \int_0^1 L_k(t)dF_n(t) \Big\} \end{aligned}$$

(m-fold integral of empirical distribution function)

- $A_n^{[m]}$ is well-defined (the integral exists) whenever $X_i \in (0,1)$.
- $A_n^{[0]}$ is the original Anderson-Darling.

II. Limiting null distribution

Main results — Limiting null distribution

- $W(\cdot)$: the Winer process on $[0,1]$
  $B(x) = W(x) - xW(1)$: Brownian bridge
- Let $B_n^{[m]}(x) = \int_0^1 h^{[m]}(t;x)dB_n(x)$.
  We can prove that as $n \to \infty$,

  $$B_n^{[m]}(\cdot) \xrightarrow{d} B^{[m]}(\cdot) \text{ in } L^2, \text{ where } B^{[m]}(x) = \int_0^1 h^{[m]}(t;x)dB(t)$$

  and hence (by continuous mapping)

  $$A_n^{[m]} = \int_0^1 \frac{B_n^{[m]}(x)^2}{\{x(1-x)\}^{m+1}}dx \xrightarrow{d} A^{[m]} := \int_0^1 \frac{B^{[m]}(x)^2}{\{x(1-x)\}^{m+1}}dx$$

- We will examine these limiting distributions $B^{[m]}(\cdot)$ and $A^{[m]}$.

Main results — Limiting null distribution (cont)

Theorem (Karhunen-Loève expansion)

$$\frac{B^{[m]}(x)}{\{x(1-x)\}^{(m+1)/2}} = \sum_{k=m+1}^{\infty} \sqrt{\frac{(k-m-1)!}{(k+m+1)!}} L_k^{(m+1)}(x)\, \xi_k$$

(uniformly in x, with prob. 1), where

$$\xi_k = \int_0^1 L_k(t) dB(t), \quad i.i.d.\ N(0,1)$$

$L_k^{(m+1)}$ *is the associate Legendre function.* □

Corollary (Limiting null distribution of $A_n^{[m]}$)

$$A^{[m]} = \int_0^1 \frac{B^{[m]}(x)^2}{\{x(1-x)\}^{m+1}} dx = \sum_{k=m+1}^{\infty} \frac{(k-m-1)!}{(k+m+1)!} \xi_k^2$$

$\xi_k^2 \sim \chi^2(1)$ *i.i.d.* □

III. Moment generating function

Moment generating function

- The moment generating function (Laplace transform) of

$$A^{[m]} = \sum_{k=1}^{\infty} \frac{1}{\lambda_k} \xi_k^2, \quad \lambda_k = k(k+1)\cdots(k+2m+1)$$

is

$$E\left[e^{sA^{[m]}}\right] = \prod_{k=1}^{\infty} \left(1 - \frac{2s}{\lambda_k}\right)^{-\frac{1}{2}}$$

Theorem

Let $x_j(s)$ $(j = 0, 1, \ldots, 2m+1)$ *be the solution of*

$\lambda_x - 2s = 0$, *i.e.,* $x(x+1)\cdots(x+2m+1) - 2s = 0$

Then

$$E\left[e^{sA^{[m]}}\right] = \prod_{j=0}^{2m+1} \sqrt{\frac{\Gamma(1 - x_j(s))}{j!}}$$

□

Moment generating function ($m = 0$)

- When $m = 0$ (Anderson-Darling), $\lambda_k = k(k+1)$
  The equation
  $$x(x+1) - 2s = 0$$
  has a solution
  $$x_0(s), x_1(s) = -\frac{1}{2} \pm \sqrt{1+8s}$$
- Hence,
  $$E\left[e^{sA^{[0]}}\right] = \sqrt{\Gamma(1-x_0(s))\Gamma(1-x_1(s))}$$
  $$= \sqrt{\frac{2\pi s}{-\cos\frac{\pi}{2}\sqrt{1+8s}}}$$
  (Anderson and Darling, 1952)
- Euler's reflection formula $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$ is used.

Moment generating function ($m = 1$)

- When $m = 1$. $\lambda_k = k(k+1)(k+2)(k+3)$.
- The equation

$$x(x+1)(x+2)(x+3) - 2s = 0$$

has the explicit solution , because by letting $x = y - 3/2$,

$$\begin{aligned}\text{LHS} =& (y-3/2)(y-1/2)(y+1/2)(y+3/2) - 2s \\ =& \{y^2 - (3/2)^2\}\{y^2 - (1/2)^2\} - 2s\end{aligned}$$

is a quadratic equation in $y^2 = (x+3/2)^2$.

- As a result,

$$x_0(s), x_1(s), x_2(s), x_3(s) = \frac{1}{2}\left(\pm\sqrt{5 \pm 4\sqrt{2s+1}} - 3\right),$$

$$E\left[e^{sA^{[1]}}\right] = \frac{\pi s}{\sqrt{3\cos(\frac{\pi}{2}\sqrt{5-4\sqrt{1+2s}})\cos(\frac{\pi}{2}\sqrt{5+4\sqrt{1+2s}})}}$$

Moment generating function ($m = 2$)

- When $m = 2$,

$$E\left[e^{sA^{[2]}}\right] = \frac{(\pi s)^{3/2}}{\sqrt{-4320\cos(\pi\sqrt{\eta_1})\cosh(\pi\sqrt{\eta_2})\cosh(\pi\sqrt{\eta_3})}}$$

where

$$\eta = \sqrt[3]{27s + 80 + 3\sqrt{81s^2 + 480s - 1728}}$$

and

$$\eta_1 = \frac{1}{12\eta}\left(4\eta^2 + 35\eta + 112\right)$$

$$\eta_2 = \frac{1}{12\eta}\left(4e^{-\pi i/3}\eta^2 - 35\eta + 112e^{\pi i/3}\right)$$

$$\eta_3 = \frac{1}{12\eta}\left(4e^{\pi i/3}\eta^2 - 35\eta + 112e^{-\pi i/3}\right)$$

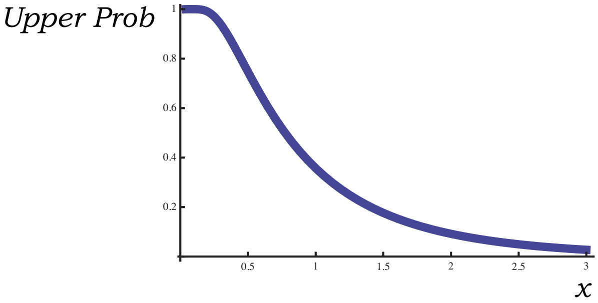
Calculation of upper prob.

- Finite representation is useful in numerical calculation.

$$P\left(A^{[m]} > x\right) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{\pi} \int_{\lambda_{2k-1}/2}^{\lambda_{2k}/2} \frac{e^{-xs}}{s\sqrt{\left|E\left[e^{sA^{[m]}}\right]\right|}} ds$$

(Smirnov-Slepian technique, see Slepian (1958)).

- The case $m = 0$:



IV. Statistical power

Statistical power

- We have the sample counterpart of the KL-expansion:

$$\frac{B_n^{[m]}(x)}{\{x(1-x)\}^{(m+1)/2}} = \sum_{k \geq m+1} \sqrt{\frac{(k-m-1)!}{(k+m+1)!}} L_k^{(m+1)}(x)\, \widehat{\xi}_k$$

where

$$\widehat{\xi}_k = \int_0^1 L_k(x) dB_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_k(X_i), \quad k \geq 1$$

- The (extended) Anderson-Darling statistics are also written in terms of $\widehat{\xi}_k$'s as

$$A_n^{[0]} = \sum_{k \geq 1} \frac{1}{k(k+1)} \widehat{\xi}_k^2 = \frac{1}{2}\widehat{\xi}_1^2 + \frac{1}{6}\widehat{\xi}_2^2 + \frac{1}{12}\widehat{\xi}_3^2 + \cdots$$

$$A_n^{[1]} = \sum_{k \geq 2} \frac{1}{(k-1)k(k+1)(k+2)} \widehat{\xi}_k^2 = \frac{1}{24}\widehat{\xi}_2^2 + \frac{1}{120}\widehat{\xi}_3^2 + \cdots$$

Statistical power (cont)

- First two components:

$$\widehat{\xi}_1 = \sqrt{12n} m_1, \quad \widehat{\xi}_2 = 6\sqrt{5n} \times (m_2 - 1/12),$$

  where $m_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - 1/2)^k$ (the sample kth moment around 1/2)
- $A_n^{[0]} = \widehat{\xi}_1^2/2 + \cdots$ has much power for mean-shift alternative, and
- $A_n^{[1]} = \widehat{\xi}_2^2/24 + \cdots$ has much power for dispersion-change alternative

V. Extension of Watson's statistic

- Similar extensions are possible for Watson's statistic. Let

$$U_n^{[m]} = \int_0^1 C_n^{[m]}(x)^2 dx, \quad C_n^{[m]} = \int_0^1 h^{[m]}(t;x) dF_n(t),$$

where

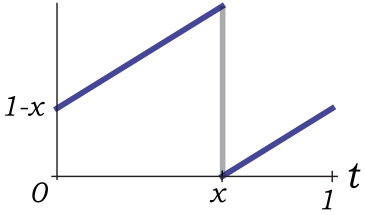
$$h^{[m]}(t;x) = \frac{(t-x)^m}{m!} \mathbb{1}(t - x \leq 0) + \frac{1}{(m+1)!} b_{m+1}(t-x)$$

- $b_m(y)$ is the Bernoulli polynomial, which satisfies

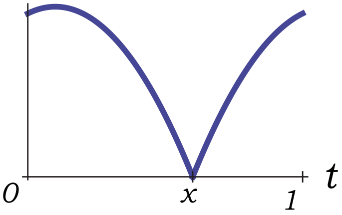
$$b_m(y+1) = b_m(y) + my^{m-1}.$$

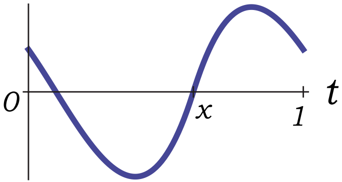
- $U_n^{[0]}$ is the original Watson statistic.
- $U_n^{[1]}$ is proposed by Henze and Nikitin (2002).

- $h^{[0]}(\cdot; x)$



- $h^{[1]}(\cdot; x)$ and $h^{[2]}(\cdot; x)$





Limiting null distribution

- Let
$$C_n^{[m]}(\cdot) \xrightarrow{d} C^{[m]}(\cdot) \quad \text{and} \quad U_n^{[m]} \xrightarrow{d} U^{[m]}.$$
- KL-expansion of $C^{[m]}(x)$:
$$C^{[m]}(x) = \sum_{k=1}^{\infty} \frac{1}{(2k\pi)^{m+1}} \left\{ l_{2k-1}^{[m]}(x)\xi_{2k-1} + l_{2k-1}^{[m]}(x)\xi_{2k} \right\},$$
where
$$l_{2k-1}^{[m]}(x) = \sin\left(2k\pi x - \frac{m+1}{2}\pi\right), \quad l_{2k}^{[m]}(x) = \cos\left(2k\pi x - \frac{m+1}{2}\pi\right),$$
$$\xi_{2k-1} = \int_0^1 \sin(2k\pi x) dB(x), \quad \xi_{2k} = \int_0^1 \cos(2k\pi x) dB(x).$$
- Consequently,
$$U^{[m]} = \sum_{k=1}^{\infty} \frac{1}{(2k\pi)^{2(m+1)}} \{\xi_{2k-1}^2 + \xi_{2k}^2\}.$$

Summary

- We proposed a class of extended Anderson-Darling statistics $A_n^{[m]}$ $(m \geq 0)$ based on m-fold integrated empirical distribution function.
- The limiting null distribution $A^{[m]}$ is explicitly derived as weighted infinite sums of chi-square random variables.
- We provided moment generating function of $A^{[m]}$ without using infinite product.
- The same-type extension for Watson's statistic U_n is possible.

- Acknowledgment: The authors thank Y. Nishiyama of ISM.

References

- Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, **23** (2), 193–212.
- Henze, N. and Nikitin, Ya. Yu. (2002). Watson-type goodness-of-fit tests based on the integrated empirical process. *Math. Methods. Stat.*, **11** (2), 183–202.
- Slepian, D. (1957). Fluctuations of random noise power. *Bell. Syst. Tech. J.*, **37**, 163–184.
- Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika*, **48** (1,2), 109–114.