Two-sample Kolmogorov-Smirnov Type Tests and Local Levels

H. Finner and V. Gontscharuk

Institute for Biometrics and Epidemiology,
German Diabetes Center (DDZ),
Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf

Adaptive Designs and Multiple Testing Procedures Workshop 2016

April 28-29, 2016,

University of Padua – Padua, Italy

Outline

Goodness-of-Fit (GOF) Problems

One-sample GOF problem:

Let $X_1, \ldots, X_m$ be iid with an unknown continuous cdf F and let F_0 be a continuous cdf. Testing problem:

$$H_0 : F(t) = F_0(t) \text{ for all } t \in \mathbb{R} .$$

Two-sample GOF problem:

Let $X_1, \ldots, X_m$ be iid and $Y_1, \ldots, Y_n$ be iid with unknown continuous cdf's F and G, respectively. Testing problem:

$$H_0 : F(t) = G(t) \text{ for all } t \in \mathbb{R} .$$

Acceptance regions of KS type tests

One-sample GOF problem:

$$\begin{aligned} A_m &= \{c(t) \leqslant \hat{F}_m(t) \leqslant d(t)\ \forall t\ \} \\ &= \{c_i \leqslant X_{i:m} \leqslant d_i\ \forall i = 1, \ldots, m\} \end{aligned}$$

Two-sample GOF problem:

$$\begin{aligned} A_{m,n} &= \{c(t) \leqslant \hat{F}_m(t) - \hat{G}_n(t) \leqslant d(t)\ \forall t\} \\ &= \{c_s \leqslant V_{m,s} \leqslant d_s\ \forall s = 1, \ldots, m+n-1\} \end{aligned}$$

with $V_{m,s} = m\hat{F}_m(t_s)$, where $t_1 < \cdots < t_{m+n}$ are the jump points of the ecdf $\hat{H}_{m+n}$ of the combined sample.

Assumption: No ties in the combined sample.

Connection to one and two-sample binomial tests

Consider local tests for $H_{0t} : F(t) = F_0(t)$ and $H_{0t} : F(t) = G(t)$.

One-sample KS type tests: Local tests based on

$$m\hat{F}_m(t) \sim B(m, F(t))$$

are one sample binomial tests.

Two-sample KS type tests: Local tests based on

$$m\hat{F}_m(t) \sim B(m, F(t)),\ m\hat{G}_n(t) \sim B(n, G(t))$$

are two-sample binomial tests.

Hence, KS type tests can be viewed as union-intersection tests based on local binomial tests.

Local Levels

Local levels are defined as local rejection probabilities under H_0:

One-sample KS type tests:

$$\begin{aligned} \alpha_i &= \mathbb{P}(U_{i:m} < c_i) + \mathbb{P}(U_{i:m} > d_i\} \\ &= \alpha_i^{low} + \alpha_i^{up} \text{ (say)}, \ i = 1, \ldots, m, \end{aligned}$$

with $U_{1:m} \leqslant \cdots \leqslant U_{m:m}$ *uniform order statistics*.

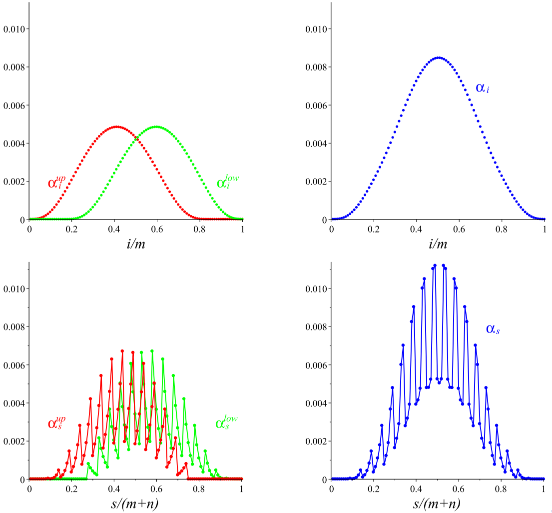
Two-sample KS type tests:

$$\begin{aligned} \alpha_s &= \mathbb{P}(V_{m,s} < c_s) + \mathbb{P}(V_{m,s} > d_s) \\ &= \alpha_s^{low} + \alpha_s^{up} \text{ (say)}, \ s = 1, \ldots, m+n-1, \end{aligned}$$

with $V_{m,s} \sim Hyp(s, m, n)$ under H_0.

Local Levels

Example. Local levels of one sample ($m = 100$) and two-sample ($m = 20$, $n = 80$) KS tests ($\alpha = 0.05$):



Local levels of one sample KS type tests

In three recent papers we studied local levels of various one sample KS type tests.

These investigations were motivated by a series of papers on higher criticism (HC) statistics (which are weighted KS-type statistics) in connection with detectability and sparsity.

Unfortunately, the asymptotic results for HC don't fit the finite world, even not for extremely large sample sizes.

Among others, we designed and studied a new higher criticism test with **equal local levels** with favourable asymptotic properties.

Asymptotics of equal local levels $\alpha_1 = \cdots = \alpha_{m+n-1}$:

$$\lim_{m\to\infty} \frac{\alpha_i}{\alpha_m^*} = 1 \quad \text{for} \quad \alpha_m^* = \frac{-\log(1-\alpha)}{\log(m)\log(\log(m))}.$$

Proper two-sample KS type tests

Noting that $\boxed{V_{m,s} \leqslant V_{m,s+1} \leqslant V_{m,s} + 1}$, meaningful acceptance regions

$$A_{m,n} = \{(x_1, \ldots, x_{m+n-1}) : c_s \leqslant x_s \leqslant d_s \; \forall s = 1, \ldots, m+n-1\}$$

of two-sample KS type tests should satisfy

$$\boxed{c_s \leqslant c_{s+1} \leqslant c_s + 1} \text{ and } \boxed{d_s \leqslant d_{s+1} \leqslant d_s + 1} \; \forall s = 1, \ldots, m+n-2. \tag{1}$$

For unconditional 2×2 table tests (like Fisher's exact test) the latter property is sometimes referred to as Barnard convexity.

Moreover, $V_{m,s} \sim Hyp(s, m, n)$ yields the natural restrictions

$$\max(0, s-n) \leqslant c_s \leqslant d_s \leqslant \min(s, m). \tag{2}$$

We call two-sample KS type tests with (1) and (2) **proper** GOF tests.

Proper two-sample KS type tests

Let $V_m = (V_{m,1}, \ldots, V_{m,m+n-1})$.

Result: Any acceptance region $A_{m,n}$ can be replaced by a proper acceptance region $A^*_{m,n} \subseteq A_{m,n}$ with

$$\mathbb{P}(V_m \in A^*_{m,n}) = \mathbb{P}(V_m \in A_{m,n}).$$

In other words, the critical value collections $(c_s : s = 1, \ldots, m+n-1)$ and $(d_s : s = 1, \ldots, m+n-1)$ should be possible paths of V_m.

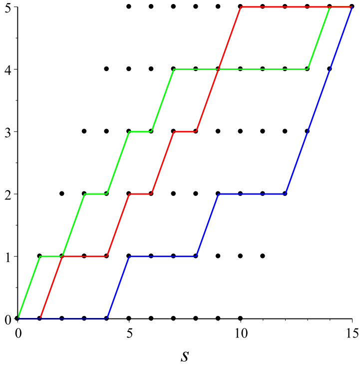
Conclusion: *Barnard convexity* is an inherent property of two-sample KS type tests.

For 2×2 table tests, a proof of Barnard convexity can be hard, cf. Finner & Strassburger (2002, JSPI) for UMPU tests.

DDZ

Paths of $V_{m,s}$

Some realizations of $(V_{m,s} : \ s = 0, \ldots, m+n)$ for $m = 5$ and $n = 10$



Effective level and α exhaustion

The effective level can easily be calculated by making use of the recursion

$$\mathbb{P}(V_{m,s+1} = j) = \frac{m-j+1}{m+n-s}\mathbb{P}(V_{m,s} = j-1) + \frac{n-s+j}{m+n-s}\mathbb{P}(V_{m,s} = j).$$

Due to the discreteness of $V_{m,s}$, the effective level is typically smaller than the prespecified nominal level α.

The difference can be large, even for larger values of m, n.

Example: Classical two-sample KS test with e.g. $m = n$.

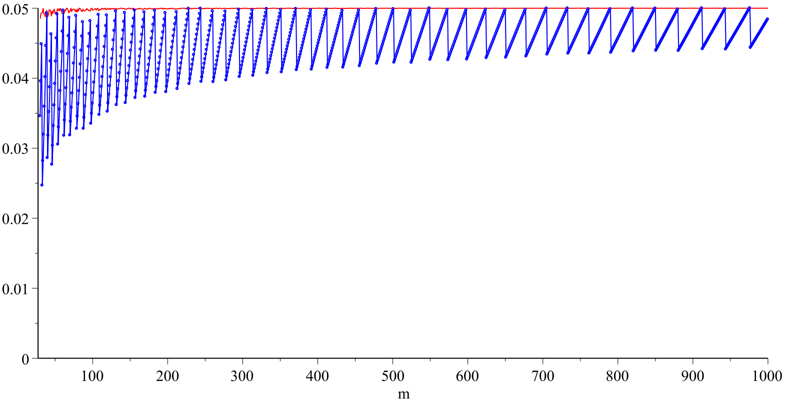
For e.g. $m = n + 1$, the effective level is typically much closer to the nominal level.

An additional maximization step as suggested by Barnard for 2×2 tables can resolve this problem (not treated here).

Effective level and α exhaustion: Example

Two-sample two-sided KS tests ($\alpha = 0.05$):

Effective levels for $m = n$ and $n = m + 1$ for $m = 30, \ldots, 1000$.



Normal approximation

For suitable values of m, n and s, the standardized statistics

$$V^*_{m,s} \equiv \frac{V_{m,s} - \mathbb{E}[V_{m,s}]}{\sqrt{\mathbb{V}ar[V_{m,s}]}} \tag{3}$$

can be approximated by the standard normal distribution. This typically leads to asymptotic expressions for the corresponding local levels, too.

Example (two-sample KS):

$$\alpha_s \to 2\Phi\left(\frac{-b_\alpha}{\sqrt{\eta(1-\eta)}}\right), \ \frac{s}{m+n} \to \eta,$$

with b_α asymptotic critical value of KS test.

Weighted KS tests

For $\nu \in [0,1]$, two-sample weighted KS statistics are defined by

$$\boxed{KS^{\nu}_{m,n} = \sup_{t\in\mathbb{R}} \sqrt{\frac{mn}{m+n}} \frac{|\hat{G}_n(t) - \hat{F}_m(t)|}{(\hat{H}_{m+n}(t)(1-\hat{H}_{m+n}(t)))^{\nu}}}$$

Reject H_0 if $KS^{\nu}_{m,n} > b_{m,n}$, where $b_{m,n}$ denotes a suitable critical value.

These tests can be rewritten in terms of $V_{m,s}$ as proper KS type tests.

For $\nu = 0$ we have the two-sample KS test.

The case $\nu = 1/2$ can be seen as a two-sample version of the higher criticism statistic. Thereby,

$$KS^{1/2}_{m,n} = \sqrt{\frac{m+n}{m+n-1}} \sup_{s=1,\dots,m+n-1} \frac{|V_{m,s} - \mathbb{E}[V_{m,s}]|}{\sqrt{\mathbb{V}ar[V_{m,s}]}}.$$

Weighted KS tests will be investigated in the next talk.

Minimum p-value (minP) tests

In the one-sample case, the minP test results in equal local levels and leads to an attractive alternative to the HC tests.

Due to the discreteness, equal local levels are not possible in the two-sample case.

However, we can try to designs a test with approximately equal local levels by applying the minimum p-value concept.

Thereby, we have to define local p-values p_s based on $V_{m,s}$, e.g., p-values related to two-sample binomial tests like Fisher's exact tests or derandomized UMPU-tests.

Minimum p-value (minP) tests

In two-sided testing, the definition of p-values depends on the weights for the tails.

The probably most popular approach is the *double the tails version*, which leads to

$$\boxed{p_s = 2\min\{1/2, F_{Hyp}(V_{m,s}|s,m,n), 1 - F_{Hyp}(V_{m,s}-1|s,m,n)\}.}$$

Then H_0 is rejected if $\boxed{\min_s p_s \leqslant \alpha^{loc}_{m,n}}$ for a suitable $\boxed{\alpha^{loc}_{m,n}}$.

All local levels α_s of this (proper) GOF test are bounded by $\alpha^{loc}_{m,n}$.

We can also apply derandomized UMPU tests for each s with an appropriate upper bound for all local levels. For $m = n$, this version coincides with the *double the tails version*, for $m \neq n$ the tests typically differ.

Animation: minP / UMPU

Design of new KS type tests

Idea: Choose some local level boundary function depending on some tuning parameter b. Choose b such that α is maximally exhausted.

For example, combine the local levels of different tests or design completely new tests.

Ex1: Maximum of local levels of KS test and minP test, both at level b.

Ex2: Sum of local levels of KS test and minP test, both at level b.

Ex3: Try something, e.g.,

$$\alpha_b(\eta) = 2\Phi\left(-\sqrt{-\log(b\eta(1-\eta))}\right), \quad \eta = s/(m+n).$$

Animation: Ex1 – Ex3

The case $m = n$: A slightly modified KS test

Idea: Define a local level boundary function by asymptotic KS local levels:

$$\alpha_b(\eta) = 2\Phi\left(-\frac{b}{\sqrt{\eta(1-\eta)}}\right), \quad \eta = s/(m+n).$$

Choose local levels α_s as large as possible with

$$\alpha_s \leqslant \alpha_b\left(\frac{s}{m+n}\right).$$

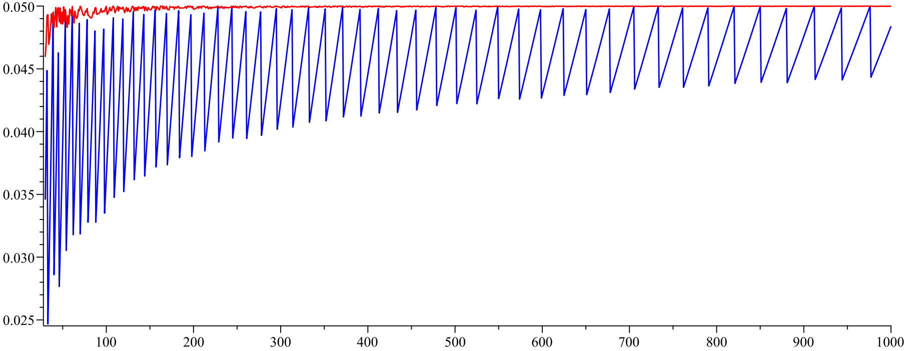
Choose b such that α is maximally exhausted.

The case $m = n$: A slightly modified KS test

Effective levels of KS and modified KS

Two-sample two-sided KS and modified KS tests ($\alpha = 0.05$):

Effective levels for $m = n = 30, \ldots, 1000$:



The modified KS test improves the original KS test in all cases !

Summary

- Barnard convexity (here: proper) is an inherent property of two-sample KS type tests
- Local levels offer a new point of view on KS type tests
- Close connection between KS type and binomial tests
- UMPU binomial tests lead to minP KS type tests with approximately equal local levels
- Local levels can be used to design new tests

References

- **Barnard G. A. (1945)**. A new test for 2×2 tables. *Nature*, **156**, 783–784.
- **Barnard G. A. (1947)**. Significance tests for 2×2 tables. *Biometrika*, **34**, 123–138.
- **Berk, R. and Jones, D. (1979)**. Goodness-of-fit test statistics that dominate the Kolmogorov Statistics. *Z. Wahrscheinlichkeit.*, **47**, 47–59.
- **Gontscharuk, V., Landwehr, S. and Finner, H. (2015)**. Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli*, **22**, 1331–1363.
- **Gontscharuk, V., Landwehr, S. and Finner, H. (2015)**. The intermediates take it all: asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biometrical J.*, **57**, 159–180.
- **Gontscharuk, V. and Finner, H. (2016)**. Asymptotics of goodness-of-fit tests based on minimum p-value statistics. *Commun. Stat. A-Theor.*, available at `http://www.tandfonline.com/doi/full/10.1080/03610926.2015.1041985`.
- **Kolmogorov, A. (1933)**. Sulla determinazione empirica di una legge di distribuzione. *G. Inst. Ital. Attuari*, **4**, 83–91. Translated in English in: Breakthroughs in Statistics II (1992), pp. 106–113, Eds. Kotz, S. and Johnson N. L.
- **Smirnov, N. (1939)**. Sur les ècarts de la courbe de distribution empirique. *Rec. Math. [Mat. Sbornik] N.S.*, **6(48)**, 3–26.