Adv. App. Stat. Presentation

On a paradoxical property of the Kolmogorov–Smirnov two-sample test

Stefan Hasselgren
Niels Bohr Institute

Bias of Kolmogorov g.o.f. test

Draw a sample $X_1, \ldots, X_n$ with unknown d.f. F. Based on these, we want to test the hypothesis

$$H_0: \; F = F_0,$$

where F_0 is a fixed d.f.

A test is unbiased if the probability of rejecting the null hypothesis

A test is unbiased if the probability of rejecting the null hypothesis

(a) is greater than (or equal to) the significance level when the alternative is true

A test is unbiased if the probability of rejecting the null hypothesis

(a) is greater than (or equal to) the significance level when the alternative is true

and

(b) is less than (or equal to) the significance level when the null hypothesis is true.

A test is unbiased if the probability of rejecting the null hypothesis

(a) is greater than (or equal to) the significance level when the alternative is true

and

(b) is less than (or equal to) the significance level when the null hypothesis is true.

The test is biased for the alternative hypothesis, if (a) is not true while (b) is still true.

Now consider a test, which has the following properties:

Now consider a test, which has the following properties:

1. Reject the null hypothesis if $d(G_n, F_0) > \delta_\alpha$.

Now consider a test, which has the following properties:

1. Reject the null hypothesis if $d(G_n, F_0) > \delta_\alpha$.

- G_n is the sample d.f. of the sample $X_1, \ldots, X_n$,

Now consider a test, which has the following properties:

1. Reject the null hypothesis if $d(G_n, F_0) > \delta_\alpha$.

- G_n is the sample d.f. of the sample $X_1, \ldots, X_n$,
- $d(G_n, F_0)$ is the 'distance' in the space of d.f.'s.

Now consider a test, which has the following properties:

1. Reject the null hypothesis if $d(G_n, F_0) > \delta_\alpha$.

- G_n is the sample d.f. of the sample $X_1, \ldots, X_n$,
- $d(G_n, F_0)$ is the 'distance' in the space of d.f.'s.
- δ_α is a 'distance' associated with the significance level α.

Now consider a test, which has the following properties:

1. Reject the null hypothesis if $d(G_n, F_0) > \delta_\alpha$.
    - G_n is the sample d.f. of the sample $X_1, \ldots, X_n$,
    - $d(G_n, F_0)$ is the 'distance' in the space of d.f.'s.
    - δ_α is a 'distance' associated with the significance level α.
2. The test is distribution free. Essentially, no assumptions are made about the underlying distribution of the sample.

Now consider a test, which has the following properties:

1. Reject the null hypothesis if $d(G_n, F_0) > \delta_\alpha$.

   - G_n is the sample d.f. of the sample $X_1, \ldots, X_n$,
   - $d(G_n, F_0)$ is the 'distance' in the space of d.f.'s.
   - δ_α is a 'distance' associated with the significance level α.

2. The test is distribution free. Essentially, no assumptions are made about the underlying distribution of the sample.

3. Such a test is called a "distance-based test."

Now, let's get technical. For some distribution function F;

Now, let's get technical. For some distribution function F;

- take all the d.f.'s with distance d from F (so all F_i where $d(F_i, F) = d$),

Now, let's get technical. For some distribution function F;

- take all the d.f.'s with distance d from F (so all F_i where $d(F_i, F) = d$),

- put them in a metric space (somehow),

Now, let's get technical. For some distribution function F;

- take all the d.f.'s with distance d from F (so all F_i where $d(F_i, F) = d$),

- put them in a metric space (somehow),

- then make a 'ball' in this space with radius $\delta > 0$ and centre at F.

Now, let's get technical. For some distribution function F;

- take all the d.f.'s with distance d from F (so all F_i where $d(F_i, F) = d$),
- put them in a metric space (somehow),
- then make a 'ball' in this space with radius $\delta > 0$ and centre at F.
- Label this ball $\mathcal{B}(F, \delta)$.

Now we take a d.f. F_0, and then suppose that (for some $\alpha > 0$) there exists a d.f. F_a, such that

- the ball $\mathcal{B}(F_a, \delta_\alpha)$ is strictly inside the ball $\mathcal{B}(F_0, \delta_\alpha)$, Now we take a d.f. F_0, and then suppose that (for some $\alpha > 0$) there exists a d.f. F_a, such that

- the ball $\mathcal{B}(F_a, \delta_\alpha)$ is strictly inside the ball $\mathcal{B}(F_0, \delta_\alpha)$,

  and

- there is a non-zero probability of the sample d.f. being in $\mathcal{B}(F_0, \delta_\alpha)$, but not in $\mathcal{B}(F_a, \delta_\alpha)$, given that F_a is the true d.f.

Then the distance-based test is biased for the alternative hypothesis F_a.

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}$

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}$

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}\{G_n$

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_a, \delta_\alpha)\}$

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_a, \delta_\alpha)\} \geq 1 - \alpha$.

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_a, \delta_\alpha)\} \geq 1 - \alpha$.

  But since the ball $\mathcal{B}(F_a, \delta_\alpha)$ is strictly inside the ball $\mathcal{B}(F_0, \delta_\alpha)$, then

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_a, \delta_\alpha)\} \geq 1 - \alpha$.

  But since the ball $\mathcal{B}(F_a, \delta_\alpha)$ is strictly inside the ball $\mathcal{B}(F_0, \delta_\alpha)$, then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_0, \delta_\alpha)\} > 1 - \alpha$,

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_a, \delta_\alpha)\} \geq 1 - \alpha$.

  But since the ball $\mathcal{B}(F_a, \delta_\alpha)$ is strictly inside the ball $\mathcal{B}(F_0, \delta_\alpha)$, then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_0, \delta_\alpha)\} > 1 - \alpha$,

  which is the same as

- $\mathbb{P}_{F_a}\{d(G_n, F_0) > \delta_\alpha\} < \alpha$.

Proof

The proof is fairly simple. Start by taking a sample $X_1, \ldots, X_n$ from F_a. This sample has sample d.f. G_n. Then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_a, \delta_\alpha)\} \geq 1 - \alpha$.

  But since the ball $\mathcal{B}(F_a, \delta_\alpha)$ is strictly inside the ball $\mathcal{B}(F_0, \delta_\alpha)$, then

- $\mathbb{P}_{F_a}\{G_n \in \mathcal{B}(F_0, \delta_\alpha)\} > 1 - \alpha$,

  which is the same as

- $\mathbb{P}_{F_a}\{d(G_n, F_0) > \delta_\alpha\} < \alpha$.

But this is strictly against the demand (a) for an unbiased test! And so the distance-based test is biased for the alternative F_a.

Bias of the KS two-sample test for different sample sizes

Take two samples $X_1, \ldots, X_m$ from F and $Y_1, \ldots, Y_n$ from G. The null hypothesis is then $H_0: \; F = G$.

Bias of the KS two-sample test for different sample sizes

Take two samples $X_1, \ldots, X_m$ from F and $Y_1, \ldots, Y_n$ from G. The null hypothesis is then $H_0: \ F = G$.

We can then easily choose or assume F and G to satisfy the conditions above. The article proves, that for equal sample size $n = m$, the KS two-sample test is unbiased.

Bias of the KS two-sample test for different sample sizes

Take two samples $X_1, \ldots, X_m$ from F and $Y_1, \ldots, Y_n$ from G. The null hypothesis is then $H_0: \; F = G$.

We can then easily choose or assume F and G to satisfy the conditions above. The article proves, that for equal sample size $n = m$, the KS two-sample test is unbiased.

However, for sufficiently large nm, the KS test is biased for the alternative $F = F_a \neq G$, since in the limit for $m \to \infty$ we obtain the Kolmogorov g.o.f. test.

Thank you!

Example

$$F_0(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

Example

$$F_0(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

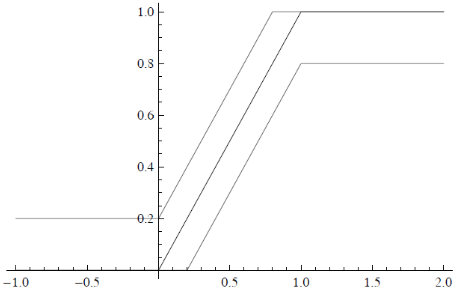

FIG 1. *The ball* $\mathcal{B}(F_0, \delta_\alpha)$.

Example

$$F_a(x) = \begin{cases} 0, & x < \delta_\alpha/2, \\ 2x - \delta_\alpha, & \delta_\alpha/2 \le x < \delta_\alpha, \\ x, & \delta_\alpha \le x < 1 - \delta_\alpha, \\ 2x - (1 - \delta_\alpha), & 1 - \delta_\alpha \le x < 1 - \delta_\alpha/2, \\ 1, & x \ge 1 - \delta_\alpha/2. \end{cases}$$

Example

$$F_a(x) = \begin{cases} 0, & x < \delta_\alpha/2, \\ 2x - \delta_\alpha, & \delta_\alpha/2 \le x < \delta_\alpha, \\ x, & \delta_\alpha \le x < 1 - \delta_\alpha, \\ 2x - (1 - \delta_\alpha), & 1 - \delta_\alpha \le x < 1 - \delta_\alpha/2, \\ 1, & x \ge 1 - \delta_\alpha/2. \end{cases}$$

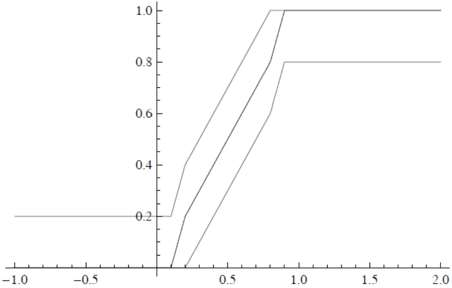

FIG 2. *The ball* $\mathcal{B}(F_a, \delta_\alpha)$.