# Tail-sensitive goodness-of-fit

## The Calibrated Kolmogorov-Smirnov test

**Amit Moscovich Eiger, Boaz Nadler** Weizmann Institute of Science **Clifford Spiegelman** Texas A&M University

## Background

Let $x_1, \dots, x_n$ be i.i.d samples from a distribution $F$.
**Goodness-of-fit problem:** decide between the null hypothesis $H_0 : F = F_0$ and an alternative hypothesis $H_1 : F \neq F_0$.

Many tests for **continuous** goodness-of-fit follow these steps:

1. Transform the samples: $u_i = F_0(x_i)$.
2. Sort $u_1, \dots, u_n$ to obtain $u_{(1)} \leq \dots \leq u_{(n)}$.
3. Compute some test statistic $T(u_{(1)}, \dots, u_{(n)})$.
4. $T >$ threshold $\implies$ reject the null hypothesis.

**Example.** Kolmogorov-Smirnov test [1933]:

$$K_n^+ = \sqrt{n} \max_i \left( \tfrac{i}{n} - u_{(i)} \right) \quad K_n^- = \sqrt{n} \max_i \left( u_{(i)} - \tfrac{i-1}{n} \right)$$

$$K_n = \max(K_n^-, K_n^+)$$

**Limitation:** Lack of sensitivity at the tails of the distribution.

## Motivation

**When is tail sensitivity important?**

1. **Detecting rare and weak contaminations** using multiple sensors. Suppose that in the event of a contamination the output of a small set of sensors is slightly shifted upwards. In that case, contaminations manifest themselves as a small change in the upper tail of the null distribution.
2. **Analyzing heterogeneous experiments.** We test a drug on $N$ patients, obtaining $p$-values $p_1, \dots, p_N$. Suppose this drug affects only $n \ll N$ patients with specific unknown genes, and has no effect on the rest. In that case, most of the $p$-values are drawn i.i.d $U[0,1]$ but a small subset is drawn from some stochastically smaller distribution.
3. **Modeling extreme events** such as 100-year floods. In such cases, recorded data is used to fit the parameters of a *generalized extreme value*, power-law, or similar distribution. The result is used to estimate the probability of extreme events, even those larger than any recorded sample. Goodness-of-fit statistics can be used to give a numerical score for the match between the resulting distribution and the data at hand. Clearly, in such cases we care mostly about the upper tail of the distribution.

Following the 2011 Tōhoku earthquake, an enormous 14 meter high tsunami wave destroyed the emergency power generators of the Fukushima Daiichi nuclear power plant resulting in the meltdown of 3 nuclear reactors. The plant was designed to withstand waves of up to 5.7 meters only.

**Figure 1:** Aftermath of the 2011 Tōhoku earthquake and tsunami. Photo by Douglas Sprott licensed under Creative Commons (CC BY-NC 2.0)

## Analysis

**Ancient lemma.** Let $U_1, ..., U_n \sim U[0,1]$ be i.i.d random variables and let $U_{(1)} \leq ... \leq U_{(n)}$ denote their order statistics. Then

$$U_{(i)} \sim \text{Beta}(i, n-i+1).$$

In particular, $\mathbb{E}[U_{(i)}] = \frac{i}{n+1}$, $\text{Var}(U_{(i)}) = \frac{i(n-i+1)}{(n+1)^2(n+2)}$.

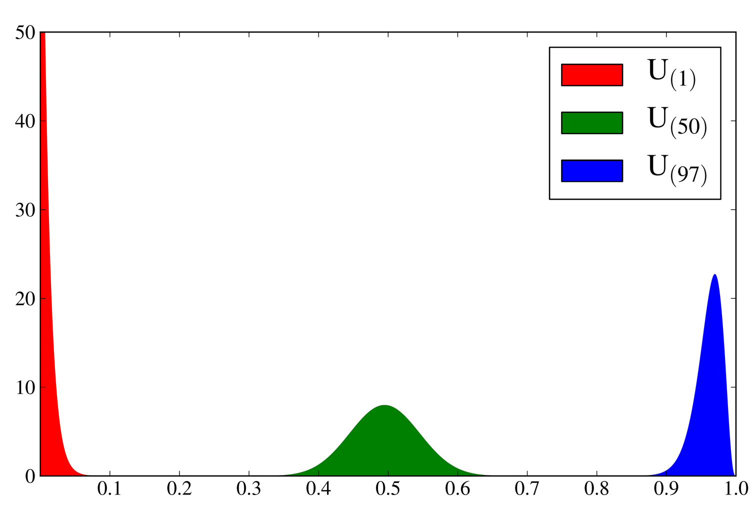


**Figure 2:** Distributions of some order statistics ($n = 100$)

Kolmogorov-Smirnov:

$$K_n = \sqrt{n} \max_i \left| \mathbb{E}[U_{(i)}] - u_{(i)} \right| + O(1/\sqrt{n}).$$

**Limitation:** $\text{Var}(U_{(i)})$ depends on $i$, largest for $i = n/2$.
$\implies$ Low tail sensitivity.

Higher-Criticism [2004]:

$$\text{HC} = \max_i \frac{\mathbb{E}[U_{(i)}] - u_{(i)}}{stdev[U_{(i)}]} (1 + O(1/n)).$$

**Limitation:** standardized **shape** of $U_{(i)}$ depends on $i$. Hence, HC does not equally calibrate deviations at different indices.

**Theorem.** [Keilson&Sumita]

1. For any fixed $c \in (0,1)$, taking $i = cn$ and letting $n \to \infty$, the distribution $\text{Beta}(i, n-i+1)$ converges to a Gaussian.
2. Fixing $i$, as $n \to \infty$, $\text{Beta}(i, n-i+1)$ does **not** converge to a Gaussian but instead to an extreme value distribution.

## New test statistic

**Idea:** Compute the $p$-value of each $u_{(i)}$. Rather than considering the largest deviation, look for the most **statistically significant** deviation

$$\text{CKS}^+ = \min_i \Pr\left[\text{Beta}(i, n-i+1) < u_{(i)}\right]$$
$$\text{CKS}^- = \min_i \Pr\left[\text{Beta}(i, n-i+1) > u_{(i)}\right]$$
$$\text{CKS} = \min(\text{CKS}^-, \text{CKS}^+)$$

## Asymptotics

**Theorem.** If $X_1, \dots, X_n \overset{i.i.d.}{\sim} F_0$ then

$$\Pr\left[\frac{1}{\log n (\log\log n)^{0.5+\epsilon}} < \text{CKS} < \frac{1}{\log n \sqrt{\log\log n}}\right] \xrightarrow{n\to\infty} 1.$$

If $X_1, \dots, X_n \overset{i.i.d.}{\sim} F \neq F_0$ then

$$\Pr\left[CKS < \frac{1+\epsilon}{4\|F - F_0\|_\infty^2} \cdot \frac{1}{n}\right] \xrightarrow{n\to\infty} 1.$$

**Corollary:** CKS is consistent.

## Computing p-values

Following classical results [Durbin 1973], we can express the null distribution of $\text{CKS}^+$ using repeated integration.

**Theorem.** Let $L_i^n(c)$ denote the inverse regularized incomplete Beta function, satisfying $\int_0^{L_i^n(c)} f_{i,n-i+1}(x)dx = c$. Then

$$\Pr(\text{CKS}^+ \geq c | \mathcal{H}_0) = n! \int_{L_n^n(c)}^1 dU_{(n)} \int_{L_{n-1}^n(c)}^{U_{(n)}} dU_{(n-1)} \cdots \int_{L_1^n(c)}^{U_{(2)}} dU_{(1)}.$$

Using this formula, a new $O(n^2)$ numerical procedure for computing $p$-values of the one-sided $\text{CKS}^+$ statistic was derived.

For computing $p$-values of the two-sided CKS one may either:

1. Perform exact computation via Noé's recursion - $O(n^3)$.
2. Approximate the $p$-value, via our one-sided $O(n^2)$ algorithm, based on the following theorem.

**Theorem.** Let $q_c = \Pr[\text{CKS}^+ \leq c \mid \mathcal{H}_0]$. Then

$$2q_c - q_c^2 \leq \Pr[\text{CKS} \leq c \mid \mathcal{H}_0] \leq 2q_c.$$

Furthermore, $\Pr[\text{CKS} \leq c \mid \mathcal{H}_0] \xrightarrow{n\to\infty} 2q_c - q_c^2$.

## Handling rounded samples

Suppose the original observations are rounded to the nearest integer.
**Question:** Can we still apply continuous goodness-of-fit?

Let $\mathbf{x} = (x_1, \dots, x_n)$ be an i.i.d vector of samples and let $\tilde{\mathbf{x}}$ denote the rounded samples. Define

$$x_i^{\downarrow} = \tilde{x}_i - \tfrac{1}{2} \quad \text{and} \quad x_i^{\uparrow} = \tilde{x}_i + \tfrac{1}{2}.$$

Clearly, $\mathbf{x}^{\downarrow} \leq \mathbf{x} \leq \mathbf{x}^{\uparrow}$. Due to the monotonicity of $\text{CKS}^+$

$$\text{CKS}^+(\mathbf{x}^{\downarrow}) \leq \text{CKS}^+(\mathbf{x}) \leq \text{CKS}^+(\mathbf{x}^{\uparrow}),$$

and similarly

$$\text{CKS}^-(\mathbf{x}^{\uparrow}) \leq \text{CKS}^-(\mathbf{x}) \leq \text{CKS}^-(\mathbf{x}^{\downarrow}).$$

Thus,

$$\min\{\text{CKS}^-(\mathbf{x}^{\uparrow}), \text{CKS}^+(\mathbf{x}^{\downarrow})\} \leq \text{CKS}(\mathbf{x}) \leq \min\{\text{CKS}^-(\mathbf{x}^{\downarrow}), \text{CKS}^+(\mathbf{x}^{\uparrow})\}.$$

**Conclusion:** the $p$-value of a set of samples can be bounded given only rounded observations.

## Finite sample results

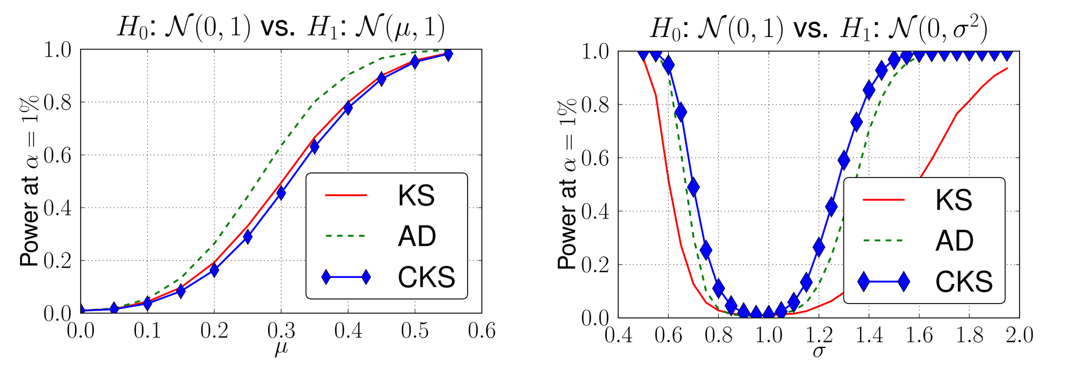


**Figure 3:** Power comparisons for detecting lack-of-fit to a standard Gaussian distribution (at significance level $\alpha = 1\%$) with $n = 100$ samples. (left panel) change in the mean of the distribution; (right panel) change in the variance.

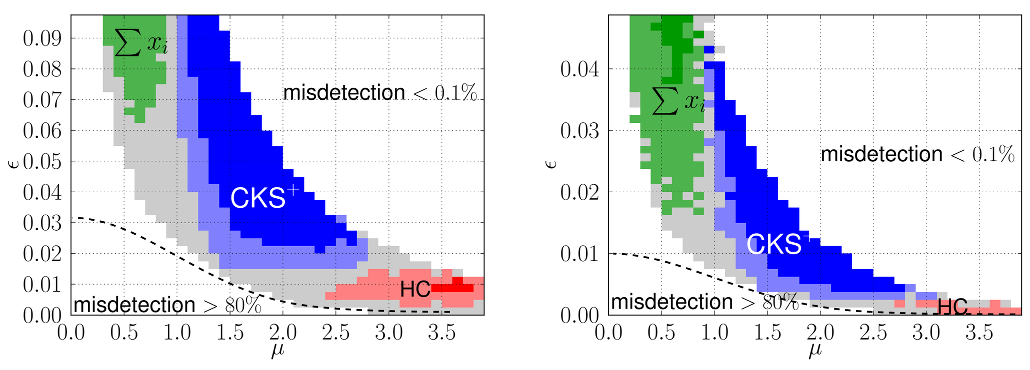


**Figure 4:** Comparison of tests for detecting rare-weak normal mixtures $(1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu,1)$ vs. $\mathcal{N}(0,1)$. Colored blobs represent regions where the misdetection rate of the second-best test divided by that of the best test was larger than 1.1. The dark centers signify regions where this ratio was larger than 1.5. The grey band delineates the zone where misdetection is in the range $0.1\% - 80\%$. The dotted line is the asymptotic detection boundary.

## Main contributions

- CKS, a consistent two-sided goodness-of-fit test with good empirical performance and asymptotically optimal detection of certain types of tail perturbations.
- New $O(n^2)$ algorithm for computing $p$-values of the $\text{CKS}^+$ and $\text{CKS}^-$ statistics. This algorithm applies to other one-sided goodness-of-fit statistics as well.
- Method for estimating the $p$-value of CKS in $O(n^2)$ operations.
- New technique for handling rounded samples. Extends to other goodness-of-fit tests and more complex transformations

### Supporting material

Preprint and code available at
http://www.wisdom.weizmann.ac.il/ amitmo/

### Acknowledgements

This work was supported by a Texas A&M-Weizmann research grant from Paul and Tina Gardner, and by a grant from the Citi Foundation.

### References


[1] Amit Moscovich Eiger, Boaz Nadler, and Clifford Spiegelman. The calibrated kolmogorov-smirnov test. *arXiv preprint arXiv:1311.3190*, 2013.

[2] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[3] J. Keilson and U. Sumita. A decomposition of the beta distribution, related order and asymptotic behavior. *Annals of the Institute of Statistical Mathematics*, 35(1):243–253, 1983.

[4] J. Durbin. *Distribution theory for tests based on the sample distribution function*, volume 9. Society for Industrial Mathematics, 1973.

[5] M. Noe. The calculation of distributions of two-sided kolmogorov-smirnov type statistics. *The Annals of Mathematical Statistics*, pages 58–64, 1972.