How to compare distributions?
or rather
Has my sample been drawn from this distribution?
or even
Kolmogorov-Smirnov and the others...

Stéphane Paltani
ISDC/Observatoire de Genève
The Problem... (I)

Sample of observed values (random variable): \( \{X_i\}, \ i = 1, \ldots, N \)

Expected distribution: \( f(x), \ X_{\text{min}} \leq x \leq X_{\text{max}} \)

→ Is my \( \{X_i\}, \ i = 1, \ldots, N \) sample a probable outcome from a draw of \( N \) values from a distribution \( f(X), \ X_{\text{min}} \leq X \leq X_{\text{max}} \)?
An X-ray detector collects photons at specific times: \( \{X_i\} \equiv \{t_i\}, i=1,\ldots,N \)

Is the source constant?

Expected distribution:

\[
f(t) = \frac{1}{(t_{\text{max}} - t_{\text{min}})}, \quad t_{\text{min}} \leq t \leq t_{\text{max}}
\]
The Problem... (II)

Sample of observed values (random variable): \( \{X_i\}, i=1,...,N \)

Sample of observed values (random variable): \( \{Y_j\}, j=1,...,M \)

→ Are the \( \{X_i\}, i=1,...,N \) distributed the same way as the \( \{Y_j\}, j=1,...,M \)?
We measure the amplitude of variability of different types of active galactic nuclei.

Seyfert 1 galaxies: \( \{ \sigma_i^S \}, \quad i=1,...,N \)

QSOs: \( \{ \sigma_j^Q \}, \quad j=1,...,M \)

BL Lacs: \( \{ \sigma_k^B \}, \quad k=1,...,P \)

Do these different types of objects have different variability properties?
For **continuous** distributions, we can easily compare visually how well two distributions agree by building the cumulatives of these distributions. The cumulative of a distribution is simply its integral:

\[ F(x) = \int_{X_{\text{min}}}^{x} f(y) \, dy \]

The cumulative of a sample can be calculated in a similar way:

\[ C_{\{X_i\}}(x) = \sum_{i} H(x - X_i) \]

(Heaviside function)

Different approaches can then be followed to obtain a quantitative assessment.
The simplest one: We determine the maximum separation between the two cumulatives:

\[ D = \max_{X_{\min} \leq x \leq X_{\max}} \left| C_{X_i} - F(x - X_{\min}) \right| \]
It is possible to estimate the expected distribution of $D$ in the case of a uniform distribution:

$$P(\lambda > D) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \mu^2}, \quad \mu = \left( \sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}} \right) D$$

$P(\lambda > D)$ is a one-sided probability distribution, so the smaller $D$, the larger the probability.
D is preserved under any transformation $x \rightarrow y = \psi(x)$, where $\psi(x)$ is an arbitrary strictly monotonic function.

$\rightarrow P(\lambda > D)$ does not depend on the shape of the underlying distributions!
Correct usage of Kolmogorov-Smirnov

- Decide for a threshold \( \alpha \), for instance 5%
- If \( P(\lambda > D) > \alpha \), then one cannot exclude that the sample has been drawn from the given distribution
- One can never be sure that the sample has been actually drawn from the given distribution

DO NOT SAY:

The probability that the sample has been drawn from this distribution is X%

SAY:

If \( P(\lambda > D) > \alpha \):
We cannot exclude that the sample has been drawn from this distribution

If \( P(\lambda > D) < \alpha \):
The probability of such a high KS value to be obtained is only X%
No correlation in the sample is allowed: Sample must be in the Poisson regime

There is a general problem among non-frequentists (the Bayesians) with the notion of null hypothesis: Such frequentist tests implicitly specify classes of alternative hypotheses and exclude others...

Furthermore, one may reject the hypothesis if the model parameters are slightly wrong
Caveats (II)

In cases where the model has parameters, if one estimates them using the same data, one cannot use KS statistics anymore!

→ Use Monte Carlo simulations
→ For Gaussian distribution, use Shapiro-Wilk normality test
Caveats (III)

Sensitivity of the Kolmogorov-Smirnov test is not constant between $X_{\text{min}}$ and $X_{\text{max}}$ since the variance of $C_{\{X_i\}}$ is proportional to $F(x-X_{\text{min}})(1-F(x-X_{\text{min}}))$, which reaches a maximum for $F(x-X_{\text{min}}) = 0.5$
Non-uniformity Correction

Anderson-Darling's statistics:

\[ D^* = \max_{X_{min} \leq x \leq X_{max}} \frac{|C_{X_i} - F(x - X_{min})|}{\sqrt{F(x - X_{min})(1 - F(x - X_{min}))}} \]

Or:

\[ D^{**} = \int_{X_{min} \leq x \leq X_{max}} \frac{|C_{X_i} - F(x - X_{min})|}{\sqrt{F(x - X_{min})(1 - F(x - X_{min}))}} dF \]

One can think of other ones...

However, none of them provides a simple (or even workable) expression for \( P(\lambda > D^{*}(\star)) \)...
The simplest modification to KS is the best one!

\[ V = D^+ + D^- = \max_{X_{\min} \leq x \leq X_{\max}} \left( C\{X_i\} - F(x - X_{\min}) \right) + \max_{X_{\min} \leq x \leq X_{\max}} \left( F(x - X_{\min} - C\{X_i\}) \right) \]

And:

\[ P(\lambda > V) = 2 \sum_{j=1}^{\infty} \left( 4j^2 \mu^2 - 1 \right) e^{-2j^2 \mu^2}, \quad \mu = \left( \sqrt{N} + 0.155 + \frac{0.24}{\sqrt{N}} \right) V \]
Kuiper on a Circle

If the distribution is defined over a periodic domain, Kuiper's statistics does not depend on the choice of the origin, contrarily to KS.
Standard test is Rayleigh test, which is nothing else than the Fourier power spectrum for signals expressed in the form:

\[ S(t) = \sum_i \delta(t - t_i) \]
Interrupted Observations

But X-ray observations are often interrupted because of various instrumental problems, so the expected distribution is **not** uniform!

Kuiper test can take into account very naturally the non-uniformity of the phase exposure map.
Examples with simulated data

Continuous case | Real GTIs | Real GTIs with signal

![Graph showing examples with simulated data.](image)
Examples with real data

EX Hya

UW Pic
**Ad nauseam: Kuiper for the fanatical... (I)**

\[
\text{Prob}(V \geq z/\sqrt{N}) = \sum_{m=1}^{\infty} 2(4m^2z^2 - 1)e^{-2m^2z^2} - \frac{8z}{3\sqrt{N}} \sum_{m=1}^{\infty} m^2(4m^2z^2 - 3)e^{-2m^2z^2} + O\left(\frac{1}{N}\right). \quad (3)
\]

\[
\text{Prob}(V \leq z) = N! \left(z - \frac{1}{N}\right)^{N-1}, \text{ if } \frac{1}{N} \leq z \leq \frac{2}{N}, \quad (4)
\]

\[
\text{Prob}(V \leq z) = \frac{(N-1)! \left(\beta^{N-1}(1-\alpha) - \alpha^{N-1}(1-\beta)\right)}{N^{N-2}(\beta - \alpha)} \quad (5)
\]

if \(\frac{2}{N} \leq z \leq \frac{3}{N}\)

\[
\alpha, \beta : \quad t^2 - (Nz - 1)t + \frac{1}{2}(Nz - 2)^2 = 0
\]

\[
\text{Prob}(V \geq z) = \sum_{t=0}^{M} \binom{N}{t} \left(1 - \frac{t}{N}\right)^{N-t-1} T_t \quad (6)
\]

with:

\[
T_t = y^{t-3} \left(y^3 N - y^2 t \frac{3-2/N}{N} - \frac{t(t-1)(t-2)}{N^2}\right) , \quad y = z + \frac{t}{N},
\]

if \(z \geq 1/2\), if \(N\) is even, and if \(z \geq (N-1)/(2N)\), if \(N\) is odd.
Ad nauseam: Kuiper for the fanatical... (II)

Tail probabilities

Expected

Best

Asymptotic equation, instead of analytical expression
References

- Any text book in statistics...
- Press W. et al., Numerical Recipes in C,F777, ..., Sect. 14
- Stephens M.A., 1965, Biometrika 70, 11