Two-sample goodness-of-fit tests to location-scale families using OWA-based skewness coefficients

Marina Iturrate–Bobes, Ignacio Montes and Raúl Pérez–Fernández

University of Oviedo

Universidad de Oviedo

21st International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)
June 15th–19th 2026

Index

Index

Location-scale families

Let $\mathcal{L}(\Omega, \mathcal{A}, P)$ denote the set of random variables $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Location-scale family

The **location-scale family** of $X \in \mathcal{L}(\Omega, \mathcal{A}, P)$ is

$$\mathcal{F}_X = \{Y \in \mathcal{L}(\Omega, \mathcal{A}, P) \mid \exists c \in \mathbb{R}, d \in \mathbb{R}^+ \text{ such that } Y \overset{\mathcal{D}}{\equiv} c + dX\}.$$

It is a set of random variables closed under location and scale transformations.

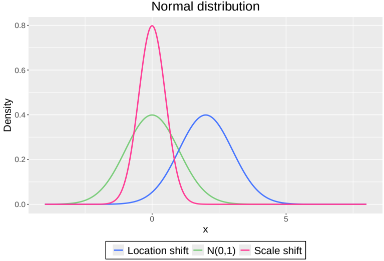
- $X, Y \in \mathcal{L}(\Omega, \mathcal{A}, P)$ belong to the same location-scale family if $c \in \mathbb{R}$ and $d \in \mathbb{R}^+$ exist such that

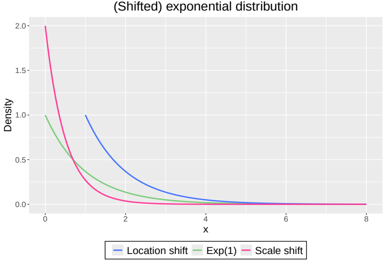
$$F_Y(y) = F_X\left(\frac{y-c}{d}\right), \quad \forall y \in \mathbb{R}.$$

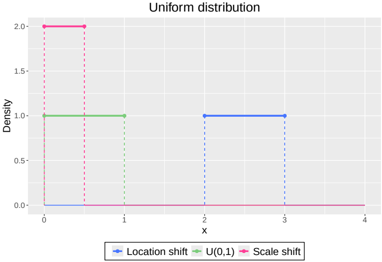
- Moreover, if X and Y are absolutely continuous:

$$f_Y(y) = \frac{1}{d} f_X\left(\frac{y-c}{d}\right), \quad \forall y \in \mathbb{R}.$$

Location-scale families







Symmetry in a random variable

- A random variable X is called **symmetric about** $x_0 \in \mathbb{R}$ if

$$P(X \leq x_0 - x) = P(X \geq x_0 + x)\,, \text{ for any } x \in \mathbb{R}.$$

- If there exists no point $x \in \mathbb{R}$ such that $P(X = x) > 0$, then X is symmetric about $x_0 \in \mathbb{R}$ if and only if

$$F(x_0 - x) = 1 - F(x_0 + x)\,, \text{ for any } x \in \mathbb{R}.$$

- If X is an absolutely continuous random variable with a density function f, then X is symmetric about $x_0 \in \mathbb{R}$ if and only if

$$f(x_0 - x) = f(x_0 + x)\,, \text{ for any } x \in \mathbb{R}.$$

A random variable X is called **symmetric** if there exists $x_0 \in \mathbb{R}$ about which X is symmetric.

Skewness coefficients

A skewness coefficient is a function

$$\begin{aligned} \gamma :\ \mathcal{L}(\Omega, \mathcal{A}, P) &\longrightarrow \mathbb{R} \\ X &\hookrightarrow \gamma(X)\,. \end{aligned}$$

An axiomatization of the notion of symmetry (Oja, 1981)

Let $X, Y \in \mathcal{L}(\Omega, \mathcal{A}, P)$ be two random variables:

1. If X is symmetric, then $\gamma(X) = 0$.
2. $\gamma(cX + d) = \gamma(X)$ where $c, d \in \mathbb{R}$ with $c > 0$.
3. $\gamma(-X) = -\gamma(X)$.
4. If $X \precsim Y$, *i.e.*, $F_Y^{-1}(F_X(x))$ is convex for all $x \in \mathbb{R}$, then $\gamma(X) \leq \gamma(Y)$.

Skewness coefficients

Some skewness coefficients in the literature

Let $X \in \mathcal{L}(\Omega, \mathcal{A}, P)$:

- $S_P(X) = \dfrac{\mu - Mo(X)}{\sigma} \longrightarrow$ Pearson (1895).
- $\gamma_1(X) = \dfrac{E\left[(X-\mu)^3\right]}{\sigma^3} \longrightarrow$ Charlier and Edgeworth (1905).
- $\gamma'_{Me(X)}(X) = 3 \cdot \dfrac{\mu - Me(X)}{\sigma} \longrightarrow$ Yule (1912).
- $\gamma_p(X) = \dfrac{C_{1-p}(X) - 2\,C_{0.5}(X) + C_p(X)}{C_{1-p}(X) - C_p(X)}$ for certain $p \in]0, 0.5[\longrightarrow$ Hinkley (1975).

Index

Index

A new family of skewness coefficients

Let $\mathcal{L}'(\Omega, \mathcal{A}, P)$ denote the set of absolutely continuous r.v. defined over $(\Omega, \mathcal{A}, P)$ whose cdf is strictly increasing on the preimage of $]0,1[$ and twice differentiable.

A new family of skewness coefficients

Let $X \in \mathcal{L}'(\Omega, \mathcal{A}, P)$:

$$\gamma(X) = \frac{(\mathbf{v} - 2\mathbf{w} + \mathbf{v}^r)^T \mathbf{C_p}(X)}{(\mathbf{v} - \mathbf{v}^r)^T \mathbf{C_p}(X)} = \frac{\mathbf{a}^T \mathbf{C_p}(X)}{\mathbf{b}^T \mathbf{C_p}(X)},$$

is a family of skewness coefficients if:

- $k \in \mathbb{N}$ is odd.
- $\mathbf{C_p}(X) \in \mathbb{R}^k$ are increasingly-ordered quantiles of X such that if $C_q(X) \in \mathbf{C_p}(X)$ for some $q \in]0,1[$, then $C_{1-q}(X) \in \mathbf{C_p}(X)$.
- $\mathbf{v}_i = 0$ for any $i \in \{1, \ldots, \lceil k/2 \rceil\}$.
- $\mathbf{w}_{\lceil \frac{k}{2} \rceil} = 1$ and $\mathbf{w}_i = 0$ for any $i \neq \lceil k/2 \rceil$.

Index

Sample version of the new family of coefficients

Sample version

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T \in \mathbb{R}^n$ be a simple random sample from an absolutely continuous random variable X. The sample version of γ is defined as:

$$\hat{\gamma}(\mathbf{X}) = \frac{(\mathbf{v} - 2\mathbf{w} + \mathbf{v})^T \, \widehat{\mathbf{C}_\mathbf{p}}(\mathbf{X})}{(\mathbf{v} - \mathbf{v}^r)^T \, \widehat{\mathbf{C}_\mathbf{p}}(\mathbf{X})} = \frac{\mathbf{a}^T \widehat{\mathbf{C}_\mathbf{p}}(\mathbf{X})}{\mathbf{b}^T \widehat{\mathbf{C}_\mathbf{p}}(\mathbf{X})},$$

where $\widehat{C_q}(\mathbf{X}) = \mathbf{X}_{(\lceil nq \rceil)}$ for any $q \in]0, 1[$.

A sufficient condition for it being well-defined:

$$n \geq \frac{1}{\min_{i=1,\ldots,m} \mathbf{p}_{i+1} - \mathbf{p}_i}.$$

Sample version of the new family of coefficients

Asymptotic distribution (1 sample)

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ be a random sample drawn from $X \in \mathcal{L}'(\Omega, \mathcal{A}, P)$. Then:

$$\sqrt{n}\frac{\hat{\gamma}(\mathbf{X}) - \gamma(X)}{\tau_X} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)\,, \qquad \text{with}$$

$$\tau_X^2 = \frac{4}{\left[(\mathbf{v}^T - \mathbf{v}^{r^T})\mathbf{C}_\mathbf{p}(X)\right]^4} \sum_{j=1}^{k}\sum_{i=1}^{k} \frac{\min\{\mathbf{p}_i, \mathbf{p}_j\}(1 - \max\{\mathbf{p}_i, \mathbf{p}_j\})}{f(\mathbf{C}_{\mathbf{p}_i}(X))f(\mathbf{C}_{\mathbf{p}_j}(X))}$$
$$\left[\mathbf{v}_i(\mathbf{w}^T - \mathbf{v}^{r^T}) + \mathbf{v}_i^r(\mathbf{v}^T - \mathbf{w}^T) + \mathbf{w}_i(\mathbf{v}^{r^T} - \mathbf{v}^T)\right]\mathbf{C}_\mathbf{p}(X)$$
$$\cdot\left[\mathbf{v}_j(\mathbf{w}^T - \mathbf{v}^{r^T}) + \mathbf{v}_j^r(\mathbf{v}^T - \mathbf{w}^T) + \mathbf{w}_j(\mathbf{v}^{r^T} - \mathbf{v}^T)\right]\mathbf{C}_\mathbf{p}(X)\,.$$

An analogous result holds when estimating τ_X such that $\hat{\tau}^2(\mathbf{X}) \xrightarrow{\mathcal{P}} \tau_X^2$, which can be achieved by estimating $\mathbf{C}_\mathbf{p}(\mathbf{X})$ with $\mathbf{X}_{(\lceil n\mathbf{p}\rceil)}$, and f with its kernel density estimator f_n.

Sample version of the new family of coefficients

Theorem 1: Asymptotic distribution (2 samples)

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)^T$ be two independent random samples drawn from two independent random variables $X, Y \in \mathcal{L}'(\Omega, \mathcal{A}, P)$. Then:

$$\frac{(\hat{\gamma}(\mathbf{X}) - \gamma(X)) \pm (\hat{\gamma}(\mathbf{Y}) - \gamma(Y))}{\sqrt{\frac{\tau_X^2}{n} + \frac{\tau_Y^2}{m}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \, ,$$

where τ_X^2 and τ_Y^2 are defined as before.

Remark: We assume $n = g(m)$ with $g : \mathbb{N} \to \mathbb{N}$ an increasing function such that $\lim_{m\to\infty} \frac{m}{g(m)} = \lambda \in]0, \infty[$, *i.e.*, both sample sizes grow to infinity at comparable rates.

Index

Two-sample goodness-of fit tests to a location-scale family

Framework

In the following, we assume that

- $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)^T$ are two independent random samples drawn from two independent random variables $X, Y \in \mathcal{L}'(\Omega, \mathcal{A}, P)$.
- $n = g(m)$ with $g : \mathbb{N} \to \mathbb{N}$ increasing such that $\lim_{m\to\infty} \frac{m}{g(m)} = \lambda \in]0, \infty[$.

We aim to test whether the underlying distributions of **X** and **Y** belong to the same location-scale family, which might be known or unknown.

Index

Goodness-of fit to a known location-scale family

Setting up the hypotheses of the test ($\mathcal{F}$ known)

The hypotheses in a goodness-of-fit test to a known location–scale family $\mathcal{F}_0$ are the following:

$$\begin{cases} H_0 : \text{X and Y belong to the same location-scale family } \mathcal{F}_0 \, . \\ H_1 : \text{X and Y } \textbf{do not} \text{ belong to the same location-scale family } \mathcal{F}_0 \, . \end{cases}$$

Known location scale family

- From Theorem 1, it follows that $\frac{(\hat{\gamma}(\mathbf{X})-\gamma(X))+(\hat{\gamma}(\mathbf{Y})-\gamma(Y))}{\sqrt{\frac{\tau_X^2}{n}+\frac{\tau_Y^2}{m}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.
- Under H_0, it holds that $\mathcal{F}_X = \mathcal{F}_Y = \mathcal{F}_0$. Therefore, by Axiom 2,
  - $\gamma(X) = \gamma(Y) \Longrightarrow \gamma(X) + \gamma(Y) = 2\,\gamma_{\mathcal{F}_0}$.
  - $\tau_X = \tau_Y = \tau_{\mathcal{F}_0}$.

Test statistic and rejection region

To perform the test, the following statistic is used

$$\frac{\sqrt{nm}}{\sqrt{n+m}} \frac{(\hat{\gamma}(\mathbf{X}) + \hat{\gamma}(\mathbf{Y}) - 2\,\gamma_{\mathcal{F}_0})}{\tau_{\mathcal{F}_0}} \xrightarrow[H_0]{\mathcal{D}} \mathcal{N}(0,1)\,.$$

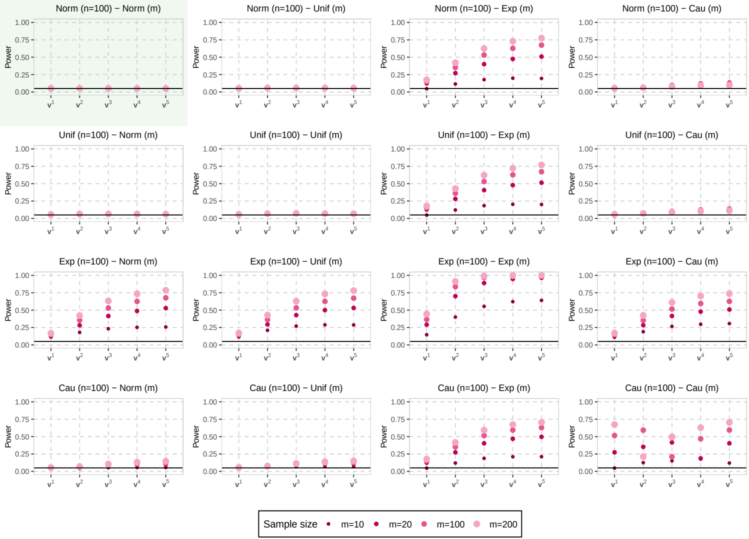
$$\text{R.R.} = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m} \,\middle|\, |\hat{\gamma}(\mathbf{x}) + \hat{\gamma}(\mathbf{y}) - 2\,\tau_{\mathcal{F}_0}| > \tau_{\mathcal{F}_0} \sqrt{\frac{n+m}{n\,m}} z_{1-\alpha/2} \right\} .$$

Goodness-of fit to a known location-scale family

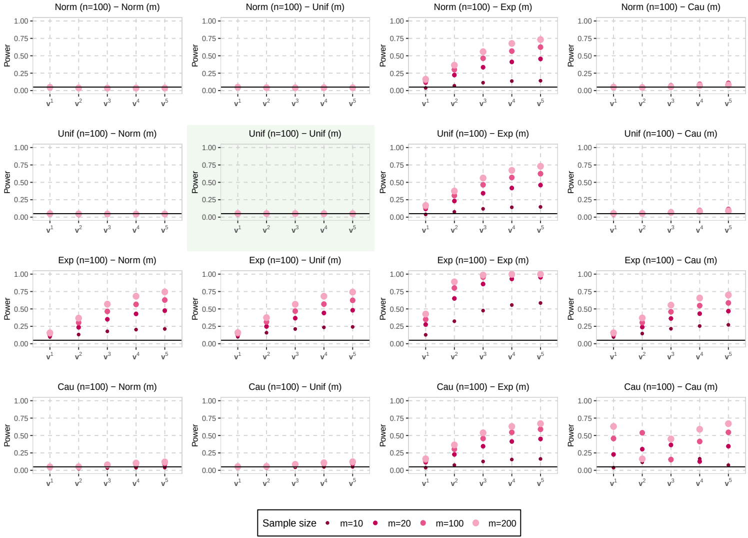
Experimental framework

- Goodness-of fit to the normal, uniform and (shifted) exponential distributions.
- Significance level $\alpha = 0.05$.
- The power of the tests is estimated by Monte Carlo simulation (10^5 replications).
- Expected behaviour:
  1. Under H_0: Power $\in [0; 0.0511]$.
  2. Under H_1: Power ≈ 1.
- $n = 100$.
- $m \in \{10, 50, 100, 200\}$.
- $\mathbf{w} = (0, 0, 1, 0, 0)^T$.
- $\mathbf{v} = \lambda\,(0, 0, 0, 1, 0)^T + (1 - \lambda)\,(0, 0, 0, 0, 1)^T$, with $\lambda \in \{0, 1/4, 1/2, 3/4, 1\}$
- $\mathbf{C_p}$ such that $\mathbf{p} = (1/6, 2/6, 3/6, 4/6, 5/6)^T$.

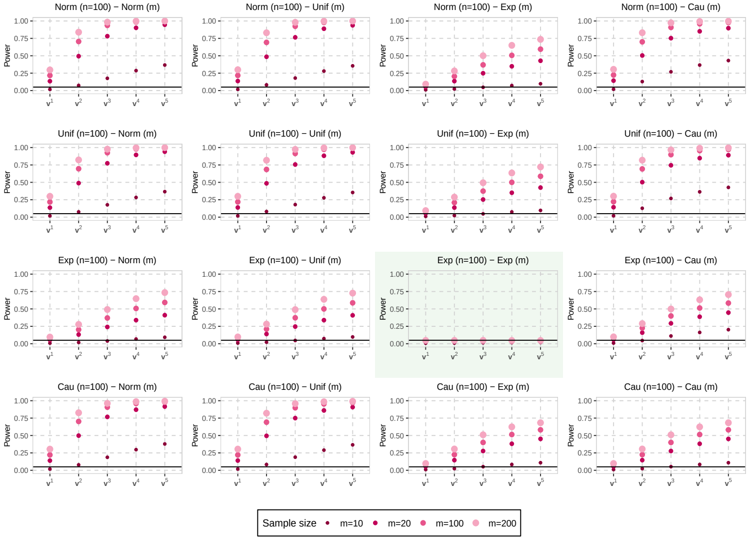
Goodness-of fit to the normal distribution



Goodness-of fit to the uniform distribution



Goodness-of fit to the (shifted) exponential distribution



Index

Goodness-of-fit to an unknown location-scale family

Setting up the hypotheses of the test ($\mathcal{F}$ unknown)

The hypotheses in a goodness-of-fit test to an unknown location–scale family are the following:

$$\begin{cases} H_0 : \text{X and Y belong to the same location-scale family}\,. \\ H_1 : \text{X and Y } \textbf{do not} \text{ belong to the same location-scale family}\,. \end{cases}$$

Goodness-of-fit to an unknown location scale family

- From Theorem 1, it follows that $\frac{(\hat{\gamma}(\mathbf{X})-\gamma(X))-(\hat{\gamma}(\mathbf{Y})-\gamma(Y))}{\sqrt{\frac{\tau_X^2}{n}+\frac{\tau_Y^2}{m}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.
- We estimate $\hat{\tau}(\mathbf{X})$ and $\hat{\tau}(\mathbf{Y})$ such that $\hat{\tau}^2(\mathbf{X}) \xrightarrow[m]{P} \tau_X^2$ and $\hat{\tau}^2(\mathbf{Y}) \xrightarrow[m]{P} \tau_Y^2$.
- Under H_0: $\gamma(X)=\gamma(Y) \implies \gamma(X)-\gamma(Y)=0$ because $\mathcal{F}_X=\mathcal{F}_Y$ and Axiom 2.

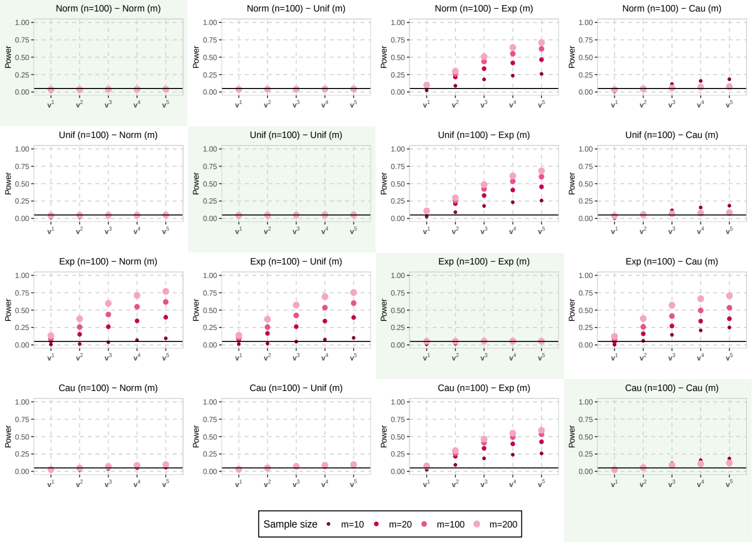
Test statistic and rejection region

To perform the test, the following statistic is used

$$\frac{\hat{\gamma}(\mathbf{X})-\hat{\gamma}(\mathbf{Y})}{\sqrt{\frac{m\,\hat{\tau}^2(\mathbf{X})+n\,\hat{\tau}^2(\mathbf{Y})}{n\,m}}} \xrightarrow[H_0]{\mathcal{D}} \mathcal{N}(0,1)\,.$$

$$\text{R.R.} = \left\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{n+m} \;\middle|\; \left|\frac{\hat{\gamma}(\mathbf{x})-\hat{\gamma}(\mathbf{y})}{\sqrt{m\,\hat{\tau}^2(\mathbf{x})+n\,\hat{\tau}^2(\mathbf{y})}}\right| > \frac{z_{1-\alpha/2}}{\sqrt{n\,m}}\right\}.$$

Goodness-of fit to an unknown location-scale family



Index

Conclusions

Conclusions of the experimental setup

- **Under** H_0: Significance level $\alpha = 0.05$ is preserved for any sample size m and weight vector $\mathbf{v}$.
- The tests have difficulty distinguishing location-scale families whenever the skewness coefficient takes the same value.
- **Under** H_1**:** The performance is more accurate when the location-scale family under H_0 and the samples involved in the tests differ in their skewness coefficient.
- **Choice of weight vector v:**
  - Under H_0: Irrelevant
  - Under H_1: Right-weighted vectors $\mathbf{v}$.

Future work

To be continued...

1. To extend the definition of the new family of skewness coefficients to a **wider range of weight vectors**.
2. To introduce a statistical measure (kurtosis) that allows us to **distinguish between random variables** with the same value of skewness $\rightarrow$ Jarque-Bera test.

Funding

Funding from grant PID2022-140585NB-I00 funded by MICIU/AEI/10.13039/501100011033 and "FEDER/UE".

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Cofinanciado por la Unión Europea

AGENCIA ESTATAL DE INVESTIGACIÓN

Two-sample goodness-of-fit tests to location-scale families using OWA-based skewness coefficients

Marina Iturrate–Bobes, Ignacio Montes and Raúl Pérez–Fernández

University of Oviedo

Universidad de Oviedo

21st International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)
June 15th–19th 2026