Bayesian Tests for Goodness of Fit

Ronald van Nooijen[1] **and Alla Kolechkina**[1,2]

[1]Faculty of Civil Engineering and Geosciences, Delft University of Technology

[2]Aronwis, Den Hoorn Z.H., The Netherlands

r.r.p.vannooyen@tudelft.nl, a.g.kolechkina@tudelft.nl

1. Introduction

In hydrology one often wants to know whether or not a series of observations corresponds to a series of samples from a given probability distribution. Examples are high flows, precipitation sums and so on. Traditionally such questions are phrased in terms of *p*-values. Given the popularity of the Bayesian approach elsewhere in hydrology, it seems natural to examine how it could be used in this context.

2. Conventions

Upper case S, T, X, Y, Z: random variables
$\vec{x}$: vector
$\vec{X}$: random vector
$\theta, \vec{\theta}$: parameter or parameter *v*ector of probability distribution

3. Problem statement

We have a random variable X distributed according to an unknown probability measure P. Our hypothesis to be tested is: $H_0 =$ "P is a member of a family of probability measures $\mathscr{P}$". We assume that to each element $P \in \mathscr{P}$ we can assign parameters $\vec{\theta} \in L \subset \mathbb{R}^m$ and that for each P the random variable X induces a probability density function (pdf) $f\left(x \mid \vec{\theta}\right)$ and a cumulative distribution function (cdf) $F\left(x \mid \vec{\theta}\right)$ on $\mathbb{R}$.

Now suppose $\vec{Y}$ is a random vector of length n with independent identically distributed (iid) random variables, each with the same distribution as X. We wish to make a statement about our hypothesis based on a sample $\vec{y}$ of $\vec{Y}$. To do this we assume we have a function s_n from $\mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}$ such that

$$\Pr\left(s_n\left(\vec{Y}, \vec{\theta}_0\right) > t | P \in \mathscr{P} \text{ and } \vec{\theta} = \vec{\theta}_0\right)$$

is decreasing for increasing t.

4. A practical example

Let $H_0 =$ "P is a member of the Gumbel distribution family with location and scale parameters", so

$$F\left(x \mid \vec{\theta}\right) = F\left(x \mid \langle \xi, \zeta \rangle\right) = \exp\left(-\exp\left(-\frac{x-\xi}{\zeta}\right)\right)$$

For s_n we take the Kolmogorov-Smirnov distance D_{KS} between the empirical cdf (ecdf) and the Gumbel cdf.

$$D_{KS}(\vec{y}, \langle \xi, \zeta \rangle) = \sup_x |F\left(x \mid \langle \xi, \zeta \rangle\right) - F_n(x, \vec{y})|$$

where

$$F_n(x, \vec{y}) = \frac{\text{number of } y_i \text{ smaller than } x}{n}$$

We used R 2.15.0 (R Development Core Team, 2012) and the additonal package package `evd` (Stephenson, 2002) to perform the experiments.

4.1 Classical *p*-value

Classical *p*-value for known parameters

$$\Pr\left(s_n\left(\vec{Y}, \langle 0, 1 \rangle\right) > s_n\left(\vec{y}, \langle 0, 1 \rangle\right) \mid H_0 \text{ and } \xi = 0, \zeta = 1\right)$$

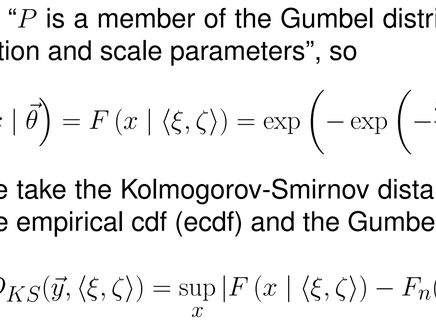

Figure 1: *ecdf of p-values Gumbel with* $\langle \xi, \zeta \rangle = \langle 0, 1 \rangle$ *(sample of 10000 vectors with* $n = 32$*)*

Plug-in *p*-value. Let $\langle \hat{\xi}, \hat{\zeta} \rangle$ be the maximum likelihood estimate of the parameters based on the sample $\vec{y}$.

$$\Pr\left(s_n\left(\vec{Y}, \langle \hat{\xi}, \hat{\zeta} \rangle\right) > s_n\left(\vec{y}, \langle \hat{\xi}, \hat{\zeta} \rangle\right) \mid H_0, \xi = \hat{\xi}, \zeta = \hat{\zeta}\right)$$

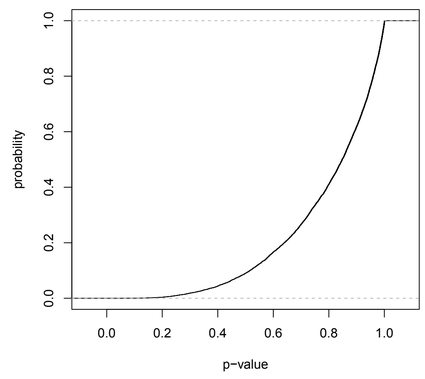

Figure 2: *ecdf of plug-in p-values for samples from Gumbel with* $\langle \xi, \zeta \rangle = \langle 0, 1 \rangle$ *(sample of 10000 vectors with* $n = 32$*)*

The problem (the *p*-value is no longer equal to the probability of its occurrence) is caused by the double use of the sample. Instead of using the distribution of

$$s_n\left(\vec{Y}, \langle \hat{\xi}(\vec{y}), \hat{\zeta}(\vec{y}) \rangle\right)$$

we need find the distribution of

$$s_n\left(\vec{Y}, \langle \hat{\xi}(\vec{Y}), \hat{\zeta}(\vec{Y}) \rangle\right)$$

One way to recalibrate: Let our original sample be $\vec{y}_0$. Determine $\hat{\xi}_0 = \hat{\xi}(\vec{y}), \hat{\zeta}_0 = \hat{\zeta}(\vec{y}_0)$. Calculate $d_0 = s_n(\vec{y}_0, \langle \hat{\xi}_0, \hat{\zeta}_0 \rangle)$. Next generate samples $\vec{y}_k$ with $k = 1, 2, \ldots, K$ from Gumbel with $\xi = \hat{\xi}_0, \zeta = \hat{\zeta}_0$ and calculate $d_k = s_n(\vec{y}_k, \langle \hat{\xi}(\vec{y}_k), \hat{\zeta}(\vec{y}_k) \rangle)$. The recalibrated *p*-value is

$$\frac{\text{number of } d_k \text{ smaller than } d_0}{K}$$

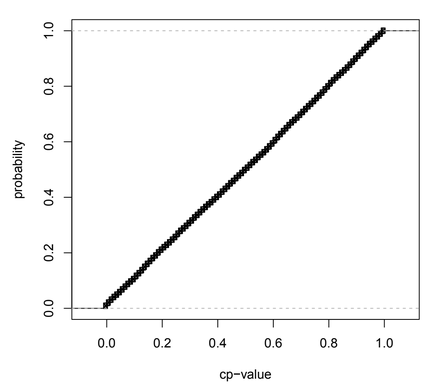

Figure 3: *ecdf of recalibrated plug-in p-values for samples from Gumbel with* $\langle \xi, \zeta \rangle = \langle 0, 1 \rangle$ *(sample of 10000 vectors with* $n = 32$*,* $K = 100$*)*

4.2 Posterior predictive *p*-value

For a sample $\vec{y}_0$ the posterior predictive p value is defined as

$$\text{ppp}\left(\vec{y}_0\right) = \Pr\left(s_n\left(\vec{Y}, \vec{\Theta}\right) \geq d\left(\vec{y}_0, \vec{\Theta}\right) \mid \vec{y}_0\right)$$

where $\left\langle \vec{Y}, \vec{\Theta} \right\rangle$ has pdf

$$f\left(\left\langle \vec{y}, \vec{\theta} \right\rangle \mid \vec{y}_0, H_0\right) \propto f\left(\vec{y} \mid \vec{\theta}\right) \pi\left(\vec{\theta} \mid \vec{y}_0\right)$$

and

$$\pi\left(\vec{\theta} \mid \vec{y}_0\right) \propto f\left(\vec{y}_0 \mid \vec{\theta}\right) \pi\left(\vec{\theta}\right)$$

Again the sample is used twice, first in the calculation s_n and then again in the determination of the posterior distribution of $\left\langle \vec{Y}, \vec{\Theta} \right\rangle$.

Again we take a sample from the Gumbel distribution with parameters $\langle 0, 1 \rangle$. For the purpose of this example we use a proper prior,

$$\pi\left(\langle \xi, \zeta \rangle\right) = \begin{cases} 0 & \xi < -4 \text{ or } \xi > 4 \text{ or } \zeta < 0.25 \text{ or } \zeta > 4 \\ \frac{1}{8 \times 3.75} \frac{1}{\zeta} & -4 \leq \xi \leq 4 \text{ and } 0.25 < \zeta < 4 \end{cases}$$

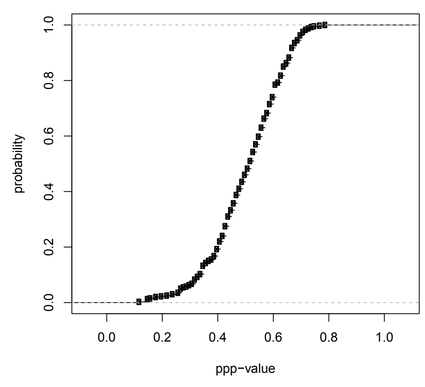

Figure 4: *ecdf of ppp-values for samples from Gumbel with* $\langle \xi, \zeta \rangle = \langle 0, 1 \rangle$ *(400 vectors with* $n = 32$*)*

Again we recalibrate, we use the procedure outlined in (Hjort et al., 2006),

For given $\vec{y}_0$ take J samples $\left\{\vec{\theta}_j\right\}_{j=1}^{J}$ from $\pi\left(\vec{\theta} \mid \vec{y}_0\right)$.
Next for each j sample $\vec{y}_j$ from $f_{\vec{Y}}\left(\vec{y} \mid \vec{\theta}_j\right)$.
This results in a sample $\left\{\left(\vec{y}_j, \vec{\theta}_j\right)\right\}_{j=1}^{J}$ from a distribution with density $\propto f_{\vec{Y}}\left(\vec{y} \mid \vec{\theta}, H_0\right) \pi\left(\vec{\theta} \mid \vec{y}_0\right)$.
Define

$$p^{\text{MC}}_{\text{ppp}}\left(\vec{y}_0\right) = \frac{\text{number of } s_n\left(\vec{y}_j, \vec{\theta}_j\right) \geq s_n\left(\vec{y}_0, \vec{\theta}_j\right)}{J}$$

Algorithm 1: *Posterior predictive p-value by Monte Carlo*

For given $\vec{y}_0$ determine $p^{\text{MC}}_{\text{ppp}}\left(\vec{y}_0\right)$ (Algorithm 1).
Take K samples from $\pi\left(\vec{v}\right)$, resulting in $\left\{\vec{\theta}_k\right\}_{k=1}^{K}$.
For each k, sample $\vec{y}_k$ from $f_{\vec{Y}}\left(\vec{y} \mid \vec{\theta}_k\right)$.
Result: a sample $\left\{\left(\vec{y}_k, \vec{\theta}_k\right)\right\}_{k=1}^{K}$ from $\propto f_{\vec{Y}}\left(\vec{y} \mid \vec{\theta}\right) \pi\left(\vec{\theta}\right)$.
For each $\vec{y}_k$ determine $p^{\text{MC}}_{\text{ppp}}\left(\vec{x}_k\right)$ (Algorithm 1) and define

$$p^{\text{MC}}_{\text{cppp}}\left(\vec{y}_0\right) = \frac{\text{number of } p^{\text{MC}}_{\text{ppp}}\left(\vec{y}_k\right) \geq p^{\text{MC}}_{\text{ppp}}\left(\vec{y}_0\right)}{K}$$

Algorithm 2: *Calibrated posterior predictive p-value by Monte Carlo*

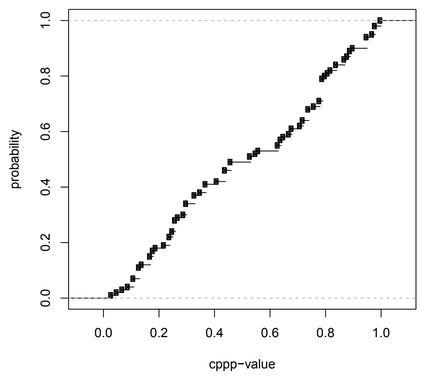

Figure 5: *ecdf of cppp-values for samples from Gumbel with* $\langle \xi, \zeta \rangle = \langle 0, 1 \rangle$ *(400 vectors of length 32)*

5. Conclusions

The literature is clear on the fact that frequentist *p*-values for composite hypothesis need to be interpreted carefully, see for example D'Agostino and Stephens (1986). Bayesian *p*-values are not as well known, but here too the literature advises caution, see also Bayarri and Berger (2000), Hjort et al. (2006). We presented an example where both the frequentist and Bayesian *p*-values needed an additional calibration.

References

M. J. Bayarri and James O. Berger. P values for composite null models. *Journal of the American Statistical Association*, 95(452):pp. 1127–1142, 2000.

Ralph B. D'Agostino and Michael A. Stephens, editors. *Goodness-of-fit techniques*. Statistics: Textbooks and Monographs. Marcel Dekker Inc., New York, 1986.

Nils Lid Hjort, Fredrik A. Dahl, and Gunnhildur Högnadóttir Steinbakk. Post-processing posterior predictive p values. *Journal of the American Statistical Association*, 101(475):pp. 1157–1174, 2006.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

A. G. Stephenson. evd: Extreme value distributions. *R News*, 2(2), June 2002.