On Determining the Distribution of a Goodness-of-Fit Test Statistic

Sean van der Merwe[a,*]

[a]University of the Free State, Box 339, Bloemfontein, 9300, South Africa
[*]Corresponding author

May 12, 2014

Abstract

We consider the problem of goodness-of-fit testing for a model that has at least one unknown parameter that cannot be eliminated by transformation. Examples of such problems can be as simple as testing whether a sample consists of independent Gamma observations, or whether a sample consists of independent Generalised Pareto observations given a threshold. Over time the approach to determining the distribution of a test statistic for such a problem has moved towards on-the-fly calculation post observing a sample. Modern approaches include the parametric bootstrap and posterior predictive checks. We argue that these approaches are merely approximations to integrating over the posterior predictive distribution that flows naturally from a given model. Further, we attempt to demonstrate that shortcomings which may be present in the parametric bootstrap, especially in small samples, can be reduced through the use of objective Bayes techniques, in order to more reliably produce a test with the correct size.

Keywords: Bayes, Distribution, Gamma, GPD, Hypothesis Testing, Objective Bayes, p-value, Predictive Posterior, Simulation

1 Introduction

1.1 Distribution tests where the null model is completely specified

Well-known tests for determining whether a sample could have arisen from a specific distribution include the Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and similar tests based on the empirical distribution function (see Darling, 1957 for a historical introduction); however, these tests in their base form assume that all parameters in the null model are known.

Specifically, let $\mathbf{X}$ be an i.i.d. random sample of size n from an unknown distribution, and let $S = S(\mathbf{X}|m)$ be a test statistic for testing the null hypothesis that $\mathbf{X}$ follows a specific distribution. In general, the test statistic S depends on the parameters $\boldsymbol{\theta}$ of the distribution to be tested, as is the case for the KS and AD statistics. Thus, $S = S(\mathbf{X}|\boldsymbol{\theta}, m)$ is a function of $\boldsymbol{\theta}$ in general, so that in their base form many distribution tests are suitable for testing a null hypothesis of the form: $\mathbf{X}$ follows a specific distribution with parameters $\boldsymbol{\theta}_0$ fixed. The test statistic used for testing a null-hypothesis of this form is then $S(\mathbf{X}|\boldsymbol{\theta}_0, m)$.

Note that for fixed (or known) $\boldsymbol{\theta}_0$, the exact distribution of $S(\mathbf{X}|\boldsymbol{\theta}_0, m)$ can often be determined, if necessary by simulation. Hence, we can have an exact test of the null hypothesis that $\mathbf{X}$ follows a specific distribution with parameters $\boldsymbol{\theta}_0$. By the test being exact, in the classical sense, we mean that the test has correct type 1 error, that is, for a given significance level α^* (say) the test falsely rejects the null hypothesis with probability α^*. This condition is equivalent to the p-value having a Uniform distribution under the null hypothesis.

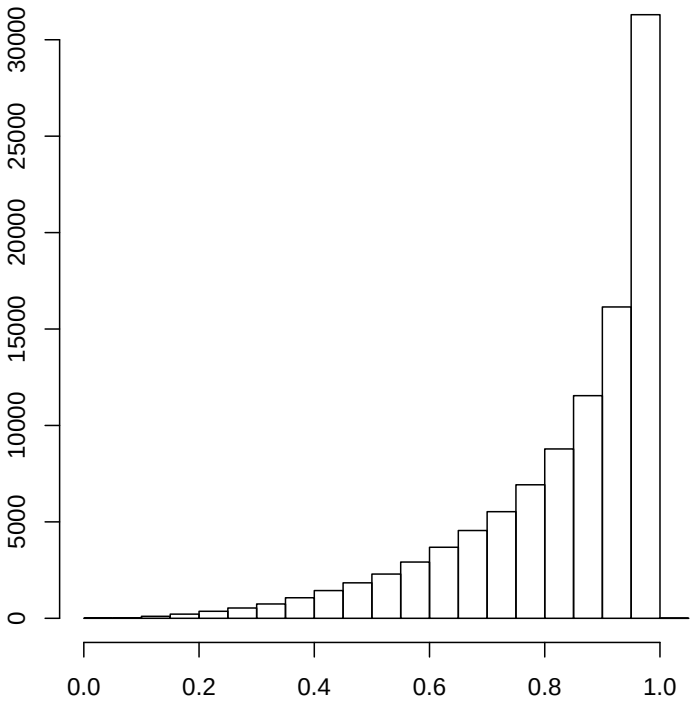

Figure 1: Calculated p-values from Gamma samples using Gamma test of G. Marsaglia and J. Marsaglia (2004).

1.2 Parameters of null model not specified

In practice it is often of interest to test whether $\mathbf{X}$ follows a specific distribution, but without specifying the parameters $\boldsymbol{\theta}$ of the distribution. If, nevertheless, the relevant test statistic $S = S(\mathbf{X}|\boldsymbol{\theta}, m)$ is a function of $\boldsymbol{\theta}$, then S might be calculated as $S = S(\mathbf{X}|\hat{\boldsymbol{\theta}}, m)$, where $\hat{\boldsymbol{\theta}}$ is some estimate of $\boldsymbol{\theta}$. When doing so in general, the problem arises that the distribution of the test statistic $S(\mathbf{X}|\hat{\boldsymbol{\theta}}, m)$ might not only depend on the distribution family and sample size, but also on the values of the unknown parameters $\boldsymbol{\theta}$ (D'Agostino and Stephens, 1986, p. 102, Darling, 1957). The distribution of the test statistic might even be affected by the method of estimation of the unknown parameters.

To illustrate this problem 100,000 samples from a $Gamma(6, 2)$ distribution were simulated and the Gamma test of G. Marsaglia and J. Marsaglia (2004) in the ADGofTest package (Bellosta, 2011) in the statistical software R (R Core Team, 2013) was performed. This test is based on the principle of replacing the unknown parameters of the Gamma distribution by their estimates. The histogram of the simulated p-values is given in Figure 1 below. The p-values in Figure 1 are clearly not uniformly distributed. Therefore, the Gamma test of G. Marsaglia and J. Marsaglia (2004) is not exact and fails to reject far too often, which results in a lack of power or false confidence in a chosen model.

If the test statistic can be standardised in some way so that it is parameter invariant (its value and distribution do not depend on the parameters of the model) then it is usually possible to simulate accurately the distribution of the test statistic under the null hypothesis. This strategy works for location-scale and log-location-scale distributions (D'Agostino and Stephens, 1986, p. 102). For the Gamma distribution, as a convenient counterexample, it is not possible to eliminate the shape parameter and thus a different approach is needed.

Gelman, Meng, et al. (1996) describe a general, Bayesian approach for calculating a posterior predictive p-value (ppp) based on the ideas of Rubin et al. (1984). More recently, authors have considered various problems using the methodology of Gelman, Meng, et al. (1996), including multivariate data (Crespi and Boscardin, 2009), discrete data (Gelman, Goegebeur, et al., 2000), hierarchical models (Sinharay and Stern, 2003), pharmacokinetic models (Yano et al., 2001), *etc.*

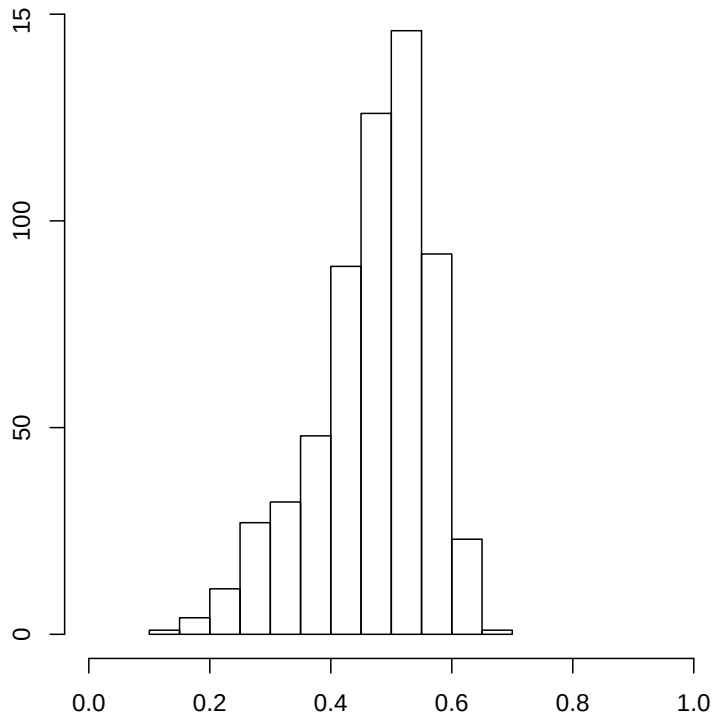

Figure 2: Calculated posterior predictive p-values from Gamma samples.

However, the approach of Gelman, Meng, et al. (1996) has been criticised — see, for example, the comment by Rubin on Gelman, Meng, et al. (1996), or Bayarri and James O. Berger (2000). Of note, when the test statistic chosen depends on the parameters of the model then the resulting test is not exact in the classical sense explained above. To illustrate this we simulated 600 $Gamma(4, 8)$ samples of size 12 and calculated the ppp based on the AD statistic for each one. The resulting histogram is given in Figure 2. It is clear that the p-values are pulled toward 0.5 and require calibration.

This problem of non-Uniform p-values is explained in detail in Robins et al. (2000), along with some methods of addressing it asymptotically. Among these methods is what Robins et al. (2000) refer to as the "double parametric bootstrap", which in turn is based on an idea of Beran (1988), who called it pre-pivoting. The approach described in Section 2 of this paper is a fully Bayesian adaptation of these ideas.

1.3 Objectives and outline of the present paper

In this paper the generic problem of testing whether a sample originates from a hypothesized model where some or all parameters of the model are unknown is addressed. A new test is introduced, based on the posterior and posterior predictive distributions, that produces valid p-values in the classical sense. The new test is compared to the parametric bootstrap in two examples:

1. The first is the independent and identically distributed Gamma observations model (Section 3) which is an example where the parametric bootstrap approach works well and we show that the new test performs equally well, both in terms of achieving the target significance level and in terms of power.
2. The second is the independent and identically distributed Generalised Pareto observations model given a known threshold (Section 4) where we note that the new test procedure comes much closer to achieving the desired significance level than the parametric bootstrap approach, and as a result, achieves higher power for the same test statistic.

In Section 2 we motivate the new test, and present an algorithm for its implementation. The core of the algorithm rests on the idea that in order to arrive at an accurate test we must make full use of all

information that can be obtained from the sample. This goal can be achieved through an objective Bayes framework.

In Section 5 and Section 6 we briefly summarise the results of the experiments and give concluding remarks.

2 New suggested methodology

2.1 Mathematical motivation

Let $\mathbf{X}$ be a random variable of dimension n from an unknown distribution, $\mathbf{x}$ an observation of $\mathbf{X}$, (that is, $\mathbf{x}$ is the observed sample), and m a hypothesized model with unknown parameter values $\boldsymbol{\theta}$. Throughout, we will use bold font to denote vectors. We denote observed quantities with lower case letters and random variables with upper case letters, except for $\boldsymbol{\theta}$ which we consider to be a scalar or vector random variable throughout. Assume that, after having observed $\mathbf{x}$, the parameter uncertainty is captured in a posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, m)$ for $\boldsymbol{\theta}$. First choose a summary statistic S that compares the sample $\mathbf{X}$ to the model m in a meaningful way; we write $S = S(\mathbf{X}|m)$. In principle, no restriction is placed on the form of the test statistic other than the notion that it should be a function of the sample, and optionally of the parameter values, given a model, and that it should increase as the discrepancy between the sample and the model increases.

In general, in order to calculate S the parameters $\boldsymbol{\theta}$ need to be specified, as is the case for the KS and AD statistics. Thus, we can usually only calculate the statistic in the form $S = S(\mathbf{X}|\boldsymbol{\theta}, m)$.

In order to remove the dependence of $S(\mathbf{X}|\boldsymbol{\theta}, m)$ on $\boldsymbol{\theta}$ we replace it by its expectation under the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X}, m)$ for $\boldsymbol{\theta}$ given $\mathbf{X}$. That is, we define $S(\mathbf{X}|m)$ as

$$S(\mathbf{X}|m) = E_{\boldsymbol{\theta}}[S(\mathbf{X}|\boldsymbol{\theta}, m)] = \int S(\mathbf{X}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|\mathbf{X}, m)d\boldsymbol{\theta} \tag{1}$$

The statistic $S(\mathbf{X}|m)$ is a random variable, and we can determine its distribution if we can determine the distribution of $\mathbf{X}$. Under a Bayesian approach, given an observed sample $\mathbf{x}$ from model m, we work with the posterior predictive distribution of $\mathbf{X}$, namely

$$p(\mathbf{X}|\mathbf{x}, m) = \int f(\mathbf{X}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|\mathbf{x}, m)d\boldsymbol{\theta} \tag{2}$$

where $f(\mathbf{X}|\boldsymbol{\theta}, m)$ is the likelihood implied by the model. In many cases f can be expressed explicitly, but this is not a requirement – being able to simulate from the model as well as from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X}, m)$ is sufficient for implementing this step.

Given an observed value

$$s(\mathbf{x}|m) = E_{\boldsymbol{\theta}}[s(\mathbf{x}|\boldsymbol{\theta}, m)] = \int s(\mathbf{x}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|\mathbf{x}, m)d\boldsymbol{\theta} \tag{3}$$

of the test statistic $S(\mathbf{X}|m)$, we now calculate the probability that $S(\mathbf{X}|m) \geq s(\mathbf{x}|m)$, namely

$$P = P(S(\mathbf{X}|m) \geq s(\mathbf{x}|m)) = \int_{S(\mathbf{X}|m) \geq s(\mathbf{x}|m)} p(\mathbf{X}|\mathbf{x}, m)d\mathbf{X} \tag{4}$$

The expression in Equation 4 is a p-value that behaves as we would expect from a classic hypothesis test.

The key difference between Equation 4 and the corresponding expression in Gelman, Meng, et al. (1996, Equation 5 on p. 738) is the order of integration. In Equation 4 every term, namely $p(\mathbf{X}|\mathbf{x}, m)$ as in Equation 2, $S(\mathbf{X}|\boldsymbol{\theta}, m)$ as in Equation 1 and $s(\mathbf{x}|\boldsymbol{\theta}, m)$ as in Equation 3, is first integrated over $\boldsymbol{\theta}$, using the appropriate posterior distribution. Then, importantly, the desired p-value is obtained by integrating over

the predictive posterior distribution for $\mathbf{X}$. Gelman, Meng, et al. (1996) integrate a conditional p-value, namely conditional on $\boldsymbol{\theta}$, over the posterior for $\boldsymbol{\theta}$.

Vital to understanding this difference is understanding that the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, m)$ in Equation 3 is not the same as the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X}, m)$ in Equation 1, which is determined in practice based on replicate samples drawn from Equation 2.

Note that we have placed no restrictions on the model so far, other than being able to simulate from the model itself, given parameter values, and being able to simulate from the posterior distribution of the model parameters.

In Section 3 we will consider a specific model and some of the technicalities that may arise. For example, in many cases the expectation $s(\mathbf{x}|m) = E_{\boldsymbol{\theta}}[s(\mathbf{x}|\boldsymbol{\theta}, m)]$ cannot be derived explicitly, and its empirical calculation may be slow. However, the statistic $s(\mathbf{x}|m)$ can be approximated by: $s(\mathbf{x}|m) = E_{\boldsymbol{\theta}}[s(\mathbf{x}|\boldsymbol{\theta}, m)] \approx s(\mathbf{x}|E_{\boldsymbol{\theta}}[\boldsymbol{\theta}], m) \approx s(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}), m)$, where $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is an estimate of $\boldsymbol{\theta}$ based on the sample. Similarly, $S(\mathbf{X}|m)$ can be approximated as

$$S(\mathbf{X}|m) = E_{\boldsymbol{\theta}}[S(\mathbf{X}|\boldsymbol{\theta}, m)] \approx S(\mathbf{X}|\hat{\boldsymbol{\theta}}(\mathbf{X}), m) \tag{5}$$

When the p-value in Equation 4 is calculated through simulation, approximation 5 can eliminate the need to obtain the posterior distribution ($p(\boldsymbol{\theta}|\mathbf{X}^*, m)$) for each draw $\mathbf{X}^*$ of $\mathbf{X}$, which increases execution speed. Both the efficiency and effectiveness of this approximation can differ dramatically from one model to another. It turns out, however, that in the case of the i.i.d. Gamma model the approximation is particularly useful.

2.2 The parametric bootstrap and posterior predictive check methods

The parametric bootstrap is implemented as follows:

1. Obtain base parameter estimates $\hat{\boldsymbol{\theta}}$.
2. Calculate the base statistic $S(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}), m)$.
3. Draw N new samples $\mathbf{x}_i,\ i = 1, \ldots, N$ from the model $f(\mathbf{X}|\hat{\boldsymbol{\theta}}, m)$ given the parameter estimates from Step 1.
4. Calculate N new statistics $S(\mathbf{x}_i|\hat{\boldsymbol{\theta}}(\mathbf{x}_i), m)$ corresponding to each new sample drawn in Step 3. The parameter estimation procedure must be repeated for each new sample.
5. Calculate the proportion of the new statistics (from Step 4) that exceed the base statistic (from Step 2) and report this result as a p-value.

The posterior predictive check adapts the parametric bootstrap as follows:

1. Replace Step 1 above with: Simulate N^* sets of parameter values ($\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N^*}$) from the posterior distribution given the base sample ($p(\boldsymbol{\theta}|\mathbf{x}, m)$).
2. Repeat Steps 2 to 5 above for each draw $\boldsymbol{\theta}_i$ to obtain N^* p-values.
3. Average these N^* p-values and report this result as a p-value.

2.3 Sketch of the new algorithm

The proposed new test can be implemented through the following simulation algorithm:

1. Derive an objective prior for model m.
2. Given the objective prior derived in Step 1, and given an observed sample $\mathbf{x}$, simulate replicate parameters $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^*$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, m)$ for $\boldsymbol{\theta}$.

3. Calculate the observed test statistic $s(\mathbf{x}|m)$. If $s(\mathbf{x}|m)$ depends on the parameters $\boldsymbol{\theta}$ then, using the replicates $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^*$ simulated in Step 2, calculate $s(\mathbf{x}|m)$ as the average of the statistics $s(\mathbf{x}|\boldsymbol{\theta}_i^*, m),\ i = 1, \ldots, N$ (refer to Equation 3).

4. Simulate replicate samples $\mathbf{X}^*, \ldots, \mathbf{X}^*$ from the posterior predictive distribution $p(\mathbf{X}|\mathbf{x}, m)$ of $\mathbf{X}$. That is, for each replicate parameter $\boldsymbol{\theta}_i^*,\ i = 1, \ldots, N$ from Step 2, simulate a replicate sample X_i^* from the distribution $f(\mathbf{X}|\boldsymbol{\theta}_i^*, m)$ (refer to Equation 2).

5. For each replicate sample $X_i^*, i = 1, \ldots, N$, calculate the test statistic $S_i^* = S(X_i^*|m)$. If $S(X_i^*|m)$ depends on the parameters $\boldsymbol{\theta}$ then do Steps 5a and 5b below:

   (a) Given the objective prior derived in Step 1, and given the replicate sample X_i^* from Step 4, simulate parameters $\boldsymbol{\theta}_{i1}^{**}, \ldots, \boldsymbol{\theta}_{iN}^{**}$ from the posterior distribution $p(\boldsymbol{\theta}|X_i^*, m)$ for $\boldsymbol{\theta}$.

   (b) Calculate $S(X_i^*|m)$ as the average of the statistics $S(X_i^*|\boldsymbol{\theta}_{ij}^{**}, m),\ j = 1, \ldots, N$ (refer to Equation 1).

6. To calculate the p-value, compare the observed test statistic $s(\mathbf{x}|m)$, from Step 3, with its simulated distribution $[S_1^*, \ldots, S_N^*]$ from Step 5 (refer to Equation 4). Explicitly, use a continuity adjustment and calculate the p-value as $P = [count(S_1^*, \ldots, S_N^* > s) + 0.5]/(N + 1)$.

3 Implementation for the Gamma distribution

Consider the following form of the pdf of the Gamma distribution:

$$f(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x > 0 \tag{6}$$

In terms of the notation of the previous section, the parameter vector is $\boldsymbol{\theta} = (\alpha, \lambda)$.

3.1 Objective prior distribution

The maximal data information (MDI) prior (Zellner, 1997, pp. 112–116) is used as an objective prior. To quote Zellner, the MDI prior provides "maximal prior average data information relative to the information in the prior distribution".

Alternative priors such as the Jeffreys prior (Jeffreys, 1998; Yang and J. O. Berger, 1998) can be used but one must be careful of additional restrictions placed on the parameters. The test procedure may malfunction or fail when the parameters estimated from the sample fall inside or near the restricted area of their domain. The MDI prior does not create such restrictions, which further motivates its use.

The MDI prior is defined as $\exp\{E[\log f(x)]\}$, which works out to:

$$\pi(\alpha, \lambda) = \frac{\lambda}{\Gamma(\alpha)} e^{(\alpha-1)\psi(\alpha)-\alpha} \tag{7}$$

where $\psi()$ is the digamma function.

3.2 Posterior distribution

Given the MDI prior (Equation 7) and observations $\mathbf{x}$, the posterior distribution is

$$p(\alpha, \lambda|\mathbf{x}) \propto \lambda^{(n\alpha+2)-1} e^{-\lambda \sum x_i} \Gamma(\alpha)^{-(n+1)} e^{\alpha\left(\sum \log x_i + \psi(\alpha) - 1\right) - \psi(\alpha)} \tag{8}$$

Therefore, $\lambda|\alpha, \mathbf{x} \sim Gamma(n\alpha + 2, \sum x_i)$, and thus

$$p(\lambda|\alpha, \mathbf{x}) = \frac{(\sum x_i)^{n\alpha+2}}{\Gamma(n\alpha + 2)} \lambda^{(n\alpha+2)-1} e^{-\lambda \sum x_i} \tag{9}$$

so that

$$p(\alpha|\mathbf{x}) \propto \Gamma(\alpha)^{-(n+1)} e^{\alpha(\sum \log x_i + \psi(\alpha) - 1) - \psi(\alpha)} \Gamma(n\alpha + 2) \left(\sum x_i\right)^{-(n\alpha+2)} \tag{10}$$

The fastest way to simulate accurately from the posterior distribution (Equation 8) appears to be as follows: First simulate values of α from $p(\alpha|\mathbf{x})$ in Equation 10 and then, given the α values, simulate corresponding values for λ from $p(\lambda|\alpha, \mathbf{x})$ in Equation 9.

3.3 Test statistic

Since the object of our comparison is to compare methods of obtaining the distribution of the test statistic and not to investigate or compare the power of statistics, we will focus only on one statistic going forward. We will use the AD statistic as it is well known and has good power. When testing for Normality, which has been heavily studied, the AD statistic (along with the Shapiro-Wilk statistic) has been shown to have high power against the general alternative (Razali and Wah, 2011). While less comparisons have been done in the case of the Gamma distribution, we refer to Henze et al. (2012) who show that the AD statistic has the highest power among the well-known statistics in the case of the Gamma distribution.

Given an i.i.d. sample $\mathbf{x} = (x_1, \ldots, x_n)$ with CDF $F_X(\boldsymbol{\theta})$, and corresponding order statistics $x_{(1)}, \ldots, x_{(n)}$, the AD statistic A^2 is defined as:

$$A^2 = s(\mathbf{x}|\boldsymbol{\theta}) = -n - \sum_{k=1}^{n} \frac{2k-1}{n} [\log F_X(x_{(k)}|\boldsymbol{\theta}) + \log(1 - F_X(x_{(n+1-k)}|\boldsymbol{\theta}))] \tag{11}$$

Given a specific time constraint, it is possible to achieve higher power for the proposed test using approximation 5, through increased sampling from the predictive posterior distribution. Furthermore, for the sake of speed, parameter estimation is performed throughout using the method of moments.

3.4 Type 1 error and power of proposed tests

We calculate, through simulation, the type 1 error and power of the proposed tests over a large number of samples from various distributions.

3.4.1 Type 1 error

First we consider data from Gamma distributions (null hypothesis is true) and determine to what extent each test works as expected from a classical hypothesis test. Specifically, we expect that if a significance level is chosen as α^* (say) then the test will falsely reject the null hypothesis proportion α^* of the time. As stated above, this is equivalent to the p-value being uniformly distributed.

3.4.2 Power

Second we consider data from alternative distributions. We show that the test will correctly reject the null hypothesis more often than proportion α^* (significance level) in all cases. We discuss the effect that failing to achieve the correct significance level as stated in Section 3.4.1 on power comparisons.

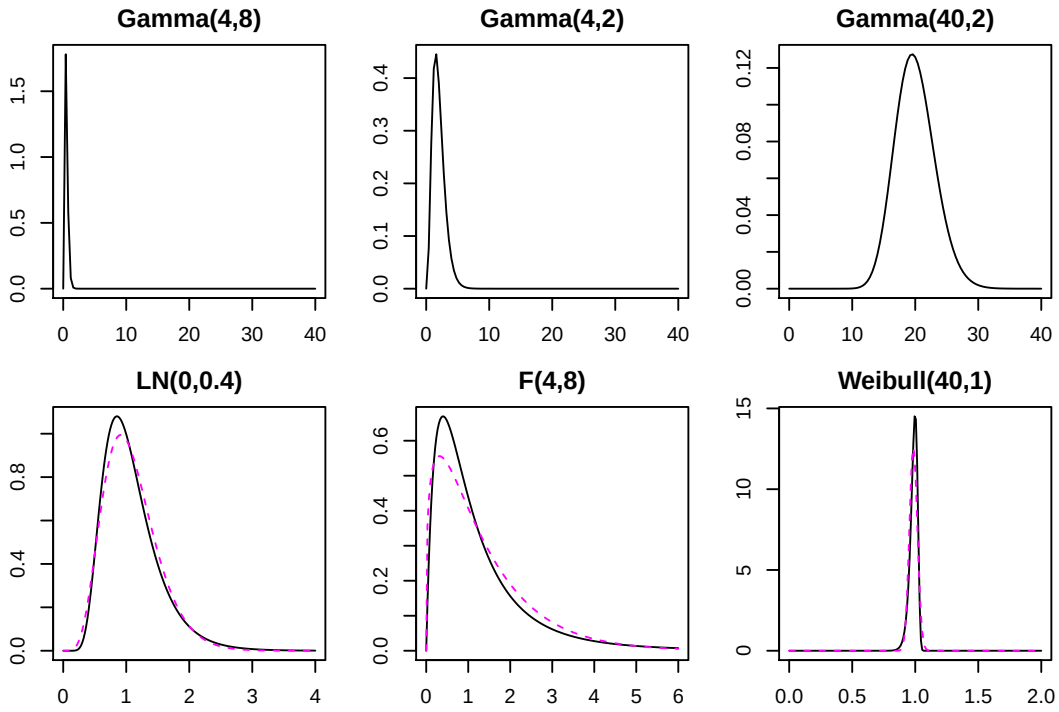

Figure 3: Illustration of various distributions used (solid lines) along with Gamma fit (dotted lines).

3.4.3 Design of simulation study

Only two small sample sizes are used for illustration: 12 and 24. This is to highlight the fact that our proposed approach (Section 2) is non-asymptotic. For each combination of sample size and distribution, 112,000 samples are simulated and the test is performed on each sample independently. The p-values are recorded and summarised in the form of rejection rates.

The distributions used under the null hypothesis are Gamma(4, 8), Gamma(4, 2) and Gamma(40, 2). The distributions used under the alternative hypothesis are the Log-Normal(0, 0.4), F(4, 8) and Weibull(40, 1). All these distributions are illustrated in Figure 3 along with the fitted Gamma approximations. The Gamma fit is to highlight the extent to which the alternative distribution differs from the Gamma distribution; note that the Log-Normal distribution is quite close to the best-fitting Gamma distribution.

3.4.4 Results

The results of the simulation study are summarized in Table 1. Clearly there are no significant differences between the parametric bootstrap method (Section 2.2) and the Bayes method (Section 2.3) as the minor discrepancies at the fourth decimal are all less than one standard deviation under the null hypothesis of Uniform p-values.

The reason for the lack of discrepancy is most likely because of the low posterior variance (or accuracy of the parameter estimation). It is for this reason that we now go on to consider a case where the posterior variance is much larger, namely the GPD.

4 Implementation for the Generalised Pareto Distribution

The Generalised Pareto Distribution is used to model the tail (extreme values) of a distribution beyond a given threshold. If the threshold is known we can subtract it from all the observations and consider it to be zero. This makes the GPD a 2-parameter distribution.

Distribution and Method	Sample Size 12	Sample Size 24
G(4,8) Bayes	0.0493	0.0492
G(4,8) ParBoot	0.0496	0.0495
G(4,2) Bayes	0.0502	0.0513
G(4,2) ParBoot	0.0500	0.0513
G(40,2) Bayes	0.0498	0.0494
G(40,2) ParBoot	0.0497	0.0492
LN(0,0.4) Bayes	0.0666	0.0862
LN(0,0.4) ParBoot	0.0666	0.0865
F(4,8) Bayes	0.0971	0.1572
F(4,8) ParBoot	0.0976	0.1576
W(40,1) Bayes	0.1560	0.3028
W(40,1) ParBoot	0.1562	0.3026

Table 1: Rejection rates at $\alpha = 5\%$ for Gamma tests for different sampling distributions and sample sizes

$$f(x|\gamma,\sigma) = \frac{1}{\sigma}\left[1+\frac{\gamma x}{\sigma}\right]^{-\frac{1}{\gamma}-1}, x < -\frac{\sigma}{\gamma} \tag{12}$$

We investigate the differences between the parametric bootstrap approach and the Bayes approach with respect to testing the hypothesis that a sample consists of independent GPD observations above a known threshold. The implementation proceeds in the same order as for the Gamma distribution, with only minor differences highlighted in Section 4.2.

4.1 Posterior distribution

We will simulate from this posterior using the Metropolis-Hastings algorithm with proposal $\gamma_c \sim N(\gamma_j, 0.05^2)$ and $\log\sigma_c \sim N(\log\sigma_j, 0.1^2)$. See Robert and Casella (2004, pp. 267–301) for an in-depth general discussion of this algorithm.

4.2 Problems with Maximum Likelihood estimation

In all cases we calculate the AD statistic for each replicate sample using the Maximum Likelihood (ML) method as implemented in the evir package in R (Pfaff and McNeil, 2012). This approach has the drawback that roughly 0.35% of the time the parameter estimation fails. In these cases we consider the statistic as missing and ignore it for p-value calculations.

As far as each original simulated sample is concerned, where ML fails we consider the p-value missing for the parametric bootstrap approach but calculate it using the posterior mean in the Bayes approach. We compared these results with the results from dropping these cases entirely and noticed no difference; thus, we can safely assume that this problem has no impact on the outcome of the experiment.

4.3 Design of simulation study

In this case only one sample size was used, namely 24. We chose this sample size to illustrate that the new approach is of most value for smaller samples.

Null distributions considered are the $GPD(\gamma = 0.25, \sigma = 1)$ and $GPD(\gamma = -0.10, \sigma = 1)$. Alternative distributions considered are the $Gamma(\alpha = 0.5, \lambda = 1)$ and $Lognormal(\mu = 1, \sigma = 1)$. These are illustrated in Figure 4.

Again we used only the well known AD statistic (Equation 11).

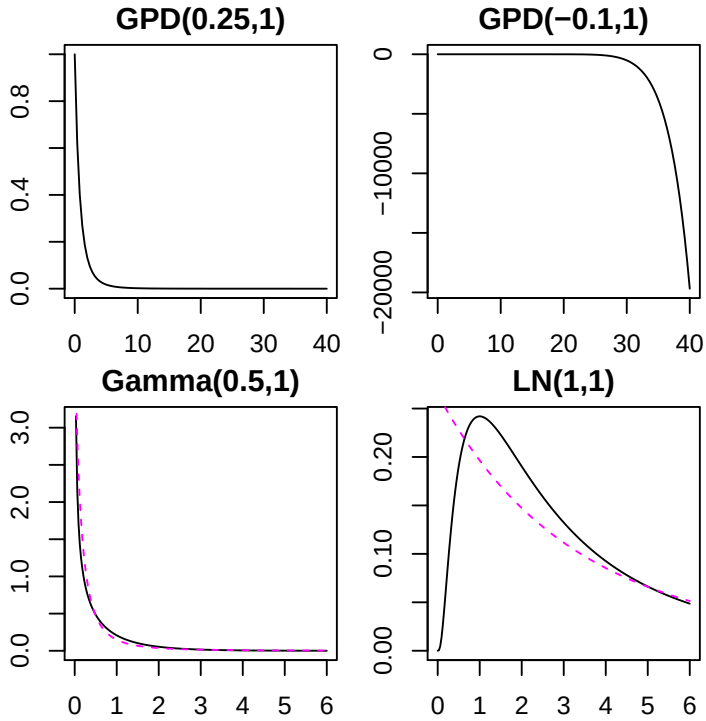

Figure 4: Illustration of various distributions used (solid lines) along with GPD fit (dotted lines).

Distribution and Method	Sample Size 24
GPD(0.25,1) Bayes	0.0432
GPD(0.25,1) ParBoot	0.0316
GPD(-0.1,1) Bayes	0.0385
GPD(-0.1,1) ParBoot	0.0220
Gamma(0.5,1) Bayes	0.4753
Gamma(0.5,1) ParBoot	0.4658
LN(1,1) Bayes	0.1305
LN(1,1) ParBoot	0.1134

Table 2: Rejection rates at $\alpha = 5\%$ for GPD tests for different sampling distributions

4.3.1 Results

The results of the simulation study are summarized in Table 2. Here the difference is marked in that the Bayes method (Section 2.3) comes much closer to achieving the desired significance level. This can be seen even more clearly in Figure 5.

5 Discussion

The differences between the observed results of the Gamma experiment and the GPD experiment may be because in the case of the Gamma distribution (especially with large values of the first parameter) the parameter estimation is relatively accurate and straightforward, in stark contrast with the GPD, for which parameter estimation is an open research topic.

The remaining departures of the p-value distribution from the Uniform might be associated with the choice of prior distribution or imperfect posterior simulation. Ideally one would derive a prior distribution such that the test produces perfectly Uniform p-values, but this does not seem mathematically tractable.

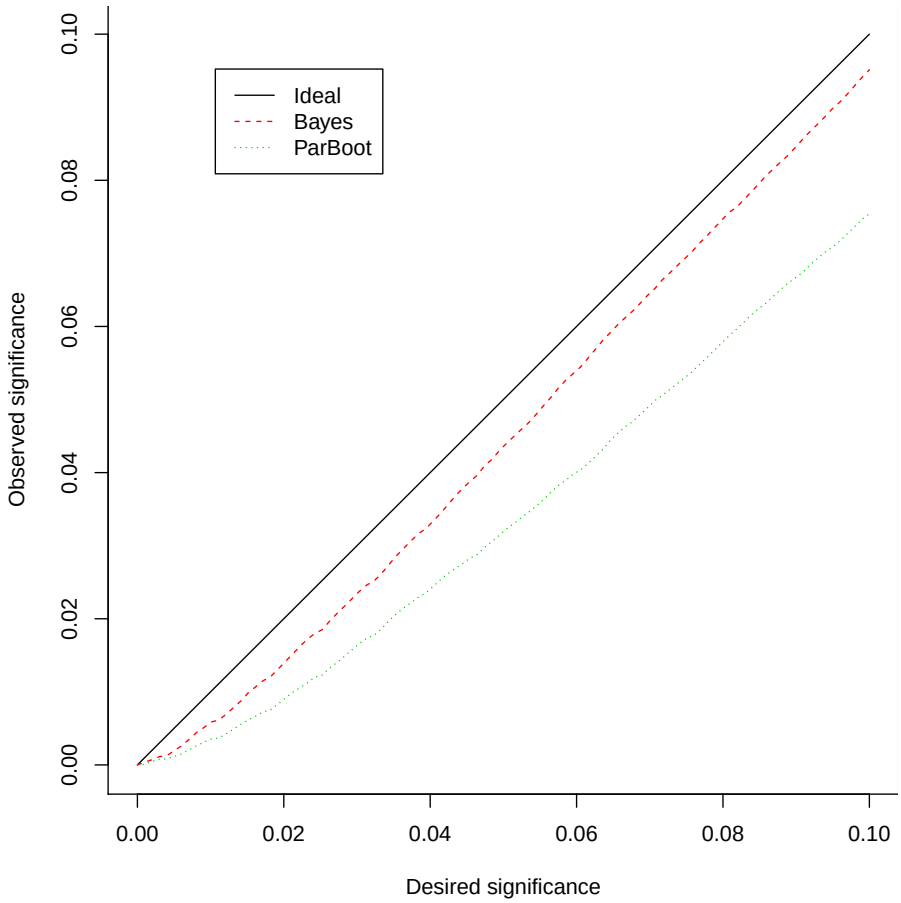

Figure 5: Observed coverage of 112,000 GPD tests using AD statistic and sample size 24. Solid line is perfect coverage, dashed line is the Bayes approach and dotted line at the bottom is the parametric bootstrap.

6 Conclusion

It is clear from the results of the simulations that the new test constructed in this paper performs very well and helps address the problem of testing for a distribution with unknown parameters. The new test procedure comes closer to achieving the correct size in a classical hypothesis testing framework, especially for small samples. Furthermore, the power of the test may be higher that what is was in the parametric bootstrap framework and will never be lower. The key difference between the test presented here and previous work is that the Bayesian adaptation works better when faced with problems where the parameter estimation is difficult and carries much uncertainty. The classic approach injects certainty where there is none and this can create false confidence in a chosen model.

Acknowledgements

The author wishes to thank Profs Schall, van der Merwe and De Waal as well as Dr van Zyl for asking all the right questions.

References

Bayarri, M. J. and Berger, James O. (2000), "P Values for Composite Null Models", *Journal of the American Statistical Association* 95.452, pp. 1127–1142, DOI: 10.1080/01621459.2000.10474309, eprint: http://dx.doi.org/10.1080/01621459.2000.10474309, URL: http://dx.doi.org/10.1080/01621459.2000.10474309.

Bellosta, Carlos J. Gil (2011), *ADGofTest: Anderson-Darling GoF test*, R package version 0.3, URL: http://CRAN.R-project.org/package=ADGofTest.

Beran, Rudolf (1988), "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements", *Journal of the American Statistical Association* 83.403, pp. 687–697, DOI: 10.1080/01621459.1988.10478649, eprint: http://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478649, URL: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1988.10478649.

Crespi, Catherine M. and Boscardin, W. John (2009), "Bayesian model checking for multivariate outcome data", *Computational Statistics & Data Analysis* 53.11, pp. 3765–3772, ISSN: 0167-9473, DOI: http://dx.doi.org/10.1016/j.csda.2009.03.024, URL: http://www.sciencedirect.com/science/article/pii/S0167947309001339.

D'Agostino, R. B. and Stephens, M. A. (1986), *Goodness-of-Fit Techniques*, vol. 68, Statistics: textbooks and monographs, Marcel Dekker Inc., ISBN: 0-8247-7487-6.

Darling, D. A. (1957), "The Kolmogorov-Smirnov, Cramer-von Mises Tests", English, *The Annals of Mathematical Statistics* 28.4, pages, ISSN: 00034851, URL: http://www.jstor.org/stable/2237048.

Gelman, Andrew, Goegebeur, Y., Tuerlinckx, F., and Van Mechelen, I. (2000), "Diagnostic checks for discrete data regression models using posterior predictive simulations", *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49.2, pp. 247–268, ISSN: 1467-9876, DOI: 10.1111/1467-9876.00190, URL: http://dx.doi.org/10.1111/1467-9876.00190.

Gelman, Andrew, Meng, Xiao-Li, and Stern, Hal (1996), "Posterior predictive assessment of model fitness via realized discrepancies", *Statistica sinica* 6.4, pp. 733–760.

Henze, Norbert, Meintanis, Simos G., and Ebner, Bruno (2012), "Goodness-of-Fit Tests for the Gamma Distribution Based on the Empirical Laplace Transform", *Communications in Statistics - Theory and Methods* 41.9, pp. 1543–1556, DOI: 10.1080/03610926.2010.542851, eprint: http://dx.doi.org/10.1080/03610926.2010.542851, URL: http://dx.doi.org/10.1080/03610926.2010.542851.

Jeffreys, H. (1998), *The Theory of Probability*, 3rd ed., OUP Oxford, ISBN: 9780191589676, URL: http://books.google.co.za/books?id=vh9Act9rtzQC.

Marsaglia, George and Marsaglia, John (2004), "Evaluating the anderson-darling distribution", *Journal of Statistical Software* 9.2, pp. 1–5.

Pfaff, Bernhard and McNeil, Alexander (2012), *evir: Extreme Values in R*, R package version 1.7-3, URL: http://CRAN.R-project.org/package=evir.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org/.

Razali, Nornadiah Mohd and Wah, Yap Bee (2011), "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests", *Journal of Statistical Modeling and Analytics* 2.1, pp. 21–33.

Robert, Christian and Casella, George (2004), *Monte Carlo Statistical Methods*, 2nd ed., Springer, ISBN: 978-0387212395.

Robins, James M, Vaart, Aad van der, and Ventura, Valérie (2000), "Asymptotic distribution of P values in composite null models", *Journal of the American Statistical Association* 95.452, pp. 1143–1156.

Rubin, Donald B et al. (1984), "Bayesianly justifiable and relevant frequency calculations for the applies statistician", *The Annals of Statistics* 12.4, pp. 1151–1172.

Sinharay, Sandip and Stern, Hal S (2003), "Posterior predictive model checking in hierarchical models", *Journal of Statistical Planning and Inference* 111.1, pp. 209–221.

Yang, R. and Berger, J. O. (1998), *A Catalog of Noninformative Priors*, Technical Report, stats.org.uk, URL: http://www.stats.org.uk/priors/noninformative/YangBerger1998.pdf.

Yano, Yoshitaka, Beal, Stuart L, and Sheiner, Lewis B (2001), "Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check", *Journal of pharmacokinetics and pharmacodynamics* 28.2, pp. 171–192.

Zellner, A. (1997), *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, Economists of the Twentieth Century Series, Edward Elgar Pub, ISBN: 9781858982205, URL: http://books.google.co.za/books?id=ICW7AAAAIAAJ.