Statistical analysis 1

L. Guillemot

Max-Planck-Institut für Radioastronomie, Bonn, Germany

11 October 2011

Outline

Let $(x_1, x_2, \ldots, x_n)$ be a set of n independent realizations of a random variable taking values on a circle C.

How to identify periodicity in this dataset?

One can apply a **test for uniformity on the circle**. If we note H_0 the *null hypothesis* (*i.e.*, absence of a periodic signal), we want to test the *alternative hypothesis* H_A of presence of a periodic signal against H_0, *i.e.*:

$$H_0 \colon p = 0 \text{ against } H_A \colon p > 0$$

where p denotes the signal strength.

Introduction

Several tests for uniformity on the circle exist and are adapted to different situations. Ideally, a good test for uniformity should however:

- be independent of any smoothing parameter (number of bins, number of harmonics, etc.).
- be invariant by rotation.
- be sensitive over a wide range of realistic signal shapes.
- be *consistent*, *i.e.*, for a given signal shape the significance should improve as more data are added.
- etc.

In this presentation we review some commonly-used tests for uniformity on the circle. **Not a comprehensive list though!**

Pearson's χ^2 test

A very commonly-used test for uniformity on the circle is the χ^2 test. The χ^2 statistic is defined by:

$$\chi^2 = \sum_{i=1}^{K} \frac{(X_i - n_i)^2}{n_i}$$

Where:

- K is the total number of data bins,
- X_i is the number of events observed in the i^{th} bin,
- n_i is the number expected according to a known distribution. In the null hypothesis H_0, n_i is the average number of events in each bin, *i.e.*, $n_i = n/K$ where n is the total number of data points.

χ^2 test (continued)

If the n_i are large enough ($n_i \gtrsim 5$), then this statistic follows the χ^2 probability density function (PDF) with the number of degrees of freedom equal to $n_d = K - n_f$, where n_f is the number of fitted parameters (only 1 in our case). The probability value for the hypothesis H_A is:

$$P\left(\chi^2 |\, n_d\right) = \int_{\chi^2}^{\infty} \frac{1}{2^{n_d/2} \times \Gamma(n_d/2)} x^{n_d/2-1} e^{-x/2} dx$$

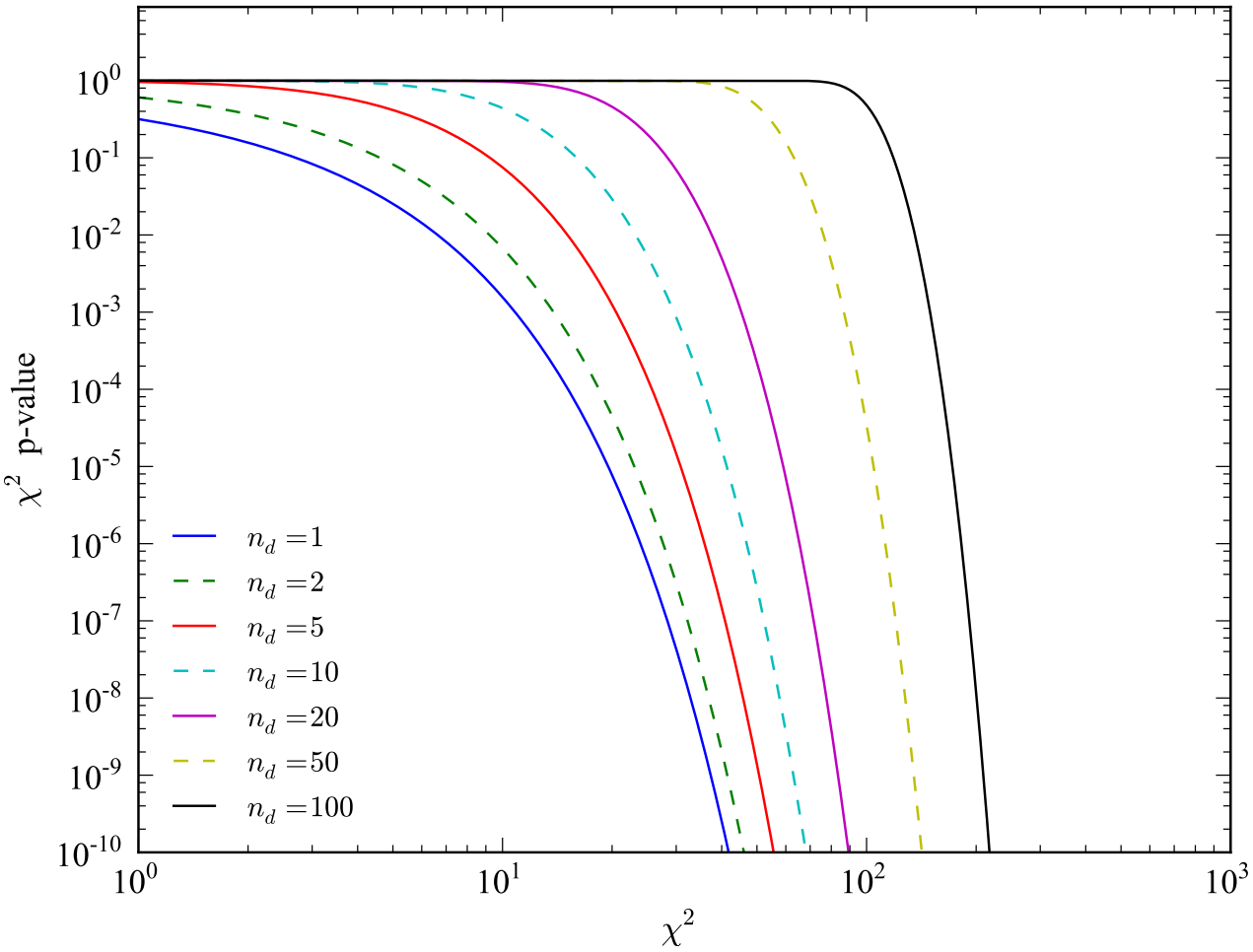
Since $n_d = K - 1$, we have:

$$P(\chi^2) = \int_{\chi^2}^{\infty} \frac{1}{2^{(K-1)/2} \times \Gamma\left(\frac{K-1}{2}\right)} x^{\frac{K-1}{2}-1} e^{-x/2} dx$$

Caveats: the results are sensitive to how the binning is performed (K parameter), they depend on the choice of the origin, and the bin contents have to be large enough ($\gtrsim 5$).

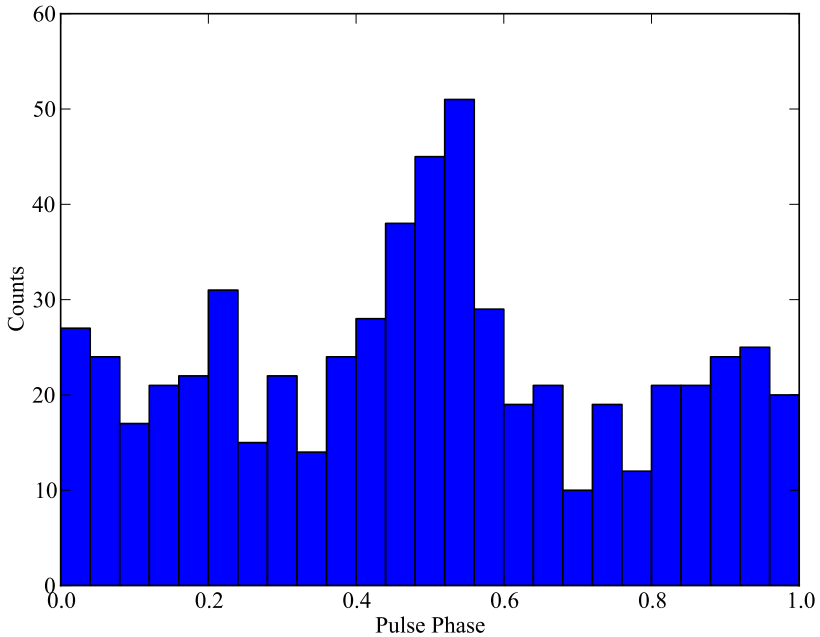
χ^2 test (continued)

χ^2 p-value as a function of χ^2, for different values of the number of degrees of freedom, n_d.

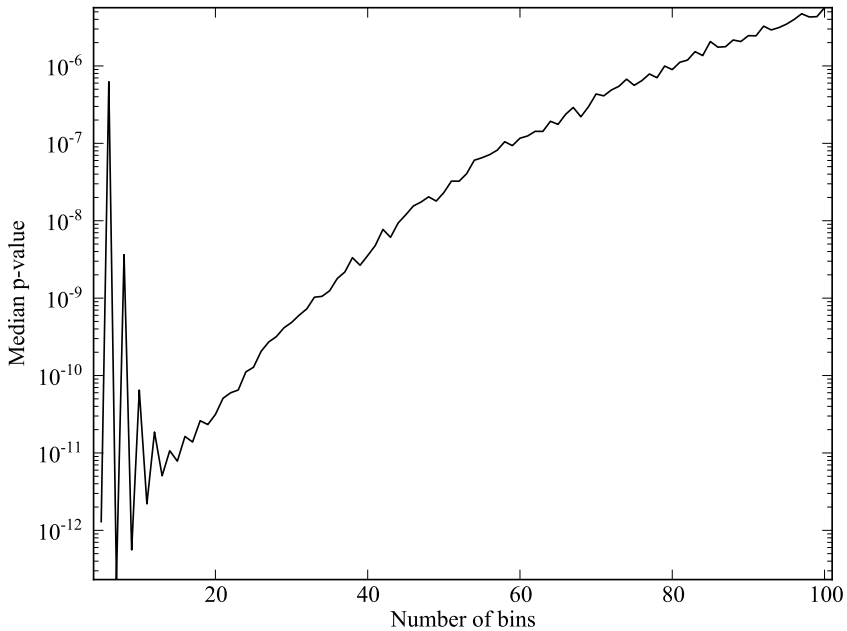


χ^2 test example

100 events were simulated based on a Gaussian distribution with mean 0.5 and standard deviation 0.1, in addition to 500 background events. The simulation was repeated 1000 times, and for each realization the p-value was computed, using different values of the number of bins, K.



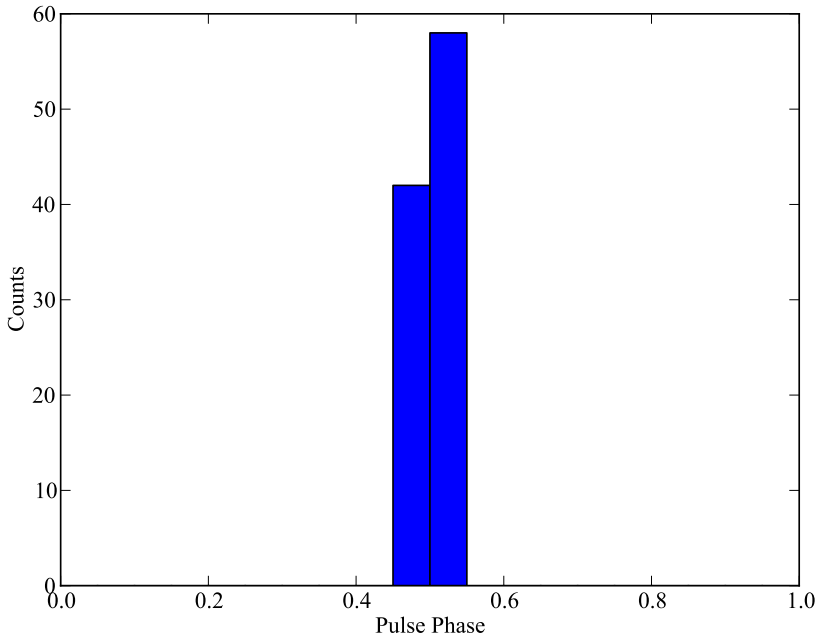
Example of simulated pulse profile.



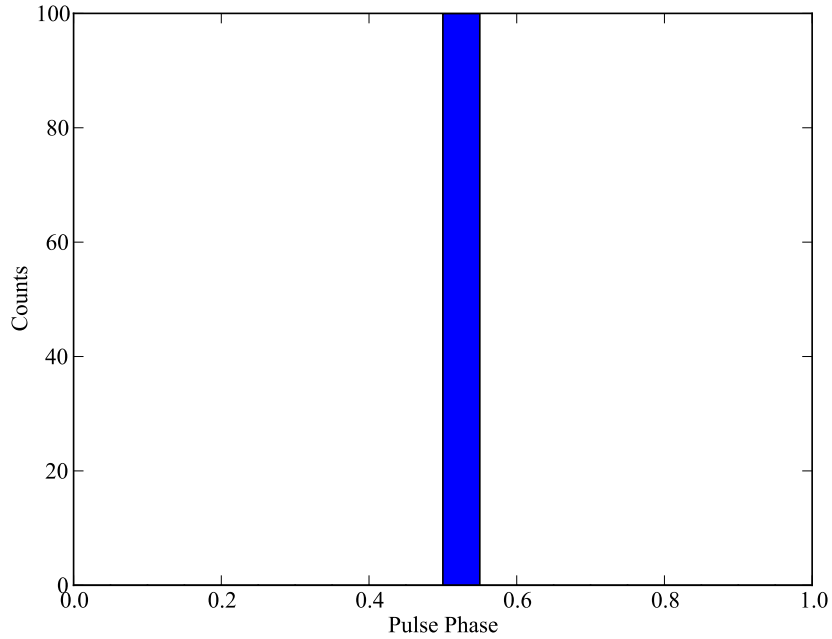
Median p-value, as a function of K.

χ^2 test example

100 events were simulated based on a Gaussian distribution with mean 0.5 and standard deviation 0.01. The χ^2 test statistic is computed using 20 bins.



$\chi^2 = 965.8$

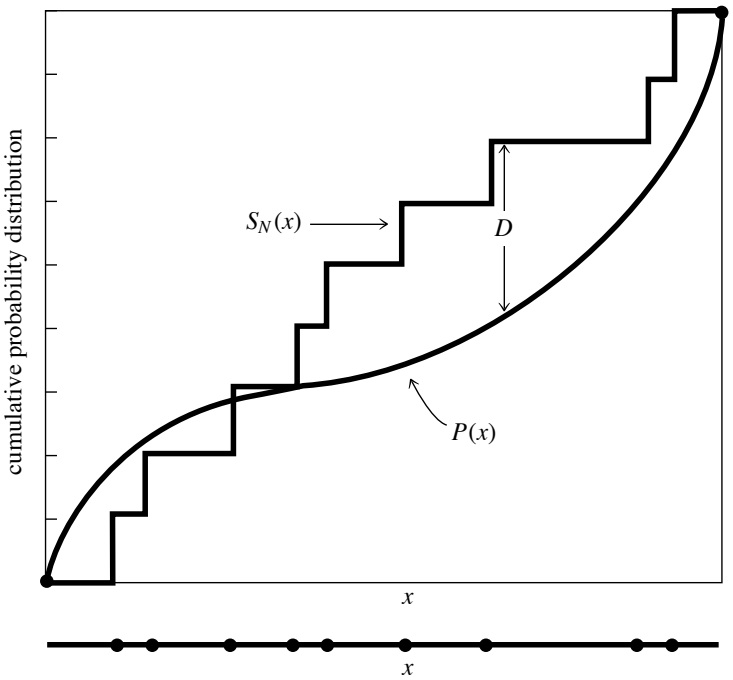


Data points shifted by half a bin. $\chi^2 = 1900$!

The calculated χ^2 values are very sensitive to how the binning is performed.

Kolmogorov-Smirnov test

The *Kolmogorov-Smirnov* statistic D is the maximum value of the absolute distance between two cumulative distribution functions (CDFs).



Let $S_n(x)$ be the CDF of a list of data points x_i, and $P(x)$ the CDF of a known distribution. The KS statistic is:

$$D = \max_{-\infty<x<\infty} |S_n(x) - P(x)|$$

In the case of an uniformly-distributed variable, we have $P(x) = x$.

KS test (continued)

It can be shown that the probability value for a given D is given by:

$$P(D) \simeq Q_{KS}\left(\left[\sqrt{n} + 0.12 + 0.11/\sqrt{n}\right] \times D\right)$$

where:

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty} (-1)^{j-1}\, e^{-2j^2\lambda^2}$$

with the limiting values $Q_{KS}(0) = 1$ and $Q_{KS}(\infty) = 0$. Note that the approximation for $P(D)$ is accurate for $n \geq 4$ only.

Advantage: no smoothing parameter required.

Caveats: the sensitivity is not independent of x! The KS test is most sensitive around the median value and less sensitive at the ends of the distribution.

Variants on the KS test : Kuiper's statistic

An example of variant on the KS test is the Kuiper statistic V, defined by:

$$V = D_+ + D_- = \max_{-\infty<x<\infty} [S_n(x) - P(x)] + \max_{-\infty<x<\infty} [P(x) - S_n(x)]$$

V is the sum of the maximum distance of $S_n(x)$ above and below $P(x)$. The significance level is given by:

$$P(V) = Q_{KP}\left(\left[\sqrt{n} + 0.155 + 0.24/\sqrt{n}\right] \times V\right)$$

where Q_{KP} is defined by:

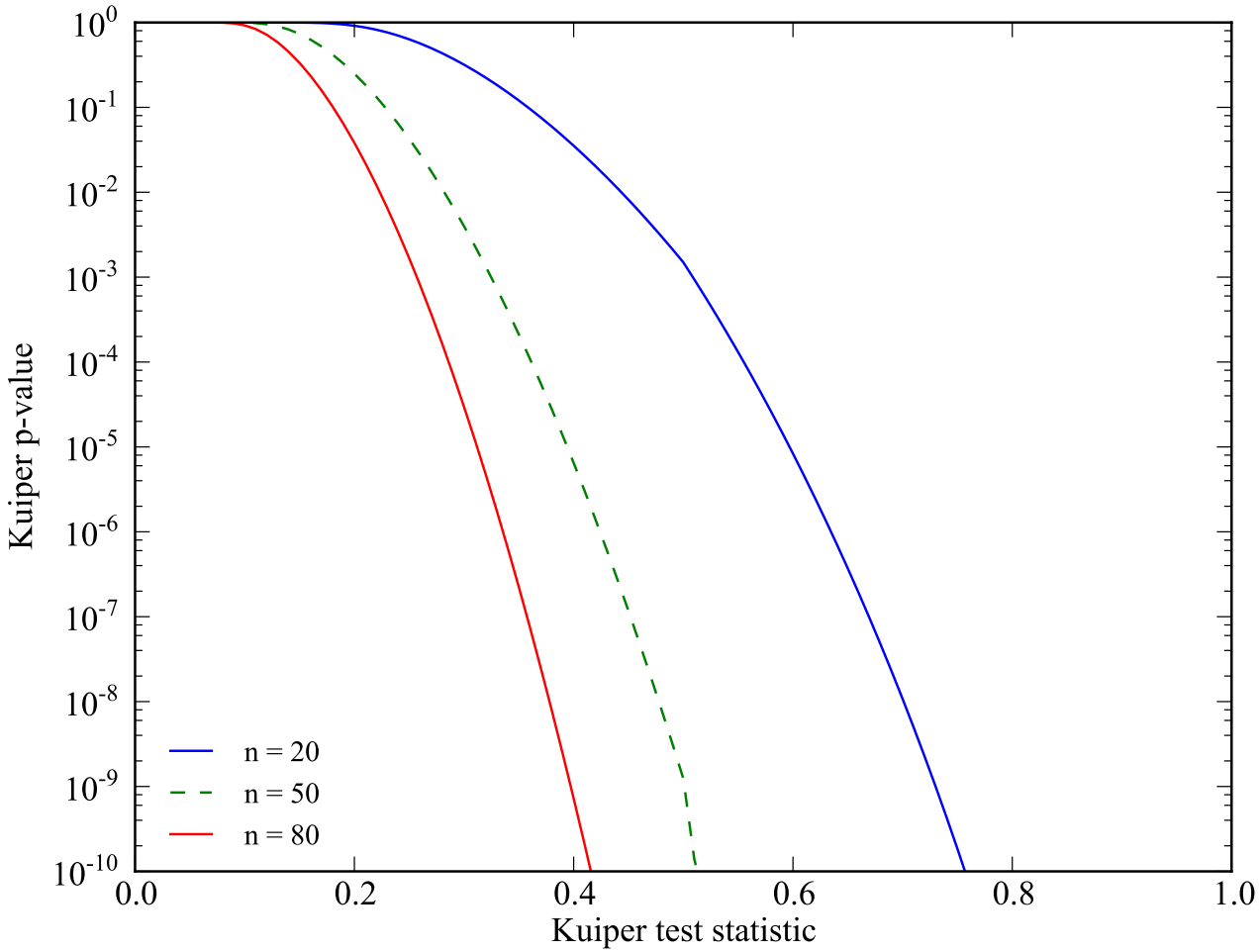
$$Q_{KP}(\lambda) = 2\sum_{j=1}^{\infty}\left(4j^2\lambda^2 - 1\right)e^{-2j^2\lambda^2}$$

with the limiting values $Q_{KP}(0) = 1$ and $Q_{KP}(\infty) = 0$.

Advantages: no smoothing parameter required, equal sensitivities at all values of x.

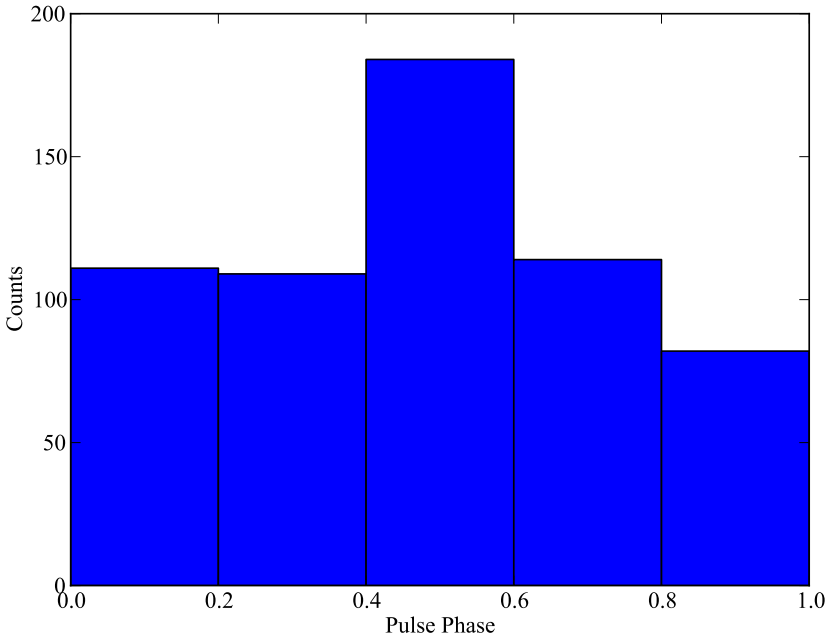
Kuiper test (continued)

Kuiper p-value as a function of V, for different numbers of data points.

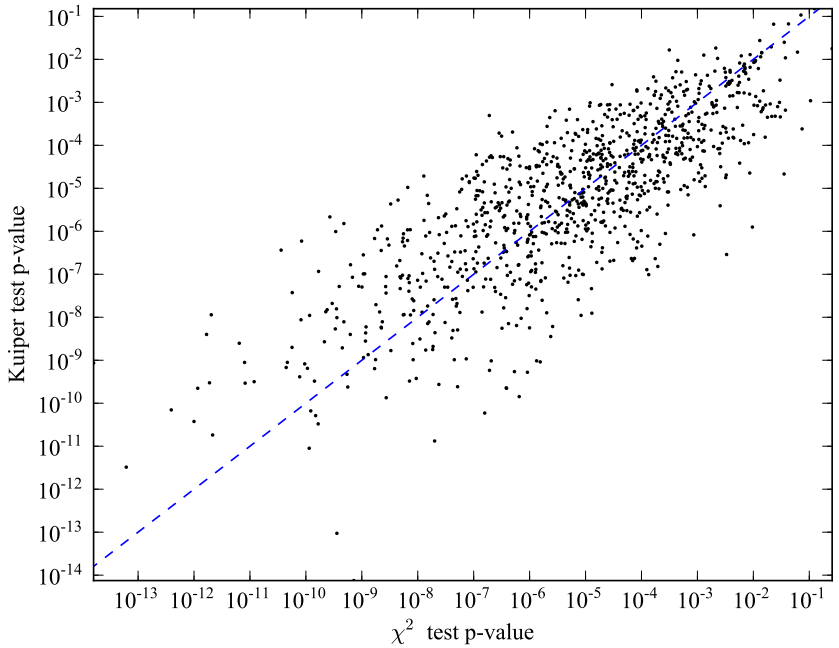


Kuiper test example

100 events were simulated based on a Gaussian distribution with mean 0.5 and standard deviation 0.1, in addition to 500 background events. The simulation was repeated 1000 times (5 bins were used for calculating the χ^2 test statistic).



Example of simulated pulse profile.



Kuiper test p-value vs. χ^2 test p-value.

Beran statistics

Let x be a variable taking value on $[0; 2\pi)$, $f_s(x)$ represent the intensity of the periodic signal in the absence of noise, and p the signal strength (where the case $p = 0$ represents pure noise, and $p = 1$ corresponds to no noise). The observed signal can be written as:

$$f(x) = p \times f_s(x) + \frac{1-p}{2\pi}$$

A measure of the distance between $f(x)$ and the uniform density $g(x) = \frac{1}{2\pi}$ is given by the Beran statistic (Beran, 1969):

$$\psi(f) = \int_0^{2\pi} \left(f(x) - \frac{1}{2\pi} \right)^2 dx = p^2 \int_0^{2\pi} \left(f_s(x) - \frac{1}{2\pi} \right)^2 dx$$

We can see from the above expression that $\psi(f) \to 0$ when $p \to 0$ and $\psi(f) = 0$ if $f_s = 1/2\pi$. The null hypothesis (*i.e.*, uniformity hypothesis) is therefore rejected if $\psi(f)$ is large.

Beran statistics (continued)

Beran (1969) showed that a locally (small p value) most powerful invariant test is to reject H_0 when $\psi(f)$ is large.

However, $f(x)$ is unknown and therefore $\psi(f)$ is unknown! How to define $f(x)$? One can replace f by an estimator $\hat{f}$ and use $\psi(\hat{f})$ as a test statistic.

It can be shown that under certain hypotheses and within constants, the χ^2 and the Kuiper statistics can be derived from ψ.

Other examples are the Rayleigh test (Rayleigh, 1919), the Watson U^2 test (Watson, 1961), the Ajne test (Ajne, 1968), and the Z_m^2 test (Buccheri, 1983).

Rayleigh test

Any density function on a circle can be written as a Fourier series (see e.g. Mardia, 1972). In the case of discrete data, we can define:

$$\hat{f}_m(x) = \frac{1}{2\pi}\left[1 + 2\sum_{k=1}^{m}\left(\hat{\alpha}_k \cos(kx) + \hat{\beta}_k \sin(kx)\right)\right]$$

with the empirical trigonometric moments:

$$\begin{aligned}\hat{\alpha}_k &= \frac{1}{n}\sum_{i=1}^{n}\cos(kx_i)\\ \hat{\beta}_k &= \frac{1}{n}\sum_{i=1}^{n}\sin(kx_i)\end{aligned}$$

Rayleigh test (continued)

Limiting the functional $\hat{f}_m$ to the first harmonic, we have:

$$\hat{f}_1(x) = \frac{1}{2\pi}\left[1 + \hat{\alpha}_1 \cos(x) + \hat{\beta}_1 \sin(x)\right]$$

and therefore:

$$\psi(\hat{f}_1) = \int_0^{2\pi} \left(\hat{f}_1(x) - \frac{1}{2\pi}\right)^2 dx = \frac{1}{\pi}\left(\hat{\alpha}_1^2 + \hat{\beta}_1^2\right) = \frac{\hat{L}_1^2}{\pi}$$

where $\hat{L}_1$ is the length of the 1^{st} harmonic, given by $\hat{L}_1 = \sqrt{\hat{\alpha}_1^2 + \hat{\beta}_1^2}$.
The quantity $\hat{L}_1^2$ can therefore be used as an unbiased estimator. The Rayleigh test statistic is given by:

$$2nR^2 = 2n\hat{L}_1^2 = \frac{2}{n}\left[\left(\sum_{i=1}^{n}\cos(x_i)\right)^2 + \left(\sum_{i=1}^{n}\sin(x_i)\right)^2\right]$$

Rayleigh test (continued)

Provided $n > 100$, the probability distribution function of the Rayleigh test statistic matches that of the χ^2, with 2 degrees of freedom. In that case the probability of H_0 being true is given by:

$$P(2nR^2) = e^{-nR^2}$$

If $n \leq 100$, the probability level needs to be computed using the result of Greenwood and Durand (1955). If we note $K = nR^2$, we have:

$$\begin{aligned} P(2nR^2) \quad = \quad & e^{-K} \times [1 + (2K - K^2)/4n - (24K - 132K^2 + 76K^3 - 9K^4)/288n^2 \\ & -(1440K + 1440K^2 - 8280K^3 + 4890K^4 - 870K^5 + 45K^6)/17280n^3] \end{aligned}$$

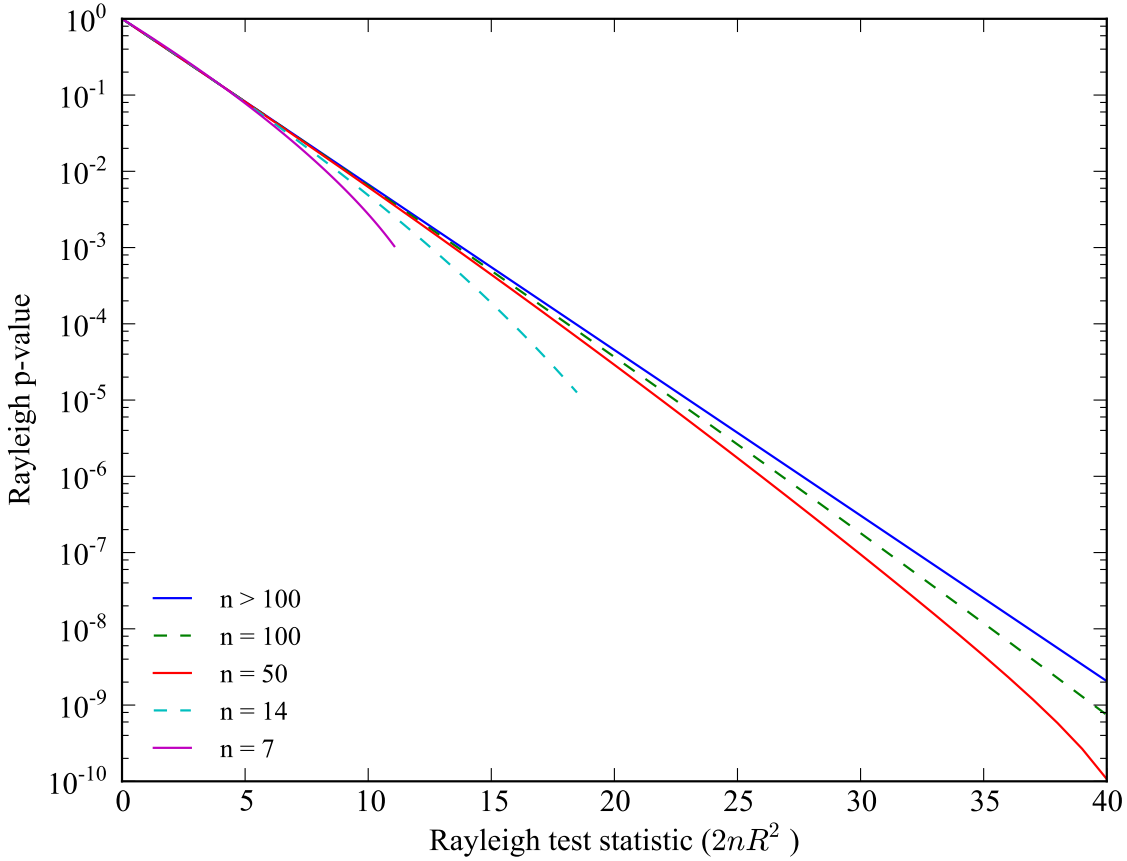
Note that the above formula is valid down to 10^{-5} for $n \lesssim 14$, and down to 10^{-3} for $n \lesssim 7$.

Advantages: unlike the χ^2 test, the Rayleigh test does not depend on any smoothing parameter and is rotation-invariant. Also, it is very sensitive to broad, sinusoidal profiles.

Caveat: in the case where two or more peaks are expected, the respective mean vectors may cancel each other, so that the distribution of R will be close to that under uniform density.

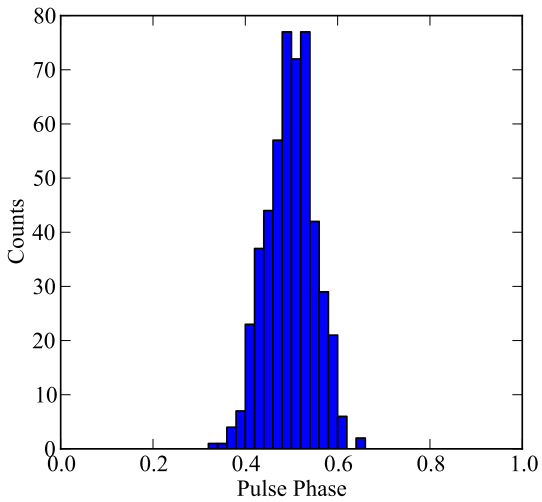
Rayleigh test (continued)

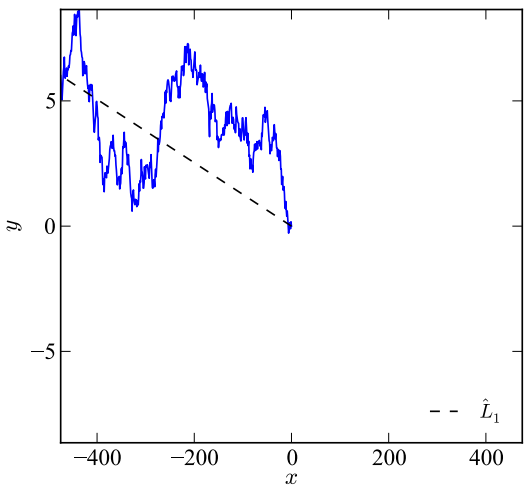
Rayleigh test p-value as a function of $2nR^2$, for different numbers of data points.

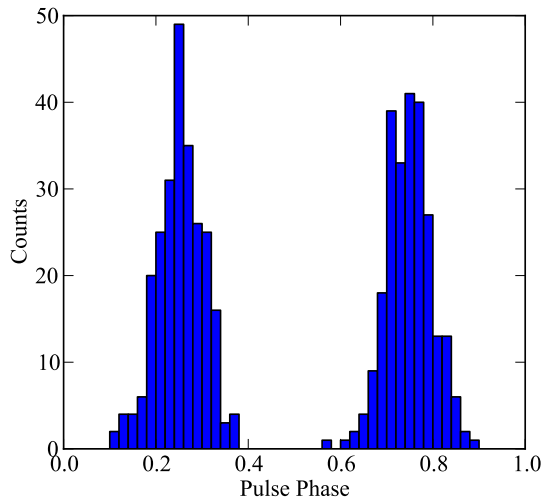


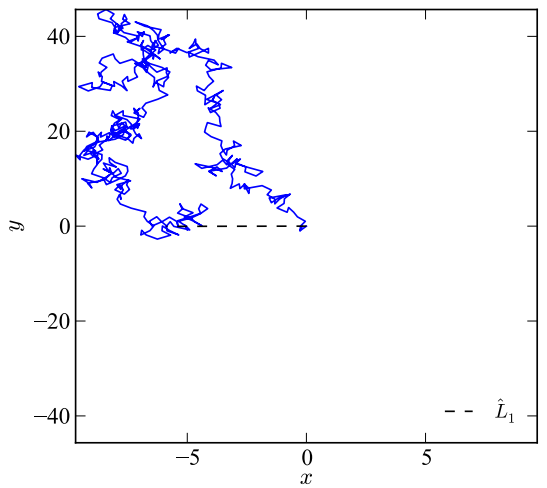
Rayleigh test (continued)

Illustration of the loss of sensitivity of the Rayleigh statistic for bimodal profiles.









Z_m^2 test

Extending the development of the Fourier series to harmonics m higher than 1, *i.e.*:

$$\hat{f}_m(x) = \frac{1}{2\pi}\left[1 + 2\sum_{k=1}^{m}\left(\hat{\alpha}_k \cos(kx) + \hat{\beta}_k \sin(kx)\right)\right]$$

we have:

$$\psi(\hat{f}_m) = \int_0^{2\pi}\left(\hat{f}_m(x) - \frac{1}{2\pi}\right)^2 dx = \frac{1}{\pi}\sum_{k=1}^{m}\left(\hat{\alpha}_k^2 + \hat{\beta}_k^2\right) = \frac{1}{\pi}\sum_{k=1}^{m}\hat{L}_k^2$$

where $\hat{L}_k$ is the length of the k^{th} harmonic.

Z_m^2 test (continued)

The Z_m^2 statistic, introduced by Buccheri et al. (1983), is defined by:

$$\begin{aligned} Z_m^2 = 2\pi n \psi(\hat{f}_m) &= 2n \sum_{k=1}^{m} \left(\hat{\alpha}_k^2 + \hat{\beta}_k^2 \right) \\ &= \frac{2}{n} \sum_{k=1}^{m} \left[\left(\sum_{i=1}^{n} \cos(k x_i) \right)^2 + \left(\sum_{i=1}^{n} \sin(k x_i) \right)^2 \right] \end{aligned}$$

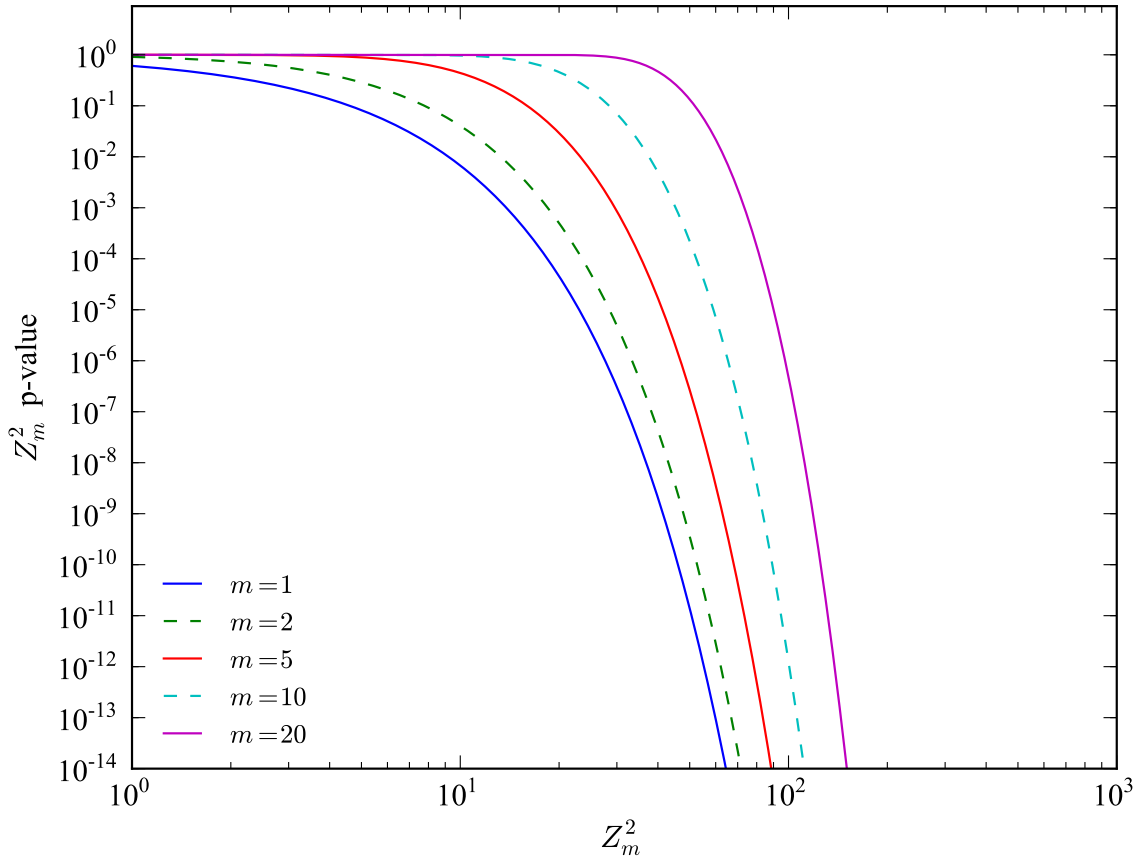
It is clear that in the case $m = 1$, Z_1^2 corresponds to the Rayleigh statistic.

Advantages: like the Rayleigh statistic, the Z_m^2 statistic is rotation-invariant. However, it is sensitive to a wider range of profile shapes.

Caveat: the optimal number of harmonics m to be selected depends on the profile shape and is therefore not known *a priori*!

Z_m^2 test (continued)

If $n \gtrsim 100$, the probability distribution function of the Z_m^2 statistic matches that of the χ^2, with $2m$ degrees of freedom.



H-test

The *H*-test, introduced by de Jager et al. (1989), uses the Z_m^2 statistic as basic and ameliorates the difficulty of determining the optimal number of harmonics m by using Hart's rule (1985), which calculates the value of m minimizing an estimator of the mean integrated squared-error (MISE) between $\hat{f}_m(x)$ and the true unknown light curve, $f(x)$:

$$\mathrm{MISE}(m) = E \int_0^{2\pi} \left(\hat{f}_m(x) - f(x) \right)^2 dx$$

Following the procedure defined by Hart's rule, de Jager et al. showed that the optimal number of harmonics, M, is defined by:

$$\max_{1 \leq m < \infty} \left(Z_m^2 - 4m + 4 \right) = Z_M^2 - 4M + 4 \geq 0$$

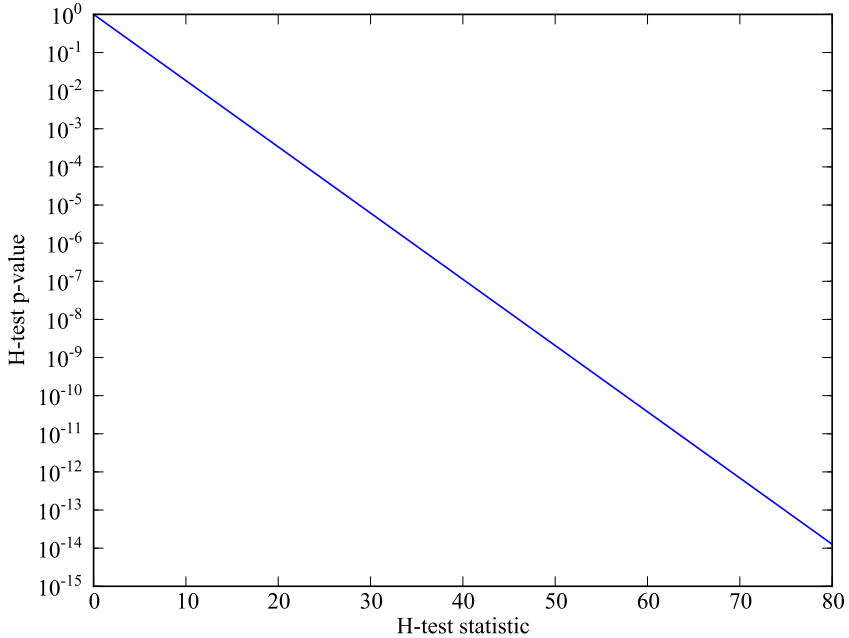
For practical reasons, they recommended truncating after 20 harmonics, and defined the H statistic as:

$$H = \max_{1 \leq m \leq 20} \left(Z_m^2 - 4m + 4 \right)$$

H-test (continued)

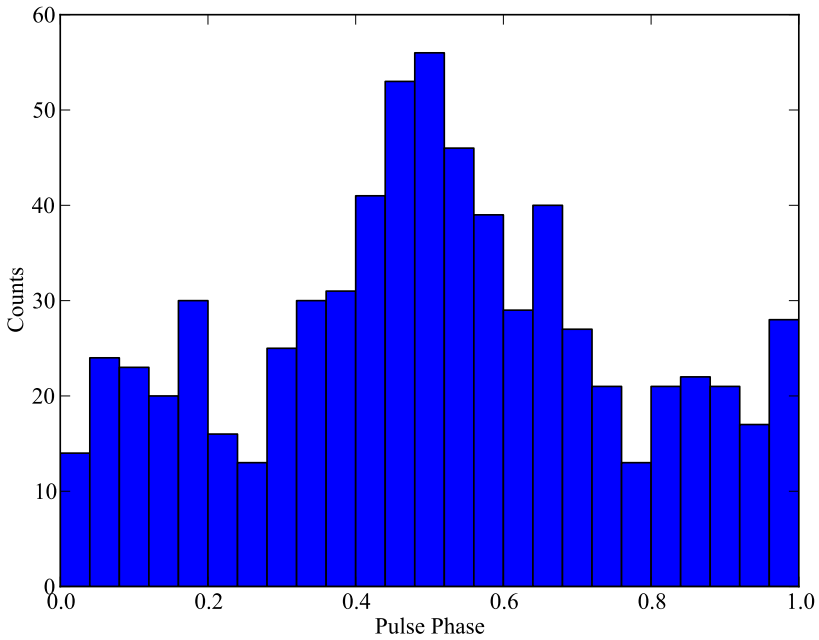
The p-value for the *H*-test was estimated by de Jager et al. (2010) by means of Monte Carlo simulations, and is good down to $\sim 10^{-14}$:

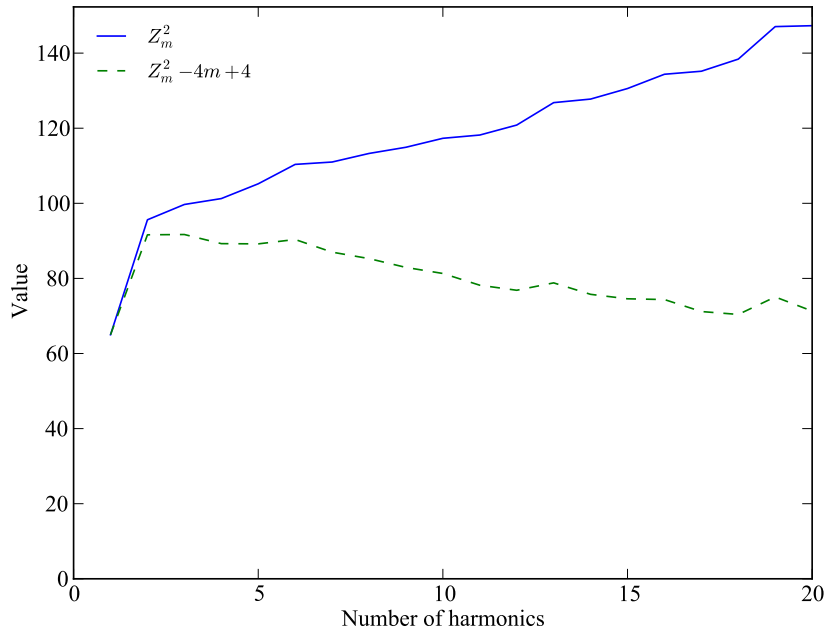
$$P(H) = e^{-0.4 \times H}$$



H-test example

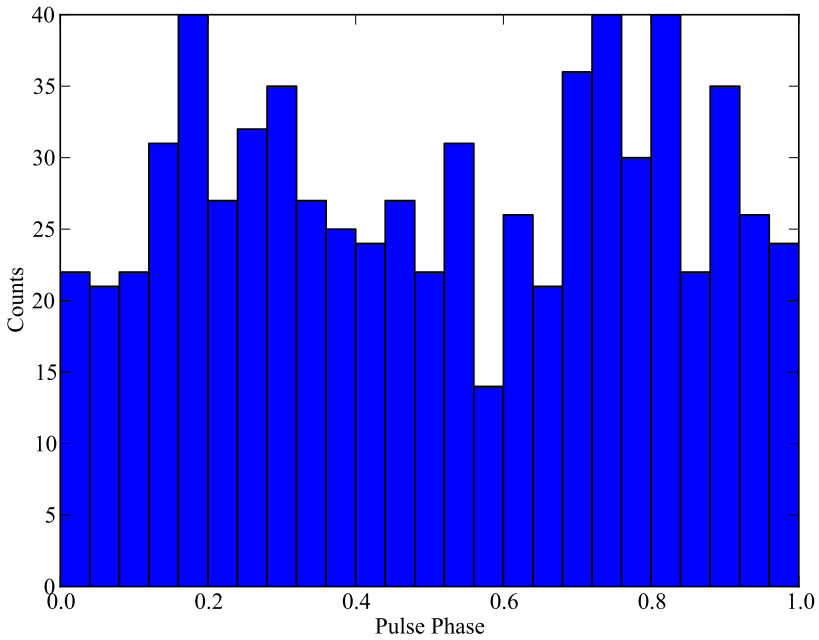
200 events were simulated based on a Gaussian distribution with mean 0.5 and standard deviation 0.1, in addition to 500 background events.

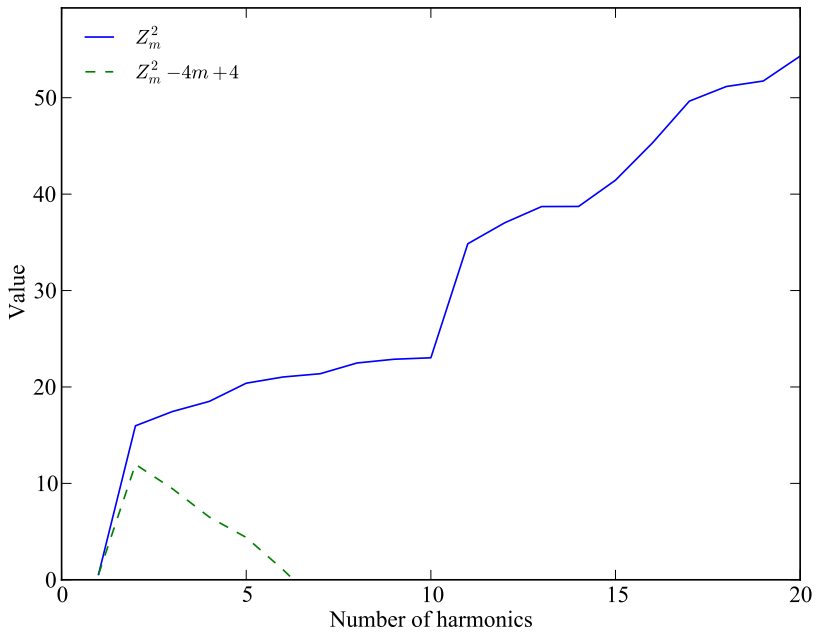




H-test example (continued)

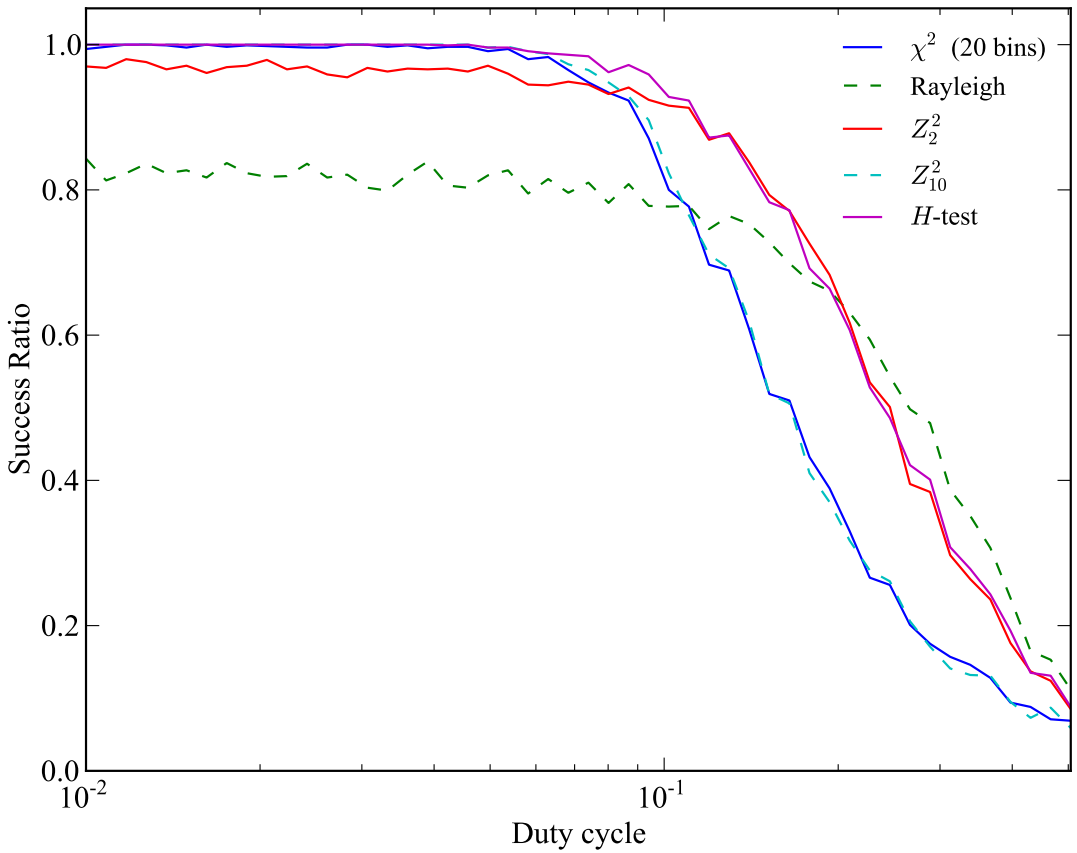
In this example, two sets of 100 events were simulated based on Gaussian distributions with respective mean 0.25 and 0.75, and standard deviation 0.1, in addition to 500 background events.





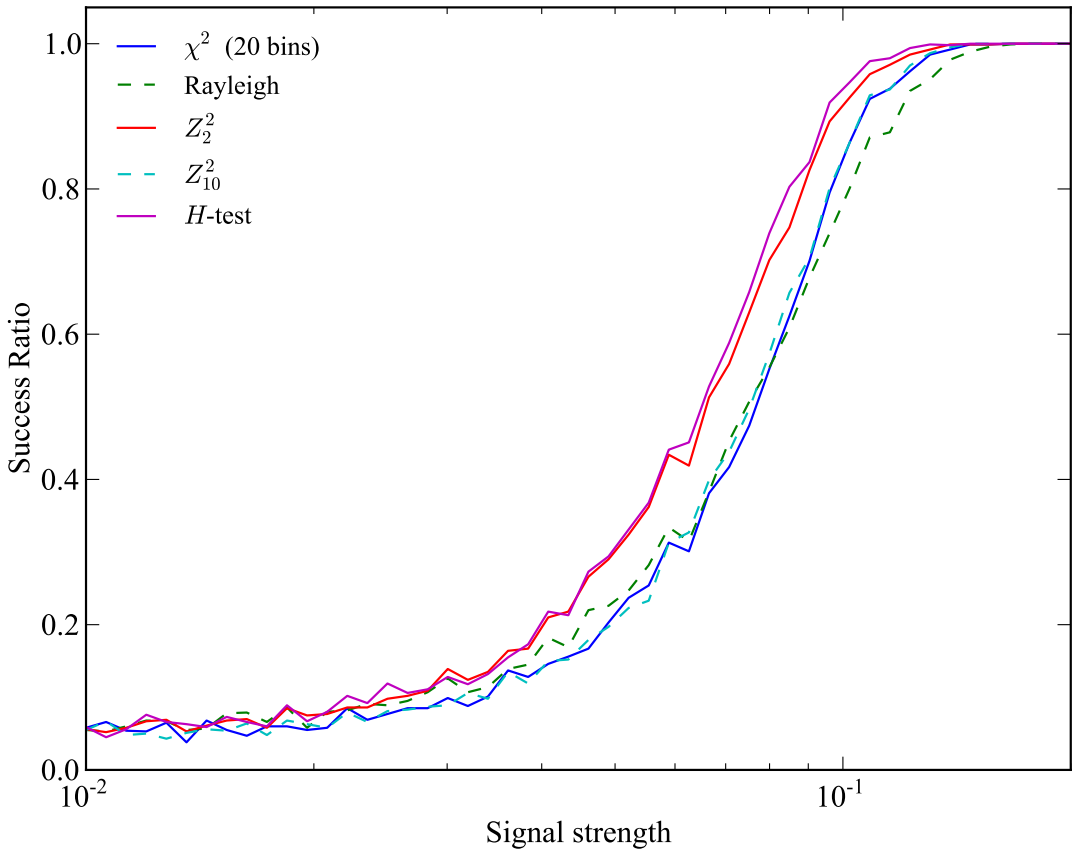
Periodicity tests comparisons

Success ratio as a function of duty cycle for 50 signal events simulated based on a Gaussian distribution with mean 0.5 and 450 background events ($p = 0.1$). For each value of the duty cycle, 1000 realizations of the simulation were performed, and the success ratio was measured using a detection threshold of 5%.



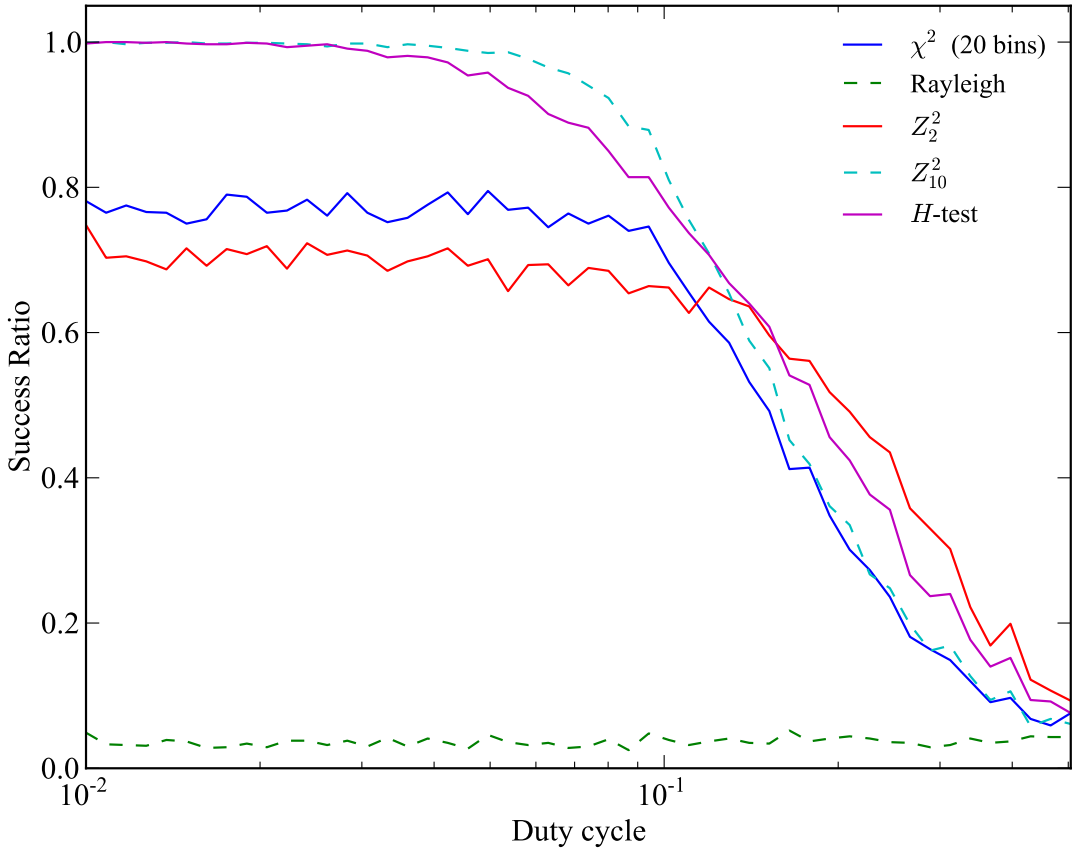
Periodicity tests comparisons (continued)

Same, as a function of the signal strength, p. $(p \times n)$ events were simulated based on a Gaussian distribution with mean 0.5 and standard deviation 0.05, and $(1 - p) \times n$ background events were generated. For each value of the signal strength, 1000 realizations of the simulated were performed.



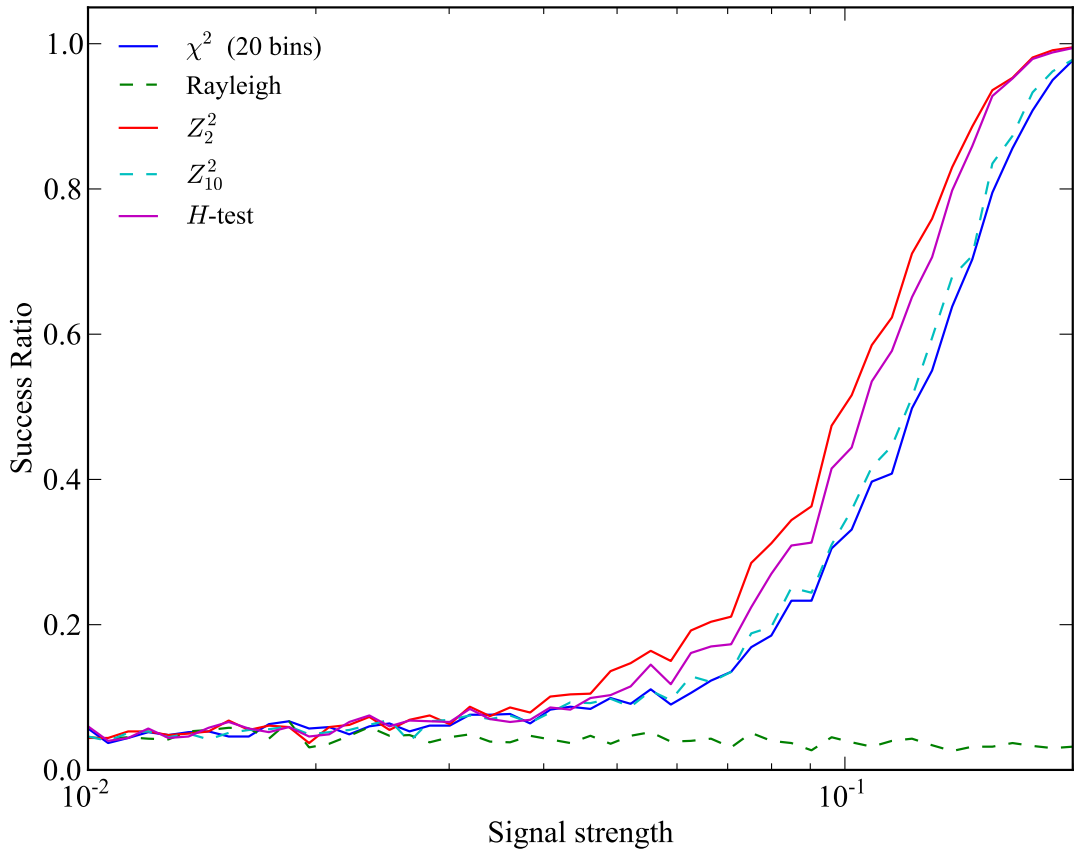
Periodicity tests comparisons (continued)

Success ratio as a function of duty cycle for 50 signal events distributed in two Gaussian distributions with respective means 0.25 and 0.75, in addition to 450 background events ($p = 0.1$). For each value of the duty cycle, 1000 realizations of the simulation were performed, and the success ratio was measured using a detection threshold of 5%.



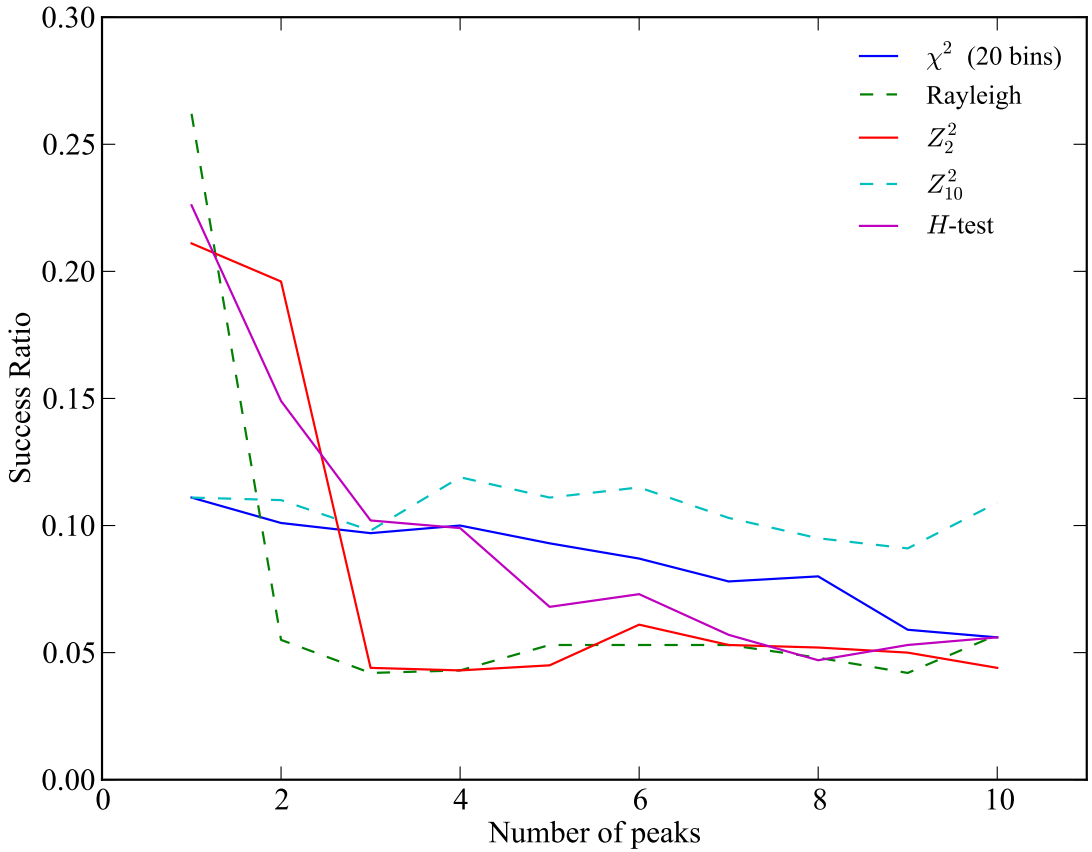
Periodicity tests comparisons (continued)

Same, as a function of the signal strength, p. $(p \times n)$ events were simulated based on two Gaussian distributions with means 0.25 and 0.75 and standard deviation of 0.05, and $(1 - p) \times n$ background events were generated. For each value of the signal strength, 1000 realizations of the simulated were performed.



Periodicity tests comparisons (continued)

Success ratio as a function of the number of peaks in the periodic signal. For each value of the number of peaks n_p, 500 events were simulated using a density function $f(x) = \frac{0.312}{1+0.2\cos(n_p x)}$, and the procedure was repeated 1000 times.



On the usage of weighting techniques

In some situations, it is possible to estimate the probability w_i (taking value on $[0; 1]$) that a given event is due to the analyzed source.

The following sources of information can for instance be used for calculating probabilities:

- the angular separation between the reconstructed direction of the incoming event and the direction of the analyzed source, in conjunction with the Point Spread Function (PSF).
- the energy of the incoming events and the spectrum of the analyzed source.
- etc.

Kerr (2011) proposed modified versions of the Z_m^2 and the H-test statistics, taking event probabilities into account.

Weighted Z^2_m and H-test statistics

If we denote w_i the probability that the ith event originates from the studied source (and therefore, $1 - w_i$ the probability that it is due to background), one can define a weighted Z^2_m statistic as:

$$Z^2_{mw} = \frac{2}{n}\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2\right)^{-1} \sum_{k=1}^{m}\left(\hat{\alpha}_k + \hat{\beta}_k\right)$$

where the weighted trigonometric moments $\hat{\alpha}_k$ and $\hat{\beta}_k$ are now given by:

$$\hat{\alpha}_k = \sum_{i=1}^{n} w_i \cos\left(k x_i\right) ; \hat{\beta}_k = \sum_{i=1}^{n} w_i \sin\left(k x_i\right)$$

We can note that if all probabilities are equal to 1, then the Z^2_{mw} is equal to Z^2_m.

Like the Z^2_m statistic, the Z^2_{mw} is distributed as a χ^2 with $2m$ degrees of freedom.

Weighted Z_m^2 and H-test statistics (continued)

By extension, one can define a weighted H-test statistic based on the weighted Z_m^2 statistic, as:

$$H_{mw} = \max_{1 \leq m \leq 20} = \left(Z_{mw}^2 - 4m + 4 \right)$$

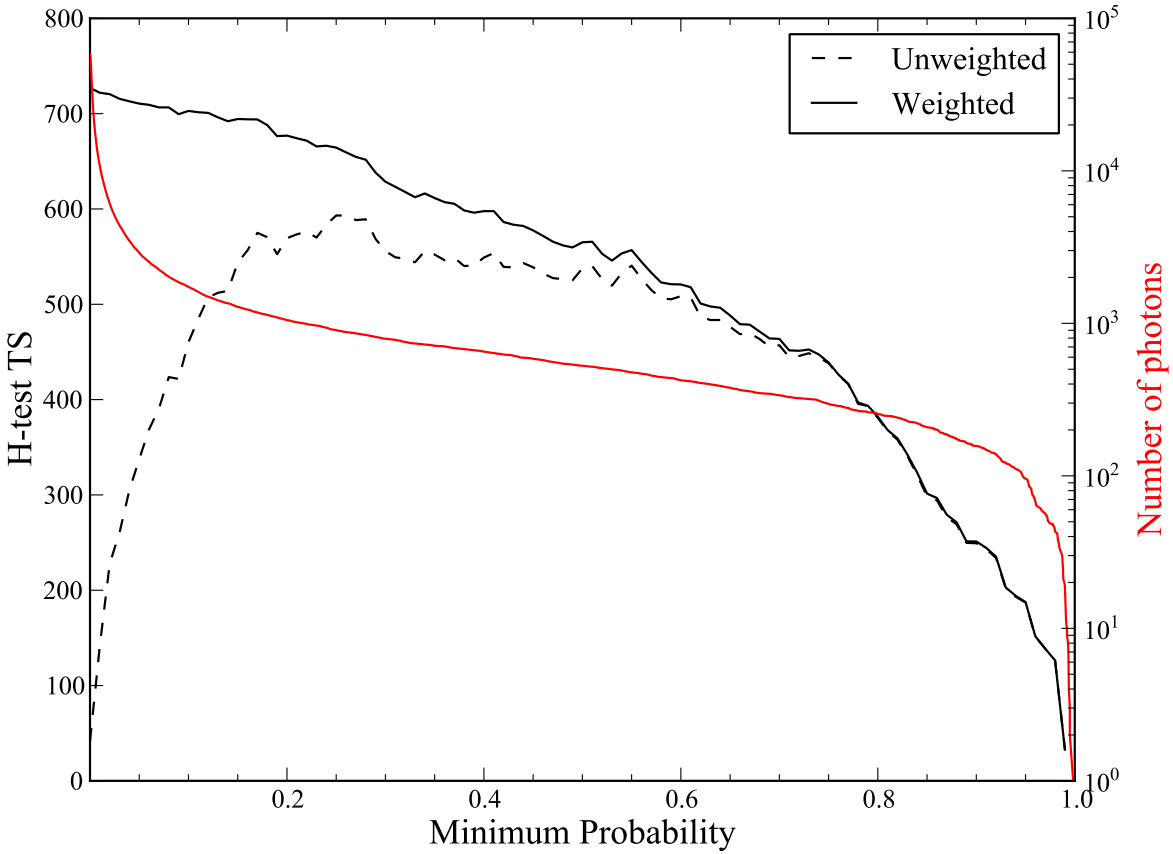
The asymptotic calibration of the weighted H-test is the same as for the unweighted one:

$$P(H_{mw}) = e^{-0.4 \times H_{mw}}$$

Again, if all probabilities are equal to 1, then the weighted H-test statistic is equal to H.

Weighted *H*-test example

Weighted and unweighted *H*-test statistics for a typical gamma-ray pulsar, observed with the *Fermi* LAT, PSR J2017+0603. In this example photon probabilities were calculated by accounting for the reconstructed directions and energies of the gamma-ray photons, and the instrument response functions of the LAT.



Conclusions

- Many tests for uniformity (or non-uniformity) on the circle exist, and are adapted to different situations (expected shape of the modulation, signal strength, etc.)
- The *H*-test seems to be a good omnibus test, for pulsar-like profile shapes. Very common in recent high-energy pulsar papers.
- Additional sensitivity can be achieved by using weighting techniques.

Thanks for your attention!

References

- Ajne, B., Biometrika **55**, 343 (1968)
- Beran, R.J., Ann. Math. Stat. **40**, 1196 (1969)
- Buccheri, R., et al., A&A **128**, 245 (1983)
- Greenwood, J.A., & Durand, D., Ann. Math. Stat. **26**, 233 (1955)
- de Jager, O.C., et al., A&A **221**, 180 (1989)
- de Jager, O.C., & Büsching, I., A&A Lett. **517**, 9 (2010)
- Kerr, M., ApJ **732**, 38 (2011)
- Mardia, K.V., Statistics of directional data (1972)
- Press, W., et al., Numerical recipes 3rd Ed. (2007), and references therein!
- Rayleigh, Lord, Phil. Mag. **37**, 321 (1919)
- Watson, G.S., Biometrika **48**, 109 (1961)