Comparison of different Goodness of fit tests

B. Aslan and G. Zech
Universität Siegen

Durham Conference on
Advanced Statistical Techniques
in Particle Physics

18-22 March 2002

Goodness of Fit Tests

Given n measurements x_i, following an unknown distribution $f(x)$

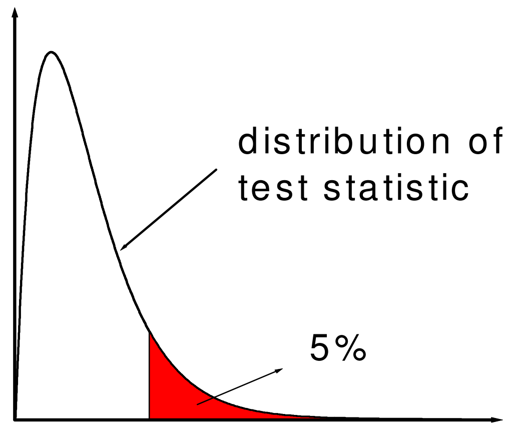
To test the simple hypothesis

$$H_0: \quad f(x) = f_0(x)$$

is a goodness of fit problem

(we consider only simple hypothesis)

- $f_0(x)$ is some specified distribution



Procedure:

1. compute a test statistic based on the deviation of the data from the model
2. compare it to a theoretical or empirical distribution, assuming H_0 is true

Types of Tests

binning required

chi square type

- Pearson
- loglikelihood ratio
- Freeman-Tukey
- Rao-Robson
- Power-Divergence Statistics

binning free

Tests based on

- EDF
- Spacings
- Order Statistics
- Entropy
- Likelihood
- window width (new)
- Energy (new)

most tests described in, Goodness-of-fit techniques,
ed. D‘Agostino R.B., Stephens M.A., Marcel Dekker (´86)

Pearson´s χ^2 Test

compares expected frequencies to observed frequencies n_i
in discrete cells:

$$X^2 = \sum_{i=1}^{B} \frac{(n_i - np_i)^2}{np_i}$$

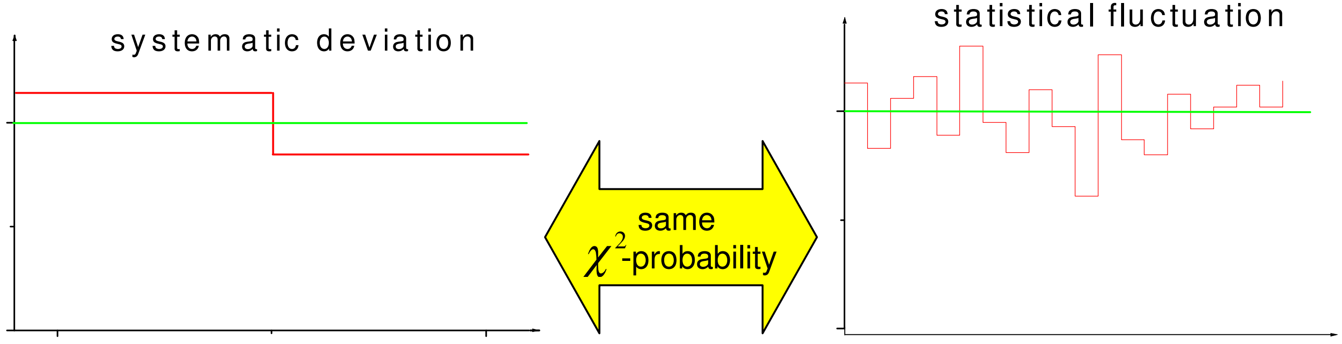
n : number of events

p_i : cell probabilities

When H_0 is true, X^2 approximately follows χ^2 distribution
with B-1 degrees of freedom

Drawback:

- does not take into account the same sign deviations of adjacent bins



- arbitrary choice of the bin widths

  - wide bins $\Rightarrow$ localized systematic deviations aren´t detectable
  - high number of bins $\Rightarrow$ systematic deviations are hidden by the statistical fluctuations

Probability Integral Transformation (PIT)

in some cases it is more convenient to work with data after they have been transformed such that

$$\boxed{z_{(i)} = F(x_i)} \quad \text{PIT} \qquad (x_i > x_{i-1})$$

if the model (H_0) is true Z will be uniformly distributed on (0,1)

$z_{(i)}$ is the i th-ranked transformed data point

a test can then be reduced to a test for uniform distribution

(powerfull tests like smooth-test exist for uniform distribution)

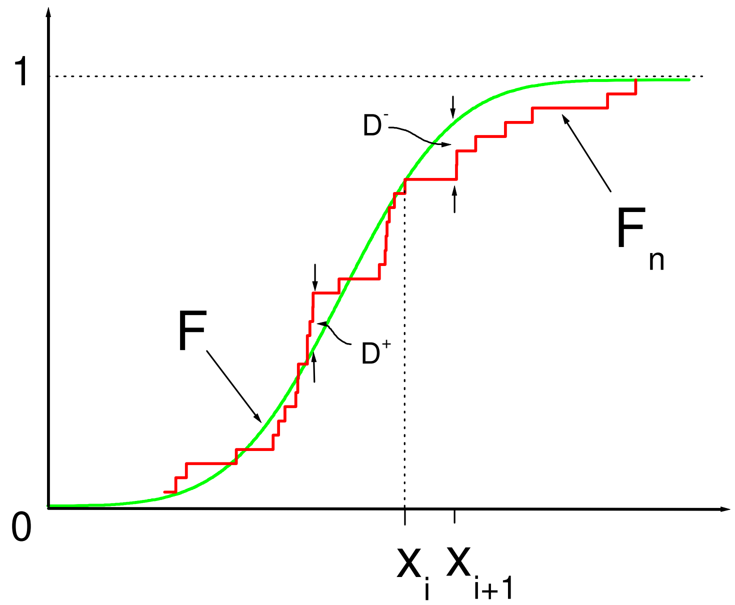
Empirical Distribution Function (EDF) Tests

based on the deviation of the EDF from the CDF (cumulative distribution function) of the model distribution

EDF is constructed by ranking the observations and then computing

$$F_n(X) = \frac{\#\ \ of\ \ observations \leq x}{n}$$

step function that increases $1/n$ in height at each observation



$F_n(X)$ is compared to the model CDF, $F(X)$

EDF-Tests

based on supremum statistics

- Kolmogorov-Smirnov

$$D^{+} = \max_{x}\{F_n(x) - F(x)\} = \max_{1\le i\le n}\left\{\frac{i}{n} - z_{(i)}\right\}$$

$$D^{-} = \max_{x}\{F(x) - F_n(x)\} = \max_{1\le i\le n}\left\{z_{(i)} - \frac{(i-1)}{n}\right\}$$

distribution free in the sense that the critical values do not depend on the specific distribution being tested

$$D = \max_{x}|F_n(x) - F(x)| = \max(D^{+}, D^{-})$$

- Kuiper

$$V = D^{+} + D^{-}$$

is useful for observation on a circle

based on quadratic statistics

- Cramer-von Mises family

$$Q = n\int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 \psi(x) \; dF(x)$$

weighting function

$$\psi(x) \equiv 1 \qquad \text{Cramer-von Mises}$$

$$\psi(x) = \frac{1}{F(x)(1 - F(x))} \qquad \text{Anderson-Darling}$$

Anderson-Darling test gives more weight to the tails

- Watson

$$U^2 = n\int_{-\infty}^{\infty} \left\{ F_n(x) - F(x) - \int_{-\infty}^{\infty} [F_n(x) - F(x)]\, dF(x) \right\}^2 dF(x)$$

for points on a circle Watson
or Kuiper should be calculated

Tests based on order statistics

$z_{(i)}$ follows a beta distribution with parameters i and $n-i+1$

$$\Rightarrow \quad E\left(Z_{(i)}\right)=\frac{i}{n+1} \quad ; \quad Var\left(Z_{(i)}\right)=\frac{i(n-i+1)}{(n+1)^2(n+2)}$$

Test quantity: $$\boxed{v_i = z_{(i)} - \frac{i}{n+1}}$$

Pearson-Hartley $$C = \sup_i |v_i|$$

Hegazy-Green $$T_1 = \sum_i \frac{v_i^{\,2}}{n} \quad ; \quad T_2 = \sum_i \frac{|v_i|}{n}$$

Lohrding $$T_3 = \sup_i |d_i| \quad ; \quad T_4 = \sum_i |d_i| \quad ; \quad T_5 = \sum_i d_i^{\,2}$$ where

$$d_i = v_i (n+1)\sqrt{n+2}\left[i(n-i+1)\right]^{-1/2}$$

Spacing tests

measure the distance between consecutive events

spacings D_i are defined by : $D_i = z_{(i)} - z_{(i-1)}$

Greenwood $$G = \sum_{i=1}^{n+1} D_i^2$$

Cressie $$S_m = \sum_i \left(nG_{i,m}\right)^2 \quad , \quad G_{i,m} = z_{(i+m)} - z_{(i)}$$

Vasicek $$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx = \begin{cases} 0 & \text{if } H_0 \text{ true} \\ < 0 & \text{otherwise} \end{cases}$$ entropy

$$\Rightarrow \quad H_{m,n} = \frac{1}{n} \sum_{i=1}^{n} \log\left[\frac{n\left(z_{(i+m)} - z_{(i-m)}\right)}{2m}\right]$$ entropy estimate

Neyman Smooth Test

alternative hypothesis is parametrized by

$$f(x) = C(\theta_1, \theta_2, \ldots, \theta_k) \exp\left[1 + \sum_{i=1}^{k} \theta_i l_i(x)\right]$$

$l_i(x)$ Legendre polynomials

θ_i parameters

$C(\theta_1, \theta_2, \ldots, \theta_k)$ normalizing constant

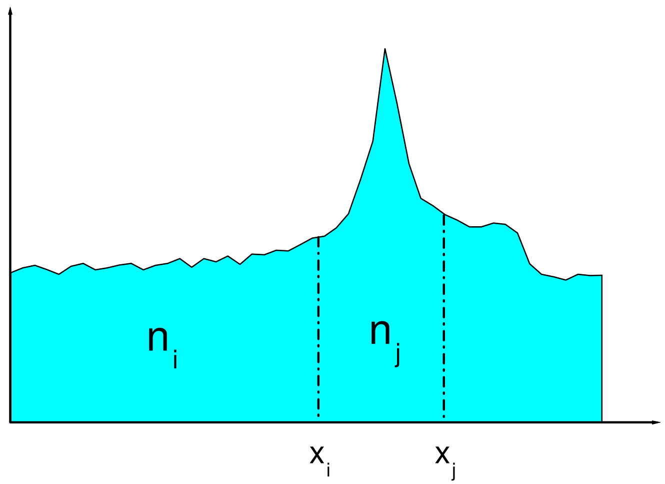
Neyman found by likelihood ratio method

$$N_k = \frac{1}{n} \sum_{i=1}^{k} \left(\sum_{j=1}^{n} l_i(z_i) \right)^2$$

asymptotic distribution of N_k is $\chi_k^{\,2}$

Three region test

is constructed for testing against a clustering or bump type alternative



test quantity is

$$O = \sup_{i,j}\left\{(n_i - np_i)^2 + (n_j - np_j)^2 + \left((n_i + n_j) - n(p_i + p_j)\right)^2\right\}$$

test distribution is constructed by MC simulation

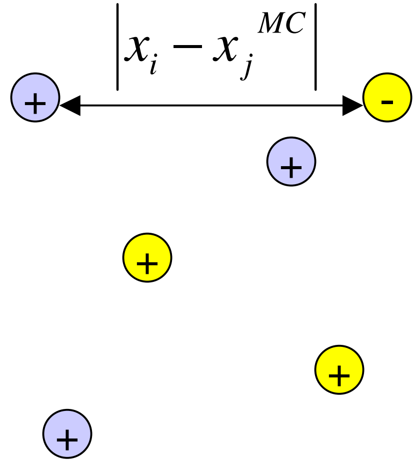
Energy test

based on the fact that the potential energy of charge distributions of opposite total charge is minimum when their distributions are equal

How to compute potential energy Φ ?

$$\Phi = \sum_{i<j}^{n} \frac{1}{|x_i - x_j|} - \sum_{i,j}^{n} \frac{1}{|x_i - x_j^{MC}|} + \cancel{\sum_{i<j}^{n} \frac{1}{|x_i^{MC} - x_j^{MC}|}}$$

irrelevant



X negative experimental sample

X^{MC} positive Monte Carlo sample (H_0)

same number of experimental and MC elements

more general:

test quantity Φ contains two terms Φ_1 , Φ_2

$$\Phi_1 = \frac{1}{n^2} \sum_{i<j} R(|x_i - x_j|) \qquad \Phi_2 = -\frac{1}{nm} \sum_{i,l} R(|x_i - x_l^{MC}|)$$

n: size of exp sample

$$\boxed{\Phi = \Phi_1 + \Phi_2}$$

m: size of MC sample

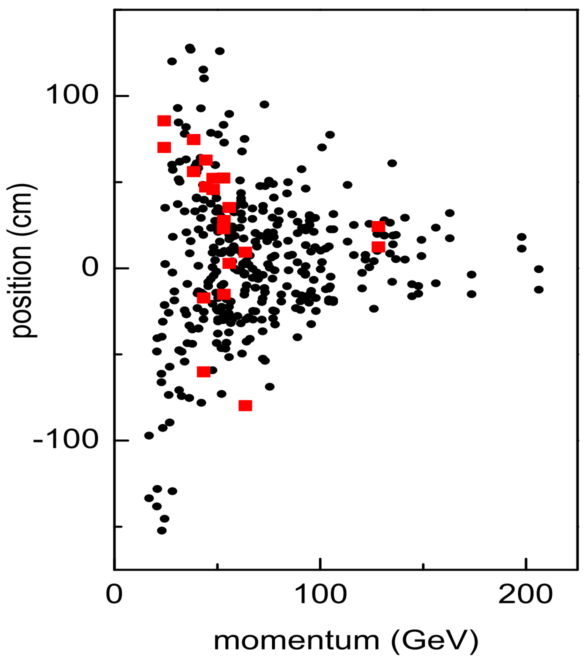
we investigated following correlation functions:

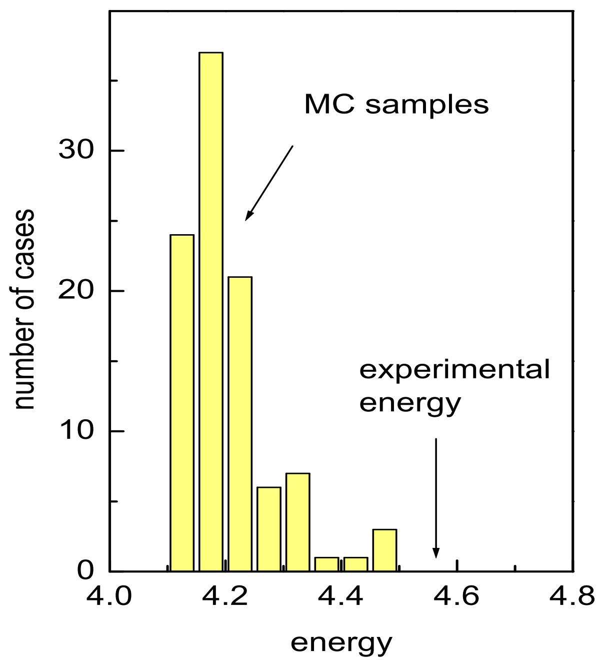
$$R(r) = \begin{cases} 1/r^{\kappa} & \\ -\ln r & \text{long rangle} \\ e^{-r^2/(2\alpha^2)} & \text{short range, related to } \chi^2 \text{ test} \end{cases}$$

more described in, A new class of binning free, multi-dimensional goodness-of-fit tests: the energy test, available on http://xxx.lanl.gov/list/hep-ex/0203

Example

the positions of tracks from ψ decays are plottet against the momentum of ψ



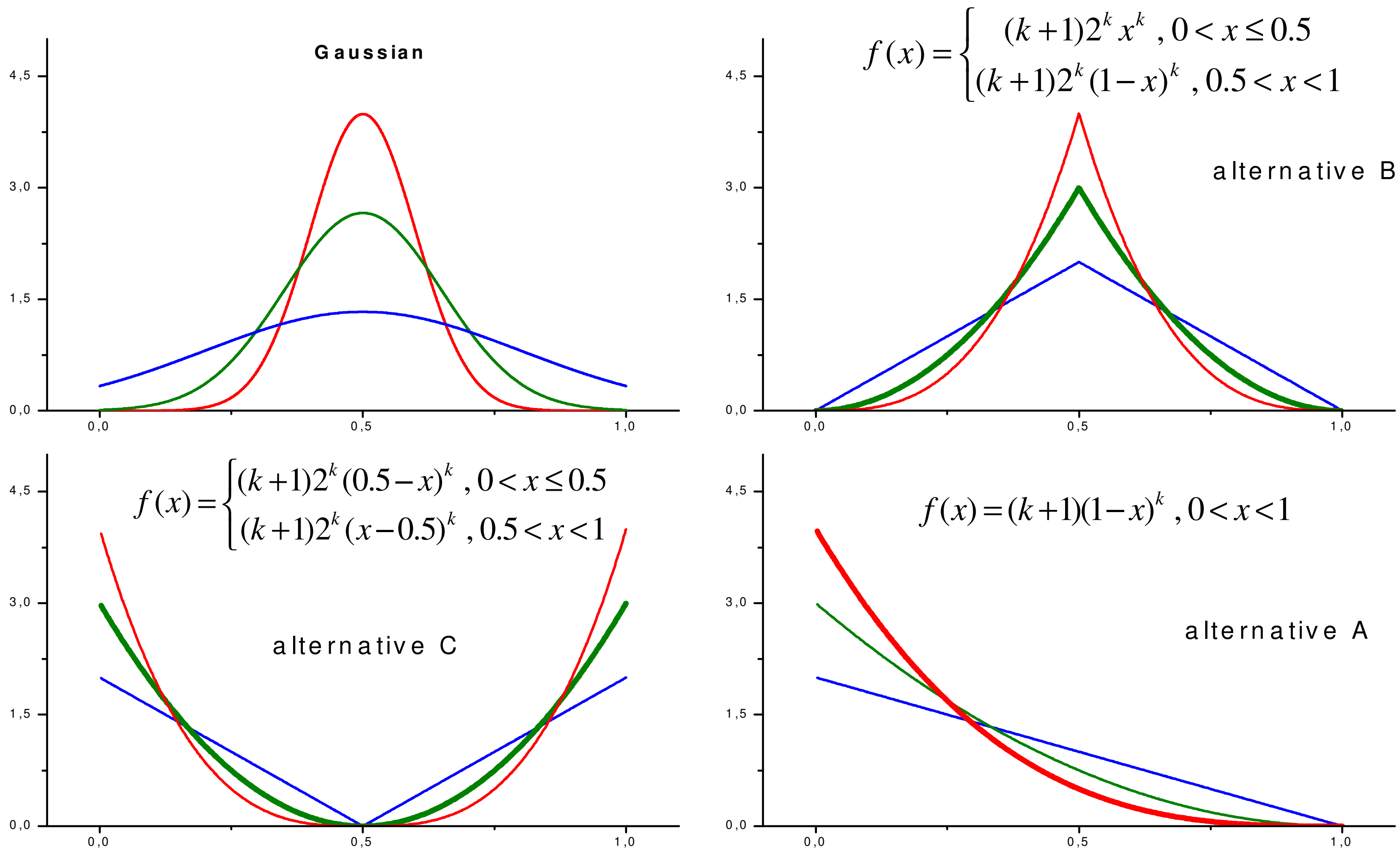


Power comparisons: Test for uniformity (univariate)

the statistics are

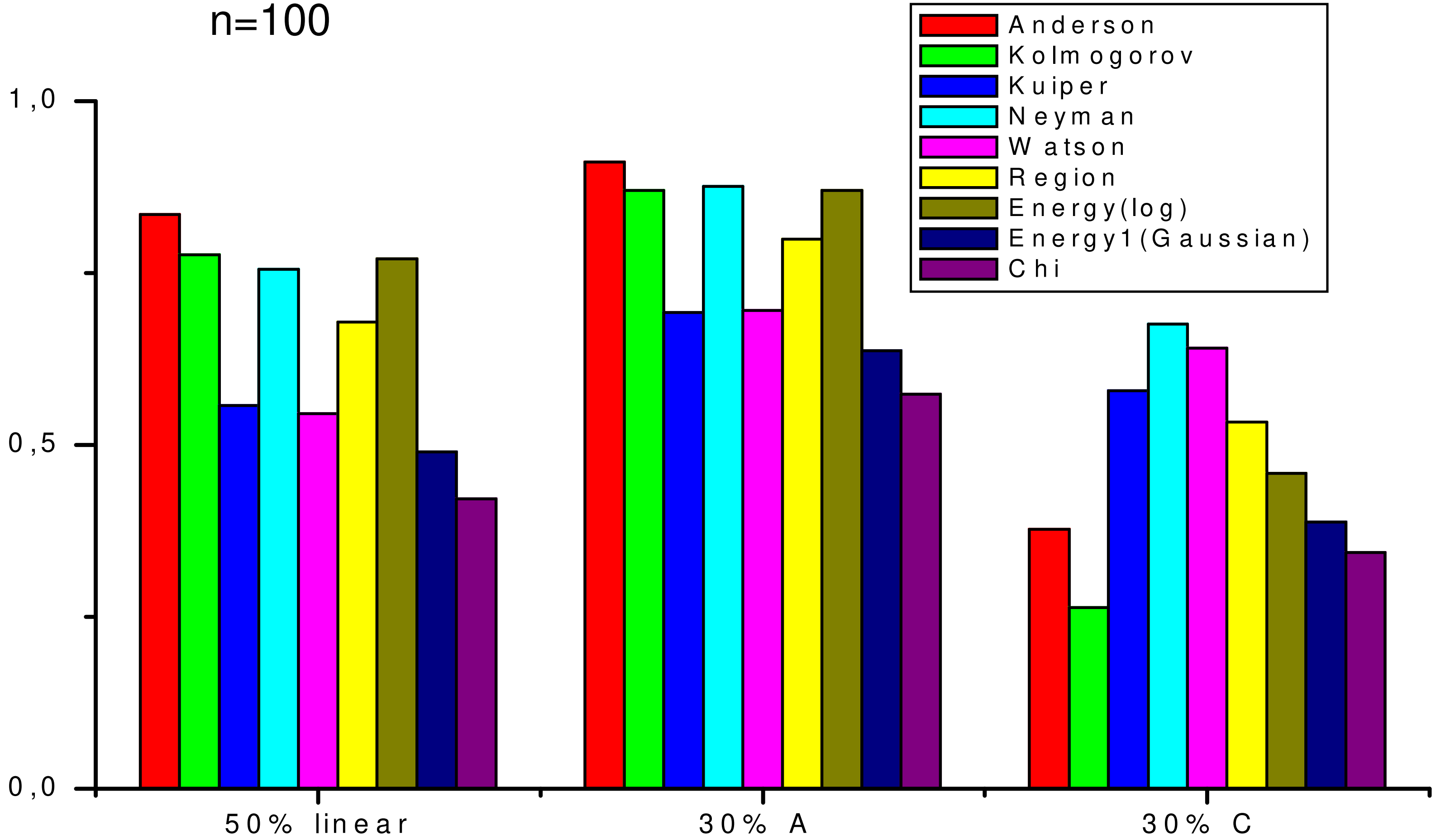
1. Anderson-Darling
2. Energy (log)
3. Energy (Gaussian)
4. Kolmogorov
5. Kuiper
6. Chi-squared
7. Neyman (k=2)
8. Watson
9. 3 Region

Alternative



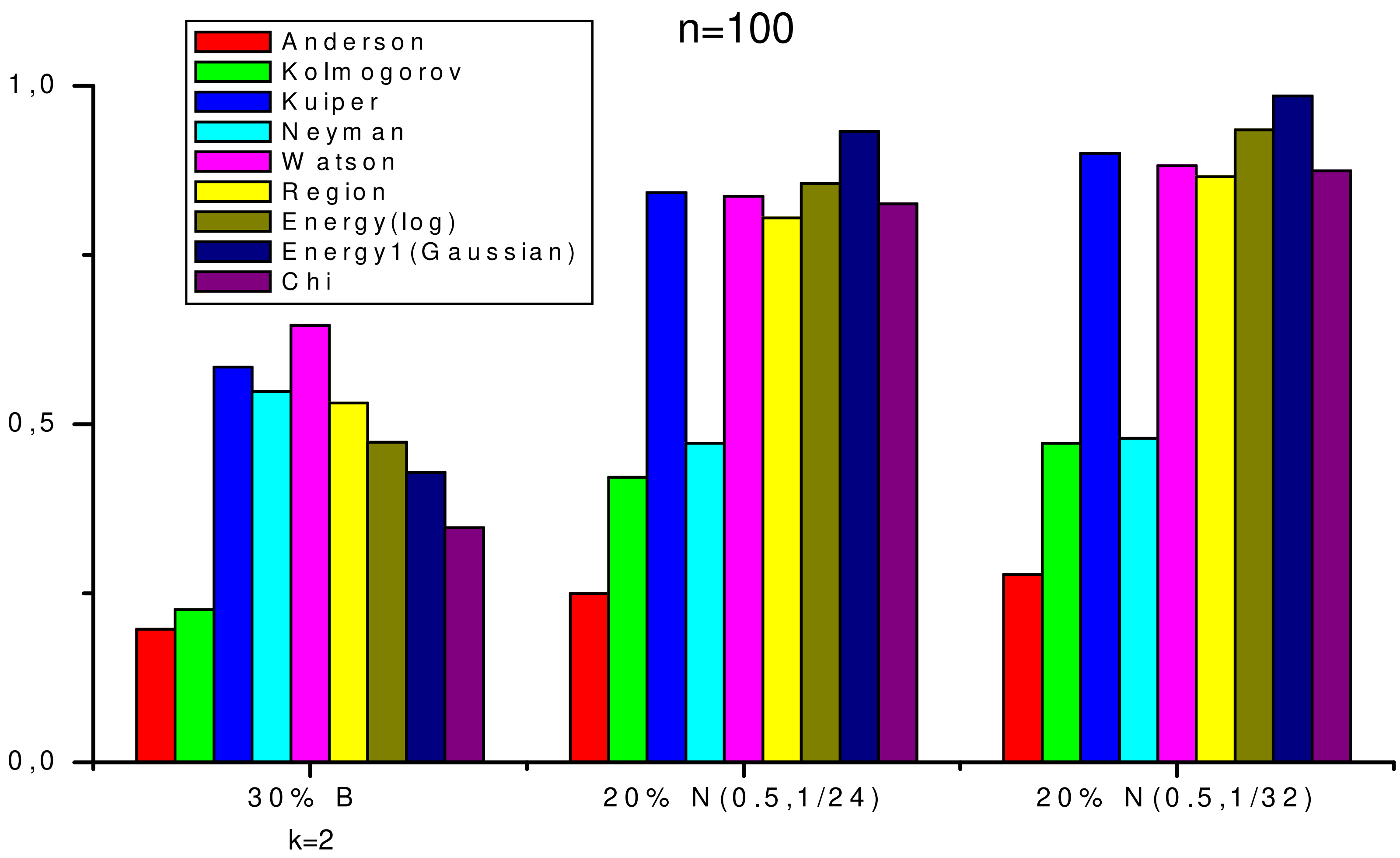
Power comparisons

n=100



k=3 k=2

Power comparisons



Multivariate Tests

very few references can be found in the literature about multivariate goodness of fit

we investigated following test procedures for bivariate normality and compared their powers

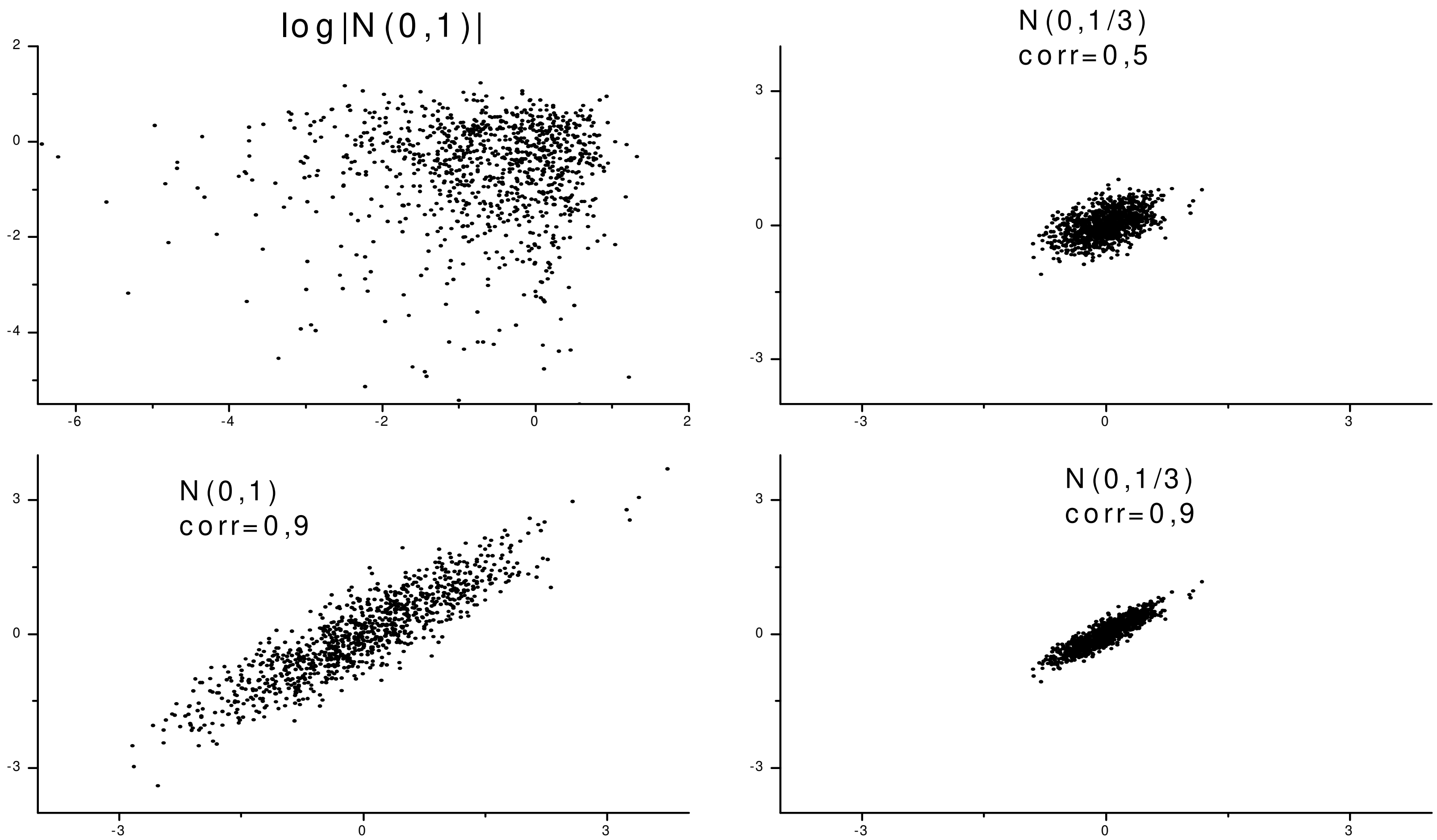
1. Energy test
2. Neyman smooth test
3. Mardia test

are restricted to Gaussian distribution

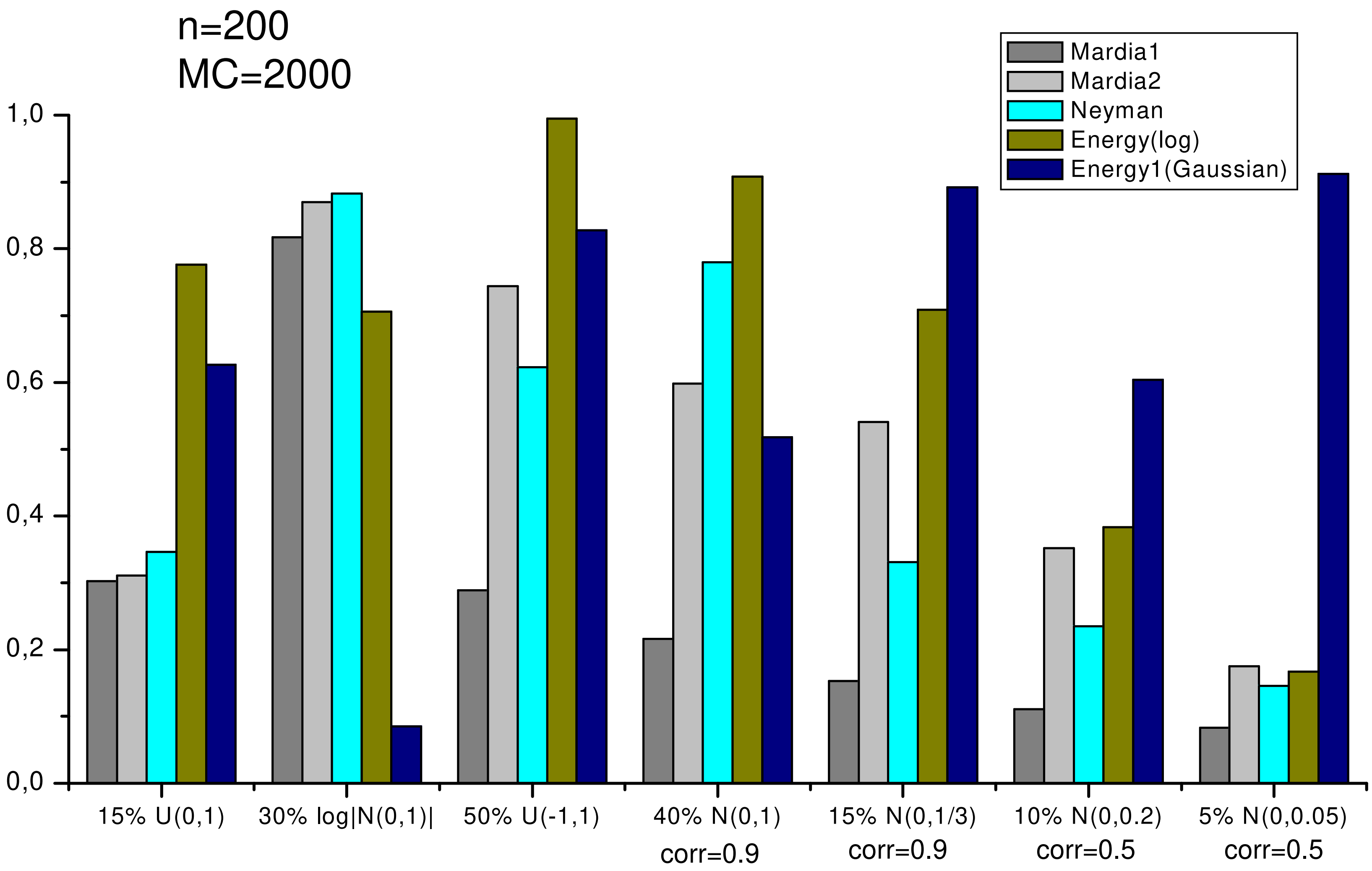
tests described in

- Goodness-of-fit Techniques, D´Agostino R.B., Stephens M.A., Marcel Dekker (1986)
- Smooth Tests of Goodness of Fit, Rayner J.C.W., Best D.J., Oxford Uni. Press (1989)

Alternative



Power comparisons



Conclusions

- no test statistic is better (concerning all alternatives) than its competitors
- there is no general theory that tells us how to choose a test
- for univariate distributions there are more powerful tests than the χ^2 test
- Neyman, Anderson and Kolmogorov are sensitive to a shift of the mean
- Watson and Kuiper are sensitive to a change in variance rather than in mean
- Energy test (log as corr. function) detects long range distortions
- Energy test (Gaussian as corr. function) useful to detect short range distortions
- Gaussian energy test is superior to χ^2 test (no arbitrary binning, more powerful)
- Energy test can be extended to the multivariate case
  (powerful for testing multivariate normality)