AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Netherlands School of Economics
ECONOMETRIC INSTITUTE

GIANNINI FOUNDATION OF
AGRICULTURAL ECONOMICS
LIBRARY
WITHDRAWN
MAR 1 3 1968

Report 6713

ON THE POWER OF THREE GOODNESS-OF-FIT TESTS

(A Simulation Approach)

by A.P.J. Abrahamse and B.S. van der Laan

November 20, 1967 Preliminary and Confidential

ON THE POWER OF THREE GOODNESS-OF-FIT TESTS

(A Simulation Approach)

by A.P.J. Abrahamse and B.S. van der Laan

Contents

	Page
1. Introduction	1
2. Three Goodness-of-Fit Tests	2
2.1. The Likelihood-Ratio Test	2
2.2. The Chi-square Test	3
2.3. The Chi-square Test on Transformed Observations	4
3. The Quality of Tests	4
4. The Simulation Method	6
5. Results and Conclusions	7
5.1. The Type-I Error	7
5.2. The Type-II Error	9
5.3. Summary of the Results	16

1. INTRODUCTION

Sometimes one uses the chi-square test or the likelihood-ratio test for testing goodness of fit. Serious problems exist with regard to these two tests, problems which can easily be overlooked but which are nevertheless of importance for the practical research worker.

One problem is that no analytical expression is known for the distribution of the test-statistics concerned, even not under the null-hypothesis H_0. But fortunately this distribution is asymptotically known, for large samples which does not alter the fact that we have to be content with a critical region determined on the basis of an only approximate distribution. Little is known about the degree of approximation for small N.

Another problem is concerned with the power of the above tests. Even in the case where an alternative hypothesis H_1 is specified, the power is completely unknown because one does not know the distribution of the chi-square and LR-test-statistics under H_1, even not asymptotically. To this day all efforts to find this distribution have failed because of the serious analytical difficulties which arise.

As was mentioned before, the above two problems are of importance for the practical research worker and hence for the econometrician. In fact, when using a test, knowledge about the degree of approximation of the distribution under H_0 and

[1] We are indebted to Prof. J. Koerts for many valuable suggestions.

the power of this test is indispencible. Therefore we shall investigate these problems by means of a simulation study. Perhaps, from an analytical point of view, this approach is not very satisfactory. But nevertheless we need more insight in the properties of these goodness-of-fit tests and thus, for the time being, we shall have to take resort to a simulation approach.

The order of discussion is as follows:
In section 2 we describe the tests that will be investigated. After that, in section 3, the general criteria for the quality of tests are discussed. Next, section 4 is devoted to the method of simulation we used, whereas the last section gives the results and the conclusions.

2. THREE GOODNESS-OF-FIT TESTS

2.1. The Likelihood-Ratio Test

Let $x_1, \ldots, x_N$ be a sample of N independent observations on a random variable with unknown distribution function F(x). Suppose that we want to test the null-hypothesis

$$H_0: F(x) = F_0(x) \tag{2.1}$$

against the alternative hypothesis

$$H_1: F(x) = F_1(x) \tag{2.2}$$

We shall assume that $F_0(x)$ is completely specified, in other words (2.1) is a simple hypothesis. The range of the variate X is divided into K mutually exclusive classes. We may then calculate the probability of an observation falling in each class, given the null-hypothesis is true, since $F_0(x)$ is completely specified. These probabilities will be denoted by p_{0k}, $k = 1, \ldots, K$. The numbers of observations falling in the K classes are given by n_k with $\sum_{k=1}^{K} n_k = N$. If the true probability of an observation falling in the kth class is denoted by p_k the n_k, $k = 1, \ldots, K$, are multinomially distributed and hence the likelihood of the sample is

$$L(n_k, p_k) = \prod_{k=1}^{K} p_k^{n_k} \tag{2.3}$$

It is easy to verify that the likelihood (2.3) is maximized when we substitute the Maximum-Likelihood estimators for p_k

(2.4) $$\hat{p}_k = \frac{n_k}{N}$$

Under the null-hypothesis the likelihood is

(2.5) $$L(n_k, p_{0k}) = \prod_{k=1}^{K} p_{0k}^{n_k}$$

so that the LR-statistic for testing H_0 against H_1 is

(2.6) $$\lambda = \frac{L(n_k, p_{0k})}{L(n_k, \hat{p}_k)} = N^N \prod_{k=1}^{K} \left(\frac{p_{0k}}{n_k}\right)^{n_k}$$

The critical region is $0 < \lambda < A$, where A is chosen to give the desired probability of a Type-I error $p(I)$.

The exact distribution of (2.6) is unknown but we do know that as $N \to \infty$, $-2 \log \lambda$ is asymptotically distributed in the chi-square form with $K - 1$ degrees of freedom.

2.2. The Chi-square Test

Another test commonly used for testing H_0 is the chi-square test which was originally proposed by Karl Pearson. The test-criterion is

(2.7) $$X^2 = \sum_{k=1}^{K} \frac{(n_k - Np_{0k})^2}{Np_{0k}}$$

and is, just as $-2 \log \lambda$, for large samples approximately chi-square distributed with $K - 1$ degrees of freedom. This follows from the fact that the n_k are multinomially distributed with parameters p_{0k}, $k = 1, \ldots, K$. It is well-known that the multinomial distribution approaches the normal distribution for large N. Thus for large N the n_k are normally distributed with means Np_{0k}. The exponent of this normal distribution proves to be a quadratic form which is exactly the chi-square test-statistic (2.7). Thus X^2 has an asymptotic chi-square distribution with $K - 1$ degrees of freedom. The right-hand tail of this distribution is taken as the place of the critical region.

Furthermore, the chi-square test is asymptotically equivalent with the LR-test, that is to say, for large samples X^2 takes the same value as $-2 \log \lambda$. From time to time the suggestion has been made that, for small samples, the LR-test is to be preferred to the chi-square test in view of possible higher power [1]. In section 5.2, however, we shall see that this happens only in a special case.

2.3. The Chi-square Test on Transformed Observations

We can also apply the usual chi-square test to the transformed observations. One has used e.g. the well-known probability-integral transformation. Thus if $x_1, \ldots, x_N$ is a given sample and if the null-hypothesis to be tested is H_0: $F(x) = F_0(x)$, we make the transformation

$$y_i = F_0(x_i) \qquad (i = 1, \ldots, N) \tag{2.8}$$

The y_i $(i = 1, \ldots, N)$ have then a uniform distribution over the interval $(0, 1)$. It is suggested that this finite range has some advantage over an infinite range $(-\infty, \infty)$ in view of the division into subclasses. We may divide the interval $(0, 1)$ into K classes of equal length which means at the same time that these classes have equal probabilities of containing an observation y_i. These probabilities are all equal to 1/K. The number of y_i's falling in class k will be called n_k^*. The test statistic is the ordinary chi-square test-statistic applied to the n_k^* and has the form

$$X^{*2} = \frac{\sum_{k=1}^{K} (n_k^* - \frac{N}{k})^2}{\frac{N}{k}} \tag{2.9}$$

$$= \frac{K}{N} \sum_{k=1}^{K} n_k^{*2} - N$$

This test-statistic has also an asymptotic chi-square distribution with K - 1 degrees of freedom just as the ordinary chi-square test-statistic and the critical region should be chosen in the right-hand tail.

3. THE QUALITY OF TESTS

For the determination of a critical region for the tests (2.6), (2.7) and (2.8) one uses the chi-square distribution with K - 1 degrees of freedom. With the help of tables of the χ^2-distribution a critical region is chosen corresponding to a

desired probability of a Type-I error p(I). The true probability of a Type-I error, say $p^*(I)$, however, is in general only equal to p(I) if the true distribution of the test-statistic concerned is identical with the distribution on which the critical region is based. We know that, for the three tests we are considering, this is only true for large samples. Thus for small N there will be a difference between p(I) and $p^*(I)$. It is not very clear how large the sample should be in order to make the approximation precise enough.

As is already mentioned above the data should be grouped into classes: the essence of the χ^2-test and of the LR-test is to reduce the problem to the multinomial distribution. As to the χ^2-test, it is generally believed that the use of the approximating chi-square distribution is precise enough for practical purposes if all the $Np_k \geqq 10$ [3, p. 420]. If some of the $Np_k < 10$ it would be advisable to pool these groups so that every group contains at least 10 expected observations. It stands to reason that the sample should be large enough to make this pooling possible. However, there is no general agreement about the admissable minimum of the theoretical probabilities Np_k. Kendall [2, p.440], for example, mentions a number of 5. Cochran [1], on the contrary, has shown that one or two frequencies may be allowed to fall to 1 or even lower, if X^2 has at least 6 degrees of freedom, without disturbing the test with p(I) = .05 or .01. For an econometrician this is a crucial problem because in economics there are seldom more than 20 or 30 observations in a particular situation.

However, the quality of a test is not only determined by p(I) but it is equally important to know the probability of a Type-II error p(II). In practical applications this reverse of the medal is often neglected. This may have serious consequences. In many cases a research worker uses a test the probability of a Type-I error of which is e.g. .05 without knowing or bothering about the probability of a Type-II error. The man is very happy because he may be sure that on the average in only five of a hundred samples he will reject the null-hypothesis when it is true. But often he is not aware of the fact that in perhaps fifty or more of a hundred samples he accepts the null-hypothesis while it is not true.

How is it possible that so many people who test a hypothesis do not pay attention to the power of their test? The answer to this question is not very difficult: the probability of a Type-I error can be determined by using the distribution of the test-statistic under the null-hypothesis. Of many test-statistics that are used in practice these distributions are known. If they are not exactly known an approximate distribution can often be used for large samples, such as in the LR-test and the chi-square test. In order to obtain the probability of a Type-II error, however, it is necessary to know the distribution of the test-statistic given that the alternative hypothesis H_1 is true. For most tests that are used these distributions have not been found, even not for large samples.

In the next section we are going to describe the simulation method by which we shall try to elucidate the problems posed in this section to a certain extent.

4. THE SIMULATION METHOD

We want to examine the probability of a Type-II error, that is the probability of accepting the null-hypothesis when it is not true. As the probability of a Type-II error equals 1 minus the power it does not make any difference whether we investigate the probability of a Type-II error or gather information about the power of a test. The power of a test is thus the probability of rejecting the null-hypothesis when it is not true. In this section we shall describe the method with which we tried to wretch some information about the power of our three tests from the secrecy.

If we want to get some insight into the power of a certain test it is necessary to specify the alternative hypothesis. As the chi-sqaure test, the LR-test and the X^{*2}-test are almost always used without specifying a clear alternative hypothesis[1] we investigated the power under some more or less different alternative hypotheses. To get concrete results specification of the shape of the distributions is unavoidable. We have chosen the normal form. Of course this is a restrictive procedure but there are no reasons to assume that the results would have been very different when another form of the distributions would have been chosen. We examined the power of testing the null-hypothesis H_0: $F(x) = N(0, 1)$ against the alternative hypothesis H_1: $F(x) = N(\mu, \sigma)$ for different values of μ and σ. It should be kept in mind that in practice we should never use the t^2-test, t^{*2}-test. and LR-test in such a case. For if the distribution under H_0 and that under H_1 differ only in location and scale we are able to find more powerful tests. The three tests are typically tests of the form: H_0: $F(x) = F_0(x)$ against H_1: $F(x) \neq F_0(x)$ and hence no alternative hypothesis is specified.

A hundred samples were drawn from a normal distribution with a mean μ and a standard deviation σ, so that the alternative hypothesis was true. Using the tables of the chi-square distribution a critical region was determined for a given probability of a Type-I error. The test statistic was calculated and the number of times that the test statistic took a value in the critical region was determined. This number divided by hundred gives us an estimate of the power of the test which is equal to 1 - p(II). This was done for the three tests described in section 2, for 5 different values of p(I), for samples consisting of 20, 30, 50, and 100 observations, and for several alternative hypotheses.

---

[1] For a goodness-of-fit test is typically not a test on the parameters of a known distribution but a test on the shape of the distribution.

The range of the variate X was divided into 4, 6, 10, and 20 classes respectively, where the classes were taken to cover equal ranges of the variate, except at an extreme where the range of the variate is infinite. We could also have chosen classes with equal hypothetical probabilities. But this would not necessarily have improved the power of the tests. For if the number of classes is not too large the equal-probabilities method may well result in a loss of sensivity at the extremes of the range of the variate. Nevertheless, the grouping of the observations into classes is a somewhat obscure problem. In the derivation of the asymptotic chi-square distribution of the chi-square test-statistic for example, the fact is used that the multinomial distribution approaches the normal distribution for large N, but this is only true for not too small Np_k.

In section 3 we touched already upon the problem of the influence of the magnitude of the Np_k on the difference between the probability of a Type-I error on which the critical region is founded p(I) and the true probability of a Type-I error p*(I). In section 5.1 we shall first show some by simulation obtained results with respect to this last problem. After that, in section 5.2, we shall turn the power.

5. RESULTS AND CONCLUSIONS

5.1. The Type-I error

A Type-I error is made when the null-hypothesis is rejected though it is true. If the distribution of the test-statistic is known under the null-hypothesis a critical region can be determined for a given probability of a Type-I error. Mostly this distribution is not exactly known, however. Then sometimes an approximate distribution for large samples may be used to determine a critical region. As was mentioned already in the previous sections the three test-statistics we consider in this paper only have such approximate distributions. They are all asymptotically χ^2-distributed with K - 1 degrees of freedom, where K denotes the number of classes into which the range of the variate is divided. So we used the tables of the χ^2-distribution to determine a critical region for a given p(I). Owing to this approximative character of the χ^2-distribution the true probability of a Type-I error, say p*(I), will in general differ from p(I). It is well-known that one factor which is responsible for the order of magnitude of this difference is the average number of observations falling in each class (see section 3). In order to get an idea about the size of difference between p(I) and p*(I) and about the influence of the average number of observations per class we estimated p (I) by simulation for several cases. We shall denote our estimates by $\hat{p}(I)$. We have drawn a hundred samples of N observations from a normal distribution with mean zero and standard deviation one and we tested the null-hypothesis H_0: $x \approx N(0, 1)$ and counted the

number of times that the test-statistic was larger than the critical value which was taken from the appropriate χ^2-tables. This number divided by 100 gives $\hat{p}(I)$. This was done for different values of N, K and p(I). The results are shown in Table 1. The first column gives the number of times that the null-hypothesis was rejected for the χ^2-test, the second for the LR-test, and the third for the χ^{*2}-test.

Table 1

$100 \times \hat{p}(I)$ for $H_0: \mu = 0 \quad \sigma^2 = 1$

p(I) = .005

K	N = 20			N = 30			N = 50			N = 100		
4	1	0	1	0	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	0	0	1	0	0	0	0	1	2	0
20	0	0	0	0	0	0	0	0	0	0	0	0

p(I) = .01

K	N = 20			N = 30			N = 50			N = 100		
4	1	0	1	0	0	1	0	0	0	0	0	0
6	3	0	1	0	0	0	0	0	0	0	0	0
10	1	1	1	1	1	0	0	0	0	2	2	0
20	0	0	0	0	0	0	0	0	0	0	1	0

p(I) = .025

K	N = 20			N = 30			N = 50			N = 100		
4	2	0	3	1	1	3	0	0	1	1	0	1
6	4	0	3	2	2	2	0	0	0	1	0	0
10	3	2	1	1	3	1	2	1	0	3	3	1
20	1	0	0	2	2	0	0	0	0	0	1	0

p(I) = .05

K	N = 20			N = 30			N = 50			N = 100		
4	2	1	5	6	1	4	0	3	3	2	0	3
6	6	5	4	5	3	4	1	0	0	4	0	3
10	5	4	2	1	5	2	8	4	0	5	5	2
20	4	1	1	3	4	0	2	1	0	1	2	0

p(I) = .10

K	N = 20			N = 30			N = 50			N = 100		
4	12	4	10	8	3	9	6	5	13	2	1	6
6	9	9	7	11	7	9	3	3	8	5	6	8
10	8	6	5	10	7	6	8	13	2	7	9	7
20	5	4	7	5	13	4	7	5	0	4	3	2

The figures of this table show that the $\hat{p}(I)$ are low, even for small samples and this is just what we want. That the average number of observations in each class $\frac{N}{K}$ does not have a significant influence on $\hat{p}(I)$ can be seen from the figures of the next table where $\hat{p}(I)$ is shown for a given p(I) = .025.

Table 2

Estimated $p^*(I)$ for the case where the theoretical $p(I) = 0.25$

	$\hat{p}(I)$			
K	N = 20	N = 30	N = 50	N = 100
4	.02	.01	.0	.01
6	.04	.02	.0	.01
10	.03	.01	.02	.03
20	.01	.02	.0	.0

Thus we may conclude that in our case the approximate character of the chi-square distribution does not have serious consequences for the quality of the three tests we used.

5.2. Type-II error

In this section we discuss the power of the three tests we investigate, for different alternative hypotheses. Again a hundred samples were drawn from a normal distribution of mean μ and standard deviation σ. We tested the null-hypothesis H_0: $X \sim N(0, 1)$ against the alternative hypothesis H_1: $X \sim N(\mu, \sigma^2)$.[1] We chose a critical region corresponding to a given $p(I)$ and counted the number of times that the value of the test-statistic fell into the critical region. We did this for different sample sizes and for different numbers of classes. This procedure was repeated for alternative values of $p(I)$ equal to .005, .025, .05, and .1, respectively, and for $\mu = 0,1$ and for $\sigma^2 = .5$, 1.5, and 2.[2]

Let us start by considering the case where $\mu = 0$ and $\sigma^2 = .5$. The first remarkable thing to notice is that the power is very low when the number of observations is not at least equal to 100. Another interesting fact is that the LR-test gives the highest power in almost all cases and that this power reaches its highest value when the number of classes is chosen to be equal to 4 or 6. Table 4 gives the error-probabilities for a special case.

---

[1] No use was made of the fact that the sample was drawn from an $N(\mu, \sigma^2)$-distribution.

[2] Negative values of μ were not considered because testing H_1: $\mu = -\mu_0$, $\sigma^2 = \sigma_0^2$ against H_0: $\mu = 0$, $\sigma^2 = 1$ will in general give the same results as testing H_1: $\mu = \mu_0$, $\sigma^2 = \sigma_0^2$ against H_0: $\mu = 0$, $\sigma^2 = 1$.

Table 3

Error-probabilities of the LR-test with H_0: $\mu = 0$, $\sigma^2 = .5$ where $N = 50$ and $K = 6$

p(I)	$\hat{p}$(II)
.005	.77
.01	.70
.025	.49
.05	.37
.1	.28

We can also use the figures of Table 3 to illustrate the necessity of knowing the power of a test when determining an optimal critical region and to show that it is not enough to confine to the probability of a Type-I error alone.

The most important reason why the power of a test is not taken into consideration when specifying a critical region, is that in many cases the distribution of the test-statistic under the alternative hypothesis and thus the power of the test are not known. If the power of a test is unknown we have no indication for the quality of the test-procedure. Furthermore we cannot be sure to have chosen an optimal value for p(I) as p(I) and p(II) should be determined simultaneously.

In most practical situations one takes p(I) equal to .025 or .05. In Table 3 we see that the corresponding estimated $\hat{p}$(II)'s are .49 and .37 respectively. But we can also choose a p(I) equal to .01 or .1 giving values of $\hat{p}$(II) equal to .70 and .28 respectively. Which combination of p(I) and p(II) should be chosen depends on the relative importance of the errors of both types in each particular situation. This may be described by means of a loss function giving weights to the errors. If these weights are c_1 and c_2 for p(I) and p(II) respectively, the expected loss will be equal to $L = c_1\, p(I) + c_2\, p(II)$. Table 4 shows the various expected losses for alternative values of c_1 and c_2.

Table 4

Expected losses for H_1: $\mu = 0$, $\sigma^2 = .5$ with $N = 50$ and $K = 6$.

p(I)	$\hat{p}$(II)	L		
		$c_1 = \frac{1}{4}$, $c_2 = \frac{3}{4}$	$c_1 = \frac{1}{2}$, $c_2 = \frac{1}{2}$	$c_1 = \frac{3}{4}$, $c_2 = \frac{1}{4}$
.005	.77	.58	.39	.20
.01	.70	.53	.36	.18
.025	.49	.37	.26	.14
.05	.37	.29	.21	.13
.1	.28	.24	.19	.15

From this table we can draw the conclusion that, if the loss-function were as we have specified it and if our decision-criterion is to minimize the expected loss, we have to choose a critical region of size .1 for the case where $c_1 = \frac{1}{4}$, $c_2 = \frac{3}{4}$, $p(I) = .1$ for the case where $c_1 = \frac{1}{2}$ and $c_2 = \frac{1}{2}$, and $p(I) = .05$ for the case where $c_1 = \frac{3}{4}$, $c_2 = \frac{1}{4}$. Notice that in none of these three cases the value of .025 for p(I) (which is often chosen in practical situations) is optimal. In some cases when the distribution of the test-statistic under the alternative hypothesis is known, one often uses a so-called "best" test. This is a test which minimizes p(II) for a given value of p(I) and for a fixed N. Nevertheless one should realize that the concept of a loss function is ignored here, for there is still one degree of freedom: p(I). We believe that in many cases one has at least some intuitive ideas about a loss function. To give an example:

Suppose that a manufacturer tests the quality of the products leaving his factory and that he tests the null-hypothesis H_0: "the quality is good enough to sell the goods" against the alternative H_1: "the quality is too bad". Let us further assume that the test is based on a sample of size N, for a certain lot of goods. We can imagine that a higher p(I) means higher costs. For p(I) is the probability of rejecting H_0 when it is true, in other words: the goods from which the sample is drawn are rejected though the quality is good. This leads to extra costs for the producer. Thus in the above case we can identify p(I) with a kind of risk for the producer. On the other side, in this way of thinking is p(II) to be interpreted as a sort of risk for the consumers of the goods, for p(II) is the probability of accepting H_0 though it is not true, in other words: the goods from which the sample is drawn are accepted and sold though the quality is bad. It is reasonable to assume that a higher risk for the consumers lowers the demand for the goods of our producer and thus we can think of p(II) as a variable in the demand function for the product which moves inversely with the quantity demanded. Thus because p(I) and p(II) cannot be varied independently of each other for a given value of N, c_1 and c_2 will depend on the structure and on the parameters of the demand-function and of the cost-function. No doubt an entrepreneur will often have some idea about his demand-function and about his cost-function and hence about his loss function.

We shall now turn to the case where H_1: $\mu = 0$ and $\sigma^2 = 1.5$. Now the χ^2-test gives the best results in contrast with the previous case (H_1: $\mu = 0$, $\sigma^2 = .5$) where the LR-test yields the highest power. Table 5 illustrates this point clearly.

Table 5

Estimated powers of three tests in the case p(I) = .025 and K = 6

	$\mu = 0, \sigma^2 = .5$			$\mu = 0, \sigma^2 = 1.5$		
N	χ^2	LR	χ^{*2}	χ^2	LR	χ^{*2}
20	.03	.09	.06	.26	.11	.05
30	.08	.21	.09	.33	.15	.07
50	.24	.51	.24	.41	.17	.15
100	.74	.85	.63	.57	.33	.23

If we compare the figures under H_1: $\mu = 0$, $\sigma^2 = .5$ with those under H_1: $\mu = 0$, $\sigma^2 = 1.5$ we notice that the change of σ^2 from .5 to 1.5 rises the power of the χ^2-test and lowers that of the LR-test with only two exceptions: for N = 20 where the power of the LR-test rises with .02 and for N = 100 where the power of the χ^2-test falls considerably. The power of the χ^{*2}-test shows the same pattern as that of the LR-test, that is to say: it falls when σ^2 is increased from .5 to 1.5.

Table 6

Error-probabilities for the χ^2-test in H_1: $\mu = 0$, $\sigma^2 = 1.5$ where N = 50 and K = 6

p(I)	$\hat{p}$(II)
.005	.74
.01	.70
.025	.59
.05	.53
.1	.44

We discover, in comparing the figures with those of the first two columns of Table 4 which gives the error-probabilities of the LR-test, that for H_1: $\mu = 0$, $\sigma^2 = 1.5$, the power is higher or the same for all values of p(I) except for p(I) = .005. Nevertheless it will be clear that the power is very low, too low for practical applications, even if N = 100, as becomes clear from Table 7.

Table 7

Error-probabilities of the χ^2-test with H_1: $\mu = 0$, $\sigma^2 = 1.5$ where N = 100 and K = 6.

p(I)	$\hat{p}$(II)
.005	.59
.01	.54
.025	.43
.05	.37
.1	.30

As is to be expected the quality of the tests rises considerably when we have an alternative hypothesis H_1: $\mu = 0$, $\sigma^2 = 2.0$. Again the χ^2-test turns out to be the best . The $\hat{p}(II)$ are lowest for K = 6 or 10. A sample consisting of 50 observations gives a rather good power. Even in the case when N = 30 we get a power which is much higher than that connected with the two alternative hypotheses discussed above, with an N equal to 100. The following table may illustrate this point.

Table 8

Error-probabilities of the χ^2-test with H_1: $\mu = 0$, $\sigma^2 = 2$ where K = 6.

	N = 30	N = 50
p(I)	$\hat{p}$(II)	$\hat{p}$(II)
.005	.37	.25
.01	.28	.22
.025	.23	.12
.05	.20	.07
.1	.19	.05

It may be interesting to consider the same kind of table for 10 instead of 6 classes:

Table 9

Error-probabilities of the χ^2-test with H_1: $\mu = 0$, $\sigma^2 = 2$ where K = 10

	N = 30	N = 50
p(I)	$\hat{p}$(II)	$\hat{p}$(II)
.005	.37	.22
.01	.32	.14
.025	.27	.09
.05	.21	.09
.1	.13	.07

If we take K equal to 20 we get the same picture as above, hence we conclude that it does not make much difference for the power of the test if we divide the range into 6, 10 or 20 classes.

Let us now consider an alternative hypothesis with $\mu = 1$ and $\sigma^2 = .5$. We again show the error-probabilities in the following table.

Table 10

Error-probabilities of the χ^2-test with H_1: $\mu = 1$, $\sigma^2 = .5$ where $K = 6$

	N = 20	N = 30	N = 50
p(I)	$\hat{p}$(II)	$\hat{p}$(II)	$\hat{p}$(II)
.005	.09	.01	.0
.01	.05	.0	.0
.025	.03	.0	.0
.05	.01	.0	.0
.1	.01	.0	.0

In this case it is not necessary to have more than 20 observations in order to obtain an acceptable power. We also noticed that a different number of classes chosen will hardly change the error-probabilities, except for $K = 4$ which gives worse results. For table 10 we have again chosen the chi-square test because it remains a little better than the other two tests, though the difference in quality is no longer very important.

It is interesting to compare Table 10 with Table 2 where H_1 was $\mu = 0$ and $\sigma^2 = .5$. The variances of the two cases are the same but the mean has been shifted to the right with a unit. We see that the power is increased by this shift. This was to be expected because, in Table 10, besides the variance also the mean differs from H_0.

The next hypothesis which we want to discuss is H_1: $\mu = 1$, $\sigma^2 = 1$ and differs from its predecessor only in the variance. That this change in variance of the alternative hypothesis has only a minor influence on p(II) may be seen by comparing the figures of Table 10 with those of the following table.

Table 11

Error-probabilities of the χ^2-test with H_1: $\mu = 1$, $\sigma^2 = 1$ where $K = 10$

	N = 20	N = 30
p(I)	$\hat{p}$(II)	$\hat{p}$(II)
.005	.11	.02
.01	.09	.01
.025	.06	.01
.05	.06	.01
.1	.06	.0

We had to choose the χ^2-test again because it is still slightly better than the other two tests. The number of classes should be taken equal to 6, 10 or 20. For practical purposes the differences between the $\hat{p}(II)$ in Table 10 and those in Table 11 may be neglected. Thus we see that in our case, where the mean of the distribution under H_1 differs from that of the distribution under H_0 by a whole unit, the power of the χ^2-test is rather robust against a change in the variance of H_1.

The next table deals with a H_1 with $\mu = 1$ but now we take σ^2 larger than 1, in fact equal to 1.5.

Table 12

Error-probabilities of the χ^2-test with H_1: $\mu = 1$, $\sigma^2 = 1.5$ where K = 10.

	N = 20	N = 30
p(I)	$\hat{p}$(II)	$\hat{p}$(II)
.005	.09	.02
.01	.11	.01
.025	.07	.01
.05	.04	.00
.1	.05	.00

Here we also take the χ^2-test because it is superior to the other two tests and we see that a sample size equal to 20 is sufficient. As to the number of classes to be chosen, the worst we can do is to take K equal to 4. It does not make much difference, however, whether we take 6, 10 or 20 classes. The $\hat{p}(II)$ are again hardly influenced by the change of the variance. This strengthens our conclusion about the insensibility of the χ^2-test to changes in the variance of H_1, which was drawn in the previous paragraph. We may illustrate this by means of the following table where we compare four different cases.

Table 13

Error-probabilities of the χ^2-test where N = 20 and K = 10

	$\hat{p}$(II)			
p(I)	$\mu = 1, \sigma^2 = .5$	$\mu = 1, \sigma^2 = 1$	$\mu = 1, \sigma^2 = 1.5$	$\mu = 1, \sigma^2 = 2$
.005	.07	.11	.09	.08
.01	.06	.09	.11	.08
.025	.03	.06	.07	.06
.05	.03	.06	.04	.05
.1	.01	.06	.05	.03

The last column of Table 13 gives the error-probabilities for a H_1: $\mu = 1$, and $\sigma^2 = 2$, the last case which is investigated. Again it does not make much difference whether we take K equal to 6, 10 or 20 and the chi-square test yields the best error-probabilities. A sample size of N = 20 proves to be sufficient.

5.3. Summary of the Results

Resuming our results we can say that the LR-test had a higher power than the χ^2-test and than the χ^{*2}-test in only one case of the 7 we investigated, namely that with $\mu = 0$ and $\sigma^2 = .05$. In all other cases the χ^2-test dominates the other two. tests. Next we give the following table.

Table 14

H_1		N	K	Test to be used	$\frac{1}{5}\sum_{i=1}^{5} \hat{p}(II)$		
μ	σ^2				N	K	
0	.5	-	4 or 6	LR	100	6	.52
0	1.5	-	6 or 10	χ^2	50	6	.45
0	2	50	4	χ^2	50	6	.14
1	.5	20	4	χ^2	20	6	.04
1	1	20	4	χ^2	20	10	.08
1	1.5	20 or 30	4	χ^2	20	10	.07
1	2	20 or 30	4	χ^2	20	10	.06

where the first two columns specify the alternative hypothesis. The third column gives the number of observations required to obtain a satisfactory power.[1] The 4th and 5th column show the number of classes to be chosen and the test to be used respectively. To give an idea of the qualities of the tests the last column gives the average power (averaged over the 5 p(I)'s that are considered) for values of N and K specified in the two previous columns.

Usually, in econometrics, we do not have more than 20 or 30 observations. But from Table 14 we learn that for N equal to 20 or 30, the χ^2-test has only a high power if the "true" distribution differs from the hypothetical distribution in location. If these two distributions have the same mean the results of the χ^2-test are very bad.

[1] The first two places of column 3 are empty because an N equal to 100 (the highest we observed) did not yet give an acceptable power.

The fourth column of Table 14 shows that in the relevant situations one should take K larger than 4. Our results indicate that it is much more dangerous to choose K too small than to choose K too large.

Of course our conclusions only hold for normal distributions, for our maintained hypothesis was: F(x) is normal. But it is not probable that choosing a different shape for the distribution of X will yield better results. Another aspect we did not investigate is the influence on the power of first estimating the mean and the variance of the distribution from the sample, before specifying the null-hypothesis. But here also we may expect that estimation will not improve our results. In the opposite, we believe that the powers will be much lower, especially when the mean is estimated accurately.

Nevertheless a next paper will be devoted to the effect of estimation on the power and we shall use there non-normal distributions.