On approximation of the distribution for Pearson statistic

Nikolai Dokuchaev

Abstract—The paper considers the classical Goodness of Fit test. It suggests to use the Gamma distribution for the approximation of the distribution of the Pearson statistics with unknown parameters estimated from raw data. The parameters of these Gamma distribution can be estimated from the first two moments of the statistic after averaging over a distribution of the unknown parameter over its range. This allows to simplify calculation of the quantiles for the Pearson statistic, as is shown in some simulation experiments with medium and small sample sizes.

Keywords: goodness of fit test, Pearson statistic, probability distributions

MSC classification: 62F03, 62G05, 62G10.

I. INTRODUCTION

The classical statistical Goodness of Fit test addresses the problem of estimation the parameters of a parametric family of distributions from observed data with unknown $d$-dimensional parameter that has to be fitted from the data. The Pearson statistics is commonly used to estimate the error; see, e.g., the literature review in [1, 2, 3, 4, 5]. Let $n$ be the number in intervals where the observations are counted in the Pearson statistic. The limit distribution of this statistics for infinitely increasing sample size is known, given some mild conditions; see, e.g. [1, 2, 3, 4, 5]. The quantiles for its limit distribution are often used as the critical values for the test. If the parameter is fitted from the raw (ungrouped) data using a consistent estimator, then the limit distribution is different; see, e.g. [1], p.24. The actual distribution of statistic for the finite samples is a discreet distribution and depends on the choice of the counting intervals and other parameters of the experiment.

This sort paper suggests to use the Gamma distribution for a simplified approximation of the distribution of the Pearson statistic with small and medium sample sizes. The parameters of these Gamma distribution can be estimated from the first two moments of the sample distribution of simulated Pearson statistic with parameter values randomized over a domain for the unknown true parameter. Some computer experiments with medium and small sample sizes shows that this helps to reduce the bias for the calculation of the quantiles for the Pearson statistic.

II. PROBLEM SETTING

Let $F(\cdot|\theta)|_{\theta \in D}$ be a give family of distributions, where $\theta \in D$ is a $d$-dimensional parameter, and where $D \subset \mathbb{R}^d$ is a domain.

Assume that we are testing a hypothesis about a population distribution for a given independent and identically distributed sample $X = (X_1, ..., X_N)$ from the distribution $F(\cdot|\theta_0)$, where $\theta_0 \in D$. Let $\hat{\theta}$ be the estimate of $\theta$ obtained using a consistent estimator $\hat{\theta} = T(X)$, where $T : \mathbb{R}^N \rightarrow \mathbb{R}^d$ is a mapping. We assume that $\hat{\theta}$ is observable.

For a given integer $n > 0$, consider a system of mutually disjoint intervals $\{I_i\}_{i=1}^n$ such that $\cup_i I_i = \mathbb{R}$ (two of these intervals are semi-infinite). Let $C_i$ be observed sample counts in the intervals $I_i$, calculated from a sample $x_1, ..., x_N$. In particular, we have that

$$\sum_{i=1}^n C_i = N.$$ 

The values $C_i$ are supposed to be observable.

Consider two hypotheses:

$H_0$: The sample come from the distribution $F(\cdot|\hat{\theta})$.

$H_A$: The sample does not come from this distribution.

A hypothesis has to be accepted or rejected based on observed $\hat{\theta}$ and observed counts $\{C_i\}$, given that the family of the distributions $F(\cdot|\theta)|_{\theta \in D}$, and the domain $D$ of possible values of $\theta$ are known.

Let $Q_i \overset{\Delta}{=} \mathbb{E}[C_i|\theta = \hat{\theta}]$.

Let us consider Pearson’s statistic

$$\chi^2 \overset{\Delta}{=} \sum_{i=1}^n \frac{(C_i - Q_i)^2}{Q_i}.$$ 

If the computed value of $\chi^2$ is large, then we reject hypothesis $H_0$. In this case, the observed and expected values are not close and the model is a poor fit to the data.

If the parameter is fitted from the grouped data using a consistent estimator based on counting in the intervals, then the limit distribution is a known $\chi^2_{n-d-1}$-distribution, given some mild conditions; see, e.g. [1, 2, 3, 4, 5]. If the parameter is fitted from the raw (ungrouped) data using a consistent estimator, then the limit distribution is

$$\chi^2_{n-d-1} + \sum_{k=n-d}^{n-1} \nu_k Z_k,$$

(1)

where $Z_k$ are independent standard normal variables and $\nu_k \in [0, 1]$ (Chernoff and Lehmann [3]; see also [1], p.24). However, the values $\{\nu_k\}$ depend on the intervals, on the population distribution, and on the estimator.

The distribution of $\chi^2$ is discreet and depends on the choice of $(F(\cdot), D, \theta_0, \{I_i\}_{i=1}^n, T(\cdot))$. The standard approach for the approximation of the distribution of $\chi^2$ for large $N$ is its approximation $\chi^2_{n-1-d}$ distribution, i.e., by the $\chi^2$-distribution.
with \( n - 1 - d \) degrees of freedom (see, e.g., [2]). In the literature, \( X^2 \) is called chi square statistic or Pearson’s statistic. This \( \chi^2_{n-1-d} \) distribution is independent on the choice of the set \( \{I_i\} \). However, the actual distribution of \( X^2 \) is not easy to describe; it depends on \( (F(\cdot), D, \theta_0, \{I_i\}, T(\cdot)) \). Therefore it is not easy to calculate quantiles used for the hypothesis testing. On the other hand, some numerical examples given below show that use of quantiles for the \( \chi^2_{n-1-d} \) distribution as a substitution for quantiles of \( X^2 \) could lead to significant bias for the critical values.

The distribution of \( X^2 \) is discreet and depends on the choice of \( \{n, \{I_i\}_{i=1}^n, F(\cdot|\theta)\} \). It is known that, under some mild assumptions,

\[
\text{the distribution of } X^2 \text{ converges to } \chi^2_{n-1-d} \quad \text{as } N \to +\infty.
\]

(See, e.g., [2]). The limit distribution here is the \( \chi^2 \)-distribution with \( n - 1 - d \) degrees of freedom. In the literature, \( X^2 \) is called chi square statistic or Pearson’s statistic. This limit distribution is independent on \( (F(\cdot), D, \theta_0, \{I_i\}, T(\cdot)) \). However, the actual distribution of \( X^2 \) is not easy to describe for a given finite \( n \). Therefore it is not easy to calculate quantiles for the hypothesis testing. On the other hand, some numerical examples given below show that use of quantiles for the \( \chi^2_{n-1-d} \) distribution as a substitution for quantiles of \( X^2 \) could lead to significant bias for the critical values.

There are several known approaches to deal with this bias; see, e.g. [4]. We suggest one more approach that seems to provide a reasonably close approximation for the distribution of the tests statistics with medium and small sample sizes.

### III. APPROXIMATION BY GAMMA DISTRIBUTION

In some numerical experiments, we have found that the Gamma distribution \( \Gamma(\alpha, \lambda) \) with the density

\[
I_{\{x>0\}} \Gamma(\alpha) \Gamma^{-1}(\alpha) x^{-\alpha-1} e^{-\lambda x}
\]

can be effectively used as a close approximation of the distribution of \( X^2 \).

For the distribution \( \Gamma(\alpha, \lambda) \), the expectation is \( \alpha / \lambda \), and the variance is \( \alpha^2 / \lambda \).

Technically, the distribution of \( X^2 \) and as well as \( (\alpha, \lambda) \) depend on the choice of \( (F(\cdot), D, \theta_0, \{I_i\}_{i=1}^n, T(\cdot)) \). However, we need an approximation that does not use \( \theta_0 \). Therefore, we suggest to estimate these parameters via matching them with the first two moments of a sample of random values \( X^2 \) simulated under the compound distribution \( F(\cdot|\theta) \) with a random \( \Theta \) given some preselected probability distribution for \( \Theta \). This removes dependence of \( (\alpha, \lambda) \) on the true parameter \( \theta \). For example, one can use a non-informative uniform distribution over a bounded domain \( D \) containing the true parameter \( \theta \).

Let \( E \) and \( \sqrt{ar} \) be the sample mean and the sample variance, respectively, over the Monte-Carlo trials for simulation of \( (\Theta, X) \) generating the implied statistic \( X^2 \).

**The procedure for fitting \( (\alpha, \lambda) \):**

(i) Run \( M \) Monte-Carlo simulations of \( \Theta \). For each simulated \( \Theta \), simulate an i.i.d. sample \( X = (X_1, ..., X_N) \) with the terms distributed under \( F(\cdot|\Theta) \).
(ii) Calculate \( X^2 \) for each simulation of \( (\Theta, X) \).
(iii) Calculate \( a = E X^2 \) and \( v = \sqrt{ar} X^2 \).
(iv) Find \( \alpha \) and \( \lambda \) such that

\[
\frac{\alpha}{\lambda} = E X^2, \quad \frac{\alpha}{\lambda^2} = \sqrt{ar} X^2,
\]

i.e.

\[
\alpha = \frac{(E X^2)^2}{\sqrt{ar} X^2}, \quad \lambda = \frac{E X^2}{\sqrt{ar} X^2}.
\]

(v) Use quantiles for \( \Gamma(\alpha, \lambda) \) as approximations for quantiles for \( X^2 \).

It seems that this approach allows to achieve a significant reduction of the bias for quantiles for the sample sizes.

### IV. NUMERICAL EXAMPLES

Let illustrate the difference between the limit distribution and actual distribution of \( X^2 \) using the following numerical example.

This would correspond to the setting with \( d = 1 \) and \( n - d - 1 = 1 \).

We run Monte-Carlo experiments with the sample size \( M = 10^6 \) for \( X^2 \). We run these experiments for four cases with different sets of parameters. These cases are listed below.

**Case A:**

For this case, we simulated \( X^2 \) for the sample \( X \) from exponential distribution \( \text{Exp}(\theta_0) \), i.e. with the density

\[
I_{\{x>0\}} \theta_0 e^{-\theta_0 x}
\]

This corresponds to the case of non-random \( \Theta = \theta_0 \). We have used \( \theta_0 = 1 \), and we have used the estimate \( \hat{\theta} = T(X) = 1/X \), where \( X \sim \frac{1}{N} \sum_{i=1}^N X_i \). This is a maximum likelihood estimate as well as the estimate implied by the method of moments. It is known that this estimate is consistent.

**Case A(i):** The sample size for the underlying process \( X \) is \( N = 10 \), the number of intervals is \( n = 3 \), and \( I_1 = (-\infty, a_1] \), \( I_2 = (a_1, a_2] \), \( I_3 = [a_2, \infty) \). The numbers \( a_1 < a_2 \) are such that \( P(X_k \in I_k|\theta_0) = 1/3 \). This choice corresponds to the most basic case where of equal probabilities for the intervals. For this case, we found in the experiments with \( 10^6 \) Monte-Carlo trials that \( E X^2 = 1.35800 \) and \( \sqrt{ar} X^2 = 2.0845822 \).

As can be seen, it is quite far for the expectation and the variance for the \( \chi^2_{n-1-d} = \chi^2_2 \) distribution; these parameters are \( n - d - 1 = 1 \) and \( 2(n - d - 1) = 2 \) respectively.

**Case A(ii):** The sample size for the underlying process \( X \) is \( N = 1000 \); the remaining parameters are the same as for Case A(i). For Case A(ii), we found in the experiments with \( 10^6 \) Monte-Carlo trials that \( E X^2 = 1.350675 \) and \( \sqrt{ar} X^2 = 2.245898 \).

Table [1] shows sample quantiles for \( X^2 \) for Cases A(i)-(ii). This example shows that use of quantiles for the limit distribution as a substitution for quantiles of \( X^2 \) for finite samples could lead to a bias. Approximation by Gamma function helps to reduce the bias, as is shown in examples described below.

We have also considered cases where the parameters \( (\alpha, \lambda) \) have been fitted to the sample \( X^2 \) simulated according to the procedure describes above.

**Case B:** For this case, we consider the family the exponential distribution \( \text{Exp}(\theta) \) with the density \( I_{\{x>0\}} \theta^{-1} e^{-\theta x} \). We assumed that \( \theta \in D = [0.5, 1.5] \). For the step (i) of
Table II(i) shows quantiles for $X$ in Case A(i); the sample size for $X$ is $N = 10$

| Quantiles | 0.75 | 0.95 | 0.99 |
|-----------|------|------|------|
| $X^2$     | 5.09 | 8.33 | 11.36 |

Case A(ii): the sample size for $X$ is $N = 1000$

| Quantiles | 0.75 | 0.95 | 0.99 |
|-----------|------|------|------|
| $X^2$     | 1.801390 | 3.052967 | 4.146487 |
|           | 6.279877 | 7.753290 | 11.07638 |

TABLE I

QUANTILES FOR CASES A(I)-A(II).

For this procedure, we have used $\Theta$ uniformly distributed on the domain $D = [0.2, 2]$. Further, the sample size for the underlying process $X$ for this case is $N = 20$, the number of intervals is $n = 3$, and $I_1 = (-\infty, 0.5]$, $I_2 = (0.5, 1.5]$, and $I_3 = [1.5, \infty)$. For this case, we have $\mathbb{E}X^2 = 1.323495$, $\mathbb{Var}X^2 = 2.142576$, and the corresponding parameters for $\Gamma(\alpha, \lambda)$ are

$\alpha = 0.8175386$, $\lambda = 0.617712$.

Table II(ii) shows quantiles for $X^2$, for the fitted distribution $\Gamma(\alpha, \lambda)$, and for the limit distribution $\chi^2_{n-d-1}$.

Case C: For this case, we consider $I_1 = (-\infty, 1]$, $I_2 = (1, 2]$, and $I_3 = [2, \infty)$. All other parameters are the same as for Case B.

For this case, we have $\mathbb{E}X^2 = 1.294582$, $\mathbb{Var}X^2 = 2.183124$, and the corresponding parameters for $\Gamma(\alpha, \lambda)$ are

$\alpha = 0.7676803$, $\lambda = 0.5929949$.

Table II(iii) shows quantiles for $X^2$, for the fitted distribution $\Gamma(\alpha, \lambda)$, and for the limit distribution $\chi^2_{n-d-1}$.

Case D: For this case, we consider $N = 1000$. All other parameters are the same as for Case C.

For this case, we have $\mathbb{E}X^2 = 1.281905$, $\mathbb{Var}X^2 = 2.191206$, and the corresponding parameters for $\Gamma(\alpha, \lambda)$ are

$\alpha = 0.7499429$, $\lambda = 0.5850224$.

Table II(iv) shows quantiles for $X^2$, for the fitted distribution $\Gamma(\alpha, \lambda)$, and for the limit distribution $\chi^2_{n-d-1}$.

Case E: For this case, we consider the family of normal distributions $N(\mu, \sigma^2)$ with $\mu \in [-0.5, 0.5]$ and $\sigma \in [1, 2]$ and $\theta = (\mu, \sigma)$. The random parameter $\Theta$ as a random vector with independent components distributed uniformly on $[-0.5, 0.5]$ and $[1, 2]$ respectively. We used $(\hat{\mu}, \hat{\sigma}) = T(X)$ such that $\hat{\mu}$ is the sample mean of $X$ and $\hat{\sigma}^2$ is the sample variance of $X$. The number of intervals is $n = 4$, and the intervals are $I_1 = (-\infty, -1]$, $I_2 = (-1, 0]$, $I_3 = (0, 1]$, and $I_4 = (1, \infty)$. For this case, we have $\mathbb{E}X^2 = 1.772562$, $\mathbb{Var}X^2 = 2.852873$, and the corresponding parameters for $\Gamma(\alpha, \lambda)$ are

$\alpha = 1.101338$, $\lambda = 0.621325$.

Table II(v) shows quantiles for $X^2$ and for the fitted distribution $\Gamma(\alpha, \lambda)$.

Case F: For this case, we consider the family the normal distributions $N(\mu, \sigma^2)$ with $\mu \in [-1, 1]$ and $\sigma \in [0.5, 4]$ and $\theta = (\mu, \sigma)$. The random parameter $\Theta$ as a random vector with independent components distributed uniformly on $[-1, 1]$ and $[0.5, 4]$ respectively. The intervals and the estimators are the same as in Case E. $I_1 = (-\infty, -1]$, $I_2 = (-1, 0]$, $I_3 = (0, 1]$, and $I_4 = (1, \infty)$.

For this case, we have $\mathbb{E}X^2 = 1.922529$, $\mathbb{Var}X^2 = 3.251841$, and the corresponding parameters for $\Gamma(\alpha, \lambda)$ are

$\alpha = 1.136623$, $\lambda = 0.5912125$.

Figures 1-2 show smoothed histograms for $X^2$ and $\Gamma(\alpha, \lambda)$ for Cases D,E, and F, respectively, constructed from the histograms for Monte-Carlo samples of the size $M = 10^6$ using the standard command $\text{densities}$ in R programming language.

These figures demonstrate quite close approximation.

We have used R programming language for calculations; calculation of $(\alpha, \lambda)$ for $N = 20$ and $M = 10^6$ takes less than a minute for a standard desktop computer. For $N = 1000$ and $M = 10^6$, it takes about 10 minutes.

V. CONCLUSION

The paper suggest to approximate the distribution of the Pearson statistic by the Gamma distributions with parameters fitted to simulated statistics a given configuration of cells where the sample occurrences are being counted. Feasibility of this approach is demonstrate with some numerical experiments. So far, the range of the parameters for these experiment was quite limited. It would be interesting to extend these experiments on more general choices of the parameters, especially $n$ and $N$. We leave this for the future research.

REFERENCES

[1] Balakrishnan, N., Voinov, V., Nikulin M. S. (2013). Chi-Squared Goodness of Fit Tests With Applications, Academic Press.

[2] Birch, M. W. (1964). A New Proof of the Pearson-Fisher Theorem. Ann. Math. Statist., 35, No. 2, 817-824.

[3] Chernoff, H., Lehmann, E.L. (1954). The use of maximum likelihood estimates in tests for goodness of fit. The Annals of Mathematical Statistics 25, 579-589.

[4] Greenwood, C., Nikulin, M. S. (1996). A guide to chi-squared testing, New York: Wiley.

[5] Plackett, R.L. (1983). Karl Pearson and the Chi-Squared Test. International Statistical Review, 51, 59-72.
(i) Quantiles for Case B; the sample size for $X$ is $N = 20$

| Quantiles | 0.75 | 0.9 | 0.95 | 0.99 |
|-----------|------|-----|------|------|
| For $X^2$ | 1.787257 | 3.111514 | 4.296282 | 6.762272 |
| For fitted $\Gamma(\alpha, \beta)$ | 1.831157 | 3.204561 | 4.262158 | 6.760412 |

(ii) Quantiles for Case C; the sample size for $X$ is $N = 20$

| Quantiles | 0.75 | 0.9 | 0.95 | 0.99 |
|-----------|------|-----|------|------|
| For $X^2$ | 1.690018 | 2.993245 | 4.200237 | 6.907514 |
| For fitted $\Gamma(\alpha, \beta)$ | 1.781864 | 3.178560 | 4.257779 | 6.811008 |

(iii) Quantiles for Case D; the sample size for $X$ is $N = 1000$

| Quantiles | 0.75 | 0.9 | 0.95 | 0.99 |
|-----------|------|-----|------|------|
| For $X^2$ | 1.707008 | 3.094988 | 4.231155 | 7.007469 |
| For fitted $\Gamma(\alpha, \beta)$ | 1.765294 | 3.161543 | 4.250052 | 6.850396 |

(iv) Quantiles for Case E; the sample size for $X$ is $N = 1000$

| Quantiles | 0.75 | 0.9 | 0.95 | 0.99 |
|-----------|------|-----|------|------|
| For $X^2$ | 2.397755 | 3.939819 | 5.124125 | 7.927041 |
| For fitted $\Gamma(\alpha, \beta)$ | 2.453660 | 3.982921 | 5.12967 | 7.796123 |

(v) Quantiles for Case F; the sample size for $X$ is $N = 1000$

| Quantiles | 0.75 | 0.9 | 0.95 | 0.99 |
|-----------|------|-----|------|------|
| For $X^2$ | 2.602056 | 4.228039 | 5.479738 | 8.478611 |
| For fitted $\Gamma(\alpha, \beta)$ | 2.656888 | 4.284709 | 5.499895 | 8.278232 |

TABLE II
QUANTILES FOR CASES B, C, D.