310/2004

Raport Badawczy

Research Report

RB/8/2004

# Statistical Tests for Comparing Pareto Charts

P. Grzegorzewski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

**POLSKA AKADEMIA NAUK**

**Instytut Badań Systemowych**

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 8373578

fax: (+48) (22) 8372772

Kierownik Pracowni zgłaszający pracę:
Prof. dr hab. inż. Olgierd Hryniewicz

Warszawa 2004

# Statistical Tests for Comparing Pareto Charts

Przemysław Grzegorzewski

Systems Research Institute, Polish Academy of Sciences

Newelska 6, 01-447 Warsaw, Poland

e-mail: pgrzeg@ibspan.waw.pl

**Abstract**

A statistical test for comparing Pareto charts is suggested. This new test, based on the Shannon entropy, is more general than the procedures applied so far.

**Key words:** entropy, hypotheses testing, Pareto chart, statistical quality control.

## 1 Introduction

One of the fundamental ideas of the Statistical Process Control (SPC) is an empirical observation of the Italian sociologist Vilfredo Pareto stating that "in a system, a relatively few failure reasons are responsible for the catastrophically many failures" (see Thompson and Koronacki, 1993). The Pareto chart is one of the most useful of the "magnificent seven", i.e. of the seven major SPC problem-solving tools. Through this chart a user can quickly and visually identify the most frequently occurring types of defects, failures delays etc. It should be stressed that the Pareto chart does not automatically identify the most important defects but rather only those that occur most frequently (see Montgomery, 1991).

The Pareto chart is just a bar chart for attribute data arranged by a category in which the categories are ordered by the number of occurrences. Additionally, a polygon corresponding to cumulative frequency distribution (ogive) is often also attached. A typical Pareto chart is shown in Figure 1.

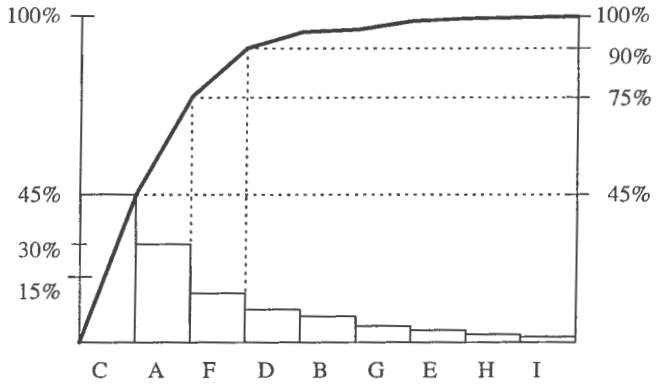

Figure 1: The Pareto chart for nine defect categories

Visual inspection of the chart is a first step in the data analysis and the Pareto principle typically provides much of the insight. Additional analysis can yield more knowledge. In particular the incorporation of statistical tests would help to distinguish the randomness due to "common" causes and due to "assignable" causes.

In more advanced Pareto analysis is sometimes important to compare given Pareto chart to a "standard" one. Moreover, sometimes one would like to decide whether there is a significant difference between two Pareto charts, e.g. to compare given Pareto chart with another Pareto chart constructed over a different time, period or process. This problem was considered by Kenett (1991). However, his test could be applied for Pareto charts with the same defect categories only.

In the present paper we propose a statistical test for comparing two Pareto charts based on the Shannon entropy. Our approach enables to omit the restrictions of the Kenett procedure, mentioned above.

The paper is organized as follows: In Sec. 2 we introduce the notation, while in Sec. 3 we express the main problem of this contribution. Next we recall basic information on the Shannon entropy (Sec. 4). Then we propose a statistical test for comparing Pareto charts with the same number of cate-

gories (Sec. 5.1), a test for comparing Pareto charts with different number of categories (Sec. 5.2) and a test for comparing Pareto chart with a "standard" one (Sec. 5.3). The suggested tests are illustrated by examples.

## 2 Notation

Let $S = \{S_1, \ldots, S_m\}$ denote a set of categories, e.g. corresponding to different defects, and let $P = \{p_1, \ldots, p_m\}$, where $p_i \geq 0$, $\sum_{i=1}^{m} p_i = 1$, denote a probability distribution defined on $S$, i.e. $p_i = \Pr(S_i)$.

Suppose $o_1, \ldots, o_n$ is a set of observed defects, i.e. $o_j \in S$ for each $j = 1, \ldots, n$. Moreover, let

$$X_i = \#\{o_j : o_j \in S_i,\ j = 1, \ldots, n\}. \tag{1}$$

stand for the number of observations belonging to $i$-th category.

It is easily seen that each Pareto chart could be identified with a set of frequencies $X_1, \ldots, X_m$ while a theoretical discrete distribution corresponding to that chart by an ordered pair $(S, P)$. Therefore, for brief, we will write $X_1, \ldots, X_m \sim (S, P)$.

## 3 Goodness-of-fit tests for comparing Pareto charts

Suppose we have two Pareto charts with identical categories $S = \{S_1, \ldots, S_m\}$, based on $n_1$ and $n_2$ observations, respectively, i.e. we have $X_1, \ldots, X_m \sim (S, P)$ and $Y_1, \ldots, Y_m \sim (S, Q)$, where $\sum_{i=1}^{m} X_i = n_1$ and $\sum_{i=1}^{m} Y_i = n_2$. We are interested whether there is a significant difference between these two Pareto charts. This problem comes down to the following statistical hypothesis testing problem:

$$\begin{aligned} H &: P = Q, \\ K &: P \neq Q, \end{aligned} \tag{2}$$

that two underlying distributions are equal against the are not equal. Thus, in fact, we face the goodness-of-fit testing problem for two discrete distributions.

To verify hypothesis stated above one can apply well-known goodness-of-fit chi-squared test. Fuchs and Kennet (1980) suggested another method, called the $M$-test, and showed that their test has larger asymptotic power than the chi-squared test in the presence of outliers, for moderate and large number of categories. Kenett (1991) also showed how to apply the $M$-test for comparing Pareto charts.

The $M$-test enables to compare a given Pareto chart to a "standard" one with the same defect categories. Unfortunately, these assumptions make a strong limitation for possible applications. Sometimes it would be desirable to compare two Pareto charts which differ in categories. Moreover, it may happen that none of the charts under study could be distinguished as a "standard" one. Hence we propose another goodness-of-fit technique free from these assumptions. We consider both situations where categories are different but the numbers of categories in two charts are equal, and the more general situation where the numbers of categories may also differ. Our method utilizes the notion of the entropy described in the next section.

The need to compare two Pareto charts with identical categories seems to be quite obvious – e.g. we may be interested in veryfying whether undertaken corrective actions have caused a desired improvement. However, one may ask what is the reason for comparing Pareto charts with nonidentical categories because in such a case thay are noncomparable. Although it is – in some sense – true, we may be interested in the comparison of the shape of two Pareto charts even abstracting from the underlying categories. For example, we may ask whether the distribution of undesirable effects that appear in two distinct sections are "more or less" similar or one distribution is more flat (uniform) while the other is more steep and so on.

# 4 Entropy

Shannon (1948) introduced the entropy of a random variable as a measure of information and uncertainty. Now the entropy is a characteristic playing a fundamental role not only in information theory and communication, but also in classification, pattern recognition, statistical physics, stochastic dynamics, statistics, etc. Just in statistics Shannon's entropy is used as a descriptive pa-

rameter (measure of dispersion), for testing normality (Vasicek, 1976; Arizono and Ohta, 1989), exponentiality (Grzegorzewski and Wieczorkowski, 1998) and uniformity (Dudewicz, et al., 1995). Goodness-of -fit test for uniformity based on the sample entropy are extensively utilized in evaluating random number generators. One also know the maximum entropy principle applicable widely in problems of inference on the basis of incomplete data.

Consider a discrete random variable $(S, P)$ assuming values from $S = \{S_1, \ldots, S_m\}$ with corresponding probabilities $P = \{p_1, \ldots, p_m\}$, where $p_i = \Pr(S_i)$. The quantity $H(P)$ is defined as

$$H(P) = -\sum_{i=1}^{m} p_i \log p_i \tag{3}$$

is called the Shannon entropy of given discrete random variable.

We adopt the convention that $0 \log 0 = 0$. The negative sign in (3) makes the entropy nonnegative and allows the logarithm to be taken with arbitrary base greater than one. When a base 2 logarithm is used, the unit of entropy is called a *bit* (binary digit). Due to this assumption the entropy of a random variable that assumes two values $S_1$ and $S_2$ with equal probabilities of $\frac{1}{2}$ is

$$H(P) = -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = 1\ [bit].$$

By the tradition, we also assume that $\log = \log_2$.

It can be shown easily that for a random variable $(S, P)$ such that $\#S = m$ we have

$$0 \leq H(P) \leq \log m.$$

Let us notice that the entropy of a discrete random variable depends only on the number of values and their probabilities and does not depend on the values themselves. This property makes the entropy a tool desired for comparing Pareto charts with nonidentical defect categories.

However, since the entropy depends on the number of values, while comparing Pareto charts with different number of categories we cannot apply the entropy defined by (3) but its standardized version $H_S(P)$, called a standardized entropy, and defined as follows

$$H_S(P) = \frac{H(P)}{\log m}. \tag{4}$$

It is obvious that for any discrete distribution $(S, P)$ we have

$$0 \leq H_S(P) \leq 1.$$

# 5 Comparing Pareto charts via entropy-based test

## 5.1 Comparing Pareto charts with the same number of categories

Suppose now we have two Pareto charts with not necessarily identical categories $S = \{S_1, \ldots, S_m\}$ and $R = \{R_1, \ldots, R_m\}$, based on $n_1$ and $n_2$ observations, respectively, i.e. we have $X_1, \ldots, X_m \sim (S, P)$ and $Y_1, \ldots, Y_m \sim (R, Q)$, where $\sum_{i=1}^m X_i = n_1$ and $\sum_{i=1}^m Y_i = n_2$. We are interested if these two Pareto charts are significantly different. However now, instead of (2) we would rather consider a following statistical hypothesis testing problem

$$H_1 \ : \ H(P) = H(Q), \tag{5}$$

$$K_{11} \ : \ H(P) \neq H(Q). \tag{6}$$

Contrary to (2) we can consider not only two-sided alternative hypothesis but also one-sided hypotheses:

$$K_{12} : H(P) < H(Q), \tag{7}$$

or

$$K_{13} : H(P) > H(Q). \tag{8}$$

Let us define a following test statistic

$$T_1 = \frac{\widehat{H}(P) - \widehat{H}(Q)}{\sqrt{\widehat{V}(P) + \widehat{V}(Q)}}, \tag{9}$$

where the empirical entropies $\widehat{H}(P)$ and $\widehat{H}(Q)$ given by

$$\widehat{H}(P) \ = \ -\sum_{i=1}^{m} \widehat{p}_i \log \widehat{p}_i, \tag{10}$$

$$\widehat{H}(Q) \ = \ -\sum_{i=1}^{m} \widehat{q}_i \log \widehat{q}_i, \tag{11}$$

are natural estimators of the true entropies $H(P)$ and $H(Q)$, respectively, $\widehat{p}_i$ and $\widehat{q}_i$ $(i = 1, \ldots, m)$ are estimators of the probabilities $p_i = \Pr(S_i)$ and $q_i = \Pr(R_i)$ of belongingness to the categories corresponding to the Pareto charts under study, i.e.

$$\widehat{p}_i = \frac{X_i}{n_1}, \tag{12}$$

$$\widehat{q}_i = \frac{Y_i}{n_2}, \tag{13}$$

and, finally, where

$$\widehat{V}(P) = \frac{1}{n_1}\left[\sum_{i=1}^{m} \widehat{p}_i \log^2 \widehat{p}_i - \left(\widehat{H}(P)\right)^2\right], \tag{14}$$

$$\widehat{V}(Q) = \frac{1}{n_2}\left[\sum_{i=1}^{m} \widehat{q}_i \log^2 \widehat{q}_i - \left(\widehat{H}(Q)\right)^2\right]. \tag{15}$$

The following lemma yields information about the distribution of the suggested test statistic (9).

**Lemma 1** *Test statistic $T_1$ given by (9) is asymptotically normal provided the null hypothesis $H_1$ holds.*

**Proof:**
Let $X_1, \ldots, X_m$ denote a sample such that $\sum_{i=1}^{m} X_i = n$. It was shown (Basharin, 1959) that the mean and variance of the empirical entropy $\widehat{H}(P)$ given by (10) which is the estimator of the true entropy $H(P)$ for given discrete distribution $(S, P)$, where $P = \{p_1, \ldots, p_m\}$, are as follows

$$E\left(\widehat{H}(P)\right) = H(P) - \frac{m-1}{2n}\log e + O\left(\frac{1}{n^2}\right), \tag{16}$$

$$Var\left(\widehat{H}(P)\right) = \frac{1}{n}\left[\sum_{i=1}^{m} p_i \log^2 p_i - (H(P))^2\right] + O\left(\frac{1}{n^2}\right). \tag{17}$$

Therefore, by the central limit theorem, we may conclude that

$$\frac{\widehat{H}(P) - H(P)}{\sqrt{\sum_{i=1}^{m} p_i \log^2 p_i - (H(P))^2}}\sqrt{n} \underset{n\to\infty}{\sim} N(0,1),$$

i.e. the empirical entropy $\widehat{H}(P)$ is asymptotically normally distributed with the asymptotic mean $H(P)$ and asymptotic variance $V(P)$, where

$$V(P) = \frac{1}{n}\left[\sum_{i=1}^{m} p_i \log^2 p_i - (H(P))^2\right]. \tag{18}$$

One may notice that the variance (18) might be estimated by

$$\widehat{V}(P) = \frac{1}{n}\left[\sum_{i=1}^{m} \widehat{p}_i \log^2 \widehat{p}_i - \left(\widehat{H}(P)\right)^2\right], \tag{19}$$

where $\widehat{p}_i = \frac{X_i}{n}$.

Thus for two independent samples $X_1, \ldots, X_m \sim (S, P)$ and $Y_1, \ldots, Y_m \sim (R, Q)$, such that $\sum_{i=1}^{m} X_i = n_1$ and $\sum_{i=1}^{m} Y_i = n_2$, we conclude that the difference of the empirical entopies $\widehat{H}(P) - \widehat{H}(Q)$ is also asymptotically normal with the asymptotic mean $H(P) - H(Q)$ and the asymptotic variance $V(P) + V(Q)$. Of course, one may estimate this asymptotic variance by $\widehat{V}(P) + \widehat{V}(Q)$. Hence, assuming that the null hypothesis $H_1$ holds, we conclude that the test statistic (9) is also asymptotically standard normal, which completes the proof. ■

Since the test statistic distribution is standard normal, we reject $H_1$ in favor of the alternative hypothesis $K_{1i}$ $(i = 1, 2, 3)$ if $T_1 \in W_i$, where $W_i$ is a critical region corresponding to $K_{1i}$, given as follows

$$W_1 = (-\infty, u_{1-\alpha/2}] \cup [u_{1-\alpha/2}, +\infty), \tag{20}$$

$$W_2 = (-\infty, u_{1-\alpha}], \tag{21}$$

$$W_3 = [u_{1-\alpha}, +\infty). \tag{22}$$

where $u_\gamma$ is a quantile of order $\gamma$ from the standard normal distribution $N(0, 1)$.

One may ask whether these asymptotical results are not too restrictive for practitioners. However, in real life we always have at least a few dozen observations, which turns out completely satisfactory.

### Example 1

Suppose that two Pareto charts (see Fig. 2 and Fig. 3) for the defective components before and after some corrective actions are given. The figures are drawn as follows: the vertical axeses are the numbers of defective components

while the horizontal axeses show the codes of the defective components (the data are shown in Table 1 and Table 2, respectively).

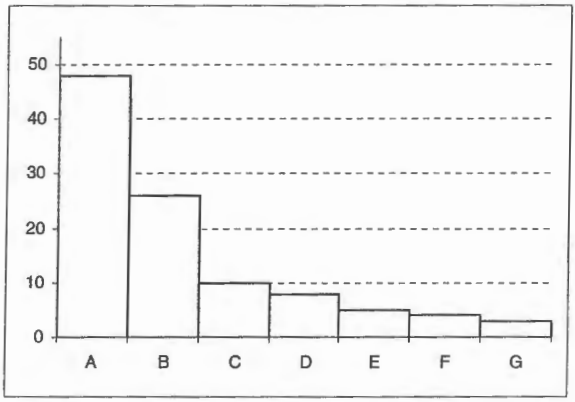


Figure 2: Pareto chart for the data obtained before the corrective action

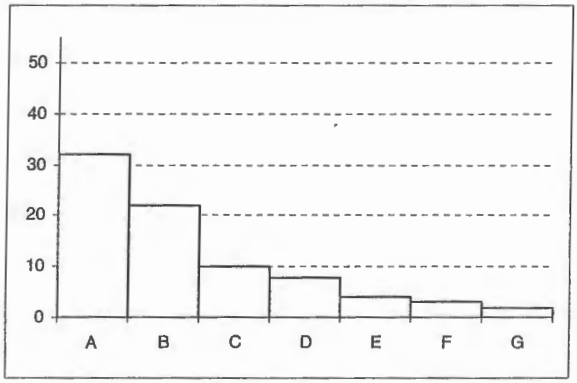


Figure 3: Pareto chart for the data obtained just after the corrective action

| Component code | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| Number of defective components | 48 | 26 | 10 | 4 | 5 | 4 | 3 |

Table 1: Data obtained before a corrective action

| Component code | A | B | C | D | E | F | W |
|---|---|---|---|---|---|---|---|
| Number of defective components | 32 | 22 | 10 | 8 | 4 | 3 | 2 |

Table 2: Data obtained just after the corrective action

Suppose we want to check not only whether these two Pareto charts differ significantly but even more, that the distribution $Q$ of the number of defective components obtained after the corrective action is more flat than the distribution $F$ obtained before that corrective action has been undertaken. Thus we verify a null hypothesis

$$H_1 : H(P) = H(Q)$$

against

$$K_2 : H(P) < H(Q).$$

Let us adopt a significane level 0.05.

We have $m = 7$. After some easy computations we get: $\widehat{H}(P) = 2.1632$, $\widehat{H}(Q) = 2.2648$ which lead to $T_1 = -4.725$. Since $T_1 \in W_1 = (-\infty, 1.u_{0.95}] = (-\infty, 1.64]$ thus we reject the null hypothesis which means that there is a significant difference between these two Pareto charts. Moreover, the entropy of the distribution corresponding to the Pareto chart obtained after the corrective action has greater entropy than that obtained before the corrective action. ■

## 5.2 Comparing Pareto charts with different number of categories

At the beginning of this section we have assumed that the numbers of categories in two charts are equal. We can omit this assumption and consider a following two Pareto charts $X_1, \ldots, X_m \sim (S, P)$ and $Y_1, \ldots, Y_t \sim (R, Q)$, where $\#S = m$, $\#R = t$ and $\sum_{i=1}^{m} X_i = n_1$, $\sum_{i=1}^{m} Y_i = n_2$. However, now we have to apply the standardized entropy given by (4).

We verify the null hypothesis

$$H_2 : H_S(P) = H_S(Q), \tag{23}$$

against one of the following alternatives

$$K_{21} \ : \ H_S(P) \neq H_S(Q), \tag{24}$$

$$K_{22} \ : \ H_S(P) < H_S(Q), \tag{25}$$

$$K_{23} \ : \ H_S(P) > H_S(Q). \tag{26}$$

Now let us consider a following test statistic

$$T_2 = \frac{\widehat{H}(P)\log t - \widehat{H}(Q)\log m}{\sqrt{\widehat{V}(P)\log^2 t + \widehat{V}(Q)\log^2 m}}, \tag{27}$$

where the empirical entropies $\widehat{H}(P)$, $\widehat{H}(Q)$ and sample variances $\widehat{V}(P)$, $\widehat{V}(Q)$ are given by (10), (11) and (14), (15), respectively.

The following result may be proved in much the same way as Lemma 1.

**Lemma 2** *Test statistic $T_2$ given by (27) is asymptotically normal provided the null hypothesis $H_2$ holds.*

**Proof:**
It is easily seen that formula (27) for test statistic $T_2$ is equivalent to

$$T_2 = \frac{\frac{\widehat{H}(P)}{\log m} - \frac{\widehat{H}(Q)}{\log t}}{\sqrt{\frac{\widehat{V}(P)}{\log^2 m} + \frac{\widehat{V}(Q)}{\log^2 t}}}. \tag{28}$$

Thus, realizing that $\frac{\widehat{H}(P)}{\log m}$ is an estimator of the standardized entropy $H_S(P)$, one may perform the same reasoning as in the proof of Lemma 1. This leads to the desired conclusion that $T_2$ is also asymptotically normal, provided (23) holds. ■

Therefore, critical regions for testing hypotheses of the equality of standardized entropies are given - as before - by (20)-(22).

**Example 2**

Now let us consider two Pareto charts prepared by the quality-improvement team which try to identify problems that appear in a manufacturing process. Through the cause and effect analysis they have distinguished two major categories of potential defects caused by the machines and materials. They have also identified various subcauses in each of these major categories. The first Pareto (Fig. 4) chart shows the number of defects due to machine causes, while the second Pareto chart (Fig. 5) shows the number of defects due to material causes (data are shown in Table 3 and Table 4, respectively). Our goal is to ascertain whether there is a significant difference between the distributions corresponding to these two sources of defects.

| Subcause code | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Number of defective components | 45 | 26 | 10 | 8 | 5 | 3 |

Table 3: Defects due to machine causes

| Subcause code | V | W | X | Y | Z |
|---|---|---|---|---|---|
| Number of defective components | 48 | 30 | 12 | 7 | 4 |

Table 4: Defects due to material causes

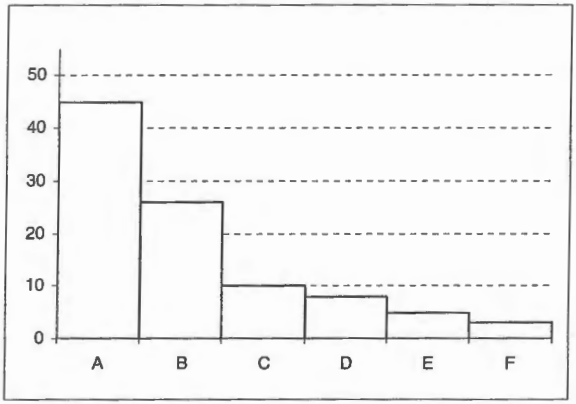


Figure 4: Pareto chart for defects due to machine causes

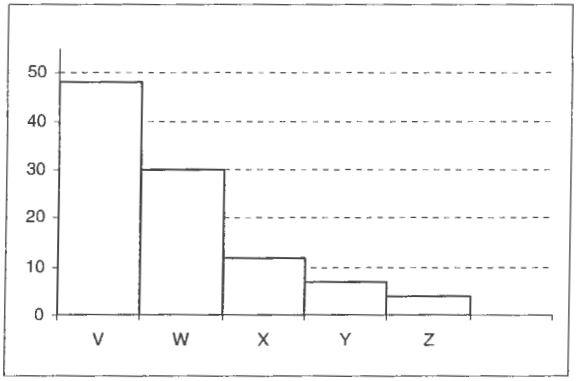


Figure 5: Pareto chart for defects due to material causes

It is seen that now we have to compare two Pareto charts which differ both in categories and in the number of categories as well. Thus we have to consider a test for standardized entropies. We verify a null hypothesis

$$H_2 : H_S(P) = H_S(Q)$$

against the alternative hypothesis

$$K_{21} : H_S(P) \neq H_S(Q).$$

Now we have $m = 6$ and $t = 5$. Substituting the data to (27) we get $T_2 = -0.3425$. If we consider, as before, 5% significance level then we get $W_1 = (-\infty, -1.96] \cup [1.96, +\infty)$. Since $T_2 \notin W_1$ thus we conclude that there is no significant difference between our two Pareto charts. ■

## 5.3 Comparing Pareto chart with the standard one

Our entropy-based test could also be used for comparing given Pareto chart with the "standard" one. Namely, one may ask whether the distribution corresponding to the Pareto chart under study $X_1, \ldots, X_m \sim (S, P)$, where $\#S = m$

and $\sum_{i=1}^{m} X_i = n$, differs significantly from the given distribution characterized by the entropy $h_0$, where $0 \leq h_0 \leq \log m$. This point reduces to testing the null hypothesis

$$H_3 : H(P) = h_0, \tag{29}$$

against one of the alternatives

$$K_{31} \ : \ H(P) \neq h_0, \tag{30}$$

$$K_{32} \ : \ H(P) < h_0, \tag{31}$$

$$K_{33} \ : \ H(P) > h_0. \tag{32}$$

A desired test statistic is given by

$$T_3 = \frac{\widehat{H}(P) - h_0}{\sqrt{\widehat{V}(P)}}, \tag{33}$$

that is asymptotically normal provided (29) holds. And consequently, critical regions are given by (20)-(22).

# 6 Conclusions

SPC is a strategy for the stepwise optimization of the production process through the sequential identification and elimination of potential problems (defects, failures, etc.). The Pareto chart is a powerful tool for the quality improvement. We believe, that statistical tests proposed in the present paper would also be useful in a more advanced analysis that enables to compare Pareto charts constructed in different time moments or even for different production processes.

Moreover, the entropy-based test suggested in this paper could be applied not only in statistical quality control. Actually they are just another goodness-of-fit tests for discrete random variables.

BIBLIOGRAPHY

Arizono I., Ohta H. (1989), A test for normality based on Kullback-Leibler information, Amer. Statist., **43**, 20-22.

Basharin G.P. (1959), On a statistical estimate for the entropy of a sequence of independent random variables, Teor. Verojatnost. i Primenen., **4**, 361-364.

Dudewicz E. J., van der Meulen E. C., SriRam M. G., Teoh N. K. W. (1995), Entropy-based random number evaluation, Amer. J. Math. Management Sci., **15**, 115-153.

Fuchs C., Kenett R. S. (1980), A test for detecting outlying cells in the multinomial distribution and two-way contingency tables, JASA, **75**, 395-398.

Grzegorzewski P., Wieczorkowski R. (1999), Entropy-based goodness-of-fit test for exponentiality, Commun. Statist. Theory and Methods, **28**, 1183-1202.

Kenett R. S. (1991), Two methods for comparing Pareto charts, JQT, **23**, 27-31.

Montgomery D.C. (1991), Introduction to Statistical Quality Control, Wiley.

Shannon C. E. (1948), A mathematical theory of communication, Bell. System Tech. J., **27**, 379-423, 623-656.

Thompson J. R., Koronacki J. (1993), Statistical Process Control for Quality Improvement, Chapman and Hall, New York.

Vasicek O. (1976), A test for normality based on sample entropy, J. Roy. Statist. Soc. Ser. B, **38**, 54-59.