

HOMOGENEITY TESTING PROBLEMS IN BIRD STRIKE DATA PROCESSING
WHEN SAMPLE SIZES ARE SMALL

Presented by Victor Y. Biryukov and Nikolai A. Nechval

HOMOGENEITY TESTING PROBLEMS IN BIRD STRIKE DATA PROCESSING
WHEN SAMPLE SIZES ARE SMALL

Viktor V. Biryukov and Nikolai A. Nechval
Civil Aviation Engineers Institute
Riga, USSR

SUMMARY

This paper deals with testing the homogeneity of bird strike data when sample sizes are small. The traditional statistical approaches developed for large sample data processing will usually not be applicable in the above case. Using a method of conditioning on a sufficient statistic of the likelihood function of bird strike data, we develop some new homogeneity tests. The present paper undertakes a statistical analysis with respect to homogeneity testing problems in the Poisson and two-parameter exponential distributions. Tests are recommended on the basis of certain optimality power properties. The illustrative examples are given.

1. INTRODUCTION

Bird strike reports from various countries for a number of years. The statistics are not always in sight into the exactness of the comparatively rare events. It is not possible for statistical analysis of the traditional data processing methods.

The motivation for the use of statistical methods in data processing when sample sizes are small is that certain conditions are not sufficient statistics. A set of random variables with unknown (nuisance) parameters is utilized in the data processing. The parameters are considered as random variables.

2. A METHOD OF CONDITIONING

This method consists of conditioning on a random variable. Let X^n be a random variable with probability density function $f(x^n; \theta)$. Let $T_n = T_n(X^n)$ be a sufficient statistic of X^n with probability density function $g(t_n; \theta)$. Let $t_n = (t_{n1}, \dots, t_{nk})$ be a vector of functions of x^n which is sufficient for the parameter θ . Then X^n can be transformed into T_n and V_n where V_n is a vector of functions of x^n which is independent of T_n .

$$L(x^n; \theta) = f(t_n, v_n; \theta)$$

where $f(t_n, v_n; \theta)$ is the joint probability density function of T_n and V_n .

$$f(v; t_n) = \frac{f(t_n, v; \theta)}{g(t_n; \theta)}$$

is the conditional probability density function of V_n which does not depend on θ . The property of $f(v; t_n)$ is that it is independent of θ .

By relying upon the method of conditioning of Rosenblatt, the function $f(v; t_n)$ can be transformed into a function of X_1, \dots, X_n into a function of X_1, \dots, X_n which is independent of θ on the interval $[0, 1]$.

To obtain a simple test, we can use the method of conditioning.

1. INTRODUCTION

Bird strike reporting systems have been in operation in many countries for a number of years and it is generally accepted that the statistics arising from these reports provide a valuable insight into the aviation bird hazard. However, bird strikes are comparatively rare events with the result that the data set available for statistical analysis is reasonably small. Consequently the traditional statistical approaches developed for large sample data processing will usually not be applicable in the above case.

The motivation for this paper is to focus attention on an exact statistical method which can be applied for bird strike data processing when sample sizes are small. This method is based on that certain conditional distributions, obtained by conditioning on a sufficient statistic, can be used to transform a data sample into a set of random variables whose distribution does not depend on the unknown (nuisance) parameters. The given distribution can be utilized in the solution which is strictly applicable when the parameters are completely specified.

2. A METHOD OF CONDITIONING ON A SUFFICIENT STATISTIC

This method consists in the following. Let $X^n = (X_1, \dots, X_n)$ be a set of independent random variables that represent observations on a random variable X and are identically distributed with probability density function $f(x; \theta)$ indexed by an unknown (nuisance) parameter θ (in general, vector), $\theta \in \theta^0$ (parameter space). Suppose that $T_n = t_n(X^n)$ is a sufficient statistic for θ with the probability density function $g(t_n; \theta)$. Let $v_{r+1}(x^n), \dots, v_n(x^n)$ be new functions of x^n such that the transformation $x^n = (x_1, \dots, x_n) \rightarrow t_n = (t_{n1}, \dots, t_{nr}), v = (v_{r+1}, \dots, v_n)$ is one-to-one and smooth enough for the Jacobian to exist. Then the likelihood function of x^n can be transformed as

$$L(x^n; \theta) = f(t_n, v; \theta) \left| \frac{\partial(t_n, v)}{\partial x^n} \right| = g(t_n; \theta) f(v; t_n) \left| \frac{\partial(t_n, v)}{\partial x^n} \right|, \quad (1)$$

where $f(t_n, v; \theta)$ is the joint probability density function of T_n and V ,

$$f(v; t_n) = \frac{f(t_n, v; \theta)}{g(t_n; \theta)} = \frac{L(x^n(t_n, v); \theta) \left| \frac{\partial x^n}{\partial(t_n, v)} \right|}{g(t_n; \theta)} \quad (2)$$

is the conditional probability density function of V given $T_n = t_n$ which does not depend on the unknown parameter θ (in virtue of the property of a sufficient statistic).

By relying upon the multivariate probability integral transformation of Rosenblatt (1952), an absolutely continuous density function $f(v; t_n)$ can be used to transform a set of random variables X_1, \dots, X_n into a smaller set of random variables that are identically and independently distributed with uniform distributions on the interval from zero to one.

To obtain a simple procedure for transforming a set of random va-

Theorem 2.1. Let X_1, \dots, X_n be a sample of independent random variables that are identically distributed with probability density function $f(x; \theta)$ indexed by an unknown (nuisance) parameter θ (in general, vector), $\theta \in \theta^0$ where θ^0 is a given parametric set. Let $T_i = t_i(X_1, \dots, X_i)$, $i = 1, \dots, n$, be the sufficient statistics for θ such that $T_i \perp_{\theta} T_{i+1}, \dots, T_n$, $i = 1, \dots, n-1$, and

...הוא נשאל

is the conditional probability density function of X_{i+1} given $T_{i+1} = t_{i+1}$. Then

Proof.

where r is the minimum size of sample required for constructing a sufficient statistic T_r for θ .

Corollary 2.1.1. The procedure of transforming $X^n = (X_1, \dots, X_n)$ is defined by

3. AN ITERATIVE

Let X_1, \dots
variable X w
bility densi
homogeneity
thesis H_0 th
inst unspeci
of the hypot
sequel $f_0(x$;
density func

Case 1. The distribution functional functional as to the va

Case 2. The functional values are

The given pa
H₀ only the
of the param
se is in fac
practice whe
is completely
arson chi-sq
fit this cas
But it posses
group bounda
that its val
i.e., it does
test in the s
the null hyp

In this paper
above objecti
pendent obser
tribution wi
other words,

The proposed
ning on a suf
the test of t
dependent obs
lollowing sequen
the first two
three observa
vations. Here
accepted, $j=3$
ons is accept
accepted. Let
statistics. T
not only mutu
function of t
last conditio
 $H_0(j)$ is reje
point; thus t

3. AN ITERATED PROCEDURE FOR TESTING THE HOMOGENEITY

Let X_1, \dots, X_n be a random sample of n observations on a random variable X with cumulative distribution function $F(x)$ and probability density function $f(x)$. The general problem of testing the homogeneity of n observations consists in testing the null hypothesis H_0 that $F(x) = F(x; \theta_1, \dots, \theta_s)$ for every x , $i=1(1)n$, against unspecified alternatives. Here the θ_i 's denote the parameters of the hypothesized cumulative distribution function F_0 . In the sequel $f_0(x; \theta_1, \dots, \theta_s)$ will denote the hypothesized probability density function. Two cases of this problem are of interest.

(3) Case 1. The hypothesis H_0 is simple, that is the cumulative distribution function F_0 under H_0 is completely specified as to its functional form, such as normality, exponentiality, etc., as well as to the values of the parameters θ_q involved.

(4) Case 2. The hypothesis H_0 is composite. This is when only the functional form of F_0 is given, and some or all of the parametric values are left unspecified.

(5) The given paper deals with Case 2, that is we assume that under H_0 only the functional form of F_0 is given to us but one or more of the parameters θ_q ($q=1, \dots, s$) are left unspecified. This case is in fact of more relevance since situations are very rare in practice where the cumulative distribution function to be tested is completely specified. It is well known that the classical Pearson chi-squared test of goodness-of-fit (χ^2) can be modified to fit this case by properly estimating the unspecified parameters. But it possesses an element of arbitrariness in the choice of group boundaries and one of the objections to this procedure is that its validity is questionable when the sample size n is small, i.e., it does not have the desirable property of being an exact test in the sense of giving exact probabilities of rejection when the null hypothesis is true.

In this paper we propose an iterated procedure (free from the above objections) for testing the null hypothesis H_0 that n independent observations X_1, \dots, X_n come from a common specified distribution with a common but unspecified parameter, i.e., in other words, for testing the homogeneity of n observations.

(6) The proposed procedure is based on the above method of conditioning on a sufficient statistic and consists in that we consider the test of the null hypothesis H_0 of the homogeneity of $n \geq 2$ independent observations. We resolve this hypothesis into the following sequence of nested hypotheses: $H_0(2)$, the homogeneity of the first two observations; $H_0(3)$, the homogeneity of the first three observations; \dots ; $H_0(n)$, the homogeneity of all n observations. Here, the test of $H_0(j)$ is not made unless $H_0(j-1)$ is accepted, $j=3, 4, \dots, n$. Then the homogeneity of the n observations is accepted if and only if all of $H_0(2), H_0(3), \dots, H_0(n)$ are accepted. Let W_2, W_3, \dots, W_n denote, respectively, the $n-1$ test statistics. These test statistics are so selected that they are not only mutually stochastically independent but each W_j is a function of the first j observations alone, $j=2, 3, \dots, n$. This last condition is extremely important to our procedure because if $H_0(j)$ is rejected, using W_j , $j < n$, we stop the testing at that point; thus there is no need to perform the test for the last $n-j$

(7)

observations. For example, suppose $H_0(2)$, and thus $H_0(n) \equiv H_0$, is rejected; we then are not required to go to the trouble and expense of performing the test for the third through n observations since we do not need to compute the statistics W_3, \dots, W_n .

Let a_j be the significance level of the test of $H_0(j)$, $j=2, 3, \dots, n$. The mutual independence of W_2, W_3, \dots, W_n implies that the significance level of the test of the homogeneity of the n observations by this iterated scheme is

$$a = 1 - \prod_{j=2}^n (1 - a_j). \quad (8)$$

It is important to emphasize at this point that this probability is the significance level of this overall test even though the sequence of tests is truncated with the test of $H_0(j)$, $j < n$. For example, if $H_0(2)$ is rejected, we have that H_0 is rejected at significance level

$$a = 1 - \prod_{j=2}^n (1 - a_j), \quad (9)$$

not simply a_2 . Moreover, we have some reason as to why all n observations are not homogeneous; namely, it seems that the first two observations are not homogeneous. Now at this point, it is quite possible that the experimenter would desire to formulate a new hypothesis, such as the homogeneity of the last $n-1$ observations. This can then be tested in the manner outlined above with n replaced by $n-1$.

We illustrate this procedure with two important applications.

4. HOMOGENEITY TESTING FOR THE POISSON PROCESS

Let $X(u)$, $u \geq 0$, be the Poisson process with probability mass function

$$f(X(u)=x; b) = \frac{(bu)^x}{x!} e^{-bu} \quad (b > 0, x \geq 0), \quad (10)$$

where $X(u)$ represents the number of events occurring in the interval $(0, u)$, $X(0)=0$ with probability 1, b is the rate parameter.

To introduce the Poisson homogeneity testing problem, we suppose that we observe n independent random variables $X_1(u_1), \dots, X_n(u_n)$. Under the null hypothesis of homogeneity, each $X_i(u_i)$ follows a distribution (10) governed by the same parameter b , i.e.,

$$f_0(X_i(u_i)=x_i; b) = \frac{(bu_i)^{x_i}}{x_i!} e^{-bu_i}, \quad \forall i. \quad (11)$$

In some problems b may be known, in others unknown. The u_i 's, however, denote constants which are always given rather than unknown. For example, if $X_i(u_i)$ is the number of bird strike incidents incurred during i th of n intervals then u_i could be the number of aircraft movements (in terms of 10,000 movements) or flying hours) and b the average bird strike incident rate per 10,000

aircraft move

We wish to test hypothesis of homogeneity of events. The null hypothesis is that the events are homogeneous. If the events are not homogeneous, then the null hypothesis is rejected.

An iterated procedure is used to test the hypothesis of homogeneity of events. The procedure is as follows: We wish to test the hypothesis of homogeneity of events. The null hypothesis is that the events are homogeneous. If the events are not homogeneous, then the null hypothesis is rejected.

$$\prod_{i=1}^n f_0(X_i(u_i))$$

where

$$f(x_i; t_i, p_i)$$

$$t_i = \sum_{q=1}^i x_q$$

$$p_i = u_i / \sum_{q=1}^i u_q$$

Here, at i th step, the hypothesis is accepted if

$$x_i \in [x_L, x_U]$$

where x_L and

$$\begin{cases} x_L - 1 \\ \sum_{x_i=0}^{x_L-1} f(x_i) \\ x_L \\ \sum_{x_i=0}^{x_L} f(x_i) \end{cases}$$

and

$$\begin{cases} t_i \\ \sum_{x_i=x_U+1}^{t_i} f(x_i) \\ t_i \\ \sum_{x_i=x_U}^{t_i} f(x_i) \end{cases}$$

respectively.

Note that the

aircraft movements (or flying hours).

We wish to test the null hypothesis (11) against an alternative hypothesis of non-homogeneity. By non-homogeneity we mean that, roughly speaking, the $X_i(u_i)$'s are more "spread out" than under the null hypothesis, either as a result of b being different for different i or else as a result of some kind of non-independence of events.

An iterated procedure for testing the homogeneity of n observations from the Poisson process (for the case of unknown b) is based on the transformation

$$\prod_{i=1}^n f_0(X_i(u_i)=x_i; b) \longrightarrow \prod_{i=2}^n f(x_i; t_i, p_i), \quad (12)$$

where

$$f(x_i; t_i, p_i) = \binom{t_i}{x_i} p_i^{x_i} (1-p_i)^{t_i-x_i}, \quad 0 \leq x_i \leq t_i, \quad (13)$$

$$t_i = \sum_{q=1}^i x_q, \quad (14)$$

$$p_i = u_i / \sum_{q=1}^i u_q. \quad (15)$$

Here, at i th stage, the hypothesis $H_0(i)$, $i \in \{2, \dots, n\}$, is accepted if

$$x_i \in [x_L, x_U], \quad (16)$$

where x_L and x_U satisfy the relations

$$\begin{cases} \sum_{x_i=0}^{x_L-1} f(x_i; t_i, p_i) \leq a_i/2 \\ \sum_{x_i=0}^{x_L} f(x_i; t_i, p_i) > a_i/2, \end{cases} \quad (17)$$

and

$$\begin{cases} \sum_{x_i=x_U+1}^{t_i} f(x_i; t_i, p_i) \leq a_i/2 \\ \sum_{x_i=x_U}^{t_i} f(x_i; t_i, p_i) > a_i/2, \end{cases} \quad (18)$$

respectively.

Note that the sample range (when the sample size n is small) is

also useful in testing the homogeneity of n observations from a common Poisson process (10), since it is known that the conditional distribution of n observations x_i from (10) subject to

$$t_n = \sum_{i=1}^n x_i = \text{constant}$$

is the multinomial,

$$f(x_1, \dots, x_n; t_n, p_1, \dots, p_n) = \frac{t_n!}{x_1! \dots x_n!} p_1^{x_1} \dots p_n^{x_n}, \quad (19)$$

where

$$p_i = u_i / \sum_{q=1}^n u_q, \quad i=1(1)n. \quad (20)$$

The exact distribution of the range r conditional upon $t_n = \text{constant}$ can be computed from (19) for a variety of n and nominal levels of significance α , giving values r_α such that

$$\Pr(r \geq r_\alpha) = \sum_{r \geq r_\alpha} f(x_1, \dots, x_n; t_n, p_1, \dots, p_n) \leq \alpha. \quad (21)$$

For the sake of illustration, let us suppose that we observe two independent random variables $X_1(u_1)=6$ and $X_2(u_2)=0$, where $u_1=u_2=3$, with the range $r=6$. It follows from (19) that, for $n=2$ and $t_2=6$, $r=6$ is significant at the probability level equal to 0.03125. If the nominal level of significance $\alpha=0.05$, there is evidence against the assumption of a common Poisson process. Note that the same result, in this case, can be obtained by the iterated procedure.

The test based on the Poisson range supplements the usual index of dispersion of equation

$$\chi^2_{n-1} = \sum_{i=1}^n \frac{[X_i(u_i) - u_i \bar{x}]^2}{u_i \bar{x}}, \quad (22)$$

where

$$\bar{x} = T_n / \sum_{i=1}^n u_i \quad (23)$$

and

$$T_n = \sum_{i=1}^n X_i(u_i), \quad (24)$$

which is approximately distributed as χ^2 with $n-1$ degrees of freedom (see, e.g. Rao, 1952, pp. 205-6).

Table 1 (see below) includes the bird strike data taken from Thorpe and Wessum (1982). Using the statistic (22) for testing the homogeneity of 4 observations from Table 1 for the Poisson process (10) we obtain $\chi^2_3=4.975$. From tabulations of the statistic χ^2_3 we get $\Pr(\chi^2_3 \geq 7.81)=0.05$. Since our computed χ^2_3 is smaller than 7.81 (at the nominal level of significance $\alpha=0.05$) we con-

clude that the non Poisson pr

$$X_i(u_i) \sim$$

where $\delta=3.2656$

TABLE 1. Natio

Country

i

1. Austria
2. Denmark
3. France
4. United King

5. HOMOGENEITY

Considerable a problem of tes variables Y_1 , distributed ex function

$$f(y; \theta) =$$

for some unspe the hypothesis ponential dist

$$f(y; \hat{\alpha}, \theta) =$$

for unspecific

For testing th nential distri mation of the the null hypot observations i many known tes used.

Let Y_1, \dots, Y_n exponential di the paper to $Y(1) \leq Y(2) \leq$

$$f(y_{(1)}, \dots$$

clude that there is no evidence against the assumption of a common Poisson process with the common parameter b , i.e.,

$$\lambda_i(u_i) \sim \frac{(bu_i)^{x_i}}{x_i!} e^{-bu_i}, \quad i=1(1)4, \quad (25)$$

where $\hat{b}=3.2656931$ represents the maximum likelihood estimate of b .

TABLE 1. National Reporting - 1980

Country	Number of bird strike incidents	Number of aircraft movements (in terms of 10,000 movements)
i	x_i	u_i
1. Austria	21	7.0000
2. Denmark	51	15.6463
3. France	134	47.7637
4. United Kingdom	356	101.6821

5. HOMOGENEITY TESTING FOR THE EXPONENTIAL DISTRIBUTION

Considerable attention has been given in the literature to the problem of testing a composite null hypothesis H_0 that a set of variables Y_1, \dots, Y_n represents a set of independent identically distributed exponential random variables with probability density function

$$f(y;\theta) = (1/\theta)\exp(-y/\theta), \quad y \geq 0, \quad (26)$$

for some unspecified common parameter $\theta > 0$. Some authors discuss the hypothesis that the variables have a common two-parameter exponential distribution with probability density function

$$f(y;\hat{a},\theta) = (1/\theta)\exp(-(y-\hat{a})/\theta), \quad y \geq \hat{a}, \quad (27)$$

for unspecified $\hat{a} \in (-\infty, \infty)$ and $\theta > 0$.

For testing the homogeneity of n observations from a common exponential distribution, a procedure is used that involves transformation of the data resulting in $(n-2)$ new variables, which under the null hypothesis of homogeneity are distributed as independent observations from the uniform distribution on $[0,1]$. Thus, the many known tests of this completely specified distribution can be used.

Let Y_1, \dots, Y_n be a random sample of size n from a two-parameter exponential distribution (27). It will be convenient throughout the paper to denote the order statistics of Y_1, \dots, Y_n as $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$. Then

$$f(y_{(1)}, \dots, y_{(n)};\hat{a},\theta) = n! \prod_{i=1}^n f(y_{(i)};\hat{a},\theta) \longrightarrow \prod_{i=2}^n f(z_i;\theta), \quad (28)$$

where

$$f(z_i; \theta) = (1/\theta) \exp(-z_i/\theta), \quad z_i \geq 0, \quad (29)$$

$$z_i = (n-i+1)(y_{(i)} - y_{(i-1)}), \quad i=2(1)n. \quad (30)$$

Using the method of conditioning on a sufficient statistic, we have

$$\prod_{i=2}^n f(z_i; \theta) \rightarrow \prod_{i=3}^n f(z_i; t_i), \quad (31)$$

where

$$f(z_i; t_i) = \frac{i-2}{t_i} \left[1 - z_i/t_i \right]^{i-3}, \quad 0 \leq z_i \leq t_i, \quad (32)$$

$$t_i = \sum_{a=2}^i z_a, \quad i=3(1)n. \quad (33)$$

For the probability integral transformation defined by

$$w_i = F(z_i; t_i) = \int_0^{z_i} f(z_i; t_i) dz_i = 1 - (1 - z_i/t_i)^{i-2}, \quad i=3(1)n, \quad (34)$$

is used, we obtain

$$\begin{aligned} \Pr(W_i \leq w_i; i=3(1)n) &= \int_{\{z_i: F(z_i; t_i) \leq w_i; i=3(1)n\}} \prod_{i=3}^n \frac{n}{t_i} dF(z_i; t_i) \\ &= \int_0^{w_n} \dots \int_0^{w_3} \prod_{i=3}^n \frac{n}{t_i} dw_i = \prod_{i=3}^n w_i, \end{aligned} \quad (35)$$

where $0 \leq w_i \leq 1$, $i=3(1)n$. Hence W_3, \dots, W_n are uniformly and independently distributed on $[0,1]$ random variables.

To test the null hypothesis H_0 we can use, for example, K. Pearson's probability product test

$$P_{n-2} = -\ln \prod_{i=3}^n w_i, \quad (36)$$

a $\Gamma(n-2, 0, 1)$ random variable, or, equivalently,

$$2P_{n-2} = -2\ln \prod_{i=3}^n w_i, \quad (37)$$

a $\chi^2_{2(n-2)}$ random variable.

To illustrate the above procedure for executing a test of the ho-

mogogeneity for the
from the paper of

TABLE 2. National

Country
Austria
Belgium
Denmark
Federal Republic of Germany
Finland

France

United Kingdom

Applying (37) we

$$2P_9 = -2\ln \prod_{i=3}^{11} w_i$$

$2P_9$ is a χ^2_{18} ran-
two-parameter ex-
ficance, we get
the observed val-
ler than 28.87,
ponentiality wit-
ficient certain-

REFERENCES

- Dahl, H. (1982).
vention measu-
Moscow, Worki-
Rao, C.R. (1952).
arch. New Yor-
Rosenblatt, M. (
Ann. Math. St-
Thorpe, J. and va-
European regi-

homogeneity for the exponentiality, the following data were taken from the paper of H. Dahl (1982) (see Table 2).

TABLE 2. National Reporting - 1982

Country	Airports	Costs per average yearly bird strike (\$)
	i	Y_i
Austria	1. Vienna	2,400
Belgium	2. Brussels	1,800
Denmark	3. 12 airports	3,200
Federal Republic of Germany	4. Civil airports	8,000
Finland	5. Helsinki-Vantaa	1,700
	6. Lyon	2,350
	7. Charles de Gaulle	7,400
France	8. Orly	2,500
	9. Marseille	9,000
	10. Nice	2,000
United Kingdom	11. Military airfields	7,000

Applying (37) we have

$$2P_9 = -2 \ln \prod_{i=3}^{11} W_i = 15.12. \quad (38)$$

$2P_9$ is a χ^2_{18} random variable iff the Y_i are drawn from a common two-parameter exponential distribution. At the 5% level of significance, we get from χ^2 -tables that $\Pr(\chi^2_{18} < 28.87) = 0.95$. Since the observed value 15.12 of our χ^2_{18} random variable is much smaller than 28.87, the conclusion about the homogeneity for the exponentiality with common parameters λ and θ can be made with sufficient certainty.

REFERENCES

- Dahl, H. (1982). Economical and operational aspects of bird prevention measures. 16th Meeting Bird Strike Committee Europe. Moscow, Working Paper 18.
- Rao, C.R. (1952). Advanced Statistical Methods in Biometric Research. New York: John Wiley and Sons, Inc.
- Rosenblatt, M. (1952). Remarks on a multivariate transformation. Ann. Math. Statist., Vol. 23, pp. 470-472.
- Thorpe, J. and van Wessum, R. (1982). Bird strikes during 1980 to European registered civil aircraft. BSCE 16. Moscow, WP 14.