Ecological Archives **E088-196-A2**

Christophe Bonenfant, Jean-Michel Gaillard, Stéphane Dray, Anne Loison, Manuela Royer, and Daniel Chessel. 2007. Testing sexual segregation and aggregation: Old ways are best. *Ecology* 88:3202–3208.

Appendix B. Mathematical development of the SSAS.

Appendix B
Mathematical developments of the SSAS

In the following section, we present the mathematical developments and related underlying statistical tenets of our paper. Contrary to segregation coefficients, this theoretical background ensures the validity of the proposed test (hereafter called $SSAS$ standing for Sexual Segregation and Aggregation Statistic).

Context

Qualitative variables

We first set the discussion in a probabilistic context. The problem deals with the study of an eventual dependence between the factor *Sex* (male or female) and the distribution of animals among *Groups*. In other words, at a fixed instant, for any individual, we consider two qualitative variables :

- its *Sex* that can take one of the 2 modalities *male* or *female*,
- its *Group* of appartenance that can take one of the theoretical modalities valued in $\mathbb{N}^*$.

We check that at the time of observation, any individual presents one and only one modality for each variable.

Observed Contingency Table

We now use the qualitative variables Sex and Group previously defined, in order to describe a population consisting of N individuals. Actually, there may not exist more Groups than the number N of individuals in the data set. Considering a modality that does not occur in the data set has no sense. So we reduce the modalities of the variable Group to the number k ($k \geq 2$) of observed Groups. We can therfore construct a standard contingency table as follows:

$Sex \setminus Group$	1	2	$\cdots$	j	$\cdots$	k	Total
$Male$	x_1	x_2	$\cdots$	x_j	$\cdots$	x_k	x
$Female$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_k	y
Total	n_1	n_2	$\cdots$	n_j	$\cdots$	n_k	N

Probabilistic approach

Studied Random variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider 2 random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ denoted by S and G, valued respectively in $\{1, 2\}$ and $\{1, 2, \cdots, k\}$. Suppose their distribution

defined as follows :

$$\begin{array}{lll} \mathbb{P}(S=1)=p\,, & \mathbb{P}(S=2)=1-p\,. & \\ \mathbb{P}(G=1)=p_1^{\mathrm{G}}\,, & \mathbb{P}(G=2)=p_2^{\mathrm{G}}\,, & \cdots \quad \mathbb{P}(G=k)=p_k^{\mathrm{G}}, \end{array}$$

with $p \in]0,1[$ and $p_1^G, \cdots p_k^G > 0$ such that $\displaystyle\sum_{j=1}^{k} p_j^G = 1$.

We aim to test if the qualitative variables *Sex* and *Group* are independent. We will test the null hypothesis $\underline{H_0}$: "S and G are independent" against the alternative hypothesis $\underline{H_1}$: "S and G are not independent". We shall first recall the following basic result.

Definition B.1

Random variables S and G are independent if and only if $\forall j \in \{1,2,\cdots,k\}$,

$$\begin{aligned} \mathbb{P}(S=1 \cap G=j) &= \mathbb{P}(S=1)\times\mathbb{P}(G=j) = p\; p_j^G\,, \\ \mathbb{P}(S=2 \cap G=j) &= \mathbb{P}(S=2)\times\mathbb{P}(G=j) = (1-p)\; p_j^G\,. \end{aligned}$$

We now consider a sample of random variables $S_1,\cdots,S_N$ (resp. $G_1,\cdots,G_N$) drawn from S (resp. G). The observations of the factor *Sex* (resp. *Group*) for the N individuals of the population give a realization $s_1,\cdots,s_N$ (resp. $g_1,\cdots,g_N$) of these random variables. Indeed, s_i reveals the *Sex* of the i^{th} individual, and g_i its *Group* of appartenance.

Remark: During the sampling procedure of the data, we need to assume the same condition as in Conradt's work (Conradt 1998), p. 222: "This applies only if individual animals in the population move between groups independently of other individuals".

Construction of a Statistic

Notations Let X (resp. Y) be the number of males (resp. females) in the population :

$$X \;=\; \sum_{i=1}^{N} \mathbb{1}\!\mathrm{l}_{S_i=1} \;\text{ and }\; Y \;=\; \sum_{i=1}^{n} \mathbb{1}\!\mathrm{l}_{S_i=2}\,.$$

We also define N_j as the number of individuals in *Group* j : $\forall\, j \in \{1,2,\cdots,k\}$,

$$N_j \;=\; \sum_{i=1}^{N} \mathbb{1}\!\mathrm{l}_{G_i=j}\,.$$

By definition, $\displaystyle\sum_{j=1}^{k} N_j = N$. We can also introduce the random variable X_j defined as the number of males belonging to *Group* j, and Y_j the number of females belonging to the same *Group* (variables X_j and Y_j satisfy $0 \le X_j, Y_j \le N_j$). Obviously, $\forall\, j \in \{1,2,\cdots,k\}$, $X_j + Y_j \;=\; N_j$, and $X \;=\; \sum_{j=1}^{k} X_j \;\;,\;\; Y \;=\; \sum_{j=1}^{k} Y_j \;=\; N - X$.

Principle of the test Under hypothesis $\underline{H_0}$, $\forall j \in \{1, 2, \cdots, k\}$ $\forall i \in \{1, 2, \cdots, N\}$,

$$\begin{aligned} \mathbb{P}(S_i = 1 \cap G_i = j) &= \mathbb{P}(S_i = 1) \times \mathbb{P}(G_i = j) = p\, p_j^{\mathrm{G}}, \\ \mathbb{P}(S_i = 2 \cap G_i = j) &= \mathbb{P}(S_i = 2) \times \mathbb{P}(G_i = j) = (1-p)\, p_j^{\mathrm{G}}, \end{aligned}$$

Hence, if $\underline{H_0}$ is satisfied, the observed frequencies should verify $\forall j \in \{1, 2, \cdots, k\}$,

$$\frac{x_j}{N} \simeq p\, p_j^{\mathrm{G}} \quad \text{and} \quad \frac{y_j}{N} \simeq (1-p)\, p_j^{\mathrm{G}}.$$

In our case, the probabilities p and p_j^{G} are unknown $\forall j \in \{1, 2, \cdots, k\}$. Nevertheless, the law of large numbers allows the approximation of p by $\frac{x}{N}$, and p_j^{G} by $\frac{n_j}{N}$ $\forall j \in \{1, 2, \cdots, k\}$. Hence the question can be reformulated in this way : Are observed frequencies $\frac{x_j}{N}$ and $\frac{y_j}{N}$ close enough from $\frac{x}{N} \times \frac{n_j}{N}$ and $\frac{y}{N} \times \frac{n_j}{N}$ respectively, to conclude that factors S and G are independent ? In terms of observed numbers, we are led to compare x_j (resp. y_j) and $\frac{x\, n_j}{N}$ (resp. $\frac{y\, n_j}{N}$).

χ^2-Statistic Actually, we only study populations of size N, with both males and females, distributed in k groups. Hence, random variables X, Y and N_j ($\forall\, j \in \{1, 2, \cdots, k\}$) satisfy X, Y, $N_j \geq 1$ $\mathbb{P}a.s.$. In such situations, the convenient mathematical statistic is defined by

Definition B.2

$$\boldsymbol{X^2} = \sum_{j=1}^{k} \left(\frac{(X_j - \frac{XN_j}{N})^2}{\frac{XN_j}{N}} + \frac{(Y_j - \frac{YN_j}{N})^2}{\frac{YN_j}{N}} \right), \quad \textit{is called the } \chi^2\textit{-statistic.}$$

Theorem B.3
Under $\underline{H_0}$, $\boldsymbol{X^2} \xrightarrow[N \to +\infty]{d} \chi^2_{k-1}$.

Proof.
The proof is very technical. For details, see Kendall and Stuart (1961). We just recall here that the degree of freedom is computed as $(2-1) \times (k-1)$. □

Theorem B.3 means that independence test of χ^2 is an asymptotical test, so that a large N (size of the population) is required.

Statistical Test of independence

Whenever applying any statistical test, we compute the value of the statistic from the observed sample. In our case, we compute

$$\boldsymbol{X^2}_{\text{observation}} = \sum_{j=1}^{k} \left(\frac{(x_j - \frac{x\, n_j}{N})^2}{\frac{x\, n_j}{N}} + \frac{(y_j - \frac{y\, n_j}{N})^2}{\frac{y\, n_j}{N}} \right).$$

The Theorem B.3 provides the asymptotic distribution of the statistic $\boldsymbol{X^2}$. We shall recall that such a distribution is available only if the sample size n is large. In the current framework, it does not seem reasonable to allow n to grow to infinity whereas the number of groups k keeps fixed. This is the reason why we do not use the asymptotic distribution of $\boldsymbol{X^2}$ but advise to use a randomization test instead (Manly 1991). In other words, we still consider the χ^2-statistics, whose exact distribution is unknown. The value of $\boldsymbol{X^2}_{\text{observation}}$ is compared with the distribution of χ^2-statistic obtained by randomly reordering the data. An observation ω_0 gives an observed contingency table, and so $X(\omega_0)$, $Y(\omega_0)$, $N_j(\omega_0)$, $\forall j \in \{1, 2, \cdots, k\}$. The randomization procedure considers any $\omega \in \Omega$ such that $X(\omega) = X(\omega_0)$, $Y(\omega) = Y(\omega_0)$, $N_j(\omega) = N_j(\omega_0)$ $\forall j \in \{1, 2, \cdots, k\}$ (i.e. any contingency tables with marginal totals equal to those of the observed table). The decision is taken by comparing the value of the statistical $\boldsymbol{X^2}_{\text{observation}}$ and the quantiles (associated to the probability $1 - \alpha$) of the randomized distribution. If hypothesis $\underline{H_0}$ is not accepted, we conclude (with risk α) that factors *Sex* and *Group* are not independent. In the paper, a Monte-Carlo version of the test is used by considering a random sampling of the randomization distribution.

Definition of an index detecting dependence

The χ^2-Statistic is convenient for testing independence of factors *Sex* and *Group*. In our context, we can prove that variable $\boldsymbol{X^2}$ can take another expression.

Proposition B.4

$$\boldsymbol{X^2} \;=\; N \left(1 - \frac{N}{XY} \sum_{j=1}^{k} \frac{X_j\, Y_j}{N_j} \right)$$

Proof.
This result is proved in the Appendix of our paper. □

Hence we can define an index called SSAS by $\frac{\boldsymbol{X^2}}{N}$.

Definition B.5
We call Sexual Segregation and Aggregation Statistic the following quantity :

$$SSAS \;=\; \left(1 - \frac{N}{X\,Y} \sum_{j=1}^{k} \frac{X_j\, Y_j}{N_j} \right)$$

Properties of *SSAS*

Given that index SSAS stems from a strong mathematical theory, several properties can be deduced easily.

Theorem B.6

i) Cramer's variable $V = \sqrt{\frac{\boldsymbol{X^2}}{N \times min(2-1, k-1)}}$ *takes values in* $[0, 1]$.

ii) $\mathbb{E}\Big(\frac{\boldsymbol{X^2}}{N} \Big) = \frac{k-1}{N-1}$.

Proof.

i) See Cramer (1999).

ii) The proof lies on the fact that $\forall\, j \in \{1, \cdots k\}$,

$$\mathbb{E}(\, X_j\, Y_j \mid X = x,\, Y = y,\, N_1 = n_1,\, \cdots,\, N_k = n_k\,) \;=\; \frac{x\, y\, n_j\, (n_j - 1)}{N(N-1)}\,. \qquad (1)$$

We refer to Haldane (1940) for details.

□

Corollary B.7

i) Index SSAS takes values in $[0, 1]$.

ii) $\mathbb{E}\Big(\frac{k-1}{N-1}\Big)$.

iii) Index SC_{social} *introduced by Conradt (1998) is the observation of a random variable whose expectation is* $\frac{-1}{N-1}$ *and not 0 contrary to what Conradt claimed in her paper.*

iv) Let us define $Z \;=\; 1 \,-\, \dfrac{N-1}{X\, Y} \displaystyle\sum_{j=1}^{k} \frac{X_j\, Y_j}{N_j - 1}$.

Conradt (1999) introduced SC as $\sqrt{Z_{observation}}$.

This index is not well-defined because $\mathbb{P}\Big(\;1 \,-\, \dfrac{N-1}{X\, Y} \displaystyle\sum_{j=1}^{k} \frac{X_j\, Y_j}{N_j - 1} \;<\; 0 \Big) \;>\; 0$.

Proof.

i) It derives from Theorem B.6 i).

ii) It derives from Theorem B.6 ii).

iii) We use the equality (1), which remains available under the probability set Ω such that $\forall\, j \in \{1, \cdots k\},\ N_j \geq 2$.

iv) The same calculation provides $\mathbb{E}(Z) = 0$. Obviously, if Index SC is well-defined, Z would be a nonnegative random variable with null expectation. In other words, Z would be null $\mathbb{P}$-almost surely. To use a null variable to define a measure of segregation is not acceptable.

Actually, we can prove that Z is not a nonnegative random variable, so that the index SC is mathematically wrong. First of all, we fix $x, y, n_1, \cdots, n_k$ such that

$$x + y = \sum_{j=1}^{k} n_j = N \quad \text{and} \quad \forall\, j \in \{1, \cdots k\}, \quad n_j > 1 \tag{2}$$

Then by definition of variables S and G in Subsection ,

$$\mathbb{P}(\ X = x,\, Y = y,\, N_1 = n_1,\, \cdots, N_k = n_k\ )\ >\ 0 \tag{3}$$

We refer to Haldane (1940) to state that

$$\mathbb{P}(\ X_1 = x_1,\, \cdots, X_k = x_k \mid X = x,\, Y = y,\, N_1 = n_1,\, \cdots, N_k = n_k\ )\ >\ 0$$

In particular, the following conditional probability is positive :

$$\begin{aligned} &\mathbb{P}(\, \exists i_1, i_2 \text{ s.t. } X_{i_1} = 1, X_{i_2} = 1 \text{ and } X_j = 0\, \forall\, j \neq i_1, i_2 \\ &\qquad \mid X = 2,\ Y = N - 2,\ N_1 = n_1,\, \cdots,\, N_k = n_k\,)\,. \end{aligned} \tag{4}$$

And Equation (3) gives that the joint probability

$$\begin{aligned} &\mathbb{P}(\, \exists i_1, i_2 \text{ s.t. } X_{i_1} = 1, X_{i_2} = 1 \text{ and } X_j = 0\, \forall\, j \neq i_1, i_2 \\ &\qquad \cap\ X = 2,\ Y = N - 2,\ N_1 = n_1,\, \cdots,\, N_k = n_k\,) \end{aligned} \tag{5}$$

is positive as the product of two positive terms.

Let us denote by $\mathcal{A}$ the set of events $\omega \in \Omega$ such that

$$\begin{aligned} &\exists i_1, i_2 \text{ s.t. } X_{i_1} = 1, X_{i_2} = 1 \text{ and } X_j = 0\, \forall\, j \neq i_1, i_2 \\ &\qquad \cap\ X = 2,\ Y = N - 2,\ N_1 = n_1,\, \cdots,\, N_k = n_k\,. \end{aligned} \tag{6}$$

Then $\mathcal{A}$ is a nonnegligible set and for such events $\omega \in \mathcal{A}$,

$$Z(\omega)\ =\ 1\ -\ \frac{N-1}{2(N-2)}\left(\frac{n_{i_1}-1}{n_{i_1}-1} + \frac{n_{i_2}-1}{n_{i_2}-1}\right)\ =\ -\frac{1}{N-2}\ <\ 0\,.$$

As a conclusion, we obtain $\mathbb{P}(\, Z\ <\ 0\,) > 0$

□

Remark 1 : When rejecting hypothesis H_0, we do it with a reasonable risk (generally equal to 5 %) that factors *Sex* and *Group* were independent nevertheless. This is the usual principle of test.

Remark 2 : Index *SSAS* must not be used to measure any degree of segregation. It has no sense to compare two values of *SSAS*, and conclude that segregation is higher in one situation than in the other one. We show here that it is not convenient to measure segregation by comparing p-values.

Remark 3 : *Remark 2* also holds for indexes SC_{social} and SC successively introduced by Conradt (1998, 1999). In addition, indexes SC_{social} and SC do not refer to any statistical test.

Confidence Interval for distance to independence ?

The question is : "Can we measure a degree of segregation"? In other words, we wish to measure how far from independence our observation are, when $\underline{H_0}$ is rejected. To deal with this problem, we need to know the (asymptotic) distribution of random variable $\boldsymbol{X^2}$ under $\underline{H_1}$. We recall that when $\underline{H_0}$ is satisfied and under the assumption that k is a constant, the asymptotic distribution of $\boldsymbol{X^2}$ is given by Theorem B.3: a χ^2 with degrees of freedom $k-1$. But when $\underline{H_0}$ is rejected, the probability that an individual satisfies both $Sex = 1$ (resp. $Sex = 2$) and $Group = j$ is no longer $p\, p_j^G$ (resp. $(1-p)\, p_j^G$).

Let us denote, $\forall j \in \{1, 2, \cdots, k\}$

$$\begin{aligned} \mathbb{P}(S = 1 \cap G = j) &= p_{1j}\,, \\ \mathbb{P}(S = 2 \cap G = j) &= p_{2j}\,. \end{aligned} \tag{7}$$

Under assumption $\underline{H_1}$, the true distribution of the couple of random variables (S, G) is defined by $\{p_{1j},\, p_{2j}\}_{j \in \{1,2,\cdots,k\}}$. Then the asymptotic distribution of $\boldsymbol{X^2}$ is a non-central χ^2 with degrees of freedom $k-1$ and non-central parameter λ (see Kendall and Stuart (1961)) :

$$\lambda = N \left(\sum_{j=1}^{k} \frac{\left(p_{1j} - p\, p_j^G\right)^2}{p\, p_j^G} + \sum_{j=1}^{k} \frac{\left(p_{2j} - (1-p)\, p_j^G\right)^2}{(1-p)\, p_j^G} \right) \tag{8}$$

Nevertheless, such a proprety requires an important condition :

$$\forall j \in \{1, 2, \cdots, k\},\ \exists\, a_j \text{ and } b_j \in \mathbb{R} \text{ such that}$$

$$\begin{aligned} p_{1j} - p\, p_j^G &= \frac{a_j}{\sqrt{N}}\,, \\ p_{2j} - (1-p)\, p_j^G &= \frac{b_j}{\sqrt{N}}\,. \end{aligned} \tag{9}$$

When Condition (9) is satisfied, the non-central parameter λ is well-defined and can be estimated by the value of $\boldsymbol{X^2}$ itself. Bulmer (1958) discusses confidence limits for $\lambda^{1/2}$, which is a natural parameter for distance to independence.

Theorem B.8
When Condition (9) holds, $\boldsymbol{X^2} \xrightarrow[N \to +\infty]{d} \chi^2_{(k-1,\lambda)}$,
where λ is the non-centrality paramater defined in (8).

Under $\underline{H_0}$, condition (9) is obviously satisfied with $a_j = b_j = 0$. And in this case, parameter λ equals 0. Thus the result given by Theorem B.8 under $\underline{H_0}$ is equivalent to the result enounced in Theorem B.3. We emphasize the fact that Condition (9) is essential in the proof of Theorem B.8. We then should not use any confidence interval for λ without proving that Condition (9) is true. Unfortunately, such a condition is very difficult to verify in practice, because groups are not identifiable and the experiment not repeatable.

Literature Cited

Bulmer, M., 1958. Confidence intervals for distance in the analysis of variance. Biometrika **45**:360–369.

Conradt, L., 1998. Measuring the degree of sexual segregation in group-living animals. Journal of Animal Ecology **67**:217–226.

Conradt, L., 1999. Social segregation is not a consequence of habitat segregation in red deer and feral soay sheep. Animal Behaviour **57**:1151–1157.

Cramer, H., 1999. Mathematical Methods of Statistics. Princeton University Press.

Haldane, J., 1940. The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. Biometrika **31**:346–355.

Kendall, M. and A. Stuart, 1961. The advanced theory of statistics – Inference and relationship, volume 2. Griffin & Co, London.

Manly, B., 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.