GINI'S EARLY INVESTIGATION ABOUT THE DISTRIBUTION OF SEXES IN HUMAN BIRTHS

Antonio Forcina[1]

Economics Department, University of Perugia, Italy

Abstract *Two early contributions by C. Gini, containing intuitions of some interest to modern readers, are reviewed. In an attempt to investigate relevant features in the distribution of sexes in human births, Gini formulated and tried to test informally several models which can he conceived as possible alternatives to the binomial distribution, leading to over or under dispersion. One of these models allows for serial correlation within births from the same couple and, in an attempt to provide a testing procedure of this alternative hypothesis, Gini exploits an intuitive notion of finite exchangeability. Informal testing is usually based on comparing expected and observed frequencies. Two interesting applications based on data from Saxony and the town of Dresden in 18th century births will be described in some detail.*

Keywords: *sex ratio, binomial distribution, over-dispersion, finite exchangeability.*

1. INTRODUCTION

The book by Gini (1908), an extended version of his dissertation at the University of Bologna, contains the results of a massive investigation about the distribution of sexes in human births based on data from different countries and historical periods. The present paper concentrates on certain crucial portions of Chapters V where he discusses a collection of models for binary data which differ in various ways from the binomial distribution. Some applications of these results to several data sets on human births contained in Chapter X will also be outlined.

Gini (1911) is a paper which appeared in a collection of studies published at the University of Cagliari where he held a position between 1909

[1] e-mail: forcina@stat.unipg.it

and 1913; the paper was later reprinted in Metron. It deals with the computation of what, in modern terminology, would be called *predictive probability*: the probability that the next child is a male, knowing that the same couple had already x males out of n births. This paper contains two main results: an expression for computing the predictive probability based on an assumption which is, essentially, equivalent to finite exchangeability and an informal test of that same assumption. Gini (1911) contains also some sharp criticism of the assumption of uniform prior probabilities which, apparently, was commonly used at the time for computations based on Bayes' theorem.

The present work is an attempt to present a selected set of results from these early contributions by translating Gini's terminology and line of argument into modern language. In particular, the collection of models which he proposed as alternatives to the binomial distribution, here are rephrased into alternatives to the multinomial distribution, without changing their original structure. Though, clearly, he did not have such extensions in mind, they are quite straightforward, in addition, this approach makes the presentation simpler.

While certain passages of these early works are easy to follow, others may appear a little obscure to a modern reader, both in the terminology and the conceptual framework current at his time but rather obsolete after over one century. Passages where conclusions may depend on subjective interpretations are discussed in some detail.

2. THE 1908 BOOK

This is a book of about 500 pages which begins with a detailed list of the sources of the examined data, followed by 14 pages of references. Chapter IV presents and discusses the main results of the *Theory of dispersion*, the name used at the beginning of last century to indicate the procedures developed by Lexis, Dormoy and others to test whether certain collections of binary data conformed to the binomial distribution, an issue which had important implications, (see Stigler, 1986, Chapter 6).

Section 1 of Chapter V of Gini's book contains a discussion of possible applications of the binomial model to the distribution of sexes in human births. He seems to say that, if we assumed that the probability of a male birth was constant and events were independent, then we could derive several results like, for instance, assess whether the proportion of males in a given town for a given year was or not too unusual for being due to random variation. In this respect he says, p. 20, "We must decide how small must be the probability of an event for us to say that it is unlikely to happen" and admits that the problem cannot have an objective solution but suggests that events having probability less than 1/20 may be considered as rather unusual. He also mentions that similar tools could be used to detect errors in the data, a kind of outlier detection.

The tools Gini mentions in Section 1 are often based on the normal approximation to the binomial distribution, when the number of observations in his applications to sex in human births are sufficiently large. For instance, on page 140 he recalls the expressions for computing the standard error of a proportion or the difference between two proportions. From the review material in Chapter IV it is quite clear that "normal dispersion", a term used often in the book, means a shape of the distribution compatible with what would be expected under the assumption of constant probability of success and independence. This is checked both by comparing observed and expected summary measures of variability as well as by what he calls "enumeration", that is the expected and observed frequencies of certain deviations from the mean. In the analyses that he presents in Chapter X, when studying the distribution of males within families with n births, to determine whether a given distribution was more or less "than normal", a terminology probably borrowed from the work of Lexis (see Stigler, 1986, Chapter 6), Gini compares observed and expected (under the binomial model) frequencies near the mean and on the tails of the distribution.

2.1 A COLLECTION OF ALTERNATIVES TO THE BINOMIAL MODEL

Section 2 in Chapter V begins with a challenging question which may be translated as follows: suppose that a given collection of observations concerning the proportion of male births in different families (or different years or regions) have distributional features compatible with those of the binomial model, can we infer that the probability of a male birth is constant and events are independent ? According to Gini, Lexis, Poisson and von Bortkiewiscz had already explained that, if both the number of births and the probability of a male birth within each separate data set varied at random, see model Ac) below, the overall distribution, obtained by merging the data, would still be compatible with the binomial model. In other words, when the binomial model is violated in different directions, the consequences on the dispersion of the overall distribution may cancel (p. 151). At this point he asserts that there are several other violations of the binomial model and describes (pp 152-156) six different alternatives to the binomial model.

Each of these models is first outlined in words and then by describing how the corresponding data could be generated at random by drawing balls from certain collections of urns; these urn models, though expressed in words, are very accurate and leave no ambiguity. After describing each model, Gini adds a short statement specifying whether the "dispersion" to be expected in the data generated by that model is "normal", "more than normal" or "less than normal". Though each of Gini's statement is indeed correct, as shown below, he does not provide any proof nor references. Thus, it is not clear whether he reached his conclusions by actually drawing balls from his boxes, or simply by intuition.

Below, when presenting each of the six alternative models, the binomial distribution is replaced with the multinomial and stated formally in modern language; in addition, an expression for the covariance matrix is derived and then compared with the covariance matrix of the multinomial distribution. Models are identified with the same headings as in the book; all of his urns contain balls of two different colours and in each case he specifies clearly that drawings are with replacement. In model Aa) over-dispersion is due to variations in the underlying probability; for instance, if the probability of having a male birth varies from couple to couple, the distribution of sexes in, say, families with 5 births should exhibit over-dispersion compared to the binomial model. Models Ab), Ac) and Ad) are about mixtures of binomial distributions and could be applied, for instance, to the distribution of the total number of male births in different communities, by aggregating data from different populations. Model Be) is about serial correlation and could be applied when studying the distribution of males births in a given town or region for a collection of years. Finally, models Bf) and Bg) are about serial correlation between observations within a given distribution and could be applied to investigate whether the sex of a new baby is affected by the sex of babies in previous births in the same family, an issue investigated in detail in Gini (1911).

Recall that if $\boldsymbol{x}$ is a vector of frequencies having a multinomial distribution $(n, \boldsymbol{p})$, where n is the total number of observations and $\boldsymbol{p}$ is the vector of probabilities, then the covariance matrix $\text{Var}(\boldsymbol{x})$ is equal to $n[\text{diag}(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}']$; in the following $\boldsymbol{\Omega}(\boldsymbol{p})$ will be used as a shorthand for $\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}'$.

Aa "We perform several sets of 100 draws ... for each set we take an urn with a different proportion between the number of balls of the two colours"; in this model he states that "dispersion is more than normal". Let $\boldsymbol{p}$ denote the vector whose elements are the proportion of balls of different colours in a given urn; then the above model may be translated by first assuming that $\boldsymbol{x}$, the vector containing the number of balls of different colours obtained by drawing n balls with replacement, has a multinomial distribution $(n, \boldsymbol{p})$. Because the composition of the urn is not constant, $\boldsymbol{p}$ is itself a random variable; assume, without loss of generality, that $\text{E}(\boldsymbol{p}) = \boldsymbol{\pi}$ and $\text{Var}(\boldsymbol{p}) = \boldsymbol{\Sigma}$. Because of the assumed multinomial, $\text{Var}(\boldsymbol{x} \mid \boldsymbol{p}) = n\boldsymbol{\Omega}(\boldsymbol{p})$; by using standard results for computing the marginal variance, it follows that

$$\text{Var}(\boldsymbol{x}) = n\boldsymbol{\Omega}(\boldsymbol{\pi}) + (n-1)\boldsymbol{\Sigma},$$

in words, as long as each sample contains more than one observation, the multinomial variance increases in a way that depends on how much the proportions $\boldsymbol{p}$ vary from one sample to the other; this is a general model of over-dispersion.

Ab "We have 10 urns with a different proportion between the balls of the two colours; we draw several samples of 100 balls obtained by taking 10 balls from each urn"; the dispersion in this model will be "less than normal". The model may be translated by assuming that the vector of observed frequencies may be written as $\boldsymbol{y} = \sum \boldsymbol{x}_i$, where $\boldsymbol{x}_i$ is the frequency vector for the balls drawn from a single urn and has a multinomial distribution $(n_i, \boldsymbol{p}_i)$; let $n = \sum n_i$, $r_i = n_i/n$ and $\boldsymbol{\pi} = \sum r_i \boldsymbol{p}_i$. Then, $\text{Var}(\boldsymbol{y}) = \sum \text{Var}(\boldsymbol{x}_i) = n[\sum r_i \text{diag}(\boldsymbol{p}_i) - \sum r_i \boldsymbol{p}_i \boldsymbol{p}_i']$. Let $\boldsymbol{\Sigma} = \sum r_i(\boldsymbol{p}_i - \boldsymbol{\pi})(\boldsymbol{p}_i - \boldsymbol{\pi})'$, then, by simple calculations, it follows that $\text{Var}(\boldsymbol{y}) = n[\boldsymbol{\Omega}(\boldsymbol{\pi}) - \boldsymbol{\Sigma}]$. This result indicates that mixing different multinomial distributions causes under-dispersion.

Ac "We have 10 numbered urns ... each urn contains balls of the two colours in a different proportion ... the number of balls we draw from each urn is not constant ... it is determined beforehand ... as follows: take an other urn containing 10 balls marked with digits from 1 to

10; from this urn we make 100 draws ... the number of balls drawn with the digit 1 indicates the number of draws to be made from urn 1 ...". According to Gini, the dispersion here is "normal". This model may be translated by assuming, as above, that $\boldsymbol{y} = \sum \boldsymbol{x}_i$ with $\boldsymbol{x}_i$ multinomial $(n_i, \boldsymbol{p}_i)$, however, because the number of draws from each urn is not constant, let $\boldsymbol{n}$ be the vector with elements n_i and assume, in addition, that $\boldsymbol{n}$ follows, in turn, a multinomial $(n, \boldsymbol{\tau})$, where the elements of $\boldsymbol{\tau}$ determine the probability of using each different urn; in Gini's example the different urns has the same probability of being selected. Let also $\boldsymbol{P}$ denote the matrix whose ith column is equal to $\boldsymbol{p}_i$ then $\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\tau}$. Notice that $\mathrm{E}(\boldsymbol{y} \mid \boldsymbol{n}, \boldsymbol{P}) = \boldsymbol{P}\boldsymbol{n}$ and also that $\mathrm{E}(\boldsymbol{y})$ $= n\boldsymbol{P}\boldsymbol{\tau} = n\boldsymbol{\pi}$. Then, by using the results for the previous model, the conditional variance may be written as

$$\mathrm{Var}(\boldsymbol{y} \mid \boldsymbol{n}, \boldsymbol{P}) = \mathrm{diag}(\boldsymbol{P}\boldsymbol{n}) - \boldsymbol{P}\mathrm{diag}(\boldsymbol{n})\boldsymbol{P}',$$

because $\mathrm{Var}(\boldsymbol{P}\boldsymbol{n}) = n[\boldsymbol{P}\mathrm{diag}(\boldsymbol{\tau})\boldsymbol{P}' - \boldsymbol{\pi}\boldsymbol{\pi}']$, the marginal variance becomes

$$\mathrm{Var}(\boldsymbol{y} \mid \boldsymbol{P}) = n[\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{P}\mathrm{diag}(\boldsymbol{\tau})\boldsymbol{P}' + \boldsymbol{P}\mathrm{diag}(\boldsymbol{\tau})\boldsymbol{P}' - \boldsymbol{\pi}\boldsymbol{\pi}'] = n\boldsymbol{\Omega}(\boldsymbol{\pi}).$$

In this model the under-dispersion due to mixing is exactly balanced by over-dispersion caused by randomness of the sample sizes in the components of the mixture.

Ad The description of this model is rather similar to the one above, except that 20 rather than 100 balls are drawn from the auxiliary urn and the number of draws from urn 1 equals the number of balls marked with 1 multiplied by 5. Dispersion here is "more than normal". Let m_i denote the number of balls drawn with the digit i, then the number n_i of balls to be drawn from the ith urn will be five time bigger, that is $n_i = 5m_i$ so that overall each sample will again be composed of 100 balls. Suppose, in general, that $n_i = km_i$, then $\mathrm{Var}(\boldsymbol{n}) = k^2\mathrm{Var}(\boldsymbol{m})$ $= k^2(n/k)\boldsymbol{\Omega}(\boldsymbol{\tau})$; in words, the variance of the component sample sizes is inflated and thus exceeds the under-dispersion produced by mixing.

Be "Suppose we make several draws of 100 balls ... After each set of 100 draws, the proportion of the balls inside the urn is changed by increasing, in a certain proportion, the balls of the colour which in the last draw came out with a lower frequency ... or, on the contrary, the balls of the colour which came out with the higher frequency". Dispersion here is "more than normal" irrespective of the sign of the correlation. This model could be translated by assuming the $\boldsymbol{y}_t$, the vector of frequencies at time t, has a multinomial distribution $(n, \boldsymbol{p}_t)$, where $\boldsymbol{p}_t$ denotes the actual proportions of balls in the urn at time t and suppose that $\boldsymbol{p}_t = \boldsymbol{\pi} + \alpha(\boldsymbol{y}_{t-1}/n - \boldsymbol{\pi})$; in words, if $\alpha < 0$, we increase the proportion of balls for the colours that came out less frequently than expected (in the long run) and decrease the proportion for the other colours. If $\alpha > 0$, changes are in the opposite direction. Recalling that $\mathrm{E}(\boldsymbol{p}_t) = \boldsymbol{\pi}$, $\mathrm{E}(\boldsymbol{p}_t\boldsymbol{p}_t') = \boldsymbol{\pi}\boldsymbol{\pi}' + \mathrm{Var}(\boldsymbol{p}_t)$, the marginal variance may be computed as follows

$$\mathrm{Var}(\boldsymbol{y}_t) = n[\mathrm{E}(\boldsymbol{\Omega}(\boldsymbol{p}_t)) + \left(\frac{\alpha}{n}\right)^2 \mathrm{Var}(y_{t-1})] = n[\boldsymbol{\Omega}(\boldsymbol{\pi}) + \alpha^2 \frac{n-1}{n} \mathrm{Var}(\boldsymbol{y}_{t-1})].$$

It follows that the variance of $\boldsymbol{y}_t$ is greater than that under multinomial sampling, irrespective of the sign of α, that is for positive or negative autocorrelation.

Bf, Bg In these two models, he assumes that each sample of n observations is taken sequentially and that, after drawing each ball, the proportions of balls in the urn are changed to induce positive or negative serial correlation within each sample. Here, according to Gini, dispersion is "more than normal" when correlation is positive and "less than normal" when correlation is negative. Here it is convenient to generalize, slightly, Gini's settings concerning the specific mechanism for inducing positive correlation. Let $\boldsymbol{\pi}$ be the vector containing the proportions of balls of different colours at the beginning and $\boldsymbol{x}_t$ be the result of the tth draw; this is a vector of 0's except for a 1 in the position corresponding to the colour of the ball drawn at time t. Assume that $\boldsymbol{x}_t$ has a multinomial distribution with total 1 and vector of probabilities $\boldsymbol{p}_t = \boldsymbol{\pi} + \alpha(\boldsymbol{x}_{t-1} - \boldsymbol{\pi})$, with $\mid \alpha \mid < 1$ and $\boldsymbol{x}_0 = \boldsymbol{\pi}$. Let $\boldsymbol{y} = \sum \boldsymbol{x}_t$; it can be easily shown by recursion that $\mathrm{E}(\boldsymbol{x}_t) = \boldsymbol{\pi}$, so that $\mathrm{E}(\boldsymbol{y}) = n\boldsymbol{\pi}$. Again by a recursive argument it can be shown that $\mathrm{Cov}(\boldsymbol{x}_t, \boldsymbol{x}_{t-k}) = \alpha^k \boldsymbol{\Omega}(\boldsymbol{\pi})$. Then, by using the standard expression for computing the variance of a sum of correlated variables, we obtain

$$\begin{aligned}\mathrm{Var}(\boldsymbol{y}) &= \sum_1^n \mathrm{Var}(\boldsymbol{x}_t) + \sum_2^n Cov(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) + \sum_3^t Cov(\boldsymbol{x}_t, \boldsymbol{x}_{t-2}) + \cdots \\ &= n\boldsymbol{\Omega}(\boldsymbol{\pi}) + \alpha(n-1)\boldsymbol{\Omega}(\boldsymbol{\pi}) + \alpha^2(n-2)\boldsymbol{\Omega}(\boldsymbol{\pi}) \ldots \ldots\end{aligned}$$

The result follows because the second term has the sign of α, the third term is smaller than the second in absolute value and each new term is smaller than the previous one in absolute value.

In the discussion that follows the presentation of the above models, Gini calls features of convergence those which produce lower dispersion and features of divergence those who produce increased dispersion and notes that, if a given data set has normal dispersion but we have detected that certain features of divergence are operating, then there should also be features of convergence operating in the opposite direction, and a careful investigation should reveal them.

2.2 TESTING FOR SERIAL CORRELATION

Three different ways of testing against model Model Be) are presented in Gini (1908), pages 159-165; they are based on data from the town of Berlin. By breaking the 6 years from 1899 to 1904 into intervals of 10 days each, he obtains a time series of 219 observations made of the proportions, say Y_t, of male births out of 100 female births. The main objective was to show the lack of negative autocorrelation that Gini calls "compensatory tendency", this feature, if present, could have been interpreted as evidence that "nature" was operating for correcting her own mistakes to prevent the sex ratio to become unbalanced, as suggested by Arbuthnot (1712)(Gini, 1908, pag. 87).

Two summary measures of serial dependence are computed: the correlation coefficient between Y_{t-1} and Y_t and a measure which Gini calls the "correlation index", apparently of his own invention. This index is computed by counting the proportion of times that Y_{t-1} and Y_t are both smaller or greater than the overall average divided by the total number of 218 pairs.

After grouping observations into 5 adjacent categories, say $C_1, \ldots C_5$, as in the first column of his Table XXXV, he also computes the conditional averages $E(Y_t \mid Y_{t-1} \in C_j)$, $j = 1, \ldots 5$, displayed in the second column of Table XXXV, together with the marginal mean in the last row. He argues that, under the assumption of a "compensatory tendency", that is negative correlation, the conditional averages should be roughly decreasing, which does not happen.

Tavola XXXV.ª

Quando in una decade nacquero % femmine maschi	Il rapporto sessuale (m. % f.) della decade seguente è in media di
1	2
I. — 90 - 97	105.06
II. — 97 - 102	106.97
III. — 102 - 106	105.46
IV. — 106 - 113	106.42
V. — 113 - 122	105.62
Media dei rapporti sessuali delle 219 decadi	106.10

Figure 1: Reproduction from Gini (1908): for the number Y_{t-1} of female births for every 100 males during each period of 10 days intervals from 1899 to 1904 in the town of Berlin, the table gives the average of Y_t (column 2) when Y_{t-1} belongs to a given category as specified in column 1.

To collect further evidence against model Be), a 5×5 contingency table with categories of Y_{t-1} by row and those of Y_t by column is constructed together with an independence table that he derives by a direct argument. Then observed frequencies are compared with the expected ones in an informal way leading him to conclude that there is no evidence of a "compensatory tendency", that is, in modern words, there is no evidence of negative autocorrelation. Negative autocorrelation, if present, could have been interpreted as evidence that some kind of Divine Providence was acting to keep sex ratio within a narrow band, as claimed by Arbuthnot (1712).

2.3 TESTING FOR OVER-DISPERSION

The initial part of Chapter X is devoted to an informal test of model Aa) by using data on births from Saxony and the town of Dresden. Both data sets contain records of families who had a child in a given decade. Gini is aware of the fact that the same family may appear more than once in these records, however, he explains that, when this happens, that family will be counted within distributions with different totals. He also describes in detail his procedure: He first computes an overall estimate of the probability of a male birth by using the overall proportion of male births within each data set (Saxony or Dresden). Then, for the families who, at some time point, reached the number of n children, he computes the probability distribution under the binomial model, P_1 in the table below, together with the ratios between the observed proportions, P_2, and the binomial probabilities. An extract of the table for Dresden is reprinted in Figure 2. The results were also visualized in a plot where each horizontal line, indexed by the number of children, displays, in the vertical axis, variations in the ratio between observed frequencies and binomial probabilities. Note that to save space both axes are translated to accommodate 5 different curves on the same plot.

Tavola XCIII.[a]

Frequenza delle varie combinazioni dei sessi nelle famiglie con 2-6 figli in Dresda (1891-1901)

Combinazione dei sessi	Valori di P_1	Valori di P_2	Valori di $\frac{P_2}{P_1}$
1	2	3	4
Famiglie con 2 figli			
2 M.	0.265689	0.269855	1.015
1 M. 1 F.	0.499522	0.495419	0.991
2 F.	0.234789	0.231726	0.999
Famiglie con 3 figli			
3 M.	0.136949	0.142788	1.042
2 M. 1 F.	0.386219	0.389102	1.007
1 M. 2 F.	0.363065	0.352703	0.971
3 F.	0.113767	0.115407	1.014
Famiglie con 4 figli			
4 M.	0.070590	0.072427	1.026
3 M. 1 F.	0.265434	0.271070	1.021
2 M. 2 F.	0.374285	0.362135	0.967
1 M. 3 F.	0.234565	0.233957	0.997
4 F.	0.055126	0.060411	1.096
Famiglie con 5 figli			
5 M.	0.036386	0.039585	1.087
4 M. 1 F.	0.171022	0.173667	1.015
3 M. 2 F.	0.321541	0.318127	0.989
2 M. 3 F.	0.302266	0.291214	0.963
1 M. 4 F.	0.142074	0.143012	1.006
5 F.	0.026711	0.034395	1.287
Famiglie con 6 figli			
6 M	0.018755	0.019719	1.051
5 M. 1 F.	0.105785	0.101680	0.961
4 M. 2 F.	0.248607	0.240226	0.966
3 M. 3 F.	0.311605	0.319445	1.025
2 M. 4 F.	0.219696	0.211763	0.963
1 M. 5 F.	0.082609	0.083477	1.071
6 E.	0.012943	0.018690	1.444

Figure 2: Reproduction from Gini (1908) concerning the town of Dresden: for the families with n births, $2 \leq n \leq 6$, and a given composition in males and females, P_1 is the expected frequency under the binomial model and P_2 is the observed proportion.

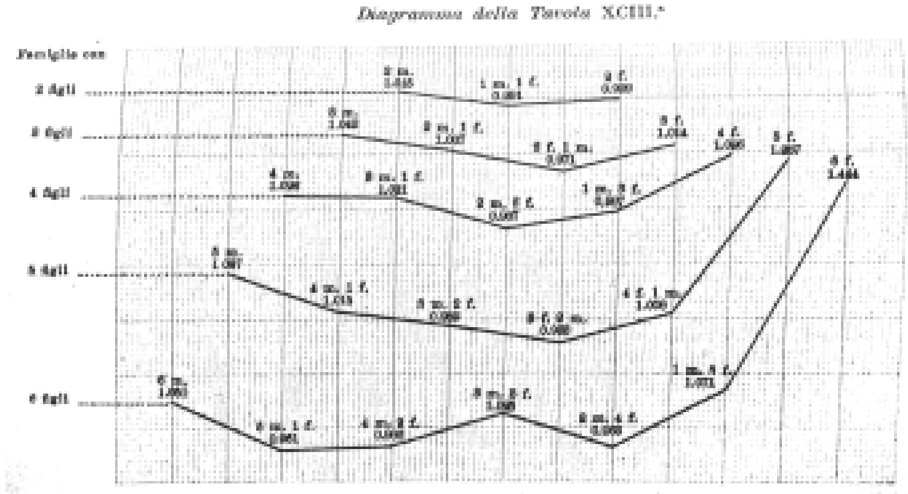

Figure 3: Reproduction from Gini (1908): for families that, at some point, had a number of children between 2 and 6, the lines connect the values of the ratio between observed proportions and binomial probabilities on the *Y* axis depending on number of females on the *X* axis; data from Dresden.

3. THE 1911 PAPER ON PREDICTIVE PROBABILITY

Gini (1911) starts by recalling, in his own way, known results concerning the predictive probability for dichotomous outcomes to be applied to data on sexes in human births. Let S_n denote the number of times that event A happens in n trials, for instance, the number of males out of n births for a given couple. The predictive probability may be written as $P(S_m = y \mid S_n = x)$, in words, the probability that a couple who had x males in the first n births will have y males in the next m births; let also θ denote the probability of a male birth in a specific context. His equation (III) on page 79, which he derives by applying rules of conditional probability and Bayes theorem, may be translated as follows:

$$P(S_m = y \mid S_n = x) = \frac{P(p = \theta)P(S_m = y \mid p = \theta)P(S_n = x \mid p = \theta)}{\sum_\theta P(p = \theta)P(S_m = y \mid p = \theta)P(S_n = x \mid p = \theta)}.$$

He explains why, in his opinion, this formula is useless in practical applications: for each possible value taken by θ we should know (or be able to assign values to) the prior probability $P(p = \theta)$ that this happens; an almost impossible task from a non subjectivist point of view. He recalls that, to overcome the problem, it is usual to assume that θ can take any possible value between 0 and 1 with the same probability, that is $P(p = \theta)$ is constant, an assumption which, later in his paper, he shows to be restrictive and unjustified most of the times.

In an attempt to find an alternative route which does not require assumptions on the prior probabilities, by simple manipulations, he writes

$$P(S_n = x) = \sum_{y=0}^{m} P(S_n = x, S_m = y) = \sum_{y=0}^{m} P(S_{n+m} = x+y)P(S_m = y \mid S_{n+m} = x+y),$$

to obtain his basic equation

$$P(S_m = y \mid S_n = x) = \frac{P(S_{n+m} = x + y)P(S_m = y \mid S_{n+m} = x + y)}{\sum_{y=0}^{m} P(S_{n+m} = x + y)P(S_m = y \mid S_{n+m} = x + y)}. \quad (1)$$

3.1 TESTING FINITE EXCHANGEABILITY

To understand the importance of equation (1), it is useful to consider the application that Gini had in mind: to compute the probability that a couple with x males in the first n births had a male in the next birth. Because he expected to be able to estimate the $P(S_{n+m} = x+y)$s from recoded data on families with a given number of births in a given region and a given period, he remained with the problem of estimating the conditional probabilities $P(S_m = y \mid S_{n+m} = x + y)$. The solution that he devised constitutes, probably, the most intriguing and innovative contribution of his paper.

Essentially, he states that, under the assumption that things remain unchanged during the $n + m$ trials, $P(S_m = y \mid S_{n+m} = x + y)$ equals the probability that, if one draws without replacement m balls from an urn containing $x + y$ white balls and $n + m - x - y$ black balls, y of these will be white, that is the formula for the hypergeometric distribution. It is well known that the above result is implied by the assumption of finite exchangeability within the $n + m$ events; on this basis, Forcina, in the discussion of Draper et al. (1993), suggests that Gini's argument may be seen as an anticipation of the assumption of exchangeability.

The phrasing of Gini's paper may appear a little obscure and in 18th century style compared with De Finetti (1937). The note at the bottom of page 82 may give the impression that Gini, to derive the hypergeometric formula, is using the assumption of a binomial distribution. However, looking at his equation (XIII) on page 87, it seems that he is considering the binomial distribution conditionally on a given realisation of the prior distribution. Both his equation (XIII) and the derivation on top of page 89, seem to imply some intuitive form of De Finetti's representation theorem. Further evidence in this direction is provided by his discussion of the possible values of the prior probability of an event when, on top of page 79, he says that the case when this probability has a well defined value corresponds to the very special case where the distribution assigning prior probabilities to the range of possible values is degenerate.

It is quite clear (see Sec. 7 on page 85) that, in Gini's opinion, finite exchangeability was not something to be taken for granted: "We can expect that these probabilities", that is those computed by assuming finite exchangeability, "should not differ in a systematic way from the observed frequencies, if the assumption we started from is correct". In particular, on the basis of the large data sets available for certain cities and regions in Germany, he could compute, from observed frequencies, among the families with x males in the first n births, the proportion of those with a males in the next birth. Again, on page 85 he clarifies that "the comparison between observed frequencies and theoretical probabilities gives us a way to verify whether the probability of the event A remains constant during the $n + m$ observations". The issue here is whether the probability of having a male birth by a given couple does or not depend on the sex of previous births.

The results of these comparisons are presented in Table I (data from Saxony) and Table II (city of Dresden) in the second part of the paper. Both tables display data obtained from records of families who reported the birth of a child during a decade; a partial reproduction of Table II is shown in Figure 4 for the families who had n births before the new one, column 1 lists the different possible outcomes, column 2 contains the frequency distribution, that is the number of families with a given outcome; column 3 contains the conditional probabilities of a new male birth under the assumption of finite exchangeability while column 4 contains the same conditional probabilities estimated directly from the data.

Except for families with two children, it emerges rather clearly that among families with an excess of males in the initial n births, the probability of a male in the next birth computed under the assumption of finite exchangeability is always greater than the observed proportions. In addition, the inequality is reversed among families with an excess of females in the initial n births, that is discrepancies are clearly not at random but seem

102

TAVOLA II.

Nascite nella città di Dresda (1891-1901)

Combinazione dei sessi nelle nascite antecedenti		Numero delle coppie osservate	Frequenza dei maschi nella prossima nascita: in teoria π'	Frequenza dei maschi nella prossima nascita: nel fatto π''	$\pi'' - \pi'$
1		2	3	4	5
1 m.	—	15,330	*0,52139*	0,52253	+ 0,00114
—	1 f.	14,357	*0,51346*	0,51459	
2 m.	—	5,062	*0,52401*	0,52232	— 0,00169
1 m.	1 f.	9.222	*0,52453*	0,51908	— 0,00544
—	2 f.	4,233	*0,50464*	0,49516	— 0,00948
3 m.	—	1,774	*0,51662*	0.49944	— 0,01718
2 m.	1 f.	4.678	*0,52892*	0,51903	— 0,00989
1 m.	2 f.	4,322	*0,50785*	0,50440	— 0,00345
—	3 f.	1,459	*0,49192*	0,49349	+ 0,00157
4 m.	—	617	*0,53264*	0,53160	— 0,00104
3 m.	1 f.	2,197	*0,52194*	0,52344	— 0,00850
2 m.	2 f.	3,012	*0,52208*	0,52756	— 0,00552
1 m.	3 f.	1,926	*0,50450*	0,51402	+ 0,00952
—	4 f.	534	*0,45402*	0,46629	+ 0,01227

Figure 4: Reproduction of Table II from Gini (1911): predictive probabilities that the new child will be a male estimated under the assumption of finite exchangeability, that is according to theory and from raw frequencies, that is de facto.

to follow a well defined pattern. These results are interpreted by Gini as evidence that there should be some kind of natural balancing mechanism in the sense that couples with an initial excess of boys were more likely to have a girl as the next child and viceversa for couples with an excess of girls.

To strengthen his conclusions, Tables VII and VIII concentrate on certain comparisons between expected probabilities and observed proportions. For instance, within families where the number of males is twice the number of girls, the difference between observed and expected is negative and increases in absolute value with n. This is interpreted as follows: not only, due to serial correlation, the probability of another male in families with more males in previous births is smaller than expected under finite exchangeability, but we need bigger families for the discrepancy to emerge more clearly.

It is interesting to compare these results with those in Chapter X in Gini (1908) where, essentially, the same data were used, though they are arranged in a different way. Gini (1908) studies the frequency distribution of sexes among the families who, having had a child in a given period, had a total of v males within a total of n births. Having detected over-dispersion within most of his distributions, he concluded that model Aa) must be operating, that is the overall probability of a male birth differs from one couple to the other. Both the model and the testing procedure in Chapter X ignore the order with which males and females were born within each couple. Gini (1911) instead is about testing models Bf) and Bg), that is serial correlation within each family.

3.2 EXTENSIONS

In order to allow a wider set of comparisons between theoretical probabilities computed under the assumption of finite exchangeability and the corresponding observed proportions, Gini (1911) devised an additional expansion as follows

$$\begin{aligned} P(S_n = x, S_m = y) &= \sum_{z=0}^{t} P(S_n = x, S_{m+t} = y + z) \\ &= \sum_{z=0}^{t} P(S_{n+m+t} = x + y + z, S_{m+t} = y + z) \\ &= \sum_{z=0}^{t} P(S_{n+m+t} = x + y + z) P(S_{m+t} = \\ &\qquad = y + z \mid S_{n+m+t} = x + y + z). \end{aligned}$$

This expression may be computed for different values of t additional births for which data may be available. In this way, for instance, Gini could compute several versions of the predictive probability that a family with two males in the first two births had another male in the third birth, by using observed proportions for families with at least 3 children; clearly, applications of this procedure may be limited because the number of large families is not large enough to provide reliable estimates of $P(S_{n+m+t} = x + y + z)$.

3.3 ON THE ASSUMPTION OF UNIFORM PRIOR DISTRIBUTION

Gini (1911) contains also a critical assessment of the assumption of uniform priors in the context of binary data. First he shows that the expression for $P(S_m = y \mid S_n = x)$ obtained under the assumption of uniform priors coincides with the expression derived under the assumption that the marginal probabilities $P(S_{n+m} = x + y)$ are constant for $y \in [0, m]$. In other words, the same result is implied by two completely different assumptions, a matter worth of further investigation.

First, on page 87, Gini shows, by a counter example, that his assumption that $P(S_{n+m} = x + y)$ does not depend on y, does not imply the assumption of uniform prior distribution. On the other hand, in the special case of $m = 1$, the case of more interest in his applications, he shows that, under the assumption of uniform priors, we must have

$$P(S_{n+1} = 0) = P(S_{n+1} = 1), \tag{2}$$

see page 88. From these results he concludes that assumption (2) is weaker than the assumption of uniform priors, though they both imply an identical expression for the predictive probability.

ACKNOWLEDGMENT

I would like to thank S. Stigler, University of Chicago, for insightful comments

REFERENCES

Arbuthnot, J. (1712). ‘An argument for Divine Providence; taken from the constant regularity observed in the births of both sexes’, *Philosophical Transactions of the Royal Society of London* **27**, 186–190.

De Finetti, B. (1937). ‘La prévision: ses lois logiques, ses sources subjectives’, *Annales de l’Insitut H: Poincaré* **7**, 1–68.

Draper, D., Hodges, J., Mallows, C. and Pregibon, D. (1993). ‘Exchangeability and data analysis (with discussion)’, *Journal of the Royal Statistical Society, A* **156**, 9–37.

Gini, C. (1908). *Il sesso dal punto di vista statistico. Le leggi della produzione dei sessi,* Sandron, Firenze.

Gini, G. (1911). Considerazioni sulla probabilità a posteriori e applicazioni al rapporto dei sessi nelle nascite umane, *Studi economico-giuridici dell’Università di Cagliari,* reprinted in Metron, 1949, vol. 15, 133–172.

Stigler, S. (1986). *The History of Statistics, the Measurement of Uncertainty before 1900,* Belknap and Harward University Press, Cambridge, Massachusetts.