

The voices of boy and girl cathedral choristers – Is there a perceptible difference? And does it matter?

**A.E.Saunders, D.Sc. (Lond.), ASIS., Hon.FRPS.,
Visiting Fellow, Bournemouth University.**

Foreword

In bygone days, sex, politics and religion were regarded as unseemly topics for discussion in polite company. Today, we should perhaps add a fourth topic, mathematics, or more specifically anything that looks remotely like closely reasoned argument. In recent years, attempts have been made to prove by experiment that there is hardly any perceptible difference between the singing of boy and girl choristers. Standard statistical methods were used to analyse the data but the experiment was so poorly designed that these methods are inapplicable, a fact which was not only admitted but subsequently ignored. It is the aim of the present article to provide a critique of the experiment, to present a rigorous method of data analysis, and hence to prove that the original conclusion, that there is a perceptible difference, was seriously understated. It is now quite clear that unless the misapplication of statistical methods and the misleading conclusions thereby obtained go unchallenged, the choral tradition of this country could be further damaged.

Introduction

The questions which form the title of the present article were prompted by two publications [1,2] describing a statistical experiment designed to compare the voices of boy and girl cathedral choristers. A preliminary assessment of these papers indicated that the experiment was flawed and that the conclusions reached must therefore be suspect. It is the purpose of this article to give detailed reasons for these assertions, no trivial task in that both the experiment design and the analysis are fraught with technical difficulty. As far as possible, therefore, the issues central to the aims of those wishing to preserve the choral tradition of this country are addressed in the main text of the article whilst the more formal statistical aspects of the work are to be found in the appendices.

The experiment

The experiment consisted of producing CD recordings of twenty short pieces of choral music which were then played to 189 listeners. The two cathedral choirs participating in the experiment were as nearly identical as possible, differing only in that girls sang the top line in one and boys in the other; the lower parts being sung by the same layclerks. Each choir sang ten out of the twenty pieces, which were played in random order to the listeners. Each listener was requested to state whether a given piece was sung by boys or girls; ‘null’ answers were forbidden. Great attention was paid to the quality of both recording and listening conditions so as to minimize the effect of ‘nuisance’ factors that might tend to obscure any real difference between the sounds produced by the two choirs.

At first sight, the experiment appears to have been designed in an orthodox fashion in spite of some very considerable practical difficulties. However, more careful consideration shows that there are some problems with the design that could have affected the results obtained, the way in which they were analyzed and therefore the validity of the conclusions. In detail, the pieces of music performed by the two choirs were different and may have reflected a gender bias (a

point made by the original authors). Most musicians will recognize that the conductor can have a major effect on the way in which a piece of music is performed and therefore perceived by the listener; in the present case, no mention was made about the conductor, or whether the choirs may have been treated differently. Any pieces that were accompanied, even though the organ accompaniment was provided by the same instrument, may not have involved the same registration. The two commercially-made recordings from which the twenty pieces were selected were recorded at different times of year so that the vocal quality of the two top lines may have been differently affected by such seasonal phenomena as sore throats or allergies. It should therefore be added that randomization is not always advantageous, particularly if nuisance factors are correlated with main effects [3] as might possibly have occurred in the present case.

Lastly, it is difficult to appreciate exactly how the individual trials could have been performed without another form of complication, usually referred to as ‘sampling without replacement’ [4]. Presumably, the listeners had been informed that of the twenty pieces of music, half were of each choir and that the order of playing had been randomized. In other words, anyone listening to the set of twenty would have had to ensure that they ended up with ten of each kind. Therefore, towards the end of each trial of twenty pieces, the listeners’ decisions may no longer have been entirely based on perceived sound quality but might have been influenced by the need to ‘balance the numbers’. The more usual approach to this kind of testing would be to ‘block’ the samples in pairs, one with each choir, but in random order within each block [5]. Ideally, the only difference *within* each block would have been the choirs, *not* the pieces being sung, which would have been the only difference *between* the blocks.

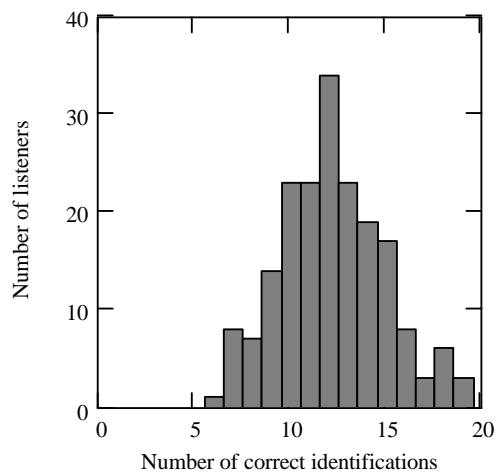


Figure 1. The **height** of each column of the histogram represents the number of listeners out of the total of 189. The **position** of a column along the horizontal axis indicates the number of times out of twenty that correct identifications were made of the sex of the top line. (Data published by Howard and Szymanski [2].)

An example of such a blocked design could have the form: BG GB GB BG BG GB GB BG GB, again involving a total of twenty samples, ten with each choir, but only ten different pieces. Ideally, the individual blocks of two would also be presented in a different randomized order to each listener so as to minimize the effect of any systematic trend, e.g., listener fatigue. The procedure would then be to invite listeners to decide the order within each block. In this design, the maximum number of correct identifications would be ten - the second of

each pair being determined by the first, so that the problem of sampling without replacement, would be eliminated, if indeed the original design was so affected.

Experimental results.

Given that the experiment was actually designed and conducted in what may now seem to have been a somewhat curious fashion, we now turn to the possibility of finding a more orthodox approach to the analysis of the data. The results obtained in the experiment outlined above were published in two equivalent but entirely consistent forms, as shown in figs.1 and 2. In the former, the bell-shape of the histogram is what a statistician would expect for an experiment in which the number of *listeners* who correctly identified the singers of a given number of *pieces* is shown as a function of that number. He would expect such a shape because, roughly speaking, the probability of *guessing* the correct answer for any particular piece of the twenty is exactly 0.5 and therefore the probability of a listener correctly guessing a certain number of times out of the twenty is readily calculable. The probability distribution thus deduced is usually called a *binomial distribution*, and is always bell-shaped.

In the present case, however, instead of the histogram being centred on ten correct identifications, as it would be if the identifications were purely a matter of guesswork, it is actually shifted a little to the right. It is this shift that is all important in attempting to reach a statistically significant judgement as to whether the choirs are distinguishable.

Statistical inference

According to Mark Twain, "there are three kinds of lies; lies, damned lies, and statistics". In essence, there are also three methods commonly used amongst statisticians for arriving at deductions based on experimental data. A detailed account of these methods would be inappropriate here, particularly as they are very well documented in the technical literature. However, one point that should not be overlooked is that results deduced from experiment by these methods are a little less definite than simple facts. In the present context, we mean statements that may well be true, but with some indication of just how credible they are. This lack of certainty is because the experimental data itself is indeterminate and is affected by chance. For this reason it is necessary to gather together a large body of data so that to some extent the nuisance factors mentioned earlier can be averaged out. In the present case, therefore, possible statements will take the form "that with a certain (stated) probability, there is a perceptible difference between the choirs". The three methods mentioned above provide the means of calculating this probability and fall under the headings *significance testing*, *confidence intervals*, and the '*goodness-of-fit*' between two collections of data.

The bell-shaped distribution of experimental results shown in fig.1 applies to a trial involving 189 listeners. It was found that the mean number of correct identifications was 12.164 out of the twenty-one possibilities (including zero) and the width of the bell shape (the standard deviation) was found to be 2.724. However, for each listener of a sample of twenty pieces the mean of *their particular sample* will generally be different from 12.164. It is found that the *mean* of the sample means is also 12.164 but the standard deviation of the distribution of such means will be only $2.724/\sqrt{20}$ or 0.609. This reduction in the width of what is still a bell-shaped distribution, the tendency of sample means to gather about the overall mean, is a consequence of the Central Limit Theorem which not only quantifies this tendency but does so in a way that allows further calculations to be made. In particular, the narrowing of the bell

shape allows better discrimination to be made between two sets of experimental data, or between experimental and purely hypothetical data sets.

In the first paper cited, it was stated that “Of the twenty test items, the mean number of correct answers was 12.21 (with a standard deviation of 2.76), being significantly better than chance ($p = 0.05$).” The form of this statement suggests that significance testing was the method of inference used. Although the previous analyses of the experimental data are somewhat equivocal on the matter of statistical inference - in one case appearing to assign a significance level of 99% and in the other of 95%, which incidentally is often regarded as inconclusive [6] - neither can be regarded as entirely satisfactory because the full role of the null hypothesis was not considered. Although it is customary to adopt a significance level that corresponds to the probability representing a given degree of skepticism, it is “always best to state the probability itself rather than to say that the result is significant or not significant at some conventional level”[7].

In the second paper cited, an algebraic expression is given for the “ $100(1-\alpha)\%$ confidence interval for the true mean” and that “all listener groups represented have an ability that is better than chance at the 0.01 level to identify the sex of the choristers singing the top line”. Remarkably, these two papers have an author in common and use the same experimental data, but use rather different methodology and terminology to present what appears to be exactly the same conclusion, *viz.* that the two choirs are perceptibly different. Even more remarkably, the usual statistical approach of testing the so-called ‘null hypothesis’, *that there is actually no perceptible difference between the choirs*, was ignored in both papers.

Testing the null hypothesis

If there were no perceptible differences between the two choirs, and the null hypothesis were valid, the bell-shaped histogram shown in fig.1 would be centred on ten correct identifications. (The null hypothesis is usually written as $H_0: \mu = \mu_0$ in which μ and μ_0 are the mean values of the experimental and hypothetical distributions of correct answers, respectively.) Rather than consider separately either of the narrower distributions obtained as a result of sampling, we can consider the sampling distribution of the difference between the experimental and hypothetical means (Appendix 1). The appropriate test statistic is then given by dividing the difference between the two *sample* means by its standard error [8]. The value of the test statistic found in this way (2.78) indicates that if the null hypothesis is correct, we would expect this value or greater to occur *purely by chance* in only about three cases out of a thousand (equivalent to a confidence level of 99.7%). We therefore reject the null hypothesis so that, by default, we are then inclined to accept the alternative hypothesis ($H_1: \mu > \mu_0$), that *there is a perceptible difference between the two choirs*.

A dilemma

We have now arrived at what appears to be an impasse in that we have three rather different answers to the question of whether the two choirs are perceptibly different. Fortunately, there is another method of data analysis that is well-known to statisticians. The answer lies in the work of two Victorians, the mathematician Karl Pearson (1857-1936) and the biologist, W.F.R.Weldon (1860-1906), the joint founders [9], in 1901, of the journal *Biometrika*. The question that Weldon set out to answer concerned the possibility of determining whether dice are biased. To this end, he threw twelve dice 26,306 times counting fives and sixes as ‘successes’. The probability of such a success for unbiased dice is expected to be exactly a

third but Weldon found that for his dice a probability of 0.3377 gave a better fit to the experimental data. According to one authority [10], “the agreement (between experiment and the results expected for unbiased dice) looks good, but for such extensive data it is really very bad. Statisticians usually judge closeness of fit by the chi-square criterion. According to it, deviations as large as those observed would happen with true dice only four times out of 10,000”.

In our case, instead of throwing twelve dice thousands of times, twenty pieces of music were judged 189 times. And instead of counting fives and sixes as successes (with a notional probability of a third), a correct identification of the top line was deemed a success (with a notional probability of a half). Although the experiments on dice and choirs may appear to be totally different, in terms of statistical theory they are almost identical. The only difference is that whereas the dice experiment involved putatively identical dice, and therefore having *equal* probabilities of success, that with the choirs involved different pieces of music, and therefore *different* probabilities of success. Whilst the former is a set of independent Bernoulli trials of *equal* probability, the latter would be regarded as a set of such trials but with *different* probabilities, what we might call an *inhomogeneous* Bernoulli trial (Appendix 2).

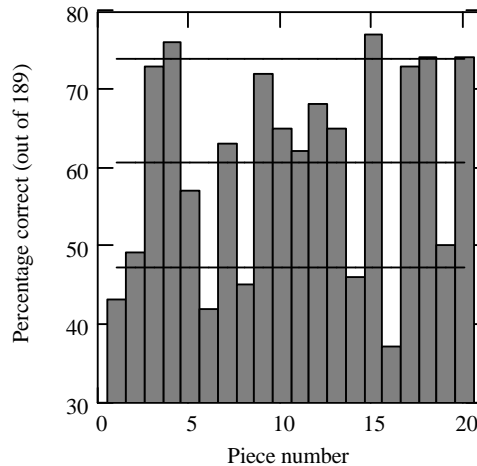


Figure 2. The height of each column of the histogram represents the percentage of correct identifications made by the 189 listeners of each sample piece. The position of the column along the horizontal axis indicates the number of the piece of music. (Data published by Welch and Howard[1].) The three horizontal lines indicate the mean (centre) and the mean plus and minus the standard deviation.

It is hardly surprising to find that data obtained from a flawed design can only be adequately analysed by using a rather more elaborate statistical technique than is to be found in standard texts. However, although this difficulty was recognized in the present case [2] but was subsequently ignored, it can be satisfactorily resolved. A rigorous theoretical treatment of the inhomogeneous Bernoulli trial shows that Pearson’s chi-square (χ^2) test can be used to examine the validity of the alternative hypothesis. To this end, we use the mean and standard deviation determined experimentally to set up a new hypothesis H_2 , viz., that there *is* a perceptible difference between the two choirs, but that the experiment was actually an *inhomogeneous* rather than a simple homogeneous Bernoulli trial having the same probability

($12.164/20 = 0.6082$) for all the individual trials (Appendix 3). The hypothesis that the experiment conforms to the latter trial is denoted by H_3 .

A χ^2 test in which the experimental frequencies are compared with those based on H_2 can then be readily carried out (Appendix 4). It will be recalled that H_0 was rejected because had it been valid the probability of obtaining purely by chance the experimentally-determined value of the test statistic was extremely small (0.0027). In contrast, the value obtained for the χ^2 statistic (10.064) corresponds to a probability of 0.26 (or almost a hundred-fold greater) so that we should certainly accept H_2 . This conclusion is reinforced by the fact that in testing H_3 , the hypothesis that the experiment was a simple homogeneous Bernoulli trial, the value obtained for the χ^2 statistic (35.904) corresponds to a probability of 4.12×10^{-5} , indicating that the experiment is much better described as an *inhomogeneous* rather than a homogeneous trial. Since H_2 is merely a refinement of the alternative hypothesis H_1 , we not only reject H_0 , the original null hypothesis, *but now have positive grounds for accepting H_1 rather than accepting it by default.*

Conclusion

From the foregoing analysis, there is little doubt that the answer to the first question posed in the title is that there *is* a perceptible difference between the voices of boy and girl cathedral choristers. Although the original analyses of a poorly-designed experiment were unsatisfactory, they nonetheless led to the same general conclusion reached here, albeit remarkably understated. Those who support the traditional cathedral choir or, for that matter, the Anglican choral tradition as it existed in the parishes, are often accused of being ‘sexist’ by those who use such puerile name-calling as a substitute for rational argument. An example of this attitude was attributed to one of the authors cited below [1,2] who is quoted [11] as having said that “this argument is primarily sexist” and is “nothing to do with music”. A more objective view would surely have been that even if there is a perceptible difference between the voices of boy and girl cathedral choristers, one is no ‘better’ than the other, just different. However, as we have now established that there really is a perceptible difference, it is necessary to make the point that most of the music that constitutes the traditional choral repertoire was intended to be sung by boys and men. To replace the boys with women or girls is therefore to present the audience with a sound other than that intended by composers, the first rank of whom would have regarded the boy’s voice as something very special and particularly appropriate for the rendition of ecclesiastical music.

With regard to the second question, it may well be the case that many will be unconcerned about the admission of girls into cathedral choirs, perhaps arguing that such choirs are elitist, of interest to only a tiny minority of the general population and therefore hardly worth thinking about. However, those who *are* concerned about the choral tradition will find the claim, that boys’ and girls’ voices are indistinguishable, not only false but will object to its use in promoting second, unnecessary choirs, particularly as the establishment of such choirs has led to a host of other problems both musical and financial.

As with so many social issues, it is all too easy to consider any particular issue in isolation, thereby arriving at both a simplistic view of the problem and its possible solutions. We are told that girls should have the same opportunities as boys, regardless of the long-term or indirect consequences. Unfortunately, equality of opportunity seldom, if ever, leads to equality of outcome, particularly if the outcome in one generation proliferates and feeds on itself to produce even greater inequalities in subsequent generations. A processes of this kind

is generally unstable and in social terms might be expected to result in unfairness and a growing sense of injustice. It is for this most fundamental of reasons that the mindless insistence on equal opportunities, and political correctness in all its forms, should be opposed.

References

1. Welch, G.F. and Howard D.M., *Gendered Voice in the Cathedral Choir*, Psychology of Music, ed. S.Hallam, Soc.Res.Psych.Mus.Mus.Ed., **30**, 102-120 (2002).
2. Howard, D.M. and Szymanski, J.E., *Listener Perception of Girls and Boys in an English Cathedral Choir*, Proceedings of 6th International Conference on Music Perception and Cognition, Keele University, (2000).
3. Chatfield, C., *Statistics for Technology*, Chapman & Hall, London, 3rd Ed., (1983), p.230.
4. Stuart, A and Ord, J.K., *Kendall's Advanced Theory of Statistics*, Griffin, London, (1983), vol.1, pp.166 and 315.
5. Box, G.E.P., Hunter, W.G. and Hunter, J.S., *Statistics for Experimenters*, Wiley, New York, (1978), p.97.
6. Chatfield, C., *ibid.*, p.140.
7. Box, G.E.P., Hunter, W.G. and Hunter, J.S., *ibid.*, p.109.
8. Bulmer, M.G., *Principles of Statistics*, Dover, New York, 1979, p.151.
9. Bell, E.T., *The Development of Mathematics*, McGraw-Hill, New York, 1940, p.543.
10. Feller, W., *Probability Theory and its Applications*, Wiley, New York, 1950, vol.1, p.106.
11. Utton, T., *Proof that choirgirls are not inferior to choirboys*, Daily Mail, 9th September 2003.

Appendices

1. Comparison of sample means

If there were no perceptible difference between the two choirs - the *null hypothesis* - we would expect the mean m_0 to be 10 because the decision as to whether boys or girls sang a particular piece in the selection of twenty would have been no more than a guess. On average, therefore, the decision would have been correct in half the cases. In other words, the probability p of being correct (or incorrect) would have been 0.5. The standard deviation of this reference data, i.e., the discrete values corresponding to those obtained by experiment, is given by $\sigma_0 = \sqrt{np[1-p]}$, in which n is the number of pieces in the recording. As this number is twenty, we find that $\sigma_0 = \sqrt{5}$ or 2.24. The corresponding mean m and standard deviation σ for the experimental data are reported [1,2] as 12.21 and 2.76, respectively. An appropriate test statistic [Bulmer, M.G., *Principles of Statistics*, Dover, New York, 1979, p.151] is then

$$t = (m - m_0) / \sqrt{(\sigma^2/n + \sigma_0^2/n)}$$

As the sample size n , taken to be the same for each sample, is sufficiently large, t is approximately a standard normal deviate z since the variance of $(m - m_0)$ is $(\sigma^2/n + \sigma_0^2/n)$. We then find that the test statistic has the value 2.75. The probability of obtaining this value or greater is then found from standard statistical tables to be about 0.003. (It should be noted that based on the published data $m = 12.164$ rather than the figure given of 12.21 and $\sigma = 2.724$ rather than 2.760, minor discrepancies that only reduce the probability from 0.0030 to 0.0027.) Although it is customary to adopt a 'significance level' that corresponds to the probability representing a given degree of skepticism, it is "always best to state the probability itself rather than to say that the result is significant or not significant at some

conventional level". [Box, G.E.P., Hunter, W.G. and Hunter, J.S., *Statistics for Experimenters*, Wiley, New York, (1978), p.109.]

2. Independent Bernoulli trials with different probabilities

If a trial can have only one of two different outcomes, success or failure, it is said to be a Bernoulli trial. In the present case, we would be making the assumption that the pieces of music are all equally likely to be correctly identified. If the probabilities of success and failure are denoted by p and q , respectively, then $p + q$ must equal unity. For n independent trials, the probability of r successes is the coefficient of t^r in the binomial expansion of a function usually referred to as the probability generating function (pgf) given by

$$g(t) = (pt + q)^n = \sum_{r=0}^n {}^n C_r p^r q^{n-r} t^r \quad (1)$$

Let us suppose that instead of the probability of success p remaining constant throughout the n trials, each trial has a different probability of success denoted by p_i . In this case, the n factors of the probability generating function in eq.1 are all different so that for such an *inhomogeneous* Bernoulli trial

$$g(t) = \prod_{i=1}^n (p_i t + q_i) \quad (2)$$

Since $p_i + q_i$ is unity, this product can be rewritten as

$$g(t) = \prod_{i=1}^n (p_i t + 1 - p_i) = \prod_{i=1}^n \{1 + (t-1)p_i\} = \prod_{i=1}^n \{1 + (t-1)p + (t-1)\pi_i\} \quad (3)$$

in which the p_i have been replaced with $(p + \pi_i)$ and p is now the mean value of the individual probabilities p_i . We can now remove a term $[1 + (t-1)p]$ from each factor of the product to obtain

$$\begin{aligned} g(t) &= \{1 + (t-1)p\}^n \prod_{i=1}^n \{1 + [(t-1)\pi_i]/[1 + (t-1)p]\} \\ &= \{1 + (t-1)p\}^n \cdot \{1 + \sum_{i=1}^n \pi_i [(t-1)]/[1 + (t-1)p] + \sum'_{i \neq j} 2\pi_i \pi_j [(t-1)^2]/[1 + (t-1)p]^2 + \dots\} \\ &= \{1 + (t-1)p\}^n \cdot \{1 + [(t-1)]/[1 + (t-1)p] \sum_{i=1}^n \pi_i + 2[(t-1)^2]/[1 + (t-1)p]^2 \sum'_{i \neq j} \pi_i \pi_j + \dots\} \\ &= \{1 + (t-1)p\}^n \cdot \{1 + 2[(t-1)^2]/[1 + (t-1)p]^2 \sum'_{i \neq j} \pi_i \pi_j + \text{terms in } \pi^3 \text{ etc.}\} \quad (4) \end{aligned}$$

since $\sum \pi_i$ is zero. The remaining summation in eq.4 is over all products of π_i and π_j for which $i \neq j$. If $\sum' 2\pi_i \pi_j$ is replaced with θ and eq.4 is rearranged, we have that

$$g(t) = \{1 + (t-1)p\}^n + \theta(t-1)^2 [1 + (t-1)p]^{n-2} + \dots \quad (5)$$

An important property of the pgf is that if the parameter t is replaced with $\exp(\alpha)$ and the resulting function is expanded as a power series in α , the coefficient of $\alpha^r/r!$ is the r th moment of the probability distribution; the function $g(e^\alpha)$, usually denoted by $m(\alpha)$, is called the moment generating function (mgf). When these operations are carried out, eq.5 becomes

$$\begin{aligned} m(\alpha) &= 1 + \{1 + (\alpha + \alpha^2/2! + \dots)p\}^n + \theta(\alpha + \alpha^2/2! + \dots)^2 [1 + (\alpha + \alpha^2/2! + \dots)p]^{n-2} + \dots \\ &= 1 + (1 + n(\alpha + \alpha^2/2! + \dots)p + \frac{1}{2}n(n-1)(\alpha + \alpha^2/2! + \dots)^2 p^2 + \dots + 2\theta \alpha^2/2! + \text{terms in } \alpha^3 \text{ etc.} \end{aligned}$$

so that by inspection we have that the first and second moments are, respectively,

$$\mu_1 = np \quad (6)$$

$$\text{and } \mu_2 = np + n(n-1)p^2 + 2\theta \quad (7)$$

The mean is therefore unaffected by the differences in probability of the individual trials provided p is taken to be the mean of these probabilities. On the other hand, the variance of the number of successes out of the total number of individual trials is given by

$$\sigma^2 = \mu_2 - \mu_1^2 = np + n(n-1)p^2 + 2\theta - n^2p^2 = np - np^2 + 2\theta = npq + 2\theta = \sigma_0^2 + 2\theta \quad (8)$$

in which σ_0^2 is the variance when all the trials have the same probability, for when all the terms π_i are zero, θ is also zero. It also follows that a simple measure of the variability of probability amongst the individual trials of an inhomogeneous Bernoulli trial is given by half the difference between the *observed* variance σ^2 and the hypothetical variance $\sigma_0^2 = npq$ of a homogeneous Bernoulli trial. As the means of the two distributions are equal, from eq.6, and hence $np = m$ (the mean of the *experimental* data), $p = m/n$ and therefore from eq.8,

$$\theta = \frac{1}{2} [\sigma^2 - \sigma_0^2] = \frac{1}{2} [\sigma^2 - npq] = \frac{1}{2} [\sigma^2 - np(1-p)] = \frac{1}{2} [\sigma^2 - m + m^2/n] \quad (9)$$

Using the experimental values of m and σ , we find that the numerical value of θ is 1.326.

3. A test of the alternative hypothesis

We can now make use of the foregoing analysis to test the acceptability of the *alternative* hypothesis, *viz.* that there is a perceptible difference between the choirs. For a total number B of Bernoulli trials, each of which includes n pieces of music, let us suppose that the number of trials in which there are r successful identifications of the top line is denoted by n_r . We then have that

$$\sum_{r=0}^n n_r = B \quad (10)$$

The relative frequency with which n_r pieces are correctly identified is then given by $f_r = n_r/B$, and if B is very large, f_r tends to the probability of obtaining r successes. The *alternative* hypothesis we shall test concerns the ‘goodness of fit’ between the experimentally observed frequencies and those deduced hypothetically. The most useful measure of agreement between the two sets of numbers is the χ^2 criterion devised by Karl Pearson in 1900 - perhaps in connection with Weldon’s dice data. To ensure that the underlying statistical theory is valid, it is often necessary to combine two or more frequencies so that the combined frequency exceeds about five. [Chatfield, C., *Statistics for Technology*, Chapman & Hall, London, 3rd Ed., (1983), p.150.] On this account, the number of ‘classes’ into which the data fall may well be less than n . In the present case, we shall take this number k to be 11 rather than the 21 correct identifications possible (including zero).

It is now necessary to deduce the relative frequencies for the hypothetical case of an inhomogeneous Bernoulli trial in which the mean and variance are taken to be those for the experimental data. Since $\pi_i < p < 1$, the sums of terms of higher degree than the second in π_i

will be very small compared with the first two terms of eq.5 so that, to a first approximation we have that

$$g(t) \approx \{1 + (t-1)p\}^n + \theta(t-1)^2 [1 + (t-1)p]^{n-2}$$

$$= (pt + q)^n + \theta(t-1)^2 [pt + q]^{n-2} \quad (11)$$

It should be noted that although this expression for the pgf is an approximation, to ensure that the sum of the probabilities is unity, it is necessary that $g(1)$ is unity, a condition that is clearly satisfied.

As mentioned earlier, the probability p_r of r successes is the coefficient of t^r in the binomial expansion of $g(t)$. We therefore apply the binomial theorem to the expressions in eq.11 to obtain

$$g(t) = \sum_{r=0}^n {}^n C_r p^{n-r} q^r t^r + \theta(t^2 - 2t + 1) \sum_{s=0}^{n-2} {}^{n-2} C_s p^{n-2-s} q^s t^s$$

$$= \sum_{r=0}^n {}^n C_r p^{n-r} q^r t^r + \theta \left\{ \sum_{s=0}^{n-2} {}^{n-2} C_s p^{n-2-s} q^s t^{s+2} - 2 \sum_{s=0}^{n-2} {}^{n-2} C_s p^{n-2-s} q^s t^{s+1} + \sum_{s=0}^{n-2} {}^{n-2} C_s p^{n-2-s} q^s t^s \right\}$$

If we now substitute in the second summation r for $(s+2)$ [so that its limits become 2 and n], and in the third summation we substitute r for $(s+1)$ [so that its limits become 1 and $(n-1)$], and in the fourth summation we substitute r for s , then we can express $g(t)$ as

$$= \sum_{r=0}^n {}^n C_r p^{n-r} q^r t^r + \theta \left\{ \sum_{r=2}^n {}^{n-2} C_{r-2} p^{n-r} q^{r-2} t^r - 2 \sum_{r=1}^{n-1} {}^{n-2} C_{r-1} p^{n-r-1} q^{r-1} t^r + \sum_{r=0}^{n-2} {}^{n-2} C_r p^{n-r-2} q^r t^r \right\}$$

$$= \sum_{r=0}^n t^r \left[{}^n C_r p^{n-r} q^r + \theta \{ a_{2,r} {}^{n-2} C_{r-2} p^{n-r} q^{r-2} - 2 a_{1,r} {}^{n-2} C_{r-1} p^{n-r-1} q^{r-1} + a_{0,r} {}^{n-2} C_r p^{n-r-2} q^r \} \right]$$

in which the coefficients $a_{i,r}$ all take the value unity except $a_{0,n-1}$, $a_{0,n}$, $a_{1,0}$, $a_{1,n}$, $a_{2,0}$ and $a_{2,1}$ which are all zero. The probability of r successes is then given by

$$p_r = \left[{}^n C_r p^{n-r} q^r + \theta \{ a_{2,r} {}^{n-2} C_{r-2} p^{n-r} q^{r-2} - 2 a_{1,r} {}^{n-2} C_{r-1} p^{n-r-1} q^{r-1} + a_{0,r} {}^{n-2} C_r p^{n-r-2} q^r \} \right]$$

$$= [B(n,r) + \theta \{ a_{2,r} B(n-2, r-2) - 2 a_{1,r} {}^{n-2} B(n-2, r-1) + a_{0,r} B(n-2, r) \}] \quad (12)$$

in which the functions $B(n,r)$ are the Bernoulli probabilities ${}^n C_r p^{n-r} q^r$.

Eq.12 is now in a convenient form for the computation of the hypothetical probabilities of the inhomogeneous Bernoulli trial since p is given by the experimental mean m (12.164) divided by n (the number of individual trials), $q = 1 - p$ and θ is given from eq.9 as $\frac{1}{2} [\sigma^2 - m + m^2/n]$ in which σ is the experimental standard deviation (2.724).

4. The χ^2 "goodness of fit" test

The required statistic is defined in terms of the observed and expected frequencies, o_k and e_k respectively, of given numbers of correct identifications, i.e.,

$$X^2 = \sum_{k=0}^{10} (o_k - e_k)^2 / e_k \quad (13)$$

in which the summation is over the k classes mentioned earlier. In the present case, there are two values of X^2 that we should consider. The first, X_1^2 , involves the expected frequencies for

the *homogeneous* Bernoulli trial using the expression $B^n C_r (m/n)^r (1 - m/n)^{n-r}$, and the second, X_2^2 , the expected frequencies Bp_r for the *inhomogeneous* trial using eq.12. These frequencies are shown in table 1. The frequencies for the eleven classes, obtained by combining the first eight and the last four, are shown in table 2. We then find that the appropriate X_1^2 and X_2^2 values are **35.904** and **10.064**, respectively, values that must now be compared with the theoretical probability distributions, the χ^2 distributions, for the appropriate number of degrees of freedom v . For X_1^2 , this number is $(k-2)$ since there are two restrictions on the frequencies, the first because the sums of the frequencies are both B (189), and the second because the means are both m (12.164). We now find from standard tables that $\chi^2_{0.005, 9} = 23.59$, considerably less than the 35.904. As expected, therefore, we reject the idea that the experimental results can be explained on the hypothesis that not only are the choirs distinguishable but that the probabilities of correct identifications are all equal (the homogeneous case). For the inhomogeneous case, the number of degrees of freedom is further reduced to eight because the frequencies are subject to a third restriction in that the standard deviations of the experimental and hypothetical distributions are both σ (2.724). We then find from tables that $\chi^2_{0.5, 8} < X_2^2 < \chi^2_{0.2, 8}$ (numerical values $7.34 < 10.064 < 11.03$). Therefore, the probability that the experimental and calculated frequencies are drawn from the same population (assumed to be roughly normally-distributed) lies between 20% and 50%. On this basis, it seems reasonable to accept the alternative hypothesis.

Table 1

	0	1	2
0	0	0	0
1	0	0	0
2	0	0.001	0.004
3	0	0.006	0.034
4	0	0.039	0.181
5	0	0.193	0.723
6	1	0.749	2.205
7	8	2.322	5.266
8	7	5.853	10.017
9	14	12.104	15.456
10	23	20.652	19.916
11	23	29.119	22.543
12	34	33.874	23.931
13	23	32.331	24.627
14	19	25.073	23.572
15	17	15.556	19.321
16	8	7.54	12.575
17	3	2.752	6.102
18	6	0.711	2.057
19	3	0.116	0.429
20	0	0.009	0.042

A =

Table 2

	0	1	2
0	9	3.31	8.413
1	7	5.853	10.017
2	14	12.104	15.456
3	23	20.652	19.916
4	23	29.119	22.543
5	34	33.874	23.931
6	23	32.331	24.627
7	19	25.073	23.572
8	17	15.556	19.321
9	8	7.54	12.575
10	12	3.588	8.63

B =

Table 1. Relative frequencies of obtaining the number of correct identifications shown in the left-hand column. The experimental frequencies are shown in column 0 (mean $m = 12.164$,

standard deviation $\sigma = 2.724$); those for the homogeneous Bernoulli trial are shown in column 1 (mean m , standard deviation 2.183); those for the inhomogeneous Bernoulli trial are shown in column 2 (mean m , standard deviation σ).

Table 2. Relative frequencies for the eleven classes used in the χ^2 test obtained by combining the first eight and the last four rows shown in table 1.

We can also calculate the probability $P_v(X^2 \geq \xi)$ that the statistic X^2 exceeds any particular value ξ for v degrees of freedom if the frequencies are drawn from the same population. For the homogeneous Bernoulli trial $P_9(X^2 \geq 35.904) = 4.12 \times 10^{-5}$, and for the inhomogeneous trial $P_8(X^2 \geq 10.064) = 0.261$.

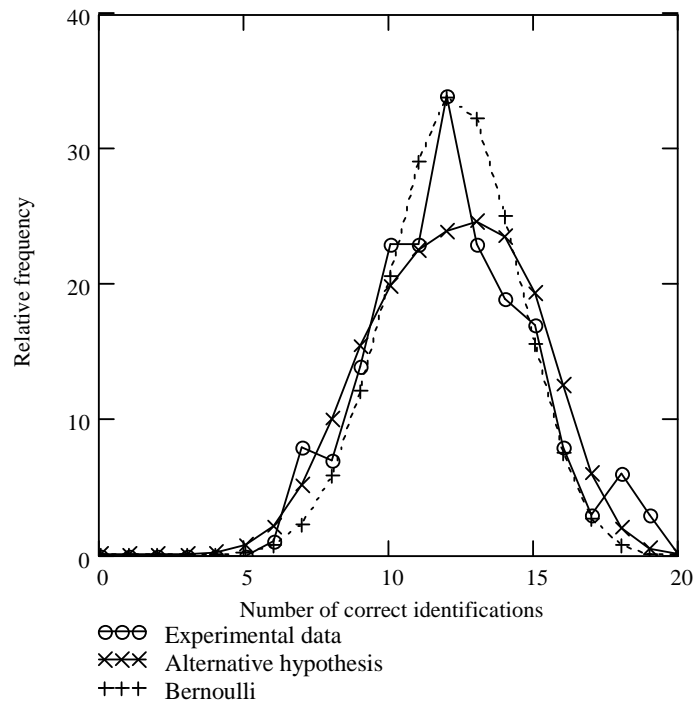


Figure 3. Relative frequencies out of a total of 189 trials of the number of correct identifications out of 20 individual trials. The experimental data and the data for the inhomogeneous Bernoulli trial (in which the probability varies amongst the individual trials) have the same means (12.164) and standard deviations (2.724). The latter constitutes the alternative hypothesis whereas the homogeneous Bernoulli trial, also shown, has the same mean but smaller standard deviation (2.183).

© A.E.Saunders 4th April 2009

saunders201@btinternet.com