Journal of Biometrics & Biostatistics

Fellman, J Biomet Biostat 2012, 3:5
DOI: 10.4172/2155-6180.1000146

Research Article

Open Access

Analysis of Sex-Linked Recessive Traits: Optimal Designs for Parameter Estimation and Model Tests

J. Fellman[1,2]*

[1]Folkhälsan Institute of Genetics, Department of Genetic Epidemiology, Helsinki, Finland
[2]Hanken School of Economics, Helsinki, Finland

Abstract

The estimation of the gene frequency of sex-linked recessive traits is reconsidered. The estimation formulae for mixed, male, and female samples are presented and compared. Optimal designs for efficient estimation are studied. Male samples are optimal for gene frequencies below 1/3 and female samples for frequencies above 1/3. Mixed samples are never optimal. The model testing problem is discussed. Mixed samples are necessary for model testing. We analyse the loss in efficiency when both estimation and testing must be performed. In general, our results indicate that mixed samples should contain an excess of males. The results obtained are applied to empirical data found in the literature [1,2].

Keywords: Maximum likelihood estimation; Model testing; Efficiency; Colour vision; xg Blood groups

Introduction

In the literature, abundant studies exist concerning probabilistic models in genetics. These have mainly investigated model building and the statistical estimation of gene frequencies. However, in to our opinion, experimental design problems have not been examined sufficiently. Against this background, this study is performed. We evaluate the estimation of the gene frequencies of sex-linked recessive traits and our basic assumption is that the trait is monogenic and recessive. Such a trait has markedly different phenotype frequencies in the male and female segments of the population. This is caused by the fact that if the trait is recessive and has a gene frequency *p* in the total population, then the frequency of affected individuals is *p* among males and p^2 among females. Consequently, direct comparisons of phenotype frequencies between males and females are of no value; e.g. the genes for colour-blindness and for blood group *Xg* are sex-linked, being located on the *X* chromosome.

We discuss and compare the maximum likelihood estimators of the gene frequency for mixed, male, and female samples. Among geneticists there is consensus that colour-blindness is not a monogenic trait. Kalmus (1985, p. 63) discussed whether the genes responsible for protan or deutan defects represent one common series of alleles on the X-chromosome or two separate series. He stated that the two-loci hypothesis seems better supported. The possibility to test the genetic model is crucial, and we give alternative methods for model testing. We analyse the loss in efficiency when both estimation and testing must be performed. The results obtained are applied to empirical data found in the literature [3,4].

Methods

Maximum likelihood estimation

The model: We consider a monogenic sex-linked recessive trait. We assume that we have a sample of size *N* consisting of *M* males and *F* females and that there are m_1 males with a recessive phenotype, m_2 males with a dominant phenotype, f_1 females with a recessive phenotype and f_2 females with a dominant phenotype. If the gene frequency of the recessive trait is *p* among both males and females [5,6], then the genetic model is given in Table 1.

	Males			Females		
	Number	Affected	Not affected	Number	Affected	Not affected
Observed	*M*	m_1	m_2	*F*	f_1	f_2
Theoretical	*M*	*Mp*	*M*(1-*p*)	*F*	Fp^2	$F(1-p^2)$

Table 1: Observed and expected number of subjects according to a monogenetic recessive sex-linked trait. Affected individuals have recessive and not affected individuals have dominant phenotypes.

A mixed sample: If we ignore a proportionality factor independent of *p*, we obtain from Table 1 the likelihood function $L(p) = p^{m_1}(1-p)^{m_2}\left(p^2\right)^{f_1}(1-p^2)^{f_2}$, with the restriction $0 < p < 1$. The function $L(p)$ can be written

$$L(p) = p^{m_1+2f_1}(1-p)^{m_2+f_2}(1+p)^{f_2}. \tag{1}$$

The log-likelihood function $l(p) = log(L(p))$ is

$$l(p) = (m_1+2f_1)log(p) + (m_2+f_2)log(1-p) + f_2 log(1+p) \quad 0 < p < 1 \tag{2}$$

If the log-likelihood function is written

$$l(p) = \left(m_1 log(p) + m_2 log(1-p)\right) + \left(2f_1 log(p) + f_2 log(1-p^2)\right) = l_M(p) + l_F(p), \tag{3}$$

the first parentheses ($l_m(p)$) contain the contribution of the male data and the second parentheses ($l_F(p)$) the contribution of the female data. When we maximize $l(p)$ in (2), we obtain

$$\frac{dl(p)}{dp} = \frac{(m_1+2f_1)}{p} - \frac{(m_2+f_2)}{(1-p)} + \frac{f_2}{(1+p)}. \tag{4}$$

The condition $\frac{dl(p)}{dp} = 0$, yields an algebraic equation of second degree

$$p^2 + \frac{m_2}{M+2F}p - \frac{m_1+2f_1}{M+2F} = 0. \tag{5}$$

***Corresponding author:** Folkhälsan Institute of Genetics, Department of Genetic Epidemiology, POB 211, FIN-00251 Helsinki, Finland, E-mail: fellman@hanken.fi

Received May 24, 2012; **Accepted** June 19, 2012; **Published** June 20, 2012

Citation: Fellman J (2012) Analysis of Sex-Linked Recessive Traits: Optimal Designs for Parameter Estimation and Model Tests. J Biomet Biostat 3:146. doi:10.4172/2155-6180.1000146

Copyright: © 2012 Fellman J. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

This equation has two roots, one negative outside the admissible region (0,1) and one positive. The positive root is

$$\hat{p} = -\frac{m_2}{2(M+2F)} + \sqrt{\frac{m_2^2}{4(M+2F)^2} + \frac{m_1+2f_1}{M+2F}}. \tag{6}$$

The upper limit of $\hat{p}$ is

$$\hat{p} < -\frac{m_2}{2(M+2F)} + \sqrt{\frac{m_2^2}{4(M+2F)^2}} + \sqrt{\frac{m_1+2f_1}{M+2F}} = \sqrt{\frac{m_1+2f_1}{M+2F}} < 1$$

Consequently, $0 < \hat{p} < 1$ and $\hat{p}$ belongs to the admissible interval (0,1). This estimation result was given by Haldane (1963). One obtains

$$\frac{d^2l(p)}{dp^2} = -\frac{(m_1+2f_1)}{p^2} - \frac{(m_2+f_2)}{(1-p)^2} - \frac{f_2}{(1+p)^2} \tag{7}$$

and $\frac{d^2l(p)}{dp^2} \le 0$. Consequently, the unique solution $\hat{p}$ maximizes *l*(*p*) (and *L*(*p*)). If we accept the model, we can estimate the estimator variance. We have

$$Var(\hat{p}) = \left[-E\left(\frac{d^2l(p)}{dp^2}\right)\right]^{-1}. \tag{8}$$

From (7) and (8), we get the information

$$\begin{aligned} I = -E\left(\frac{d^2l(p)}{dp^2}\right) &= \frac{Mp+2Fp^2}{p^2} + \frac{M(1-p)+F(1-p^2)}{(1-p)^2} \\ &+ \frac{F(1-p^2)}{(1+p)^2} = \frac{M}{p(1-p)} + \frac{4F}{1-p^2} \end{aligned}. \tag{9}$$

If we introduce $x = \frac{M}{N}$ and $1-x = \frac{F}{N}$, we obtain

$$\begin{aligned} -E\left(\frac{d^2l(p)}{dp^2}\right) = I(x,p) &= \frac{xN}{p(1-p)} + \frac{4(1-x)N}{1-p^2} \\ &= N\left(\frac{x}{p(1-p)} + \frac{4(1-x)}{1-p^2}\right) \end{aligned}. \tag{10}$$

We note that for high values of *x* (a majority of males) the information is high for low values of *p* and that for low values of *x* (a majority of females) the information is high for high values of *p*. Later, we will discuss this observation in more detail.

The inverse of *I*(*x*,*p*) yields the variance

$$\begin{aligned} Var(\hat{p}) = V(x,p) = I(x,p)^{-1} &= \left(\frac{xN}{p(1-p)} + \frac{4(1-x)N}{1-p^2}\right)^{-1} \\ &= \frac{1}{N}\left(\frac{x}{p(1-p)} + \frac{4(1-x)}{1-p^2}\right)^{-1} \end{aligned} \tag{11}$$

The estimator $\hat{p}$ is asymptotic normal and the variance $V(\hat{p})$ can be estimated by using $\hat{p}$ instead of *p* in (11). Haldane (1963, formula (5)) gives a slightly different estimate of $Var(\hat{p})$. His formula contains the observed frequencies and is, in to our opinion, not altogether satisfactory. In fact, he estimates *p* with $\hat{p}_M$ given below in (13) in the "male part" of the formula and with $\hat{p}_F$ given in (16) in the "female part" of the variance formula (c.f. formula (11)).

A male sample: If we consider a male sample and ignore the proportionality factor, which is independent of *p*, we obtain from (2) the log-likelihood function

$$l_M(p) = m_1 log(p) + m_2 log(1-p). \tag{12}$$

When we maximize $l_M(p)$, we get the "male" estimator

$$\hat{p}_M = \frac{m_1}{M}, \tag{13}$$

with the information $I_M = \frac{M}{p(1-p)}$ and the well-known variance

$$Var(\hat{p}_M) = \frac{p(1-p)}{M}. \tag{14}$$

The estimator $\hat{p}_M$ is asymptotic normal and the variance $V(\hat{p}_M)$ in (14) can be estimated by using $\hat{p}_M$ instead of *p*.

A female sample: If we consider only the female part of the sample and ignore the proportionality factor, which is independent of p, we obtain from (3) the log-likelihood function [7,8]

$$l_F(p) = 2f_1 log(p) + f_2 log(1-p^2). \tag{15}$$

If we maximize the log-likelihood function, we obtain the "female" estimator

$$\hat{p}_F = \sqrt{\frac{f_1}{F}}, \tag{16}$$

with the information $I_F = \frac{4F}{1-p^2}$ and the variance

$$Var(\hat{p}_F) = \frac{1-p^2}{4F}. \tag{17}$$

The estimator $\hat{p}_F$ is consistent, efficient and asymptotic normal and the variance $V(\hat{p}_F)$ in (17) can be estimated by using $\hat{p}_F$ instead of *p*. According to Huether and Murphy (1980), it is not clear how rapidly these asymptotic properties are approached with increasing sample size. The log likelihood equation (15) yields an unbiased estimate $\breve{p}_F^2 = \frac{f_1}{F}$ of p^2, but in (16) is biased with a negative bias. Haldane (1956) proposed an improved estimate [9,10]

$$\overset{\frown}{p}_F = \sqrt{\frac{4f_1+1}{4F+1}}. \tag{18}$$

In order to improve the ML estimates, Huether and Murphy proposed a jackknife procedure. Their estimate is, using our notations.

$$\overleftrightarrow{p}_F = F\sqrt{\frac{f_1}{F}} - (F-1)\left(\frac{f_1}{F}\sqrt{\frac{f_1-1}{F-1}} + \frac{F-f_1}{F}\sqrt{\frac{f_1}{F-1}}\right). \tag{19}$$

How these improvements influence our gene estimates will be discussed in the Discussion section. Eq. (9) indicates that the information obtained for the whole data set is $I(x,p) = I_M + I_F$. This is a consequence of the male and female data sets being independent.

Model testing

A mixed sample: In the mixed data set, there are two degrees of freedom because the row sums for males and females are fixed. After the estimation of p, one degree of freedom remains. According to Table 1, the model can be tested by the quantity [11]

$$\chi^2 = \frac{(m_1-\hat{m}_1)^2}{\hat{m}_1} + \frac{(m_2-\hat{m}_2)^2}{\hat{m}_2} + \frac{(f_1-\hat{f}_1)^2}{\hat{f}_1} + \frac{(f_2-\hat{f}_2)^2}{\hat{f}_2}, \tag{20}$$

where $\hat{m}_1 = M\hat{p},\ \hat{m}_2 = (1-\hat{p})M,\ \hat{f}_1 = F\hat{p}^2$ and $\hat{f}_2 = F\left(1-\hat{p}^2\right)$.

Under the null hypothesis that the model holds, this quantity is approximately χ^2 distributed with one degree of freedom.

The model can also be tested by the Likelihood Ratio Test (LRT). Consider

$$\Lambda = \frac{\sup\limits_{p_M=p_F} L(p_M,p_F)}{\sup\limits_{p_M,p_F} L(p_M,p_F)},$$

Where

$$L(p_M,p_F) = p_M^{m_1}(1-p_M)^{m_2}\,p_F^{2f_1}(1-p_F^2)^{f_2}.$$

The maximizations give

$$\Lambda = \frac{\hat{p}^{m_1+2f_1}(1-\hat{p})^{m_2+f_2}(1+\hat{p})^{f_2}}{\hat{p}_M^{m_1}(1-\hat{p}_M)^{m_2}\,\hat{p}_F^{2f_1}(1-\hat{p}_F)^{f_2}(1+\hat{p}_F)^{f_2}}, \tag{21}$$

Where $\hat{p}$, $\hat{p}_M$, and $\hat{p}_F$ are given in (6), (13), and (16), respectively. Under the null hypothesis, $-2\log\Lambda$ is approximately χ^2 distributed with one degree of freedom. In situations not far from the null hypothesis, the χ^2 tests based on (20) and (21) give similar results. In the applications, the formula (20) is used.

Separate male and female samples: If we estimate p separately for the male and female series, there is no degree of freedom left in either series. Consequently, if we test the hypothesis $\hat{p}_M = \hat{p}_F$, we must consider the difference $\hat{p}_M - \hat{p}_F$ with the variance

$$Var(\hat{p}_M) + Var(\hat{p}_F) = \frac{p(1-p)}{M} + \frac{1-p^2}{4F}. \tag{22}$$

Under the null hypothesis, $z = \frac{\hat{p}_M - \hat{p}_F}{\sqrt{Var(\hat{p}_M) + Var(\hat{p}_F)}}$ is standard normal.

If we accept the null hypothesis $E(\hat{p}_M) = E(\hat{p}_F) = p$, then we can obtain a weighted estimate of the common gene frequency p. To minimize the variance of the weighted estimate, the weights should be the inverses of the variances in (14) and (17). The weighted estimate is

$$\tilde{p} = \frac{\left(\frac{4F}{1-p^2}\right)\hat{p}_F + \left(\frac{M}{p(1-p)}\right)\hat{p}_M}{\frac{4F}{1-p^2} + \frac{M}{p(1-p)}}, \tag{23}$$

and its theoretical variance is $V(\tilde{p}) = \left(\frac{M}{p(1-p)} + \frac{4F}{1-p^2}\right)^{-1}$, which is identical to (11). The estimator $\hat{p}$ maximizes $L(p)$ and $L(\tilde{p}) \le L(\hat{p})$, but the weighted estimator $\tilde{p}$ and the Haldane estimator $\hat{p}$ have asymptotically the same efficiency. Consequently, both estimators are best asymptotic normal (BAN).

Design of experiments

In connection with another type of genetic problem, Brown (1975) considers efficient experimental designs for the estimation of genetic parameters. We start from the same basic idea, but use different methods. In his book on colour-blindness, Kalmus (1965, p. 85) states, without further comments, that the population frequency for rare sex-linked recessive traits must be based on male samples. Now we study this problem in more detail. We apply experimental design theory using the inference results in the preceding sections [12].

Designs for parameter estimation: Let us assume that we intend to investigate N (fixed) individuals and that the gene frequency is p. Now our problem is in what proportion $M : F$ shall we include males and females in our sample in order to minimize the variance given in (11) or, alternatively, to maximize the information measure (9). We study the information $I(x, p)$ and the variance $V(x, p)$ as functions of p and x. From (9) we get

$$I(x,p) = \frac{xN}{p(1-p)} + \frac{4(1-x)N}{1-p^2} = \frac{N(x(1-3p)+4p)}{p(1-p^2)}. \tag{24}$$

The function (24) is a linear function of x. For $p < \frac{1}{3}, I(x,p)$ is an increasing function of x and the maximum is obtained for x = 1, i.e. for a male sample. For $p > \frac{1}{3}, I(x,p)$ is a decreasing function of x and the maximum is obtained for x = 0, i.e. for a female sample. For $p = \frac{1}{3}, I(x,p)$ is constant and all samples are equally good. Our optimal experimental design for parameter estimation is hence

(i) Use a male sample if $p < \frac{1}{3}$

(ii) Use an arbitrary sample if $p = \frac{1}{3}$

(iii) Use a female sample if $p > \frac{1}{3}$.

We observe that the optimal design of the experiment depends on the true parameter value. This is common in non-linear situations, but in this case the dependence is very simple. In different populations, the frequency of colour blindness is about 0.08 so the rule (i) is in good agreement with Kalmu´s (1985) statement.

In practice, the problem is not so simple. Often when we start an investigation, we do not know the gene frequency. If we have prior information (from earlier studies) that the gene frequency is far in a known direction from one-third we can with confidence use a male or a female sample. If, however, we have no prior information or if the gene frequency is known to be in the neighbourhood of $\frac{1}{3}$, then it is difficult to decide whether to use a male or female sample. We can see in Table 2 that for the Xg blood group p is close to $\frac{1}{3}$, and this is a good example of this problem.

Let us now analyse the efficiency of a mixed sample in more detail. Assume that the true gene frequency is p. Now, we have to compare $V(x,p) = \frac{p(1-p^2)}{N(x(1-3p)+4p)}$ with $V_M(p) = \frac{p(1-p)}{N}$ if $p < \frac{1}{3}$ and with $V(x,p) = \frac{(1-p^2)}{4N}$ if $p > \frac{1}{3}$, and we obtain the relative efficiencies for the mixed sample

$$\begin{aligned} E(x,p) &= \frac{x(1-3p)+4p}{(1+p)} && \text{for } p < \frac{1}{3} \\ &= 1 && \text{for } p = \frac{1}{3} \\ &= \frac{x(1-3p)+4p}{4p} && \text{for } p > \frac{1}{3}. \end{aligned} \tag{25}$$

	N	Recessive	Dominant	$\hat{p}/SD$ a)	χ^2	$\hat{p}_M/\hat{p}_F$ b)	SD	$\tilde{p}/SD$ c)	Reference
Males	154	59	95	0.356021	0.88	0.383117	0.039175	0.355486	Mann et al., 1962
Females	188	21	167	0.025542		0.334219	0.034369	0.025836	
Males	1751	620	1131	0.341226	2.62	0.354083	0.019206	0.334715	Noades et al., 1966
Females	1667	179	1488	0.008075		0.327687	0.011570	0.009911	
Males	3513	1209	2304	0.340577	0.41	0.344150	0.013664	0.338743	Sanger et al., 1971
Females	3271	371	2900	0.005731		0.336780	0.008232	0.007051	

a) Maximum likelihood estimate on the upper line and SD on the lower line

b) Male estimate on the upper line and female estimate on the lower line

c) Weighted estimate of $\hat{p}_M$ and $\hat{p}_F$ on the upper line and SD on the lower line

Table 2: X_g in different studies.

If we must test the model, it is necessary to include in the sample both males and females. If this is done, there is a loss of efficiency relative to the best (but unknown) design. In general, if we compare a male sample, a female sample, and a mixed sample of the same size, then the efficiency of the mixed sample is always between the efficiencies of the single-sex samples.

In Figure 1, we see how the efficiencies depend on the gene frequency for the single-sex samples (x = 0 and x = 1) and for some mixed samples (x = 0.3333, 0.5155, and 0.6667). The choice of x = 0.5155 and x = 0.6667 will be explained later. We observe that for small values of p the efficiency strongly depends on the true value of p. For $p<\frac{1}{3}$, the male sample is most efficient. For $p>\frac{1}{3}$, the female sample is most efficient but the efficiency of a female sample is not as good as the efficiency of the male sample for $p<\frac{1}{3}$. Therefore, Figure 1 supports the conclusion that, independently of the true value of p, if we want to play safe a mixed sample should contain an excess of males.

This result can also be obtained in the following way. We consider the efficiency $E(x,p)$ for a mixed sample as a function of p for a given x. For $p<\frac{1}{3}$, we have $E(x,p)=\frac{x(1-3p)+4p}{(1+p)}$ and $\frac{\partial E}{\partial p}=\frac{4-4x}{(1+p)^2}\geq 0$, with equality for $x = 1$, i.e. the sample contains only male subjects. Hence, $E(x,p)$ is an increasing function of p and $E(x,p)\geq E(x,0)=x$ for $p<\frac{1}{3}$.

Similarly, we obtain for $p>\frac{1}{3}$ $E(x,p)=\frac{x(1-3p)+4p}{4p}$ and $\frac{\partial E}{\partial p}=-\frac{x}{4p^2}\leq 0$. Now, $E(x,p)$ is a decreasing function of p and $E(x,p)\geq E(x,1)=1-\frac{x}{2}$ for $p>\frac{1}{3}$. From these results, it follows that $E(x,p)\geq E_m=\min\left(x,1-\frac{x}{2}\right)$. Hence, $\max_x \min_p E(x,p)=\frac{2}{3}$, and this value is obtained for $x=\frac{2}{3}$ and $p=0$ or 1.

Speaking in terms of game theory, the strategy of nature is the choice of p and our strategy is the choice of x, and $E(x,p)$ is the pay-off of the game. The $\max_x \min_p E(x,p)$ solution indicates that we are playing safe. We expect the worst, i.e. that nature has chosen one extreme p value, and consequently, we prepare for it and choose the strategy that maximizes our gain (the efficiency). From this point of view, we should use a sample with $\frac{2}{3}$ males and $\frac{1}{3}$ females. This mixed sample guarantees at least the efficiency $\frac{2}{3}$ (cf. Figure 1).

Designs for model testing: A sample consisting of both males and females is necessary if we have doubts about the model. The doubts may concern the simple recessive inheritance (cf. colour blindness), absence of selection (the same gene frequency in males and females), exact typing independent of the sex, or the non-existence of border cases that are difficult to type. If we have a mixed sample, we can then test the model as we have noted above. This is not possible with a male-, or female-only sample. This problem is a good example of the common situation that an experimental strategy, which is optimal for parameter estimation, is too restricted to be of any value for model testing.

If we want to test the model and to use $W(x,p)=Var(\hat{p}_M-\hat{p}_F)$ given in (22) most efficiently under the null hypothesis, then we have to consider the variance

$$W(x,p)=\frac{p(1-p)}{xN}+\frac{1-p^2}{4(1-x)N} \qquad (26)$$

and to pursue $\min_x \max_p W(x,p)$. This solution indicates that we are again playing safe. We expect the worst situation, i.e. that nature has chosen a p value that maximizes the variance, and consequently, we prepare for it and choose the strategy (x) that minimizes our loss (the variance). In other words, we want to answer the question: Which sample mixture

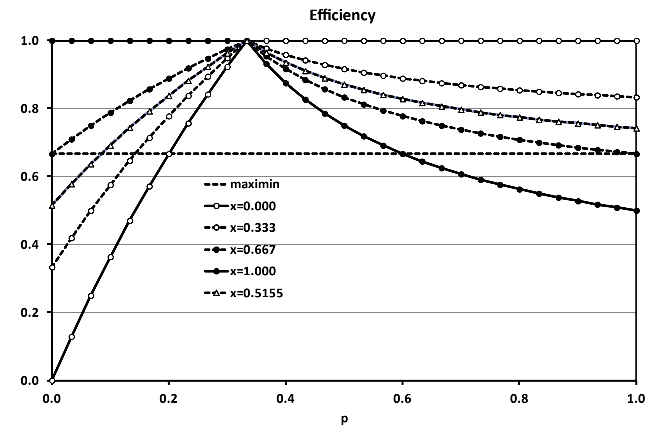

Figure 1: Efficiency as a function of the gene frequency p for different sample compositions (x = *0, 0.3333, 0.5155, 0.6667, 1*), where $x=\frac{M}{M+F}$. Maximin corresponds to $\max_x \min_p E(x,p)=\frac{2}{3}$. For details, see the text.

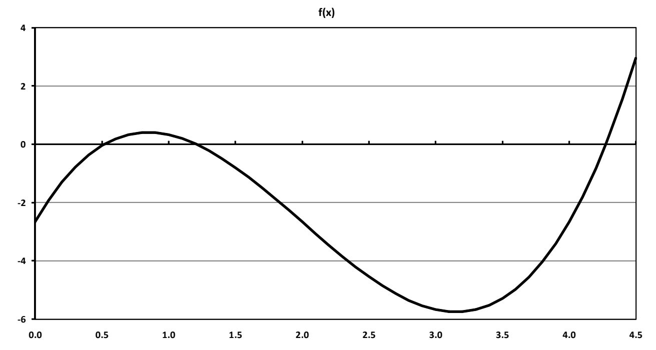

Figure 2: Graphic representation of the function $f(x)=x^3-6x^2+8x-\frac{8}{3}$. The root $\hat{x}=0.5155$ is easily recognized.

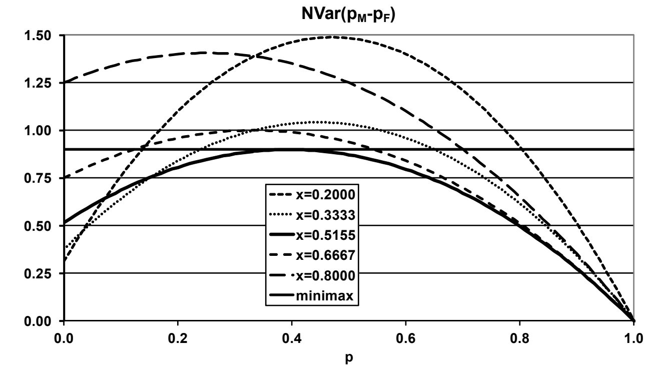

Figure 3: Variance of the difference $\hat{p}_M-\hat{p}_F$ as a function of the gene frequency p for different sample compositions $\left(x=0.2000, 0.3333, 0.5155, 0.6667, 0.8000\right)$. Minimax corresponds to $\min_x \max_p W(p,x)=\frac{0.8990667}{N}$. For details, see the text.

x: (1-*x*) minimizes the $\max_p W(x,p)$? For a given *x*, we obtain

$\frac{\partial W(x,p)}{\partial p} = \frac{1-2p}{xN} + \frac{-2p}{4(1-x)N}$ and $\frac{\partial W(x,p)}{\partial p} = 0$ gives $p(x) = \frac{2(1-x)}{4-3x}$.

The corresponding *W* value is a maximum for $\frac{\partial^2 W(x,p)}{\partial p^2} < 0$. This maximum $W_{max}(x)$, which depends on *x*, is

$$W_{max}(x) = W\left(p(x),x\right) = \frac{p(x)\left(1-p(x)\right)}{xN} + \frac{1-p(x)^2}{4(1-x)N}.$$

Now, we minimize $W_{max}(x)$ by using the derivative $\frac{dW_{max}(x)}{dx} = \frac{\partial W}{\partial p}\frac{dp}{dx} + \frac{\partial W}{\partial x}$.

If we use the condition that $\left(\frac{\partial W}{\partial p}\right)_{p(x)} = 0$, the derivative reduces to

$$\frac{dW_{max}(x)}{dx} = \frac{\partial W}{\partial x} = -\frac{p(1-p)}{x^2 N} + \frac{1-p^2}{4(1-x)^2 N}.$$

Now, we solve the equation $\frac{dW_{max}(x)}{dx} = 0$ under the restriction $p = \frac{2(1-x)}{4-3x}$. The equation simplifies to

$$x^3 - 6x^2 + 8x - \frac{8}{3} = 0. \qquad (27)$$

This equation of third degree satisfies the conditions $f(0) = -\frac{8}{3} < 0$ and $f(1) = \frac{1}{3}$. Consequently, the equation has one root or three roots within the interval (0,1). The case three roots within this interval are impossible because the product of the roots has to be $\frac{8}{3} > 1$. Thus, there is only one root within the interval (0,1). In Figure 2, we present the function $f(x) = x^3 - 6x^2 + 8x - \frac{8}{3}$ in order to locate the roots. An iterative calculation yields the numerical root $\hat{x} = 0.5155$, and the corresponding *p* value is $\hat{p} = 0.3949$. Finally, we obtain

$$\min_x \max_p W(p,x) = \frac{0.8991}{N}. \qquad (28)$$

The solution $\hat{x} = 0.5155$ is our best testing strategy in order to meet nature's worst alternative $\hat{p} = 0.3949$. This minimax solution of the testing problem does not coincide with the maximin solution of the efficiency problem. Figure 3 shows how $NW(p,x)$ depends on *p* for different values of *x*. The minimax property of $\hat{x} = 0.5155$ is easily seen.

Applications

We apply our theoretical results to empirical data. We consider both colour vision and X_g blood group data. In Table 2, we present the results of the analyses of blood group data, and in Table 3 the results of the colour vision data. The results obtained by the mixed sample and obtained by combined estimates of male and female samples are fairly similar.

Discussion

The reduction of the biases in the female estimates in the Tables 2 and 3 is presented in Table 4. The comparison between the maximum likelihood estimates and the improved estimates indicates that the MLE has a negative bias, but the sample sizes result in ignorable errors. The improvements proposed by by Haldane (1956) and Huether & Murphy (1980) yield almost identical estimates.

If our minimax design $\tilde{x} = 0.5155$ is used for an estimation problem, then the minimum efficiency is 0.5155, which is obtained for *p* = 0. If we compare this value with the maximin solution *x* = 0.6667 for the estimation problem, we observe how much we have to "pay" for the hypothesis testing. On the other hand, if our primary goal is estimation and we choose the design $x = \frac{2}{3}$, then the corresponding maximal variance is $\max_p V(p,\frac{2}{3}) = \frac{1}{N}$, which is obtained for $p = 0.3333$. This can be compared with the earlier obtained $\min_x \max_p W(p,x) = \frac{0.8991}{N}$. Hence, if our target is parameter estimation, then the efficiency of the model test is reduced in the proportion $0.8991:1$.

The common opinion of today is that colour blindness is not a one-locus trait. Waaler´s, Smith´s, and Koliopoulo´s data show statistically significant differences from the one-locus model. The common finding in this study is that the estimate $\hat{p}_M$ is less than $\hat{p}_F$, and this result supports the two-loci hypothesis. However, the other colour vision data, especially the female data, are very limited. NZHTA Report 7 (1998) presents colour vision data collected from different sources and the value of this study is this collection. In addition, that study

	N	Recessive	Dominant	$\hat{p}\,/\,SD$ a)	χ^2	$\hat{p}_M\,/\,\hat{p}_F$ b)	SD	$\hat{p}\,/\,SD$ c)	Reference
Males	9049	725	8324	0.077226	**4.76**	0.080119	0.002854	0.076979	Waaler, 1927
Females	9072	40	9032	0.00247		0.066402	0.005238	0.002506	
Males	6863	532	6331	0.074505	**4.89**	0.077517	0.003228	0.074141	Schmidt, 1936
Females	5604	20	5584	0.002862		0.059740	0.006667	0.002905	
Males	21231	1687	19544	0.078034	**5.62**	0.079459	0.001856	0.077898	Koliopoulos et al., 1976
Females	8754	37	8717	0.001740		0.065013	0.005333	0.001753	

[a)]Maximum likelihood estimate on the upper line and SD on the lower line
[b)]Male estimate on the upper line and female estimate on the lower line
[c)]Weighted estimate of $\hat{p}_M$ and $\hat{p}_F$ on the upper line and SD on the lower line

Table 3: Colour blindness in different studies.

	ML estimate	Haldane, 1956	Huether & Murphy, 1980
X_g			
Mann et al., 1962	0.33422	0.33598	0.33603
Noades et al., 1966	0.32769	0.32789	0.32789
Sanger et al., 1971	0.33678	0.33688	0.33688
Colour vision			
Whaaler, 1927	0.06640	0.06661	0.06661
Schmidt, 1936	0.05974	0.06011	0.06012
Koliopoulos et al., 1976	0.06501	0.06523	0.06523

Table 4: Comparison between the maximum likelihood estimates and the improved estimates proposed by Haldane (1956) and Huether & Murphy (1980).

presents tests of the sex differences in the distribution between subjects with colour deficiency and normal sight. The tests indicate strong sex differences, but the tests have ignored the effect of the sex-linked property of colour blindness, and consequently, these results are of minor interest.

Acknowledgements

The author is grateful to an anonymous referee for very constructive suggestions concerning the biasedness among the “female estimates” and has tried to consider these remarks in the text. This work was supported in part by a grant from the Magnus Ehrnrooth Foundation.

$\min_x \max_p W(p,x) = \frac{0.8990667}{N}$. For details, see the text.

References

1. Brown AH (1975) Efficient experimental designs for the estimation of genetic parameters in plant populations. Biometrics 31: 145-160.
2. Haldane JB (1956) Almost unbiased estimates of functions of frequencies. Sankhyā 17: 201-208.
3. Haldane JB (1963) Tests for sex-linked inheritance of population samples. Ann Hum Genet 27: 107-111.
4. Huether CA, Murphy EA (1980) Reduction of bias in estimating the frequency of recessive genes. Am J Hum Genet 32: 212-222.
5. Kalmus H (1965) Diagnosis and genetics of defective colour vision. Pergamon Press.
6. Koliopoulos J, Iordanides P, Palmeris G, Chimonidou E (1976) Data concerning colour vision deficiencies amongst 29,985 young Greeks. Mod Probl Ophthalmol 17: 161-164.
7. Mann JD, Cahan A, Gelb AG, Fisher N, Hamper J (1962) A sex-linked blood group. Lancet 1: 8-10.
8. Noades J, Gavin J, Tippett P, Sanger R, Race RR (1966) The X-linked blood group system Xg tests on British, Northern American, and northern European unrelated people and families. J Med Genet 3: 162-168.
9. NZHTA Report 7 (1998) Colour Vision Screening. A critical appraisal of the literature. New Zealand Health Technology Assessment.
10. Sanger R, Tippett P, Gavin J (1971) The X-linked blood group system Xg. Tests on unrelated people and families of Northern European ancestry. J Med Genet 8: 427-433.
11. Schmidt I (1936) Result of a mass examination of color sense with anomaloscope. Z Bahnärtztex 44-53.
12. Waaler GH (1927) Color blindness on the Erblichkeitsverhältnisse of different types of congenital. Ztschr F Induct Lineage-u Vererbungsl 45: 279-333.