

EVIDENCE FOR THE UNDERESTIMATION OF THE PROPORTION OF NORTH
AMERICAN SALMON IN THE WEST GREENLAND FEEDING AGGREGATE

Daphne J. Fairbairn

Summary

Payne (1973) developed a method of estimating the proportion of North American salmon in the West Greenland aggregate from transferrin phenotypes. This method is found to be sensitive to violations of the assumption of genetic equilibrium of transferrin phenotypes, with positive values of F (heterozygote deficiency, population subdivision) leading to a negative bias in the estimate of number of North American salmon. It is shown that F will be positive for West Greenland samples, and that Payne's estimator can be expected to have a negative genetical bias of 20%-35% for these samples. Small sample sizes and low frequencies of the rare transferrin allele (Tf4) also lead to a statistical bias in Payne's estimator. Although both the direction and magnitude of this bias depend upon sample size and allele frequency, for the West Greenland samples the bias has probably been small and positive. It is suggested that the 10%-20% discrepancy between estimates of the proportion of North American salmon at West Greenland obtained from scale analysis and from Payne's estimator is mainly due to an overall negative bias in Payne's estimator, resulting from a large negative genetical bias and smaller positive statistical bias. Because this bias can neither be accurately assessed nor eliminated, it is suggested that Payne's estimator no longer be applied to the West Greenland populations.

Introduction

With the expansion of the West Greenland salmon fishery in the mid 1960s came the need for an accurate assessment of the effects of this fishery on North American stocks of salmon. A first step in this assessment has been to determine the proportion of the West Greenland salmon that originate in North American rivers. One approach to this problem is based on observed differences between North American and European salmon in serum transferrin phenotypes (Payne 1973). Five transferrin phenotypes can be distinguished by starch gel electrophoresis and four of these are characteristic of the continent of origin of the salmon (Table 1). Because approximately 80% of salmon at West Greenland have the phenotype Tf1 (Payne, Lear and Jensen 1976; Lear and Payne 1975; Payne 1973), the total number of North American salmon cannot be determined directly. However, Payne (1973) developed an estimator of the number of North American salmon using the observed transferrin phenotypes and the assumption that these phenotypes are in genetic equilibrium (distributed accord-

ing to the Hardy-Weinberg equation) in the American component of the West Greenland stock complex. This estimator was incorporated into the assessments of the contribution of American salmon to the West Greenland fishery from 1970 to 1976 (Lear and Sandeman 1979; Payne and Cross 1977; Payne, Lear and Jensen 1976; Lear and Payne 1975; Payne 1973). It is the thesis of this paper that Payne's estimator has a persistent negative bias, and therefore, that the tendency in the past will have been to underestimate the proportion of North American salmon in the West Greenland aggregate.

Theoretical Considerations

The negative bias in Payne's technique derives from the basic assumption that the phenotypes observed among the American salmon at West Greenland will be distributed according to the Hardy-Weinberg equilibrium. Theoretical considerations give us good reason to doubt this assumption. The North American component of the West Greenland aggregate is known to be composed of a mixture of adult salmon from populations with radically differing gene frequencies (Payne 1974, Møller 1970). Samples from such an aggregate cannot be expected to conform to the Hardy-Weinberg distribution of phenotypes. Rather, they should show the familiar Wahlund effect, with an apparent deficiency of heterozygotes (Wallace 1968, Workman 1969, Cavalli-Sforza and Bodmer 1971). The expected distributions of genotypes, and thus of phenotypes, under the assumption of Hardy-Weinberg and under the assumption of population mixing or subdivision, are shown in Table 2. F is the 'fixation index' (Workman 1969) and is algebraically equivalent to the coefficient of inbreeding (Wallace 1968, Cavalli-Sforza and Bodmer 1971, Workman 1969). It describes the deviation of genotype frequencies from Hardy-Weinberg proportions. F will be positive for subdivided or mixed populations such as those in the West Greenland aggregate. If $F = 0$, the genotypic distribution follows Hardy-Weinberg expectations.

The mixing of populations with different gene frequencies is not the only factor that would lead us to expect a positive F for samples taken from the West Greenland aggregate. Payne (1974) noted that samples of adult salmon taken from rivers along the North American coast, from Maine to Labrador, tended to show a deficiency of heterozygotes. An analysis of the data from that paper yields a mean F of 0.074 (S.D. 0.037) for the anadromous salmon sampled by Payne in 1972, and 0.048 (S.D. 0.096) for those sampled by Møller in 1969. Thus, although the variance between rivers is very high, the contribution to the total F of these individual population distributions can be expected to accent rather than counteract the heterozygote deficiency due to the Wahlund effect.

Having established that we should expect $F > 0$ for samples from the West Greenland aggregate, let us look at the effects of a positive F on the estimations made following Payne's formulae. In order to estimate the number of North American salmon in his samples, Payne initially estimated

the frequency of the Tf4 allele (q) in the North American component according to the following formula:

$$\hat{q} = \frac{2}{\left(\frac{Tf1/4}{Tf4}\right) + 2} \quad (1)$$

Assuming Hardy-Weinberg equilibrium (see Table 2):

$$\hat{q} = \frac{2}{\frac{2pq}{q^2} + 2} = 2q^2 / (2pq + 2q^2) = q / (p+q) = q$$

However, if we assume $F \neq 0$:

$$\begin{aligned} \frac{Tf1/4}{Tf4} &= \frac{2pq(1-F)}{(q^2 + pqF)} \quad (\text{from Table 2}) \\ &= \frac{2(1-q)(1-F)}{q + (1-q)F} = \frac{2(1-q)}{\frac{F}{(1-F)} + q} \end{aligned}$$

Substituting into equation (1):

$$\hat{q} = \frac{2}{\frac{2(1-q)}{\frac{F}{(1-F)} + q} + 2} = q + F(1-q) \quad (2)$$

Thus, when F is positive, $\hat{q} > q$. Fig. 1 illustrates the relationship between \hat{q} and the true q, for various values of F and q. It can be seen that Payne's formula will greatly overestimate q if F is high or the true value of q is low.

Payne used \hat{q} to estimate the number of North American salmon in his samples from West Greenland in the following manner:

Assuming Hardy-Weinberg equilibrium:

$$N = (Tf1/4) / 2pq = (Tf1/4) / 2q(1-q),$$

where N is the number of North American salmon in the sample.

Now, if $F \neq 0$, the true relationship will be:

$$N = \frac{(Tf1/4)}{2q(1-q)(1-F)} \quad (\text{From Table 2})$$

The N estimated by Payne's method will be:

$$\hat{N} = \frac{(Tf1/4)}{2\hat{q}(1-\hat{q})} \quad (3)$$

To determine the relationship between \hat{N} and N , let us look at \hat{N}/N :

$$\hat{N}/N = \frac{q(1-q)(1-F)}{\hat{q}(1-\hat{q})}$$

This relationship is illustrated in Fig. 2, for different values of q and F . It can be seen that if F is positive \hat{N} will be smaller than N .

Estimates Using North American Samples

The tendency of Payne's formulae to overestimate q and underestimate N is dramatically illustrated by the examples in Table 3. Since these populations are exclusively North American, the true N and q are known; \hat{q} and \hat{N} were calculated using formulae (1) and (3). Only samples with N and q large enough to preclude statistical bias were used (see appendix). In spite of the relatively high frequencies of the Tf4 allele in the statistically acceptable samples from individual rivers (samples 1-7), the estimates of \hat{q} and \hat{N} are clearly dependent upon F , with the best estimates obtained when F approaches 0.00. The most representative values of q and F are seen in examples 8, 9 and 10. The fixation indexes for these combined samples include a large component due to the Wahlund effect, in addition to the contribution of the individual F 's. In this manner these samples more closely mimic the West Greenland population than do the individual river samples. Payne's formulae overestimate q by between 36% and 50%, and underestimate N by between 26% and 33% for these combined samples.

Although these North American samples cannot be considered truly representative of the salmon found in the West Greenland aggregate, they can yield crude estimates of the degree of bias to be expected in samples from the latter. The mean qTf_4 for all anadromous salmon populations sampled by Payne and Møller (Payne 1974, Møller 1970) is 0.27 (S.D. 0.17). This is the same as the qTf_4 for all of these samples combined (Table 3). Thus, 0.27 is our best estimate of qTf_4 at West Greenland. From equations (1) and (3), assuming $0.10 < F < 0.20$, as indicated by the combined samples in Table 3, we calculate $1.27 < (\hat{q}/q) < 1.54$ and $0.65 < (\hat{N}/N) < 0.79$. Thus we can expect Payne's technique to underestimate the number of North American salmon in the West Greenland samples by 20%-35%.

Comparison with Scale Analysis

The foregoing analysis has been based on theoretical considerations and on data gathered from North American populations at their source rivers. However, the data from the West Greenland aggregate also suggest that Payne's method has underestimated the proportion of North American fish. Salmon from the West Greenland aggregate can be assigned to their continent of origin by means of a discriminant function based on scale

characters (Lear and Sandeman 1979; Payne, Lear and Jensen 1976; Lear and Payne 1975). This discriminant function is able to discriminate European from North American salmon with an accuracy of 90%-100% when applied to the salmon at West Greenland (Lear and Sandeman 1979). Further, the scale analysis is completely independent of the electrophoretic analysis and is less likely than the electrophoretic analysis to be influenced by small samples (Lear and Payne 1975, and see appendix).

Table 4 shows the proportion of North American salmon estimated by Payne's method and by scale analysis. The 1971 electrophoretic data are not accurate because a large proportion of the European salmon were accidentally excluded from the electrophoretic sample (Lear and Sandeman, 1979). No data are available for 1973. In all five of the years for which reliable data are available, the estimates from scale analysis are greater than those from Payne's analysis, with the mean ratio of electrophoretic to scale estimates being 0.76. During the latter three years (1974-76) the same samples were used for both analyses. Over these years the mean ratio of electrophoretic to scale estimates is 0.86.

This discrepancy is not as large as that predicted by the analysis of anadromous river populations. This may indicate that at West Greenland the true F is lower or the true q is higher than in our combined estimates. Selection favoring heterozygotes, or differences between males and females in gene frequencies could result in a lower F than calculated. The frequency of the $Tf4$ allele may be higher than predicted, if Gulf of St. Lawrence stocks, not discussed by Payne (1974), have a high $qTf4$. However, for the sample sizes used from the West Greenland analysis, and for $qTf4$ less than 0.3, Payne's estimator of the number of North American salmon will have a positive statistical bias which is independent of the negative genetical bias considered above (see appendix). This statistical bias may well account for the discrepancy between the \hat{N}/N estimated from known North American samples and that estimated from scale analysis of West Greenland samples.

Conclusions

This analysis of the technique previously used to estimate the proportion of North American salmon off West Greenland from electrophoretic data has demonstrated that the assumption of Hardy-Weinberg equilibrium, which underlies this technique, should not hold for the populations concerned. This error leads to a significant underestimation of the proportion of North American salmon. Unfortunately, the degree of bias cannot be determined for specific samples from the West Greenland aggregate because it depends upon two parameters, $qTf4$ and F , which cannot be determined from the sample. The degree of bias is further obscured by interaction with a complex statistical bias. While increasing the size of the samples taken at West Greenland can be expected to reduce the statistical bias, it will have little effect on the genetical bias, which is dependent mainly upon population structure. Thus we are in the unfortunate

situation of knowing that we have a biased estimator but being unable to accurately determine the magnitude of the bias, or to eliminate or correct for it through sampling or computational changes. Our conclusions with regard to Payne's estimator are thus twofold. First, when applied to the West Greenland aggregate, it will have a negative bias of the order of 10%-30%. Second, this bias can neither be accurately determined nor eliminated. We therefore suggest that this estimator should not be included in future assessments of the North American contribution to the West Greenland fishery, and further, that past assessments which included this estimator should be reviewed in the light of the above conclusions.

Acknowledgements

I would like to thank John Pippy, Dave Reddin, Henry Lear, Paul Odense, Rex Porter and Ron Payne for their advice during the course of this project and for their many helpful criticisms of earlier drafts of the manuscript. Bill Doubleday deserves special thanks for suggesting consideration of the statistical bias, and Derek Roff's help with the graphics and simulation modelling was invaluable.

Table 1. Transferrin phenotypes observed in salmon caught at West Greenland (Payne 1973).

Phenotype	Genotype	Continent of origin
Tf1	Tf1/Tf1	Europe and North America
Tf1/4	Tf1/Tf4	North America
Tf4	Tf4/Tf4	North America
Tf1/2	Tf1/Tf2	Europe
Tf2	Tf2/Tf2	Europe

Table 2. Expected genotypic distribution under the assumption of Hardy-Weinberg equilibrium, and under the assumption of population subdivision (the Wahlund effect).

Genotype	Expected frequency*	
	Hardy-Weinberg (F=0)	Wahlund (F>0)
Tf1/Tf1	p^2	$p^2 + pqF$
Tf1/Tf4	$2pq$	$2pq(1 - F)$
Tf4/Tf4	q^2	$q^2 + pqF$

*p = frequency of the Tf1 allele, q = frequency of the Tf4 allele.

Table 3. \hat{N}/N and \hat{q}/q for anadromous river populations, based on the formulae from Payne 1973. (See text for definition of symbols.)

#	N	q	F	\hat{q}/q	\hat{N}/N	Source
1	338	0.26	0.12	1.38	0.74	a
2	120	0.32	-0.03	0.96	1.06	a
3	356	0.37	0.15	1.25	0.80	a
4	240	0.41	0.04	1.07	0.94	a
5	95	0.44	0.02	1.03	0.98	a
6	95	0.37	0.07	1.14	0.88	b
7	196	0.36	0.00	1.02	0.99	b
8	1564	0.31	0.16	1.36	0.74	c
9	881	0.20	0.12	1.50	0.67	d
10	2445	0.27	0.16	1.44	0.70	e

- a Sampled by Møller in 1969 (Møller 1970).
- b Sampled by Payne in 1972 (Payne 1974).
- c Combined anadromous populations sampled by Møller in 1969.
- d Combined anadromous populations sampled by Payne in 1972.
- e All anadromous populations.

Table 4. The proportion of North American salmon in the West Greenland aggregate as estimated from scale and electrophoretic data.

Year	Proportion North American		Reference
	Scales	Electrophoresis	
1970	0.34	0.23	1, 2
1971	0.35	0.53	1, 2
1972	0.36	0.20	1, 2
1974	0.34	0.27	1, 3
1975	0.44	0.41	1, 4
1976	0.41	0.35	5

1. Lear and Sandeman 1979.
2. Payne 1973.
3. Lear and Payne 1975.
4. Payne, Lear and Jensen 1976.
5. Payne and Cross 1977.

References

- Cavalli-Sforza, L.L. and W.F. Bodmer. 1971. The genetics of human populations. pp. 397-399. W.H. Freeman & Co., San Francisco.
- Lear, W.H. and R.H. Payne. 1975. MS. A comparison of scale analysis and serum electrophoresis as methods of determining the stock composition of Atlantic salmon off West Greenland in 1974. ICES Anadromous and Catadromous Fish Committee. C.M. 1975/MS.
- Lear, W.H. and E.J. Sandeman. 1979. Use of scale characters and a discriminant function for identifying continental origin of Atlantic salmon. ICES Res. Doc. (in press).
- Møller, D. 1970. Transferrin polymorphism in Atlantic salmon (Salmo salar). J. Fish. Res. Board Can. 27:1617-1625.
- Payne, R.H. 1974. Transferrin variation in North American populations of the Atlantic salmon, Salmo salar. J. Fish. Res. Board. Can. 31:1037-1041.
- Payne, R.H. 1973. MS. The use of serum transferrin polymorphism to determine the stock composition of Atlantic salmon in the West Greenland fishery. ICES Anadromous and Catadromous Fish Committee C.M. 1973/M:8, 7p.
- Payne, R.H. and T.F. Cross. 1977. MS. Liver aspartate aminotransferase polymorphism: a new tool for estimating proportions of European and North American salmon at West Greenland. ICES Anadromous and Catadromous Fish Committee C.M. 1977/M:10.
- Payne, R.H., W.H. Lear and J. Møller Jensen. 1976. MS. The stock composition of Atlantic salmon off West Greenland in 1975. ICES Anadromous and Catadromous Fish Committee C.M. 1976/M:8.
- Wallace, B. 1968. Topics in population genetics. pp. 100-103 and 112-114. W.W. Norton & Co. Inc. New York.
- Workman, P.L. 1969. The analysis of simple genetic polymorphisms. Human Biology 41:97-114.

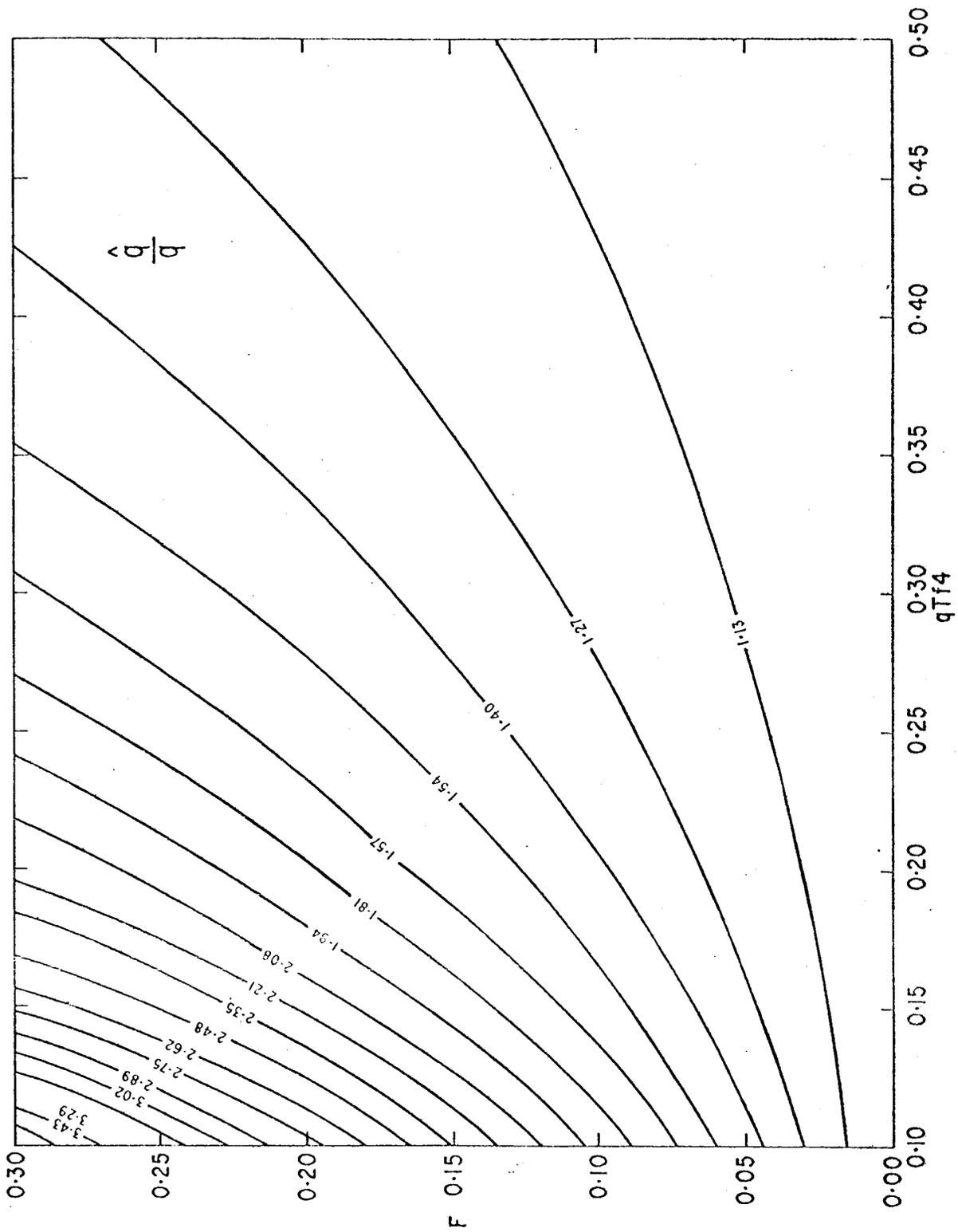


Fig. 1. The relationship between the estimated and true frequencies of the Tf4 allele for various values of F and q. (F = coefficient of deviation from Hardy-Weinberg expectations, q = true frequency of Tf4 allele, and, \hat{q} = estimated frequency of Tf4 allele (Payne 1973)).

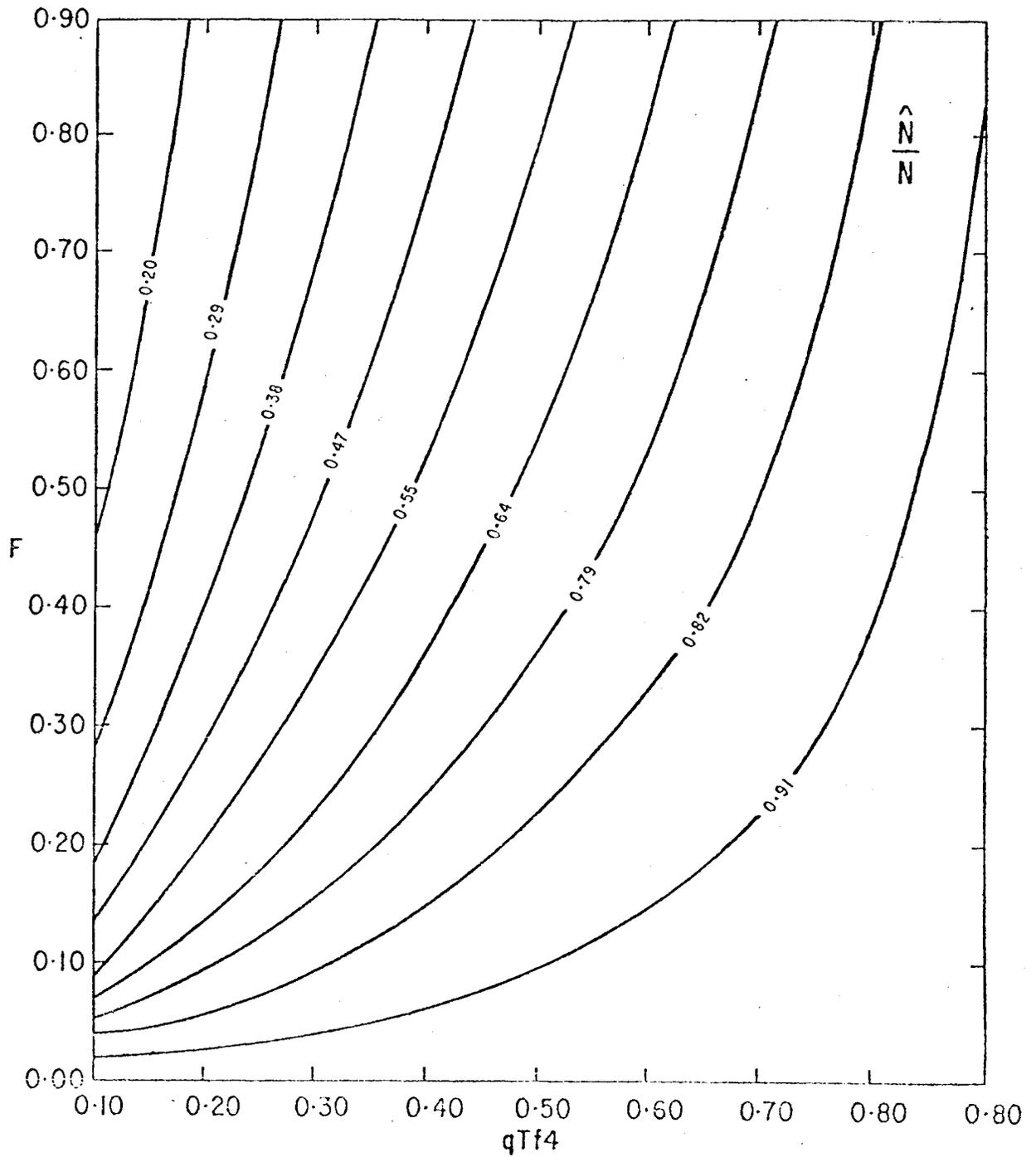


Fig. 2. The relationship between the estimated and true numbers of North American salmon in a sample for various values of F and q . (q , F , and \hat{q} are as in Fig. 1, N = true number of North American salmon and \hat{N} = number of North American salmon estimated according to Payne (1973))

Statistical Appendix

by

Daphne J. Fairbairn and Derek Roff
Research and Resource Services
Newfoundland Region
Department of Fisheries and Oceans
AIC 5X1

Because Payne's estimator of the number of North American salmon (\hat{N}) is a ratio estimator, it can be expected to have a statistical bias in addition to the genetical bias previously outlined. For simple ratio estimators, this bias is positive and of the order of $1/n$, where n is the sample size of the denominator (Cochrane 1977). It therefore will be of significance only for small samples. However, sampling error associated with sampling from a binomial distribution ($n(p+q)^2$), when both q and n are small, will also influence estimates of N . Therefore, it is difficult to predict, a priori, the direction and magnitude of the statistical bias associated with Payne's estimator. For this reason, we examined the relationship between sample size (N), frequency of the Tf4 allele (q), and estimated number (\hat{N}) of North American salmon via the following simulation.

The simulation population was in Hardy-Weinberg equilibrium ($F=0$) for alleles Tf1 and Tf4 at the transferrin locus. A random sample of size N was taken from this population, and the numbers of individuals with phenotypes Tf1/4 and Tf4 were used to estimate \hat{N} according to Payne's formulae. This process of sampling and calculating \hat{N} was repeated 500 times for each sample size and frequency of the Tf4 allele.

Since our population was in Hardy-Weinberg equilibrium for the Tf4 allele, it represented a panmictic population composed entirely of North American salmon. If Payne's estimator, \hat{N} , is an unbiased estimator, it should be distributed with both mean and median equal to N , the sample size, i.e.:

$$E(\hat{N}) = N$$

$$P(\hat{N} > N) = P(\hat{N} < N) = 0.5$$

Therefore, for each sample size we recorded the mean estimated \hat{N} , and the proportion of times that $\hat{N} > N$. This process was repeated using sample sizes from 10 to 190, at intervals of 20, for allele frequencies of 0.1, 0.2, 0.3, and 0.4.

The results of these simulations (a total of 20,000 sampling experiments) can be seen in Fig. a1 and a2. For $q = 0.1$, Payne's estimator of number of North American salmon has a strong negative bias, particularly at sample sizes less than 100. At sample sizes of 30 or less this negative bias is evident for $q = 0.2$ as well. For $q \geq 0.2$ and $N > 30$, the situation is more complex. While the median \hat{N} is close to N over this range (Fig. a1),

particularly as N and q increase, the mean of \hat{N} tends to be larger than N, particularly for $q = 0.2$. The mean \hat{N} is close to N only for $q \geq 0.3$, and $N > 90$. This difference between the mean and median values for \hat{N} suggests that the distribution of \hat{N} is skewed, with overestimates ($\hat{N} > N$) being more dispersed than underestimates. Thus, although the estimator is not biased with respect to the probability of an over- or an underestimate, it has a positive bias with respect to the magnitude of the error of the estimate: overestimates will tend to deviate more from N than will underestimates.

This complex statistical bias can be most easily understood if we substitute equation (1) for \hat{q} in equation (3), to give:

$$\hat{N} = \frac{(Tf1/4 + 2Tf4)^2}{4Tf4} \quad (a1)$$

From this equation we can see the importance of Tf4 in Payne's estimate of N. The expected value of Tf4 is q^2N which approaches 0.0 for small q and N. However, \hat{N} is undefined for $Tf4 = 0$ and, since Tf4 is an integer, can be calculated only when $Tf4 > 1$, even if $E(Tf4) \ll 1.0$. Thus, the values of Tf4 used to estimate \hat{N} will be positively biased with respect to q^2N . This positive bias in the denominator results in a negative bias in \hat{N} when both N and q are small. The positive bias in \hat{N} at larger values of N and q also derives from the presence of Tf4 in the denominator of equation (a1), and is the common positive bias associated with ratio estimators. This is illustrated in Table a1, for a population with $q = 0.25$, and a sample size of 40. Because $E(Tf4) = q^2N$, it remains small, and therefore causes this positive bias, until both N and q are quite large. At $q > 0.4$, sample sizes of 90 or more give a relatively unbiased estimate, while for $q = 0.2$, the estimate is still strongly biased at samples as large as 200.

The foregoing analysis indicates that in order to determine the degree of statistical bias associated with the samples taken at West Greenland, we must know both the number of North American salmon in the samples and the actual frequency of the Tf4 allele. The frequencies of the Tf4 allele estimated by Payne's technique (Table a2) can be expected to have a positive genetical bias of the order of 20%-40% (Fig. 1). Thus, although these estimates yield a mean \hat{q} of approximately 0.35, the true mean $qTf4$ is probably closer to 0.27, the estimate calculated previously from all anadromous salmon sampled from North American rivers. Similarly, the number of North American salmon calculated by Payne's technique (Table a2) includes a negative genetical bias. However, it is apparent from these estimates that the number of North American salmon sampled has always exceeded the range of negative statistical bias, given that $qTf4 > 0.2$. Therefore, we can expect a positive statistical bias in Payne's estimates of the number of North American salmon at West Greenland. This bias will be largest when $qTf4 < 0.3$, as in 1971 and 1974, but will be negligible when $qTf4 > 0.3$.

Reference

- Cochran, W.G. 1977. Sampling techniques. 3rd ed., John Wiley and Sons, Toronto.

Table a1. Illustration of the positive bias in \hat{N} for $N=40$, $q=0.25$ and $Tf1/4 = 2pqN = 15$. (The expected value of $Tf4$ is 2.5.)

Observed $Tf4$	\hat{N}	\hat{N}/N	$\bar{\hat{N}}$	$\bar{\hat{N}}/N$
1	72.3	1.80	46.8	1.17
2	45.1	1.13		
3	36.8	0.92		
4	33.1	0.83		

Table a2. Total sample sizes, estimated number of North American salmon (\hat{N}), and estimated frequency of the $Tf4$ allele (\hat{q}) for West Greenland samples.

Year	Total sample size (North American & European)	\hat{q}	\hat{N}		Reference
			Scales	Electrophoresis	
1970	987	0.44		226	1
1971	173	0.26		92	1
1972	475	0.43		96	1
1974	511	0.28	174	138	2
1975	522	0.37	228	214	3

1. Payne 1973.
2. Lear and Payne 1975.
3. Payne, Lear and Jensen 1976.

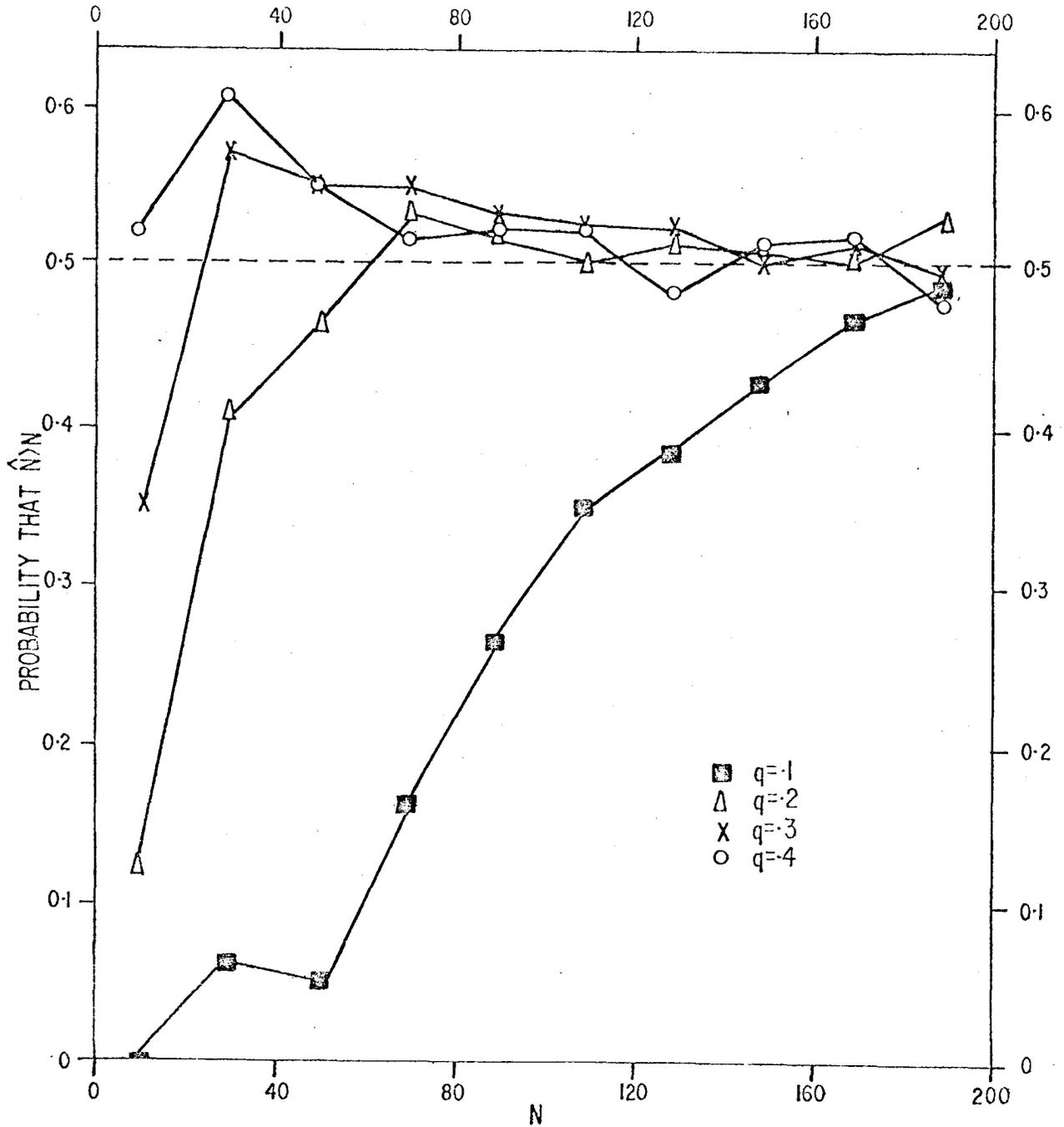


Fig. a1. Probability of an overestimate of number of North American salmon as a function of sample size. (Number of iterations per point = 500.)

Fig. a2. Ratio of mean estimated number of North American salmon (\bar{N}) to actual number as a function of sample size. (Number of iterations per point = 500.)

