International Council for the Exploration of the Sea

C.M. 1963

Symposium on the Measurement of Abundance of Fish Stocks

No. 4

Statistical Aspects of Abundance Estimates

by

K. P. Andersen
Danmarks Fiskeri- og Havundersøgelser
Charlottenlund, Denmark

The most important abundance estimate - in any case the most used - is the catch per unit effort.

When data for catch and effort have been collected the first task to be done is a statistical analysis of the figures in order to find the best estimates of the quantities of interest for the study of population dynamics.

This analysis of catch and effort data is completed so to say, when the statistical distributions in question are specified, and this paper will deal with some distributions that are often met with in practice, and with some seldom realized difficulties that arise in connection with the distributions. The paper contains very little new and is only thought as an attempt to clear things up a little.

The dream of a student of population dynamics - as John Gulland has already mentioned - is a situation, where the instantaneous fishing mortality coefficient is proportional to the effective overall fishing intensity. This situation is realized, if the mathematical expectation of the catch y is proportional to the product of effort and total number of fish N in the (small) area in question. This we write:

$$E(y) = k \times f \times N$$

where k is a factor of proportionality.

This is of course a question of defining f in the right way, and that is exactly the question with which Gulland dealt in his paper. I shall, therefore, simplify the situation by assuming that f is well defined for the individual ships, which means that f can be measured accurately except for fishing power.

To fix ideas let us take trawling as our standard example. We will suppose that hours of fishing is an exact measure of effort except for fishing power. The model can now be written:

$$E(y) = \Pi \times \tau \times N$$

where Π is the fishing power of the ship in question, τ number of trawling hours, and N the total number of fish in the area. k has been taken as 1, which means that all questions of vulnerability, availability, etc. are set aside and the effort is simply $\Pi \times \tau$.

We will write a single catch as:

$$y = \rho \times \tau \times N + \varepsilon = \omega \times \tau \times \quad \times N \times \Delta$$

where ρ and ω are related to fishing power, and ε and Δ are stochastic components. We shall first deal with fishing power and examine the consequences of three different distributions of y:-

A) y normally distributed (m,σ). The distribution function is:

$$\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(y-m)^2/2\,\sigma^2}$$

B) y log-normally distributed (α, γ). Distribution function:

$$\frac{1}{\sqrt{2\pi}\,\gamma}\;\frac{1}{y}\; e^{-(\log y-\alpha)^2/2\gamma^2}$$

C) y negative-binomially distributed. The distribution is discrete and determined by two parameters m and k. The probabilities are:

$$P(y=r) = \left(1 + \frac{m}{k}\right)^{-k} \binom{k+r-1}{r} \left(\frac{m}{m+k}\right)^r ; \quad (r = o,1,2,\ldots\ldots)$$

and

$$E(y) = m$$

$$V(Y) = m + \frac{m^2}{k}$$

If in case A $m = \rho\,\tau\,N$ and $\sigma = \varphi\sqrt{\tau}$ the expectations for two ships with identical fishing times are:

$$E(y_1) = \rho_1 \times \tau \times N \equiv \pi_1 \times \tau \times N$$

$$E(y_2)\;\; \rho_2 \times \tau \times N \equiv \pi_2 \times \tau \times N$$

or $\pi \equiv \rho$. The assumption $V(y) = \varphi^2\tau$ says that one haul of τ hours duration is statistically equal to τ one hour hauls.

If we want to compare the fishing powers of two ships, we have to estimate $\pi_1/\pi_2 = \rho_1/\rho_2$, which means that we have to estimate the ratio of the means of two normal distributions.

If the ships have taken n τ-hour hauls each in the area in question, the simplest estimate for ρ_1/ρ_2 is

$$z = \frac{\frac{1}{n}\sum y_1}{\frac{1}{n}\sum y_2} = \frac{\bar{y}_1}{\bar{y}_2}$$

The distribution of z is

$$f(z) = \frac{m}{2\pi\,\varphi_1\,\varphi_2\,\tau}\left[\int_0^{\infty} e^{-n(zw-\rho_2\tau N)^2/2\varphi_2^2\tau} \times e^{-n(w-\rho_1\tau N)^2/2\varphi_1^2\tau}\,dw\right.$$

$$\left.-\int_{-\infty}^{0} e^{-n(zw-\rho_2\tau N)^2/2\varphi_2^2\tau} \times e^{-n(w-\rho_1\tau N)^2/2\varphi_1^2\tau}\,dw\right]$$

when $\rho_2\tau N \gg \varphi_2\sqrt{\tau}$ this is very nearly:

$$\frac{n}{2\pi\,\varphi_1\varphi_2\,\tau}\int_{-\infty}^{\infty} e^{-n(zw-\rho_2\tau N)^2/2\varphi_2^2\tau} \times e^{-n(w-\rho_1\tau N)^2/2\varphi_1^2\tau}\,dw$$

$$= \frac{\sqrt{n}}{\sqrt{2\pi}} \times \frac{\rho_2 \tau N \varphi_1^2 \tau + \rho \tau N \varphi_2^2 \tau Z}{(\varphi_1^2 \tau + \varphi_2^2 \tau z^2)^{\frac{3}{2}}} \times e^{-\frac{n}{2} \times \frac{\rho_1 \tau N - \rho_2 \tau N z}{\varphi_1^2 \tau + \varphi_2^2 \tau z^2}}$$

As $f(z) \sim 1/z^2$ for $z \to \pm \infty$ it follows that $E(z)$ does not exist, and it canbe shown that there is no better estimate for ρ_2/ρ_1.

The situation is not hopeless though, as confidence limits can be found for ρ_1/ρ_2 by means of z.

As $\bar{y} - \bar{y}_2\ (\rho_1/\rho_2)$ is normally distributed $(0, \sqrt{\varphi_1^2\tau/n + (\rho_1/\rho_2)^2\,\varphi_2^2\tau/m})$ we have that

$$\frac{(y_1 - (\rho_1/\rho_2)\, y_2)}{s(\,1/n + (\rho_1/\rho_2)^2/n)^{\frac{1}{2}}}$$

where s is the combined estimate for $\varphi\sqrt{\tau}$,[1] is t-distributed with $2n - 1$ degrees of freedom, and this gives us the following confidence intervals:

$$\frac{z \pm \dfrac{ts}{y_2}\,(1 + z^2)/n - t^2s^2/n^2y_2^2\,)^{\frac{1}{2}}}{(1 - t^2s^2/n^2y_2^2\,)}$$

If $my_2 \gg s$ we get:

$$z \pm ts(1 + z^2)^{\frac{1}{2}}/y_2\sqrt{m}$$

which is the same as the "naive" formula

$$V(z) = \varphi^2\tau/ny_2^2 + \varphi^2\tau\, y_1^2/ny_2^4 = \varphi^2\tau(1 + z^2)/ny_2^2$$

gives.

In case B and $(\alpha, \chi) = (\log \omega \tau N, \sigma)$ - Δ log-normally distributed $(0,\sigma)$ - our assumptions give

$$E(y_1) = \omega_1 \times \tau \times N\ E(\Delta_1) = \omega_1 \times \tau \times N\ e^{\frac{\sigma^2}{2}} \equiv \Pi_1 \times \tau \times N$$

$$E(y_2) = \omega_2 \times \tau \times N\ E(\Delta_2) = \omega_2 \times \tau \times N\ e^{\frac{\sigma^2}{2}} \equiv \Pi_2 \times \tau \times N$$

or $\Pi = \omega\, e^{\sigma^2/2}$.

When comparing two ships we want to estimate $\Pi_1/\Pi_2 = (\omega_1/\omega_2)\, e^{(\sigma^2 - \sigma^2)/2}$ and it follows that $\Pi_1/\Pi_2 = \omega_1/\omega_2$ only if $\sigma_1 = \sigma_2$.

The best estimate of $\log(\omega_1/\omega_2)$ is of course:

$$\zeta = \overline{\log y_2} - \overline{\log y_1} = \frac{(\log \Pi\, y_2)/n}{(\log \Pi\, y_1)/n}$$

which is normally distributed $(\log(\omega_1/\omega_2), \sqrt{(\sigma_1^2 + \sigma_2^2)/n})$ and thus e^{ζ} has the expectation:

$$E(e^{\zeta}) = \frac{\sqrt{n}}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} \int_{-\infty}^{\infty} e^{\zeta}\, e^{-\frac{n(\zeta - \log\frac{\omega_1}{\omega_2})^2}{2(\sigma_1^2 + \sigma_2^2)}}\, d\zeta = \frac{\omega_1}{\omega_2}\, e^{\frac{\sigma_1^2 + \sigma_2^2}{2n}}$$

---

1) For simplicity we assume $\varphi_1 = \varphi_2$.

This formula shows that e^{ζ} is unbiased for $\frac{\omega_1}{\omega_2}$ if n is great. If this is not the case the estimate

$$e^{\zeta}\ e^{-\frac{s\,s}{2n}}$$

has much smaller bias.

In the first part I have been thinking of a situation where two ships have been fishing randomly in an area. Let us now think of comparative trawlings and formulate as follows:

$$y_{1i} = \pi_1 \times \tau \times A_i + \varepsilon_{1i} = \omega_1 \times \tau \times A_i \times \Delta_{1i}$$

$$y_{2i} = \pi_2 \times \tau \times A_i + \varepsilon_{2i} = \omega_2 \times \tau \times A_i \times \Delta_{2i}$$

with obvious conventions of notation.

In case A it is natural to try to use regression analysis, and we must distinguish between two cases:

1) N_i is a stochastic variabel.
2) N_i is not a stochastic variabel.

Let us in case 1) suppose that (y_1, y_2) is normally distributed

$$\left(\pi_1 \tau A,\ \pi_2 \tau A,\ \sqrt{\pi_1^2\tau^2\Sigma^2+\sigma_1^2},\ \sqrt{\pi_2^2\tau^2\Sigma^2+\sigma_2^2}\ \frac{\pi_1\pi_2\tau^2\Sigma^2}{\sqrt{(\pi_1^2\tau^2\Sigma^2+\sigma_1^2)\ (\pi_2^2\tau^2\Sigma^2+\sigma_2^2)}}\right)$$

The regression coefficient is

$$\beta_{y_1|y_2} = \frac{\pi_1\pi_2\tau^2\Sigma^2 \times \sqrt{\pi_1^2\tau^2\Sigma^2+\sigma_1^2}}{\sqrt{(\pi_1^2\tau^2\Sigma^2+\sigma_1^2)\ (\pi_2^2\tau^2\Sigma^2+\sigma_2^2)}\ \sqrt{\pi_2^2\tau^2\Sigma^2+\sigma_2^2}} = \frac{\pi_1\pi_2\tau^2\Sigma^2}{\pi_2^2\tau^2\Sigma^2+\sigma_2^2}$$

$$= \frac{\pi_1/\pi_2}{1 + \frac{\sigma_1^2}{\pi_2^2\tau^2\Sigma^2}}$$

and thus the regression coefficient is an underestimate of the fishing power unless $\sigma_2 \ll \pi^2\tau^2\Sigma^2$.

In case 2) we have to suppose $\sigma_2 \simeq 0$, which gives:

$$y_{1i} \approx \frac{\pi_1}{\pi_2}\, y_{2i} + \varepsilon_{2i}$$

and this demonstrates that the normal regression coefficient is an unbiased estimate of the fishing power.

If $\sigma_2 \neq 0$ the regression coefficient is a biased estimate of the fishing power, but as far as I know it is very difficult to do something rational to repair this. It is, however, clear that $\sigma_2 \ll$ (variation in y_2) is just as good as $\sigma_2 \simeq 0$.

In case B) with Δ log-normally distributed $(0, \sigma)$ we get:

$$\log \frac{y_{1i}}{y_{2i}} = \log \frac{\omega_1}{\omega_2} + (\log \Delta_{1i} - \log \Delta_{2i})$$

which says that y_{1i}/y_{2i} is log-normally distributed $(\log \frac{\omega_1}{\omega_2}, \sqrt{\sigma_1^2 + \sigma_2^2})$.
As

$$E\left(e^{\overline{\log y_1} - \overline{\log y_2}}\right) = E\left(\sqrt[n]{\prod \frac{y_2}{y_1}}\right) = \frac{\omega_1}{\omega_2} e^{\frac{\sigma_1^2 + \sigma_2^2}{2n}}$$

the situation is quite analogous to the former log-normal case.

The third distribution, the negative binomial distribution, is often found when the fish are forming shoals. One can show that the best estimates for k and m are determined by

$$\hat{m} = \frac{\sum y}{n} \quad ; \quad \left(V(\hat{m}) = \frac{m + m^2/K}{n} \right)$$

and

$$n \log \left(1 + \frac{\sum y}{\hat{k} n}\right) = \sum_{r=1}^{\infty} n_r \sum_{i=0}^{r-1} \frac{1}{\hat{k} + i}$$

As far as I know there is no simple way to handle the negative binomial distribution in respect to fishing power. But if the material is great enough it might be possible to pool the data and operate with the catch in n hauls, and in this case the central limit theorem leads over in the normally distributed case.

I have dwellt quite a lot on the question of the power factor, and the reason for this is that it gives a good opportunity to specify the distributions. I shall now take for granted that we are able to specify the fishing power, which again means that we are able to measure the real fishing effort accurately.

The catch model is simply:

$$E\left(\frac{y_i}{f_i}\right) = N$$

We shall now examine the best estimates of N that the different distributions give rise to.

If in case A) it is supposed that $\sigma^2 = \varphi^2 f$ it follows that:

$$E(y|f) = fN$$

and

$$V(y|f) = \varphi^2 f$$

N is consequently the regression coefficient in an ordinary regression with zero interception. The best estimate of N is:

$$\hat{N} = \frac{\sum y}{\sum f} \quad ; \quad \left(V(\hat{N}) = \frac{\varphi^2}{\sum f} \right)$$

If it on the other hand is supposed that:

$$V(y|f) = \varphi^2 f^2$$

the best estimate is:

$$\hat{N} = \frac{\sum \frac{y}{f}}{n} \quad ; \quad \left(V(\hat{N}) = \frac{\varphi^2}{n} \right)$$

The difference between the two hypothesis can be interpretated in the following way:

Assuming $V(y|f) = \varphi^2 f$ is equivalent to saying that the catch taken by an effort f is statistically equal to f catches each caught by means of an effort 1. This will be the case if the fish are distributed at random.

If on the other hand $V(y|f) = \varphi^2 f^2$ the fish are clustered, which causes that a single haul with effort f is statistically equivalent to f x y and not to $\sum y$.

For a log-normal distribution the effort was $f = \omega e^{\sigma^2} \tau$, but if $f^* = \omega \tau$ is used instead of f, one get:

$$\log \frac{y}{f^*} = \log N + \log \Delta$$

and supposing that σ is independent of f the following best estimates are found:

$$\widehat{\log N} = \overline{\log \frac{y}{f^*}}$$

$$\hat{N} = e^{\widehat{\log N}} e^{-\frac{\sigma^2}{2n}}$$

The assumption that σ is independent of f is the same as the assumption $V(y|f) = \varphi^2 f$.

As mentioned earlier the negative binomial distribution is found when the fish are forming shoals, which again means that heavy clustering takes place. One would expect the following catch probabilities:

$$P(y = n) = \left(1 + \frac{fN}{k}\right)^{-K} \left(\frac{k + n - 1}{n}\right) \left(\frac{fN}{fN + k}\right)^n$$

with

$$E(y|f) = fN$$
$$V(y|f) = fN + f^2N^2/k$$

If k is great one get:

$$V(y|f) \approx fN = E(y|f)$$

and the distribution is actually a Poison distribution, which is very nearly normal with $\sigma^2 = \varphi^2 f$ and we are back in a known situation.

If k is small:

$$V(y|f) \approx f^2N^2/k$$

and it seems natural to use the formulaes from the other normal case.

This paper has only dealt with very simple situations, and it has by no means exhausted the matter. It is, however, my hope that it has cleared up some often discussed points, and I also hope it can provoke a useful discussion.

References

Kendall, M. G. & Stuart, A. "The advanced theory of statistics". Vols. 1 and 2. Charles Griffin & Co. Ltd., London.

Geary, R. C. "The frequency distribution of the quotient of two normal variables". J.Roy.Statist.,Soc., 93, 442 (193o).

Fieller, E. C. "The biological standardization of insulin". Suppl. J.Roy. Stat. Soc., 7, 1 (194o).

Finney, D. J. "On the distribution of a variate whose logarithm is normally distributed". Suppl. J.Roy.Stat.Soc., 7,155 (1945).

Anscombe, F. J. "Sampling theory of the negative binomial and logarithmic series distribution". Biometrika, 37, 358 (195o).
