November 11, 1941

Dear Mr Corbet,

The reason I have not before answered your letter of October 22nd, enclosing the interesting data you have gathered on frequency of capture of 620 species of Malayan butterflies, is that it raised in my mind some questions which could only be resolved by fairly elaborate calculations.

The agreement of your series of frequencies with the Harmonic Progression $\frac{1}{n}$ is rather striking, except for small values of n, over the range for which your data are reliable, i.e., up to 24 captures. It is obvious, however, that the law breaks down hopelessly, both for higher frequencies, since the series

$$\frac{1}{n} + \frac{1}{n+1} + \frac{1}{n+2} \dots$$

is divergent, so that your expectation wuld be an infinite number of species with more than 24 captures, while it also gives an infinite expectation for zero captures.

It seems to me that the reason why your series agrees well with this formula in the middle region is that the harmonic series, for a finite run of terms, rather closely mimics the series known as the negative binomial when the papameter k representing the exponent is small.

The general formula in this series for the number of species giving <u>n</u> captures is

$$\frac{(k+n-1)!}{n!(k-1)!} \quad \frac{p^n}{(1+p)^{k+n}}$$

This expression for the expected frequencies of species giving <u>n</u> captures is related to an intelligible frequency distribution for the effective density of different species, which is best expressed in terms of the expectation of captures in a given collection appropriate to different densities. Thus, if <u>m</u> is the number of captures of a given species to be expected from its actual abundance, then <u>m</u>, unlike <u>n</u>, need not be a whole number, and for the rarer species may be a small fraction. For the common species, of course, <u>n</u> would not often differ from <u>m</u>, at least by any large multiple of the square root of <u>m</u>; so that if <u>m</u> were 100 the actual number of captures <u>n</u> would usually lie between 80 and 120. The distribution of the expectation <u>m</u>, which corresponds with the negative binomial as the distribution for <u>n</u>, [redacted] has the frequency element

$$df = \frac{1}{(k-1)!} \quad p^{-k} \quad m^{k-1} \quad e^{-\frac{m}{p}} \quad dm\,.$$

If <u>k</u> is less than 1, this decreases steadily as <u>m</u> increases; this is the case with your series. If <u>k</u> is greater than 1, it would commence by increasing and later decrease, and there would be a maximum, or mode, in the distribution of the expectation. Clearly the actual expectations are proportional to the size of the collection, supposing collection to continue using the same methods, and this is assured by the parameter <u>p</u> being proportional to the total number of insects captured. The other parameter <u>k</u> is intrinsic to the natural distribution of abundance within the group.

Putting $n = 0$ in the formula for the negative binomial, it appears that the fraction of species not represented in the collection will be

$$(1-p)^{-k}$$

This fraction cannot, of course, be observed unless the region had been already exhaustively studied. It may, however, be estimated, though roughly, from your series of frequencies. I have, therefore, fitted the negative binomial series to your data on two suppositions:

(a) that there are really 1050 species of butterflies in the region, of which 430 have not been captured, and

(b) on the supposition that there are 1200 species in all, of which 580 have not been captured. In the first case you will see that the fit is not very close, the numbers observed being in excess ~~a~~ for one and two captures, and also to a less extent for more than 15 captures. In the second trial the agreement is considerably better, though the deviations throughout are still in the same directions~~x~~ in each group - though much smaller. In fact, if the data were used to ~~establish~~ estimate the total number of species, the estimate would certainly be somewhat in excess of 1200, although the deviations from the supposition that the number is 1200 do not seem sufficiently great for anyone to assert that ~~the~~ a higher number is actually required. It is worth while comparing these two frequency distributions with that derived from the harmonic series, which, if we ignore the absurdity of an infinite number of species with more than 24 caught, may be taken to represent the limiting hypothesis $k = 0$. The deviations here are larger, and in the important region up to 10 captures in the opposite direction, showing that the data are better fitted by a finite value of k, about

$\frac{1}{5}$ and a finite number of species not much less than 1200.

Yours sincerely,

Expected Numbers of Species

1050 k .2484091	deviations	1200 .1947627	deviations	Harmonic series	deviations	Observed
103.879		110.260		132.682		118
63.059		64.292		66.341		74
45.962		45.910		44.227		44
36.300		35.791		33.170		24
29.995	+9.805	29.308	+3.439	26.536	−13.956	29
25.517		24.768		22.114		22
22.151		21.395		18.955		20
19.518		18.781		16.585		19
17.396		16.691		14.742		20
15.647	−4.229	14.980	−0.615	13.268	+10.336	15
14.177		13.552		12.062		12
12.923		12.340		11.057		14
11.842		11.299		10.206		6
10.898		10.394		9.477		12
10.067	−9.907	9.601	−7.186	8.845	− 1.647	6
9.331		8.899		8.293		9
8.673		8.275		7.805		9
8.082		7.716		7.371		6
7.549		7.212		6.983		10
7.066		6.756		6.634		10
6.626		6.347		6.318		11
6.223		5.963		6.031		5
5.854		5.616		5.769		3
5.515	+1.081	5.299	+3.922	5.528	+5.268	3
115.750	+ 3.250	118.560	+0.440		~~[illegible]~~ $-\infty$	119