Biology 300

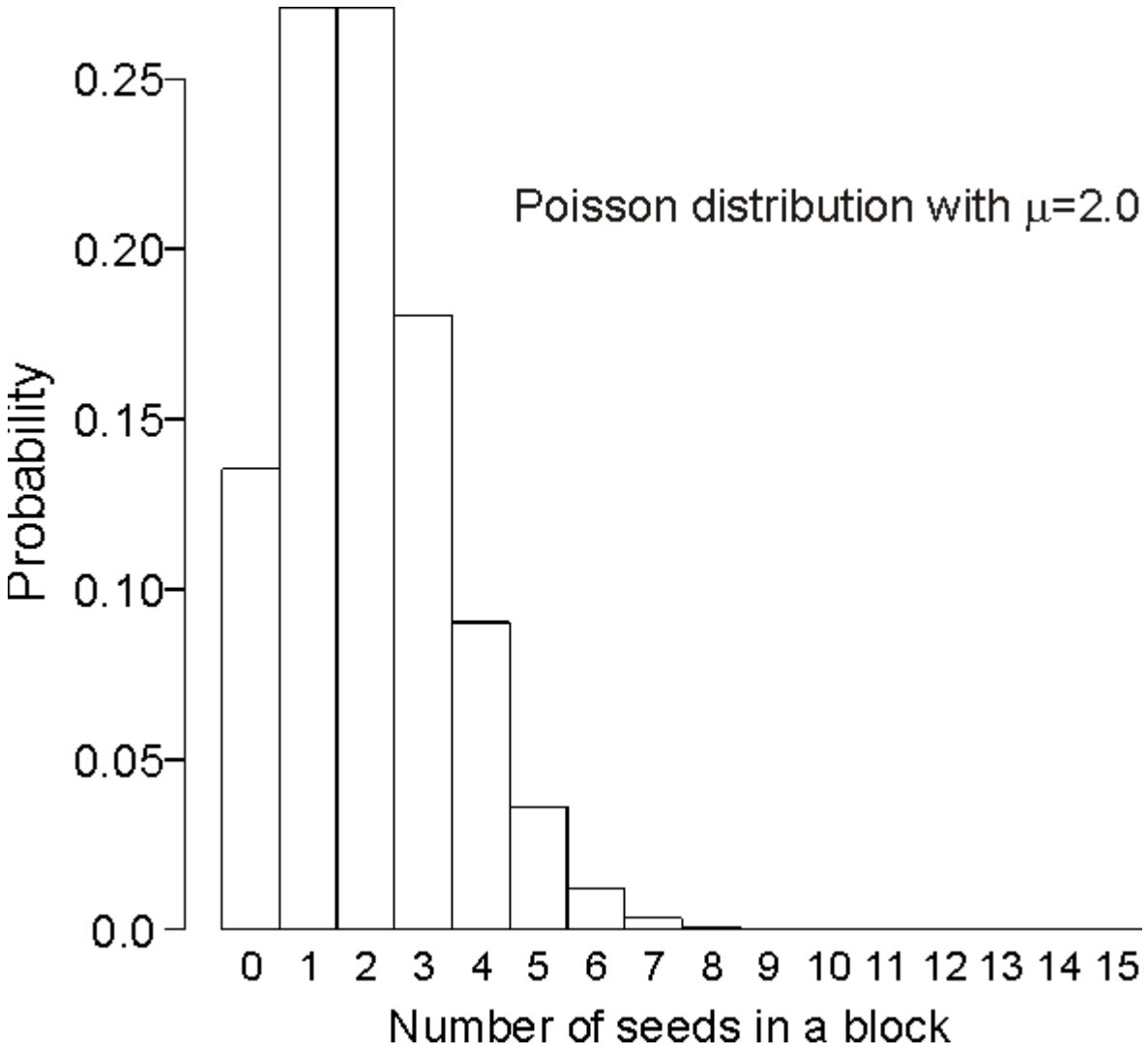
Notes on the Poisson distribution

The Poisson distribution has a number of uses in biology. Its foremost utility is in providing us with a tool to test whether the pattern of events in space or time is "random" or not.

1) Seed example.
For example, imagine that you were to scatter seeds over a vast field from a plane. Imagine also that you have divided the field up into blocks of equal size, say 10 by 10 metres in area. If the probability that a given square millimetre of soil receives a seed is low (you haven't dropped a trillion seeds, just a few thousand), and if this probability is the same everywhere across the entire field, and if seeds are independent of each other then the number of seeds per block should follow a Poisson distribution:

$$P(X) = \frac{e^{-u}u^{x}}{X!}$$

Here is what the Poisson distribution would look like if the mean number of seeds per block is 2.0:



If the probability that a given square millimetre of soil receives a seed is NOT the same everywhere across the entire field, or if seeds are NOT independent of each other, then the number of seeds per block should NOT follow a Poisson distribution. In this case either of two nonrandom patterns of seed distribution may be observed, "clumped" or "dispersed". Clumped means that seeds are somehow aggregated, such that some blocks have excessive numbers of seeds and others have too few, compared with the random expectation. Dispersed means that the seeds are distributed too evenly across the field, such that each block has about the same number.

Note that unlike the binomial distribution, we are not counting the number of successes in *n* independent trials here. We are simply counting the number of events in a block.

The example above examined a *spatial* pattern of events. The Poisson distribution also applies to the pattern of events in blocks of *time*, as the next example shows.

2) Is the distribution Poisson?
The main use of the Poisson distribution is in providing us with a null hypothesis for events in space or time. Deviations from this hypothesis tell us something about real processes in nature. For example, consider the distribution of extinction events over intervals of time across millions of years of fossil history. The best record of extinctions through earth's history come from fossil marine invertebrates, because they have shells and preserve well. Below I list the number of recorded extinctions of marine invertebrate families (a high-level taxonomic category) in 76 intervals of time, all of equal duration:

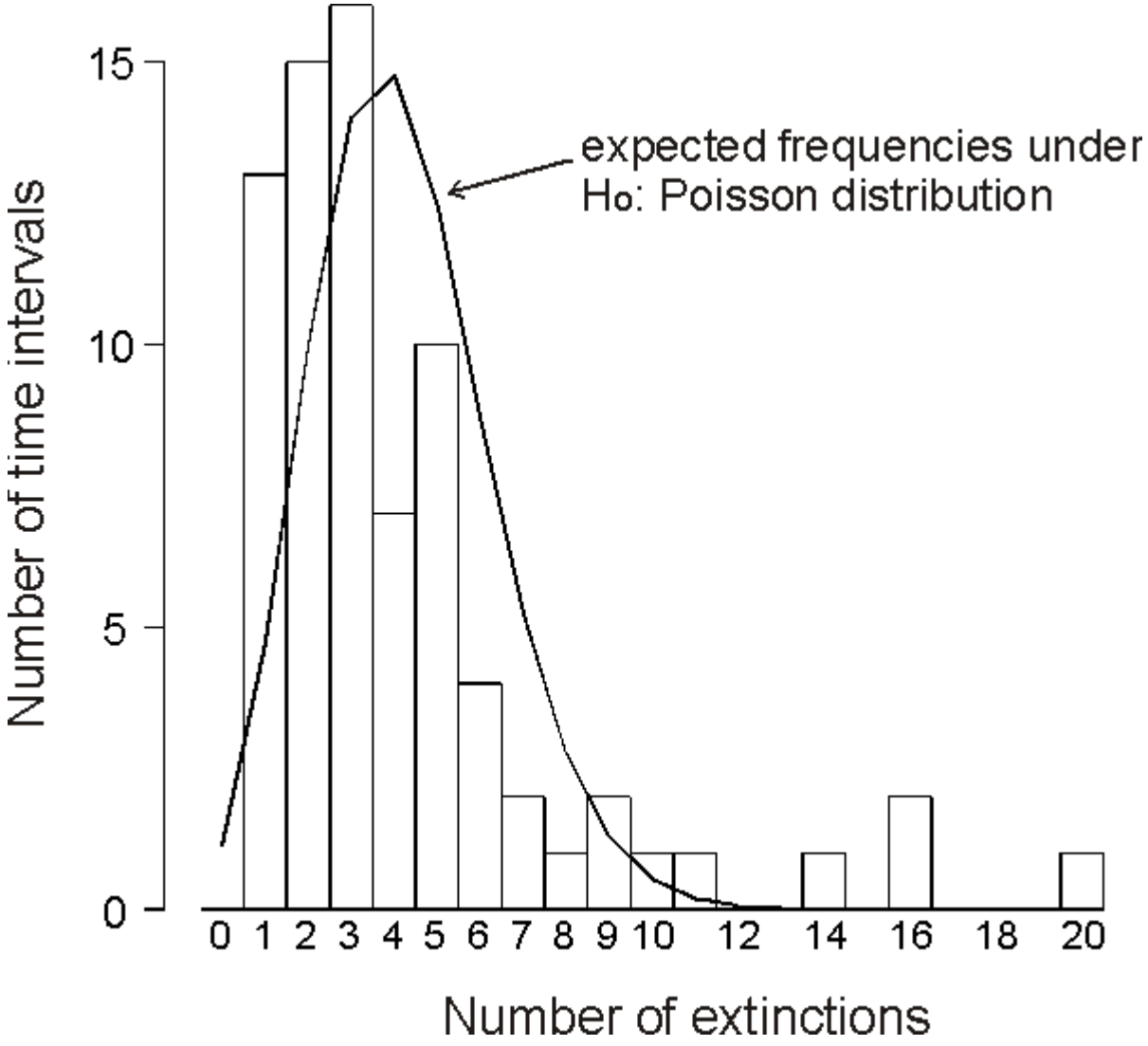
Number of extinctions recorded	Number of time intervals
0	0
1	13
2	15
3	16
4	7
5	10
6	4
7	2
8	1
9	2
10	1
11	1

12	0
13	0
14	1
15	0
16	2
17	0
18	0
19	0
20	1
Total	**76**

The question of interest is whether the pattern of extinction events through the fossil record is "random" in time, or whether instead extinctions tend to be clumped and occur in bursts ("mass extinctions"). Alternatively, extinctions may occur in a dispersed pattern. The easiest way to test this is to compare the frequency distribution of extinctions to that expected from a Poisson distribution using a chi-square goodness of fit test. Our hypotheses are:

Ho: The number of extinctions per time interval has a Poisson distribution
Ha: The number of extinctions per time interval does NOT have a Poisson distribution.

To begin the test, we need to estimate the mean number of extinctions per time interval. This is obtained by summing the measurements (13 intervals had 1 extinction, 15 had 2, 16 had 3, ...) and dividing by the total number of intervals, 76, yielding a sample mean of 4.21. This sample mean is used in place of μ in the formula for the Poisson distribution to generate the expected frequencies. I will write the table below and give more details, but first look at the result:



The histogram gives the observed frequencies, whereas the line gives the expected frequencies under the null hypothesis. Clearly there is a discrepancy. Compared with the Poisson distribution, the data show too many time intervals having many extinctions and too many having too few. But is the discrepancy between the observed and expected distributions statistically significant? We will use the chi-square goodness of fit test to determine this.

Below I have tabulated the observed and expected frequencies, and noted how these were calculated. I have grouped $X \geq 10$ extinctions, as the larger X values are infrequent. All but the last expected frequency is computed by applying the formula for the Poisson distribution to get the probability and then multiplying this probability by the total number of intervals, 76. *The expected frequency for the final category is computed by subtracting the sum of all the previous expected values from 76.*

Number of extinctions X	**Observed number of time intervals** f_i	**Expected number of time intervals** $\hat{f}_i$	**How expected frequencies were computed**
0	0	1.13	$(e^{-4.21}\ 4.21^0 / 0!)\ 76$

1	13	4.75	$(e^{-4.21}\ 4.21^{1}\ /\ 1!)\ 76$
2	15	10.00	$(e^{-4.21}\ 4.21^{2}\ /\ 2!)\ 76$
3	16	14.03	etc
4	7	14.77	etc
5	10	12.44	etc
6	4	8.72	etc
7	2	5.24	etc
8	1	2.76	etc
9	2	1.29	$(e^{-4.21}\ 4.21^{9}\ /\ 9!)\ 76$
≥10	6	0.86	76-(1.13+4.75+10.00+...+1.29)
Total	**76**	**76**	

However, too many of the expected frequencies are too small for the chi-square test. To meet our rule of thumb (no expected frequencies less than 1.0 and no more than 20% less than 5), I have grouped categories in the following way:

Number of extinctions ***X***	**Observed number of time intervals** f_i	**Expected number of time intervals** $\hat{f}_i$
0 or 1	13	5.88
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
≥8	9	4.91

Total	**76**	**76**

Using the standard formula for the chi-square statistic I computed $\chi^2 = 23.94$. The degrees of freedom for this statistic are 6. There are k=8 categories, which would normally provide us with k-1=7 degrees of freedom. However, when we computed our expected frequencies we had to use the data to estimate μ , since μ was not provided for us by the null hypothesis. This step cost us one more degree of freedom, leaving 6. The critical value for χ^2 with 6 degrees of freedom and α =0.05 is 12.59. Our χ^2 statistic of 23.94 exceeds this critical value, therefore our *P*-value is less than 0.05. We reject the null hypothesis. Extinctions in the fossil record do not fit a Poisson distribution.

A final useful comparison is between the sample mean number of extinctions, 4.21, and the sample variance in number of extinctions, which I computed as 13.72. Since the sample variance greatly exceeds the sample mean, the distribution of extinction events in time is "clumped". That is, extinctions tend to occur in bursts ("mass extinctions") rather than randomly in time.