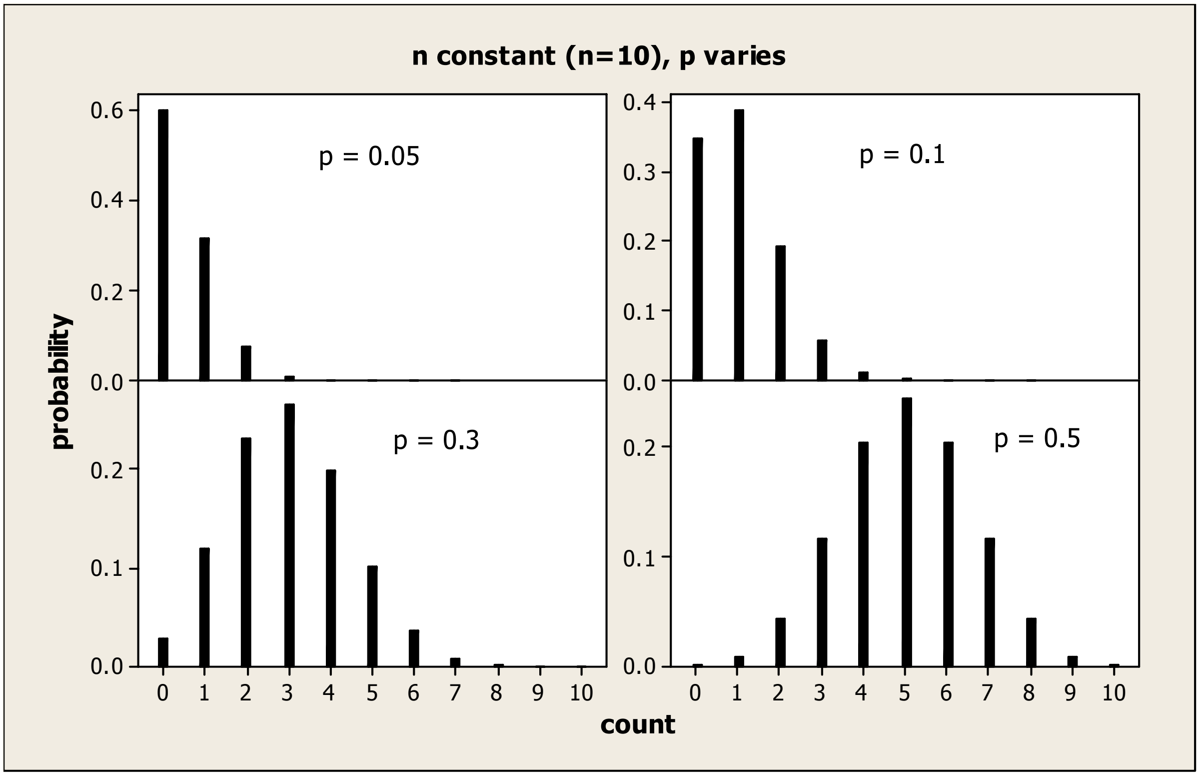
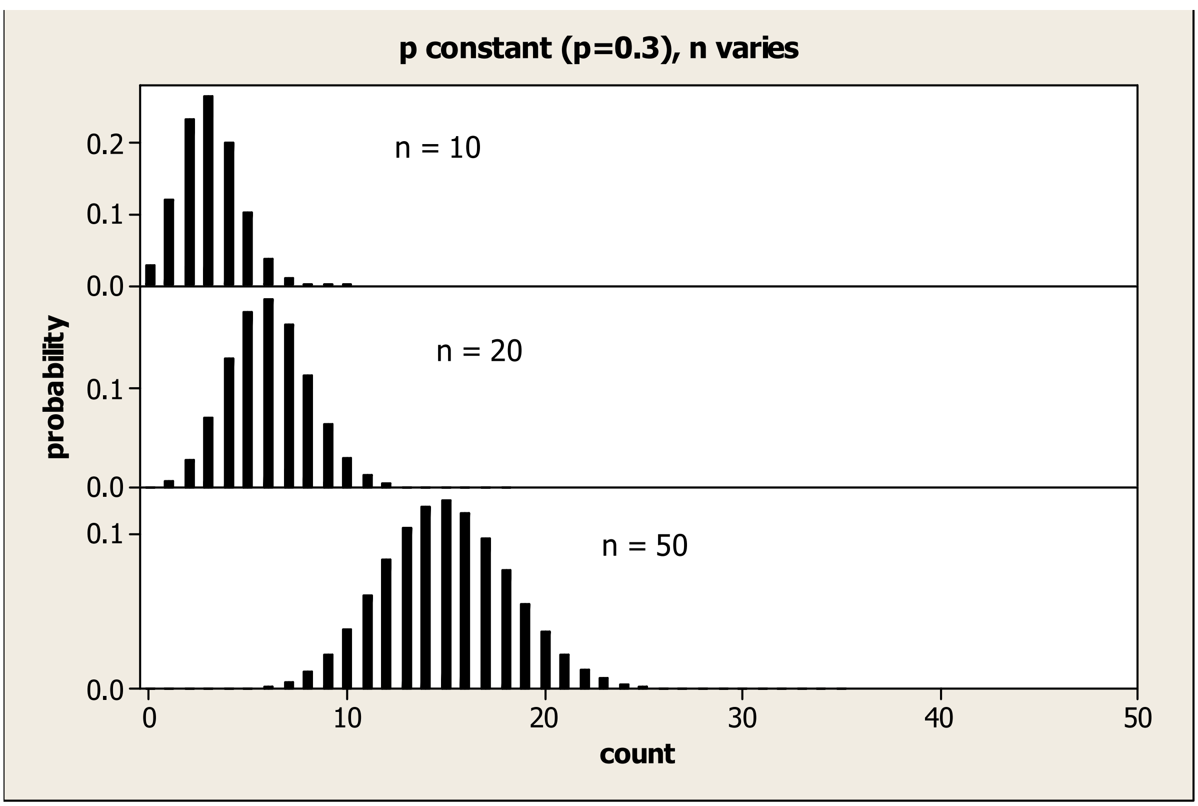
Binomial and Other Distributions

Binomial Distributions





Poisson Distribution

Derivation:

The Poisson distribution represents the extreme case of a binomial distribution in which the number of trials (n) is very large but the probability of "success" (p) is very small, so that the mean, np, remains moderate.

Applications:

Counts of events per spatial or temporal unit, *i.e.* the distribution of things in time or space.

Examples:

- plants per quadrat
- parasites per host individual
- mutations per DNA site, or per generation at one site
- cells / square in microscope slide

The idea is that there conceivably could be very many of the objects or events in a given unit; all the sites or times of a potential occurrence represent the large number of "trials" (large n). However, usually few of the potential sites or times actually have an occurrence, due to a very low probability for each "trial" (small p). For instance, a quadrat could be thought of as having very many possible sites where a plant could occur (especially if the plants are small relative to the quadrat), but in fact most sites are empty and there are only a small number of plants in the quadrat.

Parameters and formulae:

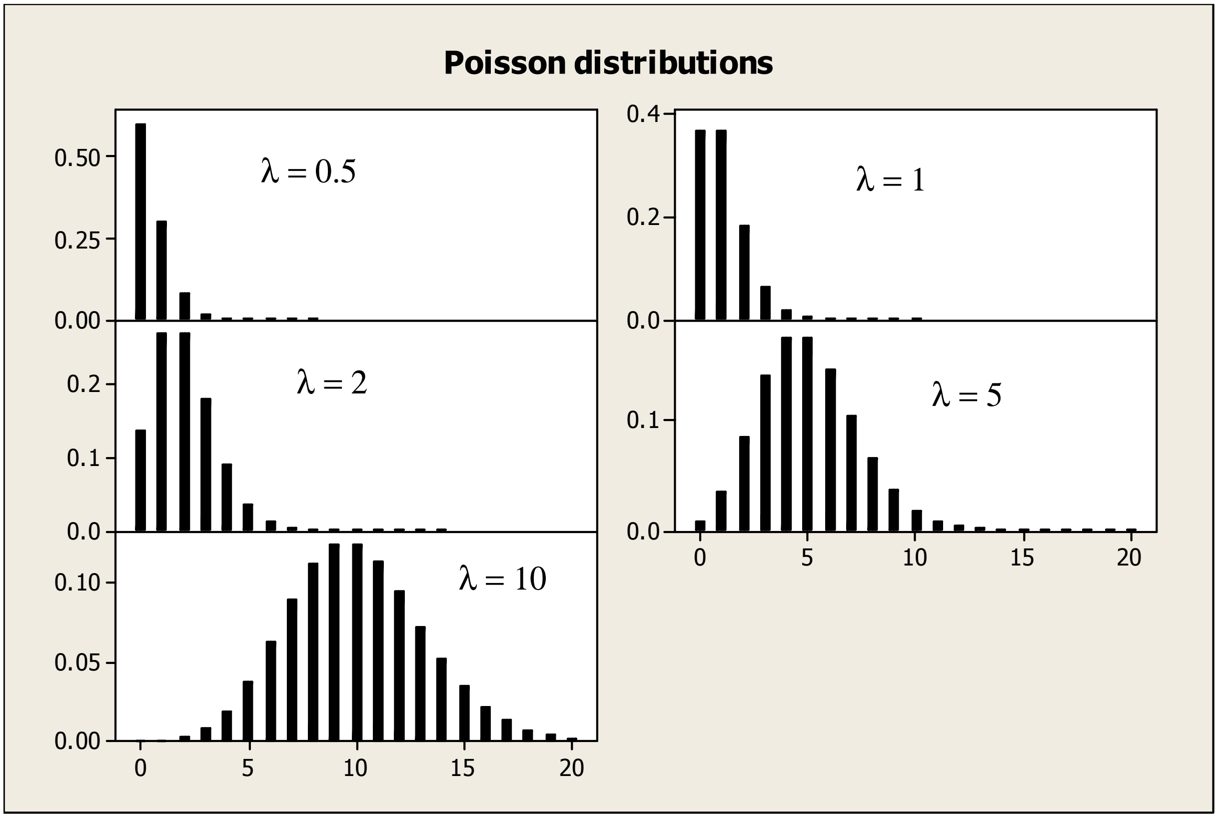
There is only one parameter, λ ("lambda").

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

To compute these probabilities, there is an iterative shortcut: Start with $P(X = 0) = e^{-\lambda}$. Then use the relationship $P(X = x) = \frac{\lambda}{x} \cdot P(X = x - 1)$ to calculate P(X=1), then P(X=2), etc.

Mean and variance:

$$\mu_x = \lambda \qquad \text{and} \qquad \sigma_X^2 = \lambda$$



Poisson distributions
λ = 0.5
λ = 1
λ = 2
λ = 5
λ = 10
0.50
0.25
0.00
0.4
0.2
0.0
0.2
0.1
0.0
0.1
0.0
0.10
0.05
0.00
0
5
10
15
20

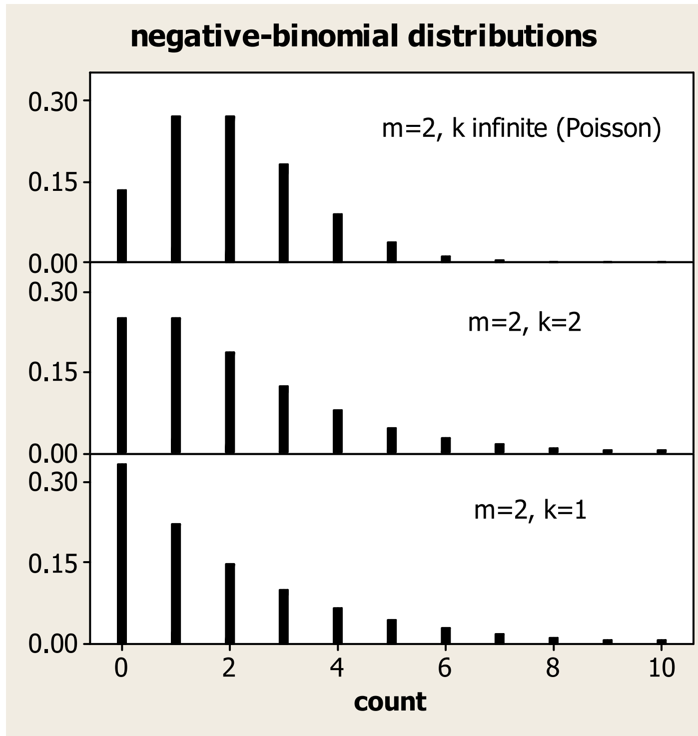
"Non-random" Distributions

Since the Poisson distribution assumes <u>independent</u> trials (*e.g.* whether a particular site within an quadrat is occupied is not affected by whether any other sites are occupied), it represents a "random" spatial or temporal distribution.

A "nonrandom" spatial or temporal distribution would be one in which "trials" are not independent; for instance, occupation of a site within a quadrat is affected by which other sites are occupied. There are two forms of such non-randomness.

<u>Aggregated distributions</u>

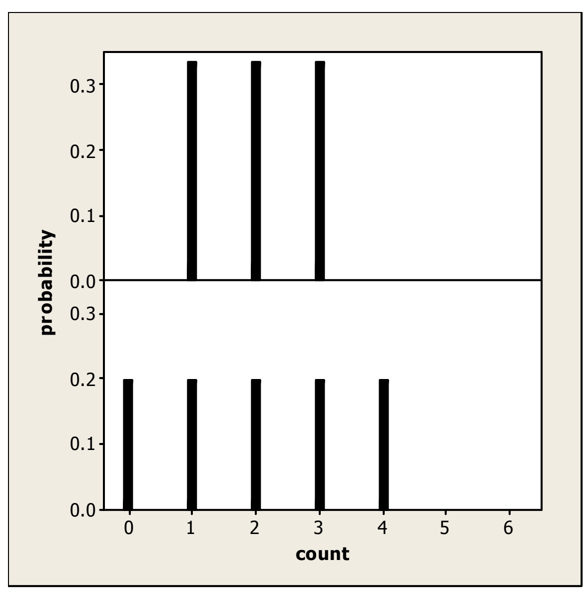
Occurrences ("successes" of trials) may tend to cluster, yielding "aggregated" or "contagious" distributions. In such distributions there are greater probabilities of both large values and low values, and those lower probabilities of intermediate values, compared to a Poisson distribution. For instance, some quadrats will be crowded, many will be empty or nearly so, and relative few will have moderate densities.



One commonly used aggregated distribution is the **negative-binomial**. As typically presented in ecology, this has two parameters: the mean m and a clumping parameter k. Smaller values of k indicate <u>greater</u> clumping; a negative binomial with $k = \infty$ is a Poisson distribution.

<u>Uniform distributions</u>

Alternatively, "successes" may be negatively related, *i.e.* spread as evenly as possible across units, representing repulsion among occurrences. One such distribution is the **discrete uniform** (or integer) distribution, in which all values within some range have equal probability.



Multinomial Distribution

Derivation:

The multinomial distribution generalizes the binomial distribution to the case in which there are more than two possible outcomes. It describes the number of occurrences, in some fixed number of independent trials, of each of several outcomes (X_i, i = 1 to k), each having some fixed probability of occurrence.

Applications:

Since many categorical variables have more than two categories, multinomial distributions are used commonly. Contingency table analyses often are based on a multinomial distribution.

Examples:

- numbers of barnacles dying from different causes, at a particular site
- distributions of phenotypes or genotypes, in a population sample or a set of offspring
- numbers of particular mutations (*i.e.* A to T, A to G, A to C, etc.)

Parameters and formula:

There are k categories of outcomes, each with probability p_i (i = 1 to k), and n independent trials. There then are k random variables X_i, and the joint probability that $X_1 = x_1$, $X_2 = x_2$, etc., is:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

Note that the p_i's must sum to 1, and the x_i's must sum to n.

Mean and variance:

Each random variable X_i has mean and variance as for the binomial distribution:

$$\mu_{x_i} = np_i \qquad \text{and} \qquad \sigma^2_{X_i} = np_i(1 - p_i)$$

Note that although the trials are independent, the X_i's are not, since they must sum to 1.

Hypergeometric Distribution

Derivation and Applications:

This distribution describes the count of occurrences of a particular outcome ("success") when sampling is without replacement and the sample is more than a small fraction of the population, so that trials cannot be considered independent and the binomial distribution is not appropriate.

Parameters and formula:

If the size of the total population is N, of which M are in the category of interest ("successes"), then the probability of getting x successes in a sample of size n is:

P(X=x) = { [the number of ways of choosing x successes out of a population of M successes]
x [the number of ways of choosing (n-x) "failures" out of a population of (N-M)] }
÷ [the number of ways of choosing n items out of a population of N]

$$P(X = x) = \frac{\left[\binom{M}{x} \cdot \binom{N-M}{n-x}\right]}{\binom{N}{n}}$$

$$= \frac{\left[\frac{M!}{x!(M-x)!}\right]\left[\frac{(N-M)!}{(n-x)!(N-M-n+x)!}\right]}{\left[\frac{N!}{n!(N-n)!}\right]}$$

(This formula is difficult to work with for any but small samples, since the factorials get both very awkward and very large numerically.)

Mean and variance:

If $p = M/N$ (*i.e.* the proportion of "successes" in the population), then

$$\mu_X = np \qquad \text{and} \qquad \sigma^2_X = np(1-p)(N-n)/(N-1)$$

Lognormal Distribution (NOTE: this is a continuous distribution)

Derivation:

The lognormal distribution is called this because if a random variable X has this distribution, its natural logarithm ($\ln X$) has a normal distribution. Or, to put it the other way around, if a random variable Y has a $N(\mu, \sigma^2)$ distribution, then $X = \exp(Y)$ has a lognormal (μ, σ^2) distribution.

Applications:

Lognormal distributions take only positive values, and are skewed to the right. They are generally useful for phenomena with these characteristics, and especially for multiplicative processes such as exponential growth, in which effects of random factors are additive on the log scale.

Examples:

- population abundances
- income or wealth of individuals or countries

Parameters and formula:

The best way to work with a lognormally distributed variable is to take is logarithm, getting a normally distribution variable. The formulae and tables for normal distributions then can be used.

Mean and variance:

If $Y = \ln(X)$ is $N(\mu, \sigma^2)$, then X (the lognormally distributed variable) has mean and variance

$$\mu_X = \exp(\mu + \sigma^2/2) \quad \text{and}$$

$$\sigma^2_X = [\exp(2\mu + \sigma^2)][\exp(\sigma^2) - 1] = (\mu_X)^2[\exp(\sigma^2) - 1]$$