Chapter 5

Discrete Random Variables

Random variables and their associated probability distributions are a basic component of statistical analyses. A statistician will examine the experiment or study and determine the type of observations or data it produces (continuous, discrete, or categorical) and then select a random variable and its distribution to model these data. We examine here three discrete random variables, the binomial, Poisson, and negative binomial, and their probability distributions. There are other discrete random variables but these three are the most commonly encountered in practice. These variables only take integer values and are typically used to model discrete or count data. We will also see how to calculate the mean and variance for a discrete random variable, using its probability distribution and a quantity called the **expected value**.

The basic concept of a **random variable** is to map the outcome of some random event into a number. For example, consider the dice cube example from Chapter 4. Define a number Y that is the number of spots showing on the dice – Y is a random variable. The sample space for Y would be $S = \{1, 2, 3, 4, 5, 6\}$ and the events any combination of these values. One requirement for Y to be a random variable is that events of the form $Y \leq y$ for any real number y are events in the probability space (Mood et al. 1974). For example, suppose that $y = 3.5$ for the dice cube example. The set defined by $Y \leq 3.5$ corresponds to the event $A = \{1, 2, 3\}$ and so is a member of the probability space for this example. This requirement is necessary in order to calculate probabilities for the random variable, and there is always a probability distribution associated with a particular random variable.

5.1 Binomial distribution

Binomial random variables are commonly used to model categorical observations or data that have two outcomes or states. For example, suppose we are sampling animals and classifying them into two age classes, say either adult (an event A) or juvenile (J). If we sample a single individual and classify it, the sample space would be $S = \{A, J\}$. We could then define a probability distribution such that $P[\{A\}] = p$ and $P[\{J\}] = 1 - p$, where p is the probability of observing an adult. Then, a random variable Y equal to the **number** of adults would be a binomial random variable. The random variable Y would have a sample space $S = \{0, 1\}$ corresponding to the number of adults. We could write the probability distribution for these two events as

$$P[Y = y] = p^y(1 - p)^{1-y}, \tag{5.1}$$

where $y = 0$ or 1. To see how this formula works, suppose we want the probability for $Y = 1$, so that $y = 1$. Inserting $y = 1$ in the above formula, we obtain

$$P[Y = 1] = p^1(1 - p)^{1-1} = p^1(1 - p)^0 = p. \tag{5.2}$$

To find the probability for $Y = 0$, we insert $y = 0$ in the formula to find

$$P[Y = 0] = p^0(1 - p)^{1-0} = p^0(1 - p)^1 = 1 - p. \tag{5.3}$$

Suppose that we now sample two animals and let Y again be the number of adults. The sample space for Y would now be $S = \{0, 1, 2\}$. What would be the probability distribution for this random variable? Assuming the two animals sampled are independent events, the probability of seeing two adults ($Y = 2$) in a row would be $p \times p = p^2$, while two juveniles ($Y = 0$) would be $(1 - p) \times (1 - p) = (1 - p)^2$. There are two ways of having one adult and one juvenile, a adult first and a juvenile second, or vice versa. The probability for each is $p \times (1 - p)$, so the probability of seeing one adult would be twice that, or $2p(1 - p)$. A general formula describing the probability distribution for this variable would be

$$P[Y = y] = \binom{2}{y} p^y(1 - p)^{2-y}. \tag{5.4}$$

where

$$\binom{2}{y} = \frac{2!}{y!(2 - y)!}. \tag{5.5}$$

The quantity $\binom{2}{y}$, known as a binomial coefficient, provides a way of calculating the number of ways y adults can occur among 2 sampled animals. It is often read as '2 choose y'. It makes use of factorials, which are defined for an integer j as the product $j \times (j-1) \times (j-2)... \times 1$. For example, $4! = 4 \times 3 \times 2 \times 1$. By convention, $0! = 1$.

To see how this distribution works, we will calculate the probability for different values of y. We have

$$\begin{aligned} P[Y=0] &= \binom{2}{0} p^0 (1-p)^{2-0} = \frac{2!}{0!(2-0)!}(1-p)^2 && (5.6) \\ &= \frac{2 \times 1}{1(2 \times 1)}(1-p)^2 && (5.7) \\ &= \frac{2}{2}(1-p)^2 = (1-p)^2 && (5.8) \end{aligned}$$

and

$$\begin{aligned} P[Y=1] &= \binom{2}{1} p^1 (1-p)^{2-1} = \frac{2!}{1!(2-1)!}p(1-p) && (5.9) \\ &= \frac{2 \times 1}{1(1)}p(1-p) && (5.10) \\ &= \frac{2}{1}p(1-p) = 2p(1-p). && (5.11) \end{aligned}$$

Finally, we have

$$\begin{aligned} P[Y=2] &= \binom{2}{2} p^2 (1-p)^{2-2} = \frac{2!}{2!(2-2)!}p^2 && (5.12) \\ &= \frac{2 \times 1}{(2 \times 1)1}p^2 && (5.13) \\ &= \frac{2}{2}p^2 = p^2. && (5.14) \end{aligned}$$

Do these probabilities sum to 1, satisfying this requirement for a probability distribution? We have $(1-p)^2 + 2p(1-p) + p^2 = (1-p)(1-p) + 2p - 2p^2 + p^2 = 1 - 2p + p^2 + 2p - 2p^2 + p^2 = 1$.

Suppose that we continue to sample l different animals, and let Y be the number of adults. The sample space for this binomial random variable would be $S = \{0, 1, 2, ..., l\}$. The probability distribution for this random variable is called the **binomial distribution**, and can be written using the formula

$$P[Y = y] = f(y) = \binom{l}{y} p^y (1-p)^{l-y} \tag{5.15}$$

where $y = 0, 1, 2, ..., l$ (Mood et al. 1974). The notation $f(y)$ is often used to denote a probability distribution, which is a function of y given the parameter values.

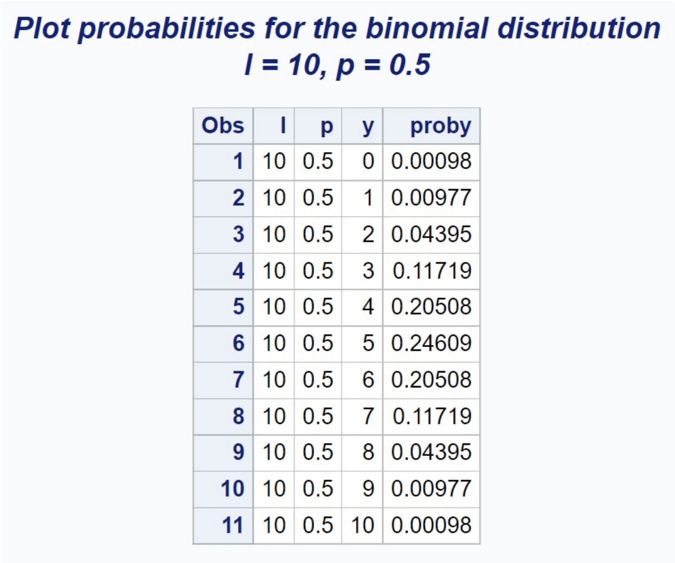
5.1.1 Binomial distribution - SAS demo

The SAS program below calculates and plots the binomial probabilities for different values of y using the SAS function `pdf`, given the values of the binomial parameters l and p. The probabilities are plotted for three different values of p, with $l = 10$. We see that for $p = 0.5$ the probability distribution has a peak at $y = 5$ (Fig. 5.2), indicating that five adults is the most likely outcome in 10 sampled animals. For $p = 0.25$ an adult occurs only 25% of the time, and so the probability distribution shifts to the left, with $y = 2$ having the highest probability (Fig. 5.3). For an adult almost certain, $p = 0.9$, then the probability distribution is shifted to the right with the peak at $y = 9$ (Fig. 5.4).

SAS Program

```
* binom_plot.sas;
title "Plot probabilities for the binomial distribution";
title2 "l = 10, p = 0.5";
data binom_plot;
    * Binomial parameters here;
    l = 10;
    p = 0.5;
    do y=0 to l;
        * Binomial distribution function;
        proby = pdf(’binomial’,y,p,l);
        * Output y and proby to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=binom_plot;
run;
* Plot probabilities;
```

```
proc gplot data=binom_plot;
    plot proby*y=1 / vref=0 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=dot c=red width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot probabilities for the binomial distribution
l = 10, p = 0.5

Obs	l	p	y	proby
1	10	0.5	0	0.00098
2	10	0.5	1	0.00977
3	10	0.5	2	0.04395
4	10	0.5	3	0.11719
5	10	0.5	4	0.20508
6	10	0.5	5	0.24609
7	10	0.5	6	0.20508
8	10	0.5	7	0.11719
9	10	0.5	8	0.04395
10	10	0.5	9	0.00977
11	10	0.5	10	0.00098

Figure 5.1: `binom_plot.sas - proc print`

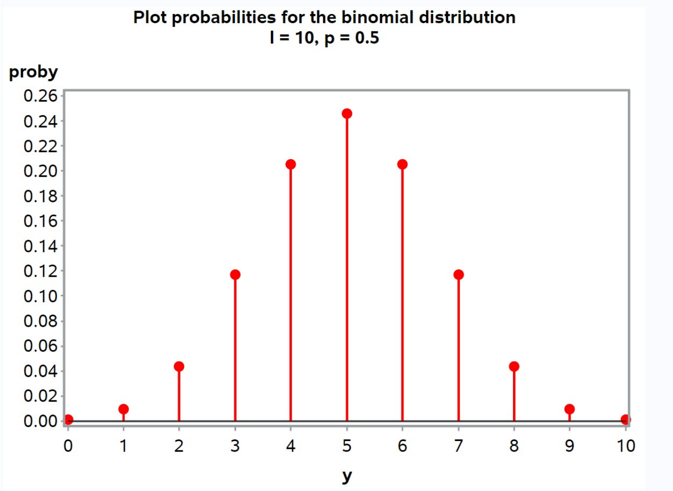


Figure 5.2: `binom_plot.sas - proc gplot`

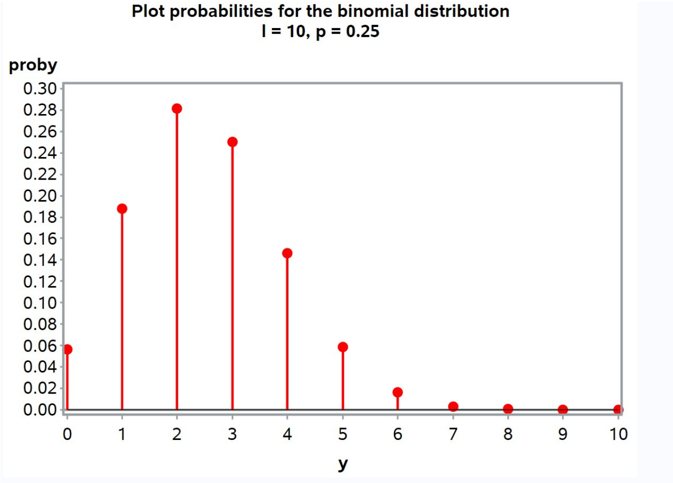


Figure 5.3: `binom_plot.sas - proc gplot`

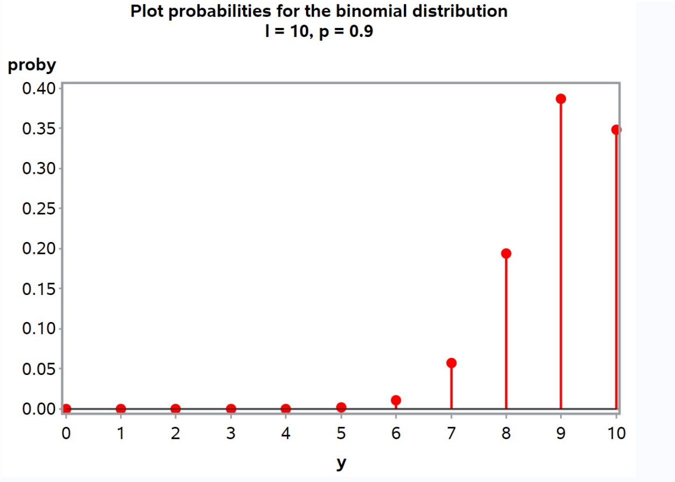


Figure 5.4: `binom_plot.sas - proc gplot`

5.2 Poisson distribution

Poisson random variables are commonly used to model counts of organisms or events in either space or time. For example, a Poisson random variable could be used to model the number of organisms in a sampling quadrat, or the number of flu infections per week in a city. The sample space for a Poisson random variable Y is $S = \{0, 1, 2, ..., \infty\}$, implying there is no upper limit on the counts. The Poisson distribution is given by the formula

$$P[Y = y] = f(y) = \frac{e^{-\lambda}\lambda^y}{y!} \tag{5.16}$$

where $y = 0, 1, 2, ..., \infty$. The parameter λ controls the shape of the distribution and is equal to the mean value of Y. For example, suppose the $\lambda = 2$. We have

$$P[Y = 0] = f(0) = \frac{e^{-2}2^0}{0!} = \frac{0.13534(1)}{1} = 0.13534, \tag{5.17}$$

$$P[Y = 1] = f(1) = \frac{e^{-2}2^1}{1!} = \frac{0.13534(2)}{1} = 0.27068, \tag{5.18}$$

$$P[Y = 2] = f(2) = \frac{e^{-2}2^2}{2!} = \frac{0.13534(4)}{2} = 0.27068, \tag{5.19}$$

$$P[Y = 3] = f(3) = \frac{e^{-2}2^3}{3!} = \frac{0.13534(8)}{6} = 0.18045, \tag{5.20}$$

$$P[Y = 4] = f(4) = \frac{e^{-2}2^4}{4!} = \frac{0.13534(16)}{24} = 0.09023 \tag{5.21}$$

and so forth.

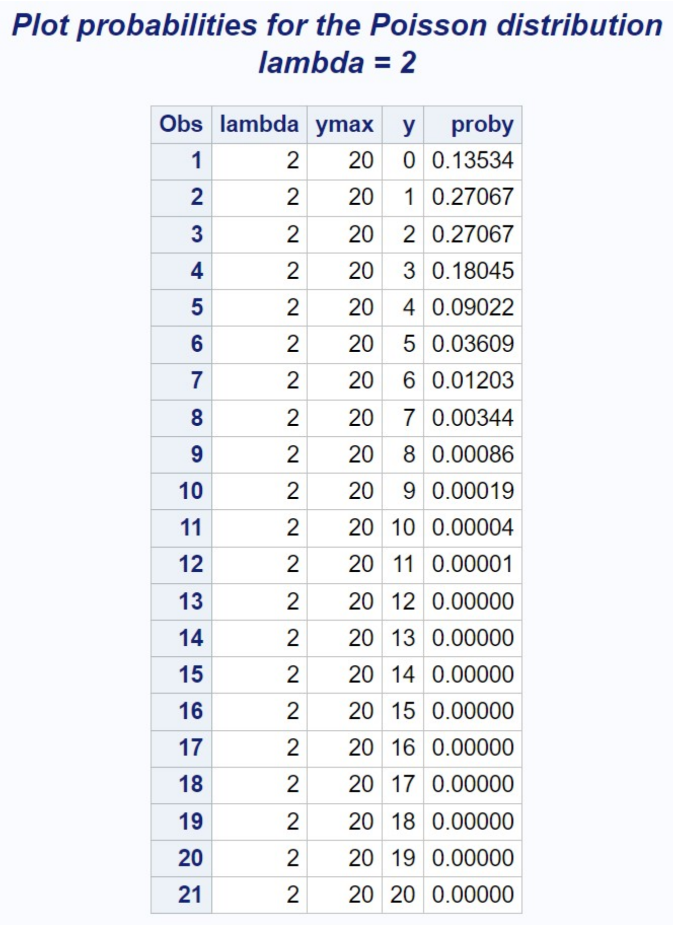
The Poisson distribution can arise in nature if certain assumptions hold true about the underlying process generating the data or observations (Mood et al. 1974, Snyder & Miller 1991). Suppose that we define an occurrence as a plant being present in a quadrat, or a case of disease occurring in a particular interval of time. **For the distribution of occurrences to be Poisson, we first need the probability of more than one occurrence to be small relative to the probability of exactly one occurrence, for a sufficiently small area of space (or short period of time).** In other words, two events are unlikely to occur in a small area or period of time. **Second, the number of occurrences in different areas of space (or time intervals) should be independent.** Another way of obtaining the Poisson distribution is as a limiting case of the binomial distribution. It can be shown that if lp is held constant (by making p small) while $l \to \infty$, the binomial distribution approaches a Poisson with $\lambda = lp$.

5.2.1 Poisson distribution - SAS demo

The following SAS program illustrates how the Poisson distribution varies for different values of λ. It is similar to the binomial distribution program, using the SAS function `pdf` to again find the probabilities (see below). We see that as λ increases, the Poisson distribution shifts to the right (Fig. 5.6, 5.7).

SAS Program

```
* Poisson_plot.sas;
title "Plot probabilities for the Poisson distribution";
title2 "lambda = 2";
data poisson_plot;
    * Poisson parameter here;
    lambda = 2;
    * Maximum value of y for plot;
    ymax = 20;
    do y=0 to ymax;
        * Poisson distribution function;
        proby = pdf('poisson',y,lambda);
        * Output y and proby to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=poisson_plot;
run;
* Plot probabilities;
proc gplot data=poisson_plot;
    plot proby*y=1 / vref=0 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=dot c=red width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot probabilities for the Poisson distribution
lambda = 2

Obs	lambda	ymax	y	proby
1	2	20	0	0.13534
2	2	20	1	0.27067
3	2	20	2	0.27067
4	2	20	3	0.18045
5	2	20	4	0.09022
6	2	20	5	0.03609
7	2	20	6	0.01203
8	2	20	7	0.00344
9	2	20	8	0.00086
10	2	20	9	0.00019
11	2	20	10	0.00004
12	2	20	11	0.00001
13	2	20	12	0.00000
14	2	20	13	0.00000
15	2	20	14	0.00000
16	2	20	15	0.00000
17	2	20	16	0.00000
18	2	20	17	0.00000
19	2	20	18	0.00000
20	2	20	19	0.00000
21	2	20	20	0.00000

etc.

Figure 5.5: `Poisson_plot.sas - proc print`

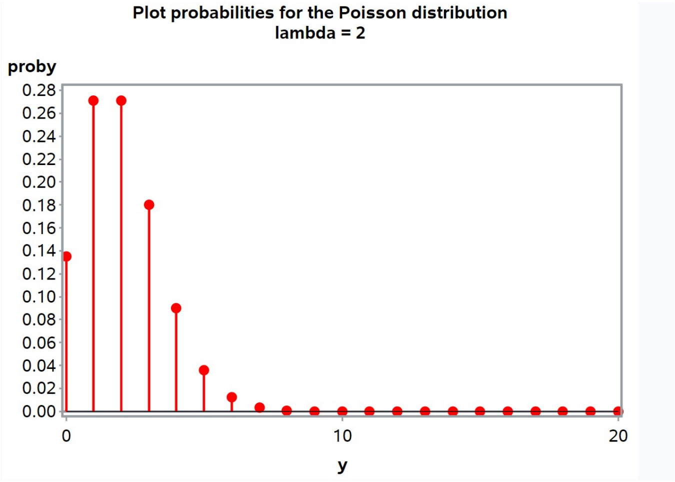


Figure 5.6: `Poisson_plot.sas - proc gplot`

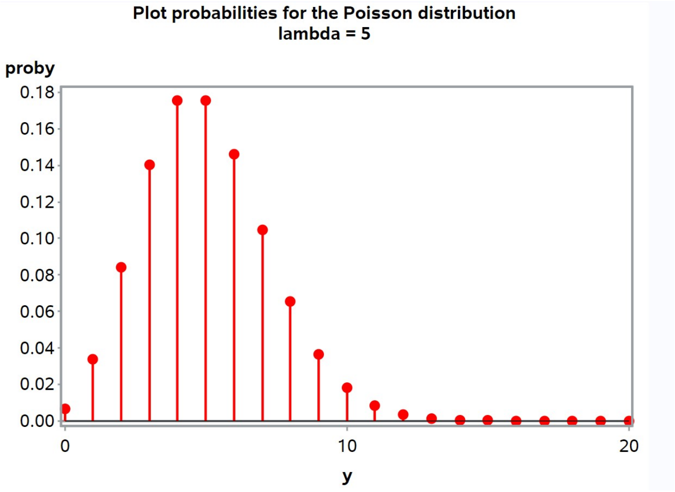


Figure 5.7: `Poisson_plot.sas - proc gplot`

5.3 Negative binomial distribution

Another useful tool for modeling count data is the negative binomial distribution. **It can be thought of as a mixture of Poisson distributions, each with a different value of λ.** For example, suppose that we are sampling insects in a forest across a number of locations. At the *ith* location the distribution of insects might be Poisson with parameter λ_i, but λ_i also differs among locations. Then the distribution of insects, considered across all locations, may have a negative binomial distribution. Because the density of most organisms typically varies in space, the negative binomial distribution often provides a better description of count data than the Poisson. The sample space for a negative binomial random variable Y is $S = \{0, 1, 2, ..., \infty\}$, the same as the Poisson. The probability distribution for the negative binomial is given by the formula

$$P[Y = y] = f(y) = \frac{\Gamma(k + y)}{\Gamma(y + 1)\Gamma(k)} \frac{(m/(k + m))^y}{(1 + m/k)^k} \tag{5.22}$$

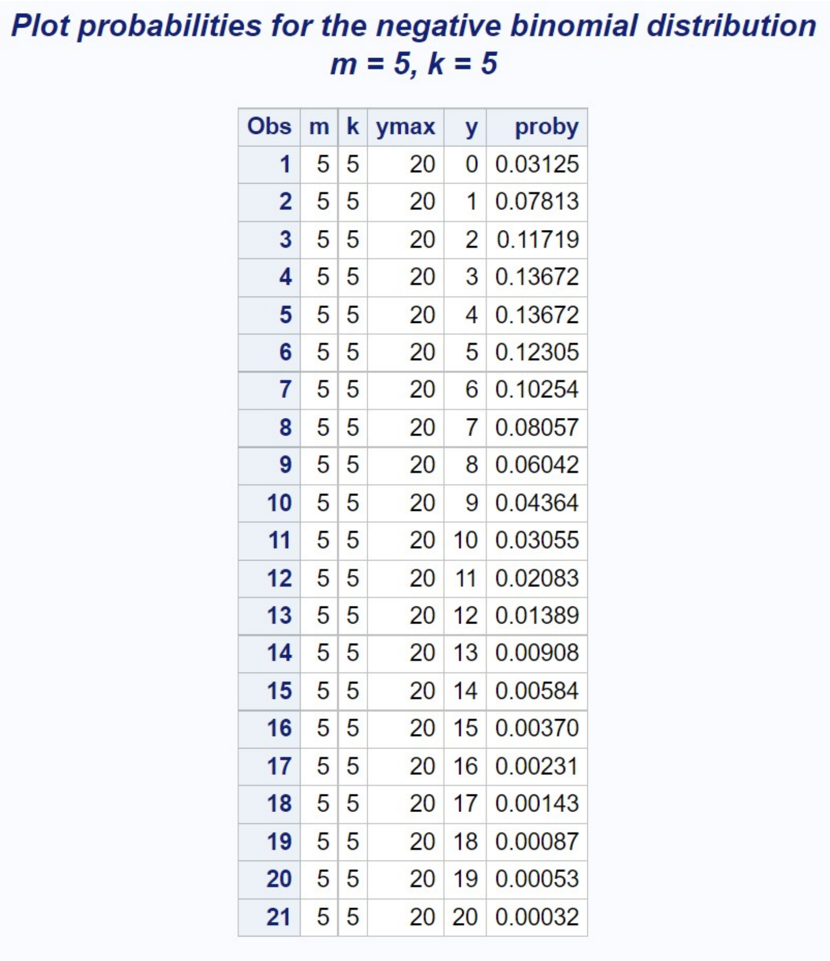
where $y = 0, 1, 2, ..., \infty$. The Γ symbol stands for the gamma function, which behaves like the factorial function but can be applied to non-integer quantities. The negative binomial distribution has two parameters, m and k, with m the mean of the distribution and k controlling its shape. For large values of k the negative binomial distribution approaches the Poisson distribution, while for small k the distribution becomes increasingly skewed to the right. See Bliss and Fisher (1953) for further information on this distribution.

5.3.1 Negative binomial distribution - SAS demo

The SAS program below shows how the shape of the negative binomial distribution varies with the parameter k. The program directly calculates the probabilities using the formula above, rather than the SAS `pdf` function, because we are using a different parameterization of the distribution. We see that distribution becomes more skewed to the right as k decreases (Fig. 5.9, 5.10).

SAS Program

```
* negbin_plot.sas;
title "Plot probabilities for the negative binomial distribution";
title2 "m = 5, k = 5";
data negbin_plot;
    * negative binomial parameters here;
    m = 5; k = 5;
    * Maximum value of y for plot;
    ymax = 20;
    do y=0 to ymax;
        * Negative binomial distribution function;
        proby = (gamma(k+y)/(gamma(y+1)*gamma(k)))*((m/(k+m))**y/(1+m/k)**k);
        * Output y and proby to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=negbin_plot;
run;
* Plot probabilities;
proc gplot data=negbin_plot;
    plot proby*y=1 / vref=0 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=dot c=red width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot probabilities for the negative binomial distribution
m = 5, k = 5

Obs	m	k	ymax	y	proby
1	5	5	20	0	0.03125
2	5	5	20	1	0.07813
3	5	5	20	2	0.11719
4	5	5	20	3	0.13672
5	5	5	20	4	0.13672
6	5	5	20	5	0.12305
7	5	5	20	6	0.10254
8	5	5	20	7	0.08057
9	5	5	20	8	0.06042
10	5	5	20	9	0.04364
11	5	5	20	10	0.03055
12	5	5	20	11	0.02083
13	5	5	20	12	0.01389
14	5	5	20	13	0.00908
15	5	5	20	14	0.00584
16	5	5	20	15	0.00370
17	5	5	20	16	0.00231
18	5	5	20	17	0.00143
19	5	5	20	18	0.00087
20	5	5	20	19	0.00053
21	5	5	20	20	0.00032

etc.

Figure 5.8: `negbin_plot.sas - proc print`

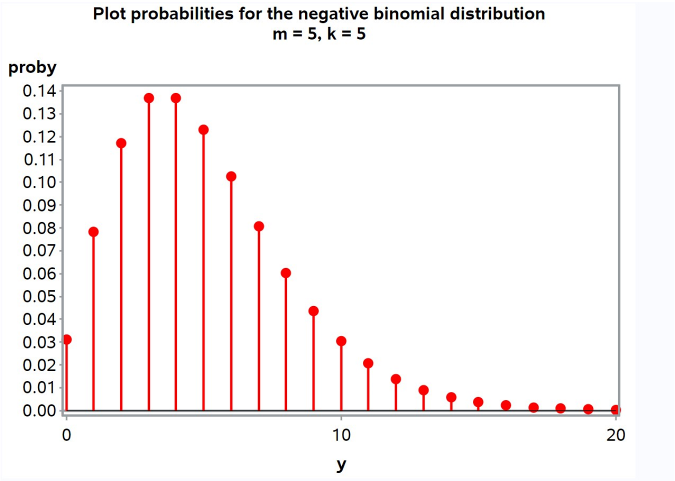


Figure 5.9: `negbin_plot.sas - proc gplot`

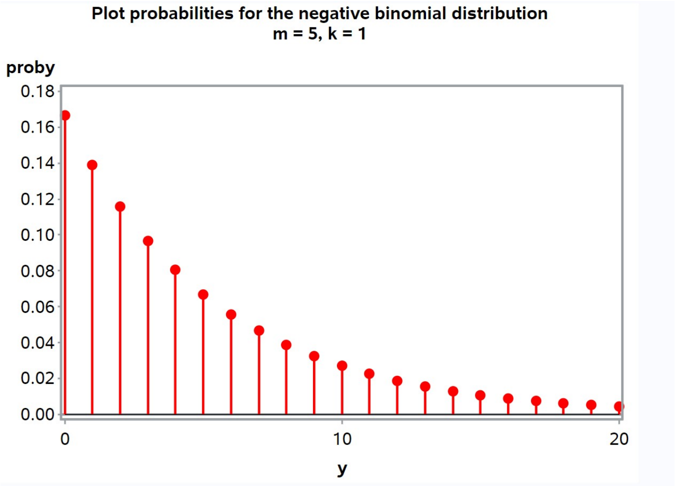


Figure 5.10: `negbin_plot.sas - proc gplot`

5.4 Expected values for discrete distributions

We have already seen how to calculate the mean, variance, and standard deviation for a set of observations (see Chapter 3). It is possible to calculate analogous quantities for probability distributions, such as the binomial, using the concept of an **expected value**.

Let Y be a random variable with some discrete probability distribution, such as the binomial, Poisson, or other distribution. The expected value or theoretical mean of Y, denoted by the expression $E[Y]$, is defined by the equation

$$E[Y] = \sum_y yP[Y = y] = \sum_y yf(y). \tag{5.23}$$

Here the summation is taken over all possible values of y for the probability distribution. **The expected value is a weighted average of each possible value of y, with the weights being the probability associated with each y.** It is a measure of the central location of the distribution of Y, in analogy to the sample mean $\bar{Y}$ for a data set. The expected value of Y can also be thought of as the sample mean $\bar{Y}$ of an infinitely large number of observations of Y.

For example, let Y have a binomial distribution with $l = 5$ and $p = 0.2$. We will first calculate some probabilities for the binomial distribution, then use them to calculate the expected value of Y, or $E[Y]$. We have

$$\begin{aligned}
P[Y = 0] = f(0) &= \binom{5}{0} 0.2^0 (1 - 0.2)^{5-0} && (5.24)\\
&= \frac{5!}{0!(5-0)!} 1(0.8^5) && (5.25)\\
&= \frac{120}{1(120)} 0.32768 && (5.26)\\
&= 0.32768. && (5.27)
\end{aligned}$$

$$
\begin{aligned}
P[Y=1]=f(1) &= \binom{5}{1}0.2^1(1-0.2)^{5-1} && (5.28)\\
&= \frac{5!}{1!(5-1)!}0.2(0.8^4) && (5.29)\\
&= \frac{120}{1(24)}0.08192 && (5.30)\\
&= 0.40960. && (5.31)
\end{aligned}
$$

$$
\begin{aligned}
P[Y=2]=f(2) &= \binom{5}{2}0.2^2(1-0.2)^{5-2} && (5.32)\\
&= \frac{5!}{2!(5-2)!}0.04(0.8^3) && (5.33)\\
&= \frac{120}{2(6)}0.02048 && (5.34)\\
&= 0.20480. && (5.35)
\end{aligned}
$$

$$
\begin{aligned}
P[Y=3]=f(3) &= \binom{5}{3}0.2^3(1-0.2)^{5-3} && (5.36)\\
&= \frac{5!}{2!(5-2)!}0.008(0.8^2) && (5.37)\\
&= \frac{120}{2(6)}0.00512 && (5.38)\\
&= 0.05120. && (5.39)
\end{aligned}
$$

$$
\begin{aligned}
P[Y=4]=f(4) &= \binom{5}{4}0.2^4(1-0.2)^{5-4} && (5.40)\\
&= \frac{5!}{4!(5-4)!}0.0016(0.8^1) && (5.41)\\
&= \frac{120}{24(1)}0.00128 && (5.42)\\
&= 0.00640. && (5.43)
\end{aligned}
$$

$$\begin{aligned}P[Y=5]=f(5) &= \binom{5}{5}0.2^5(1-0.2)^{5-5} && (5.44)\\ &= \frac{5!}{5!(5-5)!}0.00032(0.8^0) && (5.45)\\ &= \frac{120}{120(1)}0.00032 && (5.46)\\ &= 0.00032. && (5.47)\end{aligned}$$

These probabilities sum to 1, indicating our calculations are correct. Alternately, we could use the SAS program `binom_plot.sas` to find these probabilities.

We will now calculate $E[Y]$ using these probabilities and the formula for $E[Y]$ given above. We have

$$\begin{aligned}E[Y]=\sum_y yf(y) &= 0(0.32768)+1(0.40960)+2(0.20480) && (5.48)\\ &+3(0.05120)+4(0.00640)+5(0.00032) && (5.49)\\ &= 0+0.40960+0.40960 && (5.50)\\ &+0.15360+0.02560+0.00160 && (5.51)\\ &= 1.00000 && (5.52)\end{aligned}$$

So, $E[Y]=1$ for the binomial distribution with $l=5$ and $p=0.2$.

For the binomial distribution in general, it can be shown that

$$E[Y]=lp \tag{5.53}$$

for any value of l and p. Thus, the expected value or theoretical mean for the binomial distribution can be easily calculated given the parameters of this distribution. Plugging $l=5$ and $p=0.2$ into this equation, we obtain $E[Y]=5\times 0.2=1.0$, the same value as obtained using the expected value formula.

Other probability distributions would have a different formula for the expected value or theoretical mean, but the formula always involves the parameters of the distribution. For the Poisson distribution it can be shown that $E[Y]=\lambda$, while for the negative binomial distribution $E[Y]=m$.

5.4.1 Variance for discrete distributions

We can also define the theoretical variance for a random variable Y using expected values. This variance measures the dispersion of Y, and can also be thought of as the sample variance s^2 of an infinite number of observations. The variance of a discrete random variable Y, denoted by $Var[Y]$, is defined as

$$Var[Y] = E[(Y - E[Y])^2] = \sum_y (y - E[Y])^2 P[Y = y] \tag{5.54}$$

$$= \sum_y (y - E[Y])^2 f(y). \tag{5.55}$$

Note that this formula makes use of $E[Y]$, so it must be calculated first. As an example, let Y have the same binomial distribution as before, with $l = 5$ and $p = 0.2$, for which $E[Y] = 1$. Using the probabilities calculated above, we have

$$Var[Y] = \sum_y (y - E[Y])^2 f(y) \tag{5.56}$$

$$= (0-1)^2(0.32768) + (1-1)^2(0.40960) + (2-1)^2(0.20480) \tag{5.57}$$

$$+ (3-1)^2(0.05120) + (4-1)^2(0.00640) + (5-1)^2(0.00032) \tag{5.58}$$

$$= 1(0.32768) + 0(0.40960) + (1)0.20480 \tag{5.59}$$

$$+ 4(0.05120) + 9(0.00640) + (16)0.00032 \tag{5.60}$$

$$= 0.32768 + 0 + 0.20480 + 0.20480 + 0.05760 + 0.00512 \tag{5.61}$$

$$= 0.8. \tag{5.62}$$

For the binomial distribution, it can be mathematically shown that for any value of l and p, we have

$$Var[Y] = lp(1 - p). \tag{5.63}$$

Thus, the theoretical variance for the binomial distribution can also be calculated using the parameters of this distribution. Plugging $l = 5$ and $p = 0.2$ into this equation, we obtain $Var[Y] = 5(0.2)(1 - 0.2) = 0.8$, the same value as obtained using the variance formula.

Other probability distributions would have a different formula for the theoretical variance. For the Poisson distribution it can be shown that $Var[Y] = \lambda$. Because $E[Y] = \lambda$ for the Poisson, this implies the mean and variance of a Poisson random variable are equal. For the negative binomial distribution, $Var[Y] = m + m^2/k$, while $E[Y] = m$. This implies the variance of the negative binomial is always greater than its mean. The theoretical standard deviation is simply $\sqrt{Var[Y]}$.

5.5 Discrete random variables and samples

Discrete random variables like the binomial and Poisson are used to model real observations that are counts. But how well do these mathematical quantities match the behavior of the observations? We will now develop a graphical method of comparing the observed data with the pattern expected for discrete random variables, in particular the Poisson and negative binomial distributions. There are also statistical procedures called goodness-of-fit tests that are used for this purpose, but we defer this to Chapter 20.

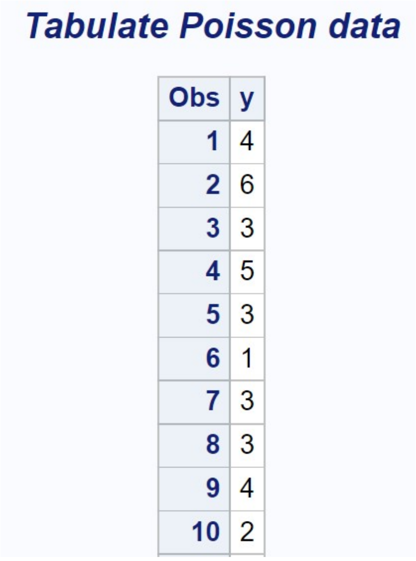
5.5.1 Parasitic wasps - SAS demo

Small insects are often sampled using sticky-traps, which are small cards covered with a substance called Tanglefoot®(The Tanglefoot Company, Grand Rapids, MI). For example, Reeve & Cronin (2010) used this method to sample populations of the parasitic wasp *Anagrus columbi*, which attacks eggs of the planthopper *Prokelisia crocea*. Suppose $n = 100$ traps are deployed for some period of time, then the traps collected and the wasps counted. If individual wasps are randomly and independently distributed across the field, we would expect the number of wasps per trap to have a Poisson distribution. We can then compare the observed distribution with the expected one for the Poisson distribution, to see if they resemble one another. If so, we can say the Poisson distribution provides an adequate description of these observations.

The first step in this procedure is simply to tabulate the number of traps with $0, 1, 2, 3, ...$ wasps, which is the observed frequency distribution. We can use `proc freq` in SAS to accomplish this task as in the following program. The numbers listed as data here are the number of wasps for each of the $n = 100$ sticky-traps. The statement `tables y` tells `proc freq` to count the number of observations for each value of `y` in the data set. The output generated is a table of these frequencies.

SAS Program

```
* poisson_freq.sas;
title 'Tabulate Poisson data';
data poisson;
    input y @@;
    datalines;
4 6 3 5 3 1 3 3 4 2
4 0 2 3 1 3 4 6 5 1
3 3 4 3 2 3 7 4 3 3
4 3 4 3 4 0 3 0 3 3
4 8 2 2 4 2 5 3 3 2
1 4 1 1 5 2 4 1 2 6
3 3 3 1 1 2 1 5 3 5
3 2 4 3 4 1 2 3 1 3
4 4 4 6 6 2 0 1 4 2
2 2 3 4 3 0 1 1 0 2
;
run;
* Print observations;
proc print data=poisson;
run;
* Tabulate data into frequencies;
proc freq data=poisson;
    tables y;
run;
quit;
```

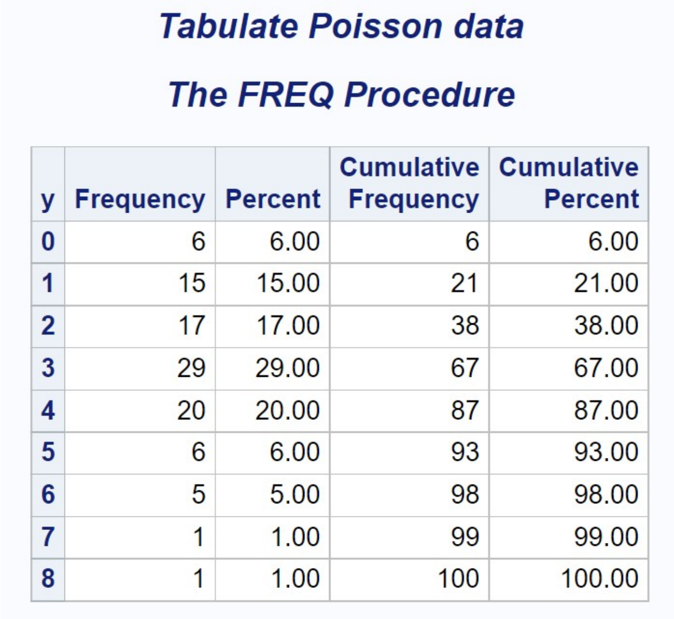


Tabulate Poisson data

Obs	y
1	4
2	6
3	3
4	5
5	3
6	1
7	3
8	3
9	4
10	2

etc.

Figure 5.11: `Poisson_freq.sas - proc print`



Tabulate Poisson data

The FREQ Procedure

y	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6	6.00	6	6.00
1	15	15.00	21	21.00
2	17	17.00	38	38.00
3	29	29.00	67	67.00
4	20	20.00	87	87.00
5	6	6.00	93	93.00
6	5	5.00	98	98.00
7	1	1.00	99	99.00
8	1	1.00	100	100.00

Figure 5.12: `Poisson_freq.sas - proc freq`

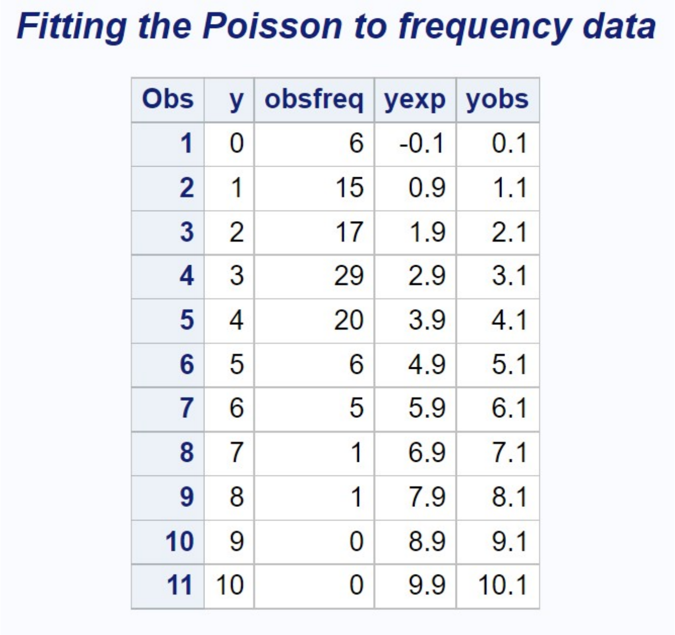
We now want to compare these observed frequencies with those expected for the Poisson distribution. We first need to estimate the Poisson parameter λ from the observed data using $\bar{Y}$ (see Chapter 8 for a justification). We then calculate the Poisson probabilities for $\lambda = \bar{Y}$, obtaining $P[Y = y]$ for values of y that spans or better exceeds the range of y values in the data set. Because $P[Y = y]$ is the probability or proportion of observations that take the value y, the expected frequency with n observations is therefore equal to $n \times P[Y = y]$. We can then compare the observed frequencies with the expected ones generated using the Poisson distribution. These calculations can be automated using the SAS program listed below. The program first uses `proc univariate` to find n, $\bar{Y}$, and the sample variance s^2 for the observed frequencies. We let `proc univariate` know that the data are in the form of frequencies (the variable `obsfreq`), rather than individual observations, by adding the command `freq obsfreq`.

The program then passes these results to a `data` step where the Poisson probabilities and expected frequencies are calculated, which are then plotted across a range of y values using `proc gplot`. See SAS output and graph below. We first see that sample mean and variance are similar in magnitude ($\bar{Y} = 2.910$ vs. $s^2 = 2.628$), suggesting these data are close to Poisson (recall that $E[Y] = Var[Y] = \lambda$ for this distribution). In addition, the observed and expected frequencies are quite similar, again implying an adequate fit by the Poisson distribution. There are some small differences in the observed and expected frequencies, which is to be expected because the observed ones are random quantities.

SAS Program

```
* Poisson_fit.sas;
title ’Fitting the Poisson to frequency data’;
data poisson;
    input y obsfreq;
    * Generate offset y values for plot;
    yexp = y - 0.1; yobs = y + 0.1;
    datalines;
0   6
1  15
2  17
3  29
4  20
5   6
6   5
```

```
7   1
8   1
9   0
10  0
;
run;
* Print data set;
proc print data=poisson;
run;
* Descriptive statistics, save ybar, n, and var to data file;
proc univariate data=poisson;
    var y;
    histogram y / vscale=count;
    freq obsfreq;
    output out=stats mean=ybar n=n var=var;
run;
* Print output data file;
proc print data=stats;
run;
* Calculate expected frequencies using ybar;
data poisfit;
    if _n_ = 1 then set stats;
    set poisson;
    poisprob = pdf('poisson',y,ybar);
    expfreq = n*poisprob;
run;
* Print observed and expected frequencies;
proc print data=poisfit;
run;
* Plot observed and expected frequencies;
proc gplot data=poisfit;
    plot expfreq*yexp=1 obsfreq*yobs=2 / overlay legend=legend1 vref=0 wvref=3
    vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=circle c=red width=3 height=2;
    symbol2 i=needle v=square c=blue width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```

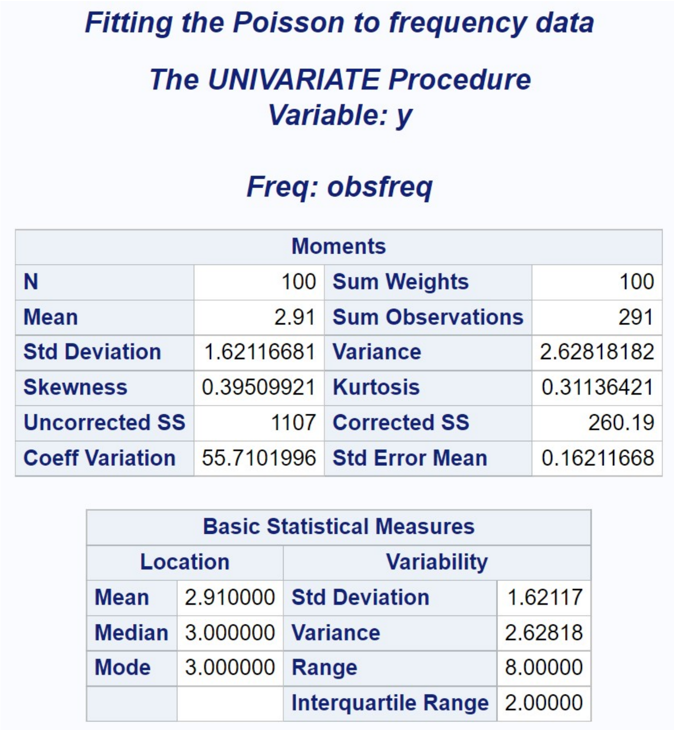


Fitting the Poisson to frequency data

Obs	y	obsfreq	yexp	yobs
1	0	6	-0.1	0.1
2	1	15	0.9	1.1
3	2	17	1.9	2.1
4	3	29	2.9	3.1
5	4	20	3.9	4.1
6	5	6	4.9	5.1
7	6	5	5.9	6.1
8	7	1	6.9	7.1
9	8	1	7.9	8.1
10	9	0	8.9	9.1
11	10	0	9.9	10.1

etc.

Figure 5.13: `Poisson_fit.sas - proc print`



Fitting the Poisson to frequency data

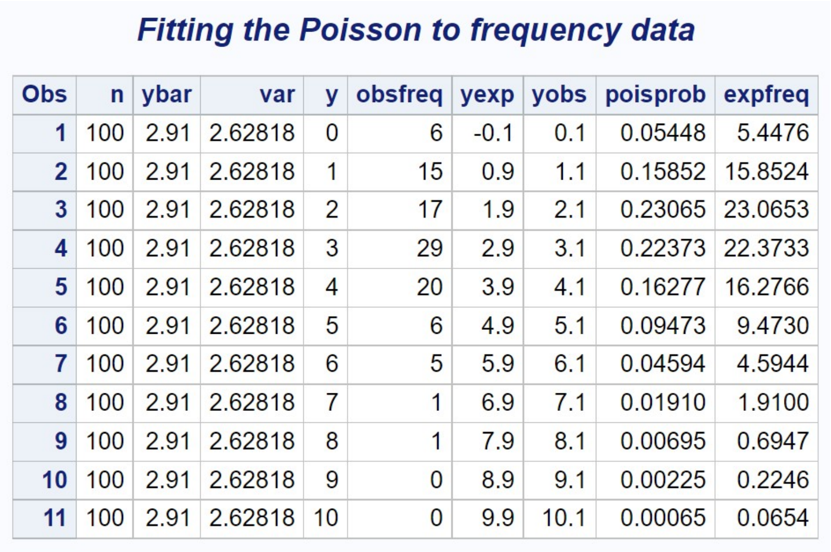
The UNIVARIATE Procedure
Variable: y

Freq: obsfreq

Moments			
N	100	Sum Weights	100
Mean	2.91	Sum Observations	291
Std Deviation	1.62116681	Variance	2.62818182
Skewness	0.39509921	Kurtosis	0.31136421
Uncorrected SS	1107	Corrected SS	260.19
Coeff Variation	55.7101996	Std Error Mean	0.16211668

Basic Statistical Measures			
Location		Variability	
Mean	2.910000	Std Deviation	1.62117
Median	3.000000	Variance	2.62818
Mode	3.000000	Range	8.00000
		Interquartile Range	2.00000

Figure 5.14: `Poisson_fit.sas - proc univariate`



Fitting the Poisson to frequency data

Obs	n	ybar	var	y	obsfreq	yexp	yobs	poisprob	expfreq
1	100	2.91	2.62818	0	6	-0.1	0.1	0.05448	5.4476
2	100	2.91	2.62818	1	15	0.9	1.1	0.15852	15.8524
3	100	2.91	2.62818	2	17	1.9	2.1	0.23065	23.0653
4	100	2.91	2.62818	3	29	2.9	3.1	0.22373	22.3733
5	100	2.91	2.62818	4	20	3.9	4.1	0.16277	16.2766
6	100	2.91	2.62818	5	6	4.9	5.1	0.09473	9.4730
7	100	2.91	2.62818	6	5	5.9	6.1	0.04594	4.5944
8	100	2.91	2.62818	7	1	6.9	7.1	0.01910	1.9100
9	100	2.91	2.62818	8	1	7.9	8.1	0.00695	0.6947
10	100	2.91	2.62818	9	0	8.9	9.1	0.00225	0.2246
11	100	2.91	2.62818	10	0	9.9	10.1	0.00065	0.0654

Figure 5.15: `Poisson_fit.sas - proc print`

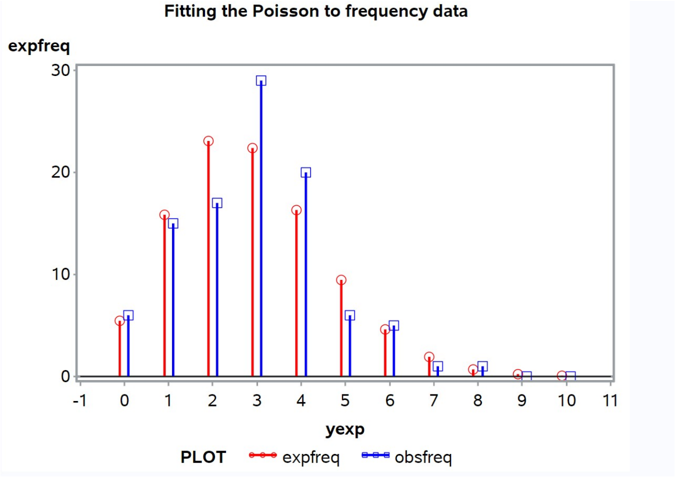


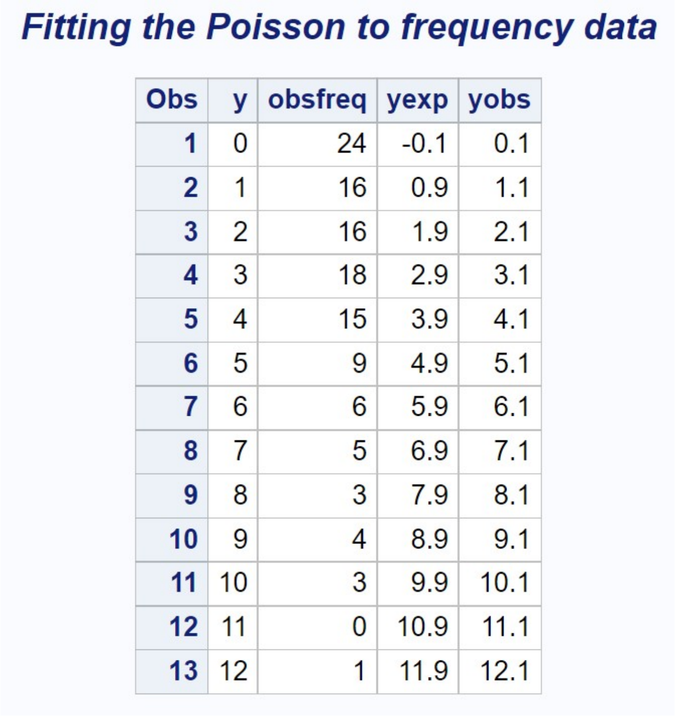
Figure 5.16: `Poissonfit.sas - proc gplot`

5.5.2 Corn borers - SAS demo

We now examine the spatial distribution of an insect pest, the European corn borer *Ostrinia nubilalis*, as reported by Bliss and Fisher (1953). The number of borers was recorded for 120 hills in which corn was planted. These data are already tabulated and can be directly inserted in the SAS program `poisson_fit2.sas` (see below). For this example, we see that the Poisson distribution provides a relatively poor fit (see Fig. 5.20) - there are more zeroes ($y = 0$) and large values ($y \geq 7$) in the observed frequencies than predicted by the Poisson. We also note that the sample variance $s^2 = 7.770$ is considerably larger than the mean $\bar{Y} = 3.167$, while for the Poisson these two quantities should be equal. This finding also suggests that these data are not Poisson in distribution.

SAS Program

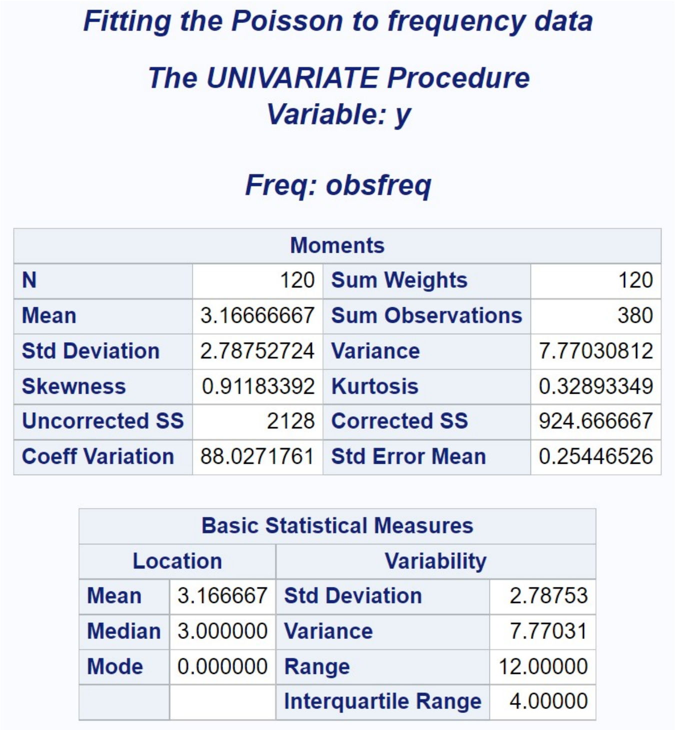
```
* Poisson_fit2.sas;
title 'Fitting the Poisson to frequency data';
data poisson;
    input y obsfreq;
    * Generate offset y values for plot;
    yexp = y - 0.1; yobs = y + 0.1;
    datalines;
0  24
1  16
2  16
3  18
4  15
5   9
6   6
7   5
8   3
9   4
10  3
11  0
12  1
;
run;
* Print data set;
proc print data=poisson;
run;
* Descriptive statistics, save ybar, n, and var to data file;
proc univariate data=poisson;
    var y;
    histogram y / vscale=count;
    freq obsfreq;
    output out=stats mean=ybar n=n var=var;
run;
* Print output data file;
proc print data=stats;
run;
* Calculate expected frequencies using ybar;
data poisfit;
    if _n_ = 1 then set stats;
    set poisson;
    poisprob = pdf('poisson',y,ybar);
    expfreq = n*poisprob;
run;
* Print observed and expected frequencies;
proc print data=poisfit;
run;
* Plot observed and expected frequencies;
proc gplot data=poisfit;
    plot expfreq*yexp=1 obsfreq*yobs=2 / overlay legend=legend1 vref=0 wvref=3
    vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=circle c=red width=3 height=2;
    symbol2 i=needle v=square c=blue width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```



Fitting the Poisson to frequency data

Obs	y	obsfreq	yexp	yobs
1	0	24	-0.1	0.1
2	1	16	0.9	1.1
3	2	16	1.9	2.1
4	3	18	2.9	3.1
5	4	15	3.9	4.1
6	5	9	4.9	5.1
7	6	6	5.9	6.1
8	7	5	6.9	7.1
9	8	3	7.9	8.1
10	9	4	8.9	9.1
11	10	3	9.9	10.1
12	11	0	10.9	11.1
13	12	1	11.9	12.1

Figure 5.17: `Poisson_fit2.sas - proc print`



Fitting the Poisson to frequency data

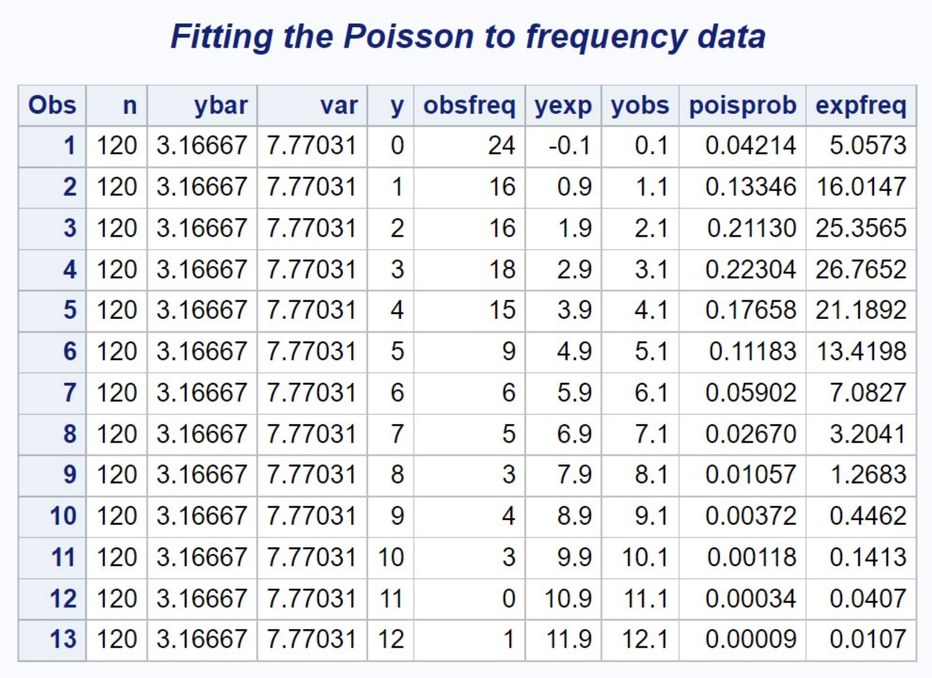
The UNIVARIATE Procedure
Variable: y

Freq: obsfreq

Moments			
N	120	Sum Weights	120
Mean	3.16666667	Sum Observations	380
Std Deviation	2.78752724	Variance	7.77030812
Skewness	0.91183392	Kurtosis	0.32893349
Uncorrected SS	2128	Corrected SS	924.666667
Coeff Variation	88.0271761	Std Error Mean	0.25446526

Basic Statistical Measures			
Location		Variability	
Mean	3.166667	Std Deviation	2.78753
Median	3.000000	Variance	7.77031
Mode	0.000000	Range	12.00000
		Interquartile Range	4.00000

Figure 5.18: `Poisson_fit2.sas - proc univariate`



Fitting the Poisson to frequency data

Obs	n	ybar	var	y	obsfreq	yexp	yobs	poisprob	expfreq
1	120	3.16667	7.77031	0	24	-0.1	0.1	0.04214	5.0573
2	120	3.16667	7.77031	1	16	0.9	1.1	0.13346	16.0147
3	120	3.16667	7.77031	2	16	1.9	2.1	0.21130	25.3565
4	120	3.16667	7.77031	3	18	2.9	3.1	0.22304	26.7652
5	120	3.16667	7.77031	4	15	3.9	4.1	0.17658	21.1892
6	120	3.16667	7.77031	5	9	4.9	5.1	0.11183	13.4198
7	120	3.16667	7.77031	6	6	5.9	6.1	0.05902	7.0827
8	120	3.16667	7.77031	7	5	6.9	7.1	0.02670	3.2041
9	120	3.16667	7.77031	8	3	7.9	8.1	0.01057	1.2683
10	120	3.16667	7.77031	9	4	8.9	9.1	0.00372	0.4462
11	120	3.16667	7.77031	10	3	9.9	10.1	0.00118	0.1413
12	120	3.16667	7.77031	11	0	10.9	11.1	0.00034	0.0407
13	120	3.16667	7.77031	12	1	11.9	12.1	0.00009	0.0107

Figure 5.19: `Poisson_fit2.sas - proc print`

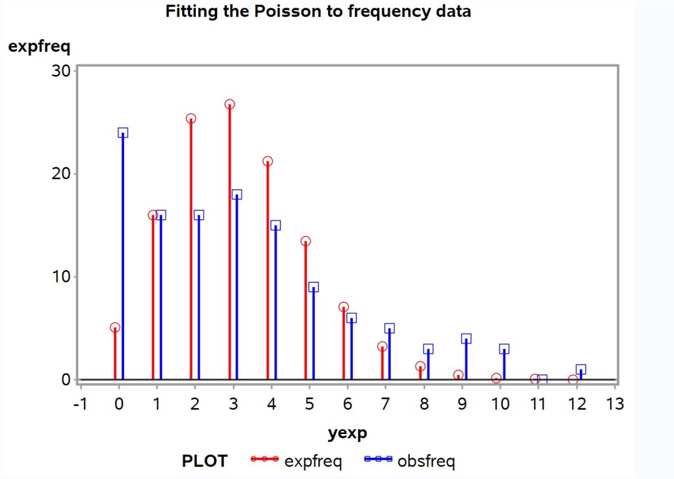


Figure 5.20: `Poissonfit2.sas - proc gplot`

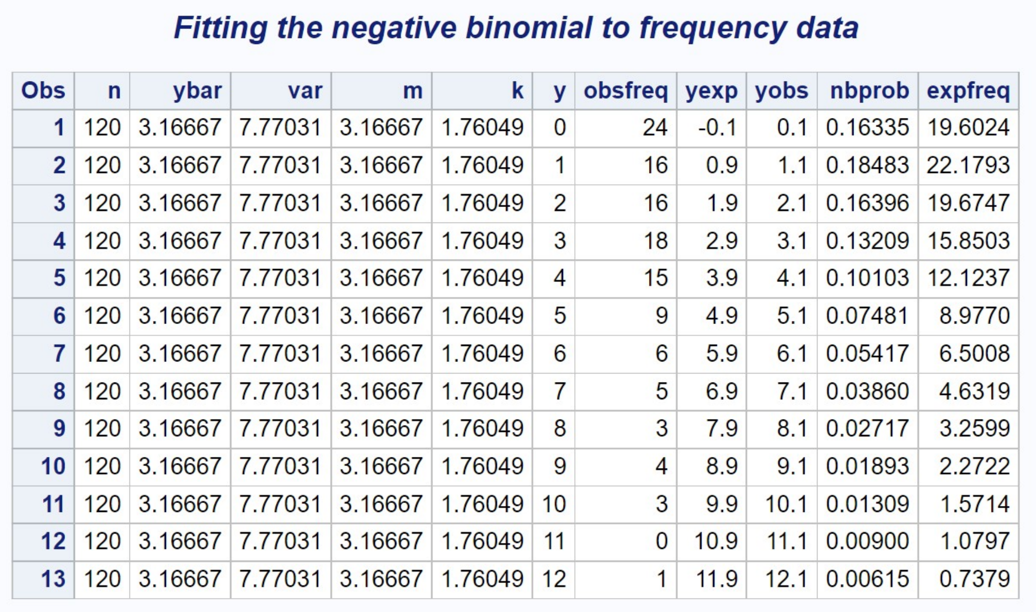
As an alternative to the Poisson, we can try fitting the negative binomial distribution using a similar SAS program. This distribution has two parameters, m and k, that must also be estimated before we can fit the distribution. The parameter m can be estimated using $\bar{Y}$ as with the Poisson, but k is best estimated using a technique called maximum likelihood (see Chapter 8). We will use a SAS procedure that can model count data using the negative binomial distribution, `proc genmod`, in order to estimate k (SAS Institute Inc. 2018). The output of `proc genmod` is manipulated in several `data` steps to combine these estimates with the observed frequency data, and then the negative binomial probabilities and expected frequencies calculated and plotted. See SAS program and output below.

We see that the expected frequencies for the negative binomial distribution provide a better match to the observed ones for this data set (Fig. 5.22). We also note that the variance predicted for the negative binomial distribution is close to the observed variance. From the negative binomial fit, we have $m = 3.167$ and $k = 1.760$, and so the estimated variance is $m + m^2/k = 3.167 + 3.167^2/1.760 = 7.459$, while the observed variance is $s^2 = 7.770$. This further implies the negative binomial provides a better fit to these data than the Poisson distribution.

SAS Program

```
* negbin_fit2.sas;
title 'Fitting the negative binomial to frequency data';
data negbin;
    input y obsfreq;
    * Generate offset y values for plot;
    yexp = y - 0.1; yobs = y + 0.1;
    datalines;
0  24
1  16
2  16
3  18
4  15
5   9
6   6
7   5
8   3
9   4
10  3
11  0
12  1
;
run;
* Print data set;
proc print data=negbin;
run;
* Descriptive statistics, save ybar, n, and var to data file;
proc univariate data=negbin;
    var y;
    histogram y / vscale=count;
    freq obsfreq;
    output out=stats mean=ybar n=n var=var;
run;
* Print output data file;
proc print data=stats;
run;
* Estimate m and k for the negative binomial distribution;
proc genmod data=negbin;
    model y = / dist=negbin;
    freq obsfreq;
    ods output ParameterEstimates=params;
run;
* Pick out value of m from genmod output;
data m;
```

```
    set params;
    if _n_ = 1;
    m = exp(Estimate);
    keep m;
run;
* Pick out value of k from genmod output;
data k;
    set params;
    if _n_ = 2;
    k = 1/Estimate;
    keep k;
run;
* Put m and k in one data file;
data params;
    merge m k;
run;
* Calculate expected frequencies using m and k;
data nbfit;
    if _n_ = 1 then set stats;
    if _n_ = 1 then set params;
    set negbin;
    nbprob = (gamma(k+y)/(gamma(y+1)*gamma(k)))*((m/(k+m))**y/(1+m/k)**k);
    expfreq = n*nbprob;
run;
* Print observed and expected frequencies;
proc print data=nbfit;
run;
* Plot observed and expected frequencies;
proc gplot data=nbfit;
    plot expfreq*yexp=1 obsfreq*yobs=2 / overlay legend=legend1 vref=0 wvref=3
    vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=circle c=red width=3 height=2;
    symbol2 i=needle v=square c=blue width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```



Fitting the negative binomial to frequency data

Obs	n	ybar	var	m	k	y	obsfreq	yexp	yobs	nbprob	expfreq
1	120	3.16667	7.77031	3.16667	1.76049	0	24	-0.1	0.1	0.16335	19.6024
2	120	3.16667	7.77031	3.16667	1.76049	1	16	0.9	1.1	0.18483	22.1793
3	120	3.16667	7.77031	3.16667	1.76049	2	16	1.9	2.1	0.16396	19.6747
4	120	3.16667	7.77031	3.16667	1.76049	3	18	2.9	3.1	0.13209	15.8503
5	120	3.16667	7.77031	3.16667	1.76049	4	15	3.9	4.1	0.10103	12.1237
6	120	3.16667	7.77031	3.16667	1.76049	5	9	4.9	5.1	0.07481	8.9770
7	120	3.16667	7.77031	3.16667	1.76049	6	6	5.9	6.1	0.05417	6.5008
8	120	3.16667	7.77031	3.16667	1.76049	7	5	6.9	7.1	0.03860	4.6319
9	120	3.16667	7.77031	3.16667	1.76049	8	3	7.9	8.1	0.02717	3.2599
10	120	3.16667	7.77031	3.16667	1.76049	9	4	8.9	9.1	0.01893	2.2722
11	120	3.16667	7.77031	3.16667	1.76049	10	3	9.9	10.1	0.01309	1.5714
12	120	3.16667	7.77031	3.16667	1.76049	11	0	10.9	11.1	0.00900	1.0797
13	120	3.16667	7.77031	3.16667	1.76049	12	1	11.9	12.1	0.00615	0.7379

Figure 5.21: `negbin_fit2.sas - proc print`

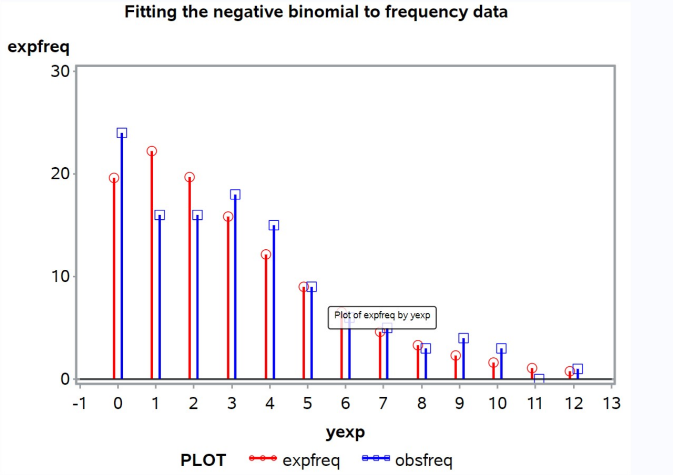


Figure 5.22: `negbin_fit2.sas - proc gplot`

5.6 Classifying spatial or temporal patterns

The spatial distribution of organisms, or the temporal occurrence of events like cases of disease, is often compared with the Poisson distribution. This distribution essentially assumes a random, independent distribution of organisms or events, and if the observed distribution differs from the Poisson then this could indicate some interesting biology. For example, if the observed frequencies have a distribution with more extreme values (low or high) than the Poisson, with $s^2 > \bar{Y}$, this implies organisms are unevenly distributed in space, or events in time. A pattern like this is often called an **overdispersed** distribution, or alternatively a clumped, aggregated, or contagious distribution (Pielou 1977, Begon et al. 2006). One method of quantifying the level of overdispersion is to fit the negative binomial distribution to the data and use the value of k as an index. Small values of k (say $k < 5$) imply an overdispersed distribution, while larger ones indicate a distribution close to Poisson. More rarely, an observed distribution may have fewer extreme values than the Poisson, with $s^2 < \bar{Y}$, implying the organisms are evenly distributed in space (or events in time). This is called an **underdispersed distribution**, also known as a regular, even, or repulsed distribution.

The figures below provide examples of spatial distributions that are overdispersed, Poisson, or underdispersed. Note the obvious clusters of organisms in the overdispersed example (Fig. 5.23). This might occur because the clusters are offspring from a single parent, the organisms are responding to resources that are clumped in space, or because the organisms are attracted to one another. The Poisson data also show a few clusters (Fig. 5.24), but these are chance occurrences. If we were to divide this graph into quadrats and count the number of organisms per quadrat, we would find the frequency distribution is close to Poisson. In contrast to the other examples, the organisms are spaced apart to some extent in the underdispersed example (Fig. 5.25). This could occur because they are territorial, compete for resources, or otherwise regulate their numbers in some fashion (Ridout & Besbeas 2004).

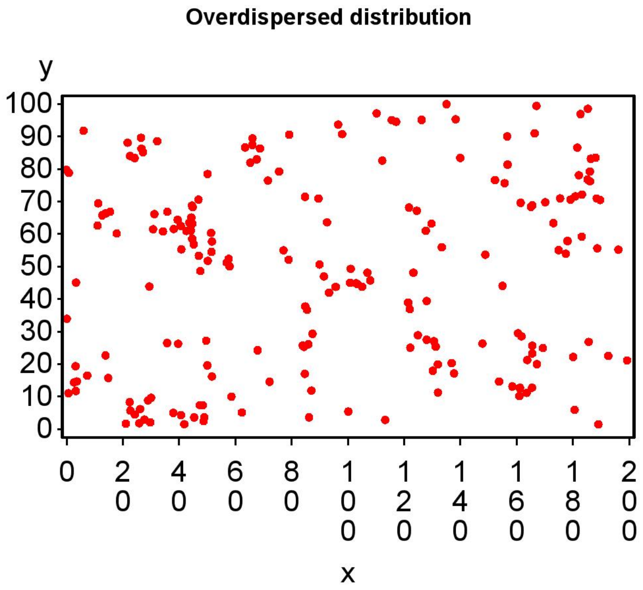


Figure 5.23: Overdispersed distribution of organisms in space

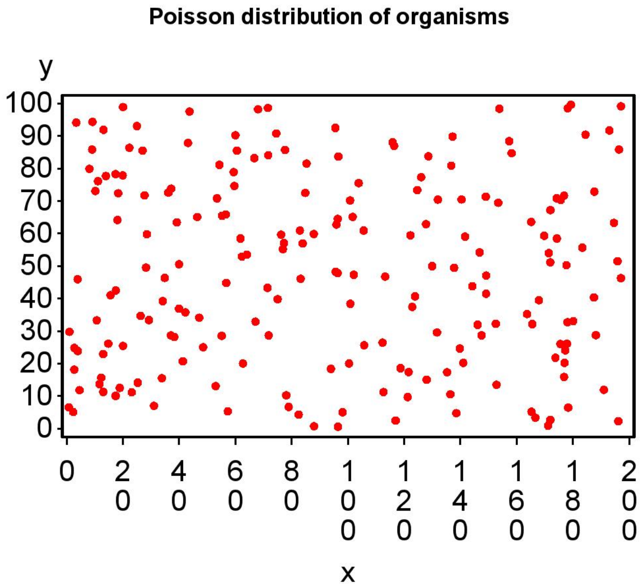


Figure 5.24: Poisson distribution of organisms in space

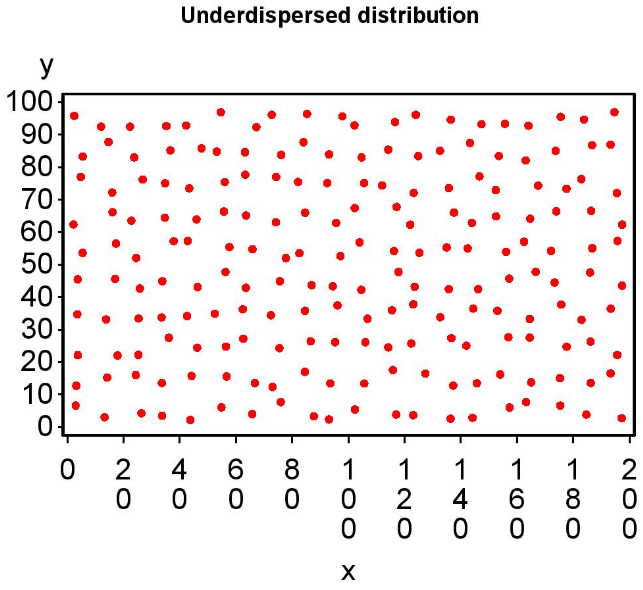


Figure 5.25: Underdispersed distribution of organisms in space

5.7 References

Begon, M., Townsend, C. R. & Harper, J. L. (2006) *Ecology: From Individuals to Ecosystems.* Blackwood Publishing, Malden, MA.

Bliss, C. I. & Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics* 9: 176-200.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

Pielou, E. (1977) *Mathematical Ecology.* John Wiley & Sons, Inc., New York, NY.

Reeve, J. D., and J. T. Cronin (2010) Edge behaviour in a minute parasitic wasp. *Journal of Animal Ecology*, 79: 483-490.

Ridout, M. S. & Besbeas, P. (2004) An empirical model for underdispersed data. *Statistical Modelling* 4: 77-89.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Snyder, D. L. & Miller, M. I. (1991) *Random Point Processes in Time and Space*, 2nd edition. Springer-Verlag New York Inc., New York, NY.

5.8 Problems

1. Consider the dice cube example from Chapter 4, and define a random variable Y that is the number of spots showing on the dice cube. Find $E[Y]$ and $Var[Y]$ for this random variable. Show your work.

2. Suppose that a random variable Y has a discrete distribution with the following probabilities:

y	$P[Y = y]$
0	0.5000
1	0.2500
2	0.1250
3	0.0625
4	0.0625

   (a) What is the expected value of Y, or $E[Y]$?

   (b) What is the variance of Y, or $Var[Y]$?

3. An entomologist studies the spatial distribution of aphids in a field. They randomly select 100 locations within the field and count the number of aphids on the plants at each location. The following observed frequency distribution was obtained:

Aphids (y)	Frequency
0	19
1	22
2	16
3	10
4	11
5	11
6	6
7	2
8	1
9	1
10	1
11	0

(a) Use the SAS program `Poisson_fit.sas` to calculate $\bar{Y}$ and s^2, and generate a plot of the observed frequencies vs. those expected for the Poisson distribution. Attach your SAS program and output.

(b) Based on the above results, do the data have a Poisson distribution? Explain your answer using the pattern of observed and expected frequencies, and the values of $\bar{Y}$ and s^2. Is the pattern random (Poisson), overdispersed, or underdispersed?

(c) What are some possible biological explanations for this pattern?

4. A field is surveyed for golden mice (*Ochrotomys nuttalli*) using a grid of baited traps. A total of 100 traps were deployed and the number of mice counted in each trap. The following frequency distribution was obtained:

Mice (y)	Frequency
0	55
1	21
2	10
3	7
4	4
5	2
6	1
7	0
8	0

(a) Use the program `Poisson_fit.sas` to calculate to calculate $\bar{Y}$ and s^2, and generate a plot of the observed frequencies vs. those expected for the Poisson distribution. Attach your program and output.

(b) Based on the above results, do the data have a Poisson distribution? Explain your answer using the pattern of observed and expected frequencies, and the values of $\bar{Y}$ and s^2. Is the pattern random (Poisson), overdispersed, or underdispersed?