Probability distributions

Probability distributions are important for working with **random variables**. Examples of random variables are:

- the number of ponds where we detect frogs,
- the number of muntjac we see during a survey,
- the weight of a randomly selected squirrel.

The result is not the same in every experiment, but if we repeat the experiment many times we can investigate the **distribution** of the result, and select a suitable model and parameter values to describe the population.

Here we will review three of the most useful probability distributions: the binomial and Poisson distributions for count data, and the normal or Gaussian distribution for continuous measurements.

Binomial distribution

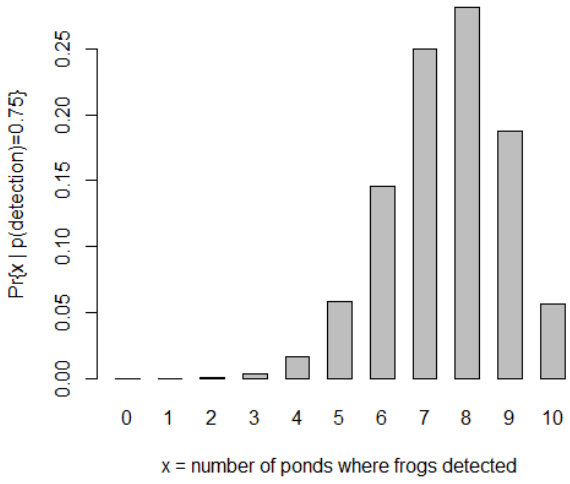
The binomial distribution applies to a **series of identical, independent trials**, each with only **two possible outcomes** (the "bi" in binomial indicates "two"). In wildlife biology the outcomes might be presence/absence, occupied/unoccupied, detected/not detected, captured/not captured, survived/died, etc.

For example, we go to 10 ponds and listen for the calls of a certain species of frog. We record the number of ponds where we hear frogs. The result will be a number between 0 and 10.

The data are nonnegative **integers** (whole numbers), with a known **upper limit**. We usually need to estimate the **probability of success** in each trial.

The binomial distribution gives the probability of obtaining x successes in n independent trials with the same probability of success (p) in each trial. We can use it to calculate the probability of detecting frogs at 0, 1, 2, …, 10 out of 10 ponds when the detection probability at each pond is 0.75. The table and graph are given below:

x = no. of ponds where frogs detected	Probability of detecting frogs at x ponds given p(detect) = 0.75
0	0.00000
1	0.00003
2	0.00039
3	0.00309
4	0.01622
5	0.05840
6	0.14600
7	0.25028
8	0.28157
9	0.18771
10	0.05631



The table was done in Excel using BINOMDIST(A…, 10, 0.75, FALSE) and the bar graph was in R using `dbinom(0:10, 10, 0.75)`.

The formula for the binomial distribution is:

$$P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X}$$

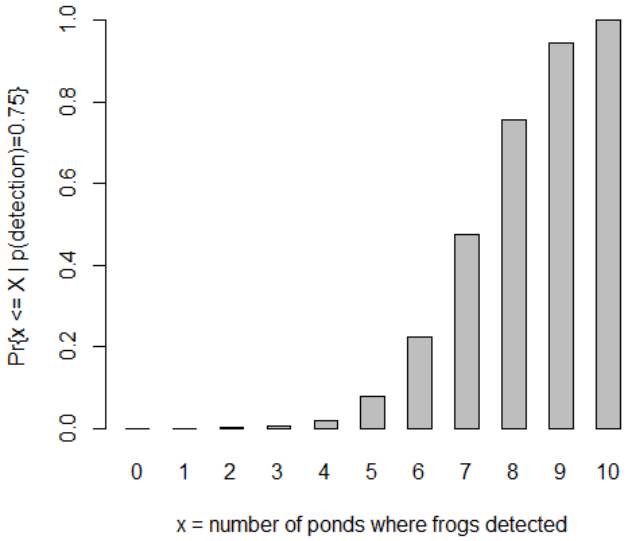
('*n*!' is '*n* factorial', the product of *n* and all the integers less than *n*; so 5! = 5 x 4 x 3 x 2 x 1 = 120.)

The Expected Value (the mean of a very large number of experiments) is E(*X*) = *np* and the variance var(*X*) or $\sigma^2 = np(1-p)$.

Cumulative probability distribution

It's often useful to know the probability of getting a certain number of successes *or fewer*: for example, we may want to know the probability of detecting frogs at *up to 3 ponds* if *p*(detect) = 0.75. We could calculate the value for 0, 1, 2 and 3 and add them up. But both Excel and R have ways of calculating the cumulative distribution. In Excel we use BINOMDIST(A…, 10, 0.6, **TRUE**) and in R we use **p**`binom` instead of **d**`binom`. The results are shown below:

x = no. of ponds where frogs detected	Probability of detecting frogs **at** $\leq$ **x ponds** given *p*(detect) = 0.75
0	0.00000
1	0.00003
2	0.00042
3	0.00351
4	0.01973
5	0.07813
6	0.22412
7	0.47441
8	0.75597
9	0.94369
10	1.00000



Poisson distribution

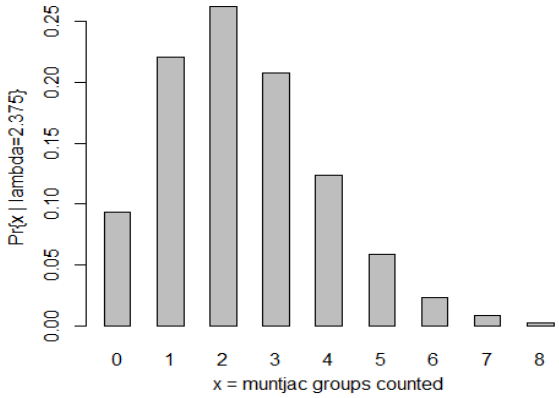
The Poisson distribution is often used to model **count data** for populations that are **randomly dispersed**. For example, during surveys at Batang Ai we counted the following numbers of groups of muntjac at different sites where the survey effort was equal.

Akup	5
Bebiyong	2
Bedegum	1
Bratik	2
Kapok	4
Pantar	1
Sebarik	4
Teliting	0

The counts are nonnegative **integers** (whole numbers) but, unlike the data for the binomial distribution, there is **no upper limit** to the number of groups of muntjac we might see. Large numbers just become less and less probable. We usually need to estimate the **expected number of animals** (λ, lambda) at each site from the data.

For the muntjac data above, $\bar{x}$ = 2.375, and the Poisson probability distribution for λ = 2.375 is shown below:

No. of muntjac groups seen	Probability
0	0.093014
1	0.220909
2	0.26233
3	0.207678
4	0.123309
5	0.058572
6	0.023185
7	0.007866
8	0.002335
9	0.000616
>9	0.000186



In Excel, we used POISSON(A…, 2.375, FALSE) to produce the table, and in R, `dpois(0:8, 2.375)` for the bar graph.

The equation for the Poisson distribution is:

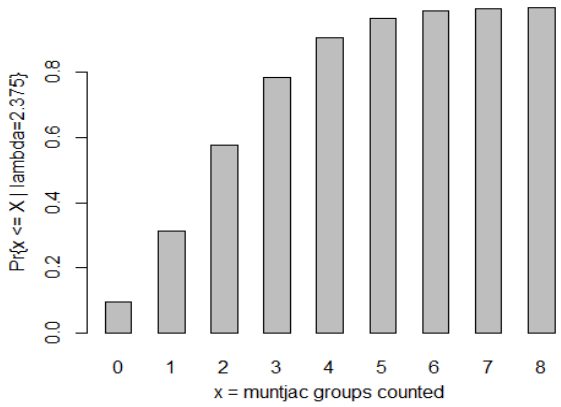
$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

where λ is the average number of muntjac per site. The expected value of x, E(x) = λ and variance of x is also equal to λ.

Cumulative probability distribution

We can also calculate and plot cumulative probabilities for Poisson distributed variables:

No. of muntjac groups seen	Cumulative probability
0	0.093014
1	0.313924
2	0.576254
3	0.783932
4	0.90724
5	0.965812
6	0.988997
7	0.996863
8	0.999198
9	0.999814



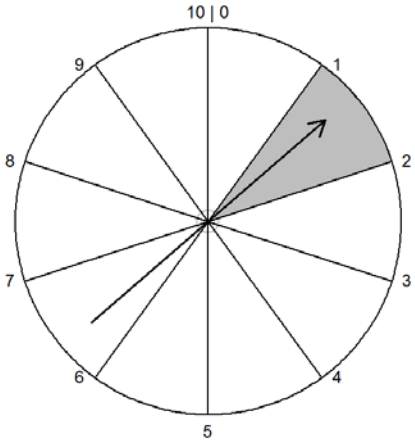
Notice that the cumulative probability for x ≤ 9 is still < 1, but as x gets larger the cumulative probability gets closer and closer to 1.

Continuous variables and probability density

The binomial and Poisson distributions deal with **discrete variables** – integers or whole numbers – where it makes sense to talk about the probability of a specific observation, eg. the probability of hearing frogs at (exactly) 6 ponds out of 10.

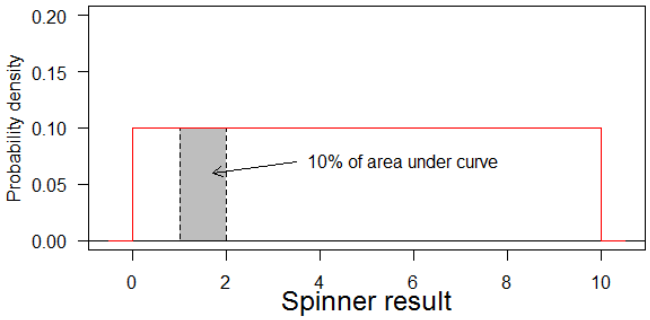
This won't work for **continuous variables**, such as the heights of people or the weights of squirrels. The probability that a squirrel will weight *exactly* 980g, that is 980.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000g is very nearly zero. And it's zero for any other exact value we might choose. But the probability that it falls in a given interval, say between 975 and 985, is not zero and we can calculate that using the concept of **probability density**.

To simulate a continuous variable, think of a spinner (see diagram right) where you flick the arrow to make it spin and see where it stops, which may be anywhere between 0 and 10.



Suppose we divide the circle into 10 sectors, as indicated by the tick marks. What's the probability that the arrow will stop between 1 and 2? Assuming it's fair, ie. all possible stopping positions are equally probable, this probability is 1/10 (10%). If we divide the circle into *N* equal sectors, the probability that the arrow will stop in a specified sector is 1/*N*; we can make *N* as big as we like and 1/*N* as small as we like.

Now think about the ratio of probability to sector width. This is (1/*N*) / (10/*N*) = 0.1, and doesn't change when we change N. This ratio is the **probability density**. (We sometimes refer to plain old probability as **probability mass** to distinguish it from probability density.) If we draw a graph of the probability density, it looks like this:



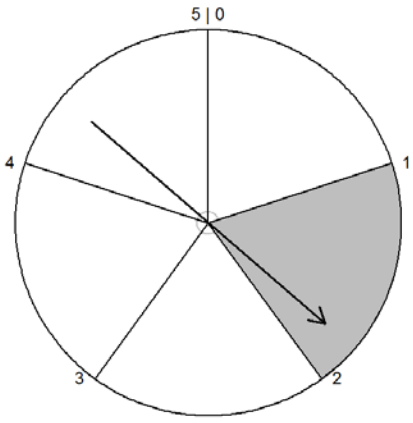
The red line plots the probability density function (pdf). The grey rectangle represents the range 1 to 2. The probability of a result in that range is:

probability density x width of range = 0.1 x (2 – 1) = 0.1 x 1 = 0.1 = 1/10.

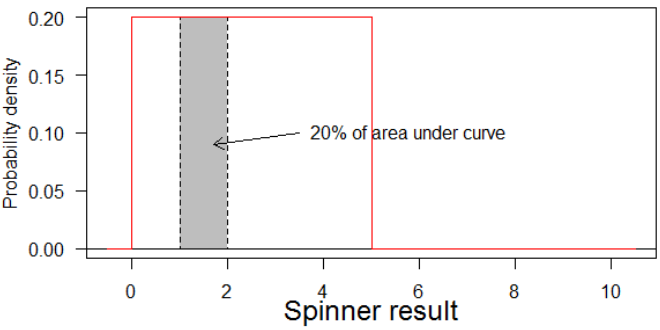
which is equal to the area of the grey rectangle. It's a general rule that the probability that a result falls in a given range equals the area under the pdf curve for the relevant range.

The total area under the red curve is just 0.1 x 10 = 1. In fact, the total area under a pdf curve = 1 always, as the probability that the result is somewhere in the range from –∞ to +∞ is 1.

Now think of a spinner marked out from 0 to 5 instead of 0 to 10, the probability that the arrow will stop between 1 and 2 is now 1/5 (20%), although the sector width is still 2 – 1 = 1.



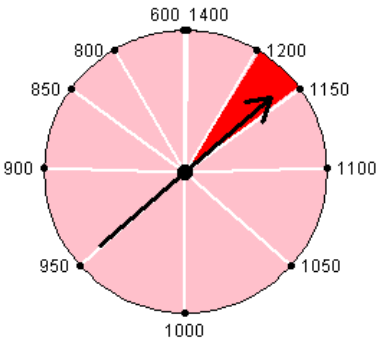
Here the probability density = 0.2, instead of 0.1, over the range from 0 to 5; the probability of a result below 0 or above 5 is zero. And the area under the red curve is 1.



Next we'll look at some more interesting pdf curves.

Normal or Gaussian distribution

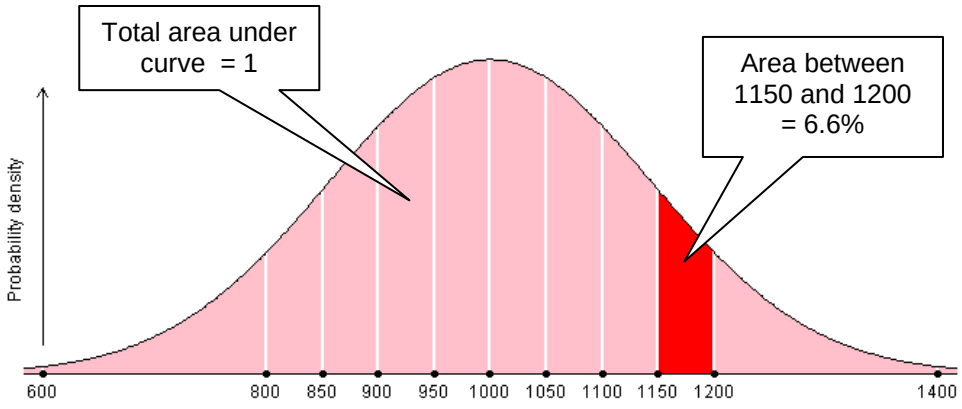
This is the 'bell curve' often used to model measurements such as the heights of people or the weights of squirrels. The spinner for a normal distribution with mean μ = 1000 and standard deviation σ = 145 is shown on the right. The scale actually goes from –∞ to +∞, though the probability of the spinner landing below 600 or above 1400 is tiny.



Probability that spinner lands between 1150 and 1200 = area of red segment

The probability of a result between 1150 and 1200g equals the area of the dark segment, which is 0.066 or 6.6%.

Unfolding[1] the spinner produces the graph below, which is the usual plot of the normal probability density function (pdf), with mean μ = 1000g and standard deviation σ = 145g:



Probability that squirrel weight is between 1150 and 1200 = dark red area

The total area under the curve from –∞ to +∞ is 1, as it must be. The dark area corresponds to weights between 1150 and 1200g. Its area = 0.066, ie. the probability that a squirrel picked at random weighs between 1150 and 1200g is 0.066 or 6.6%.

It may be useful to know that the probability of a value lying within 1 SD of the mean (between 855 and 1145g in our example) is 68% or about 2/3, and within 2×SD (810 to 1290g) it's 95%.

The equation for the normal pdf is:

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

That looks horrible, but in practice we use the functions NORMDIST in Excel or `dnorm` in R. We are usually more interested in the area under the curve than the height of the curve, and we can get that with NORMDIST (setting cumulative to TRUE) or **p**`norm` in R, which give the area under the curve to the left of the value entered.

For example, in R, using `pnorm(1200, 1000, 145)` gives the probability that a randomly drawn squirrel will weigh less than 1200g, which is 0.916. Similarly `pnorm(1150, 1000, 145)` gives the probability that a randomly drawn squirrel will weigh less than 1150g, which is 0.850. So the probability that our squirrel weighs between 1150 and 1200g is

[1] An mp4 video showing the unfolding is at http://mikemeredith.net/blog/1309_Normal_pdf_animation.htm

```
pnorm(1200, 1000, 145) - pnorm(1150, 1000, 145)
```

= 0.916 – 0.850 = 0.066.

Points to recall

- In biology we are often dealing with **random variables**.
- We deal with randomness by **modelling** the variable as a **probability distribution** and estimating the **parameters**.
- For **discrete variables** we use probability distributions such as the binomial or Poisson distribution.
- With continuous variables, probability is only meaningful for a range of values (eg. $1150 < x < 1200$), so we use **probability density functions** such as the normal (Gaussian).