Chapter 6

Continuous Random Variables

We previously examined several different probability distributions for discrete random variables, in particular the binomial, Poisson, and negative binomial distributions. These distributions are suitable for modeling observations that are counts of some type, such as the number of plants in a quadrat or the number of females vs. males in a sample. Many variables in biology are continuous, however, such as the length and weight of organisms, quantities associated with populations such as birth, mortality, and growth rates, and chemical concentrations. We will now examine continuous random variables and their associated distributions that are used to model these quantities, in particular the **uniform and normal distributions**. The uniform distribution is often used to generate random sampling points in one- and two-dimensional areas. For example, we could use the uniform distribution to select a random point along a transect to sample, or a random x, y coordinate within a field to place a sampling quadrat. It also a useful starting point for understanding continuous distributions because of its simplicity. We then turn to the normal distribution, which forms the basis of many statistical procedures. Many biological variables have a distribution close to normal, or if initially non-normal can often be transformed to more closely resemble the normal distribution.

Discrete random variables have a function $f(y)$ that directly provides the probabilities for events that are integers, such as $Y = 0$, $Y = 3$, and so forth (see Chapter 5). However, events for continuous random variables are in the form of intervals. For example, we will be interested in finding the probability for events like $1 < Y < 3$ or $Y > 5$. Continuous random variables use a different kind of function, called a **probability density function**, to find the probabilities for events. For an event like $1 < Y < 3$, probabilities are found by integrating the probability density function (finding the area under the function) over this interval. This process will be explained in more detail below. For many continuous random variables, such as the normal distribution, there exist tables of these integrals and probabilities for certain useful intervals. Note that events like $Y = 3$ have zero probability for continuous random variables, because this implies an interval of zero width and so the integral is zero. This makes some intuitive sense, because it is unlikely that a continuous quantity Y would take a value exactly equal to 3 to many decimal places.

6.1 Uniform distribution

Suppose that we have two constants, a and b, with $a < b$. A random variable Y has a uniform distribution if an observation is equally likely to occur anywhere between a and b, but never occurs outside this interval. The probability density for the uniform distribution is defined by the equation

$$f(y) = \frac{1}{b-a} \tag{6.1}$$

for $a \leq y \leq b$ (Mood et al. 1974). Outside of this interval, we have $f(y) = 0$. The quantities a and b are the parameters of the uniform distribution. The uniform distribution for $a = 0$, $b = 1$ is shown below (Fig. 6.1). The uniform distribution gets its name from the fact that its density is uniform over the interval a to b.

Note that the density simply describes a square with a length and width of one, implying an area equal to one. This is an important property of probability density functions in general – the area under $f(y)$ is always equal to one. Also shown is the uniform density for $a = 0$ and $b = 2$ (Fig. 6.2). It is lower but wider than the previous example, and also has an area of one.

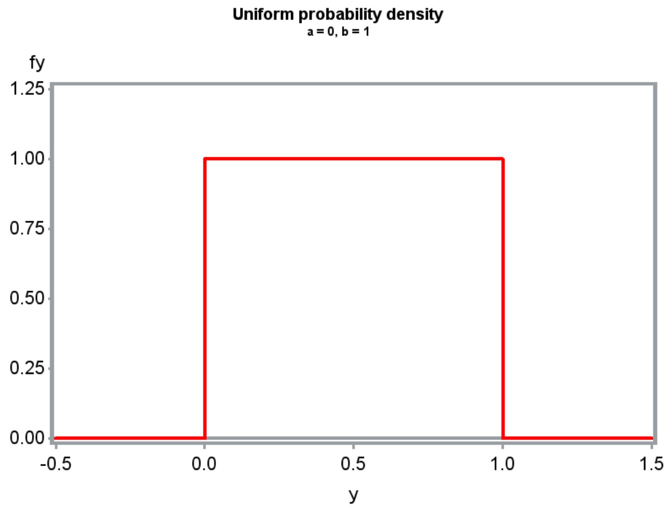

Figure 6.1: Uniform probability density for $a = 0, b = 1$

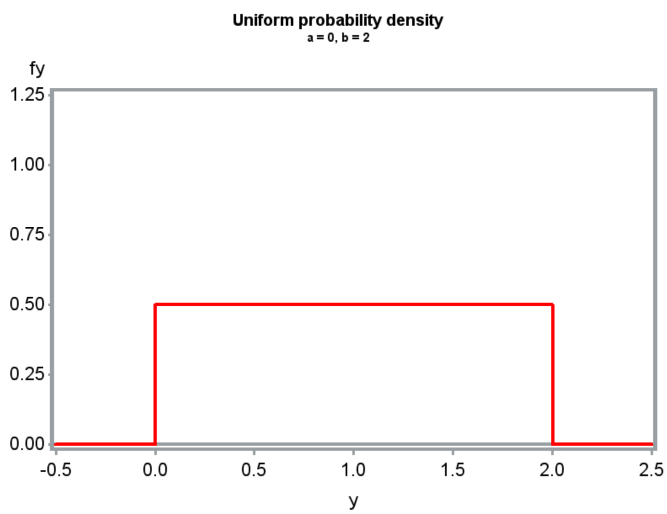

Figure 6.2: Uniform probability density for $a = 0, b = 2$

Probabilities for the uniform distribution are calculated by finding the area under the probability density function, using integration (see Chapter 2). This is relatively easy to do because of the simple form of the probability density. Suppose Y is a uniform random variable, and $a = 0$ and $b = 1$. What is the probability that an observed Y lies within the interval 0.5 to 0.75? We have

$$P[0.5 < Y < 0.75] = \int_{0.5}^{0.75} \frac{1}{b-a} dy \tag{6.2}$$

$$= \int_{0.5}^{0.75} \frac{1}{1-0} dy = y|_{0.5}^{0.75} \tag{6.3}$$

$$= 0.75 - 0.5 = 0.25. \tag{6.4}$$

We could also have found this probability without any calculus. It is just the area under $f(y)$ between 0.5 and 0.75, calculated as length $\times$ height $= (0.75 - 0.5) \times 1 = 0.25$.

Here are two more examples. Suppose that for $a = 0$ and $b = 2$, we want to find the probability that $0.2 < Y < 0.4$. The height of the density function in this case is $1/(b - a) = 1/(2 - 0) = 0.5$. We therefore have $P[0.2 < Y < 0.4] = (0.4 - 0.2) \times 0.5 = 0.1$. Now suppose we want the probability that $0 < Y < 2$. We have $P[0 < Y < 2] = (2-0) \times 0.5 = 1$. This also follows from the fact that $f(y)$ is a probability density function which has an area of one, and the interval $0 < Y < 2$ encompasses the entire range of $f(y)$.

The **cumulative distribution function** for a continuous random variable is defined as the quantity

$$F(y) = P[Y < y] = \int_{-\infty}^{y} f(z)dz. \tag{6.5}$$

This function is just the probability to the left of y. The function $F(y)$ increases from 0 to 1 as y increases. If we carry out this integral for the uniform distribution, we get the function

$$F(y) = \frac{y-a}{b-a} \tag{6.6}$$

for $a \leq y \leq b$. In addition, $F(y) = 0$ for $y < a$, and $F(y) = 1$ for $y > b$. Figure 6.3 shows the cumulative distribution function for the uniform distribution corresponding to Fig. 6.2. Note that it increases linearly between a and b, as the probability to the left of y accumulates. The cumulative distribution function has many uses in statistics, especially for continuous random variables.

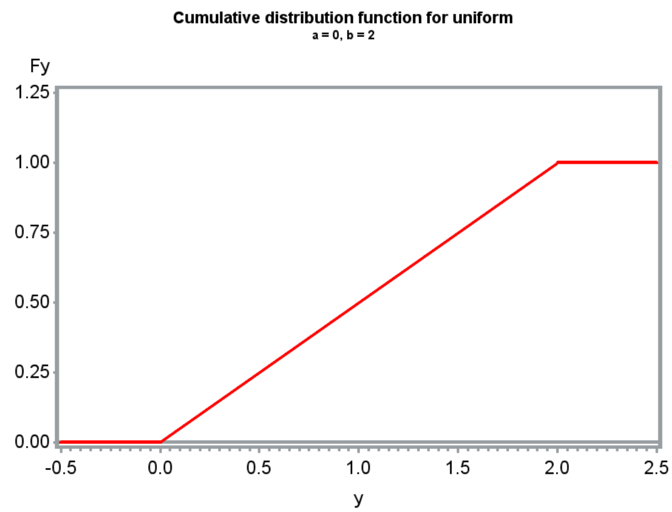

Figure 6.3: Cumulative distribution function for the uniform distribution, with $a = 0, b = 2$

The uniform distribution has a number of common applications. It is possible to generate a stream of random numbers that have a uniform distribution using software, and from these values produce random observations for other distributions, including discrete distributions as well as the normal distribution. The uniform distribution can also be used to generate random sampling points along a transect for ecological studies, or random x, y coordinates for placing quadrats within an area (see below). It can also be used to randomly sample from a population, or to randomize the order of treatments in an experiment.

6.1.1 Random sampling coordinates - SAS demo

A common application of the uniform distribution is to generate random sampling coordinates. SAS can produce random observations with a uniform distribution using the function `ranuni`. For this function, the parameter values of the uniform distribution are set at $a = 0$ and $b = 1$.

However, we will often want observations for other parameter values, especially other values of b. It can be shown that if Y has a uniform distribution with $a = 0$ and $b = 1$, then the variable $Y' = cY$ has a uniform distribution with $a = 0$ and $b = c$, where c is any positive number. This fact enables us to generate uniform random variables with any value of b.

For example, suppose we want to produce random sampling coordinates along a 100 m transect using the uniform distribution. If Y has a uniform distribution with $a = 0$ and $b = 1$, then $Y' = 100Y$ has a uniform distribution with $a = 0$ and $b = 100$. Values of Y generated in this fashion will give us sampling coordinates uniformly distributed between 0 and 100 m.

We will illustrate this process using a SAS program to generate random sampling coordinates for a 100 m transect and also a 200×100 m rectangular area. A call to `gplot` is used to plot the random coordinates. See SAS program and output below.

SAS Program

```
* randcoords.sas;
title "Generate random sampling coordinates";
* Generate n random coordinates along a c m transect;
data transect;
    * Sample size n;
    n = 20;
    * Multiplying by c gives a uniform random variable with a=0, b=c;
    c = 100;
    do i = 1 to n;
        x = c*ranuni(0);
        output;
    end;
    drop i;
run;
* Print coordinates;
proc print data=transect;
run;
* Generate n random coordinates within a 200 x 100 m area;
data coords;
    * Sample size n;
    n = 200;
    * Multiplying by c_x gives a uniform random variable with a=0, b=c_x;
    c_x = 200;
    * Multiplying by c_y gives a uniform random variable with a=0, b=c_y;
```

```
    c_y = 100;
    do i = 1 to n;
        x = c_x*ranuni(0);
        y = c_y*ranuni(0);
        output;
    end;
    drop i;
run;
* Print first 25 coordinates;
proc print data=coords(obs=25);
run;
* Show coordinates as a scatterplot;
proc gplot data=coords;
    plot y*x / vaxis=axis1 haxis=axis2;
    symbol1 v=dot c=red;
    axis1 order=(0 to 100 by 10) label=(height=2) value=(height=2)
    width=3 major=(width=2) minor=none;
    axis2 order=(0 to 200 by 20) label=(height=2) value=(height=2)
    width=3 major=(width=2) minor=none;
run;
quit;
```

Generate random sampling coordinates

Obs	n	c	x
1	20	100	45.6949
2	20	100	73.6408
3	20	100	38.8120
4	20	100	89.1029
5	20	100	34.9528
6	20	100	38.4595
7	20	100	7.6446
8	20	100	92.3383
9	20	100	92.3485
10	20	100	55.6395
11	20	100	8.8543
12	20	100	16.5194
13	20	100	94.9774
14	20	100	21.7407
15	20	100	84.6223
16	20	100	63.1644
17	20	100	48.5556
18	20	100	52.2919
19	20	100	17.1289
20	20	100	23.8348

Figure 6.4: `randcoords.sas - proc print`

Generate random sampling coordinates

Obs	n	c_x	c_y	x	y
1	200	200	100	47.102	82.2807
2	200	200	100	33.231	85.3004
3	200	200	100	112.908	31.4955
4	200	200	100	17.164	50.1056
5	200	200	100	120.141	6.1346
6	200	200	100	33.265	75.7294
7	200	200	100	33.709	3.0551
8	200	200	100	47.762	80.3206
9	200	200	100	128.141	35.2206
10	200	200	100	22.302	55.5951
11	200	200	100	43.010	4.0473
12	200	200	100	176.505	89.8507
13	200	200	100	125.940	18.4065
14	200	200	100	47.596	41.2316
15	200	200	100	25.479	76.9636
16	200	200	100	156.142	62.2666
17	200	200	100	140.374	8.6684
18	200	200	100	133.532	75.4055
19	200	200	100	158.624	15.3123
20	200	200	100	129.904	28.8471

etc.

Figure 6.5: `randcoords.sas - proc print`

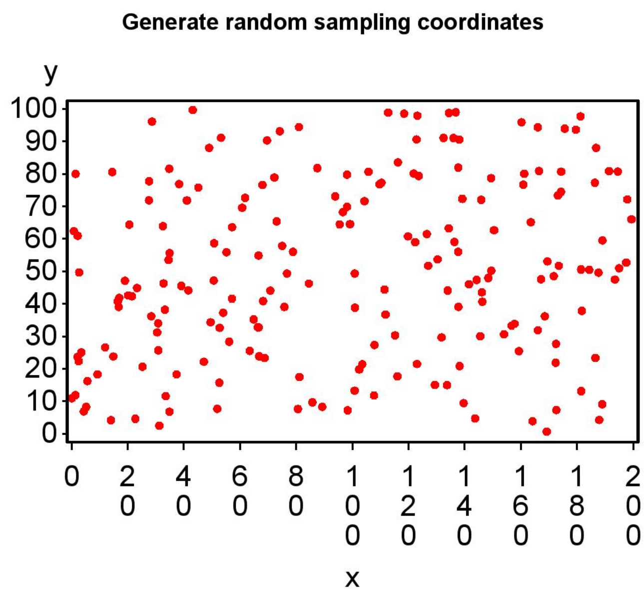

Figure 6.6: `randcoords.sas - proc gplot`

6.2 Normal distribution

The normal distribution plays an important role in statistics, with good reason. Biological variables often have a distribution that can be approximated by the normal or can be transformed to be normal. The normal distribution is thus a valid choice for modeling many variables encountered in practice. Many statistical quantities will also have a distribution approaching the normal for large sample sizes. For example, the distribution of the sample mean $\bar{Y}$ will approach the normal distribution as the sample size n increases, thanks to the central limit theorem (see Chapter 7). So, even if the underlying data are non-normal, statistics like $\bar{Y}$ will be normally-distributed for sufficiently large n.

The probability density for the normal distribution is defined by the function

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{6.7}$$

for $\infty < \mu < \infty$ and $\sigma^2 > 0$ (Mood et al. 1974). The normal distribution has two parameters, μ and σ^2. The parameter μ is the mean of the distribution and basically controls its location, while σ^2 is its variance and determines its dispersion or spread. A random variable Y with a normal distribution is often written as $Y \sim N(\mu, \sigma^2)$, where the symbol '$\sim$' stands for 'is distributed as' while 'N' signifies the normal. A random variable with a **standard normal distribution** assumes that $\mu = 0$ and $\sigma^2 = 1$, or $Y \sim N(0, 1)$. The symbol Z is often used to denote a standard normal random variable.

Figure 6.7 shows the bell-shaped normal distribution for three different sets of μ and σ^2 values, and illustrates how these parameters affect its location and shape. As μ is increased the distribution shifts to the right, while an increase in σ^2 causes the distribution to spread out.

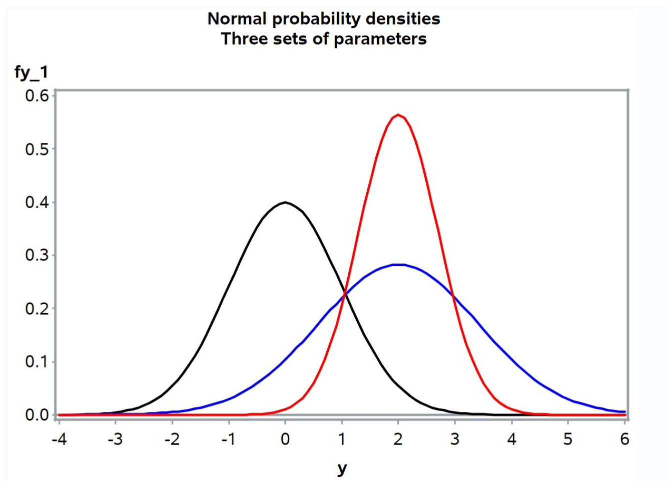

Figure 6.7: `normal_plot3.sas - proc gplot`

6.2.1 Normal distribution - SAS demo

The SAS program used to generate Fig. 6.7 is listed below. Three different sets of μ and σ^2 values are given in the `data` step of the program (feel free to experiment with other values). The different curves are specified in the `plot` statement for `proc gplot`. The `overlay` option is used to generate a single graph with all three curves, each with different colors specified by the `symbol` statements.

SAS Program

```
* normal_plot3.sas;
options pageno=1 linesize=80;
goptions reset=all;
title "Normal probability densities";
title2 "Three sets of parameters";
data normal_plot;
    * Three sets of normal parameters here;
    mu_1 = 0; sig2_1 = 1;
    mu_2 = 2; sig2_2 = 2;
    mu_3 = 2; sig2_3 = 0.5;
    * Minimum and maximum values of y;
    ymin = -4;
    ymax = 6;
    * Divisions between ymin and ymax (more = smoother graph);
    ydiv = 100;
    * Calculate step length;
    ylength = (ymax-ymin)/ydiv;
    * Find y and f(y) values for the plot;
    do i=0 to ydiv;
        y = ymin + i*ylength;
        * normal probability density function;
        fy_1 = (1/sqrt(2*3.14159*sig2_1))*exp(-((y-mu_1)**2)/(2*sig2_1));
        fy_2 = (1/sqrt(2*3.14159*sig2_2))*exp(-((y-mu_2)**2)/(2*sig2_2));
        fy_3 = (1/sqrt(2*3.14159*sig2_3))*exp(-((y-mu_3)**2)/(2*sig2_3));
        * Output y and fy1, fy2, fy3 to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=normal_plot;
run;
* Plot probability density function;
proc gplot data=normal_plot;
    plot fy_1*y=1 fy_2*y=2 fy_3*y=3 / vref=0 wvref=3 vaxis=axis1 haxis=axis1 overlay;
```

```
    symbol1 i=join v=none c=black width=3;
    symbol2 i=join v=none c=blue width=3;
    symbol3 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```

The cumulative distribution function for the normal distribution is defined as the quantity

$$F(y) = P[Y < y] = \int_{-\infty}^{y} f(z)dz = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz. \quad (6.8)$$

The values of this integral have to be numerically calculated. Fig. 6.8 shows the cumulative distribution functions for the three normal distributions shown in Fig. 6.7. Note that the mean and variance for the different normal distributions affect the overall location and shape of $F(y)$.

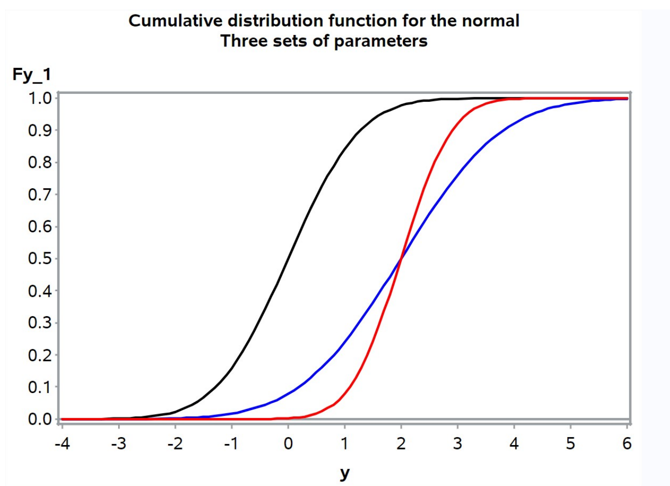

Figure 6.8: Cumulative distribution function for three normal distributions

Like other continuous random variables, events for the normal distribution are in the form of intervals. We can calculate the probabilities for events by finding the area under the normal density function corresponding to the interval. This process is more difficult than for the uniform distribution because $f(y)$ has a more complex shape. However, there exist tables of the area under $f(y)$ for certain intervals that can be used for this purpose, as well as the SAS function `probnorm`. Table Z gives the probabilities for intervals of the form $Z < z$, where Z has a standard normal distribution and $z \geq 0$ (see Chapter 23). The first two digits of z are specified in the left-most column of Table Z, while the third digit is the top row. The values within the table correspond to the probability that $Z < z$, or $P[Z < z]$, i.e., the cumulative distribution function for the standard normal.

6.2.2 Sample calculations - standard normal distribution

We illustrate how Table Z is used to calculate the probabilities for various events listed below. The general strategy is to sketch the interval on the standard normal bell curve, and deduce from this picture how to obtain the probability using Table Z.

1. Find the probability that $Z < 0.55$, or $P[Z < 0.55]$. From Table Z, we see that $P[Z < 0.55] = 0.7088$. See Fig. 6.9 for an illustration of this probability.

2. Find the probability that $0.40 < Z < 1.96$. In this case, the interval is not the same as shown in Table Z, and additional calculations are required. We first find the probabilities for the intervals $Z < 1.96$ and $Z < 0.4$ using Table Z. The probability for $0.40 < Z < 1.96$ should then be the difference between these two probabilities (see Fig. 6.10). We have $P[Z < 1.96] = 0.9750$ and $P[Z < 0.40] = 0.6554$ from Table Z, so $P[0.40 < Z < 1.96] = P[Z < 1.96] - P[Z < 0.40] = 0.9750 - 0.6554 = 0.3196$.

3. Find the probability that $Z > 0.55$. We will use the complement rule to obtain this probability (see Chapter 4). For any event A, we have $P[A^c] = 1 - P[A]$. If A is the event $Z < 0.55$, then A^C corresponds to $Z > 0.55$. Therefore, $P[Z > 0.55] = 1 - P[Z < 0.55] = 1 - 0.7088 = 0.2912$. See also Fig. 6.11.

4. Find the probability that $Z < -1.23$. This problem makes use of the symmetry of the standard normal distribution around zero, as well as the complement rule. By symmetry, we have $P[Z < -1.23] = P[Z > 1.23]$. The complement of $Z < 1.23$ is $Z > 1.23$, and so

$P[Z > 1.23] = 1 - P[Z < 1.23] = 1 - 0.8907 = 0.1093$. See Fig. 6.12.

5. Find the probability that $-0.44 < Z < 2.15$. This problem can also be handled using symmetry and the complement rule. We first have $P[Z < 2.15] = 0.9842$ using Table Z (Fig. 6.13). We then have $P[Z < -0.44] = P[Z > 0.44] = 1 - P[Z < 0.44] = 1 - 0.6700 = 0.3300$ by symmetry (Fig. 6.14). Therefore, $P[-0.44 < Z < 2.15] = P[Z < 2.15] - P[Z < -0.44] = 0.9842 - 0.3300 = 0.6542$.

6. Find a number z_0 such that $P[Z < z_0] = 0.95$. This problem is the inverse of the previous ones. Here, we want to find a value z_0 that gives a certain probability, rather than z_0 being a given quantity and determining the probability. To find z_0, we scan Table Z until we find a value that gives a probability close 0.95. We see that $z_0 = 1.64$ or 1.65 give approximately the right probability.

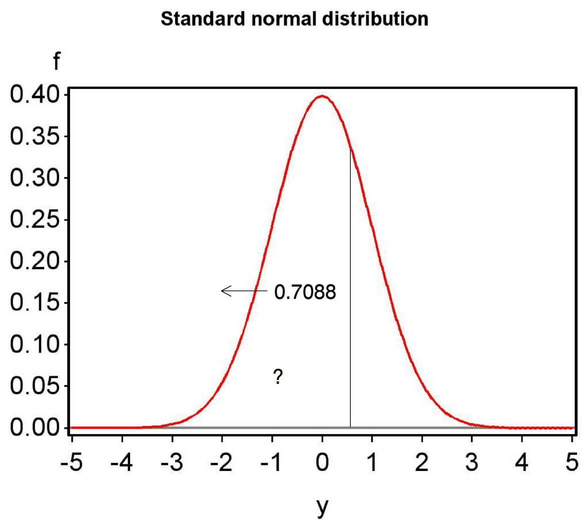

Figure 6.9: Sample calculation 1

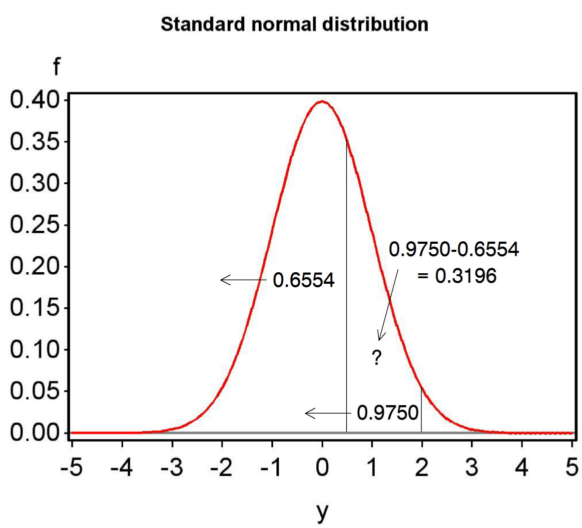

Figure 6.10: Sample calculation 2

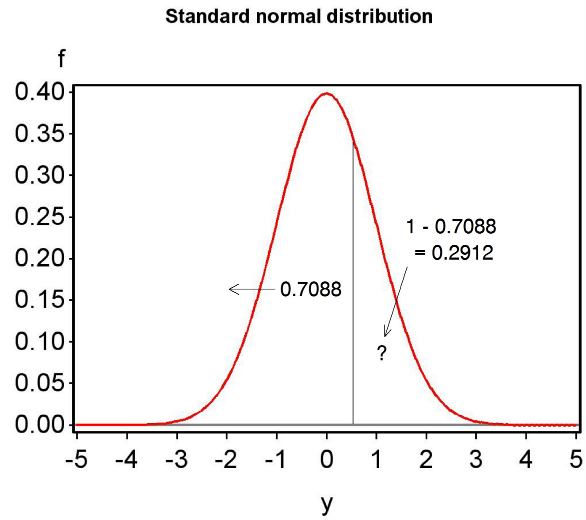

Figure 6.11: Sample calculation 3

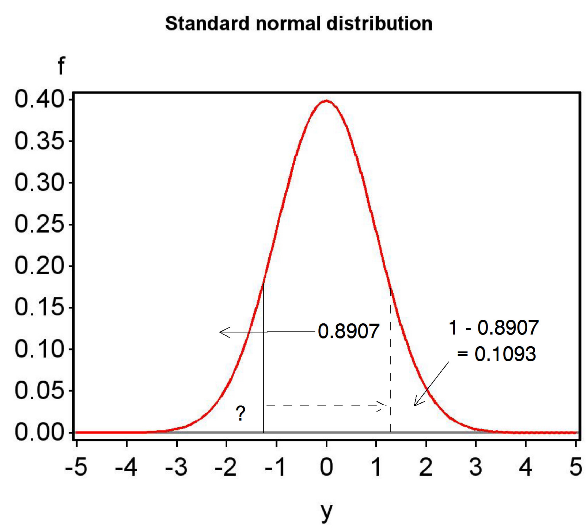

Figure 6.12: Sample calculation 4

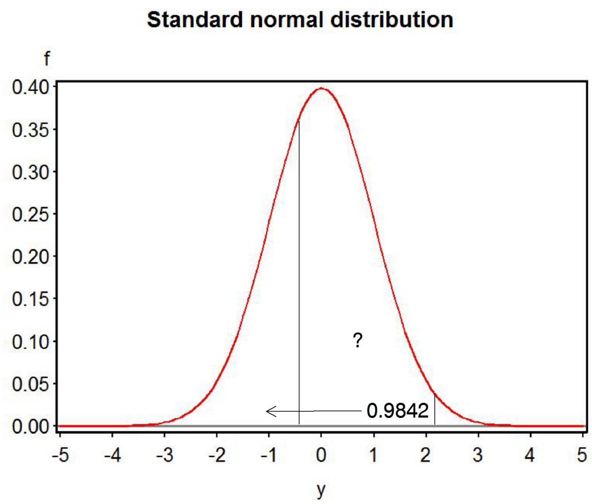

Figure 6.13: Sample calculation 5 - part 1

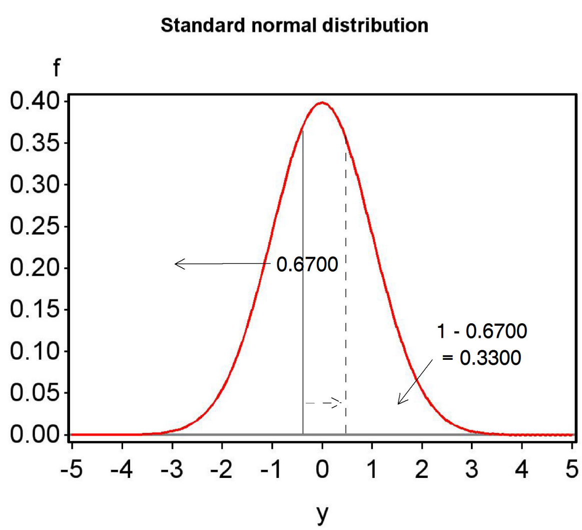

Figure 6.14: Sample calculation 5 - part 2

6.2.3 Sample calculations - other normal distributions

We now examine how probabilities can be calculated for normal distributions that are not standard normal. If $Y \sim N(\mu, \sigma^2)$, it can be shown that the quantity

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1) \tag{6.9}$$

Thus, a random variable Y with a normal distribution having any μ or σ^2 can be transformed to a standard normal Z. The transformation works by first centering the random variable Y around zero by subtracting μ, and then dividing by σ so that it has a standard deviation and variance of one. Once Y is transformed to a standard normal Z, we can find probabilities for any event involving Y using Table Z. This process is illustrated below in several sample calculations.

1. Suppose that $Y \sim N(50, 16)$. Find the probability that $Y < 55$. First, we find $\sigma = \sqrt{\sigma^2} = \sqrt{16} = 4$. Using the above equation, we then have

$$\begin{aligned} P[Y < 55] &= P\left[Y - \mu < 55 - \mu\right] && (6.10) \\ &= P\left[\frac{Y - \mu}{\sigma} < \frac{55 - \mu}{\sigma}\right] && (6.11) \\ &= P\left[Z < \frac{55 - 50}{4}\right] && (6.12) \\ &= P[Z < 1.25]. && (6.13) \end{aligned}$$

   We then use Table Z to find that $P[Z < 1.25] = 0.8944$, and so $P[Y < 55] = 0.8944$.

2. Find the probability that $52 < Y < 56$, assuming $Y \sim N(50, 16)$. To find this probability, we first convert the problem to one involving Z. We have

$$\begin{aligned} P[52 < Y < 56] &= P\left[52 - \mu < Y - \mu < 56 - \mu\right] && (6.14) \\ &= P\left[\frac{52 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{56 - \mu}{\sigma}\right] && (6.15) \\ &= P\left[\frac{52 - 50}{4} < Z < \frac{56 - 50}{4}\right] && (6.16) \\ &= P[0.50 < Z < 1.50]. && (6.17) \end{aligned}$$

We next find the probabilities for the intervals $Z < 1.50$ and $Z < 0.50$ using Table Z, and then substract them to obtain $P[0.50 < Z < 1.50]$. We have $P[Z < 1.50] = 0.9332$ and $P[Z < 0.50] = 0.6915$, so $P[0.50 < Z < 1.50] = 0.9332-0.6915 = 0.2417$. Thus, $P[52 < Y < 56] = 0.2417$.

3. Find the probability that $Y > 54$. We have

$$\begin{aligned} P[Y > 54] &= P\left[Y - \mu > 54 - \mu\right] && (6.18)\\ &= P\left[\frac{Y-\mu}{\sigma} > \frac{54-\mu}{\sigma}\right] && (6.19)\\ &= P\left[Z > \frac{54-50}{4}\right] && (6.20)\\ &= P[Z > 1.00]. && (6.21) \end{aligned}$$

We next use the complement rule to obtain this probability. We have $P[Z > 1.00] = 1 - P[Z < 1.00] = 1 - 0.8413 = 0.1587$, so $P[Y > 54] = 0.1587$.

4. Find the probability that $Y < 46.5$. We have

$$\begin{aligned} P[Y < 46.5] &= P\left[Y - \mu < 46.5 - \mu\right] && (6.22)\\ &= P\left[\frac{Y-\mu}{\sigma} < \frac{46.5-\mu}{\sigma}\right] && (6.23)\\ &= P\left[Z < \frac{46.5-50}{4}\right] && (6.24)\\ &= P[Z < -0.88]. && (6.25) \end{aligned}$$

By symmetry, we have $P[Z < -0.88] = P[Z > 0.88]$. The complement of $Z < 0.88$ is $Z > 0.88$, and so $P[Z > 0.88] = 1 - P[Z < 0.88] = 1 - 0.8106 = 0.1093$. So, $P[Y < 46.5] = 0.1093$.

5. Find the probability that $46 < Z < 52$. We have

$$\begin{aligned} P[46 < Y < 52] &= P\left[46 - \mu < Y - \mu < 52 - \mu\right] && (6.26)\\ &= P\left[\frac{46-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{52-\mu}{\sigma}\right] && (6.27)\\ &= P\left[\frac{46-50}{4} < Z < \frac{52-50}{4}\right] && (6.28)\\ &= P[-1.00 < Z < 0.50]. && (6.29) \end{aligned}$$

We then use symmetry and the complement rule to find this probability involving Z. We first have $P[Z < 0.50] = 0.6915$ using Table Z. We then have $P[Z < -1.00] = P[Z > 1.00] = 1 - P[Z < 1.00] = 1 - 0.8413 = 0.1587$ by symmetry. Therefore, $P[-1.00 < Z < 0.50] = P[Z < 0.50] - P[Z < -1.00] = 0.6915 - 0.1587 = 0.5328$, and so $P[46 < Y < 52] = 0.5328$.

6. Find a number y_0 such that $P[Y < y_0] = 0.70$. This problem can also be handled by converting it to one involving Z. We have

$$\begin{aligned} P[Y < y_0] &= P[Y - \mu < y_0 - \mu] && (6.30) \\ &= P\left[\frac{Y-\mu}{\sigma} < \frac{y_0-\mu}{\sigma}\right] && (6.31) \\ &= P\left[Z < \frac{y_0-50}{4}\right] && (6.32) \\ &= P[Z < z_0] && (6.33) \end{aligned}$$

where $z_0 = \frac{y_0-50}{4}$. We then search for a value of z_0 such that $P[Z < z_0] = 0.70$, and obtain $z_0 = 0.52$ from Table Z. We then solve for y_0 as follows:

$$\begin{aligned} z_0 &= \frac{y_0-50}{4} && (6.34) \\ 0.52 &= \frac{y_0-50}{4} && (6.35) \\ 4(0.52) &= y_0 - 50 && (6.36) \\ 2.08 &= y_0 - 50 && (6.37) \\ 2.08 + 50 &= y_0 && (6.38) \\ 52.08 &= y_0. && (6.39) \end{aligned}$$

So, $y_0 = 52.08$ is the answer. In general, one would have $z_0 = \frac{y_0-\mu}{\sigma}$, so $y_0 = \sigma z_0 + \mu$ for any σ and μ.

6.3 Expected values and variance for continuous distributions

We saw earlier how a theoretical mean, variance, and standard deviation could be calculated for a discrete random variable, using the concept of expectation and its probability distribution. The same concepts can be extended to continuous random variables and probability densities.

Let Y be a continuous random variable with some probability density. The expected value of Y, or its theoretical mean, is defined by the equation

$$E[Y] = \int_{-\infty}^{\infty} y f(y) dy \tag{6.40}$$

where $f(y)$ is the probability density of Y, and the integral is carried out over the interval $-\infty$ to ∞ (Mood et al. 1974). This equation is analogous to the definition of expected value for a discrete random variable, except that we use integration rather than summation to make the calculation.

Similar to discrete random variables, we can also define the theoretical variance of a continuous random variable using expectation. The variance of a continuous random variable Y is defined as

$$Var[Y] = E[(Y - E[Y])^2] = \int_{-\infty}^{\infty} (y - E[Y])^2 f(y) dy. \tag{6.41}$$

We can directly calculate these quantities for the uniform distribution. Recall from calculus that $\int u du = u^2/2$. We therefore have

$$E[Y] = \int_{-\infty}^{\infty} y f(y) dy = \int_{a}^{b} \frac{y}{b-a} dy \tag{6.42}$$

$$= \frac{1}{b-a} \frac{y^2}{2} |_a^b = \frac{1}{b-a} \frac{b^2 - a^2}{2} \tag{6.43}$$

$$= \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2} \tag{6.44}$$

Thus, the expected value (or theoretical mean) of a uniform random variable is located at the center of the interval, midway between a and b. It can also be shown using the above formula that the variance of the uniform distribution is

$$Var[Y] = \frac{(b-a)^2}{12} \tag{6.45}$$

The theoretical standard deviation is just the square root of this quantity.

What are these quantities for the normal distribution? Recall that the normal distribution is specified by the two parameters μ and σ^2. If $Y \sim N(\mu, \sigma^2)$, it can be shown (by evaluating the above integrals using the normal density) that

$$E[Y] = \mu \tag{6.46}$$

and

$$Var[Y] = \sigma^2. \tag{6.47}$$

Thus, the parameters μ and σ^2 for this distribution are the theoretical mean and variance $E[Y]$ and $Var[Y]$.

6.4 Continuous random variables and samples

Suppose we have a set of observations and want to determine if they can be modeled using the normal distribution. We now develop a graphical method of comparing these observed data with the pattern expected for the normal distribution, called a **normal quantile plot**. These plots exist for other continuous distributions as well, and are generally called quantile-quantile plots. The idea is to plot the quantiles for the observed data vs. the quantiles for the normal distribution, with the quantiles for the normal on the x-axis and the data quantiles on the y-axis. If the data are normally distributed, then this plot will resemble a straight diagonal line. This is because we are essentially plotting the quantiles for one normal distribution (the data) vs. the quantiles for the normal distribution itself (Wilk & Gnanadesikan 1968). This is like plotting the function $y = ax$, which is the equation of a line with slope a. See Chapter 3 for a review of quantiles such as the median, the 25% and 75% quartiles, and so forth.

We will illustrate the calculations for a normal quantile plot using a small data set. Suppose we have $n = 9$ data points that take the values 5.33, 4.98, 5.80, 4.37, 3.83, 2.76, 3.82, 4.02, and 3.09. We first order or rank the data points from smallest to largest, similar to finding the median (Table 6.1). We then find the proportion p of observations less than each data point, using the formula $p = (j - 3/8)/(n + 1/4)$, where j is the order of the data point and n is the sample size. Note that the median of these data (the value 4.02) corresponds to $p = 0.5$. The values 3/8 and 1/4 in the formula are there to prevent p from taking the value 0 or 1 for the largest and smallest observations.

Table 6.1: Calculations for a normal quantile plot

j (order)	$Y_{[j]}$	p	z
1	2.76	0.068	-1.49
2	3.09	0.176	-0.93
3	3.82	0.284	-0.57
4	3.83	0.392	-0.27
5	4.02	0.500	0.00
6	4.37	0.608	0.27
7	4.98	0.716	0.57
8	5.33	0.824	0.93
9	5.80	0.932	1.49

We then determine the quantiles of the standard normal distribution that correspond to the values of p for these data. For example, suppose we want to find a value z such that $P[Z < z] = 0.5$, the median of the standard normal distribution. We see from Table Z that $z = 0$ give the correct probability. For $p = 0.932$, we find that $z = 1.49$ gives close to the correct probability. We can similarly find the values of z for the other values of p, giving the last column in Table 6.1. The final step is then to plot the ordered data vs. the normal quantiles (Fig. 6.15). If the data are normally distributed, there should be a linear relationship between the observed data and the normal quantiles, and the normal quantile plot will be a diagonal line. This appears to be the case for these data. If the data are non-normal, however, all manner of curved relationships are possible.

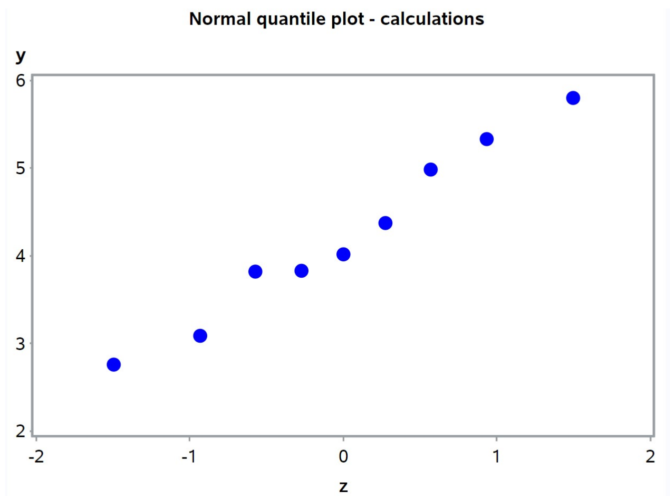

Figure 6.15: Normal quantile plot using Table 6.1

6.4.1 Elytra lengths - SAS demo

We previously examined a data set involving the elytra lengths of male and female *T. dubius* beetles and calculated various descriptive statistics using `proc univariate` (see Chapter 3). We now examine whether these data are normally-distributed using normal quantile plots. A normal quantile plot is requested by adding the command `qqplot` with the `normal` option to the program (see below). A histogram and fitted normal curve can also be generated using the `histogram` command with the `normal` option. Separate analyses are requested for male and female beetles using a `class` statement, because the two sexes differ in size and could also have potentially different distributions. We observe that the normal quantile plots for female beetles is close to linear, suggesting a normal distribution, while the males show some curvature.

SAS Program

```
* normal_quantile_plot.sas;
title 'Fitting the normal to elytra data';
data elytra;
    input sex $ length;
    datalines;
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate plots data=elytra;
    * Separate analyses for each sex;
    class sex;
    var length;
    histogram length/ vscale=count normal;
    qqplot length / normal;
run;
```

```
quit;
```

Fitting the normal to elytra data

The UNIVARIATE Procedure
Variable: length
sex = F

Moments			
N	60	Sum Weights	60
Mean	4.94	Sum Observations	296.4
Std Deviation	0.48544929	Variance	0.23566102
Skewness	-0.521146	Kurtosis	0.16125847
Uncorrected SS	1478.12	Corrected SS	13.904
Coeff Variation	9.82690878	Std Error Mean	0.06267123

Basic Statistical Measures			
Location		Variability	
Mean	4.940000	Std Deviation	0.48545
Median	5.000000	Variance	0.23566
Mode	5.200000	Range	2.20000
		Interquartile Range	0.70000

Figure 6.16: `normal_quantile_plot.sas - proc univariate`

Fitting the normal to elytra data

The UNIVARIATE Procedure
Variable: length
sex = M

Moments			
N	70	Sum Weights	70
Mean	4.71285714	Sum Observations	329.9
Std Deviation	0.44977335	Variance	0.20229607
Skewness	-0.896502	Kurtosis	1.00307174
Uncorrected SS	1568.73	Corrected SS	13.9584286
Coeff Variation	9.5435388	Std Error Mean	0.0537582

Basic Statistical Measures			
Location		Variability	
Mean	4.712857	Std Deviation	0.44977
Median	4.800000	Variance	0.20230
Mode	5.000000	Range	2.40000
		Interquartile Range	0.50000

Figure 6.17: `normal_quantile_plot.sas - proc univariate`

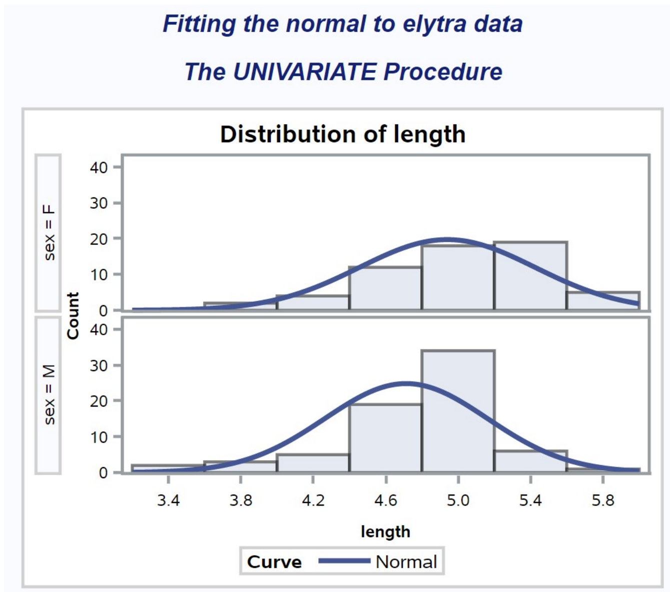

Figure 6.18: `normal_quantile_plot.sas - proc univariate`

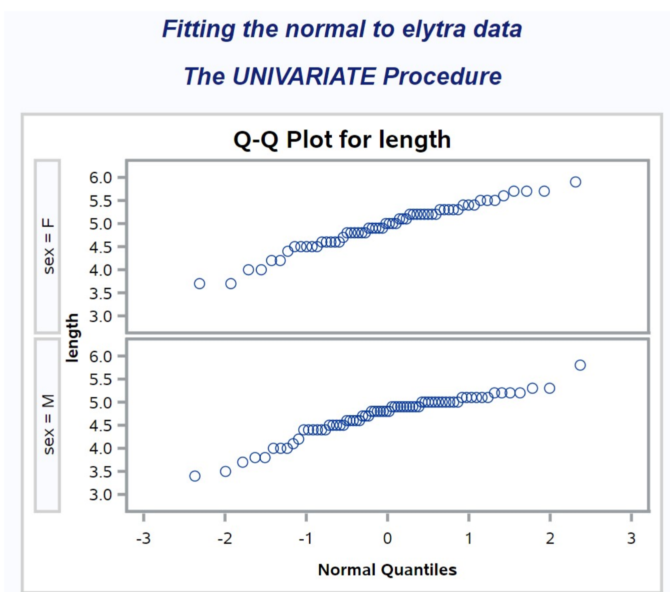

Figure 6.19: `normal_quantile_plot.sas - proc univariate`

6.4.2 Development time - SAS demo

We now examine a data set involving the development time of *T. dubius* beetles in various stages, in particular the time from the larval to prepupal stage, and then from the prepupal to adult stage (Reeve et al. 2003). See program below for details of this analysis. We see that the normal quantile plots for both stages are quite nonlinear, suggesting a distribution different from normal. This is a reflection of the skewed distributions of development time we saw earlier for these data (Chapter 3). Skewed and nonnormal distributions are a common feature of insect development data (Wagner et al. 1984).

SAS Program

```
* normal_quantile_plot_2.sas;
title 'Fitting the normal to development data';
data devel_time;
    input time_pp time_adult;
    datalines;
34  65
31  48
29   .
30  55
32  62

etc.

29   .
29 108
31 103
33   .
29  92
;
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate plots data=devel_time;
    var time_pp time_adult;
    histogram time_pp time_adult / vscale=count normal;
    qqplot time_pp time_adult / normal;
run;
quit;
```

Fitting the normal to development data

The UNIVARIATE Procedure
Variable: time_pp

Moments			
N	96	Sum Weights	96
Mean	31.3541667	Sum Observations	3010
Std Deviation	3.32764866	Variance	11.0732456
Skewness	0.75038358	Kurtosis	0.04666776
Uncorrected SS	95428	Corrected SS	1051.95833
Coeff Variation	10.6130987	Std Error Mean	0.33962672

Basic Statistical Measures			
Location		Variability	
Mean	31.35417	Std Deviation	3.32765
Median	31.00000	Variance	11.07325
Mode	30.00000	Range	14.00000
		Interquartile Range	5.00000

Figure 6.20: `normal_quantile_plot_2.sas - proc univariate`

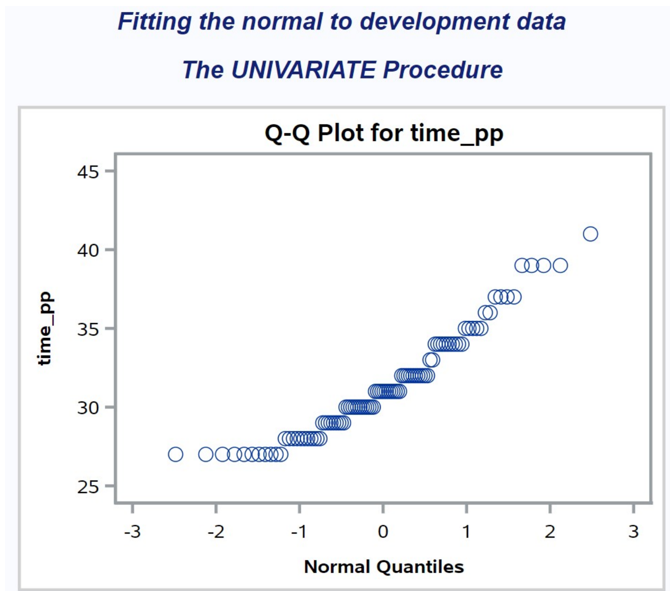

Figure 6.21: `normal_quantile_plot_2.sas - proc univariate`

Fitting the normal to development data

The UNIVARIATE Procedure
Variable: time_adult

Moments			
N	68	Sum Weights	68
Mean	75.3529412	Sum Observations	5124
Std Deviation	26.3465791	Variance	694.14223
Skewness	0.51461555	Kurtosis	-0.6244048
Uncorrected SS	432616	Corrected SS	46507.5294
Coeff Variation	34.9642346	Std Error Mean	3.19499201

Basic Statistical Measures			
Location		Variability	
Mean	75.35294	Std Deviation	26.34658
Median	68.00000	Variance	694.14223
Mode	42.00000	Range	105.00000
		Interquartile Range	46.50000

Figure 6.22: `normal_quantile_plot_2.sas - proc univariate`

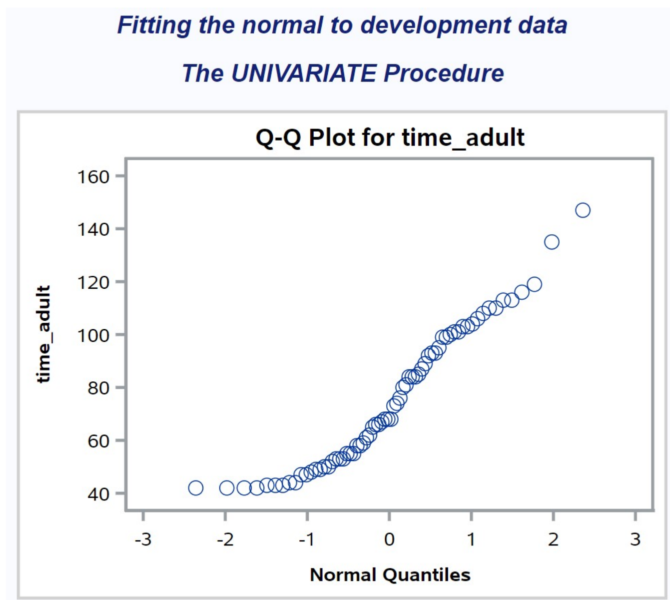

Figure 6.23: `normal_quantile_plot_2.sas - proc univariate`

6.5 References

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

Reeve, J. D., Rojas, M. G. & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

SAS Institute Inc. (2016) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

Wagner, T. L., Wu, H., Sharpe, P. J. H. & Coulson, R. N. (1984) Modeling distributions of insect development time: A literature review and application of the Weibull function. *Annals of the Entomological Society of America* 77: 475-487.

Wilk, M. B. & Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. *Biometrika* 55: 1-17.

6.6 Problems

1. A random variable Y has a uniform probability density with $a = 0$ and $b = 2$.
   (a) What is the expected value of Y, or $E[Y]$? What is the variance of Y, or $Var[Y]$?
   (b) What are the 25%, 50%, and 80% quantiles or percentiles of Y?
   (c) Find the probability that $Y < 0.05$.
   (d) Find a symmetric interval centered around $y = 1$ that has a probability of 0.95.

2. Suppose that Y has a normal distribution with $\mu = 1$ and $\sigma^2 = 3$, or $Y \sim N(1, 3)$. Find the following quantities using Table Z.
   (a) The probability that $Y > 2$.
   (b) The probability that $1 < Y < 3$.
   (c) The probability that $Y < 0.5$.
   (d) The probability that Y is not inside the interval given in b.
   (e) A value of y_0 such that the probability that $Y < y_0$ is 0.9.

3. Suppose that Y has a normal distribution with $\mu = 2$ and $\sigma^2 = 4$, or $Y \sim N(2, 4)$. Find the following quantities using Table Z:
   (a) The probability that $Y < 2.5$.
   (b) The probability that $0.5 < Y < 2.5$.
   (c) The probability that $Y < 1$.
   (d) The probability that Y is not inside the interval given in b.
   (e) A value of y_0 such that the probability that $Y < y_0$ is 0.4.