Chapter 3 : Normal Distribution

Recall in the Chapter 1 notes we looked at an example of acid rain. One of the plots was a relative frequency histogram of the pH levels with a curve overlaid on this histogram. Normal distributions can describe a surprisingly large variety of distributions or histograms observed in the *real world*, but not all types of distributions or real world histograms.

The curve is a member of the normal family of curves (normal probability density functions)

$$f(x) = \frac{1}{\sqrt{2\pi b^2}} e^{-\frac{(x-a)^2}{2b^2}} \tag{1}$$

where a, b are two numbers, also called parameters. The number $\sqrt{2\pi b^2}$ in front of the expression makes the area under the curve equal to 1. This is also what is done for the relative frequency histogram so that the area in the boxes adds up to 1.

In the example in Chapter 1 we use $a = 4.57$ and $b^2 = .0836$ (or equivalently $b = \sqrt{b^2} = 0.289$). Here specifically we used a equal to the sample mean of the data and b^2 equal to the sample variance of the data.

π and e are special and important numbers in mathematics. π is the circumference of a circle with radius 1. e is the so called natural base. Both of these are so called irrational numbers, and to 6 decimal places are

$$\pi = 3.141593 \text{ and } e = 2.718282\ .$$

In your normal arithmetic you use base 10. This means for example that the symbol 123 has a specific interpretation as a number, that is

$$3 * 10^0 + 2 * 10^1 + 1 * 10^2 = 3 + 2 * 10 + 1 * 100$$

Computers use base 2 for internal calculations, but the results are usually printed on the screen in base 10 for ease of human reading. In base 2, the number symbols 100 and 10100 are

$$0 * 2^0 + 0 * 2^1 + 1 * 2^2 (= 4 \text{ in our base 10 symbol})$$

and

$$0 * 2^0 + 0 * 2^1 + 1 * 2^2 + 0 * 2^3 + 1 * 2^4 (= 20 \text{ in our base 10 symbol}) \ .$$

Aside : π-day is celebrated on date 3.14159, that is month 3, day 14, at 1:59 PM.

A normal curve with parameters a and b^2 have the properties

- the curve is centred at a; the curve has mean $= a$
- the curve is spread out (that is width) that corresponds to variance b^2 or equivalently standard deviation $b = \sqrt{b^2}$

This is one of the main reasons that statisticians are interested in variance as a measure of spread of a distribution. It turns out to be the most natural description of spread, but this reasoning is beyond the scope of this course.

In many areas of application of statistics the mean parameter is often written as μ and the variance parameter is written as σ^2. μ is a Greek letter, pronounced *mu* or *mew* and σ is a Greek letter pronounced *sigma.* These two letters correspond to the English letters m and s respectively. We thus sometimes speak about the normal distribution with mean μ and variance σ^2, and write this in shorthand notation as $N(\mu, \sigma^2)$.

For those who have seen integrals these are given by the integrals involving the normal curve given in expression or equation (1)

$$\begin{aligned} a &= \int_{-\infty}^{\infty} x f(x) dx \\ b^2 &= \int_{-\infty}^{\infty} (x-a)^2 f(x) dx \ . \end{aligned}$$

For those who do not know what integrals are, please note the two formula above are not used anywhere in this course.

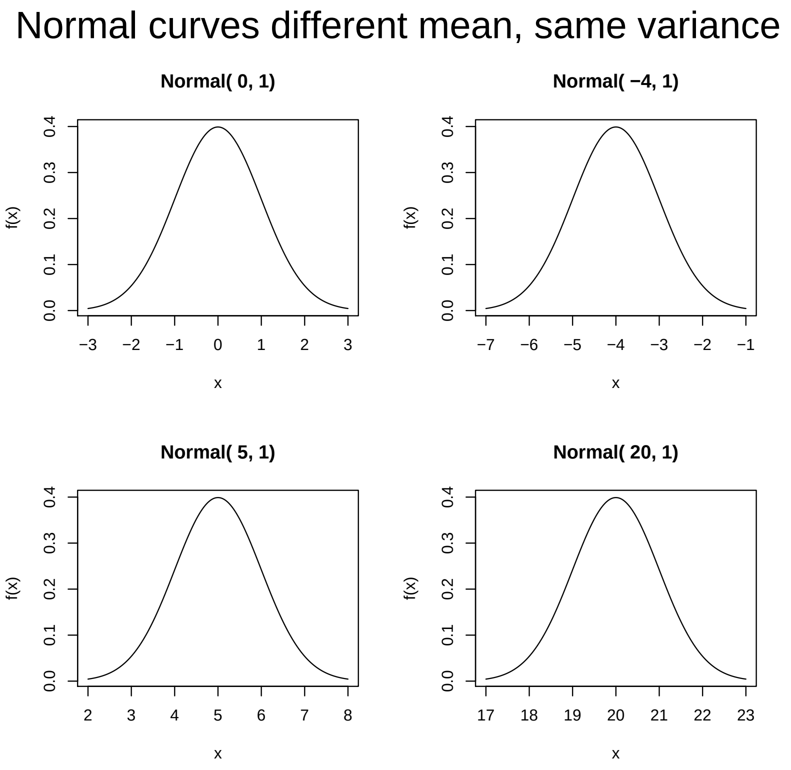

Figure 1: Normal distributions, different variances and same mean

In Figure 1 we see each normal curve is centred at the corresponding mean. Thus the curve does not change shape, only shifts according to its centre.

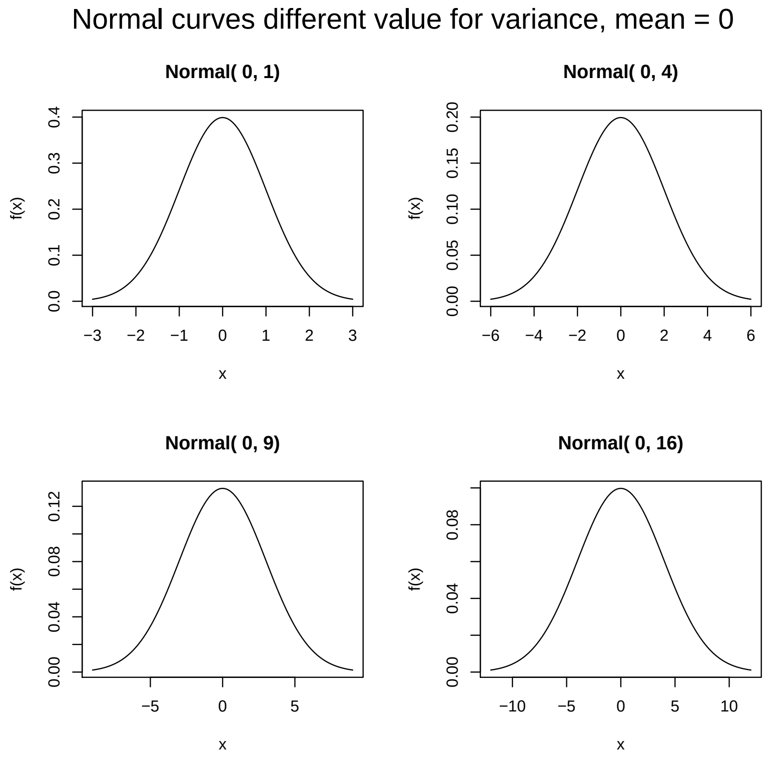

Figure 2: Normal distributions, same variance and different means

In Figure 2 notice that each curve is centred at the same value, zero, but the width of the curve changes. You will see this by looking at the x-axis. Notice the shape is the same type but it is the width that changes. The curve gets *wider* when the variance is larger. When the variance is wider the curve is not as high or tall; see the scale on the y-axis (or the $f(x)$ scale on the plots).

There are some important numbers that often come up for normal curves. These give intervals containing a certain amount of the area under the normal curve

Table 1: Area under curve for various intervals

area or probability	N(0,1) standard normal	Normal mean = a, variance = b^2
.9974	-3, 3	a -3 b , a + 3 b
.95	-1.96, 1.96	a -1.96 b, a + 1.96 b
.90	-1.645, 1.645	a -1.645 b, a + 1.645 b
.683	-1, 1	a -b, a + b

Approximately 99% of the normal area is within 3 standard deviations of the mean, that is in the interval $a \pm 3b$.

Approximately 95% of the normal area is within 2 standard deviations of the mean, that is in the interval $a \pm 2b$.

Approximately 68% of the normal area is within one standard deviation of the mean, that is in the interval $a \pm b$.

See Table A (page 684 in the 4-th Edition) which tabulates the area under the standard normal curve.

If Z has a standard normal distribution Table A gives values

$$Pr(Z \leq z)$$

for various values of z.

For example

$$\begin{aligned} Pr(Z \leq -3.0) &= .0013 \\ Pr(Z \leq -2.0) &= .028 \\ Pr(Z \leq -1.96) &= .025 \\ Pr(Z \leq 1.96) &= .975 \\ Pr(Z \leq 2.0) &= .972 \\ Pr(Z \leq 3.0) &= .9987 \end{aligned}$$

From this Table we can calculate for example

$$Pr(a < Z \leq b) = Pr(Z \leq b) - Pr(Z \leq a)$$

Thus

$$Pr(-1.96 < Z \leq 1.96) = Pr(Z \leq 1.96) - Pr(Z \leq -1.96) = .975 - .025 = .95$$

and

$$Pr(-3.0 < Z \leq 3.0) = Pr(Z \leq 3.0) - Pr(Z \leq -3.0) = .9987 - .0013 = .9974$$

If X is a random variable with a Normal distribution with mean μ and variance σ^2 (equivalently standard deviation $\sigma = \sqrt{\sigma^2}$) then

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution.

Using the table top measurement normal approximation (see the data and histogram in the Chapter 2 handout) we then obtain that about 95% of the observations are between $128.9 - 1.96 * \sqrt{1.17} = 126.8$ and $128.9 + 1.96 * \sqrt{1.17} = 131.0$. This is calculated using only the information that we are using a normal approximation and the values for the mean = 128.9 and variance = 1.17. We do not need to use the entire set of 30 observations for this. The sample first and third quartile are 128.1 and 129.8. Using the normal approximation we would predict that the 25%-tile and 75%-tile are 128.2 and 129.7, almost exactly what we in fact did observe.

We also can state that we are nearly guaranteed (that 99.7% chance) that the next table measurement will be in the interval

$$128.9 \pm 3 * \sqrt{1.17} = [125.66, 132.15] \ .$$

This is quite a powerful statement for an experiment with random outcomes. See the next page for details.

In fact if the model is reasonable we can predict where future observations will fall, at least in the sense of giving probabilities of falling into intervals.

If X represents a table top measurement we approximate the distribution of X as normal, mean $\mu = 128.9$ and variance $= \sigma^2 = 1.17$ ($\sigma = \sqrt{1.17} = 1.081$).

Thus with probability .997 we have

$$\begin{array}{rcccl}
-3 & \leq & \frac{X-128.9}{\sqrt{1.17}} & \leq & 3 \\
-3 & \leq & \frac{X-128.9}{1.081} & \leq & 3 \\
-3 * 1.081 & \leq & X - 128.9 & \leq & 3 * 1.081 \\
-3.245 & \leq & X - 128.9 & \leq & 3.245 \\
128.9 - 3.245 & \leq & X & \leq & 128.9 + 3.245 \\
125.66 & \leq & X & \leq & 132.15
\end{array}$$

At home verify that X falls between 126.8 and 131.0 with probability 0.95.