Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

3.5 The normal distribution

The **normal distribution** is by far the most important probability distribution in statistics. This is for three broad reasons.

- Many variables have distributions which are *approximately* normal, for example heights of humans or animals, and weights of various products.
- The normal distribution has extremely convenient mathematical properties, which make it a useful default choice of distribution in many contexts.
- Even when a variable is not itself even approximately normally distributed, functions of several observations of the variable ('sampling distributions') are often approximately normal, due to the **central limit theorem** (covered in Section 5.5). Because of this, the normal distribution has a crucial role in statistical inference. This will be discussed later in the course.

The equation of the normal distribution curve is:

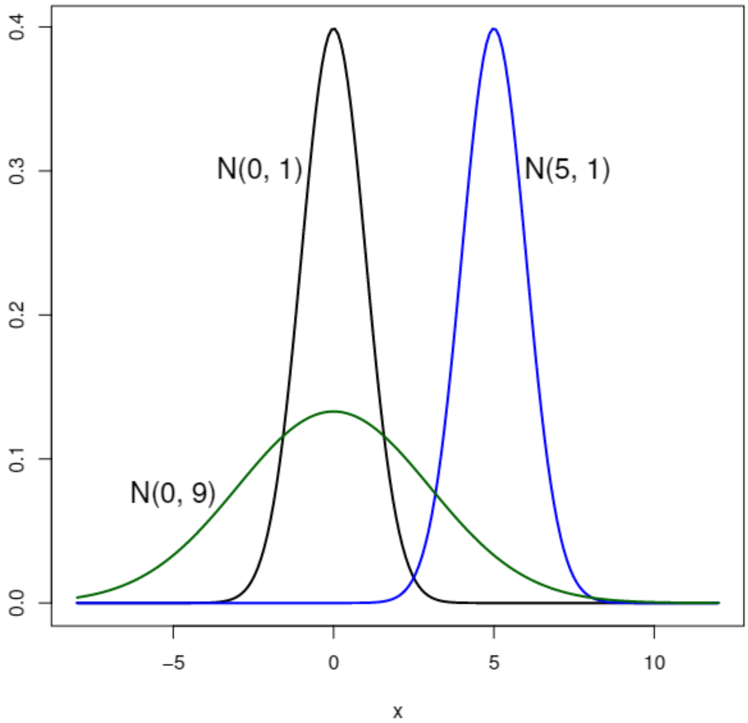
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{for } -\infty < x < \infty$$

where π is the mathematical constant (i.e. $\pi = 3.14159\ldots$), and μ and σ^2 are parameters, with $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

A random variable X with this function is said to have a normal distribution with a mean of μ and a variance of σ^2, denoted $X \sim N(\mu, \sigma^2)$. The mean can also be inferred from the observation that the normal distribution is *symmetric* about μ, which also implies that the median of the normal distribution is μ.

- **The mean μ determines the location of the curve.**
- **The variance σ^2 determines the dispersion (spread) of the curve.**

The figure below shows three normal distributions with different means and/or variances.



- $N(0, 1)$ and $N(5, 1)$ have the same dispersion but different location: the $N(5, 1)$ curve is identical to the $N(0, 1)$ curve, but shifted 5 units to the right.
- $N(0, 1)$ and $N(0, 9)$ have the same location but different dispersion: the $N(0, 9)$ curve is centered at the same value, 0, as the $N(0, 1)$ curve, but spread out more widely.

Linear transformations of the normal distribution

We now consider one of the convenient properties of the normal distribution. Suppose X is a random variable, and we consider the **linear transformation $Y = aX + b$**, where a and b are constants.

Whatever the distribution of X, if it has a mean of μ and a variance of σ^2 then it is true that Y has a mean of $a\mu + b$ and a variance of $a^2\sigma^2$.

Furthermore, if X is *normally* distributed, then so is Y. In other words, if $X \sim N(\mu, \sigma^2)$, then:

$$\boldsymbol{Y = aX + b \sim N(a\mu + b,\, a^2\sigma^2).} \tag{1}$$

This type of result is *not* true in general. For other families of distributions, the distribution of $Y = aX + b$ is *not always* in the same family as X.

Let us apply (1) with $a = 1/\sigma$ and $b = -\mu/\sigma$, to get:

$$\boldsymbol{Z = \frac{1}{\sigma}\, X - \frac{\mu}{\sigma} = \frac{X - \mu}{\sigma} \sim N\left(\frac{1}{\sigma}\,\mu - \frac{\mu}{\sigma},\ \left(\frac{1}{\sigma}\right)^2 \sigma^2\right) = N(0,\ 1).}$$

The transformed variable $Z = (X - \mu)/\sigma$ is known as a **standardised variable** or a ***z*-score**.

The distribution of the z-score is $N(0,\,1)$, i.e. the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ (and, therefore, a standard deviation of $\sigma = 1$). This is known as the **standard normal distribution**.

The figure below shows tail probabilities for the standard normal distribution. The shaded areas are $P(Z \leq -z) = P(Z \geq z)$, by symmetry of the distribution about zero.