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

Module Leader: Dr James Abdey

3.6 Variance of random variables

One very important average associated with a distribution is the expected value of the square of the deviation of the random variable from its mean, μ. This can be seen to be a measure – not the only one, but the most widely used by far – of the dispersion of the distribution and is known as the variance of the random variable. We distinguish between two different types of variance:

- the **sample variance**, S^2, which is a measure of the dispersion in a sample dataset
- the **population variance**, $\text{Var}(X) = \sigma^2$, which reflects the variance of the whole population, i.e. the variance of a probability distribution.

We have previously defined the sample variance as:

$$\boldsymbol{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.}$$

In essence, this is simply an average. Specifically, the average squared deviation of the data about the sample mean.[1] We define the population variance in an analogous way, i.e. we define it to be the average squared deviation about the population mean.

Recall that the population mean is a **probability-weighted average**:

$$\boldsymbol{\mathrm{E}(X) = \sum_{i=1}^{N} x_i\, p(x_i).}$$

The concept of a probability-weighted average (or expected value) can be extended to *functions* of the random variable. If X takes the values $x_1, x_2, \ldots, x_N$ with corresponding probabilities $p(x_1), p(x_2), \ldots, p(x_N)$, then:

$$\mathrm{E}\left(\frac{1}{X}\right) = \sum_{i=1}^{N} \frac{1}{x_i}\, p(x_i) \quad \text{for all } x_i \neq 0$$

[1]The division by $n-1$, rather than by n, ensures that the sample variance estimates the population variance correctly *on average* – known as an 'unbiased estimator'.

and:

$$\mathrm{E}(\ln(X)) = \sum_{i=1}^{N} \ln(x_i)\, p(x_i) \quad \text{for all } x_i > 0$$

also:

$$\mathrm{E}(X^2) = \sum_{i=1}^{N} x_i^2\, p(x_i).$$

So, if we consider the function $(X-\mu)^2$, i.e. the squared deviation about the population mean, the expectation of this (its probability-weighted average) is:

$$\boldsymbol{\sigma^2 = \mathrm{Var}(X) = \mathrm{E}((X-\mu)^2) = \sum_{i=1}^{N} (x_i-\mu)^2\, p(x_i)}$$

and this represents the **dispersion of a (discrete) probability distribution**.

Example

Returning to the example of a fair die, we had the following probability distribution:

$X = x$	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

We now compute the mean and variance of X as follows.

$X = x$	1	2	3	4	5	6	Total
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6	1
$x\,P(X = x)$	1/6	2/6	3/6	4/6	5/6	6/6	$21/6 = 3.5 = \mu$
$(x-\mu)^2$	25/4	9/4	1/4	1/4	9/4	25/4	
$(x-\mu)^2\,P(X = x)$	25/24	9/24	1/24	1/24	9/24	25/24	$70/24 = 2.92$

Hence $\mu = \mathrm{E}(X) = 3.5$, $\sigma^2 = \mathrm{E}((X-\mu)^2) = 2.92$ and hence the standard deviation is $\sigma = \sqrt{2.92} = 1.71$.

Probabilities for any normal distribution

Consider a normal distribution $X \sim N(\mu, \sigma^2)$, for any μ and σ^2. What if we want to calculate, for any $a < b$, $P(a < X \leq b)$?

Remember that:

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

If we apply this **transformation** to all parts of the inequalities, we get:

$$\begin{aligned} P(a < X \le b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} < Z \le \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

where $\Phi(k) = P(Z \le k)$ for some value k and is known as a cumulative probability. (Note that this also covers the cases of the one-sided inequalities $P(X \le b)$, with $a = -\infty$, and $P(X > a)$, with $b = \infty$.) This process is known as **standardisation**.

Example

Let X denote the diastolic blood pressure of a randomly selected person in England. This is approximately distributed as $X \sim N(74.2,\, 127.87)$.

Suppose we want to know the probabilities of the following intervals:

- $X > 90$ (high blood pressure)
- $X < 60$ (low blood pressure)
- $60 \le X \le 90$ (normal blood pressure).

These are calculated using standardisation with $\mu = 74.2$, $\sigma^2 = 127.87$ and, therefore, $\sigma = 11.31$. So here:

$$\frac{X - 74.2}{11.31} = Z \sim N(0,\, 1)$$

and we can determine values of this standardised variable either from statistical tables or (more conveniently) from a computer.

$$\begin{aligned} P(X > 90) &= P\left(\frac{X-74.2}{11.31} > \frac{90-74.2}{11.31}\right) \\ &= P(Z > 1.40) \\ &= 1 - \Phi(1.40) \\ &= 1 - 0.9192 \\ &= 0.0808 \end{aligned}$$

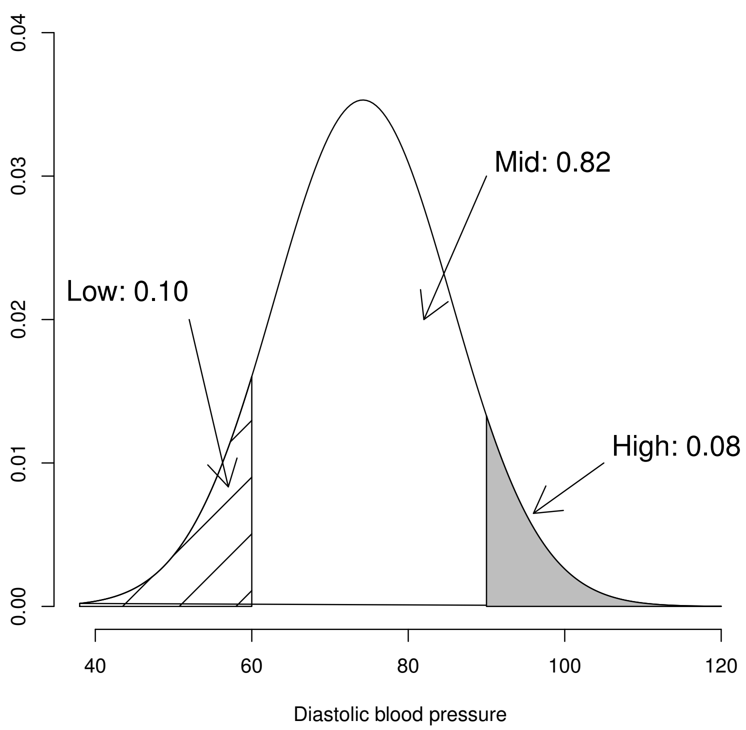
and:

$$
\begin{aligned}
P(X<60) &= P\left(\frac{X-74.2}{11.31} < \frac{60-74.2}{11.31}\right) \\
&= P(Z<-1.26) \\
&= P(Z>1.26) \\
&= 1-\Phi(1.26) \\
&= 1-0.8962 \\
&= 0.1038.
\end{aligned}
$$

Finally:

$$P(60 \le X \le 90) = P(X \le 90) - P(X < 60) = 0.8152.$$

These probabilities are shown in the figure below.



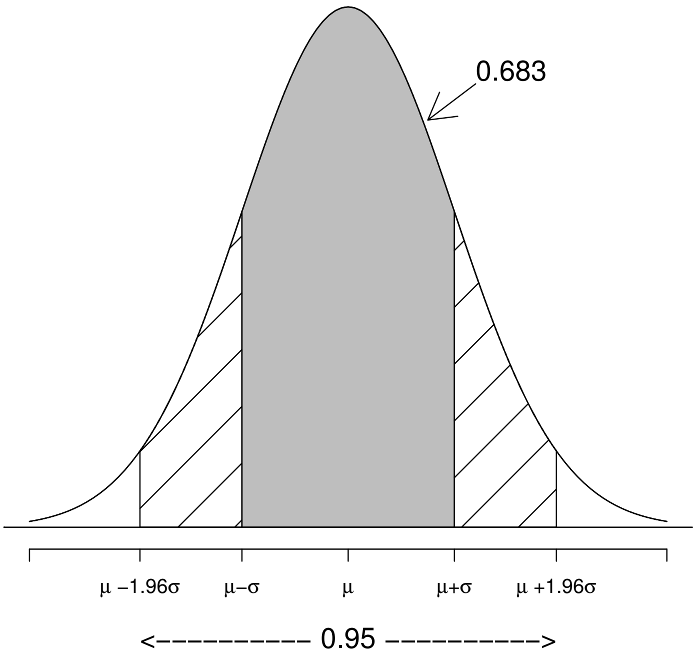
Some probabilities around the mean

The following results hold for all normal distributions.

- $P(\mu-\sigma < X < \mu+\sigma) = 0.683$. In other words, about 68.3% of the total probability is within 1 standard deviation of the mean.
- $P(\mu - 1.96\times\sigma < X < \mu + 1.96\times\sigma) = 0.950$.
- $P(\mu - 2\times\sigma < X < \mu + 2\times\sigma) = 0.954$.

- $P(\mu - 2.58 \times \sigma < X < \mu + 2.58 \times \sigma) = 0.990$.
- $P(\mu - 3 \times \sigma < X < \mu + 3 \times \sigma) = 0.997$.

The first two of these are illustrated graphically in the figure below.



Of course, when dealing with a standard normal distribution, $N(0,\, 1)$, where $\mu = 0$ and $\sigma = 1$, we have:

$$P(-1 \leq Z \leq 1) \approx 0.683$$

$$P(-2 \leq Z \leq 2) \approx 0.950$$

$$P(-3 \leq Z \leq 3) \approx 0.997.$$

Hence, on a standardised basis, it is very easy to determine whether a value is 'extreme', as only 5% of the time would a standardised value be expected to be beyond ± 2 (which we could classify as an **outlier**), and only 0.3% of the time beyond ± 3 (which we could classify as an **extreme outlier**). Values beyond four standard deviations from the mean (i.e. beyond ± 4 on a standardised scale) could be considered as **black swan events**.