8 Statistics Solutions

1) A continuous probability density function has the form:

\[ p(x) \]

\[ x^3 \]

a) What is the value of the probability density function when \( x = 2 \)?

We can write this probability density function as: \( p(x) = A(x - 1) \), since it is a straight line which has a value of zero when \( x = 1 \). The total area under the probability density function is therefore:

\[
\int_1^3 A(x - 1) \, dx = A \left[ \frac{x^2}{2} - x \right]_1^3 = A \left\{ \frac{9}{2} - 3 - \frac{1}{2} + 1 \right\} = 2A
\]

However, the area under any probability density function must be one – it always is. Therefore \( A = \frac{1}{2} \), and the probability density function is \( p(x) = (x - 1)/2 \). Therefore, when \( x = 2 \), \( p(x) = 0.5 \).

Of course, there’s a much simpler way of doing this that avoids the need to do any integration: the area of a triangle is half the base times the height. The base of this one is of length two, and the area is one, so the height must be one. Where \( x = 2 \), \( p(x) \) is half-way up the height, so it must be at 0.5.

b) What is the mean value of this distribution? The mode? The median?

I’ll take the easiest one first: the mode is three. The mode is the most likely value, which means the value of \( x \) where \( p(x) \) is the greatest. That’s at \( x = 3 \).

The mean can be determined by the standard formula for a mean:

\[
\int_{-\infty}^{\infty} x \, p(x) \, dx = \int_{1}^{3} x \left( \frac{x - 1}{2} \right) \, dx = \int_{1}^{3} \frac{x^2}{2} - \frac{x}{2} \, dx
\]

\[
= \left[ \frac{x^3}{6} - \frac{x^2}{4} \right]_1^3 = \left( \frac{27}{6} - \frac{9}{4} - \frac{1}{6} + \frac{1}{4} \right) = \frac{7}{3}
\]

and the median by the standard formula for the median:
The negative result makes no sense, since the probability distribution runs from one to three only, so we take the positive result: the median is at \(1 + \sqrt{2} = 2.4142\ldots\)

There’s an easier way to do this one as well, if you can spot it. Consider the diagram below:

The median divides the probability density function in two – with the size of the triangle on the left equal to half the size of the whole probability density function (the large triangle). Since the small triangle and the large triangle are the same shape, for one to have half the area of the other, the linear dimensions (the base and height) must be smaller by a factor of \(\sqrt{2}\). So the base is of length \(2 / \sqrt{2} = \sqrt{2}\), and therefore the median is at \(1 + \sqrt{2}\).

c) Derive the equation of the CDF (cumulative probability distribution) for this probability distribution, and plot the result.

The CDF is the integral from the start of the probability density function to the value \(x\), and is given by:

\[
P(x) = \int_{-\infty}^{x} p(t)dt = \int_{-\infty}^{x} \frac{(t-1)}{2} dt = \left[ \frac{t^2}{4} - \frac{t}{2} \right]_{-\infty}^{x}
= \left( \frac{x^2}{4} - \frac{x}{2} + \frac{1}{4} \right) = \frac{1}{4} \left( x^2 - 2x + 1 \right)
\]

This is a quadratic form, and looks like this:
The final value is one, as is always the case with CDFs.

2) Consider the weighted die described in section 8.2.3. What’s the standard deviation of this distribution?

We already know what the mean is – that’s 22/7. The standard deviation is the square root of the variance, and the variance is the mean of the squares of the distribution minus the square of the mean, so all we need to know is the mean of the squares. That’s easy enough to calculate:

\[
\overline{i^2} = \sum_{i=1}^{6} p(i)i = \frac{2}{7} 1^2 + \frac{1}{7} 2^2 + \frac{1}{7} 3^2 + \frac{1}{7} 4^2 + \frac{1}{7} 5^2 + \frac{1}{7} 6^2 \\
= \frac{92}{7} = 13.1429...
\]

Which interestingly is exactly 10 greater than the mean.

The variance is therefore \( \frac{92}{7} - \left(\frac{22}{7}\right)^2 = \frac{160}{49} \),

and the standard deviation is just the square root of this: \( \sigma = \sqrt{\frac{160}{49}} \approx 1.807 \)

3) Another weighted die always comes up with either a one or a six, and never a two, three, four or five. What is the difference between the mean and median of this die, and the mean and median of a fairly-weighted die which is equally likely to end up with any of the six sides pointing up?

There’s no difference at all. The mean and median of both of these dice are the same. It’s obvious (I hope) that the medians are the same, and it can be easily shown that the means are the same too:

For the fair die: \( \text{mean} = \frac{1}{6} 1 + \frac{1}{6} 2 + \frac{1}{6} 3 + \frac{1}{6} 4 + \frac{1}{6} 5 + \frac{1}{6} 6 = \frac{21}{6} = 3.5 \)

For the unfair die: \( \text{mean} = \frac{1}{2} 1 + \frac{1}{2} 6 = \frac{7}{2} = 3.5 \)

What is the difference in the standard deviation of the results from rolling these two dice?
We would expect the standard deviation to be different, as the unfair die gives more results further away from the mean. To work them out exactly, we’ll need the second moment of these distributions (the mean value of the square of the die rolls):

For the fair die:

\[ x^2 = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{9}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{25}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{91}{6} \]

For the unfair die:

\[ x^2 = \frac{1}{2} + \frac{1}{2} + \frac{36}{2} = \frac{37}{2} \]

So the two standard deviations (the square root of the variance) are:

For the fair die:

\[ \sigma = \sqrt{\frac{91}{6} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{91}{6} - \frac{49}{4}} \approx 1.7078 \]

For the unfair die:

\[ \sigma = \sqrt{\frac{37}{2} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{37}{2} - \frac{49}{4}} = 2.5 \]

and as expected, the unfair die has a greater standard deviation.

4) Prove that for any probability distribution that is symmetrical about the median, the mean and median are equal. What can be said about the mode in these cases?

It’s not quite as simple as you might think, this one. Consider a symmetric distribution with a median value of \( m \):

Expressing the symmetry in mathematical terms, we can write \( p(2m - x) = p(x) \).

Then calculate the mean:

\[
\text{mean} = \int_{-\infty}^{\infty} x \, p(x) \, dx = \int_{-\infty}^{m} x \, p(x) \, dx + \int_{m}^{\infty} x \, p(x) \, dx
\]

We can substitute in for the second integral using \( t = 2m - x \), to get:
\[
\text{mean} = \int_{-\infty}^{\infty} x \, p(x) \, dx + \int_{m}^{\infty} (2m - t) \, p(2m - t) \, dt \\
= \int_{-\infty}^{m} x \, p(x) \, dx + \int_{-\infty}^{m} (2m - t) \, p(t) \, dt \\
= \int_{-\infty}^{m} x \, p(x) \, dx - \int_{-\infty}^{m} t \, p(t) \, dt + 2m \int_{-\infty}^{m} p(t) \, dt \\
= 2m \int_{-\infty}^{m} p(t) \, dt 
\]

Now this integral, from minus infinity to the median, must be equal to one-half by definition of the median (since half the time the random variable has a value less than one-half). Therefore:

\[
\text{mean} = 2m \int_{-\infty}^{m} p(t) \, dt = 2m \frac{1}{2} = m 
\]

and the result is proved.

As to the mode: all you can say is that there are either an even number of modes (equal numbers on both sides of the median), or there are an odd number of modes, one of which is equal to the mean and the median. Only if you are told that there is only one mode can you say that it must be equal to the mean and the median.

5) Every day, suppose you leave your house, and start taking steps, each step being in a different, entirely random direction. After a very large number of steps, how far are you away from the house? That’s a statistical variable, and it has a Rayleigh distribution, which has the form:

\[
p(x) = \frac{2x}{r^2} \exp\left(-\frac{x^2}{r^2}\right) 
\]

where \( r \) is the rms (root mean square) value of the distribution. What’s the median value?

Standard formula for the median:

\[
\int_{-\infty}^{m} p(x) \, dx = \frac{1}{2} \\
\int_{0}^{m} \frac{2x}{r^2} \exp\left(-\frac{x^2}{r^2}\right) \, dx = \frac{1}{2} 
\]

Note that I’m integrating from zero here. The Rayleigh distribution is a distribution of positive numbers only. It’s impossible to end up a negative distance from your home. This integral is pretty easy if you make the substitution \( t^2 = x^2 / r^2 \):

\[
\int_{0}^{m} \frac{2 \sqrt{x}}{r^2} \exp\left(-\frac{x^2}{r^2}\right) \, dx = \frac{1}{2} 
\]

and since this must be equal to one half for the median:
How about the mode?

The mode is the most likely value, so it’s where the probability density function has a maximum:

\[
\frac{dp(x)}{dx} = \frac{2}{r^2} \exp \left( -\frac{x^2}{r^2} \right) + \frac{2x}{r^2} \exp \left( -\frac{x^2}{r^2} \right) \left( -\frac{2x}{r^2} \right) = 0
\]

\[
\frac{2}{r^2} \exp \left( -\frac{x^2}{r^2} \right) = \frac{2x}{r^2} \exp \left( -\frac{x^2}{r^2} \right) \left( \frac{2x}{r^2} \right)
\]

\[
1 = x \left( \frac{2x}{r^2} \right)
\]

\[
x = \frac{r}{\sqrt{2}}
\]