2. Continuous Probability Distributions

1. Introduction

Last week we discussed the idea of a discrete random variable and its associated probability function, using the binomial distribution as an example.

Today I’d like to move on to the idea of a continuous random variable. As we’ll see, certain problems and subtleties arise when we start thinking about continuous variables. I will point out some of these subtleties and give some indication as to how they are resolved. I’ll then discuss the most important continuous distribution in statistics and experimental science— the normal, or Gaussian—distribution and explain (at least to some degree) where it comes from.

2. Continuous Random Variables

We have already met the idea of a discrete random variable, for example the sum of the numbers on two dice. Such a variable takes on a certain number of values, and we can associate a specific probability with each such value.

**Question:** Can a discrete random variable take on an infinite number of values?

Let’s now consider a continuous quantity such as a person’s height. For definiteness, let’s say we are measuring the heights of a group of N people, where N is a large number, and that their heights all lie within the range 1.5m and 2.0m.

Our continuous random variable, r, is height. This variable can take on any values between 1.5 and 2.0. In the case of discrete random variables we were able to associate a definite probability with every possible value of the random variable. This is more tricky with our continuous variable r. For example, what is the probability that r takes on the value 1.8?

To put it another way, what fraction of our group of N people are exactly 1.8m tall?

A moment’s thought should show that we are never going to find someone exactly 1.8m tall although there may be plenty of people of approximately this height. Therefore, we cannot assign probabilities to specific values of r, as we did in the discrete case.

In a way, the question is meaningless, since we can’t in any case measure heights to arbitrary accuracy. Let’s suppose we can measure a height to an accuracy of 1mm. It then makes more sense to ask: What is the probability that r lies between 1.799 and 1.801. It would seem that we should be able to find a sensible answer to this question, since if our group is large enough there must be some people with heights in this range.

However, it is not obvious how to calculate the answer (except in one simple case).

**Question:** What is the probability that r lies between 1.2 and 1.8?

In the discrete case, we found the probability that a random variable lies in a specific range by adding up the probabilities for each value in that range. For example the probability that s, the sum of two dice lies between 4 and 8 is given by:

\[ p(4 \leq s \leq 8) = p(4) + p(5) + p(6) + p(7) + p(8) \]
We can’t do this in the case of our continuous variable \( r \), since we can’t define the individual probabilities. In order to make progress, we need to approach the problem from a different point of view.

3. Probabilities as Areas

Let’s go back to the binomial distribution. Figure (2.1) shows a histogram of the binomial probability function:

![Histogram of Binomial Probability Function](image)

**Figure 2.1:** The binomial probability function for \( n = 5 \) and \( p = 0.5 \)

The histogram has \( n \) bars and the bars have been arranged to be one unit wide, so that each probability is numerically equal to the area of its bar.

This observation is not of great significance in the context of a discrete distribution, since we can obtain the probability for any value of the random variable directly from the probability function. However it offers a hint as to how to define probabilities for continuous random variables. Let’s call the binomial random variable \( x_i \) and define a new random variable by

\[
y_i = \frac{x_i}{n}
\]

**Question:** What is the mean of \( y_i \)? What is its variance?

This variable ranges between 0 and 1 and has the same probability function as \( x_i \) since we get it from \( x_i \) by just dividing by a constant:

\[
p(y_i) = p(x_i)
\]

However the width of each bar is now \( h = 1/n \) and so the area of each bar no longer represents the associated probability. If we want to maintain the notion of area as probability we need to rescale \( p(y_i) \) as well.

If we define a new function:

\[
\rho(y_i) = \frac{p(y_i)}{h} = np(y_i)
\]

§3. Probabilities as Areas
and plot $\rho(y_i)$ against $y_i$, the area of each bar of the new histogram is given by

$$A_i = \rho(y_i) = np(y_i) \times h = np(y_i) \times \frac{1}{n} = p(y_i) = p(x_i)$$

This new function $\rho(x_i)$ is called the \textit{probability density function} (or just the density function) since it the probability “per unit width” along the horizontal axis.

To summarise:

We were previously able to interpret the area under the binomial probability function as a probability, provided we fixed the width of each histogram bar as 1. By rescaling our random variable and its probability function we were able to define a new function (the density function). This function no longer represents the probability; however the area of each bar of our new histogram now automatically equals the probability (we don’t have to arrange for the widths of the bars to be 1).

A plot of the density function corresponding to the distribution (2.1) is shown in Figure (2.2).

![Figure 2.2: The binomial density function for $n = 5$ and $p = 0.5$](image)

Notice that the histogram now ranges from 0 to 1 (instead of from 0 to $n$) and that the heights of the bars have changed; however the \textit{area} of each bar still represents the corresponding probability, as before. Notice also that, unlike the probability function, the density function can take values greater than 1; the only requirement is that the total area of all the histogram bars equals 1.

\section*{4. The Binomial Density for Large $n$}

What does the density function look like when $n$ gets very big? The graph for the case $n = 50$ is shown in Figure (2.3).
Figure 2.3: The binomial density function for $n = 50$ and $p = 0.5$

The horizontal range still stretches from 0 to 1, but there are many more histogram bars and each bar is narrower. The heights of the bars have also increased, however, so that the total area is still equal to 1.

As the number of contributing variables gets bigger and bigger the histogram approximates more and more closely to a smooth bell-shaped curve.

In order to see where this shape comes from, let’s think about a Bernoulli coin-tossing experiment with 50 trials. The random variable $x_i$ is the number of heads which appear.

Now think about the outcome of tossing a single coin. This is a random variable $r_i$ with two possible values, 1 (heads) and 0 (tails). Since we are conducting 50 trials, we have 50 of these variables, which are all independent of each other.

**Question:** What is the mean and variance of each of the $r_i$?

The value of our Bernoulli variable $x_i$ is given by adding the observed values of these 50 independent variables:

$$x_i = \sum_i r_i$$

and the rescaled variable $y_i$ plotted in Figure (2.3) is given by

$$y_i = \frac{1}{n} \sum_i r_i$$

So $y_i$ is just the arithmetic mean of the $n$ independent $r_i$. Let’s call a variable like $y_i$ which is the arithmetic mean of lots of independent variables an additive random variable. (This term is non-standard; I’ve just made it up).

There is an amazing theorem in probability theory (called the central limit theorem), which states, in effect, that if any additive random variable $R$ is the arithmetic mean of lots of independent random variables, $r_i$, then the probability distribution of $R$ looks like Figure (2.3).

Why do I call this amazing? Well, the result doesn’t depend on the probability distributions of the component random variables. Roughly speaking, one can take the arithmetic mean $R$ of any
(large) set of independent random variables and the density function of $R$ will look like Figure (2.3). Moreover, it will look more and more like Figure (2.3) as the number of independent variables increases. So the more random variation we introduce into a system, the more closely the probability density of the arithmetic mean approaches this bell curve!

Here is Francis Galton on the subject:

*I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.*

Francis Galton (“Natural Inheritance”, 1889)

5. The Normal Density

So far we have just talked vaguely about the “bell shape” that the binomial density assumes as $n$ gets large. Figure (2.4) shows the smooth curve that the binomial density approaches in this limit.

![Figure 2.4: The normal density as a limit of the binomial density](image)

The probability that our binomial random variable $y_i$ falls between, say 20 and 30 is given by summing up the areas of all the histogram bars between 20 and 30. For large $n$, this area is nearly the same as the corresponding area under the smooth curve. Let us therefore abstract away the scaffolding that we have used to build our first continuous probability distribution and just look at the curve itself, in Figure (2.5):
2. Continuous Probability Distributions

Figure 2.5: The normal density function

This is our first example of a continuous probability distribution, the celebrated normal or Gaussian distribution. The curve itself is also known as the normal error curve.

The random variable $x$ plotted along the horizontal axis is now a continuous distribution, and the probability that $x$ lies between two values, $a$ and $b$ is given by the area under the curve between $a$ and $b$. So we have solved the problem we started at the beginning: how to determine the probability that the value of a continuous random variable lies in a certain range?

The weird thing about the normal distribution is how it shows up everywhere. The central limit theorem assures us that if a random variable can be expressed as the arithmetic mean of many independent random variables, it will tend to a normal distribution, but it is not at all obvious that this condition is always satisfied.

To take our original example of heights; if you measure the heights of a large population, you will find that the heights are normally distributed. This would make sense if a person’s height was the arithmetic mean of a large number of factors acting independently, but is this in fact the case? Do the factors affecting height act independently? Can their effect be combined in an arithmetic mean? Nevertheless, we find in practice that in a large class of experimental situations, the measurements we make are normally distributed about (presumably) the “true” value of the quantity we are trying to measure.

Since the normal distribution is so ubiquitous, people have a habit of assuming normality even in cases where it isn’t warranted, and many statistical tests make an implicit or explicit assumption of normality. Be careful....

Everybody believes in the exponential law of errors: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation

M. Lippmann (letter to Poincare)

Question: What is the shape of the indentation in the stone under the Front Arch of TCD? Why?
6. Functional Form of the Normal Density

The central limit theorem specifies the exact form of the normal density curve; you have probably seen the formula for it.

In this section, I’ll try to give some motivation for this result. (I’ll need to use a small bit of calculus). Let’s look at the normal density curve again:

![Figure 2.6: The shape of the normal density function](image)

The mean of this density, \( \mu \), is the value of \( x \) at the peak of the distribution.
We want to find the density, \( \rho \) as a function of the random variable \( x \).
We can describe the curve by its slope, \( S \), at a point, so what information do we have available to specify a point on the curve? Well, we can specify its horizontal distance from the mean \( (x - \mu) \) and its vertical distance, \( \rho(x) \); these are the only parameters that we have to work with. We need to express the behaviour of the slope in terms of these two parameters.
What can we say about the slope? Well, it’s flat (zero) in the middle at the peak. We know that \( (x - \mu) \) is zero here, so maybe \( S \) is proportional to \( (x - \mu) \).

\[
S \propto (x - \mu)
\]

However, \( S \) is also small far away from the mean, where \( (x - \mu) \) is large. Fortunately our second parameter \( \rho(x) \) is small here, so we need:

\[
S \propto \rho(x)
\]

Finally, we know that \( S \) is positive to the left of the mean, where \( (x - \mu) \) is negative and negative to the right of the mean, where \( (x - \mu) \) is positive. Putting this all together, we get:

\[
S = -k^2 \rho(x)(x - \mu)
\]

where \( k^2 \) is a positive constant.
Once we know the slope of the curve everywhere, we can recover the curve itself:

**Calculus Alert!**

\[
\frac{d\rho}{dx} = -k^2 \rho(x)(x - \mu) \Rightarrow \int \frac{d\rho}{\rho} = \int -k^2 (x - \mu) dx \Rightarrow \log \rho = -\frac{1}{2} k^2 (x - \mu)^2 + C \Rightarrow \rho = Ae^{-\frac{1}{2} k^2 (x - \mu)^2}
\]

**All Clear**

So the density \( \rho \) is given as a function of \( x \) by:

\[
\rho(x) = Ae^{-\frac{1}{2} k^2 (x - \mu)^2}
\]

(2.1)

Here \( e \) is the natural base of logarithms (\( e \approx 2.718 \)) and the constants \( A \) and \( k^2 \) remain to be determined.

The exponent in (2.1) has to be a dimensionless number, whereas \( (x - \mu) \) has dimensions of whatever \( x \) is measuring. We only have one parameter to play with, \( (x - \mu) \), so \( k^2 \) must depend inversely on the square of this. However, \( k^2 \) is a constant and can’t depend on \( x \), so we have to add up all the values of \( (x - \mu)^2 \) (actually take the weighted average). But this is just the variance, \( \sigma^2 \). Hence:

\[
\rho(x) = Ae^{-\frac{\frac{1}{2} (x - \mu)^2}{\sigma^2}}
\]

(2.2)

The area under the curve (2.2) can be calculated and turns out to be:

\[
\text{area} = \sqrt{2\pi \sigma^2} \times A
\]

Since this area is the total probability, it must be equal to 1, and hence:

\[
A = \frac{1}{\sqrt{2\pi \sigma^2}}
\]

and so we obtain finally:

\[
\rho(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}
\]

The normal density function has two parameters, the mean, \( \mu \), which determines the location of the peak, and the variance, \( \sigma^2 \), which determines its width.

I think I’ll finish there, leaving you with one more thought...

*My own impression . . . is that the mathematical results have outrun their interpretation and that some simple explanation of the force and meaning of the celebrated integral . . . will one day be found . . . which will at once render useless all the works hitherto written.*

Augustus de Morgan, 1838

§6. Functional Form of the Normal Density