Continuous Probability Distributions

When we talked about the probability distribution of the sum of two dice, we assigned a probability to each of the 11 possibilities:
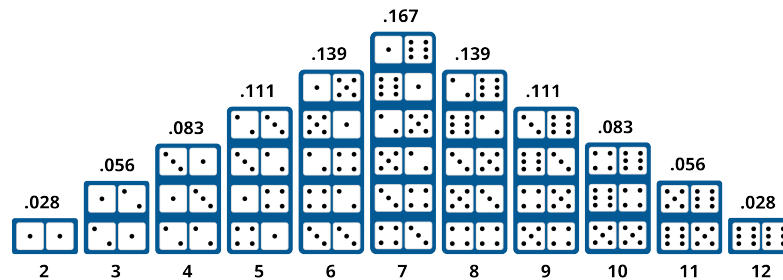

Figure 1.1: Rolling two dice probability distribution.

The probabilities all added up to be 1.0. That is a way of saying "100% of the times you throw the dice, the sum will be an integer between 2 and 12."

Now we need to talk about probabilities of properties that are continuous, not discrete. For example, we might want to ask the question, "If I randomly pick a cow from all the cows in the world, what is the probability that it will weigh less than 597.34 kg?" What does a probability distribution for a continuous variable look like?

1.1   Cumulative Distribution Function

Imagine that you live in ancient times. You buy, sell, and ship cows. A lot of people come into your office and brag about their cows: "Bessie is heavier than 99% of the cows in world!" So, you need to develop some statistics on cow weights.

You have ten cows. You weight them:

Cow	Mass in kg
Cow 1	580.22
Cow 2	540.07
Cow 3	538.20
Cow 4	512.39
Cow 5	589.75
Cow 6	456.91
Cow 7	583.09
Cow 8	493.56
Cow 9	489.97
Cow 10	704.15

If someone comes into your shop and says, "My Bessie is an astonishing 530 kg!" it would be cool to have a list on the wall that would let you yell back, "Half my cows are heavier than that, Silly!" So, you sort the cows by weight. For each weight, you say how many of your cows are lighter than that:

Proportion of cows lighter	Mass in kg
0.00	456.91
0.10	489.97
0.20	493.56
0.30	512.39
0.40	538.20
0.50	540.07
0.60	580.22
0.70	583.09
0.80	589.75
0.90	704.15

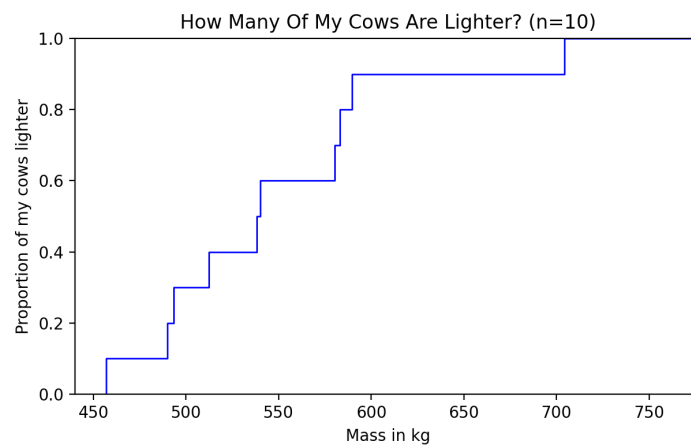In fact, for easy reference, you make a plot:

Figure 1.2: 10 cow cumulative distribution function.

Now, for any weight, you can quickly look up what proportion of your cows are lighter (and, if you subtract that from 1, what proportion of your cows are heavier).

See how jagged that graph is? That is because you only have the data for 10 cows. However, as the years pass and you weight thousands of cows, the plot will become smoother. Because it always accumulates more cows as you move from left to right, this is known as a *cumulative distribution function*, or CDF:
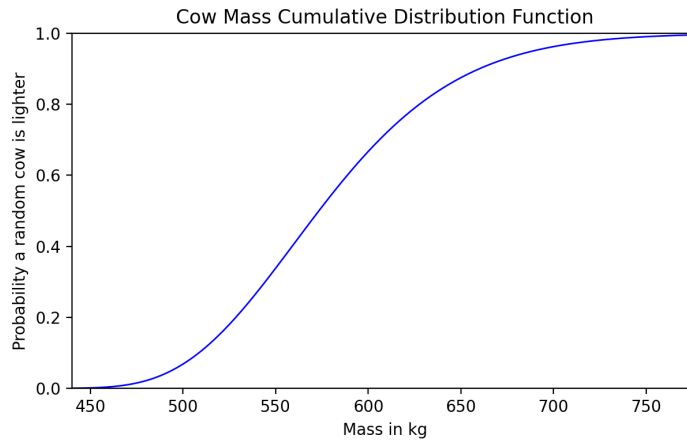

Figure 1.3: Cumulative distribution function for cow weights.

A cumulative distribution function always starts at 0 and ends at 1. On that journey, it never decreases. Reading the graph tells you what proportion of the population is *less than* any given value. If your data is normally distributed, the CDF will look like an S-curve, being symetric closer to the probability 0.5.

Let's say you want to know what proportion of cows weigh between 620 kg and 635 kg. Using the CDF, you could figure out that 77% of all cows weigh less than 620 kg and 83% of all cows weigh less than 635 kg. Thus 6% of all cows must weigh more than 620 kg and less than 635 kg.
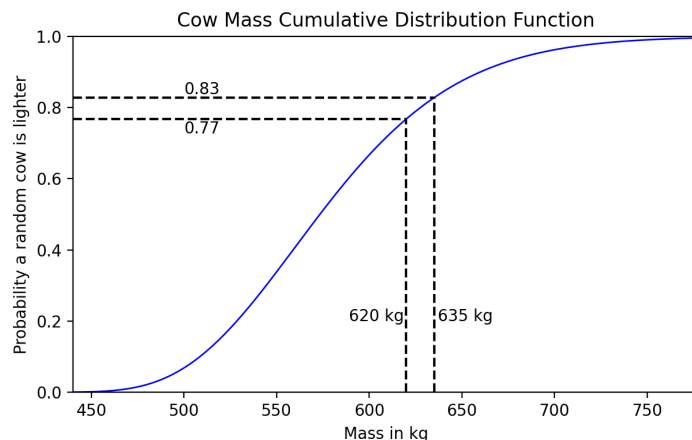
Figure 1.4: Cumulative distribution function with bounds from 620 kg to 635 kg.

1.2   Probability Density Function

The cumulative density function is handy, but some of its information can be hard to see. For example, how would you answer the question, "What is the most common weight of a cow?" You would squint at the CDF and try to determine where it was steepest.

To make these sorts of questions easier to answer, we take the derivative of the CDF to get the *probability density function* (or PDF). For the cows, it would look like this:
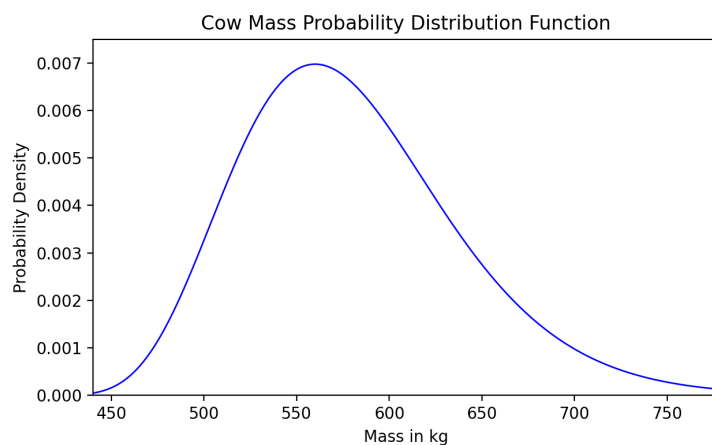
Figure 1.5: Probability density function for cow weights.

Now you can easily see that the CDF was steepest at about 560 kg. We call the highest point on the PDF the *most likely estimator*. For example, you might say "560 kg is the most likely estimator of cow mass." Sometimes we just say "the MLE".

Note that the MLE is often different from the mean or the median. In this case, for example, the distribution is skewed right — there are more cows that are heavier than the MLE than there are cows that are lighter than the MLE. The MLE would be less than the mean or the median.

Once again, let's say you want to know what proportion of cows weight between 620 kg and 635 kg. This is more difficult with a PDF than it is with a CDF. With a PDF, you have to find the area under the curve between $x = 620$ and $x = 635$.
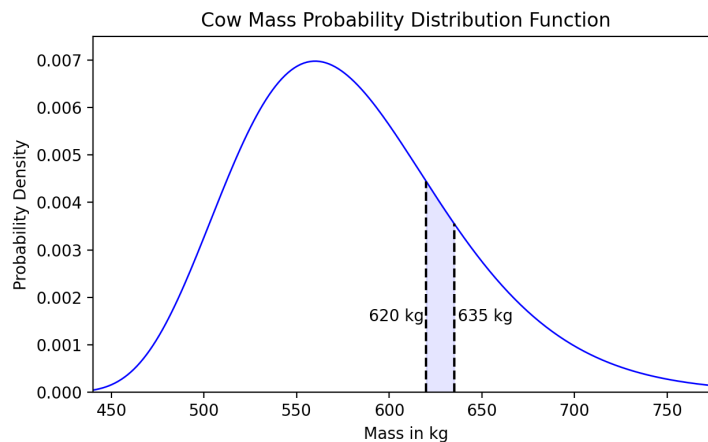

Figure 1.6: Probability density function with the integral from 620 kg to 635 kg.

This is why it is called a "probability density" — to get a true probability you need to multiply the density by the width of the region.

What is the area under the entire curve? If you integrated it, you would get the CDF. The CDF goes from 0 to 1.0. The area under a PDF is *always* 1.0.

1.3   The Continuous Uniform Distribution

The most simple continuous distribution is the uniform distribution between two numbers a and b. The CDF is a straight line from zero at a to 1 at b. For example, here is the CDF for the uniform distribution between 6 and 9:
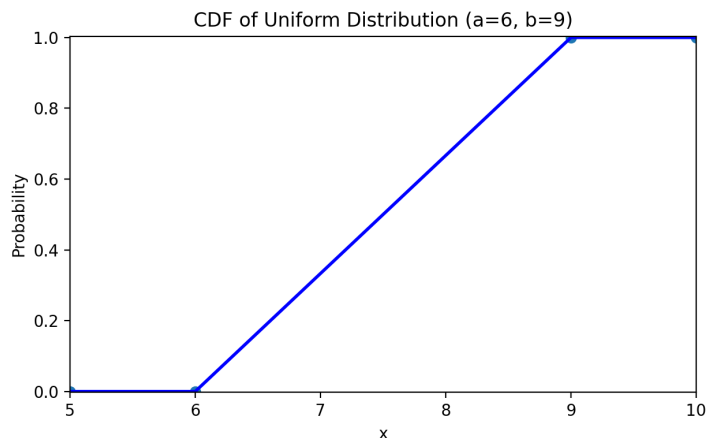
Figure 1.7: Cumulative distribution function for a uniform distribution between 6 and 9.

That line goes from 0 to 1 over a distance of 3, so its slope is $\frac{1}{3}$ between 6 and 9 and zero everywhere else. Thus, the PDF (its derivative) looks like this:
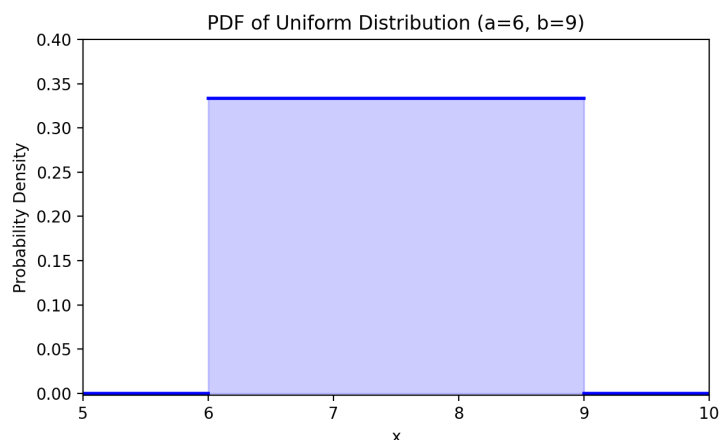

Figure 1.8: Probability density function for a uniform distribution from the continuous uniform distribution.

So, we can write the probability distribution of a continuous uniform distribution between a and b as:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b. \end{cases}$$

Notice that if a and b are less than 1 apart, the value of $p(x)$ will be greater than 1. This is an important difference between a probability and a probability density:

- A probability will always be in the interval $[0, 1]$.

- A probability density will never be less than 0, but can be much larger than 1.

That said, the probability density will always integrate to 1.

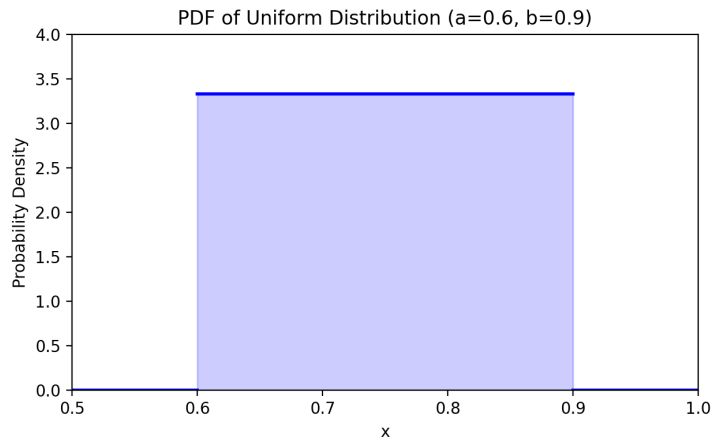Here is the PDF for a uniform distribution between 0.6 and 0.9:

Figure 1.9: Uniform probability density function.

The mean and median of a uniform distribution between a and b is its midpoint: $\frac{a+b}{2}$.

The variance (σ^2) is $\frac{(b-a)^2}{12}$.

1.4   Continuous Distributions In Python

The SciPy library has functions that let a programmer work with a large collection of different probability distributions.

For example, if you wanted to work with a continuous uniform distribution between 6 and 9, you would import the relevant functions like this:

```
from scipy.stats import uniform
```

Now, if you wanted a numpy array containing a sample of 300 numbers generated randomly from that distribution:

```
samples = uniform.rvs(loc=6,  scale=3,  size=300)
```

The `loc` argument is a. The `scale` argument is $b - a$.

If you wanted to know the value of the probability density function at 8 and 10, you could use the `pdf` function:

```
x_values = np.array([8, 10])
p_values  = uniform.pdf(x_values,  loc=6,  scale=3)
```

Now, `p_values` contains 0.33333 and 0.0.

To get the value of the cumulative distribution function at those points, you would use the `cdf` function:

```
cdf_values = uniform.cdf(x_values,  loc=6,  scale=3)
```

Now, `cdf_values` contains 0.666667 and 1.0.

The inverse of the CDF is very useful. It answers questions like "How heavy does a cow have to be to be in top 1%?"

```
bottom_top_percentiles = np.array([0.01,  0.99])
boundaries = uniform.ppf(bottom_top_percentiles,  loc=6,  scale=3)
```

Now, `boundaries` contains 6.03 and 8.97.

The SciPy library supplies these functions (`rvs`, `pdf`, `cdf`, and `ppf`) for over a hundred common continuous probability distributions.

The common "bell curve" shaped distribution is called a Gaussian or Normal distribution. It is described by its mean (the midpoint of the bell) and its standard deviation. For the normal distribution, the standard deviation is the distance you have to go from the mean to reach 68% of the population. We will talk a lot more about the normal distribution in other chapters, but let's take this opportunity to plot the CDF and PDF of a normal distribution with a mean of 32 and a standard deviation of 8.

Create a file called `plot_norm.py` and add the following lines:

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

# Constants
MEAN = 32
STD = 8
```

```python
# Plottin from the 0.5 percentile to the 99.5 percentile
x_min = norm.ppf(0.005, loc=MEAN, scale=STD)
x_max = norm.ppf(0.995, loc=MEAN, scale=STD)

# Make 200 points between x_min and x_max
x_values = np.linspace(x_min, x_max, 200)

# Get CDF for each x value
cdf_values = norm.cdf(x_values, loc=MEAN, scale=STD)

# Get PDF for each x value
pdf_values = norm.pdf(x_values, loc=MEAN, scale=STD)

# What is the highest density?
max_density = norm.pdf(MEAN, loc=MEAN, scale=STD)

# Make a figure with two axes
fig, axs = plt.subplots(nrows=2, sharex=True, figsize=(8, 8), dpi=200)
axs[0].set_xlim(left=x_min, right=x_max)

# Draw the CDF on the first axis
axs[0].set_title("CDF of Normal Distribution (mean=32, std=8)")
axs[0].set_ylim(bottom=0.0, top=1.0)
axs[0].plot(x_values, cdf_values)

# Draw the PDF on the second axix
axs[1].set_title("PDF")
axs[1].set_ylim(bottom=0.0, top=max_density * 1.1)
axs[1].plot(x_values, pdf_values)

# Add lines for mean,  mean-std, and mean+std
axs[1].vlines(MEAN - STD, 0, max_density, "r", linestyle="dashed")
axs[1].vlines(MEAN + STD, 0, max_density, "r", linestyle="dashed")

# Save out the figure
fig.savefig("norm_32_8.png")
```
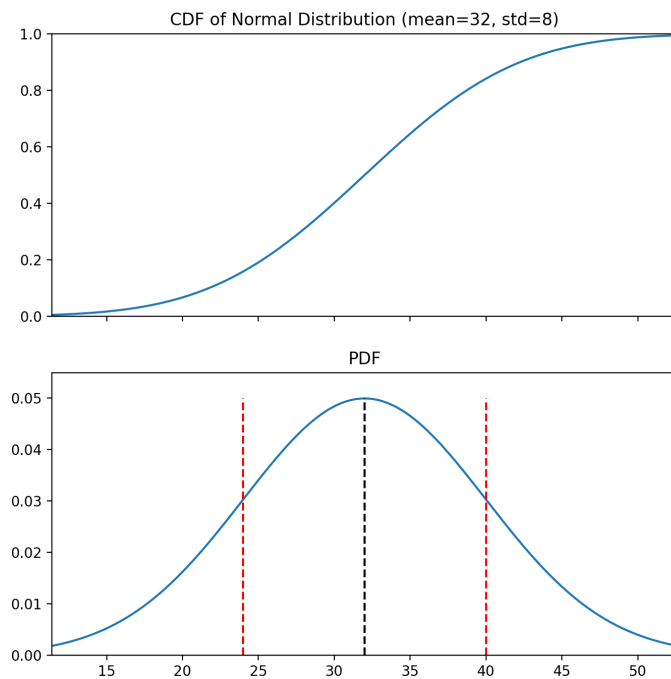
The resulting plot should look like this:

What do those vertical lines mean? An ornithologist might tell you, "The wingspan of adult robins are normally distributed with a mean of 32 cm and a standard deviation of 8 cm." This means 68% of the population of adult robins would have wingspans between the two red lines.

Exercise 1   **SciPy Stats**

Globally, the height of adult women is approximately normally distributed. The mean is 164.7 cm. The standard deviation is 7.1 cm.

Use Python and SciPy stats to answer these questions:

- To be in the tallest decile (the top 10%) of adult women, how tall does one need to be?

- What percentage of adult women are between 160 cm and 165 cm?

(In case you are wondering: For men the mean is 178.4 cm and the standard deviation is 7.6 cm.)

This is a draft chapter from the Kontinua Project. Please see our website (`https://kontinua.org/`) for more details.

Answers to Exercises

Answer to Exercise 1 (on page 11)

```python
from scipy.stats import norm

# Constants
MEAN = 164.7
STD = 7.1

# What is the cutoff for the top decile?
cutoff = norm.ppf(0.9, loc=MEAN, scale=STD);
print(f"To be in the top 10 percent, you must be at least {cutoff:.2f} cm")

# What proportion of women are between 160cm and 165cm?
shorter_than_160 = norm.cdf(160, loc=MEAN, scale=STD)
shorter_than_165 = norm.cdf(165, loc=MEAN, scale=STD)
between = shorter_than_165 - shorter_than_160
print(f"{between * 100.0:.2f}% of adult women are between 160 and 165 cm.")
```

When run, this will give you:

```
> python3 women.py
To be in the top 10 percent, you must be at least 173.80 cm
26.29% of adult women are between 160 and 165 cm.
```

INDEX