Week 2: Probability Distributions as Models

Introductory Statistics Notes

1 From Histograms to Probability Distributions

In Week 1, we described data using *histograms* and summary numbers like the mean $\bar{x}$ and standard deviation s.

Now we take the next step:

- A histogram is about one *sample*.
- A **probability distribution** is a *model* for the whole *population*.

We will think of a probability distribution as a shape that tells us:

- which values are more likely,
- how spread out the values are,
- how to find probabilities like "what is the chance X is between 5 and 8?"

We will look at two main types:

- **Continuous distributions:** smooth curves where probability is measured by area.
- **Discrete distributions:** bar graphs where probability is attached to each value.

Running examples for this week

We will use two running examples:

- **Continuous example (resting heart rate).**
  Let X = resting heart rate (beats per minute) for a randomly chosen student. We will model X by a smooth curve on the number line.
- **Discrete example (weekly tutoring visits).**
  Let Y = number of times a randomly chosen student visits the tutoring center in a week (0,1,2,3,4,...). We will model Y using a probability table.

2 Continuous Distributions as Population Models

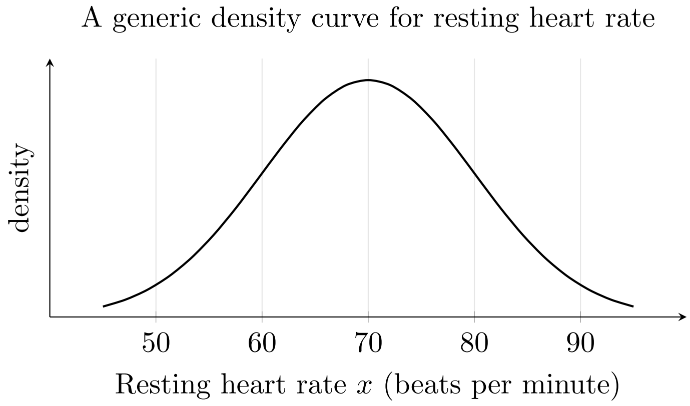
A **continuous probability distribution** is described by a smooth curve that sits on top of the number line.

We will think of the curve as a **population model**:

- The horizontal axis is the possible values of X (for example, heart rate).
- The vertical axis shows how *dense* the population is around those values.
- The **area under the curve** represents probability.

2.1 Density curves and area

A typical density curve for resting heart rate might look like this:


A generic density curve for resting heart rate

Key facts about density curves:

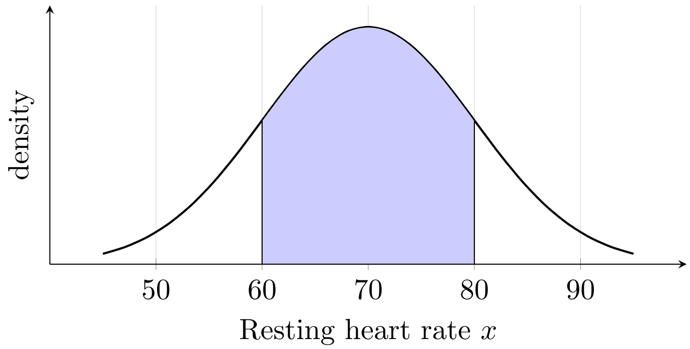
- The curve never goes below the horizontal axis.
- The total area under the curve is 1 (100% of the probability).
- The probability that X falls in an interval is the *area* over that interval.

2.2 Area as probability

If X is resting heart rate, and we want

$$P(60 \leq X \leq 80),$$

we imagine shading the region under the curve from 60 to 80 and reading off the area.



We interpret:

- $P(60 \leq X \leq 80)$ is the proportion of students in the population whose resting heart rates lie between 60 and 80 beats per minute.
- Later, when we study specific distributions (like the normal), technology will compute this area for us.

2.3 Mean as a balance point (center of mass)

For continuous distributions, the **mean** μ is still the “center” but now we think of it as a **balance point**:

- Imagine the density curve is a thin sheet of metal.
- The mean μ is the point where the sheet would balance on a fulcrum.

We do not need calculus in this class, but (for curiosity) the continuous mean is defined by an *integral*:

$$\mu = \int_{-\infty}^{\infty} x\, f(x)\, dx,$$

where $f(x)$ is the height of the density curve.

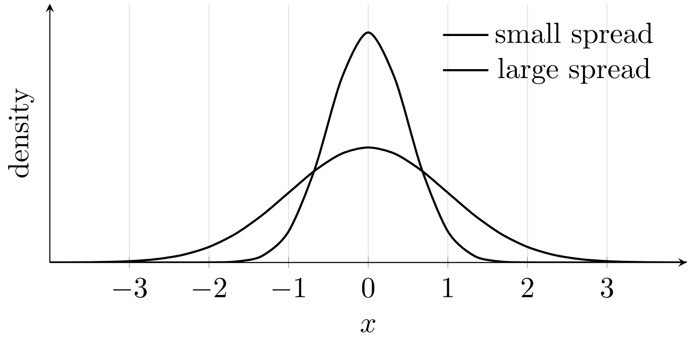
For our purposes, it is enough to know:

- μ tells us the center of the distribution,
- many common density curves are symmetric around μ.

2.4 Standard deviation as width

The **standard deviation** σ of a continuous distribution measures how spread out the curve is around its mean:

- larger $\sigma \Rightarrow$ wider, flatter curve,
- smaller $\sigma \Rightarrow$ narrower, taller curve.



We will see specific values of μ and σ when we work with concrete distributions (like the normal distribution) later in the course.

2.5 From histograms to curves

If we take a large sample of resting heart rates and draw a histogram, and then increase the sample size, the histogram becomes smoother. A density curve is an idealized limit of this process:

- The histogram describes *one sample.*
- The density curve describes the *long-run population model.*

3 Discrete Distributions and Sample Measurements

Not all random variables are continuous. Some take only specific values, like 0,1,2,3,. . . .

3.1 Discrete random variables

A **discrete random variable** takes isolated values, often counts:

- number of classes a student is taking,
- number of tutoring visits in a week,
- number of text messages sent in an hour.

Our discrete running example:

$$Y = \text{number of tutoring center visits in a week for a randomly chosen student.}$$

Suppose the population behaves like this:

y (visits)	0	1	2	3	4	5
$P(Y = y)$ (probability)	0.40	0.30	0.15	0.08	0.05	0.02

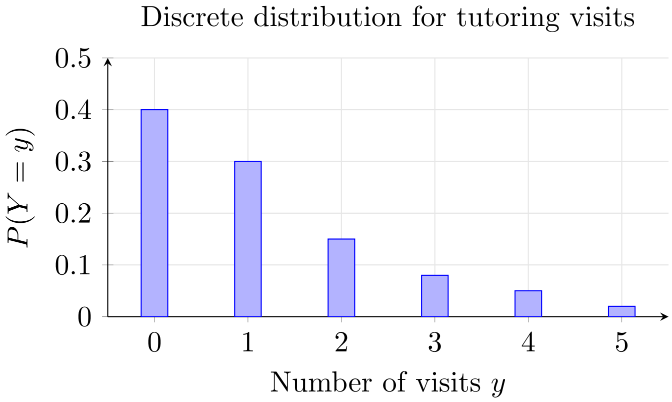
Check that the probabilities sum to 1:

$$0.40 + 0.30 + 0.15 + 0.08 + 0.05 + 0.02 = 1.00.$$

This table is a **discrete probability distribution** for Y.

3.2 Bar graph for a discrete distribution

We can draw a bar graph for $P(Y = y)$:



We can now answer questions like:

- $P(Y = 0) = 0.40$ (40% of students do not visit at all).
- $P(Y \geq 2) = P(Y = 2) + P(Y = 3) + P(Y = 4) + P(Y = 5) = 0.15 + 0.08 + 0.05 + 0.02 = 0.30.$

3.3 Mean of a discrete distribution (population level)

The **mean** of a discrete probability distribution is like a weighted average of the possible values, using the probabilities as weights:

$$\mu_Y = \sum_y y \cdot P(Y = y).$$

For our tutoring example:

$$\begin{aligned}\mu_Y &= 0 \cdot 0.40 + 1 \cdot 0.30 + 2 \cdot 0.15 + 3 \cdot 0.08 + 4 \cdot 0.05 + 5 \cdot 0.02 \\ &= 0 + 0.30 + 0.30 + 0.24 + 0.20 + 0.10 \\ &= 1.14.\end{aligned}$$

Interpretation:

- On average, students visit the tutoring center 1.14 times per week,
- If we could see the whole population, $\mu_Y = 1.14$ would be the true population mean.

In this sense, discrete distributions are also **population models**.

4 Empirical (Sample) Distributions and Sampling Variability

So far in this week:

- continuous distributions: density curves with area = probability,
- discrete distributions: probability tables and bar graphs with heights = probability.

Now we connect these models back to data.

4.1 Empirical distribution from a sample

Suppose we collect a sample of $n = 10$ students and record their weekly tutoring visits:

$$0,\ 0,\ 1,\ 1,\ 1,\ 2,\ 2,\ 3,\ 4,\ 5.$$

We can build a **frequency table** and turn it into *relative frequencies* (proportions):

y (visits)	0	1	2	3	4	5
Frequency in sample	2	3	2	1	1	1
Relative frequency	0.20	0.30	0.20	0.10	0.10	0.10

These relative frequencies are an **empirical distribution** for the sample.

Compare this to the population model:

y	0	1	2	3	4	5
Population $P(Y = y)$	0.40	0.30	0.15	0.08	0.05	0.02
Sample relative frequency	0.20	0.30	0.20	0.10	0.10	0.10

They are not the same, but they are *similar*: some sample bars are higher, others lower.

4.2 Sample statistics change from sample to sample

From this sample, the sample mean is

$$\bar{y} = \frac{0+0+1+1+1+2+2+3+4+5}{10} = \frac{19}{10} = 1.9.$$

Compare:

- **Population mean:** $\mu_Y = 1.14$ (from the probability model).
- **Sample mean:** $\bar{y} = 1.9$ (from this particular sample).

They are different because a sample is only part of the population.

If we took a different random sample of 10 students, we would almost surely get a different list of visits, a different empirical distribution, and a different sample mean $\bar{y}$.

This natural wobbling of sample statistics is called **sampling variability**.

Big picture for Week 2

By the end of this week, you should be comfortable with:

- Reading a density curve as a population model: area under the curve is probability.
- Interpreting the mean μ as a center (balance point) and the standard deviation σ as a measure of spread (width of the curve).
- Reading a discrete probability table and computing probabilities like $P(Y \geq 2)$.
- Computing the mean of a discrete distribution as a weighted average.
- Understanding that a sample has its own empirical distribution and sample statistics, which will differ from the population model and change from sample to sample.

Later we will study specific, named distributions (like the uniform, normal, geometric, and binomial) and use this language to work with them.