MAT120: Lecture 14 Handout

The Normal Distribution

Last lecture we began statistics by learning how to describe data, and we introduced the two core ideas of central tendency and variation. In this lecture we study the most famous continuous probability distribution: the normal distribution. We will learn what the normal curve is, what the standard normal distribution is, and how z-scores let us move between different normal distributions. We will also learn why probability for continuous variables is measured using *area under the curve* rather than bar heights. Next lecture we will use these tools to begin hypothesis testing.

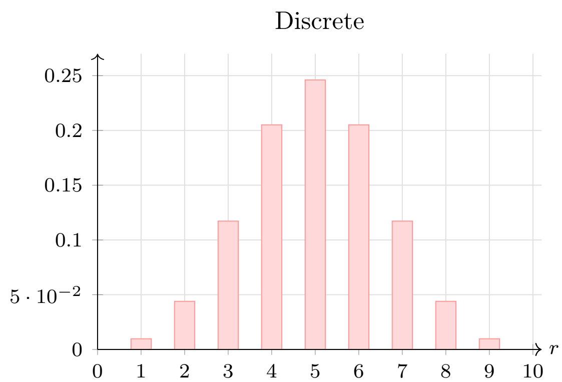
1 What is a Normal Distribution?

1.1 Discrete and continuous random variables

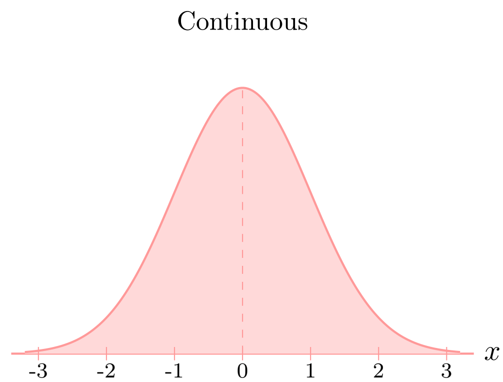
A *random variable* is a variable whose value is determined by chance. At this stage there are two important types to keep in mind:

- **Discrete random variable:** can take separate values such as $0, 1, 2, 3, \ldots$.
- **Continuous random variable:** can take any real value in an interval.

In Lecture 12 we studied a probability distribution built from a discrete random variable, namely the binomial distribution. Today we move to a probability distribution built from a continuous random variable instead.



A Binomial Distribution



A Normal Distribution

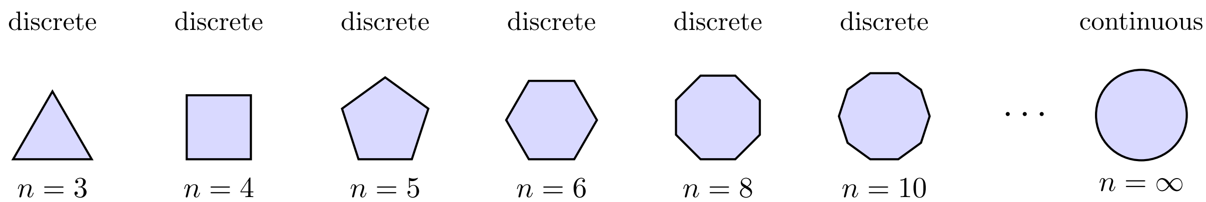
Continuous probability distributions have more structure than the discrete distributions we have seen so far. That means there is more to talk about, and therefore they are potentially more confusing. So, our strategy will be to learn the continuous case by using *limits* to relate them to what we already know.

1.2 A metaphor for limits

Suppose that we have a bunch of logs whose cross-sections have different shapes:

In the case of the first log, having a triangular cross-section will make it very awkward to roll. If we were to move on to the next log, which has a square cross-section, then we would notice that it is slightly easier. After this, we would notice that the log with a pentagonal cross-section would be even easier, and again the log with a hexagonal cross-section would be easier still. We notice that as the number of sides of the cross-section increase, the log gets easier and easier to roll. Of course, a perfectly circular log would be the easiest.

If we were to have many many logs and kept increasing the number of sides, we would get a sequence of logs that may look something like this:

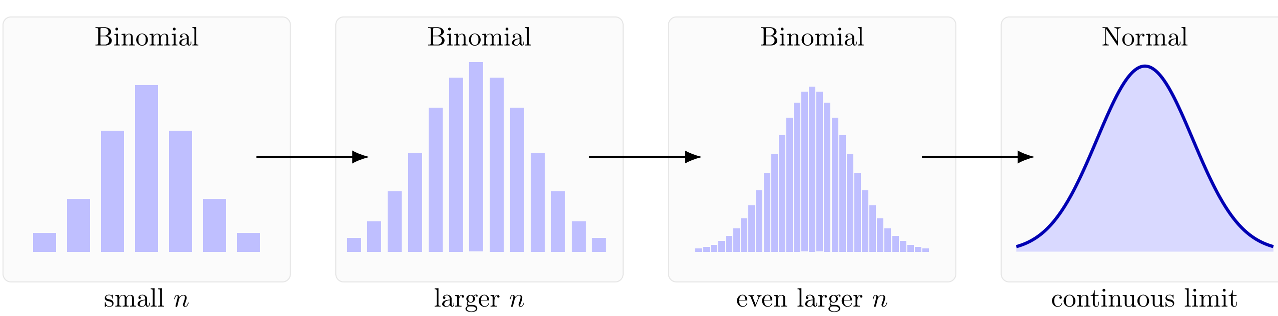


The limit of this procedure would be a circle, which we can think of as a shape with infinitely-many sides. If we interpret the number of sides as a discrete quantity, we see that in the process of adding more sides to our shapes, the polygons approach a continuous shape in the limit.

1.3 The limiting process

Our analogy of rolling logs has a deeper philosophy that may be applied to random variables: we can "approach" a normal distribution by considering a sequence of binomial distributions in which we increase the number of trials, and hence the number of bars. Put differently: a continuous curve can be described as a limit of many discrete bars.

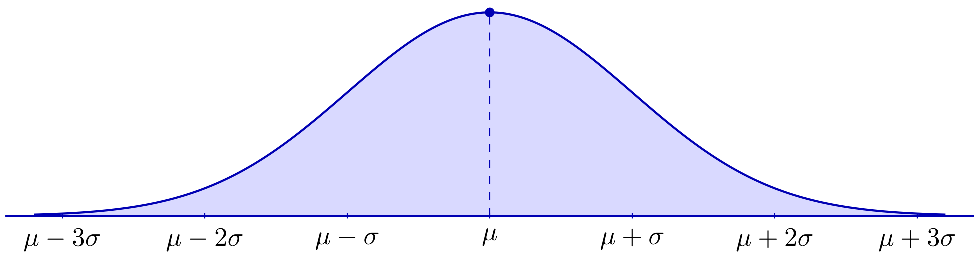
Recall that in a binomial distribution, the total number of bars in the graph is given by $n + 1$, since the possible values run from 0 to n. As n increases, the bars in the binomial distribution become thinner and more numerous. The overall distribution appears smoother and smoother, and eventually begins to resemble a curve:



This limiting bell-shaped curve is called the *normal distribution*.

1.4 The normal distribution and the normal curve

The graph of the normal density function is called a *normal curve*. Because of its shape, it is also often called a *bell curve*.



The normal distribution is one of the most important objects in statistics. In practice, we see the normal distribution in human characteristics, physical measurements, industrial data, measurement error, and many other contexts.

A normal distribution has several important features.

Important properties of a normal distribution

1. The curve is bell-shaped, with its highest point above the mean μ.
2. The curve is symmetric about the vertical line through μ.
3. The tails approach the horizontal axis but never touch or cross it.
4. The total area under the entire curve is 1.

1.5 Mathematical description

Since a normal distribution is represented by a graph, we can ask what the associated function is. This function has a special name: it is called the *normal density function*. Although we will never need to work with this function directly, it is written below:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Despite the obvious complexity of this function, we should make a few important observations. First, notice that the function has a single variable x. This is the continuous random variable that becomes the horizontal axis of the graph. Next, we see that apart from x, the function $f(x)$ needs two important numbers as parameters:

- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.

As you can see from the function above, the numbers μ and σ are the only pieces of parameter information in the function (apart from constants such as π, of course). Therefore every normal distribution is completely *characterized* by the numbers μ and σ. This means that different choices of μ and σ give different normal curves. For this reason, we often write a normal distribution as:

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

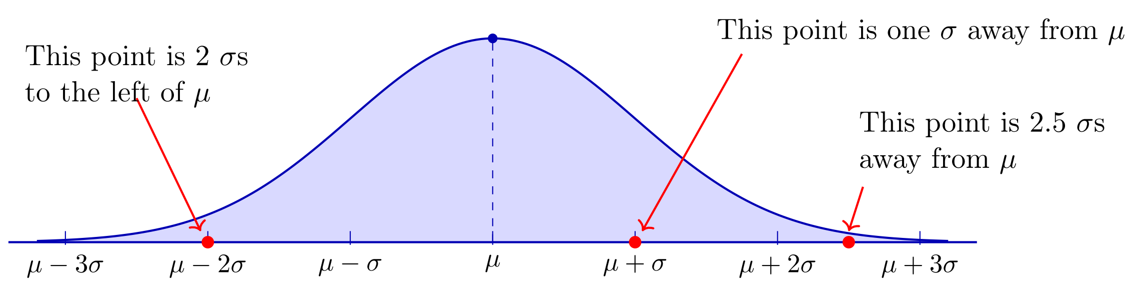
In words, this should be read "the continuous random variable X follows a normal distribution with mean μ and variance σ^2."

This is quite a special feature of the normal distribution. The population mean μ tells us where the centre of the distribution is, and the population standard deviation σ is a measure of how spread out the data is. Usually, these are just two numbers that help describe a dataset. However, in the case of the normal distribution, the numbers μ and σ *completely describe* the distribution. So, the symbols μ and σ are all the information we need in order to navigate the curve.

2 The Standard Normal Distribution

2.1 The z-score

According to our discussion so far, we see that any normal distribution $\mathcal{N}(\mu, \sigma^2)$ has an x-axis that can be described in terms of μ and σ. If we randomly select a value of the random variable X, then this will land somewhere on the x-axis. Since the x-axis is centred at μ and measured in units of σ, we can express where our variable is in terms of distance from the centre. For example:



The *z-score*, also called the *standard score*, turns this information into a pure number, by telling us how many standard deviations a measurement lies from the mean.

Definition of the z-score

If x is a value from a normal distribution with mean μ and standard deviation σ, then its z-score is

$$z = \frac{x - \mu}{\sigma}.$$

This formula can be interpreted one piece at a time:

- $x - \mu$ measures the distance x is from the mean;
- dividing by σ converts that distance into a count of standard deviations;
- the result may be positive or negative, and this corresponds to the right or left of the mean, respectively.

For example, in the diagram above:

- the point that is at $\mu - 2\sigma$ would have a z-score of -2,
- the point that is at $\mu + \sigma$ would have a z-score of 1, and
- the point that is at $\mu + 2.5\sigma$ would have a z-score of 2.5.

Sometimes we know the z-score and want to recover the original value x. Starting with the formula

$$z = \frac{x - \mu}{\sigma},$$

we can rearrange to get:

$$x = \mu + z\sigma.$$

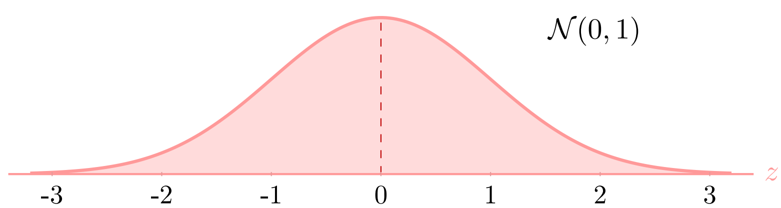
The z-score and the raw score contain the same information, just written in two different ways.

2.2 The standard normal distribution

The *standard normal distribution* is the normal distribution with mean 0 and standard deviation 1. We usually write this as

$$Z \sim \mathcal{N}(0, 1).$$

Note that we use the special letter Z for a standard normal random variable, instead of the letter X that we wrote before. As a convention, we will draw the standard normal distribution $\mathcal{N}(0, 1)$ in red, so that we can visually distinguish it from general normal distributions, which we will draw in blue. The standard normal distribution is depicted below.



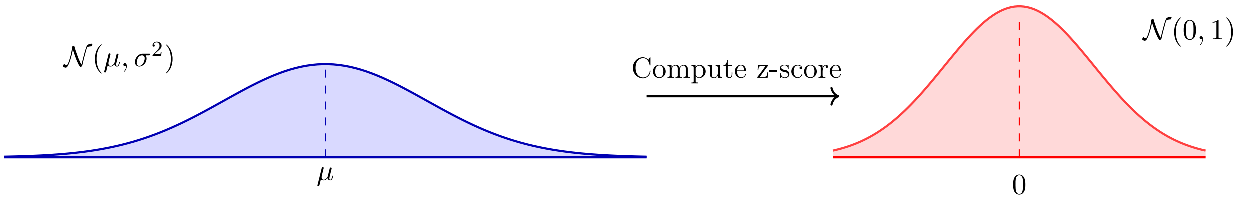
Observe the horizontal axis here: we are centred at 0 and we count up or down in increments of 1.

2.3 Why the standard normal is useful

Any normal distribution can be converted into the standard normal distribution by converting its measurements to z-scores. That is why the standard normal distribution is so important: instead of learning a different area table for every possible normal curve, we use one universal reference curve.

The big idea

Every normal distribution can be translated into the standard normal world. Once that has been done, probabilities can be read from a single table of standard normal areas.



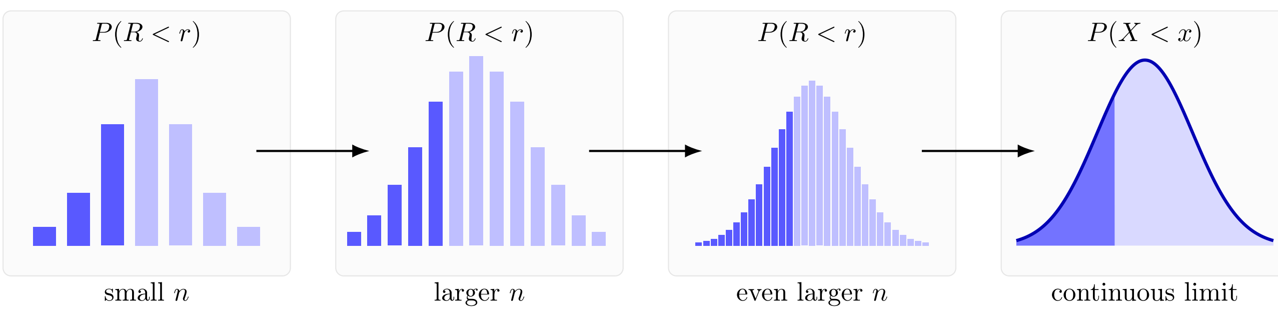
3 Areas and Probability

3.1 From bar heights to area

When working with a binomial distribution, the height of each bar gives the exact probability that the discrete random variable will take on that value. For example, in the binomial distribution for $n = 10$ and $p = 0.5$:

- the probability of exactly $r = 5$ is the height of the bar at $r = 5$, which is about 0.24;
- the probability of $r < 5$ is obtained by adding the heights of the bars at $r = 0, 1, 2, 3, 4$, which gives about 0.38.

For larger and larger values of n, this process of adding many thin bars begins to look more and more like a continuous area:



As the above picture demonstrates, for continuous random variables, we think of probabilities in terms of *areas under the curve*, rather than (a sum of) the heights of bars.

Discrete versus continuous probability

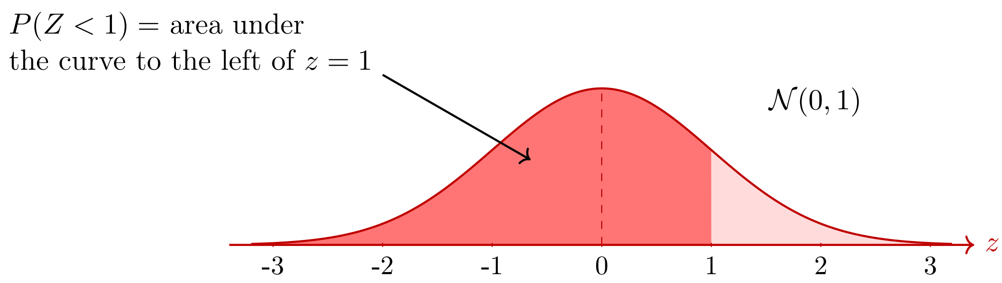
- In the **discrete** case, probability is found by adding the probabilities of separate outcomes.
- In the **continuous** case, probability is found by measuring area under a curve.

3.2 How areas are found

Consider again the standard normal distribution $\mathcal{N}(0, 1)$. According to the discussion above, when we write

$$P(Z < z),$$

we mean the area under the standard normal curve to the left of the point z. For example, the probability $P(Z < 1)$ would be given by the shaded area:



If you know some calculus, then these areas can be found using integration. However, this is often difficult, so in practice we often use a *z-table* instead.

We can think of a z-table as a list of all of the answers for common regions of the standard normal distribution. In this course, we will mostly use the table:

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z < z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

Here, the second row gives the area to the left of z for particular values of z. For example, in the diagram above, the table tells us that the shaded area is 0.84, meaning that 84% of the total area under the curve lies to the left of $z = 1$. In probabilistic terms, this says that our random variable Z has about an 84% chance of taking a value less than 1.

3.3 Understanding Areas

A z-table lists areas under the standard normal curve for the standard normal random variable Z. We interpret the table as giving an approximate answer for the probability

$$P(Z < z),$$

that is, the area to the *left* of z. Using the basic properties of the normal distribution, we can use this table to describe other areas as well.

How to find other areas

- **Areas to the left of z:** read off $P(Z < z)$ directly.
- **Areas to the right of z:** use $P(Z > z) = 1 - P(Z < z)$.
- **Areas between two values:** use $P(a < Z < b) = P(Z < b) - P(Z < a)$.
- **Negative values:** by symmetry, for $a > 0$,

$$P(Z < -a) = 1 - P(Z < a).$$

For a continuous random variable, the probability of any *single exact point* is zero. So, for example,

$$P(Z = 0.5) = 0.$$

That does not mean the point is impossible. It only means that a single point has zero width, and therefore zero area.

3.4 Computing Probabilities for Other Normal Distributions

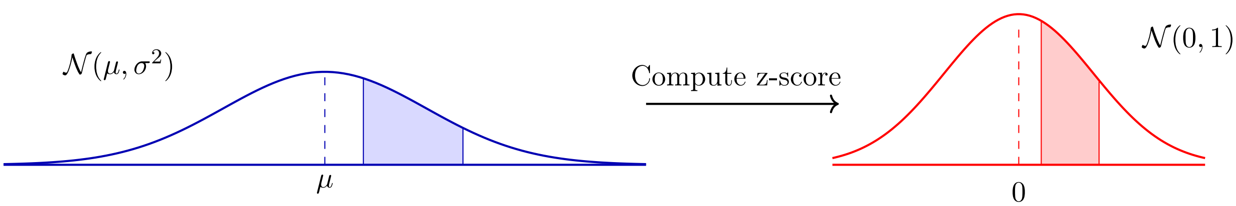
In Section 2 we mentioned that we can use the z-score formula

$$z = \frac{x - \mu}{\sigma}$$

to convert any normal distribution $\mathcal{N}(\mu, \sigma^2)$ into the standard normal distribution $\mathcal{N}(0, 1)$. What is useful here is that this transformation process also preserves the proportions of areas, meaning that we can determine probabilities in $\mathcal{N}(\mu, \sigma^2)$ by using z-scores and the z-table above. For a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we have:

$$P(X < x) = P\left(Z < \frac{x - \mu}{\sigma}\right).$$

In other words, to determine the value of $P(X < x)$, we simply compute the z-score of the point x, and then look at the corresponding area in the z-table above. This notion also works for all other areas. For example, a probability like $P(x_1 < X < x_2)$ can be equally determined using z-scores for x_1 and x_2:



This is the true strength of z-scores: they allow us to handle *any* normal distribution using the standard normal table.

Worked example

Let $X \sim \mathcal{N}(10, 2^2)$. Find the probability that a randomly selected value lies between 11 and 14. In symbols, find $P(11 \leq X \leq 14)$.

Solution

Step 1: Convert the x-values to z-values.

$$z_1 = \frac{11-10}{2} = 0.5, \qquad z_2 = \frac{14-10}{2} = 2.0.$$

Step 2: Read the left-tail areas from the table.

$$P(Z < 0.5) = 0.69, \qquad P(Z < 2.0) = 0.98.$$

Step 3: Subtract to get the interval probability.

$$P(11 \le X \le 14) = P(Z < 2.0) - P(Z < 0.5) = 0.98 - 0.69 = 0.29.$$

So the probability is

$$0.29.$$

In words, there is about a 29% chance that a randomly selected value from this normal model lies between 11 and 14.

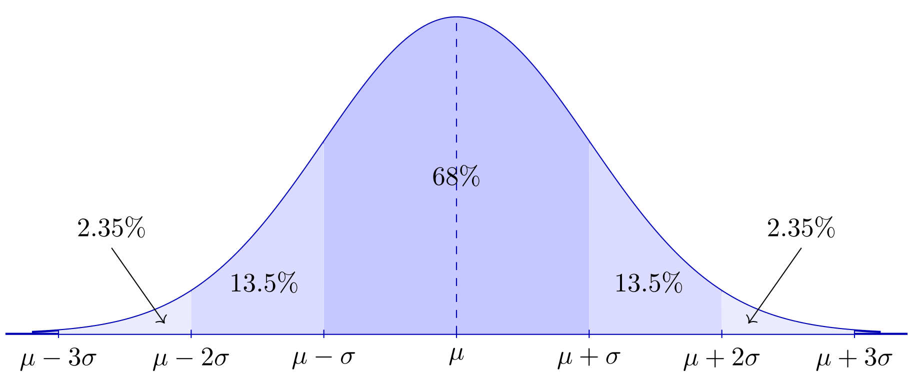
3.5 The 68-95-99.7 rule

For a normal distribution, the standard deviation gives a very useful geometric guide to the data.

Empirical rule

For a normal distribution:

- about 68% of the data lie within 1 standard deviation of the mean;
- about 95% of the data lie within 2 standard deviations of the mean;
- about 99.7% of the data lie within 3 standard deviations of the mean.



The remaining 0.3% lies in the two extreme tails, about 0.15% in each tail.

3.6 Worked exercise: areas under the standard normal curve

Use the mini z-table above, together with symmetry, to find the following areas.

Exercise

(a) Find the area to the left of $z = -1.00$.
(b) Find the area to the left of $z = 2$.
(c) Without using the z-table, find the area to the left of $z = 0$.
(d) Without using the z-table, find the area underneath the entire curve.
(e) Find the area to the right of $z = 1$.
(f) Find the area of the interval $-1 < Z < 2$.

Solution

(a) From the table,
$$P(Z < -1.00) = 0.16.$$
(b) From the table,
$$P(Z < 2) = 0.98.$$
(c) By symmetry, half of the total area lies to the left of 0, so
$$P(Z < 0) = 0.50.$$
(d) The total area under the curve is always 1.
(e)
$$P(Z > 1) = 1 - P(Z < 1) = 1 - 0.84 = 0.16.$$
(f)
$$P(-1 < Z < 2) = P(Z < 2) - P(Z < -1).$$
Using the table,
$$P(-1 < Z < 2) = 0.98 - 0.16 = 0.82.$$