CS 70 Discrete Mathematics and Probability Theory
Fall 2025 Course Notes **Note 24**

1 Normal Distribution

Another continuous distribution we will consider, and by far the most prevalent in applications, is called the *normal* or *Gaussian* distribution. It has two parameters, μ and σ^2, which are the mean and variance of the distribution, respectively.

Definition 24.1 (Normal Distribution)**.** *For any $\mu \in \mathbb{R}$ and $\sigma > 0$, a continuous random variable X with p.d.f.*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

is called a normal random variable with parameters μ and σ^2, and we write $X \sim N(\mu, \sigma^2)$. In the special case $\mu = 0$ and $\sigma = 1$, X is said to have the standard normal distribution.

Let's first check that this is a valid definition of a probability density function. Clearly $f(x) \geq 0$ from the definition. Then we verify that the integral of the probabilities is 1:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}\,\mathrm{d}x = 1. \tag{1}$$

The fact that this integral evaluates to 1 is a standard exercise in (mutivariable) integral calculus (or feel free to look it up in any standard text on probability or on the internet).

A plot of the p.d.f. f reveals a classical "bell-shaped" curve, centered at (and symmetric around) $x = \mu$, and with "width" determined by σ. Figure 1 shows that the normal densities with different values of μ and σ are very similar to each other. Indeed, the normal distribution has the following nice property with respect to shifting and rescaling.

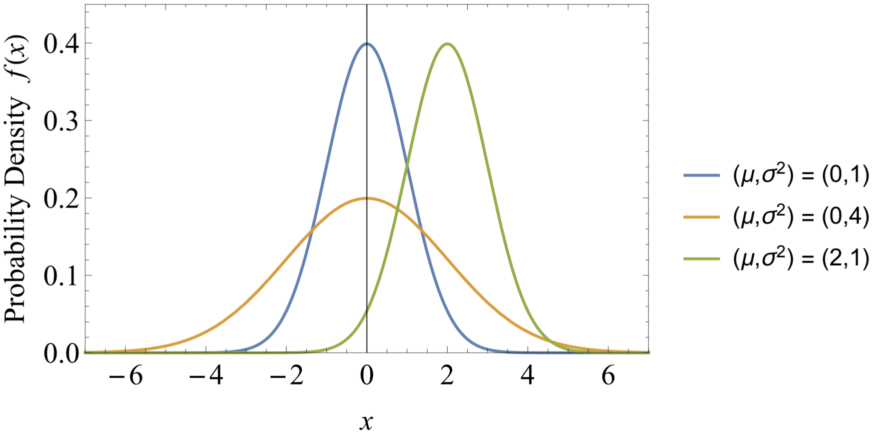

Figure 1: The density function for the normal distribution with several different choices for μ and σ^2.

Lemma 24.1. *If $X \sim N(\mu, \sigma^2)$, then $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$. Equivalently, if $Y \sim N(0,1)$, then $X = \sigma Y + \mu \sim N(\mu, \sigma^2)$.*

Proof. Given that $X \sim N(\mu, \sigma^2)$, we can calculate the distribution of $Y = \frac{X-\mu}{\sigma}$ as:

$$\mathbb{P}[a \le Y \le b] = \mathbb{P}[\sigma a + \mu \le X \le \sigma b + \mu] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\sigma a+\mu}^{\sigma b+\mu} e^{-(x-\mu)^2/(2\sigma^2)}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-y^2/2}\,\mathrm{d}y,$$

by a simple change of variable $x = \sigma y + \mu$ in the integral. Hence Y is indeed standard normal. Note that Y is obtained from X just by shifting the origin to μ and scaling by σ. □

1.1 Mean and Variance of a Normal Random Variable

Let us now calculate the expectation and variance of a normal random variable.

Theorem 24.1. *For $X \sim N(\mu, \sigma^2)$,*

$$\mathbb{E}[X] = \mu \qquad \textit{and} \qquad \mathrm{Var}(X) = \sigma^2.$$

Proof. First consider the case when $X \sim N(0,1)$. By definition, its expectation is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{0} x e^{-x^2/2}\,\mathrm{d}x + \int_{0}^{\infty} x e^{-x^2/2}\,\mathrm{d}x \right) = 0.$$

The last step follows from the fact that the function $e^{-x^2/2}$ is symmetrical about $x = 0$, so the two integrals are the same except for the sign. For the variance, we have

$$\begin{aligned} \mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \left(-x e^{-x^2/2}\right)\Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}x = 1. \end{aligned}$$

In the first line here we used the fact that $\mathbb{E}[X] = 0$; in the second line we used integration by parts; and in the last line we used (1) in the special case $\mu = 0$, $\sigma = 1$. So the standard normal distribution has expectation $\mathbb{E}[X] = 0 = \mu$ and variance $\mathrm{Var}(X) = 1 = \sigma^2$.

Now consider the general case when $X \sim N(\mu, \sigma^2)$. By Lemma 24.1, we know that $Y = \frac{X-\mu}{\sigma}$ is a standard normal random variable, so $\mathbb{E}[Y] = 0$ and $\mathrm{Var}(Y) = 1$, as we have just established above. Therefore, we can read off the expectation and variance of X from those of Y. For the expectation, using linearity, we have

$$0 = \mathbb{E}[Y] = \mathbb{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{\mathbb{E}[X]-\mu}{\sigma},$$

and hence $\mathbb{E}[X] = \mu$. For the variance we have

$$1 = \mathrm{Var}(Y) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{\mathrm{Var}(X)}{\sigma^2},$$

and hence $\mathrm{Var}(X) = \sigma^2$. □

The bottom line, then, is that the normal distribution has expectation μ and variance σ^2. This explains the notation for the parameters μ and σ^2.

The fact that the variance is σ^2 (so that the standard deviation is σ) explains our earlier comment that σ determines the “width” of the normal distribution. Namely, by Chebyshev’s inequality, a constant fraction of the distribution lies within distance (say) 2σ of the expectation μ.

Note: The above analysis shows that, by means of a simple origin shift and scaling, we can relate any normal distribution to the standard normal. This means that, when doing computations with normal distributions, it’s enough to do them for the standard normal. For this reason, books and online sources of mathematical formulas usually contain tables describing the density of the standard normal. From this, one can read off the corresponding information for any normal r.v. $X \sim N(\mu, \sigma^2)$ from the formula

$$\mathbb{P}[X \le a] = \mathbb{P}[Y \le \tfrac{a-\mu}{\sigma}],$$

where Y is standard normal.

The normal distribution is ubiquitous throughout the sciences and the social sciences, because it is the standard model for aggregate data that results from averaging a large number of independent observations of the same random variable (such as the weights of mosquitos in Berkeley, or the outcome of a physical experiment). Such averaged data, as is well known, tends to cluster around its mean in a “bell-shaped" curve, with the correspondence becoming more accurate as the number of observations increases. A theoretical explanation of this phenomenon is the Central Limit Theorem, which we discuss in Section 2.

1.2 Sum of Independent Normal Random Variables

An important property of the normal distribution is that the sum of *independent* normal random variables is also normally distributed. We begin with the simple case when X and Y are independent standard normal random variables. In this case the result follows because the joint distribution of X and Y is rotationally symmetric. The general case follows from the translation and scaling property of normal distribution in Lemma 24.1.

Theorem 24.2. *Let $X \sim N(0,1)$ and $Y \sim N(0,1)$ be independent standard normal random variables, and suppose $a, b \in \mathbb{R}$ are constants. Then $Z = aX + bY \sim N(0, a^2 + b^2)$.*

Proof. [1] Since X and Y are independent, we know that the joint density of (X,Y) is simply the product of the marginal densities:

$$f(x,y) = f(x) \cdot f(y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}.$$

The key observation is that $f(x,y)$ is rotationally symmetric around the origin, i.e., $f(x,y)$ only depends on the value x^2+y^2, which is the distance of the point (x,y) from the origin $(0,0)$; see Figure 2.

Thus, $f(T(x,y)) = f(x,y)$ where T is any rotation of the plane $\mathbb{R}^2$ about the origin. It follows that for any set $A \subseteq \mathbb{R}^2$,

$$\mathbb{P}[(X,Y) \in A] = \mathbb{P}[(X,Y) \in T(A)] \tag{2}$$

where T is a rotation of $\mathbb{R}^2$. Now given any $t \in \mathbb{R}$, we have

$$\mathbb{P}[Z \le t] \;=\; \mathbb{P}[aX + bY \le t] \;=\; \mathbb{P}[(X,Y) \in A]$$

[1] The following proof and figures are adapted from *“Why Is the Sum of Independent Normal Random Variables Normal?”* by B. Eisenberg and R. Sullivan, Mathematics Magazine, Vol. 81, No. 5.

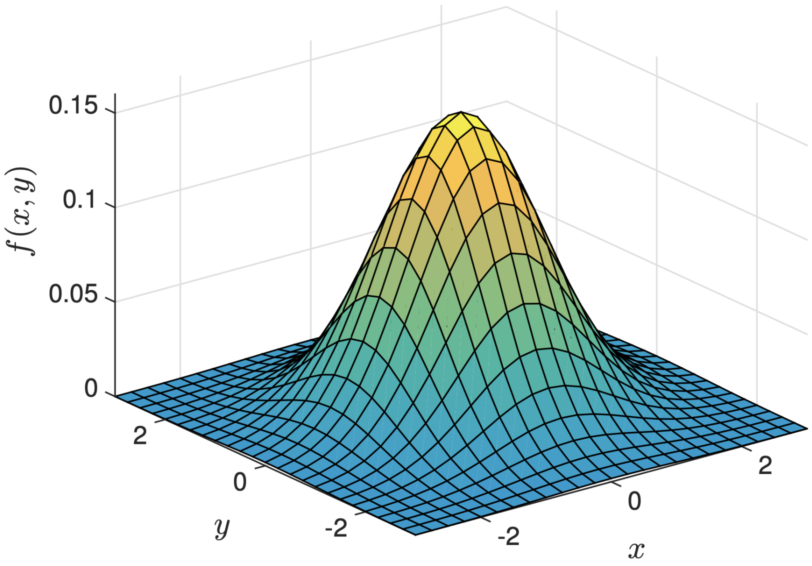

Figure 2: The joint density function $f(x,y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}$ is rotationally symmetric.

where A is the half plane $\{(x,y) \mid ax+by \le t\}$. The boundary line $ax+by=t$ lies at a distance $d = \frac{t}{\sqrt{a^2+b^2}}$ from the origin. Therefore, as illustrated in Figure 3, the set A can be rotated into the set

$$T(A) = \left\{ (x,y) \;\middle|\; x \le \frac{t}{\sqrt{a^2+b^2}} \right\}.$$

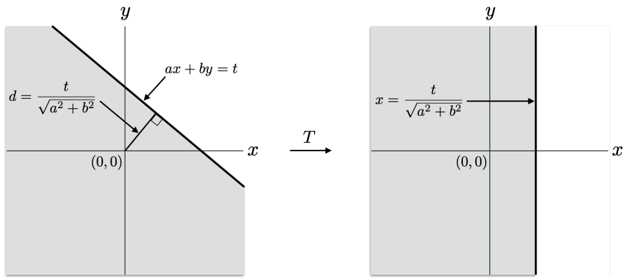
y
$ax+by=t$
$d = \frac{t}{\sqrt{a^2+b^2}}$
x
$(0,0)$
T
y
$x = \frac{t}{\sqrt{a^2+b^2}}$
x
$(0,0)$

Figure 3: The half plane $ax+by \le t$ is rotated into the half plane $x \le \frac{t}{\sqrt{a^2+b^2}}$.

By (2), this rotation does not change the probability:

$$\mathbb{P}[Z \le t] = \mathbb{P}[(X,Y) \in A] = \mathbb{P}[(X,Y) \in T(A)] = \mathbb{P}\left[X \le \frac{t}{\sqrt{a^2+b^2}}\right] = \mathbb{P}\left[\sqrt{a^2+b^2}\,X \le t\right].$$

Since the equation above holds for all $t \in \mathbb{R}$, we conclude that Z has the same distribution as $\sqrt{a^2+b^2}\,X$. Since X has standard normal distribution, we know by Lemma 24.1 that $\sqrt{a^2+b^2}\,X$ has normal distribution with mean 0 and variance a^2+b^2. Hence we conclude that $Z = aX+bY$ also has normal distribution with mean 0 and variance a^2+b^2. □

The general case now follows easily from Lemma 24.1 and Theorem 24.2.

Corollary 24.1. *Let $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ be independent normal random variables. Then for any constants $a, b \in \mathbb{R}$, the random variable $Z = aX + bY$ is also normally distributed with mean $\mu = a\mu_X + b\mu_Y$ and variance $\sigma^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$.*

Proof. By Lemma 24.1, $Z_1 = (X - \mu_X)/\sigma_X$ and $Z_2 = (Y - \mu_Y)/\sigma_Y$ are independent standard normal random variables. We can write:

$$Z = aX + bY = a(\mu_X + \sigma_X Z_1) + b(\mu_Y + \sigma_Y Z_2) = (a\mu_X + b\mu_Y) + (a\sigma_X Z_1 + b\sigma_Y Z_2).$$

By Theorem 24.2, $Z' = a\sigma_X Z_1 + b\sigma_Y Z_2$ is normally distributed with mean 0 and variance $\sigma^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$. Since $\mu = a\mu_X + b\mu_Y$ is a constant, by Lemma 24.1 we conclude that $Z = \mu + Z'$ is a normal random variable with mean μ and variance σ^2, as desired. □

2 The Central Limit Theorem

Recall from an earlier note the Law of Large Numbers for i.i.d. random variables $\{X_i\}$: it says that the probability of *any* deviation $\varepsilon > 0$, however small, of the sample average $\frac{S_n}{n}$, where $S_n = \sum_{i=1}^n X_i$, from the mean tends to zero as the number of observations n in our average tends to infinity. Thus, by taking n large enough, we can make the probability of any given deviation as small as we like.

Actually we can say something much stronger than the Law of Large Numbers: namely, the distribution of the sample average $\frac{S_n}{n}$, for large enough n, looks like a *normal distribution* with mean μ and variance $\frac{\sigma^2}{n}$. (Of course, we already know that these are the mean and variance of $\frac{S_n}{n}$; the point is that the distribution becomes normal!) The fact that the standard deviation decreases with n (specifically, as $\frac{\sigma}{\sqrt{n}}$) means that the distribution approaches a sharp spike at μ.

Recall from the last section that the density of the normal distribution is a symmetrical bell-shaped curve centered around the mean μ. Its height and width are determined by the standard deviation σ as follows: the height at the mean $x = \mu$ is $\frac{1}{\sqrt{2\pi\sigma^2}} \approx \frac{0.4}{\sigma}$; 50% of the mass is contained in the interval of width 0.67σ either side of the mean, and 99.7% in the interval of width 3σ either side of the mean. (Note that, to get the correct scale, deviations are on the order of σ rather than σ^2.)

To state the Central Limit Theorem precisely (so that the limiting distribution is a constant rather than something that depends on n), we standardize $\frac{S_n}{n}$ as

$$\frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

The Central Limit Theorem then says that the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to the *standard normal* distribution.

Theorem 24.3 (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common finite expectation $\mathbb{E}[X_i] = \mu$ and finite variance $\mathrm{Var}(X_i) = \sigma^2$. Let $S_n = \sum_{i=1}^n X_i$. Then, the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to $N(0,1)$ as $n \to \infty$. In other words, for any constant $c \in \mathbb{R}$,*

$$\mathbb{P}\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq c\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-x^2/2}\,\mathrm{d}x \quad \textit{as } n \to \infty.$$

The Central Limit Theorem is a very striking fact. What it says is the following: If we take an average of n observations of any arbitrary r.v. X, then the distribution of that average will be a bell-shaped curve centered

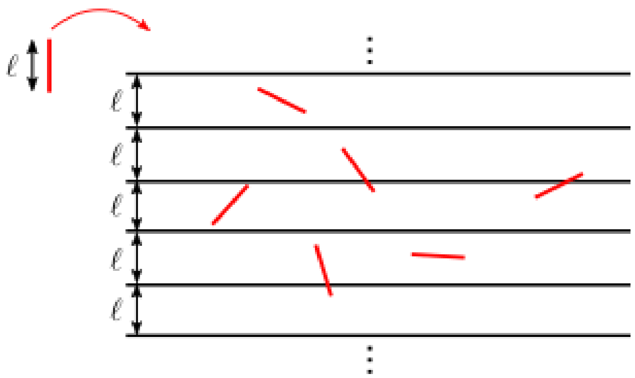

Figure 4: Buffon's Needle.

at $\mu = \mathbb{E}[X]$. Thus all trace of the distribution of X disappears as n gets large: all distributions, no matter how complex,[2] look like the normal distribution when they are averaged. The only effect of the original distribution is through the variance σ^2, which determines the width of the curve for a given value of n, and hence the rate at which the curve shrinks to a spike.

The Central Limit Theorem immediately tells us that *averaged* data tends to look normal, even when the distribution it is drawn from is not itself normal. Since much of the data we work with in real life is the result of averaging (mean wealth, mean height, mean temperature etc.), it's not surprising that we deal with the normal distribution so often. But there's actually a deeper reason why even (say) the height of a population itself tends to follow a normal distribution: this is because a person's height is actually itself the result of averaging many factors (nutrition, environment, various genetic factors, etc.), so intuitively at least we would expect it to follow a normal distribution.

3 Buffon's Needle

Here is a simple yet interesting application of continuous random variables to the analysis of a classical procedure for estimating the value of π; this is known as *Buffon's needle* problem, after its 18th century inventor Georges-Louis Leclerc, Comte de Buffon.

As illustrated in Figure 4, we are given a needle of length ℓ, and a board ruled with horizontal lines at distance ℓ apart. The experiment consists of throwing the needle randomly onto the board and observing whether or not it crosses one of the lines. We shall see below that (assuming a perfectly random throw) the probability of this event is exactly $2/\pi$. This means that, if we perform the experiment many times and record the *proportion* of throws on which the needle crosses a line, then the Law of Large Numbers tells us that we will get a good estimate of the quantity $2/\pi$, and therefore also of π; and we can use Chebyshev's inequality as in the other estimation problems we considered in an earlier note to determine how many throws we need in order to achieve specified accuracy and confidence.

3.1 Integrating a Joint Density Function

To analyze the experiment, let's consider what random variables are in play. Note that the position where the needle lands is completely specified by two random variables: the vertical distance Y between the midpoint of the needle and the closest horizontal line, and the angle Θ between the needle and the vertical. The r.v. Y

[2] We do need to assume that the mean and variance of X are finite.

ranges between 0 and $\ell/2$, while Θ ranges between $-\pi/2$ and $\pi/2$. Since we assume a perfectly random throw, we may assume that their *joint distribution* has density $f(y,\theta)$ that is uniform over the rectangle $[0,\ell/2]\times[-\pi/2,\pi/2]$. Since this rectangle has area $\frac{\pi\ell}{2}$, the density should be

$$f(y,\theta)=\begin{cases}\frac{2}{\pi\ell}, & \text{for } (y,\theta)\in[0,\ell/2]\times[-\pi/2,\pi/2],\\ 0, & \text{otherwise.}\end{cases} \tag{3}$$

Equivalently, Y and Θ are independent random variables, each uniformly distributed in their respective range. To check our answer, let's verify that the integral of this density over all possible values is indeed 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y,\theta)\,\mathrm{d}y\,\mathrm{d}\theta = \int_{-\pi/2}^{\pi/2}\int_{0}^{\ell/2}\frac{2}{\pi\ell}\,\mathrm{d}y\,\mathrm{d}\theta = \int_{-\pi/2}^{\pi/2}\frac{2y}{\pi\ell}\bigg|_{0}^{\ell/2}\,\mathrm{d}\theta = \int_{-\pi/2}^{\pi/2}\frac{1}{\pi}\,\mathrm{d}\theta = \frac{\theta}{\pi}\bigg|_{-\pi/2}^{\pi/2} = 1.$$

Since we have a joint distribution, rather than the area under the curve $f(x)$, we are now computing the area under the “surface" $f(y,\theta)$.

Now let E denote the event that the needle crosses a line. How can we express this event in terms of the values of Y and Θ? Well, by elementary geometry the vertical distance of the endpoint of the needle from its midpoint is $\frac{\ell}{2}\cos\Theta$, so the needle will cross the line if and only if $Y\le\frac{\ell}{2}\cos\Theta$. Therefore we have

$$\mathbb{P}[E]=\mathbb{P}[Y\le\tfrac{\ell}{2}\cos\Theta]=\int_{-\pi/2}^{\pi/2}\int_{0}^{(\ell/2)\cos\theta} f(y,\theta)\,\mathrm{d}y\,\mathrm{d}\theta.$$

Substituting the density $f(y,\theta)$ from (3) and performing the integration we get

$$\mathbb{P}[E]=\int_{-\pi/2}^{\pi/2}\int_{0}^{(\ell/2)\cos\theta}\frac{2}{\pi\ell}\,\mathrm{d}y\,\mathrm{d}\theta = \int_{-\pi/2}^{\pi/2}\frac{2y}{\pi\ell}\bigg|_{0}^{(\ell/2)\cos\theta}\,\mathrm{d}\theta = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2}\cos\theta\,\mathrm{d}\theta = \frac{1}{\pi}\sin\theta\bigg|_{-\pi/2}^{\pi/2} = \frac{2}{\pi}.$$

This is exactly what we claimed at the beginning of the section!