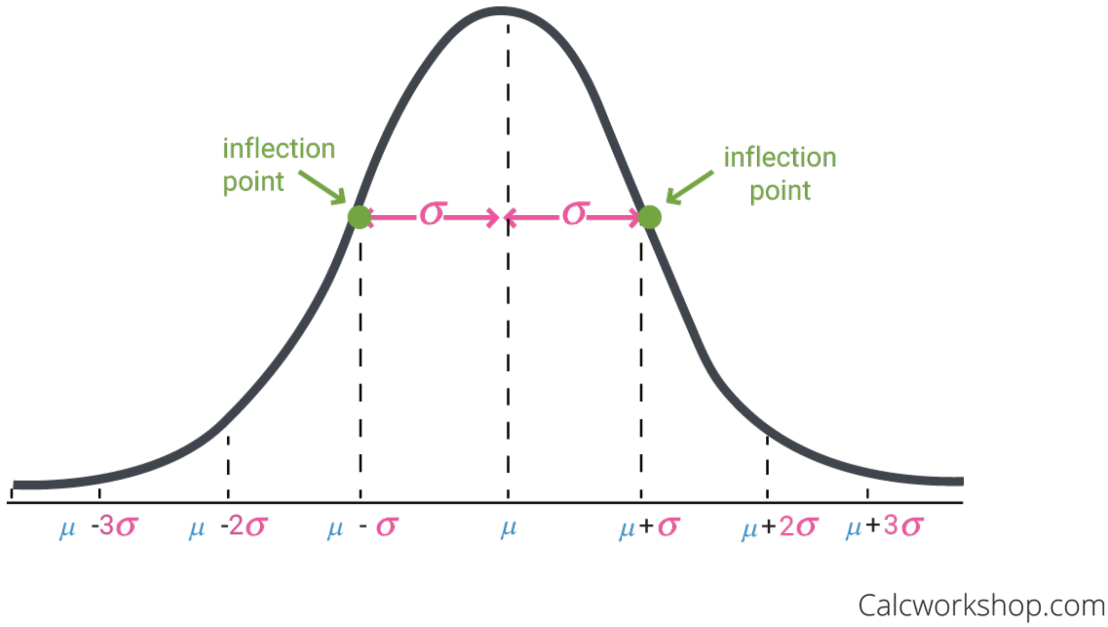

Figure 1.23: The density function of the normal distribution (the "bell curve"). The mean μ is the location of the single peak, and the standard deviation σ is the distance between the peak and the inflection points.

1.6.2 Normal distribution

Definition 1.16. X has the **normal** or **Gaussian distribution** on $\text{Range}(X) = \mathbb{R}$ if its density is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{1.21}$$

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

Remark 1.12. We call $\mathcal{N}(0,\ 1)$ the **standard normal** distribution, and typically use the letter Z for the random variable $Z \sim \mathcal{N}(0, 1)$.

Geometric properties

- The mean and variance of $X \sim \mathcal{N}(\mu, \sigma^2)$ are

  $$\mathbb{E}[X] = \mu, \qquad \text{var}(X) = \sigma^2.$$

  We will do this calculation later. Thus μ controls the center of the distribution, and σ controls the spread; see Figure 1.24.

- Although technically, $\text{Range}(X) = \mathbb{R}$, it is extremely rare to see values of X that are far from μ. This is because 99.7% of the mass lies within 3 standard deviations of the mean! In fact,

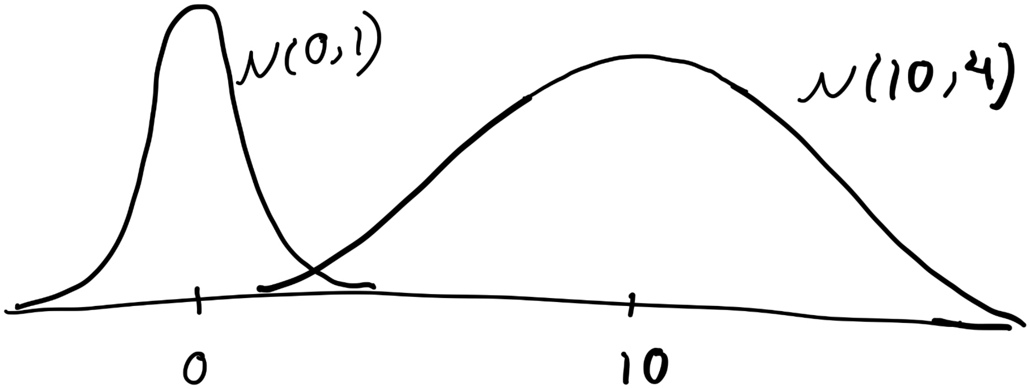

Figure 1.24

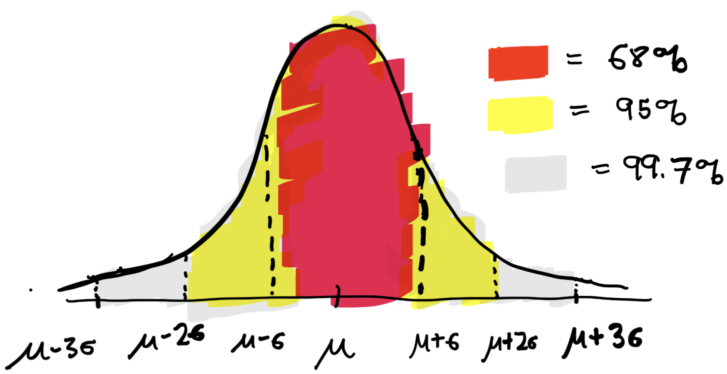

Figure 1.25: The area contained within one, two, and three standard deviations of the mean.

68%, 95%, and 99.7% of the Gaussian density's mass lies in the intervals $(\mu-\sigma, \mu+\sigma)$, $(\mu-2\sigma, \mu+2\sigma)$, and $(\mu-3\sigma, \mu+3\sigma)$, respectively. See Figure 1.25. Since most of the mass is contained in a bounded region, it can still make sense to model positive random variables (such as a midterm grade distribution...) as Gaussian, even though technically, the Gaussian can take negative values as well.

The standard Gaussian cdf. If we plug in $\mu = 0, \sigma = 1$ to (1.21), we get

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

Since densities must integrate to 1, we must have

$$\sqrt{2\pi} = \int_{-\infty}^{\infty} e^{-x^2/2} \mathrm{d}x.$$

We won't prove this identity, but you are free to use it (see Section **??** for more on using integral identities). But what about the standard Gaussian cdf, which is also given by an integral of the pdf?

It turns out that there does not exist a nice, closed-form expression for the cdf. Instead, we just give it a name, Φ, and don't simplify further. So

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t.$$

The function Φ is called a "special function" in mathematics. Although the integral cannot be simplified further, we have algorithms for approximating $\Phi(x)$ arbitrarily well. So if you've reduced some probability calculation to Φ, you're good to go. For example, suppose you want to compute $P(-1 < Z < 2)$ for $Z \sim \mathcal{N}(0,1)$. Since we get the probability by integrating the pdf, we have

$$P(-1 < Z < 2) = \int_{-1}^{2} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t.$$

This integral is also not computable in closed-form (the basic problem both here and with the cdf is that we can't write down the antiderivative of $e^{-t^2/2}$). But we can use that

$$P(-1 < Z < 2) = \Phi(2) - \Phi(-1), \tag{1.22}$$

and then compute the value of Φ numerically. On a test, if I ask you to compute some Gaussian probability, you should always write it in terms of Φ, and that's your final answer.

Linear transformations of Gaussian random variables and standardization. But what about probabilities involving nonstandard Gaussians? Do we need a separate special function Φ_{μ,σ^2} for each cdf? It turns out the answer is no, because we can always transform $\mathcal{N}(\mu,\sigma^2)$ to $\mathcal{N}(0,1)$. To explain this, I start with the following fact:

$$\textbf{Fact:} \quad \text{if } X \sim \mathcal{N}(\mu,\sigma^2), \text{ then } aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2). \tag{1.23}$$

The important thing to note here is that if X is Gaussian then so is $aX + b$ for any a and b. This is not true for all families of distributions! For example, if X is Bernoulli, then $1.5X + 2$ is certainly not Bernoulli. You will prove (X Gaussian) $\implies$ ($aX + b$ Gaussian) in Lab 7. Once we know this to be true, the two *parameters* of the normal distribution describing $aX + b$ are easy to derive. We know the first parameter is the mean and the second is the variance, so

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b = a\mu + b, \qquad \mathrm{var}(aX + b) = a^2\mathrm{var}(X) = a^2\sigma^2.$$

Now, how do we bring $X \sim \mathcal{N}(\mu,\sigma^2)$ to a standard normal? We map X to $aX + b$, and choose a, b to ensure $\mathbb{E}[aX + b] = 0$ and $\mathrm{var}(aX + b) = 1$. Since linear transformations of Gaussians are Gaussian, we must then have $aX + b \sim \mathcal{N}(0,1)$. Solving

$$0 = \mathbb{E}[aX + b] = a\mu + b, \qquad 1 = \mathrm{var}(aX + b) = a^2\sigma^2$$

for a, b gives $a = 1/\sigma$ and $b = -\mu/\sigma$. Putting it together, we have shown that

$$X \sim \mathcal{N}(\mu,\sigma^2) \quad \implies \quad \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

This is known as *standardization*.

Returning to the question of how to compute some probability like $P(a < X < b)$ where $X \sim \mathcal{N}(\mu, \sigma^2)$: we will *standardize* X to take advantage of Φ as follows:

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

To get the second line, we have simply renamed $(X-\mu)/\sigma$ as Z, which is the usual letter for a random variable that has a standard Gaussian distribution.

1.6.3 Exponential and gamma distribution

Definition 1.17. A random variable X has the **exponential distribution** with **rate** parameter β if its density is

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \beta e^{-\beta x}, & x \geq 0. \end{cases} \tag{1.24}$$

We write $X \sim \text{Exp}(\beta)$.

Waiting times are often modeled as exponential distributions; for example, the time until the first customer of the day arrives at a shop. By integrating f_X, we get that the cdf of $\text{Exp}(b)$ is

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-bx}, & x \geq 0. \end{cases}$$

The exponential distribution has the memorylessness property (work it out using the cdf!):

$$P(X > t + s \mid X > t) = P(X > s).$$

This makes sense: if a customer has not come to the shop in the first t minutes, this should not affect whether or not a customer will come to the store in the next s minutes.

The exponential distribution is a special case of the following family of distributions:

Definition 1.18. X has a **gamma distribution** if its density is

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0, \end{cases} \tag{1.25}$$

where $\alpha > 0$ is the **shape** parameter, and $\beta > 0$ is the **rate** parameter. We denote this $X \sim \text{Gamma}(\alpha, \beta)$.

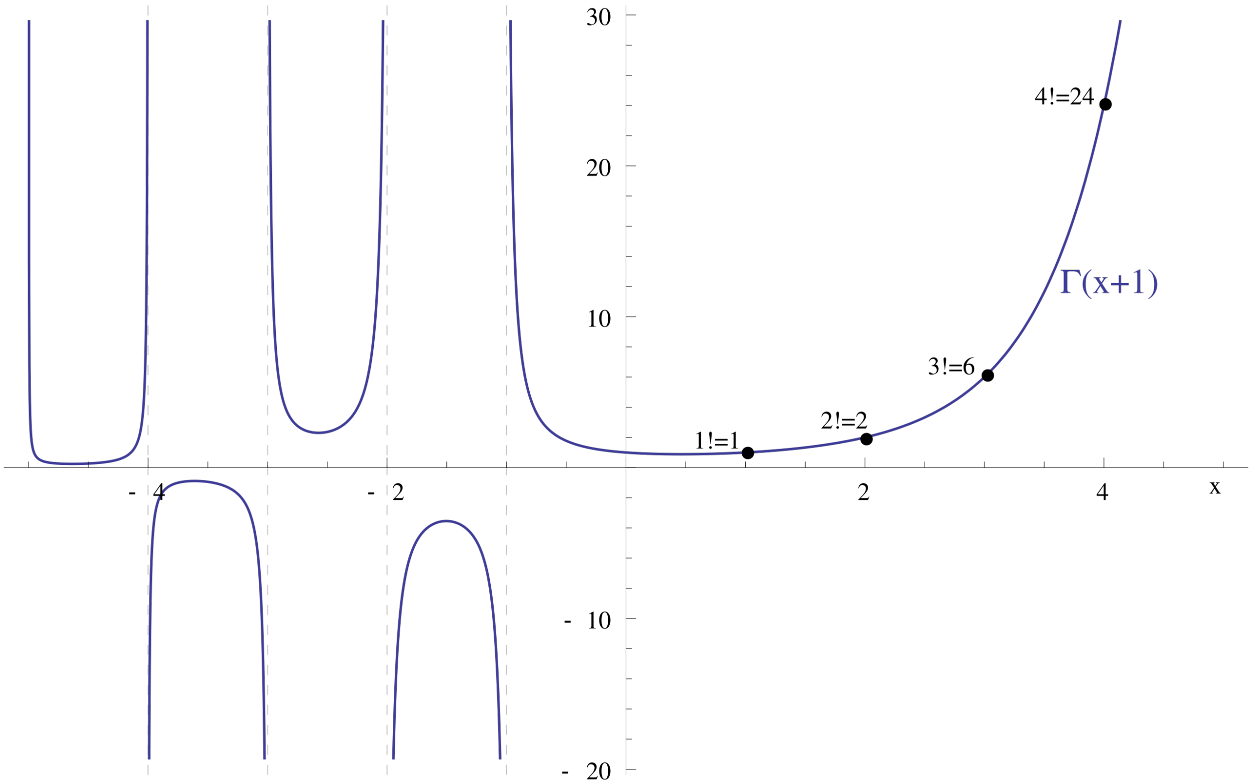

Figure 1.26: The gamma function is a continuous function that interpolates the factorials.

In this definition, $\beta^\alpha/\Gamma(\alpha)$ is a normalizing constant which ensures the density integrates to 1. The function $\Gamma(\alpha)$ is defined as follows.

Definition 1.19. The **gamma function** $\Gamma : (0, \infty) \to \mathbb{R}$ is given by

$$\Gamma(a) = \int_0^\infty y^{a-1}e^{-y}\,\mathrm{d}y, \quad a > 0. \tag{1.26}$$

Plugging in $a = 1$ we get $\Gamma(1) = \int_0^\infty e^{-y}\mathrm{d}y = 1$. Therefore, according to Definition 1.18, a Gamma$(1, \beta)$ distribution has pdf

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \beta e^{-\beta x}, & x > 0. \end{cases}$$

This is precisely the Exp(β) distribution we saw before. So Gamma$(1, \beta)$ = Exp(β).

Theorem 1.12. The gamma function satisfies the recursion

$$\Gamma(a + 1) = a\Gamma(a). \tag{1.27}$$

Proof. We apply integration by parts with

$$\begin{aligned}
u &= x^a \\
\mathrm{d}u &= ax^{a-1}\,\mathrm{d}x \\
v &= -\exp(-x) \\
\mathrm{d}v &= \exp(-x)\,\mathrm{d}x.
\end{aligned}$$

Then for any $a > 0$, we have

$$\begin{aligned}
\Gamma(a+1) &= \int_0^\infty x^a \exp(-x)\,\mathrm{d}x \\
&= \underbrace{\left[-x^a \exp(-x)\right]_0^\infty}_{=0} - \int_0^\infty -\exp(-x)ax^{a-1}\,\mathrm{d}x \\
&= a\int_0^\infty \exp(-x)x^{a-1}\,\mathrm{d}x \\
&= a\Gamma(a).
\end{aligned}$$

$\square$

Corollary 1.13. If $n = 1, 2, 3, \ldots$, then

$$\Gamma(n) = (n-1)! \tag{1.28}$$

Proof sketch. A rigorous proof follows by mathematical *induction*, but informally, if you have an integer $n \geq 2$, then Theorem 1.12 implies that

$$\begin{aligned}
\Gamma(n) &= \Gamma(n-1+1) \\
&= (n-1)\Gamma(n-1) \\
&= (n-1)(n-2)\Gamma(n-2) \\
&= (n-1)(n-2)(n-3)\Gamma(n-3) \\
&\vdots \\
&= (n-1)(n-2)(n-3)\cdots 3\cdot 2\cdot \Gamma(2) \\
&= (n-1)(n-2)(n-3)\cdots 3\cdot 2\cdot 1\cdot \Gamma(1) \\
&= (n-1)!
\end{aligned}$$

The last line follows because $\Gamma(1) = \int_0^\infty \exp(-x)\,\mathrm{d}x = 1$ (check this!). $\square$

This means that the gamma function is a continuous function that interpolates the factorials, as you see in Figure 1.26. Neat!

Example 1.16. Let's check that the pdf integrates to 1:

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x &= \int_0^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\,\mathrm{d}x \\
&= \int_0^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} (u/\beta)^{\alpha-1} e^{-u} \frac{1}{\beta}\mathrm{d}u && x = u/\beta,\ \mathrm{d}x = \mathrm{d}u/\beta \\
&= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} u^{\alpha-1} e^{-u}\,\mathrm{d}u \\
&= \frac{1}{\Gamma(\alpha)}\Gamma(\alpha) \\
&= 1.
\end{aligned}$$

Remark 1.13. We will defer the calculation to later, but $\mathbb{E}[X] = \alpha/\beta$ and $\mathrm{var}(X) = \alpha/\beta^2$.

1.6.4 "Massage and squint"

Here are two technical points that are useful when performing calculations in probability:

- every probability density is an integral identity in disguise;
- instead of antidifferentiating, we will often simplify integrals by massaging them with constants and then squinting at them until we recognize that it is the pdf of a familiar distribution. Then we apply an identity. So massage + squint.

In general, all densities have the following format:

$$f(x;\,\boldsymbol{\theta}) = \underbrace{c(\boldsymbol{\theta})}_{\text{normalizing constant}} \underbrace{k(x;\,\boldsymbol{\theta})}_{\text{kernel}}.$$

Here, $\boldsymbol{\theta}$ is generic notation for the *parameters* of the distribution. So $\boldsymbol{\theta} = (\mu,\,\sigma^2)$ for the normal, or $\boldsymbol{\theta} = (\alpha,\,\beta)$ for the gamma. $k(x;\,\boldsymbol{\theta})$ is the **probability density kernel** (**pdk**) or *unnormalized* density. It's the part of the density formula that actually depends on the argument x. $c(\boldsymbol{\theta})$ is the normalizing constant that hangs out in the front and serves just to make sure that the whole thing integrates to 1. Since

$$\int_{-\infty}^{\infty} f(x;\,\boldsymbol{\theta})\,\mathrm{d}x = \int_{-\infty}^{\infty} c(\boldsymbol{\theta})k(x;\,\boldsymbol{\theta})\,\mathrm{d}x = c(\boldsymbol{\theta})\int_{-\infty}^{\infty} k(x;\,\boldsymbol{\theta})\,\mathrm{d}x = 1,$$

then

$$\int_{-\infty}^{\infty} k(x;\,\boldsymbol{\theta})\,\mathrm{d}x = \frac{1}{c(\boldsymbol{\theta})},$$

So every probability density function gives you a free integral identity that you can use.

Example 1.17. For the Gaussian density:

$$\begin{aligned}
c(\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \\
k(x;\,\boldsymbol{\theta}) &= \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right).
\end{aligned}$$

So you get

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

Example 1.18. For the gamma density:

$$c(\boldsymbol{\theta}) = \frac{\beta^\alpha}{\Gamma(\alpha)}$$
$$k(x;\, \boldsymbol{\theta}) = x^{\alpha-1}e^{-\beta x}$$

So you get

$$\int_0^{\infty} x^{\alpha-1}e^{-\beta x}\, \mathrm{d}x = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

Every time we encounter a "famous" distribution family, its density will imply a new integral identity that we can use in subsequent computations.