Chapter 13

Normal Distribution

Review

A continuous random variable, X can take any value in an interval. We assign probabilities to the intervals as areas under a curve $f(x)$.

$f(x)$ is called a probability density function (PDF) and must satisfy two conditions:

$$f(x) \geq 0, \quad \forall x, \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

The CDF $F(x)$ is defined as usual:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

To go from CDF to PDF, we use the Fundamental Theorem of Calculus:

$$f(x) = \frac{d}{dx}F(x).$$

Review

We can calculate the mean, variance of a continuous random variable using similar formulas as before, except the summations are replaced by integration.

$$\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx,$$

$$\sigma^2 = Var[X] = E\left[X^2\right] - \mu^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \mu^2.$$

Review

For a continuous random variable, X, we often also calculate specific percentiles.

- The $100 \times p$ percentile is the number q such that

$$P(X < q) = p.$$

- A popular percentile to calculate is the 50th (or median).

The normal random variable

The normal random variable (also called the Gaussian random variable) plays a very important role in statistics.

It provides a good model for many numerical populations. For example, biological measurements such as height, weight and measurement error in scientific experiments are well approximated by a normal.

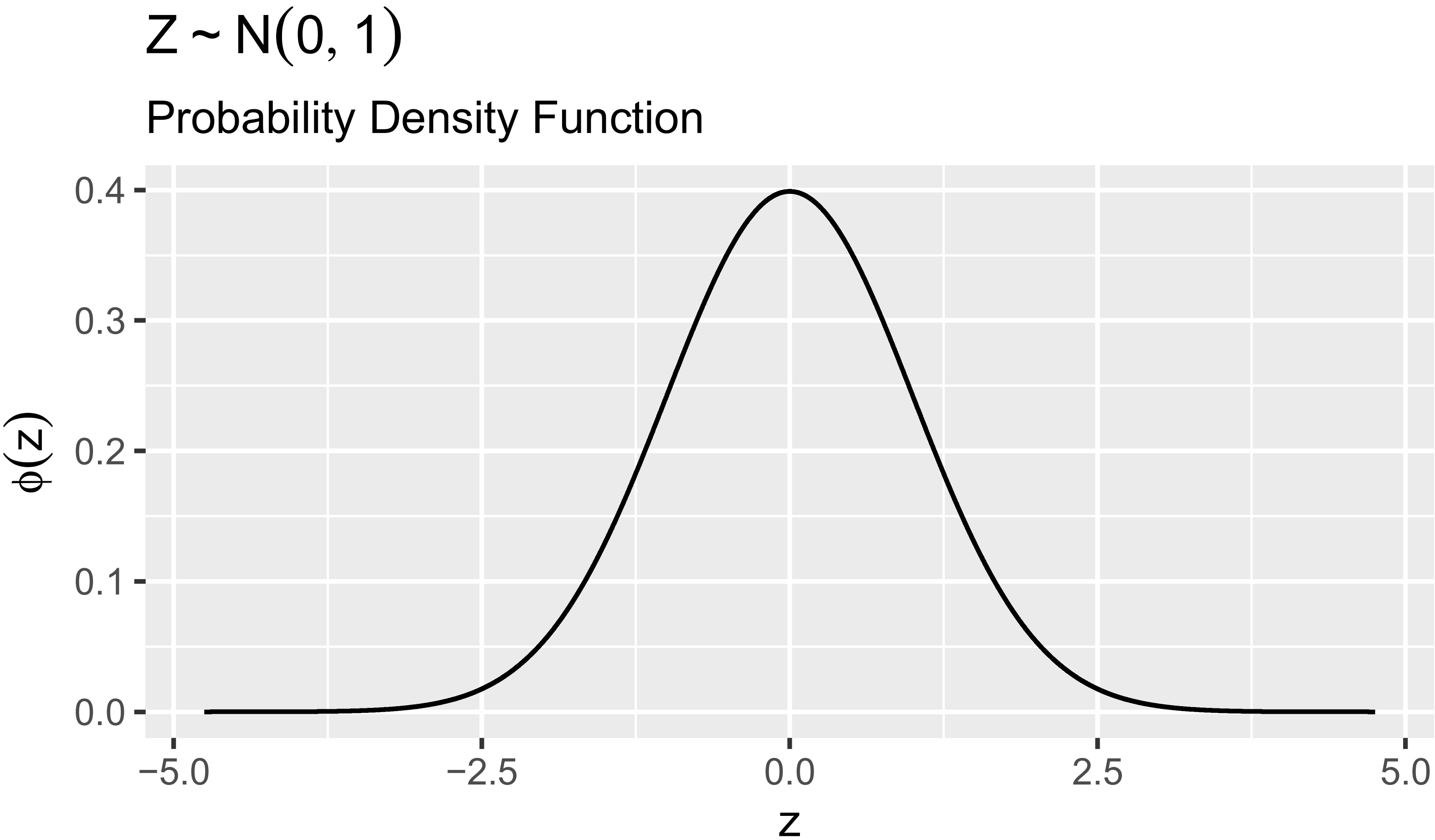
Many discrete distributions are also approximately normal, such as the binomial and Poisson, provided certain conditions on n, π, λ are met. This is a consequence of the **Central Limit Theorem**

The standard normal random variable

Definition 13.1 A continuous random variable Z has the **standard normal distribution** (denoted by $Z \sim Norm(0,1)$) if its PDF is given by

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \quad -\infty < z < \infty.$$

The use of the uppercase letter Z to denote a standard normal random variable and its PDF by $\phi(z)$ (instead of $f(z)$) is traditional in statistics.


$Z \sim N(0, 1)$
Probability Density Function
0.4
0.3
0.2
0.1
0.0
$\phi(z)$
−5.0
−2.5
0.0
2.5
5.0
z

We will take for granted that the standard normal PDF integrates to 1 over the whole real line. The approach involves converting to polar coordinates.

It is interesting to note however that even though it is possible to show that the total area under a standard normal PDF is 1, it is not possible to calculate the area over a smaller interval in closed form.

Probabilities for a standard normal random variable must therefore be evaluated using numerical integration in R (`pnorm`, `qnorm` will do this for you)

Calculations for standard normal in R

```
dnorm(x = 0)                         #returns value of phi(x)
## [1] 0.3989423

pnorm(q = 1)                         #returns P(Z <= 1)
## [1] 0.8413447

pnorm(q = 1) - pnorm(q = -1) #returns P(-1 < Z <= 1)
## [1] 0.6826895
pnorm(2) - pnorm(-2)
## [1] 0.9544997
pnorm(3) - pnorm(-3)
## [1] 0.9973002

qnorm(p = 0.5)                        #returns 100*p percentile
## [1] 0
```

Example 13.1

Suppose $Z \sim Norm(0,1)$. In other words, its PDF is given by

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \quad -\infty < z < \infty$$

a. Draw a diagram of $\phi(z)$ and shade the area corresponding to the integral

$$\int_{-0.44}^{1.33} \phi(z) dz.$$

What probability does this integral above represent? How will you calculate this in R?

Example 13.1

Suppose $Z \sim Norm(0, 1)$. In other words, its PDF is given by

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \quad -\infty < z < \infty$$

b. Which number is larger? Or are they the same?

$$A = \int_1^2 \phi(z) dz \ \text{ or } \ B = \int_{-2}^{-1} \phi(z)\, dz$$

Example 13.1

Suppose $Z \sim Norm(0,1)$. In other words, its PDF is given by

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \quad -\infty < z < \infty$$

c. Which number is larger? Or are they the same?

$$A = \int_{1.5}^{2.5} \phi(z)\, dz \quad \text{or} \quad B = \int_{0.5}^{1.5} \phi(z)\, dz$$

Example 13.1

Suppose $Z \sim Norm(0, 1)$. In other words, its PDF is given by

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \quad -\infty < z < \infty$$

d. For what value of q is the following statement true? How will you calculate it in R?

$$P(Z < q) = 0.33$$

Mean and variance of a standard normal

Theorem 13.1 Let $Z \sim Norm(0, 1)$. Then

a. $E[Z] = 0$

b. $Var[Z] = 1$

Theorem 13.1 explains the notation $Norm(0, 1)$; 0 and 1 are the *mean* and *standard deviation* of the standard normal distribution respectively.

Proof of Theorem 13.1 part a

These results follow from the fact that the standard normal PDF is an even function - symmetric about 0. In other words, for any $z > 0$ we have:

$$\phi(z) = \phi(-z)$$

By definition of the expected value:

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty} z \cdot \phi(z) dz, \\ &= \underbrace{\int_{-\infty}^{0} z \cdot \phi(z) dz}_{I_1} + \underbrace{\int_{0}^{\infty} z \cdot \phi(z) dz}_{I_2}. \end{aligned}$$

Proof of Theorem 13.1 part a.

In the integral denoted as I_1, we make the substitution

$$u = -z \quad \Rightarrow du = -dz$$

to obtain

$$\begin{aligned}
I_1 &= \int_{-\infty}^{0} z \cdot \phi(z) dz = \int_{\infty}^{0} -u \cdot \phi(-u) \cdot -du, \\
&= -\int_{0}^{\infty} u \cdot \phi(-u) \cdot du, \quad \text{flip limits} \\
&= -\int_{0}^{\infty} u \cdot \phi(u) \cdot du \qquad evenfunction \\
&= -I_2.
\end{aligned}$$

Hence

$$E[Z] = -I_2 + I_2 = 0.$$

$Norm(\mu, \sigma)$ random variable

Definition 13.2 A continuous random variable X is said to be normally distributed with parameters μ and σ if its PDF is given by

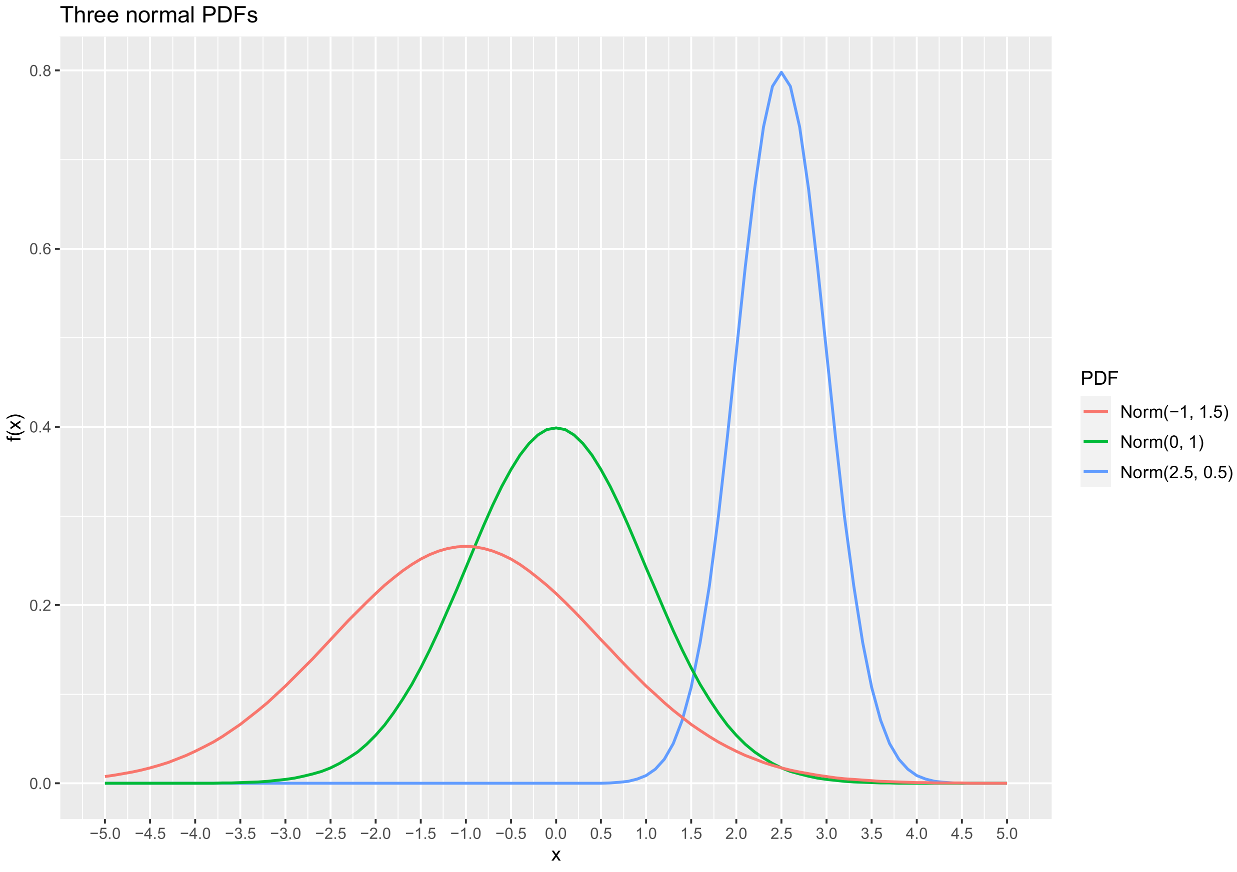
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

We write $X \sim Norm(\mu, \sigma)$[1].

Note: if we substitute $\mu = 0$ and $\sigma = 1$ in the PDF of X, we get the PDF of the standard normal variate Z.

[1] Some authors use σ^2 in place of σ in *Norm*

The PDF of X is shown for three different choices of μ and σ. The PDF is always symmetric around the value of μ and σ controls the spread of the curve.



Some results for *Norm*(μ, σ) random variable

Lemma 13.1 Let $X \sim Norm(\mu, \sigma)$. Then

a. $Z = \frac{X - \mu}{\sigma} \sim Norm(0, 1)$

b. $E[X] = \mu$

c. $Var[X] = \sigma^2$

Proof of Lemma 13.1 a

Let $X \sim \mathit{Norm}(\mu, \sigma)$. Then we want to show that the random variable Z defined as

$$Z = \frac{X - \mu}{\sigma} \sim \mathit{Norm}(0, 1)$$

We show the result by using the so-called CDF method.

This means we will first work out the CDF of Z and then differentiate it to get the PDF of Z.

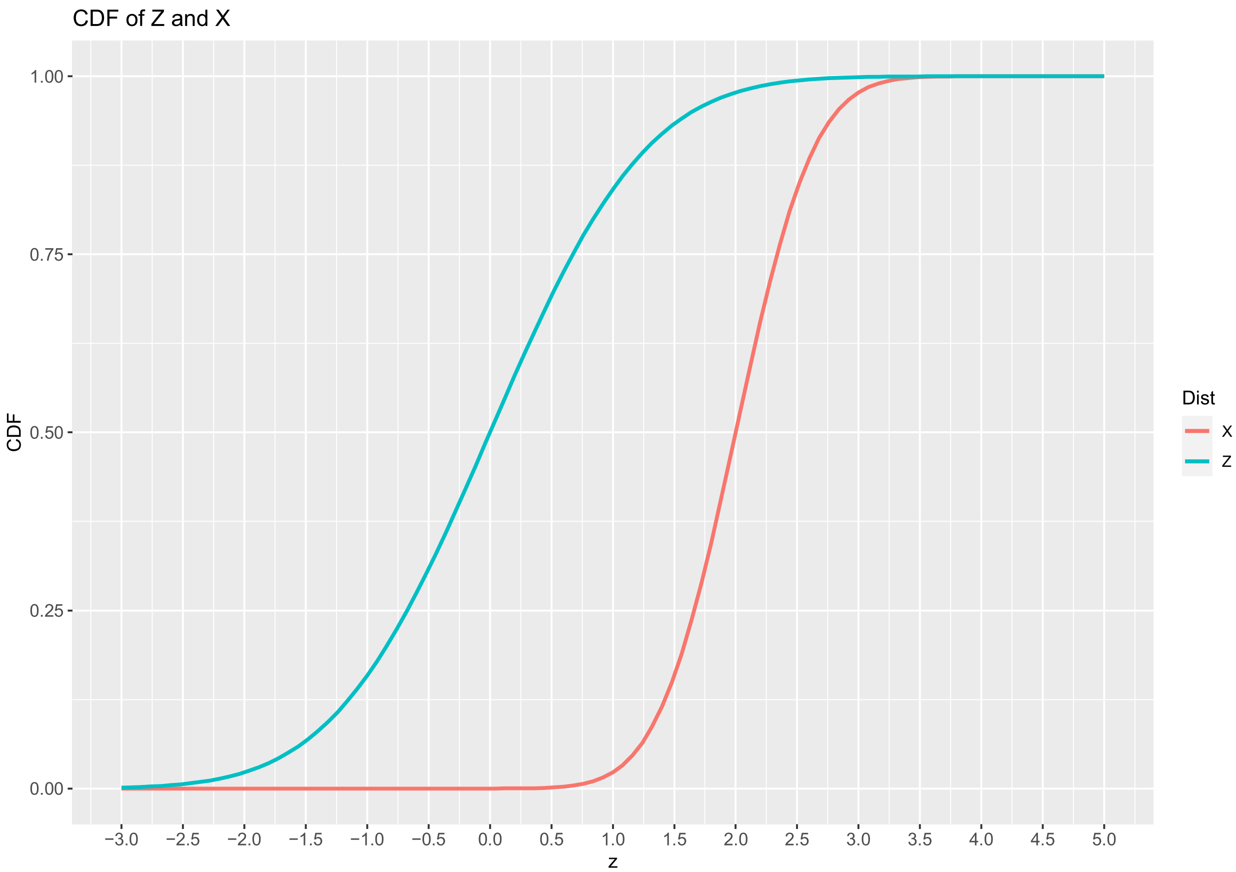
If the PDF resembles the PDF of a standard normal, then we are done.

First notice that the CDF of $Z = \frac{X-\mu}{\sigma}$ and the CDF of X are related to each other as follows.

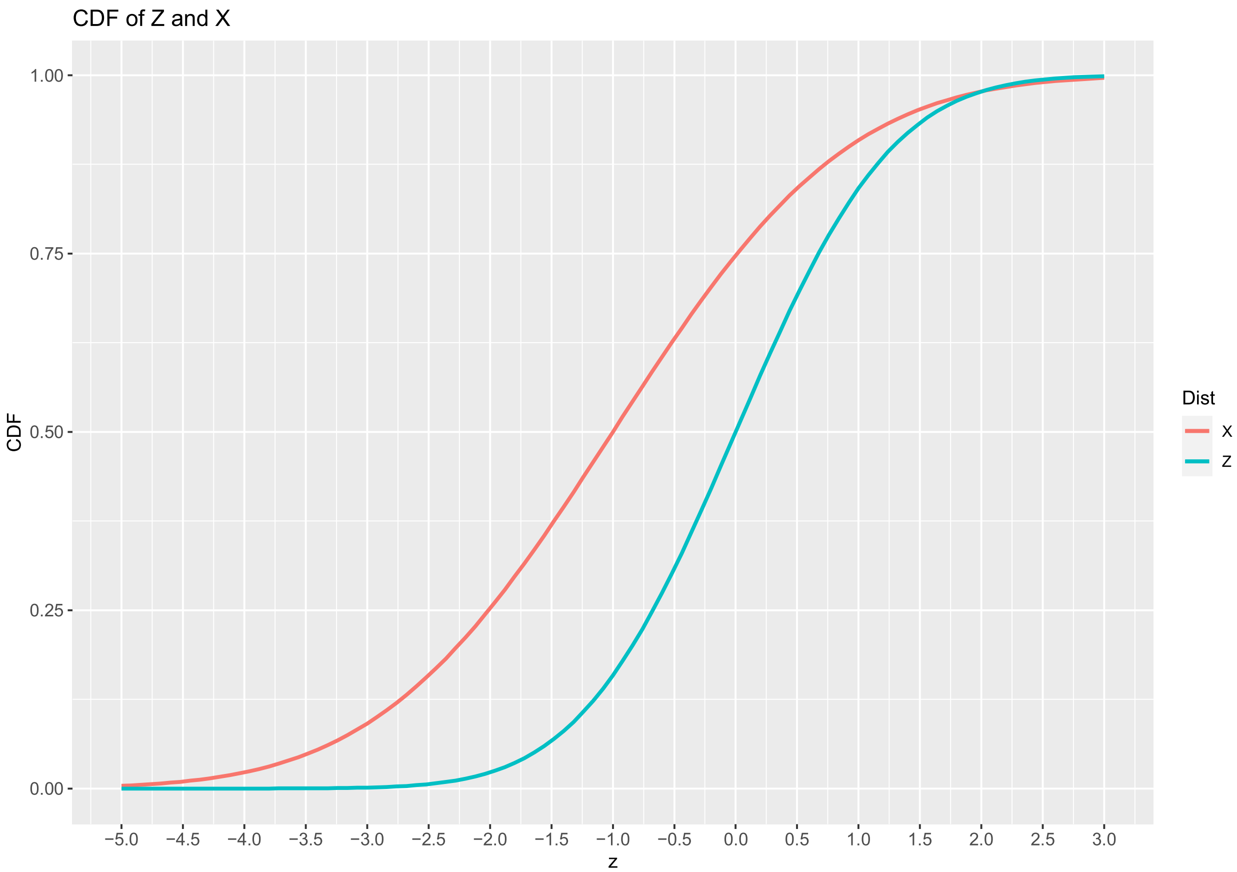
$$\begin{aligned} F_Z(z) &= \text{CDF of Z at a value z} \\ &= P(Z \leq z) \\ &= P\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= P(X \leq \mu + \sigma z), \\ &= F_X(\mu + \sigma z). \end{aligned}$$

In other words, the CDF of Z at a value z is equal to the CDF of X at the value $\mu + \sigma z$.

This is illustrated for an example with $\mu = 2$ and $\sigma = 0.5$.



Here is another example with $\mu = -1$ and $\sigma = 1.5$.



By the Fundamental Theorem of Calculus, a PDF of Z can now be found by differentiating its CDF:

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_Z(z) \\ &= \frac{d}{dz} F_X(\mu + \sigma z). \end{aligned}$$

Since the argument inside F_X involves $v(z) = \mu + \sigma z$, we need to use the chain rule of differentiation which states:

$$\begin{aligned} \frac{d}{dz} F(v(z)) &= \frac{d}{dv(z)} F(v(z)) \, \frac{dv(z)}{dz}, \\ &= f(v(z)) \cdot \sigma. \end{aligned}$$

Using the chain rule for differentiation, we obtain

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_X(\mu + \sigma z) \\ &= f_X(\mu + \sigma z) \times \sigma \\ &= \frac{1}{\sigma} \phi(z) \times \sigma \qquad \text{(why?)} \\ &= \phi(z) \end{aligned}$$

Therefore, we have shown that $Z = \frac{X-\mu}{\sigma} \sim Norm(0, 1)$.

Note that Lemma 13.1 a shows us that any normal random variable X is simply a linear transformation of the standard normal variable Z.

$Z = \frac{X-\mu}{\sigma} \Rightarrow X = \mu + \sigma Z$.

The expected value and variance claims in parts b and c therefore simply follow from the algebra of expectations.

Calculations in R for $X \sim Norm(\mu, \sigma)$

We can calculate normal probabilities using the general version of `pnorm` in R that has arguments for specified means and standard deviations.

```
pnorm(q = 5, mean = 3, sd = 2)  # Finds P(X <= 5) for X~Norm(3,2)
```

```
## [1] 0.8413447
```

It can also be useful to relate the probabilities for any arbitrary normal random variable to the standard normal random variable as shown below.

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

```r
#these two should return the same value

pnorm(q = 5, mean = 3, sd = 2)
```

```
## [1] 0.8413447
```

```r
pnorm(q = (5-3)/2 )
```

```
## [1] 0.8413447
```

The ratio $\frac{x-\mu}{\sigma}$ is called the z-score for x. It tells us how many standard deviations above or below the mean the value x falls.

For example, if $X \sim Norm(3, 2)$ then an $x = 5$ is 1 standard deviation above the mean. Hence, its z-score is 1.

For the purposes of computing probabilities, the z-scores suffice.

Example 13.2

Suppose $X \sim \mathit{Norm}(\mu, \sigma)$. Calculate the probability that X lies within 1 standard deviation of the mean. Does your answer depend on the value of μ or σ?

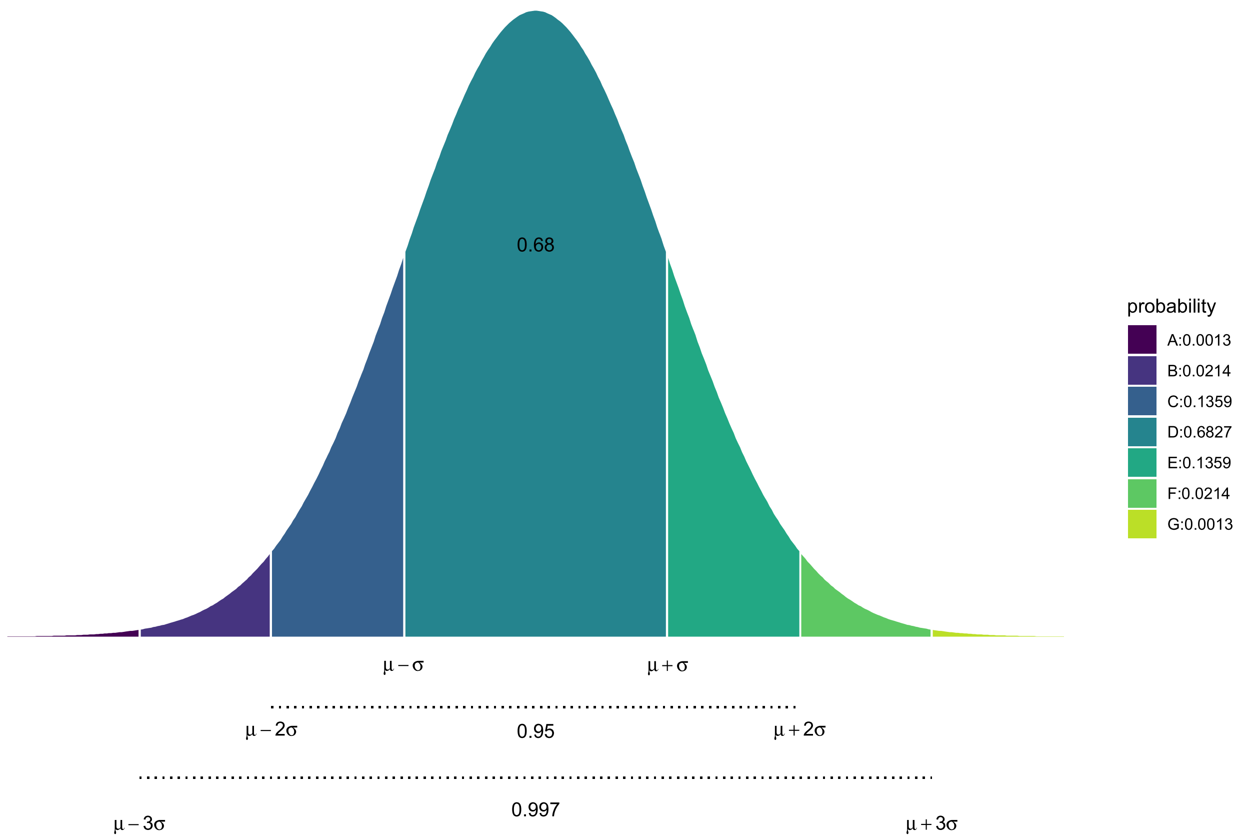
From Example 13.2 we can generalize that for any normal distribution, the area contained within k standard deviations of the mean is the same as the area within $\pm k$ under a standard normal distribution.

In particular, for *any* normal distribution, we can therefore state that

- approximately 68% of the distribution lies within 1 standard deviation of the mean
- approximately 95% of the distribution lies within 2 standard deviations of the mean
- approximately 99.7% of the distribution lies within 1 standard deviation of the mean

These benchmarks are known as the **empirical rule** or the 68-95-99.7 rule.

68-95-99.7 rule



Calculations in R for $X \sim Norm(\mu, \sigma)$

The function `qnorm` can be used to find percentiles of any normal distribution.

```
#20th percentile of X ~ Norm(3,2)


qnorm(p = 0.2, mean = 3, sd = 2)
```

```
## [1] 1.316758
```

```
#20th percentile of Z ~Norm(0,1)
qnorm(p=0.2)
```

```
## [1] -0.8416212
```

```
#how the two are related
qnorm(p=0.2)*2 + 3  #x = sigma*z + mu
```

```
## [1] 1.316758
```

Example 13.3

In most states a motorist is legally drunk, or driving under the influence (DUI), if their measured blood alcohol concentration (BAC)is found to be 0.08% or higher. Experience has shown that repeated breath analyzer measurements taken from the same person produce a distribution of responses that can be described by a normal PDF with μ equal to the person's true blood alcohol concentration and σ equal to 0.004%.

Suppose a driver's true BAC is 0.075%. What are the chances they will be incorrectly booked on a DUI charge?

Example 13.4

It is estimated that 80% of all St. Bernard dogs have weights between 134.4 and 185.6 pounds. Assuming that the weight distribution can be described by a normal distribution, and assuming that 134.4 and 185.6 are equally distant from the mean, determine the standard deviation of the distribution.

Example 13.5

A company produces jam in cardboard containers. An empty container weighs 1.5 ounces. They fill the container by putting it on a scale and pour in jam until the scale shows the value m. However, suppose the scale has a measurement error and therefore the actual amount of jam in the container is $(m - 1.5) + X$ where X is a random variable which has a normal distribution with mean 0 and a standard deviation of 0.25 ounces.

How should you choose m if you want that 95% of the containers should contain at least 16 ounces of jam?