Chapter 13

Other Continuous Distributions

References: Pruim 3.4

As with discrete random variables, we will now catalog some famous continuous distributions that have great potential as models for data. We begin with the familiar bell-shaped curve also known as the normal distribution.

13.1 Normal Distribution

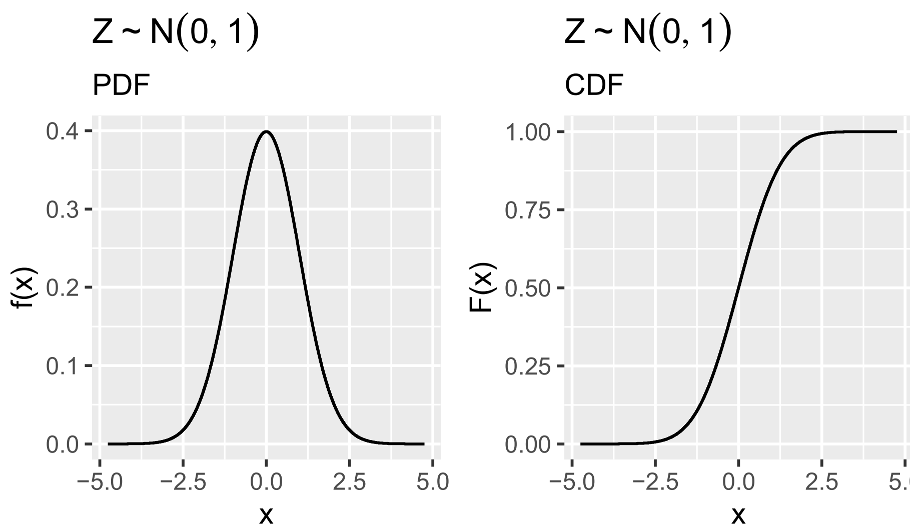
The normal distribution (also called the Gaussian distribution) plays a very important role in statistics. It provides a good model for many numerical populations. For example, biological measurements such as height, weight and measurement error in scientific experiments are well approximated by a normal. 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** which we will discuss in STAT 341.

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 Z for a standard normal random variable, and of ϕ for its PDF, and Φ for its CDF is traditional in statistics.

As the graphs below show, the standard normal PDF is symmetric around 0 and follows a bell-shaped curve. The CDF is a strictly increasing, continuous function with a familiar sigmoidal (or S) shape.



> We will take for granted that the normal PDF integrates to 1 over the whole real line. The approach involves converting to polar coordinates. Please see Theorem 3.4.2 on page 156 if you are interested in the proof.

It is important to note however that even though it is possible to show that

$$\int_{-\infty}^{\infty} \phi(z)dz = 1,$$

it is not possible to evaluate the CDF

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \phi(z)dz$$

in closed form. Probabilities for the standard normal random variable must therefore be evaluated using numerical integration.

As usual R supplies the d-, p-, q- and r- functions **dnorm**, **pnorm**, **qnorm** and **rnorm**. It is useful to memorize some benchmark values for a Norm(0,1).

```
qnorm(p = 0.975)                    #returns z: F(z) = 0.975
## [1] 1.959964
qnorm(p = 0.025)                    #returns z: F(z) = 0.025
```

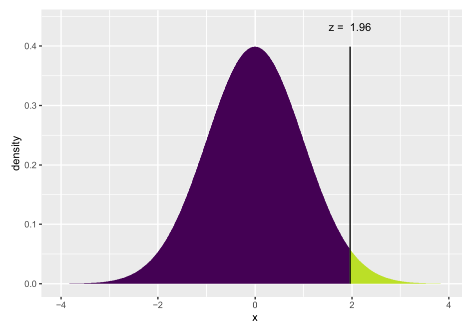
```
## [1] -1.959964

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

An enhanced version of these functions from the `mosaic` package will additionally draw the curve and shade in relevant areas.

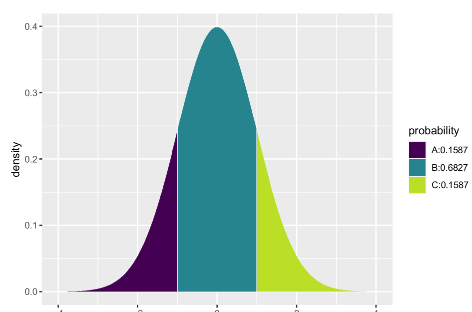
```
library(mosaic)

xqnorm(p=0.975)
```




```
## [1] 1.959964
```

```
xpnorm(q = c(-1,1))
```




```
## [1] 0.1586553 0.8413447
```

....

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

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

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

b. Which number is larger? Calculate it in R to verify your thinking. $\int_{1}^{2} \phi(z)dz$ or $\int_{-2}^{-1} \phi(z)dz$?

c. Which number is larger? $\int_{1.5}^{2.5} \phi(z)dz$ or $\int_{0.5}^{1.5} \phi(z)dz$? Verify your thinking using R.

d. For what value of q is the following statement true? Use R to find out.

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

We can calculate the mean and variance of a standard normal random variable and this is shown below.

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

- $E[Z] = 0$
- $Var[Z] = 1$

Proof. 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:

$$f(z) = f(-z)$$

By definition of the expected value:

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

In the integral denoted as I_1, we make the substitution

$$u = -z$$

which then implies

$$du = -dz$$

to obtain

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

Hence

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

> A function, g is said to be odd if $g(x) = -g(x)$ for $x > 0$. The integral of an odd function over a symmetric range results in 0. The expected value $E[Z]$ is an odd function and hence the result.

We will now calculate the variance using the short-cut formula:

$$Var[Z] = E[Z^2] - (E[Z])^2 .$$

We know $E[Z] = 0$ and all we need to do is find $E[Z^2]$.

$$\begin{aligned} E[Z^2] &= \int_{-\infty}^{\infty} z^2 \cdot \phi(z) \cdot dz \\ &= \int_{-\infty}^{\infty} z^2 \cdot \underbrace{\frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz}_{\phi(z)} . \end{aligned}$$

Using integration by parts with the following definitions:

$$u = z, \qquad dv = ze^{-z^2/2}$$

$$du = dz, \quad v = \int ze^{-z^2/2} dz = -e^{-z^2/2}$$

we have:

$$\begin{aligned} \int_{-\infty}^{\infty} z^2 \cdot e^{-z^2/2} &= -z \cdot e^{-z^2/2}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-z^2/2} dz. \\ &= 0 + \int_{-\infty}^{\infty} e^{-z^2/2} dz. \end{aligned}$$

The first term is equal to 0 since $e^{-z^2/2} \to 0$ much faster than any polynomial in z. Therefore

$$\begin{aligned} E[Z^2] &= \int_{-\infty}^{\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \\ &= 1, \end{aligned}$$

since the standard normal PDF integrates to 1 over the whole real line.

Hence

$$Var\left[Z\right] = E\left[Z^2\right] = 1.$$

□

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

We can obtain normal distributions with any mean and standard deviation we desire by using a linear transformation of Z.

Definition 13.2. A continuous random variable X has a **normal distribution** with parameters μ and σ (> 0) if its PDF is given by

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

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.

Some results for "arbitrary" normal distributions are stated below.

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

- $Z = \frac{X-\mu}{\sigma} \sim Norm(0, 1)$
- $E\left[X\right] = \mu$
- $Var\left[X\right] = \sigma^2$

Proof. The second and third claims follow from the results about the expected value and variance of linear transformations:

$$E\left[X\right] = \mu + \sigma E\left[Z\right] = \mu.$$
$$Var\left[X\right] = \sigma^2 Var\left[Z\right] = \sigma^2.$$

In order to prove the first statement, we begin by deriving the CDF of Z. For any $z \in \mathbb{R}$, we have

$$\begin{aligned} F_Z(z) &= P(Z \leq z), \\ &= P\left(\underbrace{\frac{X-\mu}{\sigma}}_{Z} \leq z\right) \\ &= P(X \leq \mu + \sigma z), \\ &= F_X(\mu + \sigma z). \end{aligned}$$

where F_Z is the CDF of Z and F_X the CDF of X.

By the Fundamental theorem of Calculus, a PDF for Z can then be found by differentiating $F_Z(z)$ with respect to z:

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_Z(z) \\ &= \frac{d}{dz} F_X(\mu + \sigma z), \\ &= f_X(\mu + \sigma z) \times \sigma && \text{(chain rule of differentiation)} \\ &= \frac{1}{\sigma}\phi(z) \times \sigma \\ &= \phi(z) \end{aligned}$$

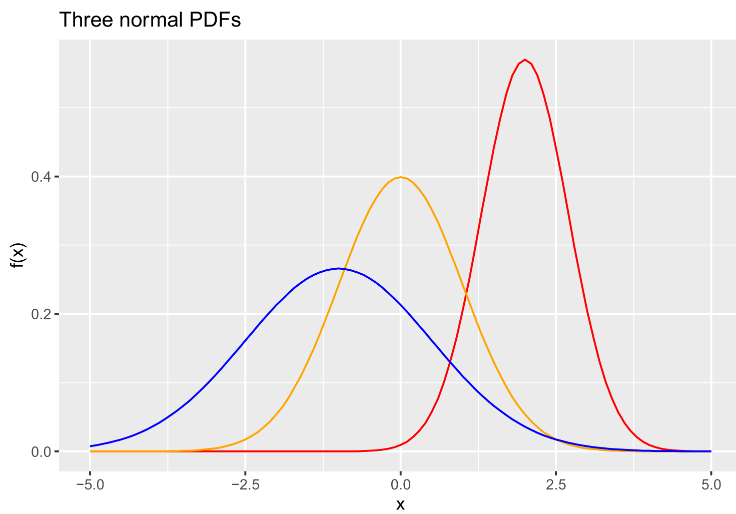
□

The normal distribution is somewhat special in the sense that its two parameters provide us with complete information on the shape and location of the distribution. This property is not unique to the normal PDF, but rather is shared by a family of PDFs known as the location-scale family. All normal distributions have a symmetric, bell-shaped PDF. The inflection point (where the second derivative changes sign) of a normal curve is always 1 standard deviation from the mean.

Three normal curves are shown below.

```
ggplot() +
     geom_function(fun = dnorm,
                   args = list(mean=2, sd = 0.7),
                   colour = "red",
                   xlim = c(-5,5)) +
     geom_function(fun = dnorm,
                   args = list(mean = 0, sd = 1),
                   colour = "orange",
                   xlim = c(-5,5)) +
     geom_function(fun = dnorm,
```

```
                  args = list(mean = -1, sd = 1.5),
                  colour = "blue",
                  xlim = c(-5,5)) +
     labs(title = "Three normal PDFs",
          x = "x",
          y = "f(x)")
```

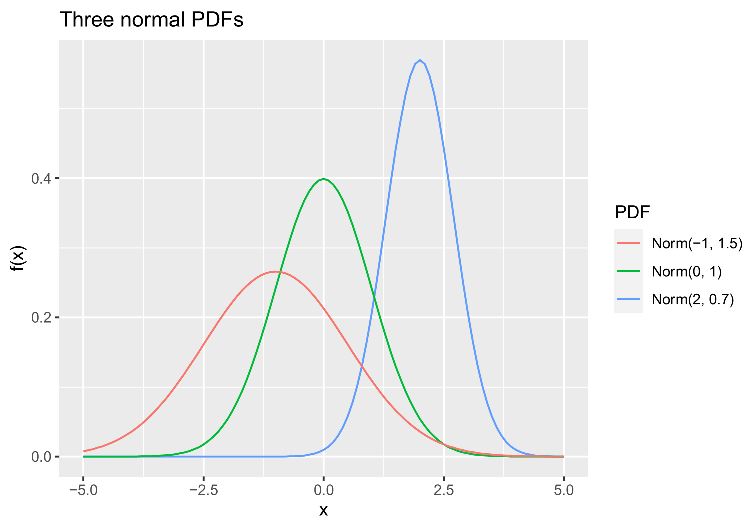


Notice how we don't have a legend telling us which color corresponds to which curve. According to the book ggplot2: Elegant graphics for Data Analysis, legends are guides for certain aesthetics - such as shape, fill, color, linetype - but only if they are created inside the `mapping = aes(...)`.

Let's revise the code above to map the color for each curve to a string. I have chosen to describe the distribution each curve corresponds to. The resulting graph is good enough for our purposes. But one could use a `scale_color_*()` layer to customize the legend, for instance, choosing the color palette, changing the heading of the legend etc.

```
ggplot() +
     geom_function(fun = dnorm,
                  args = list(mean=2, sd = 0.7),
                  mapping = aes(colour = "Norm(2, 0.7)"),
                  xlim = c(-5,5)) +
     geom_function(fun = dnorm,
                  args = list(mean = 0, sd = 1),
                  mapping = aes(colour = "Norm(0, 1)"),
                  xlim = c(-5,5)) +
     geom_function(fun = dnorm,
```

```
                 args = list(mean = -1, sd = 1.5),
                 mapping = aes(colour = "Norm(-1, 1.5)"),
                 xlim = c(-5,5)) +
    labs(title = "Three normal PDFs",
         x = "x",
         y = "f(x)",
         color = "PDF")
```



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) #5 is 1 SD above the mean
```

```
## [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.

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

```
#these two should return the same value

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

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

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

. . . .

Example 13.2. Suppose $X \sim Norm(\mu, \sigma)$. Calculate the probability that X lies within 1 standard deviation of the mean. Does your answer depend on the value of μ or σ?

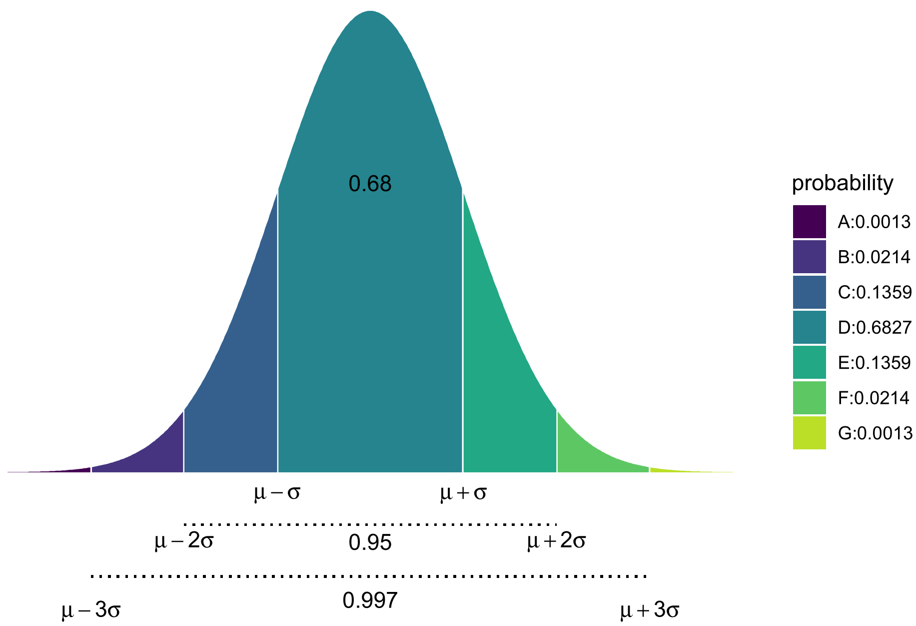
. . . .

We can generalize from Example 13.2 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 <u>any</u> 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 which is illustrated below.



Similarly, the function `qnorm` can be used to find the *pth* percentile of any normal distribution.

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

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

```
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?

. . . .

13.1.1 Practice Problems

1. Assume that $Z \sim Norm(0, 1)$. For what values of z are the following true?

a. $P(Z \geq z) = 0.2236$

b. $P(-1 \leq Z < z) = 0.5004$

c. $P(-z \leq Z < z) = 0.8$

2. Mensa is an international society whose membership is limited to those scoring above the 98th percentile on the Stanford-Binet IQ test. The general population has an average IQ of 100, with a standard deviation of 16. What is the lowest IQ score that will qualify an individual for membership in the society?

3. A college professor teaches Chemistry 101 each Autumn quarter to a large class of first-year students. For tests, she uses standardized exams that she knows from past experience produce bell-shaped grade distributions with a mean of 70 and a standard deviation of 12. Her grading imposes standards that in the long run yields 14% A, 20% B, 32% C , 20% D and 14% F. (You may assume grades above 100 are possible as are negative grades – although obviously these will occur rarely)

a. What is the minimum grade necessary in order to earn an A?

b. What is the minimum grade necessary in order to earn a B?

c. Suppose the mean score for a different instructor of Chemistry 101 is 68. What percentile does this score correspond to on our college professor's grade distribution?