**Normal probability distribution**

The most useful of all probability distributions for a continuous random variable is the normal distribution. The mathematical function for it is



where σ > 0. Notice that μ and σ are within the function. Thus, the mean and standard deviation control the shape of the curve that the function produces!

Again



and



In general, this is what a graph of the distribution looks like.





* The plotted curve consists of (y, f(y)) connected points.
* The distribution is symmetric about μ. Thus, P(Y > μ) = P(Y < μ) = 0.5. The parameter μ is often called a location parameter since it gives the central location of the distribution.
* The area under the curve is 1.
* The left and right sides of the curve extend out to - and + without touching the x-axis (although they will get very close). Note the plot above may be a little misleading with respect to this. The left and right sides of the distribution are often called the “tails” of the distribution.
* σ controls the scale of the distribution. The larger σ is the more spread out the distribution (large variability). The smaller σ is the less spread out the distribution (small variability). Below are three normal distributions demonstrating this (3\_normal.R).

> # Find f(y)

> dnorm(x = 24.3, mean = 24.3, sd = 0.6)

[1] 0.6649038

> # Plot

> curve(expr = dnorm(x = x, mean = 24.3, sd =

 0.6), xlim = c(20, 30), col = "darkgreen",

 lwd = 2, ylab = "f(y)", xlab = "y")

> curve(expr = dnorm(x = x, mean = 24.3, sd =

 1.3), xlim = c(20, 30), col = "blue", add =

 TRUE, lwd = 2)

> curve(expr = dnorm(x = x, mean = 23.1, sd =

 0.6), xlim = c(20, 30), col = "red", add =

 TRUE, lwd = 2)

> abline(h = 0)

> legend(x = 25, y = 0.6, legend = c("mu = 24.3,

 sigma = 0.6", "mu = 24.3, sigma = 1.3", "mu =

 23.1, sigma = 0.6"), lty = c(1,1,1), col =

 c("darkgreen", "blue","red"), bty = "n", lwd =

 c(2,2,2))



Note that the dnorm() function in R finds f(y). I recommend you use this program to investigate how other values of μ and σ affect the distribution.

* Some books refer to a random variable having a normal distribution as Y ~ N(μ, σ2), where “~” is read as “is distributed as” and “N” stands for “normal distribution.”
* A VERY IMPORTANT specific case of a normal distribution is the standard normal distribution. This distribution has μ = 0 and σ = 1. Therefore,



Typically, “Z” is used instead of “Y” to denote a standard normal random variable. This will be discussed more later.

* The CDF is . There is not a nice closed form expression for it. The pnorm() function in R calculates it for us at a particular value of y.
* Showing =1 involves making a transformation to polar coordinates. More simply, we can use Sage (see normal.ipnyb):



* To find E(Y) and Var(Y), we can again use Sage (normal.ipnyb):



Finding probabilities with the normal distribution

Example: Car MPG (car\_mpg.R)

Suppose that it is reasonable to assume a car’s MPG has a normal distribution with a mean of μ = 24.3 and a standard deviation of σ = 0.6. Let Y denote the MPG for one tank of gas. Answer the following questions.

1. Find the probability that a randomly selected car of the same type gets less than 23 MPG for one tank of gas.

We need to find P(Y < 23). This is the area to the left of the red line underneath the curve.



This probability can be found by

 .

Using methods from previous sections:

> pdf <- function(y, mu, sigma) {

 # Could use dnorm(x = y, mean = mu, sd = sigma)

 1/(sigma\*sqrt(2\*pi)) \* exp(-(y-mu)^2 / (2 \*

 sigma^2))

 }

> integrate(f = pdf, lower = -Inf, upper = 23, mu =

 24.3, sigma = 0.6)

0.01513014 with absolute error < 8.5e-07

and



Instead, a much easier way is to use the pnorm() function in R that evaluates the CDF.

> pnorm(q = 23, mean = 24.3, sd = 0.6)

[1] 0.01513014

The q argument stands for “quantile”; this word is used in a similar way as we saw in earlier in the course!

Notes:

* Does this probability value make sense relative to what we see in the plot?
* This probability is the CDF evaluated at y = 23: F(23)
* The pnorm() function gives the “area to the left underneath the curve”
1. Suppose σ is increased to σ = 1.3. What do you expect to happen to P(Y < 23)?

> pnorm(q = 23, mean = 24.3, sd = 1.3)

[1] 0.1586553



1. Suppose σ = 0.6 again, but μ is decreased to μ = 23.1. What do you expect to happen to P(Y < 23)?

> pnorm(q = 23, mean = 23.1, sd = 0.6)

[1] 0.4338162



Below is a nice comparative graph for the 3 examples above.



1. Suppose σ = 0.6 and μ = 24.3 again. What is P(23 < Y < 25)?

Note that P(23 < Y < 25) = P(Y < 25) – P(Y < 23) = F(25) – F(23).

> pnorm(q = 25, mean = 24.3, sd = 0.6) - pnorm(q =

 23, mean = 24.3, sd = 0.6)

[1] 0.8631974



1. Suppose σ = 0.6 and μ = 24.3 again. What is P(Y > 23)? Think in terms of the CDF!
2. Suppose σ = 0.6 and μ = 24.3 again. What is P(Y < 23 or Y > 25)?

Use complement: P(Y < 23 or Y > 25) = 1 - P(23 < Y < 25) = 1 - 0.8632 = 0.1368

1. What MPG is at least required for a car to be in the top 5% of all these same types of cars? Suppose σ = 0.6 and μ = 24.3 again.

This problem requires going in the opposite direction. We are now given a probability and need to find the corresponding value of q in the expression

P(Y > q) = 0.05.

In terms of integration, we are trying to find y in the equation below:



Equivalently,



We can simply find this in R using the qnorm() function.

> qnorm(p = 0.95, mean = 24.3, sd = 0.6)

[1] 25.28691



Thus, 25.29 is the 0.95 quantile from a normal distribution with μ = 24.3 and σ = 0.6.

Example: Grading (grade\_bell.R)

Suppose the set of test #1 grades in a class has a normal distribution with μ = 73% and σ = 8%. Let Y be a student’s grade. Answer the following.

1. What is the probability that a randomly chosen student in the class received a grade of 90% or better?

We need to find P(Y > 90).

> curve(expr = dnorm(x = x, mean = 73, sd = 8), n

 = 1000, from = 50, to = 100, col =

 "darkgreen", lwd = 2, ylab = "f(y)", xlab =

 "y")

> abline(h = 0)

> segments(x0 = 90, y0 = 0, x1 = 90, y1 = dnorm(x

 = 90, mean = 73, sd = 8), col = "red", lwd =

 5)

> 1 - pnorm(q = 90, mean = 73, sd = 8)

[1] 0.01679331



1. What percentage of students scored between a 70% and 90%?

We need to find P(70 < Y < 90) = P(Y < 90) – P(Y < 70)

> pnorm(q = 90, mean = 73, sd = 8) - pnorm(q = 70,

 mean = 73, sd = 8)

[1] 0.6293765



1. Suppose that the instructor curves the test #1 grades and that ONLY the top 10% of test scores receive A’s. Would a student be better off with a test #1 grade of 81% (still with μ = 73% and σ = 8%) or a grade of 68% on a different test #1 that has a normal distribution with μ = 62% and σ = 3%?

Find the 0.90 quantile for the two distributions.



> qnorm(p = 0.90, mean = 73, sd = 8)

[1] 83.25241

> qnorm(p = 0.90, mean = 62, sd = 3)

[1] 65.84465

A student would prefer the second test because an A would be received.

Rule of thumb for the number of standard deviations all data lies from its mean:

We discussed previously that approximately all observations should lie within 2 to 3 standard deviations from their mean. Given μ and σ for a normal distribution here, how could we make use of this?

Find P(μ - 2σ < Y < μ + 2σ) = 0.9545 and P(μ - 3σ < Y < μ + 3σ) = 0.9973. The end result is what corresponds to the empirical rule.

Question: Suppose Y has a normal distribution. What is P(Y = μ)?

Standard normal probability distribution

Before there were readily assessable software package or calculators with functions for the normal distribution, people used tables based on the standard normal distribution to find probabilities associated with ANY normal distribution. Because tables are still used at times, it is helpful to briefly go through their use here. Also, this will help you understand better how to find probabilities with the distribution.

Below is an excerpt from a table that I constructed which is meant to duplicate what is often found in books.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **z** | **0.00** | **0.01** | **0.02** | **0.03** | **0.04** | **0.05** | **0.06** | **0.07** | **0.08** | **0.09** |
| **-3.4** | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0002 |
| **-3.3** | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0003 |
| **-3.2** | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0005 |
| **-3.1** | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0007 | 0.0007 |
| **-3.0** | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| **-2.9** | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| **-2.8** | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| **-2.7** | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| **-2.6** | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| **-2.5** | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| **-2.4** | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| **-2.3** | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| **-2.2** | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| **-2.1** | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| **-2.0** | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| **-1.9** | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |

The table gives:

P(Z < -3.41) = 0.0003,

P(Z < -3.03) = 0.0012,

P(Z < -2.57) = 0.0051,

where Z is a standard normal random variable. These probabilities can be found using the pnorm() function using mean = 0 and sd = 1 as arguments.

Why are we concerned with this table of standard normal probabilities?

A simple transformation can be made from ANY normal distribution to the standard normal distribution using the following formula:



where Y has a normal distribution with mean μ and standard deviation σ and Z is a standard normal random variable with mean 0 and standard deviation 1.

Therefore, using this one table, we can find all normal distribution probabilities without R or other software.

Example: Car miles per gallon (car\_mpg.R)

1. Find the probability that a randomly selected car of this type gets less than 23 MPG for one tank of gas.

Using the tables, P(Y < 23)

= 

= P(Z < -2.1667)

≈ P(Z < -2.17)

= 0.0150.

Observing a sample from a population characterized by a normal PDF

Suppose a population can be characterized by a normal PDF. What characteristics would you expect for a sample taken from that population?

Example: Car miles per gallon (car\_mpg.R)

A random sample of size n = 1000 is taken from a population characterized by the normal distribution with μ = 24.3 and σ = 0.6.

> set.seed(3290)

> y <- rnorm(n = 1000, mean = 24.3, sd = 0.6)

> head(y)

[1] 25.34755 24.76892 23.66707 23.82414 24.96230

 24.07366

> tail(y)

[1] 22.67635 25.03712 23.63818 24.52084 24.23200

 24.50111

> # Estimates of E(Y) and sqrt(Var(Y))

> mean(y)

[1] 24.29725

> sd(y)

[1] 0.5918041

> # Estimate of P(Y < 23)

> mean(y < 23)

[1] 0.015

> # PDF

> hist(y, main = "", freq = FALSE)

> curve(expr = dnorm(x = x, mean = 24.3, sd = 0.6),

 col = "red", add = TRUE)



> # CDF

> plot.ecdf(x = y, lwd = 2, panel.first = grid(),

 ylab = "Probability", xlab = "x", col = "blue",

 main = "")

> curve(expr = pnorm(q = x, mean = 24.3, sd = 0.6),

 col = "red", add = TRUE, n = 1000)

> legend(x = 22, y = 0.8, legend = c("ECDF", "CDF"),

 lty = 1, col = c("blue", "red"), lwd = 2, bty =

 "n")



Notice how similar the calculations based on the sample are to those based on the probability distribution!

Questions:

* What would you expect to occur with the sample mean and sample standard deviation as the sample size increases? What about if the sample size decreases?
* Answer the same question in the previous bullet relative to the histogram/PDF and ECDF/CDF.

Validity of the normal distribution assumption

All of the probabilities found using the normal PDF ASSUME it is the correct probability distribution for the random variable. What if this assumption is incorrect? The probabilities found using this assumption are WRONG!

Example: Car miles per gallon (car\_mpg.R)

Suppose Y really has a distribution as shown in pink. Notice there is still an area of 1 underneath of its “curve” (use area of rectangle formula: base×height = 4×0.25 = 1).

For this setting, the alternative distribution gives P(Y < 23) = 0.7×0.25 = 0.175. With the normal assumption of μ = 24.3 and σ = 0.6, the probability was found to be 0.0151.



How does one know when the normal probability distribution assumption is valid?

Rarely, if ever, will it be 100% correct.

If a sample from the population is possible, construct a histogram of the observed values and check to see if it has the shape of a normal distribution. In addition, calculate the sample mean and variance to see if they are close to the population mean and variance (if they are known). If the histogram does have a similar shape to a normal distribution and the sample and population mean and variance are about the same (if the population values are known), then the normal distribution assumption is a reasonable approximation.

Suppose a histogram was constructed and the sample did not appear to come from a population characterized by a normal distribution. What can you do?

You could use another probability distribution. Alternatively, you can still use the normal distribution with the sample mean provided the sample size is large enough. The central limit theorem is used here in order to make a normal distribution approximation. This topic will be discussed soon!