

EXPECTATION AND VARIANCE, BINOMIAL DISTRIBUTION, CONTINUOUS DISTRIBUTIONS

Contents

1.	Expectation and variance	1
1.1.	Sample averages and expectation	1
1.2.	Conditional expectation	2
1.3.	Some additional properties of expectation	2
2.	Binomial Distribution	4
3.	Law of large numbers and “soft” CLT	5
4.	Continuous random variables [1, Ch. 3]	6
4.1.	Cumulative density function	7
4.2.	Conditional probability and expectation with Continuous variables	8
4.3.	The Gaussian curve	8
	References	8

1. Expectation and variance

1.1. **Sample averages and expectation** [1, Ch. 3].

- Usually it is difficult to obtain the full distribution of a random variable from data and it may not even be that relevant for the questions we are asking. Instead, we would like to summarize properties of a random variable from [statistics](#). The simplest example is the [sample mean](#) or [sample average](#). If we have many samples Y_1, Y_2, \dots, Y_n of a random variable Y (e.g. answers to a survey question), the sample mean is defined as

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

More generally, we might look at the average of some function of a random variable

$$\overline{f(Y)} = \frac{1}{n} \sum_{i=1}^n f(Y_i)$$

- These sample means can be computed from the probability distribution. Suppose each Y_i can take on outcomes $Y = 1, 2, 3, \dots, m$. If n is large, then the fraction of samples for which $Y_1 = y$ will be $P(Y_1 = y)$, thus the sample mean converges to the true mean:

$$\bar{Y} = \frac{1}{n} \sum_i Y_i = \frac{1}{n} \sum_{y=1}^m y n_y = \sum_{y=1}^m y \frac{n_y}{n} \approx \sum_{y=1}^m y P(Y = y)$$

- Sometimes we write $P(Y_i = y_i)$ and sometimes we write The expression on the right is the definition of the mean, or [expectation](#) denoted $E[Y]$.
- In many cases we would like to measure deviations from the mean. For this we look at the [variance](#),

$$\text{var}(Y) = E[(Y - E[Y])^2]$$

Another way to write this is

$$\text{var}(Y) = E[Y^2] - 2E[Y]^2 + E[Y]^2 = E[Y^2] - E[Y]^2$$

We can of course estimate this from a large number of samples, although as we will later see the obvious way of doing this by replacing the expectations with sample averages is not optimal. The square root of the variance is called the [standard deviation](#).

Example 1 (Mean and variance of Bernoulli random variable). *Let Y be a Bernoulli random variable with parameter q . We will use the convention that $Y = 1$ with probability q .*

Question: What is $E[Y]$ and $\text{var}(Y)$?

Solution: Using the definitions above

$$E[Y] = P(Y = 0) \times 0 + P(Y = 1) \times 1 = q$$

similarly you should be able to see that $\text{var}(Y) = q(1 - q)$. In the python notebook we test the formula $E[Y] = q$.

1.2. Conditional expectation.

- We define the conditional expectation as the expectation of the conditional variable; that is,

$$\mathbb{E}[X|Y = y] = \sum_x xP(X|Y = y)$$

With samples

$$\{(x_1, y_1), \dots, (x_n, y_n)\}$$

we could obtain this from a sample by taking the sample mean of X among just the samples where $y_i = y$; that is, let

$$n_1 = \text{number of samples where } y_i = y$$

then

$$\mathbb{E}[X|Y = y] \approx \frac{1}{n_1} \sum_{i=1}^n 1_{\{y_i=y\}} x_i$$

where $1_{\{y_i=y\}}$ is the [indicator function](#)

$$1_{\{y_i=y\}} = \begin{cases} 0 & \text{if } y_i \neq y \\ 1 & \text{if } y_i = y \end{cases}$$

- As we already noted, the conditional probabilities can tell us whether two variables are independent. That is, $P(X|Y) = P(X)$ if and only if X and Y are independent. If X and Y are independent, then $\mathbb{E}[X|Y = y] = \mathbb{E}[X]$ for all y but the converse is false: **it is possible that this is true but X and Y are not independent!** We will say about this later.

Example 2 (Computing conditional from probability model). Consider the pair of random variables (Y_A, Y_B) defined by the probability distribution we saw in week 1:

$$(1) \quad P(Y_A, Y_B) = \begin{cases} 1/2 & \text{if } Y_A = 0 \text{ and } Y_B = 0 \\ 1/8 & \text{if } Y_A = 0 \text{ and } Y_B = 1 \\ 1/8 & \text{if } Y_A = 1 \text{ and } Y_B = 0 \\ 1/4 & \text{if } Y_A = 1 \text{ and } Y_B = 1 \end{cases}$$

Question: Compute $\mathbb{E}[Y_A|Y_B = 1]$

Solution: We can obtain the conditional distribution of Y_A as

$$P(Y_A = 1|Y_B = 1) = \frac{P(Y_A = 1, Y_B = 1)}{P(Y_B = 1)} = \frac{1/4}{3/8} = \frac{2}{3}$$

Note that this means $P(Y_A = 0|Y_B = 1) = 1/3$ and so the conditional distribution of Y_A is

$$Y_A|Y_B = 1 \sim \text{Bernoulli}(2/3)$$

which means

$$\mathbb{E}[Y_A|Y_B = 1] = \frac{2}{3}.$$

Example 3 (Computing conditional expectation from data). Consider the following data containing children's test scores and some other information. Let Y be the test score and X be a binary variable representing whether the mother graduated high school.

Question: Compute $\mathbb{E}[Y|X = 0]$ and $\mathbb{E}[Y]$. Based on this, do you think it's likely that X and Y are independent?

Solution: See Python notebook.

1.3. Some additional properties of expectation.

- Previously, I introduced the notation of expectation and we saw how to compute expectations in Python. Now we cover some additional properties.

(1) **Linearity:** For two random variables X and Y with discrete sample spaces S_X and S_Y ,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Proof. We have

$$\begin{aligned} \mathbb{E}[X + Y] &= \sum_{y \in S_Y} \sum_{x \in S_X} (x + y)P(X = x, Y = y) \\ &= \sum_{x \in S_X} \sum_{y \in S_Y} xP(X = x, Y = y) + \sum_{x \in S_X} \sum_{y \in S_Y} yP(X = x, Y = y) \\ &= \sum_{x \in S_X} x \left(\sum_{y \in S_Y} P(X = x, Y = y) \right) + \sum_{y \in S_Y} y \left(\sum_{x \in S_X} P(X = x, Y = y) \right) \\ &= \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

□

- (2) **Multiplication by a constant:** If a is a constant (meaning it is not random), then

$$E[aX] = aE[X]$$

Proof. Left as an exercise.

□

- (3) **Factoring for independent variables:** If X and Y are independent, the

$$E[XY] = E[X]E[Y]$$

Proof. Using independence, we have

$$\begin{aligned} E[XY] &= \sum_{x \in S_X} \sum_{y \in S_Y} xy P(X = x, Y = y) \\ &= \sum_{x \in S_X} \sum_{y \in S_Y} x P(X = x) y P(Y = y) \\ &= \left(\sum_{x \in S_X} x P(X = x) \right) \left(\sum_{y \in S_Y} y P(Y = y) \right) = E[X]E[Y] \end{aligned}$$

□

- (4) **Tower property:** Let X and Y be two random variables,

$$E[E[X|Y]] = E[X]$$

where by $E[X|Y]$ we mean the random variable constructed by taking the conditional expectation of X given a random value of Y . Another way to define this is to introduce the deterministic function $f(y) = E[X|Y = y]$ which outputs a number for every value $y \in Y$. Then we define the random variable $E[X|Y] = u(Y)$. Therefore

$$E[E[X|Y]] = E[f(Y)]$$

Proof. Left as an exercise.

□

Example 4 (Calculating probabilities). Consider the model probability model for a variable X (similar to Equation 4)

$$P(X) = \begin{cases} 1/2 & \text{if } X = 1 \\ 1/8 & \text{if } X = 2 \\ 1/8 & \text{if } X = 3 \\ 1/4 & \text{if } X = 4 \end{cases}$$

and define

$$Y = X \cdot Z, \quad Z = \text{Geometric}(1/2).$$

In this example we consider the joint distribution (Y, X) .

Question: Using simulations of this model, illustrate properties of expectation in Python.

Solution: See Python notebook. The one that requires the most explanation is the tower property. We will use simulations to show. Let $(X_1, Y_1), \dots, (X_N, Y_N)$ be our samples. We will show that $E[X] = E[E[X|Y]]$. The left hand side is easy to approximate with samples – just take the mean of the X values. To compute $E[X|Y]$ we need to compute $E[X|Y = y]$ for each y . If n_y is the number of samples such that $Y_i = y$ then

$$E[X|Y = y] \approx \frac{1}{n_y} \sum_{\{i: Y_i = y\}} X_i$$

The sum is over all the Y_i for which $Y_i = y$. We could also write this sum as $\sum_{i=1}^N X_i 1_{Y_i=y}$. This means

$$E[E[X|Y]] \approx E \left[\frac{1}{n_y} \sum_{\{i: Y_i = y\}} X_i \right] = \sum_{y \in S_Y} P(Y = y) \underbrace{\left(\frac{1}{n_y} \sum_{\{i: Y_i = y\}} X_i \right)}_{\approx E[X|Y=y]}$$

Now use that $P(Y = y) \approx n_y/N$, therefore

$$E[E[X|Y]] \approx \sum_{y \in S_Y} \frac{1}{n_y} \sum_{\{i: Y_i = y\}} X_i \frac{n_y}{N}$$

Notice that when we cancel the n_y terms we just get the usual sample mean of X :

$$E[E[X|Y]] \approx \frac{1}{N} \sum_{y \in S_Y} \sum_{\{i: Y_i = y\}} X_i = \frac{1}{N} \sum_{i=1}^N X_i = \bar{X} \approx E[X]$$

2. Binomial Distribution

- Suppose

$$Y_i \sim \text{Bernoulli}(q), \quad i = 1, \dots, N$$

are independent. We will use the convention that $Y_i = 1$ with probability q . Let

$$Y = \sum_{i=1}^N y_i$$

Then we say Y follows [binomial distribution](#) and write

$$Y \sim \text{Binomial}(N, q)$$

Example 5 (Calculating probabilities). Let $N = 3$ and $k = 2$.

Question: Calculate $P(Y = 2)$?

Solution: There are 3 possible sequences that give $Y = 2$.

$$(1, 0, 1), (1, 1, 0), (0, 1, 1)$$

The probability that we see any particular one of these is $(1 - q)q^2$. For example,

$$\begin{aligned} P(y_1 = 1, y_2 = 0, y_3 = 1) &= P(y_1 = 1)P(y_2 = 0)P(y_3 = 1) \\ &= q(1 - q)q = q^2(1 - q). \end{aligned}$$

Therefore the chance to observe $Y = 2$ with $N = 3$ is

$$P(Y = 2) = P((1, 0, 1)) + P((1, 1, 0)) + P((0, 1, 1)) = 3q^2(1 - q).$$

Note that the binomial distribution has two parameters, N and q , representing the number of flips and probability of success respectively.

- Now let's think about what the probability distribution will look like. The chance to find any **particular** configuration of k ones is

$$q^k(1 - q)^{N-k}$$

because they are independent.

However, we need to account for the fact that there are many configurations with k ones. In general, there are

$$\binom{N}{k} = \frac{N!}{k!(N - k)!} = \frac{N \times (N - 1) \times (N - 2) \times \dots \times (N - k + 1)}{k \times (k - 1) \times (k - 2) \times \dots \times 1}$$

way to have k ones among N samples.¹

This implies

$$P(Y = k) = \binom{N}{k} q^k (1 - q)^{N-k}.$$

A graph of this function looks like a bell curve when N is large – we will see this in Python. Let's explore binomial random variables in python.

Example 6 (Expectation of binomial). Let Y be a binomial random variable.

Question: What are $E[Y]$ and $\text{var}(Y)$?

Solution:

$$E[Y] = \sum_{k=1}^N k P(Y = k) = \sum_{k=1}^N k \binom{N}{k} q^k (1 - q)^{N-k} = \dots$$

A much easier way is to use the definition of a Binomial random variable and expectations

$$\begin{aligned} E[Y] &= E \left[\sum_{j=1}^N X_j \right] \\ &\stackrel{(1)}{=} \sum_{j=1}^N E[X_j] = Nq \end{aligned}$$

¹To better understand the formula above, let $C_{N,k}$ denote the number of sequences with k ones. We can break $C_{N,k}$ up into the number of terms for which 1 appears as the first element of the sequence and those for which zero is the first. If one comes first, we have $N - 1$ remaining slots to place $k - 1$ ones, thus there are $C_{N-1,k-1}$ of these sequences. Similarly, if zero comes first, we have $N - 1$ slots but now all k ones to place, thus there are $C_{N-1,k}$ of these. It follows that

$$C_{N,k} = C_{N-1,k-1} + C_{N-1,k}.$$

Notice that the quantity $C_{N,k}$ will be smallest when $k = 1$ or $k = N$, since in these cases there is only one way to configure the sequence: $C_{N,1} = C_{N,N} = 1$. You can solve the recursion to obtain the formula.

where we are using the fact that averages are additive (property (1)). Similarly,

$$\begin{aligned} E[Y^2] &= E \left[\left(\sum_{i=1}^N X_i \right)^2 \right] = E \left[\sum_{i=1}^N \sum_{j=1}^N X_i X_j \right] \\ &\stackrel{(1)}{=} \sum_{i=1}^N \sum_{j=1}^N E[X_i X_j] \stackrel{(3)}{=} \sum_{i=1}^N \sum_{j \neq i}^N q^2 + Nq(1-q) + Nq^2 \\ &= N(N-1)q^2 + Nq(1-q) + Nq^2 = Nq(1-q) + N^2q^2 \end{aligned}$$

Therefore

$$\text{var}(Y) = E[Y^2] - E[Y]^2 = Nq(1-q)$$

To summarize what we learned in Example 6

$$(2) \quad E[Y] = qN \quad \text{var}(Y) = Nq(1-q).$$

- The important observation that the mean grows much faster with N than the variance is also captured by the coefficient of variation:

$$\text{CV} = \frac{\sqrt{\text{var}(Y)}}{E[Y]} = \sqrt{\frac{(1-q)}{q} \frac{1}{N}}.$$

The idea is that we are measuring the variation *relative* to the average. This is relevant for many applications where we only care about the relative deviations.

Example 7 (Election modeling). Consider a model of votes in an election involving two candidate. Let q be the fraction of people in the population who support candidate one and suppose N people vote at the election (you can assume N is much less than the total number of people in the population, as voter turnout is low). Then the number of people, M , who vote for the first candidate can be modeled as

$$M \sim \text{Binomial}(N, q)$$

Think about the assumption we are making when we use this model.

Question: Suppose there is a city in which a fraction $q = 0.51$ of people support a candidate for city council. If $N = 1000$ people turnout for the election, what is the chance that the actual vote share, $\phi = M/N$, differs from the actual fraction of support throughout the population by more than 1%?

Solution: See the Python notebook.

- In the problem above the vote share is the same as the sample mean:

$$\bar{X}_i = \frac{M}{N} = \frac{1}{N} \sum_{i=1}^N X_i$$

You should be able to see that $E[\phi] = q$. What about the variance?

$$\text{var}(\phi) = \text{var}(Y/N) = \frac{1}{N^2} \text{var}(Y) = \frac{q(1-q)}{N}$$

Notice that this will tend towards zero as $N \rightarrow \infty$. Meanwhile, $E[\phi]$ has no dependence on N . This is a consequence of the fact that the CV is decreasing with N and it allows us to determine q by approximating $E[\phi]$ with the sample mean.

3. Law of large numbers and “soft” CLT

- The binomial distribution illustrates a very basically principle that we have already used a number of times: When we sum over a large number of independent random variables and divide by the total number, the result is close to the mean. This is the Law of Large Numbers (LLN).

Theorem 1 (Law of Large numbers). Let X_i be independent and identically distributed and set

$$S_N = \sum_{i=1}^N X_i.$$

If $E[X_i] < \infty$, then $S_N/N \rightarrow E[X_i]$.

This is not very precise, since we should really be specific about what it means for a random number to converge to something, but for our purposes it will suffice to think of this as saying that for large enough N , S_N/N will not differ from $E[X_i]$ very much. Another way to say this is that for iid random variables X_i , $i = 1, \dots, N$, the sample average \bar{X} approach $E[X_i]$.

- The binomial distribution actually tell us more, it tell us that the variation around $E[X_i]$ is proportional to $1/\sqrt{N}$. It is natural to ask whether this is also true for other random variables. The key is that the dependence on N in Equations 2 does not depend on the distribution of X_i ! So if X_i is the roll of a dice, or a geometric distribution, we expect the same thing to hold.

- The behavior of random sums is in-fact even more universal than this argument suggests. We can actually describe the distribution of any² random sum with a single distribution. In order to describe this distribution, we need to introduce the notion of continuous random variables.

4. Continuous random variables

- Here we will talk about random variables which take on continuous values. We will find that it is often easier to work with such distribution and for certain variables they are more natural models. But most importantly, the Normal model which is the foundation of so much statistical theory is a continuous variable. It emerges as an approximation of both continuous and discrete models.

- A uniform random variable, denoted

$$Y \sim \text{Uniform}(a, b)$$

has an equal chance of taking any number in the interval $[a, b]$ (we assume $a < b$). Let $L = b - a$. This is distinct from other distributions we have encountered in that it is a continuous distribution, rather than discrete, so we need to think slightly differently about how to define the probability distribution.

- For the uniform distribution,

$$P(y_1 \leq Y \leq y_2) = \frac{y_2 - y_1}{L}$$

for $a < y_1 < y_2 < b$. That is, the chance for Y to fall in any interval is simply the length of that interval. This insures that that the probability of Y being somewhere in $[a, b]$ is one: $P(a \leq Y \leq b) = 1$. Note that as $y_2 \rightarrow y_1$, $P(y_1 \leq Y \leq y_2) \rightarrow 0$. This tells us that the chance for Y to take any specific value is 0. Indeed, there are simply too many number (uncountably many) in any interval to assign positive probability to each.

- For continuous variables, it is sometimes useful to work with the density, $f(y)$ (we will use lower case letters for density and uppercase for probability distributions). The density $f(y)$ of a continuous random variable is a positive function such that

$$P(a < X < b) = \text{area under } f(y) \text{ between } a \text{ and } b$$

- We can think of $f(y)$ as the the probability per unit Y , meaning that if we look in a small interval

$$f(y)dy = P(y \leq Y \leq y + dy) = \frac{dy}{L}.$$

Thus, for uniform distribution the density is $1/L$ if $y \in [a, b]$ and 0 otherwise. Note the following properties of $f(y)$

(1) It is positive

(2) The area under $f(y)$ is one (this is the continuous version of the condition that discrete probability distributions must sum to one).

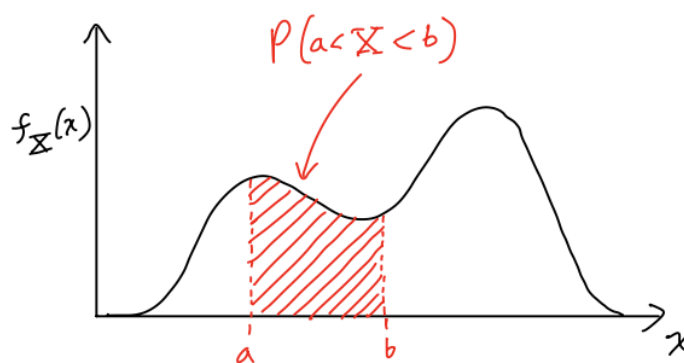


Figure 1. The density and its relationship to probabilities

Example 8 (Condition with continuous random variables). If Y is uniform on $[0, 1]$.

Question: What is the density of $Y|(Y < 1/2)$? Check the answer with simulations.

Solution: We can start with the definition of density

$$P(y_1 < Y < y_2 | Y < 1/2) = \frac{P(y_1 < Y < y_2, Y < 1/2)}{P(Y < 1/2)}$$

²With the caveat that here we only deal with the case where $\text{var}(X_i) < \infty$

What is the think on the top? We will assume $y_1 > 0$ and $y_2 < 1/2$, then the numerator is $y_2 - y_1$, since $Y < 1/2$. The key here is that if $Y \in [y_1, y_2]$ $Y < 1/2$ is automatically true, so the chance that BOTH of these things are true in a sample is the chance that the more restrictive one is true.

The denominator is $P(Y < 1/2) = 1/2$. This means

$$P(y_1 < Y < y_2 | Y < 1/2) = 2(y_2 - y_1)$$

This means the density is

$$f(y|Y < 1/2) = 2$$

Thus

$$Y|(Y < 1/2) \sim \text{Uniform}(0, 1/2)$$

We also check this with simulations in the Python notebook.

- Suppose we want to model that time before a component of a machine fails. We will assume that the rate of failure – that is, the chance that it fails per unit time – is a constant λ . In other words, for a small time interval dt , the probability for the component to fail in a small time interval $[t, t + dt)$ given that it has not yet failed is λdt . Or, in mathematical notation If T is the time of failure, then the density of $f_T(t)$ is

$$f_T(t) = \lambda e^{-\lambda t}.$$

T is an exponentially distributed random variable, and we write

$$T \sim \text{Exponential}(\lambda).$$

An exponential variable has mean $E[T] = 1/\lambda$ and variance $\text{var}(T) = 1/\lambda^2$.

Example 9 (Heterogeneous failure rate). Suppose that the machine is defective with probability 0.1. We can introduce a variable X which indicates whether the machine is defective and will fail with a rate 10. In other words, our model is

$$X \sim \text{Bernoulli}(0.1)$$

$$T|(X = x) \sim \text{Exponential}(x10 + (1 - x))$$

Question: What is $E[T]$? What about $\text{var}(T)$? Does T follow an exponential distribution?

Solution: Using the tower property of expectation

$$\begin{aligned} E[T] &= E[E[T|X]] \\ &= E[T|X = 0]P(X = 0) + E[T|X = 1]P(X = 1) \\ &= 1 \cdot (1 - 0.1) + \frac{0.1}{10} = 0.9 + 0.01 = 0.91 \end{aligned}$$

The variance

$$\text{var}(T) = E[T^2] - E[T]^2$$

and

$$E[T^2] = E[E[T^2|X]] = (1 - 0.1) \times E[T^2|X = 0] + 0.1 \times E[T^2|X = 1]$$

Note that, from the variance formula and the fact that T is exponential,

$$E[T^2|X = 1] = \text{var}(T|X = 1) + E[T|X = 1]^2 = \frac{1}{10^2} + \frac{1}{10^2} = \frac{2}{10^2}$$

Hence

$$E[T^2] = E[E[T^2|X]] = 0.9 \cdot 2 + 0.1 \cdot \frac{2}{10^2} = 1.802$$

$$\text{var}(T) = 1.802 - 0.91^2 = 0.9739$$

If T is exponential, then

$$\lambda = \frac{1}{E[T]} = \frac{1}{0.91}.$$

We know that $\text{var}(T) = E[T]^2$ for an exponential distribution, but

$$\frac{1}{\lambda^2} = 0.91^2 = 0.8281 \neq 0.9739 = \text{var}(T).$$

4.1. Cumulative density function.

- Sometimes it is useful to characterize a continuous distribution not by the density, but by the cumulative distribution function (CDF) defined as

$$F(y) = P(Y < y).$$

4.2. Conditional probability and expectation with Continuous variables.

- The definition of expected value can be generalized to continuous variables by replacing the sums with integrals. That is, for a variable X with density f_X , we have

$$E[X] = \int x f_X(x) dx$$

- Suppose X and Y are two variables on the sample spaces $S_X = S_Y = \mathbb{R}$. Then we can define a joint density $f_{X,Y}(x, y)$. From this, we can compute things like

$$P(X > x, Y > y) = \int_x^\infty \int_y^\infty f_{X,Y}(x, y) dx dy$$

If X and Y are independent, then $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ and

$$\begin{aligned} P(X > x, Y > y) &= \int_x^\infty \int_y^\infty f_{X,Y}(x, y) dx dy \\ &= \int_x^\infty f_X(x) dx \int_y^\infty f_Y(y) dy = P(X > x)P(Y > y) \end{aligned}$$

4.3. The Gaussian curve.

- One of the most important probability densities is the Gaussian:

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

- Despite the simplicity of the density function, calculating probabilities for Normal random variable by computing the area under the curve (integrating) is difficult. Instead we can remember some rough estimates based on the following figure.

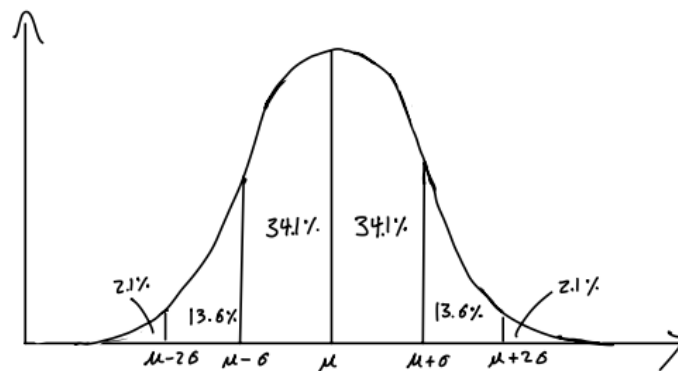


Figure 2. Probabilities in the Normal distribution

Example 10 (Calculating probabilities for Normal distribution). We use the curve above to calculate probabilities of events in the Normal distribution. Suppose

$$Y \sim \text{Normal}(5, 4)$$

Question: What is (approximately) $P(Y > 7)$?

Solution: Note that $5 + 2 = 7$, so this is asking how likely it is that a Normal variable is greater than 1 standard deviation above the mean. This about $13.5 + 2 = 15.5\%$. We can always easily compute these probabilities in python as well.

Question: What is

$$P(Y > 3 | Y < 7)?$$

Solution: In this case we would use

$$P(Y > 3 | Y < 7) = \frac{P(Y > 3, Y < 7)}{P(Y < 7)}$$

Notice that $3 = 5 - 2 = \mu - \sigma$ and we already saw $7 = 5 + 2 = \mu + \sigma$, so $P(Y < 7) \approx 0.839$ and $P(Y > 3, Y < 7) \approx 0.682$, thus the result is about 0.81.

References

- [1] John Tabak. *Probability and statistics: The science of uncertainty*. Infobase Publishing, 2014.