6 Continuous distributions (P.59-62)

So far we've talked about probabilities for discrete random variables, like number of times a coin comes up heads. In practice, data that represent measurable quantities are rounded to a specified number of decimal places, due to the limitations of measuring scales. For example, the weight of a person may be recorded as 60Kg or 60.5Kg. However, this weight may be 60.54321753..... In other words, the actual weight cannot be restricted to integer value. Such data are known as *continuous data*. In this case, the difference between any two possible data values can be arbitrarily small. This chapter considers the probability distributions and applications of such continuous data.

Examples: The following variables can be considered as continuous variables:

1. Time
2. Temperature
3. Height of young men
4. The concentration of a pollutant

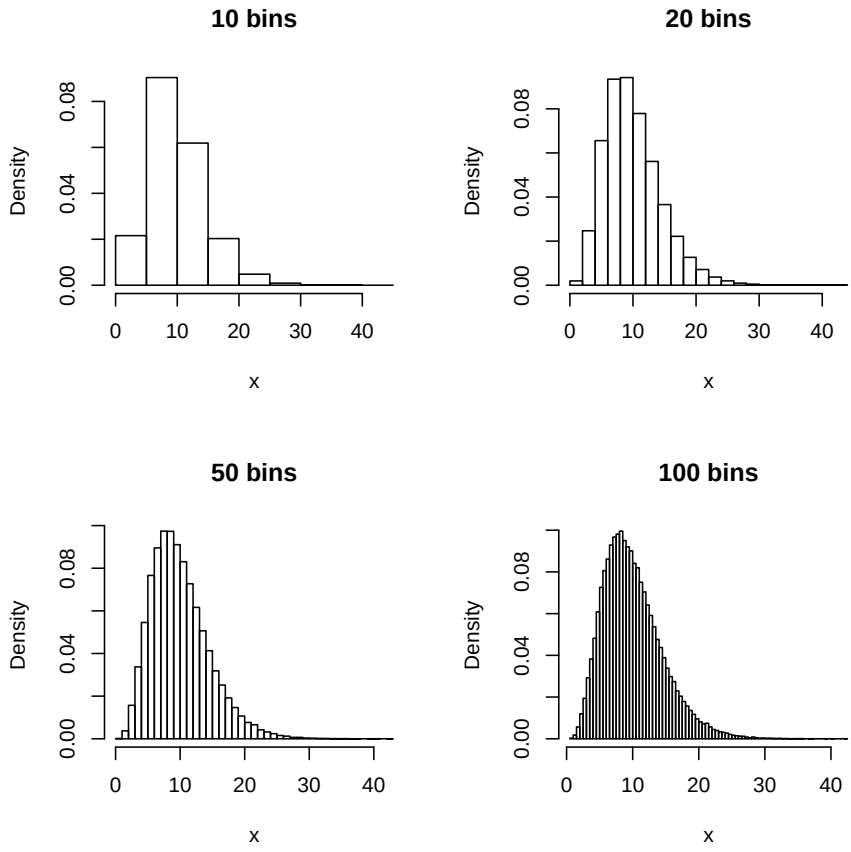
Note: These measurements can take fractional values, decimal numbers etc.

Example: Suppose that a biologist wants to investigate the distribution of lengths of certain insects for a research project. It is clear that the length, X, is a continuous random variable (rv). Why?

Answer: X can take any real value within a certain interval.

6.1 Probability density curves

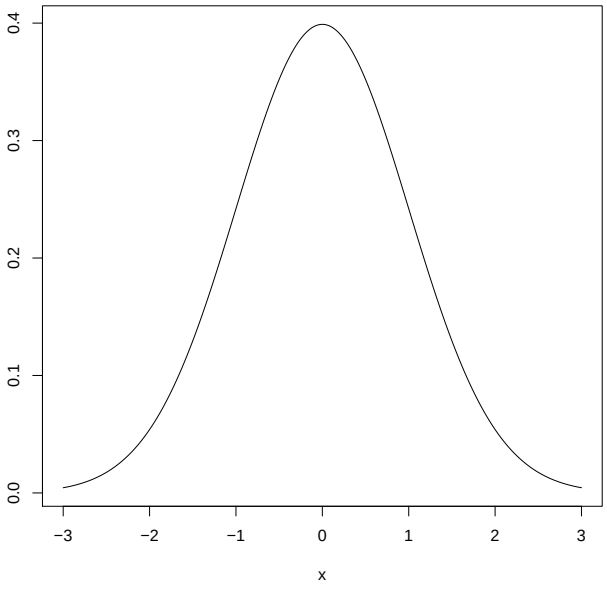
Suppose that we have a large set of continuous data measured on a particular characteristic. For example, the weights of students in a large class. That is, these measurements have been taken on a continuous random variable X. In this case, we can divide the data in to large number of intervals or classes or bins and draw histograms using relative frequencies as their heights as follows:



Recall: The frequency associated with each interval (bin) is proportional to the bar's area. Further, each rectangle represents the proportion of observations fall in each interval. Therefore, it is clear that the total area under this histogram is 100% or 1.

A smooth version of the histogram is called the *probability density curve*. The associated function or mathematical equation that describes the curve is called the *probability density function* (pdf). In this course, we do not consider any mathematical equation to explain the pdf. However, we need to know a number of important curves and their shapes to complete this course successfully.

Example: A typical probability density curve from a large set of continuous data

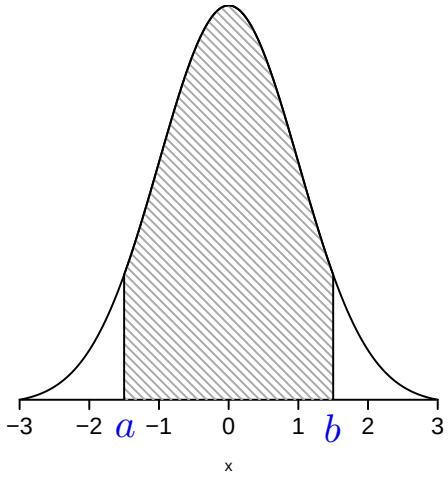


Note: The total area under this probability density curve is one unit. This corresponds to the total probability of all possible events is 1. This can be written as

$$\mathrm{P}(-\infty < X < \infty) = \text{Total area under the curve} = 1.$$

6.2 Area under the probability density curve between two points

It can be shown that the area under the density curve between any two points a and b is equal to the probability that the random variable X falls between a and b, that is, $\mathrm{P}(a \leq X \leq b)$.



Shaded area $= \mathrm{P}(a < X < b) = \mathrm{P}(a \leq X \leq b)$.

Note: Since X is a continuous random variable it can take any real value measured to any accuracy. In other words, the variable X cannot take a given fixed value. This indicates that the probability associated with any particular value of X is zero. For example, what is the probability of finding a student from this class whose weigh *exactly* 50.5 Kg? Clearly, no one can find a student with exactly 50.5Kg and therefore,

$$P(X = 50.5) = 0.$$

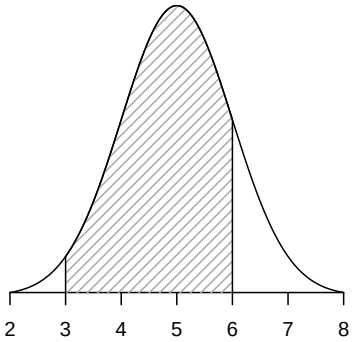
That is, the effect of the equal sign in this case is null unlike in the binomial distribution (or any other discrete distribution).

For example, if X is any continuous random variable, then $\mathrm{P}(1 < X < 3) = \mathrm{P}(1 \leq X \leq 3) = \mathrm{P}(1 \leq X < 3) = \mathrm{P}(1 < X \leq 3)$.

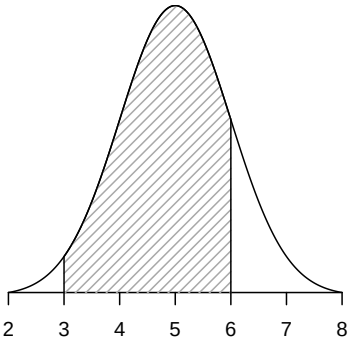
Examples: Suppose that X is a continuous rv with a pdf, symmetric about $x = 5$. Shade each area below to represent the probability:

1. $P(3 < X < 6)$
2. $P(3 \leq X \leq 6)$
3. $P(X > 7)$
4. $P(X \leq 4)$

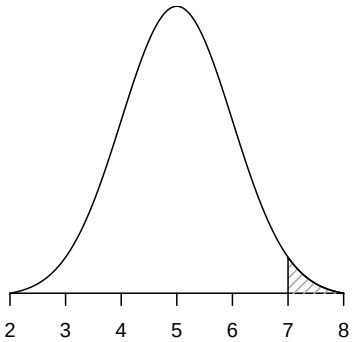
Solutions:



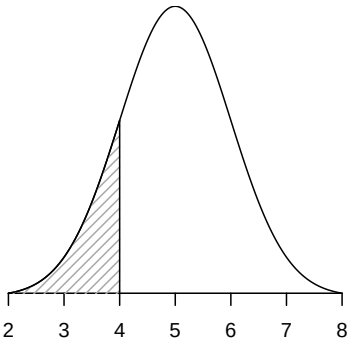
1.



2.



3.



4.

6.3 The Normal Distribution - P.62-66

Now we consider a very special continuous distribution, called the *normal distribution*. This is the most important, widely used continuous distribution in practice with many real-world applications.

The pdf given by the formula

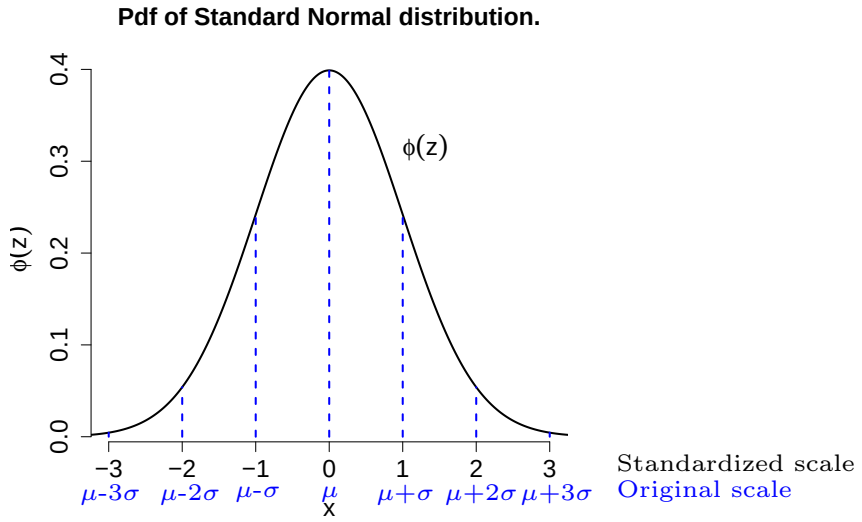
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

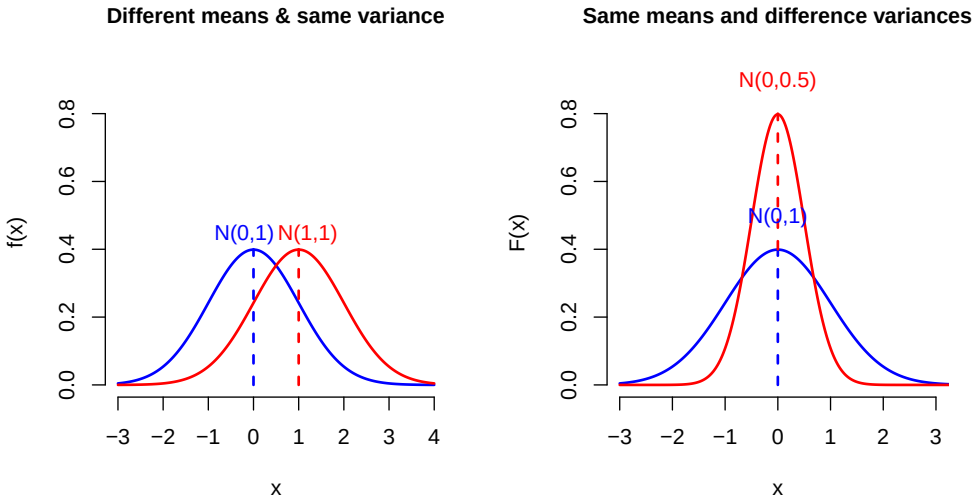
represents a symmetric, bell-shaped curve of a probability distribution that approximates very well commonly observed real-world large data sets (large samples).

It also plays a central role in most statistical theory. The curve is located at μ (or symmetric about μ) and σ indicates the spread or width of the curve. In this case we say that "X is normally distributed with mean μ and variance, σ^2" and is denoted by

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

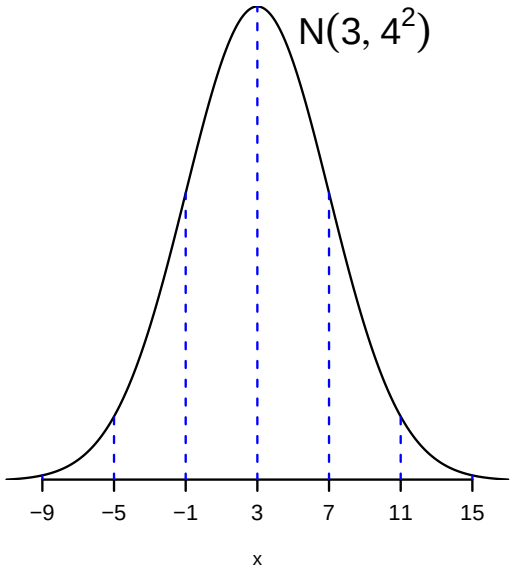
Diagram:





Example: $X \sim N(3, 4^2)$ means X is normally distributed with mean $\mu = 3$ and variance $\sigma^2 = 16$ ($\Rightarrow \sigma = 4$).

Diagram:



6.4 Standardized random variables

Let X be a random variable with mean μ and standard deviation σ. Then the random variable given by

$$Z = \frac{X - \mu}{\sigma}$$

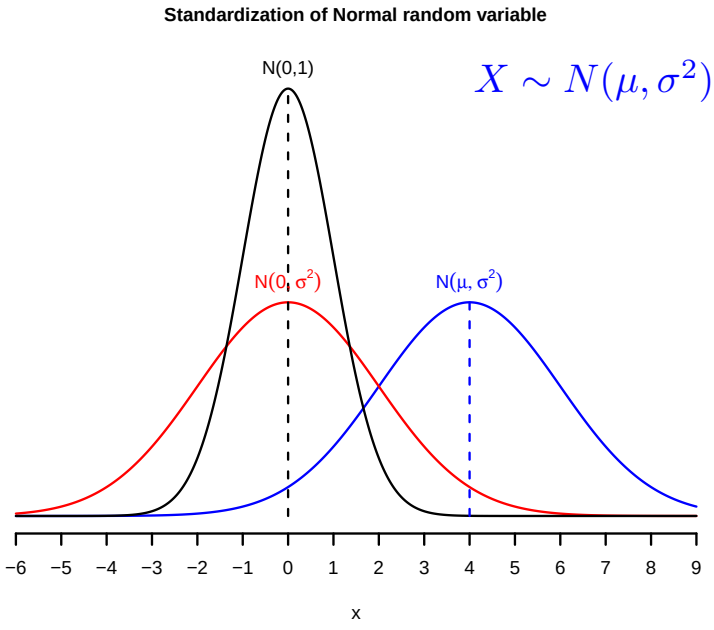
is called the standardized version of X.

Intuitively, it appeals that the standardized random variable Z will have zero mean and unit variance or standard deviation.

If the random variable X is normally distributed with mean μ and standard deviation σ (or variance σ^2), then the random variable Z

$$Z = \frac{X - \mu}{\sigma},$$

is called the standard (or standardized) normal random variable. As before $\mathrm{E}(Z) = 0$ and $\mathrm{SD}(Z) = 1$.



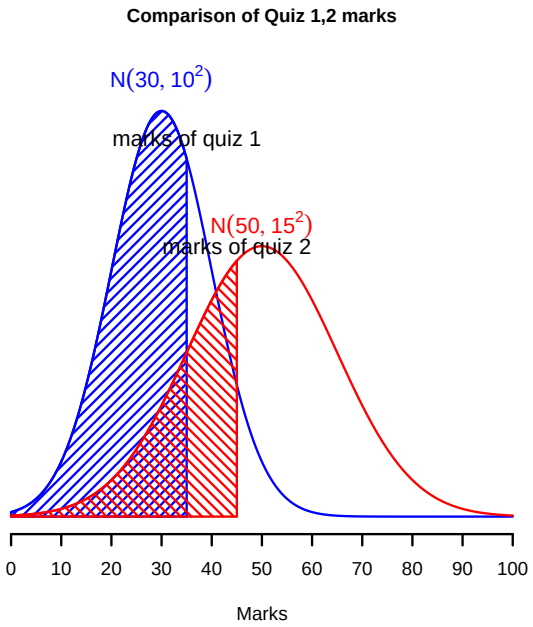
$$X \sim N(\mu, \sigma^2) \longrightarrow X - \mu \sim N(0, \sigma^2)$$
$$\longrightarrow \frac{X-\mu}{\sigma} \sim N(0, 1)$$

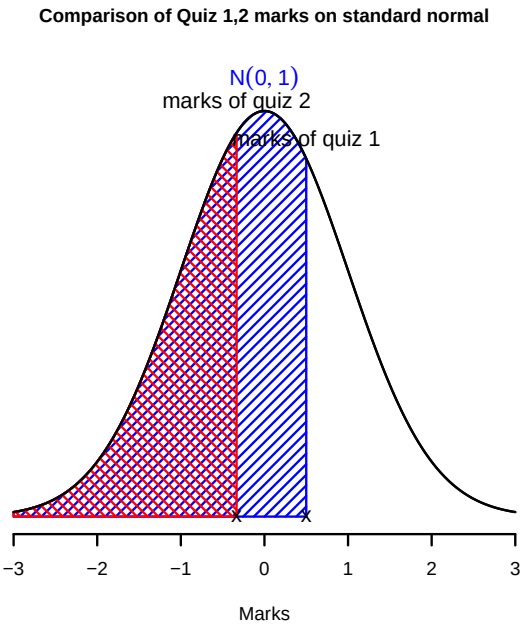
Question: Why do we need standardized variables?

Answer: After this transformation (or standardization), all the data will be reduced to a distribution *zero mean* and *unit standard deviation*. This helps us to compare the data on a standardized scale. In other words, this facilitates comparison of observations taken from different normal distributions.

Example: Suppose that the distribution of marks for Quiz 1 and 2 are $N(30, 10^2)$ and $N(50, 15^2)$ respectively. If a student got 35 and 45 marks respectively, does the student show improvement in the second quiz?

To compare the marks, one should standardize the marks to a distribution with zero mean and unit variance. The standardized marks from the standard normal $N(0, 1)$ distribution are 0.5 and -0.33 respectively and so the percentile of marks from Quiz 2 drops!





6.4.1 Standardized normal random variables

The normal random variable with mean zero and variance 1 (or SD = 1) or the standard normal variable is denoted by

$$Z \sim N(0, 1).$$

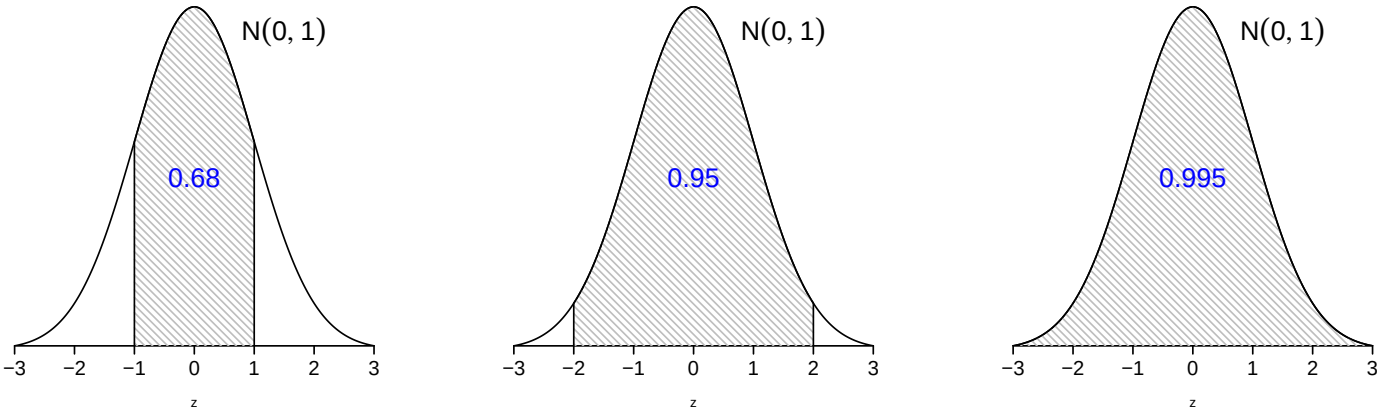
Suppose that $X \sim N(\mu, \sigma^2)$. Then it is clear that

$$Z = \frac{X - \mu}{\sigma}$$

is a standard normal variable and $Z \sim N(0, 1)$.

The shape of the standard normal distribution:

This is a bell shaped curve symmetric about zero. The shaded area (probability) covers approximately 95% of the total area (probability) as shown below:



Note: For our convenience, the area under the standard normal distribution (or probability) is tabulated in the normal table. See P.269 of the textbook, or the course website for a copy. Bring a copy of this table with you. A copy of this table will be given in all examinations.

Now we look at how to find the probabilities using normal table.

Note: It is always a good idea to shade the required area (or region) first and then look at normal table.

Examples: Suppose that $Z \sim N(0, 1)$.
1. Shade each area to represent the following probabilities and
2. find their values using the standard normal table.

(i) $P(Z > 0.00)$,

(ii) $P(Z > 1.31)$,

(iii) $P(0.00 < Z < 2.52)$,

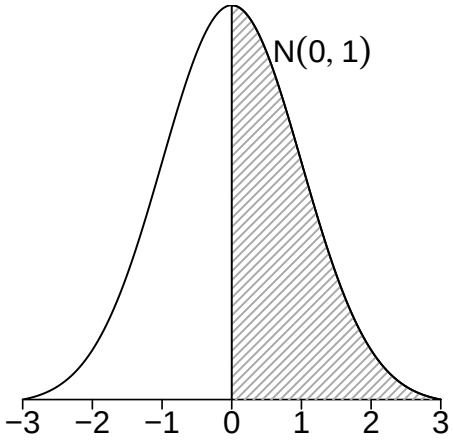
(iv) $P(-1.02 < Z < 0)$,

(v) $P(-0.71 < Z < 0.71)$,

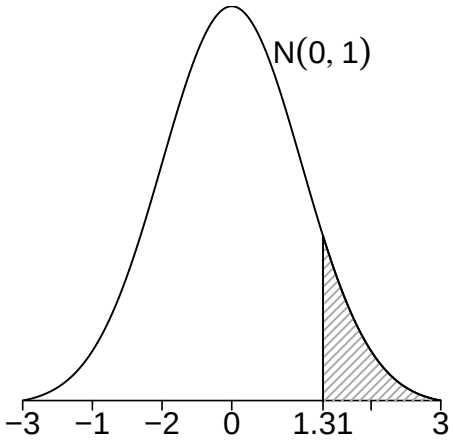
(vi) $P(-1.56 < Z < 1.33)$.

Solution:

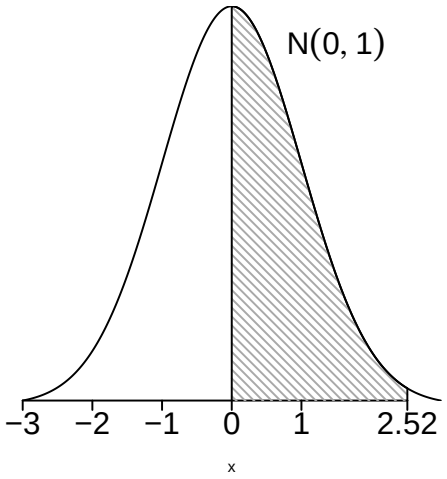
(i)



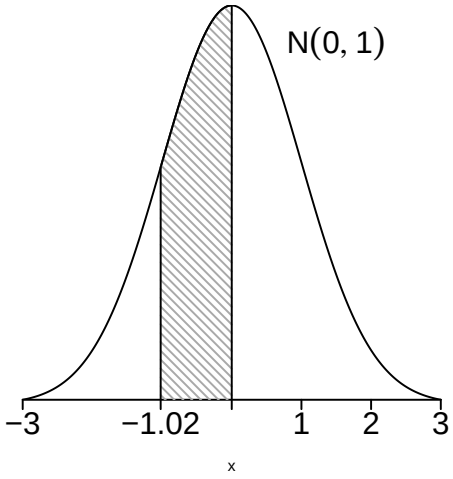
(ii)



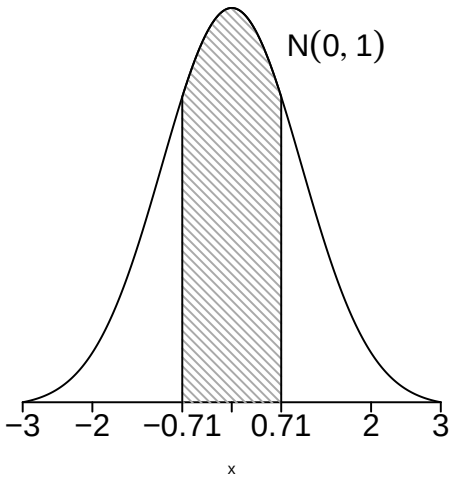
(iii)



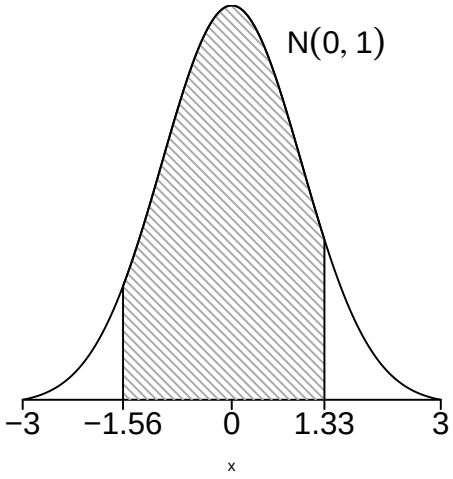
(iv)



(v)



(vi)



Key points to remember:

- Total area = 1;
- Curve is symmetric about $z = 0$.
- $P(Z \geq 0) = P(Z \leq 0) = 0.5$.
- Table 2 reports probabilities, $P(Z \leq z)$ for positive values of z only.
- $P(Z \geq z) = P(Z \leq -z)$ and $P(Z \geq z) = 1 - \mathrm{P}(Z \leq z)$.

Solution:

(i) $P(Z > 0.00)$= ___

(ii) $P(Z > 1.31) =$ _____

(iii) $P(0.00 < Z < 2.52) =$ _____

$=$ _____

(iv) Negative values of z cannnot be handled directly. We use the property of symmetry of the curve as follows:

$P(-1.02 < Z < 0) =$ _____

(v) In this case note that the values are equal with opposite signs.

$P(-0.71 < Z < 0.71) =$ _____

$=$ _____

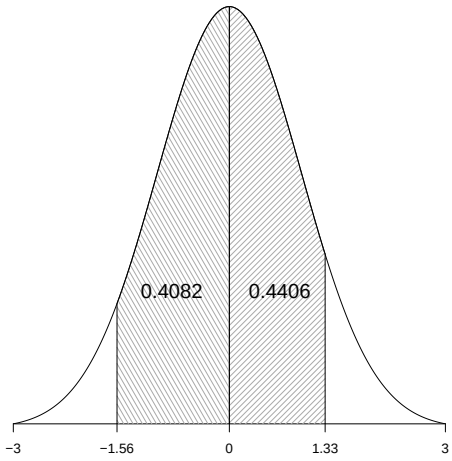
$=$ _____

Table 2: Standard Normal Distribution Table

Lower tail probabilities $P(Z < z)$ where Z follows a standard normal distribution.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
⋮										

(vi) In this case note that one value is negative and the other is positive.



$P(-1.56 < Z < 1.33) =$ _____

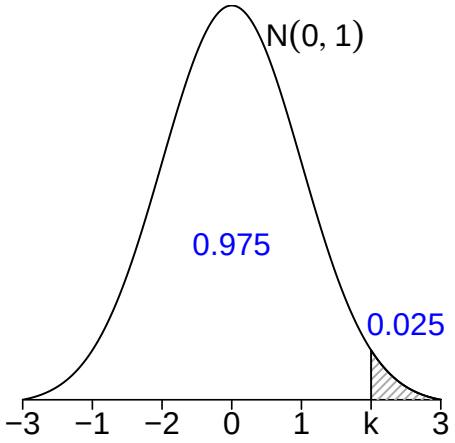
$=$ _____

Read: The examples in P.67-69.

Note: In some applications, we need to find the cut-off value from a normal distribution for a given probability. For example, suppose that a teacher wants to allocate high distinctions (HD) to the top 5% of students. What is the cut-off marks for such achievers?

Example: What is the top 2.5% cut-off value of the standard normal distribution? That is, find k such that $\mathrm{P}(Z \geq k) = 0.025$.

Solution:



Note that

$$\mathrm{P}(Z \leq k) = 1 - 0.025 = 0.975.$$

Now we need to look at the area (probability) 0.975 *inside* the z-table. We notice that 0.975 appears across 1.9 and below 0.06. Therefore,

$$k = \_\_\_\_$$

6.5 Problems related to non-standard normal distributions - P.70

It is clear that not all practical problems come from the standard normal distribution. For example, a board of examiners may claim that the distribution of marks for statistics B course in 2010 approximately follows a normal distribution with a mean of 65 and a sd of 12. A normal distribution of this type is called a non-standard normal distribution since it has mean not equal to 0, and standard deviation not equal to 1. Now we look at such problems in detail.

Suppose that the distribution of the variable X is normal with mean μ and variance σ^2. That is

$$X \sim N(\mu, \sigma^2).$$

In this case we first standardise or convert $N(\mu, \sigma^2)$ *distribution to a standard normal* $N(0, 1)$ *distribution*, using

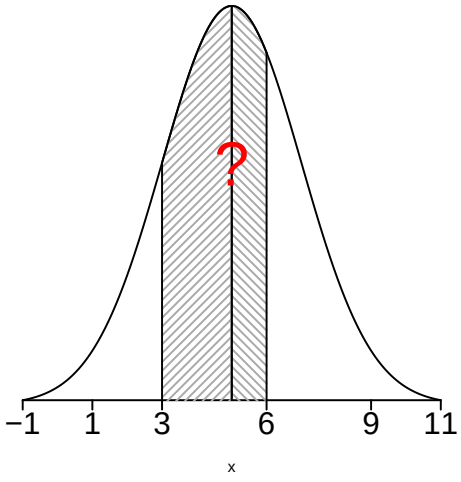
$$Z = \frac{X - \mu}{\sigma}.$$

Note: Once we have this standardised normal variable, it is straightforward to read the required probabilities from the Table 2. In other words, we can reduce *any* normal distribution question to a question about the standard normal distribution.

Example: A scientist found that the distribution of the height X (in meters) of a certain native plant is $X \sim N(5, 2^2)$. Find the probability that a randomly selected tree from this type is between 3 and 6 meters. That is, find $P(3 < X < 6)$.

Solution: Clearly $\mu = 5$ and $\sigma = 2$ and this is a non-standard normal distribution.

Diagram



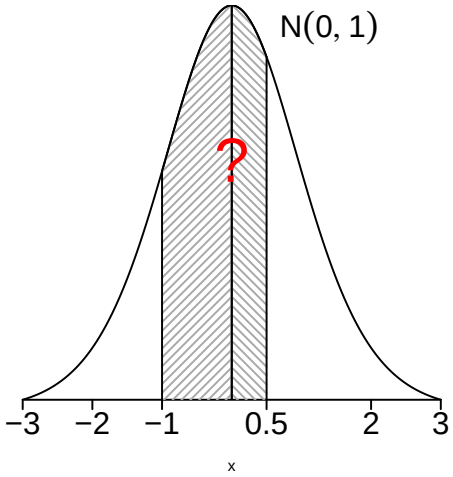
To calculate the required probability, we need to standardise each endpoint by subtracting the mean and dividing by the standard deviation. That is,

$$\frac{X - 5}{2}$$

$$P(3 < X < 6) = \underline{\qquad\qquad\qquad\qquad}$$

$$= \underline{\qquad\qquad\qquad\qquad}$$

Diagram to find $\mathrm{P}(-1.00 < Z < 0.50)$:



By symmetry and using Normal Table, we have

$$P(-1.00 < Z < 0.50) = \underline{\qquad\qquad\qquad\qquad}$$
$$= \underline{\qquad\qquad\qquad\qquad}$$
$$= \underline{\qquad}$$

6.6 Applications

Now we look at some examples of the normal distribution.

Example: Assume that cabbage yields are normally distributed with mean $\mu = 1.4$ kg/plant and standard deviation of $\sigma = 0.2$ kg/plant. Find the probability that a randomly selected cabbage yield is less than 1 kg.

Solution: Let X be the weight of a cabbage yield selected at random. Therefore, the distribution is

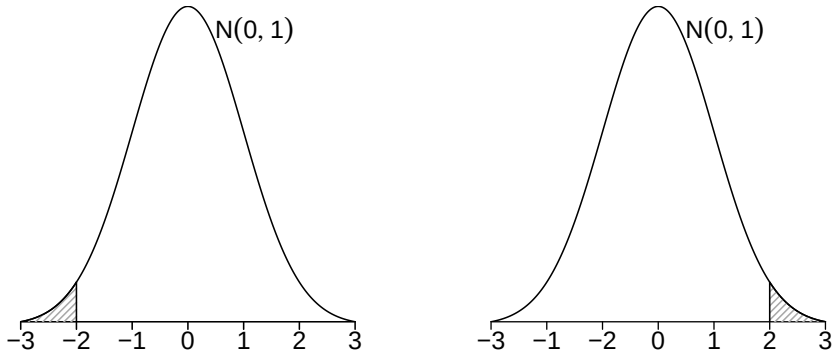
$$X \sim \underline{\qquad\qquad}$$

We need $\mathrm{P}(X < 1)$. Now,

$$P(X < 1) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

By symmetry, $\mathrm{P}(Z < -2) = \mathrm{P}(Z > 2)$.

Diagram:



From Table 2 with $z = 2$ gives 0.9772. Therefore,

$$P(Z \leq 2) = \underline{\quad} \quad \text{or} \quad P(Z \geq 2) = \underline{\qquad\qquad}$$
$$\Rightarrow P(X < 1) = \underline{\quad}$$

Read example on P.71-72 and check your answers with Normal Table.

6.7 Linear Combinations of Random Variables - P.72-75

In many real world applications of statistics, we need to consider the sums and differences of random variables. A good application is given below:

Example: Suppose that the distribution of weights of students is approximately $N(70, 5^2)$ (in Kg). What is the probability that a group of 9 students weigh more that 650Kg?

Solution: It is sensible to think that the total weight is normally distributed with mean $70 \times 9 = 630$ Kg and variance $25 \times 9 = 225 = 15^2$ Kg^2.

Therefore, the total weight, $T \sim N(630, 15^2)$. Now,

$$P(T > 650) = \_\_\_\_\_$$
$$= \_\_\_\_\_$$

General Formula

Suppose that $X_1, X_2, \ldots, X_n$ are n *independent* normal random variables from $\mathrm{N}(\mu, \sigma^2)$. Then

$$T = \sum_{i=1}^{n} X_i \sim \mathrm{N}(n\mu, n\sigma^2).$$

Even More General Formula

Suppose that $X_1, X_2, \ldots, X_n$ are n *independent* normal random variables from $\mathrm{N}(\mu_1, \sigma_1^2)$, $\mathrm{N}(\mu_2, \sigma_2^2)$, ..., $\mathrm{N}(\mu_n, \sigma_n^2)$, respectively. Suppose further that $a_1, \ldots, a_n$ are given constants. Then

$$T = \sum_{i=1}^{n} a_i X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

Example: (Multiple Choice) Suppose X and Y are independent normally distributed random variables. The variance of X is equal to 16; and the variance of Y is equal to 9. Let $Z = X - Y$. What is the standard deviation of Z?

(a) 2.65

(b) 5.00

(c) 7.00

(d) 25.0

(e) It is not possible to answer this question, based on the given information

Solution: The correct answer is (b). We recognize that variable Z is a sum of two *independent* normal random variables. The coefficients are $a_1 = 1, a_2 = -1$. As such, the variance of Z is equal to the variance of X plus the variance of Y:

$$\text{Var}(Z) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$
$$\text{SD}(Z) = \underline{\qquad\qquad}$$

Omit Normal approximation to the binomial.

Extra Problems to Practice

Try: Q1 to Q11 (P.86-89)