Stat 113 (Introduction to Mathematical Statistics)

Lesson 3.4 The Normal Distribution

Learning Outcomes

At the end of the lesson, students must be able to

1. Describe the key properties of a random variable having a normal distribution such as the mean, variance, and moment generating function,
2. Transform a normal random variable Y into the standard normal random variable Z, and
3. Compute probabilities associated with random variables having a normal distribution.

Introduction

The most widely used continuous probability distribution is the normal distribution, a distribution with a bell shape curve. Many random variables have distributions that are closely approximated by a normal probability distribution. Many of the techniques used in applied statistics are based upon the normal distribution.

Definition:

A random variable Y is said to have a normal distribution with mean μ and variance σ^2 if its probability density function is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \ -\infty < y < \infty, \ -\infty < \mu < \infty, \ \sigma > 0 \\ 0, \ \text{elsewhere} \end{cases}$$

We use the shorthand notation $Y \sim N(\mu, \sigma^2)$ to denote that Y has a normal distribution with mean μ and variance σ^2.

Properties of the normal curve:

1. All normal curves are **bell-shaped** with points of inflection at $\mu \pm \sigma$. A point of inflection is where the graph changes from moving upward with increasing steepness to downward with decreasing steepness.

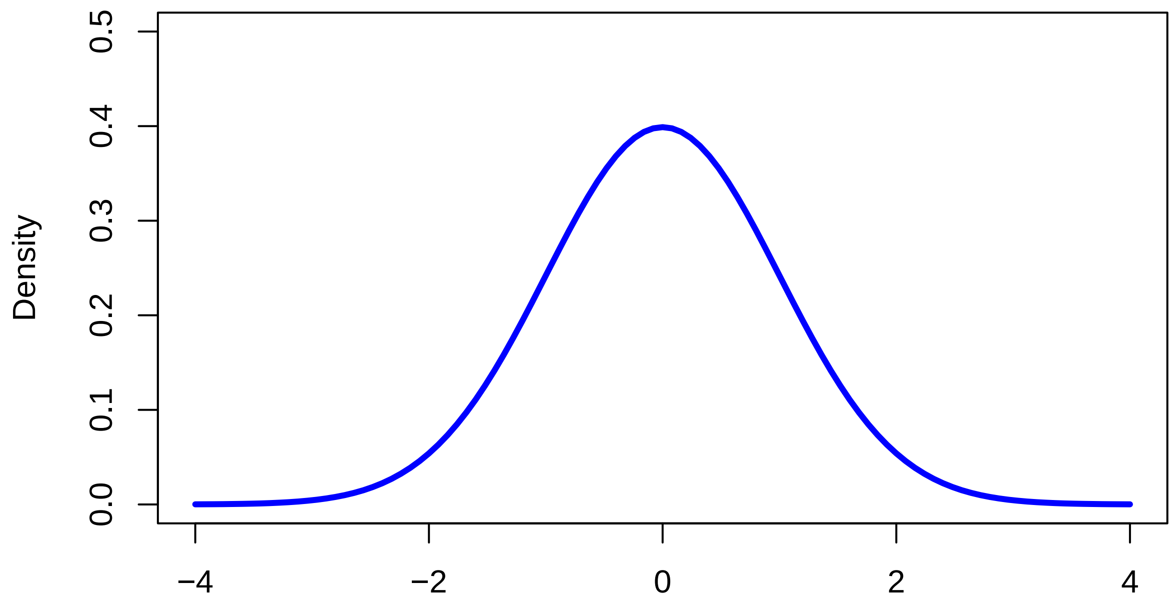


Figure 1: N(0,1) distribution

2. All normal curves are **symmetric** about the mean μ, that is, $f_Y(\mu + y) = f_Y(\mu - y)$, for all y.

3. All normal curves are positive for all y. That is, $f_Y(y) > 0$ for all y.

4. The normal curve is asymptotic to the x-axis. That is,

$$\lim_{y \to -\infty} f_Y(y) = 0, \text{ and}$$
$$\lim_{y \to \infty} f_Y(y) = 0$$

6. The height of any normal curve is maximized at $y = \mu$.

7. The area under the normal curve is 1.
8. The shape of any normal curve depends on its mean μ and standard deviation σ . This is illustrated in the graph below.

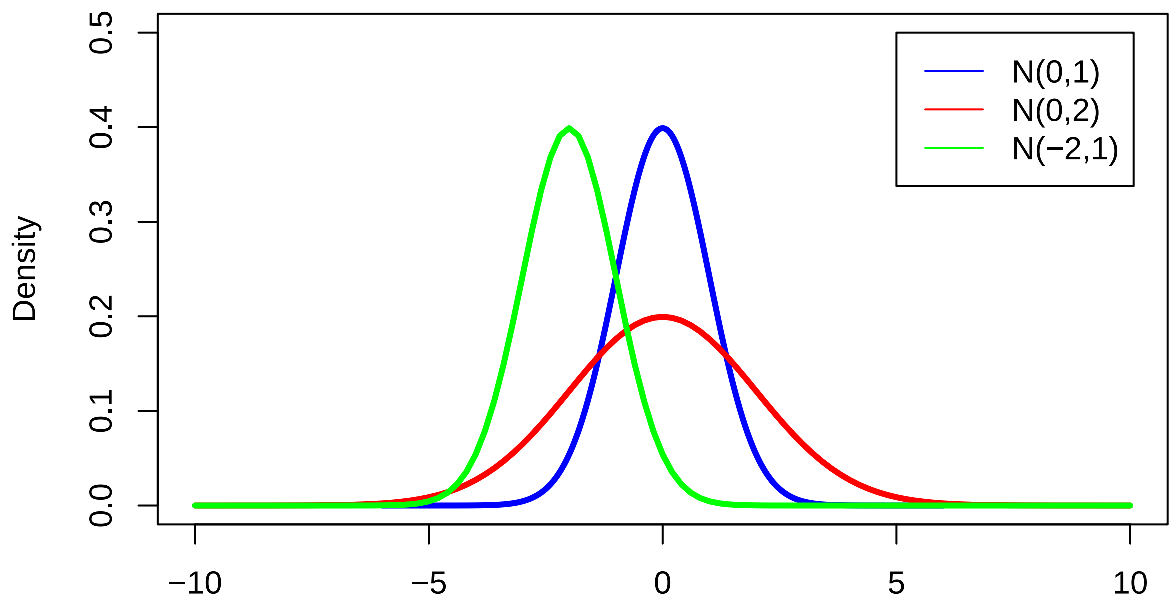


Figure 2: Various Normal Distributions

Theorem:

If $Y \sim N(\mu, \sigma^2)$, then

$$E(Y) = \mu$$
$$V(Y) = \sigma^2$$
$$m_Y(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} = exp\left[\mu t + \frac{\sigma^2 t^2}{2}\right]$$

The Standard Normal Distribution

The standard normal probability distribution is a normal distribution with mean equal to zero and variance equal to 1. We say that a random variable Z is said to have a standard normal distribution if its pdf is given by

$$f_Z(z) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, & -\infty < z < \infty \\ 0, & \text{elsewhere} \end{cases}$$

We write $Z \sim N(0,1)$ if Z has a standard normal distribution.

How do we obtain $Z \sim N(0,1)$ from $Y \sim N(\mu, \sigma^2)$? Any normal random variable Y can be "converted" into a standard normal random variable by a result known as standardization. We do this by applying the transformation

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

The above result can then be used to calculate probabilities under the normal curve as shown below. Suppose $Y \sim N(\mu, \sigma^2)$, then for any real numbers $a < b$ we have

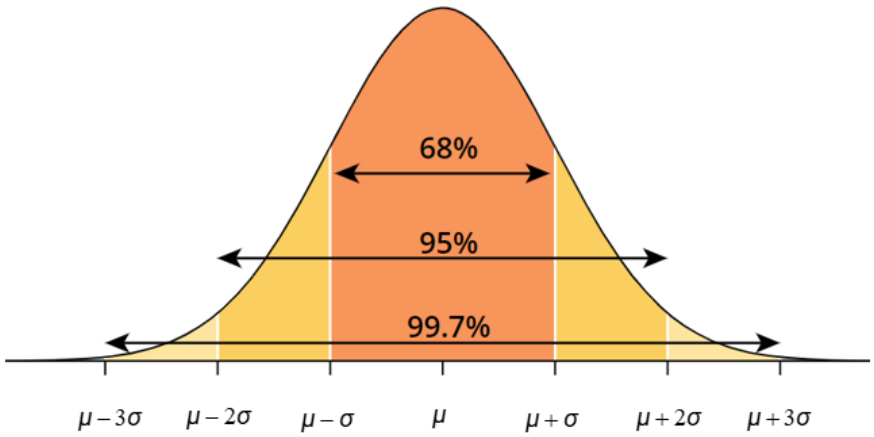
$$\begin{aligned} P(a < Y < b) &= P\left(\frac{a-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) \\ &= P\left(z_1 < Z < z_2\right), \text{ where: } z_1 = \frac{a-\mu}{\sigma} \text{and } z_2 = \frac{b-\mu}{\sigma} \\ &= F_Z(z_2) - F_Z(z_1) \end{aligned}$$

where $F_Z(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\ dt$ is the cumulative distribution function of the standard normal distribution.

This integral does not exist in closed form. However, probability tables exist that catalogue its value for different values of z (which are determined using numerical integration methods). We call these tables as Z tables. Before computing packages like R, these tables were needed. However, they are now somewhat outdated; for example, the R command $pnorm(y, \mu, \sigma)$ calculates the CDF of any $Y \sim N(\mu, \sigma^2)$ random variable at the value y. As in the previous lessons, we rely on the R program to calculate probabilities and quantiles. But for purpose of knowing how to use a Z table, you may watch this YouTube video: https://rb.gy/r6qvh7.

The Empirical Rule

The Empirical Rule states that in a normal distribution about 68% of data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean. This is illustrated in the diagram below.



Example 3.4.1:

Suppose the amount of time, in minutes, that a person must wait before he/she can be served by a teller in a bank is normally distributed with mean 15 minutes and standard deviation 2.5 minutes. Calculate the following.

a. The probability that a person needs to wait for more than 10 minutes before he/she can be served.

b. The probability that a person needs to wait between 12 and 18 minutes before he/she can be served.

SOLUTION

Let Y be the waiting time (in minutes). Then $Y \sim N(15, 2.5^2)$.

a. The probability that a person needs to wait for more than 10 minutes before he/she can be served is given by

$$\begin{aligned} P(Y > 10) &= 1 - P(Y \leq 10) \\ &= 1 - P\left(\frac{Y - 15}{2.5} \leq \frac{10 - 15}{2.5}\right) \\ &= 1 - P(Z \leq -2) \\ &\approx 1 - 0.0227, \text{ from Z table} \\ &= 0.9773 \end{aligned}$$

If you have an R compiler app in your phone, you can just type the following code *1 - pnorm(-2)* to get the answer. Or without transforming the value of Y into Z, you can also just type *1 - pnorm(10, 15, 2.5)* and you will get the same answer.

b. The probability that a person needs to wait between 12 and 18 minutes before he/she can be served is given by

$$
\begin{aligned}
P(12 < Y < 18) &= P\left(\frac{12-15}{2.5} < Z < \frac{18-15}{2.5}\right) \\
&= P\left(-1.2 < Z < 1.2\right) \\
&= P(Z < 1.2) - P(Z < -1.2) \\
&\approx 0.8849 - 0.1151, \text{ both were obtained from the Z table} \\
&= 0.7698
\end{aligned}
$$

You can use either $pnorm(1.2) - pnorm(-1.2)$ or $pnorm(18, 15, 2.5) - pnorm(12, 15, 2.5)$ in R to get the above probability.

Example 3.4.2

A company that manufactures and bottles apple juice uses a machine that automatically fills 16-ounce bottles. There is some variation, however, in the amounts of liquid dispensed into the bottles that are filled. The amount dispensed has been observed to be approximately normally distributed with mean 16 ounces and standard deviation 1 ounce. What proportion of bottles will have more than 17 ounces dispensed into them?

SOLUTION

Let Y be the volume of 16-oz bottles. We have $Y \sim N(16, 1)$. The probability that a bottle is filled with more than 17 ounces is

$$
\begin{aligned}
P(Y > 17) &= 1 - P(Y < 17) \\
&= 1 - P\left(\frac{Y-16}{1} < \frac{17-16}{1}\right) \\
&= 1 - P\left(Z < 1\right) \\
&\approx 1 - 0.8413, \text{ from Z table} \\
&= 0.1587
\end{aligned}
$$

Therefore, $0.1587 \times 100\% = 15.87\%$ of bottles are filled more than 17 ounces.

Example 3.4.3

Scores on an examination are assumed to be normally distributed with mean 80 and variance 36.

a. What is the probability that a person taking the examination scores below 75?

b. Suppose that students scoring in the top 10% of this distribution are to receive a grade of 1.0. What is the minimum score a student must get to earn a grade of 1.0?

c. What must be the cutoff point for passing the examination if the examiner wants only the top 28.1% of all scores to be passing?

SOLUTION

Let Y be the score of a student in an examination. We have $Y \sim N(80, 36)$.

a. The probability that a person taking the examination scores below than 75 is

$$\begin{aligned} P(Y < 75) &= P\left(\frac{Y-80}{6} < \frac{75-80}{6}\right) \\ &= P\left(Z < -0.83\right) \\ &\approx 0.2033, \text{ using the R code } \ pnorm(-0.83) \end{aligned}$$

b. Suppose we let y be the minimum score of students who will be given a grade of 1.0. That is, y is the minimum score of a student who belongs to the top 10% of the class. This means that

$$\begin{aligned} P(Y \geq y) &= 0.10, \text{ or} \\ \implies P(Y < y) &= 1 - P(Y \geq y) \\ &= 1 - 0.10 \\ &= 0.90 \end{aligned}$$

The score y here is referred to as a quantile, specifically, the 90% percentile.

We learned before that

$$P(Y < y) = P(Z < z)$$

Therefore,

$$P(Z < z) = 0.90$$

To get z we use the R code $qnorm(0.90)$. This will give us $z \approx 1.28$. Then, we calculate y using the transformation

$$z = \frac{y - \mu}{\sigma} \implies 1.28 = \frac{y - 80}{6}$$

Using algebra, we have

$$\begin{aligned} 1.28 &= \frac{y-80}{6} \\ 1.28 \times 6 &= y - 80 \\ 1.28 \times 6 + 80 &= y \\ y &= 87.68 \end{aligned}$$

Therefore, a student must get a score of 87 or higher inorder to get a grade of 1.0.

c. The solution is similar to (b). Let y be the cut-off score in order to pass the examination. This means,

$$P(Y \geq y) = 0.281$$

$$\begin{aligned} P(Y \geq y) &= 0.281, \text{ or} \\ \implies P(Y < y) &= 1 - P(Y \geq y) \\ &= 1 - 0.281 \\ &= 0.719 \end{aligned}$$

Based on the identity $P(Y < y) = P(Z < z)$ and using the R code $qnorm(0.719)$, we get $z \approx 0.58$. Then we compute y as follows:

$$\begin{aligned} 0.58 &= \frac{y - 80}{6} \\ 0.58 \times 6 &= y - 80 \\ 0.58 \times 6 + 80 &= y \\ y &= 83.48 \end{aligned}$$

Therefore, a student must get a score equal to 83 or greater inorder to pass the examination.

Learning Task

Instruction: Answer the following as indicated.

The average time a person spends at the grocery store in a medium-sized mall is 96 minutes with standard deviation of 17 minutes. Assume the variable is normally distributed.

a. If a visitor is selected at random, find the probability that he or she will spend at least 120 minutes at the grocery store.

b. If a visitor is selected at random, find the probability that he or she will spend at most 80 minutes at the grocery store.

c. At most how long (in minutes) will 90% of customers spend in the grocery store?