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.

The normal distribution is the most important probability distribution in statistics because many continuous data in nature resembles this bell-shaped curve when compiled and graphed.

Definition

A continuous 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

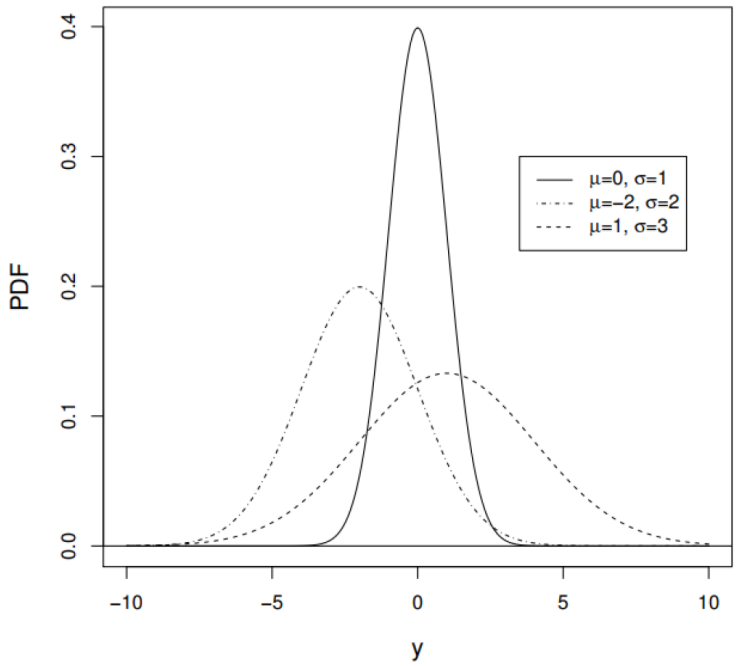
$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}},\ -\infty < y < \infty;\ -\infty < \mu < \infty;\ \sigma^2 > 0$$

We write $Y \sim N(\mu, \sigma^2)$. The above PDF can also be expressed as

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right],\ -\infty < y < \infty;\ -\infty < \mu < \infty;\ \sigma^2 > 0$$

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 downward with increasing steepness to downward with decreasing steepness.
2. All normal curves are symmetric about the mean μ, that is, $f_Y(\mu - y) = f_Y(\mu + y), \forall y$.
3. The area under an entire normal curve is 1.
4. The limit of $f_Y(y)$ as y goes to infinity is 0, and the limit of $f_Y(y)$ as y goes to negative infinity is 0. That is,
$$\lim_{y \to -\infty} f_Y(y) = \lim_{y \to \infty} f_Y(y) = 0$$
5. The height of any normal curve is maximized at $y = \mu$.
6. The shape of any normal curve depends on its mean (μ) and standard deviation (σ) . This is illustrated in the graph below.



Theorem

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

$$E(Y) = \mu$$
$$V(Y) = \sigma^2$$
$$m_Y(t) = exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

Proof: Left as an exercise!

The Standard Normal Distribution

The standard normal probability distribution is a normal distribution with mean equal to zero and variance equal to 1.

Definition

If a random variable Z has PDF

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{z^2}{2}\right],\ -\infty < z < \infty$$

Then, we write $Z \sim N(0,1)$.

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 Z by a process 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. Suppose $Y \sim N(\mu, \sigma^2)$,then

$$\begin{aligned} P(a < Y < b) &= P\left(\underbrace{\frac{a-\mu}{\sigma}}_{z_1} < \underbrace{\frac{Y-\mu}{\sigma}}_{Z} < \underbrace{\frac{a+\mu}{\sigma}}_{z_2}\right) \\ &= P(z_1 < Z < z_2), \\ &= F_Z(z_2) - F_Z(z_1) \end{aligned}$$

where $F_Z(z)$ is the CDF $Z \sim N(0,1)$ and can be calculated as

$$F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-t^2}{2}\right)\ dt$$

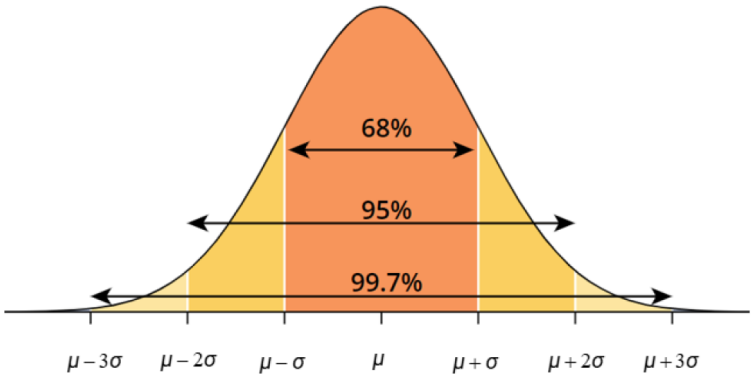
This integral does not exist in closed form. However, probability tables exist that catalogue its value for different values of z (which are determined using numerical integration methods). We call these tables as **Z** tables.

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 $N(\mu, \sigma^2)$ random variable at the value y.

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 1:

The World Health Organization uses a normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$ to describe the systolic blood pressure (SBP) of males (aged 18 and over). SBP is measured in millimeters of mercury (mm Hg). Let Y denote the SBP of an individual selected from this population.

a. An SBP of 90 mm Hg or less is generally considered to be "low." Find $P(Y \leq 90)$.

b) Find the 80^{th} percentile of this distribution.

SOLUTION: Left as classroom exercise.

Example 2:

The grade point average (GPA) of a large population of college students is approximately normally distributed with mean equal to 2.4 and standard deviation of 0.8. What fraction of the students possesses a GPA in excess of 3.0?

SOLUTION: Left as classroom exercise.

Example 3:

Scores on an examination are assumed to be normally distributed with mean 78 and variance 36.

a. What is the probability that a person taking the examination scores higher than 72?

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 achieve 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?

d. Approximately what proportion of students have scores 5 or more points above the score that cuts off the lowest 25%?

e. If it is known that a student's score exceeds 72, what is the probability that his or her score exceeds 84?

SOLUTION: Left as classroom exercise.