1.3. The Normal Distributions

1. Density curves

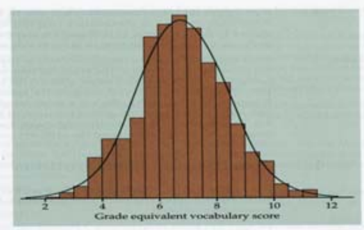

FIGURE 1.20 Histogram of the Iowa Test vocabulary scores for Gary, Indiana, seventh graders, showing the approximation of the distribution by a normal curve.

A density curve is a curve that

- **Is always on or above the horizontal axis and**
- **Has a area exactly 1 underneath it.**

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the relative frequency of all observations that fall in that range.

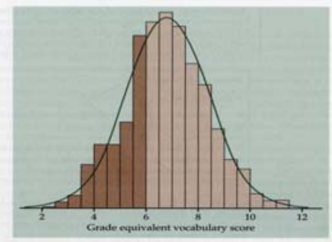

FIGURE 1.21(a) The relative frequency of scores less than or equal to 6.0 from the histogram is 0.303.

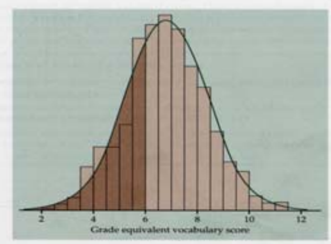

FIGURE 1.21(b) The relative frequency of scores less than or equal to 6.0 from the density curve is 0.293.

2. Normal distribution

i). What is a normal distribution?

- One particularly important class of density curves are symmetric, unimodal, and bell-shaped. They are called normal curves.
- The height of the normal curve at any point x is given by:

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

$$where \quad -\infty < \mu < \infty \quad and \quad \sigma > 0$$

ii). Properties of normal distribution

a). A normal distribution is uniquely determined by its mean μ and standard deviation σ. Its mean equals to median.

b). Normal curve is symmetric, unimodal, and bell-shaped. We can locate its mean μ and standard deviation σ by eye on the graph. Mean μ is at the center of the curve, σ is the distance from μ to the point at which changes of curvature take place.

c). The **68--95--99.7 rule**

The 68-95-99.7 rule:

Approximately 68% of the observations fall in [μ-σ, μ+σ]

Approximately 95% of the observations fall in [μ-2σ, μ+2σ]

Approximately 99.7% of the observations fall in [μ-3σ, μ+3σ]

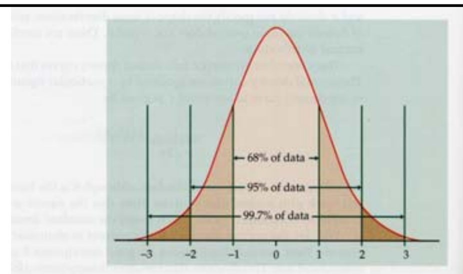

FIGURE 1.25 The 68–95–99.7 rule for normal distributions.

iii). Standardizing observations

If x is an observation from a normal distribution that has mean μ and standard deviation σ, the standardized value of x is

$$z = \frac{x - \mu}{\sigma}$$

A standardized value is called a **z-score.**

Note: we will use uppercase letters to denote variables, and lower case letters for values of variables. $X \sim N(\mu, \sigma)$

Generally we have: If $X \sim N(\mu, \sigma)$, then $X_{new}=aX+b \sim N(a+b\mu, |b|\sigma)$ I.e. linear transformation of a normal distribution variable remains normal.

iv). The standard normal distribution

- The standard normal distribution is the normal distribution N(0, 1) with mean 0 and standard deviation 1.
- If a variable X has any normal distribution N(,) with mean and standard deviation , then the standardized variable

$$Z = \frac{X - \mu}{\sigma} \propto N(0,\ 1)$$

- The standard normal table
  - A table of areas under the standard normal curve. The table entry for each value z is the area under the curve to the left of z.
- Problems:

  a). What proportion of observations on a standard normal variable Z take values less than 1.2?

  b). What proportion of observations on Z take values greater than –2.05?

  c). What proportion of observations on Z take values between –1.05 and 1.16?

v). Normal distribution calculations

Example 1:

The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined mathematics and verbal parts of the SAT exam in order to compete in their first college year. (Higher scores are required for students with poor high school grades.) In 2000, the scores of the more than one million students taking the SATs were approximately normal with mean 1019 and standard deviation 209. What percent of all students had SAT scores of at least 820:

Example 2:

The NCAA considers a student a "partial qualifier" eligible to practice and receive an athletic scholarship, but not to compete, if the combined SAT score is at least 720. What percent of all students who take the SAT would be partial qualifiers?

Example 3:

Scores on the SAT verbal test in recent years follow approximately the N (505, 110) distribution. How high must a student score in order to place in the top 10% of all students taking the SAT?

Summary for normal distribution calculations:

In all, we have two types of problems for normal calculations.

To calculate proportions (probabilities)	To calculate the critical values
a). State the problem Write down in the forms: $P(X < a)$ or $P(X > b)$ or $P(b < X < a)$	a). State the problem Write down in the forms: to find x such that $P(X < x) = p_0$ or $P(X > x) = p_0$ (change to $P(X < x)=1- p_0$)
b). Standardize $P(X < a) = P(\frac{X-\mu}{\sigma} < \frac{a-\mu}{\sigma})$ $= P(Z < \frac{a-\mu}{\sigma})$	b). Use the table to find z such that $P(Z < z) = p_0$ (or $1 - p_0$)
c). Use the table	c). Unstandardize $x = \mu + \sigma z$ Then x is what we want.

vi). The normal quantile plots

- We use a normal quantile plot to judge whether data are approximately normal.
- How to look at a normal quantile plot?
  - If points lie close to a straight line, we say that data are normal.
  - If there is a systematic deviation from a straight line, data are non-normal
  - Points that are far away from the overall pattern are outliers.