Chapter 1 - Looking at Data

Read Chapter 1 first, then read the notes and try the WEBCT assignment questions. If you need more practice, try the practice questions with answers available on the web.

Exercises:

In this chapter you will learn how scientists describe data. You will be introduced to some definitions and then learn how data is displayed graphically and numerically. The concept of a distribution will be introduced and the most popular distribution in science – **the normal** – will be discussed. But, first we will consider a few important terms.

View the video-What is statistics?

A **variable** is a measured characteristic of an object. There are two types of variables:

- **quantitative** variables take on **numerical** values, for example, height, weight etc.

- **qualitative** (or **categorical**) variables place an individual into one of several categories, for example, gender.

  We can code gender Female = 1, Male = 0 or use any other coding scheme since the numbers just distinguish the two genders.

In general, it makes sense to do computations with quantitative variables. A **data set** is a collection of values of one or more variables.

Next we see how to display qualitative or categorical data in a table and follow with how to display quantitative data.

Distribution of a Categorical Variable

A distribution of a categorical variable is a table giving each possible value of the variable and the number of times that value occurred.

It is also called a **frequency table**.

Sometimes proportions are reported in the table in which case it is referred to as a **relative frequency table**.

Example: Consider first year social work. Assume we define our categorical variable as the gender of first year social work students (i.e., M for male, F for female). The frequency table is:

M	140
F	160

The total number of observations is:

$$n = 160 + 140 = 300$$

We can also report the proportion of women as:

$$\frac{160}{300} = \frac{8}{15} =$$

.53 or 53% ***proportion of women***

<u>NOTE</u>: Frequency tables may have more than two categories.

Sometimes a chart of the percents for each category is produced and displayed in a bar graph. The heights of the bar give the percents as can be seen below.

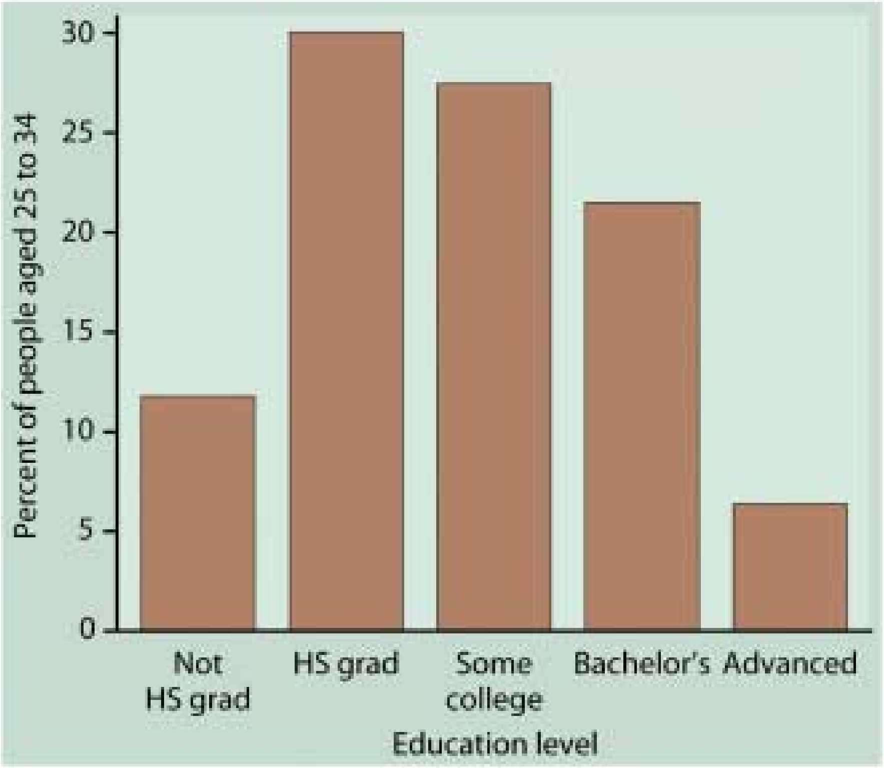


Figure 1.1a. Bar graph of the educational attainment of people aged 25 to 34 years

Distribution of a Quantitative Variable

Example: Let's examine the distribution of the numbers of some data set. For example a Psychological Survey Study Habits and Attitudes (SSHA). The SSHA evaluates college students' motivation, study habits, and attitudes towards school.

There are n = 18 women, first year university students. The values vary. This is what we mean by **<u>variation</u>**. Is there a pattern to the <u>variation</u>? Look at the variable's **<u>distribution</u>**. Scores of 18 women are listed below in two columns:

154	103
109	126
137	126
115	137
152	165
140	165
154	129
178	200
101	148

<u>Stem Plot</u>

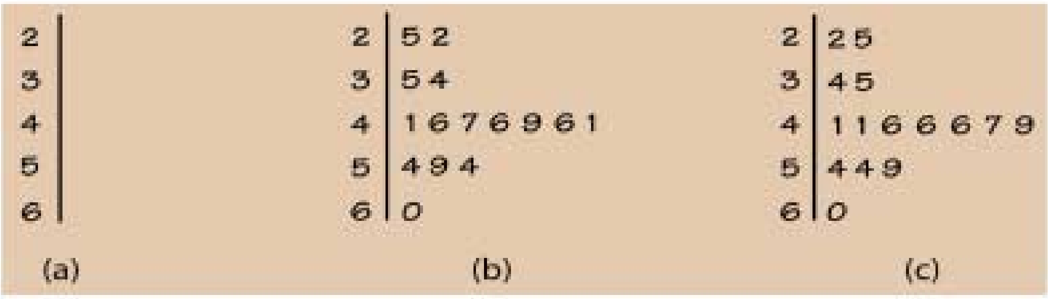
First, consider a stem plot. The number is divided into two parts, a stem and a leaf. For example, the score of 154 has a stem of 15 and a leaf of 4; the number 109 has a stem of 10 and a leaf of 9, etc. The complete plot is given below:

stem	leaf
10	139
11	5
12	669
13	77
14	08
15	244
16	55
17	8
18	
19	
20	0

remember 154 is

$\underline{15}$ $\underline{4}$

stem leaf



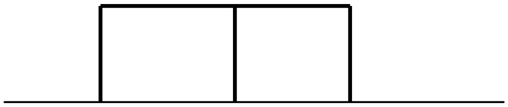
```
2 |              2 | 5 2             2 | 2 5
3 |              3 | 5 4             3 | 4 5
4 |              4 | 1 6 7 6 9 6 1   4 | 1 1 6 6 6 7 9
5 |              5 | 4 9 4           5 | 4 4 9
6 |              6 | 0               6 | 0
 (a)                (b)                 (c)
```

Figure 1.2 Making a stem plot of the data in Example 1.4. of the text (a) Write the stems. (b) Go through the data writing each leaf on the proper stem. (c) Arrange the leaves on each stem in order out from the stem.

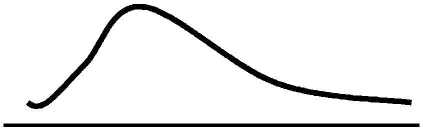
The **median** is the middle of the distribution (around 140). We will define the median precisely shortly.

Here are some common distributions:

Symmetric Distribution



right side = left side

Skewed to Right  tail to right

Another way to view the data is a histogram. **Stem plots work best with small data sets** (say, less than 100 observations) and **histograms work best with large data sets**.

Histogram

In a histogram we use intervals to display values. The intervals give a sense of the shape/pattern of the distribution of the data.

How to construct a histogram:

(1) Divide range into equal widths.

Largest value = 200   Smallest value = 101

Range = Largest – Smallest = 200 – 101 =99

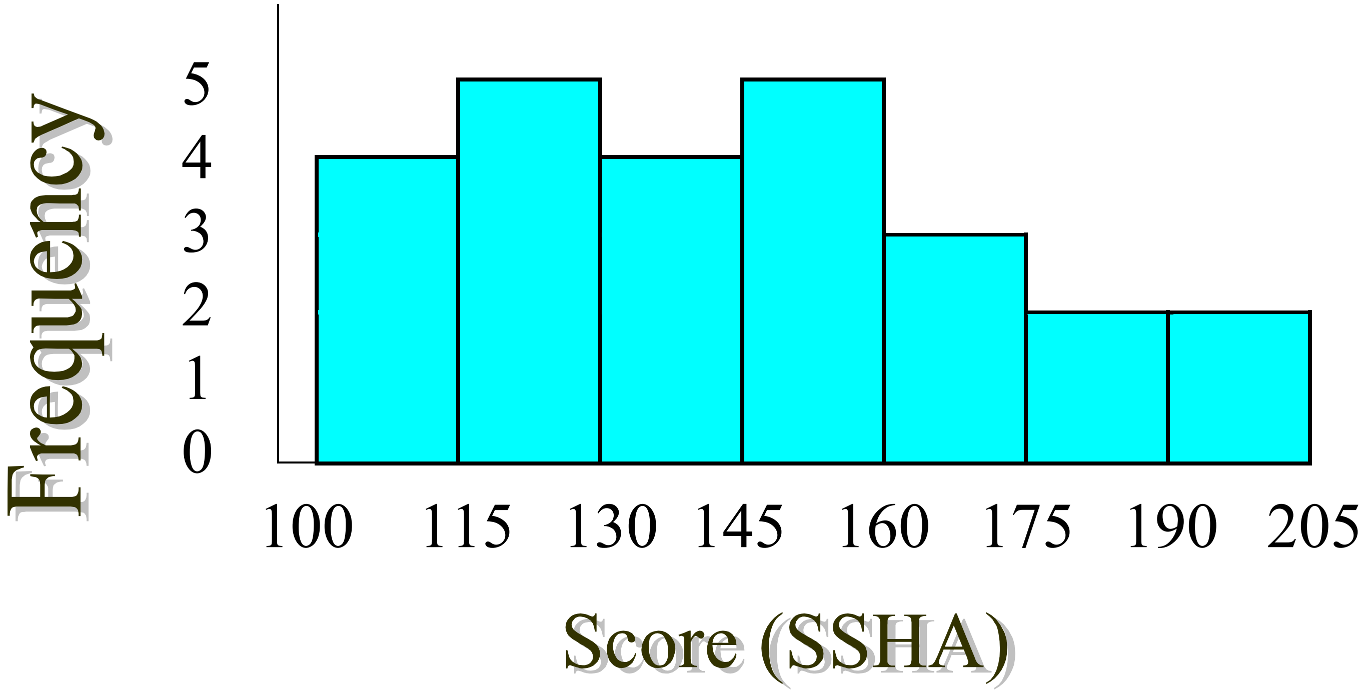
Say you want 7 classes. Take the range and divide by the number of classes. In our example, we have 99 ÷ 7 = 14 approximately as the length of each class. The class intervals are given below.

(2) Count the number of observations in each interval.

Class Interval	Count	Frequency	Percent
100-114	III	3	17%
115-129	IIII	4	22%
130-144	III	3	17%
145-159	IIII	4	22%
160-174	II	2	12%
175-189	I	1	5%
190-204	I	1	5%
		18	100%

The total count (18) should equal the number of observations in the data set. **The percent may not add to 100 due to rounding.**

The histogram clearly shows the frequency of the observations and what range they fall into. However, it does not give the actual data value as in the stem plot. For example, there is only one observation between 190-204 according to the histogram, but from the stem plot we know the value as 200.



The height of the bars gives the frequency of scores in the given range of the SSHA. If we divide the frequency by the total number of observations (n) we obtain the **relative frequency** or percent in the previous table. Next we look at numerical summary values of a distribution in contrast to graphical methods.

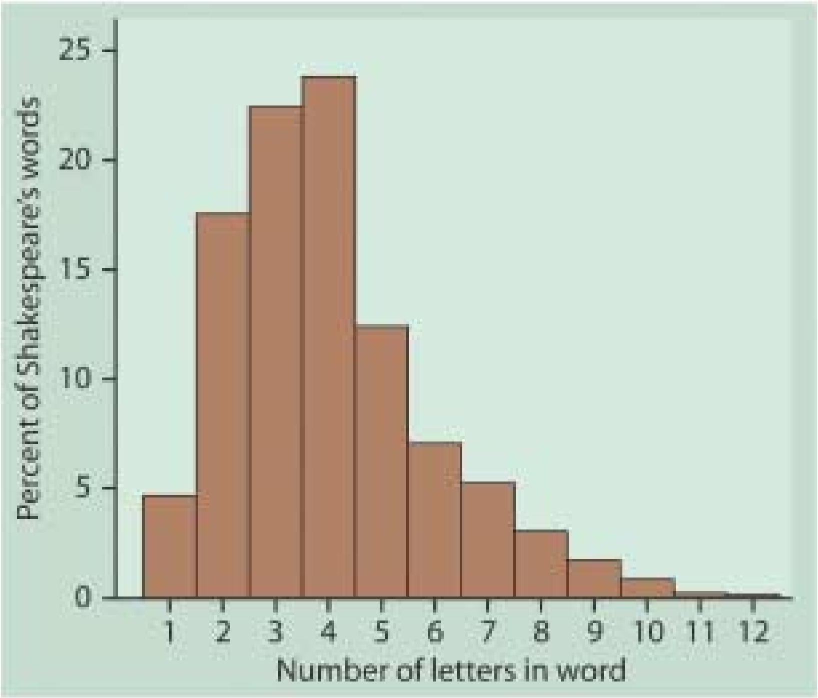


Figure 1.11 Histogram of lengths of words used in Shakespeare's play, for Exercise 1.18

Describing Distributions

The **mean** or average is the center of a distribution.

$$\mathbf{MEAN} = \bar{\mathbf{X}} = \frac{\mathbf{1}}{\mathbf{n}}\left(\mathbf{X_1 + X_2 + \ldots + X_n}\right)$$

$$\overline{X} = \frac{1}{n} \sum X_i$$

For example, in our previous data set:

$$\overline{X} = \frac{1}{18}\left(154 + 109 + \ldots + 148\right) = \frac{2539}{18} = 141.06$$

The **median** is the center of the distribution such that half the data is below the median and half above the median. Here are the steps in order to calculate the median:

(1) Sort the numbers in size from smallest to largest.

(2) If n is odd then the median is [(n+1)÷2] observations from the bottom counting from the smallest observation.

(3) If n is even then the median is the average of the 2 center observations, [(n+1)÷2] from bottom.

Example: Let’s try this with the SSHA scores data set. First examine the stem plot. There are n=18 observations, n is even.

Median is $\frac{(18 + 1)}{2} = \frac{19}{2} = 9.5$ observations.

There is no 9.5 observation so we take the average of the 9^{th} and 10^{th} observations. From stem plot the 9^{th} and 10^{th} observations are averaged to obtain $\frac{137+140}{2} = 277 = 138.5$ is our median.

Percentiles

Median = 50th percentile (halfway point)

pth percentile - p percent of the data fall at or below it

quartiles - 1st quartile = 25th percentile
3rd quartile = 75th percentile

How do we calculate quartiles?

1) Locate Median

2) Find the median of all the observations below the Median. This is called the 1st Quartile (Q_1).

3) Find the median of all observations above the Median. This is called the 3rd Quartile (Q_3)

<u>Example</u>: For our SSHA example we saw that the Median = 138.5

Find 1st Quartile. We know the number of observations below the median is 9. So we have n = 9, an odd number of observations below 138.5. Substitute in our formula:

$$\frac{(9+1)}{2} = \frac{(10)}{2} = \mathbf{5^{th}\ Observation}$$

Next, count 5 observations from the bottom of the stem plot to get:

$$Q_1 = 126$$

Find the 3rd Quartile. Again n = 9 odd:

$$\frac{(9 + 1)}{2} \quad = \quad \frac{(10)}{2} \quad = \quad \textbf{5}^{\textbf{th}}\ \textbf{Observation}$$

So we count 5 observations down from the top of the stem plot to get $Q_3 = 154$.

The **interquartile range** or **IQR** gives the center 50% of the data. If the IQR is small, then the researcher knows the center 50% is quite concentrated or has small spread.

$$\text{IQR} = Q_3 - Q_1 = 154 - 126 = 28$$

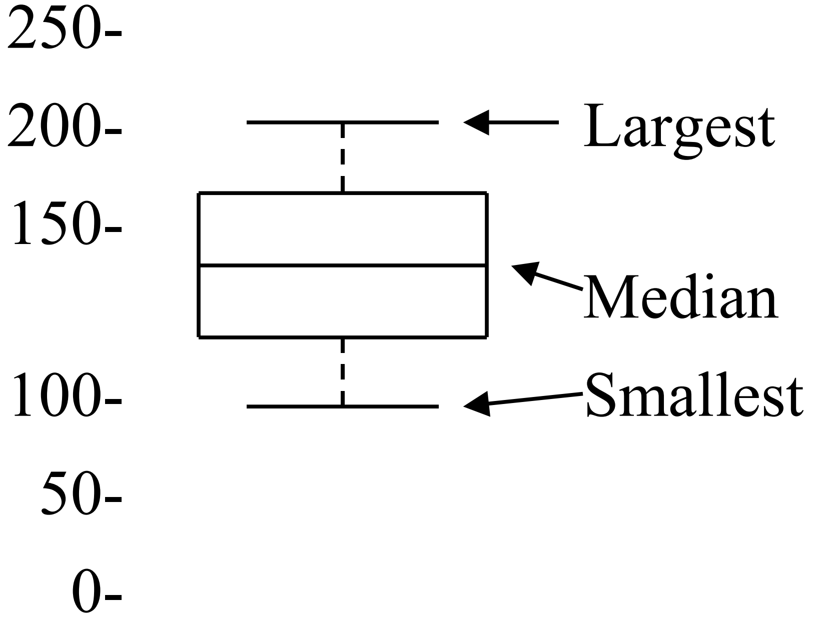
Box Plots **-** Useful numerical display to describe or compare data sets. You need the

following 5 numbers to construct a box plot.
For a variable in a data set calculate:

- largest and smallest values
- the median
- the 1st and 3rd quartiles

Example: For the SSHA data we calculated:

smallest	**101**	**– end of box plot (bottom)**
1st quart	**126**	**– next line of box plot**
median	**138.5**	**– middle**
3rd quart	**154**	**– next line of box plot (top)**
largest	**200**	**– end of box plot**



please see below for another example of boxplots and their use. Note how the highway and city milage for both types of car can easily be compared.

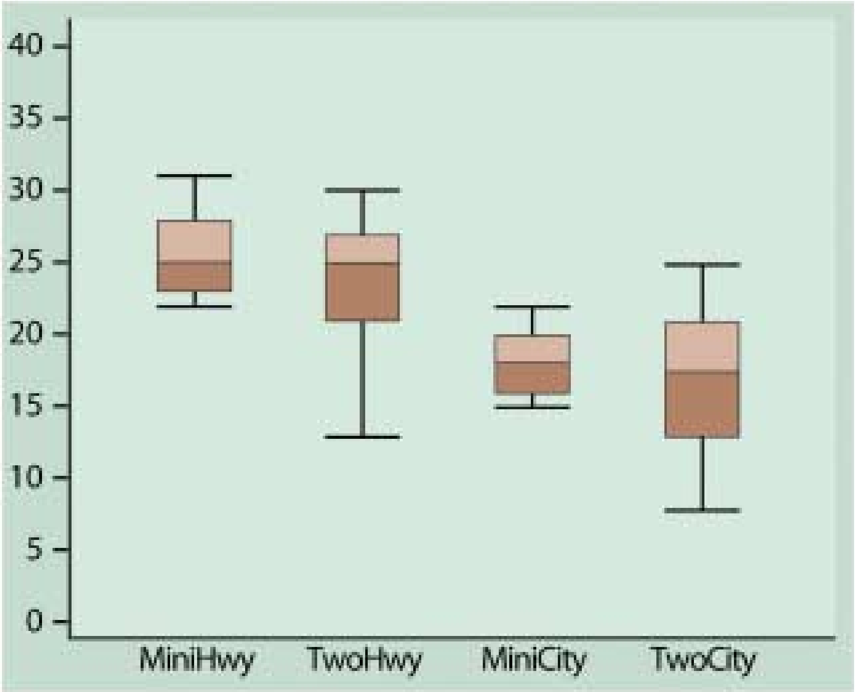


Figure 1.15 Boxplots of the highway and city gas mileages for cars classified as two-seaters and as minicompacts by the Environmental Protection Agency.

Outliers

Sometimes the boxplot is modified by plotting the values of quartiles using the formula below.

Values that lie **1.5 x IQR** units beyond the quartiles are classified as outliers, since they lie far from the central mass of the data. They deserve a closer look. For example for our data:

e.g. 1.5 x (28) = 42

Our new values are:

Q_1 126 - 42 = 84

Q_3 154 + 42 = 196

Values that are less than 84 or more than 196 would be considered outliers since they lie outside the central mass of the data. We see that 200 is an outlier using this rule. This high SSHA score deserves a closer look. Perhaps the data was typed incorrectly, or this student has a very successful approach to university that could be investigated, etc.

Standard Deviation and Variance

Standard deviation is a common measure of **spread** about the mean. **Variance** is defined as:

$$S^2 = \left(\frac{1}{(n-1)}\right)\left(\sum (X - \overline{X})^2\right)$$

The standard deviation (s) is the square root of the variance (s^2)

$$S = \sqrt{S^2}$$

SSHA example (continued), the mean is necessary for the calculation of the variance since it is the center of the distribution

$$\overline{X} = 141.06$$

We then sum and square each observation's deviation from the mean as given below. The standard deviation of 26.44 could now be compared to other research conducted on the SSHA measure.

$$S^2 = \frac{1}{17}\ ((101 - 141.06)^2 + \ldots + (200 - 141.06)^2$$

$$= (26.44)^2 \quad = \quad 699.07 \quad \text{THEN} \quad s = 26.44$$

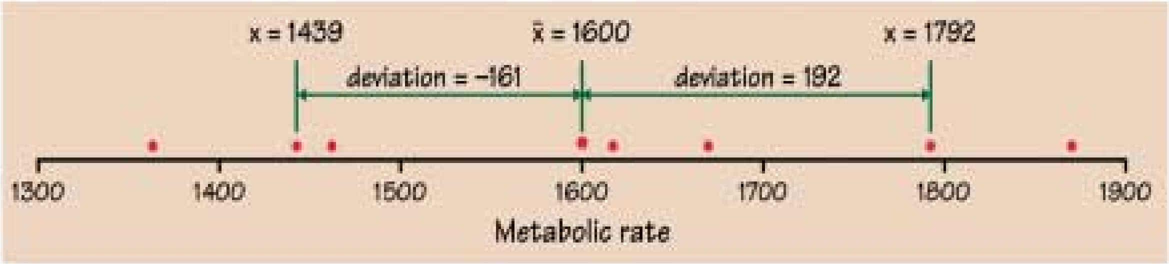


Figure 1.17 Metabolic rates for seven men, with the mean (1600) and the deviations of two observations from the mean. Deviations can be negative or positive but the variance will always be a positive quantity.

SPSS

Obtain the SPSS Manual from the website. Follow the instructions in the manual on how to enter data and run SPSS programs. It is not necessary to understand all of the SPSS output (you will learn this later).

It is more important to gain some experience with SPSS. Templates for running SPSS programs will be provided in each chapter as:

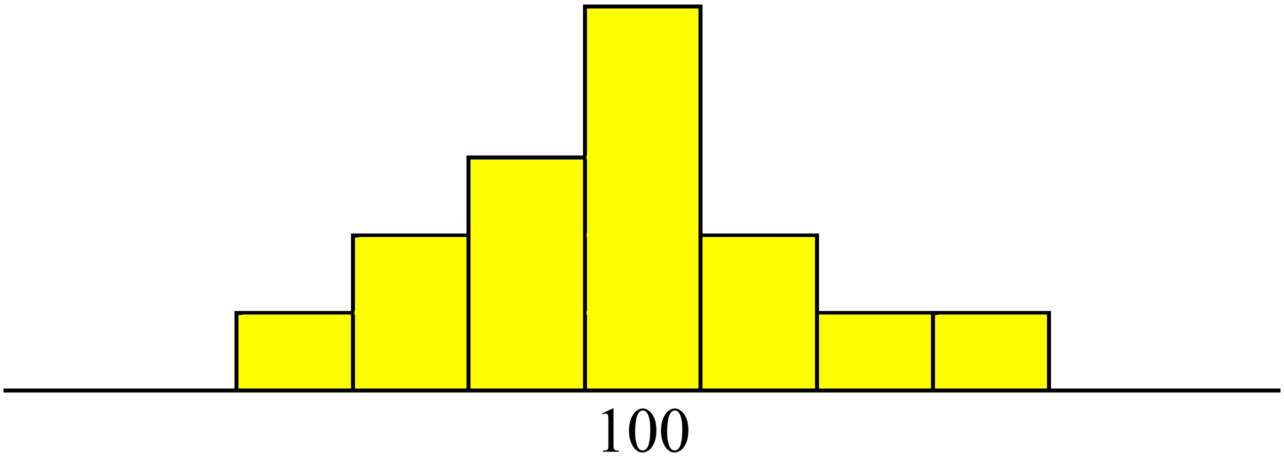
SPSS Example One

Click on the above SPSS example to find out how to calculate basic descriptive statistics.

View the video-Normal Distribution

The Normal Distribution

Consider the histogram (relative frequency) of IQ scores for adults below



The total area under the histogram is equal to 1. Look at the curve below. Area under the curve equals the relative frequency.

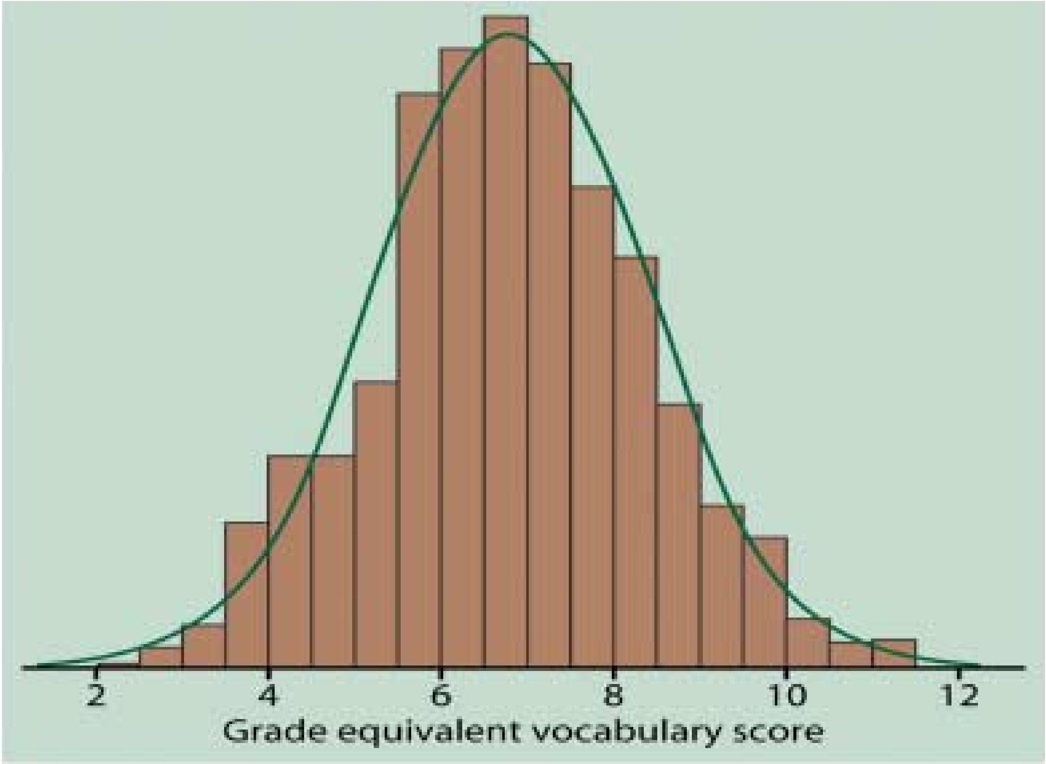


Figure 1.20 Histogram of the IOWA Test vocabulary scores for Gary, Indiana, seventh graders, showing the approximation of the distribution by a normal curve.

A density curve describes the distribution of the data. The most popular curve in all of science is the **normal density curve**. It can be defined in terms of its mean μ and standard deviation σ.

Given the mean and standard deviation of the normal density curve calculations concerning the data can be made easily.

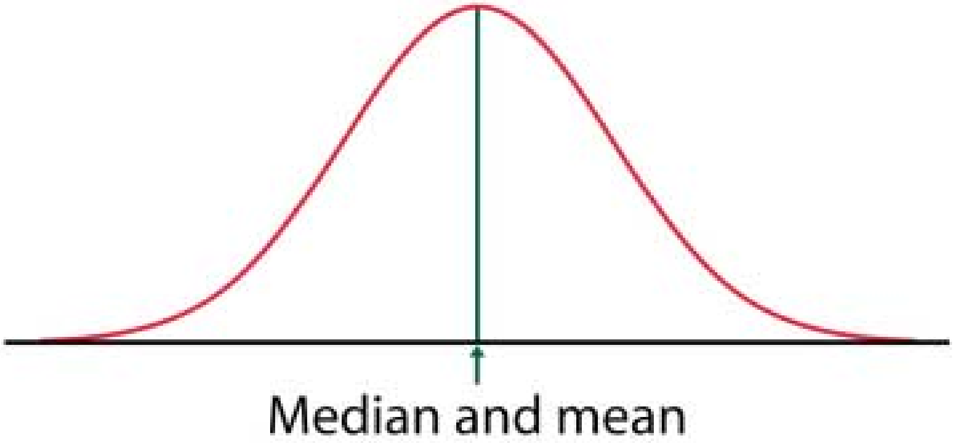


Figure 1.22a. A symmetric density curve with its mean and median marked.

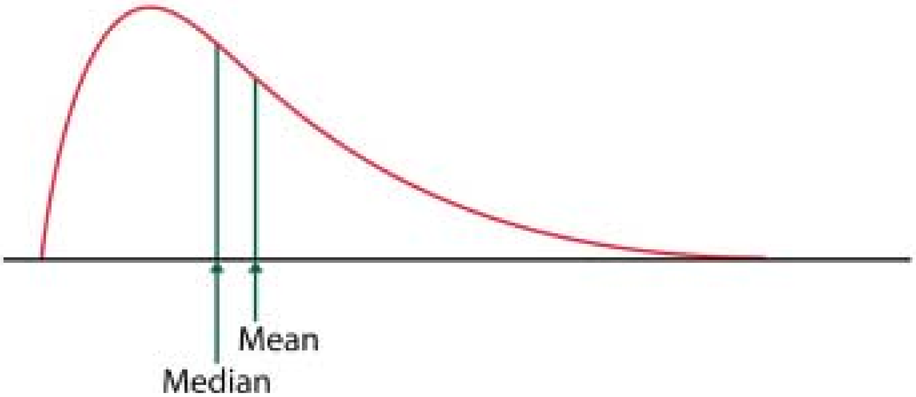


Figure 1.22b A right-skewed density curve with its mean and median marked.

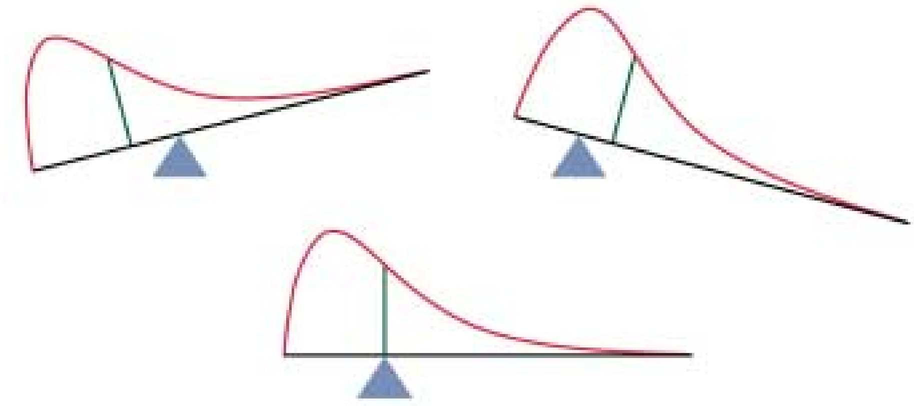


Figure 1.23 The mean of a density curve is the point at which it would balance.

Some characteristics of the normal density curve are that it is symmetric, single peaked, and bell-shaped. Note the mean is identical for the two normal distributions below but the standard deviation is larger for the first than the second.

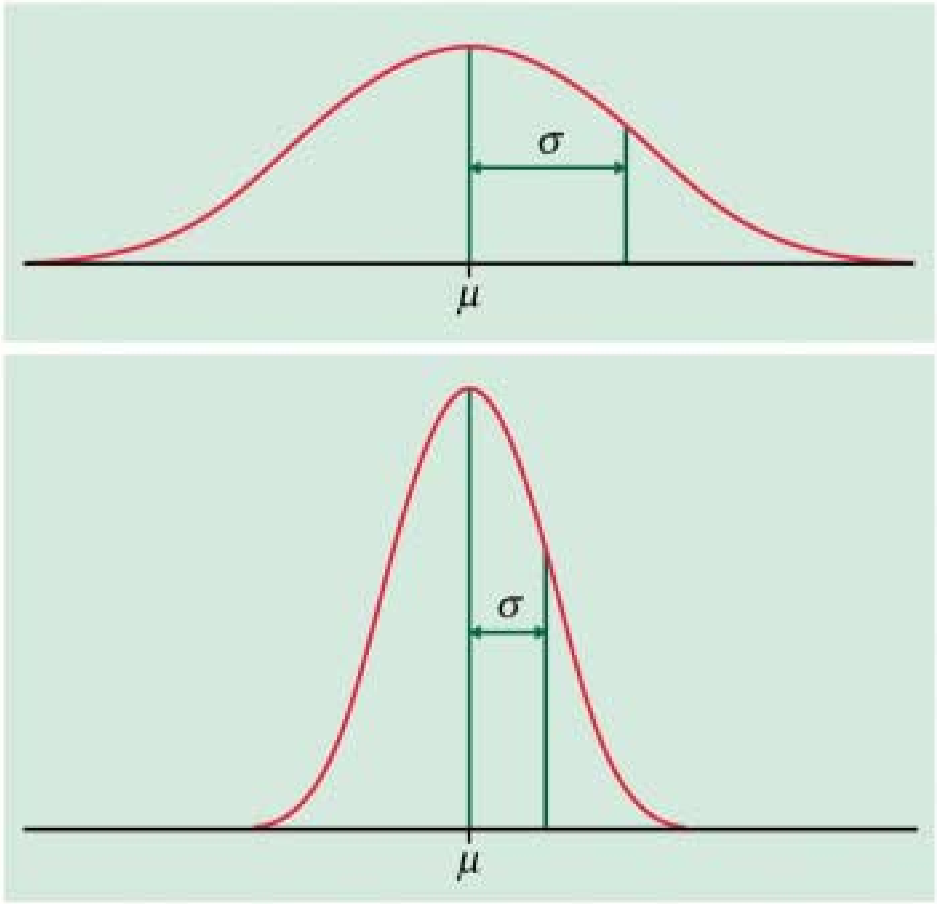
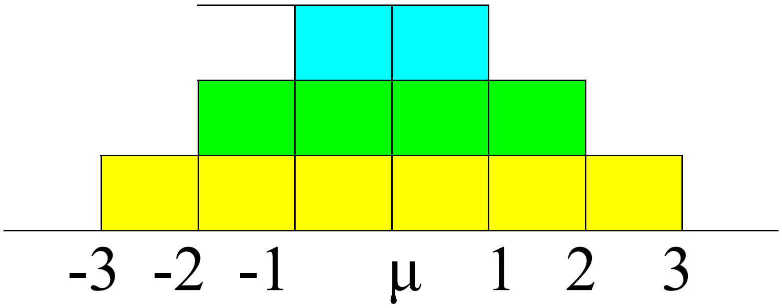


Figure 1.24 Two normal curves showing the mean μ and the standard deviation σ.

We conclude smaller standard deviations describe data that is more compact around the mean whereas larger standard deviations describe data that has larger ‘spread’ about the mean. For any normal distribution an important property is:

- **68% of observations lie within one σ of μ**
- **95% of observations lie within 2 σ of μ**
- **99.7% of observation lie within 3 σ of μ**



View the video-Normal Calculations

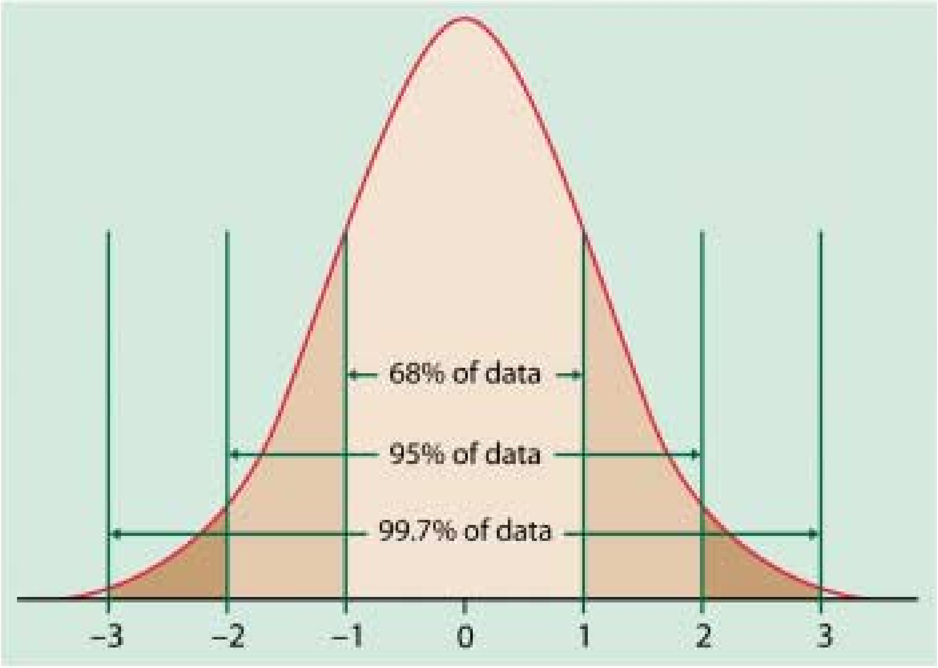


Figure 1.25 The 68-95-99.7 rule for normal distributions.

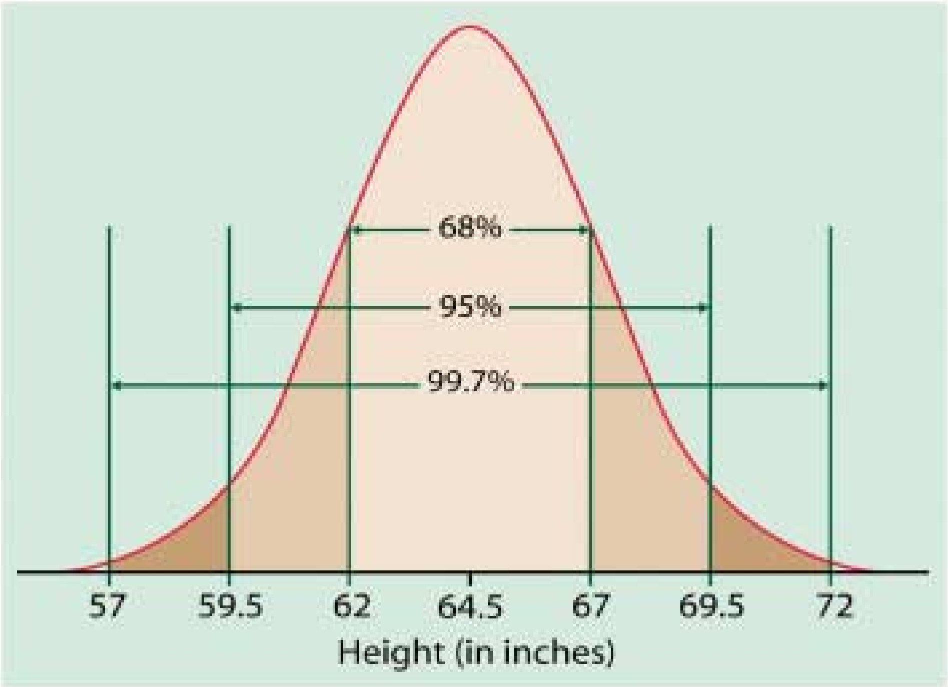


Figure 1.26 The 68-95-99.7 rule applied to the heights of young women.

Normal Calculations

If a variable, X, has a normal distribution with mean μ and standard deviation σ **(N (μ,σ))** then the standardized variable, Z defined as:

$$\mathbf{Z} = \frac{\mathbf{x - \mu}}{\boldsymbol{\sigma}}$$

"Z" has a normal distribution N(0,1) with mean 0 and standard deviation 1. We use this information to solve some important questions below using the Z tables in your text (see cover of text for a copy of Z table).

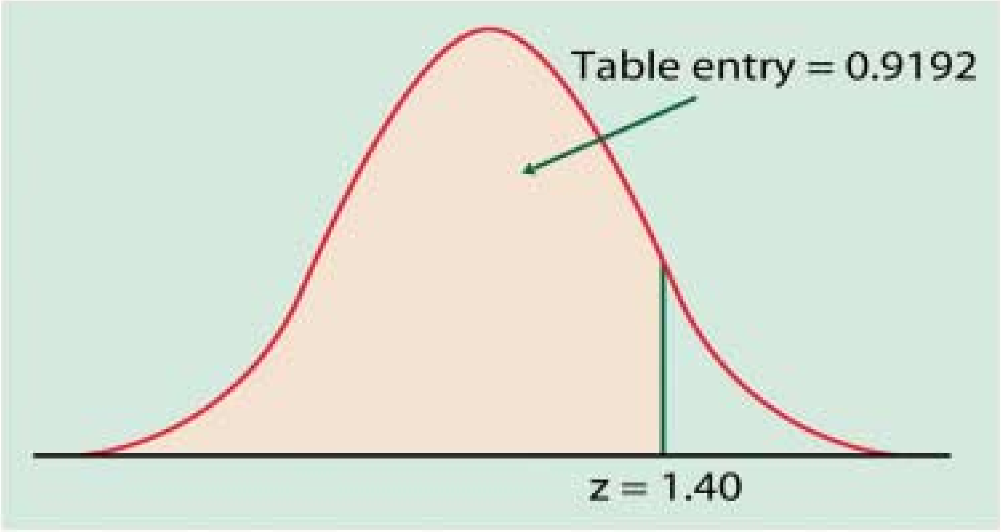


Figure 1.27a The area under a standard normal curve to the left of the point z = 1.40 is 0.9192. Table A gives areas under the standard normal curve.

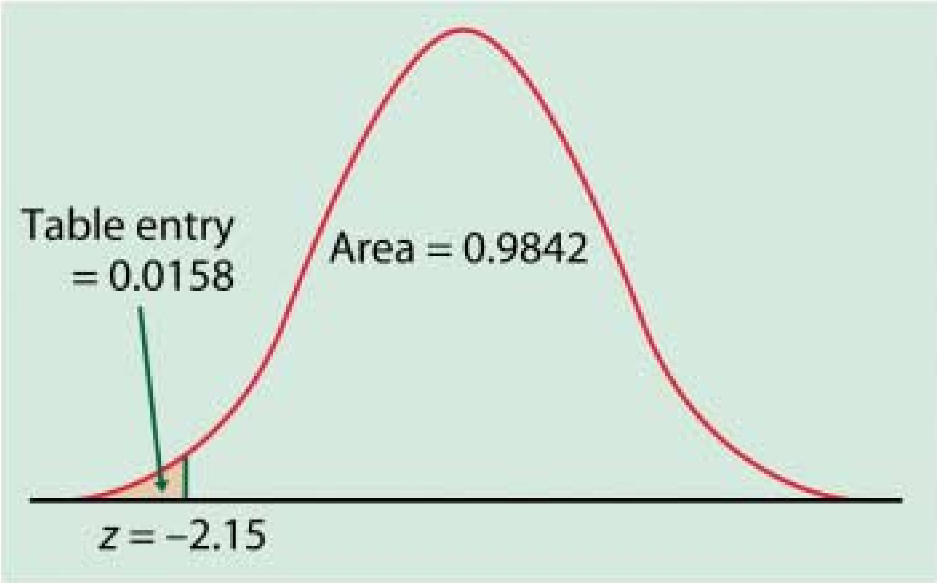


Figure 1.27b Areas under the standard normal curve to the right and left of z = -2.15. Table A gives areas to the left.

Below are several examples of the types of questions, that can be solved when we know the mean and standard deviation of the normal.

Example 1: The Law School admission test has a known normal distribution with μ=65.5 and σ=2.5 units.

The standardized score (Z) for an observation from this distribution is:

$$\mathbf{Z = \frac{\text{score} - 65.5}{2.5}}$$

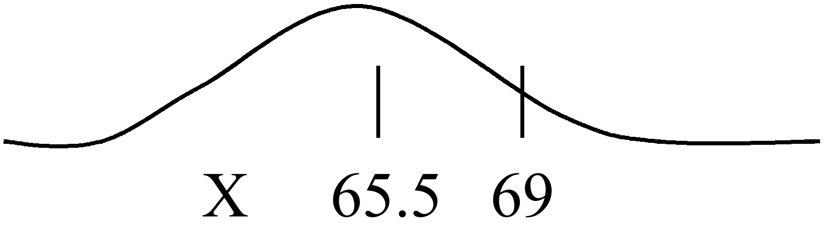
A score of 69 is standardized as a Z-score where

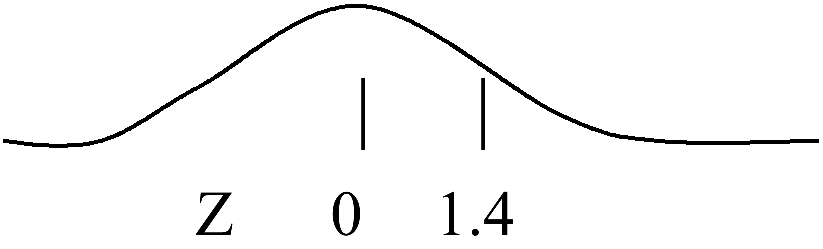
$$Z = \frac{\mathbf{69 - 65.5}}{\mathbf{2.5}} = 1.4$$

or 1.4 standard deviation above the mean.

Looking at the distributions below, we can relate the raw score of 69 to the Z = 1.4 in the distribution of Z below it.

A raw score (X) is 69 is a standardized (Z) score of 1.4

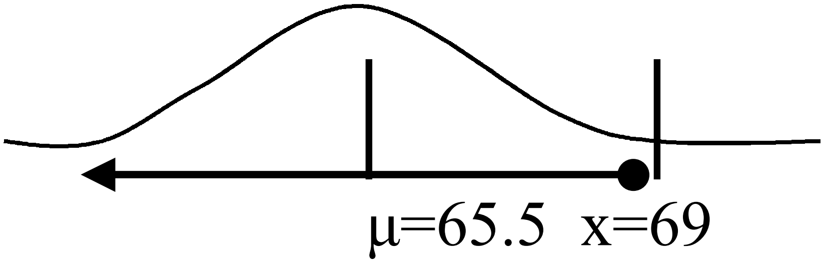


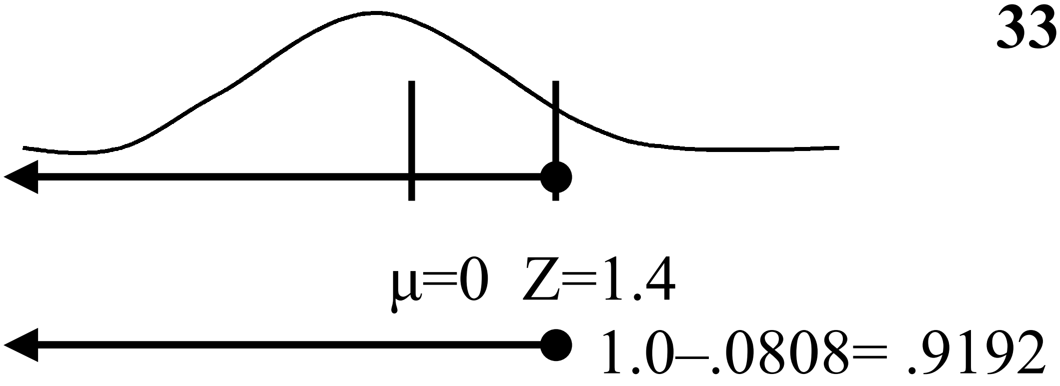


Note how the raw score of 69 above the mean of 65.5 is transformed into the Z score of 1.4 standard deviations above the mean of 0.

Question

What fraction/percentage of people who take the LSAT (Law School Admission Test) have scores less than 69? (i.e., P(X<69)=?)





Go to Z tables (inside cover of text book). Area to the left of 1.4 is .9192.

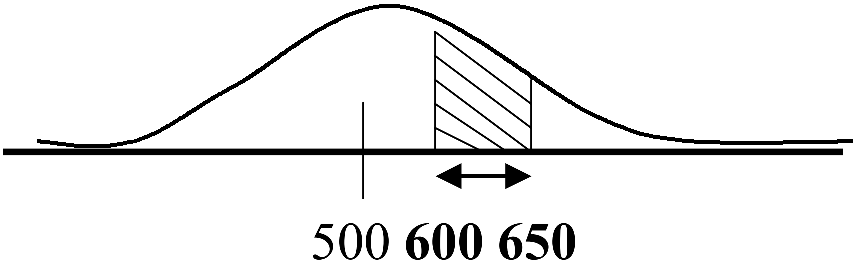
In other words, about 92% of the people score below 69. P(X<69)=.92. What percent of the people score above 69?

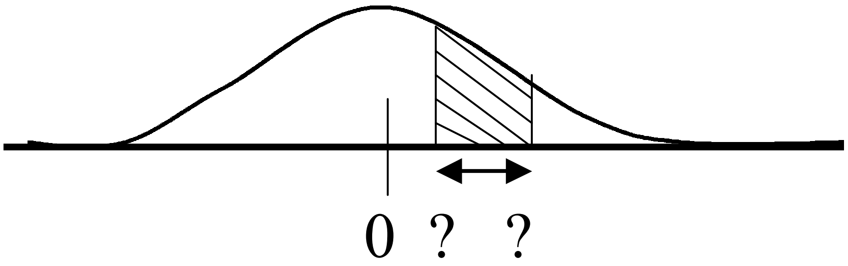
Recall the total area under the curve is 1.0. Therefore, P(X>69) = 1.0-.92 = .08 or 8%.

Remember **positive Z-scores indicate the raw data is above the mean** and **negative Z-scores indicate the data is below the mean**.

Example 2

The GRE test is normal with μ=500 and σ=100. What is the probability of randomly selecting a person with a score between X=600 and X=650 (i.e., P(600<X<650)?





What are the corresponding Z scores? Calculate the Z scores for each value.

For X = 600 $Z = \frac{X - \mu}{\sigma} = \frac{600-500}{100} = \frac{100}{100} = 1$

X = 650 $$Z = \frac{650 - 500}{100} = \frac{150}{100} = 1.5$$

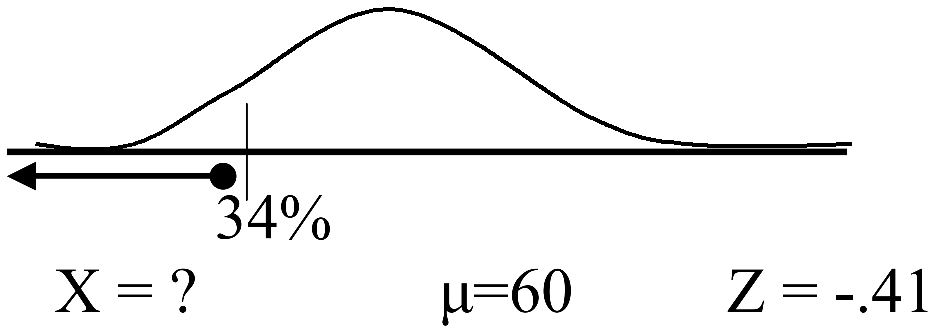
Go to the Z tables and look up Z = 1.0 (.8413) and Z = 1.5 (.9332).

Now subtract .9332 - .8413 = .0919 since we want the area between 600 and 650.

P(600<X<650) = .0919 or about 9%

Example 3

Medcat scores are normally distributed with μ=60 and σ=5. What is the 34th percentile?



First look up Z in the table for the 34^{th} percentile or .34. This Z is approximately -.41 (noted above). Substitute this in the equation below.

$$Z = \frac{x-\mu}{\sigma} = -.41 = \frac{x-60}{5}$$

multiply x5 $$5(-.41) = 5\left(\frac{x-60}{5}\right)$$

add +60 $$60+5(-.41) = X-60+60$$

Therefore $$\begin{aligned} X &= 60 + 5\,(-.41) \\ &= 60 - 2.05 \\ &= 57.95 \text{ is the 34th percentile} \end{aligned}$$

SPSS Example Two

Click above to see how normal calculations can be done on the computer without the use of tables.

Assessing Normality

Researchers routinely want to know if their data is normally distributed. A primary method for assessing normality is the normal probability plot. We use a computer to calculate this graphical display. The key feature to see in the plot (if your data is normal) is an approximately straight line. If the normal probability plot graphs the data points in a reasonably straight line, then your data is normal.

Briefly, the computer rank orders data from smallest to largest and calculates the corresponding percentile. (For data set of n = 10, the smallest is at the 10% point, second smallest 20%, etc).

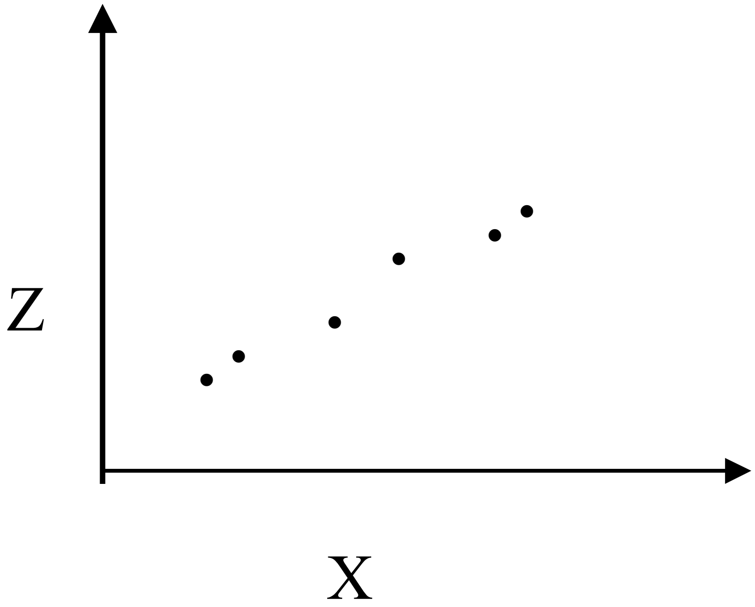
The computer then calculates the percentile z-scores. For example, the 10% point z = -1.28, etc. Plot the data on the y-axis against the corresponding z on the x-axis. If the data is close to normal you will produce a straight line with this plot. If the data is from the standard normal (z) then you will see a 45 degree line with y=z.

See your text for more examples of this plot. If the data is not normal, then the points will not be a straight line but some other curve.

See below an example of a plot of normal data and a plot of non-normal data.

Remember:
If the data is normal then points will fall close to a straight line. A computer is used to produce the plot.

A straight line indicates normal data.



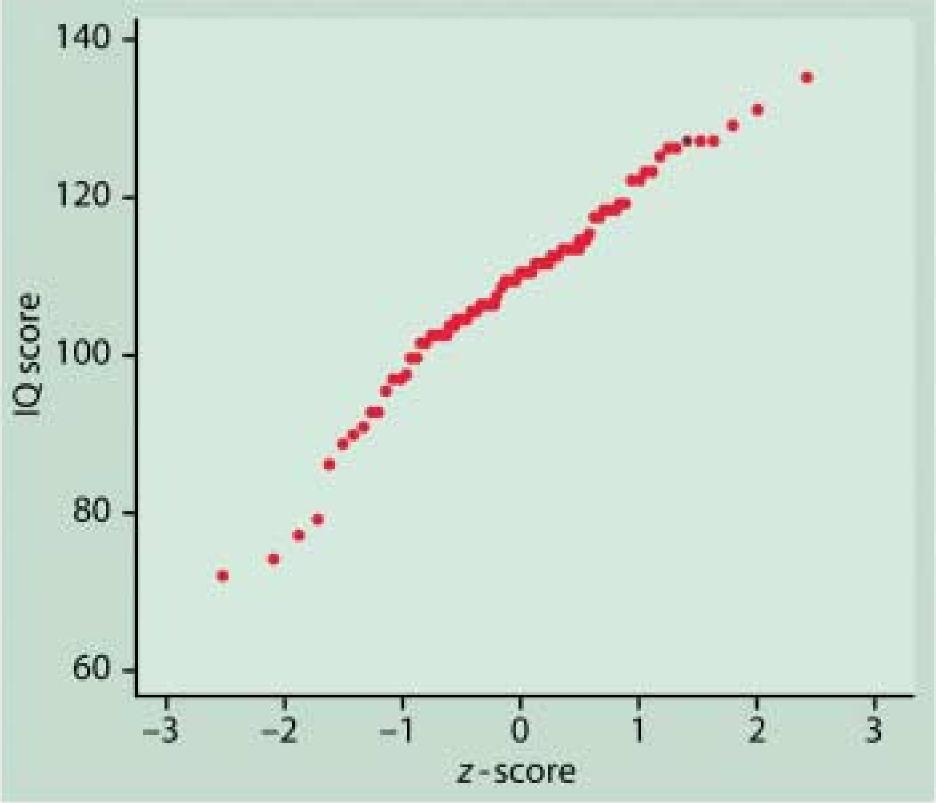
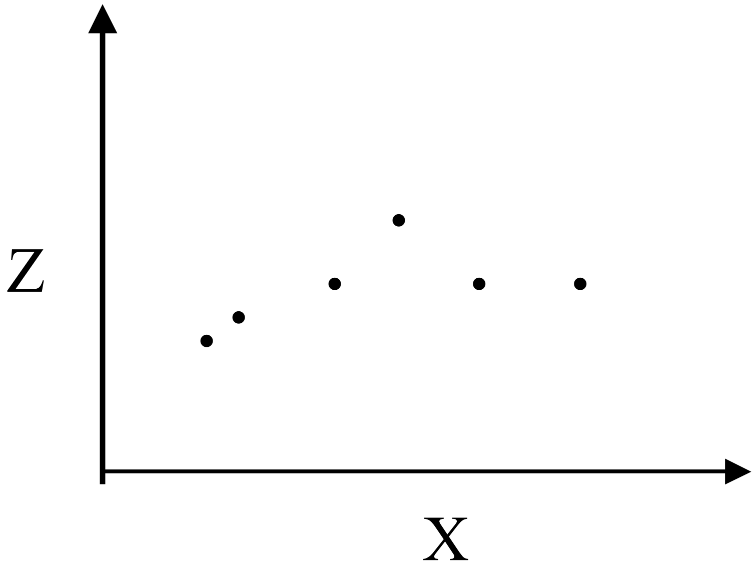


Figure 1.34 Normal quantile plot of the IQ scores of 78 seventh-grade students. The straight pattern shows that this distribution is close to normal.

No straight line: non-normal data.



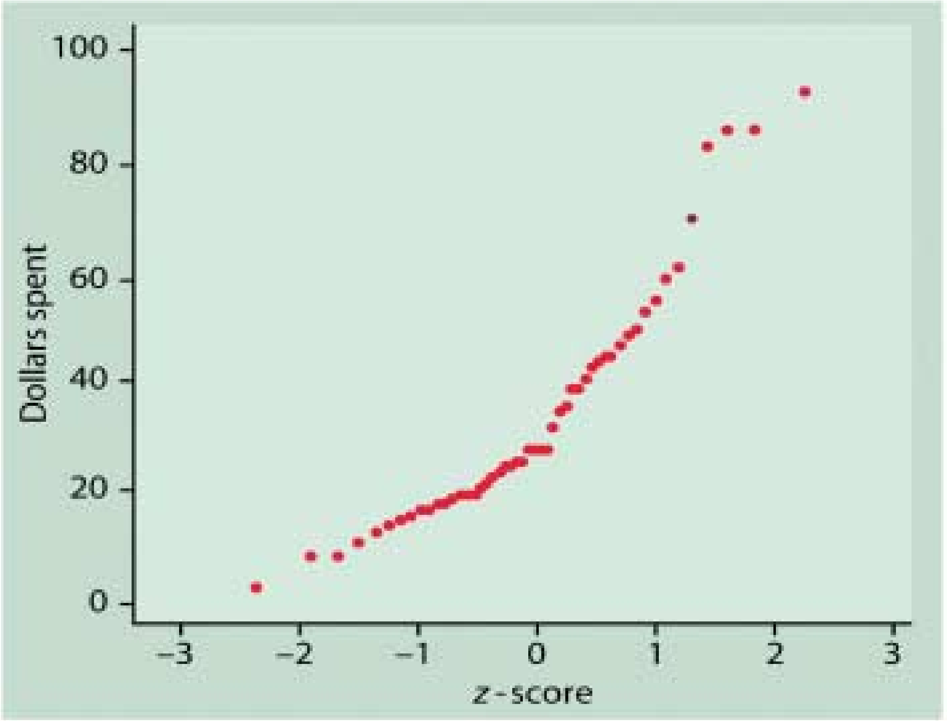


Figure 1.33 Normal quantile plots of the supermarket spending data. The pattern bends up at the right, showing right skewness.

<u>Example</u>:

- effect of calcium in diet on BP in text. One group of men was given a daily calcium supplement, while the control group was given a placebo. The seated systolic blood pressure of all the men was measured before the treatments began and again after 12 weeks. Here are the initial blood pressure readings for the two groups:

Calcium Group

107 110 123 129 112 111 107 112 136 102

Placebo Group

123 109 112 102 98 114 119 112 110 117 130

SPSS is used to calculate a normal probability plot for this example.

There are two groups, an experimental and a control group. The data for both groups appears non-normal <u>but</u> n is small in both examples. In small samples, points from the normal distribution do not always fall close to a line.

Small samples have greater variability than larger ones. One could also look at a histogram and stem plot of the data to assess normality.

SPSS Example Three

Click above to see how to use SPSS to calculate the normal probability plots for the above example. You have now completed the chapter one notes.

Given you have read the relevant textbook material and notes; you should try the assignment questions. If you want more practice questions, try the Exercises listed at the beginning of the chapter.

Remember to try the link to the virtual statistics lab and the Freeman website more help in mastering the material.

Chapter One Summary

The pattern of your data can usually be described by a **density curve**. Area under the curve gives the **relative frequency** for the distribution.

The **mean**, **median** and **quartiles** can be approximately located by eye on the curve. The **standard deviation** gives the spread of the data.

The normal distribution is the most popular distribution in science.

The mean and standard deviation completely define the normal distribution. To **standardize** any normal observation X we create a Z score.

All normal distributions satisfy the **68-95-99.7** rule. The normal **quantile** or **probability plot** is used to assess normality.

Remember: Tips for Success

1) Read the text.

2) Read the notes.

3) Try the assignments.

4) If needed, try the exercise questions.

5) Try the simulations and view the videos if you need more help with a concept

6) Try the self tests for practice on each chapter of the text at **www.whfreeman.com/ips**

Steady Work = SUCCESS !!!