Chapter 1:

Cases are the objects described by a set of data. Cases may be customers, companies, subjects in a study, units in an experiment, or other objects.

A **label** is a special variable used in some data sets to distinguish the different cases.

A variable is a characteristic of a case.

Different cases can have different values of the variables.

A **categorical variable** places a case into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

Categorical variables: Bar graphs and pie charts
distribution of a categorical variable
count percent proportion
The values of a categorical variable are labels for the categories, such as "yes" and "no." The distribution of a categorical variable lists the categories and gives either the count or the percent of cases that fall in each category. An alternative to the percent is the proportion, the count divided by the sum of the counts. Note that the percent is simply the proportion times 100.

You should always consider the best way to order the values of the categorical variable in a bar graph. Note that a bar graph using counts will look the same as a bar graph using percents. A pie chart naturally uses percents.

To make a pie chart, you must include all the categories that make up a whole. A category such as "Other" in this example can be used, but the sum of the percents for all the categories should be 100%. This constraint makes bar graphs more flexible.

Quantitative variables: Stemplots and histograms
A **stemplot** (also called a stem-and-leaf plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0. Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by the number system rather than by judgment.

Histograms do not have these limitations. A histogram breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should

choose classes of equal width. Making a histogram by hand requires more work than a stemplot. Histograms do not display the actual values observed. For these reasons, we prefer stemplots for small data sets.

Although **histograms** resemble **bar graphs**, their details and uses are distinct. A histogram shows the distribution of counts or percents among the values of a single variable. A bar graph compares the counts or percents of different items. The horizontal axis of a bar graph need not have any measurement scale but simply identifies the items being compared.

In any graph of data, look for the overall pattern and for striking deviations from that pattern. You can describe the overall pattern of a distribution by its shape, center, and spread. An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

For now, we can describe the center of a distribution by its **midpoint**, the value with roughly half the observations taking smaller values and half taking larger values.

We can describe the **spread** of a distribution by giving the smallest and largest values.

Stemplots and histograms display the shape of a distribution in the same way. Just imagine a stemplot turned on its side so that the larger values lie to the right.

Some things to look for in describing shape are:

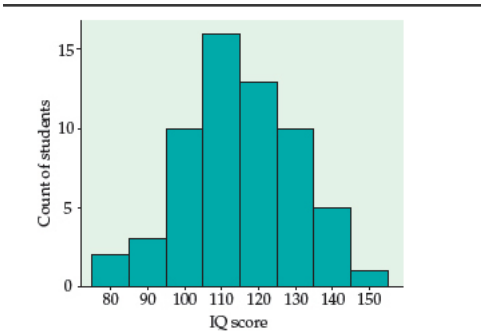
(1) modes unimodal

Does the distribution have one or several major peaks, called modes? A distribution with one major peak is called unimodal.

(2) symmetric skewed

Is it approximately symmetric or is it skewed in one direction? A distribution is symmetric if the pattern of values smaller and larger than its midpoint are mirror images of each other. It is skewed to the right if the right tail (larger values) is much longer than the left tail (smaller values).

Example:



Shape: The distribution is roughly symmetric with a single peak in the center. We don't expect real data to be perfectly symmetric, so in judging symmetry, we are satisfied if the two sides of the histogram are roughly similar in shape and extent.

Center: You can see from the histogram that the midpoint is not far from 110. Looking at the actual data shows that the midpoint is 114.

Spread: The histogram has a spread from 75 to 155. Looking at the actual data shows that the spread is from 81 to 145. There are no outliers or other strong deviations from the symmetric, unimodal pattern.

Example:

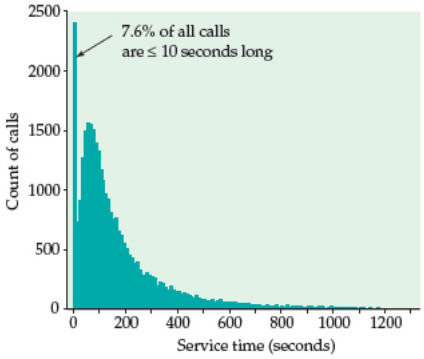


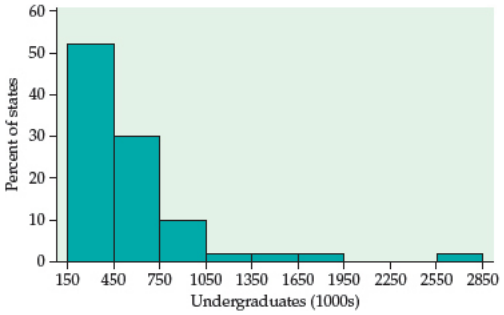
Figure 1.8

The distribution of call lengths in Figure 1.8, on the other hand, is strongly **skewed to the right. (Asmy a droite)** The midpoint, the length of a typical call, is about 115 seconds, or just under 2 minutes. The spread is very large, from 1 second to 28,739 seconds.

The longest few calls are outliers. They stand apart from the long right tail of the distribution, though we can't see this from Figure 1.8, which omits the largest observations. The longest call lasted almost 8 hours—that may well be due to equipment failure rather than an actual customer call.

Dealing with outliers

College students. How does the number of undergraduate college students vary by state? Figure 1.9 is a histogram of the numbers of undergraduate students in each of the states.6 Notice that more than 50% of the states are included in the first bar of the histogram. These states have fewer than 300,000 undergraduates. The next bar includes another 30% of the states. These have between 300,000 and 600,000 students. The bar at the far right of the histogram corresponds to the state of California, which has 2,685,893 undergraduates. California certainly stands apart from the other states for this variable. It is an outlier.



College students per 1000.

To account for the fact that there is large variation in the populations of the states, for each state we divide the number of undergraduate students by the population and then multiply by 1000. This gives the undergraduate college enrollment expressed as the number of students per 1000 people in each state. Figure 1.10 gives a stemplot of the distribution. California has 60 undergraduate students per 1000 people. This is one of the higher values in the distribution, but it is clearly not an outlier.

```
3 | 8
4 | 1 1 1 1 1 3 3
4 | 5 5 5 5 6 6 6 7 7 7 7 7 8 8 8 8 9
5 | 0 0 1 1 1 2 2 2 4 4 4 4
5 | 5 6 6
6 | 0 0 0 0 1
6 | 7 9
7 | 1 2
7 | 7
```

If you are interested in marketing a product to undergraduate students, the unadjusted numbers would be of interest because you want to reach the most people. On the other hand, if you are interested in comparing states with respect to how well they provide opportunities for higher education to their residents, the population-adjusted values would be more suitable. Always think about why you are doing a statistical analysis, and this will guide you in choosing an appropriate analytic strategy.

TIME PLOT

Whenever data are collected over time, it is a good idea to plot the observations in time order. Displays of the distribution of a variable that ignore time order, such as stemplots and histograms, can be misleading when there is systematic change over time.

A time plot of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale.

THE MEAN

Numerical description of a distribution begins with a measure of its center or average. The two common measures of center are the mean and the median. The mean is the "average value" and the median is the "middle value." These are two different ideas for "center," and the two measures behave differently. We need precise recipes for the mean and the median.

To find the **mean** $\bar{x}$ of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, . . . , x_n, their mean is

$$\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n}\Sigma x_i$$

The mean is sensitive to the influence of a few extreme observations. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a **resistant measure of center**.

A measure that is resistant does more than limit the influence of outliers. Its value does not respond strongly to changes in a few observations, no matter how large those changes may be. The mean fails this requirement because we can make the mean as large as we wish by making a large enough increase in just one observation. A resistant measure is sometimes called a **robust measure**.

THE MEDIAN

THE MEDIAN *M*

The **median** *M* is the midpoint of a distribution. Half the observations are smaller than the median and the other half are larger than the median. Here is a rule for finding the median:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations *n* is odd, the median *M* is the center observation in the ordered list. Find the location of the median by counting (n + 1)/2 observations up from the bottom of the list.
3. If the number of observations *n* is even, the median *M* is the mean of the two center observations in the ordered list. The location of the median is again (*n* + 1)/2 from the bottom of the list.

The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median.

Measuring spread: The quartiles

We are interested in the spread or variability of incomes and drug potencies as well as their centers. The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.

QUARTILE

We can describe the spread or variability of a distribution by giving several percentiles. The median divides the data in two; half of the observations are above the median and half are below the median. We could call the median the **50th percentile**. The upper quartile is the median of the upper half of the data. Similarly, **the lower quartile is the median of the lower half of the data**. **With the median, the quartiles divide the data into four equal parts; 25% of the data are in each part.**

PERCENTILE

We can do a similar calculation for any percent. The **p^{th} percentile** of a distribution is the value that has p percent of the observations fall at or below it. To calculate a percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list.

Our definition of percentiles is a bit inexact because there is not always a value with exactly p percent of the data at or below it. We will be content to take the nearest observation for most percentiles, but the quartiles are important enough to require an

exact rule.

THE QUARTILES Q_1 AND Q_3

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile** Q_1 is the median of the observations whose positions in the ordered list are to the left of the location of the overall median.
3. The **third quartile** Q_3 is the median of the observations whose positions in the ordered list are to the right of the location of the overall median.

THE FIVE-NUMBER SUMMARY

The **five-number summary** of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

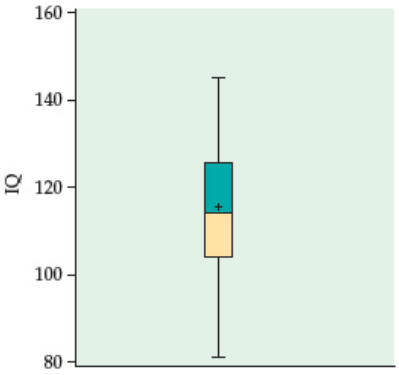
The five-number summary leads to another visual representation of a distribution, the boxplot.

BOXPLOT

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles Q_1 and Q_3.
- A line in the box marks the median M.
- Lines extend from the box out to the smallest and largest observations.

Example:



IQ scores. In Example 1.14 (page 14), we used a histogram to examine the distribution of a sample of 60 IQ scores. A boxplot for these data is given in Figure 1.14. Note that the mean is marked with a "+" and appears very close to the median. The two quartiles are each approximately the same distance from the median, and the two whiskers are approximately the same distance from the corresponding quartiles. All these characteristics are consistent with a symmetric distribution, as illustrated by the histogram in Figure 1.7.

The 1.5 × *IQR* rule for suspected outliers

If we look at the data in Table 1.2 (page 17), we can spot a clear outlier, a call lasting 2631 seconds, more than twice the length of any other call. How can we describe the spread of this distribution? The smallest and largest observations are extremes that do not describe the spread of the majority of the data. The distance between the quartiles (the range of the center half of the data) is a more resistant measure of spread than the range. This distance is called the *interquartile range*.

THE INTERQUARTILE RANGE *IQR*

The **interquartile range *IQR*** is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$

The quartiles and the IQR are not affected by changes in either tail of the distribution. They are resistant, therefore, because changes in a few data points have no further effect once these points move outside the quartiles.

However, no single numerical measure of spread, such as IQR, is very useful for describing skewed distributions. The two sides of a skewed distribution have different spreads, so one number can't summarize them. We can often detect skewness from

the five-number summary by comparing how far the first quartile and the minimum are from the median (left tail) with how far the third quartile and the maximum are from the median (right tail). The interquartile range is mainly used as the basis for a rule of thumb for identifying suspected outliers.

THE 1.5 × *IQR* RULE FOR OUTLIERS

Call an observation a suspected outlier if it falls more than 1.5 × *IQR* above the third quartile or below the first quartile.

MODIFIED BOXPLOT

Two variations on the basic boxplot can be very useful. The first, called a modified boxplot, uses the 1.5 × IQR rule. The lines that extend out from the quartiles are terminated in whiskers that are 1.5 × IQR in length. Points beyond the whiskers are plotted individually and are classified as outliers according to the 1.5 × IQR rule.

SIDE-BY-SIDE BOXPLOTS

The other variation is to use two or more boxplots in the same graph to compare groups measured on the same variable. These are called side-by-side boxplots. The following example illustrates these two variations.

Measuring spread: The standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the standard deviation to measure spread, or variability. The standard deviation measures spread by looking at how far the observations are from their mean.

THE STANDARD DEVIATION *s*

The variance s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$$

or, in more compact notation,

$$s^2 = \frac{1}{n-1}-\Sigma(x_i-\bar{x})^2$$

The **standard deviation** s is the square root of the variance s^2:

$$s = \sqrt{\frac{1}{n-1}\Sigma(x_i-\bar{x})^2}$$

PROPERTIES OF THE STANDARD DEVIATION

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread.* This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s, like the mean $\bar{x}$, is not resistant. A few outliers can make s very large.

CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and s for reasonably symmetric distributions that are free of outliers.

LINEAR TRANSFORMATIONS

A linear transformation changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{new} = a + bx$$

Adding the constant a shifts all values of x upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant b changes the size of the unit of measurement.

EFFECT OF A LINEAR TRANSFORMATION

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by b.
- Adding the same number a (either positive or negative) to each observation adds a to measures of center and to quartiles and other percentiles but does not change measures of spread.

DENSITY CURVE

A density curve is a curve that

- Is always on or above the horizontal axis.
- Has 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 proportion of all observations that fall in that range.

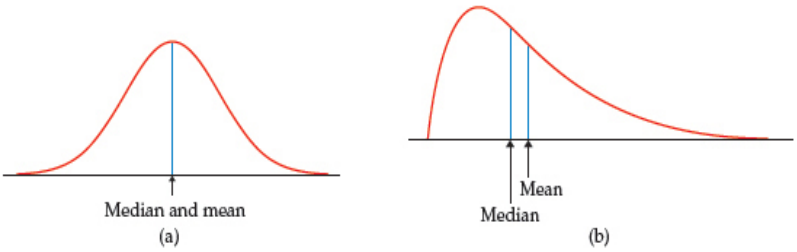


FIGURE 1.22 (a) A symmetric Normal density curve with its mean and median marked. (b) a right-skewed density curve with its mean and median marked.

The **median** of a density curve is the equal-areas point, the point that divides the area under the curve in half.

mean μ standard deviation σ

The **mean** of a density curve is the balance point, at which the curve would balance if made of solid material.

The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

NORMAL DISTRIBUTIONS

Normal curves

Normal distributions

These density curves are symmetric, unimodal, and bell-shaped. They are called Normal curves, and they describe Normal distributions. All Normal distributions have the same overall shape.

The exact density curve for a particular Normal distribution is specified by giving the distribution's **mean μ** and its **standard deviation σ**. **The mean is located at the center of the symmetric curve and is the same as the median.** Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its spread.

The standard deviation s controls the spread of a Normal curve. Figure 1.24 shows two Normal curves with different values of σ. The curve with the larger standard deviation is more spread out.

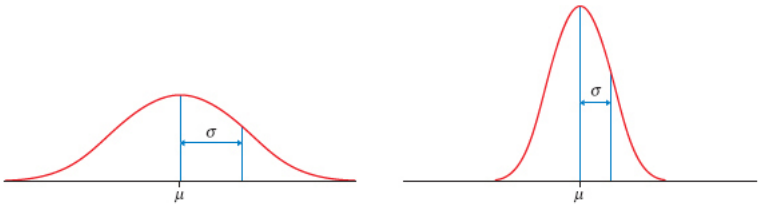


FIGURE 1.24 Two Normal curves, showing the mean μ and the standard deviation σ.

The standard deviation σ is the natural measure of spread for Normal distributions. Not only do μ and σ completely determine the shape of a Normal curve, but we can locate σ by eye on the curve. Here's how. As we move out in either direction from the center μ, the curve changes from falling ever more steeply.

There are other symmetric bell-shaped density curves that are not Normal. The Normal density curves are specified by a particular equation. The height of the density curve at any point x is given by

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

THE 68–95–99.7 RULE

In the Normal distribution with mean μ and standard deviation σ:

- Approximately **68**% of the observations fall within σ of the mean μ.
- Approximately **95**% of the observations fall within 2σ of μ.
- Approximately **99.7**% of the observations fall within 3σ of μ.

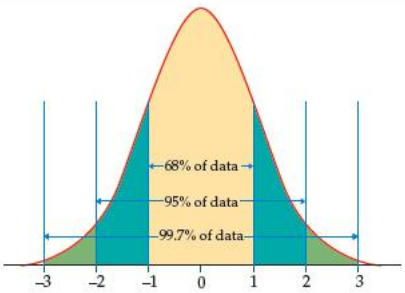
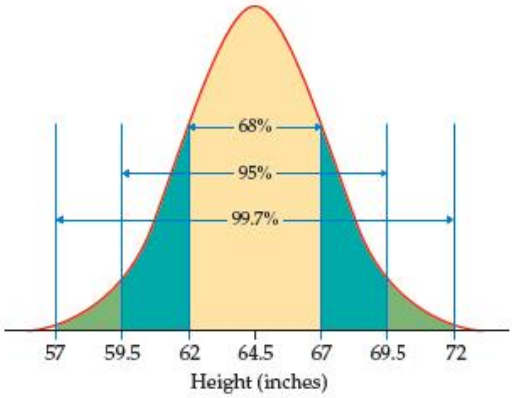


FIGURE 1.25 The 68–95–99.7 rule for Normal distributions.

$N(\mu, \sigma)$

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$. For example, the distribution of young women's heights is $N(64.5, 2.5)$.



STANDARDIZING OBSERVATIONS

As the 68–95–99.7 rule suggests, all Normal distributions share many properties. In fact, all Normal distributions are the same if we measure in units of size σ about the mean μ as center. Changing to these units is called standardizing. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation.

STANDARDIZING AND z-SCORES

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

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

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

A z-score tells us how many standard deviations the original observation falls away from the mean, and in which direction. **Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.**

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 $N(\mu, \sigma)$ with mean μ and standard deviation σ, then the standardized variable

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

has the standard Normal distribution.

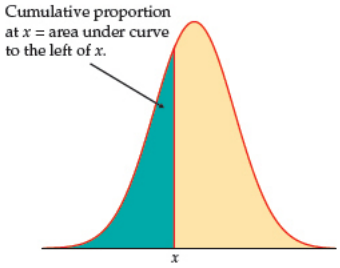


FIGURE 1.27 The cumulative proportion for a value x is the proportion of all observations from the distribution that are less than or equal to x. This is the area to the left of x under the Normal curve.

Normal distribution calculations

cumulative proportion

Areas under a Normal curve represent proportions of observations from that Normal distribution. There is no formula for areas under a Normal curve. Calculations use either software that calculates areas or a table of areas. The table and most software calculate one kind of area: **cumulative proportions**. A cumulative proportion is the proportion of observations in a distribution that lie at or below a given value. When the distribution is given by a density curve, the cumulative proportion is the area under the curve to the left of a given value. Figure 1.27 shows the idea more clearly than words do.

The key to calculating Normal proportions is to match the area you want with areas that represent cumulative proportions. Then get areas for cumulative proportions either from software or (with an extra step) from a table. The following examples show the method in pictures.

Example:
We assume that the distribution of the combined Critical Reading and Mathematics scores is approximately Normal with **mean 1010** and **standard deviation 225**.

Eligibility for aid and practice. What proportion of all students who take the SAT would be eligible to receive athletic scholarships and to practice with the team but would not be eligible to compete in the eyes of the NCAA? That is, what proportion of students have SAT scores between 620 and 800? First, sketch the areas, exactly as in Example 1.41. We again use X as shorthand for an SAT score.

1. *Standardize.*

$$\begin{array}{rcccl} 620 & \le & X & < & 800 \\ \frac{620-1010}{225} & \le & \frac{X-1010}{225} & < & \frac{800-1010}{225} \\ -1.73 & \le & Z & < & -0.93 \end{array}$$

2. *Use the table.*

$$\begin{aligned} \text{area between } -1.73 \text{ and } -0.93 &= (\text{area left of } -0.93) - (\text{area left of } -1.73) \\ &= 0.1762 - 0.0418 = 0.1344 \end{aligned}$$

As in Example 1.41, about 13% of students would be eligible to receive athletic scholarships and to practice with the team.

Sometimes we encounter a value of z more extreme than those appearing in Table A. For example, the area to the left of $z = -4$ is not given in the table. The z-values in Table A leave only area 0.0002 in each tail unaccounted for. For practical purposes, we can act as if there is zero area outside the range of Table A.

TABLE A

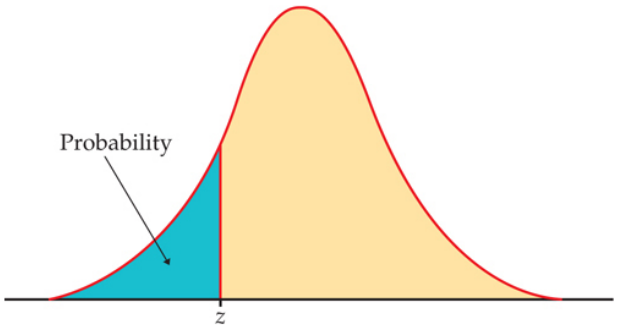


Table entry for z is the area under the standard Normal curve to the left of z.

TABLE A Standard Normal Probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776

−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

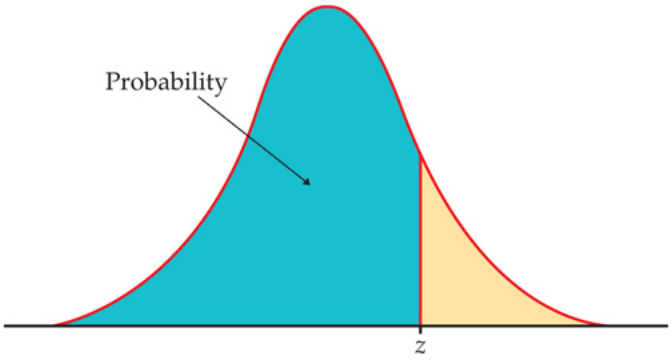


Table entry for z is the area under the standard Normal curve to the left of z.

TABLE A **Standard Normal Probabilities (continued)**

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Example:

How high for the top 10%? Scores for college-bound students on the SAT Critical Reading test in recent years follow approximately the $N(500, 120)$ distribution.[33] How high must a student score to place in the top 10% of all students taking the SAT?

Again, the key to the problem is to draw a picture. Figure 1.29 shows that we want the score x with an area of 0.10 above it. That's the same as area below x equal to 0.90.

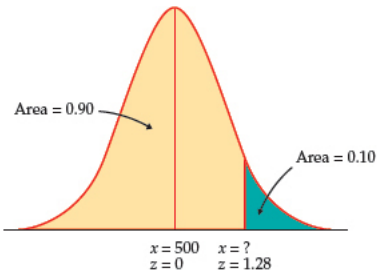


FIGURE 1.29 Locating the point on a Normal curve with area 0.10 to its right, Example 1.45.

Statistical software has a function that will give you the x for any cumulative proportion you specify. The function often has a name such as "inverse cumulative probability." Plug in mean 500, standard deviation 120, and cumulative proportion 0.9. The software tells you that $x = 653.786$. We see that a student must score at least 654 to place in the highest 10%.

Without software, first find the standard score z with cumulative proportion 0.9, then "unstandardize" to find x. Here is the two-step process:

1. *Use the table.* Look in the body of Table A for the entry closest to 0.9. It is 0.8997. This is the entry corresponding to $z = 1.28$. So $z = 1.28$ is the standardized value with area 0.9 to its left.
2. *Unstandardize* to transform the solution from z back to the original x scale. We know that the standardized value of the unknown x is $z = 1.28$. So x itself satisfies

$$\frac{x-500}{120} = 1.28$$

Solving this equation for x gives

$$x = 500 + (1.28)(120) = 653.6$$

This equation should make sense: it finds the x that lies 1.28 standard deviations above the mean on this particular Normal curve. That is the "unstandardized" meaning of $z = 1.28$. The general rule for unstandardizing a z-score is

$$x = \mu + z\sigma$$

Normal quantile plot

A histogram or stemplot can reveal distinctly non-Normal features of a distribution, such as outliers, pronounced skewness, or gaps and clusters. If the stemplot or histogram appears roughly symmetric and unimodal, however, we need a more sensitive way to judge the adequacy of a Normal model. The most useful tool for assessing Normality is another graph, the **Normal quantile plot**.

Here is the basic idea of a Normal quantile plot. The graphs produced by software use more sophisticated versions of this idea. It is not practical to make Normal quantile plots by hand.

1. Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies. For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.

Normal scores

2. Do Normal distribution calculations to find the values of z corresponding to these same percentiles. For example, $z = -1.645$ is the 5% point of the standard Normal distribution, and $z = -1.282$ is the 10% point. We call these values of Z **Normal scores**.
3. Plot each data point x against the corresponding Normal score. If the data distribution is close to any Normal distribution, the plotted points will lie close to a straight line.

Any Normal distribution produces a straight line on the plot because standardizing turns any Normal distribution into a standard Normal distribution. Standardizing is a linear transformation that can change the slope and intercept of the line in our plot but cannot turn a line into a curved pattern.

USE OF NORMAL QUANTILE PLOTS

If the points on a **Normal quantile plot** lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot. An optional line can be drawn on the plot that corresponds to the Normal distribution with mean equal to the mean of the data and standard deviation equal to the standard deviation of the data.

Figures 1.30 and 1.31 are Normal quantile plots for data we have met earlier. The data x are plotted vertically against the corresponding standard Normal z-score plotted horizontally. The z-score scale generally extends from -3 to 3 because almost all of a standard Normal curve lies between these values. These figures show how Normal quantile plots behave.

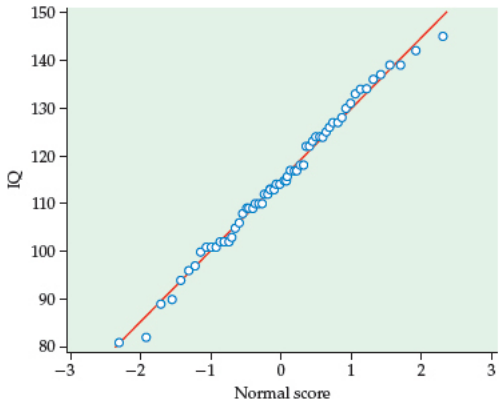


FIGURE 1.30 Normal quantile plot of IQ scores, Example 1.46. This distribution is approximately Normal.