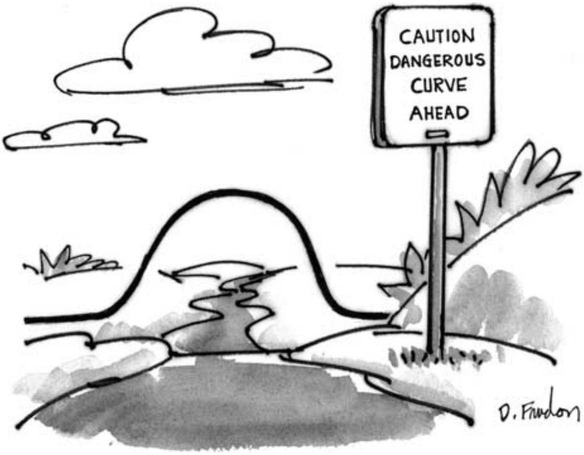
3

The Data Normalization

1. THE NORMAL CURVE

The normal curve was discovered around 1720 by Abraham de Moivre, while he was developing the mathematics of chance. (His work will be discussed again in parts IV and V.) Around 1870, the Belgian mathematician Adolph Quetelet had the idea of using the curve as an ideal histogram, to which histograms for data could be compared.

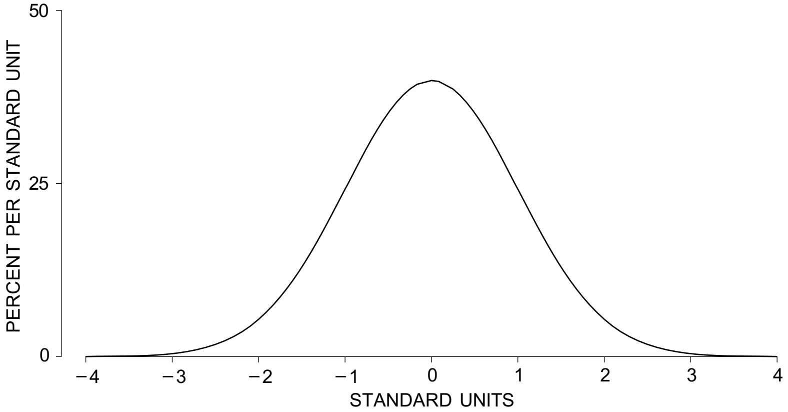


The normal curve has a formidable-looking equation:

$$y = \frac{100\%}{\sqrt{2\pi}} e^{-x^2/2}, \text{ where } e = 2.71828 \ldots .$$

This equation involves three of the most famous numbers in the history of mathematics: √2, π, and e. This is just to show off a little. You will find it is easy to work with the normal curve through diagrams and tables, without ever using the equation. A graph of the curve is shown in figure 1.

Figure 1. The normal curve.



Several features of this graph will be important. First, the graph is symmetric about 0: the part of the curve to the right of 0 is a mirror image of the part to the left. Next, the total area under the curve equals 100%. (Areas come out in percent, because the vertical axis uses the density scale.) Finally, the curve is always above the horizontal axis. It appears to stop between 3 and 4, but that's only because the curve gets so low there. Only about 6/100,000 of the area is outside the interval from –4 to 4.

It will be helpful to find areas under the normal curve between specified values. For instance,

- the area under the normal curve between −1 and +1 is about 68%;
- the area under the normal curve between −2 and +2 is about 95%;
- the area under the normal curve between −3 and +3 is about 99.7%.

Finding these areas is a matter of looking things up in a table, or pushing a button on the right kind of calculator; the table will be explained in section 2.

Many histograms for data are similar in shape to the normal curve, provided they are drawn to the same scale. Making the horizontal scales match up involves *standard units*.[1]

> A value is converted to standard units by seeing how many SDs it is above or below the average.

Values above the average are given a plus sign; values below the average get a minus sign. The horizontal axis of figure 1 is in standard units.

For instance, take the women age 18 and over in the HANES5 sample. Their average height was 63.5 inches; the SD was 3 inches. One of these women was 69.5 inches tall. What was her height in standard units? Our subject was 6 inches taller than average, and 6 inches is 2 SDs. In standard units, her height was +2.

Example 1. For women age 18 and over in the HANES5 sample—

(a) Convert the following to standard units:
(i) 66.5 inches (ii) 57.5 inches (iii) 64 inches (iv) 63.5 inches

(b) Find the height which is −1.2 in standard units.

Solution. Part (a). For (i), 66.5 inches is 3 inches above the average. That is 1 SD above the average. In standard units, 66.5 inches is 1. Fo+r (ii), 57.5 inches is 6 inches below the average. That is 2 SDs below average. In standard units, 57.5 inches is 2. For− (iii), 64 inches is 0.5 inches above average. That is 0.5/3 0.17 SD≈s. The answer is 0.17. For (iv), 63.5 inches is the average. So, 63.5 inches is 0 SDs away from average. The answer is 0. (Reminder: "≈" means "nearly equal.")

Part (b). The height is 1.2 SDs below the average, and 1.2 × 3 inches = 3.6 inches. The height is

$$63.5 \text{ inches} - 3.6 \text{ inches} = 59.9 \text{ inches}.$$

That is the answer.

Standard units are used in figure 2. In this figure, the histogram for the heights of the women age 18 and over in the HANES5 sample is compared to the normal curve. The horizontal axis for the histogram is in inches; the horizon- tal axis for the normal curve is in standard units. The two match up as indicated in example 1. For instance, 66.5 inches is directly above +1, and 57.5 inches is directly above –2.

There are also two vertical axes in figure 2. The histogram is drawn relative to the inside one, in percent per inch. The normal curve is drawn relative to the outside one, in percent per standard unit. To see how the scales match up, take the top value on each axis: 60% per standard unit matches 20% per inch because there are 3 inches to the standard unit. Spreading 60% over an SD is the same as spreading 60% over 3 inches, and that comes to 20% per inch—

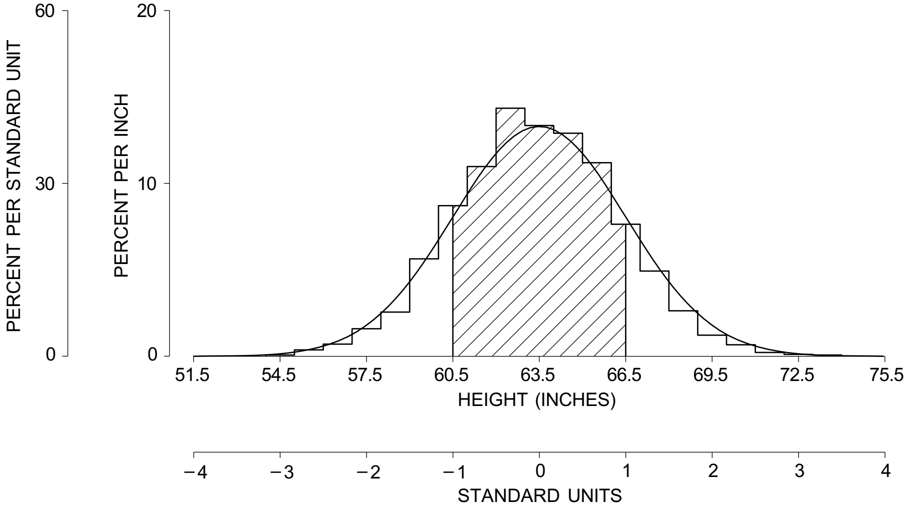
$$\begin{aligned} 60\% \text{ per standard unit} &= 60\% \text{ per } 3 \text{ inches} \\ &= 60\% \div 3 \text{ inches} = 20\% \text{ per inch.} \end{aligned}$$

Similarly, 30% per standard unit matches 10% per inch. Any other pair of values can be dealt with in the same way.

The last chapter said that for many lists, roughly 68% of the entries are within one SD of average. This is the range

$$\text{average} - \text{SD} \quad \text{to} \quad \text{average} + \text{SD}.$$

Figure 2. A histogram for heights of women compared to the normal curve. The area under the histogram between 60.5 inches and 66.5 inches (the percentage of women within one SD of average with respect to height) is about equal to the area between −1 and +1 under the curve—68%.

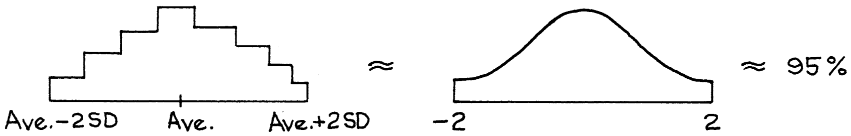


To see where the 68% comes from, look at figure 2. The percentage of women whose heights are within one SD of average equals the area under the histogram within one SD of average. This area is shaded in figure 2. The histogram follows the normal curve fairly well. Parts of it are higher than the curve, and parts of it are lower. But the highs balance out the lows. And the shaded area under the histogram is about the same as the area under the curve. The area under the normal curve between −1 and +1 is 68%. That is where the 68% comes from.

For many lists, roughly 95% of the entries are within 2 SDs of average. This is the range

average − 2 SDs to average + 2 SDs.

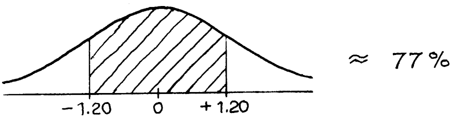
The reasoning is similar. If the histogram follows the normal curve, the area under the histogram will be about the same as the area under the curve. And the area under the curve between −2 and +2 is 95%:



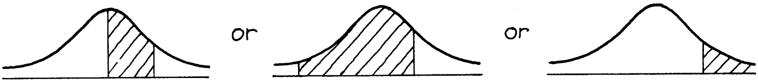
The normal curve can be used to estimate the percentage of entries in an interval, as follows.[2] First, convert the interval to standard units; second, find the corresponding area under the normal curve. The method for getting areas will be explained in section 2. Finally, section 3 will put the two steps together. The whole procedure is called the *normal approximation*. The approximation consists in replacing the original histogram by the normal curve before finding the area.

2. FINDING AREAS UNDER THE NORMAL CURVE

At the end of the book, there is a table giving areas under the normal curve (p. A104). For example, to find the area under the normal curve between– 1.20 and 1.20, go to 1.20 in the column marked *z* and read off the entry in the column marked *Area*. This is about 77%, so the area under the normal curve between −1.20 and 1.20 is about 77%.



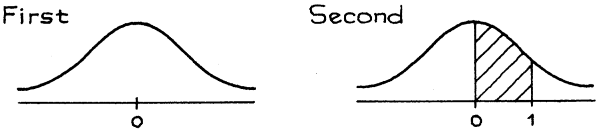
But you are also going to want to find other areas:



The method for finding such areas is indicated by example.

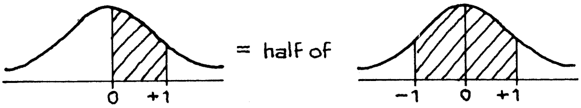
Example 2. Find the area between 0 and 1 under the normal curve.

Solution. First make a sketch of the normal curve, and then shade in the area to be found.



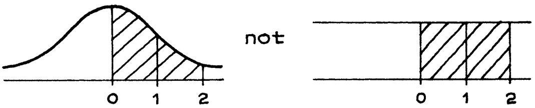
The table will give you the area between −1 and +1. This is about 68%. By symmetry, the area between 0 and 1 is half the area between −1 and +1, that is,

$$\frac{1}{2} \times 68\% = 34\%$$



Example 3. Find the area between 0 and 2 under the normal curve.

Solution. This isn't double the area between 0 and 1 because the normal curve isn't a rectangle.

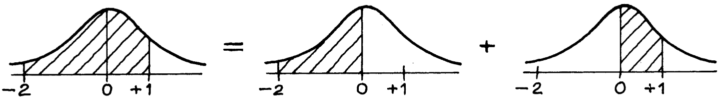


The procedure is the same as in example 2. The area between—2 and 2 can be found from the table. It is about 95%. The area between 0 and 2 is half that, by symmetry:

$$\frac{1}{2} \times 95\% \approx 48\%.$$

Example 4. Find the area between −2 and 1 under the normal curve.

Solution. The area between —2 and 1 can be broken down into two other areas—

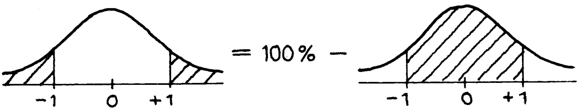


The area between ~~2~~and 0 is the same as the area between 0 and 2, by sym- metry, and is about 48% (example 3). The area between 0 and 1 is about 34% (example 2). The area between −2 and 1 is about

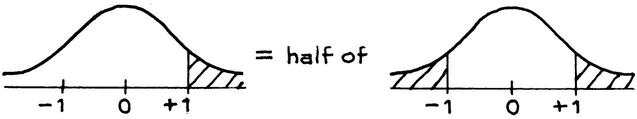
$$48\% + 34\% = 82\%.$$

Example 5. Find the area to the right of 1 under the normal curve.

Solution. The table gives the area between —1 and 1, which is 68%. The area outside this interval is 32%.

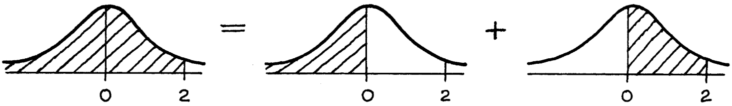


By symmetry, the area to the right of 1 is half this, or 16%.



Example 6. Find the area to the left of 2 under the normal curve.

Solution. The area to the left of 2 is the sum of the area to the left of 0, and the area between 0 and 2.



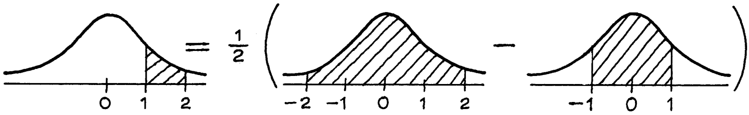
The area to the left of 0 is half the total area, by symmetry:

$$\frac{1}{2} \times 100\% = 50\%$$

The area between 0 and 2 is about 48%. The sum is 50% + 48% = 98%.

Example 7. Find the area between 1 and 2 under the normal curve.

Solution.



The area between −2 and 2 is about 95%; the area between −1 and 1 is about 68%. Half the difference is

$$\frac{1}{2} \times (95\% - 68\%) = \frac{1}{2} \times 27\% \approx 14\%.$$

There is no set procedure to use in solving this sort of problem. It is a matter of drawing pictures which relate the area you want to areas that can be read from the table.

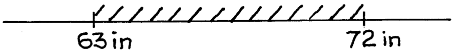
3.THE NORMAL APPROXIMATION FOR DATA

The method for the normal approximation will be explained here by example. The diagrams look so simple that you may not think they are worth drawing. However, it is easy to lose track of the area that is wanted. Please draw the diagrams.

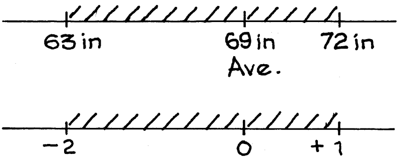
Example. The heights of the men age 18 and over in HANES5 averaged 69 inches; the SD was 3 inches. Use the normal curve to estimate the percentage of these men with heights between 63 inches and 72 inches.

Solution. The percentage is given by the area under the height histogram, between 63 inches and 72 inches.

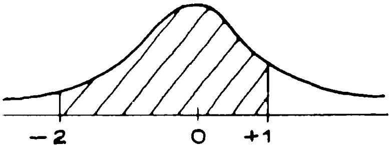
Step 1. Draw a number line and shade the interval.



Step 2. Mark the average on the line and convert to standard units.

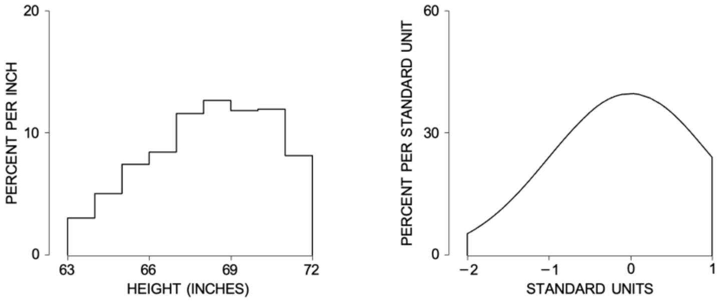


Step 3. Sketch in the normal curve, and find the area above the shaded standard-units interval obtained in step 2. The percentage is approximately equal to the shaded area, which is almost 82%.



Using the normal curve, we estimate that about 82% of the heights were between 63 inches and 72 inches. This is only an approximation, but it is pretty good: 81% of the men were in that range. Figure 3 shows the approximation.

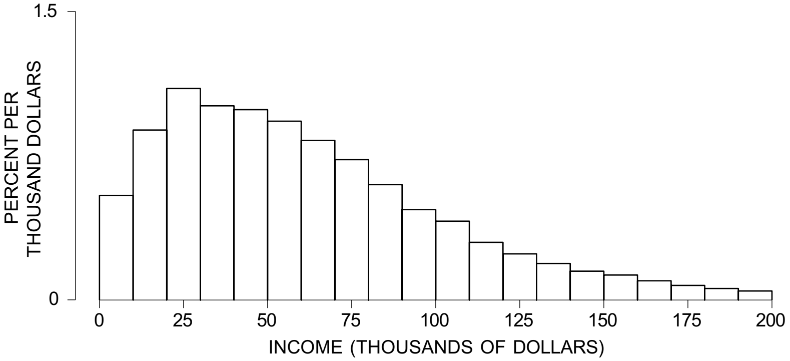
Figure 3. The normal approximation consists in replacing the original histogram by the normal curve before computing areas.



3. PERCENTILES

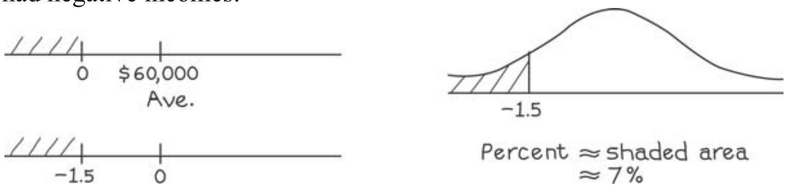
The average and SD can be used to summarize data following the normal curve. They are less satisfactory for other kinds of data. Take the distribution of family income in the U.S. in 2004, shown in figure 5.

Figure 5. Distribution of families by income: the U.S. in 2004.



Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census. Primary families.

The average income for the families in figure 5 was about $60,000; the SD was about $40,000.[3] So the normal approximation suggests that about 7% of these families had negative incomes:



The reason for this blunder: the histogram in figure 5 does not follow the normal curve at all well, it has a long right-hand tail. To summarize such histograms, statisticians often use *percentiles* (table 1).

Table 1. Selected percentiles for family income in the U.S. in 2004.

1	$0
10	$15,000
25	$29,000
50	$54,000
75	$90,000
90	$135,000
99	$430,000

Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census. Primary families.

The 1st percentile of the income distribution was $0, meaning that about 1% of the families had incomes of $0 or less, and about 99% had incomes above that level. (Mainly, the families with no income were retired or not working for some other reason.) The 10th percentile was $15,000: about 10% of the families had incomes below that level, and 90% were above. The 50th percentile is just the median (chapter 4).

By definition, the *interquartile range* equals

$$\text{75th percentile} - \text{25th percentile}.$$

This is sometimes used as a measure of spread, when the distribution has a long tail. For table 1, the interquartile range is $61,000.

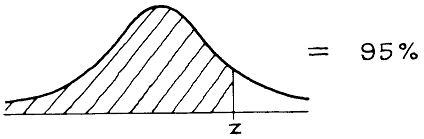
For reasons of their own, statisticians call de Moivre's curve "normal." This gives the impression that other curves are abnormal. Not so. Many histograms follow the normal curve very well, and many others—like the income histogram—do not. Later in the book, we will present a mathematical theory which helps explain when histograms should follow the normal curve.

5. PERCENTILES AND THE NORMAL CURVE

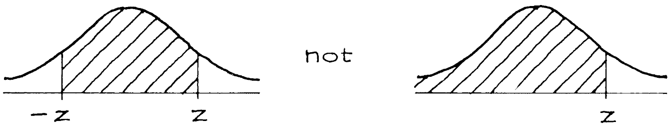
When a histogram does follow the normal curve, the table can be used to estimate its percentiles. The method is indicated by example.

Example 10. Among all applicants to a certain university one year, the Math SAT scores averaged 535, the SD was 100, and the scores followed the normal curve. Estimate the 95th percentile of the score distribution

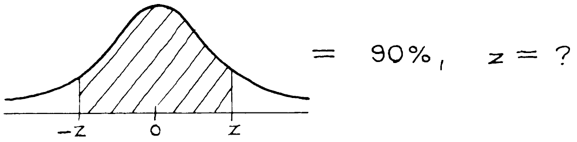
Solution. This score is above average, by some number of SDs. We need to find that number, call it *z*. There is an equation for *z*:



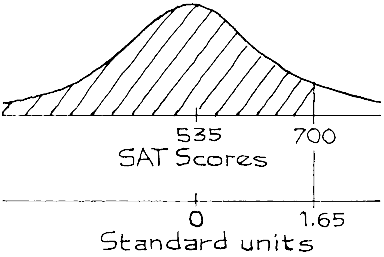
The normal table cannot be used directly, because it gives the area between — *z* and *z* rather than the area to the left of *z*.



The area to the right of our *z* is 5%, so the area to the left of −*z* is 5% too. Then the area between −*z* and *z* must be 100% − 5% − 5% = 90%.



From the table, $z \approx 1.65$. You have to score 1.65 SDs above average to be in the 95th percentile of the Math SAT. Translated back to points, this score is above average by 1.65 × 100 = 165 points. The 95th percentile of the score distribution is 535 + 165 = 700.



The terminology is a little confusing. A *percentile* is a score: in example 10, the 95th percentile is a score of 700. A *percentile rank*, however, is a percent: if you score 700, your percentile rank is 95%. There is even a third way to say the same thing: a score of 700 puts you at the 95th percentile of the score distribution.

6. SUMMARY

1. The *normal curve* is symmetric about 0, and the total area under it is 100%.

2. *Standard units* say how many SDs a value is, above (+) or below (−) the average.

3. Many histograms have roughly the same shape as the normal curve.

4. If a list of numbers follows the normal curve, the percentage of entries falling in a given interval can be estimated by converting the interval to standard units, and then finding the corresponding area under the normal curve. This procedure is called the *normal approximation*.

5. A histogram which follows the normal curve can be reconstructed fairly well from its average and SD. In such cases, the average and SD are good summary statistics.

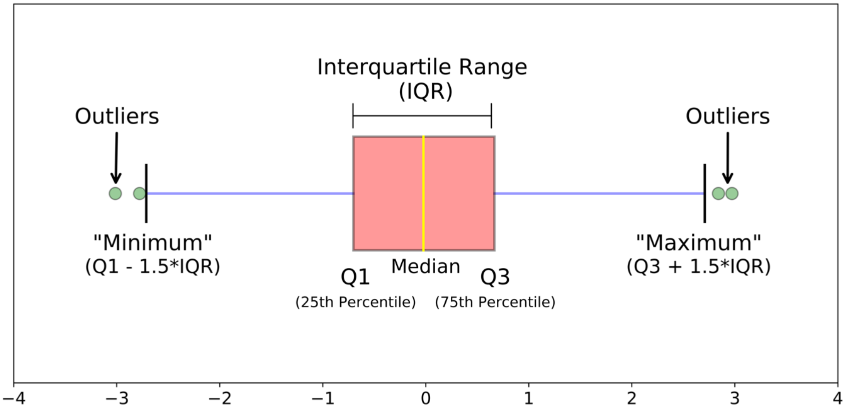
6. All histograms, whether or not they follow the normal curve, can be summarized using *percentiles*.

7. If you add the same number to every entry on a list, that constant just gets added to the average; the SD does not change. If you multiply every entry on a list by the same positive number, the average and the SD just get multiplied by that constant. (If the constant is negative, wipe out the sign before multiplying the SD.)

"LOOK, FRED! THIS SEEMS TO BE THE SAME THING, SUMMARIZED."

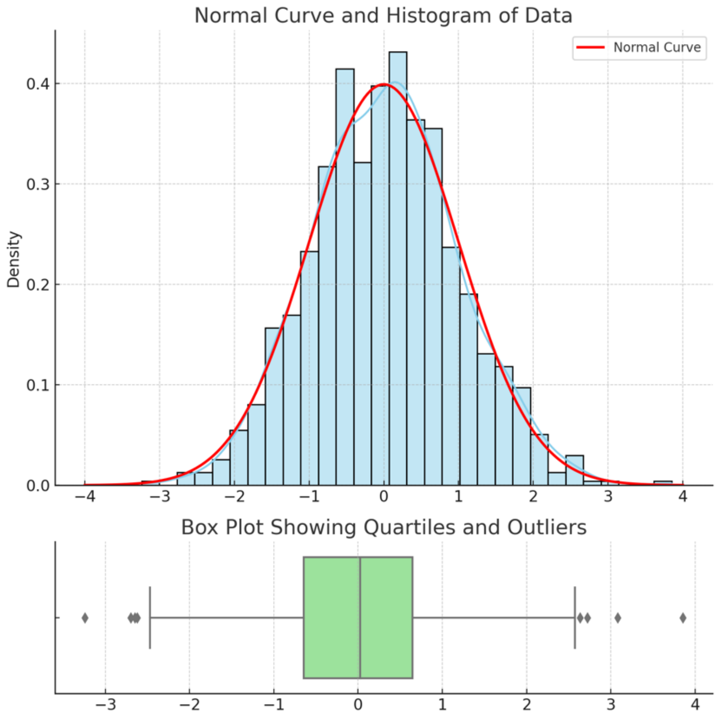
Outliers, Box Plots, and Their Relationship to the Normal Curve

A **box plot** (or whisker plot) is a graphical tool used to summarize the distribution of a dataset. It highlights the **median, quartiles, and potential outliers**. Outliers are data points that fall unusually far from the rest of the distribution. In a box plot, they are typically identified using the **1.5 × IQR rule**, where IQR (interquartile range) is the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Any value greater than Q3 + 1.5 × IQR or less than Q1 – 1.5 × IQR is flagged as an outlier.



Outliers in the box plot often correspond to the **extreme tails of the normal curve**. For example, values beyond 3 standard deviations from the mean are very rare under normality (less than 0.3% of cases) and would almost always appear as outliers in the box plot.

In practice, comparing a **box plot** with the **normal curve** helps assess how well the data aligns with normality. If the box plot shows symmetry, a centered median, and few outliers, the data may approximate a normal distribution. If there are many extreme points or skewness, the box plot suggests deviations from normality.



Here’s a visual that shows how a **normal distribution** (top) relates to its **box plot** (bottom):

- The **histogram with the red curve** represents a normal distribution.
- The **box plot** shows the **median, quartiles (Q1, Q3), whiskers, and outliers**.

Finding Percentiles and Relating to Box Plots

1. **Percentiles**:
    - A percentile indicates the value below which a certain percentage of data falls.
    - For example:
        - The **25th percentile (Q1)** means 25% of the data lies below this value.
        - The **50th percentile (median)** splits the data into two equal halves.

- The **75th percentile (Q3)** means 75% of the data lies below this value.

2. **Box Plot Construction**:
    - The **box** spans from Q1 to Q3 (the interquartile range, IQR).
    - The **line inside the box** marks the median (50th percentile).
    - The **whiskers** typically extend to the last data point within **1.5 × IQR** of the quartiles.
    - Any point beyond the whiskers is flagged as a **potential outlier**.
3. **Connection to the Normal Curve**:
    - In a normal distribution, Q1 and Q3 are symmetric around the mean, and the box plot appears balanced.
    - Most values fall within the whiskers, corresponding to the **bulk of the bell curve**.
    - Outliers in the box plot map to the **extreme tails of the normal curve** (rare events far from the mean).

Further Read on Outliers and their source: Measurement Error

Jesus: I am come to bear witness unto the truth.
Pilate: What is truth?

1. INTRODUCTION

In an ideal world, if the same thing is measured several times, the same result would be obtained each time. In practice, there are differences. Each result is thrown off by chance error, and the error changes from measurement to measurement. One of the earliest scientists to deal with this problem was Tycho Brahe´ (1546–1601), the Danish astronomer. But it was probably noticed first in the market place, as merchants weighed out spices and measured off lengths of silk.

There are several questions about chance errors. Where do they come from? How big are they likely to be? How much is likely to cancel out in the average? The first question has a short answer: in most cases, nobody knows. The second question will be dealt with later in this chapter, and the third will be answered in part VII.

2. CHANCE ERROR

This section will discuss chance errors in precision weighing done at the National Bureau of Standards.[1] First, a brief explanation of standard weights. Stores weigh merchandise on scales. The scales are checked periodically by county weights-and-measures officials, using county standard weights. The county standards too must be *calibrated* (checked against external standards) periodically. This is done at the state level. And state standards are calibrated against national standards, by the National Bureau of Standards in Washington, D.C.

This chain of comparisons ends at the International Prototype Kilogram (for short, The Kilogram), a platinum-iridium weight held at the International Bureau of Weights and Measures near Paris. By international treaty—The Treaty of the Meter, 1875—"one kilogram" was defined to be the weight of this object under standard conditions.[2] All other weights are determined relative to The Kilogram. For instance, something weighs a pound if it weighs just a bit less than half as much as The Kilogram. More precisely,

The Pound = 0.4539237 of The Kilogram.

To say that a package of butter weighs a pound means that it has been connected by some long and complicated series of comparisons to The Kilogram in Paris, and weighs 0.4539237 times as much.

Each country that signed the Treaty of the Meter got a national prototype kilogram, whose exact weight had been determined as accurately as possible relative to The Kilogram. These prototypes were distributed by lot, and the United States got Kilogram #20. The values of all the U.S. national standards are determined relative to K_{20}.

In the U.S., accuracy in weighing at the supermarket ultimately depends on the accuracy of the calibration work done at the Bureau. One basic issue is reproducibility: if a measurement is repeated, how much will it change? The Bureau gets at this issue by making repeated measurements on some of their own weights. We will discuss the results for one such weight, called NB 10 because it is owned by the National Bureau and its nominal value is 10 grams—the weight of two nickels. (A package of butter has a "nominal" weight of 1 pound; the exact weight will be a little different—chance error in butter; similarly, the people who manufactured NB 10 tried to make it weigh 10 grams, and missed by a little.)

NB 10 was acquired by the Bureau around 1940, and they've weighed it many times since then. We are going to look at 100 of these weighings. These measurements were made in the same room, on the same apparatus, by the same technicians. Every effort was made to follow the same procedure each time. All the factors known to affect the results, like air pressure or temperature, were kept as constant as possible.

The first five weighings in the series were

9.999591 grams
9.999600 grams
9.999594 grams
9.999601 grams
9.999598 grams

At first glance, these numbers all seem to be the same. But look more closely. It is only the first 4 digits that are solid, at 9.999. The last 3 digits are shaky, they change from measurement to measurement. This is chance error at work.[3]

NB 10 does weigh a bit less than 10 grams. Instead of writing out the 9.999 each time, the Bureau just reports the amount by which NB10 fell short of 10 grams. For the first weighing, this was

0.000409 grams.

The 0's are distracting, so the Bureau works not in grams but in micrograms: a *microgram* is the millionth part of a gram. In these units, the first five measurements on NB 10 are easier to read. They are

409 400 406 399 402.

All 100 measurements are shown in table 1. Look down the table. You can see that the results run around 400 micrograms, but some are more, some are less. The smallest is 375 micrograms (#94); the largest is 437 micrograms (#86). And there is a lot of variability in between. To keep things in perspective, one microgram is the weight of a large speck of dust; 400 micrograms is the weight of a grain or two of salt. This really is precision weighing!

Even so, the different measurements can't all be right. The exact amount by which NB 10 falls short of 10 grams is very unlikely to equal the first number

Table 1. One hundred measurements on NB 10. Almer and Jones, National Bureau of Standards. Units are micrograms below 10 grams.

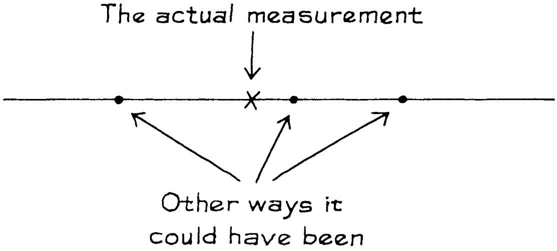
No.	*Result*	*No.*	*Result*	*No.*	*Result*	*No.*	*Result*
1	409	26	397	51	404	76	404
2	400	27	407	52	406	77	401
3	406	28	401	53	407	78	404
4	399	29	399	54	405	79	408
5	402	30	401	55	411	80	406
6	406	31	403	56	410	81	408
7	401	32	400	57	410	82	406
8	403	33	410	58	410	83	401
9	401	34	401	59	401	84	412
10	403	35	407	60	402	85	393
11	398	36	423	61	404	86	437
12	403	37	406	62	405	87	418
13	407	38	406	63	392	88	415
14	402	39	402	64	407	89	404
15	401	40	405	65	406	90	401
16	399	41	405	66	404	91	401
17	400	42	409	67	403	92	407
18	401	43	399	68	408	93	412
19	405	44	402	69	404	94	375
20	402	45	407	70	407	95	409
21	408	46	406	71	412	96	406
22	399	47	413	72	406	97	398
23	399	48	409	73	409	98	406
24	402	49	404	74	400	99	403
25	399	50	402	75	408	100	404

in the table, or the second, or any of them. Despite the effort of making these 100 measurements, the exact weight of NB 10 remains unknown and perhaps unknowable.

Why does the Bureau bother to weigh the same weight over and over again? One of the objectives is quality control. If the measurements on NB 10 jump from 400 micrograms below 10 grams to 500 micrograms above 10 grams, something has gone wrong and needs to be fixed. (For this reason, NB 10 is called a *check weight*; it is used to check the weighing process.)

To see another use for repeated measurements, imagine that a scientific laboratory sends a nominal 10-gram weight off to the Bureau for calibration. One measurement can't be the last word, because of chance error. The lab will want to know how big this chance error is likely to be. There is a direct way to find out: send the same weight back for a second weighing. If the two results differ by a few micrograms, the chance error in each one is only likely to be a few micrograms in size. On the other hand, if the two results differ by several hundred micrograms, each measurement is likely to be off by several hundred micrograms. The repeated weighings on NB 10 save everybody the bother of sending in weights more than once. There is no need to ask for replicate calibrations because the Bureau has already done the work.

No matter how carefully it was made, a measurement could have come out a bit differently. If the measurement is repeated, it will come out a bit differently. By how much? The best way to answer this question is to replicate the measurement.



The SD of the 100 measurements in table 1 is just over 6 micrograms. The SD tells you that each measurement on NB 10 was thrown off by a chance error something like 6 micrograms in size. Chance errors around 2 or 5 or 10 micrograms in size were fairly common. Chance errors around 50 or 100 micrograms must have been extremely rare. The conclusion: in calibrating other 10-gram weights by the same process, the chance errors should be something like 6 micrograms in size.

The SD of a series of repeated measurements estimates the likely size of the chance error in a single measurement.

There is an equation which helps explain the idea:

individual measurement = exact value + chance error.

The chance error throws each individual measurement off the exact value by an amount which changes from measurement to measurement. The variability in repeated measurements reflects the variability in the chance errors, and both are gauged by the SD of the data. Mathematically, the SD of the chance errors must equal the SD of the measurements: adding the exact value is just a change of scale (pp. 92–93).

To go at this more slowly, the average of all 100 measurements reported in table 1 was 405 micrograms below 10 grams. This is very likely to be close to the exact weight of NB 10. The first measurement in table 1 differed from the average by 4 micrograms:

$$409 - 405 = 4.$$

This measurement must have differed from the exact weight by nearly 4 micrograms. The chance error was nearly 4 micrograms. The second measurement was below average by 5 micrograms; the chance error must have been around −5 micrograms. The typical deviation from average was around 6 micrograms in size, because the SD was 6 micrograms. Therefore, the typical chance error must have been something like 6 micrograms in size.

Of course, the average of all 100 measurements (405 micrograms below 10 grams) is itself only an estimate for the exact weight of NB10. This estimate too must be off by some infinitesimal chance error. Chapter 24 will explain how to figure the likely size of the chance error in this sort of average.

Figure 1. The U.S. national prototype kilogram, K_{20}.

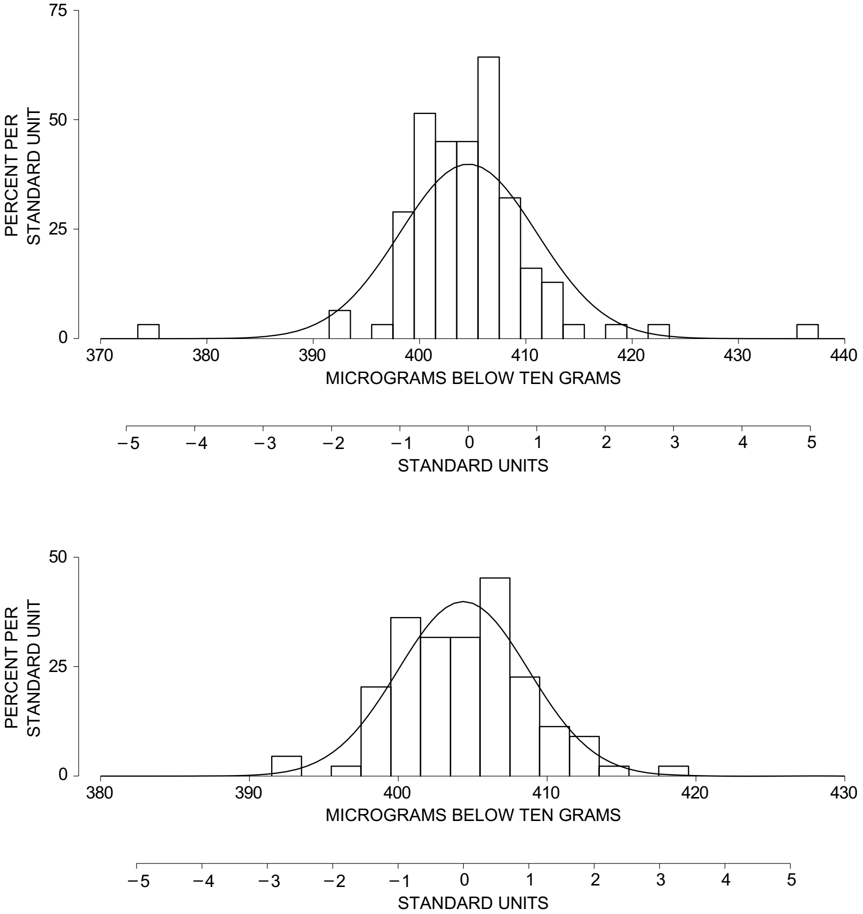
Source: National Institute of Science and Technology.

3. OUTLIERS

How well do the measurements reported in table 1 fit the normal curve? The answer is, not very well. Measurement #36 is 3 SDs away from the average; #86 and #94 are 5 SDs away—minor miracles. Such extreme measurements are called *outliers*. They do not result from blunders. As far as the Bureau could tell, nothing went wrong when these 3 observations were made. However, the 3 outliers inflate the SD. Consequently, the percentage of results falling closer to the average than one SD is 86%—quite a bit larger than the 68% predicted by the normal curve.

When the 3 outliers are discarded, the remaining 97 measurements average out to 404 micrograms below 10 grams, with an SD of only 4 micrograms. The average doesn't change much, but the SD drops by about 30%. As figure 2 shows,

Figure 2. Outliers. The top panel shows the histogram for all 100 measurements on NB 10; a normal curve is drawn for comparison. The curve does not fit well. The second panel shows the data with 3 outliers removed. The curve fits better. Most of the data follow the normal curve, but a few measurements are much further away from average than the curve suggests.



the remaining 97 measurements come closer to the normal curve. In sum, most of the data have an SD of about 4 micrograms. But a few of the measurements are quite a bit further away from the average than the SD would suggest. The overall SD of 6 micrograms is a compromise between the SD of the main part of the histogram—4 micrograms—and the outliers.

In careful measurement work, a small percentage of outliers is expected. The only unusual aspect of the NB 10 data is that the outliers are reported. Here is what the Bureau has to say about *not* reporting outliers.[4] For official prose, the tone is quite stern.

> A major difficulty in the application of statistical methods to the analysis of measurement data is that of obtaining suitable collections of data. The problem is more often associated with conscious, or perhaps unconscious, attempts to make a particular process perform as one would like it to perform rather than accepting the actual performance Rejection of data on the basis of arbitrary performance limits severely distorts the estimate of real process variability. Such procedures defeat the purpose of the program. Realistic performance parameters require the acceptance of all data that cannot be rejected for cause.

There is a hard choice to make when investigators see an outlier. Either they ignore it, or they have to concede that their measurements don't follow the normal curve. The prestige of the curve is so high that the first choice is the usual one—a triumph of theory over experience.

4. BIAS

Suppose a butcher weighs a steak with his thumb on the scale. That causes an error in the measurement, but little has been left to chance. Take another example. Suppose a fabric store uses a cloth tape measure which has stretched from 36 inches to 37 inches in length. Every "yard" of cloth they sell to a customer has an extra inch tacked onto it. This isn't a chance error, because it always works for the customer. The butcher's thumb and the stretched tape are two examples of *bias*, or *systematic error*.

> Bias affects all measurements the same way, pushing them in the same direction. Chance errors change from measurement to measurement, sometimes up and sometimes down.

The basic equation has to be modified when each measurement is thrown off by bias as well as chance error:

$$\text{individual measurement} = \text{exact value} + \text{bias} + \text{chance error}.$$

If there is no bias in a measurement procedure, the long-run average of repeated measurements should give the exact value of the thing being measured: the chance

> errors should cancel out. However, when bias is present, the long-run average will itself be either too high or too low.

Usually, bias cannot be detected just by looking at the measurements them- selves. Instead, the measurements have to be compared to an external standard or to theoretical predictions. In the U.S., all weight measurements depend on the connection between K_{20} and The Kilogram. These two weights have been com- pared a number of times, and it is estimated that K_{20} is a tiny bit lighter than The Kilogram—by 19 parts in a billion. All weight calculations at the Bureau are re- vised upward by 19 parts in a billion, to compensate. However, this factor itself is likely to be just a shade off: it too was the result of some measurement pro- cess. All weights measured in the U.S. are systematically off, by the same (tiny) percentage. This is another example of bias, but not one to worry about.