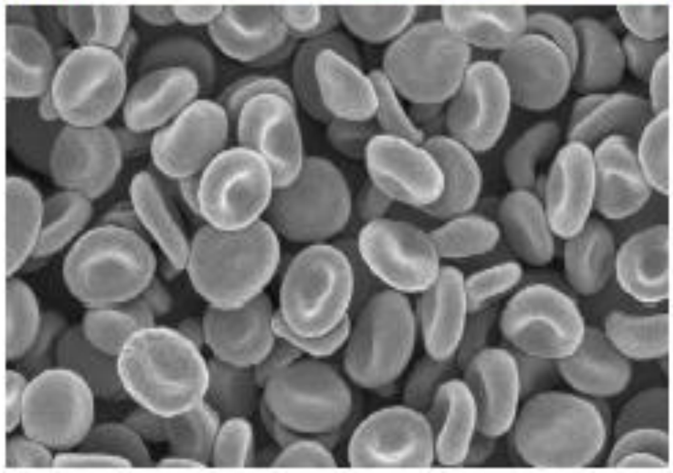
4 Statistics

IS MY RED BLOOD CELL COUNT HIGH TODAY?



Red blood cells, also called erythrocytes, transport oxygen bound to the protein hemoglobin. [Susumu Nishinaga/Photo Researchers.]

All measurements contain experimental error, so it is never possible to be completely certain of a result. Nevertheless, we often seek the answers to questions such as "Is my red blood cell count today higher than usual?" If today's count is twice as high as usual, it is probably truly higher than normal. But what if the "high" count is not excessively above "normal" counts?

Count on "normal" days		Today's count
5.1	$\times 10^6$ cells/μL	5.6×10^6 cells/μL
5.3		
4.8		
5.4		
5.2		
Average = 5.16		

The number 5.6 is higher than the five normal values, but the random variation in normal values might lead us to expect that 5.6 will be observed on some "normal" days.

We will learn in Section 4-3 that there is only a 1.3% random chance of observing a value as far from the average as 5.6 on a "normal" day. It is still up to you to decide what to do with this information.

Experimental measurements always contain some variability, so no conclusion can be drawn with certainty. Statistics gives us tools to accept conclusions that have a high probability of being correct and to reject conclusions that do not.[1]

4-1 Gaussian Distribution

We say that the variation in experimental data is *normally distributed* when replicate measurements exhibit the bell-shaped distribution in Figure 4-1. It is equally probable that a measurement will be higher or lower than the mean. The probability of observing any value decreases as its distance from the mean increases.

If an experiment is repeated a great many times and if the errors are purely random, then the results tend to cluster symmetrically about the average value (Figure 4-1). The more times the experiment is repeated, the more closely the results approach an ideal smooth curve called the **Gaussian distribution.** In general, we cannot make so many measurements in a lab experiment. We are more likely to repeat an experiment 3 to 5 times than 2 000 times. However, from the small set of results, we can estimate parameters that describe the large set. We can then make estimates of statistical behavior from the small number of measurements.

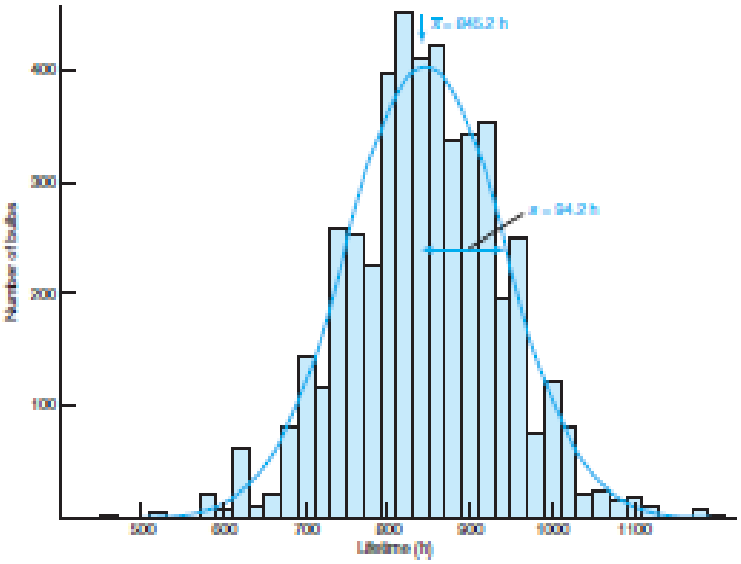


FIGURE 4-1 Bar graph and Gaussian curve describing the lifetimes of a hypothetical set of incandescent light bulbs. The smooth curve has the same mean, standard deviation, and area as the bar graph. Any finite set of data, however, will differ from the bell-shaped curve. The more measurements we make, the closer the results will come to the smooth curve.

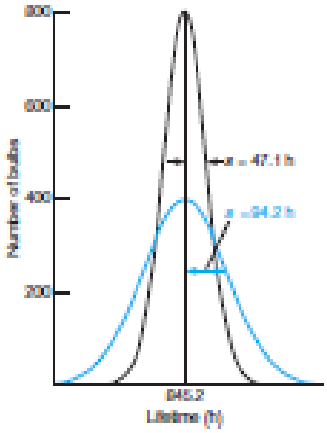


FIGURE 4-2 Gaussian curves for two sets of light bulbs, one having a standard deviation half as great as the other. The number of bulbs described by each curve is the same.

Mean Value and Standard Deviation

In the hypothetical case in Figure 4-1, a manufacturer tested the lifetimes of 4 768 electric light bulbs. The bar graph shows the number of bulbs with a lifetime in each 20-h interval. Lifetimes approximate a Gaussian distribution because variations in the construction of light bulbs, such as filament thickness and quality of attachments, are random. The smooth curve is the Gaussian distribution that best fits the data. Any finite set of data will vary somewhat from the Gaussian curve.

Light bulb lifetimes, and the corresponding Gaussian curve, are characterized by two parameters. The arithmetic **mean**, $\bar{x}$—also called the **average**—is the sum of the measured values divided by n, the number of measurements:

Mean:

$$\bar{x} = \frac{\sum_i x_i}{n} \tag{4-1}$$

The mean gives the center of the distribution. The standard deviation measures the width of the distribution.

where x_i is the lifetime of an individual bulb. The Greek capital sigma, Σ, means summation: $\sum_i x_i = x_1 + x_2 + x_3 + \cdots + x_n$. In Figure 4-1, the mean value is 845.2 h.

The **standard deviation**, s, measures how closely the data are clustered about the mean. *The smaller the standard deviation, the more closely the data are clustered about the mean* (Figure 4-2).

An experiment that produces a small standard deviation is more precise than one that produces a large standard deviation. Greater precision does not necessarily imply greater accuracy, which is nearness to the "truth."

Standard deviation:

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}} \tag{4-2}$$

In Figure 4-1, $s = 94.2$ h. A set of light bulbs having a small standard deviation in lifetime is more uniformly manufactured than a set with a large standard deviation.

For an *infinite* set of data, the mean is designated by the lowercase Greek letter mu, μ (the population mean), and the standard deviation is written as a lowercase Greek sigma, σ (the population standard deviation). We can never measure μ and σ, but the values of $\bar{x}$ and s approach μ and σ as the number of measurements increases.

As the number of measurements increases, $\bar{x}$ approaches μ, *if there is no systematic error.*

The quantity $n - 1$ in Equation 4-2 is called the **degrees of freedom**. The square of the standard deviation is called the **variance**. The standard deviation expressed as a percentage of the mean value ($= 100 \times s/\bar{x}$) is called the *relative standard deviation* or the *coefficient of variation*.

Coefficient of variation $= 100 \times \frac{s}{\bar{x}}$

4-2 Confidence Intervals

Student's *t* is a statistical tool used most frequently to express confidence intervals and to compare results from different experiments. It is the tool you could use to evaluate the probability that your red blood cell count will be found in a certain range on "normal" days.

"Student" was the pseudonym of W. S. Gosset, whose employer, the Guinness breweries of Ireland, restricted publications for proprietary reasons. Because of the importance of Gosset's work, he was allowed to publish it (Biometrika **1908**, *6*, 1), but under an assumed name.

Calculating Confidence Intervals

From a limited number of measurements (n), we cannot find the true population mean, μ, or the true standard deviation, σ. What we determine are $\bar{x}$ and s, the sample mean and the sample standard deviation. The **confidence interval** is computed from the equation

$$\text{Confidence interval} = \bar{x} \pm \frac{ts}{\sqrt{n}} \qquad (4\text{-}6)$$

where t is Student's t, taken from Table 4-2 for a desired level of confidence, such as 95%. The meaning of the confidence interval is this: If we were to repeat the n measurements many times to compute the mean and standard deviation, the 95% confidence interval would include the true population mean (whose value we do not know) in 95% of the sets of n measurements. We say (somewhat imprecisely) that "we are 95% confident that the true mean lies within the confidence interval."

EXAMPLE Calculating Confidence Intervals

The carbohydrate content of a glycoprotein (a protein with sugars attached to it) is found to be 12.6, 11.9, 13.0, 12.7, and 12.5 wt% (g carbohydrate/100 g glycoprotein) in replicate analyses. Find the 50% and 90% confidence intervals for the carbohydrate content.

TABLE 4-2 Values of Student's *t*

Degrees of freedom	Confidence level (%)						
	50	90	95	98	99	99.5	99.9
1	1.000	6.314	12.706	31.821	63.656	127.321	636.578
2	0.816	2.920	4.303	6.965	9.925	14.089	31.598
3	0.765	2.353	3.182	4.541	5.841	7.453	12.924
4	0.741	2.132	2.776	3.747	4.604	5.598	8.610
5	0.727	2.015	2.571	3.365	4.032	4.773	6.869
6	0.718	1.943	2.447	3.143	3.707	4.317	5.959
7	0.711	1.895	2.365	2.998	3.500	4.029	5.408
8	0.706	1.860	2.306	2.896	3.355	3.832	5.041
9	0.703	1.833	2.262	2.821	3.250	3.690	4.781
10	0.700	1.812	2.228	2.764	3.169	3.581	4.587
15	0.691	1.753	2.131	2.602	2.947	3.252	4.073
20	0.687	1.725	2.086	2.528	2.845	3.153	3.850
25	0.684	1.708	2.060	2.485	2.787	3.078	3.725
30	0.683	1.697	2.042	2.457	2.750	3.030	3.646
40	0.681	1.684	2.021	2.423	2.704	2.971	3.551
60	0.679	1.671	2.000	2.390	2.660	2.915	3.460
120	0.677	1.658	1.980	2.358	2.617	2.860	3.373
∞	0.674	1.645	1.960	2.326	2.576	2.807	3.291

In calculating confidence intervals, s may be substituted for σ in Equation 4-6 if you have a great deal of experience with a particular method and have therefore determined its "true" population standard deviation. If σ is used instead of s, the value of t to use in Equation 4-6 comes from the bottom row of Table 4-2.

Values of t in this table apply to two-tailed tests illustrated in Figure 4-9a. The 95% confidence level specifies the regions containing 2.5% of the area in each wing of the curve. For a one-tailed test, we use values of t listed for 90% confidence. Each wing outside of t for 90% confidence contains 5% of the area of the curve.