1

Some Basic Concepts

Objectives

At the end of this chapter you should be able to:

- Describe the reasons for conducting occupational and environmental health science (IH/EHS) exposure measurements
- Distinguish between physical sampling and statistical sampling
- Discuss the importance of representative statistical sampling
- Define precision as it relates to IH/EHS measurements
- Calculate joint, marginal, and conditional probabilities and test for independence of events
- Recognize the characteristics of the binomial, normal, and chi-square probability distributions
- Perform calculations related to the binomial, normal, and chi-square probability distributions

1.1 Introduction

Industrial hygiene and environmental health sciences (IH/EHS) practitioners measure things – it is what we do. The range of risks to health and the environment includes those due to chemicals' hazards (irritants, corrosives, carcinogens, reproductive toxins, central nervous system depressants, asphyxiants, heavy metals, etc.), physical energy hazards (extreme heat and cold, vibration, ionizing and nonionizing radiations, noise), biological hazards (airborne infectious agents, bloodborne pathogens, contact infection transmission hazards, allergens, opportunistic pathogens), and ergonomic hazards (cumulative trauma due to repetitive motions, musculoskeletal soft tissue injury, stress-inducing positions). Toxic chemicals may be encountered in indoor occupational environments, ambient outdoor environments, and the water and food we consume. Quantifying these hazards is a crucial step in determining the degree of risk and developing strategies for reducing it if necessary.

The what, when, where, who, and how of our measurements is driven by the *why*. Why do we measure things? Obviously, it is to answer a question. Some common reasons for measuring are as follows:

- To demonstrate regulatory compliance
  a. Are worker exposures exceeding OSHA Permissible Exposure Limits (PELs) or other occupational exposure guidelines such as the ACGIH Threshold Limit Values (TLV®)?
  b. What are the chemical concentrations in air, water, soil, food, or other media?
  c. Are environmental discharges exceeding emissions permit limits?

Statistical Tools for the Comprehensive Practice of Industrial Hygiene and Environmental Health Sciences,
First Edition. David L. Johnson.

- To establish baseline exposure information for specific exposure sources
  a. What are the sources of worker or public exposures to occupational or environmental insults?
  b. At what rates are pollutants being released?
  c. What is the spatial and temporal distribution of exposure levels around a source?
- To evaluate the effectiveness of control measures
  a. Is there a difference in the effectiveness of alternative controls?
  b. How effective is an intervention (engineering control, employee training, process change, etc.), that is, is the exposure or emission different (hopefully reduced!) after its implementation?
  c. Is an engineering control working as well now as it was previously?
- To characterize the frequency distribution of potential exposures or events
  a. What is the range of potential exposures or emissions?
  b. How frequently do different exposure levels occur?
  c. What fraction of worker exposures or process emissions is likely to exceed allowable levels?
  d. Is my company's rate of adverse events (overexposures, emissions exceedances, accidents, etc.) typical of the industry?
  e. Are adverse events occurring more frequently now than in the past?
- To explore associations between exposure variables
  a. What are the process factors and environmental factors that contribute to exposures?
  b. Which factors have the greatest influence on exposures?
  c. Do interactions between different contributors influence exposures?

Other good reasons for conducting measurements are to document exposure levels or emissions that are known to be within allowable limits to protect your employer against unfounded liability claims or undeserved regulatory penalties, or to reassure workers or the public. The first is just good business from an economic standpoint, but an additional benefit is that measurements documenting lower level exposures contribute to the epidemiological data from which improved dose–response relationships and associated exposure guidelines are developed. Measuring to reassure workers or the public is similarly good business because it promotes good public and labor-management relations.

Clearly, answering these very different questions requires different types and amounts of information, so the measurement strategy must be crafted to provide enough data, of high enough quality, to reliably answer the specific question. What "enough data," "high enough quality," and "reliably answer" mean are explored as we proceed.

1.2 Physical versus Statistical Sampling

We should distinguish at the outset the difference between "sample" as usually used in IH/EHS practice and "sample" as used in statistics. To the IH/EHS practitioner, "sample" usually means a physical quantity of something, such as a volume of soil, air, water, or other environmental media. In statistics, though, "sample" means a subset of all possible measurements that could be made of a quantity, from which we draw inferences about the whole population. A *census* is the special case where all possible values are in fact measured. It may seem trivial to point this out, but we should be clear on this distinction from the beginning.

We employ *statistical sampling* because there are never enough resources – as in time, people, and money – to measure all possible exposures of IH/EHS interest, where by *exposure* we mean any relevant physical, chemical, biological, or other measure of workplace or ambient environmental quality. For example, airborne contaminant concentrations due to industrial processes will typically vary up and down and over a wide range in the course of a workday, work week, season, and so on, and will likely differ as well across workers performing what appears to be the same task. To measure even one worker's full range of exposures would be extremely resource intensive, and the utility of the information would be limited because only one worker was measured, thus our reliance on statistical sampling to minimize the amount of measurement needed to draw inferences about the whole population of potential exposures. In this case, we might use our professional judgment and experience to identify groups of workers we believe likely to have similar exposures and conduct measurements on a subset of individuals randomly selected from within the group. Results for members of each s*imilar exposure group* (*SEG*) might then be used to draw inferences about the entire group's exposures.

1.3 Representative Measures

Reliable inferences about a population of exposures hinge on the measurements being *representative of the population as a whole*. Nonrepresentative sampling leads to *inferential bias*, because we are not measuring what we expect. *Representative sampling* means we have conducted our measurements such that no potential influences have been excluded or weighted differently than they occur in the exposed population. A simplistic example would be if we wanted to characterize sediment contaminant levels in a lake downstream from a pollution source but only measured sediment collected from readily accessible areas such as the shoreline or off the end of a fishing dock. This would surely give an incomplete if not outright misleading picture regarding the environmental impact of contaminant inflow to the lake. Another example might be selecting subjects only from among the day shift when measuring worker exposures at a multishift facility – there might very well be systematic differences in exposure that are influenced by the shift time, such as the rate of work, tasks performed, training and experience levels of the workers, or environmental conditions. Such differences comprise *systematic errors* in our characterization, and systematic errors result in bias.

Systematic measurement errors and associated bias can also result even when the sampling strategy is completely correct, through bias in the actual measurements themselves. When using any type of measurement instrument, it is essential to verify that it is working properly and that it is *accurate*, that is, the indicated result represents "truth." Repeated measurements of an unchanging quantity using an improperly calibrated instrument may give the same answer every time (plus or minus a bit of random variation), but the answer will be consistently off in one direction. Such systematic measurement errors can be eliminated through good quality control practices and procedures and, therefore, should not occur. If they do occur, it may be difficult if not impossible to correct for the bias afterward or to even know that it has occurred. In all further discussion, we assume that measurements are performed using accurate measurement techniques, so that the measurement results are unbiased.

1.4 Strategies for Representative Sampling

Inferences drawn from statistical samples can only be valid if the samples are representative of the broader population of all possible measurements we could have made. We must choose where or when or who to measure in such a way that we avoid introducing bias at the outset.

Consider the situation in which we are interested in estimating the average body mass index (BMI) of high school students (9th–12th grades) in our city. We might choose the nearest high school, get a list of all of the student names, and randomly select *n* names from the list, that is, we would collect a *random sample* from the student roster. We then measure each student's *BMI* and average them. Is this a good approach? Perhaps not. Random sampling is good, but choosing the school because it is nearby, that is, conducting a *convenience sample*, may not be. What if this school's study body was predominantly of one race? Average *BMI* is known to differ across ethnic groups, so our measurements would not be generalizable to the entire city unless all high schools in the city had the same ethnic distribution. Or what if for some reason this particular school had an unusually unbalanced distribution of male versus female students? Average *BMI* certainly varies across gender, so our average *BMI* measure would only be valid for this particular gender ratio and would not be representative of all high school students in the city.

Another sampling approach to avoid is chain-referral or *snowball sampling*, in which new subjects are referred by current subjects from among their circle of acquaintances or family members. It should be obvious that such groups are likely to have more in common with each other than with the general population. There are many types of sampling that do not involve random selection, and they should generally all be avoided, though there are exceptions for particular circumstances or study types.

It is tempting to think that the ideal approach would be to get a listing of all high school students in the city, randomly select *n* names from the list, and conduct the measurements. That would provide an unbiased estimate of the mean *BMI* for high school students in the city but would provide no information on how *BMI* varied across gender, ethnicity, grade level, socioeconomic status, or other demographics that might allow more meaningful interpretation of the data, comparison of our city's values with those of other cities, or comparison of future measurements with the current data.

Sampling strategies that allow more representative and informative measurements include *cluster sampling* and *stratified sampling*. Cluster sampling involves identifying relatively homogenous naturally occurring groups so that random sampling can be conducted within each group. Stratified sampling is slightly different in that the population is divided into groups based on an important demographic (e.g., gender), and representative measurements are made within each of the strata.

The key to obtaining representative measures is first to clearly define the question being asked, decide what needs to be measured, carefully assess all of the known or suspected factors that could influence the measurement, then develop a sampling strategy that measures the quantities of interest while allowing the influence of potential interferences to be "accounted for" in the statistical analysis.

1.5 Measurement Precision

A fundamental concept to understand in applying statistical methods is that of *precision*, or the similarity of replicate measures. An old joke maintains that "*A man with one watch always knows what time it is. A man with two watches is never quite sure.*" The joke, of course, is that the two watches will not show *exactly* the same time, so the owner cannot be certain which one is "true" (assuming of course that one or the other is in fact correct).

A more relevant and illustrative example of precision in IH/EHS measurements might be the bubble tube "frictionless piston" calibration of air flow through a sampling train. A typical calibration setup is shown in Figure 1.1. With the pump set at a fixed speed, the amount of

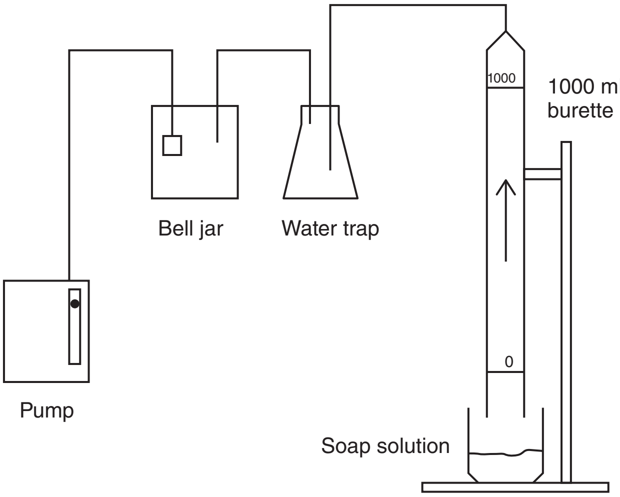

Figure 1.1 Typical bubble tube calibration apparatus for a sampling train comprised of a personal air sampling pump and filter cassette.

time taken for a soap film to transit the distance between two volume marks on the bubble tube, representing a known displaced air volume, is measured with a stopwatch. In this case, the operator starts the watch as the film passes 0 and stops it as the film passes 1000 mL, so that the air flow during that time period is 1000 mL or 1 L. Dividing the volume in liters by the time in minutes gives the flow rate in liters per minute (L/min). However, even if the pump's flow rate does not vary the measured times will not all be exactly the same due to small random variations in starting and stopping the watch twice in this manual process. The difference in any two measures can be termed the *discrepancy* in the measures (Bevington and Robinson, 1992, p. 5). By random we mean that the variation is not consistently higher or lower than the "true" value and also that the magnitude of the variation is not constant.

The pattern of variation for 10 trials might look like that shown in the table and frequency graph of Figure 1.2, which is roughly symmetric about the 30.0 seconds value (representing 2.0 L/min pump flow rate if a 1 L burette is used). Such random variations should in fact be symmetrically distributed about the "true" value, which is the arithmetic average or "mean" of the distribution. The differences of the individual measured values from the mean are termed "deviations" and taken together they reflect the measurement technique's precision. More on these descriptive measures is provided later. Only 6 time values occur, with several occurring more than once. The number of times a value occurs is its *frequency*. When the frequency of occurrence is plotted versus the measured value as shown in Figure 1.2, we can visualize the distribution of the measurements over their range.

The form of the *frequency distribution*, or the pattern of how often the different values occur, becomes clearer as the number of measures increases, so that the distribution for an experiment with 1000 trials might look like the distribution in Figure 1.3.

Other types of measurements involving counts over a measurement interval may involve the aforementioned types of random variation as well as statistical variation due to randomness in the occurrence of events being counted. Bevington and Robinson term these *instrumental uncertainty* and *statistical uncertainty*, respectively (Bevington and Robinson, 1992, pp. 38–40). For example, a count detector such as a light-scattering aerosol photometer may detect particles very reliably and have essentially no instrumental uncertainty for dilute aerosols, but due to the spatial variation in the air's particle concentration there may be discrepancies in repeated

Measurement data (s)	
30.2	30.3
29.8	29.9
30.0	30.1
30.1	30.0
30.0	29.9

n = 10

Value (s)	Frequency of occurrence
29.8	1
29.9	2
30.0	3
30.1	2
30.2	1
30.3	1
	10

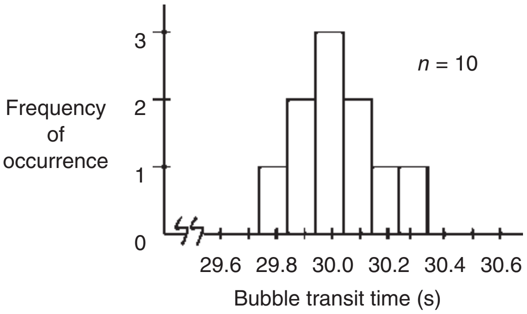

Figure 1.2 Frequency table and graph of soap bubble film transit times for 10 trials.

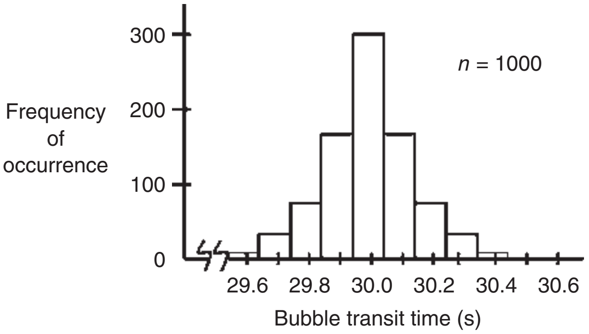

Figure 1.3 Frequency graph of soap bubble film transit times for 1000 trials.

measurements simply due to the stochastic nature of the number of particles passing through the sensing zone during any given observation period.

1.6 Probability Concepts

Probability, or the likelihood of an event occurring, is a straightforward concept with which everyone has had personal experience. The simple statement "I probably won't run into any major traffic delays today" reflects the experience of a commuter that on most days traffic is smooth, that is, that it is more likely than not that there will be no traffic delays on any given day. How likely it is that traffic will be smooth can be quantitatively expressed with probability. The more definite statement that "90 percent of the time I don't run into bad traffic" is an expression of the probability of smooth traffic occurring on a given day.

Probability is the foundation of statistical analysis techniques, and probability theory is a field of mathematics unto itself. We introduce some basic concepts and terminology as they relate to statistics; for an in-depth exploration of probability, texts devoted to the subject should be consulted.

1.6.1 The Relative Frequency Approach

Someone who states "Nine days out of ten there is nothing or only junk mail in my mailbox" is speaking from personal experience in which they have observed that on average they receive useful mail only once in every 10 days of mail delivery. Intuition tells us that on any randomly chosen day there is a 9 in 10 chance that there will be nothing or only junk mail in the mailbox, that is, there is a *90% probability* of nothing or only junk mail. This type of *probability*, estimated from past observation, is termed the *relative frequency* of an event – the number of times it occurs, x, expressed as a fraction of the total number of observations, n:

$$\textit{relative frequency} = \frac{\textit{number of times a particular outcome is observed}}{\textit{total number of observations}} = \frac{x}{n}. \tag{1.1}$$

Relative frequency provides an *estimate* of the probability of finding only junk mail on any given day. It is an estimate because these were observations for only a subset of all the mail delivery days that might be observed, that is, it is a sample of the total population of all possible mail delivery days. If nothing or only junk mail is found in the mailbox 9 out of 10 days, the relative frequency of finding only junk mail is 9/10 or 0.90, and the probability of finding nothing or only junk mail on any given day is estimated to be 90%. The probability of observing an event A is expressed as $P\{A\}$.

1.6.2 The Classical Approach – Probability Based on Deductive Reasoning

The probability of an event occurring may also be deduced from an understanding of the process involved, with the classic example being the roll of a six-sided die. There are only six possible outcomes (1–6), so that the chance of rolling any particular number is 1/6 if the die is "honest" (not biased in some way by uneven shape or weighting). Even if we had never seen a six-sided die, we could use deductive reasoning to predict the probability. Four-sided, eight-sided, ten-sided, twelve-sided, and even twenty-sided polyhedral dice can be purchased, and we can easily deduce the probability of rolling a particular number (or symbol on some dice) as 1 divided by the number of sides. In this case, we are not estimating the probability – we know it *exactly* from first principles. This *classical approach* based on deductive reasoning is the approach mathematicians used to develop early probability theory.

1.6.3 Subjective Probability

Probability estimated from relative frequency is based on measurements, and probability known exactly is based on an understanding of the process and deductive reasoning. These are objective probability assessments. However, the probability of an event occurring can also be *subjectively* estimated based on one's *degree of belief* about the likelihood of an occurrence. This is termed the *personalistic approach* to probability (Sheskin, 2011, pp. 372–375). This belief may not be based solely or at all on measurements but on other knowledge or experience. A *prior* belief, perhaps based on results of previous studies but perhaps only on professional opinion, is updated with data to reach a modified, or updated, *posterior* belief. Such probabilities are used in *Bayesian methods*, developed independently by Thomas Bayes and Pierre-Simon Laplace in the mid-1700s and early 1800s, respectively. The application of Bayesian methods in exposure assessment decision-making is discussed in Chapter 13.

1.6.4 Complement of a Probability

The total probability of all possible events is 1.0. If two black marbles, two green marbles, and one white marble are placed in a hat and one is drawn out without looking, deductive reasoning tells us that the probability of drawing a black marble is $p = P\{black\} = 2/5 = 0.40$. The probability of *not* drawing a black marble is the *complement* of this probability, $q = P\{not\,black\} = 1 - p = 0.60$, which deductive reasoning again tells us is the probability of drawing a green marble ($2/5 = 0.40$) plus the probability of drawing a white marble ($1/5 = 0.20$). More formal rules for determining probability are discussed in subsequent sections.

1.6.5 Mutually Exclusive Events

In the aforementioned examples, none of the x occurrences could happen at the same time. In any roll of a die, only one number can come up, and on any given day "no mail or junk mail only" will either occur or not – there cannot be useful mail and no mail/junk mail only on the same day. That is, the events are *mutually exclusive*. For two mutually exclusive events A and B, the probability of one *or* the other occurring is the sum of their individual probabilities:

$$P(A \cap B) = P(A) + P(B), \tag{1.2}$$

where the ∪ symbol indicates "union" or "or," that is, $P(A \cup B) =$ the probability of either A or B occurring (but not both, since they are mutually exclusive). Equation 1.2 is the *Addition Rule* for the probability of one or the other of two mutually exclusive events occurring and is the applicable rule in the marble illustration in which only one marble was drawn.

Example 1.1

On average, 15% of a cell phone owner's incoming calls are from scammers, 10% are from legitimate marketers, 25% are from friends, 20% are from family members, 10% are from business contacts, and 20% are from other sources. What is the probability that any given incoming call will be from either a scammer or a legitimate marketer (neither of which will be a welcome intrusion)?

Solution

Note that the probabilities sum to 1.0, as they must if all types of incoming phone calls are represented.

These are mutually exclusive events, assuming friends and family are not also scammers, marketers, or business contacts, so $P(A \cup B) = P(A) + P(B)$. The probability of a scammer call is 0.15, and of a legitimate marketing call is 0.10, so

$$P(A \cup B) = P(A) + P(B) = 0.15 + 0.10 = 0.25\,\text{or}\,25\%.$$

1.6.6 Independent Events

If the probability of one event A is not influenced by the occurrence of a second event B, the two events are said to be *independent*. If a coin is flipped and a Head turns up, and then is flipped again, the outcome of the second flip is not influenced by the fact that a Head turned up on the first flip. The two flips are independent.

If two events are independent, then the probability of *both* events occurring is the product of their individual probabilities:

$$P(A \cap B) = P(A)P(B), \tag{1.3}$$

where the ∩ symbol indicates "intersection" or "and," that is, $P(A \cap B) =$ the probability of both A and B occurring. This is termed the *joint probability* of A and B occurring together.

Example 1.2
A delivery customer observes over time that on any given day in June there is a 10% probability that a package will be left on his open porch. He also observes that during June there is a 15% probability of having a thunderstorm on any given day. What is the probability that the customer will have a delivery rained on during any given day in June?

Solution
Having a package delivered and experiencing a rainstorm are independent events, so the probability of both events occurring on the same day is the joint probability:

$$P(A \cap B) = P(A)P(B) = 0.10(0.15) = 0.015$$

or only about 1.5%, so it is highly unlikely that a package will get wet.

1.6.7 Events that Are Not Mutually Exclusive

The delivery and thunderstorm events in Example 1.2 are not mutually exclusive events since they can occur at the same time. For two nonexclusive events, the Addition Rule includes a third term. The probability of observing one, *or* the other, *or both* is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.4}$$

If the two events are independent, $P(A \cap B) = P(A)P(B)$ in the equation.

Example 1.3
For the situation in Example 1.2, what is the probability that on any given June day the customer will receive a package or experience a thunderstorm or both receive a package and experience a thunderstorm?

Solution
The probability of experiencing either or both of these independent events is

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B) \\ &= 0.10 + 0.15 - 0.10(0.15) = 0.235 \text{ or } 23.5\%. \end{aligned}$$

1.6.8 Marginal and Conditional Probabilities

Table 1.1 displays the typical number of hours per month devoted to various woodworking tasks during manufacture of custom dining room tables and chairs in a one-person Early American furniture shop. Construction of both products involves the same basic tasks of sawing wood planks to size and shape using a table saw or band saw, turning table and chair legs and chair spindles on a lathe, assembling the parts using glue and clamps, and sanding and finishing. The amount of time spent in each task area depends on which furniture item is being made. Only one type of furniture is made at a time and only one type of task is performed at a time in this small shop.

The probability that at any given time the shop will be engaged in sawing is $P(sawing) = 42/180 = 0.233$, the total number of hours spent sawing divided by the total hours of observation. Similarly, the probability that at any given time the shop will be making chairs is $P(chairs) = 103/180 = 0.572$. These are termed *marginal probabilities* because the frequencies are the sums in the table margins.

Table 1.1 Woodworking hours per month for various tasks during chair and table manufacture in a small furniture shop.

Woodworking operation	Chairs	Tables	Totals
Sawing	19	23	42
Turning legs and spindles	42	13	55
Assembly	38	14	52
Sanding/finishing	21	10	31
Totals	103	60	180

Conditional probability is the probability that an event will occur *given* that some other event has already occurred, that is, given that some condition has been met. The probability that A will occur given that B has occurred is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{1.5}$$

where $P(A \cap B)$ is the joint probability of A and B occurring together and $P(B)$ is the marginal probability for B. The | symbol represents "given," and $P(A|B)$ is read "the probability of A given B."

Example 1.4
What is the probability that during a visit to the shop we would find them turning parts on the lathe *given* that we know they are working on tables?

Solution
This is a conditional probability – the probability of observing turning given that tables are being built. The probability would be the joint probability of being engaged in turning while working on tables divided by the marginal probability of working on tables:

$$P(turning|tables) = \frac{P(turning\ and\ tables)}{P(tables)} = \frac{{}^{13}/_{180}}{{}^{60}/_{180}} = 0.217.$$

Note that this solution is equivalent to ignoring all of the hours associated with working on chairs, so that the "population" of total hours under consideration is only those associated with making tables, that is, 60 hours. Within those 60 hours there will be 13 hours devoted to turning, so the probability of observing turning while tables are being made is 13/60 = 0.217 as mentioned earlier.

Example 1.5
From the furniture shop data given in Table 1.1, if a glance into the shop indicates that wood planks are being sawed, what is the probability that chairs are being built?

Solution
This conditional probability would be the probability of being engaged in sawing while making chairs divided by the marginal probability of sawing:

$$P(chairs|sawing) = \frac{P(chairs\ and\ sawing)}{P(sawing)} = \frac{19}{42} = 0.452.$$

Example 1.6
What would be the probability of observing *either* sawing *or* sanding/finishing during a visit to the shop given that they are making tables?

Solution
Extending the aforementioned logic, we restrict our consideration to only the hours involved in making tables. Then the probability of observing either sawing or sanding/finishing is the sum (since they are mutually exclusive) of the two conditional probabilities of sawing given tables and sanding/finishing given tables:

$$P = \frac{23}{60} + \frac{10}{60} = \frac{33}{60} = 0.550.$$

1.6.9 Testing for Independence

Are the events in the furniture shop (furniture type and work task) independent? Intuition should tell us that they are not, because the probability of being engaged in assembly, for example, is different for chair work than for table work. We can state this more formally as a test. There are actually three tests that must be met to show independence:

$$P(A|B) = P(A)P(B|A) = P(B)P(A \cap B) = P(A)P(B). \tag{1.6}$$

For the probabilities associated with A = chairs and B = sawing, it was shown earlier that $P\{chairs|sawing\} = \frac{19}{42} = 0.452$. However, the marginal probability of working on chairs was shown to be $P(chairs) = \frac{120}{180} = 0.667$. The two are not equal, so the test $P(A|B) = P(A)$ is not met. Similarly, $P(sawing|chairs) = \frac{19}{120} = 0.158$, while $P(sawing) = \frac{42}{120} = 0.350$, so the test $P(B|A) = P(B)$ is not met. Finally, $P(chairs \cap sawing) = \frac{19}{180} = 0.106$ while $P(chairs)P(sawing) = 0.667(0.350) = 0.233$, so the third test is not met. Any of the three tests would show that the events are not independent.

Example 1.7
A survey of a bat species is conducted to determine whether gender influences whether the bat carries a particular parasite. Test gender and parasite status for independence.

	Parasite	No parasite	Totals
Male	12	48	60
Female	16	64	80
Totals	28	112	140

Solution
Any combination of the two parameters could be chosen. Using Parasite (yes) and Male gender:

$$P(Parasite|Male) = \frac{12}{60} = 0.200 \text{ and } P(Parasite) = \frac{28}{140} = 0.200$$

$$P(Male|Parasite) = \frac{12}{28} = 0.429 \text{ and } P(Male) = \frac{60}{140} = 0.429$$

$$P(Parasite \cap Male) = P(Parasite|Male)P(Male) = 0.200(0.429) = 0.086 \text{ and}$$

$$P(Parasite)P(Male) = 0.200(0.429) = 0.086.$$

All three tests of independence are met, so the gender of the bat has no influence on whether it carries the parasite – they are independent events.

1.7 Permutations and Combinations

Permutations and combinations are important probability-related concepts. “To permute” means to rearrange the order or sequence of things, so *permutations* are the different ways of ordering a group of objects. A *combination* is a selection of objects with no consideration given to the order of the selection.

1.7.1 Permutations for Sampling without Replacement

Suppose we would like to measure some characteristic of 5 different processes and feel it would be a good idea to randomize the order in which we do the measures. How many different orderings are possible? We might write the name of each process on a slip of paper, put the 5 slips in a jar and shake them up, close our eyes, and draw the slips out one at a time. A simple but effective randomization technique! There are 5 possible outcomes for the first draw, only 4 for the second, 3 for the third, and so on, so the total number of ways we could order the 5 trials would be 5(4)(3)(2)(1) = 120 permutations. Generally, the number of ways of ordering n distinct objects is then

$$permutations = n(n-1)(n-2)\dots(1) = n! \tag{1.7}$$

or “n factorial,” in which $0! \equiv 1$. This is an example of *sampling without replacement* since once a process is drawn from the jar it cannot be drawn again.

In some cases, we might not need all n of the n objects, that is, we might want to draw only a subset of objects from the total. For example, let us say that for our 5 process measurements task we can only measure 2 of them in one day. How many possible orderings are there of 2 processes drawn from the 5 (without replacement)? We could choose any of the 5 for the first one, then any of the 4 remaining for the second one, or 5(4) = 20 permutations. The more general form of Equation 1.7 for the case of ordering only r of the n objects is then the number of permutations

$${}_nP_r = n(n-1)(n-2)\dots(n-r+1) = \frac{n!}{(n-r)!}. \tag{1.8}$$

Equation 1.7 corresponds to the special case of Equation 1.8 for $r = n$, that is, all of the objects are ordered. On the second day, there would be only 3 processes left, so there would be 3(2) = 6 possible orderings of the two measurements that day. Obviously, on the third day there is only one ordering for the single process left to measure. The total of these would be 20(6)(1) = 120.

Excel provides the *PERMUT* function for calculating permutations in sampling without replacement. The function has the form PERMUT(n, k), so for the situation just described the cell equation would read =PERMUT(5,2) for the first day and PERMUT(3,2) for the second day.

Example 1.8
Our personal protective equipment manufacturing company is ready to begin marketing a new line of noise-canceling hearing protective devices and plans to number each unit with a unique 8-digit number. How many unique numbers can be created, with the restrictions that the first digit cannot be zero and no digit can be used more than once?

Solution
The first digit cannot be zero, so there are only 9 ways (1–9) to select the first number. For the second digit, there are also 9 choices because 0 is now available but the previously chosen number is not. For the third there are 8 choices, for the fourth 7 choices, and so on. The total

of number of unique permutations is 9(9)(8)(7)(6)(5)(4)(3) = 1, 632,960 numbers. We can sell quite a few devices without having to come up with a new system.

The number of permutations ${}_nP_r$ of n objects made up of groups of which n_i are alike is

$$ {}_nP_r = \frac{n!}{n_1!(n_2!)(n_3!)\ldots}. \tag{1.9}$$

Permutations in which the objects in a like group are switched are considered identical.

Example 1.9
How many possible arrangements are there of the letters in the word Mississippi?

Solution
The i appears 4 times, the s appears 4 times, the p appears 2 times, and m appears 1 time. The number of possible permutations is

$$\frac{n!}{n_1!(n_2!)(n_3!)\ldots} = \frac{11!}{4!4!2!1!} = 34{,}650\,\textit{possible arrangements}.$$

1.7.2 Permutations for Sampling with Replacement

If objects were replaced and could be drawn again, there would be n choices for the first draw, n for the second, n for the third, and so on, so that for r draws:

$$\textit{permutations} = n^r. \tag{1.10}$$

This is an example of *sampling with replacement*.

Example 1.10
For the situation in Example 1.8, how many unique numbers could we produce, still with the restriction that the first number cannot be 0, if we can reuse the digits?

Solution
Still with the restriction that the first number cannot be 0, we would have 9 ways (1–9) to select the first number but then 10 choices for each of the next 7 numbers. The total number of possible permutations would then be

$$9(10^7) = 9 \times 10^7\,\textit{possible numbers}$$

We could sell a *lot* of devices before needing a new system!

1.7.3 Combinations

Consider a situation in which we would like to divide our class of 10 students into lab partner pairs. How many ways are there of forming the first pair? That is, how many *combinations* of $n = 10$ people taken $r = 2$ at a time are possible? This is abbreviated $\binom{n}{r}$ or ${}_nC_r$ and is calculated as

$$\binom{n}{r} = {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!}. \tag{1.11}$$

Example 1.11
A particular wastewater discharge sampling task requires using integrating water samplers that periodically draw volumes from the discharge stream over a 24-hour period and combine them

into a single integrated sample. If you have 6 samplers in your inventory but need only 3 for the task, how many ways are there to select 3 of the 6? The order of selection is irrelevant.

Solution

From Equation 1.11, there are $\binom{6}{3} = \frac{6!}{3!(6-3)!} = \frac{6!}{3!\,3!} = \frac{6(5)(4)(3)(2)(1)}{3(2)(3)(2)} = 20$ possible combinations of the 6 samplers taken 3 at a time.

Excel provides the *COMBIN* function for calculating combinations. The function has the form COMBIN(n, k), so for the situation in Example 6.15, the cell equation would read =COMBIN(6,3).

1.8 Introduction to Frequency Distributions

Figure 1.2 graphically shows the distribution of frequencies with which individual measurement values occurred in the 10-measurement bubble tube experiment previously discussed. Figure 1.3 shows what the distribution might look like if we had a lot more measurements. The pump's air flow rate was (in theory) constant, so the variation in measurements had to be due to random differences in exactly when the stopwatch was started and stopped during each trial. The highest frequency is in the middle of the measurement range, with frequencies symmetrically distributed to either side and decreasing with distance from the center. But why does it have this "bell" shape?

In general, barring any systematic bias the difference in a measurement value and the true value of the quantity being measured will be due to the combined effect of perhaps several independent random influences. Each has the potential to randomly nudge the measure in a negative or positive direction to varying degrees, and so they tend to offset one another to some extent. The result is that small net differences are more likely to occur than large net differences. Why this occurs is a matter of probabilities, which we use to illustrate sources of imprecision.

1.8.1 The Binomial Distribution

Consider the situation of flipping a coin, and let a Head outcome be scored as +1 and a Tail outcome as −1. For an "honest" coin flipped an infinite number of times, we would expect the average of all the outcomes to be 0 since flipping a Head is equally as likely as flipping a Tail – each has a probability of occurrence of 0.5 or 50%. For a more limited number of flips ("trials," in statistical terms), intuition might suggest that it is more likely than not that there will be more Heads than Tails or vice versa, so that the average has a net positive or negative value. Intuition might also suggest that in a series of 10 flips, for example, it is highly unlikely that there will be either no Heads or 10 Heads.

Let the random variable X denote the total number of Heads observed in such an experiment. The probability of observing exactly x Heads in n trials can be calculated from

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \tag{1.12}$$

where p is the probability of a Head (0.5) and $q = 1 - p = 0.5$ is the probability of a Tail since these are mutually exclusive events with only these two possible outcomes. Recall that the "!" symbol denotes "factorial," for example, $n! = n(n-1)(n-2)(n-3)\ldots$ (with $0! \equiv 1$).

If Equation 1.12 is used to calculate the probabilities for x values from 0 to 10 in $n = 10$ trials it will be found that 5 Heads is the most likely outcome at $P\{X = 5\} = 0.2461$, and that 0 or 10

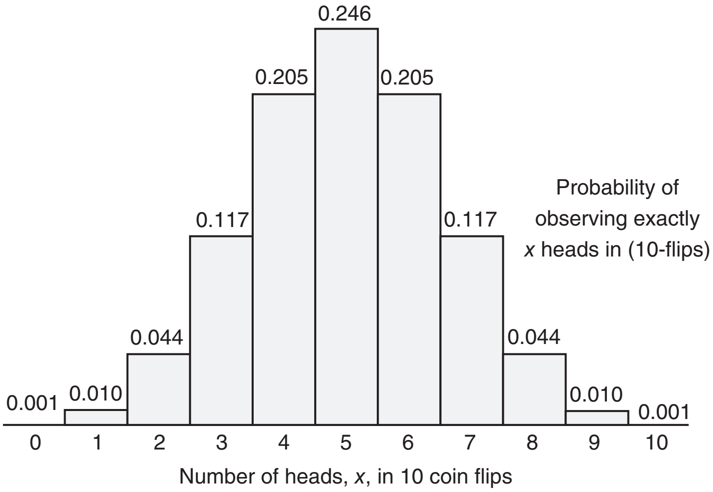

Figure 1.4 Probabilities of the binomial distribution for equal likelihood outcomes ($p = q = 0.5$), as applied to the coin flip example. Small net differences in the number of heads and tails are more likely than large net differences.

Heads is the least likely at $P\{X = 0\} = P\{X = 10\} = 0.0010$ (rounded to 4 decimal places), as shown in Figure 1.4. Note that the probabilities are symmetric about the center and add up to a total probability $P = 1.0$ since these 11 outcomes are the only ones possible for 10 flips (we assume that the coin never lands on its edge!).

The equation and figure represent the *binomial probability distribution* for *dichotomous* outcome events, that is, trials with only two possible outcomes. The binomial is a *discrete data distribution* because its measures can take on only specific values (integer counts in this case). The outcome of the flip is therefore a *discrete random variable.* The shape of the distribution depends on the values of p and q, and is only symmetric when $p = q = 0.5$. More detailed discussion of probability theory underlying the binomial distribution can be found in basic statistics texts such as Zar (2010) and Daniel and Cross (2013).

Example 1.12

You are preparing dilutions of a colorimetric stock solution prior to calibrating a spectrophotometer. You need to pipette 10 mL into a vial but have only a 1 mL volume pipette to work with. If it takes 10 pipettings to make up the 10 mL, and each pipetting has a potential random error of ±0.05 mL (for the purposes of this illustration, we assume that you never get exactly 1.0 mL volume, and that the random error is exactly +0.05 mL or −0.05 mL each time), what is the probability that you will end up with exactly 10 mL?

Solution

In order to end up with exactly 10 mL, there would have to be 5 over-pipettings and 5 under-pipettings, that is, $x = 5$ for $n = 10$ trials. The probability of this happening ($P(x = 5)$) is 0.246 as shown in Figure 1.4, or only about 25%.

Example 1.13

For the situation in Example 1.12, how likely would it be to end up with a volume that is off by 0.1 mL or more in either direction (i.e., 9.9 mL or less, or 10.1 mL or more)?

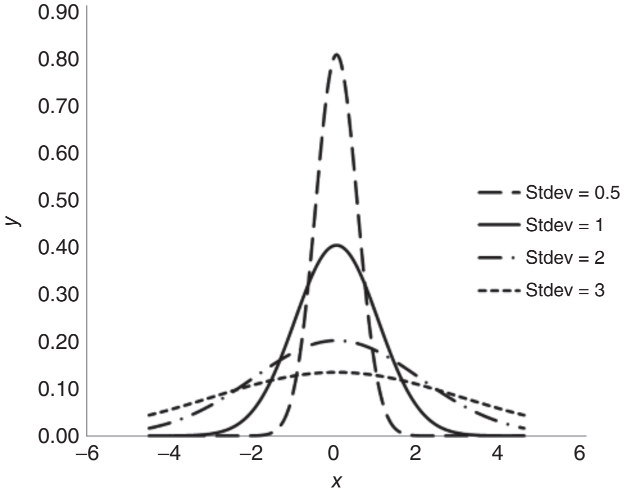

Figure 1.5 The normal distribution for mean $\mu = 0$ and various standard deviations σ. The curve for $\mu = 0$ and $\sigma = 1$ is the standard normal distribution.

Solution

The probability of being off by 0.1 mL or more in 10 pipettings is equivalent to there being 4 or more under-pipettings to get 9.9 mL or less or 6 or more over-pipettings to get 10.1 mL or more. From Figure 1.4 for all values of $x \leq 4$ (or of $x \geq 6$), $P(x \leq 4) = P(x \geq 6) = 0.001 + 0.010 + 0.044 + 0.117 + 0.205 = 0.377$ from the Addition Rule.

A more convenient approach to obtaining the probability in Example 1.9 would be to obtain the *cumulative probability* for $x \leq 4$ from Excel's *BINOM.DIST function.* This is the sum of all of the discrete probabilities for values $\leq x$ in Figure 1.4. The cell equation would be =BINOM.DIST(x, n, p, 1) = BINOM.DIST(4, 10, 0.5, 1), where x is the number of "successes" in n trails, p is the probability of a success, and the 1 is a logical operator telling Excel we want the cumulative (summed) probability, not the probability for exactly x. Excel will return the value 0.37659. If we had used the 0 logical operator the function would return the same value as Equation 1.12.

Here, we have used the binomial distribution to illustrate why the net result of a number of random influences acting together tend to offset one another to some extent, so that the net random error in a measurement is more likely to be smaller than larger. The outcome of any one component of the error occurring in either the positive direction or in the negative direction is an example of a binomial process in which there are only two possible outcomes. We revisit such binomial processes in later chapters.

1.8.2 The Normal Distribution

In the bubble tube example, the minimum difference in measurement times was 0.1 seconds – the limit of resolution of the stopwatch – so that net errors could only take on discrete values. In the more general case, random errors can take on any value over some range, and so are *continuous variables.* The resulting net error is therefore also a continuous random variable, as is its probability distribution. That distribution is the bell-shaped *normal* or *Gaussian distribution*, examples of which are shown in Figure 1.5. The normal distribution is a *continuous probability distribution* because the x values are continuous.

The entire normal distribution is described by only two parameters, the *population mean* μ and the *population standard deviation* σ, whose square is the *population variance* σ^2. The mean is a measure of *central tendency* in the measurements and standard deviation is a measure of *dispersion* (deviations) of measurements about the central value. Figure 1.5 illustrates three normal distributions with the same mean (0) but different standard deviations (0.5, 1, 2, and 3).

If the mean μ and standard deviation σ parameters of a normal data distribution are not known exactly but are estimated from measurement data, they are represented by the *sample mean* $\bar{x}$ and *sample standard deviation* s (and sample variance s^2) statistics. The sample mean $\bar{x}$ is simply the arithmetic average of the individual measurement values:

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum(f_i x_i)}{\sum f_i}, \tag{1.13}$$

where x_i is any one of the n measured values and f_i is the number of times that value occurs, that is, its frequency. The sample standard deviation is calculated as

$$s = \sqrt{\frac{1}{n-1}\sum f_i (x_i - \bar{x})^2}. \tag{1.14}$$

The square of the sample standard deviation is the sample variance, s^2.

The $n - 1$ in Equation (1.14) is termed the degrees of freedom (*df*), an important concept that we will encounter frequently as we proceed. In this application, it should be evident that the deviations of the measurement values from the mean, $x_i - \bar{x}$, must sum to zero, so that if we know all but one, that is, $n - 1$, of the deviations we can calculate the last one. This is termed a *linear constraint* on the deviations (Box *et al*., 2005, pp. 26–27). Equation 1.8 uses one estimate of a population parameter (the sample mean estimates the population mean) to estimate another population parameter (the sample standard deviation estimates the population standard deviation), so there is one linear constraint on the estimate. In later applications, we will see more than one estimated population component being used in a calculation, so that there will be p linear constraints and therefore $n - p$ degrees of freedom.

Example 1.14

A Nebraska stream's dissolved oxygen concentration (mg/L) is measured daily for 3 weeks during the winter with the results shown. What are the mean and standard deviation of the oxygen concentrations?

	O_2 conc. (mg/L)		
	Week 1	**Week 2**	**Week 3**
Monday	9.7	11.6	6.4
Tuesday	9.0	10.5	9.2
Wednesday	12.6	7.7	9.0
Thursday	10.9	8.1	12.2
Friday	12.4	12.0	10.3
Saturday	10.2	8.5	11.6
Sunday	19.8	6.6	12.2

Solution

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{21}(9.7 + 9.0 + 12.6 + \ldots + 11.5 + 12.2) = \frac{220.5}{21} = 10.5\,\text{mg/L}$$

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

$$= \sqrt{\frac{1}{21-1}\left[(9.7 - 10.5)^2 + (9.0 - 10.5)^2 + \ldots + (12.2 - 10.5)^2\right]} = 2.84\,\text{mg/L}.$$

A random variable X that follows the normal distribution with mean μ and variance σ^2 is often represented as $X \sim N(\mu, \sigma^2)$.

The height of the normal distribution curve y at any point x is given by

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{1.15}$$

Equation (1.15) is called the *probability density function* (*pdf*) of the normal distribution. The value of the probability density (the height of the curve) at any x-value can be obtained from the *NORM.DIST function* (NORMDIST in older versions of Excel), where the cell equation is NORMDIST(x, μ, σ) with the x-value of interest andμ and σ for the distribution.

For the special case of the mean $\mu = 0$ and variance $\sigma^2 = 1$ (so that the standard deviation $\sigma = 1$ as well), we have the *standard normal distribution,* often designated by $N(0, 1)$. Relative to the standard normal distribution, normal distributions with other values of μ are simply shifted left or right of 0 on the number line, whereas other values of σ change the height and breadth of the curve as shown in Figure 1.5. The value of the probability density at any x-value of the standard normal distribution can be obtained from the *NORM.S.DIST function* (NORMSDIST in older versions of Excel) as NORM.S.DIST(z, 0), where 0 a logical operator that tells Excel we want the probability density (the height of the curve) and not the cumulative probability (the area under the curve) as described later. The same value is provided by the NORM.DIST function if NORM.DIST(x, 0, 1) is specified.

The *area* under the normal distribution curve between two values x and $x + \Delta x$ is the probability that a measured value will fall within that range. For a given x value, the area under the curve to the left of that point is the probability that a measured value will be less than or equal to x, and the area to the right of the point is the probability that a measured value will be greater than x. For a given point *exactly* equal to x (i.e., $\Delta x = 0$), there is no area under the curve and the probability is 0; this can be a difficult concept to grasp. The total area under any normal distribution curve between $-\infty$ and $+\infty$ is 1.0, that is, 100% total probability since it includes all possible values of x.

For the $N(0, 1)$ distribution, the horizontal distance from zero to any point on the line is the *standard normal deviation, z,* also called the *z-score* or *z-value. z* can range from 0 to $-\infty$ on the left side and from 0 to $+\infty$ on the right side. An x-value in any normal distribution $N(\mu, \sigma)$ can be transformed to a standard normal deviation by

$$z = \frac{x - \mu}{\sigma}. \tag{1.16}$$

From examination of Equation (1.16), it will be seen that z is just the number of standard deviations the x value is from the distribution mean.

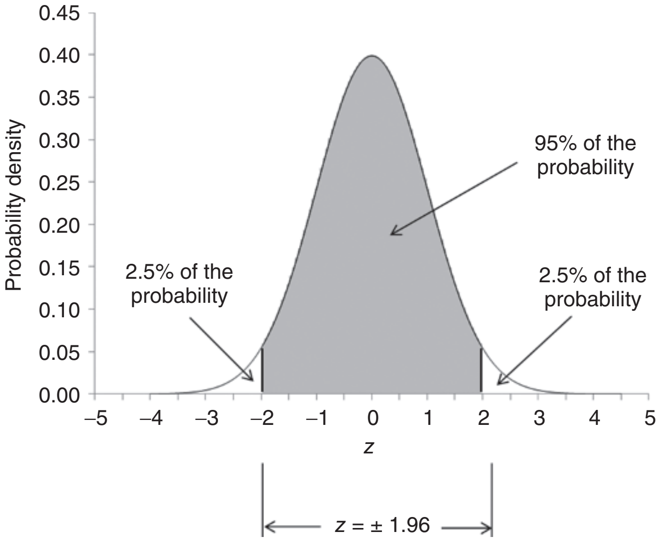

Figure 1.6 For any normal distribution, 95% of the probability (area under the curve) is within the range $\pm$1.96 standard deviations of the mean. For the standard normal curve with $\sigma = 1$, this is equivalent to $z = \pm 1.96$.

Tables of probability under the standard normal curve are given in Tables A1 and A2. Table A1 is a table of the cumulative probability from $z = -\infty$ to x, where x is non-negative. The first value in this table is $P = 0.5000$ for $z = 0.00$, indicating that half of the total probability, for $-\infty < z \leq 0$ on the left side of the distribution, has already been accounted for. Table A2 provides the probability under the normal curve in the range $-z$ to $+z$. The first value in this table is therefore 0.0000 when $z = 0$.

For any normal distribution, 95% of the area under the curve, that is, of the probability, is contained in the interval from $x = \mu - 1.96\sigma$ to $x = \mu + 1.96\sigma$. For standard normal distribution in which $\mu = 0$ and $\sigma = 1$, this is equivalent to $z = -1.96$ to $z = +1.96$ as shown in Figure 1.6 (see also Table A2). One could observe as well that 95% of the probability is also contained in the interval from $-\infty$ to $z = +1.645$ (see Table A1).

The z-value associated with a given cumulative probability from $-\infty$ to z can be calculated in Excel using the *NORM.S.INV* function (or NORMSINV in older versions of Excel), and the cumulative probability associated with a given z-value can be calculated from the NORM.S.DIST using z and the logical operator 1 to obtain the cumulative probability as noted earlier. These functions can be typed directly into a cell or inserted using the Formulas/Insert Function tab. For example, NORM.S.INV(0.975) would return 1.96 and NORM.S.DIST(1.96, 1) would return 0.9750, reflecting all of the probability to the left of the $z = +1.96$ point in Figure 1.6. Other statistical utilities that perform more complex analyses are available in the *Analysis ToolPak* add-in[1]. A glossary of some particularly useful Excel utilities and cell functions is included in the addendum to this chapter.

[1] The statistical functions in Analysis ToolPak are accessed by clicking on the Data tab in the header bar. If a Data Analysis tab does not appear in the Data page toolbar it just means the ToolPak utility has not been added. This can be quickly done by going to the File/Options/Add-ins page and clicking on the Manage button with Excel Add-Ins selected. Analysis ToolPak and several other options will appear in a dialog box – just mark the Analysis ToolPak option (not Analysis ToolPak VBA – which is for business statistics), and click OK. When you go back to the Data page you should see Data Analysis as the last tab in the toolbar. Clicking on Data Analysis tab will bring up the statistical utility options.

Example 1.15
For a normal distribution of values with $\mu = 1.5$ and $\sigma = 0.55$, what is the value x that is larger than 90% of all the values, that is, what is the 90th percentile value?

Solution
We have a table of standard normal z-values and their cumulative probabilities (Table A1), so we can readily find the z-value for the 90th percentile. Then, given the sample mean and standard deviation we can transform this to a corresponding x-value in the sample data distribution using Equation (1.16).

From Table A1, the 90th percentile point in the standard normal distribution occurs between $z = 1.28$ ($p = 0.8997$) and $= 1.29$ ($p = 0.9015$).

z	P
1.28	0.8997
?	0.9000
1.29	0.9015

By linear interpolation

$$z_{90\%} = 1.28 + (1.29 - 1.28)\frac{(0.9000 - 0.8997)}{(0.9015 - 0.8997)} = 1.2817.$$

To complete the transformation, we rearrange $z = \frac{x-\mu}{\sigma}$ to get $x = \mu + z\sigma = 1.5 + 1.2817(0.55) = 2.205$.

In Excel, NORM.S.INV(0.90) returns 1.281552 for z, for which the corresponding $x = 2.207$. The difference in results is due to applying linear interpolation to a difference in probabilities between $z = 1.28$ and $z = 1.29$ that is actually not linear. This is clear from inspection of the probability density *curve*.

The normal probability distribution is an extremely useful tool having many applications in measurement data analysis. Beginning in Chapter 3, we describe how the properties of the normal distribution are applied in *parametric* analytical techniques.

1.8.3 The Chi-Square Distribution

Another important probability distribution that underlies several statistical techniques used in this text is the *chi-square distribution* (also designated by χ^2) (read "kye square"). It can be shown that the sum of the squares of values of a standard normal random variable (such as z-scores) follows the chi-square distribution, whose shape is determined by its degrees of freedom. The degrees of freedom k for this application are equal to the number of squared measures.

The probability density function of the chi-square distribution with k degrees of freedom is described by (Daniel and Cross, 2013, p. 602)

$$f_X(x|k) = \frac{x^{[k/2]-1}e^{-x/2}}{\Gamma\left(\frac{k}{2}-1\right)2^{k/2}}, \tag{1.17}$$

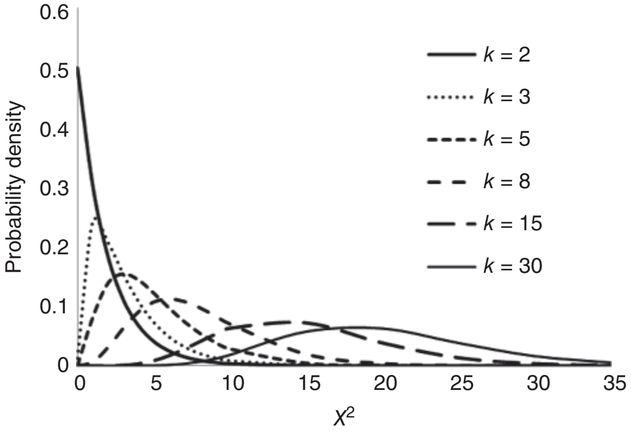

Figure 1.7 The shape of the chi-square distribution for various degrees of freedom, k. The distribution approximates the normal distribution when k is large.

for $x > 0$, where k is the number of degrees of freedom and Γ is the gamma function. When k is an even numbered integer ≥ 2, $\Gamma\left(\frac{k}{2}\right) = \left(\frac{k}{2} - 1\right)!$ If k is an odd numbered integer or < 2 then the value of the Γ function must be taken from a table of its values, which can be found via the Internet or in a book of mathematical tables. Recall that the symbol | in Equation 1.17 represents "given," so that $f_X(x|k)$ is the probability density of the chi-square distributed variable x given that there are k df. The shape of the distribution changes dramatically as the number of degrees of freedom increases, from highly skewed at $k = 2$ to something approaching the normal distribution as k exceeds 10 or so (Figure 1.7). Thus, the chi-square distribution is "*asymptotically normal*" – its shape approaches that of the normal distribution as the degrees of freedom grows large.

The cumulative probability, that is, $P(X \leq x)$, of the chi-square distribution can be calculated using the *CHISQ.DIST* function in Excel as CHISQ.DIST(x, k, 1). The CHISQ.DIST(x, k, *TRUE or FALSE*) function returns either the value of the probability density function at x for k df if FALSE (0) or the cumulative (left-tailed) probability up to x if TRUE (1). Figure 1.7 was created using pdf values calculated with the former. The area under the curve to the right of x is returned by the *CHISQ.DIST.RT* function, that is, CHISQ.DIST.RT(x, k). This right-tail probability is equivalent to 1−CHISQ.DIST(x, k, 1).

Given a set of k measures drawn from a normal distribution of all possible measures that could have been made, for which the sample mean is $\bar{x}$ and the sample standard deviation is s, we can transform the measured values to z-scores using Equation (1.16) as $z_i = \frac{x_i - \bar{x}}{s}$. The sum of the squares of the z_i values will then be chi-square distributed with k df.

Example 1.16

Four measurements of contaminant concentration in a body of water are 55, 50, 51, and 53 mg/L. The mean μ and standard deviation σ of the population of all previous measurements are 48 and 3, respectively. What is the probability associated with the X^2 value for these measures?

Solution

Each of the measures is transformed to a z-score using Equation (1.16), the z-scores are squared and summed, and this value is used to find the probability.

x	*z*-Score	*z*-Squared	
55	2.33	5.44	
50	0.67	0.44	
51	1.00	1.00	
53	1.67	2.78	
μ	48	9.67	$=X^2$
σ	3		

From Appendix B, the probability associated with $X^2 = 9.67$ for $k = 4\,df$ may be estimated as 0.9055 by linear interpolation, so that the probability of exceeding the 9.67 value (the probability in the tail of the distribution to the right of 9.67) is $1 - 0.9528 = 0.0472$ or about 4.7%. However, using the CHISQ.DIST.RT function, CHISQ.DIST.RT$(9.67, 4) = 0.046$ or about 5%. This difference is due to linear interpolation of a nonlinear curve of values, and demonstrates that using the Excel function is more accurate as well as more convenient than interpolating between table values.

We use the chi-square distribution in a number of applications in the following chapters.

1.9 Confidence Intervals and Hypothesis Testing

We have discussed how random variation will cause replicate measures of a fixed parameter to span a range of values. For a great many trials, the measurement values will exhibit a frequency distribution – typically a normal distribution described by its mean and standard deviation. For normally distributed measures, we can identify x-values of the distribution between which we would expect some fraction of many replicate measurements to fall. The range between these two x-values, within which we would expect the true (population) parameter value to fall, is termed the *confidence interval*. The width of the confidence interval is determined by the fraction we choose, typically 95%, and is termed the *confidence level* or *confidence coefficient*. The confidence level is the probability, over repeated sampling, that any given confidence interval will contain the true parameter value. For example, if we were to take 100 samples and construct a confidence interval on each result, for a confidence level of 95% we would expect 95 out of the 100 intervals to contain the true parameter value being measured. Thus, when comparing our measured value to some hypothesized parameter value, if the parameter value does not fall in the confidence interval around the measured value we conclude that it is unlikely that we were measuring that parameter value. This is an example of *hypothesis testing*, in which we have hypothesized that we were measuring the specified parameter value and tested – with a confidence interval – whether it appears likely that hypothesis is correct. Statistical analyses to answer questions often take the form of hypothesis tests.

An IH/EHS example of this type of test, in which we compare a measured value to some known value, might be in comparing a company's incidence of worker injuries to an industry average. We would form a *null hypothesis* H_o such as "There is no difference between the company's injury incidence and the industry average." The statistical test then makes a probabilistic assessment of whether the null hypothesis is likely to be correct. If the test rejects the null hypothesis as being unlikely to be correct, we accept an *alternate hypothesis* H_A that in this example might be "The company's injury incidence is different from the industry average." Exactly how this is done is discussed in Chapter 4.

An understanding of what we mean by "unlikely to be correct" is essential. We expect to have some random variability among individual measurements of a quantity, so that there is some uncertainty as to what the "true" value is that we are measuring. Each measurement provides an *estimate* of the true value, which we understand to be the true value of what we are measuring plus or minus some amount of random variation. Statistical tests basically answer this question: "Given the assumed null hypothesis, how probable is it that we would have gotten the measurement that we did?"

In the following chapters, we explore the use of confidence intervals and learn a variety of parametric and nonparametric hypothesis testing techniques. Each is illustrated with examples from the IH/EHS field.

1.10 Summary

IH/EHS measurements are made for a variety of reasons, and the *who, what, when, where,* and *how* of measurements is driven by the *why*, that is, the question one is trying to answer. There is always variability in measurements, and statistics provides tools that allow us to account for this variability when making inferences, that is, to distinguish the "truth" obscured by the fog of imprecision. Many of these tools are based on the properties of the normal distribution or the chi-square distribution. An important question is: how much of the right kind of high quality data is needed to address a specific question? The answer of course is: "it depends." It depends on the question being asked, the characteristic being measured, and the statistical method used to conduct the data analysis. There must be enough data to overcome the inherent variability in the measured characteristic and to satisfy the minimum data requirements of the statistical technique. Failure to satisfy either of these requirements will result in insufficient statistical power in the analysis, which may be expressed as the ability to statistically demonstrate a difference when a real difference exists.

1.11 Addendum: Glossary of Some Useful Excel Functions

Statistical functions related to data distributions

BINOM.DIST(x, n, p, 0) Returns the binomial probability of observing exactly x successes in n trials, where the probability of success in any one trial is p.

BINOM.DIST(x, n, p, 1) Returns the cumulative binomial probability of observing *x or fewer* successes in n trials, where the probability of success in any one trial is p.

CHISQ.DIST(X^2, df, 0) Returns the value of the chi-square probability density function (the height of the curve) at X^2 for a function with df degrees of freedom.

CHISQ.DIST(X^2, df, 1) Returns the cumulative probability of the chi-square distribution at X^2 for a function with df degrees of freedom.

CHISQ.INV(p, df) Returns the inverse (the value of X^2) of the left-tailed probability p of the chi-squared function with df degrees of freedom, that is, the X^2 value for probability p to the left of X^2.

CHISQ.INV.RT(p, df) Returns the inverse (the value of X^2) of the right-tailed probability p of the chi-squared function with df degrees of freedom, that is, the X^2 value for probability p to the right of X^2.

F.DIST(F, df_1, df_2, 0) Returns the value of the F-distribution probability density function (the height of the curve) at F for a function with df_1 and df_2 degrees of freedom.

F.DIST(F, df_1, df_2, 1) Returns the cumulative probability to the left of F for the F-distribution with df_1 and df_2 degrees of freedom.

F.DIST.RT(F, df_1, df_2) Returns the probability to the right of F for the F-distribution with df_1 and df_2 degrees of freedom.

F.INV(p, df_1, df_2) Replaces FINV. Returns the inverse of the F-distribution (the value of F) for a probability p for a distribution that has df_1 and df_2 degrees of freedom.

NORM.DIST(z, μ, σ, 0) Replaces NORMDIST. Returns the value of the normal probability density function (the height of the curve) at position z from the mean for a normal distribution with mean μ and standard deviation σ. Provides the same return as NORM.S.DIST if $\mu = 0$ and $\sigma = 1$ are specified.

NORM.DIST(z, μ, σ, 1) Replaces NORMDIST. Returns the cumulative probability to the left of position z from the mean for a normal distribution with mean μ and standard deviation σ. Provides the same return as NORM.S.DIST if $\mu = 0$ and $\sigma = 1$ are specified.

NORM.S.DIST(z, 0) Replaces NORMSDIST. Returns the value of the standard normal probability density function (the height of the curve) at position z from the mean.

NORM.S.DIST(z, 1) Replaces NORMSDIST. Returns the cumulative probability to the left of z in the standard normal distribution. This function was used to generate Table A1. To obtain the probability contained between $\pm z$, use the cell equation NORM.S.DIST(z, 1)− NORM.S.DIST($-z$, 1). This approach was used to generate Table A2.

NORMS.S.INV(p) Replaces NORMSINV. Returns the inverse (the value of z) of the standard normal distribution for a cumulative probability of p to the left of z.

T.DIST(t, df, 0) Returns the value of the probability density function (the height of the curve) at t for a t-distribution with df degrees of freedom.

T.DIST(t, df, 1) Returns the cumulative probability to the left of t for a t-distribution with df degrees of freedom.

T.DIST.2T(t, df) Returns the two-tailed p-value for a calculated t-value for df degrees of freedom. You do not need to look up the critical t-value to decide significance – significance is shown if $p < \alpha$.

T.DIST.RT(t, df) Returns the probability p to the right of a calculated t-value for df degrees of freedom. This could be used for a one-sided test – significance is shown if $p < \alpha$.

T.INV(p, df) Replaces TINV. Returns the inverse of the t-distribution (the t-value) for a cumulative probability p with df degrees of freedom.

T.INV.2T(p, df) Returns the two-tailed critical t-value for $p = \alpha$ and df degrees of freedom, or the one-tailed critical t-value for $p = 2\alpha$ and df degrees of freedom.

Statistical functions that calculate statistics from a data set

AVERAGE(*cell*1 : *cell*2) Returns the arithmetic average (mean) of the range of numbers between two cells. The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6). AVERAGE can also be used for specific cells separated by commas, for example, AVERAGE(*cellx*, *celly*, *cellz*).

COVARIANCE.S(*array*1, *array*2) Returns the covariance of two columns or rows of paired numbers.

GEOMEAN(*array*) Returns the geometric mean of the range of numbers between two cells. The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

MAX(*array*) Returns the maximum value in the range of numbers between two cells. The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

MODE(*array*) Returns the most frequently occurring value in the range of numbers between two cells. The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

MEDIAN(*array*) Returns the median value in the range of numbers between two cells. The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

MIN(*array*) Returns the minimum value in the range of numbers between two cells. The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

PERCENTILE.EXC(*array*, *k*) Returns the *k*th percentile value in the distribution of numbers between two cells (*k* is a decimal fraction between 0 and 1). The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6). *k* is a fraction between 0 and 1, and the function calculates a new value if the percentile falls between two values in the distribution (see, e.g., the manual calculation in Figure 2.8). This function is *not the same* as the PERCENTILE.INC (or its older version PERCENTILE) function, which returns different values.

QUARTILE.EXC(*array*, *quart*) Returns the specified quartile value in the distribution of numbers between two cells. The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6). quart can have the value 0 (the minimum), 1 (first quartile or 25th percentile), 2 (second quartile or 50th percentile or median), 3 (third quartile or 75th percentile), or 4 (maximum value). This function is *not the same* as the QUARTILE.INC (or its older version QUARTILE) function, which returns different values and results in a narrower IQR.

RANK.AVG(*cellx*, *cell*\$1 : *cell*\$2, 1) Returns the rank of the value in a specified cell (*cellx*) of an array of cells between *cell*1 and *cell*2. Ranking is from *smallest to largest* value, and tied values are given average ranks. This is the one to use when constructing NED plots from Weibull plotting positions. Note the \$ to anchor the data array range when copying the equation to all the rows of the data matrix.

RANK.AVG(*cellx*, *cell*\$1 : *cell*\$2, 0) Returns the rank of the value in a specified cell (*cellx*) of an array of cells between *cell*1 and *cell*2. Ranking is from *largest to smallest* value, and tied values are given average ranks. Note the \$ to anchor the data array range when copying the equation to all the rows of the data matrix.

VAR.S(*array*) Replaces VAR. Returns the variance of the range of numbers between two cells, treating the data as a sample (not as a census). The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

VAR.P(*array*) Replaces VARP. Returns the variance of the range of numbers between two cells, treating the data as a census (not as a sample). The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6). This will rarely be used because we seldom work with census measures.

STDEV.S(*array*) Replaces STDEV. Returns the standard deviation of the range of numbers between two cells, treating the data as a sample (not as a census). The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

STDEV.P(*array*) Replaces STDEVP. Returns the standard deviation of the range of numbers between two cells, treating the data as a census (not as a sample). The range can be in a row (e.g., A1:A10), in a column (e.g., C5:C18), or it could be an array of numbers between two cells in different rows and columns (e.g., A1:D6).

STANDARDIZE($x, \overline{x}, s$) Calculates a z-score for the specified value x given the sample mean $\overline{x}$ and sample standard deviation s, that is, $z = (x - \overline{x})/s$.

Statistics related to correlation and regression

CORREL($x1_1, x2_n$) Same as PEARSON. Calculates the Pearson correlation coefficient for two columns of numbers $x1$ and $x2$. The columns must of course be the same size. $x1_1$ is the top value in the left column and $x2_n$ is the last value in the right column.

INTERCEPT(y_1, x_n) Calculates the y-axis intercept of a least squares regression line fitted to a column of y-values and their corresponding x-values. The columns must of course be the same size. y_1 is the top value in the left (y-values) column and x_n is the last value in the right (x-value) column.

PEARSON($x1_1, x2_n$) Same as CORREL. Calculates the Pearson correlation coefficient for two columns of numbers $x1$ and $x2$. The columns must of course be the same size. $x1_1$ is the top value in the left column and $x2_n$ is the last value in the right column.

RSQ(y_1, x_n) Calculates the coefficient of determination (R^2) of a least squares regression line fitted to a column of y-values and their corresponding x-values. The columns must of course be the same size. y_1 is the top value in the left (y-values) column and x_n is the last value in the right (x-values) column.

SLOPE(y_1, x_n) Calculates the slope of a least squares regression line fitted to a column of y-values and their corresponding x-values. The columns must of course be the same size. y_1 is the top value in the left (y-values) column and x_n is the last value in the right (x-values) column.

Other useful functions

COMBIN(n, k) Calculates the number of possible combinations of n objects taken k at a time.

PERMUT(n, k) Calculates the number of possible permutations (orderings) of n things taken k at a time.

PRODUCT(*array*) Calculates the product (Π) of the values in the array.

SUM(*array*) Calculates the sum (Σ) of the values in the array.

SUMSQ(*array*) Calculates the sum of the squares of each of the values in the array.

1.12 Exercises

1 In the bubble tube calibration example provided in the text, what are some of the minor random variations that could have contributed to the distribution of measured values? Note that there are several components involved – the air pump, the watch, and the observer – each of which is a system in itself.

2 A corporate industrial hygienist supports three manufacturing plants with different types of products and occupational hazards. Her field time spent per year at different types of tasks in the three plants is, on average, as shown in the table. There is no "multitasking" – only one task is performed at a time.

	Plant A	Plant B	Plant C	Totals
Air sampling	384	192	288	864
Noise surveys	240	96	220	556
PPE[a] fit testing	144	72	144	360
Worker training	82	90	68	240
Totals	850	450	720	2020

[a] Personal protective equipment such as air purifying respirators or ear plugs

(a) What is the marginal probability that on any given day she will be doing noise surveys?
Answer: 0.275

(b) What is the joint probability that she will be doing noise surveys in Plant B?
Answer: 0.0475

(c) What is the conditional probability that she will be doing fit testing if she is in Plant C that day?
Answer: 0.200

(d) Formally test whether the plant and task events are independent.

3 In an Excel spreadsheet, generate a list of z-values from −4 to +4 in Column A, compute the corresponding standard normal density in Column B via =NORM.S.DIST(z, 0), and produce a plot to visualize the probability density curve.

4 Using the appropriate z-table in Appendix A, determine the z-value corresponding to 0.995 cumulative probability. If you like, also calculate the value using Excel®. What is the probability under the normal distribution curve corresponding to the range $-z$ to $+z$ for this z-value?
Answer: 2.575; 0.9900

5 Using the appropriate z-table in Appendix A, determine the probability under the curve for the z range (a) $-\sigma$ to $+\sigma$ (± 1 standard deviation from the mean), (b) -2σ to $+2\sigma$, and (c) -3σ to $+3\sigma$.
Answer: 0.6826; 0.9544; 0.9974

6 For a normal probability distribution curve with $\mu = 3$ and $\sigma = 2.5$:

(a) Calculate the height of the curve at $x = 3.5$ and at $x = 3.7$
Answer: 0.15641; 0.15344

(b) Estimate the area under the curve between these two x-values (hint: the area is estimated from the area of the rectangle with sides $\Delta x = 0.2$ and the average height of the interval).
Answer: 0.030985

(c) Calculate a z-value corresponding to each of the x values using Equation 1.10 ($z = \frac{x-\mu}{\sigma}$).
Answer: 0.28

(d) Again estimate the area under the curve by taking the difference in cumulative probabilities for these two z-values, using either Appendix A or Excel, and compare it to your answer from part (b).
Answer: 0.0310

7 What is the cumulative probability associated with an $X^2 = 12$ if there are 3 df? Specifically, what is $P(X^2 \leq 12|k = 3)$?
Answer: 0.9926

8 What X^2 value would correspond to 95% cumulative probability if there are 5 df, that is, $P(X^2 \leq x|k = 5)$? What X^2 value would correspond to 99% cumulative probability?
Answer: For 95%, 11.0705 from Appendix B, or 15.0863 from Excel

9 For a population with a known mean $\mu = 105$ and standard deviation $\sigma = 22.5$, what is the X^2 value associated with measures of 95, 100, 115, and 125 from this population?
Answer: 1.235

10 You would like to conduct breathing zone air sampling on 3 workers from a SEG of 8 workers. How many combinations of 3 of the 8 workers are possible?
Answer: 56 groupings of 3 workers

11 You have identified 6 workers on whom you would like to conduct noise dosimetry measures over an 8-hour work period. However, you have only 1 dosimeter so you must spread the work over 6 days. How many ways are there of ordering the 6 workers?
Answer: 720

References

Bevington, P.R. and D.K. Robinson. Data Reduction and Error Analysis for the Physical Sciences, Second Edition. Boston: WCB McGraw-Hill, 1992.

Box, G.E.P., J.S. Hunter, and W.G. Hunter. Statistics for Experimenters: Design, Innovation, and Discovery, Second Edition. New York: John Wiley & Sons, 2005.

Daniel, W.W. and C.L. Cross. Biostatistics – A Foundation for Analysis in the Health Sciences, 10th Edition. New York: John Wiley & Sons, 2013.

Sheskin, D. Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition, Boca Raton, FL: Chapman & Hall, 2011.

Zar, J.H. Biostatistical Analysis, Fifth Edition, Upper Saddle River, NJ: Prentice Hall, 2010.