Probability is mostly organized common sense. However, being able to be specific about what probability is enables us to more accurately calculate probabilities and to employ theoretical statistical distributions to address con- fidence limits on data-derived quantities.

3.1 Probability Basics

In data analysis and hypothesis testing we are concerned with separating the probable from the possible. First, let us have a look at possibilities. In many situations we can either list all the possibilities or say how many such outcomes there are. In evaluating possibilities, we are often concerned with finding all the possible choices that are offered. Studying these choices leads us to the "multiplication of choices" rule:

> If a choice consists of k steps, of which the first can be made in n_1 ways and the k^{th} in n_k ways, the total number of choices is Πn_i, $i = 1,k$.

This can often be seen most clearly with a tree diagram (Figure 3.1). The number of choices here are $3 \times 4 = 12$.

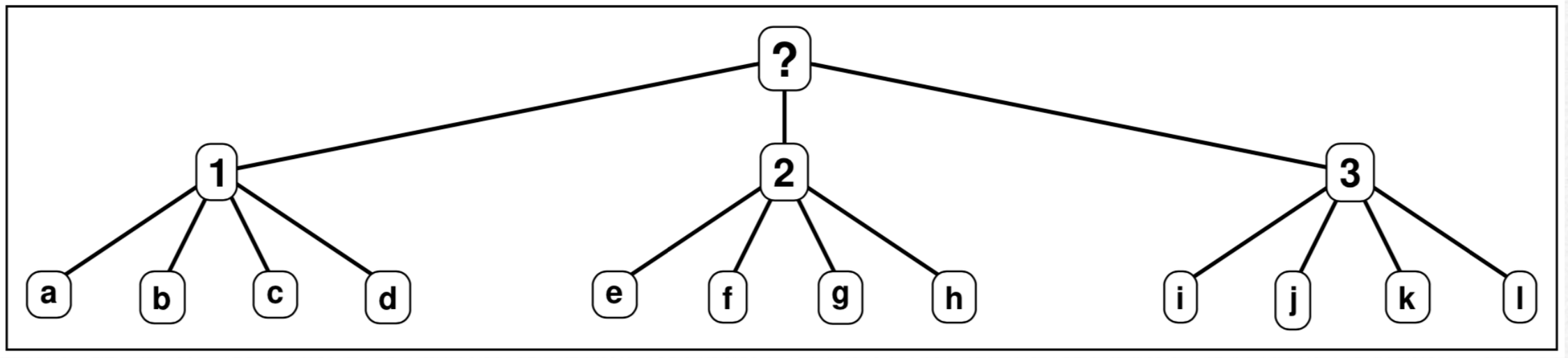

Figure 3.1: Tree diagram for illustrating all possible choices.

STATISTICAL CONCEPTS

3.1.1 Permutations

How many ways can we arrange *r* objects selected from a set of *n* distinct objects? This question applies to numerous statistical and probabilistic situations. We will first consider a simple example.

> Example 3–1. We have a tray with 20 water samples. How many ways can you select three samples from the 20? The first sample can be any of 20, the second will be any of the remaining 19, while the third is one of the remaining 18. The total ways must therefore be 20 × 19 × 18 = 6840.

We can write the number of choices as 20 × (20 − 1) × (20 − 2), and by induction we find

$$\text{ways} = n(n-1)(n-2)\ldots(n-r+1) = {}_nP_r. \tag{3.1}$$

It is convenient to introduce the factorial *n*!, defined as

$$n! = \prod_{i=1}^{n} i. \tag{3.2}$$

For convenience, we also define 0! to equal 1. We can then rewrite (3.1) as

$${}_nP_r = \frac{n(n-1)(n-2)\ldots(n-r+1)(n-r)(n-r-1)\ldots 1}{(n-r)(n-r-1)\ldots 1} = \frac{n!}{(n-r)!}. \tag{3.3}$$

This quantity is called the number of *permutations* of *r* objects selected from a set of *n* distinct objects.

> Example 3–2. We wish to determine how many different hands one can be dealt in a game of poker. With *n* = 52 (total number of cards in the deck) and *r* = 5 (number of cards in a hand), we find
>
> $${}_{52}P_5 = \frac{52!}{(52-5)!} = \frac{52!}{47!} = 48\cdot 49\cdot 50\cdot 51\cdot 52 = 3\cdot 10^8. \tag{3.4}$$
>
> However, this calculation assumes that the *order* in which you receive the cards is important.

STATISTICAL CONCEPTS

3.1.2 Combinations

In many situations we do not care about the exact ordering of the *r* objects, i.e., *abc* is the same choice as *acb* for our purpose. In general, *r* objects can be arranged in *r*! different ways ($_rP_r = r!$). Since we are only concerned about *which r* objects have been selected and not their order, we can use $_nP_r$ but must now normalize the result by *r*!, i.e.,

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!} = \binom{n}{r}. \tag{3.5} $$

The quantity $_nC_r$ is called the number of *combinations*, and the factors $\binom{n}{r}$ are called the *binomial coefficients*. After $_r$ picking the *r* objects, *n* − *r* objects are left, so consequently there are as many ways of selecting *n* − *r* objects from *n* as there are of selecting *r* objects, i.e.,

$$ \binom{n}{r} = \binom{n}{n-r}. \tag{3.6} $$

Example 3-3. How many ways can you select three tide gauge records from 10 available stations? This is a question of combinations.

$$ {}_{10}C_3 = \binom{10}{3} = \frac{10!}{3!7!} = \frac{8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3} = 8 \cdot 3 \cdot 5 = 120. \tag{3.7} $$

Likewise, per (3.6), there are also 120 ways to select 7 tide gauge records from the same 10 stations.

3.1.3 Probability

So far we have studied only what is *possible* in a given situation. We have listed all possibilities or determined how many possibilities there are. However, to be of use to us we need to be able to judge which of the possibilities are *probable* and which are *improbable*. The basic concept of probability can be stated thus: If there are *n* possible outcomes or results, and *s* of those are regarded as favorable (or as “successes”), then the probability of success is given by

$$ P = s/n. \tag{3.8} $$

This classical definition applies only when all possible outcomes are *equally likely*.

STATISTICAL CONCEPTS

Example 3–4. What is the probability of drawing an ace from a deck of cards? *Answer*: $P = 4/52 = 1/13 = 7.7\%$. How about getting a 3 *or* a 4 with a balanced die? *Answer*: $s = 2$ and $n = 6$, so $P = 2/6 = 33\%$

While equally likely possibilities are found mostly in games of chance, the classical probability concept also applies to random selections, such as making selections to reduce a large set of data down to a manageable quantity without introducing sampling bias.

Example 3–5. If three of 20 water samples have been contaminated and you select four random samples, what is the probability of picking one of the bad samples?

Answer: We have $\binom{20}{4} = 3 \cdot 5 \cdot 17 \cdot 19 = 4845$ ways of making the selection of our four samples. The number of "favorable" outcomes is $\binom{17}{3}$ [we pick three good samples of the 17 good ones] times $\binom{3}{1}$ [we pick one of the three bad samples] = 2040. It then follows that the probability is $P = s/n = 2040/4845 = 42\%$. Here we used the rule of multiplicative choices.

Obviously, the classical probability concept will not be useful when some outcomes are more likely than others. A better definition would then be

The probability of an event is the proportion of the time that events of the same kind will occur in the long run.

So, when the National Weather Service says that the chance of rain on any day in June is 0.2, it is based on past experiences that on average June had 6 days of rain. Another important probability theorem is the *law of large numbers*, which states

If a situation, trial, or experiment is repeated again and again, the proportion of successes will tend to approach the probability that any one outcome will be a success.

which is basically our probability concept in reverse.

Coin tosses illustrate the law of large numbers nicely. We toss the coin and keep track of how many times we get "heads" versus the total number of tosses. For a nice symmetric coin we expect the proportion of heads to total tosses to approach 0.5 over the long haul, but initially we are not surprised that there can be large departures from this expectation. Figure 3.2 shows how the proportion may oscillate for a small number of tosses but eventually it will approach the expected value.

STATISTICAL CONCEPTS

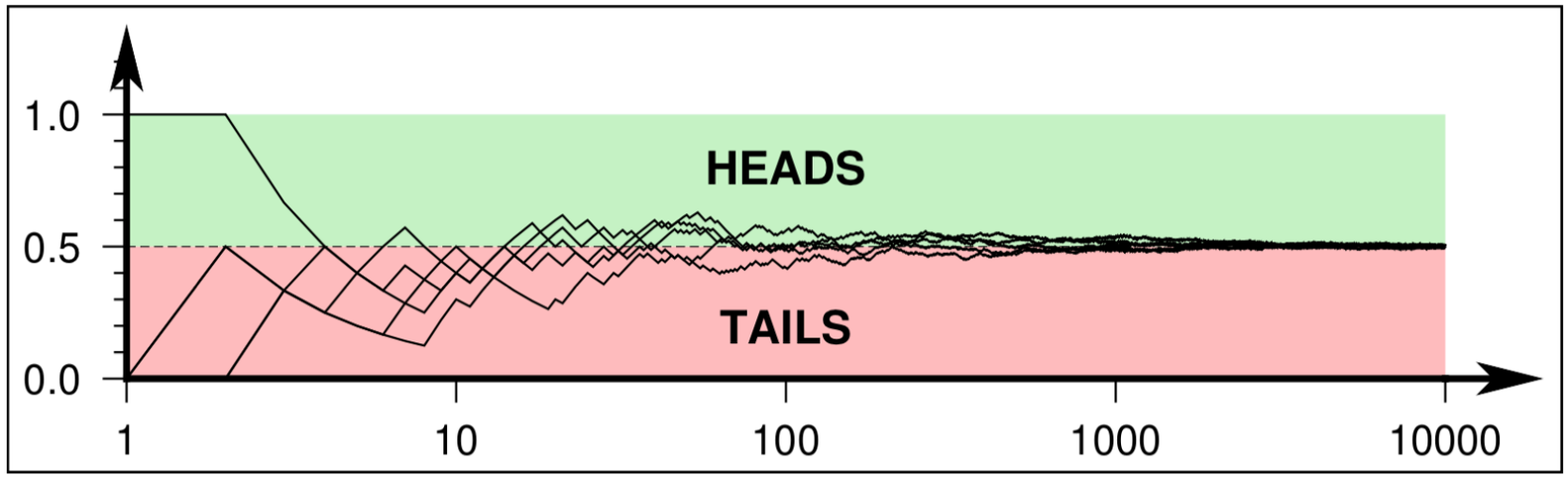

Figure 3.2: Proportion of heads in a series of coin tosses. The more tosses we complete, the closer the ratio of heads to total tosses will approach 0.5. Shown are five separate sequences. They differ considerably for small numbers but all converge on the expected proportion.

3.1.4 Some rules of probability

In statistics, the set of all possible outcomes of an experiment is called the *sample space*, usually denoted by the letter *S*. Any subset of *S* is called an *event*. An event may contain more than one item. Sample spaces may be finite or infinite. Two events that have no elements in common are said to be *mutually exclusive*, meaning they cannot both occur at the same time. There are only positive (or zero) probabilities, symbolically written

$$P(A) \geq 0 \tag{3.9}$$

for any event *A*. Every sample space has probability 1, so that

$$P(S) = 1, \tag{3.10}$$

where $P = 1$ means absolute certainty. If two events are mutually exclusive, the probability that one *or* the other will occur equals the sum of their probabilities

$$P(A \cup B) = P(A) + P(B). \tag{3.11}$$

STATISTICAL CONCEPTS

Regarding the notation, ∪ means *union* (which we read as "OR"), ∩ means *intersection* ("AND"), and ′ (the prime symbol) means *complement* ("NOT"). We can furthermore state that

$$P(A) \leq 1, \tag{3.12}$$

since absolute certainty is the most we can ask for. Also,

$$P(A) + P(A') = 1, \tag{3.13}$$

since it is certain that an event either will or will not occur.

3.1.5 Probabilities and odds

Bookmakers in London use a slightly different system of reporting probabilities. If the probability of an event is p, then the *odds* for its occurrence are

$$a : b = \frac{p}{1-p}. \tag{3.14}$$

The inverse relation gives

$$p = \frac{a}{a+b}. \tag{3.15}$$

If you are still reading this book then odds are you will pass this course!

3.1.6 Addition rules

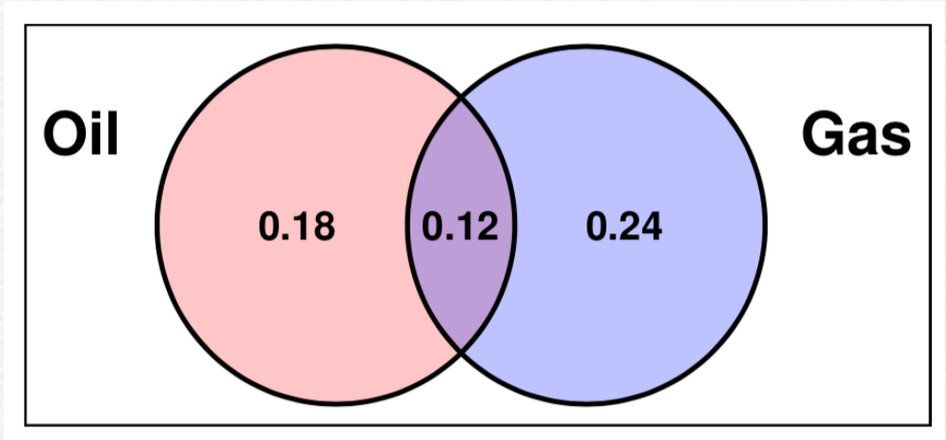

Figure 3.3: A Venn diagram illustrating the probabilities of finding hydrocarbons. The overlapping magenta wedge graphically represents the probabilities of finding *both* oil and gas.

STATISTICAL CONCEPTS

The addition rules demonstrated above only holds for *mutually exclusive events*. Let us now consider a more general case. The sketch in Figure 3.3 is a *Venn diagram*, a handy graphical way of illustrating the various combinations of possibilities and probabilities. The diagram illustrates the probabilities associated with finding hydrocarbons during a hypothetical exploration campaign. We see from the diagram that

$$\begin{aligned} P(\text{oil}) &= 0.18 + 0.12 = 0.3, \\ P(\text{gas}) &= 0.24 + 0.12 = 0.36, \\ P(\text{oil} \cup \text{gas}) &= 0.18 + 0.12 + 0.24 = 0.54. \end{aligned} \tag{3.16}$$

Now, if we used the simple addition rule (3.11), we would find

$$P(\text{oil} \cup \text{gas}) = P(\text{oil}) + P(\text{gas}) = 0.3 + 0.36 = 0.66. \tag{3.17}$$

This value overestimates the probability, because finding oil and finding gas are *not* mutually exclusive since we might find both. We can correct the equation by writing

$$P(\text{oil} \cup \text{gas}) = P(\text{oil}) + P(\text{gas}) - P(\text{oil} \cap \text{gas}) = 0.3 + 0.36 - 0.12 = 0.54. \tag{3.18}$$

The general addition rule for probabilities thus becomes

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{3.19}$$

Note that if the events *are* mutually exclusive then $P(A \cap B) = 0$ and we recover the original rule.

3.1.7 Conditional probability and Bayes basic theorem

We must sometimes evaluate the probability of an event *given that another event already has occurred*. We write the probability that A will occur given that B already has occurred as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.20}$$

STATISTICAL CONCEPTS

In our exploration example, we can find the probability of finding oil given that gas already has been found as

$$P(\text{oil}|\text{gas}) = \frac{P(\text{oil} \cap \text{gas})}{P(\text{gas})} = \frac{0.12}{0.36} = \frac{1}{3}. \tag{3.21}$$

We can now derive a general multiplication rule from (3.20) by multiplying it by $P(B)$ and exchange A and B, which gives

$$\begin{aligned} P(A \cap B) &= P(B)P(A|B) \\ P(A \cap B) &= P(A)P(B|A) \end{aligned} \tag{3.22}$$

and implies that the probability of both events A and B occurring is given by the probability of one event occurring multiplied by the probability that the other event will occur given that the first one already has occurred (occurs, or will occur). This rule is called the *joint probability* or *Bayes basic theorem*. Now, if the events A and B are independent events, then the probability that A will take place is not influenced by whether B has taken place or not, i.e.

$$P(A|B) = P(A). \tag{3.23}$$

Substituting this expression into (3.22) we obtain

$$P(A \cap B) = P(A) \cdot P(B). \tag{3.24}$$

That is, the probability that two independent events A and B both will occur equals the product of their probabilities. In general, for n independent events with individual probability p_i, the probability that all n events occur is

$$P = \prod_{i=1}^{n} p_i. \tag{3.25}$$

Example 3–6. What is the probability of rolling three ones in a row with a balanced die?
Answer: With n=3 and p=1/6,

$$P = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \approx 0.005. \tag{3.26}$$

STATISTICAL CONCEPTS

While $P(A|B)$ and $P(B|A)$ may look similar, they can be vastly different. For example, let A be the event of a death on the Bay Bridge connecting San Francisco and Oakland, and B the event of a magnitude 8 earthquake in the area. Then, $P(A|B)$ is the probability of a fatality on the Bay Bridge *given* that a large earthquake has taken place nearby, while $P(B|A)$ is the probability that we will have a magnitude 8 quake *given* that a death has been reported on the bridge. Clearly $P(A|B)$ seems more likely than $P(B|A)$ since we know the former to have happened in the past. On the other hand, we can list many causes of fatalities on the freeway other than earthquakes (e.g., traffic accidents, heart attacks, old, age, road rage, talk radio rants, and so on).
We can arrive at a relation between $P(B|A)$ and $P(A|B)$ by equating the two expressions for $P(A \cap B)$ in (3.22). We obtain $P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$, or

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}. \tag{3.27}$$

This is a useful relation since we may sometimes know one conditional probability but are interested in the inverse relationship. For example, we may know that salt domes (known as potential traps for hydrocarbons) often are associated with large curvatures in the gravity field. However, we may be more interested in the converse: Given that large curvatures in the gravity field exist, what is the probability that salt domes are the cause of such anomalies?

3.1.8 Bayes general theorem

If there are more than one event B_i (all mutually exclusive) that are conditionally related to an event A, then $P(A)$ is simply the sum of the conditional probabilities of the events B_i times their individual probabilities, i.e.

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i). \tag{3.28}$$

Substituting (3.28) into (3.27) gives, for any of the n events B_i,

$$P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{\sum_{j=1}^{n} P(A|B_j) \cdot P(B_j)}. \tag{3.29}$$

STATISTICAL CONCEPTS

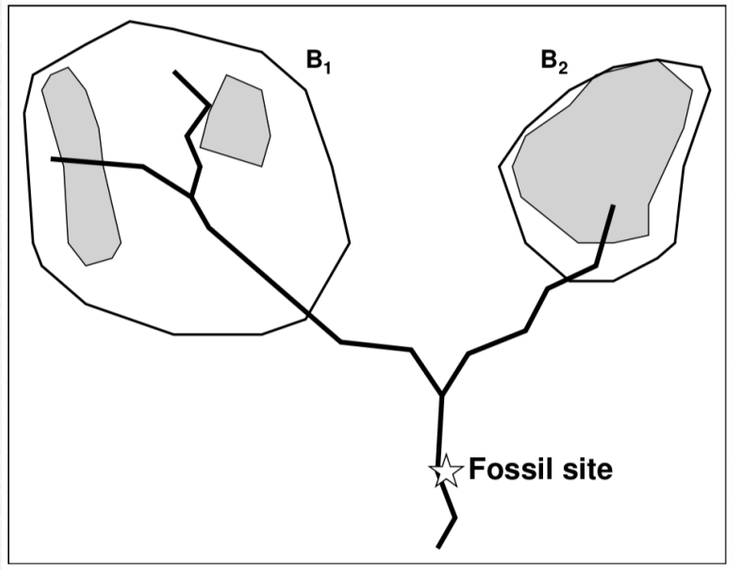

Figure 3.4: Location of a fossil discovery with respect to the two drainage basins from which it must have originated. Bayes theorem provides a formal way to assign likelihood to the possible origins.

This is the general *Bayes theorem*.

Example 3–7. Let us assume that an unknown marine fossil fragment was found in a dry stream bed in northern Sahara. Excited, a paleontologist would like to send out an expendable graduate student field party to search for a more complete specimen of the unknown species. Unfortunately, the source of the fragment cannot be identified uniquely since it was found several kilometers below the junction of two dry stream tributaries (Figure 3.4). The drainage basin B_1 of the larger stream covers 407.5 km^2, while the other basin (B_2) covers only 207.5 km^2. Based on this difference in basin size alone we might expect the probabilities that the fragment came from one of the basins are

$$P(B_1) = \frac{407.5}{615} = 0.66, \qquad P(B_2) = \frac{207.5}{615} = 0.34, \tag{3.30}$$

based solely on the proportion of each basin's area to the combined area. However, inspecting an ancient British- produced geological map reveals that only 31% of the outcropping rocks in the larger basin B_1 are marine, whereas almost 85% of the outcrops in basin B_2 are marine. We can now state two conditional probabilities:

$P(A|B_1) = 0.31$ (Probability of a marine fossil, given it was derived from basin B_1.)
$P(A|B_2) = 0.85$ (Probability of a marine fossil, given it was derived from basin B_2.)

With these probabilities and Bayes general theorem (3.29) we can find the conditional probability that the fossil came from basin B_1 given that the fossil is marine:

$$P(B_1|A) = \frac{P(A|B_1)\cdot P(B_1)}{P(A|B_1)\cdot P(B_1) + P(A|B_2)\cdot P(B_2)} = \frac{0.31\cdot 0.66}{0.31\cdot 0.66 + 0.85\cdot 0.34} = 0.41. \quad (3.31)$$

Consequently, the probability of the fossil coming from the smaller basin B_2 is the complimentary probability

$$P(B_2|A) = 0.59. \quad (3.32)$$

It therefore seems somewhat more likely that the smaller basin was the source of the fossil and that this area should be the initial target for the student-led expedition. However, $P(B_1|A)$ and $P(B_2|A)$ are not dramatically different and depends to some extent on the assumptions used to select $P(B_1)$ and $P(A|B_1)$ in the first place. Bayes general theorem is extensively used in such search and find scenarios and the probabilities that go into the procedure are constantly being revised as more is learned during the search.

3.2 The M&M's of Statistics

When discussing exploratory data analysis we mentioned that it is useful to be able to present large data sets using just a few parameters. We saw the box-and-whisker diagram graphically summarized a data distribution. However, it is often desirable to represent a data set by a *single* number which, in its way, is descriptive of the entire data set. We will see there are several ways to select this "representative" value. We will mostly be concerned with measures that somehow describe the center or middle of the data set. These are called estimates of *central location*.

3.2.1 Population and samples

If a data set consists of all conceivably possible (or hypothetically possible) observations of a certain phenomenon then we call it a *population*. A population can be finite or infinite. Any subset of the population is called a *sample*. Thus, a series of 12 coin-tosses is a sample of the potentially unlimited number of tosses in the population. We will most often find that we are analyzing samples taken from a much larger population, and our aim will be to learn something about the population by studying the smaller sample set (Figure 3.5).

STATISTICAL CONCEPTS

3.2.2 Measures of central location (mean, median, mode)

The best known estimate of central location is called the *arithmetic mean*, defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (3.33)$$

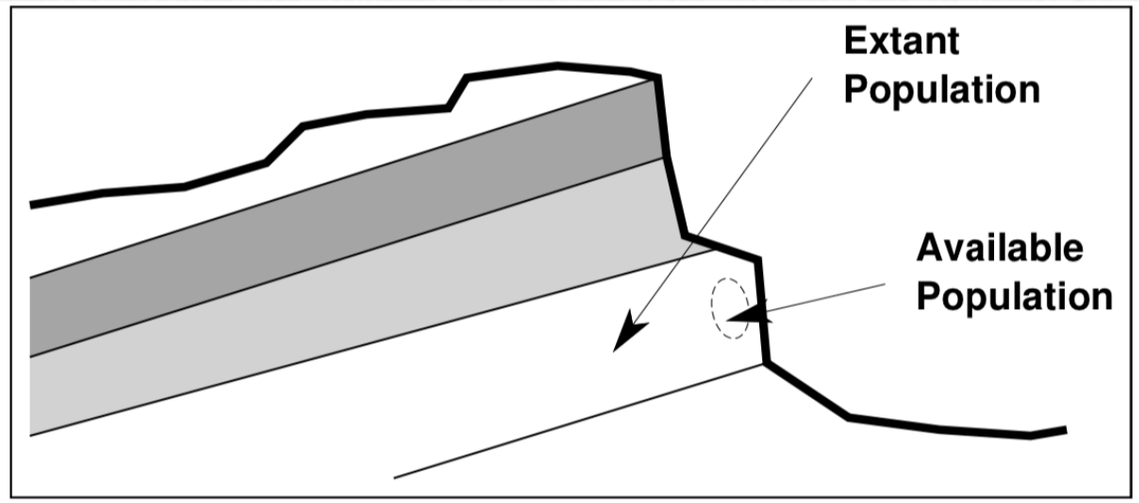

Figure 3.5: We must always try to select an unbiased sample from the population. In this example we are sampling the weathered outcrop of a sedimentary layer, which most likely is not representative of the entire formation.

The mean is also loosely called the “average.” Resist being that sloppy! When reporting the mean value, always say “mean” and not “average” so that the reader knows exactly what you have done. We call $\bar{x}$ the *sample mean* to distinguish it from the true mean of the population, denoted

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad (3.34)$$

which likely will remain unknown to us. The mean has many useful properties, which explains its common use:

- It can always be calculated for any numerical data, i.e., it always exists.
- It is unique and straightforward to calculate.
- It is relatively stable and does not fluctuate much from sample to sample taken from the same population.
- It lends itself to further statistical treatment: several $\bar{x}$ estimates from subgroups can later be combined into an overall grand mean.
- It takes into account every data value.

STATISTICAL CONCEPTS

However, the last property can sometimes be a liability. Should a few points deviate excessively from the bulk of the data then it does not make sense to include them in the sample. A better estimate for the central location may then be the *sample median*:

$$\text{median } x_i = \tilde{x} = \begin{cases} x_{n/2+1}, & n \text{ is odd} \\ \frac{1}{2}(x_{n/2+1} + x_{n/2}), & n \text{ is even} \end{cases} \tag{3.35}$$

Here, the data first must be sorted into ascending (or descending) order. We then choose the middle value (or mean of the two middle values for even n) as our median estimate.

Consider this sample of sandstone densities: {2.30, 2.20, 2.35, 2.25, 2.30, 23.0, 2.25}, n = 7. The median density can be found to be $\tilde{x}$ = 2.30, a reasonable value, while the mean density $\bar{x}$ = 5.24, which is a rather useless estimate since it is clearly far outside the bulk of the data *and* outside the range of known sandstone densities anywhere. For this reason we say that the median is a *robust* estimate of central location. Here it is rather obvious that the value 23.0, which probably is a clerical error, threw off the mean and we could correct for that by excluding it from the calculation and find $\bar{x}$ = 2.28 instead. However, in many cases our data set will be very large and we must anticipate that some values may be erroneous.

The disadvantage of the median is the need to sort the data, which can be slow. (Do you think this is really a valid reason not to use it?). However, like the mean, the median always exists and is unique.

Our final traditional estimate for central location is the *mode*. The mode is defined as the observation that occurs the most frequently. For defining the central location the mode is at a disadvantage since it may not exist (perhaps no two values are the same) or it may not be unique (our densities actually have two modes). Of course, if our data set is expected to have more than one "peak," modal estimates are important, and we will return to that later. The mode will be denoted as $\hat{x}$. The mean, median and mode of a distribution typically are related as indicated in Figure 3.6.

STATISTICAL CONCEPTS

Returning to the mean, it is occasionally the case that some measurements are considered more important than others. It could be that some observations were made with a more precise instrument, or simply that some values are not as well documented as others. These are examples of situations where we should use a *weighted mean*

$$\bar{x} = \sum_{i=1}^{n} w_i x_i \Big/ \sum_{i=1}^{n} w_i \ , \qquad (3.36)$$

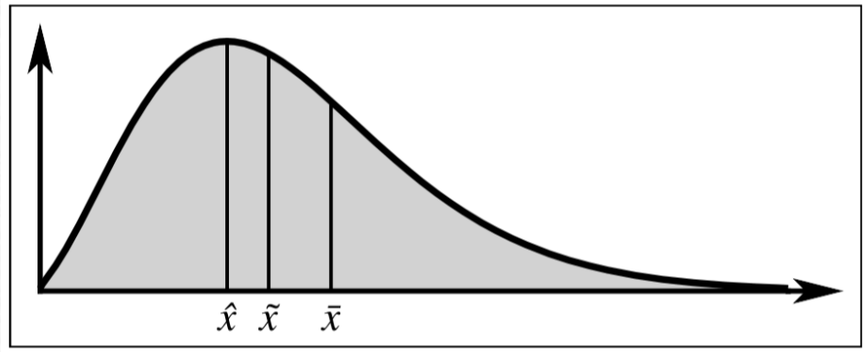

Figure 3.6: The relationship between the mean, median, and mode estimates of central location for a skewed data distribution. These estimates will all coincide for a perfectly symmetric and unimodal distribution.

where w_i is the weight of the i'th data value. If all $w_i = 1$ then we recover the original definition for the mean (3.33). This general equation is also convenient when we need to compute the overall, or *grand mean* based on the individual means from several data sets. The grand mean based on m data sets may be written as

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}, \qquad (3.37)$$

where the sample sizes n_i take the place of the weights in (3.36).

3.2.3 Measures of variation

While a measure of central location is an important attribute of our data, it says little about how the data are distributed. We need some way of representing the *variation* of our observations about the central location. In the EDA section, we used the *range* and *hinges* to indicate data variability. Another way to define the variability would be to compute the deviations from the mean,

$$\Delta x_i = x_i - \bar{x}\ , \qquad (3.38)$$

STATISTICAL CONCEPTS

and take the average of the sum of deviations, $\frac{1}{n}\sum_{i=1}^{n}\Delta x_i$. Sadly, it turns out that this sum is always zero, which makes it rather useless for our purposes. A more useful quantity might be the mean of the absolute value (*AD*) of the deviations:

$$AD = \frac{1}{n}\sum_{i=1}^{n}|\Delta x_i|. \tag{3.39}$$

Because of the absolute value sign this function is nonanalytic and often completely ignored by statisticians. You will find very superficial treatment of medians and absolute deviations in most elementary statistics books. However, when dealing with real data that include occasional bad values, the *AD* is useful, just as the median can be more useful than the mean. However, the most common way to describe variation of a population is to define it as the average *squared* deviation. Hence, the population *variance* is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2, \tag{3.40}$$

and the population *standard deviation* is therefore

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}. \tag{3.41}$$

Most often we will be working with samples rather than entire populations, and we hope (and will later test) that the sample is representative of the population. The sample standard deviation *s* is given by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{3.42}$$

Note that we are dividing by $n-1$ rather than by n. This is done because $\bar{x}$ must first be *estimated* from the sample rather than being a *given* parameter of the population, such as μ and N. This reduces the degrees of freedom by one; hence we divide by $n-1$ (we will have more to say about degrees of freedom in Section 4.2.1). We can now show one property of the mean: It is clear that s^2 depends on the choice for $\bar{x}$. Let us find the value for $\bar{x}$ in (3.42) that gives the smallest value for s^2.

STATISTICAL CONCEPTS

Consider

$$f(\bar{x}) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.43}$$

The function *f* has a minimum where $df/d\bar{x} = 0$ and $d^2f/d\bar{x}^2 > 0$,so we find

$$\frac{df}{d\bar{x}} = \frac{\sum_{i=1}^{n} -2(x_i - \bar{x})}{n-1} = \frac{-2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) = 0, \tag{3.44}$$

which gives

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0. \tag{3.45}$$

We can solve this equation and find

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.46}$$

Since

$$\frac{d^2f}{dx^2} = \frac{2n}{n-1} > 0, \tag{3.47}$$

we know that *f* has a minimum for this value of $\bar{x}$. Thus, we have shown that the value $\bar{x}$ that minimizes the standard deviation equals the mean we defined earlier in (3.33). This is a very useful and important property of the mean. Because $\bar{x}$ minimizes the squared "misfit", it is also called the *least-squares estimate* of central location (or L_2 estimate for short). When computing the mean and standard deviation on a computer we do not normally use (3.42) since it requires two passes through the data: One to compute the $\bar{x}$ and another to solve (3.42). Rather, we rearrange (3.42) to give

STATISTICAL CONCEPTS

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

$$= \sqrt{\frac{n\sum x_i^2 - 2n\bar{x}\sum x_i + n\sum \bar{x}^2}{n(n-1)}} = \sqrt{\frac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)}}. \quad (3.48)$$

3.2.4 Robust estimation

We found that the arithmetic mean is the value that minimizes the sum of the squared deviations from the central value. Can we apply the same argument to the mean absolute deviation and find what the best value for $\tilde{x}$ may be? In other words, let

$$\frac{d}{d\tilde{x}}\left(\frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \tilde{x}}{|x_i - \tilde{x}|} = 0. \quad (3.49)$$

The term inside the summation can only take on the values −1, 0, or +1. Thus, the only $\tilde{x}$ that can satisfy (3.49) is a value chosen such that half the x_i are smaller (giving −1) and half the x_i are larger (giving +1), and for odd sample sizes we also get one or more exact zeros. Thus, we have proven that the median is the location estimate that minimizes the mean absolute deviation. The median is also called the L_1-estimate of central location.

The discussion of mean and median brings up the general issue of *robust estimation*: How to calculate a stable and reasonable estimate of central location in the presence of contaminated data? As an indicator of how robust a method is, we will introduce the concept of "breakdown point." It is the *smallest fraction* of the observations that must be replaced by outliers in order to throw the estimator outside reasonable bounds.

STATISTICAL CONCEPTS

We have already seen that even a single bad value is enough to throw the mean way off. For our densities of sandstone, we had $\rho = \{2.2, 2.25, 2.25, 2.3, 2.3, 2.35, 23.0\}$, with $n = 7$. If we realized that 23.0 should be 2.3, we find $\bar{\rho} = 2.28 \pm 0.05$, while if we included $\rho_7 = 23.0$ we would find $\bar{\rho} = 5.24 \pm 7.8$. The second estimate is obviously far outside the 2.20–2.35 range we first determined. We can therefore say that the least squares estimate (i.e., the mean) has a breakdown value of $1/n$; it only takes one outlier to ruin our day. On the other hand, note that the median is ~ 2.3 in both cases, well inside the acceptable interval. It is found that the breakdown point of the median is 50%: We would have to replace half the data with bad outliers to move the estimate of the median outside the range of the original (good) data values.

Apart from the central location estimator, we also want a robust estimate of the spread of the data. Clearly, the classical standard deviation is problematic since only one bad value will make it biased due to the x^2 effect. From the success of taking the median of a string of numbers rather than summing them up, could we do something similar with the deviations? Consider what value of $\tilde{x}$ would minimize the median of $\{|x_i - \tilde{x}|\}$. You can probably see for yourselves that the $\tilde{x}$ must equal our old friend the median. Because of the robustness of the median operator, we will often use the quantity called the *median absolute deviation* (*MAD*) as our robust estimate of "spread" or variation. Note: Many textbooks and software packages (such as MATLAB) use *MAD* to indicate *mean absolute deviation* instead, as defined in (3.39) and called *AD* in these notes. Thus, we define

$$MAD = 1.4826 \text{ median } |x_i - \tilde{x}|, \tag{3.50}$$

where the factor 1.4826 is a correction term that makes the *MAD* equal to the standard deviation of normally distributed data*. Like the median, the *MAD* has a breakdown point of 50%. The *MAD* for our example was 0.07 and it remained unchanged by using the contaminated value. Having robust estimates of central location and scale, we can attempt to identify *outliers*. We may compute the robust *standard units*

$$z_i = \frac{x_i - \tilde{x}}{MAD} \tag{3.51}$$

and compare them to a cutoff value: If $|z_i| > z_{cut}$ we say we have detected an outlier. The choice for z_{cut} is to a certain extent arbitrary. It is, however, quite standard to choose $z_{cut} = 2.5$. Chances that any z_i will exceed z_{cut} is very small if the z_i's came from a normal distribution. Our normalized densities (including the contaminated value) using $\bar{x}$ and s to compute z_i gives

*This factor equals $1/P^{-1}(0.75)$, where $z = P^{-1}(p)$ is the inverse cumulative normal distribution.

$$z_{L2} = \{-0.39, -0.38, -0.38, -0.377, -0.377, -0.37, 2.28\}, \tag{3.52}$$

where none of the values qualify as an outlier. Using the median and *MAD* instead, we find

$$z_{L1} = \{-1.35, -0.68, -0.68, 0.0, 0.0, 0.68, 280.0\}, \tag{3.53}$$

and we see that the bad observation gives a huge *z*-value two orders of magnitude larger than any other. Clearly, the least-squares technique alone is not trustworthy when it comes to detecting bad points. The outlier-detecting scheme presents us with an elegant two-step technique: First find and remove the outliers from the data, then use classical *least-squares* techniques on the remaining data points. The resulting statistics are called the *least trimmed squares* estimates (LTS). We will return to the concept of robustness when discussing regression in Chapter 6.

3.2.5 Central limit theorem

How well does our sample mean, $\bar{x}$, compare to the true population mean, μ? An important theorem, called the *central limit theorem*, states

> *If n (the sample size) is large, the theoretical sampling distribution of the mean can be approximated closely with a normal distribution.*

This is rather important since it justifies the use of the normal distribution in a wide range of situations. It simply states that the sample mean $\bar{x}$ is an *unbiased estimate* of the population mean and that the scatter about μ is *normally distributed*. It can be shown that the standard deviation of the sampling mean, $s_{\bar{x}}$, is related to the population deviation, σ, by

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3.54}$$

or

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{3.55}$$

Depending on whether the population is infinite (3.54) or finite of size *N* (3.55). Thus, as *n* grows large, $s_{\bar{x}} \to 0$.

Furthermore, the sample variance s_2 has the mean value σ_2 with standard deviation

$$\sigma_s^2 = \frac{2\sigma^4}{n-1}, \tag{3.56}$$

which also → 0 for large *n*. For our analysis we will substitute the sample standard deviation *s in lieu* of the unknown population standard deviation σ, since *s* is an *unbiased estimator* of σ.

3.2.6 Covariance and correlation

We found earlier that the sample variance was defined as

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{n-1}. \tag{3.57}$$

It is often the case that our data set consists of pairs of properties, such as sets of (depth, pressure), (time, temperature), concentrations of two elements, and more. Denoting the paired properties by *x* and *y*, we can compute the variance of each quantity separately. For instance, for *y* we find

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})}{n-1}. \tag{3.58}$$

We can now define the *covariance* between *x* and *y* in a similar way as

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}. \tag{3.59}$$

While s_x and s_y tell us how the *x* and *y* values are distributed *individually*, s_{xy} tells us how the *x* and *y* values vary *together*.

STATISTICAL CONCEPTS

Because the value of the covariance clearly depends on the units of x and y, it is difficult to state what covariance values are meaningful. This difficulty is overcome by defining the Pearson *correlation coefficient* r, which normalizes the covariance to yield correlations in the ±1 range, i.e.,

$$r = \frac{s_{xy}}{s_x s_y}. \tag{3.60}$$

If $|r|$ is close to 1, then the variables are strongly correlated or anti-correlated. Values of r close to 0 mean that there is little significant correlation between the data pairs. Figure 3.7 shows some examples of data pairs and their correlations. We see that in general, r will tell us how well the data are "clustered" in some direction. Note in particular example (f), which presents data that are clearly correlated (i.e., all pairs lie on a circle), yet $r = 0$. This occurs because r is a measure of a *linear* relationship between values; a nonlinear relationship may not register a significant correlation. Thus, we must be careful with how we use r to draw conclusions about the interdependency of paired values. For example, if our (x, y) data are governed by a $y = \sqrt{x}$ law then we may find a fairly good correlation between x and y, but we would be wrong to conclude that x and y have a *linear* relationship (plotting y versus $\sqrt{x}$ *would give* a linear relationship and a much higher value of r). We will return to correlation under the rubrics of curve fitting and multiple regression in Chapter 6.

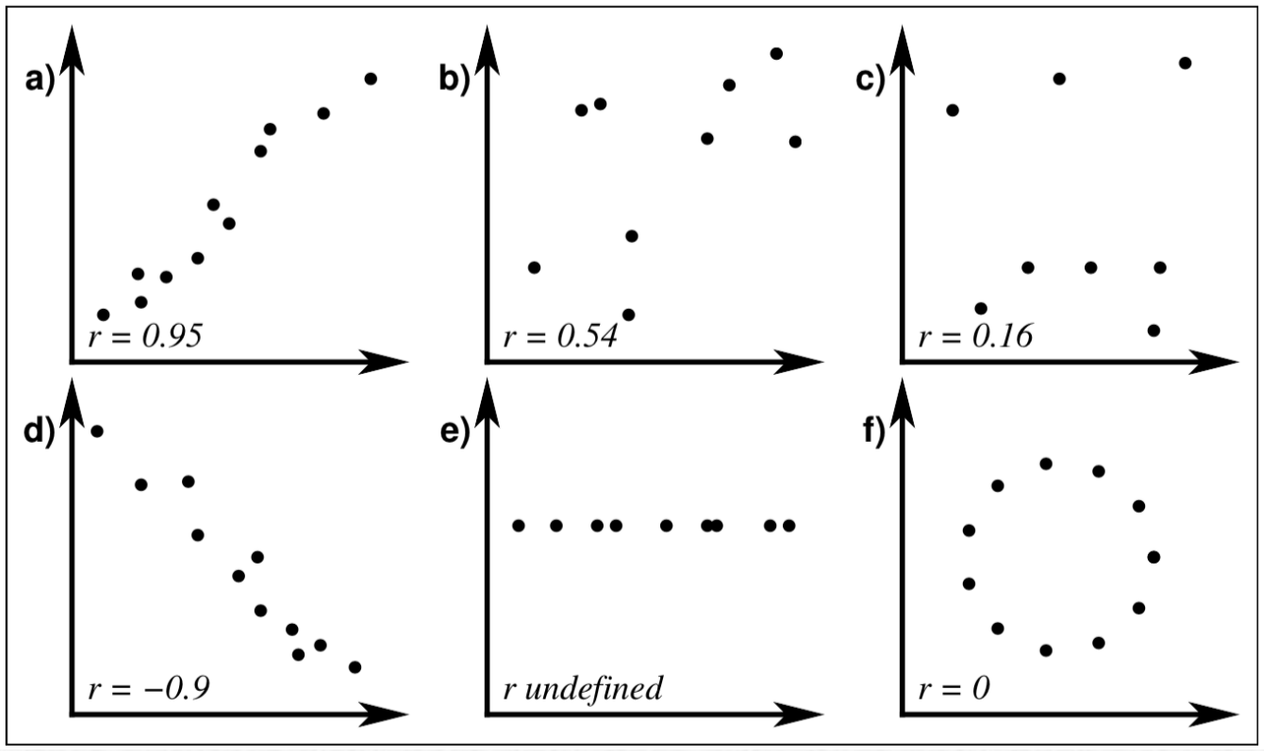

Figure 3.7: Some examples of data sets and their correlation coefficients. Note that the perfect circular correlation in (f) gives a zero linear correlation coefficient. While clearly x and y are correlated, their relationship is not *linear*.

STATISTICAL CONCEPTS

3.2.7 Moments

Returning to the L_2 estimates, we will briefly introduce the concept of *moments*. In general, the *r*'th moment is defined as

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^r, \tag{3.61}$$

except for *r* = 1 where it is customary to use the "raw moment" about zero instead. From this definition it can be seen that the mean and variance are the first (raw) and second (central) moments, respectively. We will look at two higher order (central) moments that one may encounter in the literature. The first is called the *skewness* (*SK*) and it is the third central moment, given by

$$SK = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{1}{n}\sum_{i=1}^{n} z_i^3, \tag{3.62}$$

where we normalize by *s* to get dimensionless values for *SK*. The skewness is used to investigate our data sets' *degree of symmetry* about the mean. A positive *SK* means we have a longer tail to the right of the mean than to the left, and vice versa for a negative *SK* (Figure 3.8).

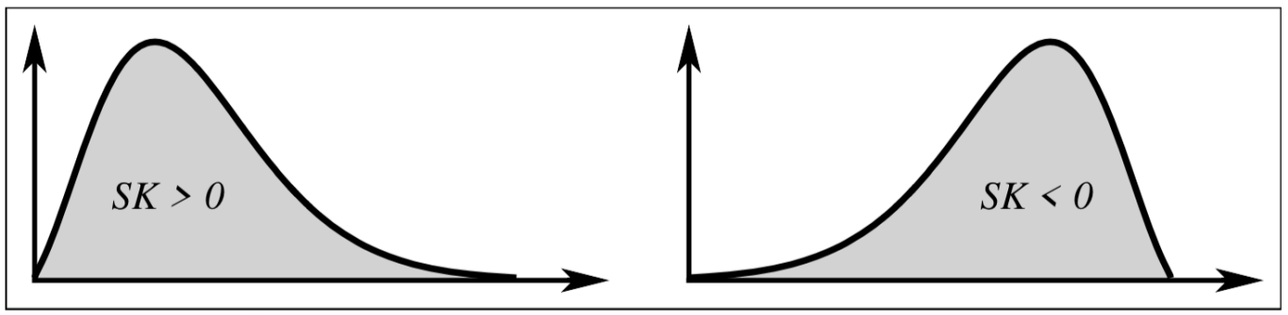

Figure 3.8: Examples of data distributions with positive and negative skewness. The sign of the skewness indicates which side of the distribution is long-tailed.

Unfortunately, if the data contain outliers then the *SK* will be very sensitive to these values and consequently be of little use to us. A more robust estimate of skewness is the *Pearson coefficient of skewness*,

$$SK_p = \frac{3(\bar{x} - \tilde{x})}{s}, \tag{3.63}$$

STATISTICAL CONCEPTS

The *Pearson coefficient of skewness* basically compares the mean and the median. An even higher-order central moment is the *kurtosis*,

$$K = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - 3 = \left\{ \frac{1}{n} \sum_{i=1}^{n} z_i^4 \right\} - 3. \qquad (3.64)$$

The correction term −3 makes $K = 0$ for a normal distribution, which we will discuss shortly. The kurtosis K attempts to quantify a data distribution's "sharpness" ($K > 0$) or "flatness" ($K < 0$; Figure 3.9). However, for most real data K can be almost infinite and should be used only with "well-behaved" data.

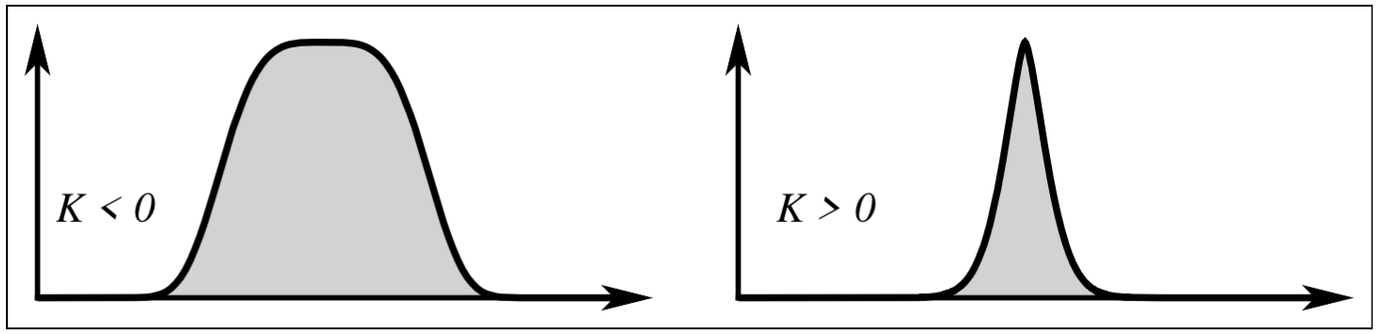

Figure 3.9: Examples of distributions with different kurtosis. Distributions with negative K are called *platykurtic*, while a positive K is called *leptokurtic*. You will of course be immensely pleased to learn that an intermediate case is called *mesokurtic*.

3.3 Discrete Probability Distributions

An important concept in statistics and probability is the notion of a *probability distribution*. It is a function $P(x)$, which indicates the probability that the event x will take place. $P(x)$ can be a discrete or continuous function. As an example of a discrete function, consider the function $P(x)$, $x = 1,2,..,6$, that gives the probability of throwing an x with a balanced die:

$$P(x) = 1/6,\ x = 1,2,...,6, \qquad (3.65)$$

or for flipping a coin:

$$P(x) = 1/2,\ x = \{H, T\} \qquad (3.66)$$

Staying with the throws of the die, we can relate $P(x)$ to the area under the curve in Figure 3.10.

STATISTICAL CONCEPTS

Two important properties shared by all discrete probability distributions are

$$0 \leq P(x_i) \leq 1, \text{ for all } x_i, \tag{3.67}$$

$$\sum_{i=1}^{n} P(x_i) = 1. \tag{3.68}$$

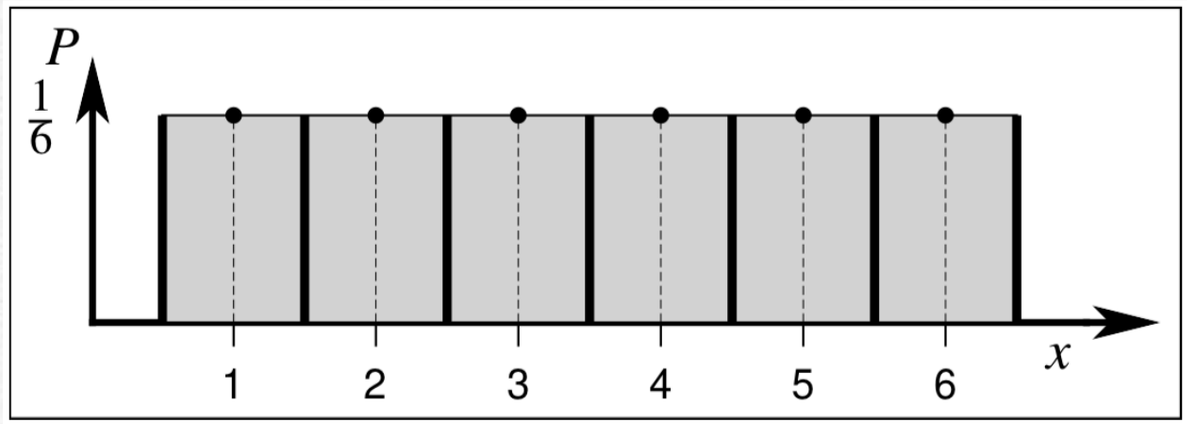

Figure 3.10: Probability of throwing any number on a die is a constant 1/6, unless the die is "loaded".

3.3.1 Binomial probability distribution

Often we are more interested in knowing the probability of a certain outcome after *n* repeated tries, such as "what is the probability of receiving junk mail three days in one week?" To derive such a function, we will assume that each event is independent and has the same probability, *p*. Then, the probability that an event *does not* occur is the complement, $q = 1 - p$. Consequently, the probability of getting *x* successes in *n* tries (and thus $n - x$ failures) is

$$P_1(x) = p^x q^{n-x}. \tag{3.69}$$

However, this probability applies to a *specific order* of all possible outcomes. Since we may not care about the order in which the successful *x* events occurred, we must scale $P_1(x)$ by the number of possible combinations of *x* successes in *n* tries. We already know this amount to be given by $\binom{n}{x}$, so our discrete probability function becomes

$$P_{n,p}(x) = \binom{n}{x} p^x q^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n. \tag{3.70}$$

STATISTICAL CONCEPTS

This expression is known as the binomial probability distribution or simply the *binomial distribution* (Figure 3.11) and it is used to predict the probability that *x* events out of *n* tries will be successful, given that each independent *x* has the probability *p* of success.

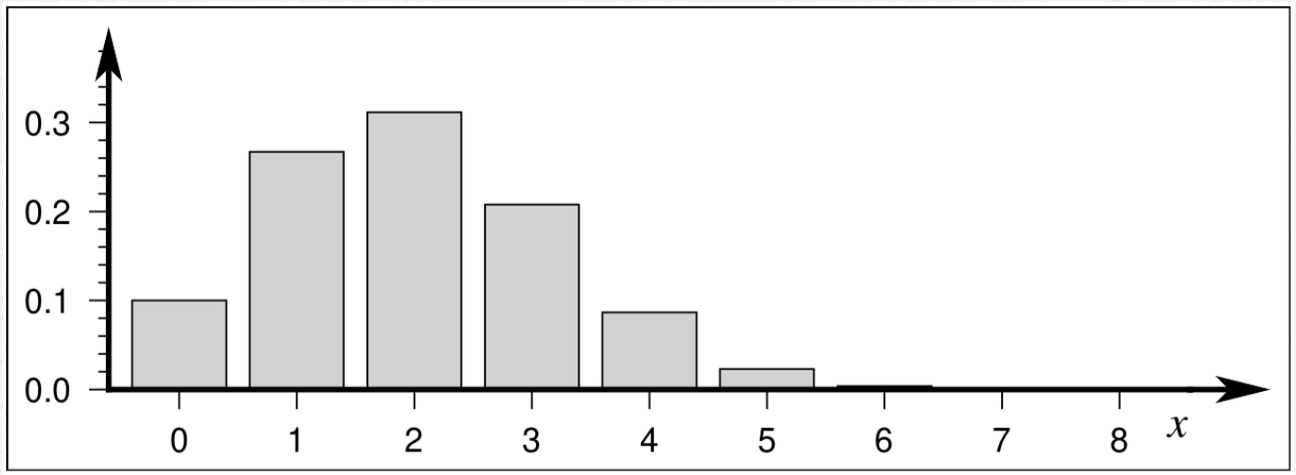

Figure 3.11: Binomial probability distribution $P_{n,p}(x)$, which shows the probability of having *x* successful outcomes out of a total of *n* tries, when each try has the probability *p* of success (and $q = 1 - p$ of failure). Here, $p = 0.25$ and $n=8$.

Example 3–8. What are the chances of drawing three red cards in six tries from a deck (assuming we place the card back into the deck after each try)? Here $p = 1/2$, so

$$P_{6,0.5}(3) = \frac{6!}{3!3!}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{6-3} = 0.31. \qquad (3.71)$$

One might have thought that getting half red and half black cards would have a higher probability, but remember that we require *exactly* 3 reds. If we compute the probability of getting 1, 2, or 3 reds separately and used the summation rule to compute the probability that we would draw 1, 2, or 3 red cards then *P* would be much higher.

The binomial probability distribution can also be used to assess the likelihood of more serious scenarios, such as the next example presents.

Example 3–9. A silver-tonged con artist approaches you on a street in New York City with a simple proposition: He has 10 beads — 9 black and one white. You get to pick one bead from his bag. You are given six opportunities to draw a bead (the bead is returned to the bag after each try), and if anytime during the six tries you pick the white bead then you have won and he will give you $20. However, if you have not picked the white bead after six tries then you owe him $20 instead.

Is this a good deal? Answer: Clearly, the probability of picking the white bead is fixed at $p = 0.1$. To lose the bet you will have to come up empty-handed six times in a row. For $n = 6$ and $r = 0$ the chances of that is simply

$$P_{6,0.1}(0) = \binom{6}{0} 0.1^0 (1-0.1)^6 = 0.53. \tag{3.72}$$

So while it is close to 50–50 the con-artist will most likely win, at least in the long run. (You probably should also be concerned that there might be something else going on as well, such as sleight-of-hand removal of the white bead before each try...)

3.3.2 The Poisson distribution

In some situations, the binomial distribution can be approximated by simpler expressions. One such case arises when the probability p for one event is very small and n is large. Such events are called *rare*, and the discrete distribution may then be approximated by

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, n \tag{3.73}$$

where $\lambda = np$ is the *rate of occurrence*. The Poisson distribution can be used to evaluate the probabilities for the occurrence of rare events such as large earthquakes, volcanic eruptions, and reversals of the geomagnetic field. For instance, the number of floods occurring in a 50-year period has been shown to follow a Poisson distribution with $\lambda = 2.2$. What is the probability that we will have at least one flood in the next 50 year period? Here, $P = 1 - P_0$, the probability of having no flood. Plugging in for $x = 0$ and $\lambda = 2.2$ we find $P_0 = 0.1108$, so $P = 0.8892$.

Example 3–10. A student is monitoring the radioactive decay of a certain sample that is expected to undergo three decays per minute. The student observes the number of decays over 100 individual one-minute periods and constructs the summary shown in Table 3.1. Does the data support the expected decay rate?

Decays	0	1	2	3	4	5	6	7	8	9
Observed	5	19	23	21	14	12	3	2	1	0

Table 3.1: Number of decays observed in one-minute interval.

STATISTICAL CONCEPTS

We make a histogram of the data by normalizing the observed frequencies by the total count and superimposing the Poisson distribution for the expected rate. The result (Figure 3.12) shows a very good fit.

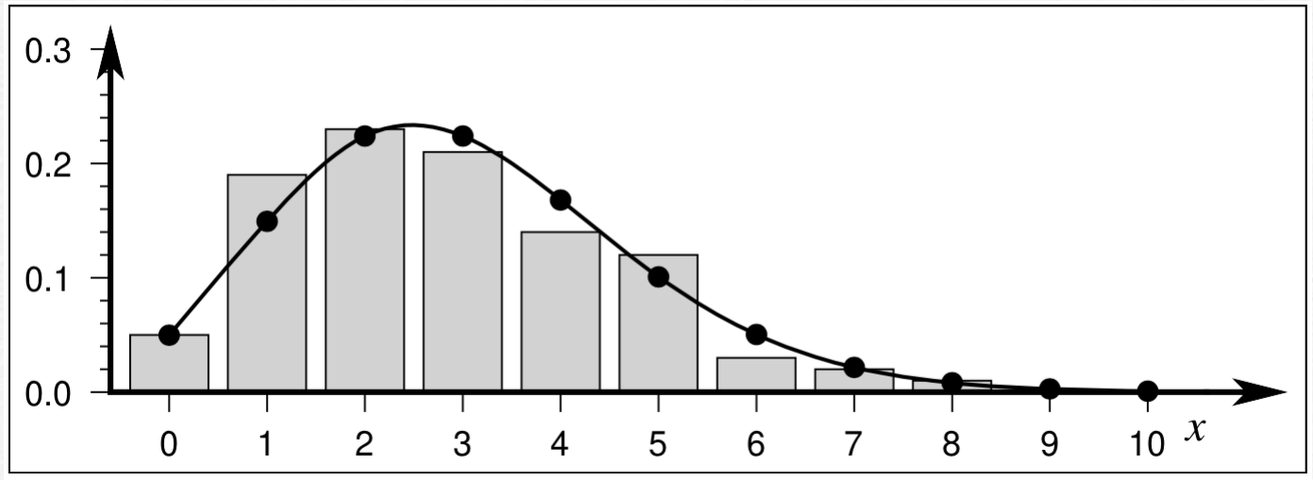

Figure 3.12: Histogram of observed decay rate frequencies (bars) and the theoretical Poisson distribution (circles) for the expected rate λ = 3.

3.4 Continuous Probability Distributions

While many populations are of a discrete nature (e.g., outcomes of coin tosses, numbers of microfossils in a core, etc.), we are very often dealing with observations of a phenomenon that can take on any of a continuous spectrum of values. We may sample the phenomenon at certain points in space-time and thus have discrete observations. Nevertheless, the underlying probability distribution is continuous (e.g., Figure 3.13).

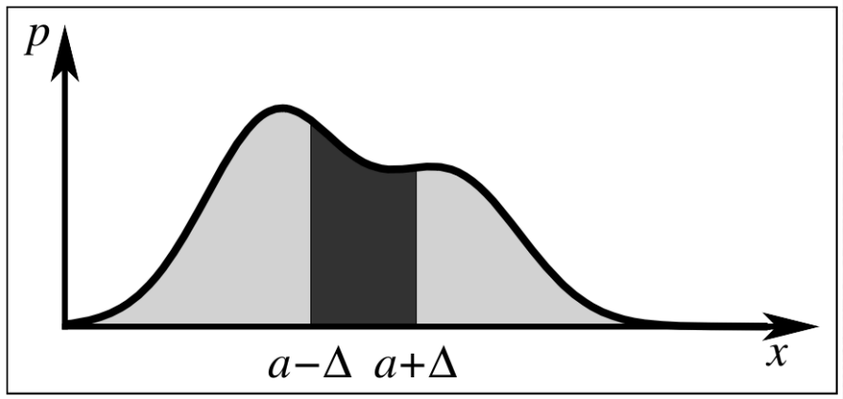

Figure 3.13: Example of a continuous probability density function (pdf). The area under any pdf must equal 1. The finite probability identified in (3.75) is indicated in dark gray.

Continuous distributions can be thought of as the limit for discrete distributions when the “spacing” between events shrinks to zero.

STATISTICAL CONCEPTS

Hence, we must replace the summation in (3.68) with the integral

$$\int_{-\infty}^{\infty} p(x)dx = 1. \tag{3.74}$$

Because of their continuous nature, functions such as $p(x)$ in (3.74) are called *probability density functions* (pdf). The probability of an event is still defined by the area under the curve, but now we must integrate to find the area and hence the probability. E.g., the probability that a random variable will take on a value between $a - \Delta$ and $a + \Delta$ is

$$P(a \pm \Delta) = \int_{a-\Delta}^{a+\Delta} p(x)dx. \tag{3.75}$$

As $\Delta \to 0$ we find that the probability goes to zero. Thus, the probability of getting exactly $x = a$ is nil.
The *cumulative distribution function* (cdf) gives the probability that an observation less than or equal to a will occur. We obtain the integral expression for this distribution by replacing the lower limit by $-\infty$ and the upper limit by a, finding

$$P_c(a) = \int_{-\infty}^{a} p(x)dx. \tag{3.76}$$

Obviously, as $a \to \infty$, $P_c(a) \to 1$. Given the cumulative distribution function we can revisit (3.75) and instead state

$$P(a \pm \Delta) = P_c(a + \Delta) - P_c(a - \Delta). \tag{3.77}$$

3.4.1 The normal distribution

So far the function $p(x)$ has been arbitrary. Any continuous function with unit area under the curve (i.e., 3.74) would qualify. We will now turn our attention to the best known and most frequently used pdf: the *normal distribution*. Its study dates back to 18th century investigations into the nature of experimental error. It was found that repeat measurements of the same quantity displayed a surprising degree of regularity.

STATISTICAL CONCEPTS

In particular, the German scientist K. F. Gauss played a major role in developing the theoretical foundations for the normal distribution, hence its other name: the *Gaussian* distribution. It is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{3.78}$$

where μ and σ have been defined previously. The constant term before the exponential normalizes the area under the curve to unity (Figure 3.14). As discussed in Section 3.2.4, it is often convenient to transform your data into so-called *standard scores:*

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{3.79}$$

in which case (3.78) reduces to

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \tag{3.80}$$

which has zero mean and unit standard deviation.

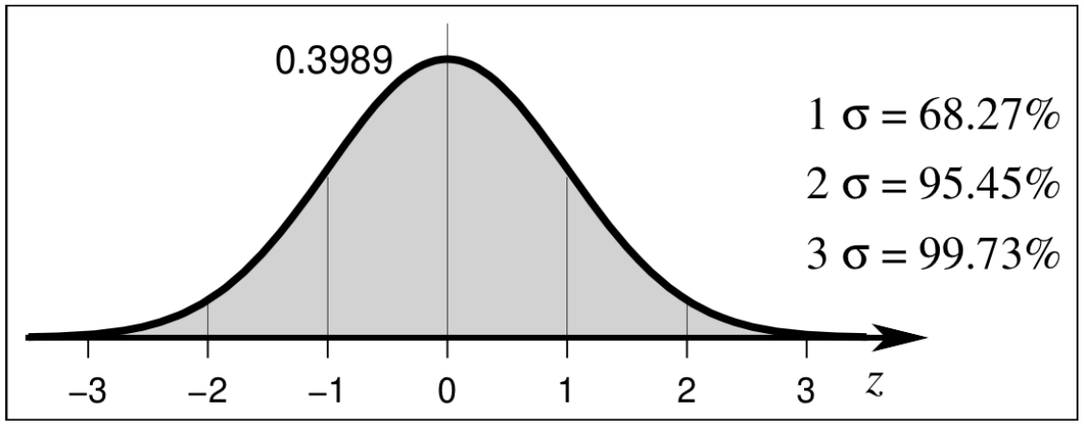

Figure 3.14: A normally distributed data set will have almost all of its values within ±3σ of the mean (this corresponds to 99.73% of the data; see legend for percentages corresponding to other multiples of ±σ).

Given the functional form of $p(z)$ we can evaluate the probability that an observation z will be ≤ a:

$$P_c(a) = \int_{-\infty}^{a} p(z) = \int_{-\infty}^{0} p(z) + \int_{0}^{a} p(z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{a} e^{-\frac{z^2}{2}} dz. \tag{3.81}$$

STATISTICAL CONCEPTS

Let

$$u^2 = \frac{z^2}{2}, \text{ hence } dz = \sqrt{2}du, \tag{3.82}$$

then

$$P_c(a) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{\frac{a}{\sqrt{2}}} e^{-u^2} du = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \frac{\sqrt{\pi}}{2} \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right)\right]. \tag{3.83}$$

It follows that, for any value *z*, the cumulative distribution function is

$$P_c(z) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right]. \tag{3.84}$$

Here, erf represents the *error function* and it is defined by the definite integral in (3.83) and tabulated in Table A.1. Furthermore, the probability that *z* falls between *a* and *b* must necessarily be

$$P_c(a \le z \le b) = P_c(b) - P_c(a) = \frac{1}{2}\left[\operatorname{erf}\left(\frac{b}{\sqrt{2}}\right) - \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right)\right]. \tag{3.85}$$

Example 3–11. Investigations into the strength of olivine have provided estimates of Young's modulus (*E*) that follow a normal distribution given by $\mu = 1.0 \cdot 10^{11}$ Pa and $\sigma = 1.0 \cdot 10^{10}$ Pa. What is the probability that a single estimate *E* will lie in the interval $9.8 \cdot 10^{10}$ Pa < E < $1.1 \cdot 10^{11}$ Pa? We convert the limits to normal scores and find they correspond to the interval $-0.2 \le z \le 1.0$. Using these values for *a* and *b* in (3.85) (or using Table A.1) we find the probability to be 0.4206.

Approximate binomial distribution

Like the Poisson distribution, the normal distribution may also serve as an approximation to the binomial distribution when *n* is large. More specifically, this approximation holds when both *np* and (1 − *p*)*n* exceed 5. Under those circumstances, the mean and standard deviation of the approximate normal distribution become

$$\mu = np, \quad \sigma = \sqrt{np(1-p)}, \tag{3.86}$$

leading to the simplified distribution

$$P_b(x) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left\{\frac{-(x-np)^2}{2np(1-p)}\right\}. \tag{3.87}$$

Example 3–12. What is the probability that at least 70 of 100 sand grains will be larger than 0.5 mm if the prob- ability that any single grain is that large is $p = 0.75$? Using the approximation (3.86) we find $\mu = np = 75$ and $s = \sqrt{n\,p(1-p)} = 4.33$. Converting 69.5 (halfway between 69 and 70) to a z score gives −1.27, and we find via Table A.1 that the probability becomes 0.898 or about 90%.

3.4.2 The exponential distribution

Another important probability density distribution is the *exponential* distribution. It is given by

$$p_e(x) = \lambda e^{-\lambda x} \tag{3.88}$$

for some constant λ. However, most of the time we will see it used as a cumulative distribution function:

$$P_c(x) = 1 - e^{-\lambda x}. \tag{3.89}$$

Eq. (3.89) gives the probability that the observation a will be in the range $0 \le a \le x$.

Example 3–13. It has been reported that the heights (z) of Pacific seamounts follow an exponential distribution defined as

$$P_c(z \le h) = 1 - e^{-h/340}, \tag{3.90}$$

which gives the probability that a seamount is shorter than h meters. Equation (3.90) then predicts that we might expect that

$$P_c(1000) = 1 - e^{-1000/340} \approx 95\% \tag{3.91}$$

of them are less than one km tall.

STATISTICAL CONCEPTS

3.4.3 Log-normal distribution

Many data sets, such as grain-sizes of sediments and geochemical concentrations, have very skewed and long-tailed distributions (e.g., Figure 3.15). In general, such distributions arise when the observed quantities have errors that depend on *products* rather than *sums*. It therefore follows that the *logarithm* of the data may be normally distributed. Hence, taking the logarithm of your data may make the transformed distribution look normal. If this is the case, you can apply standard statistical techniques applicable to normal distributions to the logarithm of your data and convert the results (e.g., mean, standard deviation) back to get the proper units. The log-normal probability density distribution is therefore given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2}. \qquad (3.92)$$

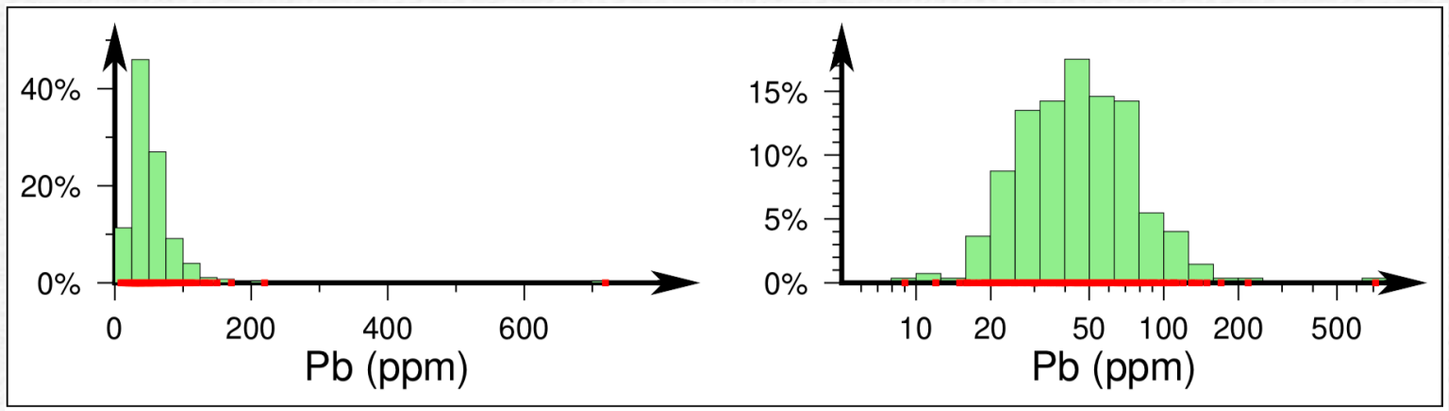

Figure 3.15: (left) The concentration of Pb in soil is very long-tailed and clearly not normally distributed. The red squares indicate individual sample values. (right) The same distribution after taking the logarithm of the data values. The resulting distribution is approximately normal, hence a log-normal distribution might be suitable to describe the data.

3.5 Inferences about Means

The central limits theorem states that the mean of a large sample taken from any distribution will be normally distributed even if the data themselves are not normally distributed, and furthermore it says that the sample mean is an unbiased estimator of the population mean. We can then use our knowledge of the normal distribution to quantify our faith in the precision of our sample mean. We already know that $s_{\bar{x}} = \sigma/\sqrt{n}$, so we can state with probability $1 - \alpha$ that $\bar{x}$ will differ from μ by at most E, which is given by

STATISTICAL CONCEPTS

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \tag{3.93}$$

where s is our estimate of σ. In other words, the chance that $\bar{x}$ exceeds the $\pm z_{\alpha/2}$ confidence interval is α. These error estimates apply to large samples ($n \geq 30$) and infinite populations. In those cases we can use our sample standard deviation s in place of σ, which we usually do not know. Here, (3.93) can be inverted to yield the sample size necessary to be confident that the error in our sample mean is no larger than E, and we find

$$n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2. \tag{3.94}$$

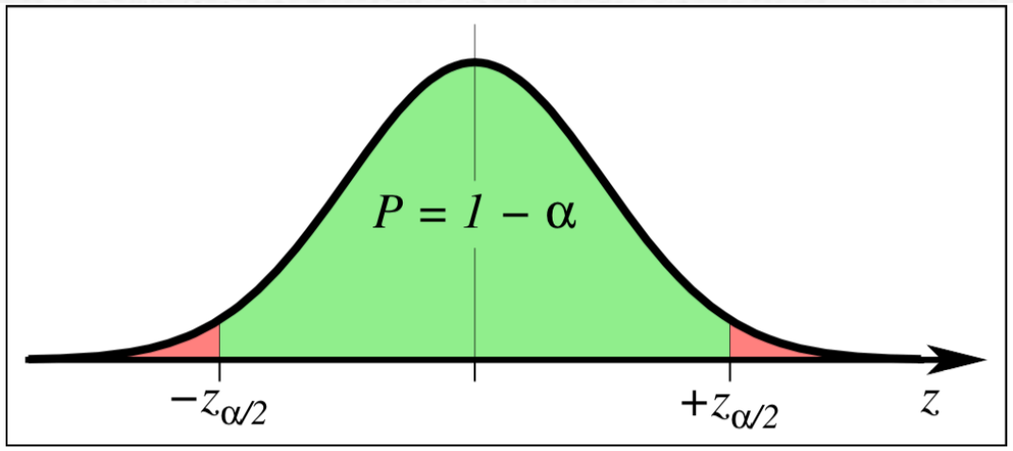

Figure 3.16: Probability is α that a value will fall in one of the two tails of the normal distribution, and α/2 that it will fall in a specific tail.

The normal score for our sample mean is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}. \tag{3.95}$$

Since this statistic is normally distributed we know that the probability is 1 − α that z will take on a value in the interval $-z_{\alpha/2} < z < +z_{\alpha/2}$. Plugging in for the limits on z,

$$-z_{\alpha/2} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < +z_{\alpha/2} \tag{3.96}$$

or

$$\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}. \tag{3.97}$$

Rearranging, we find

$$\mu = \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}. \tag{3.98}$$

Eq. (3.98) shows the *confidence interval* on μ at the $1 - \alpha$ confidence level. Very often, our confidence levels will be 95% (~2σ) or 99% (~3σ).

3.5.1 Small samples

The previous section dealt with large ($n \geq 30$) samples, where we could assume that $\bar{x}$ would be normally distributed as dictated by the central limits theorem. For smaller samples we must assume instead that the *population we are sampling* is normally distributed. We can then base our inferences on the statistic

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \tag{3.99}$$

whose distribution is called the *Student's t*-distribution (Figure 3.17). It is similar to the normal distribution but its shape depends on the degrees of freedom, $\nu = n - 1$. For large n (and hence ν) the t statistics approach the z statistics. As for z statistics, one can find tables with t values for various combinations of confidence levels and degrees of freedom (see Table A.2). For insight into what the t-distribution and others really are, get *Numerical Recipes* by Press et al. This excellent book gives both theory and computer code (in C++, C, FORTRAN, Java and a host of legacy languages).

Example 3–14. Given our sandstone density estimates from earlier, i.e., {2.2, 2.25, 2.25, 2.3, 2.3, 2.3, 2.35}, what is the 95% confidence interval on the population mean?

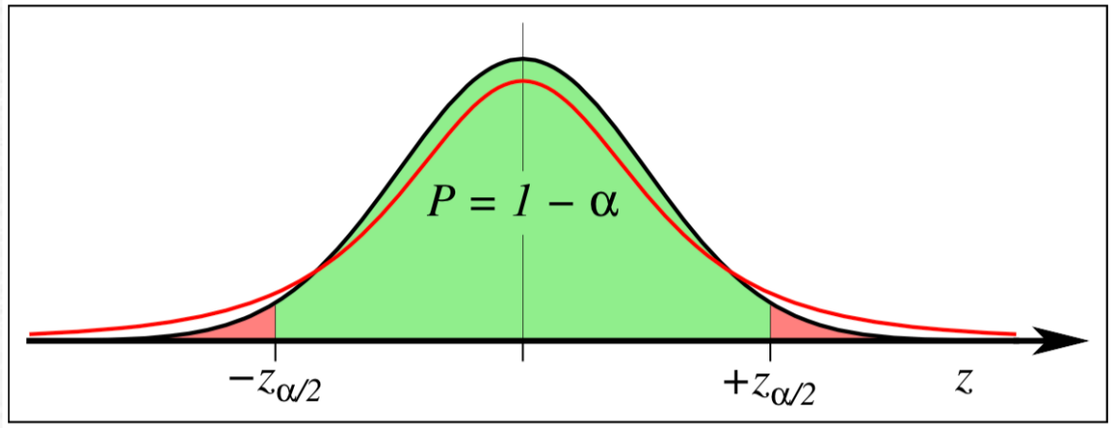

Figure 3.17: The same normal distribution and critical tails as in Figure 3.16, overlain by the Student's *t*-distribution for ν = 3 degrees of freedom (red line). For small samples the probability distribution becomes wider.

Answer: We have $\bar{x}$ = 2.28 with s = 0.05, and α = 1−95% = 0.05. The degrees of freedom ν = n−1 = 6. Table A.2 gives $t_{\alpha/2,\nu} = t_{0.025,6}$ = 2.447. Using (3.97), we find our sample mean brackets the population mean, thus (with $t_{\alpha/2}$ instead of $z_{\alpha/2}$ and s instead of σ)

$$2.28 - 2.447 \cdot \frac{0.05}{\sqrt{7}} < \mu < 2.28 + 2.447 \cdot \frac{0.05}{\sqrt{7}} \qquad (3.100)$$

or (since the bounds are symmetrical)

$$2.234 < \mu < 2.326 \qquad (3.101)$$

or

$$\mu = 2.280 \pm 0.046. \qquad (3.102)$$

A.1 Cumulative Probabilities for the Normal Distribution

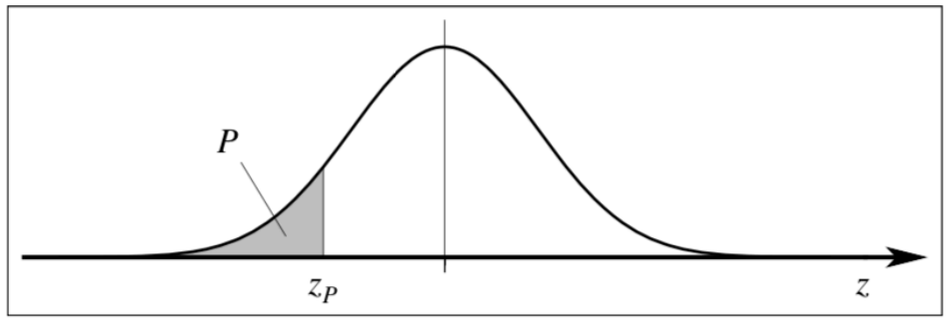

Figure A.1: Given a chosen z_P-value, the probability P (gray area under the curve from $-\infty$ to z_P) can be read from this table. Here, z_P is given in the format *-a.bc*, where *-a.b* and *0.0c* correspond to a unique row and column combination.

z_P	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-3.4	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005
-3.2	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007
-3.1	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010
-3.0	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013
-2.9	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019
-2.8	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026
-2.7	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035
-2.6	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047
-2.5	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062
-2.4	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082
-2.3	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107
-2.2	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139
-2.1	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179
-2.0	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228
-1.9	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287
-1.8	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359
-1.7	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446
-1.6	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548
-1.5	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668
-1.4	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808
-1.3	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968
-1.2	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151
-1.1	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357
-1.0	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587
-0.9	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841
-0.8	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119
-0.7	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420
-0.6	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743
-0.5	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085
-0.4	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446
-0.3	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821
-0.2	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207
-0.1	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602
-0.0	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000

Table A.1: Normal cumulative distribution function. For $z > 0$ use $P(z) = 1 - P(-z)$.