Chapter 4

Modeling Data as Random Variables and Populations as Probability Distributions

Chapter Objectives

- Distinguish between discrete and continuous probability distributions.
- Interpret the mean and standard deviation of a probability distribution.
- Recognize binomial and Poisson random variables.
- Obtain probabilities from binomial and Poisson distributions.
- Recognize normal and lognormal random variables.
- Obtain probabilities from normal and lognormal distributions.
- Find and interpret percentiles of normal and lognormal distributions.
- Find and interpret the limit of detection for a given measurement error standard deviation.

Key Takeaways

- Random variables are values determined by chance. They can be discrete or continuous.
- The probability distribution of a random variable indicates the probabilities with which the variable takes different values.
- Discrete probability distributions are represented by probability functions. Continuous ones are represented by probability density functions.
- Binomial and Poisson distributions are used to model discrete counts.
- Normal distributions are bell-shaped, continuous distributions.
- Lognormal distributions are right-skewed, continuous distributions.
- Measurement error often follows a normal distribution with mean zero.
- The limit of detection is a value that separates nondetect measurements from detects.

4.1 Introduction and Notation

The *E. coli* level in a lake can vary from one surface location to the next, and for a *randomly selected* location, the resulting *E. coli* value will be determined by chance. Any numerical variable whose value is determined by chance is called a ***random variable***.

If we knew the *E. coli* level at *every* location on the lake, we'd know which values we might end up with and we could determine the probabilities of those values. The set of values a random variable might take and the probabilities associated with those values together form what's called the ***probability distribution*** of the variable. Probability distributions are used in statistics to represent *populations*. Random variables are used to represent *sample values*.

Throughout this chapter, we'll use the following notational conventions:

- Upper case letters such as $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ will be used to denote *random* variables whose values *have yet to be determined.* For example X might represent the *E. coli* level in a water specimen that hasn't been drawn yet.
- Probabilities involving a random variable X will be denoted $\boldsymbol{P(X = 3)}$, $\boldsymbol{P(X \leq 6.5)}$, and so on.
- Occasionally, a lower case letter such as $\boldsymbol{x}$ will be used in place of a number like 3 or 6.5 to denote a generic probability, for example $P(X = x)$ or $P(X \leq x)$.

A random variable is ***discrete*** or ***continuous*** depending on whether its set of possible values consists of isolated numbers (e.g. integers) or a continuum. We'll look at discrete random variables in Section 4.2, followed by some special discrete variables that are *counts* in Section 4.3. Then in in Section 4.4, we'll turn to continuous random variables, with some special ones covered in Section 4.5.

4.2 Discrete Probability Distributions

4.2.1 Introduction

When working with a discrete random variable X, it's convenient to use the shorthand notation $\boldsymbol{p(x)}$ for $P(X = x)$, and then think of $p(x)$ as a function of x. The function $p(x)$ is called the ***probability function*** of the random variable X, and it characterizes the probability distribution of X. A graph of the probability function, as described in the next example, is called the ***probability histogram*** of the distribution.

Example 4.1: Discrete Probability Distribution

The table below shows the vehicle occupancy rates on urban arterials and freeways in Miami-Dade County, Florida (based on reports for 154,152 vehicles involved in accidents) [8].

Number of Occupants	Percentage of Vehicles
1	82 %
2	12 %
3	4 %
4	2 %

The percentage of vehicles with five or more occupants was negligibly small, so it's left out of the table above.

We can interpret the percentages as the probabilities that a randomly selected vehicle will have 1, 2, 3, and 4 occupants, respectively.

Letting X denote the number of occupants in the randomly chosen vehicle, X is a discrete random variable whose possible values are 1, 2, 3, and 4. The probability distribution of X is shown in the table below.

x	1	2	3	4
$p(x)$	0.82	0.12	0.04	0.02

The table says, for example, that the probability of a randomly selected vehicle having only one occupant is $p(1) = P(X = 1) = 0.82$.

The probability distribution is shown below as a *probability histogram*, with the possible values of X on the horizontal axis and bars whose heights are the probabilities of those values. This probability distribution represents the *population* of vehicles in Miami-Dade County.

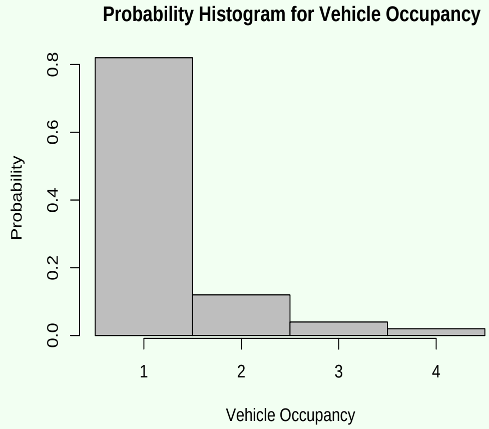


Figure 4.1: Probability histogram for vehicle occupancy.

The last example illustrates the use of a probability distribution to represent a *population* from which an elementary unit is randomly selected and a variable measured upon. If the distribution accurately represents the population, and a *random sample* is drawn from the population, a histogram of the data would approximately resemble the probability histogram.

Properties of Probability Functions: Every discrete probability function $p(x)$ has the following properties:

1. $0 \leq p(x) \leq 1$ for all values of x.
2. $\sum_{\text{all } x} p(x) = 1$, where the summation is over all possible values x of the random variable X.

4.2.2 The Mean and Standard Deviation of a Discrete Probability Distribution

We measure the *center* of a probability distribution by its ***mean***, denoted $\boldsymbol{\mu}$. If the bars in a probability histogram were weights, μ is the point along the horizontal axis at which they'd balance. We can compute the mean using the following.

Mean of a Discrete Probability Distribution: For a discrete random variable X whose probability function is $p(x)$, the mean of the distribution is

$$\mu \;=\; \sum_{\text{all } x} x\, p(x), \tag{4.1}$$

where the summation is over all possible values x of X.

Thus μ is computed by multiplying each of the possible values of the random variable by its probability and then summing the results.

The mean μ has two interpretations:

1. It's the *mean of the population* when the probability distribution represents a population.
2. It's a *typical value* of the variable.

Example 4.2: Mean of a Discrete Distribution

The mean of the probability distribution given in Example 4.1 is

$$\begin{aligned}\mu &= \sum_{\text{all } x} x\, p(x) \\ &= 1(0.82) + 2(0.12) + 3(0.04) + 4(0.02) \\ &= 1.26.\end{aligned}$$

Thus the population mean number of occupants is 1.26, which is the balancing point of the probability histogram in Fig. 4.1. On average, a randomly selected vehicle will have 1.26 occupants.

We measure the *spread* in a probability distribution (that is, variation in the values the random variable takes) by its ***standard deviation***, denoted $\boldsymbol{\sigma}$. A larger value of σ corresponds to a more spread-out distribution. To compute the standard deviation, we use the following.

Standard Deviation of a Discrete Probability Distribution: For a discrete random variable X whose probability function is $p(x)$, the standard deviation is

$$\sigma = \sqrt{\sum_{\text{all } x} (x-\mu)^2\, p(x)}, \tag{4.2}$$

where the summation is again over all possible values x of X.

Thus σ is computed by multiplying each possible squared deviation of the random variable away from the mean by its probability, summing the results, and then taking the square root. The square of the standard deviation, $\boldsymbol{\sigma^2}$, is called the ***variance*** of the distribution.

The standard deviation σ measures variation in the values that the random variable takes. It has two interpretations:

1. It's the *standard deviation of the population* when the probability distribution represents a population.
2. It's the size of a *typical deviation* of the variable away from the mean μ.

Example 4.3: Standard Deviation of a Discrete Distribution

Continuing from the previous example, the standard deviation of the probability distribution is

$$\begin{aligned}\sigma &= \sqrt{\sum_{\text{all } x} (x-\mu)^2\, p(x)} \\ &= \sqrt{(1-1.26)^2(0.82) + (2-1.26)^2(0.12) + (3-1.26)^2(0.04) + (4-1.26)^2(0.02)} \\ &= 0.63.\end{aligned}$$

Together, the mean from Example 4.1 and standard deviation from above tell us that on average in the population, a vehicle will have 1.26 occupants, plus or minus about 0.63 occupants.

4.3 Probability Distributions for Counts

In the vehicle occupancy example (Example 4.1) the probability distribution was based on accurate information about the population of vehicles. In the absence of such accurate information, we have to choose from a set of stock *theoretical distributions* the one that we *think* describes the population. Then probabilities involving randomly selected elementary units can be obtained from the theoretical distribution.

The first step is to identify whether the random variable is *discrete* or *continuous.* Discrete variables are most often *counts.* Two commonly used theoretical distributions for counts are:

1. The binomial distribution
2. The Poisson distribution

The first is used for counts of the number of times a particular outcome occurs among several trials or occasions. The second is used for counts of the number of times an event occurs over a given time period or over a given spatial region.

4.3.1 The Binomial Distribution

The following are examples of *binomial random variables*:

- The number of animal traps, among the 15 traps set, that catch animals.
- The number of fish, among 10 tested, that test positive for a certain disease.
- The number of water specimens, among the seven tested, that test positive for a certain trace chemical.

In general, a ***binomial*** random variable is one for which the following conditions are met:

Conditions for a Binomial Random Variable:

1. There are a certain number of *trials*, $\boldsymbol{n}$.
2. Each trial has *two possible outcomes*, *success* and *failure*, say.
3. The trials all have the same probability $\boldsymbol{p}$ of resulting in a success. Thus the probability of a failure is $\mathbf{1-\boldsymbol{p}}$.
4. The trials are *independent*, meaning their outcomes don't influence each other.
5. The random variable X is the count of the number of successes among the n trials.

Here are some more examples.

Example 4.4: Binomial Distribution

Human consumption of mercury (Hg) can impair brain and nervous system development in fetuses, infants, and children. Mercury is known to accumulate in fish, and the World Health Organization (WHO) suggests that fish with Hg concentrations greater than 0.5 mg/kg are unsafe for human consumption.

In the U.S., much of the fish consumed comes in the form of canned tuna, which is sometimes sold in packages of four cans. In a given package of four, the number of cans that have an unsafe Hg level could be modeled as a binomial random variable.

Example 4.5: Binomial Distribution

Studies of animal populations sometimes involve visiting several sites, each one on multiple occasions, and recording whether the animal was seen or not during each visit.

Repeated visits to a given site yield a ”detection history” for that site, for example (0, 1, 0, 1, 0, 0) if the site was visited six times and the animal seen only on the second and fourth visits. The variable of interest is how many of the visits led to a sighting (two in the example just given). This *count* could be modeled as a *binomial* random variable.

Example 4.6: Binomial Distribution

A *biological assay*, or *bioassay* for short, is an experiment to determine the toxicity of a substance.

In a bioassay of wastewater treatment plant effluent, 10 aquatic organisms are placed in an effluent-filled aquarium. After a set period of time, each organism is either dead or still alive. The variable of interest is the number of deceased organisms, and it could be modeled as a *binomial* random variable.

The Binomial Probability Function

To determine, in Example 4.4, how likely it would be to end up with, say, two unsafe tuna cans out of a package of four, we can use the ***binomial probability function***:

Binomial Probability Function: If X is a binomial random variable with n trials, each of which results in a *successes* with probability p and *failure* with probability $1-p$, then

$$p(x) \;=\; \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad \text{for} \quad x = 0, 1, 2, \ldots, n, \tag{4.3}$$

where the notation $\boldsymbol{n!}$ (read ”***n factorial***”) is defined (for each positive integer n) as

$$n! \;=\; n(n-1)(n-2)\cdots(1)$$

and, by definition, $0! = 1$.

This probability function characterizes the ***binomial distribution***, and we refer to n and p as its ***parameters***. Later we’ll see how the probability function was derived, and we’ll see that its parameters control its shape, center, and spread.

We’ll use the notation

$$\boldsymbol{X \sim \textbf{Binomial}(n,\, p)}$$

to mean that the random variable X follows a binomial distribution with parameters n and p.

Example 4.7: Binomial Distribution

Continuing with the canned tuna example (Example 4.4), we have $n = 4$ cans (trials), and each is either unsafe (*success*) or safe (*failure*). According to one study, 10% of tuna cans sold are unsafe, so the probability that a given can will be unsafe is $p = 0.1$ [10].

If the safety statuses of the four cans are independent of each other, the number of unsafe cans X in a given package is a binomial variable with parameters $n = 4$ and $p = 0.1$, which we write as

$$X \sim \text{Binomial}(4,\ 0.1).$$

Using (4.3), the chance that exactly two of the four cans will be unsafe (and the other two safe)

$$\begin{aligned} p(2) &= \frac{4!}{2!(4-2)!} 0.1^2 (1-0.1)^{4-2} \\ &= 0.049. \end{aligned}$$

The probabilities for each of the values $x = 0, 1, \ldots, 4$ are computed in a similar manner, and are shown below.

x	0	1	2	3	4
$p(x)$	0.656	0.292	0.049	0.004	0.000

A probability histogram of this binomial distribution is below.

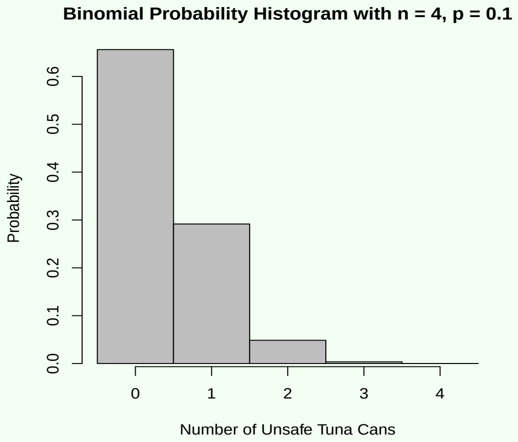


Figure 4.2: Probability histogram for the number of unsafe tuna cans in a package of four.

The next example shows that the value of p affects the shape, center, and spread of the binomial distribution.

Example 4.8: Binomial Distribution

Suppose in the bioassay study of Example 4.6 that 10 organisms are placed in each of two aquariums, one containing treatment plant effluent and the other fresh water. Suppose also that a given organism will die with probability $p = 0.7$ in the effluent aquarium and $p = 0.15$ in the fresh water one.

If their death occur *independently* of each other, the numbers of deaths in the two aquariums at the end of the study follow Binomial(10, 0.7) and Binomial(10, 0.15) distributions, respectively. These distributions are shown below.

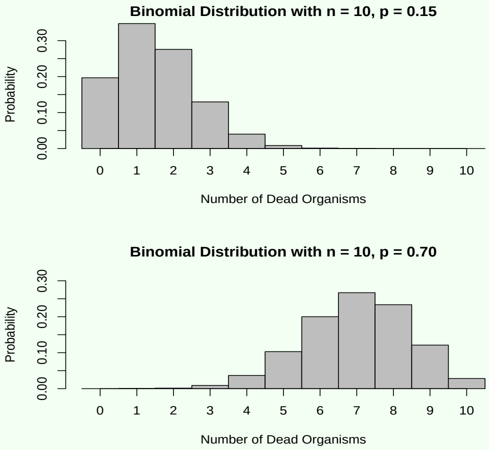


Figure 4.3: Probability histograms for the number of organisms that die in the control aquarium (top) and the treatment plant effluent aquarium (bottom).

The death-count distribution for the effluent aquarium ($p = 0.7$) lies farther to the right than the one for the fresh-water aquarium ($p = 0.15$). Also, the two distributions have different shapes and different amounts of spread.

Mean and Standard Deviation of a Binomial Distribution

The mean and standard deviation of a binomial distribution *could* be computed by constructing a table of probabilities like the one in Example 4.7, and then using formulas (4.1) and (4.2). But it turns out there's an easier way. It's clear from the last example that the mean (center) and standard deviation (spread) of the distribution depend on the value of the parameter p. In fact, they also depend on the value of the parameter n.

Here are shortcut formulas for computing the mean and standard deviation of a binomial distribution.

Mean and Standard Deviation of the Binomial Distribution: The mean $\boldsymbol{\mu_{\text{bin}}}$ of a Binomial(n, p) distribution is

$$\mu_{\text{bin}} \;=\; np \tag{4.4}$$

and the standard deviation $\boldsymbol{\sigma_{\text{bin}}}$ is

$$\sigma_{\text{bin}} \;=\; \sqrt{np(1-p)}. \tag{4.5}$$

In the next example, we'll see that the shortcut formulas give the same results as (4.1) and (4.2).

Example 4.9: Binomial Distribution

Consider again the safety statuses of tuna cans (Examples 4.4 and 4.7). Using the shortcut formula (4.4), the mean of the binomial distribution with with $n = 4$ and $p = 0.1$ is

$$\begin{aligned} \mu_{\text{bin}} &= np \\ &= 4\,(0.1) \\ &= 0.4. \end{aligned}$$

This says that the average number of unsafe cans in a package of four is 0.4.

Now using the more general formula (4.1) with the table of probabilities in Example 4.7, we get

$$\begin{aligned}\mu &= \sum_{\text{all } x} xp(x) \\ &= 0(0.656) + 1(0.292) + 2(0.049) + 3(0.004) + 4(0.000) \\ &= 0.4.\end{aligned}$$

Thus the two results are the same.

The standard deviation of the binomial distribution, from the shortcut formula (4.5), is

$$\begin{aligned}\sigma_{\text{bin}} &= \sqrt{np(1-p)} \\ &= \sqrt{(4)(0.1)(1-0.1)} \\ &= 0.6.\end{aligned}$$

This says that that actual number of unsafe cans in a package will typically differ from the mean number, 0.4, by about 0.6 of a can. One standard deviation below the mean extends into negative territory because the distribution is right skewed, as seen in Fig. 4.2.

It can be shown that we'd get the same value for σ_{bin} using the more general formula (4.2) with the table of probabilities in Example 4.7.

The Binomial Distribution When n is Large

We've seen that the shape of the binomial distribution depends on the value of p (Example 4.8). What's effect of n on the shape? Fig. 4.4 shows binomial distributions with three different values of n. As n increases, the distribution becomes more and more bell-shaped. It also shifts to the right and becomes more spread out – the mean (4.4) and standard deviation (4.5) both increase as n does. If n is large enough, the binomial distribution becomes indistinguishable from a perfect bell-shaped *normal distribution*, described later in this chapter.

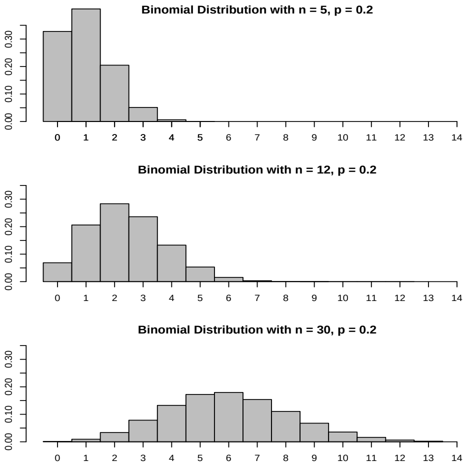


Figure 4.4: Binomial distributions having different values of n: $n = 5$ (top), $n = 12$ (middle), and $n = 30$ (bottom). In each case, $p = 0.2$.

Derivation of the Binomial Probability Function Formula

To see how the binomial probability function (4.3) was derived, we'll use the safety statuses of four cans of tuna as an example. We saw in Example 4.7 that the probability that exactly two of the four cans will be unsafe is

$$p(2) \quad = \quad \underbrace{\frac{4!}{2!(4-2)!}}_{\substack{\text{Number of ways}\\\text{two of the four}\\\text{cans can be unsafe}}} \quad \underbrace{0.1^2(1-0.1)^{4-2}}_{\substack{\text{Probability of each}\\\text{of those ways}}} \tag{4.6}$$

The first part, $4!/(2!(4-2)!)$, can be shown to be the number of test-outcome sequences for which two cans out of four are unsafe. For example, one sequence is

$$(0,\ 1,\ 0,\ 1),$$

meaning that the second and fourth cans are unsafe. Another is

$$(1,\ 0,\ 0,\ 1)$$

meaning that the first and fourth are unsafe. All $4!/(2!(4-2)!) = 6$ sequences are shown in Fig. 4.5.

The second part of the right side of (4.6), $0.1^2(1-0.1)^{4-2}$, is the probability of each of the six sequences of Fig. 4.5. For example, the sequence

$$(0,\ 1,\ 0,\ 1)$$

has probability

$$(0.9)(0.1)(0.9)(0.1) \ = \ 0.1^2(1-0.1)^{4-2},$$

and the sequence

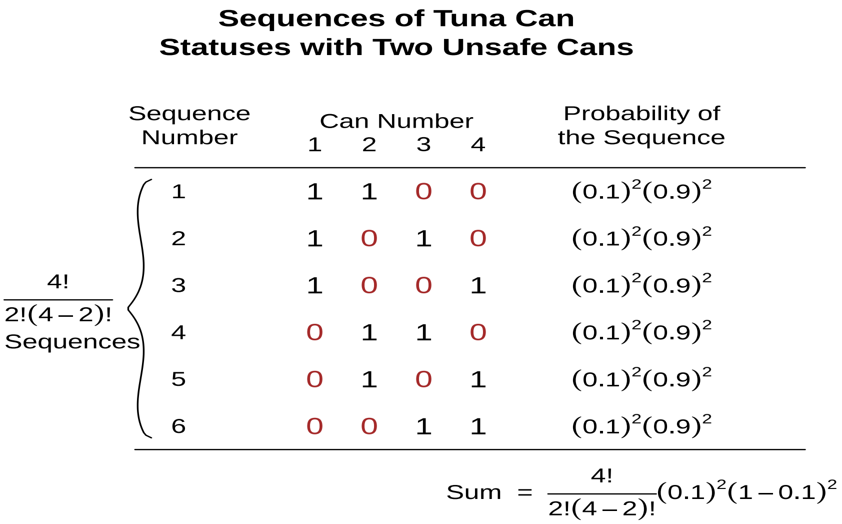
$$(1,\ 0,\ 0,\ 1)$$

has probability

$$(0.1)(0.9)(0.9)(0.1) \ = \ 0.1^2(1-0.1)^{4-2}$$

too. Similarly, each of the other four sequences has this same probability.

The probability that exactly two of the four cans will be unsafe is the probability that one or another of the six sequences will result, and is obtained by summing the six sequences' probabilities, as shown on the bottom right of Fig. 4.5 and used in (4.6).



Sequences of Tuna Can Statuses with Two Unsafe Cans

Sequence Number	Can Number 1	2	3	4	Probability of the Sequence
1	1	1	0	0	$(0.1)^2(0.9)^2$
2	1	0	1	0	$(0.1)^2(0.9)^2$
3	1	0	0	1	$(0.1)^2(0.9)^2$
4	0	1	1	0	$(0.1)^2(0.9)^2$
5	0	1	0	1	$(0.1)^2(0.9)^2$
6	0	0	1	1	$(0.1)^2(0.9)^2$

$\frac{4!}{2!(4-2)!}$ Sequences

$$\text{Sum} = \frac{4!}{2!(4-2)!}(0.1)^2(1-0.1)^2$$

Figure 4.5: All six ways in which two tuna cans out of four can be unsafe (0 = safe, 1 = unsafe) and their probabilities.

In general, the number of sequences of n outcomes for which x are *successes* and the other $n-x$ *failures* is called the number of ***combinations*** of n outcomes taken x at time, and is given by the following.

Combinations: The number of combinations of n outcomes taken x at a time is

$$\text{Number of Combinations} = \frac{n!}{x!(n-x)!}.$$

4.3.2 The Poisson Distribution

The following are examples of *Poisson random variables*:

- The number of flash floods during a 100-year period.
- The number of shooting stars in the night sky during a one hour period.
- The number of patients admitted for respiratory problems to a hospital in a given month.
- The number of trees of a certain species on a 100 m^2 plot of land.
- The number of beetles in a 1 m^2 quadrat.

In general, ***Poisson*** random variables are counts of events occurring over a given time period or spatial area and for which the following conditions are met:

Conditions for a Poisson Random Variable:

1. Events occur at random time points or at random spatial points. The (temporal) rate or (spatial) density of their occurrence is approximately constant (doesn't change over time as they're being counted or across locations inside the study region).

2. The events occur independently of each other in time or space, i.e. the likelihood of an event occurring at a given time or location doesn't depend on whether others have occurred in that (temporal or spatial) vicinity.
3. The random variable X is the count of the number of occurrences of the event in a specified time period or spatial area.

According to the second condition, the Poisson distribution would be appropriate for modeling spatial counts of plants or animals that disperse independently of each other, but not plants that grow in clumps or animals that congregate in groups, both common in nature.

Here's an example in which the Poisson distribution is used to model counts of events in time.

Example 4.10: Poisson Distribution

On average, 1.68 hurricanes make landfall on the continental U.S. per year according to records dating back to 1850 [7]. In any given year, though, the number of hurricanes is a random variable. Some researchers use the Poisson distribution, with mean 1.68, to compute probabilities of various numbers of hurricanes. This distribution is depicted below.

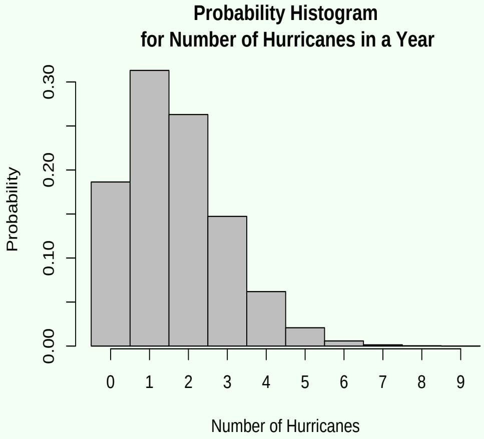


Figure 4.6: Probability histogram for number of hurricanes making landfall on the continental U.S. in a year.

Notice that this *theoretical* distribution is centered on 1.68, the long-run average number of hurricanes. According to the distribution, it would be highly unlikely for the U.S. to experience more than four hurricanes in a given year.

Here's an example in which the Poisson distribution is used to model counts of events in space.

Example 4.11: Poisson Distribution

In a study of the centipede *Lithobius muticus*, the number of centipedes in a randomly selected 1 m^2 quadrat was modeled by a Poisson distribution with mean 10.5 centipedes [12]. This distribution is shown below.

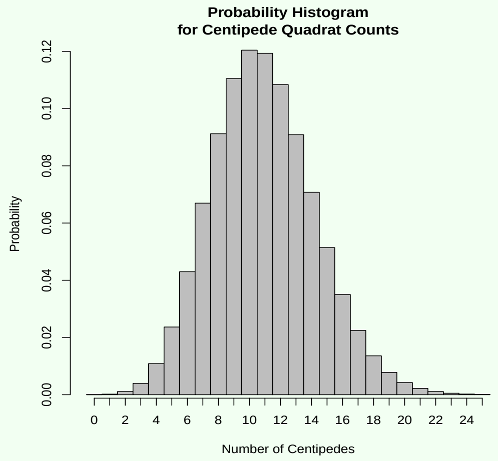


Figure 4.7: Probability histogram for the number of centipedes in a randomly selected quadrat.

Note that this *theoretical* distribution is centered on 10.5, its mean, and that most of the probability lies between about 6 and 16.

The Poisson Probability Function

To compute how likely it would be to end up with, say, three hurricanes in a given year (in Example 4.10), we use the ***Poisson probability function***:

Poisson Probability Function: If X is a random variable that follows a Poisson distribution, then

$$p(x) \;=\; \frac{\mu^x e^{-\mu}}{x!} \qquad \text{for} \quad x = 0, 1, 2, \ldots, \tag{4.7}$$

where μ is a positive constant (the distribution's mean) and $e = 2.7182\ldots$ is the exponential constant.

This probability function characterizes the ***Poisson distribution***. We refer to μ, the mean of the distribution, as its ***parameter***. It controls not only the distribution's center, but its shape and spread too. We write

$$\boldsymbol{X \sim \textbf{Poisson}(\mu)}$$

to mean that X follows a Poisson distribution with mean μ. The figure below shows Poisson distributions with three different values of μ.

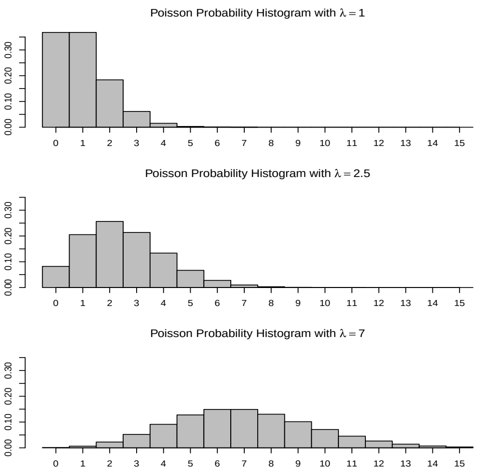


Figure 4.8: Poisson probability histograms with $\mu = 1$ (top), $\mu = 2.5$ (middle), and $\mu = 7$ (bottom).

Poisson distributions are all (at least slightly) right skewed. The larger μ is, the farther to the right the distribution's center will lie and the more spread out and bell-shaped the distribution will be. There's no rigid upper limit as to how large a Poisson variable might be, but very large values have negligibly small probabilities (see Figs. 4.8, 4.6, and 4.7). Example 4.12 shows how to compute probabilities using the probability function.

Mean and Standard Deviation of a Poisson Distribution

As mentioned, the parameter μ is the mean of the Poisson distribution, and is the distribution's center ("balancing point"). As μ increases, the distribution shifts to the right, but it also becomes more spread out (Fig. 4.8). In fact, μ is also the variance (squared standard deviation) of the distribution, that is, the mean and variance are equal:

Mean and Standard Deviation of the Poisson Distribution: The mean $\boldsymbol{\mu}_{\mathbf{pois}}$ of a Poisson(μ) distribution is

$$\mu_{\text{pois}} = \mu$$

and the standard deviation $\boldsymbol{\sigma}_{\mathbf{pois}}$ is

$$\sigma_{\text{pois}} = \sqrt{\mu}.$$

For counts of events in time periods of a given length, μ is the average count per time period. Longer periods have larger μ values and shorter periods smaller ones. For one-unit time periods, μ is the average number of events per unit of time, that is, it's the ***rate*** at which the events occur. Similarly, for counts of events over one-unit spatial areas, μ is the average number of events per unit area, that is, it's the spatial ***density*** of the events.

In practice, the exact value of μ will be unknown, so it will need to be estimated, for example using historical data. This was the approach used in Example 4.10, where the mean number of hurricanes (μ) was estimated to be 1.68 per year.

The next example shows how to compute probabilities using the probability function.

Example 4.12: Poisson Mean, Standard Deviation, and Probabilities

Fig. 4.6 is a Poisson distribution with mean $\mu = 1.68$, used to model the number of hurricanes in a year. Thus the standard deviation is $\sqrt{\mu} = \sqrt{1.68} = 1.30$ hurricanes, so a typical year will have about 1.68 hurricanes, plus or minus 1.30.

We can calculate the probability of any number of hurricanes using the probability function (4.7) with $\mu = 1.68$. For example, the probability that there will be three hurricanes is

$$\begin{aligned} p(3) &= \frac{\mu^3 e^{-\mu}}{3!} \\ &= \frac{1.68^3 e^{-1.68}}{3!} \\ &= 0.1473, \end{aligned}$$

and the probability that there won't be any is

$$\begin{aligned} p(0) &= \frac{1.68^0 e^{-1.68}}{0!} \\ &= 0.1864. \end{aligned}$$

Letting X denote the number hurricanes, the probability that there will be between two and four hurricanes, inclusive, is

$$\begin{aligned} P(2 \leq X \leq 4) &= p(2) + p(3) + p(4) \\ &= \frac{1.68^2 e^{-1.68}}{2!} + \frac{1.68^3 e^{-1.68}}{3!} + \frac{1.68^4 e^{-1.68}}{4!} \\ &= 0.4722. \end{aligned}$$

The Poisson Distribution When μ is Large

Poisson distributions are all (at least slightly) right skewed, but the larger μ is, the more bell-shaped they will be (Fig. 4.8). They'll also be shifted to the right and more spread out. If μ is large enough, the Poisson distribution will be indistinguishable from a perfect bell-shaped *normal distribution*, described later in this chapter.

4.4 Continuous Probability Distributions

4.4.1 Introduction

Continuous random variables, recall, can take *any* value over an entire continuum. Their probability distributions are represented by smooth curves called ***probability density functions*** (or ***curves***), denoted $\boldsymbol{f(x)}$. A density curve can be thought of as a smooth histogram of a population from which an elementary unit is *randomly* selected. The higher the curve is above the x-axis, the more prevalent the x values below the curve are in the population and therefore the more likely they are to result from the random selection.

Example 4.13: Continuous Probability Distribution

Consider again exposing a radon detector to 100 pCi/L of radon (Example 1.3 in Chapter 1). Even if the detector is properly calibrated, random *measurement error* makes it unlikely that the reading X will equal 100 exactly. Readings that are subject to measurement error are usually modeled by a bell-shaped density curve (the so-called *normal* distribution) like the one below.

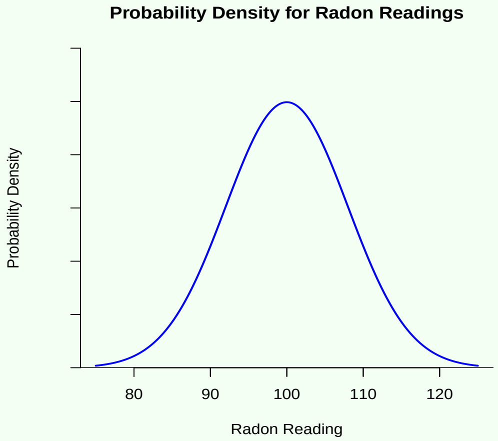


Figure 4.9: Probability density curve for radon detector readings, with measurement error, when exposed to 100 pCi/L of radon.

The curve is centered on 100, and its height indicates that readings close to this value are more likely than those farther from it. The symmetric shape indicates that X is equally likely to fall a given distance above or below the true value 100.

Example 4.14: Continuous Probability Distribution

Problems 2.6 and 3.20 in Chapters 2 and 3 described a study of contaminants in eggs of herons and egrets near Hong Kong.

The authors of the study indicate that the chlordane concentration X (ng/g) in a randomly selected heron egg from the A Chau egretry can be modeled by the right-skewed distribution density curve (the so-called *lognormal* distribution) shown below.

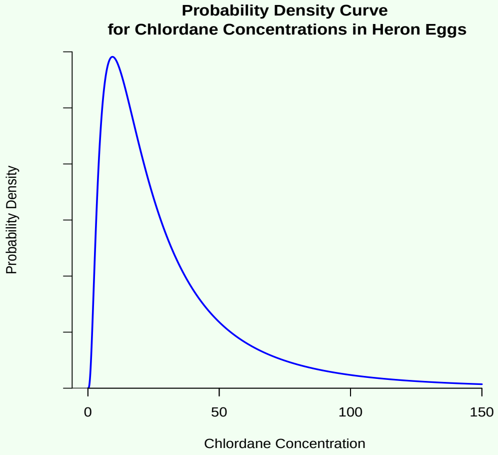


Figure 4.10: Probability density curve for chlordane concentrations in randomly selected heron eggs.

The sharp peak near zero indicates that chlordane concentrations are likely to be close to zero, but the long tail extending to the right indicates that they can occasionally be very large. In fact, for the eggs sampled in the study, most concentrations were close to zero, but a few were substantially larger (up to 75 ng/g), which lends credibility to the density curve shown in the graph.

In general, if a probability distribution accurately represents a *population* and we were to select a *random sample* from that population, we'd expect a histogram of the data to approximately resemble the probability density curve.

For a continuous random variable X, the probability that X will fall in a given interval on the x-axis is the *area under the density curve* over that interval.

Probabilities Involving Continuous Random Variables: For a continuous random variable X and any numbers a and b (with $a < b$),

$$\begin{aligned} P(a < X < b) &= \text{Area under the density curve between } a \text{ and } b. \\ P(X < b) &= \text{Area under the density curve to the left of } b. \\ P(X > b) &= \text{Area under the density curve to the right of } b. \end{aligned}$$

The figure below illustrates.

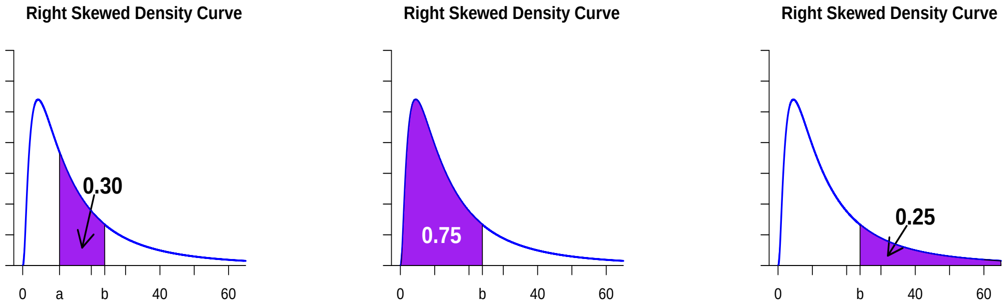


Figure 4.11: Probability density curves representing a right skewed population. Thirty percent percent of the population values are *between* a and b. Seventy-five percent are *less* than b. Twenty-five percent are greater than b.

Example 4.15: Continuous Probability Distribution

The density function from Example 4.13 is shown again below, this time with a shaded area under the curve corresponding to the probability that a radon reading will fall between 90 and 110 pCi/L.

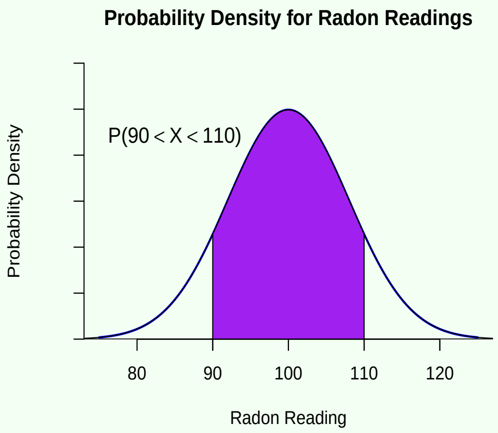


Figure 4.12: Probability density curve for radon detector readings, with $P(90 \leq X \leq 110)$ shown as the shaded area.

Because probabilities are areas under a density curve, all density curves satisfy the following conditions.

Properties of Probability Density Functions: Every continuous probability density curve $f(x)$ has the following properties:

1. The density curve lies on or above the x-axis, that is, $f(x) \geq 0$ for all x.
2. The total area under the curve is one, which is the probability of the random variable falling somewhere in its range of possible values.

Note: When working with *continuous* random variables, it makes no difference whether we write probabilities using "≤" or "<". The same is true for "≥" versus ">". Thus, for example, for any number a,

$$P(X \leq a) \;=\; P(X < a).$$

The reason is that if X is a *continuous* random variable and a is a single number, then $P(X = a) = 0$. For example, the chance that the chlordane concentration in a randomly selected heron egg will *exactly* equal, say, 45.7941 ng/g is zero. This makes intuitive sense because a single number (like 45.7941) is an infinitesimally small subset of the range of values that the random variable could take.

4.4.2 The Mean and Standard Deviation of a Continuous Probability Distribution

Recall that we measure the *center* of a probability distribution by its ***mean*** $\boldsymbol{\mu}$. For a *continuous* variable, μ is the point along the horizontal axis at which the density curve would balance if the area beneath it was weight resting on the axis.

The mean μ has the same interpretations as it did for discrete distributions:

1. It's the *mean of the population* when the probability distribution represents a population.
2. It's a *typical value* of the variable.

Recall also that we measure the *spread* in a probability distribution (that is, variation in values of the random variable) by its ***standard deviation*** $\boldsymbol{\sigma}$, and we call the square of the standard deviation, $\boldsymbol{\sigma^2}$, the ***variance*** of the distribution. A larger value of σ corresponds to a more spread-out distribution, and this is true for *continuous* distributions just as it was for discrete ones.

The standard deviation of a *continuous* distribution has the same interpretations as it did for discrete ones:

1. It's the *standard deviation of the population* when the probability distribution represents a population.
2. It's the size of a *typical deviation* of the variable away from the mean μ.

4.4.3 Percentiles of a Continuous Probability Distribution

The ***median***, or ***50th percentile***, of a continuous distribution, denoted by $\boldsymbol{\tilde{\mu}}$, is the value below which 50% of the population lies (and above which the other 50% lies). Thus the random variable X has a fifty-fifty chance of falling above or below $\tilde{\mu}$.

Whereas the mean μ would be the "balancing point" if a distribution was weight resting on the horizontal axis, the median $\tilde{\mu}$ is the "equal areas point" in the sense that half of the area under the density curve lies to the left of $\tilde{\mu}$ and the other half to the right, as seen in the left graph of Fig. 4.13. Notice in the graph that the mean is to the *right* of the median, a consequence of the right-skewness of the distribution. In general:

- For a *symmetric* distribution, the mean and median will be the same, i.e. $\mu = \tilde{\mu}$.
- For a *right skewed* distribution, the mean will be greater than the median, i.e. $\mu > \tilde{\mu}$.
- For a *left skewed* distribution, the mean will be less than the median, i.e. $\mu < \tilde{\mu}$.

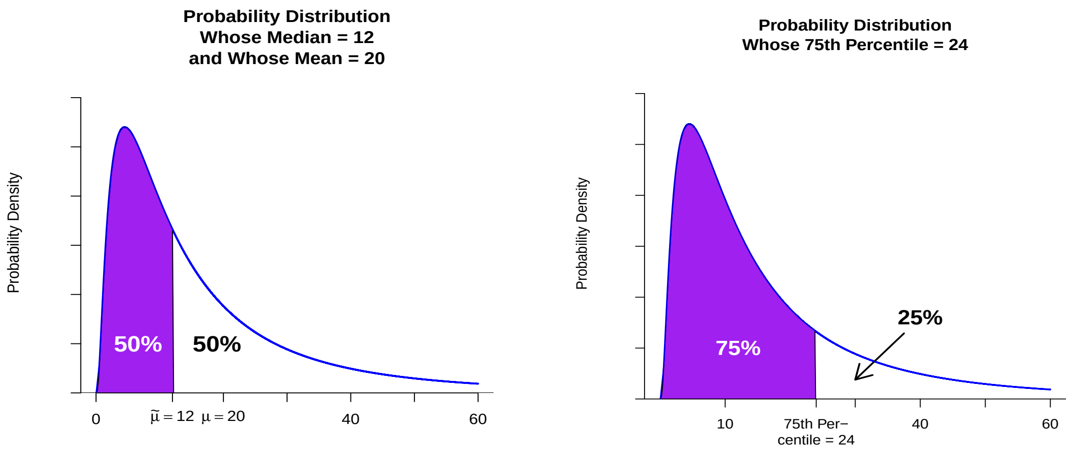


Figure 4.13: Probability distribution whose median is $\tilde{\mu} = 12$ and whose mean is $\mu = 20$ (left); the same distribution, whose 75th percentile is 24 (right).

The median of a distribution is sometimes called the distribution's *50th percentile* because 50% of the distribution lies to its *left*. Other ***percentiles*** are defined analogously. For example, the 75*th percentile* is the value below which 75% of the distribution lies (and above which the other 25% lies), and is marked on the horizontal axis in the right plot of Fig. 4.13.

A familiar example of a percentile is the so-called "100-year flood level" of a river. To see, note that the river's annual peak height is a random variable, X. The "100-year flood level" is the height for which there's only a 1 in 100 chance, or 0.01 probability, of being exceeded in any given year. So in 99% of years, the river's height remains below the "100-year flood level". In other words, the "100-year flood level" is the *99th percentile* of the distribution of X.

4.5 Probability Distributions for Continuous Measurements

In the absence of accurate information about a population's histogram shape, we choose from a set of stock *theoretical distribution* the one that we *think* describes the population. Then probabilities involving randomly selected elementary units and values of percentiles can be obtained from the theoretical distribution.

When the random variable is a *numerical measurement*, for example of a chemical contaminant concentration, it's usually *continuous*. Two commonly used continuous theoretical distributions are:

1. The normal distribution
2. The lognormal distribution

The first of these is the familiar bell-shaped distribution, and the second is a right skewed distribution commonly used in environmental science to model contaminant concentrations.

4.5.1 The Normal Distribution

Many populations follow the bell-shaped ***normal distribution***. Fig. 4.9 in Example 4.13 shows an example of a ***normal probability density curve***. Several others are shown in Fig. 4.14.

The Normal Probability Density Curve

The ***mean*** $\boldsymbol{\mu}$ and ***standard deviation*** $\boldsymbol{\sigma}$ determine, respectively, the center and spread of the normal distribution along the horizontal axis. They're referred to as the ***parameters*** of the distribution. To illustrate, Fig. 4.14 shows several normal density curves with different values of μ and σ.

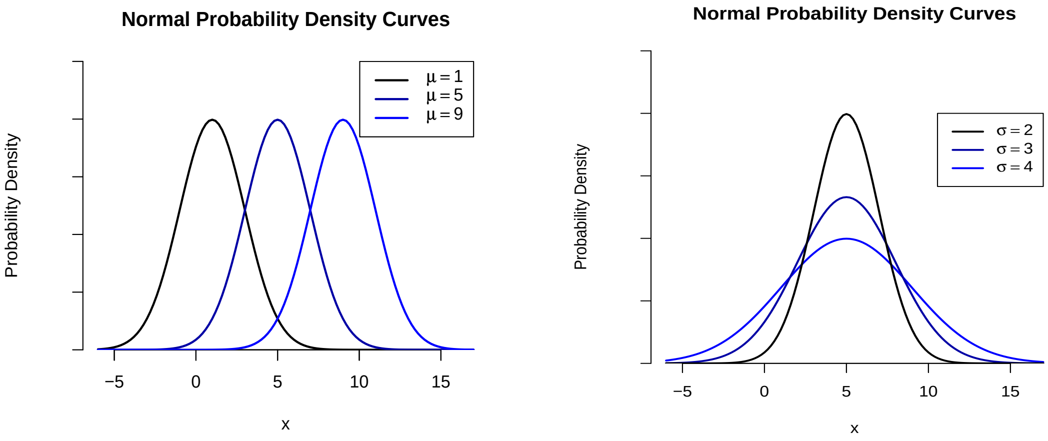


Figure 4.14: Normal distributions with different values of μ, but constant σ (left), and with constant μ, but different values of σ (right).

Note that μ is the balancing point, and a larger σ value corresponds to a more spread-out distribution.

We'll use the notation

$$\boldsymbol{X \sim \mathrm{N}(\mu,\, \sigma)}$$

to mean that the random variable X follows a normal distribution with mean μ and standard deviation σ.

Mean and Standard Deviation of the Normal Distribution

As mentioned, the parameters μ and σ are the mean and standard deviation of the normal distribution.

Mean and Standard Deviation of the Normal Distribution: The mean $\boldsymbol{\mu}_{\textbf{norm}}$ of a N(μ, σ) distribution is

$$\mu_{\text{norm}} \;=\; \mu$$

and the standard deviation $\boldsymbol{\sigma}_{\textbf{norm}}$ is

$$\sigma_{\text{norm}} \;=\; \sigma.$$

Also, because the distribution is symmetric, the mean and median are equal.

Median of the Normal Distribution: The median $\tilde{\boldsymbol{\mu}}_{\textbf{norm}}$ of a N(μ, σ) distribution is

$$\tilde{\mu}_{\textbf{norm}} \;=\; \mu.$$

The following fact tells us, among other things, that a normally distributed variable rarely falls more than about two standard deviations away from the mean, and it would be extremely rare for it to fall more than three standard deviations away.

The 68-95-99.7 Rule: If $X \sim \text{N}(\mu, \sigma)$, then

1. X will fall within one σ of μ with probability about 0.68.
2. X will fall within two σ's of μ with probability about 0.95.
3. X will fall within three σ's of μ with probability about 0.997.

This is depicted graphically below.

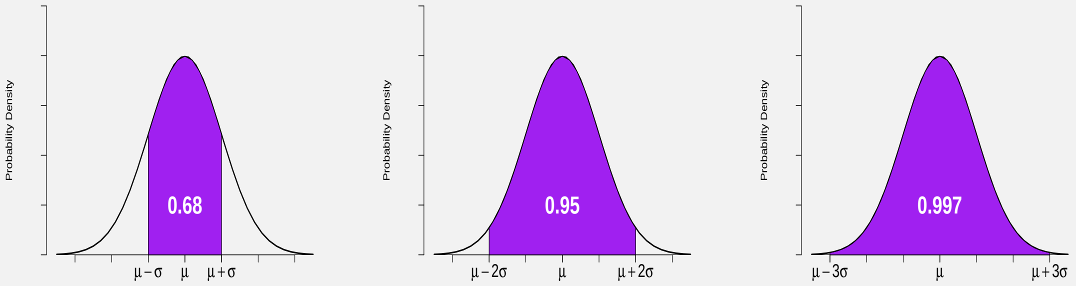


Figure 4.15: A normal random variable will be within one, two, and three standard deviations of the mean with probabilities (shaded areas) 0.68, 0.95, and 0.997, respectively.

Standardized Values and the Standard Normal Distribution

The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the ***standard normal distribution*** and denoted **N(0, 1)**. We can convert *any normal random variable* to a *standard normal* one using the following fact.

Fact 4.1 If $X \sim \text{N}(\mu, \sigma)$, and we convert X to a variable Z via

$$Z \;=\; \frac{X-\mu}{\sigma}, \tag{4.8}$$

then $Z \sim \mathrm{N}(0,\, 1)$.

For insight, the left histogram in the figure below shows a random sample $X_1, X_2, \ldots, X_n$ from a normal population with mean $\mu = 8$ and standard deviation $\sigma = 2$. It's centered on its mean, eight, and extends about three standard deviations (six units) away from eight in each direction.

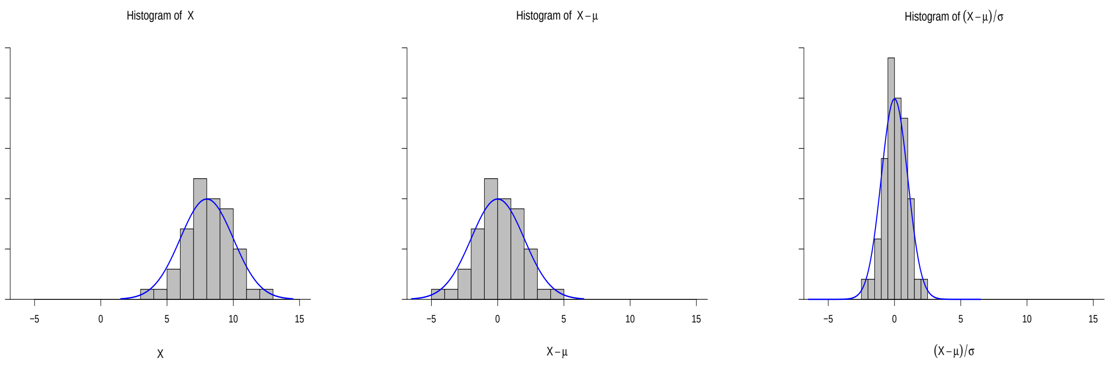


Figure 4.16: Histogram of a random sample from a N(8, 2) distribution (left); histogram after subtracting eight from each observation (center); histogram after subtracting eight from each observation and dividing by two (right).

Now suppose we use (4.8) to convert each X_i in the data set to a value Z_i, with

$$Z_i \;=\; \frac{X_i - 8}{2},$$

but we do it in two steps. In the first step we subtract eight from each X_i. This would shift the entire distribution to the left by eight units, centering it on zero, as in the middle histogram of Fig. 4.16. In the second step we divide each $X_i - 8$ by two, which would contract both ends of the zero-centered histogram, leading to the histogram shown on the right. That histogram is centered on zero and extends about three units in each direction. According to Fact 4.1, these data, $Z_1, Z_2, \ldots, Z_n$, can be treated as a random sample from a *standard normal* distribution.

When we convert a value X to a value Z using (4.8), we say that X has been ***standardized***, or converted to a ***z-score***. A standardized value, or z-score, is measured in standard deviations above or below the mean, also called ***standard units***. A z-score will be positive or negative depending on whether X is greater than or less than the mean μ. By the 68-95-99.7 Rule, it would be uncommon for the z-score of a normal variable X to differ from zero by more than about two, and extremely rare for it to differ from zero by more than three.

Finding Normal Probabilities

Probabilities involving a *standard normal* random variable can be obtained from a ***standard normal table*** (or ***z table***), which gives probabilities of the form $P(Z \leq z)$. For a general *normal* random variable X, probabilities can be obtained by converting to *standard units* via (4.8) and then using the *standard normal table*. More specifically, we have the following.

.

Finding Normal Distribution Probabilities: Suppose $X \sim \mathrm{N}(\mu,\, \sigma)$. Then for any values a and b (with $a < b$),

- $P(X < b) \;=\; P\left(Z \;<\; z_b\right)$.
- $P(X > a) \;=\; P\left(Z \;>\; z_a\right) \;=\; 1 \;-\; P\left(Z \;<\; z_a\right)$.
- $P(a < X < b) \;=\; P(z_a < Z < z_b) \;=\; P\left(Z \;<\; z_b\right) - P\left(Z \;<\; z_a\right)$.

where

$$z_a \;=\; \frac{a-\mu}{\sigma} \qquad \text{and} \qquad z_b \;=\; \frac{b-\mu}{\sigma}.$$

The following example illustrates.

Example 4.16: Normal Distribution Probabilities

Suppose $X \sim \mathrm{N}(8,\, 2)$. The probability that X will be less than 4.08 is the shaded tail area in the left graph below.

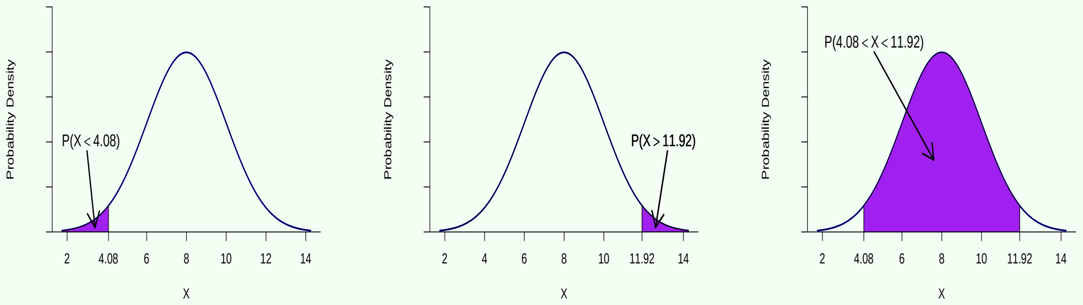


Figure 4.17: Probability density curve for X, with probabilities as shaded areas under the curve.

This probability is

$$\begin{aligned} P(X \;<\; 4.08) \;&=\; P(Z \;<\; -1.96) \\ &=\; 0.0250, \end{aligned}$$

where -1.96 is the z-score associated with 4.08, i.e.

$$z_{4.08} \;=\; \frac{4.08-\mu}{\sigma} \;=\; \frac{4.08-8}{2} \;=\; -1.96,$$

and the value 0.0250 was obtained from a standard normal table.

Similarly, the probability that X will be greater than 11.92 is the shaded area in the middle graph, and is given by

$$\begin{aligned} P(X \;>\; 11.92) \;&=\; P(Z \;>\; 1.96) \\ &=\; 1 - P(Z \;<\; 1.96) \\ &=\; 1 - 0.9750 \\ &=\; 0.0250, \end{aligned}$$

where 1.96 is the z-score associated with 11.92, i.e.

$$z_{11.92} \;=\; \frac{11.92-\mu}{\sigma} \;=\; \frac{11.92-8}{2} \;=\; 1.96,$$

and the value 0.9750 was obtained from a standard normal table.

Finally, the probability that X will fall between 4.08 and 11.92 is the shaded area in the right graph, and is given by

$$\begin{aligned} P(4.08 < X < 11.92) &= P(-1.96 < Z < 1.96) \\ &= P(Z < 1.96) - P(Z < -1.96) \\ &= 0.9750 - 0.0250 \\ &= 0.9500. \end{aligned}$$

Finding Normal Percentiles

Recall that a *percentile* of a distribution is a value below which a specified percentage of the distribution lies. Some percentiles of the standard normal distribution are shown below.

Some N(0, 1) Percentiles	
50th	0.00
95th	1.64
97.5th	1.96
99.5th	2.58

The latter three are depicted below.

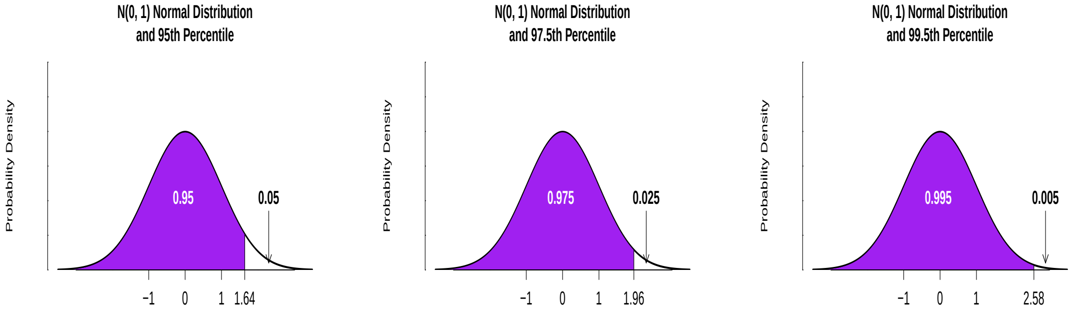


Figure 4.18: Standard normal distribution and its 95th, 97.5th, and 99.5th percentiles (1.64, 1.96, and 2.58, respectively).

To find other standard normal percentiles, scan the main body of the z table for the appropriate probability (e.g. 0.90 for the 90%th percentile) and determine the z value associated with that probability from the table margins.

We can use *percentiles* to characterize *middle percentages* of the standard normal distribution, as shown in the following table and the subsequent figures.

N(0, 1) Percentiles (Cont'd)	
Middle 90%	Between ±1.64
Middle 95%	Between ±1.96
Middle 99%	Between ±2.58

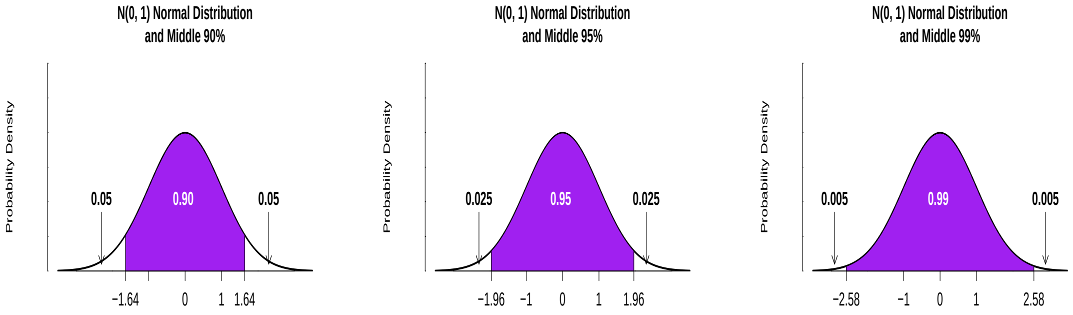


Figure 4.19: Standard normal distribution and percentiles demarcating its middle 90%, 95%, and 99% (1.64, 1.96, and 2.58, respectively).

Percentiles of other normal distributions are obtained by ”unstandardizing” the corresponding percentile of the *standard normal* distribution using the following.

Percentiles of a Normal Distribution: A percentile $\boldsymbol{x}$ of a N(μ, σ) distribution is

$$x \;=\; \mu + z\sigma, \tag{4.9}$$

where $\boldsymbol{z}$ is the corresponding percentile of the N(0, 1) distribution.

The equation for ”unstandardizing” z (4.9) was obtained by solving $z = (x - \mu)/\sigma$ for x. The intuition behind the right side of the equation is that z is measured in standard units, or standard deviations above the mean, so x needs to be that many standard deviations above the mean.

Example 4.17: Normal Distribution Percentiles

We’ll find the 5th and 95th percentiles of the N(8, 2) distribution. These are the values marked x in the two graphs below, and together they characterize the middle 90% of the distribution.

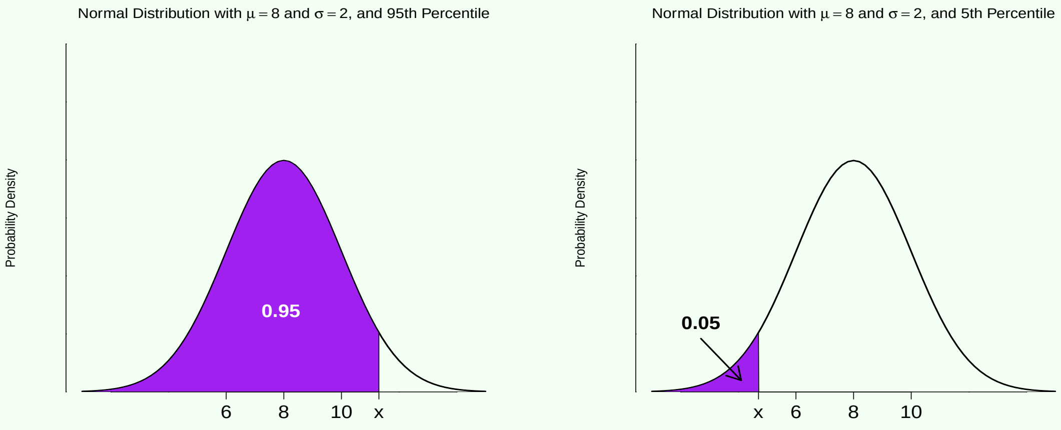


Figure 4.20: N$(8, 2)$ distribution with the 95th percentile marked x on the horizontal axis (left), and with the 5th percentile marked x on the horizontal axis (right).

To find the 95th percentile of the distribution, we "unstandardize" the 95th percentile of the standard normal distribution, $z = 1.64$, using (4.9). This gives

$$x \;=\; 8 + (1.64)(2) \;=\; 11.28.$$

To find the 5th percentile, we "unstandardize" the 5th percentile of the standard normal distribution, which is $z = -1.64$. This gives

$$x \;=\; 8 + (-1.64)(2) \;=\; 4.72.$$

Because 5% of the distribution lies to the left of 4.72 and 5% to the right of 11.28, the *middle 90%* lies between these two values.

Linear Transformations of Normal Random Variables

We now turn to some facts that are relevant to converting between measurement scales for the normally distributed variables. Each fact is preceded by an example illustrating a situation to which the fact may be applied.

Example 4.18: Linear Transformations of Random Variables

Morphological (shape-related) features of fish are sometimes used as biological indicators of pollution in streams. Such features are easy to measure, so deformities or other morphological changes over time can alert us to a pollution problem that might otherwise go undetected.

One bioindicator of pollution is fish fin lengths. The fin length X, in inches, of a randomly selected fish is a random variable. There are 2.54 centimeters in an inch, so its fin length in centimeters is $2.54X$. If the population mean fin length in inches is $\mu_{\text{in}} = 0.25$, and the standard deviation is $\sigma_{\text{in}} = 0.07$, what could be said about the population mean and standard deviation if fin lengths were measured in centimeters?

The conversion from inches to centimeters just described is an example of a ***linear transformation*** of the random variable X, that is, a conversion of the form $\boldsymbol{aX + b}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are constants. The following fact gives the mean and standard deviation of the distribution of the variable after a linear transformation.

Fact 4.2 If X is *any* random variable whose distribution has mean μ_X and standard deviation σ_X, then for any constants a and b, the new random variable $aX + b$ follows a distribution whose mean is

$$\mu_{aX+b} \;=\; a\,\mu_X + b \tag{4.10}$$

and whose standard deviation is

$$\sigma_{aX+b} \;=\; |a|\,\sigma_X. \tag{4.11}$$

Example 4.19: Linear Transformations of Random Variables

Continuing with Example 4.18, randomly selecting a fish from a population whose fin lengths are measured in inches, *then* converting it to centimeters, is equivalent to measuring the population of fin lengths in centimeters, *then* randomly selecting a fish.

The conversion of a randomly selected fish's fin length X from inches to centimeters is the linear transformation $2.54X$. By the previous fact, if the population mean and standard deviation of fin lengths measured in inches are $\mu_{\text{in}} = 0.25$ and $\sigma_{\text{in}} = 0.07$, then the population mean and standard deviation of fin lengths measured in centimeters are

$$\mu_{\text{cm}} \;=\; 2.54\mu_{\text{in}} \;=\; 2.54(0.25) \;=\; 0.635$$

and

$$\sigma_{\text{cm}} \;=\; |2.54|\,\sigma_{\text{in}} \;=\; 2.54(0.06) \;=\; 0.178.$$

Fact 4.2 shows how to determine the *mean* and *standard deviation* of the distribution of a linearly transformed random variable, but doesn't say anything about its *shape*.

Example 4.20: Linear Transformations of Random Variables

Continuing from the previous two examples, suppose the fin length X of a randomly selected fish, in inches, follows a *normal* distribution. What can be said about distribution of the fin length $2.54X$ in centimeters?

The next fact tells us that if the original random variable is normally distributed, any linear transformation of that random variable will be normally distributed too.

Fact 4.3 Suppose $X \sim \text{N}(\mu_X,\, \sigma_X)$. Then for any constants a and b,

$$aX + b \;\sim\; \text{N}\left(\mu_{aX+b},\; \sigma_{aX+b}\right), \tag{4.12}$$

where μ_{aX+b} and μ_{aX+b} are given by (4.10) and (4.11). In other words, any linear transformation $aX + b$ of a normally distributed random variable X is also normally distributed, with mean and standard deviation given in Fact 4.2.

Example 4.21: Linear Transformations of Random Variables

If the fin length X of a randomly selected fish, in inches, follows a N(0.25, 0.07) distribution, then by the previous fact and the results from Example 4.19, the fin length $2.54X$ in centimeters follows a N(0.635, 0.178) distribution.

4.5.2 The Lognormal Distribution

Environmental variables such as pollutant concentrations often exhibit right skewed distributions, with most values being bunched together near zero but a small fraction of values being quite large. A useful theoretical probability distribution for describing random variables that exhibit this behavior is the ***lognormal distribution***. Lognormal distributions are right skewed and lie entirely to the right of zero.

A few quotes indicate how widespread their use has become for modeling environmental data:

> ”Concentrations of various measured substances have distributions that are lognormal, or nearly so. Examples include radionuclides in soil, pollutants in ambient air, indoor air quality, trace

> metals in streams, metals in biological tissue, and calcium in human remains” [13].
>
> ”The assumption of lognormality for environmental data is fairly universal” [16].
>
> ”The lognormal distribution has become a common choice to represent intrinsically positive and often highly skewed environmental data in statistical analysis” [2].
>
> ”It is commonly the case that environmental data are lognormal or well approximated by a lognormal distribution” [4].

The Lognormal Probability Density Function

The center, spread, and aspects of the shape of the lognormal distribution are controlled by the values of the distribution’s two ***parameters***, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, but these are *not* its mean and standard deviation. We’ll see later how they’re interpreted. The graphs in Fig. 4.21 depict ***lognormal probability density curves*** for different values of μ and σ.

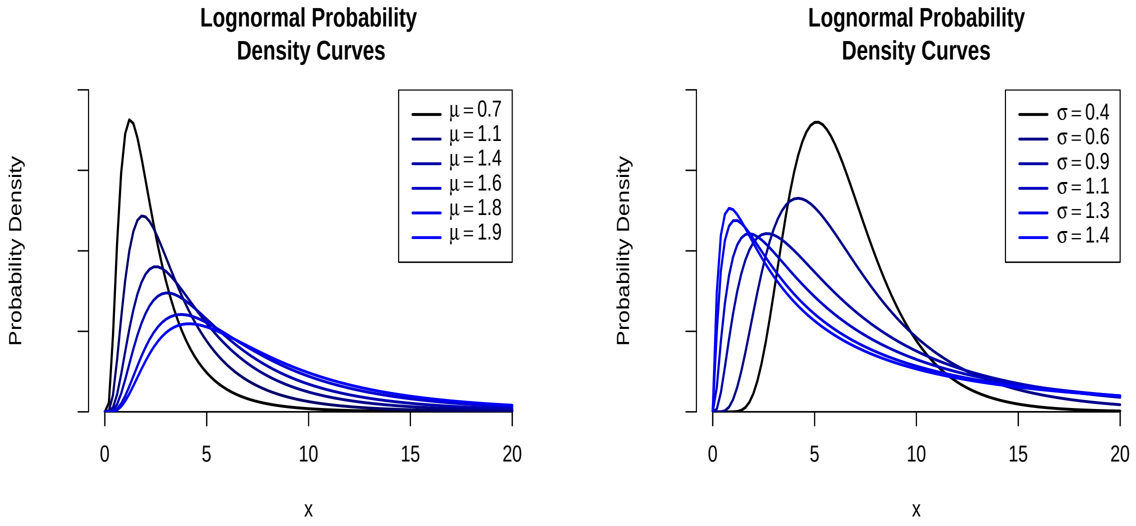


Figure 4.21: Lognormal distributions with different values of μ but constant σ (left), and with constant μ but different values of σ (right).

As can be seen, lognormal distributions can have slightly different shapes, but they’re all right skewed. In fact, the density curve describing chlordane concentrations in Fig. 4.10 of Example 4.14 is also a lognormal distribution.

We write

$$\boldsymbol{X \sim \mathrm{LN}(\mu,\, \sigma)}$$

to mean that the random variable X follows a lognormal distribution with parameters μ and σ.

The following fact explains how the lognormal distribution gets its name, and provides an interpretation of the parameters μ and σ. It says we can convert a *lognormal* random variable to a *normal* random variable by taking it’s (natural) log. In words, if X is lognormal, then its *log is normal.*

Fact 4.4 If $X \sim \mathrm{LN}(\mu,\, \sigma)$, and we make the (natural) log transformation

$$Y \;=\; \log(X),$$

then Y is a new random variable and

$$Y \sim \mathrm{N}(\mu, \sigma). \tag{4.13}$$

Thus μ and σ are the mean and standard deviation of the *normal* distribution of the *log-transformed* variable.

Example 4.22: Lognormal Distribution

To illustrate the effect of the making the *log transformation* on right skewed, lognormal data, the following $n = 50$ observations were obtained from a LN$(5, 1)$ distribution using a computer random number generator.

202.7	347.2	300.5	812.3	38.6	83.9	157.5	35.3	180.6	152.4
90.4	95.5	234.7	618.9	149.2	169.6	427.6	89.1	204.3	90.9
681.5	55.4	625.5	45.7	68.9	828.4	21.3	561.4	315.8	97.4
95.6	69.5	650.0	77.1	367.1	49.2	478.9	182.3	273.8	33.2
313.9	107.9	86.4	287.3	194.3	203.0	164.9	1307.0	209.4	164.7

A histogram of these observations is shown below on the left along with the LN$(5, 1)$ density curve representing the population from which the sample was generated.

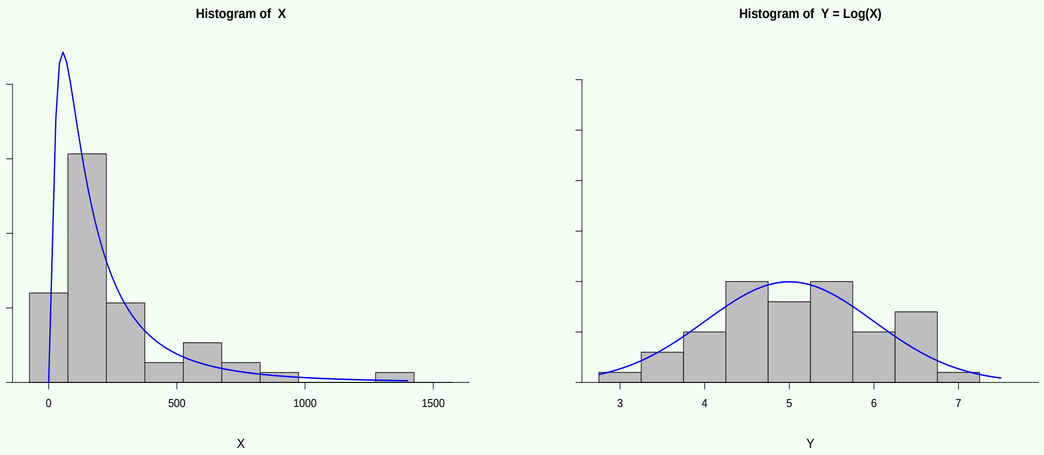


Figure 4.22: Histogram of a random sample from a LN$(5, 1)$ distribution with the LN$(5, 1)$ density curve superimposed (left). Histogram of the sample after making the log transformation of each value, with the N$(5, 1)$ density curve superimposed (right).

After making the (natural) log transformation of each of the 50 observations, the data values (now on the log scale) are:

5.31	5.85	5.71	6.70	3.65	4.43	5.06	3.57	5.20	5.03
4.50	4.56	5.46	6.43	5.01	5.13	6.06	4.49	5.32	4.51
6.52	4.01	6.44	3.82	4.23	6.72	3.06	6.33	5.76	4.58
4.56	4.24	6.48	4.35	5.91	3.90	6.17	5.21	5.61	3.50
5.75	4.68	4.46	5.66	5.27	5.31	5.11	7.18	5.34	5.10

A histogram of these log-transformed values is shown on the right in Fig. 4.22 along with the N$(5, 1)$ curve. The log-transformed data can be treated as a random sample from a N$(5,\ 1)$ distribution.

Mean and Standard Deviation of the Lognormal Distribution

Although the parameters μ and σ are *not* themselves the mean and standard deviation of the lognormal distribution, they do determine the values of the mean and standard deviation, according to the following relationships.

Mean and Standard Deviation of the Lognormal Distribution: The mean $\boldsymbol{\mu_{\text{lnorm}}}$ of the LN(μ, σ) distribution is

$$\mu_{\text{lnorm}} \;=\; e^{\mu+\frac{\sigma^2}{2}}$$

and the standard deviation $\boldsymbol{\sigma_{\text{lnorm}}}$ is

$$\sigma_{\text{lnorm}} \;=\; \sqrt{(e^{\sigma^2}-1)e^{2\mu+\sigma^2}}$$

where e is the exponential constant 2.718282....

As for other probability distributions, the mean of the lognormal distribution is its ”balancing point”, and the standard deviation measures its spread and represents a typical deviation of the random variable away from its mean.

The median of the lognormal distribution is also determined by the parameter μ, and is given by the following.

Median of the Lognormal Distribution: The median $\boldsymbol{\tilde{\mu}_{\text{lnorm}}}$ of the LN(μ, σ) distribution is

$$\tilde{\mu}_{\text{lnorm}} \;=\; e^{\mu}. \tag{4.14}$$

where once again e is the exponential constant 2.718282....

Comment: The median e^{μ} of the lognormal distribution resembles the geometric mean $e^{\bar{Y}}$ of Chapter 3, where $\bar{Y}$ was the mean of the *logs* $Y_1, Y_2, \ldots, Y_n$ of a right skewed sample $X_1, X_2, \ldots, X_n$. We now know that if the right skewed sample is from a LN(μ, σ) distribution, their logs can be treated as a sample from a N(μ, σ) distribution. In this case, $\bar{Y}$ is an estimator of μ, and so the geometric mean $e^{\bar{Y}}$ is an estimator of the median e^{μ} of the lognormal distribution. But the sample median (of the right skewed sample $X_1, X_2, \ldots, X_n$) is another estimator of the population median e^{μ}. Hence the comment in Chapter 3 that for right skewed data, the geometric mean and sample median are approximately equal.

Finding Lognormal Probabilities

We can find probabilities involving a lognormal random variable X by first making the log transformation and then using the normal distribution.

For example, suppose $X \sim$ LN$(\mu,\, \sigma)$ and we want to find, say, $P(X\,<\,b)$ for some value $b > 0$. Letting $Y = \log(X)$, we know that $Y \sim$ N$(\mu,\, \sigma)$, and we have

$$\begin{aligned} P(X\,<\,b) \;&=\; P\left(\log(X)\,<\,\log(b)\right) \\ &=\; P\left(Y\,<\,\log(b)\right) \\ &=\; P\left(Z\,<\,z_{\log(b)}\right) \end{aligned}$$

where

$$z_{\log(b)} \;=\; \frac{\log(b)-\mu}{\sigma}.$$

The probability on the last line can be obtained from a standard normal table. Probabilities of the form $P(X\,>\,b)$ and $P(a\,<\,X\,<\,b)$ can be found in a similar manner.

Finding Lognormal Percentiles

Percentiles of a lognormal distribution can be obtained by taking the *antilog* of the percentile of the corresponding normal distribution.

To see, suppose x represents the 95th percentile of a LN(μ, σ) distribution, so 95% of *this* distribution lies to the left of x. Then 95% of the N(μ, σ) distribution lies to the left of $\log(x)$. But this means, from (4.9), that

$$\log(x) \;=\; \mu + z\sigma,$$

where $z = 1.64$ is the 95th percentile of the standard normal distribution. Taking the *antilog* of both sides (recall that the *antilog* of a number a is e^a, and the *antilog* of $\log(x)$ is just x), we get the desired percentile x. To summarize, we have the following.

Percentiles of a Lognormal Distribution: A percentile $\boldsymbol{x}$ of a LN(μ, σ) distribution is

$$x \;=\; e^{\mu+z\sigma},$$

where $\boldsymbol{z}$ is the corresponding percentile of the N(0, 1) distribution.

4.6 Measurement Error and the Limit of Detection

4.6.1 A Model for Measurement Error

As mentioned in Example 4.13, the normal distribution is used to model measurements that are subject to measurement error, such as a radon detector reading. The model assumes that a measurement X follows a N(μ, σ) distribution, that is,

$$X \sim \mathrm{N}(\mu,\, \sigma),$$

where $\boldsymbol{\mu}$ is the true (unknown) concentration of the substance being measured and $\boldsymbol{\sigma}$ is determined by the *precision* of the instrument used to make the measurement, with smaller σ values reflecting more precise measurements. The model says that *on average*, measurements will equal the true value μ, but any *particular* measurement is a random variable typically differing from μ by an amount σ.

The ***measurement error***, denoted $\boldsymbol{\epsilon}$, is defined as the difference between the observed measurement X and the true value μ:

$$\epsilon \;=\; X - \mu. \tag{4.15}$$

The measurement error will be positive or negative depending on whether the measurement X is larger or smaller than μ. Using (4.15), we can express X as the true value plus measurement error,

$$X \;=\; \mu + \epsilon\,,$$

where, for consistency with the assumption that $X \sim \mathrm{N}(\mu,\, \sigma)$, we assume that

$$\epsilon \sim \mathrm{N}(0,\, \sigma),$$

which specifies that *on average*, the measurement error ϵ will equal zero. The next example puts all of this into context.

Example 4.23: A Model for Measurement Error

In Example 4.13, the normal distribution was used to model a radon detector reading X when exposed to 100 pCi/L of radon. The density curve in Fig. 4.9 shows a N(100, 8) distribution, considered to be a reasonable model for X. Thus we write

$$X \sim \mathrm{N}(100,\, 8).$$

This can be expressed as

$$X \;=\; 100 + \epsilon,$$

where

$$\epsilon \sim \mathrm{N}(0,\, 8).$$

In practice, the value of σ (eight in this example) would either be provided by the manufacturer of the radon detector or estimated from repeated exposures of the detector to a fixed quantity of radon.

4.6.2 The Limit of Detection

The presence of measurement error when measuring a chemical concentration means it's possible to obtain a non-zero, positive measurement even if in fact the true concentration is zero (that is, even if the chemical analyte isn't present). For this reason, measurement values that are very close to zero, yet still positive, are usually considered to be indistinguishable from zero.

A ***limit of detection*** (***LOD***) is a value below which measurement readings are called ***nondetects*** and are considered to be indistinguishable from zero. Readings above the LOD are considered to be legitimate, non-zero values. The LOD is based the notion that the number of *false positives* should be kept in check. A ***false positive*** (or ***type I error***) occurs when a ***blank*** specimen (one whose true concentration is zero) erroneously produces a reading above the LOD. False positives are less likely when a high LOD value is used. The figure below depicts false positives as values above the LOD and shows the normal distribution as a model for measurement error.

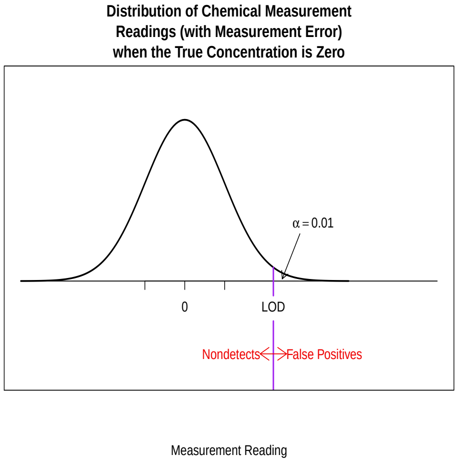


Figure 4.23: Normal distribution as a model for measurements (with measurement error) made on a blank specimen. The limit of detection (LOD) is chosen so that the probability of a false positive (denoted α) is 0.01.

The LOD is established so as to keep the probability of a false positive at some specified small value $\boldsymbol{\alpha}$. Usually α is chosen to be 0.01, so that only 1% of all measurements made on a blank specimen will result in false positives. Using a $\mathrm{N}(0,\, \sigma)$ distribution to model measurements made on a blank specimen, where σ is determined by the precision of the measurement instrument, the LOD is the 99th percentile of this distribution. Thus, from (4.9),

$$\begin{aligned} \text{LOD} &= 0 + z\sigma \\ &= 0 + 2.33\sigma \qquad (4.16) \\ &= 2.33\sigma. \end{aligned}$$

(The value $z = 2.33$ is the 99th percentile of the standard normal distribution and was obtained from a standard normal table.)

The LOD will be close to zero when the measurement instrument is precise, that is, when σ is small. For example, if $\sigma = 0.009$ mg/L, then the LOD is 0.021 mg/L, and in this case readings below 0.021 mg/L would be considered nondetects and would be flagged as such in a data set.

The general formula for the LOD is the following.

> **Limit of Detection When $\boldsymbol{\sigma}$ is Known**: For a desired false positive probability α, the limit of detection is
>
> $$\text{LOD} = z_{1-\alpha}\sigma,$$
>
> where $\boldsymbol{z_{1-\alpha}}$ is the $100(1-\alpha)$th percentile of the N(0, 1) distribution (usually $\alpha = 0.01$ or $\alpha = 0.05$) and $\boldsymbol{\sigma}$ is the measurement standard deviation as determined by the instrument's precision.

The presence of measurement error when measuring a chemical concentration can also produce a measurement value that's zero (or very close to it) even if in fact the true concentration is greater than zero (that is, even if the chemical analyte is present). A ***false negative*** (or ***type II error***) occurs when a specimen whose true concentration is *greater than zero* erroneously produces a reading *below* the LOD. For example, the figure below depicts false negatives as values below the LOD and shows that when the true concentration is equal to the LOD, the probability that a measurement will be a false negative, denoted $\boldsymbol{\beta}$, is 50%.

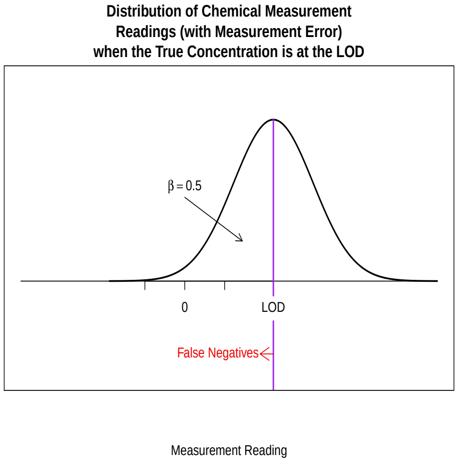


Figure 4.24: Normal distribution as a model for measurements (with measurement error) made on a specimen whose true concentration is at the limit of detection (LOD). The probability of a false negative (denoted β) is 0.5.

There's a balancing act that's performed when a LOD value is determined. On the one hand, the LOD should be large enough that false positives will be rare (e.g. they occur for only 1% measurements made on blank specimens). But on the other hand, a large LOD can result in a high rate of false negatives, especially when the true concentration being measured is close to (but still greater than) zero. It's generally accepted that using $\alpha = 0.01$ or 0.05 gives a reasonable balance between false positive and false negative rates.

Note: Often in practice, the value of the true standard deviation σ won't be known and must be estimated by measuring the concentration in a blank specimen several times and using the sample standard deviation S as an estimate of σ. When S is used in place of σ, the value $z = 2.33$ in (4.16) should be replaced by the 99th percentile of the so-called t *distribution*. This is discussed in Chapter 6.

Comment: There are several other detection criteria that may be used instead of the LOD for distinguishing measured values from zero (e.g. the "LLD," "IDL," and "MDL"). These are reviewed in [5].

4.7 Problems

4.1 In each situation below, decide whether it would more appropriate to model X as a binomial or a Poisson random variable.

a) The wind turbines at a wind power plant kill an average of 150 bats per year, on average. X is the number of bats killed in a given year.

b) To control otter populations, a state wildlife management department places fifteen live otter traps at various sites along streams throughout the region. After a day, each trap will be occupied or not. X is the number of traps that are occupied among the fifteen that were set.

c) During a thunderstorm, X is the number of lightning strikes over specific 20 minute period.

d) In a randomly selected $1\,\mathrm{m}^2$ quadrat, X is the number of Armadillidiidae (rolly pollies or pill bugs) present.

e) Eight water specimens collected randomly from a stream are tested for the presence or absence of *E. Coli.* X is the number that test positive among the eight.

4.2 According to the *Colorado Springs Gazette* (Dec. 1, 2006), 8.6% of all cars tested for emissions in El Paso, Larimer and Weld counties fail the test. Let X be the number of cars that fail the emissions test out of the next 10 cars that are tested.

a) What kind of probability distribution does X follow? Give the name of the distribution and the value(s) of its parameter(s).

b) Find $P(X = 3)$, the probability that exactly three cars fail the test.

c) Find $P(X = 1)$, the probability that exactly car one fails the test.

4.3 One way of assessing the toxicity of water is to rear aquatic organisms in the water and observe their rates of survival over a specified period of time.

It's been shown that the probability of a single individual *Ceriodaphnia dubia* (a freshwater microcrustacean) surviving for seven days in water from streams near the the Oak Ridge National Laboratory in Tennessee is 0.94 [15].

Suppose that 10 *Ceriodaphnia dubia* will be reared in the water and kept there for seven days. Let X denote the number of *Ceriodaphnia* that survive among the 10.

a) What kind of probability distribution does X follow? Give the name of the distribution and the value(s) of its parameter(s).

b) Find $P(X = 8)$, the probability that exactly 8 *Ceriodaphnia* will survive.

c) Find $P(X = 10)$, the probability that all 10 *Ceriodaphnia* will survive.

4.4 The cuckoo (*Cuculus canorus* L.) is a bird species that lays its eggs in the nests of other species of birds called hosts. The female cuckoo lays one egg per host nest. If the cuckoo egg isn't rejected by the host, it usually hatches before the host eggs do, and when it does the newly hatched cuckoo chick balances the host eggs on its neck one by one and ejects them from the nest.

A host may reject a cuckoo egg if its appearance is different from the host's own eggs. A study of various host bird species found that a reed warbler (*Acrocephalus scirpaceus* Herman) will reject a cuckoo egg with probability $p = 0.62$ and a meadow pipit (*Anthus pratensis* L.) will reject one with probability $p = 0.48$ [6].

a) If a cuckoo egg is laid in each of 12 reed warblers' nests, find the probability that exactly 8 of them will be rejected.

b) If a cuckoo egg is laid in each of 10 meadow pipits' nests, find the probability that exactly 6 of them will be rejected.

4.5 A botanist is investigating the germination (sprouting) properties of a new variety of hybrid wheat. The probability that any given hybrid seed will germinate under standard growing conditions is 0.6.

The botanist randomly selects five of the seeds and plants them under standard growing conditions. Let X denote the number of seeds that germinate among the five.

a) What kind of distribution does X follow? Give the name of the distribution and the value(s) of its parameter(s).

b) Calculate $P(X = 4)$, the probability that exactly four seeds germinate.

c) Now calculate the probabilities of 0, 1, 2, 3, 4, and 5 seeds germinating, and use these to draw a probability histogram showing the distribution of X.

d) Calculate the mean and standard deviation of the probability distribution of X, and mark these on the horizontal axis of the probability histogram.

4.6 A study suggests suggests that the number of *Philonthus fuscipennis* beetles X in a $1\,\mathrm{m}^2$ area follows a Poisson distribution with parameter $\mu = 7.75$. Thus, on average, there are 7.75 beetles per $1\,\mathrm{m}^2$ area.

Find $P(X = 10)$, the probability that a $1\,\mathrm{m}^2$ quadrat will have exactly 10 beetles.

4.7 On average, about 43.2 people are killed by lightning each year in the U.S. (see Problem 3.2 in Chapter 3). The actual number of deaths X in any given year is a random variable that could be modeled by a Poisson distribution with parameter $\mu = 43.2$.

a) Find $P(X = 41)$, the probability that there will be exactly 41 lightning deaths in a given year.

b) Find $P(41 \leq X \leq 45)$, the probability that between 41 and 45 people will be killed by lightning in a given year. **Hint**: Find the probabilities of 41, 42, 43, 44, and 45 people being killed and then add those probabilities together.

c) Recall that the mean and variance of a Poisson(μ) distribution are both equal to the value of the parameter μ. What is the value of the mean and *standard deviation* of the distribution of the number of people killed X?

4.8 The Furnas Volcano in the Azores Islands (950 miles off the coast of Portugal) erupts about 3.3 times every 1,000 years, on average [11]. However, in any given 1,000 year period, the actual number of eruptions X is a random variable.

One choice of a model for the probability distribution of X is the Poisson distribution with parameter $\mu = 3.3$. Use this distribution to find the following probabilities.

a) Find $P(X = 2)$, the probability that there will be exactly 2 eruptions in the next 1,000 years.

b) Find $P(X = 0)$, the probability that there will be no eruptions in the next 1,000 years.

c) Find $P(X \geq 1)$, the probability that there will be at least one eruption in the next 1,000 years.

4.9 Although large meteorite and asteroid impacts on Earth are rare on a human time scale, they're not unheard of:

> On 30 June 1908, a huge fireball was observed over Europe and Russia, and a large detonation recorded by seismometers. Eyewitnesses in a remote part of Siberia reported feeling a powerful shock wave. Later expeditions found a huge area of forest, over 2,000 km^2 in extent, flattened by the force of the blast, now thought to be a small asteroid or comet fragment (only 20 - 30 m across), exploding in the atmosphere. The object was small, but the energy released was huge, equivalent to 10 megatons of TNT (similar to a large nuclear explosion) [3].

Assessing the risk of such impact events to human populations requires estimates of the rates at which they occur. Obtaining such estimates is difficult, though, because the crater record on Earth has been nearly erased by geologic processes (erosion, volcanism, and plate tectonism). Nonetheless estimates exist, either from crater observations on the Moon and the other terrestrial planets or from data on meteor activity in the Earth's upper atmosphere.

The table below shows estimates of the impact rates for intermediate sized objects ($10^3 - 10^{11}$ kg), as reported in [3].

Size of Object	Mean Number of Impacts Per 100 Years	Equivalent Years Per Impact
10^5 kg	1	100
10^7 kg	0.0185	5,400
10^9 kg	0.0025	40,000

An object whose mass is 2×10^9 kg would have a diameter of about 80 m if composed of iron, and about 100 m if composed of stone. Objects as small as 10^5 kg can form craters, and are capable of widespread devastation if their impact is near a populated area.

The number of impacts by objects of a given size in a 100 year period is a Poisson random variable with parameter μ, where μ is the mean number of impacts of that size per 100 years.

a) Let X be the number of impacts by 10^5 kg objects in the next 100 years. Find $P(X = 0)$, the probability that no 10^5 kg objects will impact Earth in the next 100 years.

b) Find $P(X \geq 1)$, the probability that at least one 10^5 kg object will impact Earth in the next 100 years. **Hint**: The probability that at least one is one minus the probability of none.

c) Now let X be the number of impacts by 10^7 kg objects in the next 100 years. Find $P(X = 1)$, the probability that exactly one 10^7 kg object will impact Earth in the next 100 years.

d) Now let X be the number of impacts by 10^9 kg objects in the next 100 years. Find $P(X \geq 1)$, the probability that at least one 10^9 kg object will impact Earth in the next 100 years. **Hint**: See the hint for *b*.

4.10 In each situation below, decide whether it would more appropriate to model X as a normal or a lognormal random variable.

a) The pH in a stream varies from day to day above and below its mean pH level, which is 7.3. The fraction of days on which the pH is above it's mean is about the same as the fraction on which it's below average. X is the pH on a randomly selected day.

b) Concentrations of copper (Cu) in surface waters vary from one lake to the next. Most lakes have concentrations near zero, but a small fraction of them that are close to industrialized areas have concentrations that are substantially higher. X is the Cu concentration in a randomly selected lake.

4.11 According to the U.S. Environmental Protection Agency, chloroform, which in its gaseous form is suspected to be a cancer-causing agent, is present in small quantities in all of the country's 240,000 public water sources. If the mean and standard deviation of the amounts of chloroform present in water sources are 34 and 53 μg/L, respectively, explain why chloroform amounts do not follow a normal distribution. **Hint**: Recall that a normal distribution extends three standard deviations above and below it's mean.

4.12 A radon detector is exposed to 100 pCi/L of radon, but the actual detector reading is subject to measurement error. Suppose that the reading X follows a N(100, 8) distribution.

a) Find $P(X \leq 80)$, the probability that the reading will be 80 or less.

b) Find $P(X > 115)$, the probability that the reading will be greater than 115.

c) Find $P(80 < X < 115)$, the probability that the reading will be between 80 and 115.

d) Find the 95th percentile of the distribution of X.

e) Find the 10th percentile of the distribution of X.

4.13 Biological measurements on fish are sometimes used as indicators of environmental pollution. A study investigating the use of blood glucose levels in fish as an indicator of the insecticide dieldrin suggests that glucose levels in the small freshwater fish johnny darter (*Etheostoma nigrum* Rafinesque) in White Clay Creek, Chester County, Pennsylvania follow a normal distribution with mean 37.5 mg/100 ml and standard deviation 15.3 mg/100 ml [14].

Let X denote the blood glucose level (mg/100 ml) in a randomly selected johnny darter.

a) Find $P(X > 35)$, the probability that the glucose level will be greater than 35.

b) Find $P(X < 20)$, the probability that the glucose level will be 20 or less.

c) Find $P(20 < X < 35)$, the probability that the glucose level will be between 20 and 35.

d) Find the glucose below which 97.5% of johnny darter glucose levels fall (that is, the 97.5th percentile of the distribution of glucose levels).

e) Find the two glucose levels between which the middle 95% of johnny darter glucose levels fall (that is, the 2.5th and 97.5th percentiles of the distribution of glucose levels).

4.14 An instrument for measuring chromium in floodwater produces readings that follow a N(0, 0.008) distribution when exposed to a blank water specimen, that is, one containing no chromium.

a) Find the value of the limit of detection for which the probability of a false positive is $\alpha = 0.05$.

b) Find the value of the limit of detection for which the probability of a false positive is $\alpha = 0.01$.

4.15 The Cave of Crystals is a deep underground cave, discovered in the year 2000 in Naica, Mexico, whose walls are lined with spectacular giant crystals of the mineral gypsum, some as long as 36 ft. The crystals were deposited by hot mineral-rich water over the course of millions of years.

A study was carried out to estimate the temperature of the water during crystal growth [9]. Small crystal fragments containing fluid inclusions trapped during crystal growth were collected from the cave and heated to the homogenization temperature, the temperature at which gas bubbles within the fluid inclusion disappear leaving only liquid. This homogenization temperature is taken to be the estimated temperature of the fluid during crystal formation.

The cited study suggests that the homogenization temperature X in a randomly selected crystal specimen, in degrees Celsius, is a random variable that follows a normal distribution with mean $\mu = 52.5$ C° and standard deviation $\sigma = 4.6$ C°.

a) The homogenization temperature in degrees Fahrenheit, $\frac{9}{5}X + 32$, is a new random variable. What kind of distribution does it follow? Give the name of the distribution and the values of its mean and standard deviation.

b) Find the probability that a homogenization temperature (in degrees Fahrenheit) will be greater than 140 F°.

4.16 In a U.S. Environmental Protection Agency report, the lognormal distribution is suggested for modeling nickel concentrations (ppb) in groundwater [1]. Suppose that in certain region, they follow a LN($\mu = 3.9$, $\sigma = 1.8$) distribution.

a) Recall that μ and σ determine the mean and standard deviation of the lognormal distribution, but they themselves *aren't* the mean and standard deviation of that distribution. Compute the mean and standard deviation of the distribution of nickel concentrations.

b) Compute the median of the distribution of nickel concentrations.

c) What characteristic of the shape of the lognormal distribution explains why its mean is greater than its median?

d) Let X denote the nickel concentration in a randomly selected groundwater specimen. Compute $P(X > 25.0)$, the probability that the concentration will be greater than 25.0 ppb.

e) Compute $P(10.0 \leq X \leq 25.0)$, the probability that the concentration will be between 10.0 and 25.0 ppb.

f) Compute the 25th and 75th percentiles of the lognormal distribution of nickel concentrations. **Hint**: First find the 25th and 75th of the N(3.9, 1.8) distribution using $\mu + z\sigma$ as in (4.9), then convert these to the 25th and 75th percentiles of the LN(3.9, 1.8) distribution by taking their antilogs $e^{\mu+z\sigma}$.

Bibliography

[1] Statistical analysis of ground-water monitoring data at RCRA facilities. Technical Report EPA/530-R-93-003, United States Environmental Protection Agency, Office of Solid Waste, Washington, D.C., 1992.

[2] L. G. Blackwood. The lognormal distribution, environmental data, and radiological monitoring. *Environmental Monitoring and Assessment*, 21(3):193–210, June 1992.

[3] Phillip A. Bland. The impact rate on Earth. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 363(1837):2793 – 2810, 2005.

[4] R. K. Boeckenhauer et al. Statistical estimation and visualization of ground-water contamination data. Technical Report EPA/600/R-00/034, United States Environmental Protection Agency, August 2000.

[5] L. A. Currie. Detection: Overview of historical societal, and technical issues. In L. A. Currie, editor, *Detection in Analytical Chemistry*, volume 361 of *ACS Symposium Series*, chapter 1, pages 1–62. Oxford University Press, Washington, DC: American Chemical Society, 1988.

[6] N. B. Davies and M. De L. Brooke. An experimental study of co-evolution between the cuckoo, *Cuculus canorus*, and its hosts. I. Host egg discrimination. *The Journal of Animal Ecology*, 58(1):207–224, February 1989.

[7] James B. Elsner and Brian H. Bossak. Bayesian analysis of U.S. hurricane climate. *Journal of Climate*, 14(23), December 2001.

[8] Albert Gan, Kaiyu Liu, and Rax Jung. Vehicle occupancy data collection methods (phase II). Technical report, Lehman Center for Transportation Research, August 2007.

[9] Juan Manuel García-Ruiz, Roberto Villasuso, Carles Ayora, Angels Canals, and Fermín Otálora. The formation of natural gypsum megacrystals in Naica (Mexico). Supplementary Information deposited into the GSA Data Repository. Data Repository Item.

[10] Abua Ikem and Nosa O. Egiebor. Assessment of trace elements in canned fishes (mackerel, tuna, salmon, sardines and herrings) marketed in Georgia and Alabama (United States of America). *Journal of Food Composition and Analysis*, 18:771 – 787, 2005.

[11] G. Jones, D. K. Chester, and F. Shooshtarian. Statistical analysis of the frequency of eruptions at Furnas Volcano, Sao Miguel, Azores. *Journal of Volcanology and Geothermal Research*, 92:31–38, 1999.

[12] Monte Lloyd. Mean crowding. *Journal of Animal Ecology*, 36(1):1–30, 1967.

[13] W.R. Ott. A physical explanation of the lognormality of pollutant concentrations. *J Air Waste Manage Assoc*, 40(10):1378–1383, October 1990.

[14] Ellen K. Silbergeld. Blood glucose: A sensitive indicator of environmental stress in fish. *Bulletin of Environmental Contamination and Toxicology*, 11(1):20 – 25, 1974.

[15] A. J. Stewart. Ambient bioassays for assessing water-quality conditions in receiving streams. *Ecotoxicology*, 5:377 – 393, 1996.

[16] C. C. Travis and M. L. Land. Estimating the mean of data sets with nondetectable values. *Environmental Science and Technology*, 24(7):961, 1990.