Chapter 6

One-Sample Confidence Intervals

Chapter Objectives

- Distinguish the situation for which a t distribution should be used instead of a standard normal distribution.
- Compute and interpret one-sample z and one-sample t confidence intervals for a mean.
- Determine the sample size needed to keep the margin of error no bigger than a desired value.
- Find and interpret the limit of detection when an instrument's standard deviation isn't known.
- Assess normality of data using graphs, and transform non-normal data to normality.
- Compute and interpret a tolerance limit and a confidence interval for a percentile.
- Compute and interpret a nonparametric confidence interval for a median.
- Compute and interpret a one-sample z confidence interval for a proportion.

Key Takeaways

- The t distribution is the distribution of the sample mean standardized using the sample standard deviation instead of the population standard deviation.
- Confidence intervals are used to estimate unknown population means (or other population parameters) with a chosen level of confidence.
- A confidence interval provides a set of plausible estimates for an unknown population mean (or other parameter).
- The margin of error of a confidence interval indicates the precision of the estimate of the population mean (or other parameter).
- Confidence intervals can be one-sided. A one-sided confidence interval gives an upper (or lower) bound for an unknown population mean (or other parameter).
- One-sample z and t confidence intervals for a mean require either that the sample is from a normal population or the sample size is large. We can assess normality by graphing the data. A log transformation can make right skewed data more normal.
- Tolerance limits are used as threshold values for identifying unusually high contaminant levels relative to the distribution of background levels.
- A nonparametric confidence interval for a population median doesn't require an assumption of normality of the data.

6.1 Introduction

In most practical problems, the values of population parameters such as μ and σ won't be known and instead are *estimated* from sample data.

Estimates comes in two forms:

1. Point estimates (single values)
2. Interval estimates (ranges of values)

A ***point estimate*** is a single-valued statistic used to estimate a population parameter. Some examples of point estimates are:

	Sample Estimate	**Population Parameter**
Mean	$\bar{X}$	μ
Std Dev	S	σ
Median	$\tilde{X}$	$\tilde{\mu}$
Proportion	$\hat{P}$	p

An ***interval estimate***, or ***confidence interval***, is an entire range of values, all of which are considered to be reasonable estimates of (that is, plausible values for), an (unknown) population parameter in light of the data.

The interval's endpoints are computed from sample data in a way that allows us to state *how confident* we are that the parameter value is in the interval. In fact, the first step in constructing the interval will be to *choose* a so-called ***level of confidence***, expressed as a percent, that represents how confident we want to be that the interval will contain the true parameter value.

Example 6.1: Point Estimate and Confidence Interval for μ

Rocky Flats was a nuclear weapons production plant located 16 miles northwest of Denver, Colorado that was in operation from 1952 until 1989 [5], [6]. Its hazardous waste spills, plutonium fires, and leaking barrels of radioactive waste contaminated soil in the area. The plant's operators later pleaded guilty to criminal violations of environmental law.

Cleanup of the site took 10 years and cost $7 billion. It required establishing so-called soil action levels, threshold values above which concentrations of radioactive plutonium in soil would trigger an ”action” such as removal, containment, or stabilization. Public concern about the action levels, which were established by U.S. and Colorado government agencies, prompted an independent assessment by a private contractor, which completed its work in 2000 [15]. The table below shows background soil radiation concentrations (the plutonium isotope ^{239,240}Pu, in Bq/kg) obtained by the contractor from $n = 10$ uncontaminated sites along the Front Range of the Colorado Rocky Mountains.

Background Soil
Radiation Concentrations

Site	^{239,240}Pu
Z01	1.20
Z02	2.10
Z03	1.46
Z04	2.10
Z05	2.10
Z06	1.14
Z07	3.29
Z08	3.22
Z09	2.07
Z10	2.70

The *point estimate* of μ, the true (unknown) mean background concentration along the Front Range, is the sample mean,

$$\bar{X} \;=\; 2.14.$$

A *confidence interval* for μ, computed using a 95% level of confidence, has endpoints

$$(1.60,\ 2.68).$$

This interval serves as an entire range of estimates for μ, and we can be 95% confident that μ is contained in this interval somewhere. We'll see how the interval was computed in Example 6.3.

Confidence intervals for a population mean are computed differently depending on whether value of the population standard deviation σ is *known* or *unknown*, as we'll see in Sections 6.2 and 6.5. Although the focus of this chapter is on confidence intervals for a population mean, we'll also look at confidence intervals for a population median and for other population percentiles, *tolerance limits* (threshold values used for identifying unusually high contaminant levels), and a confidence interval for a population proportion.

6.2 One-Sample *Z* Confidence Interval for a Population Mean

In this section, we derive the formula for a confidence interval for a population mean μ under the scenario that the population standard deviation σ is known. Although this situation is rare, it will provide intuition for the more common situation in which σ is unknown, which is covered in Section 6.5.

Suppose then that we have a random sample from the population, and that either the population is *normal* or the sample size n is *large*. Then from Chapter 5,

$$\bar{X} \sim \mathrm{N}\left(\mu,\, \sigma_{\bar{X}}\right) \qquad \text{where} \qquad \sigma_{\bar{X}} \;=\; \frac{\sigma}{\sqrt{n}}.$$

In this case

$$Z \;=\; \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \sim \mathrm{N}(0,\, 1). \tag{6.1}$$

The 2.5th and 97.5th percentiles of the standard normal distribution are -1.96 and 1.96, so there's a 95% chance that Z will fall between these two values, that is,

$$0.95 \;=\; P\left(-1.96 \;<\; \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \;<\; 1.96\right).$$

Rearranging terms isolates μ between two numbers:

$$\begin{array}{rcll} 0.95 & = & P\left(-1.96\,\sigma_{\bar{X}} \;<\; \bar{X}-\mu \;<\; 1.96\,\sigma_{\bar{X}}\right) & \text{(Multiplied each term by } \sigma_{\bar{X}}\text{)} \\ & = & P\left(-\bar{X}-1.96\,\sigma_{\bar{X}} \;<\; -\mu \;<\; -\bar{X}+1.96\,\sigma_{\bar{X}}\right) & \text{(Subtracted } \bar{X} \text{ from each term)} \\ & = & P\left(\bar{X}-1.96\,\sigma_{\bar{X}} \;<\; \mu \;<\; \bar{X}+1.96\,\sigma_{\bar{X}}\right). & \text{(Multiplied each term by -1,} \\ & & & \text{which flips inequality directions)} \end{array} \tag{6.2}$$

The first line above says that 95% of the time, the *sampling error* of $\bar{X}$ will be no larger than 1.96 *standard errors*. The last line says we can be 95% confident that μ will be contained in the interval

$$\bar{X} \;\pm\; 1.96\,\sigma_{\bar{X}}, \tag{6.3}$$

which is known as a ***95% one-sample z confidence interval for μ***.

Example 6.2: One-Sample z Confidence Interval

In a study of truck emissions and air quality in California [2], the engine idle time (minutes) per day was recorded for $n = 13$ trucks, giving a sample mean

$$\bar{X} \;=\; 29.6.$$

Suppose that in the *population* of trucks, the standard deviation of their idle times is $\sigma = 10.0$ minutes. Then the standard error of $\bar{X}$ is

$$\sigma_{\bar{X}} \;=\; \frac{\sigma}{\sqrt{n}} \;=\; \frac{10.0}{\sqrt{13}} \;=\; 2.8.$$

The 95% z confidence interval for the true (unknown) *population* mean idle time μ is

$$\begin{aligned} \bar{X} \;\pm\; 1.96\,\sigma_{\bar{X}} \;&=\; 29.6 \;\pm\; 1.96(2.8) \\ &=\; 29.6 \;\pm\; 5.5 \\ &=\; (24.1,\; 35.1). \end{aligned}$$

We can be 95% confident that μ is in this range (somewhere).

The most commonly used level of confidence is 95%, but sometimes other levels are used. For these other confidence levels, a different value than 1.96 is used in (6.3). To make our confidence interval notation generic, we'll write it in terms of the value $\boldsymbol{\alpha}$ for which

$$\textbf{Level of Confidence} \;=\; \mathbf{100(1-\boldsymbol{\alpha})\%}.$$

For example, $\alpha = 0.05$ for a 95% level of confidence. This leads to the following general confidence interval procedure.

One-Sample Z Confidence Interval: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population whose mean is μ and whose standard deviation is σ, where σ is *known.* A **$100(1-\alpha)\%$** ***one-sample z confidence interval for μ*** is

$$\bar{X} \;\pm\; z_{\alpha/2}\,\sigma_{\bar{X}} \qquad \text{where} \qquad \sigma_{\bar{X}} \;=\; \frac{\sigma}{\sqrt{n}}. \tag{6.4}$$

and the $z_{\alpha/2}$ value is discussed below.

The confidence interval is valid if either the population is *normal* or the sample size n is *large.*

The value $\boldsymbol{z_{\alpha/2}}$ is called the ***z critical value*** associated with the $100(1-\alpha)\%$ level of confidence. It's the $100(1-\alpha/2)$th percentile of the standard normal distribution. For example, for a 95% level of confidence, $\alpha = 0.05$ and $z_{0.025} = 1.96$ is the 97.5th percentile. As depicted in Fig. 6.1, the middle $100(1-\alpha)\%$ of the distribution lies between $-z_{\alpha/2}$ and $z_{\alpha/2}$.

For the three most commonly used confidence levels, 90%, 95%, and 99%, the z critical values are

$$\begin{aligned} z_{0.05} &= 1.64 && \text{(for a 90\% confidence level)} \\ z_{0.025} &= 1.96 && \text{(for a 95\% confidence level)} \\ z_{0.005} &= 2.58 && \text{(for a 99\% confidence level)} \end{aligned}$$

The choice of a level of confidence affects the confidence interval width via the z critical value. More precisely, *a higher level of confidence* will result in a *wider interval.* The intuition is that if we want to be *more confident* that our interval contains μ, we'll need a *wider interval.*

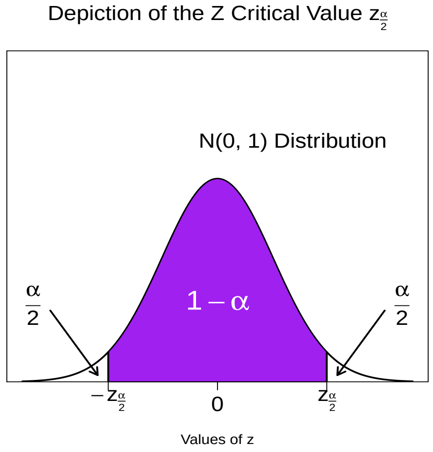


Figure 6.1: Graphical depiction of the meaning of the z critical value $z_{\alpha/2}$.

The "plus or minus" part of a confidence interval is called the ***margin of error***.

Margin of Error: For the one-sample z confidence interval for μ, the margin of error is

$$\text{Margin of Error} \;=\; z_{\alpha/2}\,\sigma_{\bar{X}} \;=\; z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}. \tag{6.5}$$

The margin of error measures the degree of precision in the estimate $\bar{X}$ of μ, and is interpreted as the largest amount by which our estimate might be off its mark. A smaller margin of error means a more precise estimate. In Example 6.2, the margin of error in the estimate of the truck population mean idle time was 5.5 minutes.

The margin of error will be small if either:

- The population standard deviation σ is small, or
- The sample size n is large.

Although the value of σ is beyond our control, we *can* choose the sample size n. Using a *larger sample size* results in a *more precise estimate* of μ.

6.3 Properties and Interpretation of Confidence Intervals

Because the endpoints of a confidence interval depend on the sample data (via the sample mean $\bar{X}$), they'll vary from one random sample to the next. This has implications for interpreting a confidence interval. To illustrate, the figure below shows twenty 90% confidence intervals (based on computer-generated random samples), with each $\bar{X}$ value shown as a green triangle in the center its interval. At the top is the sampling distribution of $\bar{X}$, which is centered on the population mean μ. In the long-run, 90% of such intervals will cover μ.

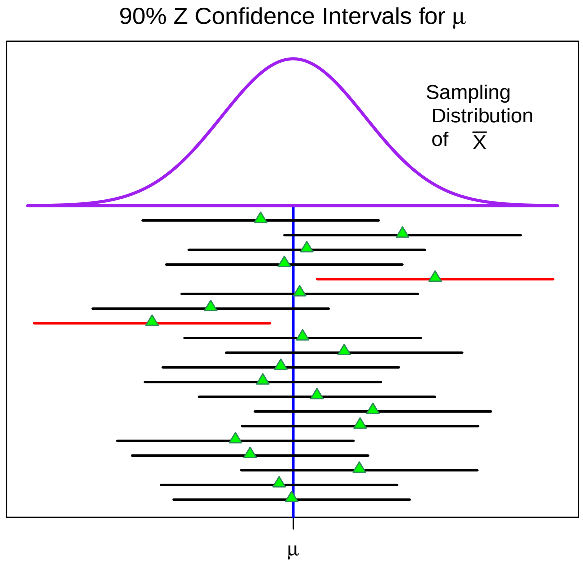


Figure 6.2: 90% one-sample z confidence intervals based on random samples from a population whose mean is μ (vertical blue line). At the top is the sampling distribution of $\bar{X}$. For each sample, the green triangle represents the observed value of $\bar{X}$.

The intervals shown in the figure are of the form

$$\bar{X} \ \pm \ 1.64\,\sigma_{\bar{X}},$$

so the ones that *don't* contain μ correspond to samples whose $\bar{X}$ value lies more than 1.64 standard errors away from μ.

Properties and Interpretation of Confidence Intervals: We can summarize the important properties and correct interpretation of confidence intervals as follows.

1. A confidence interval for an unknown population parameter (e.g. its mean μ) gives a set of plausible values for the parameter.
2. The level of confidence is the long-run percentage of intervals that would contain the parameter value if random samples of the same size were repeatedly drawn from the population. It therefore measures how confident we can be that any *particular* interval contains the parameter value.
3. A higher level of confidence will result in a wider confidence interval.
4. A larger sample size will result in a smaller margin of error and therefore a narrower confidence interval.

6.4 The *t* Distribution

In practice, we usually don't know the value of the population standard deviation σ, so we can't use the interval defined by (6.4). Instead, we replace σ by its estimate S, which gives the ***estimated standard error*** of $\bar{X}$, denoted $\boldsymbol{S_{\bar{X}}}$ and defined as

$$\boldsymbol{S_{\bar{X}}} \ = \ \frac{\boldsymbol{S}}{\sqrt{\boldsymbol{n}}}.$$

But this uncertainty in the value of σ carries over to uncertainty about how far we think $\bar{X}$ might fall away from μ. We incorporate this extra uncertainty into the confidence interval's width by using a critical value from a new distribution, the *t distribution*, in place of the z critical value. The t distribution arises when we standardize $\bar{X}$ using the estimate S in place of the true value σ.

Fact 6.1 Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population whose mean is μ. If the population is *normal*, the random variable

$$T = \frac{\bar{X} - \mu}{S_{\bar{X}}}, \qquad \text{where} \qquad S_{\bar{X}} = \frac{S}{\sqrt{n}}, \tag{6.6}$$

follows a distribution known as the ***t* *distribution*** with $\boldsymbol{n-1}$ ***degrees of freedom***.

Furthermore, even if the population is *non-normal*, the variable T still follows the t distribution *approximately* as long the sample size n is *large*.

We write

$$\boldsymbol{\frac{\bar{X} - \mu}{S_{\bar{X}}} \sim t(n-1)}$$

to mean that the random variable $(\bar{X} - \mu)/S_{\bar{X}}$ follows a t distribution with $n - 1$ degrees of freedom.

Because the denominator $S_{\bar{X}}$ is an estimate of the true standard error $\sigma_{\bar{X}}$, the value of T indicates approximately how many standard errors $\bar{X}$ is away from μ. The *degrees of freedom* is a parameter of the t distribution that controls its spread. Some t density curves with various degrees of freedom are shown in Fig. 6.3 along with the standard normal curve (labeled as a t curve with ∞ degrees of freedom for reasons that will be given later).

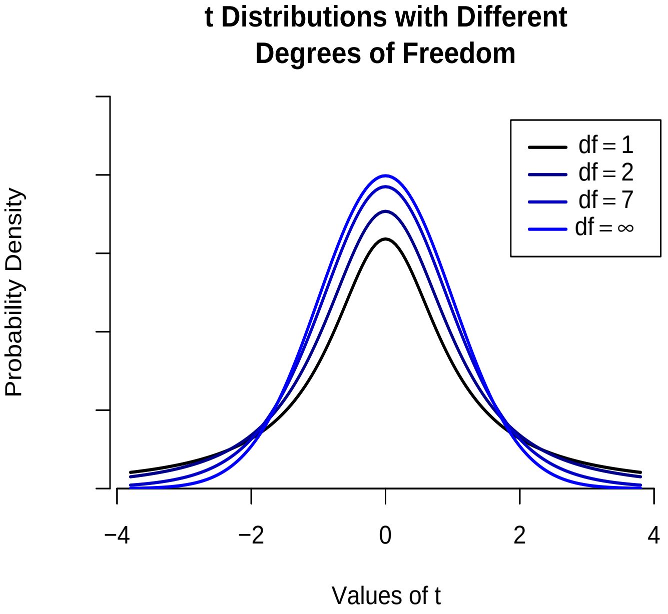


Figure 6.3: t distributions with different degrees of freedom. The t distribution with ∞ degrees of freedom is the standard normal curve.

Note: Usually $n \geq 30$ is large enough for the variable T to follow a t distribution (approximately) when the sample is from a non-normal population. However, if the population is *very* skewed, an even larger sample size may be needed. On the other hand, if the population distribution is fairly symmetric, and therefore closer to a normal distribution, a sample size of $n = 10$ or 15 may suffice.

Properties of the *t* distribution: t distributions have the following properties.

1. The t density curve is centered on zero, and resembles the standard normal curve.
2. The tails of the t curve are more spread out than those of the standard normal curve, reflecting the sample-to-sample variability in the value of S in (6.6) not present in the value of σ in (6.1).
3. As the degrees of freedom parameter increases, the t curve resembles the standard normal curve more and more. If the degrees of freedom are greater than about 40, the t curve is practically indistinguishable from the standard normal curve. A t distribution with ∞ degrees of freedom is identical to the standard normal distribution.

Comment: The reason why the single parameter of the t distribution is called its "degrees of freedom" and is $n - 1$ has to do with the fact that because the deviations $X_i - \bar{X}$ used to calculate S (for (6.6)) always sum to zero (Chapter 3), only $n - 1$ of them are "free to vary" – the remaining one is determined by the values of the other $n - 1$.

6.5 One-Sample *t* Confidence Interval for a Population Mean

When we *don't* know the population standard deviation σ, we can't use the z confidence interval of Section 6.2. Instead, we replace σ in the confidence interval formula (6.4) by its estimate S. But then, as we'll see, we also need to replace the *z critical value* by a *t critical value*.

One-Sample *t* Confidence Interval: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population whose mean is μ. A **$100(1-\alpha)\%$ *one-sample t confidence interval for μ*** is

$$\bar{X} \pm t_{\alpha/2,n-1} S_{\bar{X}}, \qquad \text{where} \qquad S_{\bar{X}} = \frac{S}{\sqrt{n}} \tag{6.7}$$

and the *t critical value* $t_{\alpha/2,n-1}$ is discussed below.

The confidence interval is valid if either the population is *normal* or the sample size n is *large*.

Like all confidence intervals, this one has the properties and interpretations given in Section 6.3.

This interval's formula is derived (at the end of this section) like the z interval's was, but using the t distribution instead of the standard normal. The value $\boldsymbol{t_{\alpha/2,n-1}}$ is the ***t critical value*** associated with the $100(1-\alpha)\%$ level of confidence. It's the $100(1-\alpha/2)$th percentile of the $t(n-1)$ distribution. Thus for a 95% level of confidence, $\alpha = 0.05$ and $t_{0.025,n-1}$ is the 97.5th percentile, but its value will depend on the degrees of freedom $(n-1)$. As depicted in Fig. 6.4, the middle $100(1-\alpha)\%$ of the t distribution lies between $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$. The t critical values can be obtained from a t distribution table or using statistical software.

The *margin of error* is the "plus or minus" part of the confidence interval.

Margin of Error: For the one-sample t confidence interval for μ, the margin of error is

$$\text{Margin of Error} = t_{\alpha/2,n-1} S_{\bar{X}} = t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}. \tag{6.8}$$

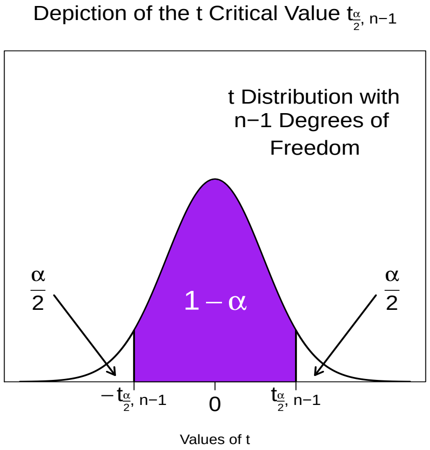

Depiction of the t Critical Value $t_{\frac{\alpha}{2}, n-1}$
t Distribution with n−1 Degrees of Freedom
$\frac{\alpha}{2}$
$1-\alpha$
$\frac{\alpha}{2}$
$-t_{\frac{\alpha}{2}, n-1}$
0
$t_{\frac{\alpha}{2}, n-1}$
Values of t

Figure 6.4: Graphical depiction of the meaning of the t critical value $t_{\alpha/2,n-1}$.

As for the z interval, a smaller margin of error in the t interval indicates that $\bar{X}$ is a more precise estimate of μ. It will be small if either the population standard deviation σ is small (in which case S should be small too) or n is large.

Example 6.3: One-Sample *t* Confidence Interval

Below is a histogram of the background soil radiation data given in Example 6.1.

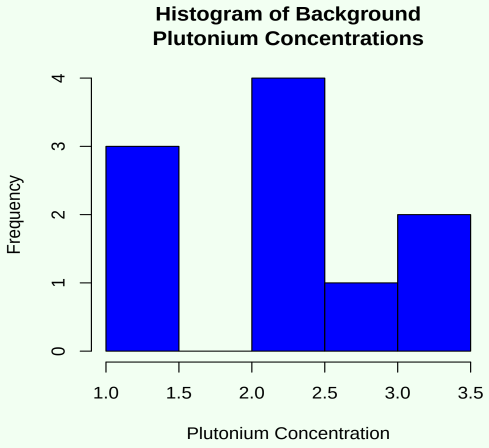


Figure 6.5: Histogram of soil radiation levels along the Front Range of the Rocky Mountains in Colorado.

Based on the shape of the histogram, there aren't strong indications that the sample came from a non-normal population, so it's appropriate use the data to construct a one-sample t confidence interval for estimating the population mean background radiation level μ.

The sample mean and standard deviation are

$$\bar{X} = 2.14 \qquad \text{and} \qquad S = 0.76,$$

so the (estimated) standard error is

$$S_{\bar{X}} \;=\; \frac{S}{\sqrt{n}} \;=\; \frac{0.76}{\sqrt{10}} \;=\; 0.24.$$

A 95% t confidence interval for μ is

$$\begin{aligned} \bar{X} \;\pm\; t_{\alpha/2,n-1}\, S_{\bar{X}} \;&=\; 2.14 \;\pm\; 2.262\,(0.24) \\ &=\; 2.14 \;\pm\; 0.54 \\ &=\; (1.60,\; 2.68), \end{aligned}$$

where the t critical value, $t_{\alpha/2,n-1} = t_{0.025,9} = 2.262$, was obtained from a t distribution table.

We can be 95% confident that μ is in this range somewhere. It's plausible, for example, that μ is as high as 2.50 Bq/kg, but not as high as 3.00 Bq/kg.

The margin of error in the estimate 2.14 of μ is 0.54 Bq/kg, and we're 95% confident that the estimate isn't off its mark by more than this amount.

The next example reminds us that the level of confidence used to construct a confidence interval affects the interval's width.

Example 6.4: One-Sample t Confidence Interval

Continuing from the last example, if instead we used a 99% level of confidence, the t critical value would be $t_{0.005,9} = 3.250$, and the confidence interval would become

$$\begin{aligned} \bar{X} \;\pm\; t_{\alpha/2,n-1}\, S_{\bar{X}} \;&=\; 2.14 \;\pm\; 3.250\,(0.24) \\ &=\; 2.14 \;\pm\; 0.78 \\ &=\; (1.36,\; 2.92). \end{aligned}$$

The higher level of confidence resulted in a larger margin of error (0.78) and a wider interval. Thus the price we pay to gain confidence that our interval contains μ a is widening of the interval.

Note: Usually $n \geq 30$ is large enough to justify the use of the one-sample t confidence interval procedure when the sample is from a non-normal population. However, if the population is *very* skewed, an even larger sample size may be needed. On the other hand, if the population distribution is fairly symmetric, and therefore closer to a normal distribution, a sample size of $n = 10$ or 15 may suffice.

To see how the formula (6.7) for the one-sample t confidence interval was derived, from Fact 6.1, we have

$$1 - \alpha \;=\; P\left(-t_{\alpha/2,n-1} \;<\; \frac{\bar{X} - \mu}{S_{\bar{X}}} \;<\; t_{\alpha/2,n-1}\right).$$

After rearranging terms (as was done for the z interval in (6.2)), we get

$$1 - \alpha \;=\; P\left(\bar{X} - t_{\alpha/2,n-1}\, S_{\bar{X}} \;<\; \mu \;<\; \bar{X} + t_{\alpha/2,n-1}\, S_{\bar{X}}\right),$$

which leads to the t interval formula (6.7).

6.6 One-Sided t Confidence Intervals for a Population Mean

Occasionally we'll want to be able to declare, with, say, 95% confidence, that a population mean is *less than* some value (rather than *between* two values). For example, because low levels of radiation in the environment are desirable, we may want to affirm, with 95% confidence, that the true mean level lies *below* some value. Other times, we may want a value *above* which μ lies with 95% confidence.

These are called *one-sided t confidence intervals for μ.*

One-Sided t Confidence Intervals for μ: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population whose mean is μ.

A **$100(1-\alpha)\%$ *one-sided t confidence interval for μ*** with an ***upper bound*** is

$$\left(-\infty,\quad \bar{X} + t_{\alpha,n-1}\, S_{\bar{X}}\right).$$

A **$100(1-\alpha)\%$ *one-sided t confidence interval for μ*** with a ***lower bound*** is

$$\left(\bar{X} - t_{\alpha,n-1}\, S_{\bar{X}},\quad \infty\right).$$

The confidence intervals are valid if either the population is *normal* or the sample size n is *large.*

For a given study, only one of these intervals would be used, and we could be $100(1-\alpha)\%$ confident that it would contain the (unknown) population mean μ.

Example 6.5: One-Sided t Confidence Interval

Consider again the $n = 10$ background soil radiation levels from along the Front Range (Example 6.1). The sample mean and standard deviation (from Example 6.3) are

$$\bar{X} \;=\; 2.14 \qquad \text{and} \qquad S \;=\; 0.76,$$

and the (estimated) standard error is

$$S_{\bar{X}} \;=\; \frac{S}{\sqrt{n}} \;=\; \frac{0.76}{\sqrt{10}} \;=\; 0.24.$$

The upper bound for a 95% one-sided t confidence interval for the true (unknown) mean concentration μ is

$$\begin{aligned}\bar{X} \;+\; t_{\alpha,n-1}\, S_{\bar{X}} \;&=\; 2.14 \;+\; 1.833\,(0.24)\\ &=\; 2.581,\end{aligned}$$

so the confidence interval is

$$(-\infty,\;\; 2.581).$$

The critical value, $t_{\alpha,n-1} = t_{0.05,9} = 1.833$, was obtained from a t distribution table.

We can be 95% confident that μ is in the range from $-\infty$ to 2.581 Bq/kg.

To see how the one-sided interval formulas were derived, consider first the upper bound. Letting $t_{\alpha,n-1}$ be the $100(1-\alpha)$th percentile of the $t(n-1)$ distribution, we have

$$1-\alpha \;=\; P\left(\frac{\bar{X}-\mu}{S_{\bar{X}}} \;>\; -t_{\alpha,n-1}\right),$$

which, after rearranging terms, is equivalent to

$$1-\alpha \;=\; P\left(\mu \,<\, \bar{X} \;+\; t_{\alpha,n-1}\, S_{\bar{X}}\right).$$

It follows that we can be $100(1-\alpha)\%$ confident that μ will be *less than* the upper bound

$$\bar{X} \;+\; t_{\alpha,n-1}\, S_{\bar{X}}.$$

Note that the t critical value here uses α, *not* the $\alpha/2$ used for the upper endpoint of a two-sided interval (Section 6.5).

A similar line of reasoning is used to derive the formula for the lower bound.

6.7 Sample Size Determination

Planning a study will often involve deciding how large the sample size should be. We know from Section 6.3 that larger samples tend to produce more precise estimates of population parameters such as μ, as measured by the margin of error of the estimate. But larger sample sizes usually require more time, and perhaps money, to collect. In this section, we'll see how to determine the smallest sample size that's still large enough to ensure an acceptably small margin of error.

Suppose we want the margin of error in an estimate of μ to be no bigger than some value, call it $\boldsymbol{B}$. In other words, we want n to be large enough so that

$$\text{Margin of Error} \;\leq\; B.$$

If the population standard deviation σ is known, then from (6.5), for a given level of confidence $100(1-\alpha)\%$, we require n to be large enough so that

$$z_{\alpha/2}\,\sigma_{\bar{X}} \;\leq\; B,$$

which is to say, large enough so that

$$z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \;\leq\; B.$$

Solving for n gives the required sample size.

Sample Size Determination: If we want the margin of error in a $100(1-\alpha)\%$ confidence interval for μ to be no bigger than B, the sample size should satisfy

$$n \;\geq\; \frac{(z_{\alpha/2}\,\sigma)^2}{B^2}, \tag{6.9}$$

which should be *rounded up* to the nearest integer.

Using this sample size, we can be sure, with a certain level of confidence, that our estimate $\bar{X}$ will be within B of the true value μ. The right side of (6.9) depends on the population standard deviation σ, which usually won't be known. In practice, we use in its place a reasonable guess its value, for example based on a ***pilot study*** (small-scale preliminary study) or on the results of preexisting studies. Even when we replace σ by a guess, we still use the z critical value, not the t critical value, because using the t critical value would require knowing the degrees of freedom, $n-1$, which depends on the still-undetermined n.

Example 6.6: Sample Size Determination

A wildlife management study is being planned for the purpose of estimating the true (unknown) mean weight μ of salmon caught by an Alaskan fishing company. It's desired that the estimate be within 0.2 lb of the true value. More specifically, the researchers want the margin of error in a 99% confidence interval for μ to be no bigger than 0.2 lb.

In a pilot study, a random sample of 10 freshly caught salmon is weighed, giving a sample standard deviation of 1.15 lb. Using this as our guess for σ in (6.9), and the z critical value $z_{0.005} = 2.58$, the sample size should satisfy

$$\begin{aligned} n &\geq \frac{(2.58 \cdot 1.15)^2}{0.2^2} \\ &= 220.08 \end{aligned}$$

salmon. Thus, rounding up, at least $n = 221$ fish should be weighed.

6.8 The Limit of Detection when σ is Unknown

In Section 4.6 of Chapter 4, the limit of detection (LOD) was defined so as to make the probability of a false positive (detection on a blank specimen) small. More specifically, under the N(0, σ) measurement error model, the LOD was defined, for a desired false positive probability α (such as $\alpha = 0.01$ or $\alpha = 0.05$), to be

$$\text{LOD} = z_\alpha \, \sigma.$$

But in practice, the true measurement error standard deviation σ is often unknown. When σ *isn't* known, it's common practice to *estimate* it by taking n repeated measurements on a blank specimen and using the sample standard deviation of these, S. Then the LOD is the following.

Limit of Detection When σ is Unknown: If we make n repeated measurements $X_1, X_2, \ldots, X_n$ on a blank specimen, then for a desired false positive probability α, the limit of detection is

$$\text{LOD} = t_{\alpha,n-1} \, S, \tag{6.10}$$

where S is the sample standard deviation of the n repeated measurements.

Typically, $n = 7$ measurements are used to obtain the estimate S of σ. In this case, for $\alpha = 0.01$, $t_{0.05,6} = 3.14$ (from a t distribution table), and for $\alpha = 0.05$, $t_{0.05,6} = 1.94$.

To see why (6.10) gives the desired false positive probability, we'll need the following fact.

Fact 6.2 Suppose $X \sim \text{N}(\mu, \sigma)$, and S is the sample standard deviation of a sample $X_1, X_2, \ldots, X_n$ from that same normal distribution. If X is independent of $X_1, X_2, \ldots, X_n$, then

$$\frac{X - \mu}{S} \sim t(n-1).$$

Thus if X is an individual measurement made on a blank specimen, for which the true concentration of the measured substance is $\mu = 0$, then by the above fact

$$\begin{aligned} \alpha &= P\left(\frac{X - 0}{S} > t_{\alpha,n-1}\right) \\ &= P\left(X > t_{\alpha,n-1} S\right). \end{aligned}$$

In other words, when the LOD is as in (6.10), the probability of a false positive is α, as desired.

6.9 Checking Normality of Data

6.9.1 Introduction

Because the one-sample t confidence interval procedure and so many other statistical procedures rest on an assumption that the data are a random sample from a normal distribution, it is useful to have a few ways to decide whether this assumption is met. Two commonly used tools are:

1. Histograms
2. Normal probability plots

6.9.2 Checking Normality Using a Histogram

The most obvious tool for checking normality is a histogram. If the histogram looks reasonably symmetric and bell-shaped, then the normality assumption is tenable. If it looks skewed or shows other obvious departures from the normal bell-shape, then the normality assumption isn't justified.

But assessing normality from a histogram can be difficult, especially when the sample size is small, because the data are grouped into class intervals (or bins), which dilutes some of the information the data contain. Instead, therefore, it's often preferable to assess normality using a *normal probability plot*, which retains all the information the data have to offer.

6.9.3 Checking Normality Using a Normal Probability Plot

A normal probability plot can be thought of as a plot of the observed sample values (y-axis) versus the values we'd expect to get if the sample came from a normal distribution (x-axis). If the points in the plot lie near a straight line, it tells us that the observed data values are close to what we'd expect them to be if in fact the sample came from a normal population. The plots will be defined in more detail later. For now, we'll just look at a few to see how they're used.

Using Normal Probability Plots to Assess Normality: If a set of data is a random sample from *some* normal distribution, then the points in a normal probability plot should fall close to a straight line. Departures from a straight line pattern are indications that the sample came from a non-normal distribution.

To illustrate, Figs. 6.6 - 6.10 show histograms and normal probability plots for five different data sets, one of which is normal, one right skewed, one left skewed, one "heavy tailed", and the last one "light tailed".

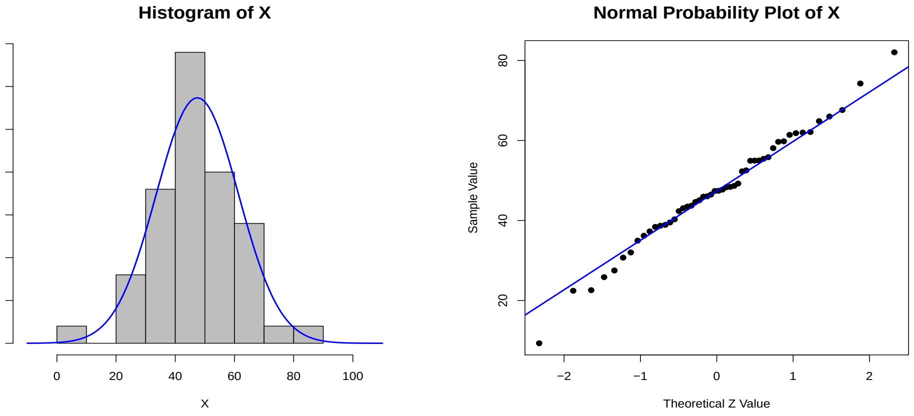


Figure 6.6: Histogram of symmetric, approximately normal data (left). Normal probability plot of the same data (right).

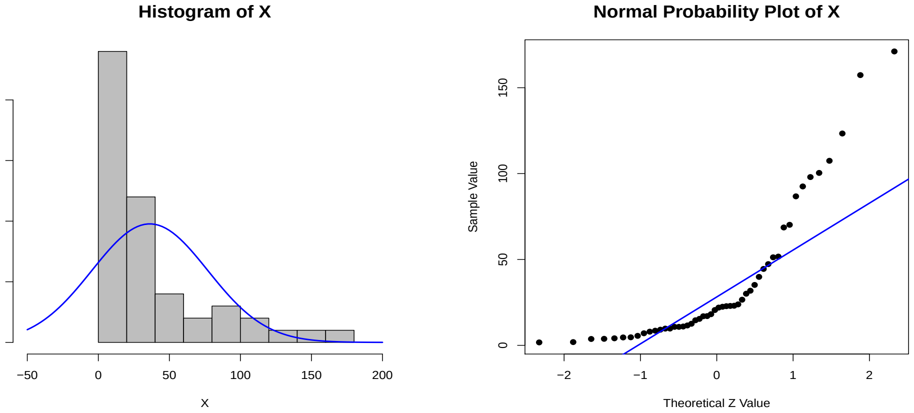


Figure 6.7: Histogram of non-normal, right skewed data (left). Normal probability plot of the same data (right).

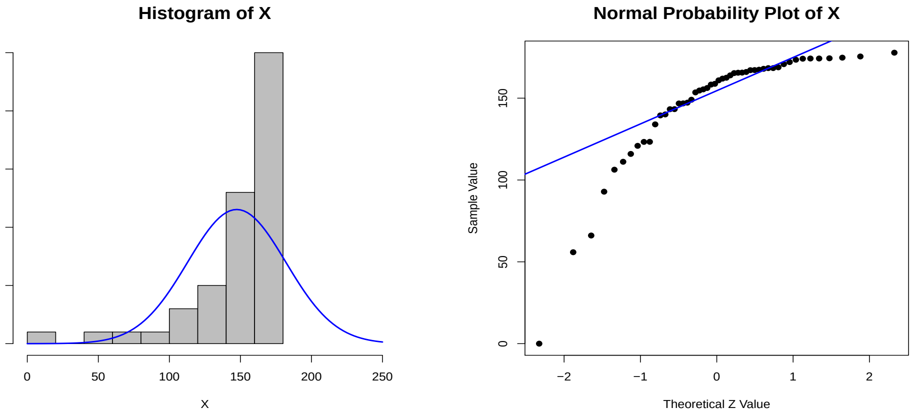


Figure 6.8: Histogram of non-normal, left skewed data (left). Normal probability plot of the same data (right).

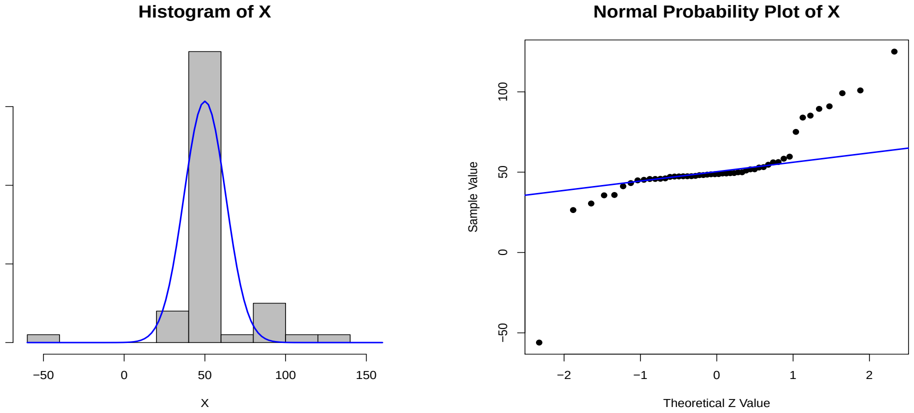


Figure 6.9: Histogram of non-normal, "heavy tailed" data (left). Normal probability plot of the same data (right).

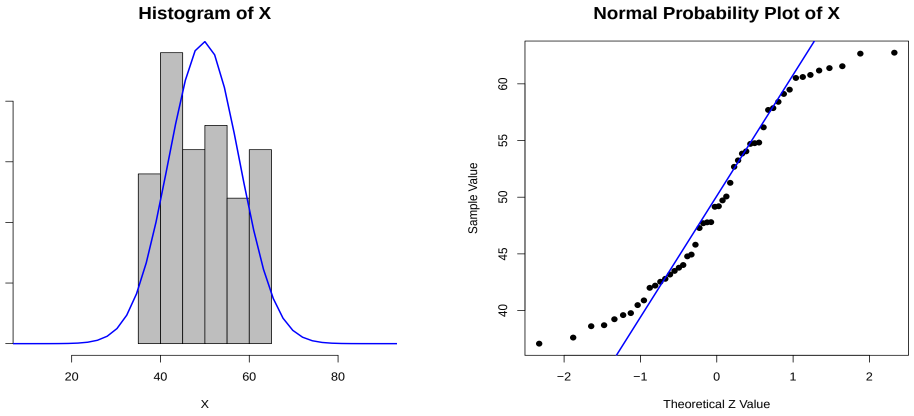


Figure 6.10: Histogram of non-normal, "light tailed" data (left). Normal probability plot of the same data (right).

Notice that for the normally distributed data in Fig. 6.6, the points in the normal probability plot hug the line. For the right and left skewed data in Figs. 6.7 and 6.8, the points follow curved patterns that face in opposite directions. And for the "heavy" and "light" tailed data sets of Figs. 6.9 and 6.10, the points follow backward and forward "S" shapes. In general, patterns like the ones above in normal probability plots can be used to deduce distribution shapes in the manner just described.

6.9.4 Details About the Construction of Normal Probability Plots

Recall that the 100pth percentile of a probability distribution is the value below which 100p% of the distribution lies, where p is a number between zero and one. We can define a ***sample percentile***, in a similar manner, as a value below which below which 100p% of the *sample* observations lie. For example, the sample median $\tilde{X}$ is the 50th sample percentile and the 3rd quartile Q_3 is the 75th sample percentile.

If we sort a set of data from smallest to largest, each value in the ordered list can be considered to be a sample percentile. The next example illustrates.

Example 6.7: Normal Probability Plots

Consider the following $n = 10$ observations, sorted from smallest to largest.

$$7 \quad 9 \quad 13 \quad 18 \quad 19 \quad 22 \quad 23 \quad 25 \quad 33 \quad 35$$

The fraction of observations that lies below the 3rd smallest one, 13, is 2/10. However, this fraction changes to 2.5/10, or 25%, if we consider 13 to be "halfway below" itself and "halfway above" itself. For this reason we define 13 to be the 25th sample percentile.

Now consider a sample $X_1, X_2, \ldots, X_n$ of size n, sorted from smallest to largest. Thus X_1 is the smallest observation, X_2 the second smallest, and so on. If we consider the ith smallest observation, X_i, to be "halfway below" itself and "halfway above" itself, then the fraction of observations that lie below that one is $(i - 0.5)/n$. It follows that the ith smallest data value X_i is the $100(i - 0.5)/n$th *sample percentile*.

Example 6.8: Normal Probability Plots

Consider again the sorted data in Example 6.7. Since $n = 10$,

$$\begin{array}{rl} 7 & \text{is the 5th sample percentile} \\ 9 & \text{is the 15th sample percentile} \\ 13 & \text{is the 25th sample percentile} \\ & \vdots \\ 35 & \text{is the 95th sample percentile} \end{array}$$

In a random sample from a population, we'd expect each sample percentile to be roughly equal to its corresponding population percentile. For example, if a random sample of size $n = 10$ was drawn from a N(0, 1) distribution, after sorting the data, we'd expect the values in the sorted list to be roughly equal to the 5th, 15th, ..., 95th percentiles of the N(0, 1) distribution. These percentiles (obtained from a standard normal table) are shown below and on the horizontal axis in Fig. 6.11.

$$\begin{array}{rl} -1.64 & \text{is the 5th N(0, 1) percentile} \\ -1.04 & \text{is the 15th N(0, 1) percentile} \\ -0.67 & \text{is the 25th N(0, 1) percentile} \\ & \vdots \\ 1.64 & \text{is the 95th N(0, 1) percentile} \end{array}$$

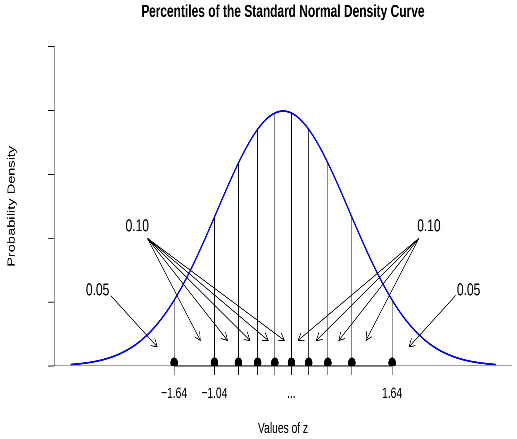


Figure 6.11: Standard normal distribution and its 5th, 15th, ..., 95th percentiles.

A ***normal probability plot*** is a plot of the sorted observations in a data set, or sample percentiles, on the y-axis versus their corresponding percentiles of the N(0, 1) distribution on the x-axis. If the sample came from a N(0, 1) distribution, we'd expect the points in the plot to fall close to the line $y = x$ (that is, the line through the origin with a slope of one). It can be shown that if the sample came from *some* N(μ, σ) distribution (but not necessarily the *standard* normal one), the points in the plot should come close to *some* straight line (specifically, the line $y = \mu + \sigma x$).

In practice, normal probability plots are easily made using statistical software.

Example 6.9: Normal Probability Plots

For the data given in Example 6.7, the normal probability plot is a plot of the x, y pairs:

Percentile	N(0, 1) Percentile x	Sorted Observation (Sample Percentile) y
5th	-1.64	7
15th	-1.04	9
25th	-0.67	13
35th	-0.39	18
45th	-0.13	19
55th	0.13	22
65th	0.39	23
75th	0.67	25
85th	1.04	33
95th	1.64	35

The plot is shown below (having been obtained using statistical software).

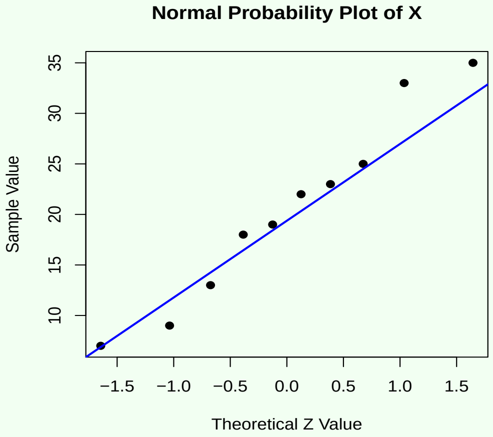


Figure 6.12: Normal probability plot of the data from Example 6.7.

Because the points in the plot hug a straight line fairly closely, it's reasonable to consider the data to be a sample from a normal distribution.

For more intuition about nonlinear patterns in normal probability plots, and in particular curved ones like the one in Fig. 6.7 that indicate right skewed distributions, that plot is reproduced below, this time with the x and y coordinates of the points projected into histograms on their respective axis margins.

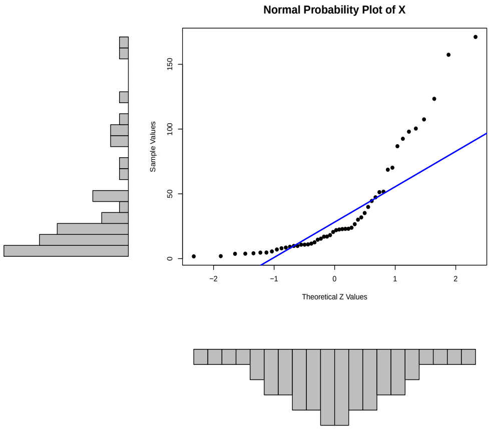


Figure 6.13: Normal probability plot of a right skewed data set. The x and y coordinates of the points in the plot are shown as histograms in the plot margins.

Notice that the histogram of the *observed* sample values in the left margin is right skewed and the histogram of *expected* values under normality at the bottom is bell-shaped.

Comment: Some software uses a slightly different formula than $(i - 0.5)/n$ to determine which N(0, 1) percentiles to use in the normal probability plot. This doesn't dramatically affect the overall appearance of the plot, though, and its interpretation remains the same.

6.10 Transforming Data to Normality

6.10.1 Introduction

Many of the statistical inference procedures that have been developed for use with normally distributed data, including the t confidence interval procedure, are fairly ***robust*** to mild departures from the normality assumption, meaning that they're still approximately valid, even when the sample size n is small, as long as the population from which the sample was drawn has a fairly symmetric, perhaps even mildly skewed distribution. However, when the population is moderately or severely skewed and n isn't large, alternative procedures should be used (such as so-called ***nonparametric*** ones that don't require normality) or the data should be ***transformed*** first, for example by taking their logs, so that the transformed data are more normally distributed.

In this section, we'll focus on transforming data to normality. Nonparametric procedures will be covered in Section 6.11 and throughout later chapters of this book.

6.10.2 The Log Transformation to Normality

Environmental data sets commonly exhibit right skewed distributions and can often be modeled as a random sample from a *lognormal* distribution. We saw in Chapter 4.1 that if a sample is drawn from a lognormal distribution, their (natural) logs can be treated as a sample from a *normal* distribution. Thus we can often apply statistical procedures designed for normally distributed data to the *logs* of right skewed data, as in the next example.

Example 6.10: Log Transformation to Normality

Polychlorinated biphenyls (PCBs) are a collection of compounds no longer produced in the U.S. but still found in the environment. They cause harmful health effects if consumed, and because they can accumulate in fish, consumption limits are recommended in areas where fish contain high PCB levels. In a U.S. Environmental Protection Agency study, PCBs were measured in fish from $n = 69$ U.S. lakes [1]. There are more than 200 types of PCBs. The data below are measured values of the total sum of all PCBs (ppb) found in the fish.

20.0	6.1	25.0	37.4	30.2	20.8	41.4	29.5	24.2	26.3
8.6	36.4	66.4	30.6	25.5	68.6	23.1	43.0	39.5	36.5
26.5	22.1	19.2	33.0	9.1	42.0	48.8	55.9	31.8	60.1
97.3	18.4	27.5	79.0	97.8	44.9	58.2	57.4	57.5	33.6
115.7	14.8	91.6	92.2	37.0	87.7	111.4	48.0	38.1	122.8
113.1	79.2	98.3	33.0	64.2	119.4	80.7	171.4	132.9	91.8
32.6	59.5	200.5	65.4	198.5	89.4	210.3	246.7	318.7	

We want a 95% confidence interval for the true (unknown) mean fish PCB concentration μ in U.S. lakes. A histogram and normal probability plot of the data are below.

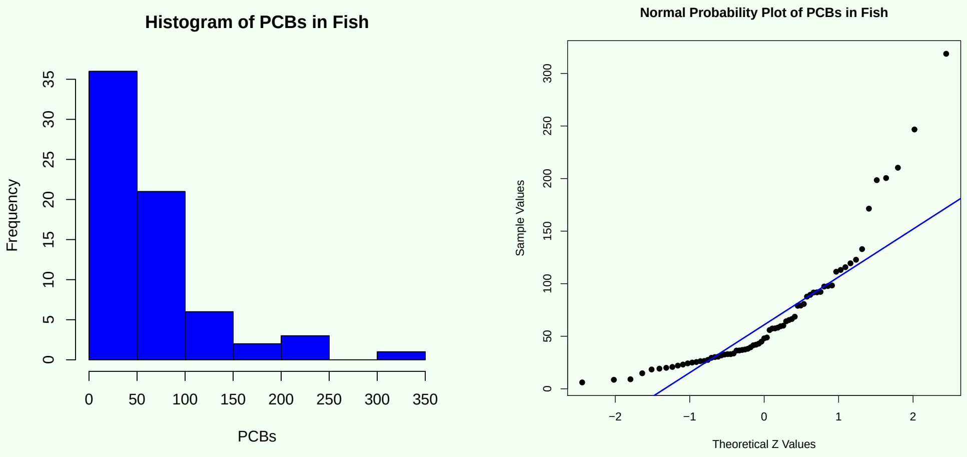


Figure 6.14: Histogram (left) and normal probability plot (right) of PCB concentrations in fish in U.S. lakes.

Based on the plots, it would be unreasonable to assume the data are a sample from a normal distribution. At this point, we have two options:

1. We could proceed with a one-sample t confidence interval, relying on the fact that $n = 69$ is large enough to justify the use of the procedure.
2. We could take logs of the data, rendering them normally distributed, then compute a one-sample t confidence interval using the *log* transformed data.

We'll use the second approach in this example. Here are the logs of the data.

3.0	1.8	3.2	3.6	3.4	3.0	3.7	3.4	3.2	3.3
2.2	3.6	4.2	3.4	3.2	4.2	3.1	3.8	3.7	3.6
3.3	3.1	3.0	3.5	2.2	3.7	3.9	4.0	3.5	4.1
4.6	2.9	3.3	4.4	4.6	3.8	4.1	4.1	4.1	3.5
4.8	2.7	4.5	4.5	3.6	4.5	4.7	3.9	3.6	4.8
4.7	4.4	4.6	3.5	4.2	4.8	4.4	5.1	4.9	4.5
3.5	4.1	5.3	4.2	5.3	4.5	5.3	5.5	5.8	

A histogram and normal probability plot of these values are below.

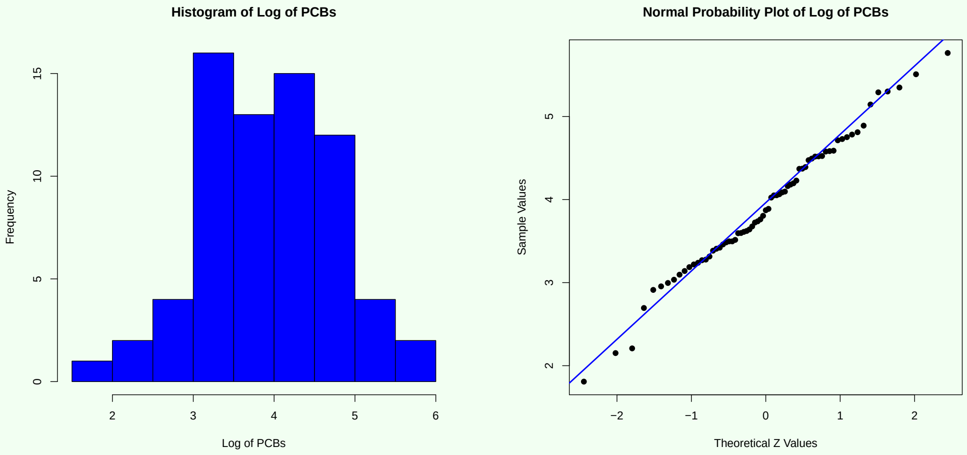


Figure 6.15: Histogram (left) and normal probability plot (right) of the natural logs of PCB concentrations in fish in U.S. lakes.

It's clear that the *logs* of the PCB concentrations follow a bell-shaped distribution, so it seems reasonable to treat them as a random sample from a normal population. We'll let μ denote the mean of this log-scale population.

The sample mean and standard deviation of the log PCB concentrations are

$$\begin{aligned} \bar{Y} &= 3.92 \\ S &= 0.80, \end{aligned}$$

so the (estimated) standard error is

$$S_{\bar{Y}} = \frac{0.80}{\sqrt{69}}.$$

Thus the point estimate for the (unknown) true mean log PCB concentration μ is $\bar{Y} = 3.92$ (in log ppb), and The 95% one-sample t confidence interval for μ is

$$\begin{aligned} \bar{Y} \pm t_{0.025,68}\, S_{\bar{Y}} &= 3.92 \pm 1.995 \left(\frac{0.80}{\sqrt{69}} \right) \\ &= 3.92 \pm 0.19 \\ &= (3.73,\ 4.11) \end{aligned}$$

where the t critical value $t_{0.025,68} = 1.995$ was obtained from t distribution table. We can be 95% confident that the true mean *log* PCB concentration μ is between 3.73 and 4.11 log ppb.

6.10.3 Back-Transforming to the Original Scale

When we perform a statistical analysis on log transformed data, the results pertain to the transformed measurement scale for the data. In the last example, the confidence interval was for the population mean *log* PCB concentration (in log ppb), *not* the mean PCB concentration (in ppb). Sometimes it's possible *back-transform* the results to the original scale, as illustrated in the next example.

Example 6.11: Back-Transforming Results to the Original Scale

The sample mean *log* PCB concentration in fish (from the previous example) is $\bar{Y} = 3.92$ (log ppb). We can convert $\bar{Y}$ back to the original scale (ppb) by taking its antilog:

$$\begin{aligned} e^{\bar{Y}} &= e^{3.92} \\ &= 50.4 \end{aligned}$$

(ppb). This the *geometric mean* (Chapter 3) of the original, untransformed PCB concentrations. It's easier to interpret than $\bar{Y}$ because it's measured on the original scale (ppb). It's an estimate of e^{μ} and, as pointed out in Chapter 3, can be thought of as an estimate of the *median* of the original PCB concentration distribution. The reason for this is that the original PCB distribution can be modeled by a lognormal distribution, ant the median of a lognormal distribution, as mentioned in Chapter 4, is e^{μ}.

Now consider the 95% confidence interval

$$(3.73,\ \ 4.11)$$

for the true mean *log* PCB concentration μ (from Example 6.10). We can take the antilogs of the interval endpoints, giving the new interval

$$(e^{3.73},\ \ e^{4.11}) \ = \ (41.7,\ \ 60.9).$$

The values 41.7 and 60.9 are measured in ppb, so they're easier to interpret than when they were on the log scale. These values give a 95% confidence interval for e^{μ}, the true *median* PCB concentration. This will be revisited in Subsection 6.11.1.

6.10.4 Other Transformations: The Ladder of Powers

Although the log transformation often results in data having a bell-shaped distribution, allowing the application of statistical procedures designed for normally distributed data, some data sets still exhibit skewness even after a log transformation has been made.

For such data sets, there are a number of other transformations we can try. The transformations are often of the form

$$Y \ = \ X^{\theta},$$

where X is the original data value, Y is the transformed value, and θ is a power chosen so as to make the distribution of Y as close to a normal distribution as possible. The following ***Ladder of Powers*** can be used as a guide for determining an appropriate choice for θ:

The Ladder of Powers

Use	$\boldsymbol{\theta}$	**Transformation**	**Name**	**Comment**
Left-skewness	3	$Y = X^3$	Cube	
Left-skewness	2	$Y = X^2$	Square	
	1	$Y = X$	Original Measurement Scale	No transformation.
Right-skewness	1/2	$Y = \sqrt{X}$	Square Root	
Right-skewness	1/3	$Y = \sqrt[3]{X}$	Cube Root	
Right-skewness	"0"	$Y = \log(X)$	Logarithm	Commonly used.
Right-skewness	-1/2	$Y = \frac{-1}{\sqrt{X}}$	Reciprocal Square Root	The minus sign preserves the order of the observations.
Right-skewness	-1	$Y = \frac{-1}{X}$	Reciprocal	The minus sign preserves the order of the observations.

To use the Ladder of Powers, starting with the original measurement scale ($\theta = 1$), we try transforming the data using values of θ successively farther and farther away from one in the appropriate direction (increasing θ if the data are left skewed, decreasing it if they're right skewed). With each transformation, we examine a normal probability plot and also perhaps a histogram to decide if the transformation was successful.

Example 6.12: Transforming Data Using the Ladder of Powers

To illustrate the use of the Ladder of Powers for choosing a transformation, we'll apply it to the PCB data of Example 6.10.

We begin by looking again at the normal probability plot of the original, untransformed data (ppb), shown on the top left below.

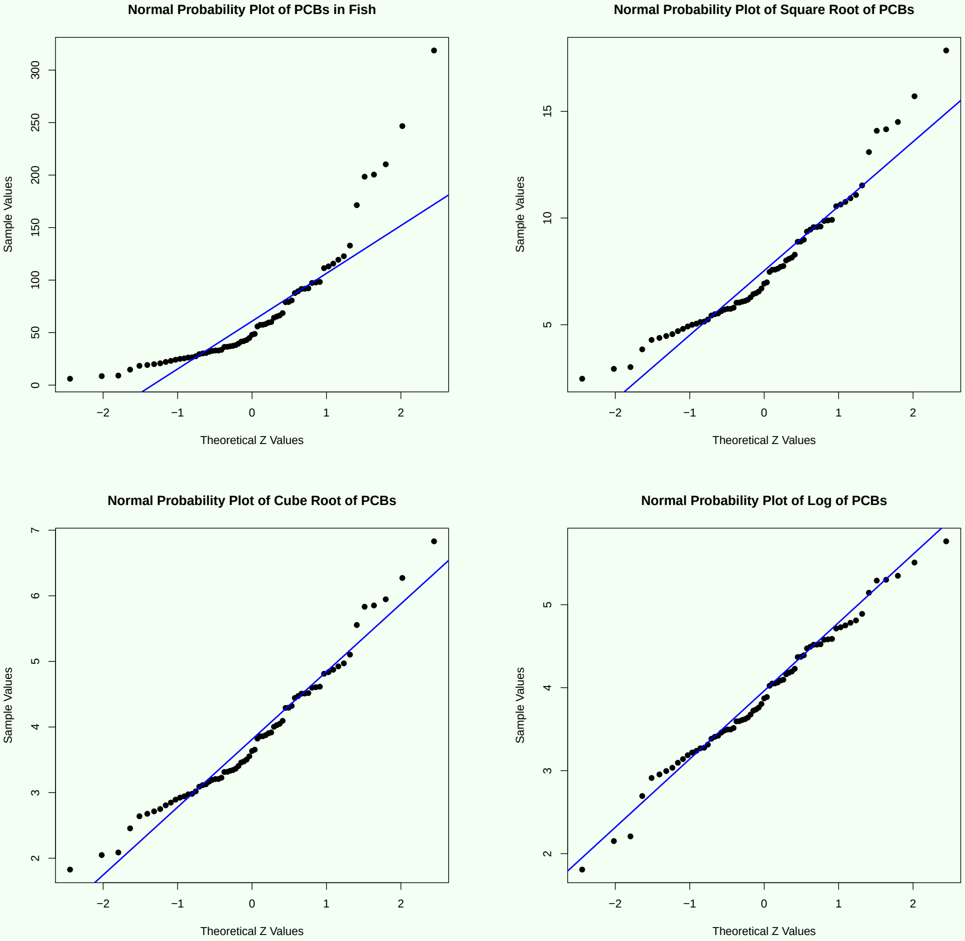


Figure 6.16: Normal probability plots. The original data on PCBs in fish from U.S. lakes (top left); the data after a square root transformation (top right); after a cube root transformation (bottom left); after a natural log transformation (bottom right).

Since the original PCB data are right-skewed, we first move ”down the ladder” one step and try the square root transformation. A normal probability plot of the square roots of the PCB concentrations is shown on the top right of Fig. 6.16. It’s apparent that the data are still right skewed, but the skewness is less severe.

Moving ”down the ladder” another step, we next try the cube root transformation. The normal probability plot after making this transformation is shown on the bottom left of Fig. 6.16. There’s still a bit of skewness, so we move ”down the ladder” one more step and try the log transformation. A normal probability plot of the logs of the data is shown in the bottom right of Fig. 6.16 and a histogram in Fig. 6.15. The plots indicate that the logs of the PCB concentrations can be treated as a random sample from a normal distribution. This was the transformation we used in Example 6.10 to validate the use of the one-sample t procedure.

6.11 Confidence Interval for a Population Median

6.11.1 Confidence Interval for a Population Median

Recall that the *median* of a continuous population, $\tilde{\mu}$, is the value below which 50% of the population lies and above which the other 50% lies. For a *normal* population, the median and mean are one and the same, so the sample mean $\bar{X}$ is a point estimate of both $\tilde{\mu}$ and μ and, the one-sample t confidence interval for μ is also a confidence interval for $\tilde{\mu}$.

For a sample from a *lognormal* population, we can obtain a confidence interval for the population median as follows. First, we take the logs of the data, so that the resulting values can be treated as a sample from a *normal* population. Then we compute a one-sample t confidence interval for the mean μ of this *normal* population. Finally, we back-transform the endpoints of the interval by taking their antilogs, as in Example 6.11. We arrive at a confidence interval for the *antilog* of μ, that is for e^{μ}, which, recall (from Chapter 4), is the *median* of the original lognormal population.

6.11.2 Nonparametric Confidence Interval for a Population Median

For non-normally distributed data or data whose distribution is unknown, a point estimate for the population median $\tilde{\mu}$ is the sample median $\tilde{X}$, and we can construct a *nonparametric* confidence interval for $\tilde{\mu}$. The values that make up the confidence interval answer the question: ”What set of values are plausible for $\tilde{\mu}$ in light of the observed data?”

To answer this question, note that in a randomly selected sample of size n from the population, each observation X_i is equally likely to fall below or above $\tilde{\mu}$ (because it's the 50th percentile), so we'd expect about half of the sample observations to fall below $\tilde{\mu}$ and the other half above it.

The actual number of observations falling below $\tilde{\mu}$, though, is a random variable. And because it's a count among n observations, each equally likely to fall above or below $\tilde{\mu}$, it follows a binomial distribution with parameters n and $p = 0.5$. Thus when the value of $\tilde{\mu}$ is unknown, we can assess the plausibility of any candidate value by counting the number of sample observations that are less than that value and comparing that count to the binomial distribution. If the count is smaller or larger than could be expected to occur just by chance, then the candidate value for $\tilde{\mu}$ is deemed implausible and is excluded from the confidence interval.

Example 6.13: Nonparametric Confidence Interval for a Median

The following data, sorted from smallest to largest, are measurements of percent total organic carbon (TOC) in clay from rivers in Spain [10].

0.21 0.22 0.33 0.34 0.35 0.35 0.37 1.28 1.34 1.82 1.88 1.91 2.26 2.57

Looking at the data, we can assess the plausibility of various values for the population median TOC percent, $\tilde{\mu}$. It doesn't seem plausible, for example, that $\tilde{\mu}$ is as large as, say, 2.6 because that would mean that the entire sample of $n = 14$ observations would be less than $\tilde{\mu}$. From the binomial(14, 0.5) distribution, the chance of this happening is only 0.00006. Similar reasoning rules out values of $\tilde{\mu}$ below 0.21 at the other end of the data set.

On the other hand, 1.3 is a plausible value for $\tilde{\mu}$ because in this case 8 of the $n = 14$ sample values would be below $\tilde{\mu}$, and the chance of this happening is high, 0.1833 to be exact (from the same binomial distribution).

The probability histogram of the binomial(14, 0.5) used in the previous example is shown below.

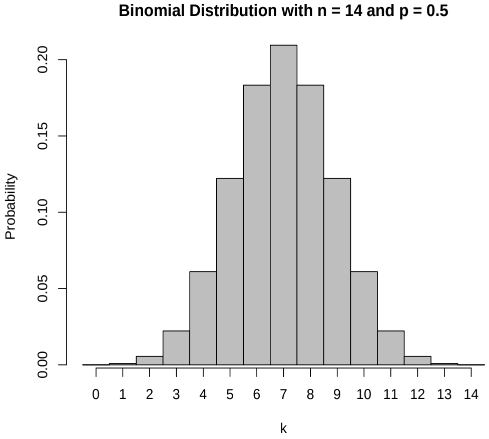


Figure 6.17: Probability distribution of the number of observations in a random sample of size $n = 14$ falling below the median $\tilde{\mu}$ of the population. This is just the binomial(14, 0.5) distribution.

It can be shown, using the probability histogram above or the binomial probability function of Chapter 4, that in 94.3% of all random samples of size $n = 14$, no fewer than 4 and no more than 10 observations will fall below population median $\tilde{\mu}$. This is the middle 94.3% of the distribution in Fig. 6.17.

This is the same as saying that 94.3% of the time, $\tilde{\mu}$ will be no smaller than the 4th smallest observation in the data set and no larger than the 11th smallest. It follows that for a random sample of size $n = 14$ from a population, a 94.3% *confidence interval for the population median* $\tilde{\mu}$ has endpoints

$$(4\text{th smallest observation},\ 11\text{th smallest observation}).$$

Example 6.14: Nonparametric Confidence Interval for a Median

Continuing from the previous example, if the TOC percents were a sample from a *normal* population, the population median $\tilde{\mu}$ and mean μ would be equal, and in this case the sample mean $\bar{X}$ would estimate both, and a t confidence interval for μ would also be a confidence interval for $\tilde{\mu}$.

But a histogram and normal probability plot of the TOC data reveal that the normality assumption isn't tenable – the data are right-skewed.

In this case, the appropriate point estimate of the true (unknown) population median TOC percent $\tilde{\mu}$ is the sample median $\tilde{x} = 0.825$. The 94.3% *nonparametric* confidence interval for $\tilde{\mu}$ is

$$(0.34,\ 1.88),$$

the 4th and 11th smallest observations in the sorted data set. We can be 94.3% confident that $\tilde{\mu}$ is in this interval somewhere.

Note: Due to the discreteness of the binomial distribution, we usually can't compute a confidence interval having the exact level of confidence that we desire (such as 95%). Instead we must be satisfied with the discrete set of available choices for the confidence level. For example, in Example 6.14 the confidence interval that ranged from the 4th to 11th smallest data values had a confidence level 94.3%. An interval ranging from the 3rd to 12th would have had a confidence level 98.7%. We couldn't construct a confidence interval with a confidence level between these two levels.

Here's the general procedure for constructing a nonparametric confidence interval for a population median.

Nonparametric Confidence Interval for a Population Median: Suppose $X_1, X_2, \ldots, X_n$ are a random sample from *any continuous* population (not necessarily normal) whose median is $\tilde{\mu}$. Then for a desired confidence level $\mathbf{100(1-\alpha)\%}$, the ***nonparametric confidence interval for*** $\boldsymbol{\tilde{\mu}}$ is

$$((k_1+1)\text{th smallest observation},\ k_2\text{th smallest observation})\,, \tag{6.11}$$

where k_1 is the largest value among $0, 1, \ldots, n$ for which

$$P(X \le k_1) \;\le\; \frac{\alpha}{2},$$

with $X \sim \text{binomial}(n,\, 0.5)$, and k_2 is the smallest value among $0, 1, \ldots, n$ for which

$$P(X \ge k_2) \;\le\; \frac{\alpha}{2}.$$

The values for k_1 and k_2 can be determined using statistical software or obtained from a table of the binomial distribution. We an be $100(1-\alpha)\%$ confident that $\tilde{\mu}$ will fall in this range.

Note: Using the above procedure, the *actual* confidence level is guaranteed to be no smaller than $100(1-\alpha)\%$. In other words, we can be *at least* $100(1-\alpha)\%$ confident that $\tilde{\mu}$ will be contained in the interval (6.11). The actual confidence level is given by the probability that a binomial(n, 0.5) random variable will fall between k_1+1 and k_2-1, inclusive.

6.12 Tolerance Limits and Confidence Interval for a Population Percentile

6.12.1 Introduction

Environmental compliance monitoring often involves comparing the concentration of a contaminant at a cleanup site or in a monitoring well to a threshold value above which the concentration would be deemed to be in violation of environmental standards.

The determination of an appropriate value for such a threshold must take into account that background concentrations of the contaminant vary naturally from one day to the next and from one location to the next. In particular, the threshold value should be high enough that naturally varying background concentrations would exceed it only very rarely, say, no more than 1% of the time.

Recall that the ***100pth percentile*** of a population is the value below which $100p\%$ of the population lies, where p is between zero and one. Thus the 99th percentile of the naturally varying background concentrations is the value that would be exceeded just by chance only 1% of the time.

We'll denote the $100p$th percentile of a population by $\boldsymbol{x_{1-p}}$. For example, the 99th percentile will be denoted by $x_{0.01}$. For a *normal* population, we know (Chapter 4) that

$$x_{1-p} \;=\; \mu \,+\, z_{1-p}\sigma,$$

where μ and σ are the population mean and standard deviation, and $\boldsymbol{z_{1-p}}$ is the $100p$th percentile of the standard normal distribution. Values of z_{1-p}, for specified p, can be obtained from a standard normal table.

If background concentrations followed a normal distribution, and μ and σ were known, we could use the 99th percentile of the distribution, $x_{0.01}$, as the threshold for environmental compliance monitoring. This value would be exceeded by only 1% of background concentrations.

In practice, though, μ and σ won't be known, so population percentiles such as $x_{0.01}$ can't be determined. If we have a sample of background concentrations, though, we can estimate μ and σ from the data, which leads to the following *point estimator* of x_{1-p}, denoted $\boldsymbol{\hat{X}_{1-p}}$.

Estimator of a Population Percentile: Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a *normal* population. Then an estimator of the 100pth percentile of the population is

$$\hat{X}_{1-p} \;=\; \bar{X} + z_{1-p} S\,,$$

where $\bar{X}$ and S are the sample mean and sample standard deviation.

This estimator *could* be used as the threshold for environmental compliance monitoring, but because it's based on statistics ($\bar{X}$ and S) that vary from one random sample to the next, it's value is subject to sampling error and therefore not entirely reliable.

Example 6.15: Point Estimate of a Population Percentile

Consider once again the $n = 10$ background soil radiation levels from along the Front Range (Example 6.1), and suppose we want to use them to establish the soil action level for monitoring the Rocky Flats cleanup.

The sample mean and standard deviation (from Example 6.3) are

$$\begin{aligned} \bar{X} &= 2.14 \\ S &= 0.76. \end{aligned}$$

Assuming the background radiation level population is a normal distribution, the point estimate of the 99th percentile of the distribution, $x_{0.01}$, is

$$\hat{X}_{0.01} \;=\; 2.14 + 2.33(0.76) \;=\; 3.91 \;\; \text{Bq/kg},$$

where the percentile $z_{0.01} = 2.33$ was obtained from a standard normal table.

This value 3.91 *could* be used as the soil action level, but because it's just based on one random sample of 10 background measurements, its reliability is in question.

Instead of using the point estimate of $x_{0.01}$ as the compliance threshold, more often a higher value, called an *upper tolerance limit*, that takes sampling error into consideration is used. In the next subsection we'll see how to construct tolerance limits, and in Subsection 6.12.3 confidence intervals for percentiles, when the population is normal. For data from a lognormal population, the procedures can be used after taking the logs of the data, and the resulting tolerance limit (or confidence interval endpoints) back-transformed by taking the antilog(s), as in Subsection 6.10.3.

6.12.2 Tolerance Limits

An upper ***tolerance limit*** with ***level of confidence*** $\boldsymbol{100(1-\alpha)\%}$ and ***coverage*** $\boldsymbol{100p\%}$ is a one-sided upper confidence limit for a population percentile x_{1-p}. In other words, it's a value below which we can be $100(1-\alpha)\%$ confident that x_{1-p}, and therefore *at least* $100p\%$ of the population, lies. A lower tolerance limit is defined analogously.

The tolerance limits we'll look at use critical values from a distribution called the ***noncentral t distribution***, which has two parameters, its ***degrees of freedom*** and a ***noncentrality parameter***. Here's how the limits are computed.

Tolerance Limits: Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a *normal* population.

Then an ***upper tolerance limit***, with ***level of confidence* $\boldsymbol{100(1-\alpha)\%}$** and ***coverage* $\boldsymbol{100p\%}$**, is

$$\bar{X} \, + \, t_{\alpha, n-1, ncp} \, S_{\bar{X}},$$

and a ***lower tolerance limit*** with the same confidence level and coverage is

$$\bar{X} \, + \, t_{1-\alpha, n-1, ncp} \, S_{\bar{X}},$$

where

$$S_{\bar{X}} \;=\; \frac{S}{\sqrt{n}},$$

and the *t critical values* $t_{\alpha,n-1,ncp}$ and $t_{1-\alpha,n-1,ncp}$ are the $100(1-\alpha)$th and 100αth percentiles, respectively, of the noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $ncp = \sqrt{n}\, z_{1-p}$.

The critical values from the noncentral t distribution can be obtained from tables or using statistical software. We can be $100(1-\alpha)\%$ confident that the true (unknown) population percentile x_{1-p} will be below the upper tolerance limit, that is, that at least $100p\%$ of the population will be below that limit. In practice, this means that for an upper tolerance limit with 95% level of confidence and 99% coverage, we can be 95% confident that *at most* 1% of all background contaminant concentrations exceed the tolerance limit.

The lower tolerance limit has an analogous interpretation. We can be $100(1-\alpha)\%$ confident that x_{1-p} will be above that limit, that is, that at least $100(1-p)\%$ of the population will lie above it. Thus for a *lower* tolerance limit with 95% level of confidence and 99% coverage, we can be 95% confident that *at most* 1% of all background contaminant concentrations are *below* the tolerance limit.

The derivation of the tolerance limit formulas is given in Subsection 6.12.4.

Example 6.16: Tolerance Limits

Continuing from the last example, the upper tolerance limit for radiation concentrations in the Rocky Flats cleanup, with 95% confidence level and 99% coverage, is

$$2.14 \, + \, 12.59 \left(\frac{0.76}{\sqrt{10}} \right) \;=\; 5.17 \;\; \text{Bq/kg},$$

where the critical value $t_{0.05,9,\sqrt{10}(2.33)} = 12.59$ was obtained using statistical software.

We can be 95% confident *no more than* 1% of all background soil radiation concentrations along the Front Range are above 5.17 Bq/kg.

Any measured soil radiation concentration above 5.17 Bq/kg at the Rocky Flats cleanup site would trigger an ”action” such as removal or containment. Note that this threshold value is higher than the estimate 3.91 Bq/kg of $x_{0.01}$ from Example 6.15 because it accounts for sampling error in that estimate.

6.12.3 Confidence Interval for a Population Percentile

If our goal is merely to *estimate* a population percentile x_{1-p}, and we want to include a margin of error with the estimate, we can compute a *two-sided* confidence interval for x_{1-p}. Here's the interval.

Confidence Interval for a Population Percentile: Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a *normal* population. Let x_{1-p} be the 100pth percentile of the distribution, where p is between zero and one.

Then a $\boldsymbol{100(1-\alpha)\%}$ ***confidence interval for*** $\boldsymbol{x_{1-p}}$ is

$$\left(\bar{X} + t_{1-\alpha/2,n-1,ncp}\, S_{\bar{X}},\ \ \bar{X} + t_{\alpha/2,n-1,ncp}\, S_{\bar{X}}\right). \tag{6.12}$$

where

$$S_{\bar{X}} = \frac{S}{\sqrt{n}},$$

and $t_{\alpha/2,n-1,ncp}$ and $t_{1-\alpha/2,n-1,ncp}$ are the $100(1-\alpha/2)$th and $100\alpha/2$th percentiles, respectively, of the noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $ncp = \sqrt{n}\, z_{1-p}$.

We can be $100(1-\alpha)\%$ confident that the true (unknown) population percentile x_{1-p} will be contained in this interval.

6.12.4 Derivation of the Tolerance Limit and Confidence Interval Formulas

The derivation of the formula (6.12) for the confidence interval for a percentile is based on the random variable

$$\frac{\hat{X}_{1-p} - x_{1-p}}{S} = \frac{(\bar{X} + z_{1-p}\, S) - x_{1-p}}{S}, \tag{6.13}$$

which follows a certain probability distribution from which values a and b can be obtained so as to satisfy

$$1 - \alpha = P\left(a < \frac{(\bar{X} + z_{1-p}\, S) - x_{1-p}}{S} < b\right)$$

(that is, a and b capture the middle $100(1-\alpha)\%$ of the distribution of the random variable (6.13)). After rearranging terms, we can write this as

$$1 - \alpha = P\left(\bar{X} + (z_{1-p} - b)\, S < x_{1-p} < \bar{X} + (z_{1-p} - a)\, S\right). \tag{6.14}$$

It can be shown that a and b are the values for which

$$(z_{1-p} - b) = \frac{t_{1-\alpha/2,n-1,ncp}}{\sqrt{n}} \qquad \text{and} \qquad (z_{1-p} - a) = \frac{t_{\alpha/2,n-1,ncp}}{\sqrt{n}},$$

which, together with (6.14), confirms the confidence interval formula (6.12). The tolerance limits are just one-sided confidence intervals for x_{1-p}, and so their formulas can be derived in a similar manner.

6.12.5 Nonparametric Confidence Interval for a Population Percentile

For non-normally distributed data or data whose distribution is unknown, we can use a sample percentile, which is just an observation in the sorted data set (Subsection 6.9.4), as a point estimate of the corresponding population percentile.

We can also construct a *nonparametric* confidence interval for a percentile x_{1-p} of a *non-normal* population using a method similar to that used to construct the nonparametric interval for the population median (Subsection 6.11.2). Details can be found in [7].

6.13 One-Sample Z Confidence Interval for a Population Proportion

Consider now data on a dichotomous categorical variable that takes values *success* and *failure*, and suppose the data are a sample from a population whose proportion of successes is p. The (point) estimate of p is the sample proportion $\hat{P}$. A confidence interval for p is given by the following.

One-Sample Z Confidence Interval (for a Proportion): Suppose we have a random sample of size n from a dichotomous population. Let p denote the proportion of *successes* in the population. A **$100(1-\alpha)$%** ***one-sample z confidence interval for p*** is

$$\hat{P} \;\pm\; z_{\alpha/2}\, S_{\hat{P}}, \qquad \text{where} \qquad S_{\hat{p}} \;=\; \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

The confidence interval is valid if the sample size n is *large* (as defined in Chapter 5).

The *margin of error* is the ”plus or minus” part of the confidence interval.

Margin of Error: For the one-sample z confidence interval for p, the margin of error is

$$\text{Margin of Error} \;=\; z_{\alpha/2}\, S_{\hat{P}} \;=\; z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

Example 6.17: Confidence Interval for a Proportion

In a study of bee populations in southeast Asia [9], a honey solution was sprayed on vegetation along several transects. Bees attracted to the honey were caught with an insect net, and their species later identified.

The study found that among the 1,631 individual bees caught, 546 were of the species *Trigona* (*Tetragonula*) *laeviceps*. Thus the sample proportion is

$$\hat{P} \;=\; \frac{546}{1,631} \;=\; 0.33,$$

and so the (estimated) standard error is

$$S_{\hat{P}} \;=\; \sqrt{\frac{0.33(1-0.33)}{1,631}} \;=\; 0.01.$$

A 95% one-sample z confidence interval for the true (unknown) population proportion that are *T.* (*T.*) *laeviceps* is

$$\begin{aligned} \hat{P} \;\pm\; z_{0.025}\, S_{\hat{P}} &= 0.33 \pm 1.96\,(0.01) \\ &= 0.33 \pm 0.02 \\ &= (0.31,\, 0.35). \end{aligned}$$

The estimate (0.33) of the true proportion has margin of error 0.02, and we can be 95% confident that the true proportion is between 0.31 and 0.35.

To see how the confidence interval formula was derived, recall (Chapter 5) that if $\hat{P}$ is the proportion of *successes* in a sample for which a dichotomous variable (taking values *success* and *failure*) is measured on each individual, then if n is large,

$$\hat{P} \sim \mathrm{N}\left(\mu_{\hat{p}},\, \sigma_{\hat{p}}\right)$$

(approximately), where

$$\mu_{\hat{P}} \;=\; p \qquad \text{and} \qquad \sigma_{\hat{P}} \;=\; \sqrt{\frac{p(1-p)}{n}}.$$

and p is the proportion of *successes* in the population. Using an argument similar to that used in (6.2) of Section 6.2, it can be shown that we can be 95% confident that p will be contained in the interval

$$\hat{P} \;\pm\; 1.96\,\sigma_{\hat{p}}. \tag{6.15}$$

Unfortunately, the standard error $\sigma_{\hat{p}}$ depends on the (unknown) population proportion p. To get around this, we use the ***estimated standard error***, denoted $\boldsymbol{S_{\hat{p}}}$, obtained by plugging the *estimate* $\hat{P}$ in for p in $\sigma_{\hat{p}}$,

$$S_{\hat{p}} \;=\; \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

Thus the 95% confidence interval is

$$\hat{P} \;\pm\; 1.96\,S_{\hat{p}}.$$

For a generic confidence level, 1.96 is replaced by the z critical value $z_{\alpha/2}$.

6.14 Problems

6.1 Based on 43 years of hydrological data from 1959-2001, the U.S. Geological Survey estimates that a stream's mean yearly discharge is 4,300 million m^3 with a margin of error of 300 million m^3. Assuming a 95% confidence interval was used, for each of the following, state whether it's a legitimate interpretation of the confidence interval. For any that aren't legitimate, say why.

a) For 95% of all years, the stream's discharge will be between 4,000 million and 4,600 million m^3.

b) Water management officials can be 95% sure that the stream's average yearly discharge for the 43 years from 1959-2001 was between 4,000 million and 4,600 million m^3.

c) Water management officials can be 95% sure that the stream's average yearly discharge for all years is between 4,000 million and 4,600 million m^3.

6.2 The pesticide DDT was measured (ng/g) in the blood of each of a random sample of northern fur seals on St. George Island, Alaska [3]. The confidence intervals below, for the true mean DDT concentration μ, were both computed from the same set of data:

$$(4.28,\;11.72) \qquad (5.21,\;10.79)$$

a) One of them is a 95% confidence interval and the other a 99% confidence interval. Which is which? Explain how you know.

b) What's the value of the sample mean DDT concentration $\bar{X}$?

6.3 Studies of the accumulation of PCBs (polychlorinated biphenyls) in dolphins have found that for males, accumulation continues throughout the animal's lifetime, but for females, concentrations tend to decline with reproductive activity through transfer across the placenta and via lactation.

In one capture-and-release study, PCB concentrations (ppm) were measured in the blubber of 25 female dolphins near Sarasota Bay, Florida [17]. Eight of the dolphins were nulliparous (had never given birth) and 17 were parous (had given birth at least once). Here are the summary statistics.

Dolphin Group	Sample size	Mean PCB in blubber	Standard deviation
Nulliparous	8	27.7	10.67
Parous	17	6.8	5.45

a) Compute a 95% confidence interval for the true mean PCB concentration in nulliparous Sarasota Bay dolphins.

b) Compute a 95% confidence interval for the true mean PCB concentration in parous Sarasota Bay dolphins.

c) If the distributions of PCB concentrations in the two dolphin populations were right skewed, would the confidence intervals in parts *a* and *b* be valid? Explain your answer. **Hint**: A *t* confidence interval will only be valid if either 1) the sample is from a normal population, or 2) the sample size is large.

6.4 Perchlorate is a chemical used in rocket and missile fuels, fireworks, road flares, blasting agents and automobile airbags. Due to improper past disposal practices, it's been detected as a contaminant in water supplies throughout the U.S., and also in lettuce, tomatoes, cucumbers, cantaloupe, and dairy milk. Human exposure to perchlorate inhibits iodide uptake by the thyroid. Iodide is a key component in certain hormones used to regulate cell respiration, energy production, growth, and maturation of body tissues.

A toxicity study was carried out to investigate the perchlorate-induced reduction of thyroidal iodide uptake [16], [8]. Volunteer human subjects were assigned to treatment groups given different doses of perchlorate through drinking water, and the percent change in thyroidal uptake after 24 hours was measured. Summary statistics are given below for two of the dose groups, one given 0.007 mg/kg/day in drinking water (Low Dose) and the other 0.020 mg/kg/day (High Dose).

Perchlorate dose group	Number of subjects in the dose group	Mean change (%) in thyroid iodide uptake after 24 hr of dosing	Standard deviation
Low Dose	7	-1.8	22.0
High Dose	10	-16.4	12.8

a) Compute a 99% confidence interval for the true mean percent change in thyroidal uptake at the low dose.

b) Does the confidence interval of part *a* provide convincing evidence that exposure at the low dose results in a decrease in thyroidal uptake? Explain your answer. **Hint**: Does the entire confidence interval lie below zero?

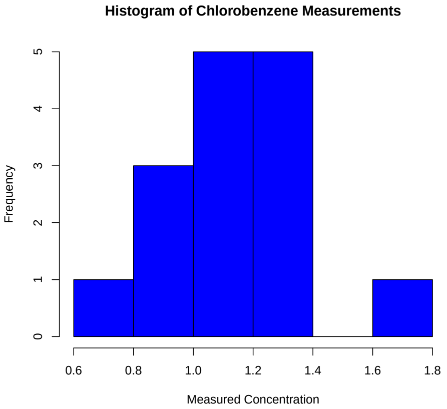
c) Compute a 99% confidence interval for the true mean percent change in thyroidal uptake at the high dose.

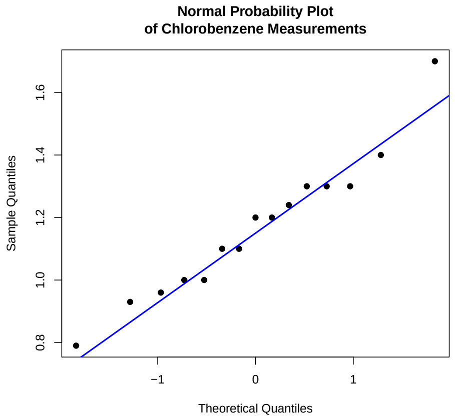
d) Does the confidence interval of part *c* provide convincing evidence that exposure at the high dose results in a decrease in thyroidal uptake? Explain your answer. **Hint**: See the hint for part *b*.

e) If percent changes followed right skewed distributions, would the confidence intervals in parts *a* and *c* be valid? Explain your answer.

6.5 Problem 3.4 in Chapter 3 described a study investigating the use of a capillary column gas chromatography method for measuring organic compounds in water. Each of 15 laboratories used the method to measure chlorobenzene (μg/L) in a reference water specimen certified as having a true concentration 1.10 μg/L of chlorobenzene.

The sample mean and sample standard deviation of the 15 measurements are 1.17 and 0.22, respectively. A histogram and normal probability plot of the data are shown below.





a) Based on the plots, is it reasonable to assume that the data are a random sample from a normal distribution?

b) If the chromatography method is systematically inaccurate, the true mean reading μ will differ from 1.10 μg/L. Compute a 95% confidence interval for the true mean reading μ.

c) Based on the confidence interval of part *b*, is there any convincing evidence that the method is systematically inaccurate? Explain your answer. **Hint**: Does 1.10 lie outside the confidence interval?

6.6 A pilot study of concentrations of the pesticide DDT (ng/g) in crocodile eggs found that a reasonable guess for the population standard deviation σ is 6.95 ng/g [12]. If you were to carry out a follow-up study, and you wanted the margin of error in a 90% confidence interval for the true mean DDT concentration μ to be no greater than 2.0 ng/g, how many crocodile eggs should you sample?

6.7 A study is to be carried out to estimate the population mean phosphorus concentration μ along a creek. A reasonable guess for the population standard deviation σ in this creek is 0.102 mg/L. If it's desired that the margin of error in a 95% confidence interval for μ be no greater than 0.03 mg/L, in how many water specimens should be sampled?

6.8 A study is to be carried out to estimate the true mean pH μ of rainfalls in an area that suffers from heavy pollution due to the discharge of smoke from a power plant. Assume that the true standard deviation σ is in the neighborhood of 0.5 pH and that we want the margin of error in a 95% confidence interval for μ be no greater than 0.1.

a) How many rainfalls should be included in the sample (one pH reading per rainfall)?

b) Would it be valid to select all of the water specimens from a single rainfall? Explain. **Hint**: The z and t confidence interval procedures are only valid if the pH readings in the sample are true *replicates* (that is, *independent* of each other). Readings made too close together (in space or time) are *pseudoreplicates* (Chapter 2).

6.9 The table below shows groundwater monitoring data for benzene measured quarterly from 1996 through 1999 under the regulatory oversight of the Los Angeles Regional Water Quality Control Board in California [14].

Groundwater Benzene

Sampling Date	Benzene (μg/L)
Mar 1996	810
Jun 1996	1200
Sep 1996	200
Dec 1996	980
Feb 1997	340
May 1997	320
Jul 1997	3.8
Oct 1997	480
Mar 1998	180
Jun 1998	38
Aug 1998	55
Dec 1998	110
Mar 1999	35
May 1999	230
Aug 1999	130
Dec 1999	120

a) Make a histogram and a normal probability plot of the benzene concentrations.

b) Based on the plots of part *a*, which distribution, the normal or the lognormal, would be a better model for describing benzine concentrations in this stream?

c) Base on your answer to part *b*, does it appear that the normality assumption required for the one-sample t confidence interval procedure is met?

d) Take the natural log of each benzene concentration and make a histogram and a normal probability plot of the log concentrations.

e) Does the normality assumption appear to met by the log concentrations?

f) Using the log concentrations, compute a 95% confidence interval for the true (unknown) mean log benzene concentration.

g) Back-transform the endpoints of the confidence interval of part *f* to the original measurement scale (μg/L) by taking their antilogs.

6.10 The following measurements of chromium (Cr), zinc (Zn), and manganese (Mn) were made in sediments in the Guanabara Bay, Rio de Janeiro, Brazil [13].

Metals in Guanabara Bay Sediments

Cr	Zn	Mn
9.1	57.2	89.6
16.0	79.3	364.2
15.8	89.9	318.1
17.2	60.2	272.3
16.9	68.4	177.7
19.9	90.6	276.5
22.4	100.9	441.7
27.2	203.2	339.2
32.9	336.8	381.4
18.7	97.7	167.9
1.0	94.3	19.7
18.7	193.4	133.0
22.0	106.3	258.9
2.7	8.2	61.9
25.1	199.5	166.1

a) Compute a 96.48% nonparametric confidence interval for the true (unknown) median chromium concentration $\tilde{\mu}_{\mathrm{Cr}}$ in the bay's sediments.

b) Compute a 99.26% nonparametric confidence interval for the true (unknown) median zinc concentration $\tilde{\mu}_{\mathrm{Zn}}$ in the bay's sediments.

c) Compute an 88.16% nonparametric confidence interval for the true (unknown) median manganese concentration $\tilde{\mu}_{\mathrm{Mn}}$ in the bay's sediments.

6.11 The background radiation concentrations made along the Colorado Front Range and reported in Example 6.1 were measured at soil depths up to 3 cm. The private contractor also measured the radiation at soil depths shallower than 0.3 cm. The table below shows these concentrations for the plutonium isotope ^{239,240}Pu (Bq/kg).

Background Soil Radiation Concentrations

Site	^{239,240}Pu
Z01	0.86
Z02	1.52
Z03	1.62
Z04	1.29
Z05	2.33
Z06	0.96
Z08	1.51
Z09	1.43
Z10	2.47

For one of the original ten sites, no measurement was made at this shallower depth, so the sample size here is only $n = 9$. The sample mean and standard deviation of the data are

$$\bar{X} = 1.55 \qquad \text{and} \qquad S = 0.54\,.$$

a) We want to determine a value *below* which the true (unknown) mean background radiation level μ lies with 95% confidence. Compute a lower 95% one-sided t confidence interval for μ.

b) Give the value of the estimate $\hat{X}_{0.01}$ of the 99th percentile, $x_{0.01}$, of the population of background radiation concentrations (assuming the population is normal). Interpret the value of the estimate – what does it tell us about the population of background radiation concentrations?

c) The background radiation measurements were made for use in establishing soil action levels, or threshold values above which a measured soil radioactivity concentration would trigger a cleanup. One way to determine such a threshold value is to use the data to construct an upper tolerance limit.

Here are the critical values $t_{\alpha,n-1,\sqrt{n}\,z_{1-p}}$ for constructing several upper tolerance limits, all with 95% confidence levels but different coverage percentages.

$$\begin{aligned} t_{0.05,8,\sqrt{9}\times 2.33} &= 12.447 \quad \text{(confidence level 95\% and coverage 99\%)} \\ t_{0.05,8,\sqrt{9}\times 1.64} &= 9.070 \quad \text{(confidence level 95\% and coverage 95\%)} \\ t_{0.05,8,\sqrt{9}\times 1.28} &= 7.354 \quad \text{(confidence level 95\% and coverage 90\%)} \end{aligned}$$

These were obtained using statistical software from the noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $\sqrt{n}\,z_{1-p}$.

Use the critical values to compute and interpret the three associated upper tolerance limits.

6.12 Public involvement in government environmental management decisions is frequently difficult to obtain.

A survey was carried out to determine the reasons for public nonparticipation in decision making processes involving an environmental assessment of a proposed hog slaughtering facility and associated wastewater treatment plant in Brandon, Manitoba, Canada [4]. The proposed facility would slaughter up to 54,000 hogs per week and its effluent treated at the wastewater treatment facility and then discharged into the Assiniboine River.

A sample of $n = 79$ people who did not participate were presented with several possible reasons for not participating, and asked whether each was important (Yes or No). One of the reasons presented was that "The ultimate decisions were foregone," meaning that their involvement would have little impact on final decisions. Of the 79 people surveyed, 51 said this was an important reason for their nonparticipation.

Compute and interpret a 95% one-sample z confidence interval for the true (unknown) population proportion p for whom "The ultimate decisions were foregone" is important in their decision not to participate.

6.13 Farmers use biosolids (sludge) from wastewater treatment plants to fertilize soil. A study was carried out to assess the risk of farmers' exposure to salmonella through the application of biosolids to farmlands in Ohio [11]. In a sample of $n = 92$ biosolids specimens, 22 tested positive for salmonella.

Compute and interpret a 95% z confidence interval for the (unknown) population proportion p of biosolids specimens that would test positive for salmonella.

Bibliography

[1] The national study of chemical residues in lake fish tissue. Technical Report EPA-823-R-09-006, United States Environmental Protection Agency, Sept 2009.

[2] Battelle. Heavy-duty truck activity data. Technical report, Office of Highway Information Management, Office of Technology Applications, Federal Highway Administration, 1999.

[3] Kimberlee B. Beckmen et al. Factors affecting organochlorine contaminant concentrations in milk and blood of northern fur seal (*Callorhinus ursinus*) dams and pups from St. George Island, Alaska. *The Science of the Total Environment*, 231:183–200, 1999.

[4] Alan Diduck and John A. Sinclaire. Public involvement in environmental assessment: The case of the nonparticipant. *Environmental Management*, 29(4):578 – 588, 2002.

[5] Electa Draper. Feds raided Rocky Flats 25 years ago, signaling the end of an era. *The Denver Post*, Jun. 1 2014.

[6] Electa Draper. Former Rocky Flats site stirs concerns for some neighbors. *The Denver Post*, Feb. 9 2014.

[7] Richard O. Gilbert. *Statistical Methods for Environmental Pollution Monitoring*. John Wiley and Sons, 1987.

[8] M. A. Greer, G. Goodman, R. C. Pleus, and S. E. Greer. Health effects assessment for environmental perchlorate contamination: the dose-response for inhibition of thyroidal radioiodine uptake in humans. *Environmental Health Perspectives*, 110:927–937, 2002.

[9] Lee Hsiang Liow, Navjot S. Sodhi, and Thomas Elmqvist. Bee diversity along a disturbance gradient in tropical lowland forests of south-east Asia. *Journal of Applied Ecology*, 38(1):180 – 192, February 2001.

[10] C. H. Nelson and P. J. Lamothe. Heavy metal anomalies in the Tinto and Odiel river and estuary system, Spain. *Estuaries*, 16(3A):496 – 511, 1993.

[11] Abramo C. Ottolenghi and Vincent V. Hamparian. Multiyear study of sludge application to farmland: Prevalence of bacterial enteric pathogens and antibody status of farm families. *Applied and Environmental Microbiology*, 53(5):1118 – 1124, May 1987.

[12] Christopher B. Pepper et al. Oganochlorine pesticides in chorioallantoic membranes of Morelet's crodocile eggs from Belize. *Journal of Wildlife Diseases*, 40(3):493–500, 2004.

[13] G. Perin et al. A five-year study on the heavy-metal pollution of Guanabara Bay sediments (Rio de Janeiro, Brazil) and evaluation of the metal bioavailability by means of geochemical speciation. *Water Resources*, 31(12):3017 – 3028, 1997.

[14] Yue Rong. Statistical methods and pitfalls in environmental data analysis. *Environmental Forensics*, 1:213 – 220, 2000.

[15] John E. Till, George G. Killough, Arthur S. Rood, Jill Weber Aanenson, Kathleen R. Meyer, Helen A. Grogan, and Warren K. Sinclair. Final report, task 5: Independent calculation. Technical Report RAC Report No. 16-RSALOP-RSAL-1999-FINAL, Risk Assessment Corporation, Feb 2000. Report submitted to the Radionuclide Soil Action Level Oversight Panel.

[16] David Ting, Robert A. Howd, Anna M. Fan, and George V. Alexeeff. Development of a health-protective drinking water level for perchlorate. *Environmental Health Perspectives*, 114(6):881–886, June 2006.

[17] Randall S. Wells et al. Integrating life-history and reproductive success data to examine potential relationships with organochlorine compounds for bottlenose dolphins (*Tursiops truncatus*) in Sarasota Bay, Florida. *Science of the Total Environment*, 349:106–119, 2005.