4

Chapter

Statistics for Food Analysis

Andrew P. Neilson (✉) • Sean F. O'Keefe
Department of Food Science and Technology,
Virginia Polytechnic Institute and State University,
Blacksburg, VA, USA
e-mail: andrewn@vt.edu; okeefes@vt.edu

S.S. Nielsen, *Food Analysis Laboratory Manual*, Food Science Text Series,
DOI 10.1007/978-3-319-44127-6_4, © Springer International Publishing 2017

4.1 INTRODUCTION

This chapter is a review of basic statistics and will demonstrate how statistics is used in the context of food analysis. It is meant to be a "survival guide" for reference. Students taking a food analysis course should have already taken an undergraduate statistics course or be taking one concurrently. A foundation for this chapter is Chap. 4, Evaluation of Analytical Data, in the *Food Analysis* textbook. The learning objectives for this chapter are to be able to

1. Calculate a mean, standard deviation, *Z*-score, and *t*-score for a sample dataset.
2. Determine whether a population is significantly different from a given value using a one-sample *t*-test.
3. Calculate a confidence interval for a sample mean using *t*-scores.
4. Determine if two populations are significantly different using a two-sample *t*-test.

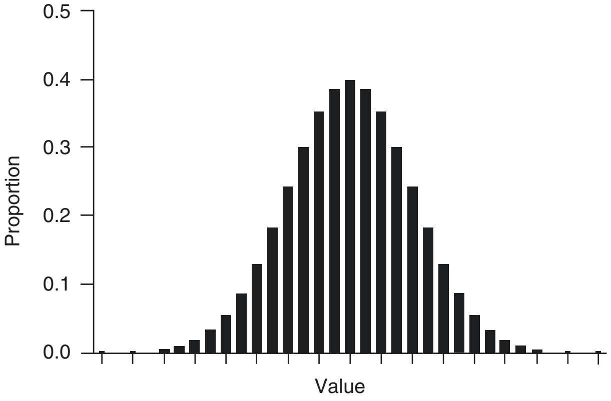
A significant component of a typical food analysis course and the use of food analysis in a professional setting (industry, research, regulatory/government) is evaluation of analytical data. In food analysis laboratory experiments, you first collect data (values), or you are given data to work with (i.e., for laboratory exercises, homework, exams, etc.). Then, you evaluate and manipulate the data, answering questions such as these:

1. Are the values significantly different from a desired value or not?
2. Are two values significantly different from each other or not?

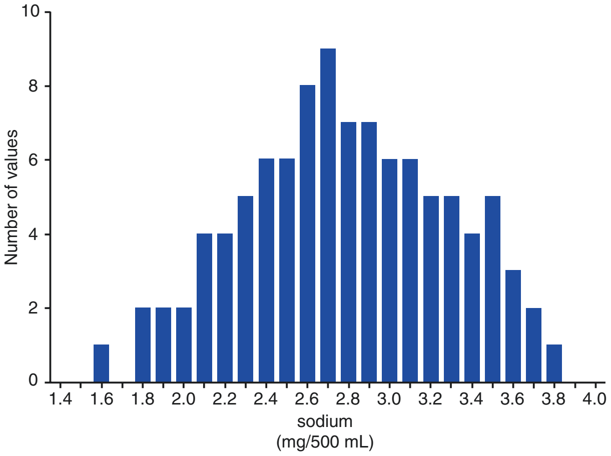
In your food analysis course, you will use statistics to solve problems typically encountered in the food industry, research, and regulatory or government scenarios. Data analysis concepts will be used in this chapter.

4.2 POPULATION DISTRIBUTIONS

The **distribution** of a given value in a **population** refers to how many members of a population have each value of the parameter measured. You can examine distributions by plotting a **histogram**, which is a graph with parameter values on the *x*-axis and number of members of the population having that value of the parameter (i.e., the "frequency" of each value) on the *y*-axis. Suppose there were 100 bottles of water in a population and you knew the sodium content of all of them. The histogram of this population might look like Fig. 4.1. Many populations are "normally distributed" (also known as **normal** population). Normal populations have histograms that look roughly like Fig. 4.2. Note that the *y*-axis of a histogram could be the absolute number of population members with a given value, the proportion (fraction or decimal) of population members with a given value, or the percentage of population members with a given value. Some populations are not normally distributed, but that is a topic beyond the scope of this chapter. For the purposes of this chapter, we will assume a normal distribution. Normal population distributions are defined by two parameters (Fig. 4.3):



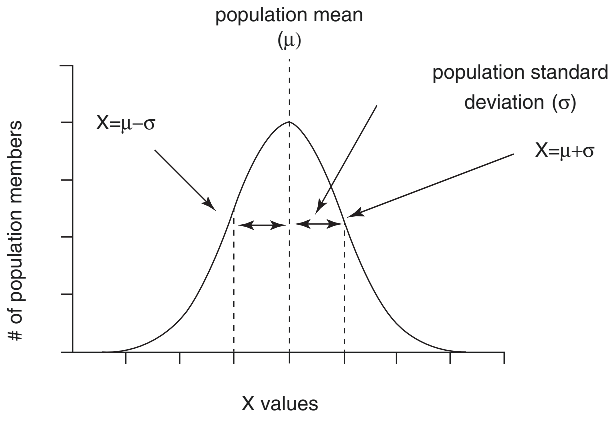
4.1 figure Example normal distribution



4.2 figure Example population histogram

1. Population **mean** (μ): center of the plot
2. Population **standard deviation** (σ): a measure of the spread of the plot

For the example sodium data, you can plot the histogram of the data with the population mean (μ) and population standard deviation (σ). If you define sodium content as (x), you can calculate the mean and standard deviation of x:



4.3 figure Shape of a normal population

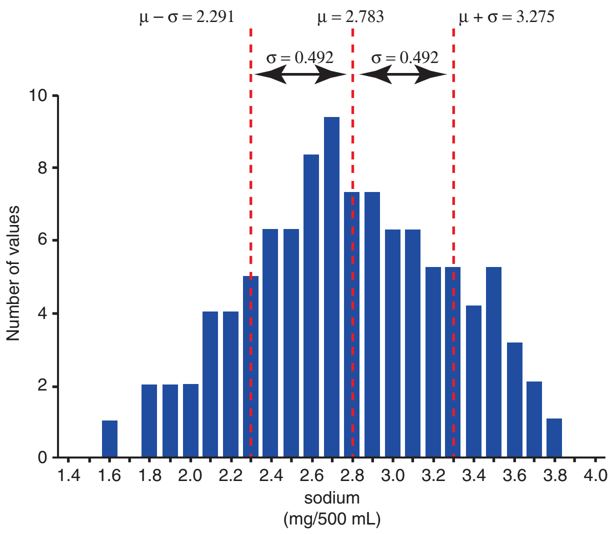
$$\text{Population mean of } x = \mu_x = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_{n-1} + x_n}{n} \quad (4.1)$$

$$\text{Standard deviation of } x = \sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad (4.2)$$

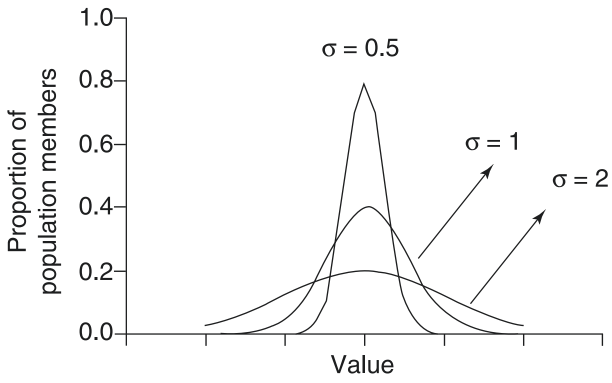
$$\text{Population mean} = \mu_x = 2.783 \text{ standard deviation} = \sigma_x = \sqrt{\frac{\sum (x_i - 2.783)^2}{100}} = 0.492$$

You can use these values to calculate the center and spread of the data, as shown in Fig. 4.4, for example, sodium data. Normal distributions can have different shapes but are still always defined by $N(\mu, \sigma)$. The notation "$N(\mu, \sigma)$" indicates a normally distributed (N) population with center μ and standard deviation σ. The shape of the curve is dictated by the standard deviation. Figure 4.5 shows three populations, each with the same mean but different standard deviations. For a normally distributed population, if you randomly pick (or **sample**) a member of the population, the randomly selected member can have any value in the population. Since most of the population is clustered around the mean, the randomly selected value is most likely to be close to the mean. The further a value is from the mean, the less often it occurs in the population and the less likely it is to be randomly selected from the population. If we know μ and σ, we can predict the probability that any given value will be selected at random from the population, as shown in Fig. 4.6. Normal distributions have

1. 50 % of values $> \mu$ and 50 % $< \mu$
2. 68, 95, and 99.7 % of values within ±1, 2, and 3 standard deviations (σ) of μ



4.4 figure Example population histogram with mean and standard deviation

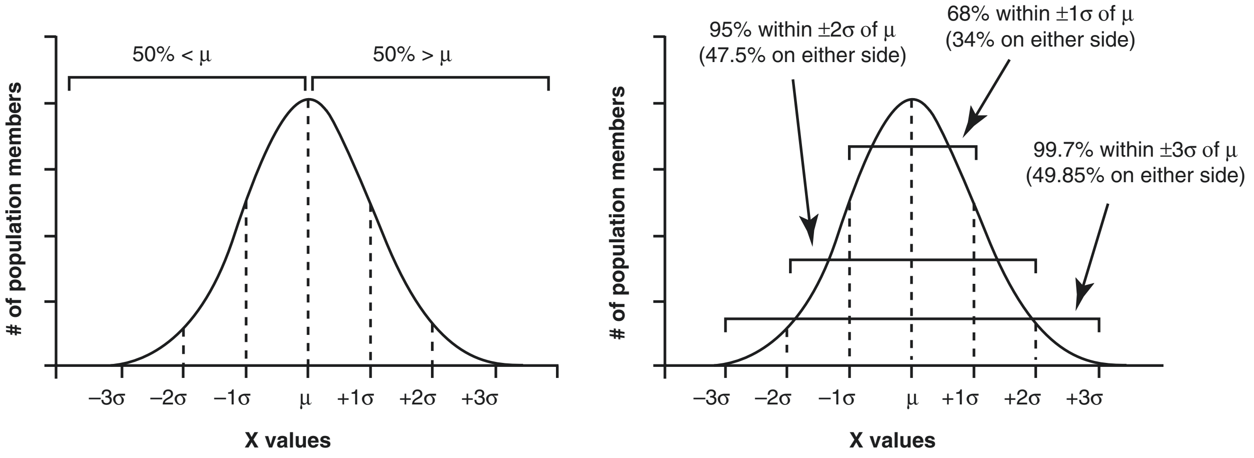


4.5 figure Normal populations with the same mean but distinct standard deviations

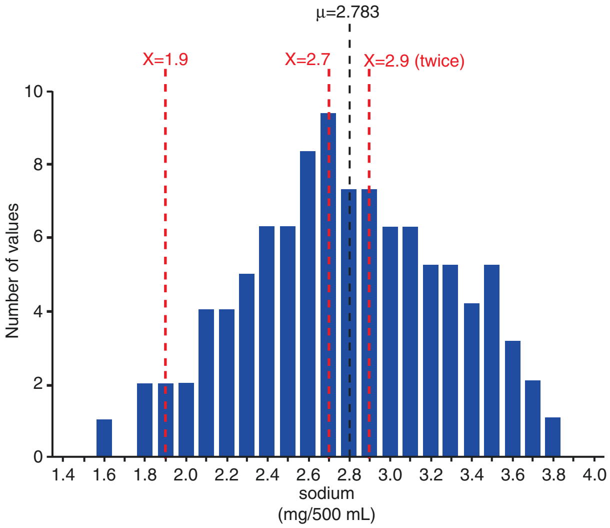
Assume you randomly sample four values from the 100 water bottle population, and 1.9, 2.7, 2.9, and 2.9 mg/500 mL were selected. As seen in Fig. 4.7, three of the randomly chosen values were close to the mean, and one value was relatively far from the mean. Remember, any of the 100 values could have been selected at random. However, the values closest to the mean have the highest probability of being chosen.

4.3 Z-SCORES

Each normal population has a different μ, σ. However, this is inconvenient for statistical calculations. To make normal distributions easier to work with, you can transform, or "standardize," all normally distributed



4.6 figure Densities of normal distributions



4.7 figure Random values from the example distribution

populations by converting the population value (x) to a standardized variable, called Z:

$$z = \frac{x-\mu}{\sigma} \tag{4.3}$$

The variable x is normally distributed with mean μ and standard deviation σ [$x \sim N(\mu, \sigma)$]. The distribution of the **Z-scores** is $Z \sim N(0, 1)$, i.e., "the standard normal distribution" (Fig. 4.8):

1. The center of the Z-distribution is $\mu = 0$.
2. The spread of the Z-distribution is defined by $\sigma = 1$.

Each x value in a population has a corresponding Z-score. For any x value, the corresponding Z-score is the number of standard deviations that X is away from μ. The more standard deviations (i.e., multiples of σ or 1) that x (or Z) is from μ (or 0), the less likely that value is to be randomly selected from the population. You can calculate a Z-score for any x value.

Example C1 For the sodium data, calculate the Z-scores for a hypothetical x value close to the mean (2.45) and one that is far from the mean (3.8).

$$x = 2.45,\ Z = \frac{x-\mu}{\sigma} = \frac{2.45-2.783}{0.492} = \frac{-0.333}{0.492} = -0.677$$

$$x = 3.8,\ Z = \frac{x-\mu}{\sigma} = \frac{3.8-2.783}{0.492} = \frac{1.017}{0.492} = 2.067$$

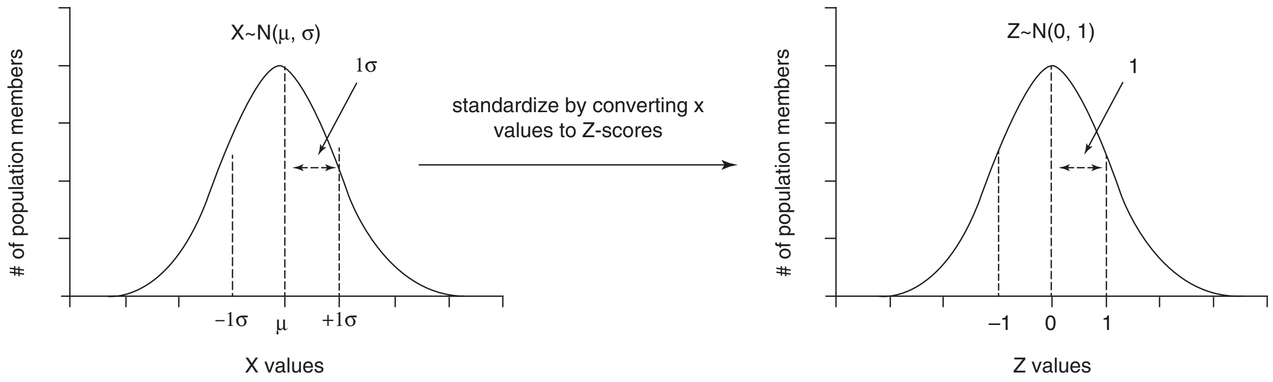
Therefore, for $X \sim N(2.783, 0.492)$, x values of 2.45 and 3.8 are 0.68 standard deviations below and 2.07 standard deviations above the mean, respectively.

The same distributions that apply to $X \sim N(\mu, \sigma)$ also apply to the $Z \sim N(1, 0)$:

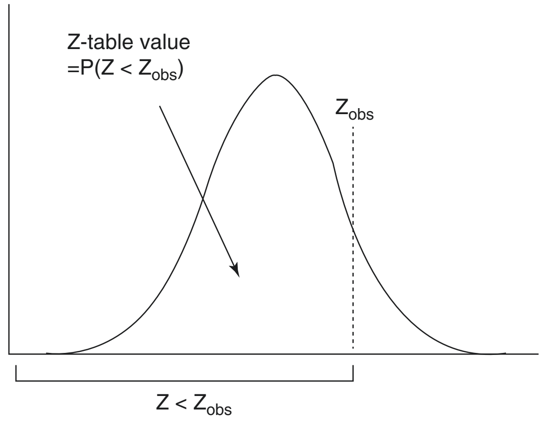
1. 0% of Z values are > 0 and 50% are < 0.
2. 68, 95, and 99.7% of Z values are within ±1, 2, and 3 of 0.

Some other important properties of normal populations:

1. The population curve = 100% of population values
2. The area under the population curve = 100% (percent) or 1 (fraction)
3. The % area under the distribution curve between any two points = probability of randomly selecting



4.8 figure Transformation of *x* values to *Z*-scores



4.9 figure Meaning of *Z*-table values

a value in that range from the population (see below)

Based on these properties of normal populations, statisticians have developed a "Z-table" (www.normaltable.com), which contains the **probability** (P) of getting a *Z*-score smaller than the observed *Z*-score if $Z \sim N(1, 0)$ (Fig. 4.9). To use the *Z*-table:

1. Find your observed *x* value.
2. Calculate a *Z*-score (Z_{obs}) from the *x* value.
3. Find the corresponding table value $= P(Z < Z_{\text{obs}})$ if $Z \sim N(1, 0) = P(x < x_{\text{obs}})$ if $x \sim N(\mu, \sigma)$.

For ease of use, the *Z*-table is organized with the first two digits of *Z* on the left and the third digit (second decimal place: 0.00, 0.01, 0.02, etc.) across the top. Therefore, to find $P(Z < 1.08)$, you would find the row with 1.0 on the left, and then go across that row to the 0.08 column. The steps listed above can be applied to analysis of the sodium data.

Example C2 Calculate the probability of observing sodium values of *less* than 2.783 mg/500 mL (the mean) and *less* than 3.5 mg/500 mL.

Calculate Z-scores for both:

$$\text{When } x = \mu = 2.783,\ z = \frac{x - \mu}{\sigma} = \frac{2.783 - 2.783}{0.492} = \frac{0}{0.492} = 0$$

$$\text{When } x = 3.5,\ z = \frac{x - \mu}{\sigma} = \frac{3.5 - 2.783}{0.492} = \frac{0.717}{0.492} = 1.4$$

Find table values for those *Z*-scores:

$P(Z < 0) = \text{table value for } Z = 0 \rightarrow 0.50\ (50\,\%)$
$P(Z < 1.46) = \text{table value for } Z = 1.46 \rightarrow 0.9279\ (92.79\,\%)$

For the sodium population (normally distributed, $\mu = 2.783$, $\sigma = 0.492$), the probability of observing sodium values < 2.783 and 3.5 is 50 % and 92.79 %, respectively.

Standardization to *Z*-scores eliminates the need for a different *Z*-score table for each population. Although the *Z*-table contains *P* values of getting a *Z*-score smaller than the observed *Z*-score, you can calculate the *P* value of getting a *Z*-score larger than the observed *Z*-score. The sum of the area = 1 (100 %). Thus, the *P* value of getting a *Z*-score smaller than the observed *Z*-score plus the *P* value of getting a *Z*-score larger than the observed *Z*-score equals 100 %:

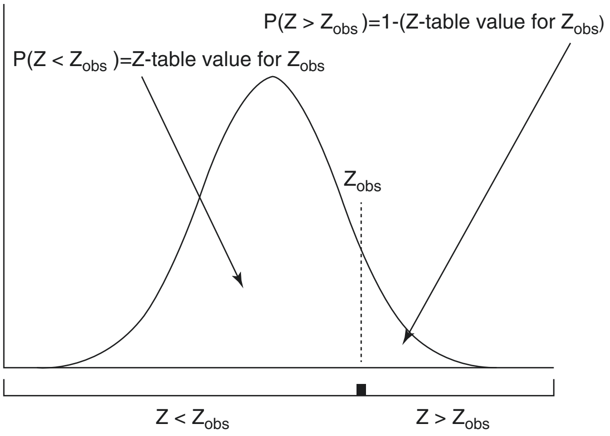
$$P(Z > Z_{\text{obs}}) + P(Z < Z_{\text{obs}}) = 1(100\%) \quad (4.4)$$

Therefore:

$$P(Z > Z_{\text{obs}}) = 1 - P(Z < Z_{\text{obs}}) \quad (4.5)$$

The full expression then becomes:

$$P(Z > Z_{\text{obs}}) + P(Z < Z_{\text{obs}}) + P(Z = Z_{\text{obs}}) = 1(100\%) \quad (4.6)$$


P(Z > Z$_{obs}$)=1-(Z-table value for Z$_{obs}$)
P(Z < Z$_{obs}$)=Z-table value for Z$_{obs}$
Z$_{obs}$
Z < Z$_{obs}$
Z > Z$_{obs}$

4.10 figure The probability of observing Z smaller than Z_{obs}

This equation covers all possible Z-scores. Thus, the P value of getting a Z-score larger than the observed Z-score is simply 1− P value of getting a Z-score larger than the observed Z-score. Therefore, to find the P value of getting a Z-score larger than the observed Z-score (Fig. 4.10):

1. From the x value, calculate a Z-score (Z_{obs}).
2. From Z_{obs}, find table value = $P(Z < Z_{\text{obs}})$.
3. From $P(Z < Z_{\text{obs}})$, calculate $P(Z > Z_{\text{obs}}) = 1 -$ table value for $Z_{\text{obs}} = P(X > X_{\text{obs}})$ if $X \sim N(\mu, \sigma)$.

Example C3 Calculate the probability of observing sodium values of greater than 2.9 mg/500 mL, again following the steps listed previously.

Calculate the Z-score:

$$\text{When } x = 2.9,\ z = \frac{x-\mu}{\sigma} = \frac{2.9-2.783}{0.492} = \frac{0.117}{0.492} = 0.24$$

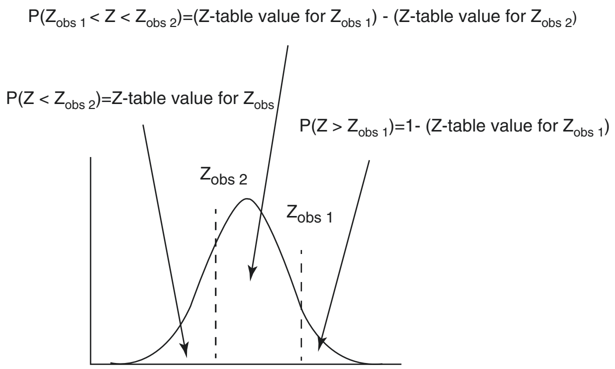
Find the table value for the Z-score: $P(Z < 0.24)$ = table value for $Z = 0.24 \rightarrow 0.5948$ (59.48 %).

Convert to $P(Z > Z_{\text{obs}})$ by subtracting from 1:

$$\left(Z > Z_{\text{obs}}\right) = 1 - P\left(Z < Z_{\text{obs}}\right) = 1 - 0.5948$$
$$= 0.4052 = 40.52\%$$

So, for the sodium population, normally distributed with $\mu = 2.783$ and $\sigma = 0.492$, the probability of observing a sodium value > 2.9 is 40.52 %.

How do you calculate the probability of getting an x value (and corresponding Z-score) between two x values (with corresponding Z-scores)? The probability that a Z-score lies between two given Z-scores is simply the difference between the two table values for $P(Z < Z_{\text{obs}})$ (i.e., subtract the smaller table value from larger value) (Fig. 4.11):


P(Z$_{obs\ 1}$ < Z < Z$_{obs\ 2}$)=(Z-table value for Z$_{obs\ 1}$) - (Z-table value for Z$_{obs\ 2}$)
P(Z < Z$_{obs\ 2}$)=Z-table value for Z$_{obs}$
P(Z > Z$_{obs\ 1}$)=1- (Z-table value for Z$_{obs\ 1}$)
Z$_{obs\ 2}$
Z$_{obs\ 1}$

4.11 figure The probability of observing Z between two Z_{obs} values

$$P\left(Z_1 < Z < Z_2\right) = P\left(Z < Z_1\right) - P\left(Z < Z_2\right) \tag{4.7}$$

Example C4 Calculate the probability of observing sodium values between 1.9 and 3.1 mg/500 mL.

Calculate the Z-scores:

$$\text{When } x = 1.9,\ z = \frac{x-m}{s} = \frac{1.9-2.783}{0.492} = \frac{-0.883}{0.492} = -1.79$$

$$\text{When } x = 3.1,\ z = \frac{x-\mu}{\sigma} = \frac{3.1-2.783}{0.492} = \frac{0.317}{0.492} = 0.64$$

Find the table values for the Z-scores:

$P(Z < 0.64)$ = table value for $Z = 0.64 \rightarrow 0.7389$
$P(Z < 0.64)$ = table value for $Z = -1.79 \rightarrow 0.0376$

Calculate the area between the two Z-scores:

$$P\left(Z_1 < Z < Z_2\right) = P\left(Z < Z_2\right) - P\left(Z < Z_1\right)$$
$$= 0.7389 - 0.0376 = 0.7013 = 70.13\%$$

So, for the sodium population which is normally distributed with $\mu = 2.783$ and $\sigma = 0.492$, the probability of observing a sodium values between 1.9 and 3.1 mg/500 mL is 70.13 %.

4.4 SAMPLE DISTRIBUTIONS

You almost never know the actual population values. This is due to several factors. First, the population is often too large to sample all members. Second, sampling often means that the product is no longer available for use. You typically sample a few members (n) and estimate population parameters from sample parameters. Population parameters are as follows:

$$\text{Population } (x) \text{ mean} = \mu_x \tag{4.8}$$

$$\text{Population}\,(x)\,\text{standard deviation} = \sigma_x \qquad (4.9)$$

Sample parameters are as follows:

$$\text{Sample size} = n \qquad (4.10)$$

$$\text{Mean of } \bar{x} \text{ distribution} = \mu_{\bar{x}} \approx \mu_x \qquad (4.11)$$

$$\text{SD of } \bar{x} \text{ distribution} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \qquad (4.12)$$

Sample parameters are related to population parameters:

1. Sample mean ≈ population mean
2. Sample standard deviation < population standard deviation

This implies that sampling can be used to estimate population parameters.

Example D1 Determine the mean and standard deviation of the sample mean distribution if you sample five or ten members of the population.

For both means:

$$\mu_{\bar{x}} \approx \mu_x = 2.783$$

The standard deviations of the sample means:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \quad \text{for } n = 5, \quad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.492}{\sqrt{5}} = 0.220$$

$$\text{for } n = 10, \quad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.492}{\sqrt{10}} = 0.156$$

You can see that:

1. The sample mean standard deviation < population standard deviation.
2. Larger n size decreases sample mean standard deviation.

Sampling from a normal population results in a normally distributed population of all possible sample means ($\bar{x}$). The population distribution refers to the individual population values of x. The sample mean distribution refers to the values of sample means ($\bar{x}$) generated by sampling x. Because the sample mean is an *average* of the population members:

1. The average sample mean is the same as the average population member.
2. The sample mean is less widely distributed than the population (outliers in the population get diluted by taking a sample *mean*).

The larger the sample size (n), the tighter (SD) the sample mean distribution becomes:

$$\text{SD of } \bar{x} \text{ distribution} = s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

$$\text{Therefore as } n\uparrow,\ s_{\bar{x}}\downarrow$$

Everything you learned about Z-scores for population values applies to sample mean values. You can transform the sample mean distribution to the Z-distribution:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \rightarrow Z \sim N(0,1) \qquad (4.13)$$

You can transform an observed sample mean into the corresponding Z-score:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (4.14)$$

You can use the Z-table to calculate the probability of getting a sample mean smaller or greater than an observed mean or between two observed means (same math as before).

Example D2 How likely would it be to have a sample mean of greater than 3.0 mg/500 mL if you sampled four bottles?

Transform the $\bar{x}$ into a Z-score:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3 - 2.783}{\frac{0.492}{\sqrt{4}}} = \frac{0.217}{0.246} = 0.88$$

Find the table value for the Z-score (area to the left of Z_{obs}):

$P(Z < 0.88)$ = table value for $Z = 0.88$: 0.8106

Find $P(Z > Z_{\text{obs}})$:

$$P(Z>0.88) = 1 - P(Z<0.88) = 1 - 0.8106 = 0.1894$$
$$= 18.94\%$$

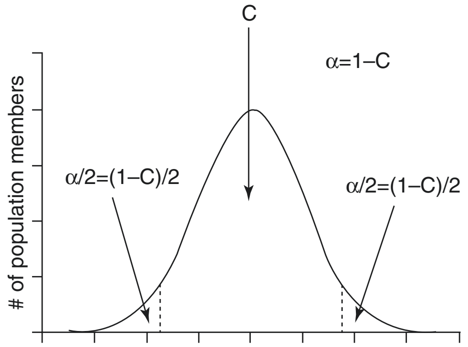
So, for the sodium data, the probability of sampling $n = 4$ bottles and getting a mean sodium content of >3.0 mg/500 mL is 18.9%.

4.5 CONFIDENCE INTERVALS

How confident are you that the observed sample mean is close to the actual population mean? And how close? You can calculate the probability of getting a sample mean as extreme as the one you observe, *if* you assume the population is normally distributed and the sample mean equals the population mean.

Procedure

1. Calculate Z-score for $\bar{x}$
2. Find area between $+Z$ and $-Z = P(Z < |Z_{\text{obs}}|) = C$



4.12 figure Probability of getting a sample mean as extreme as the observed mean

The area (C) is the probability of getting a sample mean as extreme as the one you observed (Fig. 4.12). You can calculate the probability of getting a sample mean more extreme than the one you observed (α) on either end, or getting a sample mean more extreme on a single end ($\alpha/2$).

$$C = P(-Z_{\text{obs}} < Z < +Z_{\text{obs}}) \\ = P(Z < +Z_{\text{obs}}) - P(Z < -Z_{\text{obs}}) \quad (4.15)$$

$$\alpha = 1 - C \quad (4.16)$$

$$\frac{\alpha}{2} = \frac{1-C}{2} \quad (4.17)$$

Example E1 Suppose you sample $n=3$ bottles and get a mean of 2.4 mg/500 mL. Calculate the probabilities of getting a sample mean this extreme, a sample mean smaller than 1.8 mg/500 mL, and a sample mean larger than 1.8 mg/500 mL.

Calculate the Z-score for the observed population mean:

$$Z_{\text{obs}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.4-2.783}{\frac{0.492}{\sqrt{3}}} = \frac{-0.383}{0.284} = -1.35$$

Find the $+Z$ and $-Z$ values:

$Z = -1.35$, so you are interested in the range between $Z = 1.35$ and -1.35

Find $C = P(-Z_{\text{obs}} < Z < +Z_{\text{obs}})$ from the table:

$$C = P(-Z_{\text{obs}} < Z < +Z_{\text{obs}}) \\ = P(Z < +Z_{\text{obs}}) - P(Z < -Z_{\text{obs}})$$

$$C = P(Z < +Z_{\text{obs}}) - P(Z < -Z_{\text{obs}}) \\ = \text{table value}(+Z_{\text{obs}}) - \text{table value}(-Z_{\text{obs}})$$

$$C = 0.9115 - 0.0885 = 0.823 = 82.3\%$$

For the sodium population which is $N(2.783, 0.492)$, there is an 82.3% chance that you would observe a sample mean within 1.35 standard deviations ($-1.35 < Z < +1.35$) on either side of the mean if you sampled three bottles from the population.

Now, find α:

$$\propto = 1 - C = 1 - 0.823 = 0.177 = 17.7\%$$

For the sodium population which is $N(2.783, 0.492)$, there is a 17.7% chance that you would observe a sample mean > 1.35 standard deviations ($Z < -1.35$ or $Z > +1.35$, i.e., $Z > |1.35|$) on either side of the mean if you sampled three bottles from the population.

Now, find $\alpha/2$:

$$\frac{\propto}{2} = \frac{1-C}{2} = \frac{1-0.823}{2} = 0.0885 = 8.85\%$$

For the sodium population which is $N(2.783, 0.492)$, there is an 8.85% chance that you would observe a sample mean > 1.35 standard deviations ($Z > +1.35$) above the mean if you sampled three bottles from the population.

Usually you do not know the actual population mean (μ) or standard deviation (σ). You sample and calculate sample mean (observed $\bar{x}$) and sample standard deviation (sample SD = $\sigma\bar{x}$) to estimate these parameters (observed $\bar{x} \approx \mu\bar{x} = \mu x$ and SD $\approx \sigma\bar{x} = \sigma x/\sqrt{n}$). It is highly unlikely that the observed sample mean is *exactly* equal to the population mean. You can calculate a range (i.e., an "interval") around our observed sample that is likely to cover the true population mean with a given level of statistical **confidence**. The **confidence interval** (CI) gives a margin of error around our sample mean:

$$\text{CI} : \bar{x} \pm \text{margin of error} \quad (4.18)$$

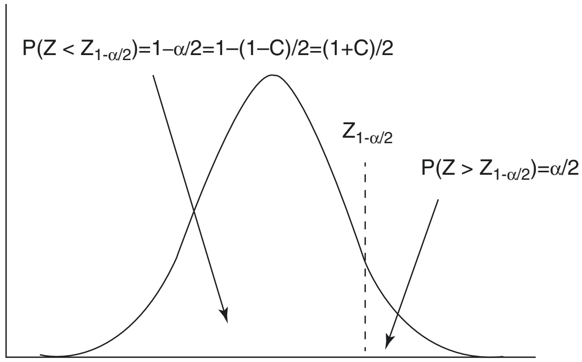
Suppose that you want to generate a CI that is within a specified number of standard deviations (i.e., number of Z-scores) from the mean. Using the Z-score formula:

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow Z\left(\frac{\sigma}{\sqrt{n}}\right) = \bar{x} - \mu$$

Since it could be on either side of the mean, change to:

$$\pm Z\left(\frac{\sigma}{\sqrt{n}}\right) = \bar{x} - \mu$$

Now, you have to choose how wide the sample margin of error should be. You do this by selecting the number of standard deviations you desire to be within on each side, which gives a maximum absolute value of Z. You could decide that you want our CI to cover the true mean with



4.13 figure Relationship between C, $\alpha/2$, and Z-table values

a given percent confidence (C). This means that the probability that the CI does *not* cover the true mean would be $1-C$ (which was defined previously as α):

$$\propto = 1-C \rightarrow C = 1-\propto$$

Then, you would find the "critical" Z-scores where:

1. $P\ (Z_{\text{crit}}<Z<Z_{\text{crit}})=C$ (the desired confidence level)
2. $P(Z>\left|Z_{\text{crit}}\right|)=\alpha=1-C$ (the probability that the CI does *not* cover the true mean)

The easiest way to do this is to find $\alpha/2$ and then find the corresponding Z-score:

First, determine $\alpha/2$ from the desired confidence (C):

$$\frac{\propto}{2}=\frac{1-C}{2}=\frac{1-0.823}{2}=0.0885=8.85\%$$

$\alpha/2$ is an area to the *right* of the Z_{crit} needed to determine the area to the left of Z_{crit} (aka $Z_{1-\alpha/2}$), i.e., $P(Z<Z_{1-\alpha/2})$ (Fig. 4.13):

$$P(Z<Z_{\text{crit}})=P\left(Z<Z_{1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}=1-\frac{1-C}{2}=\frac{1+C}{2}$$

$P(Z<Z_{1-\alpha/2})$ is the table value for $Z_{1-\alpha/2}$. Use this to find $Z_{1-\alpha/2}$.

Now you can calculate the "margin of error" for the CI:

$$\text{Margin of error}:\ \pm Z_{1-\frac{\alpha}{2}}\, x \frac{\sigma}{\sqrt{n}} \qquad (4.19)$$

The confidence interval, with confidence level C, is:

$$\text{CI}:\bar{x} \pm \text{margin of error} \rightarrow \text{CI}:\bar{x}\pm Z_{1-\frac{\alpha}{2}}\, x\left(\frac{\sigma}{\sqrt{n}}\right) \qquad (4.20)$$

Do not get flustered by the statistics details. You just need to be able to do the following:

1. Determine the desired C.
2. Calculate $\alpha/2$.
3. Calculate 1_$\alpha/2$.
4. Use this value to find $Z_{1-\alpha/2}$.
5. Calculate the CI.
6. Get from $C\rightarrow\alpha/2\rightarrow 1-\alpha/2\rightarrow Z_{1-\alpha/2}$.
7. Plug these values into the formula and calculate the CI.

Typically, you use 90, 95, or 99 % confidence in food analysis. We can choose any desired level of confidence that we want. However

Example E2 Suppose you sample $n=5$ cans of soda, and the mean caffeine content is 150 mg/can. You know that the population has a standard deviation (σ) of 15 mg caffeine/can. What is the 95% confidence interval for the mean caffeine content per can?

Calculate $\alpha/2$: $\frac{\alpha}{2}=\frac{1-C}{2}=\frac{1-0.95}{2}=\frac{0.05}{2}=0.025$

Calculate $1-\alpha/2$:

$$P\left(Z<Z_{1-\frac{\alpha}{2}}\right)=1-\frac{\alpha}{2}=1-0.025=0.975$$

$1-\alpha/2=P(Z<Z_{1-\alpha/2})=$ table value for $Z_{1-\alpha/2}$… use this value to find $Z_{1-\alpha/2}$:

$$\text{Table value}=0.975\ \rightarrow Z_{1-\frac{\alpha}{2}}=1.96$$

Calculate the CI: $\bar{x}\pm Z_{1-\frac{\alpha}{2}}\, x\frac{\sigma}{\sqrt{n}}$

$$150\,\text{mg / can} \pm 1.96\, x \frac{15\text{mg / can}}{\sqrt{5}} \rightarrow$$

$$150\text{mg / can} \pm 13.1\text{mg / can}$$

$$\text{Lower and upper limits} = 150\text{mg/can} \pm 13.1\text{mg/can}$$
$$= 136.9 \text{ and } 163.1\text{ mg / can}$$

Therefore, you estimate that the true population mean is somewhere between 136.9 and 163.1 mg caffeine/can with 95 % confidence.

1. The more confidence you want, the broader the interval gets.
2. The less confidence you are willing to accept, the narrower the range gets.

Example E3 For the soda data, calculate the 90 % CI of the sample mean.

Calculate $\alpha/2$: $\frac{\alpha}{2}=\frac{1-C}{2}=\frac{1-0.90}{2}=\frac{0.10}{2}=0.05$

Calculate $1-\alpha/2$: $1-\frac{\alpha}{2}=1-0.05=0.95$

$1-\alpha/2 = P(Z<Z_{1-\alpha/2})$ = table value for $Z_{1-\alpha/2}$.

Use this value to find $Z_{1-\alpha/2}$:

Table value = 0.95 → $Z_{1-\frac{\alpha}{2}} = 1.645$

Calculate the CI: $\bar{x} \pm Z_{1-\frac{\alpha}{2}}\, x \frac{SD}{\sqrt{n}}$

$$150\text{ mg / can} \pm 1.645\, x \frac{15\text{ mg / can}}{\sqrt{5}} \rightarrow$$

$$150\text{ mg / can} \pm 11.0\text{ mg / can}$$

$$\text{Lower and upper limits} = 150\text{ mg / can} \pm 11.0\text{ mg / can} = 139.0\text{ and } 161.0\text{ mg / can}$$

Therefore, you estimate that the true population mean is between 139.0 and 161.0 mg/can with 90 % confidence. You see that our interval is tighter, but you are therefore less confident that it contains the true population mean.

4.6 *t*-SCORES

Usually you do not have large enough n and do not know σ, which are requirements for using the Z-distribution. Therefore, statisticians developed the ***t*-score**:

$$t = \frac{\bar{x}-\mu}{\frac{\text{SD}}{\sqrt{n}}} \tag{4.21}$$

Recall that the sample standard deviation (SD) is calculated as follows:

$$\text{SD}_n = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}} \; or \; \text{SD}_{n-1} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} \quad \text{(4.22 and 4.23)}$$

You should use the "$n-1$" formula for SD when $n<25$–30. The t-score is similar to the Z-score. However, there are some critical differences. The t-score:

1. Uses the SD of the sample (known) instead of the population σ (usually unknown)
2. Is more conservative than Z-distribution (for a given mean and SD, t is larger than Z)
3. Is typically only presented for a few selected values of $\alpha/2$ (typically 0.1, 0.05, 0.025, 0.01, and 0.005, which correspond to the commonly used confidence values of 80, 90, 95, 98, and 99 %, respectively)
4. Is only listed in positive values, not negative values

Typically, you should use the t-distribution (and sample SD) for $n<25$–30, and the Z-distribution and sample SD for $n>25$–30. The t-distribution is presented in a table (www.normaltable.com), similar to Z. One major difference between t and Z is that the t is based on a value called **degrees of freedom (df)**. This value is based on the sample size:

$$\text{df} = n-1 \tag{4.24}$$

Therefore, for each sample size value (n), there is a unique t-score. Note that the t-table is divided into columns for each $\alpha/2$ values and, within those columns, for each df value. Also, note that the t-table gives the area to the *right* of t (Fig. 4.14), whereas the Z-table gives the area to the *left* of Z. Although this is confusing at first, it actually makes t easier to use for CIs than Z. First, determine $\alpha/2$ from the desired confidence (C):

$$\frac{\alpha}{2} = \frac{1-C}{2}$$

This is the table value for t. From the table value, find $t_{\alpha/2}$, df = n – 1. Then, use this value to calculate the CI:

$$\text{CI}: \bar{x} \pm t_{\frac{\alpha}{2},\text{df}=n-1}\, x \frac{\text{SD}}{\sqrt{n}} \tag{4.25}$$

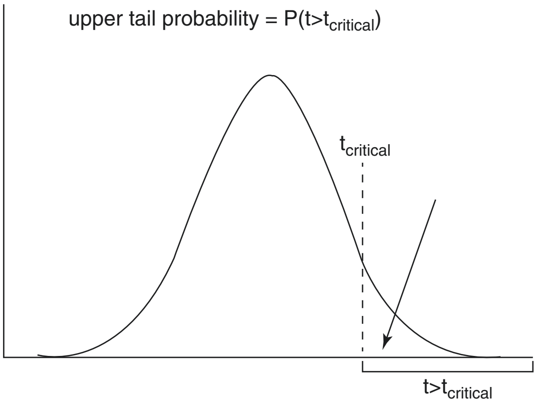
In a food analysis course, t-scores (as opposed to Z-scores) are used almost exclusively, and they have practical use in the food industry. Because t is more conservative for Z, the same confidence level will give a wider interval if calculated using t.

Example F1 Suppose that you sample $n=7$ energy bars and obtain a mean total carbohydrate content of 47.2 % with a sample standard deviation of 3.22 %. Calculate the 99 % confidence interval.

Use the t-distribution because n is relatively small.

Determine $\alpha/2$: $\frac{\alpha}{2} = \frac{1-C}{2} = \frac{1-0.99}{2} = \frac{0.01}{2} = 0.005$

Therefore, you would look in the 0.005 column on the t-table.



4.14 figure Interpretation of t-table values

Then, find $t_{\alpha/2}$, df = n – 1: df = n–1=6–1=5

Therefore, within the 0.005 column on the t-table, you will look at the row for df = 5:

$$\frac{\alpha}{2} = 0.005 \text{ and df} = 5 \ \rightarrow t_{0.005,\ \text{df}=5} = 4.032$$

Calculate the CI:

$$\text{CI}: \bar{x} \pm t_{\frac{a}{2},\text{df}=n-1}\ x\frac{\text{SD}}{\sqrt{n}} \rightarrow 47.2\% \pm 4.032\,x\frac{3.22\%}{\sqrt{7}}$$
$$\rightarrow 47.2\% \pm 4.91$$

$$\text{CI}: \bar{x}\ \pm\ t_{\frac{a}{2},\text{df}=n-1}\ x\frac{\text{SD}}{\sqrt{n}} \rightarrow 47.2\% \pm 4.032\,x\frac{3.22\%}{\sqrt{7}}$$
$$\rightarrow 47.2\% \pm 4.91$$

$$\text{Upper limit} = 47.2\% + 4.91\% = 52.1\%$$
$$\text{Lower limit} = 47.2\% - 4.91\% = 42.3\%$$

Therefore, you estimate that the true population mean is between 42.3 and 52.5% with 99% confidence.

The take-home message is that a sample mean is a **single point estimate** that is probably not exactly correct. Therefore, a single point estimate is almost useless in food analysis because it is almost always incorrect. So, why even sample and test if it is almost always wrong?

Example F2 Suppose your company's specification says that the cheese powder to be used in your products must contain 17.0% protein by weight. You measure n = 10 bags of the cheese powder and calculate a mean of 16.8% by weight and a SD of 0.3%. The observed sample mean does not equal the specification value (and it is unlikely to ever fall exactly on 17.0%). Do you reject the lot because it is not exactly 17.0%? Do you keep sampling until the observed mean is 17.0%? However, suppose your company specifies that the raw ingredient must meet the specification, from n = 10 sampled units, using a 95% CI.

This CI is:

$$\bar{x}\ \pm\ t_{\frac{a}{2},\ \text{df}=n-1}\ x\frac{\text{SD}}{\sqrt{n}} \quad \rightarrow \quad 16.8\% \pm\ t_{0.025,\text{df}=4}\ x\frac{0.3\%}{\sqrt{10}}$$
$$\rightarrow \quad 16.8\% \pm 2.776\,x\frac{0.3\%}{\sqrt{10}}$$

$$\text{CI}: 16.8\% \pm 0.263\% \quad \text{or} \quad 16.537\% - 17.063\%$$

Therefore, although the point estimate does not equal the specification value, you are 95% confident that the true mean lies somewhere between 16.537% and 17.063%, which contains the specification value (17.05). Therefore, based on this, you would accept the raw material.

The Example F2 demonstrates the utility of a CI. It is a range of values that have a specified likelihood of containing the true population mean, whereas the point estimate has a very small likelihood of being the true mean. You may wonder what use a CI is if it is a range of values. However, you can often get very "tight" CIs (i.e., a narrow range that is useful) by random sampling of a population and choosing an appropriate level of confidence. As discussed in the textbook Chap 4, you can use the concept of a CI to determine how large the n size should be to obtain a CI of a specific width. In addition to CIs, another technique that is often used to express the observed sample mean as a range of values is to calculate the sample mean ± SD or the mean ± SEM. The SEM, or **standard error of the mean**, is simply:

$$\text{SEM} = \frac{\text{SD}}{\sqrt{n}} \qquad (4.26)$$

You will notice that the SEM term appears in the t-test and CI formulae Eqs. 4.21 and 4.25. When you express the observed sample mean ± SD or SEM, you are not giving a probability-based estimate (that would require the t-term multiplier used for t-tests and CIs) but rather simply presenting the data as an estimate of the mean based on the sample variability. The range calculated based on SEM will always be narrower, and hence less conservative, than the t-test and CI. This may not be true for the range calculated based on SD, depending on the sample size. For the cheese powder data (Example F2), you could express the estimate as:

$$\text{Mean} \pm \text{SD} = 16.8\% \pm 0.3\% = 16.5 - 17.1\%$$

$$\text{Mean} \pm \text{SEM} = 16.8\% \pm \frac{0.3\%}{\sqrt{10}} = 16.8\% \pm 0.0949$$
$$= 16.705 - 16.895\%$$

Therefore, based on these ranges, we would accept the raw material based on mean ± SD but reject it based on mean ± SEM. Therefore, it is very important to know how to use your analytical data once it is obtained. What are the specific rules you will use to evaluate data and make decisions? Typically, these are specified by company lab quality assurance/quality control manuals, industry standards, or regulatory agencies.

4.7 *t*-TESTS

Sometimes you need to determine if the observed sample mean indicates that the population is the same, or different from, a chosen value. You can use a procedure called a t-test if you have the sample mean ($\bar{x}$), standard deviation (SD), sample sizes (n), and a desired population mean (μ) that you want to compare the same mean to. First, determine $\alpha/2$ from the desired confidence (C) and the df value, as before:

$$\frac{\alpha}{2} = \frac{1-C}{2} \text{ and df} = n-1$$

From these values, find the table value $t_{\alpha/2,\,\mathrm{df}=\mathrm{n}-1}$. This is the "critical value" for a t-test, labeled as "t_{critical}." Then, calculate the observed t-score ("t_{obs}") for the sample mean, SD and n:

$$t = \frac{\bar{x} - \mu}{\frac{\mathrm{SD}}{\sqrt{n}}}$$

Then, the t_{critical} and t_{obs} values are compared.

$$\mathit{If}\ |t_{\mathrm{obs}}| > t_{\frac{\alpha}{2},\mathrm{df}=n-1} \rightarrow \text{sample mean is significantly different from } \mu \text{ with confidence}(C)$$

$$\mathit{If}\ |t_{\mathrm{obs}}| < t_{\frac{\alpha}{2},\mathrm{df}=n-1} \rightarrow \text{sample mean is not significantly different from } \mu \text{ with confidence}(C)$$

A few notes are worth mentioning:

1. The more confidence you want that you have the correct answer, the larger t_{critcal} becomes, and the larger t_{obs} must be to provide evidence that the sample mean is significantly different from μ.
2. The absolute value of t_{obs} is compared vs. t_{critical}, as the table only lists positive t-scores.
3. You select μ to compare to the sample mean. You typically do not know the population mean, but you may have a "target value" that you are trying to reach or think the population should have, so you use that as μ.

Example G1 Suppose that your company makes a multivitamin with a label value of 35 mg vitamin E/capsule. Per FDA requirements, the label value needs to be within a certain amount of the actual value. You sample eight capsules and measure the vitamin E content. The observed sample mean is 31.7 mg/capsule, and the sample standard deviation is 3.1 mg/capsule. Can you say that the sample mean includes the label value with 99% confidence?

Determine $\alpha/2$ and df:

$$\frac{\alpha}{2} = \frac{1-C}{2} = \frac{1-0.99}{2} = \frac{0.01}{2} = 0.005 \quad \text{and}$$
$$\mathrm{df} = n - 1 = 8 - 1 = 7$$

From these values, find the t_{critical} value ($t_{\alpha/2,\,\mathrm{df}=n-1}$): $t_{0.005,\mathrm{df}=7} = 3.499$

Calculate t_{obs} from the sample data:

$$t_{\mathrm{obs}} = \frac{\bar{x}-\mu}{\frac{\mathrm{SD}}{\sqrt{n}}} = \frac{31.7-35}{\frac{3.1}{\sqrt{8}}} = -3.01 \quad \text{and} \quad |t_{\mathrm{obs}}| = 3.01$$

Since $|t_{\mathrm{obs}}|$ (3.01) < $t\alpha_{/2,\,\mathrm{df}=n-1}$ (3.499), you can say the sample mean is not significantly different from the label value (35 mg/capsule) with 99% confidence. If $|t_{\mathrm{obs}}|$ had been > $t_{\alpha/2,\,\mathrm{df}=n-1}$ (3.499), you would say that you have evidence the sample mean was significantly different from 35 mg/capsule. Note here that you are concerned with the absolute value of t_{obs} ($|t_{\mathrm{obs}}|$) vs. t_{critical}. Given your chosen confidence, you have a t_{critical} value of 3.499, so as long as your $|t_{\mathrm{obs}}|$ is <3.499 you can conclude that the observed sample mean is not significantly different from the chosen μ value. Therefore, t_{obs} could be anywhere on the interval (−3.499, +3.499) and still be considered to be not significantly different from the label value. So, the observed mean could be above or below the chosen μ, as long as it is not "too far" in either direction.

Confidence intervals and t-tests provide the same information for a chosen confidence level. Either procedure may be used to compare the sample mean and chosen μ value. Sometimes you need to determine if the means of two samples are "different" (as opposed to one sample and a chosen μ). Sample means are "point estimates." Just because they are not exactly the same does not mean that the populations are significantly different. You need a statistical basis for determining if the sample means are far enough apart such that you can say they are different with some level of confidence (or not). Given two sample means ($\bar{x}_1$ and $\bar{x}_2$), standard deviations (SD_1 and SD_2), and sample sizes (n_1 and n_2), first we decide on a confidence level (C), and we calculate $\alpha/2$ as before. Then, we determine df. For two sample means, df is calculated as follows:

$$\mathrm{df} = n_1 + n_2 - 2 \tag{4.27}$$

Then you find the critical t-value as before, using the $\alpha/2$ and df values:

$$t_{\frac{\alpha}{2},\mathrm{df}=n_1+n_2-2}$$

Once you have t_{critical}, you calculate t_{obs} as follows:

$$t_{\mathrm{obs}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p{}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{4.28}$$

This formula differs from the one-sample t-test in that both n_1 and n_2 are used, and you employ a value known as the **pooled variance** ($s_p{}^2$) instead of the sample SD. The pooled variance is a weighted average of two sample SD values. The pooled variance is calculated as follows:

$$\text{Pooled variance} = s_p^2 = \frac{(n_1-1)\,\mathrm{SD}_1^2 + (n_2-1)\,\mathrm{SD}_2^2}{n_1+n_2-2} \tag{4.29}$$

Variance is simply standard deviation squared. Therefore, although you will not use this value, it may be useful to understand that the "pooled standard deviation" (s_p), or the weighted average of the two sample SD values, is simply the square root of the pooled variance:

$$\text{Pooled standard deviation} = \sqrt{\text{pooled variance}}$$
$$= \sqrt{s_p^2} = \sqrt{\frac{(n_1 - 1)\text{SD}_1^2 + (n_2 - 1)\text{SD}_2^2}{n_1 + n_2 - 2}} \qquad (4.30)$$

Once you have obtained the t_{critical} and t_{obs} values, you compare them as for the one-sample t-test:

$$t_{\text{obs}} > t_{\frac{\alpha}{2},\text{df}=n_1+n_2-2} \rightarrow \text{means } \textit{are} \text{ significantly different with specified confidence}(C)$$

$$t_{\text{obs}} < t_{\frac{\alpha}{2},\,\text{df}=n_1+n_2-2} \rightarrow \text{means } \textit{are not} \text{ significantly different with specified confidence}(C)$$

Example G2 Suppose your company is concerned that two production lines are producing spaghetti sauces with different acid contents. You measure the titratable acidity in samples from both lines (Line 1, 2.1; 2.0; 2.1; 2.2; 2.3; and 2.4 %; Line 2, 2.7; 2.3; 2.2; 2.2; 2.4; 2.6; and 2.5 %). Do the two production lines appear to be producing significantly different acidity levels with 90 % confidence?

First, calculate the n, x^{-}, and SD value for each sample:
Line 1: $n = 6$, $\bar{x}$=2.183 and SD = 0.147
Line 2: $n = 7$, $\bar{x}$=2.4, and SD = 0.195

What is the confidence level (C): 90 % or 0.90?

Next, calculate $\alpha/2$: $\frac{\alpha}{2} = \frac{1-C}{2} = \frac{1-0.90}{2} = \frac{0.1}{2} = 0.05$

Forthe$t$-table,df: $\text{df} = n_1 + n_2 - 2 = 6 + 7 - 2 = 11$

Find the critical t-value on the table:

$$t_{\frac{\alpha}{2},\text{df}=n_1+n_2-2} = t_{0.05,11} = 1.796$$

Calculate a pooled variance (s_p^2):

$$s_p^2 = \frac{(n_1 - 1)\text{SD}_1^2 + (n_2 - 1)\text{SD}_2^2}{n_1 + n_2 - 2}$$
$$= \frac{(6-1)(0.147)^2 + (7-1)(0.195)^2}{6+7-2}$$
$$= \frac{0.1083 + 0.2286}{11} = 0.03063$$

Calculate t_{obs}:

$$t_{\text{obs}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_p{}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|2.183 - 2.4|}{\sqrt{0.03063\left(\frac{1}{6} + \frac{1}{7}\right)}}$$
$$= \frac{|-0.217|}{0.09737} = 2.23$$

Decision:

$$t_{\text{obs}} = 2.23 \quad \text{and}$$
$$t_{\frac{\alpha}{2},\,\text{df}=n_1+n_2-2} = 1.796 \rightarrow t_{\text{obs}} > t_{\frac{\alpha}{2},\,\text{df}=n_1+n_2-2}$$

Thus, Lines 1 and 2 produce sauces with significantly different acidities (with 90 % confidence).

4.8 PRACTICAL CONSIDERATIONS

Here are some practical considerations to help you use statistics in a food analysis course and in a career in which decisions are based on analytical data:

4.8.1 Sample Size

1. A larger sample size enables more accurate approximation of the true population values.
2. Larger sample sizes decrease the sample SD.
3. Use t for $n < 25$–30; use Z for $n > 25$–30.
4. Sample size determines df, which determines t_{critical} for a specified $\alpha/2$ value.
5. Sampling more units is often useful but not always practical or cost-effective.
6. A "happy medium" needs to be reached that provides "adequate" approximation of the mean with a fixed amount of confidence. This is done by calculating the minimum needed sample size (this is covered in the textbook chapter on sampling).

4.8.2 Confidence

1. Confidence values are chosen based on the acceptable risk to consumers (safety, poor quality, etc.), the company (quality, profit margin, formulation accuracy, etc.), or government regulatory agencies (e.g., labeling accuracy).
2. Confidence determines $\alpha/2$, which determines t_{critical} within a specified df value.
3. The more confidence is required (or desired):

(a) The wider a CI will be.
(b) The larger t_{critical} will be.
(c) The larger t_{obs} must be to be > t_{critical}.
(d) The larger the n size must be to show significance.

4. Increasing desired confidence increases the statistical burden required to shown significant differences; 90–99 % confidence is typically used (95 % is fairly standard).
5. Increasing the desired confidence is not always practical or cost-effective.

4.8.3 What Test to Use?

Table 4.1 gives "rules of thumb" for what types of tests to use for various scenarios:

4.1 table Examples of when specific statistical tools should be used for food analysis

Question	*Test/calculation*	*Examples*
Is an observed sample mean statistically similar to or different from a chosen value?	One-sample t-test	Is the actual value in agreement with the label value? Does the permitted level differ from the allowed level? Does the raw material meet company specifications? Is the composition of the finished product acceptable based on company specifications? Is a pipette delivering the desired volume? Is a packaging machine filling packages to the desired level?
What is the actual population value, based on the sample?	One-sample CI	What is the actual concentration of a compound of interest?
Based on samples from two populations, are these populations different?	Two-sample t-test ($\mu = 0$)	Are two lots different? Are two lines, plants, etc. producing products with different composition? Is your product the same or different from a competitor's product?

4.9 PRACTICE PROBLEMS

(*Note*: *Answers are at end of laboratory manual.*)

1. As QA manager for a canned soup manufacturer, you need to make sure that chicken noodle soup has the sodium content indicated on the label (343 mg/cup). You sample six cans of soup and measure the sodium content: 322.8, 320.7, 339.1, 340.9, 319.2, and 324.4 mg/cup. What is the mean of the observations (mg/cup)? What is the standard deviation of the observations (mg/cup)? Calculate the 96 % confidence interval for the true population mean, and determine both the upper and lower limits of the confidence interval. Determine if the sample mean provides strong enough evidence that the population is "out of spec" with 99 % confidence.
2. You perform moisture analysis on sweetened condensed milk using a forced-draft oven. The following data are obtained. Lot A: 86.7, 86.2, 87.9, 86.3, and 87.8 % solids; Lot B: 89.1, 88.9, 89.3, 88.8, and 89.0 %. Determine if the lots are statistically different with 95 % confidence

4.10 TERMS AND SYMBOLS

Confidence (*C*) Statistical probability of being correct based on chance alone.

Confidence interval (*CI*) A range of values (based on a point estimate plus a statistically determined margin of error) used to predict information about a population.

Degrees of freedom (*df*) The number of values that is free to vary independently.

Distribution All values in a population and the relative or absolute occurrence of each value.

Histogram A plot of all values in a population and the relative occurrence of each value.

Mean (μ) Average value.

Normal (*N*) A population distribution whose shape is defined by a mean and standard deviation.

Pooled variance (*sp2*) A measure of variability for two individual samples, incorporating both the sample sizes and sample standard deviations of both.

Population All individuals of interest.

Population mean (μx) The average value of all the members of a population.

Population standard deviation (σ) The average difference between the value of each population member and the population mean.

Probability (*P*) Statistical likelihood of an event or state occurring by chance alone.

Sample Selected members of a population, from which inferences are made about the entire population.

Sample mean ($\mu\bar{x}$) The average value of all the members of a sample.

Sample standard deviation (*SD*, $\sigma\bar{x}$) The average difference between the value of each sample member and the sample mean.

Single point estimate A single value (mean or single data point) used to predict information about a population.

Standard deviation (σ) The average difference between the value of each population member and the population mean.

Standard error of the mean (*SEM*) A measure of variability of a sample, incorporating the sample standard deviation and the sample size.

T-score (t) A normalized value describing how many standard deviations a value is from the mean but more conservative than the Z-score.

Z-score (Z) A normalized value describing how many standard deviations a value is from the mean.