Box 1. Means, variances, medians, and modes

In the main text of the book, we discuss proportions and rates, which are appropriate measures when the health indicator of interest is the presence or absence of disease. However, a health indicator is often not measured as present versus absent; rather, it is measured in terms of gradations. Examples include blood pressure, cholesterol, body mass index, birth weight, lung function, number of symptoms of depression or anxiety disorders, and many others. In these cases, we need to describe the data in terms of measures of centrality and spread of the data.

For example, a study was conducted to determine the average body mass index (BMI) of adults In Farrlandia. A random sample (see Chapter 4) was taken of 31 individuals. These are the data obtained:

Box 1, Table 1. BMI in a random sample of 31 Farrlandians

Subject	BMI
1	22.1
2	22.6
3	23.4
4	23.4
5	23.4
6	25.4
7	26.2
8	26.5
9	26.6
10	26.7
11	27.4
12	27.4
13	27.6
14	28.1
15	28.6
16	28.7
17	29.1
18	29.2
19	29.4
20	29.6
21	30.1
22	30.1
23	31.2
24	31.3
25	32.1
26	37
27	40.3
28	47.9
29	48

30	50.6
31	55

Mean

The population mean is estimated by summing the outcomes for all individual and dividing that sum by the total number of individuals. It is perhaps the most frequently used measure of centrality for continuous outcomes in epidemiology. The mean provides the average value of the outcome for all sample participants. In our study of BMI in Farrlandia, we would calculate the mean as follows:

$$\frac{\begin{gathered}22.1+22.6+23.4+23.4+23.4+25.4+26.2+26.5+26.6+26.7+27.4+\\ 27.4+27.6+28.1+28.6+28.7+29.1+\\ 29.2+29.4+29.6+30.1+30.1+31.2+31.3+32.1+\\ 37+40.3+47.9+48+50.6+55\end{gathered}}{31} = 31.1$$

Thus, the mean BMI in our sample is 31.1.

Note that we can consider the mean as an extension of the proportion that we discussed earlier for estimating prevalence and risk. Why is this? Going back to our data on BMI in Farrlandia, let us create a new variable that is 1 if an individual is considered obese (BMI>30) and a 0 if not. Now our data would look like the following table:

Box 1, Table 2. BMI in a random sample of 31 Farrlandians

Subject	BMI	BMI Over 30? (0 no, 1 yes)
1	22.1	0
2	22.6	0
3	23.4	0
4	23.4	0
5	23.4	0
6	25.4	0
7	26.2	0
8	26.5	0
9	26.6	0
10	26.7	0
11	27.4	0
12	27.4	0
13	27.6	0
14	28.1	0
15	28.6	0
16	28.7	0
17	29.1	0
18	29.2	0
19	29.4	0
20	29.6	0
21	30.1	1

22	30.1	1
23	31.2	1
24	31.3	1
25	32.1	1
26	37	1
27	40.3	1
28	47.9	1
29	48	1
30	50.6	1
31	55	1

Taking the mean of the BMI present or not variable, we get:

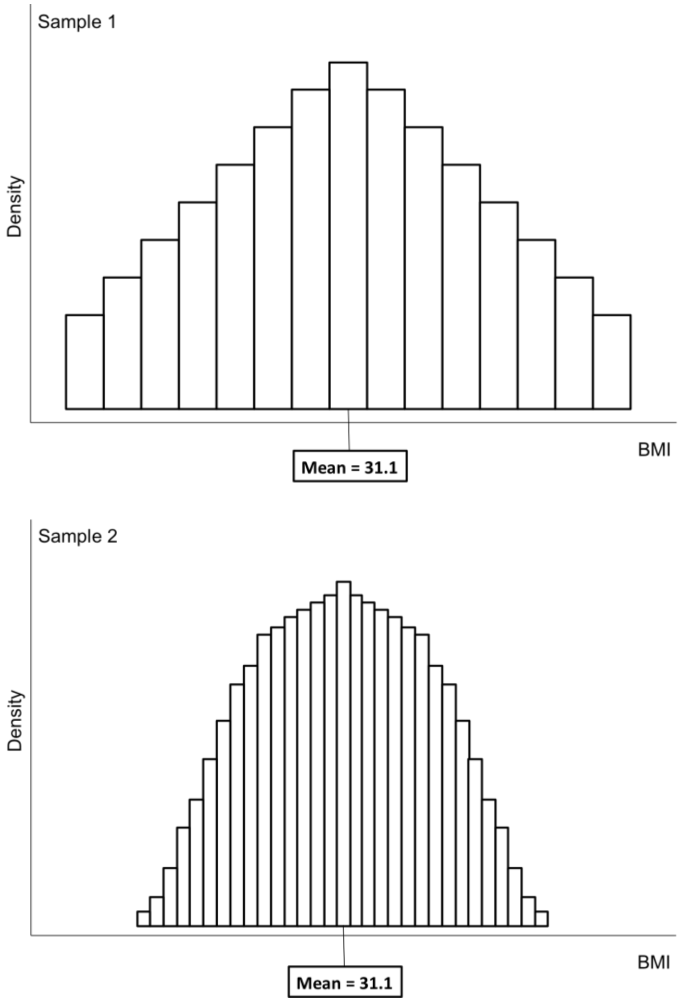
$$\frac{\begin{array}{c}0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+\\0+1+1+1+1+1+1+1+1+1+1+1\end{array}}{31} = 0.35$$

We calculated the mean of this series the same way we calculated the mean of BMI in the previous table: we added all of the values and divided by the total. The only difference is that all of the values happened to be either 0 or 1 in this case. What we find is that the mean of the series is equal to the proportion. The numerator is the number of cases of obesity (11) and the denominator is the sample size (31). The mean (0.35) can be interpreted as the proportion of people with obesity, which equals 35.0%.

Variability and spread

In addition to estimating the mean of a continuous variable, it is important to estimate how close all of the individual values are to that mean. For example, suppose we sampled two populations and obtained the following histograms of their risk of disease, shown in Figure 4 below.

Box 1, Figure 1. Mean BMI in two populations



Both of these samples have the same mean BMI 31.1. But in Sample 2, the values of BMI for each individual are all much closer to the mean than the values on in Sample 1. The spread of the individual values around the mean is a measure of the variance of the data. The size of the variance gives us important information about the distribution of the variable of interest within the sample. A large variance tells us that while the mean may be 31.1, there is a wide range of total values across the whole sample (and, if a representative sample, across the underlying population). A small variance tells us that there is little variability in the sample (and, if a representative sample, in the underlying population) with respect to the variable of interest.

A standard measure of the variability or spread of a variable in a sample is known as the standard deviation (see Box 6 in the online material that accompanies Chapter 5 for a short example of how to calculate the standard deviation). With the standard deviation we can make a direct comparison across samples on how the health indicator of interest is distributed. For example, suppose in Sample 1 we measure BMI and estimate a mean of 30 and a standard deviation of 10, compared to another sample with a mean of 35 and a standard deviation of 5. While the first sample has lower mean BMI than the second, it also has a greater spread of BMIs around that mean in the first sample. Further, with a mean and standard deviation, we can standardize across multiple measures to quantify the number of standard deviations each observation is from the mean (this is called a z-score, see Box 5 in the online material that accompanies Chapter 5).

The standard deviation provides additional information about the distribution of values when the distribution of values is normally distributed. A normally distributed variable is one in which each value has a high probability of being close to the mean and a symmetrically lower probability of being farther from the mean on either side (normal distributions typically look like a bell shaped curve). For example, if the mean of a normally distributed variable is 100, the probability that any individual value is close to 100 is high. When a variable is normally distributed, approximately 68% of all of the values in the sample will fall within one standard deviation of the mean. Approximately 95% will fall within two standard deviations, and approximately 99.7% will fall within three standard deviations. This becomes important as we construct confidence intervals around sample estimates (see Box 4).

The main limitation of the mean as a measure of centrality is that it can be influenced by extreme values in the data. For example, consider the situation in which one of the data collectors miscoded individual 31 in our dataset as having a BMI of 550 instead of 55. The mean would increase to 47.1 from than 31.1 due to the one extreme observation. In general, when the outcomes are not evenly distributed across a full range of potential values and instead are aggregated at the low end or the high end, the mean may not be the most informative measure of centrality. For example, suppose we would like to measure the mean number of cigarettes smoked per day among a sample of adolescents. We obtain the following data:

Box 1, Table 3. Number of cigarettes smoked per day among a random sample of 17 adolescents

Subject	Number of cigarettes smoked per day
1	1
2	1
3	1
4	2
5	2
6	2
7	2
8	2
9	2
10	3
11	3
12	3
13	3
14	10
15	20
16	40
17	60

The mean for this series would be 9.24. However, most of the values in the data are between 1 and 3. Therefore, reporting that the average number of cigarettes smoked by these teens is 9.24 is not very informative. We need additional measures of centrality that allow us to better capture results such as these, as will be discussed below with median and mode.

Median

The median of a set of ordered values is the numerical value that falls in the exact middle of the ordered set of values; it is the value for which 50% of the remaining values are above and 50% are below. Considering first a simple case:

$$3 \;\; 5 \;\; 7$$

The median value is 5, because there is one observation below and one observation above. Consider another case:

$$3 \;\; 3 \;\; 5 \;\; 7 \;\; 9 \;\; 9\,11$$

The median value of this set is 7, because there are three observations above and three observations below.

Considering our smoking variable shown in Table 3, the median value would be 2. There are eight observations that fall below 2 in this set, and eight that fall above 2. Thus, whereas the mean number of cigarettes smoked was 9.24, the median was 2. This signals that the distribution is quite skewed by a few heavy smokers.

In the above examples, the sets of values that we used for calculating medians each had an odd number of observations. When there is an even number of observations, how do we calculate a median? Consider this series of values:

$$1 \;\; 1 \;\; 3 \;\; 4 \;\; 7 \;\; 9$$

There are six observations in this set; there is no single value that falls directly in the middle. In this case, we take the mean of the two values most centered. Since 3 and 4 are the most centered values (two observations fall below, and two observations fall above), the median of this set is the mean of 3 and 4:

$$\frac{3+4}{2} = 3.5$$

Mode

One simple measure of centrality is the most frequently observed value, which is labeled the mode. Returning to our example of cigarette smoking from Table 3, we can determine the following: 3 students reported smoking 1 cigarette per day, 6 students reported 2 cigarettes per day, 4 reported 3 cigarettes per day, 1 student reported 10 per day, 1 student reported 20 per day, 1 student reported 40 per day, and 1 reported 60 per day. The mode is the value that is occurs most frequently; given that 6 students reported 2 cigarettes per day, the mode is 2.

Box 2. Basic reproductive rate, net reproductive rate, and herd immunity

The measures of disease occurrence and frequency we have discussed in Chapter 5 are useful whether the health indicator of interest is infectious or not. However, if the health indicator of interest is an infectious disease, additional measures of occurrence and frequency become useful. A central measure that is important to discuss is the reproductive rate. There are two versions of the reproductive rate we will introduce in this section: the basic reproductive rate and the net reproductive rate.

The basic reproductive rate of disease is a measure of the average number of people affected by each infectious case in a population among whom all individuals are susceptible to the disease (e.g. no vaccination available and no immunity of any kind). The reproductive rate is estimated as a function of three basic parameters:

- Risk of disease per contact
- Number of contacts in a given time period
- Duration of infectiousness

Considering a hypothetical infectious disease, we can conceptualize how changes in these parameters would affect the number of cases that on average arise from any given case. For example, at a given risk of disease per contact and duration of infectiousness, increasing the number of contacts will increase the number of secondary cases. Similarly, given a fixed number of contacts and duration of infectiousness, infections that have a higher risk of disease per contact will produce more secondary cases. These concepts lead directly to control measures. By reducing the number of contact (e.g., through quarantine), we can limit the number of secondary cases per infectious person. By treating infections quickly as they arise and therefore reducing the duration of infectiousness, we can also reduce the number of secondary cases. Finally, by promoting measures such as condom use and other barrier protection, we can reduce the risk of disease per contact for sexually transmitted infections.

Basic reproductive rates vary substantially across many types of infections. Infections such as measles and pertussis have very high average basic reproductive rates as the probability of infection per contact is very high.

Where do concepts such as vaccination fit into with this? Within many populations there will be individuals who are immune to the disease, either through vaccination, previous infection, or other means. The net reproductive rate, then, is the basic productive rate multiplied by the proportion that is susceptible to infection (also equal to 1 minus the proportion that is immune) or:

$$R = R_0 \times (1 - p)$$

Where p is the proportion of individuals immune to the disease, R_0 is the basic reproductive rate, and R is the net reproductive rate.

The magnitude of the net reproductive rate provides important information about the potential for disease epidemics. When the net reproductive rate is greater than 1, the disease will be epidemic in the population. That is, the disease will spread and cases will continue to accumulate. This makes theoretical sense, because if each case of disease is infecting more than one other person, on average, the cases will quickly multiply exponentially in number. When the reproductive rate is 1, the disease will continue to be prevalent in the population but without an exponential epidemic. Diseases that are present in the population without exponential epidemics are endemic. Finally, when the net reproductive rate is less than 1.0, the disease will eventually be eradicated.

Using the idea of basic and net reproductive rate, we can estimate the proportion of people in the population who need to be immune to the disease in order to prevent an epidemic. We know that an epidemic will not occur if the net reproductive rate is 1 or below. The proportion of people that need to be immune in order to prevent an epidemic, then, is the proportion that satisfies the condition that R is at most equal to 1. Therefore, we solve for p in the formula above, setting R_0 at a known value and R at 1.

To illustrate this, we will use the example of measles, which has a high basic reproductive rate of 15. What proportion would need to have immunity to measles in order to prevent an epidemic? We solve for p such that

$$1 = 15 \times (1 - p)$$

The proportion that satisfies the above formula is 93.3%. Therefore, 93.3% of the population needs to be immune to measles in order to prevent an epidemic of measles when there is a single case in the population.

An important concept illustrated here is that of herd immunity. Notice that the proportion of people who need to be immune to measles to prevent an epidemic is less than 100%. Thus the entire population can be protected from measles despite not everyone in the population having personal immunity to measles. As long as the net reproductive rate is 1 or below, epidemics of disease will not occur. Thinking about a public health approach to vaccination, for example, this concept indicates that as long as a high proportion of the population is vaccinated, the entire population will be protected even if not all members of the population are vaccinated.
The basic reproductive rate provides a quantitative measure of how quickly an infectious disease will spread in a population in which no one is immune. Within many populations, however, there will be individuals who are immune to the disease, either through vaccination, previous infection, or other means.

Box 3. Disability adjusted life years (DALYs)

One way in which to measure health is through the potential years of life lost due to disability associated with the occurrence of sickness, disease, symptoms, or syndromes. The average disability-adjusted life years associated with a particular health indicator is comprised of two components: the Years of Life Lost (YLL) and the Years Lost due to Disability (YLD). YLL can be conceptualized as a measure of the number of additional years that an individual would have survived without the presence or level of the particular health indicator. YLD can be conceptualized as the years of productive life lost due to reduced functioning because of the particular health indicator. For example, if an individual survives but cannot participate in the labor market due to illness, this lack of productivity would be counted as YLD. Basic formulae for YLD and YLL are:

$$\mathrm{YLL} = \mathrm{N} * \mathrm{L}$$

Where N is the number of deaths attributable to the health indicator of interest in a specific population in a given year, and L is the difference between the average life expectancy in the overall population and the average age of death among those who died of the disease under study. For example, in a population of 100,000 individuals, if 500 deaths are attributable to HIV/AIDS and those 500 people had an average age of death of 60 compared to the population average life expectancy of 80 years, the YLL is (80-60)*500. Thus, there are 10,000 years of life lost due to HIV in this population.

$$\mathrm{YLD} = \mathrm{I} * \mathrm{L} * \mathrm{DW}$$

Where I is the number of incident cases involving the health indicator, L is the average duration of the disease or other health indicator, and DW is a disability weight. A disability weight is an empirically determined measure of the severity of a health indicator on a scale of 0 to 1, with 0 being no effect on health and 1 being death. Thus, a disease that does not cause immediate death but reduces ability to function in daily life would receive a disability weight greater than 0 and less than 1. The YLD would then increase in proportion to the duration of the disease and the number of incident cases.

Continuing with our previous example, assuming that in a population there are 100 incident cases of HIV in a given year, an average duration of illness of 10 years, and a disability weight of about 0.14, the YLD is 140. Thus, there are 140 years lost due to disability associated with HIV in this population.

Using the YLD and the YLL, we then estimate the DALYs associated with a particular health indicator as:

$$\mathrm{DALY} = \mathrm{YLD} + \mathrm{YLL}$$

That is, the sum of the years of life lost and the years lost due to disability forms the disability-adjusted life years due to a health indicator of interest. In our example, the DALY would be 10,000+140, or 10,140 disability-adjusted life years lost due to HIV.

The DALY as a measure of disease burden in populations has been criticized when there is a lack of empirical information on which to base parameter estimates. For example, an accurate measure of DALY requires knowledge of the average age of death among those with the health indicator, average life expectancy, incidence, and duration. Each of those parameters can be difficult to estimate accurately without large-scale data collection projects. Even when these parameters are known, estimating the disability weight can be difficult and empirical data on which to base the weight can be hard to identify. Finally, accurate DALYs require causal attribution; that is, we need to specify how many deaths and how much disability are directly attributable to the health indicator. Diseases and other illnesses often co-occur, and it is difficult to correctly attribute death and disability to a specific cause. For example, if an obese individual dies following a myocardial infarction, is the death attributable to obesity or myocardial infarction? Because of the co-occurrence of many diseases and other health indicators of interest, DALYs are often over-estimated for any particular disease or health indicator.

Box 4. Standard error and confidence interval for a mean

The standard error for a mean is calculated as follows:

$$SE = \frac{SD}{\sqrt{n}}$$

Where SD is the sample standard deviation (see Box 6 in the online material that accompanies Chapter 5) for the mean and *n* is the sample size.

95% confidence interval for a mean:

$$\text{Mean} \pm 1.96(SE)$$

Where SE is the standard error of the mean.

Box 5. What is a z-score and why does it matter?

Thus far, we have learned about the mean as a measure of centrality when a variable is normally distributed, and the standard deviation as a measure of the spread around that mean. Approximately 68, 95, and 99% of normally distributed data values will fall within 1, 2, and 3 standard deviations of the mean, respectively. What if we want to know the probability that a data value will fall below 1.75 standard deviations of the mean? Or above 2.77 standard deviations of the mean? We could calculate these probabilities, but the calculations are messy and difficult. Instead, what we typically do is transform all variables into one standard distribution where the probabilities at each level have been well characterized.

To do this, we take the data values and standardize them to a z-score, which is based on a z-distribution. A z-distribution is a normal distribution with a mean of 0 and a variance of 1. We also call this distribution the standard normal distribution. We can take any normal distribution and standardize it to a z-distribution. The z-distribution has many uses and is central to decisions we make about the statistical importance of associations in epidemiology and biostatistics.

Standardizing the data values to obtain z-scores is straightforward. For each data value, we subtract the mean and divide it by the standard deviation. For example, in the table below we take a variable from a hypothetical sample of 15 individuals. In the second column is the value of the variable, and in the fourth column is the value once standardized to a z-distribution.

Box 5, Table 1. Hypothetical sample for z-distribution analysis

Participant number	Original value	Standardization	Standardized value
1	20	(20–24.96)/2.39	–1.51
2	20	(20–24.96)/2.39	–1.51
3	21	(21–24.96)/2.39	–1.20
4	22	(22–24.96)/2.39	–0.90
5	23.40	(23.40–24.96)/2.39	–0.47
6	24	(24–24.96)/2.39	–0.29
7	25	(25–24.96)/2.39	0.01
8	25.30	(205.30–24.96)/2.39	0.10
9	25.70	(25.70–24.96)/2.39	0.22
10	26	(26–24.96)/2.39	0.32
11	26	(26–24.96)/2.39	0.32
12	27	(27–24.96)/2.39	0.62
13	29	(29–24.96)/2.39	1.23
14	30	(30–24.96)/2.39	1.53
15	30	(30–24.96)/2.39	1.53
Mean	24.96		0.00
Standard deviation	3.29		1.00

Notice that the original mean is 24.96 with a standard deviation of 3.29; once we standardize to a z-distribution, the mean is 0 and the standard deviation is 1.0. Similar to any normal distribution, 68, 95, and 99% of data points will fall within 1, 2, and 3 standard deviations. Because the standard deviation is 1, this means that 68% will fall between –1 and 1, 95% will fall between –2 and 2 and 99% will fall between –3 and 3.

Now, suppose we want to know the answer to the question - What proportion of this population would have a value greater than 26? We would do this in several steps.

Step 1. Calculate the z-score of the value

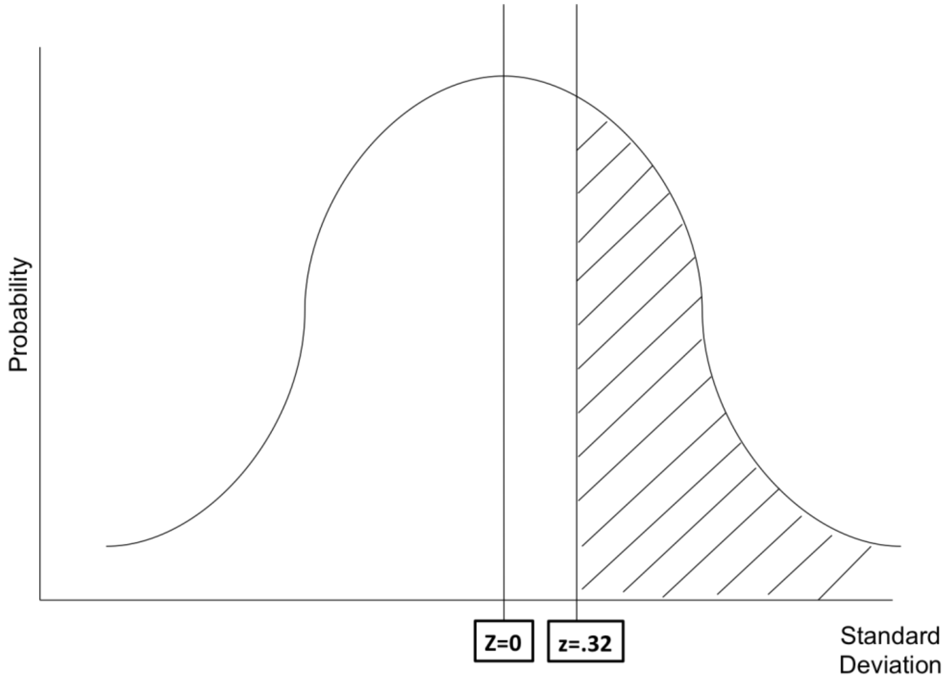
The z-score is the value on the z-distribution for the data point of interest, that is, the value of interest subtracting the mean, and divided by the standard deviation. In our example, the z-score would be:

$$\frac{26-24.96}{3.29}=0.32$$

Step 2: Estimate the desired probability based on the z-distribution

In our example, we want to find the probability that a value could be greater than 0.32 on the z-distribution. A first useful step is to graph the z-distribution and shade the area under the curve for the probability we wish to estimate. We do this in Box 5, Figure 1 below.

Box 5, Figure 1. A standard normal z-distribution with a shaded area denoting values of 0.32 or greater.



Thus, our problem is to calculate the probability of a value falling above 0.32 on the z-distribution. Because this distribution is standardized, the probability of falling above or below each value has already been calculated, and in Appendix 1 of this book we provide those probabilities.

Using Appendix 1, we can see that the probability corresponding to 0.32 is = 0.3745. Thus, we would expect that 37.45% of the population has a BMI greater than 26.

The connection between the z-distribution and confidence intervals

Let us now return to the issue of estimating confidence intervals. The most common confidence interval that we see in epidemiologic studies is the 95% confidence interval. We know that a 95% confidence interval gives a range of plausible values for the population parameter. What if we want to estimate the 93% confidence interval? If we desire 93% confidence then we are saying: if we could perform repeated sampling and construct many confidence intervals, 93% of the intervals would contain the population parameter and 7% wouldn't. Furthermore, about 3.5% of the intervals would consist of values that are all larger than the population parameter and about 3.5% of the intervals would consist of values that are all smaller than the population parameter. Thus, we can examine the z-distribution probabilities for the z-score that would correspond to a probability of 0.965 (1 – 0.035). This z-score is 1.82. If we wanted to estimate the 93% confidence interval for a mean or a proportion, we would multiply the standard error by 1.82, then add and subtract this value from the sample estimate of the mean or proportion.

Let us now return to the issue of estimating confidence intervals. The most common confidence interval that we see in epidemiologic studies is the 95% confidence interval. Suppose that we are interested in learning about a population mean. For large random samples, the sampling distribution of the sample mean ($\bar{X}$) is approximately normal with mean equal to the population mean and standard deviation equal to $\frac{SD}{\sqrt{n}}$, otherwise known as the standard error. If we standardize these sample mean values then these z-scores follow a z-distribution. The middle 95% of values in this distribution fall between –1.96 and 1.96. You can check this using Appendix 1. The fact that there is a 95% chance of z being between –1.96 and 1.96 is equivalent to saying that there is a 95% chance of observing a sample mean that is within 1.96x$\frac{SD}{\sqrt{n}}$ of the population mean. And if the sample mean is within 1.96x$\frac{SD}{\sqrt{n}}$ of the population mean, this means that the interval $\bar{X} - 1.96\frac{SD}{\sqrt{n}}$ to $\bar{X} + 1.96\frac{SD}{\sqrt{n}}$ will capture the population mean (and this will happen in 95% of all possible samples). But if the sample mean is farther than $1.96\frac{SD}{\sqrt{n}}$ from the population mean (which will happen in about 5% of the samples we draw) then the interval will not contain the population mean.

We can use the same thought process to construct confidence intervals with various degrees of confidence. For example, if we want a 93% confidence interval then we need to determine the values from the z-distribution that bound the middle 93% of the distribution. It might be easier to think of these as the values that cut off 3.5% in the lower tail and 3.5% in the upper tail of the z-distribution. From Appendix 1 we see that these values are –1.82 and 1.82 respectively. Q 93% confidence interval for the population mean ranges from $\bar{X} - 1.82\frac{SD}{\sqrt{n}}$ to $\bar{X} + 1.82\frac{SD}{\sqrt{n}}$.

Box 6. Calculating the standard deviation of the mean

Consider the following 10 values:

2 4 5 6 8 10 12 13 16 19

To estimate the mean, we sum the values and divide by the sample size:

$$\frac{2+4+5+6+8+10+12+13+16+19}{10} = 9.50$$

Step 1: Calculate the difference between each value and the mean
The variance of a variable is a measure of the spread around the mean. Thus, the first step in quantifying that spread is to calculate how close each value is to the mean. We do this by subtracting each value from the mean (see Column 2 in the table below). For example:

$$2 - 9.50 = -7.50$$
$$4 - 9.50 = -5.50$$
etc.

Step 2: Square each difference
We then square each of these values so that they are all on a positive number scale (see column 3 in the table below). For example:

$$-7.5^2 = 56.25$$
$$-5.5^2 = 30.25$$
etc.

Step 3: Sum the squared values and divide by the sample size minus 1

In Step 3 we essentially take the mean of these squared differences. We sum the differences together and divide by the sample size minus 1. We subtract 1 from the sample size as a small correction factor.

$$\frac{56.26 + 30.25 + 20.25 + 12.25 + 2.25 + .25 + 6.25 + 12.25 + 42.25 + 90.25}{10 - 1} = 30.28$$

The calculation that we obtain in Step 3 is the variance.

Step 4: Take the square root of the variance.

To obtain the standard deviation, take the square root of the variance.

$$\sqrt{30.28} = 5.50$$

Thus, the set of values in this example has a mean of 9.50 and a standard deviation of 5.50.

Box 2, Table 1. Differences and squared differences for each value

Value	Difference from Mean	Squared Difference
2	–7.50	56.25
4	–5.50	30.25
5	–4.50	20.25
6	–3.50	12.25
8	–1.50	2.25
10	0.50	0.25
12	2.50	6.25
13	3.50	12.25
16	6.50	42.25
19	9.50	90.25