Chapter 1

Statistics

(SOLUTIONS)

The theory of probability can be applied to the study of *statistics*, which may be defined as *the branch of mathematics concerned with the collection, classification, analysis, and interpretation of numerical facts, for the purposes of drawing inferences from their quantifiable probability.*

The big idea here is that natural phenomena are, in general, not completely *deterministic*. That is, they do not evolve according to precisely predictable formulas or recipes. Still, deterministic models like those examined in prior chapters often give good approximations to what happens "in real life." And we can apply statistical analyses to get a sense of *how well* these models reflect reality. And we can thereby quantify, in rigorous ways, statements like "we have this much confidence that this given situation will yield an outcome in the following prescribed range."

In this chapter, we present a sketch of the above ideas, with just enough detail that we can investigate two main (related) tools in statistical inference: *hypothesis testing* and *confidence intervals*. The reader should note, along the way, how our development of these ideas mirrors, and uses, some previously studied concepts – concepts of definite integrals, areas, Riemann sums, and the Fundamental Theorem of Calculus.

1.1 Relative frequency density

Flipping coins; the central limit theorem

Consider the binomial experiment of flipping six coins, and recording the number of heads that come up. The outcome of such an experiment will, of course, be one of the integers 0, 1, 2, 3, 4, 5, 6.

A certain Calculus class at the University of Colorado Boulder repeated this experiment 3,444 times. The results of these 3,444 trials of this experiment are summarized in the following "frequency table."

Table 1. Six coins, flipped 3,444 times

Number of heads	0	1	2	3	4	5	6
Frequency	66	341	825	1048	780	333	51

We can compile this data into a *histogram*, with "frequency" on the vertical axis and "number of heads" on the horizontal. We get a figure like this:

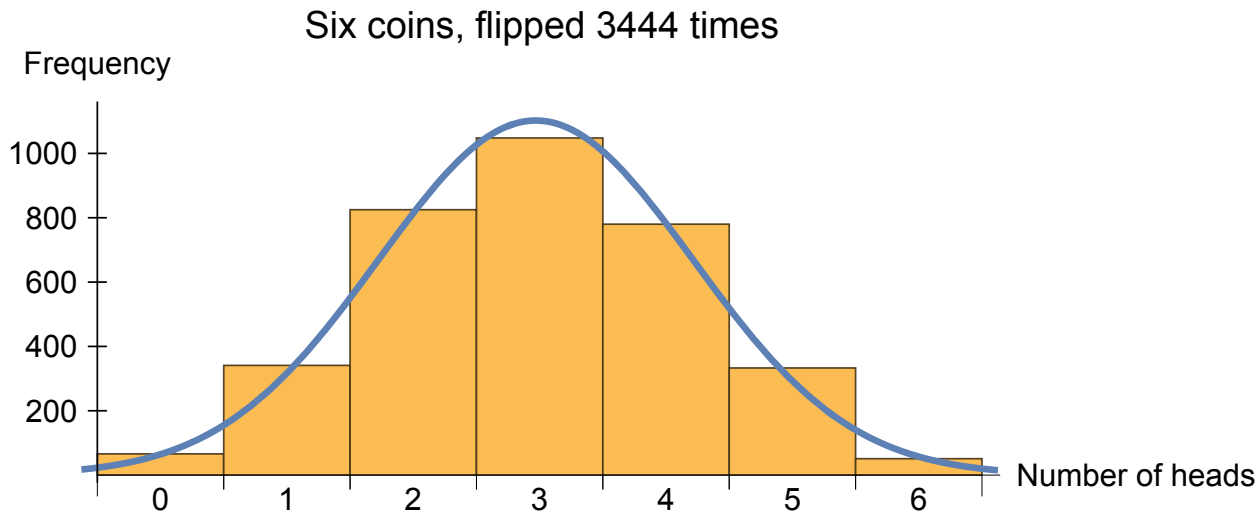


Figure 1. Histogram for trials of a coin-flipping experiment

We've superimposed, on the above histogram, a certain bell-shaped curve that approximates the "shape" of the data. The curve we've used is a particular kind of bell-shaped curve, known as a *normal* curve, and the normal curve we've chosen is one that is especially well-suited to the data. We'll explain the meaning of all of this in the next section.

(Note that the bars of our histogram are *contiguous*; there is no space between them. This is a convention that we will always follow, and that will be important to our interpretation, later in this section, of histograms in terms of area.)

The "normal" shape of a coin-flipping experiment becomes even more evident if each trial of the experiment entails a larger number of coins, and if many more trials are performed. For example, if an experiment comprises the flipping of 50 coins, and this experiment is repeated 300,000 times, then the result might look like this:

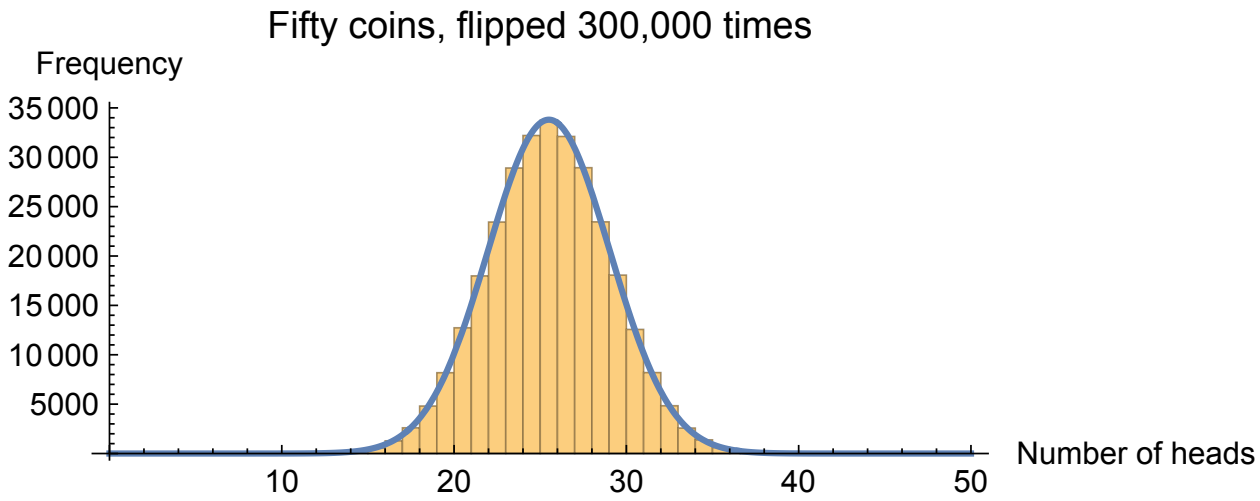


Figure 2. Histogram for trials of another coin-flipping experiment

(The data for the above histogram were obtained through a *simulation*. In other words, this data was *not* obtained by actually flipping 50 coins, 300,000 times. Rather, a computer mathematics package was used to pick, at random, 50 numbers, where each number could either be a zero – representing a coin turning up "heads" – or a one – representing "tails." The number of zeroes resulting from this was recorded, and this process was repeated 300,000 times.)

As before, we have fit the histogram with a certain "normal" curve that is particularly well-suited to the data. In this case, the fit is quite close.

The above discussions illustrate a *huge* result in probability, which will be central (pun intended) to our discussions of statistical inference in the next section.

If each trail of an experiment comprises many small, independent factors, all of which behave similarly, and many trials are performed, then (under some mild technical conditions) the outcomes of the experiment will follow a roughly *normal* distrbution.

The Central Limit Theorem

We will not prove this result; proofs may be found in most advanced texts on probability and statistics. We do take a moment, though, to note how this theorem reflects our discussions above. Consider, in particular, the scenario encapsulated by Figure 6.2. There, the 50 coins being tossed are the "many small, independent factors, all of which behave similarly" of the theorem. (They are *independent* in the sense that no one coin affects the behavior of any others. Of course, the coins might bounce against each other on the way down, but we can assume that any effects of this contact cancel each other out, in terms of the probability of any coins coming up heads. Alternatively, we can imagine that the fifty coins are flipped one at a time.) And the 300,000 repetitions are the "many trials" cited in the theorem.

We have not been specific about *how close* to normal one's distribution will be, or the manner in which this might depend on *how many factors* there are, or *how many trials* are performed.

Nor will we elaborate on this much. The important idea, for our purposes, is that *more factors* and *more trials* tend to produce distributions that are *more normal.* This idea is exemplified by comparing the scenarios and histograms of Figures 1 and 2 above. (The importance of having numerous factors and numerous trails may also be appreciated by considering some rather extreme cases. Specifically, imagine flipping a *single* coin, any number of times, or a *huge* number of coins, just once. In neither case will the histogram thus obtained look at all bell-shaped!)

Mean and standard deviation

Ultimately, we wish to draw stronger connections between histograms (like the ones in Figures 1 and 2 above) and curves that "fit" them (like the ones superimposed on the histograms in the above figures). To this end, we'll need formulas for certain quantities related to the "shape" of a data set.

We begin with the following.

Definition 1.1.1. Let X be a set consisting of n numerical (not necessarily distinct) data points, labeled $x_1, x_2, x_3, \ldots, x_n$. That is,

$$X = \{x_1, x_2, \ldots, x_n\}.$$

Then:

(a) We define the *mean* $\bar{x}$ and *standard deviation* s of X by:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \qquad s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}.$$

(b) Suppose the data points in X take on only m **distinct** values; labeled $y_1, y_2, y_3, \ldots, y_m$. Further, let f_j denote the number of times that a given value y_j occurs in X, for $1 \leq j \leq m$. (That is: various different x_k's can have the same value y_j; we denote number of x_k's that do so by f_j. So $f_1 + f_2 + \cdots + f_m =$ the total number of data points in $X = n$.) Then the above quantities $\bar{x}$ and s can be computed using the following formulas:

$$\bar{x} = \frac{f_1 y_1 + f_2 y_2 + \ldots + f_n y_n}{n}, \qquad s = \sqrt{\frac{f_1(y_1 - \bar{x})^2 + f_2(y_2 - \bar{x})^2 + \ldots + f_n(y_n - \bar{x})^2}{n-1}}.$$

The mean $\bar{x}$ is considered a measure of the *average*, or *center*, or *central tendency*, of the data in the set X. To see that this is a reasonable way to think of the mean, let's consider the formula for $\bar{x}$ given in part (a) of the above definition What this formula tells us is this: Suppose you have n perhaps unequal parts, of sizes $x_1, x_2, \ldots, x_n$, which constitute a whole of size $x_1 + x_2 + \ldots + x_n$. If you want to divide this whole into n *equal* parts, then each part must have size $\bar{x}$.

Similarly, the standard deviation s may be considered a measure of the *spread* of the data in X. The rationale for this way of thinking comes essentially from the quantities $(x_k - \bar{x})^2$ appearing in the above (first) definition of s. The idea here is that the magnitude of $x_k - \bar{x}$ tells us how far

the data point x_k is from the center $\bar{x}$ of the data; so adding up all the $(x_k - \bar{x})^2$'s gives us a sense of how far the data is *collectively* from this center.

The squaring of each $x_k - \bar{x}$, in our definition of s, ensures that all of our summands are positive, so that negative terms won't cancel out positive ones. We divide by $n - 1$ to "level the playing field" among data sets of different sizes, so that adding data points to a set does not automatically increase its standard deviation. (In some definitions of s, the division is by n rather than $n - 1$. This is for technical reasons that we will not discuss here.) And finally, we take the square root at the end to compensate, in some sense, for the squaring that we applied to each $x_k - \bar{x}$.

There is another measure of spread in a data set called *mean absolute deviation.* This definition looks like the above definition of s, except that $(x_k - \bar{x})^2$ is replaced by $|x_k - \bar{x}|$ for each k, and no square root is taken at the end. The primary advantage of standard deviation over mean absolute deviation is that the latter entails absolute values, which are not differentiable everywhere. This makes calculus much harder to apply, and makes mean absolute deviation unwieldy from a mathematical perspective.

To highlight the above definitions, and in particular, how they work for data that's "grouped" into distinct values, let's return to our "six coins, flipped 3,444 times" data. The data set X in this case has size $n = 3{,}444$; a data point x_k tell us how many heads were observed on a particular trial (say, the kth trial) of the flipping experiment.

Of course, each x_k must be an integer from 0 to 6. That is, our data groups into distinct values $y_1 = 0$, $y_2 = 1$, $y_3 = 2$, and so on. The frequency f_j with which each of these y_j's occurs is given by Table 1. We may therefore use the formulas from part (b) of Definition 1.1.1 to compute $\bar{x}$ and s, as follows:

$$\bar{x} = \frac{(66 \times 0) + (341 \times 1) + \cdots + (333 \times 5) + (51 \times 6)}{3444} = 2.96922,$$

$$s = \sqrt{\frac{66(0 - \bar{x})^2 + 341(1 - \bar{x})^2 + \cdots + 333(5 - \bar{x})^2 + 51(6 - \bar{x})^2}{3443}} = 1.24663.$$

It's not surprising that our computed value of $\bar{x}$ is close to 3. The mean measures central tendency, or average, and if we flip six (fair) coins, then we would expect that, on the "average" flip, half of those coins (that is, three of them) should come up heads.

There is no equally simple, intuitive interpretation of standard deviation. However, there is a good "reality check" on the number we obtained for s above. Namely: it's known that, if data has an approximate normal distribution, then all or nearly all of that data should fall *within three standard deviations of the mean.* In mathematical terms this means that, for such a distribution, most or all of the data lies in the interval $(\bar{x} - 3s, \bar{x} + 3s)$.

Is this the case for our coin-flip data set X? Here, we have

$$(\bar{x} - 3s, \bar{x} + 3s) = (2.96922 - 3 \times 1.24663, 2.96922 + 3 \times 1.24663) = (-0.77067, 6.70911).$$

Since all data values lie between 0 and 6 inclusive, this interval *does*, in fact, capture all of the data – and does so without too much room to spare. That is, the interval does not overshoot the

actual range of data values by much. All of this tells us that our computed value of s is at least in the right ballpark.

A different kind of histogram

A cornerstone of probability theory is the interpretation of probability as an *area under a graph*. Such an interpretation allows the full force of Calculus – Riemann sums, antiderivatives, the Fundamental Theorem, and so on – to be brought to bear on the study of probability.

Histograms, as considered above, are a first step towards realizing this interpretation. To go further, we will next need to *rescale*, or *renormalize*, these histograms, through the concept of *relative frequency density* (also called *probability density*). Here's the definition.

Definition 1.1.2. Let X be a data set, consisting of n real number data points. Consider any *bin*, meaning simply an interval of real numbers. Let f denote the number of data points in X that lie in the bin, and let B denote the length of the bin. Then we define the *relative frequency density* of the bin, denoted RFD, by the formula

$$\text{RFD} = \frac{f}{B \times n}. \tag{1.1.1}$$

In short, the relative frequency density of a bin is the number of data points in that bin, divided by the product of the length of the bin and the size of the data set.

Before discussing the importance of relative frequency density, we make the definition concrete by means of an example.

Example 1.1.1. Consider a data set X of $n = 68$ exam scores, distributed as follows.

Bin	F	RFD $= F/(B \times 68)$
[40,60)	11	$11/(20 \times 68) = 0.0081$
[60,70)	16	$16/(10 \times 68) = 0.0235$
[70,80)	14	$14/(10 \times 68) = 0.0206$
[80,85)	13	$13/(5 \times 68) = 0.0382$
[85,90)	9	$9/(5 \times 68) = 0.0265$
[90,100)	5	$5/(10 \times 68) = 0.0074$

We can draw a *relative frequency density histogram*, which is like the histograms drawn above, but now, the vertical axis denotes relative frequency density.

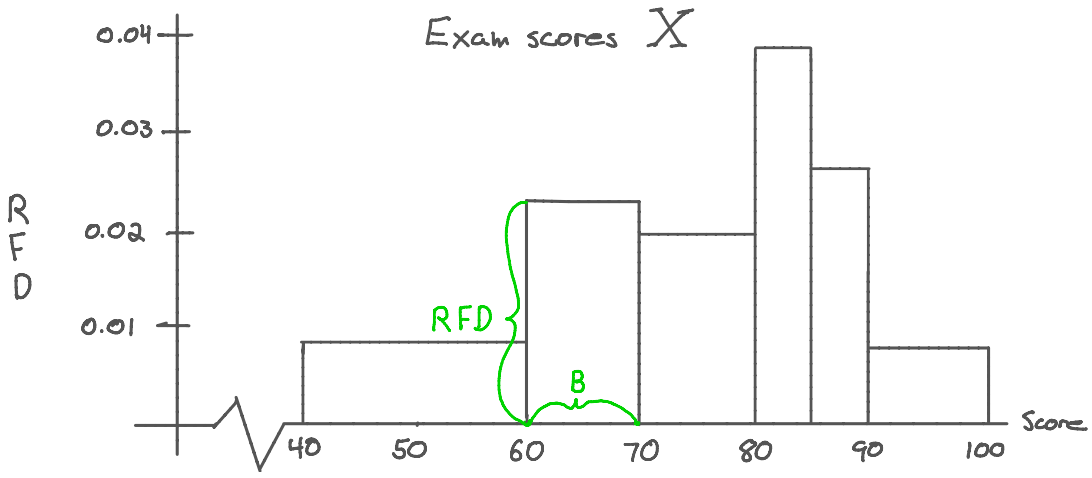


Figure 3. A relative frequency density histogram

Let's now see why relative frequency density is such a useful construct. To do this, we take the above definition (1.1.1) of RFD, and multiply both sides by B, to get

$$\text{RFD} \times B = \frac{f}{n}. \tag{1.1.2}$$

Let's look carefully at both sides of equation (1.1.2). The left-hand side gives the *height* times the *baselength* – that is, the *area* – of the RFD histogram "bar" that lies over the range in question. See Figure 3 above.

The quantity f/n on the right-hand side of (1.1.2) gives the number of points in X that lie in the given bin, divided by the *total* number of points in X. This quotient is just the *fraction*, or *proportion*, of the data that lies in the given bin. Or, put differently: f/n is the *probability* that a data point in X, chosen at random, lies in the given range.

Let's write $P(a \leq X < b)$ to denote the probability that a data point in X, chosen at random, lies in the interval $[a,b)$. Since the two sides of (1.1.2) are, in fact, equal, we then have the following conclusion.

In a relative frequency density histogram, the area of a bar over an interval $[a,b)$ equals $P(a \leq X < b)$.

Relationship between area and probability, in an RFD histogram

The bottom line is this: by plotting relative frequency density (rather than just frequency) on the vertical axis, we obtain histograms where area represents probability. This is a powerful idea, which we will exploit more fully in the next section.

In the meantime, we note that this idea can be applied "several bars at a time." For example,

using the data and histogram from Example 1.1.1 above, we can compute that

$$\begin{aligned} P(60 \leq X < 85) &= \text{area enclosed by the bars covering the range [60,85)} \\ &= \text{sum of areas of bars over [60,70), [70,80), and [80,85)} \\ &= \text{sum of (height times baselength) of these bars} \\ &= \text{sum of (RFD times baselength) of these bars} \\ &= (0.0235 \times 10) + (0.0206 \times 10) + (0.0382 \times 5) = 0.6320. \end{aligned}$$

That is, 63.2% of the exam scores lie in the interval [60,85).

Of course, we could have argued more simply. Specifically, we could have used the relative frequency density table above to conclude that the proportion of data in [60,85) is $(16+14+13)/68 = 0.632353$. Example 1.1.1 helps to illustrate the mechanics of Definition 1.1.1, but the real *value* of relative frequency density will not be seen until the next section, where we use this construct in contexts where we can't simply "count data points."

In any case, our computation of $P(60 \leq X < 85)$ has taken advantage of the fact that the interval [60,85) is precisely spanned by three of our given bins. Were this not the case, we might only be able to *approximate* probabilities. For example, given the above information concerning our set X of exam scores, we might estimate that

$$\begin{aligned} P(63 \leq X < 82) &\approx \text{area enclosed by the bars covering the range [63,82)} \\ &= \text{sum of areas of bars over [63,70), [70,80), and [80,82)} \\ &= \text{sum of (height times baselength) of these bars} \\ &= \text{sum of (RFD times baselength) of these bars} \\ &= (0.0235 \times 7) + (0.0206 \times 10) + (0.0382 \times 2) = 0.4469. \end{aligned}$$

So approximately 44.69% of the exam scores lie in the interval [63,82).

The above answer is approximate because we don't know that the exam scores are evenly distributed across each bin. For example, knowing that 16 data points lie in the interval [60,70), of length 10, clearly does not imply that $0.7 \times 16 = 11.2$ data points lie in the interval [63,70), of length 7.

We could get better estimates if we had narrower bins. In the next section, we'll consider bins that can, at least in theory, be made arbitrarily thin. This will lead us to the study of *probability density functions*, which are central to probability theory and statistical inference.

1.2 Statistical inference

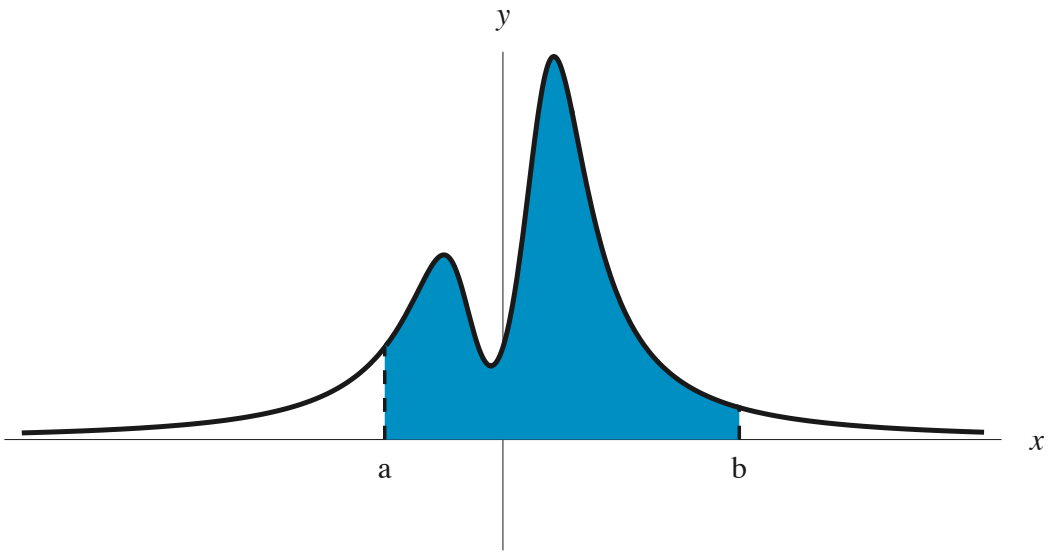
Please note that this section is in "DIY" (Do It Yourself) format: there are various blanks to be filled in, and questions to be answered, *along the way*, rather than being left to a set of exercises at the end. (SOLUTIONS IN RED)

A. Random variables and pdf's

Let X be a random variable, meaning, essentially, a way of assigning a real number to each possible outcome of an experiment. We say that X has *probability density function*, or *pdf*, given by $f(x)$ if

$$P(a < X < b) = \int_a^b f(x)\,dx$$

for any numbers a and b in the domain (set of possible values) of X. (Again, $P(a < x < b)$ denotes the probability that, if a value x is chosen from X at random, that value will lie between the numbers a and b.)



Exercise A1. Fill in the blanks: the mean μ and standard deviation σ of a pdf $f(x)$ can be obtained as follows. Draw a relative frequency ___density___ histogram corresponding to a sample of points from X. Compute the ___mean___ $\overline{x}$ and the ___standard deviation___ s of the histogram data. Repeat for larger and larger samples, using narrower and narrower bin widths. Then the tops of the bars of the histogram will smooth out to give you the graph of your pdf $y =$ ___$f(x)$___, and your numbers $\overline{x}$ and s will converge to ___μ___ and ___σ___, respectively.

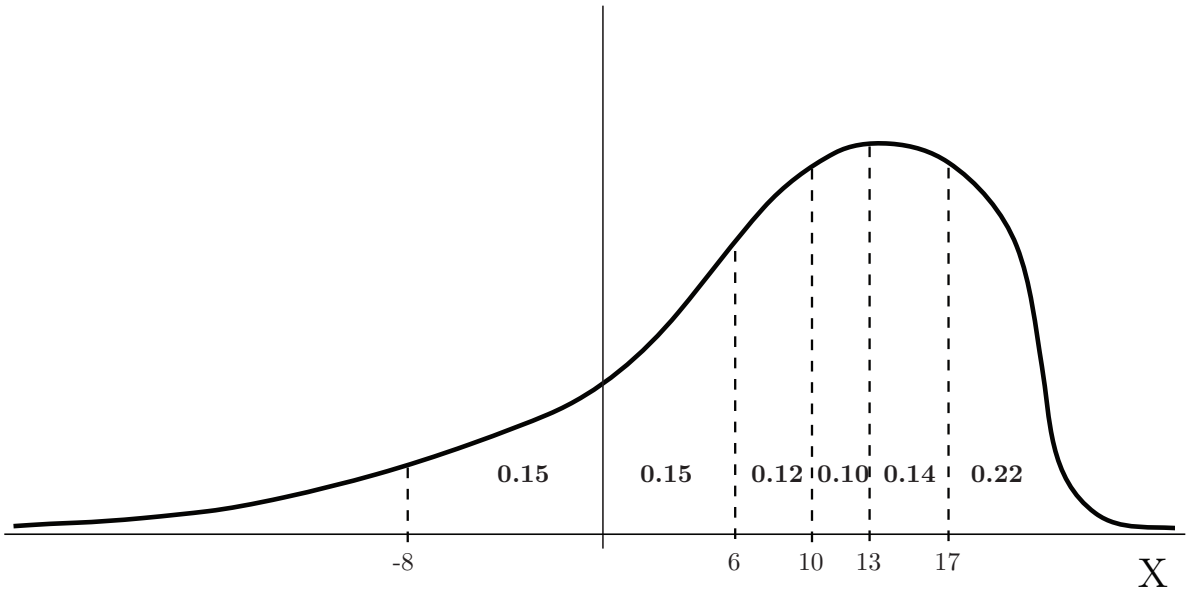
Also recall that, for any pdf $f(x)$, with domain (c,d), we must have

- $f(x) \geq 0$ for all x in (c,d) (since probabilities can't be negative), and

- $\int_c^d f(x)\,dx = 1$ (since the probability that a data point in X lies somewhere in X must equal 100%, or 1).
- $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$ for all a and b in (c,d) (probabilities are the same whether or not you include endpoints, since the area under a single point on the graph of a function is zero).

Remark. Often, the domain (c,d) of a pdf will be taken to be of *infinite* extent, meaning $c = -\infty$ or $d = +\infty$, or both.

Exercise A2. Consider the following probability density function for a random variable X. The regions delineated by dashed lines have areas as shown.



Find:

(a) $P(X < -8)$. 0.12

(b) $P(10 < X < 17)$. 0.24

(c) The number c such that 36% of all data values of X are at least 6 and at most c. $c = 17$

B. Standard normal random variables and pdf's

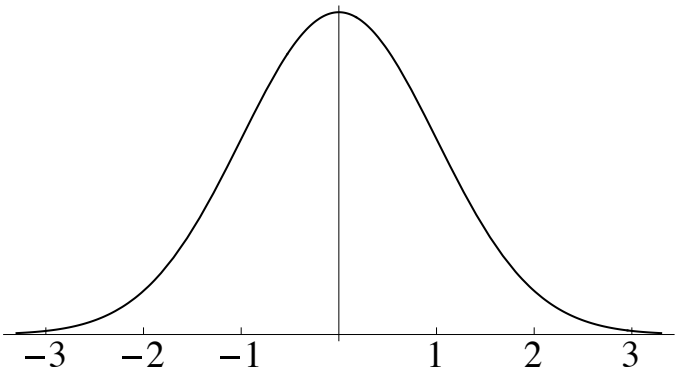
A random variable X is said to have a *standard normal distribution* if the pdf for X is given by

$$f_{0,1}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

(where x can be any real number).

For such an X, we say "X is $N(0,1)$." In other words, to say X is $N(0,1)$ is to say that, for any real numbers a, b with $a < b$,

$$P(a < X < b) = \int_a^b f_{0,1}(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx.$$



FACT: The pdf $f_{0,1}(x)$ has mean $\mu = 0$ and standard deviation $\sigma = 1$. (That's why we call it $f_{0,1}(x)$.)

Exercise B1. Fill in the blanks:

(a) It can be computed that $\int_{-1}^{1} f_{0,1}(x)\,dx \approx 0.683$. In other words: if X is $N(0,1)$, then about __68.3__% of the values of X lie within one __standard deviation__ of the mean.

(b) It can be computed that $\int_{-2}^{2} f_{0,1}(x)\,dx \approx 0.955$. In other words: if X is $N(0,1)$, then about __95.5__% of the values of X lie within __two__ standard deviations of the mean.

(c) It can be computed that $\int_{-3}^{3} f_{0,1}(x)\,dx \approx 0.997$. In other words: if X is $N(0,1)$, then about __99.7__% of the values of X lie within three __standard deviations__ of the mean.

(d) It can be computed that $\int_{-1.96}^{1.96} f_{0,1}(x)\,dx \approx 0.950$. In other words: if X is $N(0,1)$, then about __95__% of the values of X lie within __1.96__ standard deviations of the mean.

(e) It can be computed that $\int_{-2.33}^{2.33} f_{0,1}(x)\,dx \approx 0.980$. In other words: if X is $N(0,1)$, then about __98__% of the values of X lie within __2.33__ standard deviations of the mean.

(f) It can be computed that $\int_{-2.58}^{2.58} f_{0,1}(x)\,dx \approx 0.990$. In other words: if X is $N(0,1)$, then about __99__% of the values of X lie within __2.58__ standard deviations of the mean.

C. Other normal random variables and pdf's

If the pdf for a random variable Z follows the basic shape of the standard normal curve, but has mean μ (instead of 0) and standard deviation σ (instead of 1), we say "Z is $N(\mu,\sigma)$." Let's denote

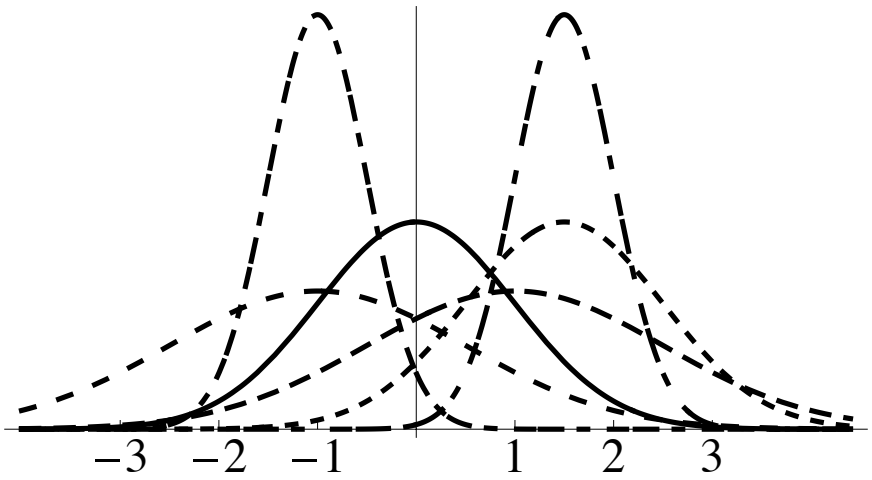
such a pdf by $f_{\mu,\sigma}(x)$. Then: to say Z is $N(\mu,\sigma)$ is to say that, for any real numbers a, b with $a < b$,

$$P(a < Z < b) = \int_a^b f_{\mu,\sigma}(x)\,dx.$$

The precise formula for $f_{\mu,\sigma}(x)$ is

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

But *we won't need* this formula, because we are going to *translate* $N(\mu,\sigma)$ variables to $N(0,1)$ variables, shortly.



Exercise C1. Recall that the mean of a data set measures the "central tendency," and that the standard deviation measures the "spread" (large standard deviation means large spread, and conversely). Given all this, and also assuming that the solid curve on the graph above is $N(0,1)$, identify, on the above graph, which of the dashed curves is $N(1.5,1)$; which is $N(1,1.5)$; which is $N(1.5,0.5)$; which is $N(-1,1.5)$; and which is $N(-1,0.5)$. Please explain your reasoning briefly, in the space below.

The two tallest curves are the least spread out, so they must have the smallest of the standard deviations, which is 0.5. The leftmost of these taller curves is centered at -1 and therefore has mean -1; the rightmost, similarly, has mean 1.5. Similar arguments apply to the other curves.

D. Translation between $N(0,1)$ variables and $N(\mu,\sigma)$ variables

We have the following NISNID ("Normal Is Standard Normal In Disguise") Fact, which we present without proof (but which is not hard to show, using the above formula for the $N(\mu,\sigma)$ pdf $f_{\mu,\sigma}(x)$):

NISNID Fact. If Z is $N(\mu,\sigma)$, then $\dfrac{Z-\mu}{\sigma}$ is $N(0,1)$.

In stats texts, you will typically find tables of $N(0,1)$ variables, but not other $N(\mu,\sigma)$ variables. Now we know why: we can *compute* probabilities associated with $N(\mu,\sigma)$ random variables if *all we know* are probabilities associated with $N(0,1)$ random variables.

Here's an example showing how.

Example. Suppose Z is $N(8,1.5)$. Find $P(5 < Z < 11)$.

Solution. Since, in this case, $\mu = 8$ and $\sigma = 1.5$, we have

$$\begin{aligned} P(5 < Z < 11) &= P\left(\frac{5-8}{1.5} < \frac{Z-8}{1.5} < \frac{11-8}{1.5}\right) \\ &= P\left(-2 < \frac{Z-8}{1.5} < 2\right) = 0.955. \end{aligned}$$

The last step is by the NISNID Fact, and by exercise **B1**(b) above. Using the strategy of the above example (and using part **B** above where necessary), complete the following exercises.

Exercise D1. Suppose Z is $N(-2,0.3)$. Find $P(-2.3 < Z < -1.7)$.

$$\begin{aligned} P(-2.3 < Z < -1.7) &= P\left(\frac{-2.3-(-2)}{0.3} < \frac{Z-(-2)}{0.3} < \frac{-1.7-(-2)}{0.3}\right) \\ &= P\left(-1 < \frac{Z-(-2)}{0.3} < 1\right) = 0.683. \end{aligned}$$

Exercise D2. Suppose Z is $N(2,2)$. Find $P(-1.92 < Z < 5.92)$.

$$\begin{aligned} P(-1.92 < Z < 5.92) &= P\left(\frac{-1.92-2}{2} < \frac{Z-2}{2} < \frac{5.92-2}{2}\right) \\ &= P\left(-1.96 < \frac{Z-2}{2} < 1.96\right) = 0.95. \end{aligned}$$

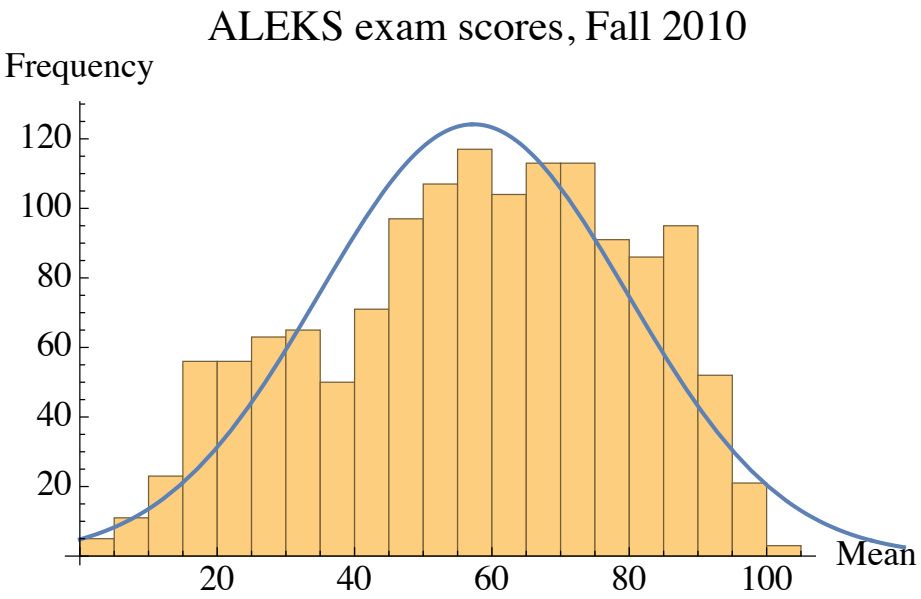
Exercise D3. Suppose Z is $N(\mu,\sigma)$. Find $P(\mu - 3\sigma < Z < \mu + 3\sigma)$.

$$P(\mu - 3\sigma < Z < \mu + 3\sigma) = P\left(\frac{\mu - 3\sigma - \mu}{\sigma} < \frac{Z - \mu}{\sigma} < \frac{\mu + 3\sigma - \mu}{\sigma}\right)$$
$$= P\left(-3 < \frac{Z - \mu}{\sigma} < 3\right) = 0.997.$$

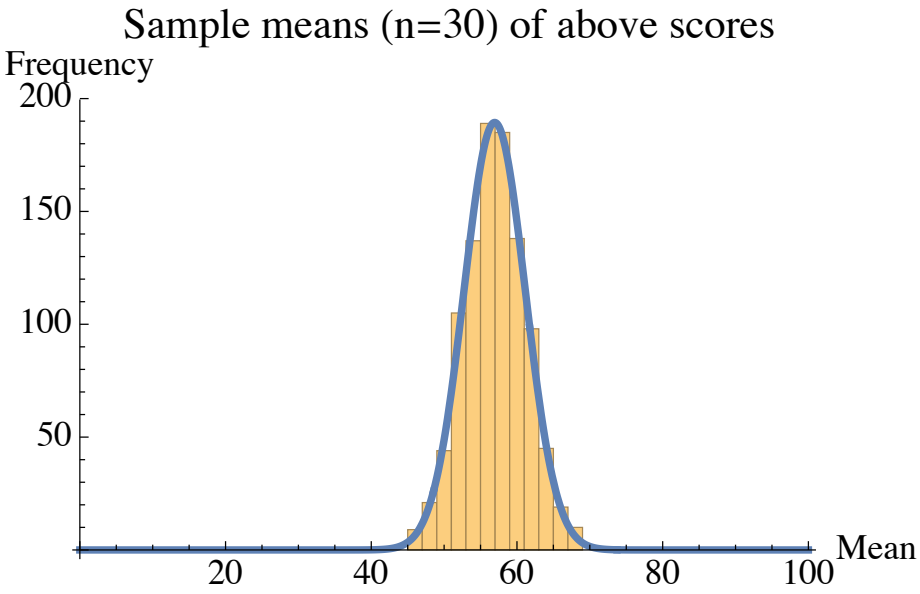
Exercise D4. What exercise **D3** directly above says is: if Z is *any* normal random variable, then __99.7__% of the data lies within three standard __deviations__ of the __mean__.

E. The sampling distribution of the mean

Consider a dataset of scores on the ALEKS exam taken by 1,399 CU students, at the start of the Fall 2010 semester. Here's a histogram for the data (with a normal curve fit to the data as well as possible):



Note that the data is not especially bell-shaped (well, it's kind of like a "skewed" bell). But now, let's do something a bit different. Let's choose a *random sample* of 30 ALEKS scores $x_1, x_2, x_3, \ldots, x_{30}$ out of the 1,399, and compute the mean $\overline{x} = (x_1 + x_2 + x_3 + \cdots + x_{30})/30$. Actually, let's do this *many, many times*, to get a whole *bunch* of sample means $\overline{x}$ (all corresponding to the same sample size $n = 30$). Here is a histogram (and a best-fit normal curve) for a large set of sample means that we obtained in this way (with the help of Mathematica).



Exercise E1. Fill in the blanks: the mean of the above set of 1,399 ALEKS exams scores looks like it's roughly ____equal____ to the mean of the above set of sample means. (Both numbers look like they're somewhere around 57 or so.) However, the standard deviation of the sample means dataset looks much ____smaller____ than that of the original dataset, because the sample means seem much ____less____ spread out (that is, they seem ____more____ tightly clustered about the central value).

Also, even though the original ALEKS data is only very roughly normal in shape, the mean score data fits a ____normal____ curve more closely.

The above observations exemplify the following theorem, which is called The Sampling Distribution of the Mean, or SDM. This result follows from the Central Limit Theorem, and is critical to "hypothesis testing" and "confidence intervals" (which we'll study in the remainder of this assignment).

Theorem (The Sampling Distribution of the Mean, or SDM). Let X be a (*not necessarily normal*) random variable, with mean μ and standard deviation σ. Fix a sample size n, and assume n is at least 30. Then the random variable $\overline{X}$ consisting of means $\overline{x}$ of *all possible* random samples of X, of size n, *IS* roughly normal, with mean $\overline{\mu} = \mu$ and standard deviation $\overline{\sigma} = \sigma/\sqrt{n}$. That is, for such X, $\overline{X}$ is roughly $N(\overline{\mu}, \overline{\sigma}) = N(\mu, \sigma/\sqrt{n})$.

Exercise E2. Our original "random variable" X of ALEKS scores data has mean $\mu = 56.81$ and standard deviation $\sigma = 22.47$. Based on the above Theorem, what are the mean $\overline{\mu}$ and standard deviation $\overline{\sigma}$ of the set $\overline{X}$ of all possible sample means of ALEKS scores (for samples of size $n = 30$)?

$\overline{\mu} = \mu = 56.81$; $\overline{\sigma} = \sigma/\sqrt{n} = 22.47/\sqrt{30} = 4.10$

Remark. There are roughly $\binom{1,399}{30} \approx 6.52941 \cdot 10^{61}$ *possible* samples of size 30 from a set of size 1,399. We couldn't possibly compute the mean for every one of these samples! (Unless, for example, we were to start at the beginning of the universe, and compute a billion billion sample means every billionth of a billionth of a second. If we had a billion people working on this simultaneously, all working on different samples, we could do it somewhere around seven times. But let's not.)

For the above histogram of sample means, we computed considerably fewer means – about 1,000, in fact. A thousand is a lot smaller than $6.52941 \cdot 10^{61}$, but it's large enough to give us a good qualitative idea of what's going on.

F. Hypothesis testing of a population mean

Here's the BIG IDEA. Suppose we have some population, represented by a random variable X. Suppose that, in the absence of any compelling evidence to the contrary, we are willing to accept that the mean μ of X is (more or less) equal to some known, specified number μ_0. The question is: what, mathematically speaking, might constitute "compelling evidence to the contrary"?

FOR EXAMPLE: Suppose we know, because of a long history of experimentation and practice, that the average lifespan of a rat is 684 days. Suppose we now administer a restricted diet to a group of 105 rats. Let's assume (although such things are almost never really true in practice) that these 105 rats represent a random sample of the population of all rats who could conceivably receive this restricted diet.

Exercise F1. Fill in the blanks: the burden of proof is on us to show that the restricted diet has any pronounced effect compared to an unrestricted diet. So, until proven otherwise, we assume that the two diets are essentially the same. That is, we're assuming that the mean μ of survival times X of the population of all rats getting the restricted diet is given by $\mu =$ 684 (in days) (your answer should be a *number*).

Suppose this assumption is true. Suppose we also know, somehow, that our survival times X for all rats on the restricted diet have standard deviation $\sigma = 286$. (Remark: in practice, you will almost never know the population standard deviation σ directly; if you did, then most likely, you'd know the mean μ as well, and you wouldn't have to hypothesize about it, and you'd be done. So in practice, one often lets the standard deviation s of the *sample* stand in for σ. In fact, that's what we've done here.) Then we know, by the SDM Theorem from part **E** above, and the fact that our sample size $n = 105$ is at least 30, that the random variable $\overline{X}$ of *sample means* from this population, for samples of size 105, will have a normal pdf, with mean

$$\overline{\mu} = \mu = \text{684} \text{ (fill in a number)}$$

and standard deviation

$$\begin{aligned} \overline{\sigma} = \sigma/\sqrt{n} &= \frac{286}{\sqrt{105}} \text{ (fill in the correct numbers for } \sigma \text{ and } n) \\ &= 27.9 \text{ (compute } \overline{\sigma}). \end{aligned}$$

But then we know, by the NISNID Fact of part **D** above, that the random variable $\dfrac{\overline{X} - \overline{\mu}}{\overline{\sigma}}$ is

standard normal – that is, this variable is $N($____0____,____1____$)$. This tells us, by part (f) of exercise **B1** above, that 99% of all possible values of the random variable $\frac{\overline{X}-\overline{\mu}}{\overline{\sigma}}$ fall between the numbers ____-2.58____ and ____2.58____.

In particular, suppose we actually *compute* a sample mean $\overline{x}$ from a sample of X, of size $n = 105$, and find that $\frac{\overline{x}-\overline{\mu}}{\overline{\sigma}}$ is *not* between the above two numbers. Well, by the above paragraph, this is pretty unlikely, if it's really true that $\mu = 684$. SO, in such a situation, we might conclude that μ is *not* equal to 684. That is: in this particular case, we would *reject* the "null hypothesis" $H_0 : \mu = 684$, and accept the "alternative hypothesis" $H_A : \mu \neq 684$, meaning we'd accept the conclusion that the restricted diet leads to substantially *different results* than the unrestricted diet. Also, we'd say that we accepted this alternative hypothesis "at the 99% level." What this means is: there's at most a 1% chance (1%=100%-99%) that we'd get sample data this far away from the hypothesized mean, if this hypothesized mean of 684 really were the true mean.

Let's wrap this up with a particular case study (actual data collected from a 1988 experiment). In this study, the mean lifespan, in days, of a group of 105 rats given the restricted diet was $\overline{x} = 968$. The standard deviation s was 286, as alluded to above. Question: is this enough for us to accept, at the 99% level, the *alternative* hypothesis that the restricted diet yields lifespans significantly different from those of the unrestricted diet? To answer:

Exercise F2. Compute $\frac{\overline{x}-\overline{\mu}}{\overline{\sigma}}$, for this particular value of $\overline{x}$ and for the $\overline{\mu}$ and $\overline{\sigma}$ computed in exercise **F1** above.

$$\frac{\overline{x}-\overline{\mu}}{\overline{\sigma}} = \frac{968-684}{27.9} = 10.175.$$

Exercise F3. Is the number you computed in exercise **F2** above between -2.58 and 2.58? Based on your answer to this, do we reject the null hypothesis $H_0 : \mu = 684$, and accept the alternative hypothesis $H_A : \mu \neq 684$, at the 99% level? Or do we *not* reject the null hypothesis? Please explain. No it's not. So we *reject* the null hypothesis, and *accept* the alternative hypothesis, at the 99% level.

Exercise F4. In general (that is, considering again any general population, not necessarily that of exercises **F1**–**F3** above), how would your test *change* if you wanted to test the null hypothesis at the 95% level, or the 98% level, instead of the 99% level? Hint: you only need to change the numbers you're comparing things to in exercise **F3** above.

Compare

$$z = \frac{\overline{X} - \overline{\mu}}{\overline{\sigma}}$$

to 1.96 or to 2.33, instead of 2.58.

Exercise F5. Back to our rats: Based on your answer to exercise **F4** above, do we reject the above null hypothesis $H_0 : \mu = 684$, and accept the alternative hypothesis $H_A : \mu \neq 684$, at the 95% level? At the 98% level? Please explain. Yes. If we reject the null hypothesis at the 99% level, we will certainly reject it at any lower level.

G. Confidence intervals for a population mean

In Section **F** above, we considered the question: Is the mean μ of a certain random variable X equal to a certain, given, "hypothesized" number μ_0? (Or, perhaps more accurately: is there enough evidence to conclude that μ is *not* equal to μ_0?) In this section we ask a slightly different – and, some would say, more plausible and useful – question, namely: within what *range* of values can we say, with a reasonable degree of confidence, a certain population mean lies? In other words, we investigate how, based on sample data, we can say things like "We are 95% confident that the mean μ of our population lies between this number and that number."

The procedure for arriving at such statements is relatively straightforward. It consists of five steps, as delineated below. **Please note:** we are going to assume, in outlining these steps, a "95% confidence level." We'll explain what this means, and will consider how to proceed for different confidence levels, a bit later.

Exercise G1: fill in the blanks.

STEP 1. Take a sample of values of X, with sample size n, where n is at least 30. Compute the mean ___$\overline{x}$___ and ___standard deviation___ s of this sample.

STEP 2. We know, from the SDM Theorem of section **E** above, that the sample means $\overline{X}$ are $N(\overline{\mu},\overline{\sigma})$, so that, by the NISNID Fact of section **D** above,

$$\frac{\overline{X}-\overline{\mu}}{\overline{\sigma}} \quad \text{is} \quad N(\underline{\quad 0 \quad},\underline{\quad 1 \quad}).$$

Therefore, a randomly chosen sample mean $\overline{x}$ satisfies (by exercise **B1(d)** above):

$$P\left(-1.96 < \frac{\overline{x}-\overline{\mu}}{\overline{\sigma}} < 1.96\right) = \underline{\quad 0.95{=}95\% \quad}.$$

If we multiply everything in parentheses through by $\overline{\sigma}$, and then subtract $\overline{x}$ from all terms in parentheses, we get

$$P\left(-\overline{x} - 1.96\,\overline{\sigma} < -\overline{\mu} < -\overline{x} + 1.96\,\overline{\sigma}\right) = 0.95 = 95\%.$$

Multiplying everything in parentheses by -1 (and remembering that mutliplying by a negative number switches the direction of an inequality), we get

$$P\left(\overline{x} + 1.96\,\overline{\sigma} > \overline{\mu} > \overline{x} - 1.96\,\overline{\sigma}\right) = 0.95 = 95\%.$$

Finally, just reverse the order in which the stuff in parentheses is written, to get

$$P\left(\overline{x} - 1.96\,\overline{\sigma} < \overline{\mu} < \overline{x} + 1.96\,\overline{\sigma}\right) = \underline{\quad 0.95{=}95\% \quad}. \qquad (*)$$

STEP 3. Now recall, from the SDM, the formulas for $\overline{\mu}$ and $\overline{\sigma}$ in terms of μ, σ, and n:

$$\overline{\mu} = \underline{\quad \mu \quad} \quad \text{and} \quad \overline{\sigma} = \underline{\quad \frac{\sigma}{\sqrt{n}} \quad}.$$

So equation $(*)$ can be rewritten:

$$P\left(\overline{x} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = \underline{\quad 0.95{=}95\% \quad}. \qquad (**)$$

Or in other words: there is a ___95___% chance that, if a mean $\overline{x}$ is computed from a random sample of size n, then μ will lie between $\overline{x} - 1.96\,\frac{\sigma}{\sqrt{n}}$ and $\overline{x}+$ ___$1.96\,\frac{\sigma}{\sqrt{n}}$___.

STEP 4. Now as discussed in part **F** above, we typically don't know the population standard deviation σ, so we *approximate* it with s, which is the sample ___standard deviation___. Then equation $(**)$ reads:

$$P\left(\overline{x} - 1.96\,\frac{s}{\sqrt{n}} < \mu < \overline{x} + 1.96\,\frac{s}{\sqrt{n}}\right) \approx \underline{\quad 0.95{=}95\% \quad}.$$

STEP 5. Because of the reasoning outlined above, we call the interval

$$\left(\overline{x} - 1.96\,\frac{s}{\sqrt{n}}, \overline{x} + 1.96\,\frac{s}{\sqrt{n}}\right)$$

a *95% confidence interval* for the population mean μ.

Exercise G2. Here are the "navel ratios," meaning the ratios

$$\frac{\text{height}}{\text{VUD}}$$

(VUD stands for "vertical umbilical displacement," or belly-button height) of a random (well, not really random, but let's pretend) sample of 48 CU students.

1.60	1.60	1.56	1.63	1.62	1.63	1.65	1.65	1.65	1.67	1.68	1.63
1.60	1.66	1.59	1.64	1.61	1.65	1.62	1.64	1.67	1.56	1.58	1.58
1.58	1.70	1.59	1.61	1.67	1.63	1.58	1.57	1.67	1.66	1.67	1.63
1.68	1.59	1.55	1.54	1.60	1.60	1.66	1.58	1.66	1.66	1.65	1.61

(a) Find the mean $\overline{x}$ and standard deviation s of the above navel ratio data. Write your answers to three decimal places.

$\overline{x} = 1.623$, $s = 0.040$

(b) Use the information from part (a) above to construct a 95% confidence interval for the mean navel ratio μ of all CU students.

$$\left(\overline{x} - 1.96\,\frac{s}{\sqrt{n}}, \overline{x} + 1.96\,\frac{s}{\sqrt{n}}\right) = \left(1.623 - 1.96 \cdot \frac{0.040}{\sqrt{48}}, 1.623 + 1.96 \cdot \frac{0.040}{\sqrt{48}}\right)$$
$$= (1.612, 1.634).$$

Exercise G3. In general (that is, considering again any general population, not necessarily that of exercise **G2** above), suppose you wanted, instead of the 95% interval of **STEP 5** above, a **98%** confidence interval. How would the interval described in **STEP 5** above change? In other words, what would a 98% confidence interval for μ look like, in terms of $\overline{x}$, s, and n? What about a 99% confidence interval? Please explain. Hint: consider exercises **B5** and **B6** above.

The 98% and 99% confidence intervals are, respectively,

$$\left(\overline{x} - 2.33\,\frac{s}{\sqrt{n}}, \overline{x} + 2.33\,\frac{s}{\sqrt{n}}\right)$$

and

$$\left(\overline{x} - 2.58\,\frac{s}{\sqrt{n}}, \overline{x} + 2.58\,\frac{s}{\sqrt{n}}\right).$$

Exercise G4. Construct 98% and 99% confidence intervals for the mean navel ratio μ of all CU students. The intervals are (1.610,1.636) and (1.608,1.638) respectively.

Exercise G5. Using the sample data from part **F** above, construct 95%, 98%, and 99% confidence intervals for the mean survival time μ of rats fed the restricted diet described there.

$$95\% : \quad \left(\overline{x} - 1.96\,\frac{s}{\sqrt{n}}, \overline{x} + 1.96\,\frac{s}{\sqrt{n}}\right) = (913.3,1022.7).$$

$$98\% : \quad \left(\overline{x} - 2.33\,\frac{s}{\sqrt{n}}, \overline{x} + 2.33\,\frac{s}{\sqrt{n}}\right) = (903.0,1033.0).$$

$$99\% : \quad \left(\overline{x} - 2.58\,\frac{s}{\sqrt{n}}, \overline{x} + 2.58\,\frac{s}{\sqrt{n}}\right) = (896.1,1039.9).$$

Exercise G6. One theory says that, on average, in many populations, the "navel ratio" studied in the above exercises is about equal to the "golden ratio," which equals $(1+\sqrt{5})/2 \approx 1.618$.

Test this theory, at the 99% level, for the population of all CU students, using the above navel ratio data. Use the procedure outlined in part **F** of this section. Make sure to state clearly your null and alternative hypotheses.

$$H_0 \colon \mu = 1.618$$
$$H_A \colon \mu \neq 1.618$$
$$z = \frac{\overline{x} - \mu_0}{(s/\sqrt{n})} = \frac{1.623 - 1.618}{(0.040/\sqrt{48})} = 0.866.$$

Since $|0.866| < 2.58$, we do *not* reject H_0 at the 99% level.