6.2 Statistical inference

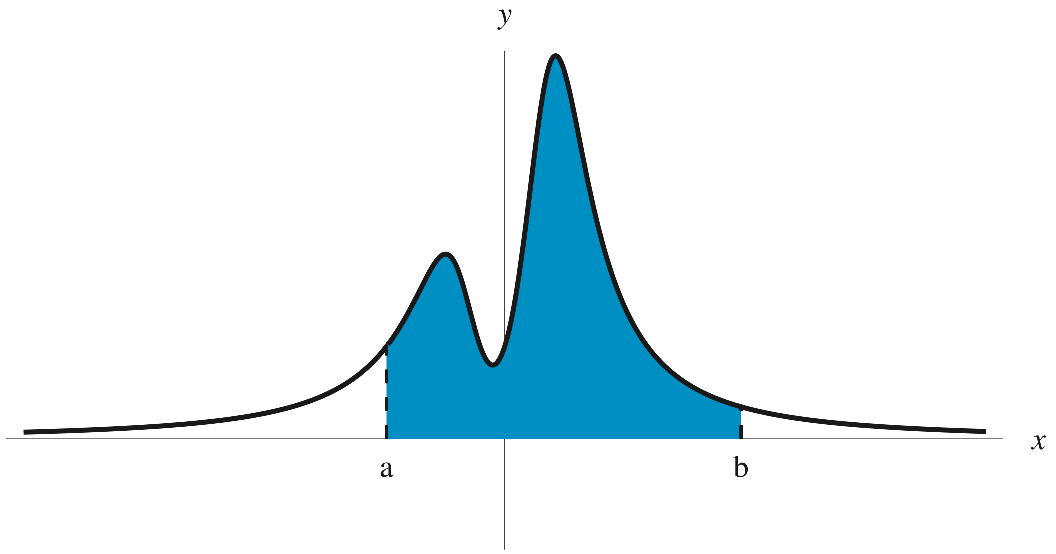
Please note that this section is in “DIY” (Do It Yourself) format: there are various blanks to be filled in, and questions to be answered, *along the way*, rather than being left to a set of exercises at the end.

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 ______________ histogram corresponding to a sample of points from X. Compute the ____________ $\overline{x}$ and the ________________________ 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 $y =$____________, and your numbers $\overline{x}$ and s will converge to _________ and _________, respectively.

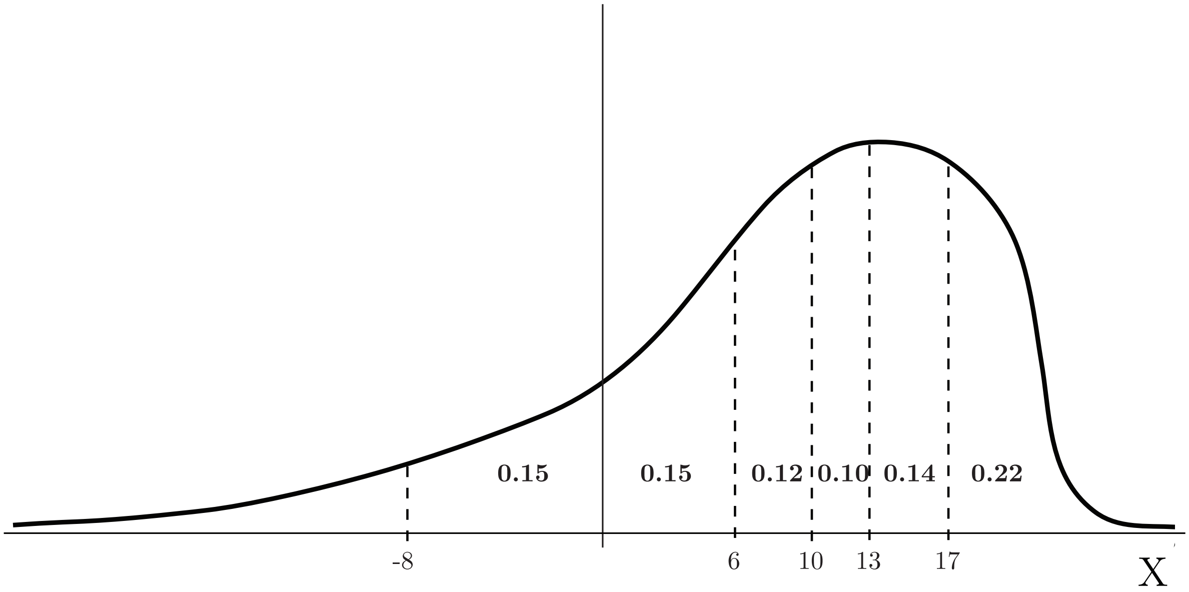
Also recall that, for any pdf $f(x)$, with domain (c,d), we must have

- $f(x) \geq 0$ 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) $P(X < -8)$. ___________

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

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

B. Standard normal random variables and pdf's

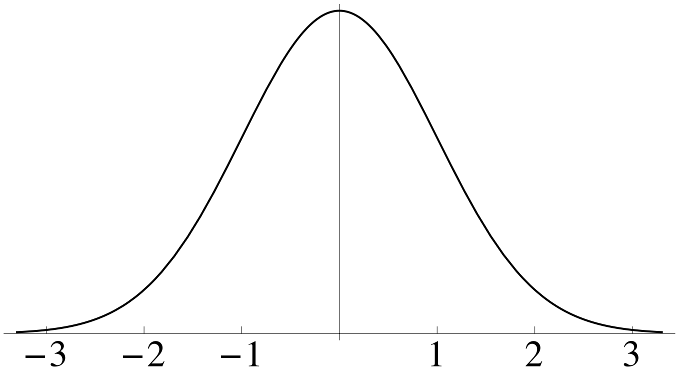
A random variable X is said to have a *standard normal distribution* if the pdf for X is given by

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

(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 __________% of the values of X lie within one ____________________ 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 __________% of the values of X lie within __________ 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 __________% of the values of X lie within three ____________________ 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 __________% of the values of X lie within __________ 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 __________% of the values of X lie within __________ standard deviations of the mean.

(f) It can be computed that $\int_{-2.576}^{2.576} f_{0,1}(x)\,dx \approx 0.990$. In other words: if X is $N(0,1)$, then about __________% of the values of X lie within __________ standard deviations of the mean.

C. Other normal random variables and pdf's

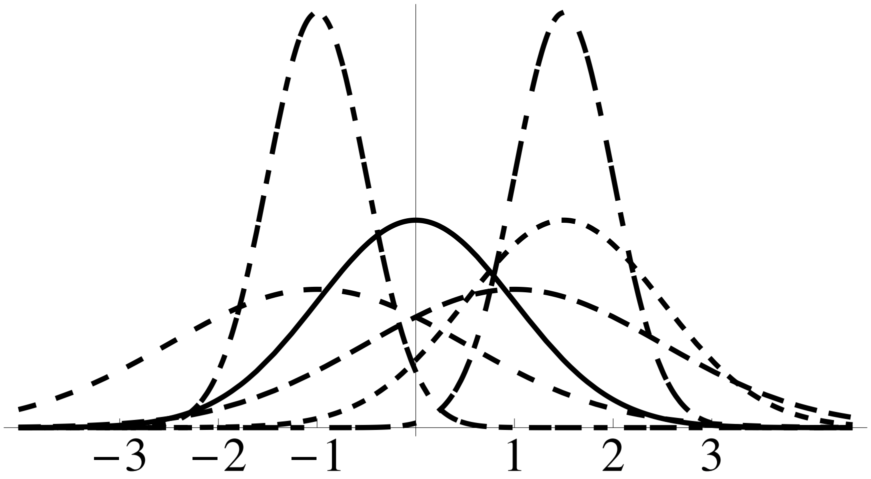
If the pdf for a random variable Z follows the basic shape of the standard normal curve, but has mean μ (instead of 0) and standard deviation σ (instead of 1), we say "Z is $N(\mu,\sigma)$." Let's denote such a pdf by $f_{\mu,\sigma}(x)$. Then: to say Z is $N(\mu,\sigma)$ is to say that, for any real 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.

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 $P(-2.3 < z < -1.7)$.

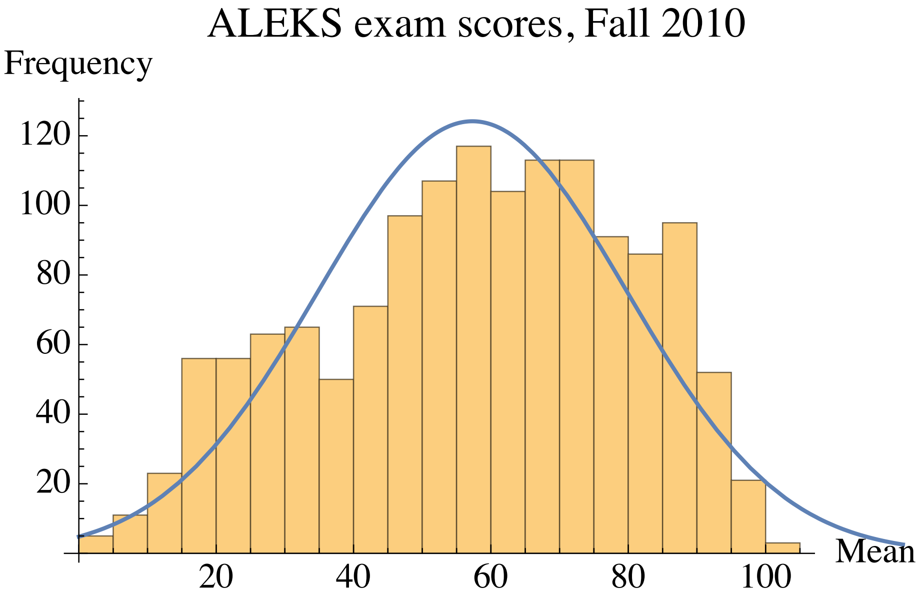
Exercise D2. Suppose Z is $N(2,2)$. Find $P(-1.92 < z < 5.92)$.

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

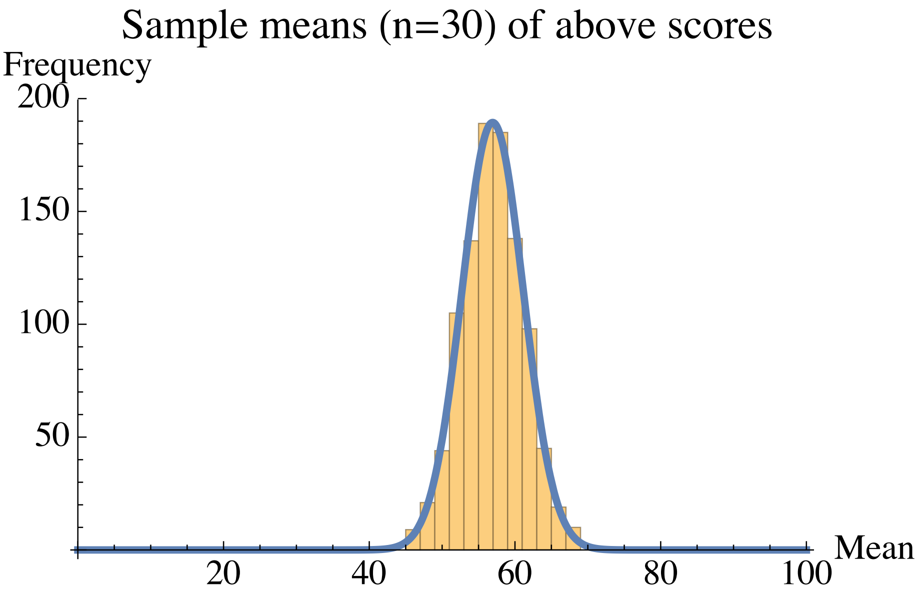
Exercise D4. What exercise **D3** directly above says is: if Z is *any* normal random variable, then ____________% of the data lies within three standard __________________ of the _____________.

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 ____________ 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 ______________ than that of the original dataset, because the sample means seem much ___________ spread out (that is, they seem ____________ 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 ______________ 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 $n = 30$)?

Remark. There are roughly $6.52941 \cdot 10^{61}$ *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 did that a hundred million times, we'd get relatively close. But let's not.)

For the above histogram of sample means, we computed considerably fewer means – about 1000, 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

OK, 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 $\mu =$__________ (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 ______________ pdf, with mean

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

and standard deviation

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

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 $N($________,________$)$. This tells us, by part (f) of exercise **B1** above, that 99% of all possible values of the random variable $\dfrac{\overline{X}-\overline{\mu}}{\overline{\sigma}}$ fall between the numbers ______________ and _____________.

In particular, suppose we actually *compute* a sample mean $\overline{x}$ from a sample of X, of size $n = 105$, and find that $\dfrac{\overline{x}-\overline{\mu}}{\overline{\sigma}}$ 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 $\dfrac{\overline{x}-\overline{\mu}}{\overline{\sigma}}$, for this particular value of $\overline{x}$ and for the $\overline{\mu}$ and $\overline{\sigma}$ computed in exercise **F1** above.

Exercise F3. Is the number you computed in exercise **F2** above between -2.576 and 2.576? 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.

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.

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.

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 _________ and _________________________ 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,

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

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

$$P\left(-1.96 < \frac{\overline{x}-\overline{\mu}}{\overline{\sigma}} < 1.96\right) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.$$

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

$$P\left(\overline{x} - 1.96\,\overline{\sigma} < \overline{\mu} < \overline{x} + 1.96\,\overline{\sigma}\right) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. \qquad (*)$$

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

$$\overline{\mu} = \_\_\_\_\_\_\_\_\_ \quad \text{and} \quad \overline{\sigma} = \_\_\_\_\_\_\_\_\_\_\_.$$

So equation $(*)$ can be rewritten:

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

Or in other words: there is a __________% 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 $\overline{x}+$ ________________.

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 _________________________. Then equation $(**)$ reads:

$$P\left(\overline{x} - 1.96\,\frac{s}{\sqrt{n}} < \mu < \overline{x} + 1.96\,\frac{s}{\sqrt{n}}\right) \approx \underline{\hspace{4cm}}.$$

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.

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

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.

Exercise G4. Construct 98% and 99% confidence intervals for the mean navel ratio μ of all CU students.

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.

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.