BIEB100. Professor Rifkin.
Notes on Section 2.2, lecture of 27 January 2014.

Do students sleep the recommended 8 hours a night on average?

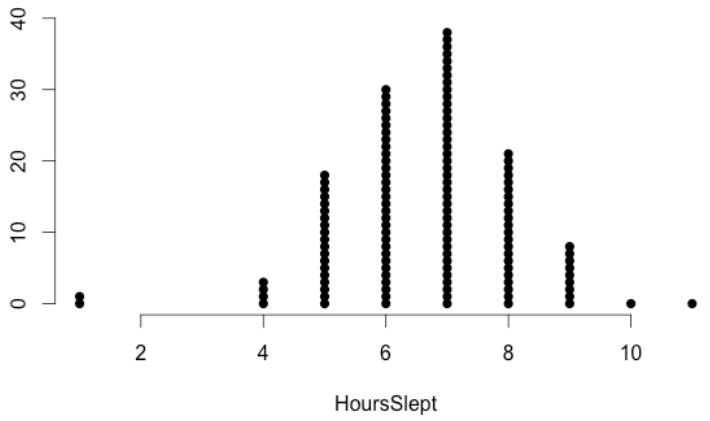
We first set up our null and alternative hypotheses:

H0: $\mu = 8$
HA: $\mu < 8$

Note that our question is about the entire population of students so we want to know about the parametric average. We are using the mean as our average in this case.

Then we collected a sample. If we really want to generalize to *all* students then we should really collect a *simple random sample.* This requires that each and every student would have an equal chance of being in our sample. Since this was an in-class demonstration, we *pretended* that the convenience sample of students in the class was in fact a simple random sample. Whether it is a simple random sample or not has its repercussions later on when we try to draw conclusions and generalize in Step 5.

Here was your data:



There were 128 students (sample size 128). Before when we were dealing with a categorical variable with 2 categories, we could summarize the data with a single number: the proportion in one of the categories. For a quantitative variable it is more complicated, we need to describe more about the data.

The average gives a good single number description of where the middle of the data is. We discussed two averages: mean and median. The mean is the sum of all the data values divided by the number of data points. It is the most commonly used average because it has some nice mathematical properties that we won't go into. The median is the data value smack in the middle of your data – sort your data in order and the median is the middle one. There is one wrinkle. Let's say you have an even number of data points, like 128. Then there is no middle one. There are 64 on the low side and 64 on the high side. In this case, you take the 64^{th} and the 65^{th} and take the mean of them. That is your median.

We talked about the shape of the distribution. Your data was fairly symmetrical. But some data (like the income distribution in the US) has a long tail to one side. This is called a *skewed* distribution. For the income distribution, the long tail is on the right side so this is called *right-skewed*. If the distribution is highly skewed, it affects the mean but not the median. A right skewed distribution will pull the mean to the right; a left-skewed distribution pulls the mean to the left. The median isn't affected by the actual values in the tails. You can pull a tail out as far as you want and it won't change. The median is *resistant* while the mean is not.

We also discussed the spread of the distribution. The standard deviation is the statistic usually used to measure the spread and it effectively measures the average distance of a datapoint in the distribution from the mean. If you look at the formula for standard deviation:

$$\sqrt{\frac{\sum_{i=1}^{i=n}(x_i - \mu)^2}{n-1}}$$

you will see that looks a little more complicated than that, but this is the idea that the formula is capturing. *i* here represents a data point.

For your data:

mean = 6.64
median = 7
standard deviation = 1.49
fairly symmetric

Now the question was whether your data were consistent with being pulled from a population with a true mean of 8 hours slept per night. There will be some variation in the population, and we ran into the question of how to represent that. I made the following assumption:

Let's say that the number of hours slept by a person in the entire population were normally distributed with a mean of 8 with the same standard deviation as in our sample. Why 8? Because we assumed that the population mean was 8 in our null hypothesis. Why the same standard deviation as in our sample? We didn't have an hypothesis about the variability of the population and were only interested in means. So we assumed that the amount of variation among you was the same as the amount of variation as in the real population, so that we could focus on deciding whether there is a meaningful difference in means.

I then simulated this population by using a computer to pick 50,000 samples from a normal distribution with mean 8 and standard deviation 1.49.

For the curious, I used the computer program R to do this. R is probably the most used statistical software in biology. You may see it in lab courses, and if you go on in biology or in any field that uses statistics, I would highly recommend you learn it. There are good, free online tutorials around, including this one: https://www.datacamp.com/

This 50,000 samples represented the population and then the goal was to randomly take a sample of size 128 from the population, calculate the mean, and repeat this many times until we get a *sampling distribution* of means. This sampling distribution represents the sample means we would get if the actual population mean were really 8. We will use this sampling distribution as our null distribution for doing inference.

Just to recap: I assumed that the population was normally distributed with a mean of 8 and a standard deviation of 1.49. What does this mean? Imagine you had infinite numbers. This is the distribution. Now imagine making a histogram or a dotplot to represent these numbers. The shape of that histogram or dotplot would be the familiar bell-curve shape of a normal distribution. This is what I mean when I say that the population was normally distributed.

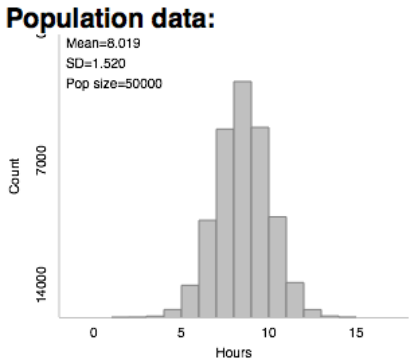
I then represented that infinitely big distribution with a finite one of 50,000. Then I picked samples of 128 repeatedly from that 50,000.

You might ask, "If you can pick numbers directly from that infinitely big one, then why bother with the intermediate step of picking 50,000 to represent it? Why not repeatedly sample 128 datapoints from the infinite one itself?"

This is an excellent question. And you are right that in concept it is a complete waste of a step. We might as well repeatedly sample 128 datapoints directly from the infinite one. The reason I didn't do this in practice is simply because the One Mean applet doesn't work that way and so the book doesn't present it that way. The applet requires you to list your population data values in the box, and it samples from those values. This is a limitation of this particular applet. A program like R could skip that step.

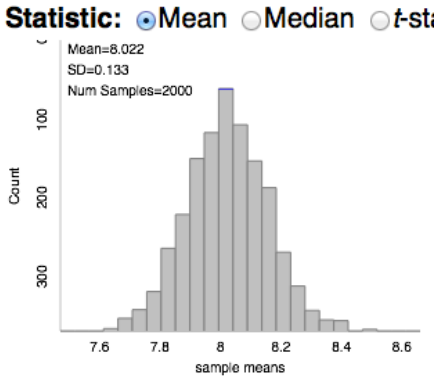
That said, 50,000 is pretty big and so *in practice* the results you get with 50,000 representing your population are going to be extremely similar to the results you would get if you actually sampled from the infinite population itself.

Now we have a population. It looked like this:



Notice that it is centered around 8 with a standard deviation of around 1.5 like we wanted (approximately, although not exactly due to the fact that although 50,000 is big, it is still finite).

We used the applet to repeatedly sample from that population. We picked a sample of 128, calculated the mean, and that was one point in our null distribution – our sampling distribution of means. We did this 2000 times and got a distribution of possible means that looked like:

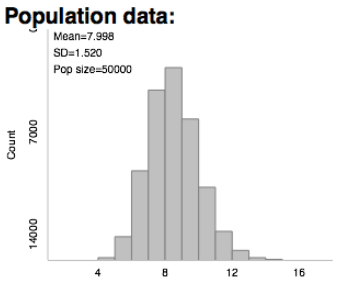


Notice that the standard deviation of this distribution is smaller than the standard deviation of the population distribution. This is obvious if you just look at the x-axes of the two plots. The standard deviation of the whole population was approximately 1.5. The standard deviation of this one is 0.133.

Immediately you can see that our actual sample mean of 6.64 is way off the left side of our null distribution. We counted the number of times out of the 2000 that we got 6.64 or less and it happened 0 times (remember that our alternative hypothesis specified "less than" so we want to look at the left tail). So our p-value is 0. Our sample (which we pretended was *representative* in a statistical sense of all students) is inconsistent with the idea that students in general get on average 8 hours of sleep a night. So we reject that hypothesis.

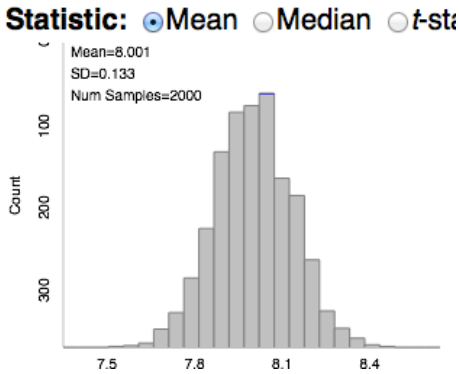
You might ask, "can we say what the actual population mean is using our sample data?" This is estimation and is the topic of chapter 3.

Distributions of means have some other curious properties. Let's say that instead of postulating a normal distribution for the population, I assumed a left skewed distribution. Below is a histogram of such a distribution that I got in a similar way as before. I used R to pick 50,000 datapoints from a skewed distribution with mean 8 and standard deviation 1.49.



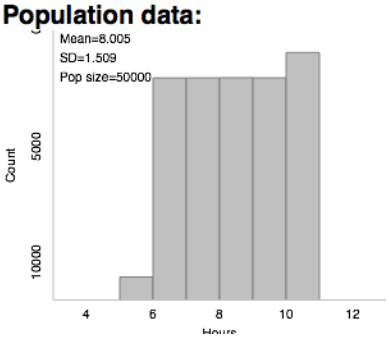
Notice the slight skew to the right.

If we repeat the same sampling process, repeatedly picking samples of size 128 from this population distribution, calculating the mean, and constructing a null distribution that way, you get the following:



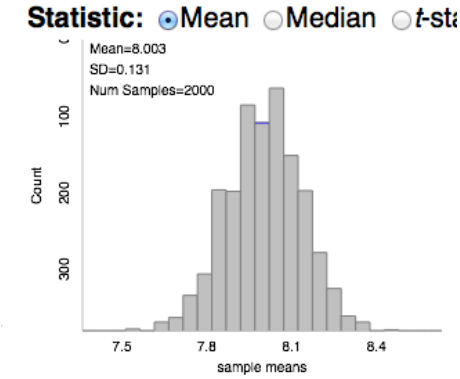
Looks pretty similar to before, doesn't it? Notice that the standard deviation is even the same. If we were to run it 10,000 times instead of just 2000 the two would look even more similar.

You might argue that the skewed population distribution was almost symmetrical. How about this one:



This one looks nothing like the other two.

We do the repeated sampling again to build up a null distribution. Here is what the null distribution of sample means looks like:

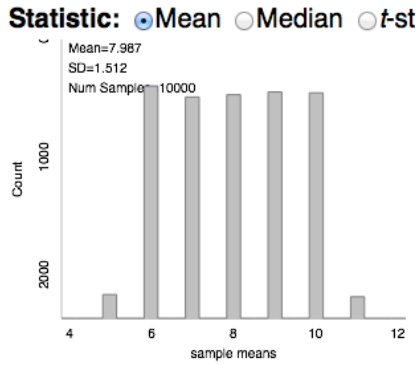


No matter what our population distribution looks like, the distribution of sample means looks the same!

Here is another curious feature of distributions of sample means.

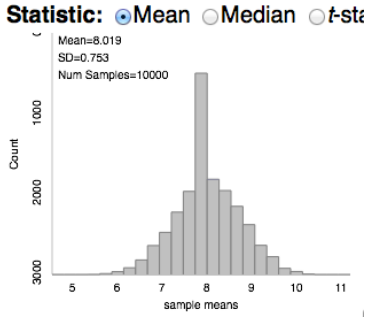
This time I am going to change the sample size. This time I am not interested in the shape of the null distribution (distribution of sample means) but only in the standard deviation of it. I'll use the uniform distribution (the flat one above) for my population, but this doesn't really matter. We would get to the same conclusion with any other one. We start out with a standard deviation in the population around 1.5

Sample size 1:



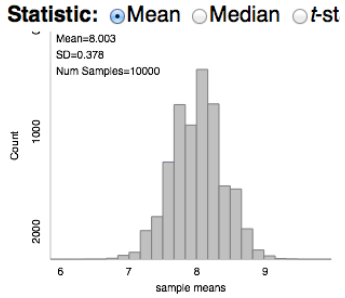
Standard deviation of the sampling distribution: ~ 1.5

Sample size 4:



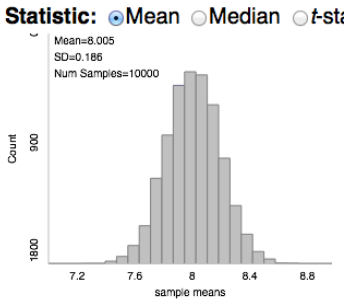
Standard deviation of the sampling distribution: ~ 0.75

Sample size 16:



Standard deviation of the sampling distribution: ~ 0.375

Sample size 64:

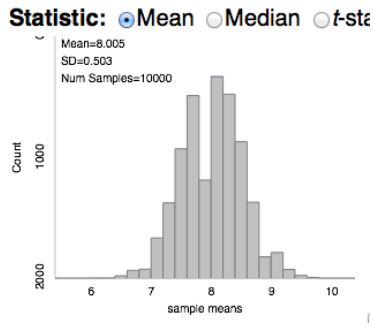


Standard deviation of the sampling distribution: ~ 0.1875

Do you see a pattern?

Each the sample size quadruples, the standard deviation of the sampling distribution goes down by half.

Watch what happens when we go from a sample size of 1 (see above) to a sample size of 9:



Standard deviation of the sampling distribution: ~ 0.5

So we increased the sample size by 9 times and the standard deviation of the sampling distribution went down by a third.

I'll leave it to you to flesh out the relationship. If you want to explore more, the One Mean applet comes with 3 populations already loaded in it that you can use.