Day 4. Histograms and Distributions: Teacher's Notes

The students have read the short article about histograms. They have come into class, each with one data point (a "datum"), representing the number of hours of sleep they had last night. We want them to become comfortable with collecting data and plotting the data onto histograms. And, we want them to know – only conceptually -- about the relationship between histograms and underlying probability distribution. We can do all this with almost no math formulas!

Collect the data. The class starts with the new data, the number of hours each student slept last night. Each student signs a piece of paper with his/her estimated number of hours of sleep, -- rounded to the nearest hour -- and hands it to you – the teacher – for collection. You promise confidentiality, no student names to be associated with any data point (possibly embarrassing ones like 2 or 14)! Before class you carefully drew on the blackboard (or whiteboard), perhaps using a ruler, an X-Y axis that a student volunteer will use to make the empirical histogram. On the X-axis we have number of hours of sleep (assumed to be an integer quantity). You might want to run the possible number of hours of sleep from 3 to 12! On the Y-axis we have student count, that is, the number of students who fit into each "bin" of the histogram. Also on the Y-axis, you clearly show the integers from 1, 2, 3, … 15.

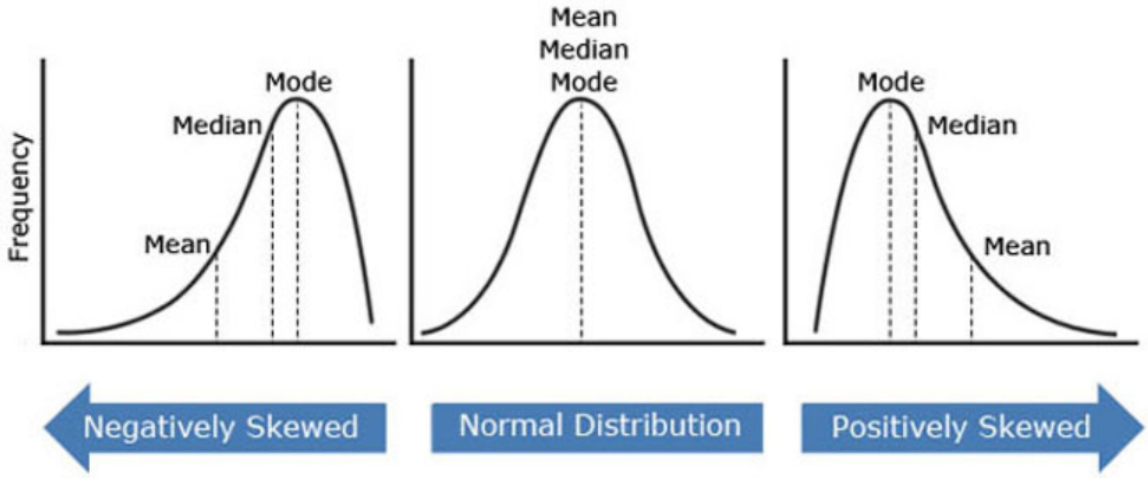
Plot the data on a histogram. Now, after you have collected all the slips of paper and shuffled them (so there is no way to attribute any reported sleep time with an individual student), one by one, you read the data entries for your student assistant to mark as she/he creates the histogram. Suppose you next read "6", representing one student who enjoyed 6 hours of sleep. The assistant would draw a unit-height block on the histogram data column representing 6 hours of sleep. She/he will increase the height of that emerging column of the histogram by one. The ultimate final histogram will be vertical collections of these unit-height boxes.

Discuss the histogram. Once the histogram is completed, another student volunteer soon computes the average sleep time, and marks it on the blackboard histogram along with mode, median and 5% tails. How do the average, mode, median compare? Encourage class discussion about shape and properties of their class histogram. You might say, "*Do we all know a lot more about class sleeping hours than we knew before? What if we were only told the mean or average value, without the histogram? Would we know just as much? Why or Why not?"* It is likely that this histogram will be a unimodal one, that is having one peak, we'd guess around 7 or 8 hours. From this peak (the mode), we'd expect drop-offs in each direction, up or down in sleep hours. You may want to discuss the "outliers," the maximum and minimum numbers of sleep hours reported. Then have a general discussion about histograms – their ease of creation, use and interpretation. Possibly add computer-based way of redoing the blackboard exercise, resulting in a perfectly beautiful professionally created histogram on the computer (using Excel).

Moving to Distributions. The discussion of histograms should now migrate to include "distributions," often called, "probability distributions, or "theoretical distributions." Draw a distribution on the blackboard (or whiteboard). Maybe it's the bell-shaped curve, otherwise known as the Gaussian or Normal distribution. Also, you might draw one that could underlie the histogram of sleep times.

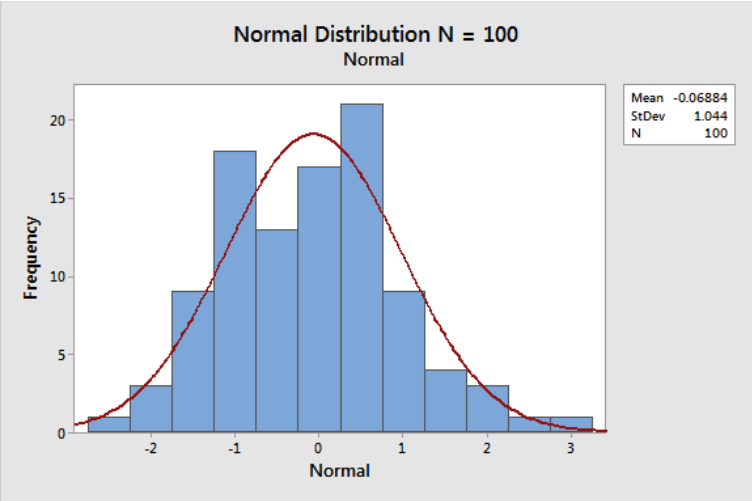
Discussion of histogram vs. "theoretical distribution." ***A distribution is a curve above or on the horizontal axis representing a "mathematical model" of the process that is generating data going into a histogram.*** It's called a *probability distribution* since the various areas under the distribution curve represent the probability that the next histogram sample will be between, say, 5 and 8 hours of sleep. Teacher: To illustrate to your students, draw this on the board, with your distribution. The total area under the curve is 1.0, since the total probability must be 1.0.

Illustrative distributions:



The middle distribution, arising often in practice, is called the Normal Distribution (or Gaussian distribution). It is symmetric around its mean or average, and here the mean = the mode = the median (a rare occurrence). The other two displayed distributions are skewed to the left or right, and we see that the mean becomes a less and less useful indicator of the outcome of the process.

Often it is good practice to display together the data as shown in the histogram and the underlying model, as shown by the probability distribution. Here is an example with the normal distribution:

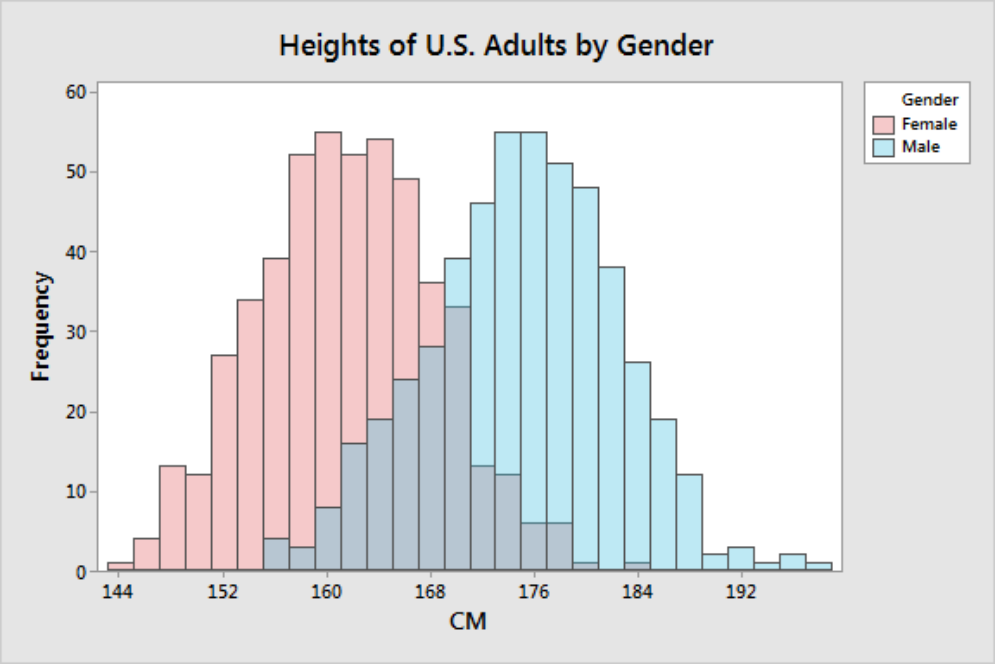


OPTIONAL SECTION. Relationship Between Histograms and Distributions. What is the relationship between histograms and distributions? As we have stated, a distribution is a conceptual model of the process creating data for your histogram. How might you visualize this, and have some fun at the same time? Draw a distribution on the board that has a very limited range, say from -2 to +5. The shape of the distribution above the x-axis between -2 and +5 can be almost anything, maybe like a roller coaster. Now, in different color chalk, enclose the entire distribution in the smallest rectangle that fully contains it, horizontally and vertically. That rectangle's bottom is located on the x-axis (or horizontal axis); its right vertical side is located at +5 and left vertical side at -2; the top side coincides with the modal or maximum value of the distribution. *The fun part*: At some distance from the board, one throws small pieces of chalk at the rectangle. Maybe, if time, each student gets a chance! The resulting chalk mark will be located at an x and a y value on the coordinate system of the board. If that mark on the board is **in the box** and **BELOW** the distribution curve, ***accept*** that chalk-toss x value as a valid sample from the distribution. If the resulting chalk mark on the board is in NOT BELOW the distribution curve, reject that chalk toss and continue. Class discussion: Why does this make sense? Since most people have bad aims with chalk, we think of the locations of chalk tosses as uniformly random over the rectangle. Then, the greatest chance of having an x-value accepted is at x values below the maximum of the distribution, that maximum equaling the mode of the distribution. Consider another part of the distribution whose height is only one half the modal value. Then, the chance that an x-value there will be accepted is only half that of one near the model value – because 50% of the time a toss at this value of x lands above the distribution line and must be rejected.

OPTIONAL SECTION. Building a Histogram. Once all the chalk pieces have been tossed and a number of them have been accepted (because they were in the box and below the distribution), we pause and ask, "How can we make a histogram with these results?" Well, we can round outcomes to integer intervals, histogram column 1: (-2, -1); histogram column 2: (-1, 0); histogram column 3: (0, +1); etc. And we can draw the histogram resulting from our chalk tosses! Let's do it…

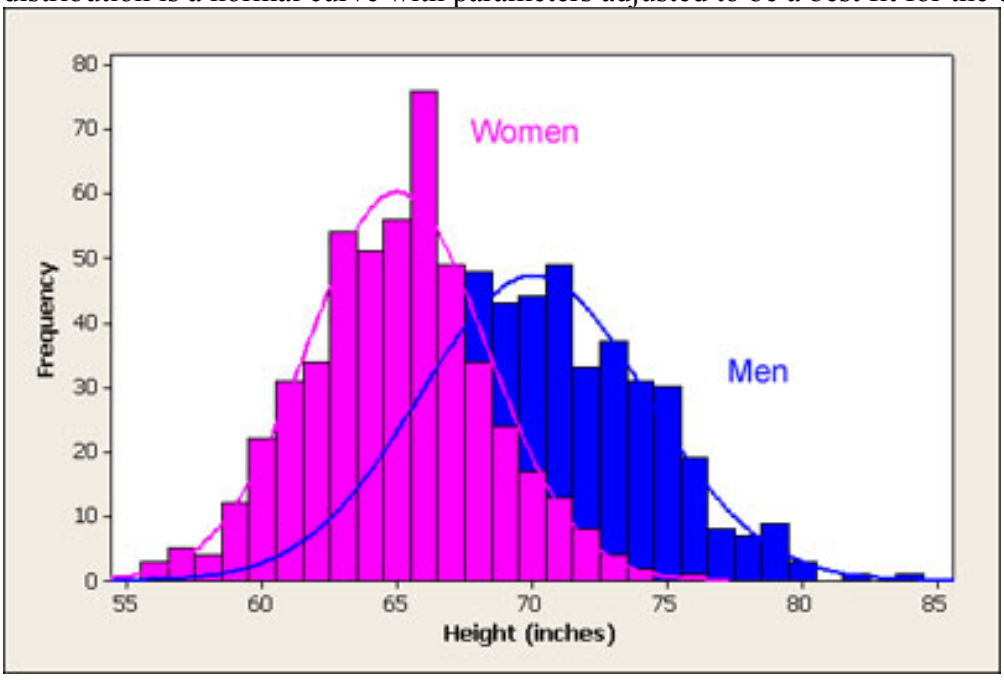
Conceptually we can think of this as the process generating data for our histograms. Each entry into a histogram can be thought of as a random accepted chalk toss onto the blackboard on which the underlying probability distribution has been drawn. That's pretty cool! And you can understand that if we have only a few data points, our resulting histogram may not look at all like the probability distribution. But as the sample size increases, things tend to stabilize, and the histogram begins to look more and more like the underlying distribution.

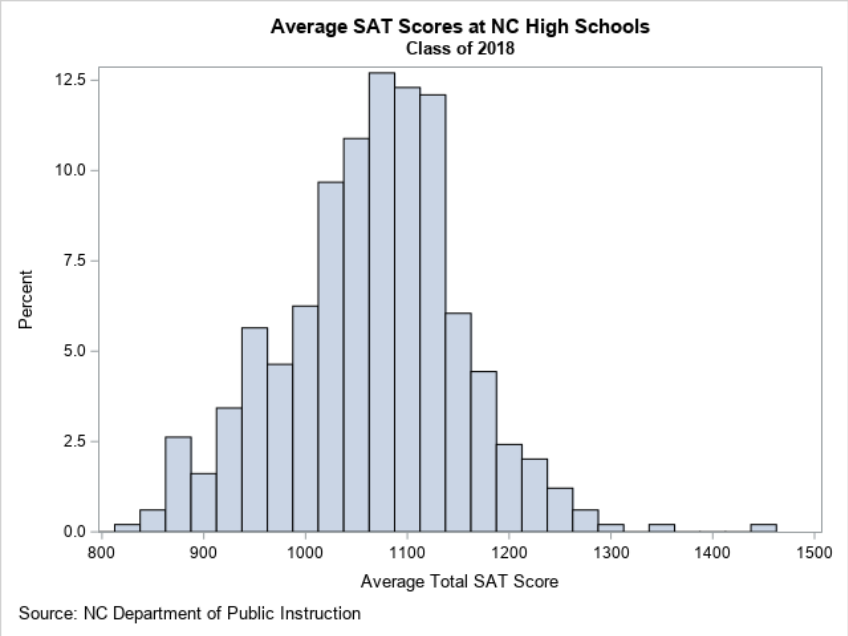
<u>Illustrative Histograms. Not optional</u>. Now we present and discuss a few illustrative histograms.



This is called a bi-modal histogram, having two peaks. Discussion how one average computed for everyone is misleading. We need two averages, one for each gender. Every time we come across a bi-modal distribution, we should think of it as displaying the results for two separate populations, and compute averages for each.

Sometimes it is useful to display a histogram together with the underlying model (distribution) that is generating the data. Here we can see that for heights of adults by gender. In each case, the distribution is a normal curve with parameters adjusted to be a best fit for the data.





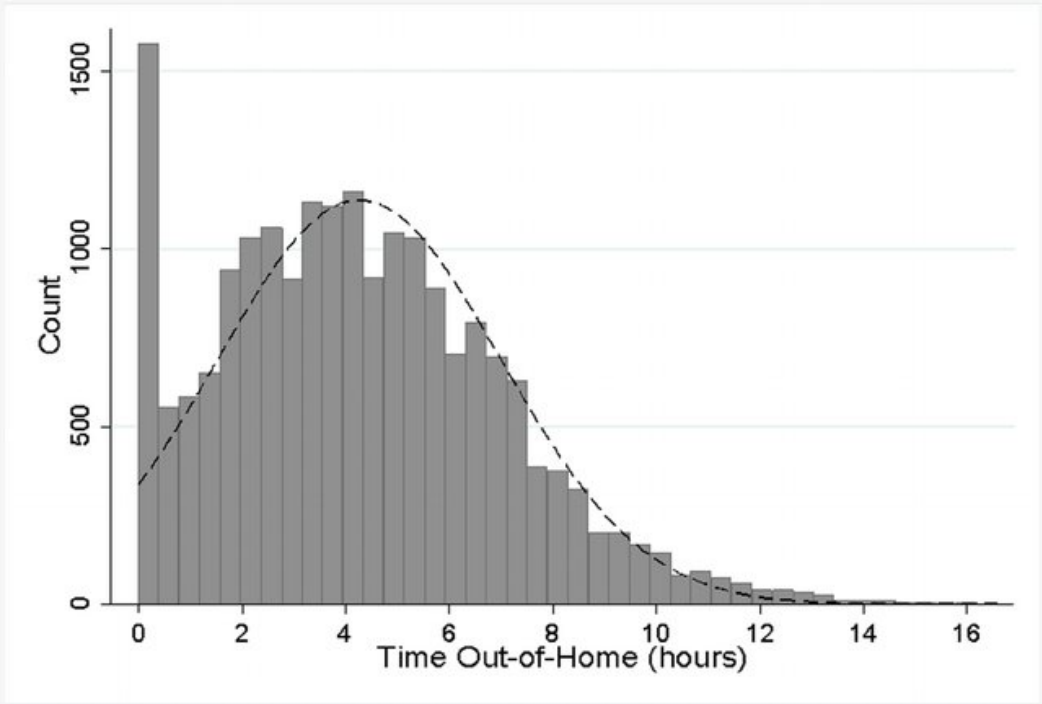
This histogram shows the distribution of the **average total SAT score for <u>schools</u>** in North Carolina. These are not individual student scores, but average scores by school. That is why the right-hand tail and left-hand tail do not contain entries that you might expect if these were scores of individual students. Always be careful to understand exactly what is being shown.

From this histogram, you can determine several facts about the data:

1. For most NC schools, the average school-wide SAT score is about 1100.
2. About 73% of NC schools have an average SAT score between 1000 and 1200.
3. There are a few schools that have much higher scores than the others. Those schools are Early College At Guilford (Total=1442), Raleigh Charter High School (Total=1356), and East Chapel Hill High (Total=1290).

Can someone from the class sketch what the histogram might look like for individual student SAT scores from Early College At Guilford? Remember, this average is 1442!

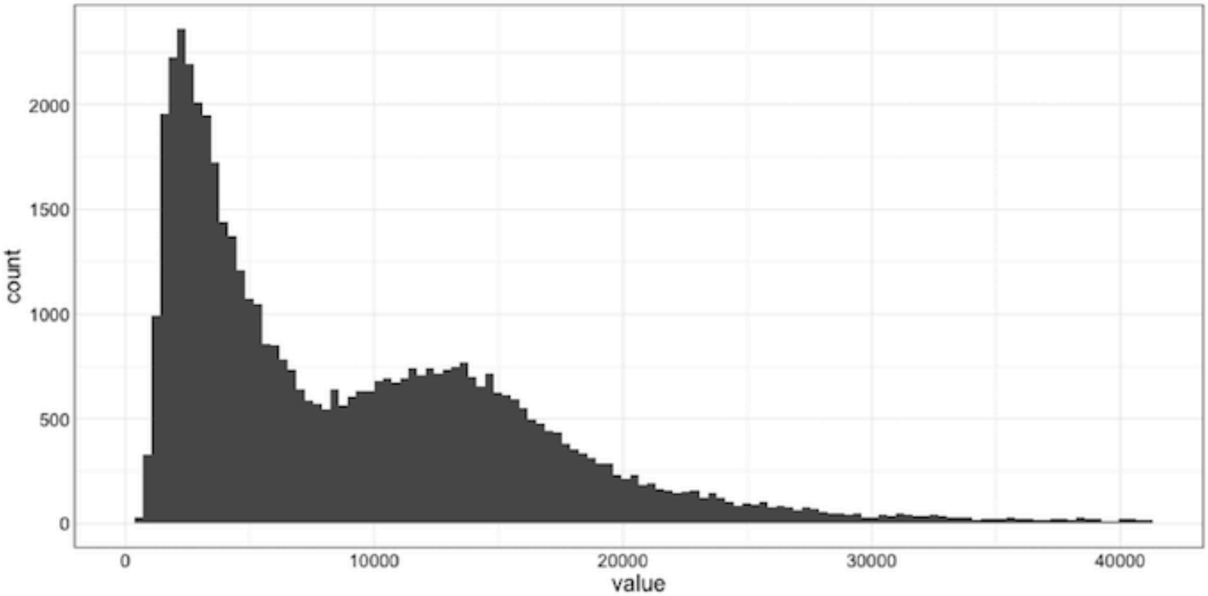
https://blogs.sas.com/content/iml/2019/03/04/visualize-sat-scores-nc.html



For older adults, histogram of the daily hours spent outside the home, showing the limit at zero. A normal distribution curve is plotted as a dashed line to show the data is approximately normally distributed except at and below zero. Here we have two separate populations displayed: Those who do get out of home and those who do not. Want to compute average for each, with one average equaling zero.

From *Time Out-of-Home and Cognitive, Physical, and Emotional Wellbeing of Older Adults: A Longitudinal Mixed Effects Model*

https://bit.ly/3hyAcXJ



Here we are seeing the histogram of response times to various internet requests. Again, we see a multi-modal distribution, suggesting that at least two different populations are combined into one display. Separate averages are required, as the global average for this histogram is misleading – not giving useful information about either subpopulation.

https://blog.newrelic.com/engineering/expected-distributions-website-response-times/

That's all for today, FOLKS!

((   ))