NUMB3RS Activity: A Stab in the Dark
Episode: "Hot Shot"

Topic: Histograms, relative frequency, and probability **Grade Level:** 7 - 12

Objective: Connect histograms and relative frequencies to probability

Time: 20 - 30 minutes

Materials: TI-83/84 Plus calculator

Introduction

In "Hot Shot," two victims of a possible serial criminal are found. Charlie uses **kernel density estimation** as a method for predicting a change in behavior to indicate a connection between the victims and the criminal. Kernel density estimation (KDE; Bowman and Azzalini 1997) is a technique whereby the "kernel" is used to produce a smooth, continuous estimate of the true density of an empirical distribution. The kernel can take on a variety of forms, but the simplest density estimation technique is the familiar histogram. Histograms divide the range of data into intervals, or "bins," and the number of data points falling into each interval is counted. The main disadvantage of the histogram is that its bin width is arbitrary, and can lead to potentially misleading results. The KDE method effectively alleviates many of the problems associated with arbitrary binning choices for histograms, although the choice of bin width (which controls the smoothing) remains subjective. This activity will focus on creating and interpreting histograms and their relationships to probability. In the Extensions, students can learn more about the idea of smoothing the data with a function to make better predictions.

A **frequency histogram** is a graphical representation of a data set in which various disjoint categories define "bins." Each piece of data is mapped onto the graph by determining the bin to which it belongs. Often, the bins are positioned on the horizontal axis. The vertical axis is used to represent the frequency or count of the data contained in each bin. Each bar in the histogram is a rectangle with the bin width and the frequency of the bin as dimensions.

A **relative frequency histogram** is a histogram based on percentages. In a frequency histogram as described in the previous paragraph, the vertical axis represents the count of items in a bin. In a relative frequency histogram, the vertical axis represents the percentage of the total count of the items in the bins. The relative frequency (or percentage) of a bin also represents the probability that the items in a bin will be selected in a random sampling of all the given items. Note that the relative frequency, or probability, for a given bin could be determined by dividing the area of the bar representing the bin by the sum of the areas of all the bars in either the frequency histogram or the relative frequency histogram.

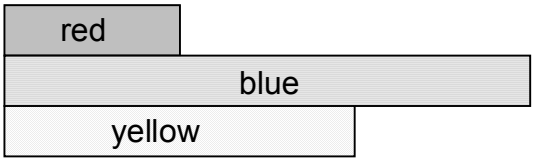
In this activity, students plot histograms and create the corresponding relative frequency histograms. Note that the activity only uses a small sample for illustrative purposes. With many samples, it would likely be more meaningful to tie the histograms to the probabilities and use the area under a smoothed curve to find those probabilities.

Terry Souhrada, University of Montana, Missoula, MT

Students will use probabilities to help predict the possible time the victims were injected with drugs in the "Hot Shot" episode. In the Extensions, they will be introduced to the concept of curve smoothing for a histogram.

Discuss with Students

1. In a histogram, the bins are all the same width. The lengths of the bars shown below represent the number of colored beads in a box. How could you use the bars to determine the probability of drawing a blue bead from the box in a single draw?



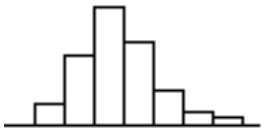
2. How would you determine the probability of drawing either a blue bead or a red bead in one draw from the box in Question 1?

3. Relative frequency is the percentage of items in a specific event out of the total number of items possible. What is the relative frequency of yellow beads in Question 1?

4. Why can one use the areas of the frequency histogram OR the area of the relative frequency histogram to determine probabilities?

Discuss with Students Answers: 1. *We are not given the number of beads of each color but can consider the bins representing the color to be indicative of the number of beads of each color. Because the width of each bin is the same, the probability is the ratio of the length of the blue bar to the sum of the lengths of all the bars. Alternately, you could divide the area of the "blue" bar by the total area contained in all three bars. This resulting quotient in either case is the probability of a blue bead being drawn.* **2.** *You would add the probability of drawing a red bead in one draw to the probability of drawing a blue bead in one draw.* **3.** *The relative frequency of yellow beads 2/6 = 1/3 ≈ 0.33* **4.** *The denominators of 100 in the decimals of a relative frequency histogram divide out resulting in the same values as the frequency histogram.*

Student Page Answers: 1a. Sample graph:

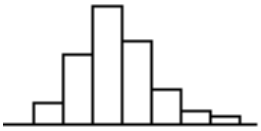


1b. The most typical is from 12 to 14 mg because the four bars in the middle contain the most data.

1c. Sample graph:  Yes; the most typical is now from 11.75 to 14.25 mg. It is harder to see a trend in the data.

Terry Souhrada, University of Montana, Missoula, MT

1d. Sample graph:  The most typical is from 12 to 13.5 mg. The bars are narrower and the information is more detailed, but some of the overall trend that was apparent in the first histogram is gone. **1e.** Charlie would likely be surprised or suspicious because only 1 person in the sample had a reading that high. However, he should remember that he is working with only a small sample.
1f. Answers vary. Here with only one sample, the first histogram might be considered best because it shows a trend in the data and the bin widths fall into the optimum categories. With only one sample, the second histogram doesn't appear as good because the bin widths use an interval that isn't helpful. Again with only one sample, the third histogram has bin widths that are too narrow and it is more difficult to analyze the general shape of the data. With many samples and to try to generalize to an entire population, narrower bin widths might be more helpful.

2a. Sample graph  **2b.** About 37/50 or 74% of the samples have more than 12.5 mg of the drug left in the system after 13 hours. **2c.** The graphs are similar and the expectations are the same. These results do tend to support Don's suspicions. **3a.** $P(X \geq 1.0) \approx 0.70 = 70\%$. The area of each bar can be approximated by measuring them. The probability is the ratio of the area of the shaded region to the total area under all the bars. **3b.** Because the majority of the people in the study have 1 mg of diazepam or higher, the two sets of data have produced similar results and they tend to support each other.

Terry Souhrada, University of Montana, Missoula, MT

Name: _____ Date: _____

NUMB3RS Activity: A Stab in the Dark

In "Hot Shot," the FBI is investigating a possible serial crime involving the deaths of two victims injected with a combination of drugs. The FBI attempts to determine at what time the victims were injected, to see if they could place the suspect at the scene at that time. Don thinks the victims were injected at least 13 hours before they were found. Charlie suggests that the amount of the drugs left in the victims' systems could indicate when they were injected. Don noted that each victim had about 12.5 mg of morphine in the blood stream and about 1.0 mg of diazepam in the system when found.

Assuming that Don was approximately correct about the time of injection, Charlie gathers information about how much morphine (in mg) would be expected in the blood after 13 hours. He is interested in levels at or below the 12.5 mg of the victims. He found the data in the table based on 50 people in a pharmaceutical study that shows how many milligrams of morphine is in their systems 13 hours after being injected with a prescribed dose of 20 mg.

11.5	13.4	12.7	12.6	12.4	12.9	12.7	12.4	13.4	12.4
12.1	12.6	14.6	11.7	12.2	14.2	13.3	13.3	12.7	12.6
12.9	14.1	12.3	13.7	12.4	13.2	13.4	12.9	12.8	13.4
13.4	13.4	12.1	12.8	12.8	12.2	13.8	11.8	13.8	12.0
12.7	13.6	12.6	13.1	13.4	12.8	13.4	13.9	12.9	12.6

1. A **frequency histogram** is a graph that shows frequency or count of data. The data is sorted into categories or bins of equal width. The height of the bin is the count of data in that bin.

a. Enter the information from the table in list L_1 on your calculator. Now, create a histogram illustrating the frequency on the vertical axis and the number of mg on the horizontal axis. To create a histogram, press [2nd] [STAT PLOT] and select Plot 1. Turn the plot on, and select the histogram icon. Press [WINDOW] and change the settings as follows: `Xmin = 11.0`, `Xmax = 15.5`, `Xscl = 0.5`, `Ymin = 0`, and `Ymax=20`. (This sets the bin width to represent 0.5 mg and the starting value of the first bin at 11.0 mg.) Press [GRAPH] to draw the histogram.

b. What is the most typical amount left in someone's system after 13 hours? How does the histogram help you see this?

c. Set the starting endpoint (`Xmin`) to 11.25 and create a new histogram. Does this affect the typical amount you described in part **b**? Why or why not?

d. Change the starting endpoint back to 11.0 and change the bin width (`Xscl`) to 0.25 and create a new histogram. Does this affect the typical amount you described you calculated in part **b**? Why or why not?

Terry Souhrada, University of Montana, Missoula, MT

e. If Charlie found a person with 14.5 mg still in his or her system, would he be surprised or suspicious? Why or why not?

f. Of the three histograms Charlie received, which best helps him answer his questions? Justify your choice.

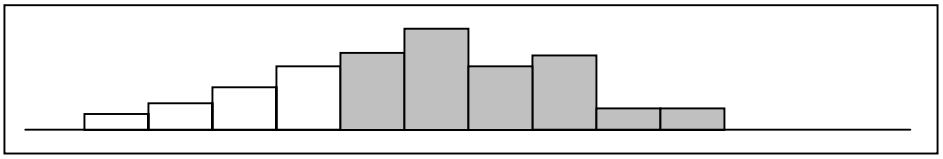
2. A **relative frequency histogram** is a histogram that uses percentages of the total count instead of the count for the height of the bin. The percentage is the relative frequency of the bin.

a. From the histogram in Question **1a**, determine the relative frequency for each bin. On a sheet of paper create a relative frequency histogram with a starting endpoint of 11.0 and a bin width of 0.5.

b. Using the relative frequency histogram with this sample, approximately what percentage of the samples have more than or equal to 12.5 mg of morphine left after 13 hours?

c. Compare the amounts you found in Parts **1b** and **2b**. Do these results tend to support Don's suspicion that the victims had been injected 13 or more hours before they were found?

3. Don knew that each victim had about 1.0 mg of diazepam in his or her system. Charlie collected similar information about diazepam as he collected about morphine. He then created a histogram showing the amounts of morphine remaining after 13 hours, which is shown along the horizontal axis of the histogram. The shaded bins represent the area of interest.



Suppose the histogram was left on the breakfast table, and Alan spilled coffee on the graph. Unfortunately, reading the frequency scale is no longer possible, nor is it possible to tell whether the graph is a frequency histogram or a relative frequency histogram.

a. Determine the probability that the victims were injected with diazepam 13 or more hours before they were found (the shaded bars). Describe your process.

b. Does this diazepam histogram support the findings of the histograms for the morphine in Question **1**? Justify your response.

Terry Souhrada, University of Montana, Missoula, MT

The goal of this activity is to give your students a short and simple snapshot into a very extensive math topic. TI and NCTM encourage you and your students to learn more about this topic using the extensions provided below and through your own independent research.

Extensions

Extension 1

Histograms are not smooth, and the width and endpoints of the bins can have a tremendous effect on its shape. Because of this, histograms are not the most reliable tools for determining probabilities. **Probability distribution functions** tend to produce better results. One of the most widely used probability distribution function is the **normal curve** (which is sometimes referred to as the "bell curve" because of its appearance).

Use the Web sites below to learn more about the normal curve.

- **http://www-stat.stanford.edu/~naras/jsm/NormalDensity/NormalDensity.html**
- **http://www.netnam.vn/unescocourse/statistics/58.htm**
- **http://mathbits.com/MathBits/TISection/Statistics2/normaldistribution.htm**
- **http://www.math.csusb.edu/faculty/stanton/m262/normal_distribution/**

1. What are some basic properties of the normal curve?

2. Do you think that the data used in this activity could be modeled with a normal curve? Give justification for your answer.

Extension 2

Oftentimes, histograms do not match a particular probability distribution function. So, statisticians find probability distribution curves using various smoothing techniques. One of these techniques is called **kernel density estimation**. Write a report about what the kernel density estimation is and how the resulting probability distribution function is constructed. Use the Web site below as a reference.

http://www.maths.uwa.edu.au/~duongt/seminars/intro2kde/

Terry Souhrada, University of Montana, Missoula, MT