Section 1.1: Three Kinds of Histograms, How Probability Histograms Converge to Density Functions for Continuous Data, and How Relative Frequency Histograms Converge to Probability Mass Functions for Discrete Data

by Professor T. Fiore

Histograms are useful for visualizing the distribution of a list of single numerical measurements, i.e. histograms are useful for visualizing *univariate numerical data.* There are three different kinds of histograms to visualize a list of single numerical measurements:

(1) frequency histograms,

(2) relative frequency histograms, and

(3) probability histograms.

The horizontal axis in all three is called the "measurement axis" and its units are the units of the measurements. The horizontal axis should be labelled by what is being measured, along with the units of that measurement. The vertical axis is different in each kind of histogram. In the three respective kinds, the vertical axis is the "frequency axis," the "relative frequency axis," or the "density axis." *Always label the axes, and remember to put arrowheads only pointing in the positive directions. No arrowheads point in the negative directions.*

To make a histogram, we subdivide the axis of measurement into intervals, called "bins". The endpoints of the intervals should not be any data point. The rectangle above a bin is different in each kind of histogram:

(1) in a *frequency* histogram, the rectangle above a bin has *height equal to the frequency* of measurements in that bin, i.e. the height above a bin is the number of measurements in that bin,

(2) in a *relative frequency* histogram, the rectangle above a bin has *height equal to the relative frequency* of measurements in that bin, i.e. the height above a bin is the proportion of measurements in that bin,

(3) in a *probability* histogram, the rectangle above a bin has *AREA equal to the relative frequency* of measurements in that bin, i.e. the AREA above a bin is the proportion of measurements in that bin.

The key feature of the probability histogram is: the probability that a randomly selected measurement from the data set is in the union of some bins is the total area above those bins! Consequently, the total area of a probability histogram is 1.

Let's now draw by hand the three kinds of histograms for the data set consisting of the seven measurements 30, 41, 47, 47, 48, 51, 52 with equal bin width 10 and first bin starting at the number 29. These measurements are ages of children in months.

None of the histograms is a bar graph! A bar graph is a different thing! A bar graph is not a histogram of any kind. In a bar graph, each bar is over a category, not an interval. In a histogram, each rectangle is over an interval.

Frequency Histogram

Draw by hand the *frequency* histogram for the data set consisting of the measurements 30, 41, 47, 47, 48, 51, 52 with equal bin width 10 and first bin starting at the number 29. These measurements are ages of children in months. Remember to label the axes, properly use arrowheads, and label tick marks on both axes.

How many measurements are between 29 and 49? *Add* the appropriate *heights* in the frequency histogram to obtain your answer, and double check it by counting in the data set.

Relative Frequency Histogram

Draw by hand the *relative frequency* histogram for the data set consisting of the measurements 30, 41, 47, 47, 48, 51, 52 with equal bin width 10 and first bin starting at the number 29. These measurements are ages of children in months. Remember to label the axes, properly use arrowheads, and label tick marks on both axes.

What proportion of measurements are between 29 and 49? *Add* the appropriate *heights* in the relative frequency histogram to obtain your answer, and double check it by counting in the data set.

Probability Histogram

Draw by hand the *probability* histogram for the data set consisting of the measurements 30, 41, 47, 47, 48, 51, 52 with equal bin width 10 and first bin starting at the number 29. These measurements are ages of children in months. Remember to label the axes, properly use arrowheads, and label tick marks on both axes.

What proportion of measurements are between 29 and 49? *Add* the appropriate *areas* in the probability histogram to obtain your answer, and double check it by find the proportion of measurements in the data set between 29 and 49.

Recall the key feature of a probability histogram: the probability that a randomly selected measurement from the data set is in the union of some bins is the total area above those bins. What is the probability that a randomly selected measurement from the data set is between 29 and 49? Compare with previous answer.

Confirm that the total area of your probability histogram is 1.

How to Make the Three Kinds of Histograms in R

To make the frequency histogram with the specified interval endpoints, we use the following code.

```
measurements<-c(30, 41, 47, 47, 48, 51, 52)  #stores the vector of measurements
endpoints<-c(29,39,49,59)                    #stores the vector of bin endpoints
hist(measurements,endpoints,main="Frequency Histogram")   #makes it
```

Warning: If bin widths are not all equal, this R command will make a probability histogram instead! Try the following!

```
hist(measurements,c(28,39,49,59),main="Probability!")   #R converts it!
```

To make the relative frequency histogram with the specified interval endpoints, we continue with the following code.

```
h<-hist(measurements,endpoints)              #stores the frequency histogram
                                             #as object h
h$density<-h$counts/sum(h$counts)            #don't explain
plot(h,freq=FALSE,main="Relative Frequency Histogram")  #makes it
```

To have the two histograms next to each other, use the following code.

```
par(mfrow=c(1,2))              #divides the graphics window in 2
hist(measurements,endpoints,main="Frequency Histogram")     #makes it
plot(h,freq=FALSE,main="Relative Frequency Histogram")      #makes it
```

To make a probability histogram with the specified interval endpoints, use the following code.

```
hist(measurements,endpoints,main="Probability Histogram",prob=TRUE)
```

Probability Histograms Converge to a Density when We Have Continuous Data with Many Data Points

See the handwritten handout.

If we have many data points, and the data is “continuous,” then we can approximate a *probability density function* by taking smaller and smaller bins in probability histograms. When taking smaller and smaller bins and more and more data points of continuous data, the probability histograms converge to a probability density function. On the other hand, the frequency histograms and relative frequency histograms of continuous data of course do *not* converge to the probability density function, see the handwritten handout for an illustration.

The key feature of a probability density function of a continuous random variable: *the probability the random variable is in a given interval is the area under the density function above that interval.* The reason is that probability histograms approach probability density functions for continuous data when we take smaller and smaller bins and more and more data!

Relative Frequency Histograms Converge to a Probability Mass Function when We Have Discrete Data with Many Data Points

See homework problem 1.1 #8 and the text of 1.1.

If we have many data points, and the data is "discrete," then we can approximate a *probability mass function* by making a relative frequency histogram with exactly one (possibly repeated) measurement number in each bin. When taking more and more data points of discrete data and keeping the bins the same so that one measurement number is in each, the relative frequency histograms converge to a probability mass function function. On the other hand, the frequency histograms and probability histograms of discrete data of course do *not* converge to the probability mass function.

Here is the solution of 1.1 #8. First realize that when all the bins have width 1, the relative frequency histogram of a univariate data set looks exactly like its probability histogram. To implement the answer to 1.1 #8 simply in R, we can just use the command for probability histogram to make the relative frequency histogram because the bin widths can be taken to be 1 this time because the only possible outcomes are 1, 2, 3, 4. Using the command for probability histogram to make the relative frequency histogram (when possible) is more convenient because the commands for relative frequency are more complicated. We take the bin endpoints to be the midpoints between whole numbers because we want the whole numbers to be in the center of the bin for easy reading, and the bin widths to be 1. For clarity, endpoints should not be measurement numbers. Non-statistical consumers of your data presentations should be able to easily follow without ambiguity.

(a)

```
x<-c(1,2,3,4)

fx<-c(2/10,3/10,3/10,2/10)
```

(b) We are sampling with replacement, so turn that option on as follows.

```
bowl<-c(1,1,2,2,2,3,3,3,4,4)

simulateddata<-sample(bowl,size=100,replace=TRUE)
```

(c)

```
plot(x,fx,type="h",col="red",xlim=c(0,5),ylim=c(0,4/10),ylab=
"PMF and Rel. Frequ.",main="PMF and Relative Frequency Histogram")

endpoints<-c(.5,1.5,2.5,3.5,4.5)

hist(simulateddata,endpoints,prob=TRUE,add=TRUE)
```