4.2: Histograms

Introduction

For displaying interval or continuously scaled data, a histogram (frequency or density distribution) is a useful graph to summarize patterns in data, and is commonly used to judge whether or not the sample distribution approximates a normal distribution. Three kind of histograms exist, depending on how the data are grouped and counted. Lump the data into a sequence of adjacent intervals or **bins** (aka **classes**), then count how many individuals have values that fall into one of the bins — the display is referred to as a **frequency histogram**. Sum up all of the frequencies or counts in the histogram and they add to the sample size. Convert from counts to percentages, then the heights of the bars are equal to the relative frequency (percentage) — the display is referred to as a **percentage histogram** (aka **relative frequency histogram**). Sum up all of the bin frequencies and they equal one (100%).

Figure 4.2.1 shows two frequency histograms of the distribution of ages for female (left panel) and male (right panel) runners at the 2013 Jamba Juice Banana 5K race in Honolulu, Hawaii (link to data set).

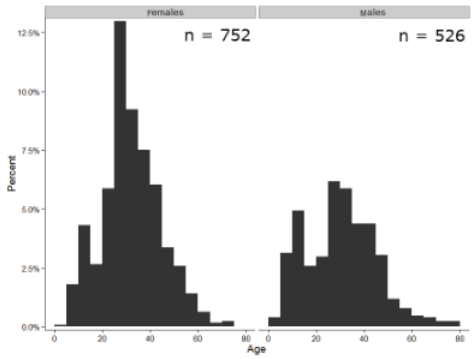

Figure 4.2.1: Histograms of age distribution of runners who completed the 2103 Jamba Juice 5K race.

The graphs in Fig. 4.2.1 were produced using R package `ggplot2` .

The third kind of histogram is referred to as a **probability density histogram**. The height of the bars are the **probability densities**, generally expressed as a decimal. The probability density is the bin probability divided by the bin width (size). The area of the bar gives the bin probability and the total area under the curve sums to one.

Which to choose? Both relative frequency histograms and density histograms convey similar messages because both “sum to one” (100%), i.e., bin width is the same across all intervals. Frequency histograms may have different bin widths; with more numerous observations, the bin width is larger than with cases with fewer observations.

Purpose of the histogram plot

The purpose of displaying the data is to give you or your readers a quick impression of the general distribution of the data. Thus, from our histogram one can see the range of the data and get a qualitative impression of the variability and the central tendency of the data.

Kernel density estimation

Kernel density estimation (KDE) is a **non-parametric** approach to estimate the probability distribution function. The “**kernel**” is a window function, where an interval of points is specified and another function is applied only to the points contained in the window. The function applied to the window is called the **bandwidth**. The **kernel smoothing function** then is applied to all of the data, resulting in something that looks like a histogram, but without the discreteness of the histogram.

The chief advantage of kernel smoothing over use of histograms is that histogram plots are sensitive to bin size, whereas KDE plot shapes are more consistent across different kernel algorithms and bandwidth choices.

Today, statisticians use kernel smoothing functions instead of histograms; these reduce the impact that binning has on histograms, although kernel smoothing still involves choices (Type of smoothing function? Default is Gaussian. Widths or bandwidths for smoothing? Varies, but the default is from the variance of the observations). Figure 4.2.2 shows a smoothed plot of the 752 age observations.

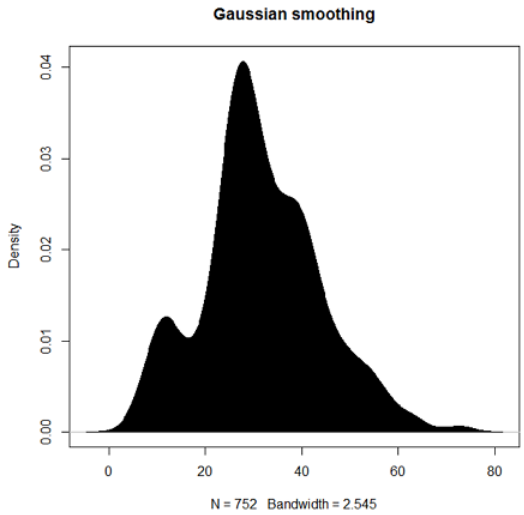

Figure 4.2.2: KDE plot of age distribution of female runners who completed the 2103 Jamba Juice 5K race in Honolulu.

Note:

Remember: the hashtag # preceding R code is used to provide comments and is not interpreted by R.

R commands typed at the R prompt were, in order:

```
d <- density(w) #w is a vector of the ages of the 752 females
plot(d, main="Gaussian smoothing")
polygon(d, col="black", border="black") #col and border are settings which allows you
```

Conclusion? A histogram is fine for most of our data. Based on comparing the histogram and the kernel smoothing graph I would reach the same conclusion about the data set. The data are right skewed, maybe kurtotic (peaked), and not normally distributed (see Ch 6.7).

Design criteria for a histogram

The X axis (horizontal axis) displays the units of the variable (e.g., age). The goal is to create a graph that displays the sample distribution. The short answer here is that there is no single choice you can make to always get a good histogram — in fact, statisticians now advise you to use a kernel function in place of histograms if the goal is to judge the distribution of samples.

For continuously distributed data the X-axis is divided into several intervals or bins:

1. The number of intervals depends (somewhat) on the sample size and (somewhat) on the range of values. Thus, the shape of the histogram is dependent on your choice of intervals: too many bins and the plot flattens and stretches to either end (over-smoothing); too few bins and the plot stacks up and the spread of points is restricted (under-smoothing). For both you lose the details of the histogram shape.
2. A general rule of thumb: try to have 10 to 15 different intervals. This number of intervals will generally give enough information.
3. For large sample size (N=1000 or more) you can use more intervals.

The intervals on the X-axis should be of equal size on the scale of measurement.

1. They will not necessarily have the same number of observations in each interval.
2. If you do this by hand you need to first determine the range of the data and then divide this number by the number of categories you want. This will give you the size of each category (e.g., range is 25; 25 / 10 = 2.5; each category would be 2.5 units).

For any given X category the Y-axis then is the number or frequency of individuals that are found within that particular X category.

Software

Rcmdr: Graphs → Histogram…

Accepting the defaults to a subset of the 5K data set, we get Fig. 4.2.3:

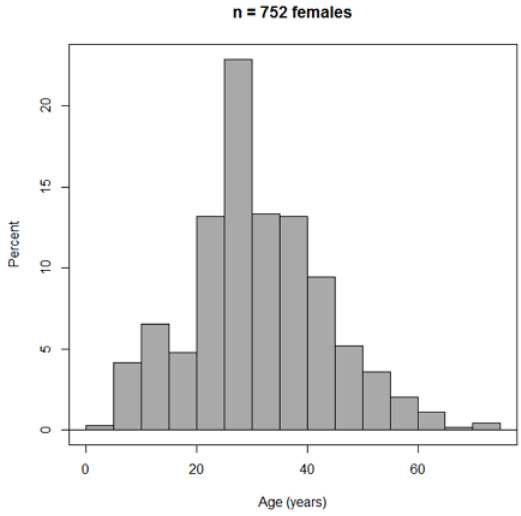

Figure 4.2.3: Histogram of 752 observations, **Sturge's rule** applied, default histogram.

The subset consisted of all females or n = 752 that entered and finished the 5K race with an official time.

R Commander plugin KMggplot2

Install the RcmdrPlugin.KMggplot2as you would any package in R. Start or restart Rcmdr and load the plugin by selecting **Rcmdr: Tools → Load Rcmdr plug-in(s)…** Once the plugin is installed select **Rcmdr: KMggplot2 → Histogram…** The popup menu provides a number of options to set to format the image. Settings for the next graph were No. of bins "Scott," font family "Bookman," Colour pattern "Set 1," Theme "theme_wsj2."

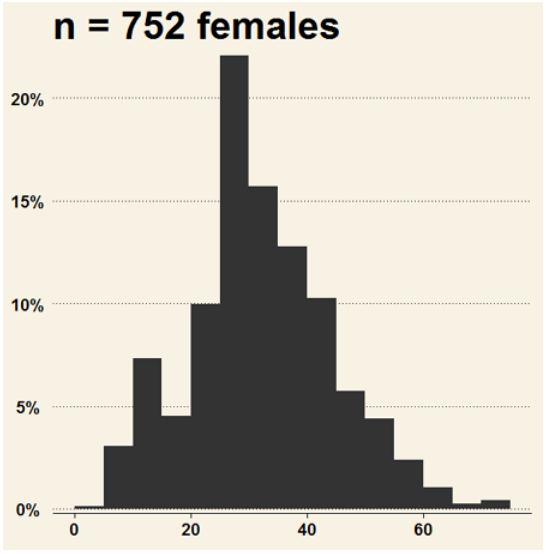

Figure 4.2.4: Histogram of 752 observations, Scott's rule applied, ggplot2 histogram.

Selecting the correct bin number

You may be saying to yourself, wow, am I confused. Why can't I just get a graph by clicking on some buttons? The simple explanation is that the software returns defaults, not finished products. It is your responsibility to know how to present the data. Now, the perfect graph is in the eye of the beholder, but as you gain experience, you will find that the default intervals in R bar graphs have too many categories (recall that histograms are constructed by lumping the data into a few categories, or bins, or

intervals, and counting the number of individuals per category => “density” or frequency). How many categories (intervals) should we use?

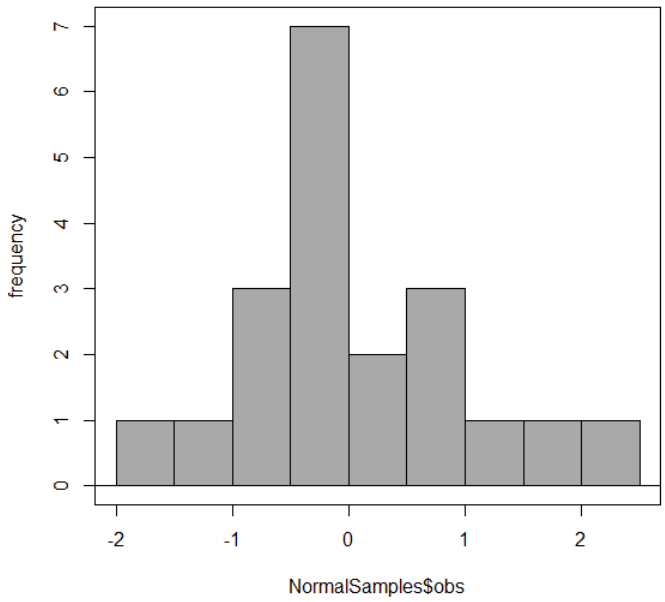

Figure 4.2.5: Default histogram with default bin size.

R’s default number for the intervals seems too much to me for this data set; too many categories with small frequencies. A better choice may be around 5 or 6. Change number of intervals to 5 (click Options, change from automatic to number of intervals = 5). Why 5? Experience is a guide; we can guess and look at the histograms produced.

Improving estimation of bin number

I gave you a rough rule of thumb. As you can imagine, there have been many attempts over the years to come up with a rational approach to selecting the intervals used to bin observations for histograms. The histogram function in Microsoft’s Excel (Data Analysis plug-in installed) uses the square root of the sample size as the default bin number. **Sturge’s rule** is commonly used, and the default choice in some statistical application software (e.g., Minitab, Prism, SPSS). **Scott’s** approach (Scott 2009), a modification to Sturge’s rule, is the default in the `ggplot()` function in the R graphics package (the Rcmdr plugin is `RcmdrPlugin.KMggplot2`). And still another choice, which uses interquartile range (IQR), was offered by **Freedman and Diacones** (1981). Scargle et al (1998) developed a method, **Bayesian blocks**, to obtain optimum binning for histograms of large data sets.

What is the correct number of intervals (bins) for histograms?

- Use the square root of the sample size, e.g., in this case the sample size $n = 20$ and $\sqrt{n} = 4.5$, round to 5.
- Follow Sturges’ rule (to get the suggested number of intervals for a histogram, let k = the number of intervals, and $k = 1 + 3.322\,(\log_{10} n)$, where n is the sample size.) I got $k = 5.32$, round to nearest whole number = 5.
- Another option was suggested by Freedman and Diacones (1981): find IQR for the set of observations and then the solution to the bin size is $k = 2 \cdot \mathrm{IQR} \cdot n^{-1/3}$, where n is the sample size.

Select Histogram Options, then enter intervals. Here’s the new graph using Sturge’s rule…

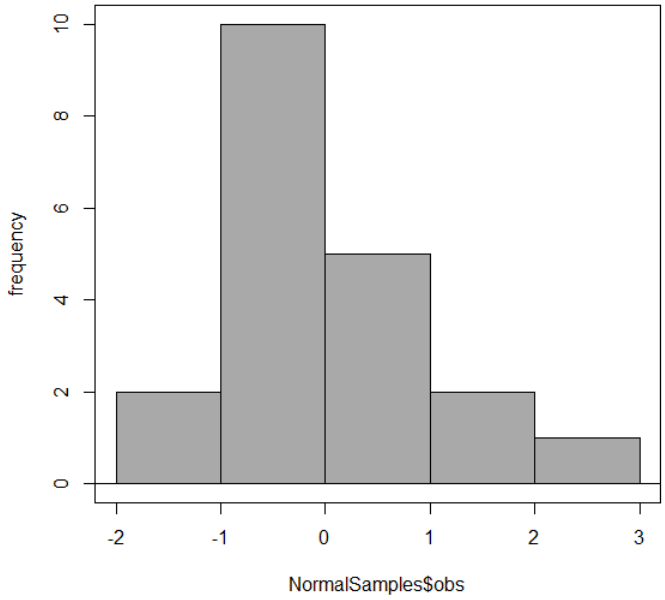

Figure 4.2.6: Default histogram, bin size set by Sturge's rule.

OK, it doesn't look much better. And of course, you'll just have to trust me on this — it is important to try to make the bin size appropriate given the range of values you have in order for the reader/viewer can judge the graphic correctly.

Questions

Example data set, comet tails and tea

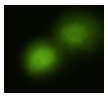

Figure 4.2.7: Examples of comet assay results.

The **Comet assay**, also called the single cell gel electrophoresis (SCGE) assay, is a sensitive technique to quantify DNA damage from single cells exposed to potentially mutagenic agents. Undamaged DNA will remain in the nucleus, while damaged DNA will migrate out of the nucleus (Figure 4.2.7). The basics of the method involve loading exposed cells immersed in low melting agarose as a thin layer onto a microscope slide, then imposing an electric field across the slide. By adding a DNA selective agent like Sybr Green, DNA can be visualized by fluorescent imaging techniques. A "tail" can be viewed: the greater the damage to DNA, the longer the tail. Several measures can be made, including the length of the tail, the percent of DNA in the tail, and a calculated measure referred to as the Olive Moment, which incorporates amount of DNA in the tail and tail length (Kumaravel et al 2009).

The data presented in Table 1 comes from an experiment in my lab; students grew rat lung cells (ATCC CCL-149), which were derived from type-2 like alveolar cells. The cells were then exposed to dilute copper solutions, extracts of hazel tea, or combinations of hazel tea and copper solution. Copper exposure leads to DNA damage; hazel tea is reported to have antioxidant properties (Thring et al 2011).

Data set, comet assay

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	10	9.7732	2.1501
Copper-Hazel	6	4.8381	0.9676
Copper-Hazel	6	3.981	0.836
Copper-Hazel	16	12.0911	2.9019

Treatment	Tail	TailPercent	OliveMoment*
Copper-Hazel	20	15.3543	3.9921
Copper-Hazel	33	33.5207	10.7266
Copper-Hazel	13	13.0936	2.8806
Copper-Hazel	17	26.8697	4.5679
Copper-Hazel	30	53.8844	10.238
Copper-Hazel	19	14.983	3.7458
Copper	11	10.5293	2.1059
Copper	13	12.5298	2.506
Copper	27	38.7357	6.9724
Copper	10	10.0238	1.9045
Copper	12	12.8428	2.5686
Copper	22	32.9746	5.2759
Copper	14	13.7666	2.6157
Copper	15	18.2663	3.8359
Copper	7	10.2393	1.9455
Copper	29	22.6612	7.9314
Hazel	8	5.6897	1.3086
Hazel	15	23.3931	2.8072
Hazel	5	2.7021	0.5674
Hazel	16	22.519	3.1527
Hazel	3	1.9354	0.271
Hazel	10	5.6947	1.3098
Hazel	2	1.4199	0.2272
Hazel	20	29.9353	4.4903
Hazel	6	3.357	0.6714
Hazel	3	1.2528	0.2506

Rat lung cells treated with Hazel tea extract and exposed to copper metal. Tail refers to length of the comet tail, TailPercent is percent DNA damage in tail, and Olive moment refer's to Olive (1990), defined as the fraction of DNA in the tail times the tail length.

Copy the table into a data frame.

1. Create histograms for tail, tail percent, and olive moment
   - Change bin size
2. Repeat, but with a kernel function.
3. Looking at the results from question 1 and 2, how “normal” (i.e., equally distributed around the middle) do the distributions look to you?

4. Plot means to compare Tail, Tail percent, and olive moment. Do you see any evidence to conclude that one of the teas protects against DNA damage induced by copper exposure?
