

Lecture 4: Histograms, Averages, Standard Deviations and the Normal Approximation for Data

Here we simply summarize the basic facts, see most statistics books, including our texts "Stat Labs" by Nolan and Speed, abbreviated NS, "Primer of Biostatistics" by Glantz, abbreviated G, or "Statistics" by Freedman, Pisani and Purves (3rd ed), abbreviated FPP.

Histograms (NS ch 1, FPP ch 3)

1. A *histogram* represents percents by area. It consists of a set of blocks. The area of each block represents the percentage of cases in the corresponding *class interval*.

To draw a histogram, begin with a distribution table of percentages in class intervals. Put down the horizontal axis, then begin to draw the blocks. To figure out the height of a block over a class interval, divide the percent by the length of the interval.

2. With the *density scale*, the height of each block equals the percentage of cases in the corresponding class interval, divided by the length of that interval.
3. With the density scale, area comes out in percent, and the total area is 100%. The area under the histogram between two values gives the percentage of cases falling in that interval.
4. We can define density curves as "histograms" drawn as curves above the horizontal axis having the interpretation given in 3.

The Average and the Standard Deviation (NS ch 1, G ch 2, FPP ch 4)

1. A typical list of numbers can be summarized by its *average* and its *standard deviation*.
2. The average of a list (also called the arithmetic mean) is the sum of the entries divided by the number of entries.
3. The average located the center of a histogram in the sense that the histogram balances when supported at the average.
4. Half the area under a histogram lies to the left of the *median* and half to the right. The median is another way to locate the center of a histogram.
5. The *r.m.s.* of a list of numbers measures how big the entries are, neglecting signs.
6. r.m.s. size of a list = $\sqrt{(\text{average of } (\text{entries})^2)}$

7. The SD measures distance from the average. Each number on a list is off the average by some amount. The SD is a sort of average size for these amounts off.
8. The SD is the r.m.s. size of the deviations from the average:

$$\text{SD} = \sqrt{(\text{average of (deviations from the average)}^2)}$$

9. Roughly 68% of the entries on a list of numbers are within one SD of the average, and about 95% are within two SDs of the average. This is true for many lists, but not all.

The Normal Approximation For Data (FPP ch 5)

1. The *normal curve* is symmetric about 0, and the total area under it is 100%.
2. *Standard units* say how many SDs a value is, above (+) or below (-) the average.
3. Many histograms have roughly the same shape as the normal curve.
4. If a list of numbers follows the normal curve, the percentage of entries falling in a given interval can be estimated by converting the interval to standard units and then finding the corresponding area under the normal curve. This procedure is called the *normal approximation*.
5. A histogram which follows the normal curve can be reconstructed fairly well from its average and SD; in such cases the average and SD are good summary statistics.

The 25th percentile of a distribution is that value having 25% of all values to the left of it and 75% of all values to the right. It is also called the *lower quartile*. The 75th percentile is defined analogously with 25% replaced by 75% and is also called the *upper quartile*. The *inter-quartile range* is the difference between the upper and the lower quartiles.

6. All histograms, whether or not they follow the normal curve, can be summarized using *percentiles*.