# Chapter 2
# Data Summaries, Histograms, Wide and Narrow Gaussian Curves

## 1 General Purpose

In order to summarize continuous data, either histograms can be plotted or Gaussian curves can be drawn. This chapter is to assess how to summarize your data with the help of a pocket calculator.

## 2 Schematic Overview of Type of Data File

| Outcome |
|---|
| . |
| . |
| . |
| . |
| . |
| . |
| . |
| . |
| . |

## 3 Primary Scientific Question

How can histograms, otherwise called frequency distributions, and wide and narrow Gaussian curves be used for summarizing continuous data?

T.J. Cleophas, A.H. Zwinderman, *Clinical Data Analysis on a Pocket Calculator*,
DOI 10.1007/978-3-319-27104-0_2

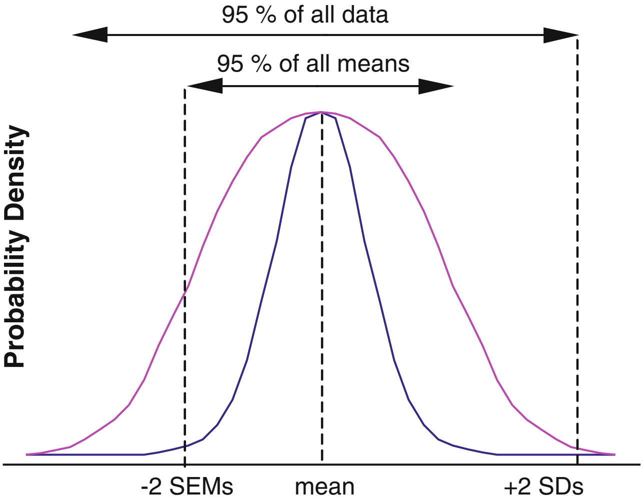
## 4 Hypothesized Data Example

Based on the same data, but with different meaning. The wide one summarizes the data of our trial. The narrow one summarizes the mean of many trials similar to our trial. It can be mathematically proven that this is so.

Continuous data can be plotted in the form of a histogram (upper graph). The upper graph is assumed to present individual cholesterol reductions after one week drug treatment in 1000 patients. The bar in the middle is observed most frequently. The bars on either side grow gradually shorter. The graph, thus, pretty well exhibits two main characteristics of your data namely the place of the mean and the distribution of the individual data against the mean. On the x-axis, frequently called z-axis in statistics, it has individual data. On the y-axis it has "how often". This graph adequately represents the data. It is, however, not adequate for statistical analyses. The lower graph pictures a Gaussian curve, otherwise called normal (distribution) curve. On the x-axis we have, again, the individual data, expressed either in absolute data or in SDs (standard deviations) distant from the mean. See Chap. 1 for calculation information of SDs. On the y-axis the bars have been replaced with a continuous line. It is now impossible to determine from the graph how many patients had a particular outcome. Instead, important inferences can be made. For example, the total area under the curve (AUC) represents:

100 % of the data,
AUC left from mean represents 50 % of the data,
left from −1 SDs it has 15.9 % of the data,
left from −2 SDs it has 2.5 % of the data,
between +2 SDs and −2 SDs we have 95 % of the data
(the 95 % confidence interval of the data).

It is remarkable that the patterns of Gaussian curves from biological data are countless, but that all of them, nonetheless, display the above percentages. This is something like a present from heaven, and it enables to make use of the Gaussian curves for making predictions from your data about future data. However, in order to statistically test your data, we will have to go one step further.



The figure underneath gives two Gaussian curves, a narrow and a wide one. Both are based on the same data, but with different meaning. The wide one summarizes the data of our trial. The narrow one summarizes the mean of many trials similar to our trial. It can be mathematically proven that this is so. However this is beyond scope of the current text. Still, it is easy to conceive that the distribution of all means of many similar trials is narrower and has fewer outliers than the distribution of the actual data from our trial, and that it will center around the mean of our trial, because our trial is assumed to be representative for the entire population. Now why should the mean of many trials be equal to the mean of our trial. The truth is, we have no certainty, but neither do we have any argument to have the overall mean on the left of right side of the measured mean of our data. You may find it hard to believe, but the narrow curve with standard errors of the mean (SEMs), or, simply, SEs on the x-axis can be effectively used for testing important statistical hypotheses, like

1. no difference between new and standard treatment,
2. a real difference,
3. the new treatment is better than the standard treatment,
4. the two treatments are equivalent.

thus, mean ± 2 SDs (or more precisely 1.96 SDs) represents 95 % of the AUC of the wide distribution, otherwise called the 95 % confidence interval of the data, which means that 95 % of the data of the sample are within. The SEM-curve (narrow one) is narrower than the SD-curve (wide one), because SEM $= \text{SD}/\sqrt{n}$ with n = sample size. Mean ± 2 SEMs (or more precisely 1.96 SEMs) represents 95 % of the means of many trials similar to our trial.

$$\text{SEM} = \text{SD}\,/\sqrt{n}$$

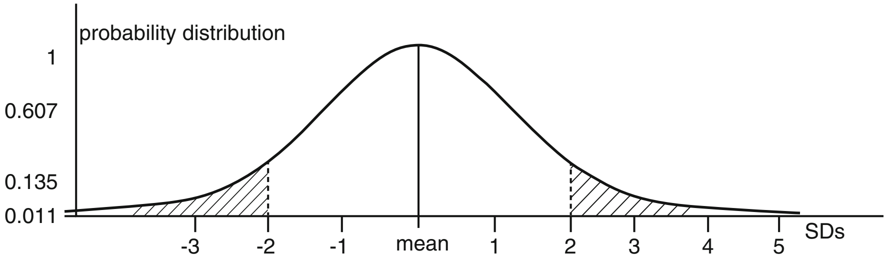
## 5 Importance of SEM-Graphs in Statistics

Why is this SEM approach so important in statistics. Statistics makes use of mean values and their standard error to test the null hypotheses of finding no difference from zero in your sample. When we reject a null hypothesis at $p < 0.05$, it literally means that there is <5 % chance that the mean value of our sample crosses the area of the null hypothesis where we say there is no difference. It does not mean that many individual data may not go beyond that boundary. Actually, it is just a matter of agreement. But it works well.

## 6 Drawing a Gaussian Curve without a Computer

The mathematical equation of a Gaussian curve is (e = Euler's constant = 2.718)

$y = e^{-1/2\ (x\wedge 2)}$
x = (individual values) / (standard deviation)
$x\wedge 2 = x^2$
at $x = 0 \rightarrow y = 1$
at $x = 1 \rightarrow y = 0.607$
at $x = 2 \rightarrow y = 0.135$
at $x = 3 \rightarrow y = 0.011$



With the help of the above equation, various y-values can, thus, be computed, and in this way a standard Gaussian curve can be drawn.

## 7 Conclusion

In order to summarize continuous data, either histograms can be plotted or Gaussian curves can be drawn. Gaussian curves can be drawn with the help of the Gaussian curve equation $y = e^{-1/2\ (x\ \hat{}2).}$. The above procedure is only entirely correct with larger samples like 100 or so. With small samples data tend to produce somewhat larger spread, and normal distributions turn into t-distributions (see the Chap. 8). But as a first step, before any analysis, histograms and Gaussian curves are convenient even with small samples.

## 8 Note

More background, theoretical and mathematical information of histograms and Gaussian curves is given in Statistics applied to clinical studies 5th edition, Chap. 1, Springer Heidelberg Germany, 2012, from the same authors.