4 A historical phenomenology

4.1 Purpose

Various authors have suggested that studying the history of a topic is good preparation for teaching that topic (Dijksterhuis, 1990; Fauvel & Van Maanen, 2000; Freudenthal, 1983b; Gulikers & Blom, 2001; Radford, 2000). The obstacles that people in the past grappled with are interesting to teachers and designers because students often encounter similar obstacles. However, students also know things that people in the past did not know. What is the relationship between the historical development of statistical and mathematical concepts (phylogenesis) and the individual development of students (ontogenesis)? For a discussion of this question we refer to Radford (2000) for a mathematics education perspective and to Cole (1996) for a cultural psychology perspective. Here we confine ourselves to the question of what we can learn from a historical study for an instruction theory for early statistics education. We used the RME heuristic of guided reinvention as a general guideline for the type of instruction we aimed for and we used a historical phenomenology as a method for studying the relation between the phenomena that were organized and the statistical concepts that historically arose for organizing such phenomena (2.1). Freudenthal (1983b) envisioned the following process of guided reinvention:

The young learner recapitulates the learning process of mankind, though in a modified way. He repeats history not as it actually happened but as it would have happened if people in the past would have known something like what we do know now. It is a revised and improved version of the historical learning process that young learners recapitulate.

‘Ought to recapitulate’—we should say. In fact we have not understood the past well enough to give them this chance to recapitulate it. (p. 1696)

Taking this cue from Freudenthal we decided to study the early history of statistics, in particular of averages, sampling, distribution, and graphs, which were the basic ingredients of the intended instructional sequence. In combination with a didactical phenomenology we could then conjecture on what “a revised and improved version of the historical learning process” for this particular topic might look like. As such this historical phenomenology also prepares the development of a hypothetical learning trajectory.

In this chapter, we first discuss the method of our historical phenomenology, and then we study the average values (mean, midrange, mode), sampling, median, distribution, and finally graphs. The last section summarizes the insights that were most informative for the instructional design.

10. Sections 4.1 to 4.3 are based on Bakker (2003).
4.2 Method

When explaining what he meant by a phenomenology, Freudenthal (1983a) started his exposé as follows:

I start with the antithesis—if it really is an antithesis—between nooumenon (thought object) and phainomenon. The mathematical objects are nooumena, but a piece of mathematics can be experienced as a phainomenon; numbers are nooumena, but working with numbers can be a phainomenon. (p. 28)

As Freudenthal wrote, there is a philosophical difficulty when distinguishing phenomena and concepts: it is hard to separate phenomena from concepts, since concepts also determine and influence how humans perceive the phenomena (e.g. Kant, 1787/1974). From a psychological perspective this distinction is also hard to make. For example, it is not exactly clear if and how humans’ perception of colors depends on the terms that are available in their language (Anderson, 1995). For an educational purpose, it is still useful to try and separate phenomena and concepts, because students do not perceive the same phenomena as we do because of our understanding of certain concepts (an example of this is given in Chapter 9). Where statisticians see a clear pattern in a graph of a signal with noise (Konold & Pollatsek, 2002), students might just see a bunch of dots. Studying history can help us see certain phenomena through the eyes of people who did not have the same concepts and techniques as we have nowadays. By analyzing the historical process we expect to be able to identify different layers and aspects of concepts that seem to be fixed products nowadays. This attention to historical development may help us take a student’s perspective and to better understand and guide the learning process.

The method of historical phenomenology requires identifying phenomena that have been organized by certain concepts and identifying concepts that have been applied to get a handle on certain phenomena. This historical phenomenology could be used to feed a didactical phenomenology, especially to find phenomena that challenge students to develop particular statistical methods or concepts. We have found no research literature in which this was already done for the statistical concepts that we were interested in, with the exception of Steinbring (1980), who writes about the development of chance and distribution with a didactical interest. Nor have we found historical phenomenologies that describe a systematic method.

Applying the method outlined above, we first collected as many historical phenomena with a statistical flavor as possible, mainly about center, sampling, distribution, and graphs (before 1900). Doing this we experienced several difficulties. One was that most histories of mathematics hardly pay any attention to the history of statistics. This need not be very surprising, because statistics arose largely from disciplines other than mathematics such as geodesy, astronomy, navigation, metallurgy, political arithmetic, medicine, anthropometry, biology, and social sciences. A second difficulty was that most historical studies of statistics start at around 1660,
which many authors mark as the start of probability and statistics (e.g. Hacking, 1975; Kendall, 1960), and focus on the nineteenth century (Porter, 1986; Stigler, 1986), whereas we had to go further back in time for the origins of the concepts we were interested in.

The next step, the selection of historical examples, was guided by the educational potential we saw in them; we only selected examples that could be regarded as preliminary stages of statistical notions with possible relevance for the design. For example, if an estimation of a number of years was reached by some method that could be interpreted as an intuitive version of an average, it was included in our phenomenology. A simple guess of a large number would not have been included. We also give an example from geometry to show that what might sound statistical (‘arithmetic mean’) need not be statistical. After many historical examples we formulate a hypothesis about students’ learning, indicated with H#. Some of these hypotheses are revisited in the didactical phenomenology and in the retrospective analyses in the forthcoming chapters, but we have not been able to test them all.

Our methodology and purpose differ from what is common in historical science. In our historical study we do not describe the historical development of concept, but try to find sources of inspiration. We certainly do not want to suggest that ‘the’ historical development of statistical concepts was an accumulation of insights or a continuous refinement of concepts. As many authors nowadays note, revolutions and ruptures occurred in the history of mathematics and statistics as well (Gillies, 1992; Krüger et al., 1989). Nor do we want to suggest that students’ learning always needs to follow the historical development. In this chapter we give a few examples in which students’ cultural knowledge, for instance about surveys, makes it inefficient to follow the historical order. Although center, distribution, sampling, and graphs are highly interrelated topics, we address them separately for reasons of readability. There are six remaining sections: 4.3 on the average values (excluding the median), 4.4 on sampling, 4.5 on the median, 4.6 on distribution, 4.7 on graphs, and 4.8 is a summary of the most important results. In Chapter 5 we revisit several issues to investigate possible didactical consequences.

4.3 Average

If we use the term ‘average’ or ‘average values’ we refer to arithmetic mean, median, mode, midrange, and precursors of those measures of center. Because we study the early history of statistics as a source of inspiration for instruction to young students, we do not make technical distinctions between, for example, the mean of a population and the sample mean to estimate the center of a distribution. The terms ‘center’ and ‘location’ are also used informally for the center of a data set or a distribution.
4.3.1 Average values to estimate a total

The oldest historical examples that we considered relevant for the historical phenomenology all had to do with estimation of large numbers. Three of them are presented below to illustrate preliminary stages of several average values.

Example 1. Number of leaves on a branch

In an ancient Indian story, which was finally written down in the fourth century AD, the protagonist Rātunāna estimated the number of leaves and fruit on two great branches of a spreading tree (Hacking, 1975). He estimated the number on the basis of one single twig, which he multiplied by the estimated number of twigs on the branches. He estimated 2095, which after a night of counting turned out to be very close to the real number. Although it is uncertain how Rātunāna chose the twig, it could well be that he chose an average-sized twig, since that would lead to a proper estimation.

The educational potential we saw in this example was that such an implicit use of a representative value could be an intuitive predecessor of the arithmetic mean because one average number represents all other twig numbers and this average number is somehow ‘in the middle’ of the others. The choice is presumably made in such a way that what is counted too much on the one hand is counted too little on the other hand. This use of an average has to do, in our modern eyes, with compensation, balance, and representativeness. Even if Rātunāna did not use the method we think he used, the problem situation inspired our instructional design to let students reinvent such a method. (For a similar example of estimating a large number, the number of years between the first and last king of Egypt, see Rubin, 1968, or Bakker, 2003).

Rubin (1971) has found other old examples of statistical reasoning in the work of one of the first scientific historians, Thucydides (circa 460-400 BC). The following two quotations are from his *History of the Peloponnesian War*. The reader is invited to decide how he or she would translate these two excerpts into modern statistical terms.

Example 2. Height of a wall of Platea (Figure 4.1)

(The problem was for the Athenians)... to force their way over the enemy’s surrounding wall... Their method was as follows: they constructed ladders to reach the top of the enemy’s wall, and they did this by calculating the height of the wall from the number of layers of bricks at a point which was facing in their direction and had not been plastered. The layers were counted by a lot of people at the same time, and though some were likely to get the figure wrong, the majority would get it right, especially as they counted the layers frequently and were not so far away from the wall that they could not see it well enough for their purpose. Thus, guessing what the thickness of a single brick was, they calculated how long their ladders would have to be...

(Rubin, 1971, p. 53)
Example 3. Crew size on ships

Homer gives the number of ships as 1,200 and says that the crew of each Boetian ship numbered 120, and the crews of Philoctetes were fifty men for each ship. By this, I imagine, he means to express the maximum and minimum of the various ships’ companies... If, therefore, we reckon the number by taking an average of the biggest and smallest ships... (Rubin, 1971, p. 53)

We interpret example 2 as an implicit use of the mode, here indicated by “the majority,” because “the majority” probably means “the most frequent value” and not necessarily “more than half.” In this situation, the Greeks probably assumed that the most frequent number would be the correct one. To find the total height of this number of bricks, they supposedly needed another estimation: the expected or the average thickness of a single brick.

Example 3 also illustrates an estimation that is based on an average value. Thucydides possibly interpreted the given numbers as the extreme values, so that the total amount of men on the ships could be estimated by taking the average of these two extremes. In fact this is called the midrange, defined as the arithmetic mean of the two extremes. This technique of averaging the extreme values of the range to obtain the midrange can be justified if certain assumptions are defensible, for instance that the underlying distribution is approximately symmetrical.

Resuming, in these historical examples we encountered phenomena that were organized by predecessors of contemporary statistical concepts. In examples 1 and 2, a kind of average similar to the arithmetic mean was probably used. In example 2, we
can also recognize the mode. In example 3, Thucydides described a method that we can call taking the midrange. In these estimation examples these notions of average were not defined or used explicitly, although many mean values were known in those days (Heath, 1981). The median, however, was absent in the early examples. Eisenhart (1974), who investigated these issues in detail, has found no possible precursors to the median before 1599. The conjecture that arises from these historical examples is the following.

H1. *Estimation of large numbers could challenge students to use intuitive notions of average.*

In Section 6.3 we describe how we tested and confirmed this conjecture in seventh-grade classes.

### 4.3.2 Mean values in Greek geometry

Explicit use of mean values and names for these values are found in ancient Greek mathematics. In Pythagoras’ time, around 500 BC, three mean values were known, namely the harmonic, geometric, and arithmetic mean (Heath, 1981; Iamblichus, 1991). Only some 200 years later, at least eleven different mean values had been defined (Heath, 1981). For a historical phenomenology it is relevant to study the phenomena that gave rise to these concepts. It turns out that the theory of the three mentioned mean values was developed with reference to music theory, geometry, and arithmetic. We provide an example of the mean values in geometry to illustrate that these mean values were not used in a statistical way. Yet we can learn from the geometrical representation and the definitions of the mean values.

This example from geometry, a theorem of Pappus, illustrates that the Greeks studied the mean values for their geometrical beauty (see Figure 4.2) and not in a statistical sense. If in the semicircle ADC with center O one has DB \perp AC and BF \perp DO, then DO is the arithmetic mean, DB the geometric mean, and DF the harmonic mean of the magnitudes AB and BC (Boyer, 1991). This theorem does clearly not belong in a statistics course at the middle school level. Yet two aspects are important: the definitions of the arithmetic mean and the representation of magnitudes.

![Figure 4.2: Theorem of Pappus on arithmetic, geometric, and harmonic mean.](image)
In addition to the Greek definition of the arithmetic mean, Aristotle (384-322 BC) defined a philosophical form of the mean, the “mean relative to us.” About the difference between the arithmetic mean and “the mean relative to us” he wrote:

By the mean of a thing I denote a point equally distant from either extreme, which is one and the same for everybody; by the mean relative to us, that amount which is neither too much nor too little, and this is not one and the same for everybody. For example, let 10 be many and 2 few; then one takes the mean with respect to the thing if one takes 6; since 10-6 = 6-2, and this is the mean according to arithmetical proportion [progression]. But we cannot arrive by this method at the mean relative to us. Suppose that 10 lb. of food is a large ration for anybody and 2 lb. a small one: it does not follow that a trainer will prescribe 6 lb., for perhaps even this will be a large portion, or a small one, for the particular athlete who is to receive it; it is a small portion for Milo, but a large one for a man just beginning to go in for athletics. (*Nichomachean Ethics*, book II, chapter vi, 5; italics added)

The description “not too much and not too little” for the average is one that students used in all of the seventh-grade teaching experiments in the context of estimation (6.3).

Figure 4.3: Greek representation of magnitudes as bars (2, 6, and 10)

In Greek mathematics, numbers and magnitudes were represented by lines. Aristotle’s example with the mean of 10 and 2 represented in the Greek way (Figure 4.3) illustrates that Greek mathematics had a different form and aim than modern mathematics: it was highly geometrical and visual. This difference between Greek and modern mathematics can also be demonstrated by the difference in definitions of the arithmetic mean. The Greek definition, as we saw in the quotation of Aristotle, is as follows: the middle number b of a and c is called the arithmetic mean if and only if a-b = b-c. Note that this definition differs in formulation from the equivalent modern one, \((a+c)/2\), and that it refers to only two values. The Greek version shows that the mean is in between the two extremes and that it is difficult to generalize, whereas the modern version emphasizes the calculation and is easy to generalize. With the didactical phenomenology in mind, it is important to note that the Greek definition shows other qualitative aspects than the modern quantitative one. For example, we can immediately see from the Greek definition that the mean is halfway between the two other values. This feature is used in Greek astronomy for interpolation (Ptolemy, 1998), but we consider this application as non-statistical. Yet we highlight this ‘in-
intermediacy’ aspect because students do not always realize that the mean is in between the extreme values (Strauss & Bichler, 1988). In a representation such as 4.3 students might be able to see that the part of the longest bar that ‘sticks out’ (compared with the middle bar) compensates the part of the shortest bar.

H2. The Greek bar representation might support the understanding that the mean is in between extreme values (intermediacy) and it might even scaffold a compensation strategy of visually estimating the mean.

In Section 6.7 we discuss how this conjecture was tested and confirmed in the teaching experiments.

4.3.3 Average, midrange, and generalization of the mean
In Section 4.3.1 we illustrated how the average was sometimes used implicitly in estimations and in Section 4.3.2 we conjectured that the Greek way of representing numbers by bars has educational potential for visual estimation of the mean. In the present section we describe how the average emerged from fair share in trade and insurance contexts, and that taking the mean of only two extreme values, the midrange, could be a predecessor of the arithmetic mean of more than two values in the context of science.

In the first millennium before Christ, the sea trade in the Mediterranean was lively (Plön & Kreutziger, 1965). During a storm, captains of small vessels with valuable merchandise sometimes needed to cut away the mast or throw some cargo overboard to avoid capsizing or to save the rest of the cargo. This act of throwing cargo overboard became known as the ‘jettison’ of cargo.

Figure 4.4: Part of the first page of a Dutch book on average by Weytsen (1641)
From about 700 BC, merchants and shippers agreed that damage to the cargo and the ship should be shared equally among themselves. What a merchant had to pay was called his ‘contribution’. This idea became part of customary law and was written down in the ‘lex Rhodia de iactu’, the Rhodian law on jettison during the codification of Roman Law in 534. The basic principle in the Digest XIV.2.1 is as follows.

The Rhodian law decrees that if in order to lighten the ship merchandise has been thrown overboard, that which has been given for all should be replaced by the contribution of all. (Lowndes & Rudolf, 1975, p. 3)

The rest of the text explains what should be done in specific situations and raises questions like, “In which proportion should compensation be paid?” Digest XIV.2.2.4 states that the equalized portion should take into account what the value of the saved and the lost cargo was. The number examples in the Latin texts are extremely simple and not very explicit. In Digest XIV.2.4.2 we read, for example:

If therefore, for instance, two persons each had merchandise valued at 20,000 sesterces and one lost 10,000 due to water damage, the one with the saved merchandise should contribute according to his 20,000, but the other on the basis of the 10,000. (Spruit, 1996; translation from Latin and Dutch)

Old Dutch books on average (e.g. Weytsen, 1641) were not very explicit either (Figure 4.4). In search of more realistic examples of calculations we resorted to books of the nineteenth century that describe how to calculate averages (e.g. Arnould & Macclachlan, 1872; Hopkins, 1859; Van der Hoeven, 1854). These averages were calculated by a so-called ‘average-adjuster’, who was a kind of accountant. This must have been a serious profession, because there was even an ‘Association of Average Adjusters’ in England in the nineteenth and early twentieth century (Lowndes & Rudolf, 1975).

From this law-historical account we can track the development of the average in maritime law. Important for this historical phenomenology is that the average’s origin is fair distribution and that proportions play a major role (cf. P11). But how did the term ‘average’ also come to signify the arithmetic mean?

The Oxford English Dictionary (Simpson & Weiner, 1989) writes that one of the meanings of ‘average’ in maritime law is “the equitable distribution of expense or loss, when of general incidence, among all the parties interested, in proportion to their several interests.” In its transferred use it came to signify the arithmetic mean:

The distribution of the aggregate inequalities (in quantity, quality, intensity, etc.) of a series of things among all the members of the series, so as to equalize them, and ascertain their common or mean quantity, etc. (...) the arithmetical mean so obtained.

From the examples in this section we see that this type of average originally arose
from the phenomena of fair share and insurance.

H3. Learning the mathematics that is involved in fair share and insurance (e.g. proportions and ratios) is good preparation for learning about the arithmetic mean. Fair share is also a suitable context to practice such skills (cf. Cortina et al., 1999).

Another possible precursor to the arithmetic mean is the midrange, which was used for example in Arabian astronomy of the ninth to eleventh century, but also in metallurgy and navigation (Eisenhart, 1974). Nowadays we model many observations and errors in those contexts with symmetrical distributions. Therefore, it is understandable that the midrange was used in those situations. Because the midrange was probably a precursor to the mean as a way to organize the center or estimate the true value, it might well be that students also use the midrange as a precursor to the mean.

H4. Students may use the midrange as a precursor to more advanced notions of average.

Not until the sixteenth century was it recognized that the arithmetic mean could be generalized to more than two cases: \( \bar{a} = (a_1 + a_2 + \ldots + a_n)/n \). Székely (1997) supposes that the invention of the decimal system by Stevin in 1585 facilitated such division calculations. This generalized mean proved useful for astronomers who wanted to know a real value, such as the position of a planet or the diameter of the moon. Using the mean of several measured values, scientists assumed that the errors added up to a relatively small number when compared to the total of all measured values. This method of taking the mean for reducing observation errors was mainly developed in astronomy, first by Tycho Brahe. From the late sixteenth century onwards, using the arithmetic mean to reduce errors gradually became a common method in other areas as well (Eisenhart, 1974; Plackett, 1970). This implies for our didactical phenomenology:

H5. Repeated measurement might be a useful instructional activity for developing understanding of the mean and distribution (cf. Konold & Pollatsek, 2002; Lehrer & Schauble, 2001). See also H12 and Section 10.4.

A question that arose was how we could benefit from the Greek definition and bar representation and still reach a general definition of the mean on \( n \) values (see Section 5.4).

4.3.4 **The mean as an entity in itself**

The historical examples until about the nineteenth century mostly had to do with approximating a real or best value, for example the number of leaves on a branch or
the diameter of the moon. In these old examples, the mean was used as a means to an end. It took a long time before the mean was used as a representative or substitute value as an entity in itself. The Belgian statistician Quetelet (1796-1874), famous as the inventor of *l’homme moyen*, the average man, was one of the first scientists to use the mean as the representative value for an aspect of a population. This transition from the real value to a representative value as a statistical construct was an important conceptual change (Porter, 1986; Stigler, 1986). What is relevant for the historical phenomenology is that there are several layers of understanding the mean as a representative value. We conjecture the following.

**H6. Using an average value in estimations of large numbers and using the mean for reduction of errors are probably easier for students than understanding the mean as an entity in itself, that is as a representative value for an aspect of a population.**

In Sections 5.1.6 and 5.1.7 we give some empirical support for this conjecture. It makes a difference if the mean as an entity in itself stands for a value that can exist or cannot exist. In 1877, Peirce—the same whom we revisit Chapters 8 and 9—wrote about this issue:

> In studies of numbers, the idea of continuity is so indispensable, that it is perpetually introduced even where there is no continuity in fact, as where we say that there are in the United States 10.7 inhabitants per square mile, or that in New York 14.72 persons live in the average house. [Footnote:] This mode of thought is so familiarly associated with all exact numerical consideration, that the phrase appropriate to it is imitated by shallow writers in order to produce the appearance of exactitude where none exists. Certain newspapers, which affect a learned tone, talk of “the average man,” when they simply mean *most men*, and have no idea of striking an average. (CP 2.646)

### 4.4 Sampling

#### 4.4.1 Estimation and sampling

Below the surface of the estimation examples, sampling issues also play a role. For instance, in the Indian story on estimating the number of leaves and fruit on a branch, the right twig had to be chosen to find an accurate average or representative value. Centuries later, John Graunt used a similar method of estimating the population of London, and Laplace to estimate the population of France (Bethlehem & De Ree, 1999), but they dealt more explicitly with sampling issues and reliability than we can infer from the examples in Section 4.3.1. Graunt knew that in parishes with reliable information about the population about three people died per eleven families per year. He also knew that there were about 13,000 funerals per year in London and he estimated the average family size as eight, which led him to 13,000 / 3 * 11 * 8 is roughly 384,000 inhabitants of London. Also in Laplace’s example, average values
were used to find a total number.

H7. More complicated estimation tasks than those of Section 4.3.1, such as those of Graunt and Laplace, might be useful to deal more explicitly with sampling issues.

Conversely, a total number can also be used to find a mean as a measure. Such an example of average in which sampling plays a role occurs in a geometry book by Köbel in 1535. Figure 4.5 shows how a rod of 16 feet should be determined by measuring the feet of sixteen men as they leave church (Stigler, 1999). This rod of 16 feet was to become a standard for other measurements in the community. In this simple measurement example we encounter several statistical issues. Were people in the past aware of the fact that the total of 16 feet is equal to 16 times the arithmetic mean of the lengths of these 16 feet? Probably not because the arithmetic mean was only defined later for more than two values. How was the sample taken? We may assume that there was no size-based criterion for selection. Although many scientists in 1535 still assumed that combining observations would amplify the errors instead of reduce them (Stigler, 1986), this example seems to be an intuitively clear way of combining measurements to reduce variation. Did the inventors of this method realize that they benefitted from compensation of errors? For the purpose of our historical phenomenology it is not necessary to know this; we simply use the example as a source of inspiration for instructional activities (6.3, 6.4, and 6.9). As the examples show, variation and sampling issues often underlie seemingly simple problems concerning averages.

H8. Examples similar to those in this section may be used to let students think about the arithmetic mean in close connection to variation, measurement, and sampling.

### 4.4.2 Decision-making

There are examples of sampling that reveal yet another origin of statistics: decision-making. We give a few examples from Jewish Law. Jewish Law deals mainly with
social, ethical, and ritual duties and is considered a rational pursuit: although rabbis accepted divine guidance, they insisted on rational methods in coming to decisions. Rabbis had to decide, among other things, how inheritances had to be distributed and whether food was kosher. If, for example, 9 out of 10 shops in a city sold kosher meat and someone found a piece of meat in that city, a rabbi could advise to consider it kosher (Rabinovitch, 1973). We interpret this as follows. If from a sample of 10 shops 9 sell kosher food, this proportion gives an indication of the chance that an arbitrary piece of meat in the city is kosher. Thus proportional or multiplicative reasoning plays an important role in the relation of sample to population.

Another decision-making example concerning multiplicative reasoning and sampling concerns the question of whether an epidemic has taken place, which is relevant to know (as it is today) for undertaking particular steps.

A town bringing forth five hundred foot-soldiers like Kfar Amiqo, and three died there in three consecutive days - it is a plague... A town bringing forth one thousand five hundred foot-soldiers like Kfar Akko, and nine died there in three consecutive days - it is a plague; in one day or in four days - it is not a plague. (Rabinovitch, 1973, p. 86)

In this example three points are interesting for the historical phenomenology. First, we see that rabbis reasoned proportionally to the total population. Second, a kind of sampling was used: the amount of foot soldiers was used as an indicator of the total population, probably because foot soldiers formed a constant percentage of the population. Third, the rabbis seemed to know typical or average death rates and they took into account how the deaths were distributed over the consecutive days.

H9. When making data-based decisions, multiplicative reasoning is an essential skill in dealing with samples versus populations (cf. P11 in Section 2.3).

4.4.3 Quality control
Apart from estimation and decision-making there are several other origins of statistics. One of them is quality control. In this section, we discuss an old secular example of sampling and quality control: the trial of the Pyx (Stigler, 1977). This trial took place at the Royal Mint of Great Britain where gold and silver coins were made. Starting from the twelfth century, every day one of the coins was put in the Pyx, which was a box in Westminster Abbey. After a few months or years, the Pyx was opened and the coins were investigated on weight and pureness. Not the single coins but the whole box was weighed and a sample of coins was melted to investigate the pureness of the coins. If the coins turned out to be good, this fact was celebrated with a banquet; otherwise the coin makers were punished. This is the first clear example of quality control we have found that is based on sampling inspection.

In Section 2.2 we wrote that we were in search for coherent knowledge of the key concepts of statistics. What we can learn from this historical phenomenology for the didactical phenomenology is that the use of averages often involves sampling issues
(4.4.1). This close link to sampling indicates one of the difficulties of learning the mean, but the link can also be used to teach sampling issues from what students already know about averages. Another thing we conjectured from this example is that sampling as carried out here, one coin per day, might be an intuitively clear way of sampling to students as well (but see Section 6.9).

H10. Problem situations similar to the trial of the Pyx may challenge students to reinvent simple sampling methods. Randomness is implicit in the trial of the Pyx.

H11. Moreover, such problems may be used to reinforce a meaningful relation between average and total, which in turn can form the basis for insight into the relation of sample and population.

4.4.4 Random sampling

From the examples in the previous sections on sampling we can infer that sampling can be used for different purposes such as finding a total number, finding a measure based on a total, and making a decision. Historically the next stage was to use sampling for getting information about a population. There are different reasons to use sampling. An important motive for the Central Bureau of Statistics in the Netherlands was to reduce the costs of its studies (Bethlehem & De Ree, 1999), and often it is also impossible to measure the whole population. Yet is sampling a relatively recent accomplishment.

Censuses were held both in ancient China and Egypt. Famous, of course, is the Roman census of Caesar August that is known from the biblical story about Jesus’s birth. Incas (1000-1500) recorded information about their people, homes, llamas, marriages, and young men that could be recruited for the army. Until late in the nineteenth century, only integral surveys were carried out because other methods were considered unreliable and discriminatory. It was considered unfair to take observations of certain human beings into account and replace those of others by calculations. This last feeling of resistance is understandable if we make a comparison with voting. Today most of us would also protest if we were not allowed to vote for a new government and a sampling method were to be used instead. Yet statisticians argue that a good sampling method is more reliable than a self-selection which results from a non-obligatory call to vote (De Mast, 2002).

A new period for statistics started in 1895 when the Norwegian Kiaer presented his ‘representative method’, which implied the deliberate selection of a representative sample, for instance as many men and women, from cities and villages, of all ages, and so on (stratified sample). It was not until 1903, however, that the International Statistical Institute accepted this method provided that the selection was carefully described. In 1906, Bowley proposed to use a process of drawing lots. The two methods coexisted until Neyman (1934) was finally able to prove that random sampling
was superior to Kiaer’s method.

One of the underlying conceptual difficulties of sampling is its close link to probability as we already hinted at in the section on decision-making. In The Probabilistic Revolution (Krüger et al., 1989) different authors underline the conceptual shift from determinism to indeterminism that was made at the end of the nineteenth century, and which proved crucial to the development of statistics and probability theory (see also Hacking, 1990; Porter, 1986; Stigler, 1986).

In contrast to people in and before the nineteenth century, students today are acquainted with surveys, which are now culturally accepted, and students might also know about random numbers from computer games. This means that students need not exactly follow the historical development of sampling, but it could still be that students think that everybody should be measured in some cases. From this historical outline we can distinguish different levels of understanding sampling.

\[ H12. \text{If the unit of thought or focus of attention is a concrete object such as a coin and if there is little variation, students may reinvent sampling methods. However, if the unit of thought or object of interest is a whole population that is influenced by multiple variables, students probably prefer stratified sampling to random sampling, because it gives the suggestion of having control of the sample.} \]

Before the seventh-grade teaching experiments we had mainly paid attention to the average values and to sampling. The historical study of distribution for example had not yielded very much. The teaching experiments in grade 7, however, urged us to reconsider the history of the median, distribution, and graphs. Because we preferred to keep the historical phenomenology of the different concepts in one chapter, the history of these concepts and graphs is discussed in the next sections. As a consequence, the hypotheses in those sections were formulated only after the seventh-grade teaching experiments and they did not play an explicit role in the hypothetical learning trajectory of these experiments.

4.5 Median

After the seventh-grade teaching experiments we had two reasons to investigate the history of the median more carefully. First, the seventh-grade students had more problems with the median as a representative value than we had expected, even after taking the results of the Nashville team into account (R16).\(^{12}\) To understand students’ problems with the median we carried out a conceptual analysis of it. Second,

---

\(^{12}\) Cobb (personal communication, February 18, 2003) and Gravemeijer assume that students’ problems with the median were due to the design (see also Cobb, McClain, & Gravemeijer, 2003), but we conjecture that there are conceptual problems with the median that are not easy to overcome.
during instructional design we could not find phenomena that beg to be organized by the median as a measure of center or a representative value. So far, our historical study had not helped us much further because we had hardly found any examples of the median before about 1840. Why did the median arise so late? Is it because it is a difficult concept? What was it used for? Can we find clues for instructional design? The historical phenomenology we undertake in this section is thus meant to find out why the median is difficult for students and to find phenomena that require using the median. Unfortunately, there turned out to be very few historical studies of the median; its history is mostly buried under bigger issues such as the normal distribution, Bayes’ theorem, the central limit theorem, and the method of least squares. Monjardet (1991) describes the median’s history with an interest in metric spaces; Godard and Crépel (1999) concentrate on the median’s statistical characteristics after 1750; and Harter’s *Chronological annotated bibliography on order statistics* (1977) provides a list of articles on order statistics, some of which deal with the question of whether the mean or median is a better measure of center (this was around 1900). Hence we had to do our own historical study of the median. We used the same method as for the previous sections.

We organize this section by focusing on the phenomena and contexts in which the median arose. It turned out that the median mostly emerged as a differentiation of the mean, and it was often the alternative measure that appeared less viable. In metaphorical terms, Ms Median appeared to be the stepsister of Ms Mean.

### 4.5.1 Theory of error

The most important context in which the average values were used was the theory of error (Sheynin, 1996). The problem at first was to find the assumed true value from the available observations, later to find the best estimate of such a value. The Greeks often used a value that fitted the theory instead of real observations to “save the phenomena,” as they called it (Pannekoek, 1961; Steinbring, 1980). Another method was to choose a value that seemed reliable, for example from a middle cluster or from values measured under favorable conditions (cf. Section 7.2). As we mentioned before (4.3.3), the midrange was used as well, for instance by Arab scientists in the ninth to eleventh century. It was not until the late sixteenth century, however, that the mean of more than two values was defined (4.3.3). Tycho Brahe seems to be the first to use the mean for reducing error and combining observations (Placket, 1970).

The first possible instance of the *median* that Eisenhart (1974) has found was in a book by Edward Wright of 1599 on navigation. Wright wrote about the determination of location with a compass (note that the letters ‘u’ and ‘v’ were used differently in those days):
Exact trueth amongst the vnconstant waues of the sea is to bee looked for, though good instruments bee neuer so well applied. Yet with heedfull diligence we come so neare the trueth as the nature of the sea, our sight and instruments will suffer vs. Neither if there be disagreement betwixt obseruations, are they all by & by to be rejected; but as when many arrows are shot at a marke, and the marke afterwards away, hee may bee thought to worke according to reason, who to find out the place where the marke stood, shall seeke out the middle place amongst all the arrowes: so amongst many different observations, the middlemost is likest to come nearest the truth.

(Eisenhart, 1974, p. 52)

It is not certain that Wright really meant the median, since he gave no numerical examples. Eisenhart argued that since Wright wrote “Neither... are they all by & by to be rejected” it is possible that he recommended the middle-most observation, the median, and not the middle place, the midrange, since then most observations would not be used. Even if this is a real example of the median, it is just a solitary example, and certainly not an indication of a common practice of using the median in navigation or any other context.

A clearer example of the median, in the context of measurement errors, is found in the work of Boscovich (around 1755). The interesting point of his work for the history of the median was the set of conditions he proposed in the search for true values, in particular a line of best fit through observations. One of these conditions was that the sum of absolute errors should be minimal; in our notation: \( \sum |x_i - x| \) is minimal. This condition turns out to be equivalent to the statistical median (David, 1998b; Eisenhart, 1977), which can be proven with differentiation. Note that the condition that the errors should add up to zero is equivalent with the arithmetic mean: 

\[
\sum (x_i - a) = 0 \iff \frac{\sum x_i}{n} = a
\]

H13. If the theory of errors (e.g. repeated measurements) is taken as a context for developing statistical ideas of center and distribution, it may be advantageous to let students formulate their own intuitions about the distribution of errors. Do the errors add up to zero? Is the chance that the measurement is too small equal to the chance that they are too large? Do errors occur symmetrically?

In fact, discussions on the mean in error theory led to the development of the concept of distribution (4.6).

4.5.2 Probability

Another context in which the median arose as a counterpart of the mean was probability theory. We give three examples of how the median is connected to probability. The first is a paradigmatic example of how an intuition of something, in this case a middle value, became differentiated into two concepts, namely median and mean life time. The second example is about birth rates and deals with quartiles and the interquartile range. The third example stems from Legendre and Laplace, who dis-
tinguished two possibilities of finding a true value, one of which we would now call the median.

1. In 1669, the Dutch brothers Christiaan and Lodewijk Huygens had an informal correspondence about their father’s life expectancy and about life expectancy in general. In 1662, just a few years earlier, they had received the famous *Bills of Mortality* by John Graunt and with the tabular data in the book they calculated their father’s and their own chances, which then evoked a flow of new mathematical problems (Véron & Rohrbasser, 2000). The brothers continued to write each other on annuities and life insurance based on these mortality tables, but they disagreed about certain calculations. It was Christiaan who realized that there was a difference between expected remaining life time and the life time that half of the people would reach. On November 28, 1669, he wrote to Lodewijk:

> There are thus two different concepts: the expectation or the value of the future age of a person, and the age at which he has an equal chance to survive or not. The first is for the calculation of life annuities, and the other for wagering. (C. Huygens, 1895, Volume 6, letter to Lodewijk Huygens; translation from French: Hald, 1990, p. 106)

Christiaan made a graph from which we can read the median life time; this graph was one of the first line graphs ever (Tufte, 2000, 2001; see also Section 4.7). In Figure 4.6 we can see that a 20-year-old person (A) had a median life time of 36 years: take the half of AB and find CD further in the graph. The French terms that Christiaan used for what we now call median life time were *apparence* (likeliness) and *vie probable* (probable life), since the person has equal chance to survive to this age or not. The chance of a half appears a natural point to look at, though it was not very useful except for chance-like problems such as wagering. More useful was what Christiaan called *espérance* and what we now call mean or expected life time. This is also what Johan de Witt and Jan Hudde used for the life annuity calculations two years later.

We conclude for the historical phenomenology that the phenomenon of predictions about life times asked for a distinction between mean and median life time due to skewed distribution (Stamhuis, 1996). The mean life time was useful for annuities, but the median was only useful for wagering.

**H14.** In this context of life times, that is in a skewed distribution, the median and mean refer to different intuitions of center. The median is connected to probability theory, in particular to halves, and the mean to expectation. When aiming at the median, it is worth trying to design problem situations in which it is reasonable to look at halves or compare halves, for instance in chance situations.
2. The second context in which we see a connection between probability and the median is birth rates. It brings us to quartiles and the interquartile range. Mathematicians such as De Moivre, Stirling, and Daniel Bernoulli studied birth phenomena with binomial distributions, not with real data. Bernoulli, for example, wanted to calculate the probability that in a binomial distribution the variable appeared between two limit values. He assumed that somewhere 2N children were born, with equal chances for boys and girls. The essential point for our median story is that he then raised the question of what the limit values were that would delimit half of the cases. A hundred more boys could be indicated by +100, 24 more girls by −24. With 2N=20,000 he found that this limit value was 47 ¼ at either side of N. In general, he wrote, it is 0.4725√N, which is close to the value 0.4769√N that is derived from the normal distribution (Hald, 1990). For the middle range between the limit values Bernoulli used the term \textit{status medius}; and for the middle limits the term \textit{limites medi} (Bernoulli, 1982, pp. 220, 385). His \textit{status medius} is the same as the interquartile range and the \textit{limites medi} are the same as the first and third quartiles. With a modern view we might interpret the middle position as the median here—
equal numbers left and right—and not the mean.
What is striking in this context of dealing with birth rates is that the median and the quartiles are more apparent than the mean and modulus (this is a precursor to the standard deviation\textsuperscript{13}; see Walker, 1931). The median is seen as the exact balance of boys and girls. Bernoulli (and later Galton) would probably have been surprised if they had heard that nowadays students learn the 68.26\% rule for standard deviations in normal distributions, and do not work with the interquartile range of 50\%.

3. The third example of the connection between the median and probability is the following. Legendre, Laplace (1812/1891), and their contemporaries used the term \textit{milieu de probabilité}, the middle of the probability, which is a suggestive name for the median in the context of probability functions.

Cournot (1843) was the first to use the term ‘median’ (\textit{valeur médiane}) for this value (Bru, 1984; David, 1995, 1998b; Stigler, 1986). He defined the median as the value \(x_0\) for which the distribution function \(F\) was \(F(x_0)=\frac{1}{2}\) and he explained that it is the value for which the area under the graph is the same on the left and on the right (Figure 4.7). Furthermore, he wrote:

Two players, one betting of the value smaller than \(x\) and the other larger than \(x\), would bet with the same chances. With a very big number, the quotient of the larger (or smaller) values than \(x\) and the total number of values will not differ much from the fraction \(\frac{1}{2}\). (Cournot, 1843, p. 83; translation from French)

\textbf{Figure 4.7:} Graph from Cournot (1843, Figure 17) with the median in a skewed distribution.

\textsuperscript{13} The name ‘standard deviation’ was introduced by Karl Pearson at the end of the nineteenth century (David, 1995; Walker, 1931).
One way to visually estimate the median in a dot plot such as in Minitool 2 is to look for which value the areas on the left and right are the same (see Section 10.5).

The central point of this section is that the median and quartiles are closely related to probability theory, especially with the chance of a half.

4.5.3 Ease of calculation and ordinal data

In 1874 Gustav Theodor Fechner (1801-1887) used the median, the Centralwerth, in an attempt to describe many sociological and psychological phenomena with methods that had proven to be useful in astronomy. He advocated the ease of calculation of the median, but he also had more theoretical reasons for using other measures of center than the mean, which we address in Section 4.5.5.

Francis Galton used the English term ‘median’ for the first time in 1882 (David, 1995) and caused the breakthrough of the concept (Godard & Crépel, 1999). As happens often in the history of mathematics and statistics (Bissell, 1996; Dijksterhuis, 1950), Galton knew the concept before he used this particular term. Before 1882 he used other terms including the middle-most value (1869) and the medium (1880), and in a lecture in 1874 he gave the following description:

The object then found to occupy the middle position of the series must possess the quality in such a degree that the number of objects in the series that have more of it is equal to that of those that have less of it. (Walker, 1931, p. 87)

Figure 4.8: Galton’s graph of the normal distribution with quartiles p and q, and median m (Galton, 1875, p. 36)

In a graph from 1875 he indicated the median and quartiles with the letters p, m, and q and the fractions $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, but not with names (Figure 4.8). We come back to this graph in Section 4.6 on distribution.
H17. A five-number summary of extreme values, quartiles, and median may be a suitable way for students to characterize distributions, once students know the median and quartiles as measures of center and spread.

Important reasons for Galton to use the median were its ease of calculation and its intuitive clarity (Stigler, 1973). Most phenomena Galton studied were roughly symmetrical, so the median would not differ much from the mean, which is laborious to calculate. Throughout his book *Natural Inheritance* (1889) he therefore used the median M and quartile distance Q, with $Q = 1/2 (Q_3 – Q_1)$, $Q_1$ being the first and $Q_3$ the third quartile (in the modern terminology). He rarely mentioned the mean and modulus ($\sqrt{2}$ times the standard deviation), probably to reduce calculation efforts and not scare away scientists without the necessary statistical background.

Apart from ease of calculation and intuitive insight, yet another reason for using the median could have played a role. Galton studied variables that he measured in an ordinal way. And indeed, with ordinal data the mean cannot be calculated, in contrast to the median.

What struck us in the historical study was that ordinal data are so rare, apart from the paradox of Borda and other theoretical voting problems that we interpret as non-statistical (Condorcet, 1785; Crépel & Godard, 1999; Goddijn, 1988; King, 1963; Monjardet, 1991). Galton seems to be one of the first to study real ordinal data, for instance in the context of intelligence and reputation. From the historical overview we conjecture the following.

H18. Although ordinality is a statistical reason to use the median as a measure of center, contexts with ordinal data are not very suitable to help students understand the median.

4.5.4 Robustness

Over the last centuries scientists have been concerned with the sensitivity of the mean to outliers, and have proposed different procedures that were more ‘robust’ as Box called it in 1953: trimmed means, weighted means, averaging different average values, but also the median (Stigler, 1973, 1980). Francis Ysidro Edgeworth (1845-1926), a younger contemporary of Galton, preferred the median to the mean because of its insensitivity to outliers, probably due to his interest in economics, which has less regular data than is common in astronomy for instance, and he was not the only one to prefer the median (Harter, 1977). Nowadays, the median’s resistance to outliers is one of the major reasons to use it, especially when the data are irregular as is common in social sciences and economics.
4.5.5 Skewed distributions
For a long time, distributions of error were assumed to be symmetrical. In 1838, Bessel was probably the first to doubt the assumption of symmetry (ESS, 1998; Steinbring, 1980). In contrast to most other scientists of his time, Fechner (1874) even assumed that most distributions of data were asymmetric. It turned out that the median minimizes the sum of absolute deviations to the first power and the mean the sum of deviations to the second power. Consequently, both measures of center are special cases of Fechner’s so-called Potenz-mittelwerthen, the values that minimize the sum of the deviations to the n-th power: \( \sum |x_i - x|^n \) is minimal. Fechner used these generalized measures of center to describe the skewness of distributions. Edgeworth also used the difference of mean and median, divided by a normalizing factor, and this measure is still used today as an indication of skewness (ESS, 1998; Stigler, 1986).

As Tukey (1977) pointed out decades later, the median is also useful in the five-number summary of unimodal distributions, consisting of the minimum value, first quartile, median, third quartile, and maximum value. In fact, this five-number summary is the basis of the box plot. The median is especially useful as a measure of center in asymmetric distributions, because it is far less influenced by extreme values than the mean (the median is more ‘robust’).

For the Nashville team, a reason to use the median was that it tends to be closer to the majority of a unimodal data set than the mean (R16). Another reason was that the median plus quartiles seemed easier than the mean plus standard deviation (P8). And third, the median and quartiles seem more appropriate when describing skewed distributions than mean and standard deviation (4.5.2).

4.5.6 Summary of the median’s history
The mean was used for reducing error from the late sixteenth century onwards, but the median was developed relatively late. We summarize a few examples. Expected life time was a useful element in calculating life annuities, whereas the median life time was considered to be just “for wagering,” as Christiaan Huygens wrote. For Daniel Bernoulli the quartiles (limites medii) were evident values to look at, but the standard deviations won in the nineteenth century.14 In the theory of error the mean

---

14. There are of course exceptions: in 1910, the Dutch botanist Tine Tammes preferred median and quartiles (De Knecht-van Eeckelen & Stamhuis, 1992).
became a popular measure for combining observations and reducing error, but the median was not used until it was acknowledged that distributions can be skewed. The contexts in which such distributions became apparent were for example the social sciences and economics (Crépel & Godard, 1999). Phenomena in these fields are far less regular than in astronomy and physics, so a measure of location that is less sensitive to outliers, such as the median, is useful. Despite the efforts of scientists like Cournot, Fechner, Galton, and Edgeworth, the median was neglected and the mean favored.

Nowadays, the median is used in order statistics, since the mean cannot be used for ordinal data. The median is also used in robust statistics, since it is far more robust than the mean. Since statistics is applied in more and more areas with irregular data, the median has become more popular (Portnoy & Koenker, 1997). The question remains, however, whether Ms Median as the stepsister of Ms Mean will ever turn out to be Cinderella.

### 4.6 Distribution

For reasons of readability, we organized the historical phenomenology of average values, sampling, and distribution into different parts, but as we argued in Section 2.2 and as the examples in previous sections show, all these concepts are intimately interwoven. Estimation using average values has to do with sampling and the median examples often involve distribution issues.

In the eighteenth century, the concept of distribution arose from the theory of errors, when the arithmetic mean as a method to reduce errors was still a topic of debate. Due to the impossibility of determining individual errors, one had to look at the relation between the errors.

> Measurements, and functions of measurements, such as their arithmetic mean, are not amenable to mathematical theory, (...) as long as individual measurements are regarded as unique entities, that is, as fixed numbers $y_1, y_2, \ldots$. A mathematical theory of measurements, and of functions of measurements, is possible only when particular measurements $y_1, y_2, \ldots$ are regarded as instances of hypothetical measurements $Y_1, Y_2, \ldots$ that might have been, or might be, yielded by the same measurement process under the same circumstances. (ESS, 1998, p. 531)

In 1756 Simpson made this shift to looking at the relation between errors when he used simple probability functions to argue that the mean of several observations was better than a single observation. The first distribution of errors he proposed was a discrete uniform distribution, that is with equal probabilities for all values $-v, -v+1, \ldots, 0, 1, \ldots, v$. Next, he assumed a discrete isosceles triangle distribution with probabilities proportional to $1, 2, \ldots, v, v+1, v, v+1, v, \ldots, 2, 1$, from which he obtained a continuous isosceles triangle distribution one year later (see Figure 4.9).
Quickly after Simpson had launched his idea of probability distributions, other scientists proposed alternative laws of error. Among them were Lagrange, Lambert, Daniel Bernoulli, Laplace, and Gauss. Note that the analytic expressions in Figure 4.10 are a modern accomplishment. Lambert, for instance, introduced his method of maximum likelihood without ever expressing his error-frequency distribution in a functional form (ESS, 1998).

Lambert (1765): flattened semicircle

\[ f(x) = \frac{1}{2} \sqrt{1 - x^2}, \quad -1 < x < 1 \]

Lagrange (1776): continuous parabolic

\[ f(x) = \frac{3}{4} (1 - x^2), \quad -1 \leq x \leq 1 \]
Lagrange (1776): continuous uniform

Lagrange (1781): cosine function

\[ f(x) = \frac{\pi}{4} \cos\left(\frac{\pi x}{2}\right) \quad -1 \leq x \leq 1 \]

Laplace (1781): log function

\[ f(x) = \frac{1}{2} \log \frac{1}{|x|} \quad -1 \leq x \leq 1 \]

Laplace (1774): double exponential

\[ f(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \]
H21. Students can reason about the shape of distributions without being bothered by analytic expressions (R14). They may implicitly assume distributions to be symmetrical.

Mathematically, there are many ways in which the normal distribution can arise. Historically the first way, and still a very common one, is to obtain the normal distribution as the limit of the binomial distribution $\text{bin}(n, p)$ with $n \to \infty$. This is a result of the De Moivre-Laplace limit theorem, which is a special case of the central limit theorem transpiring that in many cases the sum of a large number of independent random variables is approximately normally distributed.

In the context of people’s height, we can think of many factors that influence this including their parents’ height, their diet in their youth, their age, and their sports history. Even if these factors themselves are not normally distributed, their sum roughly is. This explains why so many phenomena can be described by the normal distribution (Sittig & Freudenthal, 1951; Wilensky, 1997).

This distribution and its curve are known under many names (Stigler, 1999) including ‘the law of error’, the ‘frequency law’, the ‘Gaussian curve’, ‘Laplace-Gauss’ (mainly in the French literature). One name Stigler does not mention is the ‘De Moivre distribution’, which Freudenthal (1966b) used because De Moivre was the first to define this function. An immediate result of Gauss’s work with the normal distribution was that astronomers were able to find the planetoid Ceres in the sky again (Steinbring, 1980). Quetelet then used methods that had proven successful in astronomy for anthropometrical purposes, and modeled phenomena such as the chest sizes of Scottish soldiers with a binomial distribution with the curve of the normal distribution superimposed as the so-called ‘curve of possibility’. In his tracks, Galton used the normal distribution (‘normal scheme’) for his studies in human faculties and inheritance. A famous quote is:

Gauss (1809) and Laplace (1810): law of error or normal distribution

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$
It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect from Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once. An Average is but a solitary fact, whereas if a single other fact be added to it, an entire Normal Scheme, which nearly corresponds to the observed ones, starts potentially into existence. (Galton, 1889b, p. 62)

Galton realized that he needed just two numbers for describing the whole distribution: the median and the quartile distance.

If we know the value of M [median or mean] as well as that of Q we know the entire Scheme [normal distribution]. M expresses the mean value of all the objects contained in the group, and Q defines their variability. (Galton, 1889b, p. 61)

Being interested in tribes—he traveled in Africa in 1851 (Galton, 1889a)—he recommended anthropologists to ask chiefs to arrange their people in order of height. Determining the first quartile, median, and third quartile would be sufficient statistics to describe the whole height distribution of the tribe (Walker, 1931, p. 84). Compared to the earlier large tables with observations, this use of a distribution was indeed a major leap forwards (Figure 4.11). In fact, this is why students need to develop such notions in relation to distributions:

Over-minuteness is mischievous, because it overwhelms the mind with more details than can be compressed into a single view. (1889b, p. 36)

Galton was really impressed by the normal distribution (‘law of frequency of error’), which inspired him to write poetic phrases:
I know of scarcely anything so apt to impress the imagination as the wonderful form
of cosmic order expressed by the “Law of Frequency of Error.” The law would have
been personified by the Greeks and deified, if they had known of it. It reigns with se-
renity and in complete self-effacement amidst the wildest confusion. The huger the
mob, and the greater the apparent anarchy, the more perfect is its sway. It is the su-
preme law of Unreason. Whenever a large sample of chaotic elements are taken in
hand and marshalled in the order of their magnitude, an unsuspected and most beau-
tiful form of regularity proves to have been latent all along. (Galton, 1889b, p. 66)

Several scientists compared Gauss’s law of error with observed frequencies in mass
phenomena and considered the agreement as good. Among these scientists was C.S.
Peirce (NEM III), who collected many data values on a young man’s reaction time
on consecutive days. About the curves of Figure 4.12 he wrote:

![Figure 4.12](image)

The curve has, however, not been plotted directly from the observations, but after they
have been smoothed off by the addition of adjacent numbers in the table eight times
over, so as to diminish the irregularities of the curve. The smoother curve on the fig-
ures is a mean curve for every day drawn by eye so as to eliminate the irregularities
entirely. It was found that after the first two or three days the curve differed very little
from that derived from the theory of least squares [the normal distribution].
(NEM III, p. 659)

The belief in the general applicability of the normal law was ubiquitous and slow to
die. Lippmann, a French physicist, said to Poincaré (1854-1912) about this:

All the world believes it firmly, because mathematicians imagine that it is a fact of
observation, and the observers that it is a theorem of mathematics. (Poincaré, 1892;
cited from ESS, 1998)
In Section 4.5.5 we already saw that some scientists such as Bessel, Fechner, and Edgeworth dropped the assumption of symmetry (Steinbring, 1980). Pearson, in the late nineteenth century developed a class of distributions which were transformations of the normal distribution, and which could be applied to many more phenomena.

This section shows that the concept of distribution developed over several centuries. What is important for the historical and didactical phenomenology is that concepts such as distribution change over time. This means that we have to consider these concepts in a dynamic perspective, as Steinbring (1980) writes about the concept of chance:

> It is impossible to give a definition of chance that stays the same in all grades. This implies that the concept of chance should be carefully related to suitable contexts and extended by extending the contexts. This relation between foundation and application leads to a dynamic perspective of development with respect to the concept of chance.

(p. 446; translation from German)

If we replace 'chance' by 'distribution', the quotation applies to our situation (in fact this holds for other statistical notions such as mean as well).

**H22. A notion of distribution cannot stay the same in all grades. Accordingly, the representations in which students study distributions need not stay the same.**

We allow students’ informal and possibly sloppy characterizations of distribution in our teaching experiments and allow them to work with their own informal sketches before using more advanced representations and definitions.

### 4.7 Graphs

Graphs are crucial tools in statistical investigations, because we can see patterns and trends of frequency distributions that are hard to see from a table of numbers. In this section we ask ourselves from which phenomena statistical graphs were developed. Beniger and Robyn (1978) distinguish four problem areas in which the most important graph types were developed:

1. spatial organization (17th and 18th century), for instance Halley’s map with lines of magnetic declination (1701);
2. discrete comparison (18th and early 19th century), for example Playfair’s bar chart of import and export in Scotland (published in 1786);
3. continuous distribution (19th century) with histogram and ogive-shaped line graphs;
4. multivariate distribution and correlation (late 19th and early 20th century) with three-dimensional charts and correlation diagrams.
1. Spatial organization

Descartes (1596-1650) was convinced that imagination and visualization, and in particular the use of diagrams, had a crucial part to play in scientific investigation. One of his contributions, the coordinate system, still proves powerful today. A first major success of using coordinates in a Cartesian system was Halley’s scatterplot of barometer readings against elevation above sea level (1701). This plot was an exception, though, because scientists had an obsession for tabular data (Beniger & Robyn, 1978). Between 1660 and 1800, even automatic graphs created by mechanical recorders to measure temperature, barometric readings, and tidal movements were routinely translated into tabular logs. Apparently, tables were considered clearer than graphs. It was not until the 1830s that scientific journals began to record graphs. We already know that students tend to focus on individual data values (2.2).

H23. Students initially tend to focus on tables and values. Even if they easily answer questions with the help of case-value plots, they still interpret these graphs as codifications of tables.

Figure 4.13: Playfair’s bar chart of Scotland’s import and export in 1781
(from Neeleman & Verhage, 1999, p. 21)
2. Discrete comparison
In 1765, Priestley published time-line charts with individual bars to compare the life-spans of about 2,000 celebrated persons who had lived between 1200 BC and 1750 AD. Not long after that, Playfair invented the first bar chart, which was published in 1786 and represented Scotland’s imports and exports for seventeen countries in 1781 (Figure 4.13). Ironically, he made this graph due to a lack of data. Because he had no time series data, he graphed a single year as a series of 34 bars, and apologized to the reader for that. This graph can be considered a way to make a discrete quantitative comparison of import and export.
Playfair’s bar chart was among the first graphs used and it resembles the representation of Minitool 1. Hence we conjecture:

H24. The bar chart in Minitool 1 is a representation that students easily come to understand.

This hypothesis was confirmed in the exploratory interviews and in the teaching experiments.

3. Continuous distribution
In the section on distribution we already demonstrated how several famous mathematicians had looked for functions that matched the distribution of error. Note that the graphs of Figures 4.9 and 4.10 were added for the modern reader as it was not until about 1820 that such graphs became more common. The problem of representing continuous distributions arose in vital statistics, the statistics of life information, and led to two important solutions: ogive-shaped line graphs and the histogram.
Fourier made a bar graph that represented the population of Paris by age groups and made a line graph of this, which led to the first appearance of a cumulative frequency distribution (1821). The first histogram was made by Guerry in 1833, who reorganized a bar graph to represent crime data that he had arranged in intervals of age and month. This led to a histogram. The term ‘histogram’ stems from Pearson (1895) (David, 1995; Schwartzman, 1994; Walker, 1931). Quetelet, then, was largely responsible for the further development of such graphs. In 1846, for instance, he published a symmetrical histogram with the curve of the normal distribution superimposed as the so-called ‘curve of possibility’. The histogram emerged from organizing a bar graph. This could also indicate, from a historical perspective, that a histogram is a more advanced graph than a bar graph (see Section 2.2 and 2.3).

H25. The histogram is more difficult to learn than the bar graph.

Galton pictured the normal distribution differently from what we are used to nowa-
days (Figure 4.14 left). He wrote:

I shall best explain my graphical method of expressing Distribution, which I like the more, the more I use it, and which I have latterly much developed, by showing how to determine the Grade of an individual among his fellows in respect to any particular faculty. (Galton, 1889b, p. 37) ['Grade' is a percentile; e.g., the first quartile is the 25th grade.]

This type of ogive-shaped graph was no exception in the beginning of the twentieth century, judging from Walker’s remark in 1931 that such graphs were commonly used in school textbooks for statistics. Galton wrote that though the ‘Curve of Frequency’ (right) was generally used by statisticians, but then “turned at right angles,” it was “far less convenient than that of Distribution [left]” (p. 49). He turned the frequency curve to “show more clearly its relation to the Curve of Distribution.” But, he admits, “the Curve of Frequency has other uses, of which advantage will be taken later on” (p. 49).

Galton called the type of graph on the left an ‘ogive’, after the architectural term (Bissell, 1996; Dictionary of Art, 1996). We have indeed found this ogive shape in many buildings in different countries (e.g. Figure 4.15).

H26. It is useful to let students deal with two representations of distributions similar to those of Galton, not just one (Figure 4.14).
The reader may already have noticed the resemblance of Galton’s graphs with Mini-tools 1 and 2, be they turned at an angle. Apparently, Galton found the cumulative curve (left in Figure 4.14) more useful for many of his purposes than the frequency curve (right), and he pointed at the transition between the two graphs that we also want students to see. It could indicate, but this is a tentative remark, that though the ‘Curves of Frequency’ are more common among statisticians, the ‘Curve of Distribution’ is easier for students to understand in some situations.

H27. The median is easier to conceive and develop in a representation that is similar to Galton’s Curve of distribution than in a Curve of frequency; hence it may be easier in Minitool 1-type representations than in Minitool 2-type representations. It could well be important that the bars in Galton’s distribution curve are vertical, because we conjecture that it is easier for students to read from left to right, and so to speak take the midrange, which happens to be the median, than reading from top to bottom, such as in Minitool 1.

We have not been able to test the hypotheses on the median in this study.

4. Multivariate distribution and correlation
By 1850, quantitative graphs had become accepted tools in statistics. The only graph of the problem area of multivariate data that we discuss is Galton’s first correlation diagram, because we see an analogy with Minitool 3 as a sequel to Minitool 2. Galton’s diagram, Figure 4.16, shows a bivariate distribution of head size and height. At the sides we see the univariate distributions of those variables. We interpret this example as supporting the idea of the Nashville team that a notion of univariate distribution is a prerequisite for really understanding bivariate distributions.
We jump to the late 1960s for other major contributions to statistical graphing, namely in exploratory data analysis. The most famous newcomers are the stem-and-leaf plot and the box-and-whiskers plot, which were first presented by Tukey in 1969 as ways to display and explore data sets by hand. Because the basic box plot was often misinterpreted, Tukey and colleagues also invented alternative box plots with additional information (McGill, Tukey, & Larsen, 1978). From this we conclude that box plots are not that easy: even in the statistical world they sometimes led to confusion. Hence, they should also be handled with care in education.
Chapter 4

H28. The box plot is one of the most advanced graph types used in middle schools. This is due to the incorporation of conceptual measures of center (median) and spread (quartiles).

We therefore propose to postpone the introduction of box plots until after middle school.

In addition to the four areas mentioned earlier, we cite Florence Nightingale (1820-1910) as a pioneer in graphical representations. Her main goal with those graphs was to convey to others the need to improve health care, for instance during the Crimean War (Cohen, 1984).

Table 4.1: Overview of the history of graphs, mainly from Beniger and Robyn (1978)

<table>
<thead>
<tr>
<th>Year</th>
<th>Graph Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ca. 1350</td>
<td>Proto-bar graph of a theoretical function (Nicole Oresme)</td>
</tr>
<tr>
<td>17th cent.</td>
<td>Tables of empirical data (Die Tabellen-Statistik in Germany)</td>
</tr>
<tr>
<td>ca. 1660</td>
<td>Automatic recording device producing a graph of temperature (Christopher Wren)</td>
</tr>
<tr>
<td>1686</td>
<td>Edmund Halley’s bivariate plot of barometric readings against altitude</td>
</tr>
<tr>
<td>1765</td>
<td>Measurement error as deviations from regular graphed line (Lambert)</td>
</tr>
<tr>
<td>1786</td>
<td>Playfair’s bar chart</td>
</tr>
<tr>
<td>1801</td>
<td>Playfair’s pie chart or circle graph</td>
</tr>
<tr>
<td>1821</td>
<td>Fourier’s cumulative frequency curve of inhabitants of Paris by age groupings</td>
</tr>
<tr>
<td>1828</td>
<td>Mortality curve (Quetelet)</td>
</tr>
<tr>
<td>1830-35</td>
<td>Graphical analysis of natural phenomena appears in journals</td>
</tr>
<tr>
<td>1833</td>
<td>Guerry’s first histogram of crime by age and months</td>
</tr>
<tr>
<td>1846</td>
<td>Quetelet represented urn schemata as symmetrical histograms with a “curve of possibility,” later called normal curve</td>
</tr>
<tr>
<td>ca. 1855</td>
<td>Bar graphs and polar-area graphs on mortality by Florence Nightingale</td>
</tr>
<tr>
<td>1868</td>
<td>Statistical diagrams in a school textbook (Levasseur)</td>
</tr>
<tr>
<td>1874</td>
<td>Age pyramid, bilateral histogram (F. Walker)</td>
</tr>
<tr>
<td>1875</td>
<td>Galton’s ogive graph of normal distribution</td>
</tr>
<tr>
<td>1884</td>
<td>Dot plot (see Wilkinson, 1999)</td>
</tr>
<tr>
<td>1969</td>
<td>Box plot and stem-and-leaf plot for EDA (Tukey)</td>
</tr>
</tbody>
</table>

We end this section with a chronological overview of types of graphs used in our research and graphs related to those (Table 4.1). We do not suggest that an instruction-
A historical phenomenology should follow this order. Taking Freudenthal’s cue we try to understand the past well enough to design a “revised and improved version of the historical learning process” (4.1). What we see in history is that a lot of mathematics and statistics was already known before most graphical representations were invented. In education this is different: students know much less mathematics than scientists from around 1800 but they have encountered more graphical representations.

4.8 Summary
A historical phenomenology is an analysis of the development of concepts (‘thought objects’) in relation to the phenomena that gave rise to these concepts (2.1). The essential point of didactical phenomenology is to translate such phenomena into problem situations that are meaningful for students and create the need for organization by a particular concept. Knowing the historical development of certain concepts can help to anticipate a process of guided reinvention. It can be demanding for instructional designers and teachers to put aside their knowledge of these concepts and take a student perspective. What may seem a minor step might have taken centuries to develop historically and might also be difficult to develop for students. A historical study can help to distinguish various aspects, problems, related notions and intermediate stages of the development of certain notions. In other words, it can help us look through the eyes of the students. This section is a summary of the results that turned out most useful for the didactical phenomenology and for the teaching experiments.

Estimation (H1, 7, 8)
Estimation of large numbers could well be one of the ancient origins of statistical methods. From a modern point of view we can recognize precursors to notions of average and sampling in those methods. These thought objects of average and sample are used to handle the phenomenon of variability in what is estimated. This implies that we could use estimation tasks as the starting point of a statistics unit that supports a process of guided reinvention of average and sample.

Bars representation (H2)
Magnitudes can be represented by tallies and numbers, but also by the lengths of bars (Euclid, 1956). We assume that students can easily interpret bars as representations of data values such as in a value-bar graph, especially if the variable at issue has a time dimension (life span) or a one-dimensional physical connotation (wing span, height, braking distance). Moreover, we assume that the bar representation of data values can help students in estimating means from data sets by using a compensation strategy. Better than from a table of values, they can see where the center of the data values is from a bar rep-
representation such as a value-bar graph.

Midrange (H4)
Before the arithmetic mean was used to reduce measurement errors or to summarize data (16th century), the midrange was used for such purposes (9-11th century). From a modern perspective the midrange is not a very useful measure of center, because it is too sensitive to outliers, but with symmetrical distributions (such as most error distributions) this problem is less apparent. It is likely that students will also use the midrange as an initial way to find an average, for example when estimating total numbers. With skewed distributions students can then be challenged to scrutinize their midrange strategy. Similarly, skewed distributions can be used to create a need for a distinction between mean and median (H20).

Mean as an entity in itself (H6)
Considering the mean of a variable as a representation of a specific aspect of a population (percentage of dead letters, inclination to suicide) is much more recent than the other types of means we discussed. In the nineteenth century, the mean was used more and more as an entity in itself. This was especially apparent in cases where the mean did not refer to actual situations. In 1877, for example, Peirce (CP 2.646) gave the example that in New York 14.72 persons lived in the average house. It is likely that students find this type of mean more difficult to understand than older types of means.

Sampling (H8, 10, 11, 12)
Just as there are different levels of using the mean, there are different levels of using sampling. In the estimation examples sampling was mostly implicit: the focus is on the total number and the sample helps to reduce variability in a smart way. In the Trial of the Pyx example of quality control of coins, the focus was on the weight and pureness of coins and sampling was necessary because not every coin could be melted to test its pureness. Both the total and the individual coins were clear units of thought. Later in history, however, scientists became interested in more abstract aspects of populations, as the section on the mean as an entity in itself shows. More advanced methods of sampling became necessary, because measuring populations became too expensive or even impossible. In this case, a sample is a thought object with which something can be said about the population.

Measures of spread (H15)
One way to organize the variability of a data set is by summarizing it. This can be done with a measure of the spread. Historically the oldest measure of spread is the
range (David, 1998a). However, the range is sensitive to outliers; more robust measures are the interquartile range and the standard deviation.

Distribution (H21, 22)
A more sophisticated way to summarize or model a data set is by using the thought object of ‘distribution’. In history, different distribution shapes have been proposed as summarizing the pattern in the variability of errors. Before the nineteenth century, distributions were assumed to be symmetrical. When statistics became used in more and more contexts including economics and social sciences, there was a need to make distinctions between different types of distributions (symmetrical or skewed, frequency or density distribution, sampling or population distribution).

Graphs (H23-25)
Before about 1800, scientists rarely used graphs because they preferred tabular data. Graphs are another way of summarizing data sets or patterns in variability. They can be used to represent distributions. What is interesting with respect to the normal distribution is that different representations were used; it was not only represented with the famous bell curve, but also as an ogive. This ogive shape also appears if vertical value bars are ordered by size, for example when students line up by height (Figure 7.19). It may be useful to use different representations of distributions to highlight different aspects of these distributions. By and large, the historical development of graphs is in line with the rationale of the Minitool representations.

On the basis of the present historical phenomenology, as well as prior research and exploratory interviews, a didactical phenomenology is formulated in the next chapter.