*"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." (Wilks, 1951)*

Data analysis based on only some of all in principle possible data is part and parcel of empirical work. Researchers may or indeed do hope that what they found in a partial study may give a good picture of the totality that they would ideally have researched. The paradigm of inferential statistics, as developed by, among others, Peirce, Galton, K. Pearson, Fisher, Neyman, E. Pearson, has set standards on how to approach the question on whether a result found on part of the data after all may or may not misrepresent what would have been found, had we analyzed all data.

The magic wand from this toolbox is:

* considering the totality of possible units of analysis as a population P, and research only a sample S from units of P that have been selected by a *known probability mechanism*.[[1]](#footnote-1)

There are of course myriads of possible samples S1, S2, … Smyriad that could have been produced by such a probability mechanism, and we actually got one, viz. S, out of them, just by chance. We could, under this approach, have got quite as likely a different sample S’, with different results. How different could the results from alternative samples be? Exactly through the assumption that the actually observed S is a randomly chosen instance from the whole sample space Σ= {S1, S2, …} it is made possible to study by methods from mathematical statistics the likelihood that what we find on S is a good or bad representation of P.

The dominant method in statistical inference is hypothesis testing. For the sake of conciseness we pass over the other line of inferential reasoning, the method of interval estimation (confidence intervals), which would not alter our argument.

Given a population P of cases about which we would like to gather knowledge on the basis of a sample S, generated by a specified probability mechanism, i.e. given (\*), hypothesis testing follows the line of

1. putting forward a hypothesis H0 on the population P (EXAMPLE: the average response time on calls for service in police district D is not more than 8 minutes), and an alternative hypothesis Halt , oftenthe negation of H0.
2. studying a *sample* S (EXAMPLE: a random sample of all calls for service) presumed to have been generated by the stated probability mechanism.
3. defining a *measure of deviance* between results in S and in P when H0 is true, called a *test statistic*[[2]](#footnote-2), (EXAMPLE: the observed average response time in S is 2 minutes higher than the average time in P under H0).
4. computing the probability (under the stated probability mechanism that has produced the sample) that we would get a result *at least as deviant* from H0 as has been observed in S, if indeed H0 holds true on P; that probability is called the *tail probability* of the observed test statistic, often denoted by the label “*p-value*” [[3]](#footnote-3)
5. choosing a threshold for when a tail probability (p-value) should be called low; such a threshold, often denoted by the label α, is called a significance level [[4]](#footnote-4) (EXAMPLE: chose a significance level α = 0.05)
6. interpreting p-values below the chosen significance level α as a sign that the observed S is incompatible with the conjunction of statements that (i) S is a probability sample from P ; and (ii) H0 is true (EXAMPLE: if we observe p=0.02, we conclude that, as it is smaller than 0.05, we reject H0 and state that the data of S support that Halt is true). Indeed, p-values below the chosen threshold are called “significant” or, somewhat more explicit, “significant at the level α”. Another way of formulating this is saying that in this case results on S that deviate from H0 may be generalized to P.
7. while interpreting p-values larger than the significance level α as a sign that data are compatible with H0. Such p-values are called “not significant”. Though results on S may deviate from H0, this may not be generalized to P.

In many social sciences it has become common to publish results mentioning p-values, or stating that result are significant or not significant, although it is not always clear what is the population P, or what probability mechanism is supposed to have generated S from P.

Now it is implied by the above exposé that it is completely unclear what a p-value denotes, if we have not specified the assumption formulated in the statement \* above: we must be prepared to consider the data at hand, S, as a sample from a wider population P, and as being generated by a known probability mechanism from P. A p-value simply has no meaning outside this framework, and without that framework, talking about or using p-values should be denounced as pseudo statistical inference.

Henk Elffers, written 27-10-2022

1. Often this is the assumption of random sampling, i.e. that each unit in P has an equal chance of becoming a member of S, but many other, more complex sampling schemes can be treated likewise, such as stratified sampling, multistage sampling, sampling with unequal probability of being sampled, … While the mathematics of the analysis may become more complex in those latter cases, the structure of the argument remains the same. [↑](#footnote-ref-1)
2. This test statistic is chosen with an eye on the alternative hypothesis Halt [↑](#footnote-ref-2)
3. Handbooks on mathematical statistics give recipes for those computations for many situations [↑](#footnote-ref-3)
4. Conventional choices for significance levels are α = 0.05, α = 0.01 or α=0.001. [↑](#footnote-ref-4)