DOI: 10.24425/academiaPAS.2023.147632

Trees growing at the edge of a forest usually have a different shape and height than those deeper in the forest interior. It is important to keep this in mind when choosing a sample of trees to measure or study

GENOMART/SHUTTERSTOCK.COM

Reflections in a Distorted Mirror

Mathematics offers tools renowned for their objectivity, which is a cornerstone of scientific inquiry. Yet the question arises: how accurately do statistical methods really reflect the complexities of the real world?

Dominik Tomaszewski PhD, DSc

works as an Assistant Professor at the PAS Institute of Dendrology in Kórnik. He studies the micromorphology and systematics of plants.

dominito@man.poznan.pl

Dominik Tomaszewski

PAS Institute of Dendrology in Kórnik

The field of statistics is part of the broader research methodology that uses mathematical methods to describe and understand the world. The overall picture of the world we obtain needs to be free from speculation and invention, and as objective as possible – in other words, we want it to be true and well-aligned with reality. Statistical methods can greatly assist in this. Nevertheless, at the same time, overreliance on such methods and insufficient awareness of their potential pitfalls can instead easily distort the real world.

Following the scientific method, we formulate hypotheses that can then be confirmed or refuted.

As evidence in favor of one conclusion or the other, methods and criteria based on statistical analysis are often applied. This requires an appropriate dataset. However, careful data collection is just the beginning. It must be followed by careful verification, processing, and proper analysis. Biologists often use statistical techniques in conjunction with software that performs calculations, computes correlation values and the significance level of differences, etc. However, it is up to the individual researchers to make decisions about which type of analysis to apply and how to verify the initial assumptions – such as sample size, normal distribution, or homogeneity of variance. Unfortunately, these important steps are often skipped, which is a serious mistake that undermines the entire reasoning process.

Precision

When we study a particular phenomenon, how authentic our picture of it is will depend in part on the quality of our data, in other words, on its *precision*. For example, we can measure both the height of a tree and the annual growth in its thickness, but each of these measurements will require a different level of precision. Determining a tree's height down to a fraction of a millimeter, while technically challenging, will add little of substantive value. Similarly, measuring human lifespans down to the second, or the geographical coordinates of large objects with centimeter precision, is quite simply unnecessary.

Even leaves growing on the same shoot can vary greatly in size and shape – this is something that must be borne in mind if we want to make a correct analysis of such characteristics

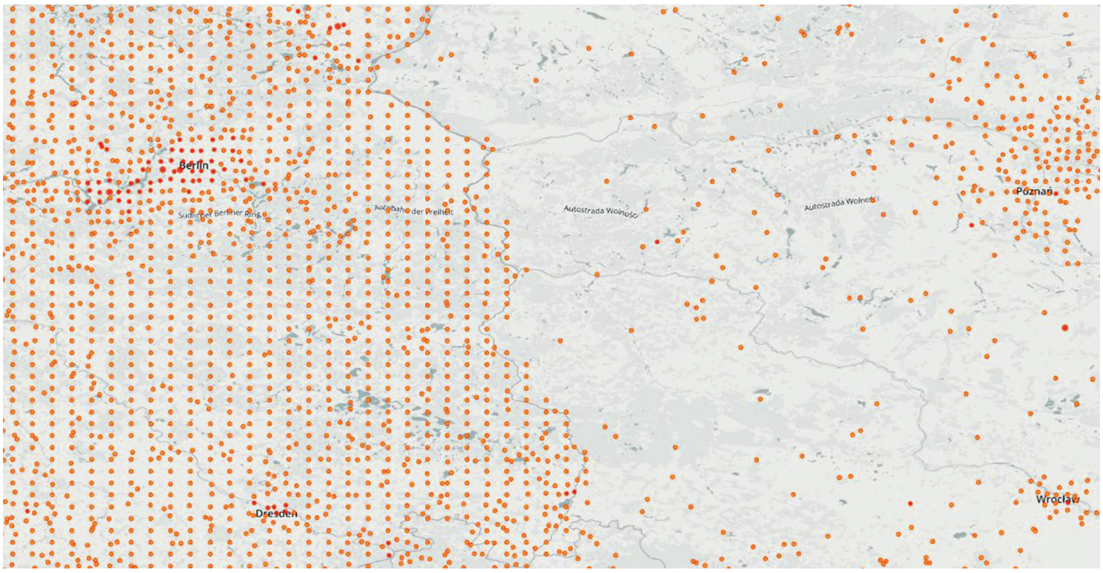
However, among the wide spectrum of possible methodological errors, excessive precision is certainly not the worst. A more critical error is using a *non-representative sample*. Rarely do we study an entire group (i.e. the general population), because statistics has developed methods that permit sound analysis using only a selected part of that group (a sample). But the caveat is, the part that is analyzed must be *representative* of the original, larger set. Only such a sample allows us to draw conclusions about the general population. For example, if we are studying the height and canopy structure of trees in a forest, using a sample consisting predominantly of edge-growing trees (or a disproportionately high share of such trees in our sample) is not a good idea, as edge trees tend to be shorter and lower-branched than those growing deeper in the forest interior. An excess of such trees in the sample will distort our resulting picture of the general tree population. This example shows that if we lack sufficient knowledge, we might make a fundamental error just by inadvertently choosing a non-representative sample. Even if all subsequent statistical analysis steps are correct, the conclusions will be unjustified.

Herbaria

One of the preconditions for a properly-selected sample is that it should be selected at random. If such randomness is achieved, we can assume the sample will be representative. However, full randomness is often difficult to attain. Studying the specimens gathered in *herbaria* – collections of pressed and dried plants that have been gathered by scholars, generally labeled with information about their systematic classification, where and when they were collected, and by whom – can serve as a good example here. Herbaria are certainly rich sources of biological data on plants: the drying process preserves their structure quite well, and in addition, genetic material can be extracted from them for research of various kinds. Herbaria currently store hundreds of millions of plant specimens from all over the world from all systematic groups. Although many such collections were mainly accumulated in the last century, quite a few of them span a much broader time range, with the oldest specimens dating back as far as nearly 500 years.

Herbarium specimens are relatively easy to access and reference, and so are often used in botanical research – such as in biometric analyses dealing with such characteristics as the dimensions of leaves, fruit, or flower parts. However, certain methodological difficulties arise here. The first problem is the issue of randomness. Ideally, a research sample should consist of *stochastically* (that is to say, randomly) collected individuals, but a collector out in the field typically violates this criterion by selecting plants that piqued his or her interest for some reason (being exceptionally small or large, easy to collect and dry, exhibiting a rare flower or leaf coloration, an unusual shape, etc.). This introduces a non-random element to the selection process, which means that the distribution of characteristic values in the sample will most likely *not* reflect that of the general population.

Another danger inherent in using herbarium specimens for research purposes derives from the drying process itself. Plant tissues contain a lot of



SOURCE: GBIF

This map of data on the occurrence of the common nettle in eastern Germany and western Poland reveals not so much an authentic picture of the species' distribution, but rather the existence of very different levels of availability of data on biodiversity

water that is removed in drying, which can significantly alter organ sizes and sometimes even the color of leaves or flowers. Experienced collectors, aware of this, often specifically record the original color on the herbarium label. For instance, in our study of about 20 species from different groups, we found that leaves lost 52–86% of their mass during drying, and their surface area decreased by 3.5–15.2% (reports in the literature indicate that the decrease can be even greater). Using methods to quantify shape changes with elliptical Fourier coefficients, it can be shown that the original form of leaves is not preserved after drying. While dried leaf blades generally do not look completely different and a botanist can still correctly identify the species, someone who does not take such changes into account and analyzes a dataset that, for instance, combines measurements of fresh and dried leaves will be making a serious methodological error. This, of course, will affect the results and their subsequent interpretation.

Big data

In many situations, problems related to the size and randomness of a sample can be solved by analyzing sufficiently large datasets ("big data"). Current IT tools and access to large amounts of data from various sources are nowadays allowing scholars to work with collections comprising not of tens or hundreds of datapoints, but hundreds of thousands or millions. In the field of biology, an excellent example of this can be found in biodiversity data. Thanks to the relatively recent processes of digitizing biological collections, the quantities of easily accessible data on the recorded occurrence of various species is growing each and every month. However, even such a huge collection is not perfect. Having even millions of points on a map indicating the occurrence of individual plants or animals still tells us nothing about the millions more that are potentially not represented there. This is illustrated well by observations of the occurrence of the common nettle (*Urtica dioica*), a species very common in Poland and Germany. The recorded distribution of the nettle looks surprising, exhibiting a large number of occurrences west of the Oder River but far fewer east of it. Each recorded observation is true, so why is the resulting overall picture nevertheless not accurate? The key lies in a bias concealed in the dataset. Since the biodiversity database in question (the world's largest: the Global Biodiversity Information Facility) happens to include an insufficient number of observations from Poland, this region is underrepresented, and so the distribution looks different on either side of the Polish-German border, which runs along the Oder River. As this example serves to show, processes of data digitization and transfer to global databases are not proceeding equally swiftly and efficiently in all countries, and so the resulting databases can be biased in various ways – the data structure is somehow skewed or distorted. Over time, of course, this should eventually work itself out, and then using big data will indeed allow for an even better, more authentic description of the world.

In summary, the landscape of biological research is fraught with challenges and potential missteps. Biologists must navigate carefully, ensuring that their methods of analysis align seamlessly with the nature of the data and the specific research questions at hand. Vigilance is key in avoiding the misinterpretation of results – even when those results themselves are accurate. Thus, in the realm of data collection and subsequent analysis, a conscious effort to sidestep these numerous pitfalls is essential for the integrity and accuracy of scientific conclusions. ■

Further reading:

Bąk J., *Statystycznie rzecz biorąc* [Statistically Speaking], 2020.

Huff D., *How to Lie With Statistics*, 1954.

Stephens-Davidowitz S., *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*, 2017.