

# **The Evolving Connection between Probability and Statistics**

**Do Statisticians Need a Probability Theory?**

**Noureddine El Karoui\***

*Prisme* N°30

**December 2014**

---

\* Noureddine El Karoui is an associate professor of statistics at the University of California, Berkeley. He was educated at École polytechnique (France) and Stanford University, where he got his Ph.D. in statistics. His main research interests are currently at the intersection of mathematical statistics and probability theory. More specifically, he has worked on random matrix theory and its applications to statistics and optimization, high-dimensional statistics and resampling/bootstrap methods in high-dimension. He is a recipient of a Sloan Fellowship in Mathematics and is a Fellow of the Institute of Mathematical Statistics. NSF support for his research is gratefully acknowledged.

## Summary

Told from the perspective of the daily life and teaching of an academic statistician, the aim of this text is to show how the field of statistics has evolved and continues to evolve, especially in relation to probability theory. The text will use two examples to illustrate that purpose. In the first case, we will look at housing data and ask whether it is possible to predict a house's future sale price based on its characteristics. In the second case, we will examine the possibility of building a SPAM filter for an e-mail account. The text will explore, at a high-level, various classical and progressively more modern statistical techniques that could be used to analyse these data, examining the role of probability theory in the development and use of these ideas, and thus illustrating the evolving connection between probability theory and statistics.

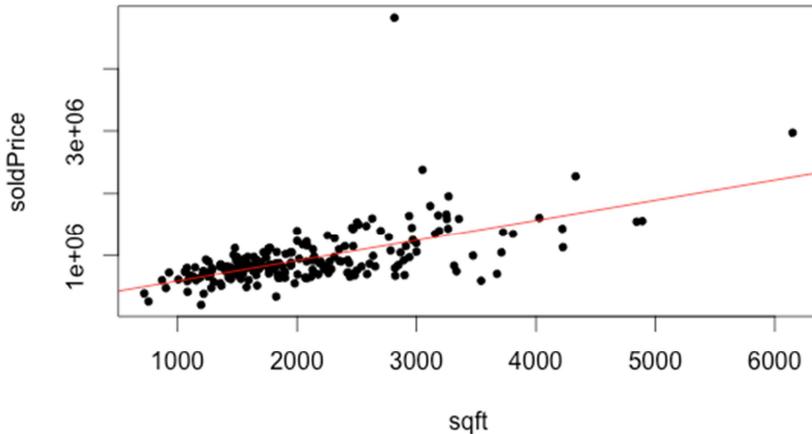
Told from the perspective of the daily life and teaching of an academic statistician, the aim of this text is to show how the field of statistics has evolved and continues to evolve, especially in relation to probability theory. Before delving further into the matter, I should also say that the talk on which this text is based was given during a one-day conference<sup>1</sup> discussing the increasing use of probabilistic ideas in many areas of science. Starting with a definition of statistics and its purpose, I would suggest a spontaneous and non-technical wording: “to make sense of and use data, exploiting them in a reasoned way”. The text will use two examples to illustrate that purpose. In the first case, we will look at housing data and ask whether it is possible to predict a house’s future sale price based on its characteristics. In the second case, we will examine the possibility of building a SPAM filter for an e-mail account. These illustrative examples were chosen, because they are easy to grasp for a non-specialist audience. Much more could be done statistically to analyse those datasets, but that would go beyond the scope of this text. The text will explore, at a high-level, various classical and progressively more modern statistical techniques that could be used to analyse these data, examining the role of probability theory in the development and use of these ideas, and thus illustrating the evolving connection between probability theory and statistics.

Let us begin with the housing example. Using *www.redfin.com*, a website that collects information about houses being sold in the United States, let us extract data from the housing market in Berkeley as if we were looking for a house to buy.

---

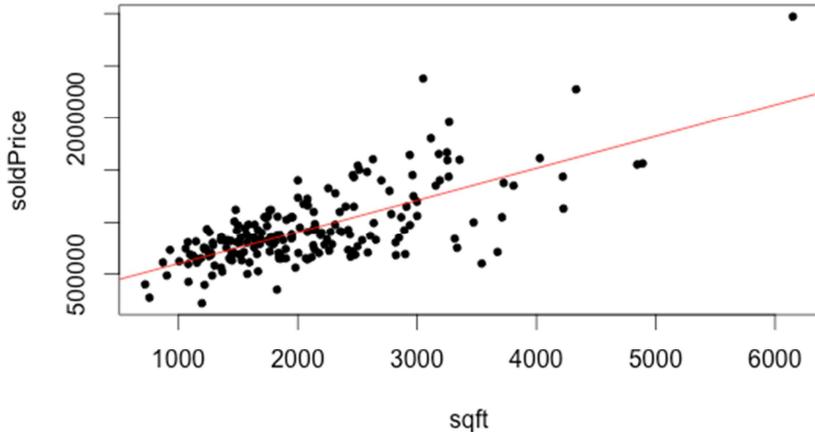
<sup>1</sup> “Are There Limits to the Probabilization of Science?”, 2 December 2013, conference organized by the Cournot Centre and Foundation and Harvard Medical School. To view the conference presentations: [http://www.centre-cournot.org/conferences\\_en.html](http://www.centre-cournot.org/conferences_en.html)

## Berkeley Housing Data: sqft vs sold price



The above data are from September 2013. Redfin gives lots of information about each house, such as sale price, short sale (or not), type of home, street address, city, state, zip code, list price (what the house is offered for), square footage, lot size, number of baths, and so on. There are about 33 attributes for each house. If a new house of interest comes on the market, we would like to know how much we should offer for it. The graph above shows a house that is significantly more expensive than all the other houses. After checking the listing, the house probably sold for around \$500,000, not \$5 million. This is most likely a reporting error, so the data point will be excluded from the data set, giving us the following graph.

**Berkeley Housing Data: sqft vs sold price  
outlier removed**



Now, we fit a line to the data. If we exclude the outlier, on average, Berkeley houses cost around \$305 per square foot, to which we need to add a base cost of around \$300,000. Based on this information, we can predict the prices of houses of various sizes (for example, houses between 1800 square feet and 2100 square feet). In the size range we are looking at, houses cost on average between \$850,000 and \$950,000. This gives us some information, but it does not tell us how accurate our prediction is. It does tell us, however, that the average price of an 1800-square-foot house in Berkeley should be somewhere between \$843,000 and \$850,000. We might, however, want to have an idea of the price of a specific house, in which case we create what is called a prediction interval. Although on average an 1800-square-foot house costs around \$850,000, if we include a measure of accuracy on top of that prediction, it tells us that the price can actually range from \$328,000 to \$1.4 million. This makes our prediction essentially worthless: the range of prices is too large to be informative. The key idea here is that the price of a house is assumed to be on average proportional to the square footage plus a base cost. We had 33 predictors at the beginning and discarded almost all of them, leaving us with only one. The intrinsic properties of each house or sale price are thus modelled as random “errors”.

Going back to the graph, the line describes the average price of a house, but each house price falls around that line. By using the stochastic nature of these errors, we are able to predict the average sale price of a house of a given size and find the

line shown in the above graph. More importantly – and that is where the strength of the stochastic argument lies – we have obtained a notion of the accuracy of our measure. This is what was used to create the prediction interval: on average, the price of an 1800-square-foot house is \$850,000, but actually it could be anywhere from \$350,000 to \$1.4 million.

The interval shown here is the interval of prices, which should contain 95 per cent of the prices of future houses with similar square footage that come on the market. This measure of accuracy tells us that knowledge alone of the average price of a house with certain characteristics gives us very little information about each individual home. Knowledge of the accuracy of our prediction tells us whether or not our prediction is valuable, or in this case worthless, because we have learned very little from them. To create this measure of accuracy, a number of probabilistic assumptions are made about the data and incorporated in the statistical software, in this case  $R$ , which produced the result we gave. For the user who may not know these assumptions,  $R$  is essentially acting as a black box. If we look more carefully at the data, these assumptions are most likely all violated, so more care is needed. It may be technically simple to get the measure of accuracy we discussed, but conceptually it requires a deep probabilistic argument. We assume that the average price of houses with similar square-footage is on the line in the graph above, and each house's price fluctuates vertically around the line. Understanding the probabilistic properties of the line that is fit to the data calls for classical probabilistic limit theory. The key import of probability theory here is to give us the measure of accuracy, without which any sort of statistical measurement is worthless. Probability theory and statistics are therefore very closely enmeshed.

Many other techniques could be used to analyse this data, however. Instead of fitting a line to the data, we could look at the house in our dataset that appears to be the most similar to the one coming on the market that we are interested in. We will say that the price of the house coming on the market will be the same as the one closest to it in our data set. That is another prediction, called the nearest neighbour method. We could also use various weighted averages. For instance, we could say that the house that we are interested in resembles a particular one, but is not that different from the others in the same area. In that case, we can average the prices of these other houses to predict the price of the new house. The weight we will give to each house in our prediction is proportional to how similar it is to the house we are

interested in. Once we are given the data, we can apply many different methods. Which method should we choose? Historically,<sup>2</sup> probabilistic analysis was used to assess *a priori* the performance of various methods, and, very importantly, to design *optimal* methods under certain modelling assumptions. “Optimal” here can be defined for instance as most accurate where the notion of accuracy is defined by the data analyst.

In this framework, the methods chosen to analyse the data thus strongly depend on the probabilistic assumptions made about the data-generating mechanism. Going back to the second graph, if we assume that the straight line accurately represents the fact that, on average, the price of the house is proportional to the square footage plus a base cost, knowing the characteristics of the random errors will tell us what method we should use to find the line. We can use different methods to fit this line to the data. If the errors are normally distributed and independent, it makes sense to use the method of least-squares, in other words, find the line that minimizes the sum of the squared vertical distances between the data points and the line we fit to the data. If the errors have a double exponential distribution, it makes sense to pick the line that minimizes the sum of absolute values of the vertical distances between the data points and the line. It is thus possible to fit different lines to the data. Having a good understanding of the stochastic characteristics of the errors or having the probability model of the data will help us find a method to pick the line, in effect telling us which line to use. To get back to our original problem, using the right method reduces our estimation error, and we obtain, to a certain degree, the optimal prediction possible.

That is what we find behind maximum likelihood methods, a very central and now old idea. The first formal developments were probably made by R.A. Fisher in the 1920s and 1930s.<sup>3</sup> The framework consists in exploiting the random mechanism we think generated the data. Let us go back now to the housing data and take, for example, the price of an 1800-square-foot house. We assume that it is a random variable with an unknown mean, denoted  $m$ , and that the price of other 1800-

---

<sup>2</sup> See, for instance, Lehmann, Erich L. (2008), *Reminiscences of a Statistician: The Company I Kept*, New York: Springer-Verlag; and Lehmann, Erich L. (2011), *Fisher, Neyman and the Creation of Classical Statistics*, New York: Springer-Verlag.

<sup>3</sup> Fisher, R.A. (1922), “On the Mathematical Foundations of Theoretical Statistics”, *Philosophical Transactions of the Royal Society of London, Series A*, 222: pp. 309–368.

square-foot houses will randomly fluctuate around that mean. Such fluctuations model the fact that some 1800-square-foot houses have more bathrooms than others, were recently remodelled, and so on. A random mechanism is thus generating the data. We can summarize, or index, this random mechanism by the parameter  $m$ , in this case the average price of an 1800-square-foot house, is unknown, but we would like to know it. If we knew the probabilistic characteristics of the fluctuations of the price of the other 1800-square-foot houses around  $m$ , also known as the error distribution,  $m$  would completely specify the distribution/probabilistic characteristics of the data.

We could also assume that the error distribution is unknown and try to estimate it from the data. That would be more complicated, but we can still deal with this complication. Similarly, the parameter indexing the data-generating process does not have to be a single number, but it could be a two or higher-dimensional parameter. Basically, the family of possible random mechanisms having generated the data is indexed by a parameter, call it  $T$ . So again, for the price of a house, if we allow the square footage to change, we model the mean price by a line, which tells us that the average price is a (linear) function of the square footage. There is a random error on top of it. The parameter  $T$  is basically the slope and the intercept for this line, so  $T$  is two-dimensional. If we assume that we know the probabilistic characteristics of the errors, if we know the parameter  $T$ , we can compute the probability of having observed the data set we are working with.<sup>4</sup> We can do this for all possible values of  $T$ . The idea is that the most natural guess for  $T$  is  $T_{MLE}$ , which maximizes the probability of seeing a data set like the one we observe.  $T_{MLE}$  is therefore inherently a probabilistic or random object, because it depends on the observed data set, which we assume is the realization of a random mechanism. That means that if we are given another data set, we will get another value for  $T_{MLE}$ . Surprisingly, it is possible to characterize the probabilistic properties of  $T_{MLE}$ , also known as its law from a single data set. We can thus say *a priori* how  $T_{MLE}$  would change if we had another data set that was generated through the same probabilistic mechanism, and if we repeated the experiments over and over again.

This characterization is somewhat tied to classical probability theory. It is possible to do so in the limiting regime where the data set is very large, that is, where

---

<sup>4</sup> I have made a technical simplification in this sentence for clarity's sake. The sentence captures the gist of the idea, but takes some liberties with formal mathematical definitions.

we have many observations. For our housing data, we can give, for example, probabilistic information about the estimate of the slope of the line we are interested in, in the limit where the number of houses we are looking at is infinite. The usual question is “how large is large”? When does this approximation become accurate? The housing market is always changing, so we cannot use data from 50 years ago to fit that line. The time period is thus limited, and for our data set, that means that we have 200 points. The question is then: is our approximation correct for 200 points? Is it giving us accurate information?

In what follows, we accept the limitations of the method, which requires us to have a model for the data generating process, and of the theory, which requires us to have a large sample size (large data set with many observations). It is nonetheless remarkable that in this context, we can use the probabilistic properties of our estimate to give both an estimate for  $T$  and a measure of accuracy of that estimate. The most important task is getting the measure of accuracy of that estimate, which is tied to how much the estimate fluctuates or changes if the experiment is repeated many times.  $T_{MLE}$  — the maximum likelihood estimator — is the most accurate estimate of  $T$ , in the limit where we have many observations.<sup>5</sup> Hence,  $T_{MLE}$  is not only a natural estimate, it also happens to be optimal. Nice results were thus achieved from the 1920s onwards, giving us a natural way to do data analysis and statistics: collect the data, perhaps designing the experiment ahead of time, and come up with a good probabilistic model for them; fit the model to the data using the maximum likelihood method. If the model is good, it should give us a near-optimal method for extracting information from the data. This is a big success story for the role of probability in science: if you have a good probabilistic model, you can obtain a near-optimal estimator of the parameters that you are interested in. This has thus become a very widely used framework.

In the late 1940s or early 1950s, Joseph Hodges at Berkeley came up with a “super-efficient estimator”<sup>6</sup> that beat the maximum likelihood estimator (MLE) in terms of accuracy at a few points and did as well everywhere else, in the setting where the size of the data set is assumed to go to infinity. The maximum likelihood

---

<sup>5</sup> This statement is true under various technical conditions on the data generating mechanism, but counter-examples do exist.

<sup>6</sup> See Le Cam, L. (1953): “On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates”, *Univ. Of Calif. Publ. In Statistics*, 1, pp. 277–330.

estimator is essentially optimal, but Hodges saw that you could outperform it at a few points, at a few values of the parameter space. His was a simple counter example, but it was a big discovery, in that it showed that there were some theoretical and conceptual problems with maximum likelihood methods.

In a related setting, Charles Stein in 1956<sup>7</sup> introduced an important and somewhat strange result: if the parameter is complicated enough — with a dimension higher than three — in the simplest possible probabilistic setting for the data generating mechanism, he could construct an estimator that was more accurate than the MLE for all values of the true parameter of interest, and for a fixed data-set size. If the previous notion of accuracy is tied to how much the estimator fluctuates, the notion of accuracy that Stein used was basically an average measure of that fluctuation.

Maximum likelihood methods have been widely used in many scientific fields, but statisticians and those working in statistical theory know that it is a natural method that has good optimality properties, but that it also suffers from some limitations. Statistical theory has required the development of many specific probabilistic tools, because statisticians are dealing with probabilistic objects that perhaps are not immediately suited to general purpose probability theory. This has created a fruitful dialogue between those working in probability theory and those working in statistical theory: at the intersection they are together creating the tools needed to understand the behaviour of a broad class of estimators and data analytic methods.

Let us turn now to the practical issues and look again at our data set. As just explained, performance and optimality are very model dependent. The assumption is that we have a good model, but what if we do not? How do we check the model? Going back to the housing example, when fitting a line to the housing data, we may become concerned, because the spread along that line seems to increase as the square footage increases. Indeed, we might expect the variability in price to change according to what segment of the housing market we are looking at. The “black box” that we ran (by using the R software somewhat naively) does not, however, assume that the variability in prices is changing with the square footage of the houses. In particular, our measure of accuracy for our estimate relies on the assumption that the

---

<sup>7</sup> Stein, C. (1956): “Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution”, *Proc. Third Berkeley Symp. Math. Statistics*, 1, University of California Press, pp. 197–206.

variability in prices is the same across all segments of the housing market, and may therefore be misleading for the data set we are looking at. Hence, a good data analysis requires knowing precisely which model is fit, being aware of its limitations and being able to adjust the metrics we care about when our data does not seem to follow the model implicitly used by the software we rely on.

Another problem is the following: does the model ever really describe the data accurately? Going back to our very simplistic housing model, we, of course, do not truly believe that the average price of a house is the sum of a fixed cost and a cost proportional to its square footage. We would need a more complicated function to accurately describe its behaviour. This is a problem of model misspecification. If we do not believe the model, can we believe the accuracy assessment that we obtain from it? Let us recall that the beauty of all this is that we are able to obtain not only an estimate, but also an estimate of accuracy of our estimate. If, however, we do not believe the model, why should we believe the measurement of accuracy? Furthermore, the asymptotic theory – which again is by now completely built into the software – requires a large data set. There are 200 observations in our housing data set. Is that enough? Do we need rather 1000, or maybe even 10 billion? How does the complexity of the parameter  $T$  affect that question? Some aspects of these questions can be tied to random matrix theory (which was discussed by other speakers at the conference): recent results in statistics<sup>8</sup> show that the properties of maximum likelihood estimators are very different when  $T$  is very complex or high dimensional than in the classical low-dimensional case; in fact it can be very suboptimal and have other undesirable properties.

One important success of statistics has thus far been to produce methods to measure quantities of interest and, at the same time, produce a measure of accuracy of those estimates. Sceptical practitioners (the author is not one of them) will tell you that they do not believe the model and trust the theory even less. Their question then becomes: can we perform this program of computing estimators and assessing their accuracy without relying on models that are doubtful (and do not in some respects fit the data) and a theory that requires a data set of essentially infinite size? They are very good practitioners, are typically very experienced, and they have a point.

---

<sup>8</sup> See, for example, El Karoui, Nouredine, Derek Bean, Peter J. Bickel, Chinghway Lim and Bin Yu (2013), “On Robust Regression with High-Dimensional Predictors”, *Proceedings of the National Academy of Sciences*, <http://www.pnas.org/content/early/2013/08/15/1307842110.full.pdf+html>.

A more modern approach was introduced at the end of the 1970s: instead of relying on probability theory and pen-and-paper computations, maybe the data set could be used more intensely. A very powerful and now widely used idea – called Bootstrap – was developed by Bradley Efron in 1979.<sup>9</sup> The Bootstrap idea is to generate many new data sets in a random fashion from the original data set. We can study the probabilistic behaviour of our estimation method by looking at how the output of the method fluctuates across the many new random data sets we have created. We then hope, or prove, that the fluctuations obtained across those newly generated data sets are the same as those we would get if we could repeat the original experiment that produced the original data set an infinite number of times. (The latter would be complicated for our housing data example....) Bootstrapping, however, would give us a good notion of the fluctuations of our estimators without the data size having to go to infinity. As the new data sets are generated from the original data set, the method should be able to take this complexity automatically into account and be less dependent on overly simplified assumptions that are sometimes required to apply an existing theory.

More specifically, if we look at the housing data, we had  $n = 204$  houses after excluding the outlier. We pick a number at random between 1 and 204, and repeat this operation 204 times. We thus had 204 points and generated 204 numbers chosen at random between 1 and 204. We call that set of numbers  $I_1$ , our index set. The same number can appear several times in  $I_1$ . From our original data set, we then create a new data set that includes only the data points indexed by the numbers appearing in  $I_1$ . So we create a new data set consisting of the original observations with indices in  $I_1$ . If  $I_1 = (10, 3, 5, 3, \dots)$ , the new data set will be  $(\text{house}_{10}; \text{house}_3; \text{house}_5; \text{house}_3, \dots)$ . We then run our estimation method on this new data set, and get a new output  $T_1^*$ . We can repeat these steps say 1000 times, getting 1000 index sets  $(I_1, I_2, \dots, I_{1000})$ , and we obtain 1000 corresponding new estimates  $(T_1^*, T_2^*, \dots, T_{1000}^*)$ , and hence 1000 different lines, which will tell us how the line we choose fluctuates when we slightly change the data set. One hopes, and can sometime prove, that the probabilistic characteristics of  $(T_1^*, T_2^*, \dots, T_{1000}^*)$  are similar to those of  $(T_1, T_2, \dots, T_{1000})$  that we would have computed had we been able to generate 1000 new data sets according to the

---

<sup>9</sup> Bradley Efron (1979), "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics* 7 (1): pp. 1–26.

(unknown) data-generating mechanism that produced our original dataset.  $(T_1, T_2, \dots, T_{1000})$  would give us a description of the probabilistic characteristics of  $T_1$  and hence could be used to answer our questions concerning its variability and accuracy, but it is in general not computable, since we do not assume that we know perfectly the data-generating mechanism. On the other hand, the bootstrap version  $(T_1^*, T_2^*, \dots, T_{1000}^*)$  is easily computable and can be used as a proxy for the uncomputable  $(T_1, T_2, \dots, T_{1000})$ .

The bootstrap brings up some provocative questions sometimes overheard in statistics classrooms among newly empowered students: if the bootstrap works, do we still need probability theory to do statistics? Do we need an asymptotic theory? Do we need the central limit theorem, a basic result in probability theory or any theorem in probability theory for that matter, if we can see or witness it repeatedly by numerical simulations? The bootstrap is known to work in many situations, but the theory remains in general asymptotic: we know it works if the sample size is large enough and if the functions of the data we are looking at are relatively “smooth” or nice in an appropriately defined technical sense. In such a case, it is a non-parametric method: we do not need to specify the probabilistic model that generated the data, but we still have to assume that the observations are independent, and so on. When addressing modern problems where the parameters estimated are high-dimensional, the bootstrap has now been shown to have serious issues or even to fail on numerous occasions, even for very “smooth” functions of the data and in very simple statistical situations.<sup>10</sup>

The next step in data-driven procedures is then to create methods that do not rely on a probabilistic model for the data and have a built-in overall accuracy assessment. Let us take the example of a SPAM data set. An email arrives, and we want to decide if it is SPAM or not. We will naturally try to use our dataset, where various characteristics of the emails are given and we are also told whether each message is SPAM or not. To solve this problem, we could try to make a probabilistic model based on the characteristics of the emails we have, trying to understand what makes a message SPAM, for instance using a technique called logistic regression (option 1). We could also run one or several methods on the data set and find a data-driven way to aggregate the methods (option 2) to help our decision-making process.

---

<sup>10</sup> See, for example, El Karoui, Nouredine, and Elizabeth Purdom, “Can We Trust the Bootstrap in High-dimension?”, Technical report 824, UC Berkeley Department of Statistics, February 2015. Under review at *JASA*.

In option 2, the accuracy will be measured globally, for example, based on the fraction of errors we make, rather than how well the method performs message by message. A standard way of measuring accuracy is to split the data at random. We keep, for example, 90 per cent of the data to “train” the method – that is, compute various parameters/characteristics that are relevant to our method on that part of the data, such as the line we found in our housing data – and we measure the accuracy of the method by seeing how it performs on the last 10 per cent of the data. We thus no longer need a theory; we can just witness how well the method performs. This idea, and its many variants such as the so-called “cross-validation” idea, is one of the most powerful techniques of accuracy assessment currently available to practitioners.

Let us now be more specific. For the SPAM dataset, there are 4601 messages labelled SPAM or not-SPAM, with 58 features (for example, the percentage of capital letters, exclamation points, dollar signs, and so on in the message). We split the data  $2/3+1/3$ , and start with the  $2/3$  called “training data”. The last third, the “test data”, will not be touched until the final step. Let us now describe “random forests”, a method proposed by Leo Breiman in 2001.<sup>11</sup> We take a random subset of the training data and a random subset of the features of the data; based on those, we create a so-called classification tree as shown below. The classification tree can be built using a method called CART.<sup>12</sup> At each node of the tree, one simply needs to perform the test described at that node to decide whether to move the email of interest left or right down the tree. Eventually, we arrive at one of the bottom nodes, a so-called leaf node, where an estimate of the probability that the message is SPAM is given.

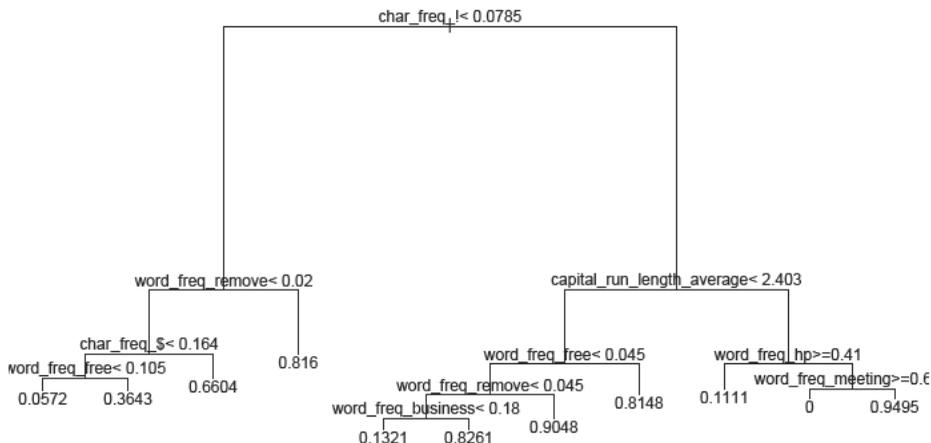
We can repeat these operations many times. In the present case, we made these random splits 1000 times and obtain 1000 classification trees – a forest of classification trees – giving us our 1000 classifiers. We then run each message in our left-out test set through the forest and obtain an empirical “probability” that it is SPAM or not. Without knowing anything about the data, it takes only a couple of minutes to perform this method on the SPAM data set and obtain a classifier with 95 per cent accuracy (that is, it classifies correctly as SPAM or not 95 per cent of the messages in the test set). From a practical point of view, this is very powerful. No

---

<sup>11</sup> Breiman, Leo (2001), “Random Forests”, *Machine Learning*, 45 (1), pp 5–32, <http://link.springer.com/article/10.1023%2FA%3A1010933404324>

<sup>12</sup> Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone (1984), *CART: Classification and Regression Trees*, Wadsworth Advanced Books and Software, Wadsworth Statistics/Probability Series.

probability model is present, except perhaps in the idea that somehow using classifiers that are somewhat uncorrelated with each other and aggregating them is better than using classifiers that are very similar to one another (this de-correlation intuition is why the 1000 classifiers are built on 1000 random subsets of the training set, using a different random set of features each time).



The dataset is used here for two purposes at once: to fit the method to the data and measure the accuracy of the procedure on future data. This frees practitioners from probabilistic modelling, because they know *a priori* that they can check performance on the data at hand and see how the procedure is going to perform on new data, assuming that the new data has similar characteristics to those present in the current dataset (a mild assumption). In essence, the only thing that matters about this data set is this particular data set. There is no theory about it. Whatever the method used, we will not need much modelling assumption; we simply see whether or not it works on our data set. The techniques coming out of the maximum likelihood ideas produce, in this context, methods to compute estimates: they have essentially no connection to any probabilistic model for generating the data, which was their original justification. We can run them on the data set at hand, but now there is essentially no probability involved. It is of course possible to run the

method even if the data do not follow the model that was first assumed to create the procedure. So the fine probability results we talked about earlier (which assume that the data follow a model specified by the user) have been replaced by numerical investigations. Probabilistic thinking is still very present in this methodology, but in a much different form than it was from the 1920s to say the mid-1970s. In particular, there is very little probability theory, even though there is a lot of randomization.

To sum up, probability and statistics are of course closely intertwined. The “refined” use of probability theory has been central in creating many statistical tools, shaping statisticians’ intuition about methods’ performance. Statistics as a field, however, is becoming more methodological: there are lots of complicated data sets to analyse; it can be difficult to create relevant probabilistic models; it takes time, and sometimes it may be impossible. The impact of probabilistic ideas is still felt in methods, but much less so in practice. In fact, a very nice and widely used textbook about modern applied statistics such as the “Elements of Statistical Learning”<sup>13</sup> has very little probabilistic content beyond basic notions that are used to frame the problems. Bayesian statistics, which I have not discussed here, naturally makes heavy use of probabilistic ideas to cast problems, but practically makes relatively little use of probability theory once the problem is set.<sup>14</sup> For a probabilistic perspective on related questions, I refer the reader to Erwin Bolthausen’s address at the 2015 Joint Statistical Meeting.<sup>15</sup>

Interesting theoretical questions remain for instance at the intersection of statistics and probability and optimization when the dimensions of the parameters we are trying to estimate are very large. As mentioned above, we can show that standard intuition about the bootstrap or maximum likelihood methods is misleading, and these methods appear to have serious problems in this modern context (failure or serious sub-optimality, respectively). In such a case, we have essentially no choice but to rely on probability theory for valid statistical inference.

---

<sup>13</sup> Hastie, T., R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, New York: Springer-Verlag.

<sup>14</sup> See for instance the textbook, *Bayesian Data Analysis*, by Gelman, A., J.B. Carlin, H.S. Stern, D. Dunson, A. Vehtari and D.B. Rubin, Boca Raton, FL, USA: Chapman and Hall/CRC Press.

<sup>15</sup> Bolthausen, Erwin, IMS Presidential Address at 2015 JSM, Seattle, “Some Thoughts about the Relations between Statistics and Probability Theory”.

Probabilistic ideas now also have a major impact on algorithms. In particular, statistics makes heavy use of linear algebra. Randomized linear algebra is becoming important: given the size of the data, the randomization step considerably speeds up the algorithm at the cost of a somewhat small loss in accuracy.

Probability theory and statistics have been and are still very closely linked. As this short text tried to show, probabilistic thinking is essential to the statistical framework. Furthermore, many widely used results in statistics rely on advanced results in probability theory. These results, however, are now embedded in the statistical software that is essential to data analysis and, possibly as a result, the importance and visibility of probability theory in the training and practice of current-day statisticians (and statistics students at the undergraduate level) is probably significantly less important than it was a few decades ago. Increased computational power has also led to the creation of many important methods for data analysis that replace the need for results in probability theory by targeted computer simulations. As explained above, these methods rely on probabilistic ideas and intuition but much less so on what most modern probabilists would consider probability theory. Probability theory is, however, starting to play a key role on the practical end in the development of provably accurate, fast and randomized statistical algorithms, and continues to be a central tool in many parts of modern theoretical statistics.