Journal of Machine Learning Research 9 (2008) 371-421 Submitted 8/07; Published 3/08

A Tutorial on Conformal Prediction

Glenn Shafer[*] GSHAFER@RUTGERS.EDU
Department of Accounting and Information Systems
Rutgers Business School
180 University Avenue
Newark, NJ 07102, USA

Vladimir Vovk VOVK@CS.RHUL.AC.UK
Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK

Editor: Sanjoy Dasgupta

Abstract

Conformal prediction uses past experience to determine precise levels of confidence in new predictions. Given an error probability ε, together with a method that makes a prediction $\hat{y}$ of a label y, it produces a set of labels, typically containing $\hat{y}$, that also contains y with probability $1-\varepsilon$. Conformal prediction can be applied to any method for producing $\hat{y}$: a nearest-neighbor method, a support-vector machine, ridge regression, etc.

Conformal prediction is designed for an on-line setting in which labels are predicted successively, each one being revealed before the next is predicted. The most novel and valuable feature of conformal prediction is that if the successive examples are sampled independently from the same distribution, then the successive predictions will be right $1-\varepsilon$ of the time, even though they are based on an accumulating data set rather than on independent data sets.

In addition to the model under which successive examples are sampled independently, other on-line compression models can also use conformal prediction. The widely used Gaussian linear model is one of these.

This tutorial presents a self-contained account of the theory of conformal prediction and works through several numerical examples. A more comprehensive treatment of the topic is provided in *Algorithmic Learning in a Random World*, by Vladimir Vovk, Alex Gammerman, and Glenn Shafer (Springer, 2005).

Keywords: confidence, on-line compression modeling, on-line learning, prediction regions

1. Introduction

How good is your prediction $\hat{y}$? If you are predicting the label y of a new object, how confident are you that $y=\hat{y}$? If the label y is a number, how close do you think it is to $\hat{y}$? In machine learning, these questions are usually answered in a fairly rough way from past experience. We expect new predictions to fare about as well as past predictions.

Conformal prediction uses past experience to determine precise levels of confidence in predictions. Given a method for making a prediction $\hat{y}$, conformal prediction produces a 95% *prediction*

*. Also at Computer Learning Research Centre, Royal Holloway.

region—a set $\Gamma^{0.05}$ that contains y with probability at least 95%. Typically $\Gamma^{0.05}$ also contains the prediction $\hat{y}$. We call $\hat{y}$ the *point prediction*, and we call $\Gamma^{0.05}$ the *region prediction*. In the case of regression, where y is a number, $\Gamma^{0.05}$ is typically an interval around $\hat{y}$. In the case of classification, where y has a limited number of possible values, $\Gamma^{0.05}$ may consist of a few of these values or, in the ideal case, just one.

Conformal prediction can be used with any method of point prediction for classification or regression, including support-vector machines, decision trees, boosting, neural networks, and Bayesian prediction. Starting from the method for point prediction, we construct a *nonconformity measure*, which measures how unusual an example looks relative to previous examples, and the *conformal algorithm* turns this nonconformity measure into prediction regions.

Given a nonconformity measure, the conformal algorithm produces a prediction region Γ^ε for every probability of error ε. The region Γ^ε is a $(1-\varepsilon)$-*prediction region*; it contains y with probability at least $1-\varepsilon$. The regions for different ε are nested: when $\varepsilon_1 \geq \varepsilon_2$, so that $1-\varepsilon_1$ is a lower level of confidence than $1-\varepsilon_2$, we have $\Gamma^{\varepsilon_1} \subseteq \Gamma^{\varepsilon_2}$. If Γ^ε contains only a single label (the ideal outcome in the case of classification), we may ask how small ε can be made before we must enlarge Γ^ε by adding a second label; the corresponding value of $1-\varepsilon$ is the confidence we assert in the predicted label.

As we explain in §4, the conformal algorithm is designed for an on-line setting, in which we predict the labels of objects successively, seeing each label after we have predicted it and before we predict the next one. Our prediction $\hat{y}_n$ of the nth label y_n may use observed features x_n of the nth object and the preceding examples $(x_1,y_1),\ldots,(x_{n-1},y_{n-1})$. The size of the prediction region Γ^ε may also depend on these details. Readers most interested in implementing the conformal algorithm may wish to look first at the elementary examples in §4.2.1 and §4.3.1 and then turn back to the earlier more general material as needed.

As we explain in §2, the on-line picture leads to a new concept of validity for prediction with confidence. Classically, a method for finding 95% prediction regions was considered valid if it had a 95% probability of containing the label predicted, because by the law of large numbers it would then be correct 95% of the time when repeatedly applied to independent data sets. But in the on-line picture, we repeatedly apply a method not to independent data sets but to an accumulating data set. After using $(x_1,y_1),\ldots,(x_{n-1},y_{n-1})$ and x_n to predict y_n, we use $(x_1,y_1),\ldots,(x_{n-1},y_{n-1}),(x_n,y_n)$ and x_{n+1} to predict y_{n+1}, and so on. For a 95% on-line method to be valid, 95% of these predictions must be correct. Under minimal assumptions, conformal prediction is valid in this new and powerful sense.

One setting where conformal prediction is valid in the new on-line sense is the one in which the examples (x_i,y_i) are sampled independently from a constant population—that is, from a fixed but unknown probability distribution Q. It is also valid under the slightly weaker assumption that the examples are probabilistically *exchangeable* (see §3) and under other on-line compression models, including the widely used Gaussian linear model (see §5). The validity of conformal prediction under these models is demonstrated in Appendix A.

In addition to the validity of a method for producing 95% prediction regions, we are also interested in its efficiency. It is efficient if the prediction region is usually relatively small and therefore informative. In classification, we would like to see a 95% prediction region so small that it contains only the single predicted label $\hat{y}_n$. In regression, we would like to see a very narrow interval around the predicted number $\hat{y}_n$.

The claim of 95% confidence for a 95% conformal prediction region is valid under exchangeability, no matter what the probability distribution the examples follow and no matter what nonconformity measure is used to construct the conformal prediction region. But the efficiency of conformal prediction will depend on the probability distribution and the nonconformity measure. If we think we know the probability distribution, we may choose a nonconformity measure that will be efficient if we are right. If we have prior probabilities for Q, we may use these prior probabilities to construct a point predictor $\hat{y}_n$ and a nonconformity measure. In the regression case, we might use as $\hat{y}_n$ the mean of the posterior distribution for y_n given the first $n-1$ examples and x_n; in the classification case, we might use the label with the greatest posterior probability. This strategy of first guaranteeing validity under a relatively weak assumption and then seeking efficiency under stronger assumptions conforms to advice long given by John Tukey and others (Tukey, 1986; Small et al., 2006).

Conformal prediction is studied in detail in *Algorithmic Learning in a Random World*, by Vovk, Gammerman, and Shafer (2005). A recent exposition by Gammerman and Vovk (2007) emphasizes connections with the theory of randomness, Bayesian methods, and induction. In this article we emphasize the on-line concept of validity, the meaning of exchangeability, and the generalization to other on-line compression models. We leave aside many important topics that are treated in *Algorithmic Learning in a Random World*, including extensions beyond the on-line picture.

2. Valid Prediction Regions

Our concept of validity is consistent with a tradition that can be traced back to Jerzy Neyman's introduction of confidence intervals for parameters (Neyman, 1937) and even to work by Laplace and others in the late 18th century. But the shift of emphasis to prediction (from estimation of parameters) and to the on-line setting (where our prediction rule is repeatedly updated) involves some rearrangement of the furniture.

The most important novelty in conformal prediction is that its successive errors are probabilistically independent. This allows us to interpret "being right 95% of the time" in an unusually direct way. In §2.1, we illustrate this point with a well-worn example, normally distributed random variables.

In §2.2, we contrast confidence with full-fledged conditional probability. This contrast has been the topic of endless debate between those who find confidence methods informative (classical statisticians) and those who insist that full-fledged probabilities based on all one's information are always preferable, even if the only available probabilities are very subjective (Bayesians). Because the debate usually focuses on estimating parameters rather than predicting future observations, and because some readers may be unaware of the debate, we take the time to explain that we find the concept of confidence useful for prediction in spite of its limitations.

2.1 An Example of Valid On-Line Prediction

A 95% prediction region is *valid* if it contains the truth 95% of the time. To make this more precise, we must specify the set of repetitions envisioned. In the on-line picture, these are successive predictions based on accumulating information. We make one prediction after another, always knowing the outcome of the preceding predictions.

To make clear what validity means and how it can be obtained in this on-line picture, we consider prediction under an assumption often made in a first course in statistics:

> Random variables $z_1, z_2, \ldots$ are independently drawn from a normal distribution with unknown mean and variance.

Prediction under this assumption was discussed in 1935 by R. A. Fisher, who explained how to give a 95% prediction interval for z_n based on $z_1, \ldots, z_{n-1}$ that is valid in our sense. We will state Fisher's prediction rule, illustrate its application to data, and explain why it is valid in the on-line setting.

As we will see, the predictions given by Fisher's rule are too weak to be interesting from a modern machine-learning perspective. This is not surprising, because we are predicting z_n based on old examples $z_1, \ldots, z_{n-1}$ alone. In general, more precise prediction is possible only in the more favorable but more complicated set-up where we know some features x_n of the new example and can use both x_n and the old examples to predict some other feature y_n. But the simplicity of the set-up where we predict z_n from $z_1, \ldots, z_{n-1}$ alone will help us make the logic of valid prediction clear.

2.1.1 FISHER'S PREDICTION INTERVAL

Suppose we observe the z_i in sequence. After observing z_1 and z_2, we start predicting; for $n = 3, 4, \ldots$, we predict z_n after having seen $z_1, \ldots, z_{n-1}$. The natural point predictor for z_n is the average so far:

$$\bar{z}_{n-1} := \frac{1}{n-1} \sum_{i=1}^{n-1} z_i,$$

but we want to give an interval that will contain z_n 95% of the time. How can we do this? Here is Fisher's answer (1935):

1. In addition to calculating the average $\bar{z}_{n-1}$, calculate

   $$s_{n-1}^2 := \frac{1}{n-2} \sum_{i=1}^{n-1} (z_i - \bar{z}_{n-1})^2,$$

   which is sometimes called the sample variance. We can usually assume that it is non-zero.

2. In a table of percentiles for t-distributions, find $t_{n-2}^{0.025}$, the point that the t-distribution with $n-2$ degrees of freedom exceeds exactly 2.5% of the time.

3. Predict that z_n will be in the interval

   $$\bar{z}_{n-1} \pm t_{n-2}^{0.025}\, s_{n-1} \sqrt{\frac{n}{n-1}}. \tag{1}$$

Fisher based this procedure on the fact that

$$\frac{z_n - \bar{z}_{n-1}}{s_{n-1}} \sqrt{\frac{n-1}{n}}$$

has the t-distribution with $n-2$ degrees of freedom, which is symmetric about 0. This implies that (1) will contain z_n with probability 95% regardless of the values of the mean and variance.

2.1.2 A NUMERICAL EXAMPLE

We can illustrate (1) using some numbers generated in 1900 by the students of Emanuel Czuber (1851–1925). These numbers are integers, but they theoretically have a binomial distribution and are therefore approximately normally distributed.[1]

Here are Czuber's first 19 numbers, $z_1, \ldots, z_{19}$:

$$17, 20, 10, 17, 12, 15, 19, 22, 17, 19, 14, 22, 18, 17, 13, 12, 18, 15, 17.$$

From them, we calculate

$$\bar{z}_{19} = 16.53, \qquad s_{19} = 3.31.$$

The upper 2.5% point for the t-distribution with 18 degrees of freedom, $t_{18}^{0.025}$, is 2.101. So the prediction interval (1) for z_{20} comes out to $[9.40, 24.13]$.

Taking into account our knowledge that z_{20} will be an integer, we can say that the 95% prediction is that z_{20} will be an integer between 10 and 24, inclusive. This prediction is correct; z_{20} is 16.

2.1.3 ON-LINE VALIDITY

Fisher did not have the on-line picture in mind. He probably had in mind a picture where the formula (1) is used repeatedly but in entirely separate problems. For example, we might conduct many separate experiments that each consists of drawing 100 random numbers from a normal distribution and then predicting a 101st draw using (1). Each experiment might involve a different normal distribution (a different mean and variance), but provided the experiments are independent from each other, the law of large numbers will apply. Each time the probability is 95% that z_{101} will be in the interval, and so this event will happen approximately 95% of the time.

The on-line story may seem more complicated, because the experiment involved in predicting z_{101} from $z_1, \ldots, z_{100}$ is not entirely independent of the experiment involved in predicting, say, z_{105} from $z_1, \ldots, z_{104}$. The 101 random numbers involved in the first experiment are all also involved in the second. But as a master of the analytical geometry of the normal distribution (Fisher, 1925; Efron, 1969), Fisher would have noticed, had he thought about it, that this overlap does not actually matter. As we show in Appendix A.3, the events

$$\bar{z}_{n-1} - t_{n-2}^{0.025}\, s_{n-1} \sqrt{\frac{n}{n-1}} \le z_n \le \bar{z}_{n-1} + t_{n-2}^{0.025}\, s_{n-1} \sqrt{\frac{n}{n-1}} \tag{2}$$

for successive n are probabilistically independent in spite of the overlap. Because of this independence, the law of large numbers again applies. Knowing each event has probability 95%, we can conclude that approximately 95% of them will happen. We call the events (2) *hits*.

The prediction interval (1) generalizes to linear regression with normally distributed errors, and on-line hits remain independent in this general setting. Even though formulas for these linear-regression prediction intervals appear in textbooks, the independence of their on-line hits was not noted prior to our work on conformal prediction. Like Fisher, the textbook authors did not have the

1. Czuber's students randomly drew balls from an urn containing six balls, numbered 1 to 6. Each time they drew a ball, they noted its label and put it back in the urn. After each 100 draws, they recorded the number of times that the ball labeled with a 1 was drawn (Czuber, 1914, pp. 329–335). This should have a binomial distribution with parameters 100 and $1/6$, and it is therefore approximately normal with mean $100/6 = 16.67$ and standard deviation $\sqrt{500/36} = 3.73$.

on-line setting in mind. They imagined just one prediction being made in each case where data is accumulated.

We will return to the generalization to linear regression in §5.3.2. There we will derive the textbook intervals as conformal prediction regions within the *on-line Gaussian linear model*, an on-line compression model that uses slightly weaker assumptions than the classical assumption of independent and normally distributed errors.

2.2 Confidence Says Less than Probability.

Neyman's notion of confidence looks at a procedure before observations are made. Before any of the z_i are observed, the event (2) involves multiple uncertainties: $\bar{z}_{n-1}$, s_{n-1}, and z_n are all uncertain. The probability that these three quantities will turn out so that (2) holds is 95%.

We might ask for more than this. It is after we observe the first $n-1$ examples that we calculate $\bar{z}_{n-1}$ and s_{n-1} and then calculate the interval (1), and we would like to be able to say at this point that there is still a 95% probability that z_n will be in (1). But this, it seems, is asking for too much. The assumptions we have made are insufficient to enable us to find a numerical probability for (2) that will be valid at this late date. In theory there is a conditional probability for (2) given $z_1,\ldots,z_{n-1}$, but it involves the unknown mean and variance of the normal distribution.

Perhaps the matter is best understood from the game-theoretic point of view. A probability can be thought of as an offer to bet. A 95% probability, for example, is an offer to take either side of a bet at 19 to 1 odds. The probability is valid if the offer does not put the person making it at a disadvantage, inasmuch as a long sequence of equally reasonable offers will not allow an opponent to multiply the capital he or she risks by a large factor (Shafer and Vovk, 2001). When we assume a probability model (such as the normal model we just used or the on-line compression models we will study later), we are assuming that the model's probabilities are valid in this sense before any examples are observed. Matters may be different afterwards.

In general, a 95% conformal predictor is a rule for using the preceding examples $(x_1,y_1),\ldots,(x_{n-1},y_{n-1})$ and a new object x_n to give a set, say

$$\Gamma^{0.05}((x_1,y_1),\ldots,(x_{n-1},y_{n-1}),x_n), \tag{3}$$

that we predict will contain y_n. If the predictor is valid, the prediction

$$y_n \in \Gamma^{0.05}((x_1,y_1),\ldots,(x_{n-1},y_{n-1}),x_n)$$

will have a 95% probability before any of the examples are observed, and it will be safe, at that point, to offer 19 to 1 odds on it. But after we observe $(x_1,y_1),\ldots,(x_{n-1},y_{n-1})$ and x_n and calculate the set (3), we may want to withdraw the offer.

Particularly striking instances of this phenomenon can arise in the case of classification, where there are only finitely many possible labels. We will see one such instance in §4.3.1, where we consider a classification problem in which there are only two possible labels, s and v. In this case, there are only four possibilities for the prediction region:

1. $\Gamma^{0.05}((x_1,y_1),\ldots,(x_{n-1},y_{n-1}),x_n)$ contains only s.
2. $\Gamma^{0.05}((x_1,y_1),\ldots,(x_{n-1},y_{n-1}),x_n)$ contains only v.
3. $\Gamma^{0.05}((x_1,y_1),\ldots,(x_{n-1},y_{n-1}),x_n)$ contains both s and v.

William S. Gossett
1876–1937

Ronald A. Fisher
1890–1962

Jerzy Neyman
1894–1981

Figure 1: **Three influential statisticians.** Gossett, who worked as a statistician for the Guinness brewery in Dublin, introduced the t-distribution to English-speaking statisticians (Student, 1908). Fisher, whose applied and theoretical work invigorated mathematical statistics in the 1920s and 1930s, refined, promoted, and extended Gossett's work. Neyman was one of the most influential leaders in the subsequent movement to use advanced probability theory to give statistics a firmer foundation and further extend its applications.

4. $\Gamma^{0.05}((x_1,y_1),\ldots,(x_{n-1},y_{n-1}),x_n)$ is empty.

The third and fourth cases can occur even though $\Gamma^{0.05}$ is valid. When the third case happens, the prediction, though uninformative, is certain to be correct. When the fourth case happens, the prediction is clearly wrong. These cases are consistent with the prediction being right 95% of the time. But when we see them arise, we know whether the particular value of n is one of the 95% where we are right or the one of the 5% where we are wrong, and so the 95% will not remain valid as a probability defining betting odds.

In the case of normally distributed examples, Fisher called the 95% probability for z_n being in the interval (1) a "fiducial probability," and he seems to have believed that it would not be susceptible to a gambling opponent who knows the first $n-1$ examples (see Fisher, 1973, pp. 119–125). But this turned out not to be the case (Robinson, 1975). For this and related reasons, most scientists who use Fisher's methods have adopted the interpretation offered by Neyman, who wrote about "confidence" rather than fiducial probability and emphasized that a confidence level is a full-fledged probability only before we acquire data. It is the procedure or method, not the interval or region it produces when applied to particular data, that has a 95% probability of being correct.

Neyman's concept of confidence has endured in spite of its shortcomings. It is widely taught and used in almost every branch of science. Perhaps it is especially useful in the on-line setting. It is useful to know that 95% of our predictions are correct even if we cannot assert a full-fledged 95% probability for each prediction when we make it.

3. Exchangeability

Consider variables $z_1,\ldots,z_N$. Suppose that for any collection of N values, the $N!$ different orderings are equally likely. Then we say that $z_1,\ldots,z_N$ are *exchangeable*. The assumption that $z_1,\ldots,z_N$ are exchangeable is slightly weaker than the more familiar assumption that they are drawn independently from a probability distribution.

In our book (Vovk et al., 2005), conformal prediction is first explained under the assumption that $z_1, \ldots, z_N$ are independently drawn from a probability distribution (or that they are "random," as we say there), and then it is pointed out that this assumption can be relaxed to the assumption that $z_1, \ldots, z_N$ are exchangeable. When we introduce conformal prediction in this article, in §4, we assume only exchangeability from the outset, hoping that this will make the logic of the method as clear as possible. Once this logic is clear, it is easy to see that it works not only for the exchangeability model but also for other on-line compression models (§5).

In this section we look at the relationship between exchangeability and independence and then give a backward-looking definition of exchangeability that can be understood game-theoretically. We conclude with a law of large numbers for exchangeable sequences, which will provide the basis for our confidence that our 95% prediction regions are right 95% of the time.

3.1 Exchangeability and Independence

Although the definition of exchangeability we just gave may be clear enough at an intuitive level, it has two technical problems that make it inadequate as a formal mathematical definition: (1) in the case of continuous distributions, any specific values for $z_1, \ldots, z_N$ will have probability zero, and (2) in the case of discrete distributions, two or more of the z_i might take the same value, and so a list of possible values $a_1, \ldots, a_N$ might contain fewer than n distinct values.

One way of avoiding these technicalities is to use the concept of a permutation, as follows:

> *Definition of exchangeability using permutations.* The variables $z_1, \ldots, z_N$ are exchangeable if for every permutation τ of the integers $1, \ldots, N$, the variables $w_1, \ldots, w_N$, where $w_i = z_{\tau(i)}$, have the same joint probability distribution as $z_1, \ldots, z_N$.

We can extend this to a definition of exchangeability for an infinite sequence of variables: $z_1, z_2, \ldots$ are exchangeable if $z_1, \ldots, z_N$ are exchangeable for every N.

This definition makes it easy to see that independent and identically distributed random variables are exchangeable. Suppose $z_1, \ldots, z_N$ all take values from the same example space $\mathbf{Z}$, all have the same probability distribution Q, and are independent. Then their joint distribution satisfies

$$\Pr(z_1 \in A_1 \;\&\; \ldots \;\&\; z_N \in A_N) = Q(A_1) \cdots Q(A_N)$$

for any[2] subsets $A_1, \ldots, A_N$ of $\mathbf{Z}$, where $Q(A)$ is the probability Q assigns to an example being in A. Because permuting the factors $Q(A_n)$ does not change their product, and because a joint probability distribution for $z_1, \ldots, z_N$ is determined by the probabilities it assigns to events of the form $\{z_1 \in A_1 \;\&\; \ldots \;\&\; z_N \in A_N\}$, this makes it clear that $z_1, \ldots, z_N$ are exchangeable.

Exchangeability implies that variables have the same distribution. On the other hand, exchangeable variables need not be independent. Indeed, when we average two or more distinct joint probability distributions under which variables are independent, we usually get a joint probability distribution under which they are exchangeable (averaging preserves exchangeability) but not independent (averaging usually does not preserve independence). According to a famous theorem by de Finetti, an exchangeable joint distribution for an infinite sequence of distinct variables is exchangeable only if it is a mixture of joint distributions under which the variables are independent (Hewitt and Savage, 1955). As Table 1 shows, the picture is more complicated in the finite case.

2. We leave aside technicalities involving measurability.

$\Pr(z_1 = \mathrm{H} \,\&\, z_2 = \mathrm{H})$	$\Pr(z_1 = \mathrm{H} \,\&\, z_2 = \mathrm{T})$
$\Pr(z_1 = \mathrm{T} \,\&\, z_2 = \mathrm{H})$	$\Pr(z_1 = \mathrm{T} \,\&\, z_2 = \mathrm{T})$

0.81	0.09
0.09	0.01

0.41	0.09
0.09	0.41

0.10	0.40
0.40	0.10

Table 1: **Examples of exchangeability.** We consider variables z_1 and z_2, each of which comes out H or T. Exchangeability requires only that $\Pr(z_1 = \mathrm{H} \,\&\, z_2 = \mathrm{T}) = \Pr(z_1 = \mathrm{T} \,\&\, z_2 = \mathrm{H})$. Three examples of distributions for z_1 and z_2 with this property are shown. On the left, z_1 and z_2 are independent and identically distributed; both come out H with probability 0.9. The middle example is obtained by averaging this distribution with the distribution in which the two variables are again independent and identically distributed but T's probability is 0.9. The distribution on the right, in contrast, cannot be obtained by averaging distributions under which the variables are independent and identically distributed. Examples of this last type disappear as we ask for a larger and larger number of variables to be exchangeable.

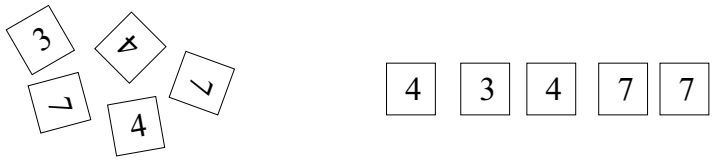


Figure 2: **Ordering the tiles.** Joe gives Bill a bag containing five tiles, and Bill arranges them to form the list 43477. Bill can calculate conditional probabilities for which z_i had which of the five values. His conditional probability for $z_5 = 4$, for example, is $2/5$. There are $(5!)/(2!)(2!) = 30$ ways of assigning the five values to the five variables; $(z_1, z_2, z_3, z_4, z_5) = (4, 3, 4, 7, 7)$ is one of these, and they all have the same probability, $1/30$.

3.2 Backward-Looking Definitions of Exchangeability

Another way of defining exchangeability looks backwards from a situation where we know the unordered values of $z_1, \ldots, z_N$.

Suppose Joe has observed $z_1, \ldots, z_N$. He writes each value on a tile resembling those used in Scrabble©, puts the N tiles in a bag, shakes the bag, and gives it to Bill to inspect. Bill sees the N values (some possibly equal to each other) without knowing their original order. Bill also knows the joint probability distribution for $z_1, \ldots, z_N$. So he obtains probabilities for the ordering of the tiles by conditioning this joint distribution on his knowledge of the bag. The joint distribution is exchangeable if and only if these conditional probabilities are the same as the probabilities for the result of ordering the tiles by successively drawing them at random from the bag without replacement.

To make this into a definition of exchangeability, we formalize the notion of a bag. A *bag* (or *multiset*, as it is sometimes called) is a collection of elements in which repetition is allowed. It is like a set inasmuch as its elements are unordered but like a list inasmuch as an element can occur more than once. We write $\wr a_1, \ldots, a_N \int$ for the bag obtained from the list $a_1, \ldots, a_N$ by removing information about the ordering.

Here are three equivalent conditions on the joint distribution of a sequence of random variables $z_1, \ldots, z_N$, any of which can be taken as the definition of exchangeability.

1. For any bag B of size N, and for any examples $a_1, \ldots, a_N$,

$$\Pr(z_1 = a_1 \;\&\; \ldots \;\&\; z_N = a_N \mid \wr z_1, \ldots, z_N \int = B)$$

   is equal to the probability that successive random drawings from the bag B without replacement produces first a_N, then a_{N-1}, and so on, until the last element remaining in the bag is a_1.

2. For any n, $1 \le n \le N$, z_n is independent of $z_{n+1}, \ldots, z_N$ given the bag $\wr z_1, \ldots, z_n \int$ and for any bag B of size n,

$$\Pr(z_n = a \mid \wr z_1, \ldots, z_n \int = B) = \frac{k}{n}, \tag{4}$$

   where k is the number of times a occurs in B.

3. For any bag B of size N, and for any examples $a_1, \ldots, a_N$,

$$\Pr(z_1 = a_1 \;\&\; \ldots \;\&\; z_N = a_N \mid \wr z_1, \ldots, z_N \int = B) = \begin{cases} \frac{n_1! \cdots n_k!}{N!} & \text{if } B = \wr a_1, \ldots, a_N \int \\ 0 & \text{if } B \neq \wr a_1, \ldots, a_N \int, \end{cases} \tag{5}$$

   where k is the number of distinct values among the a_i, and $n_1, \ldots, n_k$ are the respective numbers of times they occur. (If the a_i are all distinct, the expression $n_1! \cdots n_k!/(N!)$ reduces to $1/(N!)$.)

We leave it to the reader to verify that these three conditions are equivalent to each other. The second condition, which we will emphasize, is represented pictorially in Figure 3.

The backward-looking conditions are also equivalent to the definition of exchangeability using permutations given on p. 378. This equivalence is elementary in the case where every possible sequence of values $a_1, \ldots, a_n$ has positive probability. But complications arise when this probability is zero, because the conditional probability on the left-hand side of (5) is then defined only with probability one by the joint distribution. We do not explore these complications here.

3.3 The Betting Interpretation of Exchangeability

The framework for probability developed in Shafer and Vovk (2001) formalizes classical results of probability theory, such as the law of large numbers, as theorems of game theory: a bettor can multiply the capital he risks by a large factor if these results do not hold. This allows us to express the empirical interpretation of given probabilities in terms of betting, using what we call *Cournot's principle*: the odds determined by the probabilities will not allow a bettor to multiply the capital he or she risks by a large factor (Shafer, 2007).

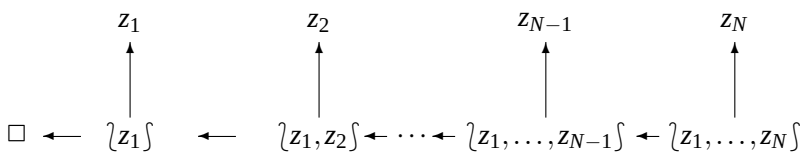


Figure 3: **Backward probabilities, step by step.** The two arrows backwards from each bag $\wr z_1, \ldots, z_n \int$ symbolize drawing an example z_n out at random, leaving the smaller bag $\wr z_1, \ldots, z_{n-1} \int$. The probabilities for the result of the drawing are given by (4). Readers familiar with Bayes nets (Cowell et al., 1999) will recognize this diagram as an example; conditional on each variable, a joint probability distribution is given for its children (the variables to which arrows from it point), and given the variable, its descendants are independent of its ancestors.

By applying this idea to the sequence of probabilities (4), we obtain a betting interpretation of exchangeability. Think of Joe and Bill as two players in a game that moves backwards from point N in Figure 3. At each step, Joe provides new information and Bill bets. Designate by $\mathcal{K}_N$ the total capital Bill risks. He begins with this capital at N, and at each step n he bets on what z_n will turn out to be. When he bets at step n, he cannot risk losing more than he has at that point (because he is not risking more than $\mathcal{K}_N$ in the whole game), but otherwise he can bet as much as he wants for or against each possible value a for z_n at the odds $(k/n) : (1-k/n)$, where k is the number of elements in the current bag equal to a.

For brevity, we write B_n for the bag $\wr z_1, \ldots, z_n \int$, and for simplicity, we set the initial capital K_N equal to \$1. This gives the following protocol:

THE BACKWARD-LOOKING BETTING PROTOCOL

Players: Joe, Bill

$\mathcal{K}_N := 1$.
Joe announces a bag B_N of size N.
FOR $n = N, N-1, \ldots, 2, 1$
  Bill bets on z_n at odds set by (4).
  Joe announces $z_n \in B_n$.
  $\mathcal{K}_{n-1} := \mathcal{K}_n +$ Bill's net gain.
  $B_{n-1} := B_n \setminus \wr z_n \int$.

Constraint: Bill must move so that his capital $\mathcal{K}_n$ will be nonnegative for all n no matter how Joe moves.

Our betting interpretation of exchangeability is that Bill will not multiply his initial capital $\mathcal{K}_N$ by a large factor in this game.

The permutation definition of exchangeability does not lead to an equally simple betting interpretation, because the probabilities for $z_1, \ldots, z_N$ to which the permutation definition refers are not determined by the mere assumption of exchangeability.

3.4 A Law of Large Numbers for Exchangeable Sequences

As we noted when we studied Fisher's prediction interval in §2.1.3, the validity of on-line prediction requires more than having a high probability of a hit for each individual prediction. We also need a law of large numbers, so that we can conclude that a high proportion of the high-probability predictions will be correct. As we show in §A.3, the successive hits in the case of Fisher's region predictor are independent, so that the usual law of large numbers applies. What can we say in the case of conformal prediction under exchangeability?

Suppose $z_1, \ldots, z_N$ are exchangeable, drawn from an example space $\mathbf{Z}$. In this context, we adopt the following definitions.

- An event E is an *n-event*, where $1 \leq n \leq N$, if its happening or failing is determined by the value of z_n and the value of the bag $\wr z_1, \ldots, z_{n-1} \int$.
- An n-event E is ε-*rare* if
$$\Pr(E \mid \wr z_1, \ldots, z_n \int) \leq \varepsilon. \tag{6}$$

The left-hand side of the inequality (6) is a random variable, because the bag $\wr z_1, \ldots, z_n \int$ is random. The inequality says that this random variable never exceeds ε.

As we will see in the next section, the successive errors for a conformal predictor are ε-rare n-events. So the validity of conformal prediction follows from the following informal proposition.

Informal Proposition 1 *Suppose N is large, and the variables $z_1, \ldots, z_N$ are exchangeable. Suppose E_n is an ε-rare n-event for $n = 1, \ldots, N$. Then the law of large numbers applies; with very high probability, no more than approximately the fraction ε of the events $E_1, \ldots, E_N$ will happen.*

In Appendix A, we formalize this proposition in two ways: classically and game-theoretically.

The classical approach appeals to the classical weak law of large numbers, which tells us that if $E_1, \ldots, E_N$ are mutually independent and each have probability exactly ε, and N is sufficiently large, then there is a very high probability that the fraction of the events that happen will be close to ε. We show in §A.1 that if (6) holds with equality, then E_n are mutually independent and each of them has unconditional probability ε. Having the inequality instead of equality means that the E_n are even less likely to happen, and this will not reverse the conclusion that few of them will happen.

The game-theoretic approach is more straightforward, because the game-theoretic version law of large numbers does not require independence or exact levels of probability. In the game-theoretic framework, the only question is whether the probabilities specified for successive events are rates at which a bettor can place successive bets. The Backward-Looking Betting Protocol says that this is the case for ε-rare n-events. As Bill moves through the protocol from N to 1, he is allowed to bet against each error E_n at a rate corresponding to its having probability ε or less. So the game-theoretic weak law of large numbers (Shafer and Vovk, 2001, pp. 124–126) applies directly. Because the game-theoretic framework is not well known, we state and prove this law of large numbers, specialized to the Backward-Looking Betting Protocol, in §A.2.

4. Conformal Prediction under Exchangeability

We are now in a position to state the conformal algorithm under exchangeability and explain why it produces valid nested prediction regions.

We distinguish two cases of on-line prediction. In both cases, we observe examples $z_1, \ldots, z_N$ one after the other and repeatedly predict what we will observe next. But in the second case we have more to go on when we make each prediction.

1. *Prediction from old examples alone.* Just before observing z_n, we predict it based on the previous examples, $z_1, \ldots, z_{n-1}$.

2. *Prediction using features of the new object.* Each example z_i consists of an object x_i and a label y_i. In symbols: $z_i = (x_i, y_i)$. We observe in sequence $x_1, y_1, \ldots, x_N, y_N$. Just before observing y_n, we predict it based on what we have observed so far, x_n and the previous examples $z_1, \ldots, z_{n-1}$.

Prediction from old examples may seem relatively uninteresting. It can be considered a special case of prediction using features x_n of new examples—the case in which the x_n provide no information, and this special case we may have too little information to make useful predictions. But its simplicity makes prediction with old examples alone advantageous as a setting for explaining the conformal algorithm, and as we will see, it is then straightforward to take account of the new information x_n.

Conformal prediction requires that we first choose a nonconformity measure, which measures how different a new example is from old examples. In §4.1, we explain how nonconformity measures can be obtained from methods of point prediction. In §4.2, we state and illustrate the conformal algorithm for predicting new examples from old examples alone. In §4.3, we generalize to prediction with the help of features of a new example. In §4.4, we explain why conformal prediction produces the best possible valid nested prediction regions under exchangeability. Finally, in §4.5 we discuss the implications of the failure of the assumption of exchangeability.

For some readers, the simplicity of the conformal algorithm may be obscured by its generality and the scope of our preliminary discussion of nonconformity measures. We encourage such readers to look first at §4.2.1, §4.3.1, and §4.3.2, which provide largely self-contained accounts of the algorithm as it applies to some small data sets.

4.1 Nonconformity Measures

The starting point for conformal prediction is what we call a *nonconformity measure*, a real-valued function $A(B,z)$ that measures how different an example z is from the examples in a bag B. The conformal algorithm assumes that a nonconformity measure has been chosen. The algorithm will produce valid nested prediction regions using any real-valued function $A(B,z)$ as the nonconformity measure. But the prediction regions will be efficient (small) only if $A(B,z)$ measures well how different z is from the examples in B.

A method $\hat{z}(B)$ for obtaining a point prediction $\hat{z}$ for a new example from a bag B of old examples usually leads naturally to a nonconformity measure A. In many cases, we only need to add a way of measuring the distance $d(z,z')$ between two examples. Then we define A by

$$A(B,z) := d(\hat{z}(B), z). \tag{7}$$

The prediction regions produced by the conformal algorithm do not change when the nonconformity measure A is transformed monotonically. If A is nonnegative, for example, replacing A with A^2 will make no difference. Consequently, the choice of the distance measure $d(z,z')$ is relatively unimportant. The important step in determining the nonconformity measure A is choosing the point predictor $\hat{z}(B)$.

To be more concrete, suppose the examples are real numbers, and write $\bar{z}_B$ for the average of the numbers in B. If we take this average as our point predictor $\hat{z}(B)$, and we measure the distance between two real numbers by the absolute value of their difference, then (7) becomes

$$A(B,z) := |\bar{z}_B - z|. \tag{8}$$

If we use the median of the numbers in B instead of their average as $\hat{z}(B)$, we get a different nonconformity measure, which will produce different prediction regions when we use the conformal algorithm. On the other hand, as we have already said, it will make no difference if we replace the absolute difference $d(z,z') = |z-z'|$ with the squared difference $d(z,z') = (z-z')^2$, thus squaring A.

We can also vary (8) by including the new example in the average:

$$A(B,z) := |(\text{average of } z \text{ and all the examples in } B) - z|. \tag{9}$$

This results in the same prediction regions as (8), because if B has n elements, then

$$|(\text{average of } z \text{ and all the examples in } B) - z| = \left| \frac{n\bar{z}_B + z}{n+1} - z \right| = \frac{n}{n+1}|\bar{z}_B - z|,$$

and as we have said, conformal prediction regions are not changed by a monotonic transformation of the nonconformity measure. In the numerical example that we give in §4.2.1 below, we use (9) as our nonconformity measure.

When we turn to the case where features of a new object help us predict a new label, we will consider, among others, the following two nonconformity measures:

Distance to the nearest neighbors for classification. Suppose $B = \wr z_1, \ldots, z_{n-1} \int$, where each z_i consists of a number x_i and a nonnumerical label y_i. Again we observe x but not y for a new example $z = (x,y)$. The nearest-neighbor method finds the x_i closest to x and uses its label y_i as our prediction of y. If there are only two labels, or if there is no natural way to measure the distance between labels, we cannot measure how wrong the prediction is; it is simply right or wrong. But it is natural to measure the nonconformity of the new example (x,y) to the old examples (x_i,y_i) by comparing x's distance to old objects with the same label to its distance to old objects with a different label. For example, we can set

$$\begin{aligned} A(B,z) :&= \frac{\min\{|x_i - x| : 1 \le i \le n-1, y_i = y\}}{\min\{|x_i - x| : 1 \le i \le n-1, y_i \ne y\}} \\ &= \frac{\text{distance to } z\text{'s nearest neighbor in } B \text{ with the same label}}{\text{distance to } z\text{'s nearest neighbor in } B \text{ with a different label}}. \end{aligned} \tag{10}$$

Distance to a regression line. Suppose $B = \wr (x_1,y_1), \ldots, (x_l,y_l) \int$, where the x_i and y_i are numbers. The most common way of fitting a line to such pairs of numbers is to calculate the averages

$$\bar{x}_l := \sum_{j=1}^{l} x_j \quad \text{and} \quad \bar{y}_l := \sum_{j=1}^{l} y_j,$$

and then the coefficients

$$b_l = \frac{\sum_{j=1}^{l}(x_j - \bar{x}_l)y_j}{\sum_{j=1}^{l}(x_j - \bar{x}_l)^2} \quad \text{and} \quad a_l = \bar{y}_l - b_l\bar{x}_l.$$

This gives the *least-squares line* $y = a_l + b_l x$. The coefficients a_l and b_l are not affected if we change the order of the z_i; they depend only on the bag B.

If we observe a bag $B = \wr z_1, \ldots, z_{n-1} \int$ of examples of the form $z_i = (x_i, y_i)$ and also x but not y for a new example $z = (x, y)$, then the least-squares prediction of y is

$$\hat{y} = a_{n-1} + b_{n-1} x. \tag{11}$$

We can use the error in this prediction as a nonconformity measure:

$$A(B, z) := |y - \hat{y}| = |y - (a_{n-1} + b_{n-1} x)|.$$

We can obtain other nonconformity measures by using other methods to estimate a line.

Alternatively, we can include the new example as one of the examples used to estimate the least squares line or some other regression line. In this case, it is natural to write (x_n, y_n) for the new example. Then a_n and b_n designate the coefficients calculated from all n examples, and we can use

$$|y_i - (a_n + b_n x_i)| \tag{12}$$

to measure the nonconformity of each of the (x_i, y_i) with the others. In general, the inclusion of the new example simplifies the implementation or at least the explanation of the conformal algorithm. In the case of least squares, it does not change the prediction regions.

4.2 Conformal Prediction from Old Examples Alone

Suppose we have chosen a nonconformity measure A for our problem. Given A, and given the assumption that the z_i are exchangeable, we now define a valid prediction region

$$\gamma^{\varepsilon}(z_1, \ldots, z_{n-1}) \subseteq \mathbf{Z},$$

where $\mathbf{Z}$ is the example space. We do this by giving an algorithm for deciding, for each $z \in \mathbf{Z}$, whether z should be included in the region. For simplicity in stating this algorithm, we provisionally use the symbol z_n for z, as if we were assuming that z_n is in fact equal to z.

The Conformal Algorithm Using Old Examples Alone

Input: Nonconformity measure A, significance level ε, examples $z_1, \ldots, z_{n-1}$, example z,

Task: Decide whether to include z in $\gamma^{\varepsilon}(z_1, \ldots, z_{n-1})$.

Algorithm:

1. Provisionally set $z_n := z$.
2. For $i = 1, \ldots, n$, set $\alpha_i := A(\wr z_1, \ldots, z_n \int \setminus \wr z_i \int, z_i)$.
3. Set $p_z := \dfrac{\text{number of } i \text{ such that } 1 \le i \le n \text{ and } \alpha_i \ge \alpha_n}{n}$.
4. Include z in $\gamma^{\varepsilon}(z_1, \ldots, z_{n-1})$ if and only if $p_z > \varepsilon$.

If $\mathbf{Z}$ has only a few elements, this algorithm can be implemented in a brute-force way: calculate p_z for every $z \in \mathbf{Z}$. If $\mathbf{Z}$ has many elements, we will need some other way of identifying the z satisfying $p_z > \varepsilon$.

The number p_z is the fraction of the examples in $\wr z_1, \ldots, z_{n-1}, z \int$ that are at least as different from the others as z is, in the sense measured by A. So the algorithm tells us to form a prediction region consisting of the z that are not among the fraction ε most out of place when they are added to the bag of old examples.

The definition of $\gamma^\varepsilon(z_1, \ldots, z_{n-1})$ can be framed as an application of the widely accepted Neyman-Pearson theory for hypothesis testing and confidence intervals (Lehmann, 1986). In the Neyman-Pearson theory, we test a hypothesis H using a random variable T that is likely to be large only if H is false. Once we observe $T = t$, we calculate $p_H := \Pr(T \geq t \,|\, H)$. We reject H at level ε if $p_H \leq \varepsilon$. Because this happens under H with probability no more than ε, we can declare $1-\varepsilon$ confidence that the true hypothesis H is among those not rejected. Our procedure makes these choices of H and T:

- The hypothesis H says the bag of the first n examples is $\wr z_1, \ldots, z_{n-1}, z \int$.
- The test statistic T is the random value of α_n.

Under H—that is, conditional on the bag $\wr z_1, \ldots, z_{n-1}, z \int$, T is equally likely to come out equal to any of the α_i. Its observed value is α_n. So

$$p_H = \Pr(T \geq \alpha_n \,|\, \wr z_1, \ldots, z_{n-1}, z \int) = p_z.$$

Since $z_1, \ldots, z_{n-1}$ are known, rejecting the bag $\wr z_1, \ldots, z_{n-1}, z \int$ means rejecting $z_n = z$. So our $1-\varepsilon$ confidence is in the set of z for which $p_z > \varepsilon$.

The regions $\gamma^\varepsilon(z_1, \ldots, z_{n-1})$ for successive n are based on overlapping sequences of examples rather than independent samples. But the successive errors are ε-rare n-events. The event that our nth prediction is an error, $z_n \notin \gamma^\varepsilon(z_1, \ldots, z_{n-1})$, is the event $p_{z_n} \leq \varepsilon$. This is an n-event, because the value of p_{z_n} is determined by z_n and the bag $\wr z_1, \ldots, z_{n-1} \int$. It is ε-rare because it is the event that α_n is among a fraction ε or fewer of the α_i that are strictly larger than all the other α_i, and this can have probability at most ε when the α_i are exchangeable. So it follows from Informal Proposition 1 (§3.4) that we can expect at least $1-\varepsilon$ of the $\gamma^\varepsilon(z_1, \ldots, z_{n-1})$, $n = 1, \ldots, N$, to be correct.

4.2.1 EXAMPLE: PREDICTING A NUMBER WITH AN AVERAGE

In §2.1, we discussed Fisher's 95% prediction interval for z_n based on $z_1, \ldots, z_{n-1}$, which is valid under the assumption that the z_i are independent and normally distributed. We used it to predict z_{20} when the first 19 z_i are

$$17, 20, 10, 17, 12, 15, 19, 22, 17, 19, 14, 22, 18, 17, 13, 12, 18, 15, 17.$$

Taking into account our knowledge that the z_i are all integers, we arrived at the 95% prediction that z_{20} is an integer between 10 to 24, inclusive.

What can we predict about z_{20} at the 95% level if we drop the assumption of normality and assume only exchangeability? To produce a 95% prediction interval valid under the exchangeability assumption alone, we reason as follows. To decide whether to include a particular value z in the interval, we consider twenty numbers that depend on z:

- First, the deviation of z from the average of it and the other 19 numbers. Because the sum of the 19 is 314, this is
$$\left| \frac{314+z}{20} - z \right| = \frac{1}{20} \left| 314 - 19z \right|. \tag{13}$$

- Then, for $i = 1, \ldots, 19$, the deviation of z_i from this same average. This is

$$\left| \frac{314 + z}{20} - z_i \right| = \frac{1}{20} \left| 314 + z - 20 z_i \right| . \tag{14}$$

Under the hypothesis that z is the actual value of z_n, these 20 numbers are exchangeable. Each of them is as likely as the other to be the largest. So there is at least a 95% (19 in 20) chance that (13) will not exceed the largest of the 19 numbers in (14). The largest of the 19 z_is being 22 and the smallest 10, we can write this condition as

$$|314 - 19z| \leq \max \left\{ |314 + z - (20 \times 22)| , |314 + z - (20 \times 10)| \right\} ,$$

which reduces to

$$10 \leq z \leq \frac{214}{9} \approx 23.8.$$

Taking into account that z_{20} is an integer, our 95% prediction is that it will be an integer between 10 and 23, inclusive. This is nearly the same prediction we obtained by Fisher's method. We have lost nothing by weakening the assumption that the z_i are independent and normally distributed to the assumption that they are exchangeable. But we are still basing our prediction region on the average of old examples, which is an optimal estimator in various respects under the assumption of normality.

4.2.2 ARE WE COMPLICATING THE STORY UNNECESSARILY?

The reader may feel that we are vacillating about whether to include the new example in the bag with which we are comparing it. In our statement of the conformal algorithm, we define the nonconformity scores by

$$\alpha_i := A(\wr z_1, \ldots, z_n \int \setminus \wr z_i \int, z_i), \tag{15}$$

apparently signaling that we do not want to include z_i in the bag to which it is compared. But then we use the nonconformity measure

$$A(B, z) := |(\text{average of } z \text{ and all the examples in } B) - z| ,$$

which seems to put z back in the bag, reducing (15) to

$$\alpha_i = \left| \frac{\sum_{j=1}^{n} z_j}{n} - z_i \right| .$$

We could have reached this point more easily by writing

$$\alpha_i := A(\wr z_1, \ldots, z_n \int, z_i) \tag{16}$$

in the conformal algorithm and using $A(B, z) := |\bar{z}_B - z|$.

The two ways of defining nonconformity scores, (15) and (16), are equivalent, inasmuch as whatever we can get with one of them we can get from the other by changing the nonconformity measure. In this case, (16) might be more convenient. But we will see other cases where (15) is more convenient. We also have another reason for using (15). It is the form that generalizes, as we will see in §5, to on-line compression models.

4.3 Conformal Prediction Using a New Object

Now we turn to the case where our example space $\mathbf{Z}$ is of the form $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$. We call $\mathbf{X}$ the *object space*, $\mathbf{Y}$ the *label space*. We observe in sequence examples $z_1, \ldots, z_N$, where $z_i = (x_i, y_i)$. At the point where we have observed

$$(z_1, \ldots, z_{n-1}, x_n) = ((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), x_n),$$

we want to predict y_n by giving a prediction region

$$\Gamma^\varepsilon(z_1, \ldots, z_{n-1}, x_n) \subseteq \mathbf{Y}$$

that is valid at the $(1-\varepsilon)$ level. As in the special case where the x_i are absent, we start with a nonconformity measure $A(B, z)$.

We define the prediction region by giving an algorithm for deciding, for each $y \in \mathbf{Y}$, whether y should be included in the region. For simplicity in stating this algorithm, we provisionally use the symbol z_n for (x_n, y), as if we were assuming that y_n is in fact equal to y.

The Conformal Algorithm

Input: Nonconformity measure A, significance level ε, examples $z_1, \ldots, z_{n-1}$, object x_n, label y

Task: Decide whether to include y in $\Gamma^\varepsilon(z_1, \ldots, z_{n-1}, x_n)$.

Algorithm:

1. Provisionally set $z_n := (x_n, y)$.
2. For $i = 1, \ldots, n$, set $\alpha_i := A(\wr z_1, \ldots, z_n \int \setminus \wr z_i \int, z_i)$.
3. Set $p_y := \dfrac{\#\{i = 1, \ldots, n \mid \alpha_i \geq \alpha_n\}}{n}$.
4. Include y in $\Gamma^\varepsilon(z_1, \ldots, z_{n-1}, x_n)$ if and only if $p_y > \varepsilon$.

This differs only slightly from the conformal algorithm using old examples alone (p. 385). Now we write p_y instead of p_z, and we say that we are including y in $\Gamma^\varepsilon(z_1, \ldots, z_{n-1}, x_n)$ instead of saying that we are including z in $\gamma^\varepsilon(z_1, \ldots, z_{n-1})$.

To see that this algorithm produces valid prediction regions, it suffices to see that it consists of the algorithm for old examples alone together with a further step that does not change the frequency of hits. We know that the region the old algorithm produces,

$$\gamma^\varepsilon(z_1, \ldots, z_{n-1}) \subseteq \mathbf{Z}, \tag{17}$$

contains the new example $z_n = (x_n, y_n)$ at least 95% of the time. Once we know x_n, we can rule out all $z = (x, y)$ in (17) with $x \neq x_n$. The y not ruled out, those such that (x_n, y) is in (17), are precisely those in the set

$$\Gamma^\varepsilon(z_1, \ldots, z_{n-1}, x_n) \subseteq \mathbf{Y} \tag{18}$$

produced by our new algorithm. Having (x_n, y_n) in (17) $1-\varepsilon$ of the time is equivalent to having y_n in (18) $1-\varepsilon$ of the time.

4.3.1 EXAMPLE: CLASSIFYING IRIS FLOWERS

In 1936, R. A. Fisher used discriminant analysis to distinguish different species of iris on the basis of measurements of their flowers. The data he used included measurements by Edgar Anderson of flowers from 50 plants each of two species, *iris setosa* and *iris versicolor*. Two of the measurements, sepal length and petal width, are plotted in Figure 4.

To illustrate how the conformal algorithm can be used for classification, we have randomly chosen 25 of the 100 plants. The sepal lengths and species for the first 24 of them are listed in Table 2 and plotted in Figure 5. The 25th plant in the sample has sepal length 6.8. On the basis of this information, would you classify it as *setosa* or *versicolor*, and how confident would you be in the classification? Because 6.8 is the longest sepal in the sample, nearly any reasonable method will classify the plant as *versicolor*, and this is in fact the correct answer. But the appropriate level of confidence is not so obvious.

We calculate conformal prediction regions using three different nonconformity measures: one based on distance to the nearest neighbors, one based on distance to the species average, and one based on a support-vector machine. Because our evidence is relatively weak, we do not achieve the high precision with high confidence that can be achieved in many applications of machine learning (see, for example, §4.5). But we get a clear view of the details of the calculations and the interpretation of the results.

Distance to the nearest neighbor belonging to each species. Here we use the nonconformity measure (10). The fourth and fifth columns of Table 2 (labeled NN for nearest neighbor) give nonconformity scores α_i obtained with $y_{25} = \mathrm{s}$ and $y_{25} = \mathrm{v}$, respectively. In both cases, these scores are given by

$$\begin{aligned}\alpha_i &= A(\wr z_1, \ldots, z_{25} \int \setminus \wr z_i \int, z_i) \\ &= \frac{\min\{|x_j - x_i| : 1 \le j \le 25 \;\&\; j \ne i \;\&\; y_j = y_i\}}{\min\{|x_j - x_i| : 1 \le j \le 25 \;\&\; j \ne i \;\&\; y_j \ne y_i\}},\end{aligned} \tag{19}$$

but for the fourth column $z_{25} = (6.8, \mathrm{s})$, while for the fifth column $z_{25} = (6.8, \mathrm{v})$.

If both the numerator and the denominator in (19) are equal to zero, we take the ratio also to be zero. This happens in the case of the first plant, for example. It has the same sepal length, 5.0, as the 7th and 13th plants, which are *setosa*, and the 15th plant, which is *versicolor*.

Step 3 of the conformal algorithm yields $p_{\mathrm{s}} = 0.08$ and $p_{\mathrm{v}} = 0.32$. Step 4 tells us that

- s is in the $1-\varepsilon$ prediction region when $1-\varepsilon > 0.92$, and
- v is in the $1-\varepsilon$ prediction region when $1-\varepsilon > 0.68$.

Here are prediction regions for a few levels of ε.

- $\Gamma^{0.08} = \{\mathrm{v}\}$. With 92% confidence, we predict that $y_{25} = \mathrm{v}$.
- $\Gamma^{0.05} = \{\mathrm{s}, \mathrm{v}\}$. If we raise the confidence with which we want to predict y_{25} to 95%, the prediction is completely uninformative.
- $\Gamma^{1/3} = \emptyset$. If we lower the confidence to $2/3$, we get a prediction we know is false: y_{25} will be in the empty set.

	Data		Nonconformity scores					
			NN		Species Average		SVM	
	sepal length	species	α_i for $y_{25} = \mathrm{s}$	α_i for $y_{25} = \mathrm{v}$	α_i for $y_{25} = \mathrm{s}$	α_i for $y_{25} = \mathrm{v}$	α_i for $y_{25} = \mathrm{s}$	α_i for $y_{25} = \mathrm{v}$
z_1	5.0	s	0	0	0.06	0.06	0	0
z_2	4.4	s	0	0	0.66	0.54	0	0
z_3	4.9	s	1	1	0.16	0.04	0	0
z_4	4.4	s	0	0	0.66	0.54	0	0
z_5	5.1	s	0	0	0.04	0.16	0	0
z_6	5.9	v	0.25	0.25	0.12	0.20	0	0
z_7	5.0	s	0	0	0.06	0.06	0	0
z_8	6.4	v	0.50	0.22	0.38	0.30	0	0
z_9	6.7	v	0	0	0.68	0.60	0	0
z_{10}	6.2	v	0.33	0.29	0.18	0.10	0	0
z_{11}	5.1	s	0	0	0.04	0.16	0	0
z_{12}	4.6	s	0	0	0.46	0.34	0	0
z_{13}	5.0	s	0	0	0.06	0.06	0	0
z_{14}	5.4	s	0	0	0.34	0.46	0	0
z_{15}	5.0	v	∞	∞	1.02	1.10	∞	∞
z_{16}	6.7	v	0	0	0.68	0.60	0	0
z_{17}	5.8	v	0	0	0.22	0.30	0	0
z_{18}	5.5	s	0.50	0.50	0.44	0.56	0	0
z_{19}	5.8	v	0	0	0.22	0.30	0	0
z_{20}	5.4	s	0	0	0.34	0.46	0	0
z_{21}	5.1	s	0	0	0.04	0.16	0	0
z_{22}	5.7	v	0.50	0.50	0.32	0.40	0	0
z_{23}	4.6	s	0	0	0.46	0.34	0	0
z_{24}	4.6	s	0	0	0.46	0.34	0	0
z_{25}	6.8	s	13		1.74		∞	
z_{25}	6.8	v		0.077		0.7		0
p_{s}			0.08		0.04		0.08	
p_{v}				0.32		0.08		1

Table 2: **Conformal prediction of iris species from sepal length, using three different nonconformity measures.** The data used are sepal length and species for a random sample of 25 of the 100 plants measured by Edgar Anderson. The second column gives x_i, the sepal length. The third column gives y_i, the species. The 25th plant has sepal length $x_{25} = 6.8$, and our task is to predict its species y_{25}. For each nonconformity measure, we calculate nonconformity scores under each hypothesis, $y_{25} = \mathrm{s}$ and $y_{25} = \mathrm{v}$. The p-value in each column is computed from the 25 nonconformity scores in that column; it is the fraction of them equal to or larger than the 25th. The results from the three nonconformity measures are consistent, inasmuch as the p-value for v is always larger than the p-value for s.

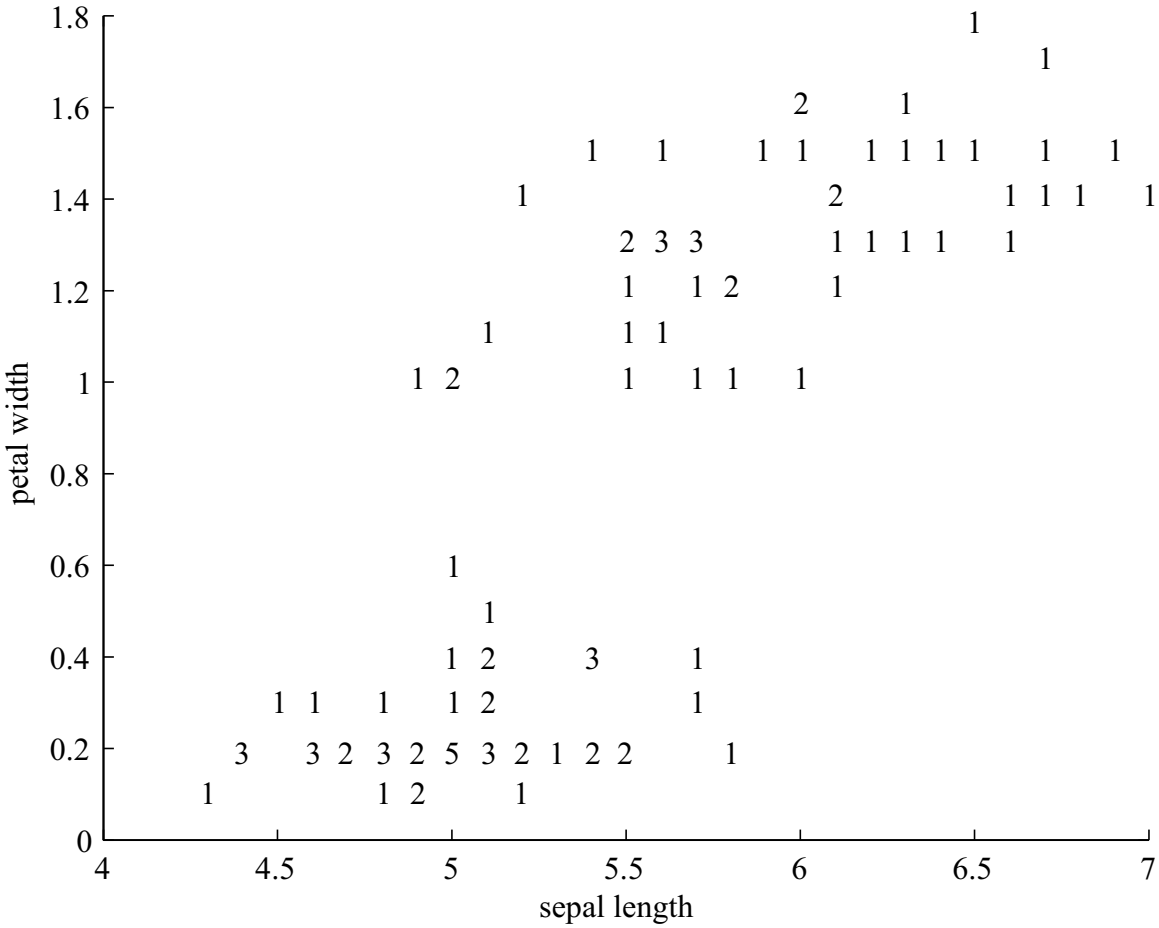


Figure 4: **Sepal length, petal width, and species for Edgar Anderson's** 100 **flowers.** The 50 *iris setosa* are clustered at the lower left, while the 50 *iris versicolor* are clustered at the upper right. The numbers indicate how many plants have exactly the same measurement; for example, there are 5 plants that have sepals 5 inches long and petals 0.2 inches wide. Petal width separates the two species perfectly; all 50 *versicolor* petals are 1 inch wide or wider, while all *setosa* petals are narrower than 1 inch. But there is substantial overlap in sepal length.

In fact, $y_{25} = \mathrm{v}$. Our 92% prediction is correct.

The fact that we are making a known-to-be-false prediction with $2/3$ confidence is a signal that the 25th sepal length, 6.8, is unusual for either species. A close look at the nonconformity scores reveals that it is being perceived as unusual simply because $2/3$ of the plants have other plants in the sample with exactly the same sepal length, whereas there is no other plant with the sepal length 6.8.

In classification problems, it is natural to report the greatest $1-\varepsilon$ for which Γ^{ε} is a single label. In our example, this produces the statement that we are 92% confident that y_{25} is v. But in order

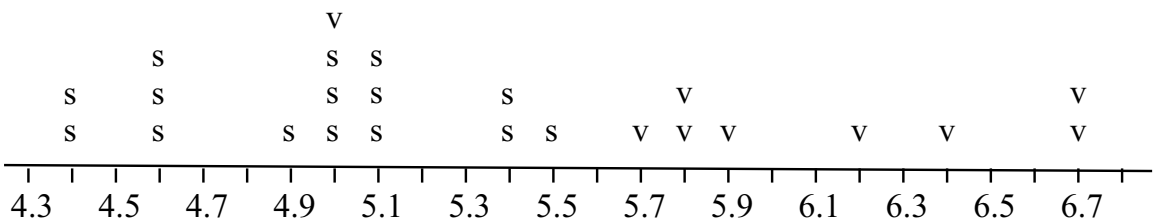


Figure 5: **Sepal length and species for the first** 24 **plants in our random sample of size** 25**.** Except for one *versicolor* with sepal length 5.0, the *versicolor* in this sample all have longer sepals than the *setosa*. This high degree of separation is an accident of the sampling.

to avoid overconfidence when the object x_n is unusual, it is wise to report also the largest ε for which Γ^ε is empty. We call this the *credibility* of the prediction (Vovk et al., 2005, p. 96).[3] In our example, the prediction that y_{25} will be v has credibility of only 32%, indicating that the example is somewhat unusual for the method that produces the prediction—so unusual that the method has 68% confidence in a prediction of y_{25} that we know is false before we observe y_{25} ($\Gamma^{0.68} = \emptyset$).

Distance to the average of each species. The nearest-neighbor nonconformity measure, because it considers only nearby sepal lengths, does not take full advantage of the fact that a *versicolor* flower typically has longer sepals than a *setosa* flower. We can expect to obtain a more efficient conformal predictor (one that produces smaller regions for a given level of confidence) if we use a nonconformity measure that takes account of average sepal length for the two species.

We use the nonconformity measure A defined by

$$A(B,(x,y)) = |\bar{x}_{B\cup\wr(x,y)\int,y} - x|, \tag{20}$$

where $\bar{x}_{B,y}$ denotes the average sepal length of all plants of species y in the bag B, and $B \cup \wr z \int$ denotes the bag obtained by adding z to B. To test $y_{25} = \mathrm{s}$, we consider the bag consisting of the 24 old examples together with $(6.8, \mathrm{s})$, and we calculate the average sepal lengths for the two species in this bag: 5.06 for *setosa* and 6.02 for *versicolor*. Then we use (20) to calculate the nonconformity scores shown in the sixth column of Table 2:

$$\alpha_i = \begin{cases} |5.06 - x_i| & \text{if } y_i = \mathrm{s} \\ |6.02 - x_i| & \text{if } y_i = \mathrm{v} \end{cases}$$

for $i = 1, \ldots, 25$, where we take y_{25} to be s. To test $y_{25} = \mathrm{v}$, we consider the bag consisting of the 24 old examples together with $(6.8, \mathrm{v})$, and we calculate the average sepal lengths for the two species in this bag: 4.94 for *setosa* and 6.1 for *versicolor*. Then we use (20) to calculate the nonconformity scores shown in the seventh column of Table 2:

$$\alpha_i = \begin{cases} |4.94 - x_i| & \text{if } y_i = \mathrm{s} \\ |6.1 - x_i| & \text{if } y_i = \mathrm{v} \end{cases}$$

for $i = 1, \ldots, 25$, where we take y_{25} to be v.

We obtain $p_\mathrm{s} = 0.04$ and $p_\mathrm{v} = 0.08$, so that

3. This notion of credibility is one of the novelties of the theory of conformal prediction. It is not found in the prior literature on confidence and prediction regions.

- s is in the $1-\varepsilon$ prediction region when $1-\varepsilon > 0.96$, and
- v is in the $1-\varepsilon$ prediction region when $1-\varepsilon > 0.92$.

Here are the prediction regions for some different levels of ε.

- $\Gamma^{0.04} = \{\text{v}\}$. With 96% confidence, we predict that $y_{25} = \text{v}$.
- $\Gamma^{0.03} = \{\text{s},\text{v}\}$. If we raise the confidence with which we want to predict y_{25} to 97%, the prediction is completely uninformative.
- $\Gamma^{0.08} = \emptyset$. If we lower the confidence to 92%, we get a prediction we know is false: y_{25} will be in the empty set.

In this case, we predict $y_{25} = \text{v}$ with confidence 96% but credibility only 8%. The credibility is lower with this nonconformity measure because it perceives 6.8 as being even more unusual than the nearest-neighbor measure did. It is unusually far from the average sepal lengths for both species.

A support-vector machine. As Vladimir Vapnik explains on pp. 408–410 of his *Statistical Learning Theory* (1998), support-vector machines grew out of the idea of separating two groups of examples with a hyperplane in a way that makes as few mistakes as possible—that is, puts as few examples as possible on the wrong side. This idea springs to mind when we look at Figure 5. In this one-dimensional picture, a hyperplane is a point. We are tempted to separate the *setosa* from the *versicolor* with a point between 5.5 and 5.7.

Vapnik proposed to separate two groups not with a single hyperplane but with a band: two hyperplanes with few or no examples between them that separate the two groups as well as possible. Examples on the wrong side of both hyperplanes would be considered very strange; those between the hyperplanes would also be considered strange but less so. In our one-dimensional example, the obvious separating band is the interval from 5.5 to 5.7. The only strange example is the *versicolor* with sepal length 5.0.

Here is one way of making Vapnik's idea into an algorithm for calculating nonconformity scores for all the examples in a bag $\wr(x_1,y_1),\ldots(x_n,y_n)\int$. First plot all the examples as in Figure 5. Then find numbers a and b such that $a \le b$ and the interval $[a,b]$ separates the two groups with the fewest mistakes—that is, minimizes[4]

$$\#\{i \,|\, 1 \le i \le n, x_i < b, \text{ and } y_i = \text{v}\} + \#\{i \,|\, 1 \le i \le n, x_i > a, \text{ and } y_i = \text{s}\}.$$

There may be many intervals that minimize this count; choose one that is widest. Then give the ith example the score

$$\alpha_i = \begin{cases} \infty & \text{if } y_i = \text{v and } x_i < a & \text{or } y_i = \text{s and } b < x_i \\ 1 & \text{if } y_i = \text{v and } a \le x_i < b & \text{or } y_i = \text{s and } a < x_i \le b \\ 0 & \text{if } y_i = \text{v and } b \le x_i & \text{or } y_i = \text{s and } x_i \le a. \end{cases}$$

When applied to the bags in Figure 6, this algorithm gives the circled examples the score ∞ and all the others the score 0. These scores are listed in the last two columns of Table 2.

4. Here we are implicitly assuming that the *setosa* flowers will be on the left, with shorter sepal lengths. A general algorithm should also check the possibility of a separation with the *versicolor* flowers on the left.

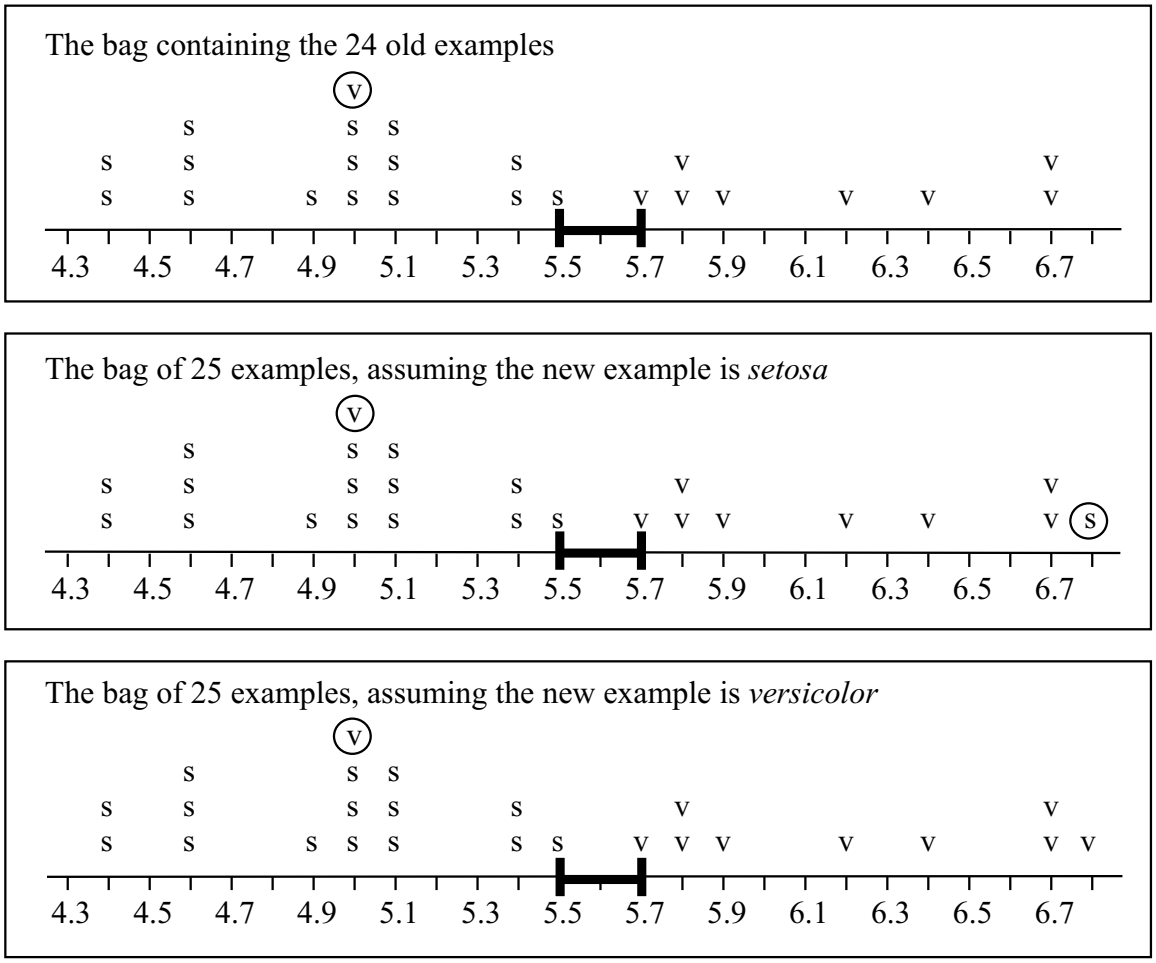


Figure 6: **Separation for three bags.** In each case, the separating band is the interval $[5.5, 5.7]$. Examples on the wrong side of the interval are considered strange and are circled.

As we see from the table, the resulting p-values are $p_{\mathrm{s}} = 0.08$ and $p_{\mathrm{v}} = 1$. So this time we obtain 92% confidence in $y_{25} = \mathrm{v}$, with 100% credibility.

The algorithm just described is too complex to implement when there are thousands of examples. For this reason, Vapnik and his collaborators proposed instead a quadratic minimization that balances the width of the separating band against the number and size of the mistakes it makes. Support-vector machines of this type have been widely used. They usually solve the dual optimization problem, and the Lagrange multipliers they calculate can serve as nonconformity scores. Implementations sometimes fail to treat the old examples symmetrically because they make various uses of the order in which examples are presented, but this difficulty can be overcome by a preliminary randomization (Vovk et al., 2005, p. 58).

A systematic comparison. The random sample of 25 plants we have considered is odd in two ways: (1) except for the one *versicolor* with sepal length of only 5.0, the two species do not overlap

	NN	Species average	SVM
singleton hits	164	441	195
uncertain	795	477	762
total hits	959	918	957
empty	9	49	1
singleton errors	32	33	42
total errors	41	82	43
total examples	1000	1000	1000
% hits	96%	92%	96%
total singletons	196	474	237
% hits	84%	93%	82%
total errors	41	82	43
% empty	22%	60%	2%

Table 3: **Performance of** 92% **prediction regions based on three nonconformity measures.** For each nonconformity measure, we have found $1,000$ prediction regions at the 92% level, using each time a different random sample of 25 from Anderson's 100 flowers. The "uncertain" regions are those equal to the whole label space, $\mathbf{Y} = \{\mathrm{s}, \mathrm{v}\}$.

in sepal length, and (2) the flower whose species we are trying to predict has a sepal that is unusually long for either species.

In order to get a fuller picture of how the three nonconformity measures perform in general on the iris data, we have applied each of them to $1,000$ different samples of size 25 selected from the population of Anderson's 100 plants. The results are shown in Table 3.

The 92% regions based on the species average were correct about 92% of the time (918 times out of 1000), as advertised. The regions based on the other two measures were correct more often, about 96% of the time. The reason for this difference is visible in Table 2; the nonconformity scores based on the species average take a greater variety of values and therefore produce ties less often. The regions based on the species averages are also more efficient (smaller); 441 of its hits were informative, as opposed to fewer than 200 for each of the other two nonconformity measures. This efficiency also shows up in more empty regions among the errors. The species average produced an empty 92% prediction region for the random sample used in Table 2, and Table 3 shows that this happens 5% of the time.

As a practical matter, the uncertain prediction regions ($\Gamma^{0.08} = \{\mathrm{s}, \mathrm{v}\}$) and the empty ones ($\Gamma^{0.08} = \emptyset$) are equally uninformative. The only errors that mislead are the singletons that are wrong, and the three methods all produce these at about the same rate—3 or 4%.

4.3.2 EXAMPLE: PREDICTING PETAL WIDTH FROM SEPAL LENGTH

We now turn to the use of the conformal algorithm to predict a number. We use the same 25 plants, but now we use the data in the second and third columns of Table 4: the sepal length and petal width for the first 24 plants, and the sepal length for the 25th. Our task is to predict the petal width for the 25th.

	sepal length	petal width	Nearest neighbor	Linear regression
z_1	5.0	0.3	0.3	$\lvert 0.003y_{25} - 0.149\rvert$
z_2	4.4	0.2	0	$\lvert 0.069y_{25} + 0.050\rvert$
z_3	4.9	0.2	0.25	$\lvert 0.014y_{25} - 0.199\rvert$
z_4	4.4	0.2	0	$\lvert 0.069y_{25} + 0.050\rvert$
z_5	5.1	0.4	0.15	$\lvert 0.008y_{25} + 0.099\rvert$
z_6	5.9	1.5	0.3	$\lvert 0.096y_{25} - 0.603\rvert$
z_7	5.0	0.2	0.4	$\lvert 0.003y_{25} - 0.249\rvert$
z_8	6.4	1.3	0.2	$\lvert 0.151y_{25} - 0.154\rvert$
z_9	6.7	1.4	0.3	$\lvert 0.184y_{25} - 0.104\rvert$
z_{10}	6.2	1.5	0.2	$\lvert 0.129y_{25} - 0.453\rvert$
z_{11}	5.1	0.2	0.15	$\lvert 0.008y_{25} + 0.299\rvert$
z_{12}	4.6	0.2	0.05	$\lvert 0.047y_{25} - 0.050\rvert$
z_{13}	5.0	0.6	0.3	$\lvert 0.003y_{25} + 0.151\rvert$
z_{14}	5.4	0.4	0	$\lvert 0.041y_{25} + 0.248\rvert$
z_{15}	5.0	1.0	0.75	$\lvert 0.003y_{25} + 0.551\rvert$
z_{16}	6.7	1.7	0.3	$\lvert 0.184y_{25} - 0.404\rvert$
z_{17}	5.8	1.2	0.2	$\lvert 0.085y_{25} - 0.353\rvert$
z_{18}	5.5	0.2	0.2	$\lvert 0.052y_{25} + 0.498\rvert$
z_{19}	5.8	1.0	0.2	$\lvert 0.085y_{25} - 0.153\rvert$
z_{20}	5.4	0.4	0	$\lvert 0.041y_{25} + 0.248\rvert$
z_{21}	5.1	0.3	0	$\lvert 0.008y_{25} + 0.199\rvert$
z_{22}	5.7	1.3	0.2	$\lvert 0.074y_{25} - 0.502\rvert$
z_{23}	4.6	0.3	0.1	$\lvert 0.047y_{25} + 0.050\rvert$
z_{24}	4.6	0.2	0.05	$\lvert 0.047y_{25} - 0.050\rvert$
z_{25}	6.8	y_{25}	$\lvert y_{25} - 1.55\rvert$	$\lvert 0.805y_{25} - 1.345\rvert$

Table 4: **Conformal prediction of petal width from sepal length.** We use the same random 25 plants that we used for predicting the species. The actual value of y_{25} is 1.4.

The most conventional way of analyzing this data is to calculate the least-squares line (11):

$$\hat{y} = a_{24} + b_{24}x = -2.96 + 0.68x.$$

The sepal length for the 25th plant being $x_{25} = 6.8$, the line predicts that y_{25} should be near $-2.96 + 0.68 \times 6.8 = 1.66$. Under the textbook assumption that the y_i are all independent and normally distributed with means on the line and a common variance, we estimate the common variance by

$$s_{24}^2 = \frac{\sum_{i=1}^{24}(y_i - (a_{24} + b_{24}x_i))^2}{22} = 0.0780.$$

The textbook $1-\varepsilon$ interval for y_{25} based on $\wr(x_1,y_1),\ldots,(x_{24},y_{24})\int$ and x_{25} is

$$1.66 \pm t_{22}^{\varepsilon/2} s_{24}\sqrt{1 + \frac{1}{24} + \frac{(x_{25} - \bar{x}_{24})^2}{\sum_{j=1}^{24}(x_j - \bar{x}_{24})^2}} = 1.66 \pm 0.311 t_{22}^{\varepsilon/2} \tag{21}$$

(Draper and Smith 1998, p. 82; Ryan 1997, pp. 21–22; Seber and Lee 2003, p. 145). Taking into account the fact y_{25} is measured to only one decimal place, we obtain $[1.0, 2.3]$ for the 96% interval and $[1.1, 2.2]$ for the 92% interval.

The prediction interval (21) is analogous to Fisher's interval for a new example from the same normally distributed population as a bag of old examples (§2.1.1). In §5.3.2 we will review the general model of which both are special cases.

As we will now see, the conformal algorithm under exchangeability gives confidence intervals comparable to (21), without the assumption that the errors are normal. We use two different nonconformity measures: one based on the nearest neighbor, and one based on the least-squares line.

Conformal prediction using the nearest neighbor. Suppose B is a bag of old examples and (x, y) is a new example, for which we know the sepal length x but not the petal width y. We can predict y using the nearest neighbor in an obvious way: We find the $z' \in B$ for which the sepal length x' is closest to x, and we predict that y will be the same as the petal width y'. If there are several examples in the bag with sepal length equally close to x, then we take the median of their petal widths as our predictor $\hat{y}$. The associated nonconformity measure is $|y - \hat{y}|$.

The fourth column of Table 4 gives the nonconformity scores for our sample using this nonconformity measure. We see that $\alpha_{25} = |y_{25} - 1.55|$. The other nonconformity scores do not involve y_{25}; the largest is 0.75, and the second largest is 0.40. So we obtain these prediction regions y_{25}:

- The 96% prediction region consists of all the y for which $p_y > 0.04$, which requires that at least one of the other α_i be as large as α_{25}, or that $0.75 \geq |y - 1.55|$. This is the interval $[0.8, 2.3]$.

- The 92% prediction region consists of all the y for which $p_y > 0.08$, which requires that at least two of the other α_i be as large as α_{25}, or that $0.40 \geq |y - 1.55|$. This is the interval $[1.2, 1.9]$.

Conformal prediction using least-squares. Now we use the least-squares nonconformity measure with inclusion, given by (12). In our case, $n = 25$, so our nonconformity scores are

$$\begin{aligned}\alpha_i &= |y_i - (a_{25} + b_{25} x_i)| \\ &= \left| y_i - \frac{\sum_{j=1}^{25} y_j}{25} - \frac{\sum_{j=1}^{25} (x_j - \bar{x}_{25}) y_j}{\sum_{j=1}^{25} (x_j - \bar{x}_{25})^2} \left(x_i - \frac{\sum_{j=1}^{25} x_j}{25} \right) \right| .\end{aligned}$$

When we substitute values of $\sum_{j=1}^{24} y_j$, $\sum_{j=1}^{24} (x_j - \bar{x}_{25}) y_j$, $\sum_{j=1}^{25} (x_j - \bar{x}_{25})^2$, and $\sum_{j=1}^{25} x_j$ calculated from Table 4, this becomes

$$\alpha_i = |y_i + (0.553 - 0.110 x_i) y_{25} - 0.498 x_i + 2.04| .$$

For $i = 1, \ldots, 24$, we can further evaluate this by substituting the values of x_i and y_i. For $i = 25$, we can substitute 6.8 for x_{25}. These substitutions produce the expressions of the form $|c_i y_{25} + d_i|$ listed in the last column of Table 4. We have made sure that c_i is always positive by multiplying by -1 within the absolute value when need be.

Table 5 shows calculations required to find the conformal prediction region. The task is to identify, for $i = 1, \ldots, 24$, the y for which $|c_i y + d_i| \geq |0.805 y - 1.345|$. We first find the solutions of

the equation $|c_i y + d_i| = |0.805y - 1.345|$, which are

$$-\frac{d_i + 1.345}{c_i - 0.805} \qquad \text{and} \qquad -\frac{d_i - 1.345}{c_i + 0.805}.$$

As it happens, $c_i < 0.805$ for $i = 1, \ldots, 24$, and in this case the y satisfying $|c_i y + d_i| \geq |0.805 - 1.345|$ form the interval between these two points. This interval is shown in the last column of the table.

In order to be in the 96% interval, y must be in at least one of the 24 intervals in the table; in order to be in the 92% interval, it must be in at least two of them. So the 96% interval is $[1.0, 2.4]$, and the 92% interval is $[1.0, 2.3]$.

An algorithm for finding conformal prediction intervals using a least-squares or ridge-regression nonconformity measure with an object space of any finite dimension is spelled out on pp. 32–33 of our book (Vovk et al., 2005).

4.4 Optimality

The predictions produced by the conformal algorithm are invariant with respect to the old examples, correct with the advertised probability, and nested. As we now show, they are optimal among all region predictors with these properties.

Here is a more precise statement of the three properties:

1. *The predictions are invariant with respect to the ordering of the old examples.* Formally, this means that the predictor γ is a function of two variables, the significance level ε and the bag B of old examples. We write $\gamma^\varepsilon(B)$ for the prediction, which is a subset of the example space $\mathbf{Z}$.

2. *The probability of a hit is always at least the advertised confidence level.* For every positive integer n and every probability distribution under which $z_1, \ldots, z_n$ are exchangeable,

$$\Pr\{z_n \in \gamma^\varepsilon(\wr z_1, \ldots, z_{n-1} \int)\} \geq 1 - \varepsilon.$$

3. *The prediction regions are nested.* If $\varepsilon_1 \geq \varepsilon_2$, then $\gamma^{\varepsilon_1}(B) \subseteq \gamma^{\varepsilon_2}(B)$.

Conformal predictors satisfy these three conditions. Other region predictors can also satisfy them. But as we now demonstrate, any γ satisfying them can be improved on by a conformal predictor: there always exists a nonconformity measure A such that the predictor γ_A constructed from A by the conformal algorithm satisfies $\gamma_A^\varepsilon(B) \subseteq \gamma^\varepsilon(B)$ for all B and ε.

The key to the demonstration is the following lemma:

Lemma 1 *Suppose γ is a region predictor satisfying the three conditions, $\wr a_1, \ldots, a_n \int$ is a bag of examples, and $0 < \varepsilon \leq 1$. Then $n\varepsilon$ or fewer of the n elements of the bag satisfy*

$$a_i \notin \gamma^\varepsilon(\wr a_1, \ldots, a_n \int \setminus \wr a_i \int). \tag{22}$$

Proof Consider the unique exchangeable probability distribution for $z_1, \ldots, z_n$ that gives probability 1 to $\wr z_1, \ldots, z_n \int = \wr a_1, \ldots, a_n \int$. Under this distribution, each element of $\wr a_1, \ldots, a_n \int$ has an equal probability of being z_n, and in this case, (22) is a mistake. By the second condition, the probability of a mistake is ε or less. So the fraction of the bag's elements for which (22) holds is ε or less. ■

	$\alpha_i = \lvert c_i y_{25} + d_i \rvert$	$-\frac{d_i + 1.345}{c_i - 0.805}$	$-\frac{d_i - 1.345}{c_i + 0.805}$	y satisfying $\lvert c_i y + d_i \rvert \geq \lvert 0.805 - 1.345 \rvert$
z_1	$\lvert 0.003y_{25} - 0.149 \rvert$	1.49	1.85	[1.49,1.85]
z_2	$\lvert 0.069y_{25} + 0.050 \rvert$	1.90	1.48	[1.48,1.90]
z_3	$\lvert 0.014y_{25} - 0.199 \rvert$	1.45	1.89	[1.45,1.89]
z_4	$\lvert 0.069y_{25} + 0.050 \rvert$	1.90	1.48	[1.48,1.90]
z_5	$\lvert 0.008y_{25} + 0.099 \rvert$	1.81	1.53	[1.53,1.81]
z_6	$\lvert 0.096y_{25} - 0.603 \rvert$	1.05	2.16	[1.05,2.16]
z_7	$\lvert 0.003y_{25} - 0.249 \rvert$	1.37	1.97	[1.37,1.97]
z_8	$\lvert 0.151y_{25} - 0.154 \rvert$	1.82	1.57	[1.57,1.82]
z_9	$\lvert 0.184y_{25} - 0.104 \rvert$	2.00	1.47	[1.47,2.00]
z_{10}	$\lvert 0.129y_{25} - 0.453 \rvert$	1.32	1.93	[1.32,1.93]
z_{11}	$\lvert 0.008y_{25} + 0.299 \rvert$	2.06	1.29	[1.29,2.06]
z_{12}	$\lvert 0.047y_{25} - 0.050 \rvert$	1.71	1.64	[1.64,1.71]
z_{13}	$\lvert 0.003y_{25} + 0.151 \rvert$	1.87	1.48	[1.48,1.87]
z_{14}	$\lvert 0.041y_{25} + 0.248 \rvert$	2.09	1.30	[1.30,2.09]
z_{15}	$\lvert 0.003y_{25} + 0.551 \rvert$	2.36	0.98	[0.98,2.36]
z_{16}	$\lvert 0.184y_{25} - 0.404 \rvert$	1.52	1.77	[1.52,1.77]
z_{17}	$\lvert 0.085y_{25} - 0.353 \rvert$	1.38	1.91	[1.38,1.91]
z_{18}	$\lvert 0.052y_{25} + 0.498 \rvert$	2.45	0.99	[0.99,2.45]
z_{19}	$\lvert 0.085y_{25} - 0.153 \rvert$	1.66	1.68	[1.66,1.68]
z_{20}	$\lvert 0.041y_{25} + 0.248 \rvert$	2.09	1.30	[1.30,2.09]
z_{21}	$\lvert 0.008y_{25} + 0.199 \rvert$	1.94	1.41	[1.41,1.94]
z_{22}	$\lvert 0.074y_{25} - 0.502 \rvert$	1.15	2.10	[1.15,2.10]
z_{23}	$\lvert 0.047y_{25} + 0.050 \rvert$	1.84	1.52	[1.52,1.84]
z_{24}	$\lvert 0.047y_{25} - 0.050 \rvert$	1.71	1.64	[1.64,1.71]
z_{25}	$\lvert 0.805y_{25} - 1.345 \rvert$			

Table 5: **Calculations with least-squares nonconformity scores.** The column on the right gives the values of y for which the example's nonconformity score will exceed that of the 25th example.

Given the region predictor γ, what nonconformity measure will give us a conformal predictor that improves on it? If

$$z \notin \gamma^{\delta}(B), \tag{23}$$

then γ is asserting confidence $1-\delta$ that z should not appear next because it is so different from B. So the largest $1-\delta$ for which (23) holds is a natural nonconformity measure:

$$A(B,z) = \sup\{1-\delta \,|\, z \notin \gamma^{\delta}(B)\}.$$

	Least-squares prediction with normal errors	Conformal prediction with two different nonconformity measures	
		NN	Least squares
96%	$[1.0, 2.3]$	$[0.8, 2.3]$	$[1.0, 2.4]$
92%	$[1.1, 2.2]$	$[1.2, 1.9]$	$[1.0, 2.3]$

Table 6: **Prediction intervals for the 25th plant's petal width, calculated by three different methods.** The conformal prediction intervals using the least-squares nonconformity measure are quite close to the standard intervals based on least-squares with normal errors. All the intervals contain the actual value, 1.4.

The conformal predictor γ_A obtained from this nonconformity measure, though it agrees with γ on how to rank different z with respect to their nonconformity with B, may produce tighter prediction regions if γ is too conservative in the levels of confidence it asserts.

To show that $\gamma_A^\varepsilon(B) \subseteq \gamma^\varepsilon(B)$ for every ε and every B, we assume that

$$z \in \gamma_A^\varepsilon(\wr z_1, \ldots, z_{n-1} \int) \tag{24}$$

and show that $z \in \gamma^\varepsilon(\wr z_1, \ldots, z_{n-1} \int)$. According to the conformal algorithm, (24) means that when we provisionally set z_n equal to z and calculate the nonconformity scores

$$\alpha_i = \sup\{1 - \delta \,|\, z_i \notin \gamma^\delta(\wr z_1, \ldots, z_n \int \setminus \wr z_i \int)\}$$

for $i = 1, \ldots, n$, we find that strictly more than $n\varepsilon$ of these scores are greater than or equal to α_n. Because γ's prediction regions are nested (condition 3), it follows that if $z_n \notin \gamma^\varepsilon(\wr z_1, \ldots, z_{n-1} \int)$, then $z_i \notin \gamma^\varepsilon(\wr z_1, \ldots, z_n \int \setminus \wr z_i \int)$ for strictly more than $n\varepsilon$ of the z_i. But by Lemma 1, $n\varepsilon$ or fewer of the z_i can satisfy this condition. So $z_n \in \gamma^\varepsilon(\wr z_1, \ldots, z_{n-1} \int)$.

There are sensible reasons to use region predictors that are not invariant. We may want to exploit possible departures from exchangeability even while insisting on validity under exchangeability. Or it may simply be more practical to use a predictor that is not invariant. But invariance is a natural condition when we want to rely only on exchangeability, and in this case our optimality result is persuasive. For further discussion, see §2.4 of our book (Vovk et al., 2005).

4.5 Examples Are Seldom Exactly Exchangeable

Although the assumption of exchangeability is weak compared to the assumptions embodied in most statistical models, it is still an idealization, seldom matched exactly by what we see in the world. So we should not expect conclusions derived from this assumption to be exactly true. In particular, we should not be surprised if a 95% conformal predictor is wrong more than 5% of the time.

We can make this point with the USPS data set so often used to illustrate machine learning methods. This data set consists of 9298 examples of the form (x, y), where x is a 16×16 gray-scale matrix and y is one of the ten digits $0, 1, \ldots, 9$. It has been used in hundreds of books and articles. In our book (Vovk et al., 2005), it is used to illustrate conformal prediction with a number of different nonconformity measures. It is well known that the examples in this data set are not perfectly exchangeable. In particular, the first 7291 examples, which are often treated as a training

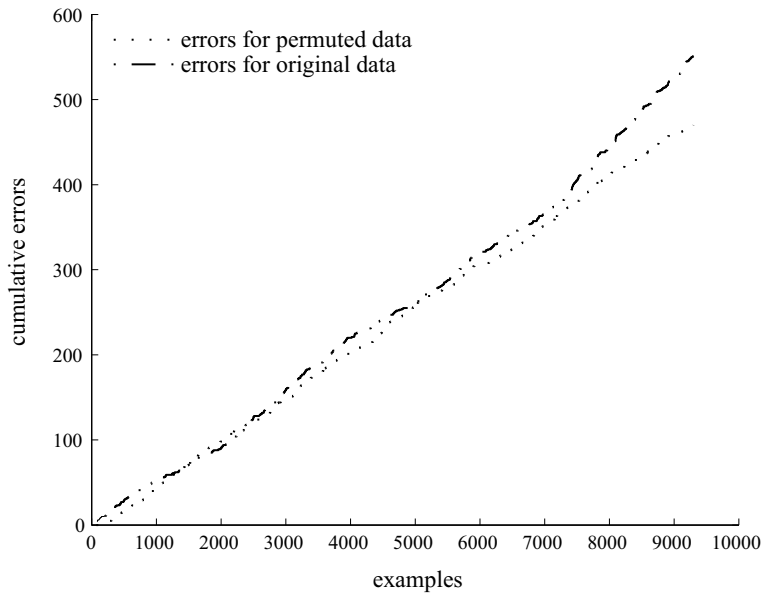


Figure 7: **Errors in** 95% **nearest-neighbor conformal prediction on the classical USPS data set.** When the 9298 examples are predicted in a randomly chosen order, so that the exchangeability assumption is satisfied for sure, the error rate is approximately 5% as advertised. When they are taken in their original order, first the 7291 in the training set, and then the 2007 in the test set, the error rate is higher, especially in the test set.

set, are systematically different in some respects from the remaining 2007 examples, which are usually treated as a test set.

Figure 7 illustrates how the non-exchangeability of the USPS data affects conformal prediction. The figure records the performance of the 95% conformal predictor using the nearest-neighbor nonconformity measure (10), applied to the USPS data in two ways. First we use the 9298 examples in the order in which they are given in the data set. (We ignore the distinction between training and test examples, but since the training examples are given first we do go through them first.) Working through the examples in this order, we predict each y_n using the previous examples and x_n. Second, we randomly permute all 9298 examples, thus producing an order with respect to which the examples are necessarily exchangeable. The law of large numbers works when we go through the examples in the permuted order: we make mistakes at a steady rate, about equal to the expected 5%. But when we go through the examples in the original order, the fraction of mistakes is less stable, and it worsens as we move into the test set. As Table 7 shows, the fraction of mistakes is 5%, as desired, in the first 7291 examples (the training set) but jumps to 8% in the last 2007 examples.

Non-exchangeability can be tested statistically, using conventional or game-theoretic methods (Vovk et al., 2005, §7.1). In the case of this data, any reasonable test will reject exchangeability decisively. Whether the deviation from exchangeability is of practical importance for prediction depends, of course, on circumstances. An error rate of 8% when 5% has been promised may or may not be acceptable.

	Original data			Permuted data		
	Training	Test	Total	Training	Test	Total
singleton hits	6798	1838	8636	6800	1905	8705
uncertain hits	111	0	111	123	0	123
total hits	6909	1838	8747	6923	1905	8828
empty	265	142	407	205	81	286
singleton errors	102	27	129	160	21	181
uncertain errors	15	0	15	3	0	3
total errors	382	169	551	368	102	470
total examples	7291	2007	9298	7291	2007	9298
% hits	95%	92%	94%	95%	95%	95%
total singletons	6900	1865	8765	6960	1926	8880
% hits	99%	99%	99%	98%	99%	98%
total uncertain	126	0	126	126	0	126
% hits	82%		82%	98%		98%
total errors	382	169	551	368	102	470
% empty	69%	85%	74%	57%	79%	61%

Table 7: **Details of the performance of** 95% **nearest-neighbor conformal prediction on the classical USPS data set.** Because there are 10 labels, the uncertain predictions, those containing more than one label, can be hits or errors.

5. On-Line Compression Models

In this section, we generalize conformal prediction from the exchangeability model to a whole class of models, which we call on-line compression models.

In the exchangeability model, we compress or summarize examples by omitting information about their order. We then look backwards from the summary (the bag of unordered examples) and give probabilities for the different orderings that could have produced it. The compression can be done on-line: each time we see a new example, we add it to the bag. The backward-looking probabilities can also be given step by step. Other on-line compression models compress more or less drastically but have a similar structure.

On-line compression models were studied in the 1970s and 1980s, under various names, by Per Martin-Löf (1974), Steffen Lauritzen (1988), and Eugene Asarin (1987; 1988). Different authors had different motivations. Lauritzen and Martin-Löf started from statistical mechanics, whereas Asarin started from Kolmogorov's thinking about the meaning of randomness. But the models they studied all summarize past examples using statistics that contain all the information useful for predicting future examples. The summary is updated each time one observes a new example, and the probabilistic content of the structure is expressed by Markov kernels that give probabilities for summarized examples conditional on the summaries.

In general, a Markov kernel is a mapping that specifies, as a function of one variable, a probability distribution for some other variable or variables. A Markov kernel for w given u, for example, gives a probability distribution for w for each value of u. It is conventional to write $P(w \mid u)$ for this

distribution. We are interested in Markov kernels of the form $P(z_1,\ldots,z_n \mid \sigma_n)$, where σ_n summarizes the examples $z_1,\ldots,z_n$. Such a kernel gives probabilities for the different $z_1,\ldots,z_n$ that could have produced σ_n.

Martin-Löf, Lauritzen, and Asarin were interested in justifying widely used statistical models from principles that seem less arbitrary than the models themselves. On-line compression models offer an opportunity to do this, because they typically limit their use of probability to representing ignorance with a uniform distribution but lead to statistical models that seem to say something more. Suppose, for example, that Joe summarizes numbers $z_1,\ldots,z_n$ by

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad \text{and} \qquad r^2 = \sum_{i=1}^{n}(z_i-\bar{z})^2$$

and gives these summaries to Bill, who does not know $z_1,\ldots,z_n$. Bill might adopt a probability distribution for $z_1,\ldots,z_n$ that is uniform over the possibilities, which form the $(n-1)$-dimensional sphere of radius r centered around $(\bar{z},\ldots,\bar{z})$. As we will see in §5.3.2, this is an on-line compression model. It was shown, by Freedman and Smith (see Vovk et al., 2005, p. 217) and then by Lauritzen (1988, pp. 238–247), that if we assume this model is valid for all n, then the distribution of $z_1,z_2,\ldots$ must be a mixture of distributions under which $z_1,z_2,\ldots$ are independent and normal with a common mean and variance. This is analogous to de Finetti's theorem, which says that if $z_1,\ldots,z_n$ are exchangeable for all n, then the distribution of $z_1,z_2,\ldots$ must be a mixture of distributions under which $z_1,z_2,\ldots$ are independent and identically distributed.

For our own part, we are interested in using an on-line compression model directly for prediction rather than as a step towards a model that specifies probabilities for examples more fully. We have already seen how the exchangeability model can be used directly for prediction: we establish a law of large numbers for backward-looking probabilities (§3.4), and we use it to justify confidence in conformal prediction regions (§4.2). The argument extends to on-line compression models in general.

For the exchangeability model, conformal prediction is optimal for obtaining prediction regions (§4.4). No such statement can be made for on-line compression models in general. In fact, there are other on-line compression models in which conformal prediction is very inefficient (Vovk et al., 2005, p. 220).

After developing the general theory of conformal prediction for on-line compression models (§5.1 and §5.2), we consider two examples: the exchangeability-within-label model (§5.3.1) and on-line Gaussian linear model (§5.3.2).

5.1 Definitions

A more formal look at the exchangeability model will suffice to bring the general notion of an on-line compression model into focus.

In the exchangeability model, we summarize examples simply by omitting information about their ordering; the ordered examples are summarized by a bag containing them. The backward-looking probabilities are equally simple; given the bag, the different possible orderings all have equal probability, as if the ordering resulted from drawing the examples successively at random from the bag without replacement. Although this picture is very simple, we can distinguish four distinct mathematical operations within it:

1. *Summarizing.* The examples $z_1,\ldots,z_n$ are summarized by the bag $\wr z_1,\ldots,z_n\int$. We can say that the summarization is accomplished by a *summarizing function* Σ_n that maps an n-tuple of examples $(z_1,\ldots,z_n)$ to the bag containing these examples:

$$\Sigma_n(z_1,\ldots,z_n) := \wr z_1,\ldots,z_n\int.$$

   We write σ_n for the summary—that is, the bag $\wr z_1,\ldots,z_n\int$.

2. *Updating.* The summary can be formed step by step as the examples are observed. Once you have the bag containing the first $n-1$ examples, you just add the nth. This defines an *updating function* $U_n(\sigma,z)$ that satisfies

$$\Sigma_n(z_1,\ldots,z_n) = U_n(\Sigma_{n-1}(z_1,\ldots,z_{n-1}),z_n).$$

   The top panel in Figure 8 depicts how the summary σ_n is built up step by step from $z_1,\ldots,z_n$ using the updating functions $U_1,\ldots,U_n$. First $\sigma_1 = U_1(\Box,z_1)$, where $\Box$ is the empty bag. Then $\sigma_2 = U_2(\sigma_1,z_2)$, and so on.

3. *Looking back all the way.* Given the bag σ_n, the $n!$ different orderings of the elements of the bag are equally likely, just as they would be if we ordered the contents of the bag randomly. As we learned in §3.2, we can say this with a formula that takes explicit account of the possibility of repetitions in the bag: the probability of the event $\{z_1 = a_1,\ldots,z_n = a_n\}$ is

$$P_n(a_1,\ldots,a_n \mid \sigma_n) = \begin{cases} \frac{n_1!\cdots n_k!}{n!} & \text{if } \wr a_1,\ldots,a_n\int = \sigma_n \\ 0 & \text{if } \wr a_1,\ldots,a_n\int \neq \sigma_n, \end{cases}$$

   where k is the number of distinct elements in σ_n, and $n_1,\ldots,n_k$ are the numbers of times these distinct elements occur. We call $P_1,P_2,\ldots$ the *full kernels*.

4. *Looking back one step.* We can also look back one step. Given the bag σ_n, what are the probabilities for z_n and σ_{n-1}? They are the same as if we drew z_n out of σ_n at random. In other words, for each z that appears in σ_n, there is a probability k/n, where k is the number of times z appears in σ_n, that (1) $z_n = z$ and (2) σ_{n-1} is the bag obtained by removing one instance of z from σ_n. The kernel defined in this way is represented by the two arrows backward from σ_n in the bottom panel of Figure 8. Let us designate it by R_n. We similarly obtain a kernel R_{n-1} backward from σ_{n-1} and so on. These are the *one-step kernels* for the model. We can obtain the full kernel P_n by combining the one-step kernels $R_n,R_{n-1},\ldots,R_1$. This is most readily understood not in terms of formulas but in terms of a sequence of drawings whose outcomes have the probability distributions given by the kernels. The drawing from σ_n (which goes by the probabilities given by $R_n(\cdot \mid \sigma_n)$) gives us z_n and σ_{n-1}, the drawing from σ_{n-1} (which goes by the probabilities given by $R_{n-1}(\cdot \mid \sigma_{n-1})$) gives us z_{n-1} and σ_{n-2}, and so on; we finally obtain the whole random sequence $z_1,\ldots,z_n$, which has the distribution $P_n(\cdot \mid \sigma_n)$. This is the meaning of the bottom panel in Figure 8.

All four operations are important. The second and fourth, updating and looking back one step, can be thought of as the most fundamental, because we can derive the other two from them. Summarization can be carried out by composing updates, and looking back all the way can be carried out by composing one-step look-backs. Moreover, the conformal algorithm uses the one-step back

$$\begin{tikzcd}
 & z_1 \arrow[d] & z_2 \arrow[d] & & z_{n-1} \arrow[d] & z_n \arrow[d] \\
\Box \arrow[r] & \sigma_1 \arrow[r] & \sigma_2 \arrow[r] & \cdots \arrow[r] & \sigma_{n-1} \arrow[r] & \sigma_n
\end{tikzcd}$$

Updating. We speak of "on-line" compression models because the summary can be updated with each new example. In the case of the exchangeability model, we obtain the bag σ_i by adding the new example z_i to the old bag σ_{i-1}.

$$\begin{tikzcd}
 & z_1 & z_2 & & z_{n-1} & z_n \\
\Box & \sigma_1 \arrow[l] \arrow[u] & \sigma_2 \arrow[l] \arrow[u] & \cdots \arrow[l] & \sigma_{n-1} \arrow[l] \arrow[u] & \sigma_n \arrow[l] \arrow[u]
\end{tikzcd}$$

Backward probabilities. The two arrows backwards from σ_i symbolize our probabilities, conditional on σ_i, for what example z_i and what previous summary σ_{i-1} were combined to produce σ_i. Like the diagram in Figure 3 that it generalizes, this diagram is a Bayes net.

Figure 8: **Elements of an on-line compression model.** The top diagram represents the updating functions $U_1, \ldots, U_n$. The bottom diagram represents the one-step kernels $R_1, \ldots, R_n$.

probabilities. But when we turn to particular on-line compression models, we will find it initially most convenient to describe them in terms of their summarizing functions and full kernels.

In general, an *on-line compression model* for an example space $\mathbf{Z}$ consists of a space $\mathbf{S}$, whose elements we call *summaries*, and two sequences of mappings:

- Updating functions $U_1, U_2, \ldots$. The function U_n maps a summary s and an example z to a new summary $U_n(s,z)$.

- One-step kernels $R_1, R_2, \ldots$. For each summary s, the kernel R_n gives a joint probability distribution $R_n(s', z \mid s)$ for an unknown summary s' and unknown example z. We require that $R_n(\cdot \mid s)$ give probability one to the set of pairs (s', z) such that $U_n(s', z) = s$.

We also require that the summary space $\mathbf{S}$ include the empty summary $\Box$.

The recipes for constructing the summarizing functions $\Sigma_1, \Sigma_2, \ldots$ and the full kernels $P_1, P_2, \ldots$ are the same in general as in the exchangeability model:

- The summary $\sigma_n = \Sigma_n(z_1, \ldots, z_n)$ is built up step by step from $z_1, \ldots, z_n$ using the updating functions. First $\sigma_1 = U_1(\Box, z_1)$, then $\sigma_2 = U_2(\sigma_1, z_2)$, and so on.

- We obtain the full kernel P_n by combining, backwards from σ_n, the random experiments represented by the one-step kernels $R_n, R_{n-1}, \ldots, R_1$. First we draw z_n and σ_{n-1} from $R_n(\cdot \mid \sigma_n)$, then we draw z_{n-1} and σ_{n-2} from $R_{n-1}(\cdot \mid \sigma_{n-1})$, and so on. The sequence $z_1, \ldots, z_n$ obtained in this way has the distribution $P_n(\cdot \mid \sigma_n)$.

On-line compression models are usually initially specified in terms of their summarizing functions Σ_n and their full kernels P_n, because these are usually easy to describe. One must then verify that these easily described objects do define an on-line compression model. This requires verifying two points:

1. $\Sigma_1, \Sigma_2, \ldots$ can be defined successively by means of updating functions:

$$\Sigma_n(z_1, \ldots, z_n) = U_n(\Sigma_{n-1}(z_1, \ldots, z_{n-1}), z_n).$$

   In words: σ_n depends on $z_1, \ldots, z_{n-1}$ only through the earlier summary σ_{n-1}.

2. Each P_n can be obtained as required using one-step kernels. One way to verify this is to exhibit the one-step kernels $R_1, \ldots, R_n$ and then to check that drawing z_n and σ_{n-1} from $R_n(\cdot \mid \sigma_n)$, then drawing z_{n-1} and σ_{n-2} from $R_{n-1}(\cdot \mid \sigma_{n-1})$, and so on produces a sequence $z_1, \ldots, z_n$ with the distribution $P_n(\cdot \mid \sigma_n)$. Another way to verify it, without necessarily exhibiting the one-step kernels, is to verify the conditional independence relations represented by Figure 8: z_n (and hence also σ_n) is probabilistically independent of $z_1, \ldots, z_{n-1}$ given σ_{n-1}.

5.2 Conformal Prediction

In the context of an on-line compression model, a *nonconformity measure* is an arbitrary real-valued function $A(\sigma, z)$, where σ is a summary and z is an example. We choose A so that $A(\sigma, z)$ is large when z seems very different from the examples that might be summarized by σ.

In order to state the conformal algorithm, we write $\tilde{\sigma}_{n-1}$ and $\tilde{z}_n$ for random variables with a joint probability distribution given by the one-step kernel $R(\cdot \mid \sigma_n)$. The algorithm using old examples alone can then be stated as follows:

The Conformal Algorithm Using Old Examples Alone

Input: Nonconformity measure A, significance level ε, examples $z_1, \ldots, z_{n-1}$, example z

Task: Decide whether to include z in $\gamma^\varepsilon(z_1, \ldots, z_{n-1})$.

Algorithm:

1. Provisionally set $z_n := z$.
2. Set $p_z := R_n(A(\tilde{\sigma}_{n-1}, \tilde{z}_n) \geq A(\sigma_{n-1}, z_n) \mid \sigma_n)$.
3. Include z in $\gamma^\varepsilon(z_1, \ldots, z_{n-1})$ if and only if $p_z > \varepsilon$.

To see that this reduces to the algorithm we gave for the exchangeability model on p. 385, recall that $\sigma_n = \wr z, \ldots, z_n \int$ and $\tilde{\sigma}_{n-1} = \wr z_1, \ldots, z_n \int \setminus \wr \tilde{z}_n \int$ in that model, so that

$$A(\tilde{\sigma}_{n-1}, \tilde{z}_n) = A(\wr z_1, \ldots, z_n \int \setminus \wr \tilde{z}_n \int, \tilde{z}_n) \tag{25}$$

and

$$A(\sigma_{n-1}, z_n) = A(\wr z_1, \ldots, z_{n-1} \int, z_n). \tag{26}$$

Under $R_n(\cdot \mid \wr z_1, \ldots, z_n \int)$, the random variable $\tilde{z}_n$ has equal chances of being any of the z_i, so that the probability of (25) being greater than or equal to (26) is simply the fraction of the z_i for which

$$A(\wr z_1, \ldots, z_n \int \setminus \wr z_i \int, z_i) \geq A(\wr z_1, \ldots, z_{n-1} \int, z_n),$$

and this is how p_z is defined on p. 385.

Our arguments for the validity of the regions $\gamma^\varepsilon(z_1, \ldots, z_{n-1})$ in the exchangeability model generalize readily. The definitions of n-event and ε-rare generalize in an obvious way:

- An event E is an *n-event* if its happening or failing is determined by the value of z_n and the value of the summary σ_{n-1}.
- An n-event E is ε-*rare* if $R_n(E \mid \sigma_n) \leq \varepsilon$.

The event $z_n \notin \gamma^\varepsilon(z_1, \ldots, z_{n-1})$ is an n-event, and it is ε-rare (the probability is ε or less that a random variable will take a value that it equals or exceeds with a probability of ε or less). So working backwards from the summary σ_N for a large value of N, Bill can still bet against the errors successively at rates corresponding to their probabilities under σ_n, which are always ε or less. This produces an exact analog to Informal Proposition 1:

Informal Proposition 2 *Suppose N is large, and the variables $z_1, \ldots, z_N$ obey an on-line compression model. Suppose E_n is an ε-rare n-event for $n = 1, \ldots, N$. Then the law of large numbers applies; with very high probability, no more than approximately the fraction ε of the events $E_1, \ldots, E_N$ will happen.*

The conformal algorithm using features of the new example generalizes similarly:

The Conformal Algorithm

Input: Nonconformity measure A, significance level ε, examples $z_1, \ldots, z_{n-1}$, object x_n, label y

Task: Decide whether to include y in $\Gamma^\varepsilon(z_1, \ldots, z_{n-1}, x_n)$.

Algorithm:

1. Provisionally set $z_n := (x_n, y)$.
2. Set $p_y := R_n(A(\tilde{\sigma}_{n-1}, \tilde{z}_n) \geq A(\sigma_{n-1}, z_n) \mid \sigma_n)$.
3. Include y in $\Gamma^\varepsilon(z_1, \ldots, z_{n-1}, x_n)$ if and only if $p_y > \varepsilon$.

The validity of this algorithm follows from the validity of the algorithm using old examples alone by the same argument as in the case of exchangeability.

5.3 Examples

We now look at two on-line compression models: the exchangeability-within-label model and the on-line Gaussian linear model.

The exchangeability-within-label model was first introduced in work leading up to our book (Vovk et al., 2005). It weakens the assumption of exchangeability.

The on-line Gaussian linear model, as we have already mentioned, has been widely studied. It overlaps the exchangeability model, in the sense that the assumptions for both of the models can

hold at the same time, but the assumptions for one of them can hold without the assumptions for the other holding. It is closely related to the classical Gaussian linear model. Conformal prediction in the on-line model leads to the same prediction regions that are usually used for the classical model. But the conformal prediction theory adds the new information that these intervals are valid in the sense of this article: they are right $1-\varepsilon$ of the time when used on accumulating data.

5.3.1 THE EXCHANGEABILITY-WITHIN-LABEL MODEL

The assumption of exchangeability can be weakened in many ways. In the case of classification, one interesting possibility is to assume only that the examples for each label are exchangeable with each other. For each label, the objects with that label are as likely to appear in one order as in another. This assumption leaves open the possibility that the appearance of one label might change the probabilities for the next label.

Suppose the label space has k elements, say $\mathbf{Y} = \{1,\ldots,k\}$. Then we define the *exchangeability-within-label model* as follows:

Summarizing Functions The nth summarizing function is

$$\Sigma_n(z_1,\ldots,z_n) := (y_1,\ldots,y_n,B_1^n,\ldots,B_k^n)\,,$$

where B_j^n is the bag consisting of the objects in the list $x_1,\ldots,x_n$ that have the label j.

Full Kernels The full kernel $P_n(z_1,\ldots,z_n \,|\, y_1,\ldots,y_n,B_1^n,\ldots,B_k^n)$ is most easily described in terms the random action for which it gives the probabilities: independently for each label j, distribute the objects in B_j^n randomly among the positions i for which y_i is equal to j.

To check that this is an on-line compression model, we exhibit the updating function and the one-step kernels:

Updating When (x_n,y_n) is observed, the summary

$$(y_1,\ldots,y_{n-1},B_1^{n-1},\ldots,B_k^{n-1})$$

is updated by inserting y_n after y_{n-1} and adding x_n to $B_{y_n}^{n-1}$.

One step back The one-step kernel R_n is given by

$$R_n(\text{summary},(x,y)\,|\,y_1,\ldots,y_n,B_1^n,\ldots,B_k^n) = \begin{cases} \frac{k}{|B_{y_n}^n|} & \text{if } y = y_n \\ 0 & \text{otherwise,} \end{cases}$$

where k is the number of xs in $B_{y_n}^n$. This is the same as the probability the one-step kernel for the exchangeability model would give for x on the basis of a bag of size $|B_{y_n}^n|$ that includes k xs.

Because the true labels are part of the summary, our imaginary bettor Bill can choose to bet just on those rounds of his game with Joe where the label has a particular value, and this implies that a 95% conformal predictor under the exchangeability-within-label model will make errors at no more than a 5% rate for examples with that label. This is not necessarily true for a 95% conformal predictor under the exchangeability model; although it can make errors no more than about 5% of

the time overall, its error rate may be higher for some labels and lower for others. As Figure 9 shows, this happens in the case of the USPS data set. The graph in the top panel of the figure shows the cumulative errors for examples with the label 5, which is particularly easy to confuse with other digits, when the nearest-neighbor conformal predictor is applied to that data in permuted form. The error rate for 5 is over 11%. The graph in the bottom panel shows the results of the exchangeability-within-label conformal predictor using the same nearest-neighbor nonconformity measure; here the error rate stays close to 5%. As this graph makes clear, the predictor holds the error rate down to 5% in this case by producing many prediction regions containing more than one label ("uncertain predictions").

As we explain in §4.5 and §8.4 of our book (Vovk et al., 2005), the exchangeability-within-label model is a *Mondrian model*. In general, a Mondrian model decomposes the space $\mathbf{Z} \times \mathbb{N}$, where $\mathbb{N}$ is set of the natural numbers, into non-overlapping rectangles, and it asks for exchangeability only within these rectangles. For each example z_i, it then records, as part of the summary, the rectangle into which (z_i, i) falls. Mondrian models can be useful when we need to weaken the assumption of exchangeability. They can also be attractive even if we are willing to assume exchangeability across the categories, because the conformal predictions they produce will be calibrated within categories.

5.3.2 THE ON-LINE GAUSSIAN LINEAR MODEL

Consider examples $z_1, \ldots, z_N$, of the form $z_n = (x_n, y_n)$, where y_n is a number and x_n is a row vector consisting of p numbers. For each n between 1 and N, set

$$X_n := \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \text{and} \qquad Y_n := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Thus X_n is an $n \times p$ matrix, and Y_n is a column vector of length n.

In this context, the *on-line Gaussian linear model* is the on-line compression model defined by the following summarizing functions and full kernels:

Summarizing Functions The nth summarizing function is

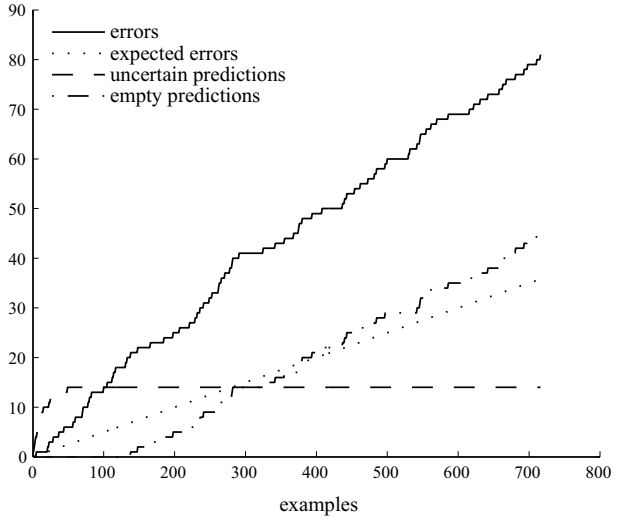
$$\begin{aligned} \Sigma_n(z_1, \ldots, z_n) &:= \left(x_1, \ldots, x_n, \sum_{i=1}^n y_i x_i, \sum_{i=1}^n y_i^2 \right) \\ &= \left(X_n, X_n' Y_n, Y_n' Y_n \right). \end{aligned}$$

Full Kernels The full kernel $P_n(z_1, \ldots, z_n \mid \sigma_n)$ distributes its probability uniformly over the set of vectors $(y_1, \ldots, y_n)$ consistent with the summary σ_n. (We consider probabilities only for $y_1, \ldots, y_n$, because $x_1, \ldots, x_n$ are fixed by σ_n.)

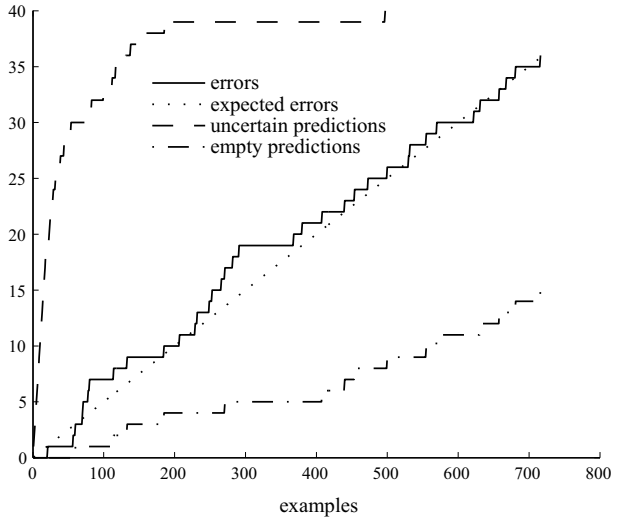
We can write $\sigma_n = (X_n, C, r^2)$, where C is a column vector of length p, and r is a nonnegative number. A vector $(y_1, \ldots, y_n)$ is consistent with σ_n if

$$\sum_{j=1}^n y_j x_j = C \qquad \text{and} \qquad \sum_{j=1}^n y_j^2 = r^2.$$

This is the intersection of a hyperplane with a sphere. Not being empty, the intersection is either a point (in the exceptional case where the hyperplane is tangent to the sphere) or a lower-dimensional



Exchangeability model



Exchangeability-within-label model

Figure 9: **Errors for** 95% **conformal prediction using nearest neighbors in the permuted USPS data when the true label is** 5**.** In both figures, the dotted line represents the overall expected error rate of 5%. The actual error rate for 5s with the exchangeability-within-label model tracks this line, but with the exchangeability model it is much higher. The exchangeability-within-label predictor keeps its error rate down by issuing more prediction regions containing more than one digit ("uncertain predictions").

sphere. (Imagine intersecting a plane and a 2-dimensional sphere; the result is a circle.) The kernel $P_n(\cdot \mid \sigma_n)$ puts all its probability on the point or distributes it uniformly over the the lower-dimensional sphere.

To see that the summarizing functions and full kernels define an on-line compression model, we must check that the summaries can be updated and that the full kernels have the required conditional independence property: conditioning $P_n(\cdot \mid \sigma_n)$ on $z_{i+1}, \ldots, z_n$ gives $P_i(\cdot \mid \sigma_i)$. (We do not condition on σ_i since it can be computed from $z_{i+1}, \ldots, z_n$ and σ_n.) Updating is straightforward; when we observe (x_n, y_n), we update the summary

$$\left(x_1, \ldots, x_{n-1}, \sum_{i=1}^{n-1} y_i x_i, \sum_{i=1}^{n-1} y_i^2 \right)$$

by inserting x_n after x_{n-1} and adding a term to each of the sums. To see that conditioning $P_n(\cdot \mid \sigma_n)$ on $z_{i+1}, \ldots, z_n$ gives $P_i(\cdot \mid \sigma_i)$, we note that conditioning the uniform distribution on a sphere on values $y_{i+1} = a_{i+1}, \ldots, y_n = a_n$ involves intersecting the sphere with the hyperplanes defined by these $n-i$ equations. This produces the uniform distribution on the possibly lower-dimensional sphere defined by

$$\sum_{j=1}^{i} y_j^2 = r^2 - \sum_{j=i+1}^{n} y_j^2 \qquad \text{and} \qquad \sum_{j=1}^{i} y_j x_j = C - \sum_{j=i+1}^{n} y_j x_j;$$

this is indeed $P_i(y_1, \ldots, y_i \mid \sigma_i)$.

The on-line Gaussian linear model is closely related to the *classical Gaussian linear model*. In the classical model,[5]

$$y_i = x_i \beta + e_i, \tag{27}$$

where the x_i are row vectors of known numbers, β is a column vector of unknown numbers (the regression coefficients), and the e_i are independent of each other and normally distributed with mean zero and a common variance. When $n-1 > p$ and $\mathrm{Rank}(X_{n-1}) = p$, the theory of the classical model tells us the following:

- After observing examples $(x_1, y_1, \ldots, x_{n-1}, y_{n-1})$, estimate the vector of coefficients β by

  $$\hat{\beta}_{n-1} := (X'_{n-1} X_{n-1})^{-1} X'_{n-1} Y_{n-1}$$

  and after further observing x_n, predict y_n by

  $$\hat{y}_n := x_n \hat{\beta}_{n-1} = x_n (X'_{n-1} X_{n-1})^{-1} X'_{n-1} Y_{n-1}.$$

- Estimate the variance of the e_i by

  $$s_{n-1}^2 := \frac{Y'_{n-1} Y_{n-1} - \beta'_{n-1} X'_{n-1} Y_{n-1}}{n-p-1}.$$

5. There are many names for the classical model. The name "classical Gaussian linear model" is used by Bickel and Doksum (2001, p. 366).

- The random variable
$$t_n := \frac{y_n - \hat{y}_n}{s_{n-1}\sqrt{1 + x_n'(X_{n-1}'X_{n-1})^{-1}x_n}} \tag{28}$$
has a t-distribution with $n-p-1$ degrees of freedom, and so
$$\hat{y}_n \pm t_{n-p-1}^{\varepsilon/2} s_{n-1}\sqrt{1 + x_n'(X_{n-1}'X_{n-1})^{-1}x_n} \tag{29}$$
has probability $1-\varepsilon$ of containing y_n (Ryan 1997, p. 127; Seber and Lee 2003, p. 132).

The assumption $\mathrm{Rank}(X_{n-1}) = p$ can be relaxed, at the price of complicating the formulas involving $(X_{n-1}'X_{n-1})^{-1}$. But the assumption $n-1 > \mathrm{Rank}(X_{n-1})$ is essential to finding a prediction interval of the type (29); when it fails there are values for the coefficients β such that $y_{n-1} = X_{n-1}\beta$, and consequently there is no residual variance with which to estimate the variance of the e_i.

We have already used two special cases of (29) in this article. Formula (1) in §2.1.1 is the special case with $p = 1$ and each x_i equal to 1, and formula (21) at the beginning of §4.3.2 is the special case with $p = 2$ and the first entry of each x_i equal to 1.

The relation between the classical and on-line models, fully understood in the theoretical literature since the 1980s, can be summarized as follows:

- If $z_1, \ldots, z_N$ satisfy the assumptions of the classical Gaussian linear model, then they satisfy the assumptions of the on-line Gaussian linear model. In other words, the assumption that the errors e_i in (27) are independent and normal with mean zero and a common variance implies that conditional on $X_n'Y_n = C$ and $Y_n'Y_n = r^2$, the vector Y_n is distributed uniformly over the sphere defined by C and r^2. This was already noted by R. A. Fisher in 1925.

- The assumption of the on-line Gaussian linear model, that conditional on $X_n'Y_n = C$ and $Y_n'Y_n = r^2$, the vector Y_n is distributed uniformly over the sphere defined by C and r^2, is sufficient to guarantee that (28) has the t-distribution with $n-p-1$ degrees of freedom (Dempster, 1969; Efron, 1969).

- Suppose $z_1, z_2, \ldots$ is an infinite sequence of random variables. Then $z_1, \ldots, z_N$ satisfy the assumptions of the on-line Gaussian linear model for every integer N if and only if the joint distribution of $z_1, z_2, \ldots$ is a mixture of distributions given by the classical Gaussian linear model, each model in the mixture possibly having a different β and a different variance for the e_i (Lauritzen, 1988).

A natural nonconformity measure A for the on-line Gaussian linear model is given, for $\sigma = (X, X'Y, Y'Y)$ and $z = (x, y)$, by
$$A(\sigma, z) := |y - \hat{y}|, \tag{30}$$
where $\hat{y} = x(X'X)^{-1}X'Y$.

Proposition 2 *When (30) is used as the nonconformity measure, the $1-\varepsilon$ conformal prediction region for y_n is (29), the interval given by the t-distribution in the classical theory.*

Proof When (30) is used as the nonconformity measure, the test statistic $A(\sigma_{n-1}, z_n)$ used in the conformal algorithm becomes $|y_n - \hat{y}_n|$. The conformal algorithm considers the distribution of this

statistic under $R_n(\cdot \mid \sigma_n)$. But when σ_n is fixed and t_n is given by (28), $|t_n|$ is a monotonically increasing function of $|y_n - \hat{y}_n|$ (see Vovk et al., 2005, pp. 202–203, for details). So the conformal prediction region is the interval of values of y_n for which $|t_n|$ does not take its most extreme values. Since t_n has the t-distribution with $n-p-1$ degrees of freedom under $R_n(\cdot \mid \sigma_n)$, this is the interval (29). ■

Together with Informal Proposition 2, Proposition 2 implies that when we use (29) for a large number of successive values of n, y_n will be in the interval $1-\varepsilon$ of the time. In fact, because the probability of error each time is exactly ε, we can say simply that the errors are independent and for this reason the classical law of large numbers applies.

In our example involving the prediction of petal width from sepal length, the exchangeability and Gaussian linear models gave roughly comparable results (see Table 6 in §4.3.2). This will often be the case. Each model makes an assumption, however, that the other does not make. The exchangeability model assumes that the xs, as well as the ys, are exchangeable. The Gaussian linear model assumes that given the xs, the ys are normally distributed.

Acknowledgments

This article is based on a tutorial lecture by Glenn Shafer at the Workshop on Information Theory and Applications at the University of California at San Diego on February 1, 2007. Glenn Shafer would like to thank Yoav Freund and Alon Orlitsky for the invitation to participate in this workshop. We thank Wei Wu for his assistance in computations. Many other intellectual debts are acknowledged in the preface of our book (Vovk et al., 2005).

Appendix A. Validity

The main purpose of this appendix is to formalize and prove the following informal proposition:

Informal Proposition 1 *Suppose N is large, and the variables $z_1, \ldots, z_N$ are exchangeable. Suppose E_n is an ε-rare n-event for $n = 1, \ldots, N$. Then the law of large numbers applies; with very high probability, no more than approximately the fraction ε of the events $E_1, \ldots, E_N$ will happen.*

We used this informal proposition in §3.4 to establish the validity of conformal prediction in the exchangeability model. As we promised then, we will discuss two different approaches to formalizing it: a classical approach and a game-theoretical approach. The classical approach shows that the E_n are mutually independent in the case where they are exactly ε-rare and then appeals to the classical weak law of large numbers for independent events. The game-theoretic approach appeals directly to the more flexible game-theoretic weak law of large numbers.

Our proofs will also establish the analogous Informal Proposition 2, which we used to establish the validity of conformal prediction in on-line compression models in general.

In §A.3, we return to R. A. Fisher's prediction interval for a normal random variable, which we discussed in §2.1.1. We show that this prediction interval's successive hits are independent, so that validity follows from the usual law of large numbers. Fisher's prediction interval is a special case of conformal prediction for the Gaussian linear model, and so it is covered by the general result

for on-line compression models. But the proof in §A.3, being self-contained and elementary and making no reference to conformal prediction, may be especially informative for many readers.

A.1 A Classical Argument for Independence

Recall the definitions we gave in §3.4 in the case where $z_1, \ldots, z_N$ are exchangeable: An event E is an *n-event* if its happening or failing is determined by the value of z_n and the value of the bag $\wr z_1, \ldots, z_{n-1} \int$, and an n-event E is ε-*rare* if $\Pr(E \mid \wr z_1, \ldots, z_n \int) \leq \varepsilon$. Let us further say that n-event E is *exactly* ε-*rare* if

$$\Pr(E \mid \wr z_1, \ldots, z_n \int) = \varepsilon. \tag{31}$$

The conditional probability in this equation is a random variable, depending on the random bag $\wr z_1, \ldots, z_n \int$, but the equation says that it is not really random, for it is always equal to ε. Its expected value, the unconditional probability of E, is therefore also equal to ε.

Proposition 3 *Suppose E_n is an exactly ε-rare n-event for $n = 1, \ldots, N$. Then $E_1, \ldots, E_N$ are mutually independent.*

Proof Consider (31) for $n = N - 1$:

$$\Pr(E_{N-1} \mid \wr z_1, \ldots, z_{N-1} \int) = \varepsilon. \tag{32}$$

Given $\wr z_1, \ldots, z_{N-1} \int$, knowledge of z_N does not change the probabilities for z_{N-1} and $\wr z_1, \ldots, z_{N-2} \int$, and z_{N-1} and $\wr z_1, \ldots, z_{N-2} \int$ determine the $(N-1)$-event E_{N-1}. So adding knowledge of z_N will not change the probability in (32):

$$\Pr(E_{N-1} \mid \wr z_1, \ldots, z_{N-1} \int \ \& \ z_N) = \varepsilon.$$

Because E_N is determined by z_N once $\wr z_1, \ldots, z_{N-1} \int$ is given, it follows that

$$\Pr(E_{N-1} \mid \wr z_1, \ldots, z_{N-1} \int \ \& \ E_N) = \varepsilon,$$

and from this it follows that $\Pr(E_{N-1} \mid E_N) = \varepsilon$. The unconditional probability of E_{N-1} is also ε. So E_N and E_{N-1} are independent. Continuing the reasoning backwards to E_1, we find that the E_n are all mutually independent. ■

This proof generalizes immediately to the general case of on-line compression models (see p. 407); we simply replace $\wr z_1, \ldots, z_n \int$ with σ_n.

If N is sufficiently large, and E_n is an exactly ε-rare n-event for $n = 1, \ldots, N$, then the law of large numbers applies; with very high probability, no more than approximately the fraction ε of the N events will happen. It is intuitively clear that this conclusion will also hold if we have an inequality instead of an equality in (31), because making the E_n even less likely to happen cannot reverse the conclusion that few of them will happen.

The preceding argument is less than rigorous on two counts. First, the proof of Proposition 3 does not consider the existence of the conditional probabilities it uses. Second, the argument from the case where (31) is an equality to that where it is merely an inequality, though entirely convincing, is only intuitive. A fully rigorous proof, which uses Doob's measure-theoretic framework to deal with the conditional probabilities and uses a randomization to bring the inequality up to an equality, is provided on pp. 211–213 of our book (Vovk et al., 2005).

A.2 A Game-theoretic Law of Large Numbers

As we explained in §3.3, the game-theoretic interpretation of exchangeability involves a backward-looking protocol, in which Bill observes first the bag $\wr z_1, \ldots, z_N \int$ and then successively z_N, z_{N-1}, and so on, finally observing z_1. Just before he observes z_n, he knows the bag $\wr z_1, \ldots, z_n \int$ and can bet on the value of z_n at odds corresponding to the probabilities the bag determines:

$$\Pr(z_n = a \mid \wr z_1, \ldots, z_n \int = B) = \frac{k}{n}, \tag{33}$$

where k is the number of times a occurs in B.

THE BACKWARD-LOOKING BETTING PROTOCOL
Players: Joe, Bill
$\mathcal{K}_N := 1$.
Joe announces a bag B_N of size N.
FOR $n = N, N-1, \ldots, 2, 1$
Bill bets on z_n at odds set by (33).
Joe announces $z_n \in B_n$.
$\mathcal{K}_{n-1} := \mathcal{K}_n +$ Bill's net gain.
$B_{n-1} := B_n \setminus \wr z_n \int$.

Bill's initial capital $\mathcal{K}_N$ is 1. His final capital is $\mathcal{K}_0$.

Given an event E, set

$$e := \begin{cases} 1 & \text{if } E \text{ happens} \\ 0 & \text{if } E \text{ fails.} \end{cases}$$

Given events $E_1, \ldots, E_N$, set

$$\mathrm{Freq}_N := \frac{1}{N} \sum_{j=1}^{N} e_j.$$

This is the fraction of the events that happen—the frequency with which they happen. Our game-theoretic law of large numbers will say that if each E_n is an ε-rare n-event, then it is very unlikely that Freq_N will substantially exceed ε.

In game-theoretic probability, what do we mean when we say an event E is "very unlikely"? We mean that the bettor, Bill in this protocol, has a betting strategy that guarantees

$$\mathcal{K}_0 \geq \begin{cases} C & \text{if } E \text{ happens} \\ 0 & \text{if } E \text{ fails,} \end{cases} \tag{34}$$

where C is a large positive number. Cournot's principle, which says that Bill will not multiply his initial unit capital by a large factor without risking bankruptcy, justifies our thinking E unlikely. The larger C, the more unlikely E. We call the quantity

$$\overline{\mathrm{P}}E := \inf \left\{ \frac{1}{C} \;\middle|\; \text{Bill can guarantee (34)} \right\} \tag{35}$$

E's *upper probability*. An unlikely event is one with small upper probability.

Proposition 4 (Game-theoretic weak law of large numbers) *Suppose E_n is an ε-rare n-event, for $n = 1, \ldots, N$. Suppose $\varepsilon < 1/2$, $\delta_1 > 0$, $\delta_2 > 0$, and $N \geq 1/\delta_1\delta_2^2$. Then*

$$\overline{\mathrm{P}}(\mathrm{Freq}_N \geq \varepsilon + \delta_2) \leq \delta_1.$$

In words: If N is sufficiently large, there is a small (less than δ_1) upper probability that the frequency will exceed ε substantially (by more than δ_2).

Readers familiar with game-theoretic probability will recognize Proposition 4 as a form of the game-theoretic weak law of large numbers (Shafer and Vovk, 2001, pp. 124–126). The bound it gives for the upper probability of the event $\mathrm{Freq}_N \geq \varepsilon + \delta_2$ is the same as the bound that Chebyshev's inequality gives for the probability of this event in classical probability theory when the E_n are independent and all have probability ε.

For the benefit of those not familiar with the concepts of game-theoretic probability used in the proof just cited, we now give an elementary and self-contained proof of Proposition 4.

Lemma 5 *Suppose, for $n = 1, \ldots, N$, that E_n is an ε-rare n-event. Then Bill has a strategy that guarantees that his capital $\mathcal{K}_n$ will satisfy*

$$\mathcal{K}_n \geq \frac{n}{N} + \frac{1}{N}\left(\left(\sum_{j=n+1}^{N}(e_j - \varepsilon)\right)^+\right)^2 \tag{36}$$

for $n = 1, \ldots, N$, where $t^+ := \max(t, 0)$.

Proof When $n = N$, (36) reduces to $\mathcal{K}_N \geq 1$, and this certainly holds; Bill's initial capital $\mathcal{K}_N$ is equal to 1. So it suffices to show that if (36) hold for n, then Bill can bet on E_n in such a way that the corresponding inequality for $n-1$,

$$\mathcal{K}_{n-1} \geq \frac{n-1}{N} + \frac{1}{N}\left(\left(\sum_{j=n}^{N}(e_j - \varepsilon)\right)^+\right)^2, \tag{37}$$

also holds. Here is how Bill bets.

- If $\sum_{j=n+1}^{N}(e_j - \varepsilon) \geq \varepsilon$, then Bill buys $(2/N)\sum_{j=n+1}^{N}(e_j - \varepsilon)$ units of e_n. By assumption, he pays no more than ε for each unit. So we have a lower bound on his net gain, $\mathcal{K}_{n-1} - \mathcal{K}_n$:

$$\begin{aligned}
\mathcal{K}_{n-1} - \mathcal{K}_n &\geq \frac{2}{N}\left(\sum_{j=n+1}^{N}(e_j - \varepsilon)\right)(e_n - \varepsilon) \\
&\geq \frac{1}{N}\left(\sum_{j=n}^{N}(e_j - \varepsilon)\right)^2 - \frac{1}{N}\left(\sum_{j=n+1}^{N}(e_j - \varepsilon)\right)^2 - \frac{1}{N} \\
&\geq \frac{1}{N}\left(\left(\sum_{j=n}^{N}(e_j - \varepsilon)\right)^+\right)^2 - \frac{1}{N}\left(\left(\sum_{j=n+1}^{N}(e_j - \varepsilon)\right)^+\right)^2 - \frac{1}{N}.
\end{aligned} \tag{38}$$

  Adding (38) and (36), we obtain (37).

- If $\sum_{j=n+1}^{N}(e_j-\varepsilon)<\varepsilon$, then Bill does not bet at all, and $\mathcal{K}_{n-1}=\mathcal{K}_n$. Because

$$\left(\left(\sum_{j=n}^{N}(e_j-\varepsilon)\right)^+\right)^2-\left(\left(\sum_{j=n+1}^{N}(e_j-\varepsilon)\right)^+\right)^2\le(\varepsilon+(e_n-\varepsilon))^2\le 1,$$

  we again obtain (37) from (36).

■

Proof of Proposition 4 The inequality $\mathrm{Freq}_N\ge\varepsilon+\delta_2$ is equivalent to

$$\frac{1}{N}\left(\left(\sum_{j=1}^{N}(e_j-\varepsilon)\right)^+\right)^2\ge N\delta_2^2. \tag{39}$$

Bill's strategy in Lemma 5 does not risk bankruptcy (it is obvious that $\mathcal{K}_n\ge 0$ for all n), and (36) says

$$\mathcal{K}_0\ge\frac{1}{N}\left(\left(\sum_{j=1}^{N}(e_j-\varepsilon)\right)^+\right)^2. \tag{40}$$

Combining (39) and (40) with the assumption that $N\ge 1/\delta_1\delta_2^2$, we see that when the event $\mathrm{Freq}_N\ge\varepsilon+\delta_2$ happens, $\mathcal{K}_0\ge 1/\delta_1$. So by (34) and (35), $\overline{\mathrm{P}}(\mathrm{Freq}_N\ge\varepsilon+\delta_2)\le\delta_1$. ■

A.3 The Independence of Hits for Fisher's Interval

Recall that if $z_1,\dots,z_n,z_{n+1}$ are independent normal random variables with mean 0 and standard deviation 1, the distribution of the ratio

$$\frac{z_{n+1}}{\sqrt{\sum_{i=1}^{n}z_i^2/n}} \tag{41}$$

is called the *t-distribution with n degrees of freedom*. The upper percentile points for this distribution, the points t_n^ε exceeded by (41) with probability exactly ε, are readily available from textbooks and standard computer programs.

Given a sequence of numbers $z_1,\dots,z_l$, where $l\ge 2$, we set

$$\bar{z}_l:=\frac{1}{l}\sum_{i=1}^{l}z_i \qquad\text{and}\qquad s_l^2:=\frac{1}{l-1}\sum_{i=1}^{l}(z_i-\bar{z}_l)^2.$$

As we recalled in §2.1.1, R. A. Fisher proved that if $n\ge 3$ and $z_1,\dots,z_n$ are independent and normal with a common mean and standard deviation, then the ratio t_n given by

$$t_n:=\frac{z_n-\bar{z}_{n-1}}{s_{n-1}}\sqrt{\frac{n-1}{n}} \tag{42}$$

has the t-distribution with $n-2$ degrees of freedom (Fisher, 1935). It follows that the event

$$\bar{z}_{n-1} - t_{n-2}^{\varepsilon/2}\, s_{n-1} \sqrt{\frac{n}{n-1}} \le z_n \le \bar{z}_{n-1} + t_{n-2}^{\varepsilon/2}\, s_{n-1} \sqrt{\frac{n}{n-1}} \tag{43}$$

has probability $1-\varepsilon$. We will now prove that the t_n for successive n are independent. This implies that the events (43) for successive values of n are independent, so that the law of large numbers applies: with very high probability approximately $1-\varepsilon$ of these events will happen. This independence was overlooked by Fisher and subsequent authors.

We begin with two purely arithmetic lemmas, which do not rely on any assumption about the probability distribution of $z_1, \dots, z_n$.

Lemma 6 *The ratio t_n given by (42) depends on $z_1, \dots, z_n$ only through the ratios among themselves of the differences*

$$z_1 - \bar{z}_n, \dots, z_n - \bar{z}_n.$$

Proof It is straightforward to verify that

$$z_n - \bar{z}_{n-1} = \frac{n}{n-1}\,(z_n - \bar{z}_n) \tag{44}$$

and

$$s_{n-1}^2 = \frac{(n-1)s_n^2}{n-2} - \frac{n(z_n - \bar{z}_n)^2}{(n-1)(n-2)}. \tag{45}$$

Substituting (44) and (45) in (42) produces

$$t_n = \frac{\sqrt{n(n-2)}(z_n - \bar{z}_n)}{\sqrt{(n-1)^2 s_n^2 - n(z_n - \bar{z}_n)^2}} \tag{46}$$

or

$$t_n = \frac{\sqrt{n(n-2)}(z_n - \bar{z}_n)}{\sqrt{(n-1)\sum_{i=1}^n (z_i - \bar{z}_n)^2 - n(z_n - \bar{z}_n)^2}}. \tag{47}$$

The value of (47) is unaffected if all the $z_i - \bar{z}_n$ are multiplied by a nonzero constant. ■

Lemma 7 *Suppose $\bar{z}_n$ and s_n are known. Then the following three additional items of information are equivalent, inasmuch as the other two can be calculated from any of the three:*

1. z_n
2. $\bar{z}_{n-1}$ *and* s_{n-1}
3. t_n

Proof Given z_n, we can calculate $\bar{z}_{n-1}$ and s_{n-1} from (44) and (45) and then calculate t_n from (42). Given $\bar{z}_{n-1}$ and s_{n-1}, we can calculate z_n from (44) or (45) and then t_n from (42). Given t_n, we can invert (46) to find z_n (when $\bar{z}_n$ and s_n are fixed, this equation expresses t_n as a monotonically increasing function of z_n) and then calculate $\bar{z}_{n-1}$ and s_{n-1} from (44) and (45). ■

Now we consider probability distributions for $z_1, \dots, z_n$.

Lemma 8 *If $z_1, \ldots, z_n$ are independent and normal with a common mean and standard deviation, then conditional on $\bar{z}_n = w$ and $\sum_{i=1}^n (z_i - \bar{z}_n)^2 = r^2$, the vector $(z_1, \ldots, z_n)$ is distributed uniformly over the $(n-2)$-dimensional sphere of vectors satisfying these conditions (the intersection of the hyperplane $\bar{z}_n = w$ with the $(n-1)$-dimensional sphere of radius r centered on the point $(w, \ldots, w)$ in $\mathbb{R}^n$).*

Proof The logarithm of the joint density of $z_1, \ldots, z_n$ is

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (z_i - \mu)^2 = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^n (z_i - \bar{z}_n)^2 + n(\bar{z}_n - \mu)^2\right), \qquad (48)$$

where μ and σ are the mean and standard deviation, respectively. Because this depends on $(z_1, \ldots, z_n)$ only through $\bar{z}_n$ and $\sum_{i=1}^n (z_i - \bar{z}_n)^2$, the distribution of $(z_1, \ldots, z_n)$ conditional on $\bar{z}_n = w$ and $\sum_{i=1}^n (z_i - \bar{z}_n)^2 = r^2$ is uniform over the set of vectors satisfying these conditions. ■

Lemma 9 *If the vector $(z_1, \ldots, z_n)$ is distributed uniformly over the $(n-2)$-dimensional sphere defined by the conditions $\bar{z}_n = w$ and $\sum_{i=1}^n (z_i - \bar{z}_n)^2 = r^2$, then t_n has the t-distribution with $n-2$ degrees of freedom.*

Proof The distribution of t_n does not depend on w or r. This is because we can transform the uniform distribution over one $(n-2)$-dimensional sphere in $\mathbb{R}^n$ into a uniform distribution over another by adding a constant to all the z_i and then multiplying the differences $z_i - \bar{z}_n$ by a constant, and by Lemma 6, this will not change t_n.

Now suppose $z_1, \ldots, z_n$ are independent and normal with a common mean and standard deviation. Lemma 8 says that conditional on $\bar{z}_n = w$ and $(n-1)s_n^2 = r^2$, the vector $(z_1, \ldots, z_n)$ is distributed uniformly over the sphere of radius r centered on $w, \ldots, w$. Since the resulting distribution for t_n does not depend on w or r, it must be the same as the unconditional distribution. ■

Lemma 10 *Suppose $(z_1, \ldots, z_N)$ is distributed uniformly over the $N-2$-dimensional sphere defined by the conditions $\bar{z}_n = w$ and $\sum_{i=1}^n (z_i - \bar{z}_n)^2 = r^2$. Then $t_3, \ldots, t_N$ are mutually independent.*

Proof It suffices to show that t_n still has the t-distribution with $n-2$ degrees of freedom conditional on $t_{n+1}, \ldots, t_N$. This will imply that t_n is independent of $t_{n+1}, \ldots, t_N$ and hence that all the t_n are mutually independent.

We start knowing $\bar{z}_N = r$ and $s_N = w$. So by Lemma 7, learning $t_{n+1}, \ldots, t_N$ is the same as learning $z_{n+1}, \ldots, z_N$. Geometrically, when we learn z_N we intersect our $(N-1)$-dimensional sphere in $\mathbb{R}^N$ with a hyperplane, reducing it to an $(N-2)$-dimensional sphere in $\mathbb{R}^{N-1}$. (Imagine, for example, intersecting a sphere in $\mathbb{R}^3$ with a plane: the result is a circle.) When we learn z_{N-1}, we reduce the dimension again, and so on. In each case, we obtain a uniform distribution on the lower-dimensional sphere for the remaining z_i. In the end, we find that $(z_1, \ldots, z_n)$ is distributed uniformly over an $(n-1)$-dimensional sphere in $\mathbb{R}^n$, and so t_n has the t-distribution with $n-2$ degrees of freedom by Lemma 9. ■

Proposition 11 *Suppose $z_1, \ldots, z_N$ are independent and normal with a common mean and standard deviation. Then $t_3, \ldots, t_N$ are mutually independent.*

Proof By Lemma 10, $t_3, \ldots, t_N$ are mutually independent conditional on $\bar{z}_N = w$ and $s_N = r$, each t_n having the t-distribution with $n-2$ degrees of freedom. Because this joint distribution for $t_3, \ldots, t_N$ does not depend on w or r, it is also their unconditional joint distribution. ■

References

Eugene A. Asarin. Some properties of Kolmogorov δ-random finite sequences. *Theory of Probability and its Applications*, 32:507–508, 1987.

Eugene A. Asarin. On some properties of finite objects random in the algorithmic sense. *Soviet Mathematics Doklady*, 36:109–112, 1988.

Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Volume I. Prentice Hall, Upper Saddle River, New Jersey, second edition, 2001.

Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

Emanuel Czuber. *Wahrscheinlichkeitsrechnung und ihre Anwendung auf Fehlerausgleichung, Statistik, und Lebensversicherung*. Volume I. Wahrscheinlichkeitstheorie, Fehlerausgleichung, Kollektivmasslehre. Teubner, Leipzig and Berlin, third edition, 1914.

Arthur P. Dempster. *Elements of Continuous Multivariate Analysis*. Addison-Wesley, Reading, Massachusetts, 1969.

Norman R. Draper and Harry Smith. *Applied Regression Analysis*. Wiley, New York, third edition, 1998.

Bradley Efron. Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64:1278–1302, 1969.

Ronald A. Fisher. Applications of Student's distribution. *Metron*, 5:90–104, 1925.

Ronald A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398, 1935.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Hafner, New York, third edition, 1973. First edition, 1956.

Alex Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *Computer Journal*, 50:151–163, 2007.

Edwin Hewitt and Leonard J. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501, 1955.

Steffen L. Lauritzen. *Extremal Families and Systems of Sufficient Statistics*. Volume 49 of *Lecture Notes in Statistics*. Springer, New York, 1988.

Erich L. Lehmann. *Testing Statistical Hypotheses*. Wiley, New York, second edition, 1986. First edition, 1959.

Per Martin-Löf. Repetitive structures and the relation between canonical and microcanonical distributions in statistics and statistical mechanics. In Ole E. Barndorff-Nielsen, Preben Blæsild, and Geert Schou, editors, *Proceedings of Conference on Foundational Questions in Statistical Inference*, pages 271–294, Aarhus, 1974. Memoirs 1.

Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society, Series A*, 237:333–380, 1937.

Geoff K. Robinson. Some counterexamples to the theory of confidence intervals. *Biometrika*, 62: 155–161, 1975.

Thomas P. Ryan. *Modern Regression Methods*. Wiley, New York, 1997.

George A. F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley, Hoboken, NJ, second edition, 2003.

Glenn Shafer. From Cournot's principle to market efficiency. In Jean-Philippe Touffut, editor, *Augustin Cournot, Economic Models and Rationality*. Edward Elgar, Cheltenham, 2007.

Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's only a Game!* Wiley, New York, 2001.

Dylan S. Small, Joseph L. Gastwirth, Abba M. Krieger, and Paul R. Rosenbaum. R-estimates vs. GMM: A theoretical case study of validity and efficiency. *Statistical Science*, 21:363–375, 2006.

Student (William S. Gossett). On the probable error of a mean. *Biometrika*, 6:1–25, 1908.

John W. Tukey. Sunset salvo. *American Statistician*, 40:72–76, 1986.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.