Statistics and common sense

November 26, 2021

Nobuyuki Hanaki
ISER, Osaka University, Japan

Jan R. Magnus[1]
Department of Econometrics and Data Science, Vrije Universiteit Amsterdam
and Tinbergen Institute, Amsterdam, The Netherlands

Donghoon Yoo
ISER, Osaka University, Japan

Abstract: Common sense is a dynamic concept and it is natural that our (statistical) common sense lags behind the development of statistical science. What is not so easy to understand is why common sense lags behind as much as it does. We conduct a survey among Japanese students and try to understand why some probabilistic and statistical questions that baffled great minds a few hundred years are now easy, while other (relatively straightforward) questions are not only difficult but even counter-intuitive.

JEL Classification: B16, C10, C91, D80, D90

Keywords: probability, statistical methods, experiment, common knowledge

Acknowledgements: We gratefully acknowledge financial support from grant-in-aid for scientific research (KAKENHI, grant number: 20H05631) from the Japan Society for the Promotion of Science (JSPS) as well as support from the Joint Usage/Research Center at ISER, Osaka University.

[1]Corresponding author. Address: Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. *E-mail addresses:* nobuyuki.hanaki@iser.osaka-u.ac.jp (Hanaki); jan@janmagnus.nl (Magnus); donghoonyoo@iser.osaka-u.ac.jp (Yoo).

La théorie des probabilités n'est au fond
que le bon sens réduit au calcul
P. S. Laplace (1814)

1 Introduction

The study of medicine, zoology, geography, astronomy, and of sculpture and music goes back thousands of years; and this applies also — and in particular — to mathematics. As a field of study, mathematics is at least 5000 years old and reached a high point in the third century BC with Euclid of Alexandria and Archimedes of Syracuse. Primary school children learn mathematics through counting, then adding and multiplying, then simple equations like $x+2=5$, and this develops into high-school and undergraduate mathematics in a perfectly natural way.[2]

In contrast, probability theory is a young science. It originated from the need to calculate the odds in games of chance, but although the Greeks presumably played games of chance (according to mythology, both Hermes and Pan gambled), there is no record of any mathematical analysis of gambling and odds until the sixteenth century. A gambler may wish to know how many times he needs to throw with one die so that the probability of obtaining 6 at least once exceeds 50%. Cardano (1501–1575) thought the answer was three (in fact, the answer is four). Or: if we throw with two fair dice, how many times do we need to throw so that the probability of obtaining at least one double-6 exceeds 50%. The Chevalier de Méré (1607–1684) thought that he needed 24 throws. This problem has become famous because it intrigued Blaise Pascal (1623–1662) and Pierre de Fermat (1601–1665), and the solution (25 throws) is contained in a letter of Pascal to Fermat dated July 29, 1654. These two examples show how eminent mathematicians struggled with problems that we now find quite trivial. Probability theory begins with Pascal, Fermat, and Christiaan Huygens (1629–1695), and we place its birth year at 1654; see Kahneman (2011) and Tijms (2021) for examples and historical details of early and current pitfalls in probability.

While probability theory is a young science, mathematical statistics is much younger and we place its birth year at 1809. In statistics and econo-

[2]One would think that kids, after learning how to add and multiply, then learn subtraction and division, in that order. But this is not so. In most countries, kids learn fractions at primary school, but not negative numbers. So they can solve $1/3+1/4$ but not $2-5$. This is because ratios are closer to daily life and common sense than negative numbers. You can cut an apple in two halves, but what does it mean to take away two apples when there is only one apple on the table?

metrics we speak about "least squares" as if this were one concept, but in fact it has two quite distinct meanings. We can think of least squares as a method of approximation (as Legendre did in 1805) or as a method of estimation (as Gauss did in 1809 and 1823). Legendre simply wanted to draw the best line through a given set of points and he defined "best" as the minimum of the sum of squared deviations. No randomness enters in Legendre's approach. In contrast, Gauss studied an estimation problem. In his 1809 monograph he studied the linear model $y = X\beta + u$ where the u_i are independent and identically distributed with mean zero and common variance. Instead of assuming a normal distribution on the u_i and then showing that the maximum likelihood estimator is equal to the least squares formula, Gauss asked the opposite question: what distribution is required so that the resulting maximum likelihood estimator is given by the least-squares formula? This is not an easy question to answer and Gauss had to make several rather heroic assumptions before he arrived at the normal distribution. Gauss needed these strong assumptions because he wanted to show that normality of the error distribution is sufficient *and* necessary. But it is only sufficient. Nevertheless Gauss showed in 1809 that the maximum likelihood estimator of β in the linear model under normality is given by the least-squares formula.[3] Not completely happy with his assumptions, Gauss considered the same model again in 1823. This time he asked a different question, namely: which linear unbiased estimator has the smallest variance? The answer is what we now call the Gauss–Markov theorem.

Probability theory and statistics are generally considered difficult fields. Of course, other fields, for example physics or law, are also difficult, but the difficulty in dealing with random variables seems to be of a different nature. When we start an undergraduate degree in physics or law we already have some basic understanding of the subject. But dealing with variables which do not take specific values (as in algebra), but rather follow some probabilistic law — this requires a new way of thinking, and our mind is apparently not very well equipped for this task.

Why not? Maybe because probability and statistics are such young fields. Durbin (1985, 1988) attempted a Darwinian approach arguing that we acquired just enough thinking capacity to ensure survival as primitive hominids millennia ago, which would explain why we can do mathematics as well as we can, but not probability theory. But why wouldn't we have needed some knowledge of risk and probability to survive?

[3]Gauss' analysis is discussed in detail in Magnus (2021), where it is shown that — contrary to what most historical commentators write — Gauss' treatment is correct within his self-imposed framework.

Suppose you are poisoned in the jungle and the only way to save yourself is to lick a special kind of frog. Only the female of that species will do; licking the male frog doesn't help. The male and female frogs look identical and appear with equal probabilities. The only difference is that the male frogs sometimes emit a distinctive croak. You spot a frog in front of you, but then you hear a croaking sound behind you. You turn around and spot two frogs there. There's only time to run to one side. Which way should you run?

Surely our ancestors would have been much helped in their survival if they could solve this and similar puzzles, which even today cause controversy among non-probabilists.[4]

Another possible explanation is provided by the idea of "morphic resonance" (Sheldrake, 1995). When laboratory rats have learned a new maze, rats elsewhere seem to learn it more easily. How can this happen? Perhaps because some form of "collective consciousness" has descended among all rats. This is not a phenomenon that conventional scientific theories can explain, and it remains a precarious argument, easily dismissed as magical thinking and pseudo-science. It is related to Carl Jung's (1936) idea that "there exists a second psychic system of a collective, universal, and impersonal nature which is identical in all individuals". Jung calls this the "collective unconscious".

An easier explanation and less precarious may be how we educate our children. While arithmetic and mathematics are basic school subjects, this is not the case for probability theory and statistics. Most children learn nothing about these subjects, and if they do it is mostly in the form of some simple tricks which do not lead to thinking like a probabilist.

What is worse is that common sense and probabilistic and statistical theory often diverge, and this is the subject of the current paper. In the quote at the top of this paper, Laplace (1814, p. 273) states that "probability theory is *au fond* nothing but common sense reduced to calculus". This may be so, but "common sense" is not a static but a dynamic concept. What is common sense now was not common sense a few hundred years ago, and what is not common sense today may be common sense sometime in the future. Some of the problems and misunderstandings that baffled such minds as Pascal, Fermat, and Leibniz no longer baffle even non-probabilists today. But, at the same time, there are many seemingly simple questions that today even people

[4]The frog in front you has a sample space (M, F) and hence the probability of a female frog is $1/2$. The two frogs behind you have a sample space (MM, MF, FM, FF), but the additional information (the croak) reduces this to (MM, MF, FM). Hence the probability of at least one female frog is $2/3$ and so you should run to the two frogs behind you.

with quantitative skills find hard to solve. And, even if they can solve such problems, they may find the outcomes counter-intuitive and unacceptable. We shall see examples of this divergence between theory and common sense as we proceed.

In 1905 Einstein published his special relativity theory about the structure of spacetime, which led to many counter-intuitive consequences. For example, two events, simultaneous for one observer, may not be simultaneous for another observer if the observers are in relative motion; or: moving clocks tick more slowly than an observer's stationary clock; or: objects are shortened in the direction that they are moving with respect to the observer. But, even though most of us don't fully understand these things, we don't find them counter-intuitive any more. Somehow they have sunk into collective consciousness. Strangely enough, the laws of probability, especially conditional probability, have only sunk into our collective consciousness to a small degree. One of the purposes of the current paper is to investigate the degree of collective consciousness.

As a thread through the paper are ten questions from a survey we conducted among students at Osaka University in Japan. In Section 2 we explain the survey design. In Section 3 we establish the student's background in probability theory and test some basic quantitative ability. We discuss unconditional probabilities in Section 4 and conditional probabilities in Section 5. Then, turning to statistical issues, we discuss prediction in Section 6, prediction intervals in Section 7, and testing in Section 8. In Section 9 we investigate to which extent a background in probability theory helps to answer the questions posed in the survey, and we distinguish between males and females, undergraduates and postgraduates, field of study, the order of asking the questions, and cognitive ability. Section 10 concludes.

2 Survey design

An online survey was conducted between July 27th and July 30th, 2021 by the Experimental Economic Laboratory of the Institute of Social and Economic Research (ISER) at Osaka University. The survey employed a web-based online recruitment system, specifically designed for organizing economic experiments, called ORSEE (Greiner, 2015). We invited 415 students from Osaka University (both undergraduates and postgraduates) who had previously participated in other online experiments, so that we know some of their individual characteristics from these previous experiments (Hanaki et al., 2021). On July 27 each student received an email with an invitation to

participate and an individually customized link to the survey site. Students were asked to fill out the questionnaire by July 30. They had one hour to answer all the questions after accessing the site.[5]

Of the 415 students, 350 students completed the survey. One student completed the survey twice (which was only possible by using two different browsers); he/she is only counted once. This leaves 349 students.

The survey contained ten questions plus one "attention-verification" question (in the middle of the survey), as follows:

Question 0: *What is the likelihood of obtaining head in a throw of a fair coin? Please select "(A) 1" so that we know you are paying attention.*

(A) 1, (B) 1/2, (C) 1/4, (D) 1/6.

Nine students did not answer (A) and these have been excluded.[6] This leaves us with 340 "clean" responses for analysis: 96% Japanese students versus 4% foreign students, 69% undergraduates versus 31% postgraduates, and 61% men versus 39% women.

Students enrolled in a variety of faculties, which we label as

STE: Science, Technology, and Engineering;

Med: Medicine (incl. public health, biology, dentistry, pharmaceutical);

HS: Humanities and Social Science (incl. literature, foreign languages, law, international public policy, economics).

The distribution of the students in our sample over the faculties was as follows:

	STE	*Med*	*HS*	Total
Undergraduate	85	35	116	236
Postgraduate	51	24	29	104
Total	136	59	145	340

The 11 questions are numbered 0–10. In order to minimize possible ordering effects, we prepared two versions of the questionnaire. In the first version the order of the questions is 1, 4, 5, 2, 3, 0, 6, 8, 9, 7, 10. In the second version the order is reversed: 10, 7, . . . , 1. Allocation to the participants was random.

[5]The hour includes the time required to read the consent form and agreeing to it.

[6]Of course, the "correct" answer is (B), so there is a possibility that those who answered (B) had not read or understood the instruction. But we excluded them anyway.

The questions fall into different categories. In Question 1 we ask about the student's background in probability theory. Questions 2 and 3 test some basic quantitative ability. In Questions 4–6 we test knowledge of (unconditional) probabilities and in Question 7 of conditional probability, which is much more difficult. The most interesting questions are 8–10 about prediction, prediction intervals, and testing.

3 Basic quantitative knowledge and ability

We first ask the students about their background in probability and statistics.

Question 1: *Did you follow and pass a probability or statistics course at university level? If so, did you enjoy it?*

	Freq.	Percent
(A) Yes, I followed such a course and I enjoyed it	80	23.5
(B) Yes, I followed such a course and I did not enjoy it	124	36.5
(C) No, I did not follow such a course	136	40.0

We see that 60% of the students received some instruction on probability theory and statistics. Most did not enjoy it.

We next ask two questions on basic quantitative ability.

Question 2: *The average annual salary for an employee at a university is ¥4,000,000. This year, the management awards the following two bonuses to every employee: an end-of-year bonus of ¥300,000 and an incentive bonus equal to 10 percent of the employee's salary. What is the average total bonus received by employees?*

	Freq.	Percent
(A) ¥300,000	1	0.3
(B) ¥400,000	15	4.4
(C) ¥700,000	322	94.7
(D) ¥1,000,000	2	0.6

The correct answer is $300,000 + 400,000 = 700,000$, and 95% of the students got this right.

Question 3: *An economist is studying the relationship between the weight of a car, its reliability rating (the higher the rating, the more reliable), and the annual cost of maintenance. The economist reports the following correlations:*

the correlation between the weight of a car and the car's reliability rating is -0.20*; and the correlation between the weight of a car and the annual maintenance cost is* 0.40*. Which of the following statements are true?*

(1) Heavier cars tend to be more reliable,
(2) Heavier cars tend to be less reliable,
(3) Heavier cars tend to cost more to maintain,
(4) Car weight is related less strongly to its reliability than to its maintenance cost.

	Freq.	Percent
(A) (1) only	2	0.6
(B) (2) only	8	2.3
(C) (1) and (3)	22	6.5
(D) (2), (3), and (4)	308	90.6

The first statement is incorrect, but the other three are correct, so that (D) is the correct answer and 91% had it right. The large majority of the students in our sample thus answered simple quantitative questions correctly, 95% for Question 2 and 91% for Question 3. Still, 5-9% of the students failed to answer even the simplest questions.

4 Unconditional probability

Next we asked three questions about basic (unconditional) probabilities.

Question 4: *One die is tossed. What is the probability that the die will land on a number that is smaller than or equal to 4?*

	Freq.	Percent
(A) 1/4	0	0.0
(B) 1/3	1	0.3
(C) 1/2	10	2.9
(D) 2/3	329	96.8

Question 4 is an easy starting question about basic probabilities. There are four "good" (1, 2, 3, 4) outcomes and two "bad" (5, 6) outcomes, and since all are equally likely, the answer is 2/3. Almost all students (97%) got this right.

One level more difficult is throwing with two dice.

Question 5: *You throw with two dice. Then you can throw any number between 2 and 12. Now, you can throw 12 only by throwing six twice. Similarly, you can throw 11 only by throwing 5 and 6 once each. Which of the following is correct?*

	Freq.	Percent
(A) The probability of throwing 11 is the same as the probability of throwing 12	50	14.7
(B) The probability of throwing 11 is twice the probability of throwing 12	285	83.8
(C) The probability of throwing 11 is three times the probability of throwing 12	5	1.5

This is a famous question, because the celebrated German mathematician Gottfried Wilhelm (von) Leibniz (1646–1716) maintained that it was equally likely to throw twelve with two dice than to throw eleven, because "l'un et l'autre ne ce peut faire que d'une seule manière" (one or the other can be done in only one way).[7] Leibniz' error is remarkable as it came some sixty years after the discoveries of Pascal and Fermat, which marked the birth of probability theory. It demonstrates just how difficult the basic concepts in probability theory are.

The correct answer is (B) because there are 36 equally likely outcomes, of which one (6-6) yields 12 and two (5-6 and 6-5) yield 11. Most of the students (84%) got this right, not because they were more clever than Leibniz but because basic probability theory has somehow sunk into "collective consciousness", at least to some extent.

Equally famous is the next question.

Question 6: *You and your friend play a simple game with one die in a tea house. If the outcome of the throw is even, you get 1 point; if it is odd, your friend gets 1 point. Each of you puts ¥60 on the table, and the first to reach 3 points wins the game and gets the money. One day, at the score 2-1 in your favor, the tea house burns down. You and your friend take the money and run. Next day you meet again (in another tea house), but you don't want to continue the game. Instead you want to divide the money. How should this be done?*

[7]Letter from Leibniz to the Swiss philosopher and mathematician Louis Bourguet, dated March 22, 1714.

	Freq.	Percent
(A) Both you and friend get ¥60	153	45.0
(B) You get ¥80 and your friend gets ¥40	127	37.3
(C) You get ¥90 and your friend gets ¥30	54	15.9
(D) You get everything, your friend gets nothing	6	1.8

Pascal and Fermat corresponded about this question, and the problem was resolved in Pascal (1665) by relating it to Pascal's triangle. Almost one-half of the students in our sample (46.8%) thought of this as an ethical or legal problem, not as a probabilistic problem, so they answered (A) or (D) which have no probabilistic basis. Our interest is in those who attempted a probabilistic solution, so let's concentrate on (B) and (C). The argument in (B) seems to correspond to common sense: you won 2/3 of the games, so you get 2/3 of the money. The problem with this argument is that it is backward-looking; it only considers the past. Pascal, on the other hand, considered the future, arguing that, since three games had already been played, a maximum of two more games needed to be played. These two games could end in two losses for you (with probability 1/4), but in every other case you win the money. So the probability that you win the match is 3/4, and this means that (C) is the correct answer.[8]

In Question 6 we see for the first time a divergence between common sense and probability theory: more than twice of the students preferred (B) over (C). Even though the correct answer has been known for over 350 years and the question should be an easy one in any introductory probability class, common sense has not yet adjusted to probabilistic truth. Most people's intuition is simply wrong.

5 Conditional probability

Much more difficult than unconditional probabilities are conditional probabilities.

Assume that each born child is equally likely to be a boy or a girl. If a family has two children, then what is the probability that both are girls, if we know that the youngest is a girl? And what is the probability, if we know that one of them is a girl?

[8]Suppose the score is 2-1 in your favor but you need n (rather than 3) wins. The probability p_n of lifting the prize decreases monotonically from 3/4 (when $n = 3$) to 1/2 (when $n \to \infty$). Hence there should be a value of n such that $p_n \approx 2/3$. In fact, $p_4 = 11/16$ and $p_5 = 42/64$, so that $p_5 < 2/3 < p_4$. So, somewhere between $n = 4$ and $n = 5$ looking forwards or backwards leads to the same outcome.

A simple question, but it has puzzled many. Let BG denote the case where the oldest child is a boy and the youngest a girl, and similar for BB, GB, and BG. Then the sample space (BB, BG, GB, GG) reduces to (BG, GG) in the first case and it reduces to (BG, GB, GG) in the second case. Hence the conditional probabilities are 1/2 and 1/3, respectively.

Similarly, the famous "Monty Hall" problem is a typical example of a counter-intuitive probability puzzle.

There are three doors. Behind one door is a car, behind the other doors is a goat. You pick a door (call it door A). You're hoping for the car, but since you know nothing the probability of success is 1/3. Monty Hall, the game show host, examines the other two doors (B and C) and opens one with a goat. (If both doors have goats, he picks randomly.) Now, do you stick with door A (your original guess) or do you switch to the unopened door? Does it matter?

The answer is that if you don't switch the probability of winning the car is 1/3, and if you switch the probability is 2/3. The mystery about this puzzle is that an extremely simple setup can cause such confusion, even with trained quantitative people. The difficulty people have in getting conditional probabilities right is quite well-known among behavioral scientists, and some have argued that presenting information in frequencies (rather than in probabilities), helps people to make the correct inference; see Gigrenzer and Edwards (2003).

Here is another counter-intuitive example.

Question 7: *You are worried about your mother's health, and you are convinced that she suffers from some rare disease you have been reading about. So, your mother visits the doctor. The doctor is not convinced, but she agrees to have a test done anyway. The disease your mother gets tested for is quite rare, occurring in only one of every 10,000 people. If your mother has the disease then there is a 99% probability that the test is positive. But if your mother does not have the disease, then the test can also be positive; this happens with a probability of 0.5%. After a few days the test result becomes available. It is positive. What do you think is the probability that your mother has the disease?*

	Freq.	Percent
(A) Less than 2%	111	32.6
(B) About 40%	16	4.7
(C) About 70%	34	10.0
(D) At least 98%	179	52.7

The right answer is (A) and 1/3 of the students got it right. The reasoning, too complex for the untrained probabilist, proceeds as follows. Let A denote the event that your mother has the disease and B the event that she tests positive; and let A^* denote the event that your mother doesn't have the disease and B^* the event that she tests negative. Then,

$$\Pr(A) = 0.0001, \qquad \Pr(B|A) = 0.99, \qquad \Pr(B|A^*) = 0.005.$$

This is all the information we have. The information suffices to obtain the complete joint distribution (in percentages):

	B	B^*	marginal
A	0.00990	0.00010	0.01
A^*	0.49995	99.49005	99.99
marginal	0.50985	99.49015	100.00

This gives

$$\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)} = \frac{0.00990}{0.50985} \approx 0.0194,$$

which is less than 2%. Question 7 is not just a probabilistic puzzle but it has potentially serious practical consequences. Only one in three of our respondents got it right. For the human mind it is counter-intuitive that if a test has a 99% probability to be positive when the patient has the disease and if the test turns out positive, then the probability that the patient actually has the disease is less that 2%. Most people simply don't believe it.

6 Prediction

So far, we have tested our respondents on their quantitative ability (Questions 1–3) and on their understanding of the basic laws of probability (Questions 4–6) and conditional probability (Question 7). We now turn to statistics. There are many counter-intuitive results in statistics and we shall discuss three of them. Of these, our first question is perhaps the most counter-intuitive.

Question 8: *Suppose the Minister of Economics needs to forecast next year's inflation. He asks two well-known experts to advise him. The first expert responds that there will be 1% inflation next year, and the second that there will be 2% inflation. The two forecasts are published in the press so that everybody knows about them. The minister then reflects. He realizes that the two experts know each other and that they base their forecasts on the same*

(or very similar) data sets. Also, from past experience, the minister has more trust in the second expert than the first. After considering the two forecasts he declares the Ministry's forecast to be 2.25%. What do you think of this?

	Freq.	Percent
(A) A forecast larger than 2% is certainly possible, given the fact that the two experts know each other	57	16.8
(B) I would have expected a forecast between 1% and 2%. Why does the minister ignore his advisors?	203	59.7
(C) Such a counter-intuitive forecast would only rarely be reasonable	80	23.5

Let x_1 and x_2 be two uncorrelated observations with a common mean q and variances σ_1^2 and σ_2^2, respectively. We wish to estimate q as an unbiased estimator with the lowest variance (BUE). Hence,

$$\hat{q} = \alpha x_1 + (1-\alpha)x_2 \tag{1}$$

with mean q and variance

$$\mathrm{var}(\hat{q}) = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 = \sigma_2^2\left(\alpha^2(1+\omega^2) - 2\alpha + 1\right), \tag{2}$$

where $\omega = \sigma_1/\sigma_2$. The variance is minimized when $\alpha = 1/(1+\omega^2)$, and we obtain

$$\hat{q} = \frac{x_1 + \omega^2 x_2}{1+\omega^2}, \qquad \mathrm{var}(\hat{q}) = \frac{\omega^2\sigma_2^2}{1+\omega^2}. \tag{3}$$

We note that estimated mean is in-between x_1 and x_2 and that its variance is smaller than both σ_1^2 and σ_2^2. Adding information reduces the variance.

But now consider the same situation assuming that x_1 and x_2 are correlated. Let $\mathrm{cov}(x_1, x_2) = \rho\sigma_1\sigma_2$, where ρ is the correlation coefficient. Again we have $\hat{q} = \alpha x_1 + (1-\alpha)x_2$, as in (1), with the same mean q but not the same variance. In fact,

$$\begin{aligned}\mathrm{var}(\hat{q}) &= \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \\ &= \sigma_2^2\left(\alpha^2(1-2\rho\omega+\omega^2) - 2\alpha(1-\rho\omega) + 1\right),\end{aligned} \tag{4}$$

which is minimized when

$$\alpha = \frac{1-\rho\omega}{1-2\rho\omega+\omega^2}, \tag{5}$$

so that α lies outside the $[0,1]$ interval if and only if $\min(\omega, 1/\omega) < \rho < 1$. The estimator takes the form

$$\hat{q} = \frac{(1-\rho\omega)x_1 + \omega(\omega-\rho)x_2}{1-2\rho\omega+\omega^2}, \tag{6}$$

which agrees with (3) when $\rho = 0$. When $x_1 = 1$ and $x_2 = 2$, then $\hat{q} = 2.25$ if and only if $\omega^2 - 6\rho\omega + 5 = 0$, which can only occur for $1 \leq \omega \leq 5$ and $\sqrt{5}/3 \leq \rho \leq 1$, for example for $\omega = 3$ and $\rho = 7/9 \approx 0.78$; see Magnus and Vasnev (2008, Appendix A) for further details and intuitions about this case.

We conclude that in the presence of positive autocorrelation it is perfectly possible (even likely) that a combined forecast lies outside the bounds indicated by the advisors. But, if the advisors' forecasts are publicly available, then it would take a courageous politician to go outside these bounds, and the outcome of our experiment shows that the public would not understand the rationale for such a deviation. Still, *not* going outside the bounds would be bad policy and would lead to suboptimal forecasts.

7 Prediction intervals

When predicting, we are not only concerned with the prediction itself but also with the reliability of the predictor. Here, we are faced with a puzzle for which there is no easy answer, and where common sense and mathematical rigor don't seem to be in line, even for experts. The next case illustrates this situation.

Question 9: *Suppose the Minister of Economics needs to know the value of some unknown quantity in order to formulate ministry policy. Let's call this quantity q. He consults an expert, who tells him that* $q = 10$. *The expert cannot be entirely sure about this number, but she is confident that q lies between 8 and 12. The minister then proceeds with policy based on this information. After some time he thinks it wise to consult a second expert. The second expert tells him that* $q = 30$. *This expert is not certain either, but he is confident that q lies between 26 and 34. The minister believes that the first expert is slightly more reliable than the second expert, but only slightly. Based on this new information the minister decides to change q from* $q = 10$ *(the old information) to* $q = 20$ *(the average of the old and the new information). But how much confidence should the minister have in this new number? Indicate below the range that the minister should feel quite confident about. Tick only one box. The minister should be quite confident that:*

	Freq.	Percent
(A) q lies between 19 and 21	15	4.4
(B) q lies between 15 and 25	62	18.2
(C) q lies between 11 and 29	89	26.2
(D) q lies between 8 and 34	174	51.2

The problem here is a conflict between two pieces of information, which happens frequently in practice. In Bayesian analysis, for example, the prior and the sample information may deliver conflicting messages. In the normal framework (normal prior, normal likelihood) this implies that the posterior mean is somewhere in-between the mean of the prior and the mean of the sample, which is reasonable. But it also implies that the posterior variance is *smaller* than the variance of the prior and the variance of the sample. This seems also reasonable because we have added information so that the precision should increase. But it is counter-intuitive (also for the professional), because the conflicting information makes us *less* confident about the result: more information leads to less confidence.

The example in Question 9 is frequentist rather than Bayesian, but the idea is the same. We have two pieces of information, say $x_1 = 10$ and $x_2 = 30$ with standard deviations which are approximately equal to $\sigma_1 = 1$ and $\sigma_2 = 2$. Then, if the two observations are uncorrelated, the average $\bar{x} = (x_1 + x_2)/2$ has variance

$$\text{var}(\bar{x}) = \frac{\sigma_1^2 + \sigma_2^2}{4} = 5/4. \tag{7}$$

The standard deviation of $\bar{x}$ is therefore $\sqrt{5}/2 \approx 1.12$ and a reasonable confidence interval for the unknown mean q would be $18 < q < 22$.

More generally, allowing for different weights and for possible correlation, we write again, as in (1), $\hat{q} = \alpha x_1 + (1-\alpha)x_2$ with mean q and variance (4). The variance is minimized when α takes the value in (5), in which case the estimator and its variance take the form

$$\hat{q} = \frac{(1-\rho\omega)x_1 + \omega(\omega-\rho)x_2}{1 - 2\rho\omega + \omega^2}, \qquad \text{var}(\hat{q}) = \frac{(1-\rho^2)\omega^2\sigma_2^2}{1 - 2\rho\omega + \omega^2}, \tag{8}$$

which reduces to (3) when $\rho = 0$. In our case, we have $x_1 = 10$, $x_2 = 30$, $\sigma_1 = 1$, $\sigma_2 = 2$, and $\omega = \sigma_1/\sigma_2 = 1/2$, so that

$$\hat{q} = \frac{35/2 - 20\rho}{5/4 - \rho}, \qquad \text{var}(\hat{q}) = \frac{1-\rho^2}{5/4 - \rho}. \tag{9}$$

For $\rho = 0$ we find $\hat{q} = 14$ which is smaller than $\bar{x} = 20$ because the first advisor is considered more reliable than the second, and $\text{var}(\hat{q}) = 4/5$. Taking correlation into account does not help to increase the variance, which achieves a maximum $\text{var}(\hat{q}) = 1$ at $\rho = 1/2$ and converges to zero as $\rho \to 1$.

Hence, from a theoretical point of view the variance remains small, even when we take correlation into account. This, however, does not correspond to common sense, as is clear from our respondents. Only 4% found it reasonable that q lies between 19 and 21, while the majority (51%) voted for q to lie between 8 and 34.

8 Testing

Testing hypotheses is another counter-intuitive enterprise. When we have an idea that perhaps a statement S is true, then the natural and common sense thing to do is to find many examples where S holds. But when we follow a first course in Statistics we learn that a statistician does the opposite. The statistician puts all effort into *rejecting* the hypothesis and only if they have tried everything and from every angle and still the hypothesis is not rejected, even then the statistician does not conclude that the hypothesis is true, but only that it cannot be rejected.

Most trained statisticians are used to this and don't find it counter-intuitive any more, but for the average citizen it remains counter-intuitive, even though we know since Popper (1962) that if we want to prove a statement like *All statisticians lie*, then searching for more and more dishonest statisticians may be useful in *formulating* the hypothesis but not in *testing* it. In order to test the hypothesis we have to search for honest statisticians. One honest statistician suffices to reject the hypothesis.

In the behavioral sciences most statements are not of the form "all A are B" but rather "most A are B". For example, we know that men run about 10% faster the women.[9] But it is easy to find women who run faster than men, and one counter-example does *not* refute the hypothesis. While the Popperian approach does not work here, the testing theory in mathematical statistics is unaffected. Unfortunately this testing theory is not in line with common sense.

In daily life, most of us have no wish to challenge our beliefs; we prefer to seek confirmation of our beliefs. We choose friends whose ideas agree with our ideas and we read newspapers that promote views that we find sympathetic. How many trained statisticians subscribe to newspapers that reflect views with which they fundamentally disagree? From a statistical point of view this would be the rational thing to do, but few of us actually do it.

Our final question illustrates this behavior.

Question 10: *Suppose you are a high-school student in your final year; next year you'll be going to university. In choosing your field of study, you waver between business economics (choice A) and Japanese literature (choice B).*

Studying business economics (choice A) will give you a qualification that will make you attractive to companies so you can obtain an amazing internship. It is also a great foundation for an MBA or a finance degree, or a

[9]The current world records are 9.58 sec versus 10.49 sec in the 100 meters (9.5%); 1 min 40.91 sec versus 1 min 53.28 sec in the 800 meters (12.3%); and 26 min 11.00 sec versus 29 min 1.03 sec in the 10,000 meters (10.8%).

degree in public policy. During your studies you'll enjoy the problem-solving and strategic thinking the discipline requires. After completing your studies it will be easy to find a job, and you will earn a good salary. A degree in business economics will be useful if you wish to start your own business, and it may help you become a successful investor. You will understand how economies work: to understand economics is to understand how the world works. You will be able to predict trends of businesses and economies based on your knowledge rather than based only on what is reported in the media. You'll also develop an informed perspective on social and political issues.

Studying Japanese literature (choice B) will help you understand Japanese history and culture. You will learn Japanese expressions that are not used in daily life. Japanese literature is rich in history and tradition, and it offers a vast array of genres, authors, and styles that you can explore. Your studies will help you communicate in more meaningful and expressive ways, and they will allow you to understand literature at a deeper level. Literature gives us glimpses of other times, places, and lives that we will never experience otherwise; it offers invaluable insights into what it means to be human. The field offers unlimited directions for creative analysis and original work. After completing your studies you may become a teacher of Japanese or possibly a famous writer, and you will enjoy a richer intellectual life.

Now, what do you choose: A or B?

Most of the respondents (76%) chose for business economics, but this is not really what interests us. After choosing A or B, the question continues:

Next I offer you some further advice. If you want advice in favor of A, click A. If you want advice in favor of B, click B.

This second question *does* interest us. There is little point in asking positive advice in favor of your preferred option, because this should not change your decision. Asking advice favoring the opposite view might change your decision, so this is the sensible thing to do. But this is not how people behave. Apparently, they wish to be confirmed in their view and they are not interested in listening to a deviating view. Of those who chose A in our sample, 61% want advice in favor of A; and of those who chose B, 73% want advice in favor of B.

Depending on their answer the advice would be revealed:

(if clicked A:) *What will your parents think? They will see that you're headed towards a well-paying job, and this will make them happy.*

(if clicked B:) *What will your parents think? They will see that you're headed towards a rewarding life where you will enjoy your work,*

and this will make them happy.
Now, what do you choose: A or B?

Very few students changed their minds. Of those who chose business economics and received affirmative advice, 99% chose business economics again; the minority of students who asked confrontational advice still remained with their original choice (94%). Of those who chose Japanese literature and received affirmative advice, 97% chose Japanese literature again; the minority of students who asked confrontational advice still remained with their original choice (82%).

Steps 2 and 3 are then repeated.

I offer you some further (final) advice. Again, if you want advice in favor of A, click A. If you want advice in favor of B, click B.

Of those who chose A, wanted advice about A and affirmed their choice A after the advice (155 students), 66% chose to receive further advice on A again. Of those who chose B, wanted advice about B and affirmed their choice B after the advice (59 students), 71% chose to receive further advice on B again. This is the largest group (214 students, 65% of the sample), and they represent the people whose interest is in affirming their prior views. They have no interest in the alternative and don't want to put their idea to the test.

At the other end of the scale are those who are willing to challenge their prior ideas. Of those who chose A, wanted advice about B, and affirmed their choice A after the advice (94 students), 55% chose to receive further advice on B again. Of those who chose B, wanted advice about A and affirmed their choice B after the advice (18 students), 33% chose to receive further advice on A again. This means that 112 students (33% of the sample) behaved rationally, following the ideas of statistical testing.

Only 14 students changed their choice after receiving advice: 4 changed their mind *in spite of* affirming advice, but 10 were apparently convinced by the argument in favor of the alternative. Of those 10 students, 7 behaved rationally by challenging their latest choice again.

Depending on their answer the advice would then be revealed:

(if clicked A:) *Studying business economics will help you become a rational person.*
(if clicked B:) *Literature is the pinnacle of civilization — studying it honors the very best humankind has to offer.*
Now, what do you choose: A or B?

Of the 257 students who chose A at the beginning, 246 (96%) chose A again at the end; and of the 83 students who chose B at the beginning 78 (94%) chose B again at the end.

9 Does it help to be trained in probability theory?

Our respondents are part of a student data base, and therefore we know something about them. We know, for example, whether they are male or female, undergraduate or postgraduate, what their field of study is, and we also know something about their "cognitive ability".[10] We now investigate the relevance of these additional pieces of information.

Gender — Female versus male. In our sample about 60% of the respondents followed a course in probability or statistics (Question 1), and this is roughly the same for men and women: 62% for men versus 60% for women. But, of those who followed such a course, men enjoyed it much more than women: 78% versus 50%. This explains, perhaps, why the men in our sample performed better than the women. Of the easy questions (Questions 2–5), men scored 93% and women 88%; while on the difficult questions (Questions 6–10), men scored 33% and women 21%.

Undergraduates versus postgraduates. The undergraduates in our sample had roughly the same exposure to a previous class in probability and statistics as the postgraduates, and their enjoyment of such a class was also roughly the same. Undergraduates performed slightly better than postgraduates: 92% versus 91% on the easy questions, and 29% versus 27% on the difficult questions. In particular in Question 10 (testing), the undergraduates proved themselves more rational than postgraduates.

Field of study. As discussed in Section 2, we distinguish between three fields of study, labeled *STE* (Science, Technology, and Engineering; 40%), *Med* (Medicine; 17%), and *HS* (Humanities and Social Science; 43%). Of the *Med* students, 85% had followed some course in probability and statistics, but only 24% of those had enjoyed the course. Of the *STE* students, fewer students (65%) followed such a course but they enjoyed it more (49%). As expected, the number of students in *HS* with a background in probability

[10] We also know whether a student is Japanese or foreign, but there are only 12 foreign students in our sample — to small a number to draw conclusions.

and statistics is relatively small (45%) and of those only 37% enjoyed the course.

The lack of enjoyment among *Med* students is reflected in how well the students performed in our survey. The *STE* students performed best, followed by the *Med* students, and the *HS* students. On the easy questions the *STE* students scored 95% (36% on the difficult questions), while the scores for the *Med* students were 90% (25%) and for the *HS* students 89% (22%). The difference between the three groups shows up most markedly in the difficult questions.

Order of asking the questions. Not much difference is detected between the two orderings. One might think that students would find "easy-to-difficult" more congenial and therefore perform better than "difficult-to-easy", but this hypothesis is rejected: "easy-to-difficult" scored 57% while "difficult-to-easy" scored 56%, and the difference is not statistically significant.

Cognitive ability. Most students in the data base have been subjected to a six-question cognitive reflection test, where the first three questions are taken from Finucane and Gullion (2010) and the last three from Toplak et al. (2014); see also Frederick (2005). The score *CRT* is the number of correct answers: $0 \leq CRT \leq 6$. For example, one question from the first group is:

Soup and salad cost 5.50 euros in total. The soup costs 5 euros more than the salad. How much does the salad cost (in euros)?

and one question from the second group:

If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together (in days)?

The correct answer for the first question is 0.25 euro, but the intuitive answer is 0.50 euro; while for the second question the correct answer is 4 days and the intuitive answer 9 days.

The *CRT* score is, not surprisingly, highly correlated to Question 1. Of those with a high *CRT* score ($CRT = 6$), 62% followed an earlier probability and statistics course, while of those with a lower score ($CRT \leq 4$) only 49% followed such a course. The *CRT* score is also positively correlated with the performance on our test, and this is especially true for the easy questions (Questions 1–4). Those who got all four questions right have a score of $CRT = 5.5$, which suggests that a correct solution to the easy questions is affected by prior education and field of study, but that difficult questions are difficult regardless of the respondents' background and cognitive ability.

10 Conclusion

The fact that probability, especially conditional probability, is difficult is well-known. In the current paper we concentrated on statistics rather than on probability, and we asked the following two questions. First, why are probability theory and statistics perceived as particularly difficult? Is it their short history (perhaps using a Darwinian argument), morphic resonance, lack of exposure through education? Second, are there common situations where theoretical results are counter-intuitive for all but the best-trained of us, that is, do statistics and common sense diverge, and if so, to what degree?

Let's start with the second question. Yes, statistics and common sense often diverge. We have seen that "easy" probabilistic questions can now be solved even by students without any background in probability and statistics. So, these untrained students can solve problems that puzzled Pascal and Leibniz, presumably because this knowledge has somehow sunk into "collective consciousness". In contrast, "difficult" questions remain difficult regardless of the respondents' background and cognitive ability. Casual observation and a little introspection seems to confirm this. It is unlikely that an academic with at least some quantitative background will make a mistake in some simple arithmetic exercise, say 321×123. But the same academic is not going to be equally confident about a simple question in probability or statistics, such as the question about the family with two children in Section 5. In our statistical questions 8–10 the lack of statistical understanding is quite remarkable. In particular, the concept of statistical testing in Section 8 remains a mystery for most of our respondents. They seek confirmation rather than allowing their view to be challenged. Even properly trained quantitative students don't understand some of the basic ideas of estimation and testing theory, and their intuition is often contrary to statistical theory.

The first question is more difficult. *Why* is statistics so difficult? Maybe because it has such a short history, maybe because we don't (or hardly) learn it at school or from our parents, maybe because it requires a way of thinking that is alien to the human mind.

A proper understanding of risk, probability, and tests is getting increasingly important in our society, and a lack of understanding can be quite dangerous. Since we cannot change the history of our field or the human mind, there is only one way to increase the understanding of random variables and statistics, and that is by introducing children to it at a young age. And there is only one way of achieving this, and that is to provide some serious probability and statistics teaching to their teachers.

References

Durbin, J. (1985). Evolutionary origins of statisticians and statistics. In: *A Celebration of Statistics: The ISI Centenary Volume* (A. C. Atkinson and S. E. Fienberg, Eds). Springer, New York.

Durbin, J.(1988). Is a unified consensus of statistics attainable? *Journal of Econometrics*, 37, 51-61.

Finucane, M. L. and C. M. Gullion (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging*, 25(2), 271–288.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium*. Perthes and Besser, Hamburg. (Translated as "Theory of Motion of the Heavenly Bodies Moving About the Sun in Conic Sections" by C. H. Davis (1857). Little Brown & Co., Boston. Reprinted in 1963, Dover, New York).

Gauss C. F. (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxia*. Dieterich, Göttingen. (Translated as "Theory of the Combination of Observations Least Subject to Errors" by G. W. Stewart (1995). SIAM, Philadelphia).

Gigerenzer, G. and A. Edwards (2003). Simple tools for understanding risks: from innumeracy to insight. *The BMJ*, 327(7417), 741–744.

Greiner, B. (2015). An Online Recruitment System for Economic Experiments. *Journal of the Economic Science Association*, 1(1), 114–125.

Hanaki, N., K. Inukai, T. Masuda, and Y. Shimodaira (2021). Participants' Characteristics at ISER-Lab in 2020. Discussion Paper Nr. 1141, Institute of Social and Economic Research, Osaka University, Osaka, Japan.

Jung, C. G. (1936). The concept of the collective unconscious. In: *Collected Works*, Vol. 9 (1959), p. 42.

Kahneman, D. (2011). *Thinking Fast and Slow*. Farrar, Straus and Giroux, New York.

Laplace, P. S. (1814). *Essai Philosophique sur les Probabilités*. Courcier, Paris, 6th edition, 1840. (Translated as "A Philosophical Essay on Probabilities" by F. W. Truscott and F. L. Emory (1902). Reprinted 1951, Dover, New York).

Legendre A. M. (1805). *Nouvelles Méthodes Pour la Détermination des Orbites des Comètes*. Courcier, Paris.

Magnus, J. R. (2021). Gauss on least-squares and maximum-likelihood estimation, submitted for publication.

Magnus, J. R. and A. Vasnev (2008). Using macro data to obtain better micro forecasts, *Econometric Theory*, 24, 553–579.

Pascal, B. (1665). *Traité du Triangle Arithmétique*. G. Desprez, Paris. www.dspace.cam.ac.uk/handle/1810/244353

Popper, K. R. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, Abingdon-on-Thames, UK.

Sheldrake, R. (1995). *A New Science of Life*. Park Street Press.

Tijms, S. (2021). *Chance, Logic and Intuition*. World Scientific, Singapore.

Toplak, M. E., R. F. West, and K. E. Stanovich (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking and Reasoning*, 20(2), 147–168.