EXPLORING BAYES WITH NO PRIOR KNOWLEDGE

Since early 2020, the novel corona virus SARS-CoV2 has been turning the world upside down. Case numbers rise, newspapers are headlining: **exponential growth**, **flatten the curve**, **stop the curve**, **reproduction rate**, **incidence**. The population turns towards their political leaders and likewise towards their epidemiologists and statisticians. How will the situation develop? Why do different analyses yield different results? Can I trust the results? Can I trust the data?

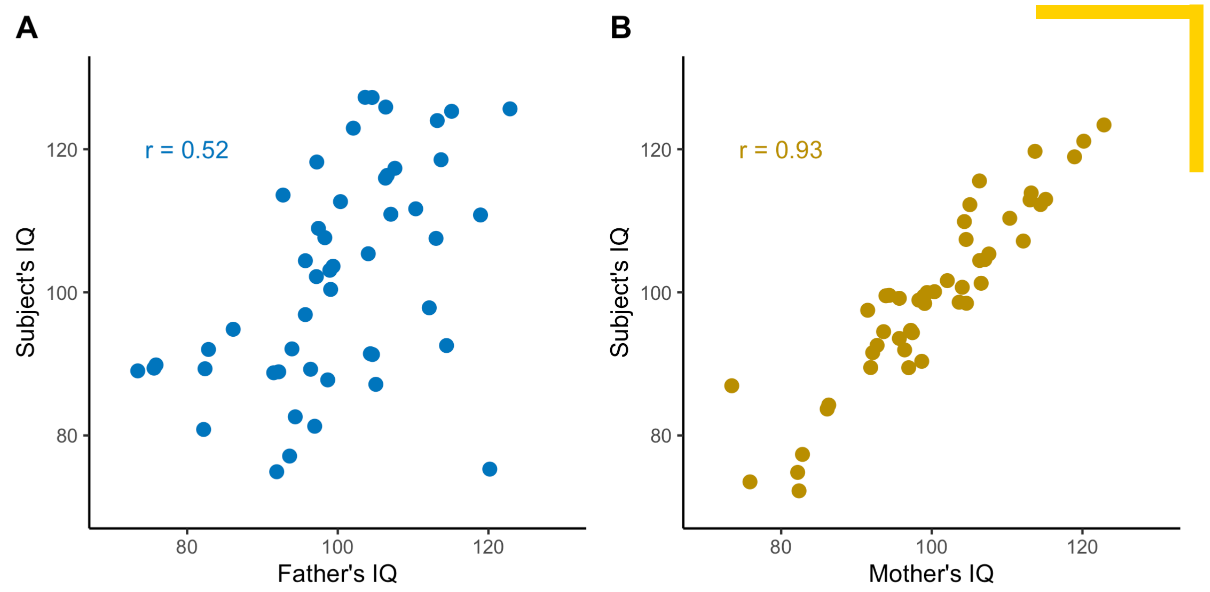
Taking a closer look at the original scientific papers, the vast number of complex statistical methods can be overwhelming. This article will illustrate how modern statistical processes that derive actionable intel from data can meet the expectations of the public: plasticity and communication of uncertainty. These concepts are key features of the so-called Bayesian framework of statistics.

This article will provide a short introduction for readers who are willing to enter the world of Bayesian statistics. It's worth your while. And you didn't know the best part yet: **There won't be a single formula!**

In the majority of psychological research projects, the statistical analysis is based on the frequentist approach. It goes back to the core ideas of Egon Pearson (1895–1980), Jerzy Neyman (1894–1981), and Ronald Fisher (1890–1962). As pioneers of quantitative psychology, Neyman and Pearson introduced the concept of the correlation coefficient in the early 20th century.

During the 20th century, these basic statistical ideas were further developed by numerous scientists, resulting in highly complex statistical methods. However, one aspect faded into the background – the core of their idea, the central assumption that Pearson, Neyman, and Fisher based their framework on. More on that later.

Thomas Bayes (1701-1761) lived long before that time. While Bayes' Theorem is widely known from school – or at latest an introductory lecture on probability theory – the fundamental aspects of Thomas Bayes' works lie deeper than his popular Theorem, which associates conditional probabilities. When he formulated Bayes' Theorem in the 18th century, the complex applications of his theory were far beyond imagination.

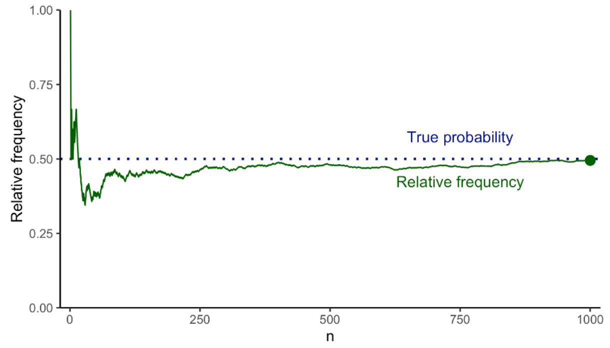


Scatterplot showing the relation between two variables.

*x-axis: Parent's IQ (**A** father, **B** mother), y-axis: Subject's IQ*

*Note the different forms of the data clouds. In plot **B**, the points seem to follow a system: Subjects with a low IQ often occur when their mother has a low IQ as well. On the other hand, when a subject has a high IQ, their mother also seems to have a higher IQ. The scatter in plot **A** appears to be less systematic. The bottom-right-most point represents a subject with an IQ < 80 when the father's IQ is roughly 120. The connection between the IQ measures seems to be low compared to plot **B**.*

*This connection is called **correlation**. The correlation is quantified with the **correlation coefficient r**. It describes the relative systematic joint variance (aka covariance) of two variables. Higher values indicate a stronger systematic relation while values towards zero indicate little systematic relation. **Side note:** The correlation coefficient can also be negative. This means that high x values are usually accompanied by low y values and vice versa. An example is the relation between sports and depression (e.g. Vilhjalmsson & Thorlindsson, 1992).*

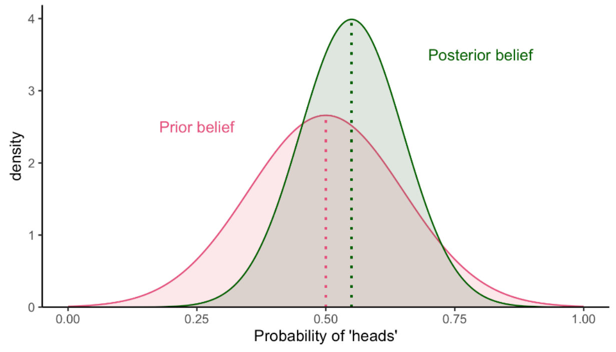


Law of Large Numbers.

Imagine you tossed a coin very often. Instead of the simple result of each toss, you calculate the ratio of heads in all the tosses you have made so far. This ratio, also called relative frequency, is illustrated in the graph to the left. Note how it is volatile at the start and converges towards the actual probability of heads: 50%.

Degree of Belief.

The Bayesian approach conceptualizes the coin's heads probability as the degree of belief in several plausible underlying probabilities. The pink curve illustrates the most likely estimate prior to the experiment (50%). Thus, we assume that the coin is fair. However, it might as well be unfair towards heads or tails. The green curve shows the posterior belief. Apparently, we have good reason to think that the coin is slightly biased towards heads!



Frequentists define probability as the relative frequency of an event after an arbitrary number of observations. On the other hand, the Bayesian approach understands probability as the degree of belief in an event. It provides a framework to update prior knowledge with data that is empirically observed. The benefit of Bayesian analyses in a time where a dynamic pandemic demands fast and dynamic responses is evident: A Bayesian analysis does not yield a single point estimate but quantifies the certainty of the estimate. This result is highly plastic. It can be easily updated with novel data that arise from new available information. Let's illustrate this with an example that does not rely on coins, tea or wine.

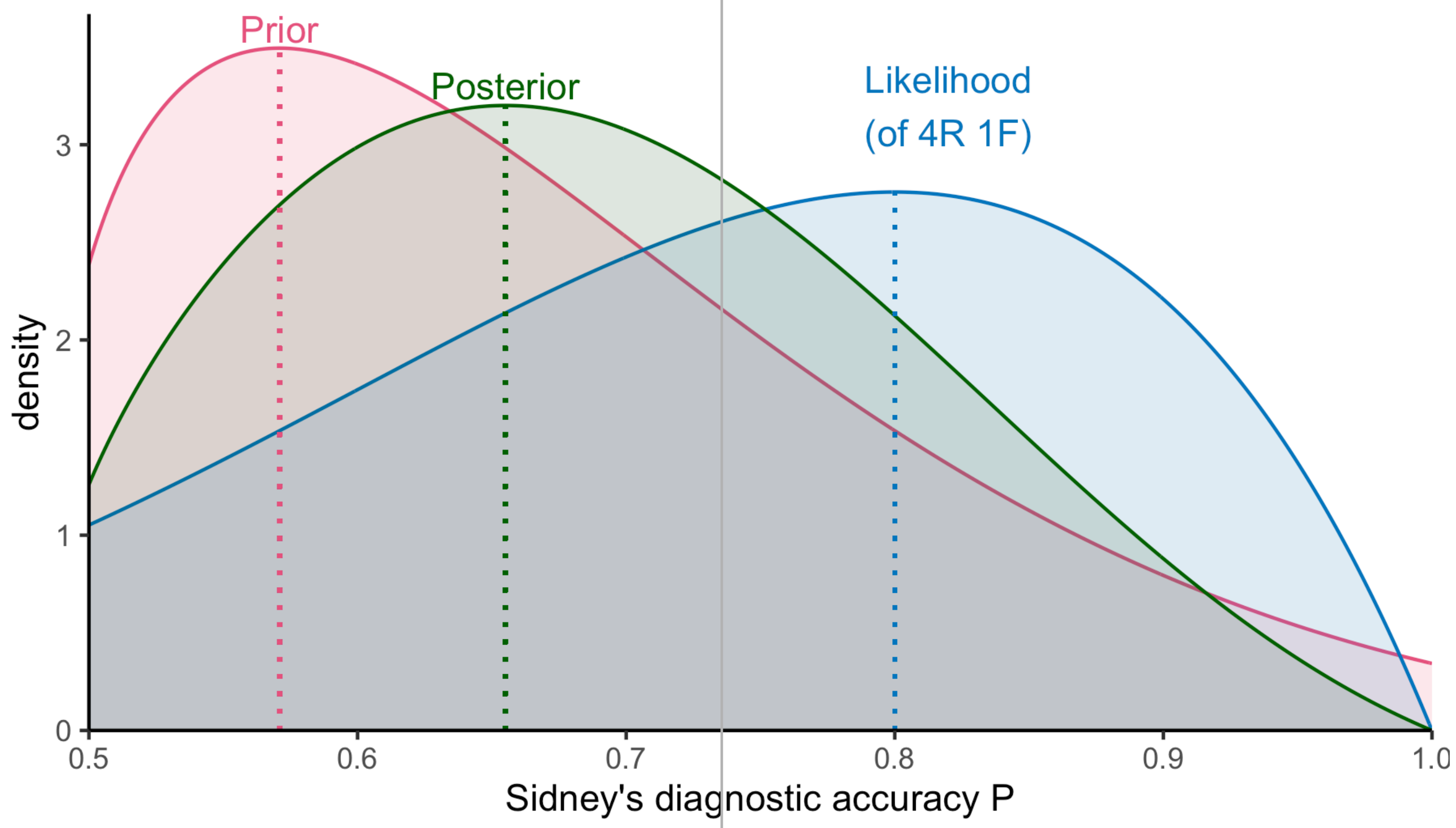
Take a physician who claims to detect covid infections with her bare hands. Let's call her Sidney. You might be curious if Sidney is a fraud or if she can really diagnose covid infections with this technique. In the following, we are going to model the scenario from a Bayesian perspective on probability.

Before the actual analysis, we must think about the probability theory behind detecting covid cases. The physician will be presented with eight pairs of patients – one of whom is covid-negative while the other one is covid-positive. Sidney has to identify the covid-positive patient in each trial. Her judgement will be denoted as right (R) or false (F). If Sidney was entirely unable to distinguish covid-negative and covid-positive patients, she would effectively flip a coin for every diagnosis. In this case, the probability of a right answer is 50% in each trial. In the Bayesian framework, we will now express our belief in the physician as a mathematical function. We call this function **PRIOR**, since it incorporates our belief prior to the analysis. Let's say that Sidney is extraordinarily charismatic, and she has somehow convinced us that she is not a fraud. However, we are still reasonably skeptical about her claim.

The prior should cover the possibility that she is purely guessing (50%) but also that she is absolutely capable of detecting covid cases (100%). This underlying accuracy corresponds to Sidney's diagnostic accuracy with respect to detecting covid infections. We will refer to this accuracy as P. Since percentages are often unhandy, we will express P as a point number between 0.5 (50%) and 1.0 (100%). If we were to commit ourselves to a fixed-point estimate for her diagnostic accuracy P, we might choose a value slightly greater than the guessing rate of P = 0.5 – remember how she somehow convinced us that she may actually possess these diagnostic abilities. The other plausible values between P = 0.5 and P = 1.0 will be filled in so that the prior distribution is somewhat smooth.

Now we conduct the actual experiment. Sidney identifies four out of five patient pairs correctly: 4R, 1W. We are now going to quantify how likely this result is for different values of her diagnostic accuracy P. This **LIKELIHOOD** represents the evidence of the data we observed.

According to the core idea of Bayesian inference, we are now going to update the prior belief with the evidence of the likelihood function. In practice, this is achieved by multiplying and normalizing the two functions. The result is a so-called **POSTERIOR** distribution that expresses the belief in Sidney after the experiment.



Sidenote. The math-savvy reader might be interested in how the likelihood function can be derived.

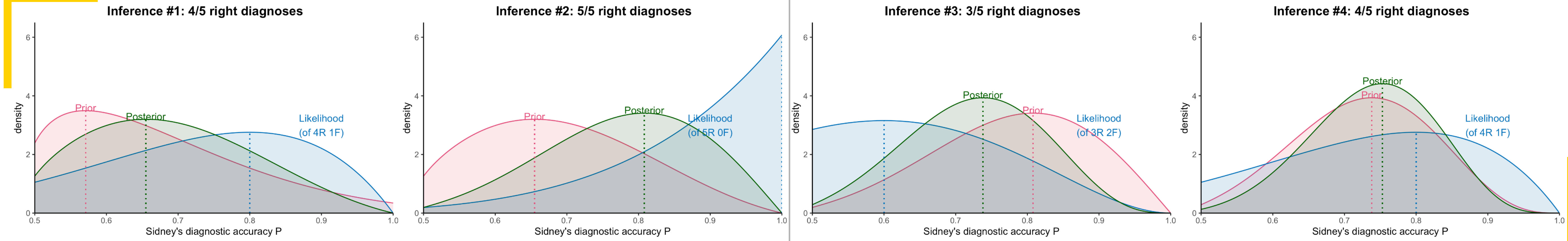
*The case of Sidney the physician can be modeled as a Bernoulli process of n = 5 trials with success rate p = P and failure rate q = 1-P. The resulting number of correctly identified patients k then follows a Binomial distribution **B(n, p)**. The likelihood curve in the graph shows the value of the probability mass function of the Binomial distribution's for fixed parameters n and k with varying success rate p, denoted as **f(p; n,k)**.*

You can play with the numbers and explore how both your prior belief and the outcome of the experiment influence the posterior: The web app accompanying this article is available at marvinschmitt.shinyapps.io/ExploringBayesWithNoPriorKnowledge/

We can see how the belief in Sidney's claim has shifted. Prior to the experiment, we were skeptical but still left the option open for Sidney to have any plausible diagnostic accuracy above the guessing-rate of P = 0.5. We assigned the majority of mass to the proximity of P = 0.6 with a soft decay towards P = 1.0. The likelihood function of Sidney's performance shows how well the data suit different plausible values of P. If we were to pick the estimate of Sidney's diagnostic accuracy P that maximizes the likelihood of the data, it would be P = 0.8. This point-estimate which is purely based on the observed data (and not the prior) is called the **Maximum Likelihood Estimate**. It maximizes the likelihood of the observed data. If you disengage from the broad statistical framework we are building here, the estimate of P = 0.8 seems cunningly apparent: If Sidney correctly identifies 4 out of 5 patients, that's a ratio of 80%. Thus, it is the most likely scenario if she always diagnoses at an accuracy of exactly 80%. However, we are navigating in a Bayesian framework. We have only **updated** our prior belief with this likelihood function to obtain the posterior belief.

We see how our belief towards Sidney's diagnostic capabilities has increased. The posterior distribution is shifted to the right of the prior distribution, indicating more belief in higher values of P. The highest point of the posterior lies at approximately P = 0.65. So, if we were to commit ourselves to a point estimate, we would probably say that she diagnoses 65% of cases correctly. Note how this is not identical to the Maximum Likelihood Estimate. That is because our prior belief in Sidney was particularly low. In fact, if the prior was completely uniform, the aforementioned fixed-point estimate of the posterior distribution would equal the Maximum Likelihood Estimate.

The following paragraphs will generalize the ideas of the toy example to a conceptual level. Bayesian statistics pro-



vide a framework to judge the plausibility of hypotheses when we take the data as a basis. This principle stands in contrast to the frequentist framework which evaluates the plausibility of data when we take a hypothesis as a basis. This seemingly fine difference is fundamental to a degree that Bayesian tests can tackle numerous scientific questions much more precisely. But this is not the only way how a Bayesian analysis can improve the evaluation of a research question.

Even before conducting an analysis, a researcher usually has some idea of the estimate to be taken. This degree of belief before the analysis is called **PRIOR**. The **LIKELIHOOD** quantifies how well the data fit the current belief. Then, the Bayesian inference is conducted: The prior is updated with the likelihood, leading to the posterior. In fact, the likelihood contains all information necessary to adjust the prior. The likelihood only depends on the data that has actually been observed. This means that the Bayesian analysis only depends on data that has been observed. This is called the **LIKELIHOOD PRINCIPLE**. On the other hand, the frequentist approach needs a definition of more extreme data. This is because the frequentist approach tries to quantify the probability of the data under the current estimate. However, this represents a point-estimate which is assigned a probability of zero by definition. Therefore, the frequentist framework quantifies the probability of the data **or more extreme data** under the estimate. This more extreme data might never be observed in reality though. Hence, the frequentist framework does not fulfill the likelihood principle.

In a Bayesian analysis, one's current belief is encoded in the prior, which is in turn updated through the likelihood to obtain the **POSTERIOR**. This means that the posterior reflects one's belief after the analysis. Suppose one was to conduct a subsequent second analysis with novel data. The belief before this second analysis is again the knowledge at this very point in the scientific process. This plasticity of results can be rephrased in Bayesian terminology:

"Today's posterior is tomorrow's prior"
(Lindley, 1972, p.2)

Let's briefly apply this updating process to the example of Sidney the Physician.

Imagine we would conduct the experiment four times:

Experiment	Result
1	4/5 correct
2	5/5 correct
3	3/5 correct
4	4/5 correct

Each time, the most recent posterior is chosen to be the prior of the subsequent experiment. Note how the posterior shifts towards the respective likelihood in each experiment. However, it is much less volatile than the Maximum Likelihood Estimate which would fluctuate more with varying performances in the experiments. This behavior illustrates the plasticity of Bayesian Inference with respect to multiple subsequent experiments.

Again, feel invited to explore the influence of the prior and likelihood functions on the posterior estimate of Sidney's diagnostic accuracy in the accompanying web app at
marvinschmitt.shinyapps.io/ExploringBayesWithNoPriorKnowledge/

Assigning a prior distribution for an analysis is an opportunity since previous findings and beliefs can be incorporated conveniently. It does however seem to come with a flipside: Will the outcome of the analysis actually depend on the choice of the prior? Short answer: **YES, IT WILL.**

Since the prior is updated – and not replaced – with the likelihood of the observed data, the prior does influence the posterior distribution. While critics may call it a weakness of the Bayesian approach, it is actually a key feature. This is because the researcher will always express their uncertainty about the prior in the prior itself. This aspect of uncertainty determines how much the prior will be influenced by the likelihood during the analysis. Remember how we modeled our prior belief in the physician's diagnostic accuracy as a slight tendency above the guessing rate P = 0.5 while distributing probability mass among the remaining plausible values.

A common terminology distinguishes **OBJECTIVE** (uninformed) priors and **SUBJECTIVE** (informed) priors. Objective priors feature large variance and express high uncertainty. Subjective priors attempt to model the researcher's current state of belief with the appropriate degree of certainty to fit the analysis at hand. This behavior renders the criticism on prior dependency irrelevant since researchers can always choose an objective prior when they do not have sufficient domain knowledge for an informed prior. In the case of a massively uninformed prior – i.e. a uniform distribution – the resulting estimate will even be identical to the regularly utilized Maximum Likelihood Estimate. That is, if it is covered by the prior distribution.

Shaping an analysis through the prior can be taken even one step further. Assume you suspected the data collection process to be faulty or unreliable. At the same time, assume you had a strong theoretical background prior to your analysis. In this case, you do not necessarily want the data to influence your belief too strongly. You might choose a prior distribution that assigns a lot of mass onto your particular prior belief – thus reducing the impact of the likelihood on the posterior distribution. Influencing the way the observed data interacts with the estimation process is a means of regularization.

As discussed earlier, the result of Bayesian inference is usually a posterior distribution over the estimated parameter. While this is oftentimes beneficial, critics may claim that there are cases where the result of an analysis should be a point estimate rather than a full distribution. For instance, reporting the 7-day incidence rate of covid cases as a full posterior distribution every day would presumably lead to confusion and mess. Imagine the radio show host desperately describing a full posterior distribution in every news section. In this case, single-value predictions are desired:

"Within the last week, the 7-day incidence rate has dropped to from 165 to 148 cases per 100,000 citizens in Berlin."

One method to conveniently extract a point estimate from a posterior distribution is the so-called **Maximum a Posteriori Estimate** (MAP). The idea is cunningly simple: Choose the parameter value that is the most likely, namely the mode of the posterior distribution. In the previous visualizations of Bayesian Inference, the MAP is shown as a green dashed line at the maximum of the posterior distribution. If the context allows for a little more detail, you might choose to add the posterior variance to the report as a measure of uncertainty.

This short introduction attempted to provide the necessary basic knowledge to conceptually understand Bayesian inference. It has become evident that two key ideas of Bayesian inference – namely uncertainty and plasticity – meet public expectations towards scientific research. This article closes with three calls to action. Firstly, researchers should model their prior assumptions into the analyses they conduct. Secondly, scientific results should always be reported along with the uncertainty tied to them – whether the analysis is Bayesian or not. Thirdly, scientific results should be reported in a way that enables interested readers to understand the entire underlying scientific process.

Notes

Since the article is typeset in a non-academic magazine style, the required citations are not in accordance with the academic APA style and will be provided hereinafter. The example of Sidney the Physician is inspired by Dennis Lindley's illustration in *The Appreciation of Tea and Wine*. The quote *"Today's Posterior is Tomorrow's Prior"* originates from an article by Dennis Lindley as well. The full citations for this article are:

Lindley, D. V. (1993). *The Analysis of Experimental Data: The Appreciation of Tea and Wine*. Teaching Statistics, *15(1)*, 22–25.

Lindley, D.V. (1972). *Bayesian statistics, a review*. Philadelphia, PA: SIAM.

Vilhjalmsson, R., & Thorlindsson, T. (1992). The Integrative and Physiological Effects of Sport Participation: A Study of Adolescents. *The Sociological Quarterly*, *33*(4), 637–647.

If you are interested in a much more in-depth introduction to a modern workflow of Bayesian Inference in statistical modeling, I can highly recommend Michael Betancourt's series of blog articles on *Probability Theory* as well as *Modeling and Inference*. The case studies can be found at betanalpha.github.io/writing/

Source Code

All visualizations have been generated with the R programming language. The main plotting utilities included the packages `ggplot2` and `cowplot`.

The accompanying web app allows the reader to explore the process of Bayesian Inference when Sidney's diagnostic accuracy towards diagnosing covid cases is estimated. The web app was designed using the `shiny` package family.

The commented code for the plots as well as the web app is available online at github.com/marvinschmitt/ExploringBayesWithNoPriorKnowledge