Bayesian Basics

Introduction

Technical objective

Understand the general premises and mechanisms of the Bayesian approach to statistical analysis

Tutorial overview

In this tutorial, we begin laying the groundwork for understanding the Bayesian approach to statistics and data analysis. We first describe frequentist statistics as a familiar framework with which to contrast Bayesian statistics. We then introduce Bayes' theorem, the key mathematical relationship underlying the Bayesian approach. Next, we preview several applied analysis methods based on Bayes' theorem. We conclude with a discussion of the philosophical merits of Bayesian methods, especially for those conducting psychological research.

Review of frequentist statistics

At some point during your education, you've probably learned about statistics, whether that be from a class, a book, or a YouTube video. When students learn about statistics in school, they typically learn what is called *frequentist statistics*. Frequentist statistics covers topics such as null hypothesis significance testing, p-values, and confidence intervals. You may be surprised to learn that this standard set of statistical methods is really more of a framework–and that there exist other plausible and useful frameworks for statistical analysis.

In order to have a comparison point for these other frameworks, it is important to first have a solid understanding of what frequentist statistics is. A key tenet of the frequentist framework is the frequentist interpretation of probability: the probability that an event occurs is defined by the long-term frequency, or observed proportion in the space of all possible relevant events, of that event (Romeijn, 2022).

Frequentist probability simulation

For example, consider rolling a fair six-sided die. Suppose we are interested in the event of rolling a six. If we roll the die, say, 10 times, we will get some number of sixes. We are likely to have one six or two sixes, and we may even have zero sixes. The proportion of sixes we see, with respect to all possible events (i.e., rolling any number from one to six), in each of these cases is $\frac{1}{10}$, $\frac{2}{10}$, and $\frac{0}{10}$, respectively. Let's run 10 simulated die rolls and see what proportion of our rolls are sixes.

First, we set a seed for consistent pseudorandom number generation and load in the tidyverse package (Wickham et al., 2019) for (simulated) data manipulation and visualization.

```r
set.seed(4)
library(tidyverse)  # for plotting with ggplot2
```

```r
six_sided_die <- c(1:6)

roll_a <- sample(six_sided_die, size = 10, replace = TRUE, prob = NULL)
prop_a <- sum(roll_a == 6) / 10
print(prop_a)
```

```
## [1] 0.2
```

Obviously, with only 10 rolls of the die, the frequency of rolling a six does not reveal the true probability of rolling a six, which we know to be $\frac{1}{6}$. However, if we roll the die many more times, we will likely roll six in close to $\frac{1}{6}$ of our rolls. If we could roll the die a number of times that approaches infinity, we would see that the proportion of rolls that are sixes converges to exactly $\frac{1}{6}$. This limit-at-infinity convergence to the true probability represents what we mean by a frequentist long-run probability. We will now run our die-rolling simulation 10,000 and 10,000,000 times and plot the proportions of sixes we get.

```r
roll_b <- sample(six_sided_die, size = 10000, replace = TRUE, prob = NULL)
prop_b <- sum(roll_b == 6) / 10000
print(prop_b)
```

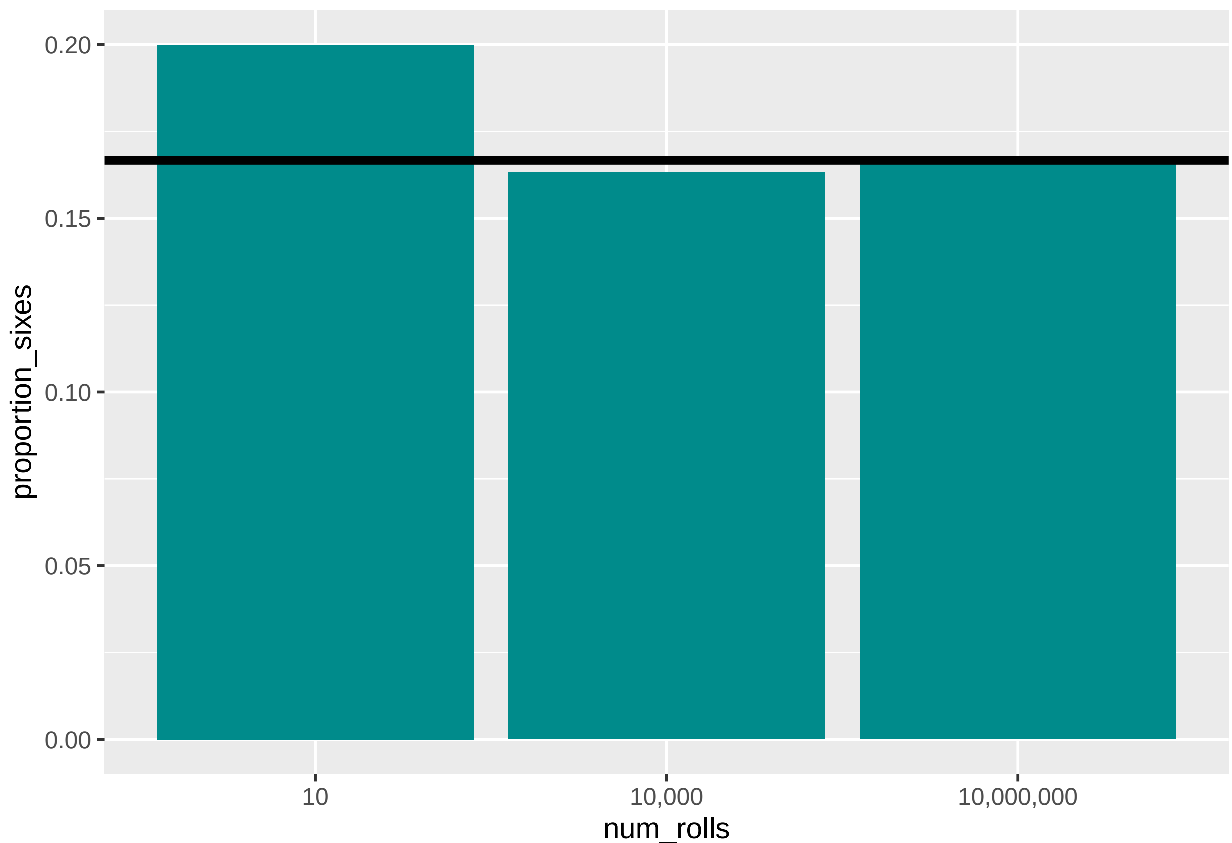
```
## [1] 0.1632
```

```r
roll_c <- sample(six_sided_die, size = 10000000, replace = TRUE, prob = NULL)
prop_c <- sum(roll_c == 6) / 10000000
print(prop_c)
```

```
## [1] 0.1666441
```

```r
die_rolls <- data.frame(matrix(data = NA, nrow = 3, ncol = 2))
colnames(die_rolls) <- c("num_rolls", "proportion_sixes")
die_rolls$num_rolls <- c("10", "10,000", "10,000,000")
die_rolls$proportion_sixes <- c(prop_a, prop_b, prop_c)

die_rolls %>% ggplot(aes(x=num_rolls, y=proportion_sixes)) +
  geom_bar(stat = "identity", fill="cyan4") +
  geom_hline(yintercept = (1/6), linewidth=1.5)
```



Note that the horizontal black line in the above plot represents a probability of exactly $\frac{1}{6}$. We observe that with more rolls, our estimate of the frequentist probability of rolling a six gets closer to the true probability of rolling a six, $\frac{1}{6}$.

In frequentist data analysis, we are generally interested in using the frequentist interpretation of probability to evaluate the plausibility of a proposed data-generating hypothesis, as compared to a null hypothesis. In other words, we want to see if there is a statistically-indicated interesting relationship between our variables, as opposed to no relationship at all. This is evaluation is made using a framework called null hypothesis significance testing (NHST). In NHST, one hopes to show that the observed data are highly improbable, i.e., occur at a very low long-run frequency, in a world where the proposed data-generating hypothesis is *not* true. A "low" long-run frequency is often defined by having a probability (p-value) less than 0.05.

Bayes' theorem

Now that we've covered the basics of frequentist probability and statistical analysis, let's shift our focus to the Bayesian approach. At its core, Bayesian analysis boils down to one simple yet powerful mathematical expression: Bayes' theorem. The expression for Bayes' theorem is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where both A and B are typically viewed as random variables. However, they can also be interpreted as point values when we are plugging in specific values of A and B at which to evaluate the expression. Since we are interested in assessing the plausibility of hypotheses given some evidence (or data), we will rewrite Bayes' theorem using the variables H = hypothesis and E = evidence:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}.$$

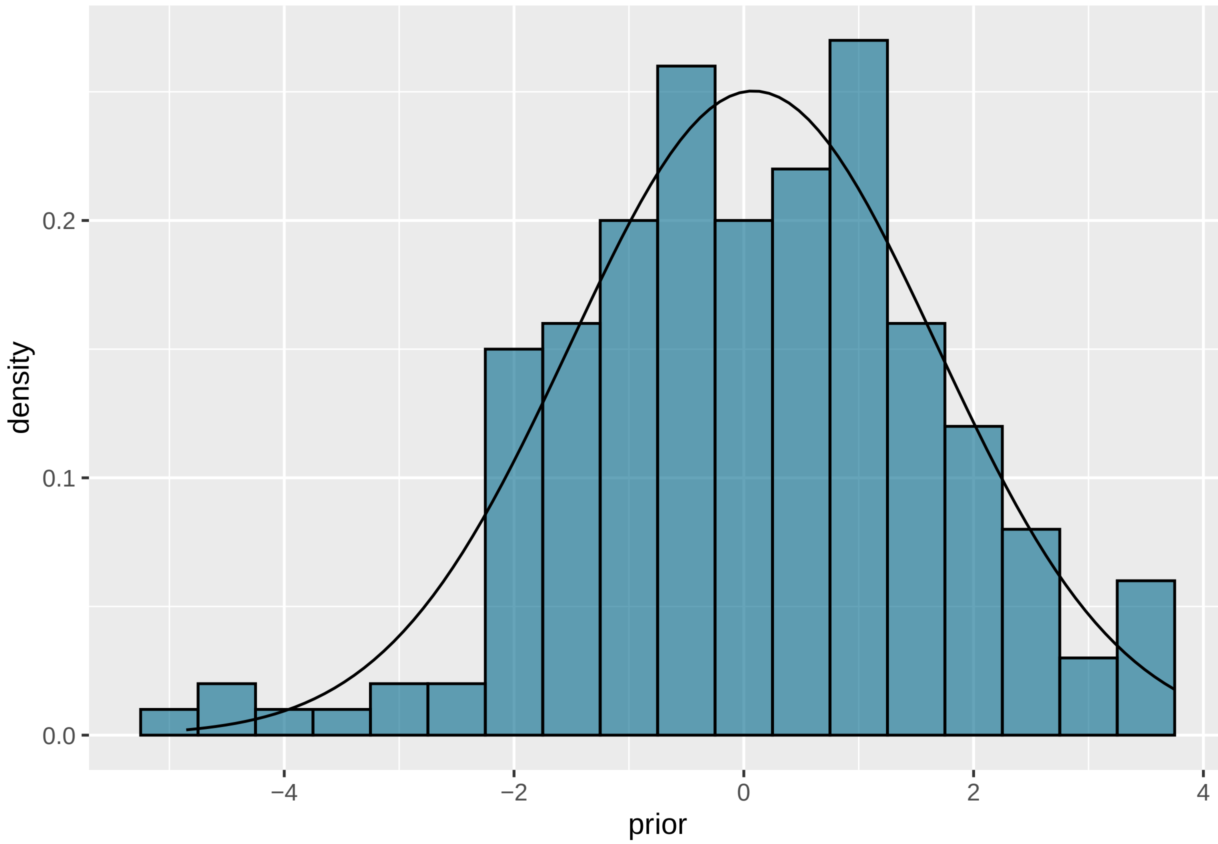
The four terms in Bayes' theorem are often referred to as the **posterior**, the **likelihood**, the **prior**, and the **normalizing constant**. The posterior, $P(H|E)$, refers to the probability of the hypothesis being true after having observed the evidence. The likelihood, $P(E|H)$, refers to the probability of observing the evidence in the case where the hypothesis is known to be true. The prior, $P(H)$, refers to the probability of the hypothesis being true without having observed any evidence. Finally, the normalizing constant, $P(E)$, refers to the probability of the evidence occurring independent of any hypothesis. More pragmatically, the normalizing constant ensures that the probabilities across all possible hypotheses sum to 1.

In general, Bayes' theorem operates under the assumption that our degree of belief in a hypothesis can be expressed in terms of (i) our prior, or existing, degree of belief in the hypothesis and (ii) the contribution of newly observed evidence. In a Bayesian context, the interpretation of a probability is more akin to a "degree of belief" than a long-run frequency (Hájek, 2023). Bayes' theorem allows us to concisely express our confidence in various different data-generating hypotheses via a probability distribution over those hypotheses.

Bayes' theorem simulation

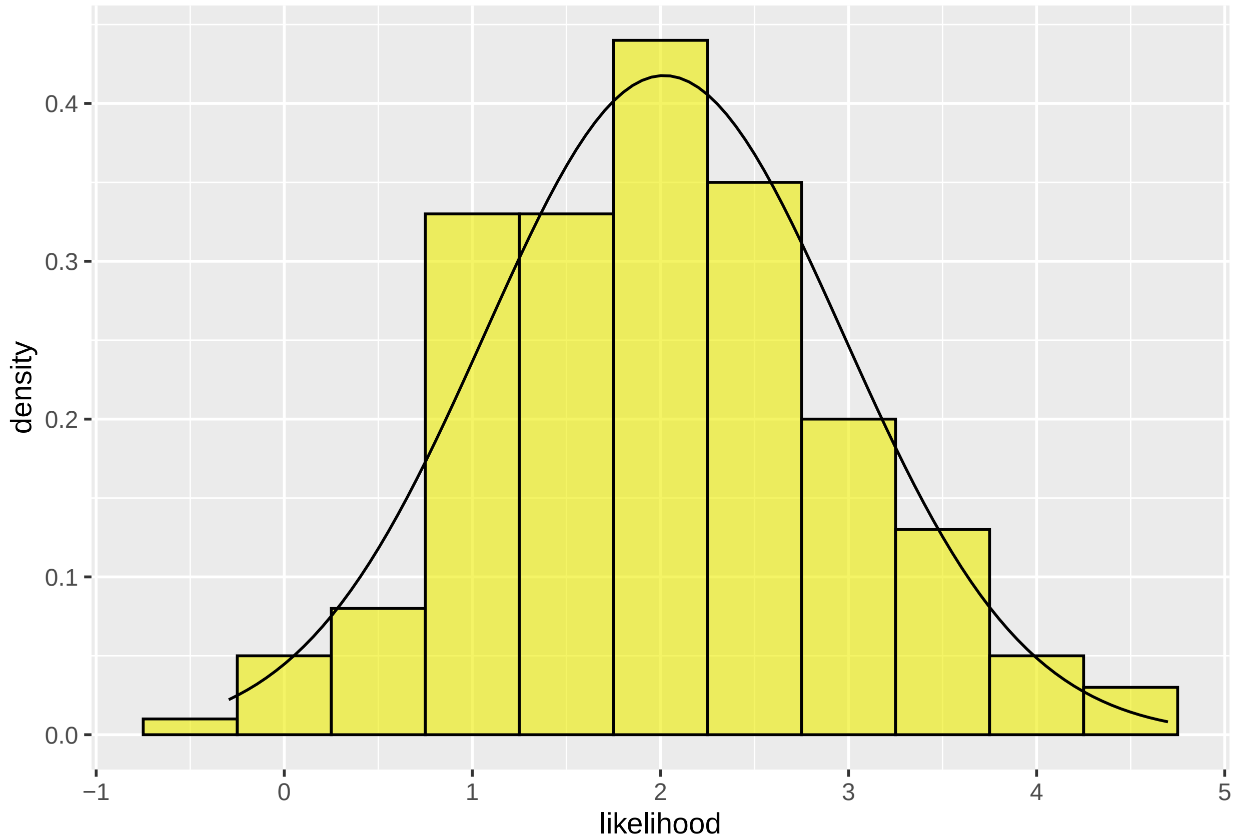
To help demonstrate Bayes' theorem more concretely, let's generate some fake data and run a simulation of Bayes' theorem. We first generate a prior distribution that is normally distributed with a mean of 0 and a standard deviation of 1.5. We plot this distribution both as a histogram, bucketing values by intervals of 0.5, and as a density curve.

```r
dist <- data.frame(prior = rnorm(n = 200, mean = 0, sd = 1.5))
dist %>% ggplot() +
  geom_histogram(mapping = aes(x = prior, y = after_stat(density)),
                 fill = "deepskyblue4", alpha = 0.6, color = "black", binwidth = 0.5) +
  stat_function(fun = dnorm, args = list(mean = mean(dist$prior), sd = sd(dist$prior)))
```



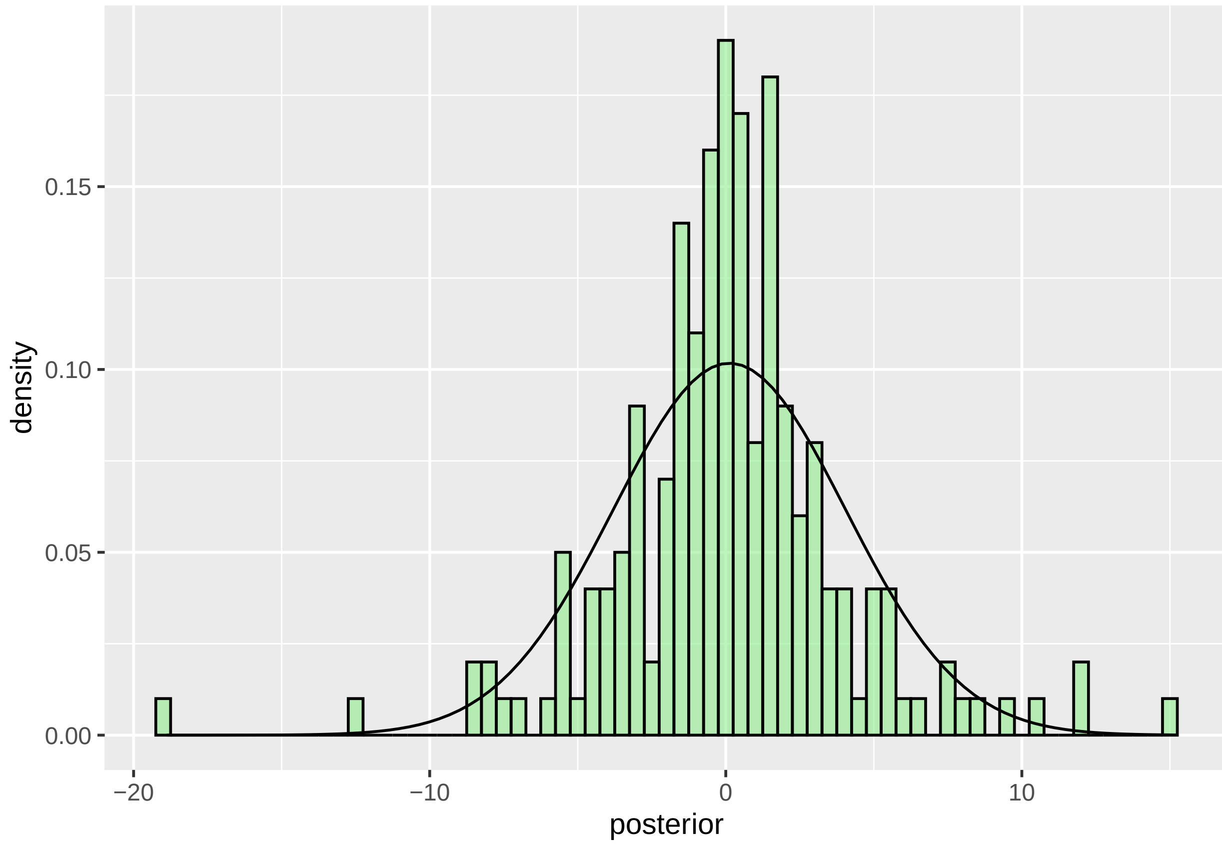
We next generate our "data," also known as the likelihood distribution. We simulate these data as being normally distributed with a mean of 2 and a standard deviation of 1. Since the standard deviation of our likelihood distribution is smaller than that of our prior distribution, we can interpret this as meaning we have greater certainty about the true value of the likelihood than the true value of the prior.

```
dist$likelihood <- rnorm(n = 200, mean = 2, sd = 1)
dist %>% ggplot() +
  geom_histogram(mapping = aes(x = likelihood, y = after_stat(density)),
                 fill = "yellow2", alpha = 0.6, color = "black", binwidth = 0.5) +
  stat_function(fun = dnorm, args = list(mean = mean(dist$likelihood), sd = sd(dist$likelihood)))
```



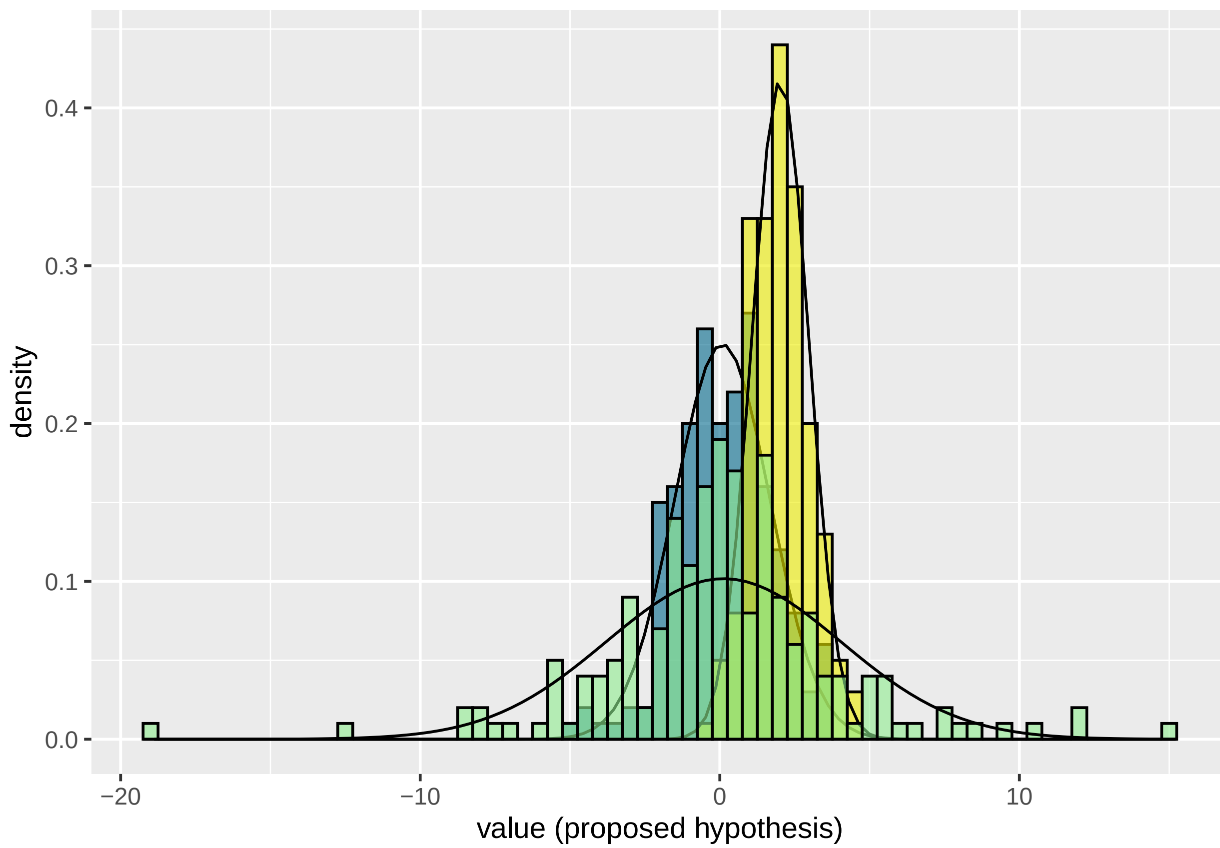
Finally, we calculate the posterior distribution by multiplying the likelihood distribution by the prior distribution. Note that we do not divide by a normalizing constant here, as such a constant has no meaning in this simplified simulation. Omitting this constant should not change our interpretation of the posterior distribution.

```
dist$posterior <- dist$likelihood * dist$prior
dist %>% ggplot() +
  geom_histogram(mapping = aes(x = posterior, y = after_stat(density)),
                 fill = "lightgreen", alpha = 0.6, color = "black", binwidth = 0.5) +
  stat_function(fun = dnorm, args = list(mean = mean(dist$posterior), sd = sd(dist$posterior)))
```



To help us visually compare the three distributions, we now plot them all together.

```
dist %>% ggplot() +
  geom_histogram(mapping = aes(x = prior, y = after_stat(density)),
                 fill = "deepskyblue4", alpha = 0.6, color = "black", binwidth = 0.5) +
  stat_function(fun = dnorm, args = list(mean = mean(dist$prior), sd = sd(dist$prior))) +
  geom_histogram(mapping = aes(x = likelihood, y = after_stat(density)),
                 fill = "yellow2", alpha = 0.6, color = "black", binwidth = 0.5) +
  stat_function(fun = dnorm, args = list(mean = mean(dist$likelihood), sd = sd(dist$likelihood))) +
  geom_histogram(mapping = aes(x = posterior, y = after_stat(density)),
                 fill = "lightgreen", alpha = 0.6, color = "black", binwidth = 0.5) +
  stat_function(fun = dnorm, args = list(mean = mean(dist$posterior), sd = sd(dist$posterior))) +
  xlab("value (proposed hypothesis)")
```



We observe that the posterior distribution (green) has a mean relatively close to that of the prior distribution (blue). However, since the posterior integrates newly observed data, i.e., the likelihood, with the prior, we see that our confidence in many of the values of H, the proposed hypothesis, has decreased. In other words, the posterior distribution has noticeably greater variance than either of the other two distributions. When we observe new data that contradicts our prior beliefs, we become more uncertain about what the true hypothesis is.

This simulation illustrates just one way in which Bayes' theorem can play out. We encourage you to play around with the distributions' parameter values, then see what happens when you apply Bayes' theorem.

Applications of Bayes' theorem

When we conduct Bayesian data analysis in practice, we typically use more advanced methods that build on the basics of Bayes' theorem outlined in the preceding section. Here we give an overview of three such applied Bayesian methods. Each method described here has its own dedicated tutorial in this series, describing the method's underlying philosophy, mathematics, and practical implementation.

Bayesian parameter estimation

One relatively straightforward and widely applicable Bayesian statistical method is **Bayesian parameter estimation**. Bayesian parameter estimation is a Bayesian alternative to frequentist model fitting, and it can be used to estimate various kinds of models. Whereas in frequentist analysis we estimate specific parameter values and provide confidence intervals based on standard errors, Bayesian parameter estimation returns a probability distribution over each estimated parameter in a model (Kruschke, 2010). To generate these posterior distributions, we first specify a prior distribution over each parameter, often guided by prior literature or knowledge. We then run a Bayesian sampling algorithm, which allows the model to learn from the data. There are various methods available for interpreting the results of a Bayesian regression to test hypotheses and explore relationships among variables.

Bayesian networks

When we are interested in modeling a complex system of variables, we might consider using a **Bayesian network model**. A Bayesian network consists of a collection of variables, each probabilistically taking on different values, and probabilistic linkages among these variables. This system is represented as a directed acyclic graph (DAG) in which the nodes are variables and the edges are relationships among the variables (Ben-Gal, 2008). Given a set of multivariate data, we can learn a Bayesian network from these data. Once we have generated a Bayesian network, we can modulate the values of particular nodes to make inferences and test hypotheses about the system modeled by the network. Bayesian networks can be especially helpful when we want to understand how several events or features affect one another and estimate the strength of the probabilistic relationships among them.

Bayesian cognitive modeling

A **Bayesian cognitive model** is a computational model that aims to simulate human cognition by representing one's understanding of the world as probabilistic (or Bayesian) inference using abstract world knowledge and evidence (Tenenbaum et al., 2011). In such models, we posit that in a given scenario, a person first specifies a prior distribution over possible states of the world (Lee & Wagenmakers, 2013). They then consider newly observed evidence, which is often noisy, and use this to update their belief distribution over possible states of the world. As more observations are made, the model can be sequentially updated to reflect this new knowledge in the posterior distribution. In their most literal interpretation, Bayesian cognitive models assert that the human mind reasons and learns using probabilistic simulations of the world. At the very least, these models assume that Bayesian inference is a good approximation of how the human mind operates.

Philosophical merits of Bayesianism

The frequentist approach is sometimes incongruent with how we conceptualize and measure social and behavioral phenomena–which are often inherently subjective and thus imperfectly measured. In such cases, we may want to consider the Bayesian alternative.

For example, the use of a prior probability distribution in Bayesian inference allows us to capture the prior knowledge available to a researcher in a particular operationalization of a construct. This may include knowledge of the validity of measurement tools or of the results of previous trials. As additional data are collected, the posterior distribution can be continually updated to reflect this new knowledge and the researcher's degree of belief in this knowledge.

Additionally, the use of probability distributions, in contrast to point estimates of probability, helps more fully capture the lingering uncertainty surrounding a hypothesis after the analysis has concluded. Instead of concluding, for example, that hypothesis H_1 fits the data well, we can set forth the more comprehensive evaluation that we have a strong, but not absolute, degree of belief in H_1 and weaker degrees of belief in additional hypotheses H_2 and H_3.

Further reading

If you would like to learn more about the Bayesian approach, here are a few external resources:

- Contrasting frequentist and Bayesian statistics: https://www.austincc.edu/mparker/stat/nov04/talk_nov04.pdf
- Moving from Bayes' theorem to Bayesian modeling: https://www.nature.com/articles/s43586-020-00001-2
- General Bayesian philosophy/epistemology: https://plato.stanford.edu/entries/epistemology-bayesian/

References

Ben-Gal, I. (2008). Bayesian networks. In F. Ruggeri, R.S. Kenett, & F.W. Faltin (Eds.), *Encyclopedia of Statistics in Quality and Reliability*. https://doi.org/10.1002/9780470061572.eqr089

Hájek, A. (2023). Interpretations of probability. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023). Metaphysics Research Lab, Stanford University.

Kruschke, J. K. (2010). Bayesian data analysis. *WIREs Cognitive Science*, *1*, 658–676. https://doi.org/10.1002/wcs.72

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Romeijn, J.-W. (2022). Philosophy of statistics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331* (6022), 1279–1285. https://doi.org/10.1126/science.1192788

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686.