Numeracy
Advancing Education in Quantitative Literacy

Volume 14 | Issue 1 Article 5

2021

COVID-19 and Quantitative Literacy: Focusing on Probability

Michael A. Lewis
Silberman School of Social Work at Hunter College, michael.a.lewis@hunter.cuny.edu

Recommended Citation

Lewis, Michael A.. "COVID-19 and Quantitative Literacy: Focusing on Probability." *Numeracy* 14, Iss. 1 (2021): Article 5. DOI: https://doi.org/10.5038/1936-4660.14.1.1362

COVID-19 and Quantitative Literacy: Focusing on Probability

Abstract

The COVID-19 pandemic is arguably the worst crisis the world has faced, so far, in this new century. We haven't seen a pandemic like this since the *1918 Flu* at the beginning of the last century, and, as of this writing, there appears to be no end in sight. What those of us who're focused on quantitative methods have noticed, in addition to the many people dying, becoming ill, and losing their livelihoods, is the importance of quantitative literacy to an understanding of what's going on. That's what this article is about. Specifically, it's about how the COVID-19 pandemic is illustrating the importance of understanding different aspects of probability theory, particularly conditional probability.

Keywords

Covid-19, probability, conditional probability, Bayes' theorem, diagnostic testing

Cover Page Footnote

Michael Anthony Lewis is Professor at the Silberman School of Social Work at Hunter College. He is the co-author of *Economics for Social Workers*, co-editor of *The Ethics and Economics of the Basic Income Guarantee*, and author of *Social Workers Count*.

Introduction

One of the things the unfolding COVID-19 pandemic makes clear is the importance of quantitative literacy. Irfan points this out in a very informative recent piece on the disease (2020). For example, if a nation's residents are to take seriously public health experts' call for social distancing, it might help if they have a basic understanding of exponential growth. Irfan's article is meant to help provide that understanding. Not only does the COVID-19 pandemic provide a mathematics lesson on exponential growth; it also provides one on probability.

For one thing, not all people exposed to the virus end up getting infected; instead, there is some non-zero probability that they'll become infected. Also, if a person becomes infected with COVID-19, they won't necessarily die from it; instead, there is some non-zero chance that they will. In fact, the chance that someone will die from COVID-19, given that they've been infected, is an example of a special type of probability called a *conditional probability*.

Generally, a conditional probability is the probability that one thing happens, given that something else has happened. Another example of a conditional probability is the chance that someone ends up being hospitalized with COVID-19, given that they've been infected. Yet a third example, is the probability that someone ends up on a ventilator, given that they've been hospitalized.

It should be clear from these examples that exponential growth isn't all there is to pandemics such as COVID-19; probabilities are just as relevant. Therefore, understanding probabilities is key to understanding this pandemic, as well as related matters, such as diagnostic testing. The rest of this paper elaborates on these points.

What Is Probability?

Even though the term *probability* is often used, there's disagreement on what it means (Hacking 2001; McGrayne 2011). To some, probability is a relative frequency or, more formally, the limit of a relative frequency (Von Mises 1957). That is, probability for these folks is thought of as

$$\frac{\text{no. of times some event occurs}}{\text{no. of times the event could possibly occur}}$$

or the number this fraction approaches as the number of times the event in question could occur approaches infinity.

To others, probability is degree of belief in a proposition about the occurrence of some event (Savage 1972). Suppose we consider the proposition "more people will die from the current COVID-19 pandemic than died during the Spanish flu pandemic of the early 20th century." Assume that one person says that there is a

20% chance that this statement is true, while a second person claims that this chance is 50%. The second person believes in this proposition to a greater degree than does the first one.

Even though there are disagreements about what *probability* means, mathematicians and statisticians widely agree on how probabilities must behave mathematically. Based on work by Andrey Kolmogorov (1956), mathematicians now define *probability* in terms certain axioms and are able to prove various theorems about probabilities on the basis of these axioms. This more mathematical treatment applies whether one thinks of probabilities as relative frequencies or degrees of belief.

An important distinction, relevant to this paper, that mathematicians make is between *unconditional* and *conditional probabilities*. Let $P(A)$ stand for the probability that event A occurs and $P(A \mid B)$ stand for the probability that event A occurs, given that another event B also occurs. Then $P(A)$ would be the unconditional probability of A and $P(A \mid B)$ would be the conditional probability of A, that is the probability of A conditional on the occurrence of B. By using "event language," for the purposes of this paper, I'm coming down on the side of the relative frequentist interpretation of probability (Hacking 2001). That is because for the examples I'll use later, it's most natural to discuss them in terms of that interpretation.

Expressed as fraction, $P(A)$ is

$$\frac{\text{no. of times event A occurs}}{\text{no. of times event A could possibly occur}}$$

and $P(A \mid B)$ is

$$\frac{\text{no. of times event } A \text{ occurs}}{\text{no. of times event } A \text{ could possibly occur, given that } B \text{ has occured}}$$

With these preliminaries out of the way, let's turn to how COVID-19 provides lessons in probability theory.

Probability, the Case Fatality Ratio, and the Infection Fatality Ratio

As of April 13, 2020, there were 554,849 confirmed cases of COVID-19 in the U.S. and 21,942 deaths from the disease (Centers for Disease Control and Prevention [CDC] 2020a). In the language of probability theory, a death from COVID-19 is an event of interest. The 554,849 confirmed cases of COVID-19 are the number of deaths, up to that time, that could have occurred, so we can estimate the probability of dying from the disease conditional on being diagnosed as

DOI: https://doi.org/10.5038/1936-4660.14.1.1362

$$\frac{21{,}942}{554{,}849} = \text{about } 0.04 \text{ or } 4\%$$

Based on these data, if someone in the United States is diagnosed with the disease, they have about a 4% chance of dying from it. Epidemiologists call this particular probability the *case fatality ratio* (CFR).

When it comes to estimating how deadly a disease like COVID-19 is, the CFR suffers from a serious problem. For one thing, the actual number of COVID-19 cases, as well as the number of deaths from COVID-19, are likely to be underestimated (CDC 2020b; Howard 2020). An underestimate of the true number of COVID-19 cases means that the denominator of COVID-19's CFR is larger than we think. For a given value of the numerator, a larger denominator means that COVID-19's CFR is actually smaller than we think. That is, COVID-19 isn't actually as deadly as it appears. An underestimate of the number of deaths from COVID-19 means that the numerator is higher than we think. For a given value of the denominator, a large numerator means that COVID-19's CFR is actually larger than we think, so COVID-19 is more deadly than it appears to be.

A key reason we might be underestimating the denominator of COVID-19's CFR has to do with testing. In order to get diagnosed with COVID-19 and, therefore, become part of the denominator of the CFR, one has to test positive for the condition, and of course, to test positive, one must first be tested. A number of people infected with the virus are mild or asymptomatic cases, which often means that they don't get tested. Since they don't get tested, they can't become part of the CFR denominator, and therefore, the denominator of COVID-19 appears smaller than it actually is (CDC 2020b).

The reason we may be underestimating COVID-19 deaths is, in part, also related to testing. Some people who died from COVID-19 weren't recorded as having done so because they died at home and, therefore, were never tested and diagnosed with the disease. Others who died in hospitals from COVID-19 were recorded as having died from something else. Also, there is often a lag time between the point at which someone dies from COVID-19 and the point at which their death is recorded as such. So, at any given point in time, there may be people who died from COVID-19 but who don't appear in the numerator of the CFR because those deaths haven't been recorded yet.

Even though we may be underestimating both the number of cases, as well as the number of deaths, there is reason to believe that the underestimate of the number of cases is much larger than the underestimate of the number of deaths. According to Weinberger et al. (2020), COVID-19 deaths may be 28% higher that the number of recorded deaths. Havers et al. (2020) tell us that some areas underestimated the number of COVID-19 cases by as low as a factor of 6, while others did so by as high a factor of 24. Since the underestimate of cases appears to be much higher than the underestimate of deaths, the CFR is likely overestimating how deadly the disease is.

To get a sense of how large this overestimate might be, let's increase the number of deaths in the numerator of our earlier CFR calculation by 28% and our denominator by a factor of 6:

$$\frac{21{,}942*(1.28)}{554{,}849*(6)} = \text{about } 0.008 \text{ or } 0.8\%$$

Recall that the CFR was about 0.04 or 4%. The calculation directly above results in a much lower fatality rate of 0.8%. Since this fatality rate is based on the estimated number of *actual infections*, rather than the number of diagnosed cases, it's called the *Infection Fatality Ratio*.

Probability and COVID-19 Testing

As I said earlier, asymptomatic or mild cases of COVID-19, because they're not tested, may escape diagnosis. Obviously, the main reason we want to test people is to see if they have the disease. However, the probability that someone actually has COVID-19, given that they've tested positive for it depends on the sensitivity of the test, the specificity of the test, and the prevalence of the disease.

The *sensitivity* of a test is the probability that a person tests positive, given that they have the disease, while the *specificity* is the probability that they test negative, given that they are not infected. That is, the sensitivity and specificity of diagnostic tests are defined in terms of conditional probabilities. The prevalence of a disease is the overall proportion of the population which has it; it can be thought of as the probability that someone in the population has the disease.

The main problems with any diagnostic test are *false positives* and *false negatives*. Regarding COVID-19, a false positive would occur if someone without the disease tested positive for it; a false negative would occur if someone with the disease tested negative. I mentioned earlier that the probability that someone has COVID-19, given that they've tested positive for it depends on test sensitivity and specificity, as well as the disease's prevalence. This can be shown using a theorem from probability theory called *Bayes' theorem*.

Bayes' theorem takes the following form:

$$P(A \mid B) = \frac{P(B \mid A)*P(A)}{P(B)}$$

and

$$P(B) = P(B \mid A)*P(A) + P(B \mid -A)*P(-A)$$

DOI: https://doi.org/10.5038/1936-4660.14.1.1362

where -*A* stands for *the probability that* A *does not occur*. In Bayes' theorem, *P(A)* is sometimes called the *base rate*. Let's rewrite Bayes' theorem using the relevant probabilities for COVID-19 testing:

$$P(\text{COVID-19} \mid \text{Positive}) = \frac{P(\text{Positive} \mid \text{COVID-19}) * P(\text{COVID-19})}{P(\text{Positive} \mid \text{COVID-19}) * P(\text{COVID-19}) + P(\text{Positive} \mid \text{No COVID-19}) * P(\text{No COVID-19})}$$

Within this context, *P*(Positive│COVID-19*)* is the sensitivity of the test, *P*(COVID-19*)* is the prevalence of the disease, and *P*(Positive│No COVID-19*)* is the probability of a false positive. *P*(Positive│No COVID-19*)* is also equal to 1 minus the specificity of the test. *P*(COVID│Positive*)* is the probability that someone has COVID-19, given that they've tested positive for it; it's also called the *positive predictive value* (PPV) of the test. See Boersma and Willard (2008) for further discussion of the relevance of probability theory to medical testing.).

I don't know the actual figures for the sensitivity and specificity of COVID-19 tests; no doubt, these will vary depending on the particular tests we're talking about. The point I want to make, however, applies to any COVID-19 test, so it's fine for me to use hypothetical numbers.

Suppose we have a test with a sensitivity and specificity of 99%. Since the probability of a false positive is 1 minus the specificity of the test, which in this case is 1 minus 0.99, the probability of a false positive is 0.01 or 1%. Suppose also, for the sake of discussion, that the number of *actually infected* cases is 554,849. The prevalence of COVID-19 could then be estimated by dividing 554,849 by the total US population of 329.4 million. That would come out to about 0.17%. We can now fill in Bayes' theorem:

$$P(\text{COVID-19} \mid \text{Positive}) = \frac{0.99*0.0017}{(0.99*0.002 + 0.01*0.9983)} = 0.14$$

So the chance that someone who tested positive actually has the illness would only be about 14%.

Suppose we assume that the 554,849 figure underestimates the number of COVID-19 cases and that, in fact, there are 6,588,000 cases. The prevalence would then be 6,588,000/329.4 million or 2%. This implies that the probability of someone not being infected would decrease to 98%. Now Bayes' theorem would look like this:

$$P(\text{COVID-19} \mid \text{Positive}) = \frac{0.99*0.02}{(0.99*0.02 + 0.01*0.98)} = 0.67$$

We see that there would now be a much greater chance (67%) that someone who tested positive for COVID-19 actually has the disease.

Now suppose we assume that 55,998,000 people in the United States currently have COVID-19, bringing prevalence up to about 17% and the probability of someone not having it down to about 83%. Bayes' theorem would now be:

$$P(\text{COVID-19} \mid \text{Positive}) = \frac{0.99*0.17}{(0.99*0.17 + 0.01*0.83)} = 0.95$$

With a prevalence this high, 95 out of every 100 people who tested positive for COVID-19 would actually have it.

Since sensitivity, specificity (indirectly as the probability of a false positive), and prevalence all appear in Bayes' theorem, we see how the PPV of a COVID-19 test (or any diagnostic test) depends on these three quantities. But a key takeaway from these calculations is that the PPV depends heavily on how prevalence is related to the chance of a false positive. If prevalence is less than the chance of a false positive, the PPV of a test will be relatively low. But if prevalence is greater than the probability of a false positive, PPV will be relatively high.

The *negative predictive value* (NPV) of a test is the probability that someone who tested negative actually doesn't have the disease or, if it's an antibody test, doesn't have antibodies to it. It can be shown that if the prevalence of a disease is less than the probability of a false positive, the NPV of the test will be relatively high. But if prevalence is greater than the chance of a false positive, the NPV of the test will be relatively low.

In the midst of a deadly global pandemic, one could ask which is more important: the PPV or NPV of a test. This moves us into the realm of decision making. When schools, businesses, and other organizations reopen, some of them will probably, in an effort to keep their employees and clients/consumers safe, implement some type of testing regimen. It's conceivable that at least some of them will reason in the following way.

If people test positive, they will be sent home or told not to come to work. If the PPV of the test is low, then there is a good chance that many of the people sent home or told not to come to work won't actually be positive. If someone tests negative, they will be allowed to remain at or come to work. If this is done when the NPV of the test is low, there is a good chance that many of the people allowed to come to work, or remain there, will be infected. Which is worse: keeping people home when they don't really have the disease or allowing people to work when they do? These are the kinds of questions employers, businesses, and even everyday citizens/residents will have to think about; familiarity with Bayes' theorem might help them do so more carefully.

Probability and COVID-19 with Natural Frequencies

The psychologist Gerd Gigerenzer (2002) is well known for his work on presenting probabilities in terms of natural frequencies instead of percentages or proportions. Everything from the previous section can be discussed in terms of such frequencies. That is, everything can be discussed in terms of relationships between magnitudes instead of percentages or proportions.

Let's translate the first calculation from the previous section into natural frequencies using Table 1 as a starting point. The entry at the top is the US population of 329,400,000 people, and 554,849 of them have COVID-19. Of these 554,849, there are 549,301 (99% sensitivity) who tested positive for the disease, while 5,548 people tested negative. Moving on, 328,845,151 of the 329,400,000 people in the United States do not have COVID-19. Out of these 328,845,151, there are 3,288,452 (1% false positive probability) who tested positive for it, while 325,556,699 (99%) tested negative.

Recall that the positive predictive value of a test is the probability of actually having the disease, given that one tested positive for it. Once natural frequencies are presented, it becomes easy to figure this out. Start with the total number of positive cases: 549,301 + 3,288,452 = 3,837,753. Since 549,301 of these positive cases actually have COVID-19, if we divide 549,301 by 3,837,753, we obtain the chance that someone has COVID-19, given that they tested positive for it—the PPV of the test—which comes out to about 0.14 or 14%, which is the same we found in the previous section using Bayes' theorem.

Table 1
COVID-19 Cases and Infections in US Population with Low Prevalence

US Population	329,400,000
COVID-19	554,849
Positive	549,301
Negative	5,548
No COVID-19	328,845,151
Positive	3,288,452
Negative	325,556,699

Table 2
COVID-19 Cases and Infections in US Population with Medium Prevalence

US Population	329,400,000
COVID-19	6,588,000
Positive	6,522,120
Negative	65,880
No COVID-19	322,812,000
Positive	3,228,120
Negative	319,583,880

Natural frequencies can be used to gain insight into the other calculations from the previous section. See Table 2 for the assumption that 6,588,000 folks in the US population have COVID-19, while 322,812,000 of them don't. To figure out the PPV given these numbers, we divide 6,522,120 by (6,522,120 + 3,228,120). The result is about 0.67 or 67%, the same as we found in the previous section.

Table 3
COVID-19 Cases and Infections in US Population with High Prevalence

US Population	329,400,000
COVID-19	55,998,000
Positive	55,438,020
Negative	559,980
No COVID-19	273,402,000
Positive	2,734,020
Negative	270,667,980

A natural frequencies version of the third calculation from the previous section can be obtained by using Table 3. This time we obtain the PPV by dividing 55,438,020 by (55,438,020 + 2,734,020); the result is about 0.95 or 95%, just as we obtained in the previous section.

The use of natural frequencies doesn't provide any new information. It's just a different way of presenting the same information, but one which may reduce the chance of mistakes. In his book *Calculated Risks*, Gigerenzer (2002) provides a great deal of evidence indicating that people have a hard time thinking in terms of Bayes' theorem. He also provides evidence that when folks think about Bayes' theorem in terms of natural frequencies, they tend to make fewer mistakes.

Probability and COVID-19 Testing with Venn Diagrams

Another way to make Bayes' theorem easier to understand is to use pictures, specifically Venn Diagrams. Consider the Venn diagram Figure 1:

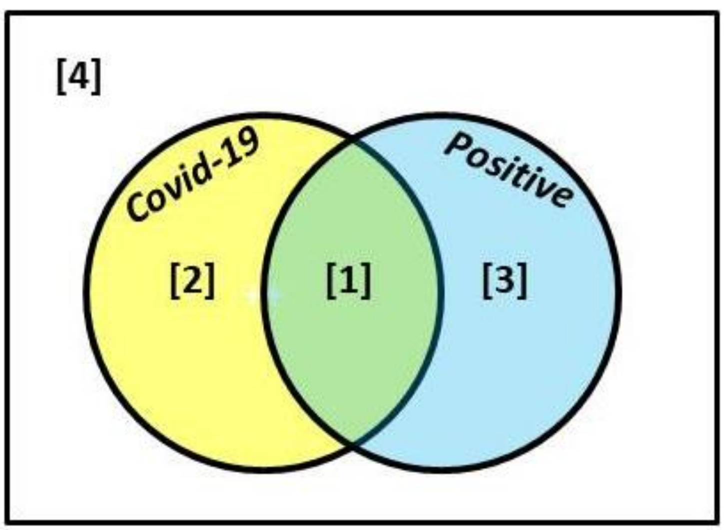

Figure 1. Venn diagram of possible cases for: Has COVID-19? Tested Positive? [1] Yes, Yes; [2] Yes, No; [3] No, Yes; [4], No, No

Now consider Bayes' theorem again:

$$P(\text{COVID-19} \mid \text{Positive}) = \frac{P(\text{Positive} \mid \text{COVID-19}) * P(\text{COVID}-19)}{P(\text{Positive} \mid \text{COVID-19}) * P(\text{COVID-19}) + P(\text{Positive} \mid \text{No COVID-19}) * P(\text{No COVID-19})}$$

In Figure 1, the sum of areas 1, 2, 3, and 4 represent all US residents. The sum of areas 1 and 2 represent all of those with COVID-19, the total of areas 1 and 3 represent all of those who tested positive for COVID-19, and area 1 represents all those with COVID-19 who tested positive for the disease.

Here are the relevant probabilities for Bayes' theorem stated in terms of the Venn Diagram:

$$P(\text{COVID-}19|\text{Positive}) = \frac{[1]}{[1]+[3]} = \text{PPV of the test}$$

$$P(\text{Positive}|\text{COVID-}19) = \frac{[1]}{[1]+[2]} = \text{sensitivity of test}$$

$$P(\text{Positive}|\text{No COVID-}19) = \frac{[3]}{[3]+[4]} = \text{probability of a false positive}$$

$$P(\text{COVID-}19) = \frac{[1]+[2]}{[1]+[2]+[3]+[4]} = \text{Prevalence of the disease}$$

$$P(\text{No COVID-}19) = \frac{[3]+[4]}{[1]+[2]+[3]+[4]}$$

$$P(\text{Positive}) = \frac{[1]}{[1]+[2]} * \frac{[1]+[2]}{[1]+[2]+[3]+[4]} + \frac{[3]}{[3]+[4]} * \frac{[3]+[4]}{[1]+[2]+[3]+[4]} = \frac{[1]+[3]}{[1]+[2]+[3]+[4]}$$

In $P(\text{Positive})$, the expression after the second equality sign follows because [1] + [2] and [1] + [3] in the expression after the first equality sign both cancel. We can now plug all these quantities into Bayes' theorem:

$$\frac{[1]}{[1]+[3]} = \frac{\frac{[1]}{[1]+[2]} * \frac{[1]+[2]}{[1]+[2]+[3]+[4]}}{\frac{[1]+[3]}{[1]+[2]+[3]+[4]}} =$$

$$\frac{[1]}{[1]+[2]} * \frac{[1]+[2]}{[1]+[2]+[3]+[4]} * \frac{[1]+[2]+[3]+[4]}{[1]+[3]}$$

The sums [1] + [2] and [1] + [2] + [3] + [4], after the second equality sign, both cancel; so we're left with $\frac{[1]}{[1]+[3]}$. Taking a look again at the Venn Diagram above, we see that [1] + [3] are the folks who tested positive for COVID-19, while [1] are those who tested positive who actually have the disease. This means that $\frac{[1]}{[1]+[3]}$ is the proportion of those who tested positive who actually have the disease. That is, $\frac{[1]}{[1]+[3]}$ is the positive predictive value of the test, which is exactly what we want.

Conclusion

As I write these lines in April of 2020, COVID-19 is wreaking havoc on the world. Whether we'll be able to control it will depend on good decision making, and good decision making will depend on an understanding of probability theory, especially Bayes' theorem. This paper has been an attempt to graphically illustrate this point.

Acknowledgment

I'd like to thank Nathan D. Grawe, Len Vacher, and two anonymous reviewers for their very helpful comments on this paper. Any remaining mistakes are my responsibility alone.

References

Boersma, Stuart and Teri Willard. 2008. "False Positives and Referral Bias: Content for a Quantitative Literacy Course." *Numeracy* 1(2): Article 5. https://doi.org/10.5038/1936-4660.1.2.5

Centers for Disease Control and Prevention. 2020a. "Cases in the U.S." https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html, retrieved April, 13.

Centers for Disease Control and Prevention. 2020b. "Study: Actual Number of COVID-19 Cases Vastly Underestimated." https://www.drugtopics.com/latest/study-actual-number-COVID-19-cases-vastly-underestimated, retrieved 4/30/20.

Gigerenzer, Gerd. 2002. *Calculated Risks: How to Know When Numbers Deceive You*. New York: Simon and Schuster.

Hacking, Ian. 2001. *An Introduction to Probability and Inductive Logic*. Massachusetts: Cambridge University Press. https://doi.org/10.1017/CBO9780511801297

Havers, Fiona P., Carrie Reed, Travis Lim, Joel M. Montgomery, John D. Klena, Aron J., Alicia M. Fry, Deborah L. Cannon, Cheng-Feng Chiang, Aridth Gibbons, Inna Krapiunaya, Maria Morales-Betoulle, Katherine Roguski, Mohammad Ata Ur Rasheed, Brandi Freeman, Sandra Lester, Lisa Mills, Darin S. Carroll, S. Michele Owen, Jeffrey A. Johnson, Vera Semenova, Carina Blackmore, Debra Blog, Shua J. Chai, Angela Dunn, Julie Hand, Seema Jain, Scott Lindquist, Ruth Lynfield, Scott Pritchard, Theresa Sokol, Lynn Sosa, George Turabelidze, Sharon M. Watkins, John Wiesman, Randall W. Williams, Stephanie Yendell, Jarad Schiffer, Natalie J. Thornburg. 2020. "Seroprevalence of Antibodies to SARS-Cov-2in 10 Sites in the United States, March 23–May 12, 2020." *JAMA Internal Medicine*, Published online July 21, 2020. https://doi.org/10.1001/jamainternmed.2020.4130

Howard, Jacqueline. 2020. "U.S. Coronavirus Death Count Likely and Underestimate. Here's Why." https://www.cnn.com/2020/04/06/health/us-coronavirus-death-count-cdc-explainer/index.html, retrieved 4/30/20.

Irfan, Umair. 2020. "The Math Behind Why We Need Social Distancing, Starting Right Now: How Just one Case of Coronavirus Could Lead to Thousands More if we Don't Limit Social Contact." *Vox*, March 15, https://www.vox.com/2020/3/15/21180342/coronavirus-COVID-19-us-social-distancing, retrieved 4/30/20.

Kolmolgorov, A.N. 1956. *Foundations of the Theory of Probability*. New York: Chelsea Publishing Company.

McGrayne, Sharon Bertsch. 2011. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Connecticut: Yale University Press.

Savage, Leonard J. 1972. *The Foundations of Statistics*. New York: Dover Publications, Inc.

Von Mises, Richard. 1957. *Probability, Statistics, and Truth*. New York: Dover.

Weinberger, Daniel M., Jenny Chen, Ted Cohen, Forrest W. Crawford, Farzad Mostashari, Don Olson, Virginia E. Pitzer, Nicholas G. Reich, Marcus Russi, Lone Simonsen, Anne Watkins, Cecile Viboud. 2020. "Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020," *JAMA Internal Medicine* 180(10):1336–1344. https://doi.org/10.1001/jamainternmed.2020.3391