Examining Algorithms

Nathaniel Forde

2021-03-03

Contents

Chapter 1

Preface

What this book is

Chapter 2

The Average Man

In the 1840s the Average man stalked the nightmares of Augustin Cournot. The mathematician was haunted by a melange of imagined parts. A Frankenstein's monster of mismatched limbs, variously soldered, stapled or sown at the joints. He worried that no one but a "physical monstrosity" could exhibit the average, weight, height and other mean attributes in one body. (Stigler, 2016) But fantastical fears were no bar to a useful mathematical fiction. The notion of averaging was a technological breakthrough - applications of averaging multiplied without cease: polling, gambling, forecasting - statistics were recorded everywhere, compounding one on another; averages of averages tenuously tethered to observable facts by layers of abstraction and scales of measurement.

Slowly, doubts began to creep back into the statistics. In the 1970s the psychologist Paul Meehl would worry that such brute approximations had stifled the development of the softer sciences and contributed to the mis-measurement of man. He would go on to elaborate twenty features of psychological science which made such measurement constructs inapt, unreliable and difficult to clearly falsify. This was progress! (Meehl, 1978) He then showed that the methods of validating such constructs were the main culprit and the likely cause of psychology's implausible claims to concrete, replicable results. Around thirty years later some of the same replication issues would come to be called a crisis. Cournot was right to fear.

Seen from the perspective of a patient the difficulties are more urgent and stark. In 2019 Esme Yang would write with hope of the comfort given by a diagnosis, the knowledge that she was not "pioneering an inexplicable experience."[1] To find yourself described in the DSM provides: a framework and the glimmer of a cure, a chance to slot yourself into a category, a community of care and a medicinal regime. Frustrating then when the categories themselves are in flux. Diagnostic

[1] All remarks on Schizophrenia in the following draw on the precise and personal essays in (Wang, 2019)

criteria arrayed over page after page attempting to capture the elusive core of a psychological dysfunction. Descriptions are derived from clinical interviews correlated and meshed with genetic markers, then poorly mapped to a family of ailments. Neither clearly bipolar disorder nor manic depression, schizophrenia, psychosis or schizo-affective disorder. But an idiosyncratic presentation (as they all must be), uniquely felt and individually suffered.

Wang's collection of schizophrenias defy easy taxonomy "because there are just so many different ways in which people can develop a syndrome that looks like schizophrenia … as we now define it." The task then becomes one of coping with uncertainty and adjusting expectations. It's not so exact a science that you can measure twice and cut once. The measurement schemes change and you need multiple cuts. You measure out the impact of treatments and the trade-offs - what's non-negotiable and how many side-effects are acceptable? What works for you versus what's advised by the professionals. Sterile cost-benefit considerations become suddenly dramatic and life changing.

2.1 …and Expected Value

There is an algorithm beloved by bureaucrats. An unsung hero of administrators and accountants. An algorithm both ubiquitous and under appreciated. It's pivotal for nearly every business and informs the actions of tech firms and policy makers the world over. It is only mildly hyperbolic to say that understanding this formula unlocks wealth and power. The algorithm lies at the heart of online A/B testing, all policy analysis, sound strategy and poker play.

$$EV(O)_p = p_1 u(o_1) + p_2 u(o_2) + ... + p_k u(o_k)$$

The expected financial value of a random process is just the sum of the utility (typically dollar outcomes) weighted by their probabilities. Outcomes can vary from deals of cards, to customer transactions, election results, continued sanity. Pascal can argue sincerely that such considerations compel even belief in God. The infinite downsides of hell at any likelihood ought to compel even the cynical sceptic. But the formula, glossed as a rule for rational action, merits your attention for more mundane wagers too. If you intend to maximise your expected value, the meaning of probability is not an idle concern. [2] While statistics are often tortured to rubber stamp decisions and probabilities are abused to fit policy prescriptions with false precision, the crisp clarity of the rule has an enduring appeal that promises to sift the murky swamps of Big Data. It's a scalpel that anyone can wield to parse the syntax of statistical jargon and carve answers from an abstract space of probabilities. "What's my expected return? My likely life-quality?" - a simple question, with a surprisingly complex answer.

[2] For a typical example of how EV is used to express decisions under uncertainty see chapter 7 in the textbook (Barber, 2012), but the paradigm owes much of its influence to decision theoretic work of Leonard Savage in (Savage, 1954)

Whenever expected utility is used as a metric of profit there is a risk of pivot, a point where an explanatory model of rational expected action is substituted for more obscure black-box optimisation techniques. The focus switches from modelling the consumer or the patient to modelling returns based on the consumer and the firm's expected value. Swapping a customer catering model for a customer-as-commodity perspective focused on profit in aggregate and customer acquisition. The tendency is common because it's easier and profit often overwhelms all other priorities, but the loss of understanding typically amounts to a longer term net-loss. Shareholders take comfort from increased profit and algorithms deployed at scale, but it's rare that any single algorithm actually or always maximises the available profit. This dynamic enforces a kind of inescapable see-saw motion where the consumer modelling exercise goes through a constant feedback loop. A good model (informal or formal) of human needs and wants feeds a better a predictive model of individual action. When the latter fails we go back to the utility curves and the algorithm of expected value because it is (if not reliably predictive) a rich and deeply explanatory model of human action under uncertainty.

Decision theory is an abstract formalism which purports to account for how you ought to reason under uncertainty, conditional on knowing your own mind, what you want and the likelihood of those outcomes. Wrestling with a diagnosis of schizophrenia you weigh your future in the face of sickness while knowing your mind could fail in the act, knowing, perhaps, you're not yourself. Schizophrenics are involuntarily detained if a medical professional deems them to have lost sufficient "insight"[3] , but the stated dysfunction assumes that these kinds of insight should be transparent when the psychiatric crisis is resolved. We'll see that the level of insight assumed by decision theory is, at best, hard won and far from transparent.

2.2 Probability: Twin Aspects

Probability emerged slowly and with a dual aspect. On one tradition probability refers to the long run tendency of a random process, on another probability is construed as the degree of belief in an outcome. On the first (frequentist) interpretation a probability distribution has certain fixed theoretical characteristics: as in a uniform probability distribution of a fair coin where all outcomes are equally likely, or as with the normal distribution where most outcomes cluster symmetrically about a central average. On the second (Bayesian) reading the characteristics of the probability distribution are learned from the data. The controversy centres around the fact that it's unclear how a frequentist could ascribe probabilities to unique events. Without appeal to a large set of observations (or known theoretical distribution) the claim that an event appears

[3] "The mind has been *taken over*. The mind has *lost the ability to make rational decisions* in (Wang, 2019) pg58

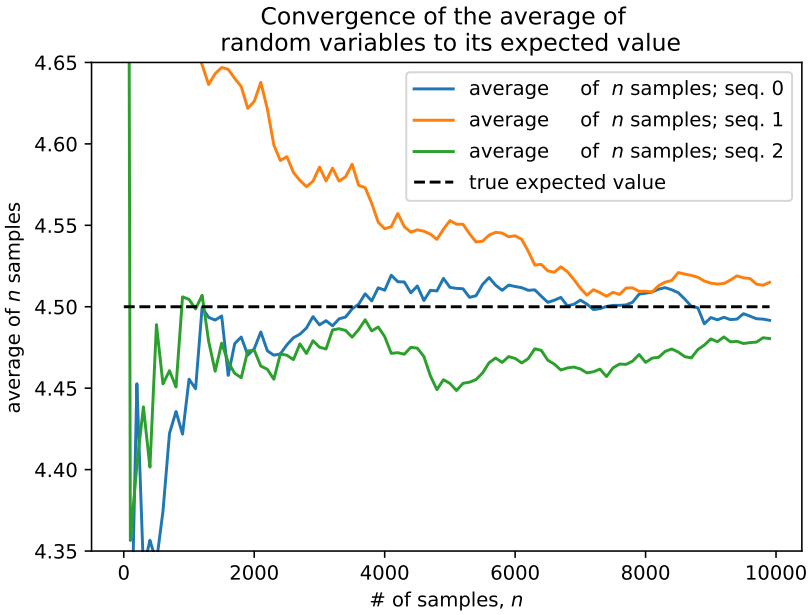
frequently or infrequently is not well defined. Consequently, tabulations of probability appear inappropriate for claims of unique or rare events. In contrast the Bayesian is content to ascribe probabilities to any all partial beliefs no matter how specific. For the Bayesian, the probability calculus is a set of edicts about how to rationally manage and modulate your beliefs. So it's acceptable to have a probabilistic belief in rare cases so long as you update those probabilities with new data when available.

These two approaches are united by the Law of Large numbers which states that as the size of our sample increases our sample average will converge to the expected realisation of the theoretical process.

$$\frac{1}{N}\sum_{i=1}^{N} O_i \text{ converges to } E(O) \text{ as } N \text{ approaches } \infty$$

In this graph we have fixed a Poisson distribution with a mean of 4.5 and can see three examples of how consecutive averaging from the increasing sample sizes results in a closer and closer convergence to the (true) population mean.

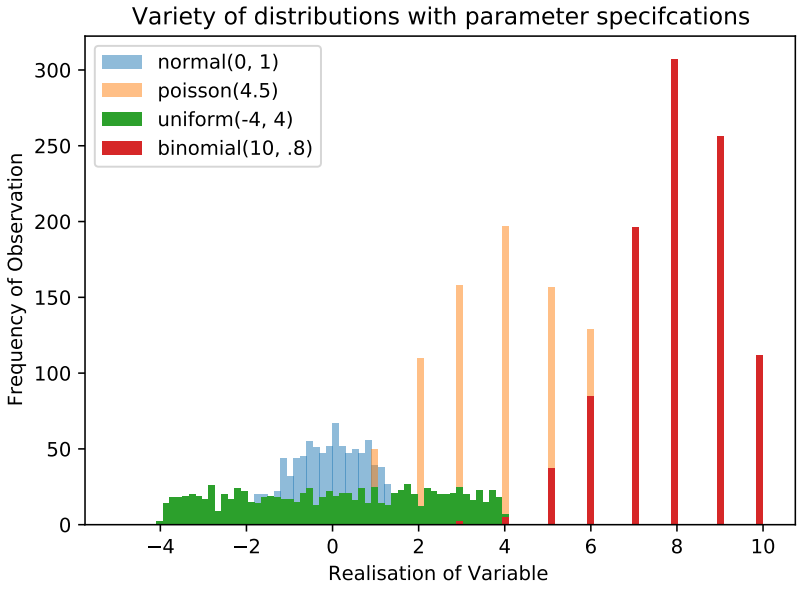
```
# set up the ground truth
np.random.seed(100)
sample_size = 10000
expected_value = lambda_ = 4.5
poi = np.random.poisson
N_samples = range(1, sample_size, 100)

for k in range(3):
    samples = poi(lambda_, sample_size)
    partial_average = [samples[:i].mean() for i in N_samples]
    plt.plot(N_samples, partial_average, lw=1.5, label="average \
    of  $n$ samples; seq. %d" % k)

plt.plot(N_samples, expected_value * np.ones_like(partial_average),
         ls="--", label="true expected value", c="k")

plt.title("Convergence of the average of \n random variables to its \
expected value")
plt.ylabel("average of $n$ samples")
plt.xlabel("# of samples, $n$")
plt.legend()
lims = plt.ylim(4.35, 4.65)
plt.show()
```



Though common knowledge today, in 1650 "the very concept of averaging is [new]… and most people could not observe an average because they did not take averages.(Hacking, 2006) Systematically grappling with the implications of observations is a somewhat recent human endeavour - one which is far from perfected. This tendency is now fundamental to the interpretation of probability. Take a game with fixed and fair odds and we see that repeated play will converge over time because of characteristics which govern the process. Dice are a homely example. In the wild we never know the characteristics which cause the observed spread of outcomes, but such is the influence of gambling on probability, that we assume there exists a stable pattern to be gamed. Partially this is pragmatic, the maths is more tractable if we can assume a well behaved underlying process. The results are compelling: The Normal (Bell Curve) distribution, the Poisson distribution the Bernoulli distribution (to name a few) are all rightly famous. Their shapes are characteristics of innumerable random processes. The distributions cleanly circumscribe and corral likely patterns of events.

```
normal = np.random.normal(0, 1, 1000)
poisson = np.random.poisson(4.5, 1000)
uniform = np.random.uniform(-4, 4, 1000)
binomial = np.random.binomial(10, .8, 1000)

bins = np.linspace(-5, 10, 100)

h1 = plt.hist(normal, bins, alpha=0.5, label='normal(0, 1)')
```

```
h2 = plt.hist(poisson, bins, alpha=0.5, label='poisson(4.5)')
h3 = plt.hist(uniform, bins, label='uniform(-4, 4)')
h4 = plt.hist(binomial, bins, label='binomial(10, .8)')
plt.legend(loc='upper left')
plt.title("Variety of distributions with parameter specifcations")
plt.xlabel("Realisation of Variable ")
plt.ylabel("Frequency of Observation")
plt.show()
```



But the gambling paradigm clouds the fact that in practice we start on the left side of the law of large numbers (with samples) and we often start with small numbers resulting from a unknown number of data-generating processes. Well behaved probability distributions are rare beasts; a tiny fraction of the world's arbitrary menagerie. The fundamental question in probability is not whether probability is a measure of belief or frequency - it is whether we can safely assume that the underlying process adheres to a known model? The Bayesian focus is to try and learn from the new data the expected characteristics of the underlying process, while the Frequentist tries to gauge the accuracy of their assumptions about the underlying process. Both are attempts to validate the structure of the model's theoretical distribution to inform inference. If we can't validate a model, we're better learning what we can from the sample, trusting to wide confidence intervals and worst scenario planning. But in all cases when we need to make a decision, the following questions are inescapable : What are your expectations based on? How do they figure in our choices, and can we use them to improve our outcomes?

2.3 Small Worlds and Statistical Inference

Anissa Weieir and Morgan Geyser, later diagnosed with schizopoty, were in 2017 trailed for the attempted murder of their friend Payton. While believing themselves to be acting at the behest of Slendar Man, they were "willing to forgo even friendship for the sake [their] version of unreality" [4]. Just as in the depth of an obsession reality can be warped by a psychotic focus, a statistical model will accentuates some parts and ignores others.

This narrowness can have devastating effects if deployed without care. But prediction is a visceral need, unavoidable, it precedes the sophistication of probability in any order of analysis. Without some regularity between X, Y we can only interpret their collisions as timorous noise. But even when there is a better than arbitrary correlation between X, Y we'll insist on ascribing a measurable probability to their association. Heuristics will kick in, and confidence will grow out of proportion to the evidence. So whether we view probabilities as a measure of credibility or frequency, the focus is always on a process which in reality reveals a pattern under repetition. Even tenuous connections can be enough to cause havoc. Models, more often than not, are deliberate simplifications of complexity, attempts to formulate a unobserved process within a mathematical structure. Lego-like versions of the world are built and rebuilt, in which we bash parts together to see what reliably sticks. Forget about the pieces we don't own. Count the pairs of blocks that: match, wedge, smash, click-into-place or break, then draw out the ratio of success and the spread of outcomes. Parse out the details of how reds go with greens, and blues with yellows. This is your sample distribution. The expected gain depends on both this uncertainty, measured in this small world, and the finer points of statistical inference.

> A [small world] is… completely satisfactory, only if it is actually a microcosm, that is, only if it leads to a probability measure … that can be written down explicitly pg 88 (Savage, 1954)] .

The danger of shrinking your world comes when you mistake the map for the territory and act on that delusion.

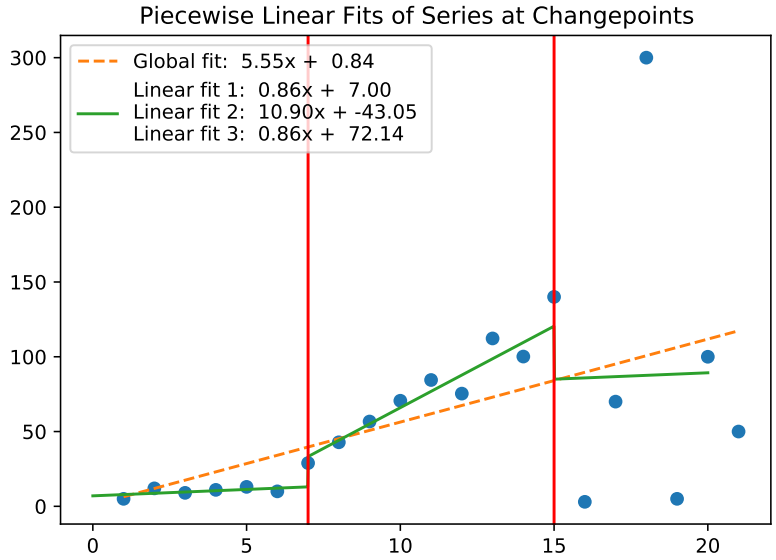
Small worlds are machines for figuring out expected values of a statistical process. We shrink the parameter set to better measure the variance and flux throughout the system. For any hypothetical system there can be multiple plausible approximations of the underlying process which need to be assessed comparatively on their "goodness" of fit. We iterate through new and improved versions, each an attempt to make a conjectural connection between $ X, Y $ mathematically precise. But once built, they embed the errors and assumptions made in their design. McElreath dubs them Golems, primed and then loosed on the world, insensitive to subtlety and context they perform only as instructed. Consider how a basic regression model tries to predict an outcome Y as linear function of some observed features X:

[4] "The Slender Man, the Nothing and Me" in (Wang, 2019)

$$Y = const + \beta X + \epsilon$$

where ϵ is a random variable representing the error (or noise). A modest notational device for disaster. While $const, \beta$ are parameters estimated by an optimisation process to ensure the equation fits the data as neatly as possible. In the plot below we have a series characterised by change. After the first shock we can refit the model so that the line tracks well with the evolving data. After the second shock we try another refit, but the range the and variance of the data makes our basic model a poor fit, i.e. the data no longer exhibits a linear relationship. This presents three examples of error in the modelling process: (i) it's difficult to identify (in the moment) those changepoints in the data which reflect structural change, (ii) the linearity assumptions that go into the model are sound but the parameters need be re-estimated based on the new data and (iii) the third linear model is simply a terrible match for the pattern in the data.

```
### Piecewise Linear Fits

x = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 2
y = np.array([5, 12, 9, 11, 13, 10,
              28.92, 42.81, 56.7, 70.59, 84.47, 75.36, 112.25, 100.14, 140.03, 3, 70, 

def piecewise_linear(x, b1, a1,  b2, a2, b3, a3):
    funcs = [lambda x:b1*x + a1,
             lambda x:b2*x + a2,
             lambda x:b3*x + a3]
    conds = [x < 7, ((x >= 7) & (x < 15)), x > 15]
    return np.piecewise(x, conds, funcs)

p , e = optimize.curve_fit(piecewise_linear, x, y, method="trf")
a, b = polyfit(x, y, 1)
xd = np.linspace(0, 20, 1000)

plt.plot(x, y, "o")
plt.plot(x, a + b * x, '--', label="Global fit: {b: .2f}x + {a: .2f}".format(b=b, a=a)
plt.plot(xd, piecewise_linear(xd, *p), label="Linear fit 1: {b: .2f}x + {a: .2f} \n"
                                             "Linear fit 2: {b1: .2f}x + {a1: .2f} \n"
                                             "Linear fit 3: {b2: .2f}x + {a2: .2f}".fo



plt.axvline(x=7, color='red')
plt.axvline(x=15, color='red')
plt.title("Piecewise Linear Fits of Series at Changepoints")
plt.legend()
plt.show()
```



Every model is a guess as to the implicit order in apparent noise. Sometimes there is no order, and other times the patterns is too subtle for a dumb model to capture. In practice you never really know whether a single new error stems from a misfit but appropriate model or an entirely inappropriate model. As we increase our number of sample fits we hope to better approximate the true linear process (if any) generating the data. Imagine now that the data points in Figure 3 are repeatedly re-speckled over the canvas. We can refit a new model for each set of scattered data points and each refit gives us a new sample values for $ const, $. If the underlying data generating process is stable, then the parameter fits will converge to the correct values of $const, \beta$; correct in the sense that they can be used to draw the line of best fit for the data. A statistically stable process is one that can be modelled with errors ϵ normally distributed around 0, so that the model will be *correct on average* because $E(\epsilon) = 0$. Our predictions will overshoot in some cases but on the whole the errors up and down will cancel each other out.

> "Typically, the assumptions in a statistical model are quite hard to prove or disprove, and little effort is spent in that direction. The strength of empircal claims made on the basis of such modeling therefore does not derive from the solidity of those assumptions. Equally, these beliefs cannot be justified by the complexity of the calculations... These observations lead to uncomfortable questions" (Freedman, 2009)

Forecasting with the parameters of best fit minimises our forecast errors because the fluctuations are stable about the centre of the line. These are the required

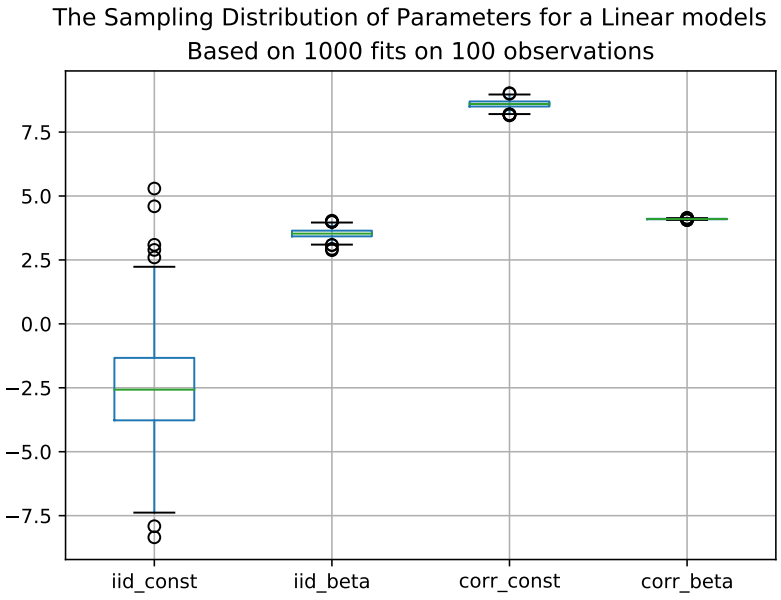
assumptions for a process to exhibit the tendency of regression towards the mean. If they're not met, we will see poor parameter estimates and wild swings away from the linear path. The fundamental statistical assumption here is about the properties of our mistakes! The model is less plausible of is our judgements are made in the grip of a delusion.

2.4 Modeling: Improper Assumptions and Skewed Expectations

Below we build two sampling distributions based on different models of an underlying processes. One in which the errors are independent, normally distributed around 0 and in the other the errors are correlated in a sine-wave like pattern, increasing and decreasing periodically. This is akin to the difference between measuring error when predicting the heights of randomly sampled people versus predicting the sales volumes on randomly selected days of the week. A random sample of daily sales risks clumping weekends together and skewing the expected values. No such risk exists when sampling from independent individuals.

```
#### Sampling Distributions of Linear Fits
#### Build True Model
N = 1000
X = np.random.uniform(0, 20, N)
uncorrelated_errors = np.random.normal(0, 10, N)
correlated_errors = np.random.uniform (0, 10) + np.sin(np.linspace(0, 10*np.pi, N)) \
          + np.sin(np.linspace(0, 5*np.pi, N))**2 \
          + np.sin(np.linspace(1, 6*np.pi, N))**2 + .6*X

Y_corr = -2 + 3.5 * X + correlated_errors
Y = -2 + 3.5 * X + uncorrelated_errors
population = pd.DataFrame({'X': X, 'Y': Y, 'Y_corr': Y_corr})

fits = pd.DataFrame(columns=['iid_const', 'iid_beta', 'corr_const', 'corr_beta'])
for i in range(0, 1000):
    sample = population.sample(n=100, replace=True)
    Y = sample['Y']; X = sample['X'] ; Y_corr = sample['Y_corr']
    X = sm.add_constant(X)
    iid_model = sm.OLS(Y, X)
    results = iid_model.fit()
    corr_model = sm.OLS(Y_corr, X)
    results_2 = corr_model.fit()
    row = [results.params[0], results.params[1], results_2.params[0], results_2.params
    fits.loc[len(fits)] = row
```

```
fits.boxplot()
plt.suptitle("The Sampling Distribution of Parameters for a Linear models")
plt.title("Based on 1000 fits on 100 observations")
plt.show()
```



Chapter 3

Literature

Here is a review of existing methods.

Chapter 4

Methods

We describe our methods in this chapter.

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.

Bibliography

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Freedman, D. A. (2009). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge University Press.

Hacking, I. (2006). *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge University Press, 2 edition.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46.

Savage, L. (1954). *The Foundations of Statistics*. Wiley.

Stigler, S. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard University Press.

Wang, E. W. (2019). *The Collected Schizophrenias*. Penguin.