Pathologies of Orthodox Statistics

Thomas P. Minka

(revised 1/26/2000)

Abstract

By rejecting the use of a prior distribution over parameters, orthodox statistics is forced to focus on *estimators*, functions which guess parameter values, and to invent heuristics for choosing among estimators. Two popular heuristics are *unbiasedness* and *maximum likelihood*. Since these heuristics are not consistent with Bayes' rule, they are also not consistent with the axioms of common sense from which Bayes' rule is derived. Hence we expect there to be situations in which they violate common sense and indeed it is not hard to find such situations. This paper reviews a few simple, realistic scenarios where pathologies occur with either the unbiasedness heuristic or the maximum likelihood heuristic.

1 Introduction

Many inference problems work like this: we observe some data and want to infer something about the process that generated it. If we have a probability distribution over possible processes, parameterized by θ, then there is general agreement that Bayes' rule solves the problem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

where x is the data we observe. Unfortunately, a probability distribution over θ is not always forthcoming, e.g. if θ describes something about our universe like the mass of an electron. An influential camp of statisticians, led by Ron Fisher, believed that any choice of $p(\theta)$ would, in this case, be arbitrary and therefore produce arbitrary answers. Their alternative plan went like this:

- Instead of computing a probability distribution over θ, directly decide on one value of θ based on x. Call this function the *estimator* for θ.
- Apply any one of a series of heuristics to design an estimator that makes intuitive sense.

Since it avoided the supposed impasse of the prior distribution, this plan suffused the practice and teaching of statistics, making it now what we call the "orthodox" approach to statistics.

The problem with heuristics is that they don't always work. Thus orthodox statistics is designed to fail occasionally; Bayes' rule never was. This note gives examples of exactly this phenomenon.

2 The Unbiasedness Heuristic

The unbiasedness heuristic says that an estimator $f(x)$ for θ should satisfy

$$\int_X f(x)p(x|\theta)dx = \theta$$

That is, the expected value of our estimator, taken over all possible data sets, should be the true value of the parameter. The following examples illustrate two different kinds of pathologies of this heuristic. The first pathology is related to the fact that the expectation in this definition is taken over all possible data sets, while one should only care about the particular data that was observed. The second pathology arises when no valid estimator satisfies the criteria.

2.1 The bias of a coin

This example appears in Lindley (1972). Suppose we flip a coin n times and get h heads. The probability of this occurrence is

$$P(\text{heads} = h|n, \theta) = \binom{n}{h} \theta^h (1-\theta)^{n-h} \tag{1}$$

The unbiased estimate of θ in this case is h/n. To check this, compute

$$\begin{aligned}
\sum_{h=0}^{\infty} f(h)p(h|n,\theta) &= \sum_{h=0}^{\infty} \frac{h}{n} \frac{n!}{(n-h)!h!} \theta^h (1-\theta)^{n-h} &(2)\\
&= \theta \sum_{h=0}^{\infty} \frac{(n-1)!}{(n-h)!(h-1)!} \theta^{h-1} (1-\theta)^{n-h} &(3)\\
&= \theta \sum_{h=0}^{\infty} p(h-1|n-1,\theta) &(4)\\
&= \theta &(5)
\end{aligned}$$

Now suppose we flip a coin as many times as it takes to get h heads. Let the number of flips be n. Here h is fixed and n varies. The probability of this occurrence is

$$P(\text{flips} = n|h, \theta) = \binom{n-1}{h-1} \theta^h (1-\theta)^{n-h} \tag{6}$$

To see why, note that this situation is equivalent to having $h-1$ heads in $n-1$ trials followed by a head on the nth trial. With respect to θ, the likelihood function in (6) is a constant times the likelihood function in (1). For any two values of θ, say θ_1 and θ_2, the relative weight that the data gives to them is the same in both scenarios:

$$\frac{P(\text{heads} = h|n, \theta_1)}{P(\text{heads} = h|n, \theta_2)} = \frac{P(\text{flips} = n|h, \theta_1)}{P(\text{flips} = n|h, \theta_2)} \tag{7}$$

According to this formula, if the data prefers θ_1 to θ_2 in the first scenario, then θ_1 must also be preferred to θ_2 in the second scenario. However, the unbiased estimate of θ is now $(h-1)/(n-1)$ (or 1 if $n=1$), by a similar derivation as above:

$$\begin{aligned}
\sum_{n=1}^{\infty} f(n)p(n|h,\theta) &= \sum_{n=1}^{\infty} \frac{h-1}{n-1}\frac{(n-1)!}{(n-h)!(h-1)!}\theta^h(1-\theta)^{n-h} && (8)\\
&= \theta\sum_{n=1}^{\infty} \frac{(n-2)!}{(n-h)!(h-2)!}\theta^{h-1}(1-\theta)^{n-h} && (9)\\
&= \theta\sum_{n=1}^{\infty} p(n-1|h-1,\theta) && (10)\\
&= \theta && (11)
\end{aligned}$$

Why does this happen? It is because the unbiasedness heuristic invokes a sum over all x, i.e. all samples that could have been observed, but were not. Therefore, changing your belief in those other samples can change the unbiased estimate, even though the likelihood function for the particular sample you observed is unchanged.

Another way to explain the pathology is due to Jaynes (1996). Since the data set tells us both n and h, it shouldn't matter whether we assumed n or h beforehand. Asserting a proposition twice has no effect on our knowledge: "A and A" is the same as "A." But the unbiasedness heuristic *does* make such a distinction between information assumed beforehand and information acquired from the data. Only the former is used to design the estimator.

Note that when $h=1$ and $n>1$, the unbiased estimator will say $\hat{\theta}=0$, even though this is impossible. The estimator only cares about getting the right expectation, and not about being logically consistent. This point is echoed in the next example.

2.2 Arrival rate

This example appears in Lindley (1972). Suppose while waiting one hour, we observe x arrivals from a Poisson process with rate θ per hour. The probability of this occurrence is

$$P(\text{arrivals}=x|\theta) = e^{-\theta}\theta^x/x!$$

We want to know the probability that there will be no occurrences in the next hour. That is, we want to estimate $e^{-\theta}$. An unbiased estimator $f(x)$ for $e^{-\theta}$ must satisfy

$$\sum_{x=0}^{\infty} f(x)e^{-\theta}\theta^x/x! = e^{-\theta}$$

for all θ. Multiplying both sides by e^{θ} gives

$$\sum_{x=0}^{\infty} f(x)\theta^x/x! = 1 \qquad (12)$$

Therefore $f(x)$ is the power series in θ for 1, which means $f(x) = 0^x$, i.e. $f(x) = 1$ if $x = 0$ and $f(x) = 0$ if $x > 0$. Therefore, if we observe no arrivals in an hour, then we expect to never observe any arrivals ($e^{-\theta} = 1$ so $\theta = 0$), and if we observe any arrivals at all, then we expect to always see arrivals ($e^{-\theta} = 0$ so $\theta = \infty$).

As another example, suppose we now want to estimate the probability of no occurrences in the next two hours, i.e. $e^{-2\theta}$. We can proceed as before by multiplying both sides by e^{θ}, so that $f(x)$ is the power series for $e^{-\theta}$ in θ. This establishes $f(x) = (-1)^x$ as the only unbiased estimate, which is absurd since $e^{-2\theta}$ must always be positive. This estimator also makes an irrelevant distinction between even and odd numbers of arrivals. Similar problems occur if we try to estimate some power of θ (Jaynes, 1996).

Here we have exploited the fact that unbiased estimators are not invariant to a change in the parameters. The unbiased estimator for θ is x, but the estimator for $e^{-\theta}$ is not e^{-x}. The reason is that when you change parameters the expectation is now being taken over a different space.

The same pathology occurs when estimating θ^2 from samples of a $\mathcal{N}(\theta, 1)$ distribution (Press, 1989). The minimum-variance unbiased estimator is $f(x) = \bar{x} - 1/n$, which can sometimes be negative even though θ^2 must be positive.

3 The Maximum Likelihood Heuristic

The maximum likelihood heuristic says that an estimator $f(x)$ for θ should have the property

$$f(x) = \text{argmax}_{\theta} p(x|\theta)$$

This heuristic gives reasonable solutions in the scenarios considered above (the ML estimates are h/n and e^{-x}, respectively). It doesn't rely on an imagined space of data sets and is invariant to reparameterization. However, this section gives two other scenarios in which the unbiasedness heuristic leads to the more sensible answer. This provides a reminder that neither heuristic can be recommended universally. More generally, is not guaranteed that even one of the heuristics will be sensible for your problem.

The maximum likelihood heuristic arises as an approximation to Bayes' rule where the likelihood function is assumed to be sharply peaked and the prior is uniform. Hence pathologies can be found by considering situations where the approximation is not valid.

3.1 Mixture of Gaussians

This example appears in Lindley (1972). Let $X = [x_1..x_N]$ be a sample of size N from the density

$$p(x) = \frac{1}{2}\mathcal{N}(x; 0, 1) + \frac{1}{2}\mathcal{N}(x; \mu, \sigma^2)$$

That is, each sample either comes from a standard normal or a normal with unknown parameters. If we set $\mu = x_i$ for some i, then as $\sigma \to 0$ the probability of the entire data set $p(X|\mu, \sigma)$ tends to infinity. Therefore all maximum likelihood estimates are of the form $(\mu = x_i, \sigma = 0)$. However, $\sigma = 0$ does not define a proper density, so this solution must be rejected. Of course, there are *local* maxima which are well-behaved, but since these are only local, they are rejected by the maximum-likelihood heuristic.

An even simpler example has a single sample from a Gaussian with unknown mean and variance. Again, the maximum likelihood estimate of the variance is invalid. If instead we use Bayes' rule with a uniform prior on m, we find that

$$\begin{aligned}
p(x|\sigma) &= \int_m p(x|m, \sigma)p(m) && (13)\\
&= \int_{-\infty}^{\infty} \mathcal{N}(x; m, \sigma^2)\ 1\ dm && (14)\\
&= 1 = p(x) && (15)\\
p(\sigma|x) &= \frac{p(x|\sigma)p(\sigma)}{p(x)} && (16)\\
&= p(\sigma) && (17)
\end{aligned}$$

so the posterior for σ is simply the prior for σ, which makes sense since one data point provides no information about the variance.

3.2 Pairs of samples

This example appears in Lindley (1972). Suppose we have K Gaussian densities from which we draw two samples each. The densities have means $\mu_1..\mu_K$ and common variance σ^2, all of which are unknown. The samples from density i are a_i and b_i. Let the collective data set which contains $2K$ points be called $\mathbf{X} = \{a_1..a_K, b_1..b_K\}$. The probability of $\mathbf{X}$ is

$$\begin{aligned}
p(\mathbf{X}|\mu_1..\mu_K, \sigma^2) &= \prod_i \mathcal{N}(a_i; \mu_i, \sigma^2)\mathcal{N}(b_i; \mu_i, \sigma^2) && (18)\\
&= \frac{1}{(2\pi\sigma^2)^K} \exp\left(-\frac{\sum_i (a_i - \mu_i)^2 + (b_i - \mu_i)^2}{2\sigma^2}\right) && (19)\\
&= \frac{1}{(2\pi\sigma^2)^K} \exp\left(-\frac{\sum_i (a_i - b_i)^2 + 4(\mu_i - \frac{a_i+b_i}{2})^2}{4\sigma^2}\right) && (20)
\end{aligned}$$

from which it is clear that the maximum likelihood estimate of $\mu_i = (a_i + b_i)/2$. Taking the logarithm, we have that the maximum likelihood estimate of σ^2 must maximize

$$-K \log(2\pi\sigma^2) - \frac{1}{4\sigma^2}\sum_i (a_i - b_i)^2$$

whose maximum occurs at $\hat{\sigma}^2 = \frac{1}{4K}\sum_i (a_i - b_i)^2$. However, this estimator is biased:

$$E[\hat{\sigma}^2] = \frac{1}{2}\sigma^2 \tag{21}$$

and remains so even as $K \to \infty$ (i.e. the number of samples goes to infinity). Therefore the estimator is not consistent.

If $K = 1$, then we have two samples from a Gaussian with unknown parameters, where we already knew that the maximum likelihood estimator is biased by a factor of two. The case $K > 1$ essentially repeats this experiment many times, obtaining a more precise estimator but one which is still biased by a factor of two.

One way to explain this is that the number of free parameters increases with K, so the overfitting caused by the maximum-likelihood heuristic never goes away. Parameters like μ_i that are not of interest in the problem are called *nuisance parameters* and are much easier to handle using Bayes' rule: they are simply integrated out. Using a uniform prior over μ_i gives

$$\begin{aligned}
p(X|\sigma^2) &= \int_{\mu_1} .. \int_{\mu_K} p(X|\mu_1..\mu_K, \sigma^2) p(\mu_1)..p(\mu_K) && (22)\\
&= \int_{\mu_1} .. \int_{\mu_K} \frac{1}{(2\pi\sigma^2)^K} \exp\left(-\frac{\sum_i (a_i - b_i)^2}{4\sigma^2}\right) \exp\left(-\frac{\sum_i (\mu_i - \frac{a_i+b_i}{2})^2}{\sigma^2}\right) 1 && (23)\\
&= \frac{(\pi\sigma^2)^{K/2}}{(2\pi\sigma^2)^K} \exp\left(-\frac{\sum_i (a_i - b_i)^2}{4\sigma^2}\right) \int_{\mu_1} .. \int_{\mu_K} \frac{1}{(\pi\sigma^2)^{K/2}} \exp\left(-\frac{\sum_i (\mu_i - \frac{a_i+b_i}{2})^2}{\sigma^2}\right) && (24)\\
&= \frac{(\pi\sigma^2)^{K/2}}{(2\pi\sigma^2)^K} \exp\left(-\frac{\sum_i (a_i - b_i)^2}{4\sigma^2}\right) && (25)
\end{aligned}$$

The maximum of this likelihood function is at $\hat{\sigma}^2 = \frac{1}{2K}\sum_i (a_i - b_i)^2$, which as an estimator is not only consistent but also unbiased. Of course, there is no need to use a maximum-likelihood solution at this point either. Since the posterior for σ^2 will approach an impulse as $K \to \infty$, any Bayesian estimate will be consistent, regardless of priors and costs.

4 Conclusions

Orthodoxy plays a shell game where if one heuristic fails, it advocates another. Unfortunately, it is not always so easy to check the answer by intuition and see that it is inconsistent or illogical.

Furthermore, appealing to the unbiasedness versus maximum likelihood heuristic is a matter of taste, leading to a multiplicity of "best" estimators for the same problem. Interestingly, these are the same criticisms that orthodoxy leveled against using priors.

Alternatively, one can simply use Bayesian probability theory, which is not heuristic in nature and dictates a single right answer for every well-posed problem. Its success hinges only on the consistency of the axioms from which it is derived: axioms which are not only compelling but which have been scrutinized for centuries without a single inconsistency ever having been found. Thus there is no rebuttal paper to this one, containing pathologies in Bayesian probability theory. The "paradoxes" of Bayesian probability theory which sometimes appear in the literature have always been found erroneous (Jaynes, 1996).

The requirement for this strong warranty, however, is that the practitioner be careful to specify all assumptions and knowledge to be used in an inference task. If certain knowledge is missing, then that fact should be encoded too. Orthodoxy believes that probability distributions can only encode knowledge, and not ignorance. But one of the achievements of this century, though it was slow in being recognized, was that it is possible to encode ignorance in a prior (Jeffreys, 1961). Since Jeffreys' original work, there is now a large body of research on how to find these priors. At this point in time, we have numerous examples of how orthodox results, when they make sense, are equivalent to choosing one such "ignorance prior."

Acknowledgements

Aaron Bobick and Rosalind Picard helped clarify the presentation.

References

[1] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Fragmentary Pre-release, 1996. `ftp://bayes.wustl.edu/Jaynes.book/`.

[2] Harold Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, third edition, 1961.

[3] D. V. Lindley. *Bayesian Statistics: A Review*. Society for Industrial and Applied Mathematics, 1972.

[4] S. James Press. *Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons, New York, 1989.