Bayes@Lund Book Club

Janes' Probability Theory

Chapter 4, Elementary hypothesis testing

What have we done so far?

- Chapter 1
  - Deductive and plausible reasoning
  - Introducing the robot
- Chapter 2
  - Probability Theory As Extended Logic
  - Probability without Kolmogorov
- Chapter 3
  - Sampling problems
  - *The model is fixed, what's the probability distribution over not-yet-seen data?*

As is clear from the basic desiderata listed in Chapter 1, the fundamental principle underlying all probabilistic inference is:

> *To form a judgment about the likely truth or falsity of any proposition A, the correct procedure is to calculate the probability that A is true:*
>
> $$P(A|E_1E_2\cdots) \tag{4.1}$$
>
> *conditional on all the evidence at hand.*

What’s in chapter 4? What’s not in chapter 4?

- Bayes rule
- Priors, posteriors, likelihood
- Odds, logodds and... decibels (?!)
- Hypothesis testing
- Multiple hypothesis testing
- Optional stopping
- Continuous distributions
- (almost) parameter estimation
- Dissing Bayes

Which of a set of hypotheses is most likely to be true?

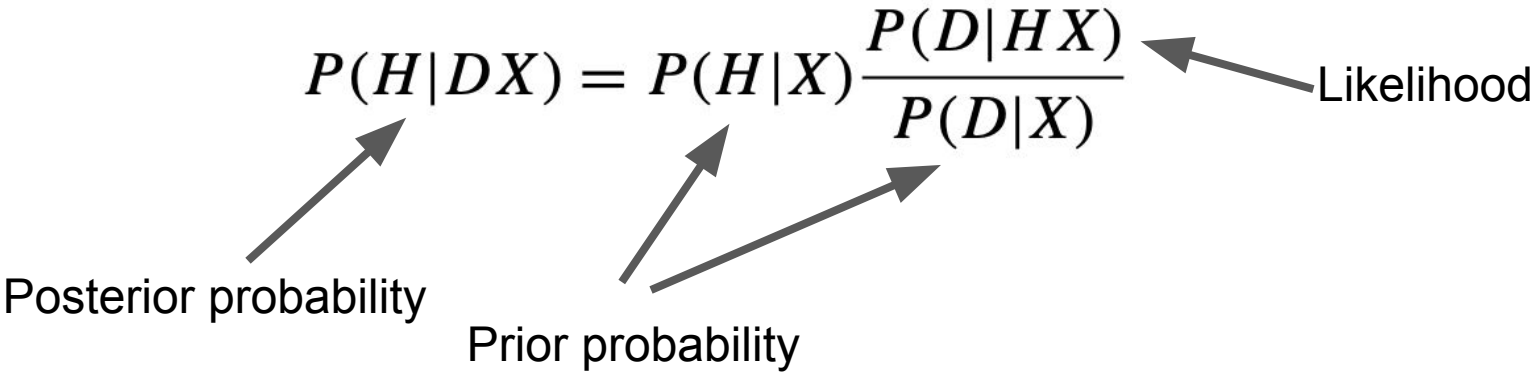
To solve this problem does not require any new principles beyond the product rule (3.1) that we used to find conditional sampling distributions; we need only to make a different choice of the propositions. Let us now use the notation

$$\begin{aligned} X &= \text{prior information}, \\ H &= \text{some hypothesis to be tested}, \\ D &= \text{the data}, \end{aligned}$$

and write the product rule in the form

$$P(DH|X) = P(D|HX)P(H|X) = P(H|DX)P(D|X). \tag{4.2}$$

$$P(H|DX) = P(H|X)\frac{P(D|HX)}{P(D|X)}. \tag{4.3}$$



A likelihood $L(H)$ is not itself a probability for H; it is a dimensionless numerical function which, when multiplied by a prior probability and a normalization factor, may become a probability.

Testing binary hypotheses with binary data

We have 11 automatic machines turning out widgets, which pour out of the machines into 11 boxes. This example corresponds to a very early stage in the development of widgets, because ten of the machines produce one in six defective. The 11th machine is even worse; it makes one in three defective. The output of each machine has been collected in an unlabeled box and stored in the warehouse.

$$A \equiv \text{we chose a bad batch (1/3 defective)},$$
$$B \equiv \text{we chose a good batch (1/6 defective)}.$$

$$P(H|DX) = P(H|X)\frac{P(D|HX)}{P(D|X)}. \tag{4.3}$$

$$P(\overline{H}|DX) = P(\overline{H}|X)\frac{P(D|\overline{H}X)}{P(D|X)}, \tag{4.4}$$

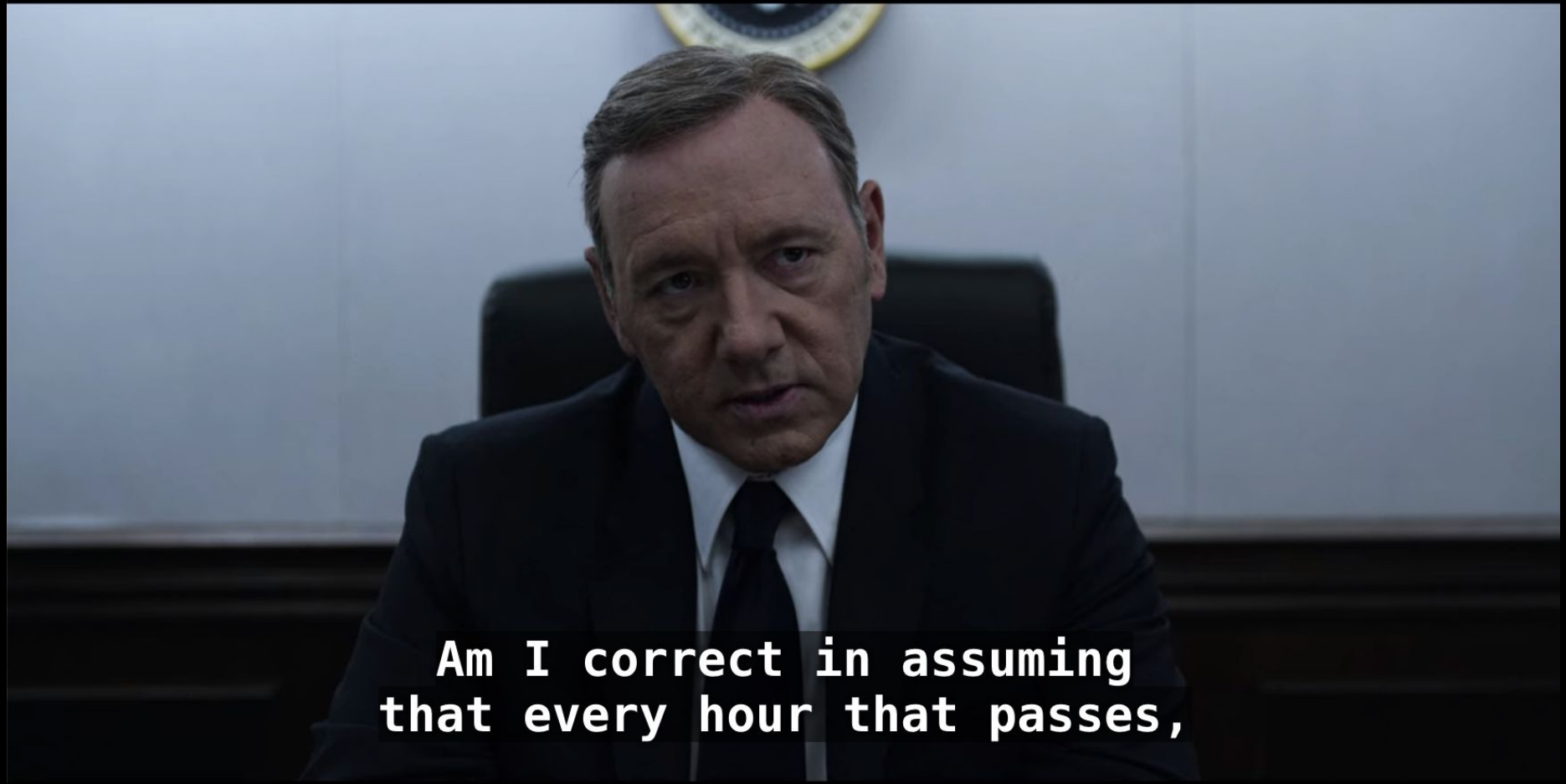
$$\frac{P(H|DX)}{P(\overline{H}|DX)} = \frac{P(H|X)}{P(\overline{H}|X)}\frac{P(D|H\ X)}{P(D|\overline{H}X)}, \tag{4.5}$$

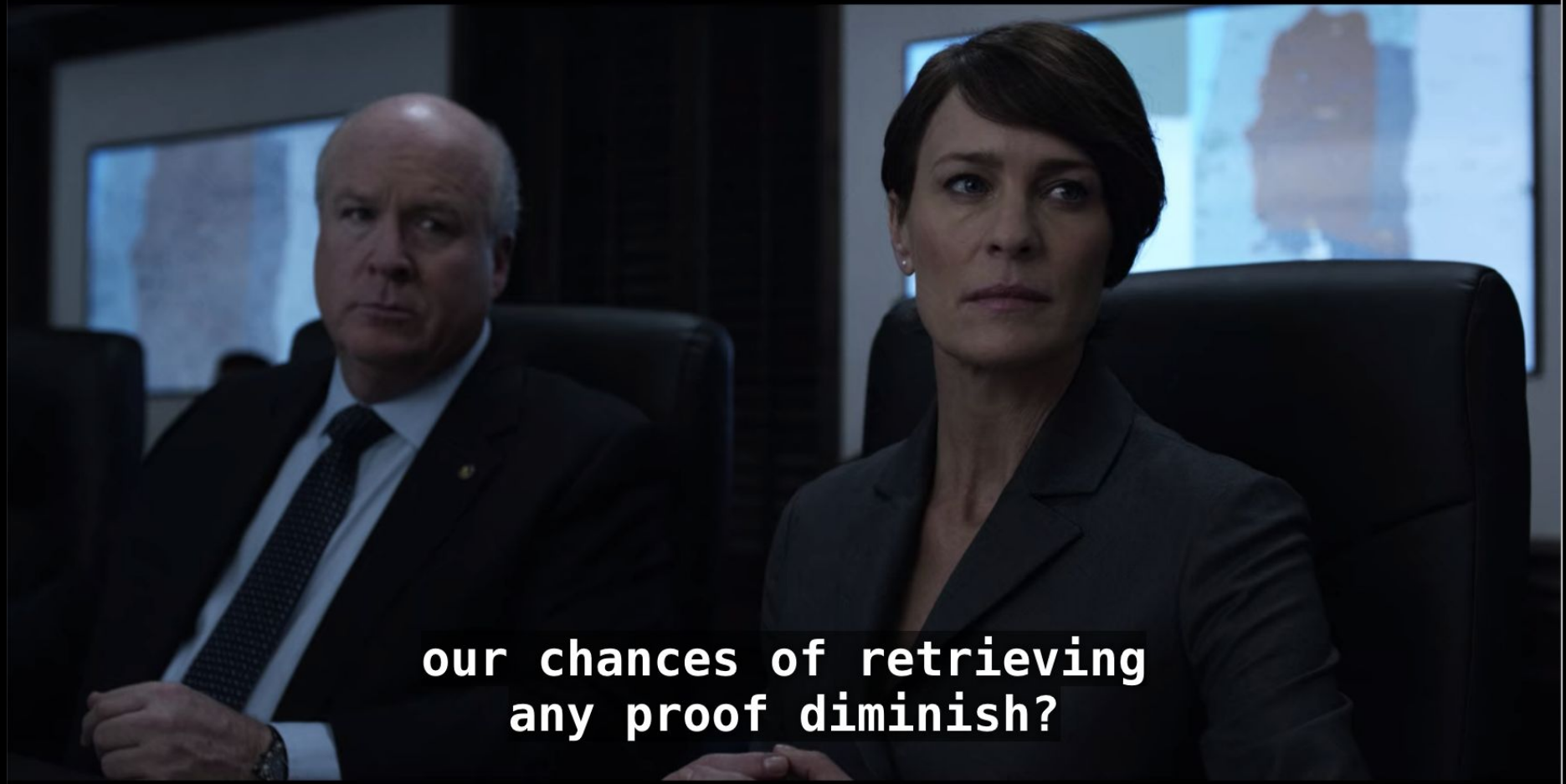
$$O(H|D\ X) \equiv \frac{P(H|D\ X)}{P(\overline{H}|DX)}, \tag{4.6}$$

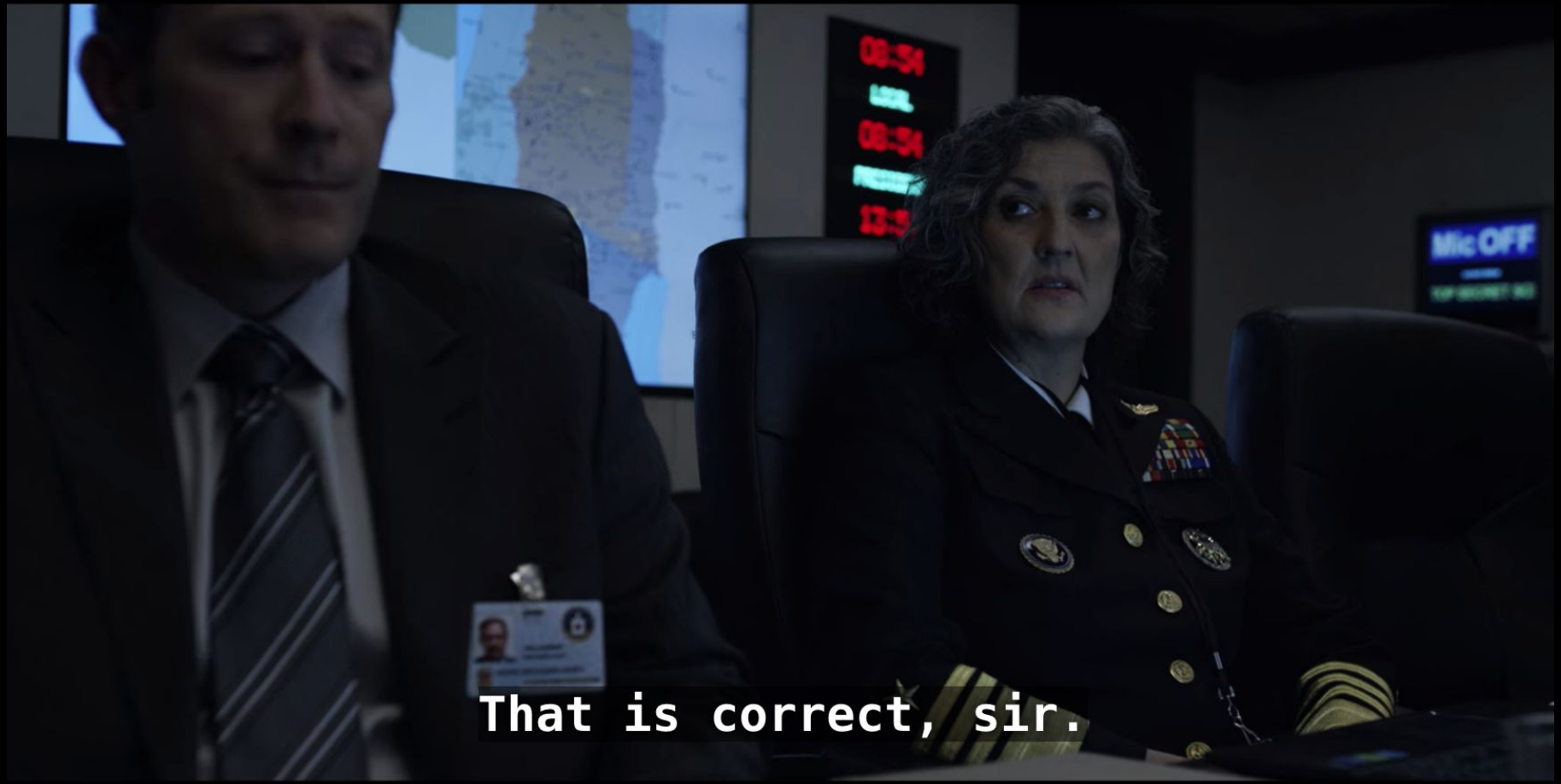
then we can combine (4.3) and (4.4) into the following form:

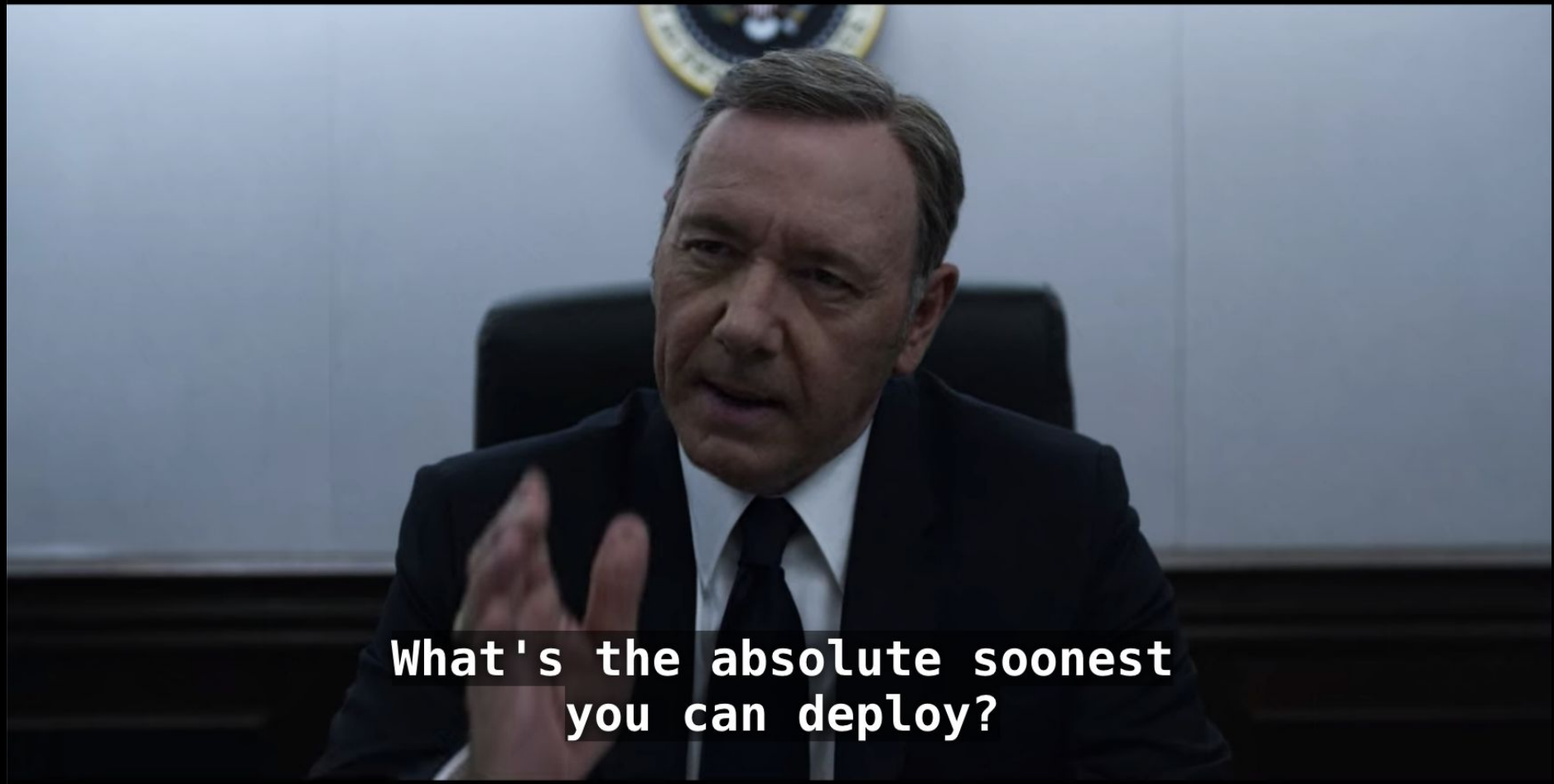
$$O(H|D\ X) = O(H|X)\frac{P(D|HX)}{P(D|\overline{H}X)}. \tag{4.7}$$

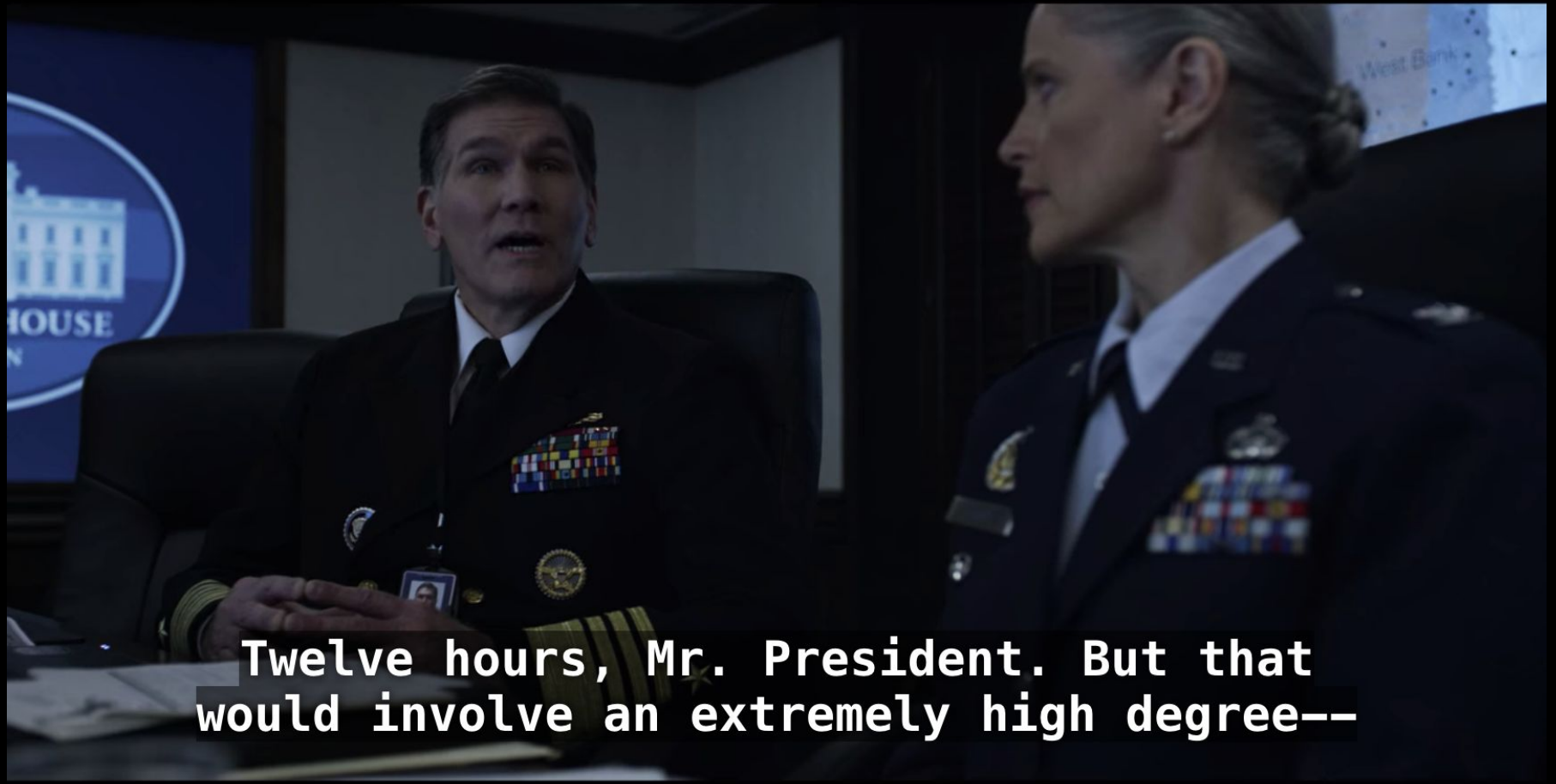
What are the odds?


Am I correct in assuming
that every hour that passes,


our chances of retrieving
any proof diminish?

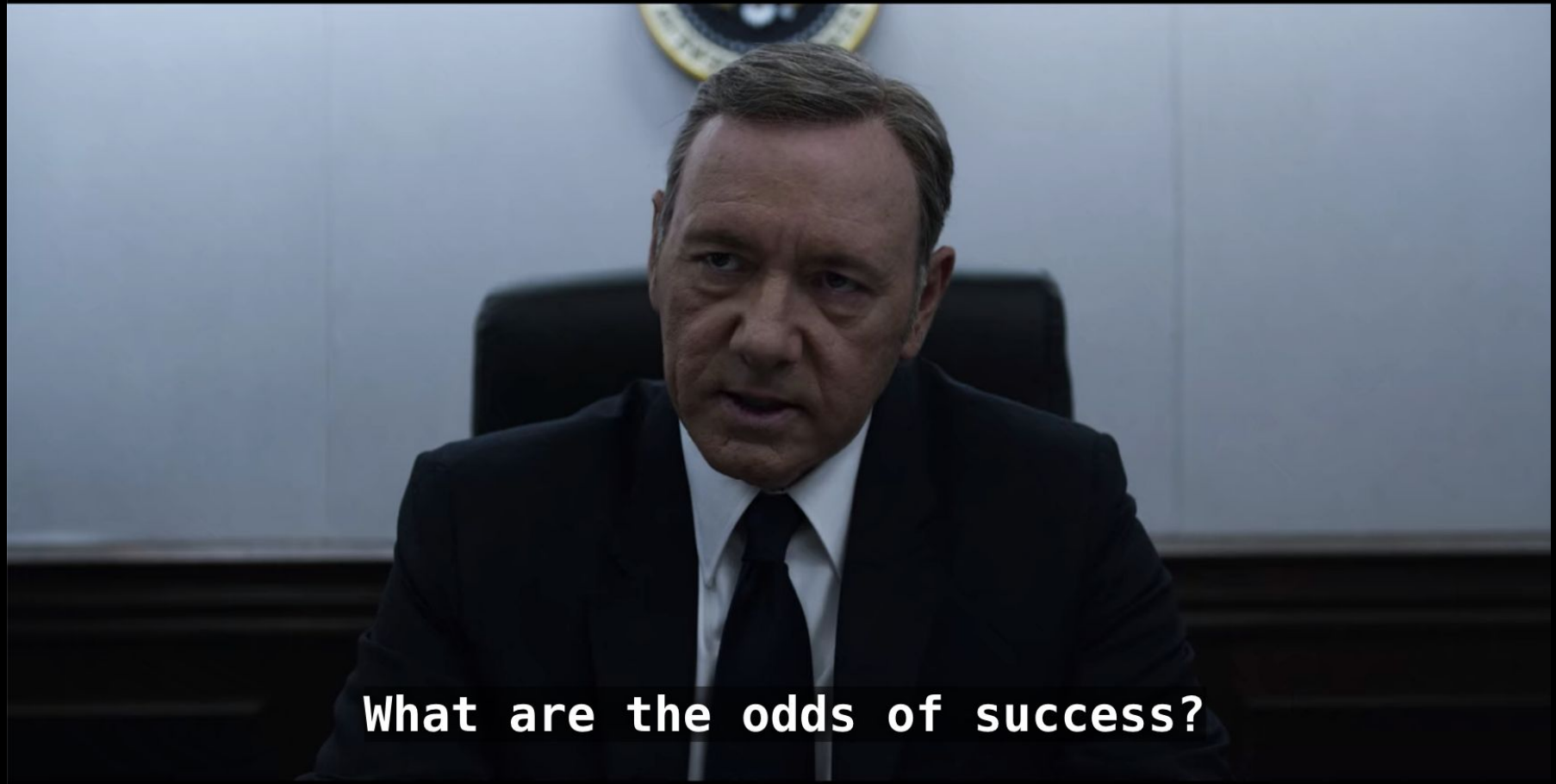

That is correct, sir.

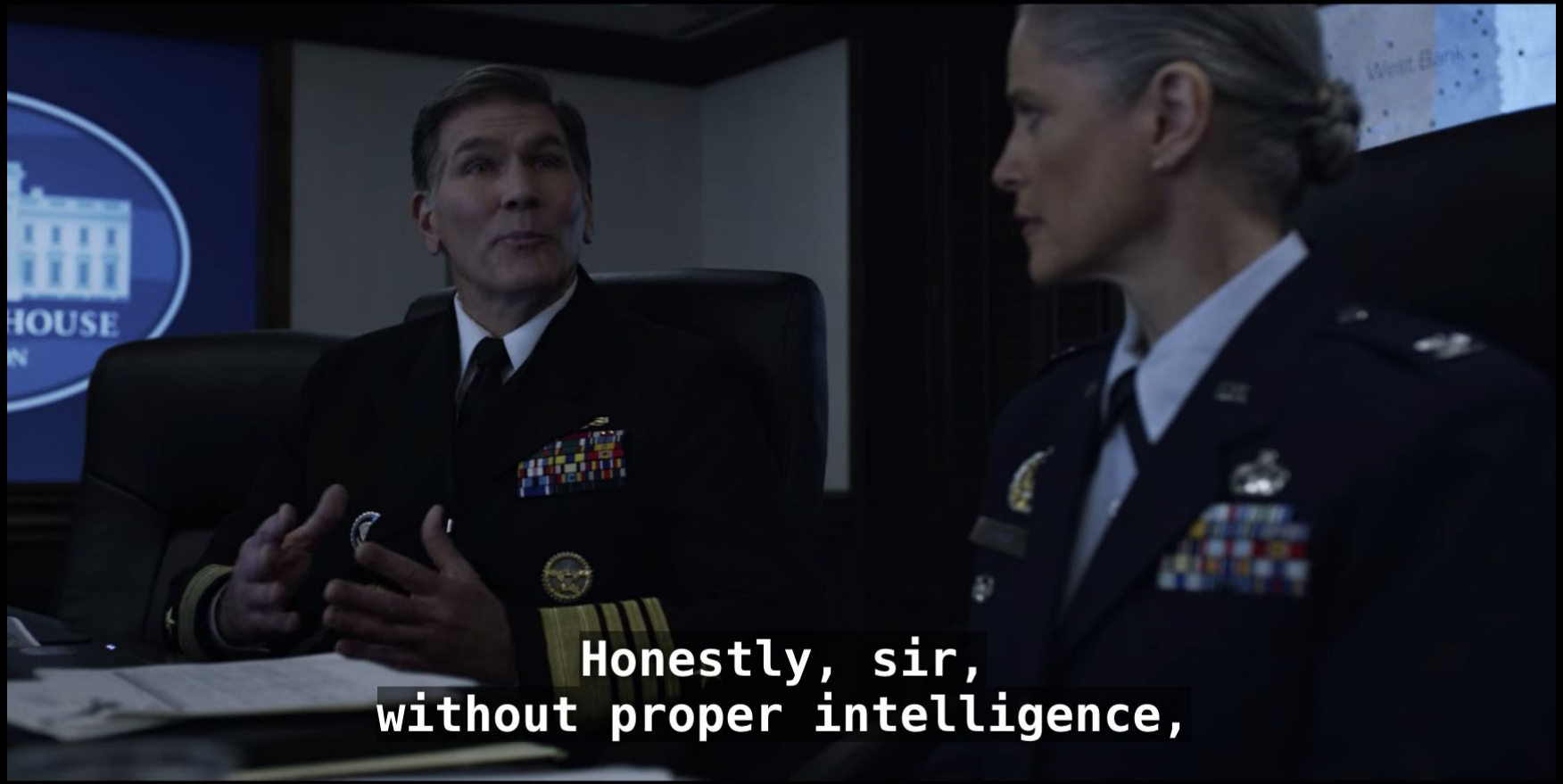

What's the absolute soonest
you can deploy?

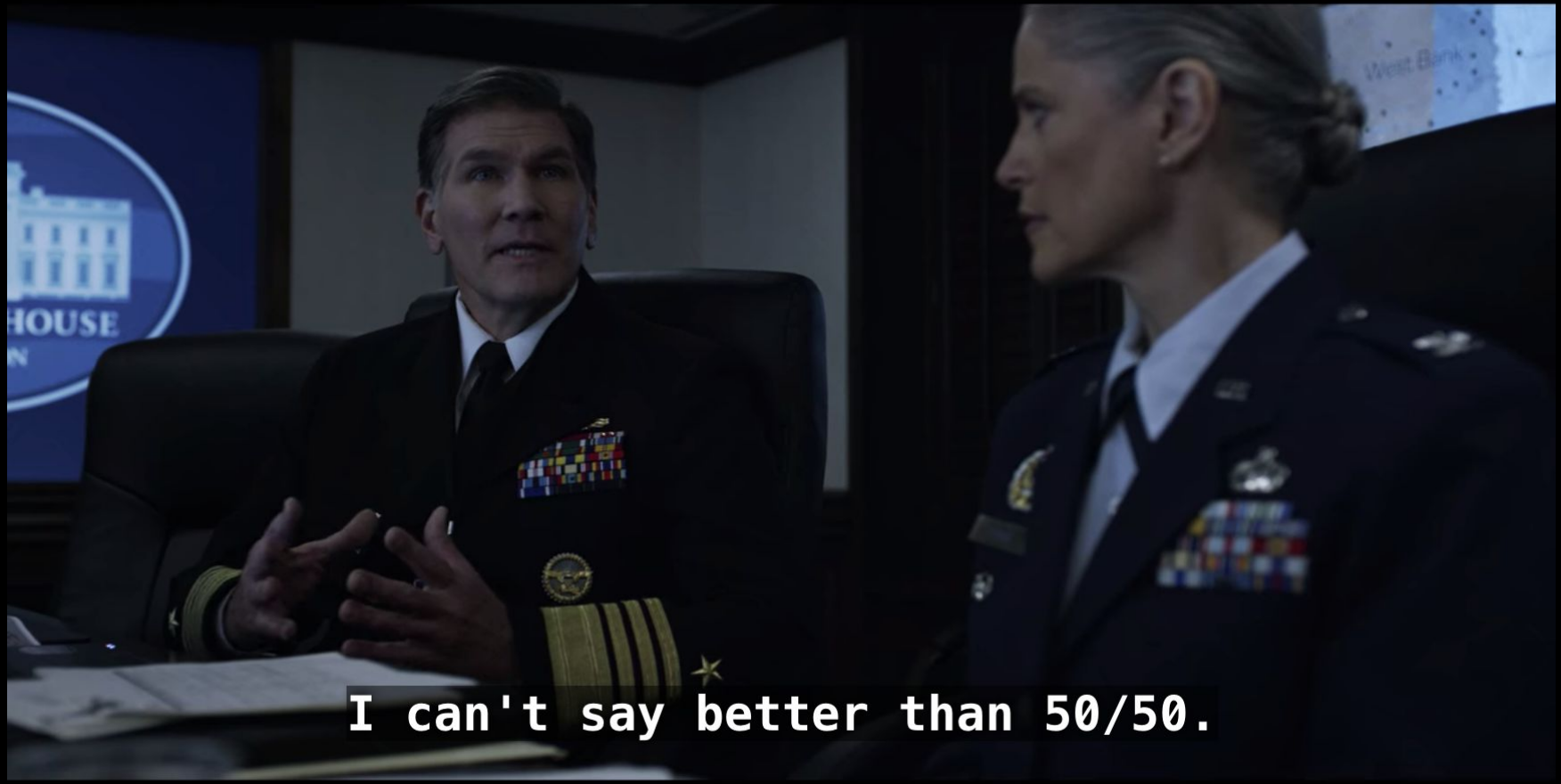

Twelve hours, Mr. President. But that
would involve an extremely high degree--


Yeah, I'm aware of the risk, Admiral.


Yeah, I'm aware of the risk, Admiral.


What are the odds of success?


Honestly, sir,
without proper intelligence,


I can't say better than 50/50.


You're certain?


Yes, I am.


Let's put it in motion.

Today, the logarithm-of-odds $\{u = \log[p/(1-p)]\}$ has proved to be such an important quantity that it deserves a shorter name; but we have had trouble finding one. Good (1950) was perhaps the first author to stress its importance in a published work, and he proposed the name *lods*, but the term has a leaden ring to our ears, as well as a nondescriptive quality, and it has never caught on.

Log-odds!


20
10
5
3
0
3
6
–
20
40
60
80
100 %
+
dB

Jaynes' Evidence

We define a new function, which we will call the *evidence* for H given D and X:

$$e(H|DX) \equiv 10 \log_{10} O(H|DX). \tag{4.8}$$

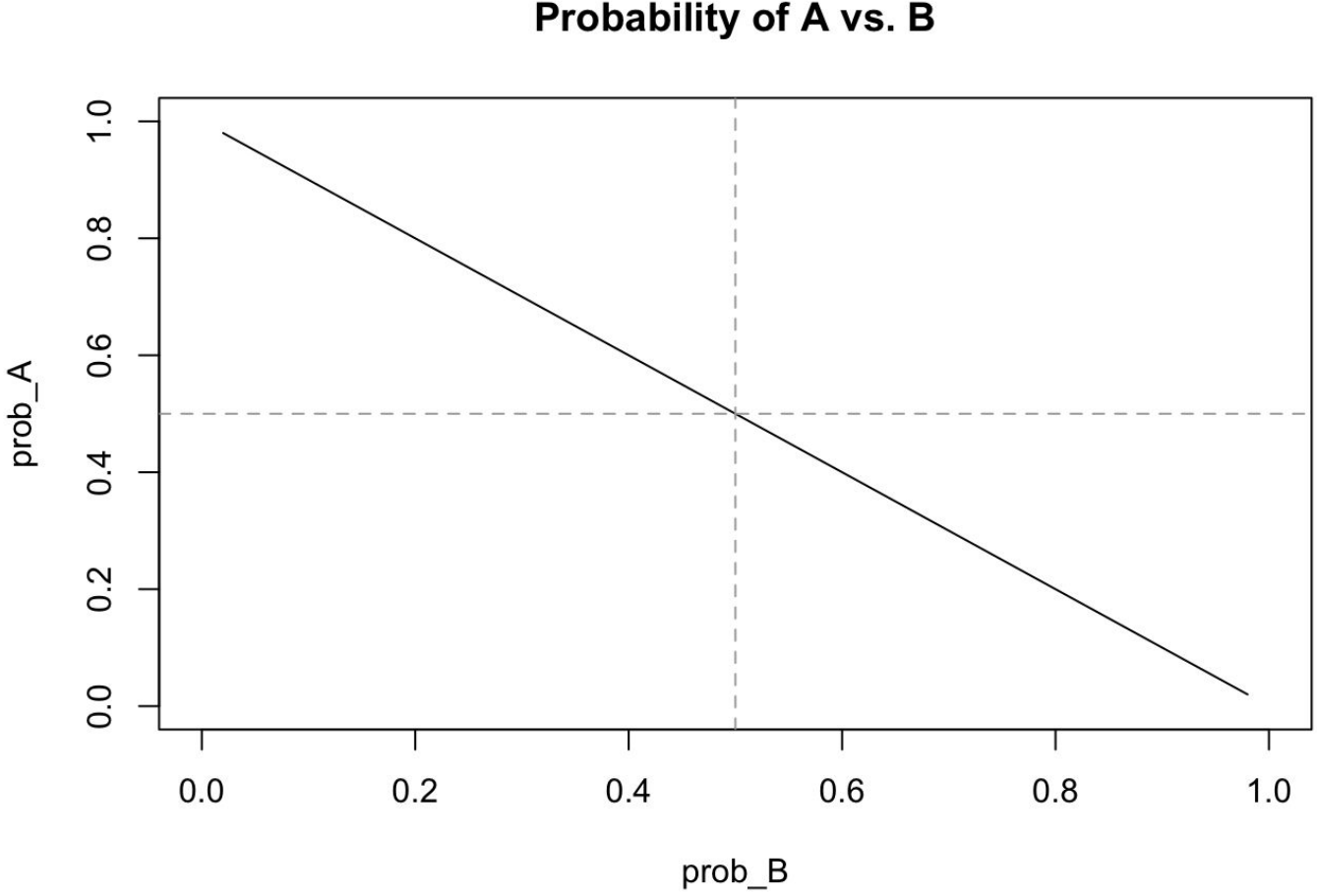
This is still a monotonic function of the probability. By using the base 10 and putting the factor 10 in front, we are now measuring evidence in *decibels* (hereafter abbreviated to db). The evidence for H, given D, is equal to the prior evidence plus the number of db provided by working out the log likelihood in the last term below:

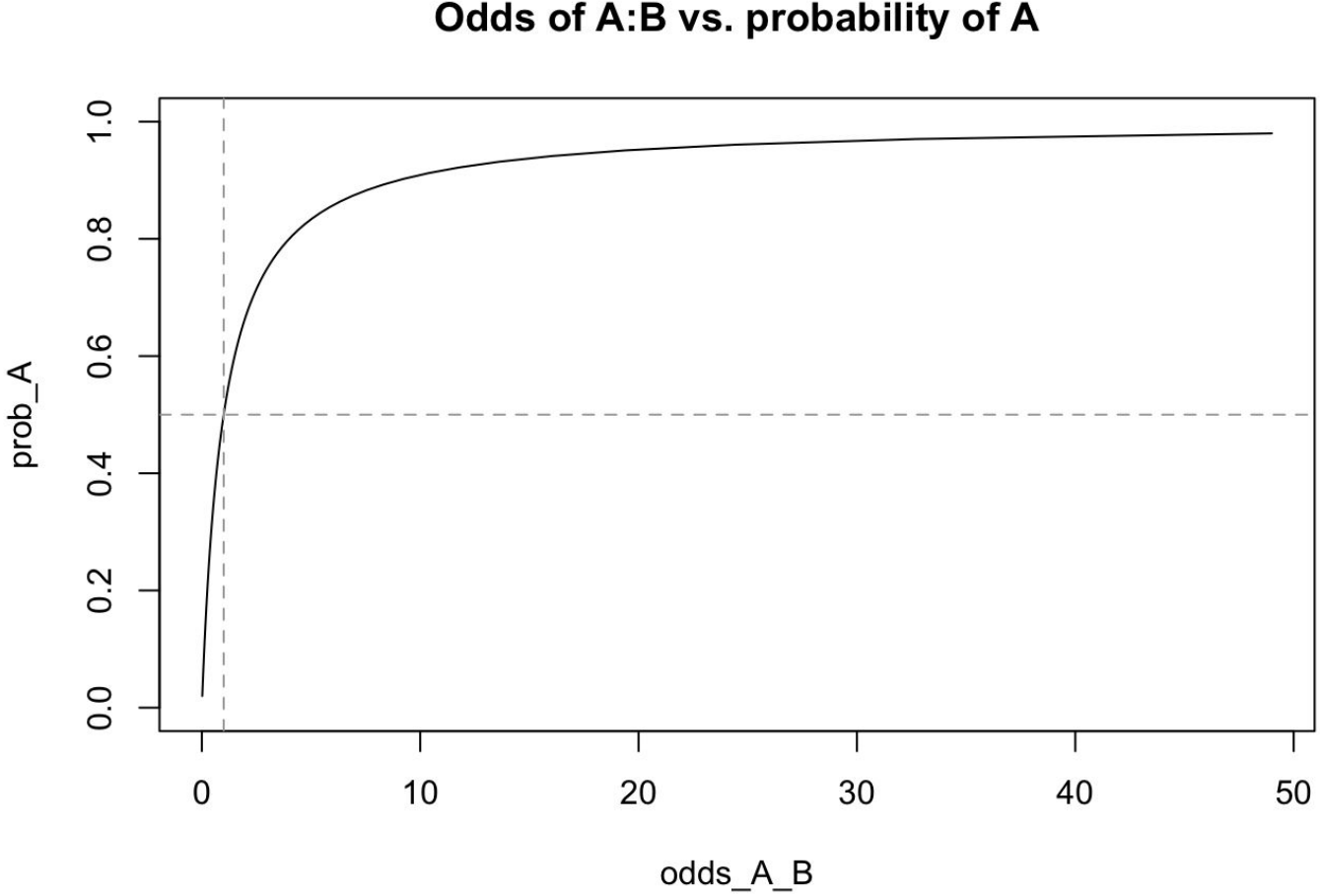
$$e(H|DX) = e(H|X) + 10 \log_{10} \left[\frac{P(D|HX)}{P(D|\overline{H}X)} \right]. \tag{4.9}$$

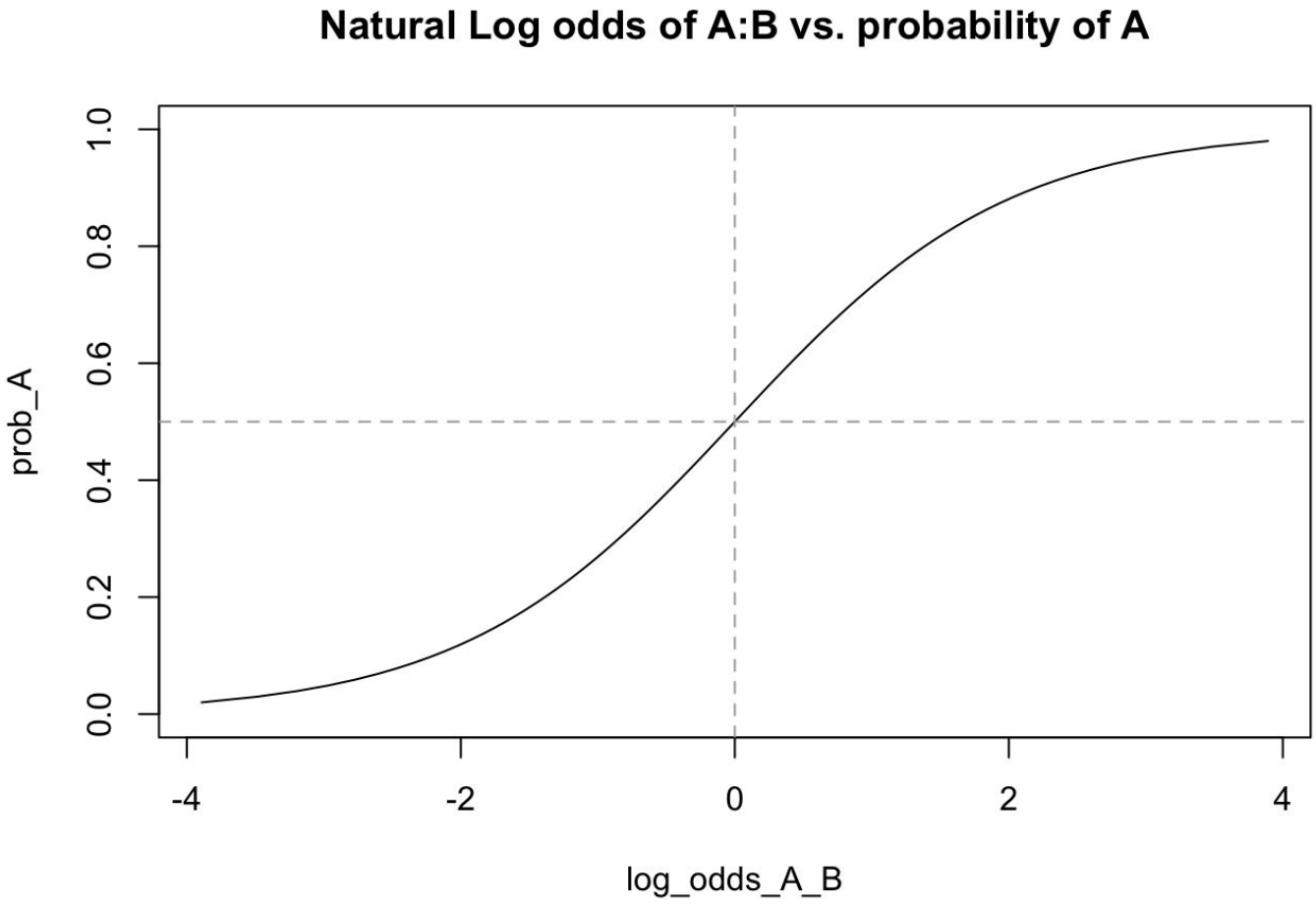
dB >> Odds >> Probability

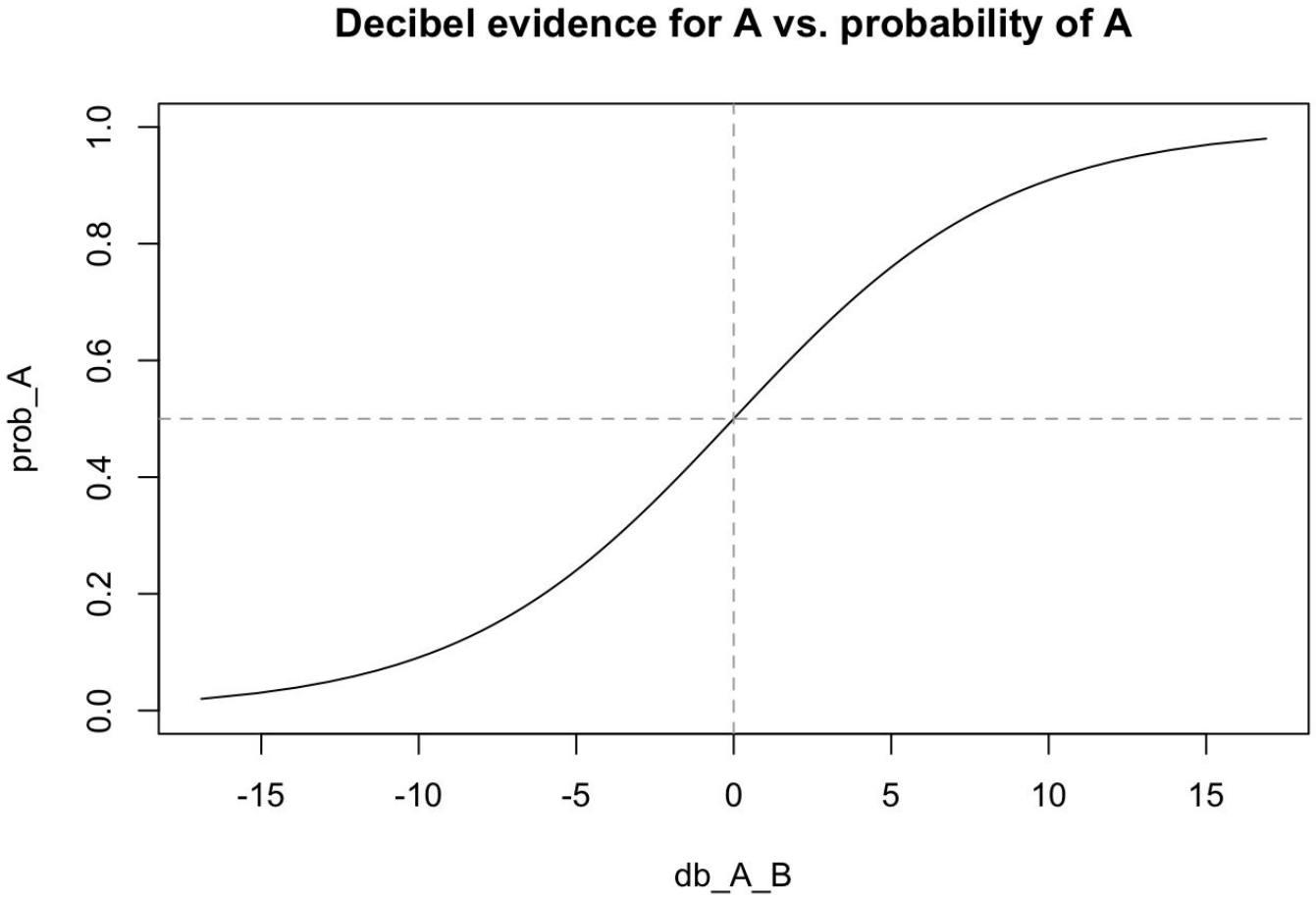
It is obvious from Table 4.1 why it is very cogent to give evidence in decibels. When probabilities approach one or zero, our intuition doesn't work very well. Does the difference between the probability of 0.999 and 0.9999 mean a great deal to you? It certainly doesn't to the writer. But after living with this for only a short while, the difference between evidence of plus 30 db and plus 40 db does have a clear meaning to us.

e	O	p
0	1:1	1/2
3	2:1	2/3
6	4:1	4/5
10	10:1	10/11
20	100:1	100/101
30	1000:1	0.999
40	10^4:1	0.9999
$-e$	$1/O$	$1-p$


Probability of A vs. B
prob_A
0.0
0.2
0.4
0.6
0.8
1.0
prob_B
0.0
0.2
0.4
0.6
0.8
1.0


Odds of A:B vs. probability of A
prob_A
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
odds_A_B


Natural Log odds of A:B vs. probability of A
prob_A
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
log_odds_A_B


Decibel evidence for A vs. probability of A
prob_A
0.0
0.2
0.4
0.6
0.8
1.0
-15
-10
-5
0
5
10
15
db_A_B

The 1 dB threshold

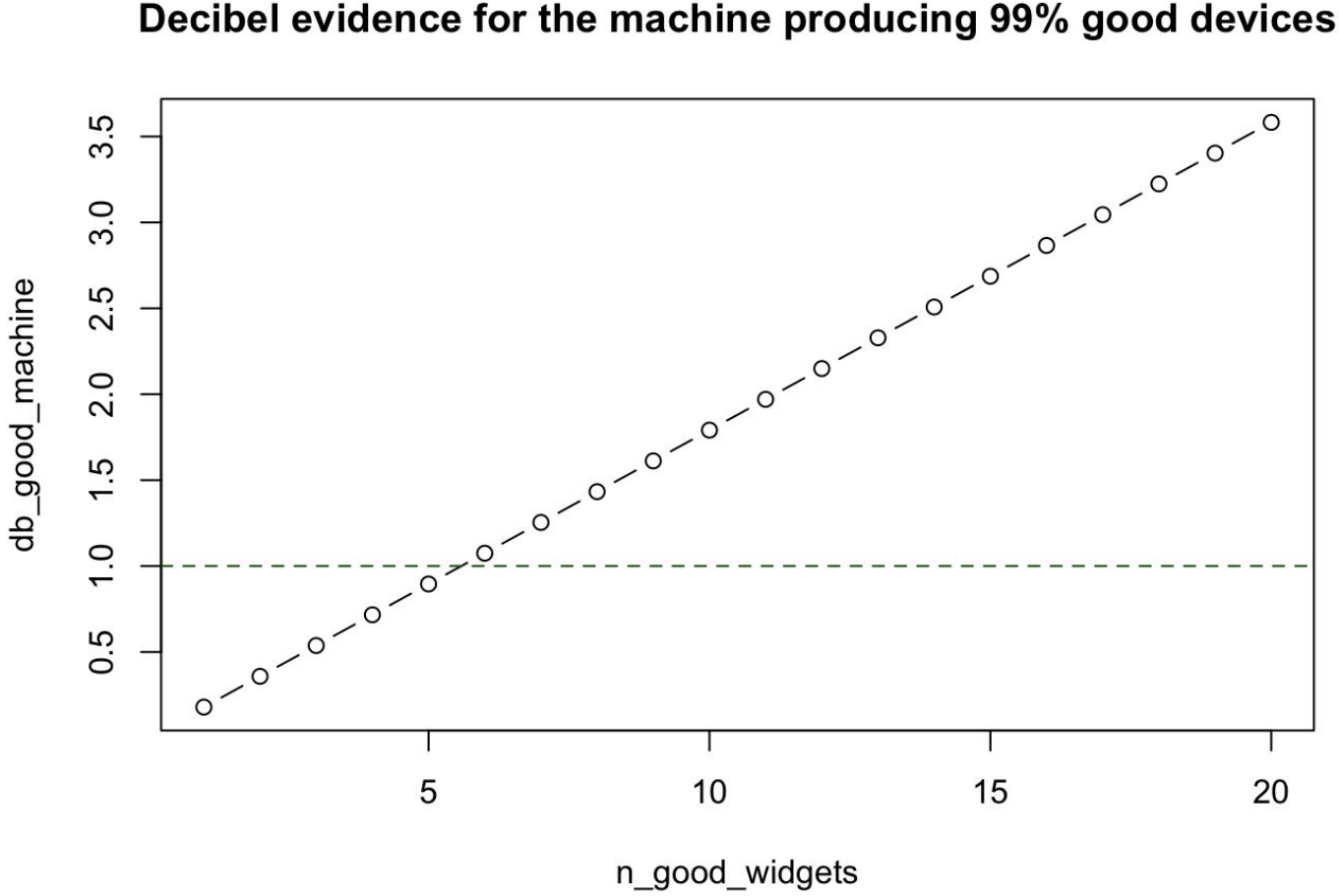
Even the factor of 10 in (4.8) is appropriate. In the original acoustical applications, it was introduced so that a 1 db change in sound intensity would be, psychologically, about the smallest change perceptible to our ears. With a little familiarity and a little introspection, we think that the reader will agree that a 1 db change in evidence is about the smallest increment of plausibility that is perceptible to our intuition.

The 1 dB threshold with two machines

The Good machine: 99/100 Widgets are correct

The Bad machine: 95/100 Widgets are correct

Which machine are we looking at?


Decibel evidence for the machine producing 99% good devices
db_good_machine
0.5
1.0
1.5
2.0
2.5
3.0
3.5
5
10
15
20
n_good_widgets

Sequential inference

The way described above is how our robot would do it if we told it to reject or accept on the basis that the *posterior probability* of proposition A reaches a certain level. This very useful and powerful procedure is called 'sequential inference' in the statistical literature, the term signifying that the number of tests is not determined in advance, but depends on the sequence of data values that we find; at each step in the sequence we make one of three choices: (a) stop with acceptance; (b) stop with rejection; (c) make another test. The term should not be confused with what has come to be called 'sequential analysis with nonoptional stopping', which is a serious misapplication of probability theory; see the discussions of optional stopping in Chapters 6 and 17.

Adults learn to make mental allowance for the reliability of the source when told something hard to believe. One might think that, ideally, the information which our robot should have put into its memory was not that we had either 1/3 bad or 1/6 bad; the information it should have put in was that some unreliable human *said* that we had either 1/3 bad or 1/6 bad.

We *do* want the robot to believe whatever we tell it; it would be dangerous to have a robot who suddenly became skeptical in a way not under our control when we tried to tell it some true but startling – and therefore highly important – new fact. But then the onus is on us to be aware of this situation, and when there is a good chance that skepticism will be needed, it is up to us to give the robot a hint about how to be skeptical for that particular problem.

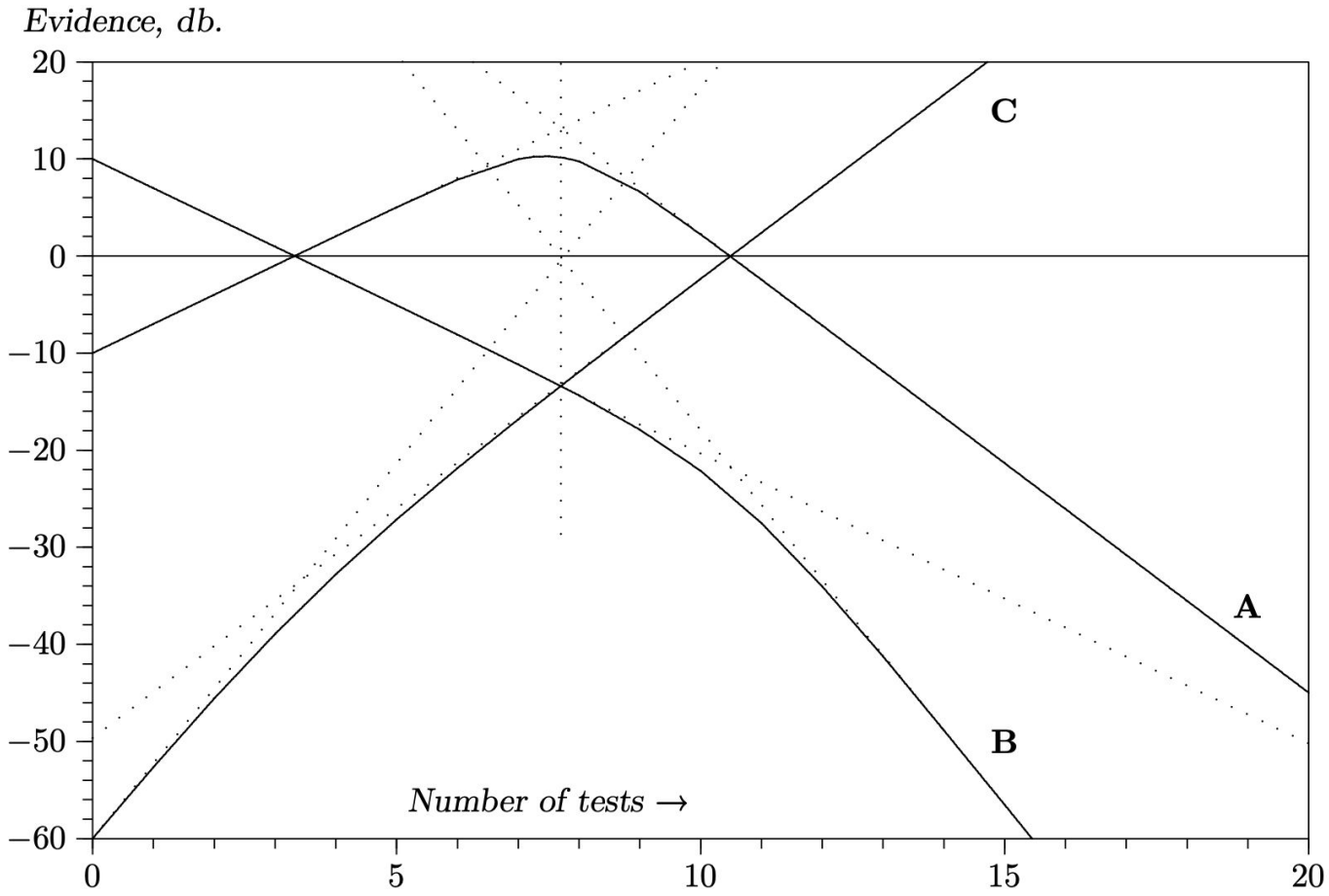


Fig. 4.1. A surprising multiple sequential test wherein a dead hypothesis (C) is resurrected.


Christmas
GIFT EXCHANGE
DICE GAME
STEAL ANY GIFT
MAKE 2 PEOPLE SWAP GIFTS
EVERYONE PASS LEFT
EVERYONE PASS RIGHT
LEAVE GAME OR STAY
UNWRAP YOUR GIFT

What is distributed?

We must warn the reader about another semantic confusion which has caused error and controversy in probability theory for many decades. It would be quite wrong and misleading to call $g(f)$ the 'posterior distribution **of** f', because that verbiage would imply to the unwary that f itself is varying and is 'distributed' in some way. This would be another form of the mind projection fallacy, confusing reality with a state of knowledge about reality. In the problem we are discussing, f is simply an unknown constant parameter; what is 'distributed' is not the *parameter*, but the *probability*. Use of the terminology 'probability distribution **for** f' will be followed, in order to emphasize this constantly.

Continuous distributions

But, in fact, if we become pragmatic we note that f is not really a continuously variable parameter. In its working lifetime, a machine will produce only a finite number of widgets; if it is so well built that it makes 10^8 of them, then the possible values of f are a finite

s ion

o ere

is em.

L ples

o like

this when one takes note of the actual, real-world situation.

It's all discrete

In spite of the pragmatic argument just given, thinking of continuously variable parameters is often a natural and convenient approximation to a real problem (only we should not take it so seriously that we get bogged down in the irrelevancies for the real world that infinite sets and measure theory generate). So, suppose that we are now testing simultaneously an

Dissing Bayes

4.6.1 Historical digression

It appears that this result was first found by an amateur mathematician, the Rev. Thomas Bayes (1763). For this reason, the kind of calculations we are doing are called 'Bayesian'. We shall follow this long-established custom, although it is misleading in several respects. The general result (4.3) is always called 'Bayes' theorem', although Bayes never wrote it; and it is really nothing but the product rule of probability theory which had been recognized by others, such as James Bernoulli and A. de Moivre (1718), long before the work of Bayes. Furthermore, it was not Bayes but Laplace (1774) who first saw the result in generality and showed how to use it in real problems of inference. Finally, the calculations we are doing – the direct application of probability theory as logic – are more general than mere application of Bayes' theorem; that is only one of several items in our toolbox.

What’s in chapter 4? What’s not in chapter 4?

- Bayes rule
- Priors, posteriors, likelihood
- Odds, logodds and... decibels (?!)
- Hypothesis testing
- Multiple hypothesis testing
- Optional stopping
- Continuous distributions
- (almost) parameter estimation
- Dissing Bayes