# Using evidence to make decisions

**Charles Jenkins, Energy Division, CSIRO Black Mountain**

charles.jenkins@csiro.au

# Jasper's view

- From Wall & Jenkins, page 1, line 1

Science is about decision. Building instruments, collecting data, reducing data, compiling catalogues, classifying, doing theory – all of these are tools, techniques or aspects which are necessary. But we are not doing science unless we are deciding something; only *decision counts*. Is this hypothesis or theory correct? If not, why not? Are these data self-consistent or consistent with other data? Adequate to answer the question posed? What further experiments do they suggest?

# Evidence, Bayes factor and decisions

If we have two exclusive models H0 and H1, then the Bayes factor links the prior and posterior odds

posterior odds = Bayes factor x prior odds

The Bayes factor **B** depends on the average likelihood over the priors:

$$\frac{\int \text{prob(data|H1) prob(model parameters|H1)}}{\int \text{prob(data|H0) prob(model parameters|H0)}}$$

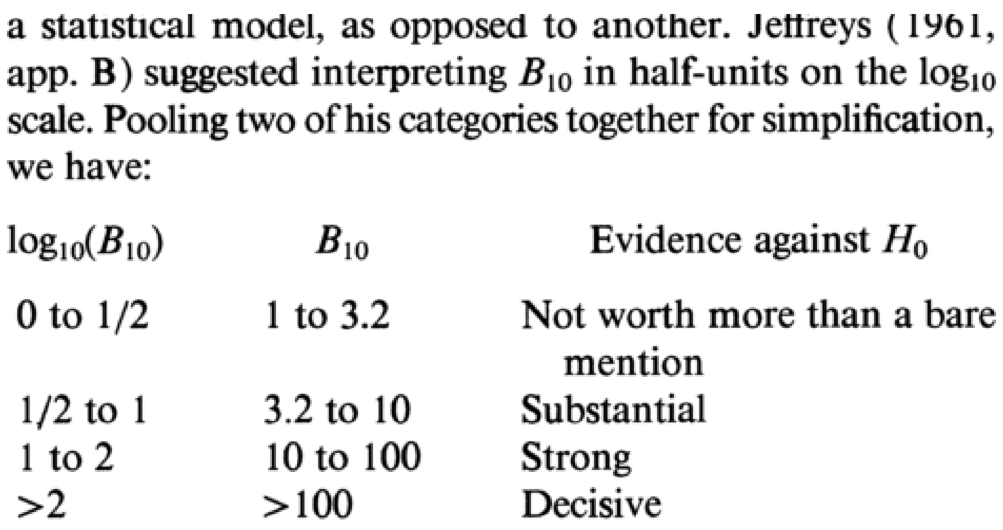
# Making choices

Because it gives the odds, the Bayes factor (or generalizations of the idea to more than two competing models) is very attractive in dealing with real-world questions of the type

“What are these data telling me to do next?”

...and it is natural to ask, what are compelling odds?

# Strength of evidence ideas



a statistical model, as opposed to another. Jeffreys (1961, app. B) suggested interpreting $B_{10}$ in half-units on the $\log_{10}$ scale. Pooling two of his categories together for simplification, we have:

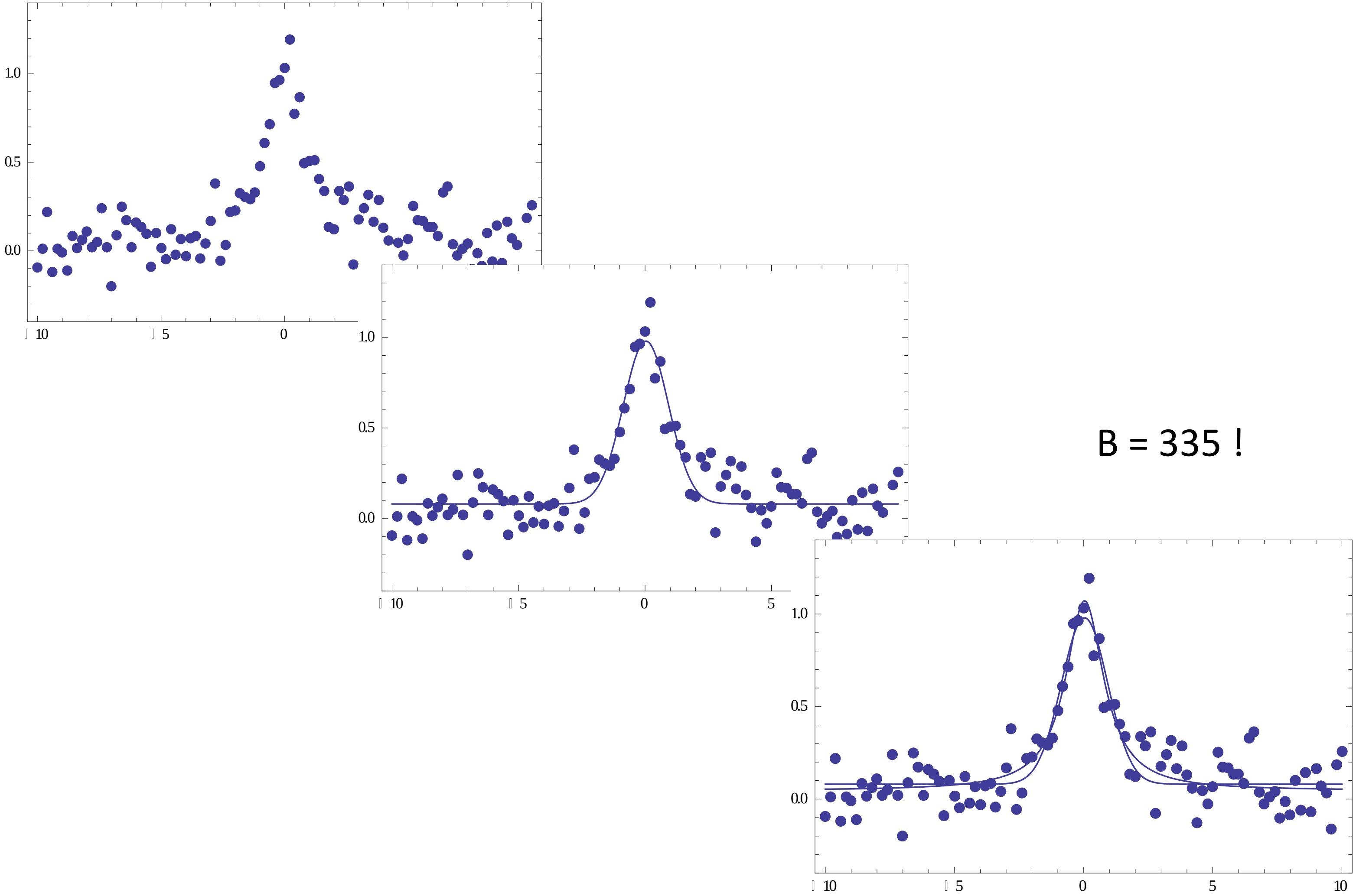
| $\log_{10}(B_{10})$ | $B_{10}$ | Evidence against $H_0$ |
|---|---|---|
| 0 to 1/2 | 1 to 3.2 | Not worth more than a bare mention |
| 1/2 to 1 | 3.2 to 10 | Substantial |
| 1 to 2 | 10 to 100 | Strong |
| >2 | >100 | Decisive |

From Kass & Raftery 1995

# A simple example

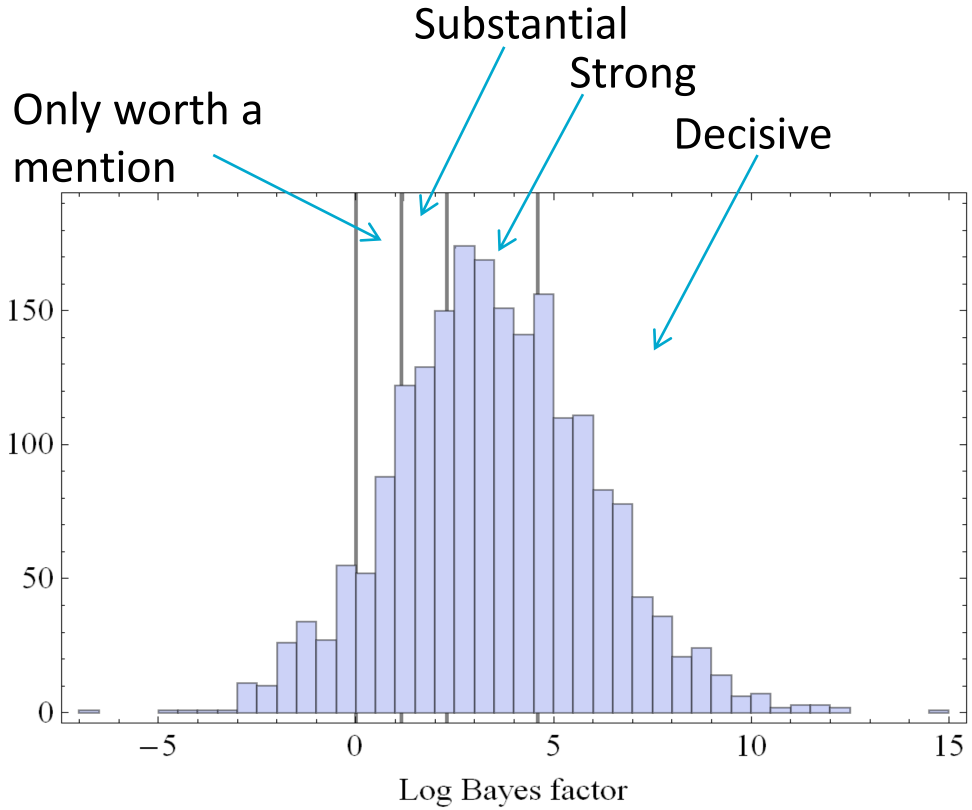
- Some fairly low SNR spectroscopic data with just one line
- We ask, is this line Gaussian? Or is it Lorentzian (power-law wings)?
- We imagine we are actually going to act on the conclusion we draw; we will not just write down the posterior odds on the Lorentzian, but we will then do something based on how good those odds are.
- Is it good enough to have strong evidence? Decisive evidence? If the risks of not acting were high enough, might we act on evidence "not worth more than a bare mention"?

# An example simulation



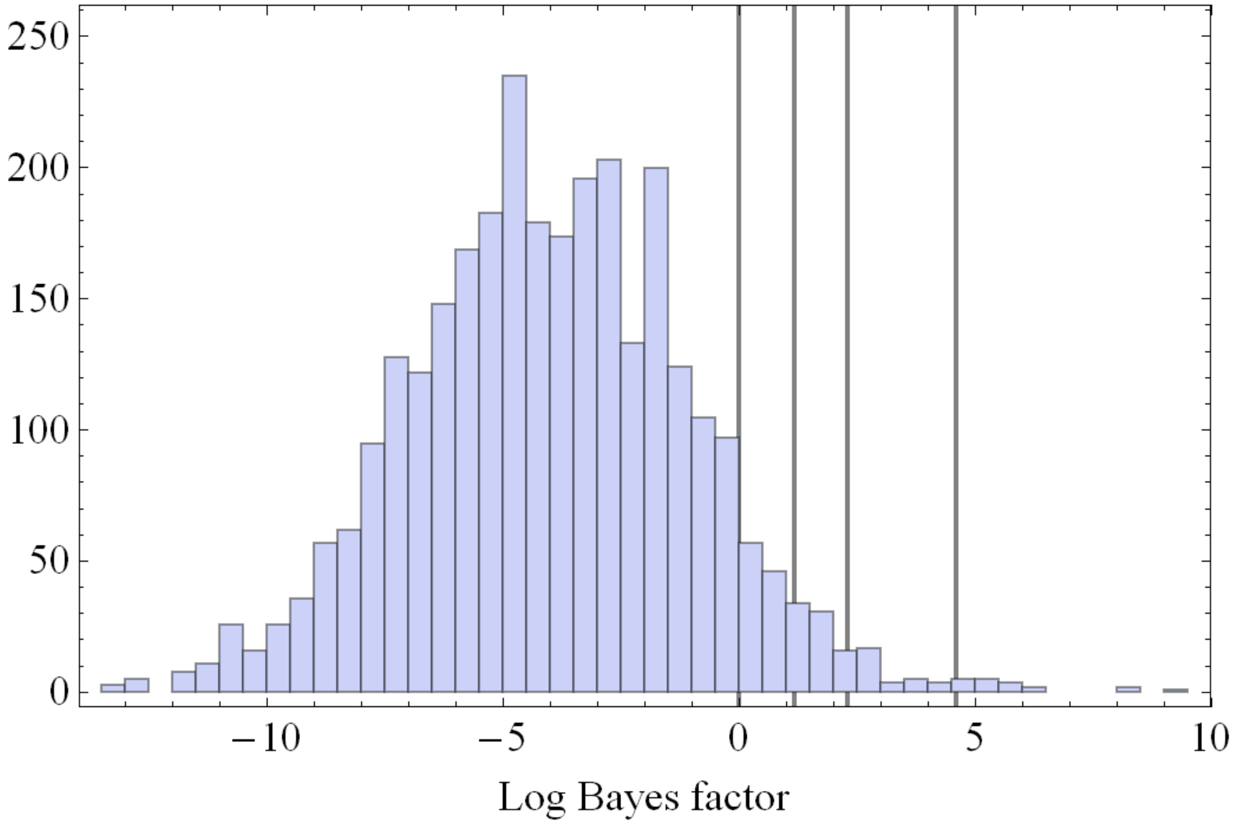
B = 335 !

# Spread in the odds on...



The line profile is actually Lorentzian; the Bayes factor is in the sense Lorentzian/Gaussian.

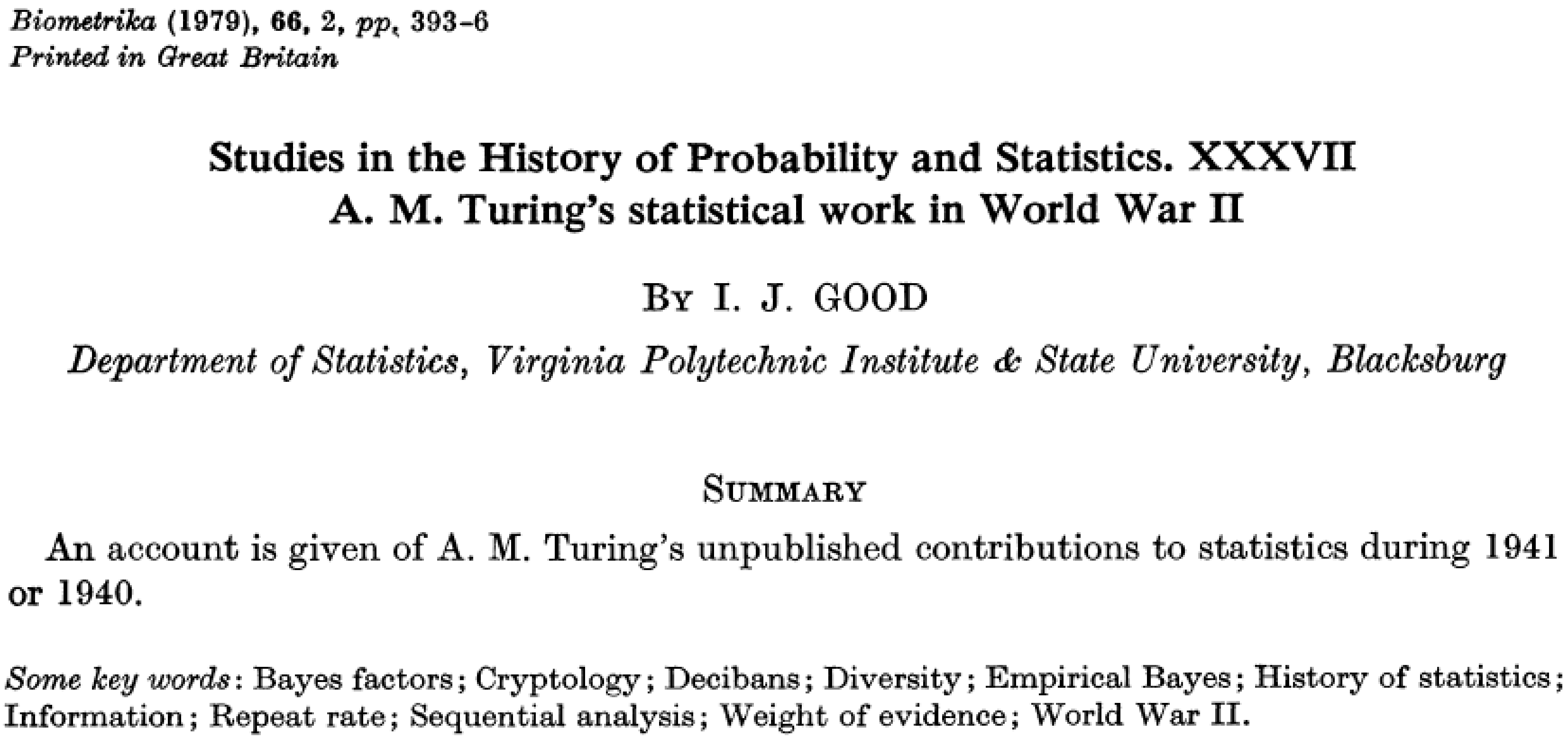
# ...and odds against



The line profile is actually Gaussian; the Bayes factor is again in the sense Lorentzian/Gaussian.

# A general result

- We see that the Bayes factor has a large spread upon repeated realizations of the data.
- Various simulations, and some analytical estimates, show that the log of the Bayes factor (the “weight of evidence”) is ~ normally distributed
- If the mean of log B is $\mu$ then the standard deviation is $\alpha\sqrt{\mu}$
- $\alpha$ is typically 1 – 2 so this is a big effect.
- It turns out Turing knew this:



*Biometrika* (1979), **66**, 2, *pp*. 393–6
*Printed in Great Britain*

# Studies in the History of Probability and Statistics. XXXVII
# A. M. Turing's statistical work in World War II

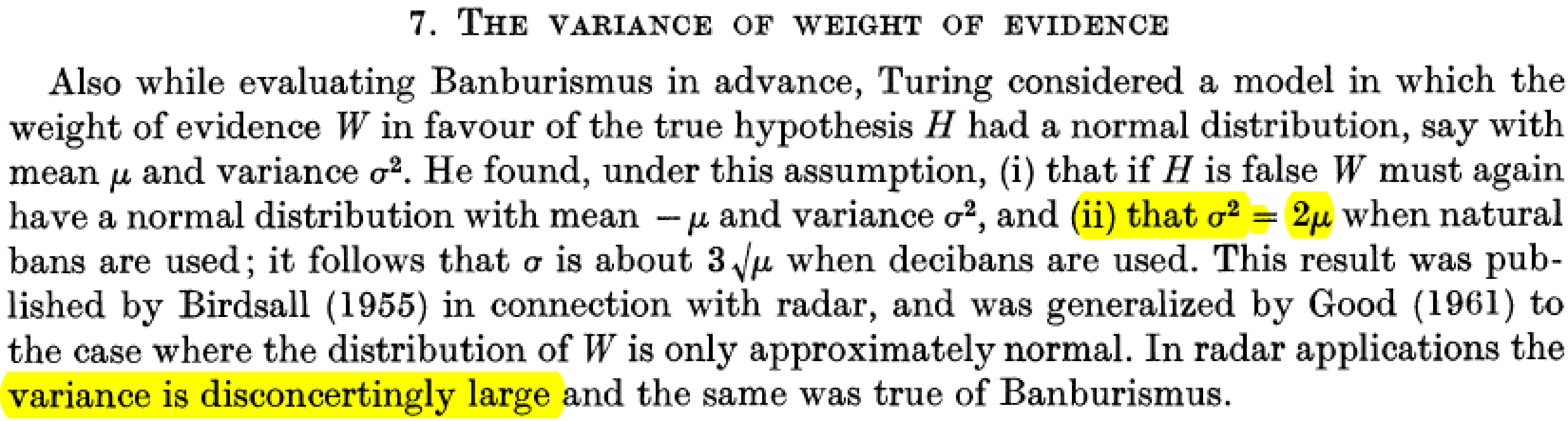
BY I. J. GOOD

*Department of Statistics, Virginia Polytechnic Institute & State University, Blacksburg*

## SUMMARY

An account is given of A. M. Turing's unpublished contributions to statistics during 1941 or 1940.

*Some key words*: Bayes factors; Cryptology; Decibans; Diversity; Empirical Bayes; History of statistics; Information; Repeat rate; Sequential analysis; Weight of evidence; World War II.



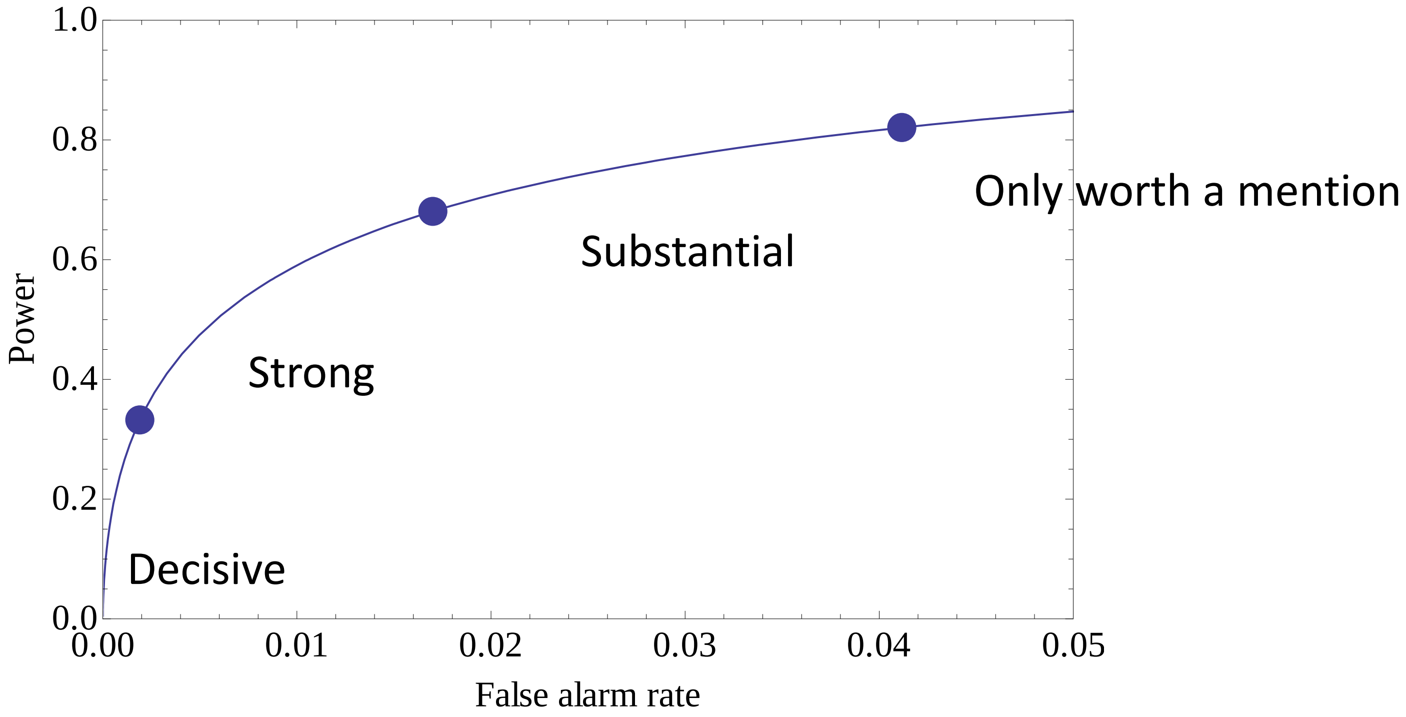
## 7. The variance of weight of evidence

Also while evaluating Banburismus in advance, Turing considered a model in which the weight of evidence $W$ in favour of the true hypothesis $H$ had a normal distribution, say with mean $\mu$ and variance $\sigma^2$. He found, under this assumption, (i) that if $H$ is false $W$ must again have a normal distribution with mean $-\mu$ and variance $\sigma^2$, and (ii) that $\sigma^2 = 2\mu$ when natural bans are used; it follows that $\sigma$ is about $3\sqrt{\mu}$ when decibans are used. This result was published by Birdsall (1955) in connection with radar, and was generalized by Good (1961) to the case where the distribution of $W$ is only approximately normal. In radar applications the variance is disconcertingly large and the same was true of Banburismus.

# What next?

- The posterior odds approach remains attractive, for all the familiar reasons of principle associated with the Bayesian approach.
- However, the wide spread may lead to bad decisions in yes/no applications: Good remarks that it is *“horrifying in relation to radar”* for example
- We will look at our toy example in two ways –
  - The classical Neyman-Pearson approach, asking about the statistical power and false alarm rate, based on a threshold in the posterior odds
  - Via the Positive Predictive Value – what’s the chance that H1 is true given that I have observed the posterior odds to be over a certain threshold?
- Each of these might be useful approaches in different circumstances
- Assume prior odds are 1 for the examples

# Neyman-Pearson and “ROC” diagram



Power: chance that the posterior odds exceed the threshold, given that a Lorentzian model applies. False alarm rate: chance that the posterior odds exceed the threshold, given that a Lorentzian model does ***not*** apply (and hence the Gaussian model must apply).

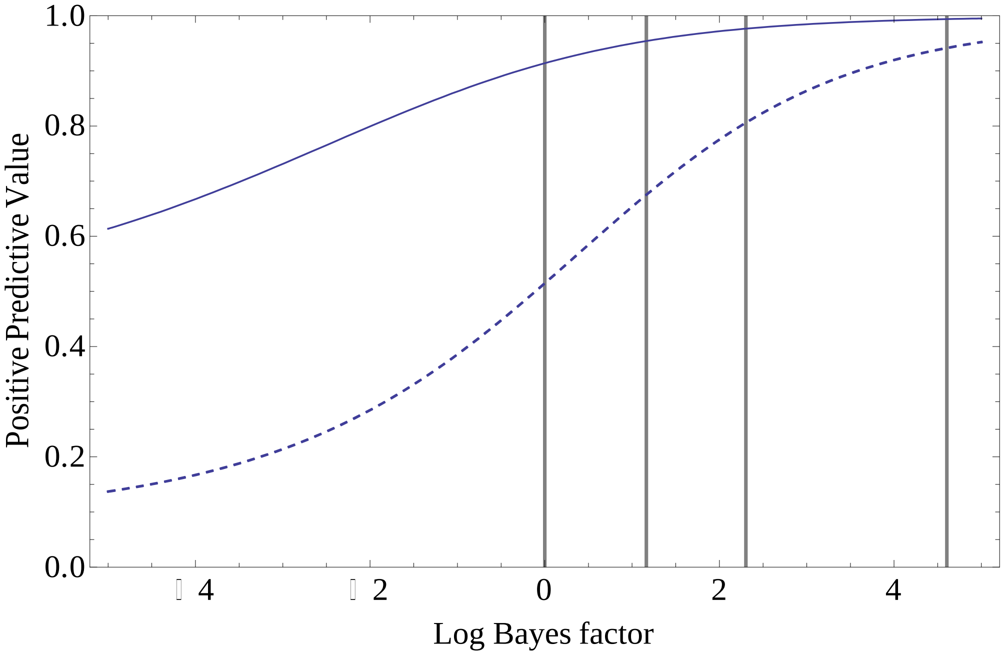
# Positive Predictive Value

Only worth a mention

Substantial

Strong

Decisive



Power

$$\mathrm{PPV} = \frac{\Theta \mathscr{P}}{\alpha + \Theta \mathscr{P}}$$

False alarm rate

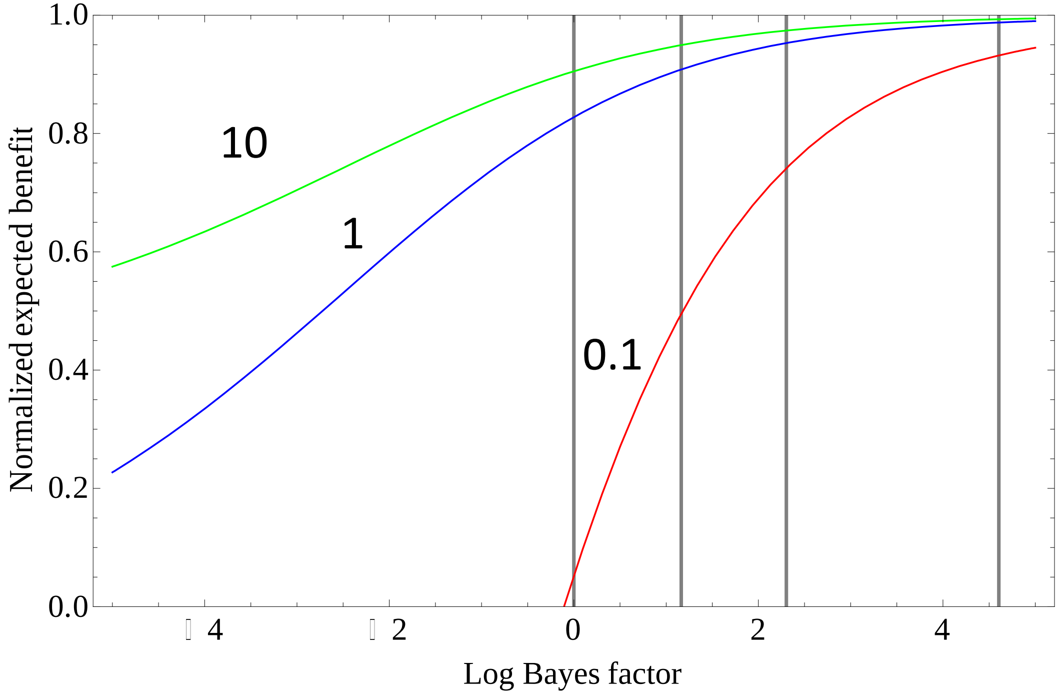
Prior odds

PPV: the chance that the Lorenzian model is right if we record a Bayes factor above the given threshold. Dashed line: prior odds are 10:1 for a Gaussian

# Values

- In the background of any discussion about decisions is the question of the "loss function" – what is the payoff for getting it right? – what is the penalty for getting it wrong?
- Good's remark "horrifying in respect to radar" illustrates the point. If we shoot down everything that we see in our airspace we have 100% success rate against the enemy.  Isn't that what we want?
- Final example: the expected gain/loss, computed from the PPV for various cases.

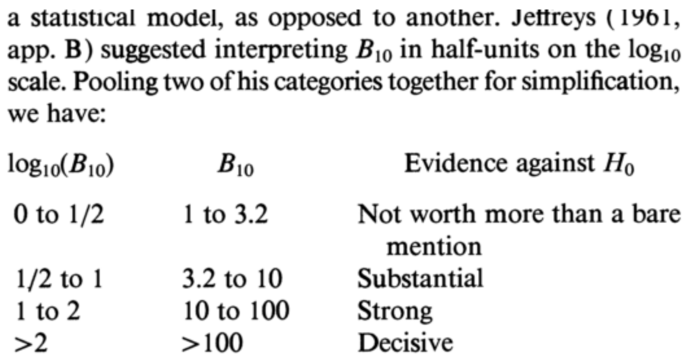
# Adding some benefit



The curves are indexed with the ratio |gain/loss|: gain if you correctly pick the Lorenztian, loss if you pick it wrongly

# Assessment

- The random spread in log B will extend across all these categories
- Decisive evidence is actually very cautious in a Neyman-Pearson sense



a statistical model, as opposed to another. Jeffreys (1961, app. B) suggested interpreting $B_{10}$ in half-units on the $\log_{10}$ scale. Pooling two of his categories together for simplification, we have:

| $\log_{10}(B_{10})$ | $B_{10}$ | Evidence against $H_0$ |
|---|---|---|
| 0 to 1/2 | 1 to 3.2 | Not worth more than a bare mention |
| 1/2 to 1 | 3.2 to 10 | Substantial |
| 1 to 2 | 10 to 100 | Strong |
| >2 | >100 | Decisive |

- Viewed through the PPV lens, the categories are discriminatory if the prior odds on H1 are small – also a kind of caution
- This is also true if the penalty for wrongly rejecting H0 is considerable – a similar kind of caution
- From this perspective the Bayes factor seems to have a connection with the classical p-value (rejecting the null) as discussed most recently by Johnson (PNAS November 2013)

# Conclusions

- The posterior odds, or equivalently the Bayes factor or weight of evidence, is an attractive method for taking principled decisions
- However, these quantities have *considerable* scatter under repeated realizations of the data
- This focuses attention on the chances of taking the wrong decision...
- ...but the weight to attach to this, and hence the weight of evidence required, is a question of values, not probability