BAYES, BANJOS, BAD TEST SCORES, AND BIG SCIENCE

Brian W Junker
Department of Statistics
Carnegie Mellon University
Professional Statistics: brian@stat.cmu.edu
Everything Else: junkerb@gmail.com

Topics

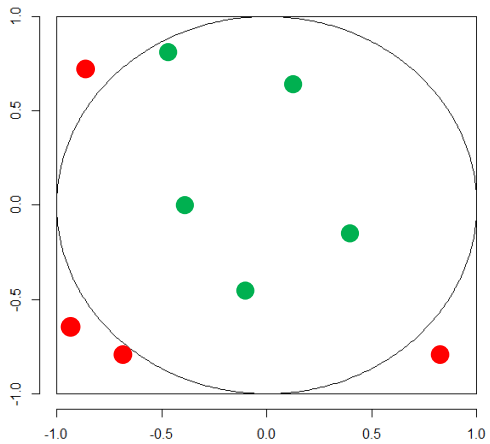
- Bayes
  - Probability on a Dartboard, Conditional Probability
  - The Reverend Thomas Bayes (not a little-known Unitarian!)
  - Bayes' Theorem
- Banjos
  - Getting the beat in music
- Bad Test Scores
  - Scoring the SAT and similar tests
  - What to do with a zero or a 100%??
- Big Science
  - Big data in science, scalable computation, and Bayes
  - Bayesian Brains?

Probability on a Dartboard

- Suppose I throw a dart at a circular target on a square board.
- Let's let Event A = "dart lands inside the circle"
  - Each dart is an independent trial on which "A" can happen
  - If darts land equally likely anywhere in the square, then the probability of "A" will be proportional to the ratio of the area of the circle to the area of the square

Probability on a Dartboard

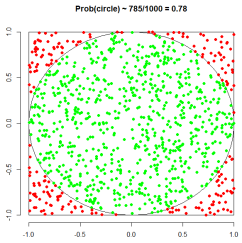
- A = dart inside circle
- Each dart leaves a mark:
  - Green dot: "A" happened
  - Red dot: "A" failed
- (R demonstration)…
  - R: prob.circ()



- If the darts are equally likely to land in any equal area,

$$P(A) = \frac{(\text{area of circle})}{(\text{area of square})} = \pi/4$$

Aside: Monte Carlo estimation of π



- If P(A) = π/4 then $\pi = 4 \times p$
- If we have estimated

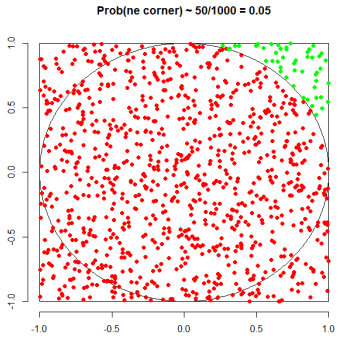
$$P(A) = \frac{\#(\text{green darts})}{\#(\text{green darts}) + \#(\text{red darts})} = 0.78$$

then $\pi \approx 4 \times 0.78 = 3.12$

- ***This method is called "Monte Carlo estimation". It was first named and applied to difficult problems by the Manhattan Project folks***
  - The estimate gets better and better the more darts we throw, by the "*Law of Large Numbers*"
  - (but that is a different presentation!)

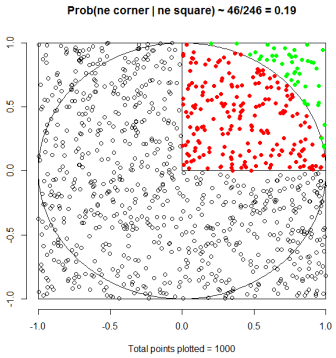
Conditional Probability

- Now, let A = "dart lands in NE corner"



- What is P(A)?
  - R: prob.ne.corner()

- Let B = "dart lands in NE square"



- What is P(A|B)?
  - R: cond.prob.ne.corner()

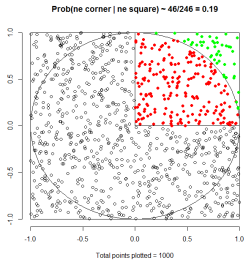
Conditional Probability

- The idea of conditional probability does not depend on darts. If A and B are two events, then we define

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

- This is entirely consistent with our estimate of P(A|B):

$$\begin{aligned} P(A|B) &\approx \frac{\#(\text{green darts})}{\#(\text{green \& red darts})} \\ &= \frac{\#(\text{green darts})/\#(\text{all darts})}{\#(\text{green \& red darts})/\#(\text{all darts})} \\ &\approx \frac{\text{P(A \& B)}}{\text{P(B)}} \end{aligned}$$



The Reverend Thomas Bayes

- 1701?—1760
- 2nd Gen Presbyterian Minister & Amateur Mathematician
- Published
  - *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures* (1731)
  - *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst* (anon.,1736)
- Wrote but did not publish
  - *An Essay towards solving a Problem in the Doctrine of Chances*

- This almost certainly is not Bayes(!)
- It is the only claimed likeness
- ***Not*** a Unitarian (or Universalist)!

Before Bayes' "Essay"

- Official statistics have been kept since 2200BC (China)
- By the mid-1700's in Europe
  - Statistical Tables were kept for taxation and other state purposes
  - Observations were combined to reduce measurement error (especially in Astronomy)
    - Calculations were developed to show how much better the mean was, than an individual observation, as a measurement
    - Least squares was used as a curve-fitting tool
    - Gauss would later (ca. 1795) develop the normal distribution to place least squares on more principled footing
- Bernoulli & deMoivre knew how to take initial conditions B and derive probabilities of events A from them ***(probability)***
- BUT they could not take a set of events A and infer the initial conditions B from them ***(inverse probability)***

Contribution of Bayes' "Essay"

- Mathematicians were convinced that there should be a principled approach to ***inverse probability***
  - But they couldn't see it because their examples were too simple (cards and games of chance).
- Bayes came up with a better example in which a principled theorem could be derived and applied
- In modern, simplified notation, ***Bayes' Theorem*** is

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

- Although the algebra is easy (next slide!) it was a big breakthrough to be able to write this down.

Bayes' Theorem on Inverse Probability

- Suppose we know P(A), P(B), and P(A|B).
- We want the "inverse probability" P(B|A).

$$P(B|A) = \frac{P(B\&A)}{P(A)} \qquad \text{Definition of } P(B|A)$$

$$= \frac{P(A|B)\cdot P(B)}{P(A)} \qquad \text{Since } P(A|B) = P(A\&B)/P(B)!$$

- So $$\boxed{P(B|A) = \frac{P(A|B)\cdot P(B)}{P(A)}}$$
- (Did you blink?)

Why was this difficult for Bayes' contemporaries?

- Let A = "observable event", B = "initial conditions"
- For decades mathematicians had been calculating
  - P(A) = P(observable event)
  - P(A|B) = P(observable event | initial conditions)

  without ever having to worry about whether

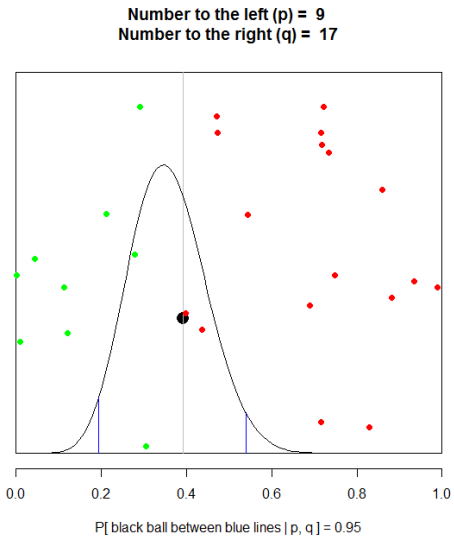
  P(B) = P(initial conditions) made sense
  - Indeed, they would "set" the initial conditions, so there didn't seem to be any probability to attach to B!
- Bayes had to argue that such expressions ***did*** make sense!
  - He did so by setting up an experiment in which the "initial conditions" were also randomly determined

Bayes' Billiard Table (he didn't call it that!)

- Let's play a game!
- I will randomly place a black ball on a pool table
  - This is the *initial condition* (it is *random!*)
- You get to blindly place a ball on the table 1000 times
  - I will only tell you if it is to the left or to the right of my ball
  - These are the *observable events!*
- Can you guess where my ball is, from
  - p = # of your balls to the left of my black ball
  - q = # of your balls to the right of my black ball
- R: billiards()

Bayes' Billiard Table

- Easy to "guess" the black ball is at 9/(9+17) = 0.35
- With his theorem, Bayes could also derive
  - A probability curve showing where the black ball was likely or not likely to be
  - A ***recipe*** for an interval where the black ball was likely to be!
- Bayes could measure *uncertainty* in the guess!



Why didn't Bayes publish?

- For years, famous statisticians claimed that Bayes was troubled by the idea of P(initial conditions).
  - ***However***, modern interpretations of Bayes' Essay suggest that he was absolutely certain he was right about that part of the problem.
    - Probability can be defined in ways other than relative frequency (think about the weather, or making book on NFL games)
- The calculation of the probability curve, and the interval where the black ball was likely to be, involved calculus
  - ***Easy hand calculation***, if p, q small
  - ***Impossible hand calculation***, if p, q large
- Bayes realized that he had a ***recipe that could not be followed***, if the curve and interval could not be calculated.

Aside: Stigler's Law of Eponymy

- ***Thomas Simpson*** beat Bayes to it: he published a partial solution to the problem of inverse probability a few years before Bayes
  - Bayes probably read Simpson's work and may have been trying to improve on it
- ***Bayes*** gave a better partial solution, but only in terms of the billiards example
  - He shared the Essay with essentially no one while he was alive
  - Posthumously edited, updated and published by ***Richard Price***
- ***Pierre-Simon Laplace*** independently discovered these ideas and published them in a complete, general form
- *And so, of course, we call it Bayes' Theorem.*

Bayesian Decision-Making

- ***States of nature:*** $B_1, B_2, \ldots, B_N$
- ***Observable event:*** A
- ***Decision:*** Which B_k given, A?

- ***Prior distribution:*** $P(B_1), P(B_2), \ldots, P(B_N)$
- ***Likelihood function:***

  $P(A|B_1), P(A|B_2), \ldots, P(A|B_N)$
- ***Posterior distribution:*** Inverse probability of each B_k,:
  - $P(B_1|A) = P(A|B_1) \cdot P(B_1) / P(A)$
  - $P(B_2|A) = P(A|B_2) \cdot P(B_2) / P(A)$
  - …
  - $P(B_N|A) = P(A|B_N) \cdot P(B_N) / P(A)$

Decision: Choose the B_k with maximum $P(B_k|A)$

Bayesian Decision-Making

- Each posterior probability has same factor 1/ P(A); can omit this and get the same decision:
  - $P(B_k|A) = P(A|B_k) \cdot P(B_k) / P(A) \propto P(A|B_k) \cdot P(B_k)$

"posterior is proportional to likelihood times prior"

- Interpretation: Bayes' Theorem lets you combine
  - Past experience about prevalence of each B_k (prior)
  - Consistency of the data A with each B_k (likelihood)

  in making a decision (in a way that minimizes expected loss [or maximizes expected gain]…)

Banjos: What is the beat?

- ***How do we recognize the "beat" in music?***
  David Temperly (2007) argues that this decision process can be modeled with Bayes Rule.
- The most likely time signature will be the one that maximizes likelihood times prior

$$P(\text{time sig} \mid \text{music}) \propto P(\text{music} \mid \text{time sig}) \cdot P(\text{time sig})$$

Posterior likelihood prior

(actually time signatures are not in 1-1 correspondence with the rhythmic "grid" of a music, any more than unreduced fractions are in 1-1 correspondence with the numbers they represent, but we will "go with it" for this talk…)

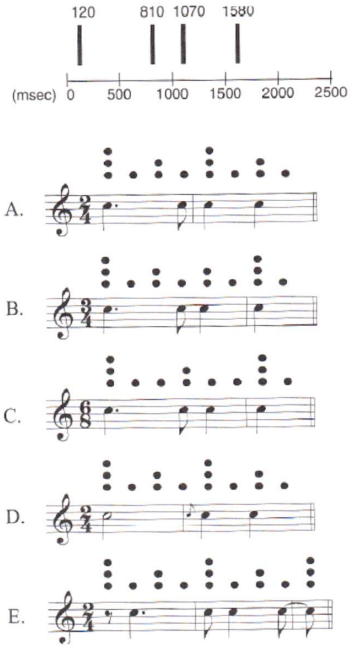
A prior distribution for beats: P(time sig)

- Our musical experience suggests some time signatures are more likely than others:
  - Most common 4/4, 2/2
  - Next most common 3/4
  - Less common 6/8, 9/8, 12/8
  - More rare 5/4, 5/8, 7/8
    - and various mixed time signatures (Blue Rondo a la Turk!)
- Suggests a prior distribution like
  - P(4/4 or 2/2) = 0.70
  - P(3/4) = 0.15
  - P(6/8) = 0.06
  - P(5/4) = 0.05
  - P(others) = 0.04

I just made these numbers up. An "actual" prior, based on a corpus of European folk songs, is in Temperly's book.

A likelihood for beats: P(music | time sig)

- Human-made music doesn't line up perfectly with rhythmic grid
- But some are more likely than others!
- Temperly develops P(music | time sig) based on this.



Can you guess the time signature?

- Example 1: Waterbound
- Example 2: Trick of the Light
- Example 3: Elk River Blues
- Think about your "recognition process" in terms of …

$$P(\text{time sig} \mid \text{music}) \propto P(\text{music} \mid \text{time sig}) \cdot P(\text{time sig})$$

Music and Probability

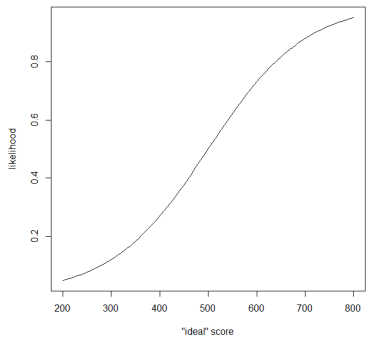
- Temperly applies the same Bayesian principles to
  - Recognizing the rhythmic grid of a music
  - Recognizing the key from a monophonic melody
  - Recognizing the key from a polyphonic melody
  - Etc
- He then discusses the consequences of this for
  - Musical expectation and error detection
  - Recognizing different musical styles
  - Scope for deviation from a musical style
  - Etc.
- (fun book! $20 at Amazon.com…)

Bad Test Scores

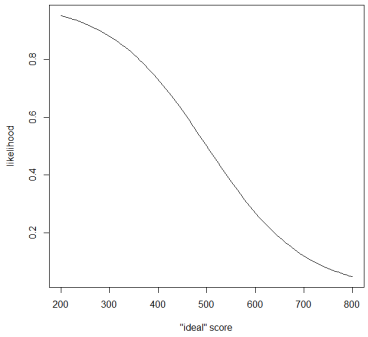
- Standardized tests are a big industry in the US; see Zwick (2002) for an overview.
- Many tests are not scored on simply the “number right”.
- Instead
  - a likelihood is formed,
  - and a parameter (kind of an “ideal test score”) is estimated from the likelihood, from each student’s test score

Building the likelihood

- The “test score” likelihood is a product of functions:
  - For each correct answer, a function like this appears in the product:



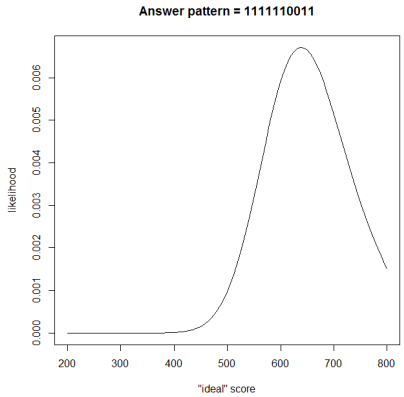
  - For each wrong answer, a function like this appears in the product:



  - R: test.score.likelihood()

Estimating the ideal test score

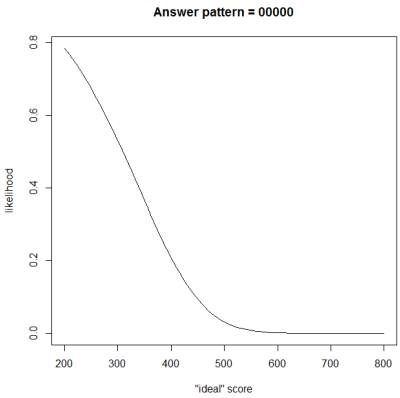
- If you get some wrong and some right, your likelihood will look like this, and the “ideal score” that maximizes the likelihood is a reasonable estimate:



- This is “maximum likelihood estimation” (MLE)

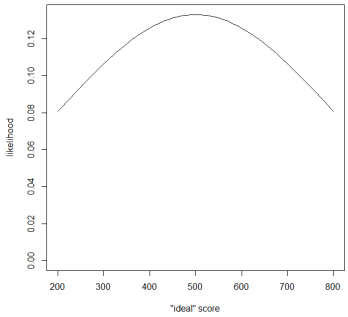
Problem: What about 0% or 100% right?

- The likelihood for 0%



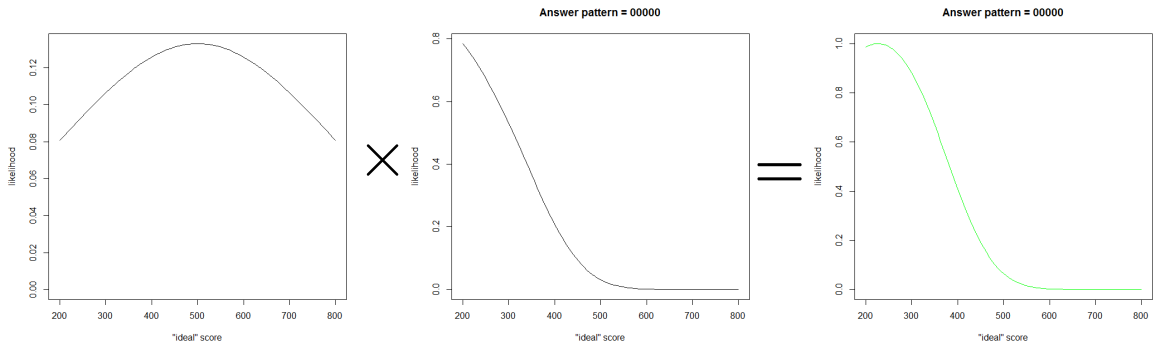
- The MLE is $-\infty$ (negative infinity)

- Most (ideal) test scores should be "in the middle", with a few low and a few high ones. We might have a prior distribution like this:



Fixing infinite test scores

- When we multiply the troublesome likelihood by this prior, we get a posterior distribution like this:



- And the maximum of this becomes our new estimate.
  - R: test.score.both()

Bayesian Regularization

- Incorporating prior "beliefs" makes the problem of finding the ideal score manageable.
  - It doesn't make infinite scores impossible, just very unlikely.
- Much practical Bayesian statistics is like this.
  - The prior distribution "regularizes" the problem enough to make it solvable.
- This "regularization" is how I first became acquainted with Bayesian methods!

Advantages of Bayes in Science

- Regularization
- Borrowing strength from past experience (or concurrent data-rich settings) to improve inferences
- Language of probability
  - Common approach to formulating problems and solutions
  - Tools to find optimal solutions
  - Vast expressive power to formulate many scientific problems in probabilistic terms

Bayesian Successes

- In small to medium-sized problems, there has been a Bayesian revolution
  - Enabled by cheap, fast computers with capacious memory
- Some examples:
  - Fair methods for scoring tests and extracting as much information as possible from them
  - Analysis of complex federal surveys
  - Analysis of face to face (or online!) social networks
  - Pooling information from multiple raters (e.g. in radiology or job performance), and learning about individual differences among raters
  - Describing and understanding trajectories toward Alzheimers' Disease
- Other examples:
  - Inferring treatment success from historical medical records
  - Nate Silver's 538 (and PEC, and …) predictions!
  - Spam detection, email/word completion, …

Bayesian Limitations

- "Computational scalability"
  - We know how to solve small problems but when the same problem gets large, we can't do the computation
- Thomas Bayes experienced this:
  - For small p, q, any freshman in calculus could solve the problem of estimating where the black ball was
  - For large p, q, this solution is tedious, and too slow for a human to carry out
- With modern computers, we can extend the reach of Bayes theorem to modestly large problems
  - N = # observations
  - P = # features per observation
- N ~ 1,000,000, P ~ 1,000 is sort of the Bayes ballpark

Big Data and Bayes

- Many current "Big Science" & "Big Commerce" projects involve truly big data (large N, large P!), e.g.
  - Sloan digital sky survey (N=500,000,000 objects)
  - Online recommender systems (N=300,000 of purchases/day)
  - Google search logs (N=400,000,000 searches/day)
  - Large Hadron Collider at Fermilab (N=10^3--10^9 or more collisions per second)
  - Microarray (gene expression) data (N=10^3 observations, P=10^6 features)
- Three problems
  - *Computation may be too slow* (Thomas Bayes' problem for supercomputers!)
  - *Likelihood may be too complex* to be represented in the computer
  - *Curse of dimensionality*: as P grows, the number of probes needed to search a volume of radius R is proportional to R^P

Strategies for Big Data

- Reduce P: Dimension Reduction & Feature Extraction
- Reduce N: Subsampling

- Focus computation on a narrow inferential goal
  - Don't try to represent the whole likelihood or the whole posterior

- Scalable computation – take at most one pass through the data!

- Hastie, Tibshirani & Freedman (2009) survey some basic methods
  - Some methods are "provably correct"; others are just heuristic

Aside: Bayesian Brains?

- "All people are Bayesians, some just haven't figured that out about themselves yet"
  - I kind of doubt it
- The same problems with representing models, and doing exact computations on them, exist for computers or human brains
  - Human brains are not especially fast or efficient at computation
- Most likely, humans adopt heuristics that reduce to (approximately) Bayesian thinking, for small problems.
  - In this sense, we probably handle inference the way big data inferences area handled

References

- Stigler, S. (1980). *The history of statistics: the measurement of uncertainty before 1900.* Cambridge MA: Harvard University Press.
- Temperly, D. (2007). *Music and probability.* Cambridge MA: MIT Press.
- Zwick, R. (2002). *Fair game? The use of standardized tests in America.* New York: Routledge-Farmer.
- Hastie, T., Tibshirani, R. & Freedman, J. (2009). *The elements of statistical learning (2nd ed).* NY: Springer. See also http://www-stat.stanford.edu/~tibs/ElemStatLearn/