Econ 8013 Notes, 2018

This course is designed to introduce the student to the theory and methods of Bayesian inference, with particular emphasis on applied work.

We will take **a (posterior) sampling approach** to statistical inference.

Texbooks are

Hahn(2014) Bayesian Methods for Management and Business, Wiley.

Kruschke (2015) Doing Bayesian Data analysis, Academic Press. (aka "Bayesian Puppies")

Software

R and RStudio

[examples can be provided in Julia, python and Matlab]

Contents

1. 8013 Intro. Notes

Deduction True or False (1 or 0, or any arbitrary values) => "I think, therefore I am." End of story!

Deductive inference: Assumptions needed to go any further – assumptions based on inductive inference.

Axioms = a consensus of expert opinion on what is self-evident/obvious (to the expert). Your deductions are only valid if your axioms are valid. You cannot deduce axioms/assumptions.

Induction/ (inductive **inference)**

```
|____________________________________________________|
0                   0 < p < 1                         1
```

definitely true - almost surely true – highly likely – maybe/don't know – highly unlikely – almost surely false – definitely false

Wikepedia (for all answers!): "**Inductive reasoning** (as opposed to *deductive* reasoning or *abductive* reasoning) is a method of reasoning in which the premises are viewed as supplying some evidence for the truth of the conclusion. While the conclusion of a deductive argument is certain, the truth of the conclusion of an inductive argument may be *probable*, based upon the evidence given. …

As a logic of induction rather than a theory of belief, Bayesian inference does not determine which beliefs are *a priori* rational, but rather determines how we should rationally change the beliefs we have when presented with evidence. We begin by committing to a prior probability for a hypothesis based on logic or previous experience, and when faced with evidence, we adjust the strength of our belief in that hypothesis in a precise manner using Bayesian logic."

Statistical inference is inductive inference.
When we talk about "inference" in statistics, we mean statistical inference aka inductive inference from data (observation).

Deductive proof (theorem proving), analytical solutions (pen and paper)
vs.
Inductive reasoning (inference), experimental evidence, simulation and numerical evaluation (computation)

[See chapter 1 of Downey (2016) Think Complexity, 2e]

Degree of plausibility of propositions or statements
(degree of belief – belief is a poor word choice as it has too much emotional baggage from religious usage)
Scientific revolution is based on inference. No inference, no science.

What is "science"? **The scientific method** and application of the scientific method

Science is a method: hypothesis | experiment/evidence | inference | decision | repeat
What is a hypothesis? A proposition (often a question posed as one possible answer to the question).

Hypothesis testing is central to the scientific method.

At the casino (of life):

Inference	=>	Decision
'S. Holmes'	then	You (the gambler)

You **first** must play detective and infer the probabilities that proposition are true or false (likelihood/chance of possible outcomes).

Only then do you make decisions based on (conditional on) the probabilities. Choosing to reject or not reject a hypothesis is a decision.

Decision making under uncertainty = decision making with incomplete information
– outcome/truth is not known with certainty.

Expected utility hypothesis = Expected outcome == $\sum_i a_i p(a_i|E)$

a_i = action/outcome i,
$p(a_i|E)$ = probability of a_i given relevant information (evidence) E. The sum is over the set of all possible outcomes.

Let p = 'degree of plausibility' of a proposition (or degree of credibility, or degree of belief – though belief is a poor word choice).

How plausible is a proposition? For proposition x = "some statement/hypothesis"
If x is definitely true given information E, $p(x|E) = 1$.
If x is definitely false given information E, $p(x|E) = 0$.
Must be true that $0 \leq p(x|E) \leq 1$.

Chance (% chance something is true) = $p(x|E) \times 100$, so just scaled from 0 to 100 instead of 0 to 1.

There is no such thing as (an objective) probability.
x is usually neither random, nor a variable. We do not need (nor necessarily want) randomization or variables.

Suppose statement x is either true or not always (i.e. it is not sometimes true and sometimes false – there is no randomness and it is not a variable). E.g. x = "The professor lived in Clifton in 2010."

[Ok, you can argue that everything is randomly determined if we go back far enough and believe everything philosophers tell you, but you can also argue that everything is deterministic if you go back far enough and believe everything physicists tell you.]

Putting aside the philosophical mumbo jumbo and trying to infer whether x is a true statement today: statement x is not a variable, it is either true or false with nothing random about it. It either happened and is true, or it didn't and so is false. **Exercise:** think about it a minute, and assign a probability (or set of probabilities) to proposition x.

Objectivity vs. subjectivity vs. scientific objectivity
Something can be subjective to the individual but scientifically objective. Objective, in the sense of beyond the experience of the individual, really has nothing to do with 'scientific objectivity' = consensus of unbiased expert opinion that something is true based on the evidence not feelings.

Subjective feelings and subjectively occurring but not based on feelings. Person A added 1+1 and got answer 2. The activity was subjective, the answer can be validated and replicated as scientifically objective. The number system (and all of mathematics) is a subjective experience – it all happens entirely in human minds. No humans = no math.

If you get hit in the head by a rock. Your pain is entirely subjective – you feel the pain, no one else does. The evidence that health care professionals gather that you feel pain is scientifically objective. We infer that someone who was whacked in the head with a rock feels pain. We cannot deduce it, and it is subjective to the individual. We inductively infer from long (painful) experience that you (most likely, almost surely) feel pain. There is still a chance that you do not actually feel pain and you are faking, but we can look at all the evidence 'objectively', and make a sound inference.

All science is subjective to the individuals involved, even when the ideal of 'scientific objectivity' is being fully maintained. Science requires a consensus of expert opinion; there is an element of democracy to it – if you are right, but no one believes you or agrees, science does not advance (you can be ignored and die a lonely and bitter outcast!).

"Science advances one funeral at a time." Max Planck, originator of quantum theory, 1918 Nobel prize in physics.

[A paraphrase of: "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." 1906]

Degrees of plausibility (aka probabilities) are subjective in the sense that they depend on the person assigning the degree of plausibility (**because the assignment depends on the information available**).

All probabilities are conditional.
S = set of all possible outcomes => $p(S|S) = 1$. Also, $p(not\ S|S) = 0$.

Must be conditional on the sample space at the least. [Draw a Venn diagram and chop parts off to see how probabilities change as the conditioning information changes.]

A probability is **not** the limit of a frequency – that concept does not apply to most (really any) situations faced in reality.

Think of rolling a 6-sided die. We don't actually infer that it is 'fair' (prob = 1/6 for all outcomes) by rolling it a large number of times; we look at it and evaluate its physical properties and the rolling conditions, then maybe run a few trials at most. We could also gather more information (though probably at great cost) about the angular momentum and spin, velocity, air speed and density, surface friction, imperfections in the die, neurological state of person tossing the die just prior to tossing (or state of machine doing the tossing), etc., and improve our predictive ability. It is only because this is generally not viable or far too costly that we throw up our hands and say p = 1/6, but as soon as the die leaves the hand, and arguably before, the outcome is certain given enough information. Also, what if we carefully place the die on a table with a particular side up, hiding the die from other observers, then ask someone to gamble on the outcome being some value, say '6'? What probability will that person assign to the proposition "the side facing upwards on the die has 6 dotes"? This is no longer a random variable, it is a fact that is either true or false, but the person asking to choose does not know the answer (yet).

2. Preliminaries: Foundational concepts

What is a probability?

Chance of hitting a baseball: 0% to 100%. You cannot "give 110%, that's impossible!
A probability is the same, only scaled from 0 to 1, so
Prob. = 0.0 if chance of statement true = 0%
Prob. = 1.0 if chance of statement true = 100%

Probability as degree of plausibility.
Degree of plausibility of any proposition/statement.

[Proposition: a statement or assertion that expresses a judgment or opinion.
Synonyms: theory, hypothesis, thesis, argument, premise, principle, theorem, concept, idea, statement]

Extension of logic from propositions being either true or false to proposition being more or less plausible. From either 0 or 1, to degree of plausibility of the truth of a proposition.

Cox, R.T. (1961), Jaynes, E.T. (2003) Probability Theory: The Logic of Science.

Consider:
"The fair 6-sided die will have one dot on the uppermost face once rolled." Instead of true or false, degrees of plausibility of that statement = 1/6

"It will rain today." Instead of "yes" or "no" on, "there is a 60% chance it will rain today."

All probabilities are conditional

We assign probabilities to propositions/statement *conditional on the information available*. There is no such thing as an unconditional probability.

Draw a Venn diagram. What is called $P\ \ A\ \ =$ Area A/Area S, is really $P\ \ A|S\ $. If we now remove some of S, or add to it, $P\ \ A$ changes.

Random variables are not relevant. Often (usually!), the statement is neither random nor a variable.

We are making statements with **incomplete information** on the truth value of the statement.

Subjective vs. Objective
There is no such thing as (an objective) probability.

De Finetti's treatise on the theory of probability begins with the
provocative statement PROBABILITY DOES NOT EXIST, meaning that probability does not exist in an objective sense. Rather, probability exists only subjectively within the minds of individuals.

Coin glued to a table. Your information set differs from mine. There is a **rational probability assignment** for each of us, and it is **different** because we have **different information** sets.

Probabilities are subjective in the sense that they depend on the information available when assigning a probability. Probabilities are representations of a state of knowledge. If knowledge is complete, then the probability (that the statement is true) is zero or one, otherwise it is somewhere between zero and one.

Probability of a hit in baseball with no information (foreign student, martian), vs. probability with some (more or less) knowledge of the game.

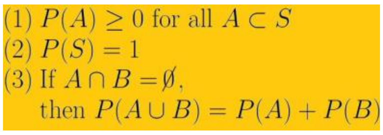
$$p\ \ A|B\ \ \ vs.\, p\ \ A|S\ \ .$$

Subjective probabilities are scientifically objective. They are conditional on the information available. They can be axiomatic – self evidently true upon reflection.

Jaynes (2003) necessarist approach: ***Two people with exactly the same information should assign the same probabilities.***

3. Bayes' rule

Follows straightforwardly from the axioms of probability (should be conditional though):



(1) $P(A) \geq 0$ for all $A \subset S$
(2) $P(S) = 1$
(3) If $A \cap B = \emptyset$,
then $P(A \cup B) = P(A) + P(B)$

Follows straight from Venn diagram with sets A and B overlapping:

$$p\ \ A|B,S\ \ = cp\ \ B|A,S\ \ p\ \ A|S\ \ .$$

This always gets written without the S, 'to simplify', but remember that it really belongs there.

The constant c is there to ensure that the probabilities sum to one. It is the inverse of the sum, over all possible outcomes (or the integral for a continuous A) of the two other terms on the RHS, i.e.

$$c = 1\ /\ \sum p\ \ B|A,S\ \ p\ \ A|S\ \ .$$

What is a probability distribution?
A table of probability values assigned to values of the quantity of interest.

From basic probability theory to statistical inference

We have some unknown quantity or proposition we want to make inferences about, call it "proposition A" and label it θ. It does not have to be a random variable, it does not have to be a variable or a parameter. As long as we are asking questions about the proposition without knowing the answer, we can assign degrees of plausibility to the various possible answer. It could be: is it going to rain at all in the next hour, or is Carlos going to get a hit when he comes up to bat in the game tonight, or is it the Dow Jones Ind. Ave. going to increase tomorrow? We reword those as propositions or statements and do some detective work to assign a degree of plausibility, based on all the evidence we have available.

Statistical inference is about reverse engineering from the evidence to the DGP. It is about playing detective.

DGP | $\rightarrow$ data (observations)

As a detective, we are interested in learning about some facet of the DGP.

For some observed outcomes, y_i, (at bats, rolls of the die, flips of the coin, stock prices, output, unemployment, sales, etc.), we

In this setting, Bayes' rule can be written as,

$$p\ \theta|y\ \ \propto p(y|\theta)p(\theta).$$

posterior is proportional to *likelihood* times *prior*.

Likelihood

Should really condition on information and assumptions made in choosing a model

$$p(y\ \theta, M, I_y),$$

where $M =$ model assumed, $I_y =$ information about the DGP for y used.

In general, there are a **lot** of restrictive assumptions in these.

The parameters of the model, θ, are just the constants in the functional form of the model chosen for $p(y|.)$.

Our models are approximations to reality. Reality is generally far more complex than our approximating model.

Prior

Given our model choice, M, the prior distribution = probability of various possible values of θ being the best choice to represent the DGP for y.

The model parameters are obviously dependent on the model chosen. "Prior information", I_θ, is our information\knowledge, before we examine the data, of what the likely values are that should be assigned to θ for our model to best approximate the DGP.

The prior distribution should really be written as,

$$p\ \theta|M, I_\theta\ \ .$$

We generally drop all the conditioning terms to keep things looking clearer, but it is important to realize that all these probabilities are conditional on the information used to assign them. The information and assumptions used to select a prior are usually pretty trivial in comparison to the information and assumptions used to select the likelihood.

$$p\ \theta|y\ \ \propto p(y|\theta)p(\theta).$$

$$p\ \theta|y\ \ = cp(y|\theta)p(\theta).$$

$$c = 1/\int p\ y|\theta\ p\ \theta\ d\theta$$

Since c is just a constant, it can usually be ignored (or computed at the end).

4. Using Bayes' rule for inference

To use the above equation, we must make three decisions [In my opinion, Hahn (2014) puts them in the wrong order].

1. Choose a model for the DGP of y

Pick a particular structure, an approximating equation (or set of equations) based on what we know (or think we know) about the process generating the observations for y.

E.g. *Bernoulli type experiment* (flipping a coin, chance of rain today, getting a hit in baseball when at bat, customer buys or not, etc.)

E.g. simple linear approximation given some explanatory variable, x.

In the Bernoulli experiment, we can show convincingly that the Binomial distribution is a very good model of the DGP. This is a bit confusing though, because we combine decisions one and two (below). The 'model' is really just than chance of success = a constant, i.e. $p\ \ s|M,I\ \ =\theta,$ where θ is a fixed parameter.

Linear approximation regression example:
For a normal distribution, the two parameters that define the distribution happen to equal the mean and the variance, so the value of 'parameter number one' and the mean happen to be the same. We choose a simple linear function $y_i=\alpha+\beta x_i+e_i$, where e_i is the approximation error.

2. Choose a "sampling distribution" aka likelihood function

"Selecting the likelihood function for our data involves thinking about what kind of data we have and what distribution might be appropriate for representing and understanding our data." [Hahn (2014), p. 13]

In order to do statistical inference, we need a probability distribution.

E.g., for the Bernoulli trials set up, the Binomial is the obvious choice

E.g., for the linear approximation, we have some very good justifications for choosing a Normal (aka Gaussian) for the approximation error, namely CLTs, information theory (the maximum entropy principle), and common sense intuition.

Often it is a bit arbitrary how we define the model vs. the distribution, so choices one and two really go together. In the Bernoulli trials case, there isn't a sensible way to really separate them.

3. Choose a prior distribution

Once we have chosen a model and distribution, we must choose a probability distribution to represent what we know (or assume to be known) about the parameters of the model and the distribution chosen above.

Often we either don't have much information about the parameters, or want to act as if we don't have much information, so that we can "*Let the data speak for themselves*." In this case, we use "uninformative" or "reference" priors that represent very little knowledge about the parameters. We can consider these priors to be "axiomatic", in that any reasonable person should be willing to agree with the information content assumed.

Axiomatic priors
What is an axiom? Self evident (to the expert after some serious thought and debate!). A proposition that needs no proof because it is "obviously" true.

An axiomatic prior then, is one that is a reasonable representation of the information available to everyone.

We can all agree that there are only 6 possible outcomes when rolling a 6 sided die (we can reroll if something weird happens), so probability of one of the sides occurring must = 1, and each side has probability between zero and one of occurring:

$$0 \leq p\ \ S_i|I_0\ \ \leq 1, \quad \sum_i p\ \ S_i|I_0\ \ = 1.0,$$

for $i = 1, \ldots ,6.$

A uniform prior on the chance of hitting a baseball or rain tomorrow is difficult to argue against. We clearly know that the chance is somewhere between 0 and 100%!

For a regression model parameter, we know that it lies somewhere on the real line (between + and – infinity), though often we can do much better than that with just a little thought.

For a variance parameter, we know that it must be positive and nonzero (you can't have negative spread, and if the variance is zero, the variable is no longer a variable, it is a constant). Also, we often have reason to expect the variance to more likely be relatively small than large (since we are attempting to minimize the approximation errors), and certainly no larger than the total variance in the dependent variable.

Check robustness of prior
We can also always (and we should always) **check how sensitive the results are to reasonable variations in the prior.**

5. Bernoulli trials example

Consider a coin toss, hitting a baseball ("hit" or not), etc.

Beware! Point of confusion: Note that this example can be confusing because the unknown parameter of interest is also the probability, so we end up making probability statements about a probability. That is, if the chance of a six on a (possibly unfair) die is θ, then it is also true that $p\ \ x = 6|\theta\ \ = \theta$.

Think of the unknown parameter in this case as simply (one of) the parameter(s) of the distribution. It just happens to be the case that the probability of an outcome is equal to that parameter value the way we build our model.

Now let's start with the proportions example of Bernoulli trials. See examples on p.12-16 of Hahn (2014). Consider a coin toss, rain or not today, chance of hitting a baseball:

There is an obvious choice for the model and likelihood in this case. We can show, with some straightforward algebra, that the DGP can be represented acutely by

$$p \;\; y|n,\theta \;\; \propto \theta^y(1-\theta^{n-y},$$

where,
$y =$ number of success
$n =$ number of trials
θ = chance of success on each trial (probability of success).

For the examples, a success is a head (or a tail, pick one!), success = it rains, success = hit.

This gives the likelihood function.

The normalizing constant can be shown to be $\binom{n}{y} = \frac{n!}{y! \;\; n-y \;!}$

An obvious choice for an uninformative prior is the uniform on [0,1], i.e.
We know it must be true that $0 \leq \theta \leq 1.$ If we allow for any of these values to be equally likely, then $I_\theta =$ set of all θ such that $0 \leq \theta \leq 1,$ and

$$p \;\; \theta|M, I_\theta \;\; = 1, \quad 0 \leq \theta \leq 1,$$

Draw this density function!

Combining likelihood × prior give $p(\theta|y,n)$, which turns out to be in the form of a beta density. See "**Intro. to Bayesian Inference.pdf**" notes.

- See **binomialexamplech2hahn.R**
  - Follows Hahn (2014) examples.

See example on p. 34, $n = 5, y = 4,$ unnormalized posterior is then likelihood × prior,

$$p \;\; \theta|y,n \;\; = c5\theta^4 \;\; 1-\theta \;\; ^1 \times 1$$

We can calculate mean, mode, SD, etc.

(i) analytically (since someone else has done all the math for us), but we can also,

(ii) let the computer figure it out for us (i.e. **evaluate numerically**).

Draw values of the unknown quantity, θ, directly from the posterior beta distribution.

Formulas and further examples are in: **Introduction to Bayesian inference.pdf**
For examples based on baseball hitting average, see: **BeroulliAtBat.R** and data set Blackmonv2.csv

6. A Brief History of Bayesian Inference

Era 1: The Analytical Era. Up until the 1950s, (almost) everything had to be done analytically, so we could only deal with analytically tractable models and distributions. This was true for all approaches to statistical inference of course.

We always start from the same place,

$$p\ \ \theta|y\ \ \propto p(y|\theta)p(\theta),$$

or,

$$p\ \ \theta|y\ \ = cp(y|\theta)p(\theta),$$

where,

$$c = 1/\int p\ \ y|\theta\ \ p\ \ \theta\ \ d\theta.$$

Find this and evaluate the integral to get c. We then evaluate different moments, etc., such as

The mean, $\mu = \int \theta p\ \ \theta|y\ \ d\theta,$

The variance, $\sigma^2 = \int\ \ \theta - \mu\ \ ^2 p\ \ \theta|y\ \ d\theta,$

etc.

A lot of integration is required, but fortunately, for most tractable models and distributions, this has already been done for us.

For example, with a binomial likelihood and beta prior (which includes the uniform and Jeffreys priors as special cases), all the results are given in many books, including Hahn (2014). “It can be shown that…”, for a **Beta prior** with parameters *a* and *b* (see also Bolstad, p.144, Fig 8.1), we get the posterior,

$$p\ \ \theta|y,a,b\ \ = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}$$

which is also a Beta distribution. For this distribution, doing the above integration,

mean $\theta\ \ = \frac{y+a}{n-y+b}$, mode $\theta\ \ = \frac{y+a-1}{n+a+b-2}\ \ \ a,b > 1\ \ $, var $\theta\ \ = \frac{(y+a)(n-y+b)}{n+a+b\ \ ^2(n+a+b+1)}$.

Choosing values for the prior parameters a and b, and given data, y and n, we can calculate these values, calculate interval estimates, etc.

Another example is the Normal likelihood and Normal-Gamma prior. That is, if we have a Normal likelihood function and we specify a normal prior for the mean, and an inverted gamma prior for the variance (i.e. a gamma prior for the precision = inverse of the variance), we get a joint posterior distribution for the mean and precision that is Normal-Gamma. The marginal posterior for the mean is then a Student-t distribution, and the marginal posterior for the precision is a Gamma (or inverted Gamma for the variance). Formulas are given in Hahn, p.62-62 (section 3.11.3).

In practice then, we use the analytical results in the same way we did using the frequentist approach. We look up the formulas and calculate the values given the data (or get the computer to do the calculating!).

Era 2: Limited numerical computation. From the 1960s to 1980s - we can extend our method to some analytically less tractable cases as follows:

1. Find the normalizing constant, c, in our likelihood $\times$ prior formula using numerical integration.
2. Evaluate $p(\theta|y)$ over many points covering the support (i.e. domain where it is not very close to zero – vanishingly small – plot it to find this!) to get: $\theta_i, p\ \ \theta_i|y\ \ , i = 1{,}2, \ldots , M, M$ large.
3. Approximate moments of the distribution using the sample, i.e.
   mean $\theta\ \ \approx \sum_{i=1}^{M} \theta_i p\ \ \theta_i|y\ \ $, etc.

See **binomialposteriors.R** for some examples

The problem with the above approach is that it suffers from "the curse of dimensionality". The computational time increases dramatically as you have more parameters to evaluate, so it is only feasible for models with a few parameters at most.

Era 3: Monte Carlo Integration, 1980s. We can now use MC integration methods, such as accept-reject methods, importance sampling (including antithetic acceleration), etc. For examples, see Hahn, chapter 2:

Accept-reject methods – analogous to throwing darts at the figure and counting how many points are below the function relative to all points. That is, draw a box around the function, then randomly draw points in the box. Evaluate which points are above and below the function. The ratio of point below (accepted) to total points (accepted + rejected) gives the Area under the curve.

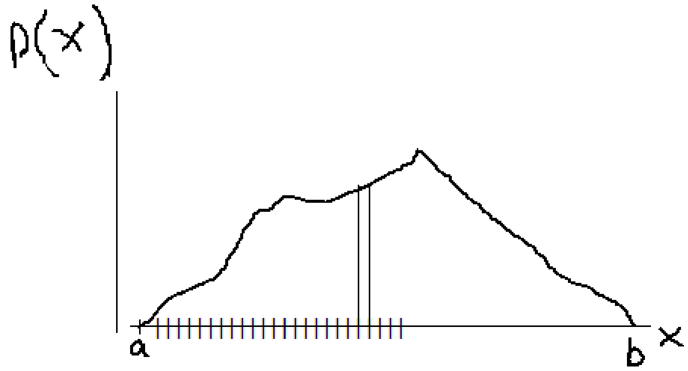
We can make this method more efficient by covering the curve with a smaller area (lowering the rejection rate), and a variety of other techniques.

[Importance sampling is a variation on this theme that involves choosing a distribution (since we are dealing with evaluating areas under probability distributions) that matches the distribution we are integrating fairly closely, and drawing from that rather than from a uniform, so we have a better acceptance rate and we draw a sample that more closely matches one from the distribution we want (the target distribution).]

Alternatively, we can use the following approach.

Consider Hahn, p. 35, formula (3.2) for example.

To evaluate the area under a function (the integral) in the interval a to b:
Partition the horizontal axis between a and b (the domain) into a large number, S, of small intervals



Each interval is $\frac{b-a}{S}$. Then evaluate each rectangle = height of function at midpoint of interval by the interval,
Areas of each rectangle at point $x = x_i$: $p\ x_i\ \ \times \frac{b-a}{S}$

Total area under the curve is then approximated by,

$$Area = \sum_{i=1}^{S} p\ x_i\ \ \times \frac{b-a}{S}$$

$$= \frac{b-a}{S} \sum_{i=1}^{S} p\ x_i \qquad (3.2)$$

This is similar to (3.2), but we have used equally spaced values in the domain $a \leq x \leq b$ to partition the horizontal axis. This is the approach used in Era 2.

Instead, we can draw a set of random points in that interval and partition based on those random points. Each small interval will not be exactly equal to $\frac{b-a}{S}$, but with enough point (i.e. a fine enough partition), they will be close, and the sum of the intervals will $= \sum \frac{b-a}{S}$. We can then still use the formula in (3.2) and we will, for a large enough number of draws of x_i, get a close approximation to the Area.

Era 4: MCMC methods, 1990-present. MCMC method were actually invented in the 1950s, but we weren't smart enough to understand what had already been invented until the 1990s! Metropolis, et al. (1953) invented the Metropolis algorithm. This was extended by Hastings (1970), giving the Metropolis-Hastings algorithm. This was then finally rediscovered and introduced to the statistics literature by Gelfand and Smith (1990).

Simulating draws from the posterior distribution is the simplest version of this: the first MC in MCMC without the second, i.e. Monte Carlo methods, but not Markov Chain. For examples of drawing a simulated sample from a binomial*beta posterior, see **binomialbetaexample.R**

Gibbs sampling is a special case of the Metropolis and MH algorithms where the acceptace rate is 100% because we can draw directly from a known distribution.

The Metropolis algorithm is still the main approach when we cannot draw directly from a known distribution. It is an accept-reject method as above, just a better version. See Hahn (2014), chapter 4 for an exposition of the Gibbs, Metropolis and MH algorithms.

Main condition for convergence of MCMC chain, positivity or “Harris positive recurrence”, can be violated if improper priors are used. Otherwise, the convergence is remarkably general.

7. Hypothesis testing

Separate belief and desire. We may have a strong emotional reaction to that statement, but that has little to no effect on the outcome unless a lot of other people think the same way as us (and I hope they do!).

Frequentist and Bayesian inference are most clearly differentiated by their approaches to precise null hypothesis testing. Even with very large samples, the frequentist and Bayesian conclusions from a point null test can be contradictory. It is possible to get a small frequentist *p*-value, strongly rejecting *H₀* but a large posterior odds or Bayes factor in favor of *H₀*.

Consider the following illustrative example (given by Stone, 1997). Suppose θ is the proportion of a specific type of particle counted in an experiment. The theory under consideration predicts that $\theta = 0.2$ exactly, so the null hypothesis is well defined, H_0: $\theta = 0.2,$ and there is no specific alternative, $H_1 : \theta \neq 0.2.$ The experimental results yield: *n* = 527,135 total particles, and *s* = 106,298 of the specific type observed. What is the evidence against H_0?

Employing frequentist methods gives a maximum likelihood estimate, $\theta = 0.201652$ with standard error = 0.0005526, This results in a *p*-value = 0.0028, indicating strong evidence against the null. Instead of using the *p*-value, the (Bayesian) physicist uses a uniform prior: $p\ \theta\ = 1, 0 < \theta < 1,$ and computes the Bayes factor, *B* = 8.27, indicating evidence in favor of the null. The Bayesian posterior distribution however, is *not* in conflict with the *p*-value, since the posterior probability given the data D, $P\ \theta > 0.2|D\ = \Phi\ 2.99\ = 1 - p$-value/2, where Φ is the standard normal cumulative distribution function. So any Bayesian using a uniform prior must have a strong posterior belief that the true value of θ is larger than 0.2. A 0.99 equal-tailed Bayesian probability interval for $\theta =$ (0.20023, 0.20308), is identical to a 99% frequentist confidence interval and excludes 0.2.

Why are Bayesian inference (posterior probabilities) and hypothesis testing in conflict? Methods of obtaining scientifically objective Bayesian posterior distributions for inference are widely accepted (usually involving noninformative or ‘reference’ priors). The real problem appears to be the hypothesis testing framework when used to test a precise null hypothesis.

See notes on Bayesian hypothesis testing.

8. MCMC algorithms in practice

Econ-8013-MMC-2.pdf for notes on Metropolis and Gibbs algorithms, etc.

See **gibbsgraph.R** (with out$betadraw from **gibbssegch3.R**) for example of how chain moves around the joint distribution one step at a time.

Examples from Hahn (2014), chapter 4

See **Bayesregexample.R** for Bayesian MCMC regression with odds output

See **postoddsmc.R** for hypothesis testing directly from MCMC sample

Appendix: R code

postoddsmc.R evaluates posterior odds using an MC sample.

See mcvanalyticalodds.R for calculation of analytical odds using t or Normal (or any density that R evaluates, Gamma, Beta, etc.):

```
# evaluate odds for Normal comparing 0 and max
max <- 1.0  # put value of max = mean here
null <- 0.0
numN <- dnorm(max,mean=max,sd=1)  # set mean and sd
denN <- dnorm(null, mean=max,sd=1)

poddsN <- numN/den

# Same for a t density with 8 df
numt <- dt(max,df=8)
dent <- dt(null,df=8)

poddst <- numt/dent
```

```
# binomialposteriors.R
#
# Evaluates and plots posteriors for binomial experiment with
# three different priors: normal, beta and uniform

# set vectors to store values
R = 1001
th = rep(0,R)
normbin = rep(0,R)
betaprior = rep(0,R)
betapost = rep(0,R)
unifpost = rep(0,R)
normprior = rep(0,R)
truncunif = rep(0,R)

# parameters for beta prior
a = 2
b = 2

# n = number of obs, y = number of 'successes' in n trials
n = 40
y = 8

# prior parameters for normal prior
m0 = 0.9
s20 = 10

# univariate numerical integration to determine normalizing constant
for normbin spec
```

```
int1 <- function(z) (exp(-((z-m0)^2)/(2*s20)))*(z^y)*(1-z)^(n-y)
cth = integrate(int1, lower = 0, upper = 1.0)
# c is the normalizing constant for use with normbin
c = cth$value
c



# uniform prior with interval a to b
# suppose a and b are:
a <- 0.10
b <- 0.4

int1 <- function(z) (z^y)*(1-z)^(n-y)
cth = integrate(int1, lower = a, upper = b)
# c is the normalizing constant for use with normbin
c2 = cth$value
c2

int2 <- function(z) ((z^(y+1))*(1-z)^(n-y))/c2
mean <- integrate(int2, lower = a, upper = b)
mean

# c is the normalizing constant for use with normbin
c2 = cth$value
c2


# Evaluate each posterior from 0 to 1, and the beta and normal priors
# normbin = posterior with normal prior
# betapost = posterior with beta prior
# unifpost = posterior with uniform prior
# betaprior = beta prior
# normprior = normal prior
for (i in 1:R)
{
th[i] = a + i*0.001*(b-a) - 0.001*(b-a)
truncunif[i] <- (1/c2)*(th[i]^y)*(1-th[i])^(n-y)
}

plot(th,truncunif,type='l')

bdraws <- rbeta(10000,y+1,n-y+1)

bd2 <- subset(bdraws,bdraws >= a & bdraws <= b)
hist(bd2)

mean(bd2)
sd(bd2)

a <- 2
b <- 4
```

```
for (i in 1:R)
{
th[i] = i*0.001 - 0.001

normbin[i] = (1/c)*(exp(-((th[i]-m0)^2)/(2*s20)))*(th[i]^y)*(1-
th[i])^(n-y)
betaprior[i] = (gamma(a+b)/(gamma(a)*gamma(b)))*(th[i]^(a-1))*(1-
th[i])^(b-1)
betapost[i] = (gamma(n+a+b)/(gamma(y+a)*gamma(n-y+b)))*(th[i]^(y+a-
1))*(1-th[i])^(n-y+b-1)
unifpost[i] = (gamma(n+2)/(gamma(y+1)*gamma(n-y+1)))*(th[i]^y)*(1-
th[i])^(n-y)
normprior[i] = (1/sqrt(2*3.142*s20))*exp(-((th[i]-m0)^2)/(2*s20))

}

# plot of everything on one graph
yy = cbind(normbin,betaprior,betapost,unifpost)
matplot(th,yy,type='l', col= 1:6,lwd=2)
legend("topright",legend=c("Norm-post","Beta-prior", "Beta-post",
"Unif-post"),col=c(1:4),lty=c(1:4),lwd=2,bty="n",cex=1.1)
# NB: use lty=c(1,1) in above (matplot and legend) for all solid lines


# plot of everything on one graph
yy = cbind(betapost,unifpost,normprior,truncunif)
matplot(th,yy,type='l', col= 1:6,lwd=2)
legend("topright",legend=c("Beta-post", "Unif-post", "Norm-
prior","trunc. unif"),col=c(1:4),lty=c(1:4),lwd=2,bty="n",cex=1.1)
# NB: use lty=c(1,1) in above (matplot and legend) for all solid lines



par(mfrow=c(1,2))

# plot of priors
# uniform prior = 1 for all 0 < theta < 1
unif = rep(1,R)
yy = cbind(unif,betaprior,normprior)
matplot(th,yy,type='l', col= 1:3,lwd=2)
legend("topright",legend=c("Uniform prior","Beta prior", "Normal
prior"),col=c(1:3),lty=c(1:3),lwd=2,bty="n",cex=1.1)
# NB: use lty=c(1,1) in above (matplot and legend) for all solid lines


# plot of posteriors
yy = cbind(unifpost,betapost,normbin)
matplot(th,yy,type='l', col= 1:3,lwd=2)
legend("topright",legend=c("Uniform post","Beta post", "Normal
post"),col=c(1:3),lty=c(1:3),lwd=2,bty="n",cex=1.1)
# NB: use lty=c(1,1) in above (matplot and legend) for all solid lines
############################################
```

```
# binomialbetaexample.R

# binomial likelihood * beta prior

# prior parameters to get a uniform
a <- 1
b <- 1

# data
n <- 100
y <- 4

# domain:
theta <- 0:1000/1000

# "DIY" evaluating the posterior density function
lnormc <- lgamma(a+b+n) - lgamma(a+y) - lgamma(n-y+b)
lkern <- (a+y-1)*log(theta) + (n-y+b-1)*log(1-theta)
postth <- exp(lnormc+lkern)

plot(theta,postth,type='l',col=4)

# show you get same from dbeta function in R
post2 <- dbeta(theta,shape1=(a+y),shape2=(n-y+b))
lines(theta,post2,col=3,lwd=2)

meanpu <- (a+y)/(a+n+b)


# credible interval
thdraws1 <- rbeta(1000000,shape1=(a+y),shape2=(n-y+b))
mean(thdraws1)
sd(thdraws1)
interval1  <-  quantile(thdraws1,probs  =  c(0.005,  0.025,  0.5,  0.975,
0.995))
interval1



# informative beta prior
a <- 4
b <- 36
prior2 <- dbeta(theta,shape1=a,shape2=b)
post2 <- dbeta(theta,shape1=(a+y),shape2=(n-y+b))
lines(theta,prior2,col="pink",lwd=2)
lines(theta,post2,col="magenta",lwd=2)

meanpb <- (a+y)/(a+n+b)

# credible interval
thdraws2 <- rbeta(1000000,shape1=(a+y),shape2=(n-y+b))
```

```
mean(thdraws2)
sd(thdraws2)
interval2 <- quantile(thdraws2,probs = c(0.005, 0.025, 0.5, 0.975,
0.995))
interval2


# less informative beta prior (but more informative than uniform)
a <- 1
b <- 9
prior3 <- dbeta(theta,shape1=a,shape2=b)
post3 <- dbeta(theta,shape1=(a+y),shape2=(n-y+b))
lines(theta,prior3,col="orange1",lwd=2)
lines(theta,post3,col="orange2",lwd=2)

meanp3 <- (a+y)/(a+n+b)



# credible interval

# credible interval
thdraws3 <- rbeta(1000000,shape1=(a+y),shape2=(n-y+b))
mean(thdraws3)
sd(thdraws3)
interval3 <- quantile(thdraws3,probs = c(0.005, 0.025, 0.5, 0.975,
0.995))
interval3

# results: means and probability or "credible" or HPD intervals
meanpu; meanpb; meanp3
interval1
interval2
interval3



# Now suppose data are (more like the baseball example)
n <- 20
y <- 4

# prior mean of about 2.50, max of around 0.40, choose a and b for
this
# informative beta prior

########## try different prior parameter values, e.g. 4, 20,

a <- 4
b <- 20

# Analytical solution
priorhits <- dbeta(theta,shape1=a,shape2=b)
```

```
plot(theta,priorhits,col="green1",lwd=2,type='l')


# posterior is a beta with parameters as below
posthit1 <- dbeta(theta,shape1=(a+y),shape2=(n-y+b))
plot(theta,posthit1,col="green3",lwd=2,type='l')
lines(theta,priorhits,col="green1",lwd=2)

# numerical simulation solution
priorhitss <- rbeta(1000000,shape1=a,shape2=b)



# posterior is a beta with parameters as below
posthit1 <- dbeta(theta,shape1=(a+y),shape2=(n-y+b))
plot(theta,posthit1,col="green3",lwd=2,type='l')
lines(theta,priorhits,col="green1",lwd=2)
lines(density(priorhitss),col="blue",lwd=2,type='l',lty=2)

# simulation solution
posthits <- rbeta(1000000,shape1=(a+y),shape2=(n-y+b))
lines(density(posthits),col="blue",lwd=2,type='l',lty=2)





# Uniform prior between a >0 and b < 1
a <- 0.10
b <- 0.41

# can use numerical integration, as in binomialposteriors.R
# Alternatively, using simulation:
# Just need to truncate posterior from U[0,1]

# draws from posterior for uniform[0,1]

# credible interval
au <- 1
bu <- 1
thdrawsu <- rbeta(1000000,shape1=(au+y),shape2=(n-y+bu))
mean(thdrawsu)
sd(thdrawsu)
intervalu <- quantile(thdrawsu,probs = c(0.005, 0.025, 0.5, 0.975,
0.995))
intervalu

# now truncate to between a and b
thdrawsab <- subset(thdrawsu,thdrawsu >= a & thdrawsu <= b)
hist(thdrawsab,freq=F)
lines(density(thdrawsab),col=5)
mean(thdrawsab)
```

```
sd(thdrawsab)
intervalab <- quantile(thdrawsab,probs = c(0.005, 0.025, 0.5, 0.975,
0.995))
intervalab

# plot all posteriors together
plot(density(posthits),col="blue",lwd=2,type='l') # using informative
beta prior
lines(density(priorhitss),col="blue",lwd=2,type='l',lty=2)#
informative beta prior
lines(density(thdrawsu),col="green",lwd=2,type='l')  #  using  uniform
[0,1] prior
lines(density(thdrawsab),col="red",lwd=2,type='l')  #  using  uniform
[a,b] prior
```

```
# postoddsmc.R
postoddsmc <- function(mcsample,null=0.0){

# use: set postoddsmc(mcsample,null=nullvalue)
#      mcsample = sample from the posterior
#      nullvalue = value under the null hypothesis

# purpose: evaluate odds vs. a point null hypothesis
#    given sample from posterior (from MC or MCMC)

# output = odds ratio vs. null of zero

# Jeff Mills and Hamed Namavari, 2016

testPoint = null   ## null hypothesis cut-off point

DenB1 <- density(mcsample,n=1024) ## posterior density
index <- which(DenB1$x-testPoint==min(abs(DenB1$x-testPoint)))
if (length(index) == 0) {
  index <- which(-DenB1$x+testPoint==min(abs(DenB1$x-testPoint)))
}

# an alternative to above?
#  if  (testPoint<min(DenB1$x)  |  testPoint>max(DenB1$x))  num  =
min(DenB1$y)

denom <-  max(DenB1$y)
num <- DenB1$y[index]
oddsratio <- denom/num

##  NEXT LINE TO HAVE oddsratio > 10000 reported
if(oddsratio > 10000) {oddsratio ="> 10000"}

list(oddsratio = oddsratio)
}
```