Boxian Bayesianism: Penalized Estimation as a Universal Tool for Data Analysis

July 2014

Sander Greenland

Department of Epidemiology and
Department of Statistics, UCLA

Please report errors and send comments to Sander Greenland at lesdomes@ucla.edu

Primary readings for this workshop:

Greenland S. Bayesian Perspectives for Epidemiologic Research. *International Journal of Epidemiology*, in 3 parts:

I. Foundations and basic methods. 2006; 35: 765-778 – reprinted as Chapter 18 of Modern Epidemiology, 3rd ed., 2008 (ME3).

II. Regression analysis. 2007; 36: 195-202.

III. Bias analysis via missing-data methods. 2009; 38: 1662-1673; corrigendum 2010; 39: 1116 – see also Ch. 19 of ME3.

Some further technical readings:

Greenland, S. (2007). Prior data for non-normal priors. *Statistics in Medicine*, 26: 3578-3590.

Greenland, S. (2008). Variable selection and shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, 167: 523-529; erratum 1142.

advanced theory for Bayes III:

Greenland, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Science*, 24: 195-210.

Some further general readings:

Greenland, S. (2009). Dealing with uncertainty about investigator bias: disclosure is informative. *Journal of Epidemiology and Community Health*, 63: 593-598.

Greenland, S. (2012). Transparency and disclosure, neutrality and balance: shared values or just shared words? *Journal of Epidemiology and Community Health*, 66: 967-970.

Greenland, S. (2012). Causal inference as a prediction problem: Assumptions, identification, and evidence synthesis. Ch. 5 in: Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.). *Causal Inference: Statistical Perspectives and Applications*. New York: Wiley, 43-58.

Preliminary Disclaimer

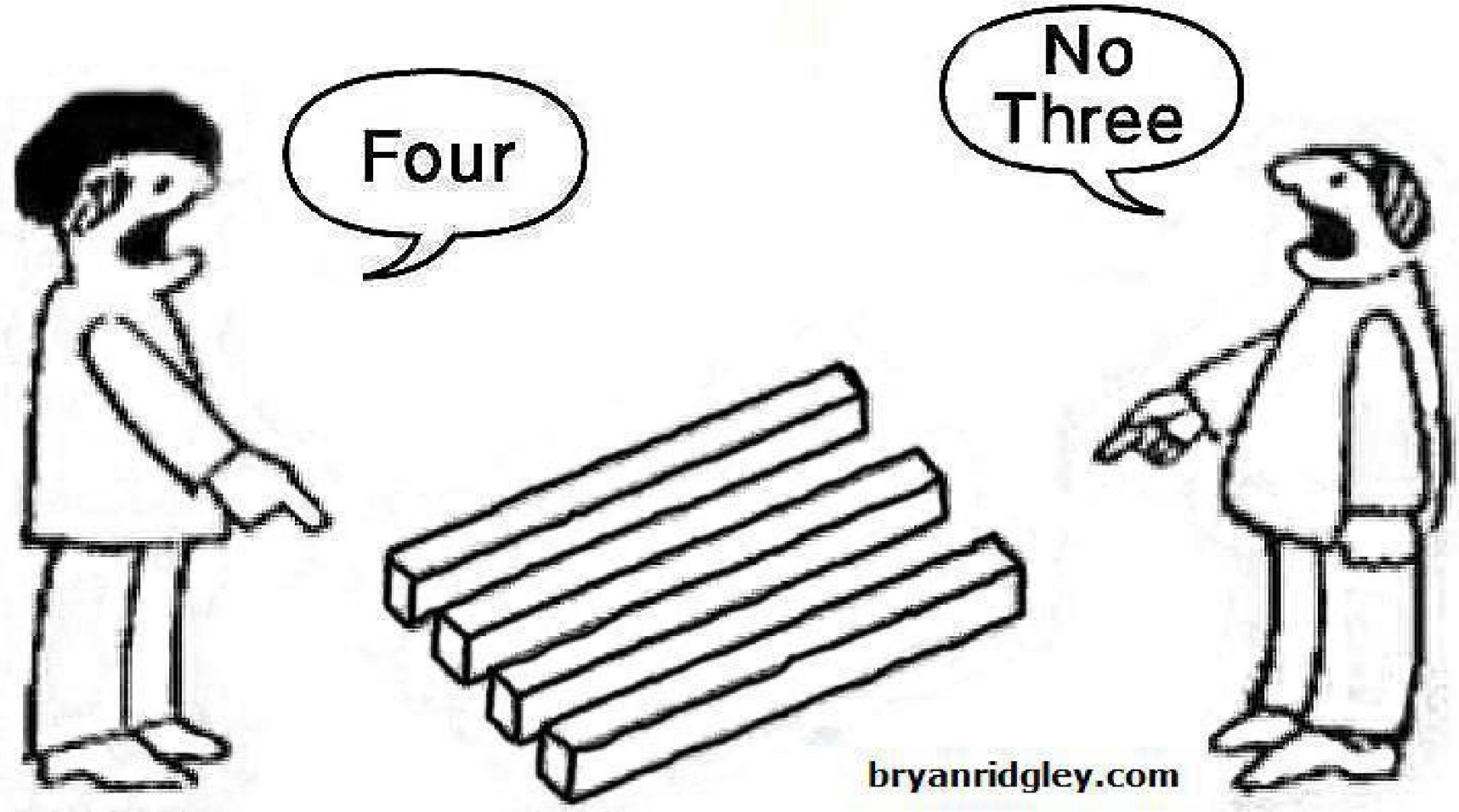
Use of frequentist or Bayesian methods does **not** require adopting a frequentist or Bayesian philosophy, any more than using a nail or a screw requires adopting a spiky or screwy philosophy.

- There are many Bayesian philosophies. I am **not** a Bayesian. I am a **pragmatist**.
- Subjective-Bayesian interpretations and methods can be **useful aids to reasoning**.

Specifically, Bayesian methods should be part of core statistics training, because they can

- facilitate use of models bigger than conventional frequentist methods allow,
- provide an alternative perspective on the inferential meaning of information.
- display the subjective elements and values that pervade analyses and reports.
- **force more thinking about the connection between statistics and reality.**

Reality can be so complex that equally valid observations from differing perspectives can appear to be contradictory.



The key divide of concern in HSS is **not** Bayesian vs. frequentist, but whether the models assumed for the data generating process (**DGP**) yield spurious identification of a target quantity or a decision from data.

- The model is the set of all assumptions being used for the math (deduction). Thus all statistics are model-based – even so-called nonparametric statistics (which assume **randomization**, an incredibly strong multivariate independence assumption).

The frequency (sampling or randomization) distributions used in HSS are riddled with so many untestable validity assumptions that they represent only personal expectations about the frequencies. Furthermore, replication is often so limited and subject to bias that one cannot rely on tests of these distributions. They are thus not 'objective' frequencies; they are instead only community conventions, which favor certain values or stakeholders and penalize others.

In OHSS, the assumptions used to derive these conventional distributions are far too shaky to claim as "objective."

- At best, these assumptions only refer to statements that have reached a certain level of ‘general acceptance as true’ (GAT status) in a scientific community that is the source of authority.
- More usually, the assumptions are simply unquestioned conventions that most users either do not see or do not understand.

It is important to question authority whenever authority offers certainties about matters for which there is far from sufficient evidence for that certainty.

- To carry out this advice, we have to come up with defensible measures of uncertainty, and methodology for constructing those uncertainty measures from statements (including data).
- That turns out to be an incredibly difficult task – perhaps too difficult for routine use. But we have to see what would be involved to appreciate how much we fall short in practice.

Scientific foundations for statistical analysis

Scientific analysis contains elements labeled "statistical," which deal with input and organization of data based on data descriptions and the hypothesized mechanisms that generated those data.

- These elements need to be integrated into reasoning laden with contextual meaning, not treated as abstract math or computing.

Subjective elements and values play a decisive role in all statistical analyses

- There is an illusory sense of objectivity induced when there is great overconfidence, as generated by elaborate mathematics, strong sense of authority, or extensive social agreement.
- Feelings of objectivity in turn feed back to create overconfidence. This is amply illustrated in history by scientists and often entire fields assigning near certainty to hypotheses later refuted.
- Classic statistician examples: Fisher on smoking and lung cancer, Jeffreys on continental drift.

Medical examples abound: Hormone replacement therapy as a general anti-aging treatment, low-fat diets for weight loss, and numerous drugs that were aggressively promoted and then discredited (with errors sometimes encouraged by deceptive trial reporting, e.g.,Vioxx).

- There are many parallels in modern statistical practice. Among them, the objectivist/frequentist hegemony produced an epidemic of significance testing that in turn has led to gross distortions of data reporting (publication bias, rampant misinterpretation of ambiguous results as null).

The ubiquity of error at all levels

Error (including systematic error) is inevitable – not only data and inference error but also conceptual error, extending to the highest authorities.

- A key to minimizing average error *cost* is uncertainty assessment, to encourage well-balanced hedging (proper allowance for alternatives to 'accepted' hypotheses).
- A key to minimizing conceptual error is to vary perspectives by applying conceptually different approaches to assessments.

"Objectivity" in statistics usually means nothing more than the following act of faith:

- The data really were derived from a study conducted in a manner that forced the assumed data-probability ("statistical") distribution Pr(data|hypothesis) to hold for all hypotheses of interest.

This deduction is *always* based on assuming interventions (selection or treatments) were applied to population units following a known and perfect randomized design protocol.

In the real world of health sciences, however such faith (absolute certainty) in researchers and publications is unwarranted, for reasons such as

- Cheating, fraud (see WSJ ’09 on 21 faked studies)
- Procedural error and bias - often undocumented, but hard to deal with even if recognized.
- More generally and perhaps more commonly, sociopolitical agendas and pure overconfidence driving interpretation and presentation of results…

  Example: Selective citation in Young and Karr in *Significance* Dec. 2011, and Wasserman blog 2012 uncritically accepting their claims.

- "Bayesian" methods can aid uncertainty assessment in the face of these problems.

However, we need to recognize and adopt other names for those methods because:

- the "Bayesian" label is too vague and too mired in useless philosophical controversy.
- Bayes' theorem is neither necessary nor the most clear representation for the methods.
- "Bayesians" are too diverse (Good, *Am Stat* 1971, described 46,656 types of Bayesian).

The "Bayes" I prefer could be called ***Boxian***:

- A model is a set of assumptions/constraints, which are called priors when their justification appeals to external information.
- Nonbayesians call soft (fuzzy) constraints "penalty functions" or "random-coefficient distributions" instead of "prior distributions"
- Priors are predictions that can be tested.
- Contextual interpretations of models can provide external diagnostics of constraints.

What are prior probabilities?

Probabilistic judgments about parameter magnitudes, derived **externally** (from anything other than the analysis data).

The usual priors in health literature are either:

- Reference or 'noninformative' priors, or
- Direct opinions expressed by a highly biased and tiny sample of 'experts.'
- **Always present!** But usually hidden, especially when they represent prejudices.

Problems with conventional priors:

- Reference priors give back numeric results close to P-values and confidence intervals, and so do not get us beyond the limits of conventional frequentist methods.
- Direct-opinion priors risk injecting severe personal biases into the analyses, and so are the source (and worthy) of the scorn heaped on Bayesian methods by frequentists.

Each approach tries to avoid the labor of creating scientific priors, which is daunting.

Value bias afflicts all decision-theoretic inference, most often as **nullism** (subtalk: 22-34)

Call a methodology *value-biased* when it incorporates assumptions about error costs that are not universally accepted (especially when hidden).

- Example: The consistent use of the null as the test hypothesis, to the point of failing to distinguish the two concepts. I call this an example of *nullism*, value bias toward the null.
- May be based on imagined costs of rejecting the null (as in product surveillance), or metaphysical beliefs (parsimony, religious, ideological).

Nullism has a long and glorious history among physics idolaters as "skepticism" (certainty):

- "Heavier than air flying machines are impossible" – Lord Kelvin, 1895
- "Continental drift is out of the question" because no [known] mechanism is strong enough – Sir Harold Jeffreys, geophysicist (and originator of spiked priors)
- "Physics shows that cell phones cannot cause cancer" because microwaves are not ionizing – Michael Shermer, Scientific American Oct. 2010

In health and social sciences there is rarely any positive scientific evidence that the null is *exactly* true, and few specialties (e.g., genomics) have credible mechanistic arguments for claiming departures from the null are probably negligible.

- DR Cox (2001) opined that in many studies "there may be no reason for expecting the effect to be null. The issue tends more to be whether the direction of an effect has been reasonably firmly established and whether the magnitude of any effect is such as to make it of…importance."

This view directly indicts a good portion of the Bayesian literature, where null spikes are used to represent the belief that a parameter "differs negligibly" from the null. In many settings, even a tightly concentrated probability near the null has no basis in genuine evidence.

Still, many scientists and statisticians exhibit quite a bit of prejudice in favor of the null based on faith in oversimplified physical models of biology. Shermer (Sci Am 2010) is a vivid example, claiming a link of cell phones to cancer is "physically impossible" because microwaves are nonionizing.

Null prejudice also arises more subtly from confusion of decision rules with inference rules, and from adoption of simplicity or parsimony as a metaphysical principle rather than a heuristic.

- In psychology, many have argued that the null hypothesis is almost never exactly true. Similar arguments apply in medical research, where medicines are pursued because they interact with systems involved in the disease process.
- We may be highly certain that any effect present is small enough so that the cost of ignoring it is acceptable; but this is a value-laden judgment.

Yet nullism has also been widely taught as an integral part of Neyman-Pearson testing – even though it is not: "**According to circumstances and according to the subjective attitudes of the research worker**, one … error may appear more important to avoid than the other. … **the error which is the more important to avoid will be called 'error of the first kind'**, …the [hypothesis] the unjust rejection of which constitutes the error of the first kind, will be called 'the hypothesis tested'." (Neyman, Synthese, 1977, p. 104; emphases added)

That is, H may be the non-null hypothesis!

Neyman continues: “From the point of view of the manufacturer [of a chemical A] the error in asserting the carcinogenicity of A is (or may be) more important to avoid than the error in asserting that A is harmless. Thus, **for the manufacturers of A, the 'hypothesis tested' may well be: 'A is *not* carcinogenic'. On the other hand, for the prospective user of chemical A the hypothesis tested will be unambiguously: 'A *is* carcinogenic'**. In fact, this user is likely to hope that the probability of error in rejecting this hypothesis be reduced to a very small value!”

Here, nullism is bias against the consumer, for the manufacturer. Multiple testing expands this bias:

- It takes the joint null as the test hypothesis (the one more important to not reject incorrectly) using the standard maximum tolerable Type-I error rate of 0.05 for the entire ensemble null.
- These tests assume ever higher relative costs of Type-I over Type II errors as the number K of hypotheses increases (the expected cost of any false positive is more than for any false negative).
- This is likely true for drug companies monitoring adverse effects, but not for patients.

Bias toward null inferences is sometimes defended with reference to the class of heuristic maxims known collectively as parsimony principles (such as "Ockham's Razor"). A parsimony principle relevant here is "model simplicity has benefits."

- Simple models can provide useful data summaries and can outperform some complex models even if the reality is complex to the point of saturation.
- But: simple models have had a dubious record in observational studies, because there are too many potentially important factors for a simple model to capture for inference (i.e., going beyond the data).

- Sometimes a simplified model is close enough to reality so that the bias incurred from its use is outweighed by the noise reduction it achieves, which is a basis for penalization and other shrinkage (regress-toward-the-mean) methods.
- Simplicity can also have practical benefits: Use of simpler models may save time, enable wider adoption, and reduce procedural errors.
- These benefits neither imply nor require that the underlying study phenomena are themselves simple, since they accrue even when modeling transcomputable phenomena like living organisms.

Especially in health risk assessment, the universal null is sometimes defended by appealing to dubious contextual or mechanistic arguments.

- One such argument claims humans have been naturally selected to resist carcinogens.

Even if true in some sense, it does not apply to the chief controversies of the field.

In almost all environmental, occupational, and medical examples (fields where litigation abounds), at least one and usually both the following facts invalidate the argument:

- The entire class to which the exposure belongs (e.g., plasticizers, pharmaceuticals) was introduced only recently. Thus it is a leap of faith to assume our bodies evolved defenses that disarm these exposures completely.
- The cancer is extremely rare until well past reproductive age, so its causes exert negligible selective pressure. Even the cancers most common before and during reproductive ages are rare at parenting ages. With this rarity, there can be only slight selective pressure from the effects at issue (RR<4).

Thus the argument is only rhetorical seduction by mechanism: We have one unsupported hypothesis (selective pressure favors the null at issue) touted *as if* it strongly supports the null. Conclusions:

- Arguments that null hypotheses are generally true are logically fallacious and empirically void.
- While predictive and modeling tasks might adopt parsimony as a heuristic, such *use* of null hypotheses does not justify *belief* in them.
- Like all hypotheses, null hypotheses need to be contrasted against evidence, and treated agnostically when (as usual) evidence is weak.

Bayesian misinterpretations play a major role in common statistical inference

- Typical human intuitions interpret probabilities *involving* hypotheses as probabilities *of* hypotheses. When the probability is instead *conditional on* the hypothesis, this interpretation is "the prosecutor's fallacy", mistakenly equating

Pr(hypothesis|evidence) = Pr(evidence|hypothesis)

where the hypothesis is of guilt (although the same fallacy can involve a hypothesis of innocence)

P is usually misinterpreted as the probability that "chance alone" produced the observed association:

- Oleckno (*Essential Epidemiology*, 2002, p. 182) says the P-value "measures the **probability that the difference is due to sampling error**;"
- Article listing Kooperberg, Lumley, Psaty (AJE, 2007) says they "conducted a permutation test to estimate the **probability of a chance finding**"
- Harris and Taylor (*Medical Statistics Made Easy*, 2nd ed, 2008, p. 24-25) state that "the P value gives the **probability of any observed differences having happened by chance**."

- Unfortunately for these misinterpretations, the hypothesis (and hence probability) that sampling error or random error or **chance alone** produced a difference or an association or a "finding" is **logically identical** to the hypothesis that **the null hypothesis is true *and* there is no bias**.
- In practice, the null P-value is rarely even close to the Bayesian posterior probability of the latter hypothesis. Thus it seems valuable to describe the posterior probabilities that P-values actually do approximate or bound.

- A P-value for a hypothesis H using a statistic T is the probability that T would have been at least as big as t, its observed value, given the hypothesis H: $P_H = \Pr(T \geq t|H)$.
- But the P-value P_H is routinely mistaken for the probability of the hypothesis given T, $\Pr(H|T=t)$ ("inversion")
- Bayes theorem shows that P_H is equal to $\Pr(H|T=t)$ only under certain special conditions.
- Without those conditions, P_H can easily be an order of magnitude **smaller** than $\Pr(H|T=t)$ (hence the claim "P-values overstate evidence")

A special case where a P-value becomes a useful posterior probability: Let P_b be the 2-sided P-value for whether a coefficient β equals b, and suppose the prior Pr(β) is diffuse relative to the likelihood. Then

- $P_b/2$ approximates the posterior probability that the estimate of β is on the wrong side of b. In particular, if b is 0,

  $$P_0/2 \approx \Pr(\hat{\beta} \text{ has wrong sign} \mid \text{data and model})$$

- This was a standard interpretation (e.g., Student, 1908) before significance-testing became dominant (largely due to Fisher's hatred of Bayesian stats and insistence on "objectivity")

"Bayesian" is somewhat misleading

As with "Big Bang Theory", the label "Bayesian Statistics" came from a critic, R.A. Fisher (Fienberg, *BA* 2006). Unfortunately, that led teaching to focus on Bayes' theorem for updating hypothesis probabilities.

- Other updating formulas (e.g., data augmentation; information addition) are more transparent and often ease computation as well.

- In the latter theories, basic concepts are laid out on a (−2) log-probability scale and Bayes' theorem is replaced by an additivity property of information measures that resembles:

external ("prior") information

+ internal study information

= total (posterior) information

- This formulation displays the symmetry between prior and study information, which should be reflected in the criteria used to judge the acceptability of each source.

- In the Dec. 2010 issue of the unfortunately titled magazine *Significance*, there's a boxed quote from the brilliant statistician Bradley Efron (inventor of the bootstrap) that is completely wrong:

  "A Bayesian prior is an assumption of an infinite amount of past relevant experience."
- There are several ways to see this statement is pure nonsense in practice, e.g., for normal priors the larger the variance the less the information, with zero information as a greatest lower bound.
- Worse, it is the data model, not the explicit prior, that assumes infinite past relevant experience!

Frequentist methods use no **explicit** prior, and so some claim the methods are "objective" or

"let the data speak for themselves."

This is pure delusion because frequentist methods are filled with implicit priors, and

DATA SAY NOTHING AT ALL!

Data are markings on paper or bits that just sit there.

If you hear the data speaking, seek psychiatric care immediately!

Example: Everyone uses the model

$$g(\mu_{xz}) = \alpha+\beta x+\gamma z \text{ where } \mu_{xz} \equiv E(Y|x,z)$$

without saying it assumes $\delta = 0$ in the model

$$g(\mu_{xz}) = \alpha + \beta x + \gamma z + \delta xz$$

Frequentists call $\delta = 0$ a constraint;

Bayesians could instead say (equivalently) that δ has a prior with $E(\delta) = var(\delta) = 0$.

But when they use the model they both forget that they used a prior with $E(\delta) = var(\delta) = 0$, unless they do "model checking".

$\delta = 0$ means parallel lines ("no interaction") on the math-convenient scale of the link $g(\mu_{xz})$.

What is the justification for this assumption?

Absolutely no contextual justification in HSS!

It need not be very misleading if we restrict our inference to marginal means, but can be quite misleading for conditional means.

By calling the assumption $\text{var}(\delta) = 0$ we can see it assumes infinite information, and that we can use instead $\text{var}(\delta) > 0$ so that we no longer assume with certainty what we do not know to be true.

Why learn Bayesian methods?

The major practical reasons:

- To see what it takes to create credible models and probabilities of hypotheses.
- To fit models bigger than conventional frequentist methods allow.
- To make some allowance for bias or uncertainty sources that frequentist models must assume are absent or identified by the data (and thus analytically controllable).

What Everyone Should Know:

- Identifiability = estimability = ability to estimate the target parameter from the data **and the rest of what is assumed known**.
- Identification is achieved only by models (assumption sets) = **asserted knowledge:**
    - Random sampling assumptions identify population associations
    - Randomization assumptions identify causal effects

What everyone should also know:

In observational studies,

- sampling is **not** random
- exposure assignment is **not** random
- measurement error is **not** random (but produces bias even if it is random).

Yet every method in common use (including "validation-study" methods) uses at least two of these assumptions (within covariate-adjustment levels); most use all three.

Data models are priors!

Conventional models are **priors for the outcomes**, given the parameters, e.g.,

the linear regression model

$$Y = \alpha + x\beta + \mathbf{z\theta} + \varepsilon \ ,$$

is a prior that the outcome is the product of **known** x,z with the unknown β,**θ** plus a constant α, a **random** error ε, and…

nothing more.

These are very strong priors, and…

We know these priors are wrong!

- Often excused by saying "all models are wrong, but some are useful"

Rotten excuse in epi, where we know these models are **very** wrong (yet they are the standard) because:

- Some confounders are not measured,
- Selection depends on X,Y,Z, and more,
- X, Y, Z are mismeasured; even if random, mismeasurements produce biases.

Example of an excusable data model:

Causal: Y_x = potential outcome|treatment x

$$= \alpha + x\beta + \mathbf{z}\boldsymbol{\theta} + \mathbf{u}\boldsymbol{\gamma} + \varepsilon$$

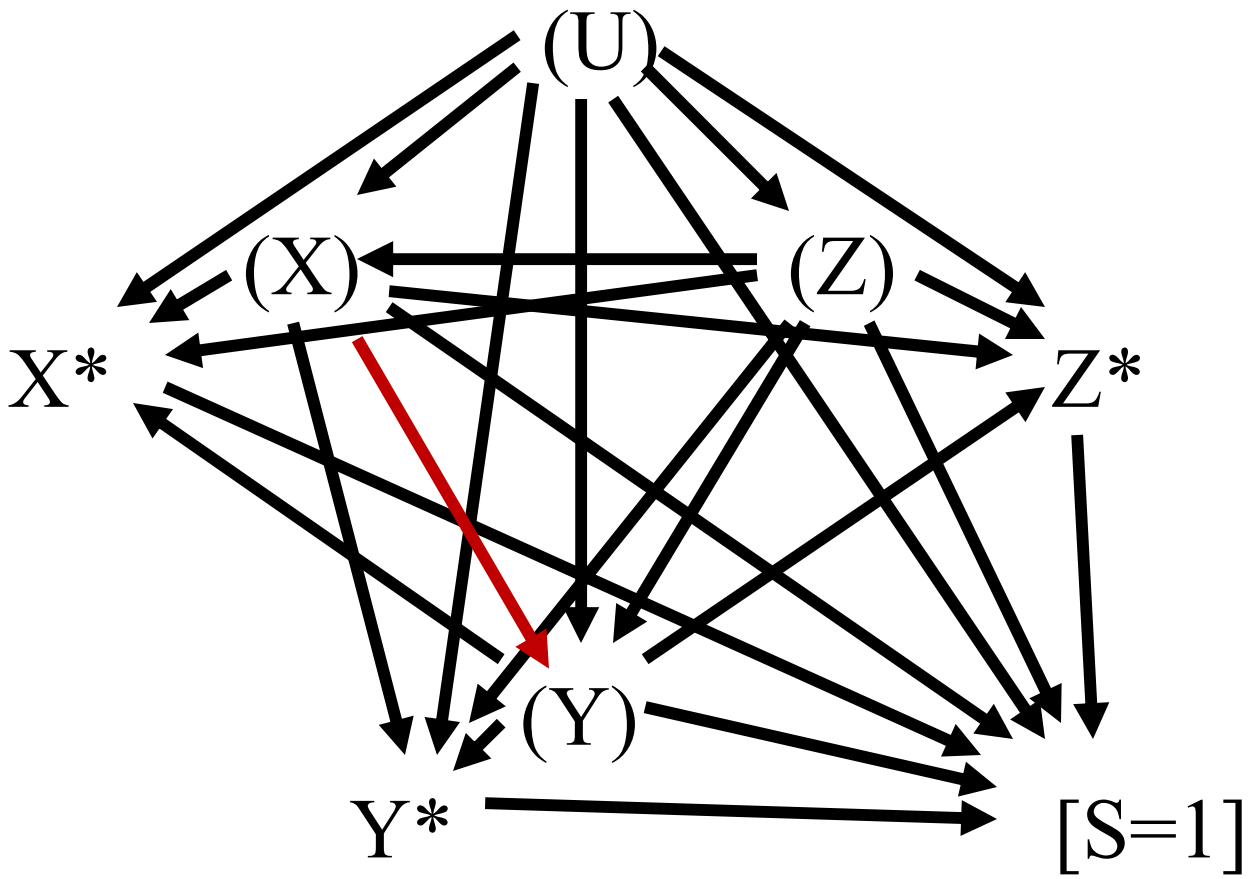
Measurement $(X^*, Y^*, Z^*) = (1, x, y, \mathbf{z})\Lambda + \mathbf{u}\Psi$

Selection rate = $\exp\{(1, \mathbf{u}, x, x^*, y, y^*, \mathbf{z}, \mathbf{z}^*)\boldsymbol{\pi}\}$

where **U**, X, Y, **Z** are missing on everyone and X*, Y*, **Z*** are just measurements.

- This is a **system** of equations with matrices Λ, Ψ and vectors $\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\pi}$ of unknowns.

Just to estimate the main effect of X on Y under the model below we would need to account for X-Y associations arising from **all 25** of its arrows:



Parameter priors:
Probabilities (bets) about parameters

These priors are **always present too!**

Use of the conventional model

$Y = \alpha + x\beta + \mathbf{z\theta} + \varepsilon$ implies 100% certainty that

$\gamma = 0$, $\Psi = \boldsymbol{\pi} = \mathbf{0}$ and Λ = identity, in

$Y_x = \alpha + x\beta + \mathbf{z\theta} + \mathbf{u\gamma} + \varepsilon$

$(X^*, Y^*, Z^*) = (1, x, y, \mathbf{z})\Lambda + u\Psi$

selection rate $= \exp\{(1, \mathbf{u}, x, x^*, y, y^*, \mathbf{z}, \mathbf{z}^*)\boldsymbol{\pi}\}$

- Frequentist: “Conventional statistics display the component of inferential uncertainty due to the play of chance” – see next slide.
- “Objective” Bayesian: “Conventional statistics provide a reference-Bayes analysis under noninformative priors” – ignores that the data model is full of dogmatic priors.
- Descriptive: “Permutation statistics display sensitivity of results to minor data alterations (perturbations)” – They display only sensitivity of *frequentist* statistics.

Claim: “Bias or not, frequentist statistics measure uncertainty due to chance”. How this claim fails:

- We can’t measure how much variation is chance when we don’t know the randomizer (or even if there is one), and don’t know all the mechanisms that lead to study inclusion and events.
- Thus, frequentist inference becomes at best a speculative exercise, producing dubious subjunctives like “if the mechanisms were as simple as this (which no one believes), here is what our inference would be.”

The Probabilistic Imperative

Inferences from conventional models like $Y=\alpha+x\beta+\mathbf{z\theta}+\varepsilon$ are often pure fantasy.

- Priors allow public, explicit summary inferences from sensitivity (bias) models.
- Hence it is imperative that we train ourselves in probabilistic uncertainty assessment (including Bayesian methods) so that we understand the logic of reasoning under uncertainty.

What are scientific priors?

- **Derived** by explicit deductive reasoning from past data – where the derivation goes beyond reported conclusions, to the data and its source, as in a critical research synthesis/meta-analysis.
- **Evaluated** in terms of 'equivalent data' (data augmentation priors) to provide a better sense of their information content, and
- **Tested** against current data.

How to fit a Bayesian model
(partial list, in order of increasing "exactness")

- Monte-Carlo Sensitivity Analysis (MCSA), an incomplete approximation
- Inverse-variance (precision) weighting
- Penalized likelihood (PL), algebraically equivalent to using data-augmentation priors (DAPs).
- Posterior sampling (MCMC/WinBUGS; importance sampling; others).

- Penalized likelihood. Can be fit with ML programs using DAPs; gives exact posterior odds, plus approximate posterior probabilities (same approximation as ML).
- Posterior sampling (MCMC/WinBUGS, importance sampling) – "simulation exact" posterior probabilities, which however include simulation error (error that can be made negligible by running much longer than defaults).

Which fitting method?

- Given the crudeness of meta-analyses and observational epidemiology, simpler methods are often adequate.
- All methods have limits and hazards.
- Most statisticians have become addicted to MCMC (one called it “Bayesian cocaine”).
- Other methods can provide better intuitive understanding and show the connection of Bayesian methods to familiar methods.

Prior-Data Equivalents

Using statistics in observational studies requires understanding priors and their ubiquity in **all** statistics.

- Data equivalents (data-augmentation priors, DAPs, or just **data priors**) are one way of understanding the logical strength of a prior distribution.
  - They also allow one to get Bayesian results from conventional software.

Example: A case-control study

Relation of maternal antibiotic use during pregnancy (X = 1) to sudden infant death syndrome (SIDS, Y = 1), NICHHD collaborative study from the 1980s.

Antibiotics might be associated with

- elevated risk (marking effects on the fetus of an infection, or via a direct effect), or
- reduced risk (by reducing presence of infectious agents).

These are weak speculations, but suppose strong effects seem unlikely.

One plausible prior for ln(OR) ≈ ln(RR) is normal, mean 0, variance ½ , because it puts

- Equal odds on causal vs. preventive association (prior median for OR of 1)
- 2:1 odds on OR between ½ and 2 ≈ $e^{\pm\sqrt{½}}$
- 95% probability on OR between ¼ and 4:

$$\exp(0 \pm 1.96\sqrt{½}) \approx ¼, 4$$

Now that you've seen the prior, here are the actual data:

	X=1	X=0
Y=1	173	602
Y=0	134	663

OR = 1.42, ln(OR) = 0.352

Standard Error for ln(OR) = 0.128

95% CL for OR = 1.11, 1.83

Posterior approximation via inverse-variance (information) weighted averaging of prior and sample means:

- Approximate mean, median, mode of ln(OR) given prior and likelihood ("data")
  $= (0/½ + .352/.128^2)/(1/½ + 1/0.128^2)$
  $= 0.341$
- Approximate variance of ln(OR)|prior+lik
  $= 1/(1/½ + 1/0.128^2) = 0.126^2$
- Approx. posterior median OR = 1.41
- Approx. 95% posterior limits = 1.10, 1.80

Technical Notes:

- The posterior odds ratio may be adjusted by adjusting the sample estimate and setting the prior on the adjusted association.
- *Partial-Bayes* (semi-Bayes, "mixed model") means that some but not all explicit model parameters are given an explicit prior
- The above analysis is partial-Bayes because it does not use an explicit prior for the exposure prevalence Pr(X=1).

- In fully Bayesian methods there is an explicit prior on all free parameters.
- In conventional frequentist methods there is no explicit prior for any explicit parameter; but the implicit priors are absurd:
  - They place no constraint on the explicit parameters.
  - They set implicit parameters to single values, as if they were known with certainty (which they aren't).

- The partial-Bayesian results above hardly differ from frequentist results, due to the large sample size, weak prior (variance ratio $½/.128^2 > 30$) , and small difference (in standard errors) between sample and prior;

BUT

- The above results (**both frequentist and Bayesian**) are fantasies as effect estimates, because they take no account of potential bias sources (which are implicit parameters).

What data would provide information equivalent to the prior?

What perfect experiment would give

- 0 as the conventional ln(OR) estimate
- ½ as its estimated variance?

Answers to such questions can be found by thought experiments, and reveal a deep connection between frequentist and Bayesian methods…

Suppose we are given the results of a randomized trial with equal allocation of N mothers to X=1 and N to X=0.

The risk ratio would then equal the ratio of the number of exposed cases A_1 to the number of unexposed cases A_0:

$$RR = (A_1/N)/(A_0/N) = A_1/A_0$$

Given the rarity of SIDS, OR ≈ RR and variance of $\ln(RR) \approx 1/A_1 + 1/A_0$.

To yield our prior, the data must satisfy

$RR = A_1/A_0 = 1$, so $A_1 = A_0 = A$

$\text{var } \ln(RR) \approx 1/A + 1/A = 2/A = ½$,

so $A_1 = A_0 = A = 4$, total $= A_+ = 8$.

Thus, a data set roughly equivalent to our normal (0,½) prior would have 4 cases in each arm of a perfect simple RCT.

NOTE: The method depends only on having N>>A; actual U.S. 1980s SIDS risk ≈ .001

Augment observed data with prior data as a separate stratum. Prior N need only be very large – only relative (not absolute) sizes of the Ns are used for calculating RR:

	$Prior_X=1$		$Prior_X=0$	
	X=1	X=0	X=1	X=0
Y=1	4	4	173	602
Y=0	100,000	100,000	134	663

$RR_{prior} = 1$ OR = 1.42

95% PL = 0.25, 4.00 95% CL = 1.11, 1.83

Now obtain a summary estimate using any uniform (constant) odds-ratio method…

Approx. posterior median and 95% limits from Mantel–Haenszel and maximum likelihood:

1.41 (1.10, 1.80),

same as the information-weighted average (the Woolf method).

To obtain conventional (frequentist) estimate and confidence limits, set $A_1 = A_0 = 0$, which shows the frequentist limits are Bayes limits in the extreme case of zero prior information.

Checking the prior

Just as you should check homogeneity before summarizing across strata, you should check the compatibility of the prior and the actual data. Simple starting check:

- test equality of the actual-data and prior estimates; equivalently
- test the product term $\text{Prior}_X \bullet X$ in a regression of Y on X, covariates, and Prior_X

Checking the prior in the example:

- For the above example, the test statistic is

  $\{\ln(1.42) - \ln(1)\}/(.128^2 + 0.500)^{1/2}$

  The normal *P*-value is 0.63, so by this check the prior and the data appear "compatible."

In more complex settings, however, passing a test does not imply compatibility along all dimensions.

For a regression, we can add the stratum as a set of **two** prior records (X=1,0) with

- an indicator $Prior_X$=1 for the two prior records, 0 for the remaining data
- other regressors set to 0 (upon recentering)

The earlier prior table then looks like

Cases	Total	X	Z_1	…	Z_J	$Prior_X$
4	100,000	1	0	…	0	1
4	100,000	0	0	…	0	1

$Prior_X$ = 0 for all other records.

Entry of prior records requires data entered as "grouped" with a column for the cases (event) counts and a column for the total represented by the record, with $\text{Prior}_X = 0$.
A single case record will then look like

Cases	Total	X	Z_1	…	Z_J	Prior_X
1	1	x	z_1	…	z_J	0

while a single noncase record will look like

Cases	Total	X	Z_1	…	Z_J	Prior_X
0	1	x	z_1	…	z_J	0

Prior-data coding check: Fit the logistic model $\Pr(Y=1| X=x) = \text{expit}(\alpha+x\beta)$ to the prior data alone, where $\text{expit}(u) \equiv e^u/(1+e^u)$:

Cases	Total	X	Z_1	…	Z_J	Prior_X
4	100,000	1	0	…	0	1
4	100,000	0	0	…	0	1

With only these two records, the program solves

$$\text{logit}(4/10^5) = \alpha+\beta,\ \text{logit}(4/10^5) = \alpha$$

$$v = \text{var}(\hat{\beta}) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}$$

to get $\hat{\alpha} = \ln(4/10^5), \hat{\beta} = 0, \text{SE}(\hat{\beta}) = \sqrt{1/2} = 0.707$

Prior RCT with Non-null center

To get $RR_{prior} \neq 1$, set $N_1/N_0 = 1/RR_{prior}$:

For an approximate lognormal prior for RR with median at 2 instead of 1, with the same ln(RR) variance of ½, use

	X=1	X=0
Y=1	4	4
Y=0	100,000	200,000

$RR_{prior} = 2$, 95% prior limits = 0.50, 8.0

General outcomes, regressors, and models (including quantitative Y)

Say the model is $g(\mu_{xz}) = \alpha + x\beta + \mathbf{z\theta}$

A data prior for β may then be an additional stratum with

- its own indicator Prior_X
- all covariates Z_j at 0 (after recentering)
- a point estimate (MLE) for β and confidence limits equal to the desired prior mode and limits.

Coefficient priors: Bit (**bi**nary digi**t** or **B**ernoulli **t**ransmission) representation

What is the simplest kind of bit string that conveys the same amount of information as in a N(0,v) distribution for a logistic coefficient?

- Approximate answer using Fisher information (precision) 1/v: Any bit string with A zeros and A ones, where A = 2/v (so v = 1/A + 1/A = 2/A).

For v < ½ and hence A > 4, the approximation is fine for epi, although it can be improved by adding "continuity corrections": A = 2/v + ½

Replacing the program-forced constant

- Requires a command to leave out the constant or intercept (Stata: **noconstant**, SAS: **NOINT**)
- In its place we enter a new variable, Const:

  Const = 1 for actual data **or** intercept-prior data,

  0 for all other prior data
- The coefficient of Const is the intercept α in the generalized linear model (GLM) for actual data

$$g(\mu_{xz}) = \alpha + x\beta + \mathbf{z\theta} = \text{Const}\bullet\alpha + x\beta + \mathbf{z\theta}$$

(To fit an actual no-intercept model $g(\mu_{xz}) = x\beta + \mathbf{z\theta}$, specify no constant and leave out Const as well)

With Const, we can force an independent N(0,v) prior on a logistic coefficient by adding one binomial record that gives back the prior.

Check: Fit the **no-intercept** model

$$\Pr(Y=1 \mid X=x) = \text{expit}(x\beta)$$

to the prior data. Data for a N(0,½) prior:

Cases	Total	X	Z_1	…	Z_J	Const
4	8	1	0	…	0	0

The program solves logit(4/8) = logit(½) = 0 = β,

v = ¼ + ¼ = ½ to get $\hat{\beta} = 0, \text{SE}(\hat{\beta}) = \sqrt{1/2} = 0.707$

The offset variable H

- An offset is a variable H added to the entire data set whose coefficient will be set to 1 by the program, so the fitted model becomes

$$g(\mu_{xzh}) = \alpha + x\beta + \mathbf{z\theta} + h$$

- Requires a command to declare H an offset (Stata: **offset(H)**, SAS: **OFFSET=H**)
- H = 0 for actual data (except for actual offsets), H= −m for prior data records (which may be 0) where m is the prior mode (most probable value)
- H is not needed if all the prior modes are zero.

Using an offset, we can force an independent N(m,v) prior on a coefficient by adding a record to the data that gives back the prior. To check, fit the **no-intercept offset** model

$\Pr(Y=1| X=x, H=h) = \text{expit}(x\beta+h)$

to the prior data. Data for a N(ln(2),½) prior:

Cases	Total	X	Z_1	…	Z_J	Const	H
4	8	1	0	…	0	0	−ln(2)

The program solves $\text{logit}(4/8) = 0 = \beta - \ln(2)$ to get

$$\hat{\beta} = \ln(2),\ \text{SE}(\hat{\beta}) = \sqrt{1/2} = 0.707$$

Entry of prior records requires data entered as "grouped" with a column for the cases (event) counts and a column for the total represented by the record; Const=1; and H=0 or real offset
A single case record with offset will look like

Cases	Total	X	Z_1	…	Z_J	Const	H
1	1	x	z_1	…	z_J	1	0

while a single noncase record will look like

Cases	Total	X	Z_1	…	Z_J	Const	H
0	1	x	z_1	…	z_J	1	0

An alternative to grouped-data entry is to use weighted records, which require two records per prior. Data for a N(ln(2),½) prior:

Y	Weight	X	Z_1	…	Z_J	Const	H
1	4	1	0	…	0	0	−ln(2)
0	4	1	0	…	0	0	−ln(2)

$$\hat{\beta} = \ln(2),\ \mathrm{SE}(\hat{\beta}) = \sqrt{1/2} = 0.707$$

Actual case and noncase records will then look like

1	1	x	z_1	…	z_J	1	0
0	1	x	z_1	…	z_J	1	0

Essential preprocessing for quantitative covariates

For every quantitative variable, as needed

- recenter so that 0 is a meaningful, relevant value that occurs within the body of the data
- rescale so that 1 is a meaningful, relevant span that occurs within the body of the data

Thus each variable X should be transformed to ensure that a comparison of X=1 to X=0 is meaningful, relevant, and not extrapolating beyond the range of X in the data.

Failure to describe the scale can produce unintelligible, useless results:

From Medscape/WebMD (Duffy et al. J Clin Oncol 2009; 27, 1969-1975):

"Physical activity [PASE: 10 points (out of 300?)*] was also associated with survival in univariate analysis and was approaching significance [P=0.085] in the multivariate model (hazard ratio 0.98; 95% CI 0.95-1.00)."

*(PASE is proprietary and data about it is difficult to find)

Priors for Quantitative Variables

- For a quantitative variable X modeled by one parameter, the prior data represent a trial comparing 1 unit of X vs. 0 units of X.

Thus, to set a reasonable prior on the coefficient of X, it is essential to recenter and rescale X so that

- 0 and 1 are contextually meaningful values for X within the data range, and
- 0 and 1 are meaningfully far apart.

- diastolic blood pressure in **cm** deviation from 8 (cm–8) rather than in mm above 0
- smoking intensity in **packs/day** rather than cigs./day (0 is already meaningful)
- hours of sleep/day in **deviations** from 8 hours rather than in total hours (units already meaningful)

We can then gauge the strength of beliefs held by those taking substantive positions...

Example: Let T = 1 microtesla average magnetic-field exposure

Some writers remain convinced that there is no effect of T on childhood leukemia.

- They seem at least 95% certain that the RR for T=1 is between 0.91 and 1.1.

That corresponds to a perfect simple RCT with at least $A = 2/\{\ln(1.1)/1.96\}^2 \approx 850$ cases per arm of a disease whose rate is only $4/10^5$ per year = about **1700** cases total.

Where did all this information come from?

At 4 cases per 10^5 child-years, a trial would have to observe over **40,000,000** child-years to get the 1700 required cases. In reality,

- No trial has been or ever will be done.

Some (not all) physicists claim that an effect would violate the laws of physics. In reality,

- Laws of physics tell us nothing about how poorly understood biologic systems will react to sustained field exposure.

Spiked priors are "semi-dogmatic" claims

A null spike of size (say) ½ represents betting even odds that background information can prove $\beta=0$ *with absolute certainty*.

- When there is no such decisive prior information, a spike represents an unscientific faith in, or commitment to, the null, with no empirical foundation in most health and social-science applications. Thus the results it produces are examples of "spinning knowledge out of ignorance" (Greenland and Poole, Epid 2013).

Opposite extreme:
"Reference Bayes", "Objective Bayes"
= "utterly ignorant of subject" Bayes

Priors in these methods correspond to $A \leq ½$:

- $A = ½$ gives a 95% prior probability that RR is between 1/648 and 648. In everyday epidemiologic terms this prior is almost total ignorance (648 is greater than the RR for some of the deadliest infections). As a consequence, the methods give intervals close to the conventional (A=0) intervals.

An application of Bayes: Too many measured covariates chasing too few data

Conventional solution: do variable selection by conventional methods (e.g., stepwise, best subsets).

Consequences: Poor to disastrous –

- May discard key confounders in favor of less relevant variables
- Grossly overstates significance of what is retained (first documented in 1944!)

Ordinary variable selection is biased, causing:

- Unnecessary residual confounding
- Interval estimates too narrow (up to a half).

A common counter-solution: Leave every variable in if possible (make *ad hoc* simplifications if not).

- A common consequence:

Sparse-data bias = estimate inflation
(and general failure of asymptotic behavior)
-- might be aggravated by PS/IPT weighting

Classic example: Appetite suppressants and PPH (Abenhaim et al. NEJM 1996):

- Conditional ML logistic regression odds ratio (OR) for >3 mos. use: 23 (CL: 6.9, 78) BUT this is from adjusting **seven** covariates with only **five** exposed controls!
- Crude OR: 18; other adjustments and higher risk expected for those who use suppressants suggest possible **upward** confounding. Thus the adjustment was expected to **decrease** the estimate, but increased it instead…

Phenylpropanolamine appetite suppressants and stroke (Kernan et al. NEJM 2000):

- CML logistic OR: 16 (CL: 1.4, 184)
  BUT from adjusting **four** covariates with only **one** exposed control!
- Crude mid-p OR is 11; again, higher risk was expected for those who use suppressants which suggests possible **upward** confounding, so adjustment was expected to **decrease** the estimate. Instead it increased the estimate and so may have worsened bias.

Bayesian and related alternatives to variable selection and maximum likelihood

As known since the 1970s (e.g., Leamer, 1978), there is a family of solutions that beats more common approaches.

It is known by many names: Stein, empirical-Bayes, partial-Bayes, shrinkage, penalized estimation; ridge regression; BLUP; hierarchical, multilevel, & random-coefficient modeling (review: Greenland, AJE 2008).

Purpose: If one is tempted to draw inferences "from the data," shrinkage methods can help moderate temptations toward excessive certainty, especially excessive certainty encouraged by conventional identified-model based results (which are saturated with opaque assumptions). Each covariate Z gets a prior e.g., 9 prior records for 9 covariates, each with

- its covariate Z, coded 1,
- zeros for all other regressors in the model.

Some earlier epidemiologic applications of partial-Bayes shrinkage for multiple exposures:

Greenland S, Finkle WD (implants and chronic disease). Annals Epidemiol 2000;10:205-213.

Aragaki C, Greenland S, et al. (genes and colon polyps). CEBP 1997;6:307-314

Witte JS, Greenland S, et al. (foods and breast cancer). Epidemiol 2000;11:684-688.

Greenland S (nutrients and breast cancer). Biometrics 2000;56:915-921.

Why null hypotheses do not deserve special prior weight in most analyses

There is rarely a **scientific** basis for asserting that a null hypothesis deserves some special favor beyond being the prior mode or median, e.g., in the form of fixing Type I error at the cost of Type II error, or having a spike (point prior mass) at 0.

- Such devices represent confusion of inference (trying to determine what is) with decision making (trying to decide what to do). Bayesian inferences require only posterior probabilities; Bayesian decisions additionally require loss functions.

An example from nutrient epidemiology:

Greenland S (Biometrics 2000;56:915-921):

- 140 female breast-cancer cases

+ 222 sister controls = 140 matched sets

- 5 forced covariates **W** (centered age, etc.)
- 87 foods **X**
- 35 nutrients **XZ** computed from the foods **X** using the diet-nutrient matrix **Z**, and hence **completely collinear** with foods …

Multilevel models for measurement

When exposure measures are constructed from more basic data (e.g., nutrients from diet histories, chemicals from job histories), we need **hierarchical (multilevel) models** to reflect this fact.

- If nutrient vector = matrix•(food vector)

  then a model saying g(risk) = $\alpha + \pi$•nutrients

  claims that all food effects are transmitted through the **measured** nutrients.

We have good reasons to doubt this model, and using it leads to grossly overconfident inferences.

To remove the unwarranted assumption, we must expand the model in a single-level ("random-coefficient") form:

$$g(\text{risk}) = \alpha + \pi \bullet \text{nutrients} + \delta \bullet \text{foods}$$

which cannot be fit without a prior because of the collinearity of nutrients and foods.

- Omitting foods is a N(0,0) prior for δ: It says we are certain there is no food effect not captured by our **measurements** of nutrients and the linear model form (the standard nutrient-epi prior), so we are sure that $\delta = 0$.

$\delta = 0$ is a junk prior that generates poorly calibrated inferences, because $\delta \neq 0$

Likewise, when chemicals = M•jobs,

$$g(\text{risk}) = \alpha + \pi \bullet \text{chemicals} + \delta \bullet \text{jobs}$$

cannot be fit without a prior because of the collinearity (complete confounding) of chemicals and jobs.

- Omitting jobs is a N(0,0) prior on δ: We are 100% certain there is no job effect not captured by our **measurements** of chemicals and the linear model form, so $\delta = 0$.

Bayes 0: Use a prior that is not contextually nonsensical (Greenland, Biometrics 2000)

Various ways to do so. Two easy ways:

- Give a betting interval for the parameter, e.g., "I'd bet 2:1 that RR = exp(δ) is between ½ and 2, with equal odds above and below this interval."
- State what you expect the parameter to be, then state what is your expectation is "worth" in terms of a perfect RCT.

Claiming a parameter is zero when there is no evidence is just poor judgment and exaggerates certainty

It is easy to do better than using "$\delta = 0$":

- "$\delta = 0$" is the same as $\delta \sim N(0,0)$,

  zero variance = infinite information about δ.
- This is equivalent to claiming your expectation is worth an infinitely large RCT -- a "pipeline to god" (NHS) prior.
- Instead, try $\delta \sim N(0,V)$, with V "small" but positive definite (a model for "such effects as present are not likely to be large").

Nutrient ORs e^{π}. Backward deletion, CML: **δ=0**.
Partial-Bayes: 95% prior limits ½, 2 (A=16.5) on e^{δ}

	Back Del (α=.10)	CML	Partial-Bayes
#cov:	15 (of 35)+5=20	35+5=40	35+87+5=127
Ω3 fatty acids	.77	.71	.58
(g/day)	**.65, .92**	**.46, 1.1**	**.17, 2.0**
Phytoestrogen	.80	.73	.73
(mg/day)	**.70, .92**	**.58, .93**	**.40, 1.3**
Alcohol	.94	.89	.93
(3 oz./day)	**.88, 1.00**	**.63, 1.3**	**.37, 2.3**
Carbohydrate	1	.97	.99
(100 g/day)	**1,1** (deleted)	**.79, 1.2**	**.58, 1.7**

At the other extreme: those who don't want to "contaminate their data" with priors.

Pure frequentist and "objective Bayes" answers flow from assuming that your background expectations - everything you think you know about the unknown model parameters - is worthless. No wonder that

- Math statisticians love these methods! And
- Some of the worst epidemiologic inferences have flowed from those slavishly devoted to frequentist methods…

Historical case-study:

Among leading statisticians in the smoking-lung cancer controversy were

- Cornfield, a nascent Bayesian working with epidemiologists in arguing for the effect,
- Fisher, an anti-Bayesian frequentist who defended cigarette smoking, asserted that the association might be pure confounding – essentially he refused to constrain δ at all, allowing it to just as well be ln(100) as 0.

What is the Bayesian meaning of "no prior" on a free parameter β in our model?

"No prior on β" corresponds to a $N(0,\infty)$ prior for β. If e^{β} is a relative risk RR, this prior gives equal odds on

- $RR = 10^{-100}$, a sufficient preventive
- $RR = 1$, no effect,
- $RR = 10^{100}$, a sufficient cause

Such a prior is **IDIOTIC** in most health and social science. It is even sillier than the priors used in "Objective Bayesian" methods.

Nutrient ORs e^{π}. Col 2: 95% PLs .80,1.25 on e^{δ} (A=154). Col 3: ¼ ,4 (A=4.5) on e^{π} and ½, 2 on e^{δ}

#cov	CML (35+5=40)	Partial-Bayes 40+87@(.8,1.25)	5+35@(¼,4) +87@(½,2)
Ω3 fatty acids	.71	.71	.68
(g/day)	**.46, 1.1**	**.39, 1.3**	**.28, 1.7**
Phytoestrogen	.73	.73	.77
(mg/day)	**.58, .93**	**.53, 1.01**	**.46, 1.3**
Alcohol	.89	.90	.91
(3 oz./day)	**.63, 1.3**	**.57, 1.4**	**.46, 1.8**
Carbohydrate	.97	.97	.98
(100 g/day)	**.79, 1.2**	**.74, 1.3**	**.66, 1.5**

Hierarchical (multilevel) modeling for foods:

Say the probability of being a case in sibship *i* if one has **X**=**x**, **W**=**w** is

$$\Pr(Y=1|\mathbf{x},\mathbf{w}) = \text{expit}(\alpha_i + \mathbf{x\beta} + \mathbf{w\gamma})$$

The food effect **β** is a function of measured-nutrient effects plus residual food effects.

If $\boldsymbol{\beta} = \mathbf{z\pi} + \boldsymbol{\delta}$, we have

$$\Pr(Y=1|\mathbf{x},\mathbf{w}) = \text{expit}(\alpha_i + \mathbf{xz\pi} + \mathbf{x\delta} + \mathbf{w\gamma}),$$

the same as the model for nutrient and residual-food effects used above.

Extending Data Priors

- Priors on product-term ("interaction") coefficients
- Warning on product-term and intercept priors
- Improving approximation accuracy
- Non-normal priors:
  - Priors with heavier tails than normal
  - Skewed priors
  - Rescaled priors
- Dependent priors

Priors on product-term coefficients

We can force an independent N(m,v) prior on a product coefficient ("interaction") by adding a record to the data that gives back the prior.

Check: Fit the **no-intercept offset product** model

$\Pr(Y=1| X=x, Z=z, H=h) = \text{expit}(xz\gamma+h)$

to the prior data. Data for a N(0,½) prior:

Cases	Total	X	Z	…	XZ	Const	H
4	8	0	0	…	1	0	0

Note: X = 0, Z =0, yet XZ = 1, **so the record is physically impossible**. But it gives back the γ prior.

Warnings on product and intercept priors:

- In ordinary regression there is a "hierarchy principle" that says: Unless you have solid physical grounds for doing so, do not enter a higher-order term without all its components as well, e.g., do not enter XZ without X and Z. (Not entering X equals a 0-mean, 0-variance prior.)
- A Bayesian generalization: Always put much weaker priors on lower-order terms than on higher-order terms, unless you have solid grounds to do otherwise.

Examples

Say the model has $x\beta+z\gamma+xz\delta$. Then δ should get the strongest (perhaps only) prior.

- $\Pr(\beta)$ is the prior for the relation of X to Y only when Z=0; the prior for the relation of X to Y when Z=1 is the implicit prior for $\beta+\delta$, which must have larger variance than β (if β and δ have independent or positively correlated priors).

Analogously, if the model has an intercept α, its prior is best left out or left very weak relative to other terms unless there is solid information about it.

Approximation Accuracy

- For logistic coefficients, the above normal approximations are adequate down to A = 4.
- For prior mode m, the exact prior distribution for $e^{\beta - m}$ implied by a binomial prior-data record with A_1 cases and A_0 noncases is $F(2A_1, 2A_0)$.

Example: $\exp(1.96v^{1/2}) = 4$ gives $A = 2/v = 4$

But, the exact prior $\Pr(1/4 < e^{\beta - m} < 4)$ is 93.3%, which is the probability that an F(8,8) variate is between ¼ and 4.

Improving Accuracy: Yates correction

Better prior intervals can be obtained by adding ½ **after** computing the desired A. Examples:

- Let A = 4.5 instead of 4. From an F(9,9) table, $\Pr(¼ < e^{\beta - m} < 4)$ is now 94.9%.
- $\exp(1.96v^{½}) = 16$ gives v = 2 so A= 2/v =1. But… from an F(2,2) table, the exact prior $\Pr(1/16 < e^{\beta - m} < 16)$ is only 88.2%. So… Let A = 1.5 instead of 1. From an F(3,3) table, $\Pr(1/16 < e^{\beta - m} < 16)$ is now 95.2%.

So the add-½ rule is fine down to an initial A of 1.

Table showing A for approx. normal and exact F:

Prior 95% limits	A from Wald	A from F(2A,2A)	
1/50, 50	0.50	0.92	
1/40, 40	0.57	0.99	[1 is OK]
1/16, 16	1	1.5	
1/8, 8	1.8	2.3	
1/7, 7	2	2.5	
1/5, 5	3	3.5	
1/4 , 4	4	4.5	
1/3 , 3	6.4	6.9	[7 is OK]
1/2 , 2	16	16.6	[16.5 is OK]
2/3 , 3/2	46.7	47.3	[47 is OK]
3/4 , 4/3	92.8	93.4	[93 is OK]
4/5 , 5/4	154.3	154.9	[155 is OK]
5/6 , 6/5	231.1	231.7	[232 is OK]

Software warning:

Some packages will truncate fractional counts (not SAS or Stata logistic procs).

In those packages, if you add a record with $A_1 = A_0 = 4.5$, you will get $A_1 = A_0 = 4$. This is not very important as long as you understand what got used. Nonetheless,

- **Always have the program print back the data to see if truncation, rounding, or some other problem occurred.**

Skewing the Prior

To skew to the upward (to right) while keeping same prior mode m, skew prior data to the X=1 column and set $H = \ln(A_1/A_0) - m$.

Hypothetical RCT generating the prior:

	X=1	X=0
Y=1	A_1	A_0
Total	N_1	N_0

mode $m = \ln(RR_{prior}) = \ln(A_1/N_1) - \ln(A_0/N_0)$

Make table $RR_{prior} = e^m$ by using $N_1/N_0 = e^m A_1/A_0$

Example: Skewing upward (positive skewing, heavier right tail)

RCT example with $A_1 = 60$, $A_0 = 3$, m = 0:

Make table RR = 1 by setting $N_1/N_0 = A_1/A_0$

	X=1	X=0
Y=1	60	3
Total	600,000	30,000

$RR_{prior} = e^0 = 1$, offset H = ln(60/3) − 0 = 2.996

To skew downward, skew prior data to X=0 (negative skewing, heavier left tail)

Data record for this prior:

Cases	Total	X	Z_1	…	Z_J	Const	H
60	3	1	0	…	0	0	2.996

Program applied to this record solves $\ln(60/3) = \beta + 2.996$, yielding $\hat{\beta} = 0$

Percentiles for this prior, from F(120,6):

Pr(RR < ¼) =	0.1%	Pr(RR > 4) =	4%
Pr(RR < ½) =	7	Pr(RR > 2) =	19
Pr(RR < ⅔) =	18	Pr(RR > 1.5) =	32
Pr(RR < 1) =	43	Pr(RR > 1) =	57

Procedural caution for using skewed priors

In general (apart from Bayesian analysis), if there is concern about data sparsity in a logistic regression, it is advisable to have your program compute profile-likelihood (likelihood-ratio) confidence intervals (Stata: **pllf** or **logprof**; SAS: **PRL=LR**).

Use of priors can greatly reduce sparse-data bias, but skewed priors may aggravate the failure of conventional Wald ($\hat{\beta} \pm 1.96\text{SE}$) intervals if either A_1 or A_0 is small. **Thus, when you use skewed priors it is especially important to also use the profile-likelihood interval option.**

Rescaling the Prior

Suppose we like the shape just shown but want to *rescale* the percentiles by a factor of S on the log scale. Then we make the prior record refer to a group in which X= 1/S; meaning we divide X by S and set $H = \ln(A_1/A_0) - m/S$:

S=2:	X= ½	X=0
Y=1	60	3
Total	600,000	300,000

$m/2 = \ln(RR_{prior}) = 0$, $H = \ln(60/3) - 0/2 = 2.996$

Data record for prior **expanded** by S=2:

Cases	Total	X	Z_1	…	Z_J	Const	H
60	63	½	0	…	0	0	2.996

Percentiles are now spread by a factor of S=2, e.g., the percents for ½, 2 are now for ¼, 4:

Pr(RR < ¼) = 7% and Pr(RR> 4) = 19%;

Before rescaling these were 0.1% and 4%.

If instead we want to **compress** percentiles by a factor S= ½, we use X= 1/S = 1/½ = 2 and

$H = \ln(A_1/A_0) - m/½ = \ln(A_1/A_0) - 2m$

Effect of rescaling on prior information (equivalent prior sample size)

The prior information is proportional to the total prior case number $A_+ = A_1 + A_0$

- The information-equivalent sample size after rescaling by S is A_+/S^2.
- In the example, expanding by S= 2 makes the prior data equivalent to a trial with only $A_+ = 63/2^2 \approx 16$ cases instead of 63 cases.
- S=½ creates equivalence to $63/½^2$=252 cases

Rescaling to improve normality

For prior symmetry, $A_1 = A_0 = A$

- If we want to preserve the variance after rescaling by S, multiply A by S^2.
- If we want accurate normality, make $S \geq 10$

Check of data for a N(ln(2),½) prior:

Cases	Total	X	Z_1	…	Z_J	Const	H
400	800	1/10	0	…	0	0	−ln(2)/10

Program solves $\text{logit}(400/800) = 0 = \beta/10 - \ln(2)/10$
$v = (1/400 + 1/400)/(1/10)^2 = ½$ to get

$$\hat{\beta} = \ln(2),\ \text{SE}(\hat{\beta}) = \sqrt{1/2} = 0.707$$

Technical footnotes (ignorable):

The F-prior for $e^{\beta - m}$ corresponds to a generalized-conjugate (GC) prior for β:

The GC prior for $\beta = \ln(RR)$ with mode m and scale S is a recentered and rescaled log-F distribution,

$$\beta \sim S \bullet \ln(F) + m,$$

where F is an F-variate with $2A_1$, $2A_0$ df

This prior corresponds to a prior RCT table with X = 1/S, 0 (instead of 1,0) and $RR = e^m$

When offsets are not a software option, we can force the coefficient for H to 1 by adding a prior-data record that has H=1 and

- $A_1 = 10^6 e = 2{,}718{,}282$ cases
- total count $A_+ = 10^6(1+e)$,

  (so noncase count is $A_0 = 10^6$)
- all other variables set to zero.

From this record, the coefficient of H will be

$$\text{logit}(A_1/A_+) = \ln(A_1/A_0) = \ln(e) = 1$$

with SE ≈ 0.001 (practically zero).

Two Poisson-regression options that use a single record per prior

1) Convert actual data to binomial form, then use logistic regression: For actual records,

a) use the observed count as number cases;

b) convert the person-time into units so small (e.g., hours) that the amount N is always more than 10^5 times the number cases;

c) use the person-time N as the total count.

Then proceed as with logistic models (replace intercept with Const, add H, etc.)

2) If all priors are normal, instead one can use Poisson regression with the offset H merged into the person-time:

a) Leave the actual data intact, except replace the intercept with Const, as in logistic.

b) In the prior record for the coefficient for X: Set the Poisson count A to S^2/v,

Person-time = A/exp(m/S) with $S \geq 30$ (needs a larger S to force normality),

X=1/S and all other covariates zero

Conditional-logistic & Cox regression

Need to add two groups of prior records, but no Const needed (these models have no intercept):

Add A_1 discordant matched pairs with

- case at X=1/S, H = −m/S,
- noncase at X=0, H=0.

Add A_0 discordant matched pairs with

- case at X=0, H=0,
- noncase at X=1/S, H = −m/S.

Cox model requires these pairs be declared strata; they can all be assigned time = 1. Or, fit Cox model by logistic regression (Efron, JASA 1988)

Dependent priors can be created by

- adding prior data in the form of a multiway table whose maximum-likelihood estimates under the model show the desired correlations (Greenland, Biometrics 2001, 2003);

or by

- modeling dependencies hierarchically using prior (2^{nd}-stage) covariates (Greenland, Biometrics 2000, JASA 2003, JRSS 2005).

Nonidentified Bias Modeling: The Missing-Data Perspective

- Random sampling implies: population members (individuals) are missing from our sample at random (no selection bias).
- Randomization implies: potential outcomes are missing at random (no confounding).

We'd like all missing data to be missing at random (MAR) –

but they're not.

The usual validity problems are all bias due to missing data

- Confounding: nonrandomly missing potential outcomes
- Selection bias: nonrandomly missing subjects
- Measurement error: missing actual variables of interest, so we use proxies in their place (which may produce bias even if the **error** is random)

What we **always** do:

Completed data = observed + missing data

- To make any inference beyond what we see (the observed), we must have a model (a set of assumptions) that projects from the observed data to the missing data (or to aspects of the missing data, like means) to get the completed data; that is,

we must always impute missing information.

The standard model (the BIG lie)

Everything is missing at random (MAR)

- This model is the basis for **every** statistical method and program in use in health sciences today.
- Its deficiencies are usually addressed by dismissive (often flawed) judgments in discussion sections.
- Sometimes, they are addressed by a limited sensitivity analyses.

Sensitivity Analysis

- Expands the model to allow for fixed known departures from MAR, using bias models = sensitivity models = nonignorable-missingness models
- For various degrees of departure from MAR (for various values of the bias parameters), sensitivity analysis displays the results obtained pretending that that degree of departure is correct.

Sensitivity analysis is valuable – and limited

- It requires specification of the precise form of bias sources (MAR departures) – a major benefit! … But:
- It becomes cumbersome as the number of bias dimensions grows, and unintelligible beyond 3 dimensions;
- As a result, it leaves us with only vague judgments based on hidden priors.

Probabilistic Sensitivity Analysis and Bayesian analysis

- Ways of forming probabilistic judgments about the target parameters, accounting for specified bias sources
- Provide a **personal** weighted average (summary) over a sensitivity analysis using **explicit** priors for weighting

Problem: Explicit priors incur objections -- yet inference without priors is impossible.

Modeling of Unmeasured Variables

The Bayesian approach to covariates can be extended to include priors for *unmeasured* covariates (e.g., Leamer JASA 1974; Gustafson 2003, Stat Sci 2005; Greenland JASA 2003, JRSS 2005, Stat Sci 2009).

- These are examples of *nonidentified bias modeling*, in which the conventional model is expanded to allow unmeasured influences (latent variables).

Consider again the SIDS study:

	X=1	X=0
Y=1	173	602
Y=0	134	663

OR = 1.42, ln(OR) = .352

Standard Error for ln(OR) = .128

95% CL for OR = 1.11, 1.83

A Misclassification Example

X above represents only mother's report of antibiotic use, which must often be mistaken.

Let T = the indicator of true antibiotic use.

There is no doubt that often $T \neq X$, and this event likely depends on the outcome Y:

We expect

- more false positives among cases, that is: more $T<X$ if $Y=1$ than if $Y=0$.
- more false negatives among controls, that is: more $T>X$ if $Y=0$ than if $Y=1$.

We want the marginal TY odds ratio OR_{TY} from the TXY table collapsed over X.

But, if we only have the XY data,

- Data on T are missing for everyone; hence T is a latent variable
- We need information that allows us to predict (impute) T from XY. That is, we need information on the **predictive values**

$$\pi_{txy} \equiv P(T=t|X=x,Y=y).$$

Information on π_{txy} would allow us to reconstruct the desired TY distribution:

	X=1		X=0	
	T=1	T=0	T=1	T=0
Y=1	$173\pi_{111}$	$173\pi_{011}$	$602\pi_{101}$	$602\pi_{001}$
Y=0	$134\pi_{110}$	$134\pi_{010}$	$663\pi_{100}$	$663\pi_{000}$

Uncorrected stats (95% limits 1.11,1.83) are obtained by pretending that

$\pi_{xxy} = P(T=x|X=x,Y=y) = 1$, i.e., that X=T

The conventional analysis assumes **no** error, an extreme prior no one holds.

We need a more realistic model, rather than routinely defaulting to a ridiculous, extreme, and false assumption.

- To more easily connect with typical prior information, reparameterize the predictive values with a logistic model:

  $\pi_{1xy} = \text{expit}(\beta_T + \beta_{TX}x + \beta_{TY}y + \beta_{TXY}xy)$

  where $\text{expit}(u) \equiv e^u/(1+e^u)$

Ordinary sensitivity analysis is unintelligible for 4 parameters

With only 3 values for each parameter, we'd have to examine a table of $3^4 = 81$ "adjusted" odds ratios.

- With no further indication of the relative plausibility of each combination, any simple summary (e.g., an average, a median, a quartile) would correspond to posterior summaries based on a prior that placed 1/81 probability on each combination – a prior no one would hold.

Why not just use a plausible prior?

The TX odds ratios equal the ROC (receiver operating characteristic) odds ratios:

TX odds ratio when Y=y:

$\exp(\beta_{TX}+\beta_{TXY}y) =$

true-positive odds/false-positive odds =

Sensitivity/False-negative rate

False-positive rate/Specificity

Note that $\beta_{TX}= \beta_{TXY} = 0$ for a worthless X (X independent of T given Y).

Specifying β_{TX} :

$\exp(\beta_{TX})$ = ROC odds ratio among noncases

$= Se_0Sp_0/Fn_0Fp_0$

$\exp(\beta_{TX})$ would be expected to be high even for a mediocre X. For example,

- for sensitivity .6 and specificity .9 among noncases, $\exp(\beta_{TX}) = (.6/.4)/(.1/.9) = 13.5$

Also, $\exp(\beta_{TX}) = .6(.8)/.4(.2) = 6$,
$.8(.8)/.2(.2) = 16$, $.8(.9)/.2(.1) = 36$

Specifying β_{TXY} :

- $\beta_{TXY}>0$ if cases report more accurately than noncases (as measured by ROC odds ratio), e.g., due to higher sensitivity among cases
- $\beta_{TXY}<0$ if vice-versa, e.g., due to higher specificity among controls
- If the misclassification is "nondifferential" (X independent of Y given T) then $\beta_{TXY} = 0$

  Nondifferential is exactly $\beta_{XY} = \beta_{TXY} = 0$

Specifying expit(β_T):

expit(β_T) = Pr(T=1|X=0,Y=0) = π_{100}
= probability among noncases that a "test negative" (X=0) is a false negative.

- For T=1 uncommon (<20%) and a highly sensitive test X (Se>90%), this probability would be small (<2%).
- We can place our prior directly on this probability, then transform this π_{100} prior back to a prior for β_T = logit(π_{100}).

Specifying β_{TY}:

$\exp(\beta_{TY})$ is the TY odds ratio when X=0.

- Under nondifferential misclassification, X is independent of Y given T, making the OR_{TY} odds ratio collapsible over X:

$$OR_{TY} = \exp(\beta_{TY}).$$

Thus, if our priors for β_{XY} and β_{TXY} are centered at zero and not diffuse, it may be reasonable to use our $\ln(OR_{TY})$ prior for β_{TY}, perhaps widened slightly.

One set (of many) of independent-normal priors suggested by the above considerations - a test case to see the sensitivity of the implausible conventional results to use of plausible priors instead:

	mean	variance	95% prior limits
β_T	logit(.1)	.16	expit(β_T): .05,.20
β_{TX}	ln(13.5)	.25	exp(β_{TX}): 5, 36
β_{TY}	0	.50	exp(β_{TY}) : ¼, 4
β_{TXY}	0	.125	exp(β_{TXY}): ½, 2

Ways to use the classification prior to make probabilistic inferences

Some methods from the risk assessment and policy literature:

- Monte-Carlo Sensitivity Analysis (MCSA; see second half of Ch. 19, ME3)
- Bayes and partial-Bayes (semi-Bayes)

MCSA equivalent to partial-Bayes analysis when (as here) the only parameters with a prior are not identified by the data model.

Many variations are possible

Advantages of MCSA: Fast, intuitive.

Advantages of Bayes: Unlike MCSA,

- Can be used with priors on any set of coefficients (priors are not limited to nonidentified parameters)
- Can be used with validation data and second-stage (two-phase) data – those data are entered as complete records
- Automatically avoids impossible values

Technical note: Let $E_{txy} = \pi_{txy}E_{+xy}$ be the expected counts at T=t, X=x, Y=y.

- The joint TXY distribution is determined by the π_{1xy} and E_{+xy}
- But, the distribution of the observables (X,Y) depends only on the marginal expectations E_{+xy}, not on the π_{txy}
- Hence, if there is no prior on the XY margin E_{+xy} or its parameters (no external constraint on the E_{+xy}), the π_{txy} distribution won't be updated and MCSA and Bayesian answers will be the same (apart from simulation error).

MCSA need involve only simple simulation. Example: a "basic bias bootstrap" algorithm

1) Draw β_T, β_{TX}, β_{TY}, β_{TXY} from their prior

2) Compute the π_{txy} from $\pi_{0xy} = 1 - \pi_{1xy}$ where

$$\pi_{1xy} = \text{expit}(\beta_T + \beta_{TX}x + \beta_{TY}y + \beta_{TXY}xy)$$

3) Fill in the missing T values using the π_{txy}, which results in an imputed TXY table

4) Collapse over X and compute $OR_{TY}^{(b)}$

5) Bootstrap the actual data to add in the random error:

Replace steps (3) and (4) by

a) Resample the observed counts once

b) Fill in the missing T values using the π_{txy}

c) Collapse over X and compute $OR_{TY}^{(r)}$

d) Correct for bias and random error:

$$OR_{TY}^{(c)} = [OR_{TY}^{(b)}]^2/OR_{TY}^{(r)}$$

Repeat 10,000+ times to get distributions of $OR_{TY}^{(b)}$ and $OR_{TY}^{(c)}$

Two Bayesian Poisson-regression methods:

Both start by coding the actual data as 4 cell counts with covariates T, TX, TY, TXY (entering missing value code as necessary), X, Y, XY, and a "person-time" of 1 for each count. Coefficients of T, TX, TY, TXY are β_T, β_{TX}, β_{TY}, β_{TXY}. Then either:

1) Translate each prior into 2 records with counts A_1,A_0, person-times $N_1/N_0 = e^{-m}$ and 1. Allows skewed priors but requires prior indicators.
2) Translate each prior into a single record with $A = S^2/v$, person-time $N=A/e^{m/S}$, $S \geq 30$. Suitable for normal priors and produces much simpler data.

Example translation of prior into data table:

Our normal prior for β_T has variance = v = 0.16,

mean = mode = m = logit(.1) = ln(1/9) = −ln(9),

so expit$\{-\ln(9) \pm 1.96(.16)^{1/2}\}$ = .048, .196 ≈ .05, .20

are the 95% prior limits for π_{100} = expit(β_T).

- Using $A_1 = A_0 = A = 2/v = 2/0.16 = 12.5$ gives the above interval 94.5% exact {F(25,25)} prior probability; adding ½ to get A=13 gives it 94.9% exact {F(26,26)} prior probability.
- Person-time ratio: $N_1/N_0 = e^{-m} = \exp(\ln(9)) = 9$

Prior counts and person times for method using two records per prior

$N_1/N_0 = \exp(-\text{prior mode}) = e^{-m}$, $A = 2/v$

	N_1/N_0	A	(A+½)	95% prior limits
β_T	.9/.1=9	12.5	(13)	expit(β_T): .05,.20
β_{TX}	1/13.5	8	(8.5)	exp(β_{TX}): 5, 36
β_{TY}	0	4	(4.5)	exp(β_{TY}): ¼, 4
β_{TXY}	0	16	(16.5)	exp(β_{TXY}): ½, 2

For skewed priors ($A_1 \neq A_0$), use A_1 for first record and A_0 for second record.

3) With 2-record method, must add the 4 prior indicators: $Prior_T$, $Prior_{TX}$, $Prior_{TY}$, $Prior_{TXY}$

4) Do loglinear Poisson regression of the counts (actual and prior) on the covariates with the N = person-time, using ML for missing data, or multiple imputation with many imputations.

5) Collapse the fitted TXY counts for the actual data over X, and compute OR_{TY}

The resulting "confidence interval" for OR_{TY} will be the approximate posterior interval.

The following tables show the augmented data for each method; "." is missing value.

Data for 2-record method (Const not shown) Prior indicators

A	N	T	X	Y	TX	TY	XY	TXY	P_T	P_{TX}	P_{TY}	P_{TXY}
173	1	.	1	1	.	.	1	.	0	0	0	0
602	1	.	0	1	0	.	0	0	0	0	0	0
134	1	.	1	0	.	0	0	0	0	0	0	0
663	1	.	0	0	0	0	0	0	0	0	0	0
13	9	1	0	0	0	0	0	0	1	0	0	0
13	1	0	0	0	0	0	0	0	1	0	0	0
8.5	1	0	0	0	1	0	0	0	0	1	0	0
8.5	13.5	0	0	0	0	0	0	0	0	1	0	0
4.5	1	0	0	0	0	1	0	0	0	0	1	0
4.5	1	0	0	0	0	0	0	0	0	0	1	0
16.5	1	0	0	0	0	0	0	1	0	0	0	1
16.5	1	0	0	0	0	0	0	0	0	0	0	1

Augmented data, offset (one-record) method with rescaling by S = 30 (. is missing value):

A	N	Const	T	X	Y	TX	TY	XY	TXY
173	1	1	.	1	1	.	.	1	.
602	1	1	.	0	1	0	.	0	0
134	1	1	.	1	0	.	0	0	0
663	1	1	.	0	0	0	0	0	0
5625	6052	0	1/30	0	0	0	0	0	0
3600	3301	0	0	0	0	1/30	0	0	0
1800	1800	0	0	0	0	0	1/30	0	0
7200	7200	0	0	0	0	0	0	0	1/30

For prior records, $A = S^2/v$,

$N = A/\exp(m/S) = S^2/\{v \bullet \exp(m/S)\}$ with S=30.

Poisson regression of A on listed covariates:

Posterior median for OR_{TY} = 1.19, 95% Wald limits 0.41, 3.4 (Monte-Carlo limits 0.37, 3.4)

- Fairly **insensitive** to reasonable choices for prior variances of β_T, β_{TX}, β_{TXY}.
- But, very **sensitive** to prior for β_{TY}, e.g., changing prior 95% limits for $\exp(\beta_{TY})$ to 1/8 and 8, the posterior 95% Wald limits for OR_{TY} become 0.25, 5.67

Absent precise information on T|X,Y, the data add little information about OR_{TY}

- The expanded-model results show that the precision of the conventional frequentist results (1.42, 95% limits 1.11, 1.83) **and** Bayesian results (1.41, 95% limits 1.10, 1.80) is due entirely to the absurdly precise (point-prior) assumptions these make about the predictive values π_{txy} (that these are 1 when T=X and 0 when T≠X).

A caution: Familiar asymptotics can fail spectacularly with nonidentified models

We have come to rely on summaries that apply only to strongly identified parameters (e.g., associations among direct measurements, adjusted for a few variables). Examples:

- Point estimates that are solutions of estimating equations (ML, GEE, IPTW, Doubly Robust),
- Wald limits: point estimate ± 1.96•SE

Without identification, these can be grossly inadequate summaries of location and uncertainty.

Summary

- Conventional methods at best produce only data descriptions and often result in garbage inferences (very biased and extremely overconfident).

Nonidentified-bias modeling

- frees us from the ludicrous assumptions implicit in conventional methods
- is essential for realistic statistical inference in observational epidemiology.

Sensitivity analysis

- Helps by allowing us to consider nonidentified models

But

- Becomes almost unintelligible beyond 3 dimensions (at least 4 dimensions are needed for most misclassification analysis)
- Generates no inference without a prior, and so raises the danger of claims based on very strong yet implicit (hidden) priors.

Issues of bias can be reduced to issues of missing data…

…hence inferences hinge on priors about relations (π_{txy}) of missing to observed data.

- The missing-data formulation leads to methods for bias analysis with standard missing-data algorithms.
- In tandem, translating priors into prior data helps gauge the strength of prior opinions about parameters.

Probabilistic (MCSA, Bayesian, etc.) analyses of nonidentified models…

- Summarize over a sensitivity analysis using explicit prior distributions
- Produce opinions derived explicitly from the entire set of priors (all data models and all parameter priors)
- **Will often reveal that the data add little about the target parameter beyond what we were already willing to assert**.