Bayesian Statistical Analysis in Medical Research

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu

www.ams.ucsc.edu/~draper

ROLE Steering Committee Meeting
New York NY

25 April 2007

The Big Picture

Statistics is **the study of uncertainty: how to measure it,** and **what to do about it.**

How to **measure** uncertainty: **probability**; two main **probability paradigms**: **frequentist** and **Bayesian**.

What to do about uncertainty: two main activities —

- **Inference**: **Generalizing outward** from a given data set (**sample**) to a larger universe (**population**), and attaching **well-calibrated** measures of **uncertainty** to the generalizations (e.g., "**Nonwhites** in the population of people at substantial risk of **HIV–1** infection are **88%** more likely to get infected if they don't receive this **rgp120 vaccine** than if they do receive it (**relative risk** of infection **1.88**, **95%** interval estimate **1.14–3.13**)").

- **Decision-Making**: Taking or recommending an **action** on the basis of available data, in spite of remaining uncertainties (e.g., "Based on this trial, for whom nonwhites were a **secondary subgroup**, it's recommended that the vaccine be studied further with nonwhites as the **primary study group**").

Frequentist Probability

• **Frequentist** probability: Restrict attention to phenomena that are **inherently repeatable** under (essentially) **identical conditions**; then, for an event A of interest, $P_F(A)$ = limiting **relative frequency** with which A occurs in the (hypothetical) repetitions, as number of repetitions $n \to \infty$.

— **Pros:** Math easier; natural approach to defining **calibration** of probability statements (e.g., "In hypothetical **repetitions** of this data-gathering activity, about **95%** of the time this method for creating **95%** **confidence intervals** for a population summary θ will generate an interval that does indeed include θ").

— **Cons:** Only applies to **inherently repeatable processes** (e.g., can't talk about $P_F(1.14 \leq \theta_{RR} \leq 3.13)$ [θ_{RR} = population relative risk] or P_F(this patient is HIV positive); **hard** to **quantify** many **relevant** sources of uncertainty (e.g., uncertainty about the **"right" statistical model** for the data); harder to **combine information** from **two or more information sources**.

NB Looks **"objective"** (call $P(A)$ **objective** if reasonable people more or less agree on its value), but in general **is not**; more on this point below.

Bayesian Probability

• **Bayesian** probability: numerical **weight of evidence** in favor of an uncertain proposition, obeying a series of **reasonable axioms** to ensure that Bayesian probabilities are **coherent** **(internally logically consistent)**.

One approach: Given your background **knowledge**, **assumptions** and **judgments** $\mathcal{B}$, imagine **betting** with someone about the truth value of some true/false proposition A, and ask yourself what **odds** $O_{\mathcal{B}}$ you would need to give or receive so that you judge the bet to be **fair**; then for you $P_B(A|\mathcal{B}) = \frac{O_{\mathcal{B}}}{1+O_{\mathcal{B}}}$.

— **Pros:** The **most general definition** of probability so far developed (applies to **any** processes, repeatable or not); **Theorem** (details below): **All optimal decisions are Bayesian**.

— **Cons:** Math harder; **good calibration not guaranteed**.

NB Clearly **"subjective"** (**different background knowledge** or different **assumptions** and **judgments** about how that knowledge bears on the question → potentially **different** Bayesian probabilities $P_B(A|\mathcal{B})$ for the same proposition A), but this is actually a **positive feature** of the paradigm.

Frequentist Statistics

Frequentist inference: (1) Think of your **data** as **like** a **random sample** from some **population**. (2) Identify some **numerical summary** θ of the population of interest, and find a reasonable **estimate** $\hat{\theta}$ of θ based on the sample. (3) **Imagine repeating** the **random sampling**, and use the **random behavior** of $\hat{\theta}$ across these **repetitions** to make **probability statements** involving θ (e.g., **confidence intervals** for θ [e.g., "I'm 95% **confident** that θ_{RR} is between 1.14 and 3.13"] or **hypothesis tests** about θ [e.g., the P **value** for testing H_0: $\theta_{RR} < 1$ against H_A: $\theta_{RR} \geq 1$ is 0.012, so I **reject** H_0).

NB **Not possible** in problems of **realistic complexity** to **avoid making assumptions** and **judgments** (e.g., with **observational** data (**not randomly sampled** from the population of greatest scientific interest), **reasonable** people may **differ** in their answers to the important question **"What is the broadest scope of valid generalizability outward from this data set?"** [i.e., what's the **largest population** from which it's reasonable to think of these data as **like a random sample**?]); this applies to **all inferential work** in statistics, frequentist or Bayesian, so in general **"objectivity" is not attainable**.

Frequentist Statistics (continued)

Frequentist decision-making: (1) Specify a collection of **decision rules**, which attach a **decision** $\delta(y)$ to each possible **hypothetical data set** y. (2) Specify a **loss function** $l[\delta(y), \theta_0]$ that quantifies how much is **lost** if decision $\delta(y)$ is taken when the **true value** of the **unknown** θ is θ_0. (3) **Evaluate** the available decision rules in terms of their average loss with respect to **all possible data sets** that **might arise** in **repeated random sampling**.

Decision-making has been **de-emphasized** in **frequentist statistics** since Wald (1950) proved a famous **theorem** that says, informally, that **all good decisions are Bayesian** and **all Bayesian decisions are good**.

— **Frequentist pros:** **Implementation** easier; since based on frequentist probability, **straightforward** to address the important scientific question "With this statistical method, **how often do I get the right answer?**"

— **Frequentist cons:** Can only make **legitimate** probability statements about $\hat{\theta}$, not θ (e.g., OK to talk about $P_F(1.14 \leq \hat{\theta}_{RR} \leq 3.13)$ but not $P_F(1.14 \leq \theta_{RR} \leq 3.13)$ or $P_F(H_0 \text{ is true}) = P_F(\theta_{RR} < 1)$, so **inferential statements** about the world are **indirect** at best.

Bayesian Statistics

Bayesian inference: (1) In the **Bayesian paradigm**, **information** about **numerical unknowns** θ is **quantified** with **probability distributions** (e.g., an **information source** equivalent to the statement "θ is **probably** between **35** and **65**, and is **almost certainly** between **0** and **100**" might well be quantified with a **bell curve** (**normal** or **Gaussian** distribution) $p(\theta)$ with **mean 50** and **standard deviation 15**). (2) Given a **data set** y, **quantify information** about θ **external** to y in a **prior distribution** $p(\theta)$, and **quantify information** about θ **internal** to y in a **likelihood distribution** $l(\theta|y)$. (3) Then use **Bayes' Theorem** to compute the **combined information** about θ both **internal** and **external** to y, which is contained in the **posterior distribution** $p(\theta|y)$:

$$p(\theta|y) = c\,p(\theta)\,l(\theta|y); \tag{1}$$

here c is a **constant** chosen to make everything **add up to 1** (as all probabilities must do to be **coherent**).

NB In the **Bayesian paradigm** θ and y can be **almost anything** (**numbers, vectors, matrices, images, movies, phylogenetic trees**, ...).

Bayesian Statistics (continued)

NB **Assumptions** and **judgments** are also **central** to this approach to **inferential statistics** (e.g., **specification** of the **likelihood distribution** is related to the issue mentioned above about **thinking** of the data set y as **like** a **random sample** from a **population**).

Prediction is a special case of **inference** in which the unknowns of interest are **observable quantities** (e.g., the **HIV viral load** for **this** nonwhite person if **not treated** with the **vaccine** [e.g., think of a **lognormal distribution** centered at **6,000** HIV copies per ml]) — creating **predictive distributions** that fully capture **all relevant sources of uncertainty** is **easy** to do in the **Bayesian** paradigm and **hard** to do in **frequentist** statistics.

Bayesian decision-making: Given an **unknown** θ and a **data set** y, (1) Specify a set of **possible actions** $\{a_1, a_2, \dots\}$. (2) Specify a **utility function** $U(a, \theta_0)$ that **quantifies** how much is **gained** if action a is taken when the **true value** of the **unknown** θ is θ_0. (3) Find the **action** that **maximizes** the **average (expected) utility** $E[U(a, \theta)]$, where the **average (expectation)** is taken over the **posterior uncertainty** about θ, **quantified** via the **posterior distribution** $p(\theta|y)$ from **Bayes' Theorem** (equation (1) above).

Bayesian Pros

(1) **Inference** is more **direct**: e.g., **no need to use** P **values**, since OK to talk about $P_B(H_0$ is true); in fact, **hypothesis testing decreases** and **interval estimation increases** in **importance**, which is a **good scientific outcome** (e.g., if the **effectiveness** of a **new treatment** for **hypertension** is at issue, it's **far more relevant** to make statements on the **scale of the data** [e.g., "With **95% (posterior) probability**, the **population mean decrease** θ in **systolic blood pressure** over a **3–month period** with **new drug** ND versus the **previous best drug** PB is between **5** and **20 mmHg**"] than on the **probability scale** [e.g., "The P **value** for testing H_0: $\theta < 0$ against H_A: $\theta \geq 0$ is **0.02**"]).

(2) **Decisions** are **optimal** (this is **important** because **many problems** that are **traditionally formulated** in **inferential** terms are really **decisions** [more on this below]).

(3) **Prediction** of **observables**, which is at the **heart** of both **good science** and **good statistical model-checking** (**good models** make **good predictions**, **bad models** make **bad predictions**), is **easy**.

Bayesian Pros and Cons

• More **Bayesian pros**:

(4) **Combining information** from **all relevant sources** (e.g., **meta-analysis** of many studies, through **Bayesian random-effects hierarchical models**) is **straightforward**.

(5) **Bayesian procedures** often have **better repeated-sampling properties** than **freqentist methods** in **complicated models** (e.g., Browne and Draper (2007): in **simulations** involving **random-effects logistic regression models** in **medical** settings, **frequentist "95%" confidence intervals** for **variance components** included **true values 0–60%** of the time but **Bayesian 95% intervals** included the **truth 93%** of the time).

• **Bayesian cons:**

(1) **Implementation harder** than with the **frequentist** approach, but **Markov chain Monte Carlo (MCMC) simulation-based methods** have now made **fitting Bayesian models straightforward** for a **wide variety** of **complicated data structures** and **experimental designs** in **medical research**.

Bayesian-Frequentist Fusion

• More **Bayesian cons**:

(2) **Good calibration is not guaranteed** with the **Bayesian** approach if **strong prior information** that's **retrospectively** seen to be **out of step with the world** is used: **prior information** in **Bayesian** work is **equivalent** to a **prior data set** with n_p observations, which is **combined** with a **sample data set** with n_s observations, and if n_p is chosen **too big** in relation to n_s and the **prior information** is **off-target**, the result will be **worse** than just working with the **sample data set**.

Since **monitoring calibration** is essentially a **frequentist** activity, this suggests looking for a way to **combine** the **Bayesian** and **frequentist** approaches.

I believe that my **job** as a **statistician** is **not** to **choose** one of the two **probability paradigms** and **defend** it against **attacks** from people who **favor the other one** (this was the **philosophical stand** taken by **many statisticians** in the **20th century**) but to find a **fusion** of the two approaches that **maximizes** the **strengths** of the fusion and **minimizes** the **weaknesses**.

Bayesian-Frequentist Fusion (continued)

For me the **best fusion** is to

(a) **reason in a Bayesian way** when **formulating** my **inferences** and **decisions** (because the **Bayesian paradigm** is the **most flexible approach** so far invented for **quantifying all relevant sources of uncertainty** in **complicated problems** and **making choices** in the face of such uncertainty), and

(b) **reason in a frequentist way** when **evaluating** the **quality** of these **inferences** and **decisions**, by keeping track of the **calibration** of my **Bayesian methods** (I do this by constructing **predictive distributions** for **observables** and **comparing** those distributions with the **actual observed values**).

In this way **Bayesian coherence** keeps me **internally free of logical contradiction** and **frequentist calibration** keeps me **honest** (in **good touch** with the **external world**).

I believe (and others agree) that **statistics in the 21st century** will be **dominated** by a **Bayesian-frequentist fusion** rather like this one.

Practical Details

• A **good book** on **Bayesian methods** in **medical research** is

Spiegelhalter DJ, Abrams KR, Myles JP (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley.

• The U.S. **Food and Drug Administration** (FDA) has wholeheartedly embraced **Bayesian methods** in the **design** and **analysis** of **clinical trials** for **medical devices** — for example, in May 2006 the FDA Center for Devices and Radiological Health issued a document called *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials – Draft Guidance for Industry and FDA Staff*, which is available at

`www.fda.gov/cdrh/osb/guidance/1601.html`.

• `SAS` has incorporated **Bayesian methods** in some of its procedures — you can now do **Bayesian inference** via **MCMC** for **generalized linear models** (`bgenmod`), **parametric survival models** (`blifereg`) and **semiparametric survival models** (`bphreg`); for details see

`www.sas.com/apps/demosdownloads/setupcat.jsp?cat=SAS%2FSTAT+Software`

Example: HIV Vaccine Efficacy

I'm currently working with colleagues at UCSC on a **Bayesian re-analysis** of data from a **randomized controlled trial** of an **rpg120 vaccine** against **HIV** (rgp120 HIV Vaccine Study Group (2005). Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV–1 infection. *Journal of Infectious Diseases*, **191**, 654–663).

5403 healthy HIV-negative volunteers at high risk of getting HIV were **randomized**, **3598** to the **vaccine** and **1805** to **placebo** (in both cases, 7 injections over 30 months), and followed for **36 months**; the **main outcome** was presence or absence of **HIV infection** at the end of the trial, with **Vaccine Efficacy** (VE) defined as 100(1 − relative risk of infection); **secondary frequentist analyses** examined **differences** in VE by **gender, ethnicity, age**, and **education** and **behavioral risk score** at baseline.

The trial found a **small decline** in infection overall (**6.7% vaccine, 7.0% placebo**) that was **neither practically nor statistically significant**; **large preventive effects** of the **vaccine** were found for some **subgroups** (e.g., **nonwhites**), but **statistical significance vanished** after adjustment for **multiple comparisons**.

Example (continued)

As mentioned on page 2, the relative risk of infection $RR = \frac{p_{\text{placebo}}}{p_{\text{vaccine}}} = \frac{100}{100-VE}$ was 1.88 (95% confidence interval, 1.14–3.13).

Category, parameter	Rate of HIV-1 infection		VE (95% CI)	P	
	Vaccine	Placebo		Unadjusted[a]	Adjusted[b]
All volunteers	241/3598 (6.7)	127/1805 (7.0)	6 (−17 to 24)	.59	>.5
Men	239/3391 (7.0)	123/1704 (7.2)	4 (−20 to 23)	.73	>.5
Women	2/207 (1.0)	4/101 (4.0)	74 (−42 to 95)	.093	.41
Race					
White (non-Hispanic)	211/2994 (7.0)	98/1495 (6.6)	−6 (−35 to 16)	.60	>.5
Men	211/2930 (7.2)	98/1468 (6.7)	−6 (−35 to 16)	.61	...
Women	0/64 (0)	0/27 (0)	...	...	...
Hispanic	14/239 (5.9)	9/128 (7.0)	15 (−96 to 63)	.70	>.5
Men	13/211 (6.2)	9/114 (7.9)	20 (−88 to 66)	.61	...
Women	1/28 (3.6)	0/14 (0)	...	...	...
Black (non-Hispanic)	6/233 (2.6)	9/116 (7.8)	67 (6 to 88)	.028	.24
Men	5/121 (4.1)	5/59 (8.5)	54 (−61 to 87)	.21	...
Women[c]	1/112 (0.9)	4/57 (7.0)	87 (−19 to 98)	.033	...
Asian (all men)	3/56 (5.4)	3/21 (14.3)	66 (−70 to 93)	.17	>.5
Other	7/76 (9.2)	8/45 (17.8)	50 (−39 to 82)	.18	>.5
Men	7/73 (9.6)	8/42 (19.0)	51 (−34 to 82)	.16	...
→ Nonwhite	30/604 (5.0)	29/310 (9.4)	47 (12 to 68)	.012	.13 ←
Men	28/461 (6.1)	25/236 (10.6)	43 (3 to 67)	.036	...
Women	2/143 (1.4)	4/74 (5.4)	74 (−43 to 95)	.10	...
Age					
<30 years	84/971 (8.7)	43/504 (8.5)	−1 (−46 to 30)	.95	>.5
>30 years	157/2627 (6.0)	84/1301 (6.5)	8 (−19 to 30)	.61	>.5
Education level[d]					
Less than a college degree	95/1409 (6.7)	52/713 (7.3)	8 (−29 to 34)	.63	>.5
College or graduate degree	146/2188 (6.7)	75/1092 (6.9)	4 (−27 to 27)	.77	>.5
Baseline behavioral risk score[e]					
Low risk	32/1211 (2.6)	11/609 (1.8)	−48 (−193 to 26)	.26	>.5
Medium risk	177/2229 (7.9)	90/1107 (8.1)	3 (−25 to 25)	.92	>.5
High risk	32/158 (20.3)	26/89 (29.2)	43 (4 to 66)	.032	.29

Example (continued)

Note that the P **value** for the **nonwhite subgroup** was **0.012 before**, but **0.13 after, multiple comparisons adjustment**.

However, **frequentist multiple comparisons methods** are an **inferential approach** to what should really be a **decision problem** (**Should** this **vaccine** be given to **nonwhite** people at high risk of getting HIV? **Should another trial** focusing on **nonwhites** be run?), and when **multiple comparison methods** are viewed as **"solutions"** to a **Bayesian decision problem** they **do not have a sensible implied utility structure**: they're **terrified** of **announcing that an effect is real when it's not** (a **"type I error"**), and have **no built-in penalty** for **failing to announce an effect is real when it is** (a **"type II error"**).

In the **frequentist** approach, **type II errors** are supposed to be **taken care of** by having done a **power calculation** at the time the **experiment** was **designed**, but this **begs the question** of **what decision should be taken, now that this study has been run**, about whether to **run a new trial** and/or **give the vaccine to nonwhite people now**.

Example (continued)

When the problem is **reformulated** as a **decision** that properly **weighs all of the real-world costs and benefits**, the **result** (interpreted in **frequentist** language) would be a **third** P **value column** in the table on page 15 (a column called **"Implied** P **from a decision-making perspective"**) that would look **a lot more like the first (unadjusted)** P **value column than the second (multiple-comparisons adjusted) column**, leading to the **decision** that a **new trial for nonwhites for this vaccine** is a **good investment**.

This can be seen in an **even simpler setting**: consider a **randomized controlled trial** with **no subgroup analysis**, and define Δ to be the **population mean health improvement** from the **treatment** T as compared with the **control condition**.

There will typically be **some point** c **along the number line** (a kind of **practical significance threshold**), which **may not be 0**, such that if $\Delta \geq c$ the **treatment** should be **implemented** (note that this is a decision problem).

The **frequentist hypothesis-testing inferential approach** to this problem would test H_0: $\Delta < c$ against H_0: $\Delta \geq c$, with **(reject** H_0**)** corresponding to the action $a_1 =$ **{implement** T**}**.