

**Incorporating Scientific  
Knowledge into Data  
Analysis:**

**Sometimes it's Necessary?**

**Wesley O. Johnson  
Department of Statistics  
UC Davis**

# Fundamental Ideas

- Bayesian statistical analysis is based on the premise that *all uncertainty should be modeled with probability* and that *statistical inferences should be logical conclusions based on the laws of probability*.
- This typically involves the explicit use of subjective information provided by the scientist, since initial uncertainty about unknown parameters must be modeled from *a priori* expert opinions. Bayesian methodology is consistent with the goals of science.
- For large amounts of data, scientists with different subjective prior beliefs will ultimately agree after (separately) incorporating the data with their “prior” information.

- On the other hand, “insufficient” data can result in (continued) discrepancies of opinion about the relevant scientific questions.
- Bayesian analysis appears to be the only logically consistent method of making statistical inferences, but not the only useful one.
- We believe that the best statistical analysis of data involves a collaborative effort between subject matter scientists and statisticians, *and*
- that it is both appropriate and necessary to incorporate the scientist’s expertise into making decisions related to the data.

## Simple Probability Computations Provide the Basis for Bayesian Statistical Inference

### **Example: Drug Screening**

- Let  $D$  indicate a drug user and  $C$  indicate someone who is clean of drugs. Let  $(+)$  indicate that someone tests positive on a drug test, and  $(-)$  indicates testing negative.
- The overall *prevalence* of drug use in the population is, say,  $\pi = \Pr(D)$  and an expert believes that drug use in this population is relatively rare; they are 95% sure that the prevalence is less than 0.25.

- The *sensitivity*,  $Se = \Pr(+|D)$ , of the drug test is unknown, but an expert's best guess is 0.98 and they are 95% certain that the sensitivity is at least 0.95.
- The *specificity*,  $Sp = \Pr(-|C)$  is also unknown but the expert guess is 0.95, and they are 95% certain that it is at least 0.90.
- 100 individuals have been sampled from the population and screened for drugs.

$$T^+ = 10 \text{ tested (+).}$$

- It is of interest to make inferences about

$$\Pr(D|+) = \frac{\pi Se}{\pi Se + (1 - \pi)(1 - Sp)}$$

and

$$\Pr(D|-) = \frac{\pi(1 - Se)}{\pi(1 - Se) + (1 - \pi)Sp}$$

- Frequentist analysis would involve setting  $Se = 0.98$ ,  $Sp = 0.95$ , ignoring prior information about the prevalence and uncertainty in these estimates

- Solve  $\hat{\pi} * 0.98 + (1 - \hat{\pi}) * 0.05 = 0.10$

$$\hat{\pi} = 0.054$$

- If  $T^+ \leq 5$  had been observed,  $\hat{\pi} \leq 0$

- Large sample CI (assuming  $Se$  and  $Sp$  are truly known) is

$$0.054 \pm 0.063$$

```
model{ x ~dbin(p,100)
p <- pi*eta + (1-pi)*(1-theta)
eta ~ dbeta(151.8,4.08)
theta ~ dbeta(99.7,6.2)
pi ~ dbeta(1,10.41)
PrDpos <- pi*eta/p
PrDneg <- pi*(1-eta)/(1-p)}
list(x = 10)
list(pi =.05,eta=.98,theta = .95)
```

node	mean	sd	2.5%	97.5%
Pr(D neg)	0.0015	0.0013	$6.8/10^5$	0.0049
Pr(D pos)	0.437	0.212	0.0381	0.808
Se	0.974	0.013	0.943	0.992
Prev	0.049	0.032	0.0032	0.121
Sp	0.942	0.02	0.899	0.977

## Example: Analysis of Trauma Data (Bedrick, Christensen and Johnson, 1997)

- Data on a random subset of 300 patients admitted to The University of New Mexico Trauma Center between the years 1991 and 1994.
- For each patient we have their injury severity score (ISS), their revised trauma score (RTS), their age (AGE), the predominant type of injury (TI), that is, whether it was blunt ( $TI = 0$ ), e.g., the result of a car crash, or penetrating ( $TI = 1$ ), e.g., gunshot wounds, and whether the patient eventually survived.

- The ISS is an overall index of a patient's injuries and can take on values from 0 for a patient with no injuries to 75 for a patient with severe injuries in three or more body areas
- The RTS is an index of physiologic injury, and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate, and Glasgow Coma Scale; it takes on values from 0 for a patient with no vital signs to 7.84 for a patient with normal vital signs.
- These data were provided by Dr. Turner Osler, a trauma surgeon at the University of Vermont
- We use a logistic regression model proposed by Dr. Osler to estimate the probability of a patient's death using an intercept, predictors ISS, RTS, AGE, and TI along with an interaction between AGE and TI.

- Dr. Osler's expert opinions formed the basis for our prior
- To induce a proper prior distribution on the 6 dimensional vector  $\beta$ , we require a joint distribution on death probabilities for 6 sets of conditions
- Based on discussions with our expert and plots of the data, we defined a  $2^4$  factorial having ISS at levels 25 and 41, RTS at levels 3.34 and 7.84, AGE at levels 10 and 60, TI at levels 0 and 1.
- The idea was to pick values of the variables that were relatively extreme within the data but still had substantial probabilities for both death and survival.
- The prior conditions were chosen as a  $1/4$  replicate of this  $2^4$  with two center points.

- Our elicitation from Dr. Osler involved obtaining first, fiftieth and 99th percentiles for each specified probability

<i>i</i>	Design for prior						Beta ( $a_i, b_i$ )	
	ISS	RTS	AGE	TI	INT	$a_i$	$b_i$	
1	1	25	7.84	60	0	0	1.1	8.5
2	1	25	3.34	10	0	0	3.0	11.0
3	1	41	3.34	60	1	60	5.9	1.7
4	1	41	7.84	10	1	10	1.3	12.0
5	1	33	5.74	35	0	0	1.1	4.9
6	1	33	5.74	35	1	35	1.5	5.5

Prior Specification

- The first probability,  $\tilde{p}_1$ , corresponds to an individual that “has good physiology, is ‘not bad hurt’, does not have a lot of reserve,” and for whom there is “added uncertainty due to age.” The median is 0.09.
- The second type of individual “has bad physiology, is very ill, but is young and resilient and is not so badly hurt.” The median is 0.20
- The third individual has “bad physiology, a pretty bad injury, and there is much more

uncertainty here due to the age factor.”  
The median is 0.80.

- Individual 4 “is young, resilient and has a big injury.” The median is 0.07.
- Dr. Osler had more difficulty with the 5th and 6th types of individuals because their conditions were both less extreme and more related than those already considered. The medians were 0.15 and 0.19, respectively, and were highly dispersed.