Contrast Between Frequentist and Bayesian Statistics

- Frequentist statistics: Uncertainty about, for example, parameter estimates is quantified by investigating how such estimates would vary one to the next in repeated sampling from the same population.

- Bayesian statistics: Uncertainty is quantified by determining how much *prior* opinions about parameter values change in light of the observed data.

To a Bayesian, data sets which might have been, but were not, observed are irrelevant to making inferences about the unknown parameters. *The only data set of any relevance is the one that was actually observed.*

In contrast, in the frequentist approach, data sets which might have been observed (but were not) are extremely relevant since they are the basis of determining measures of uncertainty.

The term *Bayesian* derives from Thomas Bayes (1702-1761), who was a British mathematician and Presbyterian minister. Bayes introduced *Bayes' theorem*, which is the starting point of our study.

Bayes' theorem Let $A_1, \ldots, A_m$ be mutually exclusive and exhaustive events. (Exhaustive means $A_1 \cup \cdots \cup A_m = \mathcal{S}$, where $\mathcal{S}$ is the sample space.) For any event B such that $P(B) > 0$,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}, \; j = 1, \ldots, m.$$

Proof By the definition of conditional probability,

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)}. \qquad (1)$$

Again by the definition of conditional probability,

$$P(A_j \cap B) = P(B|A_j)P(A_j). \tag{2}$$

Also,

$$\begin{aligned} P(B) &= P(B \cap \mathcal{S}) \\ &= P[B \cap (A_1 \cup \cdots \cup A_m)] \\ &= P[(A_1 \cap B) \cup \cdots \cup (A_m \cap B)] \\ &= \sum_{i=1}^{m} P(A_i \cap B) \\ &= \sum_{i=1}^{m} P(B|A_i)P(A_i). \end{aligned} \tag{3}$$

Substituting (2) and (3) into (1) proves the result.

By considering problems to which Bayes theorem has been applied, we can already see part of the reason why some statisticians, scientists and philosophers find Bayesian methodology controversial.

Bayes' theorem is often used *when all aspects of an experiment's outcome have been determined, but the experimenter is uncertain about some of these aspects.*

Examples:

- A polygraph is given to someone charged with a crime. The polygraph indicates that the person is lying on a given question. What is the probability that the person actually lied?

- A person tests positive for AIDs. What is the probability that the person actually has AIDs?

A frequentist might argue that it isn't appropriate to apply probability to such situations. At best, they might say that the probability is either 0 or 1 that a person who tested positive for AIDs actually has the disease. After all, he/she either has it or not.

> To a frequentist, *probability is to be applied to experiments whose outcome has not yet been determined.*

On the other hand, Bayes' theorem is a mathematical fact, and if the frequentist probabilities are available for computing $P(A_j|B)$, then why not use the theorem?

Example Suppose 5% of a given population is infected with AIDs, and that a certain AIDs test gives a positive result 4% of the time among patients who do not have AIDs and 98% of the time among patients who have AIDs. If a given person has tested positive, what is the probability that he/she actually has AIDs?

$$A_1 = \text{event person has AIDs}$$
$$B = \text{event of testing positive}$$

$$\begin{aligned}
P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_1^c)P(A_1^c)} \\
&= \frac{0.98(0.05)}{0.98(0.05) + 0.04(0.95)} \\
&= 0.563.
\end{aligned}$$

Bayes' Theorem Applied to Statistical Models

Suppose we are to observe a vector of data $\boldsymbol{Y}$, a realization of which is denoted $\boldsymbol{y}$. $\boldsymbol{Y}$ has a probability distribution depending upon an unknown vector of parameters $\boldsymbol{\theta}$, which we wish to infer.

We'll denote the probability distribution of $\boldsymbol{Y}$ by $f(\boldsymbol{y}|\boldsymbol{\theta})$, emphasizing that this is the distribution of $\boldsymbol{Y}$ *conditional on* $\boldsymbol{\theta}$ *being the true value of the unknown parameter vector*.

It's important to distinguish between the true (but unknown) value of $\boldsymbol{\theta}$, which we could call $\boldsymbol{\theta}_0$, and values of the parameter vector that are merely being entertained as possibilities for $\boldsymbol{\theta}_0$. Unless otherwise stated, $\boldsymbol{\theta}$ is a variable that denotes *possibilities* for the truth, and $\boldsymbol{\theta}_0$ is a constant that represents the truth.

The *prior* distribution of $\boldsymbol{\theta}$ is a probability distribution that represents the experimenter's opinion about the unknown parameters. This distribution will be denoted $\pi(\boldsymbol{\theta})$.

Suppose that $\boldsymbol{\theta}$ is a continuous vector and π is a probability density. Then π has the interpretation that

$$\int_A \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

represents the experimenter's degree of belief that $\boldsymbol{\theta}_0$ lies somewhere in the region A.

Bayes' theorem applied to statistical model

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{m(\boldsymbol{y})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{t})\pi(\boldsymbol{t})\, d\boldsymbol{t}},\end{aligned}$$

where Θ is the parameter space, i.e., the set of all possible values for $\boldsymbol{\theta}$.

Terminology

π: The *prior* distribution.

$f(\boldsymbol{y}|\boldsymbol{\theta})$: For given $\boldsymbol{y}$ and regarded as a function of $\boldsymbol{\theta}$, the *likelihood* function.

$\pi(\boldsymbol{\theta}|\boldsymbol{y})$: The *posterior* distribution.

m: The *marginal* distribution of $\boldsymbol{Y}$, or the *prior predictive* distribution.

The posterior distribution expresses the experimenter's updated beliefs about $\boldsymbol{\theta}$ in light of the observed data $\boldsymbol{y}$. The data may well change his opinions about the unknown parameters, and will usually sharpen them.

The marginal m is called the prior predictive distribution because (i) it is the unconditional distribution of $\boldsymbol{Y}$, and (ii) it may be used as an aid in predicting a value of $\boldsymbol{Y}$.

Let $\tilde{\boldsymbol{Y}}$ be a data vector that is yet to be observed. The *posterior predictive* distribution of $\tilde{\boldsymbol{Y}}$ given $\boldsymbol{Y}$ is

$$m(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \frac{\int_{\Theta} f(\boldsymbol{y}, \tilde{\boldsymbol{y}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{m(\boldsymbol{y})}.$$

This distribution could be used to predict a value of $\tilde{\boldsymbol{Y}}$ given that $\boldsymbol{Y}$ has been observed to be $\boldsymbol{y}$.

Subjectivity in Bayesian inference

A Bayesian analysis is *subjective* in that two different people may observe the same data $\boldsymbol{y}$ and yet arrive at different conclusions about $\boldsymbol{\theta}$. This can happen when the two people have different prior opinions about $\boldsymbol{\theta}$.

The subjectivity of the Bayesian paradigm is a source of controversy. A legitimate argument is that it seems "unscientific" for one's personal prejudices to affect the conclusions of a scientific study.

Counters to the subjectivity criticism

- In "unscientific" situations, it seems natural that one's conclusions will be affected by his/her prior opinions.

- When a large amount of data is available, the prior has little effect on the posterior, unless the prior is *extremely* sharp.

- A more objective approach in scientific situations is to use so-called *noninformative priors*. Such priors are meant to express ignorance about the unknown parameters. (We'll discuss noninformative priors much more throughout the course.)