Bayes' Law Example

- **Example**: (1975 British national referendum on whether the UK should remain part of the European Economic Community)
- Suppose 52% of voters supported the Labour Party and 48% the Conservative Party. Suppose 55% of Labour voters wanted the UK to remain part of the EEC and 85% of Conservative voters wanted this.
- What is the probability that a person voting "Yes" to remaining in EEC is a Labour voter?

$$P(L|Y) = \frac{P(Y|L)P(L)}{P(Y)}$$

Bayes' Law Example

Note
$P(Y) = P(Y, L) + P(Y, L^c) = P(Y|L)P(L) + P(Y|L^c)P(L^c)$.
So

$$\begin{aligned} P(L|Y) &= \frac{P(Y|L)P(L)}{P(Y|L)P(L) + P(Y|L^c)P(L^c)} \\ &= \frac{(.55)(.52)}{(.55)(.52) + (.85)(.48)} = 0.41. \end{aligned}$$

Bayes' Law with Multiple Events

Let $\mathbf{D}$ represent some observed data and let A, B, and C be mutually exclusive (and exhaustive) events conditional on $\mathbf{D}$. Note that

$$\begin{aligned} P(\mathbf{D}) &= P(A \cap \mathbf{D}) + P(B \cap \mathbf{D}) + P(C \cap \mathbf{D}) \\ &= P(\mathbf{D}|A)P(A) + P(\mathbf{D}|B)P(B) + P(\mathbf{D}|C)P(C). \end{aligned}$$

By Bayes' Law,

$$\begin{aligned} P(A|\mathbf{D}) &= \frac{P(\mathbf{D}|A)P(A)}{P(\mathbf{D})} \\ \Rightarrow P(A|\mathbf{D}) &= \frac{P(\mathbf{D}|A)P(A)}{P(\mathbf{D}|A)P(A) + P(\mathbf{D}|B)P(B) + P(\mathbf{D}|C)P(C)}. \end{aligned}$$

Bayes' Law with Multiple Events

- Denoting A, B, C by θ_1, θ_2, θ_3, we can write this more generally as

$$P(\theta_i|\mathbf{D}) = \frac{P(\theta_i)P(\mathbf{D}|\theta_i)}{\sum_{j=1}^{3} P(\theta_j)P(\mathbf{D}|\theta_j)}.$$

- If there are k distinct discrete outcomes $\theta_1, \ldots, \theta_k$, we have, for any $i \in \{1, \ldots, k\}$:

$$P(\theta_i|\mathbf{D}) = \frac{P(\theta_i)P(\mathbf{D}|\theta_i)}{\sum_{j=1}^{k} P(\theta_j)P(\mathbf{D}|\theta_j)},$$

- The denominator equals $P(\mathbf{D})$, the **marginal** distribution of the data.
- Note if the values of θ are portions of the continuous real line, the sum may be replaced by an integral.

Bayes' Law Example (4 Classes)

Example: In the 1996 General Social Survey, for males (age 30+):

- 11% of those in the lowest income quartile were college graduates.
- 19% of those in the second-lowest income quartile were college graduates.
- 31% of those in the third-lowest income quartile were college graduates.
- 53% of those in the highest income quartile were college graduates.

What is the probability that a college graduate falls in the lowest income quartile?

Bayes' Law Example (4 Classes)

$$P(Q_1|G) = \frac{P(G|Q_1)P(Q_1)}{\sum_{j=1}^4 P(G|Q_j)P(Q_j)}$$
$$= \frac{(.11)(.25)}{(.11)(.25) + (.19)(.25) + (.31)(.25) + (.53)(.25)} = 0.09.$$

Exercise: Find $P(Q_2|G)$, $P(Q_3|G)$, $P(Q_4|G)$ also. How does this **conditional** distribution differ from the **unconditional** distribution $\{P(Q_1), P(Q_2), P(Q_3), P(Q_4)\}$?

Statistics Using Bayes' Law

- We now consider inference about parameters, based on data.
- Generically denote an unobserved parameter of interest as θ.
- Generically denote our data as $\mathbf{D}$.
- Our probability model for the data, given a value of θ, is denoted $p(\mathbf{D}|\theta)$.
- Our model for our prior knowledge about θ is denoted $p(\theta)$.
- This could be highly specific or quite vague, depending how uncertain we are about θ.

Statistics Using Bayes' Law

- We seek to make probability statements about θ, **given** some observed data: $p(\theta|\mathbf{D})$.
- By Bayes' Law,

$$p(\theta|\mathbf{D}) = \frac{p(\theta)p(\mathbf{D}|\theta)}{p(\mathbf{D})}.$$

- Note $p(\mathbf{D})$ **does not** depend on θ and thus carries no information about θ.
- It is simply a **normalizing constant** which makes $p(\theta|\mathbf{D})$ sum (or integrate) to 1.

Statistics Using Bayes' Law

- For inference about θ, it is just as good to write

$$p(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta)$$

- The LHS is called the **posterior distribution** of θ and represents a compromise between the **prior** information about θ in $p(\theta)$ and the information from the sample about θ in $p(\mathbf{D}|\theta)$.
- Some useful **summaries** of the posterior are the **posterior mean**

$$E[\theta|\mathbf{D}] = \int \theta p(\theta|\mathbf{D})\, d\theta$$

Statistics Using Bayes' Law

and the **posterior variance**

$$
\begin{aligned}
var[\theta|\mathbf{D}] &= E\left\{(\theta - E[\theta|\mathbf{D}])^2|\mathbf{D}\right\} \\
&= \int (\theta - E[\theta|\mathbf{D}])^2 p(\theta|\mathbf{D})\, d\theta \\
&= \int \theta^2 p(\theta|\mathbf{D})\, d\theta - 2E[\theta|\mathbf{D}] \int \theta p(\theta|\mathbf{D})\, d\theta \\
&\quad + \left(E[\theta|\mathbf{D}]\right)^2 \int p(\theta|\mathbf{D})\, d\theta \\
&= E[\theta^2|\mathbf{D}] - \left(E[\theta|\mathbf{D}]\right)^2
\end{aligned}
$$

- If the values of θ are discrete, sums would replace the integrals.

CHAPTER 2 SLIDES START HERE

Some Notation

- **Notation**: We hereby denote our data as the $n \times k$ matrix $\mathbf{X}$.
- We denote the parameter(s) of interest (possibly multidimensional) to be the vector $\boldsymbol{\theta}$.
- We will denote our posterior distribution for $\boldsymbol{\theta}$ using $\pi(\cdot)$.

Likelihood Theory

- The likelihood function $L(\boldsymbol{\theta}|\mathbf{X})$ is a function of $\boldsymbol{\theta}$ that shows how "likely" are various parameter values $\boldsymbol{\theta}$ to have produced the data $\mathbf{X}$ that **were observed**.
- In classical statistics, the specific value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta}|\mathbf{X})$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$.
- In many common probability models, when the sample size n is large, $L(\boldsymbol{\theta}|\mathbf{X})$ is unimodal in $\boldsymbol{\theta}$.
- **Note**: Unlike $p(\boldsymbol{\theta}|\mathbf{X})$, $L(\boldsymbol{\theta}|\mathbf{X})$ does **not necessarily** obey the usual laws for probability distributions.
- Also, in the classical framework, all the randomness within $L(\boldsymbol{\theta}|\mathbf{X})$ is attached to $\mathbf{X}$, not to $\boldsymbol{\theta}$.

Likelihood Theory

- Mathematically, if the data $\mathbf{X}$ represent iid observations from probability distribution $p(\mathbf{X}|\boldsymbol{\theta})$, then

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{X}_i|\boldsymbol{\theta})$$

  (where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are the n data vectors).
- The **Likelihood Principle** of Birnbaum states that (given the data) all of the evidence about $\boldsymbol{\theta}$ is contained in the likelihood function.
- Likelihood Principle implies: Two experiments that yield equal (or proportional) likelihoods should produce equivalent inference about $\boldsymbol{\theta}$.

The Bayesian Framework

- Suppose we observe an iid sample of data $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$.
- Now $\mathbf{X}$ is considered fixed and known.
- We also **must** specify $p(\boldsymbol{\theta})$, the prior distribution for $\boldsymbol{\theta}$, based on any knowledge we have about $\boldsymbol{\theta}$ **before** observing the data.
- Our model for the distribution of the data will give us the likelihood

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{X}_i|\boldsymbol{\theta}).$$

The Bayesian Framework

- Then by Bayes' Law, our posterior distribution is

$$\pi(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{X})}{p(\mathbf{X})} = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{X})}{\int_{\Theta} p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{X})\, \mathrm{d}\boldsymbol{\theta}}$$

- Note that the **marginal distribution** of $\mathbf{X}$, $p(\mathbf{X})$, is simply the joint density $p(\boldsymbol{\theta}, \mathbf{X})$ (i.e., the numerator) with $\boldsymbol{\theta}$ integrated out.
- With respect to $\boldsymbol{\theta}$, it is simply a **normalizing constant** that ensures that $\pi(\boldsymbol{\theta}|\mathbf{X})$ integrates to 1.

The Bayesian Framework

- Since $p(\mathbf{X})$ carries no information about $\boldsymbol{\theta}$, for conciseness we may drop it and write

$$\pi(\boldsymbol{\theta}|\mathbf{X}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{X}).$$

- Often we can calculate the posterior distribution by multiplying the prior by the likelihood and **then** normalizing the posterior at the **last** step, by including the necessary constant.