de Finetti's Theorem for 1s and 0s

The judgment of exchangeability still seems to leave the joint distribution of the Y_i quite **imprecisely specified**.

After defining the concept of exchangeability, however, de Finetti went on to prove a **remarkable result**: if you're willing to regard the $\{Y_i, i = 1, \ldots, n\}$ as part (for instance, the beginning) of an **infinite** exchangeable sequence of 1s and 0s (meaning that every finite subsequence is exchangeable), then there's a simple way to characterize your joint distribution, if it's to be **coherent** (e.g., de Finetti, 1975; Bernardo and Smith, 1994).

(**Finite** versions of the theorem have since been proven, which say that the longer the exchangeable sequence into which you're willing to embed $\{Y_i, i = 1, \ldots, n\}$, the harder it becomes to achieve coherence with any probability specification that's far removed from the one below.)

de Finetti's Representation Theorem. If $Y_1, Y_2, \ldots$ is an **infinitely exchangeable** sequence of 0–1 random quantities with probability measure P, there exists a distribution function $Q(\theta)$ such that the joint distribution $p(y_1, \ldots, y_n)$ for $Y_1, \ldots, Y_n$ is of the form

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, dQ(\theta) \ , \quad (22)$$

where $Q(\theta) = \lim_{n\to\infty} P(\frac{1}{n}\sum_{i=1}^{n} Y_i \leq \theta)$ and $\theta \stackrel{P}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i$.

θ can also be interpreted as the **marginal probability** $P(Y_i = 1)$ that any of the Y_i is 1.

The Law of Total Probability

The distribution function Q will generally be well-behaved enough to have a **density**: $dQ(\theta) = p(\theta)d\theta$.

In this case **de Finetti's Theorem** says

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\, p(\theta)d\theta. \qquad (23)$$

Important digression. We saw in part 1 of the lecture notes that for the **true-false propositions** D and A,

$$\begin{aligned} P(D) &= P(D \text{ and } A) + P[D \text{ and (not } A)] \qquad (24) \\ &= P(A)\, P(D|A) + P(\text{not } A)\, P(D|\text{not } A). \end{aligned}$$

This is a special case of the
Law of Total Probability (LTP).

Notice that A and (not A) divide, or **partition**, the collection of all possible outcomes into two non-overlapping (**mutually exclusive**) and **exhaustive** possibilities.

Let $A_1, \ldots, A_k$ be any **finite partition**, i.e., $P(A_i \text{ and } A_j) = 0$ (mutually exclusive) and $\sum_{i=1}^{k} P(A_i) = 1$ (exhaustive).

Then a **more general** version of the LTP gives that

$$\begin{aligned} P(D) &= P(D \text{ and } A_1) + \ldots + P(D \text{ and } A_k) \\ &= P(A_1)\, P(D|A_1) + \ldots + P(A_k)\, P(D|A_k) \qquad (25) \\ &= \sum_{i=1}^{k} P(A_i)\, P(D|A_i). \end{aligned}$$

Hierarchical (Mixture) Modeling

There is a **continuous** version of the LTP: by analogy with (25), if X and Y are real-valued random variables

$$p(y) = \int_{-\infty}^{\infty} p(x)\, p(y|x)\, dx. \tag{26}$$

$p(x)$ in this expression is called a **mixing distribution**.

Intuitively (26) says that the overall probability behavior $p(y)$ of Y is a mixture (**weighted average**) of the conditional behavior $p(y|x)$ of Y given X, weighted by the behavior $p(x)$ of X.

Another way to put this is to say that you have a choice: you can either model the random behavior of Y **directly**, through $p(y)$, or **hierarchically**, by first modeling the random behavior of X, through $p(x)$, and then modeling the conditional behavior of Y given X, through $p(y|x)$.

Notice that X and Y are **completely general** in this discussion—in other words, given any quantity Y that you want to model stochastically, you're free to choose any X upon which Y depends and model Y **hierarchically** given X instead, if that's easier.

Symbolically

$$Y \quad \leftrightarrow \quad \left\{ \begin{array}{c} X \\ Y|X \end{array} \right\}. \tag{27}$$

The reason for bringing all of this up now is that (23) can be **interpreted** as follows, with θ playing the role of x:

$$\begin{aligned} p(y_1, \ldots, y_n) &= \int_0^1 p(y_1, \ldots, y_n|\theta)\, p(\theta)\, d\theta \\ &= \int_0^1 \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}\, p(\theta)\, d\theta. \end{aligned} \tag{28}$$

The Simplest Mixture Model

(28) implies that in any **coherent** expression of uncertainty about **exchangeable** binary quantities $Y_1, \ldots, Y_n$,

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}. \tag{29}$$

But (a) the left side of (29), interpreted as a function of θ for fixed $y = (y_1, \ldots, y_n)$, is recognizable as the **likelihood function** for θ given y, (b) the right side of (29) is recognizable as the likelihood function for θ in **IID Bernoulli sampling**, and (c) (29) says that these must be the **same**.

Thus, to summarize de Finetti's Theorem **intuitively**, the assumption of exchangeability in your uncertainty about binary observables $Y_1, \ldots, Y_n$ amounts to behaving **as if**

- there's a quantity called θ, interpretable as either the **long-run relative frequency of 1s** or the marginal probability that any of the Y_i is 1,

- you need to treat θ as a **random** quantity with density $p(\theta)$, and

- **conditional** on this θ the Y_i are IID B(θ).

In yet other words, for a Bayesian whose uncertainty about binary Y_i is exchangeable, the model may effectively be taken to have the simple **mixture** or **hierarchical** representation

$$\left\{ \begin{array}{ccc} \theta & \sim & p(\theta) \\ (Y_i|\theta) & \overset{\text{IID}}{\sim} & \text{B}(\theta), \; i = 1, \ldots, n \end{array} \right\}. \tag{30}$$

Conditional Independence

This is the **link** between frequentist and Bayesian modeling of binary outcomes: exchangeability implies that you should behave like a frequentist as far as the **likelihood function** is concerned (taking the Y_i to be IID Bernoulli(θ)), but a frequentist who treats θ as a random variable with a **mixing distribution** $p(\theta)$.

NB This is the first example of a general pattern:

$$Y_i \text{ exchangeable} \leftrightarrow \left\{ \begin{array}{c} Y_i \textbf{ conditionally IID} \\ \text{given one or more parameters} \end{array} \right\}. \quad (31)$$

So **exchangeability** is a special kind of **conditional independence**: binary exchangeable y_i are not independent, but they become conditionally independent given θ.

(30) is an example of the simplest kind of **hierarchical model (HM)**: a model at the top level for the underlying death rate θ, and then a model below that for the binary mortality indicators Y_i conditional on θ (this is a basic instance of (27): it's **not easy** to model the **predictive** distribution for $(Y_1, \ldots, Y_n)$ directly, but it becomes a lot easier when θ is introduced at the **top level of a 2–level hierarchy**).

To emphasize an important point mentioned above, to make sense of this in the Bayesian approach **you have to treat** θ **as a random variable**, even though logically it's a fixed unknown constant.

This is the main conceptual difference between the Bayesian and frequentist approaches: as a Bayesian you use the **machinery** of random variables to express your uncertainty about unknown quantities.

Approach	Fixed	Random
Frequentist	θ	Y
Bayesian	y	θ

Bayes' Theorem

2.4 Prior, posterior, and predictive distributions

What's the **meaning** of the mixing distribution $p(\theta)$?

$p(\theta)$ doesn't involve $y = (y_1, \ldots, y_n)$, so it must represent your information about θ before the data set y arrives—it makes sense to call it your **prior distribution** for θ.

I'll address how you might go about **specifying** this distribution below.

Q: If $p(\theta)$ represents your information about θ before the data arrive, what represents this information **after** y has been observed?

A: It has to be $p(\theta|y)$, the **conditional** distribution for θ given how y came out.

It's natural to call this the **posterior distribution** for θ given y.

Q: How do you get from $p(\theta)$ to $p(\theta|y)$, i.e., how do you **update** your information about θ in light of the data?

A: Bayes' Theorem for **continuous** quantities:

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{p(y)}. \tag{32}$$

This requires some interpreting. As a Bayesian I'm **conditioning on the data**, i.e., I'm thinking of the left-hand side of (32) as a function of θ for fixed y, so that must also be true of the right-hand side. Thus (a) $p(y)$ is just a constant—in fact, you can think of it as the **normalizing constant**, put into the equation to make the product $p(\theta)\, p(y|\theta)$ integrate to 1; and (b) $p(y|\theta)$ may look like the usual frequentist sampling distribution for y given θ (Bernoulli, in this case), but I have to think of it as a function of θ for fixed y. We've already encountered this idea (p. 15): $l(\theta|y) = c\, p(y|\theta)$ is Fisher's **likelihood function**.

Prior, Posterior, and Predictive Distributions

So **Bayes' Theorem** becomes

$$p(\theta|y) = c \cdot p(\theta) \cdot l(\theta|y) \ , \quad (33)$$

$$\text{posterior} = \left(\begin{array}{c} \text{normalizing} \\ \text{constant} \end{array} \right) \cdot \text{prior} \cdot \text{likelihood} \ .$$

You can also readily construct **predictive distributions** for the y_i before they're observed, or for future y_i once some of them are known.

For example, by the LTP, the **posterior predictive distribution** for $(y_{m+1}, \ldots, y_n)$ given $(y_1, \ldots, y_m)$ is

$$p(y_{m+1}, \ldots, y_n | y_1, \ldots, y_m) = \quad (34)$$

$$\int_0^1 p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m) \, p(\theta | y_1, \ldots, y_m) \, d\theta.$$

Consider $p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)$: if you **knew** θ, the information $y_1, \ldots, y_m$ about how the first m of the y_i came out would be **irrelevant** (imagine predicting the results of IID coin-tossing: if you somehow **knew** that the coin was perfectly fair, i.e., that $\theta = 0.5$, then getting (say) 6 heads in the first 10 tosses would be useless to you in quantifying the likely behavior of the next (say) 20 tosses—you'd just use the **known true value** of θ).

Prior, Posterior, and Predictive Distributions

Thus $p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)$ is just $p(y_{m+1}, \ldots, y_n | \theta)$, which in turn is just the **sampling distribution** under IID B(θ) sampling for the binary observables $y_{m+1}, \ldots, y_n$, namely $\prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$.

And finally $p(\theta | y_1, \ldots, y_m)$ is recognizable as just the **posterior distribution** for θ given the first m of the binary outcomes.

Putting this all together gives

$$p(y_{m+1}, \ldots, y_n | y_1, \ldots, y_m) = \tag{35}$$
$$= \int_0^1 \prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, p(\theta | y_1, \ldots, y_m) \, d\theta$$

(we can't compute (35) yet because $p(\theta | y_1, \ldots, y_m)$ depends on $p(\theta)$, which we haven't **specified** so far).

This also brings up a key difference between a **parameter** like θ on the one hand and the Y_i, before you've observed any data, on the other: parameters are inherently **unobservable**.

This makes it harder to evaluate the **quality** of your uncertainty assessments about θ than to do so about the **observable** y_i: to see how well you're doing in predicting observables you can just compare your predictive distributions for them with how they actually turn out, but of course this isn't possible with things like θ **which you'll never actually see**.

2.5 Inference and prediction. Coherence and calibration

The de Finetti approach to modeling emphasizes the **prediction** of observables as a valuable adjunct to **inference** about unobservable parameters, for at least two reasons:

• Key scientific questions are often **predictive** in nature: e.g., rather than asking "Is drug A better than B (on average across many patients) for lowering blood pressure?" (inference) the ultimate question is "How much more will drug A lower **this patient's** blood pressure than drug B?" (prediction); and

• Good **diagnostic checking** is predictive: An inference about an unobservable parameter can never be directly verified, but often you can reasonably conclude that inferences about the parameters of a model which produces poor predictions of observables are also **suspect**.

With the predictive approach parameters diminish in importance, especially those that have no physical meaning—such parameters (unlike θ above) are just **place-holders for a particular kind of uncertainty on your way to making good predictions**.

It's arguable (e.g., Draper, 1995) that the discipline of statistics, and particularly its applications in the social sciences, would be improved by a **greater emphasis on predictive feedback**.

This is not to say that parametric thinking should be **abolished**.

As the calculations on the previous pages emphasized, parameters play an important simplifying role in forming modeling judgments: the single strongest simplifier of a joint distribution is **independence** of its components, and whereas, e.g., in the mortality example the Y_i are not themselves independent, they become so conditional on θ.

Where Does the Prior Come From?

de Finetti's Theorem for 0–1 outcomes says informally that if you're trying to make **coherent** (internally consistent) probability judgments about a series of 1s and 0s that you judge exchangeable, you may as well behave like a frequentist—IID B(θ)—with a prior distribution $p(\theta)$.

But **where** does this prior **come from**?

NB Coherence doesn't help in answering this question—it turns out that **any** prior $p(\theta)$ could be part of **somebody**'s coherent probability judgments.

Some people regard the need to answer this question in the Bayesian approach as a **drawback**, but it seems to me (and to many other people) to be a **positive feature**, as follows.

From Bayes' Theorem the prior is supposed to be a summary of **what you know (and don't know)** about θ before the y_i start to arrive: from previous datasets of which you're aware, from the relevant literature, from expert opinion, ... from all **"good"** source(s), if any exist.

Such information is almost always present, and should presumably be used when available—the issue is how to do so "well."

The goal is evidently to choose a prior that you'll **retrospectively** be **proud of**, in the sense that your predictive distributions for the observables (a) are well-centered near the actual values and (b) have uncertainty bands that correspond well to the realized discrepancies between actual and predicted values. This is a form of **calibration** of your probability judgments.

There is **no guaranteed way to do this**, just as there is no guaranteed way to arrive at a "good" frequentist model (see "Where does the likelihood come from?" below).

Choosing a "good" Prior

Some general comments on arriving at a **"good"** prior:

• There is a growing literature on methodology for **elicitation** of prior information (e.g., Kadane et al., 1980; Craig et al., 1997; Kadane and Wolfson, 1997; O'Hagan, 1997), which brings together ideas from statistics and perceptual psychology (e.g., people turn out to be better at estimating **percentiles** of a distribution than they are at estimating **standard deviations** (SDs)).

• Bayes' Theorem on the **log scale** says (apart from the normalizing constant)

$$\log(\text{posterior}) = \log(\text{prior}) + \log(\text{likelihood}), \tag{36}$$

i.e., (posterior information) = (data information) + (prior information). This means that **close attention should be paid to the information content of the prior** by, e.g., density-normalizing the likelihood and plotting it on the same scale as the prior: it's possible for small n for the **prior to swamp the data**, and in general you shouldn't let this happen without a good reason for doing so.

Comfort can also be taken from the other side of this coin: with large n (in many situations, at least) the **data will swamp the prior**, and specification errors become less important.

• When you notice you're quite uncertain about how to specify the prior, you can try **sensitivity** or **(pre-posterior) analysis**: exploring the mapping from prior to posterior, before the data are gathered, by (a) generating some possible values for the observables, (b) writing down several plausible forms for the prior, and (c) carrying these forward to posterior distributions.

If the resulting distributions are similar (i.e., if "all reasonable roads lead to Rome"), you've uncovered a useful form of **stability** in your results; if not you can try to capture the prior uncertainty **hierarchically**, by, e.g., adding another layer to a model like (30) above.

• Calibration can be estimated by a form of **cross-validation:** with a given prior you can (a) repeatedly divide the data at random into modeling and validation subsets, (b) update to posterior predictive distributions based on the modeling data, and (c) compare these distributions with the actual values in the validation data.

I'll illustrate some **examples** of this idea later.

Computation: Conjugate Analysis

Note that calibration is **inherently frequentist** in spirit (e.g., "What percentage of the time do your 95% central predictive intervals include the actual value?"). This leads to a useful **synthesis** of Bayesian and frequentist thinking:

> **Coherence** keeps you internally honest; **calibration** keeps you in good contact with the world.

2.6 Conjugate analysis. Comparison with frequentist modeling

Example: Prior specification in the mortality data. Let's say (a) you know (from the literature) that the 30-day AMI mortality rate given average care and average sickness at admission in the U.S. is about **15%**, (b) You know little about care or patient sickness at the DH, but (c) You'd be somewhat surprised (e.g., on Central Limit Theorem grounds) if the "underlying rate" at the DH was much less than **5%** or more than **30%** (note the asymmetry). To quantify these judgments you seek a flexible family of densities on (0,1), one of whose members has mean **.15** and (say) **95% central interval (.05,.30)**.

A convenient family for this purpose is the **beta** distributions,

$$\text{Beta}(\theta|\alpha,\beta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad (37)$$

defined for $(\alpha > 0, \beta > 0)$ and for $0 < \theta < 1$.

We can use `Maple` to evaluate the **normalizing constant** c.

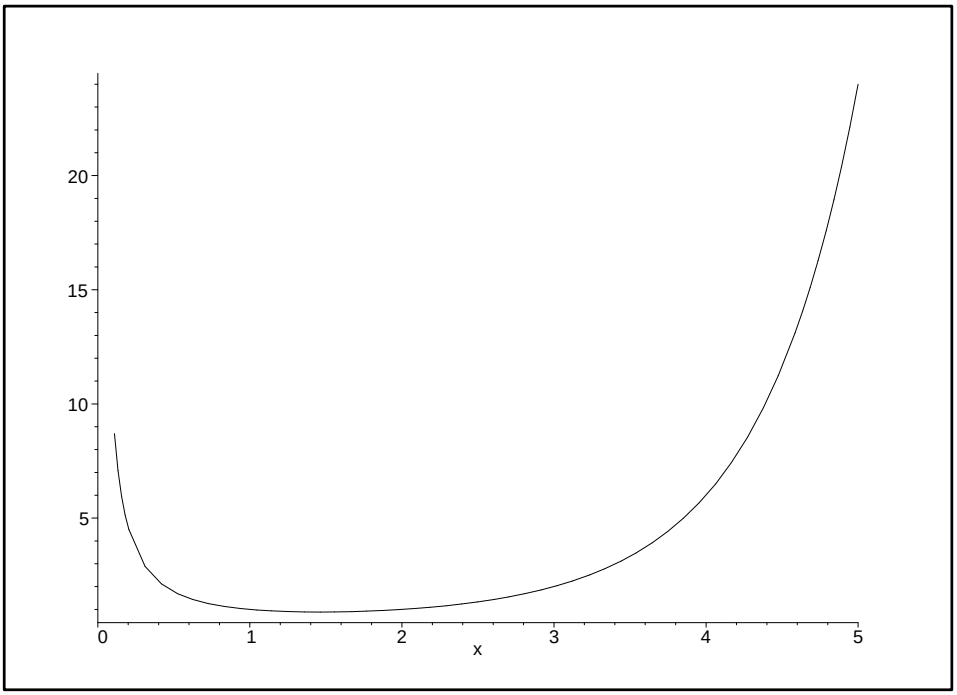
```
rosalind 3> maple

    |\^/|     Maple V Release 5 (University of California, Santa Cruz)
._|\|   |/|_. Copyright (c) 1981-1997 by Waterloo Maple Inc. All rights
 \  MAPLE  /  reserved. Maple and Maple V are registered trademarks of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.

> assume( alpha > 0, beta > 0, theta > 0, theta < 1 );
```

The Beta Distribution

```
> p1 := ( theta, alpha, beta ) -> theta^( alpha - 1 ) *
    ( 1 - theta )^( beta - 1 );

                                          (alpha - 1)            (beta - 1)
     p1 := (theta, alpha, beta) -> theta             (1 - theta)

> integrate( p1( theta, alpha, beta ), theta = 0 .. 1 );

                              Beta(alpha~, beta~)

> help( Beta );

Beta - The Beta function

Calling Sequence:
     Beta( x, y )

Parameters:
     x - an expression
     y - an expression

Description:

- The Beta function is defined as follows:

      Beta( x, y ) = ( GAMMA( x ) * GAMMA( y ) ) / GAMMA( x + y )

> help( GAMMA );

GAMMA - The Gamma and Incomplete Gamma Functions

lnGAMMA - The log-Gamma function

Calling Sequence:

     GAMMA( z )
     GAMMA( a, z )
     lnGAMMA( z )

Parameters:

     z - an expression
     a - an expression
```

The Beta Distribution (continued)

```
Description:

- The Gamma function is defined for Re( z ) > 0 by

      GAMMA(z) =  int( exp( -t ) * t^( z - 1 ), t = 0 .. infinity )

  and is extended to the rest of the complex plane,
  less the non-positive integers, by analytic continuation.
  GAMMA has a simple pole at each of the points
  z = 0, -1, -2, ... .

- For positive real arguments z, the lnGAMMA function is defined by:

      lnGAMMA( z ) = ln( GAMMA( z ) )

> plotsetup( x11 );

> plot( GAMMA( x ), x = 0 .. 5, color = black );
```



Notice that $\Gamma(1) = 1, \Gamma(2) = 1, \Gamma(3) = 2, \Gamma(4) = 6$, and $\Gamma(5) = 24$—the **pattern** here is that

$$\Gamma(n) = (n-1)! \quad \text{for integer } n. \tag{38}$$

The Beta Distribution (continued)

Thus the Gamma function is a kind of **continuous generalization** of the **factorial** function.

What all of this has shown is that the **normalizing constant** in the beta distribution is

$$c = \left[\int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} \, d\theta \right]^{-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}, \qquad (39)$$

so that the full definition of the **beta distribution** is

$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad (40)$$

for $(\alpha > 0, \beta > 0)$ and for $0 < \theta < 1$.

The beta family is **convenient** for two reasons:

(1) It exhibits a wide variety of **distributional shapes** (e.g., Johnson and Kotz, 1970):

```
> p := ( theta, alpha, beta ) -> ( GAMMA( alpha + beta ) /
    ( GAMMA( alpha ) * GAMMA( beta ) ) ) *  theta^( alpha - 1 ) *
    ( 1 - theta )^( beta - 1 );

p := (theta, alpha, beta) ->

                                (alpha - 1)            (beta - 1)
    GAMMA(alpha + beta) theta              (1 - theta)
    ------------------------------------------------------------
                      GAMMA(alpha) GAMMA(beta)

> plot( { p( theta, 0.5, 0.5 ), p( theta, 1, 1 ), p( theta, 2, 3 ),
    p( theta, 30, 20 ) }, theta = 0 .. 1, color = black );
```