1 Probability models

1.1 Observation, experiments and models

Stochastic Models[1]

Normal vs *Bernoulli* and *Poisson*: We need to distinguish between *individual* observations, governed by Bernoulli and Poisson (or if quantitative rather than all-or-none or a count, Normal) and *statistics* formed by aggregation of individual observations. If a large enough number of individual observations are used to form a statistic, its (sampling) distribution can be described by a Gaussian (Normal) probability model. So, ultimately, this probability model is just as relevant.

1.1.1 Epidemiologic [subject-matter] models [JH]

We need to also make a distinction between the quantity(quantities) that is(are) of substantive interest or concern, the data from which this(these) is(are) estimated, the *statistical* models used to get to the the quantity(quantities) and the relationships of interest.

For example, of medical, public health or personal interest/concern might be the

- level of use of cell phones while driving
- average and range [across people] of reductions in cholesterol with regular use of a cholesterol-lowering medication
- amount of time taken by health care personnel to decipher the handwriting of other health care personnel
- (average) number of times people have to phone to reach a 'live' person
- reduction in one's risk of dying of a specific cancer if one is regularly screened for it.
- appropriate-size tracheostomy tube for an obese patient, based on easily easily obtained anthropometric measurements
- length of central venous catheter that can be safely inserted into a child as a function of the child's height etc.
- rate of automobile accidents as a function of drivers' blood levels of alcohol and other drugs, numbers of persons in the car, cell-phone and other activities, weather, road conditions, etc.
- Psychological Stress, Negative Life Events, Perceived Stress, Negative Affect Smoking, Alcohol Consumption and Susceptibility to the Common Cold
- The force of mortality s a function of age, sex and calendar time.
- Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction
- Are seat belt restraints as effective in school age children as in adults?
- Levels of folic acid to add to flour, so that most people have sufficiently high blood levels.
- Early diet in children born preterm and their IQ at age eight.
- Prevalence of Down's syndrome in relation to parity and maternal age.

Of broader interest/concern might be

- the wind chill factor as a function of temperature and wind speed
- how many fewer Florida votes Al Gore got in 2000 because of a badly laid-out ballot
- a formula for deriving one's "ideal" weight from one's height
- yearly costs under different cell-phone plans
- yearly maintenance costs for different makes and models of cars
- car or life insurance premiums as a function of ...
- cost per foot^2 of commercial or business rental space as a function of ...
- Rapid Changes in Flowering Time in British Plants
- How much money the City of New York should revover from Brink's for the losses the City incurred by the criminal activities of two Brink's employees (they collected the money form the parking meters, but kept some of it!).

[1] 'Stochastic' `http://www.allwords.com/word-stochastic.html` French: stochastique(fr) German: stochastisch(de) Spanish: estocstico(es) Etymology: From Ancient Greek (polytonic,), from (polytonic,) "aim at a target, guess", from (polytonic,) "an aim, a guess". Parzen, in his text on Stochastic Processes .. page 7 says: <<The word is of Greek origin; see Hagstroem (1940) for a study of the history of the word. In seventeenth century English, the word "stochastic" had the meaning "to conjecture, to aim at a mark." It is not clear how it acquired the meaning it has today of "pertaining to chance." Many writers use the expression "chance process" or "random process" as synonyms for "stochastic process."<<

1.1.2 From behaviour of statistical 'atoms' to statistical 'molecules'

1 condition' or 1 circumstance' or 'setting' [also known as "1-sample problems"]

The smallest statistical element or unit (?atom): its quantity of interest might have a Y distribution that under sampling, could be represented by a discrete random variable with '2-point' support (Bernoulli), 3-point support, $k-$point support, etc. or interval support (Normal, gamma, beta, log-normal, ...)

The *aggregate* or summary of the values associated with these elements is often a sum or a count: with e.g., a Binomial, Negative Binomial, gamma distribution. Or teh summary might be more complex – it could be some re-arrangement of the data on the individuals (e.g., the way the tumbler longevity data were summarized). This brings in the notion of "sufficient statistics".

More complex: t, F, ...

2 or conditions' or 1 circumstances' or 'settings', indexed by possible values of 'X' variable(s). Think of the 'X' variable(s) as 'covariate patterns' or 'profiles.'

unknown conditions or circumstances Sometimes we don't have any measurable (or measured) 'X' variable(s) to explain the differences in Y from say family to family or person to person.There instead of the usual multiple regression approach, we use the concept of a hierarchical or random-effects or latent class or mixture model.

1.2 Binary data

It is worth recalling from the first semester, the concepts of states and events (transitions from one state to another).

Cohort studies with fixed follow-up time

Recall: *cohort* is another name for a closed population, with membership (entry) defined by some event, such as birth, losing one's virginity, obtaining one's first driver's permit, attaining age 21, graduating from university, entering the 'ever-married' state, undergoing a certain medical intervention, enrolling in a follow-up study, etc. Then the *event of interest* is the *exit* (transition) from a/the state that prevailed at entry. So *death* is the transition from the *living* state to the *dead* state, receiving a *diagnosis* of cancer changes one's state from 'no history of cancer since entry' to 'have a history of cancer', being convicted of a traffic offense changes one's state from 'clean record' to 'have a history of traffic offenses.' We can also envision more complex situations, with a transition from 'never had a stroke,' to 'have had 1 stroke,' to 'have had 2 strokes,' ... or 'haven't yet had a cold this winter,' to 'have had 1 cold,' to 'have had 2 colds,' etc.

Censoring: to be distinguished from *truncation*. Truncation implies some observations are missed by the data-gathering process, i.e., that the observed distribution is a systematic distortion of the true distribution. Note that we can have censoring of any quantitative variable, not just one that measures the duration until the event of interest. For example, the limits on say a thermometer or a weight scale or a chemical assay may mean that it cannot record/detect values below or above these limits. Also, the example in C&H implicitly refers to *right* censoring: one can have *left* censoring, as with lower limits of detection in a chemical assay, or *interval* censoring, as – in repeated cross-sectional examinations – with the date of sero-conversion to HIV.

Incidence studies: the word *new* means a change of state since entry.

"*Failure*": It is a pity that C&H didn't go one step more and use the even more generic term "*event*". That way, they would not have to think of graduating with a PhD (i.e., *getting out of – exiting from – here*) as "*failure*" and still being here" as "*survival*." This simpler and more general terminology would mean that we would not have to struggle to find a suitable label of the 'y' axis of the $1 - F(t)$, usually called $S(t)$, function. One could simply say *"proportion still in initial state,"* and substitute the term for the initial state, i.e., proportion still in PhD program, proportion event-free, etc.

N or n? D or d? JH would have preferred lower case, at least for the denominator. In *sampling* textbooks, N usually denotes the *population* size, and n the *sample* size. In the style manual used in *social sciences*, n is the sample size in each stratum, whereas N is the overall sample size. Thus, for example, a study might report on a sample of $N = 76$ subjects, composed of $n = 40$ females and $n = 36$ males.

Cohort studies with variable follow-up time: If every subject entered a study at least 5 years ago, then, in principle, one should be able to determine D and $N - D$, and the 5-year survival proportion. However, *losses to follow-up* before 5 years, and before the event of interest, lead to observations that are typically regarded as censored at the time of the loss. Another phenomenon that leads to censored observations is *staggered entry*, as in the JUPITER trial. Unfortunately, some losses to follow-up may be examples of *informative*' censoring.

Cross-sectional prevalence data

Recall again that prevalence refers to a *state*. Examples would include the

proportion (of a certain age group, say) who wear glasses for reading, or have undetected high blood pressure, or have high-speed internet at home, or have a family history of a certain disease, or a certain 'gene' or blood-type.

From a purely *statistical* perspective, the analysis of *prevalence* proportions of the form D/N and *incidence* proportions of the form D/N takes the same form: the underlying statistical 'atoms' are N Bernoulli random variables.

1.3 The binary probability model

JH presumes they use this heading as a shorthand for 'the probability model for binary responses' (or 'binary outcomes' or binary random variables)

... to "*predict* the outcome" : JH takes this word *predict* in its broader meaning. If we are giving a patient the probability that he will have a certain *future* event *say within the next 5 years*, we can talk about predicting the outcome: we are speaking of *pro*gnosis; but what if we are giving a woman the probability that the suspicious finding on a mammogram does in fact represent an existing breast cancer, we are speaking of the *present*, of whether a phenomenon already *exists*, and we use a prevalence proportion as an estimate of the *dia*gnostic probability. Note that prevalence and incidence refer to aggregates.

The risk parameter

Risk typically refers to the *future*, and can be used when speaking to or about one person; we don't have a comparable specialized term for *the probability that a state exists* when speaking to or about one person, and would therefore just use the generic term probability.

The odds parameter

The sex-ratio is often expressed as an odds, i.e., as a ratio of males to females. If the proportion of males is 0.51, then the male:female ratio is 51:49 or (51/49):1, i.e., approximately 1.04:1. This example is a good reason why C&H should have used a more generic pair of terms than failure and survival (or success and failure).

In betting on horse races (at least where JH comes from), odds of 3:1 are the odds *against* the horse winning; i.e., the probability of winning is 1/4. When a horse is a heavy favourite so that the probability of winning was 75%, the "bookies" would give the odds as "3:1 *on*."

Rare events

One of the tricks to make events *rare* will be to slice the time period into small slices or windows.

Death, the first of the two only sure events (taxes is the other) is also rare - in the short term!

Also, it would be more correct to speak of a *rare events*, since disease is often used to describe a process, rather than a transition. And since most transitions are rapid, the probability of a transition (an event) occurring within a given short sub-interval will usually be small.

If the state of interest being addressed with cross-sectional data is uncommon (or rare), then yes, the prevalence odds and the prevalence proportion will be very close to each other.

Supplementary Exercise 1.1. If one rounds probabilities or risks or prevalences (π's), or their corresponding odds, $\Omega = \pi/(1-\pi)$, to 1 decimal place, at what value of π will the rounded values of π and Ω be differerent? Also, why use lowercase π for proportion, and uppercase Ω for odds?

1.4 Parameter Estimation

Should you be surprised if the estimate were π were other than D/N? Consult Google or Wikipedia on "the rule of succession," and on Laplace's estimate of the probability that the sun will rise tomorrow, given that it has unfailingly risen ($D = 0$) for the past 6000 years, i.e., $N \approx 365 \times 6000$.

Supplementary Exercise 1.2. One has 2 independent observations from the model

$$E[y|x] = \beta \times x.$$

The y's might represent the total numbers of typographical errors on x randomly sample pages of a large document, and the data might be $y = 2$ errors in total in a sample of $x = 1$ page, and $y = 8$ errors in total in a separate sample of $x = 2$ pages. The β in the model represents the mean number of errors per page of the document. Or the y's might represent the total weight of x randomly sample pages of a document, and the data might be $y = 2$ units of weight in total for a sample of $x = 1$ page, and $y = 8$ units for a separate sample of $x = 2$ pages. The β in the model represents the mean weight per page of the document. We gave this 'estimation of β' problem to several statisticians and epidemiologists, and to several grade 6 students, and they gave us a variety of estimates, such as $\hat{\beta} = 3.6$/page, 3.33/page, and 3.45!

How can this be? You might wish to run the applet '2 datapoints and a model' (link from the bottom left corner of JH's home page.)

1.5 Is the model true?

I wonder if they were aware of the quote, attributed to the statistician George Box that goes something like this

> "all models are wrong; but some are more useful than others"

`http://en.wikiquote.org/wiki/George_E._P._Box`

2 Conditional probability models

2.1 Conditional probability

JH is suprised at how few textbooks used trees to explain conditional probabilities. Probability trees make it easy to see the direction in which one is preceeding, or looking, where simply algebraic symbols can not, and make it easier to distinguish 'forward' from 'reverse' probabilities.

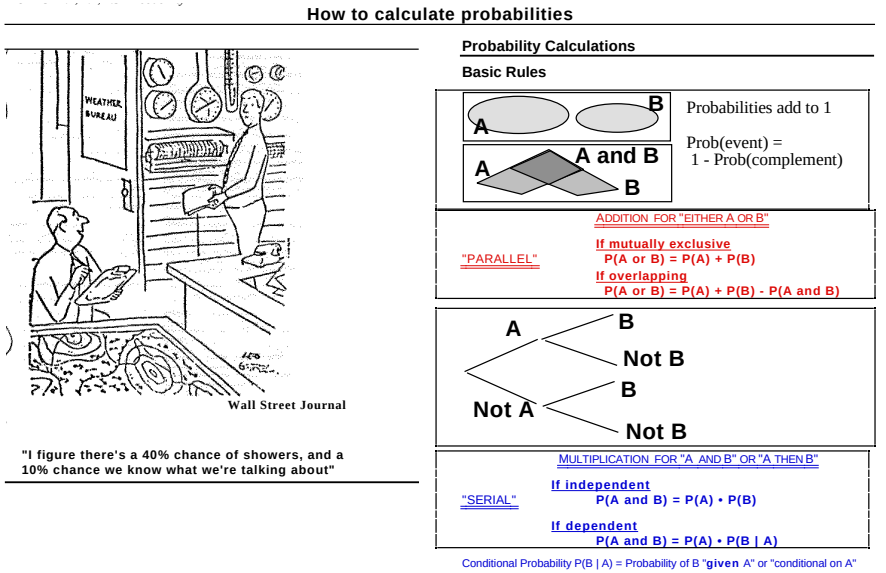

Figure 1: From JH's notes for EPIB607, introductory biostatistics for epidemiology

Trees show that the probability of a particular sequence is always a fraction of a fraction of a fraction .. , and that if we start with the full probability of 1 at the single entry point on the extreme left, then we need at the right hand side to account for all of this (i.e., the 'total') probability.

Statistical dependence and independence

JH likes to say that with independence, one doesn't have to look over one's shoulder to the previous event to know which probability to multiple by.. The illustrated example on the gender composition of 2 independent births, and of a sample of 2 persons sampled (without replacement) from a pool of 5 males and 5 females, show this distinction: in the first example, when one comes to the second component in the probability product, $Pr(y_2 = male)$ is the same

whether one has got to there via the 'upper' path, or the 'lower' one. know

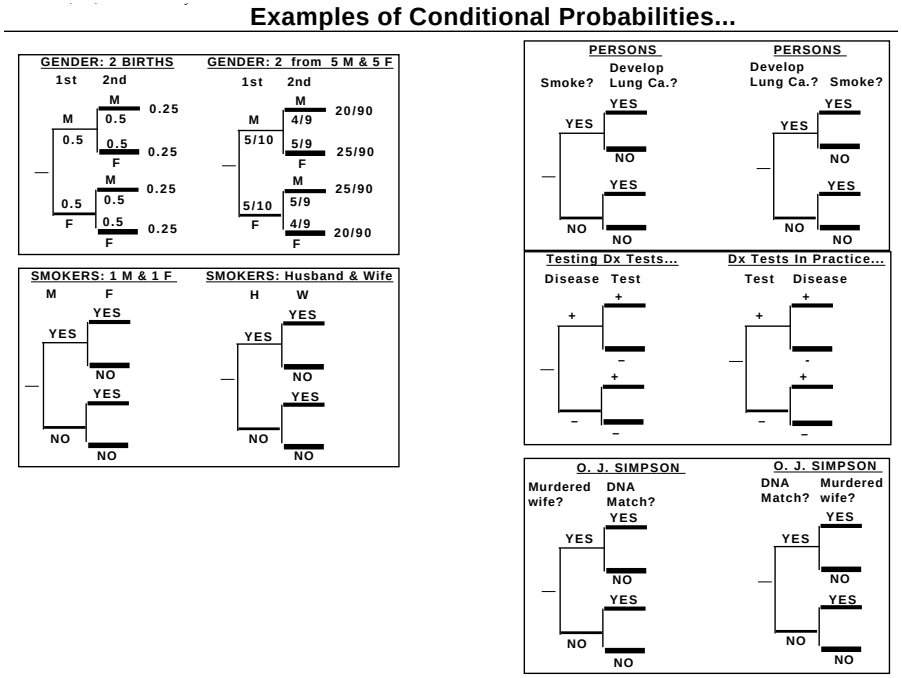

Figure 2: JH examples of independence/dependence, and 'forward'/'reverse' probabilities

2.2 Changing the conditioning: Bayes' rule

The right hand half of JH Figure 2 shows 3 examples of 'forward' (on left) and 'reverse' probabilities.

These same distinctions between 'forward' and 'reverse' probabilities is at the heart of the frequentist p-values (probabilities) versus Bayesian posterior probabilities. To state it simply,

$$Probability[data|Hypothesis] \neq Probability[Hypothesis|data]$$

or, if you prefer something that rhymes,

$$Probability[data|theta] \neq Probability[theta|data].$$

Two striking – and frightening – examples of misunderstandings about them are given on the next page.

U.S. National Academy of Sciences under fire over plans for new study of DNA statistics: Confusion leads to retrial in UK.[2]

[...] He also argued that one of the prosecution's expert witnesses, as well as the judge, had confused two different sorts of probability.

One is the probability that DNA from an individual selected at random from the population would match that of the semen taken from the rape victim, a calculation generally based solely on the frequency of different alleles in the population. The other is the separate probability that *a match between a suspect's DNA and that taken from the scene of a crime could have arisen simply by chance* – **in other words that the suspect is innocent despite the apparent match.**[3] This probability depends on the other factors that led to the suspect being identified as such in the first place.

During the trial, a forensic scientist gave the first probability in reply to a question about the second. Mansfield convinced the appeals court that the error was repeated by the judge in his summing up, and that this slip – widely recognized as a danger in any trial requiring the explanation of statistical arguments to a lay jury – justified a retrial. In their judgement, the three appeal judges, headed by the Lord Chief Justice, Lord Farquharson, explicitly stated that their decision "should not be taken to indicate that DNA profiling is an unsafe source of evidence."

Nevertheless, with DNA techniques being increasingly used in court cases, some forensic scientists are worried that flaws in the presentation of their statistical significance could, as in the Deen case, undermine what might otherwise be a convincing demonstration of a suspect's guilt.

Some now argue, for example, that quantified statistical probabilities should be replaced, wherever possible, by a more descriptive presentation of the conclusions of their analysis. "The whole issue of statistics and DNA profiling has got rather out of hand," says one. Others, however, say that the Deen case has been important in revealing the dangers inherent in the '**prosecutor's fallacy**'. They argue that this suggests the need for more sophisticated calculation and careful presentation of statistical probabilities. "The way that the prosecution's case has been presented in trials involving DNA-based identification has often been very unsatisfactory," says David Balding, lecturer in probability and statistics at Queen Mary and Westfield College in London. "Warnings about the prosecutor's fallacy should be made much more explicit. After this decision, people are going to have to be more careful."

[2] NATURE p 101-102 Jan 13, 1994.

[3] italics by JH. The wording of the italicized phrase is imprecise; the text in bold wording is much better .. if you read "despite" as "given that" or "conditional on the fact of"t

"The prosecutor's fallacy": Who's the DNA fingerprinting pointing at? [4]

Pringle describes the successful appeal of a rape case where the primary evidence was DNA fingerprinting. In this case the statistician Peter Donnelly opened a new area of debate. He remarked that

> forensic evidence answers the question "What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?"
>
> while the jury must try to answer the question "What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?" [5]

Apparently, Donnelly suggested to the Lord Chief Justice and his fellow judges that they imagine themselves playing a game of poker with the Archbishop of Canterbury. If the Archbishop were to deal himself a royal flush on the first hand, one might suspect him of cheating. Assuming that he is an honest card player (and shuffled eleven times) the chance of this happening is about 1 in 70,000.

But if the judges were asked whether the Archbishop were honest, given that he had just dealt a royal flush, they would be likely to place the chance a bit higher than 1 in 70,000 *.

The error in mixing up these two probabilities is called the "the prosecutor's fallacy," and it is suggested that newspapers regularly make this error.

Apparently, Donnelly's testimony convinced the three judges that the case before them involved an example of this and they ordered a retrial

[* Comment by JH: This is a very nice example of the advantages of Bayesian over Frequentist inference .. it lets one take one's prior knowledge (the fact that he is the Archbishop) into account.

The book 'Statistical Inference" by Michael W. Oakes is an excellent introduction to this topic and the limitations of frequentist inference.]

[4] New Scientist item by David Pringle, 1994.01.29, 51-52; cited in Vol 3.02 Chance News

[5] (JH) Donnelly's words make the contrast of the two types of probability much "crisper." The fuzziness of the wording on the previous story is sadly typical of the way statistical concepts often become muddied as they are passed on.

2.3 Examples

2.3.1 Example from genetics

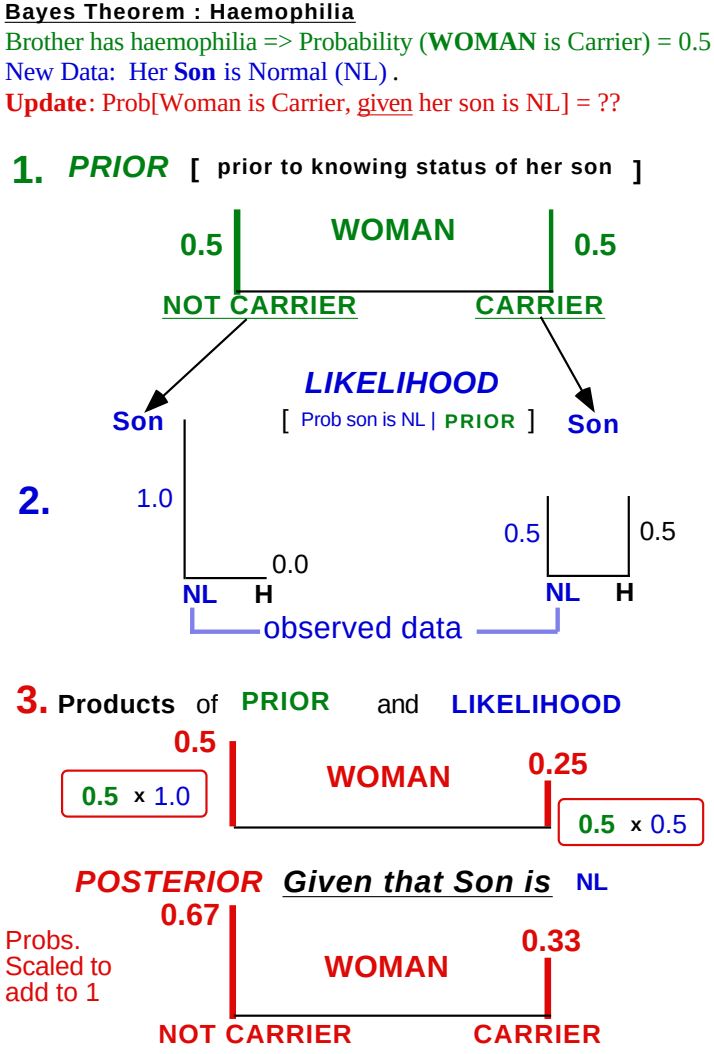

Figure 3: a simpler (but now outdated) example – nowadays there are direct tests for being a carrier: so one doesn't have to wait to have a son to alter the probabilities

2.3.2 Twins: Excerpt from an article by Bradley Efron

Modern science and the Bayesian-Frequentist controversy

> Heres a real-life example I used to illustrate Bayesian virtues to the physicists. A physicist friend of mine and her husband found out, thanks to the miracle of sonograms, that they were going to have twin boys. One day at breakfast in the student union she suddenly asked me what was the probability that the twins would be identical rather than fraternal. This seemed like a tough question, especially at breakfast. Stalling for time, I asked if the doctor had given her any more information. "Yes", she said, "he told me that the proportion of identical twins was one third". This is the population proportion of course, and my friend wanted to know the probability that her twins would be identical.
>
> Bayes would have lived in vain if I didn't answer my friend using Bayes' rule. According to the doctor the prior odds ratio of identical to nonidentical is one-third to two-thirds, or one half. Because identical twins are always the same sex but fraternal twins are random, the likelihood ratio for seeing "both boys" in the sonogram is a factor of two in favor of Identical. Bayes' rule says to multiply the prior odds by the likelihood ratio to get the current odds: in this case 1/2 times 2 equals 1; in other words, equal odds on identical or nonidentical given the sonogram results. So I told my friend that her odds were 50-50 (wishing the answer had come out something else, like 63-37, to make me seem more clever.) Incidentally, the twins are a couple of years old now, and "couldnt be more non-identical" according to their mom.

Supplementary Exercise 2.1. Depict Efron's calculations using a probability tree.

Supplementary Exercise 2.2 Use a probability tree to determine the best strategy in the Monty Hall problem

```
( http://en.wikipedia.org/wiki/Monty_Hall_problem )
```

Supplementary Exercise 2.3 A man has exactly two children: you meet the *older* one and see that it's a boy. A woman has exactly two children; you meet *one* of them [don't know if its the younger/older] and see is a boy. What is the probability of the man's younger child being a boy, and what is the probability of the woman's "other" child being a boy?