35

Random Inference Topics

▶ 35.1 What do you know if you are ignorant?

Example 35.1. A real variable x is measured in an accurate experiment. For example, x might be the half-life of the neutron, the wavelength of light emitted by a firefly, the depth of Lake Vostok, or the mass of Jupiter's moon Io.

What is the probability that the value of x starts with a '1', like the charge of the electron (in S.I. units),

$$e = 1.602\ldots \times 10^{-19}\,\mathrm{C},$$

and the Boltzmann constant,

$$k = 1.380\,66\ldots \times 10^{-23}\,\mathrm{J\,K^{-1}}?$$

And what is the probability that it starts with a '9', like the Faraday constant,
$$\mathcal{F} = 9.648\ldots \times 10^{4}\,\mathrm{C\,mol^{-1}}?$$

What about the second digit? What is the probability that the mantissa of x starts '1.1...', and what is the probability that x starts '9.9...'?

Solution. An expert on neutrons, fireflies, Antarctica, or Jove might be able to predict the value of x, and thus predict the first digit with some confidence, but what about someone with no knowledge of the topic? What is the probability distribution corresponding to 'knowing nothing'?

One way to attack this question is to notice that the units of x have not been specified. If the half-life of the neutron were measured in fortnights instead of seconds, the number x would be divided by $1\,209\,600$; if it were measured in years, it would be divided by 3×10^{7}. Now, is our knowledge about x, and, in particular, our knowledge of its first digit, affected by the change in units? For the expert, the answer is yes; but let us take someone truly ignorant, for whom the answer is no; their predictions about the first digit of x are independent of the units. The arbitrariness of the units corresponds to *invariance* of the probability distribution when x is *multiplied* by any number.

If you don't know the units that a quantity is measured in, the probability of the first digit must be proportional to the length of the corresponding piece of logarithmic scale. The probability that the first digit of a number is 1 is thus

$$p_1 = \frac{\log 2 - \log 1}{\log 10 - \log 1} = \frac{\log 2}{\log 10}. \tag{35.1}$$
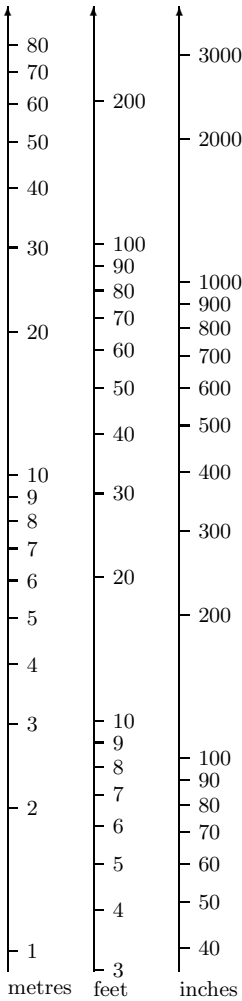
Figure 35.1. When viewed on a logarithmic scale, scales using different units are translated relative to each other.

Now, $2^{10} = 1024 \simeq 10^3 = 1000$, so without needing a calculator, we have $10 \log 2 \simeq 3 \log 10$ and

$$p_1 \simeq \frac{3}{10}. \qquad (35.2)$$

More generally, the probability that the first digit is d is

$$(\log(d+1) - \log(d))/(\log 10 - \log 1) = \log_{10}(1 + 1/d). \qquad (35.3)$$

This observation about initial digits is known as Benford's law. Ignorance does not correspond to a uniform probability distribution over d.   □

▷ Exercise 35.2.[2] A pin is thrown tumbling in the air. What is the probability distribution of the angle θ_1 between the pin and the vertical at a moment while it is in the air? The tumbling pin is photographed. What is the probability distribution of the angle θ_3 between the pin and the vertical as imaged in the photograph?

▷ Exercise 35.3.[2] Record breaking. Consider keeping track of the world record for some quantity x, say earthquake magnitude, or longjump distances jumped at world championships. If we assume that attempts to break the record take place at a steady rate, and if we assume that the underlying probability distribution of the outcome x, $P(x)$, is not changing – an assumption that I think is unlikely to be true in the case of sports endeavours, but an interesting assumption to consider nonetheless – and assuming no knowledge at all about $P(x)$, what can be predicted about successive intervals between the dates when records are broken?
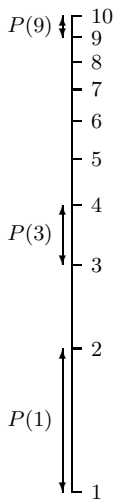
▶ 35.2 The Luria–Delbrück distribution

Exercise 35.4.[3C, p.449] In their landmark paper demonstrating that bacteria could mutate from virus sensitivity to virus resistance, Luria and Delbrück (1943) wanted to estimate the mutation rate in an exponentially-growing population from the total number of mutants found at the end of the experiment. This problem is difficult because the quantity measured (the number of mutated bacteria) has a heavy-tailed probability distribution: a mutation occuring early in the experiment can give rise to a huge number of mutants. Unfortunately, Luria and Delbrück didn't know Bayes' theorem, and their way of coping with the heavy-tailed distribution involves arbitrary hacks leading to two different estimators of the mutation rate. One of these estimators (based on the mean number of mutated bacteria, averaging over several experiments) has appallingly large variance, yet sampling theorists continue to use it and base confidence intervals around it (Kepler and Oprea, 2001). In this exercise you'll do the inference right.

In each culture, a single bacterium that is *not resistant* gives rise, after g generations, to $N = 2^g$ descendants, all clones except for differences arising from mutations. The final culture is then exposed to a virus, and the number of resistant bacteria n is measured. According to the now accepted mutation hypothesis, these resistant bacteria got their resistance from random mutations that took place during the growth of the colony. The mutation rate (per cell per generation), a, is about one in a hundred million. The total number of opportunities to mutate is N, since $\sum_{i=0}^{g-1} 2^i \simeq 2^g = N$. If a bacterium mutates at the ith generation, its descendants all inherit the mutation, and the final number of resistant bacteria contributed by that one ancestor is 2^{g-i}.

Given M separate experiments, in each of which a colony of size N is created, and where the measured numbers of resistant bacteria are $\{n_m\}_{m=1}^{M}$, what can we infer about the mutation rate, a?

Make the inference given the following dataset from Luria and Delbrück, for $N = 2.4 \times 10^8$: $\{n_m\} = \{1, 0, 3, 0, 0, 5, 0, 5, 0, 6, 107, 0, 0, 0, 1, 0, 0, 64, 0, 35\}$. [A small amount of computation is required to solve this problem.]

▶ 35.3 Inferring causation

Exercise 35.5.[2, p.450] In the Bayesian graphical model community, the task of inferring which way the arrows point – that is, which nodes are parents, and which children – is one on which much has been written.

Inferring causation is tricky because of 'likelihood equivalence'. Two graphical models are likelihood-equivalent if for any setting of the parameters of either, there exists a setting of the parameters of the other such that the two joint probability distributions of all observables are identical. An example of a pair of likelihood-equivalent models are $A \rightarrow B$ and $B \rightarrow A$. The model $A \rightarrow B$ asserts that A is the parent of B, or, in very sloppy terminology, 'A causes B'. An example of a situation where '$B \rightarrow A$' is true is the case where B is the variable 'burglar in house' and A is the variable 'alarm is ringing'. Here it is literally true that B causes A. But this choice of words is confusing if applied to another example, $R \rightarrow D$, where R denotes 'it rained this morning' and D denotes 'the pavement is dry'. 'R causes D' is confusing. I'll therefore use the words 'B is a parent of A' to denote causation. Some statistical methods that use the likelihood alone are unable to use data to distinguish between likelihood-equivalent models. In a Bayesian approach, on the other hand, two likelihood-equivalent models may nevertheless be somewhat distinguished, in the light of data, since likelihood-equivalence does not force a Bayesian to use priors that assign equivalent densities over the two parameter spaces of the models.

However, many Bayesian graphical modelling folks, perhaps out of sympathy for their non-Bayesian colleagues, or from a latent urge not to appear different from them, deliberately discard this potential advantage of Bayesian methods – the ability to infer causation from data – by skewing their models so that the ability goes away; a widespread orthodoxy holds that one should identify the choices of prior for which 'prior equivalence' holds, i.e., the priors such that models that are likelihood-equivalent also have identical posterior probabilities; and then one should use one of those priors in inference and prediction. This argument motivates the use, as the prior over all probability vectors, of specially-constructed Dirichlet distributions.

In my view it is a philosophical error to use only those priors such that causation cannot be inferred. Priors should be set to describe one's assumptions; when this is done, it's likely that interesting inferences about causation *can* be made from data.

In this exercise, you'll make an example of such an inference.

Consider the toy problem where A and B are binary variables. The two models are $\mathcal{H}_{A \rightarrow B}$ and $\mathcal{H}_{B \rightarrow A}$. $\mathcal{H}_{A \rightarrow B}$ asserts that the marginal probability of A comes from a beta distribution with parameters $(1, 1)$, i.e., the uniform distribution; and that the two conditional distributions $P(b \mid a = 0)$ and $P(b \mid a = 1)$ also come independently from beta distributions with parameters $(1, 1)$. The other model assigns similar priors to the marginal probability of B and the conditional distributions of A given B. Data are gathered, and the

counts, given $F = 1000$ outcomes, are

$$
\begin{array}{cc|c}
 & a=0 \quad a=1 & \\
b=0 & 760 \quad\quad 5 & 765 \\
b=1 & 190 \quad\quad 45 & 235 \\
\hline
 & 950 \quad\quad 50 &
\end{array}
\tag{35.4}
$$

What are the posterior probabilities of the two hypotheses?

> Hint: it's a good idea to work this exercise out symbolically in order to spot all the simplifications that emerge.

$$
\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma(x) \simeq \ln(x) - \frac{1}{2x} + O(1/x^2).
\tag{35.5}
$$

The topic of inferring causation is a complex one. The fact that Bayesian inference can sensibly be used to infer the directions of arrows in graphs seems to be a neglected view, but it is certainly not the whole story. See Pearl (2000) for discussion of many other aspects of causality.

▶ 35.4 Further exercises

Exercise 35.6.[3] Photons arriving at a photon detector are believed to be emitted as a Poisson process with a time-varying rate,

$$
\lambda(t) = \exp(a + b \sin(\omega t + \phi)),
\tag{35.6}
$$

where the parameters a, b, ω, and ϕ are known. Data are collected during the time $t = 0 \ldots T$. Given that N photons arrived at times $\{t_n\}_{n=1}^{N}$, discuss the inference of a, b, ω, and ϕ. [Further reading: Gregory and Loredo (1992).]

▷ Exercise 35.7.[2] A data file consisting of two columns of numbers has been printed in such a way that the boundaries between the columns are unclear. Here are the resulting strings.

891.10.0	912.20.0	874.10.0	870.20.0	836.10.0	861.20.0
903.10.0	937.10.0	850.20.0	916.20.0	899.10.0	907.10.0
924.20.0	861.10.0	899.20.0	849.10.0	887.20.0	840.10.0
849.20.0	891.10.0	916.20.0	891.10.0	912.20.0	875.10.0
898.20.0	924.10.0	950.20.0	958.10.0	971.20.0	933.10.0
966.20.0	908.10.0	924.20.0	983.10.0	924.20.0	908.10.0
950.20.0	911.10.0	913.20.0	921.25.0	912.20.0	917.30.0
923.50.0					

Discuss how probable it is, given these data, that the correct parsing of each item is:

(a) $891.10.0 \rightarrow 891.\ 10.0$, etc.

(b) $891.10.0 \rightarrow 891.1\ 0.0$, etc.

[A parsing of a string is a grammatical interpretation of the string. For example, 'Punch bores' could be parsed as 'Punch (noun) bores (verb)', or 'Punch (imperative verb) bores (plural noun)'.]

▷ Exercise 35.8.[2] In an experiment, the measured quantities $\{x_n\}$ come independently from a biexponential distribution with mean μ,

$$
P(x \,|\, \mu) = \frac{1}{Z} \exp(-|x - \mu|),
$$

where Z is the normalizing constant, $Z = 2$. The mean μ is not known. An example of this distribution, with $\mu = 1$, is shown in figure 35.2.

Assuming the four datapoints are

$$\{x_n\} = \{0, 0.9, 2, 6\},$$

what do these data tell us about μ? Include detailed sketches in your answer. Give a range of plausible values of μ.

Figure 35.2. The biexponential distribution $P(x \mid \mu = 1)$.
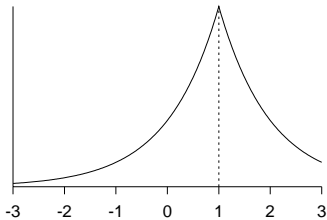
▶ **35.5 Solutions**

Solution to exercise 35.4 (p.446). A population of size N has N opportunities to mutate. The probability of the number of mutations that occurred, r, is roughly Poisson

$$P(r \mid a, N) = e^{-aN} \frac{(aN)^r}{r!}. \tag{35.7}$$

(This is slightly inaccurate because the descendants of a mutant cannot themselves undergo the same mutation.) Each mutation gives rise to a number of final mutant cells n_i that depends on the generation time of the mutation. If multiplication went like clockwork then the probability of n_i being 1 would be $1/2$, the probability of 2 would be $1/4$, the probability of 4 would be $1/8$, and $P(n_i) = 1/(2n)$ for all n_i that are powers of two. But we don't expect the mutant progeny to divide in exact synchrony, and we don't know the precise timing of the end of the experiment compared to the division times. A smoothed version of this distribution that permits all integers to occur is

$$P(n_i) = \frac{1}{Z} \frac{1}{n_i^2}, \tag{35.8}$$

where $Z = \pi^2/6 = 1.645$. [This distribution's moments are all wrong, since n_i can never exceed N, but who cares about moments? – only sampling theory statisticians who are barking up the wrong tree, constructing 'unbiased estimators' such as $\hat{a} = (\bar{n}/N)/\log N$. The error that we introduce in the likelihood function by using the approximation to $P(n_i)$ is negligible.]

The observed number of mutants n is the sum

$$n = \sum_{i=1}^{r} n_i. \tag{35.9}$$

The probability distribution of n given r is the convolution of r identical distributions of the form (35.8). For example,

$$P(n \mid r = 2) = \sum_{n_1=1}^{n-1} \frac{1}{Z^2} \frac{1}{n_1^2} \frac{1}{(n - n_1)^2} \quad \text{for } n \geq 2. \tag{35.10}$$

The probability distribution of n given a, which is what we need for the Bayesian inference, is given by summing over r.

$$P(n \mid a) = \sum_{r=0}^{N} P(n \mid r) P(r \mid a, N). \tag{35.11}$$

This quantity can't be evaluated analytically, but for small a, it's easy to evaluate to any desired numerical precision by explicitly summing over r from

$r = 0$ to some $r_{\max}$, with $P(n \,|\, r)$ also being found for each r by $r_{\max}$ explicit convolutions for all required values of n; if $r_{\max} = n_{\max}$, the largest value of n encountered in the data, then $P(n \,|\, a)$ is computed exactly; but for this question's data, $r_{\max} = 9$ is plenty for an accurate result; I used $r_{\max} = 74$ to make the graphs in figure 35.3. `Octave` source code is available.[1]
Incidentally, for data sets like the one in this exercise, which have a substantial number of zero counts, very little is lost by making Luria and Delbruck's second approximation, which is to retain only the count of how many n were equal to zero, and how many were non-zero. The likelihood function found using this weakened data set,

$$L(a) = (e^{-aN})^{11}(1 - e^{-aN})^9, \qquad (35.12)$$

is scarcely distinguishable from the likelihood computed using full information.

Solution to exercise 35.5 (p.447).   From the six terms of the form

$$P(\mathbf{F} \,|\, \alpha\mathbf{m}) = \frac{\prod_i \Gamma(F_i + \alpha m_i)}{\Gamma(\sum_i F_i + \alpha)} \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha m_i)}, \qquad (35.13)$$

most factors cancel and all that remains is

$$\frac{P(\mathcal{H}_{A \to B} \,|\, \mathrm{Data})}{P(\mathcal{H}_{B \to A} \,|\, \mathrm{Data})} = \frac{(765+1)(235+1)}{(950+1)(50+1)} = \frac{3.8}{1}. \qquad (35.14)$$

There is modest evidence in favour of $\mathcal{H}_{A \to B}$ because the three probabilities inferred for that hypothesis (roughly 0.95, 0.8, and 0.1) are more typical of the prior than are the three probabilities inferred for the other (0.24, 0.008, and 0.19). This statement sounds absurd if we think of the priors as 'uniform' over the three probabilities – surely, under a uniform prior, any settings of the probabilities are equally probable? But in the natural basis, the logit basis, the prior is proportional to $p(1-p)$, and the posterior probability ratio can be estimated by

$$\frac{0.95 \times 0.05 \times 0.8 \times 0.2 \times 0.1 \times 0.9}{0.24 \times 0.76 \times 0.008 \times 0.992 \times 0.19 \times 0.81} \simeq \frac{3}{1}, \qquad (35.15)$$

which is not exactly right, but it does illustrate where the preference for $A \to B$ is coming from.
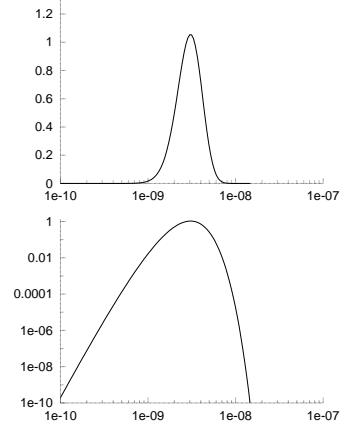

Figure 35.3. Likelihood of the mutation rate a on a linear scale and log scale, given Luria and Delbruck's data. Vertical axis: likelihood/10^{-23}; horizontal axis: a.

---

[1] `www.inference.phy.cam.ac.uk/itprnn/code/octave/luria0.m`