

## Probability and statistics exercises

**D.1** Balls are thrown independently at random into  $A$  bins. Each ball has an equal probability,  $1/A$ , of going into each bin. Call the number of balls thrown  $N$ . How big must  $N$  be, for it to be likely that *every bin has at least one ball*?

(In the special case  $A = 365$ , this question is equivalent to: ‘how big a group of people do we need, if we want every day of the year to be the birthday of someone in the group?’)

**D.2** *Thinking about paradoxes is a great way to understand a subject. Here is one of my favourite probability paradoxes, related to Waiting Times.*

Fred rolls an unbiased six-sided die once per second, noting the occasions when the outcome is a six.

- What is the mean number of rolls from one six to the next six?
- Between two rolls, the clock strikes one. (That is, the time is exactly 1.00pm.) What is the mean number of rolls from 1.00pm until the next six?
- Now think back before 1.00pm. What is the mean number of rolls, *going back in time*, until the most recent six?
- What is the mean number of rolls from the six before 1.00pm to the next six?
- Is your answer to (d) different from your answer to (a)? Explain. Make a simulation to check your explanation.

## Bayesian inference

**D.3** Imagine that a coin is selected and tossed. We believe that the coin is *either* a fair coin ( $f = 0.5$ ), *or* a biased coin having probability  $f = 0.1$  of producing heads. (As usual, Heads is called ‘1’ and tails 0.)

Call these two hypotheses  $H_0$  ( $f = 0.5$ ) and  $H_1$  ( $f = 0.1$ ). Assume that the prior probabilities of these two hypotheses are equal. Now, gradually, the data arrive. (The data are a sequence of outcomes  $x_1x_2x_3\dots$ ) For each of the sequences below, compute the posterior probability of the two hypotheses.

- $x_1x_2x_3 = 000$
- $x_1x_2x_3x_4 = 0001$
- $x_1x_2x_3x_4\dots x_9x_{10} = 0001000000$

**D.4** Box 1 contains 1 red ball and 3 black balls. Box 2 contains 1 red ball, 1 white ball, and 1 black ball. Box 3 contains 1 red ball and 1 black ball.

- A box is chosen at random and one ball is drawn. What is the probability that the ball is red?
- Given that the ball is red, what are the probabilities that the chosen box was box  $k$ , for  $k = 1, 2, 3$ ?

**D.5** Suppose that it is known that 1% of the population have a nasty disease. There is a test for this disease; the outcome of the test is either ‘positive’ or ‘negative’, and the test is known to be 98% reliable. That is, the test outcome is positive in 98% of people who have the disease, and the outcome is positive in 2% of the people who do not have it. Joe takes the test, and the outcome is positive. What is the probability that Joe has got the disease?

**D.6** A bag contains four balls: two red, and two black. Fred draws a ball from the bag (and doesn't put it back), then George draws a ball. (a) What is the probability that George's ball is red? (b) What is the probability that George's ball is red, *given that* Fred's ball is red? (c) What is the probability that Fred's ball is red, given that George's is red?

## Sampling theory

Many statisticians focus on an approach to statistics in which we *process* the data, creating objects called *estimators* and *confidence intervals*, or applying *hypothesis tests*.

Having invented these ways of processing the data, these statisticians then make probability statements about the properties of their estimators, confidence intervals, or tests.

This way of doing things (invent an estimator, then describe its properties) is called **sampling theory**. It is different from Bayesian inference. Bayesian inference does not require us to invent estimators. In Bayesian inference, we write down our assumptions, write down the data, then infer what we want to know, using the rules of probability.

Nevertheless, it is important to understand sampling theory, because so many people talk 'sampling theory' language. For answers to the following questions, see Riley (QA401.R54X) or Hodge and Seed (QA273.H655X).

**D.7** Let a source of random variables have probability distribution  $P(x)$ , which has mean  $\mu \equiv \int dx P(x)x$  and standard deviation  $\sigma$ . (This distribution may be either discrete or continuous.)

Now  $N$  points are going to be drawn from this distribution. The points are called  $x_1, \dots, x_N$ . We define the estimator

$$\bar{x} \equiv \frac{1}{N} \sum_{n=1}^N x_n.$$

This estimator is known as the **sample mean**. It is an estimator of  $\mu$ .

What is the expected value of  $\bar{x}$ ? What is the standard deviation of  $\bar{x}$ ?

**D.8** Now assume that you know  $\mu$ . We define the estimator

$$\hat{\sigma}_0^2 \equiv \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2.$$

What is the expected value of  $\hat{\sigma}_0^2$ ?

**D.9** Now assume that you do not know  $\mu$ . We define the estimator

$$\hat{\sigma}_N^2 \equiv \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2,$$

where  $\bar{x}$  was defined above. What is the expected value of  $\hat{\sigma}_N^2$ ?

Explain why calculators and spreadsheets provide another estimator,  $\hat{\sigma}_{N-1}^2$ , defined by:

$$\hat{\sigma}_{N-1}^2 \equiv \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2.$$