Chapter 2

Statistical models, sampling and sampling distributions

2.1 Objectives

1. To understand some important probability distributions—discrete and continuous.
2. To formulate statistical models for common situations and state their assumptions.
3. To appreciate the fundamentals of sampling and simulation.
4. To calculate the sampling distribution of the sample mean.

2.2 Distributions

Statistics provides models to describe variation and uncertainty. These models are founded on probability distributions, and provide useful insights into underlying patterns.

For example, let's think about what it means to say that some system is binomially-distributed. What we really mean is that we have a sample that comes from a binomial 'process', which has the following properties:

- each observation can be in exactly one of two states,
- the state of an observation is independent of all of the other observations, and
- the propensity of an observation to be in one or other of the states is the same for each observation.

For example, we might think about the tossing of a coin. If the coin is fair, and the coin-tosser is impartial, then a sequence of coin tosses could be thought of as a sequence of binomial observations (actually Bernoulli, which is the binomial distribution with $n = 1$). We might be interested in the number of tails tossed, for example. Let's invent a random variable, X, that represents the number of tails from a sequence of 10 tosses of a fair coin. We would write:

$$X \overset{\mathrm{d}}{=} \mathrm{Bi}(n = 10, p = 0.5)$$

We can ask ourselves what those outcomes could possibly look like. The top-left panel of Figure 2.1 has the "probability mass function" of the tail-count of 10 tosses of a fair coin.

In many situations we don't know the parameter, p. We might be comfortable with the idea that the broad form of the distribution is correct, but not clear on what specific values some of the key quantities take. This speaks to the challenge of modelling the distribution of the data, which generally takes the following steps:

- Formulate a possible model;
- Estimate the unknown parameters;
- Check the assumptions of the model.

In this lab we will be thinking about and formulating models. In order to help us do this, we need to know more about randomness, and the tools that are used to manipulate and understand it.

R provides functions for manipulating and examining many popular discrete probability distributions (Figure 2.1).

```
> curve(dbinom(x, p = 0.5, size = 10), from = 0, to = 10, type = "s",
+     main = "Binomial")
> curve(dgeom(x, prob = 0.2), from = 0, to = 10, type = "s", main = "Geometric")
> curve(dpois(x, lambda = 3), from = 0, to = 10, type = "s", main = "Poisson")
> curve(dnbinom(x, size = 5, prob = 0.75), from = 0, to = 10, type = "s",
+     main = "Negative Binomial")
```

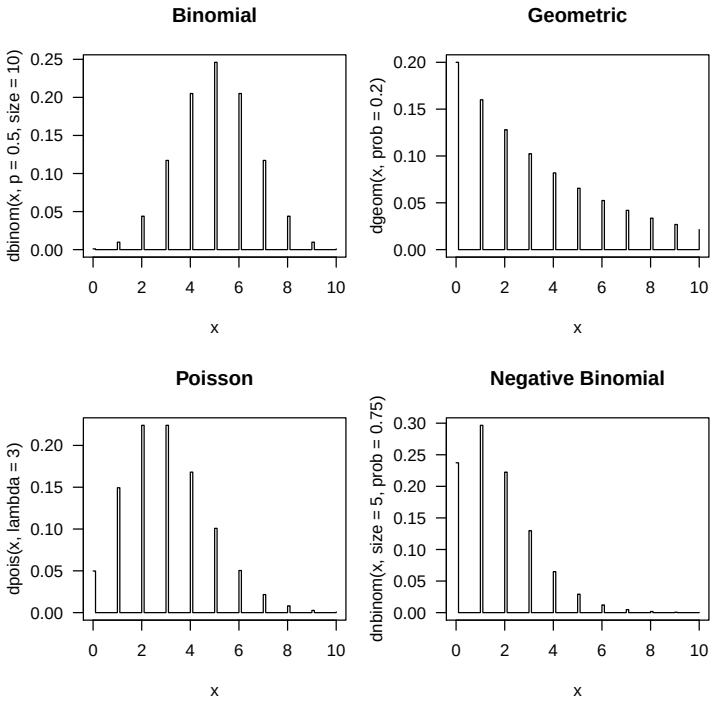

Figure 2.1: Popular probability mass functions (discrete).

In the `curve` function used above, the `type` argument gives the type of plot desired, e.g. whether to have just points (`type="p"`), lines(`type="l"`) or, as in this case, "stair steps" (`type="s"`).

R also provides functions for manipulating and examining many popular continuous probability distributions (Figure 2.2).

```
> curve(dnorm, from = -3, to = 3, main = "Normal")
> curve(dt(x, df = 5), from = -3, to = 3, main = "t (5 df)")
> curve(dchisq(x, df = 10), from = 0.01, to = 30, main = "Chi-squared (10 df)")
> curve(dgamma(x, shape = 2), from = 0.01, to = 8, main = "Gamma")
```

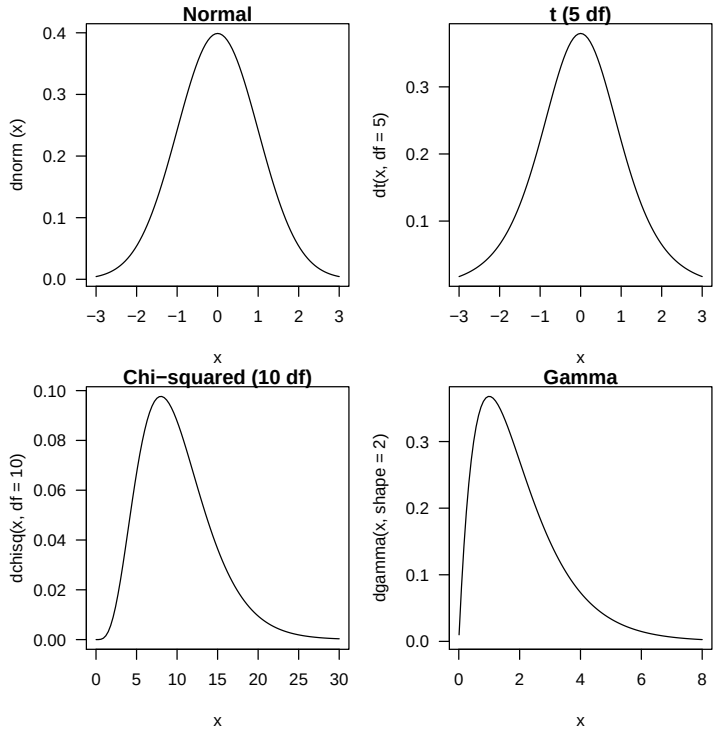

Figure 2.2: Popular probability density functions (continuous).

2.2.1 Seed

R uses pseudo-random-number generators that require an initial state. We can choose this state using `set.seed` in the following way:

```
> set.seed(83446410)
```

We can also let the computer choose the seed. The advantage of setting the seed ourselves is that we can always reproduce our results, even if they involve random numbers.

2.3 Statistical Models

In this section we begin the formal modelling process for situations that involve a numerical response variable and categorical or numerical explanatory variables. While formal treatment will be limited to one numerical or one categorical explanatory variable, an informal treatment of some more complex situations will also be presented. Previously we introduced the idea of modelling the *distribution* of a random variable, here we extend the idea to *statistical* models which enable us to allow for the *non-random* effects of the explanatory variable(s).

A model is an artificial construction that has the same essential properties as the thing being modelled, and can be used to understand and predict its behaviour. For instance, a model plane can be tested in a wind tunnel to study the properties of the real plane and predict how it will fly. Similarly, a statistical model is a mathematical description of the response variable, and can be used to describe, understand and predict the response. In both cases, there are good, bad and in-between models.

It is useful to think in terms of four stages of the modelling process (three of which we listed for modelling distributions):

1. Formulate a possible model;
2. Estimate any unknown parameters;
3. Check the assumptions of the model;
4. Use the model to estimate quantities of interest from the parameter estimates, and the uncertainty associated with these quantities—this is the heart of statistical inference.

These stages need not be done strictly in order. For instance, if at the third stage the assumptions of a model are found to be unrealistic, then we may return to the first stage and try again. It is also possible to have several similar models that seem equally good (like having similar model planes).

In this chapter we study some general concepts of modelling and use some simple statistical models to illustrate the basic ideas. We consider mainly stage 1 above, with a bit of stage 2. In later chapters we consider methods of inference (stage 4) and examination of assumptions (stage 3). This general approach includes many standard methods like t-tests and confidence intervals as special cases.

2.4 Formulating statistical models

A statistical model generally has two parts: a deterministic function that describes the patterns in the response in terms of explanatory variables, and a random part that describes the variation in the response that we have not been able to explain using the explanatory variables. The models we consider take the form

$$\textit{response} = \textit{deterministic function} + \textit{random error} \qquad (\textit{response} \sim \textit{det.fn})$$

The term `error` here is not saying that we have made a mistake. It is statistical terminology for the variation that we haven't been able to explain by the deterministic part of the model. Note that the errors can be positive or negative.

The deterministic part of the model is often fairly simple, perhaps a linear or quadratic function (straight or curved line fitted to the scatterplot), or different means for different groups or categories. For the time being, the random part will be assumed to be adequately described by a normal distribution.

To formulate a model, first examine the structure of the data (type of response and type of explanatory variables). Use graphs and summary statistics to study the patterns and try to write down equations that describe the patterns. We now illustrate this process with the three examples above. These three cases are very common, and also provide building blocks for many other situations.

2.4.1 No explanatory variables (constant function)

▷ **Example. Soil pH.** The pH level of the soil in a field was tested by taking samples from 17 scattered locations in the field. The pH readings were as follows:

```
> pH <- c(6, 5.7, 6.2, 6.3, 6.5, 6.4, 6.9, 6.6, 6.8, 6.7, 6.8,
+     7.1, 6.8, 7.1, 7.1, 7.5, 7)
```

Let's consider a simple model for soil pH. The soil pH values vary, but we have no information to explain this. A histogram is (Figure 2.3):

```
> hist(pH)
```

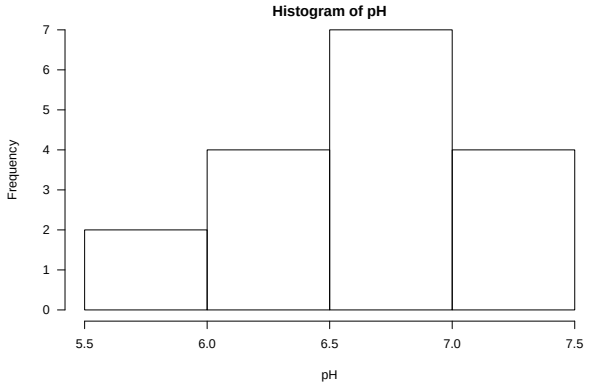

Figure 2.3: Histogram of the soil pH dataset.

The data are centred somewhere between 6.5 and 6.8, and are spread out in a way that could be like a normal distribution. If we were to try to predict the pH of other samples from the field, we would probably use the very simple model:

$$\begin{aligned} \text{pH} &= \text{mean} + error \\ y_i &= \mu + e_i \end{aligned}$$

for the ith soil sample, where $error = e_1, e_2, \ldots, e_n$ represent separate draws from a N(0, σ) distribution, i.e. a normal distribution with mean 0 and standard deviation σ. (Note that sometimes the distribution is denoted as N(0, σ^2) to state the variance σ^2.)

Make some educated guesses at these values:

$\hat{\mu} \approx$ _____ $\hat{\sigma} \approx$ _____

The hat indicates that the quantity has been estimated from the data. μ is not too hard to guess, and σ is a bit harder. A reasonable first guess at $\hat{\sigma}$ is one quarter of the range. (Why is that so?). More accurate estimates are given in the solutions to the lab exercises.

Looking ahead, this model may be too simple because the pH values were taken systematically across the field, and a pattern may emerge.

2.4.2 One numerical explanatory variable

▷ **Example. Fuel economy of cars.** The following data were extracted from the RACV web-site. Fuel consumption is measured in litres per 100 kilometres.

```
> cars <- read.csv("../data/racv.csv")
> cars
```

```
                         Make lp100km mass.kg List.price
1                  Alpha Romeo     9.5    1242      38500
2                      Audi A3     8.8    1160      38700
3            BA Falcon Futura    12.9    1692      37750
4  Chrysler PT Cruiser Classic     9.8    1412      33400
5         Commodore VY Acclaim    12.3    1558      37510
6           Falcon AU II Futura   11.4    1545      34860
7                 Holden Barina     7.3    1062      13990
8                 Hyundai Getz      6.9     980      13990
9               Hyundai LaVita      8.9    1248      23990
10                     Kia Rio     7.3    1064      14990
11                     Mazda 2     7.9    1068      17790
12                Mazda Premacy    10.2    1308      27690
13                  Mini Cooper     8.3    1050      32650
14   Mitsubishi Magna Advance     10.9    1491      33990
15       Mitsubishi Verada AWD    12.4    1643      46460
16                  Peugeot 307     9.1    1219      31490
17                Suzuki Liana      8.3    1140      19990
18            Toyota Avalon CSX    10.8    1520      34490
19       Toyota Camry Ateva V6     11.5    1505      35990
20       Toyota Corolla Ascent     7.9    1103      21790
21     Toyota Corolla Conquest     7.8    1081      23590
```

Much of the variation in the fuel consumption can be explained by differences in car weights (Figure 2.4).

```
> plot(lp100km ~ mass.kg, data = cars)
```

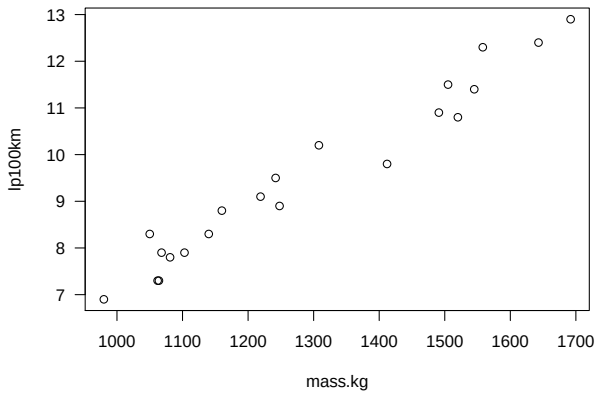

Figure 2.4: Scatterplot of fuel consumption of cars.

A sensible way to use weight to explain the variation in fuel consumption is a linear model:

$$\begin{aligned} \text{fuel consumption} &= \alpha + \beta \times weight + error \\ y_i &= \alpha + \beta x_i + e_i \end{aligned}$$

for the ith car, where e_i again represents separate draws from a $N(0, \sigma)$ distribution. This model is called *linear regression*.

Here, there are two parameters to estimate (excluding σ), the slope or gradient β and the intercept α. Make some educated guesses at these parameter estimates:

$\hat{\beta} \approx$ _____ $\hat{\alpha} \approx$ _____

The interpretation of the slope is the change in the response that is associated with one unit change in the explanatory variable (here, the increase in fuel consumption (l/100 km) for each unit (1kg) increase in weight). The intercept is the estimated response when $x = 0$. The intercept doesn't mean much here (the fuel consumption of a car that has zero weight) but it is necessary to define the line.

The weight of the cars in the data set range from just under 1000 kg to just under 1700 kg. Extrapolating from these data to cars with weights outside this range cannot be justified by the data alone; it is an act of faith that the trends continue.

2.4.3 One categorical explanatory variable

▷ **Example. Potato yield.** The following data were obtained from an experiment to investigate the effect of inorganic and organic fertilizers on the yield of potatoes.

Variety: King Edward
Area of each plot: 0.02 ha

Treatments:
1 blood + superphosphate
2 sulphate of ammonia + superphosphate
3 blood + steamed bone flour
4 sulphate of ammonia + steamed bone flour

Crop Weight in kilograms

Treatment				
1	752	762	686	787
2	621	637	670	575
3	642	667	655	660
4	645	627	596	576

```
> potatoes <- data.frame(treatment=rep(1:4,each=4),
+                        wt.kg=c(752, 762,  686, 787,
+                          621, 637,  670, 575,
+                          642, 667,  655, 660,
+                          645, 627,  596, 576))
```

The yield of potatoes varies from plot to plot, and we can use the fertilizer information (a categorical explanatory variable or factor) to try to explain some of this variation (Figure 2.5).

```
> plot(wt.kg ~ treatment, data = potatoes)
```

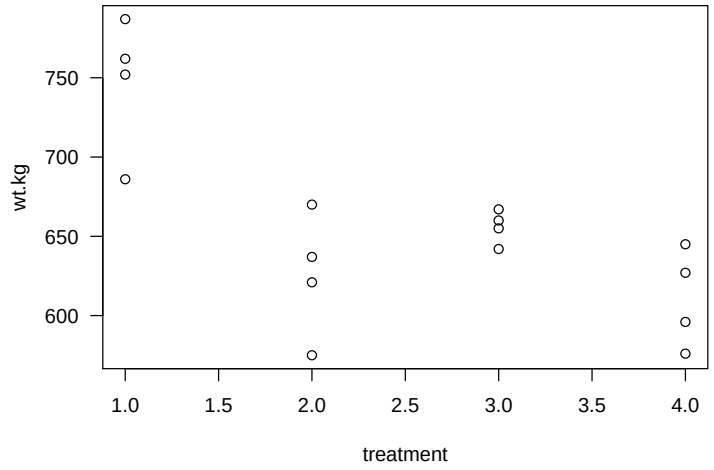

Figure 2.5: Scatterplot of treatment effects on harvest mass of potatoes.

The labelling of the x-axis of this dotplot is not very appealing, because it makes the fertilizer treatments look like a numerical variable. The graph, however, is effective in showing both the location and spread of points for each treatment. (Tidying up the x-axis could be done with a couple of lines of code.)

The treatments do seem to influence yield; in particular, the four highest yields were obtained with treatment 1. With so few observations per treatment it is difficult to say what sort of distribution might be appropriate, but there is no indication that a normal distribution is not appropriate. From the dotplot it does appear that the within treatment standard deviations may not all be the same; the spread of yields for treatments 1, 2 and 4 is considerably greater than that of the yields for treatment 3. However, a formal comparison of these spreads indicates that the differences could be just due to sampling variation. So a simple model for these data is:

$$\begin{aligned} \text{Potato yield} &= \text{mean for treatment used} + error \\ y_{ij} &= \mu_i + e_{ij} \end{aligned}$$

for the jth plot ($j = 1, 2, 3, 4$) using treatment i ($i = 1, 2, 3, 4$), where e_{ij} again represents separate draws from a N(0, σ) distribution. Note that we are assuming that each treatment may have a different mean yield but the *same* standard deviation σ. Testing this model against the simpler model of each treatment having the same mean is called *analysis of variance.*

Again, we can make some educated guesses:

$\hat{\mu}_1 \approx$ _____ $\hat{\mu}_2 \approx$ _____

$\hat{\mu}_3 \approx$ _____ $\hat{\mu}_4 \approx$ _____ $\hat{\sigma} \approx$ _____

There are four parameters to estimate in the model for the mean (i.e. excluding σ).

2.5 Simulation

R provides powerful functionality for randomly generating data. Each of the probability density functions that we have seen comes with a corresponding random-number generator, where the distribution of the process of the numbers so generated matches the nominated distribution. For example, to generate 10 standard normally-distributed numbers, use

```
> rnorm(10)

 [1]  1.35117063 -0.37641102  0.30839344  1.56904501 -0.94470867  1.74037888
 [7]  0.06946446 -0.03106227  0.95931670  0.93045383
```

If we wished to make a population of 1000 normally-distributed numbers, for later sampling, we could use, for example,

```
> normal.pop <- rnorm(1000)
```

▷ **Example.** Verify its shape using

```
> qqnorm(normal.pop)
```

and

```
> plot(density(normal.pop))
```

If the data come from a normal distribution, the first plot should be close to a straight line, and the second should approximate the characteristic bell-shaped curve.

2.6 Sampling

R provides powerful functionality for randomly sampling populations.

▷ **Example.** Have a look at the `help` for the `sample` function with

```
> ?sample
```

We can select a sample of 10 of our normally-distributed data points (without replacement) by

```
> sample(normal.pop, size = 10)

 [1] -0.21588803 -1.14363446 -0.37756562  1.41268311  1.21080235 -0.45243492
 [7]  1.07990527 -0.27537916  0.02300605 -0.15064564
```

If we want to do the sample thing with a more skewed distribution, such as the exponential, then we can use

```
> exponential.pop <- rexp(1000, rate = 1)
> exponential.sample <- sample(exponential.pop, size = 10)
```

We verify our expectation of the shapes of these datasets in Figure 2.6.

```
> plot(density(exponential.pop), main = "Exponential Population")
> qqnorm(exponential.pop, main = "")
> qqline(exponential.pop)
> plot(density(exponential.sample), main = "Just a sample")
> qqnorm(exponential.sample, main = "")
> qqline(exponential.sample)
```

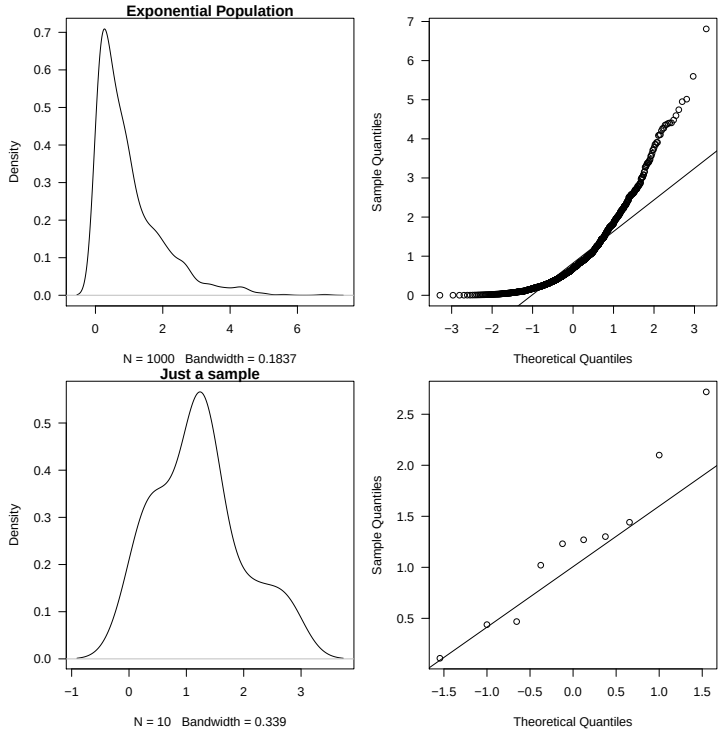

Figure 2.6: Checking the shapes of our exponential data objects.

Not surprisingly, the generated data are not consistent with having come from a normal distribution. Note that the `main` argument is for entering a main title of the graph if required.

2.7 Sampling distributions

2.7.1 The sampling distribution of the sample mean

Let's focus for the moment on the sampling distribution of the sample mean. We will take samples of size 15 from our exponential population, and compute the mean, many times. We could do this operation with a `for` loop, but in R it is more efficient to use the `sapply` function which applies the desired function, in this case the mean, to many samples.

```
> exp.means <- sapply(1:1000,
+       function(x) mean(sample(exponential.pop, size=15)))
> exp.means[1:10]

 [1] 1.1120348 0.7723983 1.2030199 1.4515720 1.3943345 1.0255221 0.9617335
 [8] 0.9893440 0.6881620 0.9554790
```

Each of the these values (we have only listed the first 10 out of 1000) is the mean of a random sample of size 15 taken from an exponential distribution. We can now examine the distribution of the sample means (Figure 2.7). This is a graphical representation of the *sampling distribution of the sample mean*.

```
> plot(density(exp.means))
```

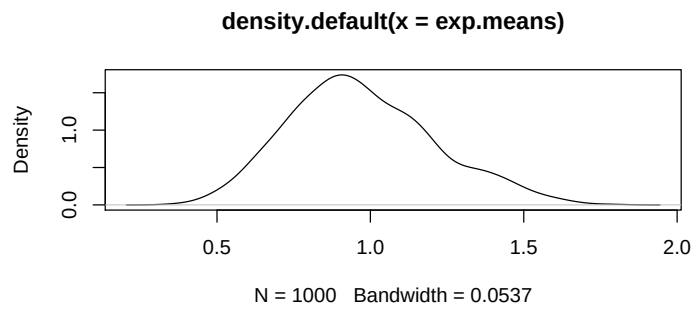

Figure 2.7: Sampling distribution of sample mean from samples of size 15, extracted from the exponential distribution with rate 1.

This looks more normally distributed than the density of the individual values examined in Figure 2.6. An obvious next step is to try it with samples of a larger size. Run the R code again with `size=50` and examine the sampling distribution of the sample mean.

We will now extend this code to compare the sampling distributions for a range of sizes.

```
> levels <- rep(c(1, 10, 20, 40, 80, 160), each = 1000)
> sample.means <- sapply(levels, function(x) mean(sample(exponential.pop,
+     size = x)))
```

Using a box plot we can get some idea of how well the sampling distributions for the six sample sizes converge to a symmetrical distribution. To further examine the normality we would need to examine each size individually.

```
> boxplot(sample.means ~ levels, xlab = "n", ylab = expression(paste(bar(x))))
> abline(h = 1, col = "darkgrey")
```

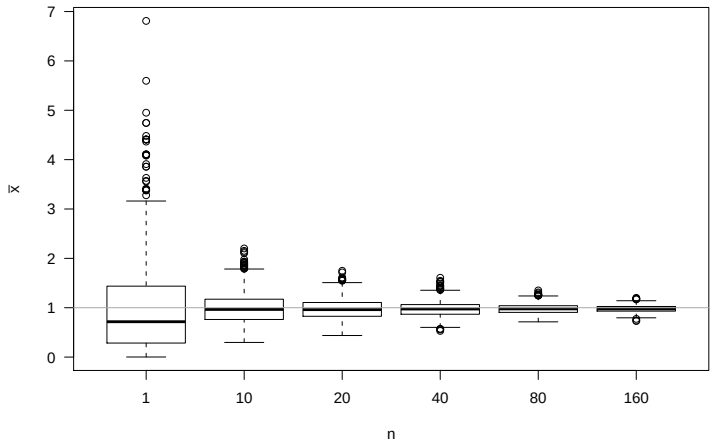

Figure 2.8: Simulated sampling distributions for the mean of data from the exponential distribution with a variety of sample sizes.

Note that the command `abline()` adds a line to an existing plot.

▷ **Example.** What does the sampling distribution of the variance look like for sample sizes of 15 and 50?

In R, the sample variance is calculated using the function `var()`.

2.8 Exercises

1. The following observations are salinity values for water specimens obtained from a certain region:

```
> salinity <- c(9.3, 10.7, 5.5, 9.6, 12.2, 16.6, 9.2, 10.5, 7.9,
+     13.2, 11, 8.8, 13.7, 12.1, 9.8)
```

   Display these data appropriately and write a suitable model. Identify the parameters to be estimated and calculate the estimates of these parameters.

2. A pollution-control inspector suspected that a riverside community was releasing semitreated sewage into a river and this, as a consequence, was changing the level of dissolved oxygen of the river. To check this, he drew 15 randomly selected specimens of river water at a location above the town and another 15 specimens below. The dissolved oxygen readings, in parts per million, are given below:

Above town	5.2	4.8	5.1	5.0	4.9	4.8	5.0	4.7	4.7	5.0	4.7	5.1	5.0	4.9	4.9
Below town	4.2	4.4	4.7	4.9	4.6	4.8	4.9	4.6	5.1	4.3	5.5	4.7	4.9	4.8	4.9

```
> pollution <- data.frame(location = rep(c("Above", "Below"), each = 15),
+     d.o. = c(5.2, 4.8, 5.1, 5, 4.9, 4.8, 5, 4.7, 4.7, 5, 4.7,
+         5.1, 5, 4.9, 4.9, 4.2, 4.4, 4.7, 4.9, 4.6, 4.8, 4.9,
+         4.6, 5.1, 4.3, 5.5, 4.7, 4.9, 4.8, 4.9))
```

   Display these data appropriately and write a suitable model. Identify the parameters to be estimated and the assumptions of the model. Estimate the parameters excluding σ. How would you estimate σ?

For exercises 3 to 6, develop an appropriate statistical model, including the necessary subscripts (i, j, etc.) You can use symbols (e.g. y_{ij}) or words (e.g. $weight_{ij}$) to denote the variables in the model.

3. Can aspirin help heart attacks? The Physician's Health Study, a large medical experiment involving 22000 male physicians, attempted to answer this question. One group of 11000 physicians took an aspirin every second day, while the rest took a placebo. After several years it was found that the subjects in the aspirin group had significantly fewer heart attacks than subjects in the placebo group.

4. New varieties of corn with altered amino acid patterns may have higher nutritive value than standard corn, which is low in the amino acid lysine. An experiment compares two new varieties, called opaque-2 and floury-2, with normal corn. Corn-soybean meal diets using each variety of corn are prepared at three different protein levels: 12%, 16% and 20%. There are thus nine diets in all. Researchers assign 10 one-day-old male chicks to each diet and record their weight gain after 21 days. The weight gain of the chicks is a measure of the nutritive value of their diet.

5. A study was conducted to compare the number of tapeworms in the stomachs of sheep treated with a drug for worms against the number in those not treated. A sample of 14 worm-infected lambs was randomly divided into 2 groups. Seven were injected with the drug and the remainder were left untreated. After a 6-month period, the lambs were slaughtered and the following worm counts recorded:

								$\bar{x}$	s
Drug-treated sheep	19	43	28	50	16	33	14	29.00	13.83
Untreated sheep	40	54	26	63	21	37	39	40.00	14.67

6. Periodic measurements of salinity and water flow were taken in North Carolina's Pamlico Sound, resulting in the following data (x = water flow, y = salinity):

x	23	24	26	25	30	24	23	22	22	24	25	22	22	22	24
y	7.6	7.7	4.3	5.9	5.0	6.5	8.3	8.2	13.2	12.6	10.4	10.8	13.1	12.3	10.4

7. Find another distribution function that interests you that wasn't covered in the notes—gamma, negative binomial, Cauchy, etc. Explore the sampling distribution of the mean of that distribution in a similar way.

* Further practice in statistical models: Return to the exercises for Lab class 1. For each exercise, write an appropriate statistical model, including both the deterministic and the stochastic portions of the model.