Chapter 2: Probability and the normal distribution

1. Imagine you are rolling a fair six-sided die (with faces numbered from 1 to 6) and answer the following questions to improve your intuition about probability:

    - What is the probability of rolling a number greater than 2?
    - What is the probability of rolling a 2 or a 5?
    - If you roll the die twice, what is the probability of getting a 2 on the first roll and a 5 on the second?
    - If you roll the die twice, what is the probability of getting at least a 5 in both rolls?
    - If you roll the die three times, what is the probability of getting at least one 6?

2. Suppose you are conducting a study where you randomly assign participants a time of day to complete a survey, which could be any hour between 9 a.m. and 5 p.m. (inclusive). Assume that each time within this interval is equally likely to be chosen.

    - What type of probability distribution does this scenario represent?
    - What is the probability that a participant is assigned a time between 12 p.m. and 2 p.m.? And how would you compute this probability in R using the function `punif()`?

3. Consider a uniform distribution over the range from 0 to 10, represented as $U(0{,}10)$.

    - What is the height of this pdf between 0 and 4? (Hint: Remember that the total area under the curve must be 1.)
    - What is the probability of drawing a value less than 2 from this distribution? Use `punif()` to help with your calculation
    - If you randomly sample a value from this distribution, what is the probability of getting a value between 2 and 3?
    - If you use the command: `punif(3, min = 0, max = 10)`, what information does this give you? And if you instead use: `1 - punif(3, min = 0, max = 10)`?
    - How would you interpret `qunif(0.5, min = 0, max = 10)` in the context of this uniform distribution?

4. Suppose that the height of the population in a certain city is normally distributed, with a mean (and median) height is 170 cm, and the standard deviation of 10 cm.

    - What is the probability that a randomly selected individual in this city is taller than 170 cm?
    - Use the `qnorm()` function to find the 95th percentile of this distribution. What does this percentile represent?
    - Use the `pnorm()` function to find the probability that a randomly selected individual has a height less than 160 cm.

5. Use the `rnorm()` function to generate a set of 1000 random values from a normal distribution with a mean of 10 and a standard deviation of 2.
    - Use `mean()` and `sd()` to calculate the mean and standard deviation of your sample. Are they close to the specified values (10 and 2)?
    - Create a plot of these values, showing an histogram of the data with overlaid the normal distribution curve computed using the population parameter (10 and 2).
6. Consider the again the trees dataset, which you can load in the workspace using `data(trees)` and which contains measurements of black cherry trees. Use the variable `Girth` (despite the name, this is the diameter of the trees in inches – see `?trees` for more info).
    - Convert `Girth` from inches to centimetres.
    - Calculate the mean and standard deviation of `Girth`, as well as the number of trees in this dataset.
    - Compute the standard error of the mean (SEM).
    - Use your results to construct a 95 per cent confidence interval for the mean diameter of black cherry trees.
7. Load the built-in `sleep` dataset (using the command `data(sleep)`), which contains data on the effects of two different drugs on sleep patterns. Focus on the `extra` variable, which records extra hours of sleep for each participant.
    - Calculate the mean and standard deviation and standard error of `extra`.
    - Use the standard error to compute a 95 per cent confidence interval for the mean extra hours of sleep.
    - Interpret this confidence interval in the context of the study.
8. Use R to explore why the normal distribution is so common by simulating the central limit theorem. Start by drawing samples from a uniform distribution.
    - Use the function `runif()` to draw a single sample of 2 values from a uniform distribution between 0 and 1. Take the mean of that sample.
    - Repeat this process 1000 times, each time storing the sample mean. (Hint: use a `for` loop or `replicate()`.)
    - Plot a histogram of the 1000 sample means. What shape does this distribution have?
    - Increase the sample size from 2 to 30 and repeat the process. How does the shape of the histogram of sample means change?
    - Try increasing the sample size further, e.g. to 100. What happens to the spread and shape of the distribution of sample means?