Homework 2 - Solution

Student

```r
library(ggplot2)  # you'll need this package for the R section
```

Instructions

Use the R Markdown version of this file to complete and submit your homework. Items in **bold** require an answer. Make sure you change the author in the header to your own name.

Conceptual Questions

1. The integrity of twenty thin metal plates is tested by bending them until they break. The number of pounds per square centimeter (cm) of force required to break each plate is measured and recorded. Five of the plates break with 50 lbs/cm^2 of pressure or less, while the remaining fifteen plates break with more than 50 lbs/cm^2 of pressure.

    a) **What is the proportion of plates that break under the lower amount of pressure?**
    The proportion is 5/20 = 0.25
    b) **Is the proportion you calculated in step a) equal to the probability of a plate breaking under 50 lbs/cm^2 of pressure or less? Why or why not? If not, what would you need to do to find the exact probability?**
    No. We will never know the exact probability. Finding it would require infinitely many tests.
    c) **As more plates are tested, what will happen to the proportion of plates that break under 50 lbs/cm^2 of pressure or less? What is this phenomenon called?**
    The proportion will approach the exact probability of a plate breaking under 50 lbs/cm^2 of pressure or less. It is called the Law of Large Numbers.
    d) Assume that the probability of a plate breaking under 70 lbs/cm^2 of pressure or less is 0.66, the probability of a plate breaking under between 50 and 80 lbs/cm^2 of pressure is 0.73, and the probability of a plate breaking under between 50 and 70 lbs/cm^2 of pressure is 0.5. **What is the probability of a plate breaking under more than 80 lbs/cm^2 of pressure?** (Hint: Drawing a picture might help)

    $$\begin{gathered} P(breakpt > 80) \\ = 1 - P(breakpt < 80) \\ = 1 - (P(breakpt <= 70 or 50 < breakpt <= 80)) \\ = 1 - (P(breakpt <= 70) + P(50 < breakpt <= 80) - P(50 < breakpt <= 70)) \\ = 1 - (0.66 + 0.73 - 0.5) = 1 - 0.89 = 0.11. \end{gathered}$$

    e) Now assume that one thousand plates are tested and 108 of them break under more than 80 lbs/cm^2 of pressure. **Does this contradict your answer in part D? Why or why not?**
    No. Even though our observed proportion was slightly off from the probability in part D, by the Law of Large Numbers, if the 0.11 really is the true probability, we would arrive at this probability with an infinitely large number of tests.

Randomness in everyday life

Come up with an example of a process from your everyday life that has a random outcome. You don't need to be able to assign probabilities to the outcome of the process you pick.

1. **Describe the process. Why do you consider it random?**
   e.g, Rolling a dice, it is random since you don't necessarilly have the same result every time.

2. **What are the possible outcomes?**
   {1,2,3,4,5,6}

3. **Define two different events based on the outcome.**
   e.g, 'Having a odd number for the next roll', 'In the next roll the number is larger than 3

4. **Do you think the events you defined are independent?** You do not need to do any calculations here, but you should justify your answer using the definition of independence.
   No, the set of numbers larger than 3 and the set of odd numbers have an non-zero intersection

5. **Describe two different random variables you could define on the outcome.**
   e.g, $X = 1$ if having a odd number for the next roll, $X = 0$ otherwise

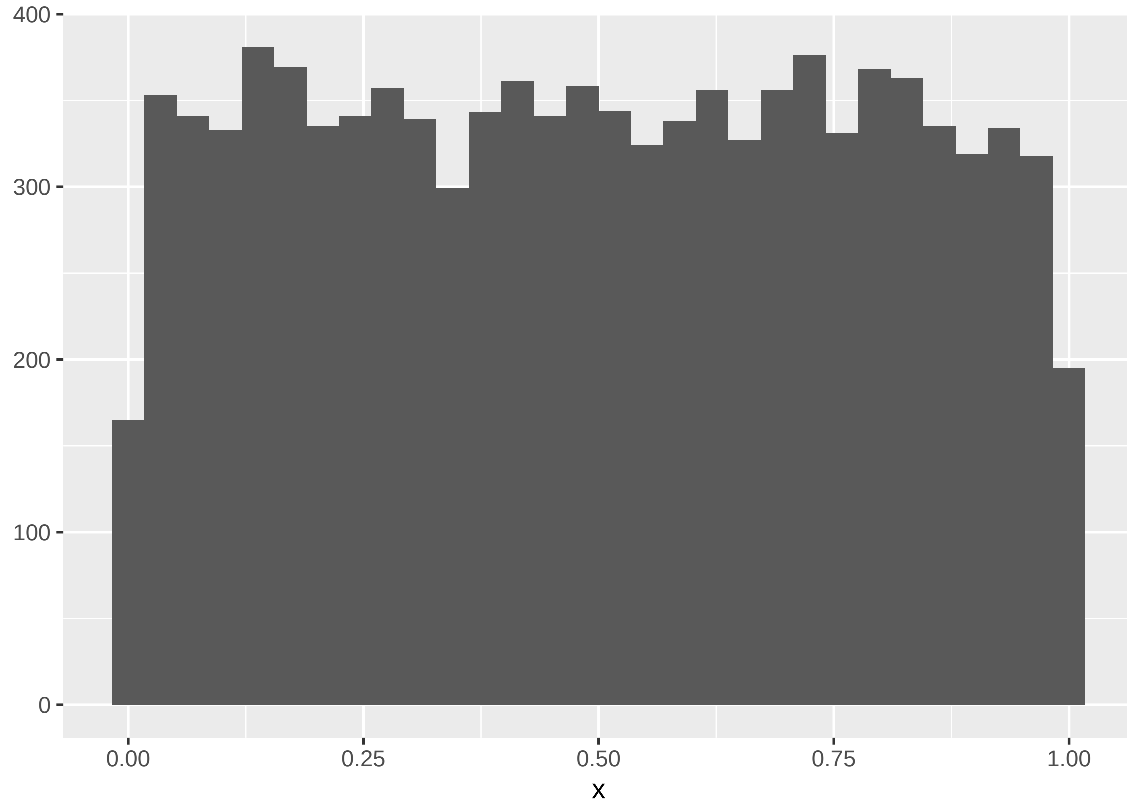
Simulation and estimation in R

1. The R function `runif()` generates random numbers from the continuous Uniform distribution. For example, `runif(1, min = 0, max = 1)` generates one random number from a Uniform(0,1) distribution, i.e. one random number between 0 and 1 with uniform probability.

   a) Create a histogram of 10000 draws from a Uniform(0, 1) distribution.

```
library(ggplot2)
x <- replicate(10000, runif(1, 0, 1))
qplot(x)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



b) Using your 10000 draws from a), estimate the probability a random variable with this distribution is between 0.5 and 0.75

```
sum(x < 0.75 & x > 0.5)/10000 # about 0.25
```

```
## [1] 0.2511
```

2. The `rbinom()` function in R simulates random variables from a Binomial Distribution. Recall a Binomial random variable is the number of successes in n trials where the probability of success is π. In `rbinom()` n is specified using the `size` argument and π by the `prob` argument. Confusingly, `rbinom()` has an argument called `n`, but this how many times we would like to simulate from a Binomial(n, π) not the value of n.

   So, for example, `rbinom(n  = 1, size = 10, prob = 0.5)` will simulate **one** random variable that has a Binomial(10, 0.5) distribution, whereas `rbinom(n = 10, size = 1, prob = 0.5)` will simulate 10 random variables each with a Binomial(1, 0.5) distribution.

   In the U.S. having a baby girl isn't as equally likely as having a baby boy, in fact if we think of "having a baby girl" as a success in a single "having a baby" trial, then the probability of success is about 0.49 (see CIA World FactBook).

   a) **Using simulation, estimate the probability that for a family who has two children both will be girls.** (Hint: Estimate $P(X = 2)$ when $X \sim \text{Binomial}(2, 0.49)$).

```
x <- rbinom(10000, size = 2, prob = 0.49) # OR
x <- replicate(10000, rbinom(1, size = 2, prob = 0.49))
sum(x == 2)/10000 # OR
```

```
## [1] 0.2347
```

```r
mean(x == 2)
```

```
## [1] 0.2347
```

b) **Using simulation, estimate the probability that for a family who has four children all are girls.**

```r
mean(rbinom(10000, size = 4, prob = 0.49) == 4)
```

```
## [1] 0.0539
```

c) **Using simulation, estimate for a family with two children one of whom is a boy, the probability that the other child is a girl.** Hint: This is tricky because it is a conditional probability. You can re-use your simulation of families with two children from part a), but extract just those where one child is a boy (i.e. $X = 0$, or 1). What proportion of those families have a boy and a girl (i.e. $X = 1$)?

```r
have_boy <- x[x == 0 | x == 1]
mean(have_boy == 1)
```

```
## [1] 0.6524239
```

d) **Why can't we estimate the probability that a randomly chosen family in the US has at least one girl? What additional information would we need to estimate this?**

If we are simulating this probability using the binomial distribution we have to know the parameter 'size', i.e. the number of children in the family. To estimate the probability that a randomly chosen family in the US has at least one girl we would need to know the distribution of the number of children in families in the US.