

Chapter 4. Probability models

- A **probability model** is an assignment of probabilities to possible outcomes.
- We don't observe these probabilities. We observe a particular dataset.
- If we treat the dataset as an outcome of a probability model, we can answer questions such as,

"If there really is no association between unemployment and life expectancy, what is the probability we would see an estimated linear model coefficient as large as the one we actually observed, due to random fluctuations in both quantities?"

- Here, we are most interested in developing a probability model for the linear model.
- First, we need some basic tools for probability models: random variables, the normal distribution, mean, variance and standard deviation.

Random variables and events

- A **random variable** X is a random number with probabilities assigned to outcomes.

Example: Let X be a roll of a fair die. A natural probability model is to assign probability of $1/6$ to each of the possible outcomes $1, 2, 3, 4, 5, 6$.

- An **event** is a set of possible outcomes.

Example: For a die, $E = \{X \geq 4\} = \{4, 5, 6\}$ is the event that the die shows 4 or more.

- We can assign probabilities to events just like to outcomes.

Example: For a die, $P(E) = P(X \geq 4) = 3/6 = 1/2$.

Question 4.1. If an experiment can be repeated many times (like rolling a die) how can you check whether the probability model is correct?

Notation for combining events

- $\{E \text{ or } F\}$ is the event that either E or F or both happens.
- Since E and F are sets, we can write this as a union, $\{E \text{ or } F\} = E \cup F$
- $\{E \text{ and } F\}$ is the event that both E and F happen.
- We can write this as an intersection,

$$\{E \text{ and } F\} = E \cap F$$

- Usually, we prefer “and/or” to “intersection/union”.

Question 4.2. When does this formal use of “and/or” agree with usual English usage? When does it disagree?

The basic rules of probability

- ① Probabilities are numbers between 0 (impossible) and 1 (certain).
- ② Let \mathcal{S} be the set of all possible outcomes. Then, $P(\mathcal{S}) = 1$.
Example: For a die, $P(X \in \{1, 2, 3, 4, 5, 6\}) = 1$.
- ③ Events E and F are called **mutually exclusive** if they cannot happen at the same time. In other words, their intersection is the empty set. In this case,

$$P(E \text{ or } F) = P(E) + P(F).$$

Question 4.3. You roll a red die and a blue die. Let

$E = \{\text{red die shows } 1\}$, $F = \{\text{blue die shows } 1\}$, $G = \{\text{red die shows } 6\}$.

(a) Are E and F mutually exclusive? (b) How about E and G ? (c) How about F and G ?

Discrete random variables

- A **discrete random variable** is one where we can list all possible outcomes. Let's call them x_1, x_2, \dots
- A discrete random variable is specified by probability that the random variable takes each possible outcome,

$$p_i = P[X = x_i], \text{ for } i = 1, 2, 3, \dots$$

- It can be helpful to plot a graph of p_i against x_i .
- This graph is called the **probability mass function**.

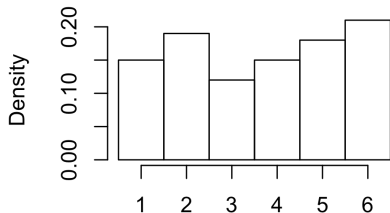
Question 4.4. Sketch the probability mass function for a fair die.

Simulating the law of large numbers

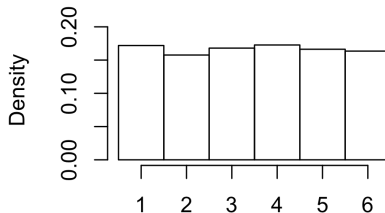
- The “law of large numbers” says that the proportion of each outcome i in a large number of draws of a discrete random variable approaches p_i .
- We can test this by simulation, using the `replicate()` command.

Worked example 4.1. In R, a random draw with replacement from $\{1, 2, 3, 4, 5, 6\}$ can be obtained by `sample(1:6, size=1)`. This is equivalent to one roll of a fair die.

```
hist(replicate(n=100, sample(1:6, size=1) ),  
     main="", prob=TRUE, breaks=0.5:6.5, xlab="n=100", ylim=c(0, 0.21))
```



n=100



n=10000