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 a least squares linear model coefficient as large as the one we actually observed, due to random fluctuations in the data?*"

- Here, we are particularly 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, $\mathrm{P}(E) = \mathrm{P}(X \geq 4) = 3/6 = 1/2$.

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" 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, $\mathrm{P}(\mathcal{S}) = 1$.
   Example: For a die, $\mathrm{P}(X \in \{1, 2, 3, 4, 5, 6\}) = 1$.
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,

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

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 $x_1, x_2, \ldots$.
- A discrete random variable is specified by probability that the random variable takes each possible outcome,

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

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

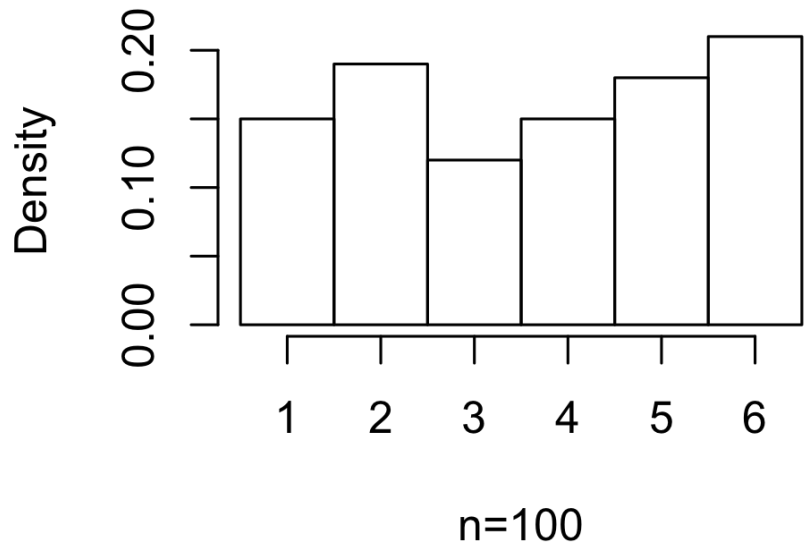
Question 4.4. Sketch the probability mass function for a fair die.

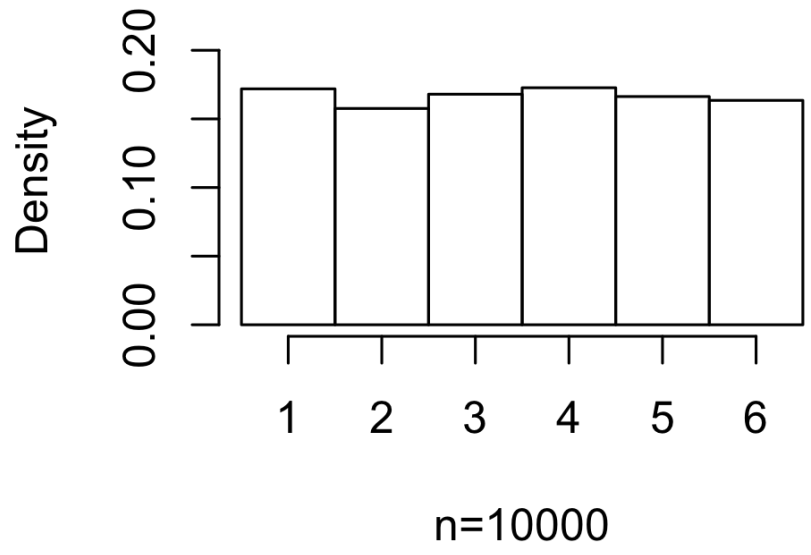
Simulating the law of large numbers

- 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))
```





Continuous random variables: the normal distribution

- A **continuous random variable** is one which can take any value in an interval of the real numbers.

  Example: physical quantities such as time and speed are not limited to a discrete set of possible values.

- We will often see the **normal distribution**.
- Let's look at **normal random variables** simulated by R using `rnorm()`.

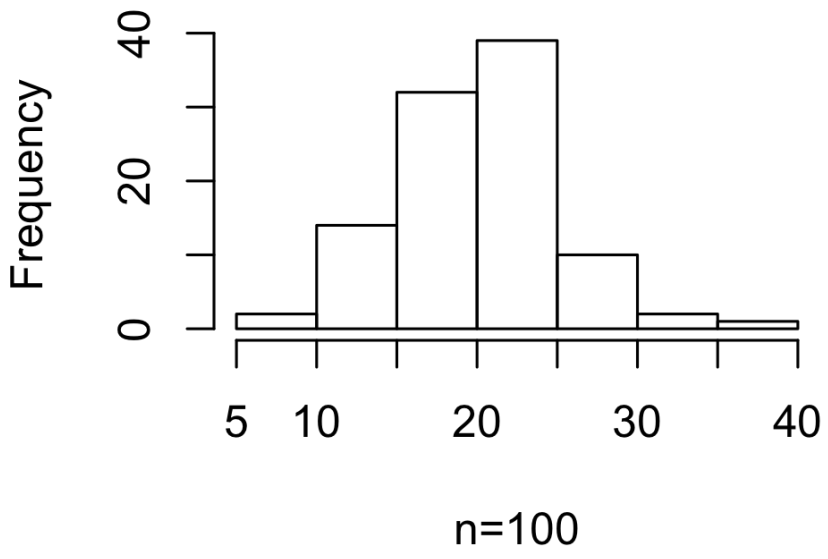
```
rnorm(n=10,mean=20,sd=5)
```

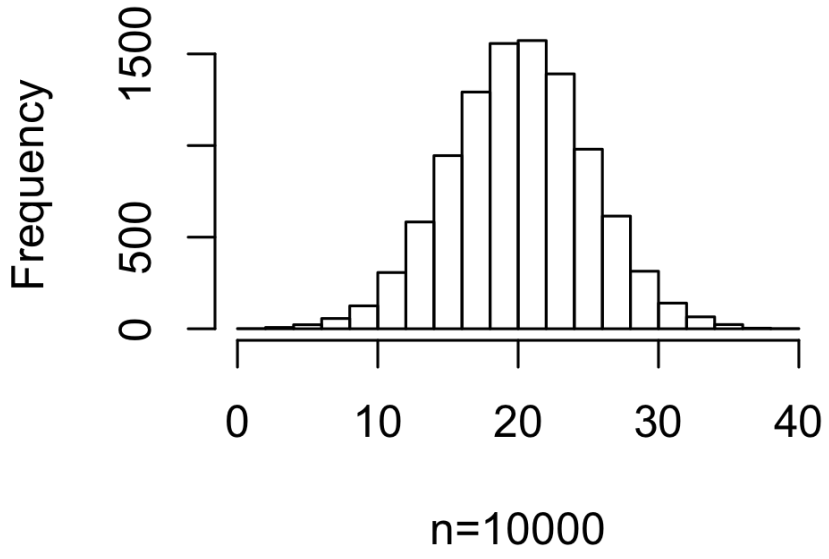
```
##  [1] 14.01141 26.18597 17.18972 21.12226 15.20211 30.34660
##  [7] 19.94277 22.05179 27.73804 25.94906
```

- The arguments `mean=20,sd=5` of `rnorm()` are the **parameters** of the normal distribution.
- A normal random variable can take any numeric value: it is continuous.
- Values are centered on the mean and are usually less than twice the standard deviation (sd) from the mean.

A histogram of normal distribution simulations

```
hist(rnorm(n=100,mean=20,sd=5),
  main="",xlab="n=100")
```





- Large samples from the normal distribution follow a **bell curve** histogram.
- From smaller samples, this is harder to see.

Finding probabilities for a continuous random variable

- A continuous random variable X has a **probability density function** $f(x)$ with

$$\mathrm{P}(a < X < b) = \int_a^b f(x)\,dx$$

- Write $X \sim \mathrm{normal}(\mu, \sigma)$ to mean X is a normal random variable with mean μ and sd σ. The probability density function of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

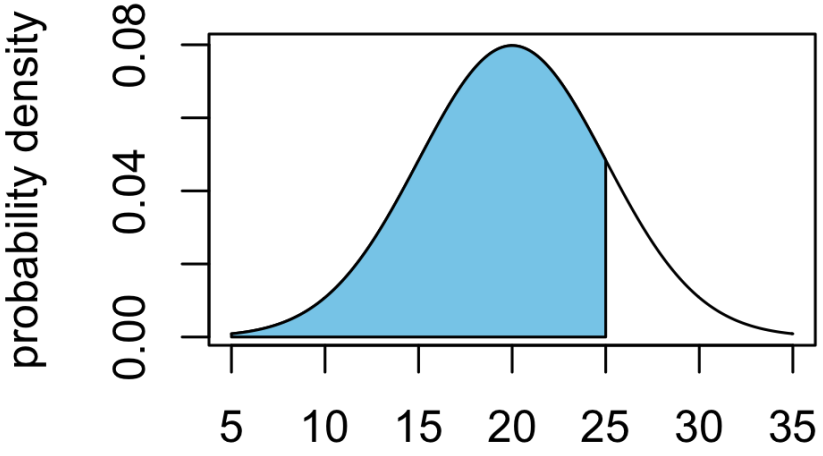
and so

$$\mathrm{P}(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\,dx$$

- This integral has no closed form solution.
- Fortunately, R provides `pnorm()` and `qnorm()` that let us work with probabilities for the normal distribution numerically.

Calculating probabilities for the normal distribution

- `pnorm()` finds the **left tail** of the normal distribution.



Example: `pnorm(25,mean=20,sd=5)` computes the shaded area above.

- We don't have to do calculus with the normal integral, but we do use the relationship between the curve, area under the curve, and probability. And we must know how to compute these things in R.
- For $X \sim \text{normal}(\mu, \sigma)$,

$$\texttt{pnorm(x,mu,sigma)} = \text{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}\, dy$$

Finding probabilities that are not a left tail

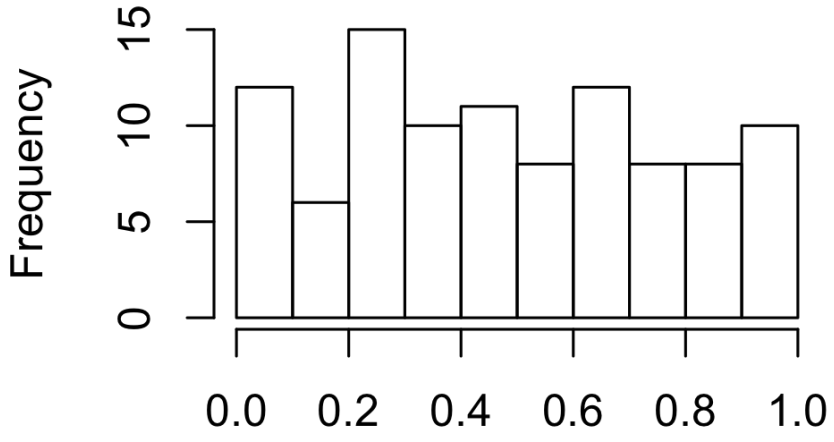
Question 4.5. Let $X \sim \text{normal}(10, 10)$. Sketch a shaded area under a curve giving $\text{P}(0 \leq X \leq 10)$. Write this probability as an integral and as R code.

Hint: To use `pnorm()`, think of the shaded area as the difference of two left tails.

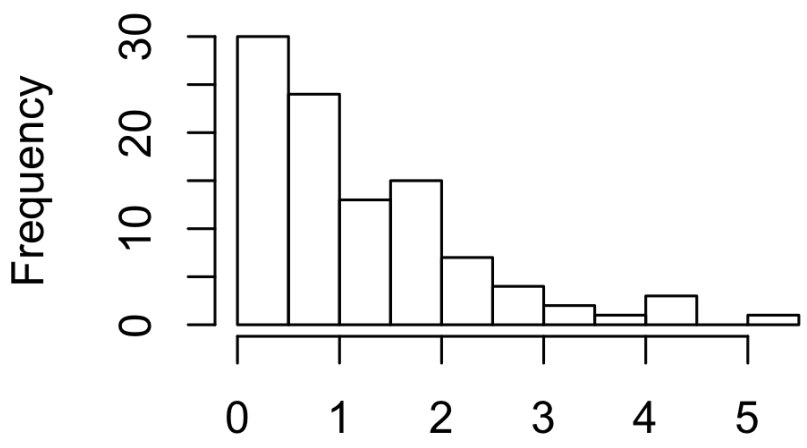
Other continuous distributions

- There are many continuous random variables that are not normally distributed.
- For example, we could explore the **uniform** or **exponential** distributions.

```
hist(runif(100))
```




```
hist(rexp(100))
```



- Normal random variables are the only ones we will need in this class.
- Let's investigate why the normal distribution is so important.