Lab 4: Normal & Binomial Distributions

Overview

In this lab we work with distributions of random variables. The normal distribution is an example of a continuous random variable, and the binomial distribution is an example of a discrete random variable. We'll explore data to arrive at empirical estimates of parameters, and then use those estimates to explore the binomial and normal distributions and calculate the probability of certain events occurring based on these distributions.

Data

The physiological cost of reproduction has been established by a reduction in the lifespan of female fruitflies. Is there a similar cost to male fruitflies? This dataset contains observations on five groups of 25 male fruitflies from an experiment designed to test if increased reproduction reduces longevity for male fruitflies. The five groups are: males forced to live alone, males assigned to live with one or eight newly pregnant females (non-receptive females), and males assigned to live with one or eight virgin females (interested females). See Partridge and Farquhar. "Sexual Activity and the Lifespan of Male Fruitflies", *Nature*, 294: 580-581, 1981, for the published research. We will use the same data, encoded in `fruitfly.csv`:

`No`: serial number (1-25) within each group of 25
`type`: Type of experimental assignment

1 = no females
2 = 1 newly pregnant female
3 = 8 newly pregnant females
4 = 1 virgin female
5 = 8 virgin females

`lifespan`: lifespan (days)
`thorax`: length of thorax (mm)
`sleep`: percentage of each day spent sleeping

Explore data

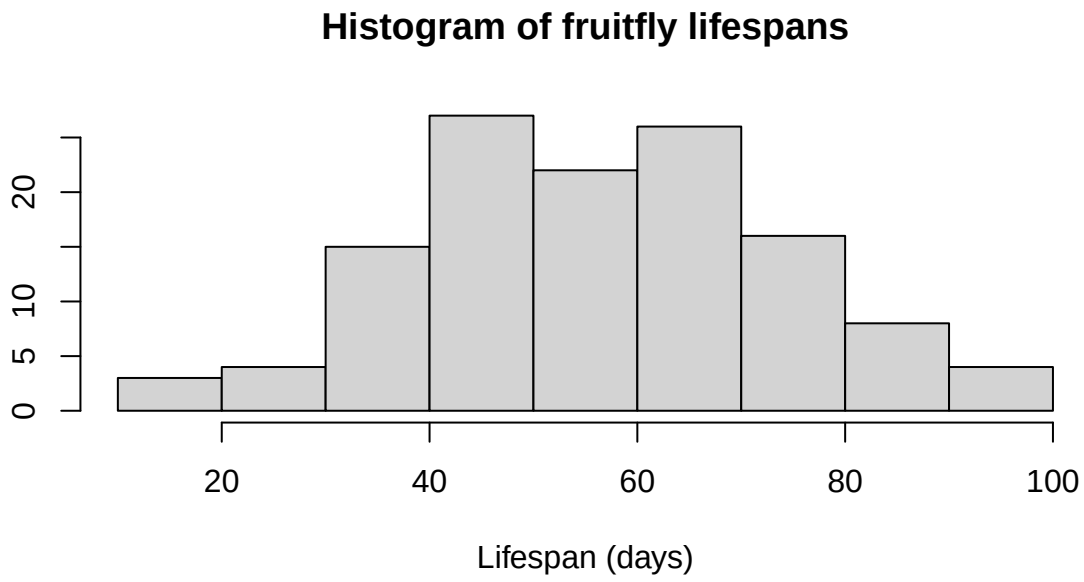
Import the dataset `fruitfly.csv`. Obtain summary statistics and examine the distribution of the overall lifespan of fruitflies.

```
#getwd() #Always begin by checking your working directory!
#setwd("location of working directory") #File path of your working directory
fruitfly <- read.csv("fruitfly.csv")  #Note that this will work ONLY IF your dataset (fruitfly.csv file
```

```
summary(fruitfly$lifespan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   46.00   58.00   57.44   70.00   97.00
```

```r
hist(fruitfly$lifespan,
     main = "Histogram of fruitfly lifespans", # main = for the title
     xlab = "Lifespan (days)",                 # xlab = for the x-axis label
     ylab = "")                                # ylab = for the y-axis label
```



The lifespan of fruitflies ranges from 16 to 97 days with an average of 57.4 days (standard deviation = 17.6). From the histogram, the distribution of lifespan appears to be approximately normal.

Normal distribution

Once you have identified a random variable, like `lifespan`, as approximately normal, you can use the normal distribution to calculate probabilities regarding certain outcomes of this variable. For example, what is the probability that a fruitfly lives less or equal than 50 days? Without statistical software, you do this by calculating a z-score and then consulting the z-table. With R, this can be done with the `pnorm` function.

```r
?pnorm() #to learn more about the pnorm function using the built-in help tool in RStudio
pnorm(q = 50, mean = 57.4, sd = 17.6)
```

```
## [1] 0.3370767
```

The `pnorm` command gives the area under the normal curve with the specified mean and standard deviation, such that the corresponding normal random variable is less or equal to a given value `q` (a lower tail area). The argument `q` stands for quantile, which represents a value from the data distribution.

Based on the normal distribution, 33.7% of fruitflies are expected to survive less or equal to 50 days. How does this compare to what we observed in our data?

```r
sum(fruitfly$lifespan <= 50) / length(fruitfly$lifespan)
```

```
## [1] 0.392
```

In the `sum` function we count how many observations survived less or equal to 50 days, and with the `length` command we obtain the total sample size. In the data, our empirical estimate is that 39.2% survived less or equal to 50 days, which is relatively close to the theoretical estimate of 33.7%.

If we wanted an **upper tail** area, such as the probability that a fruitfly survives more than 50 days, we take the **complement** of this event.

```
1 - pnorm(q = 50, mean = 57.4, sd = 17.6)
```

```
## [1] 0.6629233
```

which yields a probability of 0.663.

To calculate a probability that a normal random variable falls into an interval, for example that a fruitfly survives between 50 and 70 days, we can compute the probability that a fruitfly survives less or equal to 50 days and less or equal to 70 days, and subtract the former from the latter. The following two lines of code are equivalent:

```
pnorm(q = 70, mean = 57.4, sd = 17.6) - pnorm(q = 50, mean = 57.4, sd = 17.6)
```

```
## [1] 0.4258995
```

```
diff(pnorm(q = c(50, 70), mean = 57.4, sd = 17.6))
```

```
## [1] 0.4258995
```

Note that a vector argument `q = c(50, 70)` indicates that `pnorm` shall compute two probabilities, for 50 and 70 days, while `diff` subtracts the former from the latter.

Lastly, we can use the normal distribution to find percentiles of the normal distribution using `qnorm`. The argument `p` is the percentile of interest, entered as a number between 0 and 1. For example,

```
qnorm(p = 0.90, mean = 57.4, sd = 17.6) #returns the value of the 90th percentile
```

```
## [1] 79.95531
```

```
pnorm(q = 79.95531, mean = 57.4, sd = 17.6) #confirms that the value of 79.95531 is the 90th percentile
```

```
## [1] 0.9
```

Based on the normal distribution, 90% of fruitflies survive approximately 80 days or less.

Binomial distribution

We calculated previously that the probability that a fruitfly survives more than 50 days is 0.663. Suppose we are experimenting on a new group of 8 fruitflies, and we want to know the likelihood that a certain number of them survive more than 50 days. To answer this question, we can use the binomial distribution because:

(1) fruitflies are assumed to be independent,
(2) there is a fixed number of fruitflies,

(3) each fruitfly can either survive more than 50 days or die before or at 50 days,
(4) each fruitfly is assumed to be equally likely to survive.

The `dbinom` command gives the probability that a certain number of fruitflies survive more than 50 days. For example, the probability that exactly 5 out 8 fruitflies survive more than 50 days is calculated by

```
?dbinom() # to learn more
dbinom(x = 5, size = 8, prob = 0.663)
```

```
## [1] 0.2745651
```

which gives a probability of 0.27. The argument `x` is the number of survived fruitflies (number of successes), `size` is the total number of fruitflies (number of trials), and `prob` is the probability to survive more than 50 days (probability of success). Since we have 8 fruitflies, as few as 0 and as many as 8 could survive more than 50 days.

You can quickly calculate the probability associated with each of the outcomes by letting `x` be a vector of numbers from 0 to 8.

```
dbinom(x = 0:8, size = 8, prob = 0.663)
```

```
## [1] 0.0001663563 0.0026182603 0.0180287034 0.0709378655 0.1744503147
## [6] 0.2745651243 0.2700840911 0.1518149660 0.0373343184
```

In order to see the results more clearly, you could put the results in a **data frame** with columns representing different values of `x` and the corresponding probabilities. To do this, we will use the `data.frame` command as follows.

```
Fruitfly8 <- data.frame(x = 0:8, prob = dbinom(x = 0:8, size = 8, prob = 0.663))
Fruitfly8   # look at the data frame created above
```

```
##   x         prob
## 1 0 0.0001663563
## 2 1 0.0026182603
## 3 2 0.0180287034
## 4 3 0.0709378655
## 5 4 0.1744503147
## 6 5 0.2745651243
## 7 6 0.2700840911
## 8 7 0.1518149660
## 9 8 0.0373343184
```

Notice that `x = 0:8` means that the first column will have a name `x` and numbers from 0 to 8 as values. Similarly, the second column has the name `prob` and the corresponding binomial probabilities, calculated using the `dbinom` function, as values.

Lastly, you can calculate the probability associated with a range of values by using `sum`. The probability that less than or equal to 5 fruitflies survive more than 50 days is the probability that between 0 and 5 fruitflies survive more than 50 days. The following two lines of code are equivalent

```r
sum(dbinom(x = 0:5, size = 8, prob = 0.663))
```

```
## [1] 0.5407666
```

```r
sum(Fruitfly8$prob[Fruitfly8$x <= 5])
```

```
## [1] 0.5407666
```

The first method uses the `dbinom` function directly, while the second method explores the data frame `Fruitfly8` that we have just created. Notice that `[Fruitfly8$x <= 6]` indicates that only the probabilities corresponding to `x <= 6` are chosen. This an example of choosing a subset of observations. You will learn about this in more detail later in the class.

The probability that at least 3 fruitflies survive more than 50 days is the probability that between 3 and 8 fruitflies survive more than 50 days.

```r
sum(dbinom(x = 3:8, size = 8, prob = 0.663))
```

```
## [1] 0.9791867
```

```r
sum(Fruitfly8$prob[Fruitfly8$x >= 3])
```

```
## [1] 0.9791867
```