Assessment 04 - Normal Distribution

Gabriele Mineo - Harvard Data Science Professional

Proportions

Histograms and density plots provide excellent summaries of a distribution. But can we summarize even further? We often see the average and standard deviation used as summary statistics: a two number summary! To understand what these summaries are and why they are so widely used, we need to understand the normal distribution.

The normal distribution, also known as the bell curve and as the Gaussian distribution, is one of the most famous mathematical concepts in history. A reason for this is that approximately normal distributions occur in many situations. Examples include gambling winnings, heights, weights, blood pressure, standardized test scores, and experimental measurement errors. Often data visualization is needed to confirm that our data follows a normal distribution.

Here we focus on how the normal distribution helps us summarize data and can be useful in practice.

One way the normal distribution is useful is that it can be used to approximate the distribution of a list of numbers without having access to the entire list. We will demonstrate this with the heights dataset.

Load the height data set and create a vector x with just the male heights:

```r
library(dslabs)
data(heights)
x <- heights$height[heights$sex == "Male"]
```

Instructions

- What proportion of the data is between 69 and 72 inches (taller than 69 but shorter or equal to 72)?

```r
library(dslabs)
data(heights)
x <- heights$height[heights$sex == "Male"]
mean(x > 69 & x <= 72)
```

```
## [1] 0.3337438
```

Averages and Standard Deviations

Suppose all you know about the height data from the previous exercise is the average and the standard deviation and that its distribution is approximated by the normal distribution. We can compute the average and standard deviation like this:

```
library(dslabs)
data(heights)
x <- heights$height[heights$sex=="Male"]
avg <- mean(x)
stdev <- sd(x)
```

Suppose you only have `avg` and `stdev` below, but no access to `x`, can you approximate the proportion of the data that is between 69 and 72 inches?

Instructions

- Use the normal approximation to estimate the proportion the proportion of the data that is between 69 and 72 inches.

- Note that you can’t use x in your code, only avg and stdev. Also note that R has a function that may prove very helpful here - check out the pnorm function (and remember that you can get help by using ?pnorm).

```
library(dslabs)
data(heights)
x <- heights$height[heights$sex=="Male"]
avg <- mean(x)
stdev <- sd(x)
```

Approximations

Notice that the approximation calculated in the second question is very close to the exact calculation in the first question. The normal distribution was a useful approximation for this case.

However, the approximation is not always useful. An example is for the more extreme values, often called the “tails” of the distribution. Let’s look at an example. We can compute the proportion of heights between 79 and 81.

```
library(dslabs)
data(heights)
x <- heights$height[heights$sex == "Male"]
mean(x > 79 & x <= 81)
```

Instructions

- Use normal approximation to estimate the proportion of heights between 79 and 81 inches and save it in an object called approx.
- Report how many times bigger the actual proportion is compared to the approximation.

```
library(dslabs)
data(heights)
x <- heights$height[heights$sex == "Male"]
exact <- mean(x > 79 & x <= 81)
```

Seven footers and the NBA

Someone asks you what percent of seven footers are in the National Basketball Association (NBA). Can you provide an estimate? Let’s try using the normal approximation to answer this question.

First, we will estimate the proportion of adult men that are 7 feet tall or taller.

Assume that the distribution of adult men in the world as normally distributed with an average of 69 inches and a standard deviation of 3 inches.

Instructions

- Using this approximation, estimate the proportion of adult men that are 7 feet tall or taller, referred to as seven footers. Print out your estimate; don’t store it in an object.

```
1 - pnorm(7*12, 69, 3)
```

```
## [1] 2.866516e-07
```

Estimating the number seven footers

Now we have an approximation for the proportion, call it p, of men that are 7 feet tall or taller.

We know that there are about 1 billion men between the ages of 18 and 40 in the world, the age range for the NBA.

Can we use the normal distribution to estimate how many of these 1 billion men are at least seven feet tall?

Instructions

- Use your answer to the previous exercise to estimate the proportion of men that are seven feet tall or taller in the world and store that value as p.
- Then round the number of 18-40 year old men who are seven feet tall or taller to the nearest integer. (Do not store this value in an object.)

```r
p <- 1 - pnorm(7*12, 69, 3)
round(p * 10^9)
```

```
## [1] 287
```

How many seven footers are in the NBA?

There are about 10 National Basketball Association (NBA) players that are 7 feet tall or higher.

Instructions

- Use your answer to exercise 4 to estimate the proportion of men that are seven feet tall or taller in the world and store that value as p.
- Use your answer to the previous exercise (exercise 5) to round the number of 18-40 year old men who are seven feet tall or taller to the - nearest integer and store that value as N. Then calculate the proportion of the world's 18 to 40 year old seven footers that are in the NBA. (Do not store this value in an object.)

```r
p <- 1 - pnorm(7*12, 69, 3)
N <- round(p * 10^9)
10/N
```

```
## [1] 0.03484321
```

Lebron James' height

In the previous exerceise we estimated the proportion of seven footers in the NBA using this simple code:

```
p <- 1 - pnorm(7*12, 69, 3)
N <- round(p * 10^9)
10/N
```

Repeat the calculations performed in the previous question for Lebron James' height: 6 feet 8 inches. There are about 150 players, instead of 10, that are at least that tall in the NBA.

Instructions

- Report the estimated proportion of people at least Lebron's height that are in the NBA.

```r
## Change the solution to previous answer
p <- 1 - pnorm(6*12 + 8, 69, 3)
N <- round(p * 10^9)
150/N
```

```
## [1] 0.001220842
```

Interpretation

In answering the previous questions, we found that it is not at all rare for a seven footer to become an NBA player.

What would be a fair critique of our calculations?

Possible Answers

- Practice and talent are what make a great basketball player, not height.
- The normal approximation is not appropriate for heights.
- As seen in exercise 3, the normal approximation tends to underestimate the extreme values. It's possible that there are more seven footers than we predicted. [X]
- As seen in exercise 3, the normal approximation tends to overestimate the extreme values. It's possible that there are less seven footers than we predicted.