Normal Distribution

Casper Crause

Contents

Libraries

We only need the `tidyverse` for basic funcions

```
library(tidyverse)
```

Normal Distribution

Defining Normal Distribution

- The probability distribution function `(PDF)` is a normal curve with an area of 1 beneath it. This represents the frequency of values.

Finding the probability:

Suppose we have a class of students and we recorded their height in meters.

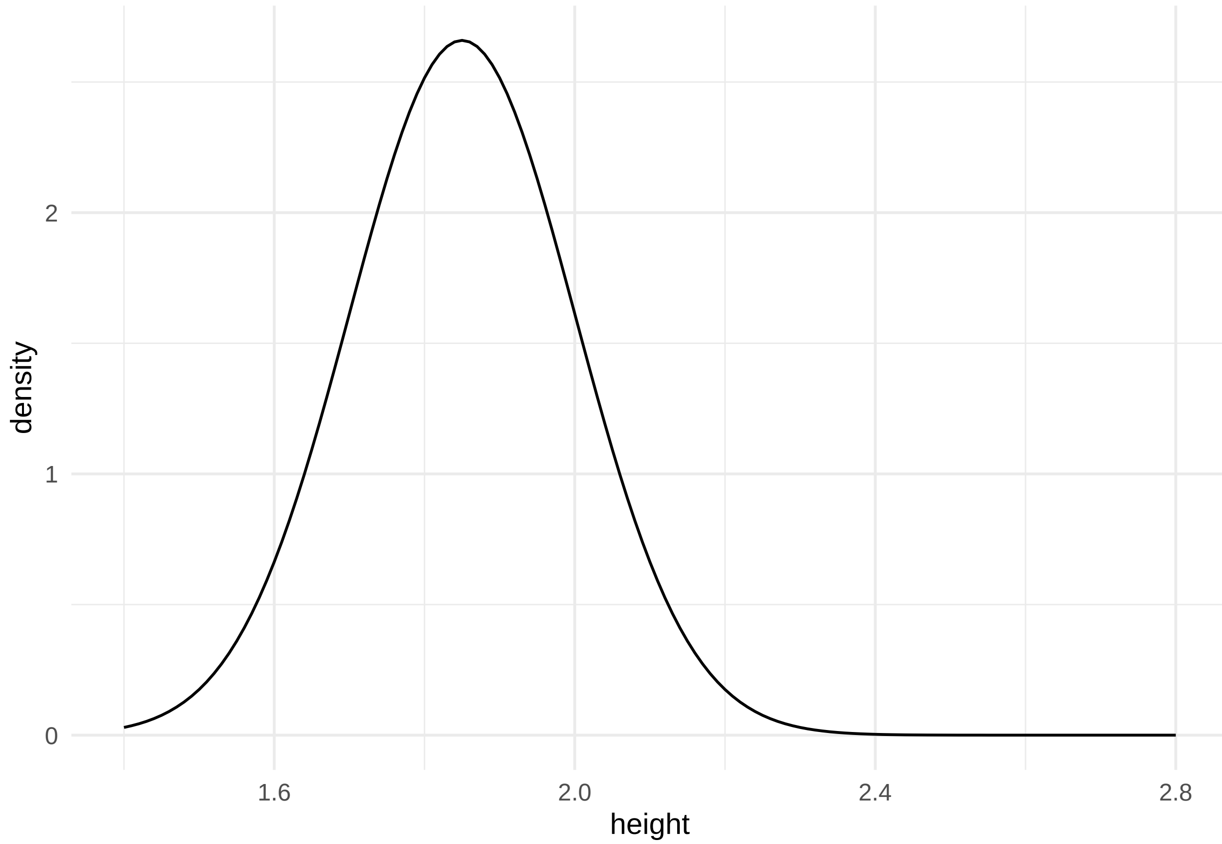
- The ***average height*** is 1.85 meters and the ***standard deviation*** in 0.15 meters

```
tibble(
     `student id`    = 1:141,
    height           = seq(1.4,2.8, by = 0.01),
    density          = dnorm(x     = height,
                        mean      = 1.85,
                        sd        = 0.15),
    cum_prob = pnorm(q     = height,
                            mean = 1.85,
                            sd   = 0.15)
     )-> data
data
```

```
## # A tibble: 141 x 4
##    `student id` height density cum_prob
##           <int>  <dbl>   <dbl>    <dbl>
##  1            1   1.4   0.0295  0.00135
##  2            2   1.41  0.0360  0.00168
##  3            3   1.42  0.0437  0.00207
##  4            4   1.43  0.0528  0.00256
##  5            5   1.44  0.0635  0.00313
##  6            6   1.45  0.0760  0.00383
##  7            7   1.46  0.0906  0.00466
##  8            8   1.47  0.107   0.00565
##  9            9   1.48  0.127   0.00682
## 10           10   1.49  0.149   0.00820
## # ... with 131 more rows
```

Plotting the PDF

```
data %>%
    ggplot(aes(height,density))+
    geom_line()+
    theme_minimal()
```



What percentage of students are shorter than 2.05 meters?

1. We can `filter` for the height at exactly 2.05m and look at the `cum_prob`
2. We can use the `pnorm()` function

```
data %>%
    filter(height==2.05)
```

```
## # A tibble: 1 x 4
##   `student id` height density cum_prob
##          <int>  <dbl>   <dbl>    <dbl>
## 1           66   2.05    1.09    0.909
```

```
pnorm(q = 2.05,mean = 1.85, sd = .15) -> p_2.05

p_2.05 %>% scales::percent(accuracy = 0.01)
```

```
## [1] "90.88%"
```

What percentage of students are taller than 2.05 meters?

If you were curious about the *cum_prob* and scrolled to page 15, you would have seen the probality is equal to 1. That is no coincidence! The sum of all probalities must add up to 1.

Using this principle we can find the percentage of students taller than 2.05m

$$p(A) + p(B) = 1$$

hence

$$1 - p(A) = p(B)$$

```
(1-p_2.05) %>%
    scales::percent(accuracy = 0.01)
```

```
## [1] "9.12%"
```

Finding the proportion of students between two diferent height measures

How to find the proportion of students that fall between 1.87m and 2.05m

1. We need to find the % of students shorter than **1.87m** and the % who are shorter than **2.05m** . We already know ***90.88%*** of students are shorter than **2.05m**.

```
pnorm(q = 1.87,mean = 1.85,sd = .15) -> p_1.87
```

```
(p_2.05 - p_1.87) %>%
    scales::percent(accuracy = 0.01)
```

```
## [1] "35.58%"
```

Working with Quantiles

What is the *68th percentile* of this group of students?

1. We can use the qnorm function

```
qnorm(p = .68,mean = 1.85,sd = 0.15) %>%
    scales::comma(suffix   = 'M',
                  accuracy = 0.01 )
```

```
## [1] "1.92M"
```

Equivalent functions in Excel

1. Returning the normal distribution for a spesified mean and standard deviation:

$$= NORM.DIST(x, mean, sd, cumulative)$$

2. Finding the value for a specified probability

$$= NORM.INV(probability, mean, sd)$$