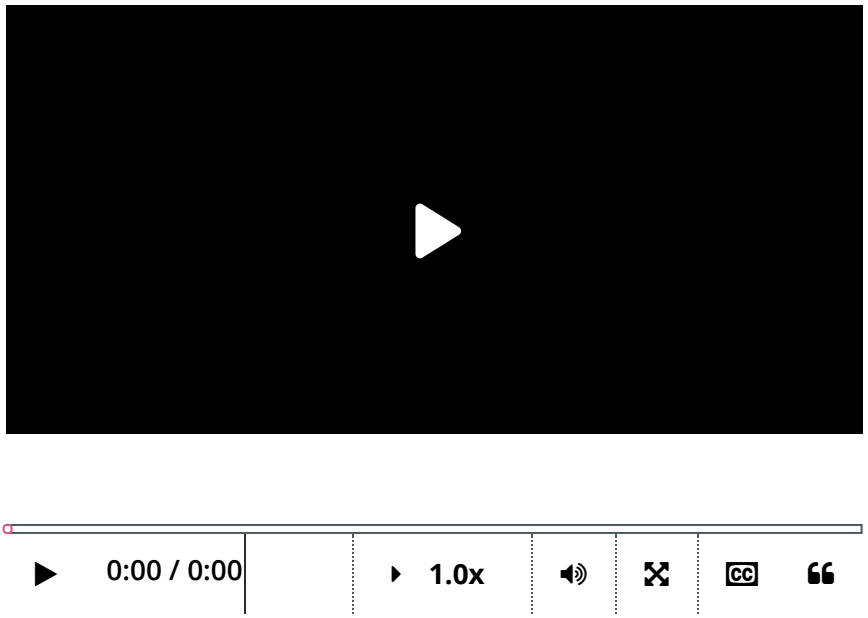
The Normal CDF and pnorm
The normal CDF and pnorm



Start of transcript. Skip to the end.

RAFAEL IRIZARRY: In the data visualization module,

we introduced the normal distribution as a useful approximation

to many naturally occurring distributions,

including that of height.

The cumulative distribution for the normal distribution

Textbook link

Here is the textbook section on the theoretical distribution and the normal approximation.

Key points

- The normal distribution has a mathematically defined CDF which can be computed in R with the function `pnorm()`.
- `pnorm(a, avg, s)` gives the value of the cumulative distribution function $F(a)$ for the normal distribution defined by average `avg` and standard deviation `s`.
- We say that a random quantity is normally distributed with average `avg` and standard deviation `s` if the approximation `pnorm(a, avg, s)` holds for all values of `a`.
- If we are willing to use the normal approximation for height, we can estimate the distribution simply from the mean and standard deviation of our values.
- If we treat the height data as discrete rather than categorical, we see that the data are not very useful because integer values are more common than expected due to rounding. This is called *discretization*.
- With rounded data, the normal approximation is particularly useful when computing probabilities of intervals of length 1 that include exactly one integer.

Code: Using `pnorm` to calculate probabilities

Given male heights `x`:

```
library(tidyverse)
library(dslabs)
data(heights)
x <- heights %>% filter(sex=="Male") %>% pull(height)
```

We can estimate the probability that a male is taller than 70.5 inches with:

```
1 - pnorm(70.5, mean(x), sd(x))
```

Code: Discretization and the normal approximation

```
# plot distribution of exact heights in data
plot(prop.table(table(x)), xlab = "a = Height in inches", ylab = "Pr(x

# probabilities in actual data over length 1 ranges containing an integ
mean(x <= 68.5) - mean(x <= 67.5)
mean(x <= 69.5) - mean(x <= 68.5)
mean(x <= 70.5) - mean(x <= 69.5)

# probabilities in normal approximation match well
pnorm(68.5, mean(x), sd(x)) - pnorm(67.5, mean(x), sd(x))
pnorm(69.5, mean(x), sd(x)) - pnorm(68.5, mean(x), sd(x))
pnorm(70.5, mean(x), sd(x)) - pnorm(69.5, mean(x), sd(x))

# probabilities in actual data over other ranges don't match normal app
mean(x <= 70.9) - mean(x <= 70.1)
pnorm(70.9, mean(x), sd(x)) - pnorm(70.1, mean(x), sd(x))
```

Learn About Verified Certificates