The Gaussian Distribution

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

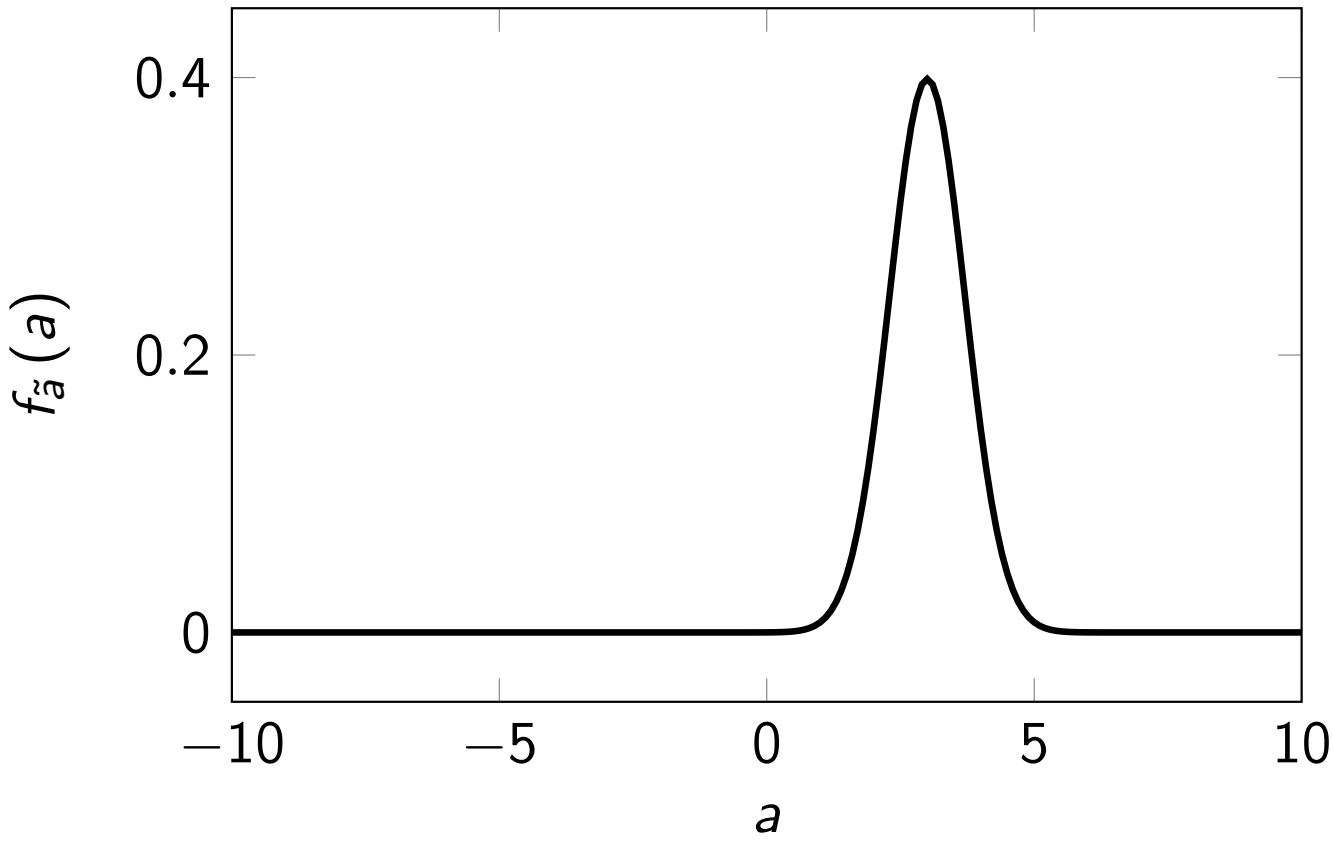
Gaussian distribution

Motivation: Sum of independent quantities is approximately Gaussian

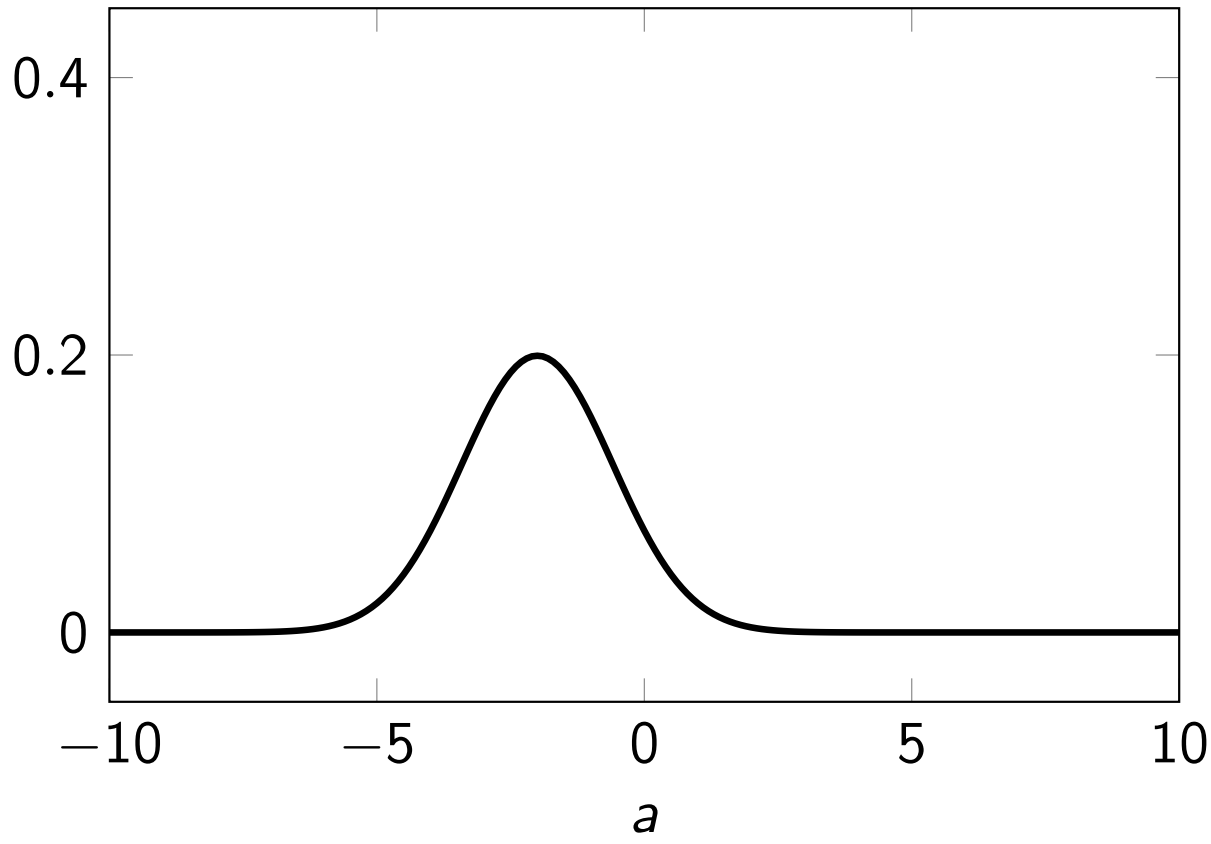
The Gaussian or normal parametric pdf with mean μ and standard deviation σ is

$$f_{\tilde{a}}(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

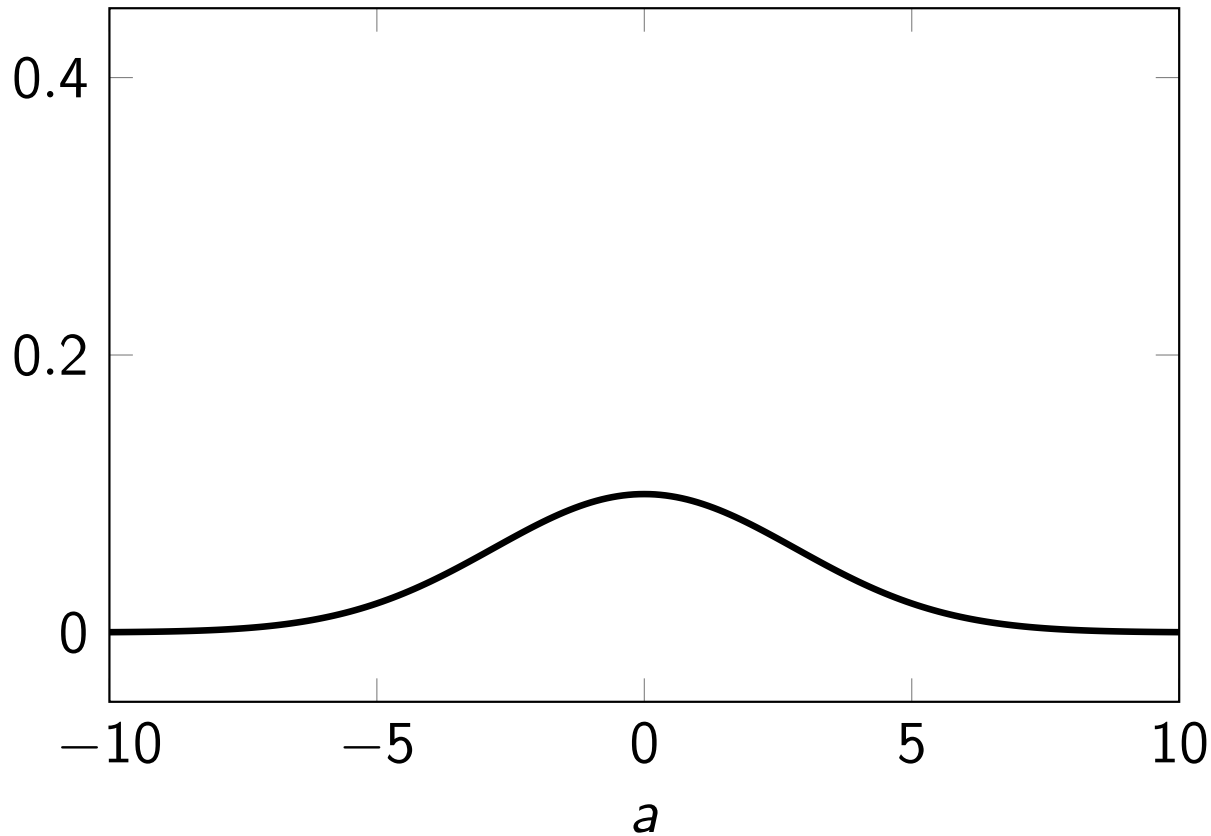
$\mu = 3,\ \sigma = 1$



$\mu = -2,\ \sigma = 2$



$\mu = 0,\ \sigma = 4$



Shifting and scaling a Gaussian

$\tilde{a}$ is a standard Gaussian

What is the pdf of $\tilde{b} := \sigma\tilde{a} + \mu$?

$$\begin{aligned} F_{\tilde{b}}(b) &= \mathrm{P}(\sigma\tilde{a} + \mu \leq b) \\ &= \mathrm{P}\left(\tilde{a} \leq \frac{b-\mu}{\sigma}\right) \\ &= \int_{-\infty}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2}{2}} \, \mathrm{d}a \end{aligned}$$

$$f_{\tilde{b}}(b) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b-\mu)^2}{2\sigma^2}}$$

Real data: Height of 25,000 people

page | discussion | view source | history

SOCR Data Dinov 020108 HeightsWeights

Contents [hide]

SOCR Data - 25,000 Records of Human Heights (in) and Weights (lbs)

Summary

Human Height and Weight are mostly hereditable, but lifestyles, diet, health and environmental factors also play a role in determining individual's physical characteristics. The dataset below contains 25,000 synthetic records of human heights and weights of 18 years old children. These data were simulated based on a 1993 by a Growth Survey of 25,000 children from birth to 18 years of age recruited from Maternal and Child Health Centres (MCHC) and schools and were used to develop Hong Kong's current growth charts for weight, height, weight-for-age, weight-for-height and body mass index (BMI). See also the Major League Baseball Players Height and Weight dataset.

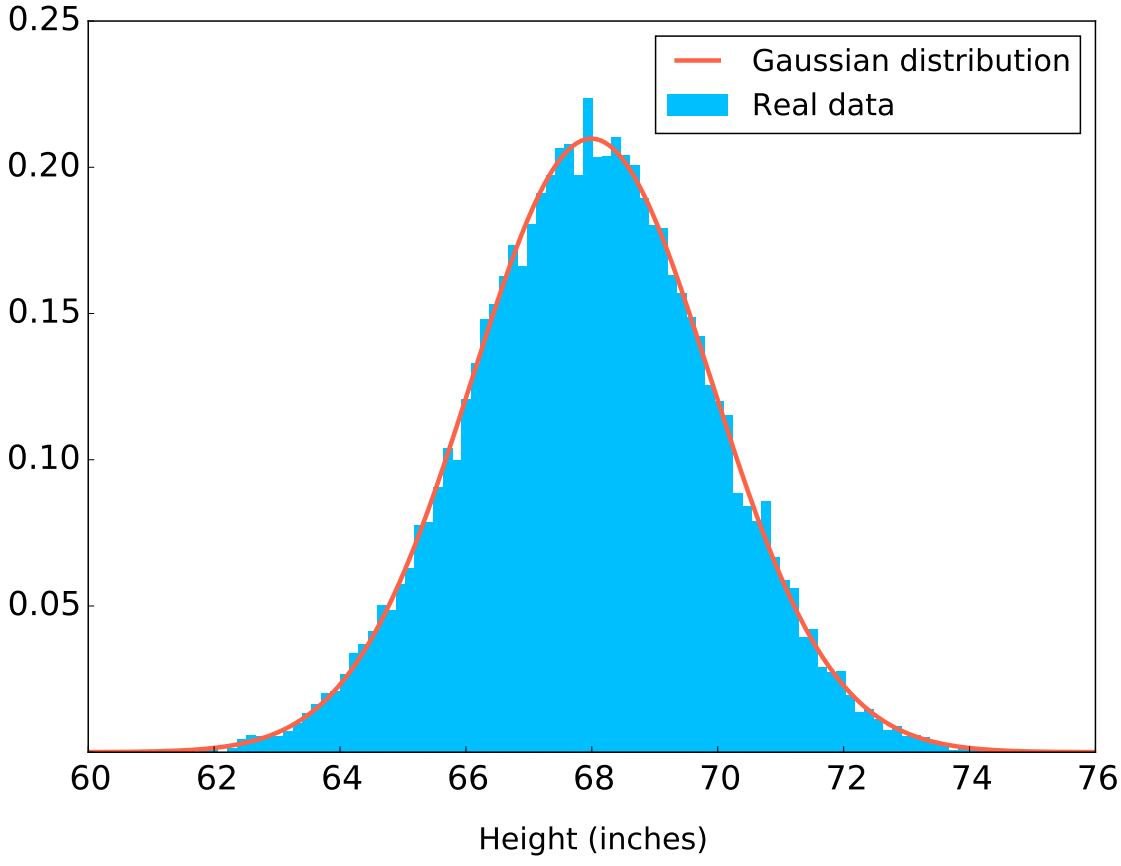
Complete Data

You can find the complete dataset here (7MB).

Sample of 200 Individuals

Index	Height(Inches)	Weight(Pounds)
1	65.78	112.99
2	71.52	136.49
3	69.40	153.03
4	68.22	142.34
5	67.79	144.30
6	68.70	123.30

Real data: Height of 25,000 people



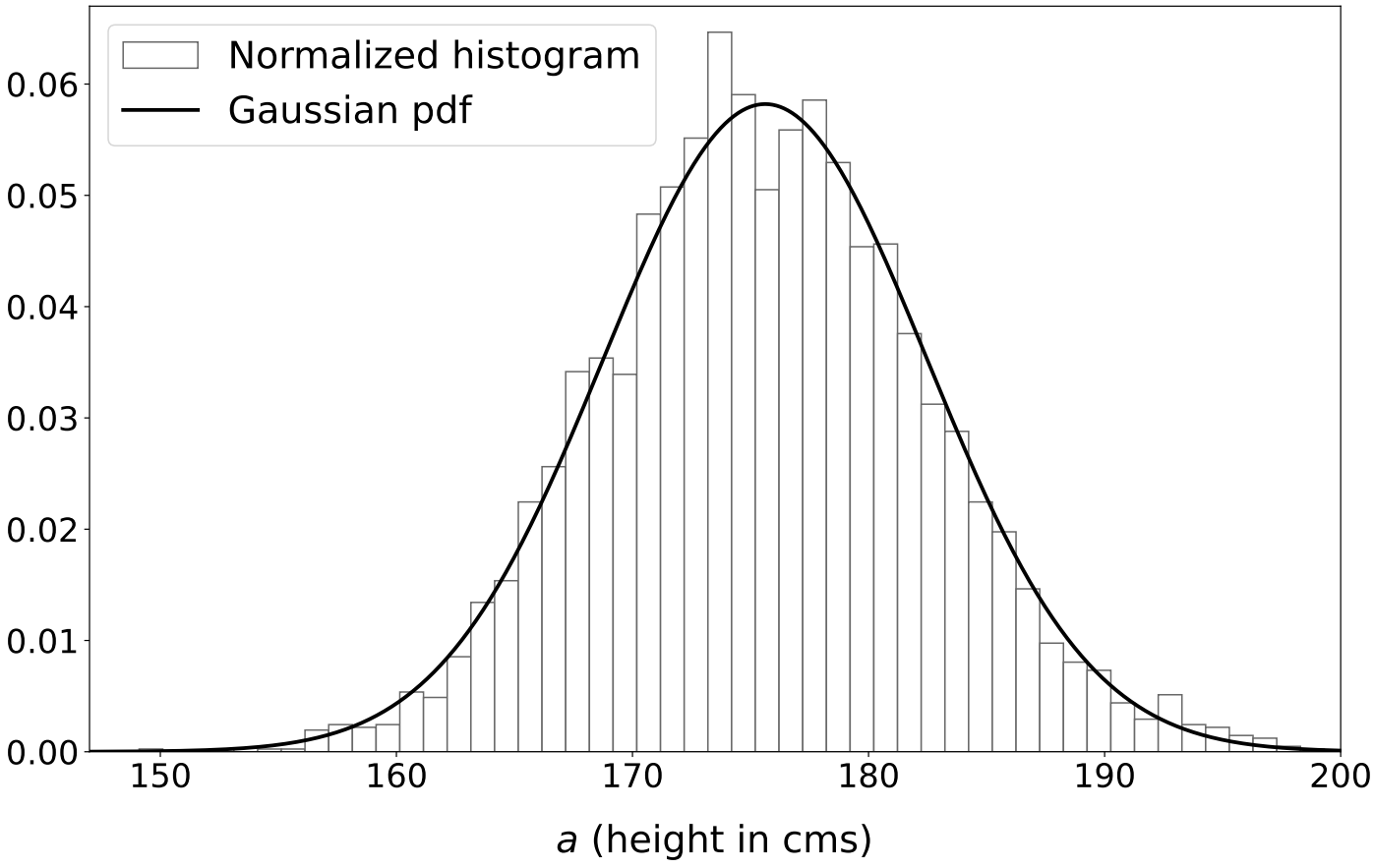
Real data?

SOCR Data - 25,000 Records of Human Heights (in) and Weights

Summary

Human Height and Weight are mostly hereditable, but lifestyles, diet, health and environmenta
human heights and weights of 18 years old children. These data were simulated based on a 1993 k

Height in US army



What have we learned?

Definition of Gaussian distribution