

Gaussian distribution and MAP estimates

Oliver W. Layton

CS251: Data analysis and visualization

Lecture 18, Fall 2018

Monday March 17

Plan

- Gaussian distribution
- Gaussian model of human height
- Multivariate Gaussian distribution
- Maximum a-posteriori (MAP) estimation

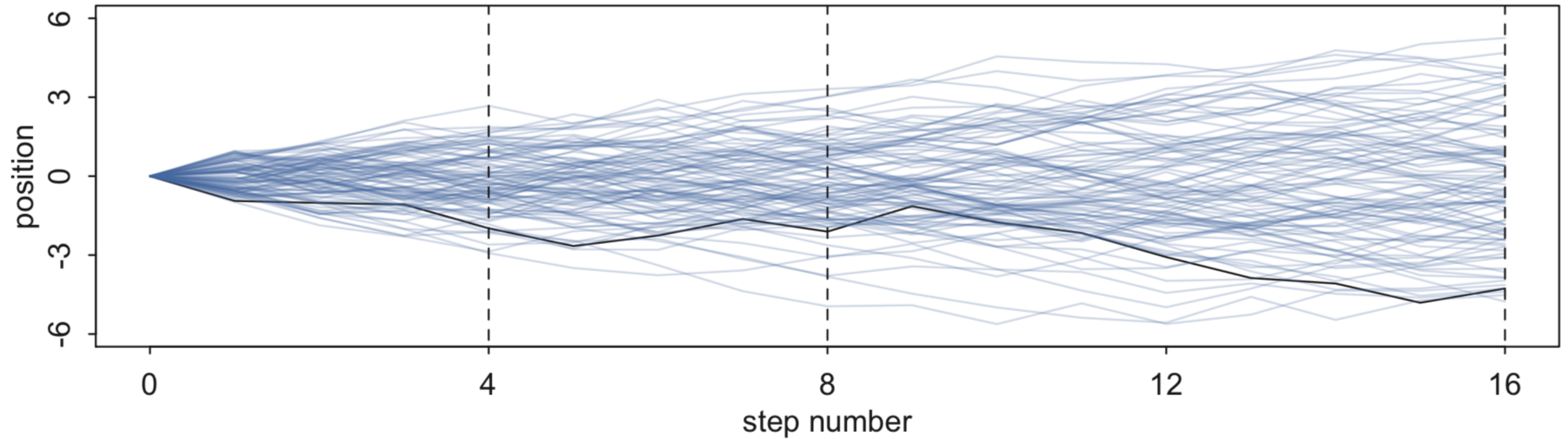
Gaussian (Normal) distribution

- Let's write the formula on the board and introduce its parameters.
- The Binomial distribution converges to the Gaussian distribution when n (number of tosses) is large (≈ 20) and p is not too near 0 or 1.
- The Binomial distribution has mean np and variance $np(1 - p)$.
- To get the approximation, we plug in $\mathcal{N} \left(np, \sqrt{np(1 - p)} \right)$.
- Let's play with the difference using Python.

Why Normal (Gaussian) distributions are common (1/2)

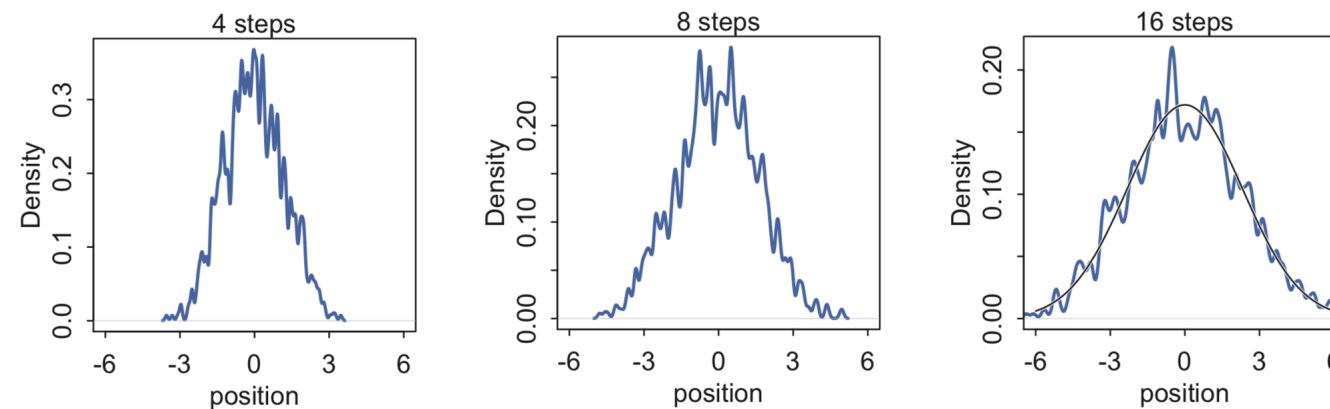
- Gaussian distributions are the most common probability distributions in science.
- Why does it model so many phenomena?
- **Random walk:** Steps on a soccer field

Random walk



Why Normal (Gaussian) distributions are common (2/2)

Adding small random fluctuations repeatedly results in a Gaussian.



- Central limit theorem: Sum of independent identically distributed (*IID*) random variables results in a Gaussian
- Repeated multiplication (e.g. growth factors)

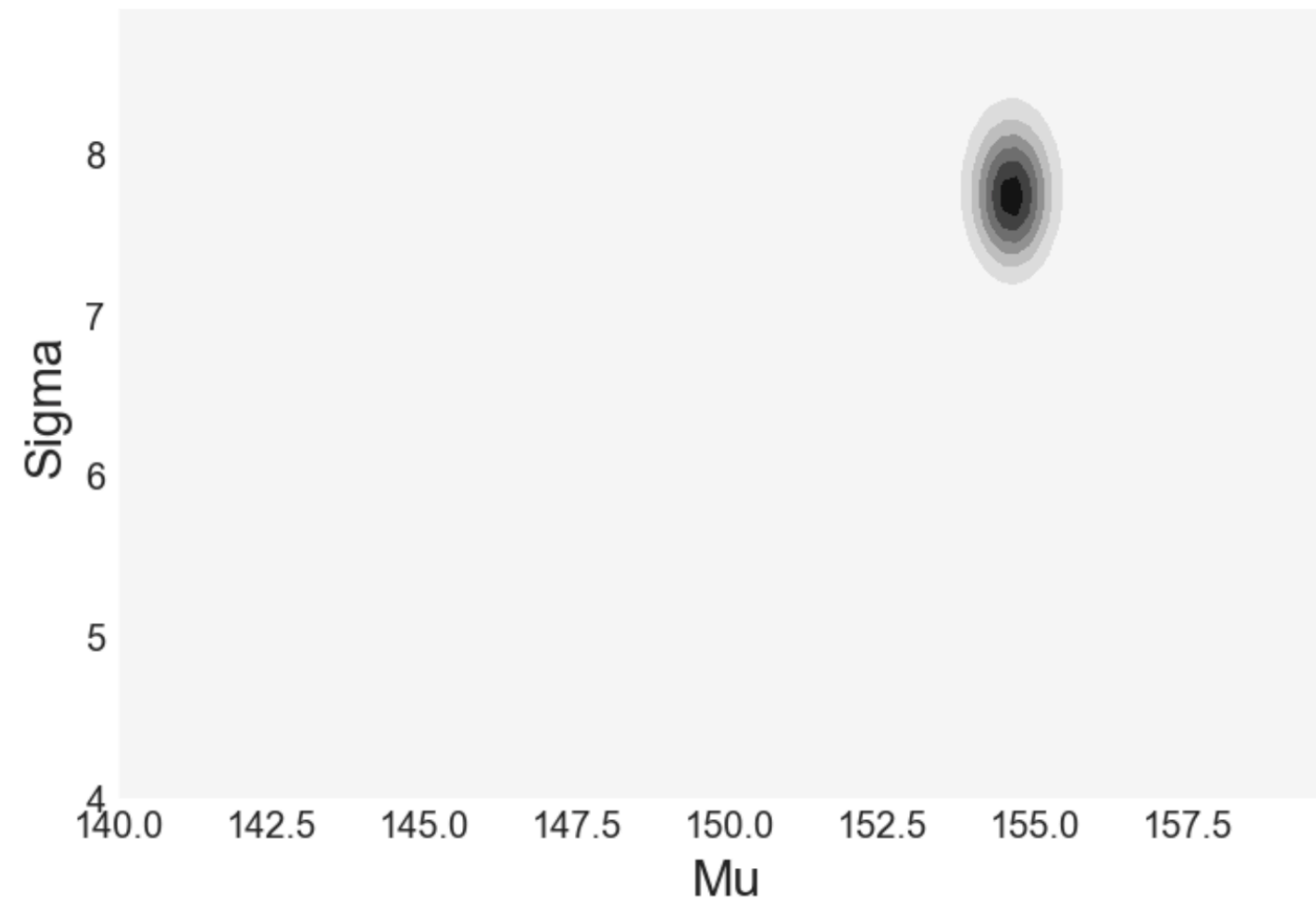
Example: Modeling human height data

Data: !Kung San (20th century foraging population) census collected by Nancy Howell in 1960s. Heights measured for 352 adults.

	height	weight	age	male
0	151.765	47.825606	63.0	1
1	139.700	36.485807	63.0	0
2	136.525	31.864838	65.0	0
3	156.845	53.041915	41.0	1
4	145.415	41.276872	51.0	0

Goal: Use a Gaussian distribution to describe the average height (μ) and variability (σ^2) across the !Kung San population.

Heatmap of Gaussian μ and σ given height data



Given the range of μ and σ values (there ISN'T a single pair), which do we pick as the "best", most representative of the height data mean and variability?

Maximum a posteriori estimate

- (μ, σ) "data" produced by the *output* of the height modeling Gaussian is ALSO Gaussian, but in 2D.
- A Gaussian in two or higher dimensions is called a **multivariate Gaussian/normal distribution**.
- It has a mean and covariance matrix that nicely summarizes plausible pairs of μ and σ in the height model. Mean only gives us heatmap peak.
- Summarizing the height model parameter space (2D: μ, σ) according to the peak of a multivariate Gaussian ($\vec{\mu}$) is called **maximum a posteriori (MAP) estimation**.
- Let's write out multivariate gaussian equation

MAP Estimates for height data

2D Gaussian Mean vector:

parameter	mean (cm)
mu	154.61
sigma	7.77

2D Gaussian Covariance matrix:

	mu	sigma
mu	0.172456	0.001159
sigma	0.001159	0.086376

A 2D Gaussian with the above μ and Σ matrix values captures the height population data.

MAP Correlation matrix (from covariance matrix)

	mu	sigma
mu	1.0	0.009494
sigma	0.009494	1.0

Tells us that knowing either mu or sigma tells us next to nothing about the other value. They are independent.