Visualizing Distributions and Relationships in Data

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Histograms, KDE plots, Rugplots to visualize univariate data

Scatter plots, Jointplots, Hexbin plots for bivariate data

Regression plots to view relationships

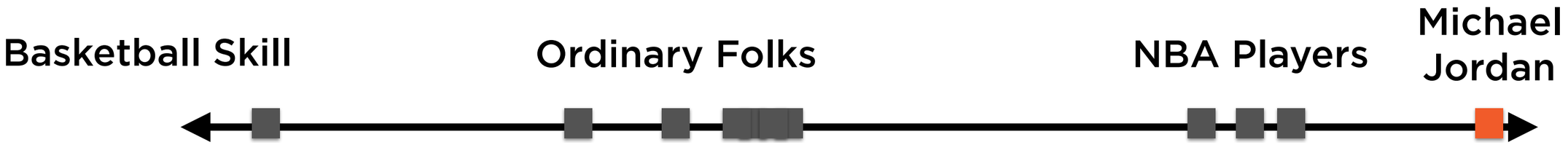
Pairwise relationships using PairGrid

Specialized plots for categorical data

Understanding KDE Plots

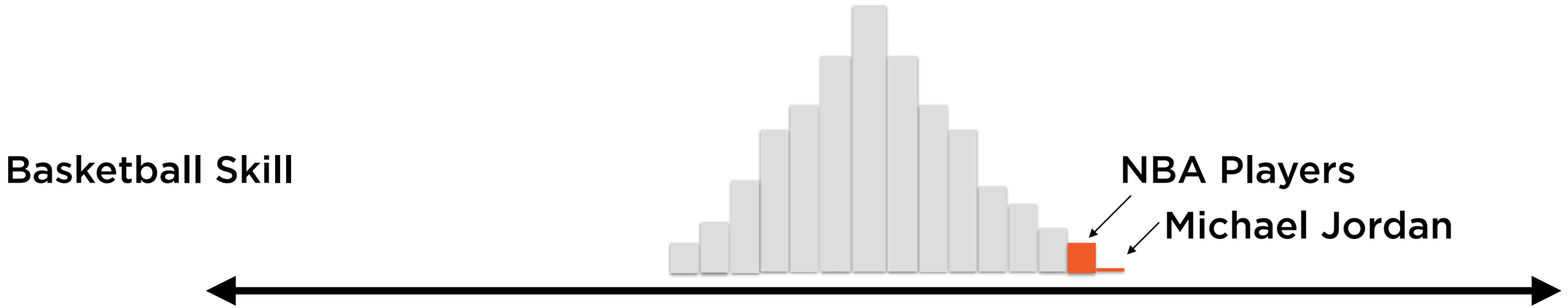
“Michael Jordan is a once-in-a-lifetime player”

Outliers



A once-in-a-lifetime player is an outlier, a point far from the pack

Outliers

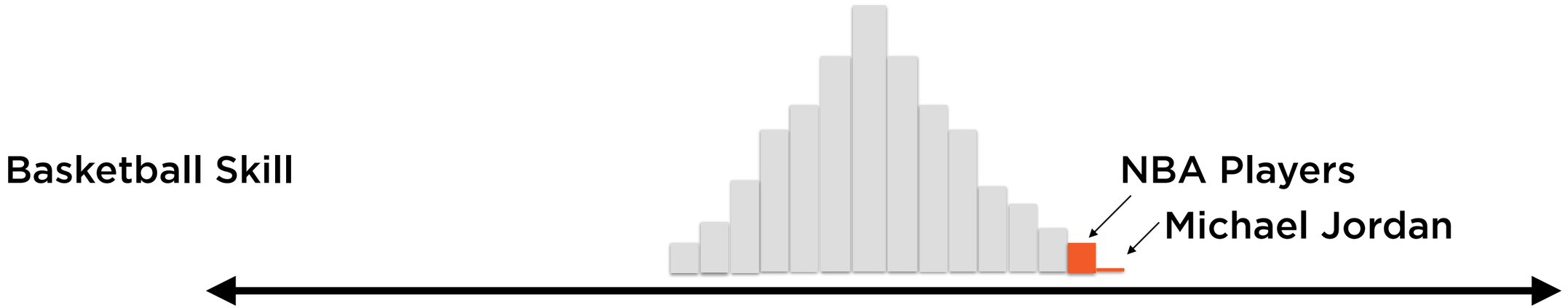


In reality, most ordinary folks would be clustered around an average level of skill

The NBA players would be outliers

Michael Jordan would be an even greater outlier

Outliers

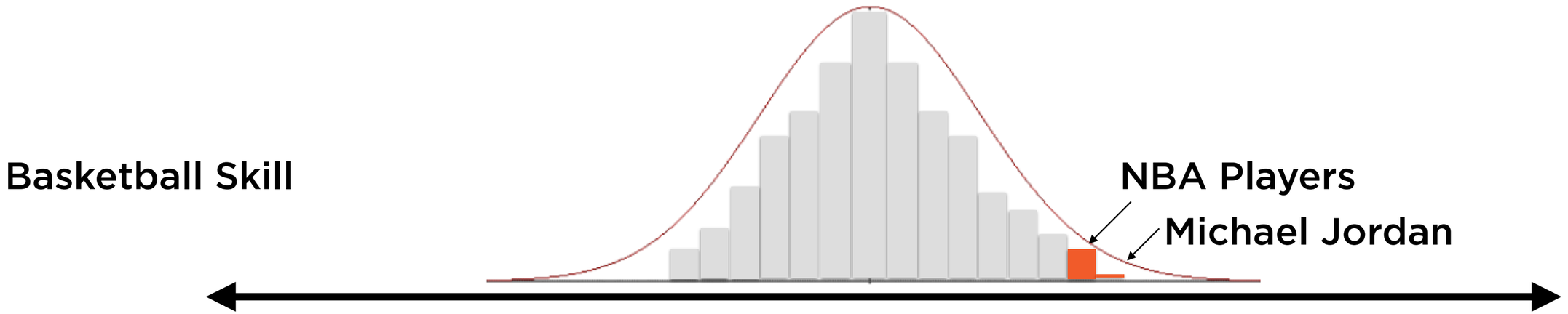


This chart above tells us how common a specific level of skill is

The shape of this chart resembles a bell

This is a Normal Probability Distribution

Outliers

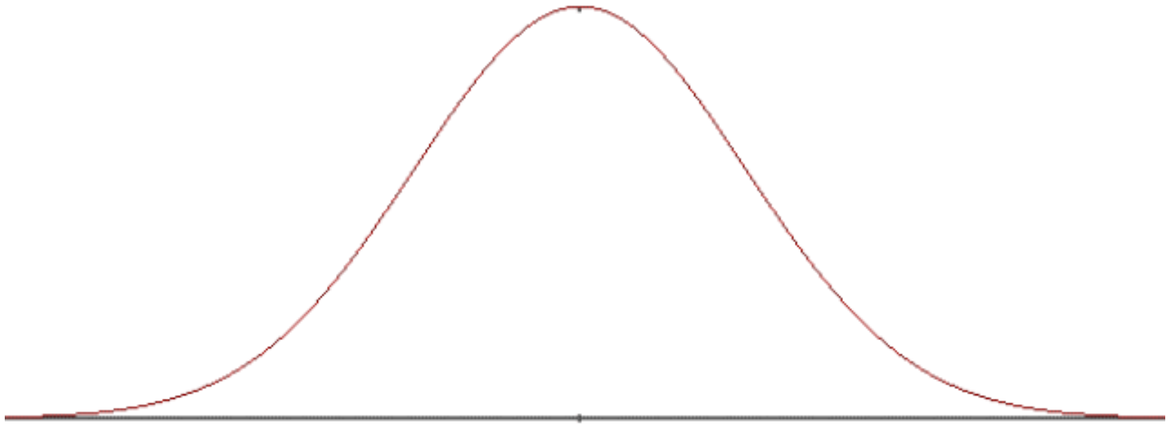


This chart above tells us how common a specific level of skill is

The shape of this chart resembles a bell

This is a Normal Probability Distribution

Outliers

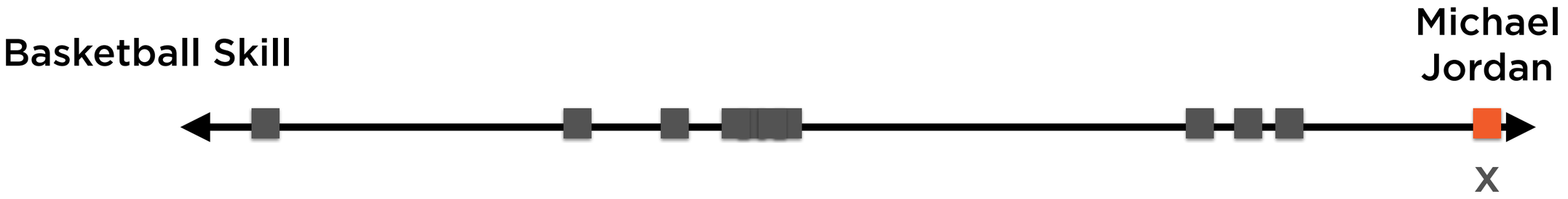


Average is common

Very high and very low are both unusual

The bell curve occurs everywhere in nature

Outliers



What is the probability of any specific value x occurring in the data?

The answer lies in a probability distribution function

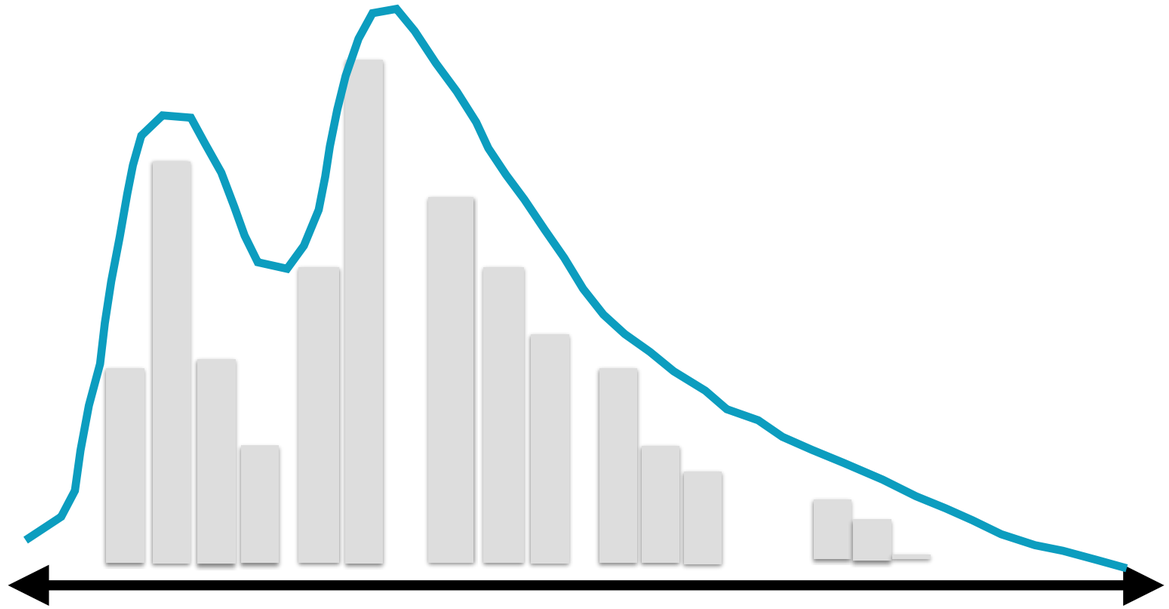
Kernel Density Estimation

A mathematical technique used to get a smooth probability distribution from a histogram of raw data

Given a set of points

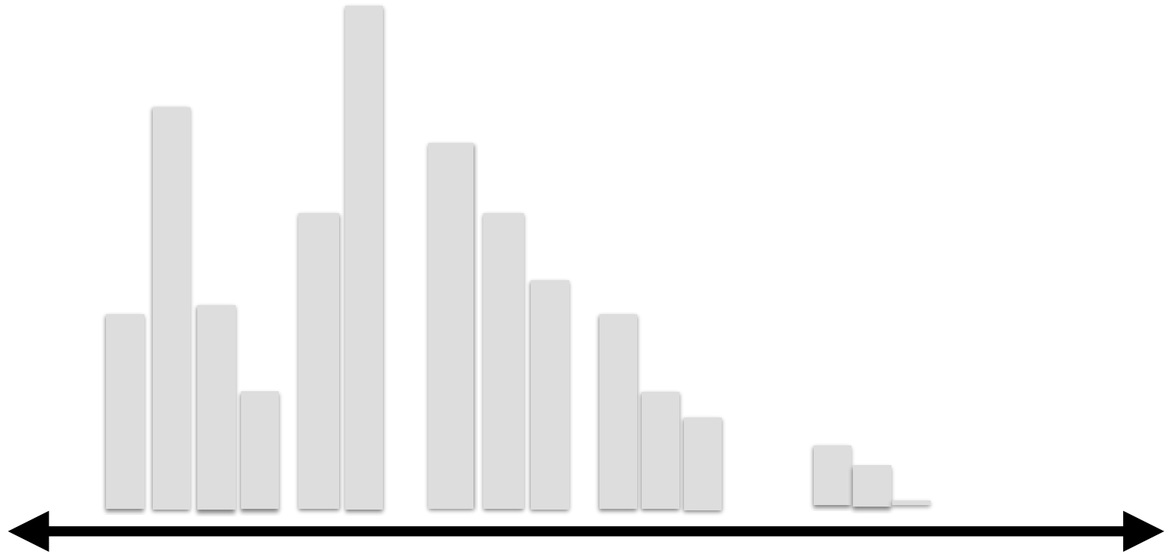
Figure out their probability distribution

Area under curve must sum to 1



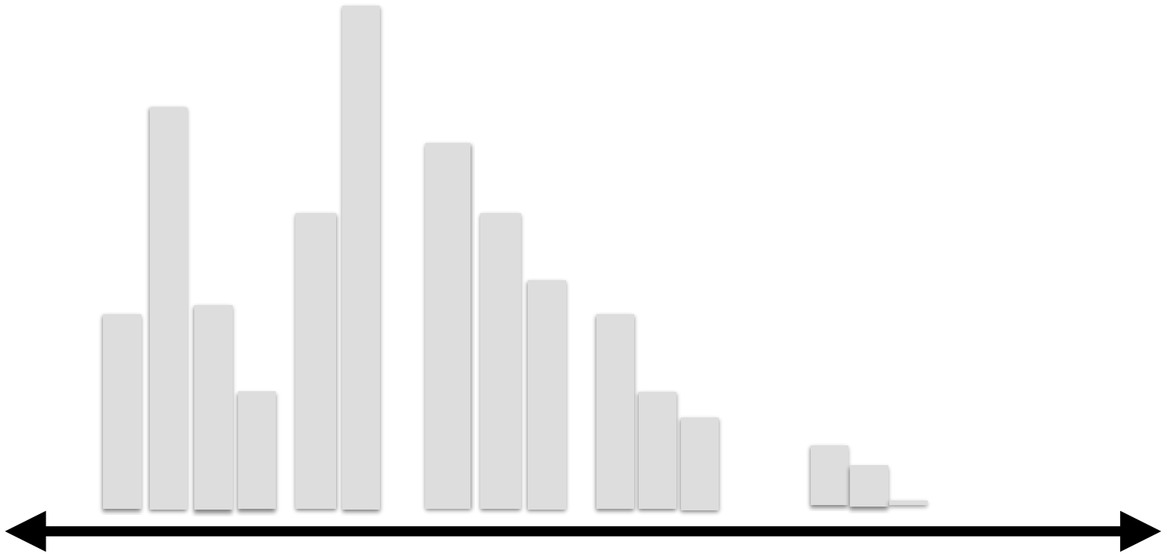
KDE is a standard technique

Non-parametric “Smoothing” technique



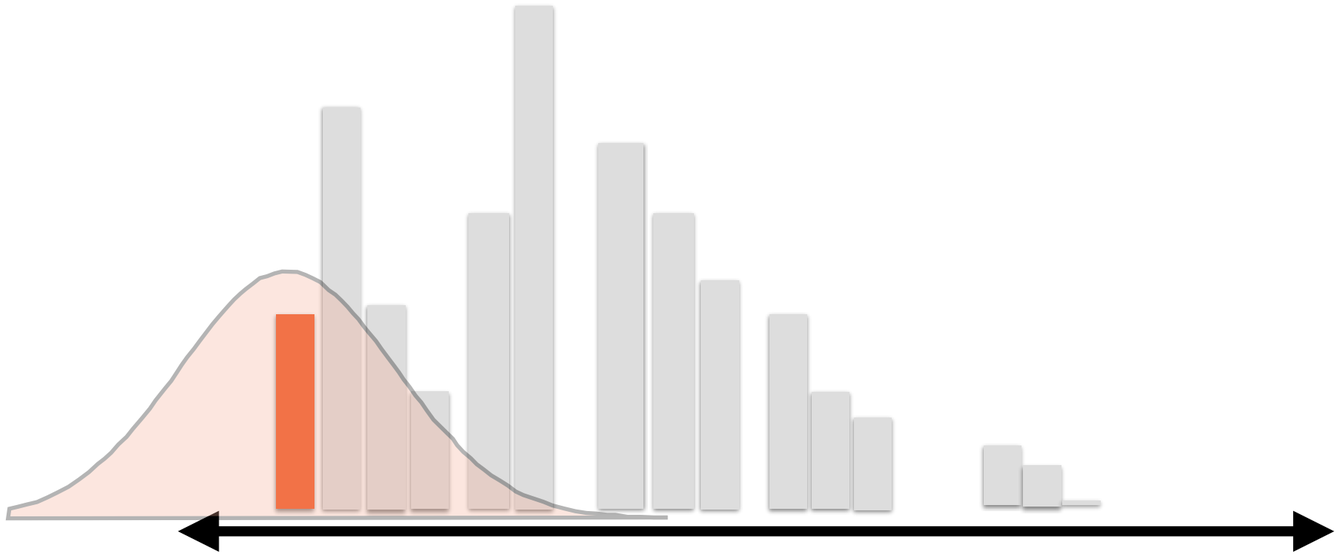
Assume points have same distribution

“Independent Identically Distributed”



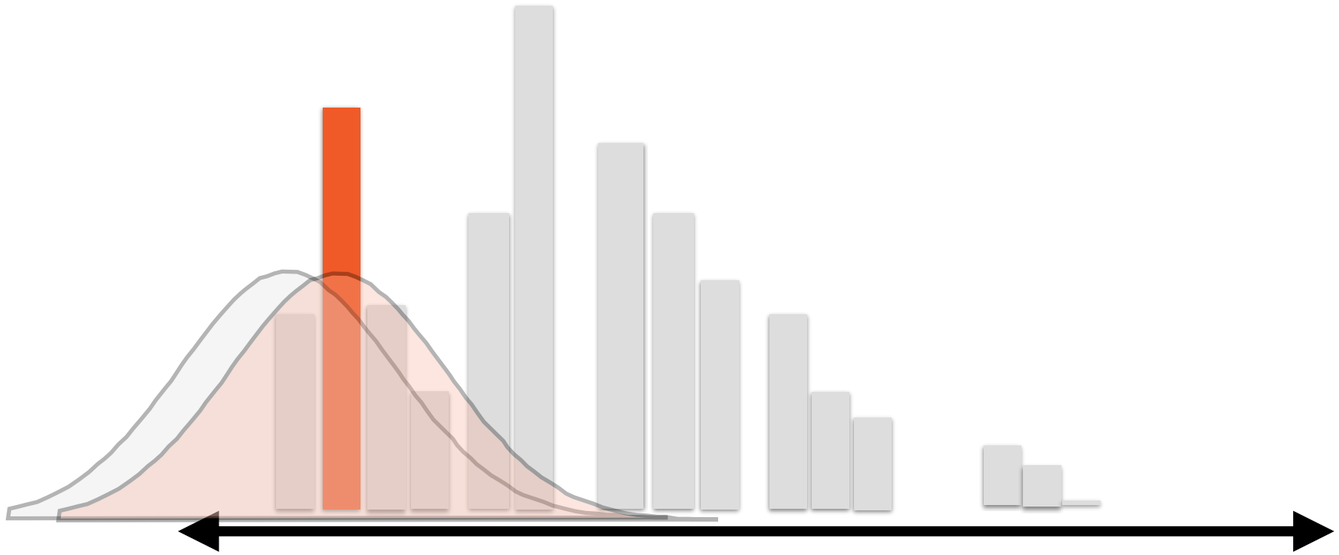
Assume points have same distribution

“Independent Identically Distributed”



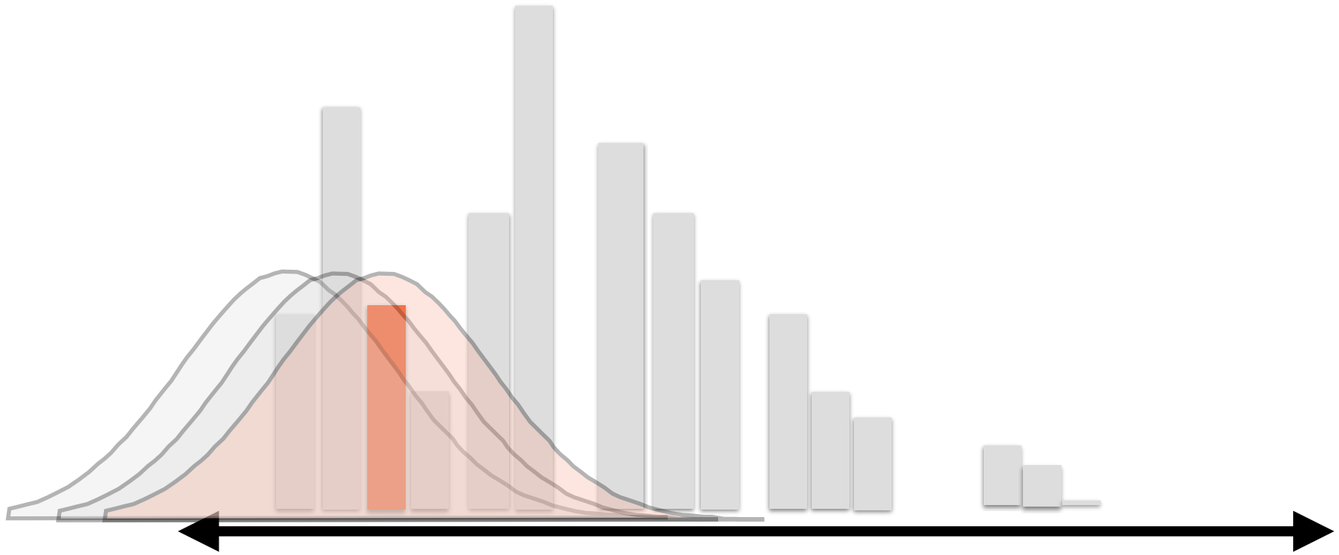
Assume points have same distribution

“Independent Identically Distributed”



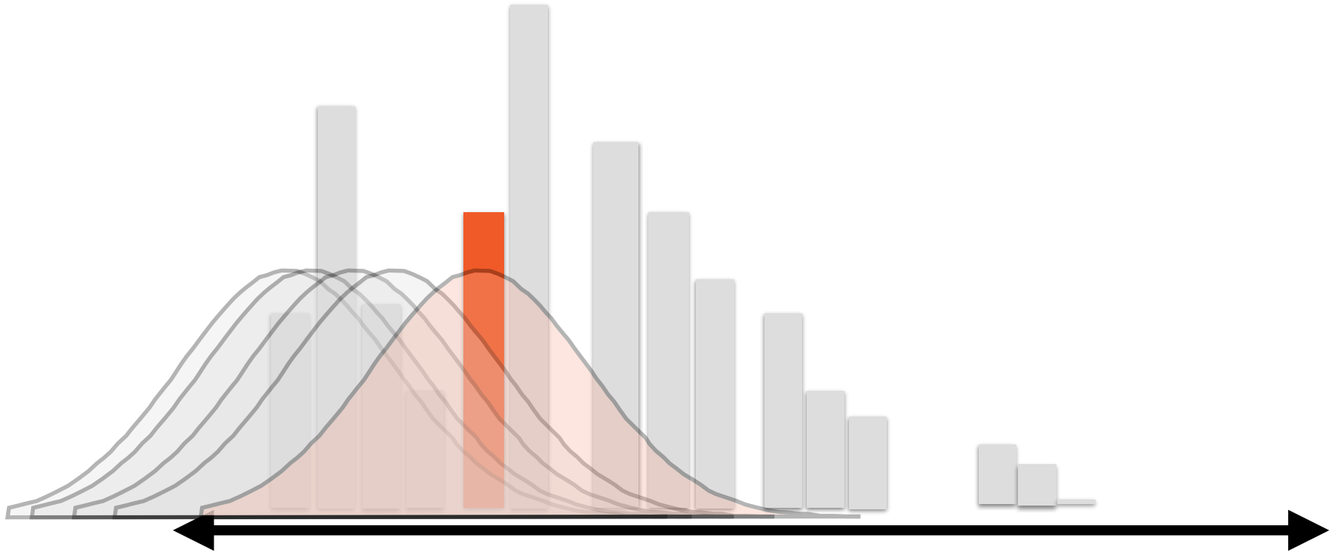
Assume points have same distribution

“Independent Identically Distributed”



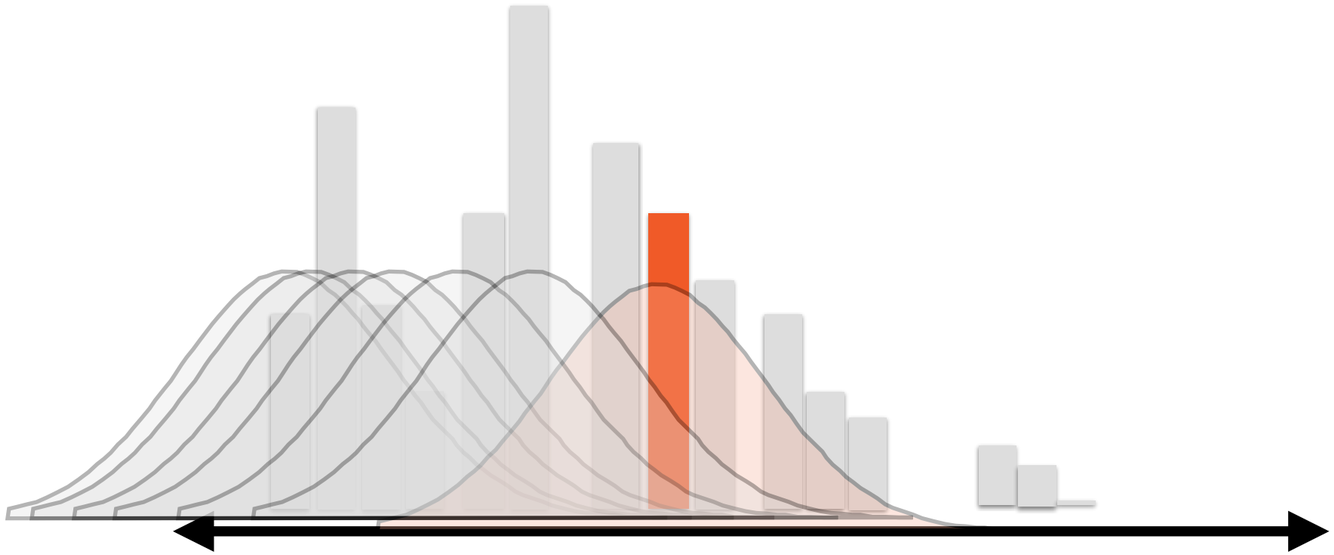
Assume points have same distribution

“Independent Identically Distributed”



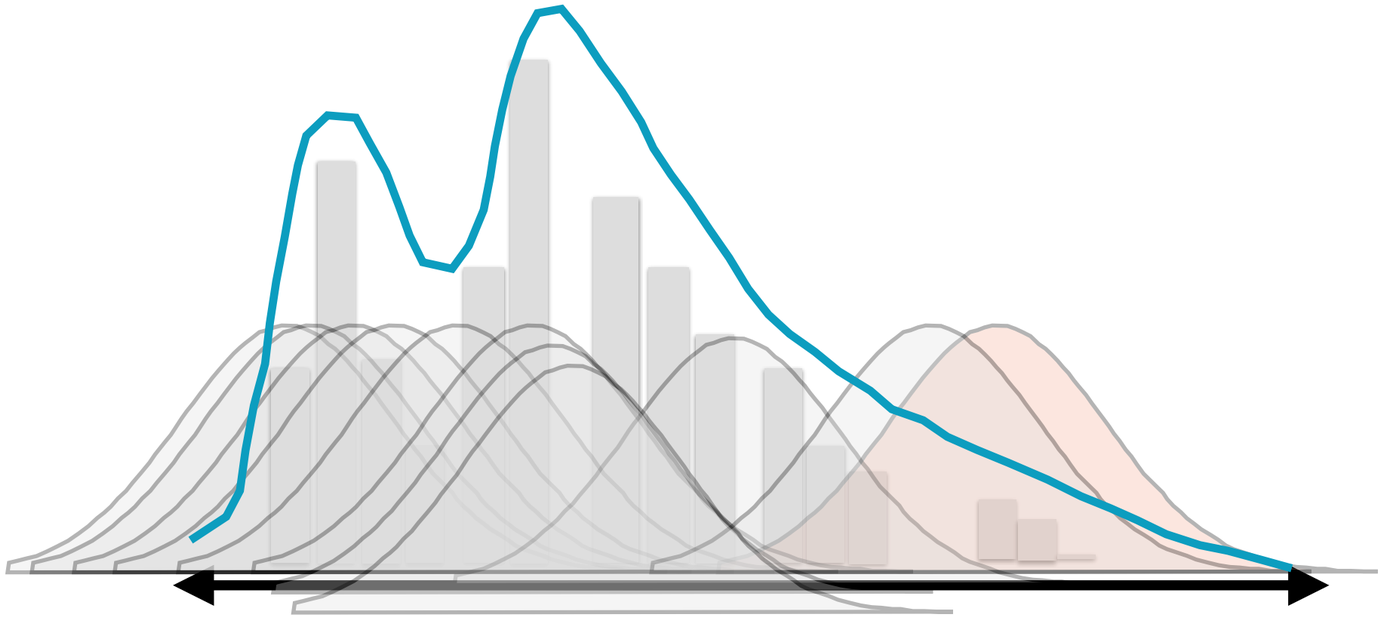
Assume points have same distribution

“Independent Identically Distributed”



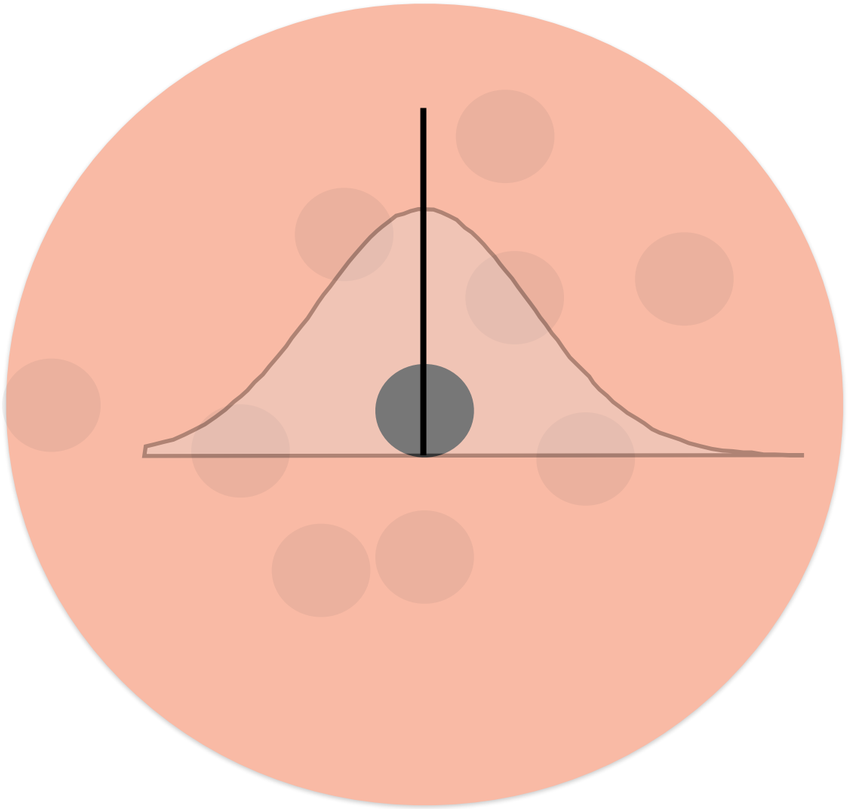
“Sum” them all up

Get resulting PDF of data

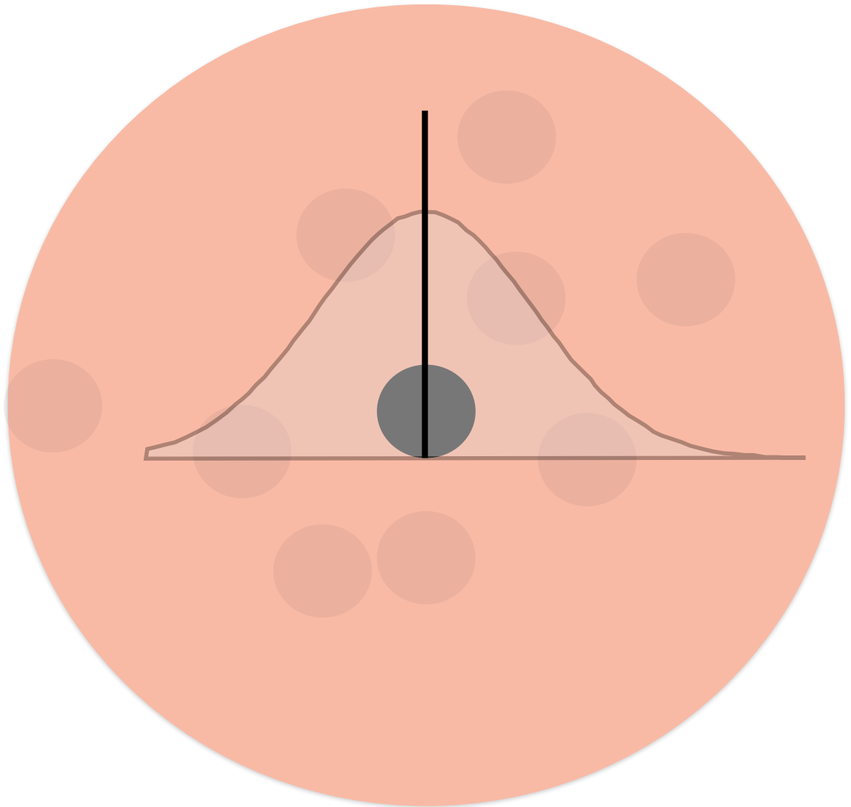


Kernel Density Estimation

Fit distribution from histogram



Gaussian Kernel



Gaussian probability distribution

Defined by

- mean μ
- standard deviation σ

Demo

Univariate and bivariate distributions

Demo

Pairwise relationships

Demo

Regression plots

Demo

Strip plots and swarm plots

Demo

Box plots, violin plots, and factor plots

Summary

Histograms, KDE plots, Rugplots to visualize univariate data

Scatter plots, Jointplots, Hexbin plots for bivariate data

Regression plots to view relationships

Pairwise relationships using PairGrid

Specialized plots for categorical data