Introduction to key Statistical Concepts

Romain Lafarguette, Ph.D. Amine Raboun, Ph.D.

Quants & IMF External Experts

romainlafarguette.github.io/ aminerabou.github.io/

Singapore Training Institute, 18 April 2023

Resources

This course is based on several free and open-source references available on line

- Mostly Harmless Econometrics: ▸ Link
- Wasserman All of Statistics ▸ Link
- Greene Econometrics ▸ Link

Outline of the Course

1. Data Concepts
2. Refresher on Probability Theory
3. Statistical Inference

Table of Contents

1 Data Concepts

2 Refresher on Probability Theory

3 Statistical Inference

Overview

Financial econometrics (including time-series econometrics) are based on four main elements:

1. A sample of data
2. An econometric model, based on a theory or not
3. An estimation method to estimate the coefficients of the model
4. Inference/testing approach to validate the estimation

Data Types

In econometrics, sets can be mainly distinguished in three types:

1. Cross-sectional data
2. Time series data
3. Panel data

Cross-Sectional Data

Cross-sectional data are the most common type of data encountered in statistics and econometrics.

- Data at the entities level: banks, countries, individuals, households, etc.
- **No time dimension**: only one "wave" or multiple waves of different entities
- Order of data does not matter: no time structure

Time Series Data

Time series data are very common in financial econometrics and central banking. They entail specific estimation methods to do the **time-dependence**.

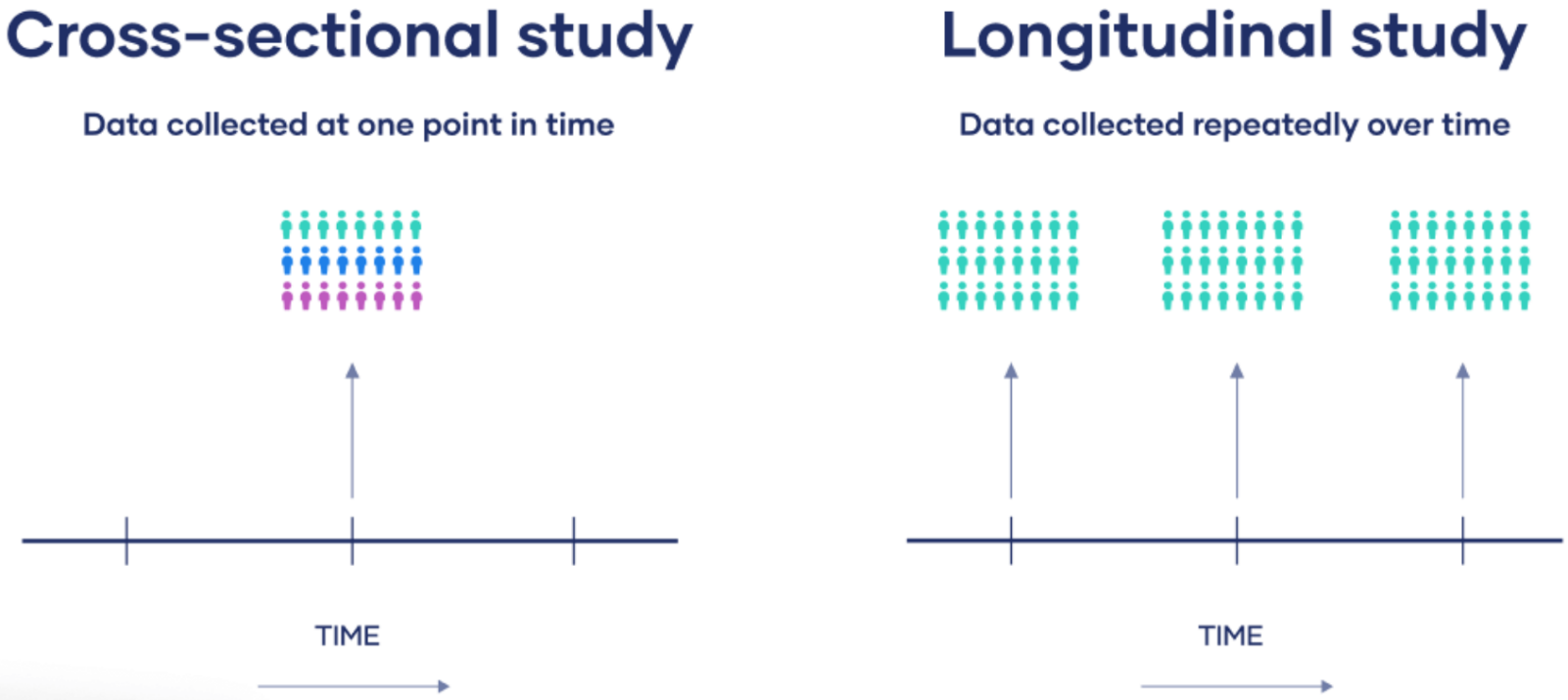
- Data for a single entity (person, bank, country, etc.) collected at multiple time periods. Repeated observations of the same variables (interest rate, GDP, prices, etc.)
- Order of data is important!
- The observations are typically not independent over time

Panel/Longitudinal Data

Panel data contain the most information and allow for more complex estimation and analysis.

- Data for multiple entities (individuals, firms, countries, banks, etc.) in which outcomes and characteristics of each entity are observed at multiple points in time
- Combine cross-sectional and time-series information
- Present several advantages with respect to cross-sectional and time series data, depending on the topic at hands

Difference between Cross-Sectional and Panel Data



Source: *cdn.scribbr.com/wp-content/uploads//2020/05/x-sectional-vs-long-graphic.png*

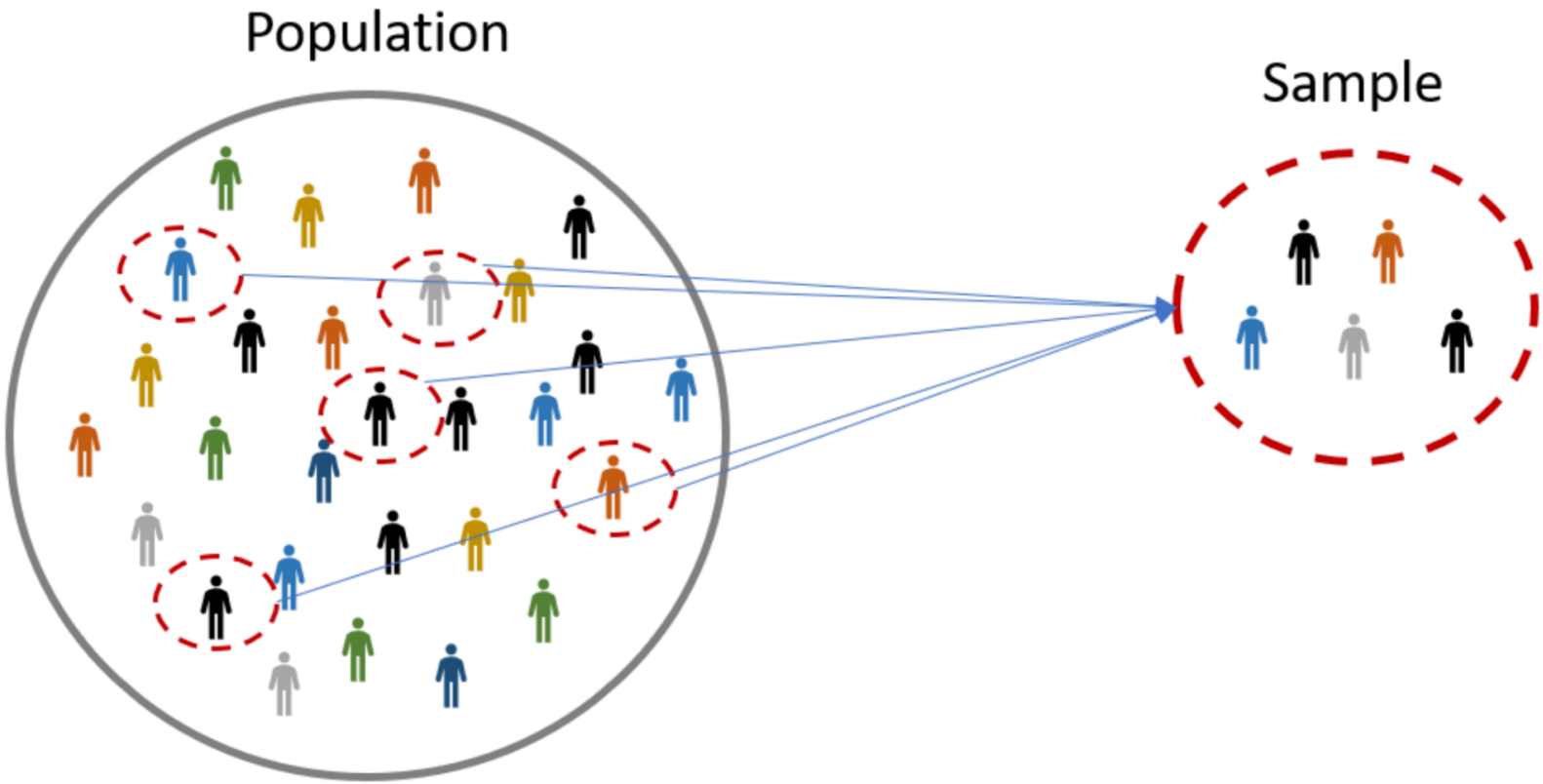
Population vs. Sample

Definition: Population

A **population** is defined as including all entities (e.g. banks or firms) or all the time periods of the processus that has to be explained

- In most cases, it is impossible to observe the entire statistical population, due to constraints (recording period, cost, etc.)
- A researcher would instead observe a **statistical sample** from the population. He will estimate an econometric model to understand the **properties on the population as a whole**.

Population vs. Sample



Credit: *medium.com/analytics-vidhya*

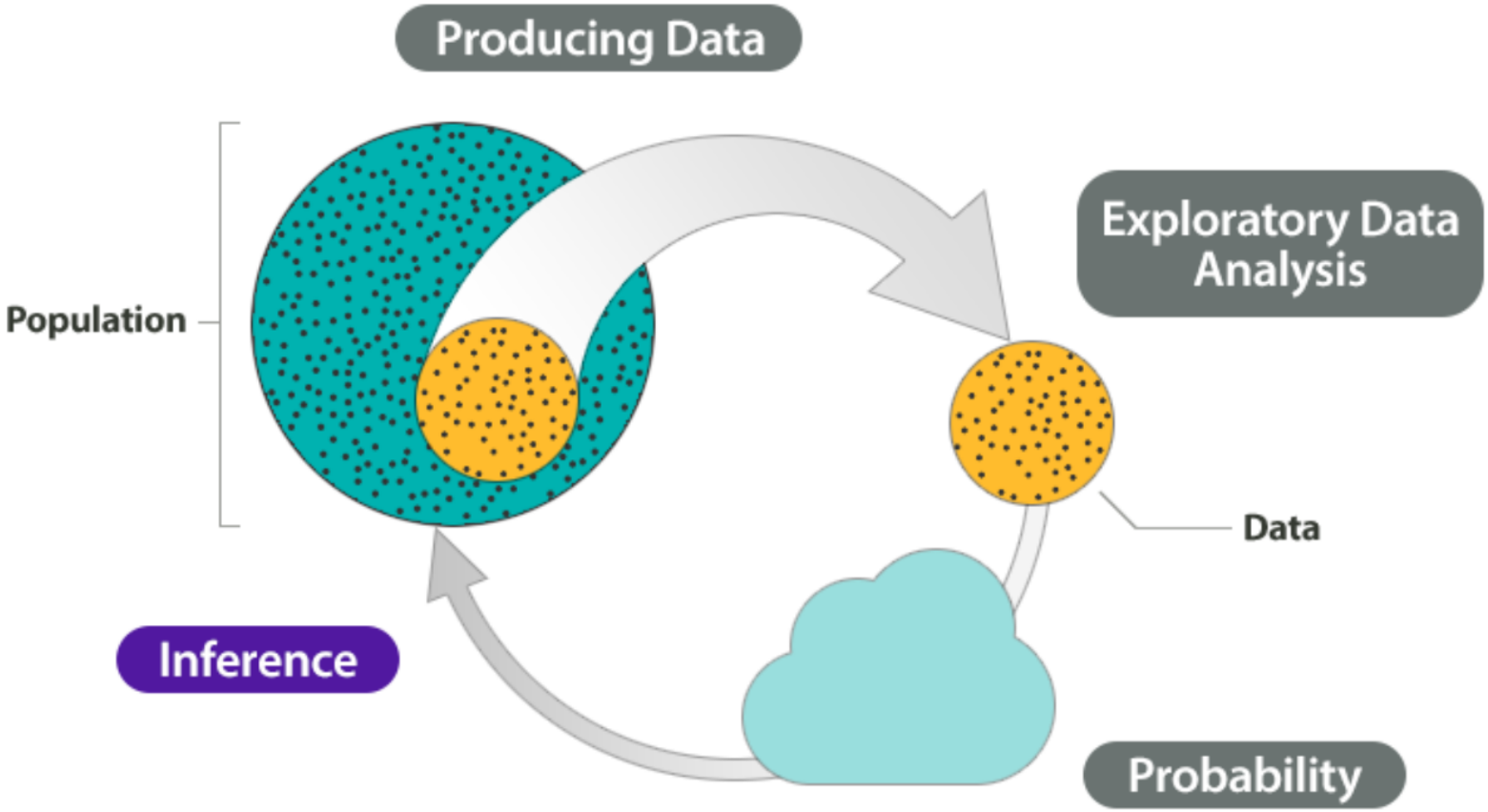
Data Generating Process

Definition: Data Generating Process

- A **Data Generating Process (DGP)** is a process in the real world that "generates" the data (or the sample) of interest.
- The process is represented by random variable (see after) X_t; the observation x_t is one possible realization of X_t
- Given that we observe a set of $x_1, \ldots, x_T$ what can we **infer** about the process $X_1, \ldots, X_T$ that has generated them?

1. DGP: "true" model that generated the data $x_1, \ldots, x_t$
2. But we only observe the time series a **finite number of times**
3. However, it is convenient to allow - theoretically - the number of observations to be **infinite**: $\{X_t\}_{t \in \mathbb{Z}}$. In this case,$\{X_t\}_{t \in \mathbb{Z}}$ is called a discrete time **stochastic process**

DGP, Population and Sample



Source: *bookdown.org/cristobalmoya*

Example: Data Generating Process

Let us assume that there is a linear relationship between interest rates in two countries (R, R^*), their forward (F) and their spot exchange rate (S).

$$\frac{F}{S} = \frac{1+R}{1+R^*}$$

This non-arbitrage relationship (CIP) can be used in the foreign exchange market to determine the forward exchange rate

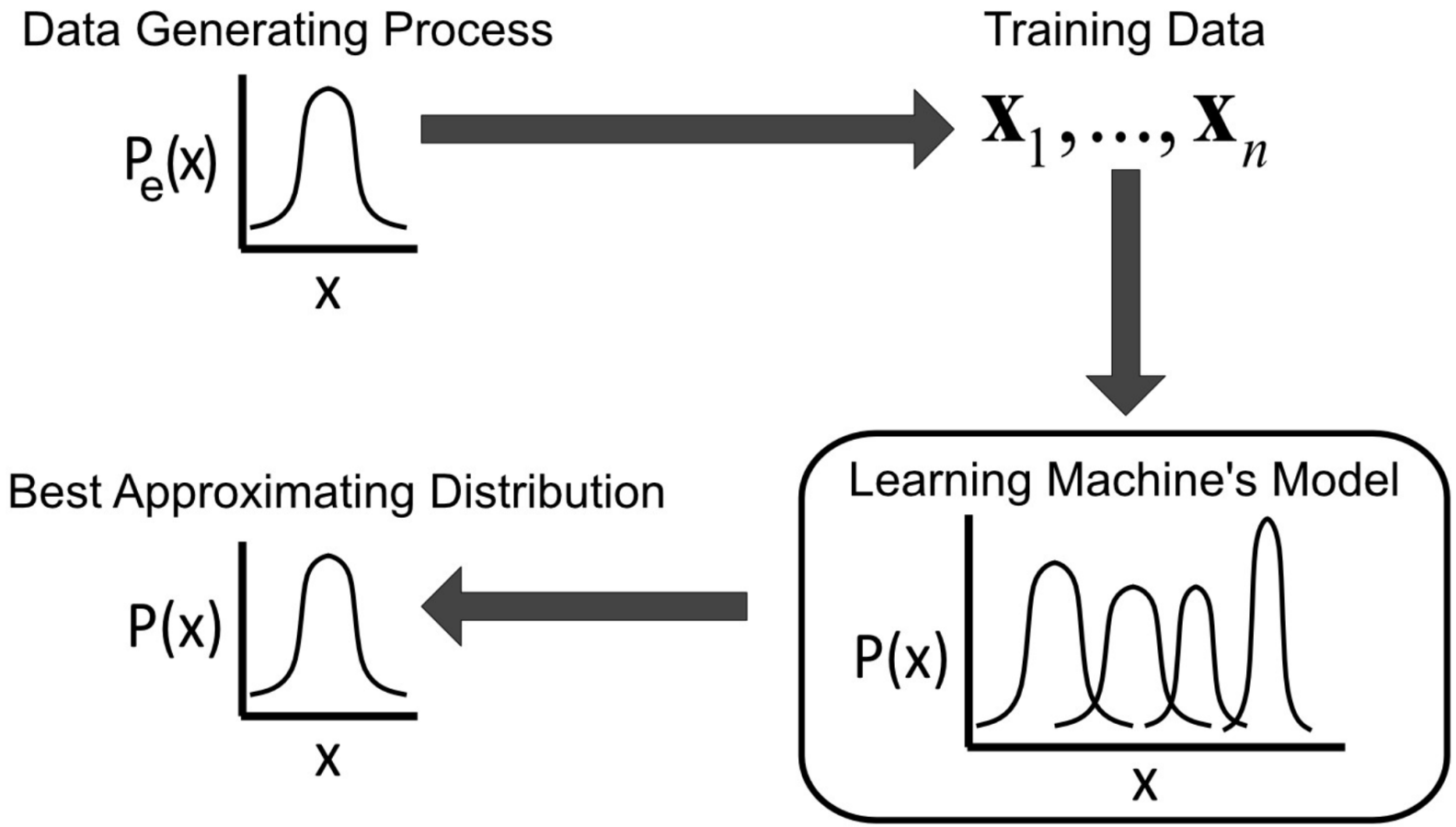
$$\mathbb{E}[F|S = s, R = r, R^* = r^*] = s * \frac{1+r}{1+r^*}$$

This relationship is the **Data Generating Process** for F

The equivalent of population for time series econometrics is the DGP.
NB: note that I use R to describe the random variable and r to describe its realization

Econometrics Challenge

The challenge of econometrics is to draw conclusions about a DGP (or population), after observing only one realization $\{x_1, \ldots X_N\}$ of a random sample (the dataset). To this end, we use a statistical - or econometrics - model



Source: *statisticalmachinelearning.com/wp-content/uploads/2019/12/SMLframework2.jpg*

Table of Contents

1 Data Concepts

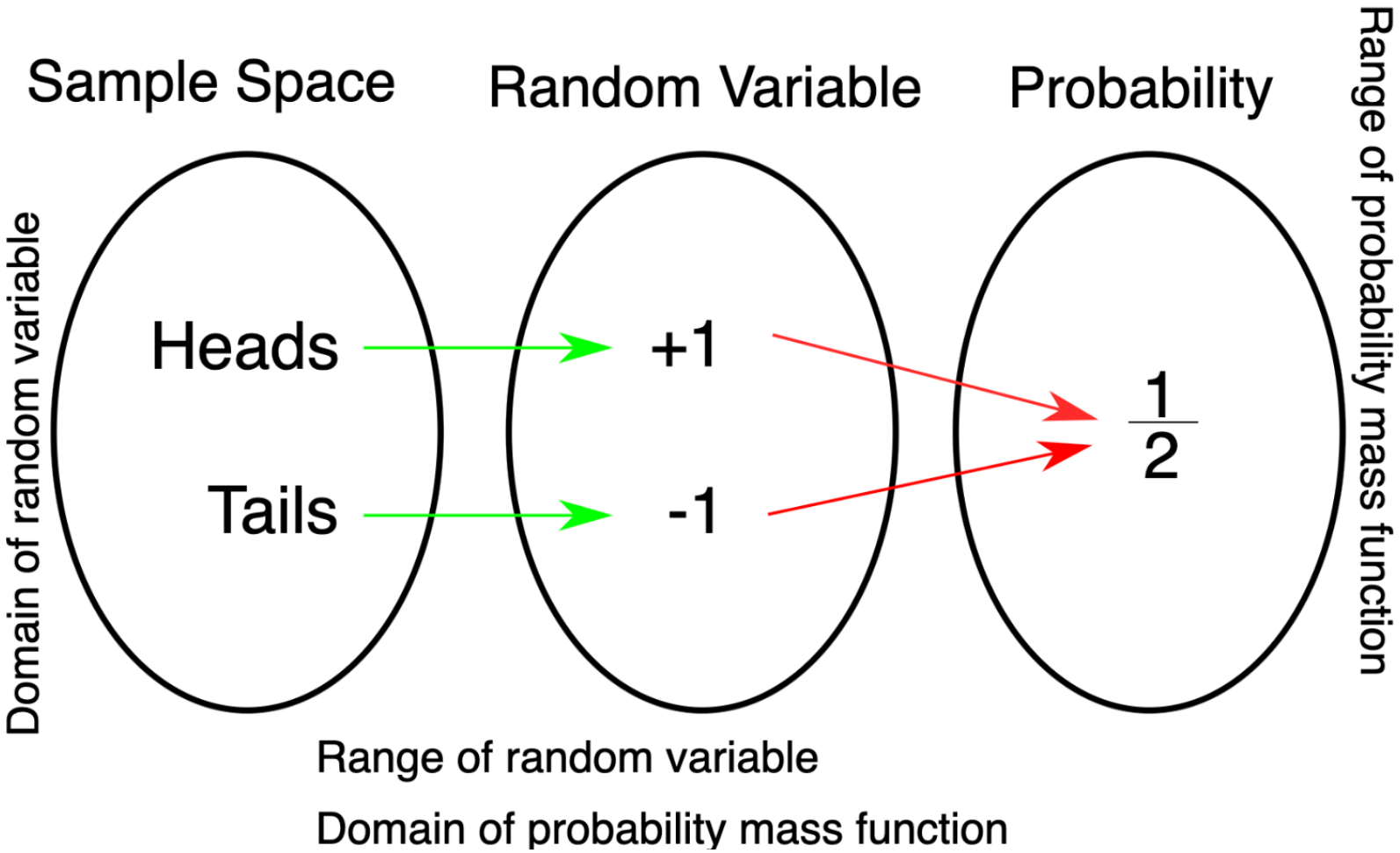
2 Refresher on Probability Theory

3 Statistical Inference

Refresher: Random Variables

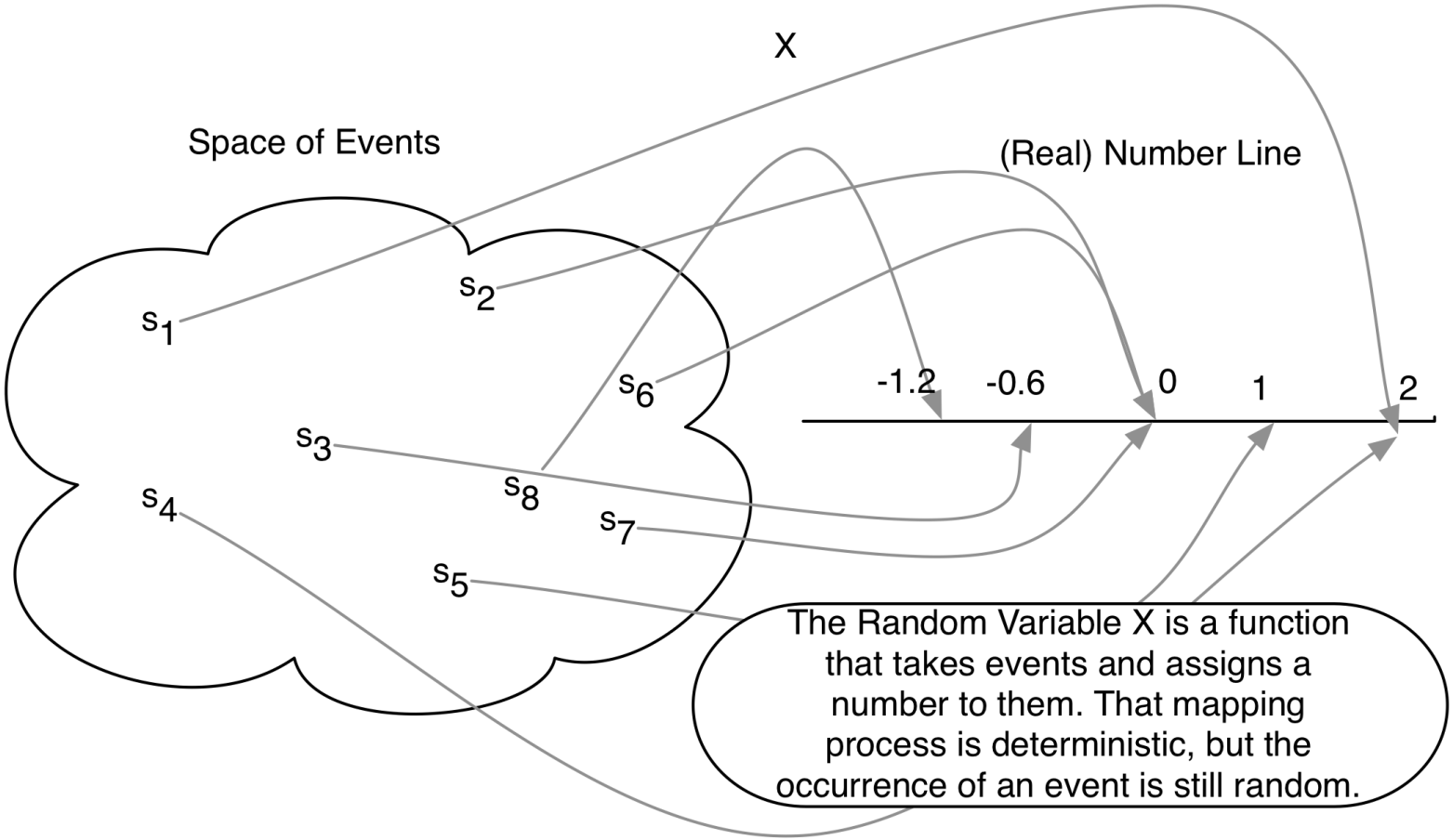
- Mathematicians are formalizing and modeling randomness via the concept of **random variables**.
- Pay attention: a random variable is neither random (it is formalized via laws and distributions), not a variable (it is a function)
- A **random variable** is a function $f : \Omega \mapsto \mathcal{R}$ that assigns to a set of outcome Ω a **value**, often a real number.
- The probability of an outcome is equal to its **measure** divided by the measure of all possible outcomes
  - Example: obtaining an even number by rolling a dice: $\{2, 4, 6\}$
  - Probability to obtain an even number by rolling a dice: $m(\{2, 4, 6\})/m(\{1, 2, 3, 4, 5, 6\}) = \frac{1}{2}$ *(here, the measure simply "counts" the outcomes with equal weights)*

Random Variables: Intuition



Source: *Wikipedia*

Random Variables: Mapping



Credit: *iqss.github.io/prefresher/images/rv.png*

Random Variables (II)

- Random variables are the "building block" of statistics:
  - Random variables are characterized by their distribution (generating function, moments, quantiles, etc.)
  - The behavior of two or more random variables can be characterized by their dependence/independence, matrix of variance-covariance, joint distribution, etc.
  - The main theorems in statistics (law of large numbers, central limit theorem, etc.) leverage the properties of random variables

Refresher: Probability Distributions for Continuous Variables

Definition: Probability Distribution

The probability distribution of a random variable describes how the probabilities of the outcomes are distributed. How more likely is one outcome compared to another? Is there more upside risk or downside risk? Etc.

Distributions are equivalently represented by their:

1. **The probability density function (pdf)**
2. **The cumulative distribution function (cdf)**

The pdf is also necessary to define distributions' moments (see after).

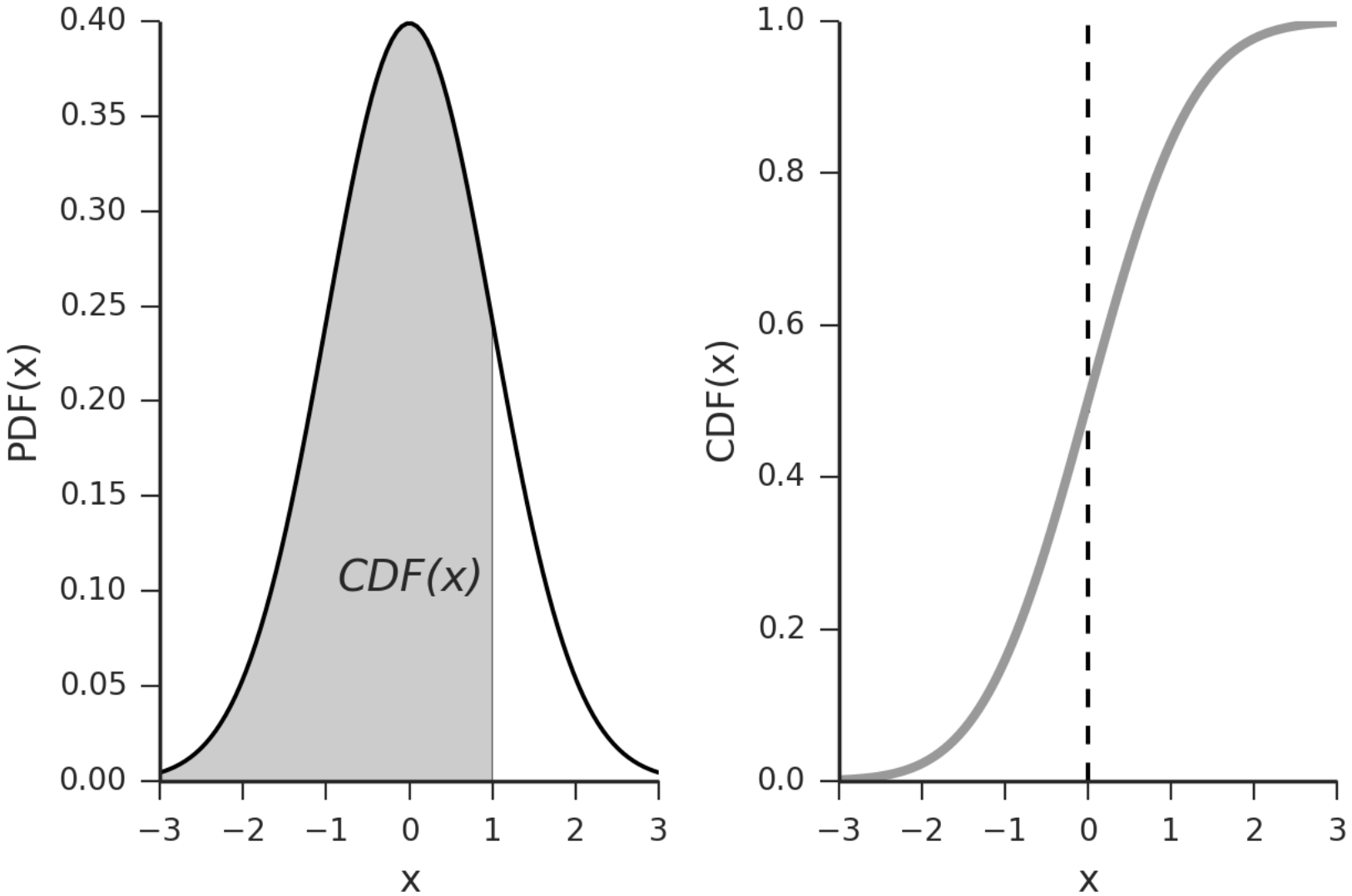
PDF and CDF

- **The probability density function (pdf)** usually denoted $f_X(x0)$ (remember than X stands for the random variable, while $x0$ stands for a realization of the random variable) represents the relative likelihood that the random variable X will fall within a small neighborhood of $x0$ (infinitesimal concept). It is easier to conceptualize the pdf via the cdf

- **The cumulative distribution function (cdf)** usually denoted $F_X(x0)$ represents the probability that the random variable will be lower than $x0$. It cumulates the pdf ("all the small neighborhoods") such that:

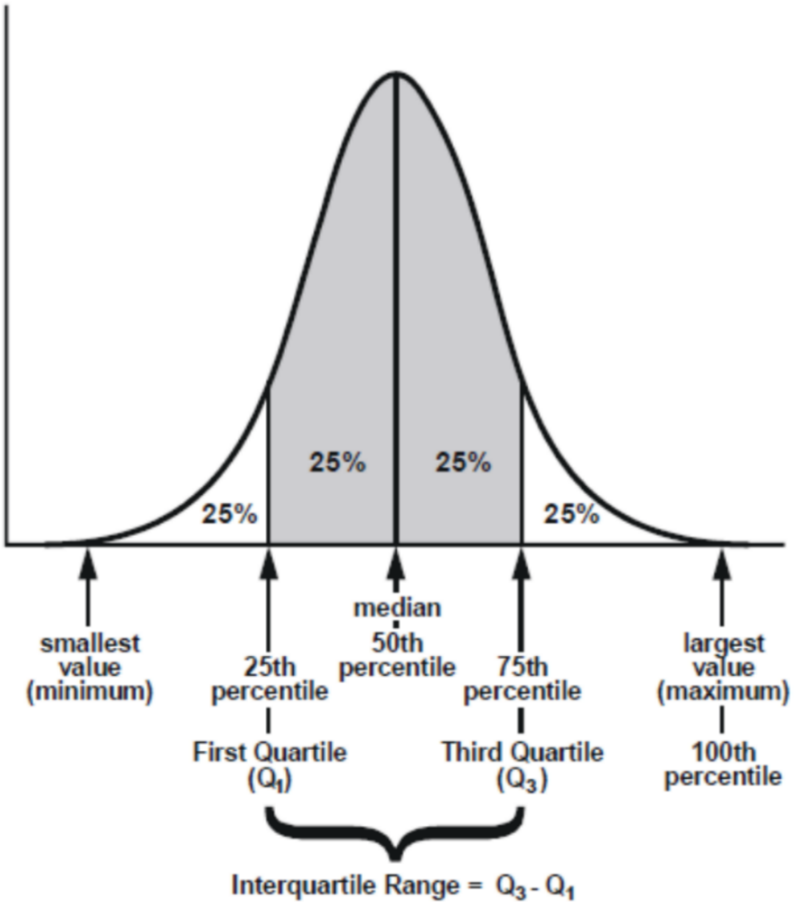
$$F_X(x0) \equiv \mathbb{P}[X \leqslant x0] = \int_{-infty}^{x0} f_X(h)dh$$

Link between PDF and CDF



Quantiles

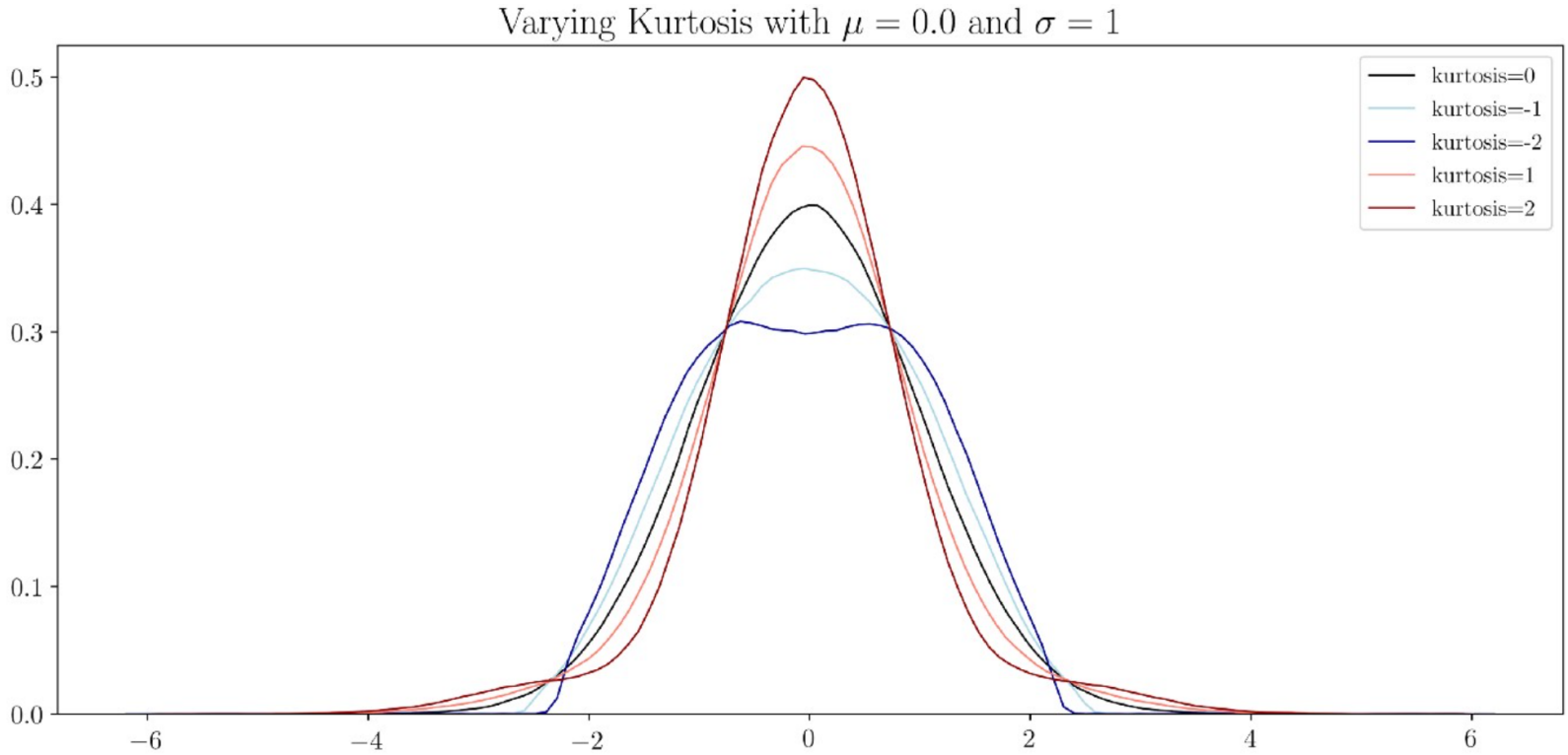
- Quantiles are cut points dividing a PDF into intervals of same probability
- Quantile Q at probability p: $\mathbb{P}[X \leq Q] = p$
  - The quantile function $\mathcal{Q}(p) = \inf\{y \in \mathbb{R} : p \leq F_X(y)\}$ is the **inverse cumulative distribution function**
- The median is the quantile at 50% (half observations below)
- Distributions can be characterized by their PDF, CDF or quantile functions (among other)



Moments: Overview

	Formula	Interpretation
Mean	$\mathbb{E}[X_t] = \mu$	Central tendency
Variance	$\mathbb{V}[X] = \mathbb{E}\left[(X_t - \mu)^2\right] = \sigma^2$	Dispersion around μ
Skewness	$\mathbb{S}[X] = \mathbb{E}\left[(X_t - \mu)^3\right] = \text{sk}$	Symmetry
Kurtosis	$\mathbb{K}[X] = \mathbb{E}\left[(X_t - \mu)^4\right] = \kappa$	Tail heaviness

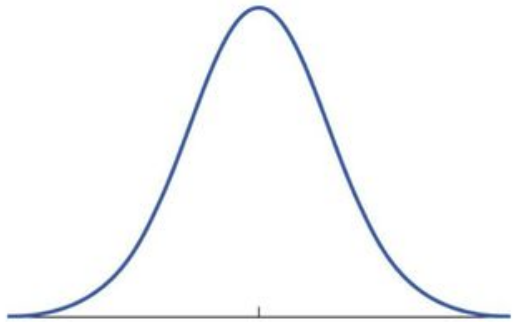
Mean, Variance, Kurtosis



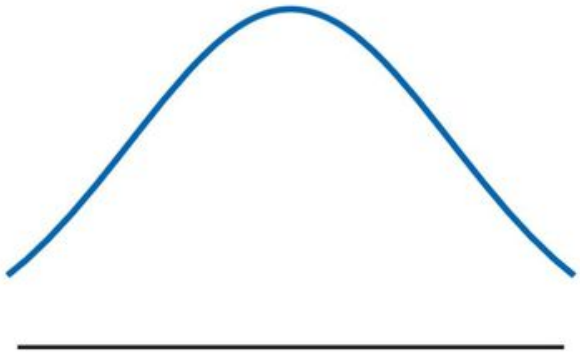
Credit: *iqss.github.io/prefresher/images/rv.png*

Kurtosis: Benchmarking against the Normal Distribution

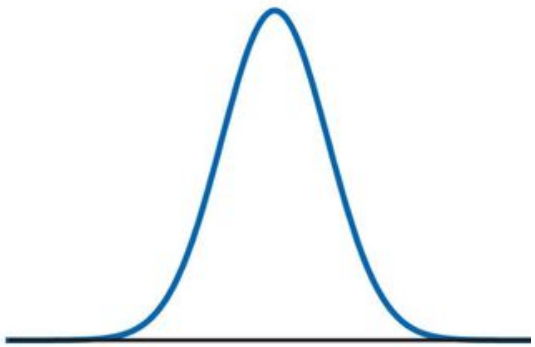
There are different shapes of kurtosis



Normal distribution

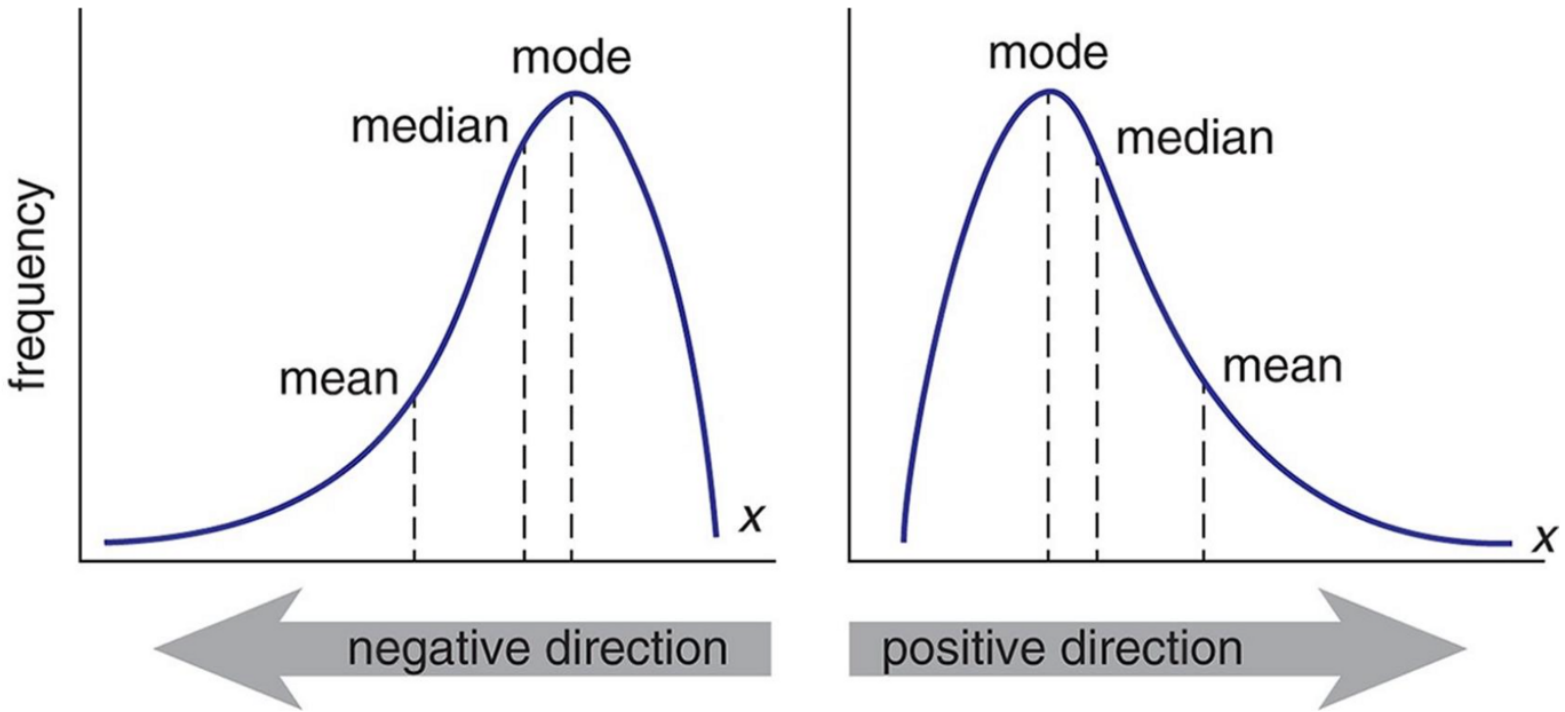


Heavy Tails



Light Tails

Skewness



Credit: *iqss.github.io/prefresher/images/rv.png*

Moments: In Practice

- The moments allow to characterize the shape of the returns distribution
- However, the theoretical moments are **unobservable** and need to be estimated
- Assume that we have a sample $\{x_1, \ldots, x_T\}$ realizations of the sequence of X_t

Table of Contents

1 Data Concepts

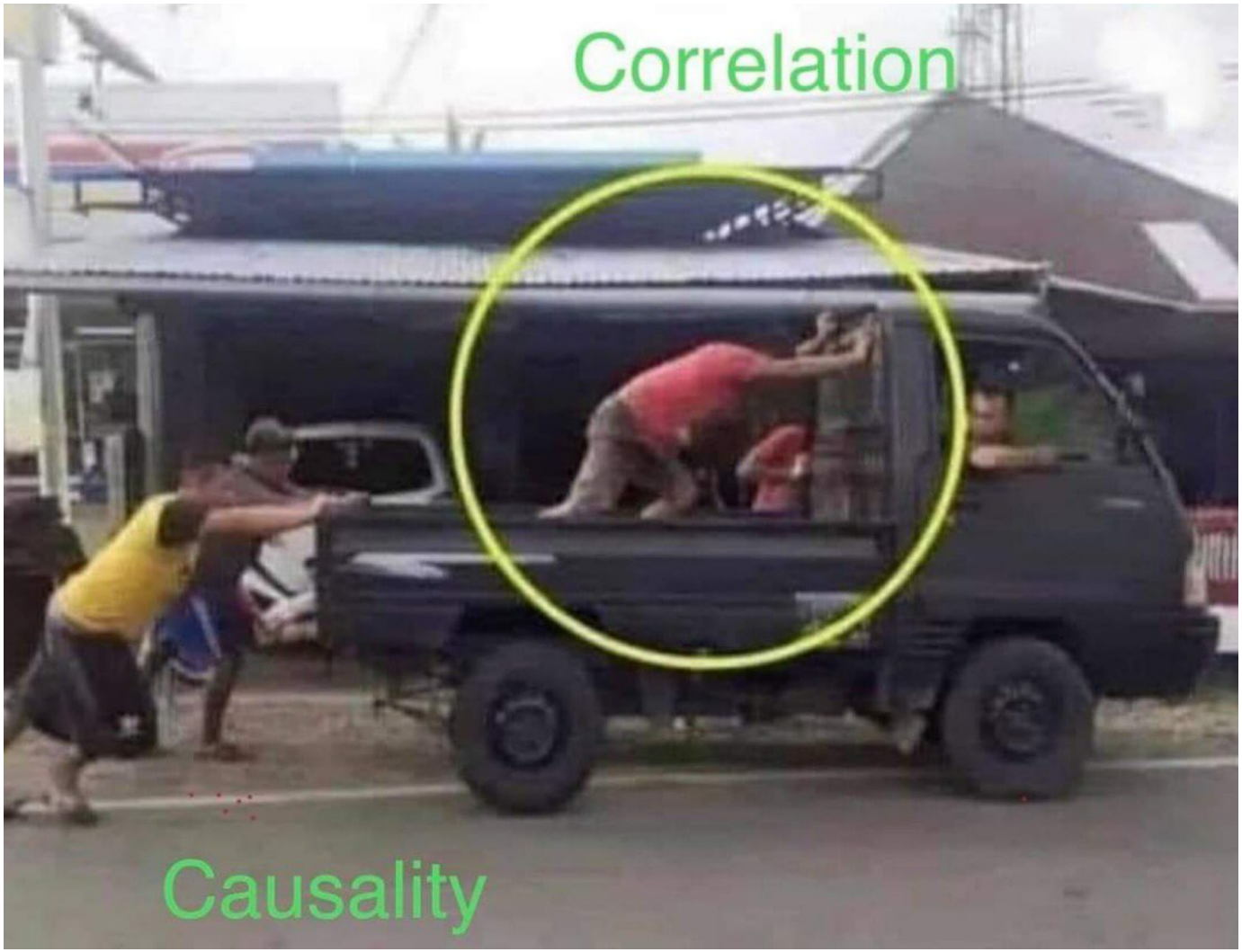
2 Refresher on Probability Theory

3 Statistical Inference

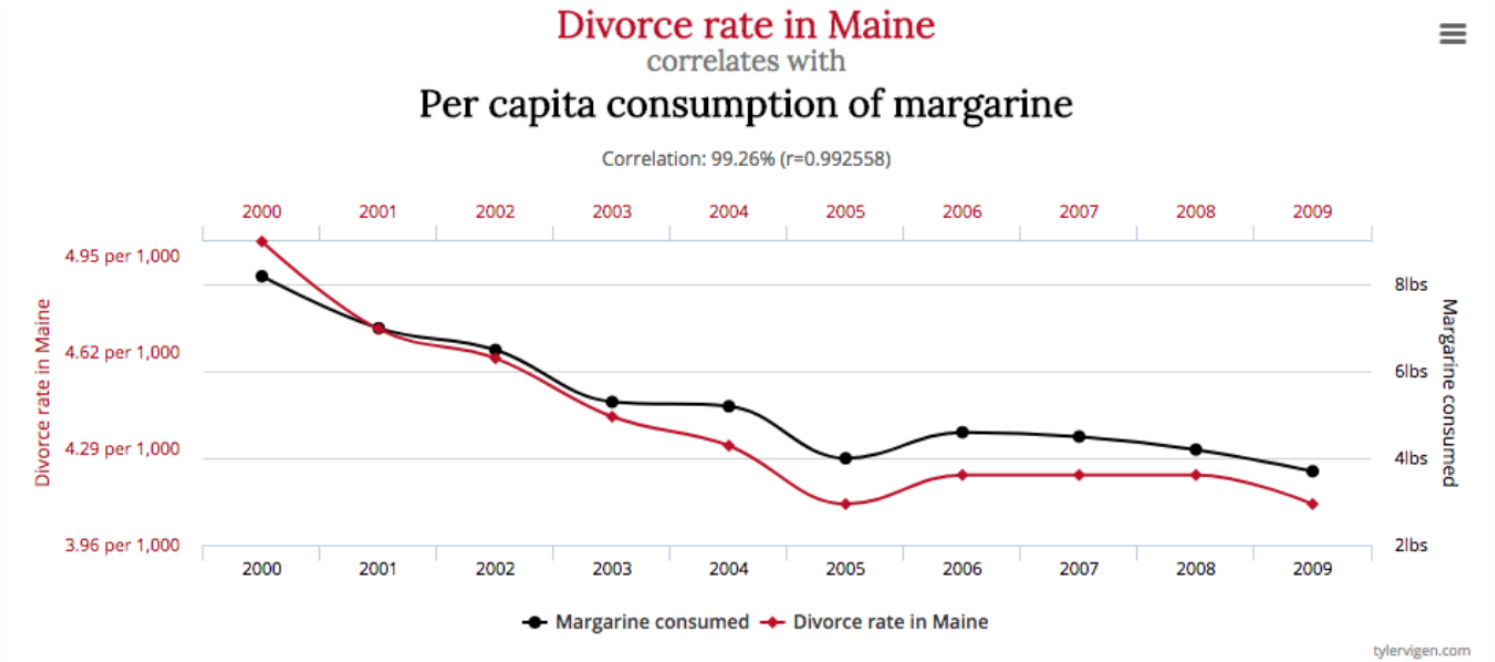
Statistical Inference from a Frequentist Point of View

- Postulate a certain simplified representation of the world
- Postulate that the data that we observe follows a certain data generating process and an error term between the model and the reality: $Y = f(X) + \epsilon$
- Statistical inference is about estimating the relationship $\hat{f}$ such that it minimizes the error term
  - For instance, in the linear model $Y = \alpha + \beta X + \epsilon$
  - Estimate the coefficients $\hat{\alpha}, \hat{\beta}$

Correlation is not Causation !



Spurious Relationship



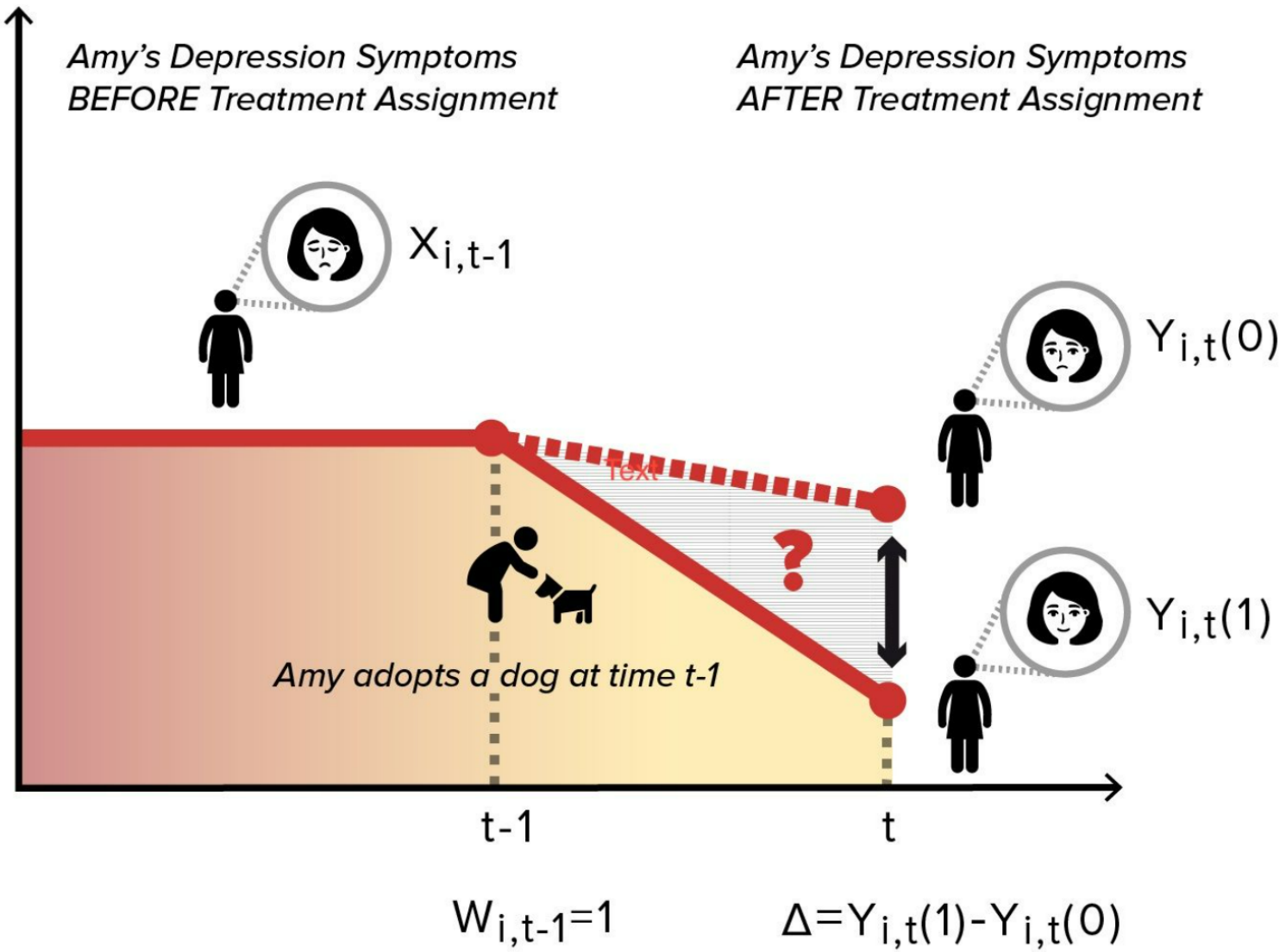
Causality

- To identify causality, more advanced approaches than standard regressions are needed

- Need to be able to isolate the effect and build a "counterfactual": what would have happened if everything else would have been the same EXCEPT the variable I am interested in?

- Need different set-up, from the "cleanest" to the more endogenous one:
  1. Randomized control experiment (in a lab/on the field)
  2. Natural experiment (e.g. Cuba and Mariel)
  3. High frequency identification (FXI on Russian data)
  4. Regression discontinuity (law of one price at the Canada/US border)
  5. Instrumental variables (storms and fish price)
  6. Synthetic matching

Counterfactual Analysis



Estimator

Definition: Estimator

An **estimator** is any function $F(x1, \ldots, x_t)$ of a sample. Note that any descriptive statistics is an estimator (a simple one)

Example: Sample Mean

The sample mean (or overage) of a sample is an estimator of the (theoretical) mean $\mathbb{E}[X_t] = \mu$.
The estimator is simply: $\hat{\mu}_t \equiv \bar{X}_t = \frac{1}{T} \sum_{t=1}^{T} x_t$

Example: Variance

Example: Sample Mean

Assume that the observations are drawn from $i.i.d$ random variables. The **sample variance** $\hat{\sigma}_T^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x}_t)^2$

Note: the denominator is equal to *T-1* as to define a sample variance corrected for the small sample bias.

Sampling Distribution

Fact

An estimator $\hat{\theta}$ is a **random variable**

Therefore, $\hat{\theta}$ has a (marginal or conditional) **probability distribution**. This sampling distribution is characterized by a probability distribution function (pdf) $f_{\hat{\theta}}(u)$

Definition: Sampling Distributions

The probability distribution of an estimator is called the **sampling distribution**

The sampling distribution is described by its moments, such as expectations, variance, skewness, etc.

Point Estimate

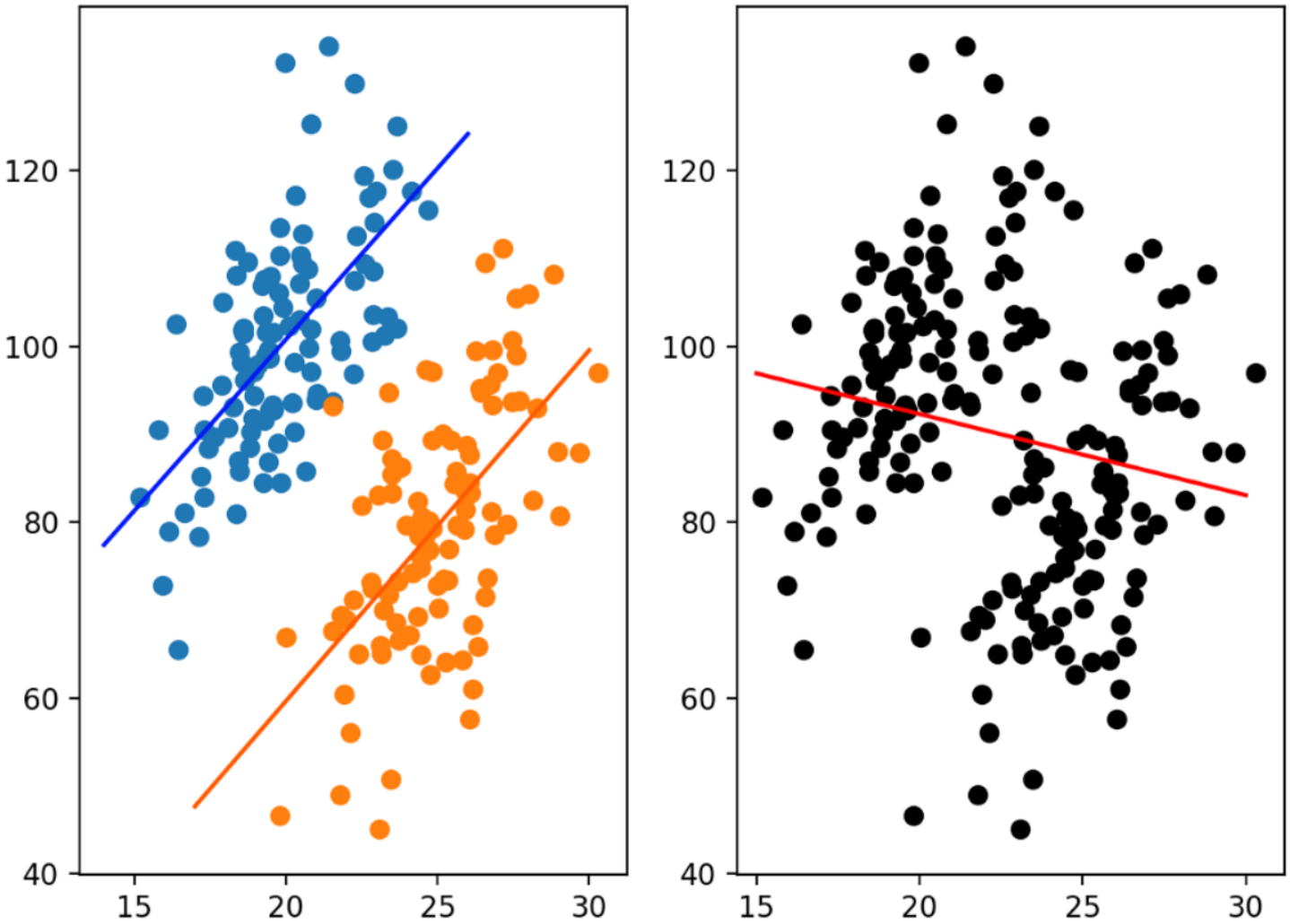
Definition: Estimate

An estimate is the realized value of an estimator (e.g. a number, in a case of a point estimate) that is obtained for a particular value $x0$. Often noted as $\hat{\theta}(x0)$ for the estimator $\hat{\theta}$

Estimate of a linear regression

- DGP $Y = \alpha + \beta * X + \epsilon$, with joint sample $y1, \ldots, y_T, x1, \ldots, x_T$.
- We have an estimator (for instance an OLS) of $\hat{\alpha}, \hat{\beta}$
- Then, for any value of $X = x_0$, we can simply project the **conditional expected estimate** $y_0 = \hat{\alpha} + \hat{\beta} * x_o + \hat{\epsilon}$
- If the estimator is unbiased, the fitted residuals $\hat{\epsilon} = y_t - \hat{\alpha} - \hat{\beta} * x_t$ are centered on **average**:$\mathbb{E}[\hat{\epsilon}] = 0$. This is why the residual disappear from the estimate of the conditional expected estimate in an OLS... but no bias doesn't mean no variance !
- ϵ & $\hat{\epsilon}$ are random variables: they determine $\hat{\alpha}, \hat{\beta}$ distributions

Simpson Paradox

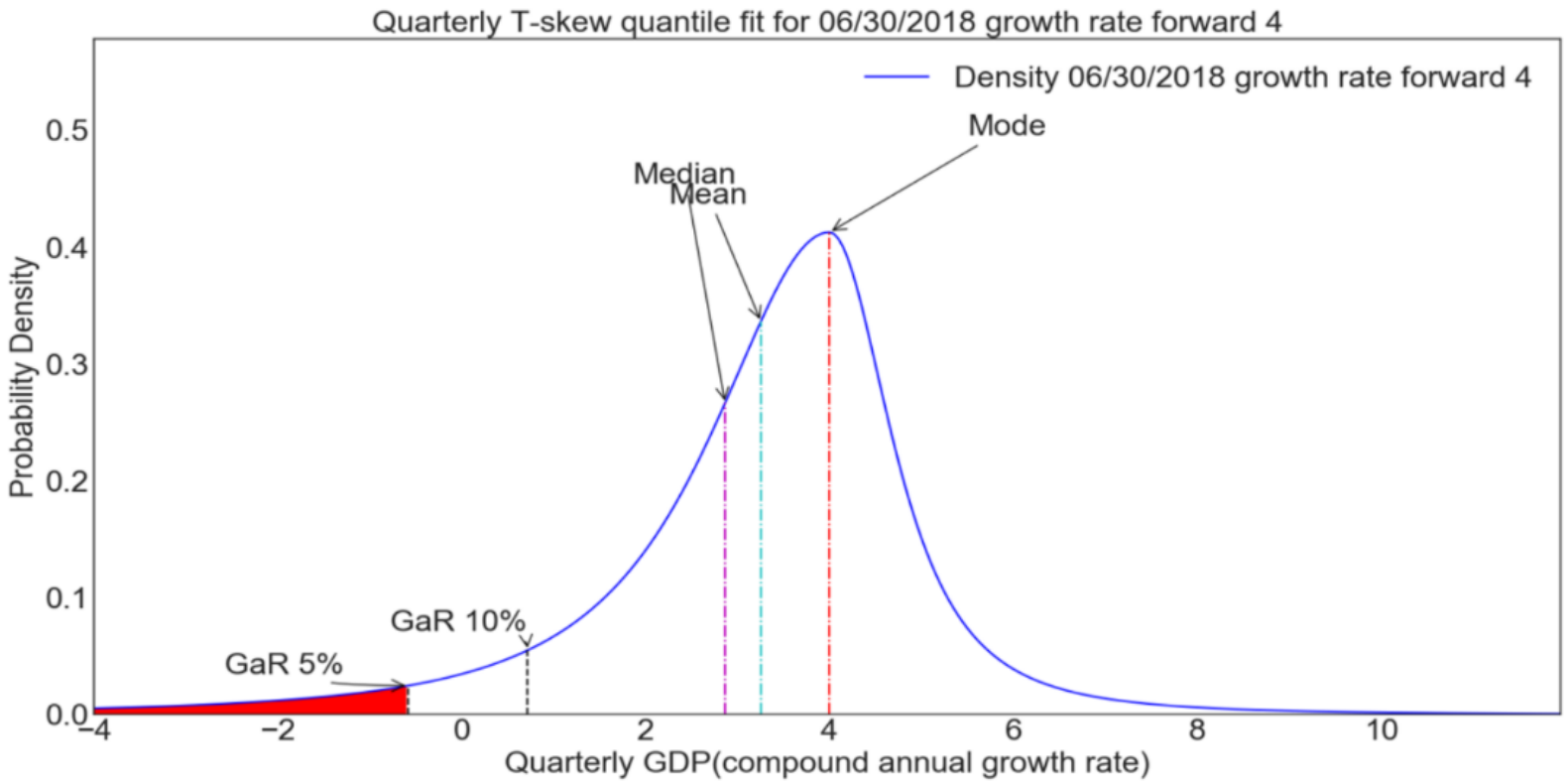


Source: *Lafarguette et al. (2019)*

Density Estimate

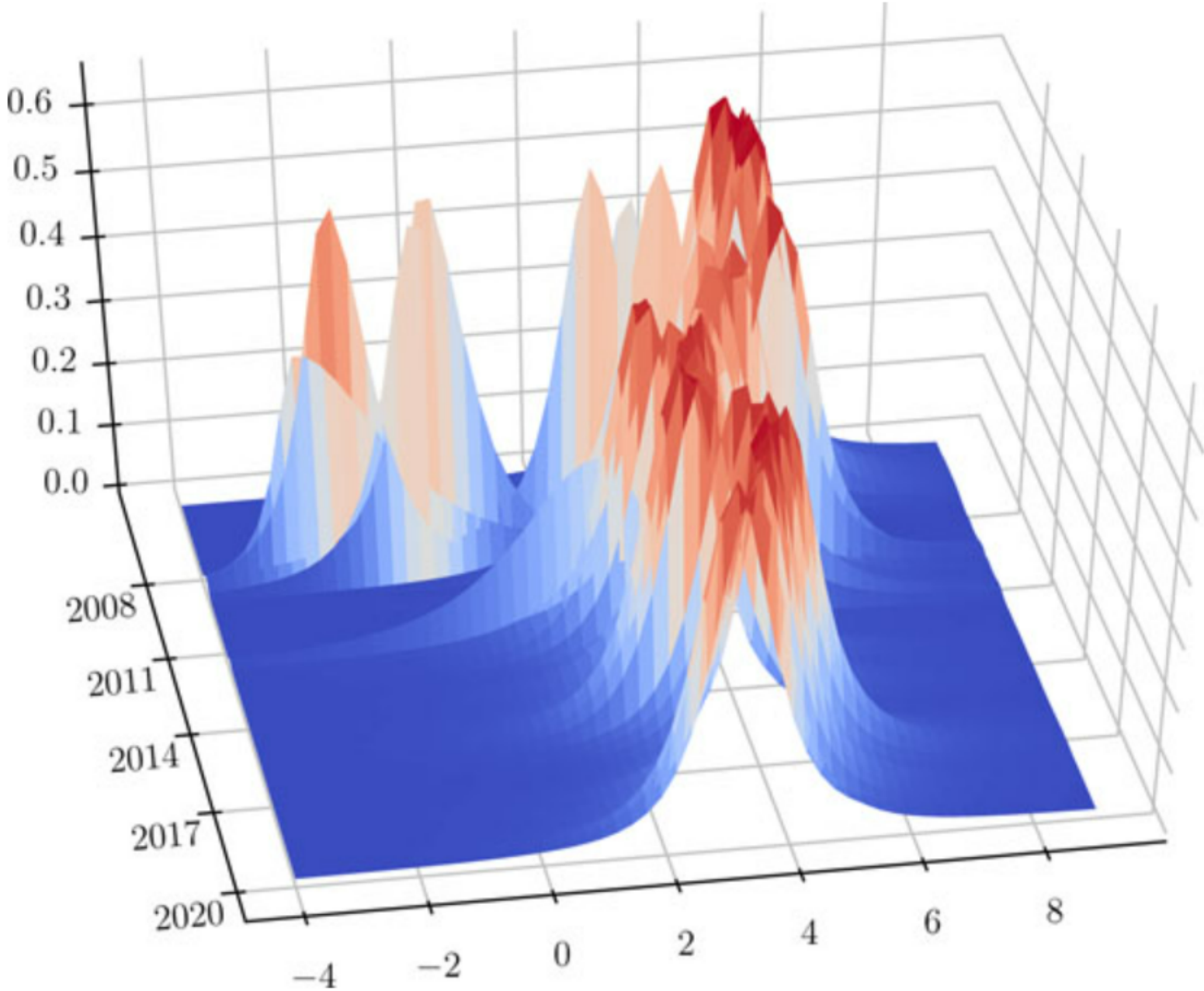
- Don't model only a point (the mean often), but the entire conditional distribution
- Interested in the **conditional future distribution** $P[Y_{t+h}|X_t]$
- Critical to estimate risks, incorporate variance, etc.
- Assume to use specialized models, that are going to forecast different points of the distribution

Density Modeling



Source: *Lafarguette et al. (2019)*

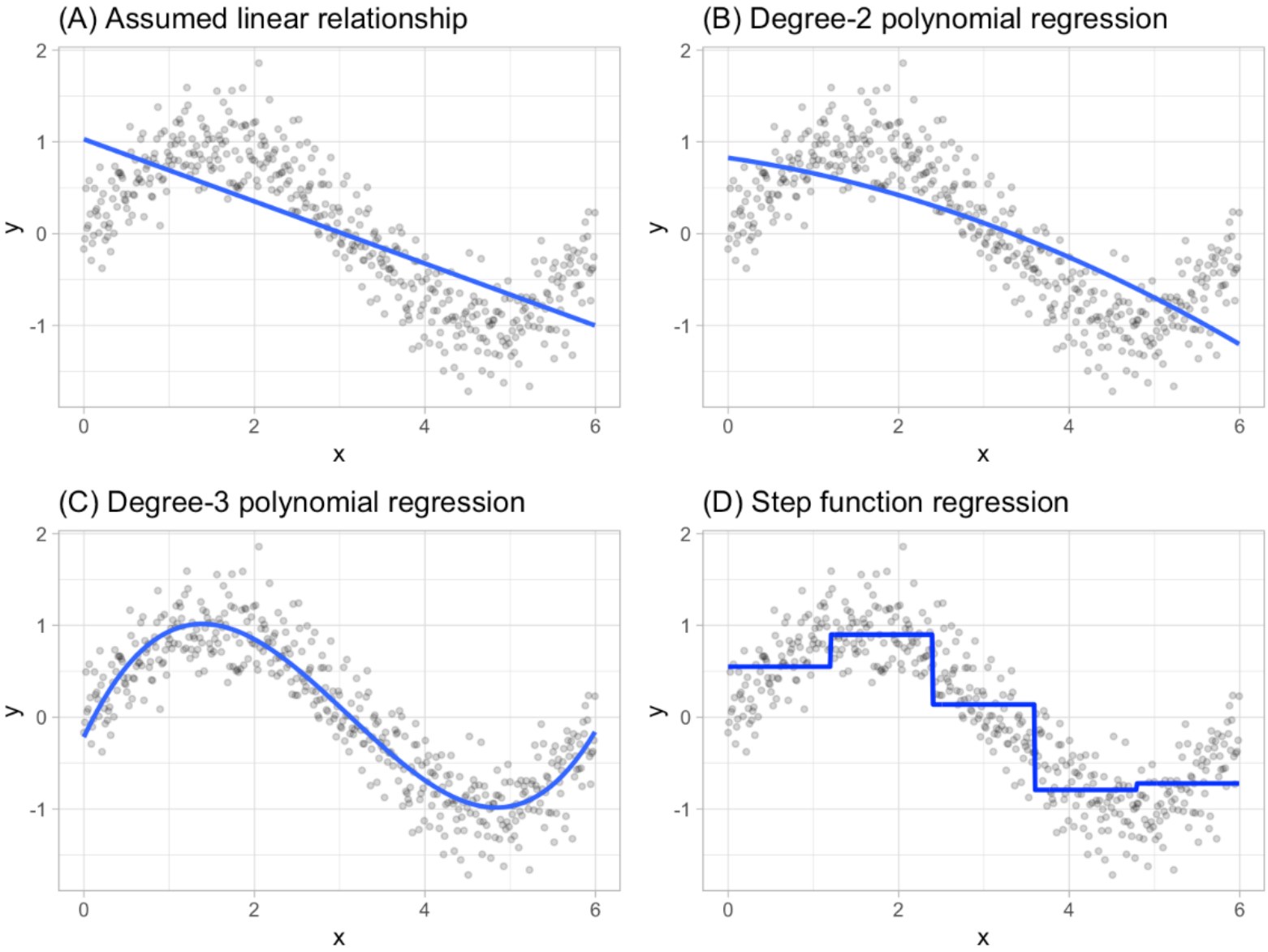
One-Quarter Ahead Conditional Distribution of GDP



Linear Model and Non-Linear Models

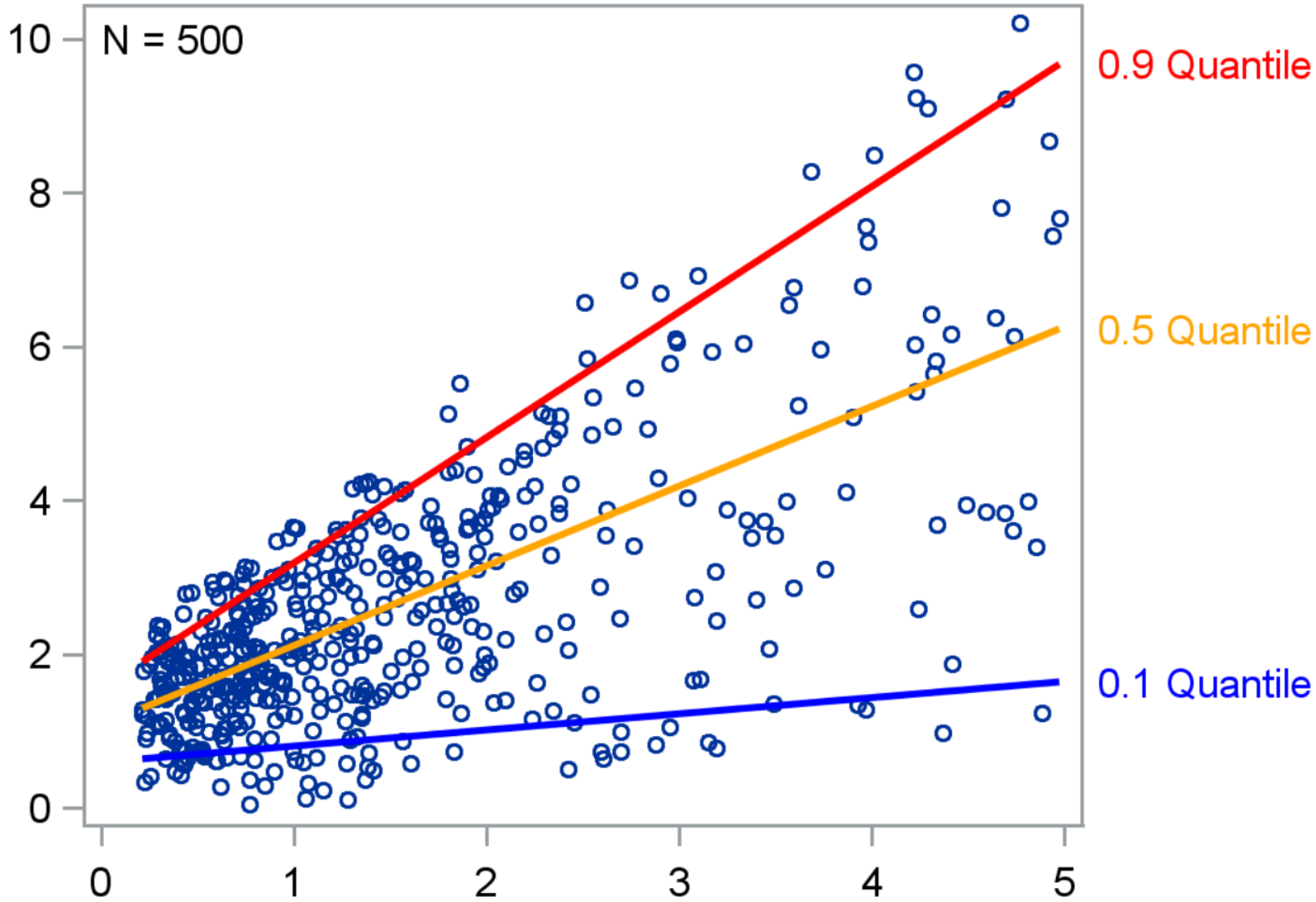
- A linear model represents the dependent variable as a linear combination (a sum) of independent variables:
  $Y = \alpha + \beta_1 \times X_1 + \beta_2 \times X_2 + \ldots \beta_3 X_3 + epsilon$
  - Importantly, in such models, the **marginal effect** of X_i on Y is a constant: $\frac{\partial Y}{\partial X_i} = \beta_i$
  - The representation of the model in a 2D is a straight, hence the name
- To rephrase Tolstoy, *"All linear models are alike, each non-linear model is different in its own way"*
- That being said, there are three main forms of non-linearities common in econometrics:
  1. In the relationship between X and Y: $Y = f(X)$ where f is non-linear, therefore the marginal effect is non constant. For instance, $Y = \alpha + \beta X^2 + \epsilon \leftrightarrow \frac{\partial Y}{\partial X} = 2\beta \times X$ (depends on X)
  2. On the distribution of Y, typically, on the quantiles: $Q(Y, q) = \alpha^q + \beta^q X$
  3. Over the horizon, where the marginal effect varies according to a term structure: $Y_{t+h} = \alpha_h + \beta_h X_t + \epsilon_{t+h}$ (works also for time-varying coefficients)

Linearities and Non-Linearities



Source: *https://bradleyboehmke.github.io/HOML/mars.html*

Non-Linearities in the Distribution of Y



Source: *SAS*

Differentiated Impact

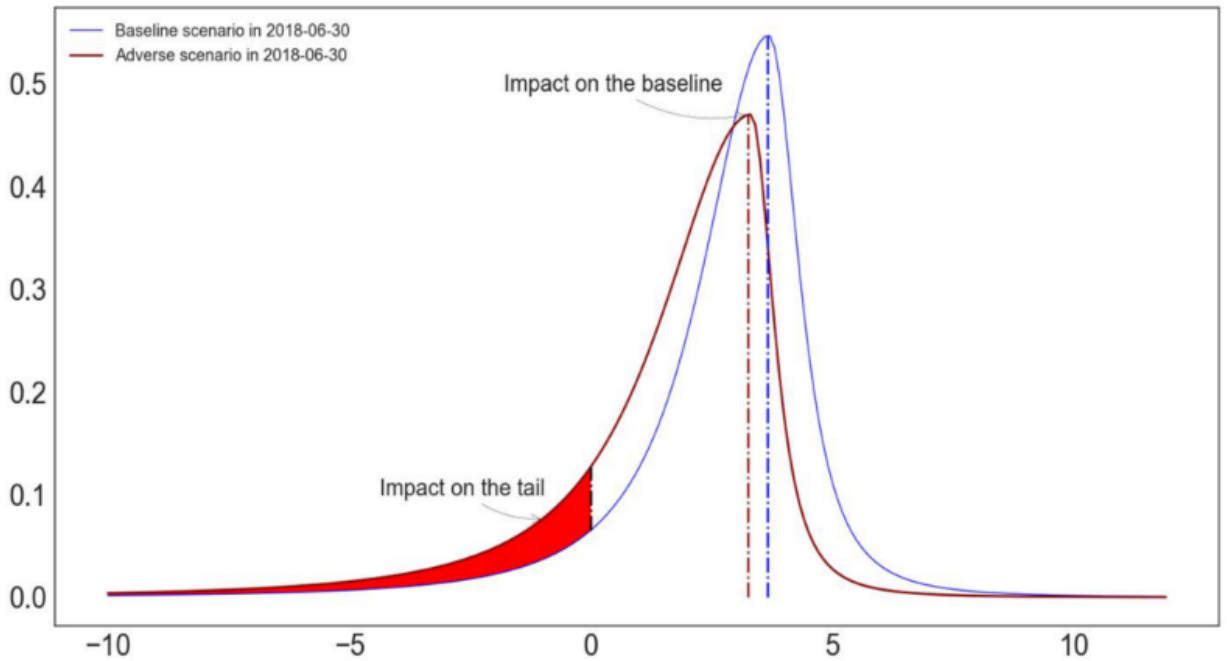
Impact of Tightened Financial Conditions on Growth

Adverse scenario assumes a 1 sd shock on price of risk

Results:

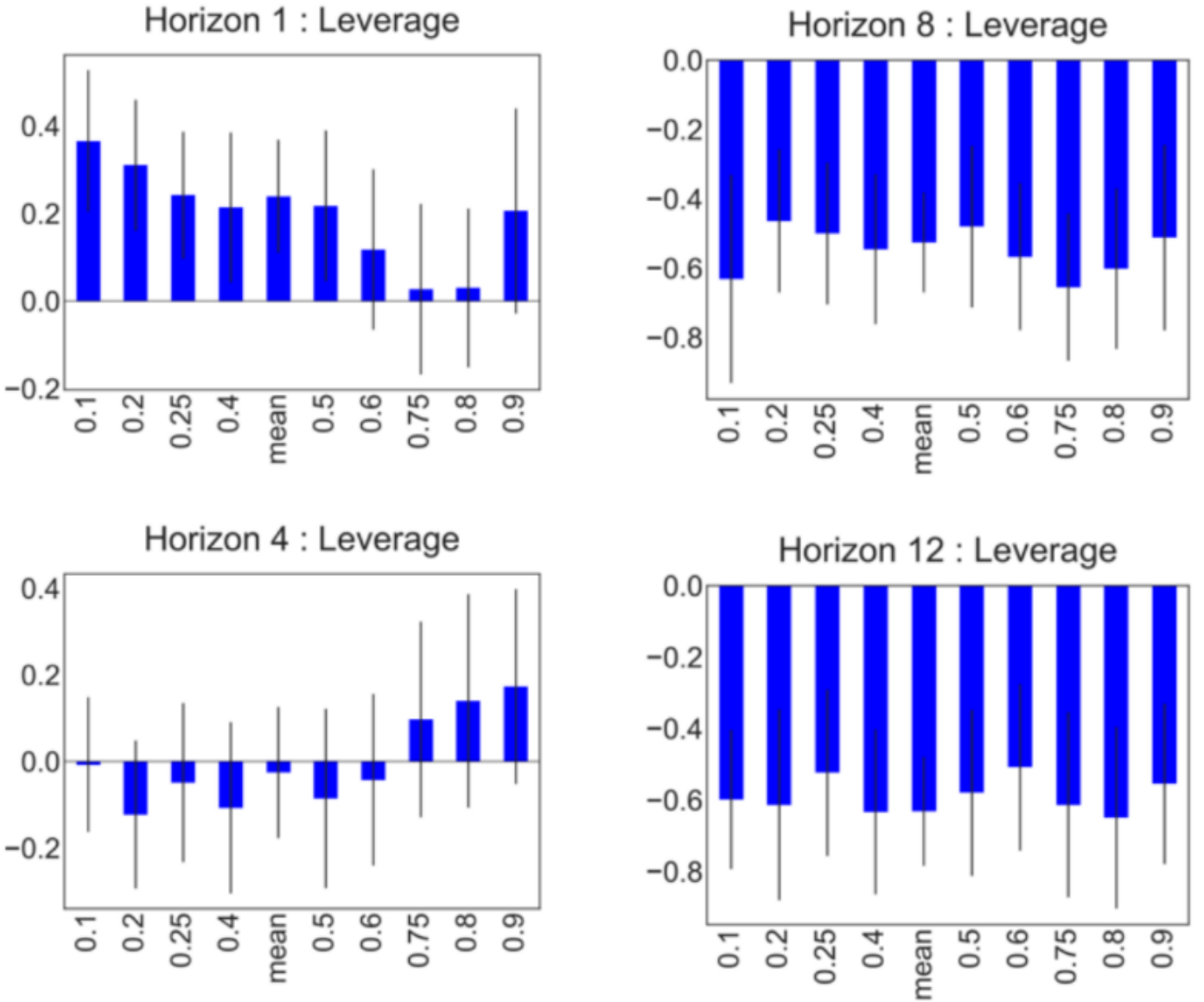
Impact on the average growth: from 3.6% to 3.25%

Impact on the cumulative probability of a recession: from 8% to 18%



Source: *Lafarguette et al. (2019)*

Term Structure



Source: *Lafarguette et al. (2019)*