# Lecture 2: The World as a Data-Generating-Process

Dr. L.S. Sanna Stephan

2023

RUG Groningen

## Learning objectives

You will learn

1. what is the **origin** of data-sets in economics, econometrics and OR and
2. what are probability distributions and why we need to know them.

LO1: The Origin of Our Data

- The Population and the Random Sample
  *(Further reading: DABEP Ch. 3.2 )*
- Probabilities, Moments and Distributions
- Sample Representativeness
  *(Further reading: DABEP Ch. 1.7, 1.8)*

LO2: Theoretical Distributions
*(Further reading: DABEP Ch. 3.9, Ch. 3.U1)*

- Discrete Distributions
- Continuous Distributions
- The Importance of Probability Distributions in Economics/Econometrics/OR

# LO1: The Origin of Our Data

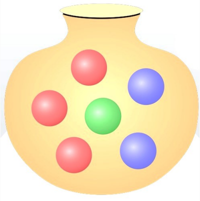
# LO1: The Origin of Our Data

## The Population and the Random Sample

*(Further reading: DABEP Ch. 3.2 )*

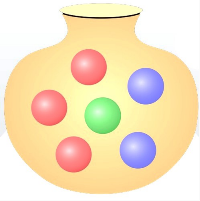
# The urn problem

As researchers, we assume that there is a truth that we want to learn: half of the balls in the urn are red

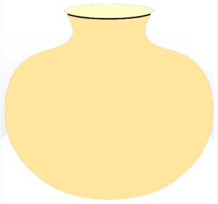


# The urn problem

As researchers, we assume that there is a truth that we want to learn: half of the balls in the urn are red

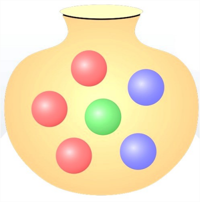


... but we cannot observe the truth!

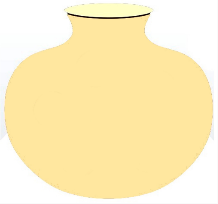


# The urn problem

As researchers, we assume that there is a truth that we want to learn: half of the balls in the urn are red



... but we cannot observe the truth!



**What we do to learn about the truth is to sample from the urn**

# Population and sample

- All the balls in the urn are the **population**.
- The true **data generating process** has created them.
- The balls we have drawn from the urn are our **sample**.
- We believe that there are infinitesimally many balls in the urn (the population).
- $\Rightarrow$ we can draw a sample as large as we want.

# Population and sample - example 1

- Example: household survey in the Netherlands.
- But the population of the Netherlands is finite ???

# Population and sample - example 1

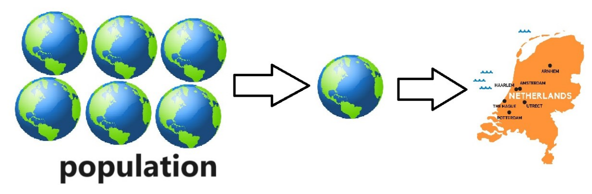
- Example: household survey in the Netherlands.
- But the population of the Netherlands is finite ???
- For the econometrician, “the population” is the population of potential residents of the Netherlands.

# Population and sample - example 1

- Example: household survey in the Netherlands.
- But the population of the Netherlands is finite ???
- For the econometrician, “the population” is the population of potential residents of the Netherlands.
- Nature has drawn a finite amount of inhabitants ...
- ... we sample a finite amount of those.

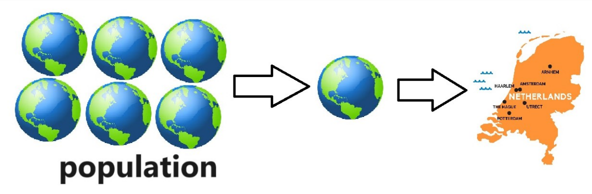
# Population and sample - example 1

Nature has drawn a finite amount of inhabitants ...

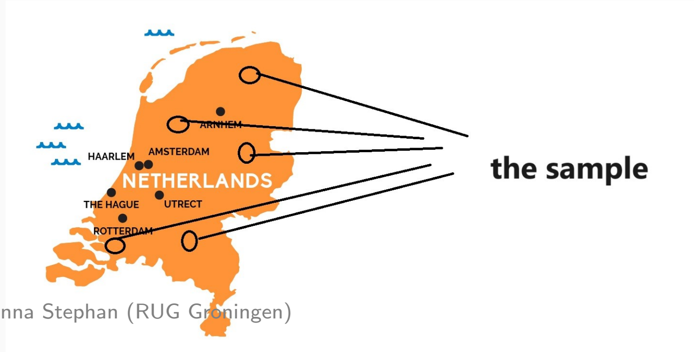


# Population and sample - example 1

Nature has drawn a finite amount of inhabitants ...




... we sample a finite amount of those.



# Population and sample - example 2

- A machine has produced a certain number of products.
- We want to check for faulty items.

## Population and sample - example 2

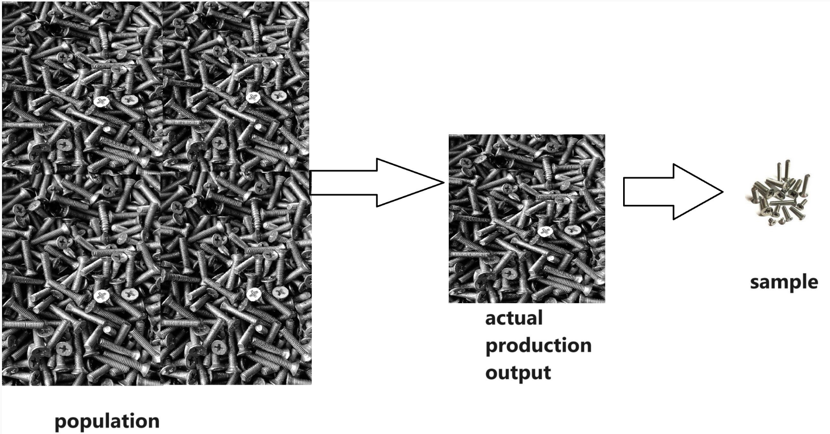
- A machine has produced a certain number of products.
- We want to check for faulty items.
- Nature has chosen the products that have actually been produced ...

# Population and sample - example 2

- A machine has produced a certain number of products.
- We want to check for faulty items.
- Nature has chosen the products that have actually been produced ...
- ... we sample some out of these products.

# Population and sample - example 2

Nature has chosen the products that have actually been produced and we sample some out of these products.



# Randomness

- We do not observe the population (the urn), but only the sample (the draws).
- Sampling is a **random experiment**
- What we observe (the actual colors of the balls we draw) are **realizations** of a random variable.
- The colors of the balls in the urn represent the different realizations that the random variable can take.
- If all balls in the urn have the same color, the process **deterministic**.

Merrian Webster Dictionary:

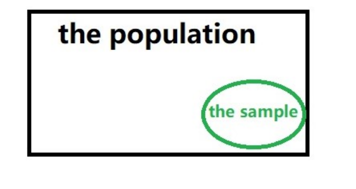
*"A random variable is a variable that is itself a function of the result of a statistical experiment in which each outcome has a definite probability of occurrence"*

| **theoretical concept** | **urn model** | **example application** |
|---|---|---|
| random variable ($X$) | balls in the urn | potential population of Dutch |
| possible realizations | colors | heights |
| observed realization ($x$) | color of a randomly drawn ball | height of a randomly selected current resident |

# LO1: The Origin of Our Data

## Probabilities, Moments and Distributions

# Probabilities and frequencies




L1: relative frequency: $rf_j = \frac{\sum_{i=1}^{N} \mathbf{I}(x_i = j)}{N}$

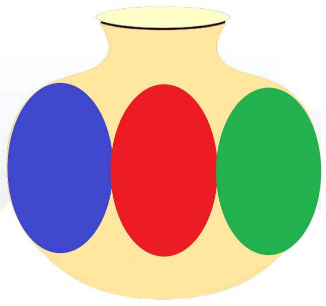
fraction of category $j$ in the **sample**

L2: probability (of occurrance): $P(X_i = j)$

fraction of category $j$ in the **population**

$\rightarrow$ probability that randomly sampled individual falls in category $j$

# Frequencies and probabilities - example



- There are infinitesimally many balls in the urn.
- Here: one third is read.
- $P_{red} = P(X_i = red) = \frac{1}{3}$
- If we draw 6 balls, we still may, for example, draw 3 red and 3 blue balls
- $rf_{red} = \frac{1}{2}$

# Moments

- **Frequency** (absolute or relative): how often does the variable $X$ take value $j$ in our **sample**?
- **Probability**: how likely is it that the random variable $X$ takes value $j$?
- **Moments** are the population equivalent to the measures of centrality and dispersion from L1

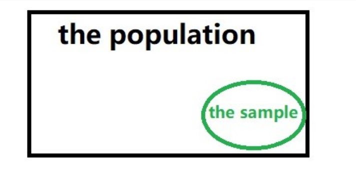
sample mean $\longrightarrow$ (population) mean (AKA expected value of the random variable $E(X)$)
sample variance $\longrightarrow$ (population) variance $\sigma^2(X)$

# LO1: The Origin of Our Data

## Sample Representativeness

*(Further reading: DABEP Ch. 1.7, 1.8)*

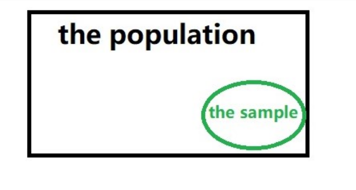
# Recap




- Characteristic (e.g. age), randomly distributed in the population.

  $\Rightarrow X$ (age) is a random variable
- $X \sim P_x$ (meaning: $X$ follows a distribution defined by the probabilities)

# Recap

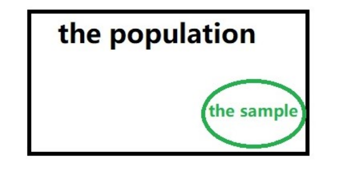



- Characteristic (e.g. age), randomly distributed in the population.

  $\Rightarrow X$ (age) is a random variable
- $X \sim P_x$ (meaning: $X$ follows a distribution defined by the probabilities)

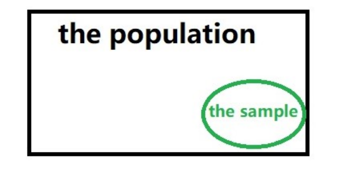
  $P(X = j)$ probability that a randomly selected individual is j years old.

# Recap



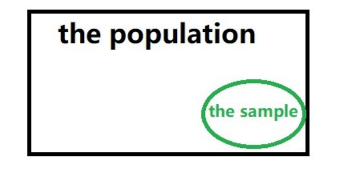

- Characteristic (e.g. age) randomly distributed in population
  $\Rightarrow X$ (age) is a random variable
- $X \sim P_x$ (meaning: $X$ follows a distribution defined by the probabilities)

# Recap




- Characteristic (e.g. age) randomly distributed in population
  $\Rightarrow X$ (age) is a random variable
- $X \sim P_x$ (meaning: $X$ follows a distribution defined by the probabilities)
- Sample $N$ individuals $\Rightarrow$ obtain $N$ realizations $x_1, x_2, \ldots, x_N$

## Recap




- Characteristic (e.g. age) randomly distributed in population $\Rightarrow X$ (age) is a random variable
- $X \sim P_x$ (meaning: $X$ follows a distribution defined by the probabilities)
- Sample $N$ individuals $\Rightarrow$ obtain $N$ realizations $x_1, x_2, \ldots, x_N$ e.g. $10, 20, 99, 76, \ldots, 5$

# Sample representativeness

- Relative frequency ($rf_j$): fraction of category $j$ in the sample
- Probability ($P_j$): fraction of category $j$ in the population
- The sample is **representative**: for all $j$, if $rf_j \neq P_j$, this is purely due to randomness (of the sampling)

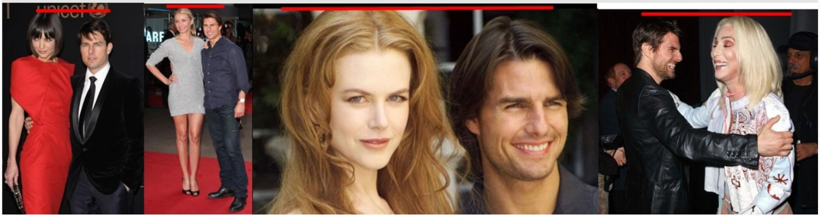
☠ The sample is not representative: **sample selection**

# Sample selection (source:https://datacolada.org)

**Women are taller than men!**

Data evidence:

Study 1: We found that (male) actor Tom Cruise is reliably shorter than his (female) partners.



**Women are taller than men!**

Data evidence:

Study 2: We found that (female) elementary school teachers were much taller than their (mostly male) students.

**Women are taller than men!**

Data evidence:

Study 3: We found that female basketball players are reliably taller than male referees.

# Sample selection (source:https://datacolada.org)

**Women are taller than men!**

Data evidence:

Sampled individuals were not randomly selected!

**Argue why there may be sample selection!**

Online restaurant ratings

# Sample selection

**Argue why there may be sample selection!**

Online restaurant ratings

Those who are very satisfied or very unsatisfied will answer.

# Sample selection

**Argue why there may be sample selection!**

Online restaurant ratings

Those who are very satisfied or very unsatisfied will answer.

Training effectiveness (comparing outcomes for participants to non-participants)

# Sample selection

**Argue why there may be sample selection!**

Online restaurant ratings

Those who are very satisfied or very unsatisfied will answer.

Training effectiveness (comparing outcomes for participants to non-participants)

Those who sign up may be particularly motivated.

The trainer/boss might choose participants that they believe will most benefit.

# Sample selection

**Argue why there may be sample selection!**

Comparing fertility rates among female top managers and female kindergarden teachers to infer which workplaces are family-friendly

# Sample selection

**Argue why there may be sample selection!**

Comparing fertility rates among female top managers and female kindergarden teachers to infer which workplaces are family-friendly

Top managers probably not that family oriented, more focused on work and desire less children (if so, low fertility is by choice)

# Sample selection?

Idea: Monetary compensation for survey participants.

# Sample selection?

Idea: Monetary compensation for survey participants.

☠ Risk: sampled individuals poorer than the average.

## Sample selection?

Idea: Monetary compensation for survey participants.

☠ Risk: sampled individuals poorer than the average.

✓ Better: pick compensation that is attractive to everybody (e.g. lottery to win an exclusive good, donation) or ethical motivation.

## Sample selection?

"Does completing the Bsc thesis in econometrics lead to stress and time pressure?"

Random sample from students that handed in thesis.

# Sample selection?

"Does completing the Bsc thesis in econometrics lead to stress and time pressure?"

Random sample from students that handed in thesis.

☠ Sample excludes students who gave up or were granted an exception.

# Sample selection?

"Does completing the Bsc thesis in econometrics lead to stress and time pressure?"

Random sample from students that handed in thesis.

☠ Sample excludes students who gave up or were granted an exception.

✓ Solution: Random sample from initially registered students.

# Sample selection?

Interviews on people's life style choices (smoking, exercise and diet choices).

# Sample selection?

Interviews on people's life style choices (smoking, exercise and diet choices).

☠ Embarrassing questions! Most likely, individuals that make choices perceived as "poor" will not participate.

## Sample selection?

Interviews on people's life style choices (smoking, exercise and diet choices).

☠ Embarrassing questions! Most likely, individuals that make choices perceived as "poor" will not participate.

✓ Solution: anonymous online survey.

# Sample selection

No test for sample selection! Observe sampling process:

- Are time and location chosen such that certain sub-groups of the population are more likely to be present?
- Is participation voluntary and do certain groups have stronger incentives to participate?
- Cross-check with other data/studies.

## LO2: Theoretical Distributions

*(Further reading: DABEP Ch. 3.9, Ch. 3.U1)*

# Distributions

Merrian Webster Dictionary:
*"A random variable is a variable that is itself a function of the result of a statistical experiment in which each outcome has a definite probability of occurrence"*

- $P_j = P(X = j) \forall j$
- All probabilities together: **probability distribution**
- $X \sim P_x$
- Probability distribution can be discrete (finite number of possibilities) or continuous

# Distributions

- Some variables are known to follow specific distributions.
- Useful for developing and estimating models!
- Distributions can take parameters
- e.g. mean ($\mu$) and variance/standard deviation ($\sigma^2/\sigma$)

# LO2: Theoretical Distributions
*(Further reading: DABEP Ch. 3.9, Ch. 3.U1)*

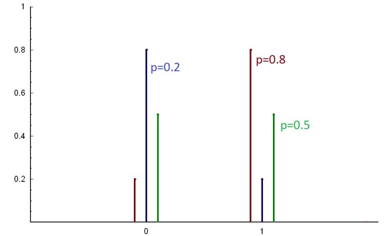
## Discrete Distributions

**Bernoulli distribution**

For binary random variables (two possible outcomes)

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$




Example: coin toss (head, tails), student (EU, non EU)

**Generalized Bernoulli distribution**

For categorical variables (K possible categories)

$$P(X = 1) = p_1$$

$$P(X = 2) = p_2$$

$$\dots P(X = K) = 1 - p_1 - p_2 - \dots - p_{K-1}$$

Example: gender (male, female, non-binary), nationality of EU students ($K = 27$)

**Binomial distribution**

Number of realizations that fall into a category if we conduct $n$ experiments, each follows a Bernoulli distribution with parameter p

Tool: **binomial coefficient**

$$\binom{n}{k}$$

"$k$ out of $n$, disregarding the order and without replacement" *

* *number of ways, disregarding order, that k objects can be chosen from among n objects when objects already selected are not replaced*

# binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)....(2)(1)}{[k(k-1)...(2)(1)][(n-k)(n-k-1)...(2)(1)]}$$

Example: 4 employees, we want to pick 2 on a committee, how many possibilities are there?

# binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)....(2)(1)}{[k(k-1)...(2)(1)][(n-k)(n-k-1)...(2)(1)]}$$

Example: 4 employees, we want to pick 2 on a committee, how many possibilities are there?

- employees A,B,C,D
- first person: 4 possibilities
- second person: 3 possibilities (for each of the previous 4 possibilities)

  $\rightarrow 4 \times 3$ committees

# binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)....(2)(1)}{[k(k-1)...(2)(1)][(n-k)(n-k-1)...(2)(1)]}$$

Example: 4 employees, we want to pick 2 on a committee, how many possibilities are there?

- employees A,B,C,D
- first person: 4 possibilities
- second person: 3 possibilities (for each of the previous 4 possibilities)
  $\longrightarrow 4 \times 3$ committees

wait: we disregard the order! AB same as BA! $\Rightarrow$ less then 12 **different** committees

namely: $\frac{4\times 3}{2\times 2} = 6$ (try it :))

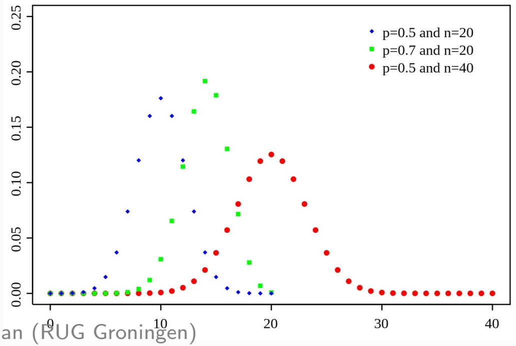
## Binomial distribution

let $X_1, \ldots, X_n \sim Bernoulli(p)$ , let $Y = \sum_{i=1}^{n} \mathbf{I}(x_i = 1)$

then $y$ can take any value from 0 to $n$ and

$$P(y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

probability that $k$ out of the $n$ random variables fall into category 1

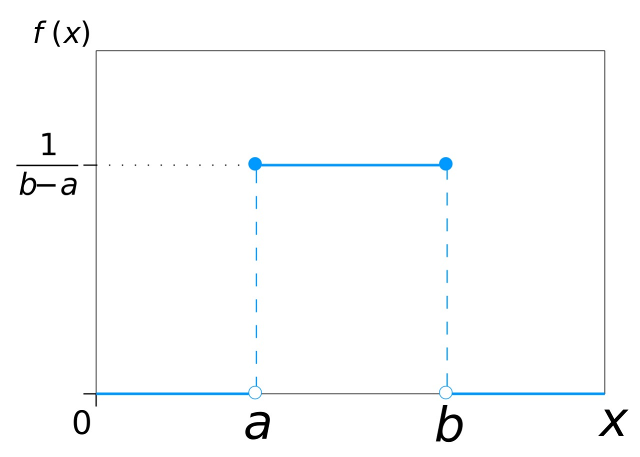


# LO2: Theoretical Distributions

*(Further reading: DABEP Ch. 3.9, Ch. 3.U1)*

## Continuous Distributions

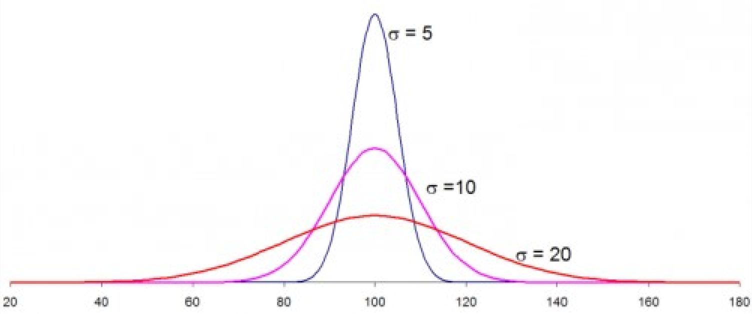
**Uniform distribution** any value in the interval $[a, b]$ is equally likely



**Normal distribution (Gaussian)**

If $\mu = 0, \sigma^2 = 1$ standard normal distribution

Nice properties and often occurs in the world: IQ, height, errors in measurement or production, (some) stock market return



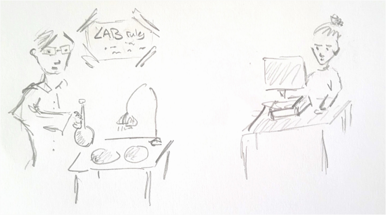
# LO2: Theoretical Distributions

*(Further reading: DABEP Ch. 3.9, Ch. 3.U1)*

## The Importance of Probability Distributions in Economics/Econometrics/OR

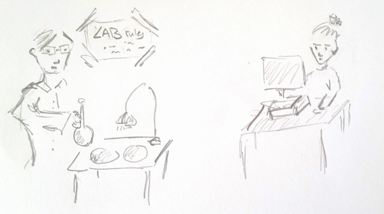
# The economist's lab is a little different...

Why do these two researchers have different data?



# The economist's lab is a little different...

Why do these two researchers have different data?




# The importance of probability distributions in econometrics/OR

- Data at hand is a **random** sample.
- Consequence: there is noise (random stuff we **don't control**)
- Too many factors at play in economic interaction, model cannot include all!
- Economists make abstract models that focus on some factor(s).
- The rest is "random stuff we **don't model**".
- Know the distribution of the "random stuff" $\Rightarrow$ derive mathematical properties and prove them!

# Model and noise

$$\mathbf{D_i = f(Price_i, X_i) + \varepsilon_i}$$

$D_i$ demand for good $i$

$Price_i$ is price of good $i$

$X_i$ are product features

# Model and noise

$$\mathbf{D_i = f(Price_i, X_i) + \varepsilon_i}$$

$D_i$ demand for good $i$

$Price_i$ is price of good $i$

$X_i$ are product features

what is $\varepsilon_i$ ???

# Model and noise

$$\mathbf{D_i = f(Price_i, X_i) + \varepsilon_i}$$

$D_i$ demand for good $i$

$Price_i$ is price of good $i$

$X_i$ are product features

what is $\varepsilon_i$ ???

$\Rightarrow$ consumer taste, trend shifts,...

# Model and noise

$$\mathbf{Q_i = f(K_i, L_i) + \varepsilon_i}$$

$Q_i$ output of factory $i$
$L_i$ labour used in factory $i$
$K_i$ capital used in factory $i$

# Model and noise

$$\mathbf{Q_i = f(K_i, L_i) + \varepsilon_i}$$

$Q_i$ output of factory $i$

$L_i$ labour used in factory $i$

$K_i$ capital used in factory $i$

what is $\varepsilon_i$ ???

## Model and noise

$$\mathbf{Q_i = f(K_i, L_i) + \varepsilon_i}$$

$Q_i$ output of factory $i$

$L_i$ labour used in factory $i$

$K_i$ capital used in factory $i$

what is $\varepsilon_i$ ???

$\Rightarrow$ worker ability, weather conditions, pre-material quality ,...

# Model and noise

$$\mathbf{S_i = f(H_i, IQ_i) + \varepsilon_i}$$

$S_i$ score student $i$
$H_i$ hours studied for student $i$
$IQ_i$ IQ score of student $i$

# Model and noise

$$\mathbf{S_i = f(H_i, IQ_i) + \varepsilon_i}$$

$S_i$ score student $i$

$H_i$ hours studied for student $i$

$IQ_i$ IQ score of student $i$

what is $\varepsilon_i$ ???

# Model and noise

$$\mathbf{S_i = f(H_i, IQ_i) + \varepsilon_i}$$

$S_i$ score student $i$

$H_i$ hours studied for student $i$

$IQ_i$ IQ score of student $i$

what is $\varepsilon_i$ ???

$\Rightarrow$ motivation, attention, capability to concentrate on test day, luck ,...

## Recap

Today we learned

- For us, the world is one big (or the accumulation of many little) data-generating-process(es) (**DGP**) from which we draw **samples** to learn about the truth.
- Because we sample, our data contains randomness.
- Because economic interaction does not take place in labs, there are often factors that we do not model.
- this is another source of randomness
- We use know probability distributions to derive properties of our model randomness which we (will) use.