

Statistics

Econometrics

Probability

Machine Learning

Big Data

Data Science

Computational SC

Statistics

Econometrics

Probability

Machine Learning

Big Data

Data Science

Computational SC

Probability

Model —————→ **Data**

Statistics

Data —————→ **Model**

Probability

Model : When we flip a coin the process that governs its behavior is one in which:

- Call the coin flipping “variable X ”
- X has 2 possible outcomes: {Head,Tail}
- $P(X = \text{Head}) : p$
- $P(X = \text{Tail}) : q = 1-p$

E.g. If $p=0.9$, then $q=0.1$ as well.

We call this a RANDOM variable: we know the “parameters” (p), but not the outcome of X

- If we flip this coin, most likely we get a Head, but not necessarily

Formally, we would say that :

“ X comes from a Bernoulli distribution with parameter $p=0.9$ ” :

$X \sim \text{Bernoulli}(0.9)$

Probability

We can also derive other theoretical results

What is the probability of getting 2 Heads if I flip a “fair” coin twice?

Model: $p=q=0.5$

Four possible outcomes:

H,H

H,T

T,H

T,T

1 out of 4 = 0.25


More formally: $p * p = 0.5 * 0.5 = 0.25$

Even more formally: What is the probability of getting n Heads if I flip the coin n times?

p^n

Probability

Model  **Data**

Model : $p = 0.5$ 

H
T
H
T
T
H
H
T
H
T

5 Head

T
H
H
T
H
T
H
T
H
T

5 Head

H
T
T
H
H
T
T
H
H
H

6 Head

Statistics

Data  Model

H
T
H
T
T
H
H
T
H
T

What is the underlying model that generates data like this?

Or, what is p ?

We don't know, but we can estimate p from the data.

We need an ESTIMATOR, let's call it θ

Which will produce an ESTIMATE of p , let's call it \hat{p} .

Statistics

Data  Model

H
T
H
T
T
H
H
T
H
T

What is an estimator?

An estimator is $f(\text{Data})$ that aim to approximate the true value of p , that is, \hat{p}

Ideas for $\theta(x_i, \dots, x_n)$?

\hat{p} = Number of occurrence of H / Total number of occurrences

$$\hat{p} = 5/10 = 0.5$$

Statistics

Data  Model

H
T
H
T
T
H
H
H
T
H
T

\hat{p} = Number of occurrence of H / Total number of occurrences

How good is our estimate? —> **Probability Theory**

Statistics

Data  Model

Let's say the data DOES come from a world in which $p=0.5$ Here are some possible data draws :

Our data

H
T
H
T
T
H
H
T
H
T

$$\hat{p} = 5/10 = 0.5$$

Other data

T
H
H
T
H
T
H
T
H
T

$$\hat{p} = 5/10 = 0.5$$

Other data

H
T
T
H
H
T
T
H
H
H

$$\hat{p} = 6/10 = 0.6$$

Statistics

+

Probability

Data —————→ **Model**

Model —————→ **Data**

**Produce ESTIMATES of
(unknown) true parameters
governing the processes we
want to study**

**Quantify UNCERTAINTY
around these estimates**

Data is usually structured in matrix form: [rows,columns] → [observations, variables]

We have a sample of :

- 10 people
- 4 men and 6 women
- each of them tossed the same coin once

| | “Coin toss” | “Gender” |
|------|-------------|----------|
| (1) | H | Male |
| (2) | T | Female |
| (3) | H | Male |
| (4) | T | Female |
| (5) | T | Female |
| (6) | H | Male |
| (7) | H | Female |
| (8) | T | Female |
| (9) | H | Female |
| (10) | T | Male |

Data[5,1] : T

Data[7,2] : Female

Do not mistake “variables” in a dataset with RANDOM VARIABLES

| | “Coin toss” | “Gender” |
|------|-------------|----------|
| (1) | H | Male |
| (2) | T | Female |
| (3) | H | Male |
| (4) | T | Female |
| (5) | T | Female |
| (6) | H | Male |
| (7) | H | Female |
| (8) | T | Female |
| (9) | H | Female |
| (10) | T | Male |

- “Coin toss” is a **VARIABLE** (column) in our data
- Each of the 10 entries in “Coin toss” are the manifestation of a **RANDOM VARIABLE** with a Bernoulli distribution and parameter p