Mind the Bias: How Selection Effects Shape Our Understanding of the Universe

Leonardo Iampieri

PHD Seminar - 05/02/2025

Istituto Nazionale di Fisica Nucleare

Introduction

- **Bias:** A systematic error that skews measurements away from the true value
  - A scale that unfailingly shows you a few pounds heavier or lighter than your actual weight.
  - A video camera that consistently adds a few inches to your waistline.

- **Selection Bias:** A systematic error that occurs when the chosen sample does not accurately represent the entire population.
  - Since this error is systematic, it can often be measured and corrected for by accounting for the sampling differences.

Diagoras, the non-believer

- Cicero's account of Diagoras of Melos offers one of the earliest recorded observations of **selection bias**.

- Cicero, De Natura Deorum 3.37:

Atque hoc etiam, Diagora, qui dictus est atheus, solebat in contione dicere, cum ei, qui vota exsolverant, pictam tabulam ostenderent, in qua e naufragio servati grates dis agerent: 'Ubi sunt, inquit, illi qui naufragio perierunt?'	And Diagoras, who was called an atheist, used to say this in public assemblies: when people showed him a painted tablet depicting those who had been saved from shipwreck, giving thanks to the gods, he would ask: '**But where are those who perished in the shipwreck?**

Selection bias, an historic challenge

- This bias has been rediscovered here and there throughout history across disciplines, often to be rapidly forgotten.

- Bacon, Novum Organum, Aphorism XLVI:

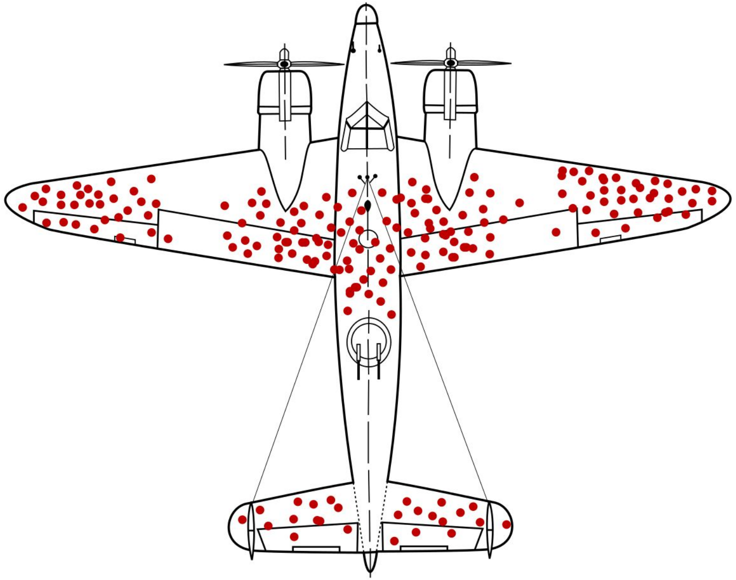
> And such is the way of all superstition, whether in astrology, dreams, omens, divine judgments, or the like; wherein men, having a delight in such vanities, mark the events where they are fulfilled, but where they fail, though this happen much oftener, neglect and pass them by.

Selection Biases in WWII

- **Military Analysis:** During WWII, American military analysts examined bullet holes on returning bomber airplanes.

- **Selection bias:** Only planes that returned were analyzed, ignoring those that were shot down.

- **Conclusion:** The areas with fewer or no bullet holes on survivors were the most critical.



How to become a Millionaire in Ten-Steps

- Numerous studies of millionaires aimed at figuring out the skills required for success follow this methodology:
  - They take a population of millionaires and look at what attributes they have in common (courage, risk taking, optimism, and so on).
  - They then infer that these traits help you become successful.

- **Biased Methodology:** They neglect to analyze whether these same traits are equally common among non-millionaires.

- **False Casual Links:** Without analyzing both groups, they wrongly inferred a connection between these attributes and success.

I Exist, Therefore I bias

- **Anthropic bias:** Our own existence produces a selection bias.
  - The condition that we are in existence imposes restrictions on the process that led us here.

- N.N. Taleb, The Black Swan, The Cosmetic Because:

> Whenever our survival is in play, the very notion of because is severely weakened. The condition of survival drowns all possible explanations […] Why didn't the bubonic plague kill more people? People will supply quantities of cosmetic explanations involving theories about the intensity of the plague and "scientific models" of epidemics. […] had the bubonic plague killed more people, the observers (us) would not be here to observe. So it may not necessarily be the property of diseases to spare us humans.

Probability: Recap

- **Probability P(A):** natural number that quantifies our **degree of belief** in the occurrence or truth of event A.

- Probability is inherently **subjective**. Probability depends on the status of information of the subject who evaluates it.

$$P(A) \to P(A|I_s(t))$$

  - where $I_s(t)$ is the information available to subject s at time t.

- Subjective does not mean arbitrary!
  - In order for our belief system to be coherent Probability must follow **rules**!

Inference and Bayes Theorem

➔ **Inference:** process of drawing conclusions about **causes** from **observed** effects.

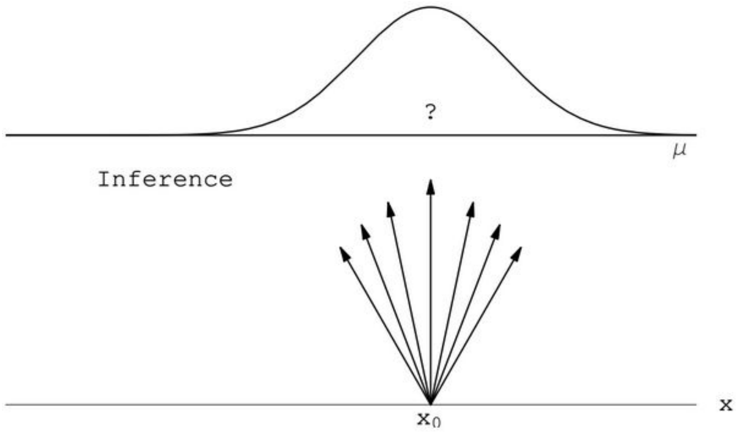
➔ **Bayes Theorem:** allows us to update the probability of cause A given observation B.

Likelihood

Posterior **Prior**

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A)$$

Evidence



Quod Videmus Testimur

- Given a set of independent observations $\{\vec{x}_i\}$ drawn from a model parameterized by $\vec{\lambda}$, the the probability of obtaining this specific dataset (the likelihood) is:

$$p(\{\vec{x}_i\}|\vec{\lambda}) = \prod_{i=1}^{N_{\rm obs}} \frac{p_{\rm pop}(\vec{x}_i|\vec{\lambda})}{\int \mathrm{d}\vec{x}\, p_{\rm pop}(\vec{x}|\vec{\lambda})}$$

- When selection bias is present, some events are more likely to be observed than others. This effect is quantified by the **detection probability** $p_{\rm det}(\vec{x})$.
  - Beware! $p_{\rm det}(\vec{x})$ is a probability (i.e. $p_{\rm det}(\vec{x}) \in [0, 1]$).

- With selection effects included, the likelihood becomes:

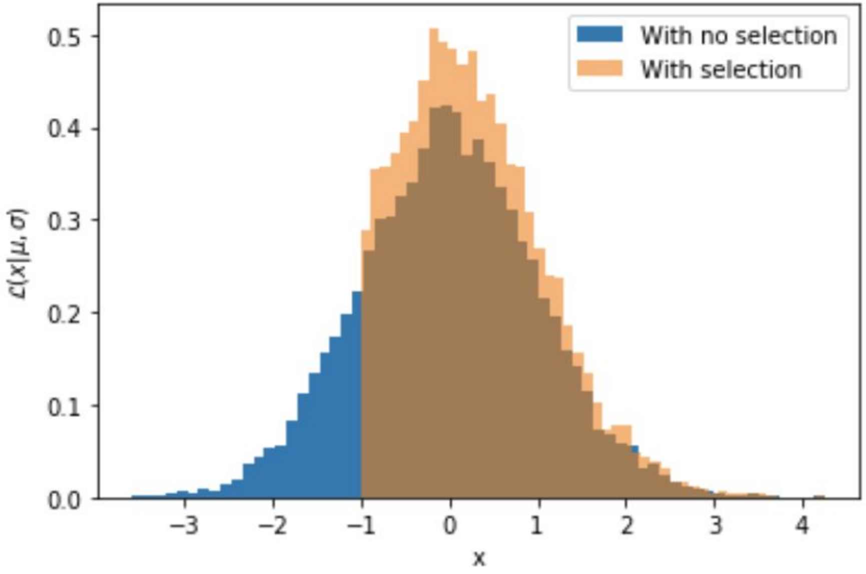
$$p(\{\vec{x}_i\}|\vec{\lambda}) = \prod_{i=1}^{N_{\rm obs}} \frac{p_{\rm pop}(\vec{x}_i|\vec{\lambda})\, p_{\rm det}(\vec{x}_i)}{\int \mathrm{d}\vec{x}\, p_{\rm pop}(\vec{x}|\vec{\lambda})\, p_{\rm det}(\vec{x})}$$

A Simple Example - I

- → Random process generates numbers from a **normal** distribution with **mean** 0 and **variance** 1.

- → **Selection bias:** Only samples with values $x > -1$ are observed.

- → How can we estimate the mean from these samples?



A Simple Example - II

➔ We must normalize the likelihood by incorporating a **detection probability.**

➔ **Detection Probability:** $p_{\mathrm{det}}(x) = \begin{cases} 0, & \text{if } x \leq -1, \\ 1, & \text{if } x > -1. \end{cases}$

➔ Likelihood before including the selection effect:

$$\mathcal{L}(x|\mu) = \frac{\exp[-(x-\mu)^2/(2\sigma^2)]}{\int_{-\infty}^{\infty} \exp[-(x-\mu)^2/(2\sigma^2)]dx} = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(x-\mu)^2/(2\sigma^2)]$$

➔ Likelihood after including the selection effect:

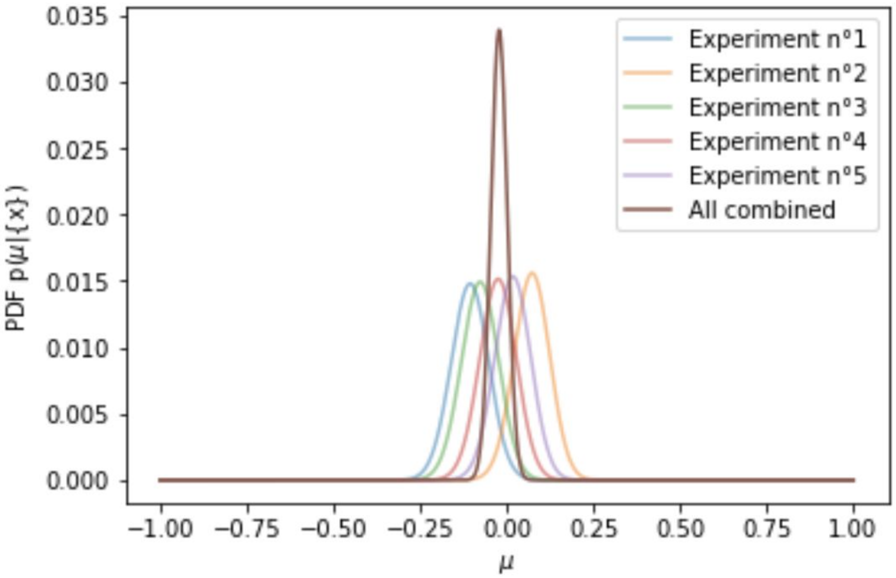
$$\mathcal{L}(x|\mu) = \frac{\exp[-(x-\mu)^2/(2\sigma^2)]}{\int_{x_{thr}}^{\infty} \exp[-(x-\mu)^2/(2\sigma^2)]dx} = \frac{\exp[-(x-\mu)^2/(2\sigma^2)]}{I(\mu, x_{thr})}$$

A Simple Example - III

- ➔ **Posterior Distribution:**

$$p(\mu|\{x\}) \propto \mathcal{L}(\{x\}|\mu)p(\mu) = p(\mu)\prod_i \mathcal{L}(x_i|\mu)$$

- ➔ By including the detection probability, the **posterior distribution** will converge to the correct mean value.



A Peek into GW Cosmology

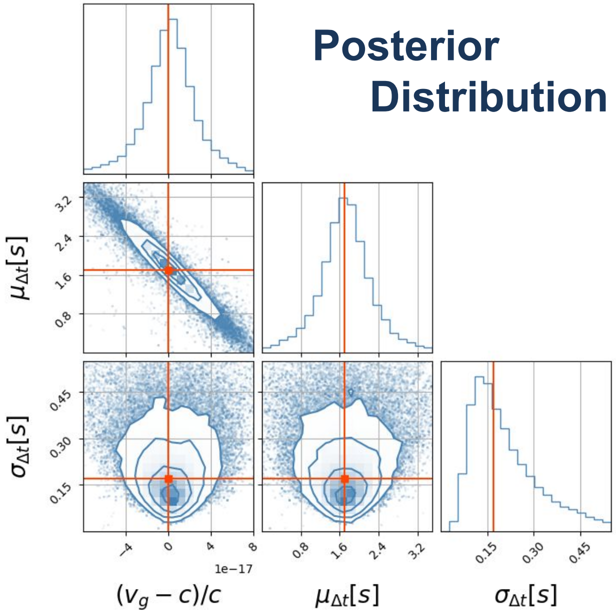
➔ Inference of astrophysical and cosmological parameters from joint observations of **Gravitational Waves** (GWs) and **short Gamma-ray burst** (sGRBs) from **Binary Neutron Star** (BNs) Mergers.

➔ Full expression of Hierarchical Likelihood:

$$\mathcal{L}(\{\vec{d}_i\}|\vec{\lambda}) \propto \prod_{i=1}^{N_{\mathrm{obs}}} \frac{\int \mathcal{L}(\vec{d}_i|D_L, \Delta t_d, \vec{\lambda}) \frac{dV_c}{dz} \frac{\psi(z;\vec{\lambda}) p_{\mathrm{pop}}(\Delta t_s|\vec{\lambda})}{(1+z)^2 \left|\frac{\partial D_L}{\partial z}\right|} \, dD_L d\Delta t_d}{\int p_{\mathrm{det}}(D_L, \Delta t_d, \vec{\lambda}) \frac{dV_c}{dz} \frac{\psi(z;\vec{\lambda}) p_{\mathrm{pop}}(\Delta t_s|\vec{\lambda})}{(1+z)^2 \left|\frac{\partial D_L}{\partial z}\right|} \, dD_L d\Delta t_d}$$

Detection Probability

Posterior Distribution



➔ The **finite sensitivities** of the GW and sGRB detectors lead to a selection bias.

Conclusion

- **Selection Bias:** Incomplete data can skew our inferences and lead us to wrong conclusions.
  - Selection bias affects diverse fields—from scientific research to everyday decision-making.
  - Our very existence introduces biases.

- How to deal with them?
  - Ensure your observations reflect the entire population.
  - Renormalize the likelihood by using a detection probability.