Statistical Inference

sabato 4 ottobre 2025 10:48

Statistical inference (or inferential statistics) is the process by which the characteristics of a *population* are induced from the observation of a part of it (called a *sample*), usually selected through a random experiment.

→ Definition

In the history of statistics, inference has gone through two major periods. The first began at the end of the 19th century and developed decisively in the first half of the 20th century, with the works of R. Fisher, K. Pearson, Jerzy Neyman, Egon Pearson, and Abraham Wald, introducing fundamental ideas concerning likelihood, the power of hypothesis tests, confidence intervals, and more.

The second major period, still ongoing, has been made possible by the increasing computational power of computers, available at ever more affordable prices.

Within statistical inference, two main schools of thought can be distinguished, tied to different conceptions —or interpretations—of the meaning of probability:

- **Classical, or frequentist inference**
- **Bayesian inference**

→ a little bit of context

The first is linked to the historical contributions of R. Fisher and K. Pearson, and represents the majority view. The second is based on the application of Bayes' theorem to statistical inference.

There also exists, in fact, a third approach, which is a criticism of the very concept of inference: **statistical subjectivism**, advocated by the engineer and mathematician Bruno De Finetti. In particular, De Finetti, by rejecting the ontological possibility of repeatable cases, denied the reliability of frequentist statistics.

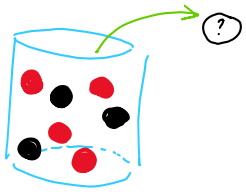
Frequentist vs Bayesian →

Both the frequentist and the Bayesian approaches share, first of all, the axioms of probability as well as the entire statistical–mathematical framework. Bayes' theorem is valid for both approaches, and in both cases one usually deals with parametric statistics. What changes is the meaning attributed to the concept of probability.

Within statistical inference, these differences emerge, on the one hand, in how (and whether) to use information known prior to "seeing" the data and how to quantify such information, and on the other hand, in the different ways of interpreting results.

An example of how the same experiment is seen by the two approaches can be illustrated with a classic textbook problem:

An urn contains balls that are identical except for color. An unknown proportion π of them are black. Drawing a ball 100 times with replacement, it happens, for instance, that 30 times the ball is black.



In both approaches a binomial distribution is assumed:

- **Frequentist approach**
  The typical frequentist approach, based on confidence intervals, would establish for the unknown value of π a 95% confidence interval between 0.21 and 0.39.
  The "95% confidence" does not mean that π is contained in the interval with probability 95% (that would be a typically Bayesian statement). Instead, it means that, given the assumptions and the method used, in 95% of cases the procedure will produce a correct statement—in the sense that the true value will indeed lie within the calculated interval.
  This approach emphasizes that the unknown value π either lies in the interval or it does not, but no probability values are assigned to this fact.

- **Bayesian approach**
  The Bayesian approach, instead, begins by formalizing prior beliefs about what the true value of π might be, assuming a discrete or continuous distribution over possible values of π.
  In the particular case where one wishes to model complete ignorance, one might use a uniform discrete distribution, or—given the relatively large sample size (100 draws)—a continuous uniform distribution between 0 and 1.
  By choosing a prior distribution for the parameter π, one obtains the posterior distribution:

$$f(\pi \mid n = 100, k = 30) = (n+1)\binom{n}{k}\pi^k(1-\pi)^{n-k}$$

  The maximum value (and thus the most probable one) is again given by k/n = 30/100 = 0.3, the same as in the frequentist approach.
  The difference is that, in the Bayesian framework, this is the most probable value *a posteriori*, given the prior beliefs and the experimental results.
  Using the posterior distribution, one can state that the probability that the unknown parameter π lies between 0.216 and 0.393 is 0.95, that is, 95%. Similarly, the probability that π lies between 0.21 and 0.39 is 95.3%.

To summarize this example: in the frequentist approach, one makes statements about how often the method produces correct results. In the Bayesian approach, one directly assigns a probability to an interval containing the parameter. This difference is often ignored in practice, but from a theoretical standpoint it is substantial.

Why we need to infer ?

In practice, it is rarely possible to study theoretical models directly or to observe an entire population of interest. Instead, we are limited to reduced, empirical observations collected from a sample. Statistical inference provides the framework that allows us to use this limited information to estimate the parameters of the population—such as the mean, the variance, or proportions—and to draw conclusions about the underlying structure.

This leads us to the theory of estimation, which focuses on how population parameters can be approximated from sample data.

Estimation process →

The entire goal of estimation theory is to arrive at an estimator, and preferably to implement one that is practically usable. An estimator is a deterministic function that, starting from the observed data, produces an estimate of the parameter.

Def: An m-dimensional statistic of a sample $X_1 \dots X_n \sim F_\theta$ is a function $T_n : \mathbb{R}^n \to \mathbb{R}^m$ of just $X_1 \dots X_n$

↝ in other words, it is a statistic if it is dependent on the sample only and not on the parameters that must be estimated

example: $X_1, X_2 \dots X_n \sim Be(\vartheta)$ iid with ϑ unknown

$\Rightarrow T(X_1 \dots X_n) := \left(\sum_{i=1}^{n} X_i \right) / n$ it's a statistic

while $\tilde{T}(X_1 \dots X_n) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \vartheta)$ it is not!

Statistics can have different characteristics and are therefore classified according to their properties:

- **Sufficient statistic**: contains all the information in the sample about the parameter.
- **Minimal sufficient statistic**: a sufficient statistic that cannot be reduced further without losing information.
- **Complete statistic**: a sufficient statistic such that no non-trivial function of it has expectation zero for all parameter values.
- **I-complete statistic**: a generalization of completeness, often used in more advanced contexts.
- **Unbiased estimator**: an estimator whose expected value equals the true parameter.
- **Efficient estimator**: among unbiased estimators, the one with the smallest variance.
- **UMVUE (Uniformly Minimum Variance Unbiased Estimator)**: an unbiased estimator that has the lowest variance for all possible values of the parameter.
- **Consistent estimator**: converges in probability to the true parameter as the sample size grows.

Def: Given $(f_\vartheta)_{\vartheta \in \Theta}$ and $(X_1 \dots X_n) \sim f_\vartheta$, $T = T(X_1 \dots X_n)$ is a sufficient statistic if the conditional distribution $(X_1 \dots X_n) \mid T = t$ does not depend on ϑ.

↝ $\mathbb{P}((X_1 \dots X_n) \in B \mid T = t) = h(B, t)$ there's no ϑ!

Theorem: factorization criteria of Fisher - Neyman

$(X_1 \dots X_n) \sim f_\vartheta$ $T = T(X_1 \dots X_n)$ is sufficient if and only if

$$f_\vartheta(x_1 \dots x_n) = g_\vartheta(T(x_1 \dots x_n)) \cdot h(x_1 \dots x_n)$$

g_ϑ: depend on ϑ; h: doesn't depend on ϑ

example $X_1 \dots X_n \sim N(\mu, \sigma^2)$ iid

obs. 1 $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ but what if $x = (x_1 \dots x_n)$ is a vector?

$$\Rightarrow f_{\mu,\sigma^2}(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

obs. 2 $\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(\bar{x}-\mu)^2 + 2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) =$

($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$)

$= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x}) =$

$\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} 1$

$= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\left[\sum_{i=1}^{n} x_i - \bar{x}\cdot n\right] =$ (the bracket $= 0$)

$= \sum_{i=1}^{n} x_i - n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i$

$= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$

$\Rightarrow f_{\mu,\sigma^2}(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right]\right\} =$

$= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\cdot n\cdot(\bar{x}-\mu)^2\right\} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} \cdot 1$

$g_{\mu,\sigma^2}(T(x))$ $\qquad$ $h(x)$

where $T(X) = \left(\bar{X}, \sum_{i=1}^{n}(X_i-\bar{X})^2\right)$ is a sufficient statistic □

obs: T sufficient statistic of ϑ, $(X_1 \dots X_n) \sim f_\vartheta$, $T: \mathbb{R}^n \to \mathbb{R}^k$ and $\varphi: \mathbb{R}^k \to \mathbb{R}^k$ bijection $\Rightarrow$ $T' = \varphi(T)$ is still sufficient

Def: A statistic T^* is minimal sufficient for $(X_1 \dots X_n) \sim f_\vartheta$ if:

1. T^* is sufficient
2. $\forall\, T'$ sufficient $\exists\, \varphi$ s.t. $T^* = \varphi(T')$

Def: A sufficient statistic T is complete if $\forall \varphi$ measurable s.t.

$\mathbb{E}_\vartheta(\varphi(T)) = 0 \ \forall \vartheta \in \Theta \Rightarrow \mathbb{P}_\vartheta(\varphi(T) = 0) = 1$

Def: $(X_1 \dots X_n) \sim f_\vartheta$, $\vartheta \in \Theta$, $\tau(\vartheta): \Theta \to \mathbb{R}^k$

$T(X_1 \dots X_n)$ is a correct estimator (or unbiased) of τ if:

$\mathbb{E}_\vartheta(T) = \tau(\vartheta) \quad \forall \vartheta \in \Theta$

example 1: $X_1 \dots X_n \sim Be(\vartheta)$ iid $\quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\Rightarrow \mathbb{E}_\vartheta(\overline{X}) = \mathbb{E}_\vartheta\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\vartheta(X_i) = \frac{1}{n}\sum_{i=1}^{n}\vartheta = \frac{1}{n}\cdot n\cdot\vartheta = \vartheta$$

example 2: $X_1 \dots X_n$ iid with average μ and variance σ^2

$$\Rightarrow S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ is an unbiased estimator of } \sigma^2$$

proof:

$$\mathbb{E}\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \sum_{i=1}^{n}\mathbb{E}\left[(X_i-\overline{X})^2\right] = n\,\mathbb{E}\left[(X_1-\overline{X})^2\right] =$$

$$= n\cdot\mathbb{E}\left[(X_1-\mu+\mu-\overline{X})^2\right] =$$

$$= n\cdot\left\{\mathbb{E}\left((X_1-\mu)^2\right) + 2\mathbb{E}\left((X_1-\mu)(\mu-\overline{X})\right) + \mathbb{E}\left((\mu-\overline{X})^2\right)\right\} =$$

$$= n\cdot\mathrm{Var}(X_1) + 2n\,\mathbb{E}\left(X_1(\mu-\overline{X})\right) - \underbrace{2n\,\mathbb{E}\left(\mu(\mu-\overline{X})\right)}_{=2n\mu\,\mathbb{E}(\mu-\overline{X})} + n\cdot\mathbb{E}\left((\mu-\overline{X})^2\right) =$$

$$= n\sigma^2 + 2n\,\mathbb{E}\left(X_1(\mu-\overline{X})\right) + n\mathrm{Var}(\overline{X}) =$$

$$= n\sigma^2 - 2n\,\mathrm{Cov}(X_1,\overline{X}) + \cancel{n}\frac{\sigma^2}{\cancel{n}} =$$

$$= \sigma^2(1+n) - 2\cancel{n}\frac{1}{\cancel{n}}\sum_{i=1}^{n}\mathrm{Cov}(X_1,X_i) =$$

$$= \sigma^2(n+1) - 2\sigma^2 = \sigma^2(n+1-2) = \sigma^2(n-1) \qquad \square$$

obs: $\mathbb{E}_\vartheta\left((T-\tau(\vartheta))^2\right) := MSE_{\tau(\vartheta)}(T) = \mathrm{Var}(T) + \left(\mathbb{E}_\vartheta(T)-\vartheta\right)^2$

Mean Squared Error

distortion / bias of T

Def: $(X_1 \dots X_n) \sim f_\vartheta$, $\vartheta \in \Theta$, $\tau : \Theta \to \mathbb{R}$,

T^* is UMVUE if and only if:

1.) $\mathbb{E}_\vartheta(T^*) = \tau$

2.) $\mathrm{Var}_\vartheta(T^*) \le \mathrm{Var}_\vartheta(T) \quad \forall\, T$ unbiased estimator of τ

Theorem: If an UMVUE exist, then it is unique.

Def: $X_1 \dots X_n \sim f_\vartheta$ iid. $\Rightarrow f_\vartheta(X_1 \dots X_n) = \prod_{i=1}^{n} f_\vartheta(X_i) = L_n(\vartheta, x_1 \dots x_n)$ (with $X_i = x_i$) → likelihood function

Def: $U_n(\vartheta, X_1 \dots X_n) := \frac{\partial}{\partial \vartheta} \log\left(L_n(\vartheta, X_1 \dots X_n)\right) = \sum_{i=1}^{n} \frac{\partial}{\partial \vartheta} \log(f_\vartheta(X_i))$

↳ contribute

Def $I_n(\vartheta) := \mathbb{E}_\vartheta\left[U_n^2(\vartheta, X)\right]$ → Fisher Information

Theorem: inequality of Cramèr-Rao

$X_1 \dots X_n \sim f_\vartheta$ iid, $T_n = T_n(X_1 \dots X_n)$ correct estimator of $\tau = \tau(\vartheta)$

$\Rightarrow$ under right regularity properties it is true that:

$$Var_\vartheta(T_n) \geq \frac{(\tau'(\vartheta))^2}{I_n(\vartheta)}$$

Moreover $Var_\vartheta(T_n) = \frac{(\tau'(\vartheta))^2}{I_n(\vartheta)} \iff T_n = \tau(\vartheta) + a(\vartheta) U_n(\vartheta, X_1 - X_n)$

Def: An estimator T_n that satisfies the Cramèr-Rao limit is said to be efficient

Up to this point we have focused on the properties of point estimators and the criteria that make them desirable. However, in practice it is often not enough to provide a single numerical estimate of a parameter. Instead, we seek to quantify the uncertainty associated with estimation, which naturally leads to the concepts of interval estimation, confidence intervals, p-values, and statistical errors.

↳

Instead of estimating ϑ with a single aleatory variable, we look for an interval $I = I(X_1 \dots X_n) \subseteq \Theta$ such that $\mathbb{P}_\vartheta(\vartheta \in I) \simeq 1$

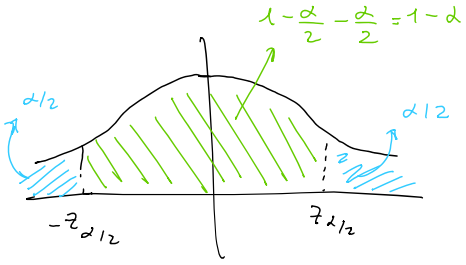
$\Rightarrow$ such interval is called confidence interval

Def: $(X_1 \dots X_n) \sim f_\vartheta$, $\vartheta \in \Theta \subseteq \mathbb{R}$, $\alpha \in (0,1)$. A $1-\alpha$ confidence interval is a random interval $I = I(X_1 \dots X_n)$ s.t. $\mathbb{P}_\vartheta(\vartheta \in I) \geq 1-\alpha \quad \forall \vartheta \in \Theta$

example:

$X_1 \dots X_n \sim N(\mu, \sigma_0^2)$ μ unknown, σ_0^2 known

$$\Rightarrow \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma_0} \sim N(0,1) \Rightarrow \mathbb{P}_\mu\left(-z_{\alpha/2} < \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma_0} < z_{\alpha/2}\right) = 1-\frac{\alpha}{2}-\frac{\alpha}{2} = 1-\alpha$$



$$\Rightarrow \mathbb{P}_\mu\left(-\frac{\sigma_0}{\sqrt{n}}z_{\alpha/2} - \bar{X} < -\mu < \frac{\sigma_0}{\sqrt{n}}z_{\alpha/2} - \bar{X}\right) = 1-\alpha$$

$$\Rightarrow \mathbb{P}_\mu\left(\bar{X} - \frac{\sigma_0}{\sqrt{n}}z_{\alpha/2} < \mu < \bar{X} + \frac{\sigma_0}{\sqrt{n}}z_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow \left[\bar{X} - \frac{\sigma_0}{\sqrt{n}}z_{\alpha/2} \;;\; \bar{X} + \frac{\sigma_0}{\sqrt{n}}z_{\alpha/2}\right] \text{ is a } 1-\alpha \text{ confidence interval for } \mu.$$

Def: F is a cumulative distribution function (CDF) and $\alpha \in [0,1]$. Then the α-level left quantile is $F^{-1}(\alpha)$ and the α-level right quantile is $F^{-1}(1-\alpha)$

$$\rightsquigarrow \mathbb{P}\left(X \leq F^{-1}(\alpha)\right) = \alpha \quad \text{and} \quad \mathbb{P}\left(X \geq F^{-1}(1-\alpha)\right) = \alpha$$

While confidence intervals provide a range of plausible values for an unknown parameter, in many situations we are instead interested in formally testing specific claims about that parameter. This leads to the framework of **parametric hypothesis testing**, where we define a null hypothesis, compute test statistics, and assess evidence through p-values and error probabilities.

In general, the situation for a parametric test is the following:

$(X_1 \dots X_n) \sim f_\vartheta$, $\vartheta \in \Theta \subseteq \mathbb{R}^d$, $\Theta_0 \subset \Theta$ and we need to decide between:

$H_0 : \vartheta \in \Theta_0$ and $H_1 : \vartheta \in \Theta_0^c = \Theta \setminus \Theta_0$

H_0: null-hypothesis; H_1: alternative hypothesis

A statistical test is a decision rule which, based on the observation of $X_1 \dots X_n$, tells me to deny or accept H_0.

Def: Given $(X_1 \dots X_n) \sim f_\vartheta$, $\vartheta \in \Theta = \Theta_0 \cup \Theta_0^c$ and a test for the hypothesis $H_0 : \vartheta \in \Theta_0$ and $H_1 : \vartheta \in \Theta_0^c$, the p-value is the smallest level α^* $(x_1 \dots x_n)$ that allows us to deny H_0 observing $x_1 \dots x_n$.

- The p-value does **not** represent the probability that the null hypothesis is true.
- Instead, it measures how **compatible the observed data are with H_0**:
  - A **small p-value** (e.g., < 0.05) suggests that the observed data would be very unlikely if H_0 were true → evidence against H_0.
  - A **large p-value** indicates that the data are consistent with H_0, so there is not enough evidence to reject it.

👉 Intuitively: the p-value is a measure of "surprise" with respect to the null hypothesis.

There will always be uncertainty, for this errors could be made:

reality \ test	accept H_0	reject H_0
H_0 is true	✓	1° type error
H_0 is false	2° type error	✓