Elements of Statistics (Part I)

Prof. Americo Cunha Jr

Rio de Janeiro State University – UERJ

americo.cunha@uerj.br

www.americocunha.org

@AmericoCunhaJr @AmericoCunhaJr

@AmericoCunhaJr @AmericoCunhaJr

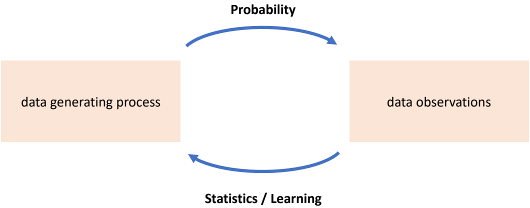
Probability vs Statistics

Probability

Given a data generating process, what are the properties of the outcomes?

Statistics

Given the outcomes, what can we say about the process that generated the data?



L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Statistical Inference

What is inference about?

Statistical inference (or learning) is the process of using data to infer the distribution that generated the data.

A typical inference question:

Given a sample $X_1, \cdots, X_n$ with distribution F_X, how to infer F_X?

Some typical inference problems:

- estimation
- confidence sets
- hypothesis testing
- clustering or classification

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Parametric vs Nonparametric

A statistical model is a set of distributions (or densities)

$$\mathfrak{F} = \left\{ p_X(x;\theta) \mid \theta \in \Theta \right\},$$

where θ is a (vector/scalar) parameter in a space of parameter Θ.

- Parametric statistics:
  - $\mathfrak{F}$ can be parametrized by a finite number of parameters (finite dimensional problem)
  - probability distribution known a priori
  - seek for distribution parameters
- Nonparametric statistics:
  - $\mathfrak{F}$ can not be parametrized by a finite number of parameters (infinite dimensional problem)
  - probability distribution unknown a priori
  - seeks for distribution shape

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Examples of statistical models

Example 1 (parametric):

$X_1, \cdots, X_n$ are observations of $X \sim \mathcal{N}(\mu, \sigma)$

$$\mathfrak{F} = \left\{ p_X(x;\mu,\sigma) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\,\sigma^2} \right\} \;\middle|\; \mu \in \mathbb{R},\;\; \sigma > 0 \right\}$$

The problem is to estimate μ and σ.

Example 2 (nonparametric):

$X_1, \cdots, X_n$ are independent observations from an unkown F_X

$$\mathfrak{F} = \{\text{set of all possible CDFs}\}$$

The problem is to estimate F_X.

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Frequentist vs Bayesian

The two dominant approaches (paradigms) for inference are:

- Frequentist (or classical):
  - probability is a limit frequency
  - parameters are fixed
  - inference based on asymptotic properties
- Bayesian:
  - probability is a degree of belief
  - data are fixed
  - inference based on posterior distribution

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Statistical Estimator

A statiscal estimator is a rule for calculating an estimate of a given quantity based on observed data.

Estimation deals with three distinct objects:

- estimand (quantity to be estimated)
- estimator (estimation rule)
- estimate (estimation result)

There are two types of estimators:

- point estimator
- interval estimator

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Point Estimator

Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed (iid) data points from some distribution F_X.

A point estimator $\widehat{\theta}_n$ for parameter θ is a random variable

$$\widehat{\theta}_n = g(X_1, \cdots, X_n).$$

This estimator can be thought a single "best guess" of some quantity of interest (a parameter in a parametric model, a CDF, a PDF, etc).

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Examples of point estimators

$X_1, \cdots, X_n$ are independent observations of $X \sim \mathcal{N}(\mu, \sigma^2)$

Point estimators for μ and σ^2 are given by:

- sample mean

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- sample variance

$$\widehat{\sigma^2}_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu}_n)^2$$

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Quantified properties of a point estimator

- Bias:

$$\texttt{bias}\left(\widehat{\theta}_n\right) = \mathbb{E}\left\{\widehat{\theta}_n\right\} - \theta$$

If $\texttt{bias}\left(\widehat{\theta}_n\right) = 0$ the estimator is said to be unbiased

Distance between the average of the collection of estimates, and the single parameter being estimated.

- Mean Square Error:

$$\texttt{MSE}\left(\widehat{\theta}_n\right) = \mathbb{E}\left\{\left(\widehat{\theta}_n - \theta\right)^2\right\} = \texttt{bias}\left(\widehat{\theta}_n\right)^2 + \texttt{var}\left(\widehat{\theta}_n\right)$$

Indicate how far, on average, the collection of estimates are from the single parameter being estimated.

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Behavioral properties of a point estimator

- $\widehat{\theta}_n$ is said to be consistent if $\widehat{\theta}_n \xrightarrow{p} \theta$

  Increasing the sample size increases the probability of the estimator being close to the population parameter.

- $\widehat{\theta}_n$ is said to be asymptotically normal if

$$\frac{\widehat{\theta}_n - \theta}{\sqrt{\mathtt{var}\left(\widehat{\theta}_n\right)}} \xrightarrow{d} \mathcal{N}(0,1),$$

A consistent estimator whose distribution around the true parameter approaches a normal distribution.

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Confidence Interval

A $1-\alpha$ confidence interval for parameter θ is a random interval $C_n = (a, b)$, where $a = a(X_1, \cdots, X_n)$ and $b = b(X_1, \cdots, X_n)$ are random variables such that

$$\mathcal{P}\left\{a \leq \theta \leq b\right\} \geq 1 - \alpha, \quad \text{for all} \quad \theta \in \Theta.$$

This random interval envelopes θ with probability $1-\alpha$.

Remark:
C_n is a random variable, while θ is fixed parameter.

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Example of confidence interval

"83% of the population favor invest more on education."

What parameter is estimated on this poll ?
p = proportion of people who favor invest more on education

"Poll is accurate to within 4 points 95% of the time."

$$\Longleftrightarrow$$

$C_n = (79, 87) = 83 \pm 4$ is a 95% confidence interval for the poll

If you form a confidence interval this way every day for the rest of your life, 95% of your intervals will contain the true parameter p.

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Hypothesis Testing

- H_0: **null hyphotesis**

  The hyphotesis to be retained or rejected

- H_1: **alternative hyphotesis**

  H_1 is rejected if H_0 is true
  H_1 is accepted if H_0 is false

Does the data provide sufficient evidence to reject H_0?

	Retain Null	Reject Null
H_0 is true	correct decision	type I error
H_1 is true	type II error	correct decision

Table: Possible outcomes of hypothesis testing.

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

An example of hypothesis test

Testing if a Coin is Fair

$$X_1, \cdots, X_n \sim \text{Bernoulli}(p)$$

$$H_0 : p = 1/2 \text{ versus } H_1 : p \neq 1/2$$

It seems reasonable to reject H_0 if

$$T = |\widehat{p}_n - 1/2| \text{ is large}$$

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

Remarks about hypothesis test

Important remarks about hypothesis test:

- Useful to see if there is evidence to reject H_0
- Not useful to prove that H_0 is true
- Failure to reject H_0 might occur because:
  - H_0 is true
  - test is not effective

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

References

S. E. Fienberg, *What Is Statistics?*, **Annual Review of Statistics and Its Application**, 1:1–9, 2014.

V. Stodden, *Reproducing Statistical Results*, **Annual Review of Statistics and Its Application**, 2:1–19, 2015.

G. Claeskens, *Statistical Model Choice*, **Annual Review of Statistics and Its Application**, 3:233–256, 2016.

L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

L. Wasserman, **All of Nonparametric Statistics**, Springer, 2007.

G. Casella, **Statistical Inference**, Thomson Press (India) Ltd, 2008.

J. Shao, **Mathematical Statistics**, Springer, 2nd Edition, 2007

J. E. Gentle, **Computational Statistics**, Springer; 2009

J. E. Gentle, W. K. Härdle, and Y. Mori, **Handbook of Computational Statistics: Concepts and Methods**, Springer, 2nd revised and updated Edition, 2012.