Elements of Statistics

Prof. Americo Cunha Jr

americocunha.org

@AmericoCunhaJr

@AmericoCunhaJr @AmericoCunhaJr

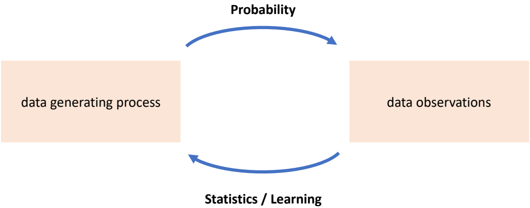
Probability vs Statistics

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

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

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.

Nonparametric Estimators

Empirical Distribution Function

$X_1 < X_2 < \cdots < X_n$ are independent observations of X

The empirical distribution function (empirical CDF) is an estimator for distribution function F_X defined by

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\left(X_i \leq x\right),$$

where

$$\mathcal{I}\left(X_i \leq x\right) = \begin{cases} 1 & \text{if} \quad X_i \leq x \\ 0 & \text{if} \quad X_i > x. \end{cases}$$

Empirical CDF is consistent and unbiased estimator

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

Histogram Estimator

$X_1 < X_2 < \cdots < X_n$ are independent observations of X

Split the support of X into a denumerable number of bins $\mathcal{B}_m$ with width h_m, i.e.,

$$\text{Supp}\, X = \bigcup_{m=-\infty}^{+\infty} \mathcal{B}_m = \big[(m-1)\, h_m, m\, h_m\big]$$

The histogram is an estimator for probability density function p_X defined by

$$\widehat{p}_n(x) = \sum_{m=-\infty}^{+\infty} \frac{\nu_m}{n\, h_m}\, \mathbb{1}_{\mathcal{B}_m}(x),$$

where ν_m is the number of samples of X in $\mathcal{B}_m$ and

$$\mathbb{1}_{\mathcal{B}_m}(x) = \begin{cases} 1 & \text{if} \quad x \in \mathcal{B}_m \\ 0 & \text{if} \quad x \notin \mathcal{B}_m. \end{cases}$$

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

Kernel Density Estimator

$X_1, X_2, \cdots, X_n$ are observations of X

The kernel density estimator for the probability density function p_X is defined by

$$\widehat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right),$$

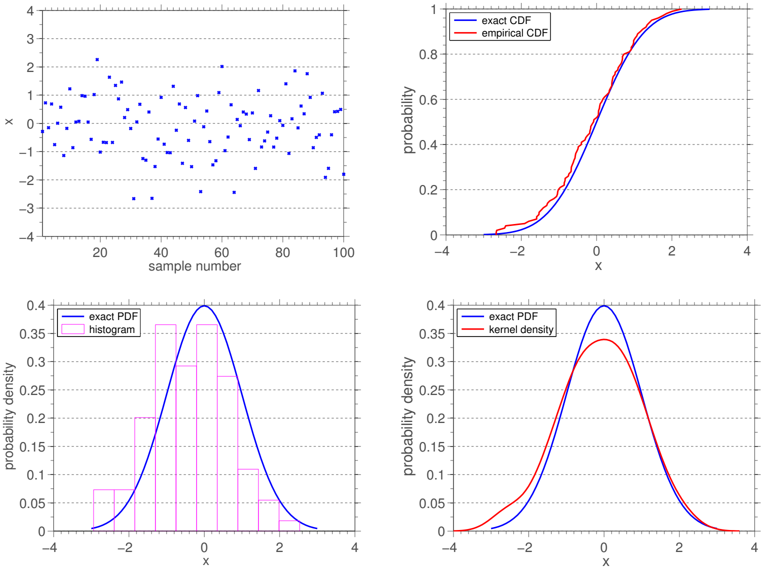
where $h > 0$ is the estimator bandwidth and the kernel K is a smooth function such that

- $K(x) \geq 0$
- $\displaystyle\int_{\mathbb{R}} K(x)\, dx = 1$
- $\displaystyle\int_{\mathbb{R}} x\, K(x)\, dx = 0$
- $\displaystyle\int_{\mathbb{R}} x^2\, K(x)\, dx > 0$

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

An example in nonparametric estimation

100 samples of $X \sim \mathcal{N}(0,1)$



Parametric Estimators

Statistical Moments

Let random variable X be parametrized by vector parameter

$$\theta = (\theta_1, \theta_2, \cdots, \theta_k).$$

For $1 \leq j \leq k$, the j-th moment of X is

$$\alpha_j(\theta_1, \theta_2, \cdots, \theta_k) = \mathbb{E}\left\{X^j\right\} = \int_{\mathbb{R}} x^j \, dF_X(x),$$

while the *j*-th sample moment is defined by

$$\widehat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j,$$

where $X_1, X_2, \cdots, X_n$ are observations of X.

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

Moments Estimator

The method of moments estimator $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k)$ is defined to be the value of $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ such that

$$
\begin{array}{rcl}
\widehat{\alpha}_1 & = & \alpha_1(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k) \\
\widehat{\alpha}_2 & = & \alpha_2(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k) \\
\vdots & \vdots & \vdots \\
\widehat{\alpha}_k & = & \alpha_k(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k).
\end{array}
$$

These estimators are very simple and consistent (under very weak assumptions), but they are often biased.

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

An example of moments estimator

$$X_1, X_2, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

- random variables first and second moment

$$\alpha_1(\mu, \sigma^2) = \mu \quad \text{and} \quad \alpha_2(\mu, \sigma^2) = \mu^2 + \sigma^2$$

- method of moments estimator

$$\widehat{\alpha}_1 = \alpha_1(\widehat{\mu}, \widehat{\sigma}^2) \quad \text{and} \quad \widehat{\alpha}_2 = \alpha_2(\widehat{\mu}, \widehat{\sigma}^2)$$

$$\Longleftrightarrow$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \widehat{\mu} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \widehat{\mu}^2 + \widehat{\sigma}^2$$

- parameters estimators

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i^2 - \widehat{\mu} \right)$$

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

Likelihood Function

$X_1, X_2, \cdots, X_n$ are independent observations of X

The likelihood function is defined by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p_X(X_i, \theta).$$

The log-likelihood function is defined by

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log p_X(X_i, \theta).$$

The likelihood function is the joint density of the data, except it is treated as a function of the parameter θ.

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

Maximum Likelihood Estimator

The maximum likelihood estimator, denoted by $\widehat{\theta}$, is the parameter vector θ that maximizes likelihood function $\mathcal{L}_n(\theta)$.

The estimatior $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k)$ is obtained from the solution of

$$
\begin{aligned}
\frac{\partial \mathcal{L}_n}{\partial \theta_1}(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k) &= 0 \\
\frac{\partial \mathcal{L}_n}{\partial \theta_2}(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k) &= 0 \\
\vdots \quad & \vdots \quad \vdots \\
\frac{\partial \mathcal{L}_n}{\partial \theta_k}(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k) &= 0.
\end{aligned}
$$

MLE is consistent and has the smallest (asymptotically) variance.

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

An example on maximum likelihood estimatation

$$X_1, X_2, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

Likelihood function:

$$\begin{aligned}
\mathcal{L}_n(\mu, \sigma) &= K \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left(-\frac{1}{2\,\sigma^2}(X_i - \mu)^2\right), \\
&= K\,\sigma^{-n} \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{n}(X_i - \mu)^2\right) \\
&= K\,\sigma^{-n} \exp\left(-\frac{n\,S^2}{2\,\sigma^2}\right) \exp\left(-\frac{n(\overline{X} - \mu)^2}{2\,\sigma^2}\right)
\end{aligned}$$

$$K = \left(\sqrt{2\,\pi}\right)^{-n}, \quad \overline{X} = n^{-1} \sum_{i=1}^{n} X_i \;\text{ and }\; S^2 = n^{-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

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

An example on maximum likelihood estimatation

Log-likelihood function:

$$\begin{aligned}\ell_n(\mu,\sigma) &= \log\left\{K\,\sigma^{-n}\,\exp\left(-\frac{n\,S^2}{2\,\sigma^2}\right)\,\exp\left(-\frac{n(\overline{X}-\mu)^2}{2\,\sigma^2}\right)\right\}\\ &= \log K - n\,log\sigma - \frac{n\,S^2}{2\,\sigma^2} - \frac{n(\overline{X}-\mu)^2}{2\,\sigma^2}\end{aligned}$$

(log-likelihood or likelihood leads to the same estimator)

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

An example on maximum likelihood estimatation

Maximum log-likelihood estimator:

$$\frac{\partial \ell_n}{\partial \mu}\left(\widehat{\mu}, \widehat{\sigma}\right) = 0 \;\; \text{and} \;\; \frac{\partial \ell_n}{\partial \sigma}\left(\widehat{\mu}, \widehat{\sigma}\right) = 0$$

$$\Longleftrightarrow$$

$$\frac{n(\overline{X} - \widehat{\mu})}{\widehat{\sigma}^2} = 0 \;\; \text{and} \;\; -\frac{n}{\widehat{\sigma}} + \frac{n\,S^2}{\widehat{\sigma}^3} + \frac{n(\overline{X} - \widehat{\mu})^2}{\widehat{\sigma}^3} = 0$$

Parameters estimators:

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \;\; \text{and} \;\; \widehat{\sigma} = S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - \widehat{\mu}\right)}$$

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

Computing Maximum Likelihood Estimates

In general MLE estimator is not known analytically.

Log-likelihood expansion around θ^j gives

$$0 = \ell_n'(\theta) \approx \ell_n'(\theta^j) + (\theta - \theta^j)\,\ell_n''(\theta^j)$$

which provides

$$\theta \approx \theta^j - \frac{\ell_n'(\theta^j)}{\ell_n''(\theta^j)}, \;\; \ell_n''(\theta^j) \neq 0$$

Newton method for MLE estimation:

$$\widehat{\theta}^{j+1} = \widehat{\theta}^j - \frac{\ell_n'(\widehat{\theta}^j)}{\ell_n''(\widehat{\theta}^j)}$$

$\widehat{\theta}^0$ defined by moments estimator

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

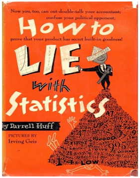
Final Remarks on Statistics

Statistical Software

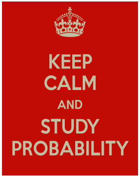
- R (programming language)
  https://www.r-project.org
- Ox (programming language)
  www.oxmetrics.net
- SciPy (Phython library)
  https://www.scipy.org
- GNU Octave
  https://www.gnu.org/software/octave
- Scilab
  http://www.scilab.org
- MATLAB
  https://www.mathworks.com

Be careful with statistics!

©

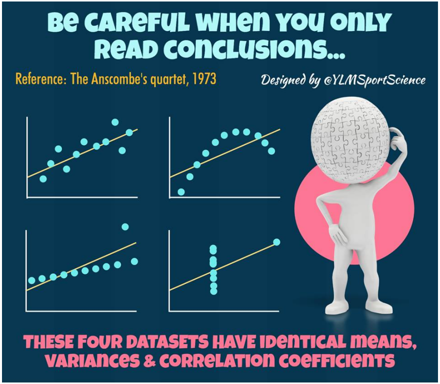


©



©

©



©

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.

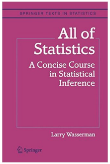
L. Wasserman, **All of Nonparametric Statistics**, Springer, 2007.

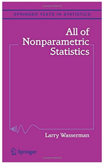
G. Casella, **Statistical Inference**, Thomson Press (India) Ltd, 2008.

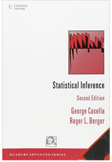
J. Shao, **Mathematical Statistics**, Springer, 2nd Edition, 2007

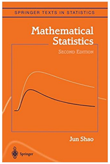
J. E. Gentle, **Computational Statistics**, Springer; 2009

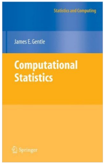
J. E. Gentle, W. K. Härdle, and Y. Mori, **Handbook of Computational Statistics: Concepts and Methods**, Springer, 2nd revised and updated Edition, 2012.

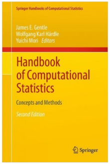












How to cite this material?

A. Cunha Jr, *Elements of Statistics*, 2021.

@AmericoCunhaJr

@AmericoCunhaJr @AmericoCunhaJr

© copyrighted material

- 76% of the statistics are made up:
  https://thecooleststuffever.com/76-statistics-are-made-t-shirt
- How to lie with statistics:
  Wikimedia Commons, File:How to Lie with Statistics.jpg — Wikimedia Commons, the free media repository https://en.wikipedia.org/wiki/File:How_to_Lie_with_Statistics.jpg
- Keep calm and study probability:
  https://keepcalms.com/p/keep-calm-and-study-probability/
- Keep calm and study statistics:
  https://www.keepcalmandposters.com/poster/791012_keep_calm_and_study_statistics
- Be careful when you only read conclusions:
  https://ylmsportscience.com/2016/07/30/stats-be-careful-when-you-only-read-conclusions-by-ylmsportscience/