STA 104 Applied Nonparametric Statistics

Chapter 1: Introduction

Xiner Zhou

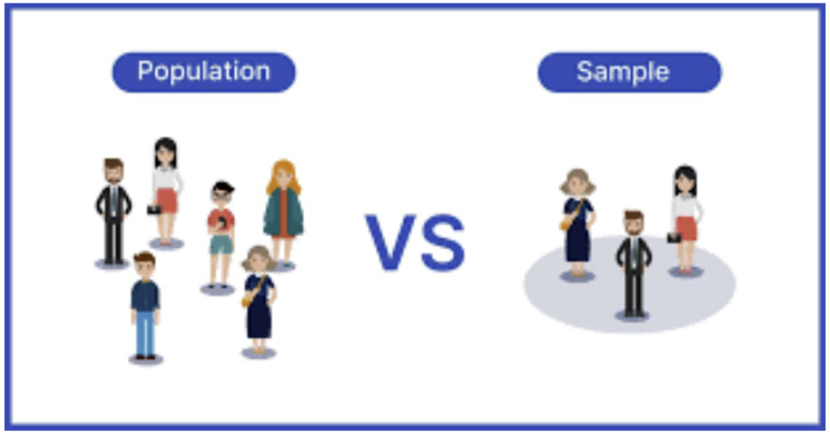
Department of Statistics, University of California, Davis

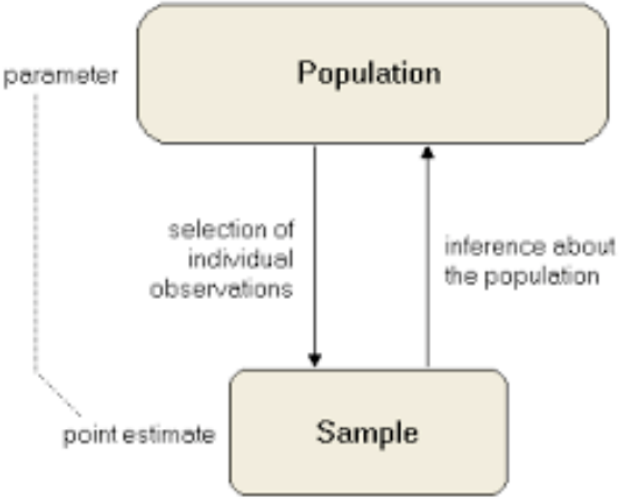
Table of contents

Parametric and Nonparametric Methods

Statistics or Statistical Inference:

to make inferences about a larger potentially observable collection of data called a population, using a sample.





We associate distributions with populations.

- families of distributions (parameters)
  - normal $\mathrm{N}\left(\mu, \sigma^2\right)$
  - $\mathrm{B}(n, p)$ distribution
  - Uniform
  - multinomial
  - Poisson
  - exponential
  - gamma
  - beta
  - Cauchy

Parametric methods:

Given a set of random sample from some population with a distribution that is assumed to be a member of a family such as the normal or binomial, to estimate or test hypotheses about the unknown parameters.

For a sample from a normal distribution

- sample mean is a point (i.e., a single value) estimate of the parameter μ
- CI: $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$
- z or t-test provides a measure of the strength of the evidence provided by a sample in support of an a priori hypothesized value μ_0 for population mean

Normal distribution? theoretical grounds, past experience

Central limit theorem justifies such a use of the normal distribution: asymptotic approximations

Parametric inference may be inappropriate or even impossible.

- no obvious family of distributions that provides our data
- no clearly defined parameters about which we can make inferences

$\Rightarrow$ Nonparametric methods= distribution-free methods

Nonparametric methods= distribution-free methods

Make inferences about parameters in wider sense:

we do not assume our samples are associated with any prespecified family of distributions

Does not mean assumption free:

always make some assumptions about the underlying population distribution

e.g. Nonparametric test: methods can be applied to samples from populations having distributions only specified in broad terms

- e.g., as being continuous, symmetric
- distribution-free property: distribution of test statistic is the same no matter what the population distribution may be

Robust:

do not depend critically on the correctness of an assumption that samples come from a distribution in a particular family

Historical Notes

- Francis Galton (1892): developed a method for classifying and assess agreement between patterns (categorical data) on fingertips

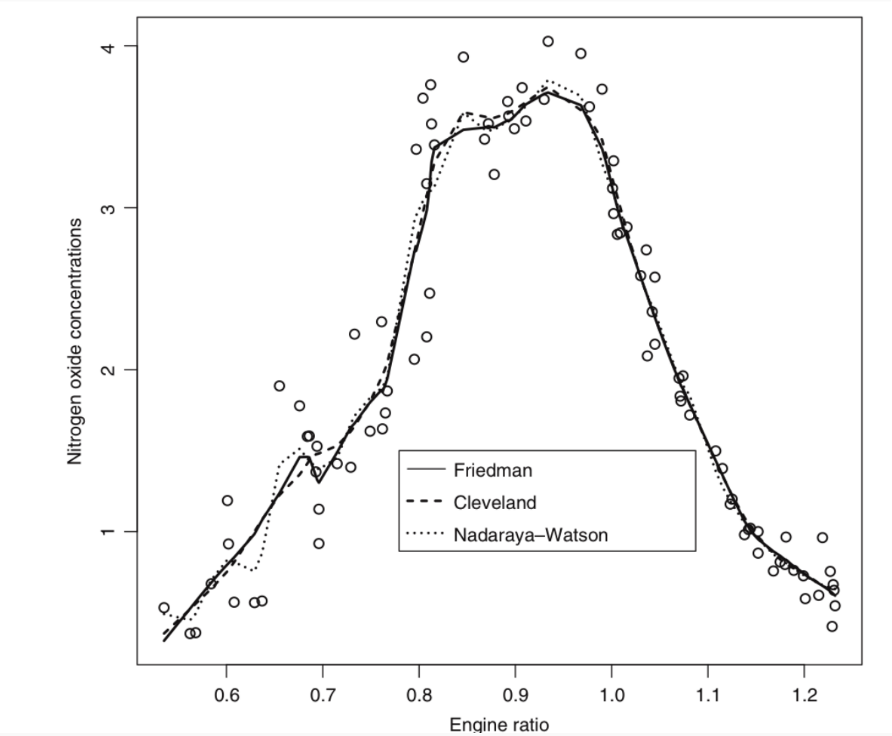
- Karl Pearson (1900): chisquared goodness-of-fit test applicable to any discrete distribution
- Spearman (1904): rank correlation coefficient
- Ronald Fisher and E.J.G. Pitman in the 1930s: permutation tests

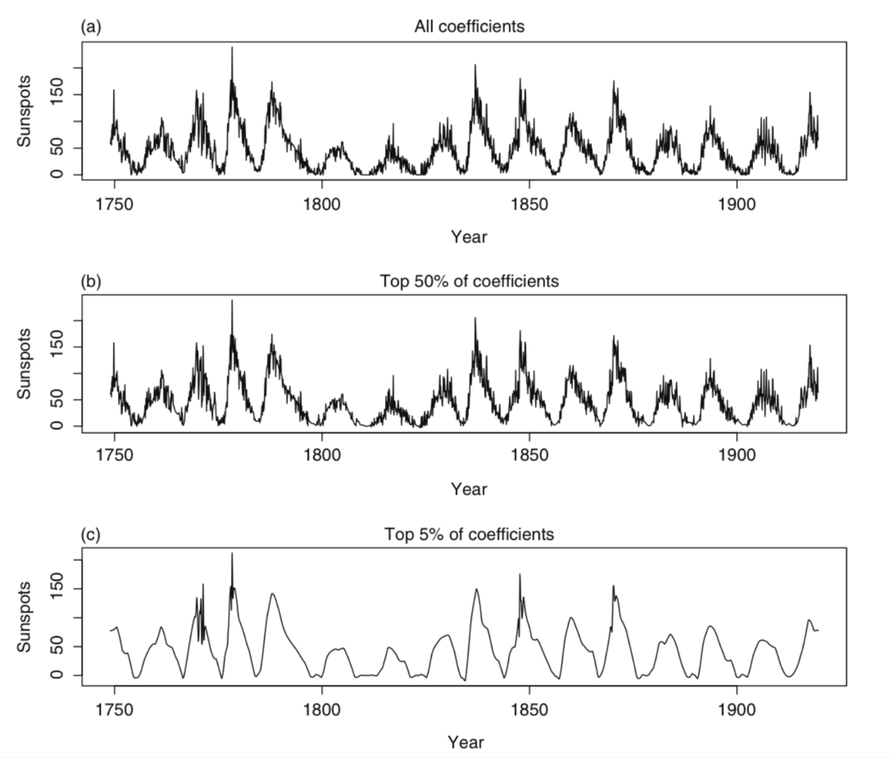
Historical Notes

- Friedman, Smirnov, Wilcoxonin 1930s
  - observations consisting simply of preferences or ranks could be used in permutation tests to make some inferences
  - even if we have precise measurements, we sometimes lose little useful information by ranking them in increasing order of magnitude and basing analyses on these ranks.
  - when assumptions of normality are not justified, analyses based on ranks may be the most efficient available and robust
- Hodges and Lehmann (1963) : interval estimation
- modern, computer intensive procedures of bootstrapping introduced by Efron (1979)

Historical Notes

- Nonparametric Regression
  - Local averaging
  - Local regression
  - Kernel smoothing
  - Wavelets





Nonparametric advantage?

ill-founded hopes that data would fit a restricted mathematical model with few parameters, and emphasis on simplifying concepts such as linearity, have often been replaced by the use of robust methods.

Strength: when insufficient theory or data to justify, or to test compatibility with, specific distributional models.