Biostatistics

Nuno Sepúlveda, 07.10.2024

About myself

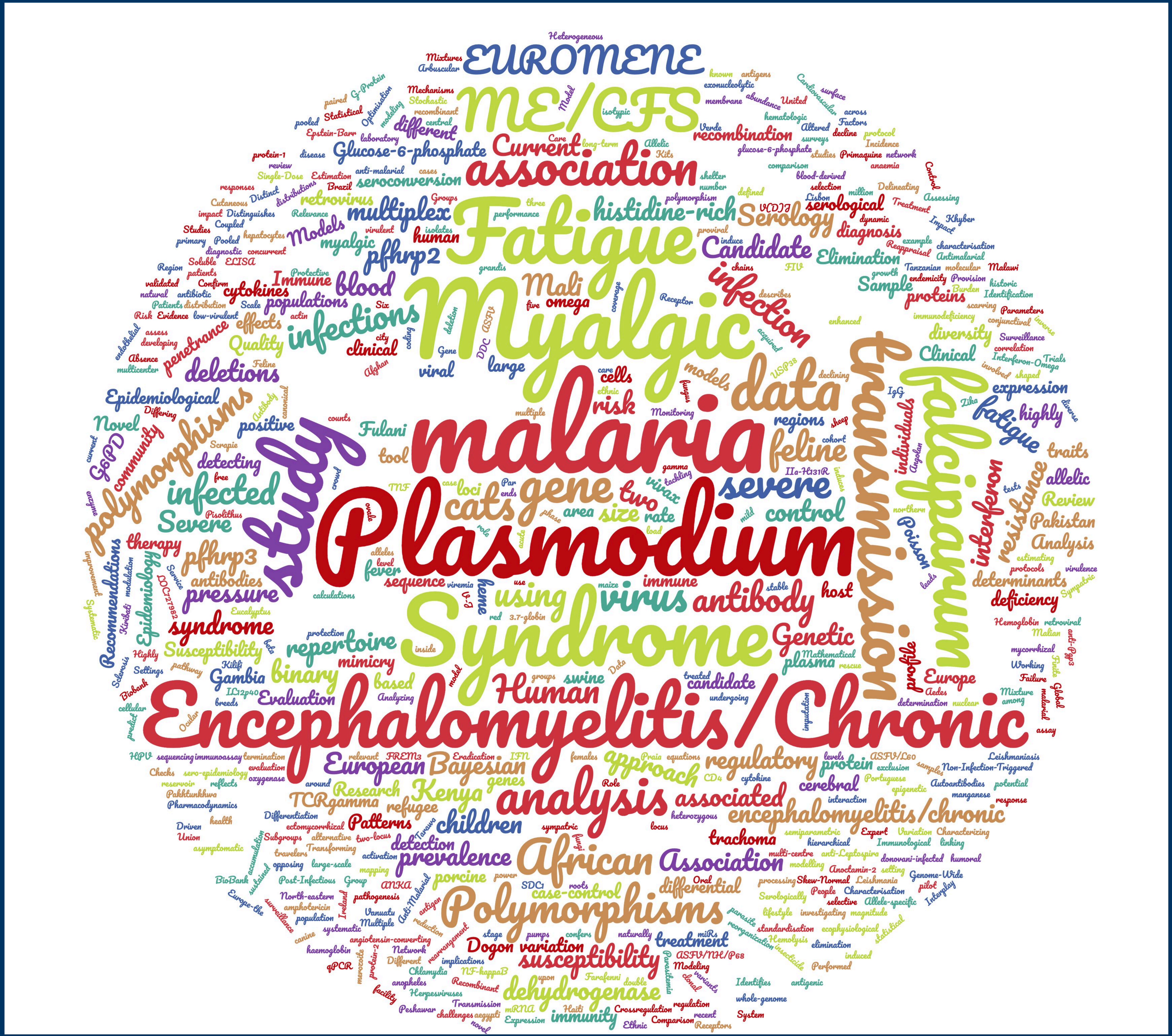
BSc and MSc in Applied Mathematics (Statistics)
PhD in Biomedical Sciences

@Gulbenkian Institute for Science (2001-2009, Portugal)
Theoretical Immunology Group / Quantitative Biology Group

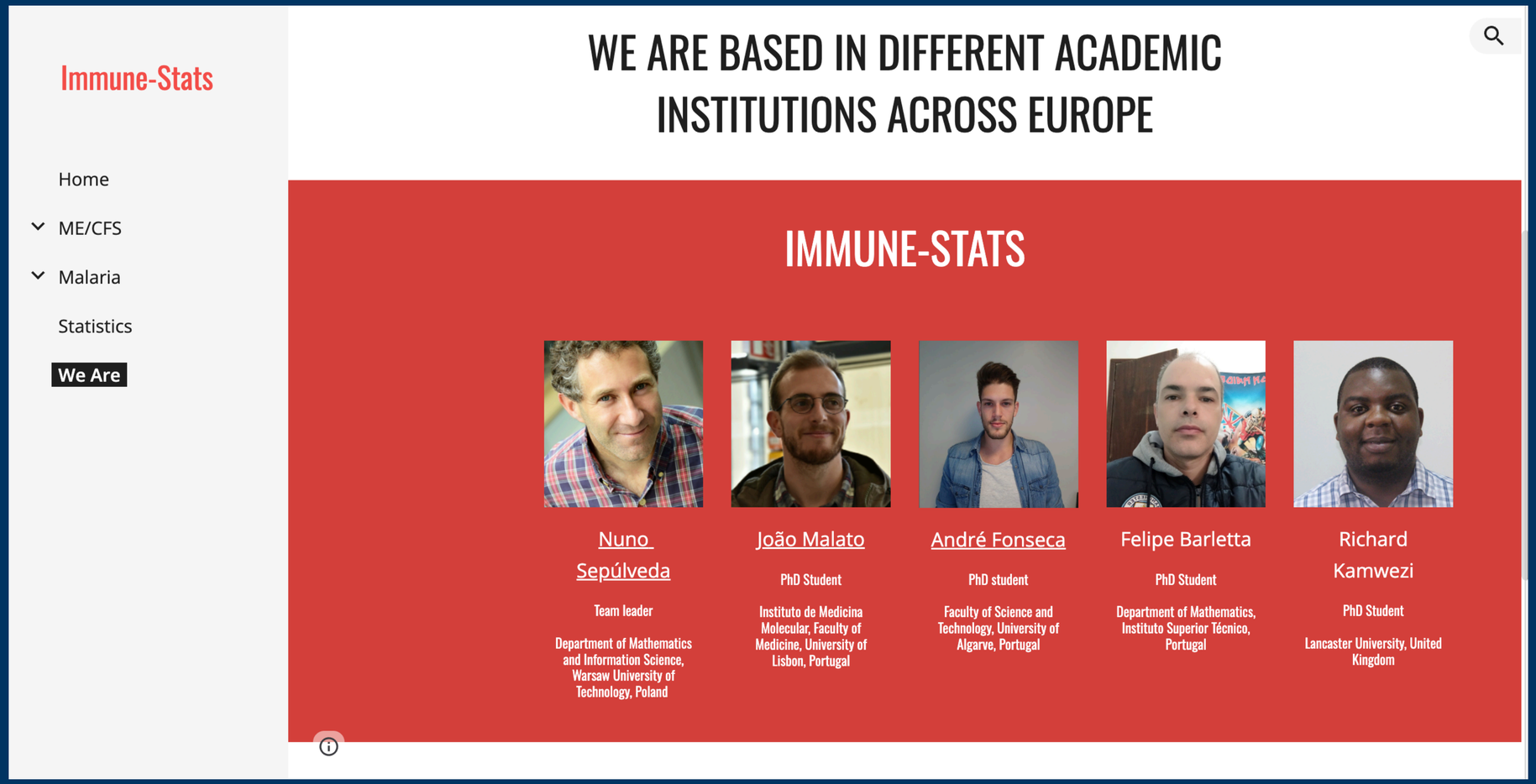
@London School of Hygiene and Tropical Medicine (2010-2019, United Kingdom)
Research Fellow in Statistical Genetics and Genetic Epidemiology
Assistant Professor in Biostatistics and Statistical Genetics

@Charité Medical University of Berlin (2020-2021, Germany)
Consultant in Bionformatics and Biostatistics

@Politechnika Warszawa (2021-2023, Poland)
Visiting Professor (ULAM Programme - NAWA)


Plasmodium
malaria
Encephalomyelitis/Chronic
Myalgic
Syndrome
Fatigue
ME/CFS
falciparum
transmission
study
EUROMENE
association
analysis
African
Polymorphisms
data
gene
infection
infections
polymorphisms
virus
antibody
infected
severe
susceptibility
dehydrogenase
Association
encephalomyelitis/chronic
regulatory
Human
approach
Kenya
European
Bayesian
fatigue
cats
feline
interferon
resistance
children
prevalence
Serology
Candidate
histidine-rich
multiplex
Glucose-6-phosphate
Current
pfhrp2
pfhrp3
syndrome
pressure
repertoire
binary
Epidemiology
Recommendations
Genetic
Clinical
diversity
control
size
loci
deletions
Epidemiological
Severe
therapy
fever
antibodies
heme
Susceptibility
Gambia
Evaluation
mimicry
based
swine
candidate
plasma
Europe
profile
deficiency
determinants
host
sequence
immune
cerebral
protein
associated
trachoma
differential
porcine
case-control
treatment
variation
Dogon
immunity
Patterns
refugee
Research
TCRgamma
detection
blood
populations
cytokines
Immune
clinical
penetrance
effects
Quality
community
positive
Novel
G6PD
detecting
tool
Fulani
Mali
risk
cells
models
two
vivax
rate
area
using
Sample
Elimination
diagnosis
serological
proteins
expression
highly
traits
allelic
Review
Pakistan
Analysis
regions
individuals
Poisson
recombination
seroconversion
retrovirus
myalgic
Models
different
human
omega
viral
large

Immune-Stats Group



www.immune-stats.net

Tell me about yourself

Syllabus

1. General review
    a. What is Biostatistics?
    b. Population/Sample/Sample size
    c. Type of Data – quantitative and qualitative variables
    d. Common probability distributions
    e. Work example – Malaria in Tanzania
2. Applications in Medicine
    a. Construction and analysis of diagnostic tools – Binomial distribution, ROC curve, sensitivity, specificity, Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model
3. Applications in Genetics, Genomics, and other 'omics data
    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values, estimation of biological age
    c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial
4. Other Topics
    a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
    b. Serological analysis – Gaussian (skew-normal) mixture models
    c. Advanced sample size and power calculations

Evaluation

Group Project + Presentation (40%)

Oral examination (60%)

Course material

https://github.com/immune-stats/Biostatistics_2024_2025/

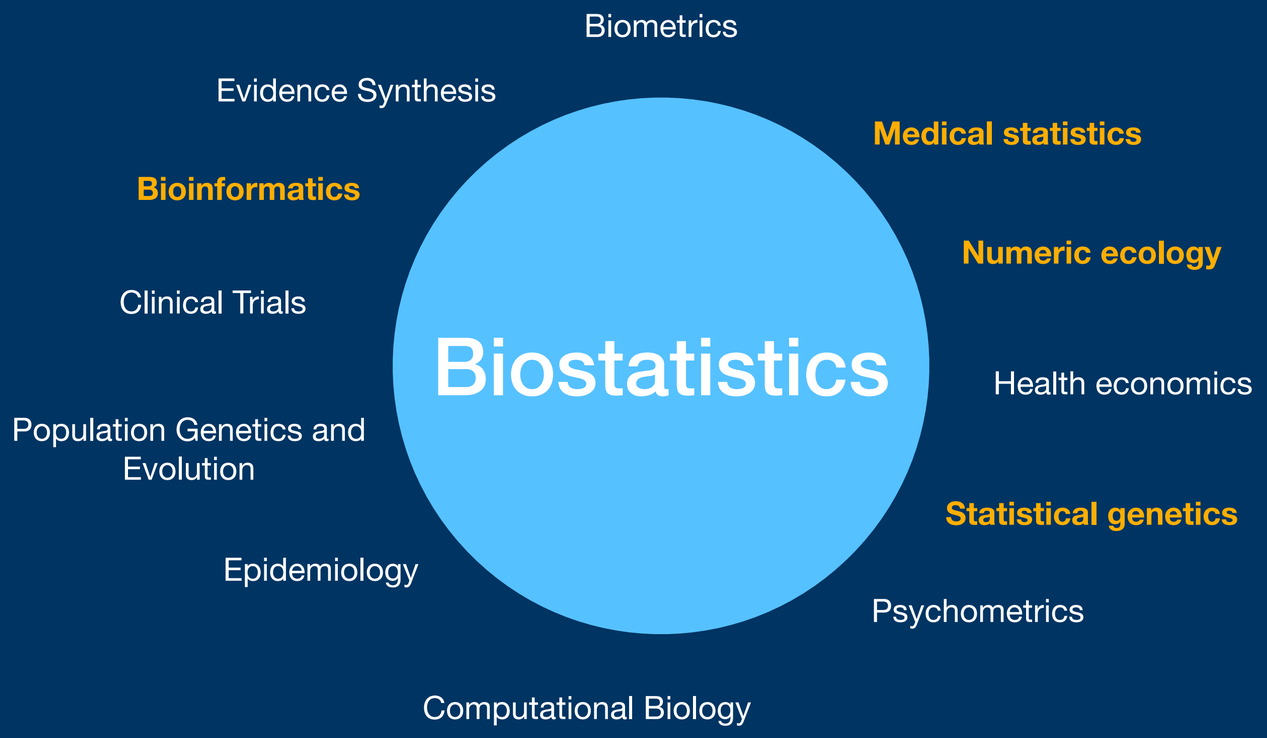
nuno.sepulveda@pw.edu.pl

Communication

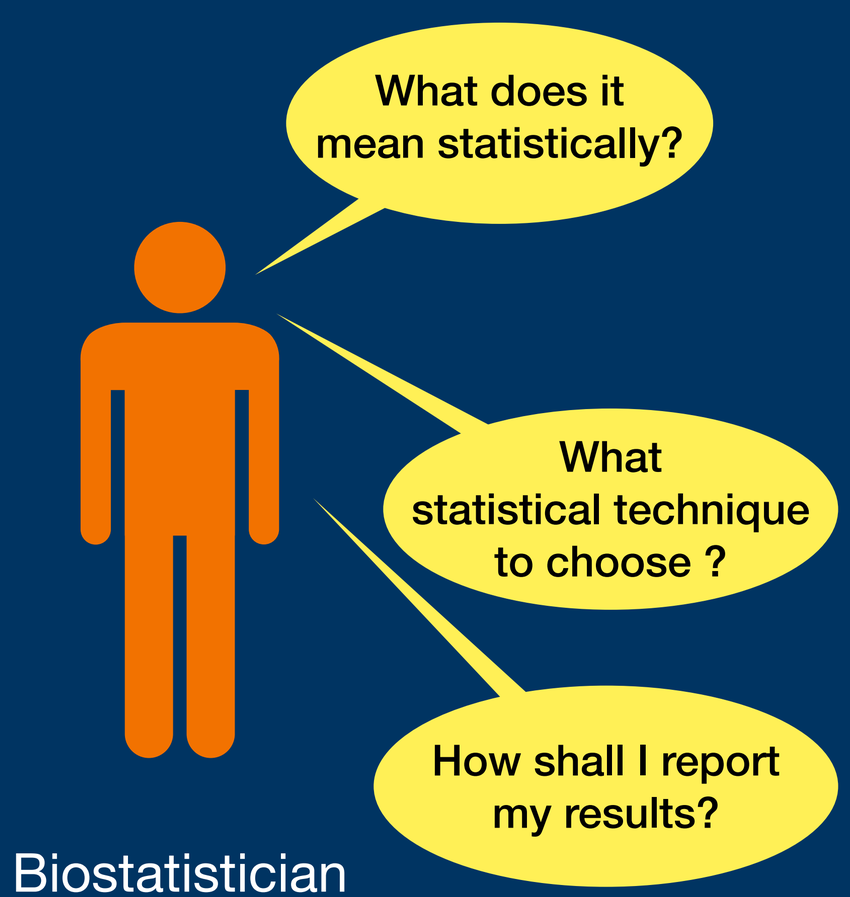
Two people should be chosen as the main contact points with me

Biostatistics

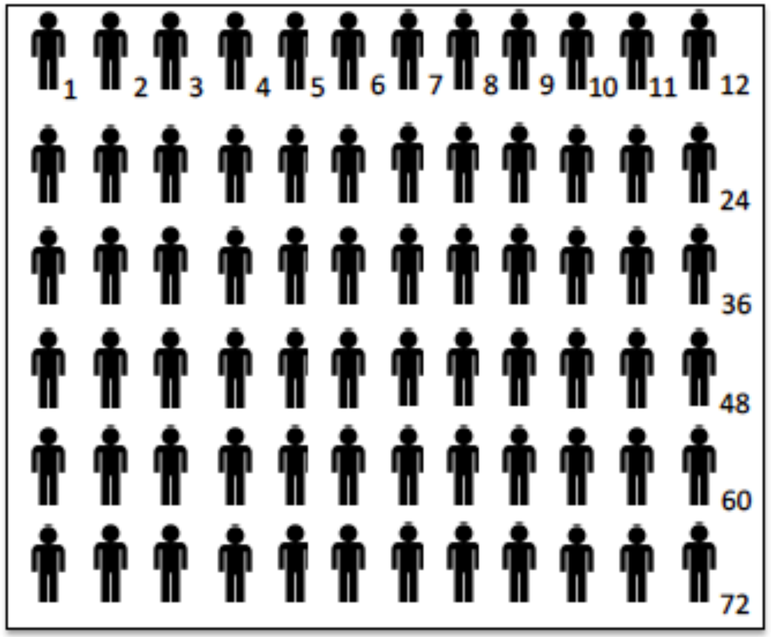
Application of statistical techniques to scientific research in health-related fields, including medicine, biology, and public health, and the development of new tools to study these areas.



Importance of communication



Population



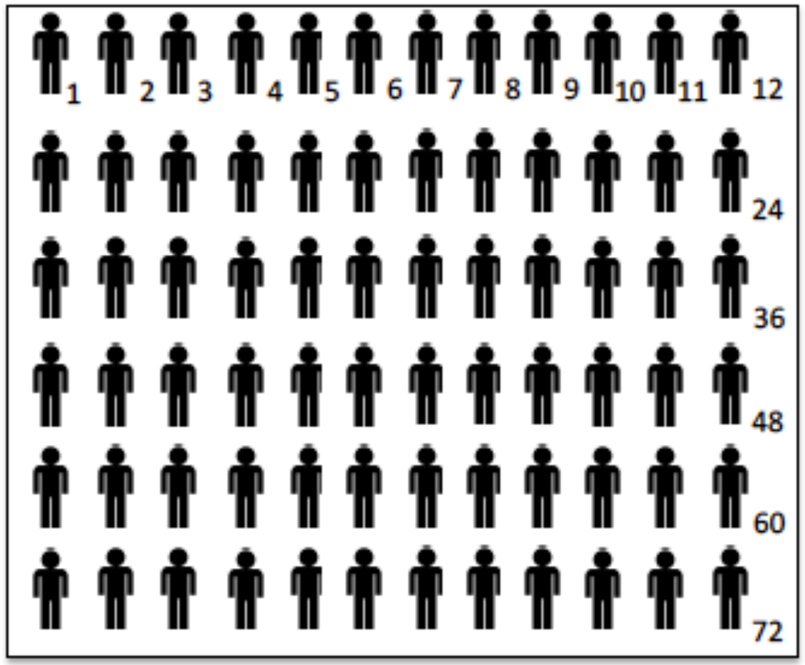
Population:

The complete set of individuals from which you what to learn something.

Population Size:

The total number of individuals in the population.

Census



Census:

It is a study conducted in the entire population.
It might require a large set of resources.

Example:

Data collected by the National Office for Statistics.

INSTITUTO NACIONAL DE ESTATÍSTICA

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Study of a rare disease.

Sample

Sampled individuals

Not-sampled individuals

Sample:

A set of individuals which it is thought to be **representative** of the whole population.

Sample size:

The total number of individuals included in the sample.

Randomized sample

Sampled individuals

Not-sampled individuals

Randomized sample:

Individuals should be **randomly** selected from the population.

Exercise:

Can you randomly select a new sample of 5 individuals from this population?

Common sampling strategies: unstratified sampling

Sampled individuals

Not-sampled individuals

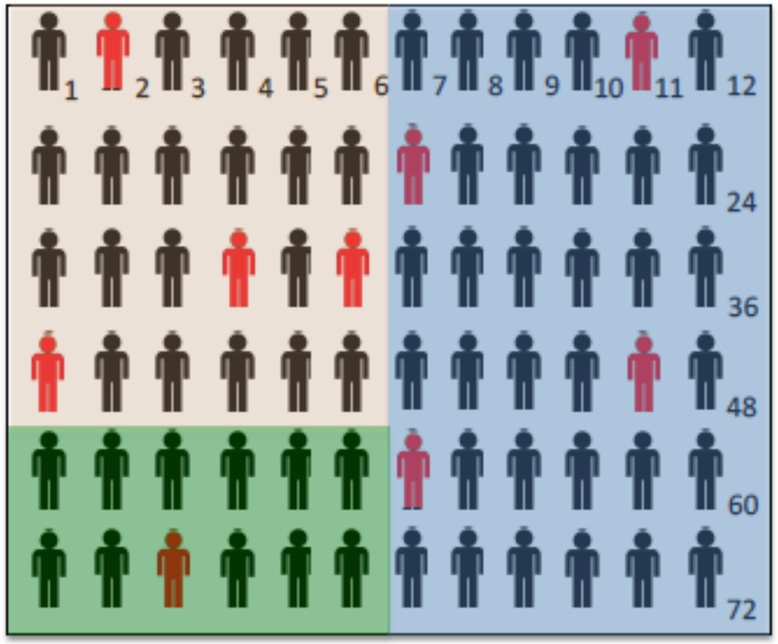
Unstratified sampling

Individuals should be **randomly** selected from the population.

It is chosen when little information is known about the population under study.

Common sampling strategies: stratified sampling

Population is composed of distinct subpopulations (stratum) with similar pattern of response.



Sampled individuals

Not-sampled individuals

Stratified sampling:

Individuals should be **randomly** selected from each stratum of the population.

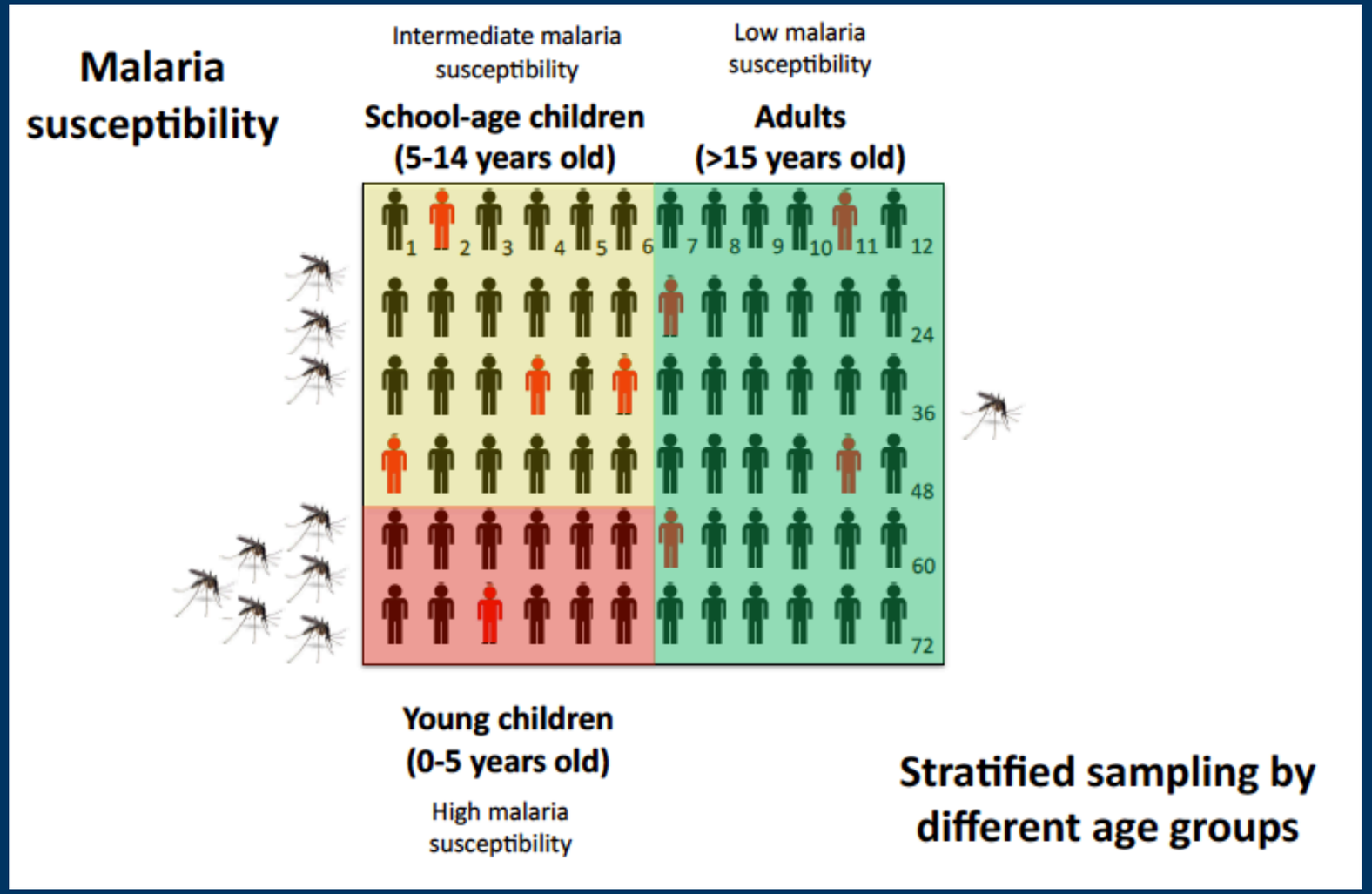
The total number of sampled individuals per stratum should scale with the respective total number in the population.

Main advantage:
Increase precision on estimates.

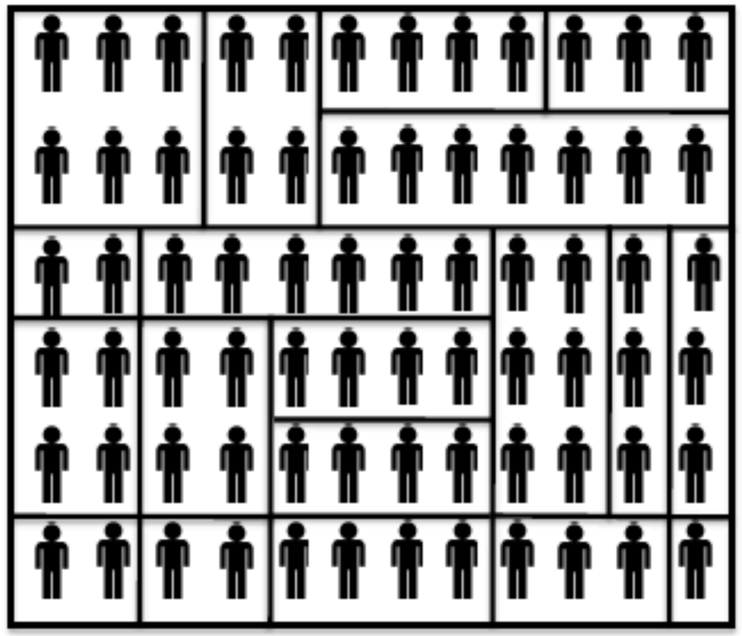
Avoids confounding and/or controls confounder effects.

Requirement:
You need to know how to define the strata!

Common sampling strategies



Common sampling strategies: cluster sampling



Sampling by clusters:

Population is divided into different clusters.

Select clusters randomly.

All individuals within a cluster are measured.

Example of clusters:

Household or compound.

Common sampling strategies: cluster sampling

Sampling by clusters:

Population is divided into different clusters.

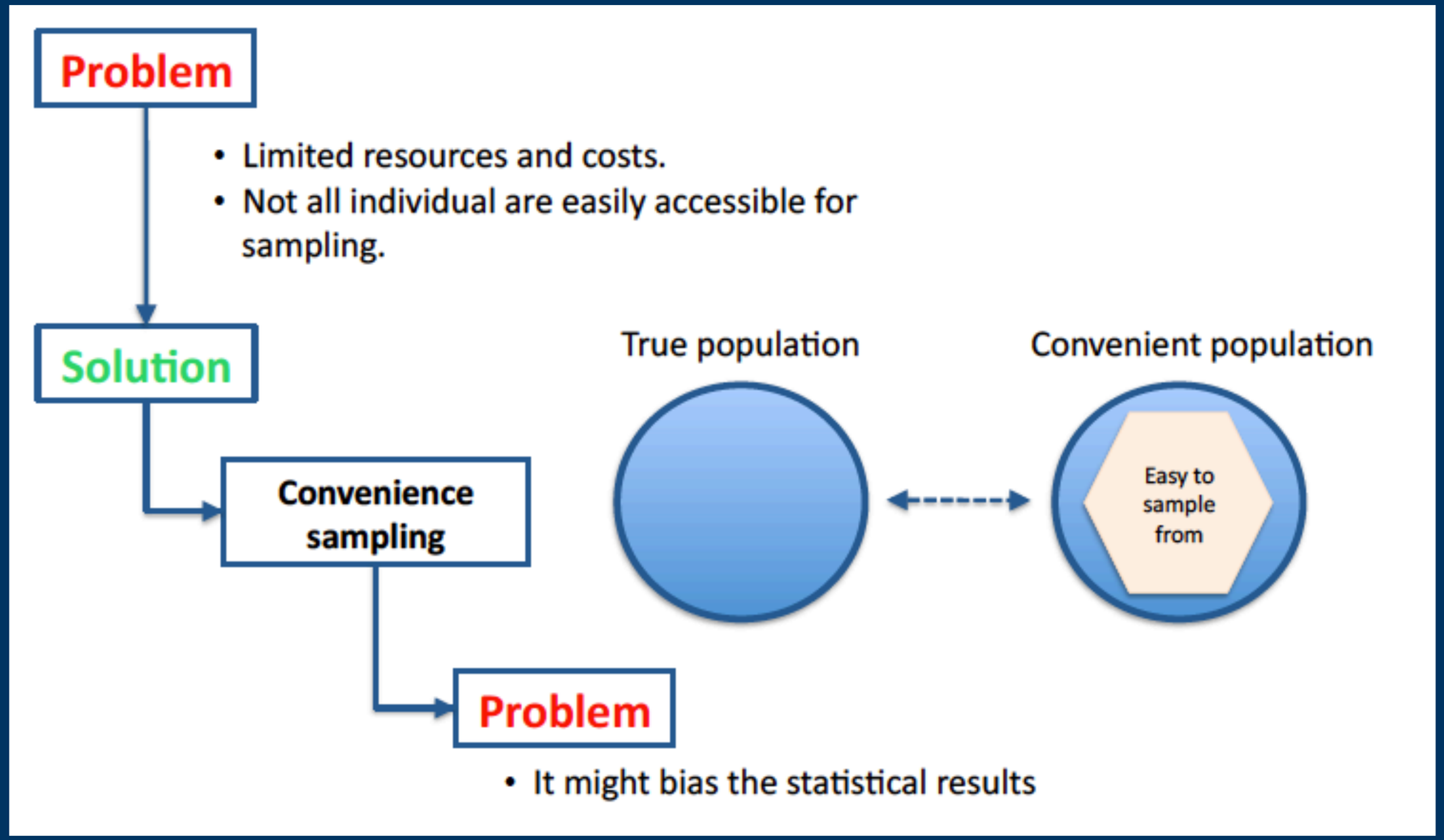
Select clusters randomly.

All individuals within a cluster are measured.

Example of clusters:

Household or compound.

In practice


Problem
• Limited resources and costs.
• Not all individual are easily accessible for sampling.
Solution
Convenience sampling
True population
Convenient population
Easy to sample from
Problem
• It might bias the statistical results

Type of variables

Quantitative

Continuous – measurements with virtual infinite precision

Ex: height, weight, time until cure, etc.

Discrete – count data

Ex: number of infection episodes per person, number of treatment doses per patient.

Qualitative

Binary – two categories

Ex: cured/not cured, presence/absence, wild type/mutated allele, etc.

Polytomous – Many categories

Ex: eye colour, genotype, socioeconomic status, ethnicity, etc.

Other types of variables

Images

Medical imaging

Videos

Symbolic data

Two cautionary notes

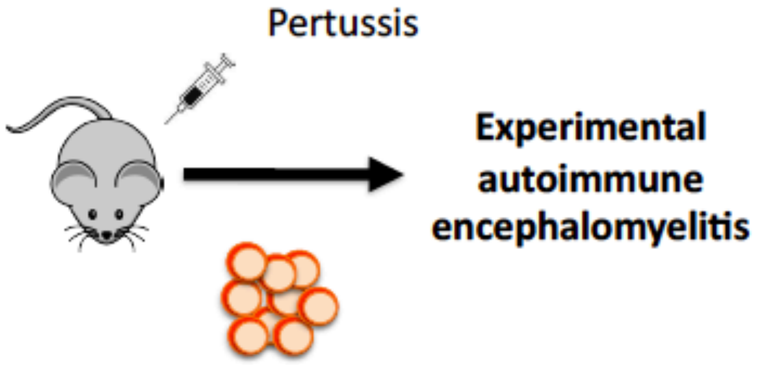
1. The true type of a variable is dependent on the unit of analysis

Define the unit of analysis as a function of the objective

2. Qualitative variables might be “hidden” in apparently quantitative variables

Always read first the data dictionary before doing any analyses

Example from the literature: EAE score



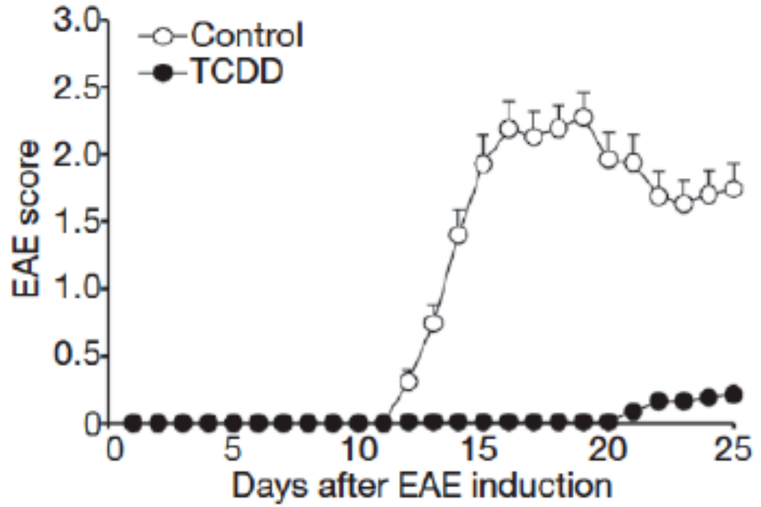
Quintana et al (2008). Nature

Vol 453|8 May 2008 doi:10.1038/nature06880 nature

ARTICLES

Control of T_{reg} and T_H17 cell differentiation by the aryl hydrocarbon receptor

Francisco J. Quintana[1], Alexandre S. Basso[1], Antonio H. Iglesias[1], Thomas Korn[1], Mauricio F. Farez[1], Estelle Bettelli[1], Mario Caccamo[2], Mohamed Oukka[1] & Howard L. Weiner[1]



(Sigma-Aldrich) intraperitoneally on days 0 and 2. Clinical signs of EAE were assessed according to the following score: 0, no signs of disease; 1, loss of tone in the tail; 2, hindlimb paresis; 3, hindlimb paralysis; 4, tetraplegia; 5, moribund.

Summarising data

Summary statistics

Maximum
Minimum
Mean
Median
Quantiles
Quartiles
Mode
Proportion
Frequencies

Standard deviation
Variance
Variation coefficient
Interquartile range

When to use each one of these summary tools?

Visualization tools

Boxplots
Scatterplots
Histograms
ECDF plots
Density plots
Strip plots

Barplots
Piecharts

Heatmaps

Common probability distributions

Which distributions do you remember?

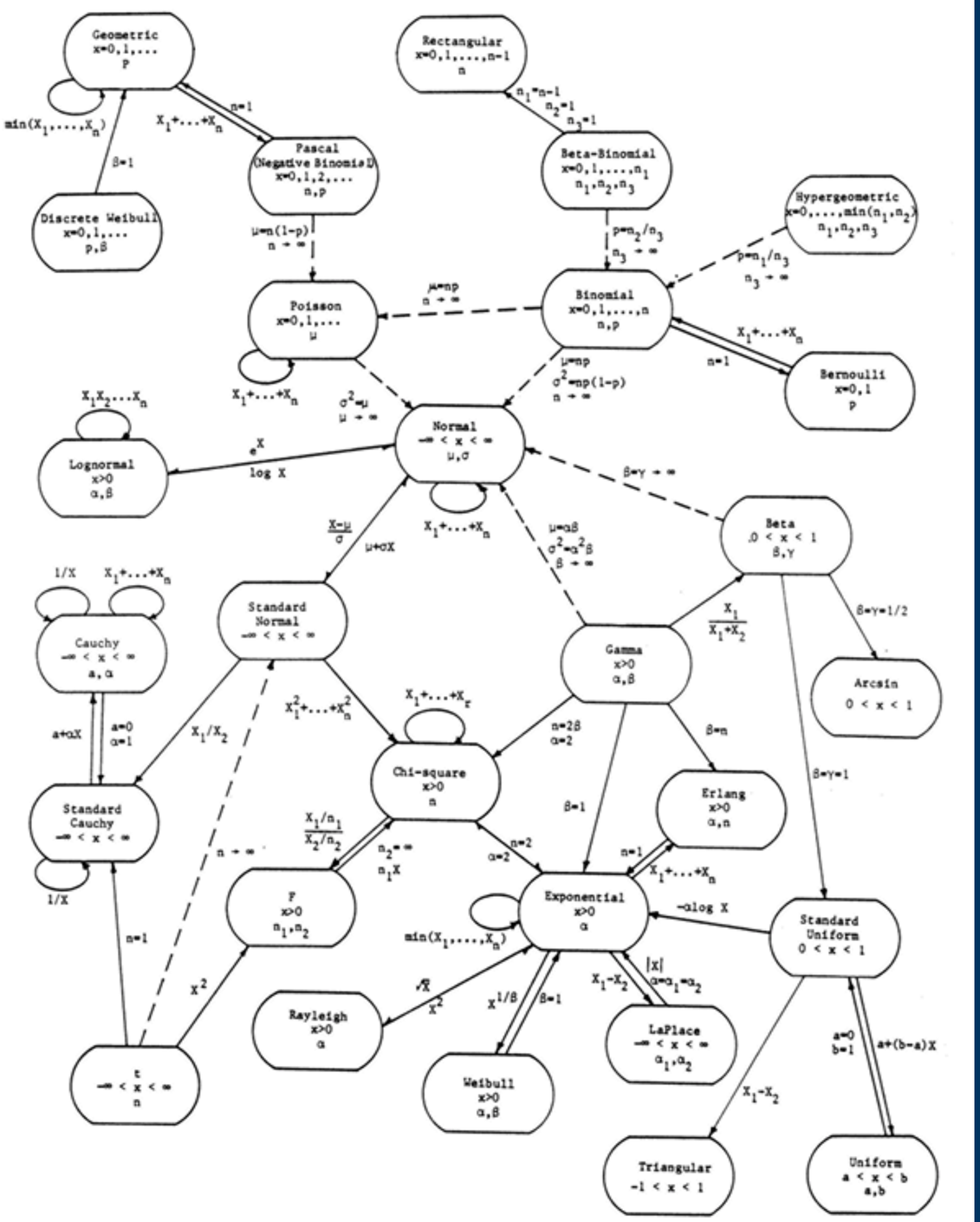


Figure 1. Relationships Among Distributions.

Leemis (1986). The American Statistician, 40 , 143

Statistical tests

Which statistical tests do you remember?

Warm-up example

PNAS

Estimating medium- and long-term trends in malaria transmission by using serological markers of malaria exposure

C. J. Drakeley*†‡, P. H. Corran*‡§, P. G. Coleman*, J. E. Tongren*, S. L. R. McDonald*, I. Carneiro*, R. Malima†¶, J. Lusingu†¶, A. Manjurano†‖, W. M. M. Nkya†‖, M. M. Lemnge†¶, J. Cox*, H. Reyburn*†, and E. M. Riley*,**

*Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom; †Joint Malaria Programme, P.O. Box 2228, Moshi, Tanzania; §National Institute for Biological Standards and Control, South Mimms EN6 3QG, United Kingdom, ‖Kilimanjaro Christian Medical Centre, P.O. Box 3010, Moshi, Tanzania; and ¶Amani Medical Research Institute, National Institute for Medical Research, P.O. Box 4, Amani, Tanzania

Edited by Louis H. Miller, National Institutes of Health, Rockville, MD, and approved February 23, 2005 (received for review November 23, 2004)

MAJOR ARTICLE

Altitude-Dependent and -Independent Variations in *Plasmodium falciparum* Prevalence in Northeastern Tanzania

Chris J. Drakeley,[1,5] Ilona Carneiro,[5] Hugh Reyburn,[1,5] Robert Malima,[1,3] John P. A. Lusingu,[3,4] Jonathan Cox,[5] Thor G. Theander,[4] Watoky M. M. M. Nkya,[2] Martha M. Lemnge,[3] and Eleanor M. Riley[5]

[1]Joint Malaria Programme and [2]Kilimanjaro Christian Medical Centre, Moshi, and [3]National Institute for Medical Research, Amani Medical Research Centre, Amani, Tanzania; [4]Centre for Medical Parasitology, University of Copenhagen, Panum Institute, Copenhagen, Denmark; [5]Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

Warm-up example

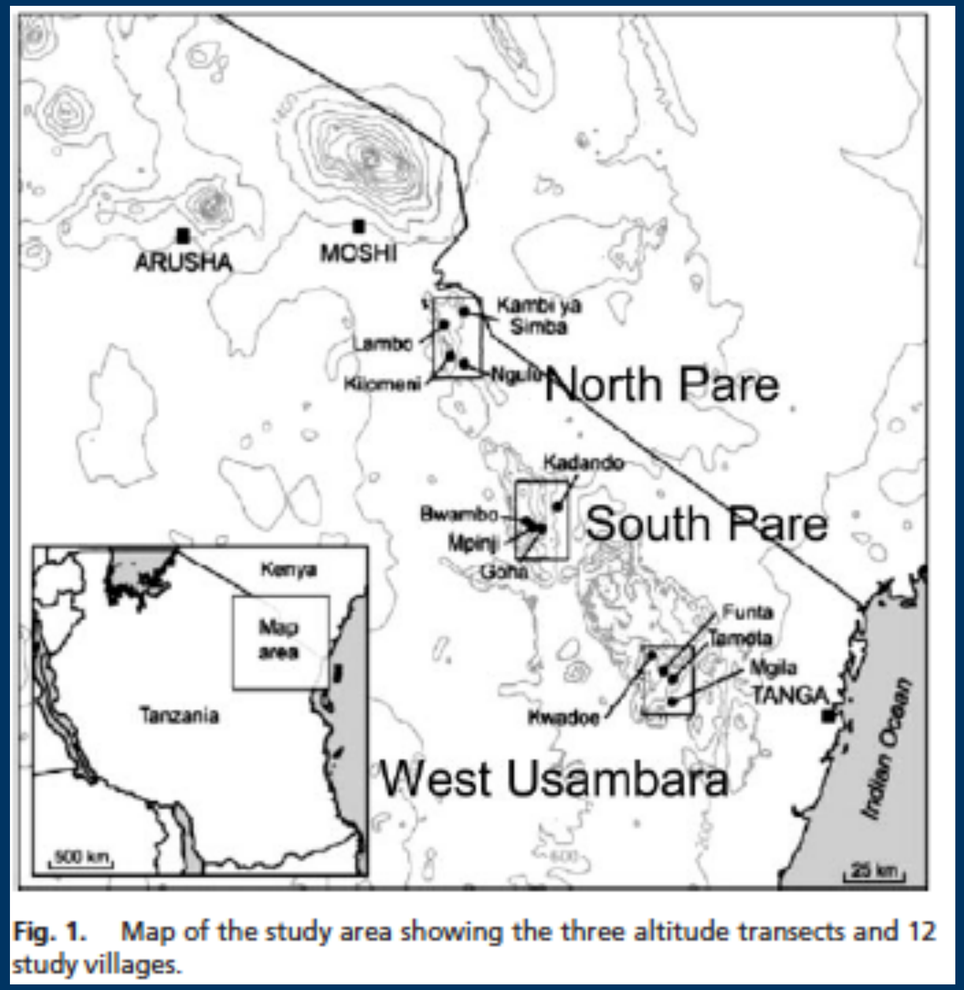


Fig. 1. Map of the study area showing the three altitude transects and 12 study villages.

Cross-sectional study

Stratified sampling (three age groups: 0-4, 5-14, 15-45

24 villages in 6 altitude transects

~8146 individuals (6 months-45 years old)

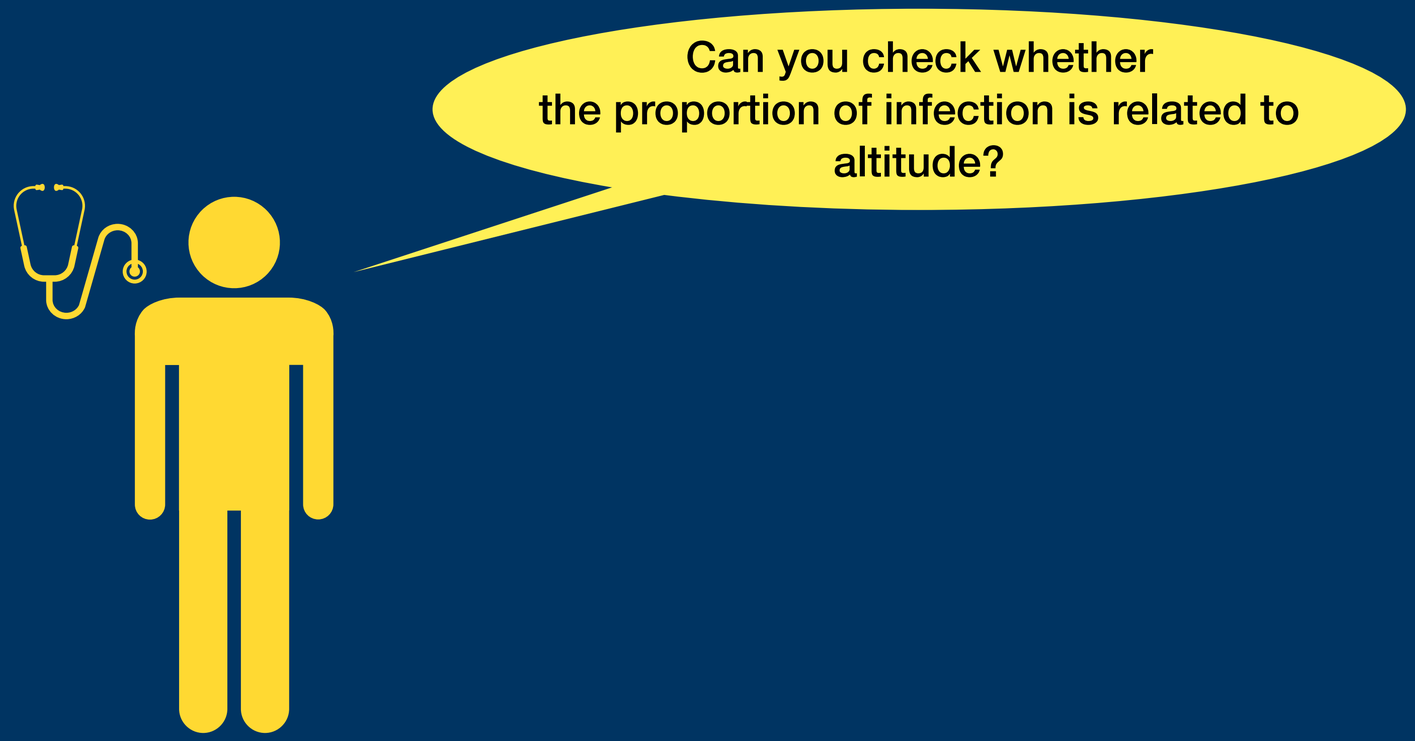
Gender and age distributions matched across villages

Github: Data/data_tanzania.csv

Warm-up example

Warm-up example

Warm-up example



Researcher