Design serological studies for influenza. Representativeness, sample size calculation

Ana Paula Rodrigues
National Institute of Health Doutor Ricardo Jorge (INSA), Portugal
Department of Epidemiology

Outline

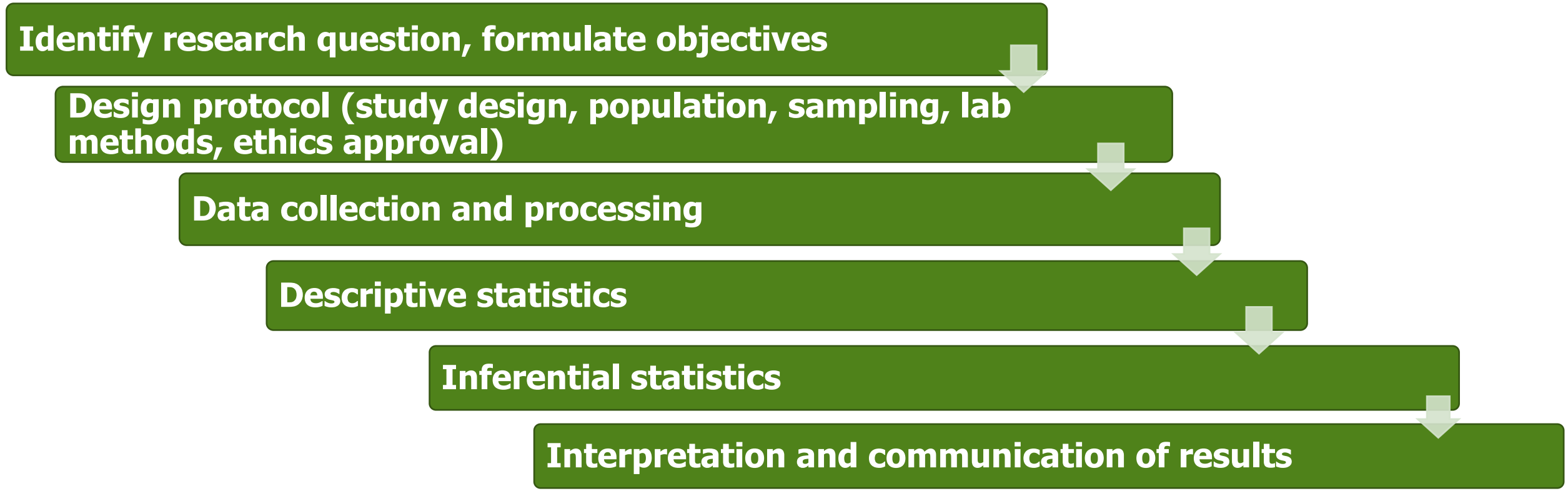
1. Study designs

2. Sampling methods

3. Sample calculation

    Prevalence study - proportion

How to design a serological survey?



Bibliography: Aday LA, Cornelius LJ. *Designing and conducting health surveys: A comprehensive guide*. 3rd ed. San Francisco: Jossey-Bass; 2006.

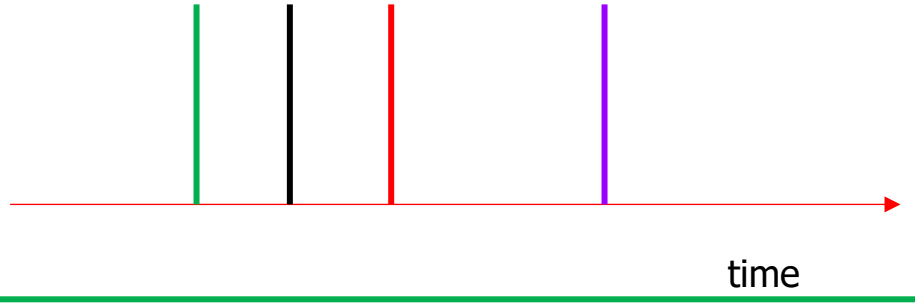
Study Designs

Cross-sectional study

(prevalence)

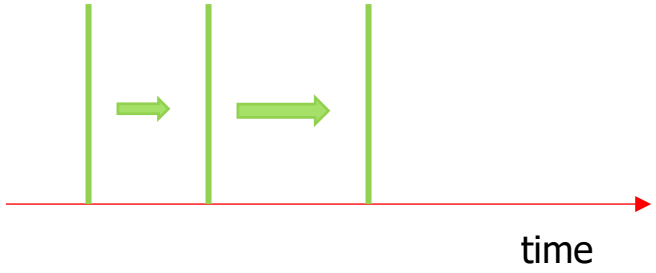
Serie of cross-sectional studies

(prevalence along time)



Longitudinal studies

(cohort, sero-incidence)



Study objective

Primary objective: To quantify the proportion of people positive for a specific antibody or measure the titre/ concentrations of an antibody.

Additional objectives: Compare groups, evaluate the effect of the intervention

How to measure outcome?

- Proportion (seroprevalence) or titre/concentrations?
- Which assay to use?

Sensitivity and specificity of the assay

- How to account for "imperfect" assay when interpreting survey results?

Target population: who, where, when?

Who?

- General population,
- Specific population subgroup (example: healthcare workers, people employed in the agricultural sector, blood donors, residents in nursing homes)

Where?

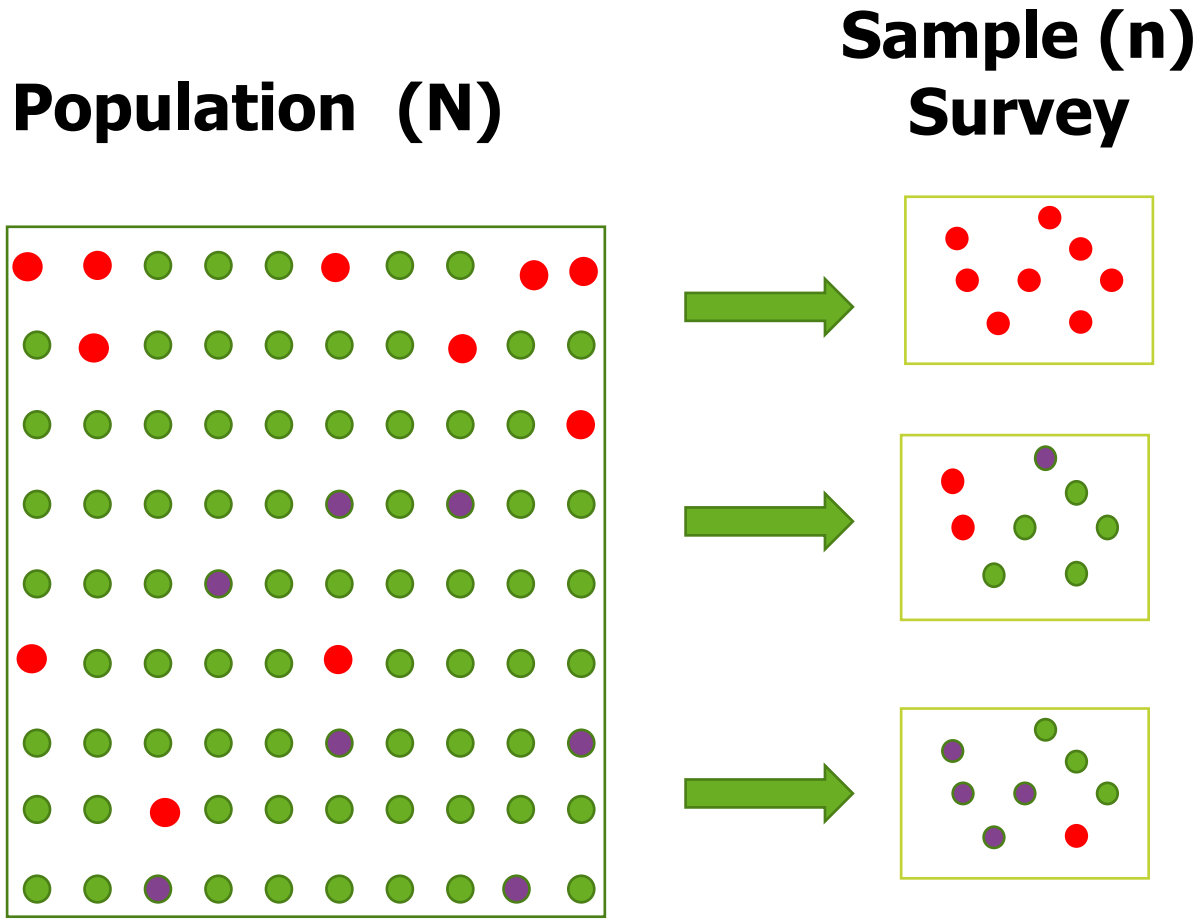
- Local,
- Regional,
- National,
- Multi-country

When?

- Seasonality, outbreak context

How to recruit participants?

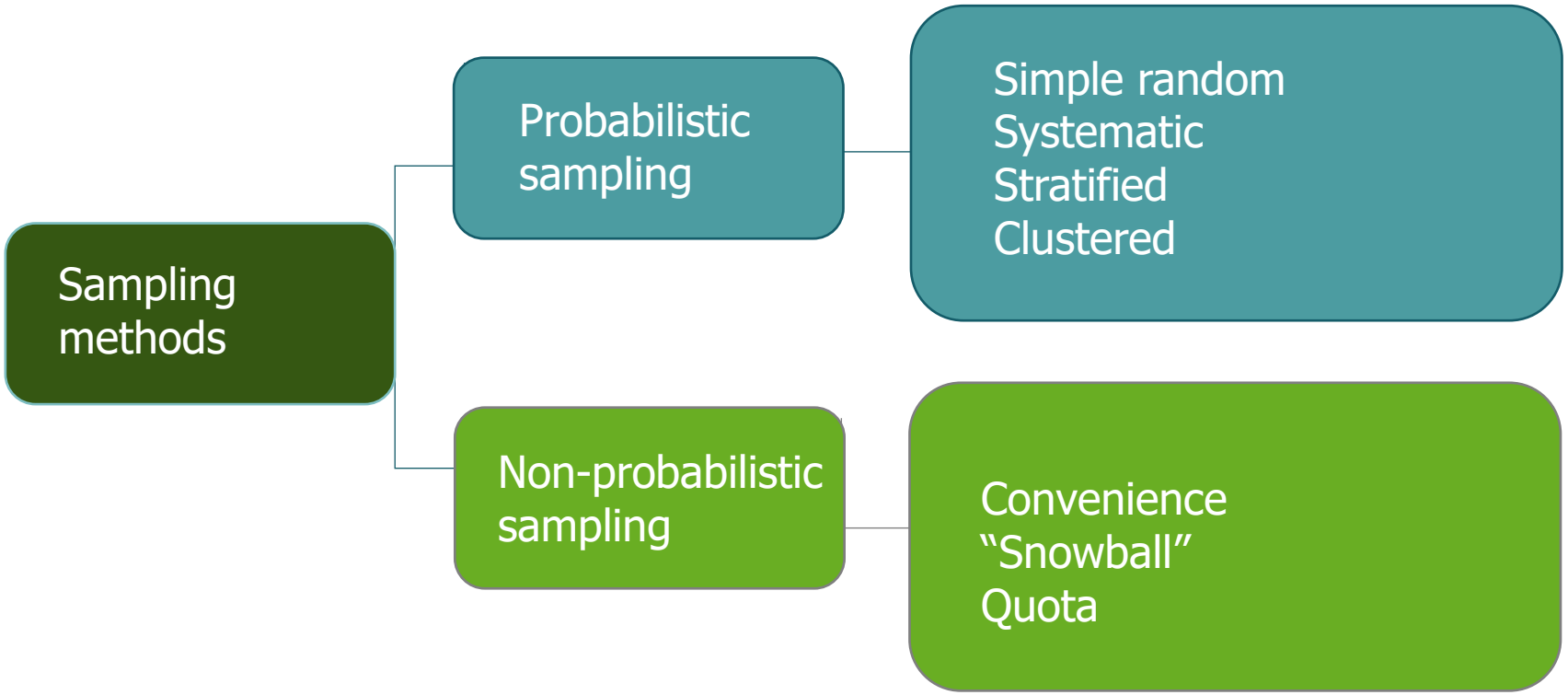
We want to learn about the population



- The sample should **resemble characteristics of the target population**
- The results estimated in that sample are **generalisable to the target population**

- **Which method** can be used to select a sample?
- **How many** units to select?
- **Where** to recruit?

Which method can be used to select a sample?

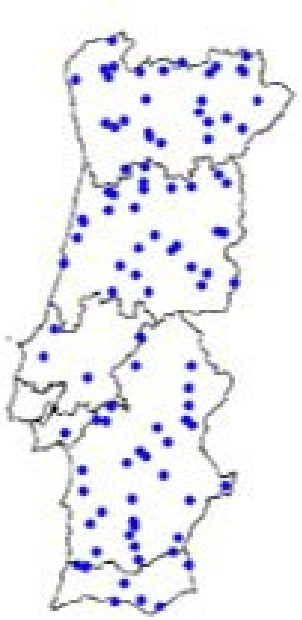


Modern trend:

Complex multi-stage sampling that combines different approaches;

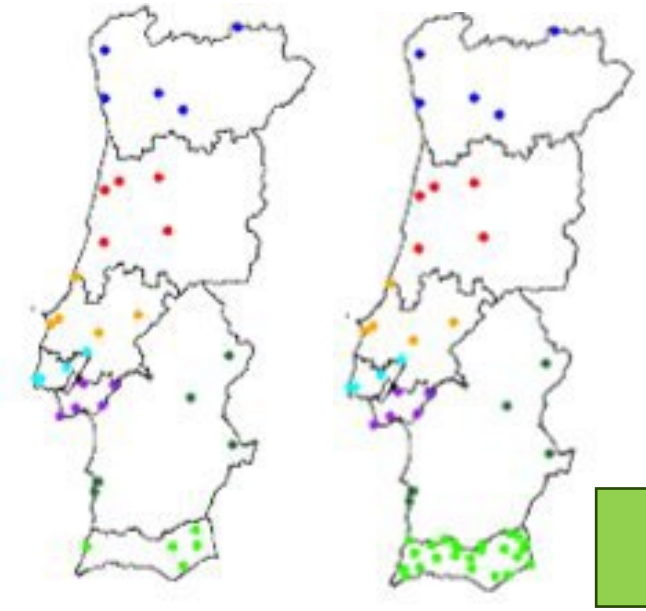
Increased use of non-probabilistic sampling strategies or a combination of probabilistic and non-probabilistic methods

Simple random sampling



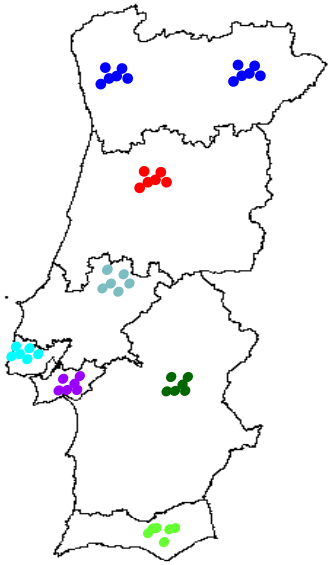
- Everyone has an equal chance to be selected.
- Most straightforward method to mimic population characteristics.
- **List of elements of the population is needed.**

Stratified random sampling



- The sample is selected independently in each strata.
- Ensures the **representation of all strata**.
- **Comparison between strata is easy**.
- Typical strata: geographical region, sex, age group.

Cluster random sampling



- Population is divided into small groups (clusters), and a **sample of clusters is selected**.
- **Reduced cost** and time.
- Overcomes the issue of a lack of sampling frames.
- Require **larger sample sizes than other** random selection methods to achive same precision of estimates.

Over sampling

Multistage stratified cluster sampling

Age-Dependent Patterns of Infection and Severity Explaining the Low Impact of 2009 Influenza A (H1N1): Evidence From Serial Serologic Surveys in the Netherlands

Anneke Steens, Sandra Waaijenborg, Peter F. M. Teunis, Johan H. J. Reimerink, Adam Meijer, Mariken van der Lubben, Marion Koopmans, Marianne A. B. van der Sande, Jacco Wallinga, and Michiel van Boven*

* Correspondence to Dr. Michiel van Boven, Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, the Netherlands (e-mail: michiel.van.boven@rivm.nl).

"Two population-based surveys were conducted using 2-stage cluster sampling (Figure 1). Of 430 municipalities in the Netherlands, **38 were randomly selected**, and an **age-stratified random sample was drawn from the municipal population registers**".

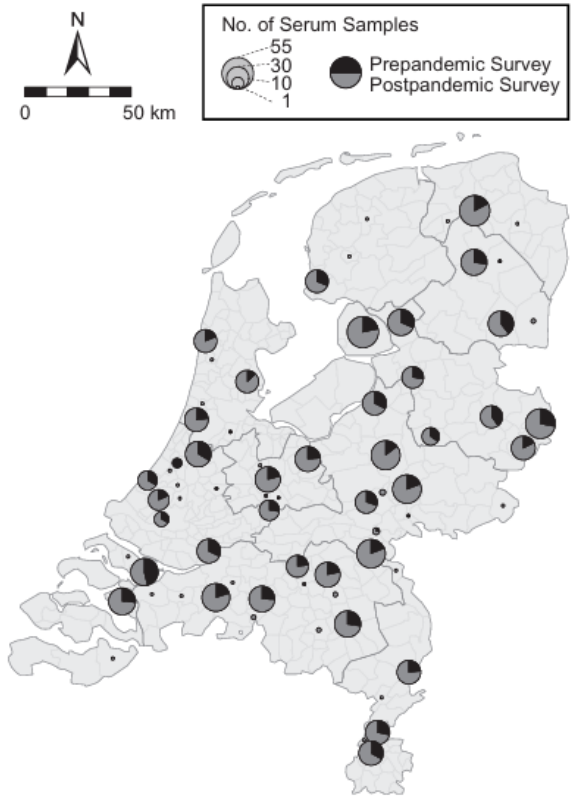
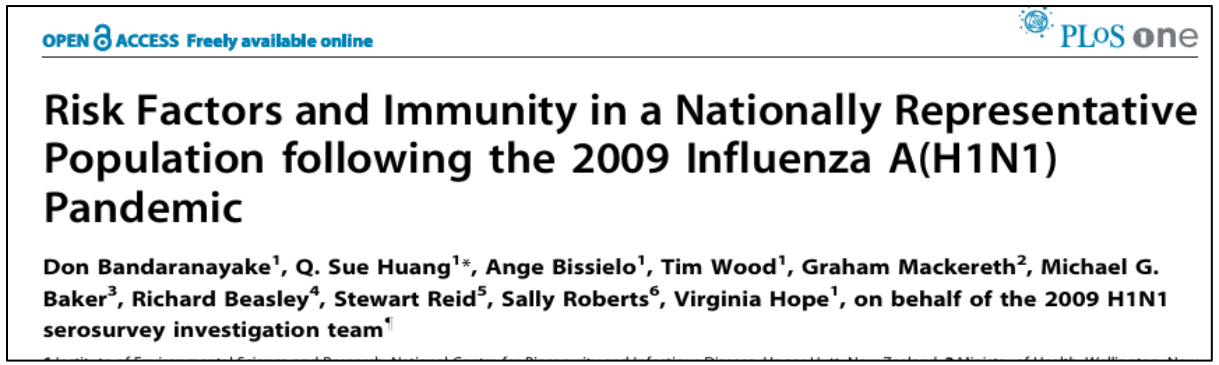


Figure 1. Geographic distribution of participants' residences in 2 influenza A (H1N1) serologic surveys, the Netherlands, 2009–2010. The sizes of the circles reflect the number of serum samples collected per municipality (dark gray: prepandemic survey (September 2009); light gray: postpandemic survey (March/April 2010)).

American Journal of Epidemiology, 2011 doi: 10.1093/aje/kwr245

Multistage stratified cluster sampling



OPEN ACCESS Freely available online

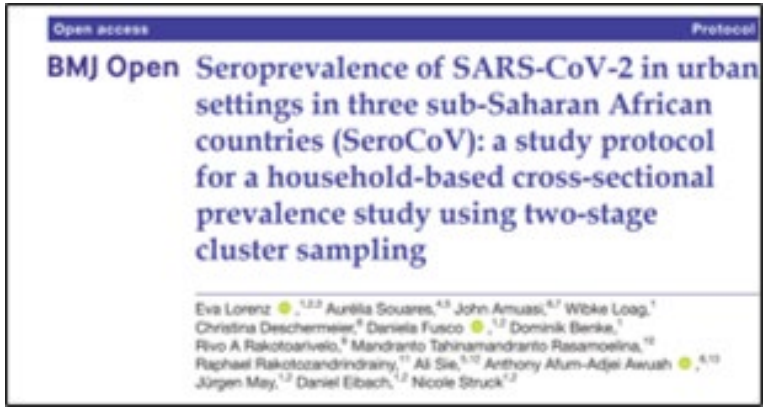
PLoS one

Risk Factors and Immunity in a Nationally Representative Population following the 2009 Influenza A(H1N1) Pandemic

Don Bandaranayake[1], Q. Sue Huang[1]*, Ange Bissielo[1], Tim Wood[1], Graham Mackereth[2], Michael G. Baker[3], Richard Beasley[4], Stewart Reid[5], Sally Roberts[6], Virginia Hope[1], on behalf of the 2009 H1N1 serosurvey investigation team[1]

"The first stage of the cross-sectional study was a purposive **cluster sample of general practices** in New Zealand, **followed by a stratified random sample of registered patients**. The study included 14 GP clinics across the country. Within each practice, registered patients were stratified by **age group and ethnicity**. Five age groups were categorised as 1 to 4, 5 to 19, 20 to 39, 40 to 59, and 60+. Ethnicity, for analytical purposes, was divided into three categories as Maori, Pacific Peoples and Other. **Within each stratum, simple random sampling was performed to select sufficient numbers of participants**".

Multistage stratified cluster sampling


Open access Protocol

BMJ Open

Seroprevalence of SARS-CoV-2 in urban settings in three sub-Saharan African countries (SeroCoV): a study protocol for a household-based cross-sectional prevalence study using two-stage cluster sampling

Eva Lorenz, Aurélia Souares, John Amuasi, Wibke Loag, Christina Deschermeier, Daniela Fusco, Dominik Benke, Rivo A Rakotoarivelo, Mandranto Tahinamandranto Rasamoelina, Raphael Rakotozandrindrainy, Ali Sie, Anthony Afum-Adjei Awuah, Jürgen May, Daniel Eibach, Nicole Struck

"In the first stage, administrative boundaries were used to allocate clusters. Shapefiles with administrative boundaries and respective population statistics were obtained from in-country or online sources.

Clusters were selected using the probability proportional to population size method. In the second stage, the sampling frame was based on geographical coordinates (geo-point sampling – any point method) within the polygon(s) where the target population lived.

Within each sampled unit, geographical **coordinates were randomly selected as households to be recruited into the study**. **Study participants will be selected in line with their country's age and gender distribution**. One person per household will be included in the study."

Multistage stratified cluster sampling

Epidemiology and Infection

cambridge.org/hyg

Original Paper

*Equal contributions.

Cite this article: Janssens H et al (2022). Pre-vaccination SARS-CoV-2 seroprevalence

Pre-vaccination SARS-CoV-2 seroprevalence among staff and residents of nursing homes in Flanders (Belgium) in fall 2020

Heidi Janssens[1,2], Stefan Heytens[2], Eline Meyers[3], Ellen De Schepper[4], An De Sutter[2], Brecht Devleesschauwer[5,6], Asangwing Formukong[3], Sara Keirse[1], Elizaveta Padalko[3,7], Tom Geens[1,*] and Piet Cools[1,*]

"The nursing homes (NH) were selected from a database of Liantis, a Belgian external occupational health service.

A subset of 210 NH employing at least ten staff members was used. The **100 NH were chosen** according to the true proportion of NH per province, resulting in 56 NH, situated in Flemish region and the Brussels capital region, and 44 NH in the Walloon region. **In every NH, n residents and 60-n staff were randomly selected using an online tool** specifically developed for this study. The **number of selected residents and staff in each NH reflected the proportion of residents and staff in that NH**".

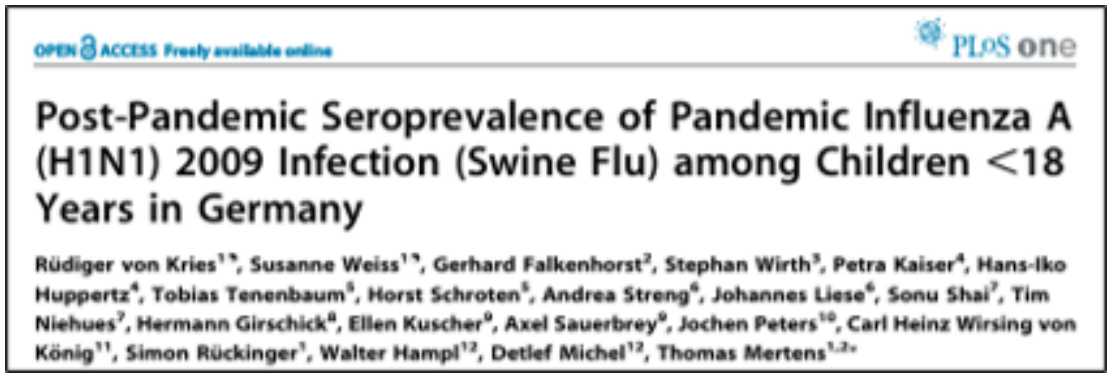
Non-probabilistic sampling methods

- The "ease" with which potential participants can be located and recruited
- Can be used for sampling rare and **difficult-to-access groups** of the population (homeless, refugees)
- **Reduce costs** and time of data collection
- Low response rates in probabilistic surveys
- Challenges in access to quality **sampling frames** (list of all units of the population)

"Price to pay"

- More **complex analysis**
- Can result in **biased** (not representative) estimates

Non-probability sampling methods



OPEN ACCESS Freely available online PLoS one

Post-Pandemic Seroprevalence of Pandemic Influenza A (H1N1) 2009 Infection (Swine Flu) among Children <18 Years in Germany

Rüdiger von Kries[1*], Susanne Weiss[1*], Gerhard Falkenhorst[2], Stephan Wirth[3], Petra Kaiser[4], Hans-Iko Huppertz[4], Tobias Tenenbaum[5], Horst Schroten[5], Andrea Streng[6], Johannes Liese[6], Sonu Shai[7], Tim Niehues[7], Hermann Girschick[8], Ellen Kuscher[9], Axel Sauerbrey[9], Jochen Peters[10], Carl Heinz Wirsing von König[11], Simon Rückinger[1], Walter Hampl[12], Detlef Michel[12], Thomas Mertens[1,2*]

"Eight German **paediatric primary care hospitals** (Bremen, Berlin, Krefeld, Wuppertal, Erfurt, Wurzburg, Mannheim, Munich) **provided sera from April 1st to July 31st 2010, obtained during blood withdrawal for routine laboratory testing** from in- or outpatients aged 1 to 17 years. Children with an illness impeding an adequate immune response were excluded, so were children with serious conditions".

Non-probability sampling methods

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

ELSEVIER

Cross-protection to new drifted influenza A(H3) viruses and prevalence of protective antibodies to seasonal influenza, during 2014 in Portugal

Raquel Guiomar [a,*], Susana Pereira da Silva [a], Patrícia Conde [a], Paula Cristóvão [a], Ana Carina Maia [a], Pedro Pechirra [a], Ana Paula Rodrigues [a], Baltazar Nunes [a], Portuguese Laboratory Network for the Diagnosis of Influenza Infection Luís Milho [b], Ana Paula Coelho [b], Aida Fernandes [b], Paula Caseiro [c], Fernando Rodrigues [c], Lurdes Correia [c], João Pereira-Vaz [c], Sofia Almeida [d], Paula Branquinho [e], Rita Côrte-Real [e], Regina Viseu [f], Maria João Peres [f], Raquel Sanches [g], Filipa Dantas [g], Ludovina Freitas [g], Graça Andrade [g], Manuel Maurílio [h], Filomena Caldeira [h], Rita Cabral Veloso [i], Luisa Mota-Vieira [i], Marta Soares [j], Ana Rita Couto [j], Jácome Bruges-Armas [j], Rita Mouro Pinto [k], Joana Sobrinho Simões [k], Maria do Rosário Costa [k], João Tiago Guimarães [k], Luís Martins [l], Mário Cunha [l]

"To study influenza immunity in the Portuguese population, a non-probabilistic sample was used. Samples were collected from **people attending hospital laboratories for other reasons aside from influenza infection**. A convenience sample of 626 sera was collected during June 2014. Sera were selected from **all age groups (0–4; 5–14; 15–64 and ≥65 years old) and both genders, in equal proportion, at 11 hospital laboratories** from the Portuguese Laboratory Network for the Diagnosis of Influenza Infection, covering all administrative health regions (HR) of Portugal: Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve and including also the Açores (São Miguel and Terceira islands) and Madeira islands".

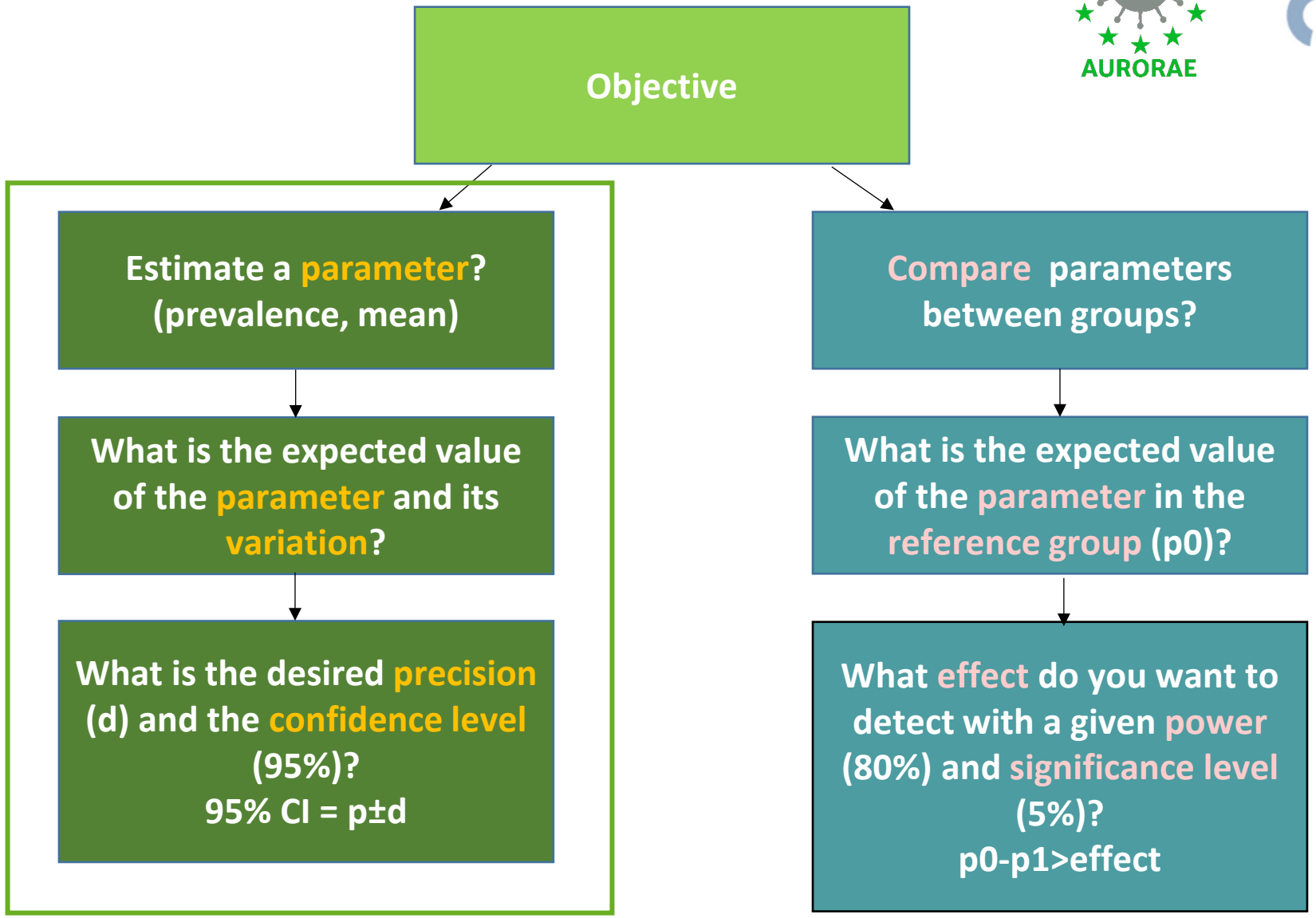
Objective

Estimate a parameter? (prevalence, mean)

- To determine the prevalence of seasonal influenza protective antibody titres for A(H3) after the 2025/26 influenza season in Portugal
- To estimate the mean levels of SARS-CoV-2 IgG antibody titers among previously infected unvaccinated individuals

Compare parameters between groups?

- To compare the seroprevalence of protective antibodies against influenza A/Texas/50/2012 among 5–14 years old and 65+ years old
- To compare the seroprevalence of SARS-CoV-2 in Portugal at different time points



What is the expected value of the parameter and its variability in the population?

Make an educated guess about the expected value of prevalence, mean, and standard deviation to determine the sample size.

How this could be achieved:

- Use of data from previous studies of the same or similar populations;
- Develop a pilot study;
- If unknown, use assumptions that maximize the sample size (50% for prevalence).

What is the desired precision (d)?

d reflects the width of confidence interval (CI).

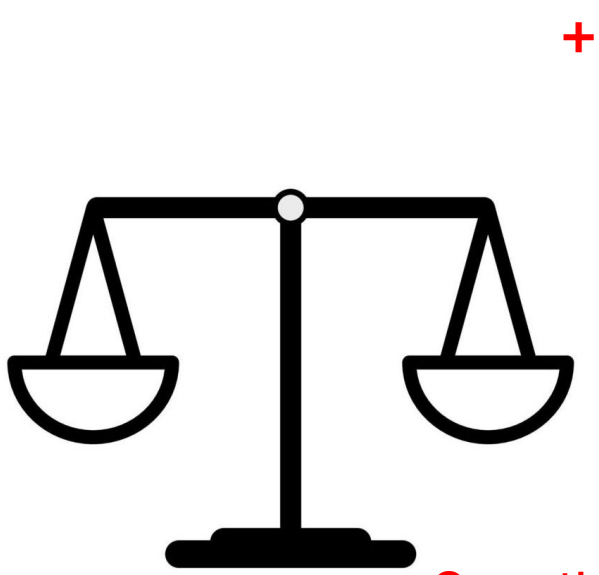
What is the desired confidence interval (CI)?

Usually we use 95 % CI

Summary of sample sizes for different expected seroprevalences and precisions

Expected prevalence	Precision (d)			
	10 %	5 %	2.5 %	1 %
50 %	97	384	1,535	9,513
25 %	73	289	1,152	7,152
10 %	35	139	553	3,446
5 %	-	73	292	1,822

Considering an design effect of 1. For complex samples consider an design effect ≥ 1.5



OpenEpi - Toolkit Shell for Developing New Applications

Expand All | Collapse

- Home
- Info and Help
- Language/Options/Settings
- Calculator
- Counts
  - Std.Mort.Ratio
  - Proportion
  - Two by Two Table
  - Dose-Response
  - R by C Table
  - Matched Case Control
  - Screening
- Person Time
  - 1 Rate
  - Compare 2 Rates
- Continuous Variables
  - Mean CI
  - Median/%ile CI
  - t test
  - ANOVA
- Sample Size
  - Proportion
  - Unmatched CC
  - Cohort/RCT
  - Mean Difference
- Power
- Random numbers
- Searches

Open Source Epidemiologic Statistics for Public Health

Now in English, French, Spanish, Italian, and Portuguese

Version 3.01 Updated 2013/04/06 ***Try it in a Smartphone browser!***

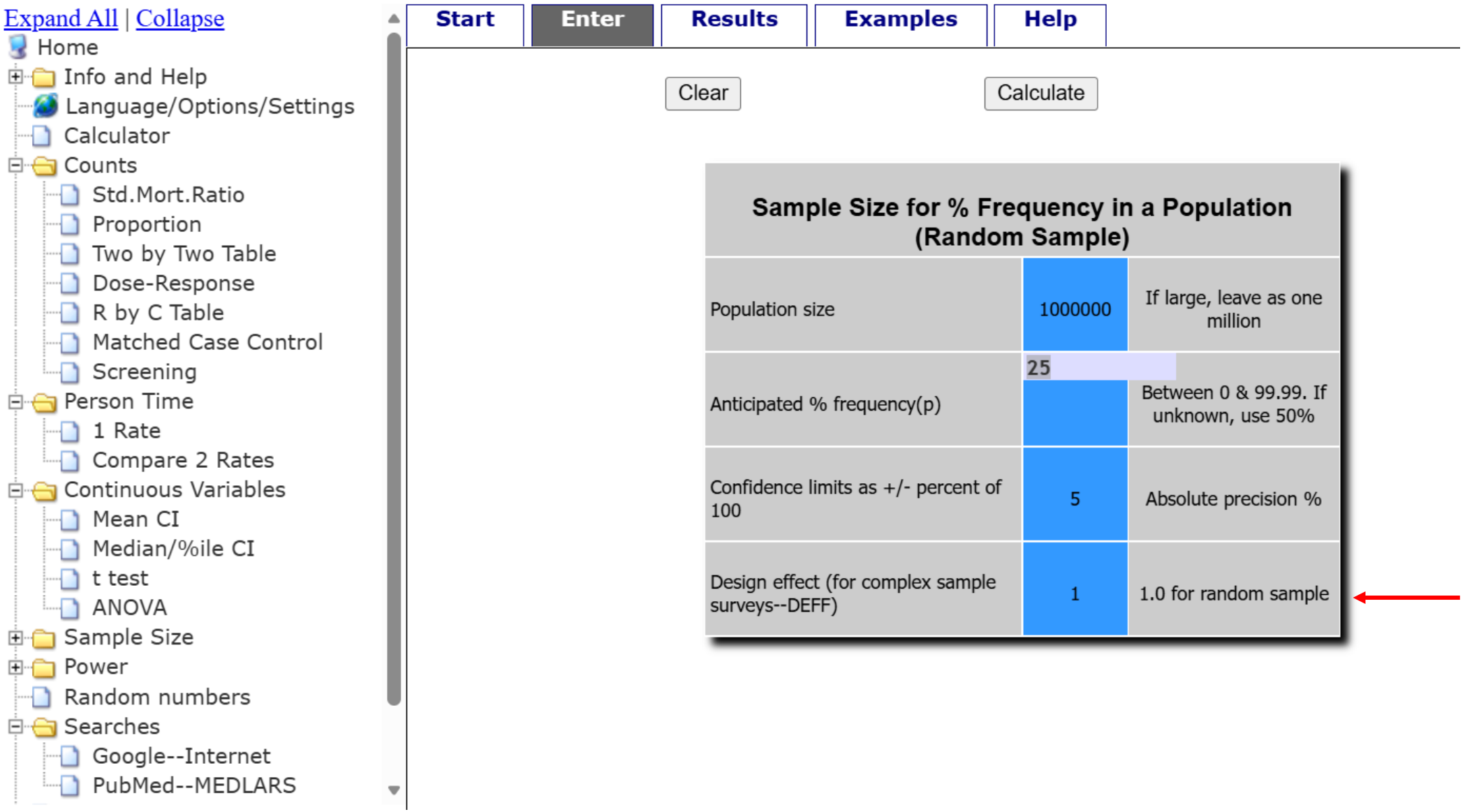
OpenEpi provides statistics for counts and measurements in descriptive and analytic studies, stratified analysis with exact confidence limits, matched pair and person-time analysis, sample size and power calculations, random numbers, sensitivity, specificity and other evaluation statistics, R x C tables, chi-square for dose-response, and links to other useful sites.

OpenEpi is free and **open source** software for epidemiologic statistics. It can be run from a web server or downloaded and run without a web connection. A server is not required. The programs are written in JavaScript and HTML, and should be compatible with recent Linux, Mac, and PC browsers, regardless of operating system. (If you are seeing this, your browser settings are allowing JavaScript.) The programs can be run in the browsers of many iPhone and Android cellphones

Test results are provided for each module so that you can judge reliability, although it is always a good idea to check important results with software from more than one source. Links to hundreds of Internet calculators are provided.

The programs have an open source license and can be downloaded, distributed, or translated. Some of the components from other sources have licensing statements in the source code files. Licenses referred to are available in full text at OpenSource.org/licenses.

OpenEpi development was supported in part by a grant from the Bill and Melinda Gates Foundation to Emory University, Rollins School of Public Health



≥1.5 for complex samples

Expand All | Collapse

- Home
- Info and Help
- Language/Options/Settings
- Calculator
- Counts
  - Std.Mort.Ratio
  - Proportion
  - Two by Two Table
  - Dose-Response
  - R by C Table
  - Matched Case Control
  - Screening
- Person Time
  - 1 Rate
  - Compare 2 Rates
- Continuous Variables
  - Mean CI
  - Median/%ile CI
  - t test
  - ANOVA
- Sample Size
- Power
- Random numbers
- Searches
  - Google--Internet
  - PubMed--MEDLARS
- Internet Links

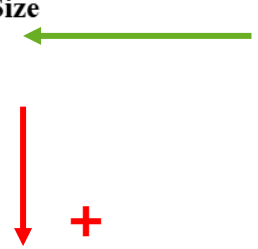
Start | Enter | **Results** | Examples | Help

Sample Size for Frequency in a Population

Population size(for finite population correction factor or fpc)(*N*): 1000000
Hypothesized % frequency of outcome factor in the population (*p*): 25%+/-5
Confidence limits as % of 100(absolute +/- %)(*d*): 5%
Design effect (for cluster surveys-*DEFF*): 1

Sample Size(*n*) for Various Confidence Levels

ConfidenceLevel(%)	Sample Size
95%	289
80%	124
90%	203
97%	354
99%	498
99.9%	812
99.99%	1135



Equation

Sample size $n = [DEFF*Np(1-p)]/ [(d^2/Z^2_{1-\alpha/2}*(N-1)+p*(1-p)]$

Results from OpenEpi, Version 3, open source calculator--SSPropor
Print from the browser with ctrl-P
or select text to copy and paste to other programs.

Almost never we get **100 %** participation **– how to account for this?**

Prevalence = 25 %, d=± 2.5 % CI 95%: 22.5 %-27.5 % **Required sample size**: n= 1,152

+

Correction factor: 10 % (how many participants we expect to lose)

Previous knowledge,
Context dependent

Number to invite $= \frac{1{,}152}{(1-0.1)}$ **= 1,280**

Messages to take

- The study design depends of the research question
- Probabilistic sampling methods are less prone to be biased, but more difficult to implement
- Non-probabilistic sampling methods are more prone to be biased and the statistical analysis is more complex, but are easier to implement and to get access to difficult-to-reach populations
- Sample size depends of the expected value, precision of the estimate and the level of confidence. Complex samples (clusters) need higher sample size

Bibliography

- Laurie KL, Huston P, Riley S, Katz JM, Willison DJ, Tam JS, Mounts AW, Hoschler K, Miller E, Vandemaele K, Broberg E, Van Kerkhove MD, Nicoll A. Influenza serological studies to inform public health action: best practices to optimise timing, quality and reporting. Influenza Other Respir Viruses. 2013 Mar;7(2):211-24. doi: 10.1111/j.1750-2659.2012.0370a.x.
- Metcalf CJE, Farrar J, Cutts FT, et al. Use of serological surveys to generate key insights into the changing global landscape of infectious disease. Lancet 2016; published online April 5. http://dx.doi.org/10.1016/S0140-6736(16)30164-7
- Dean NE, Howard DH, Lopman BA. Serological Studies and the Value of Information. Am J Public Health. 2023 May;113(5):517-519. doi: 10.2105/AJPH.2023.307245.
- Kennedy-Shaffer L, Qiu X, Hanage WP. Snowball Sampling Study Design for Serosurveys Early in Disease Outbreaks. Am J Epidemiol. 2021 Sep 1;190(9):1918-1927. doi: 10.1093/aje/kwab098.
- European Centre for Disease Prevention and Control. Sample size guidance for surveillance data. Stockholm: ECDC; 2023
- Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version. www.OpenEpi.com
- Aday LA, Cornelius LJ. *Designing and conducting health surveys: A comprehensive guide*. 3rd ed. San Francisco: Jossey-Bass; 2006

Slido questions

When calculating sample size for a seroprevalence study, which parameter has the greatest impact?