The Online R-FETPV 1st Module : Basic Epidemiology and Surveillance Data Analysis

5 April -28 May 2021

USAID FROM THE AMERICAN PEOPLE

Department of Livestock Development

Food and Agriculture Organization of the United Nations

R-FETPV Regional Field Epidemiology Training Program for Veterinarian

Sampling and sample size for surveillance and investigation data

Suwicha Kasemsuwan
Kasetsart University

USAID FROM THE AMERICAN PEOPLE

Department of Livestock Development

Food and Agriculture Organization of the United Nations

R-FETPV Regional Field Epidemiology Training Program for Vietnam

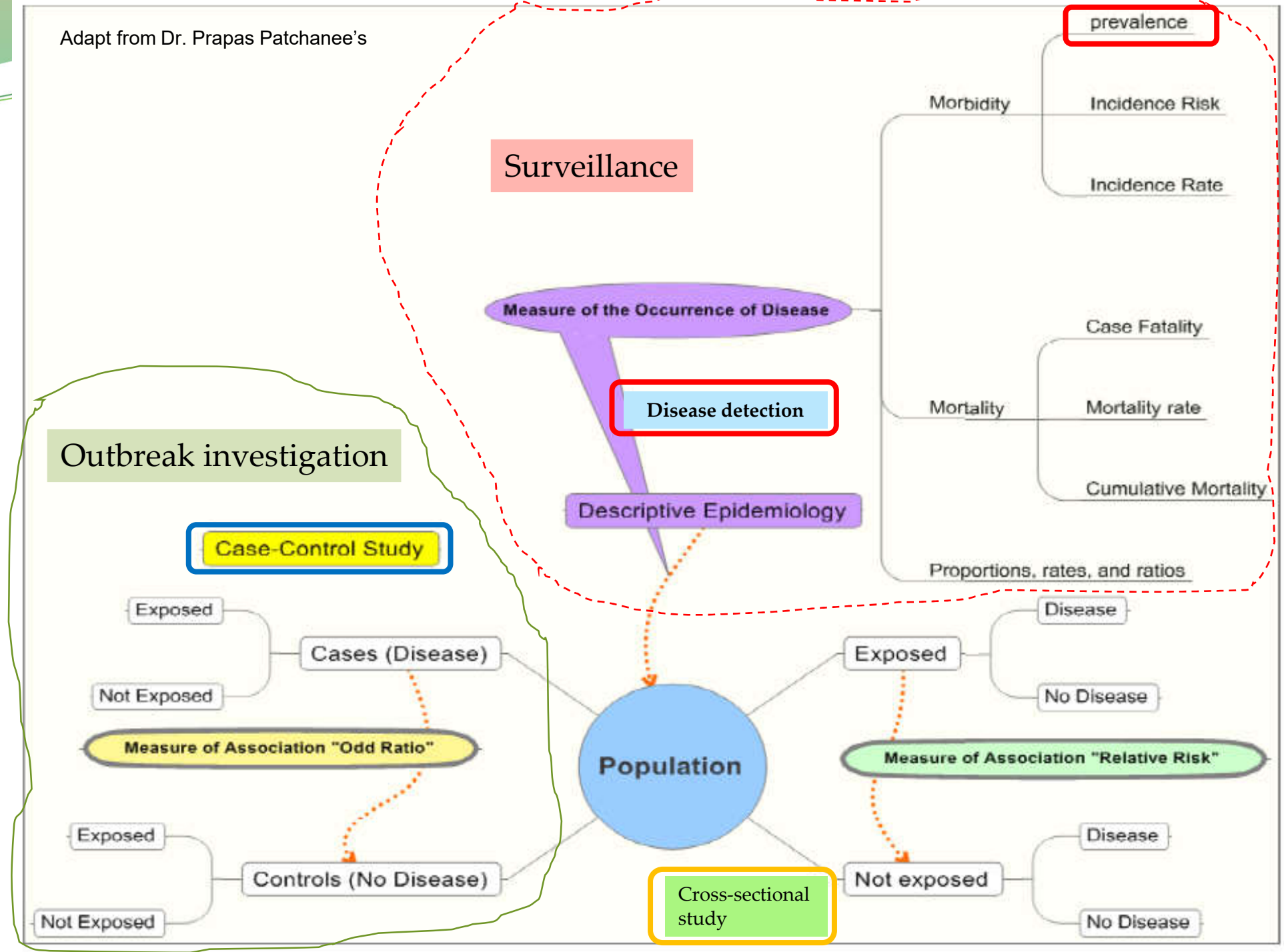
Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence determination
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

References

- Survey Toolbox manual
- ProMESA help file
- EpiTools http://epitools.ausvet.com.au/
- Openepi.com

Adapt from Dr. Prapas Patchanee's

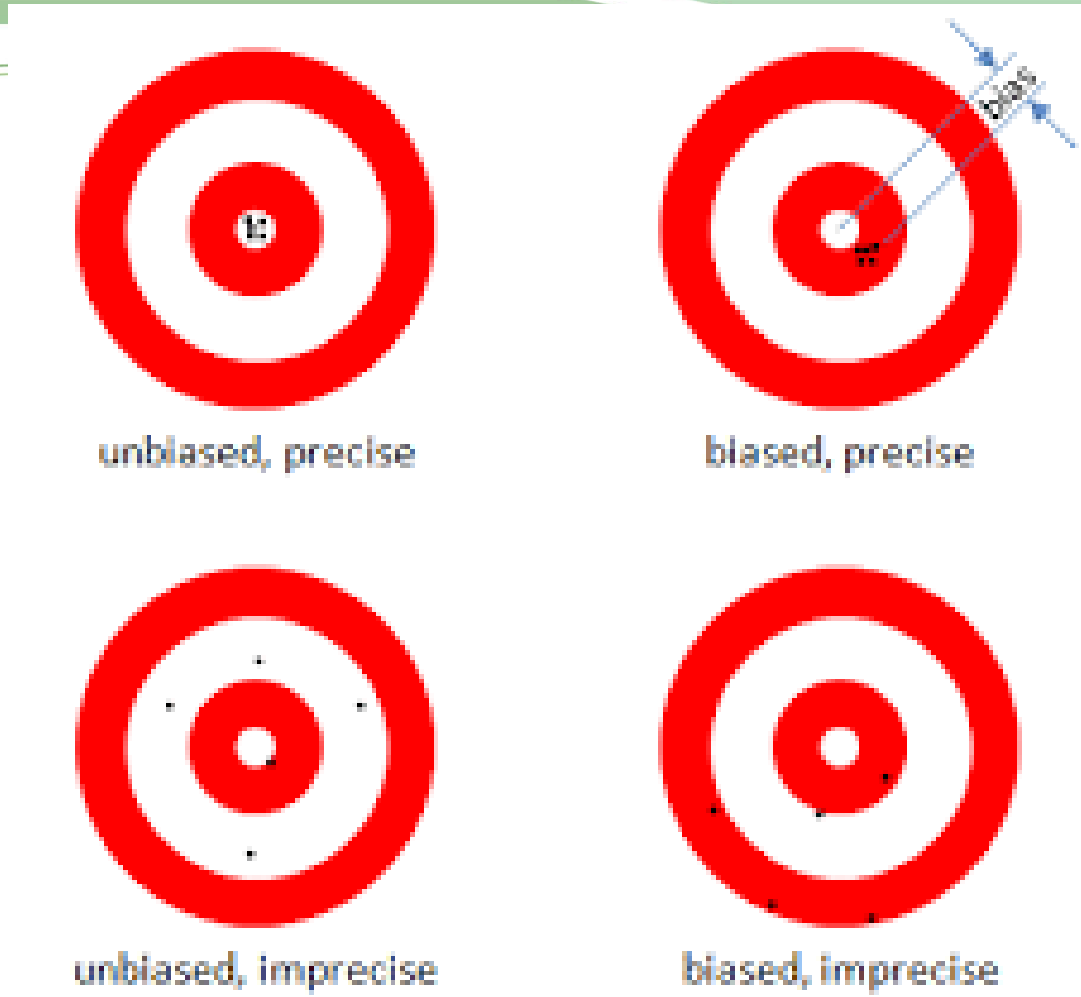


Reason

- Why do you take samples?
  - Time
  - Cost
  - Human resource
  - Harmful
  - Etc.

Samples must represent population

- Sampling – main methods
  - Simple random sampling
  - Systematic sampling



Bias: not representative samples

Imprecise: not enough samples

http://cfile10.uf.tistory.com/image/2013F650512207270E31A5

Simple random sampling

- Random sampling in a source population
- Every individual has the **same probability** of being chosen
- Representative +++++
- Need a sampling frame (population census)

- *Ex: random sampling of villages based on their names / ZIP codes*

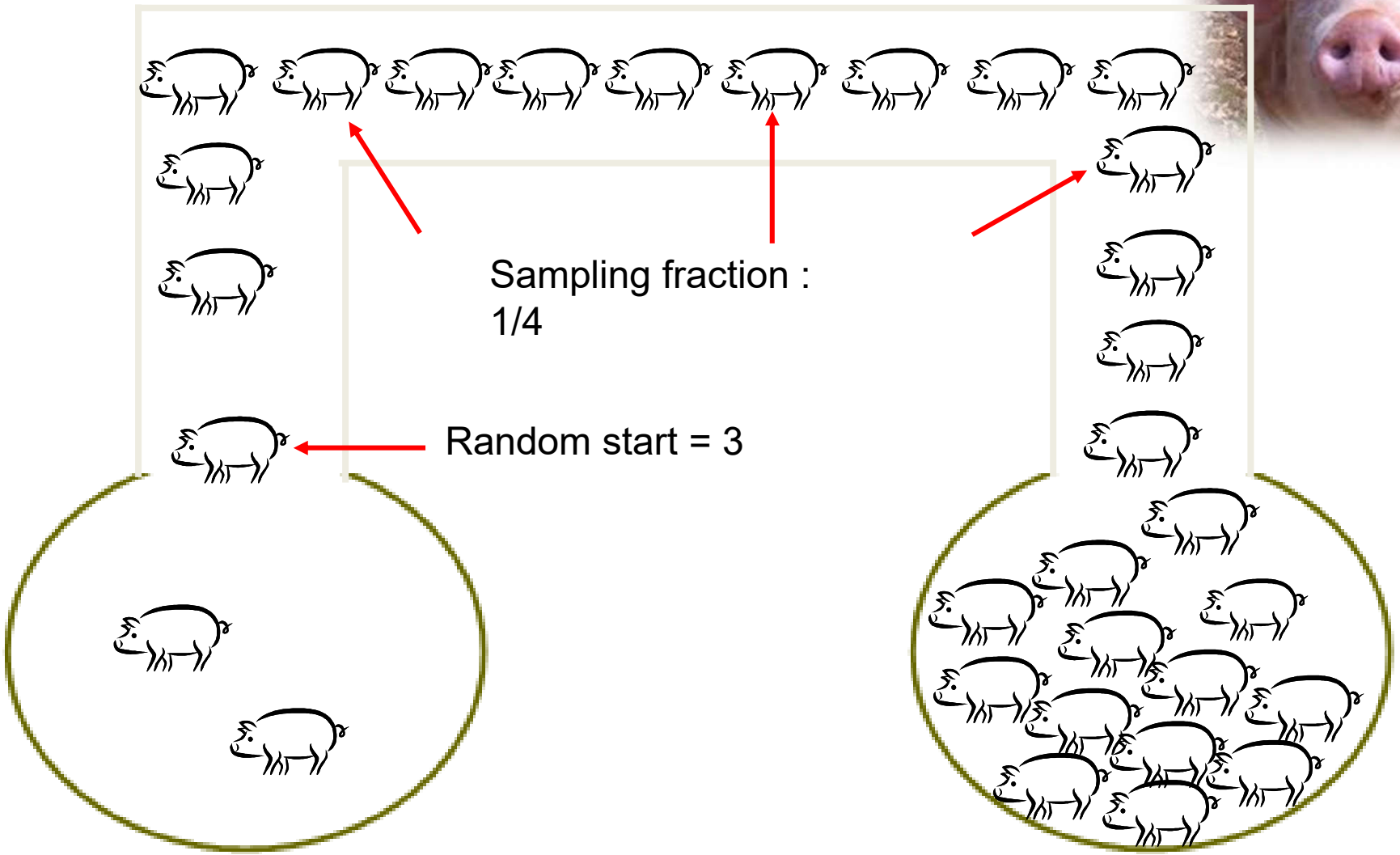
Systematic sampling

- Selecting elements according to an established rule
- Representative +++ (sample is disseminated in the population)
- Does not need a sampling frame

- *Ex: define a sampling fraction (x%), then random sampling of a figure k between 1 and n*

  ➔ *Sampling of animals k; k+x; k+2x…*

2. Systematic sampling



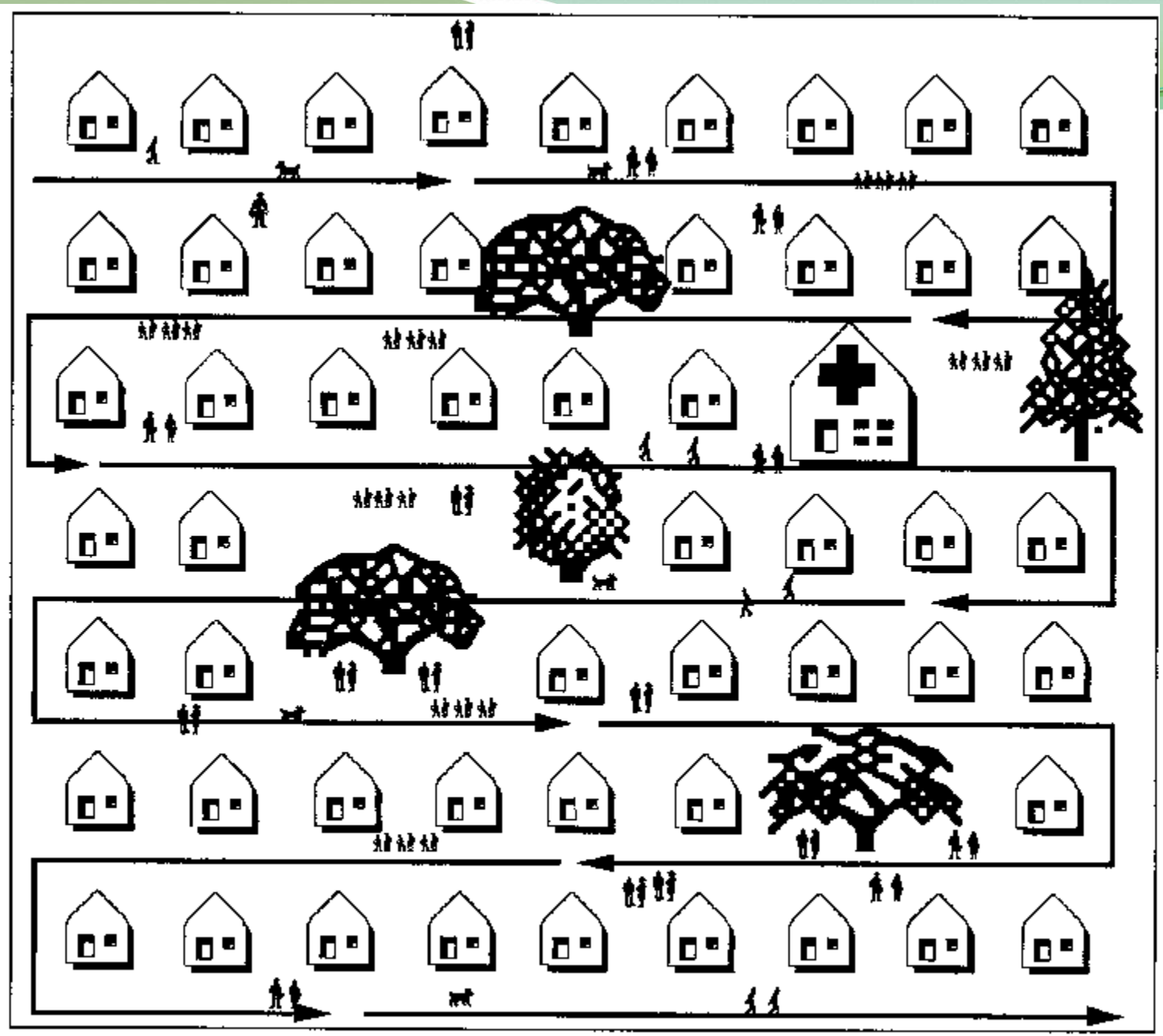
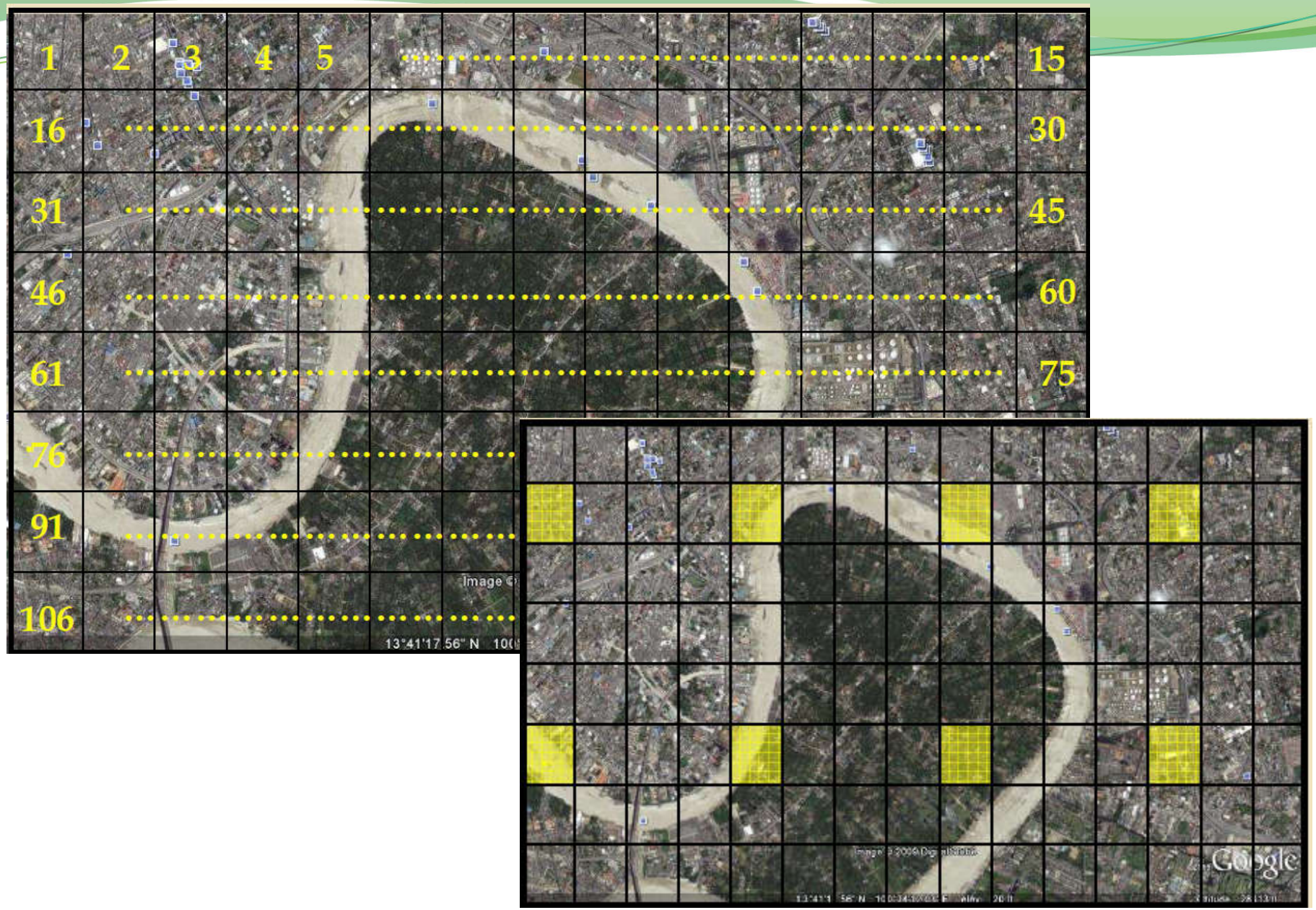


Figure 11 : *Systematic sampling, example of drawing 1 household every 7 households, starting with household N°4.*


1
2
3
4
5
15
16
30
31
45
46
60
61
75
76
91
106
Image ©
13°41'17.56" N
Google

How many samples?

- Objective of the study
  - Descriptive Epidemiology
    - Disease detection / freedom from a disease
    - Level of disease determination (prevalence determination)
  - Analytic Epidemiology
    - Observational study

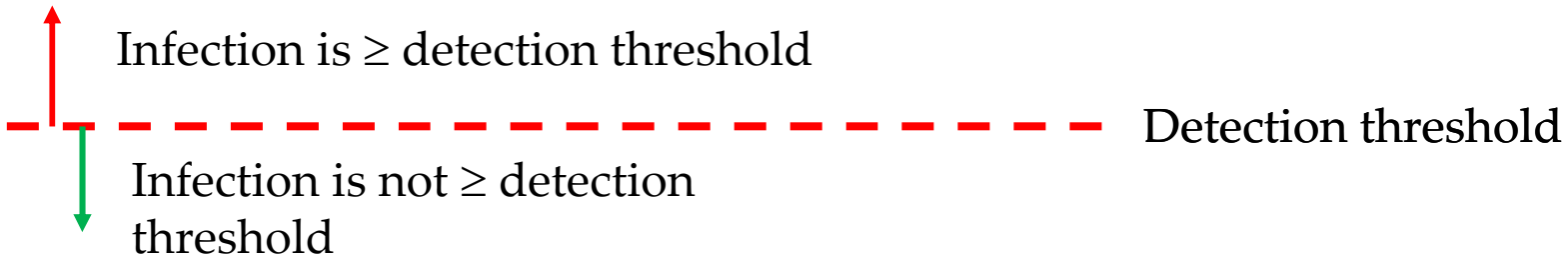
Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence determination
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Sample size:
Disease detection / freedom from infection

Objective:

- Detect disease circulation:
  - Identify infected … regions, villages, herds, …
- Need to set a detection threshold (**design prevalence**) :
  - Above, the herd/region is considered infected
  - Below : herd/region is considered free from disease



- ≥ 1 animal infected → whole unit infected
- No sample positive → whole unit considered free from the disease

- Sample size to detect disease

$$n = \frac{\ln(\alpha)}{\ln(1-p)}$$

  - p : design prevalence;
  - α: *probability of not detecting a disease = probability of being wrong when declaring the population free from a disease* $= (1-p)^n$
- Validity
  - Large population (n/N < 10%)
  - Diagnostic tests considered perfect.

- Sample size to detect disease (or proven of disease freedom)

$$n = \frac{\ln(\alpha)}{\ln(1 - \quad)}$$

Hypothesis : **infinite population**

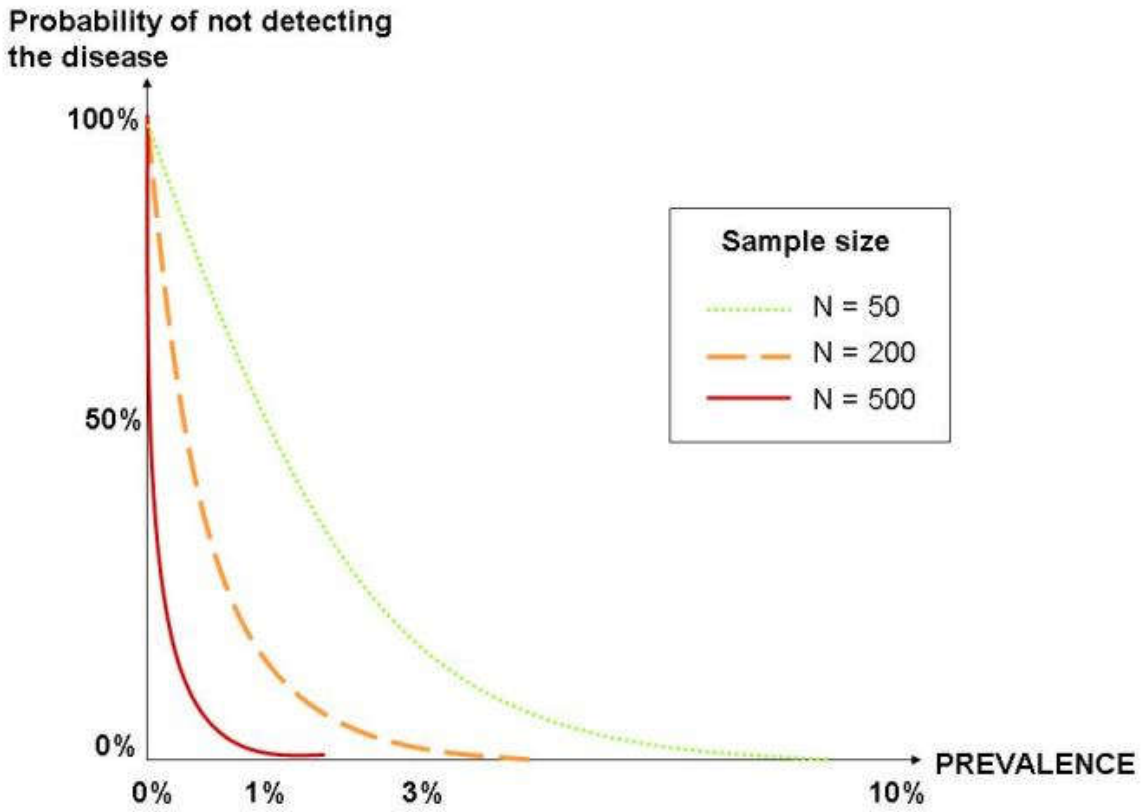
If n/N < 10% ➔ infinite population ➔ draw <u>without</u> replacement

If n/N > 10% ➔ finite population ➔ draw <u>with</u> replacement

  - *p*
  - *α* *ring the* *p*

- Validity
  - Large population (n/N < 10%)
  - Diagnostic tests considered perfect.

Probability of disease detection



When **p** ↑ **,** it is easier to detect disease

When **n** ↑ , it is also easier to detect disease

Table 2. Number needed for study to be confident that the disease will be detected if present at or above a specified prevalence based on hypergeometric sampling and assuming a perfect test.

Population size	Expected prevalence of disease in population (95%/99% confidence)				
	0.10	0.05	0.02	0.01	0.005
20	15/18	19/20	20/20	20/20	20/20
50	22/29	34/41	48/50	50/50	50/50
100	25/35	44/59	77/90	95/99	100/100
150	26/38	48/67	94/117	129/143	150/150
200	27/39	51/72	105/136	155/180	190/198
250	27/40	52/75	112/149	174/210	238/244
300	27/41	53/77	117/159	189/235	233/286
350	27/41	54/79	121/167	201/255	272/324
400	27/41	54/80	124/174	210/272	310/360
450	28/42	55/81	126/179	218/287	349/391
500	28/42	55/82	128/183	224/300	388/420
600	28/42	56/83	131/189	235/320	378/470
700	28/42	56/84	134/194	243/336	369/511
800	28/43	56/85	135/198	249/349	421/546
1,000	28/43	57/86	138/204	258/367	450/601
1,200	28/43	57/86	139/208	264/381	471/642
1,400	28/43	57/87	141/210	268/391	486/673
1,600	28/43	57/87	142/212	272/398	500/699
1,800	28/43	57/88	142/214	275/404	509/719
2,000	28/43	58/88	143/215	278/409	517/736
3,000	28/43	58/88	145/219	284/425	542/791
4,000	28/44	58/89	146/222	287/433	555/821
5,000	28/44	58/89	146/223	289/438	563/839
6,000	28/44	58/89	146/224	291/441	569/852
10,000	28/44	58/89	147/225	294/443	580/888
100,000	28/44	58/90	148/228	298/457	596/915
>100,000*	28/44	58/90	148/228	298/458	598/919

* Based on binomial model.

Online tools : EpiTools

http://epitools.ausvet.com.au/content.php?page=FreeCalc2

FreeCalc: Calculate sample size for freedom testing with imperfect tests

Input Values

Population Size:

Test Sensitivity:

Test Specificity:

Design prevalence:

Design prevalence (proportion or number of units):

Analysis options:

Desired type I error (1 - minimum population-sensitivity): 0.05

Desired type II error (1 - minimum population-specificity): 0.05

Calculation method:
(these settings can usually be left as default values)

- (•) Modified hypergeometric exact
- ( ) Simple binomial (large population)

Population threshold for binomial method: 10000

Maximum limit for sample size: 3200

Precision (significant digits): 4

Submit Reset

Calculate the required sample size and cut-point for testing to demonstrate population freedom from diseas

This utility uses the methods described by:
Cameron and Baldock (1998): A new probability formula for surveys to substantiate freedom from disease. F
Cameron (1999): *Survey Toolbox for Livestock Diseases - A practical manual and software package for act*
Canberra, Australia.

Inputs include:

- Size of the population sampled;
- Test sensitivity and specificity;
- Design prevalence (the hypothetical prevalence to be detected). Design prevalence can be specified
- Maximum acceptable Type I (1 - population-sensitivity) and Type II (1 - population-specificity) error
  popultion is diseased;
- Calculation method: hypergeometric (for small populations), or simple binomial (for large populations);
- The population size threshold, above which the simple binomial method is used regardless of which ca
- The maximum upper limit for required sample size; and
- The desired precision of results (number of digits to be displayed after the decimal point).

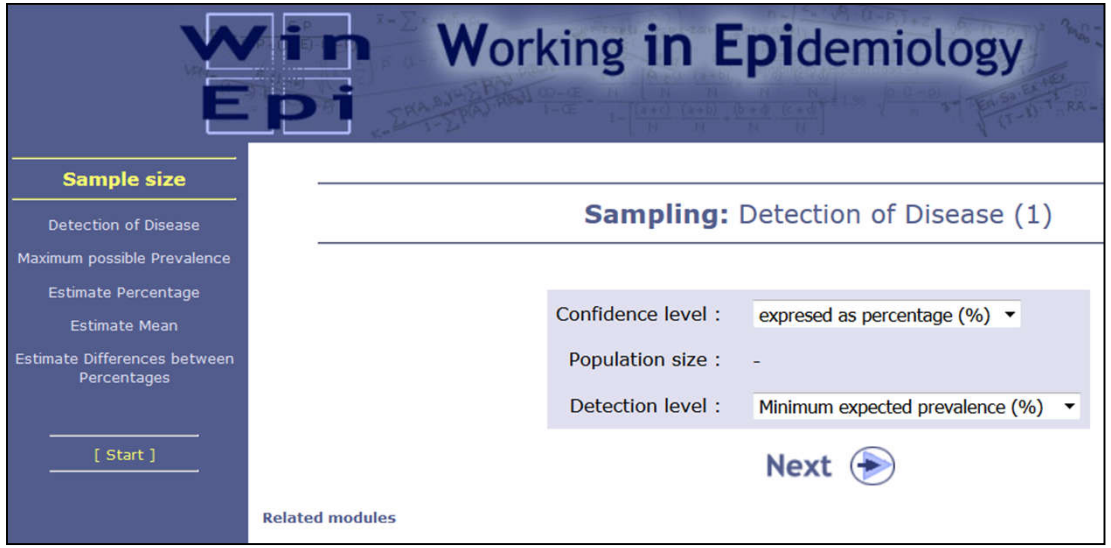
The results are presented as:

- The minimum sample size and corresponding cut-point number of positives to achieve the specified ty
- achieved Type I and Type II error levels and corresponding population-level sensitivities and specificit
- A descriptive interpretation of the results; and
- an error message if the desired error levels cannot be achieved within the limits of populatuon and/or

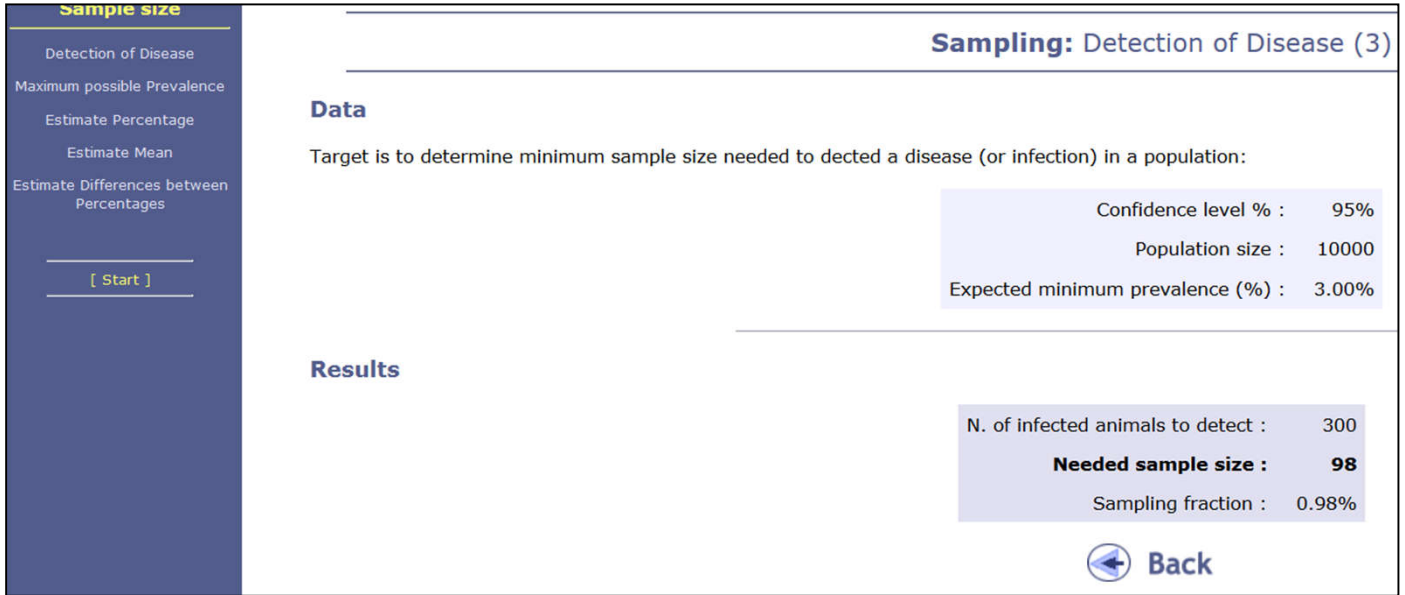
Online tools : WINEPI

http://www.winepi.net/uk/index.htm Sample size > detection of disease

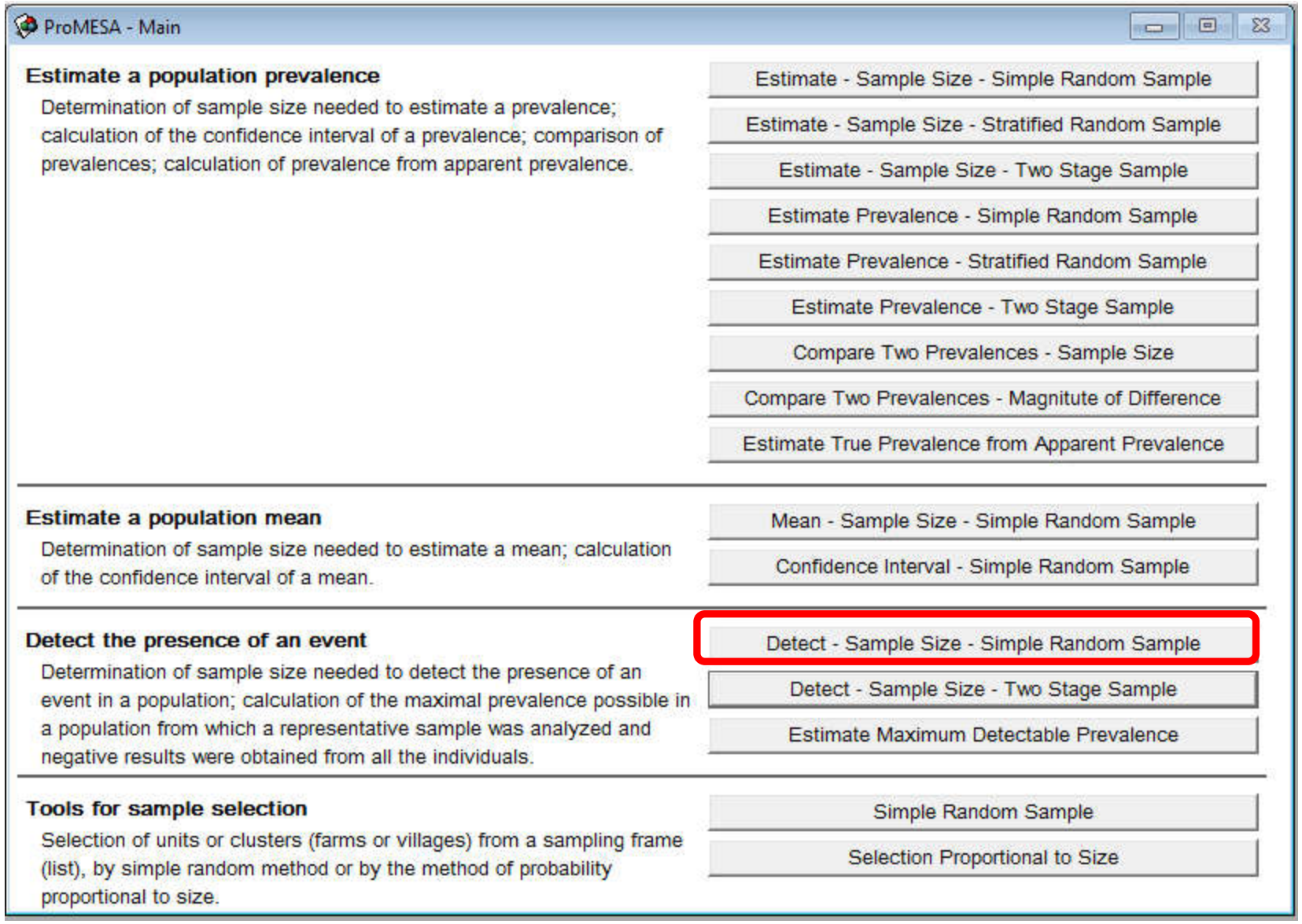
1



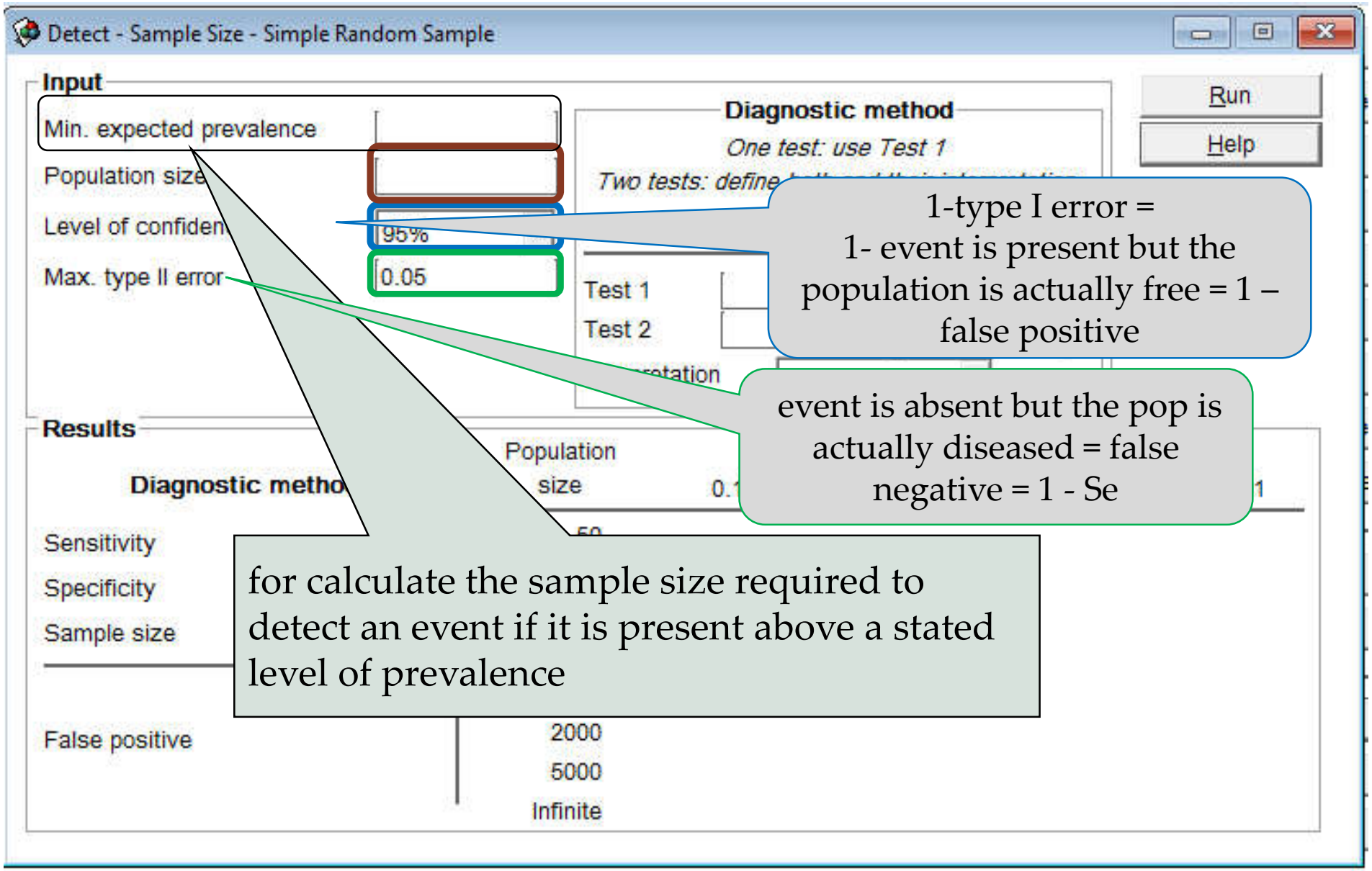
2

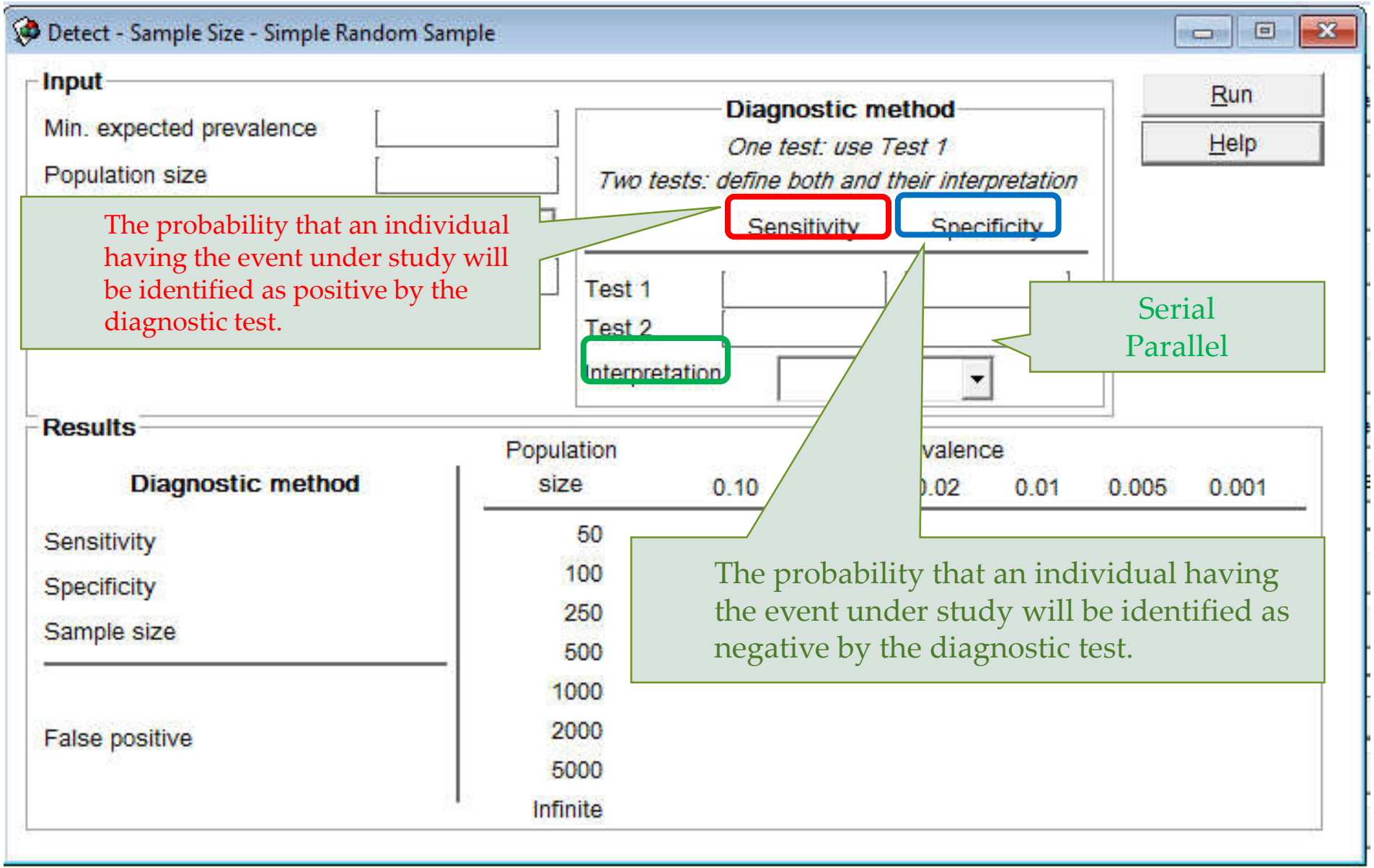


ProMESA


ProMESA - Main
Estimate a population prevalence
Determination of sample size needed to estimate a prevalence; calculation of the confidence interval of a prevalence; comparison of prevalences; calculation of prevalence from apparent prevalence.
Estimate - Sample Size - Simple Random Sample
Estimate - Sample Size - Stratified Random Sample
Estimate - Sample Size - Two Stage Sample
Estimate Prevalence - Simple Random Sample
Estimate Prevalence - Stratified Random Sample
Estimate Prevalence - Two Stage Sample
Compare Two Prevalences - Sample Size
Compare Two Prevalences - Magnitute of Difference
Estimate True Prevalence from Apparent Prevalence
Estimate a population mean
Determination of sample size needed to estimate a mean; calculation of the confidence interval of a mean.
Mean - Sample Size - Simple Random Sample
Confidence Interval - Simple Random Sample
Detect the presence of an event
Determination of sample size needed to detect the presence of an event in a population; calculation of the maximal prevalence possible in a population from which a representative sample was analyzed and negative results were obtained from all the individuals.
Detect - Sample Size - Simple Random Sample
Detect - Sample Size - Two Stage Sample
Estimate Maximum Detectable Prevalence
Tools for sample selection
Selection of units or clusters (farms or villages) from a sampling frame (list), by simple random method or by the method of probability proportional to size.
Simple Random Sample
Selection Proportional to Size

$$n = \left[1 - \left(1 - CL\right)^{\frac{1}{e}}\right] \times \left(N - \frac{e-1}{2}\right)$$



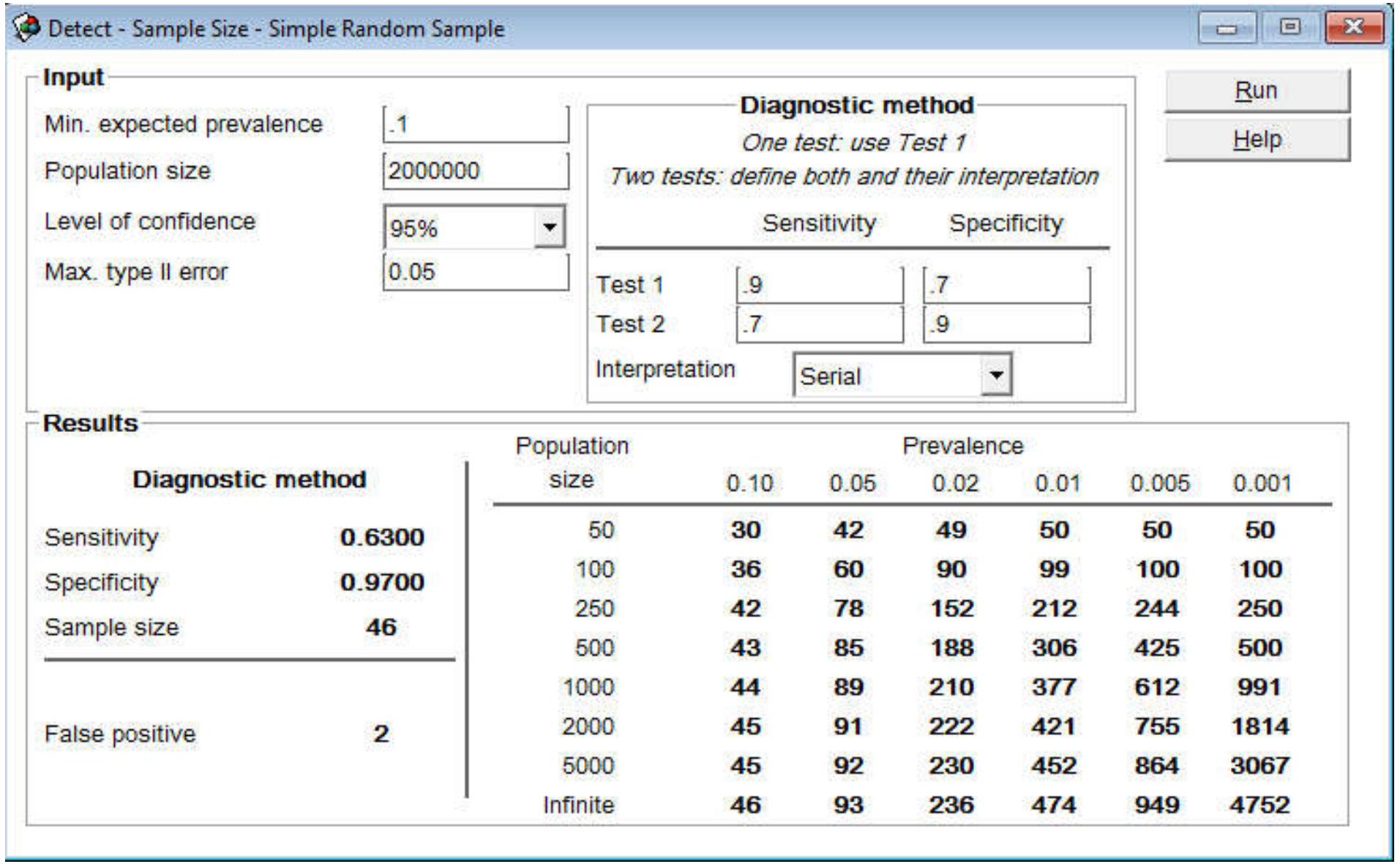

Detect - Sample Size - Simple Random Sample
Input
Min. expected prevalence
Population size
Diagnostic method
One test: use Test 1
Two tests: define both and their interpretation
Sensitivity
Specificity
Test 1
Test 2
Interpretation
Run
Help
The probability that an individual having the event under study will be identified as positive by the diagnostic test.
Serial
Parallel
Results
Diagnostic method
Sensitivity
Specificity
Sample size
False positive
Population size
0.10
0.02
0.01
0.005
0.001
50
100
250
500
1000
2000
5000
Infinite
The probability that an individual having the event under study will be identified as negative by the diagnostic test.

Combination of the tests

- Parallel: individual is positive if any or both tests = positive.
  - Increase sensitivity, decrease specificity
- Series: individual is positive if both tests = positive.
  - Decrease sensitivity, increase specificity

Diagnostic strategy	Sensitivity	Specificity
One test	= Se Test1	= Sp Test1
2 tests – parallel	= 1- (1 – Se Test1) * (1 – Se Test2)	= Sp Test1 * Sp Test2
2 tests – serial	= Se Test1 * Se Test2	= 1 – (1 – Sp Test1) * (1 – Sp Test2)

- Minimum expected prevalence
  - = used for calculate the sample size required to detect an event if it is present above a stated level of prevalence.
- Sensitivity
  - The probability that an individual having the event under study will be identified as positive by the diagnostic test.
- Specificity
  - The probability that an individual not having the event under study will be identified as negative by the diagnostic test.


Detect - Sample Size - Simple Random Sample
Input
Min. expected prevalence
.1
Population size
2000000
Level of confidence
95%
Max. type II error
0.05
Diagnostic method
One test: use Test 1
Two tests: define both and their interpretation
Sensitivity
Specificity
Test 1
.9
.7
Test 2
.7
.9
Interpretation
Serial
Run
Help
Results
Diagnostic method
Sensitivity 0.6300
Specificity 0.9700
Sample size 46
False positive 2
Population size
Prevalence
0.10 0.05 0.02 0.01 0.005 0.001
50 30 42 49 50 50 50
100 36 60 90 99 100 100
250 42 78 152 212 244 250
500 43 85 188 306 425 500
1000 44 89 210 377 612 991
2000 45 91 222 421 755 1814
5000 45 92 230 452 864 3067
Infinite 46 93 236 474 949 4752

Scenarios

1. No positive result = free of event under study
2. Number of positive results > false positives
3. Number of positive results < false positives

1. No positive test result

- Population: free of the interested event, with low risk of error
- Type II error involved: related to the lack of sensitivity of the diagnostic method.
  - event is absent but the pop is actually diseased

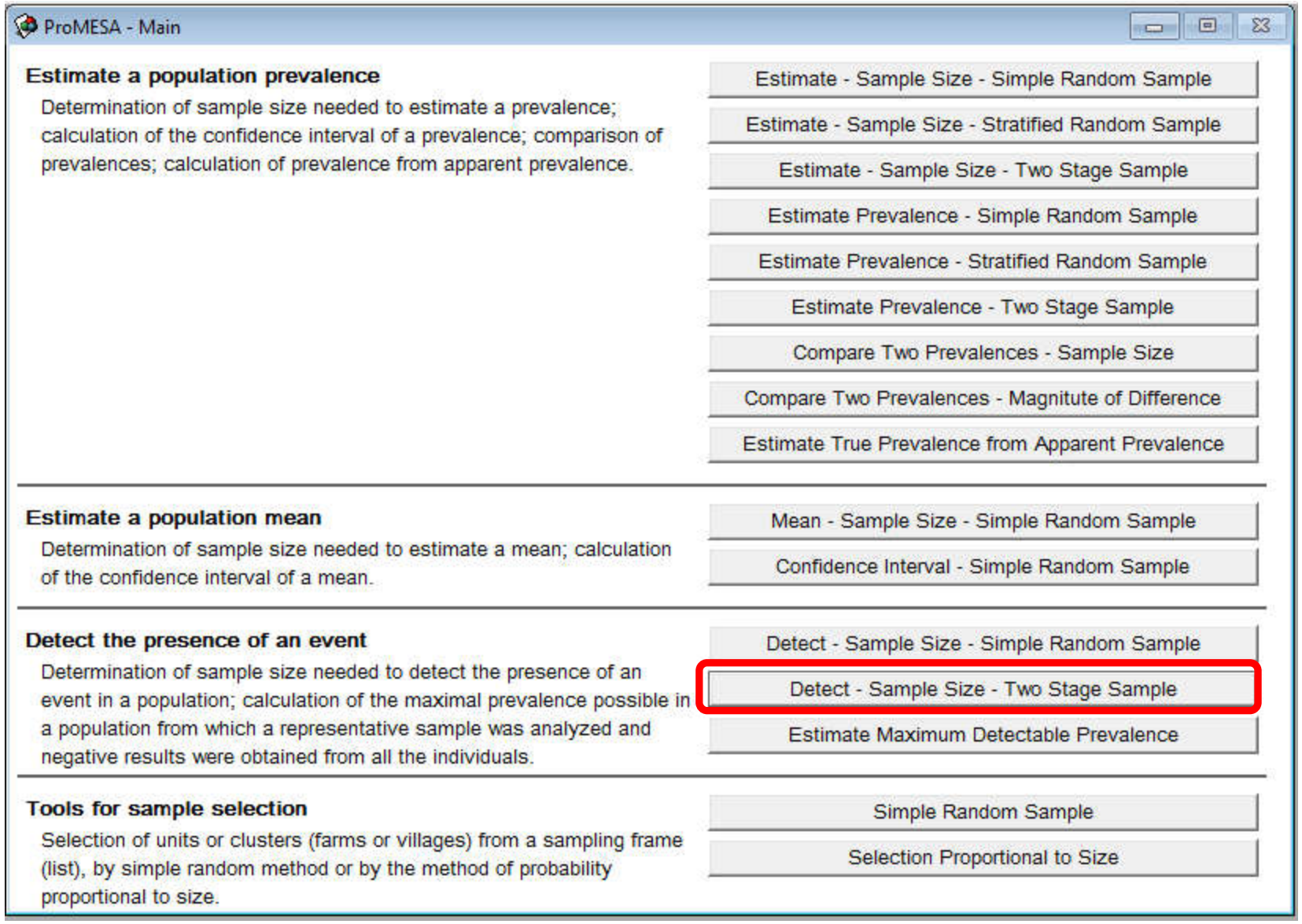
- Type I error: not related because no positive results.

(Type I error = the probability of concluding that the population under study is affected by the event when it is actually free of it.)

2. Positive results > expected number of false positives

- = event is present in the population.
- Type I error?: event is present but the population is actually free.
  - happen when diagnostic specificity < one that put in for sample size calculation.

ProMESA

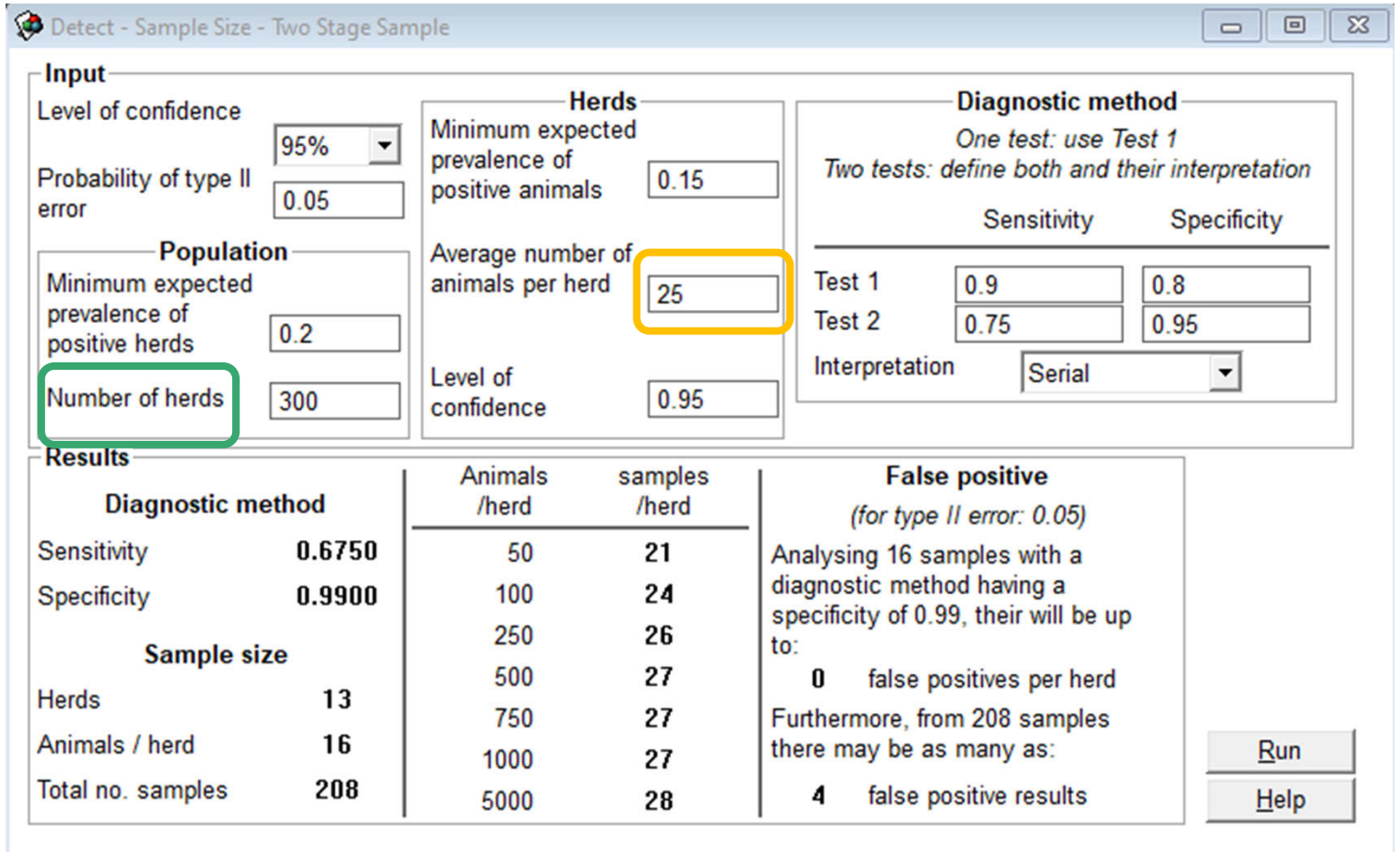

ProMESA - Main
Estimate a population prevalence
Determination of sample size needed to estimate a prevalence; calculation of the confidence interval of a prevalence; comparison of prevalences; calculation of prevalence from apparent prevalence.
Estimate - Sample Size - Simple Random Sample
Estimate - Sample Size - Stratified Random Sample
Estimate - Sample Size - Two Stage Sample
Estimate Prevalence - Simple Random Sample
Estimate Prevalence - Stratified Random Sample
Estimate Prevalence - Two Stage Sample
Compare Two Prevalences - Sample Size
Compare Two Prevalences - Magnitute of Difference
Estimate True Prevalence from Apparent Prevalence
Estimate a population mean
Determination of sample size needed to estimate a mean; calculation of the confidence interval of a mean.
Mean - Sample Size - Simple Random Sample
Confidence Interval - Simple Random Sample
Detect the presence of an event
Determination of sample size needed to detect the presence of an event in a population; calculation of the maximal prevalence possible in a population from which a representative sample was analyzed and negative results were obtained from all the individuals.
Detect - Sample Size - Simple Random Sample
Detect - Sample Size - Two Stage Sample
Estimate Maximum Detectable Prevalence
Tools for sample selection
Selection of units or clusters (farms or villages) from a sampling frame (list), by simple random method or by the method of probability proportional to size.
Simple Random Sample
Selection Proportional to Size

$$n_h = \left[1-(1-CL)^{\frac{1}{e}}\right] \times \left(N_h - \frac{e-1}{2}\right)$$

$$n_i = \left[1-(1-CL)^{\frac{1}{e}}\right] \times \left(N_i - \frac{e-1}{2}\right)$$

N_h The number of herds in the population.

N_i The number of individuals per herd.



- Minimum expected prevalence of positive herds
  - State the lower proportion of affected herds if the event really existed in the region under study.
- Number of herds
  - The total number of farms in the population under study.
- Minimum expected prevalence of positive animals
  - State the lower proportion of affected herds if the event really existed in the region under study.
- Number of animals per herd
  - The mean number of animals per herd.

Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence determination
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Sample size:
Freedom from a disease

Faculty of Veterinary Medicine
Kasetsart University

Home | Prevalence | Freedom | Studies | Diagnostics | Sampling

- 1-Stage Freedom analysis
  - Analyse results
  - Sample Size
- Representative 1-stage
  - Back-calculate design prevalence
  - Confidence of freedom for multiple time periods
  - Confidence of freedom for a single time period
  - Population sensitivity - constant unit sensitivity
  - Population sensitivity - pooled sampling
  - Population sensitivity - varying unit sensitivity
  - Sample size - perfect test specificity
  - Sample size - pooled sampling in a large population
  - Sample size for target confidence of freedom
- Representative 2-stage
  - Analyse 2-stage survey - actual data
  - Analyse 2-stage survey - fixed sample size
  - Least-cost sample sizes from sampling frame
  - Least-cost sample sizes - no sampling frame
  - Sample sizes - specified cluster sensitivity
  - Stochastic analysis - 2-stage freedom data
- Risk based 1-stage

EPITOOLS

Population size

Test sensitivity

Test specificity

Design prevalence (proportion or number of units)

Analysis options:

Desired type I error (1 - minimum population-specificity) 0.05

Desired type II error (1 - minimum population-sensitivity) 0.05

Calculation method:
(these settings can usually be left as default values)

- (•) Modified hypergeometric exact
- ( ) Simple binomial (large population)

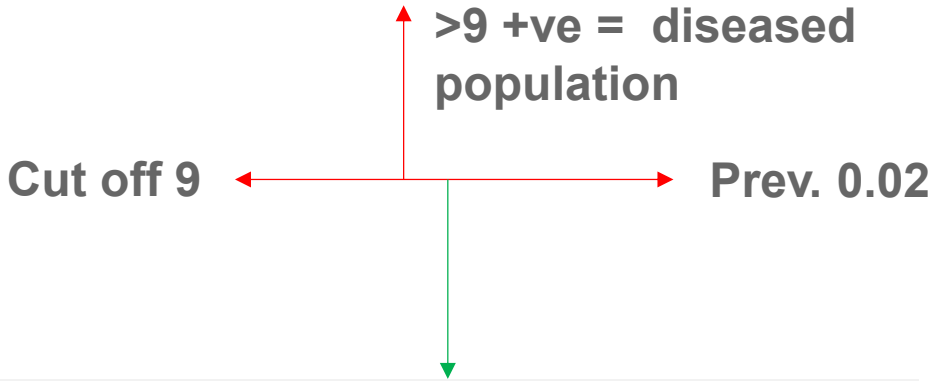
Population threshold for binomial method 10000

Maximum limit for sample size 3200

Precision (significant digits) 4

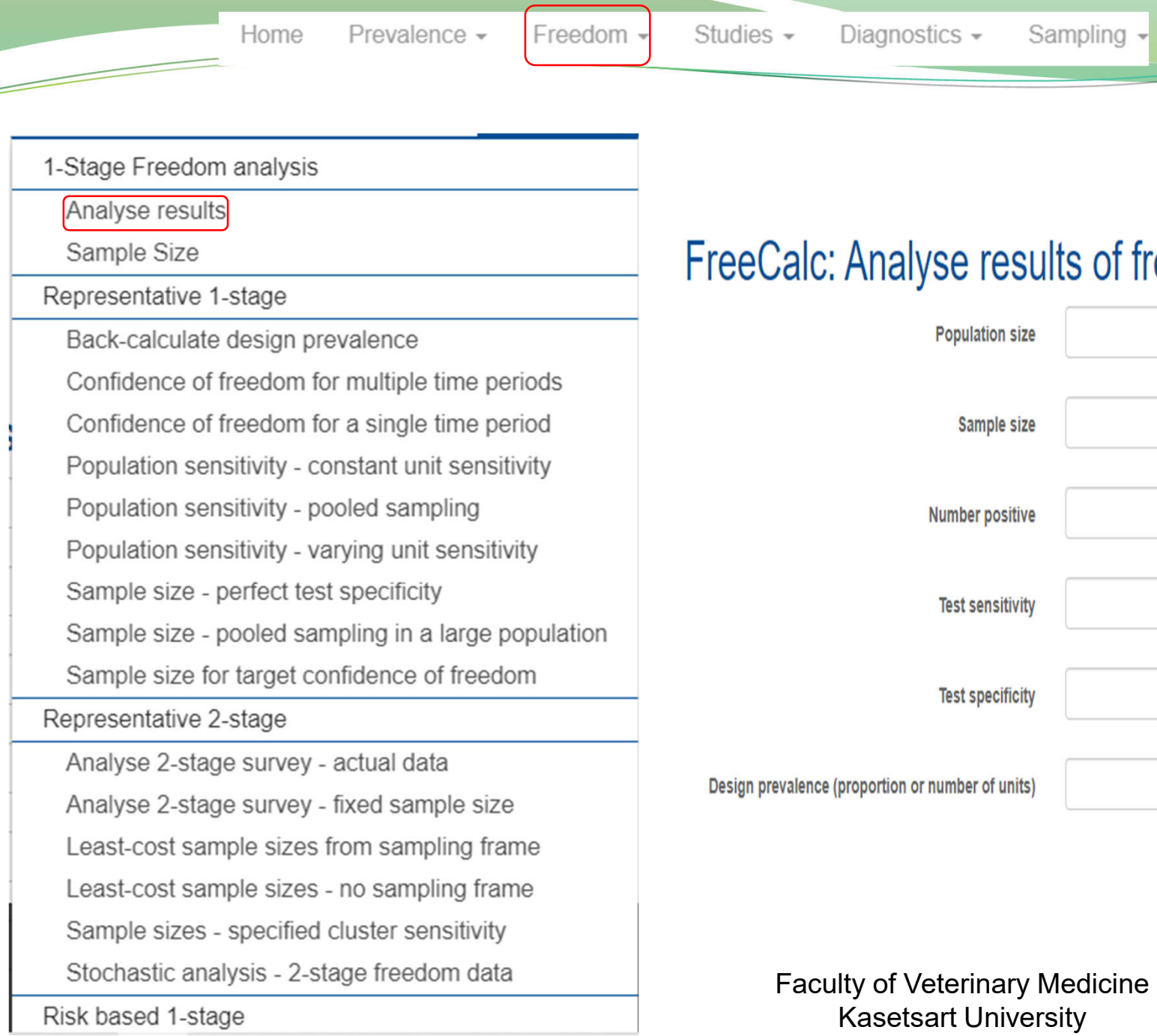
https://epitools.ausvet.com.au/

Pop size	10000
Se	0.95
Sp	0.99
Design Prev	0.02
type I error	0.05
type II error	0.05
max n	3200
required n	539
cut-point num of positive	9
type I error	0.0493
type II error	0.0474
pop level Se	0.9507
Pop level Sp	0.9526



If a random sample of **539** units is taken from a population of **10000** and **9** or fewer reactors are found, the probability that the population is diseased at a prevalence of **0.02** is **0.0493.**

Pop size	1000	1000	1000	1000
Se	0.95	0.95	0.95	0.9
Sp	0.99	0.99	0.9	0.99
Design Prev	0.02	0.2	0.2	0.2
type I error	0.05	0.05	0.05	0.05
type II error	0.05	0.05	0.05	0.05
max n	3200	3200	3200	3200
required n	539	22	55	24
cut-point num of positive	9	1	9	1
type I error	0.0493	0.05	0.0468	0.0441
type II error	0.0474	0.0202	0.0444	0.0239
pop level Se	0.9507	0.95	0.9532	0.9559
Pop level Sp	0.9526	0.9798	0.9556	0.9761



Faculty of Veterinary Medicine
Kasetsart University

Inputs

- Design prevalence: the hypothetical prevalence to be detected.
- Maximum acceptable Type I (1 - population-sensitivity) and Type II (1 - population-specificity) error values for determining whether to accept/reject the null or alternative hypothesis, assuming a null hypothesis that the population is diseased.
- Calculation method: hypergeometric (for small populations), or simple binomial (for large populations).

outputs

- Assuming a null hypothesis that the population is diseased.
- The probability of the null hypothesis is the probability of observing this many reactors or fewer ($\leq$reactors), if the population was diseased at a level $\geq$ the specified design prevalence.
  - If this probability is small, we can conclude that it is very unlikely that the population is diseased.
  - If the probability is large, then there is not enough evidence to conclude that the population is free from disease;
- The probability of the alternative hypothesis is the probability of observing this many reactors or more if the population was truly disease free.
  - If this is small, then it is very unlikely that the population is free from disease.
  - If it is large, then it is consistent with there being no disease in the population.

outputs

- If both null and alternative probabilities are small,
  - it suggests that the population is not free from disease,
  - but the prevalence is less than the design prevalence specified; and
- If both null and alternative probabilities are large,
  - the sample size was too small to distinguish a population with the specified design prevalence from a disease-free population.

Test sensitivity	0.95
Test specificity	0.99
Population size	10000
Design prevalence	0.02
Diseased elements	20
Sample size	539

Number positive < cut-off

Number positive	3 (cut-off=9)
Target Type I error (1-Se, FN)	0.05
Target Type II error (1-Sp, FP)	0.05

Null hypothesis:	Probability of observing <= 3 reactors in a sample of 539 individuals from a population with a disease prevalence of 2% = < 0.0001.
Alternative hypothesis:	Probability of observing >= 3 reactors in a sample of 539 individuals from a disease free population = 0.9057.
Conclusion:	These results are adequate to reject the null hypothesis and conclude that the population is free from disease (at the expected minimum prevalence of 2%) at the 0.9999 confidence level.
Cluster-sensitivity (SeH)	0.9999
Cluster-specificity (SpH)	0.2131

Test sensitivity	0.95
Test specificity	0.99
Population size	10000
Design prevalence	0.02
Diseased elements	20
Sample size	539

Number positive > cut-off

Number positive	15 (cut-off=9)
Target Type I error (1-Se, FN)	0.05
Target Type II error (1-Sp, FP)	0.05

Null hypothesis:	Probability of observing <= 15 reactors in a sample of 539 individuals from a population with a disease prevalence of 2% = 0.513.
Alternative hypothesis:	Probability of observing >= 15 reactors in a sample of 539 individuals from a disease free population = 0.0005.
Conclusion:	These results are not adequate to conclude that the population is free from disease (at the expected minimum prevalence of 2%). The confidence level is only 0.487. We may conclude that the population is diseased at a confidence level of 0.9995.
Cluster-sensitivity (SeH)	0.487
Cluster-specificity (SpH)	0.9999

Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence determination
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Sample size:
Prevalence determination

Faculty of Veterinary Medicine
Kasetsart University

When to determine disease prevalence

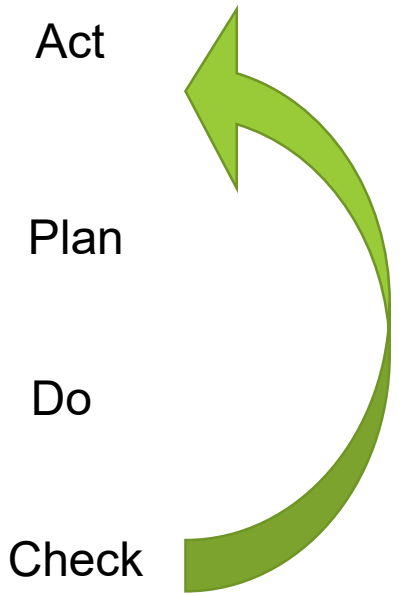
- When you want to know
  - the level of the disease

in order to

  - plan for disease control / eradication

Implementation

  - the effectiveness of control measures



Parameter to be considered

- For descriptive analysis
  - Confidence
  - Precision

- Mostly chosen levels for the confidence are 95 or 99%
- Example: Given a 99% confidence interval, true prevalence of TB in cattle will fall between 1% and 5%.

Estimate a population prevalence – sample size

- Simple random sample
- Stratified random sample
- Two stage sample

Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Simple random sampling

$$n = \frac{p * (1 - p) * z^2}{e^2}$$

p = The assumed prevalence of the event in the population under study.

Z = The critical value obtained from a standard normal distribution
90% CI → Z = 1.64
95% CI → Z = 1.96
99% CI → Z = 2.58
CI = p ± Z * SE

e = The maximum absolute error that the user is willing to accept.
ie; A prevalence of 0.40
A relative error of 0.10
The absolute error = prevalence*relative error = 0.40 *0.10 = 0.04
relative error = absolute error / prevalence
Relative error should be ≤ 0.20

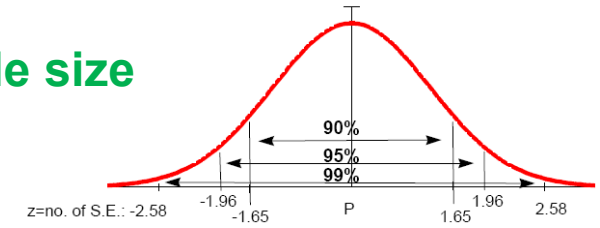
ProMESA's help

$$n = \frac{p * (1 - p) * z^2}{e^2}$$

P	P (1-P)
0.5	0.25
0.4	0.24
0.3	0.21
0.2	0.16
0.1	0.09

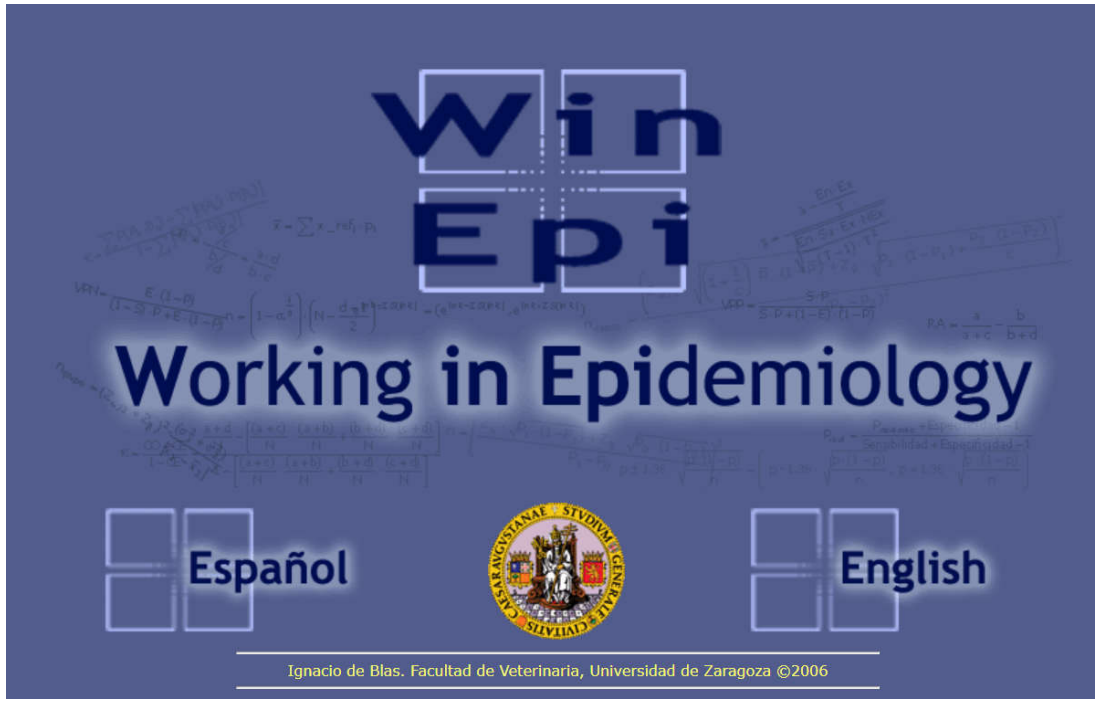
p = 0.50 = conservative expected prevalence → max. n
p at 0.60 → sample size requirement = p at 0.40

Z: at higher confidence level → bigger sample size



e = less accepted error → bigger sample size

www.winepi.net



Sample size

Detection of Disease

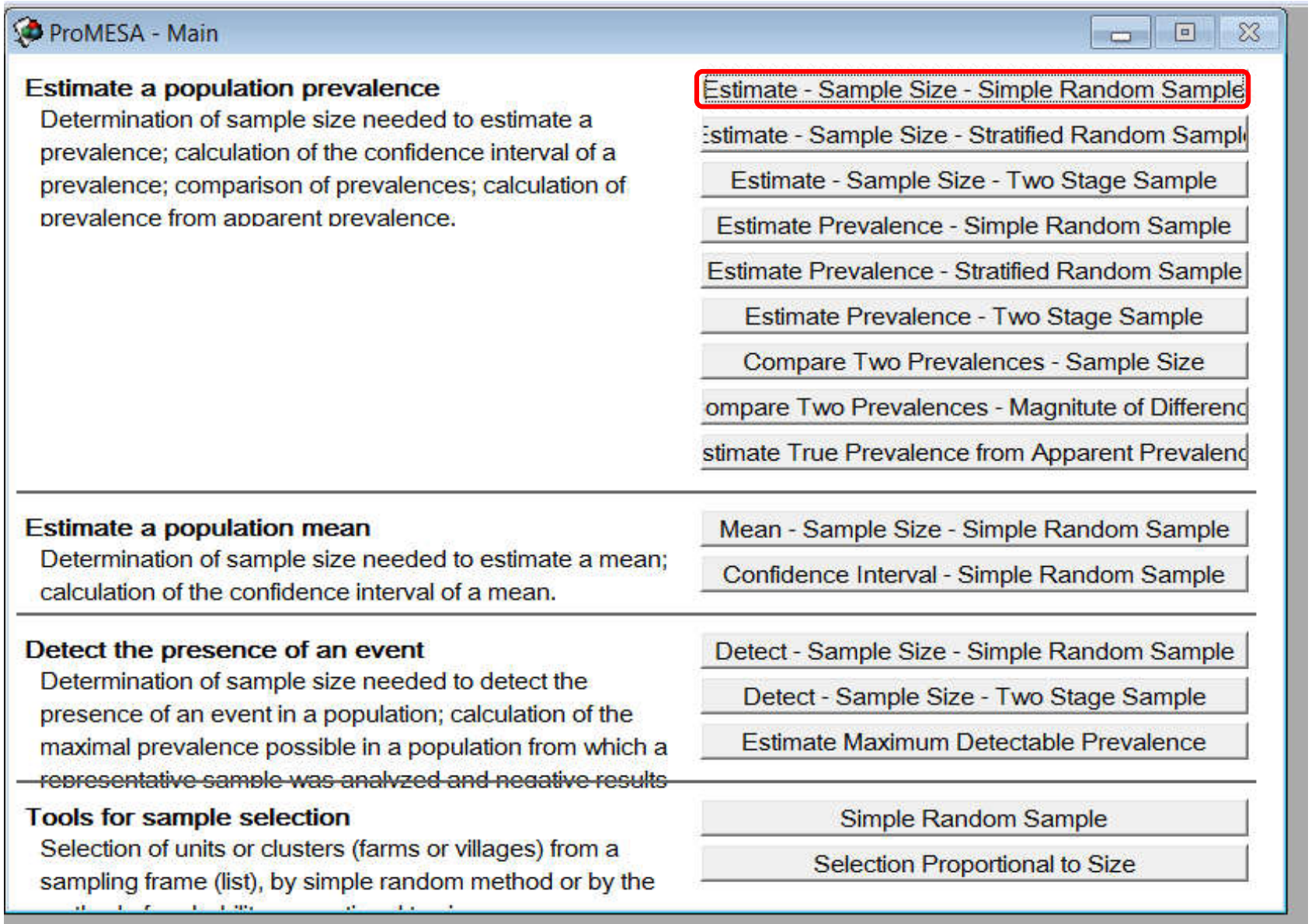
Maximum possible Prevalence

Estimate Percentage

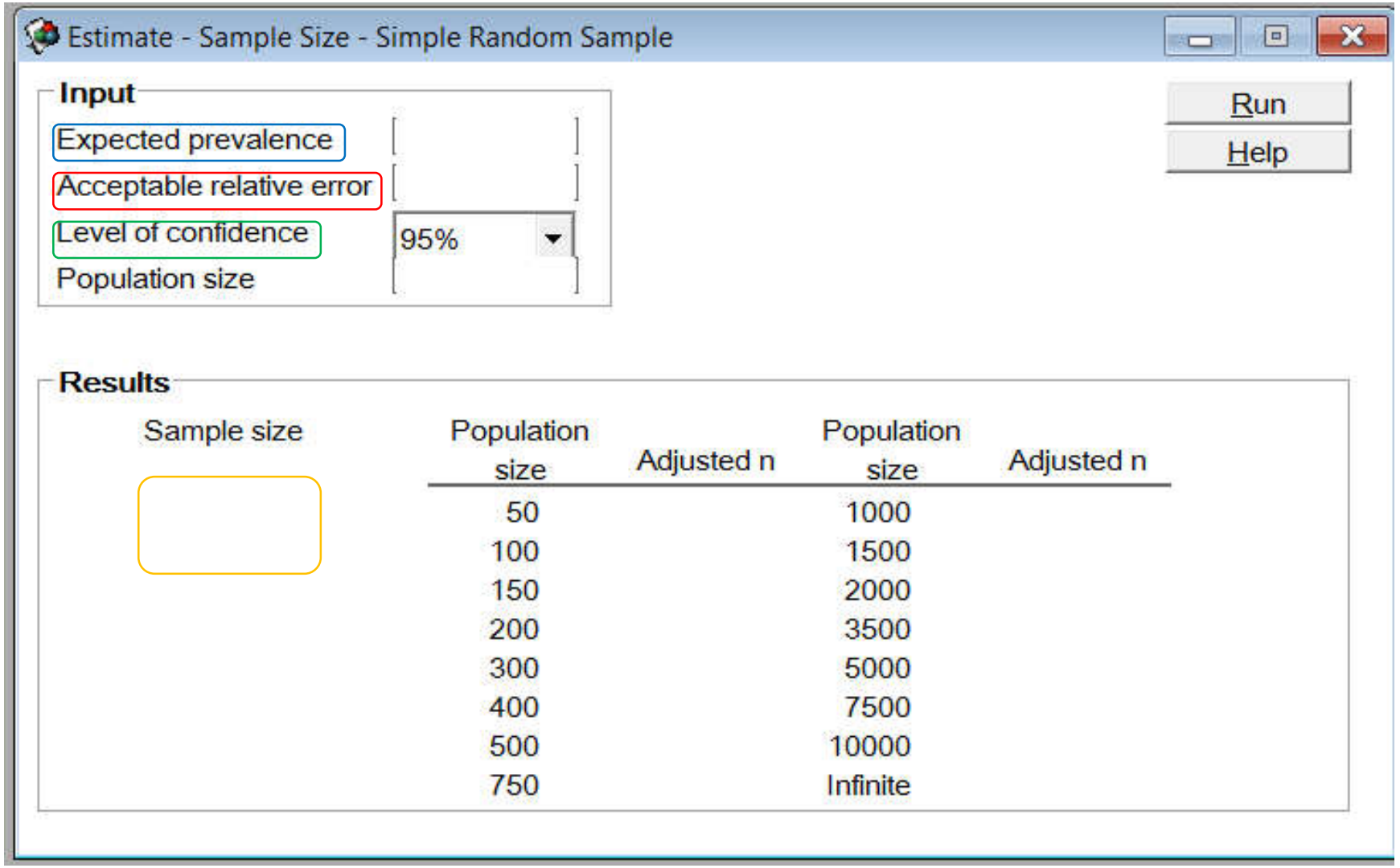
Estimate Mean

Estimate Differences between Percentages

ProMESA

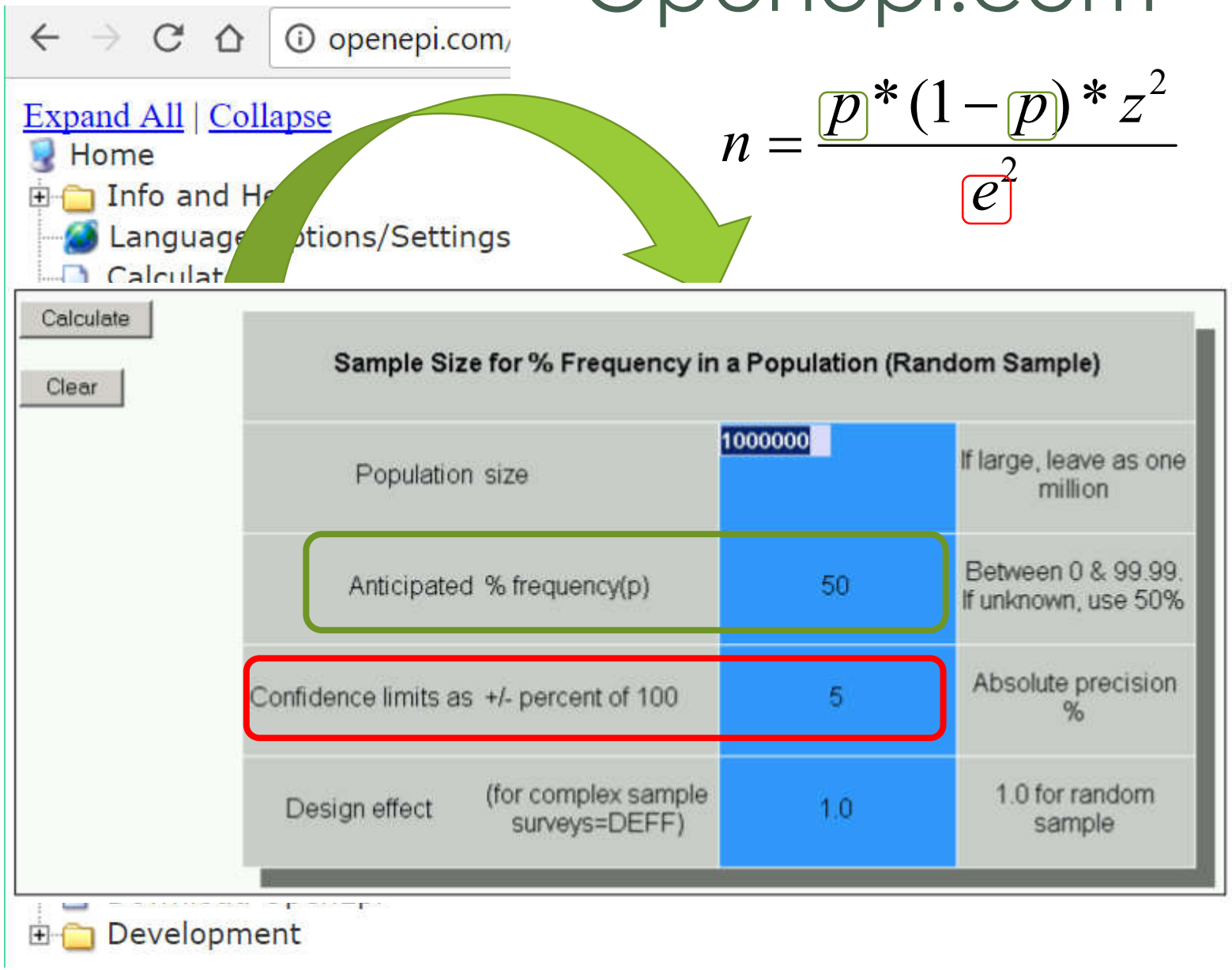

ProMESA - Main
Estimate a population prevalence
Determination of sample size needed to estimate a prevalence; calculation of the confidence interval of a prevalence; comparison of prevalences; calculation of prevalence from apparent prevalence.
Estimate - Sample Size - Simple Random Sample
Estimate - Sample Size - Stratified Random Sampl
Estimate - Sample Size - Two Stage Sample
Estimate Prevalence - Simple Random Sample
Estimate Prevalence - Stratified Random Sample
Estimate Prevalence - Two Stage Sample
Compare Two Prevalences - Sample Size
ompare Two Prevalences - Magnitute of Differenc
stimate True Prevalence from Apparent Prevalenc
Estimate a population mean
Determination of sample size needed to estimate a mean; calculation of the confidence interval of a mean.
Mean - Sample Size - Simple Random Sample
Confidence Interval - Simple Random Sample
Detect the presence of an event
Determination of sample size needed to detect the presence of an event in a population; calculation of the maximal prevalence possible in a population from which a representative sample was analyzed and negative results
Detect - Sample Size - Simple Random Sample
Detect - Sample Size - Two Stage Sample
Estimate Maximum Detectable Prevalence
Tools for sample selection
Selection of units or clusters (farms or villages) from a sampling frame (list), by simple random method or by the
Simple Random Sample
Selection Proportional to Size

$$n = \frac{p * (1 - p) * z^2}{e^2}$$

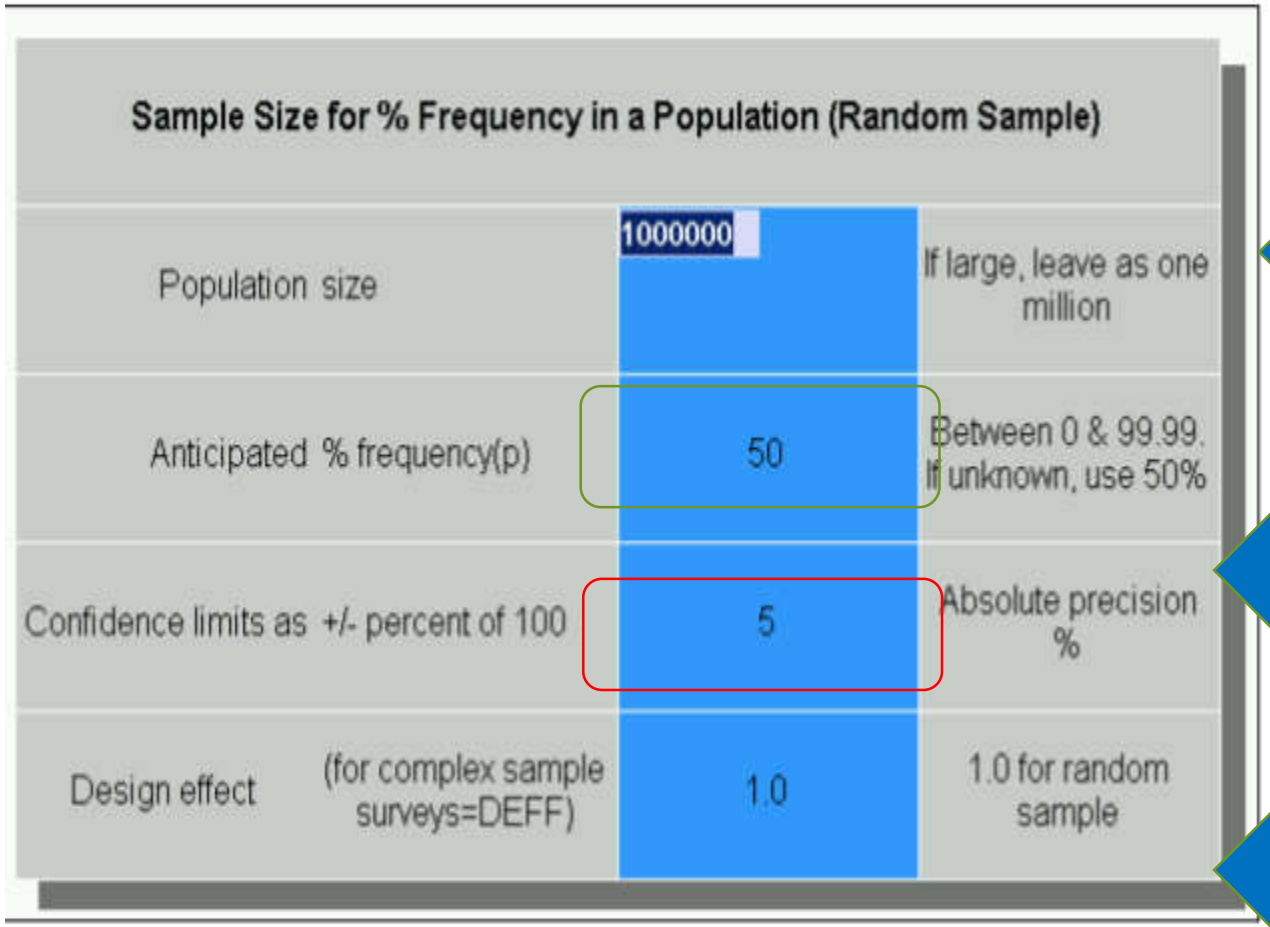


Openepi.com

$$n = \frac{p^* (1 - p) * z^2}{e^2}$$



$$n = \frac{p * (1 - p) * z^2}{e^2}$$



Sample Size for % Frequency in a Population (Random Sample)			
Population size		1000000	If large, leave as one million
Anticipated % frequency(p)		50	Between 0 & 99.99. If unknown, use 50%
Confidence limits as +/- percent of 100		5	Absolute precision %
Design effect	(for complex sample surveys=DEFF)	1.0	1.0 for random sample

>=100,000 is large

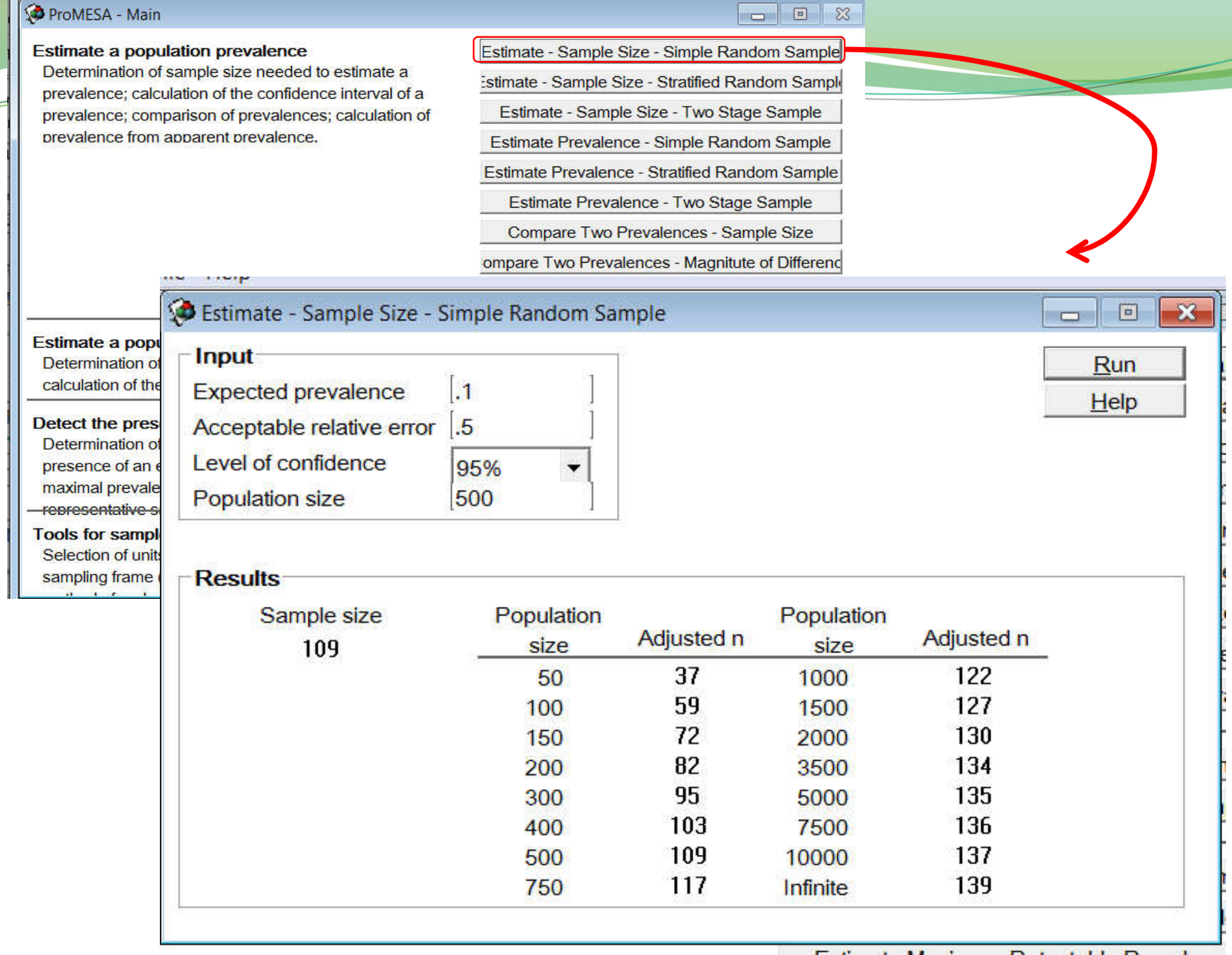
+5%, 50% +5%, i.e., (45%, 55%).

SRS=1
Cluster = 2-10
n = n of SRS*DEFF

Example 1 – sample size: SRS

The investigator has been asked to **determine the proportion** of a 500 cows herd that will yield a positive culture for *M. paratuberculosis*.

The acceptable absolute precision is +/- 5% and the expected prevalence when sampling at the slaughter house is assumed to be 10%.

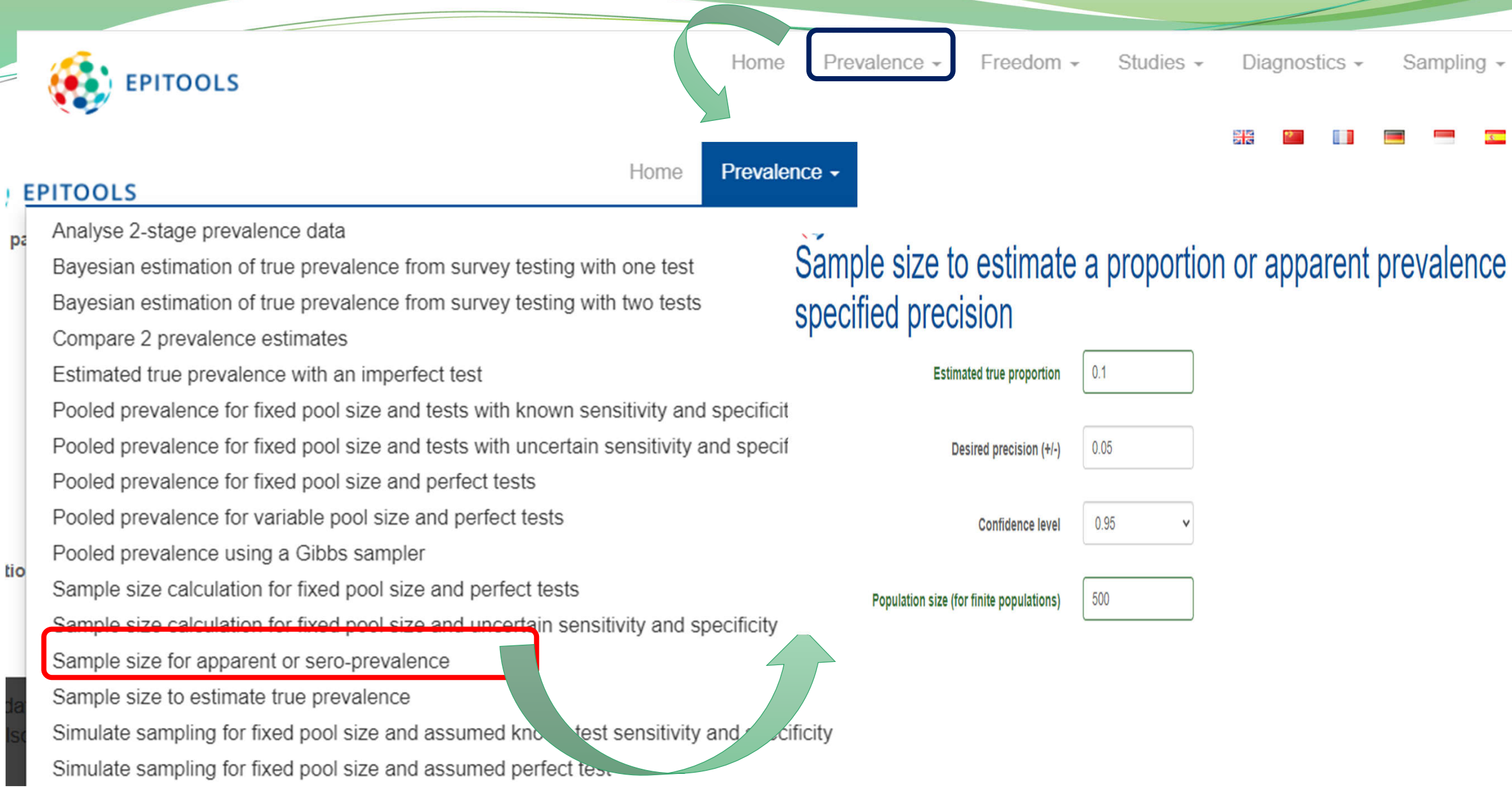

ProMESA - Main
Estimate a population prevalence
Determination of sample size needed to estimate a prevalence; calculation of the confidence interval of a prevalence; comparison of prevalences; calculation of prevalence from apparent prevalence.
Estimate - Sample Size - Simple Random Sample
Estimate - Sample Size - Stratified Random Sampl
Estimate - Sample Size - Two Stage Sample
Estimate Prevalence - Simple Random Sample
Estimate Prevalence - Stratified Random Sample
Estimate Prevalence - Two Stage Sample
Compare Two Prevalences - Sample Size
ompare Two Prevalences - Magnitute of Differenc
Estimate - Sample Size - Simple Random Sample
Input
Expected prevalence .1
Acceptable relative error .5
Level of confidence 95%
Population size 500
Run
Help
Results
Sample size
109
Population size
Adjusted n
Population size
Adjusted n
50 37 1000 122
100 59 1500 127
150 72 2000 130
200 82 3500 134
300 95 5000 135
400 103 7500 136
500 109 10000 137
750 117 Infinite 139

Absolute error

Sample size bound to some limitations such as laboratory capacity.

Calculate precision: if too large → useless to carry out the research

e = The maximum absolute error that the user is willing to accept.
= Z * SE (known n; i.e. Capacity of lab.)


EPITOOLS
Home
Prevalence
Freedom
Studies
Diagnostics
Sampling
EPITOOLS
Home
Prevalence
Analyse 2-stage prevalence data
Bayesian estimation of true prevalence from survey testing with one test
Bayesian estimation of true prevalence from survey testing with two tests
Compare 2 prevalence estimates
Estimated true prevalence with an imperfect test
Pooled prevalence for fixed pool size and tests with known sensitivity and specificit
Pooled prevalence for fixed pool size and tests with uncertain sensitivity and specif
Pooled prevalence for fixed pool size and perfect tests
Pooled prevalence for variable pool size and perfect tests
Pooled prevalence using a Gibbs sampler
Sample size calculation for fixed pool size and perfect tests
Sample size calculation for fixed pool size and uncertain sensitivity and specificity
Sample size for apparent or sero-prevalence
Sample size to estimate true prevalence
Simulate sampling for fixed pool size and assumed kno test sensitivity and cificity
Simulate sampling for fixed pool size and assumed perfect test
Sample size to estimate a proportion or apparent prevalence with specified precision
Estimated true proportion
0.1
Desired precision (+/-)
0.05
Confidence level
0.95
Population size (for finite populations)
500

Large population	139
Population = 500	109

Sample size to estimate a proportion or apparent prevalence with specified precision

Estimated true proportion: 0.1

Desired precision (+/-): 0.05

Confidence level: 0.95

Population size (for finite populations): 500

Sample sizes for varying prevalence and precision values

	AP = 0.01	AP = 0.02	AP = 0.05	AP = 0.1	AP = 0.2	AP = 0.5
Precision = 0.01	381	753	1825	3458	6147	9604
Precision = 0.02	96	189	457	865	1537	2401
Precision = 0.05	16	31	73	139	246	385
Precision = 0.1	4	8	19	35	62	97
Precision = 0.2	1	2	5	9	16	25

Sample Size for % Frequency in a Population (Random Sample)

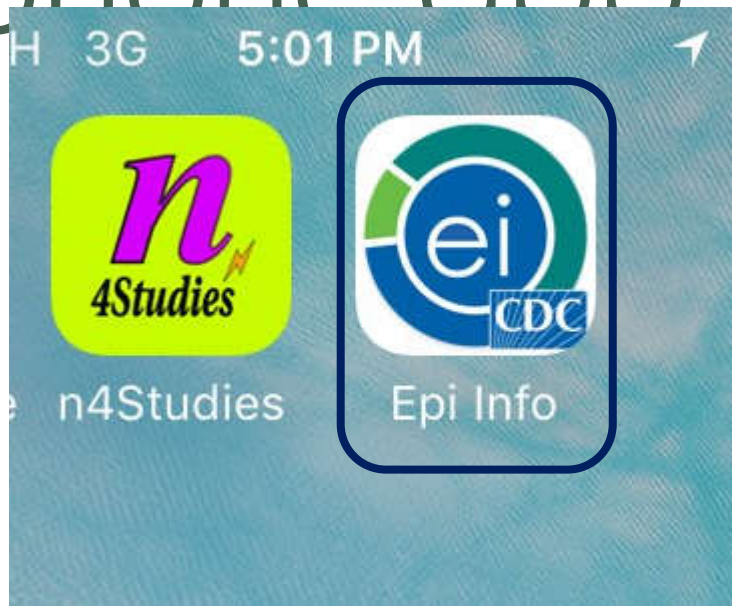
Population size	500	If large, leave as one million
Anticipated % frequency(p)	10	Between 0 & 99.99. If unknown, use 50%
Confidence 100		
Design e surveys-		

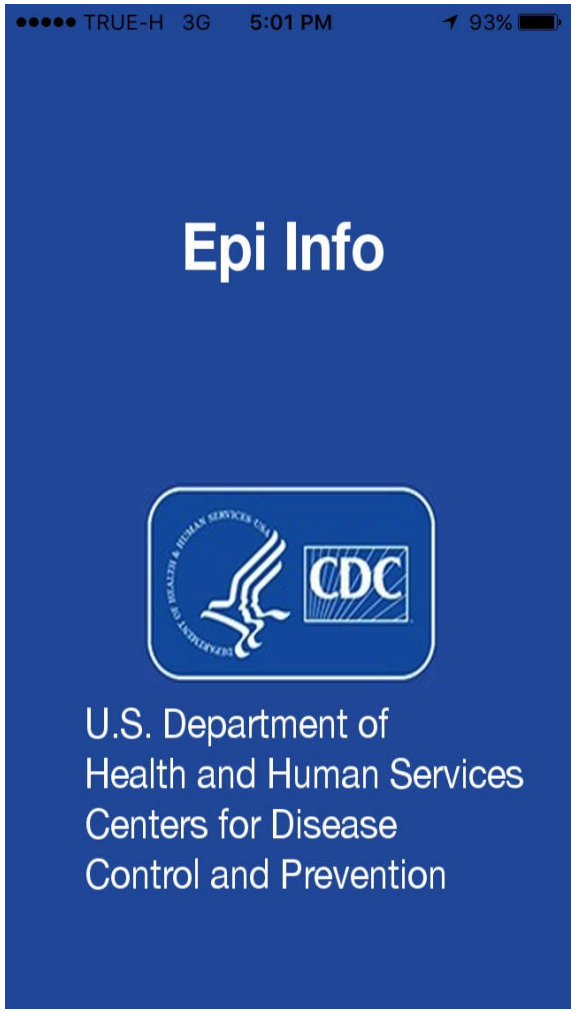
Population size(for finite population correction factor or fpc)(*N*): 500
Hypothesized % frequency of outcome factor in the population (*p*): 10%+/-5
Confidence limits as % of 100(absolute +/- %)(*d*): 5%
Design effect (for cluster surveys-*DEFF*): 1

Sample Size(*n*) for Various Confidence Levels

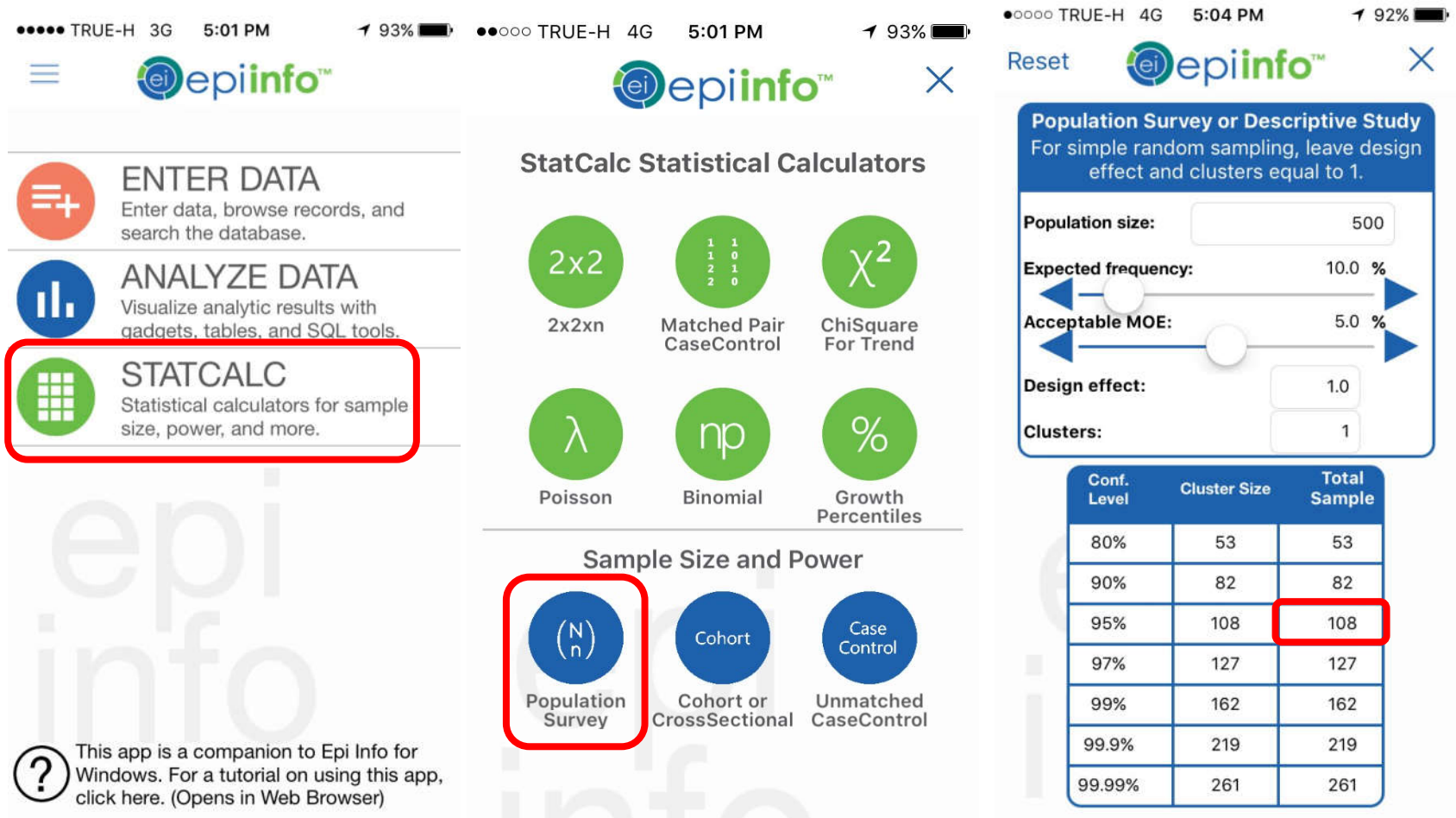
ConfidenceLevel(%)	Sample Size
95%	109
80%	53
90%	82
97%	127
99%	162
99.9%	220
99.99%	262

Smart phone app





EpiInfo


TRUE-H 3G 5:01 PM 93%
epiinfo™
ENTER DATA
Enter data, browse records, and search the database.
ANALYZE DATA
Visualize analytic results with gadgets, tables, and SQL tools.
STATCALC
Statistical calculators for sample size, power, and more.
This app is a companion to Epi Info for Windows. For a tutorial on using this app, click here. (Opens in Web Browser)
TRUE-H 4G 5:01 PM 93%
epiinfo™
StatCalc Statistical Calculators
2x2
2x2xn
Matched Pair CaseControl
ChiSquare For Trend
Poisson
Binomial
Growth Percentiles
Sample Size and Power
Population Survey
Cohort
Cohort or CrossSectional
Case Control
Unmatched CaseControl
TRUE-H 4G 5:04 PM 92%
Reset
epiinfo™
Population Survey or Descriptive Study
For simple random sampling, leave design effect and clusters equal to 1.
Population size: 500
Expected frequency: 10.0 %
Acceptable MOE: 5.0 %
Design effect: 1.0
Clusters: 1
Conf. Level | Cluster Size | Total Sample
80% | 53 | 53
90% | 82 | 82
95% | 108 | 108
97% | 127 | 127
99% | 162 | 162
99.9% | 219 | 219
99.99% | 261 | 261

Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Stratified sampling

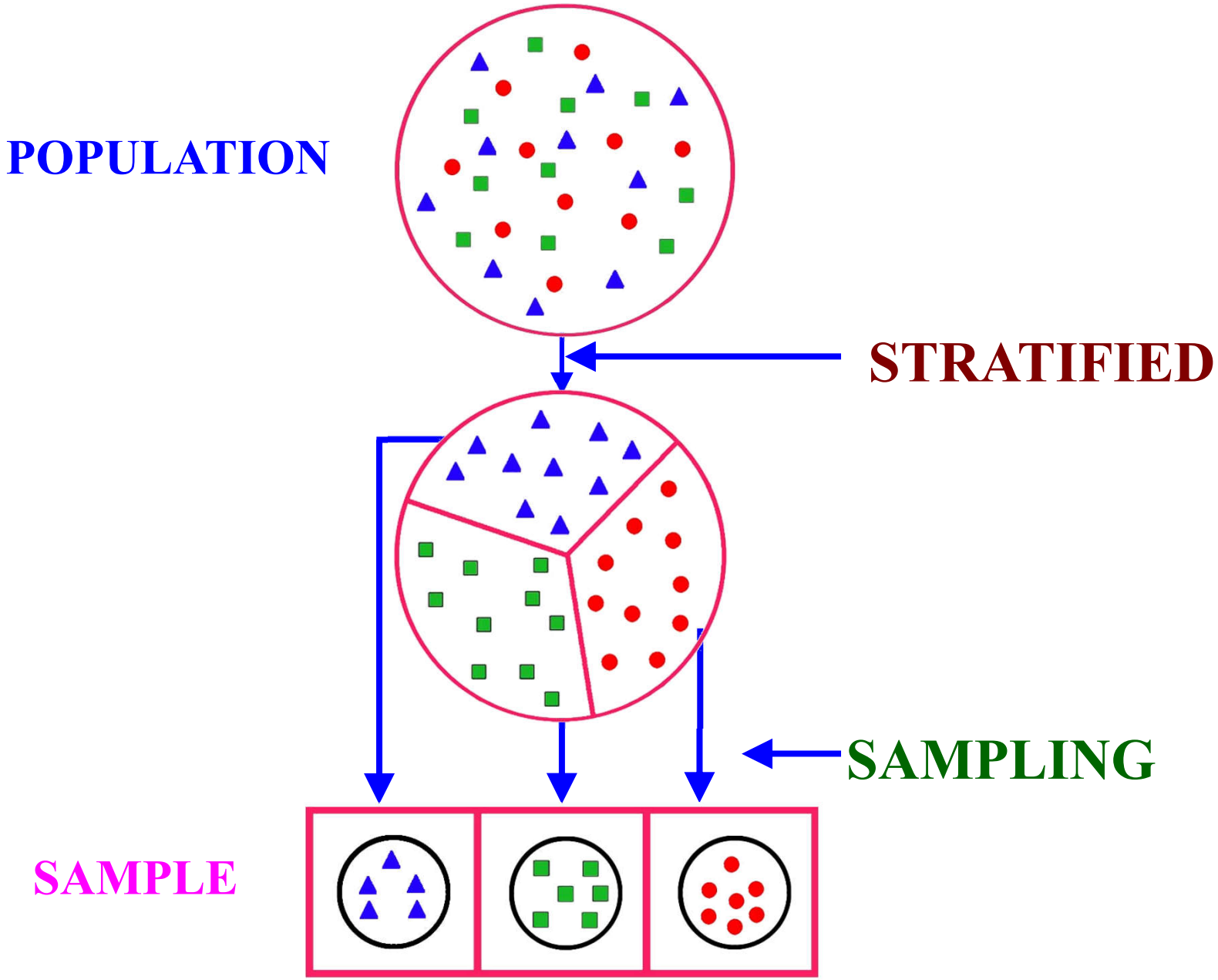
- **Strata** = groups of the population divided because of a strong **biologic hypothesis**, which implies a difference of prevalence for the disease/parameter studied.

  Ex = production type (dairy/beef), age group, species, …

- **Clusters** = groups of the population divided because of a **practical** reason for the investigator.

  Ex = hospital, school, herd…

Stratified random sampling



Stratified sampling

- Stratified sampling : simple random sampling within each strata
- **All strata** are represented in the sample
- Number of units to represent each stratum is proportional to the importance of the strata in the population

Advantages of stratified sampling

- Focus on important subpopulations and ignores irrelevant ones.

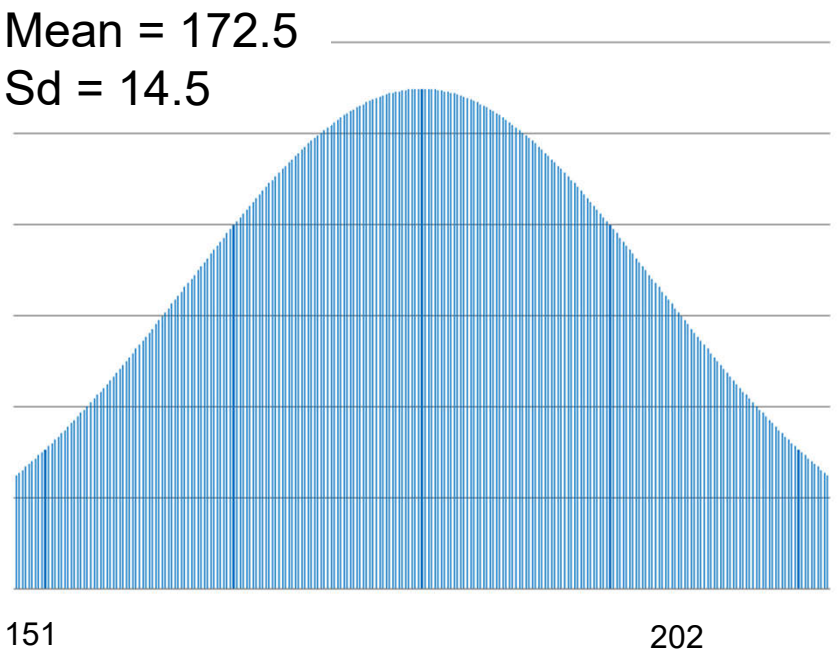
- Variability within strata is minimized
- Variability between strata is maximized

- Optimization: for a given sampling size (n), precision of the estimate is always higher than those obtained from simple random sampling
  - Helps to reduce sample size

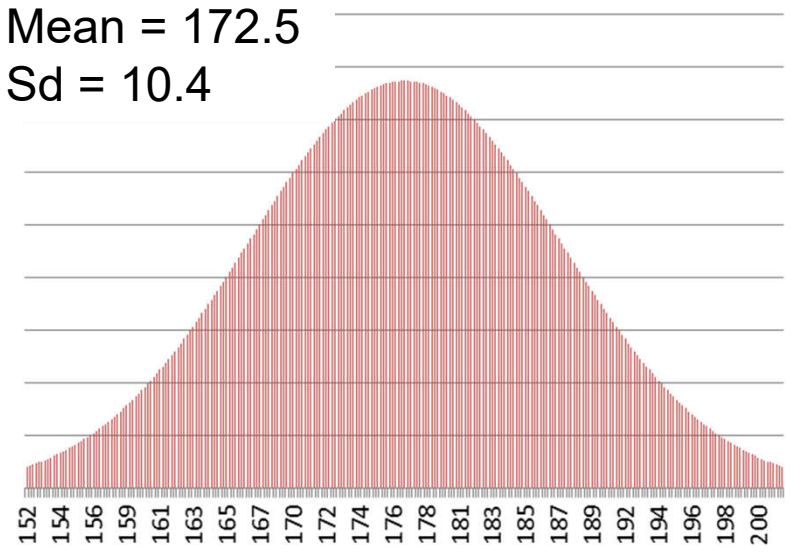
Selection of relevant stratification variables can be difficult !!!!

Exemple : size of veterinary students

Simple random sampling (n=60)



Stratified sampling (boys/girls, n= 60)



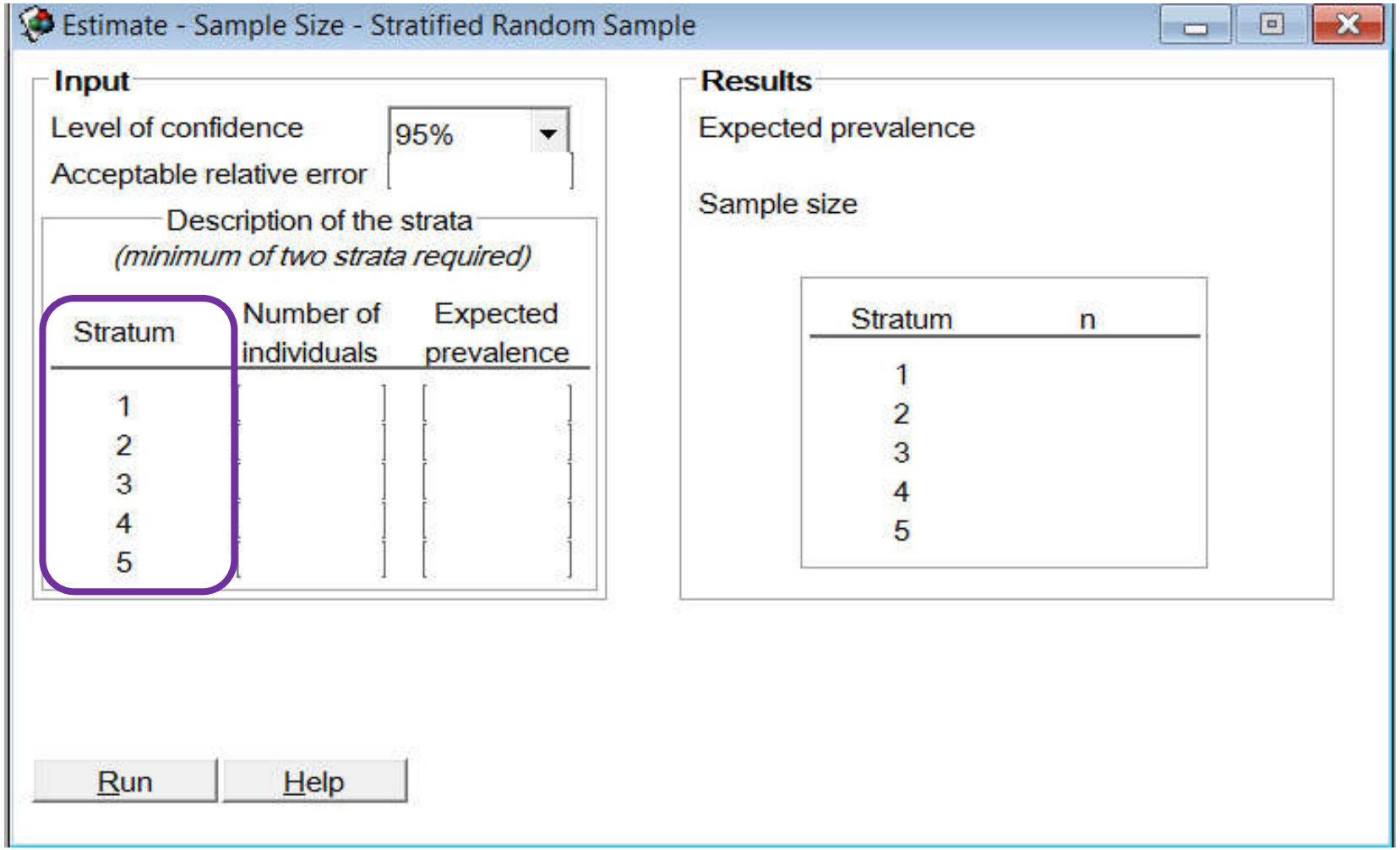
$$n = \frac{\sum_{i=1}^{e}\left[\frac{(n_i)^2 \times p_i \times (1-p_i)}{w_i}\right]}{N^2 \times \frac{AE^2}{Z^2} + \sum_{i=1}^{e}\left[n_i \times p_i \times (1-p_i)\right]}$$

Where:

e	The number of strata.
n_i	The number of individuals in strata i.
p_i	The expected prevalence in strata i.
N	The total number of individuals in the population.
AE	The acceptable absolute error.
z	The value obtained from the standard normal distribution. To each value of confidence there is a correspondent value of z. The levels of confidence more frequently used in biological studies are 90%, 95% and 99%. The values of z correspondent to them are 1.645, 1.96, and 2.58 respectively.
w_i	A weighting factor of each strata, calculated as follows:

$$w_i = \frac{n_i \times \sqrt{p_i \times (1-p_i)}}{\sum_{i=1}^{e}\left[n_i \times \sqrt{p_i \times (1-p_i)}\right]}.$$

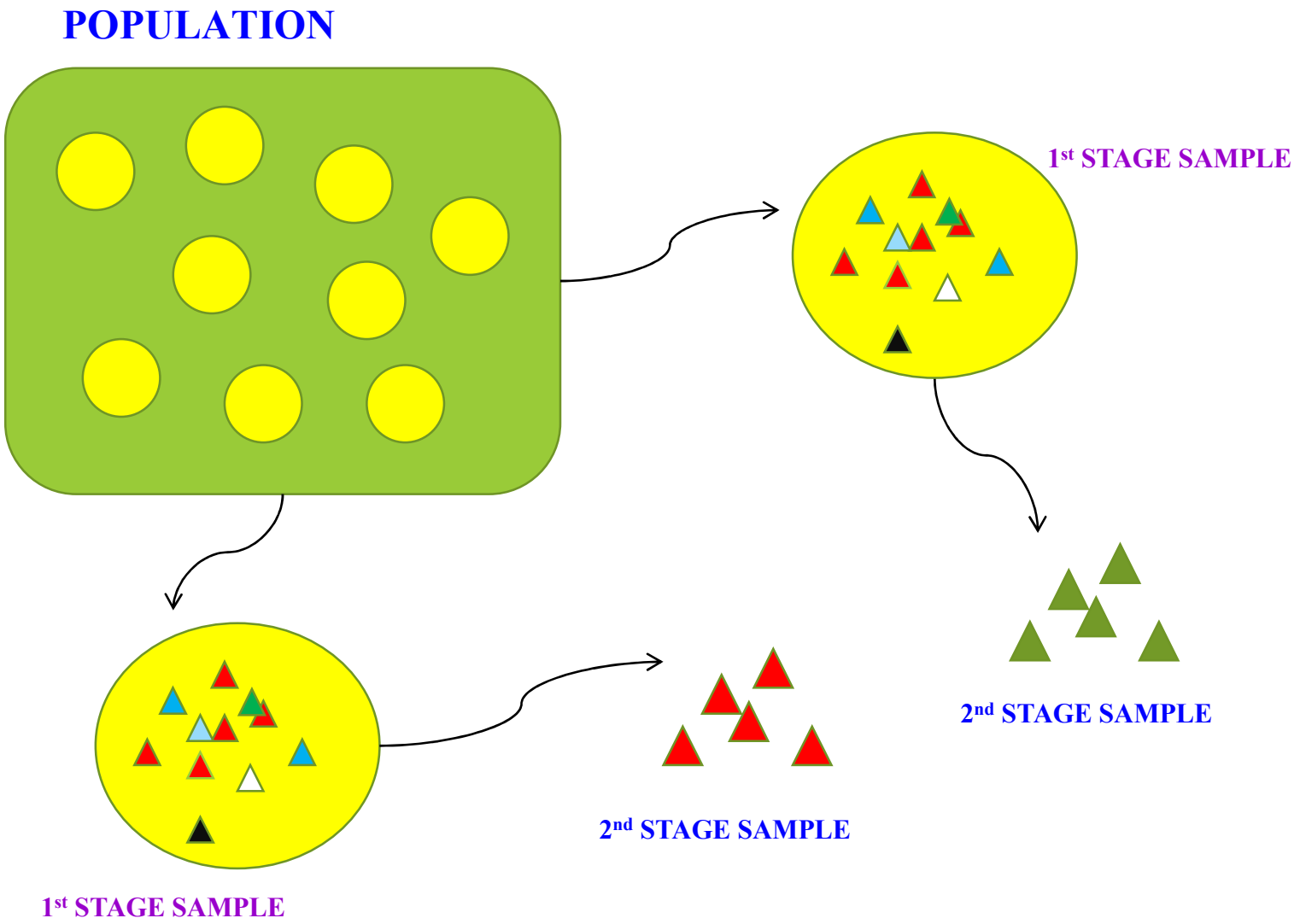
Purpose of stratified: to limit confounding factors
i.e.: breed, sex, parity, lactation, age



Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Estimate prevalence
  - Observational study

Two Stage Sample



$$n = \frac{p \times (1-p) \times D \times z^2}{e^2 \times b}$$

Where:

n	The number of clusters that have to be selected.
p	The expected prevalence.
D	The design effect.
z	The critical value obtained from a standard normal distribution. For each level of confidence there is a corresponding value of z. The levels of confidence frequently used in biological studies are 90%, 95%, and 99%. The corresponding z values are 1.64, 1.96, and 2.58 respectively.
e	The acceptable absolute error.
b	The number of individuals to select per cluster.

The design effect (D) is the ratio of the standard error using a two-stage design to the standard error on sample random sampling.

$$D = \frac{SE_{TwoSt}}{SE_{SimpleRnd}}$$

b = number of individuals (animals) to select per cluster (farm, village)

Number of samples to be taken per cluster depends on the analysis of operative factors and available resources.

It is convenient to be ≥ 5.

The smaller the number of samples to take per cluster, the greater the number of cluster to be selected.

Disease	Prevalence		Number clusters	Design effect	Rho
	%	n			
Enzootic bovine leucosis	1.51	2907	104	3.52	0.09
Enzootic bovine leucosis	11.75	945	81	2.11	0.10
Enzootic bovine leucosis	1.93	466	90	1.34	0.08
Infectious bovine rhinotracheitis	31.97	2852	104	2.76	0.07
Infectious bovine rhinotracheitis	47.88	969	82	1.71	0.07
Infectious bovine rhinotracheitis	28.11	466	90	2.62	0.39
Bovine virus diarrhoea	6.30	2799	108	6.95	0.23
Bovine virus diarrhoea	19.07	970	82	5.74	0.42
Bovine virus diarrhoea	69.74	466	90	2.76	0.42
Newcastle disease	37.89	1470	253	1.89	0.18
Infectious bursal disease	41.56	1470	253	2.56	0.37
Leptospira hardjo	38.55	2861	104	2.54	0.06
Leptospira icterohaemorrhagica	13.60	2861	104	4.24	0.12
Leptospira grippotyphosa	16.57	2861	104	3.91	0.11
Leptospira canicola	5.38	2861	104	3.04	0.08
Brucella abortus	7.74	1512	104	2.18	0.09
Brucella ovis	11.71	1529	40	6.94	0.16
Brucella ovis	10.99	1529	40	9.20	0.22
Anaplasma marginale	3.78	2909	104	2.19	0.04
Anaplasma marginale	4.32	1111	91	2.11	0.10
Trypanosoma vivax	2.75	2909	104	2.56	0.06
Trypanosoma vivax	30.87	1111	91	2.68	0.15
Trypanosoma congolense	23.94	1111	91	2.51	0.13

$$n = \frac{p \times (1-p) \times D \times z^2}{e^2 \times b}$$

The rate of homogeneity (roh)

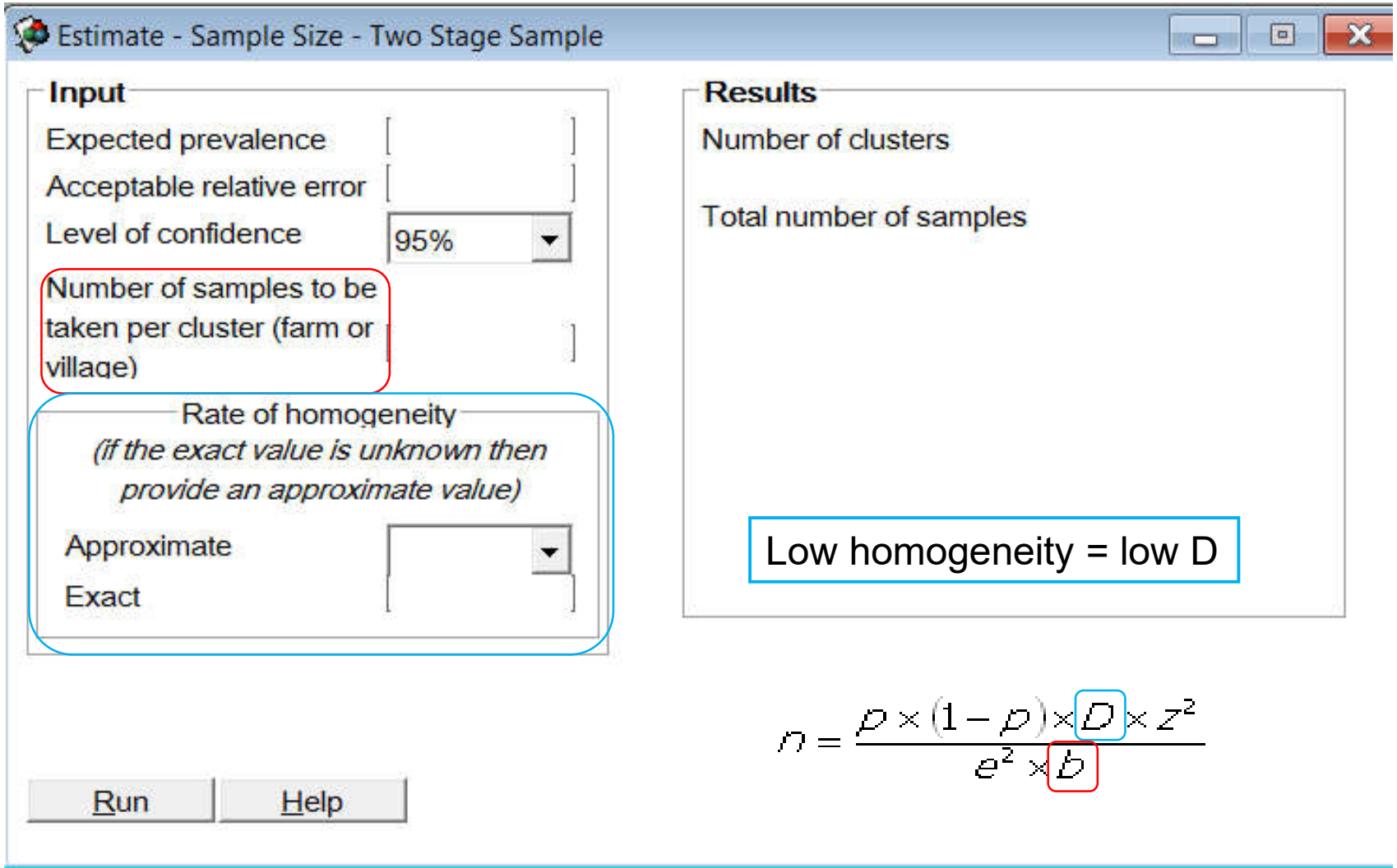
$$rho = \frac{D-1}{m-1}$$

m = the average number of individuals per cluster

Condition for implementing two-stage sample is best when rho is small (<1).

Rho > 4 → SRS

Two Stage Sample



Content/Outline

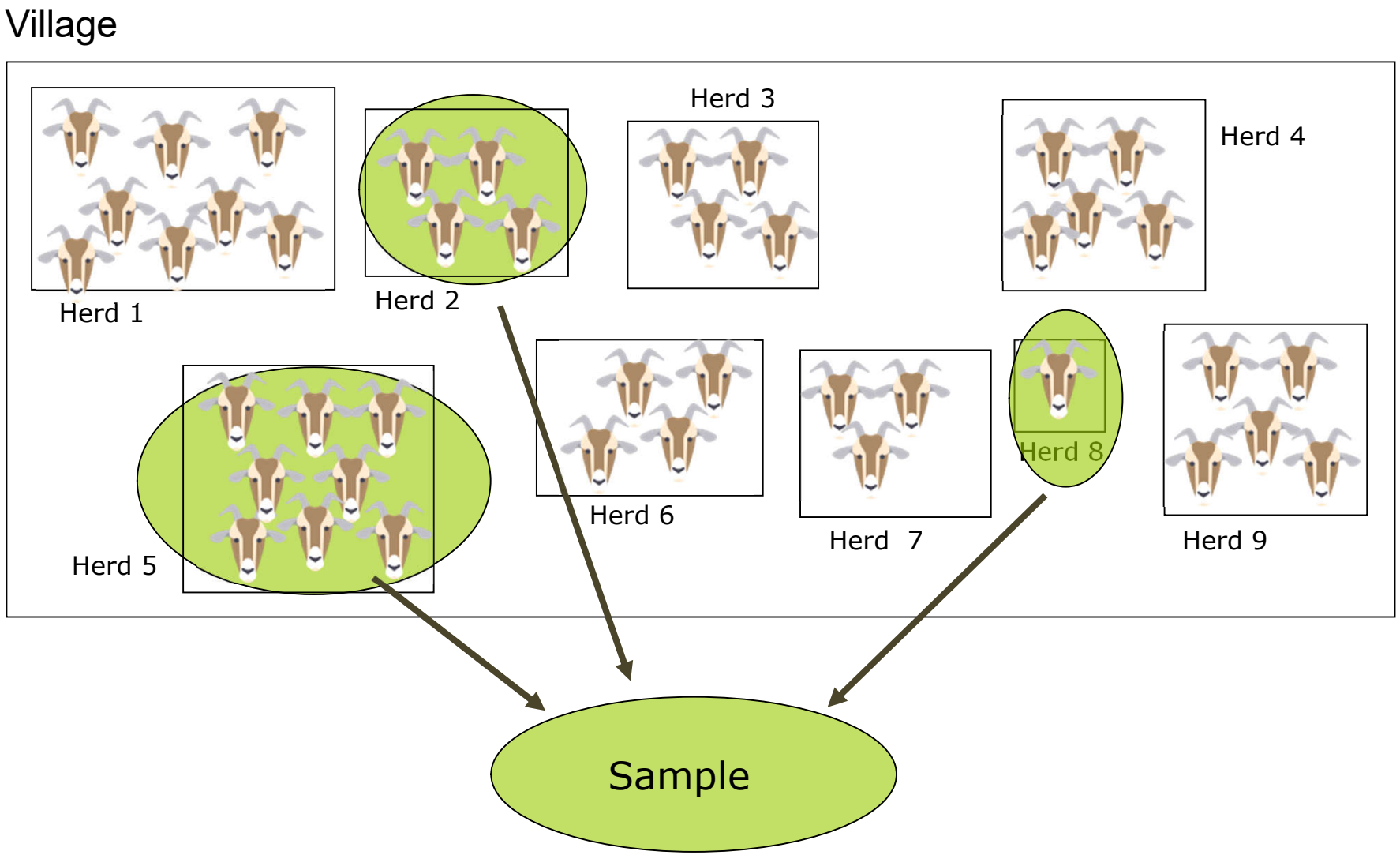
- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Cluster sampling

5. Cluster sampling

- Cluster : group of animals, of herds, ...
  - ex : villages, herds, flocks, ...
- Random sampling of clusters, then include the individuals of the clusters
- **All individuals** of the chosen cluster are present in the sample

5. Cluster sampling



Advantages and drawbacks of cluster sampling

- Limit travelling costs
- Loss of efficiency : individuals from a cluster are more similar than those of different clusters
  - Need to take it into account for statistical analysis
- It is better to have a lot of small clusters than a few big clusters !
  - To maintain variability
- Sampling size calculation is more tricky
  - Design effect

Cluster effect results in a loss of precision

Cluster effect (DEFF)

- *DEFF = Design Effect*
  - **« correcting factor»** takes into account the heterogeneity of the clusters for the considered indicator

- The higher is the expected prevalence, the higher is the DEFF
- The bigger is the number of individuals, the higher is the DEFF

Cluster effect (DEFF) : OpenEpi

Sample Size for % Frequency in a Population (Random Sample)		
Population size	1000000	If large, leave as one million
Anticipated % frequency(p)	30	Between 0 & 99.99. If unknown, use 50%
Confidence limits as +/- percent of 100	5	Absolute precision %
Design effect (for complex sample surveys--DEFF)	1	1.0 for random sample

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*):	30%+/-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%	323
80%	138
90%	228
97%	396
99%	558
99.9%	909
99.99%	1271

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.

Cluster effect (DEFF)

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*):	30%+/-5
Confidence limits as % of 100(absolute +/- %)(*d*):	5%
Design effect (for cluster surveys-*DEFF*):	1.5

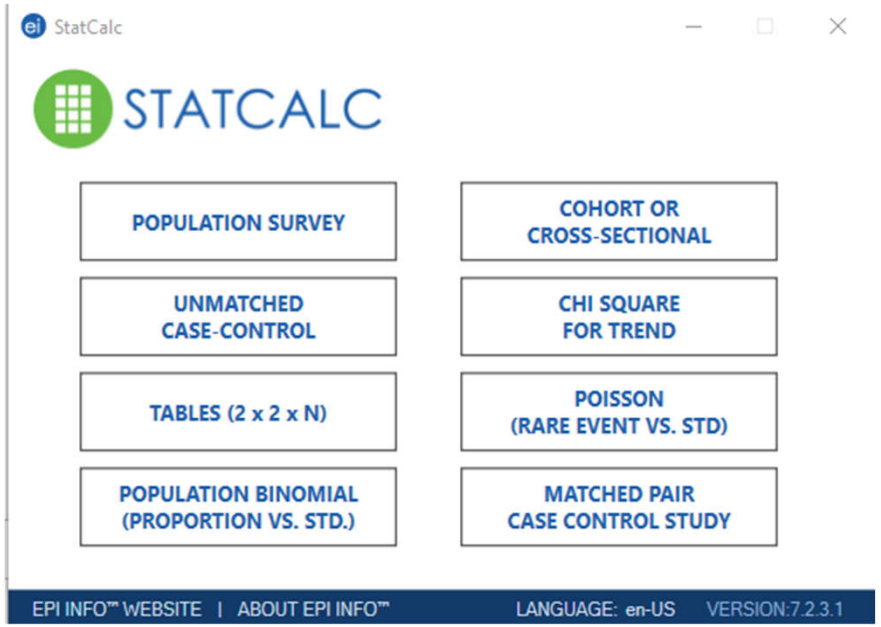
Sample Size(*n*) for Various Confidence Levels

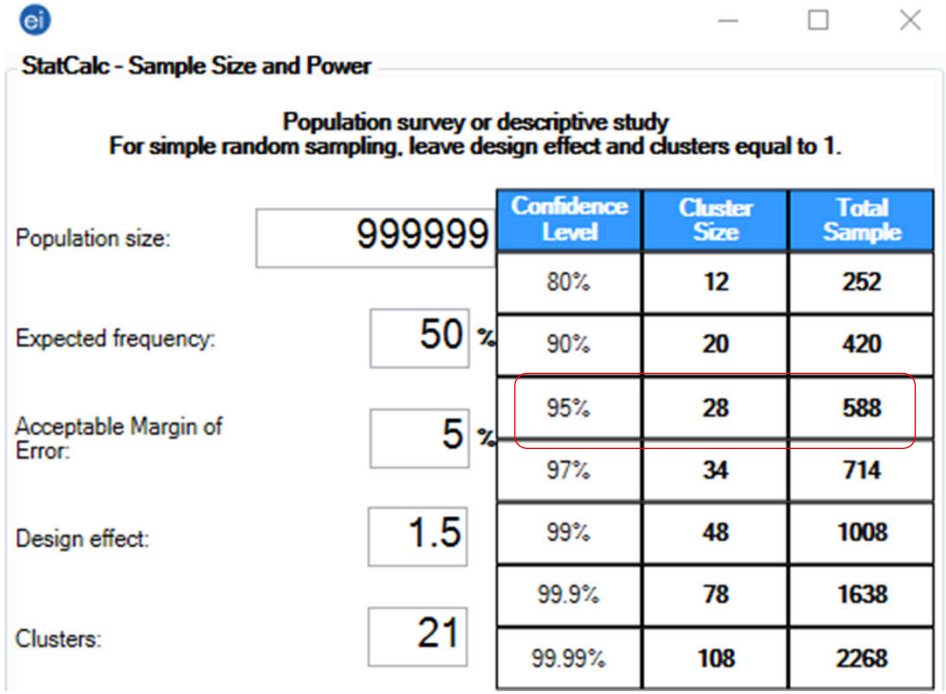
ConfidenceLevel(%)	Sample Size
95%	484
80%	207
90%	341
97%	594
99%	836
99.9%	1364
99.99%	1906

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.


StatCalc
STATCALC
POPULATION SURVEY
COHORT OR CROSS-SECTIONAL
UNMATCHED CASE-CONTROL
CHI SQUARE FOR TREND
TABLES (2 x 2 x N)
POISSON (RARE EVENT VS. STD)
POPULATION BINOMIAL (PROPORTION VS. STD.)
MATCHED PAIR CASE CONTROL STUDY
EPI INFO™ WEBSITE | ABOUT EPI INFO™
LANGUAGE: en-US
VERSION:7.2.3.1


StatCalc - Sample Size and Power
Population survey or descriptive study
For simple random sampling, leave design effect and clusters equal to 1.
Population size: 999999
Expected frequency: 50 %
Acceptable Margin of Error: 5 %
Design effect: 1.5
Clusters: 21
Confidence Level
Cluster Size
Total Sample
80% 12 252
90% 20 420
95% 28 588
97% 34 714
99% 48 1008
99.9% 78 1638
99.99% 108 2268

Disease	Prevalence		Number clusters	Design effect
	%	n		
Enzootic bovine leucosis	1.51	2907	104	3.52
Enzootic bovine leucosis	11.75	945	81	2.11
Enzootic bovine leucosis	1.93	466	90	1.34
Infectious bovine rhinotracheitis	31.97	2852	104	2.76
Infectious bovine rhinotracheitis	47.88	969	82	1.71
Infectious bovine rhinotracheitis	28.11	466	90	2.62
Bovine virus diarrhoea	6.30	2799	108	6.95
Bovine virus diarrhoea	19.07	970	82	5.74
Bovine virus diarrhoea	69.74	466	90	2.76
Newcastle disease	37.89	1470	253	1.89
Infectious bursal disease	41.56	1470	253	2.56
Leptospira hardjo	38.55	2861	104	2.54
Leptospira icterohaemorrhagica	13.60	2861	104	4.24
Leptospira grippotyphosa	16.57	2861	104	3.91
Leptospira canicola	5.38	2861	104	3.04
Brucella abortus	7.74	1512	104	2.18
Brucella ovis	11.71	1529	40	6.94
Brucella ovis	10.99	1529	40	9.20
Anaplasma marginale	3.78	2909	104	2.19
Anaplasma marginale	4.32	1111	91	2.11
Trypanosoma vivax	2.75	2909	104	2.56
Trypanosoma vivax	30.87	1111	91	2.68
Trypanosoma congolense	23.94	1111	91	2.51
Trypanosoma brucei	24.39	1111	91	2.39
Eimeria spp.	27.13	1010	104	2.53
Eimeria spp.	15.63	1113	91	4.32

www.who.int › advisory_bodies › meetings › generic_t... XLS

calculation design effect

calculation design effect. A, B, C, D, E. 1, DESIGN EFFECT: 2, The ...

$$\text{Variance (SRS)} = \frac{\text{mean prevalence} * (1-\text{mean prevalence})}{(\text{sample size}-1)}$$

$$\text{Cluster variance} = \frac{\text{Sum of squares of (prevalence - mean prevalence)}}{\text{square of (number of districts - 1)}}$$

$$\text{Design effect (DEFF)} = \frac{\text{Cluster variance}}{\text{SRS variance}}$$

$$\text{Variance (SRS)} = \frac{\text{mean prevalence} * (1-\text{mean prevalence})}{(\text{sample size}-1)}$$

district	sample size	no. positive TST	prevalence of infection
a	1200	98	0.0817
b	1100	81	0.0736
c	1050	72	0.0686
d	950	108	0.1137
e	1000	101	0.1010
f	900	87	0.0967
g	1150	83	0.0722
h	950	78	0.0821
i	1000	111	0.1110
j	900	98	0.1089
total	**10200**	**917**	**0.0899**

Mean prevalence = 917/10200 = 0.0899

Sample size = 10200

$$\text{Cluster variance} = \frac{\text{Sum of squares of (prevalence - mean prevalence)}}{\text{square of (number of districts - 1)}}$$

district	sample size	no. positive TST	prevalence of infection	square of (district prevalence - mean prevalence)
a	1200	98	0.081666666667	0.00006782
b	1100	81	0.073636363636	0.00026457
c	1050	72	0.068571428571	0.00045499
d	950	108	0.113684210526	0.00056560
e	1000	101	0.101000000000	0.00012317
f	900	87	0.096666666667	0.00004576
g	1150	83	0.072173913043	0.00031428
h	950	78	0.082105263158	0.00006079
i	1000	111	0.111000000000	0.00044513
j	900	98	0.108888888889	0.00036050
total	10200	917		
	number of districts =		mean prevalence =	sum of squares =
	10		0.089901960784	0.00270261

(0.0899- 0.0816)^2

917/10200 = 0.0899

	Simple Random Sampling	Stratified	Cluster
(+)	- Representative - Calculation and analysis = simple	- Gain in precision => ↘ sampling size	- Not need of a sampling frame - Save time and costs
(-)	- Time and travelling costs	- Analysis more complex	- Cluster effect (lack of variability) - Analysis more complex (design effect)

- Multi-stage sampling :
  - Combines advantages from different methods
  - But statistical analysis is more complex

Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence determination
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Observational studies

Case – control study

$$n = \frac{pq\left(1+\frac{1}{c}\right)\left(z_{\alpha/2} - z_{\beta}\right)^2}{\left(p_1 - p_2\right)^2}$$

n : number of cases

c : ratio of controls / cases

p_1 : proportion of controls exposed

p_2 : proportion of cases exposed

$z_{\alpha/2}$: z value for α; for α=5%, $Z_{\alpha/2}$ = 1.96

z_{β} : z value for power (1-β) ; often power = 80%, z_{β}= 0.84

n depends on :

- **Difference of proportion in the groups**
- **Frequence of exposures to the factor**

Case – control study

$$N_{field} = n + 10\% \, n$$

To anticipate the problems : no answers, losses, farmer who quit…

Case – control study

- Risk of having an accident when crossing a highway closing the eyes
  - hazard = important when a car hits you
  - risk = happens quite often

➔ Small « n »

- Risk of having an accident when crossing a country road closing the eyes
  - hazard = important when a car hits you
  - risk = happens very rarely

➔ Big « n » : you'll need a lot of crossing before finding a case

Online tools

- 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
  - Google--Internet
  - PubMed--MEDLARS
- Internet Links
- Download OpenEpi
- Development

http://www.openepi.com/Menu/OE_Menu.htm

Sample Size for Unmatched Case Control Study

Two-sided confidence level	95	(1-alpha) usually 95%
Power(% chance of detecting)	80	Usually 80%
Ratio of Controls to Cases	1.0	For equal samples, use 1.0
Percent of controls exposed	20	Between 0.0 and 99.99
Please fill in one of the following. The other will be calculated.		
Odds ratio	3	
Percent of cases with exposure		Between 0.0 and 99.99

Sample Size for Unmatched Case-Control Study

For:

Two-sided confidence level(1-alpha)	95
Power(% chance of detecting)	80
Ratio of Controls to Cases	1
Hypothetical proportion of controls with exposure	20
Hypothetical proportion of cases with exposure:	42.86
Least extreme Odds Ratio to be detected:	3.00

	Kelsey	Fleiss	Fleiss with CC
Sample Size - Cases	65	64	73
Sample Size - Controls	65	64	73
Total sample size:	130	128	146

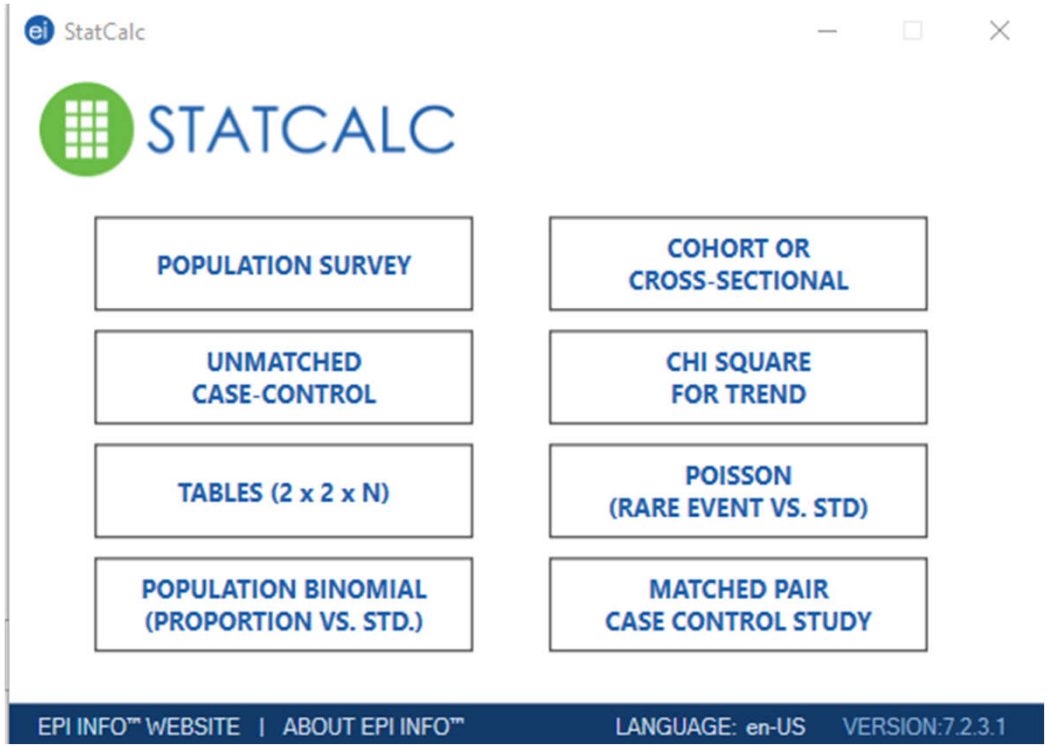
References

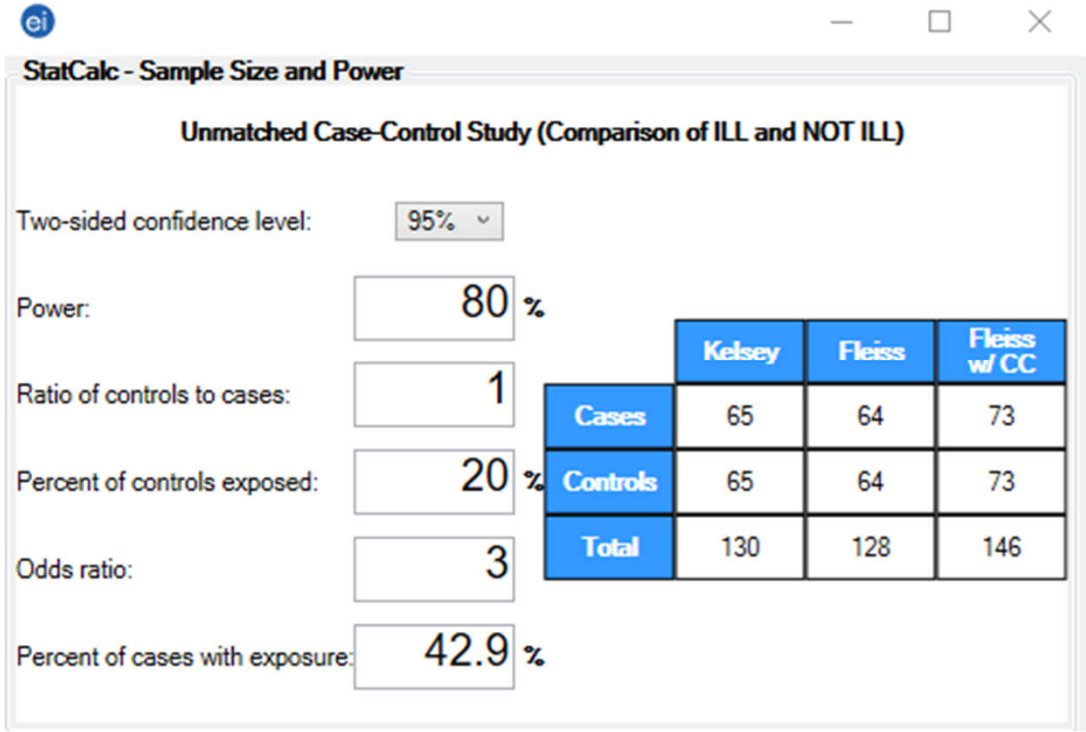
Kelsey et al., Methods in Observational Epidemiology 2nd Edition, Table 12-15
Fleiss, Statistical Methods for Rates and Proportions, formulas 3.18 &3.19

CC = continuity correction
Results are rounded up to the nearest integer.
Print from the browser menu or select, copy, and paste to other programs.

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

epiinfo™


StatCalc
STATCALC
POPULATION SURVEY
COHORT OR CROSS-SECTIONAL
UNMATCHED CASE-CONTROL
CHI SQUARE FOR TREND
TABLES (2 x 2 x N)
POISSON (RARE EVENT VS. STD)
POPULATION BINOMIAL (PROPORTION VS. STD.)
MATCHED PAIR CASE CONTROL STUDY
EPI INFO™ WEBSITE | ABOUT EPI INFO™
LANGUAGE: en-US
VERSION:7.2.3.1


StatCalc - Sample Size and Power
Unmatched Case-Control Study (Comparison of ILL and NOT ILL)
Two-sided confidence level: 95%
Power: 80 %
Ratio of controls to cases: 1
Percent of controls exposed: 20 %
Odds ratio: 3
Percent of cases with exposure: 42.9 %
Kelsey
Fleiss
Fleiss w/ CC
Cases 65 64 73
Controls 65 64 73
Total 130 128 146

Content/Outline

- Overview / purpose / sampling
- Sample size
  - Disease detection
    - Simple random sampling
    - Two-stage sampling
  - Freedom from a disease
  - Prevalence
    - Simple random sampling
    - Stratified random sampling
    - Two-stage sampling
    - Cluster sampling
  - Observational study
- Estimate prevalence

Estimate prevalence confidence interval

Estimate prevalence / confidence interval: Stratified random sample

$$p = \frac{\sum_{i=1}^{e} (N_i \times p_i)}{N}$$

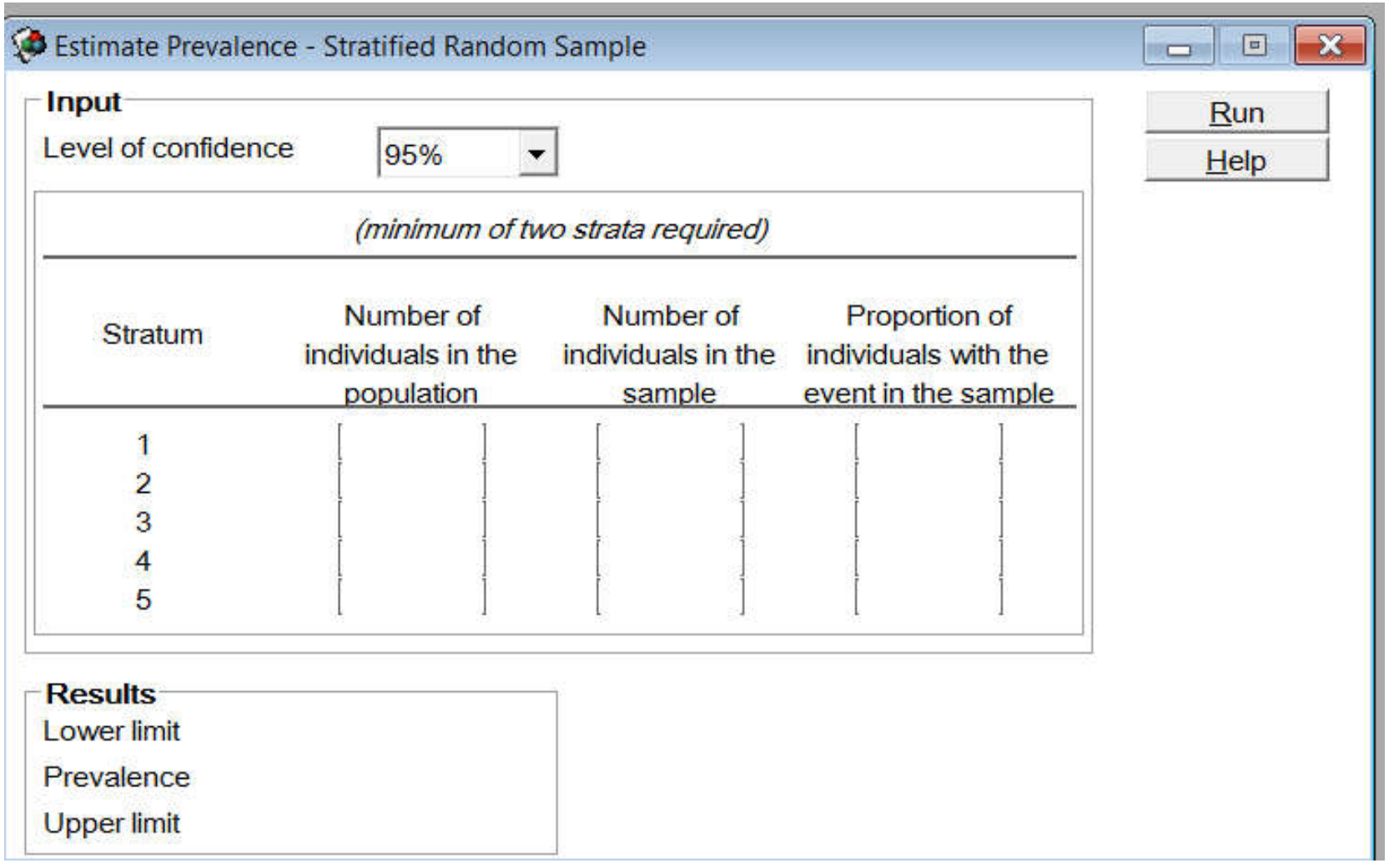
$$CI = p \pm Z * SE$$

$$SE = \frac{\sum_{i=1}^{e} \left[(N_i)^2 \times \left(\frac{N_i - n_i}{N_i} \right) \times \left(\frac{p_i \times (1 - p_i)}{n_i - 1} \right) \right]}{N^2}$$

Where:

e	The number of defined strata.
N_i	The number of individuals from the population that form part of strata i.
n_i	The number of individuals from the sample that form part of strata i.
p_i	The estimated proportion of individuals with the event in strata i.
N	The number of individuals in the population.

Estimate prevalence / confidence interval: Stratified random sample



Input

Level of confidence: 95%

(minimum of two strata required)

Stratum	Number of individuals in the population	Number of individuals in the sample	Proportion of individuals with the event in the sample
1	1000	20	0.25
2	600	20	0.5
3			
4			
5			

Results

Lower limit	**0.1975**
Prevalence	**0.3438**
Upper limit	**0.4900**

Estimate prevalence / confidence interval: Two-stage sample

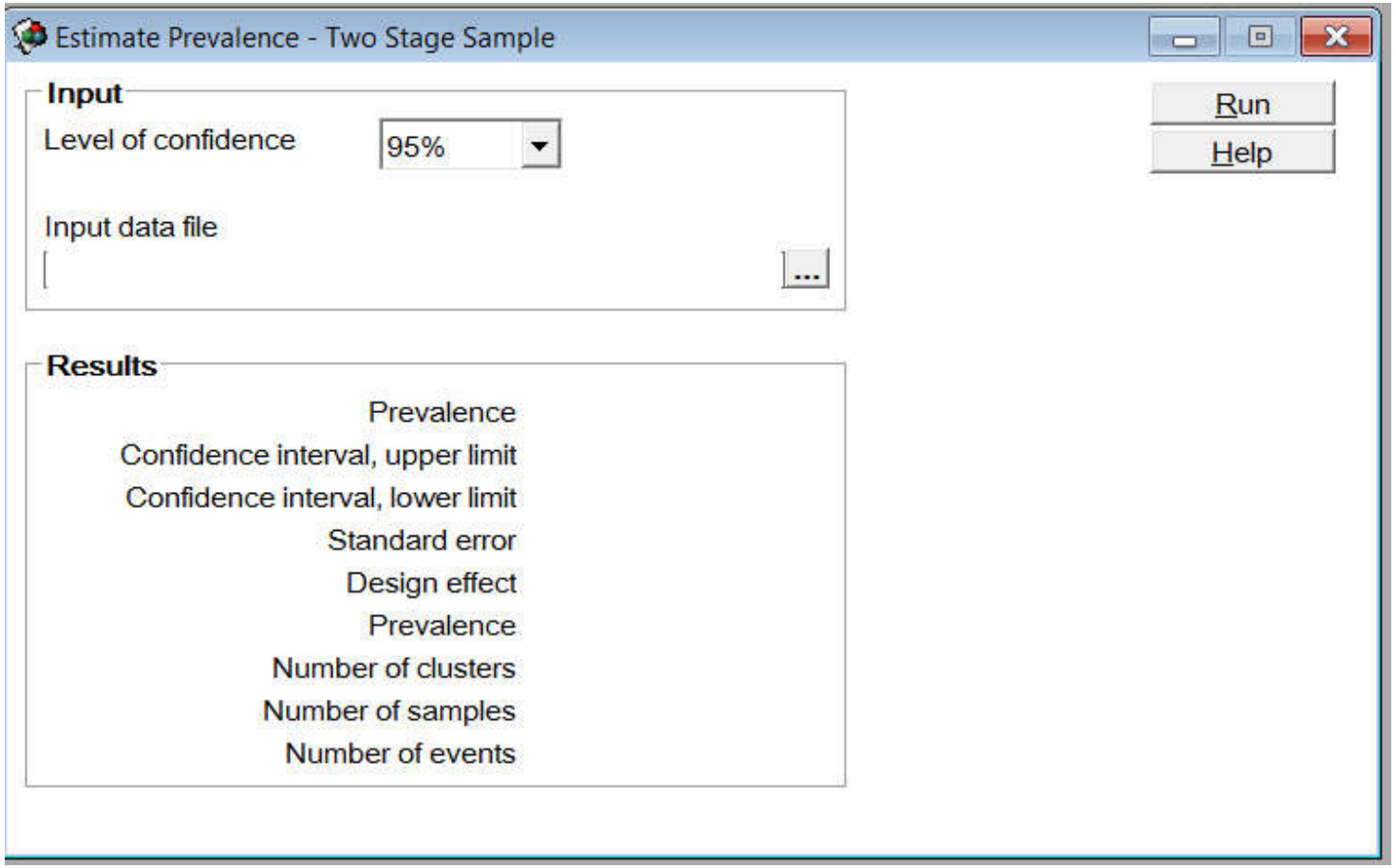
$$p = \frac{\text{number of individuals with the event}}{\text{number of individuals in the sample}}$$

$$SE_{TwoStages} = \frac{c}{n_{Total}} \times \sqrt{\frac{\sum_{i=1}^{c} (e_i)^2 - 2 \times p \times \sum_{i=1}^{c} (n_i \times e_i) + p^2 \times \sum_{i=1}^{c} (n_i)^2}{c \times (c-1)}}$$

Where:

c	The number of clusters included in the sample.
n_{total}	The total number of individuals included in the sample.
n_i	The number of individuals in the sample belonging to cluster i.
e_i	The number of individuals with the event in the sample belonging to cluster i.
p	The estimated proportion of individuals with the event in the population.

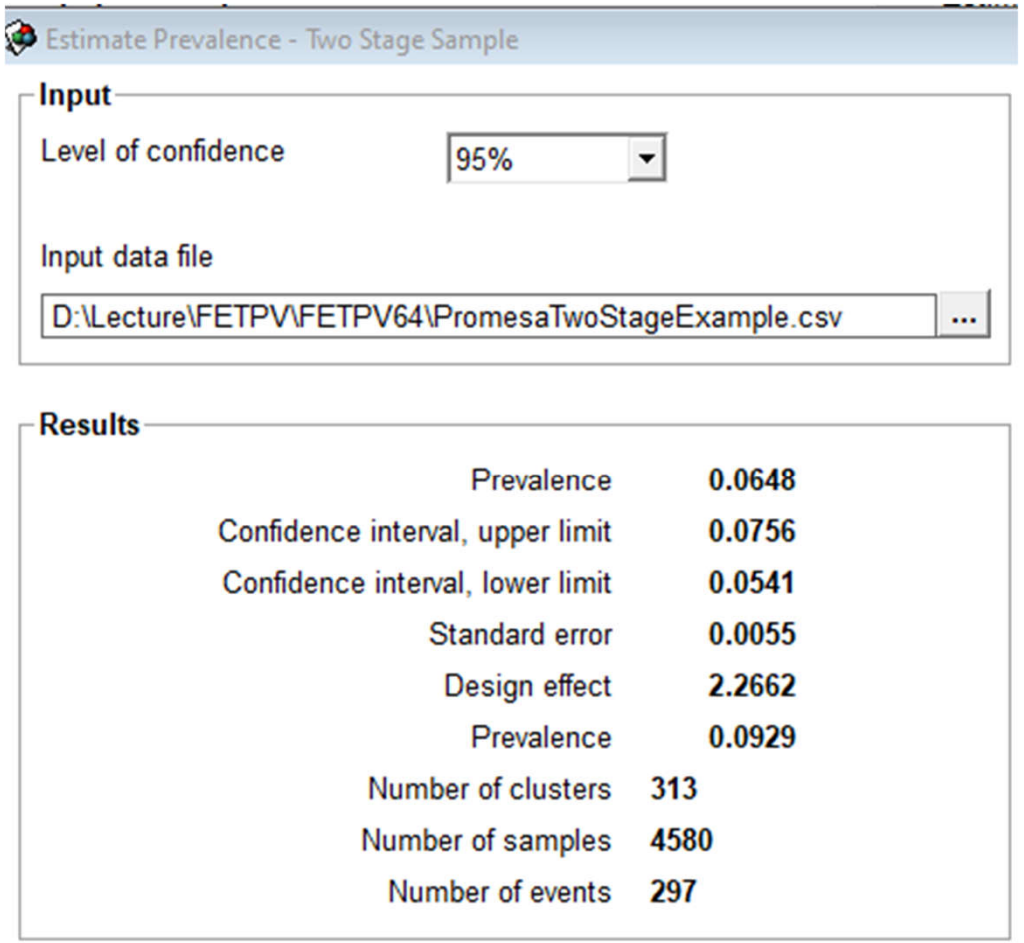
Estimate prevalence / confidence interval: Two-stage sample



ProMESA: Data file – prev. for 2-stage sample

- Input data file
  - .csv (comma separated value)
  - Requires 3 columns
  - Cluster ID, # of samples taken in the clusters, # of events detected in the clusters

	A	B	C
1	ID	# samples	# positives
2	1	13	1
3	2	9	1
4	3	44	1
5	4	23	1
6	5	12	1
7	6	21	1
8	7	19	1
9	8	15	1
10	9	12	1
11	10	42	1
12	11	20	1
13	12	11	1
14	13	27	1
15	14	38	1
16	15	30	1
17	16	30	1
18	17	7	1
19	18	2	1



Prevalence is 0.065
95% confidence interval = 0.05 – 0.075

Random sampling proportional to the size

- Needs an information about the size of each unit (ex : herd size)
- The bigger is the unit, the higher is the probability of being included in the sample
- Improves sample efficiency