EPPO Workshop for inspectors on risk-based sampling and inspection

April 26th – 28th 2023

EUROPEAN FOOD SAFETY AUTHORITY

RISK-BASED SAMPLING: THE STATISTICS BEHIND THE SCENE

José Cortinas Abrahantes

efsa
EUROPEAN FOOD SAFETY AUTHORITY

OUTLINE

- Basic Concepts and definitions used in Surveillance
  - Why is it needed?
  - General concepts used to design a surveillance
  - Definitions of several parameter used to design a surveillance
- Risk based sampling
  - Definitions of risk and surveillance
  - Probability concepts and derivations
  - Demonstrating pest freedom
  - Risk based surveillance
- Conclusions

Why is Sampling needed?

- **Feasibility**, not easy to gather information about the whole population
- **Time constrains**
- **Cost-effective**
- Gather information about a **portion** of the **population** and still be able **to infer** about the **whole population**
- Able to quantify **accuracy** mathematically

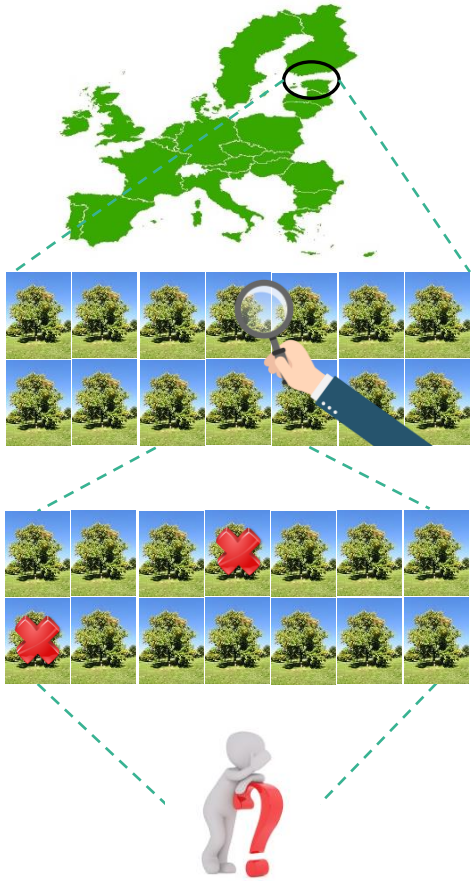
BASIC CONCEPTS AND DEFINITIONS: DESIGNING A SURVEILLANCE

What do you want to **generalized** to?

How will you assess the population?

What level can we live with?

How confident we want to be?



Target Population

Method use to detect the pest

Design prevalence

Confidence level

BASIC CONCEPTS AND DEFINITION

- Confidence level (CL): A measure of reliability of the survey. It states the confidence that a pest is absent from the surveyed area, or that its true prevalence is below the design prevalence.
- Design prevalence (DP, Tolerance level, ISPM31): Analogous to the term level of detection used in "Methodologies for sampling of consignments". The minimum pest prevalence that a survey will detect with a given confidence level.
- Detection method: The combination of methods applied from the field to the laboratory to determine if a pest is present. The detection method may include a sequence of operations such as the visual examination, trapping, sampling and testing.
- Diagnostic sensitivity (Se, Efficacy of detection, ISPM31): The probability to detect the pest in the sample, based on a specific sampling or diagnostic protocol that has been followed, given that the pest is present.
- Sampling effectiveness (Sef): The probability to collect infected/infested material in the sample of an inspection unit given that the unit is infected/infested.
- Method sensitivity (MeSe): The product between Sef and Se.

BASIC CONCEPTS AND DEFINITION

- Inspection unit (link to Sample unit, ISPM31: The elementary units on which the detection method is applied (e.g. plants, plant parts, commodities, pest vectors that are examined for detection of a pest) as part of a survey.
- Target population: The set of individual plants, commodities or vectors in which the target pest can be detected within an area of interest. The size of the target population (**N**) corresponds to the number of inspection units it contains and its geographic boundary.
- Epidemiological unit: A homogeneous area where the interactions between the pest, the host plants, the abiotic and biotic factors and conditions would result into the same epidemiology, should the pest be present.
- Risk factor: A biotic or abiotic factor (e.g. related to environment, ecosystem or a human activity) that increases the probability of infestation of an epidemiological unit by the pest .

RISK BASED SAMPLING: RISK AND SURVEILLANCE

- **Risk:** The **probability** that an event is taken place, considering also the consequences of that event given that it has happened.
  - Examples:
    - Species (susceptibility, speed of propagation)
    - Production systems (Nurseries, garden centres, Agricultural field)
    - Place of origin of a commodity
    - Spatial context

- **Surveillance:** Is the continuous investigation of a population to detect the occurrence of a pest for control purpose and in general involve testing part of the population.

- **Surveillance types: Active** or **Passive**

RISK BASED SAMPLING: PROBABILITY CONCEPTS

- **Risk-based surveillance: Looking** for something where it is **mostly likely to be found**.
- If we want to use **risk-based sampling**, you must know some **risk factors** associated to what you are looking for.
- **Risk-based surveillance** involves using knowledge of **risk factors** to improve the probability to find the pest, it uses **probability concepts**.
- **Risk-based surveillance** uses the differential risk of units in the population to increase the probability of detection.

RISK BASED SAMPLING: PROBABILITY CONCEPTS

- The probability of an event A is denoted as $P(A)$ and similarly for event B $(P(B))$, below some illustrations of probability rules assuming independence:

  - AND $\Longrightarrow$ $P(A \text{ AND } B) = P(A) \times P(B)$ (violations: $P(A \text{ AND } B) = P(A) \times P(B|A)$)
  - OR $\Longrightarrow$ $P(A \text{ OR } B) = P(A) + P(B) - P(A) \times P(B)$ (violations: $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$)
  - SUM $\Longrightarrow$ $\sum_{i=1}^{N} P(i) = 1$
  - NOT $\Longrightarrow$ $P(\text{NOT } A) = 1 - P(A)$

RISK BASED SAMPLING: PROBABILITY CONCEPTS

- Considering that the probability is denoted as $P(A)$, below some probability examples and rules used in probability:
  - Considering rolling a  , what is the probability to get a 3, $P(3)$?
  - **Answer:** $P(3) = \frac{1}{6}$
  - Considering rolling a  , what is the probability to not get a 3, $P(\text{not } 3)$?
  - **Answer:** $P(\text{NOT } 3) = P(1) + P(2) + P(4) + P(5) + P(6) = 1 - P(3) = \frac{5}{6}$
  - What would be the sum of all potential probabilities when rolling a 
  - **Answer:** $\sum_{i=1}^{6} P(i) = 1$

RISK BASED SAMPLING: PROBABILITY CONCEPTS

- Other examples of probability rules:
  - Considering rolling a  twice, what is the probability to get first a 3 and then a 6?
  - **Answer:** $P(3 \text{ AND } 6) = P(3) \times P(6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$
  - Considering rolling a  , what is the probability to get a 3 or a 6?
  - **Answer:** $P(3 \text{ OR } 6) = P(3) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

RISK BASED SAMPLING: PROBABILITY CONCEPTS

- Other examples of probability rules:
  - Considering that the objective is detecting a pest in specific host plant, and the probability to detect the pest is 10%, what is the probability of detecting the pest in at least one plant if I sample and test 8 plants?
  - **Answer:** $P(detect) = 0.1$

$$P(\text{NOT } detect) = 1 - P(detect) = 1 - 0.1 = 0.9$$

$$P(\text{NOT } detect \text{ in all } 8 \text{ plants}) = P(\text{NOT } detect \text{ in plant } 1) \times \cdots \times P(\text{NOT } detect \text{ in plant } 8)$$

$$= \left(1 - P(detect)\right)^8$$

$$P(detect \text{ in at least } 1 \text{ out of } 8 \text{ plants}) = 1 - \left(1 - P(detect)\right)^8$$

$$= 1 - (1 - 0.1)^8 = 1 - 0.9^8 = 1 - 0.43 = 0.57$$

RISK BASED SAMPLING: PROBABILITY CONCEPTS

- Other examples of probability rules:
  - If the prevalence of a pest is 15% and the test used to detect the pest has a sensitivity of 85%, what is the probability of getting a positive test results if we select a random unit in the population?
  - **Answer:** $P(infected) = \mathrm{P(D+)} = \text{prevalence} = 0.15$

$$P(test\ positive) = \text{sensitivity} = 0.85 \Longrightarrow P(T+|D+)$$

$$P(infected\ \text{AND}\ test\ positive) = P(infected) \times P(test\ positive)$$

$$= P(D+) \times P(T+|D+)$$

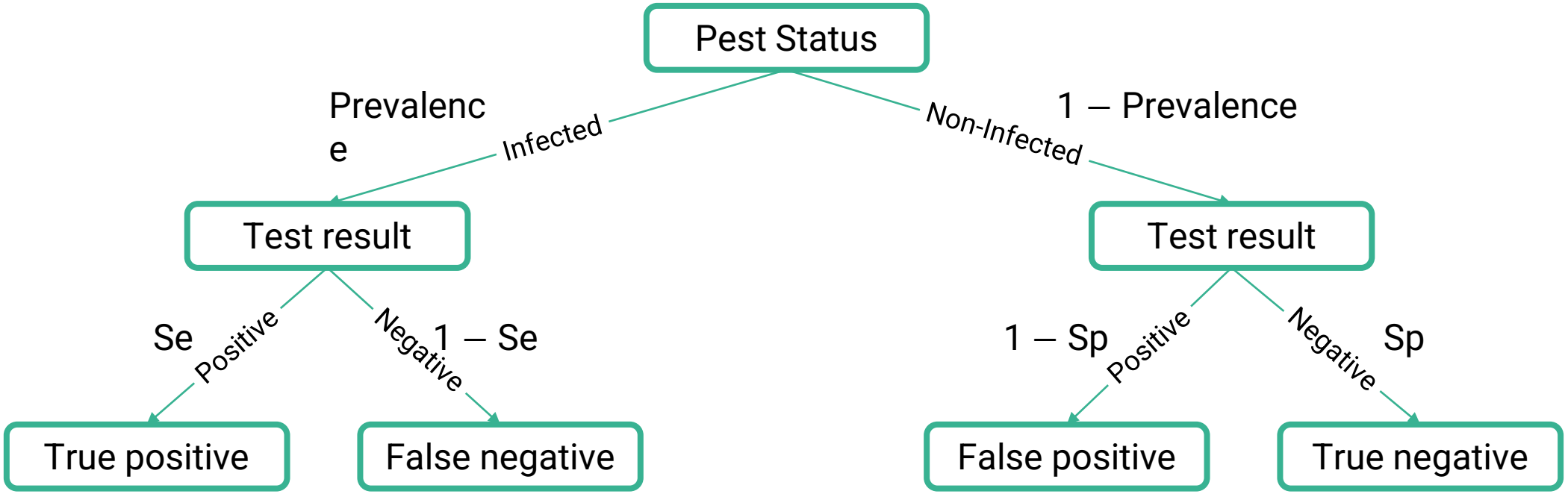
$$= 0.15 \times 0.85 = 0.1275$$

  - If the Specificity would be known then, we could also calculate:

$$P(D+|T+) = \frac{P(D+) \times P(T+|D+)}{P(D+) \times P(T+|D+) + (1 - P(D)) \times \left(1 - P(T-|D-)\right)} = \frac{P \times Se}{P \times Se + (1-P) \times (1 - Sp)}$$

RISK BASED SAMPLING: PROBABILITY CONCEPTS

- Some graphical presentation of specific probability rules:



RISK BASED SAMPLING: DEMONSTRATING PEST FREEDOM

- Considering simple random sampling in the population (e.g. host, commodities):
  - The confidence of a surveillance conducted that sampled and tested n units in the population considering specific tolerance level (design prevalence, DP) can be calculated as follow:
  - **Answer:** $P(infected) = \mathrm{P(D+)} = DP$

$$P(\textbf{NOT } infected) = 1 - DP$$

$$P(2 \text{ units } \textbf{NOT } infected) = (1 - DP) \times (1 - DP)$$

$$P(\text{ALL } n \text{ units } \textbf{NOT } infected) = (1 - DP)^n$$

$$P(\text{at least 1 unit } \textbf{IS } infected) = 1 - (1 - DP)^n = CL$$

  - If the method used to detect is imperfect ($Se < 1$) then the confidence would be:
  - **Answer:** $CL = 1 - (1 - Se \times DP)^n \Longrightarrow n = \frac{Log(1-CL)}{Log(1-Se \times DP)}$
  - In case that population size (N) is not very large:
  - **Answer:** $CL = 1 - \left(1 - \frac{n \times Se}{N - \frac{1}{2}(N \times Se \times DP - 1)}\right)^{N \times DP} \Longrightarrow n \cong \frac{\left(1-(1-CL)^{\frac{1}{N \times DP}}\right) \times \left(N - \frac{1}{2}(N \times Se \times DP - 1)\right)}{Se}$

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

- To illustrate this concept we will use the following example:
  - A Risk based surveillance should be conducted to detect pest **X** in specific host, assuming that the design prevalence (acceptable tolerance level, $DP = 0.05$, i.e. 5% of the entire host population could be infected by the pest), and it is known that the host population can be subdivided into two risk groups, in which the high risk host plants are 2 times more likely to be infected than the low risk host plants, and the high risk plants represent 20% of the total host population. The aim is to conduct a risk based surveillance to detect the pest if present, considering the specified design prevalence DP.

Subgroup	Relative Risk	Proportion
High risk	2	0.2
Low risk	1	0.8

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

➢ **Answer:** Three options could be used to conduct the risk based surveillance:

1. Conduct a surveillance considering simple random sampling ignoring the differential risk in the host population.
2. Conduct a surveillance considering both risk group, but ensuring that in both group the confidence level achieved to be the same using the procedure to combine evidence from the two groups $\left(OCL = 1 - (1 - CL_{HR}) \times (1 - CL_{LR})\right)$
3. Conduct a surveillance considering only the high risk group, since the likelihood to find it in that group is larger than for the rest of the hosts in the population under investigation.

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

- **Option 1:**

1. If the tolerable level (DP) in the entire population is 5%, a simple random sampling of the population could be conducted to detect the pest, without accounting for differential risk.

➢ **Answer:** Considering a perfect test then the sample size would be:

$$n = \frac{Log(1 - CL)}{Log(1 - DP)} = \frac{Log(1 - 0.95)}{Log(1 - 0.05)} = \frac{-2.996}{-0.051} \approx 58.4 \approx 59$$

Implying that if we sample randomly 59 host plants in the entire host population and all test results would be negative, then we could conclude pest freedom with 95% confidence, or even more accurate that if the pest **X** would be present, we could be 95% confidence that the prevalence in the population would be below 5%.

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

- **Option 2:**

1. If the tolerable level (DP) in the entire population is 5%, what would be the level for both risk groups, considering that the likelihood to be infected is twice as large in the high risk and they represent only 20% of the total population.

➢ **Answer:** The tolerable level for each risk group can be calculated using the relative risk and the proportion, using specific weighing:

$$W_{HR} = \frac{RR_{HR}}{RR_{HR} \times Prop_{HR} + RR_{LR} \times Prop_{LR}} = \frac{2}{2 \times 0.2 + 1 \times 0.8} = \frac{2}{1.2} \approx 1.67$$

$$W_{LR} = \frac{RR_{LR}}{RR_{HR} \times Prop_{HR} + RR_{LR} \times Prop_{LR}} = \frac{1}{2 \times 0.2 + 1 \times 0.8} = \frac{1}{1.2} \approx 0.83$$

$$DP_{HR} = W_{HR} \times DP = 1.67 \times 0.05 \approx 0.083$$

$$DP_{LR} = W_{LR} \times DP = 0.83 \times 0.05 \approx 0.042$$

$$\Longrightarrow DP = DP_{HR} \times Prop_{HR} + DP_{LR} \times Prop_{LR} = 0.083 \times 0.2 + 0.042 \times 0.8 = 0.05$$

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

- **Option 2:**

2. Now we can used the way to combine different evidence to calculate the confidence needed in each group (assuming being the same) ensuring a confidence level of 95% ($OCL = 0.95$).

➢ **Answer:** Considering a perfect test then the sample size would be:

$$OCL = 1 - (1 - CL_{HR}) \times (1 - CL_{LR}) = 1 - (1 - CL)^2$$

$$CL = 1 - \sqrt{1 - OCL} = 1 - \sqrt{1 - 0.95} \approx 0.78$$

3. Using the tolerable level for the each risk group (DP_{HR} and DP_{LR}) and the calculated $CL = 0.78$, the sample size for each group can be computed.

➢ **Answer:** Considering a perfect test then the sample size would be:

$$n_{HR} = \frac{Log(1 - CL)}{Log(1 - DP_{HR})} = \frac{Log(1 - 0.78)}{Log(1 - 0.083)} = \frac{-1.514}{-0.087} \approx 17.47 \approx 18$$

$$n_{LR} = \frac{Log(1 - CL)}{Log(1 - DP_{LR})} = \frac{Log(1 - 0.78)}{Log(1 - 0.042)} = \frac{-1.514}{-0.043} \approx 35.29 \approx 36$$

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

- **Option 2:**
  - **Answer:** Implying that if we sample randomly 54 host plants where 18 are belonging to the high-risk group and all test results would be negative, then we could conclude pest freedom with 95% confidence, or even more accurate that if the pest X would be present, we could be 95% confidence that the prevalence in the population would be below 5%.

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

- **Option 3:**

1. If the tolerable level (DP) in the entire population is 5%, what would be the level in the high risk group, considering that the likelihood to be infected is twice as large and they represent only 20% of the total population.

➢ **Answer:** The tolerable level for each risk group can be calculated using the relative risk and the proportion, using specific weighing:

$$W_{HR} = \frac{RR_{HR}}{RR_{HR} \times Prop_{HR} + RR_{LR} \times Prop_{LR}} = \frac{2}{2 \times 0.2 + 1 \times 0.8} = \frac{2}{1.2} \approx 1.67$$

$$W_{LR} = \frac{RR_{LR}}{RR_{HR} \times Prop_{HR} + RR_{LR} \times Prop_{LR}} = \frac{1}{2 \times 0.2 + 1 \times 0.8} = \frac{1}{1.2} \approx 0.83$$

$$DP_{HR} = W_{HR} \times DP = 1.67 \times 0.05 \approx 0.083$$

$$DP_{LR} = W_{LR} \times DP = 0.83 \times 0.05 \approx 0.042$$

$$\Longrightarrow DP = DP_{HR} \times Prop_{HR} + DP_{LR} \times Prop_{LR} = 0.083 \times 0.2 + 0.042 \times 0.8 = 0.05$$

RISK BASED SAMPLING: RISK BASED SURVEILLANCE

- **Option 3:**

  2. Now we can used the tolerable level for the high risk group (DP_{HR}) to calculate the sample size ensuring a confidence level of 95% ($CL = 0.95$).

  ➢ **Answer:** Considering a perfect test then the sample size would be:

$$n = \frac{Log(1 - CL)}{Log(1 - DP_{HR})} = \frac{Log(1 - 0.95)}{Log(1 - 0.083)} = \frac{-2.996}{-0.087} \approx 34.57 \approx 35$$

  Implying that if we sample randomly 35 host plants belonging to the high risk group and all test results would be negative, then we could conclude pest freedom with 95% confidence, or even more accurate that if the pest **X** would be present, we could be 95% confidence that the prevalence in the population would be below 5%.

CONCLUSIONS

- **Surveillance** is a **cost-effective** way to assess pest freedom in a population, making possible to **infer about the whole population** considering only a fraction of the population units.

- **Probability rules and statistical concepts** help us to conduct a scientifically sound surveillance **ensuring** specified **level of confidence** about the conclusions that can be drawn.

- **Differential risk** can be used to focus the surveillance in the group of the population which is **more likely to be infected** and this would **reduce** the **sample size** needed

THANK YOU

efsa
EUROPEAN FOOD SAFETY AUTHORITY