Sampling

Signe Andersen
Centre for Veterinary Epidemiological Research
Atlantic Veterinary College
University of Prince Edward Island
Why think about sampling?

• **Sample TOO LARGE**
  - Waste of time, money and other resources
    - You will never get another research grant...

• **Sample too small**
  - Poor precision
  - Important results may be missed
    - Consequence = wrong conclusions may be drawn about the population!
Hierarchy of populations

• External population
  » “Total” population. Whether or not the results from the study apply depends on how representative the target population is of the external population.

• Target population
  » Population the study population actually represents

• Study population
  » Samples are picked from this population

• Sampling frame
  » List of all the elements you want to sample (e.g. a list of herds or cows that might be selected for sampling)
Sampling methods

• Non-probability
  • Judgment
  • Convenience
  • Purposive

• Probability
  • Random (simple, systematic, stratified)
  • Cluster
  • Multistage
Non-Probability Sampling

• Judgment
  – Subjects picked based on investigator’s judgment of their representativeness of population. DO NOT USE!

• Convenience
  – Herds located close to vet school, producers with good compliance, etc. Generalisability is usually not as good as probability sampling, but it may be necessary to use for logistical reasons.

• Purposive
  – Used in observational studies (case-control, cohort); sampling takes place within two groups, one with and one without a risk factor, and can be done in random fashion.
Probability Sampling

• Simple random
  – Formal randomisation; requires complete enumeration of the population

• Systematic random
  – Selection of subjects at regular intervals (e.g. every 4\(^{th}\) cow); requires that population be structured in some manner (every 4\(^{th}\) cow through the chute or in the tie-stall) (BEWARE OF PATTERNS)

• Stratified random
  – A way to get specific groups of animals represented with stratum specific rates (e.g. Holsteins and Jerseys); separate dairy cows into the two breeds and sample randomly within stratum, ensuring proportionality to the source population
Probability Sampling, continued

• Cluster sampling
  – A cluster is a group of animals that are “more alike” (e.g. a litter of piglets or a pen of feedlot calves)
  – Sometimes easier to sample all animals in a single cluster than a few here and there

• Multistage sampling
  – First, sample clusters; then take a random sample of individuals within the cluster
  – This type of sampling can have many levels depending of the levels of “clustering” in real life.
  – A useful way of sampling large populations for observational studies (e.g. dry cow and lactation cohorts)
    - 15 random cows per farm in each study - in the National Cohort of Dairy Farms (NCDF)
Confidence level

• The standard normal curve with data falling between 1 and 2 standard deviations ($\sigma$); the area under the curve when the two dark blue wedges are removed represents 95% of the data

• $\alpha$ denotes the confidence level and is calculated as $1 - 95\%$ or $1 - 0.95 = 0.05$. This corresponds to the table value $z = 1.96$
Sampling to estimate a mean

Formula: \[ n = \frac{Z_{1-\alpha/2} \cdot \sigma^2}{L^2} \]

- \( n \): sample size
- \( Z \): value of the standard normal distribution corresponding to a two-sided 95% confidence level of \( 1 - \alpha/2 = Z_{1-0.05/2} = Z_{0.975} = 1.96 \)
- \( \sigma \): ‘a priori’ estimate of variance
- \( L \): precision or “maximum allowable error” (e.g. estimated daily weight gain within ± 5kg)
Sampling to estimate a proportion

Formula:

\[ n = \frac{Z_{1-\alpha/2}^2 \cdot pq^2}{L^2} \]

\( p \)  
a guess at the proportion being estimated (e.g. prevalence of a certain disease); the guess can be based on previous studies. If no information is available, use \( p=0.5 \) (\( n \) will be maximised at this value). The adjusted formula should also be used in this case, if \( n/N >0.1 \).

\( q \)  
equal to \( 1 - p \)
Adjustment for finite (small) population

- Only for descriptive studies
- Useful for simple or stratified random sample when sample fraction exceeds 10%

Formula: \[ n' = \frac{1}{\frac{1}{n} + \frac{1}{N}} \]

- \( n \) originally calculated sample size
- \( n' \) adjusted sample size
Sampling to detect disease

Formula: $n = \left(1 - \alpha^D\right) \left(N - \frac{D-1}{2}\right)$

- $\alpha$: 1 – the confidence level (usually 0.05)
- $D$: estimated minimum number of diseased animals in the group (population size × min. expected prevalence)
- $N$: population size

This formula relies on a diagnostic test with Se=100 % and Sp=100%.
Sampling to detect disease

- Infinite population
  \( N > 1000 \)

Formula:
\[
 n = \frac{\ln \alpha}{\ln(1 - p)}
\]

- \( \alpha \) 1 – confidence level (usually 0.05)
- \( p \) expected prevalence of disease

This formula relies on a perfect diagnostic test.
Examples

• Handouts with a few examples
• Use formulae in lecture handouts
• You will need a calculator…. 
• We are all here to help!
• Solutions will be handed out before the break.
• Good luck!