Sampling & Confidence Intervals

Mark Lunt

Centre for Epidemiology Versus Arthritis
University of Manchester

25/10/2022

Principles of Sampling

- Often, it is not practical to measure every subject in a population.
- A reduced number of subjects, a sample, is measured instead.
  - Cheaper
  - Quicker
  - More thorough
- Sample needs to be chosen in such a way as to be representative of the population

Types of Sample

- Simple Random
- Stratified
- Cluster

- Quota
- Convenience
- Systematic

Simple Random Sample

- Every subject has the same probability of being selected.
- This probability is independent of who else is in the sample.
- Need a list of every subject in the population (*sampling frame*).
- Statistical methods depend on randomness of sampling.
- Refusals mean the sample is no longer random.

Stratified

- Divide population into distinct sub-populations.
  - E.g. into age-bands, by gender
- Randomly sample from each sub-population.
  - sampling probability is same for everyone in a sub-population
  - sampling probability differs between sub-populations
- More efficient than a simple random sample if variable of interest varies more *between* sub-populations than *within* sub-populations.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Cluster

- Randomly sample groups of subjects rather than subjects
- Why ?
  - List of subjects not available, list of groups is
  - Cheaper and easier to recruit a number of subjects at the same time.
  - In intervention studies, may be easier to treat groups: randomise hospitals rather than patients.
- Need a reasonable number of clusters to assure representativeness.
- The more similar clusters are, the better cluster sampling works.
- Cluster samples need special methods for analysis

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Quota

- Deliberate attempt to ensure proportions of subjects in each category in a sample match the proportion in the population.
- Often used in market research: quotas by age, gender, social status.
- Variables not used to define the quotas may be very different in the sample and population.
- Proportion of men and of elderly may be correct, not proportions of elderly men.
- Probability of inclusion is unknown, may vary greatly between categories
- Cannot assume sample is representative.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Systematic & Convenience Samples

Systematic Take every n^{th} subject.

- If there is clustering (or periodicity) in the sampling frame, may not be representative.
- Shared surnames can cause problems.
- Randomly order and take every n^{th} subject: random.

Convenience Take a random sample of easily accessible subjects

- May not be representative of entire population.
- E.g. people going to G.P. with sore throat easy to identify, not representative of people with sore throat.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Estimating from Random Samples

- We are interested in what our sample tells us about the population
- We use sample statistics to estimate population values
- Need to keep clear whether we are talking about sample or population
- Values in the population are given Greek letters $\mu, \pi \ldots$, whilst values in the sample are given equivalent Roman letters $m, p \ldots$.
- Suppose we have a population, in which a variable x has a mean μ and standard deviation σ. We take a **random** sample of size n. Then
  - Sample mean $\bar{x}$ should be close to the population mean μ.
  - However, if several samples are taken, $\bar{x}$ in each sample will differ slightly.

Variation of $\bar{x}$ around μ

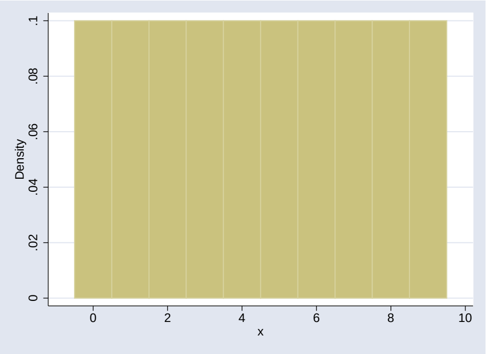
- How much the means of different samples differ depends on

  Sample Size: The mean of a small sample will vary more than the mean of a large sample.

  Variance in the Population: If the variable measured varies little, the sample mean can only vary little.
- I.e. variance of $\bar{x}$ depends on variance of x and on sample size n.

Example

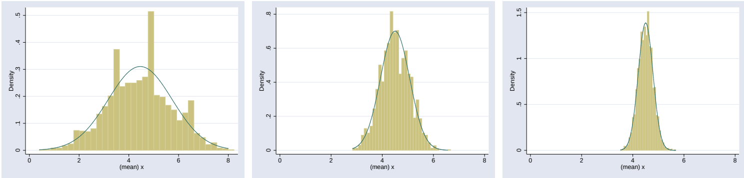
Consider consider a population consisting of 1000 copies of each of the digits 0, 1, . . . , 9. The distribution of the values in this population is



Example: Samples

- Samples of size 5, 25 and 100
- 2000 samples of each size were randomly generated
- Mean of x ($\bar{x}$) was calculated for each sample
- Histograms created for each sample size separately

Example: Distributions of $\bar{x}$



Size 5   Size 25   Size 100

Properties of $\bar{x}$

$\boldsymbol{E(\bar{x}) = \mu}$ i.e. on average, the sample mean is the same as the population mean.

Standard Deviation of $\bar{x}$ = $\frac{\sigma}{\sqrt{n}}$ i.e the uncertainty in $\bar{x}$ increases with σ, decreases with n. The standard deviation of the mean is also called the **Standard Error**

$\bar{x}$ is normally distributed This is true whether or not x is normally distributed, provided n is sufficiently large. Thanks to the *Central Limit Theorem*.

Standard Error

- Standard deviation of the *sampling distribution* of a statistic
- Sampling distribution: the distribution of a statistic as sampling is repeated
- All statistics have sampling distributions
- Statistical inference is based on the standard error

Example: Sampling Distribution of $\bar{x}$

$\mu = 4.5$ $\sigma = 2.87$

Size of samples	Mean $\bar{x}$		S.D. $\bar{x}$	
	Predicted	Observed	Predicted	Observed
5	4.5	4.47	1.29	1.26
25	4.5	4.51	0.57	0.57
100	4.5	4.50	0.29	0.30

Estimating the Variance

In a population of size N, the variance of x is given by

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{1}$$

This is the *Population Variance*
In a sample of size n, the variance of x is given by

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{2}$$

This is the *Sample Variance*

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Why $n - 1$ rather than N

Population $\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$
Sample $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$

- Use $n - 1$ rather than n because we don't know μ, only an imperfect estimate $\bar{x}$.
- Since $\bar{x}$ is calculated from the sample (i.e. from the x_i), x_i will tend to be closer to $\bar{x}$ than it is to μ.
- Dividing by n would underestimate the variance
- With a reasonable sample size, makes little difference.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Proportions

Suppose that you want to estimate π, the proportion of subjects in the population with a given characteristic. You take a random sample of size n, of whom r have the characteristic.

- $p = \frac{r}{n}$ is a good estimator for π.
- If you create a variable x which is 1 for subjects which have the characteristic and 0 for those who do not, then $p = \bar{x}$
- If the sample is large, p will be normally distributed, even though x isn't

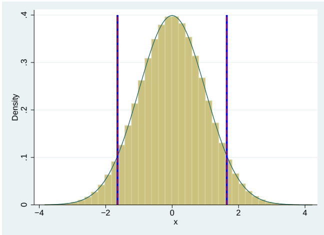
CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Reference Ranges

If x is normally distributed with mean μ and standard deviation σ, then we can find out all of the percentiles of the distribution. E.g.
Median = μ
25^{th} centile = $\mu - 0.674\sigma$
75^{th} centile = $\mu + 0.674\sigma$
Commonly, we are interested in the interval in which 95% of the population lie, which is from $\mu - 1.96\sigma$ to $\mu + 1.96\sigma$
This is from the 2.5^{th} centile to the 97.5^{th} centile

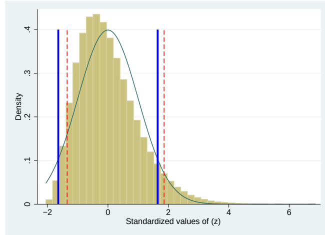
CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Reference Range Illustration



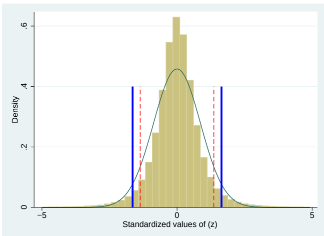
- Red lines cut off 5% of data in each tail
- 90% of data lies between lines
- Blue lines are at -1.645, 1.645

Non-normal distributions 1: Skewed distribution



- χ^2 distribution
- Red lines cut off 5% of data in each tail
- Mean $\pm$ 1.645 $\times$ S.D. covers $>$ 90% of data
- Only 2% $<$ mean - 1.645 S.D
- 6.5% $>$ mean + 1.645 S.D.

Non-normal distributions 2: Long-tailed distribution



- t-distribution
- Symmetric, but not normal
- Higher “peak”, longer tails than normal
- Red lines cut off 5% of data in each tail
- Blue lines at mean $\pm$ 1.645 S.D.
- Mean $\pm$ 1.645 $\times$ S.D. covers $>$ 94% of data

Reference Range Example

Bone mineral density (BMD) was measured at the spine in 1039 men. The mean value was 1.06g/cm^2 and the standard deviation was 0.222g/cm^2. Assuming BMD is normally distributed, calculate a 95% reference interval for BMD in men.

$$
\begin{aligned}
&\text{Mean BMD} &&= 1.06\text{g/cm}^2 \\
&\text{Standard deviation of BMD} &&= 0.222\text{g/cm}^2 \\
\Rightarrow\ &\text{95\% Reference interval} &&= 1.06 \pm 1.96 \times 0.222 \\
& &&= 0.62\text{g/cm}^2, 1.50\text{g/cm}^2
\end{aligned}
$$

Confidence Intervals

- The distribution of $\bar{x}$ approaches normality as n gets bigger.
- $\bar{x}$ can be thought of as a random draw from a distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- If samples could be taken repeatedly, 95% of the time, the $\bar{x}$ would lie between $\mu - 1.96\frac{\sigma}{\sqrt{n}}$ and $\mu + 1.96\frac{\sigma}{\sqrt{n}}$.
- As a consequence, 95% of the time, μ would lie between $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$.
- This is a 95% confidence interval for the population mean.
- If, as is usually the case, σ is unknown, can use its estimate s.

Confidence Interval Example

In 216 patients with primary biliary cirrhosis, serum albumin had a mean value of 34.46 g/l and a standard deviation of 5.84 g/l.

$$
\begin{aligned}
&\text{Standard deviation of } x &&= 5.84 \\
\Rightarrow\quad &\text{Standard error of } \bar{x} &&= \frac{5.84}{\sqrt{216}} \\
& &&= 0.397 \\
\Rightarrow\quad &\text{95\% Confidence Interval} &&= 34.46 \pm 1.96 \times 0.397 \\
& &&= (33.68, 35.24)
\end{aligned}
$$

So, the mean value of serum albumin in the *population* of patients with primary biliary cirrhosis is probably between 33.68 g/l and 35.24 g/l.

Confidence Intervals for Proportions

- p is normally distributed with standard error $\sqrt{\frac{p(1-p)}{n}}$ **provided n is large enough**.
- This can be used to calculate a confidence interval for a proportion.
- Exact confidence intervals can be calculated for small n (less than 20, say) from tables of the binomial distribution.
- A reference range for a proportion in meaningless: a subject either has the characteristic or they do not.

Confidence Interval around a Proportion: Example

100 subjects each receive two analgesics, X and Y, for one week each in a randomly determined order. They then state a preference for one drug. 65 prefer X, 35 prefer Y. Calculate a 95% confidence interval for the proportion preferring X.

$$
\begin{aligned}
&\text{Standard Error of p} &&= \sqrt{\frac{0.65 \times 0.35}{100}} \\
& &&= 0.0477 \\
\Rightarrow\quad &\text{95\% Confidence Interval} &&= 0.65 \pm 1.96 \times 0.0477 \\
& &&= (0.56, 0.74)
\end{aligned}
$$

So, in the general population, it is likely that between 56% and 74% of people would prefer X.

Confidence Intervals in Stata

- The `ci` command produces confidence intervals
- For proportions, you use the `binomial` option

Confidence Intervals and Reference Ranges

- Confidence intervals tell us about the *population mean*
- Reference ranges tell us about *individual values*
- Reference ranges require the variable to be normally distributed
- Confidence intervals do not
  - If sampling distribution of statistic of interest is normal
  - Normality may require reasonable sample size

Sample Size Calculations

- Primary outcome of a study is a statistic (mean, proportion, relative risk, incidence rate, hazard ratio etc)
- The larger the study, the more precisely we can estimate our statistic
- We can calculate how many subjects we need to achieve adequate precision if we
  - know how the distribution of the statistic changes with increasing numbers of subjects
  - Have a definition of “adequate”
- Power-based calculations are more complicated (for next week).

Sample size for precision of mean

Suppose that we want to know μ to a certain level of precision.

- We can be 95% certain that μ lies within

$$\bar{x} \pm \frac{1.96\sigma}{\sqrt{n}}$$

- The width of this interval depends on n, which we control.
- Therefore, we can select n to give our chosen width.
- Need to use an estimate for σ, for which we can use s.

Sample Size Formula

Suppose we want to fix the width of the 95% confidence interval to 2W, i.e. 95% CI = $\bar{x} \pm W$. Then

$$\begin{aligned} W &= 1.96 \times \text{Standard Error} \\ &= 1.96 \times \frac{\sigma}{\sqrt{n}} \\ \Rightarrow \quad W^2 &= \frac{1.96^2 \sigma^2}{n} \\ \Rightarrow \quad n &= \left(\frac{1.96\sigma}{W}\right)^2 \end{aligned}$$

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Sample Size Example

In the primary biliary cirrhosis example, suppose that we wish to know the mean serum albumin in cirrhosis patients to within 0.5 g/l. How many patients would we need to study (assuming a standard deviation of 5.84 g/l).

$$\begin{aligned} W &= 0.5 \\ \sigma &= 5.84 \\ \Rightarrow \quad n &= \left(\frac{1.96\sigma}{W}\right)^2 \\ &= \left(\frac{1.96 \times 5.84}{0.5}\right)^2 \\ &\approx 524 \end{aligned}$$

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS