

### Populations & Samples

David Lee

Arthritis Research UK Epidemiology Unit  
University of Manchester



11/10/2011

- 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.
- Randomly sample from each sub-population.
- More efficient than a simple random sample if variable of interest varies more *between* sub-populations than *within* sub-populations.

## 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.

## 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.
- Cannot assume sample is representative.

## Systematic & Convenience Samples

**Systematic** Take every  $n^{\text{th}}$  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  $n^{\text{th}}$  subject: random.

**Convenience** Take a random sample of easily accessible subjects

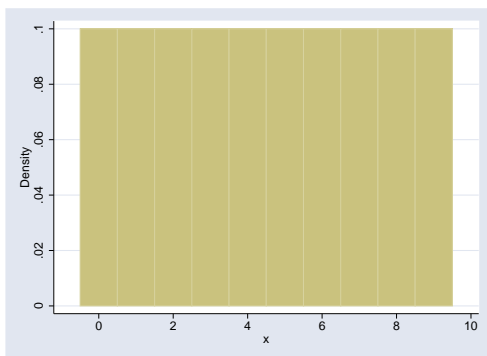
- May not be representative of entire population.

## Estimating from Random Samples

- Values in the population are given Greek letters  $\mu, \pi \dots$ , whilst values in the sample are given equivalent Roman letters  $m, p \dots$
- Suppose we have a population, in which a variable  $x$  has a mean  $\mu$  and standard deviation  $\sigma$ . We take a **random** sample of size  $n$ . Then
  - Sample mean  $\bar{x}$  should be close to the population mean  $\mu$ .
  - However, if several samples are taken,  $\bar{x}$  in each sample will differ slightly.

## 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



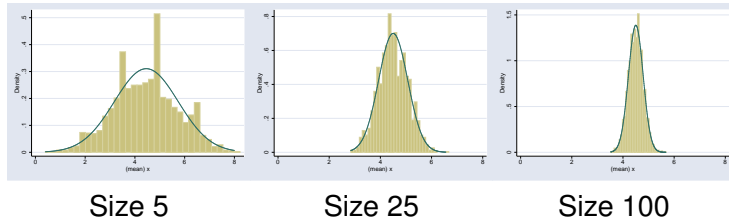
## Variation of $\bar{x}$ around $\mu$

- 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: 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 $\bar{x}$



## 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

## Properties of $\bar{x}$

$E(\bar{x}) = \mu$  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  $\sigma$ , 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*.

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

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

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

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

This is the *Population Variance*

In a sample of size  $n$ , the variance of  $x$  is given by

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

This is the *Sample Variance*

## Why $n - 1$ rather than $N$

Population  $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

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

- Use  $n - 1$  rather than  $n$  because we don't know  $\mu$ , 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  $\mu$ .
- Dividing by  $n$  would underestimate the variance
- With a reasonable sample size, makes little difference.

## Proportions

Suppose that you want to estimate  $\pi$ , 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  $\pi$ .
- 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

## Reference Ranges

**If  $x$  is normally distributed** with mean  $\mu$  and standard deviation  $\sigma$ , then we can find out all of the percentiles of the distribution. E.g.

Median =  $\mu$

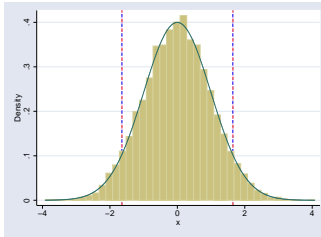
25<sup>th</sup> centile =  $\mu - 0.674\sigma$

75<sup>th</sup> 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  $\mu + 1.96\sigma$

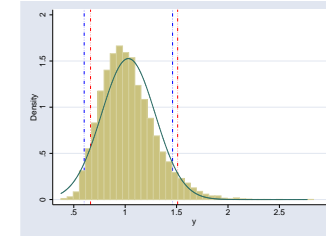
This is from the 2.5<sup>th</sup> centile to the 97.5<sup>th</sup> centile

## Reference Range Illustration



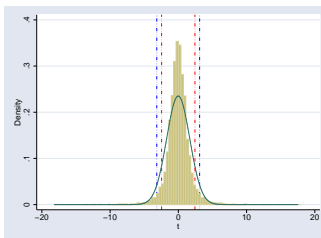
- Marked lines cut off 5% of data in each tail
- 90% of data lies between lines
- Lines are at -1.645, 1.645

## Non-normal distributions 1: Skewed distribution



- Red lines cut off 5% of data in tails
- Mean  $\pm 1.645 \times$  S.D. covers  $> 90\%$  of data
- Only 2% cut-off below mean, 6.5% above

## Non-normal distributions 2: Long-tailed distribution



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  $1.06\text{g/cm}^2$  and the standard deviation was  $0.222\text{g/cm}^2$ . 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.
- The standard deviation of  $\bar{x}$  is  $\frac{\sigma}{\sqrt{n}}$ .
- 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,  $\mu$  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,  $\sigma$  is unknown, can use its estimate  $s$ .

## 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 is meaningless: a subject either has the characteristic or they have not.

## 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.

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

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 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.

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

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 samples are large enough).

## 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

## Sample size for precision of mean

Suppose that we want to know  $\mu$  to a certain level of precision.

- We can be 95% certain that  $\mu$  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  $\sigma$ , 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

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

## 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).

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