Confidence Intervals

Dr. Osman Abdullahi

Point Estimation:

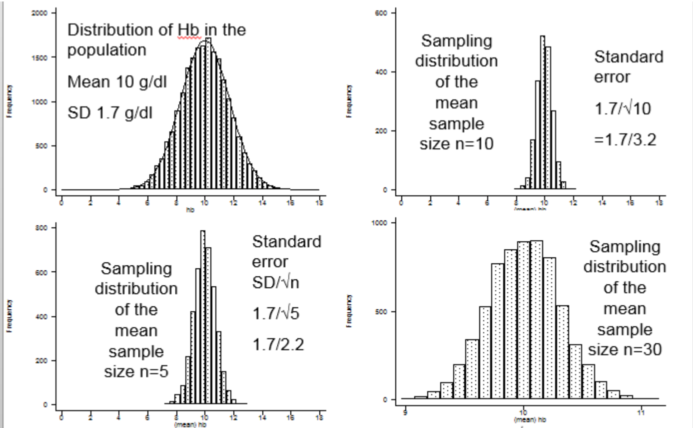
1. Provides a single value
   - Based on observations from one sample
2. Gives no information about how close the value is to the unknown population parameter
3. Example: Sample mean $x = 3$ is point estimate of unknown population mean

Interval estimation:

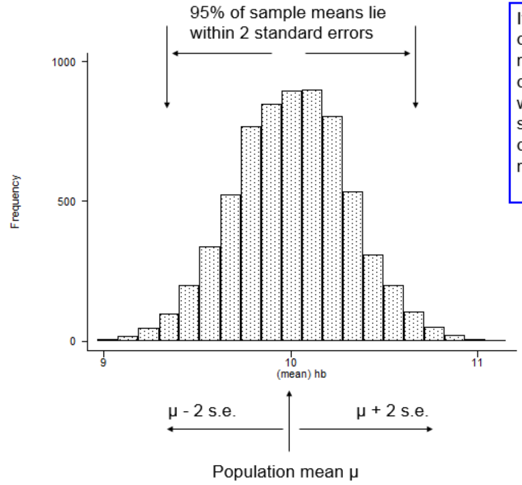
1. Provides a range of values
   - Based on observations from one sample
2. Gives information about closeness to unknown population parameter

- Stated in terms of probability
  - Knowing exact closeness requires knowing unknown population parameter

3. Example: The range between 50 and 70 contains the true unknown parameter value with 95% confidence]



- Bigger sample size, less variability in sample means: s.e.=SD√n
- Average value of sample means is equal to the population mean
- Sample means follow normal distribution if the distribution in the population is normal
- and if n is large, is roughly normal even if the population is not normal



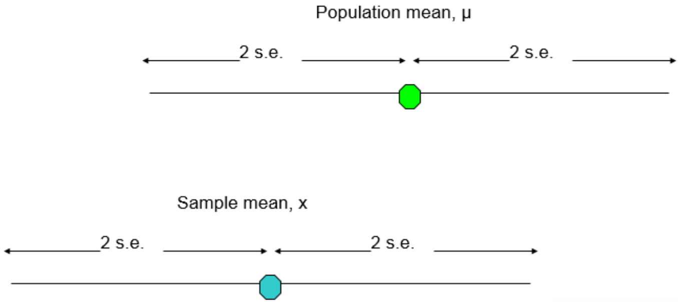
If we repeated our sampling many times, 95% of sample means would be within 2 standard errors of the population mean.

In practice we usually have only one sample but we can use the sample data to quantify the uncertainty in our single estimate

95% of sample means lie within 2 s.e. of the population mean

so we can also say:

95% of the time the population mean will lie within 2 s.e. of the our sample mean



Confidence interval on μ (σ known)

$$\Pr\left\{-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right\} = 0.95$$

$$\Pr\left\{\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

$$\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}}, \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a 95% confidence interval for μ .

A survey of haemoglobin status in children <5yrs in Kilifi district. 30 children gave a finger prick blood sample and the mean Hb was 9.6g/dl with a standard deviation of 1.5 g/dl.

Sample estimate of the SD: 1.5 g/dl

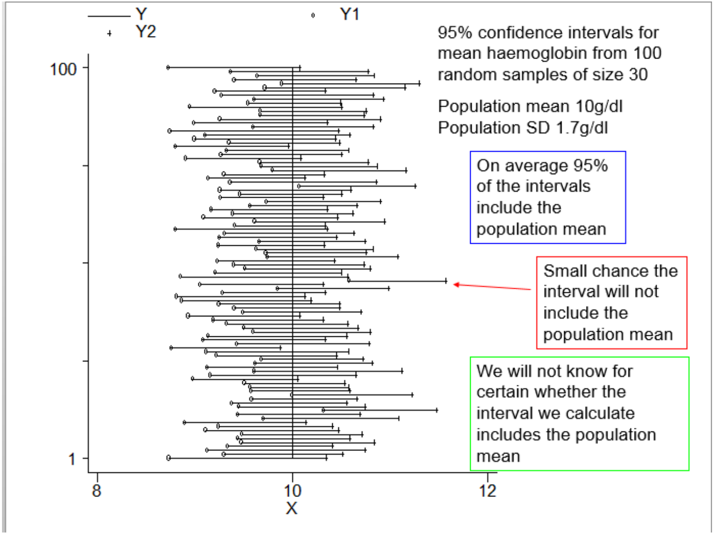
Standard error s.e. = SD/$\sqrt{n}$

s.e.=1.5/$\sqrt{30}$ = 1.5/5.477 = 0.274

95% confidence interval 9.6 – 2x0.274 to 9.6 + 2x0.274

9.1 g/dl to 10.1 g/dl

Interpretation: we can say with 95% confidence that the mean Haemoglobin concentration in children in the population could be as small as 9.1 or as big as 10.1 g/dl


Y
Y1
Y2
100
1
8
10
12
X
95% confidence intervals for mean haemoglobin from 100 random samples of size 30
Population mean 10g/dl
Population SD 1.7g/dl
On average 95% of the intervals include the population mean
Small chance the interval will not include the population mean
We will not know for certain whether the interval we calculate includes the population mean

Confidence Intervals: Factors affecting width

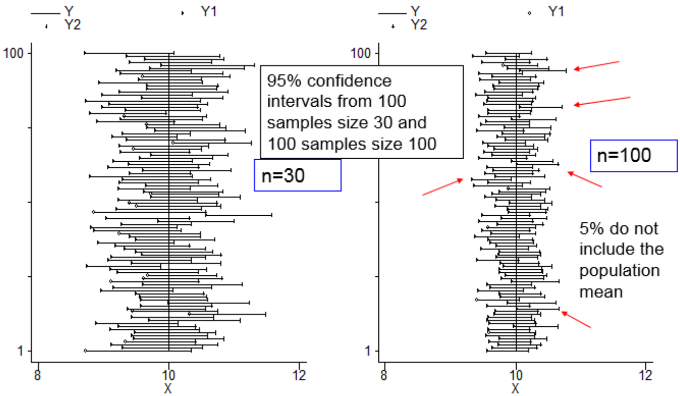
Intervals extend from

$\bar{X} - Z\sigma_{\bar{X}}$ to $\bar{X} + Z\sigma_{\bar{X}}$

1. Data dispersion
   - Measured by σ
2. Sample size

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

3. Level of confidence $(1 - \alpha)$
   - Affects Z



- 95% of the intervals include the population mean
- Larger sample size -> narrow confidence interval
- We can improve the probability of including the population mean by using 99% confidence intervals, at the cost of having wider intervals and so more uncertainty

- For large samples: x+/- 1.96s.e.
- With continuous data, if samples are small the estimated standard error tends to be an underestimate. To correct for this we use t-tables to find the multiplier:

degrees of freedom	Two-tailed probability 0.1	0.05	0.01
1	6.314	12.706	63.656
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
10	1.812	2.228	3.169
20	1.725	2.086	2.845
29	1.699	**2.045**	**2.756**
30	1.697	2.042	2.750
100	1.660	1.984	2.626
150	1.655	1.976	2.609

Sample mean 9.6 Sample SD 1.5

Standard error $1.5/\sqrt{30}=0.274$

95% confidence interval:

9.6 – **2.045** x 0.274 **to** 9.6 + **2.045**x0.274

9.04 to 10.16 g/dl

99% confidence interval:

9.6 – **2.756** x 0.274 **to** 9.6 + **2.756**x0.274

8.85 to 10.35 g/dl

Summary

- The confidence interval expresses the uncertainty in sample estimates of means, proportions, treatment efficacy etc.
- Larger the sample, the narrower the CI
- Can improve the probability of including the population mean by calculating 99% interval but at a cost of having a wider interval and thus greater uncertainty
- Should design studies to yield CI's that are narrow enough to draw conclusions
- Interpretation depends on the assumption that the sample was representative of the population

Exercise

- Suppose we've collected a random sample of 10 recently graduated students and asked them what their annual salary is. Imagine that this is the data we see, 44617,7066, 17594, 2726, 1178, 18898, 5033, 37151, 4514, and 4000.
- **Goal:** Estimate the mean salary of the graduated children. Find a 90 and 95 % confidence interval for the mean.
- **Setting 1:** Assume that incomes are normally distributed with unknown mean and SD = ksh15,000.
- **Setting 2:** Same problem, only now we do not know the value for the standard deviation.