Chapter 4: Estimation

- Estimation is the process of using sample data to draw inferences about the population

Sample information $\bar{x}, s^2$ —— Inferences ——→ Population parameters μ, σ^2

Point and interval estimates

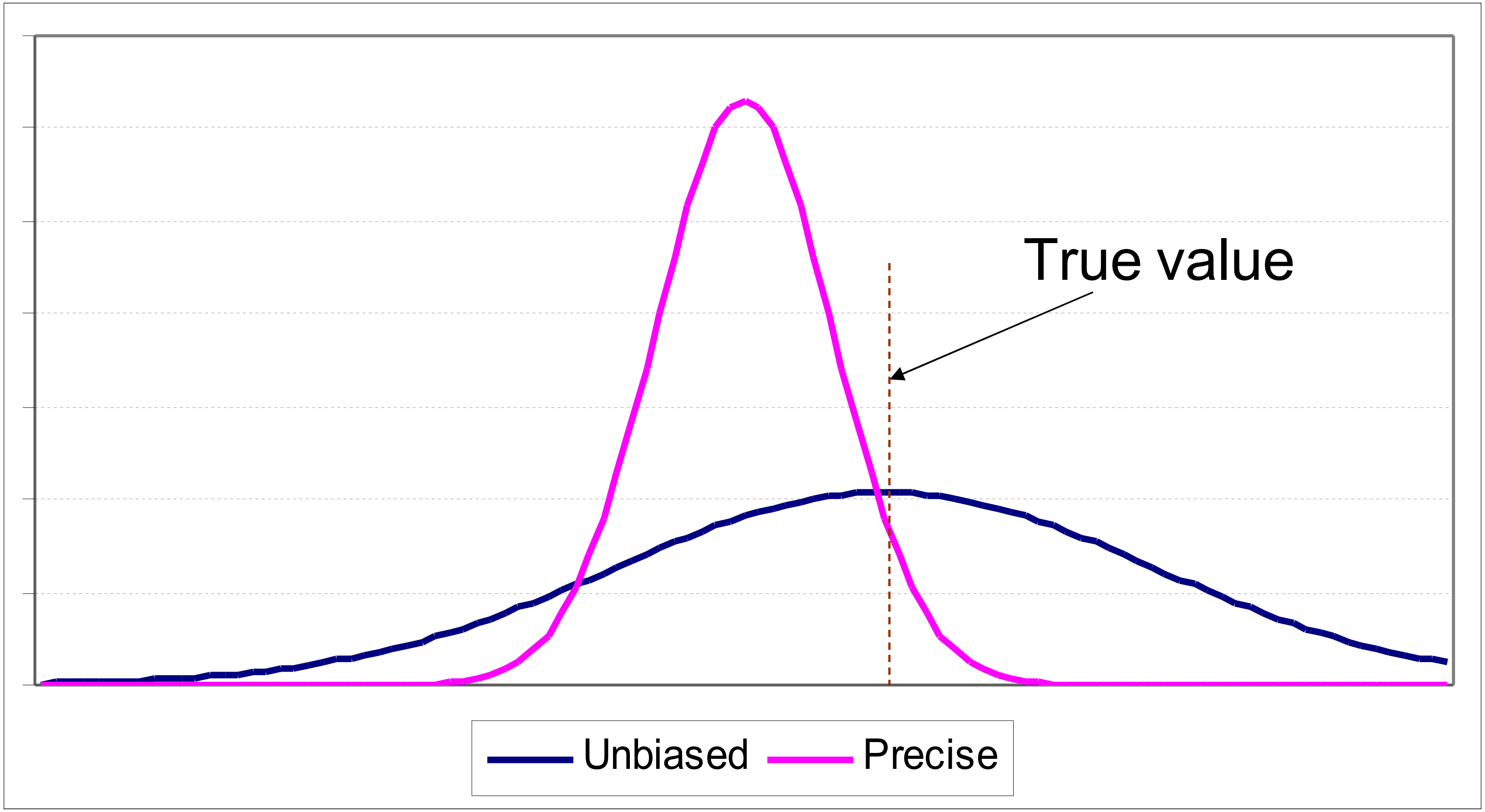
- Point estimate – a single value
  - the temperature tomorrow will be 23°
- Interval estimate – a range of values, expressing the degree of uncertainty
  - the temperature tomorrow will be between 21° and 25°

Criteria for good estimates

- Unbiased – correct on average
  - the expected value of the estimate is equal to the true value

- Precise – small sampling variance
  - the estimate is close to the true value for all possible samples

Bias and precision – a possible trade-off



Estimating a mean (large samples)

- Point estimate – use the sample mean (unbiased)
- Interval estimate – sample mean ± ‘something’
- What is the something?
- Go back to the distribution of $\bar{x}$

The 95% confidence interval

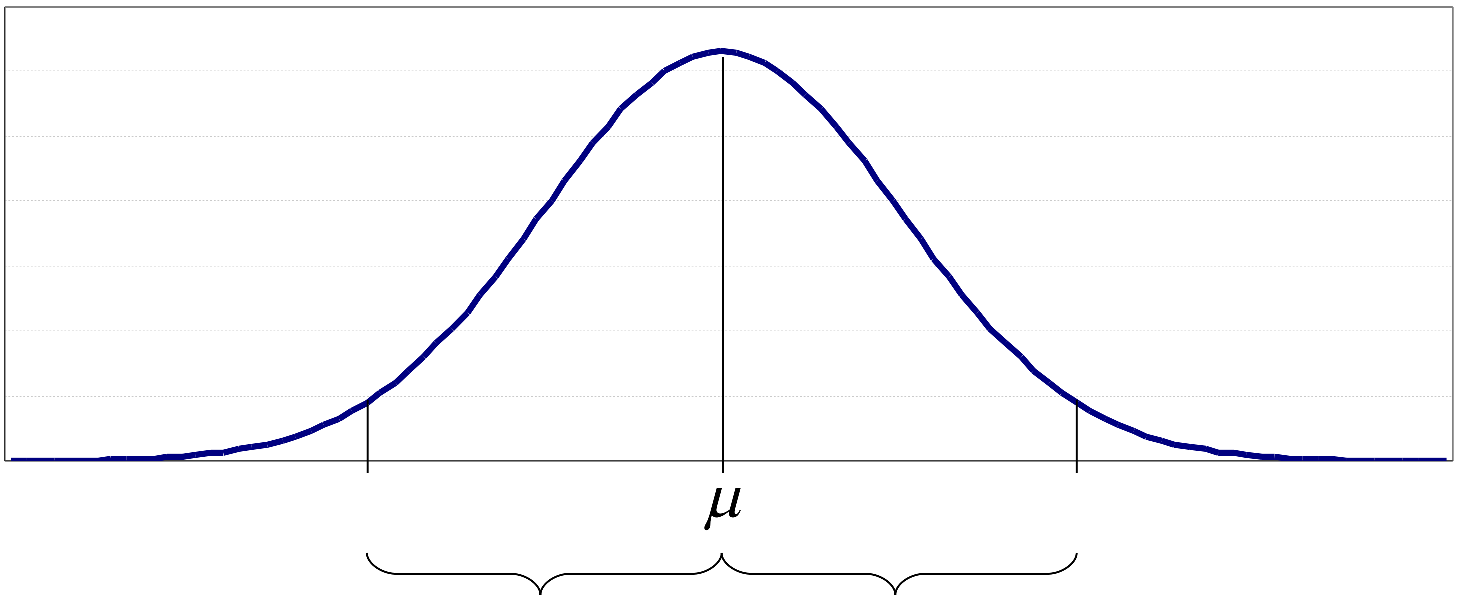
- $\bar{x} \sim N(\mu, \sigma^2/n)$ (Eqn. 3.17)
- Hence the 95% probability interval is

$$\Pr(\mu - 1.96\sqrt{\sigma^2/n} \le \bar{x} \le \mu + 1.96\sqrt{\sigma^2/n} = 0.95$$

- Rearranging this gives the 95% confidence interval

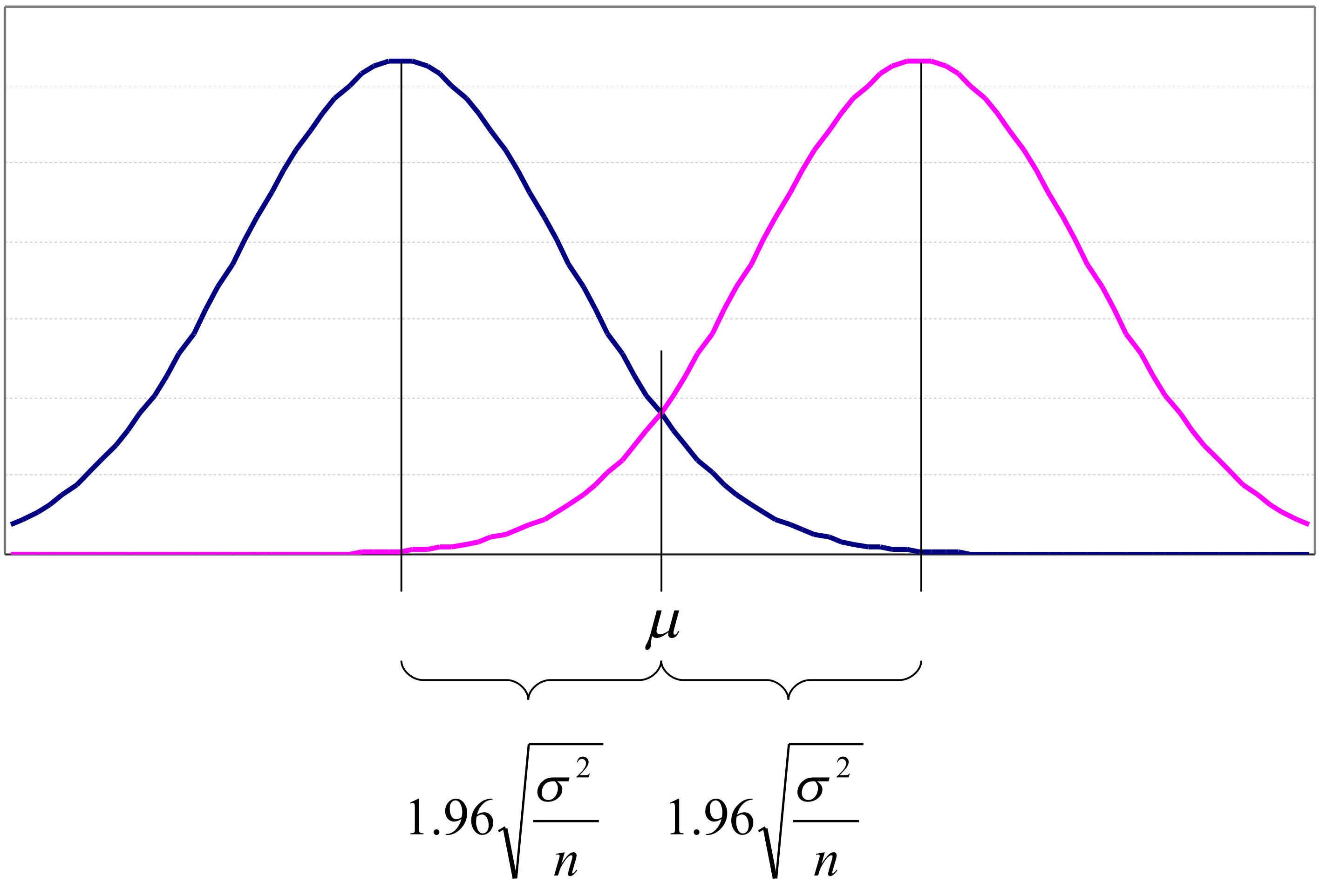
$$[\bar{x} - 1.96\sqrt{\sigma^2/n} \le \mu \le \bar{x} + 1.96\sqrt{\sigma^2/n}]$$

The 95% probability interval



$$1.96\sqrt{\frac{\sigma^2}{n}} \qquad 1.96\sqrt{\frac{\sigma^2}{n}}$$

The 95% confidence interval



Example: estimating average wealth

- Sample data:
  - $\bar{x}$ = 130 (in £000)
  - s^2 = 50,000
  - n = 100
- Estimate μ, the population mean

Example: estimating average wealth (continued)

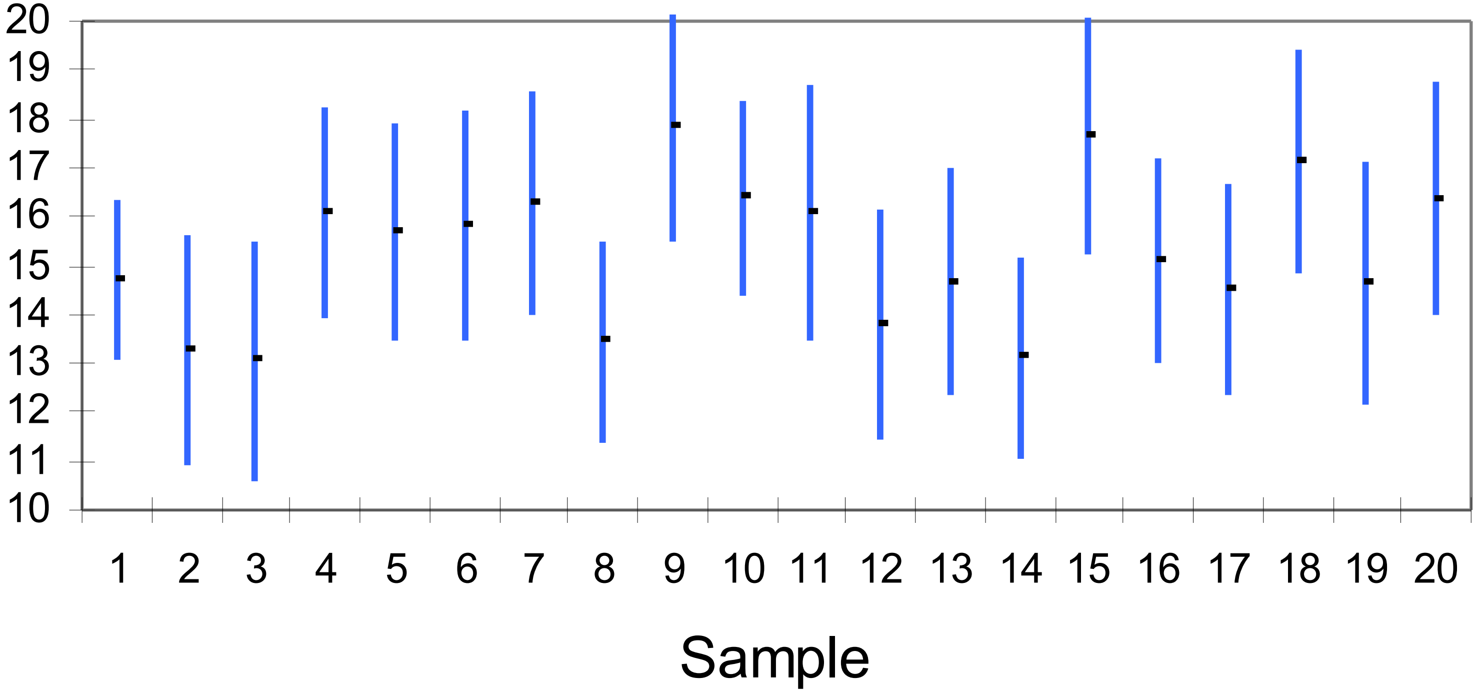
- Point estimate: 130 (use the sample mean)
- Interval estimate – use

$$\bar{x} \pm 1.96 \times \sqrt{s^2/n}$$

$$= 130 \pm 1.96 \times \sqrt{50{,}000/100}$$

$$= 130 \pm 43.8 = [86.2, 173.8]$$

What is a confidence interval?



One sample out of 20 (5%) does not contain the true mean, 15.

Estimating a proportion

- Similar principles
  - The sample proportion provides an unbiased point estimate
  - The 95% CI is obtained by adding and subtracting 1.96 standard errors

- In this case we use

$$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Example: unemployment

- Of a sample of 200 men, 15 are unemployed. What can we say about the true proportion of unemployed men?

- Sample data
  - p = 15/200 = 0.075
  - n = 200

Example: unemployment (continued)

- Point estimate: 0.075 (7.5%)

- Interval estimate:

$$p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

$$= 0.075 \pm 1.96 \times \sqrt{\frac{0.075 \times 0.925}{200}}$$

$$= 0.075 \pm 0.037 = [0.038, 0.112]$$

Estimating the difference of two means

- A survey of holidaymakers found that on average women spent 3 hours per day sunbathing, men spent 2 hours. The sample sizes were 36 in each case and the standard deviations were 1.1 hours and 1.2 hours respectively. Estimate the true difference between men and women in sunbathing habits.

Same principles as before…

- Obtain a point estimate from the samples
- Add and subtract 1.96 standard errors to obtain the 95% CI
- We just need the appropriate formulae

CE:

AQ: We have introduced the caption "Calculating point estimate". Please confirm if it is ok.

Calculating point estimate

- Point estimate – use $\bar{x}_1 - \bar{x}_2$
- For the standard error, use $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Hence the point estimate is 3 – 2 = 1 hour

CE:

AQ: We have introduced the caption "Confidence intervals". Please confirm if it is ok.

Confidence intervals

- For the confidence interval we have

$$1 \pm 1.96\sqrt{\frac{1.1}{36}+\frac{1.2}{36}}$$
$$= 1 \pm 0.7 = [0.3, 1.7]$$

- i.e. between 0.3 and 1.7 extra hours of sunbathing by women.

Using different confidence levels

- The 95% confidence level is a convention
- The 99% confidence interval is calculated by adding and subtracting 2.57 standard errors (instead of 1.96) to the point estimate.
- The higher level of confidence implies a wider interval.

Estimating the difference between two proportions

- Similar to before – point estimate plus and minus 1.96 standard errors

$$p_1 - p_2 \pm 1.96\sqrt{\frac{p_1\left(1-p_1\right)}{n_1}+\frac{p_2\left(1-p_2\right)}{n_2}}$$

Estimation with small samples: using the *t* distribution

- If:
  - The sample size is small (<25 or so), and
  - The true variance σ^2 is unknown
- Then the *t* distribution should be used instead of the standard Normal.

Example: beer expenditure

- A sample of 20 students finds an average expenditure on beer per week of £12 with standard deviation £8. Find the 95% CI estimate of the true level of expenditure of students.

- Sample data:

$$\bar{x} = 12, s = 8, n = 20$$

Example: beer expenditure (continued)

- The 95% CI is given by

$$\bar{x} \pm t_{n-1}\sqrt{s^2/n}$$
$$= 12 \pm 2.093\sqrt{8^2/20}$$
$$= 12 \pm 3.7 = [8.3, 15.7]$$

- The *t* value of t_{19} = 2.093 is used instead of z =1.96

Summary

- The sample mean and proportion provide unbiased estimates of the true values

- The 95% confidence interval expresses our degree of uncertainty about the estimate

- The point estimate ± 1.96 standard errors provides the 95% interval in large samples