SCIU4T4: Confidence intervals

Confidence in a sampled mean value

- How certain are we that our sample mean ($\bar{x}$) is close to the population mean (μ_x)?
- Should express our uncertainty clearly
- **Confidence intervals** reflect uncertainty of μ_x

Confidence in $\mu_{\mathbf{x}}$:

- **High**: Mean wing length was $\bar{x} = 7.2\ cm\ \pm 0.1$
- **Low**: Mean wing length was $\bar{x} = 7.2\ cm\ \pm 2.0$

Larger $\pm$ interval means *less* confidence in the true mean

Confidence in $\mu_{\mathbf{x}}$:

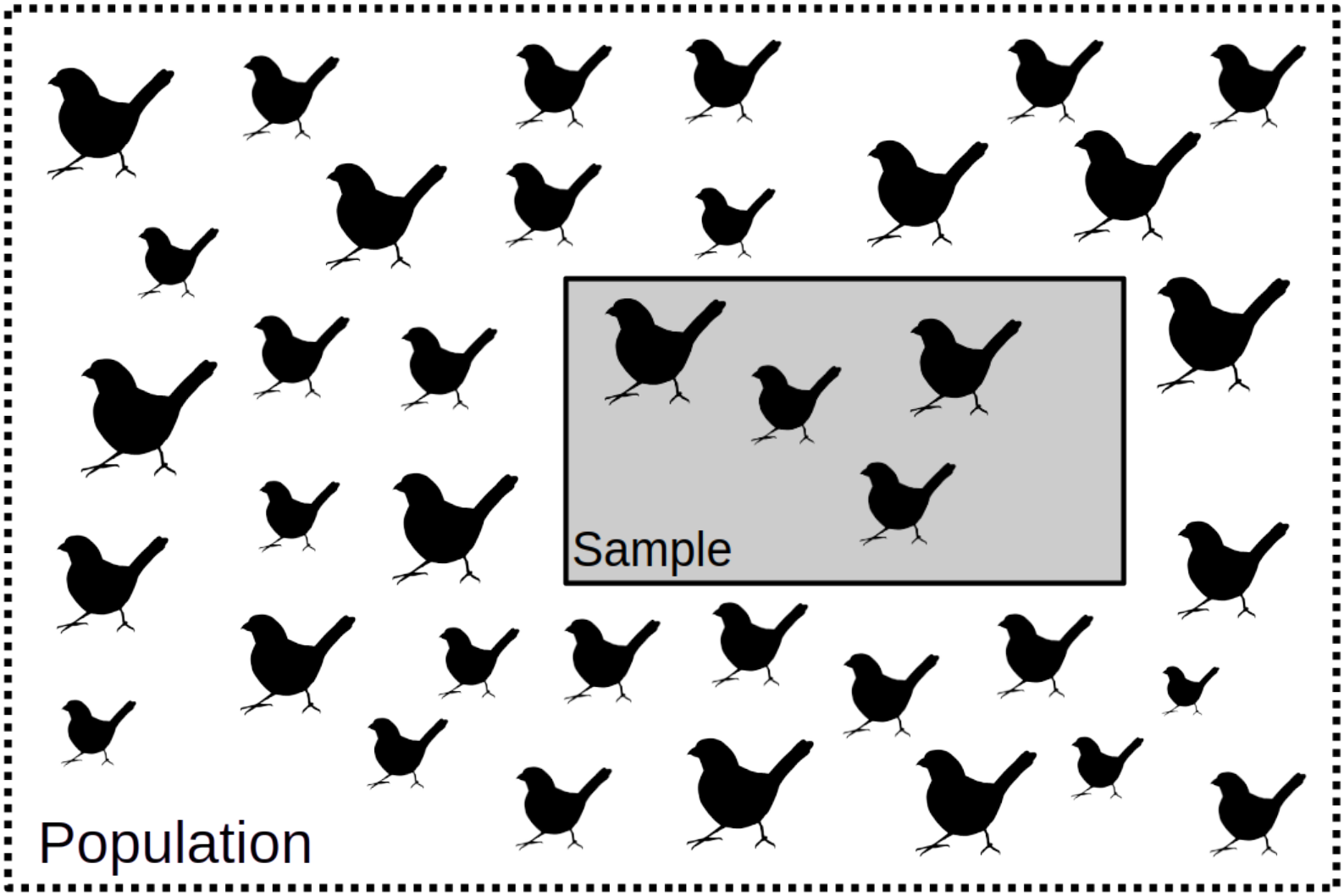
- **High**: Mean wing length was $\bar{x} = 7.2\ cm\ (7.1, 7.3)$
- **Low**: Mean wing length was $\bar{x} = 7.2\ cm\ (5.2, 9.2)$

Show the lower (7.1) and upper (7.3) confidence interval (CI)

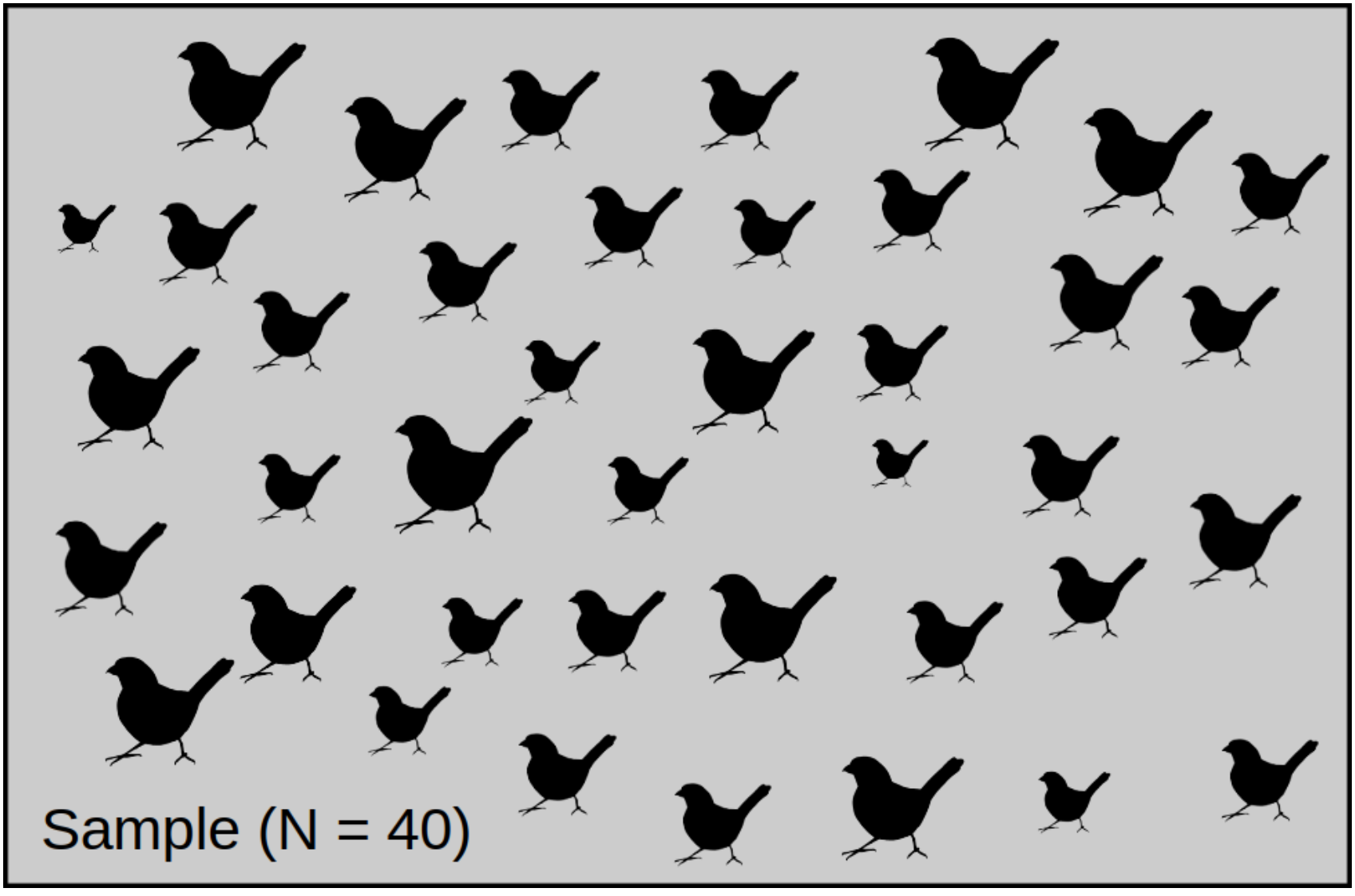
The goal of CIs is to contain uncertainty:

- How frequently will the CIs that we draw *contain* the population mean (μ_x)?
- Logic is again based on repeated resampling

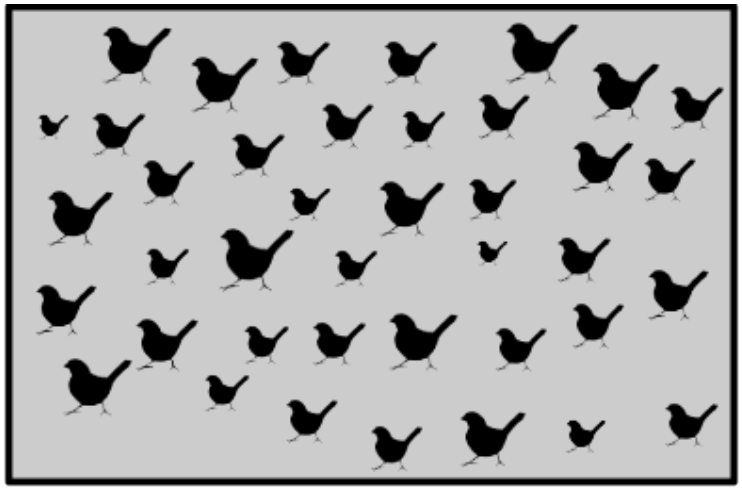
Sampling from a population (confidence intervals)



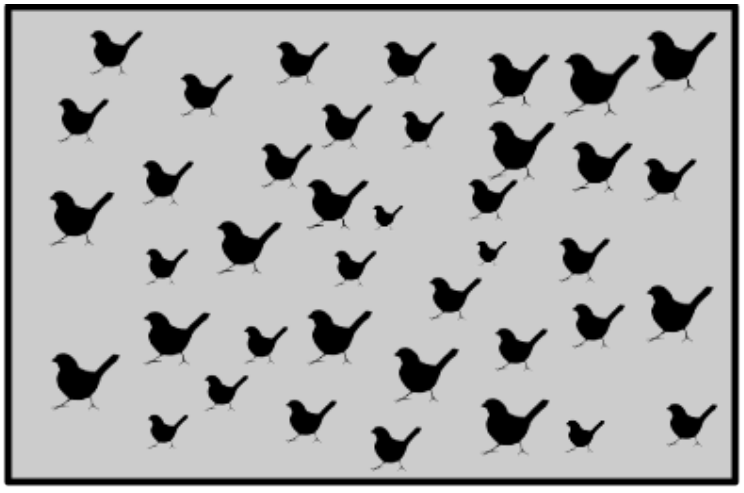
Consider the distribution of a sample of N = 40



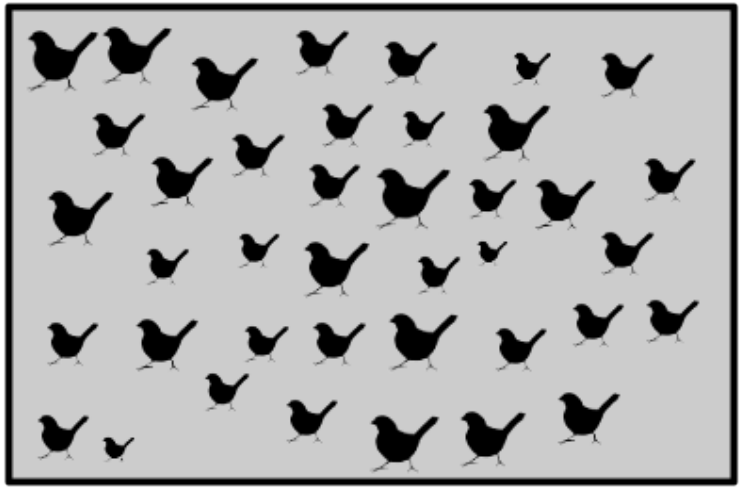
Sample means given true mean of 7.2 when N = 40



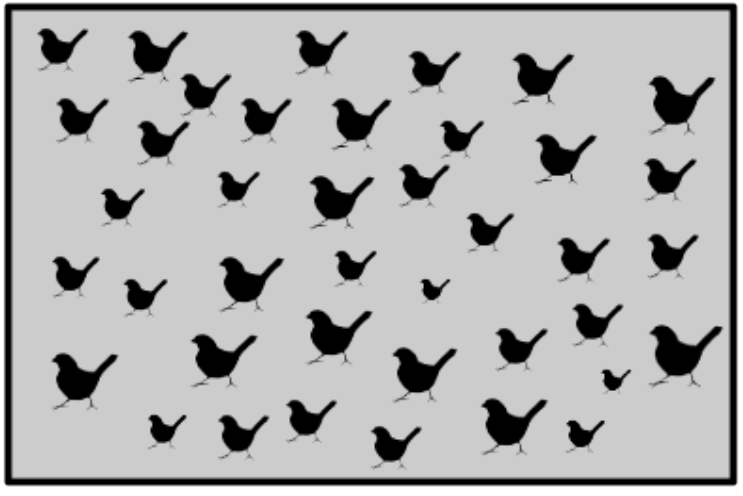
Sample mean = 6.59 cm



Sample mean = 7.45 cm

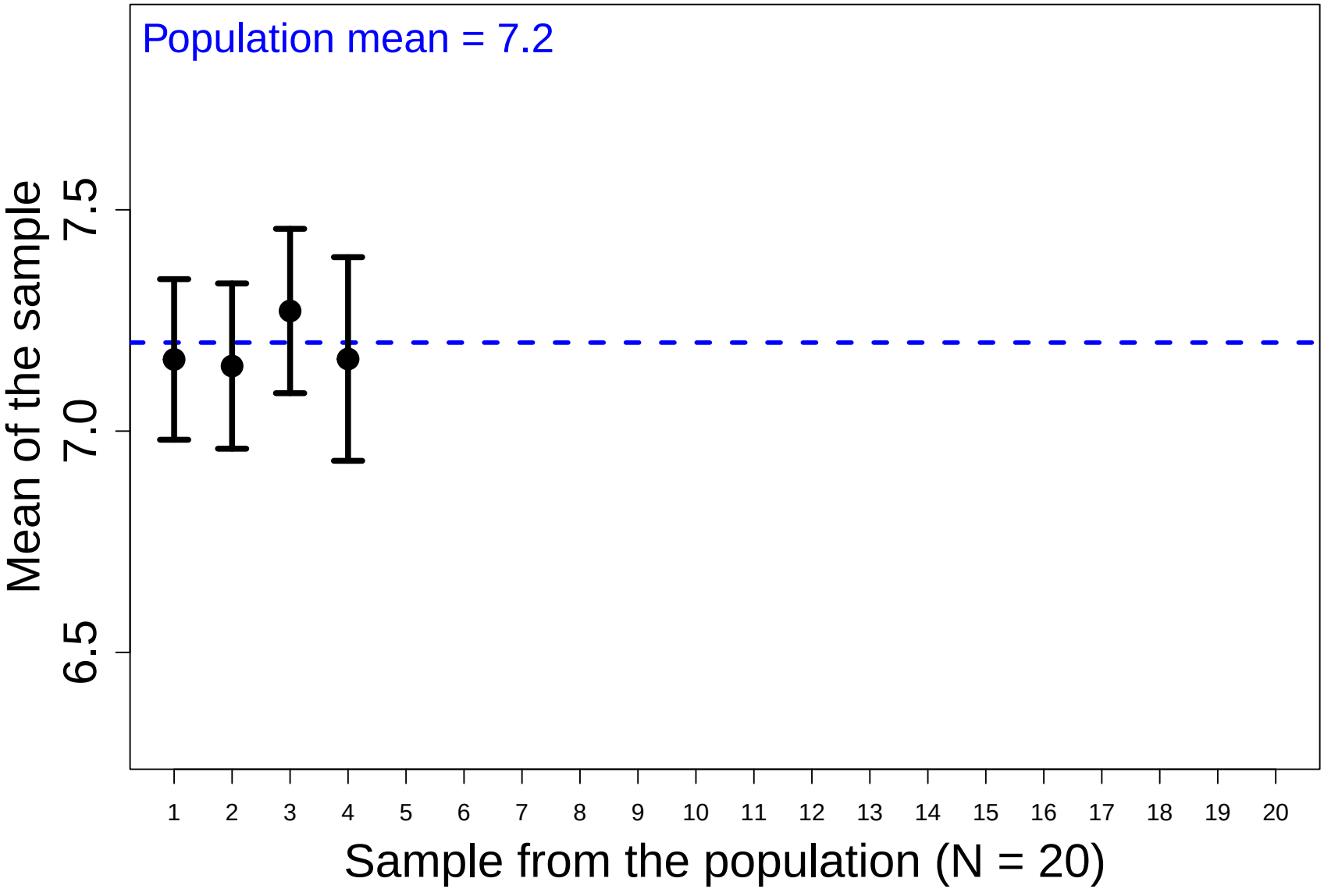


Sample mean = 7.03 cm

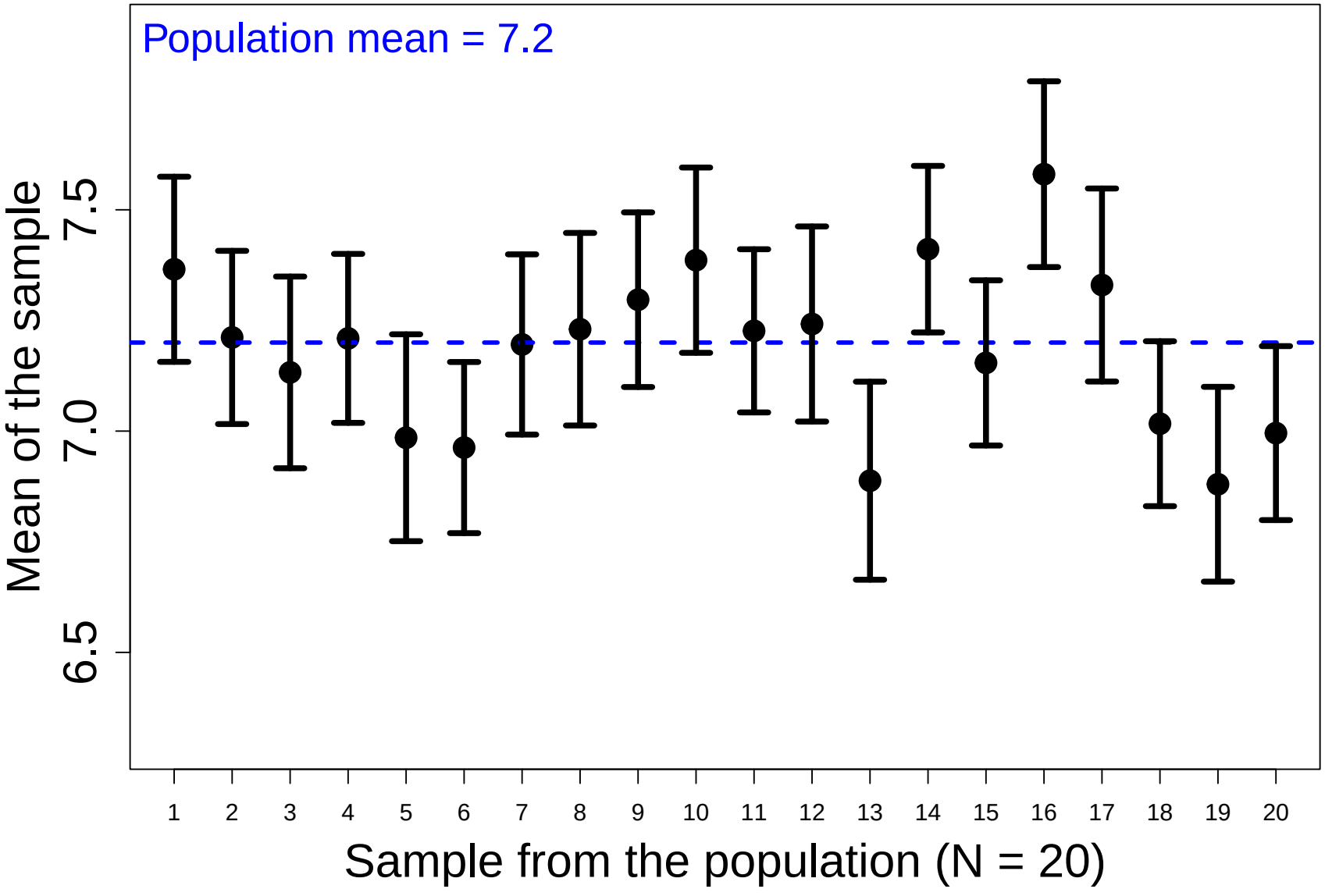


Sample mean = 6.50 cm

Confidence in a sampled mean value (80%)



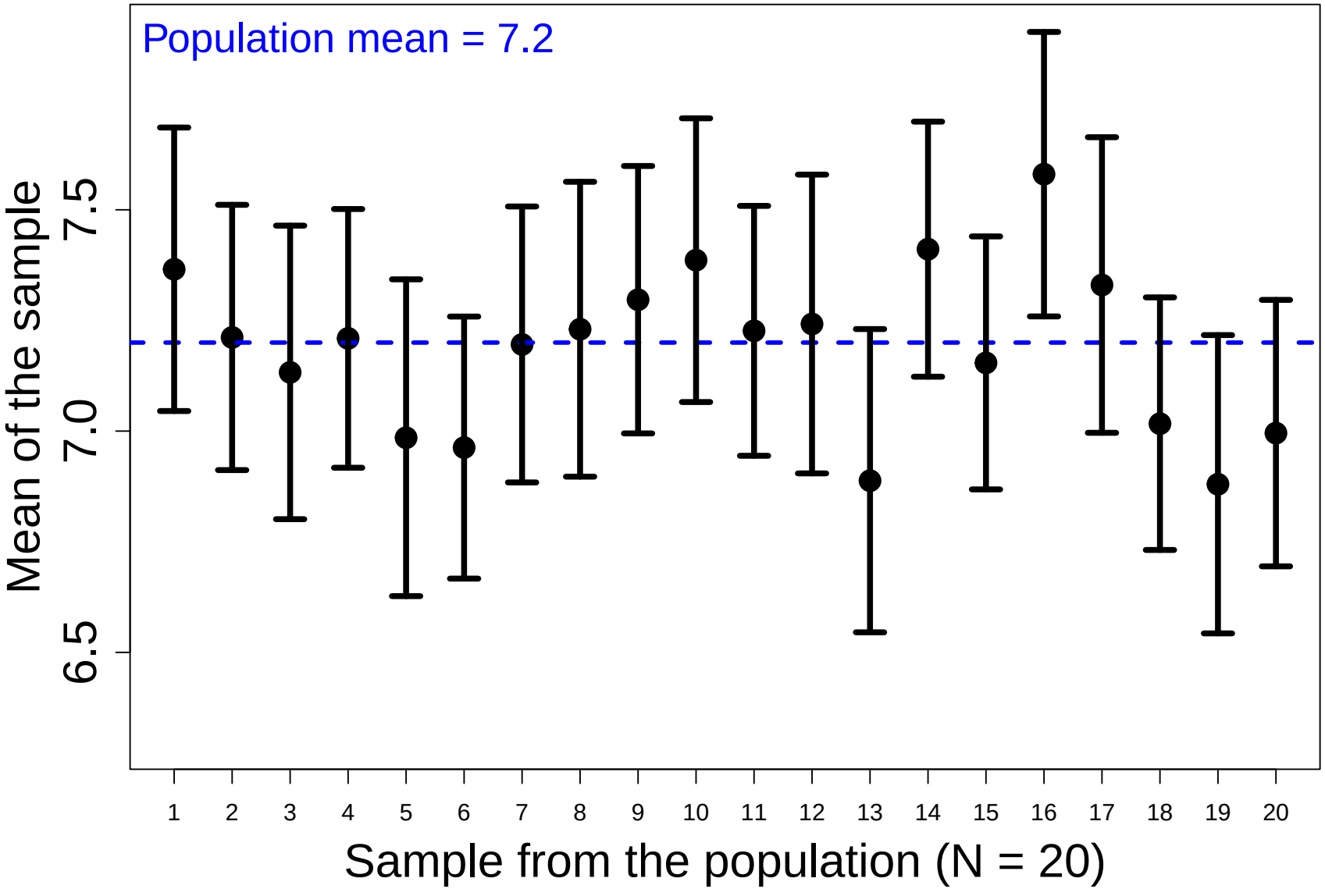
Confidence in a sampled mean value (80%)



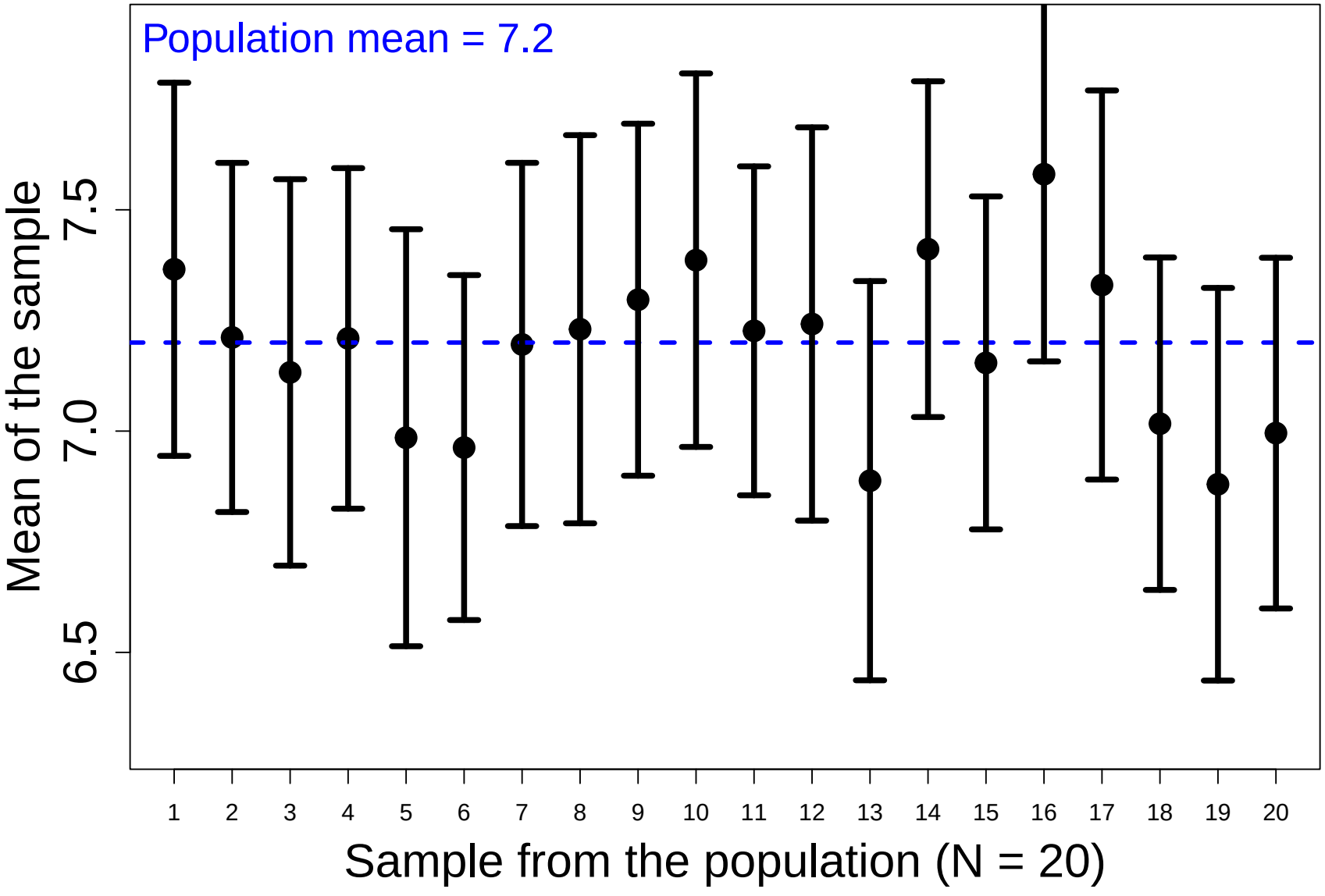
The goal of CIs is to contain uncertainty:

- How frequently will the CIs that we draw *contain* the population mean (μ_x)?
- **Having higher confidence requires wider intervals**

Confidence in a sampled mean value (95%)



Confidence in a sampled mean value (99%)



Tools used to calculate CIs:

- Distribution probabilities (e.g., z-scores)
- The standard error ($s/\sqrt{N}$)

Concrete example: mean of 7 temperatures ($^\circ C$)

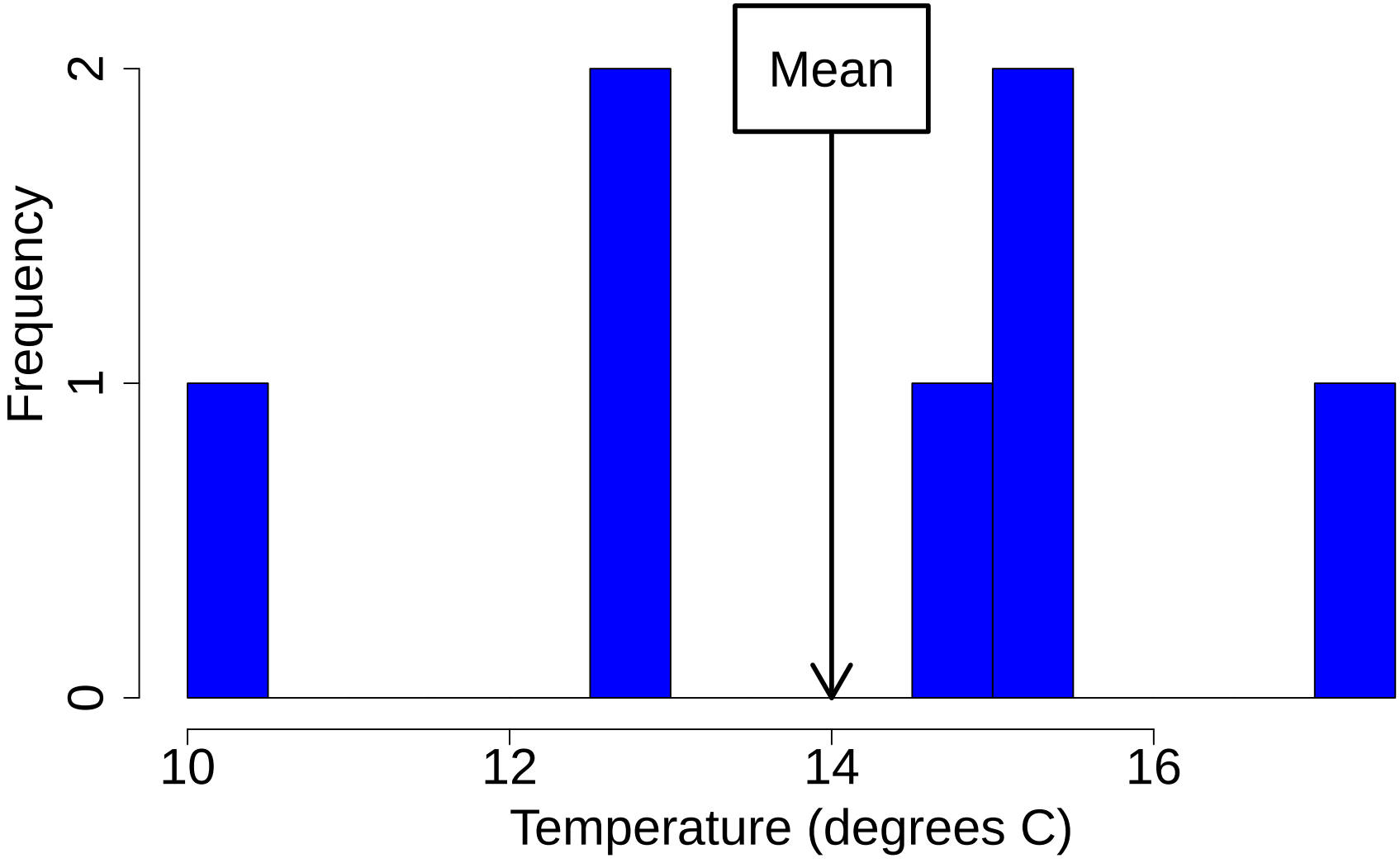
Table 1: Seven values (x) of soil temperature ($^\circ C$) at a site

x_1	x_2	x_3	x_4	x_5	x_6	x_7
17.1	15.2	14.9	12.6	15.2	10.3	12.7

$$\bar{x} = 14$$

95% Confidence in $\bar{x}$?

Sample mean could be far off the population mean



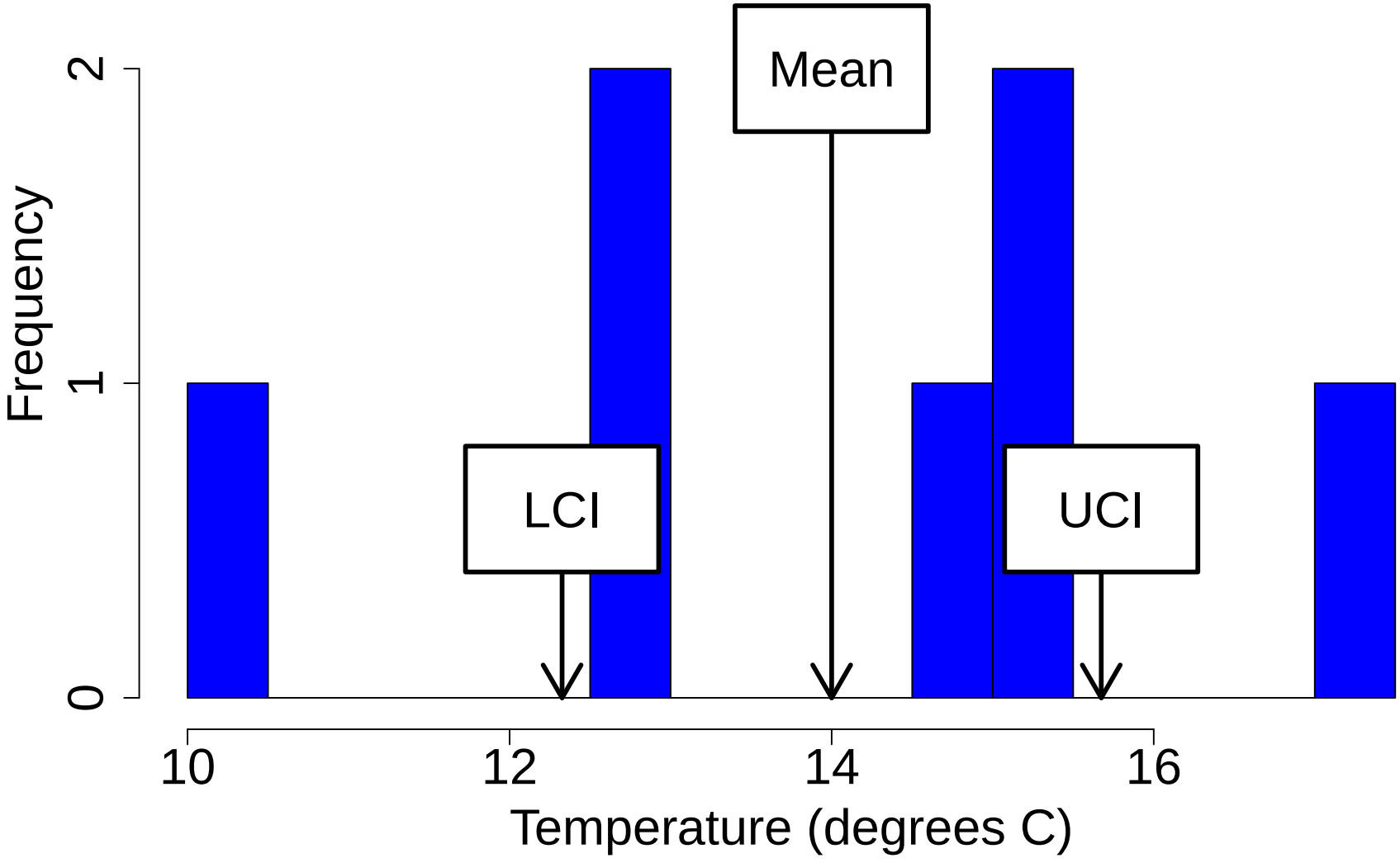
Concrete example: mean of 7 temperatures ($^\circ C$)

x_1	x_2	x_3	x_4	x_5	x_6	x_7
17.1	15.2	14.9	12.6	15.2	10.3	12.7

$$\bar{x} = 14 \pm 1.6740$$

95% CIs: (12.34, 15.67)

Confidence intervals (95% CIs) around mean of 14 cm

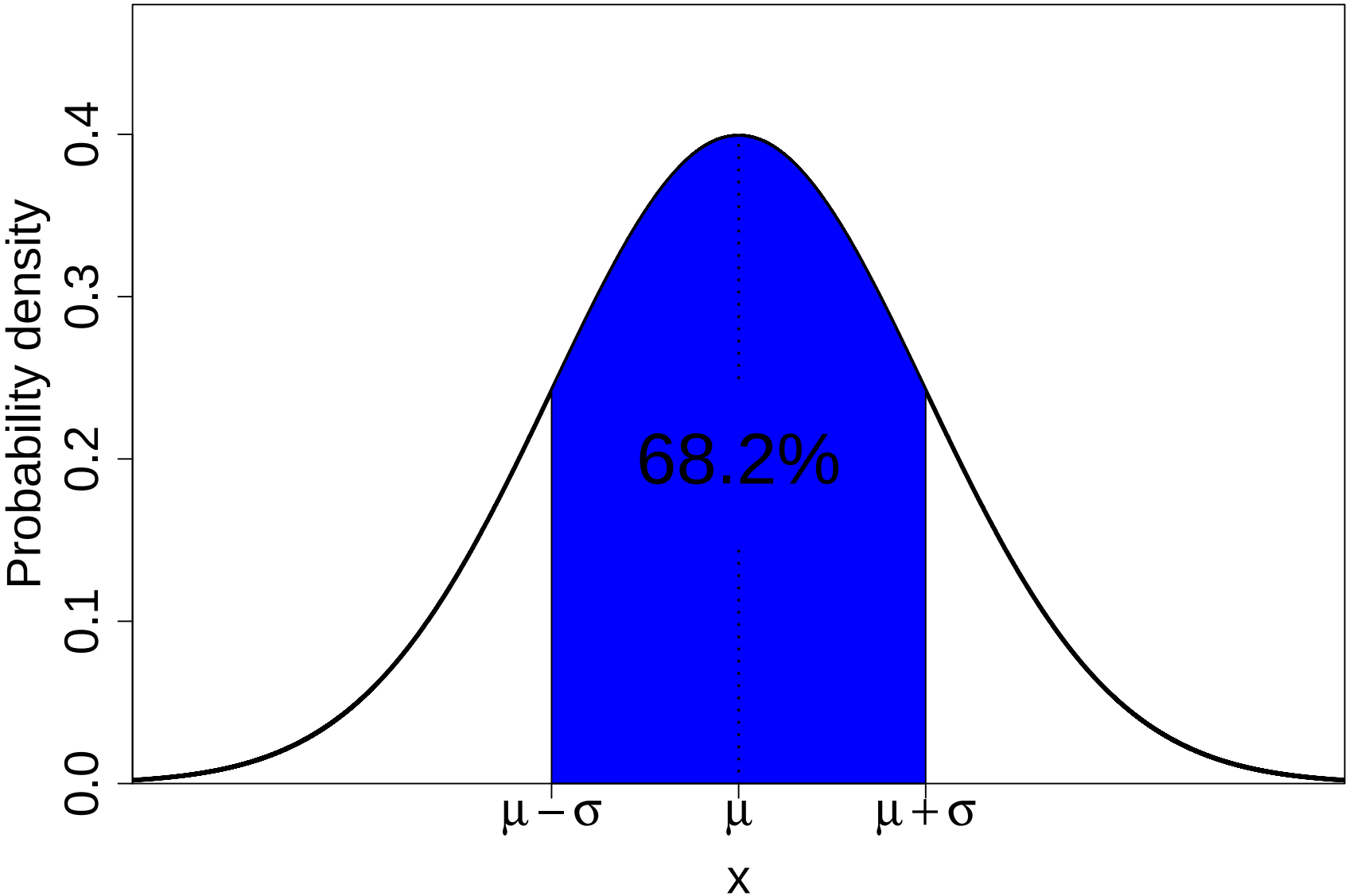


What do we actually mean by 95% confidence?

x_1	x_2	x_3	x_4	x_5	x_6	x_7
17.1	15.2	14.9	12.6	15.2	10.3	12.7

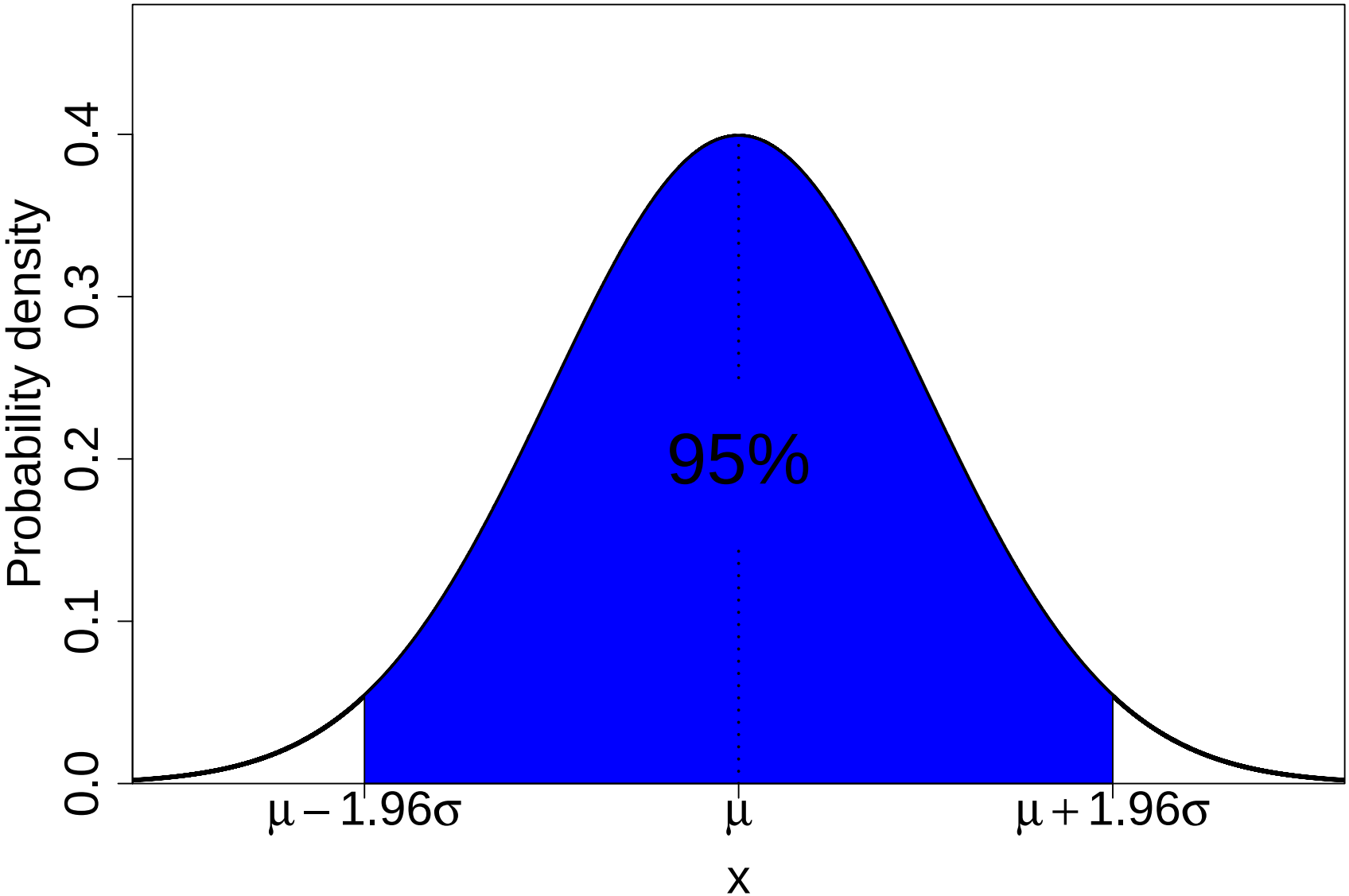
- If we were to repeat our sample (N $= 7$) and calculate CIs over and over again, 95% of the time our intervals would contain the population mean (which we cannot know)
- Not (quite) the same as 95% probability that the mean is between our CIs

But what are we actually calculating?



[1] https://bradduthie.github.io/stats/app/zandp/

But what are we actually calculating?



[1] https://bradduthie.github.io/stats/app/zandp/

How do we calculate confidence intervals?

General idea: Lower (LCI) and upper (UCI) confidence interval

Verbally

- LCI = Mean – (z-score × standard error)
- UCI = Mean + (z-score × standard error)

Mathematically

- $\text{LCI} = \bar{x} - z\left(\frac{s}{\sqrt{N}}\right)$
- $\text{UCI} = \bar{x} + z\left(\frac{s}{\sqrt{N}}\right)$

How do we calculate confidence intervals?

x_1	x_2	x_3	x_4	x_5	x_6	x_7
17.1	15.2	14.9	12.6	15.2	10.3	12.7

$N = 7 \quad \bar{x} = 14 \quad s = 2.26 \quad z = 1.96$

$$LCI = \bar{x} - z\left(\frac{s}{\sqrt{N}}\right)$$

$$= 14 - 1.96\left(\frac{2.26}{\sqrt{7}}\right)$$

$$= 12.34$$

How do we calculate confidence intervals?

x_1	x_2	x_3	x_4	x_5	x_6	x_7
17.1	15.2	14.9	12.6	15.2	10.3	12.7

$N = 7 \quad \bar{x} = 14 \quad s = 2.26 \quad z = 1.96$

$$UCI = \bar{x} + z\left(\frac{s}{\sqrt{N}}\right)$$

$$= 14 + 1.96\left(\frac{2.26}{\sqrt{7}}\right)$$

$$= 15.67$$

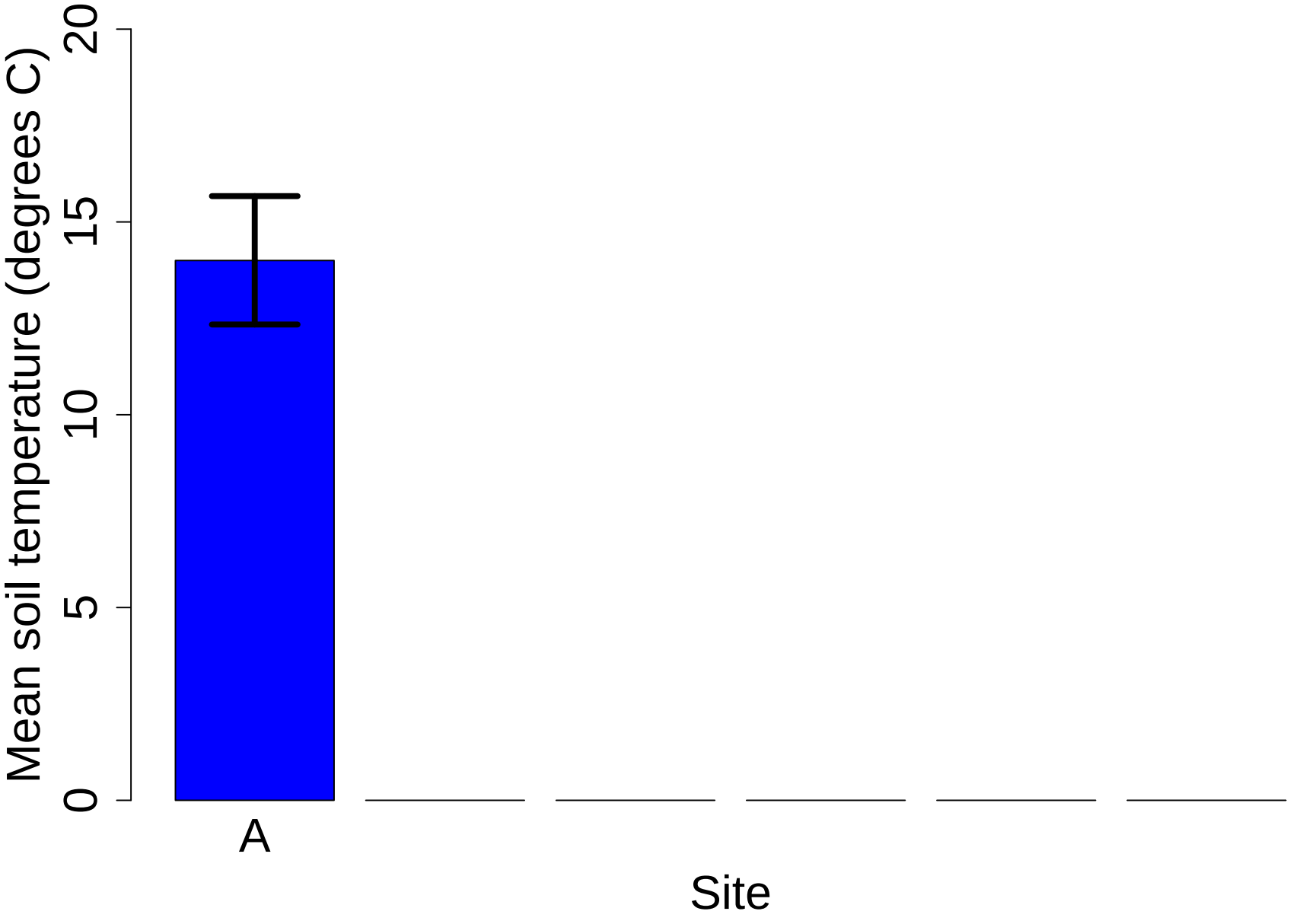
How do we calculate confidence intervals?

x_1	x_2	x_3	x_4	x_5	x_6	x_7
17.1	15.2	14.9	12.6	15.2	10.3	12.7

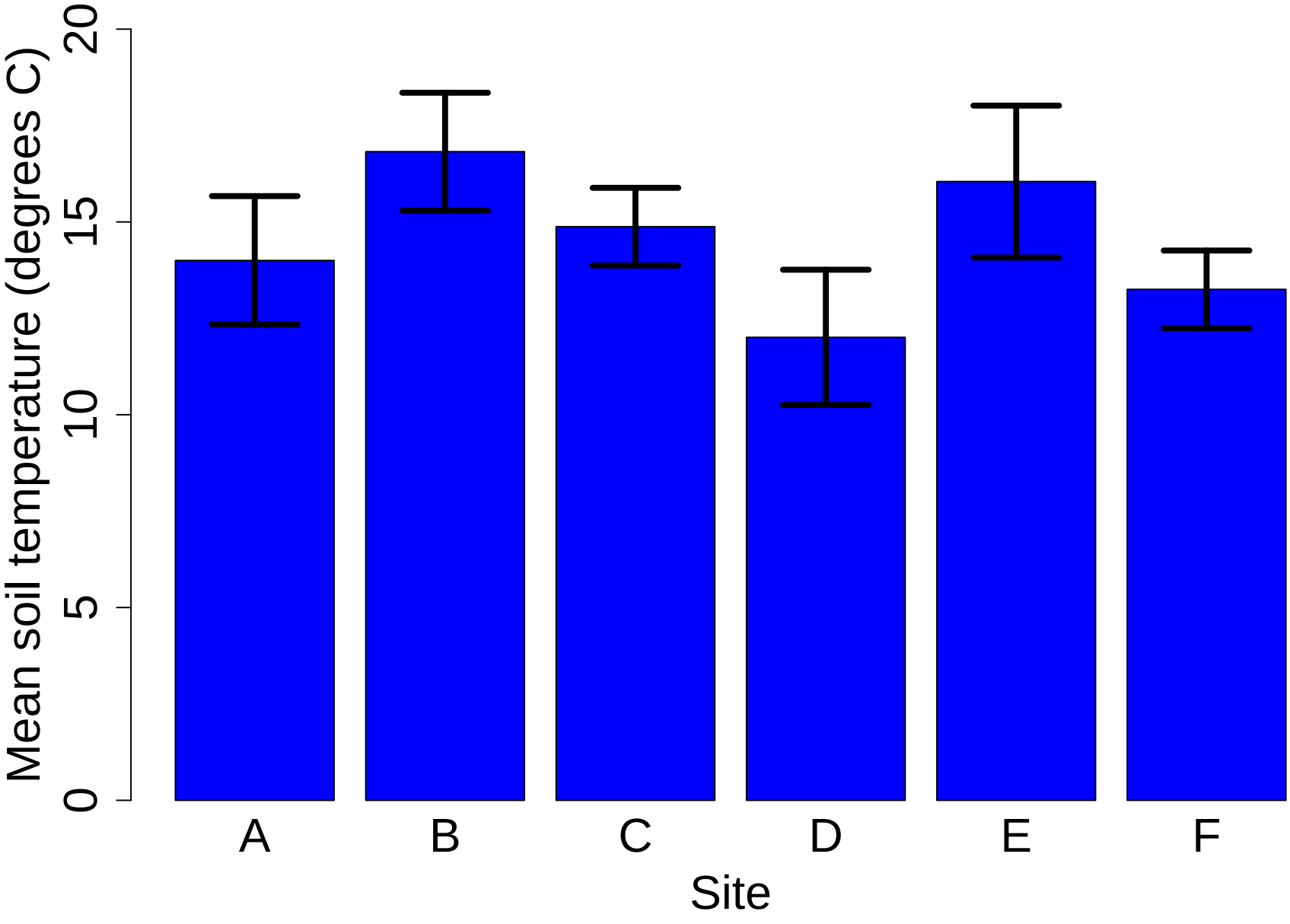
$N = 7 \quad \bar{x} = 14 \quad s = 2.26 \quad z = 1.96$

$$LCI = 12.34 \qquad UCI = 15.67$$

Visualising a confidence interval



Visualising a confidence interval



Accuracy of confidence intervals

- Assumes sample means are normally distributed around the population mean
- Applies for high sample size due to central limit theorem
- What if assumption is violated?

What about for proportions?

Player ID	OS	Dam size	Biodiversity
1	Android	small	0
2	Android	large	34
3	iPhone	small	14
4	Android	large	72
5	iPhone	medium	3

CIs around the proportion of Android users?

Confidence intervals of a proportion

Total players: $N = 74$

- 56 Android
- 18 iPhone

$$Pr(Android) = 56/74 = 0.757$$

CIs around this proportion?

Confidence intervals of a proportion: Wald interval

Assume a normal distribution?

- $\mathrm{LCI} = \bar{x} - z\left(\frac{s}{\sqrt{N}}\right)$
- $\mathrm{UCI} = \bar{x} + z\left(\frac{s}{\sqrt{N}}\right)$

Standard deviation of a binomial:

$$s = \sqrt{p\,(1-p)}$$

$$p = 56/74 = 0.757$$

Confidence intervals of a proportion: Wald interval

Assume a normal distribution?

- $\text{LCI} = p - z \times SE$
- $\text{UCI} = p + z \times SE$

Standard deviation of a binomial:

$$s = \sqrt{p\,(1-p)}$$

$$SE = \frac{s}{\sqrt{N}} = \frac{\sqrt{p\,(1-p)}}{\sqrt{N}} = \sqrt{\frac{p\,(1-p)}{N}}$$

Confidence intervals of a proportion: Wald interval

Assume a normal distribution?

- $\text{LCI} = p - z \times SE$
- $\text{UCI} = p + z \times SE$

Standard deviation of a binomial:

- $\text{LCI} = p - z \times \sqrt{\frac{p(1-p)}{N}}$
- $\text{UCI} = p + z \times \sqrt{\frac{p(1-p)}{N}}$

Confidence intervals of a proportion: Wald interval

Assume a normal distribution?

- $\text{LCI} = p - z \times SE$
- $\text{UCI} = p + z \times SE$

Standard deviation of a binomial:

- $\text{LCI} = 0.757 - 1.96\sqrt{\frac{0.757(1-0.757)}{74}}$
- $\text{UCI} = 0.757 + 1.96\sqrt{\frac{0.757(1-0.757)}{74}}$

Confidence intervals of a proportion: Wald interval

Assume a normal distribution?

- LCI $= p - z \times SE$
- UCI $= p + z \times SE$

Standard deviation of a binomial:

- LCI $= 0.659$
- UCI $= 0.855$

Confidence intervals of a proportion: Wald interval

Problems with the Wald interval

- Wald interval is often inaccurate
- Problematic when p near 0 or 1

Other ways to calculate proportion CIs

- Compare methods with simulated data
- Clopper-Pearson method used by jamovi
- Clopper-Pearson wider than necessary
- Based on the binomial distribution

A bigger problem with the normal distribution and CIs

Assuming we know population standard deviation (σ)

- Only have *sample* estimate, s
- Need to account for this uncertainty
- Normal distribution will be biased
- t-distribution will correct for bias