Estimating a mean

Sampling

Sampled data comes from a **population**.

When sampling some property of interest, there will be a *true* mean in this population. This true mean is called the **population mean**. But most of the time it is impossible to know it exactly.

So we have to *estimate* it. An estimation of the population mean from a given sample is called the **sample mean**.

The sample mean is calculated as the average of a random sample:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Example

Six data points were sampled from a distribution with unknown mean μ:

$$21, 15, 20, 13, 12, 17.$$

Estimate the mean μ based on these data points.

How can we be sure of the accuracy of this result? And how can we quantify the accuracy?

The key idea will be that, whenever a random sample is taken, the resulting sample mean is also a random variable. So correctly interpreting our data relies on also knowing the *distribution of the sample mean*.

Looking forward

The idea behind measuring the accuracy of a sample mean will be to:

- Estimate the variance. A population with large variance will yield samples with greater variation than a population with a small variance. Compare:
  - Smaller variance: $21, 15, 20, 13, 12, 17$
  - Larger variance: $28, 18, 9, 1, 14, 9$
- Use the estimated variance and the size of the sample to approximate the distribution of the sample mean. This is called the central limit theorem.
- Determine an interval describing an estimated lower and upper bound for the true mean.

Estimating variance

Estimating variance

The estimate for the variance is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

The fact that we use $n-1$ instead of n in the denominator is called **Bessel's correction**, and it ensures we have an unbiased estimation of the variance.

To be unbiased means that

$$E(S^2) = \sigma^2.$$

Roughly speaking, the reason we need Bessel's correction is because $\overline{X}$ is an estimate as well, which introduces a source of bias that would otherwise result in too small of an estimate. To counteract this, a smaller denominator is used, making the entire expression larger, and it can be mathematically proven that doing this is no longer biased.

Example

Estimate the mean μ, the variance σ^2 and the standard deviation σ of the population from which the following sample was taken:

$$18, 19, 17, 19, 13, 11, 18$$

The central limit theorem

Reminder

The idea behind measuring the accuracy of a sample mean will be to:

- Estimate the variance. A population with large variance will yield samples with greater variation than a population with a small variance. Compare:
  - Smaller variance: $21, 15, 20, 13, 12, 17$
  - Larger variance: $28, 18, 9, 1, 14, 9$
- Use the estimated variance and the size of the sample to approximate the distribution of the sample mean. This is called the central limit theorem.
- Determine an interval describing an estimated lower and upper bound for the true mean.

Central limit theorem

Theorem (The central limit theorem)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with finite mean μ and finite variance σ^2. If n is sufficiently large then

$$\overline{X} \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\overset{\text{approx.}}{\sim}$ denotes 'approximately distributed as'.

The central limit theorem tells us about the randomness of the sample mean (as long as our sample size is big enough). In particular, it tells us:

- the expected value of the sample mean is the population mean, and
- how the sample mean will vary (with variance $\frac{\sigma^2}{n}$).
  - you can think of this latter part as encoding a "margin of error".

It is important to take note that this works for "sufficiently large n" (usually $n \geqslant 30$), and really only yields an approximation. If n is small, then more sophisticated techniques are required.

Standard error

Since the central limit theorem includes the variance of the sample mean, we can quantify the margin of error of our estimate.

- Note carefully: the variance of the sample mean is NOT the same as the population variance.

The variance of the sample mean is $\frac{\sigma^2}{n}$, where σ^2 is the population variance and n is the sample size. Thus its standard deviation is $\frac{\sigma}{\sqrt{n}}$.

We use this to define the standard error:

$$\text{SE} = \frac{\sigma}{\sqrt{n}}.$$

But this depends on knowing the true population variance σ^2. Like the population mean, we cannot expect to know this. So we have to use an estimate for that as well.

Key idea: the standard error encodes the margin of error of an estimate.

Calculating confidence intervals

Confidence intervals

We define an **approximate 95% confidence interval** as

$$\overline{X} \pm 1.96 \times \text{SE},$$

where $\overline{X}$ is the sample mean, and SE is the estimate of the standard error.

The reason it is *approximate* is because of the approximation in the central limit theorem.

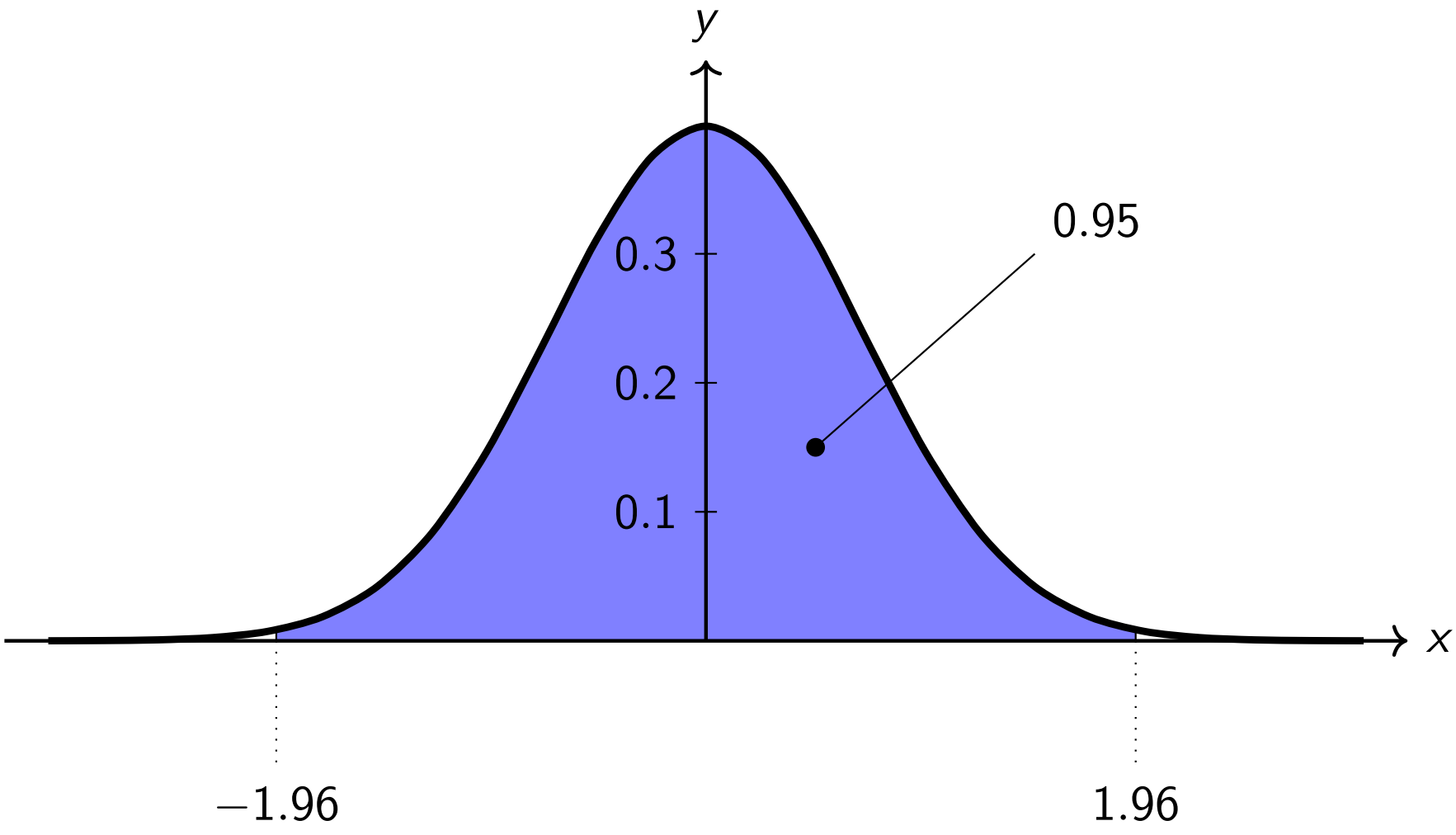
The confidence interval depends on both the sample mean $\overline{X}$ and the standard error SE.

- Both $\overline{X}$ and SE depend on a random sample.
- So $\overline{X}$ and SE are both random.
- It follows that the confidence interval also depends on random factors.

It is called a 95% confidence interval because there is a 95% chance that the interval "lands on" the true mean.

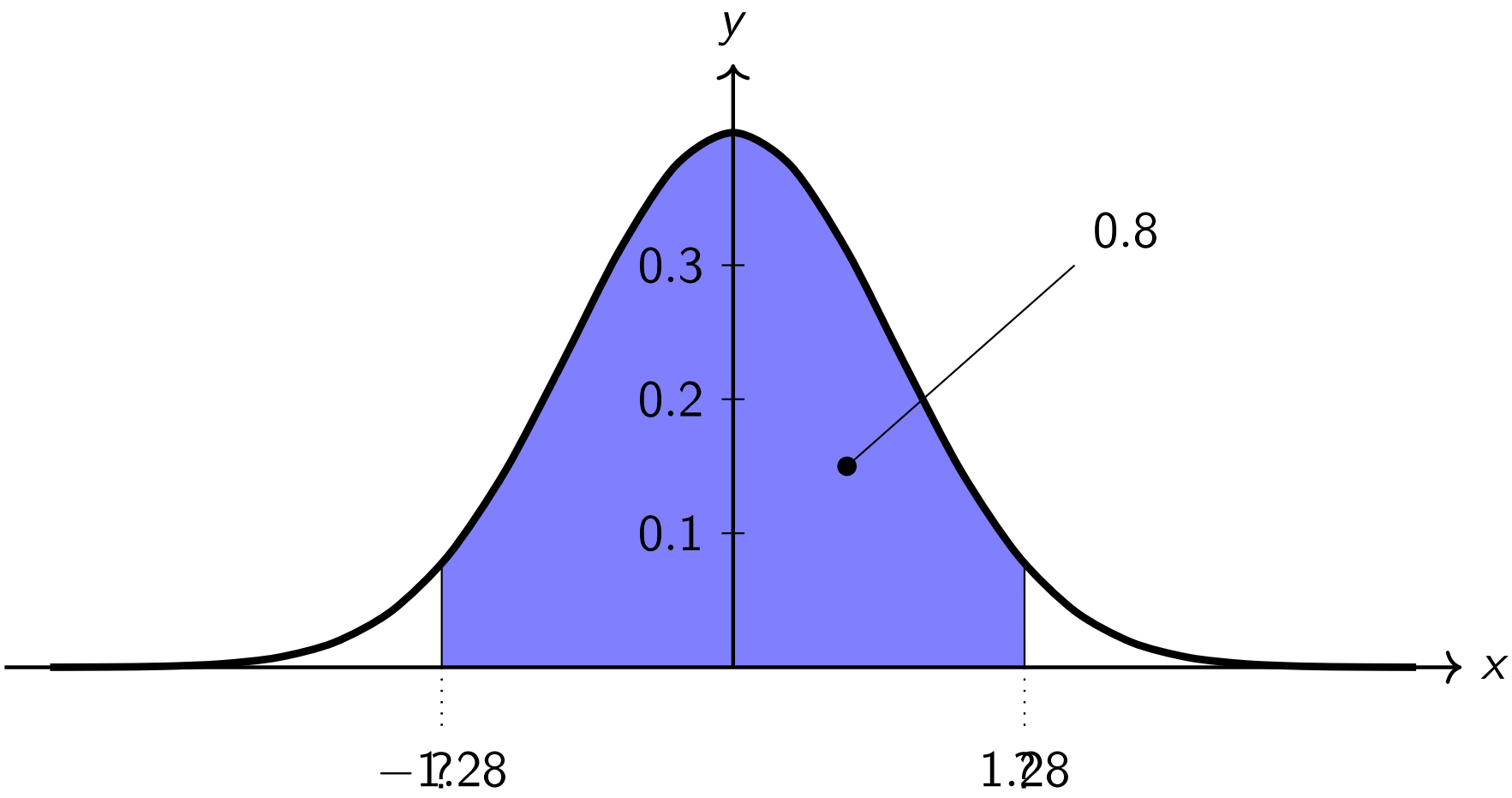
Why 1.96?

The number 1.96 used in the formula for the interval is based on a normal distribution.



Other confidence levels

For other confidence intervals, we have to replace the 1.96 with another appropriate number.



Use `qnorm` in R to obtain the right number. For an 80% confidence interval, use `qnorm((1+0.8)/2)` to obtain 1.28.

Example

Six data points were sampled from a distribution with unknown mean μ:

$$21, 15, 20, 13, 12, 17.$$

Determine an approximate 95% confidence interval for μ.

Steps:

1. Compute the sample mean.
2. Compute the sample variance.
3. Compute the standard error.
4. Substitute into the formula $\overline{X} \pm 1.96 \times$ SE.

Interpreting confidence intervals

It is important to note that a confidence interval is telling you about the margin of error of the estimate. So, for example, a correct interpretation should take on a form like this:

We are 95% confident that the mean product life is between 5 and 10 days

and **NOT** like this:

We are 95% confident that a product will last between 5 and 10 days

To illustrate this, consider the following data set:

$$2, 2, 2, 8, 8, 8$$

You would find an approximate 95% confidence interval of $(2.37, 7.63)$. None of the data points are in this interval! But the mean, 5, is.

Confidence intervals by simulation

Confidence intervals by simulation

This part is done entirely in R.

Confidence intervals for proportions

Proportions

We can also give confidence intervals for proportions and probabilities. Estimating a proportion is simple:

$$\widehat{p} = \frac{\text{number of successes}}{\text{total}}$$

We then have a different way to compute the standard error:

$$\text{SE} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

It is based on the fact that the variance of a Bernoulli trial is $p(1-p)$.

Can be used for:

- Estimating the probability for a Bernoulli trial with unknown probability.
- Estimating the proportion of individuals in a population satisfying some criteria.

Example

We will use R to simulate a weighted coin, then use the techniques above to estimate the probability of flipping heads.