Sampling variation, estimation and confidence intervals

Roger Newson (King's College, London, UK). `roger.newson@kcl.ac.uk`

- A brief recap on samples and populations.
- What a confidence interval is. (What does "95% CI" mean?)
- What decides how wide a confidence interval should be. (Without using too many formulae!)

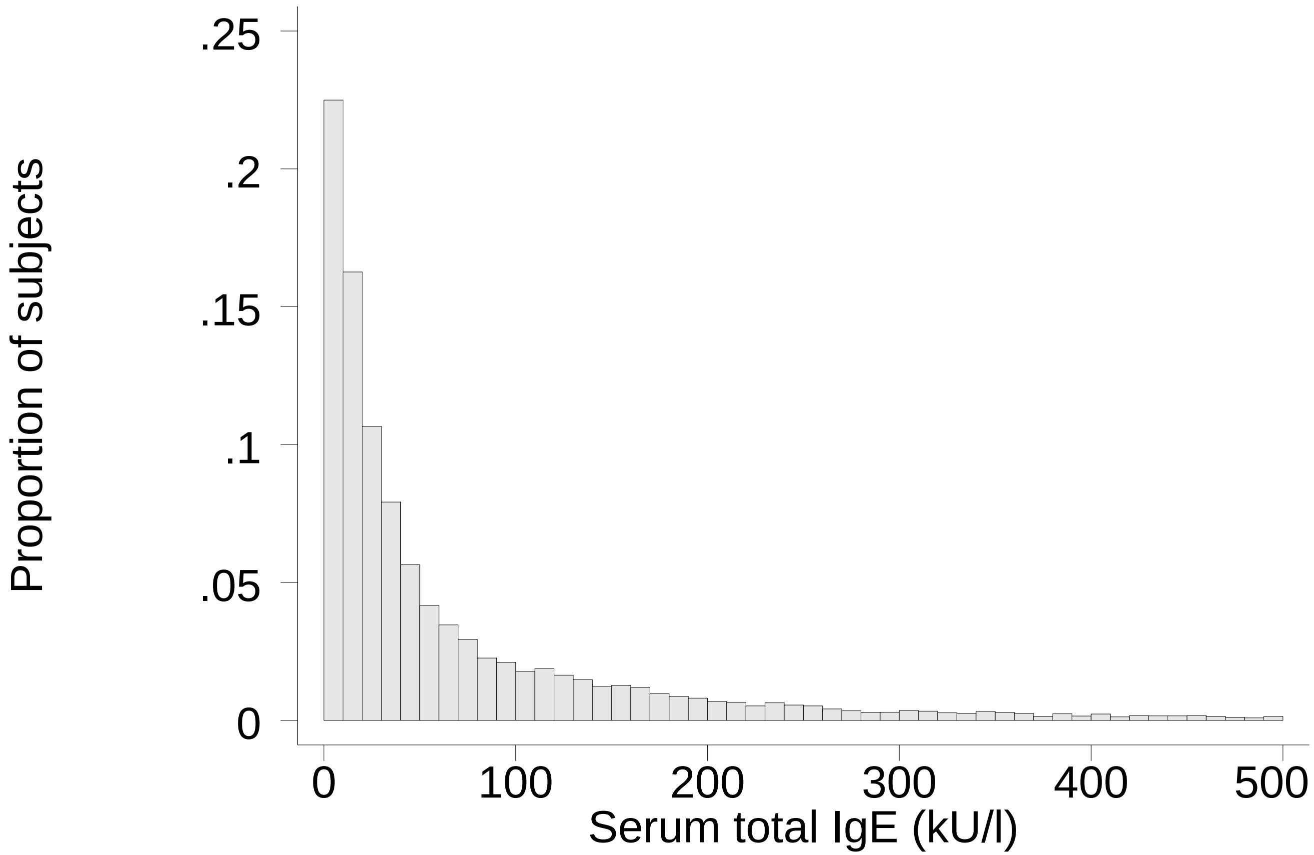
Individuals, samples and populations (recap)

- Measurements on **individuals** are called **variables**. (For instance, blood pressure (mm Hg) or allergy to cats as measured by a skin prick test (yes/no).)
- A **population** is the group of individuals we wish to know about.
- Measurements on **populations** are called **parameters**, and are typically means or proportions, and their differences or ratios.
- The act of making measurements on every individual in a population is called a **census**. *However* ...

Individuals, samples and populations (recap)

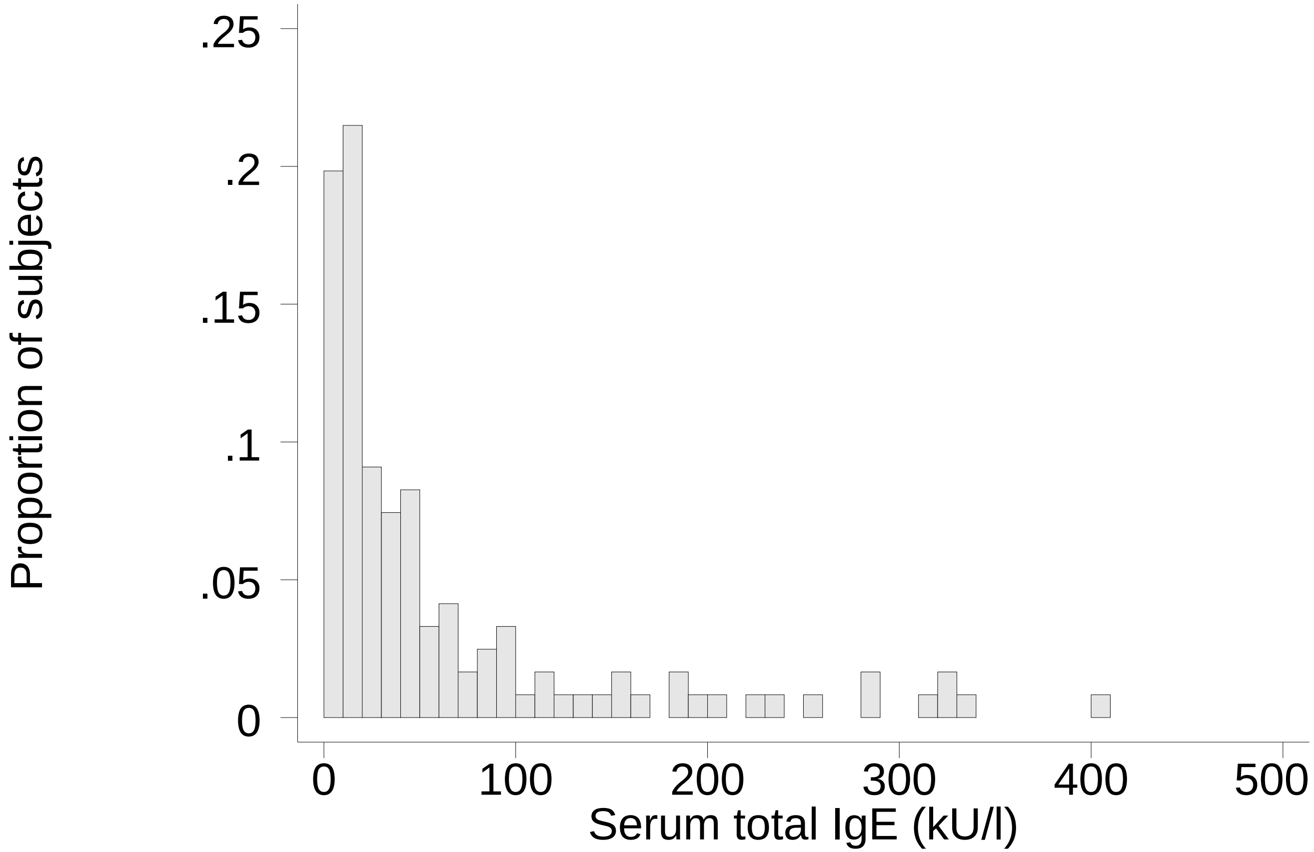
- … a census is nearly always expensive. (And sometimes impossible at any price.)
- *So* we do the next best thing, and take a **sample** of individuals from the population, and make measurements on those.
- The means and proportions, and their differences and ratios, are called **sample statistics**.
- We use them to **estimate** (or "guess") the **population parameters**.

Measurements of serum IgE in a population of 13,554 adults



The population mean is 68.08 kU/l. (We did a census.)

Measurements of serum IgE in a sample of 121 from the population of 13,554 adults



The sample mean is 64.41 kU/l (compared to a *population* mean of 68.08 kU/l).

Sample means and population means

- No matter how carefully you sample, the sample mean is unlikely to be exactly the same as the population mean.
- *However*, the sample mean is usually the best estimate we have.
- *Therefore*, it would be useful to have an idea of how accurate (or inaccurate) it is likely to be.
- (The same applies to sample proportions, sample medians, sample relative risks, and other sample statistics used to measure population parameters.)

The 95% confidence interval (CI)

In the medical literature, you frequently encounter the term "95% CI". It is important for medics to understand what it means.

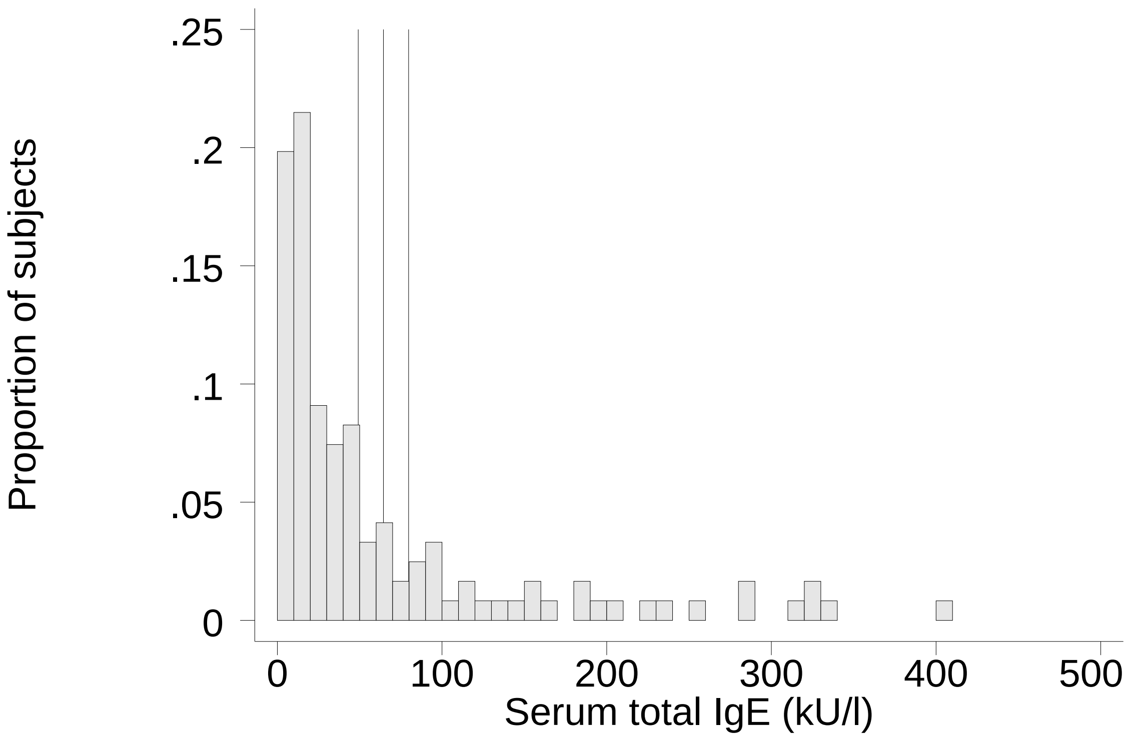
For instance, a group of clinicians might take a sample of 121 adults from the huge population we saw earlier. They might report:

"The mean IgE was 64.41 kU/l (95% CI, 49.11 to 79.71 kU/l)."

This means that:

- The *sample* mean IgE was 64.41 kU/l.
- The clinicians were 95% confident that the *population* mean IgE was between 49.11 kU/l and 79.71 kU/l.

IgE in a sample of 121 patients (with sample mean and 95% CI shown as vertical lines)



The sample mean is 64.41 kU/l (95% CI, 49.11 to 79.71 kU/l).
The *population* mean (unknown to the clinicians) is 68.08 kU/l.
(Note that the 95% CI *does not* include 95% of the sample!)

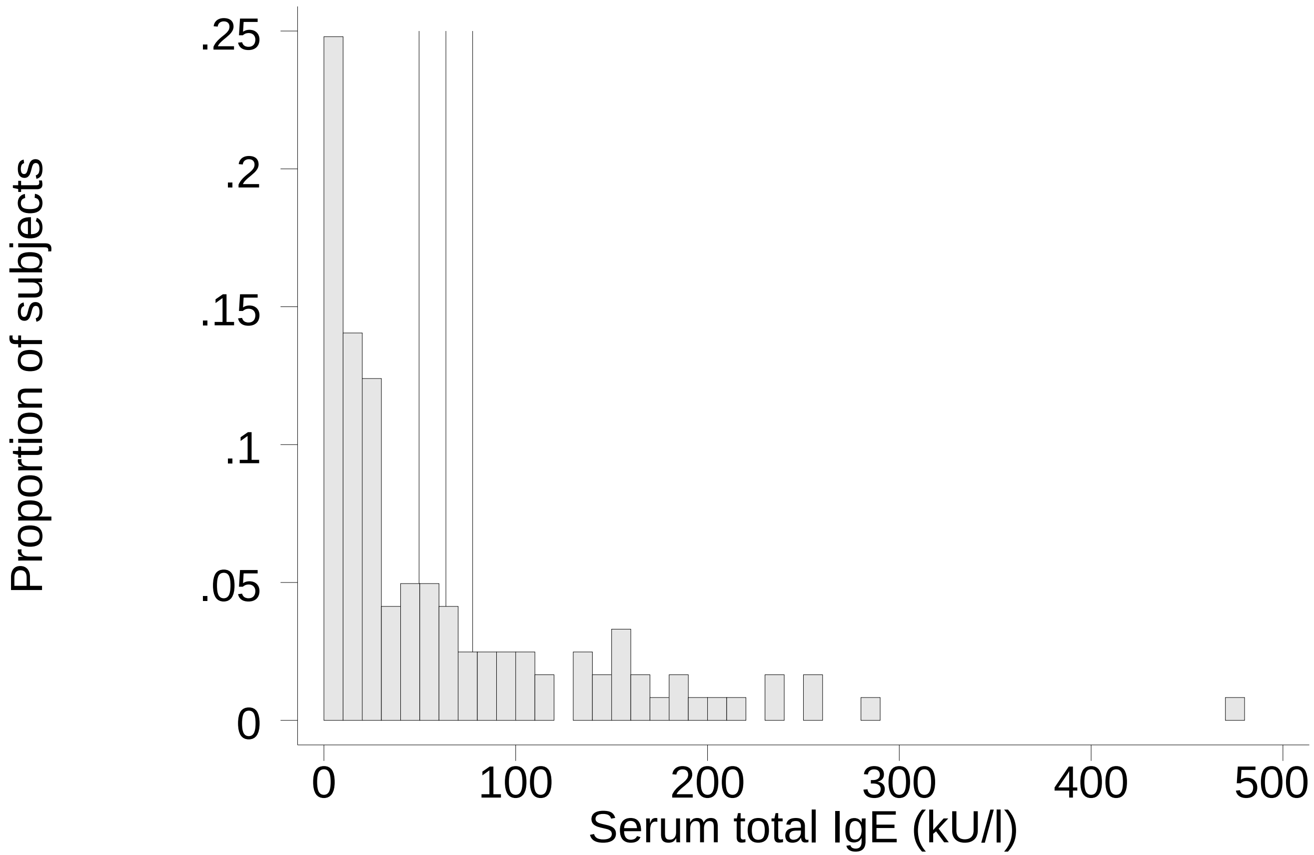
The 95% confidence interval (CI): Definition

- A **95% confidence interval (CI)** for a mean is a range of values, within which we are 95% confident that the true *population* mean will lie.
- It is like a net, attached to the *sample* mean, and spread wide enough to catch the *population* mean in 95% of samples.
- For "mean", you can read "proportion", or "relative risk", or "median difference". Or any other population parameter which we might estimate with a sample statistic.

So why are the clinicians so confident?

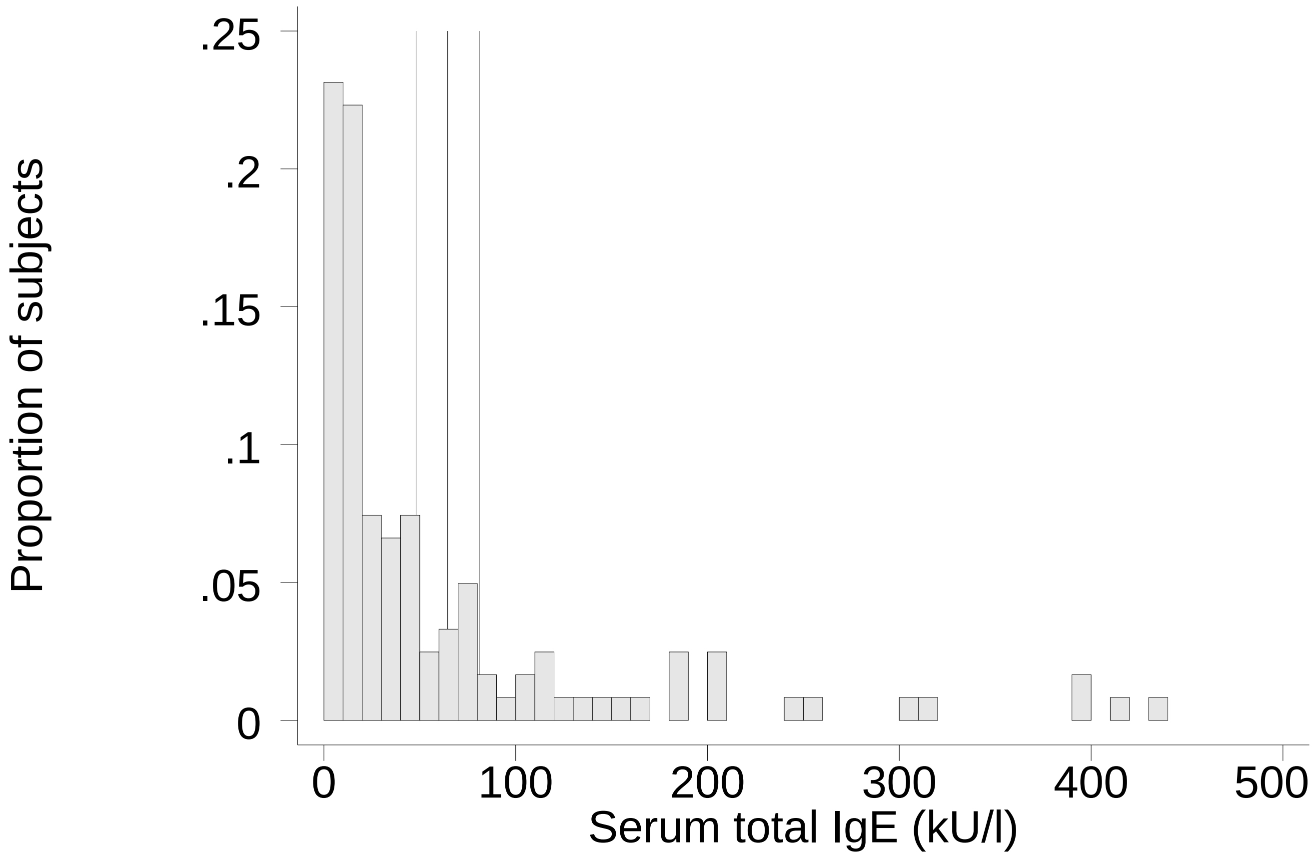
- The clinicians said that they were “95% confident” that the population mean was somewhere in their confidence interval.
- This means that, *supposing* that they had taken a huge number of samples from the same population (instead of just one), and had calculated a 95% CI from *each* of these samples using the same formula, *then* 95% of these CIs would have contained the true population mean.
- You will now see some of the samples the clinicians *might* have taken.

A second sample of 121 patients



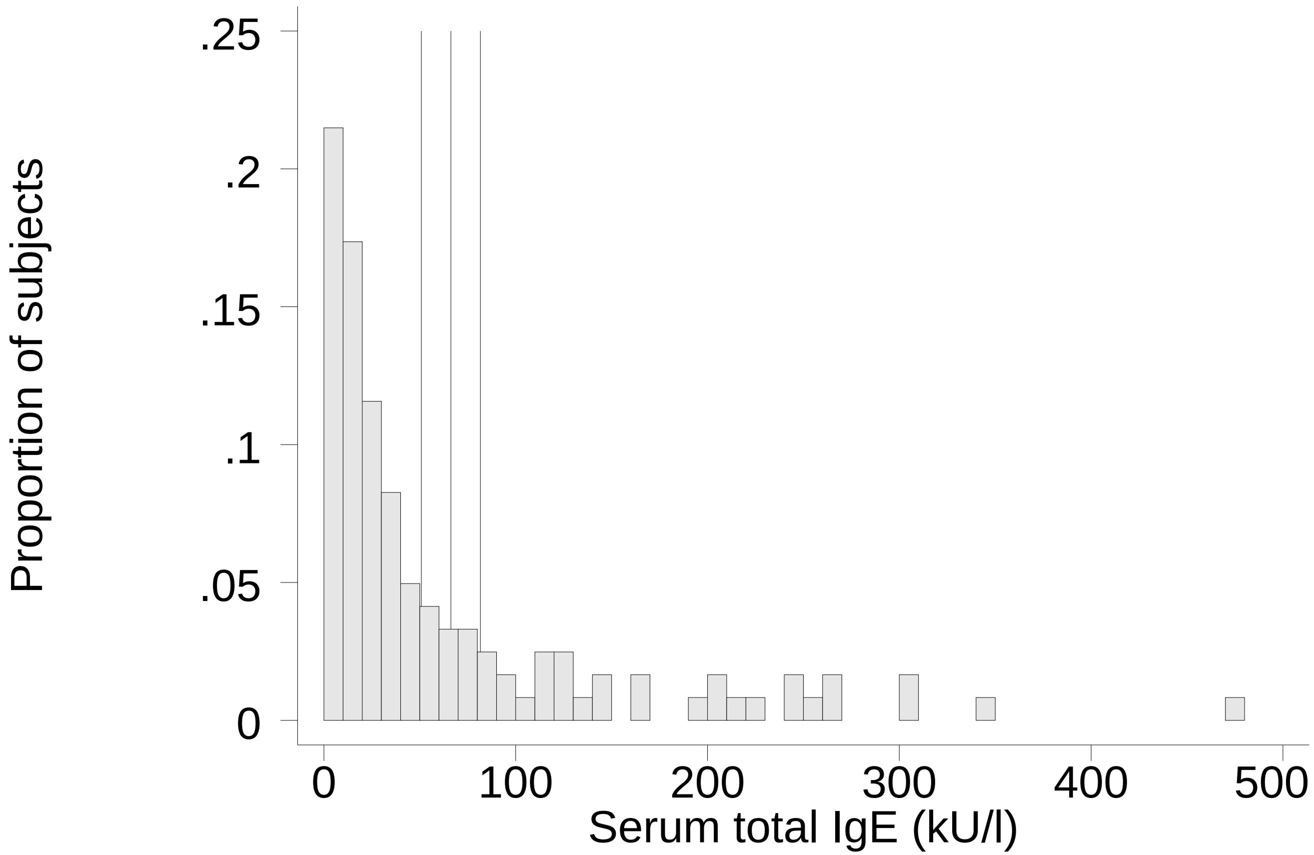
The sample mean is 63.60 kU/l (95% CI, 49.64 to 77.56 kU/l).

A third sample of 121 patients



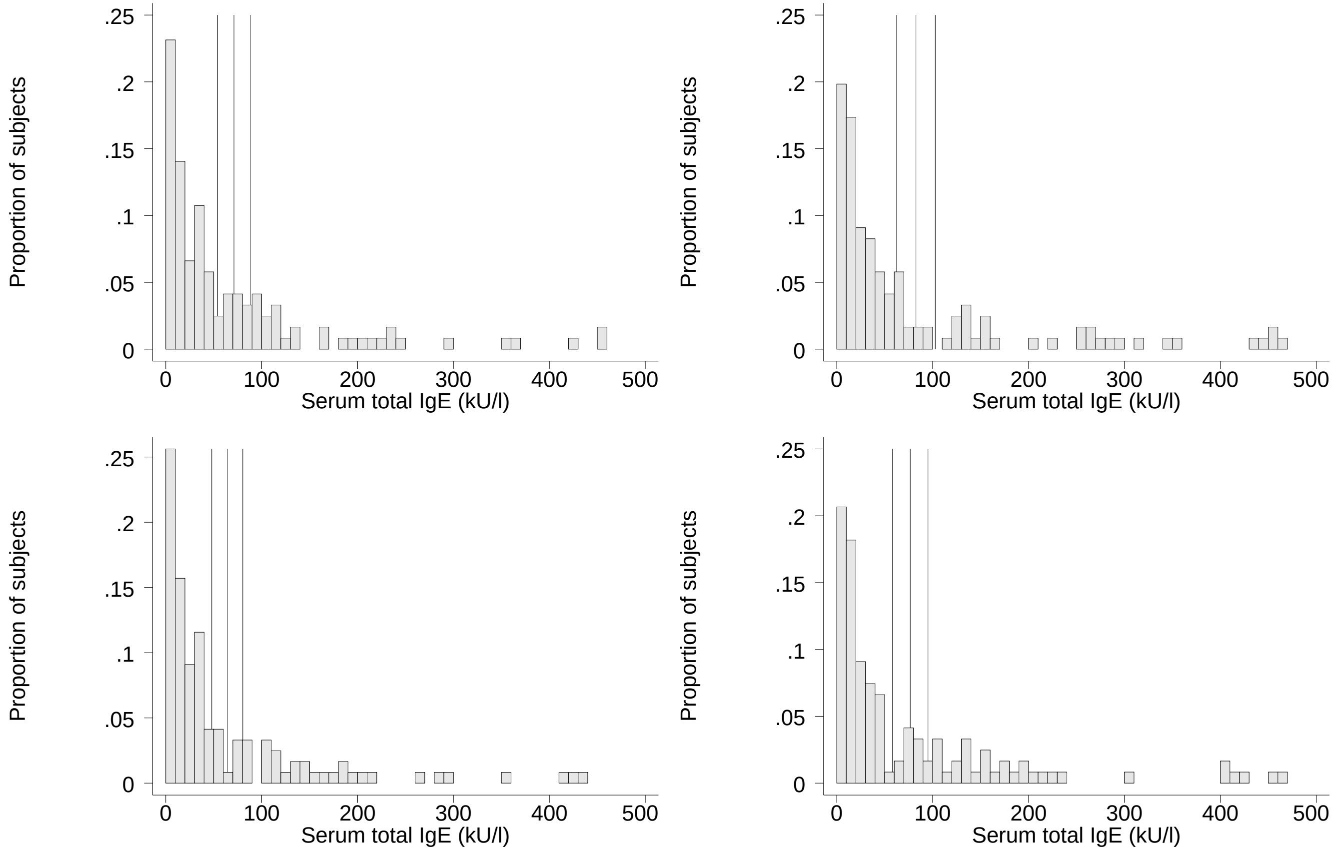
The sample mean is 64.52 kU/l (95% CI, 48.06 to 80.98 kU/l).

A fourth sample of 121 patients



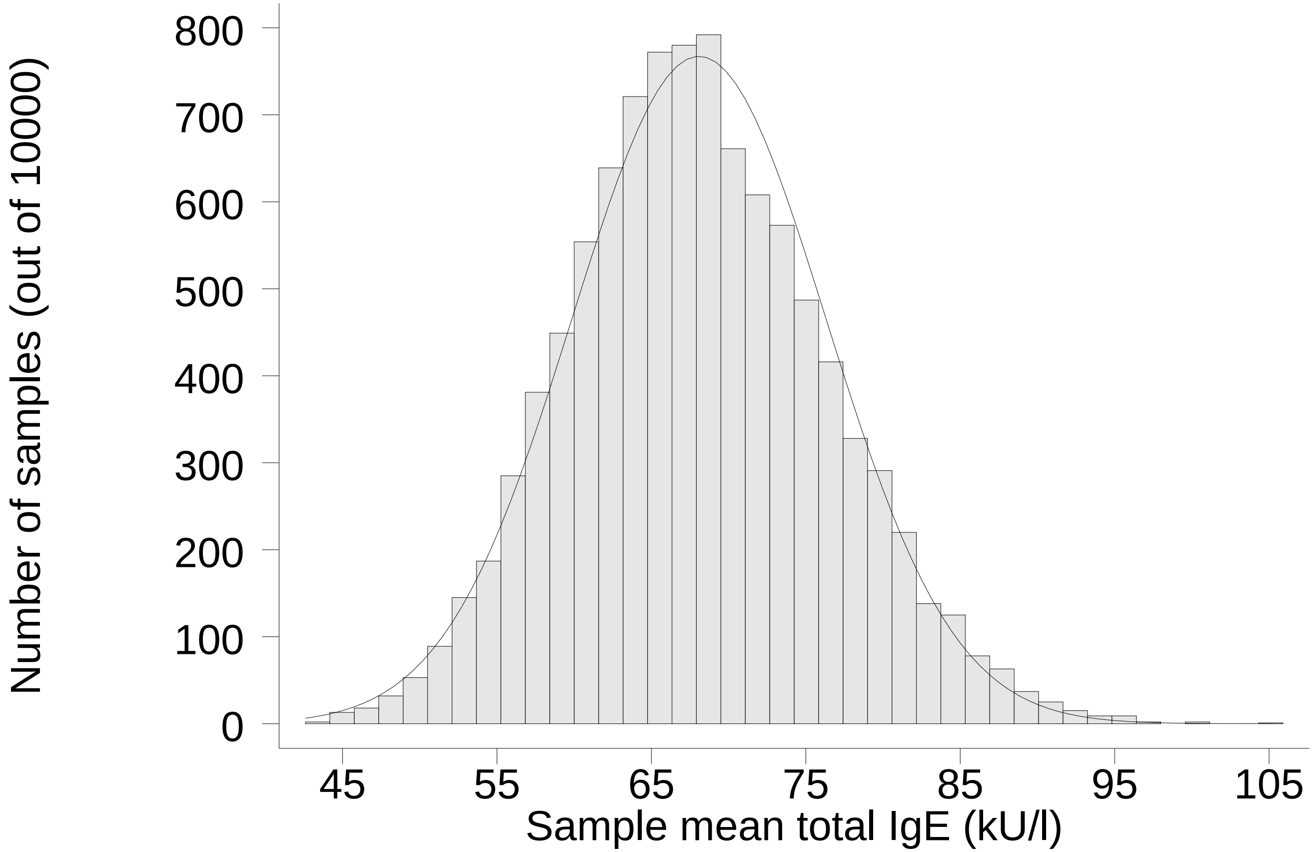
The sample mean is 66.22 kU/l (95% CI, 50.83 to 81.62 kU/l).

Four more samples of 121 patients each



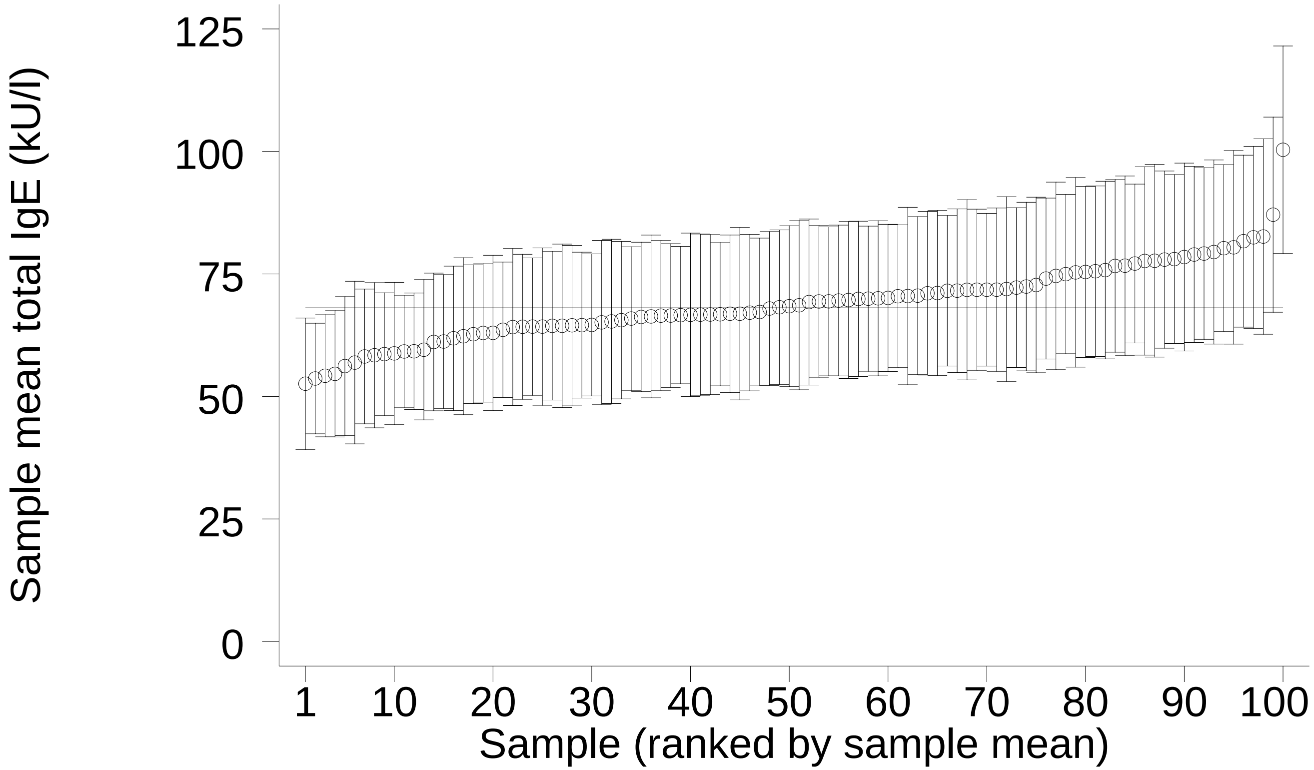
(The *population* mean is still 68.08 kU/l.)

Sample mean IgE values (kU/l) for 10,000 samples of 121 patients each



The *population* mean is 68.08 kU/l. The *sample* means are distributed (approximately) Normally around it.

Sample means and 95% confidence limits for *only 100* of the 10,000 samples, ranked by ascending sample mean



The *population* mean is 68.08 kU/l (horizontal line). All the 95% CIs contain it, except for the lowest 4 and the highest 1.

Summary of the 10,000 samples of 121 patients each

- The population distribution of the IgE values is *not* normal.
- *However*, the 10,000 *sample mean* IgE values were Normally distributed around the population mean.
- Also, of the 10,000 95% CIs, 9376 (94%), which is near enough to 95%, contained the *population* mean.
- So the clinicians had good reasons for being 95% confident that *their* 95% CI (calculated from *one* sample) contained the population mean.

How wide should a 95% confidence interval be?

Statisticians have formulae to calculate this, which medics do not have to learn. However, the width depends mainly on two things:

- The **variability** of the data in the population. (The more variable the population, the wider the CI.)

- The **size** of the sample. (The larger the sample, the *smaller* the CI.)

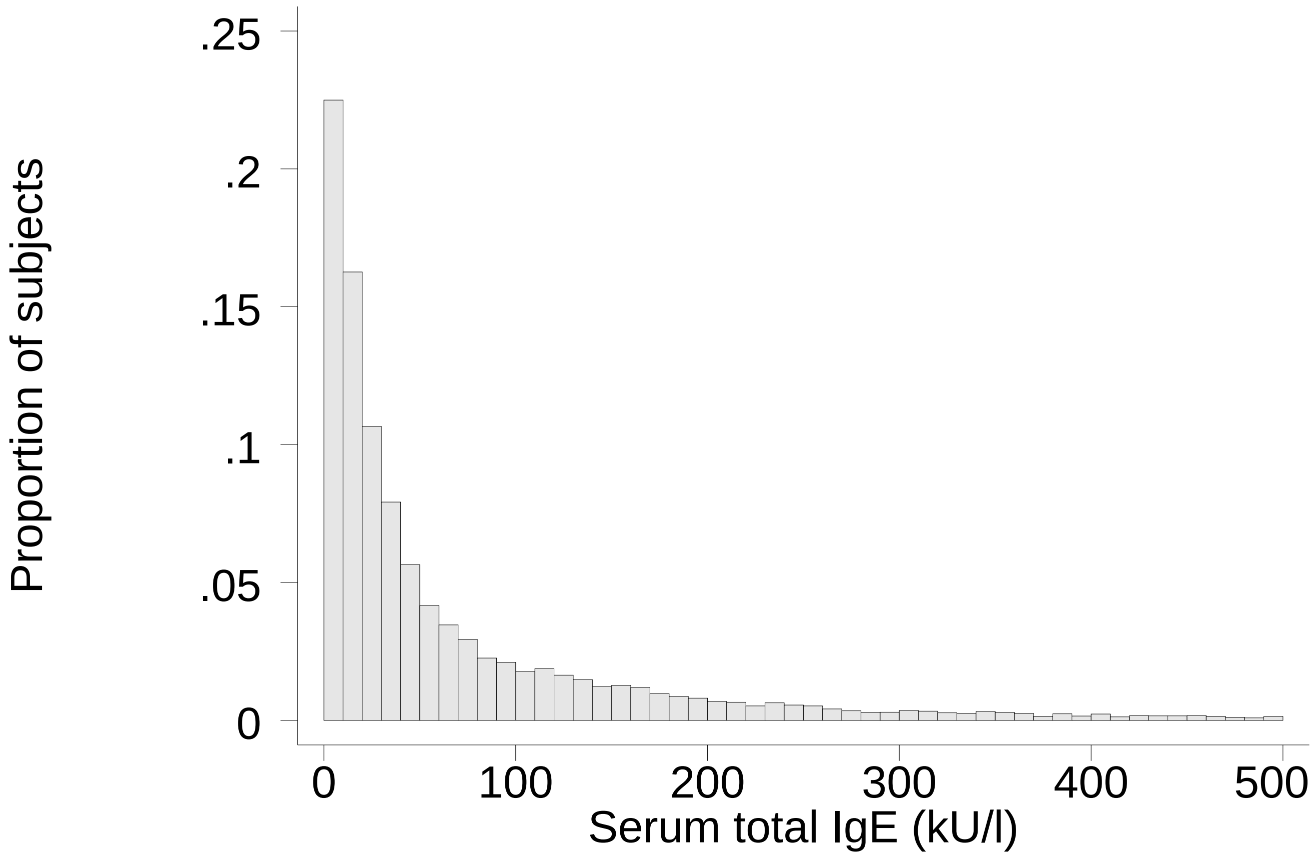
As a rule, the width of the CI obeys an **inverse square law**. This means that, to halve the width of the CI, you must *quadruple* the sample size. (Not just double it.) So, increased precision can be expensive.

Standard errors and the Central Limit Theorem

Sometimes (but not always), confidence intervals are calculated using **standard errors**. The principle behind these is called the **central limit theorem**, which states that, if we take many samples of n from a population:

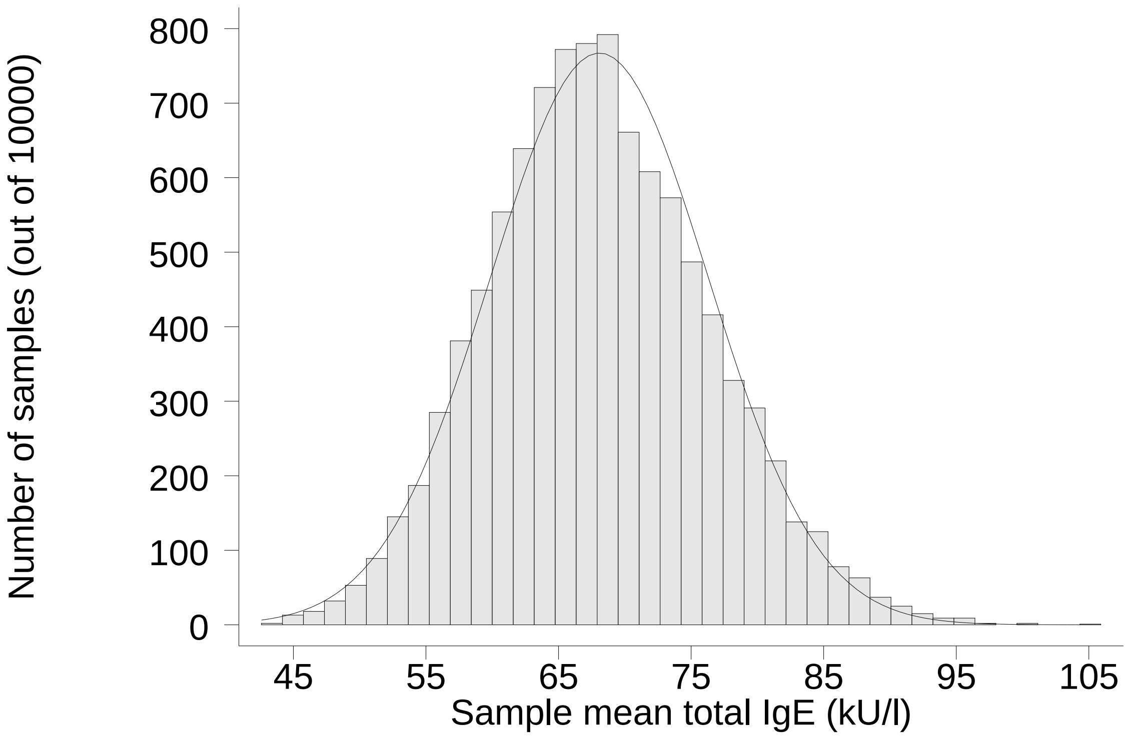
- The sample means have an approximately Normal distribution;
- This Normal distribution is centred on the *population* mean;
- The standard deviation (SD) of this Normal distribution (called the **standard error of the mean**) is equal to Population SD/$\sqrt{n}$.

Measurements of serum IgE in a population of 13,554 adults



The population mean is 68.08 kU/l. The population SD is 90.53 kU/l.

Sample mean IgE values (kU/l) for 10,000 samples of 121 patients each



The sample means are distributed (approximately) Normally, with a mean of 68.08 kU/l (the *population* mean) and a population SE of

$$\text{Population SD}/\sqrt{n} = 90.53/\sqrt{121} = 8.23 \text{ kU/l}.$$

Standard deviations and standard errors

Some people (especially students at exam time) confuse standard *errors* (SEs) with standard *deviations* (SDs). The difference is:

- The standard **d**eviation measures the variability of in**d**ivi**d**uals.
- The standard **e**rror measures the variability of sampl**e** m**e**ans.

The two are different, but are related by the formula $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$, where n is the sample number. So, to halve the standard error, you must *quadruple* the sample number, not just double it.

Standard errors and confidence intervals

- Unfortunately, clinicians (like the ones in our story) usually do not know the population mean or the population standard error (SE).

- Therefore, to calculate a confidence interval, they must *estimate* the population SE by the **sample SE**, which is equal to Sample SD/$\sqrt{n}$.

- The confidence interval then extends from a lower limit of

  $$\textbf{sample mean} - \textbf{multiplier} \times \textbf{sample SE}$$

  to an upper limit of

  $$\textbf{sample mean} + \textbf{multiplier} \times \textbf{sample SE},$$

  where “multiplier” is a number of “SE-units”. So, the bigger the SE, the wider the confidence interval.

Which multiplier to use?

- The choice of a multiplier depends on many things, and medics do not have to know the details.
- However, the most important consideration is the **confidence level**. This is usually 95%, but is sometimes higher (eg 99%) or lower (eg 90%).
- If you want a 95% confidence interval, the multiplier is usually approximately 2 "SE-units", and is *never* less than 1.96 "SE-units".
- So, if you want a 95% confidence interval *and* your sample is large (eg 121), then the confidence interval extends from

  $$\textbf{sample mean} - \mathbf{1.96} \times \textbf{sample SE}$$

  to

  $$\textbf{sample mean} + \mathbf{1.96} \times \textbf{sample SE}.$$

90%, 95% and 99% confidence intervals

If you are taking large samples (> 60), then the multiplier is:

- 1.65 for a 90% CI;
- 1.96 for a 95% CI;
- 2.58 for a 99% CI.

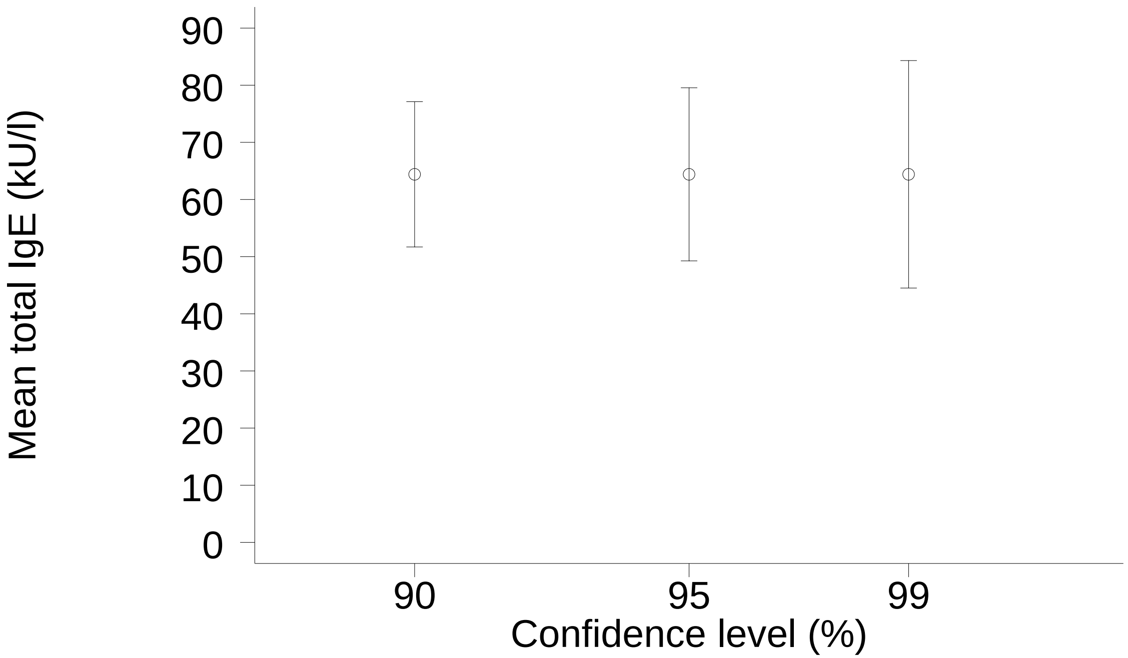
So, a 99% CI (intended to catch the *population* mean in 99% of samples) extends from

$$\textbf{sample mean} - \mathbf{2.58} \times \textbf{sample SE}$$

to

$$\textbf{sample mean} + \mathbf{2.58} \times \textbf{sample SE}.$$

The price of extra confidence: 90%, 95% and 99% CIs for the population mean total IgE (kU/l)



The CIs are all calculated from the same sample of 121, with a sample mean of 64.41 kU/l and a sample SE of 7.73 kU/l. However, the higher the confidence level, the greater is the width of the CI (in “SE-units”).

Summary (1): What is a confidence interval?

- A **95% confidence interval (CI)** around a sample mean is a range of values, in which we are 95% confident that the *population* mean lies.
- The confidence interval is usually centred on the *sample* mean, and is bounded by limits wide enough to catch the *population* mean in 95% of samples.
- For "mean", you can read "median", "proportion", "relative risk", or any other **sample statistic** used to estimate a **population parameter**.
- (And for "95%", you may read "90%", "99%", etc.)

Summary (2): What decides the width of a confidence interval?

- The **confidence level**. (Usually 95%, but sometimes 99% (wider) or 90% (narrower).)
- The **variability** of the data. (The more variable the data, the wider the CI.)
- The **number** of individuals in the sample. (The larger the sample, the *narrower* the CI.)

For people who like formulae, *some* CIs extend from a lower limit

$$\textbf{sample statistic} - \textbf{multiplier} \times \textbf{sample SE}$$

to an upper limit

$$\textbf{sample statistic} + \textbf{multiplier} \times \textbf{sample SE}.$$