Lecture 5: Estimation and Statistical Inference

Thomas Chadefaux

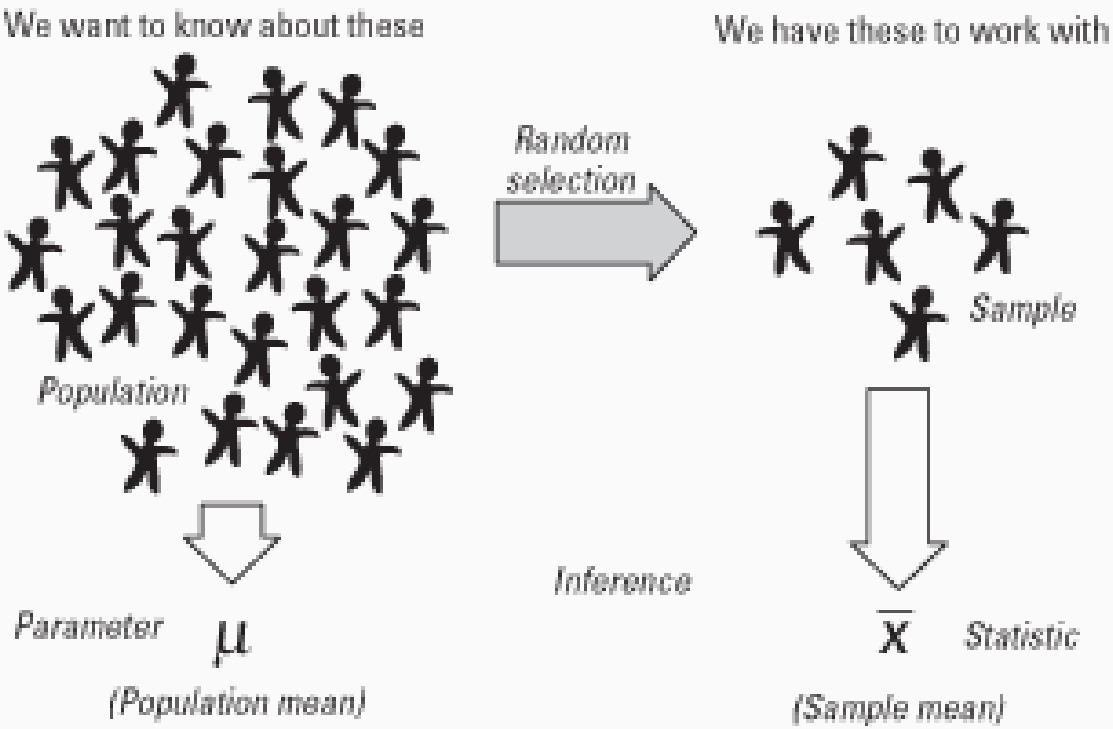
PO7001

Outline of weeks 5 & 6

- What is statistical inference?
- Confidence intervals
- tests of significance
- Bootstrapping

Inference

What is Inference?



What is Inference?

- Purpose of inference: Draw conclusions from the data.
- Using our sample, we want to draw conclusions about the general population
- This involves making probability calculations to take into account chance.

An example

- Test whether a drug is effective
- Sample of 40 patients.
  - 20 given the drug
  - 20 given the placebo
- 60% of the treatment group get better (i.e., 12 of them), whereas only 8 of the placebo group get better.

An example

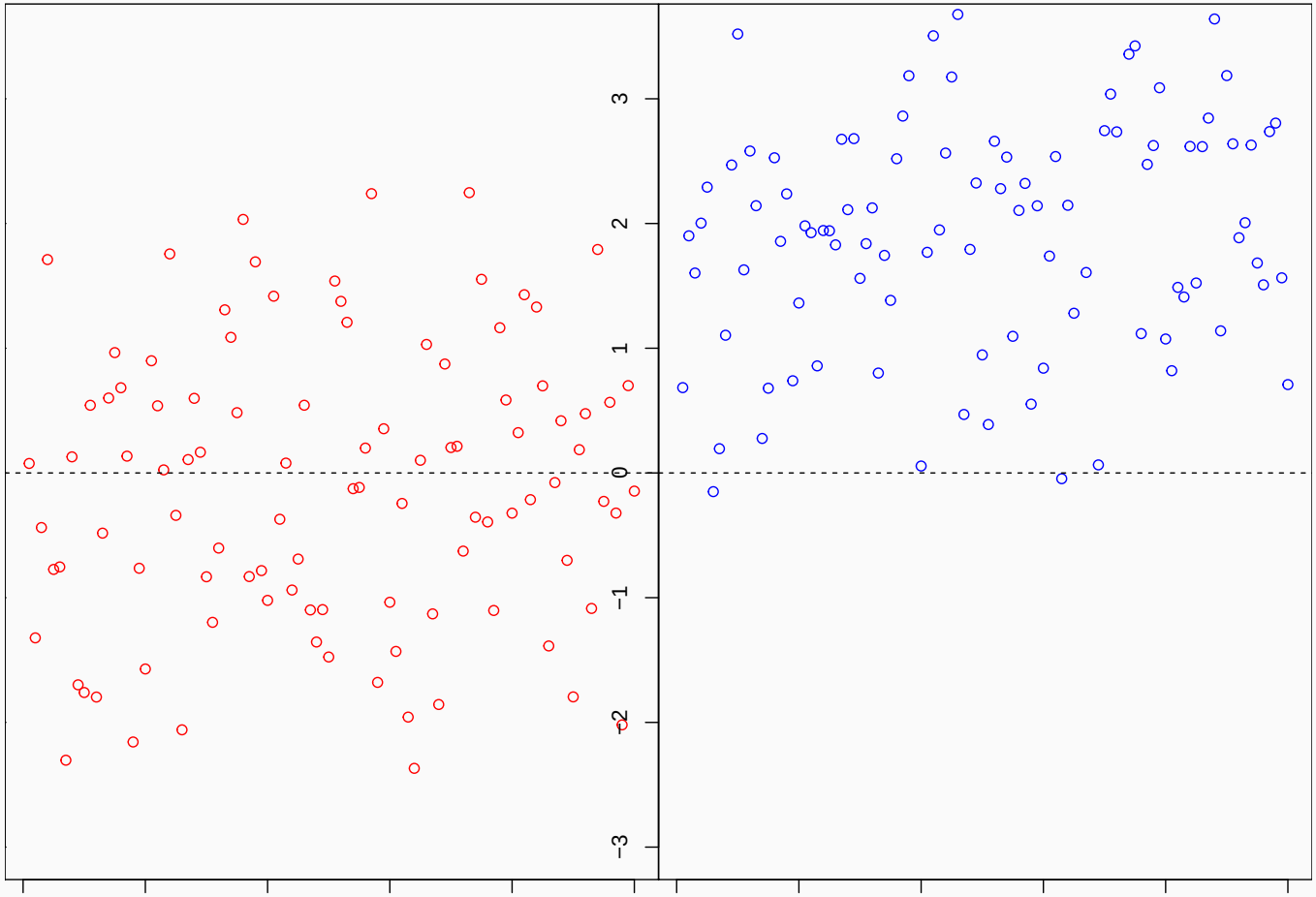
- Can we conclude that the drug is effective?
- In fact, these results would occur by chance one out of every five times by sheer chance. So we conclude that the observed difference is too likely to be due to chance, and hence fail to reject the hypothesis that the drug is ineffective.

Statistical Inference

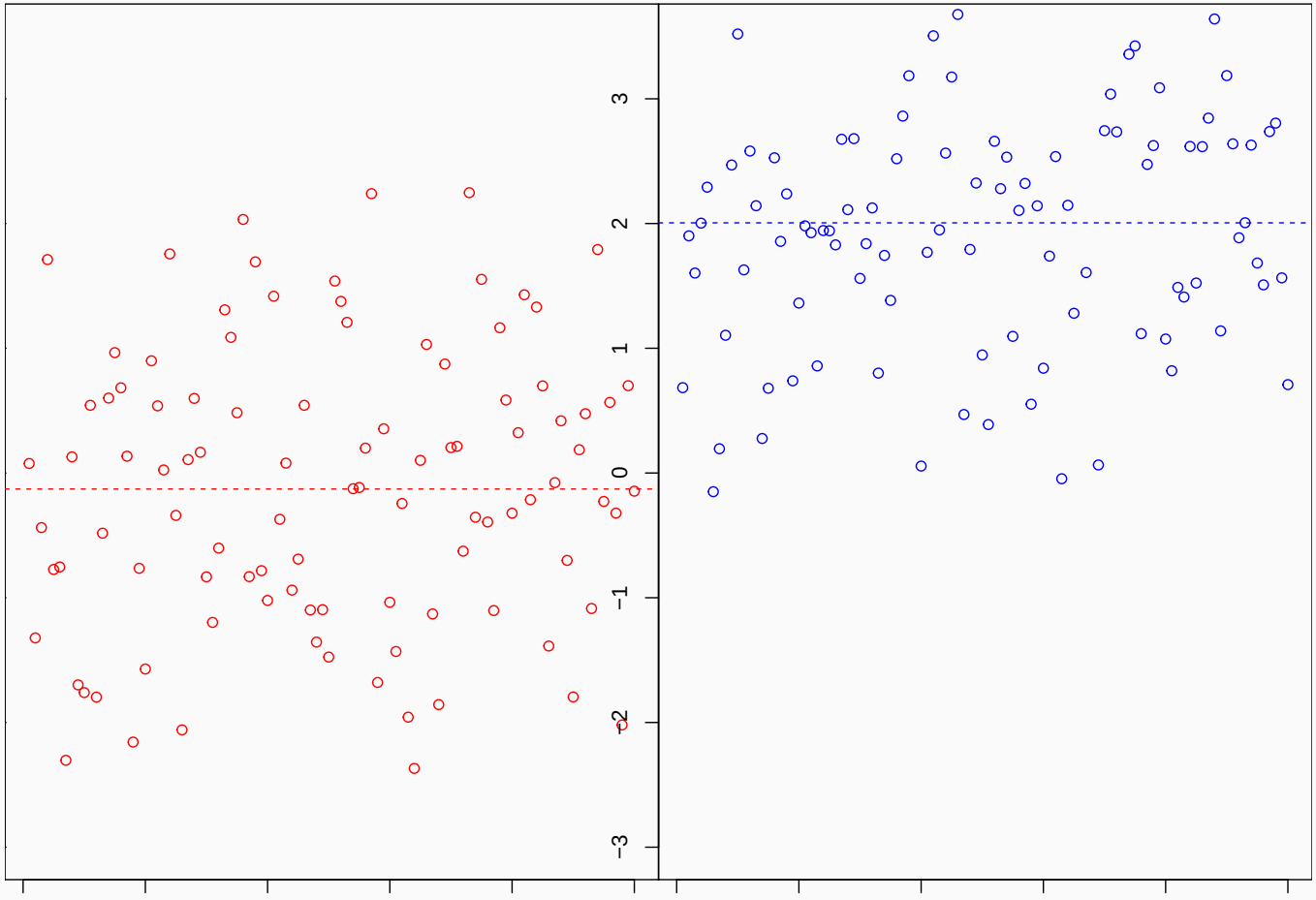
- Two main types:
  - Confidence Intervals: estimate the value of a parameter
  - Tests of significance: assess the evidence for a claim
- Both rely on assessing what would happen if we sampled many times. In particular, would we be fairly likely to obtain this result by chance?

The big picture. Pay attention, this part is key to the rest of quants 1 and quants 2!!

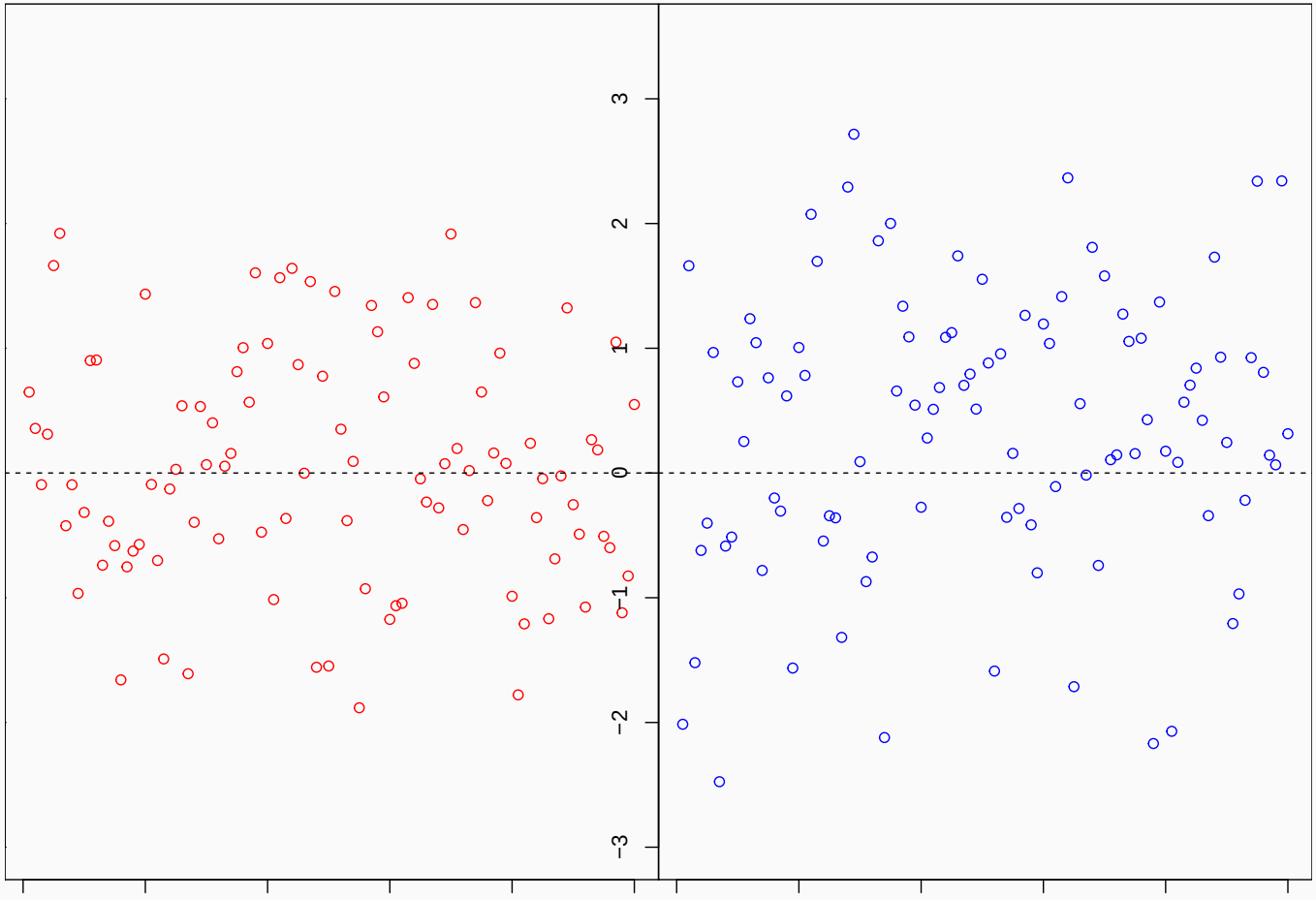
The big picture: What is big enough?



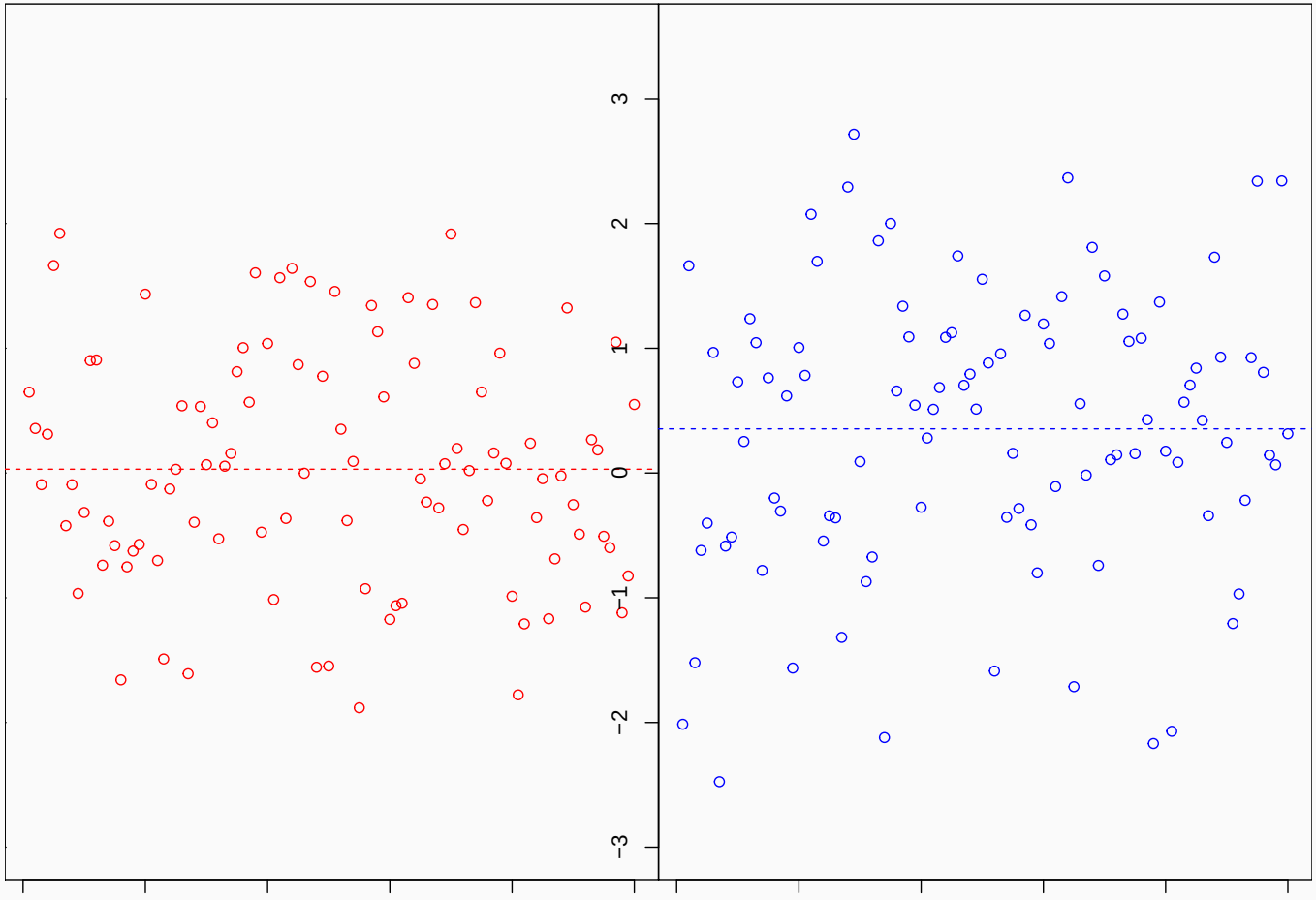
The big picture: What is big enough?



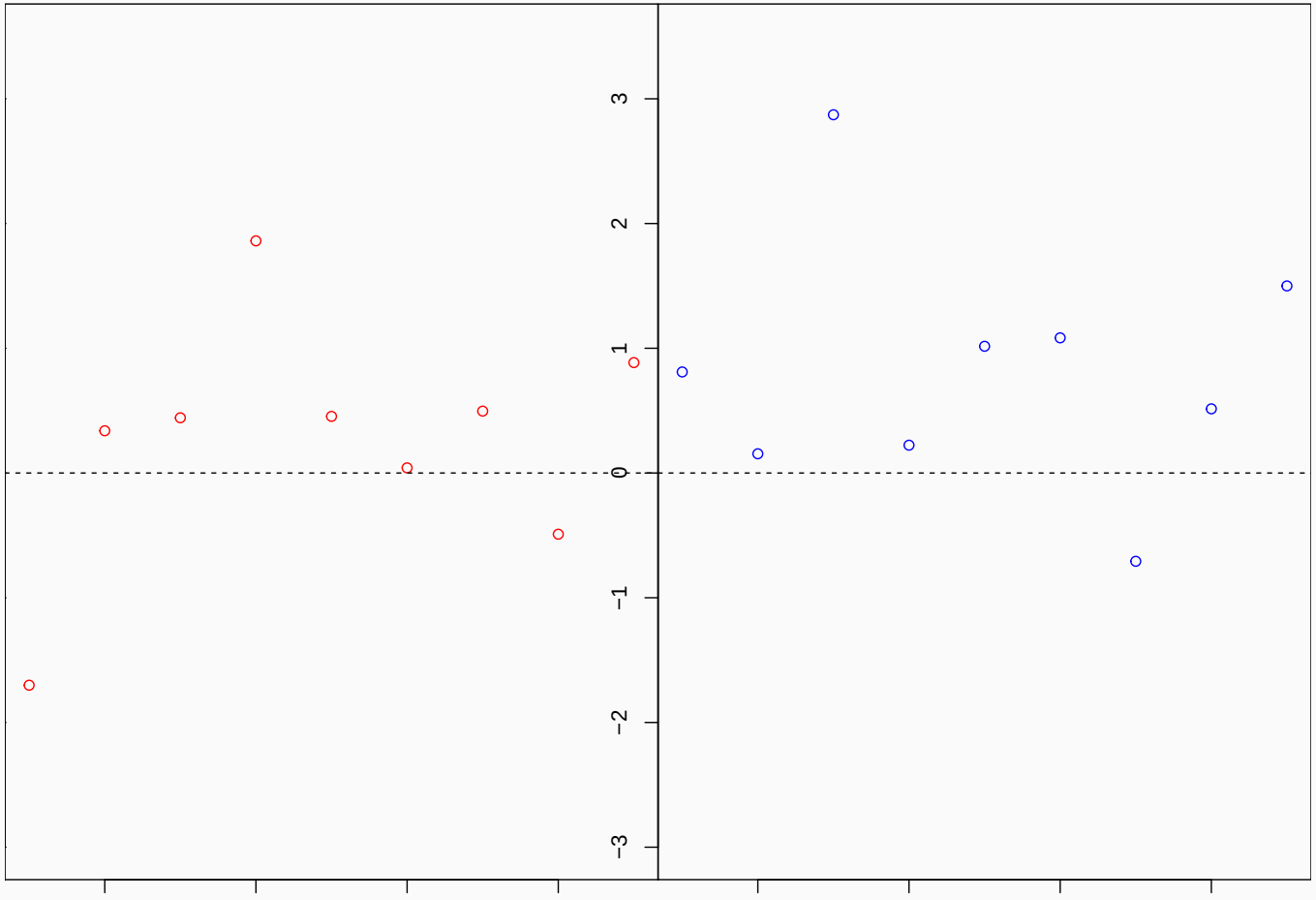
The big picture: What is big enough?



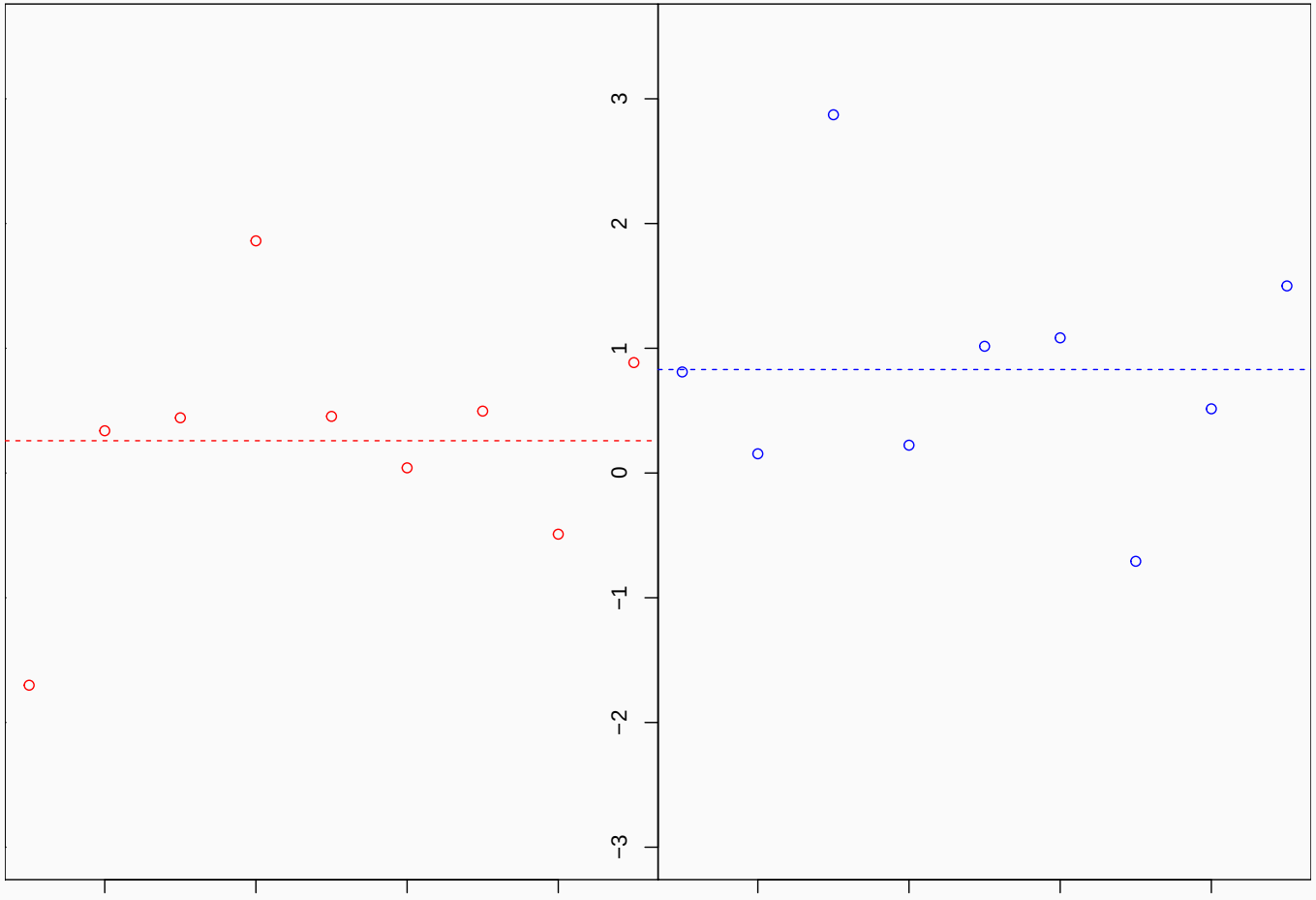
The big picture: What is big enough?



The big picture: What is big enough?

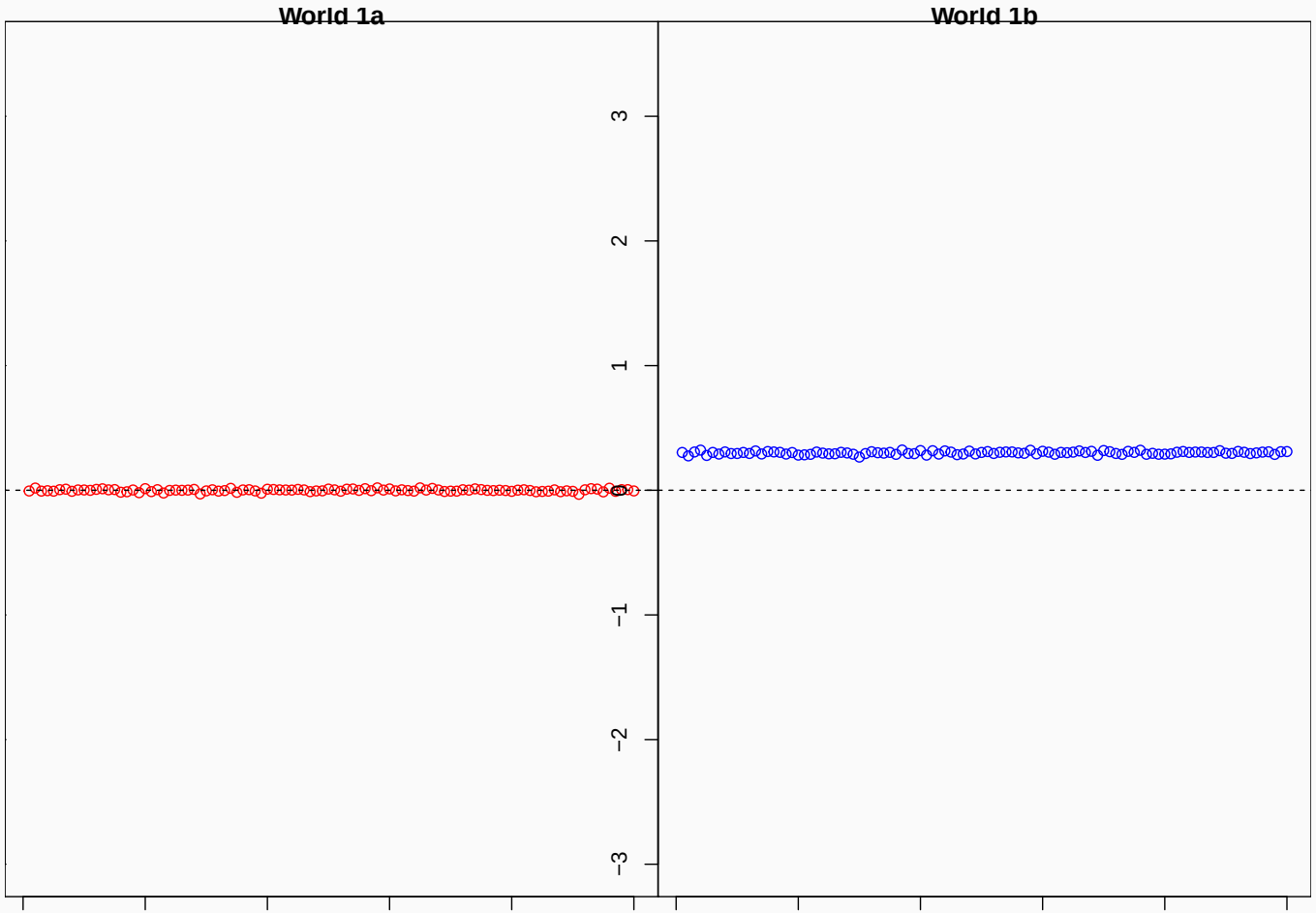


The big picture: What is big enough?



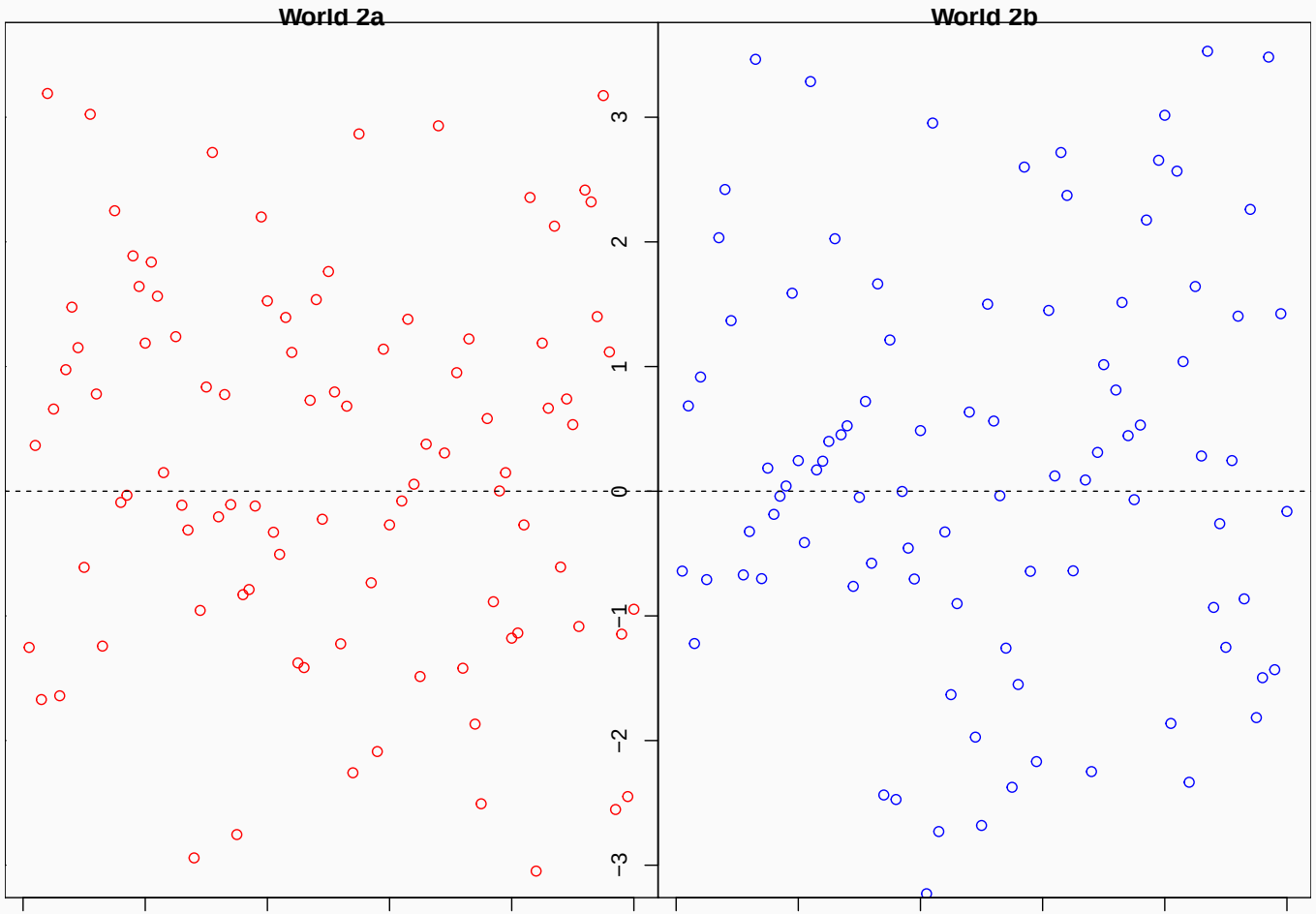
The big picture: Zooming in

Intuitively, these two samples are clearly different:



The big picture: What is big enough?

. . . but these are maybe not. Yet in the first example, the average difference was only 0.3. Here the difference is 0.6



The big picture: What is big enough?

- A bigger difference in world 2...
- ... and yet we are more confident of a difference in case 1

The big picture

- When we observe a sample, we'd like to know if that sample (e.g., its mean) is significantly different from a hypothesized number. E.g., is the mean IQ of TCD/UCD students higher than 100?
- So we collect IQs of a sample of UCD/TCD students and calculate their mean, obtain a number, say 102. But is that "significantly different" from 100?
- After all, if I took another sample of students, I would get another mean. Maybe 103, maybe 98, maybe 100, etc.
- So 102 may, or may not be surprisingly large

The big picture

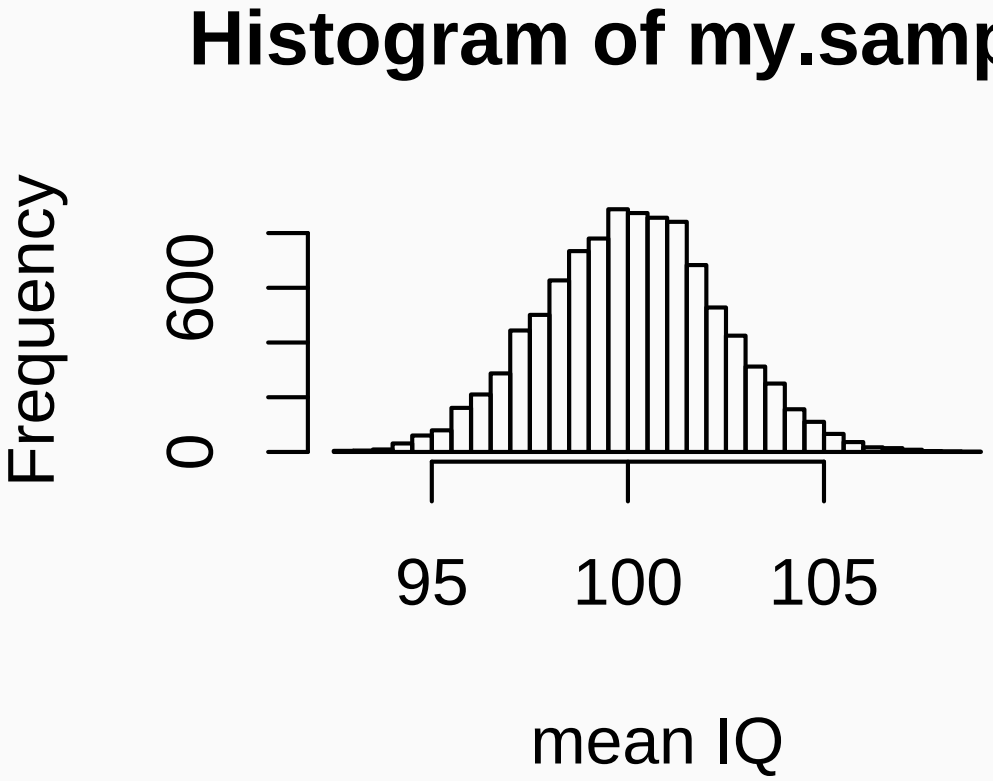
- To be able to say "significantly different", we need to be able to say with confidence that 102 is unlikely to have come up "by chance".
- I.e., we want to say that "102" would be unlikely to have been observed if the true population mean was 100.
- So the question we need to ask is: how likely would we have been to observe this sample if the truth is that our sample was actually drawn from a population (of TCD/UCD students) with mean 100.

The big picture

- To do that, we would ideally like to compare our sample to many samples drawn from a population with mean 100.
- Suppose we could do just that: draw many samples from a population with mean 100, and calculate their means. These means would have a certain distribution. We could then compare our particular sample to that distribution and say: “It looks quite different” or “it looks quite the same”.

Sampling distribution

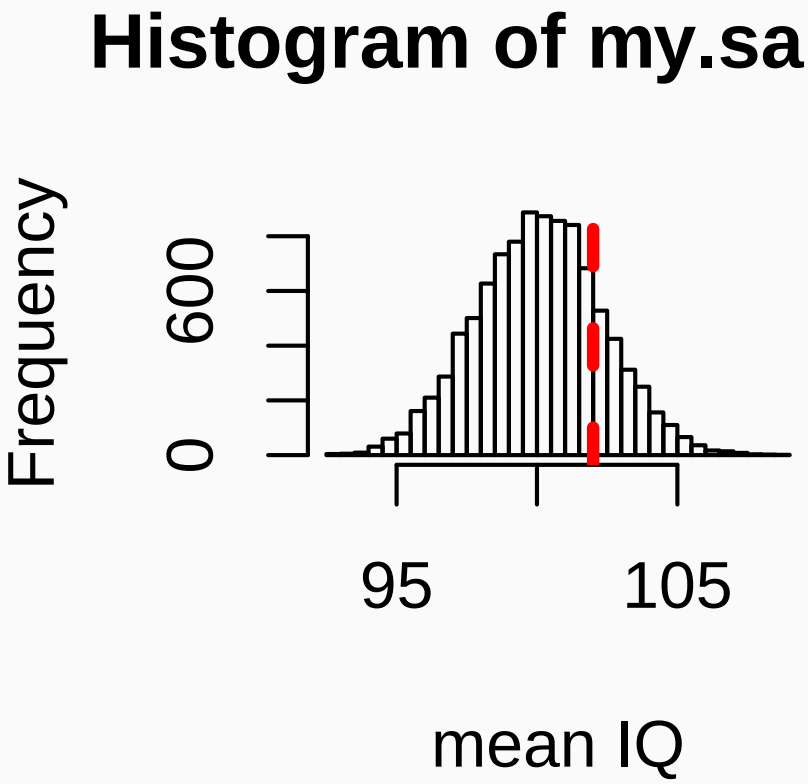
So let's do that: take random samples from a population with mean 100 and calculate their means.



Note that this is NOT a distribution of IQs. It is the distribution of MEAN IQs, i.e., what we call the **sampling distribution**.

Sampling distribution

Now we can see whether 102 was a surprisingly large mean:



Sampling distribution

Visually, we can already tell that 102 is not very large. Why? Because if we took a large number random samples from a population with mean IQ $= 100$, many of these samples would have a mean greater or equal to 102.

In other words, observing 102 is not surprising and does not lead us to conclude that TCD/UCD students's IQ must come from "another" population (i.e., one with a higher IQ)

Based on this, it is in fact perfectly "plausible" that UCD/TCD students have IQs in line with the overall population.

Sampling distribution

A little more formally, to know whether 102.5 is a lot or not, we'd like to say that it's greater than, say, 95% of the other sample means. So let's see, based on our earlier sample, whether it is.

```
length(my.samples[my.samples >= 104]) / length(my.samples[
```

```
## [1] 0.04307917
```

In short

Because inference relies on samples, they will be based on the *sampling distribution* of the data. E.g., if we sampled over and over, how would, say, the means of those samples be distributed? So we need to know about the *sampling distribution* of the sample means.

The Oh So Sad Truth: we do not know the sampling distribution!

But not so fast: The Oh So Sad Truth

- The Oh So Sad news is that no one knows the true sampling distribution. I.e., we do not have the ability to draw an infinity of samples to compare ours to.

But not so fast: The Oh So Sad Truth

- The Oh So Sad news is that no one knows the true sampling distribution. I.e., we do not have the ability to draw an infinity of samples to compare ours to.
- The good news is that there is a way to infer what it is. What we need to reconstruct the curve is a. its mean and b. its standard deviation (what we'll call the Standard error).

But not so fast: The Oh So Sad Truth

- The Oh So Sad news is that no one knows the true sampling distribution. I.e., we do not have the ability to draw an infinity of samples to compare ours to.
- The good news is that there is a way to infer what it is. What we need to reconstruct the curve is a. its mean and b. its standard deviation (what we'll call the Standard error).
- Why is that all we need? Because we assumed a normal distribution.

But not so fast: The Oh So Sad Truth

- The Oh So Sad news is that no one knows the true sampling distribution. I.e., we do not have the ability to draw an infinity of samples to compare ours to.
- The good news is that there is a way to infer what it is. What we need to reconstruct the curve is a. its mean and b. its standard deviation (what we'll call the Standard error).
- Why is that all we need? Because we assumed a normal distribution.
- Wait, what? Why did we assume a normal distribution? This is outrageous!

But not so fast: The Oh So Sad Truth

- The Oh So Sad news is that no one knows the true sampling distribution. I.e., we do not have the ability to draw an infinity of samples to compare ours to.
- The good news is that there is a way to infer what it is. What we need to reconstruct the curve is a. its mean and b. its standard deviation (what we'll call the Standard error).
- Why is that all we need? Because we assumed a normal distribution.
- Wait, what? Why did we assume a normal distribution? This is outrageous!
- We assumed a normal distribution because we are dealing with the distribution of means. And sample means are normally distributed!

But not so fast: The Oh So Sad Truth

- The Oh So Sad news is that no one knows the true sampling distribution. I.e., we do not have the ability to draw an infinity of samples to compare ours to.
- The good news is that there is a way to infer what it is. What we need to reconstruct the curve is a. its mean and b. its standard deviation (what we'll call the Standard error).
- Why is that all we need? Because we assumed a normal distribution.
- Wait, what? Why did we assume a normal distribution? This is outrageous!
- We assumed a normal distribution because we are dealing with the distribution of means. And sample means are normally distributed!
- Why? Because of the central limit theorem.

The sampling distribution of the sample mean

So we need estimates of the mean and variance of the sampling distribution. First let's try to find the sample mean—it's easy!

- The sample mean $\bar{x}$ from a sample or experiment is an estimate of the mean μ of the underlying population (remember that greek is for population, latin for sample).
- Suppose that we collect many samples and calculate each sample's mean. We now have a set of means $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_N$.
- What is the mean of those means? Note that the expected value of $\bar{x}_i$ is simply μ. So the mean of those means is simple

$$\frac{1}{N}(\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_N)$$

$$= \frac{1}{N}(\mu + \mu + \ldots + \mu) = \mu$$

The variance of the means

- A little more complicated, however, is the variance. I.e., what is

$$Var(\bar{X})$$

- First note that $Var(aX) = a^2 Var(X)$. Why? Prove it!
- Let $X_1, X_2, \ldots, X_n$ be observations from a population with standard deviation σ. Then the variance of their sum T is $n\sigma^2$
- The variance of T/n (i.e., the variance of the sample mean $\overline{x}$) must be $\frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$
- And then obviously the standard deviation of T/n is $\frac{\sigma}{\sqrt{n}}$

*The **standard deviation** of the sample means is called the standard error and is defined as*

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

(Proof that $Var(aX) = a^2 Var(X)$)

$$Var(aX) = \frac{1}{N}\sum_i (aX - a\bar{X})^2$$

$$= \frac{1}{N}\sum_i (a(X - \bar{X}))^2$$

$$= \frac{1}{N}\sum_i a^2(X - \bar{X})^2$$

$$= Na^2\frac{1}{N}\sum_i (X - \bar{X})^2$$

$$= a^2\sum_i (X - \bar{X})^2 = a^2 Var(X)$$

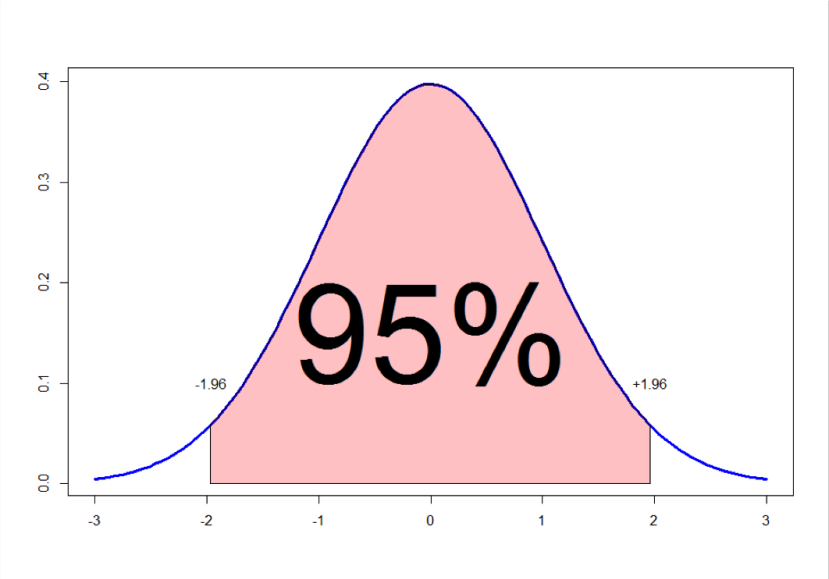
Statistical confidence

- Suppose that we observe a sample of 16 students with mean IQ 105 and standard deviation 10
- From this, we can *estimate* the standard deviation *of the sampling distribution* (i.e., the standard error) as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{10}{4} = 2.5$$

- So we're almost there
- Note that we know that for a normal distribution, 95% of the data is within $\pm 2\sigma$ of the mean (more precisely, $\pm 1.96\sigma$). Why?

Reminder: The area under the normal curve



- So 95% of all samples will be between $\mu - 2\sigma_{\bar{x}}$ and $\mu - 2\sigma_{\bar{x}}$,
- So we can say that we are 95% confident that the unknown *population* mean is between $\mu - 2\sigma_{\bar{x}}$ and $\mu + 2\sigma_{\bar{x}}$ i.e., between $105 - 2.5 = 102.5$ and $105 + 2.5 = 107.5$.

Confidence Intervals

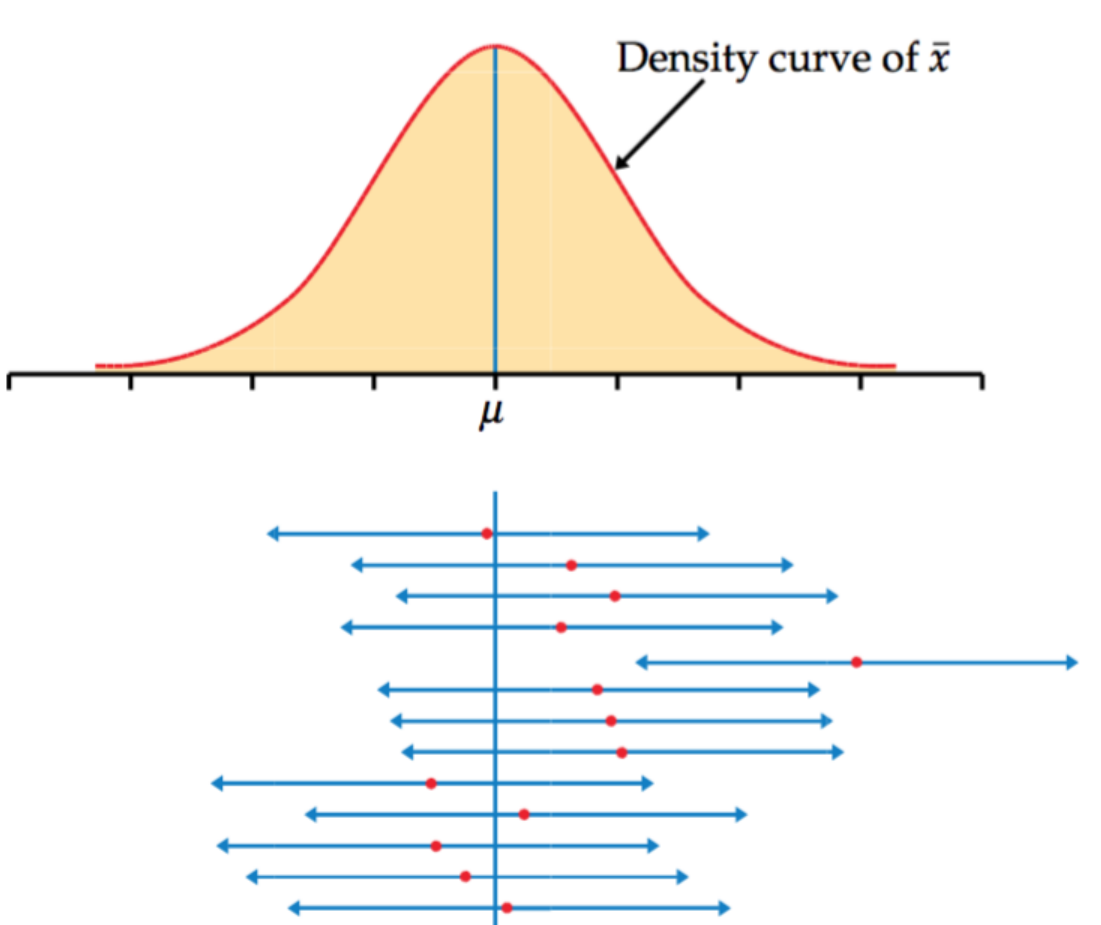
What we just calculated is a 95% confidence interval for the mean μ. Like most CIs, it has the form:

$$\text{estimate} \pm \text{margin of error}$$

The estimate, in this case, was the mean. The margin of error reflects how accurate we think our estimate is.

Interpretation

The 95% confidence interval tells us that if we sampled over and over, and every time calculated the 95% confidence interval of the mean based on that sample, then 95% of these intervals would contain the true value of μ



Calculating the CI

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

- For example: Suppose the sample mean is 20 and the standard deviation is 10. We have 900 observations. Then the confidence interval of the mean is:

$$20 \pm 1.96 \times \frac{10}{30} \approx [19.33, 20.66]$$

Calculating the CI

- Note that the size of the confidence interval depends critically on:
  - the significance level of the test (here 5%), and, correspondingly, the critical value of the test (1.96 here)
  - the sample size n. To see this, consider the example above, but this time with only 25 observations:

$$20 \pm 1.96 \times \frac{10}{5} \approx [16, 24]$$

I.e., a much larger interval