From Samples to Populations

Research Methods for Political Science

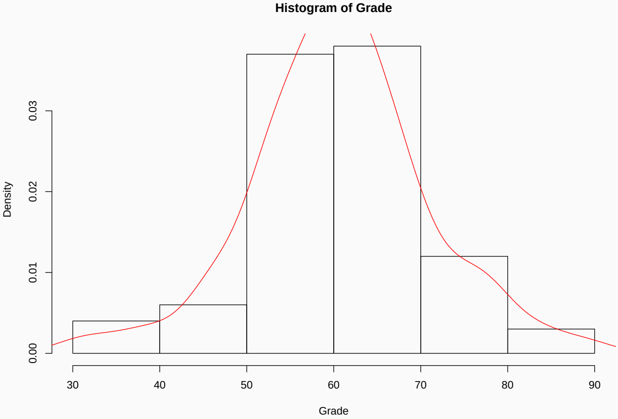
Thomas Chadefaux

Trinity College Dublin

Motivation

The shape of a sample distribution

What we learned last time: the shape of a distribution. E.g., we have a sample of 100 students' grades:



The shape of a sample distribution

- But what we really care about is the general population. E.g., is this distribution of grades for the 100 students representative of all students' grades?
- Probably not, but the two means would probably be similar. But *how* similar?

Estimates and Inferences

The numbers (mean, median, etc) we use to describe our sample are **estimates** of the numbers that describe the population. We call these numbers *statistics*.

The true numbers of the population are called *parameters*.

$\rightarrow$ a *statistic* is used to estimate a *parameter*. - E.g., we rely on the sample mean (the mean of the sample) to estimate the population mean.

Aside: notation

- When we talk about statistics, we use roman letters. E.g.:
  - the sample mean is $\overline{x}$
  - the sample standard deviation is s
  - the sample variance is s^2
  - etc.
- When we talk about population parameters, we use greek letters. E.g.
  - the population mean is μ
  - the population standard deviation is σ
  - the population variance is σ^2
  - etc.

Our first inference

Meet the Drumchiosaurus

This is Nick

Meet the Drumchiosaurus

You have just found the perfectly preserved remains of a Drumchiosaurus. No one has ever seen one before. The Drumchiosaurus is yellow and 15cm long.

Congratulations, you have a sample!

It's a sample of size 1 (we'll often say or write N=1)

Inferences about the Drumchiosaurus

You have just found the perfectly preserved remains of a Drumchiosaurus. No one has ever seen one before. The Drumchiosaurus is yellow and 15cm long.

But you probably don't care about only Nick. What is your best guess as to the average length of the Drumchiosaurus population (the Drumchio. . . sauruses, . . . sauri, . . . sauris?)

- 10cm?
- 15cm?
- 20cm?

Inferences about the Drumchiosaurus

What is your best guess of the average length of the Drumchiosaurus population (the Drumchiosaur. . . uses, Drumchiosauri, Drumchiosauris?)

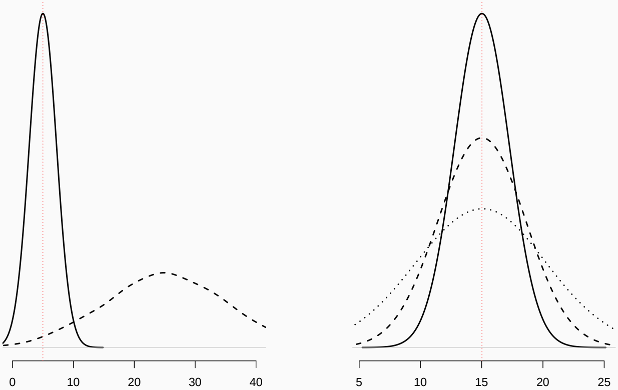
- 10cm?
- 15cm?
- 20cm?

15cm is probably your best guess, since all you have is that one specimen.

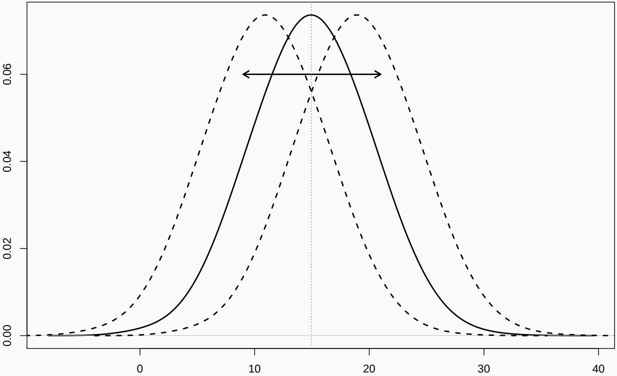
Inferences about the Drumchiosaurus

You'd also like to know about the *distribution* of Drumchiosauri's lengths.

Which of these distributions do you think our friend Nick is most likely to have come from?



Inferences about the Drumchiosaurus



The Drumchiosaurus party

Our Drumchiosaurus is now joined by its buddies

The Drumchiosaurus party

The Drumchiosauri measure 14, 15, 16.1 and 12.8cm.

So we can now make a new estimate of the mean, *and* of the variability in the population, based on our sample (N=4).

We calculate the standard deviation *s* as

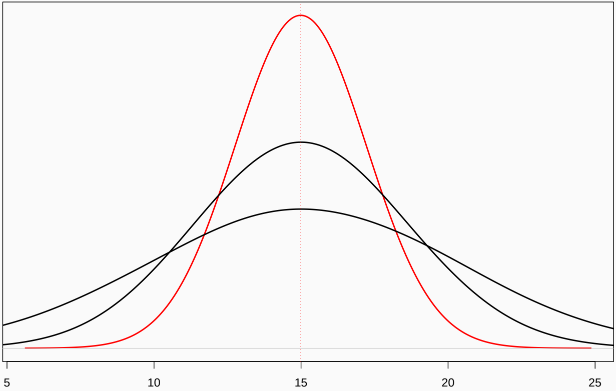
```r
sd(c(14,15,16.1,12.8))
```

```
## [1] 1.408013
```

So it looks like there is not too much variation in Drumchiosauri's height

The Drumchiosaurus party

The red curve was probably the best



The Drumchiosaurus party

In short:

- We can make inferences with very little info
- Our inferences adjust to new info
- With more info we can refine our inferences and rule out more and more of them
- We can never be certain about the true population parameter. But we can say e.g. that the mean has 90% probability of being between 13 and 17.

Sampling

Sampling

We got a little taste for sampling with our Drumchiosaurus party.

But suppose that once Nick and his buddies leave, another group of Drumchiosauri show up.

Their mean length is unlikely to be **exactly** the same as the initial one.

Sampling: example

- Drumchiosauri group 1: 14, 15, 16.1, 12.8
    - mean: 14.475
- Drumchiosauri 2: 18, 14.9, 12.3, 15.2
    - mean: 15.1
- Drumchiosauri 3: 13, 13.5, 16, 14.2
    - mean: 14.175
- Drumchiosauri 4: 16, 16.2, 15.9, 15
    - mean: 15.775
- etc.

Sampling

What we get then are multiple values of sample means: 14.475, 15.1, 14.175, 15.775, etc.

Suppose we collected a lot of samples, each a bit different. Then we would have a sample of samples, and a sample of sample means.

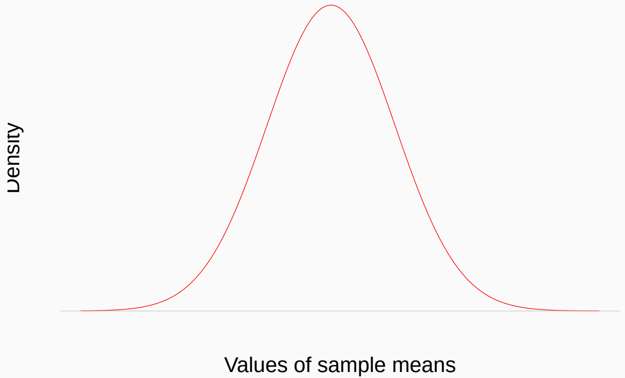
That sample of samples would have a mean, a standard deviation, etc., just like our sample of observations.

The sampling distribution

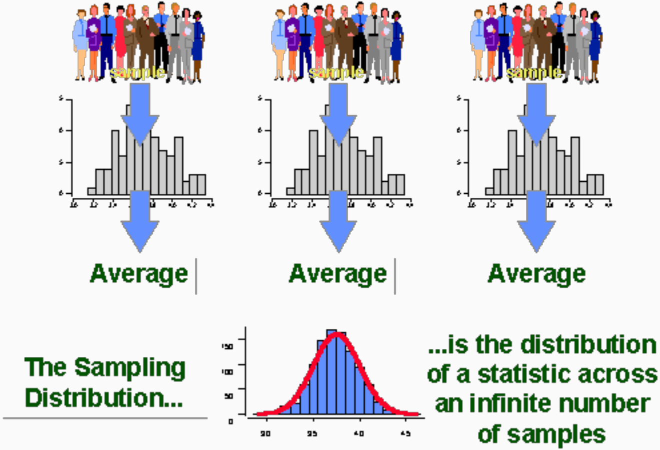
We call the distribution of these sample means the *sampling distribution of the sample mean*

A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population

The sampling distribution



The sampling distribution



The Sampling Distribution...

...is the distribution of a statistic across an infinite number of samples

The sampling distribution is normal (important)

Notice that the distribution of sample means is itself normal. This is not an accident.

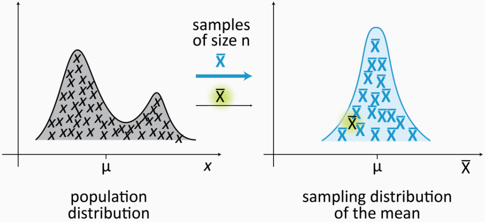
Important: Whether or not the distribution of the sample (or of the population) is normal, the distribution of the sample means (i.e. the sampling distribution) will be normally distributed

This fact is called the Central Limit Theorem

The sampling distribution is normal (important)

The central limit theorem is one of the reasons why the normal distribution is so important.

- For a large number of samples, the distribution *of the sample means* will be *normally* distributed, *regardless of the shape of the original distribution*. NB there are exceptions (e.g., highly skewed distribution, etc.)



The central limit theorem (Important)

http://onlinestatbook.com/stat_sim/sampling_dist/

Standard error

The sampling distribution has a mean and a standard deviation.

IMPORTANT: the standard deviation of the sampling distribution is called the standard error

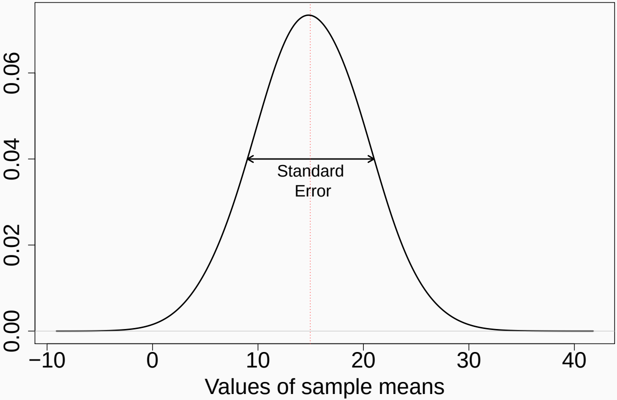
The standard error is not the standard deviation of the sample (or of the population)!

IMPORTANT: the standard error is NOT the same as the standard deviation of the sample. It is the standard deviation of the **sample means**, NOT of the sample.

Standard error

The standard error will always be smaller than the standard deviation. Why?

Standard error



Standard error

Ok, but... why do we care?

The standard error is going to be the foundation of almost EVERYTHING we do from now on.

You MUST understand the concept of the standard error.

Standard error (heads-up)

Ok, but... why do we care?

Heads-up: When we get a sample, and hence a sample mean, we'll want to know whether that sample mean is "abnormally" high.

- e.g., Are Trinity students smarter than the average population?
- To answer the question, we need to
  1. collect a sample of Trinity students
  2. calculate their mean IQ. Say we get 101.5
  3. determine whether 101.5 is "big enough" to conclude that yes, TCD students are smarter than the average population (whi has IQ 100)

Standard error (heads-up)

Step 3 is key. To determine whether it is “big enough”, we’ll need to know whether that mean IQ could or could not have come from the general population.

To know that, we’ll want to imagine that we took many samples from the general population, and determine how often we’d get an average of 101.5. If the answer is “often”, then we’ll conclude that 101.5 is not a particularly large sample average, and hence that TCD students are not necessarily smarter.

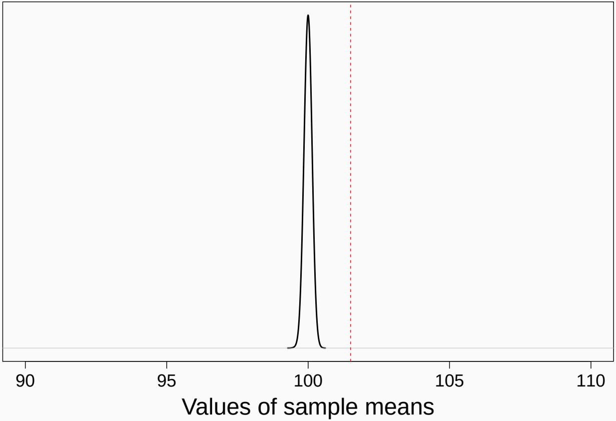
Standard error (heads-up)

Great but what does it have to do with the standard error?

The standard error is the standard deviation of sample means. I.e., it tells us how frequently we would get a sample that deviates from the hypothesized mean (here 100) by quite a bit.

Standard error (heads-up)

Suppose the standard error is very small. Then the distribution of sample means would look like that.

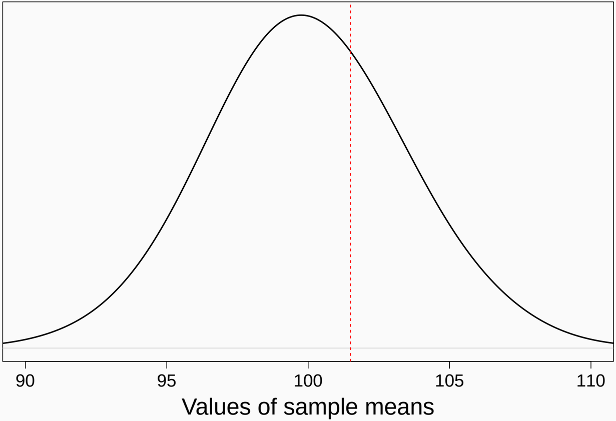


Standard error (heads-up)

In that case, it is *unlikely* that such a sample could have come from the general population, and hence we conclude that TCD students ARE smarter than the overall population

Standard error (heads-up)

If, however, the standard error is very large, then the distribution of sample means would look like that.



Standard error (heads-up)

In that case, it is *likely* that such a sample could have come from the general population, and hence we cannot conclude that TCD students are smarter than the overall population

Standard error (heads-up)

This is why the standard error is SO important!

Standard error

In practice, we only take *one* sample, but we can think of that sample as belonging to a distribution of *possible* samples. The distribution of their means will be normal

Standard error

Great. Now how do we get the standard error?

After all, we only have *one* sample, not all possible samples. Hence we have *one* mean, not the distribution of all possible sample means.

Estimating the standard error.

Estimating the standard error.

The standard error of the mean can be calculated with only 2 variables:

1. the standard deviation of the sample (s). NB: that's the SD *of the sample*
2. the size of the sample

Estimating the standard error: intuition

1. why do we need the standard deviation of the sample?

Suppose that we have a sample with a **large** standard deviation. That tells us something about the population this sample came from, namely that it probably has a large standard deviation itself. Hence, every time you take a sample, you are likely to get a pretty different mean.

A sample with a **small** SD will instead have means that are more or less always the same. So then it makes sense that the *standard error* (reminder: it is the standard deviation *of the sample means*) is a function of the standard deviation *of the sample*.

Estimating the standard error: intuition

2. why do we need the size of the sample?

Suppose you have a small sample size. Say, n=1. Then by chance you are likely to get means that are all over the place.

If instead you have a large sample size, you will get some large values, some small, but these will average to something close to the population mean. So whenever you sample (with a large n), you will get a mean that is more or less the same each time. I.e., a small standard error!

Estimating the standard error: formula

Now we are ready to estimate the standard error, at last:

$$SE = \frac{s}{\sqrt{n}}$$

(reminder: s is the sample standard deviation.)

(If you must know: OPTIONAL)

For those frustrated by formulae coming out of nowhere, here is how we got the formula above:

1. Suppose we take a sample of independent observations from a population with mean μ and variance σ^2.
2. Calculate their total: $T = x_1 + x_2 + \ldots + x_n$
3. Because each of the x_i has (expected) variance σ^2 and they are independent, the sum T will have variance

$$var(x_1) + var(x_2) + \ldots + var(x_n) = \sum_i \sigma_i^2 = n\sigma^2$$

(If you must know: OPTIONAL)(cont'd)

4. Similarly, the mean of this sample will have variance:

$$
\begin{aligned}
var(\frac{T}{n}) =& \sum_i (\frac{1}{n} T - \frac{1}{n}\overline{T}) \\
=& \frac{1}{n^2} \sum_i (T - \overline{T}) \\
=& \frac{1}{n^2} var(T) \\
=& \frac{1}{n^2} n\sigma^2 \quad \text{(see step 3)} \\
=& \frac{\sigma}{\sqrt{n}}
\end{aligned}
$$

Estimating the standard error: an example

Suppose you collect a sample of 100 student IQ, with standard deviation 11. Then the standard error is:

$$SE = \frac{s}{\sqrt{n}} \tag{1}$$

$$= \frac{11}{\sqrt{100}} \tag{2}$$

$$= \frac{11}{10} = \quad 1.1 \tag{3}$$

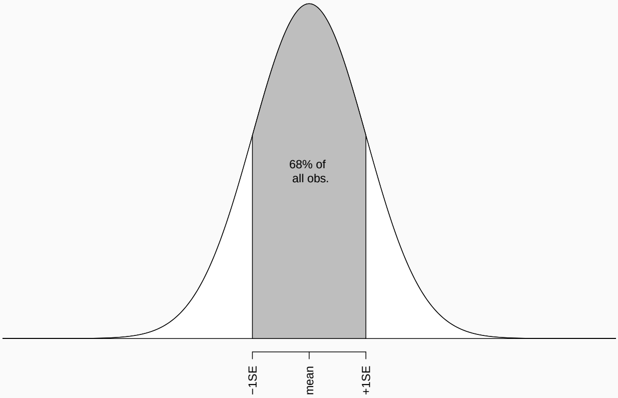
Confidence Intervals

Confidence intervals

Now that we know the standard error, we can do useful things.

Confidence intervals

In our previous class, we discussed the area under the normal curve. As a reminder:



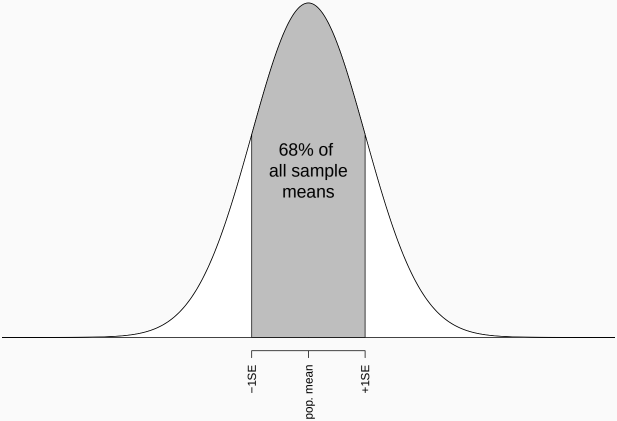
Confidence intervals

Now remember that the standard error (SE) is a standard deviation; it is the standard deviation of the sample mean. So we can use the above to say for example that the range

$$\text{Population mean} \pm 1\text{SE}$$

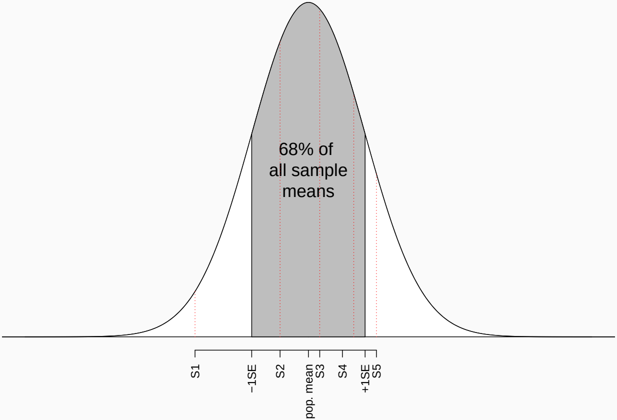
would contain 68% of all sample means.

Confidence intervals



Confidence Intervals

We can use this fact to our advantage. Suppose we collect a number of sample means:



Confidence Intervals

Now look at all the sample means that fall in the grey zone. Does the range

$$\text{sample mean in the grey zone} \pm 1SE$$

include the (true) population mean? Yes it does!

Now look at all the sample means that DO NOT fall in the grey zone. Does the range

$$\text{sample mean outside the grey zone} \pm 1SE$$

include the (true) population mean? NO it does NOT!

Confidence intervals

In fact, we know that 68% of sample means will fall within 1 SE of the population mean.

So for 68% of sample means, it will be true that the range $\text{sample mean} \pm 1SE$ contains the population mean

Confidence intervals

Of course in reality we only observe ONE sample mean. But whatever that sample mean may be, we can say that there is a 68% probability that the range $\text{sample mean} \pm 1SE$ contains the population mean

Confidence interval

Let's apply this to a practical example. Suppose you have a sample of 100 students' marks x, with $\overline{x} = 50$ and $\sigma = 15$.

Step 1: calculate the standard error of the mean:

$$SE = \frac{s}{\sqrt{n}} = \frac{15}{10} = 1.5$$

As just discussed, there is a 68% probability that the true population mean is within 1 SE of the sample mean.

So we are 68% confident that the true population mean is within

$$50 \pm 1.5$$

Confidence interval

What we have just calculated is a > Confidence interval

Here we calculated a 68% confidence interval, and that interval is [48.5 – 51.5]

Confidence intervals

Of course, that leaves a 32% chance that the true mean is NOT within these bounds. And that's probably too much of a risk.

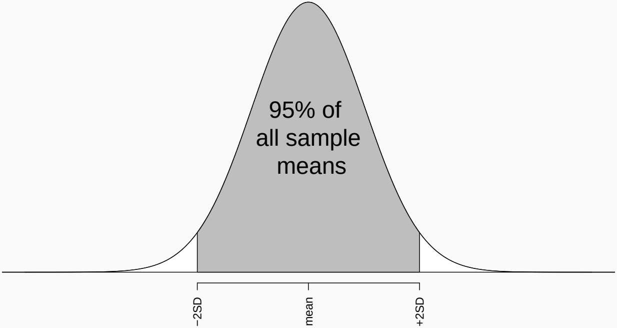
So you want MORE confidence.

Confidence intervals

To calculate a 95% confidence interval, you'd want to ask: I want a 95% probability that my sample mean ± k standard errors contains the true mean.

Confidence intervals

From last week again:



Confidence intervals

So 95% of all sample means would fall within 2SD of the true mean (technically it's 1.96... it doesn't matter much). Therefore to calculate the 95% confidence interval, I would need to calculate the range:

$$\text{sample mean} \pm 2\mathit{SE}$$

Using the example above, that would mean that my 95% confidence interval would be [47–53].

Confidence intervals

I could do the same for a 99.7% confidence interval:

$$\text{sample mean} \pm 3SE$$

which would be [45.5–54.5].

Note that the more "confidence" I get, the wider my interval. This makes sense. If I wanted 100% confidence, I could say that I am 100% confident that the true mean lies between 0 and 100. . .

Confidence interval

How can you get more confidence? Two ways:

1. increase the size of the confidence interval. But of course not ideal
2. increase the sample size. Consider our 95% confidence interval above, but imagine we had 900 students, rather than the initial 100. Now the SE becomes

$$SE = \frac{s}{\sqrt{n}} = \frac{15}{30} = 0.5$$

and therefore our 95% confidence interval is now [49–50], instead of the initial [48.5–51.5]. Much better, but of course at a cost.

Summary

Key concepts in this lecture

- Inferences: learn about a population from a sample
- Sampling $\rightarrow$ sampling distribution: a probability distribution of a statistic (here the mean)
- The sampling distribution is (almost always) normal: Central limit theorem
- The standard error of the mean
  - is the standard deviation *of the sample mean*.
  - is estimated as $\frac{s}{\sqrt{n}}$
- Confidence intervals: sample mean $\pm kSE$, where the more k, the more confidence (k=1 gives you the 68% CI, k=2 the 95% CI, k=3 the 99.7% CI, etc.)