Lies, Damn Lies and Statistics

By David Hearne, Researcher, Centre for Brexit Studies

Can statistics lie? In this blog post I want to try and carefully discuss some of the issues wrapped up in the title. Although for many mere mention of the word statistics will bring back unwanted memories of maths lessons in school, these themes can bring down governments and change lives. Indeed, as anyone who has watched The Imitation Game (and it's a film I can thoroughly recommend) will know, they can alter the course of wars.

The first issue raised by the title is the fact that there are multiple definitions of the term "statistic". To a mathematician the term is likely to conjure up thoughts of "random variables" (incidentally, a complete misnomer since they are neither random nor variables), probability distributions and stochastic processes. Most "normal people" (with apologies to mathematicians!), however, use the term rather differently.

Even those of us who are not mathematicians use the word statistics to mean subtly different things, depending on the context. A reputable dictionary gives multiple definitions, one of which is, "the science of using information discovered from studying numbers" and another, "a fact or piece of information that is shown by a number"[1].

Now, there's certainly overlap between these two things, but they are distinct. In any event, it is clear that most of us have quite a broad definition in mind. When we use the terminology, what many of us are actually referring to is the act of measurement. How many nurses are in the NHS? How much did the government borrow last year? How big is the economy? What is inflation? How profitable is Tesco? What was the unemployment rate last month?

All of these questions (and many others) share two principle challenges. How do we *define* the issue at stake and how do we *measure* it? The two are usually intertwined. Sometimes this is easy, but often it is much harder.

Take unemployment, for example. How do you define if somebody is unemployed? In the past, governments attempted to ‘hide’ or minimise the unemployment rate by changing the definition. Today, most developed countries use the International Labour Organisation’s definition, which facilitates comparisons both across countries and over time.

However, even this isn’t perfect. It requires individuals to be both available and actively looking for work. However, in reality, neither of these things are absolutes. What about individuals with caring responsibilities that restrict their availability to certain times of day? How hard do you have to look in order to be “actively looking”? What about those who have ‘given up’ or are not actively looking but would work if the right job came up?

What about those who have very occasional work? They might not be “unemployed”, but neither do they really fit most traditional notions of employment. You could be looking for work but take on odd jobs (and thus register as self-employed) in the interim. You might not count as “unemployed” if you manage to get 2 or 3 hours of work per week, but you are still looking for work.

However, even if we are happy with our definition we come to the thorny question of measurement. Typically, labour market status is determined through survey evidence. Individuals are selected and then asked a series of questions about themselves and their labour market status.

This is the arena that relates to the more narrow definition of applied statistics. How are individuals chosen? What is the sampling frame, what survey methods are used etc. Typically, the methods of statistical inference used are valid only in the case of random selection.

Yet true random selection is extremely difficult, particularly in the case of countries like the UK where there is no centralised system of identification[2]. The details of how the procedure is actually carried out are readily available in various documentation from the Office for National Statistics (ONS) and I shan’t bore readers with a detailed description.

However, no matter how good the work of a statistical agency – and the ONS does an excellent job with the resources available to it – all surveys face practical challenges and these are likely to vary over time. Aside from the question of how to select individuals in the most random way possible, there are issues of contact difficulty and refusal.

Some individuals are not contactable, others refuse to answer the survey. Then there are practical challenges around translation and identification. Amongst those who *do* agree to answer the survey there are difficulties: are all their answers truthful and have they correctly understood every question?

As a result, in spite of the huge amount of hard work that goes into collecting accurate survey data, there are inevitable issues. Responses need to be weighted to accurately reflect the UK population. If fewer young people and ethnic minorities respond than should do (to reflect the UK's demographic make-up) then those that *do* respond have a higher weight.

Moreover, it is likely that there have been changes over time in how (and whether) certain socio-demographic groups respond. Are non-responders statistically different from those that *do* respond? The upshot is that interpretation of statistics requires a great deal of care. Our best estimate of unemployment might be 5.1%, but it could be 4.8% or it could be significantly higher.

In practice, given these challenges, estimates of the overall unemployment rate in the UK show remarkable stability and accuracy most of the time (the present pandemic might be a partial exception to this fact – administrative data suggest that official figures might be underestimating it). This is a testament to those who work so hard to produce them. However, that does not change the need for extreme caution in their interpretation and use, especially over time and between different places.

Subnational comparisons or those involving specific sub-groups (e.g. many ethnic minority groups) tend to exhibit much greater uncertainty. Other things are even more difficult. Measuring prices across time and space can be fiendishly challenging. Statistics don't lie, but they can be erroneous and certainly can be misleading.

Does that mean that we shouldn’t trust any of them? No. However, context is important. Ask yourself this: exactly what is being measured and how? Often the answers are less clear than we think.

[1] https://dictionary.cambridge.org/dictionary/english/statistic

[2] Incidentally, this is one of the practical reasons that the UK was so much more lax than many other EU states in its application of conditionality to freedom of movement of labour.