DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

THE UNIVERSITY OF AUCKLAND

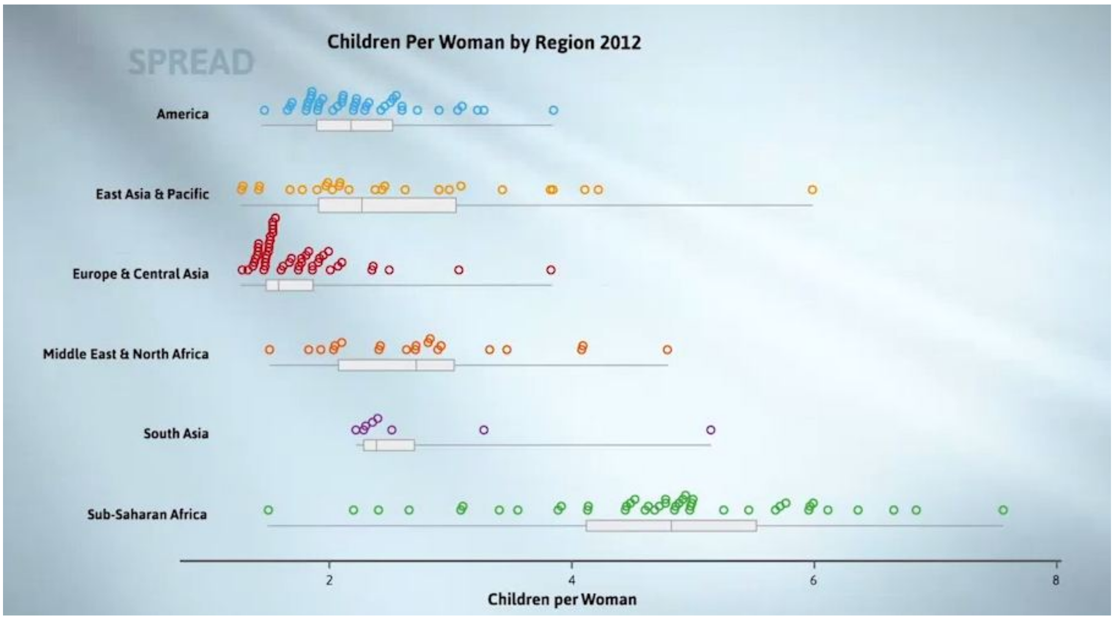
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 2

COMPARING GROUPS

Welcome back. In the last video, we were working with dot plots for a single numeric variable. I concluded by making the point that although these plots are useful for grasping how things are for a single variable, they're even more useful for getting a handle on how things change when we make comparisons. The most useful insights from data usually come from spotting an important change.

This time, we'll use country-level data Gapminder for 2012. The Gapminder variables measured demographic and socioeconomic factors like hunger, population growth, poverty, energy consumption, social factors, and education. We'll start by looking at the simple question, "What's been happening to the numbers of children women have been having around the world?" We are using the variable Children Per Woman. Technically, this is the fertility rate. Roughly, it's the average number of children a woman would have over her lifetime as estimated in a particular year.

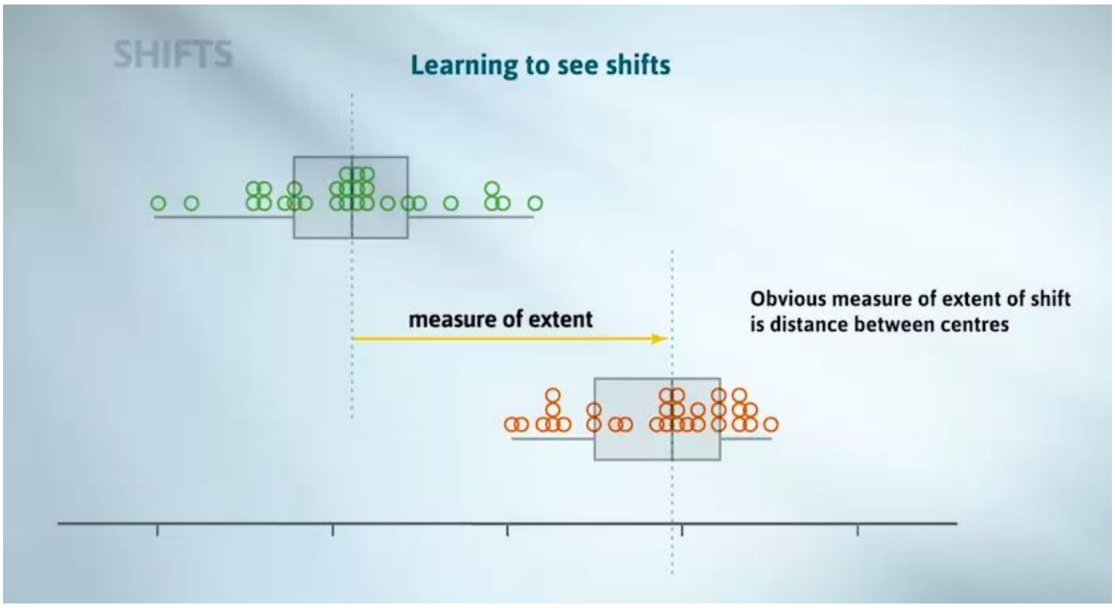


Here are the children per woman values for the world's countries in 2012. Two variables are being used, a numeric variable, Children Per Woman, and a categorical variable, Region. The plot has several sets of dots for each regional grouping, all plotted against the same scale. It helps us address the question, "How does fertility differ between regions?" What can we see?

I'm guessing that even without instruction you're already seeing quite a lot. It'd be helpful to your learning if you paused the video at this point and jotted down a few of the main features about regional differences and the number of children per woman that jump out at you.

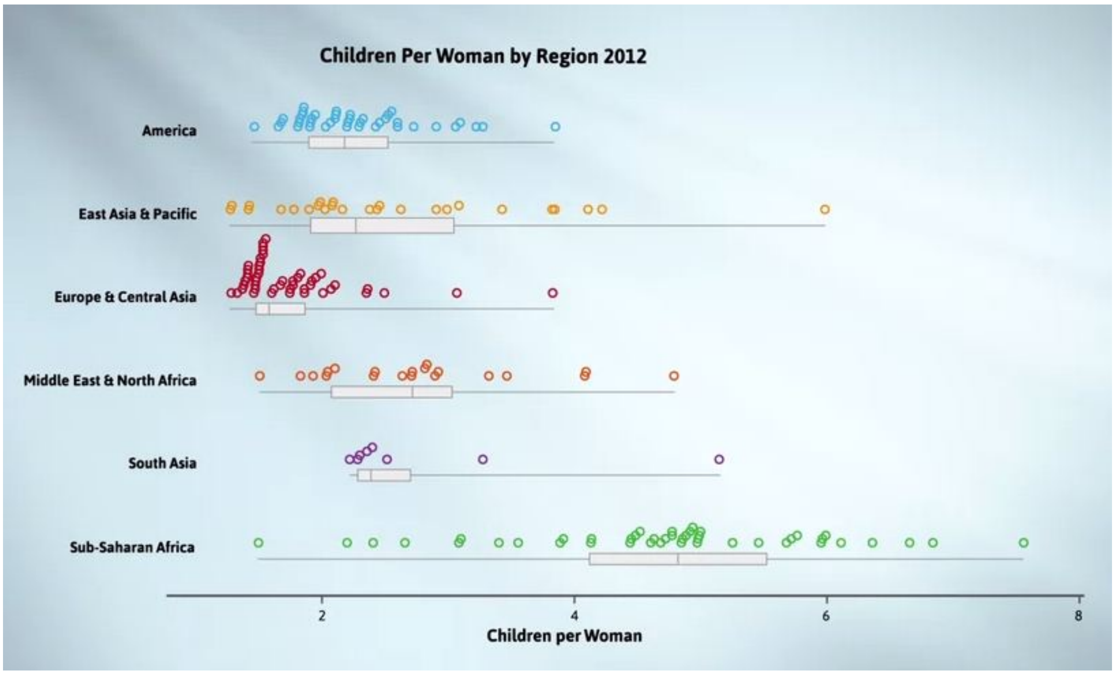
Welcome back. Well, what sort of things can we see in a plot like this one?

As usual it's all about educating your eyes. The main new things we look for are changes between groups, differences in centre, variation, or extreme differences in shape.

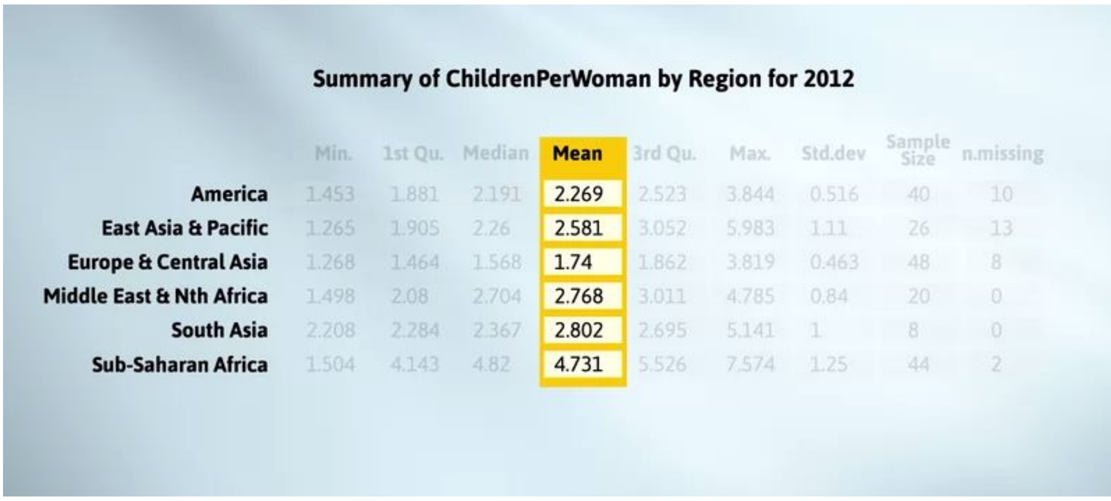


First I'll explain differences between centres or shifts. I like the language of shift because it describes well what I usually see first. It's a very visual notion.

The obvious measure of shift is the distance between centres. We could look at a change in means, or as the box plots most help us see, a change in medians. Whereas the obvious measure of the extent of shift is the change in centres, changes in centres on their own often mislead.



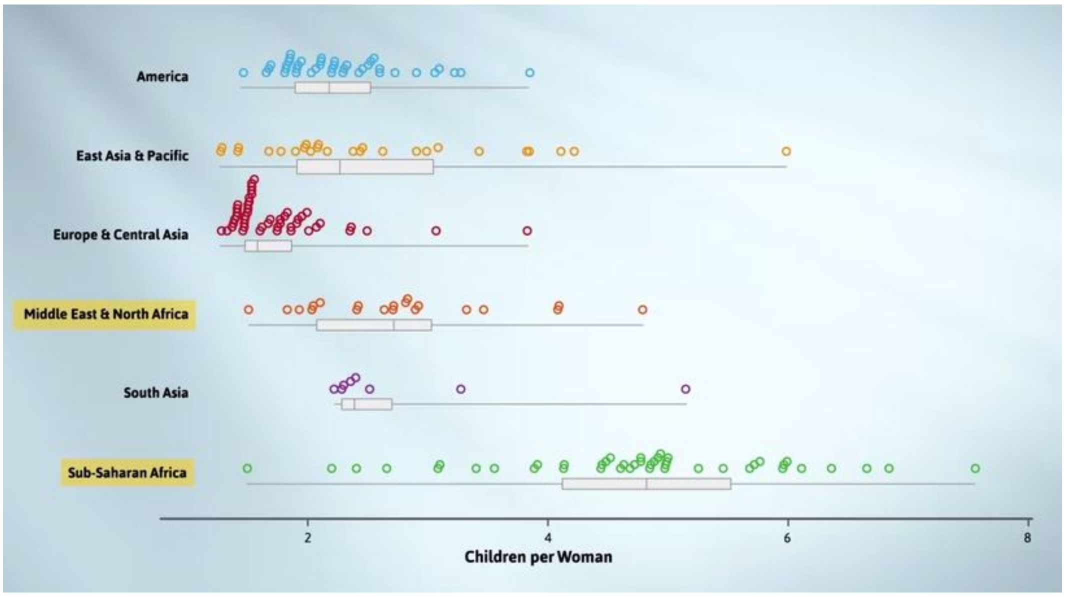
For instance, here are the regional means of Children Per Woman in 2012. News and other reports often just give averages like this and write stories from them.



Summary of ChildrenPerWoman by Region for 2012

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.dev	Sample Size	n.missing
America	1.453	1.881	2.191	2.269	2.523	3.844	0.516	40	10
East Asia & Pacific	1.265	1.905	2.26	2.581	3.052	5.983	1.11	26	13
Europe & Central Asia	1.268	1.464	1.568	1.74	1.862	3.819	0.463	48	8
Middle East & Nth Africa	1.498	2.08	2.704	2.768	3.011	4.785	0.84	20	0
South Asia	2.208	2.284	2.367	2.802	2.695	5.141	1	8	0
Sub-Saharan Africa	1.504	4.143	4.82	4.731	5.526	7.574	1.25	44	2

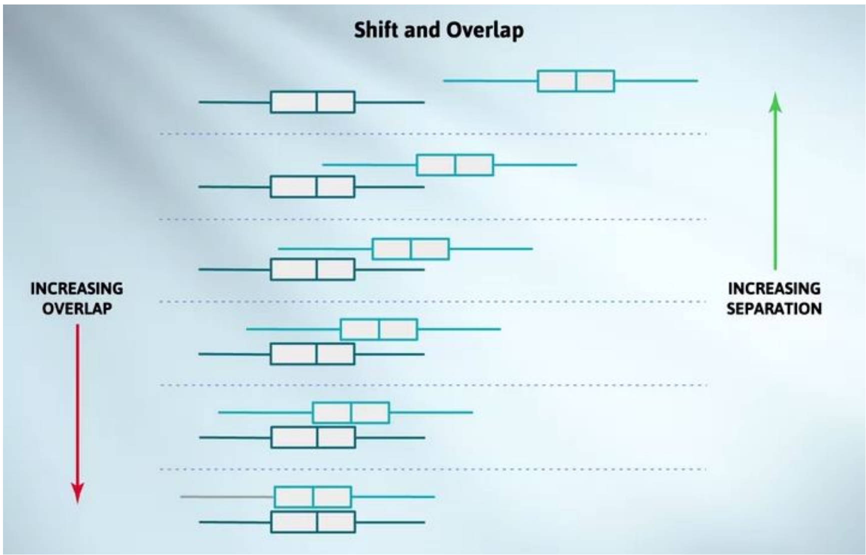
When that's all the information we have, we tend to think that what we're seeing in the averages is the way it is in general, losing sight of the fact there might be a great deal of variation around each of those averages. If we'd only been presented with the information here, we'd most probably conclude that Children Per Woman values are much bigger in Sub-Saharan Africa than in the Middle East and North Africa.

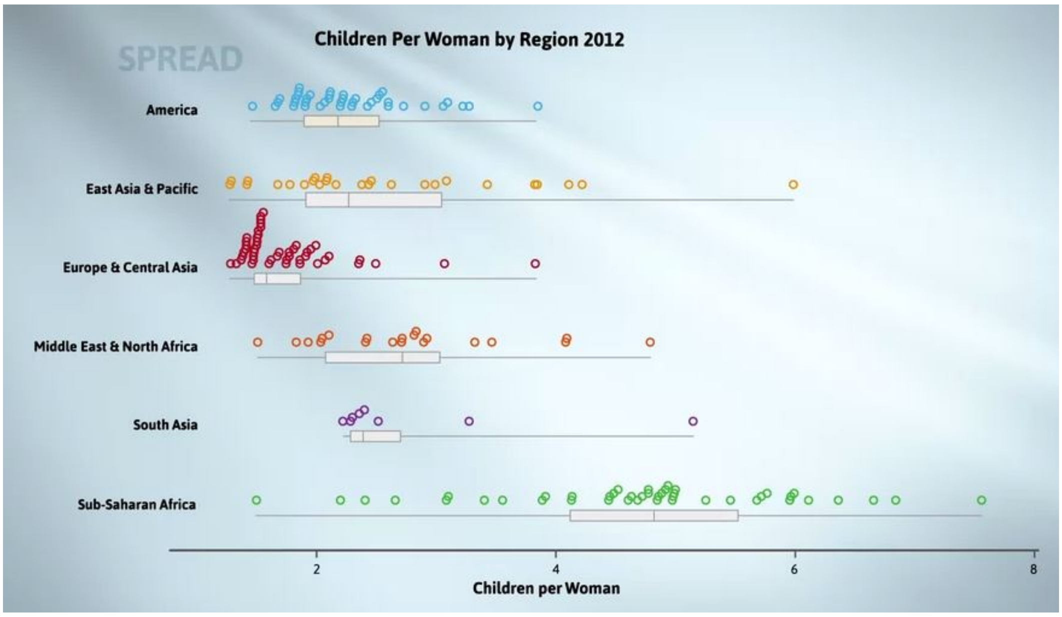


While that's usually the case, some countries in Sub-Saharan Africa, have quite small values, smaller than almost any in Middle East and North Africa. In fact, there is enormous variability in fertility in this region. They are spread all the way from about 1.5 to almost 8.

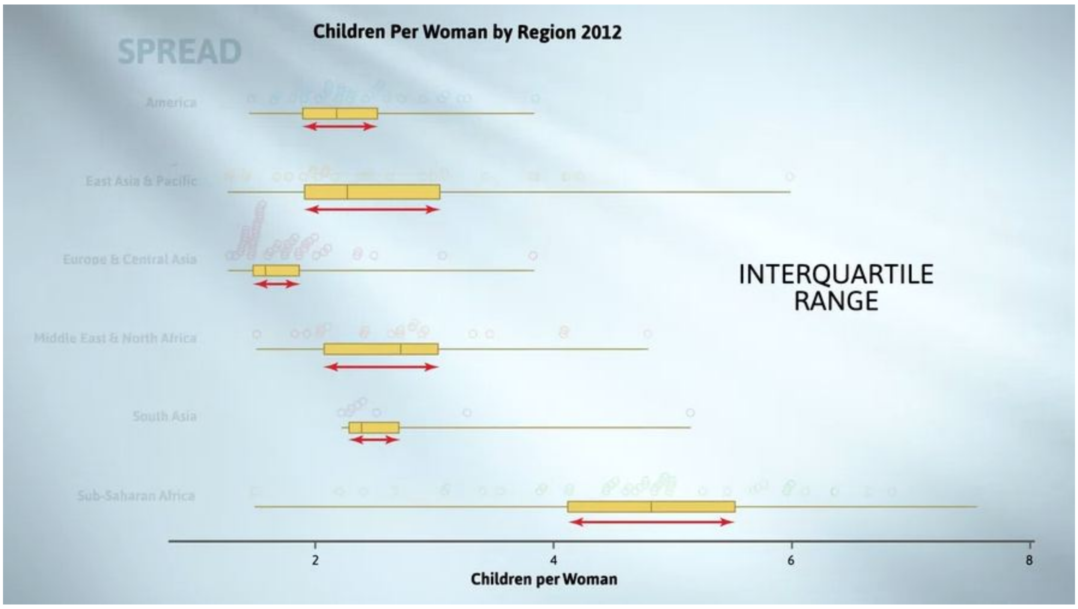
As a general rule, shifts should be considered in conjunction with background variability, which is shown by the spread of values we see. We should ask ourselves, "Is this a big difference compared with background variability or a small difference?" Differences that are very small compared to background variability are usually of no practical importance. They seldom help us make better predictions or make better decisions or design more efficient systems.

The main visual elements here are shift and overlap. Do we have complete separation, everything in one group is bigger than anything in the other? Is there a lot of overlap, shifts small compared to background variability? Or where are we in between these extremes?



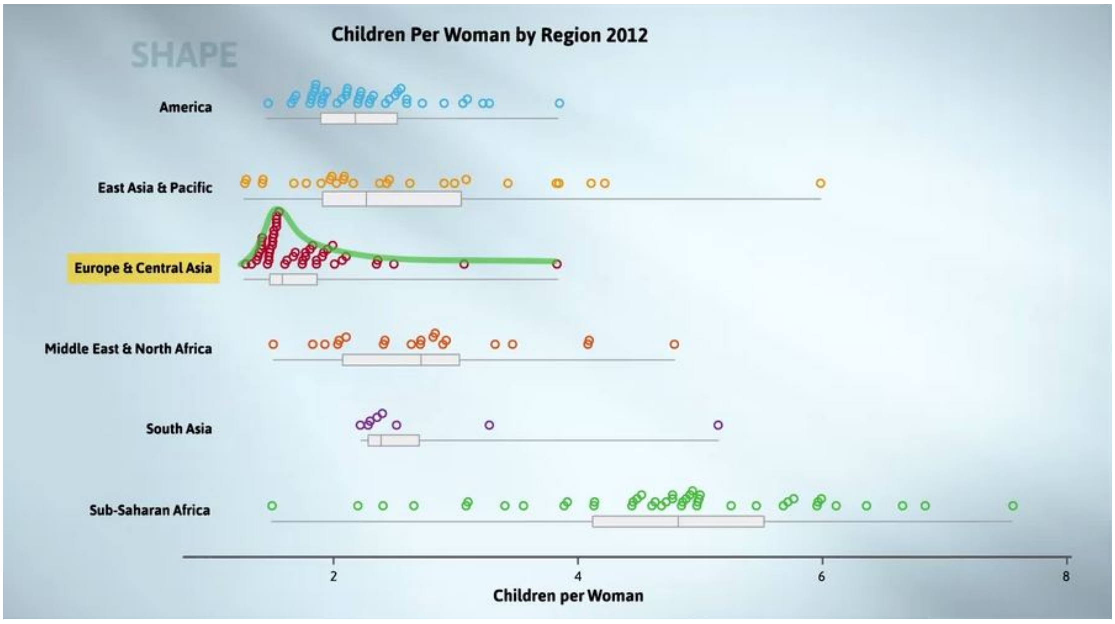


For reasons we'll get to later in the course, we tend only to pay attention to changes in spread if they're quite large or if they're being seen in quite large data sets.



The spread or variability measure on our box plots is the interquartile range. It's the length of the box in the box plot. The natural way of comparing spreads is multiplicative. In other words, "Does the second one look twice as big or half as big again as the first?" and so on. In the countries plot, only the small variability of Europe and Central Asia and South Asia are really remarkable.

We only pay attention to changes in shape when they're fairly extreme and they're being seen in very large data sets.

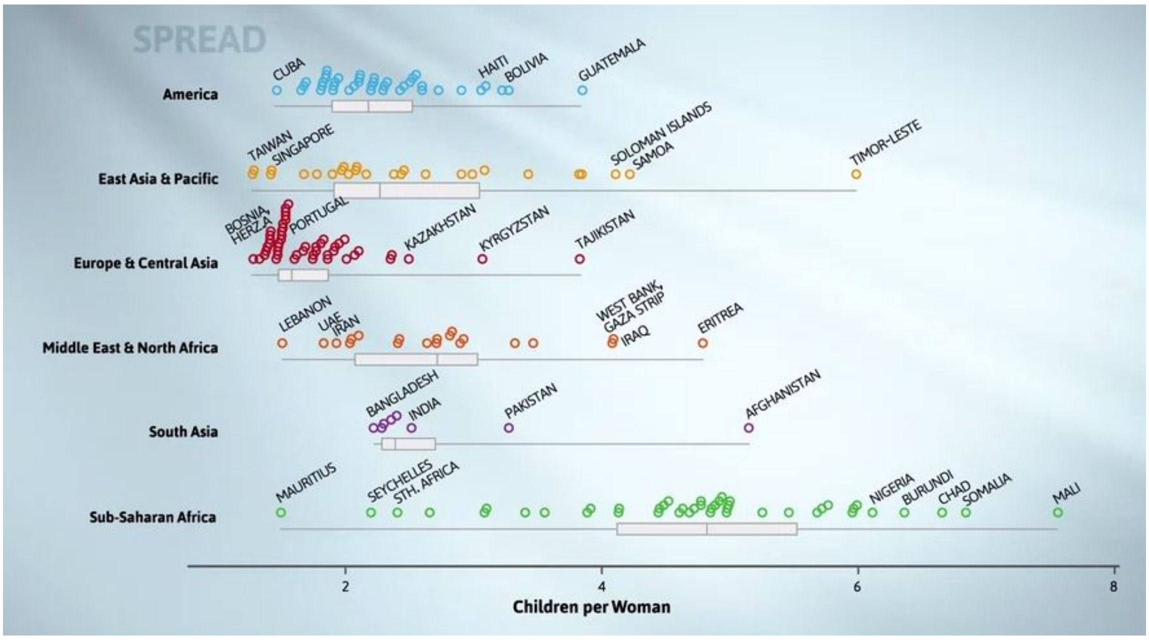


The only shape that's remarkably different here again belongs to Europe and Central Asia, where the distribution is very skewed, bunching up against 0 with a long right tail. Now let's look into this graph in more detail and see some of the things that stand out.

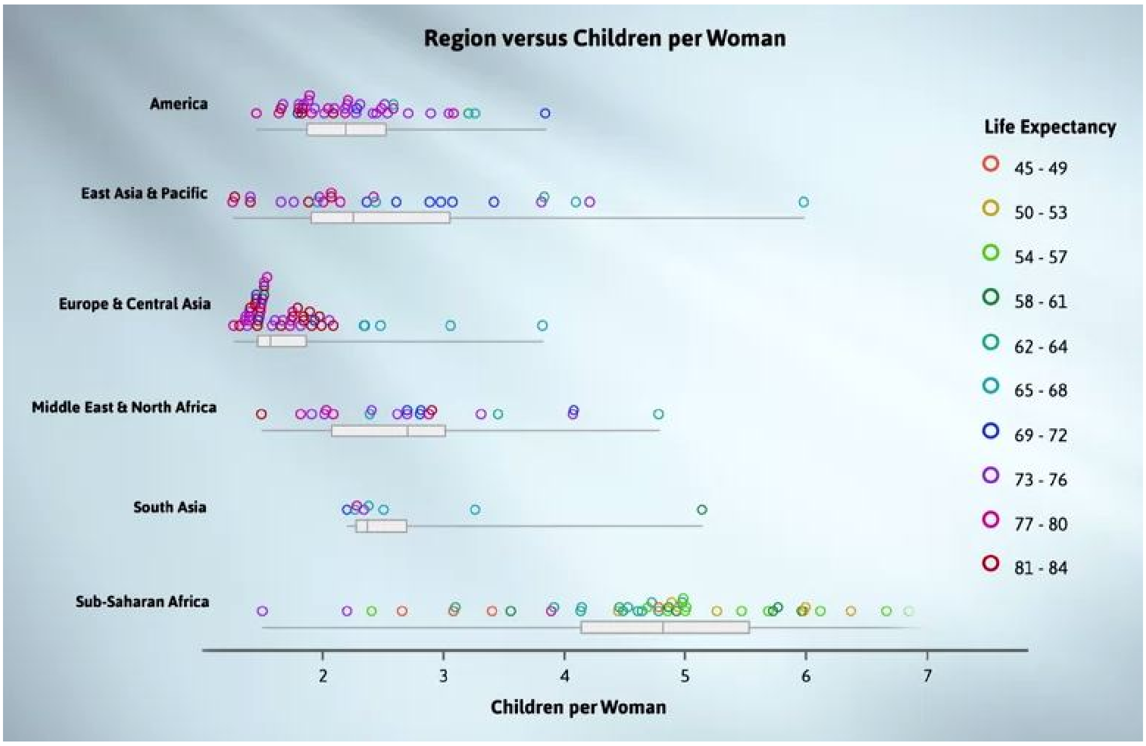
There's a lot of variation in Children Per Woman even within the same region. Put another way, even countries in the same region are very different on this variable. For example, East Asia and Pacific have one country with average Children Per Woman of about 1 at the bottom end and another country with 6 at the top end, and an almost uniform spread between 1 and 5.

There's an obvious tendency for considerably fewer children per woman in Europe and Central Asia. This group is shifted furtherest to the left. Next come America, then East Asia and Pacific, and then South Asia, with roughly similar centres in about 2.2 to 2.4. We then go up to the Middle East and North Africa, centred about 2.7. Then shifted considerably further to the right is Sub-Saharan Africa, centred at just under 5.

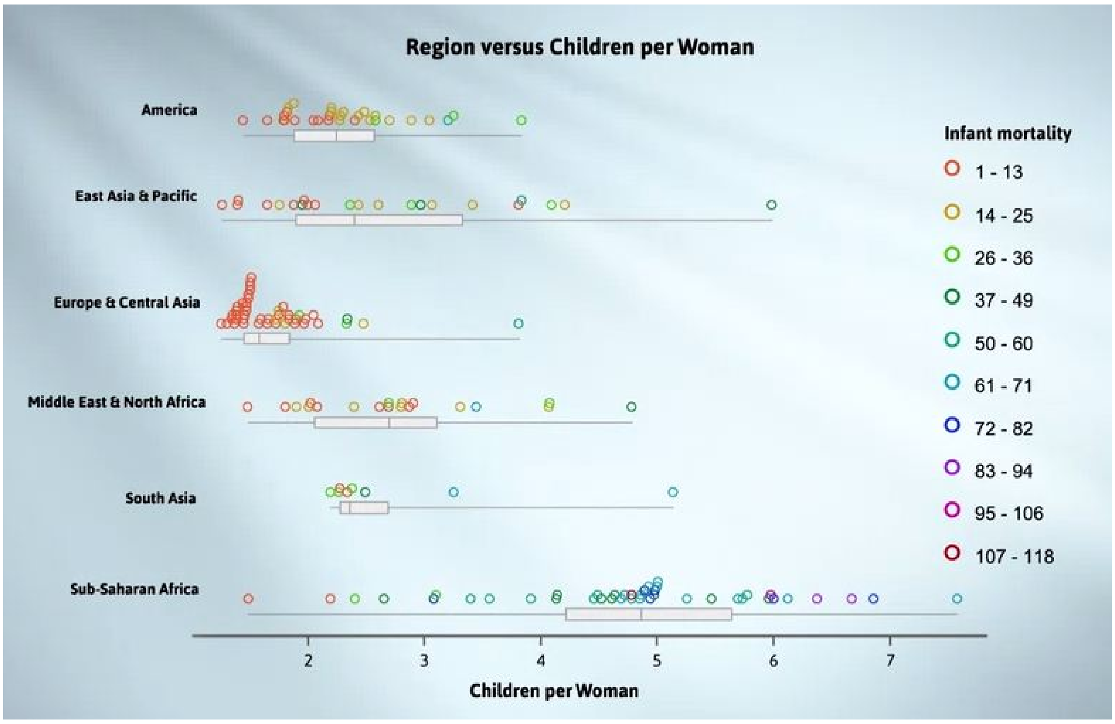
Because of all the variation in values from countries from the same region, we may guess there's a lot more than region going on here.



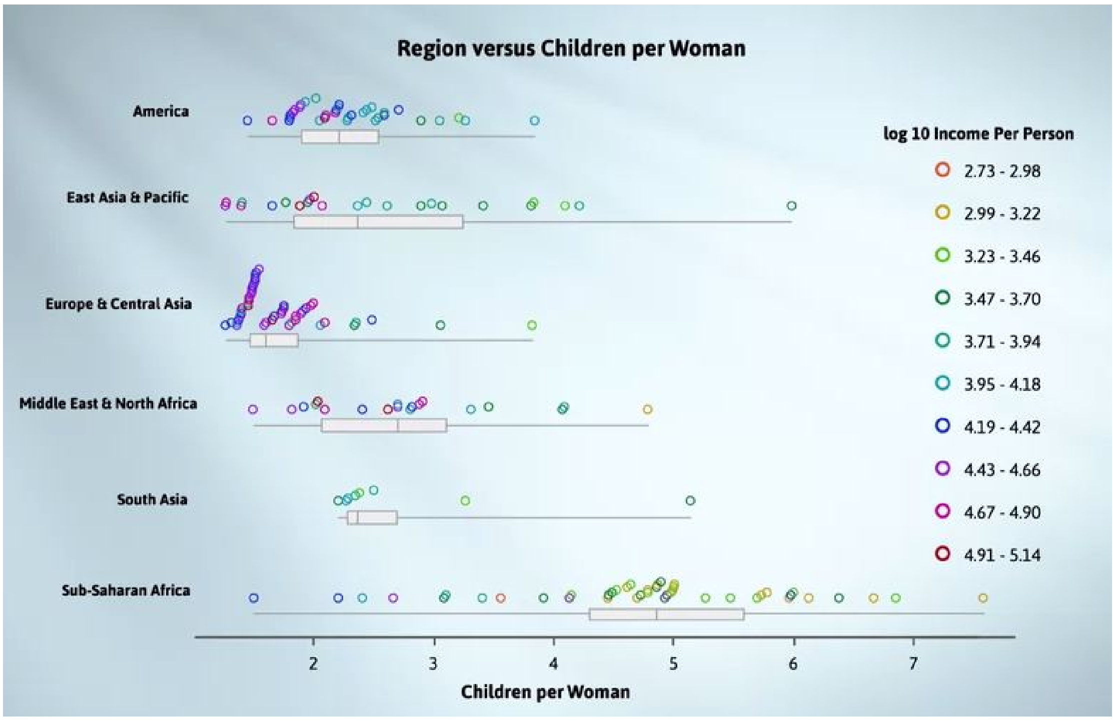
Here we've annotated the 2012 graph with the names of countries with extreme values. We might wonder what factors countries with high numbers of children per woman might have in common. I'll use colour to look at a couple.



First, I've coloured by life expectancy. The high life expectancies are the crimsons and purples. The low life expectancies are the greens into the odd brown and orange. The countries on the left-hand side are the high life expectancy purples. The countries out to the right are mainly the low life expectancy greens. So high numbers of Children Per Woman are associated with low life expectancies.



Next, I've coloured by infant mortality. The countries towards the left are the low infant mortality oranges, while those out to the right tend to be the high infant mortality greens and blues. There are very few purples. So high numbers of Children Per Woman are associated with high infant mortalities.



Last, I've coloured by income per person on a log scale. The countries towards the left-hand side tend to be the high income crimsons and purples, while those towards the right tend to be from the lower income end of the colour range. So high numbers of

Children Per Woman seem to be associated with factors that tend to go along with poverty. That brings us to the end of this video. And I'll leave you with these questions to remind you of the ideas we've just covered.

QUESTIONS

- When we compare children per woman, between regions, how many variables were involved and what were they?
- What are the main things we look for when comparing groups using dot plots?
- Why, as a way of transmitting group-comparison information, is simply quoting group averages often misleading?
- What is the natural way of measuring the extent of a shift?
- When looking at the extent of a difference between group centres, what else should we be relating that to when trying to gauge how big or important change is?
- When do we pay attention to changes in spread or variability?
- How should we measure a "change" in spreads?
- When do we pay attention to changes in shape?
- Why shouldn't we look at changes in centres in isolation?
- What can we do with sets of dot plots to explore the effect of another variable?