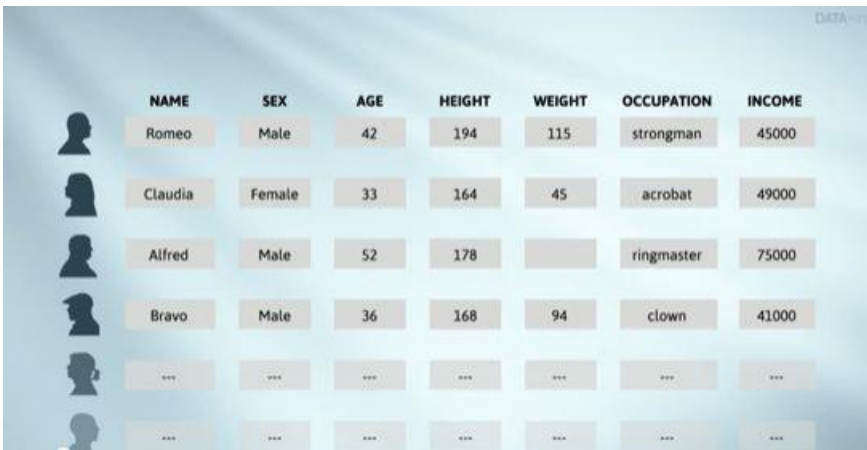








WEEK 1 1.12 DATA ORGANISATION by Chris Wild

Welcome. In this video, we'll show how statistical data is most commonly organised and stored, and learn some essential terminology. In this clip, we've been recording data on individual people. We've organised the data so that each row corresponds to a person and the cells in that row are the pieces of information we've collected about that person. We can see what sort of information each piece is by looking at the corresponding column header.



	NAME	SEX	AGE	HEIGHT	WEIGHT	OCCUPATION	INCOME
	Romeo	Male	42	194	115	strongman	45000
	Claudia	Female	33	164	45	acrobat	49000
	Alfred	Male	52	178		ringmaster	75000
	Bravo	Male	36	168	94	clown	41000
							
							

Here are the incomes for each of our people. Here are their occupations. And so on. This is a very simple and easily-understood way of organising and storing data. It's also the default organisational format expected by statistical data analysis programmes. Data organised like this is often called rectangular data.

A row corresponds to an individual thing about which data has been recorded and a column corresponds to some property that is being recorded for each of those things.

We use the database term "entities" for the things we're recording data about. And the standard statistical term "variables" for the properties being recorded. The entities we collect and store individual-level data about could be anything. In our example data set, the entities are the individual people. In other data sets, the entities could be cities, countries, animals, plants, companies, financial transactions,

telephone calls, and so on. The important point is that we have additional data about each one of these entities.

We'll be using rectangular data as our data organisation format almost exclusively. Manipulating data sets to transform them from "the way the data was organised (or disorganised!!) when it arrived" to a format that a data analysis programme likes, can take considerable amounts of time. There are special skills to be learned there. But those are skills for another course.

Our data may not be complete. There are often gaps in the data for various reasons. The Weight cell for Alfred the ringmaster is empty, perhaps because he refused to answer the weight question. This cell is said to contain a missing value. Rather than leave the cell blank as we have here, a commonly used alternative is to use some sort of code to signify that "the value for this cell is missing" (a missing-value code). For example, in addition to interpreting empty cells as representing missing values, iNZight also by default interprets a cell containing just NA or NULL as saying, "This entry is missing".

Different programmes have different default missing-value code conventions. But they also usually have a way for the users to tell the programme that they want some particular value to be treated as a missing-value code (for example, -99). We can immediately distinguish between two types of variables in our data set. Age, height, weight, and income are numeric variables. All of their values are numbers and each can easily be thought of as a type of measurement that we're making on each individual.

Sex and occupation are categorical variables. They are different ways of "putting people in boxes", or, in other words, different ways of placing people or entities into groups or categories. A categorical variable may use numbers and/or alphabetic characters to label a category. It all comes down to the investigator's intent. Do I want to treat this variable as being like a measurement? Or do I want to treat the numbers in this column as different labels that refer to different groups?

If you want a variable containing only numbers to be treated as categorical (defining groups) you will have to inform your analysis programme of this. (In iNZight, we use "Convert to Categorical" under the "Manipulate Variables" menu.) You may have noticed that we haven't classified Name in this data set. It's just an identifier that gives us a way of specifying which person or row we are talking about. There are distinctions we can make between variable types that are finer

than the simple numeric/categorical distinction. But we won't make them until they're needed.

Why should we care about different types of variables? We need to care because the ways in which we look at data on a variable, and the way computer programmes behave, are largely determined by its type. Finally, I'll leave you with these questions to remind you of the ideas we've just covered.

QUESTIONS

What does a row in a rectangular data set correspond to?

What does a column in a rectangular data set correspond to?

What is a variable?

What is an entity?

What are the two variable types we are distinguishing between?

Why do we make these distinctions?