ScienceDirect

Available online at www.indianjrheumatol.com and www.sciencedirect.com

Nuts and Bolts in Clinical Research

Data types – The first step to planning your study

Varun Dhir [a,*], *Vivek Arya* [b]

[a]*Department of Internal Medicine, PGIMER Chandigarh,* [b]*Department of Internal Medicine, PGIMER, Dr RML hospital, New Delhi, India*

ARTICLE INFO

Article history:
Received 3 February 2014
Accepted 5 February 2014
Available online 25 February 2014

Keywords:
Data
Statistics
Numerical
Categorical
Dichotomous

ABSTRACT

This first in the series on clinical research will focus on the basics of data – the types of data (with examples of a rheumatoid arthritis study) and how to numerically and graphically represent these. Knowing the types of data is indeed the first step to planning a study as it leads to a proper planning, entering of data and finally suitable analysis.

1. Introduction to the series on clinical research

A series on clinical research including statistics is not a new idea. Indeed, many journals have excellent series available. Why then here (in the IJR) and now? The purpose of this series will be to give busy clinicians (and researchers) a relook at the basics of statistics and research methodology. This series endeavors to fill these gaps. (Disclaimer: This is not meant to replace a textbook or the experts!)

2. What is data?

Anything that we measure or categorize generates 'data'. Statistics is the science of dealing with this data – involving planning of collection of data, description of that data and then analysis. We collect multiple variables (something which varies) of different cases (often patients for us). Whether we like it or not, data is hitting us in all directions in daily life. Simply put, if we are measuring or categorizing something (in anyway), we are generating data.

3. Why do we need data?

Sir Arthur Conan Doyle put it aptly and we quote "It is a capital mistake to theorize before one has data". We commonly use data to generate descriptions (descriptive statistics), as in cricket where there is a description of the average of the previous runs scored or number of 50s scored or categorizing our research by the average number of citations one gets or the h-index! However, it is drawing inferences and analysis (inferential statistics) that is the tough part (not covered here!) that actually leads to accurate comparisons of these descriptions.

* *Corresponding author.*
E-mail address: varundhir@gmail.com (V. Dhir).

0973-3698/$ – see front matter
http://dx.doi.org/10.1016/j.injr.2014.02.002

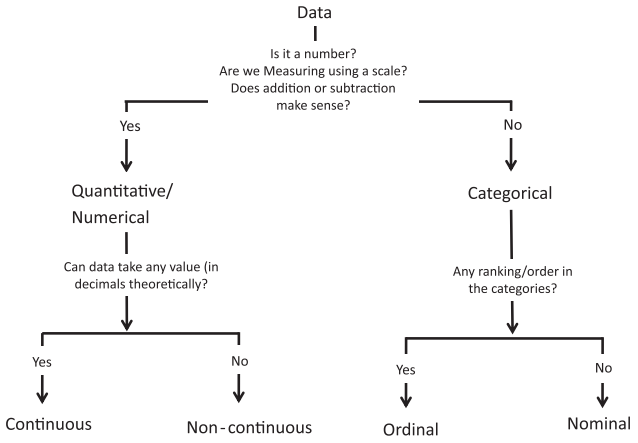

Fig. 1 – Classification of data.

4. Why do we need to know about types of data?

We need to know the type of data to be able to properly plan its collection, enter it appropriately and most importantly for analysis. The mathematical treatment of data is vastly different depending on the nature of the data. This means the kind of tests applied for describing and for analyzing data depends primarily on the kind of data. Indeed, becoming familiar with the data and its characteristics is the first and most essential step in performing statistical evaluation. An important part of this process is to *see* the data visually by converting them into graphs.

5. What are the types of data/variables?

Data can be of many types, but it can be divided into numerical or categorical (Fig. 1). It must be noted that the classification of data is not uniform. Examples of each of the data types as in a study on rheumatoid arthritis are given in Table 1.

5.1. *Numerical/quantitative data*

That which can be measured by a scale of some sort and adding, subtracting etc makes sense.

Continuous data: the measurement can assume any value along a continuum and can be meaningfully subdivided into finer and finer increments, depending upon the precision of the measurement system. It has some units, although this may or may not have any absolute meaning. As it is measured by a scale, it is often also called 'scalar'.

Non-continuous data/counts: this data is not continuous but numerical in the form of counts, i.e., takes the value only of integers. Ideally should be analyzed using Poisson distribution for count data. Sometimes, if the scale is large it is treated as continuous data. Sometimes, this data is used to generate a score, that is continuous (like tender joint count (TJC), swollen joint count (SJC) used for disease activity score 28 (DAS28)).

5.2. *Categorical/qualitative data*

This is data that is placed/categorized into certain pre-defined categories. This can be nominal or ordinal.

5.2.1. *Nominal*

This is data in categories; however, there is no hierarchy to these categories. If there are only two categories, it is called binomial or bivariate or binary. For example gender (male/female), religion or nationality.

5.2.2. *Ordinal*

There is a hierarchy or order to the categories. A good example is severity of pain scored as none, mild, moderate, severe and very severe. This can be derived from continuous data but not

Table 1 – Examples of different types of data for a rheumatoid arthritis study.

	Numerical data		Categorical		Survival/time-to-event
	Continuous data	Non-continuous/ Count	Ordinal	Nominal (including binomial)	
Demographic	Age, weight, height, duration of disease	Points scored on ACR/ EULAR 2010[a] criteria	Weight: underweight, normal, over-weight, obese	Type: early RA, advanced RA; married: yes/no; gender: male/female	
RA activity scores	Disease activity score, ESR, visual analog scale (patient or physician), grip strength	Tender joint count, swollen joint count	High, moderate, low disease activity, or remission (EULAR[b] criteria)	Active or inactive (active = yes or no)	Time to remission
RA damage scores		Ven der Hedje score, damaged joint count	Larsen scoring: grade 1–5	Erosion (yes/no)	Time to first detection of erosion
RA function	HAQ[c], SF 36[d]	WHO Quality of life	Steinbrocker functional classes		Time to loss of occupation

[a] ACR/EULAR 2010 = American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis.
[b] EULAR = European League Against Rheumatism.
[c] Health assessment questionnaire.
[d] Short form 36.