Introduction to Statistics

John Karuitha

3/17/2021

Introduction to Statistics

Background: In the mainstream media and social media, we are constantly bombarded by a lot of data. In everyday language, the term data connotes numerical types of data- population numbers, age, etc. However, much of the data today is unstructured- from the music you listen to on Sportify and YouTube Music, to videos you would watch on Netflix and YouTube. How best can we make sense of this data? Statistics is the wing of maths devoted to making sense of data? Note that not all data lends itself to statistical analysis. In this course we examine how statistics allows us to make sense of data. But first, some definitions.

Defining statistics

- The term statistics has several meanings;
- Statistics the subject: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.
- Statistics as values computed: Statistics (plural) are quantities calculated from (sample) data.

The two types of statistics - Descriptive vs Inferential

- Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data.
- Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made.
- They are simply a way to describe our data.
- For instance, when you get a dataset with height and weight of students in your class and compute averages, variance and standard deviation.
- HERE You are simply describing the data and NOT using it to make conclusions about the height and weight of ALL students at Karatina university.

The two types of statistics - Descriptive vs Inferential

- In this level of statistics, we get data from a SAMPLE and use it to make conclusions about the WHOLE population.
- Usually, we have no access to the entire population and rely on sample data to draw inferences about the population.
- Example of inferential statistics is when you take a sample of 2nd year BHR students at Karatina University, take their height and weight, and use this data as a basis to infer the height and weight of ALL students at Karatina University.

Defining data;

- What are data? (singular- datum): Data are qualitative or quantitative values along with their context.
- Modern datasets are so large; think of how much data safaricom collects in a day. Would you store that data in an excel spreadsheet. NO!
- Usually large datasets are stored in vast repositories (databases) called data warehouses, and accessed by advanced tools like SQL. Find out what SQL is!!

Big Data

- The sheer quantity of data has birthed a new name, big data.
- Big data are data so vast they cannot be handled using the traditional methods of storage (e.g. excel spreadsheet) and analysis (using SPSS) are inadequate.
- Big data connotes data that accumulates very fast (the velocity problem), from many sources (the variety problem), and is large in quantity (the volume problem) that it is challenging to tell whether or not it is authentic (the veracity problem).
- Think of a dataset with 150 billion rows of data and 1000 columns of variables. wheredo you start analysing that?

Data Analytics;

- Many companies are increasingly using past customer data to predict their future purchases, which allows them to serve customers better or target them better with adverts.
- This process of using data, especially of transactional data (data collected for recording the companies' transactions), to make decisions and predictions is sometimes called data mining or predictive analytics. The more general term business analytics (or sometimes simply analytics) describes any use of data and statistical analysis to drive business decisions from data whether the purpose is predictive or simply descriptive.

Data context

To make sense of data we need to answer several questions;

- Who is the data about? Individuals, corporations, machines, etc.
- What is the data is about? purchase history, weight, blood group, etc.
- When was the data collected? is it still relevant if it was collected 50 years ago?
- Where were the data collected? Is data collected in Somalia relevant to Kenya.
- Why was the data collected? What questions did we have in mind? For tax accounting purposes; For use in targeting adverts to consumers? In short what were your research questions?

Metadata

- Metadata is data about data.
- Check the last phone call you made.
- The metadata about the call tells you who you called, when you called, the duration of the call including the data and time, if you ask safaricom they will give you data about where you called from, including a verbatim of what you discussed (the what and why you called).
- Metadata allows us to aswer the who, what, when, where and why of data.
- Metadata provides the context of the data.

Storing data

Usually data are stored in data tables.

- Most databases are just huge data tables.
- Rows of the data table represent individual cases.
- Columns of the data represent the variables- the attributes of each case.

Example of a data table

Table 1: Data about men and women

Name	Age	Height	Weight	County	Date_of_data
Atieno	32	180	90	Mombasa	Jan 1, 2021
Mogaka	25	160	70	Laikipia	Jan 1, 2021
Etyang	18	200	74	Narok	Jan 1, 2021
Jeptoo	22	150	55	Wajir	Jan 1, 2021
Junior	10	120	35	Samburu	Jan 1, 2021

Data table

- Each row represents a case. For instance the first row tells us about Atieno, and so on.
- Each row is a record, information about an individual in a database.
- Cases go by different names, depending on the situation. Individuals who answer a survey are referred to as `respondents`. People on whom we experiment are `subjects` or (in an attempt to acknowledge the importance of their role in the experiment) `participants`, but animals, plants, websites, and other inanimate subjects are often called `experimental units`. Often we call cases just what they are: for example, customers, economic quarters, or companies. In a database, rows are called records—in this example, purchase records. Perhaps the most generic term is cases.

Data table

- The column titles (variable names) tell what has been recorded.
- Each recorded item in the column is called a variable, like name, height, weight etc.
- A general term for a data table like the one shown above is a `spreadsheet`, a name that comes from bookkeeping ledgers of financial information.
- In large datasets many datatables are linked together in a `relational database` like we do in Ms Access.

Variable Types

- When the values of a variable are simply the names of categories we call it a categorical, or qualitative, variable. Examples are Name and County in our table.
- When the values of a variable are measured numerical quantities with units, we call it a quantitative variable. Note they must have units to be quantitative.
- BE CAREFUL;
- Your National ID number is NOT quantitative!!!! Does it have units?? If you were to get the mean of the ID Numbers of your family members, how would you interpret it?
- Numbers like ID Numbers are identifier variables, just a special type of categorical or qualitative variables that you could easily confuse as quantitative.

Identify qualitative vs categorical (qualitative) variables

- Admission number; 111, 112, 113, 114, 115
- Height; Short, Medium, Tall, Tall, Short, Median
- Height; 150 cm. 180 cm, 200 cm, 120 cm.
- Weight: Underweight, Obese, Overweight, Ideal Weight, Overweight.
- Weight: 60 kg, 170 kg, 90 kg, 75 kg
- Age; Toddler, teen. teen, young adult, elder.
- Age; 15 years, 20 years, 17 years.
- Age; 15-20 years, 20-24 years, 50-60 years, 70 years and above.
- Notice how variables can be presented as qualitative or quantitative (see height/ weight/age).

Identify qualitative vs categorical (qualitative) variables

- Don't label a variable as categorical or quantitative without thinking about the data and what they represent. The same variable can sometimes take on different roles.
- Don't assume that a variable is quantitative just because its values are numbers. Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- Always be skeptical. One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. The context colors our interpretation of the data, so those who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan website. The question that respondents answered may be posed in a way that influences responses.

Variables types

- Categorical variables can either be ordinal vs nominal.
- Weight; light, medium, heavy, superheavy
- Age; Toddler, Teen, Young adult, Elder.
- Note that these two categorical variables have inherent order; the are ordinal.
- order $==$ ordinal

Variables types- Ordinal vs Nominal.

- Others have no inherent order, they are Nominal.
- Examples are gender (male, female); names of cities (Nairobi, Mombasa, Kisumu)
- Look at the county column in the table above, is there order; so are names.

Variable types- continous or discrete

- Numeric variables are either continous or discrete.
- Continous variables can take on any value inclusing fractions and zero.
- Continuous variables are usually measured like height, weight, speed.
- In many cases, variables like weight are usually shown as whole numbers but in reality can take on any values.
- We mostly present continuous values as discrete because of convenience or lack of precise tools for measuring,
- This problem of presenting continuous variables as discrete is called “discretization”.
- In reality, may be you are not 21 years old; you could be 21.01 years old if you put in the days, hours, minutes and seconds.

Variable types- continous or discrete

- Other numeric variables are discrete- they can only take on whole numbers.
- Can we have 2.5 students in a class?
- Think of discrete numbers as counts.

Data Types; Time series, cross-sectional, and logitudinal/ panel data

A time series is an ordered sequence of values of one or more quantitative variables measured at regular intervals over time. Time series are common in business. Typical measuring points are months, quarters, or years, but virtually any consistently-spaced time interval is possible.

Example: Time series

Table 2: An example of Time series data

Year	Salary	Weight
2010	10000	88.23229
2011	11000	86.12024
2012	12000	80.14054
2013	13000	86.50320
2014	14000	84.86394
2015	15000	84.79715
2016	16000	83.34149
2017	17000	89.19606
2018	18000	83.23908
2019	19000	82.95834
2020	20000	81.78926

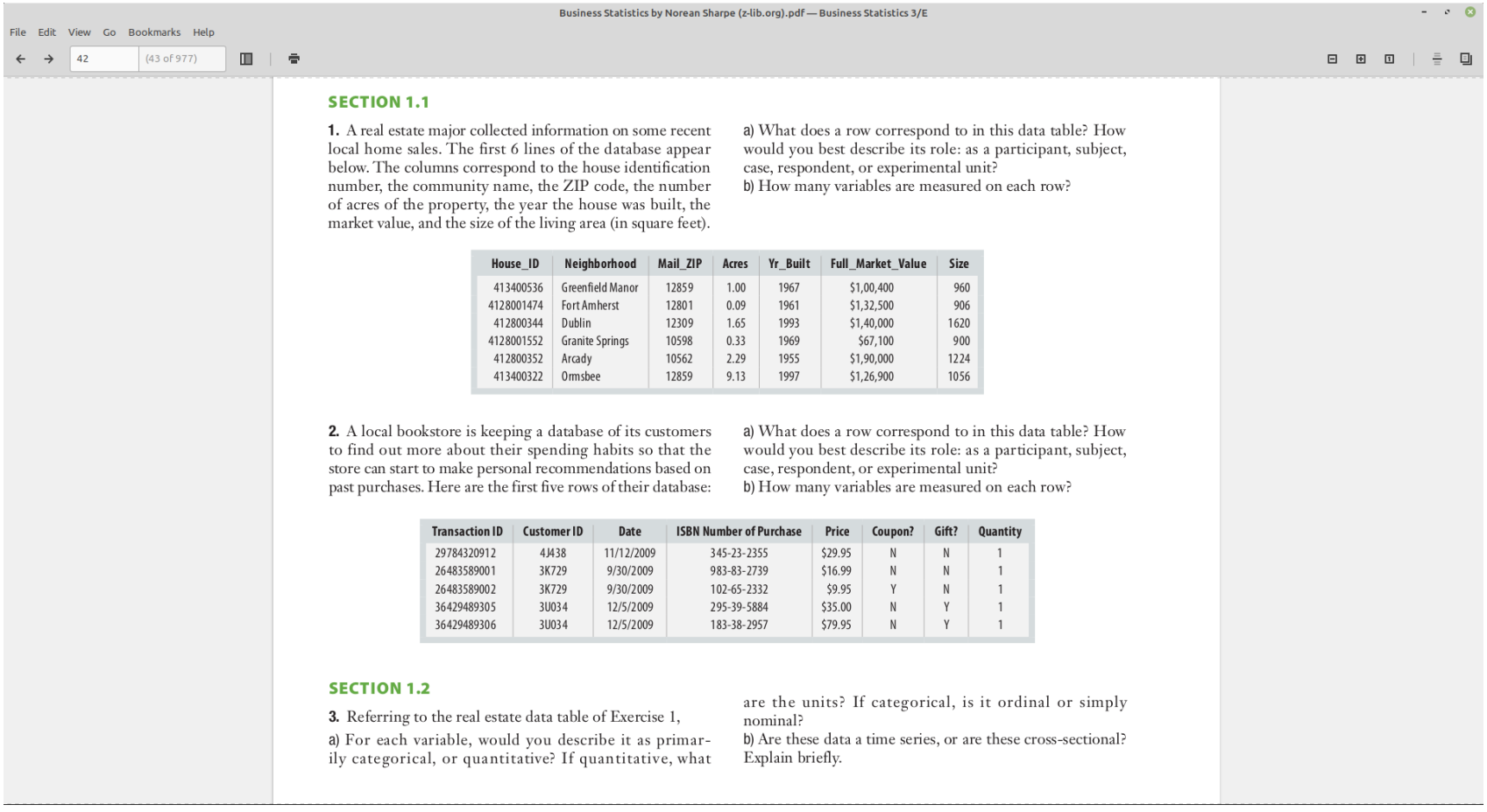
Cross sectional data

- For cross-sectional data, several variables are measured at the same time point, see table 1.
- The data in table 1 was all at a point in time, Jan 1, 2021.

Panel data

- Panel data has both a time series element and a cross-sectional element.
- Research on this.

Exercise



SECTION 1.1

1. A real estate major collected information on some recent local home sales. The first 6 lines of the database appear below. The columns correspond to the house identification number, the community name, the ZIP code, the number of acres of the property, the year the house was built, the market value, and the size of the living area (in square feet).

a) What does a row correspond to in this data table? How would you best describe its role: as a participant, subject, case, respondent, or experimental unit?
b) How many variables are measured on each row?

House_ID	Neighborhood	Mail_ZIP	Acres	Yr_Built	Full_Market_Value	Size
413400536	Greenfield Manor	12859	1.00	1967	$1,00,400	960
4128001474	Fort Amherst	12801	0.09	1961	$1,32,500	906
412800344	Dublin	12309	1.65	1993	$1,40,000	1620
4128001552	Granite Springs	10598	0.33	1969	$67,100	900
412800352	Arcady	10562	2.29	1955	$1,90,000	1224
413400322	Ormsbee	12859	9.13	1997	$1,26,900	1056

2. A local bookstore is keeping a database of its customers to find out more about their spending habits so that the store can start to make personal recommendations based on past purchases. Here are the first five rows of their database:

a) What does a row correspond to in this data table? How would you best describe its role: as a participant, subject, case, respondent, or experimental unit?
b) How many variables are measured on each row?

Transaction ID	Customer ID	Date	ISBN Number of Purchase	Price	Coupon?	Gift?	Quantity
29784320912	4J438	11/12/2009	345-23-2355	$29.95	N	N	1
26483589001	3K729	9/30/2009	983-83-2739	$16.99	N	N	1
26483589002	3K729	9/30/2009	102-65-2332	$9.95	Y	N	1
36429489305	3U034	12/5/2009	295-39-5884	$35.00	N	Y	1
36429489306	3U034	12/5/2009	183-38-2957	$79.95	N	Y	1

SECTION 1.2

3. Referring to the real estate data table of Exercise 1,

a) For each variable, would you describe it as primarily categorical, or quantitative? If quantitative, what are the units? If categorical, is it ordinal or simply nominal?
b) Are these data a time series, or are these cross-sectional? Explain briefly.

Figure 1: Check out these questions

Additional resources

- What is panel / longitudinal data
- https://www.youtube.com/watch?v=7_GdwN_iwmw