BACK TO BASICS: UNDERSTANDING THE NUMBERS BEHIND COVID-19

TEACHING NOTE

Manoj Dayal Chiba
University of Pretoria's Gordon Institute of Business Science, Johannesburg, South Africa

CASE SYNOPSIS:

The case is set during the COVID-19 pandemic and the South African government's response to the pandemic. A brief timeline is provided as part of the introduction to the case study, with the following being a timeline of the events: a. 14 March 2020, 114 South African citizens were repatriated from Wuhan the epicentre of the COVID-19 outbreak; b. 15 March 2020, South Africa's President, Cyril Ramaphosa declares a National State of Disaster, and this includes various measures to protect against the spread of COVID-19, while the healthcare system is geared up to deal with the pandemic. Among the measures implemented, travel bans from high-risk countries and closing of air-traffic, closing of land ports and banning of gatherings of more than 100 people. c. 23 March 2020, President Cyril Ramaphosa announced a national lockdown beginning on 27 March 2020 for three weeks. d. 9 April 2020, President Ramaphosa extends the national lockdown by a further two weeks. The World Health Organisation (WHO) had commended South Africa on the swift action taken to curb the spread of the virus.

Individuals and organisational leaders are grappling to make sense of the spread of the virus, and the barrage of the information that is being communicated through multiple channels, formal and informal. In order to make sense of the information, the case is premised on getting access to the raw data and conducting the analysis based on the publicly available data.

The central requirement of the case is to compare the number of positive cases per million, based on the population data contained in the dataset, of South Africa to a comparable country.

METHOD:

The case relied on secondary data which was publicly available, in the form of popular press articles, government communication channels, blog posts and popular news publishers. The reason for the choice of these sources was to serve the central themes of bombardment of information, as well as often contradictory information sources. Finally, given the World

Health Organisation’s (WHO) commendation of South Africa’s response to the COVID-19 pandemic was used to build the comparative nature of the case, as this commendation was not publicly announced to all countries response, specifically for other emerging market countries.

The case begins with a timeline of information from South African History Online which provided a timeline of the key events up to 4 May 2020, the time of writing the case. Considering this article as the starting point, further information from the South African Modelling Consortium was referred to. The redacted graphs in Exhibits 8 and 11 was then used to build the case of how data is presented serves to convey a message.

In order to build the narrative, raw data from the European Union Open Data Portal (EU ODP) was then consulted to understand the spread of the virus through different countries globally. This dataset then gave rise to understanding how the accessing credible raw data is relatively easy, but how one may establish the credibility of raw data sources is important.

ASSIGNMENT QUESTIONS:

1. What are the criteria for establishing credibility of data source?
2. Produce the relevant descriptive statistics for South Africa.
3. What would be an appropriate country to compare to South Africa? Why?
4. What is an appropriate statistical test for the comparison? Why?

TARGET AUDIENCE:

Post-graduate students learning statistics as part of a degree programme. The case assumes no prior statistics knowledge and therefore is aimed at teaching the importance of the basics of statistical analysis and then progressing to tests for differences.

LEARNING OBJECTIVES:

1. How to establish credibility of data sources;
2. Measurement scales of data
3. The importance of descriptive statistics and generating the following based on the type of data: Mean, Median, and Standard Deviation.
4. Graphical methods
5. Test for differences: t-test, ANOVA

TEACHING PLAN AND OBJECTIVES:

10-15 Minutes: INTRODUCTION TO COVID-19

The course instructor should spend 15-20 minutes providing the context of COVID-19 and a brief history (see supporting material). The pandemic timeline is important to place the case in context. Further points to be highlighted which are directly linked to the case:

1. 31 December 2019: Cluster of cases related of Pneumonia identified in Whuhan province of China.;
2. Given the case is set in South Africa, it must be noted that December and January are generally considered months in which holidays are taken, with return to work generally mid-January for South Africans. Note this is a generalization.
3. 11 March 2020: The World Health Organisation (WHO) characterizes COVID-19 as a pandemic.

10 Minutes:

The course instructor should provide a synopsis of the case.

15-20 Minutes: Objective 1: CREDIBILITY OF DATA SOURCES

The course instructor should now discuss the ubiquitous availability of data in the public domain. Students should be tasked with sharing publicly available raw data sources that they know or have come across. Some examples that the course instructor may want to show are given in the supporting materials section.

The discussion should then be focussed on how to establish credibility of the sources of data. Students should be asked how would they establish the credibility of a website. Answers may include government websites may be credible, ease of access among a host of others. The instructor should then focus the discussion on the following:

1. Why is the organisation/individual collecting the data?
2. Is there a listed author/individual/contact details that can be called/emailed? That is, is there an individual which can be contacted for queries on the dataset?
3. Where is the organisation/individual obtaining the data from?

4. Quality of the website from which the dataset is downloaded – assessing professionalism, writing style, number of adverts.

For each of the above, the dataset has been obtained and is publicly available on the https://data.europa.eu/ The establishing of credibility of the data source is therefore undeniable given the above questions and the dataset for the case.

20 – 25 minutes: Objective 2: MEASUREMENT SCALES OF DATA

The instructor should next provide some background into measurement scales:

1. Nominal;
2. Ordinal;
3. Interval; and
4. Ratio.

The instructor should then probe what type of measurement scale is the variable, number of positive cases as per the dataset and probe students on why. This is a good opportunity to present the below table such that a summary of each type can be seen.

Table 1: Measurement scale

	Nominal	**Ordinal**	**Interval**	**Ratio**
Order of values known		✓	✓	✓
Counts (frequency of distribution)	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median			✓	✓
Mean			✓	✓
Quantify the difference between each value			✓	✓

Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has “true zero”				✓

Post the discussion above, the course instructor should lead the discussion on descriptive statistics.

30-35 Minutes: Objective 3: DESCRIPTIVE STATISTICS

The descriptive statistics are central to the case, and what is meant by Back to Basics in the case.

The course instructor should now lead the discussion on the importance of beginning any statistical analysis with the **appropriate** descriptive statistics. The instructor should lead the students on the following, and provide the definitions of each:

1. Mean: The sum of the values divided by the number of observations;

The course instructor may want to lead the discussion on the sensitivity of the mean to outliers. This is not mandatory for the case.

2. Median: The middle value when data are arranged in ascending or descending order.

If the course instructor has chosen to discuss the mean and outliers, a discussion of the median and outliers not affecting this measure should be then discussed.

3. Standard deviation: Square root of variation. The course instructor should explain the importance of understanding the mean and the median in conjunction with the standard deviation.

Based on the above identified measurement scales, the students should now identify what basic descriptive statistics can be calculated given the type of data. The course instructor should be clear that the data contained is ratio data and therefore the mean, median and standard deviation can be calculated.

The course instructor should then ask for the descriptive statistics of the chosen country to be calculated. While the case deals with South Africa and the case eludes to a South Africa centric analysis, given the nature of the data the course instructor may want to calculate the descriptive statistics for any of the countries listed in the dataset. The dataset contains data for 206 countries.

10-15 Minutes: Objective 4: GRAPHICAL METHODS

The course instructor should next lead the discussion on graphical display of data. The exhibits 1-3 should be looked at – that is, a discussion on how different types of graphical methods may be used for the same data. The instructor should then lead the discussion on how the types of data dictates the method of graphical method used. It should be noted that students may request an understanding of the logarithmic graphs, and the instructor should be able to lead this discussion on why a logarithmic graph and how to interpret this.

The course instructor should then lead the discussion on alternate graphical methods that could be used for the type of data that is contained in the dataset. Given the ratio data the course instructor needs to identify that bar graphs may be **inappropriate.**

The course instructor should discuss Exhibit 6. The discussion should be around the representation of the number of positive tests. In the exhibit it should be noted that California is the highest, but more generically that a few states in the United States account for the vast majority. However if this is overlaid with the number total tests as is displayed in Exhibit 7, it becomes apparent that the states with the highest number of positive cases also have the highest testing. This is an important point to highlight how adding information may convey a different set of insights.

Finally, exhibit 8, is another important manner in which the data should be displayed, as it provides multiple metrics of COVID-19 on a single graph – that is the confirmed cases, but also the number of deaths. As exhibit 8 shows a large fraction of individuals have already recovered globally from the virus as the numbers are often communicated cumulatively rather than disaggregated.

30-45 Minutes: Objective 5: TESTS FOR DIFFERENCES

The course instructor should clearly explain tests for differences. That is, t-tests and ANOVA. When discussing t-tests the instructor needs to clearly explain the difference between dependent and independent samples t-tests. The case lends itself to an independent samples t-test. The reason is that there are no repeated measures primarily. Furthermore, the instructor should explain that t-tests are used when seeking to understand if there is a difference between **two groups**, and in context of the case, two countries. The null and alternate hypothesis for the t-tests are as follows (NOTE: Not the hypothesis of the case question, rather that of the t-test):

Ho: Mean of Group 1 (country 1) = mean of group 2 (country 2)
H1: Mean of Group 1 (country 1) ≠ mean of group 2 (country 2)

The instructor should then ask the class to select two countries to compare - based on the ability to use a t-test. The discussion should be on the comparability of the country, and in the context of the case, South Africa and India could be compared given the similarity as Emerging markets. The instructor may choose another two countries, but ensure that comparability can be done.

Prior to the execution of the tests, a new variable should be created by the students: Number of positive cases per million – theoretically this provides a rate – but more importantly allows for comparability for countries with a population of more than a million:

(Number of positive cases for Country A / Population for Country A) x 1,000,000

The instructor may then ask the class to execute the independent samples t-test. The decision point is then left to the arguments based on the countries selected and if the selected countries p-values are significant or not.

The instructor should next discuss the Analysis of Variance (ANOVA) test. Primarily, the instructor should focus on the use of the one way ANOVA test when seeking to understand if there is a difference between 3 or more groups, and in the context of the case, three or more countries. The null and alternate hypotheses for the ANOVA test should be discussed, and the below is consistent, if three countries are chosen:

Ho: Mean of group 1 (country 1) = mean of group 2 (country 2) = mean of group 3 (country 3)

H1a: Mean of group 1 (country 1) ≠ mean of group 2 (country 2) = mean of group 3 (country 3)

H1b: Mean of group 1 (country 1) = mean of group 2 (country 2) ≠ mean of group 3 (country 3)

H1c: Mean of group 1 (country 1) ≠ mean of group 2 (country 2) ≠ mean of group 3 (country 3)

The instructor should then ask the class to select three countries which could be used for the one-way ANOVA. The discussion should be on the comparability of the country, and in the context of the case, South Africa, India and Brazil could be compared given the similarity as Emerging markets. The instructor may choose another three countries, but ensure that comparability can be done.

The instructor may want to discuss the interpretation of probability values (p-values) but this is not a requirement of the case.

Prior to the execution of the tests, a new variable should be created by the students: Number of positive cases per million – theoretically this provides a rate – but more importantly allows for comparability for countries with a population of more than a million:

(Number of positive cases for Country A / Population for Country A) x 1,000,000

The instructor may then ask the class to execute the ANOVA. The decision point is then left to the arguments based on the countries selected and if the selected countries p-values are significant or not.

SUPPORTING MATERIAL:

Note: In order to teach the case, a brief history globally of the COVID-19 should be understood, as well as the brief history of South Africa's response a good timeline of events can be found at:

Globally: https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19.

South Africa: https://www.sahistory.org.za/article/covid-19-timeline-2019-2020

Examples of open-data sources:

www.gapminder.org
www.i2ifacility.org
www.statssa.gov.za
http://data.un.org/
https://www.google.com/finance

THE DATA SET:

The case uses the dataset “COVID-19 Geographic Distribution Worldwide” in Microsoft Excel. It must be noted that the data was accessed on 11 April 2020, and therefore the last date of data is 10 April 2020. It is advisable, that latest dataset be downloaded from: https://data.europa.eu/