Creating a Statistical Analysis Plan

*Introduction to Statistical Thinking and Data Analysis 2022-23  
Applied Statistics Lab — Week 1 (10 October 2022)*

An analysis plan helps you think through your data, what you will use it for, and how you will analyse it. Analysis planning can be an invaluable investment of time. It can help you select the most appropriate research methods and statistical tools. It will ensure that the way you choose or collect your data will help you get reliable analytic results.

### Learning objectives:

Given information about a noncommunicable (NCD) health problem and a request for health-related information, you will be able to create an analysis plan that includes the following:

* Research question(s) and/or hypotheses,
* Dataset(s) to be used,
* Inclusion/exclusion criteria,
* Variables to be used in the main analysis (outcome and exposure variables),
* Statistical methods (including descriptive and inferential analysis).

# Overview of the steps to analyse data

To accurately create an analysis plan, you should be familiar with the steps in analysing data.

### Descriptive analysis

An essential first step of statistical analysis is to describe the data being used. This is vital in understanding the sample population, checking the data is correct and identifying any outliers. Both graphs and tables may be used to visualise and summarise the data. For instance, in research papers the Table 1 often presents a summary of the study population. Here, the characteristics of participants in a COVID-19 vaccine clinical trial are summarised by trial arm1:

Table

Description automatically generated

Descriptive analyses are exploratory, conclusions about hypotheses should not be made at this stage. Descriptive analyses might include some or all the following:

* **Summarise the sample population characteristics, e.g., by person, place, and time.**
  + Summarise variables by frequency (counts, percentages) or using measures of central tendency (mean, mode, median) and measures of spread (range, variance, standard deviation).
  + Tables, bar plots and histograms could all be used, depending on the data
* **Determine the frequency of outcome variables.**
* **Two-way variable analysis** 
  + For example, the prevalence of obesity by sex or education level.
  + If the objective is to quantify associations between exposures and outcomes, two-variable table displays are useful. Rows represent levels of exposure (education) and columns representing presence or absence of the outcome (obesity).
  + Scatter plots and box plots could also be used.

### Inferential analyses.

By generalising from the sample population (the data), inferential statistics allows us to infer characteristics of the target population, perform hypotheses tests, understand the relationship between exposure and outcome variables and make predictions. Inferential analyses might include some or all the following:

* **Compute and interpret measures of association:**
  + Determine themagnitude of association between an exposure variable and an outcome variable.
  + For example, t-tests for continuous data, chi-square tests for categorical data, and other statistical tests as appropriate.
* **Calculate confidence intervals (and statistical significance):**
  + Confidence intervals are preferred over p-values (binary measure of statistical significance) as they provide information about the direction and magnitude of measures of association.
* **Multivariable analysis and modelling technique to address the hypotheses:**
  + Utilise the results of your exploratory analyse, the literature and previous experience when implementing your modelling strategy to determine a final model or set of models that best explain your data, including when choosing which exposure variables to include.
  + Modelling methods covered by this course include *linear regression, logistic regression,* and *survival analyses*.
* **Check model assumptions and consider sensitivity analyses** 
  + It is important to check your results are robust.
* **Assess for effect modification:**
  + Effect modification is present when an effect measure such as sex, age or geographic location is different at several levels in an exposure-disease relationship. This is evaluated through statistical assessment of interaction between variables.
* **Assess the effect of potential confounders:**
  + Confounding is an apparent association between disease and exposure resulting from a third factor that was not considered. A confounder is an independent risk factor for the disease that also happens to be associated with the exposure variable under consideration.

Note, confounding and effect modification will be covered in more detail in Principles of Epidemiology (weeks 7 and 8 respectively).

**Review of date types. Give two examples of each. Are they categorical or numerical?**

1. Continuous:
2. Binary:
3. Ordinal:
4. Nominal:
5. Discrete:

# The analysis plan

### Research questions and/or hypothesis

* Determine the general research topic or the questions you need to answer in the analysis.
* The hypothesis clearly states what is being investigated and should be testable.
* A typical hypothesis is the association of an outcome with a specific exposure or intervention.

### Dataset to be used

* Data may be primary (collected for the study) or secondary (collected for another purpose by other individuals or organisations).
* Commonly used datasets include:
  + Vital registration (number of deaths, cause of death for a country)
  + Cohort study data
  + Demographic health surveys (DHS) used in low- and middle-income countries
  + Population census
  + Randomised clinical trials
  + The National Health and Nutritional Examination survey (NHANES - U.S.)
  + Public health department case records
  + Cross-sectional studies
* Databases typically are representative of a population either through a census (all persons included) or a sample (number of people selected to represent the population). For example, NHANES 1999–2000 interviewed 9,965 persons in the United States, and the database includes hundreds of variables.
* Before attempting data analysis, it is important to determine the study design and methods. Identify whether the data include:
  + all persons in the population of interest (census)
  + a sample representative of the population
  + a sample not representative of the population
* Determine if the dataset contains the variables you need to answer the research questions.
* Assess how complete and recent are the data.

**You are asked to investigate whether there is a relationship between chocolate consumption and the academic attainment of students at school.**

**How could you address this question using primary data?**

**How could you address this question using secondary data?**

**What might be some of the strengths and weaknesses of each approach?**

### Inclusion / exclusion criteria

* Describe the criteria you will use to determine which data to analyse. For example, if you have data from an entire country or region but you work in a particular district, your inclusion criteria might include “all records of participants residing in District X.”
* Similarly, if you are assigned to the Diabetes Unit and you are analysing hospital discharge data, your inclusion criteria might be “all hospital discharge records with ICD-10 codes E10 to E14.” You might exclude readmissions within 3 days of a previous discharge (which is likely a continuation of the previous problem or a complication from the previous hospitalization rather than a new episode). If your intention is to look at discharge planning, you would exclude any patient that died while hospitalized.
* An exclusion criteria for clinical trials if often pregnancy or breast feeding.

### Variables to be used in the main analysis

* The Analysis Plan should contain a list of variables to analyse that will be kept in the analysis dataset. For example, if the original dataset contains information about income, but your analysis does not need to include income, then the analysis file would not include the income variable.
* Also list variables that are not in the original dataset but should be calculated. For example, if the hospital discharge dataset contains date of admission, date of discharge, and date of birth, but it does not include length of stay, then you need to calculate that variable. You would list the name of the variable (“HospDays”), type of variable (integer), its explanation (“number of days in hospital”), and the fact that it is a calculated variable (calculated: HospDisch – HospAdmit +1).
* The key outcome variables should also be flagged or listed.

### Statistical methods and software

* There are different statistical methods you will use depending on the research questions.
* There are many quantitative statistical software packages, this course will use R.

**Case study: developing an analysis plan to estimate hypertension.**

The past few decades have brought a new global phenomenon called the “nutrition transition” in many low- and middle-income countries2. This transition includes a large shift from traditional diets and lifestyles to one increasingly composed of pre-packaged and processed foods along with sedentary lifestyles. Infectious diseases remain a critical public health priority for many parts of the world. However, many countries now confront a ‘double disease burden’, as rates of noncommunicable disease (NCDs), such as diabetes, cardiovascular disease, and cancers account for more than half of the global burden of disease in both developing and developed countries3.

NCDs are responsible for more than 60% of all deaths worldwide, with more than 80% of NCD-related deaths occurring in low- and middle-income countries (LMICs). The burden of NCDs is expected to grow as both the world population and the proportion of persons 60 years and older continue to increase; NCDs disproportionately affect this age group.

Cardiovascular diseases (CVD) are the single largest cause of mortality worldwide. This represents nearly 30% of all deaths and about 50% of all NCDs. Common behavioural risk factors, including tobacco use, physical inactivity, unhealthy diet, and the harmful use of alcohol, are responsible for approximately 80% of the global CVD burden4.

Raised blood pressure, or hypertension, is the leading risk factor for mortality5. Hypertension has decreased significantly in nearly all high-income countries due to widespread diagnosis and access to low-cost medications. In contrast, mean blood pressure has been stable or increasing in most African countries; approximately 40% of adults in many of these countries are estimated to have high blood pressure. Most of these people remain undiagnosed, although many could be treated with low-cost medications; this would significantly reduce the risk of death and disability from heart disease and stroke3.

Effective prevention strategies for NCDs, and specifically for reducing the burden of hypertension and CVDs exist, including reduced sodium consumption and reduced tobacco use. Effective strategies, however, require specific data on risk factors to set priorities and develop and monitor interventions.

## **How One Country is Addressing the Burden of CVD**

Country X is a rapidly modernising nation of 10 million people, with a growing middle-class. Vital statistics indicate a gradually increasing overall life expectancy but a surprisingly high rate of deaths due to CVD, specifically stroke. Health officers from two southern provinces report a steady increase in the use of renal dialysis services. This is placing serious constraints on the regional health service budgets. The traditional diet of the country is generally rich in fresh vegetables, whole grains, and healthy oils; however, the younger and middle-aged segments of the population have rapidly increased their consumption of meals out of the home, which are typically higher in fat, salt, and processed foods. There has been a tremendous expansion of restaurants that serve meals not typical of the traditional diet in the country. The National Health Service can only provide limited specialty care services; however, it is able to provide affordable primary care and basic medications to citizens based on family income.

A recently assembled panel of health care leaders in the country gave recommendations to the newly appointed Minister of Health (MoH). Their report highlighted concerns CVD, stroke, chronic renal disease, and other NCD risk factors such as physical inactivity and changing diet. However, the report cited concerns about the lack of basic information on hypertension, a potential factor underlying these other conditions.

You are the leader of the MoH’s Chronic Disease Surveillance Unit. The MoH has asked you to analyse national health survey data on hypertension, which has been collected every two years over the past decade, and other provincial level hospital data and report on the findings. The most recent data were collected last year. The MoH wants to provide the report to the national and provincial decision-makers so they can better understand the magnitude of the burden of disease of hypertension and the key determinants and underlying factors that are affecting this public health burden. With this information, the MoH is hoping to target resources and support evidence-based actions and policies to improve the health of the population.

## **In small groups or pairs spend 30 minutes discussing ideas for your analysis plan. Fill in the following sections:**

**Research question and / or hypotheses:**

**Dataset(s) to be used:**

**Inclusion criteria:**

**Exclusion criteria:**

**Outcome variables:**

**Exposure variables:**

**Data analysis:**

* Descriptive analysis – which population characteristics should be summarised?
* Inferential analysis - what statistical tests and methods might be useful? (Do not worry if you do not know many yet, you will by the end of the course!)

## References

1. Baden LR, El Sahly HM, Essink B, et al. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med* 2021; **384**(5): 403-16.

2. Popkin BM. The shift in stages of the nutrition transition in the developing world differs from past experiences! *Public Health Nutr* 2002; **5**(1A): 205-14.

3. Yach D, Hawkes C, Gould CL, Hofman KJ. The global burden of chronic diseases: overcoming impediments to prevention and control. *JAMA* 2004; **291**(21): 2616-22.

4. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. *Curr Probl Cardiol* 2010; **35**(2): 72-115.

5. Asaria P, Chisholm D, Mathers C, Ezzati M, Beaglehole R. Chronic disease prevention: health effects and financial costs of strategies to reduce salt intake and control tobacco use. *Lancet* 2007; **370**(9604): 2044-53.