Problem Set 1: Sampling from a population, summarizing data, the normal distribution, t-distribution, and hypothesis testing

Introduction to Statistical Thinking and Data Analysis

MSc in Epidemiology and MSc in Health Data Analytics, Imperial College London

10 October 2022

A) Consolidating concepts

A1. What does the standard deviation measure? variability from mean

a. The spread of the middle 50% of the distribution. nope

b. The amount of variability in the sample mean.

c. The amount of variability in the population. nope because i have a sample

d. None of the above. not sure

A2. Which of the following is true of the standard normal distribution?

a. It has a mean of 0 and a standard deviation of 1.

b. It has an area equal to 0.5. nope, equal to 1

c. It has a mean of 1 and a standard deviation of 0. no, option A

d. It cannot be used to approximate any normally distributed variable. no

A3. For a normally distributed dataset, we would expect approximately:

a. 70% of observations to lie within 1 standard deviation of the mean.

b. 1 out of 20 observations to lie outside of two standard deviations of the mean. 5%, yes
95%

c. A bell shaped distribution.

d. All of the above.

A4. Given that **X** is a normally distributed variable with a mean of 5 and a standard deviation of 1.2, what is the probability that X is less than 6?

a. 0.80

z = X-mean/sd = 6-5/1.2 = 1/1.2 = 0.83
P(X<6) = P(z < 0.83) = 0.7976716

b. 0.65

c. 0.90

d. 0.77

A5. What percentage of females are taller than 170cm in a population with a mean of 165cm and standard deviation of 5.5cm?

a. 18%

P(X > 170) = P(z > 170-165/5.5) = 1 - P(z <= 170-165/5.5) = 0.1816511
pnorm(170, 165, 5.5, lower.tail = FALSE)

b. 12%

c. 24%

d. None of the above.

A6. What is the Z-score of the 95th quantile of the standard normal distribution?

a. 1.64

z = qnorm(0.95) = 1.64

b. 1.96

c. 1.28

d. 0.83

A7. Which of the following is not true of the central limit theorem?

a. Provided the sample size is sufficiently large, the distribution of a sample mean is approximately normal. also true

b. The underlying distribution of the population must be normal.

c. The sample means will be normally distributed around the population mean. also true

d. The more you run a random experiment, the more its results will follow a normal distribution. also true

A8. The normal distribution is also called:

a. Poisson distribution

b. Bernoulli's distribution

c. Gaussian distribution

d. Student's t

bell shaped

A9. The shape of the t distribution:

a. Is the same as the normal distribution.

b. Is skewed.

c. Depends on the number of degrees of freedom.

d. None of the above.

A10. A 95% confidence interval for the mean of a population is such that:

a. The population mean will fall within the confidence interval 95% of the time.

b. If we sample the same population 100 times, 95 of the confidence intervals calculated from these random samples will contain the population mean.

expect, not will contain

c. ~~It contains 95% of the values in the population.~~

d. There is a 95% probability that it contains the population mean.

B) Practicing skills

B1. The dataset `perulung_ems.csv` contains data from a study of lung function among a sample of 636 children aged 7 to 10 years living in a deprived suburb of Lima, Peru, introduced on page 27 of Kirkwood and Sterne. FEV1 is the *forced expiratory volume* in 1 second, the maximum amount of air which children could breath out in 1 second measured using a spirometer.

Variable	Description
id	Participant ID number
fev1	Forced Expiratory Volume in 1 second
age	Age in years
height	Height in centimeters
sex	Sex (0 = female, 1 = male)
respsymptoms	Presence of respiratory symptoms (0 = no symptoms; 1 = symptoms)

a. What type of variable is each variable in the dataset?

b. What shape is the frequency distribution of each of the variables?

c. What are some research questions which these data could have been collected to address?

d. Use R to create appropriate univariate graphical summaries of each of the variables.

e. Create a single table summarizing key characteristics of the sample—an appropriate 'Table 1' for a medical or epidemiologic paper. (It is probably possible to construct a full table with R commands, but you might find it easier to do calculations of summary statistics with R and copy the R output into a separate table in MS Word, Excel, or similar.)

f. In this sample of 636 children, does there appear to be an association between:

(i) sex and height,

(ii) age and height,

(iii) sex and lung function

(iv) sex and presence of respiratory symptoms,

(v) respiratory symptoms and lung function.

Support your answers with graphical or numerical evidence.

g. What is the target population to which your conclusions about these questions might generalize?

h. Calculate an estimate and 95% confidence interval for each of the following quantities:

(i) Height in the whole population.
(ii) FEV1 in the population.
(iii) Height for male and females children separately.
(iv) FEV1 for children with respiratory symptoms and those without respiratory symptoms.

For each of these outcomes calculate the large sample CI using the normal distribution and the small-sample CI using the t-distribution. Compare your calculations to the 95% CI estimate output by the `t.test()` function.

i. For each of the following research questions:

- articulate an appropriate null and alternative hypothesis,
- calculate an appropriate estimate and uncertainty range,
- determine an appropriate statistical test for your hypothesis, and
- report the results of your hypothesis test to answer the question.

(i) Is the average height of children aged 7 to 10 years in Lima greater than 124cm?
(ii) Is the average height of girls in Lima equal to 123.5cm?
(iii) Is there an association between sex and height among children in Lima?
(iv) Do children with respiratory symptoms have reduced pulmonary function compared to children with no respiratory symptoms?

j. If the null hypothesis were true for all of the above 4 questions, what is the probability of erroneously rejecting at least one null hypothesis and incorrectly concluding an association exists?

B2. The *National Health and Nutrition Examination Survey* (NHANES) is a nationally representative survey to assess the health and nutrition of adults and children in the United States. The survey was first conducted in the 1960s and has been conducted continuously since 1999 with around 5000 respondents sampled and interviewed in their homes every year. The survey consists of a combination of questionnaire responses and physical and biomarker measurements. Further information about the survey and datasets can be found here: https://www.cdc.gov/nchs/nhanes/index.htm.

The R package `NHANES` contains an extract of 75 variables about 10,000 respondents to NHANES between 2009 and 2012 abstracted for educational purposes. The actual NHANES survey datasets include sampling weights to account for non-equal sampling probability of certain population groups to increase the statistical efficiency of the survey, which is not covered in this course. The dataset of 10,000 respondents in the `NHANES` R package has been constructed such that the dataset can analysed as if it were a simple random sample from the American population. See the package documentation for information and links about accessing and analysing the actual NHANES data for research purposes; there are other R packages available on CRAN to assist with accessing and processing the actual NHANES survey data.

The objective of this exercise is to practice loading large datasets into R, understanding the structure and variables in a dataset, and conducting exploratory analysis.

a. Open and explore the NHANES dataset in R through the following steps:

- Install the `NHANES` R package: `install.packages("NHANES")`
- Load the R package: `library(NHANES)`
- Load the NHANES dataset into your workspace: `data(NHANES)`
  - Use the command `?` to access the help page for the dataset: `?NHANES`
    Review the dataset edocumentation, particularly looking

(i) What was the purpose for collecting the data?
(ii) When and how were the data in the dataset collected?
(iii) What is the target population of the sample?
(iv) What is the sample size? Who was eligible to be included in the dataset? Are there different eligibility or inclusion criteria for certain variables?
(v) What are the areas of information available in the dataset?

In the documentation, note that several of the variables are only collected for respondents of a certain age range or in one of the survey rounds but not the other. This is important to take note because it may affect what questions can be addressed by the data, or result in errant conclusions and incorrect interpretation if eligibility and inclusion criteria are not appropriately considered during analysis. For the remainder of the tutorial, we will only consider the subset of the sample who are adults aged 20 years and older.

Construct this dataset with the R command: `nhanes20pl <- NHANES[NHANES$Age >= 20, ]`

(vi) Confirm that your new dataset has 7235 respondents remaining. Use at least one different R command to achieve construction of the same subsetted dataset. Confirm that your alternate command has the same number of rows and columns.

b. *Types of variables.* Using the subsetted `nhanes20pl` dataset, identify at least one variable of each of the types of variables: continuous, discrete numeric, binary, categorical, and ordered categorical. For an identified variable of each type, create an appropriate summary of the frequency distribution and calculate an appropriate measure of central tendency and variation.
c. *Frequency distributions.* Identify at least one variable in the dataset that has a frequency distribution matching each of the shapes described in Kirkman and Sterne Figures 3.5 and 3.6 (pages 20-21). For each of the identified variables, calculated the mean, median, mode, variance, standard devation, range, and interquartile range.

B3. Statistical modelling: using the normal distribution to estimate population distributions. The purpose of this exercise is to practice applying the normal distribution to estimate the distribution of an outcome in a population using data from a sample drawn from the population. The version of the NHANES dataset we are using is constructed to approximate a simple random sample from the American population. Thus we can use the observations in our sample to make inference about the American population. For each of the questions below, consider three continuous variables:

- `Height`: Standing height in centimeters.
- `BMI`: Body Mass Index
- `AlcoholYear`: Number of days over the past year that participant drank alcoholic beverages.

a. Create a graph of the frequency distribution for each of the variables, either using histograms or density plots. Does each outcome appear consistent with arising from a normal distribution?

b. Calculate and report the sample mean and standard deviation for each of the outcomes.

c. Use the normal distribution as a model to estimate the following:

(i) The proportion of adults who are above 165cm tall.
(ii) The proportion of adults between 153cm and 160cm tall.
(iii) The height of a door in order that 90% of adults can walk under without ducking.
(iv) The proportion of adults who are obese, defined as BMI above 30.
(v) The proportion of adults who are overweight, defined as BMI between 25 and 30.
(vi) The BMI threshold at which 25% of adults are below.
(vii) The proportion adults who drink alcohol on more than 100 days per year.
(viii) The proportion of adults who drink alcohol on fewer than 10 days per year.
(ix) The interquartile range for the number of days per year that American adults drink alcohol. *Hint: use the R functions* `pnorm(...)` *and* `qnorm(...)`.

d. Check your estimates based on the normal distribution by directly calculating each of the above proportions among the observed sample. Based on this comparison, do you think that the normal distribution is a good model for the population distribution of each outcome in the population?

C) Advanced learning

C1. *Missing data.* Using the `nhanes20pl` dataset, review the amount of missing data for each of the variables pertaining to *Demographic Variables* and *Physical Measurements (Hint: the R function 'is.na' may be useful).* (i) Amongst each grouping (*Demographic Variables* and *Physical Measurements*) identify the one variable with the highest proportion of missing cases. (Do not consider variables that were available for only one of the survey rounds or not recorded for this age range when making your assessment.) (ii) For each of these two variables, in this sample, is there any relationship between `Gender` or `Race1` and the probability that data on the outcome is missing?