IEHC0046 BASIC STATISTICS FOR MEDICAL SCIENCES

Analysis of Continuous Data I: Practical

10 October, 2020

In this practical session we will learn how to use R to explore means of continuous variables and conduct one sample t-tests. We will use the ELSA dataset that you have been using throughout the course so far.

Remember to use a script to save your code and to change your working directory so you can load the ELSA dataset easily.

```
load("elsa.Rdata")
```

1. Exploring continuous variables

1.1. Calculating means and 95% confidence intervals (CIs)

In this first section we will explore some of the continuous variables in our dataset by calculating their range, mean and 95% CIs. There are many ways to obtain the mean of a continuous variable in R, but here we will use the function `t.test()`, which is from Base R. We'll also use the `descr()` function from the package `summarytools` to obtain descriptive statistics for continuous variables, so let's install and load that package.

```
install.packages("summarytools")
library(summarytools)
```

To obtain the mean and 95% CIs for a variables, use the syntax `t.test(x)` where `x` is a vector. To obtain descriptive statistics, use `descr(x)`.

Q. Describe the distribution of BMI (variable name `bmi`) in our sample based on its minimum and maximum values, mean and 95% CIs.

To get descriptive statistics, we use the `descr()` function.

```
descr(elsa$bmi)
```

```
## Descriptive Statistics
## elsa$bmi
## Label: BMI (kg/m2)
## N: 3129
##
##                         bmi
## ----------------- ---------
##              Mean     27.53
##           Std.Dev      4.49
##               Min     14.81
##                Q1     24.48
##            Median     27.03
##                Q3     29.95
##               Max     52.47
##               MAD      4.01
##               IQR      5.47
```

```
##                CV      0.16
##          Skewness      0.83
##       SE.Skewness      0.05
##          Kurtosis      1.43
##           N.Valid   2930.00
##         Pct.Valid     93.64
```

The output tells us that the are 2, 930 participants will observed BMI in our sample (`N.valid`). The mean BMI is 27.5 kg/m^2 (standard deviation = 4.5 kg/m^2). We see that the minimum BMI is 14.81 kg/m^2 and the maximum is 52.47 kg/m^2.

To get the mean and 95% CIs, we use the `t.test()` function.

```
t.test(elsa$bmi)
```

```
##
##  One Sample t-test
##
## data:  elsa$bmi
## t = 331.71, df = 2929, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  27.36890 27.69438
## sample estimates:
## mean of x
##  27.53164
```

The output tells us that the mean `bmi` is 27.53 kg/m^2, with a 95% CI confidence interval for 27.37 - 29.69 kg/m^2.

Q. Now let's do the same for diastolic blood pressure (variable name `dbp`).

```
descr(elsa$dbp)
```

```
## Descriptive Statistics
## elsa$dbp
## Label: Diastolic BP in mmHg
## N: 3129
##
##                         dbp
## ----------------- ---------
##              Mean     77.32
##           Std.Dev     11.90
##               Min     45.00
##                Q1     69.00
##            Median     77.00
##                Q3     84.50
##               Max    133.50
##               MAD     11.12
##               IQR     15.50
##                CV      0.15
##          Skewness      0.41
##       SE.Skewness      0.05
##          Kurtosis      0.83
##           N.Valid   2692.00
##         Pct.Valid     86.03
```

There are 2,692 participants with an observed DBP value in our sample. The mean DBP is 77.3 mmHg

(standard deviation = 11.9). The minimum DBP is 45 mmHg and maximum is 133.5 mmHg.

```
t.test(elsa$dbp)
```

```
##
##  One Sample t-test
##
## data:  elsa$dbp
## t = 337.22, df = 2691, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  76.87486 77.77409
## sample estimates:
## mean of x
##  77.32448
```

The 95% CIs ranges from 76.9 mmHg to 77.8 mmHg, showing 95% coverage of the population mean DBP that are compatible with the study data.

1.2. Calculating means and 95% CIs for sub-groups

Sometimes we might want to show the mean values of a continuous variable for different groups of our population. For example, showing the mean BMI of men and women separately. We can do this in R in multiple ways. Here, we'll show two approaches: subsetting and the `by()` function. If you need a recap on subsetting, please refer back to the notes from the Introduction to R session.

Q: Show the mean BMI and 95% CIs for men and women in this sample separately. What do you conclude?

First we need to check how `sex` - the variables containing participant gender - is coded.

```
str(elsa$sex)
```

```
##  Factor w/ 2 levels "male","female": 2 1 2 1 1 1 2 1 1 2 ...
##  - attr(*, "label")= chr "Sex"
```

```
levels(elsa$sex)
```

```
## [1] "male"   "female"
```

`sex` is a factor with two levels, "`male`" and "`female`".

To get means and 95% CIs for males and females separately, we can subset the vector we pass to `t.test()`.

```
t.test(elsa$bmi[elsa$sex == "male"])
```

```
##
##  One Sample t-test
##
## data:  elsa$bmi[elsa$sex == "male"]
## t = 257.05, df = 1291, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  27.27892 27.69850
## sample estimates:
## mean of x
##  27.48871
```

```
t.test(elsa$bmi[elsa$sex == "female"])
```

```
##
```

```
##  One Sample t-test
##
## data:  elsa$bmi[elsa$sex == "female"]
## t = 225.58, df = 1637, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  27.32582 27.80518
## sample estimates:
## mean of x
##   27.5655
```

The mean BMI of men is 27.5kg/m^2 and the 95% CIs range from 27.3kg/m^2 – 27.7kg/m^2. The mean BMI for women is 27.6kg/m^2 and the 95% CI is 27.3kg/m^2- 27.8kg/m^2. BMIs are very similar for men and women in our sample.

Alternatively, we could have used the `by()` function to get the same answer. The `by()` function takes the three (or more) inputs. The first argument is the vector we wish to perform an operation on (in this case `bmi`). The second is the vector which contains the groupings we wish to do an operation *by* (in this case `sex` because we want to do a `t.test()` for males and females separately). The third argument is the function that does the operation we want to do (in this case `t.test()`). (Further arguments to `by()` are passed to the operating function.)

```
by(elsa$bmi, elsa$sex, t.test)
```

```
## elsa$sex: male
##
##  One Sample t-test
##
## data:  dd[x, ]
## t = 257.05, df = 1291, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  27.27892 27.69850
## sample estimates:
## mean of x
##  27.48871
##
## ------------------------------------------------------------
## elsa$sex: female
##
##  One Sample t-test
##
## data:  dd[x, ]
## t = 225.58, df = 1637, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  27.32582 27.80518
## sample estimates:
## mean of x
##   27.5655
```

We can use these methods to explore the means of any number of different sub-groups in our sample in a descriptive manner. In the above exercise we just reported the means and 95% CIs for two sub-groups (men and women), but we could have looked at differences by a variable with more than two categories, such as physical activity, ethnicity or social class.

Q. Report the means and 95% CIs for BMI for the different physical activity categories (vari-

able physact). What do you conclude?

Again we should first check the coding of the physical activity variable (physact).

```
str(elsa$physact)
```

```
##  Factor w/ 3 levels "Group 1 -low",..: 1 1 1 1 1 2 2 2 2 2 ...
##  - attr(*, "label")= chr "3 categories of physical activity (1=low, 3=high)"
```

```
levels(elsa$physact)
```

```
## [1] "Group 1 -low"     "Group 2 - medium" "Group 3 - high"
```

There are three categories of physical activity. 1 = low, 2 = medium and 3 = high. To save time, we'll use the by() function to get the mean BMI (+ 95% CIs) for the three groups.

```
by(elsa$bmi, elsa$physact, t.test)
```

```
## elsa$physact: Group 1 -low
##
##  One Sample t-test
##
## data:  dd[x, ]
## t = 204.44, df = 1214, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  28.06098 28.60477
## sample estimates:
## mean of x
##  28.33287
##
## ------------------------------------------------------------
## elsa$physact: Group 2 - medium
##
##  One Sample t-test
##
## data:  dd[x, ]
## t = 203.56, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  26.99676 27.52232
## sample estimates:
## mean of x
##  27.25954
##
## ------------------------------------------------------------
## elsa$physact: Group 3 - high
##
##  One Sample t-test
##
## data:  dd[x, ]
## t = 178.49, df = 713, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  26.26927 26.85359
## sample estimates:
## mean of x
```

```
##  26.56143
```

The mean BMI of people with low physical activity is 28.3kg/m^2 (95% CI: 28.1, 28.6). For people reporting medium physical activity it is 26.6kg/m^2 (95% CI: 27.0, 27.5). For people reporting high physical activity is 26.6kg/m^2 (95% CI: 26.3, 26.9).

Mean BMI is lower for individuals reporting greater physical activity. You will learn how to formally test such differences in a later practical.

2. One-sample t-tests

2.1. Comparing means in our sample to population means

We can use a one sample t-test when we want to compare the estimates from our sample with those from the general population. To do this in we use the syntax `t.test(x, mu = <pop_mean>)` where `x` is a vector of values and the argument `mu` is a values to compare the mean of `x` against.

Q. We know from another study that the mean BMI of adults aged 60+ in the whole population is 28.5kg/m^2. Are adults aged 60+ different in the ELSA sample different from the general population in terms of BMI?

We need to subset the `bmi` variable to only select individuals who are aged 60 or above (the relevant variables is `age`). We also need to add 28.5 to the argument `mu`.

```
t.test(elsa$bmi[elsa$age >= 60], mu = 28.5)
```

```
##
##  One Sample t-test
##
## data:  elsa$bmi[elsa$age >= 60]
## t = -7.5821, df = 1408, p-value = 6.143e-14
## alternative hypothesis: true mean is not equal to 28.5
## 95 percent confidence interval:
##  27.41753 27.86252
## sample estimates:
## mean of x
##  27.64003
```

The mean BMI of these adults aged 60+ was 27.6kg/m^2 (95% CI: 27.4, 27.9). The output also reports that we have degrees of freedom 1,408 (`df`). In a one-sample t test, degrees of freedom is equal to observations minus 1. So there are 1,409 individuals aged 60+ in our sample who have observed BMI.

The output gives a t statistic (-7.58) and a p-value (6.1e-14) for the hypothesis that the true population mean of sample from which ELSA participants were drawn is equal to 28.5. (The p-value is presented in scientific notation - 6.1e-14 = 6.1 x 10^{-14} = 0.000000000000061 - in other words, a very small number).

The small p-val suggests that it is unlikely that we would have obtained a sample mean of 27.6 kg/m^2 if the true population mean was 28.5 kg/m^2. Our data therefore have low compatibility with the null hypothesis. We can conclude that our sample mean BMI is different from the general population mean.

Q. What about if the general population mean BMI was 27.5 kg/m^2? Do our conclusions change? Let's amend the code from the previous question.

```
t.test(elsa$bmi[elsa$age >= 60], mu = 27.5)
```

```
##
##  One Sample t-test
##
```

```
## data:  elsa$bmi[elsa$age >= 60]
## t = 1.2346, df = 1408, p-value = 0.2172
## alternative hypothesis: true mean is not equal to 27.5
## 95 percent confidence interval:
##  27.41753 27.86252
## sample estimates:
## mean of x
##  27.64003
```

In contrast to the previous results, the t statistic is low (t = 1.23) and the p values is larger than typical statistical significance levels. These findings suggest that our reasonably high compatibility with the null hypothesis that the mean BMI in the population ELSA was drawn from is the same as the general population mean.

2.2 One sample t-tests for sub-groups

We can also conduct t-tests for specific sub-groups in our sample. For instance, we might want to know whether the mean BMI of people aged 60+ is higher in those who report low physical activity compared to the whole sample.

Q: Find out whether the mean BMI of people aged 60+ who report low physical activity is higher than the mean for the whole sample who are aged 60+. You can identify those who reported low physical activity using the variable `physact`. You also need to first calculate the mean for those aged 60+.

First let's calculate the mean for people aged 60+. We can just use the `mean()` function, setting `na.rm` to `TRUE` to ignore missing values. Let's also store this result so we can use it directly in the second step.

```
m_60plus <- mean(elsa$bmi[elsa$age >= 60], na.rm = TRUE)
m_60plus
```

```
## [1] 27.64003
```

The mean BMI of adults aged 60+ in our sample is 27.6 kg/m^2 .

Now let's use this result in a one sample t-test to assess whether the mean BMIs of people who report low physical activity is different from the mean BMI for the whole sample.

Recall that the low activity group were the first category in the `physact` variable. (We can subset the results of the `levels()` function directly).

```
t.test(elsa$bmi[elsa$age >= 60 &
                  elsa$physact == levels(elsa$physact)[1]],
       mu = m_60plus)
```

```
##
##  One Sample t-test
##
## data:  elsa$bmi[elsa$age >= 60 & elsa$physact == levels(elsa$physact)[1]]
## t = 2.9943, df = 692, p-value = 0.002849
## alternative hypothesis: true mean is not equal to 27.64003
## 95 percent confidence interval:
##  27.81704 28.49133
## sample estimates:
## mean of x
##  28.15418
```

The output shows that the mean BMI for people aged 60+ who reported low physical activity is 28.2kg/m^2 (95% CI: 27.8, 28.5). The t statistic is 3.2 and the p value is < 0.001. There is low compatibility with the

null hypothesis: the results suggest that, among participants aged 60+, mean BMI for those reporting low physical activity is higher than the whole sample.

Note, the above code is hard to read and can be made clearer using `tidyverse` functions.

```
library(tidyverse)
elsa %>%
  filter(age >= 60, # Keeps the correct observations
         physact == levels(physact)[1]) %>%
  pull(bmi) %>% # Extracts the bmi variable
  t.test(mu = m_60plus)
```

```
##
##  One Sample t-test
##
## data:  .
## t = 2.9943, df = 692, p-value = 0.002849
## alternative hypothesis: true mean is not equal to 27.64003
## 95 percent confidence interval:
##  27.81704 28.49133
## sample estimates:
## mean of x
##  28.15418
```

Formative Exercise

If you would like to look at more examples of conducting one sample t-tests you might like to work through the optional exercise below.

Q: Do participants in non-manual social classes aged 60+ have a different mean BMI to the whole sample aged 60+?

Note you will need to create a new binary social class variable combining social classes I, II, IIIN as 'non-manual' and social classes IIIM, IV, V as 'manual'. (Hint: use `fct_collapse()` from the package `forcats` which is loaded with the `tidyverse` - see (**this website**)[https://forcats.tidyverse.org/reference/fct_collapse.html].)

Q: Do participants in non-manual social classes aged 60+ have a different mean DBP to the whole sample aged 60+?