Exercise Day 3 - AMH

Alessandra Meddis

Birth control pills and Anti-Mullerian hormone

In a clinical study researches wished to investigate the degree of which taking birth control pills changed the level of Anti-Müllerian hormone (AMH) in the the human body. AMH is expressed by granulosa cells of the ovary during the reproductive years, and limits the formation of primary follicles by inhibiting excessive follicular recruitment by follicle-stimulating hormone. It is thus expected that women taking birth control pills will show a significantly lower level of AMH on average. This is an observational study i.e., the participants were not randomized to either birth control or no birth control, but a random sample of 732 women were included in the study.

In this exercise we would like to create some descriptive analysis to check the researcher hypothesis **taking birth control pills lowers the level of AMH on average**.

The data set contains the following variables:

- *amh* : anti-mullerian hormone level
- *pill*: usage of birth control pills (no/yes)
- *age*: age of woman in years
- *BMI* : BMI of women
- *smoking* : if a smoker (1: yes, 0: no)

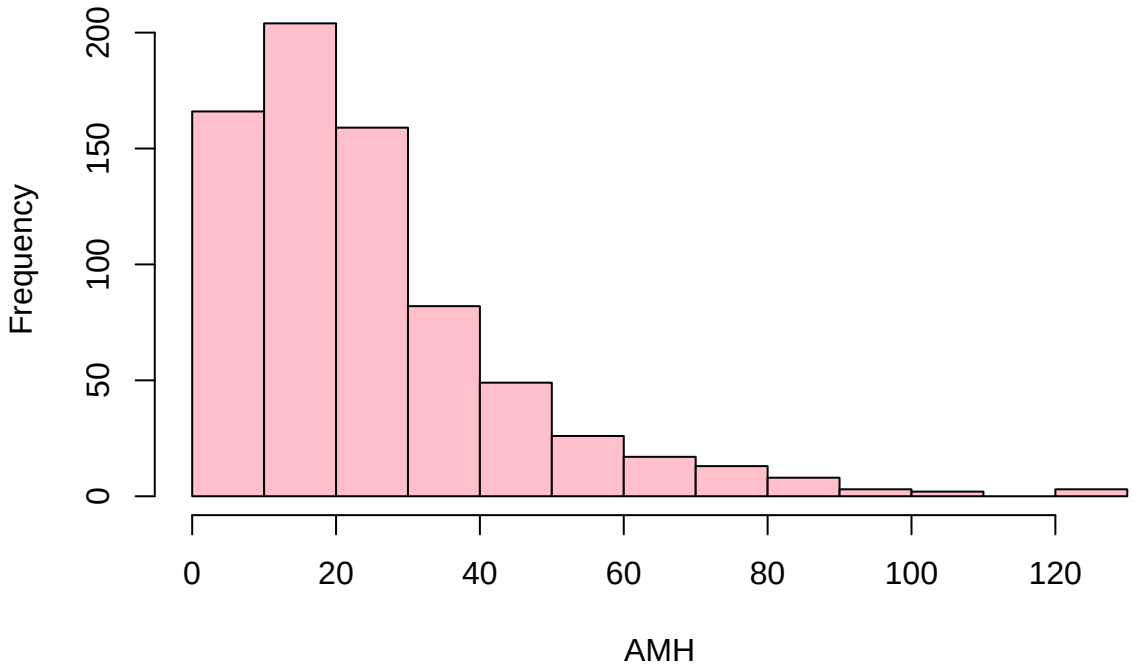
You can find the data here: https://raw.githubusercontent.com/AMeddis/IntrotoR-for-Basic-Statistics/refs/heads/main/data_exercise/amh.csv

Question 0. First look ath the data

0. Print a summary of the data and check if you have to transform some variables.
1. We want to study the AMH level by contraceptive use, it would be interesting knowing the proportion of contraceptive users.
2. We might want to understand if users and non users are similar respect to some characteristics. Calculate the descriptive statistics you believe are relevant, for this research study, such as median of age and BMI by contraceptive use.

Question 1. Distribution of AMH

We start looking at the distribution of the AMH.



The histogram shows an heavy tail, so we might prefer to use the logarithmic scale of AMH for the analysis.

1a. Add a new variable with logarithmic transformation for the AMH.

1b. Create the histogram for the log-AMH.

Can you see a difference in the distribution?

As a Note: In most statistical analyses, conclusions about continuous outcomes are often based on the mean. However, when data are skewed or have a heavy tail(like for the AMH), the mean can be misleading because it is highly sensitive to extreme values (outliers). In such cases, it is more appropriate to focus on the median, or to apply a transformation (e.g., a log transformation) that reduces the impact of the tail.

Question 2. Descriptive analysis: AMH by contraceptive use

We want to understand whether taking birth control has an impact on the AMH level. What would you calculate to check whether the AMH level either increase or decrease by use of contraceptive? (use the log-transformed variable)

1. Calculate the statistic you had in mind.
2. Create a plot (choose the one you prefer) to show the distribution of AMH level by contraceptive use.

What would be your conclusion on the hypothesis?

Question 3. AMH and age

This is an observational study, no randomization on the contraceptive use. Check the results on the median age among users and non-users(Question 0). What can you notice?

1. Create a plot (choose the one you think it is most relevant) to show the distribution of age by contraceptive use.
2. Create a scatterplot where the y-axis is the AMH level (log scale) and the x-axis is age. Add the regression line, namely the fitted line of the average AMH level by age (in year). You can add the line using the command **abline(lm(y-Variable ~ x-Variable, data), col="Choose a color")** (see slides).

What does this show?

4. Would you change the analysis? Would you consider other variables when studying if AMH level is decreasing with contraceptive use?

Question 4. Descriptive analysis: AMH and contraceptive use by age

So far, we believe that age might be relevant when assessing if AMH level is decreasing with contraceptive use. We can consider age as a categorical variable and have a stratified analysis, namely we re-run the descriptive analysis in each age-group.

1. Create a new variable of age group. You can choose the number of groups to use and the cut-off value(s).
2. Calculate the descriptive statistic (the same chosen in Question 1) by age group. What would be your conclusion now?
3. Create a plot you believe it would best explain the results. (You can choose to use either the continuous or categorical version of age)