## Simulated Outbreak Exercise

|  |  |
| --- | --- |
| Scenario Date: | Thursday 12 October (09:15-10:30) |
| Inject No: | 14 |
| Inject time: | 1 hour 15 minutes |
| Inject Title: | Univariable analysis using R |

1. **Learning outcomes:**At the end of the session, participants will be able to:

* Perform hypothesis tests.
* Estimate risk ratios (also called relative risk) for categorical data.
* Interpret the univariable results.
* (Optional) Investigate dose-response relationships in categorical data.

1. **Story/plot description:**You have just estimated risk ratios (RR) of individual food items manually to better understand the concept. RR give you an idea of which food items could be the culprit(s) of the outbreak.

Now, you will practice performing hypothesis testing and calculating RR in R to investigate the associations of suspicious food items with the disease.

Note that, although this is not an exhaustive list of tests, the following can be useful for this exercise (relevant for comparisons between two groups):

* **For continuous variables:**
  + shapiro.test() (for checking if normally distributed)
  + t.test() (for normal distributions)
  + wilcox.test() (for non-parametric testing. Used to determine if two numeric samples are from the same distribution, when their populations are not normally distributed or have unequal variance.)
* **For categorical variables:**
  + chisq.test() (if all comparisons include at least 5 cases)
  + fisher.test() (if there are < 5 cases for any comparison)

1. **Questions/assignments for the group:**

3.1. Install packages (if needed) and load libraries.

3.2. Import your data

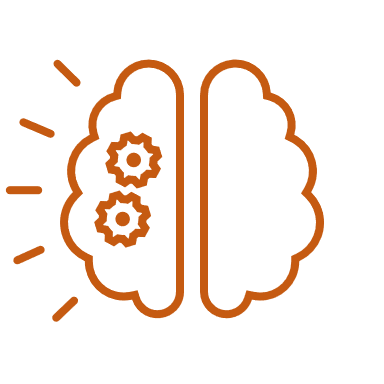
3.3. Check if the following variables are associated with being a case:

a) age (as continuous variable)

Facilitator’s note: With the Shapiro-Wilk test we check if the variables are following the normal distribution. The null hypothesis is that the data follow a normal distribution, therefore, rejecting the null hypothesis means that the data do not follow the normal distribution. A p-value below the cutoff for rejecting the null hypothesis, e.g., a p-value<0.05 means that we reject the null hypothesis that the data follow the normal distribution.

For age, the p-value is <0.05, therefore we reject the null hypothesis that the data are normally distributed. As we see in the graph most frequently reported age is <20 years.

b) sex



**Stop and reflect**:

* What do these results tell you - is there an association between sex and being a case?
* Do these results differ from what you expected when looking at the descriptive figures (case proportions stratified by sex, or the age sex pyramid)? If so, why do you think this is?

Facilitator’s note: Both variables are categorical. A chi-squared test can be used when you want to test if there is a relationship between two categorical variables (and there are no cells with a number lower than 5 – in this case you would use a Fisher’s text). P-value of the association between sex and being a case is p-val = 0.3, suggesting that there is no association at the p-val = 0.20 level.

Results are expected. At this point, we are suspecting a contaminated food item to be the culprit of the outbreak. We would think that during a dinner at a high school there will be no difference in the menu between males and females: they would all, more or less, eat the same things. And we also don’t expect that whichever pathogen that was in the contaminated food item would affect more any specific gender.

c) class

Facilitator’s note: In this case you have a 2x3 contingency table of class against disease status, we can use a chi-square here. P-value of the association between class and age is p-val = 0.09, suggesting that there is an association at the p-val = 0.20 level among at least one of the 3 different classes. Fellows may want to investigate this further, but in the remaining of the study this is not explored further.

A hypothesis could be that one of the classes sat together and, for whatever reason, they ate more of the contaminated food item at those tables. This could be further studied if we had the spatial distribution of where the people sat at the dinner. Another hypothesis is that one of the classes differ from other classes in some characteristics that made them more susceptible or more exposed to the infected food item.

d) group

Facilitator’s note: P-value of the association between group and age is p-val = 0.2, suggesting that there may be an association at the p-val = 0.20 level. Explanations are like above, with variable class.

3.4. Risk Ratios for food items

Facilitator’s note: The risk ratios of each food item (including the 2x2 table) are reported in the R-solutions. The output of the CS() command is two tables (one with the 2x2 table and one with the risk difference, the risk ratio and the attributable fraction among exposed as well as the attributable fraction among the population (and the confidence intervals for all the estimates). The Chi-square and the p-value are also reported. In the second part, a table with all the food items is printed including attack rates for exposed and unexposed as well as risk ratios and the 95% confidence intervals (CI ll and CI ul, for the lower and upper interval) and p-values.

1. To see if food items (note these are categorical variables) are associated with being a case, calculate risk ratios and 95% confidence intervals for food items. You can calculate this individually for each food item (using CSTable()), or all at once (see hint below).



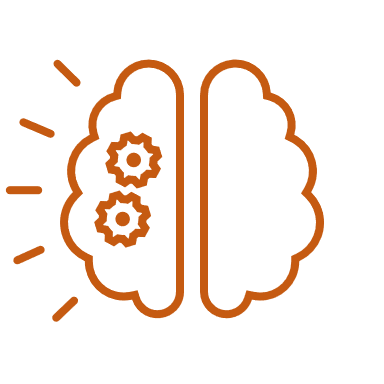
**Hint**: Create a food\_vars vector containing food variables of interest and use the function CSTable() of the EpiStats package.

*Using this hint, you will create a table with the name of exposure variables, the total number of exposed, the number of exposed cases, the attack rate among the exposed, the total number of unexposed, the number of unexposed cases, the attack rate among the unexposed, risk ratios, 95% percent confidence intervals, and p-values. Amazing, isn’t it? Have a look at ??CSTable to learn more.*

1. Prepare the above RR table for publication



**Hint**: Use flextable() and set\_header\_labels().



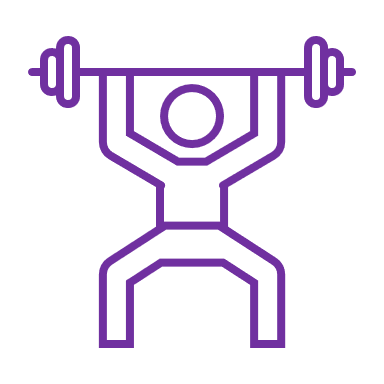
**Stop and reflect**:

* What can you infer from this table?
* Considering the relative risks, which food or drink items do you think were most likely to be the vehicle(s) of infection in this outbreak?
* Do you think there are any confounders or effect modifiers? If so, how would you investigate these further?

The interesting results here are that the food items that are most suspicious are pasta, veal and champagne. Pasta as such is unlikely to be contaminated, but as you can see in the picture (and from the dinner night), it was served with pesto! Maybe it was the pesto? Who-ho-ho!

Before one jumps into conclusions, consider that this result could be due to confounding! Maybe pasta was “clean” but eaten by all the people who ate the food item that actually was contaminated!

3.5. OPTIONAL Dose response.



**Extra**: Check for a dose-response relationship between the food items with the highest RR values (the top 3) and being a case.

**Stop and reflect (for dose-response)**:

* What do these results tell you?

Results suggest a dose response relationship of having eaten pasta, pointing towards pasta as the potential vehicle. The higher the amount of pasta they ate, the stronger is the association (RR) with getting ill/being a case.

3.6. As a summary of what you’ve done above, answer these questions:

* Is the respondents’ sex associated with being a case?
* Is the school class associated with being a case?
* Which foods increase the risk of being a case?
* (Optional) Is there a dose-response relationship between the food items and being a case?
* What do you think is the most likely culprit(s) of this outbreak at this point? Are there any risk factors you would like to highlight?

3.7 Draft the relevant paragraph(s) and table(s) and/or figure(s) in your outbreak report (Methods and results).

1. **Notes for facilitators:**

Note that fellows may still struggle with the R interface. Remind them that they always need to repeat the same steps by creating the workspace with access to the correct folders (R project) and that they must load the packages they will use and import the right dataset.

This session will be preceded by a plenary introduction to the most useful packages and functions for univariate analysis, so fellows should hopefully be able to progress.

1. **Scenario for the role play:**N/A
2. **Timing:**  
   1 hour 15 minutes
3. **Need for materials (logistics)**  
   N/A
4. **Deliverables**  
   A script with reproducible code and a paragraph in the Methods section summarizing the approach to the descriptive analysis, AND the completed section in the Results including tables, graphs and introductory text.