How People Learn

Exercises for Week 2 – solutions

Question 1

Why is it important to be able to define the different types of data?

The type of data visualization or statistical tests you will use will depend on the answer to three questions:

1. How many variables am I working with? (in all the exercises today you are working with a single variable)
2. What is the type of data?
3. Am I looking describe/visualize data or to make inferences to a population?

In this week's exercises you are always working with one variable at a time.

Nominal data is data where categories have a name but no order or hierarchy. Male/Female are categories with a name but no hierarchy. Faculty of origin (ENAC/STI/IC/SV/CDM/SB/CDH) are also categories without a hierarchy.

Ordinal data have categories and have an order. If age asked in groups (<19; 19-20; 21-22; 22>) then this could be regarded as ordinal because the data is in categories can be organized hierarchically. Likert scale questions which provide a statement and the respondent indicates if they Strongly Agree; Agree; Disagree; Strongly Disagree is also ordinal data. We assume that ordinal data does not have fixed distance between the categories – that is we do not know if Strongly Agree is twice as far from Disagree as it is from Agree.

Interval data has order, and a fixed distance between points. Age, if measured as a specific value (16; 17; 18; 19; 20; 21; 22 etc.) is interval type data. Score on a mathematics test is also interval type data. In both these cases it does make sense to say that the gap between 22 and 24 is the same as the gap between 26 and 28, i.e. that there is a fixed distance between two points.

Interval data often has a larger number of different 'categories' than other types of data and because there are fixed distances between points it can be treated as continuous data (i.e. it is logical to think of a class as having a mean average age of 22.333 or a mean average math score of 47.674). This is not true of nominal and ordinal data even if it has a large number of categories (it does not make logical sense of try to think of a group as being one third of the way between Strongly Agree and Agree).

Question 2 (a)

	Nominal	Ordinal	Interval
Frequency Table	✓	✓	✓
Bar chart	✓	✓	
Pie chart	✓		
Mode	✓	✓	✓
Median		✓	✓

Frequency tables are used when there are categories. For interval data it would generally be more useful to use some other way of representing the data such as a stem and leaf plot.

A bar chart represents data as discrete categories so is more appropriate for categorical data (nominal or ordinal). A histogram represents data as continuous so is more appropriate for interval data.

A Pie chart does not have two poles and so does not represent hierarchy (more or less) well. It works fine when there are (a few) categories, and no hierarchy. This suggest nominal data.

The mean, median and mode are three different measures of central tendency. As noted above, the mean average is suited to continuous data and so should not be used with nominal or ordinal data. The median (midpoint) makes sense to use if you can organize your data hierarchically and so works for ordinal and interval data but not for nominal data. The mode (most cited category) is the only option available for nominal data. It can be used for other types of data but generally doesn't provide much useful information.

Question 2 (b)

Table 1: Number of HPL class participants from different EPFL sections

Section	**Number**	**%**
Architecture	4	6.9
Chemistry & Chemical Engineering	1	1.7
Civil Engineering	3	5.2
Communication Systems	7	12.1
Computer Science	11	19.0
Digitial Humanities	1	1.7
Electrical Engineering	7	12.1
Financial Engineering	2	3.4
Life Sciences	3	5.2
Materials Science	5	8.6
Mathematics	3	5.2
Microengineering	3	5.2
Mechanical Engineering	4	6.9
Physics	4	6.9
Total	58	100

Note: 59 respondents, class total = 60; On-line survey data

Note: You will use tables to analyze data but also to present it to someone else. The point of this exercise is to make clear how tables should be presented in a report. All tables that you include in a report should include all of the stated information in order to make them as readable as possible by others. In this case I have:

- A table number (Table 1) so that I can refer back to it easily in the text
- A title that actually explains what the table shows
- Readable category names (Chemistry and Chemical Engineering rather than CGC)
- Clear column headings

In this case we have an n of 59. I know that there are 60 people registered but since you may not know that, you may not have included that information (which is fine). In this case there are no missing responses but there may be times that some people do not answer the question for one reason or another and in that case the number of missing responses should be in the footnote.

Question 3

Note: In Question 1 it was noted that you should ask yourself if you are describing data or making inferences. In question 2 you were describing data. In question 3 you are making inferences.

H_0: The percentage of EPFL masters students who get all three questions correct in Frederick's CRT is equal to 17%.

H_1: The percentage of EPFL masters students who get all three questions correct in Frederick's CRT is not equal to 17%.

In our study (in class) we had 41 out of 59 get all 3 correct. This is .69, with an N=59.

$$z = \frac{p - P}{\sigma_p}$$

$$Standard\ error\ of\ proportion = \ \sigma_p = \sqrt{\frac{P(1-P)}{N}}$$

We will reject H_0 if z is more extreme than or equal to plus or minus 1.96. (p=0.05).

P = Hypothesed population proportion of correct results (.17)

p = proportion of correct results found in sample (0.69)

N = 59

$$Standard\ error\ of\ proportion = \ \sigma_p = \sqrt{\frac{.17(.83)}{59}} = \sqrt{0.00239} = 0.0489$$

$$z = \frac{0.69 - 0.17}{0.0489} = 10.63$$

The chance of getting a proportion of .69 from a random sample of size 59 when the actual proportion in the population is .17 is infinitesimally small.

Interpretation:

Frederick found that 17% of a large sample of US university students got all three questions correct. In the EPFL sample the proportion was much higher: 69%. The data does not support the null hypothesis that the proportion of Master students in the EPFL student population who answer all three questions correctly is the same as that found by Frederick in the US (17%). In fact the evidence against the null hypothesis is very strong. An examination of the data suggests that the proportion of correct answers is higher in the EPFL cohort.

It should be noted that the hypothesis test is based on the assumption that the sample is a random sample. In this case a random method was not used to select the sample and there may be a bias built into the EPFL sample. For example, the EPFL data was collected from those taking a class on cognition and this bias may have affected the results from EPFL. The class chosen is oversubscribed and so students who are well organized in planning their class registrations are also likely to be over represented. Therefore, despite the very low p value, we should be very careful before generalizing from this to the cohort of EPFL Master's students as a whole. However, if we are more restrictive in how we define the population (e.g. if we treat this year's class as a random sample of all 'How People Learn students over the last six years), then our inference seems more reasonable.

(Note: You will always have to interpret results in text in your report. Your interpretations should balance (a) the fact that sampling is never according to the mathematical assumptions of "simple random sample with replacement" and (b) at the same time, sampling methods can be more or less random.)