Statistik mit R - Exercise Sheet 0

Anthony John Dsouza
7053485

Tyler Scott Lee
7054832

Yana Veitsman
7054842

1. **Voice onset time** is a **continuous** measurement because between any two measures of time there can be an intermediate value. It is also **ratio scale** because there is a natural zero when there is no onset time. Furthermore, multiplication is possible because it is sensical to say "the voice onset time of X constant is twice that of Y constant."

   The **number of words in a lexicon** is a **discrete measurement** because, for example, there is no possible value between 10000 words and 10001 words. It is **ratio scale** because it is comprehensible to say that there are zero words in a lexicon. Furthermore, you can say "this lexicon has twice as many words as that lexicon."

2. **A population could be published prose in American English, and the sample could be the Brown corpus**. This could be used to answer the question "What registers is word X used in?"

   Sampling in NLP uses various algorithms, including locally typical sampling, nucleus, and top-k sampling, the selection of which would depend on the final task to be performed (text classification, questions answering, text generation, etc.) and the objective of the investigation.

3. For the Brown corpus, texts for individual genres were randomly selected, but the genres included were manually selected. Therefore, if one wanted to study texts in an individual genre, one would have a random sample, but if one wanted to study all published prose, then the sample would not be random.

4. (a) **Independent variable: meal type (complete meal, vegetarian meal, free flow meal) – nominal scale**. Meal types are categories that cannot be ordered because, for example, there is no sense in which a complete meal is before or after a vegetarian meal.

   (b) **Dependent variable: nutritional value – ratio scale**. There is a natural zero. Furthermore, it can be multiplied or divided. For example, half of a 500-calorie meal would be 250 calories.

5. (a) **Independent variable: delay length – ratio scale**. There is a natural zero, i.e. no delay, and you can say that a 30-minute delay is twice the length of a 15-minute delay, for example.

   (b) **Dependent variable: satisfaction (measured with Likert scale) – ordinal or quasi-interval scale**. Technically, it is not clear that "disagree" is equidistant from "strongly disagree" and "neither agree nor disagree," but respondents tend to treat the five options as if that is the case (Navarro p. 18).

6. (a) **Independent variable: gender – nominal scale.**

(b) **Dependent variable: yes/no – ratio scale**. There could be zero yes-responses or zero no-responses, so there is a natural zero. Furthermore, multiplication and division are possible, because one could have twice as many yes-responses as no-responses, for example.

7. **Adaptive memory: Greater memory advantages in bilinguals' first language**

   (a) **Do bilinguals have better memory recall performance in the survival condition in their first language?**

   (b) **Bilinguals**

   (c) **127 Spanish-English bilinguals**. However, **the sample does not appear to be truly random**. Firstly, the bilinguals selected only represent the speakers of one of the language pairs and, secondly, all the participants are undergraduate students, whose background may give them an advantage at memory recall tasks.

   (d) **The bias in the study is related to the sample selection:** Spanish-English bilingual undergraduate students from a public research university are likely not representative of either all the bilinguals or even Spanish-English bilinguals. Undergraduate students come from a socio-economically privileged background and may have better memory recall abilities, for example.

   (e) The study included an **experiment** where the participants been shown two sets of words in the fixed-random order and asked to rate them based on how a) relevant the word is to them in the survival situation and b) pleasantness of the word; later they completed a survey about the words presented as well as filled a brief demographic questionnaire.

   (f) **Dependent variable: Response time**

   (g) **Independent variable: Whether the participants' words selection is based on L1 or L2.**

   (h) **Response time: Continuous variabe**
   Word selection based on L1/L2: Discrete variable

   (i) **Response time: Ratio scale**
   Word selection based on L1/L2: Ordinal scale

   (j) A comparison **T-test** - a one-way analysis of variance **(ANOVA)** - was used.

   (k) **ANOVA** is usually used when more than two groups are compared; in the case of the study presented there are four different sets of conditions in which there is one independent nominal variable (word selection based on L1/L2), and one dependent ratio-scale variable (response time).