End of Module Assessment 1: Statistical Analysis Presentation Health Survey for England 2011 Data Written Audio Transcript, References, Appendix

By William Bolton
16.02.2024
Word count: 1339 (Minimum 1000; Maximum 1500 equivalent)

Slide 1: Title slide.

Hello, my name is Will Bolton, and I am a postgraduate certificate in artificial intelligence student at the University of Essex online. My presentation today is around the analysis of health survey data for England in 2011 looking at alcoholism. The structure of my presentation will cover a brief introduction and overview of the situation with alcohol and related impact on population health in the UK, an overview of the dataset I am using, and the results of my descriptive and inferential statistical analysis.

Slide 2: Brief introduction slide

Alcohol is a major problem in population ill-health in the UK. A 2009 study in the Journal of Public Health estimated that alcohol consumption was responsible for 31,000 deaths in the UK and that alcohol consumption cost the UK NHS £3.0 billion. Alcohol consumption was responsible for 10% of all disability adjusted life years (male: 15%; female: 4%) in the UK (Balakrishnan et al, 2009). Alcohol misuse is known to cause a range of health conditions as summaries in this infographic from the UK Health Security Agency, in cancer, heart and liver disease, and mental health problems (O'Connor R, UK Health Security Agency).

Slide 3: The dataset used

Health Survey for England, or HSE, is a series of annual surveys, carried out since 1991, designed to measure health and health related behaviours in adults and children. In 2011, a core topic covered was population drinking (Health Survey for England, 2011). The survey represents a National Statistic and presents sociodemographic and behavioural information to enable understanding of the determinants of health. It helps policy makers and members of the public themselves to improve the health of the nation. In this presentation and assignment, I will be performing a range of descriptive and inferential statistics to understand this data better. I will use the statistical software R because it is an advanced language that allows users to perform a wide range of statistical computations and is well suited for graphical interpretation of data, making it easy to understand.

Slide 4: Descriptive statistics of the dataset

This table summaries descriptive statistics of the dataset relevant to the tasks in this assignment. A large majority of participants drank alcohol (although around 1600 did not answer) and just over half were women. This is a large majority and suggests drinking is prevalent in this sample. You would expect roughly equal split in the population of males and females. This suggests this dataset is likely to be representative of the wider population. One in five obtained the highest educational level which was a NVQ4/5 or degree/equivalent. A minority were divorced, and a smaller minority were separated. The majority were married or single.

Slide 5: Descriptive statistics of the dataset

Household size, BMI and age at last birthday are summarised here with some basic descriptive statistics to highlight the centre, spread and variance of the data. The majority of people seem to live in households of around 3 people, have a slightly overweight BMI being just over 25 (where normal is below 25 and above 18.5) and are around 42 years old.

Slide 6: Inferential statistics of the dataset

I wanted to conduct a significance test to find our which gender drinks more alcohol. A Chi-square test is appropriate here to examine for differences in proportions between two categories (gender being a categorical variable). Before running the test, I made a 2x2 contingency table above. The chi-square demonstrated that the association between gender and drinking status can be considered to be extremely statistically significant. Being female is associated with a positive drinking status.

Slide 7: Inferential statistics of the dataset

Next, I wanted to conduct a significance test to find out which region drinks the most alcohol. For this I also used Chi-square test as I am working with categorical variables. A Chi-square test is appropriate here to examine for differences in proportions between categorical variables. Before running the test, I another contingency table here. The chi-square statistic was 98.53 with a p-value < 0.0001 and therefore, the association between region and drinking status can be considered to be extremely statistically significant. The South East was the region that drinks the most.

Slide 8: Inferential statistics of the dataset

Now I want to investigate whether there is a statistical difference between men and women within the variables of valid height and valid weight. To decide on the statistical test, I had to run a. normality test. I chose the Shapiro-Wilk test for normality, as it is a common choice. This is constructed such that the null hypothesis is that the data are normally distributed, and the alternative hypothesis is that the data are not normally distributed. If the p-value is less than 0.05, the distribution fails the normality test, and one should use a non-parametric test to compare the medians instead. In this case, for both male and female the normality test was failed for both weight and height, as demonstrated by these R output screenshots. I therefore conducted a Mann-Whitney U test for both variables, height and weight. For differences between valid height between men and women, the statistic was 14713021, p-value < 0.0001. For differences between valid weight between men and women, the statistic was 12449400, p-value < 0.0001. This indicates a strongly statistical significance in height and weight between men and women, which is to be expected.

Slide 9: Inferential statistics of the dataset

Finally, I wanted to assess the correlation between whether a person drinks nowadays, total household income, age at last birthday and gender. To do this I asked R to build a correlation matrix presented here. This table shows us how each of the variables correlates with the others. It is important to note that correlation does not imply causation. Nonetheless, a correlation coefficient near 1 or -1 indicates a strong positive or negative linear relationship, respectively. A correlation coefficient near 0 suggests little to no linear relationship. The correlations between these variables are all quite weak, indicating that none of the examined factors have a strong linear relationship with whether a person drinks or not. The strongest correlation observed is between gender and drinking status although it remains weak. This suggests a slight difference in drinking behaviour between genders. These weak correlations highlight the complexity of drinking behaviour, suggesting it's influenced by a mix of factors not fully captured by income, age, and gender alone.

Slide 10: Discussion

From this dataset, it seems that women are more likely to be drinkers than men and region is associated with drinking status being more prevalent in the South East. Drinking is a complex behaviour, and several factors are at play. This makes analysing and interpreting data challenging. Beard et al (2019) identified that social-grade and educational attainment appear to be the strongest predictors of alcohol consumption in England, followed closely by housing tenure. These are both potentially related to region and gender. In a critical review of alcohol consumption and gender, Moinuddin et al (2016) revealed that an increase in female alcohol consumption has been observed in most societies and cultures, including in the UK. They highlight that women were more likely to be affected by alcohol abuse owing to a range of biological and sociocultural abuse, leading to what the authors describe as an ‘emerging female alcohol epidemic’. To conclude, I have two main recommendations based on this data and its relevance to the existing literature. First is that more detailed data collection and mixed methods analysis sis needed to unpick the underlying mechanisms of this complex multi-factorial behavioural problem. Second is that future research is needed to identified effective interventions designed to help women who may be affected by alcohol abuse or at risk of this.

Slide 11: References

Here are my references. A full appendix of selected R commands and output to demonstrate my working and learning is available at the end of my written audio transcript documented submitted with this recording.

Slide 12:

Thank you very much for your attention, this concludes the presentation.

References
(Not dictated verbatim)

Balakrishnan, R., Allender, S., Scarborough, P., Webster, P. and Rayner, M., (2009). The burden of alcohol-related ill health in the United Kingdom. Journal of Public Health, 31(3), pp.366-373.

Beard, E., Brown, J., West, R., Kaner, E., Meier, P. and Michie, S. (2019). Associations between socio-economic factors and alcohol consumption: a population survey of adults in England. PloS one, 14(2), p.e0209442.

Health Survey for England 2011. Online at: https://www.gov.uk/government/publications/health-survey-for-england-2011 Date accessed: 10.02.2024

Moinuddin, A., Goel, A., Saini, S., Bajpai, A. and Misra, R.(2016). Alcohol consumption and gender: a critical review. J Psychol Psychother, 6(3), pp.1-4.

O'Connor R, UK Health Security Agency. Alcohol – Some encouraging trends Online at: https://ukhsa.blog.gov.uk/2015/06/17/alcohol-some-encouraging-trends/ Date accessed: 10.02.2024.

Appendix: R commands and selected output

Loading the data

```
install.packages("haven")
library(haven)
hse_data <- read_sav("HSE 2011.sav")
```

Exploring the data

```
> n <- nrow(HSE_2011)
> print(n)
[1] 10617

> attributes(HSE_2011)

> print_labels(HSE_2011$dnnow)

Labels:
 value                label
    -9              Refusal
    -8           Don't know
    -1 Item not applicable
     1                  Yes
     2                   No

> table(HSE_2011$dnnow)

   1    2
6712 1822


> print_labels(HSE_2011$Sex)

Labels:
 value                   label
    -9                 Refusal
    -8              Don't Know
    -2 Schedule not applicable
    -1     Item not applicable
     1                    Male
     2                  Female

> table(HSE_2011$Sex)

   1    2
4852 5765


> print_labels(HSE_2011$topqual3)

Labels:
 value                     label
    -9                  Refused
    -8               Don't know
    -7       Refused/not obtained
    -6      Schedule not obtained
    -2    Schedule not applicable
    -1            Not applicable
     1  NVQ4/NVQ5/Degree or equiv
     2     Higher ed below degree
     3     NVQ3/GCE A Level equiv
     4     NVQ2/GCE O Level equiv
     5 NVQ1/CSE other grade equiv
     6              Foreign/other
     7           No qualification

> table(HSE_2011$topqual3)

   1    2    3    4    5    6    7 
2008  948 1248 1803  395  127 2037 


> print_labels(HSE_2011$marstatc)

Labels:
 value                                      label
    -9                                    Refused
    -8                                 Don't know
    -7                       Refused/not obtained
    -6                      Schedule not obtained
    -2                    Schedule not applicable
    -1                             Not applicable
     1                                     Single
     2                                    Married
     3 Civil partnership including spontaneous answers
     4                                  Separated
     5                                   Divorced
     6                                    Widowed
     7                                 Cohabitees

> table(HSE_2011$marstatc)

   1    2    3    4    5    6    7 
1613 4501    4  224  594  693  979 
```

Understanding descriptive statistics for specific variables

```
> mean(HSE_2011$HHSize, na.rm = TRUE)
> median(HSE_2011$HHSize, na.rm = TRUE)
> get_mode <- function(x) {
+     ux <- unique(x)
+     ux[which.max(tabulate(match(x, ux)))]
+ }
> mode_value <- get_mode(HSE_2011$HHSize)
> min_value <- min(HSE_2011$HHSize, na.rm = TRUE)
> max(HSE_2011$HHSize, na.rm = TRUE)
> max(HSE_2011$HHSize, na.rm = TRUE)
> range(HSE_2011$HHSize, na.rm = TRUE)
```

Creating contingency tables and performing Chi-squared tests

```
> table(HSE_2011$Sex, HSE_2011$dnnow)

      1    2
  1 3172  605
  2 3540 1217
```

1 = male
1= yes drink

```
> table_drink_sex <- table(HSE_2011$Sex, HSE_2011$dnnow)

> chi_square_result <- chisq.test(table_drink_sex)


> attributes(HSE_2011$gor1)
$label
[1] "Government Office Region - numeric"

$format.spss
[1] "F1.0"

$display_width
[1] 6

$class
[1] "haven_labelled" "vctrs_vctr"     "double"

$labels
        North East              North West
             1                       2
Yorkshire and The Humber           East Midlands
             3                       4
      West Midlands         East of England
             5                       6
          London              South East
             7                       8
        South West                 Wales
             9                      10
         Scotland
            11
```

```
> contingency_tabl1 <- table(HSE_2011$gor1, HSE_2011$dnnow)
>
> table(HSE_2011$gor1, HSE_2011$dnnow)

      1    2
  1  576  135
  2  833  270
  3  686  201
  4  624  136
  5  686  207
  6  763  172
  7  674  304
  8 1130  255
  9  740  142
```

Normality testing and non-parametric tests

```
> height_male <- HSE_2011$htval[HSE_2011$Sex == 1]
>
> height_female <- HSE_2011$htval[HSE_2011$Sex == 2]
>
> shapiro.test(height_male)

        Shapiro-Wilk normality test

data:  height_male
W = 0.73755, p-value < 2.2e-16

>
> # Shapiro-Wilk normality test for females
> shapiro.test(height_female)

        Shapiro-Wilk normality test

data:  height_female
W = 0.75777, p-value < 2.2e-16
```

Correlation matrices

```
> data_subset <- HSE_2011[, c("dnnow", "totinc", "Age", "Sex")]
>
> correlation_matrix <- cor(data_subset, use="complete.obs")
>
> print(correlation_matrix)

1.00000000 0.073253390  0.07660716  0.116688967
0.07325339 1.000000000  0.05120967  0.002515187
0.07660716 0.051209667  1.00000000 -0.015937349
0.11668897 0.002515187 -0.01593735  1.000000000
```