**Audio transcript – statistical analysis presentation**

**Slide 1**

Hi. For this final assignment of the numerical analysis module, I will present a statistical analysis report which provides an epidemiological perspective of alcohol consumption in England in the year 2011

**Slide 2**

Let’s start first with a brief background on the current landscape of alcohol consumption in England, based on the latest data of the Health Survey for England (NHS England, 2022). In 2022, 81% of English adults over the age of 16 reported drinking alcohol in the last year, and 55% at least weekly.

The survey also showed an increase in weekly average consumption to 15.3 units per week in 2022, from an average of 14.5 in 2021, and a historical minimum of 10.9 in 2015 (NHS England, 2022).

These results are particularly concerning, given that alcohol consumption is associated with a number of both short and long-term risks, the latter including cardiovascular disease, various gastrointestinal conditions, as well as several types of cancer (NHS, 2022).

For this assignment, I will be exploring demographic patterns related to alcohol consumption in England, using openly-available, anonymised, and individual-level data from the publication of the very same Health Survey for England in 2011. This dataset includes information on 58 variables, of which 12 will be used here.

Following from the assignment instructions, this presentation will include a number of descriptive statistics, where I will provide general descriptive summaries of particular variables of interest, as well as inferential statistics, where I will test different hypothesis which aim to establish relationships between select variables.

The analysis tool chosen for this assignment is R. R is an open source programming language which was developed specifically for statistical analysis and is widely used in finance and science given its simple but powerful syntax and strong support from an established community of programmers and developers.

**Slide 3**

Let’s start with descriptive statistics

**Slide 4**

Along this presentation, I have included the tasks specified in the instructions as numbered questions. First, there are 10617 people included in the sample, of which 5765 (or 54.3%) are women

**Slide 5**

Among people included in this sample, 6712 or 63.2% report drinking nowadays. However, this variable has a large number of missing values, amounting to 2083 or 19.6% of the total sample. Therefore, if those people are excluded, then the proportion of people who report drinking nowadays is 78.7% of those with complete data.

**Slide 6**

The highest educational level reported is Degree level or equivalent (corresponding to NVQ grades 4 and 5), and accounting for 2008 people or 18.9% of the total sample.

**Slide 7**

With regards to marital status, there are 594 divorced people, which accounts to 5.6% of all people included in the dataset and 6.9% of those with complete data. There are also 224 separated people, representing 2.1% of all people and 2.6% of those with complete data.

**Slide 8**

This slide shows visual representations of the distribution for age at last birthday using and histogram and density plot, and summary statistics in the tablet and boxplot. The mean was 41.6 years, median 42, and mode 43.77 (which was calculated using the probability density function). Minimum and maximum age were 0 and 100 years respectively, leading to a range of 100 years, and standard deviation 23.8.

**Slide 9**

As for body mass index or BMI, the mean was 25.9, median 25.59, and mode 25 (also calculated using the probability density function). Minimum and maximum BMI were 8.34 and 100 respectively, leading to a range of 56.94 years, and standard deviation was 6.14.

**Slide 10**

Finally, for household size I present here a barchart showing counts of people with different household sizes, as well as the summary table. The mean household size was 2.85 people, median 3, and mode 2 (as shown by the tallest bar in the chart). Minimum and maximum household size were 1 and 10 respectively, leading to a range of 9 years, and standard deviation was 1.37.

**Slide 11**

We will now proceed to inferential statistics and hypothesis testing

**Slide 12**

The first assignment question was which sex drinks more alcohol.

The barchart and contingency table show the frequency distributions of whether people drink nowadays and their sex. Both show that a nominally larger proportion of men report drinking nowadays when compared with women (84% versus 74.4%).

A chi-squared test was performed to test the hypothesis that there was a statistically significant difference in the proportions of people drinking alcohol nowadays in both sexes, and the R console output is shown at the bottom. The chi-squared statistic is 114.15 with a p-value of <0.001, showing that this difference is highly statistically significant. We can therefore conclude that men drink on average more than women.

**Slide 13**

The following question was which English region drinks the most.

The barchart and contingency table show the frequency distributions of whether people drink nowadays and their region. Both show that the region where the proportion of people who report drinking nowadays is nominally larger is the South West of England (accounting for 83.9%).

A chi-squared test was performed to test the hypothesis that there is a statistically significant difference in the proportions of people drinking alcohol nowadays across regions. The chi-squared statistic is 98.53 with a p-value of <0.001, showing that this difference is highly statistically significant. We can therefore conclude that the proportion of people who drink nowadays is different across UK regions, with those in the South West drinking the most.

**Slide 14**

We now turn to assessing differences in body composition between men and women, starting with height.

The plots show the distribution of height grouped by sex and a boxplot with summary statistics, which are also shown in the table. We can see that all components of the five-figure summary for height, as well as the mean, are higher in men.

In order to run hypothesis testing, we first need to check whether the assumptions of parametric tests of independent samples for continuous variables hold true, namely that the data are normally distributed. The result from the Kolmogorov-Smirnoff test shows that this is not the case, as the null hypothesis of this test is that the data are normally distributed, but the p-value is <0.001 and therefore we reject the null hypothesis that height is normally distributed. The Kolmogorov-Smirnoff test was used here instead of the Shapiro-Wilk test because the number of observations was very large. The absence of a normal distribution is also suggested by the heavily left-skewed distribution seen in the histogram, and in the distance between mean and median.

Based on these results, a non-parametric Wilcoxon independent samples test was performed. The hypothesis to be tested here is that there is a statistically significant difference in average height between men and women. The p-value for the resulting test statistic is <0.001, and we can therefore conclude that there is a highly statistically significant difference in height between men and women, with men being taller on average than women.

**Slide 15**

A similar analysis was performed for weight. As with height, all components of the five-figure summary for weight, as well as the mean, are higher in men.

As before, the result from the Kolmogorov-Smirnoff test shows that weight is also non-normally distributed. The absence of a normal distribution is also suggested by the bimodal and heavily right-skewed distribution seen in the histogram, and in the distance between mean and median shown in the table.

Based on these results, a non-parametric Wilcoxon independent samples test was performed. The hypothesis to be tested here is that there is a statistically significant difference in average weight between men and women. The p-value for the resulting test statistic is <0.001, and we can therefore conclude that there is a highly statistically significant difference in weight between men and women, with men being heavier on average than women.

**Slide 16**

Finally, the last task was to calculate the correlation between whether a person drinks nowadays, total household income, age at last birthday, and gender. The table shows a correlation matrix across these 4 variables. Of note, drinking nowadays was coded as 0 for No and 1 for Yes, so that a positive correlation coefficient indicates an association with drinking. Sex was coded as 1 for males and 2 for females, so that a positive correlation coefficient indicates an association with female sex.

The correlation between drinking nowadays and total household income was negative and weak (with a correlation coefficient of -0.073 with a p-value of <0.001), indicating a highly statistically significant inverse association between increasing household income and drinking nowadays.

Drinking nowadays was also negatively and weakly correlated with age at last birthday (with a correlation coefficient of -0.069 with a p-value of <0.001), indicating a highly statistically significant inverse association between increasing age and drinking nowadays.

Drinking nowadays was negatively and moderately correlated with sex (with a correlation coefficient of -0.116 and p-value of <0.001), indicating a highly statistically significant inverse association between female sex and drinking nowadays.

Total household income was positively and weakly correlated with age at last birthday (with a correlation coefficient of 0.05 and with a p-value of <0.001), indicating a highly statistically significant direct association between increasing age and increasing total household income.

On the other hand, total household income showed a positive and weak correlation with sex (with a correlation coefficient of 0.005), but this correlation was not statistically significant at an alpha of 0.05, indicating no evidence of a statistically significant association between total household income and sex.

Finally, age at last birthday was positively and weakly correlated with sex (with a correlation coefficient of 0.033 and p-value of <0.001), indicating a highly statistically significant direct association between increasing age and being of the female sex.

**Slide 17**

The main finding from this presentation in relation to sex and drinking was that, in the 2011 report of the Health Survey for England, men were more likely to report drinking nowadays than women (with a highly statistically significant difference of 84% vs 74.4%).

More updated data from the same survey for 2022 showed an identical result in terms of the proportions of men who reported drinking in the last year, but with a 3.6% absolute increase in women from 74.4% to 78%. Of note, this survey also showed that men are more likely to report drinking at least weekly, and with a higher amount of mean weekly units of alcohol consumed.

**Slide 18**

In conclusion, this presentation showed that men were more likely to be current drinkers than women in England in the year 2011, with similar findings observed in the exact same survey in 2022, but with a worsening trend for women.

These results highlight that more emphasis should be put on efforts to educate the population on alcohol related harms in order to reduce its consumption.

They also suggest that campaigns tailored specifically to men may be needed in order to bring the frequency of alcohol consumption among men closer to that seen among women, but at the same time that campaigns designed for women are also important in order to limit and reverse the secular trend of increase in the proportion of women who report drinking seen from 2011 to 2022.

**Slide 19**

These are my references.

**Slide 20**

Thank you for your attention. The full R code used for this analysis is included as an appendix in my submission and in my GitHub portfolio, together with the corresponding outputs in the form of graphs, tables, and a PDF report.

**References:**

NHS (2022) ‘Risks - Alcohol misuse’. Available from: https://www.nhs.uk/conditions/alcohol-misuse/risks/.

NHS England (2022) *Health Survey for England*. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england.