Statistical Analysis Presentation: Transcript

Maria Ingold

INTRODUCTION

[2:Intro]

The Office for National Statistics (ONS) shows that approximately twice as many males die from alcohol than females annually, and that while drinking-related deaths remained consistent through 2019, they statistically significantly increased since 2020 (Office for National Statistics, 2024). By contrast, the Health Survey for England (HSE), which annually evaluates population proportions of English alcohol intake, shows the percentage of English drinkers are generally flat, perhaps declining (NHS England Digital, 2022). Unfortunately, COVID-19 meant skipping research in 2020 and changing it in 2021, so this decline is not directly comparable.

The 2011 HSE teaching dataset, helps understand how to analyse gender, drinking and related data, through descriptive statistics—summarising, organising, and displaying—and inferential statistics—drawing conclusions and generalising to a population (Bruce, Bruce & Gedeck, 2020). The analysis used R, a free, open-source statistical scripting language with graphing capabilities (R, N.D.).

DESCRIPTIVE STATISTICS

[3:a-e]

The HSE dataset contains 10,617 unique serial numbers, indicating the number of people in the sample. Just over half—54.3%—are women.

The other statistics are based on respondents. Just over three-quarters drink alcohol nowadays. Almost a quarter reached the sample's highest educational level: the National Vocational Qualification four, five, degree or equivalent (City & Guilds, N.D.a, N.D.b). About 2.6% are separated and 6.9% are divorced.

[4:f:HHSize]

Household size counts the number of people—1 to 10—in a household. Household size is right-skewed, which makes the most common size 2 and the median 3. The lower standard deviation shows the data concentrates close to the mean, which shifts right because of outliers 8, 9 and 10, and left by a high frequency of 2. Therefore, the median—3—best represents the central tendency in this case.

[5:f:BMI]

Body Mass Index (BMI) is a continuous random variable, so mode estimation uses a probability density function plot—which is right skewed—and shows the most common—the mode—at about 25.

At just over 6, the standard deviation spreads numbers somewhat out from the mean, and the mean shifts right because of outliers above 40. So, median—25.59—is again best for central tendency.

[6:f:Age]

With ages 0 to 100, skew is close to 0. However, because it is bimodal at 42 and 64, with another minor peak at 2, it isn't normal.

Mean and median work best with a unimodal—single mode—distribution (Holmes, Illowsky & Dean, 2017). So, here the modes are most useful.

The high standard deviation of nearly 24 shows the data spreads out from the mean—there are variations in the ages—which is reflected in the histogram.

INFERENTIAL STATISTICS

[7:a:gender]

Because both gender and alcohol are categorical, a normality test is not required. The non-parametric chi-squared test of independence rejects the null hypothesis that gender and drinks nowadays are independent variables. The p-value—less than .001—confirms that there is a very highly significant association. While not implying causation, the contingency table and the chi-squared test indicate that a larger percentage of women drink nowadays than men. However, over double the percentage of women than men do not drink nowadays. This could account for the very high significance in the p-value.

[8:b:region]

Again, because region and alcohol are both categorical, there is no normality test. The chi-squared test p-value rejects the null hypothesis that region and drinks nowadays are independent. At less than .001 there is a very highly significant association between region and drinks nowadays.

[9:b:region]

The bar graph clearly shows that the South East has the highest percentage of drinkers. Given the significance, the inference is that the South East drinks the most alcohol.

[10:c:height]

While gender is categorical, height is a continuous random variable, so normality tests are required. Initial tests on mean, median and mode, per gender, show that they are not all equal. Q-Q plots are not linear, the histogram—per gender for height—confirms the data is left-skewed, and Shapiro-Wilk proves non-parametric is required.

[11:c:height]

The Mann-Whitney U test's null hypothesis is that height distributions for males and females are equal. The alternative is not equal.

With a p-value less than .001, we reject the null hypothesis that they are equal. This result is very highly significant—gender does affect height. Outliers on valid data are likely due to younger ages being shorter.

[12:c:weight]

Gender and weight are also categorical and continuous variables. Q-Q plots are not linear, the histogram has two peaks, and Shapiro-Wilk's p-value is less than .001 which is very highly significant in confirming it is not normal.

[13:c:weight]

Similarly, weight uses a Mann-Whitney U test. Null is weight distributions for males and females are equal, while the alternative is not equal. The p-value is again less than .001. We reject the null that they are equal. It is very highly significant—gender affects valid weight. Low outliers are likely younger ages, and estimated weights over 130kg are classed as valid. Unlike mean, medians are less sensitive to outliers (Bruce, Bruce & Gedeck, 2020). The medians are different. There is a statistical difference between men and women on valid weight.

[14:d:correlation]

Drinks nowadays and gender are both categorical, which do not require a normality test. However, both age—quantitative—and total household income—ordinal—have Q-Q plots, histograms, and Anderson-Darling p-value tests show they are very highly significantly not normal.

[15:d:correlation]

The only two fully appropriate tests in this correlation table are Spearman's Correlation between ordinal variable total household income and quantitative variable age, and Cramer's V between the two categoricals—which are normal—gender and drinks nowadays (Mangiafico, 2023). The remaining use Spearman's and are indicative.

Age at last birthday and total household income had the most interesting graph. While not linear, the upper right quadrant shows it is negatively, but poorly correlated, with the p-value indicating this is statistically very highly significant. Total household income and drinks nowadays, and gender and total household income, were also negatively but poorly correlated at a very high level of significance.

Age at last birthday and drinks nowadays, and gender and age at last birthday, were both positively but poorly correlated, with the former being very highly significant and the latter only highly significant.

Gender and drinks nowadays are both categorical, so Cramer's V was applied. Comparing Cramer's phi and Spearman's rho demonstrated the values were the same. Very weakly positive, very poor correlation. The p-values both indicated very high statistical significance that gender and drinks nowadays are indeed very poorly correlated.

RELEVANT LITERATURE

[16:Literature]

Because COVID-19 impacted the HSE surveys, which show a higher percentage of men drinking than women in the UK, further literature is required (NHS England Digital, 2022). Garnett et al. (2021) sampled over 22,000 drinkers during the first COVID-19 2020 lockdown. Almost half reported drinking the same, with about a quarter drinking less and a quarter drinking more than usual in the last week. Lighter drinkers and men were drinking less, possibly due to diminished peer pressure, and heavier drinkers and women more, but men were still drinking more heavily, as before lockdown (Alcohol Change UK, 2020; Morris et al., 2020; Garnett et al., 2021).

The increase in female drinking was largely due to increases in gender inequality and worsened mental health (Garnett et al., 2021). Oreffice and Quintana-Domeque's (2021) UK research for roughly the same time-period, disagrees on gender drinking, finding women were 19% less likely to drink alcohol, but corroborates women's decreased mental health and acknowledges that women's adjusted mean difference in childcare is 219% higher than men.

[17:Literature]

Which is why Oldham et al.'s (2021) self-selected survey is surprising when it discovers that both mental health and living with children during the first lockdown are much more significant factors of heavy episodic drinking for men than women. While self-selected is less reliable for generalisation, it's worth investigating these at-risk areas further (Holmes, Illowsky & Dean, 2017; Oldham et al., 2021).

CONCLUSION AND RECOMMENDATIONS

[18:Conclusion]

In Summary:

Analysis of the Health Survey for England 2011 teaching dataset revealed significant gender differences in alcohol consumption. A contingency table showed a higher proportion of women reported drinking compared to men. The chi-squared test of independence found a very highly statistically significant association between gender and drinking nowadays. However, Cramer's V showed that despite the significance, the association was weak. This could be due to sample size or uneven distribution—twice the percentage of women than men do not drink, but the samples are smaller (Learn Statistics Easily, 2024).

Recommendations:

- **Data Equality**: There are 55.7% women counted in drink nowadays, but only 44.3% men. Get more equal representation.
- **Further Analysis**: Use logistic regression to predict drinking as a function of gender.
- **Time Series Analysis**: Conduct further research across subsequent years to see if results are similar.

REFERENCES

Alcohol Change UK (2020) *Drinking during lockdown: headline findings*. Available from: https://alcoholchange.org.uk/blog/covid19-drinking-during-lockdown-headline-findings [Accessed 25 May 2024].

Bruce, P., Bruce, A. & Gedeck, P. (2020) *Practical Statistics for Data Scientists*. 2nd ed. O'Reilly Media. Available from: vbk://9781492072898 [Accessed 24 May 2024].

City & Guilds (N.D.a) *NVQs and SVQs*. Available from: https://www.cityandguilds.com/qualifications-and-apprenticeships/qualifications-explained/nvqs-svqs-keyskills-vocational-skillsforlife [Accessed 9 May 2024].

City & Guilds (N.D.b) *Qualification Comparisons - NVQ Level 1, 2, 3, 4, 5, 6, 7, 8*. Available from: https://www.cityandguilds.com/qualifications-and-apprenticeships/qualifications-explained/qualification-comparisons [Accessed 7 May 2024].

Garnett, C., Jackson, S., Oldham, M., Brown, J, & Steptoe, A. (2021) Factors associated with drinking behaviour during COVID-19 social distancing and lockdown among adults in the UK, *Drug and alcohol dependence* 219. DOI: https://doi.org/10.1016/J.DRUGALCDEP.2020.108461.

Holmes, A., Illowsky, B. & Dean, S. (2017) *Introductory Business Statistics*. Available from: https://archive.org/details/IntroductoryBusinessStatistics [Accessed 25 May 2024].

Learn Statistics Easily (2024) *Cramer's V and Its Application for Data Analysis*. Available from: https://statisticseasily.com/cramers-v/ [Accessed 26 May 2024].

Mangiafico, S.S. (2023) Summary and Analysis of Extension Program Evaluation in R.

Morris, H., Larsen, J., Catterall, E., Moss, A., Dombrowski, S. (2020) Peer pressure and alcohol consumption in adults living in the UK: A systematic qualitative review, *BMC Public Health* 20(1): 1–13. DOI: https://doi.org/10.1186/S12889-020-09060-2/FIGURES/2.

NHS England Digital (2022) *Health Survey for England, 2021 part 1*. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021 [Accessed 24 May 2024].

Office for National Statistics (2024) *Alcohol-specific deaths in the UK: registered in 2022*. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/bulletins/alcoholrelateddeathsintheunitedkingdom/registeredin2022 [Accessed 24 May 2024].

Oldham, M., Garnett, C., Brown, J., Kate, D. Shahab, L., & Herbec, A. (2021) Characterising the patterns of and factors associated with increased alcohol consumption since COVID-19 in a UK sample, *Drug and Alcohol Review* 40(6): 890–899. DOI: https://doi.org/10.1111/DAR.13256.

Oreffice, S. & Quintana-Domeque, C. (2021) Gender inequality in COVID-19 times: evidence from UK prolific participants, *Journal of Demographic Economics* 87(2): 261–287. DOI: https://doi.org/10.1017/DEM.2021.2.

R (N.D.) *What is R?* Available from: https://www.r-project.org/about.html [Accessed 25 May 2024].