University of Essex

Online

NA_PCOM7E April 2025 B

Final Presentation

University of Essex

Online

I. Introduction – Executive Summary

Introduction

As a broad overview, the task at hand is to evaluate the statistical significance of a variety demographic factors drawn from the 2011 Health Survey for England; in particular, the focus of analysis rests on finding and evaluating predictive demographic factors relevant to the consumption of alcohol by adult survey participants (as of the 2011 date of the survey).

Summary of Findings

Men were more likely to report consuming alcohol than women, but a large percentage of both genders reported consuming alcohol (well over 70% for each).

Despite high alcohol consumption rates for both genders, significant disparities were found by gender in how alcohol consumption is studied and by which investigative framework, whether medical or sociological.

Likewise, public health policies focused on alcohol consumption exhibit a lack of awareness of gender differences in modalities of consumption, which should be rectified through further study.

University of Essex

Online

I. Introduction – Preliminary Data Cleaning

For ease and efficiency of evaluation, the source data was trimmed and cleaned prior to importing it into RStudio.

Methodology

1. The source file “HSE 2011.sav” was converted to CSV format.

2. The CSV file columns were trimmed in MS Excel to include only those columns actively used throughout the assignment.

3. The retained columns were renamed with standardized human-readable variable names for ease of reference and convenience of use in R coding.

4. Placeholder values of “247” or “255,” indicating no known / usable / applicable value recorded (King-Hele, 2013), were replaced with “NA” for ease of processing, as the string ”NA” is used throughout RStudio as a standardized placeholder value for values that need to be ignored or eliminated from processing to avoid skewing the results of analysis.

5. The new data file was then imported into RStudio as a data frame.

I. Introduction – Preliminary Data Cleaning

	A	B	C	D	E	F	G	H	I	J	K
1	householdSize	sex	ageLastBirthday	highestEducationalQualification	householdIncome	maritalStatus	region	height	weight	bmi	currentlyConsum
2	2	2	100	7	10934.5791	6	2	NA	NA	NA	2
3	1	2	97	7	NA	6	9	NA	NA	NA	1
4	1	2	96	7	NA	6	8	158.39999389648438	68	27.101825714111328	1
5	1	2	96	7	23442.62305	6	2	154.60000610351562	54.5	22.80224037	2
6	1	2	96	7	NA	6	9	152	51.20000076293945	22.16066551208496	1
7	1	2	96	7	14918.0332	6	8	147.8000030517578	47.599998474121094	21.790042877197266	2
8	1	1	95	5	NA	5	1	NA	NA	NA	2
9	3	1	95	5	NA	6	3	160	62.20000076293945	24.296875	1
10	4	2	94	7	NA	6	7	NA	NA	NA	1
11	3	2	94	4	8239.436523	6	9	NA	NA	NA	1
12	1	1	94	7	NA	6	6	NA	NA	NA	1
13	1	1	94	7	10655.7373	6	8	155.60000610351562	66.59999847	27.507747650146484	1
16	1	2	94	7	NA	6	9	148.89999389648438	53.20000076293945	23.99508285522461	1
17	1	2	93	7	14918.0332	6	9	NA	NA	NA	1
18	2	1	93	5	22100	2	4	NA	70.5	NA	2
19	1	2	93	6	19180.32813	6	8	NA	NA	NA	2
20	2	2	92	NA	NA	6	3	NA	NA	NA	2
21	2	2	92	7	22100	2	4	NA	68.09999847	NA	2
22	1	1	92	1	NA	6	8	NA	NA	NA	1
23	2	1	92	7	NA	2	5	NA	NA	NA	1
24	1	1	92	3	NA	6	9	161.5	79.69999694824219	30.557180404663086	1
25	1	2	92	2	NA	6	4	154.1999969482422	60.099998474121094	25.275846481323242	1

Excerpt of the first 25 head values of the culled data frame.

University of Essex

Online

II. Descriptive Statistics

A. Total sample	10,617
B. Percent that consume alcohol	63.22%
C. Percent women	54.30%
D. Highest educational level	Category 1 – NVQ4/NVQ5/Degree or equivalent
E. Percent divorced	5.59%
F. Percent separated	2.11%

Interpretative Notes

A. 10,617 subjects participated in the survey (King-Hele, 2013).
B. 6,712 consumers of alcohol / 10,617 subjects = 0.6322 or 63.22% of the total sample.
C. 5,765 women / 10,617 subjects = 0.543 or 54.30% of the total sample.
D. The highest educational level recorded was NVQ4/NVQ5/Degree or an equivalent qualification.
E. 594 divorced subjects / 10,617 subjects = 0.0559 or 5.59% of the total sample.
F. 224 divorced subjects / 10,617 subjects = 0.0211 or 2.11% of the total sample.

II. Descriptive Statistics

Metric	Household size	BMI	Age at last birthday
Mean	2.8507	25.91202	41.56136
Median	3	25.59349	42
Mode	2	None	64
Minimum	1	8.34011	0
Maximum	10	65.27721	100
Range	1 - 10	8.34011 - 65.27721	0 - 100
Standard deviation	1.368528	6.144844	23.83203

Interpretive Note

1. Note that BMI does not have a mode value, as there are no repeated identical values in the column.

University of Essex

Online

III. Inferential Statistics – Gender

Gender	Consumes alcohol		Test value | P-value
	Yes	**No**	X^2: 114.72 P-value: < 2.2e-16 or 0.00000000000000022
Male	83.98%	16.02%	
Female	74.42%	25.58%	

Interpretive Notes

1. 3,172 men who consume alcohol / 3,777 male subjects = 0.8398 or 83.98% of sample men.
2. 605 men who consume alcohol / 3,777 subjects = 0.1602 or 16.02% of sampled men.
3. 3,540 women who consume alcohol / 4,757 female subjects = 0.7442 or 74.42% of sample women.
4. 1,217 women who consume alcohol / 4,757 female subjects = 0.2548 or 25.58% of sampled women.
5. Note that approx. 19.52% of the total sample of 10,617 did not provide a usable yes or no answer to the question of their alcohol consumption, or the question was not applicable to them (for example, because they were children too young to consume alcohol). For this reason, percentages are based solely on those participants to whom the question was applicable and who provided usable answers about their consumption (or lack of consumption) of alcohol.

University of Essex

Online

III. Inferential Statistics – Gender

Gender	Consumes alcohol		Test value | P-value
	Yes	No	X^2: 114.72 P-value: < 2.2e-16 or 0.00000000000000022
Male	83.98%	16.02%	
Female	74.42%	25.58%	

Interpretive Notes (continued)

6. The H0 or null hypothesis is that there is **not** a significant association between gender and alcohol consumption (or lack of alcohol consumption).
7. The H1 or alternative hypothesis is that there **is** a significant association between gender and alcohol consumption (or lack of alcohol consumption).
8. The p-value is nearly zero, indicating that the we must reject the null hypothesis.
9. Broadly interpreted, there **is** a significant association between gender and alcohol consumption for those participants for whom the question was applicable and who provided usable answers.
10. **Men (83.98% of men) were more likely than women (74.42% of women) to report that they consumed alcohol.**

University of Essex

Online

III. Inferential Statistics – Region

Region	Consumes alcohol		Test value | P-value
	Yes	No	X^2: 112.28 P-value: < 2.2e-16 or 0.00000000000000022
North East	81.01%	18.99%	
North West	75.52%	24.48%	
Yorkshire	77.34%	22.66%	
East Midlands	82.11%	17.89%	
West Midlands	76.82%	23.18%	
East of England	81.60%	18.40%	
London	68.92%	31.08%	
South East	81.59%	18.41%	
South West	83.90%	16.10%	

University of Essex

Online

III. Inferential Statistics – Region

Interpretive Notes

1. As before, percentage values reflect percentages across each region, not across the entirety of the sample, and exclude missing or unusable values, as well as survey participants such as small children for whom the question of alcohol consumption did not apply.
2. The H0 or null hypothesis is that there is **not** a significant association between region and alcohol consumption (or lack of alcohol consumption).
3. The H1 or alternative hypothesis is that there **is** a significant association between region and alcohol consumption (or lack of alcohol consumption).
4. The chi-square value represents the discrepancy between the observed values versus the values expected if there were no associations present between the independent and dependent variables.
5. The P-value is nearly zero, indicating that the we must reject the null hypothesis.
6. Broadly interpreted, there **is** a significant association between region and alcohol consumption for those participants for whom the question was applicable and who provided usable answers.
7. **Residents of the London metropolitan region (68.92%) were the least likely to report currently consuming alcohol, while residents of the South West region (83.90%) were the most likely to report currently consuming alcohol.**

III. Inferential Statistics – Height

Normality Test #1 – Shapiro-Wilk

1. The first step is to test for the presence or absence of a normal distribution of height values using the Shapiro-Wilk normality test. Note that this test in RStudio can only use the first 5,000 values.
2. The H0 or null hypothesis is that height values follow a normal distribution. The H1 or alternative hypothesis is that height values are skewed, meaning they do not follow a normal distribution.
3. With an extremely small p-value well under 0.05, and a W-value significantly different from 1 (or close to 1), we must reject the null hypothesis.
4. **We conclude that height values do <u>not</u> follow a normal distribution.**

```
        Shapiro-Wilk normality test

data:  as.numeric(HSE_2011$height[0:5000])
W = 0.83553, p-value < 2.2e-16
```

III. Inferential Statistics – Height

Normality Test #2 – Anderson-Darling

1. Given that the Shapiro-Wilk normality test can only use the first 5,000 row values, but the data frame contains over 10,000 row values, we confirm the result of the first normality test using the Anderson-Darling normality test, which can evaluate all available height values.
2. As before, the H0 or null hypothesis is that height values follow a normal distribution.
3. The H1 or alternative hypothesis is that height values are skewed, meaning they do not follow a normal distribution.
4. With an extremely small p-value well under 0.05, and an A-squared value significantly larger than 1 (or a comparably small number suggestive of a normal distribution), we must reject the null hypothesis.
5. **We confirm that height values do <u>not</u> follow a normal distribution.**

```
        Anderson-Darling normality test

data:  as.numeric(h)
A = 398.24, p-value < 2.2e-16
```

III. Inferential Statistics – Height

Evaluative Test – Mann-Whitney-Wilcoxon

1. The independent variable, gender, is categorical / nominal, while the dependent variable, height, is continuous.
2. We previously asserted based on Shapiro-Wilk and Anderson-Darling tests that height is not normally distributed, which makes the Mann-Whitney-Wilcoxon appropriate to look for a significant variance in mean heights between men and women.
3. The H0 or null hypothesis is that there is no significant variance in height between the genders.
4. The H1 or alternative hypothesis is that there is a significant variance in height between the genders.
5. Given an extremely low p-value approaching zero, well below a confidence level of 0.05, we must reject the null hypothesis.
6. **We conclude that there is a significant variance in height between the genders.**

```
> wilcox.test(as.numeric(HSE_2011$height) ~ HSE_2011$sex, data = HSE_2011)

        Wilcoxon rank sum test with continuity correction

data:  as.numeric(HSE_2011$height) by HSE_2011$sex
W = 14713021, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

III. Inferential Statistics – Weight

Normality Test #1 – Shapiro-Wilk

1. As for height, the first step is to test for the presence or absence of a normal distribution of values using the Shapiro-Wilk normality test.
2. The H0 or null hypothesis is that weight values follow a normal distribution.
3. The H1 or alternative hypothesis is that weight values are skewed, meaning they do not follow a normal distribution.
4. With an extremely small p-value well under 0.05, we must reject the null hypothesis – even though the W-value approaches 1, the extremely small p-value indicates that the null hypothesis is invalid.
5. **We conclude that weight values do <u>not</u> follow a normal distribution.**

```
        Shapiro-Wilk normality test

data:  as.numeric(HSE_2011$weight[0:5000])
W = 0.9696, p-value < 2.2e-16
```

III. Inferential Statistics – Weight

Normality Test #2 – Anderson-Darling

1. As before, given that the Shapiro-Wilk normality test can only use the first 5,000 row values, but the data frame contains over 10,000 row values, we confirm the result of the first normality test using the Anderson-Darling normality test, which can evaluate all available weight values.
2. The H0 or null hypothesis is that weight values follow a normal distribution.
3. The H1 or alternative hypothesis is that weight values are skewed, meaning they do not follow a normal distribution.
4. With an extremely small p-value well under 0.05, and an A-squared value significantly larger than 1 (or a comparably small number suggestive of a normal distribution), we must reject the null hypothesis.
5. **We confirm that weight values do <u>not</u> follow a normal distribution.**

```
        Anderson-Darling normality test

data:  as.numeric(w)
A = 100.8, p-value < 2.2e-16
```

III. Inferential Statistics – Weight

Evaluative Test – Mann-Whitney-Wilcoxon

1. As with height, the independent variable, gender, is categorical / nominal, while the dependent variable, weight, is continuous.
2. We previously asserted based on Shapiro-Wilk and Anderson-Darling tests that weight is not normally distributed, which makes the Mann-Whitney-Wilcoxon appropriate to look for a significant variance in mean weights between men and women.
3. The H0 or null hypothesis is that there is no significant variance in weight between the genders.
4. The H1 or alternative hypothesis is that there is a significant variance in weight between the genders.
5. Given an extremely low p-value approaching zero, well below a significance level of 0.05, we must reject the null hypothesis.
6. **We conclude that there is a significant variance in weight between the genders.**

```
> wilcox.test(as.numeric(HSE_2011$weight) ~ HSE_2011$sex, data = HSE_2011)

        Wilcoxon rank sum test with continuity correction

data:  as.numeric(HSE_2011$weight) by HSE_2011$sex
W = 12449400, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

III. Inferential Statistics – Correlations

Methodology

1. The task is to find correlation values between whether a person drinks nowadays, total household income, age at last birthday, and gender.
2. Alcohol consumption and gender are categorical binary values (yes or no, male or female). Age at last birthday is recorded as discrete integer values (full years).
3. Household income, however, is a continuous value that may not be normally distributed, and that presence or absence of normal distribution will determine which type of correlative test to use.
4. After applying the Shapiro-Wilk and Anderson-Darling normality tests to the household income column, both of which returned extremely small p-values, we conclude that household income is not normally distributed.
5. Because household income is not normally distributed, Spearman correlation is appropriate.

```
        Shapiro-Wilk normality test

data:  as.numeric(corr$householdIncome[0:5000])
W = 0.78212, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  as.numeric(hi)
A = 483.87, p-value < 2.2e-16
```

University of Essex

Online

III. Inferential Statistics – Correlations

	Consumes alcohol?	Household income	Age at last birthday	Gender
Consumes alcohol?	1.0	-0.21820942	0.05061805	0.10714980
Household income	-0.21820942	1.0	-0.12221843	-0.06328855
Age at last birthday	0.05061805	-0.12221843	1.0	-0.02807798
Gender	0.10714980	-0.06328855	-0.02807798	1.0

```
> cor(corr, use="complete.obs", method="spearman")
                      sex ageLastBirthday householdIncome currentlyConsumes
sex                1.00000000     -0.02807798     -0.06328855        0.10714980
ageLastBirthday   -0.02807798      1.00000000     -0.12221843        0.05061805
householdIncome   -0.06328855     -0.12221843      1.00000000       -0.21820942
currentlyConsumes  0.10714980      0.05061805     -0.21820942        1.00000000
```

III. Inferential Statistics – Correlations

Interpretive Notes

1. The majority of correlative values indicated fairly weak positive and negative correlations, as the majority of values are closer to zero than they are to 1 or -1.
2. The strongest correlation is between household income and consumption of alcohol.
3. As a negative correlation, we conclude that higher household income associates with a fairly weak trend toward not consuming alcohol.
4. This correlation is weak, however, as it lands well below -1.
5. The second strongest correlation is between gender and consumption of alcohol, but it is likewise quite weak, landing far closer to zero than to 1.

IV. Discussion

Interpretation of Findings

1. Men (83.98% of men) were more likely than women (74.42% of women) to report that they consumed alcohol.
2. This finding is consistent with current literature that men are more likely to report consuming any amount of alcohol and larger quantities of alcohol, as well as more likely to experience health, legal, and other harms therefrom (Cook et al, 2025).
3. A recent meta-survey of studies found that the majority of studies focused on women who consume alcohol, however, and frequently from a sociological or cultural perspective, rather than a medical perspective (Cook et al, 2025).
4. For example, alcohol consumption by women in England is often cast through the lens of “wine-mom culture,” in which culture alcohol consumption is viewed as normal, or even desirable, as a means to cope with the stresses of parenting (Hill and Mazurek, 2023).
5. In terms of public health policy, there is a lack of current examination of gender differences in modes of alcohol usage, with – as a consequence – a similar lack of examination into the effectiveness of gender-targeted public health policies for alcohol consumption (Emslie et al, 2024).

V. Conclusions and Recommendations

Areas for Further Research

1. Further research should focus on alcohol consumption by gender through a medical lens, especially for women.
2. Further research into male alcohol consumption modalities should be undertaken in general, with an eye toward identifying cultural factors that influence consumption modalities and health outcomes therefrom.
3. Public health policy development should focus on gender-aware policies that appropriately account for and respond to gender differences in alcohol consumption modalities.

VI. Appendix – R Commands

Part II. Descriptive Statistics

Importing the data frame

```
library(readxl)
HSE_2011 <- read_excel("Desktop/Essex/Essex - Numerical Analysis/HSE 2011.xls",col_types =
c("numeric", "numeric", "numeric","numeric", "numeric", "numeric","numeric", "numeric",
"numeric","numeric", "numeric", "numeric"))
View(HSE_2011)
```

Summarizing household size data

```
mean(HSE_2011[["householdSize"]])
median(HSE_2011[["householdSize"]])
library(collapse) | fmode(HSE_2011[["householdSize"]])
min(HSE_2011[["householdSize"]])
max(HSE_2011[["householdSize"]])
range(HSE_2011[["householdSize"]])
sd(HSE_2011[["householdSize"]])
```

VI. Appendix – R Commands

Part II. Descriptive Statistics (continued)

Summarizing BMI data

```
mean(replace(HSE_2011[["bmi"]], HSE_2011[["bmi"]]==255, NA), na.rm = TRUE)
median(replace(HSE_2011[["bmi"]], HSE_2011[["bmi"]]==255, NA), na.rm = TRUE)
library(collapse)
fmode(HSE_2011[["bmi"]])
min(HSE_2011[["bmi"]])
max(HSE_2011[["bmi"]])
range(HSE_2011[["bmi"]])
sd(HSE_2011[["bmi"]])
```

Summarizing age at last birthday data

```
mean(replace(HSE_2011[["ageLastBirthday "]], HSE_2011[["ageLastBirthday "]]==255, NA), na.rm = TRUE)
median(replace(HSE_2011[["ageLastBirthday "]], HSE_2011[["ageLastBirthday "]]==255, NA), na.rm = TRUE)
```

VI. Appendix – R Commands

Part II. Descriptive Statistics (continued)

Summarizing age at last birthday data (continued)

```
fmode(HSE_2011[["ageLastBirthday"]])
min(HSE_2011[["ageLastBirthday "]])
max(HSE_2011[["ageLastBirthday"]])
range(HSE_2011[["ageLastBirthday"]])
sd(HSE_2011[["ageLastBirthday"]])
```

Part III. Inferential Statistics

Alcohol consumption by gender

```
library(dplyr)
selected_data <- HSE_2011 %>% select(sex, currentlyConsumes)
contingency_table <- table(selected_data$sex, selected_data$currentlyConsumes)
chi_square_test <- chisq.test(contingency_table)
print(chi_square_test)
```

VI. Appendix – R Commands

Part III. Inferential Statistics (continued)

Alcohol consumption by gender (continued)

```
selected_data <- HSE_2011 %>% select(region, currentlyConsumes)
contingency_table <- table(selected_data$region, selected_data$currentlyConsumes)
chi_square_test <- chisq.test(contingency_table)
print(chi_square_test)
```

Height analysis

```
shapiro_test <- shapiro.test(as.numeric(HSE_2011$height[0:5000]))
print(shapiro_test)
library(nortest)
height <- na.omit(HSE_2011$height)
ad_test <- ad.test(as.numeric(height))
print(ad_test)
wilcox.test(as.numeric(HSE_2011$height) ~ HSE_2011$sex, data = HSE_2011)
```

VI. Appendix – R Commands

Part III. Inferential Statistics (continued)

Weight analysis

```
shapiro_test <- shapiro.test(as.numeric(HSE_2011$weight[0:5000]))
print(shapiro_test )
weight <- na.omit(HSE_2011$weight)
ad_test <- ad.test(as.numeric(weight))
print(ad_test)
wilcox.test(as.numeric(HSE_2011 $weight) ~ HSE_2011$sex, data = HSE_2011)
```

Multi-variable correlation

```
shapiro_test <- shapiro.test(as.numeric(corr$householdIncome[0:5000]))
print(shapiro_test)
hi <- na.omit(corr$householdIncome)
ad_test <- ad.test(as.numeric(hi))
print(ad_test)
cor(corr, use="complete.obs", method="spearman")
```

University of Essex

Online

VI. Appendix – References

Cook, M., Pennay, A., Caluzzi, G., Cooklin, A., MacLean, S., Riordan, B., Torney, A. and Callinan, S. (2025). Examining gender in alcohol research: A systematic review of gender differences in how men and women are studied in alcohol research. *International Journal of Drug Policy*, 138, p.104763. doi:https://doi.org/10.1016/j.drugpo.2025.104763.

Emslie, C., Lyons, A., Dimova, E., Kersey, K., Burrows, A., Waddell, K. and Tello, J. (2024). Gender-Responsive Approaches to the Acceptability, Availability and Affordability of Alcohol. Brief 11. [online] Glasgow Caledonian University. WHO. Available at: https://researchonline.gcu.ac.uk/en/publications/gender-responsive-approaches-to-the-acceptability-availability-an [Accessed 14 Jul. 2025].

Hill, E.M. and Mazurek, M.O. (2023). Wine-Mom Culture, Alcohol Use, and Drinking Motives: A Descriptive Study and Cross-Cultural Exploration of American and British Mothers. *Substance Use & Misuse*, 59(3), pp.1–11. doi:https://doi.org/10.1080/10826084.2023.2275572.

King-Hele, S. (2013). *Teaching Dataset Health Survey for England 2011 User Guide*. [online] Available at: https://doc.ukdataservice.ac.uk/doc/7402/mrdoc/pdf/7402hse_2011_teaching_dataset_user_guide.pdf.

UK Data Archive Study Number 7260 - Health Survey for England. (2011). Available at: https://doc.ukdataservice.ac.uk/doc/7260/mrdoc/pdf/7260userguide.pdf.