Analysis of Health Survey England data

Statistical Analysis Presentation

Introduction

- The Health Survey for England (HSE)
  - Periodic survey
  - Monitors trends in national health
  - Estimate risk factors

- Excessive alcohol consumption (Office of National Statistics, 2020; NHS Information Centre , 2019)
  - Increased alcohol-related hospital admissions
  - Increased alcohol-specific deaths
  - prescriptions for drugs used to treat alcohol dependence
  - Increased road casualties involving illegal alcohol levels

Aim and methods

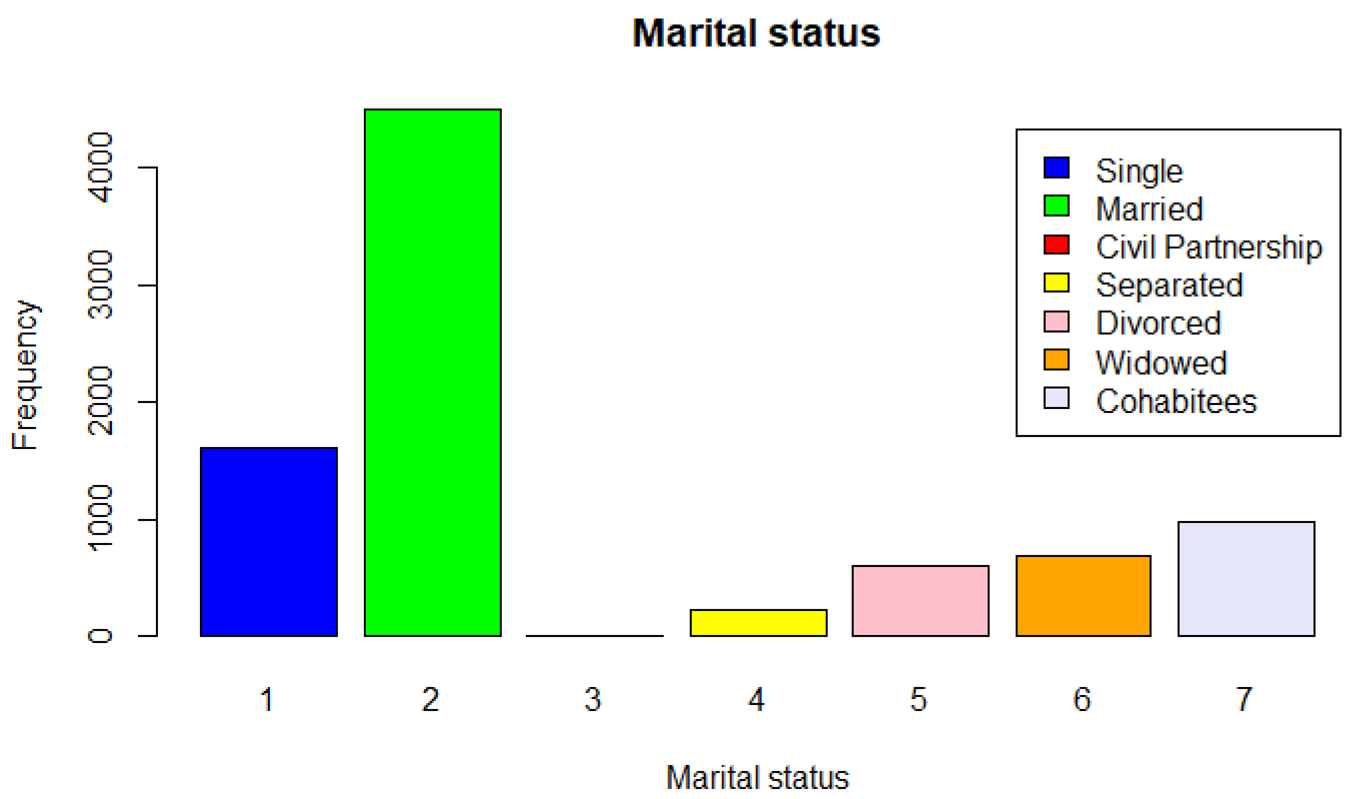
- Aim:
  - Present statistical analysis results and interpretation
- Data used
  - Health Survey for England – 2011 Publication Date:20 Dec 2012
  - Format: .sav file
- Data analysis using R
  - Descriptive statistics
  - Graphical representation
  - Inferential statistics
- R Studio
  - Free and open-source resource for data analysis
  - data visualization like pie charts, histograms, box plot, scatter plot
  - provides many statistical tests
  - has many packages, libraries of functions

Sample data

Total sample	10617
Percentage of people drinking alcohol	78.65%
Percentage of women in the sample	54.30%
Highest education level: NVQ4/NVQ5/Degree or equiv	23.44%
Percentage of Divorced people in the sample	6.90%
Percentage people live separated in the sample	2.60%

- Most people in the sample data – drink alcohol, were women, nearly quarter had highest education level

Bar chart

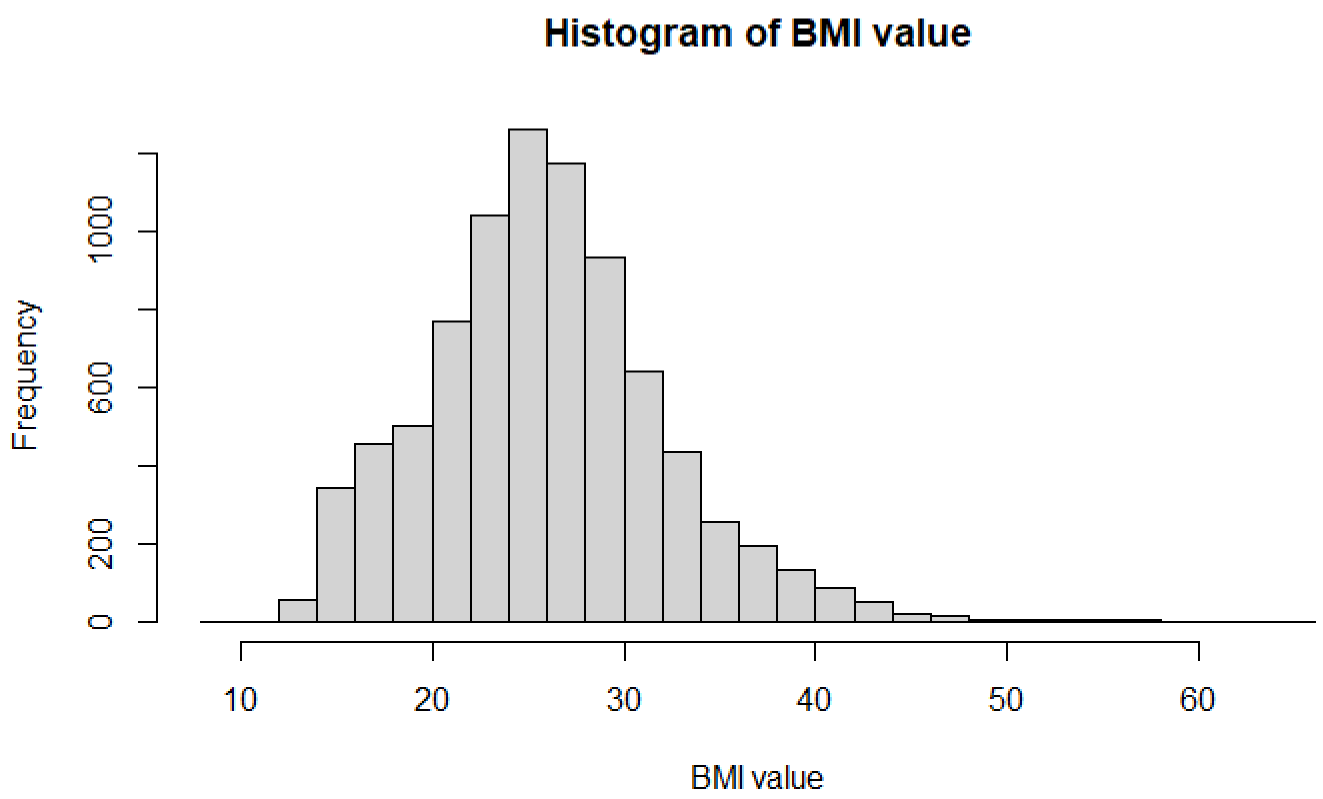


Marital status
Frequency
0
1000
2000
3000
4000
1
2
3
4
5
6
7
Marital status
Single
Married
Civil Partnership
Separated
Divorced
Widowed
Cohabitees

Descriptive statistics

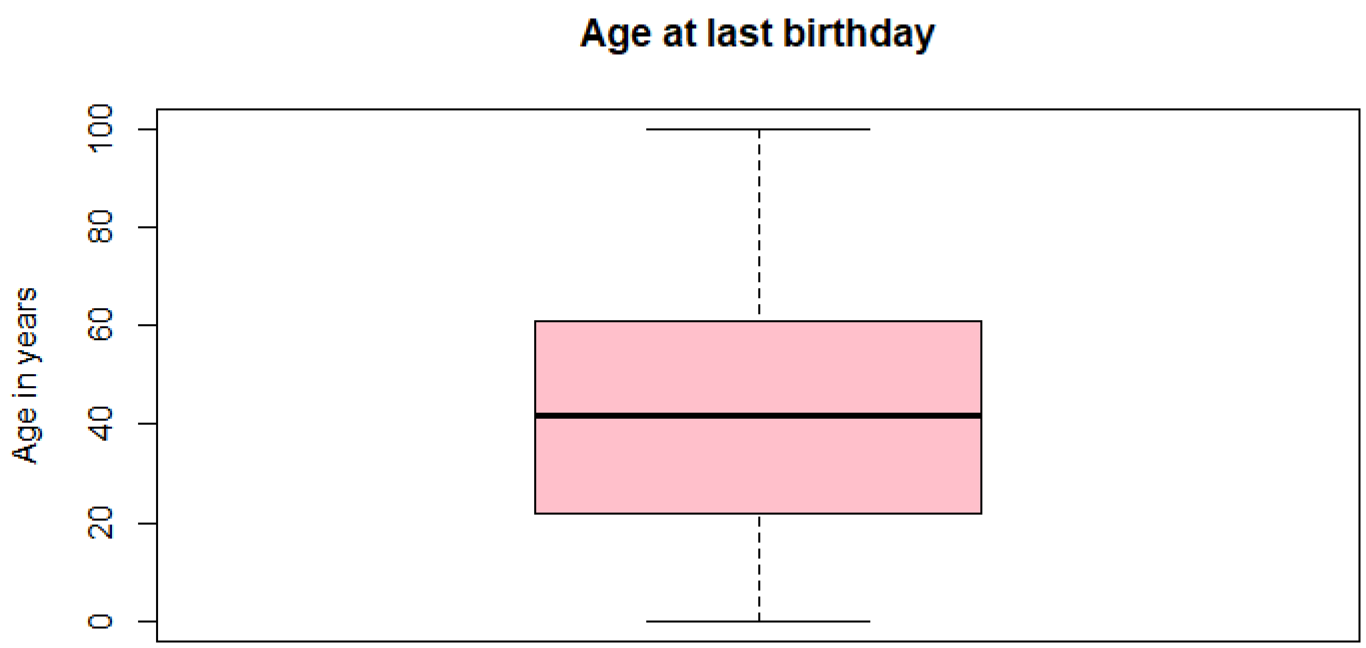
	Household size	BMI	Age at last birthday
Mean	2.85	25.92	41.56
Median	3	25.59	42
Mode	2	13.77	42
Minimum	1	8.34	0
Maximum	10	65.28	100
Range	9	56.94	100
Standard deviation	1.37	6.14	23.83

Histogram



Histogram of BMI value
Frequency
0
200
600
1000
10
20
30
40
50
60
BMI value

Boxplot



Significance test – which gender drinks more alcohol now-a-days

- Chi-square test
  - data in form of counts
  - contingency table

Gender	Drinking status in counts(and %)		Test value/p value
	1 - Yes	2 - No	
1 - Male	3172 (84 %)	605 (16 %)	114.15 / 2.2e-16
2 - Female	3540 (74.42%)	1217 (25.58%)	

  - p-value lesser than 0.005 shows very highly significant male proportion drinks alcohol compared to female

Significance test – which region drinks more alcohol now-a-days

Region	Drinking status in counts(and %)		Test value/ p value
	1 - Yes	2 - No	
1 – North East	576 (81.01%)	135(18.99%)	98.53/2.2e-16
2 – North West	833 (75.52%)	270(24.48%)	
3 – Yorkshire & Hummer	686(77.34%)	201(22.66%)	
4 – East Midlands	624(82.11 %)	136(17.89%)	
5 – West Midlands	686(77.34%)	201(22.66%)	
6 – East of England	763(81.60%)	172(18.40%)	
7 - London	674(68.92%)	304(31.08%)	
8 – South East	1130(81.59%)	255(18.41%)	
9 – South West	740(83.90%)	142(16.10%)	

p-value lesser than 0.005 shows very highly significant proportion of people in South West region drinks alcohol compared to other regions

Statistical difference between men and women on height and weight

Height - Independent two sample t-test
Null hypothesis: True difference in means of the men and women is equal to 0
Alternative hypothesis: true difference in means of the men and women is not equal to 0

- t is the t-Test statistic value (t=25.96)
- df is the degrees of freedom (df=8644)
- p-value is the significance level of the t-Test (p-value= 2.2e-16)
- Confidence interval of the mean at 95 percent (confidence interval: [9.42 , 10.96]
- Sample estimates refers to the mean value of the two samples (Mean in men group = 167.39, Mean in women group = 157.20)

The p-value of the test is 2.2e-16 which is less than the significance level alpha = 0.05. We conclude that the null hypothesis is rejected that the true difference in means of the men and women group is not equal to 0 and fail to reject the alternative hypothesis.

Weight – Independent two sample t-test
Null hypothesis: True difference in means of the men and women is equal to 0
Alternative hypothesis: true difference in means of the men and women is not equal to 0

- t is the t-Test statistic value (t=18.13)
- df is the degrees of freedom (df=8739)
- p-value is the significance level of the t-Test (p-value= 2.2e-16)
- Confidence interval of the mean at 95 percent (confidence interval: [8.48 , 10.54]
- Sample estimates refers to the mean value of the two samples (Mean in men group = 74.27, Mean in women group = 64.76)

The p-value of the test is 2.2e-16 which is less than the significance level alpha = 0.05. We conclude that the null hypothesis is rejected that the true difference in means of the men and women group is not equal to 0 and fail to reject the alternative hypothesis.

Pearson correlation, r

	Drink now-a-days	Total household income	Age at last birthday	Gender
Drink now-a-days	+1.0	+0.07	+0.07	+0.12
Total household income	+0.07	+1.0	0.05	0.00
Age at last birthday	+0.07	+0.05	+1.0	+0.03
Gender	+0.12	0.00	+0.03	+1.0

Discussion and conclusion

- Discussion
  - the percentages of high-volume drinking and high-frequency drinking - greater in men than women (Wilsnack et al. 2018; Chaiyasong et al. 2018)
  - HSE_2011 data– similar results - more male proportion drinks alcohol compared to female –using drink now-a-days variable
  - Evaluation using the variable total units of alcohol/week required

- Conclusion and Recommendation
  - Choosing the right variable and right statistic test for analysis - avoid misinterpretation of test results
  - R – great tool for statistical analysis and graphical representation

References

- Statistics on Alcohol: England 2020(Office of National Statistics, 2020)
- Smoking, drinking and drug use among young people in England in 2018 (NHS Information Centre , 2019)
- Wilsnack, R. W., Wilsnack, S. C., Gmel, G., & Kantor, L. W. (2018) Gender differences in binge drinking: Prevalence, predictors, and consequences. *Alcohol Research: Current Reviews* 39(1): 57–76.
- Chaiyasong, S., et al. (2018) Drinking patterns vary by gender, age and country-level income: Cross-country analysis of the International Alcohol Control Study. Drug Alcohol Rev., 37: S53-S62. DOI: https://doi.org/10.1111/dar.12820

Appendix – R screenshots for slide 4
Descriptive statistics of the variables dnnow, Sex, topqual3, marstatc

```
> round(prop.table(table(HSE_2011$dnnow,useNA = "no"))*100,2)

    1     2
78.65 21.35
> round(prop.table(table(HSE_2011$Sex,useNA = "no"))*100,2)

   1    2
45.7 54.3
> round(prop.table(table(HSE_2011$topqual3,useNA = "no"))*100,2)

    1     2     3     4     5     6     7
23.44 11.07 14.57 21.05  4.61  1.48 23.78
> round(prop.table(table(HSE_2011$marstatc,useNA = "no"))*100,2)

    1     2     3     4     5     6     7
18.74 52.29  0.05  2.60  6.90  8.05 11.37
>
```

Variable labels

dnnow 1 | Yes | | 2 | No

Sex 1 | Male | | 2 | Female

topqual3 - 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

marstatc 1 | Single | | 2 | Married | | 3 | Civil partnership including spontaneous answers | | 4 | Separated | | 5 | Divorced | | 6 | Widowed | | 7 | Cohabitees

Appendix – R screenshots for slide 6
Descriptive statistics of the variables HHSize, bmival, Age

```
> describe(HSE_2011$Age)
   vars     n  mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 10617 41.56 23.83     42    41.5 28.17   0 100   100 -0.02    -0.99 0.23
> describe(HSE_2011$bmival)
   vars    n  mean   sd median trimmed  mad  min   max range skew kurtosis   se
X1    1 8376 25.92 6.14  25.59   25.92 5.53 8.34 65.28 56.94 0.56     1.03 0.07
> describe(HSE_2011$HHSize)
   vars     n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 10617 2.85 1.37      3    2.75 1.48   1  10     9 0.83      1.3 0.01
> 
```

To find mode

```
> names(sort(-table(HSE_2011$bmival)))[1]
[1] "13.7670587559799"
> names(sort(-table(HSE_2011$Age)))[1]
[1] "42"
> names(sort(-table(HSE_2011$HHSize)))[1]
[1] "2"
> 
```

Appendix – R screenshots for slides 5,7&8
Bar chart for variable marstatc, boxplot for variable Age and histogram for variable bmival

```
> count<-table(HSE_2011$marstatc,useNA="no")
> barplot(count, main = "Marital status",col="darkblue")
> ?barplot
> barplot(count, main = "Marital status",col="darkblue",xlab="Marital status", ylab="Frequency")
> leg<-c("Single","Married","Civil Partnership","Separated","Divorced","Widowed","Cohabitees")
> barplot(count, main = "Marital status",col="darkblue",xlab="Marital status", ylab="Frequency",legend.
text=leg)
> barplot(count, main = "Marital status",col=c("blue","green","red","yellow","pink","orange","lavende
r"),xlab="Marital status", ylab="Frequency",legend.text=leg)
```

```
> boxplot(HSE_2011$Age, main= "Age at last birthday", ylab="Age in years", col = "pink")
> hist(HSE_2011$bmival, main="Histogram of BMI value", xlab = "BMI value", breaks=30)
```

Appendix – R screenshots for slide 9
Contingency table and chi square test – which gender drinks more alcohol

```
> Gender<-c("Male","Female","Male","Female")
> Status<-c("Yes","Yes","No","No")
> val<-c(3172,3540,605,1217)
> xyz<-data.frame(Gender,Status,val)
> print(xyz)
  Gender Status  val
1   Male    Yes 3172
2 Female    Yes 3540
3   Male     No  605
4 Female     No 1217
> tab.<-xtabs(val~Gender+Status,data = xyz)
> tab.
        Status
Gender     No  Yes
  Female 1217 3540
  Male    605 3172
> sol.chisq<-chisq.test(tab.)
> sol.chisq

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.
X-squared = 114.15, df = 1, p-value < 2.2e-16
```

Appendix – R screenshots for slide 10
Contingency table and chi square test – which region drinks more alcohol

```
> 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
> Region<-c("North East","North West","YorkshirenHummer","East Midlands","West Midlands","East of Engla
nd","London","South East","South West","North East","North West","YorkshirenHummer","East Midlands","We
st Midlands","East of England","London","South East","South West")
> Dstatus<-c("Yes","Yes","Yes","Yes","Yes","Yes","Yes","Yes","Yes","No","No","No","No","No","No","N
o","No","No")
> nval<-c(576,833,686,624,686,763,674,1130,740,135,270,201,136,207,172,304,255,142)
> dframe<-data.frame(Region,Dstatus,nval)
> print(dframe)
             Region Dstatus nval
1        North East     Yes  576
2        North West     Yes  833
3  YorkshirenHummer     Yes  686
4     East Midlands     Yes  624
5     West Midlands     Yes  686
6   East of England     Yes  763
7            London     Yes  674
8        South East     Yes 1130
9        South West     Yes  740
10       North East      No  135
11       North West      No  270
12 YorkshirenHummer      No  201
13    East Midlands      No  136
14    West Midlands      No  207
15  East of England      No  172
16           London      No  304
17       South East      No  255
18       South West      No  142
```

```
> tab.<-xtabs(nval~Region+Dstatus,data=dframe)
> tab.
                  Dstatus
Region               No  Yes
  East Midlands     136  624
  East of England   172  763
  London            304  674
  North East        135  576
  North West        270  833
  South East        255 1130
  South West        142  740
  West Midlands     207  686
  YorkshirenHummer  201  686
> sol.chisq<-chisq.test(tab.)
> sol.chisq

        Pearson's Chi-squared test

data:  tab.
X-squared = 98.53, df = 8, p-value < 2.2e-16
```

Appendix – R screenshots for slide 11
Statistical difference between men and women on height and weight

```
> t.test(htval~Sex,data=HSE_2011,var.equal=TRUE)

        Two Sample t-test

data:  htval by Sex
t = 25.964, df = 8644, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
  9.418226 10.956474
sample estimates:
mean in group 1 mean in group 2
       167.3928        157.2054

> t.test(wtval~Sex,data=HSE_2011,var.equal=TRUE)

        Two Sample t-test

data:  wtval by Sex
t = 18.125, df = 8739, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
  8.479781 10.536397
sample estimates:
mean in group 1 mean in group 2
       74.26612        64.75803

>
```

Appendix – R screenshots for slide 12
Correlation between variables dnnow, totinc, Age and Sex

```
> cor.test(HSE_2011$dnnow,HSE_2011$totinc)

        Pearson's product-moment correlation

data:  HSE_2011$dnnow and HSE_2011$totinc
t = 6.6743, df = 8257, p-value = 2.644e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05176787 0.09467113
sample estimates:
       cor
0.07325339

> cor.test(HSE_2011$dnnow,HSE_2011$Age)

        Pearson's product-moment correlation

data:  HSE_2011$dnnow and HSE_2011$Age
t = 6.3793, df = 8532, p-value = 1.871e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04775254 0.08998509
sample estimates:
       cor
0.06889968

> cor.test(HSE_2011$dnnow,HSE_2011$Sex)

        Pearson's product-moment correlation

data:  HSE_2011$dnnow and HSE_2011$Sex
t = 10.782, df = 8532, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0949586 0.1368223
sample estimates:
     cor
0.115942
```

```
> cor.test(HSE_2011$totinc,HSE_2011$Age)

        Pearson's product-moment correlation

data:  HSE_2011$totinc and HSE_2011$Age
t = 5.0693, df = 10300, p-value = 4.062e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.03060618 0.06913137
sample estimates:
       cor
0.04988733

> cor.test(HSE_2011$totinc,HSE_2011$Sex)

        Pearson's product-moment correlation

data:  HSE_2011$totinc and HSE_2011$Sex
t = 0.48221, df = 10300, p-value = 0.6297
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0145607  0.0240597
sample estimates:
        cor
0.004751272

> cor.test(HSE_2011$Age,HSE_2011$Sex)

        Pearson's product-moment correlation

data:  HSE_2011$Age and HSE_2011$Sex
t = 3.3695, df = 10615, p-value = 0.0007558
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01367304 0.05167641
sample estimates:
       cor
0.03268654
```