StatsProject

Aydan LaCroix and Eric Geisler

2024-04-15

Reading Data

```
library(readr)
library(epitools)
library(vcd)
```

```
## Loading required package: grid

##
## Attaching package: 'vcd'

## The following object is masked from 'package:epitools':
##
##     oddsratio
```

```
library(vcdExtra)
```

```
## Loading required package: gnm

##
## Attaching package: 'vcdExtra'

## The following object is masked from 'package:epitools':
##
##     expand.table
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:vcdExtra':
##
##     summarise
## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
df <- read_csv("Documents/responses.csv")
```

```
## Rows: 1010 Columns: 150

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (11): Smoking, Alcohol, Punctuality, Lying, Internet usage, Gender, Lef...
## dbl (139): Music, Slow songs or fast songs, Dance, Folk, Country, Classical ...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
df = na.omit(df)
head(df)
```

```
## # A tibble: 6 x 150
##   Music ‘Slow songs or fast songs‘ Dance  Folk Country ‘Classical music‘ Musical
##   <dbl>                      <dbl> <dbl> <dbl>   <dbl>             <dbl>   <dbl>
## 1     5                          3     2     1       2                 2       1
## 2     4                          4     2     1       1                 1       2
## 3     5                          5     2     2       3                 4       5
## 4     5                          3     4     3       2                 4       3
## 5     5                          3     2     3       2                 3       3
## 6     5                          5     5     3       1                 2       2
## # i 143 more variables: Pop <dbl>, Rock <dbl>, ‘Metal or Hardrock‘ <dbl>,
## #   Punk <dbl>, ‘Hiphop, Rap‘ <dbl>, ‘Reggae, Ska‘ <dbl>, ‘Swing, Jazz‘ <dbl>,
## #   ‘Rock n roll‘ <dbl>, Alternative <dbl>, Latino <dbl>,
## #   ‘Techno, Trance‘ <dbl>, Opera <dbl>, Movies <dbl>, Horror <dbl>,
## #   Thriller <dbl>, Comedy <dbl>, Romantic <dbl>, ‘Sci-fi‘ <dbl>, War <dbl>,
## #   ‘Fantasy/Fairy tales‘ <dbl>, Animated <dbl>, Documentary <dbl>,
## #   Western <dbl>, Action <dbl>, History <dbl>, Psychology <dbl>, ...
```

Cleaning Data

```
df$"Internet usage" <- ifelse(df$"Internet usage" %in% c("no time at all", "less than an hour a day"),
df$"Internet usage" = ifelse(df$"Internet usage" == "few hours a day", "Moderate Usage", df$"Internet u
df$"Internet usage" = ifelse(df$"Internet usage" == "most of the day", "High Usage", df$"Internet usage
```

```
df$"Happiness in life" = ifelse(df$"Happiness in life" == 1, "Very Unhappy", df$"Happiness in life")
df$"Happiness in life" = ifelse(df$"Happiness in life" == 2, "Unhappy", df$"Happiness in life")
df$"Happiness in life" = ifelse(df$"Happiness in life" == 3, "Neither Happy nor Unhappy", df$"Happiness
df$"Happiness in life" = ifelse(df$"Happiness in life" == 4, "Happy", df$"Happiness in life")
df$"Happiness in life" = ifelse(df$"Happiness in life" == 5, "Very Happy", df$"Happiness in life")
```

```
colnames(df)[colnames(df) == "Happiness in life"] <- "LifeHappiness"
colnames(df)[colnames(df) == "Internet usage"] <- "InternetUsage"
```

```
df$InternetUsage <- factor(df$InternetUsage, levels = c("Low Usage", "Moderate Usage", "High Usage"))
df$LifeHappiness <- factor(df$LifeHappiness, levels = c("Very Unhappy", "Unhappy", "Neither Happy nor Un
```

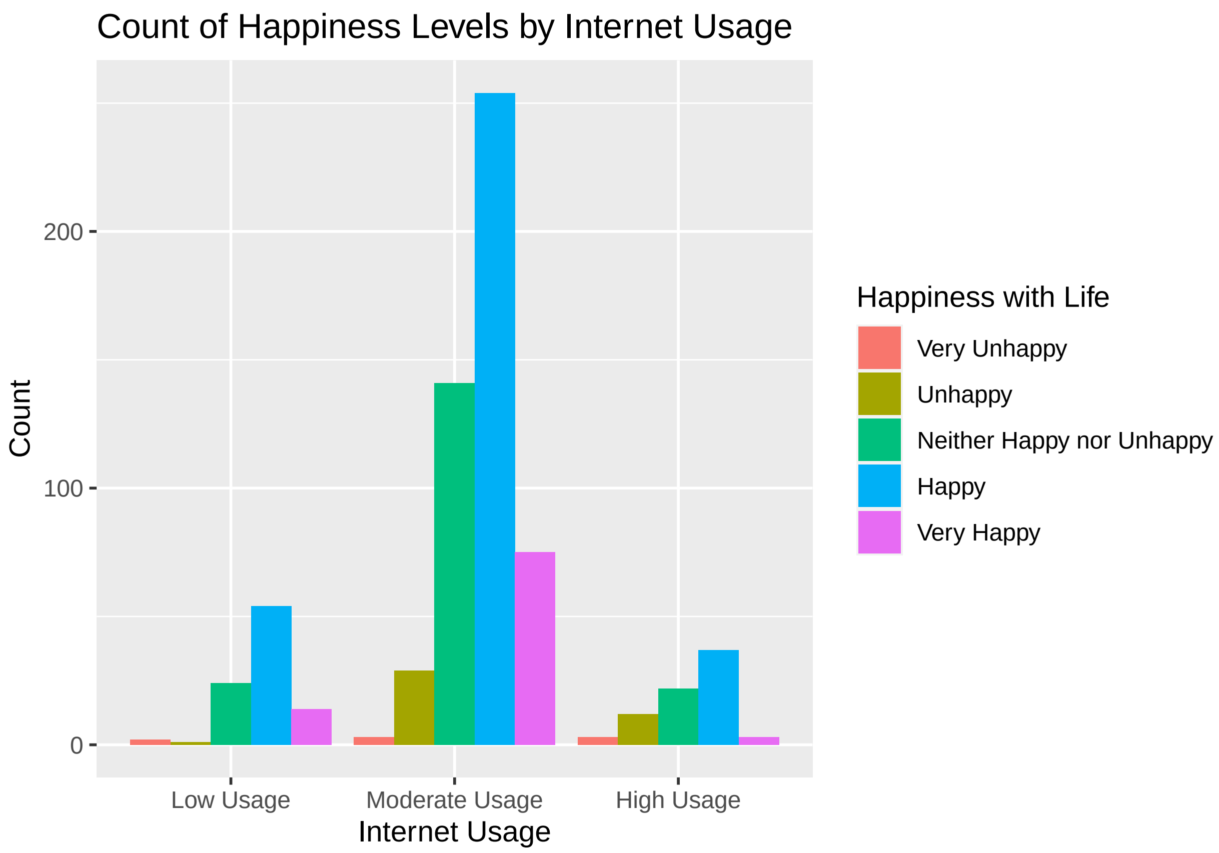
Creating a Contingency Table

```
table = table(df$InternetUsage, df$LifeHappiness)
table
```

```
##
##                  Very Unhappy Unhappy Neither Happy nor Unhappy Happy
##   Low Usage                 2       1                        24    54
##   Moderate Usage            3      29                       141   254
##   High Usage                3      12                        22    37
##
##                  Very Happy
##   Low Usage              14
##   Moderate Usage         75
##   High Usage              3
```

Data Visualization

```
ggplot(df, aes(x = InternetUsage, fill = LifeHappiness)) +
  geom_bar(position = "dodge") +
  labs(x = "Internet Usage", y = "Count", fill = "Happiness with Life") +
  ggtitle("Count of Happiness Levels by Internet Usage")
```



Conducting a CMH Test

```
results = CMHtest(table)
results
```

```
## Cochran-Mantel-Haenszel Statistics for  by
##
##                  AltHypothesis  Chisq Df       Prob
## cor        Nonzero correlation 13.318  1 0.00026289
## rmeans  Row mean scores differ 18.555  2 0.00009350
## cmeans  Col mean scores differ 18.440  4 0.00101226
## general    General association 29.015  8 0.00031524
```

Null Hypothesis: internet usage is independent of happiness with life

Conclusion: At the 0.05 level of significance, there is evidence in these data to conclude that there is a linear trend/association between internet usage and happiness with life, as the p-value of .00026 is below .05.

Analyzing the Residuals

```
psr = chisq.test(table)
```

```
## Warning in chisq.test(table): Chi-squared approximation may be incorrect
```

```
psr$stdres
```

```
## 
##                  Very Unhappy    Unhappy Neither Happy nor Unhappy      Happy
##   Low Usage         0.8917109 -2.2529949                -0.5828643  1.1897341
##   Moderate Usage   -2.4135534 -0.8340403                 0.3396148 -0.5229155
##   High Usage        2.3323871  3.6075637                 0.1721351 -0.5847598
## 
##                  Very Happy
##   Low Usage       0.3329528
##   Moderate Usage  1.6670253
##   High Usage     -2.6488914
```

Subtable 1: Low and Moderate Internet Usage

```
df_low_moderate <- df[df$InternetUsage %in% c("Low Usage", "Moderate Usage"), ]
df_low_moderate$InternetUsage <- factor(df_low_moderate$InternetUsage, levels = c("Low Usage", "Moderat
table_low_moderate <- table(df_low_moderate$InternetUsage, df_low_moderate$LifeHappiness)

table_low_moderate
```

```
## 
##                  Very Unhappy Unhappy Neither Happy nor Unhappy Happy
##   Low Usage                 2       1                        24    54
##   Moderate Usage            3      29                       141   254
## 
##                  Very Happy
##   Low Usage              14
##   Moderate Usage         75
```

```
assocstats(table_low_moderate)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 7.4481  4  0.11402
## Pearson          6.5525  4  0.16151
## 
## Phi-Coefficient   : NA 
## Contingency Coeff.: 0.104 
## Cramer’s V        : 0.105
```

Null Hypothesis: internet usage is independent of happiness with life

Conclusion: At the 0.05 level of significance, there is no evidence in these data to conclude that internet usage is not independent of happiness with life, as the p-value of .11 is above .05.

```
prob_u_low = 3/95
prob_u_moderate = 32/502
odds_u_low = prob_u_low / (1 - prob_u_low)
```

```
odds_u_moderate = prob_u_moderate / (1 - prob_u_moderate)
odds_u_low_mod = odds_u_low / odds_u_moderate
odds_u_low_mod
```

```
## [1] 0.4789402
```

```
prob_h_low = 68/95
prob_h_moderate = 329/502
odds_h_low = prob_h_low / (1 - prob_h_low)
odds_h_moderate = prob_h_moderate / (1 - prob_h_moderate)
odds_h_low_mod = odds_h_low / odds_h_moderate
odds_h_low_mod
```

```
## [1] 1.324327
```

The estimated odds of being unhappy with life while having low internet usage are about .48 times the estimated odds of being unhappy with life while having moderate internet usage.

The estimated odds of being happy with life while having low internet usage are about 1.32 times the estimated odds of being happy with life while having moderate internet usage.

Subtable 2: Moderate and High Internet Usage

```
df_moderate_high <- df[df$InternetUsage %in% c("Moderate Usage", "High Usage"), ]
df_moderate_high$InternetUsage <- factor(df_moderate_high$InternetUsage, levels = c("Moderate Usage", "
table_moderate_high <- table(df_moderate_high$InternetUsage, df_moderate_high$LifeHappiness)

table_moderate_high
```

```
##
##                   Very Unhappy Unhappy Neither Happy nor Unhappy Happy
##   Moderate Usage             3      29                       141   254
##   High Usage                 3      12                        22    37
##
##                   Very Happy
##   Moderate Usage         75
##   High Usage              3
```

```
assocstats(table_moderate_high)
```

```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 19.939  4 0.00051333
## Pearson          22.214  4 0.00018170
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.192
## Cramer's V        : 0.196
```

Null Hypothesis: internet usage is independent of happiness with life

Conclusion: At the 0.05 level of significance, there is evidence in these data to conclude that internet usage is not independent of happiness with life, as the p-value of .0005 is below .05.

```
prob_u_high = 15/77
odds_u_high = prob_u_high / (1 - prob_u_high)
odds_u_mod_high = odds_u_moderate / odds_u_high
odds_u_mod_high
```

```
## [1] 0.2814184
```

```
prob_h_high = 40/77
odds_h_high = prob_h_high/ (1 - prob_h_high)
odds_h_mod_high = odds_h_moderate / odds_h_high
odds_h_mod_high
```

```
## [1] 1.759104
```

The estimated odds of being unhappy with life while having moderate internet usage are about .28 times the estimated odds of being unhappy with life while having high internet usage.

The estimated odds of being happy with life while having moderate internet usage are about 1.75 times the estimated odds of being happy with life while having high internet usage.

Subtable 3: Low and High Usage

```
df_low_high <- df[df$InternetUsage %in% c("Low Usage", "High Usage"), ]
df_low_high$InternetUsage <- factor(df_low_high$InternetUsage, levels = c("Low Usage", "High Usage"))
table_low_high <- table(df_low_high$InternetUsage, df_low_high$LifeHappiness)

table_low_high
```

```
##
##              Very Unhappy Unhappy Neither Happy nor Unhappy Happy Very Happy
##   Low Usage             2       1                        24    54         14
##   High Usage            3      12                        22    37          3
```

```
assocstats(table_low_high)
```

```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 20.290  4 0.00043775
## Pearson          18.204  4 0.00112591
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.309
## Cramer's V        : 0.325
```

Null Hypothesis: internet usage is independent of happiness with life

Conclusion: At the 0.05 level of significance, there is evidence in these data to conclude that internet usage is not independent of happiness with life, as the p-value of .0004 is below .05.

```
odds_u_low_high = odds_u_low / odds_u_high
odds_u_low_high
```

```
## [1] 0.1347826
```

```
odds_h_low_high = odds_h_low / odds_h_high
odds_h_low_high
```

```
## [1] 2.32963
```

The estimated odds of being unhappy with life while having low internet usage are about .13 times the estimated odds of being unhappy with life while having high internet usage.

The estimated odds of being happy with life while having low internet usage are about 2.33 times the estimated odds of being happy with life while having high internet usage.