Education Level vs. Frequency of Reading the Newspaper

Introduction:

Is education level (number of years of school completed) associated with interest in newspapers (number of times newspaper is read per week)? Or are these variables independant of each other? This question becomes more important as newspaper readership has been declining over time. Target marketing to specific individuals based on education level could result in better sales.

In this case education level is the explanatory variable (discreet numerical) and frequency of reading the newspaper the response variable (categorical and ordinal).

Data:

Since 1972, the General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of American society. The GSS aims to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. The vast majority of GSS data is obtained in face-to-face interviews. Computer-assisted personal interviewing (CAPI) began in the 2002 GSS. Under some conditions when it has proved difficult to arrange an in-person interview with a sampled respondent, GSS interviews may be conducted by telephone.

Each case in the GSS data is one interview. This is an observation study, and not an experiment, so the results suggest correlation, but not neccessarily causation. Responses to surveys/interviews simply observe the data without directly interfering with how the data arises.

Since this is a random sample, the results can be generalized to the larger adult population (18-89) of the United States. There may be some non-response bias to the results for those people that could not be reached for interviews, for example people without an address or who move frequently, illegal immigrants, etc. Also, there may be a voluntary response bias by including only those who choose to be interviewed.

The two variables of interest in the GSS data are:

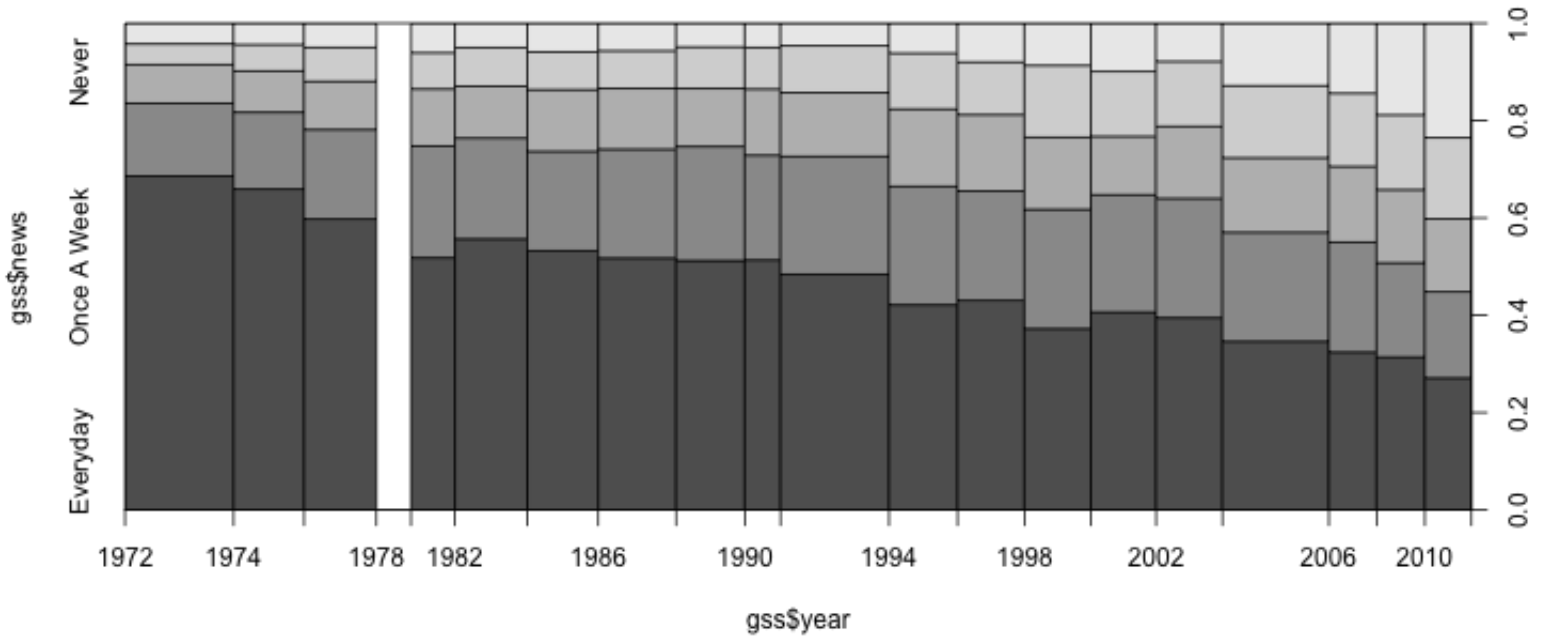
educ: Highest year of school completed. Explanatory variable. Numeric discrete.

news: How often does respondent read newspaper. Response variable. Categorical. Values include: Everyday, Few Times a Week, Once a Week, Less than Once a Week, Never

Exploratory data analysis:

If you look at the trends of newspaper readership over the years since the GSS started in 1972, you can see that the percentage of daily readers has declined and the percentage that never read the paper has increased significantly.

```
plot(gss$news ~ gss$year)
```

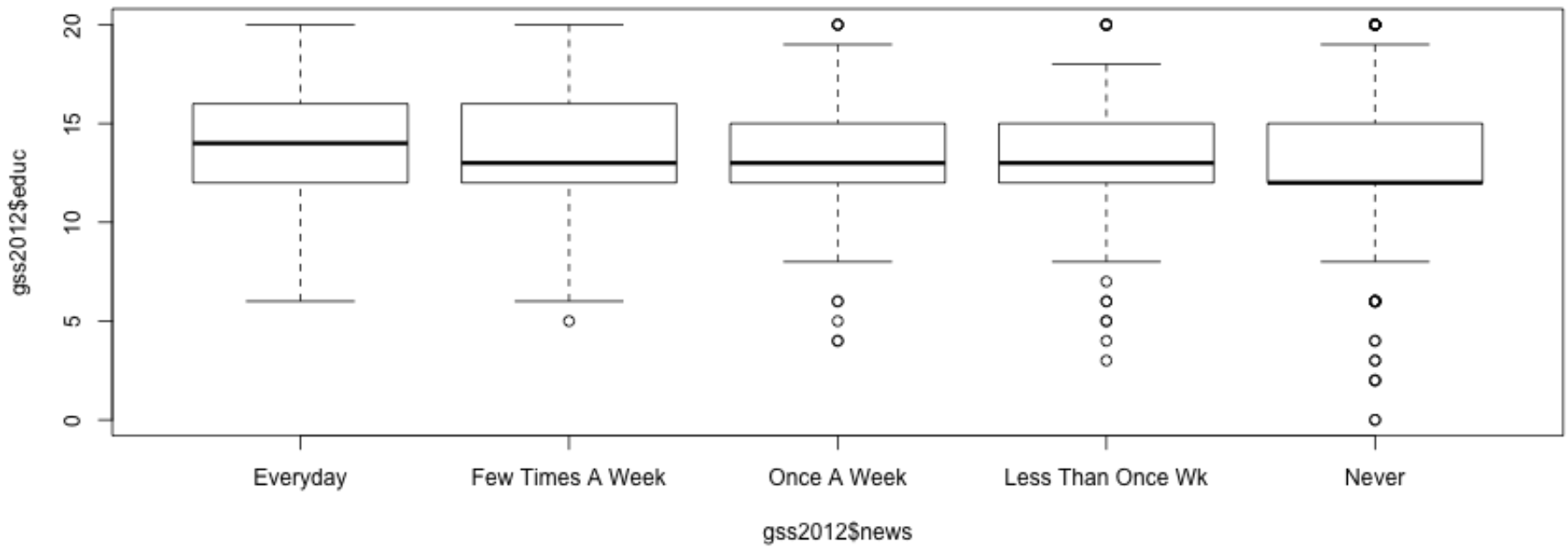


Since the percentages are so different over the years, I decided to focus on the most recent year of data - 2012. I created a subset data frame called gss2012:

```
gss2012 = subset(gss, gss$year == "2012")
```

Looking at the breakdown of education (number of years of schooling completed) vs. how often they read the newspaper, it looks like there could be a positive association:

```
plot(gss2012$educ ~ gss2012$news)
```



Inference:

Since we are comparing one numerical and one categorical varaible (with more than 2 levels), a hypothesis test seems best to compare means across the groups. The null hypothesis (H0) is that all the means are equal to each other, and there is no difference in frequency of newspaper reading based on years of education. The alternate hypothesis (HA) is that at least one mean is different.

Looking at the summary statistics for each grouping of newspaper reading, you can see that the mean years of education values are highest for reading Everyday (14.49) and get progressively smaller for each grouping down to Never (12.84 years of school). The standard deviations vary from 2.78 to 3.47.

```
by(gss2012$educ, gss2012$news, summary)
```

```
## gss2012$news: Everyday
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     6.0    12.0    14.0    14.5    16.0    20.0 
## ------------------------------------------------------------ 
## gss2012$news: Few Times A Week
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     5.0    12.0    13.0    13.8    16.0    20.0 
## ------------------------------------------------------------ 
## gss2012$news: Once A Week
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     4.0    12.0    13.0    13.2    15.0    20.0       1 
## ------------------------------------------------------------ 
## gss2012$news: Less Than Once Wk
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     3.0    12.0    13.0    13.2    15.0    20.0 
## ------------------------------------------------------------ 
## gss2012$news: Never
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    12.0    12.0    12.8    15.0    20.0
```

```
summary(gss2012$news)
```

```
##          Everyday  Few Times A Week       Once A Week Less Than Once Wk 
##               353               230               195               217 
##             Never              NA's 
##               306               673
```

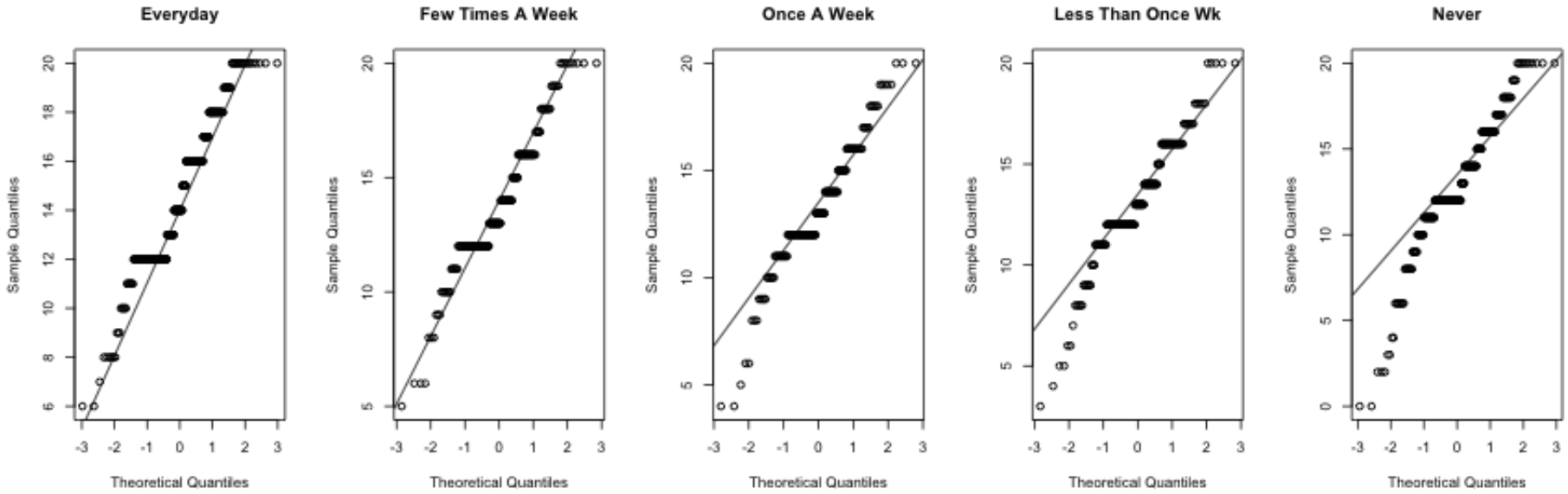
Doing an Analysis of Variance (ANOVA) testing on the means is reasonable since the data meets ANOVO conditions:

1. The observations are independant of each other within groups and between groups since the randomly selected

sample size (n=1974) is less than 10% of the adult population of the US and each respondant is independant of others (no pairing and each group sample size is less than 10% of the population).

2. The variability across groups is about equal - homoscedastic (looking at the box plot)
3. The data within each group is nearly normal, looking at the normal probability plot (QQ distribution):

```
par(mfrow = c(1, 5))
every = subset(gss2012, gss2012$news == "Everyday")
qqnorm(every$educ, main = "Everyday")
qqline(every$educ)
few = subset(gss2012, gss2012$news == "Few Times A Week")
qqnorm(few$educ, main = "Few Times A Week")
qqline(few$educ)
once = subset(gss2012, gss2012$news == "Once A Week")
qqnorm(once$educ, main = "Once A Week")
qqline(once$educ)
less = subset(gss2012, gss2012$news == "Less Than Once Wk")
qqnorm(less$educ, main = "Less Than Once Wk")
qqline(less$educ)
never = subset(gss2012, gss2012$news == "Never")
qqnorm(never$educ, main = "Never")
qqline(never$educ)
```



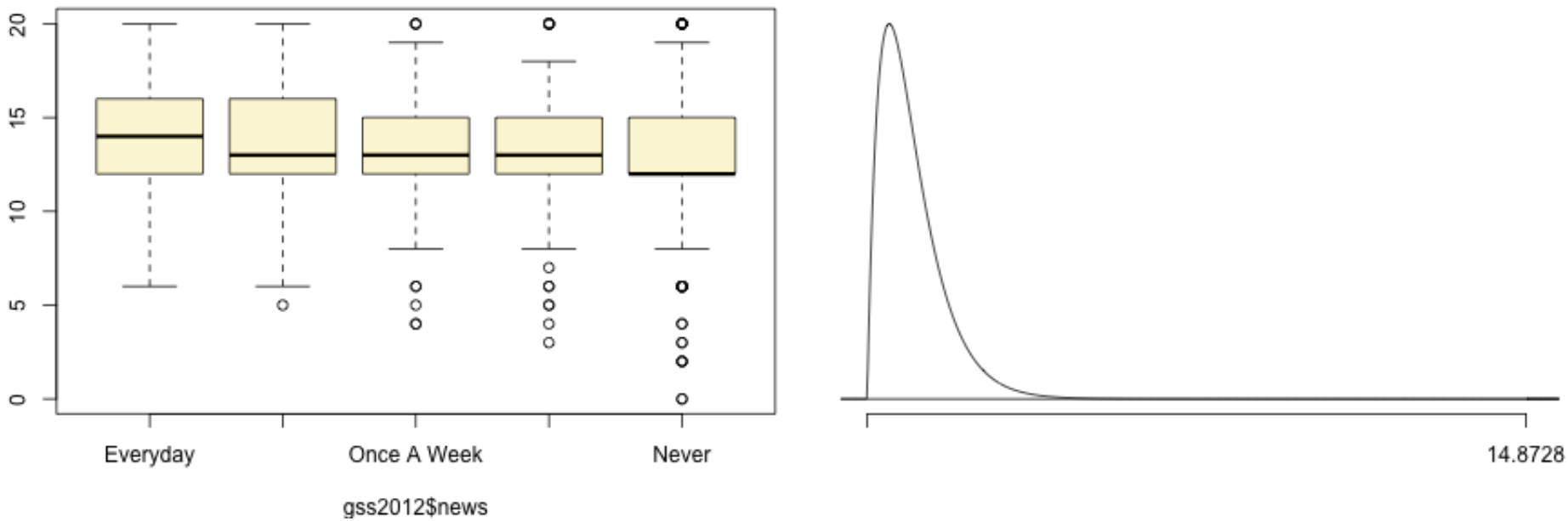
Running the ANOVA tests we get these results:

```
source(url("http://bit.ly/dasi_inference"))
inference(x = gss2012$news, y = gss2012$educ, est = "mean", type = "ht", method
= "theoretical",
    alternative = "greater")
```

```
## Warning: closing unused connection 6 (http://bit.ly/dasi_inference)
```

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
## Summary statistics:
## n_Everyday = 353, mean_Everyday = 14.49, sd_Everyday = 2.907
## n_Few Times A Week = 230, mean_Few Times A Week = 13.8, sd_Few Times A Week =
2.795
## n_Once A Week = 194, mean_Once A Week = 13.22, sd_Once A Week = 2.789
## n_Less Than Once Wk = 217, mean_Less Than Once Wk = 13.17, sd_Less Than Once
Wk = 2.841
## n_Never = 306, mean_Never = 12.84, sd_Never = 3.475
```

```
## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x            4    537     134    14.9 6.9e-12
## Residuals 1295  11692       9
##
## Pairwise tests: t tests with pooled SD
##                   Everyday Few Times A Week Once A Week Less Than Once Wk
## Few Times A Week    0.0066                NA          NA                NA
## Once A Week         0.0000            0.0466          NA                NA
## Less Than Once Wk   0.0000            0.0259      0.8647                NA
## Never               0.0000            0.0002      0.1686            0.2171
```



The p value or probability of all 5 means being equal is very small (6.895e-12) for this F value of 14.873. Therefore the null hypothesis is rejected.

Testing the means of pairs against each other requires a more stringent Bonferroni correction significance level to reduce the probability of a Type 1 error (rejecting the null hypothesis when it is actually true). The modified alpha * = alpha divided by the number of comparisons K (K = 10). In this case alpha * = .05/10 = .005. The standard error for these comparisons is the square root of the overall ANOVA mean square error divided by the sample size of each group = sqrt(MSE/n1 + MSE/n2).

Applying this alpha * significance level of .005 to each pairwise ANOVA calculated p-value can only reject four of the pairwise tests: Everyday with Once a Week, Less than Once Wk, and Never, as well as Few Times A Week with Never. The other tests have higher probability of the means being the same than .005.

Conclusion:

Based on the results of the ANOVA tests, there is a significant difference in means of education level (number of years of school completed) for each category of how often people read the newspaper. Since the overall frequency of newspaper reading has declined over the years since 1972, it is important to target advertizing to those people that would read the

paper most often. For follow-up research, it would be interesting to investigate associations between paper readership and other variables in the GSS.

References:

Data is from the GSS. General Social Survey Cumulative File, 1972-2012 Coursera Extract:

https://d396qusza40orc.cloudfront.net/statistics%2Fproject%2Fgss1.html

Additional information about the GSS:

http://www.norc.org/Research/Projects/Pages/general-social-survey.aspx

http://publicdata.norc.org:41000/gssbeta/faqs.html

Appendix:

```
head(gss2012[c("age", "educ", "news")], n = 40)
```

```
##       age educ              news
## 55088  22   16       Once A Week
## 55089  21   12              <NA>
## 55090  42   12              <NA>
## 55091  49   13  Few Times A Week
## 55092  70   16  Few Times A Week
## 55093  50   19          Everyday
## 55094  35   15             Never
## 55095  24   11              <NA>
## 55096  28    9       Once A Week
## 55097  28   17              <NA>
## 55098  55   10          Everyday
## 55099  36   16             Never
## 55100  28   12              <NA>
## 55101  59   12 Less Than Once Wk
## 55102  52    4       Once A Week
```

```
## 55103  35   13              <NA>
## 55104  36   12          Everyday
## 55105  47   13       Once A Week
## 55106  55   12              <NA>
## 55107  18   12 Less Than Once Wk
## 55108  76    0              <NA>
## 55109  39   10              <NA>
## 55110  54   14       Once A Week
## 55111  45   16 Less Than Once Wk
## 55112  71   12          Everyday
## 55113  42   17              <NA>
## 55114  22   15  Few Times A Week
## 55115  50   10             Never
## 55116  81   16  Few Times A Week
## 55117  44   13          Everyday
## 55118  78   16              <NA>
## 55119  63   14          Everyday
## 55120  73   19          Everyday
## 55121  40   16              <NA>
## 55122  42   14       Once A Week
## 55123  62   18              <NA>
## 55124  52   11             Never
## 55125  49   12  Few Times A Week
## 55126  27   17          Everyday
## 55127  30   14             Never
```