Chisq.test explorations

The question is: can `chisq.test()` be used to compare the frequency distributions of a table of data, and a subset of that table? More specifically, we have a data set that includes columns for "ideology" (Conservative, Moderate, Liberal, Refused, Don't Know), and for "voted" (Biden, Trump, Refused, Someone Else, Don't Know). We want to see if the distribution of "ideology" for the subset of rows for which "voted" = "Refused" (refused to say) is the same as the distribution of "ideology" for all rows, or if it tilts in some direction.

```
# Note: the data file `31118130.csv` is a symlink to the original file in the "../3 + 4" directory in t
raw_data <- read_csv("31118130.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   end_state1 = col_double(),
##   acarot = col_number(),
##   endq1 = col_double(),
##   q2rot = col_logical(),
##   endq2 = col_double(),
##   endq10 = col_double(),
##   q17rot = col_number(),
##   endq22 = col_double(),
##   partyrot = col_number(),
##   length = col_double(),
##   reprostat = col_logical(),
##   qctr = col_double(),
##   xctr = col_double(),
##   wt1 = col_double(),
##   weight = col_double(),
##   standwt = col_double(),
##   weight_ssrs = col_double(),
##   cdc2013 = col_double()
## )
## i Use `spec()` for the full column specifications.
```

First step: get the frequency distribution for "ideology" for the whole set, which is 1676 rows.

```
ideology_count <- table(raw_data$ideology)
ideology_count
```

```
##
## Conservative   Don't Know      Liberal     Moderate      Refused
##          527           68          424          617           40
```

Then extract the subset of rows we want to examine, based on the answer to one question, and get the frequency distribution of that subset, to compare for proportionality with the full set.

```
ideology_by_refused_voted2 <- raw_data %>%
  select(ideology, voted2) %>%
  filter(voted2 == "Refused")
```

```
voted2_refused_count <- table(ideology_by_refused_voted2$ideology)
voted2_refused_count
```

```
##
## Conservative   Don't Know      Liberal     Moderate      Refused
##           38            6           16           37           14
```

Then run a chisq.test on the two distributions. First off: is this comparison valid? The objects we're comparing are `tables`. Not sure if that's an acceptable input type for `chisq.test()`.

```
class(ideology_count)
```

```
## [1] "table"
```

```
class(voted2_refused_count)
```

```
## [1] "table"
```

```
chisq.test(ideology_count,voted2_refused_count)
```

```
## Warning in chisq.test(ideology_count, voted2_refused_count): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  ideology_count and voted2_refused_count
## X-squared = 20, df = 16, p-value = 0.2202
```

The function doesn't complain about its inputs, and returns a p-value of `0.2202`, meaning we can't reject our null hypothesis and the differences in the distributions are explainable by chance. Bu t I don't know if this result is meaningful, because I don't know if the call made sense.

To force it another way, build a manual data set with these numbers, using xtabs to build cross-tabs, manually entering the data from the above tables, which (are? aren't?) the same thing as a frequency distribution:

```
group <- c("all", "all", "all", "all", "all", "refused", "refused", "refused", "refused", "refused")
ideology <- c("Moderate", "Conservative", "Liberal",  "Don't Know", "Refused", "Moderate", "Conservativ
totals <- c(617,527,424,68,40,37,38,16,6,14)
A <- data.frame(group,ideology,totals)
A
```

```
##      group     ideology totals
## 1      all     Moderate    617
## 2      all Conservative    527
## 3      all      Liberal    424
## 4      all   Don't Know     68
## 5      all      Refused     40
## 6  refused     Moderate     37
## 7  refused Conservative     38
## 8  refused      Liberal     16
## 9  refused   Don't Know      6
## 10 refused      Refused     14
```

```
A_xtabs <- xtabs(totals~group+ideology,data=A)
A_xtabs
```

```
##          ideology
## group     Conservative Don't Know Liberal Moderate Refused
##   all              527         68     424      617      40
```

```
##   refused          38          6       16       37       14
```

Now run the `chisq.test()` on this xtab object:

```r
chisq.test(A_xtabs,correct=F)
```

```
## Warning in chisq.test(A_xtabs, correct = F): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  A_xtabs
## X-squared = 42.092, df = 4, p-value = 1.596e-08
```

This gives a very different p-value, showing that the difference in the distributions IS statistically significant and we can reject the null hypothesis.

The input this time is still a table, and also an "xtabs" object:

```r
class(A_xtabs)
```

```
## [1] "xtabs" "table"
```

Finally, run a `chisq.test()` with a manual version of the dataset and a calculated expected values (as ratios) list, based on the subset as the observed values, and the expected values as ratios of the full set values to the size of the set. I woulc expect this test to give the same p-value as the above test:

```r
obs <- c(37,38,16,6,14)
exp <- c(617/1676,527/1676,424/1676,68/1676,40/1676)
chisq.test(obs, p=exp)
```

```
## Warning in chisq.test(obs, p = exp): Chi-squared approximation may be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 54.97, df = 4, p-value = 3.297e-11
```

It returns 3.297e-11… basically zero, but still 3 orders of magnitude off the other very tiny p-value of 1.596e-08. It also complains that the X-squared approximation may be off.

Running the Fisher test gives a different, but still very tiny P-value:

```r
fisher.test(A_xtabs)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  A_xtabs
## p-value = 5.339e-06
## alternative hypothesis: two.sided
```

…what am I doing wrong?