Are Google Suggestions Sexist?

FD

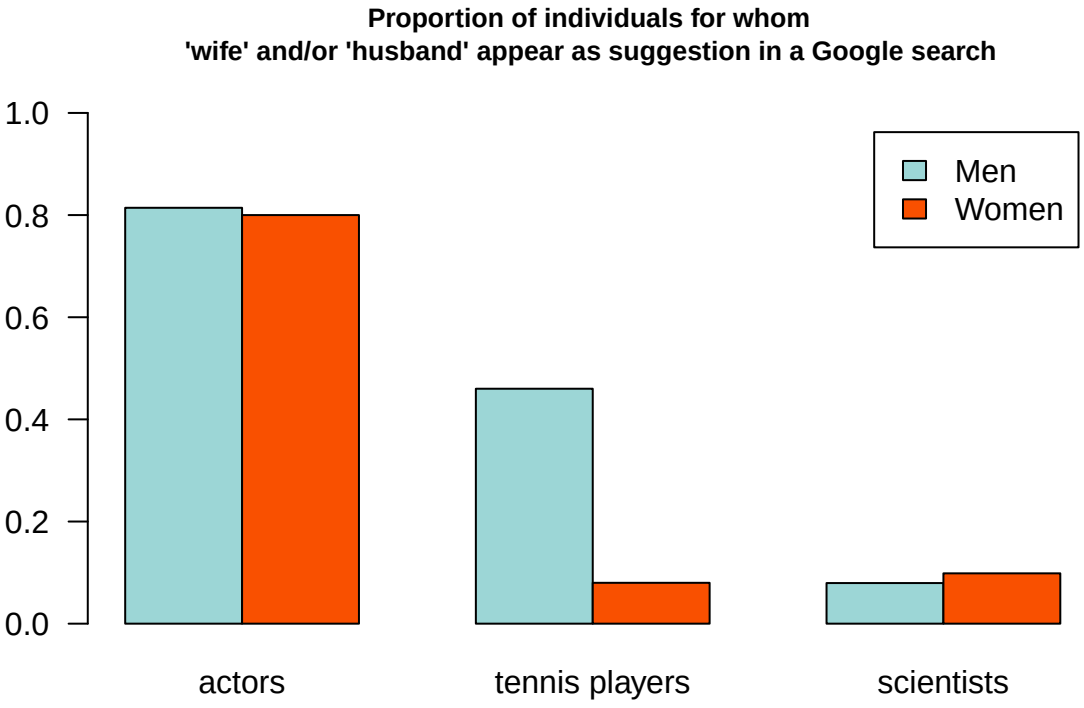
7 June 2016

Google's autocomplete feature sometimes yields disturbing results. When looking for female scientists, you may have noticed the appearance of "husband" in the list of suggestions, as reported recently on Twitter. But does this only happen to female scientists, or instead to any public figure?

To answer this question, I gathered names of high profile people with different professions (scientists, tennis players, Hollywood actors), wrote a script to fetch the list of Google suggestions, and compared the proportion of people for whom "husband" and/or "wife" are suggested as a function of their profession (see the Methods section for more details).

Results

Comparing professions



Actors

There is basically no difference between male (0.81) and female (0.8) actors in terms of frequency of "wife" or "husband" suggestions by Google:

```
summary(glm(formula = wh ~ sex, data=t.actors, family = binomial))
```

```
##
```

```
## Call:
## glm(formula = wh ~ sex, family = binomial, data = t.actors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8350   0.6410   0.6410   0.6410   0.6681
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.47810    0.30735   4.809 1.52e-06 ***
## sex         -0.09181    0.55027  -0.167    0.867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97.245  on 99  degrees of freedom
## Residual deviance: 97.217  on 98  degrees of freedom
## AIC: 101.22
##
## Number of Fisher Scoring iterations: 4
```

Tennis players

There is a strong difference between male and female tennis players, but maybe not the one you'd expect: "husband" or "wife" appear in Google's suggestions for 4 women out of the 50 in the dataset, while they appear for 23 men (out of 50). One reason may be that, for an equivalent ranking, the popularity of female tennis players is lower than the popularity of male tennis players.

```
summary(glm(formula = wh ~ sex, data=t.tennis, family = binomial))
```

```
##
## Call:
## glm(formula = wh ~ sex, family = binomial, data = t.tennis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1101  -1.1101  -0.4084   1.2462   2.2475
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1603     0.2838  -0.565 0.572019
## sex          -2.2820     0.5935  -3.845 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.652  on 99  degrees of freedom
## Residual deviance:  96.871  on 98  degrees of freedom
## AIC: 100.87
##
## Number of Fisher Scoring iterations: 5
```

Scientists

If we now turn to the scientists dataset, we first observe that the frequency at which the "wife" and "husband" suggestions appear is much lower (0.0843373 for men and women combined) than for the actors dataset for instance. What about the difference between men and women in this dataset?

```
summary(glm(formula = wh ~ sex, data=t.science, family = binomial))
```

```
##
## Call:
## glm(formula = wh ~ sex, family = binomial, data = t.science)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4555  -0.4070  -0.4070  -0.4070   2.2503
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.4491     0.1303 -18.798   <2e-16 ***
## sex           0.2356     0.2410   0.978    0.328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 624.17  on 1078  degrees of freedom
## Residual deviance: 623.24  on 1077  degrees of freedom
## AIC: 627.24
##
## Number of Fisher Scoring iterations: 5
```

We do not detect an significant difference in this dataset.

Further dissecting the "science" dataset

The suggestions highlighted on Twitter were mainly about HHMI scientists, so let's focus on this subset of our science dataset for the moment:

```
model.s <- glm(formula = wh ~ as.factor(sex), data=t.science[t.science$HHMI==1,], family = binomial)
summary(model.s)
```

```
##
## Call:
## glm(formula = wh ~ as.factor(sex), family = binomial, data = t.science[t.science$HHMI ==
##     1, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4738  -0.3167  -0.3167  -0.3167   2.4567
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)      -2.9676     0.2417  -12.28   <2e-16 ***
## as.factor(sex)1   0.8373     0.3894    2.15   0.0315 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 224.57  on 480  degrees of freedom
## Residual deviance: 220.24  on 479  degrees of freedom
## AIC: 224.24
##
## Number of Fisher Scoring iterations: 5
```

There is indeed an effect of sex on the proportion of "wife" and "husband" in Google Suggestions for HHMI scientists. But if we look closer, we realize that out of the 12 HHMI female scientists for whom "husband" or "wife" appears in Google's suggestions, 3 have given TED talks, out of the 4 female HHMI scientists who have given TED talks. Out of the 18 male HHMI scientists for whom "husband" or "wife" appears in Google's suggestions, 0 have given TED talks, out of the 1 male HHMI scientists who have given TED talks.

It may be the higher proportion of TED speakers among female HHMI scientists that is driving the result. Here is a summary of the different sample sizes among HHMI scientists:

```
##    ted
## sex  no yes
##   M 367   1
##   W 109   4
```

Now adding participation to a TED talk in the analysis of this HHMI dataset,

```
model.st <- glm(formula = wh ~ as.factor(sex) + TED, data=t.science[t.science$HHMI==1, ], family = bino
summary(model.st)
```

```
##
## Call:
## glm(formula = wh ~ as.factor(sex) + TED, family = binomial, data = t.science[t.science$HHMI ==
##     1, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4128  -0.3128  -0.3128  -0.3128   2.4665
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.9928     0.2434 -12.297   <2e-16 ***
## as.factor(sex)1   0.6409     0.4119   1.556   0.1197
## TED               2.8901     0.9595   3.012   0.0026 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 224.57  on 480  degrees of freedom
## Residual deviance: 211.85  on 478  degrees of freedom
```

```
## AIC: 217.85
##
## Number of Fisher Scoring iterations: 5
```

```
model.i <- glm(formula = wh ~ as.factor(sex) * TED, data=t.science[t.science$HHMI==1, ], family = binom
summary(model.i)
```

```
##
## Call:
## glm(formula = wh ~ as.factor(sex) * TED, family = binomial, data = t.science[t.science$HHMI ==
##     1, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6651  -0.3171  -0.3171  -0.3171   2.4556
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.9647     0.2417 -12.266   <2e-16 ***
## as.factor(sex)1       0.5568     0.4237   1.314    0.189
## TED                 -11.6014   882.7434  -0.013    0.990
## as.factor(sex)1:TED  15.1079   882.7442   0.017    0.986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 224.57  on 480  degrees of freedom
## Residual deviance: 210.27  on 477  degrees of freedom
## AIC: 218.27
##
## Number of Fisher Scoring iterations: 13
```

```
model.t <- glm(formula = wh ~ TED, data=t.science[t.science$HHMI==1, ], family = binomial)
summary(model.t)
```

```
##
## Call:
## glm(formula = wh ~ TED, family = binomial, data = t.science[t.science$HHMI ==
##     1, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3537  -0.3417  -0.3417  -0.3417   2.3956
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8112     0.1981 -14.187  < 2e-16 ***
## TED           3.2167     0.9341   3.443 0.000574 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 224.57  on 480  degrees of freedom
## Residual deviance: 214.13  on 479  degrees of freedom
## AIC: 218.13
##
## Number of Fisher Scoring iterations: 5
```

```
#install.packages("MuMIn")
library(MuMIn, quiet=TRUE)
```

```
## Warning: package 'MuMIn' was built under R version 3.2.3
```

```
model.sel(model.s, model.t, model.st, model.i)
```

```
## Model selection table
##           (Int) as.fct(sex)     TED as.fct(sex):TED df   logLik  AICc
## model.st -2.993           +   2.890                  3 -105.926 217.9
## model.t  -2.811               3.217                  2 -107.063 218.2
## model.i  -2.965           + -11.600                +  4 -105.135 218.4
## model.s  -2.968           +                        2 -110.120 224.3
##          delta weight
## model.st  0.00  0.367
## model.t   0.25  0.324
## model.i   0.45  0.293
## model.s   6.36  0.015
## Models ranked by AICc(x)
```

We will hence go back to the entire science dataset, which includes AAAS (American Academy of Arts and Sciences) members (Section II.4) and all TED speakers of the science section.

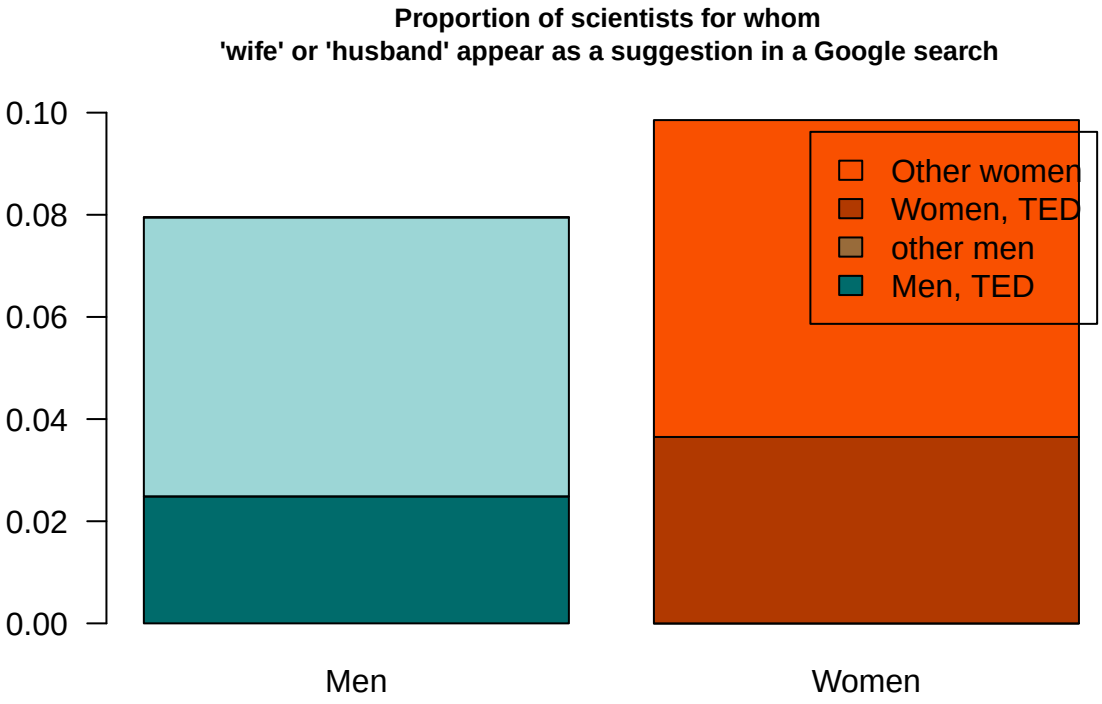
With this entire science dataset, we can explore the interactions between sex and the fact of having given a TED talk (i.e., having been given a wide exposure) on the probability of having "wife" or "husband" appear as a Google suggestion:

```
summary(glm(formula = wh ~ sex+TED, data=t.science[t.science$HHMI==1 | t.science$AAAS==1, ], family = bi
```

```
##
## Call:
## glm(formula = wh ~ sex + TED, family = binomial, data = t.science[t.science$HHMI ==
##     1 | t.science$AAAS == 1, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1851  -0.2751  -0.2751  -0.2751   2.5663
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.2551     0.2313 -14.074  < 2e-16 ***
## sex           0.7385     0.3798   1.945  0.05182 .
## TED           2.5346     0.8038   3.153  0.00161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 262.65  on 665  degrees of freedom
## Residual deviance: 250.01  on 663  degrees of freedom
## AIC: 256.01
##
## Number of Fisher Scoring iterations: 6
```

This analysis suggests that the TED effect is strongest.



Conclusion

Google's suggestions may not be that sexist against female scientists after all: the fact that "wife" and "husband" appear in the list of suggestions of a Google search may instead be a side-effect of being a public figure.

Methods

Name collection

Tennis players

I downloaded on 2016-06-06 the names of the 50 best-ranked male and female tennis players (according to the ATP and WTA rankings). (download the list of names).

Actors

I downloaded on 2016-06-06 a list of the 100 most valuable movie stars, assigned sexes based on first names and pictures. The list contains 30 women names and 70 men names. (download the list of names).

Scientists

I downloaded on 2016-06-06 three datasets:

- The names of all AAAS members in the Evolutionary and Population Biology and Ecology section (38 women, 149 men, download the list of names.);
- The names of all HHMI scientists (113 women, 368 men, download the list of names.);
- The names of all TED speakers in the Science section (127 women, 293 men, download the list of names.).

Some names appear in more than one of these lists. I combined all names in a table, removed duplicates, and indicated with 0s and 1s the affiliations of each scientist, e.g.

```
##                 name sex     cat wife husband wh AAAS HHMI TED
## 183 Edward O. Wilson   0 science    0       0  0    1    0   1
```

for a total of 1079 individual scientists (274 women and 805 men).

Google Suggestions

I wrote a bash script to do this step automatically. Google suggestions for each name in each list were downloaded via the http://suggestqueries.google.com webpage. For each set of suggestions, the words "wife" and "husband" were searched for.

Please note that although I made sure to specify `hl=en` in the query (i.e., setting the language to English), Google Suggestions may differ depending on geographic location. You can do this step again simply by typing

```
./script.sh fileprefix
```

in a terminal, where `fileprefix` is either `tennis`, `actors`, `aaas`, `hhmi`, or `ted`.

Analysis

You just need to look at the source code of this file! It is written in Rmarkdown, so the analysis code is contained in it.

Reproducibility

As mentioned above, the results will depend on where you run the analysis from, since Google suggestions depend on location. To redo everything on the already collected lists of names, simply type

```
./make.sh
```

in a terminal (in the working directory of the project.)