Index

220250711

2023-05-16

Research Question

The aim of the study from which the data was taken (Simmons et al., 2011) was to express that psychology papers contain a lot of false positive results, where analyses are statistically significant but have no real effect. They wanted to prove that by strategically analysing data they could create significance from variables that definitely have no real interaction- that song listened to by participants significantly affected their age, which is obviously impossible. My visualisation aims to address the question of whether choice of covariate in an ancova model has an effect on the significance, by analysing whether song listened to affects participant age or participant's father's age.

Data Origins

The data came from open source repository published by the following study: Simmons, J P, Nelson, L D and Simonsohn, U (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science 22*(11): 1359–1366, DOI: https://doi.org/10.1177/0956797611417632.
Variable names are responses to nuisance questions, such as 'what is your favourite football player', because the original study was just interested in age of participants ('?' column) and what song they listened to ('potato', when64, or 'kalimba'). The data from the article downloads in .txt form, which I copied and pasted into an Excel spreadsheet as it appeared neater in R. I then uploaded the data in Excel form to GitHub, which is where the data is pulled from by R. Here are the first few lines of data before processing:

```
head(s2)
```

```
## # A tibble: 6 x 17
##    aged   dad   mom female  root  bird political quarterback olddays potato
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>     <dbl>       <dbl>   <dbl>  <dbl>
## 1  7097    53    47      0     1     6         2           4      13      0
## 2  6713    47    39      1     1     7         4           2      12      1
## 3  6942    53    51      0     1     5         2           2      13      1
## 4  9938    61    59      1     1     7         1           3      14      0
## 5  7850    53    48      1     1     7         2           2      13      1
## 6  7082    42    43      0     1     7         2           2      13      0
## # i 7 more variables: when64 <dbl>, kalimba <dbl>, feelold <dbl>,
## #   computer <dbl>, diner <dbl>, cond <chr>, aged365 <dbl>
```

Data Preparation

The first step was to remove all the nuisance variables, leaving only the necessary ones for analysis- participant age, father age, potato, kalimba, when64. The song variables were coded as 1 or 0 depending on if they

listened to that song or not. These variables were then recoded to only keep rows which were listened to (1) for each song. Ultimately, this left a dataframe with 3 columns; participant age, father age, and song listened to.

```
head(df1)
```

```
##   dadage   pptage factor_song
## 1     47 18.39178      Potato
## 2     53 19.01918      Potato
## 3     53 21.50685      Potato
## 4     50 20.29589      Potato
## 5     49 19.36986      Potato
## 6     63 21.09589      Potato
```

Statistical analyses

```
## Statistical analyses ##

# Descriptive statistics for participant and father age by song
descriptive <- df1 %>%
  group_by(factor_song) %>%
  summarise(mean_age = mean(pptage),
            sd_age = sd(pptage),
            mean_dad = mean(dadage),
            sd_dad = sd(dadage))
summary(descriptive)
```

```
##   factor_song    mean_age         sd_age         mean_dad         sd_dad
##  Kalimba:1    Min.   :20.34   Min.   :1.089   Min.   :49.89   Min.   :3.727
##  Potato :1    1st Qu.:20.45   1st Qu.:1.650   1st Qu.:50.99   1st Qu.:4.708
##  When   :1    Median :20.57   Median :2.210   Median :52.09   Median :5.689
##               Mean   :20.69   Mean   :1.933   Mean   :52.35   Mean   :5.058
##               3rd Qu.:20.87   3rd Qu.:2.355   3rd Qu.:53.58   3rd Qu.:5.723
##               Max.   :21.17   Max.   :2.499   Max.   :55.07   Max.   :5.757
```

```
# ANCOVA model
# Response variable = participant age
# Group variable = song
# Covariate = father age
ancova_ppt <- aov(pptage ~ factor_song + dadage, data = df1)
ancova_pptage <- Anova(ancova_ppt, type="III")
summary(ancova_ppt)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## factor_song  2   3.63    1.81   0.597 0.55678
## dadage       1  34.26   34.26  11.287 0.00214 **
## Residuals   30  91.06    3.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ancova_pptage)
```

```
##      Sum Sq            Df          F value           Pr(>F)
##  Min.   :13.50   Min.   : 1.0   Min.   : 2.224   Min.   :0.001172
##  1st Qu.:29.07   1st Qu.: 1.0   1st Qu.: 6.755   1st Qu.:0.001655
##  Median :36.66   Median : 1.5   Median :11.287   Median :0.002139
##  Mean   :44.47   Mean   : 8.5   Mean   : 8.792   Mean   :0.043017
##  3rd Qu.:52.05   3rd Qu.: 9.0   3rd Qu.:12.076   3rd Qu.:0.063939
##  Max.   :91.06   Max.   :30.0   Max.   :12.866   Max.   :0.125740
##                                 NA's   :1        NA's   :1
```

```
# Test for Homogeneity
leveneTest(pptage~factor_song, data = df1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.5882 0.5614
##       31
```

```
#   p=0.56, test was not significant so assumption met


# Test for Independence of covariate and group
m1 <- lm(pptage ~ factor_song + dadage, data=df1)
m2 <- lm(pptage ~ factor_song * dadage, data=df1)
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: pptage ~ factor_song + dadage
## Model 2: pptage ~ factor_song * dadage
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 91.058
## 2     28 78.800  2    12.258 2.1779 0.1321
```

```
#   p=0.13, test was not significant so assumption met

# This ANCOVA meets statistical assumptions


# ANCOVA model
# Response variable = father age
# Group variable = song
# Covariate = participant age
ancova_dad <- aov(dadage ~ factor_song + pptage, data = df1)
ancova_dadage <- Anova(ancova_dad, type="III")
summary(ancova_dad)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## factor_song  2  153.9   76.95   3.833 0.03292 *
## pptage       1  226.6  226.56  11.287 0.00214 **
```

```
## Residuals   30  602.2   20.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ancova_dadage)
```

```
##      Sum Sq            Df          F value          Pr(>F)
##  Min.   :124.4   Min.   : 1.0   Min.   : 4.848   Min.   :0.002139
##  1st Qu.:177.1   1st Qu.: 1.0   1st Qu.: 5.524   1st Qu.:0.008562
##  Median :210.6   Median : 1.5   Median : 6.199   Median :0.014985
##  Mean   :286.9   Mean   : 8.5   Mean   : 7.445   Mean   :0.011891
##  3rd Qu.:320.5   3rd Qu.: 9.0   3rd Qu.: 8.743   3rd Qu.:0.016766
##  Max.   :602.2   Max.   :30.0   Max.   :11.287   Max.   :0.018548
##                                 NA's   :1        NA's   :1
```

```
# Test for Homogeneity
leveneTest(dadage~factor_song,data= df1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  1.0191 0.3727
##       31
```

```
#   p=0.37, test was not significant so assumption met


# Test for Independence of covariate and group
n1 <- lm(dadage ~ factor_song + pptage, data=df1)
n2 <- lm(dadage ~ factor_song * pptage, data=df1)
anova(n1, n2)
```

```
## Analysis of Variance Table
##
## Model 1: dadage ~ factor_song + pptage
## Model 2: dadage ~ factor_song * pptage
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 602.17
## 2     28 569.87  2    32.305 0.7936 0.4621
```

```
#   p=0.46, test was not significant so assumption met

# This ANCOVA meets statistical assumptions


# Post Hoc analyses on both ANCOVA models
# Anlayses within group differences for significance
posthoc_ppt <- glht(ancova_ppt, linfct = mcp(factor_song = "Tukey"))
summary(posthoc_ppt)
```

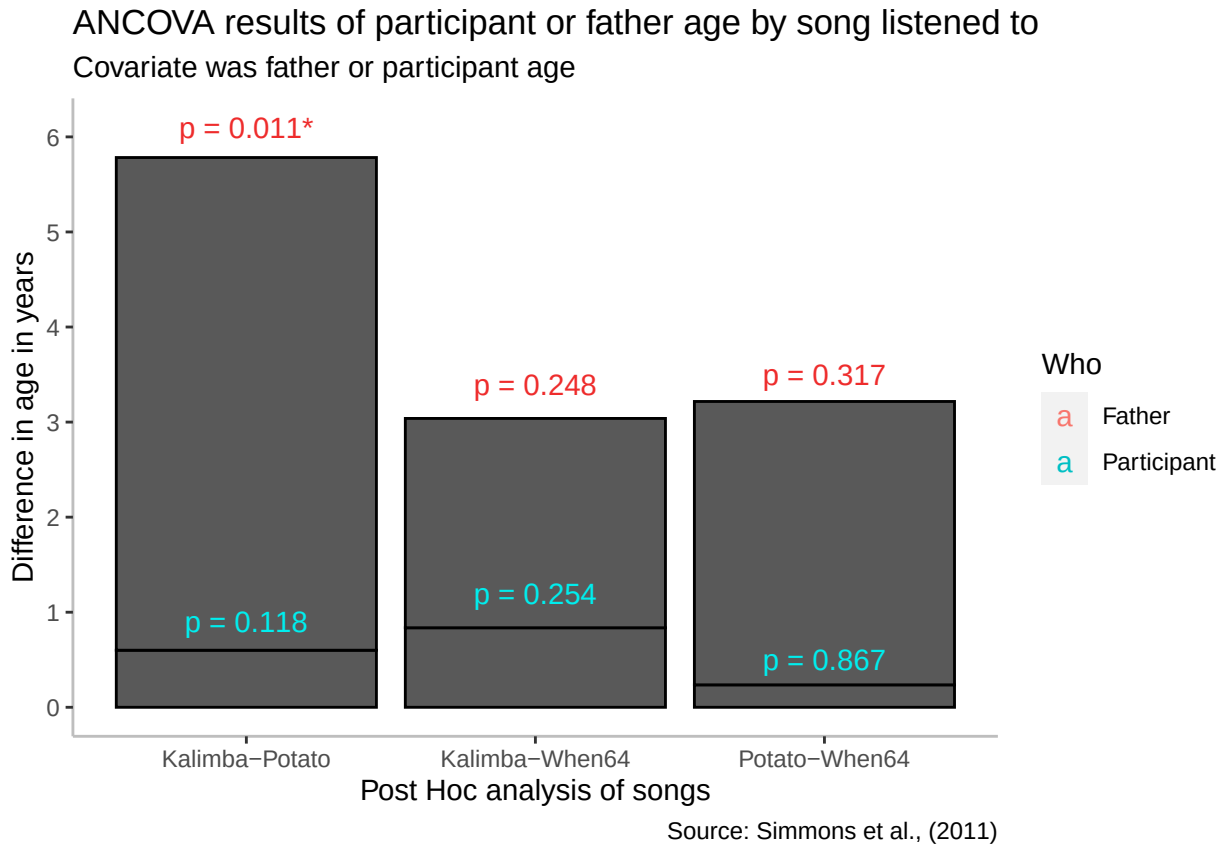
```
##
##   Simultaneous Tests for General Linear Hypotheses
```

```
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = pptage ~ factor_song + dadage, data = df1)
##
## Linear Hypotheses:
##                        Estimate Std. Error t value Pr(>|t|)
## Potato - Kalimba == 0  -1.6540     0.8077  -2.048    0.118
## When - Kalimba == 0    -1.2842     0.7943  -1.617    0.254
## When - Potato == 0      0.3698     0.7248   0.510    0.867
## (Adjusted p values reported -- single-step method)
```

```
posthoc_dad <- glht(ancova_dad, linfct = mcp(factor_song = "Tukey"))
summary(posthoc_dad)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = dadage ~ factor_song + pptage, data = df1)
##
## Linear Hypotheses:
##                        Estimate Std. Error t value Pr(>|t|)
## Potato - Kalimba == 0    5.990      1.929   3.105   0.0112 *
## When - Kalimba == 0      3.327      2.041   1.630   0.2484
## When - Potato == 0      -2.663      1.808  -1.473   0.3171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Visualisation: Difference in age, because of song listened to, or covariate?



Summary

This vizualisation shows that false significance can be created from nonesense variables through careful manipulation of statistical analyses. With more time, the study could have been stretched to include more data, such as from study 1 dataset of the original paper. It is limited in that it only shows data from 42 participants, with each group having an unequal number of participants, but this information is not displayed on the graph. This kind of data being shown would have enriched the vizualisation by further expressing how statistical analyses can hide the meaningful origins of the data. Future research could investigate the effect of unequal group size on post hoc comparisons in ANCOVAs.