R Course: Lesson 2

Paolo Cavatore

19 giugno 2016

```
rm(list=ls())
load("rcourse_lesson2_environment.RData")
```

Introduction

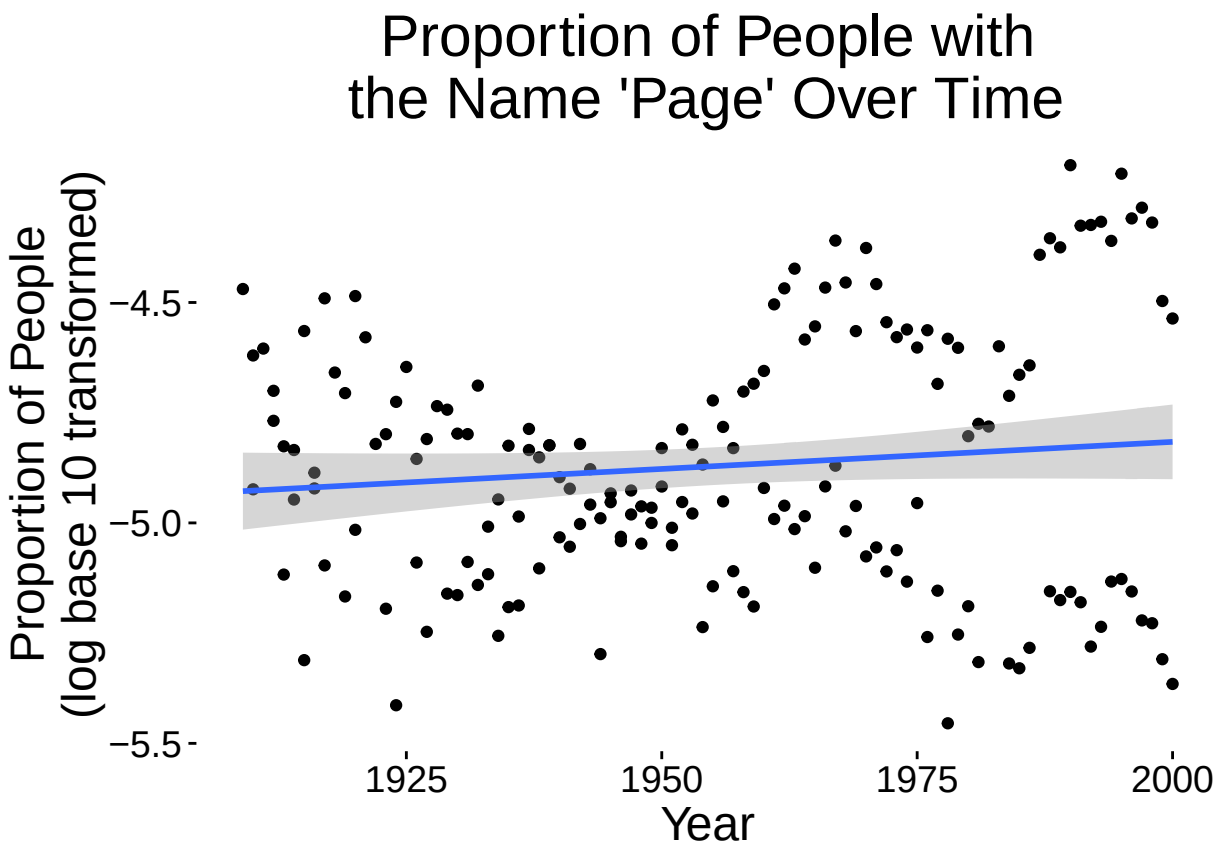
I looked at how common my name, "Page", is in the United States population both by year and sex.

Results

Prevalence by Year

Below is a plot for how the proportion of people with the name "Page" (log base 10 transformed) has changed over time. Overall the trend is pretty flat, with maybe a slight increase over time.

```
year.plot
```



To test if there is a significant effect of year a linear model was built. Proportion of the population log base 10 transformed was the dependent variable and year the independent variable. As shown below, year was not significant, although the coefficients do show a positive slope.

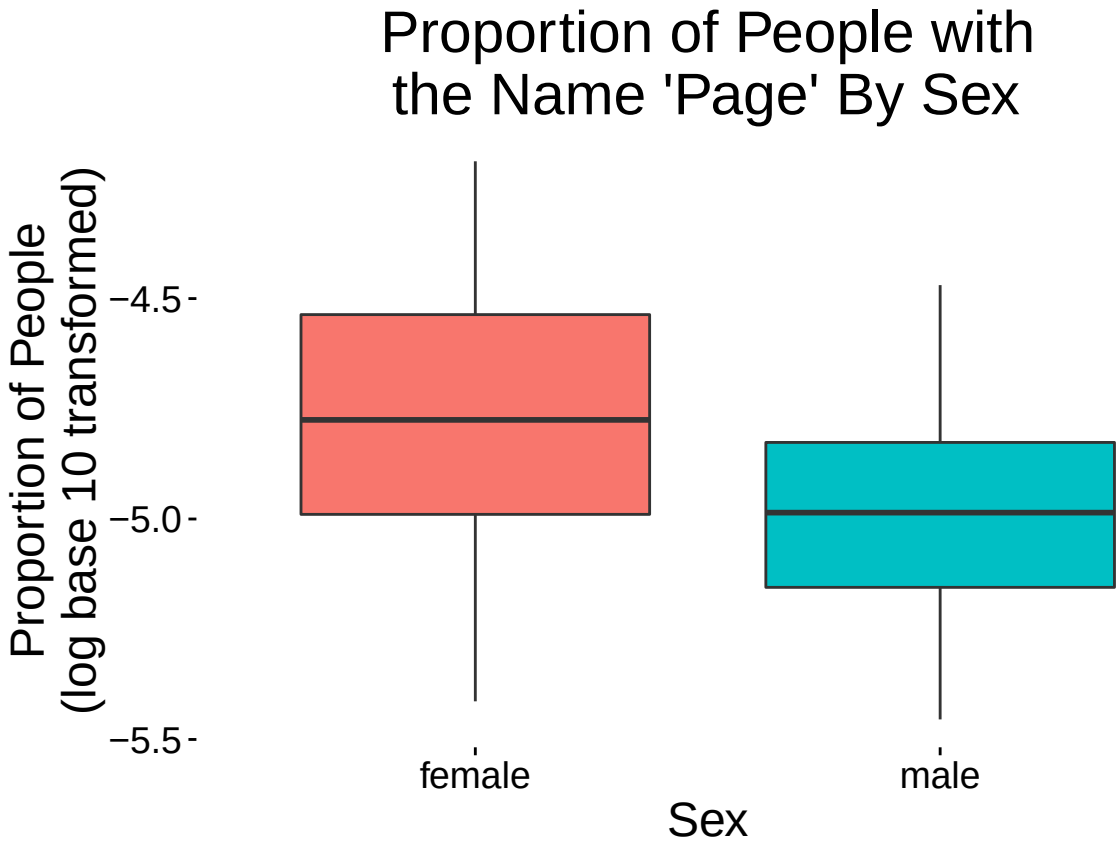
```
year.lm_sum
```

```
##
## Call:
## lm(formula = prop_log10 ~ year, data = data_stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61133 -0.21655 -0.01936  0.21082  0.63981
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.2724137  1.6252827  -4.475 1.39e-05 ***
## year         0.0012281  0.0008311   1.478    0.141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2854 on 172 degrees of freedom
## Multiple R-squared:  0.01254,    Adjusted R-squared:  0.006795
## F-statistic: 2.184 on 1 and 172 DF,  p-value: 0.1413
```

Prevalence by Sex

Below is a plot for the proportion of people with the name "Page" (log base 10 transformed) by sex. It appears there is a clear effect of sex, where it is a less popular name for males than females.

```
sex.plot
```



To test if there is a significant effect of sex a linear model was built. Proportion of the population log base 10 transformed was the dependent variable and sex the independent variable. As show below, sex was significant, with the name being less common in males than females [$R^2 = 0.15$, $F(1, 172) = 32.23$, $p < 0.001$].

```
sex.lm_sum
```

```
##
## Call:
## lm(formula = prop_log10 ~ sex, data = data_stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65869 -0.19693 -0.01146  0.18316  0.56621
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -4.75484    0.02859 -166.302  < 2e-16 ***
## sexM        -0.22695    0.03998   -5.677 5.73e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2636 on 172 degrees of freedom
## Multiple R-squared:  0.1578, Adjusted R-squared:  0.1529
## F-statistic: 32.23 on 1 and 172 DF,  p-value: 5.727e-08
```

Conclusion

The name “Page” has had pretty steady popularity over time but is more common in females than males.